Department of Physics and Astronomy
University of Heidelberg

Master Thesis

in Physics

submitted by

Tom Wolf

born in Ludwigshafen, Germany

2023

A first search for the decay $B^+ \to \mu^+ \nu_\mu e^+ e^-$

This Master Thesis has been carried out by Tom Wolf at the
Physikalisches Institut Heidelberg
under the supervision of
Prof. Dr. Stephanie Hansmann-Menzemer
and
Dr. Martino Borsato
(Milano-Bicocca University and INFN)

Abstract

This thesis presents the first search for the decay $B^+ \to \mu^+ \nu_\mu e^+ e^-$. The dataset was collected by the LHCb experiment in 2016-2018 in proton-proton collisions at a centre-of-mass energy of 13 TeV and correponds to an integrated luminosity of $5.1\,\mathrm{fb}^{-1}$. Due to the decay being rare, the chief obstacle of the analysis is the predominance of background decays. To accommodate this, strict selection requirements are imposed on the dataset and a data-driven approach is used to model the background stemming from misidentification of hadrons ($\pi^\pm$, $K^\pm$). By constructing background-only toy samples, an expected upper limit of the branching ratio of $\mathcal{B}(B^+ \to \mu^+ \nu_\mu e^+ e^-) < 1.22^{+0.36}_{-0.28} \times 10^{-7}$ is found at 95% confidence level.

Zusammenfassung

Diese Arbeit präsentiert die erste Suche nach dem Zerfall $B^+ \to \mu^+ \nu_\mu e^+ e^-$. Der benutzte Datensatz, entsprechend einer integrierten Luminosität von $5.1\,\mathrm{fb}^{-1}$ an Proton-Proton Kollisionen mit einer Schwerpunktsenergie von 13 TeV, wurde in den Jahren 2016-2018 vom LHCb Experiment aufgezeichnet. Aufgrund der Seltenheit des Zerfalls ist das größte Hindernis der Analyse die große Menge an Untergrundzerfällen. Um diese zu bewältigen werden strikte Auswahlbedingungen an den Datensatz gestellt. Weiterhin wird eine Datenbasierte Methode verwendet um den von Hadronmisidentifikation ($\pi^\pm$, $K^\pm$) stammenden Untergrund zu modellieren. Mittels der Konstruktion von Untergrundsimulationen bestimmen wir eine erwartete Obergrenze von $\mathcal{B}(B^+ \to \mu^+ \nu_\mu e^+ e^-) < 1.22^{+0.36}_{-0.28} \times 10^{-7}$ bei einem Vertrauensintervall von 95%.

Contents

Chapter 1

Introduction

Having been developed over several decades in the last century, the Standard Model of particle physics (SM) is a formidable framework encompassing a large part of our current understanding of nature. With the Nobel Prize winning discovery of the Higgs boson in 2012 as capstone [1, 2], its predictions have been extensively tested and experimentally verified. However, we know that it is not yet complete, as a host of unsolved questions remains, the two most famous probably being the integration of gravity and the explanation of dark matter. While its incompleteness is apparent, it is not clear how the SM needs to be amended. Consequently, direct searches for new physics are difficult.

A different approach to this problem is to examine processes that are described by the SM and test its predictions at ever greater precision. For this, the study of b physics is a promising approach. The bottom quark is the heaviest quark that hadronises and the decays of B mesons are expected to be sensitive to virtual contributions of possible new particles. However, B mesons are composite objects made up of quarks that interact through the strong force, which cannot be described perturbatively at low energies. To account for that, the light-cone distribution amplitude of the B meson is defined, a function that encompasses all non-perturbative effects that occur when calculating B meson decays. This function can then be probed experimentally. An important parameter for its characterisation is its first inverse moment, λ_B. So far, only a lower limit of $\lambda_B > 200\,\mathrm{MeV}$ has been set to this parameter by the Belle cooperation [3, 4].

To probe λ_B and in extension the light-cone distribution amplitude, decays of the type $B^+ \to \ell\nu_\ell\gamma^*$ are considered the ideal choice, as they are fully leptonic and thus not affected by strong-interaction effects besides those inherent in the B meson structure. So far, searches for the decays $B^+ \to \ell^+\nu_\ell\gamma$[3, 4] and $B^+ \to \mu^+\nu_\mu\mu^+\mu^-$[5] have been performed, setting upper limits for the respective branching fractions. A first search for the decay $B^+ \to \mu^+\nu_\mu e^+e^-$ is presented in this thesis, using the data set recorded by LHCb during Run 2 of the Large Hadron Collider.

In chapter 2 of this thesis, a brief theoretical background is provided. We introduce the Standard Model, which describes our current understanding of the underlying physics and the concept of light-cone distribution amplitudes, which are what the analysis aims to probe. Following this, chapter 3 gives an overview of the LHCb experiment, which provides the data used in this analysis. In chapter 4, we present the target decay $B^+ \to \mu^+\nu_\mu e^+e^-$ and the analysis strategy. Chapter 5 introduces all data and simulation samples that we use. The imposed selection requirements are explained as well. In chapter 6, we treat simulations of the signal decay. First a reweighting algorithm is employed to adapt to a change in theoretical prediction. Then the signal decay channel is normalised to a reference channel. The expected background decay contributions are described in chapter 7. A boosted decision tree classifier is trained to select against combinatorial background. To model the contribution of misidentified hadron backgrounds, a data-driven approach based on efficiencies gained from a reference sample is used. Further, in chapter 8, the insight gained into the background structure is used to construct simulated samples that reflect our knowledge of the expected data. These toy samples are used to test the stability of the used method. In chapter 9, we furthermore extract an expected upper limit from fits to the simulated samples. Lastly, chapter 10 provides an assessment of the progress achieved and future prospects of the analysis.

Chapter 2

Theoretical background

This chapter will give a short theoretical introduction to the physics necessary to understand the processes touched upon in this analysis. At first, in section 2.1, it will briefly cover our understanding of the fundamental processes and objects of our reality and how we describe them, as summarised in the Standard Model. Then in section 2.2, the concept of light-cone distribution amplitudes is introduced, which describes non-perturbative strong-interaction effects that occur during weak B-meson decays.

2.1 The Standard Model of particle physics

The Standard Model of particle physics is a gauge theory encompassing three of the four fundamental forces of nature in a uniform way. It contains a description of all elemental particles as fields and the allowed interactions between them.

2.1.1 Gauge structure of the Standard Model

Often cited as the most fundamental structure is that of gauge invariance, meaning the invariance of central objects and measurements under a certain class of transformations. The SM is locally gauge invariant with respect to the group

$$SU(3)_C \times SU(2)_L \times U(1)_Y. \tag{2.1}$$

Each part of this group corresponds to a fundamental force, carried by spin 1 vector bosons. Each of these gauge bosons corresponds to a generator of the group. Further, for each group of transformations under which local gauge invariance holds, there is a conserved charge in the theory.

Out of $SU(3)_C$ comes the strong interaction, conserving colour charge C. Accordingly, the theory describing it is called Quantum Chromodynamics (QCD). Mediators of the strong force are the eight massless gluons. The gluons themselves also carry colour charge, allowing gluon-gluon interaction. An important trait of the strong force is the way its coupling strength changes with the energy scale. As energy increases (or distance decreases), it asymptotically approaches 0, meaning colour charged particles appear as free particles (asymptotic freedom). Conversely, it grows very large for small energies (or large distances). Related to this is the phenomenon of colour confinement, stating that all free objects in nature are of neutral colour, or alternatively that it is impossible to isolate a colour charged particle such as an individual quark or gluon. The strong interaction affects only quarks and causes them to form hadrons.

The electroweak part of the SM is based on $SU(2)_L \times U(1)_Y$. Here the L denotes that the gauge bosons corresponding to the $SU(2)_L$ generators only couple to particles of left-handed chirality, which corresponds to the conservation of the weak isospin $\mathbf{T}$. The weak hypercharge, Y, is conserved by $U(1)_Y$. In contrast to $SU(3)_C$, it is not possible to disentangle $SU(2)_L$ and $U(1)_Y$. The three weak gauge bosons, $W^\pm$ and Z^0, and the electromagnetic gauge boson, the photon (often denoted just as γ), are not corresponding to the gauge bosons at generator level. Instead they are linear combinations of them, with mixing between the neutral gauge bosons resulting in the Z^0 and the photon.

Famously, the $SU(2)_L \times U(1)_Y$ symmetry is spontaneously broken to $U(1)_Q$, causing the fermions and the weak gauge bosons to be massive. Instead of weak isospin and hypercharge, only their sum $Q = T_3 + \frac{1}{2}Y$ is conserved. This is the electric charge. As a result of the mass of the weak gauge bosons, the weak interaction is strongly suppressed at low energies or large distances.

2.1.2 Particle content of the Standard Model

While fields are the fundamental objects of the standard model, particles are their incarnations that can actually be measured and searched for. Particles are classified either as fermions if they have half-integer spin or bosons if their spin is an integer. Tab. 2.1 gives an overview of the fundamental bosons. As explained above, the gauge structure of the SM implies the existence of *exactly* one gauge boson for each generator. All of the gauge

Table 2.1: The fundamental bosons of the Standard model. Masses taken from [6]

Particle:	g	$W^\pm$	Z^0	γ	H^0
Force	strong	weak	weak	e.m.	-
Charge Q [e]	0	±1	0	0	0
Mass [GeV][1]	0	80.4	91.2	0	125.3
Spin	1	1	1	1	0
Weak Isospin T_3	0	±1	0	0	-1/2

Table 2.2: The fermions of the Standard Model. Each of them further has a corresponding antiparticle. Masses taken from [6].

Generation	1st		2nd		3rd	
Lepton	ν_e	e^-	ν_μ	μ^-	ν_τ	τ^-
Mass [MeV]	$< 10^{-6}$	0.511	$< 10^{-6}$	105.7	$< 10^{-6}$	1,777
Q [e]	0	-1	0	-1	0	-1
T_3	1/2	−1/2	1/2	−1/2	1/2	−1/2
Quark	u	d	c	s	t	b
Mass [MeV]	2.2	4.7	1,270	93.4	172,700	4,200
Q [e]	2/3	−1/3	2/3	−1/3	2/3	-1/3
T_3	1/2	−1/2	1/2	−1/2	1/2	−1/2

bosons are vector bosons, having a spin of 1. Additionally, there is the Higgs, responsible for the fermion mass generation, a scalar boson with 0 spin.

Besides the gauge bosons mediating the interactions, there are also fundamental fermions. Historically the first of these to be discovered was the electron, a very light charged lepton. In addition to that there are the up and down quarks that form protons and neutrons, which, together with the electron, make up all stable matter[2] Further, the electron neutrino can be inferred from beta decay. These four particles make up the first generation of fermions. The two leptons, the electron and the electron neutrino, are in a weak isospin duplet and so are the two quarks.

In principle these particles and their respective antiparticles would be perfectly sufficient for the SM to work. However, experiment has found two further generations of fermions. Each generation has the same structure as the first, with identical charges and isospin. The only difference is the mass, with a higher generation particle being heavier than its lower generation counterpart.

An overview of the fundamental fermions is given in Tab. 2.2. As mentioned, for each fermion there is also a corresponding antiparticle[3] sharing all features, but with inverted electric charge.

2.1.3 The Lagrangian of the Standard Model

As a quantum field theory, the objects at the heart of the SM are fields that can be excited in discrete quanta. These excitations are what we refer to as particles, their qualities being defined from the equations and rules that their respective fields are obeying.

All of this can be encapsulated in one compact object, the Lagrangian density of the Standard Model, $\mathscr{L}$. When the Standard Model is declared locally gauge invariant, it is

[2]The structure of a hadron is actually not that simple. It is understood to be a dynamic object, gaining a lot of its mass from gluons.

[3]The possible exceptions are neutrinos which might be their own antiparticles.

invariance of this object that is meant. A complete discussion of the Lagrangian density is outside the scope of this work and can be found in dedicated textbooks such as [7].

To illustrate the general idea, we will consider as an example the case of quarks. Quarks are described by fermionic fields. In one generation, n, of quarks, four fields are necessary for that,

$$\text{one L-doublet : } \quad \begin{pmatrix} u_n \\ d_n \end{pmatrix}_L \equiv q_L^n \quad \text{and two R-singlets:} \quad u_R^n, d_R^n. \tag{2.2}$$

For both the u type and the d type quark, an L-chiral and an R-chiral field exist. The L-chiral fields are grouped in a doublet, while the R-chiral fields form singlets, maximally violating parity.

The dynamics fields are described by the terms in the Lagrangian that contain them. A first set describes the quarks interactions with the gauge bosons

$$\mathscr{L}^n_{\text{quark,gauge}} = i\bar{q}_L^n D^\mu \gamma_\mu q_L^n + i\bar{u}_R^n D^\mu \gamma_\mu u_R^n + i\bar{d}_R^n D^\mu \gamma_\mu d_R^n. \tag{2.3}$$

Here, γ_μ are the Dirac Matrices and D^μ is the covariant derivative, defined as

$$D^\mu = \partial^\mu + ig_s G_a^\mu T^a + igW_b^\mu \tau^b + ig' B^\mu Y. \tag{2.4}$$

Quarks interact with the gluon field, G_a^μ, and the electroweak boson fields, W_b^μ and B^μ. The respective interaction strength is determined by the coupling constants of the strong and electroweak force, g_s, g and g'. To assure correct transformation under the symmetry groups $SU(3)_c$, $SU(2)_L$ and $U(1)_Y$, their generators, T^a, τ^b and Y are included. Identity matrices in $SU(2)$ and $SU(3)$ space have been omitted for clarity.

The terms in 2.3 by themselves are perfectly sufficient to describe a particle that is charged under all interactions and propagates at the speed of light. However, as quarks are massive, a second set of terms is present in the Lagrangian, describing the generation of quark masses by coupling with the Higgs doublet,

$$\mathscr{L}^n_{\text{quark,Yukawa}} = -\sum_{n=1}^{3} \left[\bar{q}_L^n \Gamma_u^{nk} \phi^* u_R^k + \bar{q}_L^n \Gamma_u^{nk} \phi d_R^k \right] + h.c. \tag{2.5}$$

In this expression, $\phi = (h^+, v+h^0)^T$ denotes the Higgs doublet. Its presence both describes interactions between quarks and the Higgs boson and the generation of quark masses by the Higgs vacuum expectation value v. Crucially, the latter involves the 3×3 Yukawa coupling matrices Γ_q^{nk}. These matrices are in general not diagonal.

Similarly, the electroweak gauge bosons couple to the Higgs. However, the mass terms do not correspond to the $SU(2)_L \times U(1)_Y$ gauge fields, but to a linear combination of them. These linear combinations are the electroweak gauge bosons found in nature, the charged $W^{\pm}$ bosons, the neutral Z^0 boson and the photon, γ.

$$\begin{aligned} W^{\pm} &= (W_1 + W_2)/\sqrt{2} \\ Z^0 &= \cos(\theta_W) W_3 - \sin(\theta_W) B \\ \gamma &= \sin(\theta_W) W_3 + \cos(\theta_W) B \end{aligned} \tag{2.6}$$

While the choice of basis for the quark fields is free, we find that the mass Eigenstates do not correspond to the Eigenstates of the weak interaction, called the flavour Eigenstates. They are related by unitary transformations, $\mathbf{U}_{u/d}$, defined as

$$\mathbf{u}_{\text{flavour}} = \mathbf{U}_u \mathbf{u}_{\text{mass}} \quad \text{and} \quad \mathbf{d}_{\text{flavour}} = \mathbf{U}_d \mathbf{d}_{\text{mass}}. \tag{2.7}$$

Here we used $\mathbf{u}$ and $\mathbf{d}$ to denote the three vector of up type quarks, $(u, c, t)^T$, and down type quarks, $(d, s, b)^T$. With this, the terms of the Lagrangian density describing the interactions between the quarks and the $W^{\pm}$ mesons are

$$\mathscr{L}_{\text{quarks},W} = -\frac{g}{\sqrt{2}} \left[\bar{d}_L \gamma^\mu U_d^\dagger U_u u_L W_\mu^- + \bar{u}_L \gamma^\mu U_u U_d^\dagger d_L W_\mu^+ \right]. \tag{2.8}$$

We find that the $W^{\pm}$ allows up type quarks to transition into down type quarks and vice versa, even if they are in different generations. The object governing this change of flavour is the Cabbibo-Kobayashi-Maskawa (CKM), matrix

$$\mathbf{V}_{CKM} \equiv \mathbf{U}_u^\dagger \mathbf{U}_d. \tag{2.9}$$

The CKM matrix is unitary and has four free parameters that need to be determined experimentally. They are three angles and a phase that allows CP-Violation. For the moduli of the entries, the current global fit results are [8]

$$|\mathbf{V}_{CKM}| = \begin{pmatrix} |V_{ud}| & |V_{us}| & |V_{ub}| \\ |V_{cd}| & |V_{cs}| & |V_{cb}| \\ |V_{td}| & |V_{ts}| & |V_{tb}| \end{pmatrix} \approx \begin{pmatrix} 0.974 & 0.225 & 0.004 \\ 0.225 & 0.974 & 0.041 \\ 0.009 & 0.041 & 1.000 \end{pmatrix}. \tag{2.10}$$

2.2 Light-cone distribution amplitudes

Due to the running of α_s, perturbation theory can describe the strong interaction only in a limited capacity, namely at high energy scales. This necessitates alternative theoretic approaches, depending on the considered process.

Of interest to us are decays of B mesons through the weak interaction. The weak transition itself can be computed using perturbation theory, but B-mesons are dynamic composite particles, which are described by QCD. As an example, one can express the amplitude for the decay of a B meson into two mesons, M_1, M_2, as

$$\mathcal{A}(B \to M_1 M_2) = \frac{GF}{\sqrt{2}} \sum_i \lambda_i C_(\mu) \langle M_1 M_2 | \mathcal{O}_i | B \rangle (\mu). \tag{2.11}$$

where G_F is the Fermi constant, λ_i a CKM factor and $C_i(\mu)$ a coefficient function incorporating QCD effects above the scale $\mu \sim m_b$. The central objects are transition matrix elements of local operators in the weak effective Hamiltonian, $\mathcal{O}_i$, which capture non-perturbative effects and need to be treated.

One way to go about this is the QCD factorisation approach [9]. As a starting point, the heavy quark limit is adopted. With a large mass imbalance between the two constituent quarks, the B meson velocity is nearly equal to that of the b quark. For our considerations, it is fixed to be equal, i.e. the b quark is at rest in the B-meson rest frame.

In the decay, the b quark undergoes a weak transition, while the second constituent quark of the B meson does not interact weakly, hence it is termed spectator quark. However, additional strong interactions between the quarks are possible. Hard gluon exchanges are expected. We split up the transition matrix element of the operator $\mathcal{O}_i$, based on whether this interaction involves the spectator quark or not,

$$\begin{aligned} \langle M_1 M_2 | \mathcal{O}_i | \bar{B} \rangle = & \textstyle\sum_j \quad F_j^{B \to M_1}(m_2^2) \int_0^1 du\, T_{ij}^I(u) \phi_{M_2}(u) + (M_1 \leftrightarrow M_2) \\ + & \int_0^1 d\xi\, du\, dv\, T_i^{II}(\xi, u, v) \phi_B(\xi) \phi_{M_1}(v) \phi_{M_2}(u) \end{aligned} \tag{2.12}$$

In this expression, $F_j^{B \to M_{1,2}}(m_{2,1}^2)$ denotes a form factor and $m_{1,2}$ the masses of the light mesons. The hard scattering functions $T_{ij}^I(u)$ and $T_i^{II}(\xi, u, v)$ encompass the perturbatively calculable contributions to the expression, such as the weak transition and high-energy strong interactions. They are dependent on the fractions of the meson momenta carried by the light quarks, ξ, u and v. Non-perturbative QCD effects due to the internal meson dynamics are included in the light-cone distribution amplitudes $\phi_X(u)$. They

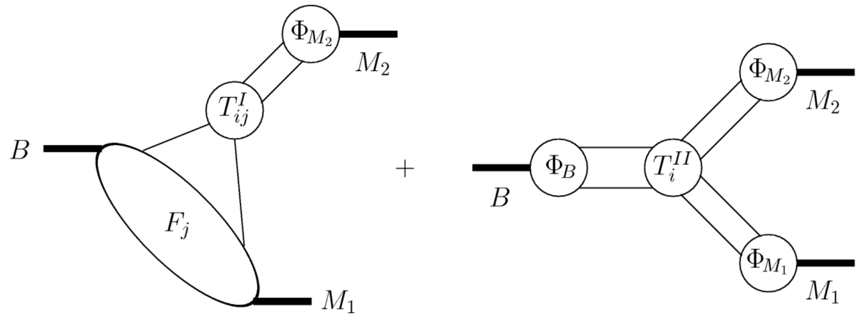


Figure 2.1: Graphical illustration of the factorisation formula 2.12. Note that for clarity, only one form factor term is depicted. Adapted from [9].

can be understood to be like parton distribution functions, describing the density of a wave function in dependence of its fraction of total hadron momentum. Both light-meson light-cone distribution amplitudes, $\phi_{M_{1,2}}(u)$, and form factors $F_j^{B\to M_{1,2}}$, can be predicted by lattice QCD. However, $\phi_B(u)$ cannot be accessed by theoretic methods and needs to be determined experimentally. The second form factor term is only necessary for decays in which the B-meson spectator-quark can end up in either meson, such as $B^+ \to \pi^0 K^+$. For a decay in which this is not the case, e.g. $B^0 \to \pi^- K^+$, the term is not needed.

In Fig. 2.1, an illustration of the factorisation is given. The image on the left corresponds to the form-factor term in equation 2.12, which describes the contribution in which the hard gluon is exchanged between the b quark and the emitted meson, M_2. On the right, the other case is sketched. The gluon is exchanged between the emitted and the light spectator quark. It thus additionally probes the structure of the B meson and the other final state meson, M_1.

The central parameter to describe the B-meson light-cone distribution amplitude is its inverse first moment, λ_B [10], defined as

$$\frac{1}{\lambda_B} = \int_0^\infty d\omega \frac{\phi_+^B(\omega)}{\omega}. \tag{2.13}$$

In this expression, ω denotes the energy of the light quark and $\phi_+^B(\omega)$ describes the quark wave-function inside the B meson. It only depends on one parameter, because the total available energy is fixed by the B-meson rest-mass. The subscript + denotes that the wave function has been transformed to light-cone coordinates. Accordingly, ω is understood to be the light-cone projection of the light quark momentum in the B-meson rest frame.

At the moment of writing, only a limit of $\lambda_B > 240\,\mathrm{MeV}$ at 90% confidence level has been established [3]. A more precise determination of this parameter would allow to use exclusive $b \to u\ell\nu_\ell$ decays, such as $B \to \pi\ell\nu_\ell$, to determine the modulus of the CKM matrix element V_{ub}.

Chapter 3

The LHCb experiment

This chapter provides an overview of the LHCb experiment. It does not aim to be a treatise on technical details, but rather convey the experimental basis of this analysis. It starts in section 3.1 with the Large Hadron Collider which provides proton-proton collisions, that are used to probe b and charm decays by the LHCb experiment. Section 3.2 covers the LHCb detector, which measures various properties of the decays. Section 3.2.3 in turn describes how the relevant part of the measured information is extracted out of an otherwise overwhelming amount of data.

3.1 The Large Hadron Collider and LHCb

The **L**arge **H**adron **C**ollider (LHC) is the most powerful particle collider that has been built to date. It is situated at the European Organization for Nuclear Research, CERN, formerly **C**onseil **E**uropéen pour la **R**echerche **N**ucléaire, close to Geneva. A circular accelerator, installed in the former LEP tunnel, it has a circumference of 26.7 km. The LHC is designed [11] to operate at centre-of-mass energies of up to $\sqrt{s} = 14\,\text{TeV}$ and is able to deliver a peak luminosity of $L = 10^{34}\,\text{MeV}^{-2}\text{s}^{-1}$ of proton-proton collisions.

During a first phase of operation called Run 1 in the years 2010-2012 the LHC was running at a centre-of-mass energy of $\sqrt{s} = 7\,\text{TeV}$ (2010-2011) and $\sqrt{s} = 8\,\text{TeV}$ (2012) [12], which was increased to $\sqrt{s} = 13\,\text{TeV}$ for Run 2 in the years 2015-2018. Run 3 started in 2022, with the LHC now operating close to the design energy of 13.5 TeV. The centre-of-mass energy can be thought of as governing what is likely to happen in any individual collision, as important quantities such as cross sections and formfactors scale with it. Further, it is also the ”energy budget” that must be adhered to by particle creation and particle kinematics.

Luminosity can be thought of as rate of particle interactions induced by the accelerator. It governs the number of events, N_{events}, of a given process, expected to be produced in a given amount of time.

$$N_{events} = \sigma_{event}(s) \times \int L(t)\, dt \tag{3.1}$$

with $\sigma_{event}(s)$ being the cross section of a given type of event at the centre-of-mass energy. The LHC reaches its high luminosity by having 2808 bunches of $\sim 10^{11}$ protons circulate in both opposing directions, with a nominal spacing of 25 ns between bunches.

The unique possibilities given by the LHC are used by a multitude of experiments. However, there are four main experiments situated at proton-proton interactions points: ATLAS, CMS, ALICE and LHCb. This analysis is based on data that was recorded during Run 2 in the years 2016-2018 at LHCb.

LHCb is a dedicated heavy flavour physics experiment [13], meaning it is focused on measurements concerning physical processes involving bottom and charm quarks. LHCb distinguishes itself among other experiments in this category by its singularly enormous number of decays, thanks to the LHC's high luminosity and a large cross section due to the high centre-of-mass energy. However, with these advantages uniquely provided by the LHC also comes the drawback of the highly energetic hadronic collisions providing more background.

An illustration of the LHCb detector can be seen in Fig. 3.1. Placed directly around the collision point, the Vertex locator (VELO) detects charged tracks and pinpoints pri-

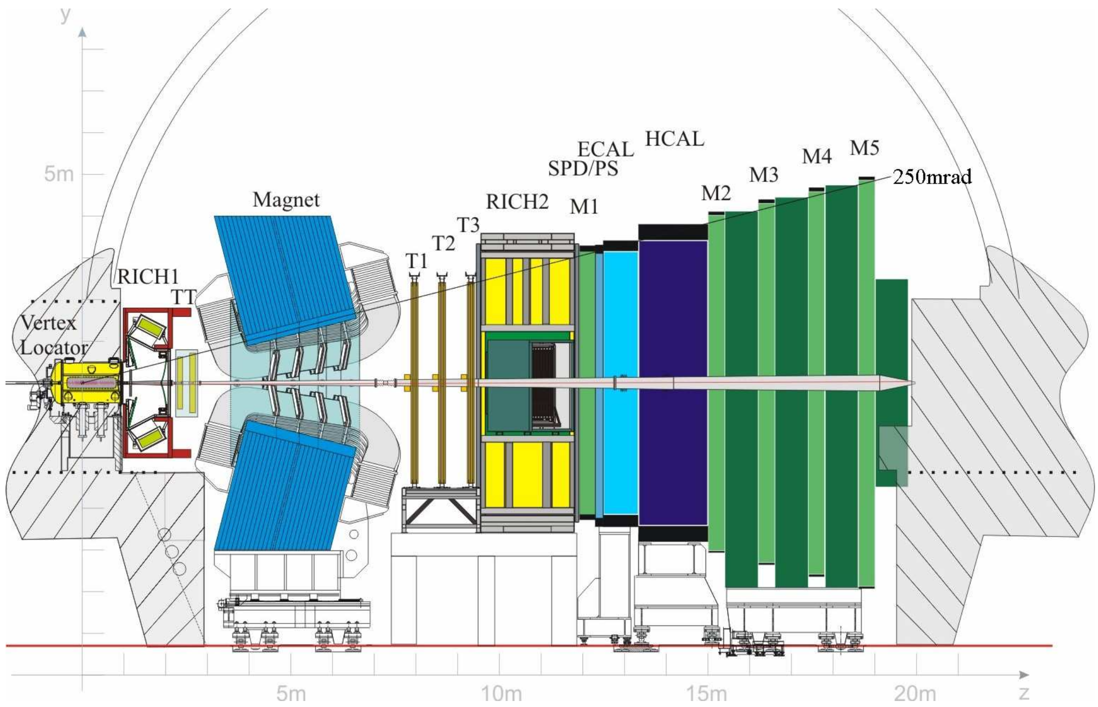


Figure 3.1: An illustration of the cross section of the LHCb detector [13].

mary and secondary decay vertices. Information about the flight path of charged particles is recorded by the Tracking stations before (TT) and after (T1-T3) the magnet. The effect of the magnet on these allows the determination of their momenta. Ring-Imagining Cherenkov detectors (RICH) provide particle hypotheses for charged particles. Electrons and photons induce showers in the Electromagnetic Calorimeter (ECAL), while the Scintillating Pad Detector (SPD) assists in separating between them. Similarly the Hadronic Calorimeters (HCAL) detects hadronic showers. Finally, the Muon stations track and identify muons.

We note the coordinate system displayed: the z-axis follows the beam through the detector, the y-axis points up and the x-axis points out of the picture, away from the centre of the LHC. When indicating a direction on the z-axis ”downstream” denotes the direction away from the interaction point towards the detector and ”upstream” denotes the opposite. LHCb stands out among the LHC experiments by being a single-arm forward spectrometer, meaning the detector is asymmetric, measuring only in one direction coming frome the interaction point. It covers $10 - 300\,\mathrm{mrad}$ in the bending and $10 - 250\,\mathrm{mrad}$ in the non-bending plane. For high energies, b- and c-flavoured hadrons are mostly produced highly boosted, with a small angle θ to the beam direction, which motivates this design. LHCb covers 4% of the solid angle, but 25% of $b\bar{b}$ production. Instead of the angle one can also use the pseudorapidity η defined as [14]

$$|\eta| = |\ln(\tan(\theta))| \tag{3.2}$$

In this parametrisation, the LHCb angular acceptance is $1.8 < \eta < 4.9$. Histograms of the $b\bar{b}$ production distribution can be seen in Fig. 3.2.

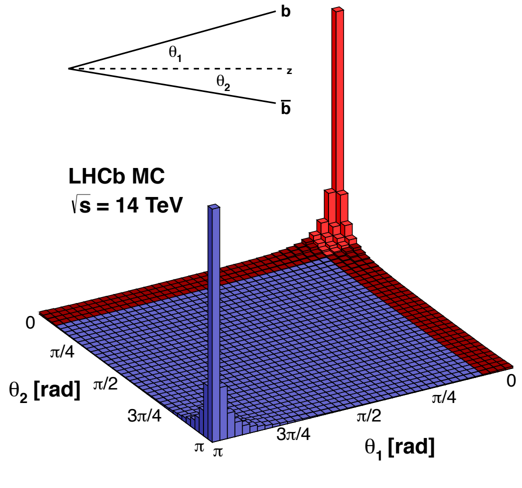
Comparing how the distribution depends on the angle to how it depends on the pseudorapidity, one finds that the distribution is more even in pseudorapidity with the peak being spread out over more bins, making it more desirable as a binning parameter.

Another parameter often practical to use is the transverse momentum p_T. It is defined as the momentum transverse to the beam direction,

$$p_T = \sqrt{p_x^2 + p_y^2}. \tag{3.3}$$

In contrast to the total momentum of a particle which still carries a portion of the initial proton momentum, the transverse momentum is fully due to physics happening in and after the collision.

A unique trait of LHCb is the usage of luminosity levelling. By beam-widening, the instant luminosity is reduced by up to two orders of magnitude. Additionally, it is kept near constant throughout a fill of the LHC, whereas naturally it would reduce over time.



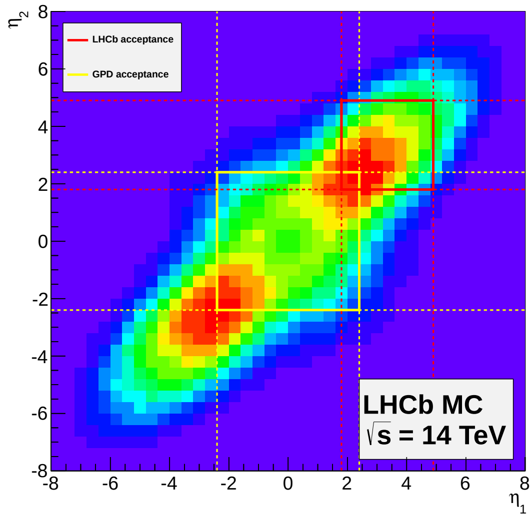


Figure 3.2: $b\bar{b}$ production distribution as a function of angle (left) and pseudorapidity (right). Both figures are based on the same Monte Carlo simulation. In the right hand plot the LHCb acceptance is marked as a red box [15].

3.2 The LHCb detector

The LHCb detector is made up of several subdetectors. Each of them provides a complementary set of measurements to achieve the necessary level of decay reconstruction. Broadly speaking, their functions can be separated into two categories: tracking, the construction of a particle track with associated momentum out of individual hits and particle identification (PID), the assignment of a particle type to a particular track. Further it is necessary to read out and collect the information in a practical manner and make it accessible.

This section covers the the components of the detector, the trigger and the following data flow, as in place from 2015-2018 during Run 2 of the LHC, to provide the background of the data that is later used by this analysis.

3.2.1 Tracking

Vertex Locator

The Vertex Locator is surrounding the proton-proton collision region. It allows reconstruction of particle trajectories close to the interaction point. As displaced secondary vertices are a prime characteristic of *b*- and *c*-flavour decays, its function of precisely identifying the vertex positions is of importance for their identification and for an efficient trigger [16]. An illustration of it can be seen in Fig. 3.3.

The VELO is comprised of 42 silicon microstrip detector modules, in turn consisting

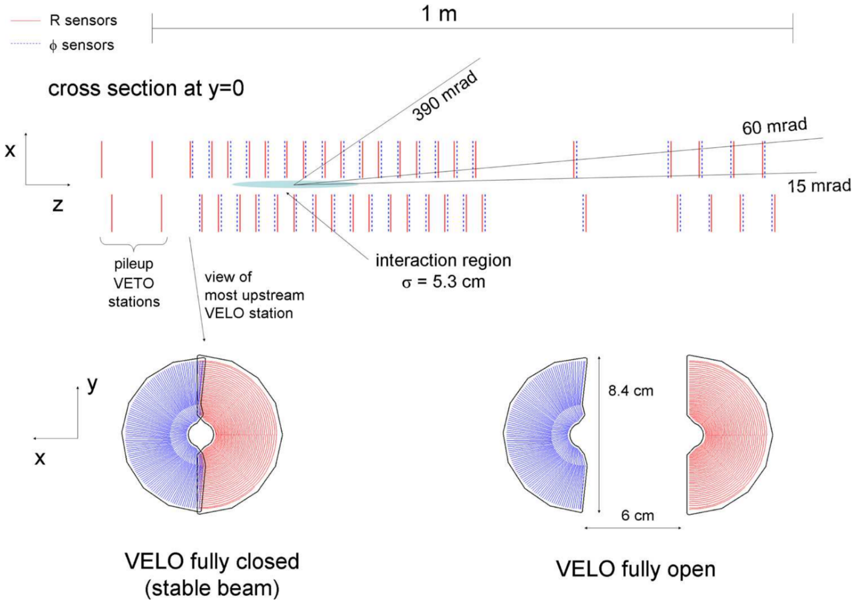


Figure 3.3: Cross section of the VELO in x-z-plane at y = 0 (top) and x-y-plane (bottom). Top showcases the geometry of modules and sensors. Note that the latter alternate. Bottom shows the VELO in both Open and Closed position [13].

of sensors for both the r and the ϕ coordinate, the radial distance from the beam line and the angle. The aperture required by the LHC during beam injection is larger than that during the subsequent operation, hence it is advantageous to have the VELO be able to move closer to the beam once it is stable. Thus the modules are made to be roughly half-circular with a slight overlap and arrayed in two halves that can then be brought together. The modules of one half are encased together by a thin corrugated aluminium foil that separates them from the machine vacuum.

Magnet

LHCb uses a warm dipole magnet to bend charged tracks. As deflection by Lorentz force is dependent on momentum, measuring the curvature of particle tracks allows reconstruction of their momenta. The Magnet is constructed in such a way that it provides an integrated magnetic field of 4 Tm for 10 m long tracks, but has its field drop off to less than 2 mT inside the RICH. The polarisation of the magnet is flipped regularly to control potential asymmetries.

Trigger Tracker

Also known as Tracker Turicensis, the Trigger Tracker (TT) is placed upstream of the magnet. Besides improving momentum resolution it also allows detection of low-momentum

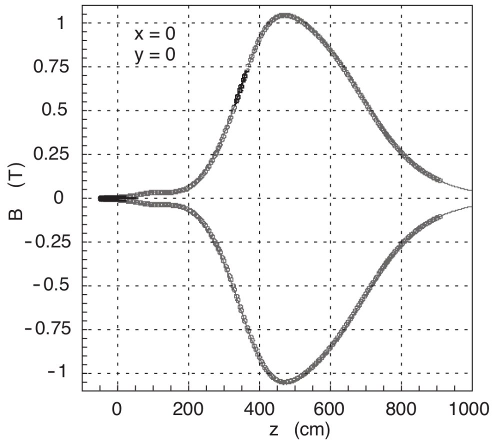


Figure 3.4: The magnetic field along the z-axis [13].

particles that are deflected out of acceptance by the magnet. It is of rectangular shape covering 150 cm × 130 cm and uses silicon microstrips sensors, with a strip pitch of 200 μm, that are arranged in four detection layers. The strips of the first and last layers are vertical, whereas those of the the second and third layer are rotated by -5° and 5° respectively. The TT's spatial resolution is measured to be about 50 μm and the overall hit efficiency is found to be greater than 99.7% [12].

Tracking stations

Because the occupancy is expected to be higher close to the beam, the tracking stations behind the magnet use a two section design. The centre of a station, directly surrounding the beam pipe, is the Inner Tracker (IT), which provides a finer granularity than the Outer Tracker (OT) making up the the larger remaining part of the station.

A silicon microstrip detector, like the TT, the Inner Tracker shares the structure of four detection layers with varying angle. However, while the TT is a single unit of plain rectangular shape, an IT station comprises four overlapping boxes covering a cross-shaped area with a width of 120 cm and a height of 40 cm.

The Outer Tracker is a drift-time detector, consisting of straw-tubes filled with a 7:3 mixture of Argon and CO_2, arrayed in two staggered layers per module. The same four layer arrangement used in the TT, with the modules in the two middle layers tilted to the vertical, is also employed for the OT. The hit efficiency in the centre half of the straw is found to be 99.2% and positional resolution is 200 μm [17].

3.2.2 Particle identification

RICH

Having already gained a measure of the momentum of the particle from the tracking, one can combine it with a measurement of its velocity to deduce its mass and provide particle identification. For that the Cherenkov effect is used: if a charged particle traverses a medium with a refractive index $n > 1$, with a velocity, v, larger than the speed of light in the medium, c/n, it emits photons at an angle, θ, to the direction of its momentum. This angle is only a function of the refractive index and the particle velocity, $\cos(\theta) = c/(nv)$, meaning once it is measured the velocity can be inferred and thus also mass and particle type. This is especially useful for charged hadrons of longer lifetime, such as pions, kaons and protons.

For any given radiator material, a certain minimum momentum is needed for a particle to emit Cherenkov radiation. Conversely, for sufficiently high momenta the differences in emission angle are too small to be distinguished by the employed detector construction. As a consequence LHCb employs two Ring Imagining Cherenkov detectors: RICH1, placed between VELO and TT, and RICH2 situated directly downstream of the Tracking stations. RICH1 covers the momentum range of $2 - 40\,\mathrm{GeV}$, while RICH2 covers the $15 - 100\,\mathrm{GeV}$ region, but only with an angular range of $15 - 120\,\mathrm{mrad}$

Both RICH detectors have a similar construction. They consist of conic vessels filled with a radiator material and with mirrors at the side to focus emitted photons towards a detector situated outside spectrometer acceptance. RICH1 uses C_4F_{10} as a radiator while RICH2 is using CF_4.

Calorimeters

The LHCb calorimeter system consists of four components. In order of increasing distance to the interaction point these are the Scintillating Pad Detector (SPD) and PreShower (PS), the Electromagnetic Calorimeter (ECAL) and the Hadronic Calorimeter (HCAL).

The SPD and PS are two layers of scintillator tiles sandwiching a 15 mm layer of lead. The purpose of the SPD is the separation of charged and neutral particles, with only the former causing a signal. It is also important for the trigger as there is a cut on the number of SPD hits, vetoing decay candidates which are too complex for analysis. As light hadrons, meaning mostly pions, and electrons differ in energy deposition when inducing showers in the lead layer, the PS behind the SPD in discriminating between them.

The ECAL is composed of alternating layers of lead absorber and scintillating tiles. It measures the energy of incoming electrons and photons, which cause electromagnetic showers in the absorber material via Bremsstrahlung and pair production. Photomultipliers measure the light emission in the scintillator layers.

The HCAL shares the principal construction concept of the ECAL, but uses iron instead of lead as absorber material. When passing through the iron layers, hadrons induce hadronic showers, allowing to measure their energy.

Muon system

Farthest away from the interaction point is the Muon System comprised of the five rectangular Muon Stations (M1-M5). The first station M1 is upstream and stations 2-5 are downstream of the HCAL. Iron absorbers with a thickness of 80 cm are placed between stations M2 to M5. The stations employ multi-wire proportional (MWPCs) chambers, 276 each, with the exception of the inner part of the first station which uses 12 Gas Electron Multiplier (GEM) detectors to cope with the enormous radiation. Muons ionize the gas when passing through, allowing detection.

All five muon stations register hit position of a potential particle track. Having higher resolution, the first three stations also use the information gained about the slope of the track and combine it with the average *pp* interaction point to extract a fast measurement of transverse momentum [18], independently of the tracking system. Taken together these are principal inputs of the L0 trigger.

A binary variable `isMuon` is assigned to tracks that hit enough muon stations. The number of required hits depends on the particle momentum and is shown in Tab. 3.1

PID variables

For each individidual sub-detector one can calculate the likelihood of the observed track being realized given a certain particle identity hypothesis. The likelihoods of different sub-detectors are then combined in one variable for the entire detector. To use these likelihoods as discriminating feature between different particle hypotheses one can then consider the difference of their logarithms, the difference in logarithmic likelihoods (DLL). As pions are the most numerous particles in the aftermath of a *pp* collision, all likelihoods are compared to the pion hypothesis as a baseline. In this analysis, the DLL variables are referred to as `PIDx`, with x being the particle hypothesis.

Momentum [GeV]	Required hits in
$3 < p < 6$	M2 & M3
$6 < p < 10$	M2 & M3 & (M4 | M5)
$10 < p$	M2 & M3 & M4 & M5

Table 3.1: Muon station hit requirement for the `isMuon` variable dependent on track momentum.

Another set of discriminating variables is gained by using an artificial neural network to provide a ”probability” (in actuality best understood to be just a score) for a track to be caused by each specific particle type, given its behaviour in the different sub-detectors. These variables are called `ProbNN`.

3.2.3 Online and offline reconstruction

Trigger

To cope with the enormous frequency of 40 MHz provided by the LHC, the LHCb experiment makes use of a triggering scheme that filters out decay candidates on site (online) that do not contain interesting physics or are not feasible to be reconstructed, before committing data to storage. During Run 2 the trigger used a three stage approach [19].

The level-zero (L0) trigger is exclusively based on the calorimeter system and the muon stations. This stage is completely implemented on local hardware. The L0 trigger asks for particle candidates to pass transverse energy thresholds, as the decay products of heavy mesons, which LHCb is interested in, are expected to come with a higher transverse energy. Further it places a cut on the number of SPD hits to veto high multiplicity candidates which are hard to reconstruct. Tracks with a hit in all muon stations are also selected. The L0 trigger reduces the rate to 1 MHz, which is low enough that the tracking system can be read out.

Following this are two software stages. They are implemented in the Moore application [20]. The first is High Level Trigger 1 (HLT1) which does a partial decay reconstruction. It reconstructs both charged tracks and primary vertices. While particle identification does not yet come fully into play, muon tracks are now assigned as such. HLT1 does reduce the rate further to 110 kHz.

In the last stage before storage, HLT2 also reconstructs the tracks of neutral particles and can use all PID features. Due to making full use of the available sub-detectors HLT2 is able to reconstruct all tracks without a minimum momentum requirement. It is also able to use multivariate classifiers (MVAs) to reject fake tracks. Ultimately, data is written to storage at a rate of 12.5 kHz, from where it can be used for further analysis (offline).

Data flow

The recorded data is not yet ready to be used directly for analysis but still in need of some preparation. This is done by running it through a chain of applications as depicted in Fig. 3.5.

Brunel [22] reconstructs the hits in various sub-detectors into tracks. Using DaVinci [23] one can then transform raw data into a format usable for analysis: ROOT. From here

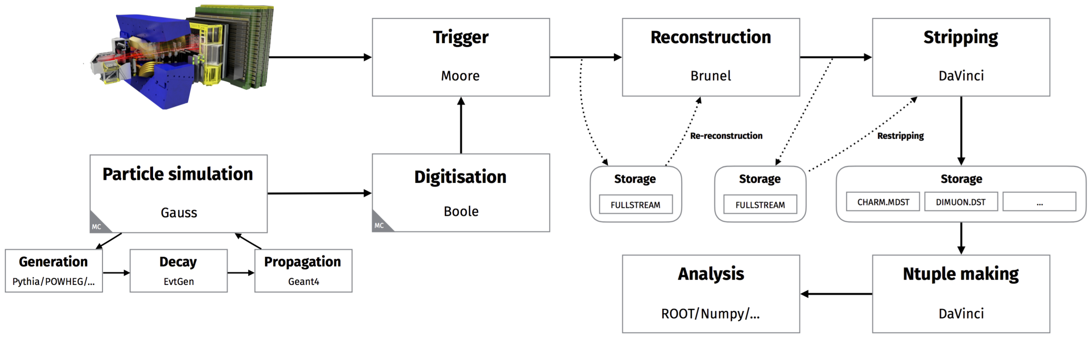


Figure 3.5: The course of data at LHCb. Note that both simulated data and recorded data are processed in the same manner starting with the Moore application [21].

on the user is free to choose how to commence the actual analysis.

Simulation

Another important component of analysis work is the usage of simulated data samples. These bridge the gap between theory and experiment by allowing direct comparison between what is measured and what one would predict to measure. They also allow to study how different ways of data-manipulation, e.g. cuts, affect efficiencies for a given process.

LHCb uses the Gauss framework [24] to implement the simulation. Proton-proton collision events are generated by Pythia [25], decays of the resulting hadrons are simulated by EvtGen [26] and interactions and propagation within the detector are governed by Geant4 [27]. Boole [28] then digitises simulated hits to look as if they are coming from the actual detector. This in turn is then fed into the same data processing machinery as real detector output, starting from the trigger application Moore.

Chapter 4

Analysis overview

After establishing the general background of the thesis, we now move on to the analysis itself. This chapter aims to provide a broad overview of both goal and method employed. Section 4.1 introduces the physical process of the decay and lays out why it is of interest. In section 4.2 an outline of the search for it is presented. Section 4.3 provides an overview of the tools used for this work.

4.1 The decay $B^+ \to \mu^+ \nu_\mu e^+ e^-$

As described in the preceding chapter, the B meson light-cone distribution amplitude is an important input for calculations describing B decays that involve hadrons. As of yet, it is poorly constrained by experiment. To probe it one wants to use a decay that is dependent on it but can otherwise be described perturbatively, such as a fully leptonic decay.

The prime candidates for this are decays of the sort $B^+ \to \ell^+ \nu_\ell \gamma$. However, the photon emitted is hard to reconstruct for LHCb. Instead, one can consider decays in which the photon is virtual and converts into a lepton pair, $B^+ \to \ell^+ \nu_\ell \ell^{+\prime} \ell^{-\prime}$, which is easier to reconstruct. A first search for $B^+ \to \mu^+ \nu_\mu \mu^+ \mu^-$ was performed by the LHCb collaboration and achieved an upper bound on the branching ratio that is close to theoretic expectations [5], although with the caveat that the three charged leptons being of the same flavour complicates further utilisation of the result.

This analysis is concerned with the decay $B^+ \to \mu^+ \nu_\mu e^+ e^-$. The branching ratio of this decay channel is predicted to be proportional to $1/\lambda_B^2$ [29]. A plot showcasing the dependency of the branching ratio of $B^- \to e^- \bar{\nu}_e \mu^- \mu^+$ on λ_B is depicted in Fig. 4.1. Due to lepton flavour symmetry and no predicted sensitivity to CP-conjugation, it is expected to be the same for $B^+ \to \mu^+ \nu_\mu e^+ e^-$. Due to constraints on the phase space considered, the predicted values in the figure deviate from that of [30], which will be used in this

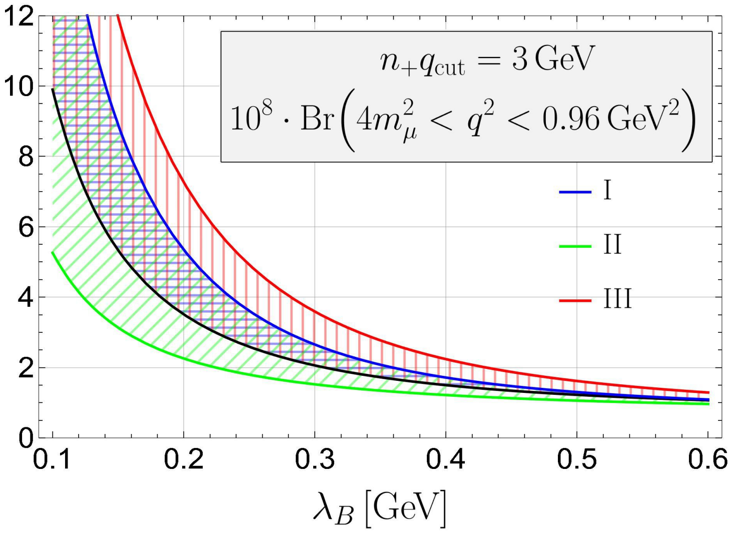


Figure 4.1: Branching fraction of the decay $B^- \to e^-\bar{\nu}_e\mu^-\mu^+$ as a function of λ_B. The four different lines correspond to changes in the radiative correction model. Taken from [29].

work.

The dominant tree level diagram of the decay is depicted in Fig. 4.2. Emission of the photon by the u quark is favoured over emission by the $\bar{b}$ quark, as the propagator connecting the photon and the W^+ vertex is proportional to $\frac{1}{m_q}$, with m_q being the mass of the respective quark. Photon emission by the muon is helicity suppressed and is hence also disfavoured.

Further, the photon propagator has hadronic contributions. These, more specifically the ρ and ω resonances, are predicted by theory [30] to be the main contributors of the virtual photon. In practice this results in the invariant dielectron mass, m_{ee}, distribution peaking at around 770 MeV.

The dielectron mass selection requirement is accordingly chosen to be around the resonance. This allows a clean separation from the decay $B^+ \to \mu^+\nu_\mu\gamma$, which in the case of

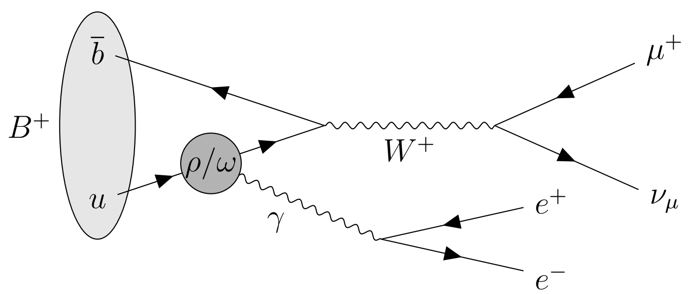


Figure 4.2: Dominant tree level Feynman diagram of $B^+ \to \mu^+\nu_\mu e^+e^-$. The photon propagator has contributions of the ρ and ω resonances.

the photon converting into an electron pair populates the low m_{ee} region and shares a final state with $B^+ \to \mu^+\nu_\mu e^+e^-$.

Due to the direct annihilation of $\bar{b}$ and u quark, the decay is suppressed by the square of $|V_{ub}|$, which is currently measured at $3.67^{+0.091}_{-0.07}$[1]. The theoretical prediction for the signal branching ratio is [30][2]

$$\mathcal{B}(B^+ \to \mu^+\nu_\mu e^+e^-) = (3.78 \pm 0.56) \times 10^{-8}. \tag{4.1}$$

For the purpose of the analysis there is no distinction made between $B^+ \to \mu^+\nu_\mu e^+e^-$ and its CP conjugate $B^- \to \mu^-\bar{\nu}_\mu e^+e^-$ as there is no difference in their treatment or any relevant parameter. This means that when dealing with data samples only the relative charges of the particles matter, e.g. for signal decay candidates there is a muon with a certain charge, an electron with opposite charge and another electron with the same charge as the muon. The charge of the $B^\pm$ meson is then taken to be that of the muon.

4.2 Analysis strategy

To achieve the goal of either detecting the decay $B^+ \to \mu^+\nu_\mu e^+e^-$ or providing an upper limit to its branching ratio, the analysis ultimately examines a mass spectrum and search for the signal signature. Due to neutrino involvement, the invariant four-body mass can not be reconstructed. Instead, an approximation of it, the corrected mass, m_{corr}, is used. It is defined as

$$m_{corr} = \sqrt{m_{\mu ee}^2 + p_\perp^2} + p_\perp, \tag{4.2}$$

with $m_{\mu ee}$ the invariant mass of the three charged leptons and $p_\perp$ the visible momentum transverse to the B^+ flight direction. A derivation of this variable can be found in A.1.1. Further, it can be shown that this type of variable is among the class of optimal variables for this sort of problem [33].

The data used for this thesis was recorded by the LHCb experiment during the years 2016-2018 of LHC Run 2. From all the data that successfully passed the trigger and was recorded, decay candidates are extracted. Each consists of three charged tracks, a dielectron pair and a muon, that can be joined at a secondary vertex clearly displaced from the primary vertex.

The resulting data sample is introduced in chapter 5, along with the selection im-

[1] There is some tension between the values for $|V_{ub}|$ extracted from inclusive and exclusive measurements [32], which needs to be resolved eventually.

[2] This prediction relies on dispersive methods instead of QCD factorisation. Thus, it does not make an assumption on λ_B, but on several form factors. The methods are assumed to be equivalent, but a translation is cumbersome.

posed, both during its production and in the following analysis. To note among these are the tight lepton PID requirements, intended to suppress the expected misidentification background. Besides the data sample containing the signal, the chapter also introduces all samples that are used for this work. They are either recorded data expected to help with estimating background contributions or simulated data of relevant decays, including the signal channel.

From this point on, the work that is carried out in this thesis is described. It sttarts with an inspection of the signal channel in chapter 6. For this a signal LHCb simulation sample is used. First, this sample is adapted to a change in the theoretical model of the signal decay by the usage of a machine-learning based reweighting algorithm. Then, the reference channel $B^+ \to K^+ J/\psi(e^+e^-)$ is used to normalize the signal decay, allowing an estimate of the number of expected decays.

Due to the low branching fraction of $B^+ \to \mu^+\nu_\mu e^+e^-$, the main problem of the search is the abundance of background. This is treated in chapter 7. The relevant contributors are expected to be combinatorial background and background due to particle misidentification. In addition to the already mentioned selection requirements, we try to curtail these backgrounds further by the usage of two gradient boosted decision tree classifiers, one constructed specifically in this work and a second, pre-made one. Furthermore, we model the remaining background. For the misidentified background this is done by weighting hadron samples, while the combinatorial background is taken to follow an exponential decay distribution.

With the ability to model the expected background, the next step is then the construction of toy models, described in chapter 8. This is done by bootstrapping decays from background proxies. These toy models reflect our knowledge of what can be expected in the observed data. The imperfection of our knowledge is captured by the fluctuations from toy to toy. By using these toys, we can optimize the selection on the classifier scores and estimate the quality of our fitting procedure.

Finally, an expected number of signal decays to which the used methodology is sensitive is calculated in chapter 9. For this, a large number of toy samples is constructed and fitted, with the methods introduced before. The individual sensitivity is then extracted from each fit and the median is taken. This in turn, combined with the normalisation performed before, allows to estimate an expected upper limit on the branching fraction of the signal decay.

4.3 Tools

4.3.1 Data management

Samples coming out of the machinery described in 3.2.3 are in `.root` format. ROOT [34] is an analysis framework, developed in the late 1990s at CERN to cope with the large amount of data faced in particle physics. Principally ROOT encompasses a suite of analysis tools, but this work only uses the file format designed for it. The `.root` format is structured hierarchically, much like a UNIX directory, which allows for memory-efficient storage of large amounts of similar objects. In `.root` data is stored in a tabular object called a tree, with column-like subobjects named branches.

While memory efficient, the ROOT framework is still limited in terms of capabilities compared to a full-fledged programming language. Due to this, all work of this thesis is done in Python. To access `.root` files in Python, the `uproot` package [35] is used, extracting samples into tabular pandas [36][37] DataFrames. These frames are the principal objects with which the work of this thesis is performed. They can be understood simply as a list of decay candidates, each corresponding to a row of parameters providing all information extracted out of the measurement apparatus described in chapter 3.2

Further the vector package [38] is used when calculating with 3D or Lorentz vectors.

4.3.2 Gradient boosted decision trees

Gradient boosted decision trees (BDTs) are a machine learning technique which gives predictions for an output variable, y, based on a set of input variables, x_i. Depending on the problem at hand the output can be among a discrete set of possibilities (classification) or on a continuous spectrum (regression). It is a case of supervised machine learning, meaning that a labelled data set, i.e. a set in which each entry includes both input and output variables, is provided to train the predictive algorithm. In this thesis the technique is used for both, first for regression in chapter 6 and then for classification in chapter 7.

Decision trees

The basic building block of this technique is the decision tree. To exemplify the concept consider a classification problem: a data set is provided, each of its N elements is characterized by continuous input variables $\boldsymbol{x}$ and a class label $y = 0, 1$. The decision tree is tasked with predicting y as a function of $\boldsymbol{x}$. To do so, it imposes a threshold in one of the input variables and splits the set into two regions. It then splits one region again using another threshold on another variable. This process then continues iteratively. The decision tree predicts the same class for all elements that share a region (leaf). The split is done in such a way as to minimize a metric which compares the predicted values, $f(\boldsymbol{x}_n)$, with the provided true labels y_n. This is called the loss function, an example would be the squared Euclidean metric

$$L(\boldsymbol{x}_n, y_n) = \sum_{n=0}^{N-1} (F(\boldsymbol{x}_n) - y_n)^2. \tag{4.3}$$

In principle a single decision tree can near-perfectly split up most encountered sets and provide nearly correct predictions. In practice, one is not interested in a perfect prediction for the training data set whose correct labels are already known. Instead, the aim is to deploy the constructed classifier on a set for which the classes are unknown. Fine-tuning the predictive algorithm to such a degree that it adapts to statistical fluctuations of the training data (overtraining) is not only inefficient, but also lessens its generality.

To circumvent this, limits are placed on the iterative process. Possibilities are demanding a minimal loss reduction for a split to be allowed, demanding a minimum number of entries per leaf or limiting the depth i.e. the maximal number of threshold cuts for one leaf.

Gradient boosting

Boosting denotes the successive construction of simple predictive models (weak learners), with each new model being improved by insight gained from evaluating the last. In the case of gradient boosting, this is done by using gradient descent. For that the gradient of the loss function $g(x_n)$ is considered.

$$g(x_n) = \frac{\partial L(x_n, F(x_n))}{\partial F(x_n)} \tag{4.4}$$

As the goal is to minimize the loss function, one can follow the gradient towards a minimum. This is done by training another decision tree, h, on the negative gradient and then adding it to the first decision tree. The idea is that at first order the gradient is proportional to the derivation from the label. For this approximation to be correct, h is multiplied with ρ a coefficient that is numerically optimized. Further, one can multiply another coefficient, η, called the learning rate or step size, which slows the movement towards the minimum, but allows more fine-grained movement and thus protects against accidentally overshooting the minimum.

For a given iteration number, the current decision tree is thus a sum of all previous corrections and the first decision tree F^0

$$F^l(x_n) = F^{l-1}(x_n) + \eta\rho^l h^l(x_n) = F^0 + \sum_{i=1}^{l} \eta\rho^i h^i(x_n). \tag{4.5}$$

The process is stopped, either after a fixed amount of iterations or when having lowered the loss below a threshold.

K-Folding

A common method to optimally use available training data for supervised machine learning is K-folding. Following this method the available labelled data is split into k subsets. A BDT is then trained on each of the k possible assortments of $k-1$ subsets. Each BDT then provides a test score for the subset of the training data it is not trained on. The final BDT is then the mean of all k BDTs. To evaluate its quality, one compares the test scores with the true labels. The advantage of this method is that the entirety of available training data is used both for training and for testing.

4.3.3 PIDCalib2

PIDCalib2 [39] is the updated version of PIDCalib [40], a set of software designed to estimate PID efficiencies in the LHCb experiment. The principal idea is to consider the effects of given PID requirements on a large amount of calibration tracks of known particle type, whose particle identity is obtained without any PID selection techniques. From this, it is possible to estimate the impact of the PID requirement on different samples.

More precisely, it works like this: in a first step, PIDCalib creates a histogram, binwise recording the efficiency of a chosen requirement on a chosen particle type. The user chooses in which features the binning is done and how the bin edges are placed. A reference file with clear particle ID is provided in the package. PIDCalib takes into account the year of data taking, the polarity of the magnet and other selections applied to the sample.

In a second step, a sample in ROOT format is entered into PIDCalib, together with the information on which particle track to apply the requirement on. PIDCalib then assigns an efficiency to each decay according to its placement in the binning scheme of the before created histogram.

The efficiency assigned to a decay candidate can then be understood thusly: Assuming the particle track is indeed corresponding to the chosen particle type, then particles of this type with the recorded values in binning variables would pass the selected PID requirement with the assigned efficiency.

In this analysis, PIDCalib2 is used prominently in chapter 7 to estimate number and distribution of misidentified particle background decays.

We use PIDCalib2 to estimate efficiencies for muons, charged pions and charged kaons. The decay $J/\psi \to \mu^+\mu^-$ is used by the muon calibration sample and $D^{*+} \to D^0(K^-\pi^+)\pi^+$ is used by the calibration sample for both pion and kaon [41]. In both cases, the decay is used because clean identification of the targeted particle, without the use of designated PID features, is possible. Throughout, binning is done in p_T and η, with logarithmic binning in the former and linear binning in the later.

As a side note it should be mentioned that PIDcalib subtracts an offset to account for noise. In low populated channels this can lead to negative efficiencies, in which case PIDcalib automatically assigns an efficiency of -999. We cut away all decays for which this happens.

4.3.4 The `iminuit` package

The python package `iminuit` [42] allows accessing the Minuit2 C++ library with python. This in turn is the updated version of the MINUIT algorithm [43] originally written in Fortran, by Fred James around 1975-1980. MINUIT was constructed specifically to use numerical methods to solve the complex minimization problems that are often faced by CERN scientists and has become a standard for this purpose. Generally the algorithm finds the minimum of a function. When wanting to fit a model to data, this function is the loss function describing how close the model prediction aligns with the data, such as negative log-likelihood or χ^2-sum.

Template fits

In addition to standard MINUIT capabilities, `iminuit` also brings with it some functionalities of its own. Among these is the implementation of template fits with included error propagation [44]. This method allows the fitting of a composite sum of histograms and continuous functions. Histograms have, at most, one free parameter, the yield, while continuous functions are allowed to have arbitrarily many. Fitting histograms is advantageous over the fitting of a continuous function in situations where (proxy) samples of processes expected to make up the data sample are available, while the probability density functions behind them are unknown. If it is necessary to fit the parameters of a known function, such as the decay constant of an exponential function, a pure histogram fit does not work. A template fit can deal with cases, which require both. Another advantage is that the template fit can take into account unusual uncertainties on the histogram bins, for example due to the decays behind them being weighted, because it is binwise by nature.

An example of a template fit is shown in Fig. 4.3. It consists of three contributions:

- A histogram with fixed yield.
- A histogram with floating yield.
- A parametric function with two floating parameters.

A composite sum of the three contributions is fitted. In this fit the parameters of the function and the yield of the non-fixed histogram are varied. The other histogram is completely fixed, but it contributes bin-wise uncertainties, which are taken into account by the fitting procedure.

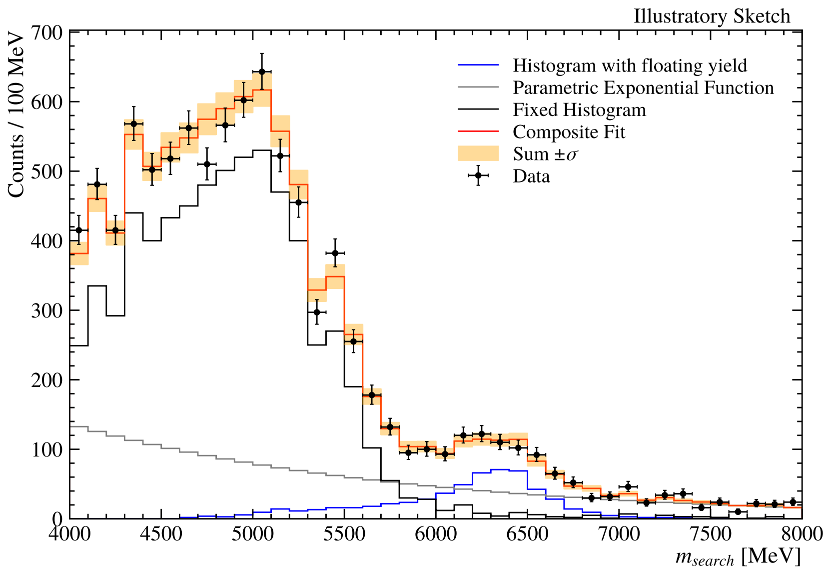


Figure 4.3: An example of a template fit.

Chapter 5

Samples and selection

This chapter documents the steps taken to select $B^+ \to \mu^+\nu_\mu e^+e^-$ candidates and which data and simulated samples are used in this analysis. In section 5.1, the selection requirements imposed on data are described. This is followed by an introduction of the simulated data samples used, given in section 5.2. Lastly, section 5.3 describes the signal window mass requirements on both dielectron mass and corrected mass.

5.1 Data selection

For this analysis data recorded by LHCb during the years 2016-2018 of Run 2 is used. This corresponds to an integrated luminosity of $5.1\,\mathrm{fb}^{-1}$ [45]. Before any work is done on recorded data, it passes through three levels of ever tighter selection constraints.

5.1.1 Trigger selection

The first is on the trigger lines met. As described in 3.2.3, LHCb uses both a hardware trigger and two subsequent software triggers to select what to record. For each level a potential decay candidate is required to fulfil any one of several selected conditions. These are listed in Tab. 5.1. We demand that the trigger conditions are satisfied by a track reconstructed as particle track in the decay candidate, as opposed to being fulfilled by a feature not belonging to the signal decay candidate. This requirement is termed TOS, **T**rigger **o**n **S**ignal. The selection of candidates according to trigger lines happens offline, i.e. after the stripping described below.

On the L0 level we demand that at least one of the three charged particle tracks in the final state passes the corresponding L0 trigger threshold, i.e. either the muon candidate passes the muon trigger or one of the electron candidates passes the electron trigger. At HLT1 we then further demand the existence of a well reconstructed track with large

Table 5.1: Selected trigger lines. For a candidate to pass through a trigger level it needs to fulfil at least one of the given conditions. For the L0 conditions, we demand that they are fulfilled by the corresponding track given in parentheses.

Trigger	Selected conditions
L0	`L0Electron`(e^+, e^-) `L0Muon`(μ^+)
HLT1	`Hlt1TrackMVA` `Hlt1TwoTrackMVA` `Hlt1TrackMuon`
HLT2	`Hlt2Topo[2,3]Body` `Hlt2TopoMu[2,3]Body`

impact parameter and at HLT2 we further specify by looking for a two or three body decay displaced from the primary vertex. L0 signal efficiency is estimated at 33%, that of HLT1 at 97% and that of HLT2 at 86%.

5.1.2 Stripping selection

The second stage of data selection is the so called stripping. A set of requirements is implemented to reduce the background pollution and thus the overall size of the data sample to a useable level. For this, the stripping line `B23MuNu_Muee` is used. Its content can be seen in Tab. 5.2. In contrast to the trigger selection, it is now demanded that a decay candidate satisfies all requirements given.

Both corrected mass and invariant three body mass are loosely constrained to regions which are expected to contain most of the signal. A first constraint on the transverse electron momentum is made as the low p_T region is high on background.

Several requirements are in place to assure the quality of the reconstructed candidates. Upper limits are placed on the $\chi^2/ndof$ values for the fits of the vertex reconstruction and the lepton track reconstruction. Similarly, the threshold on the allowed ghost probability, p_{ghost}, serves to eliminate ghost tracks, i.e. tracks arising from the combination of unrelated subdetector hits.

In Fig. 5.1, a sketch of the decay topology is given. As the B^+ meson is relatively long lived, it is expected to propagate before decaying, leading to a secondary vertex. Hence, a set of requirements is aimed at selecting candidates that involve a secondary vertex. For that the flight distance of the B^+ meson from primary to secondary vertex, FD, must be significant. This is captured by a threshold on the parameter χ^2_{FD}, which is the flight distance divided by its uncertainty. An analogous χ^2 object is defined for the impact parameter, the perpendicular distance of the particle track to the primary vertex. With the leptons originating from a displaced vertex, this must be significant too. Its definition is shown in the image on the right.

Table 5.2: Requirements of the `B23MuNu_Muee` stripping line. The χ^2 values denote the χ^2 value of the fit used for reconstruction and $ndof$ denotes the number of degrees of freedom of such a fit.

Applied on	Requirement
B^+	$m_{corr} \in [2500, 10000]$ MeV $m_{\mu ee} \in [0, 7500]$ MeV $\texttt{DIRA} > 0.99$ $p_T > 2000$ MeV $\chi^2_{FD} > 30$ $\chi^2_{vertex}/ndof < 4$
μ^+	$\chi^2_{track}/ndof < 3$ $\mathrm{p}_{ghost} < 0.35$ $\min(\chi^2_{IP}(primary)) > 9$ $\texttt{PIDmu} > 0$ $(\texttt{PIDmu} - \texttt{PIDK}) > 0$
$e^\pm$	$p_T > 200$ MeV $\chi^2_{track}/ndof < 3$ $\mathrm{p}_{ghost} < 0.35$ $\min(\chi^2_{IP}(primary)) > 25$ $\texttt{PIDe} > 2$ $(\texttt{PIDe} - \texttt{PIDK}) > 0$

A constraint on the B^+ direction angle, `DIRA`, the cosine of the angle between momentum and flight direction is also among them. Construction of the direction angle is shown in the left hand plot. For a B^+ meson originating from the primary vertex it should vanish,

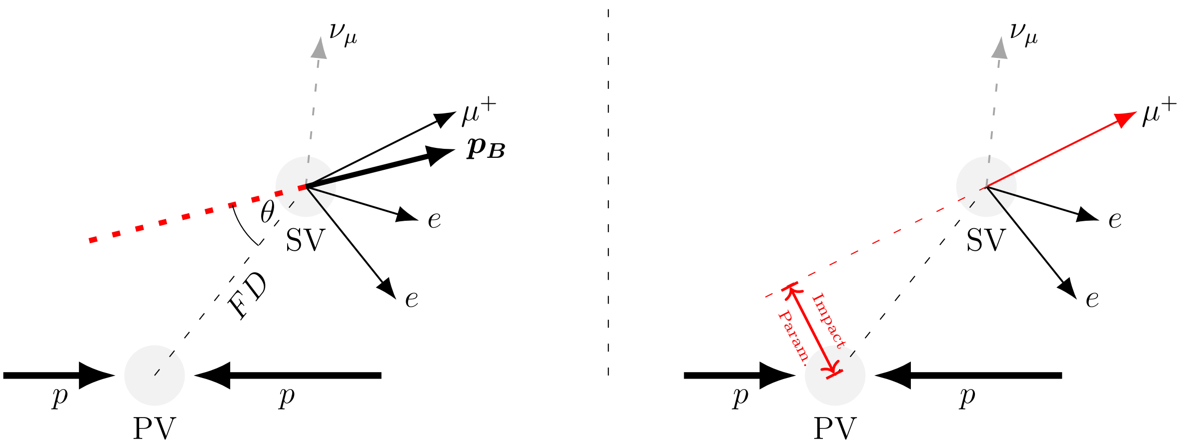


Figure 5.1: Schematic depictions of the decay topology in the x-z plane. It consists of the primary pp interaction vertex, PV, the secondary B^+ decay vertex, SV, and the decay products. On the left the direction angle, θ, between B^+-meson momentum and the line between primary and secondary vertex is shown, as is the flight distance, FD. In the right image the impact parameter of the muon to the primary vertex is indicated.

corresponding to a cosine of one. Additionally, the transverse momentum of the B^+ must be high to allow for good vertex separation.

Lastly, several requirements are placed on the `PID` variables of both the muon and the electrons. These are the afore mentioned DLL variables that combine the information gained in the various PID related subdetectors.

An important thing to keep in mind about the stripping selection is that it is done centrally and cannot be changed for a processed sample. The processed sample resulting from the given trigger and stripping selection requirements is referred to as data sample.

5.1.3 Offline selection

This is in contrast to the last round of selection, the offline selection. The thresholds imposed at this stage can be (and partially are) lifted during the analysis, if so desired. They are listed in Tab. 5.3.

A first set of requirements is in place to assure that decay candidates interact with the subdetectors as required. Electrons are expected to have hits in and be within the acceptance of the ECAL. Hits in the RICH detectors are also expected. Similarly, the muon track needs to have left hits in the muon stations. Additionally, hit multiplicity in the SPD is also restricted. In principle these requirements should mostly be fulfilled by any candidate that passes both trigger and stripping, they are imposed again to remove possible edge cases.

PIDCalib is based on the usage of reference samples and needs certain minimum (transverse) momenta for its prediction to be sensible. Requiring these has the additional advantage of combatting backgrounds, which are numerous in the low p_T region. Further, we also impose an upper limit on transverse momenta and an accepted range on pseudo-rapidity. This is done to have a uniform phase space when using PIDCalib on different samples. Lowly populated regions, where the method breaks down, are also removed.

Again, a part of the selection is made to assure that the candidate topology contains a good B^+-decay secondary vertex. Direction angle and χ^2_{FD} requirements are tightened. The impact parameter χ^2 of the combined dielectron object is demanded to be above a threshold. As a quality check of the vertex construction itself, the χ^2 of the distance of closest approach (DOCA) of the lepton tracks must be small.

Several kinematic requirements are made on the decay candidates. The measured electron transverse momentum without bremsstrahlungs reconstruction, $p_{T,track}$, is cut off at 200 MeV, again to reduce background. The square of the missing mass m^2_{miss} is also constrained, meaning the discrepancy between the three body mass and the corrected mass must not be too large. Similarly, the invariant mass squared of the muon-neutrino system, k^2, is given a lower limit.

Table 5.3: Offline selection requirements.

Type	**Applied on**	**Requirement**
Detector	μ^+	`hasMuon` $== 1$
	$e^\pm$	`hasCalo` $== 1$ `hasRich` $== 1$ `InAccECAL` $== 1$ `region`$_{ECAL} \geq 0$ $(x_{ECAL} > 363.3\,\text{mm} \mid y_{ECAL} > 282.6\,\text{mm})$
	general	`nSPDHits` < 450
PIDCalib	μ^+	$p_T \in [1200, 14000]\,\text{MeV}$ $p > 3000\,\text{MeV}$ $\eta \in [1.75, 4.5]$
	$e^\pm$	$p_T \in [500, 14000]\,\text{MeV}$ $p > 3000\,\text{MeV}$ $\eta \in [1.75, 4.5]$
Topological	B^+	`DIRA` > 0.995 $\chi^2_{FD} > 100$ $\chi^2_{DOCA}(\mu^+e^+e^-) < 9$
	dielectron	$\chi^2_{IP} > 40$
Kinematic	$e^\pm$	$p_{T,track} > 200\,\text{MeV}$
	general	$m_{\mu ee} \in [500, 6000]\,\text{MeV}$ $m^2_{miss} \in [-10, 10]\,\text{GeV}^2$ $k^2 > -5\,\text{GeV}^2$
Clone veto	$\mu^+e^\pm$	$\theta(\mu^+, e^\pm) > 5\,\text{mrad}$
Virtual photon requirement	$e^\pm$	`VeloCharge` < 1.25
	dielectron	$FD < 20\,\text{mm}$ $\chi^2_{FD} < 9$
PID	μ^+	`PIDmu` > 2 `ProbNNmu` > 0.8 `isMuon` $== 1$
	$e^\pm$	`ProbNNe` > 0.2

To protect against a track segment being reconstructed twice as part of two different tracks (cloned), a minimal opening angle of 5 mrad between the muon track and either lepton track is demanded. The requirement is not made on the electron pair, as there it is covered by the virtuality conditions of the photon.

These are made to assure that the dielectrons come from a virtual photon and not the decay of a real particle. We therefore restrict the flight distance, FD, i.e. the distance between the B^+ decay vertex and a fitted vertex of the two electron tracks, to exclude a possible tertiary vertex. Further, a limit is placed on the measured charge deposition in the VELO. The electrons are assumed to be minimum ionizing particles, meaning they induce a fixed energy deposit per hit. A higher measured deposit suggests that two tracks are collinear. As the virtual photon is massive, photons emitted from its decay should be emitted with a non zero opening angle.

Lastly, PID requirements are also tightened. They now also include usage of the `ProbNN` variables and the `isMuon` flag. We note that the additional electron PID requirements are limited to a single threshold on `ProbNNe`, as the requirements on the stripping level are already strict.

5.1.4 Further data samples

In addition to the main data sample described above, we also make use of two further samples based on LHCb measurements. They are:

- A sample without muon PID requirement at stripping level. Due to the large number of eligible decay candidates it is prescaled to 1%, i.e. it contains only 1% of all decay candidates that fulfil its set of requirements and have been measured by LHCb within the considered time frame. This sample is referred to as misIDμ sample.
- A sample with wrong relative charges of the particle tracks. Specifically, this means that the two electron tracks have the same charge, corresponding to a selection of unphysical decays $B^+ \to \mu^\pm e^+ e^+$. This sample is referred to as same sign sample.

With the exception of the described deviation, both samples make use of the same stripping and trigger selection requirements as the signal sample. This allows direct comparison to estimate expected backgrounds.

5.2 Simulation samples

For the estimation of expected distribution shapes and how they are affected by selection requirements, simulation samples are used. These samples are produced following the description in 3.2.3. The simulation takes into account the proton-proton collision, the considered decay itself, propagation and LHCb detector response. Like the recorded data samples, it also takes into account year of operation and magnet polarity. It, however, does not take into account the yield of the simulated decay. Instead, the number of generated decay events is chosen manually and has no relation to the number of decays that occurred during the corresponding time of LHCb data taking.

An overview of the used simulation samples and the number of generated decays is given in Tab. 5.4. As described in chapter 3, the LHCb detector does not cover the entire solid angle, meaning that of all decays that occur only a fraction can potentially be observed. To not waste computing power, only decays in the detector acceptance

Table 5.4: LHCb simulation samples used for this analysis. The acceptance efficiency of the signal channel is exactly 50% due to a mistake in the production.

Sample	# decays at phys. level	ϵ_{acc}[%]
$B^+ \to \mu^+\nu_\mu e^+e^-$	6040440	50^*
$B^+ \to K^+J/\psi(e^+e^-)$	15×10^6	17.25 ± 0.08
$B^+ \to \mu^+\nu_\mu\pi^0(\gamma(ee)\gamma)$	3×10^6	16.557 ± 0.037
$B^+ \to \mu^+\nu_\mu\pi^0(e^+e^-\gamma)$	2×10^6	16.484 ± 0.037
$B^+ \to \mu^+\nu_\mu\eta(\gamma(e^+e^-)\gamma)$	3×10^6	15.670 ± 0.037
$B^+ \to \mu^+\nu_\mu\eta(e^+e^-\gamma)$	2×10^6	15.692 ± 0.037
$B_c^+ \to \mu^+\nu_\mu J/\psi(e^+e^-)$	3×10^6	15.664 ± 0.026
$B_c^+ \to \mu^+\nu_\mu\chi_{c0}(XJ/\psi(e^+e^-))$, $B_c^+ \to \mu^+\nu_\mu\chi_{c1}(XJ/\psi(e^+e^-))$, $B_c^+ \to \mu^+\nu_\mu\chi_{c2}(XJ/\psi(e^+e^-))$	1.5×10^6	9.704 ± 0.018

are generated. By dividing the number of generated decays by the detector acceptance efficiency, ϵ_{acc}, the corresponding number of decays occurring in the detector is obtained. Both this number and the efficiency are given in Tab. 5.4. For the signal channel sample, the acceptance efficiency is exactly 50% due to a mistake in the production. As it is known, this has no negative effects, aside from the inefficient production.

Further, the last listed sample contains a mixture of different decays. All of them involve the B_c^+ decaying into a charmonium, which then in turn decays into a J/ψ. Their relative contributions are fixed according to theoretical predictions.

All simulation samples are truth matched. This means that the reconstructed particle tracks are compared to their true identity, which is available for simulated samples. This way it is assured that any reconstructed decay candidate does indeed correspond to a simulated decay and is not a background caused by detector or reconstruction effects.

Generator level simulation samples

Additionally, the analysis uses two simulation samples of the signal decay at generator level. These samples only depict the kinematics of the decay $B^+ \to \mu^+\nu_\mu e^+e^-$, they do not include anything else, neither a preceding pp collision, nor propagation in and response of a detector. Both samples contain 1 million decays. They differ in the theoretical model used and the tool used to generate them. One of them is based on the model of Ivanov and Melikhov [46] and generated using EvtGen. The other sample is based on the model of van Dyke et al. [30] and generated using the library EOS [47].

5.3 The signal mass window

When looking for signal, another selection requirement is imposed. This is the signal mass window constraining both dielectron and corrected mass. It requires

$$m_{ee} \in [600, 900]\,\text{MeV} \quad \text{and} \quad m_{corr} \in [4500, 7000]\,\text{MeV}.$$

The range for the dielection mass is motivated by the prediction that the virtual photon is dominated by the ω and ρ contributions at 783 MeV and 770 MeV respectively. There is an expected photon contribution falling of $\propto 1/m_{ee}$ from $2 \times m_e$ onwards. However, this region is also host to several problematic backgrounds, which can be seen in section 7.4. Further, the decay $B^+ \to \mu^+\nu_\mu\gamma$ with the photon converting into an electron pair is also expected to populate this region. With a shared final state, separation is not possible.

As already mentioned, the corrected mass is the feature whose distribution is used to search for signal signature. In the left plot of Fig. 5.2, the expected distribution for the signal decay $B^+ \to \mu^+\nu_\mu e^+e^-$ can be seen. The plot is based on the usage of the simulated sample. A peak at $m_{B^+} = 5279\,\text{MeV}$ is visible, but so is a large tail towards the lower mass region. This is expected, as the corrected mass is an approximation that is always lower than the true B^+ mass. In addition, the peak is broadened in both directions due to resolutions effect. These are primarily stemming from the quality of the vertex reconstruction and the momentum resolution of the electrons. Still, the corrected mass offers good separation power. As a naïve comparison, the invariant three body mass of the lepton tracks can be considered. Its distribution in the simulation sample can be seen in the right plot of Fig. 5.2.

For the corrected mass, the choice of region is chosen as to provide good opportunity for the mass fit. The lower limit of the considered corrected mass range is chosen as

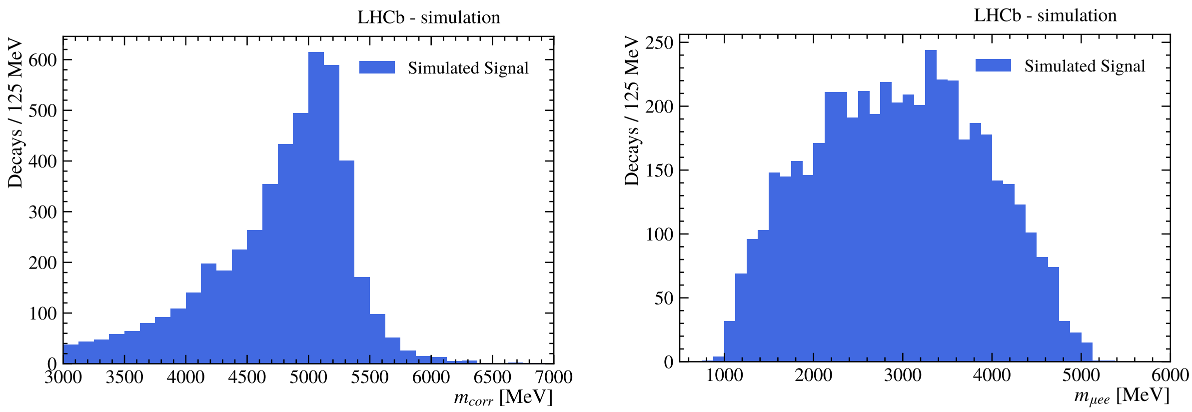


Figure 5.2: Distribution of the corrected mass distribution (right) and the invariant mass of the three lepton tracks (left).

4500 MeV, because it is expected that below this value we cannot model the background anymore. As data runs out above 7000 MeV, the considered region is cut off at this point.

Blinding strategy

Among the selection requirements, the signal mass window is a special case. To not bias the methodology involved, we avoid examining the data in this mass region until a final fit at the end of the analysis. Fits are made to the blinded region, but the fitted signal yield is kept unknown and only the fitted background is examined. This procedure is called blinding.

Instead, throughout this work, we often consider sidebands i.e. samples which explicitly do not cover this mass region. This is done by replacing either of the two mass requirements by another one excluding it, e.g. $m_{ee} \in [900, 1200]$ MeV. This allows to study the impact of the used methods on a proxy region that is close to what one would expect in the signal region.

Chapter 6

Signal studies

The purpose of this chapter is to get an understanding of the expected behaviour of the signal decay contribution. For this the simulated signal sample is examined. As a necessary first step, it is adapted to a change in the theoretic model by reweighting it, which is described in section 6.1. Then, in section 6.2, the decay channel $B^+ \to K^+ J/\psi(e^+e^-)$ is used for normalisation.

6.1 Reweighting

At the outset of this analysis process the most current theoretical description of the kinematics of $B^+ \to \ell^+\nu_\ell\ell^{+\prime}\ell^{-\prime}$ decays was given by Ivanov and Danilina [46, 48]. Consequently, the implemented simulation of this type of decay in EvtGen uses their model. Since then, this model has been superseded by that of van Dyk et al. [30].

However, this new model is not yet implemented in EventGen and hence also not in the general LHCb simulation framework, meaning it is currently not possible to use its predictions for the decay kinematics as inputs for the simulation of decay reconstruction in LHCb. To circumvent this issue, we aim to reweight simulation data that is already at hand, but based on the old model. First a reweighter is trained on generator level data, which is available for both models. Then this reweighter is applied on LHCb simulation data, assigning weights to each decay candidate based on kinematic parameters.

6.1.1 Changes of the kinematic model

As a starting point, the kinematic predictions provided by the two models are compared. For that we consider five kinematic parameters that are sufficient to fully describe the decay $B^+ \to \mu^+\nu_\mu e^+e^-$: the invariant mass of the dielectron pair, q^2, the invariant mass of the neutrino muon pair, k^2, the cosine of the angle between the same sign electron

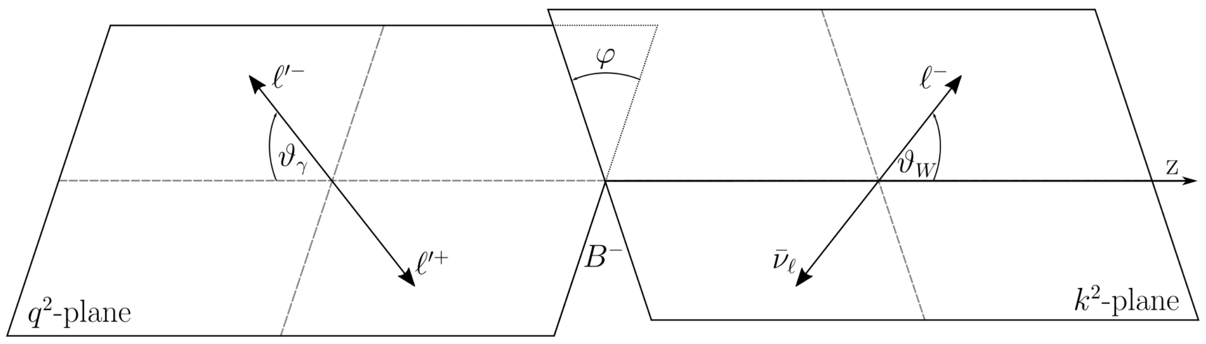


Figure 6.1: A sketch of a $B^- \to \ell'^{,-}\nu\ell^-\ell^+$ decay. The B meson moves along the z direction. Note that the angles $\vartheta_\gamma \equiv \theta_\ell, \vartheta_W \equiv \theta_W$ and ϕ are in different rest frames. Further note that charge conjugation necessitates adaption in the angle calculation. Taken from [30].

momentum and the B-meson momentum in the dielectron rest frame, θ_ℓ, the cosine of the angle between muon momentum and B-meson momentum in the neutrino-muon rest frame, θ_W and the angle between the two planes that are spanned by the respective lepton pair momentum vectors in the B rest frame, ϕ. A visualisation is given in Fig. 6.1.

By the choice of ϕ covering a 2π range, the lepton angles θ_ℓ, θ_W are automatically constrained to a range of π and can be described by their respective cosine without loss of information. The parameter k^2 describes the energy that is carried off by the neutrino-muon system and is thus constrained by the available energy, $(m_{B^+} - 2m_e)$, the difference between the B^+ mass and twice the electron mass.

In principle, q^2 is constrained similarly. However, theory does not yet give detailed predictions for $q^2 > 1\,\text{GeV}^2$. As the total contribution above this threshold is expected to be very small, it is taken as upper limit of q^2 for all purposes of this work. We note that the variable q^2 corresponds to the square of the dielectron mass m_{ee}. The considered signal window, introduced before, is then in this feature: $q^2 \in [0.36, 0.81]\,\text{GeV}^2$.

Additionally, our aim is to reweight the already produced simulated LHCb data, which has in place a lower q^2 limit of $0.026\,\text{GeV}^2$ and a lower k^2 limit of $0.01\,\text{GeV}^2$. For the purposes of reweighting, these constraints are imposed on the generator level samples as well. The total set of constraints is summarized in Table 6.1.

To compare the predictions of the two models we consider the distributions of the

Table 6.1: Considered ranges of the kinematic parameters.

Parameter	$\cos(\theta_\ell)$	$\cos(\theta_W)$	ϕ	q^2	k^2
Range	[-1,1]	[-1,1]	$[-\pi,\pi]$	$[0.026, 1]\,\text{GeV}^2$	$[0.01\,\text{GeV}^2,\ (m_{B^+} - 2m_e)^2]$

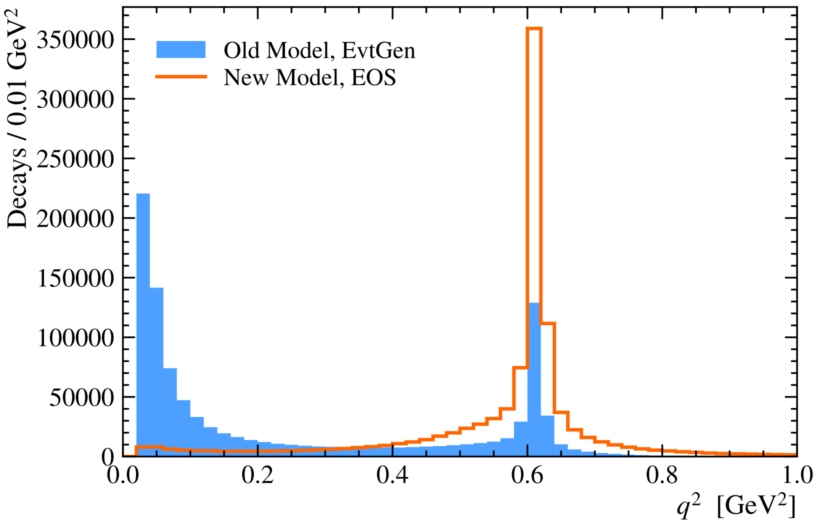


Figure 6.2: Distribution of q^2 for old model and new model at generator level.

kinematic parameters in the two generator level samples. Noticeable differences can be found, the most striking being in the q^2 distribution pictured in Fig. 6.2. It is apparent that the contribution of the ρ/ω resonance has increased relative to that of the photon pole.

Comparing the number of generated events in the signal dielectron mass window against the total number generated, one finds that the selection contains 75% of decays in the new model. This translates to a branching fraction of $(2.8 \pm 0.4) \times 10^{-8}$ for the decay limited to this dielectron mass region.

While less pronounced, there are also visible differences in the distributions of the other four parameters, which can be seen in Fig. 6.3.

An important question to consider is in how far these differences are connected. We find negligible correlation between the parameters, depicted in Appendix A.2.1, but with the most pronounced change being in q^2 we separate the events in two categories with $q^2 \leq 0.4\,\text{GeV}^2$ and $q^2 > 0.4\,\text{GeV}^2$, respectively, pictured in Fig. 6.4.

In the higher q^2 regime, where the ρ/ω resonance is the main contributor, the models nearly coincide in their predictions. The divergences stem from the description of the lower q^2 region, where it is mainly the photon pole contributing.

In this region, the new model on average predicts a higher k^2 value than the old model, meaning more momentum is carried by the average neutrino-muon pair. The nearly linear slope of $\cos(\theta_W)$ suggests further that in the neutrino-muon rest frame the muon favours being parallel to the B^+-momentum, while the neutrino favours antiparallel alignment with the B^+-momentum direction. Similarly, the shape of $\cos(\theta_\ell)$ suggests the electrons prefer emission perpendicular to the B^+-momentum in the dielectron rest frame.

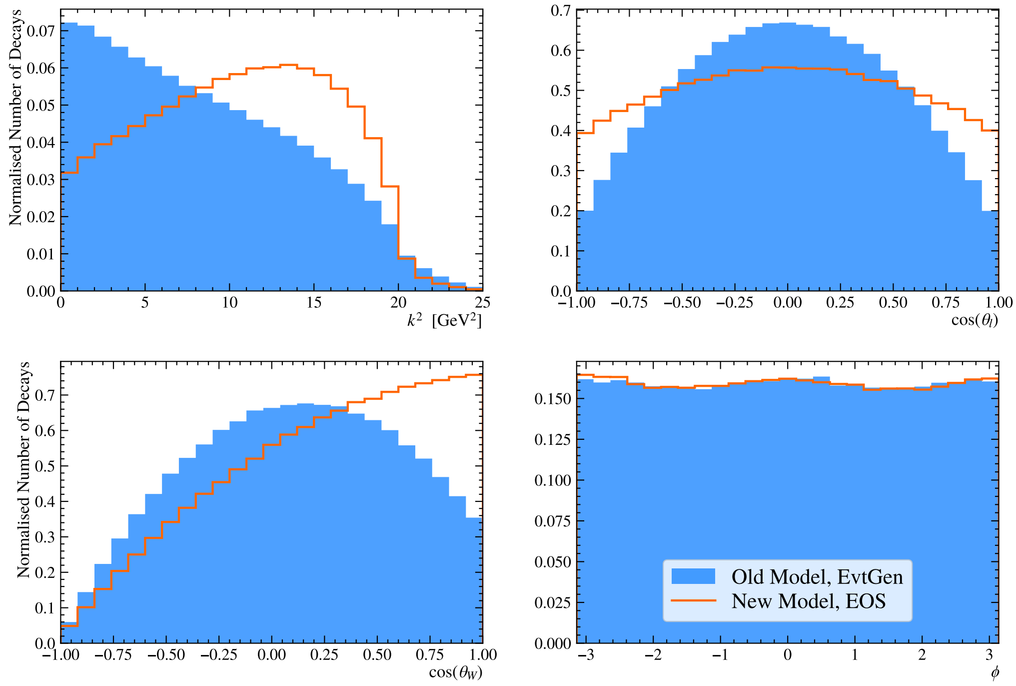


Figure 6.3: Distributions of k^2, $\cos(\theta_\ell)$, $\cos(\theta_W)$, ϕ for both old model and new model at generator level.

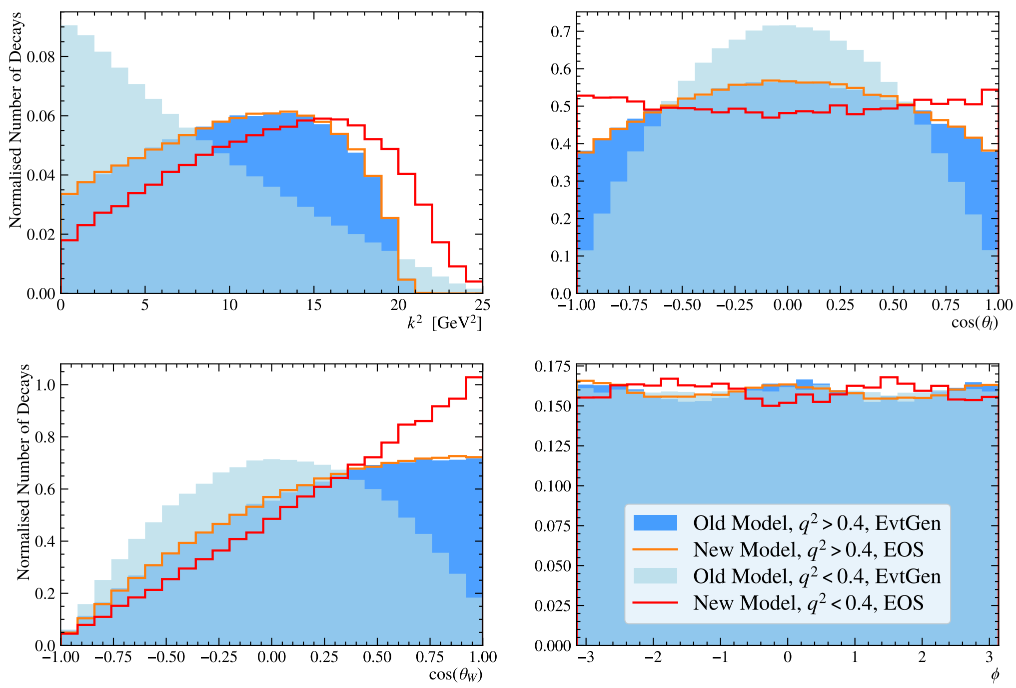


Figure 6.4: Distributions of k^2, $\cos(\theta_\ell)$, $\cos(\theta_W)$, ϕ for both old model and new model at generator level. The samples are split into $q^2 \leq 0.4\,\mathrm{GeV}^2$ and $q^2 > 0.4\,\mathrm{GeV}^2$

6.1.2 Reweighting with `hep_ml`

The idea of reweighting is simple: There are two sets of events with differing distributions in several variables and one would like them to coincide as multidimensional distributions. To achieve that, a weight is assigned to each event in one set according to its values in the considered parameters. This means that for the purpose of a distribution an event does not necessarily count as a single event any more, but is scaled by its weight when counted.

Reweighting is done using the `hep_ml` python package[49], which provides a reweighting algorithm based on gradient boosted decision trees [50]. In essence, this method works as described in 4.3.2. For training, the old model and new model generator level sample are used. The five kinematic parameters are the input features. A BDT assigns a weight, which can be any positive value, to each decay candidate of the old model sample. It aims to assign them in a manner that leads the resulting multidimensional kinematic parameter distribution of the weighted old model data set to match that of the new model.

The data is trained on in a k-fold manner, with k=2. As hyperparameters for the construction of the reweighter, we choose a number of iterations of 200, a maximum number of decision layers per tree of four and a learning rate of 0.1.

To estimate the quality of the reweighting, we can in turn train simple classifier BDTs to distinguish between new model decays and either reweighted or unreweighted old model decays. This classifier is trained using the same hyperparameters as the reweighter, i.e.

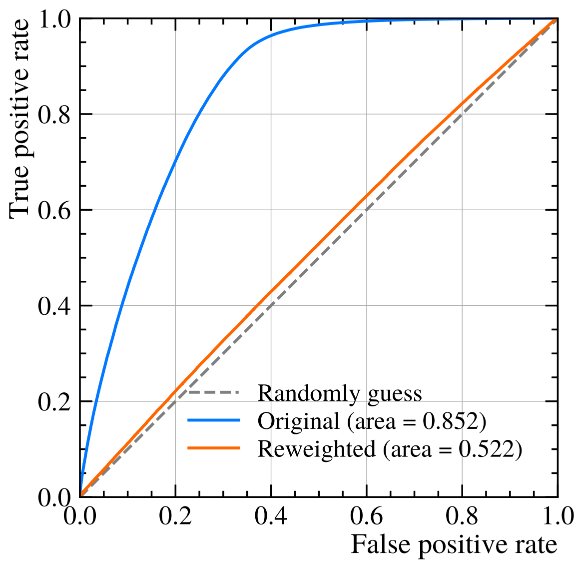


Figure 6.5: ROC curves for two classifiers. One constructed to distinguish new model decays against old model decays and the other constructed to differentiate between new model decays and reweighted old model decays. Both are constructed using 2-folding.

the BDT constructed to match and the BDTs constructed to discriminate are allowed the same level of complexity. As it is wished to apply the classifiers on the same sample that is used for training, 2-folding is used.

For evaluation the resulting **r**eceiver **o**perating **c**haracteristic (ROC) curves are examined. They can be understood thusly: A classifier assigns a score to each decay candidate, with the aim of identifying decays generated by using the new model. Then a score threshold is imposed. A fraction of new model decays passes this requirement (true positive rate), but the same applies for the old model decays (false positive rate). The ROC curve is obtained by considering all possible threshold values.

If the score offers no discriminative power, a threshold effects both type of decays equally. This would result in a perfectly diagonal curve. With the intention of the reweighting being conformance of the kinematic distributions, this is the optimal result. The outcome can be quantified in a single number the **a**rea **u**nder **c**urve (AUC), which is 0.5 in the case of no discriminative power and 1 in the case of perfect classification.

In Fig. 6.5 the ROC curves of the constructed classifiers are shown. While the classifier constructed for the unreweighted case provides some separation power with an AUC of 0.852, the classifier for the reweighted case achieves only an AUC of 0.522. This suggests that the reweighting is successful in matching the old model data to the new model.

The one-dimensional projections of the kinematic parameters of the reweighted sample can also be directly compared with those of the new model. This can be seen in Fig. 6.6. No significant deviation is apparent.

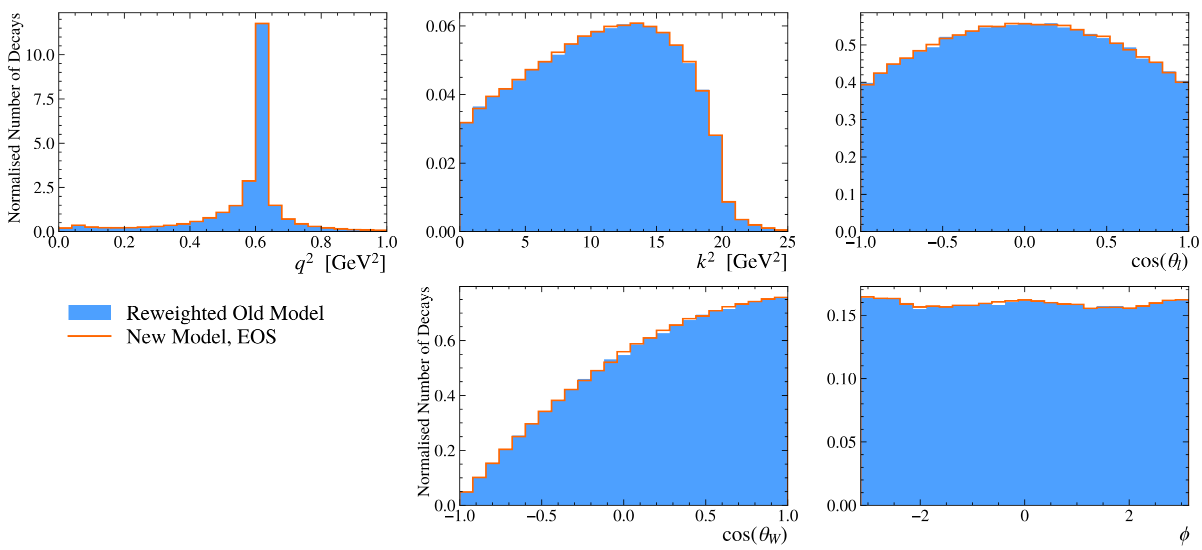


Figure 6.6: The reweighted generator level distributions in comparison to those of the new model.

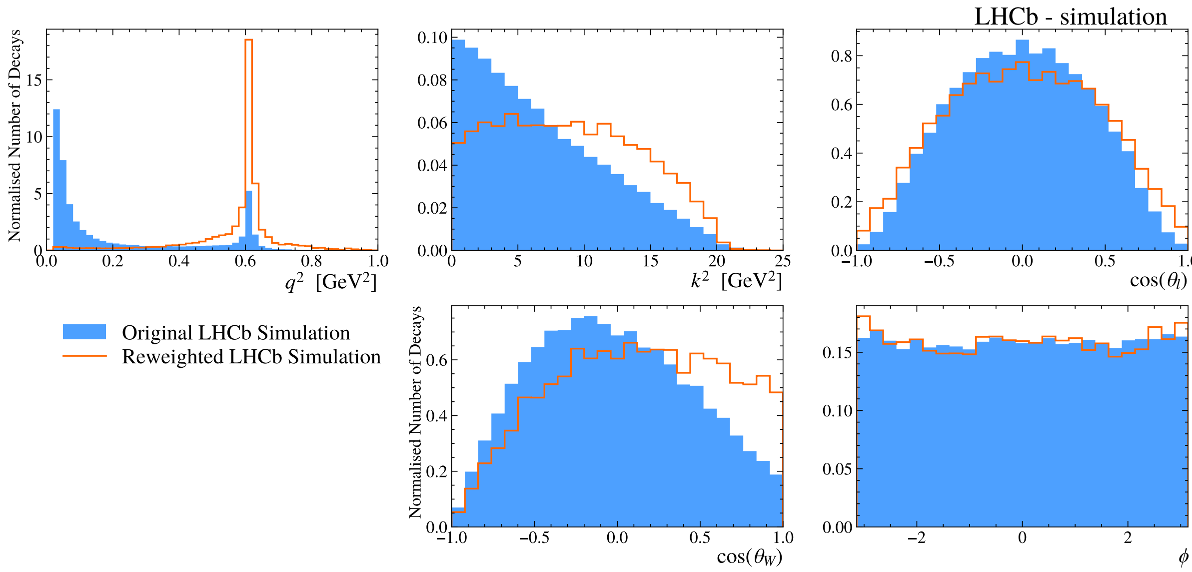


Figure 6.7: The reconstruction MC distributions before and after reweighting.

Going forward from this the constructed reweighter is then applied on the fully simulated LHCb sample. It is important to note that for the reweighting we use the true kinematic parameters before reconstruction, which are still in the sample. The result of this is seen in Fig. 6.7, where the original kinematic parameter distributions of the LHCb simulation sample are compared their reweighted counterparts. The changes broadly mirror those observed at generator level.

As it is the parameter which is ultimately of most concern for this analysis, it is also instructive to inspect how the reweighting impacts the corrected mass spectrum. The corrected mass shape before and after reweighting is depicted in Fig. 6.8. One can see that the general shape of the spectrum is retained in the reweighting process, but a slight broadening occurs. This is consistent with the changes observed beforehand in the distribution of k^2. If k^2 is higher on average that means the neutrino-muon pair is likely to carry a higher momentum. The corrected mass approximates the missing neutrino momentum, effectively by dropping the component parallel to the B^+ momentum. If the neutrino momentum is vanishing it is perfectly correct. As the neutrino carries away some energy, the corrected mass gains a tail towards lower mass. The more momentum the neutrino is expected to carry, the larger the tail.

Also important is the efficiency of the signal channel. On the generator level weights are normalized, i.e. the sum of the weights of a reweighted generator level sample is identical to its number of entries. The sum of the weights assigned to the individual decay candidates of the simulated LHCb data, however, can be higher or lower than the number of contained candidates, corresponding to an increase or decrease in efficiency.

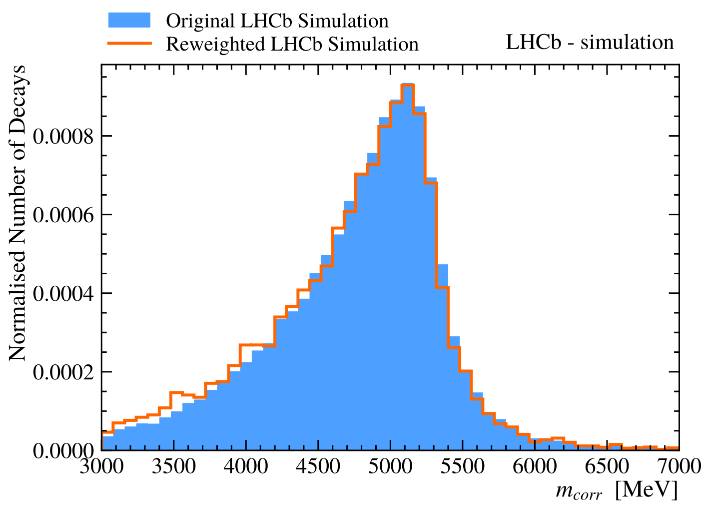


Figure 6.8: The corrected mass distribution before and after reweighting.

Indeed, on finds that the total efficiency of the reweighted sample is only 81% of that of the unreweighted sample. Having noted that the impact of the reweighting is most strongly seen in q^2, we consider the efficiency of the simulated LHCb data both before and after reweighting in bins of q^2. This is depicted in Fig. 6.9. At lower q^2 values, the efficiency of the reweighted sample is substantially lower. However, in the signal dielectron mass region the difference is far less severe, with the efficiency before and after reweighting roughly equal in the most populated bins.

With the reweighting deemed successful, its result is used from here on. Whenever the simulated LHCb data sample is used, it is understood to be weighted even if not explicitly stated.

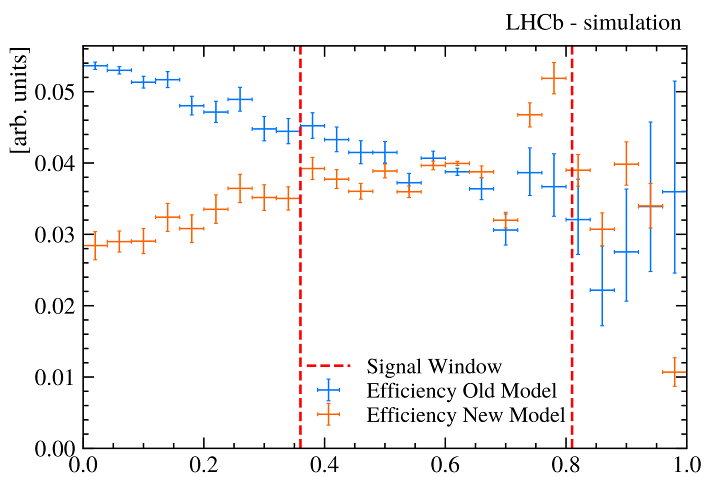


Figure 6.9: Unreweighted and reweighted simulated LHCb data efficiency in q^2-bins.

6.2 Normalisation and yield estimate

Having a theoretical prediction for the branching ratio of the signal decay, it is wished to translate it into a prediction of the observed number of events. In principle this could be done using equation 6.1

$$N_{exp} = \mathcal{L}_{int} \times \sigma_{b\bar{b}} \times 2f_u \times \mathcal{B}(B^+ \to \mu^+\nu_\mu e^+e^-) \times \epsilon. \tag{6.1}$$

The number of B mesons produced at the LHCb experiment over a given time is equal to the integrated luminosity, $\mathcal{L}_{int}$, times the $b\bar{b}$-pair production cross section times double the corresponding quark fragmentation fraction, f_u in the case of the B^+ meson. The branching ratio $\mathcal{B}(B^+ \to \mu^+\nu_\mu e^+e^-)$ then denotes how likely a B^+ meson is to decay via the signal channel. Lastly, the fraction of the occurring signal decays that are on average observed by the detector is represented by the efficiency, ϵ.

However, all of this quantities come with uncertainties. Especially the production cross-section is not known precisely. To circumvent this problem, the way taken in practice is to normalise the channel to a reference channel. For this, a similar decay with a well determined branching ratio and a clear peak is chosen.

In our case the decay $B^+ \to K^+J/\psi(e^+e^-)$ is used. Like $B^+ \to \mu^+\nu_\mu e^+e^-$, it is a weak B^+ decay resulting in three charged particles, including an electron pair. In contrast to the signal channel, the electron pair is not produced by a virtual photon on the ρ/ω resonance, but by a real J/ψ with a mass of 3069.9 MeV. Additionally, instead of a muon and a neutrino, only a kaon is produced. A tree level diagram of the decay is shown in Fig. 6.10.

Due to all final state particles of this decay being reconstructable, the appropriate variable to consider when searching for its peak is the invariant three body mass. For the calculation of this feature, we now assume that the muon candidate has the mass of the K^+. Additionally, the momenta of the electrons are scaled to be exactly on the J/ψ-mass

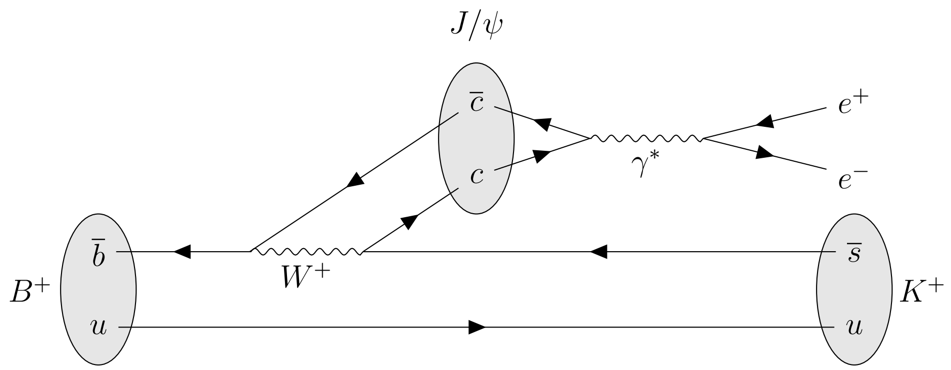


Figure 6.10: Dominant tree level Feynman contribution of $B^+ \to K^+J/\psi(e^+e^-)$.

for the calculation of the three body mass: $p'(e^{\pm}) = p(e^{\pm}) \times m(J/\psi)/m_{ee}$.

We expect the reference decay to be contained in the misIDμ sample. It has undergone the same trigger and stripping selection as the signal channel data sample, save for the muon PID requirements which it lacks. However, it has an inbuilt prescaling of 1%. The offline selection is not applied. To select specifically candidates of the reference decay, we require them to pass a `probNNk` threshold of 0.5 on the kaon candidate track. As mass windows we require $m_{ee} \in [2650, 3300]$ MeV and $m_{J/\psi(K^+e^+e^-)} \in [5200, 5500]$ MeV. Here, $m_{J/\psi(K^+e^+e^-)}$, is the invariant three body mass with the electrons scaled to the J/ψ mass.

To obtain the yield of the reference channel, we aim to fit it with a double sided Crystal Ball function, described in equation 6.2

$$f(x; N, \mu, \sigma, \alpha_L, n_L, \alpha_R, n_R) = N \times \begin{cases} A_L \times \left(B_L - \frac{x-\mu}{\sigma}\right)^{-n_L}, & \text{for } \frac{x-\mu}{\sigma} < \alpha_L \\ \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), & \text{for } \alpha_L \leq \frac{x-\mu}{\sigma} \leq \alpha_R \\ A_R \times \left(B_R - \frac{x-\mu}{\sigma}\right)^{-n_R}, & \text{for } \frac{x-\mu}{\sigma} > \alpha_R, \end{cases} \tag{6.2}$$

with

$$A_{L/R} = \left(\frac{n_{L/R}}{|\alpha_{L/R}|}\right)^{n_{L/R}} \times \exp\left(-\frac{|\alpha_{L/R}|^2}{2}\right) \quad \text{and} \quad B_{L/R} = \frac{n_{L/R}}{|\alpha_{L/R}|} - |\alpha_{L/R}|. \tag{6.3}$$

The double-sided Crystal Ball function is a differentiable function with a Gaussian peak and tails following independent power-laws. It is able to describe large tails, which are often found when energy is lost before detection, for example due to the presence of a neutrino or a large amount of Bremsstrahlungs radiation.

The fit is done in two steps using the fitting package zfit [51]. At first, only a double-sided Crystal Ball function is fitted to the simulated sample of the reference channel. The selection requirements described above are imposed on this sample as well. From this fit, a plot of which can be found in the Appendix A.2.2, the parameters describing the tails $\alpha_{L/R}$ and $n_{L/R}$ are extracted.

In a second stage, the sum of a double-sided Crystal Ball function and an exponential function are fitted to the reference channel peak in recorded LHCb data. The exponential function is chosen to cover an expected combinatorial-background contribution and has two free parameters, yield and decay constant. For the double-sided Crystal Ball function the parameters obtained in the first fit are set as fixed and only μ, σ and the yield, N, are fitted.

The fit is shown in Fig. 6.11. We find that it describes the observed data well and extract the yield of the reference channel contribution as 4300 ± 90 decays. After scaling it by 100 to account for the prescaling of the misIDμ sample, one can use this result to

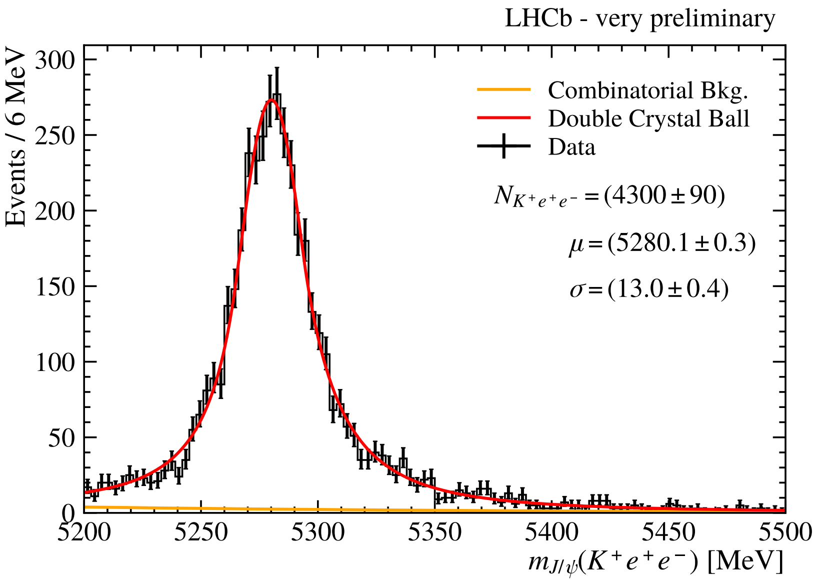


Figure 6.11: Reference channel fit of a double-sided Crystal Ball function and an exponential to LHCb data.

postulate equation 6.1 for both signal and reference channel and divide the equations. This results in equation 6.4, which relates the number of signal decays expected to be observed to the number of observed reference channel decays. Besides that, the only dependencies left are on the branching fractions and on the efficiencies.

$$N^{expected}_{\mu^+\nu_\mu e^+e^-} = N^{observed}_{K^+e^+e^-} \times \frac{\mathcal{B}(B^+ \to \mu^+\nu_\mu e^+e^-)}{\mathcal{B}(B^+ \to K^+J/\psi(e^+e^-))} \times \frac{\epsilon(B^+ \to \mu^+\nu_\mu e^+e^-)}{\epsilon(B^+ \to K^+J/\psi(e^+e^-))} \tag{6.4}$$

For both channels, the efficiency is treated as the product of two accessible partial efficiencies,

$$\epsilon = \epsilon_{acc} \times \epsilon_{selection}. \tag{6.5}$$

Of these, ϵ_{acc} describes the efficiency of generator level decays with respect to the detector acceptance, as provided in Tab. 5.4. The selection efficiency, $\epsilon_{selection}$, is obtained by dividing the number of decay candidates that pass the selection requirements by the total number of candidates that have been generated to produce the considered sample.

Importantly, it is the efficiency ratio that is relevant. Simulated data might not capture the impact of detector effects perfectly, but it can be assumed to be incorrect in a systematic fashion. By dividing the efficiencies of two similar decays, these errors of selection efficiency should roughly cancel.

The inputs of the calculation are listed in Tab. 6.2. The shown branching fraction of the signal decay is only corresponding to the signal dielectron mass window

Table 6.2: Branching ratios and efficiencies of signal and reference decay channel.

Decay channel	$\boldsymbol{\mathcal{B}} \times 10^8$	$\boldsymbol{\epsilon}_{\mathbf{acc}} \times 10^2$	$\boldsymbol{\epsilon}_{\mathbf{selection}} \times 10^3$	$\boldsymbol{\epsilon} \times 10^5$
$B^+ \to \mu^+\nu_\mu e^+e^-$	2.8 ± 0.4	50	1.35 ± 0.03	67.6 ± 1.3
$B^+ \to K^+J/\psi(e^+e^-)$	6090 ± 120	17.25 ± 0.08	34.2 ± 0.5	590.1 ± 0.8

$m_{ee} \in [600, 900]$ MeV. The efficiency of the signal decay already takes into account all selection requirements given in chapter 5, including the offline selection. While the lower efficiency of the signal decay in comparison to the reference channel results from a multitude of effects, one can be singled out and that is the presence of an additional neutrino. By itself, this already causes a loss of reconstruction quality, but it also necessitates adaptions that in turn lower efficiency, such as the usage of the corrected mass, which has a large tail in the lower mass region, which cannot be used.

After inserting extracted yield, branching ratios and efficiencies into equation 6.4, the number of expected signal decays after imposing the chosen selection requirements is obtained,

$$N^{expected}_{\mu^+\nu_\mu e^+e^-} = 23 \pm 3. \tag{6.6}$$

The uncertainty of this value mainly stems from the uncertainty of the theoretical prediction of the branching fraction. This in turn comes in a large part from the uncertainty of $|V_{ub}|$.

Chapter 7

Background

This chapter covers background contributions that are expected in our sample besides the signal. Background refers to any decay candidates that are not stemming from our signal decay. In section 7.1 combinatorial background is covered, in which a decay candidate is created from tracks that are not connected in a physical process. We construct a BDT classifier to reduce this type of background and estimate its shape to be exponential. In section 7.2 misidentified backgrounds are discussed. These are background contributions stemming from the assignment of wrong PID to a particle track. We estimate both number and distribution of these type by applying efficiency-derived weights on reference data samples. Following this, the isolation variable is introduced in section 7.3. It is the output of a pre-made BDT and provides discriminatory power against background with a surplus of charged tracks. In section 7.4 we describe background resulting from the decay of light mesons, whose final state differs from the signal one only by an additional photon. We use simulated data to estimate the expected number of these type of background decays in the signal dielectron mass window. Finally, in section 7.5, B_c^+ decays with the same final state as our channel are examined.

7.1 Combinatorial background

Combinatorial background denotes background which is not purely due to a single physical process, but a composite of particle tracks from (partially) unrelated processes that are mistakenly reconstructed as signal decay candidate.

Due to the high centre-of-mass energy of the pp collision, an abundance of particles is created in a multitude of processes. The arrangement of tracks in an order similar to that expected of the signal decay is possible and cannot be prevented. Some of these background candidates can be excluded by the usage of available kinematic variables in a multivariate analysis. However, any exclusion comes at the price of lowering signal efficiency and it is not possible to completely eliminate this background. Hence, it is necessary to estimate its size and model its shape for the mass fit.

7.1.1 Construction of a boosted decision tree classifier

We use a gradient boosted decision tree classifier (BDT), as described in section 4.3.2, implemented via XGBoost [52]. The classifier is trained on the simulated signal sample against the same sign data sample, using kinematic variables. In both cases we apply the full selection and exploit the entire corrected mass range in the signal dielectron mass window.

Choice of training samples

Any Standard Model process producing a muon and two same charge electrons is expected to be suppressed strongly by the necessary couplings. From this it follows that a decay candidate consisting of three such tracks is likely combinatorial in nature. Further, the reconstruction itself does not care about the relative sign of two tracks, suggesting that combinatorial background with same charge electron tracks and opposite charge electron tracks should be equally likely. There should also be no difference in the distributions of kinematic features.

This means that one expects identical distributions of this type of background in both signal and same sign sample, motivating the choice of the same sign sample as training data.

However, in addition to combinatorial decay candidates, the same sign sample contains further contributions. For example, decay candidates with incorrectly charged tracks can also be obtained by misidentifying the particles of a physical process. An example would be the cascade decay $B^0 \to e^+\nu_e D^- \to e^+\nu_e\mu^-\bar{\nu}_\mu K^0 \to e^+\nu_e\mu^-\bar{\nu}_\mu\pi^+\pi^-$, if the π^+ is misidentified as an electron and the π^- is missed. Generally, as the most abundant particle tracks in the detector are pion tracks, a significant overlap between misidentified background and combinatorial background is to be expected. A clean separation of these two types of background is not possible.

Hence, training against the same sign sample results in a classifier that is not just rejecting strictly combinatorial background, but more generally background that is not signal-like in its topology and kinematics. For reasons of practicality it is still referred to as combinatorial BDT.

As a stand in for the signal we use the signal simulation sample with the weights obtained in chapter 6.

Choice of training parameters

To train the BDT the following features are chosen:

- The χ^2_{vertex} value of the B^+ decay vertex. The χ^2 value of the fit of the three tracks into a secondary vertex.
- Decay time of the B^+.
- The cosine of the B^+ direction angle, i.e. the angle between the B^+ momentum and the direction from its production to its decay vertex.
- The χ^2 value of the minimum impact parameter of e^+ and e^-, χ^2_{IP}. The impact parameter is the perpendicular distance of the particle track to the primary vertex.
- Transverse momentum, p_T and pseudorapidity, η, for all three leptons.
- The χ^2 value of the fit to all three tracks, divided by the number of degrees of freedom, $\chi^2_{track}/ndof$.
- The ghost probability, p_{ghost}, for all three leptons. The output of a neural network trained to detect ghost tracks, i.e. tracks arising from the combination of unrelated subdetector hits. Takes into account information not used directly for the track fit.

These parameters are chosen for their separating power. Plots comparing the distributions between same sign data and simulated signal sample are found in A.3.1. Plots depicting their correlation in both the simulated signal sample and the same sign sample are found in A.3.2. Further considered were:

- Kinematic variables pertaining to the dielectron pair. Dropped due to strong correlation with those of the individual electrons.
- Minimal impact parameter of the muon. Dropped due to strong correlation with χ^2_{vertex} of the B^+ secondary vertex.
- χ^2 of the distance of closest approach of the lepton tracks. Dropped due to lack of separating power.
- χ^2 of the distance between primary and secondary vertex. Dropped due to lack of separating power.

After the training process described further below, we also have access to the feature importances of the final classifier, depicted in Fig. 7.1. The importance is calculated by considering every decision that is made in any of the trees used and then summing

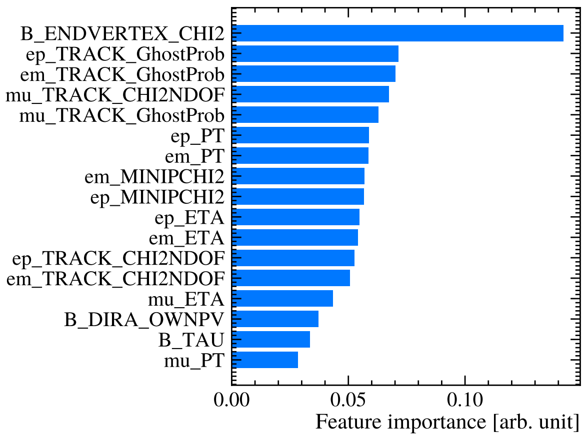


Figure 7.1: Importances of the BDT training features.

up the loss reduction due to splits in one given variable[1]. Following this metric, the most important variable is χ^2_{vertex}. As combinatorial background comes from accidental matching of tracks, it is sensible that vertex quality is often lower than for signal tracks. Similarly, χ^2_{track} and p_{ghost} can also be understood as measures of track quality. Taken together this suggests that the classifier works primarily by focusing on the reconstruction quality of decay candidates.

Hyperparameter optimization

To optimize classifier performance, we undertake a search for optimal hyperparameters. For this, different configurations of hyperparameters are used when training the classifier and then compared in separation power, the metric being the area under the ROC curve. This process is done using 4-fold cross validation. The configurations for testing are chosen in two different ways, by following a grid and by random choice. For the first method a short list of accessible values is provided for each hyperparameter and the optimization then tests each point of this grid. In the second method either a continuous space or a range of integers is provided, depending on what is possible for the hyperparameter in question, and the optimization picks a random possible value for each hyperparameter. 500 different hyperparameter configurations are covered by the random search. Implementation of both searches is done using the scikit-learn library [53]. The hyperparameters and the values considered in the optimization process were:

- Learning rate, η, which downscales the corrective term added in each iteration. Smaller means slower change, which protects from overcorrecting.

[1]This is just one metric useable to estimate the relevance of BDT features. It tends to favour features that enable lots of small/medium decisions over those that are used seldomly with a big impact. To decide which features to use, we considered loss of AUC due to feature elimination.

- Minimum loss reduction, γ, required for each new partition. Higher values disincentivise decisions that do not offer much discriminative power.
- Maximal depth of decision tree, meaning the maximal number of decisions allowed to characterize a leaf. A higher depth allows more complex models, but also risks overfitting.
- Minimal child weight, w_{min}, meaning the minimal summed weight required to be contained in any new leaf that is partitioned off. Higher values disincentivise decisions that only affect very few decays.
- Number of estimators, n_{est}, the number of iterations before the algorithm is stopped. Needs to be high enough for the approximation to be meaningful, but setting it to high may lead to overfitting.
- Subsample rate, percentage of training data that is randomly chosen to be used for a new iteration. It is recommended to stay above 50%. A lower subsample rate prevents overfitting, but also lowers the number of decays available for training.

Tab. 7.1 gives the hyperparameter values considered in the optimization process and the optimal configurations achieved. The optimal configurations agree in setting low γ, w_{min} and the subsample rate. They also place η and n_{est} at similar values. The results are either distanced from the edge of parameter space or at an edge that is given by the method itself, which suggests that the hyperparameter configurations found are indeed optima and not artifacts of a too tightly constrained range.
However, both of these configurations lead to significant overtraining, as judged by the comparison of ROC curves of test and training sample. The plot is found in the appendix

Table 7.1: The considered hyperparameters, their values considered by the search and the best configurations found. Curly brackets denote discrete sets, the shown numbers are the only possible values. Round brackets denote uniform intervals, any number or integer inside is accessible by the search with equal probability. A log() denotes logarithmically uniform intervals, values are drawn according to a distribution uniform on a logarithmic scale. The last column is the configuration chosen to suppress overtraining.

	grid range	**grid result**	**rand. range**	**rand. result**	**choice**
η	$\{0.01, 0.03, 0.07, 0.1, 0.3\}$	0.07	$\log(0.001, 0.1)$	0.043	0.043
γ	$\{0, 0.3, 1, 2, 10\}$	0.3	$\log(0.01, 30)$	0.031	10
depth	$\{4, 5, 6\}$	4	$\{4, 5, 6\}$	5	5
w_{min}	$\{1, 10, 100\}$	1	$\log(1, 100)$	2.56	15
n_{est}	$\{350, 400, 450, 500, 550\}$	450	$(250, 550)$	442	442
s. rate	$\{0.5, 0.75\}$	0.5	$(0.5, 1)$	0.51	0.5

A.3.3. To curtail this we manually select a set of hyperparameters with higher values for γ and w_{min}. This leads to a small loss of AUC score but a significant reduction of overtraining. A further decrease of overtraining is not achievable without significant loss of AUC.

Classifier output

Having chosen a hyperparameter configuration we train the BDT using five fold cross-validation. To evaluate its discriminating power, we consider the resulting ROC curve and AUC score, as well as the score distribution in both same sign and signal sample. The ROC curve is depicted in the left plot of Fig. 7.2. In the right plot we see the score distribution of the classifier on both Monte Carlo simulation sample and same sign sample. We find that both are smooth without abberations and the classifier provides an AUC of 0.91 which gives it a strong separation power. However, it is also clear that there is still some overtraining left.

The optimal threshold on the combinatorial BDT score is determined later on in subsection 8.2. The optimization takes into account both the combinatorial BDT score and the isolation variable, introduced in section 7.3. To not only work on high background samples, preliminary requirements on both of these variables are used in some of the further studies. These correspond to a combinatorial BDT signal efficiency of 80% and an isolation requirement signal efficiency of 95%, resulting in a total efficiency of 76%.

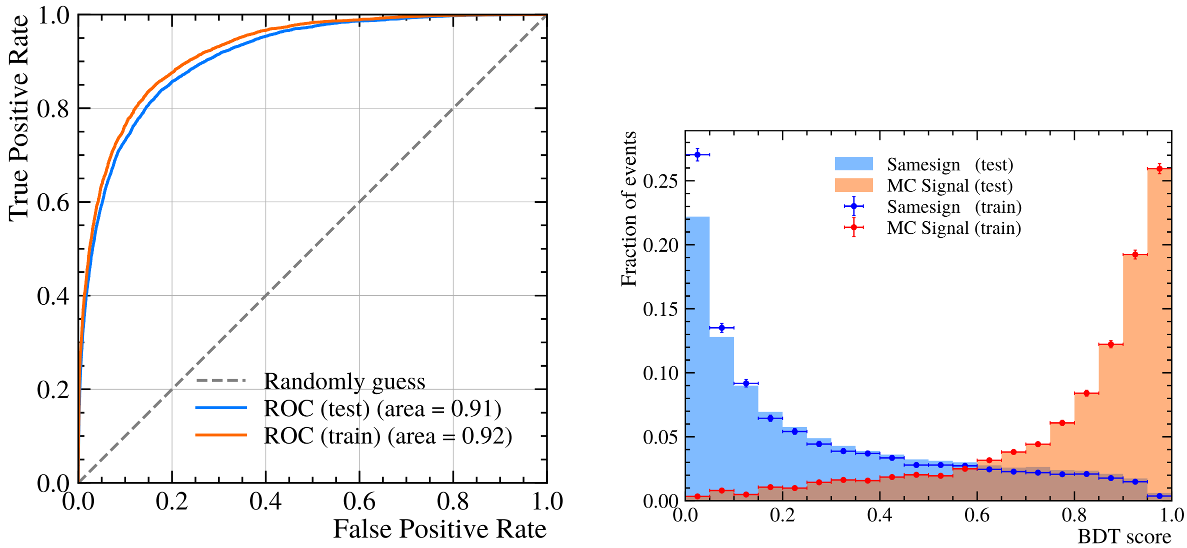


Figure 7.2: ROC curve (left) and BDT score distribution (right) of the final classifier.

7.1.2 Shape estimation

While the threshold on the combinatorial BDT score is not yet determined, it is already apparent that it is unlikely to completely eliminate all combinatorial background. Thus, we also wish to construct background toy models, both for the optimization itself and to gauge the quality of our mass fit and sensitivity estimate. Therefore it is necessary to model the combinatorial background contribution.

Often the sidebands in corrected mass, m_{corr}, are used to extrapolate the combinatorial background shape and yield. For $B^+ \to \mu^+\nu_\mu e^+e^-$ that would be $m_{corr} < 4500\,\text{MeV}$ and $m_{corr} > 7000\,\text{MeV}$. In our case this is not possible, since the high sideband is too sparsely populated and the low sideband is not expected to be a good choice, as is be shown later.

Instead we consider the same sign sample, which we used to train the BDT. As explained before, it is likely that the same sign sample consists not purely of combinatorial background. This is no issue when training the BDT, but as we also estimate the contribution of misidentified background, the usage of the same sign sample as a direct proxy for the combinatorial background runs the risk of double counting. That said, the same sign sample can still provide information about the expected shape of the combinatorial background. We assume a falling exponential. This is a common choice to model combinatorial background, known to work well with higher mass tails, but to lose modeling power in the lower mass range. An exponential model is also used in the $B^+ \to \mu^+\nu_\mu\mu^+\mu^-$ analysis [5].

We further look at a subsample, where we impose that all three leptons have the same charge. This subsample, called ”three-of-a-kind” (ToaK), can be assumed to be nearly without any physical structure. While it contains too few candidates to be used effectively as training data for the BDT[2] or as a proxy for the mass fit, it can be instructive to examine it.

The corrected mass distribution of both the same sign sample and the ToaK subsample can be seen in Fig. 7.3. We consider the mass window $m_{ee} \in [600, 900]\,\text{MeV}$ and $m_{corr} \in [3000, 9000]\,\text{MeV}$. The left hand plot depicts both the same sign distribution and the ToaK subsample. While differing in yield, the two curves roughly share their shape, with a plateau in the low mass range that leads into an exponential decay.
When examining the curve, one finds that the inflection point, at which the exponential assumption breaks down, is somewhere between 4000 MeV and 4500 MeV, justifying our lower corrected mass limit. On the other end, the number of candidates starts to run out after 7000 MeV, which fits with the exponential assumption of this area, but does not

[2] We can consider the ROC curve of the ToaK sample and find that it does only slightly differ from that of the general same sign sample A.3.4. One can consider this to be a test of the BDT on a purely combinatorial sample.

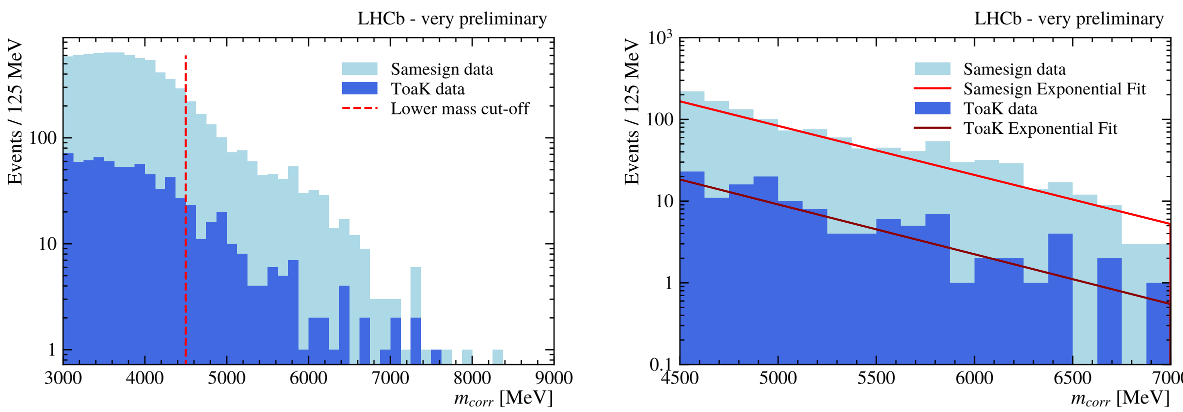


Figure 7.3: On the left, the corrected mass distribution of same sign and three-of-a-kind data with $m_{corr} \in [3000, 9000]$ MeV. On the right, the corrected mass distributions with $m_{corr} \in [4500, 7000]$ MeV and an exponential fitted to them.

confirm it.

In the right hand plot we show the fit of exponential functions to both same sign and ToaK data. We now consider only $m_{corr} \in [4500, 7000]$ MeV, as this is the range we expect an exponential to be a valid assumption. The exponential fits the data well, even though the ToaK sample is low in statistics.

From this we induce that the exponential assumption for the shape of the combinatorial contribution is justified in the signal window. However, due to the same sign sample potentially having an overlap with the misidentified background and the ToaK subsample being very low in size, we cannot use them to access yield or slope of the exponential. To get these, we later use an iterative procedure and the dielectron mass sideband $m_{ee} \in [900, 1200]$ MeV in section 8.2.

7.2 Misidentified background

Misidentified background (misID) are all types of backgrounds which include at least one particle track being actually a misidentified particle.

An example for this would be the decay $B^+ \to e^+ \nu_e \bar{D}^0 (K^+ e^- \bar{\nu}_e)$. If the kaon is mistakenly identified as a muon, the decay has the right particle composition to be recorded as a signal candidate. Cabibbo favoured, with a branching ratio of $O(10^{-4})$[6], this decay is also far more abundant than the signal channel, making it potentially dangerous, even if the misID rate is low.

Problematically, misID can be of a multitude of different origins. It can even be combinatorial in nature. This makes the usage of simulated data for estimating the contribution impractical. Instead, we use a data driven method to estimate both shape and yield of misID background in our signal data sample.

It follows the approach of the analysis on tests of lepton universality in $b \to s\ell^+\ell^-$ decays[54], which faced a similar issue concerning hadron to electron misID in the decays $B^+ \to K^+e^+e^-$ and $B^0 \to K^{*0}e^+e^-$.

7.2.1 Introduction to the method

To understand the method it is best to start by considering a simpler situation. We commence by considering a single lepton track with a single type of hadron misID background. To distinguish between them, a PID feature is used. This is depicted in Fig. 7.4. Both types of tracks have a different distribution in the PID feature and a threshold requirement is imposed at a chosen value. This requirement eliminates a large part of the hadron background, but not its entirety. A number of hadron tracks, N_{had}^{pass}, passes this threshold. It is equal to the total number of hadrons before the selection requirement, N_{had}, times the efficiency of the requirement, $\epsilon_{\texttt{lepPID}}$. Conversely, a number of hadrons, N_{had}^{fail}, does not pass this threshold and it is equal to N_{had} times $1 - \epsilon_{\texttt{lepPID}} = \epsilon_{\texttt{inverse}}$.

Having exhausted the discriminative power of the PID feature, we are left with a remnant of misidentified hadrons, which we cannot separate from the lepton tracks. Only

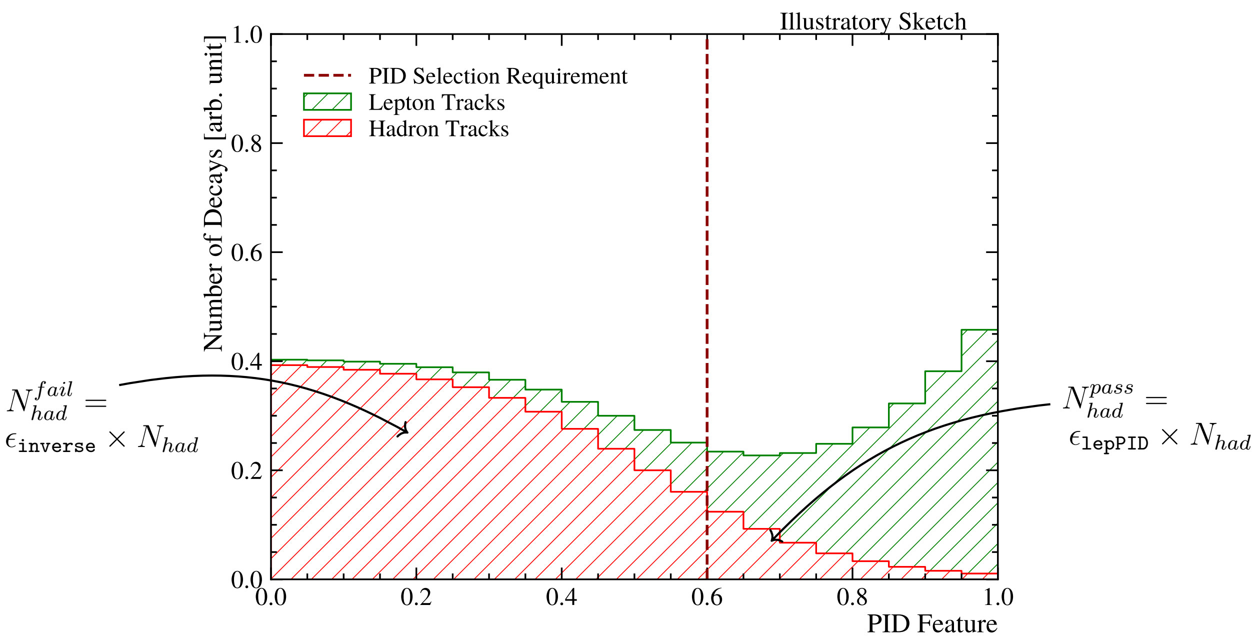


Figure 7.4: A sketch showing possible distributions of a lepton species and a hadron species in a PID feature. The imposition of a requirement on this feature splits the total number of hadron tracks into those that pass the threshold and those that fail it. The relationship is given by the corresponding efficiency.

the total number of tracks that pass, $N^{pass} = N^{pass}_{had} + N^{pass}_{lep}$, is accessible. However, we might be able to estimate the size of this contribution. For this we use the fact that the failed tracks are mostly made up of hadron background, $N^{fail} \approx N^{fail}_{had}$.

If the efficiency of the requirement is known, we can approximate the number of hadron tracks that pass the PID requirement,

$$N^{pass}_{had} = \frac{\epsilon_{\texttt{lepPID}}}{\epsilon_{\texttt{inverse}}} \times N^{fail}. \tag{7.1}$$

While this result is useful, it is not yet what we are really after. Not just the number, but also the m_{corr} shape of the misID background is needed. The shape can change under the imposition of a PID requirement, as both the efficiency of the PID system and the corrected mass depend on the kinematics of the particle track. To account for this, the PID efficiencies need to be treated as functions of the kinematic parameters of the individual tracks.

In practice, where efficiencies are gained from evaluating reference samples with limited statistics, this is done by constructing bins in transverse momentum, p_T, and pseudora-

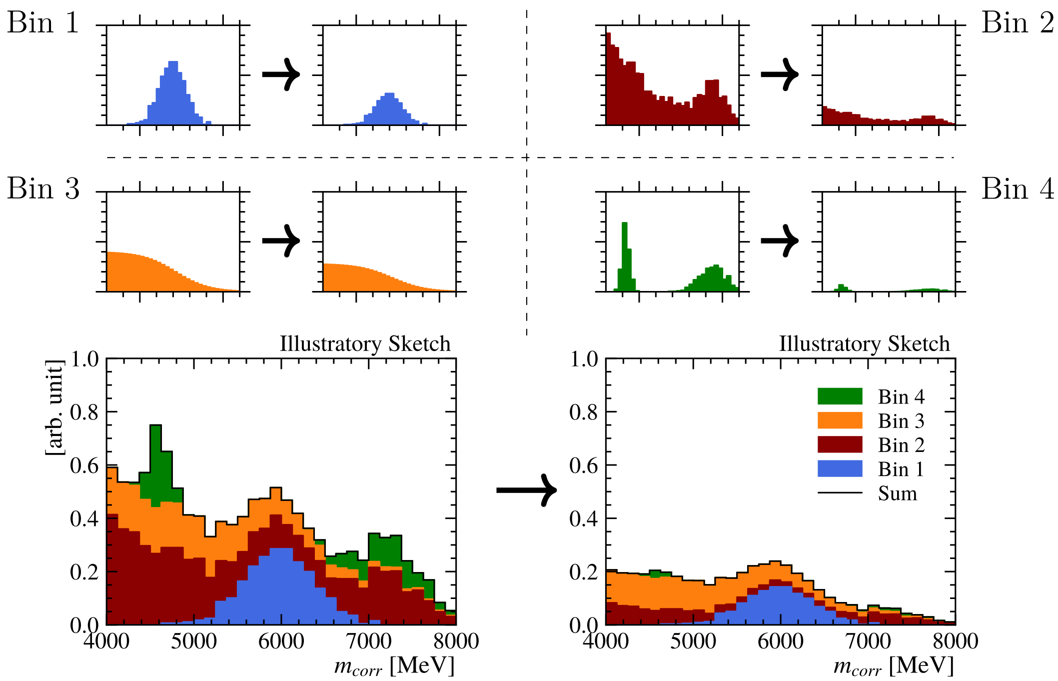


Figure 7.5: A sketch showing how the hadron contribution might look in different (p_T, η)-bins. Top half shows the individual contributions, while the lower half shows the summed contribution. In all five cases, the contribution before the PID requirement is shown on the left of the arrow and the contribution with the requirement in place is shown on the right of the arrow.

pidity, η, in which efficiencies are treated as constant. This simply means that the the estimate done before is now done several times, once for each bin. Values for $\epsilon_{\texttt{lepPID}}$ and $\epsilon_{\texttt{inverse}}$ are assigned to each individual track accordingly. The efficiency ratio is then used as the weight of the decay candidate.

The kinematic dependence of the efficiency allows an approximation of how the shape of the hadron contribution changes when imposing the PID requirement. A sketch of this is given in Fig. 7.5. It shows a possible m_{corr} distribution of the hadron contribution split into four (p_T, η)-bins. In the upper half the four individual distributions and how they are impacted by imposing the PID requirement are shown. The total number of decays decreases in each case, but the shapes of the four contributions do not change. In the lower half, the contributions of all four bins are added up. As the PID requirement efficiency is different for each bin, the shape of the summed contribution does change.

Moving on from the simplest case, we consider the case of two lepton tracks, each of which can potentially be a misidentified hadron. Again, a binning scheme in p_T and η is used, now for each track individually. We consider the situation for decays that share their binning assignment, i.e. the efficiencies for a given track are constant for all decays.

With two tracks, there are decay candidates in which both lepton tracks are misidentified. This means that the number of decays in which both tracks pass the PID requirement, $N^{p,p}$, is the the sum of four contributions in which the tracks are either leptons, a hadron and a lepton or two hadrons,

$$N^{p,p} = N^{p,p}_{\ell,\ell} + N^{p,p}_{h,\ell} + N^{p,p}_{\ell,h} + N^{p,p}_{h,h}. \tag{7.2}$$

Measured contributions are denoted by $N^{p/f,p/f}$, with the p/f denoting whether a track has passed or failed the lepton requirement. The true underlying contributions are denoted with $N^{f/p,f/p}_{h/\ell,h/\ell}$, with h/ℓ denoting whether the track is a hadron or a lepton.

To obtain an estimate of the contributions involving at least one hadron track we initially sum up the individual single misID estimates for either lepton. They are obtained, as described above, by using the measured number of decays for which one track fails the PID requirement and one track passes, $N^{f,p}$, and weighting it with the ratio of the efficiencies of the failed track, $\epsilon^{f}_{\texttt{lepPID}}/\epsilon^{f}_{\texttt{inverse}}$. However, among these decays will also be some in which there are two hadron tracks, one of which failing and one passing the PID requirement, $N^{f,p} = N^{f,p}_{had,lep} + N^{f,p}_{had,had}$. By weighting these sort of decays with the ratio of the efficiencies, ϵ^{f}, of the failed track, we also obtain the contribution of two hadron decays in which both tracks pass the requirement,

$$\frac{\epsilon^{f}_{\texttt{lepPID}}}{\epsilon^{f}_{\texttt{inverse}}} \times N^{f,p} = \frac{\epsilon^{f}_{\texttt{lepPID}}}{\epsilon^{f}_{\texttt{inverse}}} \times (N^{f,p}_{h,\ell} + N^{f,p}_{h,h}) = N^{p,p}_{h,\ell} + N^{p,p}_{h,h}. \tag{7.3}$$

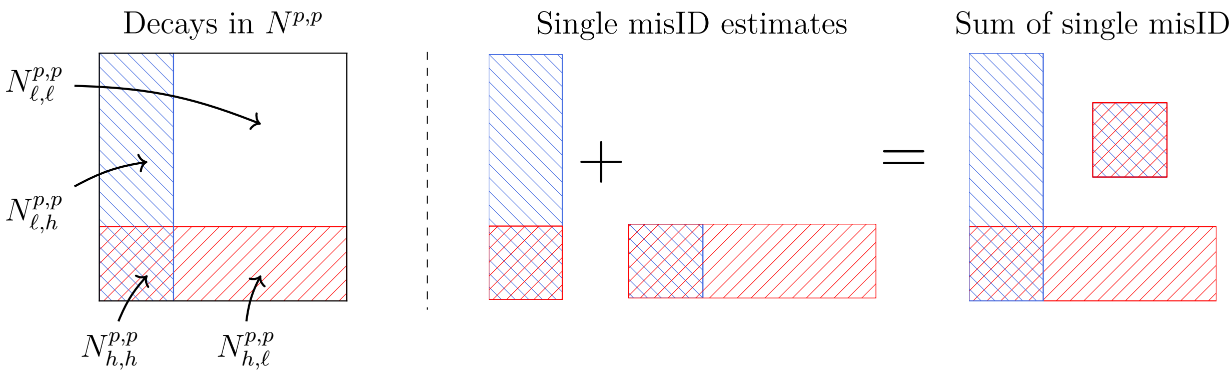


Figure 7.6: A sketch of the double misID consideration. The left image pictures the set of all decay candidates in which both tracks pass the lepton PID requirement. On the right of it the addition of the single misID estimates is depicted.

This contribution is contained in both single misID estimates. By summing them up, it is obtained twice. A sketch of this is shown in Fig. 7.6. On the left it depicts the set of all decays in which both tracks pass the PID requirement. Following equation 7.2 it consists of four contributions. On the right the two single misID estimates are summed up. As they each contain the double misID contribution once, the summed estimate has an excess contribution of exactly the size of the double misID. To recover the correct number of decays it is necessary to subtract the double misID contribution, $N^{p,p}_{h,h}$, once. Analogously to the single ID estimates, an estimate of it can be obtained by weighting the decays in which both tracks fail the lepton PID, $N^{f,f}$, with the the product of their two efficiency ratios,

$$N^{p,p}_{h,h} = \frac{\epsilon^1_{\texttt{lepPID}}}{\epsilon^1_{\texttt{inverse}}} \times \frac{\epsilon^2_{\texttt{lepPID}}}{\epsilon^2_{\texttt{inverse}}} \times N^{f,f}. \tag{7.4}$$

As in the case of single misID, the method can then be extended to also cover the shape of the estimated misID contribution by introducing an efficiency binning scheme in p_T and η for each of the tracks.

The method can be scaled up to include a third lepton track with the possibility of misidentification. One finds for the number of decays in which all tracks pass the requirement, $N^{p,p,p}$, that there are three single misID contributions, three double misID contributions and one triple misID contribution,

$$N^{p,p,p} = N^{p,p,p}_{\ell,\ell,\ell} + N^{p,p,p}_{h,\ell,\ell} + N^{p,p,p}_{\ell,h,\ell} + N^{p,p,p}_{\ell,\ell,h} + N^{p,p,p}_{\ell,h,h} + N^{p,p,p}_{h,h,\ell} + N^{p,p,p}_{h,\ell,h} + N^{p,p,p}_{h,h,h}. \tag{7.5}$$

In the case of two tracks a single misID estimate contains a double misID contribution, stemming from the misID possibility of the track passing the lepton PID requirement. In the case of three tracks, the decays in a single misID estimate have two tracks that pass the lepton PID requirement. Consequently there are two different double misID contributions in any single misID estimate. In total this results in all three possible double misID contributions being included twice in the sum of the single misID estimates. Accordingly, they need to be subtracted once.

For the triple misID, one finds that it is included once in any double or single misID contribution. As three double misID contributions are subtracted from the sum of three single misID contributions, it is not included yet and needs to be added once.

7.2.2 Implementation of the misID estimate

We now move on to use this method to estimate both yield and shape of the misID background of the signal decay $B^+ \to \mu^+\nu_\mu e^+e^-$. Initially only single misID is considered. We examine decays where one track fails the offline PID requirements, or equivalently passes their inverse.

It is assumed that the misID background is made up entirely of pions and kaons. Distinction between them is made by discriminating on `ProbNNk`: if `ProbNNk>0.1` the particle is treated as a kaon, otherwise it is treated as a pion. Though this choice of threshold is arbitrary, the systematic uncertainty introduced by it has been treated in [54] and is found to be negligible for the shape and below $O(10\%)$ when considering the yield. In principle proton misID cannot be ruled out. However, due to the higher mass of the proton it is expected to be small in size. Using `ProbNNp` as indicator, neither the misIDμ nor the data sample show any evidence of a substantial proton contribution.

As a first step, data samples are procured, in which one can find decay candidates with a single track failing the PID requirement, but the candidate passing the entirety of the selection otherwise. The data samples used for this are listed in Tab. 7.2, along with the inverted PID requirements and resulting number of decay candidates.

For the electrons, one needs to keep in mind that the given charge is relative i.e. e^+ denotes the electron track with the same charge as the muon track, while e^- denotes the track with opposite charge.

Of note is that among the positron candidates there are two times as many candidates we identify as hadrons as among the electron candidates. This is later commented upon.

The reason for the usage of the signal data sample as base for the electron misID is that, when writing this thesis, no sample was available that had the desired stripping requirements combined with a looser requirement on electron PID.

As the signal sample has strong DLL based PID requirements implemented already on stripping level[3], we are left with the option of inverting the `ProbNNe` threshold. Con-

[3]Incidentally, the reference[54] also uses a sample with PID requirements already applied on the stripping level.

Table 7.2: An overview of the samples and selection requirements used for the single misID estimation and the resulting available decay candidates. The number is given with full selection requirements except PID on the relevant lepton. It also corresponds to the entire mass range of both m_{ee} and m_{corr}.

Track	Inverted requirement	Used sample	# $\pi^{\pm}$decays	# $K^{\pm}$decays
μ^+	(`PIDmu<0.0`|`ProbNNmu<0.2`) &`isMuon==0`	misIDμ	48975	64315
e^+	`ep_ProbNNe < 0.2`	signal data	103777	11725
e^-	`em_ProbNNe < 0.2`	signal data	53437	5563

sequently only the more electron-like hadrons are considered. These are also the hadrons with the highest weights, meaning that they would likely also dominate the misID estimate of a more complete sample.

For the muon misID sample, the inverted requirement is not a total inversion of the requirement, as was done in the introduction. This is a choice made to account for the imbalance in the discriminative power of the features used. The `isMuon` variable is known to be very efficient by itself, which is why it is required to be failed.

Using these samples and requirements one can extract the quantity N^{fail} of equation 7.1. Next, the efficiencies need to be obtained.

Fore this, it is first necessary to further split up the electron and positron misID samples. This is done based on whether the particle that failed the PID requirement caused the trigger or not. The motivation of this is that efficiency is expected to differ strongly between tracks that cause the trigger and those that do not. For muons this split is not necessary, as there are no decay candidates with muon tracks that fail the PID requirement and pass the trigger.

To access the efficiencies PIDCalib2 is used. The tool is introduced and described more in depth in section 4.3.3. It can assign the efficiencies of both the PID requirement and the inverted PID requirement. This happens candidate per candidate, with dependence on η and p_T. In the upper row of Fig. 7.7 bin wise efficiency histograms are shown. The left histogram depicts the efficiency map for a pion to pass the electron PID requirement and the right histogram depicts the efficiency map of a pion to pass the pion PID requirement `e_ProbNNe<0.2&e_ProbNNk<0.1`. In both cases the electron candidate is not causing the trigger. Further histograms corresponding to different hadron to lepton misidentification possibilities are found in A.3.5.

We then divide the lepton requirement efficiency by the hadron requirement efficiency. This corresponds to a binwise dividing of the before created histograms, followed by assigning the candidates to the corresponding bins of the resulting histogram. The lower plot in Fig. 7.7 depicts the efficiency ratio histogram for pion to electron misID. For

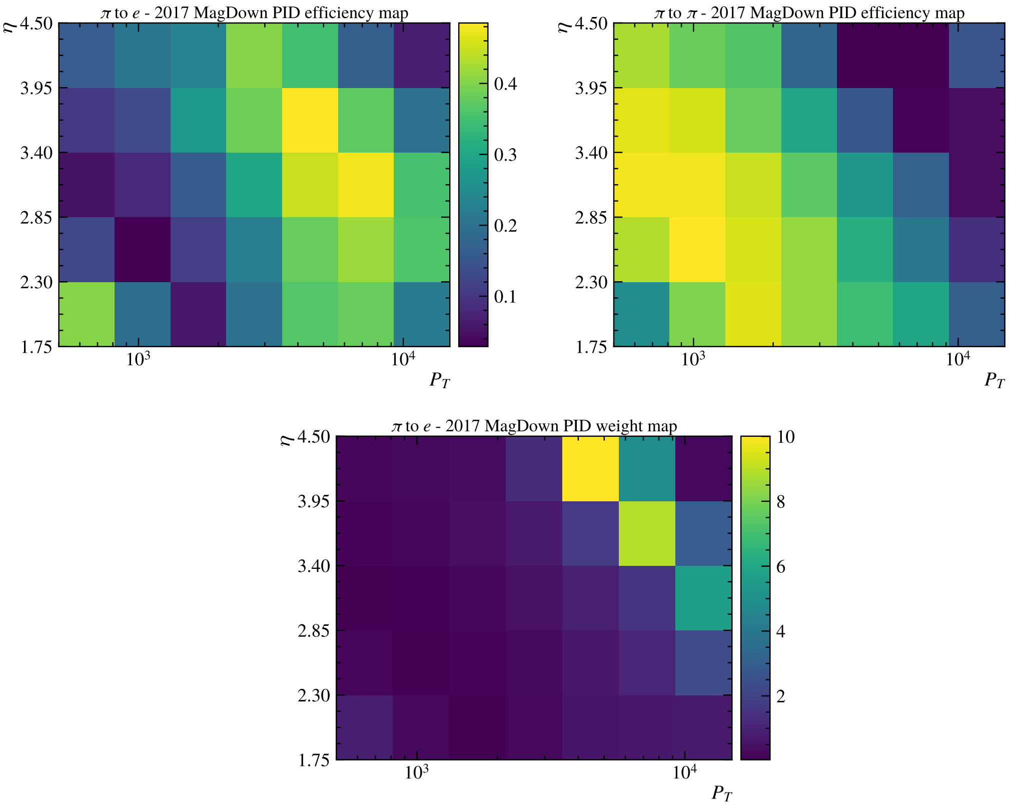


Figure 7.7: The upper left histogram shows the efficiency of the electron PID requirement on the reference pion sample, the upper right shows the efficiency of the inverted PID requirement and the lower plot shows the ratio gained from binwise dividing electron PID efficiency by hadron requirement efficiency. All plots are for data recorded in 2017 with negative magnet polarity.

the candidates of the misIDμ sample, the weights are multiplied with a factor of 100 to compensate for the prescaling of the data sample.

To estimate the contribution of multiple misID the same method is used, only applied on several leptons at the same time. The inverted requirements are imposed on two or three leptons and the assigned efficiency ratios are multiplied. For $\mu^+e^\pm$ double misID and triple misID the misIDμ sample is used and for e^+e^- double misID we use the signal sample. The choice of the former is because the muon requirements at stripping levels are already too tight to allow finding hadron to muon misID in the signal sample.

When using this method it is in principle necessary to also estimate the amount of leptons that fail the lepton PID cuts and are thus encompassed in the hadron samples, leading to an overestimate of the misID contribution. However, the expected impact of this effect scales mainly with the yield of contributions with signal final state. Both the signal channel and the B_c^+ decays sharing its final state are predicted to have low yields in the signal window. Any other $B_{u/c}^+$ decay producing at least two electrons and a muon is contributing even less due to incurring additional coupling constants. We thus neglect this effect.

7.2.3 Cross check $B^+ \to K^+ J/\psi(e^+e^-)$

To test the potency of this method when estimating expected misID background in our signal data, we consider a decay with similar topology to our signal, but a final state which differs from ours by the type of one particle track. A prime candidate for this is the decay channel also used for the normalization $B^+ \to K^+ J/\psi(e^+e^-)$, which is copiously produced around $m_{ee} = m(J/\psi) = 3096.9\,\text{MeV}$.

With the exception of the mass window, the full signal selection is implemented on the signal data sample, including the muon PID requirements. Due to the abundance of the decay, a peak of mostly $K^+ \to \mu^+$ misID might still be visible and could be compared to the prediction of our misID estimate.

As in the normalisation procedure in section 6.2, the electron momenta are scaled to be on the J/ψ-mass. To search for a signature we use the invariant three body mass, with scaled electron momenta and kaon mass hypothesis for the muon track candidate. Requirements are imposed to curtail contributions of decays of the type $B_c^+ \to \mu^+ \nu_\mu X_{c\bar{c}}(e^+e^-)$, $X_{c\bar{c}}$ being mostly J/ψ, which share the signal final state and hence have a substantially higher efficiency under signal PID selection than $B^+ \to K^+ J/\psi(e^+e^-)$. For this we use the corrected mass and the three body mass calculated with muon mass hypothesis and

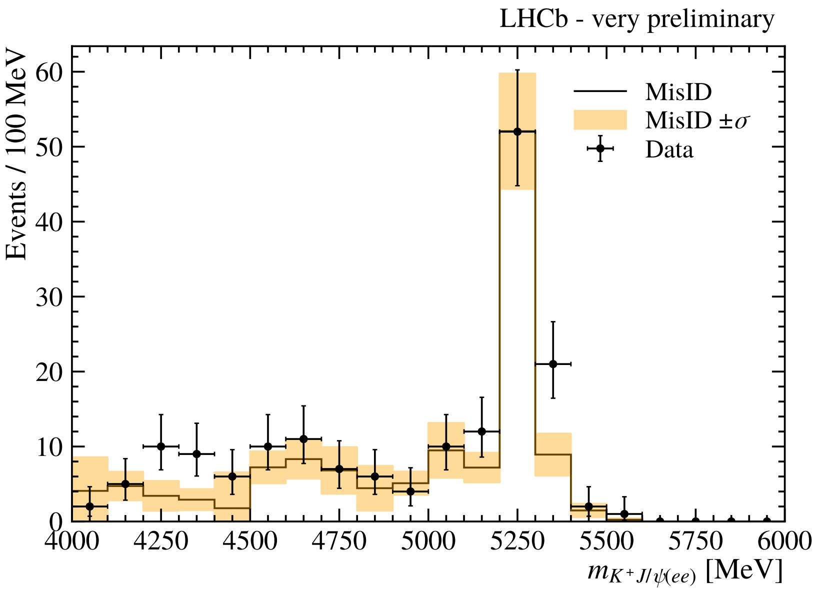


Figure 7.8: The observed $m_{K^+J/\psi(ee)}$ distribution in data compared to the estimated distribution from the weighted hadron samples.

the electron momenta fixed to the J/ψ mass. As the initial particle of this decay is a B_c^+ with higher mass than a B^+ we impose an upper limit on the corrected mass of 5500 MeV. Additionally we demand that $m_{corr} - m_{\mu ee} < 200$ MeV, to cut away those candidates where an unreconstructed particle, the neutrino, takes away too much energy. The preliminary BDT requirement is also applied.

The resulting distribution in m_{K^+ee}, with $m_{ee} \in [2700, 3300]$ MeV can be seen in Fig. 7.8. A peak in the data located at the B^+ mass is clearly visible and successfully matched by our estimate of the misID contribution, without significant discrepancy.

The estimate slightly undershoots the measured data. This is expected, as there can never be more misID candidates than there are candidates in total. Still, there is no significant disagreement between data and estimate.

This cross check showcases that our method is able to replicate the shape and yield of misID in our signal sample. A caveat to this is the low number of candidates in the considered mass region and the large uncertainties attached to our method.

Additionally there is a remnant of B_c^+ contribution that has not been completely cut away by our selection. By scaling the results of section 7.5, this contribution can be estimated. A modified plot containing this is found in A.3.6. It does not change any observation made.

As a last comment, the K^+ can decay into a muon and neutrino. As the neutrino carries away some momentum, this can lead to an underestimate of momentum if the kaon decays before the magnet. The track can then drop out of the kinematic selection of the PIDCalib reference sample. Tracks of this sort are not captured by this method, meaning that $K^+ \to \mu^+$ misID is underestimated. While it is expected that this is a minor effect, there is currently no quantitative estimate of its impact available. The issue is known to the PIDCalib community and work is in progress to alleviate it. For this cross check specifically, this is not an issue, as the momentum loss also leads to a shift in the invariant three body mass, meaning that decay candidates of this sort do no contribute to the peak seen in the data.

7.2.4 Comparison with the high m_{ee} sideband

Distributions of the weighted hadron samples should behave like those expected for the corresponding misID contributions in our signal sample. With the signal mass window blinded, the next closest mass window available to compare our estimate with data is the closest high dielectron mass sideband where the corrected mass requirement of the signal window $m_{corr} \in [4500, 7000]$ MeV is shared, but the dielectron mass range $m_{ee} \in [900, 1200]$ MeV is shifted by 300 MeV. In this subsection the preliminary BDT selection is used.

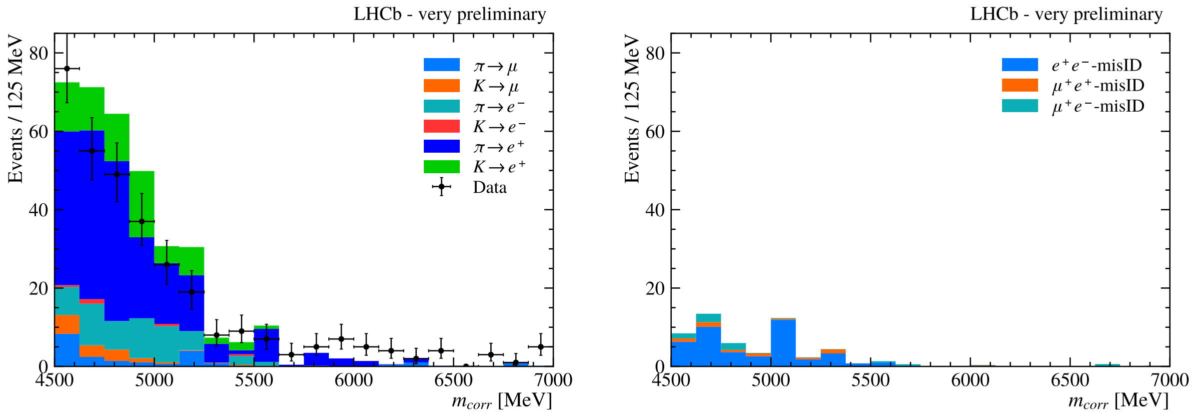


Figure 7.9: Stacked up mass distributions of different misID types. On the left is the single misID estimate compared with the observed data. On the right is the double misID estimate. The corrected mass window is the same as for the final signal selection, but dielectron mass is within $[900, 1200]$ MeV.

In Fig. 7.9, the estimated single and double misID curves are depicted. The left image shows the summed up single misID estimate, split up into its contributions and compared with the measured data in this mass region. It is apparent that hadron to electron misID is contributing more than hadron to muon misID, which fits the understanding that the muon PID capabilities of LHCb are stronger than those for electrons or light hadrons. Further, among the electrons it is the electron with the same charge as the muon, which is more likely to be a misidentified hadron.

A large source of this are decay candidates resulting from the decay chain of the $\bar{b}$ quark, such as $\bar{b} \to \bar{c} \to \bar{d}/\bar{s}$. In each transition a $W^{\pm}$ boson is emitted, which can decay into a lepton and a neutrino. In addition at least one hadron results from this chain. The two leptons are of opposite charge, the hadron is of the same charge as the B^+ meson.

In the case of electron misID, one of the leptons is a muon and the other lepton is accordingly an electron of opposite charge. If the produced hadron is misidentified, it is identified as a positron. If a random particle track matches to a good vertex with the lepton tracks, it needs to be of the same sign as the muon as well to be reconstructed as candidate. Either leads to positron misID. The higher contribution of pion misID over kaon misID is due to pion tracks being the most common sort of track in the detector.

When comparing the single misID estimate with the measured data, one finds that it overshoots slightly in the corrected mass region below the B^+ mass. There is also the need for a combinatorial contribution to match the high mass tail, which also extends into the lower mass region and thus push up the total estimate even further. This overestimation is due to the double counting of the double misID contribution. As explained in section 7.2.1, the double misID contribution is counted twice when summing up the single misID

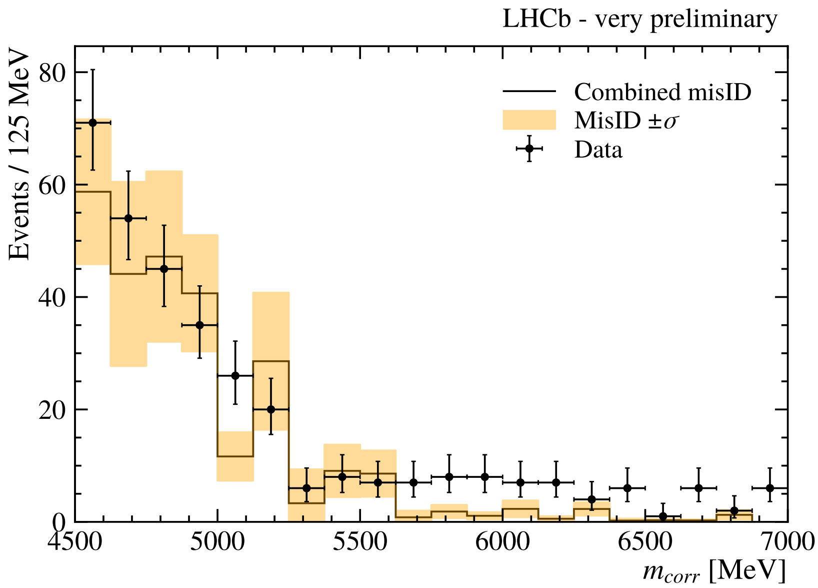


Figure 7.10: Single misID with double misID subtracted on $m_{ee} \in [900, 1200]$ MeV sideband.

estimates. It is necessary to estimate and subtract it.

The double misID estimate is depicted in Fig. 7.9. It is mostly made up of e^+e^- double misID which is consistent with electrons already being more susceptible to misID in the single misID case.

This contribution is subtracted form the summed single misID estimates, with the subtraction bounded from below, meaning that we forbid a negative estimate of the number of misID candidates. The resulting misID curve, which can be seen in Fig. 7.10, slightly undershoots the measured data across the entire corrected mass range, as desired. As this method subtracts the entries of the double misID histograms of those from the single misID histogram, it induces larger relative errors, as the entries themselves are subtracted, while the errors sum up.

Additionally, a clear splitting up of the misID contribution is no longer possible as the subtracted double misID contributions cannot be matched to corresponding single misID contributions. Both of these side-effects are inherent to this method and cannot be avoided. We now use the overshoot in the higher mass range to estimate the combinatorial contribution. This is done by employing a template fit, as described in section 4.3.4, consisting of the final misID histogram shown in Fig. 7.10 and a falling exponential function with floating yield and decay constant for the combinatorial contribution. The resulting fits can be seen in Fig. 7.11. We find that they match the data well without significant deviation. Of note is that this holds true even if we only consider the statistical uncertainty of the data itself and not that stemming from the large weights of the misID

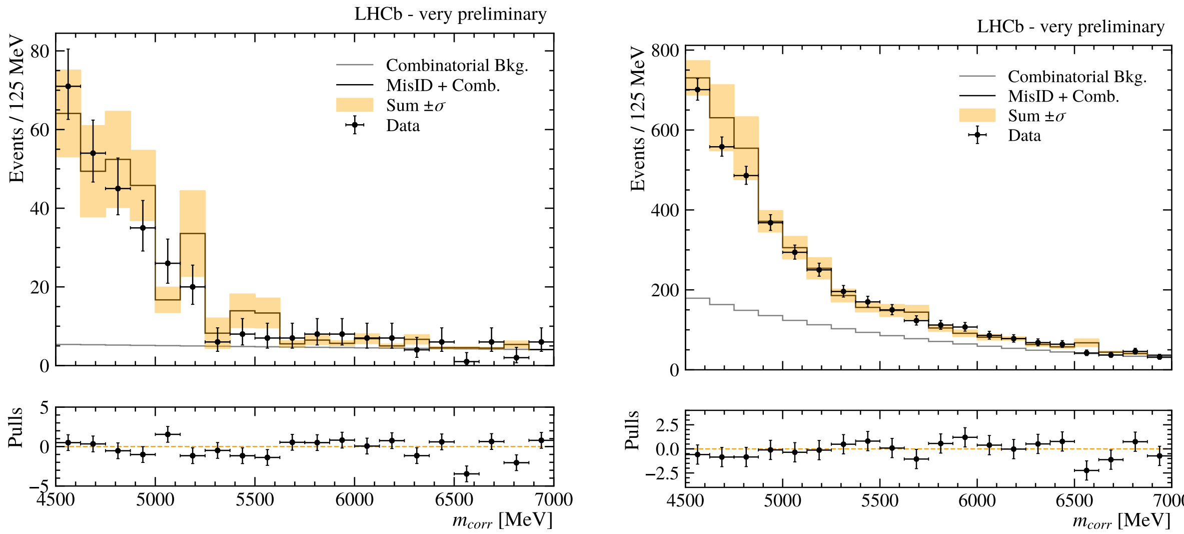


Figure 7.11: Template fit to the corrected mass distribution of the $m_{ee} \in [900, 1200]$ MeV sideband. The fit is made up of an exponential function representing the combinatorial distribution and a fixed misID contribution. On the left the preliminary BDT requirement is in place, while on the right we use none. Otherwise the same method and selection is used.

template.

From the fit we also get a yield of the combinatorial contribution of 94 ± 20 decays. Our misID estimate for the same region is 263 ± 26. While it is not necessarily the case that this ratio holds in the signal mass window, it certainly showcases that misID background is indeed a major contribution.

As all three single misID and all three subtracted double misID contributions also contain a triple misID contribution, it would be necessary to add the triple misID contribution once more to the total summed misID. To estimate the triple misID contribution, only the misIDμ data sample can be used, but we run out of statistics, rendering the strength of the estimation doubtful. To get an upper limit, we can use the observation that the triple misID contribution is at most as large as the smallest double misID contribution. With only e^+e^- misID contributing on a level above $O(1\%)$ to the total misID estimate, it is safe to treat triple misID as negligible.

7.2.5 Signal window estimate

With the merit of the method verified by its estimate matching data on both the high dielectron mass sideband and at the $B^+ \to K^+ J/\psi(e^+e^-)$ peak, we now turn to the signal mass region of $m_{corr} \in [4500, 7000]$ MeV and $m_{ee} \in [600, 900]$ MeV.

Fig. 7.12 shows the predicted misID curves, with the preliminary BDT requirement in place on the left and without any BDT requirement on the right. We find that both

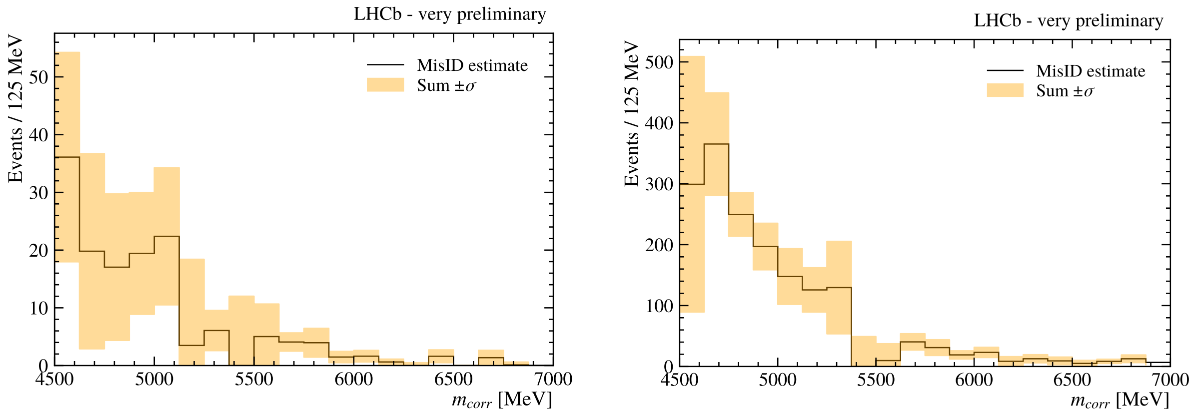


Figure 7.12: Single misID with double MisID subtracted in the signal mass window. The left hand plot uses the preliminary requirement while the right hand plot has no BDT threshold imposed.

curves share a gap, an empty bin corresponding to $m_{corr} \in [5375, 5500]$ MeV. This gap is just high enough in mass that it should not be directly at the peak of the signal channel, but in the rapidly falling off high-mass tail. Still, this could impact the fitted signal yield in the end, if the fitter lowers the combinatorial yield to match the gap and fits a too high signal yield using its long tail to compensate the underestimated combinatorial contribution in the lower mass range.

Further, we note that the uncertainties are large, especially after applying the BDT selection. To understand the structure of the total misID curve, we consider its two contributions. Fig. 7.13 depicts the corrected mass distributions of both single and double misID, split up into different components. The preliminary BDT requirement is used for both plots. We find that the double misID is the reason for both undesired traits of the combined estimate. A peak in its structure at which double misID surpasses single misID causes the gap above the B^+ mass and the high double misID in general is responsible for the large uncertainties. This also suggests that the gap is likely not what one would truly expect from physics, because the summed estimate should not be below the double misID estimate if both are correct.

Inspection of the decay candidates which cause the peak does not yield large irregularities. We find several candidates with larger weights, meaning that the peak cannot be attributed to a single outlier. When considering the double misID in general, we find that it is mostly made up of double pion misID. In addition we find that pion to positron and pion to electron single misID are similar in size, contrasting the dielectron mass sideband where positron misID is far more likely. This points to an abundant physical source of double pion misID. A likely candidate is the decay $B^+ \to \mu^+\nu_\mu\rho^0(\pi^+\pi^-)$, which one would expect to find in the signal mass window. With a measured branching ratio of

$(1.58 \pm 0.11) \times 10^{-4}$[6], it is far more abundant than our signal channel. The ρ^0 decays promptly, which does not impact vertex quality enough to fail our selection.

While not optimal, our misID estimate does fulfil its purpose and can be used both for toy construction and for mass fits. It works very well if the number of candidates is high. Once the selection requirements are tightened and candidates become scarce, relative uncertainties grow large. We find that this is a major bottleneck working with this method. At the same time it is also a honest reflection of lack of knowledge on our part, indicating the limits of the chosen methodology.

7.3 The isolation variable

A method of cutting down on background is demanding that the particle tracks are isolated. The idea behind this is simple: for a signal candidate we expect exactly three charged tracks that are well matched as a vertex and nothing more. If an additional track matches to the same vertex, this suggests that the candidate picked up is not the signal channel. A typical example of this would be background due to a partially reconstructed decay such as $B^+ \to \ell^+ \nu_\ell \bar{D}^0 (\to K^+ \pi^- \pi^- \pi^+)$. These often involve a decay of the B^+ into a D^0 or D^* meson that subsequently decays into several particles including pions or kaons. Due to the size of the involved CKM matrix elements, $|V_{cb}|$ and $|V_{cs}|$, they have large branching fractions.

To discriminate against badly isolated particle tracks, we use a pre-made BDT. It is implemented during the data processing with DaVinci, when information about charged tracks not used to construct a decay candidate (underlying tracks) is available and assigns an isolation score to each signal candidate track. We then take the minimum of these scores. This BDT has been trained to distinguish $B_s^0 \to K^- \mu^+ \nu_\mu$ from a variety

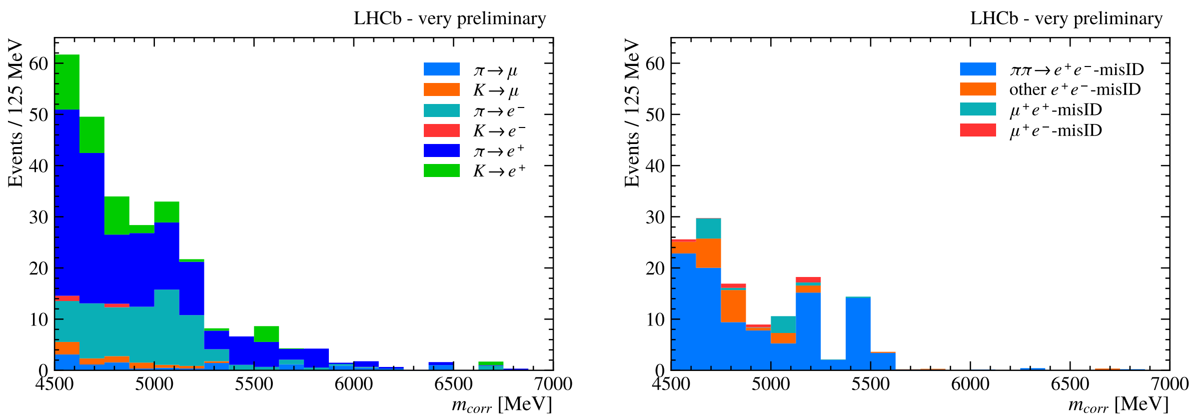


Figure 7.13: Stacked up mass spectra of different misID types within the signal mass window.

of $B_s^0 \to J/\psi X$ backgrounds. As it works on a track per track basis, it can readily be applied on our decay. It uses various kinematic variables describing the relation of each decay candidate track and each underlying track, such as the angle between candidate track and underlying track or the distance between the B^+ vertex and point of closest approach of two tracks. An introduction to the BDT is given in [55]. The variable is not used as a further input for our combinatorial BDT, because it is already the output of a BDT trained on kinematic variables whose workings are outside of our control. With the same sign sample limited in size, separation power is likely higher if this variable is used independently.

7.4 Light meson background

A potentially relevant type of background results from a B^+ decay into a muon, neutrino and light neutral meson, such as π^0, η or η'. The neutral meson can then decay into an electron pair and a photon (Dalitz decay). The final state of this decay chain $B^+ \to \mu^+\nu_\mu h^0(ee\gamma)$ differs from that of the signal only by a single photon, which can be either missed or wrongly picked up as a Bremsstrahlungs-photon. The same applies if the hadron decays into two photons and at least one converts into an electron pair.

To estimate the number of decays in the signal mass window, we consider each background individually and use equation 6.1. It gives the expected number of decays, N_{exp}, as a product of integrated luminosity, $\mathcal{L}_{int}$, the production cross section of $b\bar{b}$-pairs, $\sigma_{b\bar{b}}$, the quark fragmentation fraction, f_q, the respective branching ratio of the decay, $\mathcal{B}(B^+ \to \mu^+\nu_\mu h^0(ee\gamma))$, the generator level efficiency, ϵ_{gen}, representing the likelihood of the decay to happen within LHCb acceptance and the efficiency of the used selection $\epsilon_{selection}$. Further, we assume that the photon conversion rate is 5%, represented by a conversion factor $k_c = 1$ for the case of decays of the form $h^0 \to e^+e^-\gamma$ and $k_c = 1 - 0.95^2$ for decays of the form $h^0 \to \gamma\gamma(ee)$.

$$\begin{aligned} N_{exp} &= \mathcal{L}_{int} \times \sigma_{b\bar{b}} \times 2f_q \times \mathcal{B}(B^+ \to \mu^+\nu_\mu h^0(ee\gamma)) \times \epsilon_{gen} \times \epsilon_{selection} \times k_c \\ &= \mathcal{L}_{int} \times \sigma_{b\bar{b}} \times 2f_q \times \mathcal{B}(B^+ \to \mu^+\nu_\mu h^0(ee\gamma)) \times \epsilon_{gen} \times \frac{N_{MC,selection}}{N_{MC,generator}} \times k_c \end{aligned}$$

The value for most of these parameters, listed in Tab. 7.3, are taken from literature. For $\epsilon_{selection}$, we count the number of candidates in a Monte Carlo sample of the considered decay with the desired selection applied $N_{MC,selection}$ and divide it by the number of candidates generated for this sample $N_{MC,generator}$.

Following the expectation that the dielectron-mass distributions of these decays are to

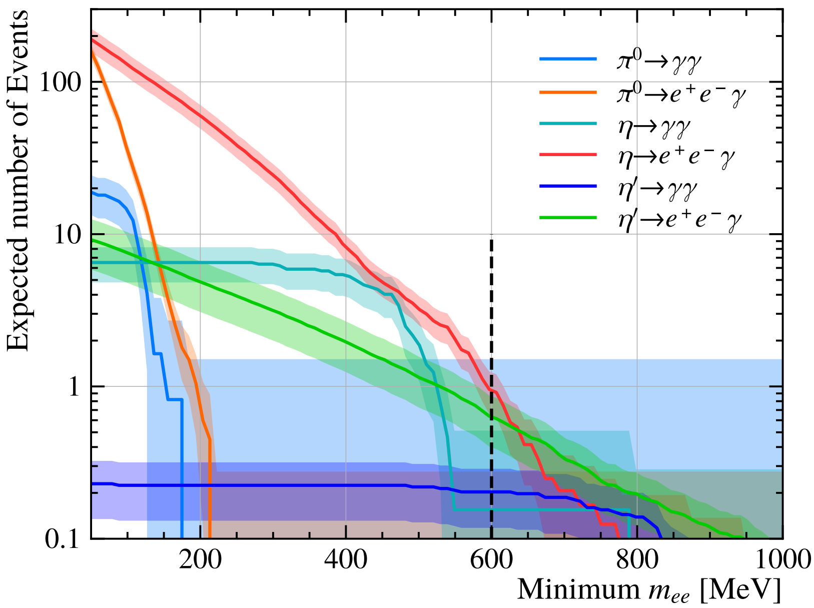


Figure 7.14: Expected number of different light meson backgrounds as a function of the m_{ee} limit. The black dotted line denotes the mass threshold of the signal window.

peak at $m_{ee} = 0$ and then fall off towards higher masses, we consider N_{exp} as a function of the lower m_{ee} threshold. The resulting curves are depicted in Fig. 7.14. An overview over the different expected contributions of the decays can be found in Tab. 7.4

It is apparent that all considered decays drop off quickly enough to be negligible in the signal window between 600 MeV and 900 MeV, with expected numbers of remnant pollution decays being below $O(1)$, before any BDT requirements. When considering the

Table 7.3: General parameters necessary for the light meson background estimates.

Integrated luminosity $\mathcal{L}_{int}$	$b\bar{b}$ cross section $\sigma_{b\bar{b}}$	Fragmentation fraction f_u
(5.1 ± 0.10) fb^{-1} [45]	$(560 \pm 50) \times 10^9$ fb [56]	0.346 ± 0.008 [57]

Table 7.4: Masses of the light meson, branching ratios and expected number of remaining decays in the signal mass window for the light meson decay considered potentially relevant. We take into account both Dalitz decay and decay into two real photons with electron pair creation.

Light Meson h^0	π^0	η	η'
Mass [MeV]	134.9768 ± 0.0005	547.862 ± 0.017	977.78 ± 0.06
$\mathcal{B}(B^+ \to \mu^+\nu_\mu h^0)$	$(7.80 \pm 0.27) \times 10^{-5}$	$(3.5 \pm 0.4) \times 10^{-5}$	$(2.4 \pm 0.7) \times 10^{-5}$
$\mathcal{B}(h^0 \to \gamma\gamma)$	$(98.82 \pm 0.03)\%$	$(39.36 \pm 0.18)\%$	$(2.31 \pm 0.03)\%$
$\mathcal{B}(h^0 \to e^+e^-\gamma)$	$(1.17 \pm 0.04)\%$	$(6.9 \pm 0.4) \times 10^{-3}$	$(4.91 \pm 0.27) \times 10^{-4}$
$N_{exp}(h^0 \to \gamma\gamma)$	0.00 ± 1.51	0.20 ± 0.36	0.20 ± 0.09
$N_{exp}(h^0 \to e^+e^-\gamma)$	0.00 ± 0.28	0.91 ± 0.29	0.64 ± 0.23

uncertainties of the estimates, it is important to keep in mind that they are bounded from below by the assumption that the observed number of candidates is of Poissonian distribution. As the number of expected decays in the signal region is low, or even zero, this means that the uncertainty is only an indicator of the generated quantity of candidates for the used simulation samples, scaled by the branching ratio of the considered decay. It is thus a very conservative uncertainty, which does not reflect any physical lack of knowledge.

For the decays of η' a simulation sample is not available, but the decays are expected to behave nearly similar to those of η. To still get an estimate of $\epsilon_{selection}(\eta')$ we scale the dielectron mass of the η Monte Carlo sample by the ratio $m_{\eta'}/m_{\eta}$.

7.5 Background from B_c^+ decays

Made up of an $\bar{b}$ and a c quark, the B_c^+ is a close relative of the B^+ meson. It is expected to feature an analogue decay channel $B_c^+ \to \mu^+\nu_\mu e^+e^-$ with the same final state particles as our decay. In addition, its c quark opens up the channel $B_c^+ \to \mu^+\nu_\mu J/\psi(e^+e^-)$, which also leads to the same final state. Due to sharing the signal decay final state and having a similar topology, both decays are not covered by any of the background suppression and modelling measures in use. Consequently a significant contribution of either decay in the signal window would necessitate a separate treatment.

The $B_c^+ \to \mu^+\nu_\mu J/\psi(e^+e^-)$ decay

In principle, the decay channel $B_c^+ \to \mu^+\nu_\mu J/\psi(e^+e^-)$ is located around the dielectron mass of $m_{J/\psi} = 3096.9\,\text{MeV}$, far away from the signal dielectron mass window. However, the peak in dielectron mass is quite broad, while the yield is very large. It is not a priori clear that the remaining contribution of this channel in the signal window is negligible, as the $B^+ \to \mu^+\nu_\mu e^+e^-$ decay is expected to have a low yield.

To estimate the remaining contribution of the background $B_c^+ \to \mu^+\nu_\mu J/\psi(e^+e^-)$ in the signal mass window, $m_{corr} \in [4500, 7000]\,\text{MeV}$ and $m_{ee} \in [600, 900]\,\text{MeV}$, we first extract its yield at around the J/ψ mass and later scale it down to the signal window based on the behaviour of Monte Carlo samples of the decay.

For this we use the $B^+ \to \mu^+\nu_\mu e^+e^-$ data sample with full selection, except the signal mass window. A dielectron mass window around the J/ψ mass, $m_{ee} \in [2700, 3300]\,\text{MeV}$ is considered. The feature of interest for this search is the corrected mass, but with the electron momentum scaled to the J/ψ mass, as done in the $B^+ \to K^+J/\psi(ee)$ cross check. To cut away B^+ background, we also impose a requirement on the three body mass, $m_{\mu J/\psi(ee)} > 4500\,\text{MeV}$.

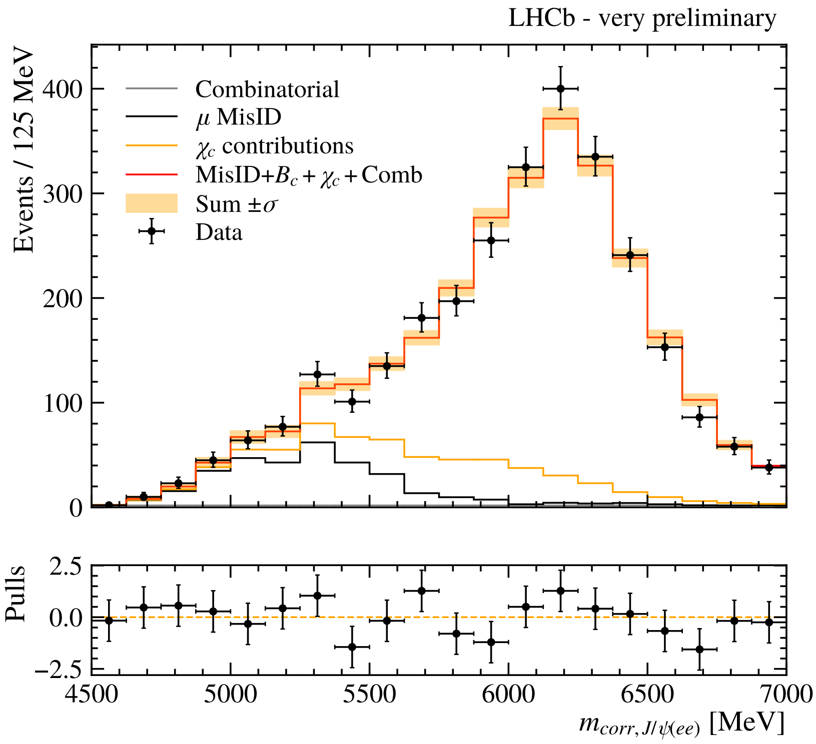


Figure 7.15: Template fit to the $B_c^+ \to \mu^+\nu_\mu J/\psi(ee)$ peak. Visible are a fixed μ misID contribution, MC samples for both the decay itself and a $B_c^+ \to \mu^+\nu_\mu\chi_c(\gamma J/\psi(ee))$ cocktail, as well as an exponential model for the combinatorial contribution. The contributions are stacked upon another.

We then perform a template fit on the data. For the decay $B_c^+ \to \mu^+\nu_\mu J/\psi(e^+e^-)$ a template is derived from a corresponding LHCb simulation sample and used with floating yield. The same we do for the decays $B_c^+ \to \mu^+\nu_\mu\chi_{c0}(\gamma J/\psi(ee))$, $B_c^+ \to \mu^+\nu_\mu\chi_{c1}(\gamma J/\psi(ee))$ and $B_c^+ \to \mu^+\nu_\mu\chi_{c2}(\gamma J/\psi(ee))$, which are expected to be found in this mass region. They are described by one shared simulation sample with their yields relative to each other fixed. To accommodate combinatorial background an exponential function with floating yield and shape is used. Lastly, a fixed muon misID template is added. The composite sum of this is then fitted.

The result of this fit is depicted in Fig. 7.15. We find that the fit matches the contributions to the data without significant deviation and positively compares to the results of [57].

For the misID background we consider only the muon single misID contribution, because the high number of electrons from the signal compromises the electron misID estimate.

From the fit the yield is extracted. It is now scaled with the ratio of the number of candidates in the simulation sample found in the signal mass window to the number found in the mass window used for the $B_c^+ \to \mu^+\nu_\mu J/\psi(e^+e^-)$ fit. The result approximates the expected contribution of $B_c^+ \to \mu^+\nu_\mu J/\psi(e^+e^-)$ as a background in the signal mass window. We find that it is 1.5 decays before any BDT requirement or 0.6 decays after the arbitrary selection. This is low enough to allow for safely neglecting this decay channel.

The $B_c^+ \to \mu^+\nu_\mu e^+e^-$ decay

A last background is $B_c^+ \to \mu^+\nu_\mu e^+e^-$, which shares both final state and topology with the target decay. Currently there is no theoretical prediction for the branching ratio of this decay. Additionally, there are nearly no experimental results on the branching ratios of the B_c in general.

While at LHCb B_u^+ are produced two orders of magnitudes more often than B_c^+ [57], the matrix element $|V_{cb}|$ is also larger by about a magnitude than $|V_{ub}|$ which is quadratically suppressing our signal channel. Therefore, one would expect naïvely that $B_c^+ \to \mu^+\nu_\mu e^+e^-$ happens at a comparable rate to our signal decay. However, because the B_c^+ has a shorter lifetime, selection efficiency is lower for this decay, than for the signal. In addition the decay is not expected to have a significant ρ^0/ω contribution, due to the lack of a u quark [58]. Hence, we do not expect it to contribute meaningfully in the signal mass window and neglect it.

Chapter 8

Toy studies

Having gained an understanding of both signal shape and background makeup, we can construct toy models, which reproduce our understanding of what is to be expected in the blinded signal region. The construction of such models in a way that they reflect expected fluctuations is done using bootstrapping and described in section 8.1. In section 8.2 these models are then used to optimize the selection threshold on both the combinatorial BDT the isolation variable and to verify the stability of the used methodology. This is followed by an estimate of the impact of a larger misID control sample size in section 8.4.

8.1 Toy construction

As explained in the last chapter, the two principal background contributions expected in the signal data are misID and combinatorial background. For misID background a proxy is constructed that mirrors the expected background in both yield and shape. It is sensitive to changes in BDT selection. For combinatorial background, the assumption of a falling exponential function is chosen and positively tested on the high dielectron mass sideband and the same sign sample. To construct useful toys, it is additionally needed to have an estimate of both the decay constant and most importantly the yield of the exponential, dependent on the BDT threshold.

This proves problematic. The most intuitive solution would be to use the ToaK sample, as it is the closest thing available to a pure combinatorial sample. Its shape could be used directly and its yield should be proportional to the real background, needing only a constant scale factor, which could be obtained by an initial blinded-fit without BDT selection applied. However, the ToaK sample is simply too small and too quickly decimated by the BDT to be used. For the same sign sample, it is not clear whether slope and yield can be directly transferred and the low sample size remains an issue.

Another option would be the usage of a blinded fit to the signal data, meaning a fit for which neither the fitted signal yield nor the data are examined. The fitted background contributions on the other hand can be extracted. However, this runs the risk of the fitted signal contribution skewing the results. By bad luck it is possible that statistical fluctuations are best described by fitting an oversized signal contribution. This can be somewhat counteracted by examining the fitted background curves themselves, but this is not feasible to do when many fits are necessary, as is the case for the working point optimization later on.

Instead the dielectron mass sideband is used. It is already considered in the last chapter and found to be well described by summing the misID estimate and an exponential combinatorial contribution. However, whether its combinatorial contribution mirrors that in the signal region is not clear a priori. To test this, it is compared with the output of a blinded fit to the signal region. Initially, this is done with no BDT requirement applied on either sample.

The fit includes a fixed misID contribution, a free exponential function and a signal contribution with floating yield, whose shape is given by the simulated signal sample. Depicted in Fig. 8.1 are the fixed misID and fitted combinatorial contribution of the blinded fit. The fit happens only in the blinded mass region, the exponential function is then extrapolated over a wider range. To make sure that the blinded fit is working as intended, it can be compared with the observed data on the corrected mass region below 4500 MeV and above 7000 MeV. The exponential assumption is likely to break down eventually, but close to the blinded region, data and extrapolated fit should still agree. Indeed, we find that they do.

The yield and decay constant of the exponential function are compared with those obtained from the fit to the high dielectron mass sideband. The yields agree, but the decay constants deviates, with the exponential fitted to the sideband being less steep. This can be accommodated by the inclusion of a constant scaling factor for the decay constant.

Hence, we choose the combinatorial contribution of the high dielectron mass sideband as proxy for the signal mass window combinatorial background contribution. The pseudo-decays making up the combinatorial contribution of the toy samples are drawn from an exponential distribution using the fitted decay constant times multiplied with the scale factor. Their number is randomly decided, based on a Poisson distribution with the fitted yield as expectation value.

To construct the misID contribution of the toy models, we use bootstrapping [59]. From each misID sample, candidates are randomly drawn and assembled in a new sample of the same size. A candidate can be drawn multiple times. The new samples are used to construct a new misID estimate distribution. From this distribution pseudo-decays are

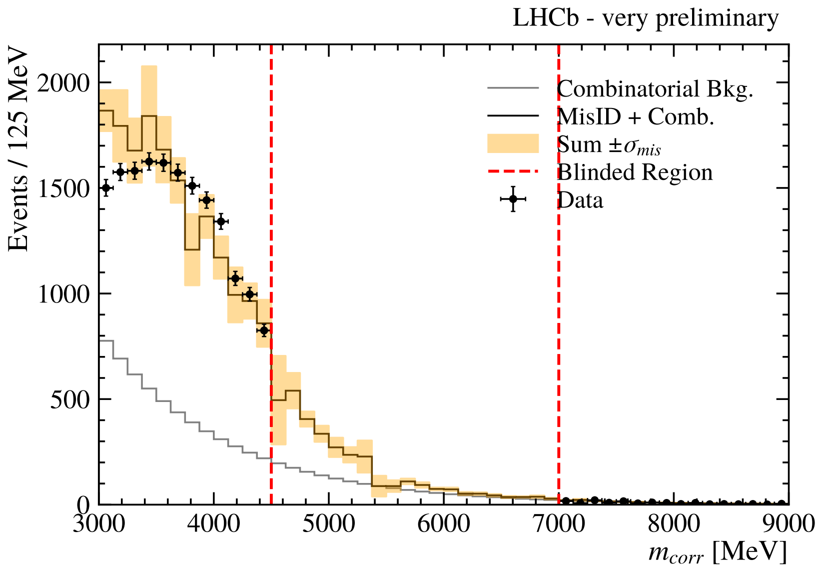


Figure 8.1: MisID contribution and fitted combinatorial of a fit to data in the signal mass window, extrapolated to cover a wider mass range. They are compared with data outside the signal mass window. The red dashed lines denote the blinded region.

randomly drawn as above. These pseudo-decays are then combined with those obtained for the combinatorial distribution to form the toy sample.

An example toy can be seen in Fig. 8.2. As explained, the toy data is based on the combinatorial fit to the dielectron mass sideband and the misID estimate for the signal window. It can be compared with the fitted background from the blinded fit and found to not deviate significantly. This suggests that, in the case of no BDT requirement, the constructed toys and chosen combinatorial proxy work as expected. However, it is possible that imposing requirements on the BDT selection impacts sideband and signal region in a different manner.

8.2 Choosing a BDT working point

With the toy models acting as stand-in for the expected data in the signal window, it is now possible to optimize the BDT selection for the final mass fit. The aim is to choose the combined requirements on combinatorial BDT score and minimum isolation score, called the working point, to maximize the sensitivity of the fitting procedure to the signal decay.

For this, a two dimensional grid of working points is considered. The range of possible threshold scores for either BDT is corresponding to a linear decrease of signal efficiency without a requirement placed on the other BDT. For the combinatorial BDT 10% increments are used, while for the isolation score the range includes an offset of 5%, i.e. the

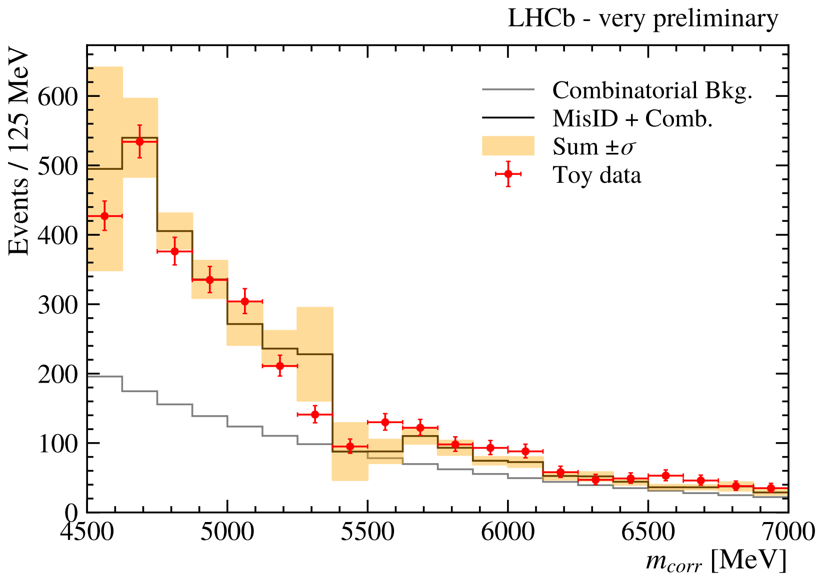


Figure 8.2: Toy data compared with the background result of the blinded fit.

possible efficiency loss values are 0%, 5%, 15%, 25% etc.. To calculate the efficiency loss the LHCb signal simulation sample is used.

At each working point on the grid, 250 toy samples are constructed. On each toy a template fit is performed. As in the final fit, the fit contains the fixed misID estimate, an unconstrained exponential for the combinatorial and a simulation based signal histogram, whose yield is free to float. Note that while each toy is different, the fitting procedure is always the same, including the same fixed misID shape. The combinatorial proxy is obtained from a fit to the high mass sideband that happens once per working point.

The figure of merit of a given working point is an approximation of the minimal amount of signal yield it would rule out at 95% confidence level. To construct this figure of merit for a given working point, we consider the upper limit of the 2σ confidence interval given by the loss profile of the MINOS algorithm and subtract the fitted yield. As is shown later in 8.3, the yield fitted to the toys does not show any bias, adhering to a Gaussian distribution centred on zero. Consequently, averaging this number is expected to provide a similar value as calculating the upper limit with respect to zero signal yield, which is computationally more expensive. Further, this is divided by the BDT requirement efficiency to give the sensitivity to signal events before the BDT selection. This allows direct comparison of different working points and with the expected number of decays calculated in section 6.2. The figure of merit is averaged over all successful fits for a given working point.

Another factor to consider is fit stability. Once sample size decreases sufficiently, the fitting algorithm is no longer able to fit all constructed toys. The reason behind this is that the underlying MINUIT algorithm needs a cost function that has a continuous derivative around its minimum. This is not given any more, once the relative uncertainties grow too large, as variation in the fitting parameters do not sufficiently change the cost function any more.

Unfortunately, `iminuit` does not reliably declare every failed fit as such. While it is possible to demand the fit to fulfil further expectations on successful fits, such as errors of sensible size or parameters that are not at their range limits, it is not possible to completely isolate against failures in such a way. This means, that once one uses a BDT selection where fits are like to fail, it becomes necessary to consider each fit individually to know for certain whether a fit is valid or not.

As this is not feasible, we are conservative in the choice of the working point. Besides minimizing the figure of merit, we further impose the condition that the failure rate is below 1%. This may seem overcautious, especially when regarding the sometimes drastically lower figure of merit values for working points with higher failure rates. However, it needs to be kept in mind that for a failed fit the confidence interval provided by `iminuit` loses its meaning and is often near zero. Consequently, the averaged figure of merit cannot be trusted any more.

Additionally, it is possible that the toy samples with failed fits are similar in some way. This would mean that by choosing a working point with higher failure rate, a certain class of toy sample would be removed from the process. This could bias the fitting procedure and skew the result of the average upper limit estimation.

While the usage of the dielectron mass sideband as stand-in for the combinatorial background is justified for the case of no BDT selection, it cannot be assumed that it is effected by the imposition of BDT requirements in the same way as the combinatorial background in the signal window. To still be able to construct toys, we proceed iteratively. This means we start out using this proxy. Having found an optimal working point under this assumption, the sideband with this BDT configuration is compared to a blinded fit and if discrepancies arise, it is adjusted accordingly and the process repeated.

Fig. 8.3 depicts percentage of failures and signal sensitivity for different working points. A curve corresponds to a fixed isolation BDT threshold, with varying combinatorial BDT requirement. Toy models with failed fits are not considered for the curves in the right-hand plot. As one can see in the left hand plot, the failure rate increases quickly when tightening the BDT requirements. The chosen working point, the stable configuration with the lowest figure of merit, is circled. For the purpose of the optimization, the standard deviation of the mean is considered the relevant uncertainty, as it captures the variation of the average figure of merit when constructing new toys. To later express lack

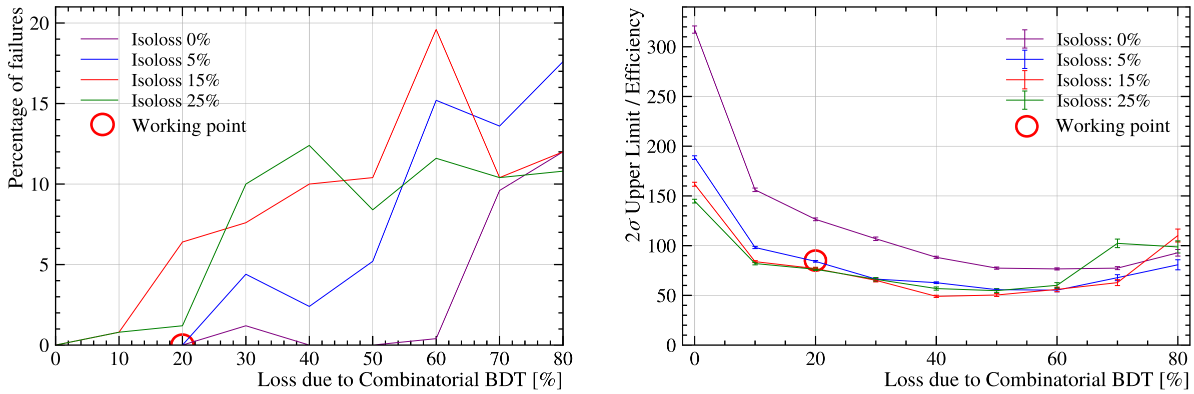


Figure 8.3: Failure percentage (left) and signal sensitivity (right) for different BDT working points. Combinatorial proxy used is the dielectron mass sideband combinatorial contribution. The uncertainty on the signal sensitivity is the uncertainty of the mean.

of knowledge about the true upper limit, the standard deviation is used.

We also note that when tightening the requirement on either BDT, the first increment brings the largest improvement in figure of merit. This holds for both combinatorial BDT, with the steepness of the curve decreasing and for the isolation BDT where the difference between the curve without any selection and that with the minimal 5% efficiency loss is the largest between neighbouring curves. Keeping in mind that fitting failures can contaminate outputs in the lower efficiency region, it is prudent to not interpret more into the plot.

Besides the working points shown in the plots, further curves are considered with the efficiency loss due to isolation at higher percentages. They are omitted for clarity and can be found in A.4.1.

With BDT requirements now in place it is necessary to revisit the assumption of the dielectron mass sideband being a correct proxy, for the combinatorial background. For that a blinded fit is done again in the signal region, now at the working point and its output compared with that of a fit to the sideband. One finds that the blinded fit gives a combinatorial yield that is 2.5 times higher than that of the sideband. Similarly the decay constant is far lower. To compensate, the sideband yield is scaled by a factor of 2.5 when using it as proxy to construct toy samples. The exponential fitted to the sideband is nearly flat rendering its decay constant useless for the task at hand. Instead the decay constant is fixed at the value given by the blinded fit. This causes the toys to be certainly wrong at the working point without BDT selection, but as it is already apparent that the optimization does not end there, this is not an issue.

With the toy sample construction adapted, the process goes through the second iteration. As before, stability and sensitivity are examined. Fig. 8.4 depicts failure percentage and figure of merit in this iteration. There are now more stable working points accessi-

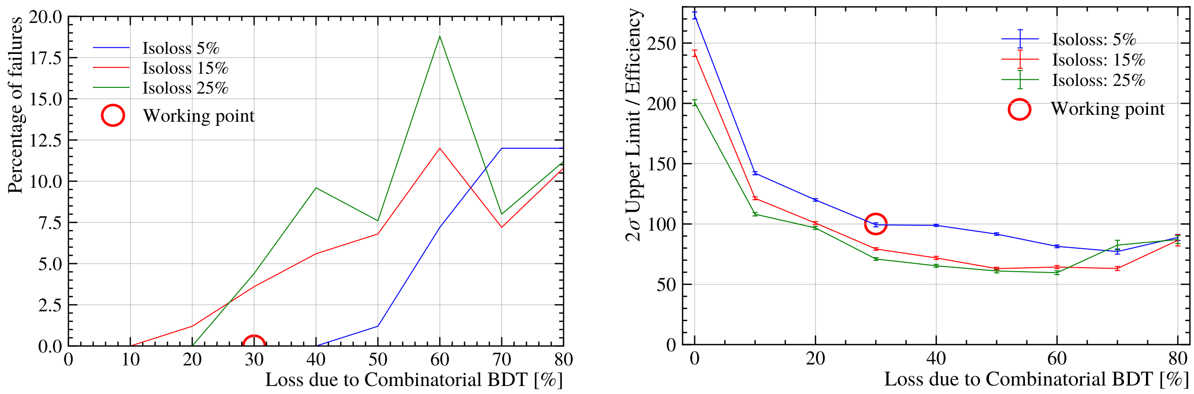


Figure 8.4: Failure percentage (left) and signal sensitivity (right) for different BDT working points. Combinatorial proxy used is the dielectron mass sideband combinatorial contribution, adapted to the results of the first iteration. The uncertainty on the signal sensitivity is the uncertainty of the mean, over all toys.

ble. At the same time, the figure of merit has increased across the board. Both of these effects are explained by the increase of the total size of background, which lowers relative uncertainties, while increasing their absolute values.

It appears as if the figure of merit curve runs into a plateau, with three stable working points of roughly the same minimal sensitivity. Of these, the one with the highest signal efficiency is chosen, as it is not on the edge of the stable configuration range and offers higher statistics.

As a side note, instead of taking the upper limit as result from a fit, one could also use the Punzi figure of merit[60], which is constructed for such a purpose. A plot depicting the Punzi figure of merit for this iteration can be found in A.4.2. It shows a similar plateau structure and suggests the same working point.

Again, yield and decay constant of the proxy are compared with a blinded fit. At this working point, they are found to be within uncertainty from another. Thus, the iterative process is stopped and the working point chosen.

Fig. 8.5 depicts the background contributions of the blinded fit to data at the BDT working point, in the signal region. They are extrapolated to cover the entire $m_{corr} \in [3000, 9000]$ MeV mass range. With the implemented BDT selection, the discrepancy between extrapolated estimate and data in the lower mass region is far more pronounced than without any selection, but at the edge of the blinded region they still roughly match, suggesting that the fit does work as intended.

At the working point, the requirement on the combinatorial BDT has an efficiency of 70% and that on the isolation BDT an efficiency of 95%, providing a combined BDT selection efficiency of $(67 \pm 3)\%$. The upper limit of the 95% confidence interval is at 66 ± 17, which would translate to the methodology being sensitive to any signal with

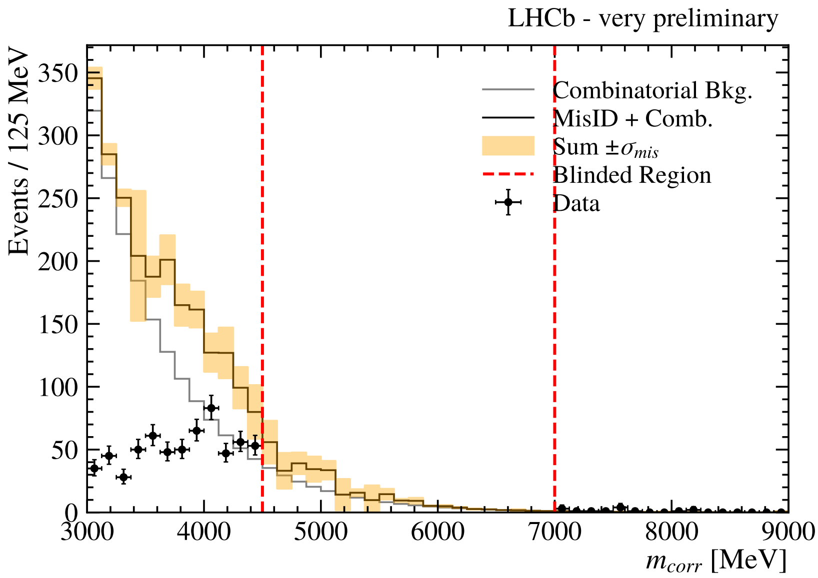
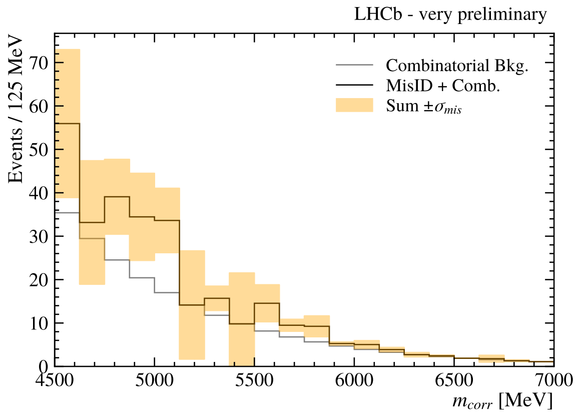


Figure 8.5: MisID estimate and fitted combinatorial of a blinded fit to data in the signal region, extrapolated over a wider mass range. Data is plotted outside the signal corrected mass window. The plotted 1σ bands correspond to the misID uncertainties.

more than 99 ± 26 decays before the BDT selection is imposed. In the signal region, 269 ± 56 background decays are fitted, of which 189 ± 46 are combinatorial background and 80 ± 32 are misID background. As a comparison, the expected number of background decays without any BDT requirement is 3320 ± 277, of which 1700 ± 259 are misID and 1620 ± 100 are combinatorial.

While this results in the remaining background being two parts combinatorial and one part misID, it needs to be kept in mind that this comes with large uncertainties. As the combinatorial estimate results from a fit, these uncertainties reflect only the number of candidates in the data sample and the uncertainties of the misID estimate. This means that even though it only contributes a third of the background, misID is still very important for both fit quality and sensitivity to signal.

In the left-hand plot of Fig. 8.6 the background contributions in the blinded region of the fit to data is depicted. The right-hand plot depicts an example fit to a toy sample. Of note are the binwise relative uncertainties of the misID estimate, given as errorbands in the left-hand plot, which are often larger than unity. Comparison with the errors provided by the template fit to the toy sample in the right-hand plot shows little change indicating that the main source of uncertainties in the template fit are indeed the misID estimate uncertainties.



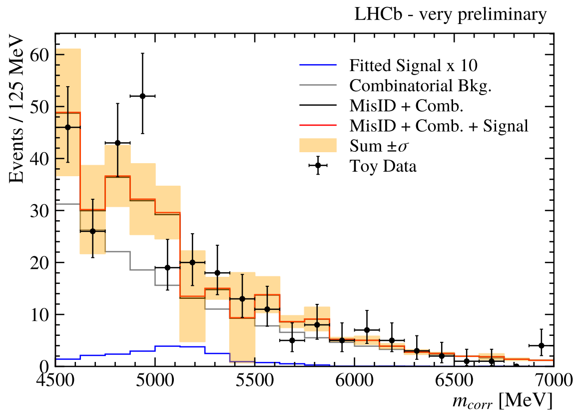


Figure 8.6: Background contributions of blinded fit (left) and template fit to toy sample (right). Both are with the optimized BDT selection. For the toyfit, the depicted 1σ bands are depicting the uncertainty given by the fit to toy data. For the blinded fit we consider only the misID uncertainty.

8.3 Testing toy and fit stability

To gauge whether this method of toy construction and fitting behaves as expected, both in terms of consistency and in terms of statistical behaviour, we construct a large number of toys, using the method explained above. Now there is the additional possibility of a signal contribution, sampled using the simulated signal sample as probability distribution. Again, a fit is performed to each toy. This is done with injected signal yields of exactly 0, 10 and 100 decays. In each case 5000 toys are constructed.

Then the pull distribution, meaning the distribution of the difference between fitted yield and injected yield divided by the uncertainty of the fitted yield, is examined. We again use uncertainties derived from the loss profile by the MINOS algorithm. This time however, we consider the 1σ confidence interval. As the loss profile is in general not symmetric, care needs to be taken whether the fitted yield is higher or lower than the injected value. In the former case the lower uncertainty is used and in the latter the higher.

As our assumption is for the fluctuations to be Gaussian, we would expect that the pull distributions are Gaussian as well. Even more importantly, there should be no bias, corresponding to the Gaussian being centred on zero.

The plotted pull distributions can be seen in Fig. 8.7. They are well described by the fitted Gaussians. There are some minor deviations, the most notable being asymmetric tails that can be seen for the zero signal case and the case of 100 signal decays. Similarly, we find that for 0, 10 and 100 injected signal decays a significantly (3σ) higher yield is fitted in 0.57%, 0.47% and 0.54% of toys, which is slightly higher than the expected 0.27%. It is not possible to say whether this hints to a flaw in the methods used or is

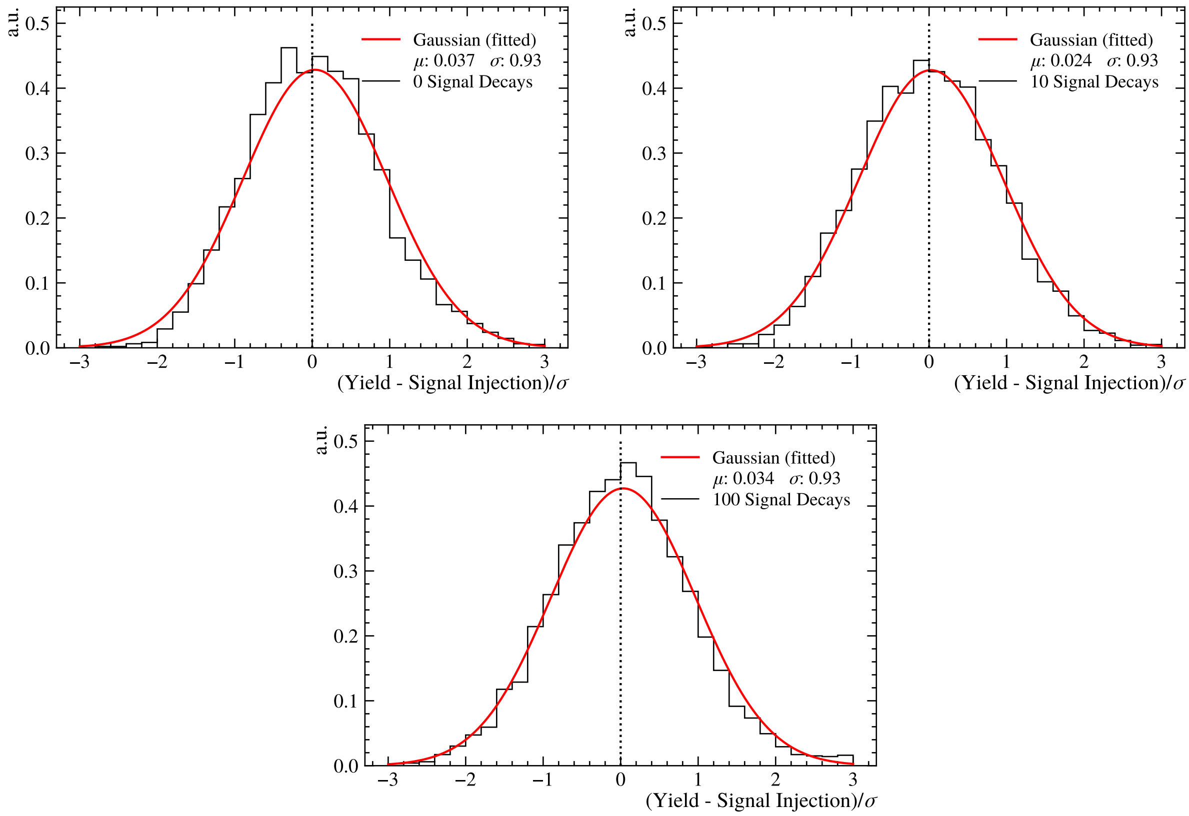


Figure 8.7: Distribution of the difference between fitted signal yield and injected signal yield divided by uncertainty for 5000 toys.

indicative that the Gaussian assumption is not exactly correct for this kind of situation.

However, the discrepancies are minor. For signal injections of 0, 10 and 100 decays, the fitted Gaussians have a mean of 0.037 ± 0.013, 0.024 ± 0.013 and 0.034 ± 0.013 respectively, which is acceptable.

8.4 High statistics limit

It was established in the last section that the uncertainty of the misID estimate is the main source of uncertainty for the template fit. This uncertainty in turn could potentially be decreased by the usage of a larger hadron sample for the misID estimate that captures the expected shape more correctly. While it is hard to predict the impact a gradual increase of statistics in the hadron samples would have on the results of the process, because this cannot be reflected in the bootstrapping, what can be considered is the limit.

We assume that the misID shape results from a sample with infinite statistics. This would accordingly set the uncertainties of the misID estimate to zero. Further, the bootstrapping now samples an infinite number of candidates from the misID sample, which

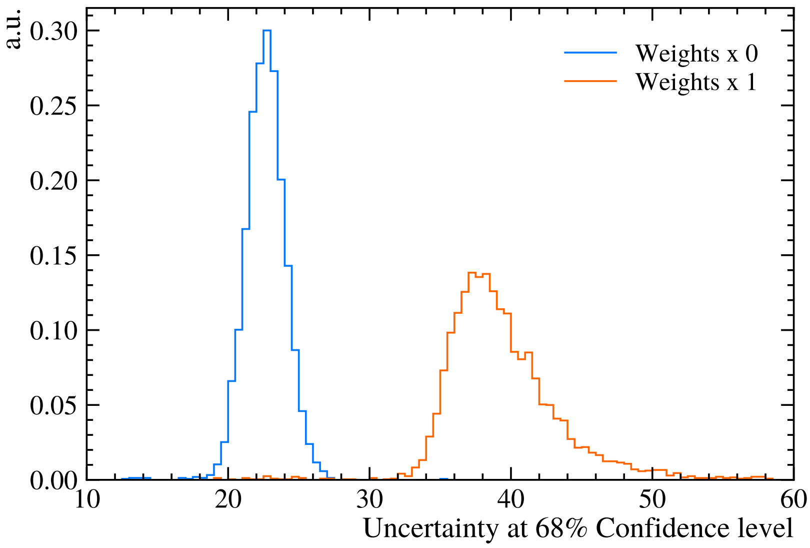


Figure 8.8: Yield uncertainties of fits to 5000 toy samples. The blue distribution assumes the misID sample to be of infinite size, letting the mean uncertainty vanish. The orange curve corresponds to the case at hand described in the chapter before.

in turn exactly captures the true underlying distribution. This can be approximated by simply not bootstrapping at all and instead drawing pseudo-decays directly from the probability distribution corresponding to the summed up misID estimate.

We do this for 5000 toys and compare the distribution of the fitted yield uncertainties with that of the limited statistics method we have described above. This is depicted in Fig. 8.8. As can be seen, in the infinite sample size limit, the average yield uncertainty of the fit decreases by roughly half. This would directly translate into a similar decrease of the minimal number of detectable signal decays.

In addition, in this limit the range of stable working points should increase, as the increasing relative uncertainty of the misID is the limiting factor for fit stability. It is possible that this would lead to a working point with stricter BDT requirements that provides a higher sensitivity to signal.

Chapter 9

Expected upper limit

As explained in chapter 4, the aim of this analysis is to search for a signature of the decay $B^+ \to \mu^+\nu_\mu e^+e^-$. If no signature is found, an upper limit of the branching ratio is established instead. As the signal region is still blinded, we will estimate the the expected upper limit.

To work towards this purpose, we normalise the signal decay to the reference channel $B^+ \to K^+J/\psi(e^+e^-)$ in chapter 6. Then, in chapter 7 the expected background contributions are examined and treated by the implementation of a BDT classifier and estimation of misID shape and yield. With this estimate we can construct toy models of what is expected in the blinded signal region. They can be used to optimize the selection and verify the stability of the used fitting method, which is done in chapter 8.

In this chapter, these methods are combined to provide an expected branching-ratio upper-limit. We follow the methodology employed by the Belle collaboration in their search for $B^+ \to \mu^+\nu_\mu\gamma$ [3]. First, in section 9.1 we explain how the sensitivity to signal yield of a given fit can be determined. Then, in section 9.2 we provide an expectation for the sensitivity of the fit to data by the using toys and convert it into an expected upper limit on the branching ratio.

9.1 Extracting the sensitivity

As explained in 4.3.4, the fitting algorithm numerically minimizes a cost function. As a basis to later construct a cost function, a likelihood function, $\mathcal{L}$, is established first. For a binned fit without weights, this function is based on the bin-wise Poissonian probability, $\mathcal{P}(n_i;\nu_i)$, for the number of observed events, n_i, to occur given an expected value, ν_i. The total likelihood is then given by the product of the per bin probabilities,

$$\mathcal{L} = \prod_{i}^{bins} \mathcal{P}(n_i; \nu_i). \tag{9.1}$$

The expected number of events in a bin, ν_i, is given by the sum of all the individual contributions to this bin, y_{ic},

$$\nu_i = \sum_c y_{ic} = \sum_c f_{ic} y_c, \tag{9.2}$$

which is equivalent to considering the total yields of each contribution, y_c, and summing them up, weighted by the fraction of them to be found in a particular bin f_{ic}. These yields, y_c, are what is varied by the fitting algorithm, to find a minimum of the cost function. For the cost function, it is desired that the parameter values that minimise it also maximise the likelihood. Thus the negative logarithm of the likelihood is used.

In the template fit framework used for this analysis, the likelihood is modified to take into account weights attached to the individual events and the usage of parametric functions, which results in a more complex expression [44].

While the likelihood describes the probability to observe a bin-wise numbers of events, $(n_1, ..., n_{i_{max}}) \equiv \boldsymbol{n}$, given a set of expected bin-wise contributions, we are interested in the probability of one of these contributions having a certain yield given the observed number of events. To obtain this, one can turn the likelihood into a probability density function, $\mathcal{F}$, by using Bayes' theorem:

$$\mathcal{F}(y_c|\boldsymbol{n}) = \frac{\mathcal{L}(\boldsymbol{n}|y_c)\pi(y_c)}{\int_0^\infty \mathcal{L}(\boldsymbol{n}|y_c)\pi(y_c)d\nu_c}. \tag{9.3}$$

Here a flat prior, $\pi(y_c)$, is used, which is one for $y_c > 0$ and zero otherwise. It corresponds to our demand that the signal contribution may not be negative. The function $\mathcal{F}(y_c|\boldsymbol{n})$ is the probability density as a function of the yield of a contribution, y_c, given the observed bin-wise numbers of events, $\boldsymbol{n}$.

In practice, this function is accessed by scanning the likelihood as a function of the yield we are interested in. A number of values for this yield are specified, as a one dimensional grid. Then, at each of this yields, a fit is done with the yield of interest fixed and other parameters floating. From it, $\mathcal{F}$ is derived.

To obtain the minimum yield of a contribution a fit is sensitive to, we impose

$$1 - \mathrm{CL} = \int_0^{y_c} \mathcal{F}(y_c|\boldsymbol{n}) dy_c. \tag{9.4}$$

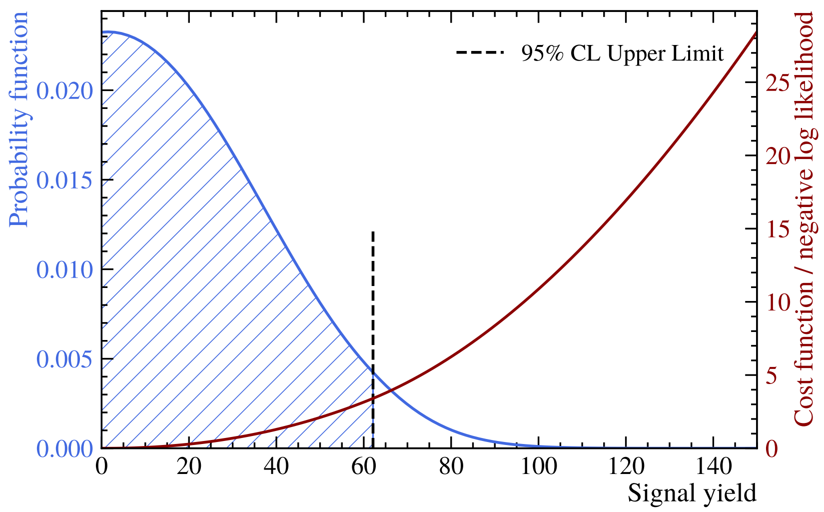


Figure 9.1: A plot of both the likelihood and the negative logarithmic likelihood against the signal yield.

Here, CL is the desired confidence level, in our case it is 95%. The value of y_c which fulfils this equation is the sensitivity of the fit.

This differs from the the sensitivity estimate used for the working point optimisation by being taken respective to zero instead of the fitted yield. This means we truly calculate the minimal amount of signal yield that is ruled out by the fit at the desired confidence level. During the optimisation, the minimal amount in addition to the fitted yield was used instead.

Both the cost function and the probability distribution function of an example fit are depicted in Fig. 9.1. As one would expect, the minimum of the former and the maximum of the latter coincide.

9.2 Calculating the expected upper limit

As our signal mass window is blinded, we cannot extract the signal sensitivity from a fit to data. Instead we use a large number of toy data samples and take the median. In total we construct 3000 toys, as described in chapter 8, based on the background estimates for the signal region at the working point. On each of them a fit is performed, including the misID estimate, a falling exponential for the combinatorial background and a signal template. Then the sensitivity to signal decays, N^{sens}, is calculated.

The distribution of this feature is depicted in Fig. 9.2. It is rather broad with a median of 67^{+19}_{-15}. The uncertainties are large due to the low number of decay candidates at the working point, which causes the toy sample distributions to fluctuate strongly.

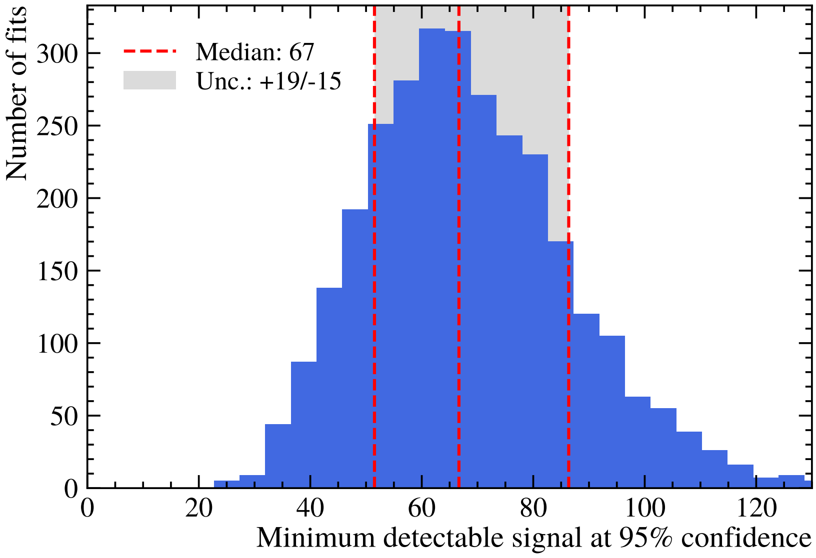


Figure 9.2: Distribution of the signal sensitivity of 3000 fits to toy samples. The confidence interval of the median is chosen as to enclose 34% of fits on either side of the median for a confidence level of 68%.

In principle, this is a fundamental consequence of being in the low statistics regime and cannot be circumvented. The effect is further aggravated by the weighted nature of the misID estimate from which we bootstrap.

We also note that sensitivity obtained by employing this method coincides with the value obtained using the approximation in place for the working point optimisation. While uncertainties are too large to allow a statement on how good the approximation is exactly, the lack of a significant discrepancy suggests that the result of the working point optimisation is not compromised by the approximation.

Having gained a measure of the sensitivity to signal of our method, it is now possible to provide an expected upper limit to the branching ratio of the decay $B^+ \to \mu^+\nu_\mu e^+e^-$ that can be established using the work of this thesis. For this, we use the normalisation performed in section 6.2.

Equation 6.4 is transformed to yield the branching ratio,

$$\mathcal{B}(B^+ \to \mu^+\nu_\mu e^+e^-) = \frac{\epsilon(K^+J/\psi(e^+e^-))}{\epsilon(\mu^+\nu_\mu e^+e^-)} \frac{N^{observed}_{\mu^+\nu_\mu e^+e^-}}{N^{observed}_{K^+e^+e^-}} \times \mathcal{B}(B^+ \to K^+J/\psi(e^+e^-)). \quad (9.5)$$

For yield, efficiency and branching ratio of the reference channel, the values given in section 6.2 are used. As requirements on the isolation variable and the combinatorial BDT are now imposed, the signal selection efficiency, given before in Tab. 6.2, is correspondingly lowered to $\epsilon_{selection} = (9.02 \pm 2.1) \times 10^{-4}$. If no significant signal yield is found, that is equivalent to $N^{observed}_{\mu^+\nu_\mu e^+e^-} < \mathrm{median}(N^{sens})$.

Combining all of this in equation 9.5, we obtain an expected upper limit on the signal branching ratio,

$$\mathcal{B}(B^+ \to \mu^+\nu_\mu e^+e^-) < 1.22^{+0.36}_{-0.28} \times 10^{-7} \quad \text{at } 95\% \text{ CL.} \tag{9.6}$$

The uncertainty on this result is mostly due to the uncertainty of the signal sensitivity.

Chapter 10

Conclusion

The first search for the rare decay $B^+ \to \mu^+\nu_\mu e^+e^-$, is presented in this thesis. It evaluates data recorded by the LHCb experiment in 2016-2018, corresponding to an integrated luminosity of $5.1\,\mathrm{fb}^{-1}$. An expected 95% confidence upper-limit of the branching ratio is estimated at

$$\mathcal{B}(B^+ \to \mu^+\nu_\mu e^+e^-) < 1.22^{+0.36}_{-0.28} \times 10^{-7}, \tag{10.1}$$

which is about three times larger than the theoretical prediction of $(3.78 \pm 0.56) \times 10^{-8}$, but still within one order of magnitude. As a first study, this work did not aim for perfection, but rather for establishing feasibility of the search. Hence, there is room for future improvement of the analysis.

At the point of writing, the analysis is predominantly limited by the low number of recorded decay candidates. The misID estimate is the main source of uncertainty for the fit. If larger control samples of decay candidates passing the inverted PID requirements are available, the uncertainty of the misID estimate diminishes. This yields an immediate improvement in fit sensitivity and thus decreases the upper limit.

In the short term, the available number of decay candidates is increased by the production of several samples tailored for the misID estimate. These are samples which share all selection requirements of the signal sample, except the PID requirement on one lepton track. One such sample is already used in this analysis, the misIDμ sample. It will be joined by samples without PID requirements on either of the electrons. However, while the misIDμ sample is prescaled to contain only 1% of all recorded decays passing the requirements, the new samples contain 5%. As the production requires a central reprocessing of the entire LHCb Run 2 data set, it is time-intensive. The process has started and the samples are expected to be ready in February 2024.

In the long term, it is additional data taking that will provide the biggest gain. With the LHCb collaboration planning to record data corresponding to an integrated luminosity of $15\,\mathrm{fb}^{-1}$ during Run 3 of the LHC, the number of available decay candidates will increase dramatically. Importantly, the LHCb trigger system is switched to a full software trigger, eliminating the L0 stage [61], thereby increasing the efficiency on electron tracks. In contrast to what can be achieved by reprocessing Run 2 data, the additional data recorded in Run 3 also increases the number of signal decays in the data. A future search for the decay, making use of the full Run 2 and Run 3 LHCb data sets, is thus in a strong position to either measure the decay or exclude it at a branching ratio below what is predicted by theory.

While the low number of decays is currently the bottleneck of the search for the target decay, there is further work necessary to fulfil the standards of a proper analysis. Chiefly, this is the estimation of systematic uncertainties. PID efficiencies are known to be imperfectly replicated in LHCb simulations. They can be corrected by comparison with reference samples, which in turn introduces a systematic uncertainty due to the limited sample size. For the misID estimate, proton misID, crossfeed between different hadron misID types and the case of real leptons failing their respective lepton PID requirement need to be considered.

Appendix A

Appendix

A.1 Analysis overview

A.1.1 Derivation of the corrected mass

It is necessary to construct a feature to search for signal signature. For this we start from the B^+ invariant mass and approximate further.

$$\begin{aligned} m_B^2 &= (p_{vis} + p_\nu)^2 = m_{vis}^2 + 2p_{vis}p_\nu + m_\nu^2 \\ &= m_{vis}^2 + 2(E_{vis}E_\nu - \vec{p}_{vis}\vec{p}_\nu) \end{aligned} \tag{A.1}$$

As the neutrino is massless its mass is dropped. We can also split up the momentum into a part that is parallel to the B flight direction and a part that is transverse. Momentum conservation then implies that the transverse momentum components of the neutrino and the visible momenta sum up to zero.

$$\begin{aligned} p_{vis} \cdot p_\nu &= E_{vis}E_\nu - \vec{p}_{vis}\vec{p}_\nu = E_{vis}E_\nu - \vec{p}_{\parallel vis}\vec{p}_{\parallel \nu} - \vec{p}_{\perp vis}\vec{p}_{\perp \nu} \\ &= E_{vis}E_\nu - \vec{p}_{\parallel vis}\vec{p}_{\parallel \nu} + \vec{p}_\perp^{\,2} \end{aligned} \tag{A.2}$$

As the product of two four-momenta is Lorentz invariant, the reference frame can be boosted such that $p_{\perp\nu}$ vanishes, without changing the expression. Further we make the approximation that visible momentum and neutrino momentum are identical in their parallel component $p_{\parallel vis} = p_{\parallel \nu}$, making the latter vanish as well.

$$\begin{aligned} p_{vis} \cdot p_\nu &\approx \sqrt{(m_{vis}^2 + \vec{p}_{\perp vis}^{\,2})(m_\nu^2 + \vec{p}_{\perp \nu}^{\,2})} + \vec{p}_\perp^{\,2} \\ &\approx \sqrt{m_{vis}^2 + \vec{p}_\perp^{\,2}} \cdot |\vec{p}_\perp| + \vec{p}_\perp^{\,2} \end{aligned} \tag{A.3}$$

From this a final expression can be constructed.

$$\begin{aligned} m_B^2 &\approx m_{vis}^2 + 2\sqrt{m_{vis}^2 + \vec{p}_\perp^2} \cdot |\vec{p}_\perp| + 2\vec{p}_\perp^2 \\ &\approx \left(\sqrt{m_{vis}^2 + \vec{p}_\perp^2} + |\vec{p}_\perp|\right)^2 \\ m_{corr} &\equiv \sqrt{m_{vis}^2 + \vec{p}_\perp^2} + |\vec{p}_\perp| \end{aligned} \tag{A.4}$$

A.2 Signal studies

A.2.1 Kinematic parameter correlation

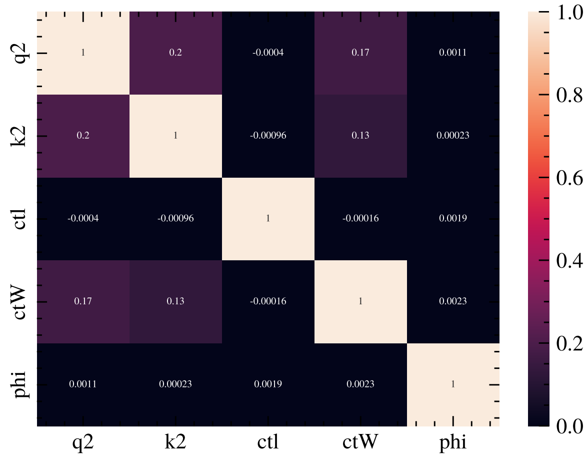


Figure A.1: Correlation of the kinematic parameters in the old model.

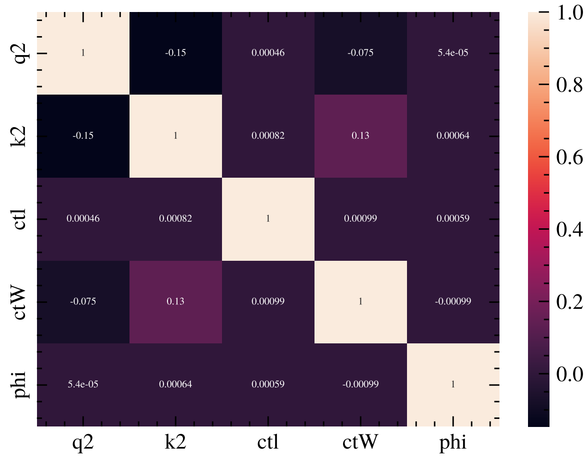


Figure A.2: Correlation of the kinematic parameters in the new model.

A.2.2 Fit of the normalisation channel

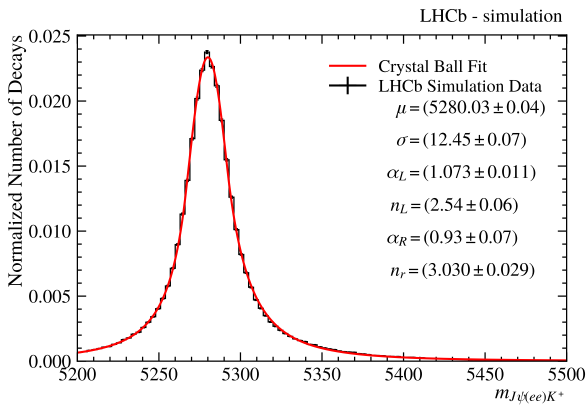


Figure A.3: A double Crystal Ball function fitted to a simulated LHCb data sample.

A.3 Background

A.3.1 BDT features

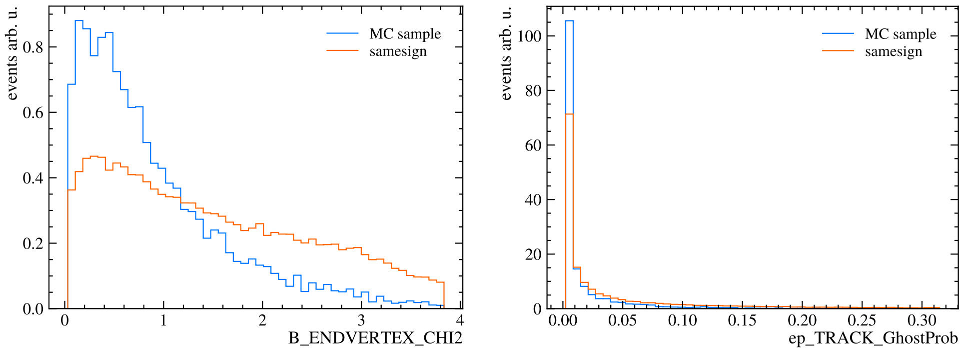


Figure A.4: χ^2_{vertex} (LH) and $p_{ghost}(e^+)$ (RH) distributions of samesign and signal simulation sample.

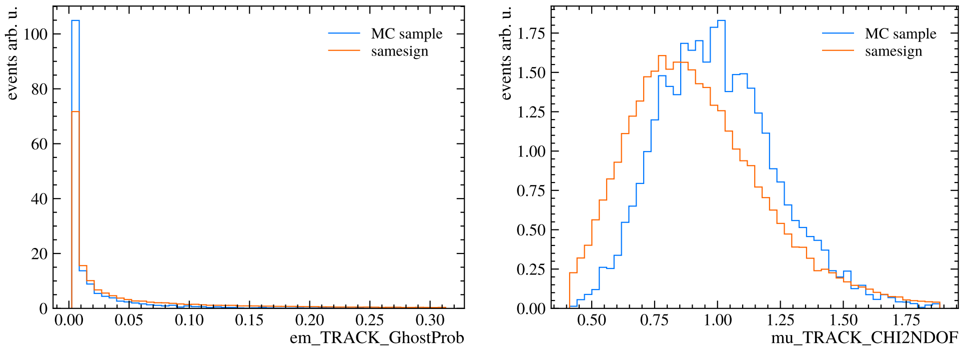


Figure A.5: $p_{ghost}(e^-)$ (LH) and $\chi^2_{Track}(\mu)$ (RH) distributions of samesign and signal simulation sample.

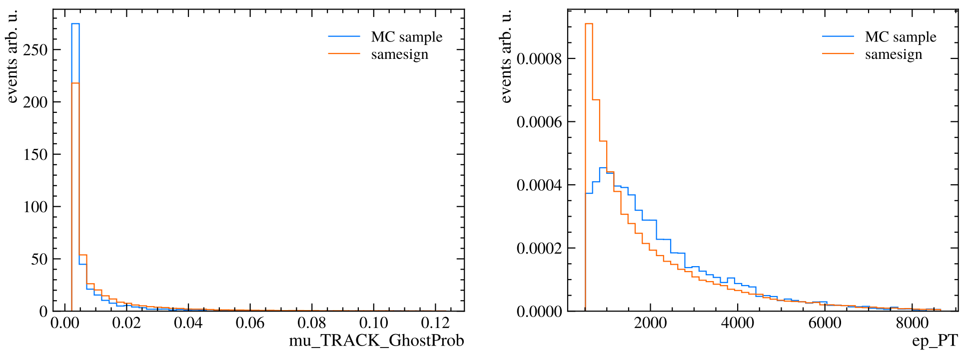


Figure A.6: $p_{ghost}(\mu^+)$ (LH) and $p_T(e^+)$ (RH) distributions of samesign and signal simulation sample.

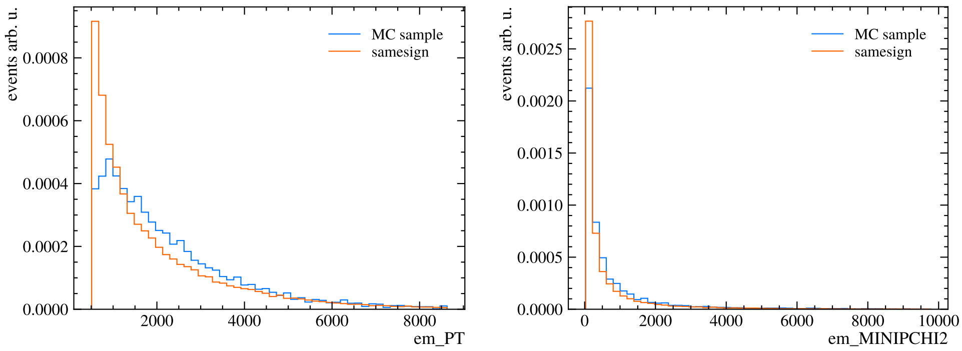


Figure A.7: $p_T(e^-)$ (LH) and Minimal impact parameter of e^- (RH) distributions of samesign and signal simulation sample.

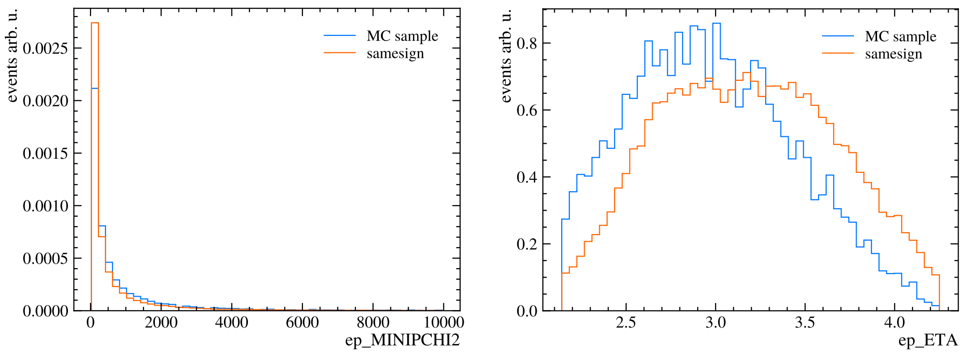


Figure A.8: Minimal impact parameter of e^+ (LH) and $\eta(e^+)$ (RH) distributions of samesign and signal simulation sample.

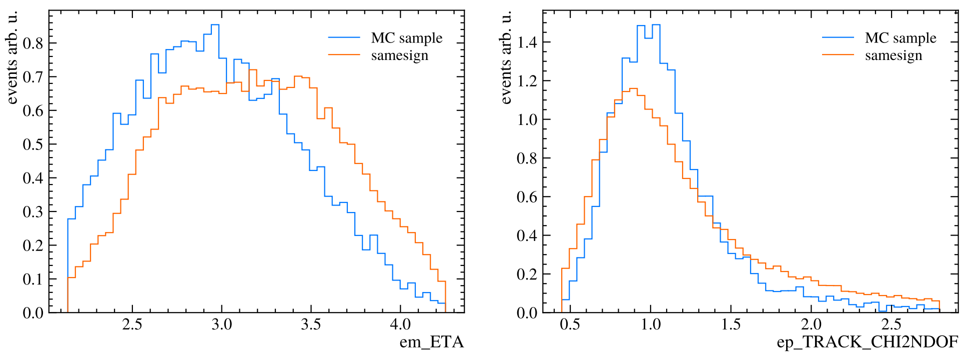


Figure A.9: $\eta(e^-)$ (LH) and $\chi^2_{Track}(e^+)$ (RH) distributions of samesign and signal simulation sample.

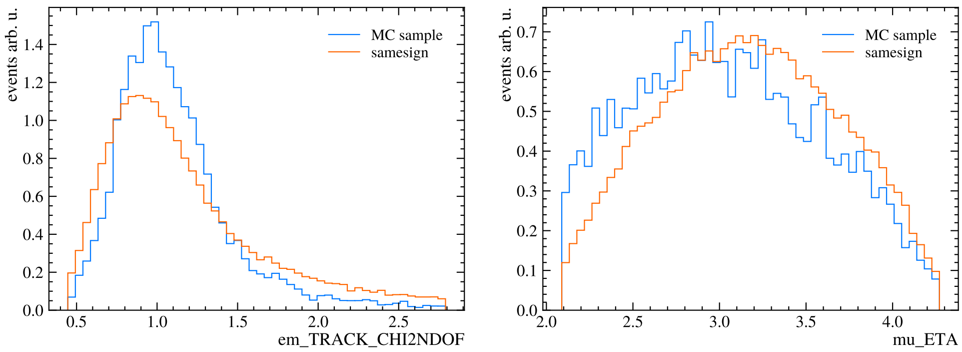


Figure A.10: $\chi^2_{Track}(e^-)$ (LH) and $\eta(\mu)$ (RH) distributions of samesign and signal simulation sample.

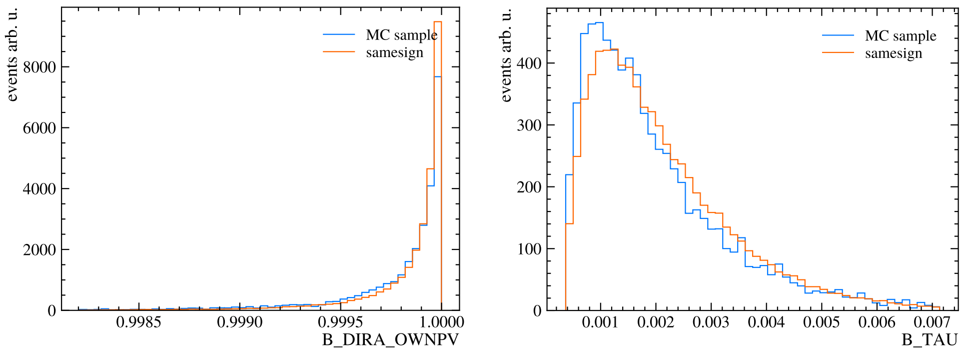


Figure A.11: Cosine of B^+ direction angle (LH) and τ_B (RH) distributions of samesign and signal simulation sample.

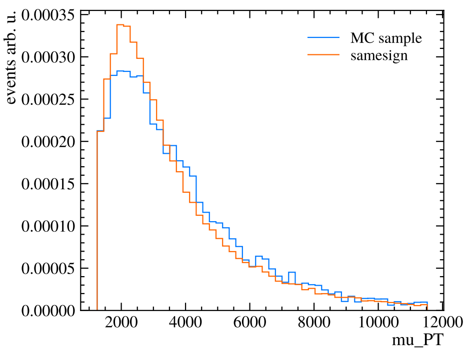


Figure A.12: $p_t(\mu)$ distribution of samesign and signal simulation sample.

A.3.2 Correlations of the BDT features

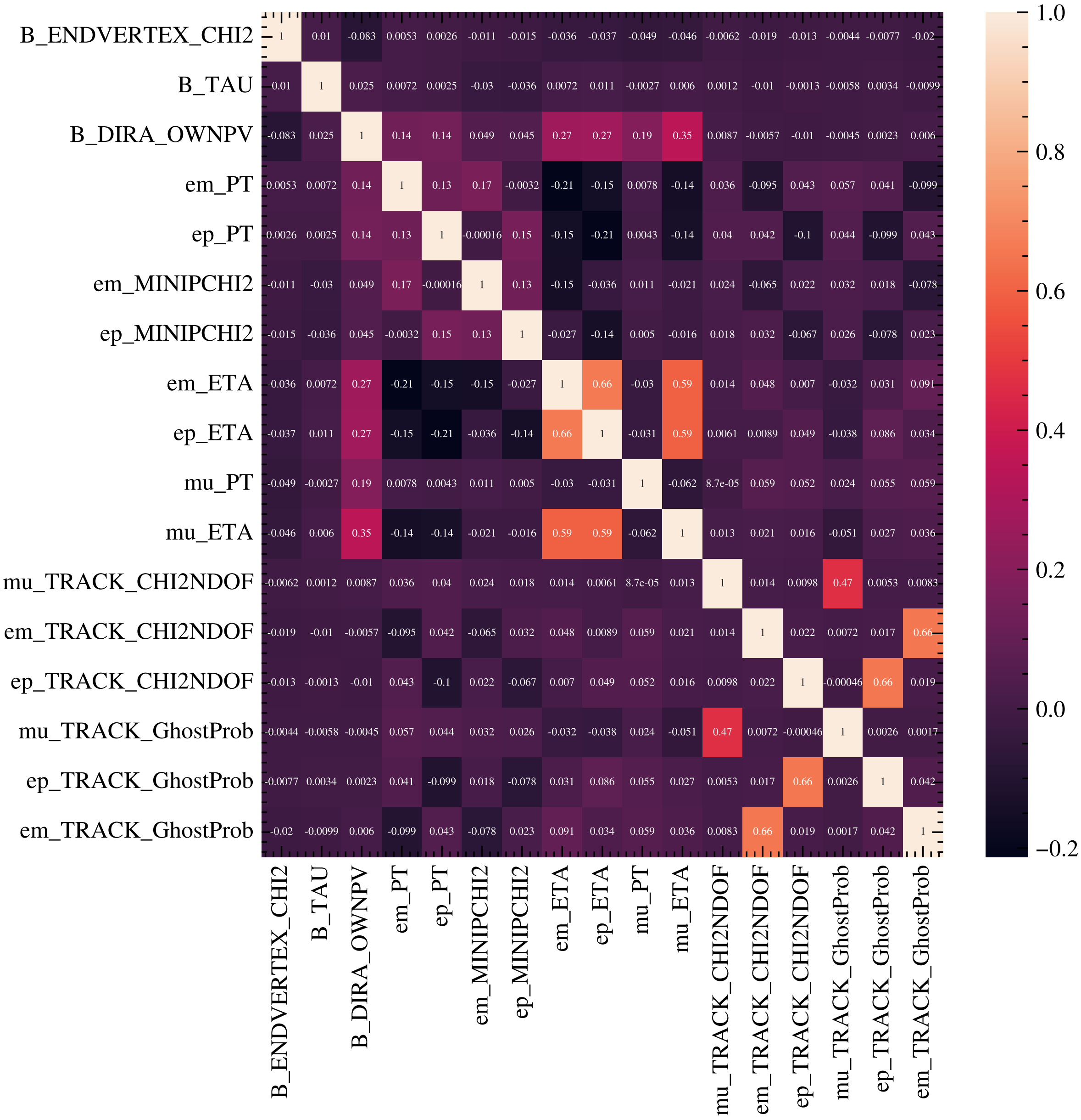


Figure A.13: Correlation matrix for the combinatorial BDT training parameters in the same sign sample.

In FigA.13 the correlation in the samesign sample of the BDT input features can be seen. We note large strong correlation between the pseudorapidities. They were still all used for the BDT, as the characterisation of tracks based on transverse momentum and pseudorapidity in the misID estimate in section 7.2 gives these features a special position. Further, for a given track ghost probability and χ^2_{track} correlate strongly. Both features were still used, as removing either led to a decrease in separation power.

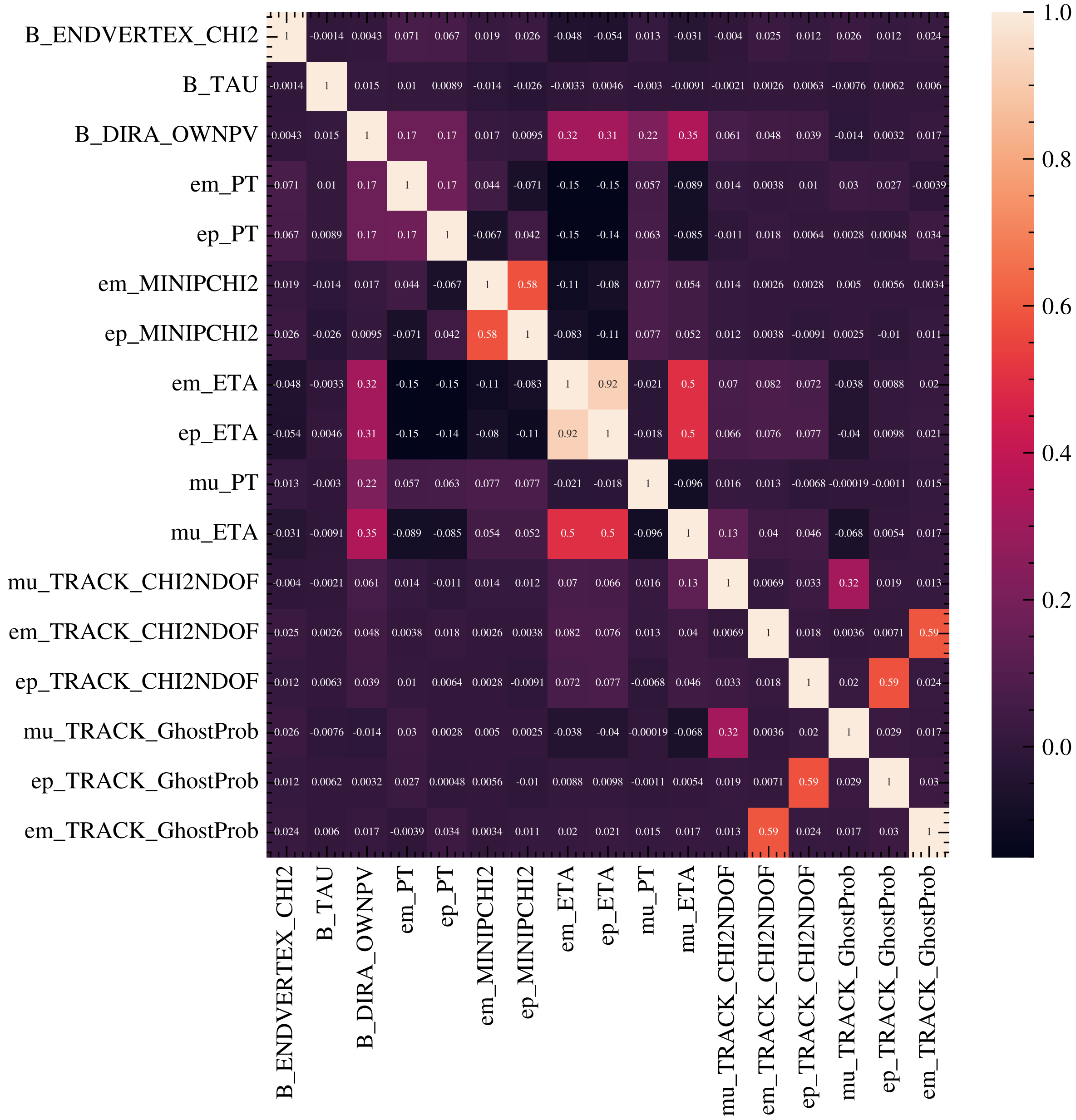


Figure A.14: Correlation matrix for the combinatorial BDT training parameters in the signal LHCb simulation sample.

A.3.3 Overfitting

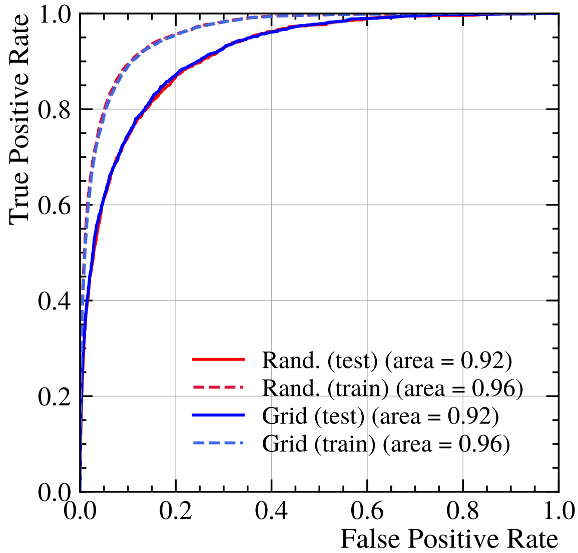


Figure A.15: ROC curves, once using the optimal hyperparameters from gridsearch and once using those from random search.

Overtraining is apparent. We also note that the differences between the two different configurations are very small.

A.3.4 ROC curve of the ToaK sample

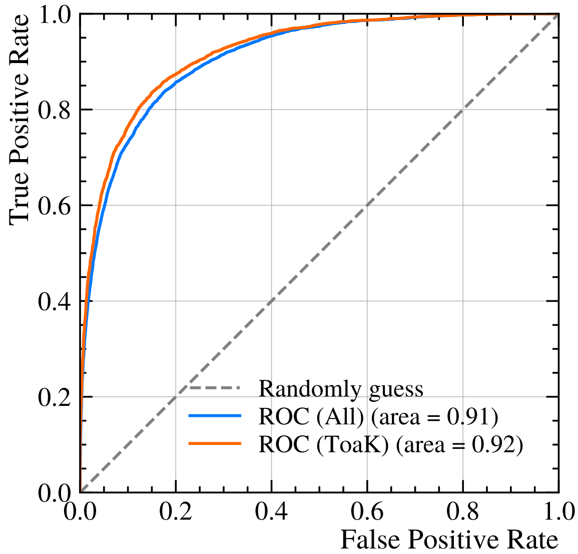


Figure A.16: ROC curves, once using the entire samesign sample and once only using the ToaK sample as background testing data. In both cases the training used the entire samesign data sample

The AUC is slightly larger if only the ToaK data is used to stand in as background for calculating the ROC curve thresholds. As the ToaK sample is expected to be purely combinatorial this is a good sign. More importantly however, is that the AUC is not less, which means that the combinatorial has not only learned to classify structure in the samesign sample which is not expected in combinatorial.

A.3.5 PIDCalib histograms

While all configurations were used, only plots for the 2017 data taking period with negative magnet polarity are shown here. In all plots the upper left histogram shows the efficiency of the electron PID cut on the reference pion sample, the upper right histogram shows the efficiency of the inverted PID cut and the lower plot shows the ratio gained from binwise dividing electron cut efficiency by hadron cut efficiency.

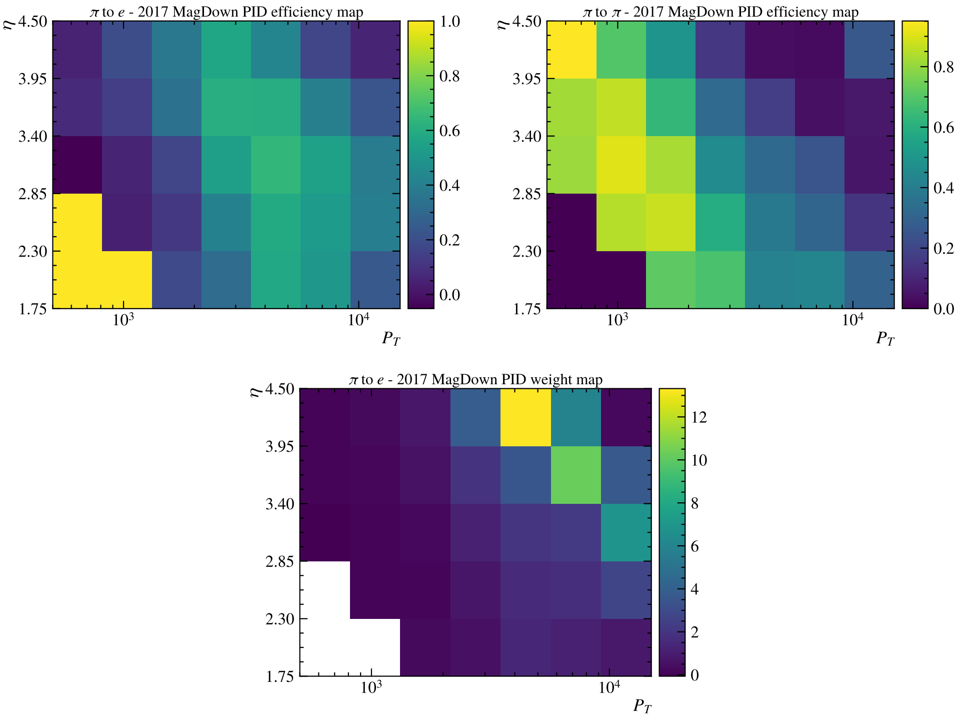


Figure A.17: Histograms for $\pi \to e$ misID, assuming that the electron candidate is triggered on.

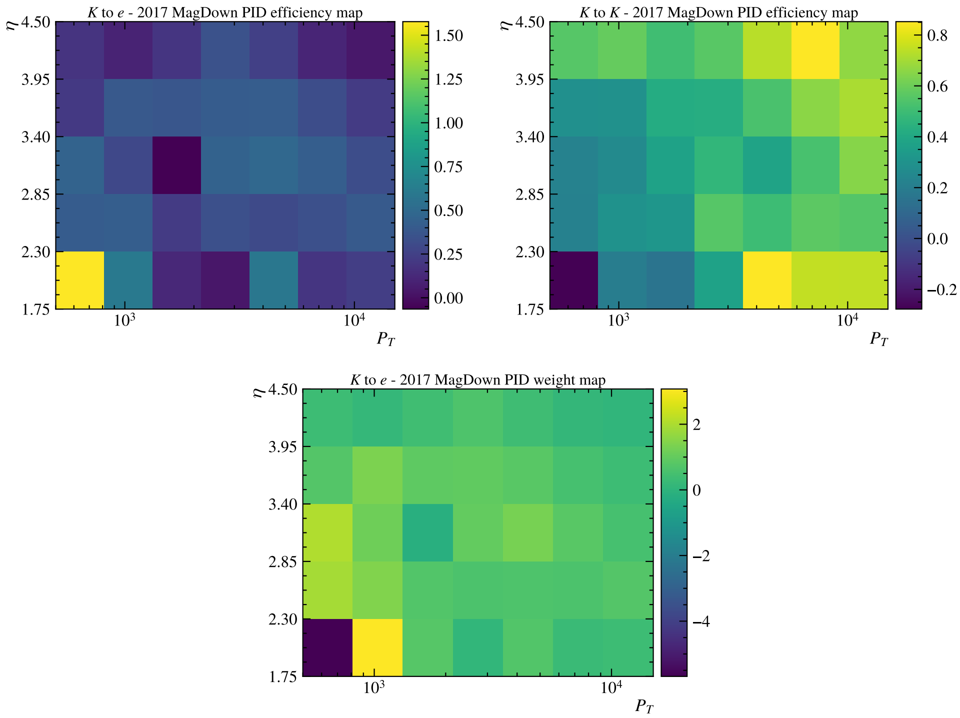


Figure A.18: Histograms for $K \to e$ misID, assuming that the electron candidate was not triggered on.

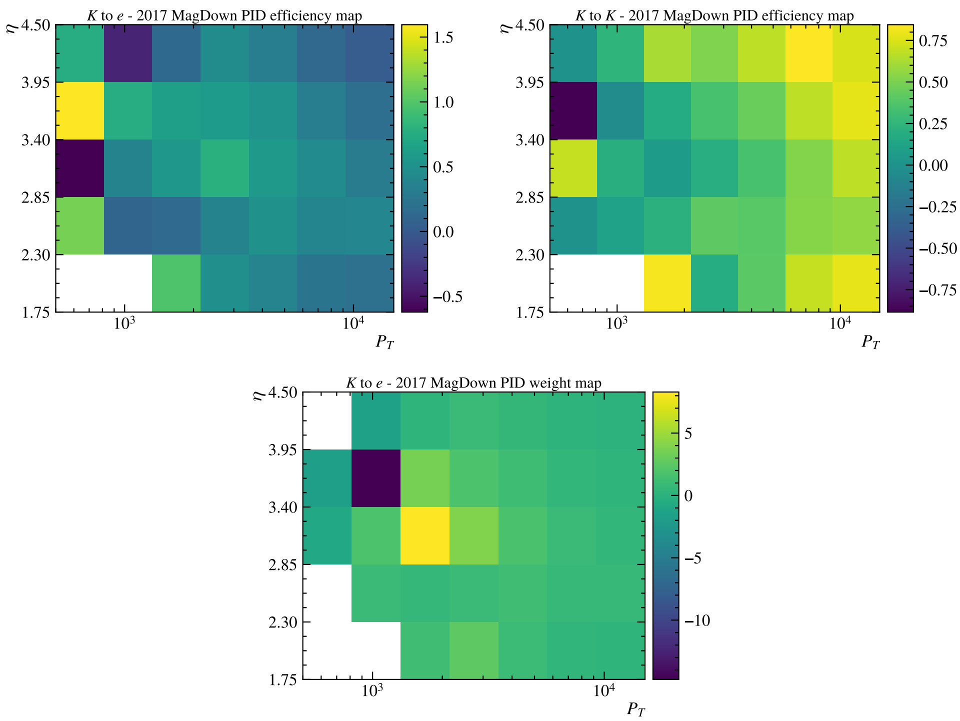


Figure A.19: Histograms for $K \to e$ misID, assuming that the electron candidate was triggered on.

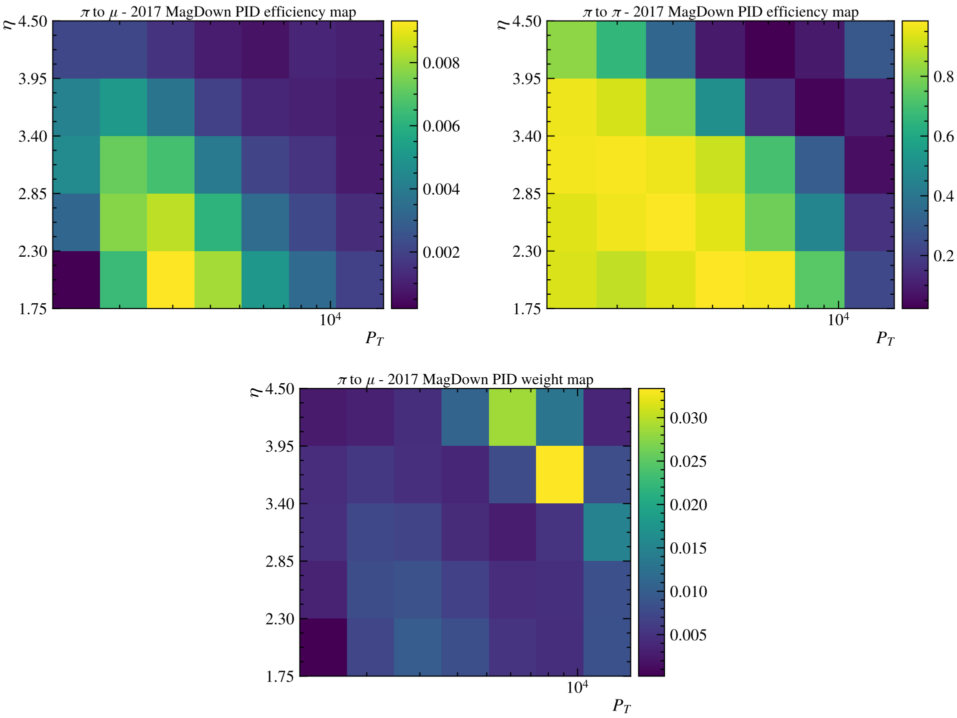


Figure A.20: Histograms for $\pi \to \mu$ misID.

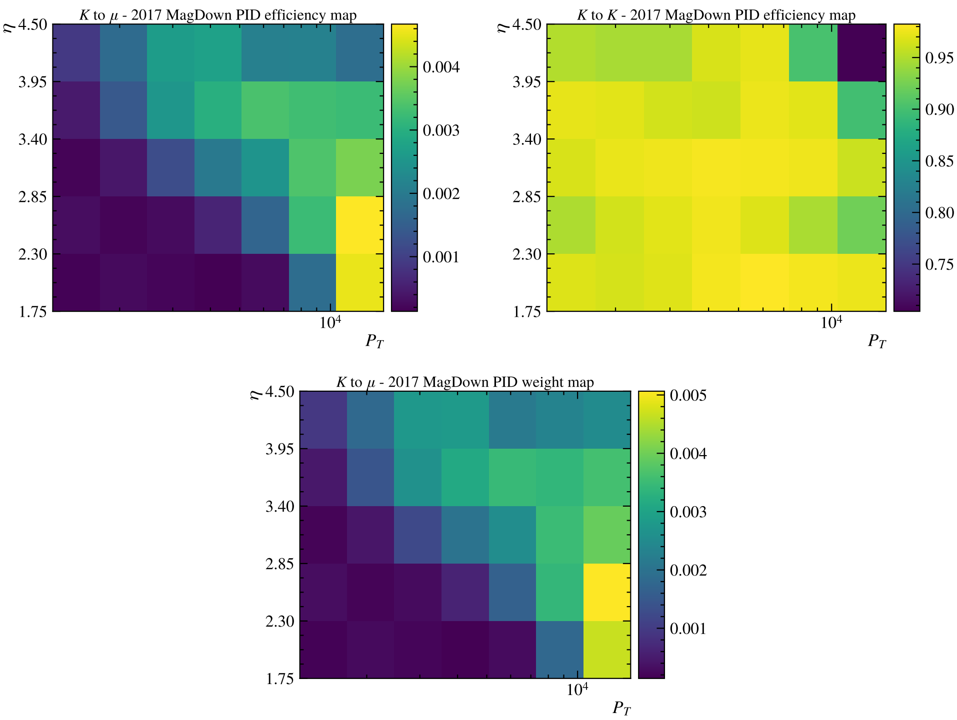


Figure A.21: Histograms for $K \to \mu$ misID.

A.3.6 Cross check $B^+ \to K^+ J/\psi(ee)$

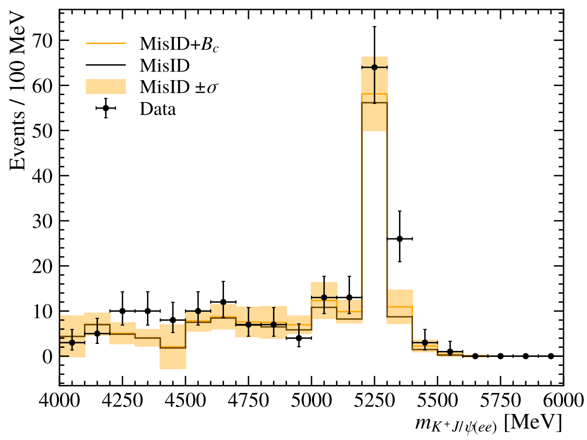


Figure A.22: The measured distribution in $m_{K^+J/\psi(ee)}$ compared to the estimated distribution from the weighted hadron samples summed with the estimated remaining B_c^+ contribution.

A.4 Toys

A.4.1 Working point optimization

In Fig. A.23 we see failure rate and figure of merit curves for all isolation BDT values considered. We note that for higher isolation BDT requirements the failure rate increases drastically, with no gain in the figure of merit.

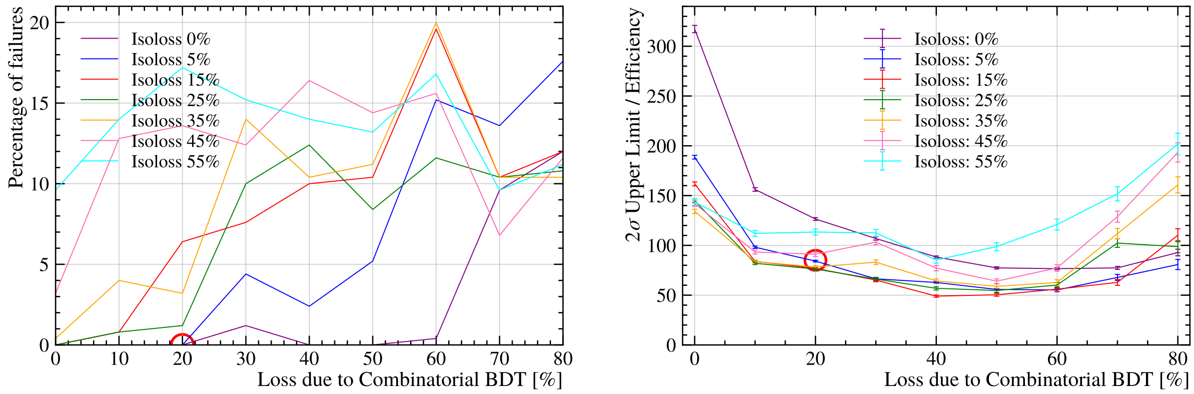


Figure A.23: Failure percentage (left) and signal sensitivity (right) for different BDT working points. Combinatorial proxy used is the dielectron mass sideband combinatorial contribution. The uncertainty on the signal sensitivity is the uncertainty of the mean.

A.4.2 Punzi figure of merit

The Punzi figure of merit is defined as

$$\frac{\epsilon}{a/2 + \sqrt{B}}, \tag{A.5}$$

with a the considered level of significance in σ, ϵ the signal efficiency and B the number of background events. As can be seen in Fig.A.23, the Punzi figure of merit also shares roughly shares the structures of the upper limit based figure of merit, in that it changes drastically with the first implemented BDT requirement, but then plateaus. Among the stable working points, the maximum of the Punzi figure of merit is at the same working as the minimum of the used figure of merit, with an isolation BDT loss of 5% and a combinatorial BDT loss of 30%.

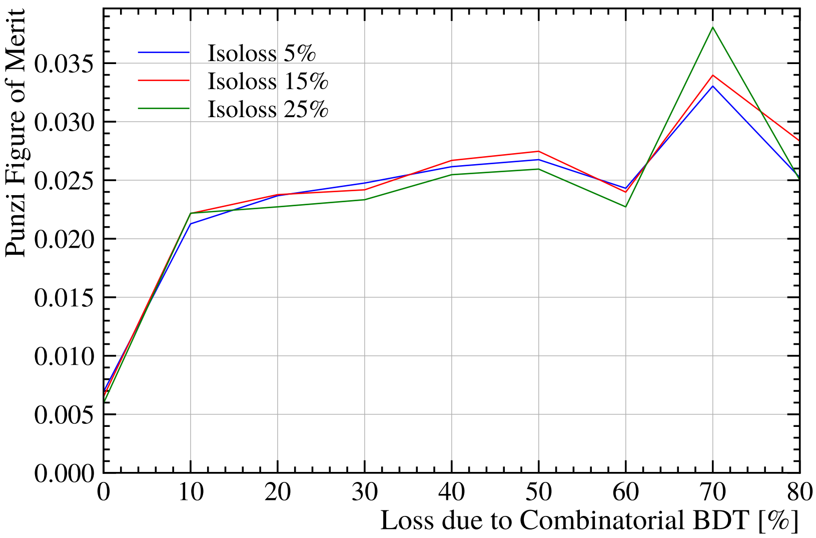


Figure A.24: Punzi figure of merit for different working points. The curves are results of the second iteration of the working point optimization.

Bibliography

[1] S. Chatrchyan et al. “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC”. In: *Physics Letters B* 716.1 (2012), pp. 30–61. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2012.08.021. URL: https://www.sciencedirect.com/science/article/pii/S0370269312008581.

[2] G. Aad et al. “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”. In: *Physics Letters B* 716.1 (2012), pp. 1–29. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2012.08.020. URL: https://www.sciencedirect.com/science/article/pii/S037026931200857X.

[3] Gelb et al. “Search for the rare decay of $B^+ \to \ell^+\nu_\ell\gamma$ with improved hadronic tagging”. In: *Phys. Rev. D* 98 (11 Dec. 2018), p. 112016. DOI: 10.1103/PhysRevD.98.112016. URL: https://link.aps.org/doi/10.1103/PhysRevD.98.112016.

[4] Andreas Heller. “Search for $B^+ \to \ell^+\nu_\ell\gamma$ decays with hadronic tagging using the full Belle data sample”. PhD thesis. Karlsruher Institut für Technologie, 2015.

[5] R. Aaij et. al. “Search for the rare decay $B^+ \to \mu^+\nu_\mu\mu^-\mu^+$”. In: *The European Physical Journal C* 79.8 (Aug. 2019). DOI: 10.1140/epjc/s10052-019-7112-x. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-019-7112-x.

[6] R. L. Workman et al. “Review of Particle Physics”. In: *PTEP* 2022 (2022), p. 083C01. DOI: 10.1093/ptep/ptac097.

[7] Paul Langacker. *The Standard Model and Beyond*. 2nd. Taylor & Francis, 2017.

[8] J. Charles et al. “CP violation and the CKM matrix: assessing the impact of the asymmetric B factories”. In: *The European Physical Journal C* 41.1 (May 2005), pp. 1–131. DOI: 10.1140/epjc/s2005-02169-1. URL: https://doi.org/10.1140%2Fepjc%2Fs2005-02169-1.

[9] M. Beneke et al. "QCD factorization for exclusive non-leptonic B-meson decays: general arguments and the case of heavy–light final states". In: *Nuclear Physics B* 591.1 (2000), pp. 313–418. ISSN: 0550-3213. DOI: https://doi.org/10.1016/S0550-3213(00)00559-9. URL: https://www.sciencedirect.com/science/article/pii/S0550321300005599.

[10] Alexander Khodjamirian, Blaženka Melić, and Yu-Ming Wang. *A guide to the QCD light-cone sum rule for b-quark decays*. 2023. arXiv: 2311.08700 [hep-ph].

[11] Lyndon Evans and Philip Bryant. "LHC Machine". In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08001. DOI: 10.1088/1748-0221/3/08/S08001. URL: https://dx.doi.org/10.1088/1748-0221/3/08/S08001.

[12] "LHCb detector performance". In: *International Journal of Modern Physics A* 30.07 (Mar. 2015), p. 1530022. DOI: 10.1142/s0217751x15300227. URL: https://doi.org/10.1142%2Fs0217751x15300227.

[13] A Augusto et al. Alves. "The LHCb Detector at the LHC". In: *JINST* 3 (2008). Also published by CERN Geneva in 2010, S08005. DOI: 10.1088/1748-0221/3/08/S08005. URL: https://cds.cern.ch/record/1129809.

[14] Mark Thomson. *Modern Particle Physics*. Cambridge University Press, 2013. DOI: 10.1017/CBO9781139525367.

[15] Christian Elsasser. *$\bar{b}b$ production angle plots*. https://lhcb.web.cern.ch/lhcb/speakersbureau/html/bb_ProductionAngles.html, accessed 15th August 2023.

[16] 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.

[17] "Performance of the LHCb Outer Tracker". In: *Journal of Instrumentation* 9.01 (Jan. 2014), P01002–P01002. DOI: 10.1088/1748-0221/9/01/p01002. URL: https://doi.org/10.1088%2F1748-0221%2F9%2F01%2Fp01002.

[18] Alves et al. "Performance of the LHCb muon system". In: *Journal of Instrumentation* 8.02 (Feb. 2013), P02022–P02022. ISSN: 1748-0221. DOI: 10.1088/1748-0221/8/02/p02022. URL: http://dx.doi.org/10.1088/1748-0221/8/02/P02022.

[19] R. Aaij et al. "Design and performance of the LHCb trigger and full real-time reconstruction in Run 2 of the LHC". In: *Journal of Instrumentation* 14.04 (Apr. 2019), P04013–P04013. ISSN: 1748-0221. DOI: 10.1088/1748-0221/14/04/p04013. URL: http://dx.doi.org/10.1088/1748-0221/14/04/P04013.

[20] *The MOORE Project*. https://lhcbdoc.web.cern.ch/lhcbdoc/moore/, accessed 14th September 2023.

[21] *The LHCb Starterkit lessons - The LHCb data flow.* https://lhcb.github.io/starterkit-lessons/first-analysis-steps/dataflow.html, accessed: 13th September 2023.

[22] *The BRUNEL Project.* https://lhcbdoc.web.cern.ch/lhcbdoc/brunel/, accessed 14th September 2023.

[23] *The DAVINCI Project.* https://lhcbdoc.web.cern.ch/lhcbdoc/davinci/, accessed 14th September 2023.

[24] I Belyaev et. al. "Handling of the generation of primary events in Gauss, the LHCb simulation framework". In: *Journal of Physics: Conference Series* 331.3 (Dec. 2011), p. 032047. DOI: 10.1088/1742-6596/331/3/032047. URL: https://dx.doi.org/10.1088/1742-6596/331/3/032047.

[25] Torbjörn Sjöstrand, Stephen Mrenna, and Peter Skands. "PYTHIA 6.4 physics and manual". In: *Journal of High Energy Physics* 2006.05 (May 2006), pp. 026–026. DOI: 10.1088/1126-6708/2006/05/026. URL: https://doi.org/10.1088%2F1126-6708%2F2006%2F05%2F026.

[26] David J. Lange. "The EvtGen particle decay simulation package". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 462.1 (2001). BEAUTY2000, Proceedings of the 7th Int. Conf. on B-Physics at Hadron Machines, pp. 152–155. ISSN: 0168-9002. DOI: https://doi.org/10.1016/S0168-9002(01)00089-4. URL: https://www.sciencedirect.com/science/article/pii/S0168900201000894.

[27] J. et. al. Allison. "Geant4 developments and applications". In: *IEEE Transactions on Nuclear Science* 53.1 (2006), pp. 270–278. DOI: 10.1109/TNS.2006.869826.

[28] *The BOOLE Project.* https://lhcbdoc.web.cern.ch/lhcbdoc/boole/, accessed 14th September 2023.

[29] Martin Beneke et al. "QCD factorization of the four-lepton decay $B^- \to \ell\bar{\nu}_\ell \ell^{(\prime)}\bar{\ell}^{(\prime)}$". In: *Eur. Phys. J. C* 81.7 (2021), p. 638. DOI: 10.1140/epjc/s10052-021-09388-y. arXiv: 2102.10060 [hep-ph].

[30] Stephan Kürten et al. "Dispersion relations for $B^- \to \ell^-\overline{\nu}_\ell \ell'^- \ell'^+$ form factors". In: *Phys. Rev. D* 107 (5 Mar. 2023), p. 053006. DOI: 10.1103/PhysRevD.107.053006. URL: https://link.aps.org/doi/10.1103/PhysRevD.107.053006.

[31] Joshua P. Ellis. "Ti k Z-Feynman: Feynman diagrams with Ti k Z". In: *Computer Physics Communications* 210 (Jan. 2017), pp. 103–123. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2016.08.019. URL: http://dx.doi.org/10.1016/j.cpc.2016.08.019.

[32] Y. et. al. Amhis. "Averages of b-hadron, c-hadron, and τ-lepton properties as of 2021". In: *Physical Review D* 107.5 (Mar. 2023). ISSN: 2470-0029. DOI: 10.1103/physrevd.107.052008. URL: http://dx.doi.org/10.1103/PhysRevD.107.052008.

[33] Chan Beom Park. "Could MT2 be a singularity variable?" In: *Journal of High Energy Physics* 2021.11 (Nov. 2021). DOI: 10.1007/jhep11(2021)042. URL: https://doi.org/10.1007%2Fjhep11%282021%29042.

[34] R Brun, F Rademakers, and S Panacek. "ROOT, an object oriented data analysis framework". In: (2000). URL: https://cds.cern.ch/record/491486.

[35] Jim Pivarski et. al. *scikit-hep/uproot: 3.12.0*. Version 3.12.0. July 2020. DOI: 10.5281/zenodo.3952728. URL: https://doi.org/10.5281/zenodo.3952728.

[36] The pandas development team. *pandas-dev/pandas: Pandas*. Version latest. Feb. 2020. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[37] Wes McKinney. "Data Structures for Statistical Computing in Python". In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.

[38] Henry Schreiner, Jim Pivarski, and Saransh Chopra. *vector*. Version v0.10.0. https://doi.org/10.5281/zenodo.7054478. Sept. 2022. DOI: 10.5281/zenodo.7054478.

[39] *PIDCalib2*. https://pypi.org/project/pidcalib2/, accessed 14th September 2023.

[40] Lucio Anderlini et al. *The PIDCalib package*. Tech. rep. Geneva: CERN, 2016. URL: https://cds.cern.ch/record/2202412.

[41] Roel Aaij et Al. "Selection and processing of calibration samples to measure the particle identification performance of the LHCb experiment in Run 2". In: *EPJ Techniques and Instrumentation* 6.1 (Feb. 2019). DOI: 10.1140/epjti/s40485-019-0050-z. URL: https://doi.org/10.1140%2Fepjti%2Fs40485-019-0050-z.

[42] Hans Dembinski and Piti Ongmongkolkul et al. "scikit-hep/iminuit". In: (Dec. 2020). https://doi.org/10.5281/zenodo.3949207. DOI: 10.5281/zenodo.3949207.

[43] F. James and M. Roos. "Minuit: A System for Function Minimization and Analysis of the Parameter Errors and Correlations". In: *Comput. Phys. Commun.* 10 (1975), pp. 343–367. DOI: 10.1016/0010-4655(75)90039-9.

[44] Hans Peter Dembinski and Ahmed Abdelmotteleb. *A fast and stable approximate maximum-likelihood method for template fits*. 2022. arXiv: 2206.12346 [stat.ME].

[45] *LHCb Luminosity Project.* https://twiki.cern.ch/twiki/bin/viewauth/LHCbPhysics/LHCbLuminosity, accessed: 24th October 2023.

[46] Mikhail A. Ivanov and Dmitri Melikhov. "Theoretical analysis of the leptonic decays $B \to \ell\ell\ell'\overline{\nu}'_\ell$, journal = Physical Review D". In: 105.1 (Jan. 2022). DOI: 10.1103/physrevd.105.014028. URL: https://doi.org/10.1103%2Fphysrevd.105.014028.

[47] D. van Dyk et. al. "EOS: a software for flavor physics phenomenology". In: *The European Physical Journal C* 82.6 (June 2022). DOI: 10.1140/epjc/s10052-022-10177-4. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-022-10177-4.

[48] Danilina, Anna, Nikitin, Nikolay, and Toms, Konstantin. "Rare four leptonic B-mesons decays with a neutrino in final state". In: *EPJ Web Conf.* 222 (2019), p. 03019. DOI: 10.1051/epjconf/201922203019. URL: https://doi.org/10.1051/epjconf/201922203019.

[49] Alex Rogozhnikov. ***hep_ml***. https://github.com/arogozhnikov/hep_ml. 2015.

[50] Alex Rogozhnikov. "Reweighting with Boosted Decision Trees". In: *Journal of Physics: Conference Series* 762 (Oct. 2016), p. 012036. DOI: 10.1088/1742-6596/762/1/012036. URL: https://doi.org/10.1088%2F1742-6596%2F762%2F1%2F012036.

[51] Jonas Eschle et al. "zfit: Scalable pythonic fitting". In: *SoftwareX* 11 (2020), p. 100508. ISSN: 2352-7110. DOI: https://doi.org/10.1016/j.softx.2020.100508. URL: https://www.sciencedirect.com/science/article/pii/S2352711019303851.

[52] Tianqi Chen and Carlos Guestrin. "XGBoost". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, Aug. 2016. DOI: 10.1145/2939672.2939785. URL: https://doi.org/10.1145%2F2939672.2939785.

[53] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[54] R. Aaij et al. "Test of lepton universality in $b \to s\ell^+\ell^-$ decays". In: *Phys. Rev. Lett.* 131.5 (2023), p. 051803. DOI: 10.1103/PhysRevLett.131.051803. arXiv: 2212.09152 [hep-ex].

[55] B. Khanji. *Charged Isolation Challenge.* https://indico.cern.ch/event/533133/contributions/2172015/attachments/1277028/1895185/C_Isolation_BK.pdf, Accessed: 21st October 2023.

[56] Aaij, R. et al. "Measurement of the b-Quark Production Cross Section in 7 and 13 TeV pp Collisions". In: *Phys. Rev. Lett.* 118 (5 Feb. 2017), p. 052002. DOI: 10.1103/PhysRevLett.118.052002. URL: https://link.aps.org/doi/10.1103/PhysRevLett.118.052002.

[57] Aaij, R. et al. "Measurement of the B_c^- meson production fraction and asymmetry in 7 and 13 TeV pp collisions". In: *Physical Review D* 100.11 (Dec. 2019). DOI: 10.1103/physrevd.100.112006. URL: https://doi.org/10.1103%2Fphysrevd.100.112006.

[58] Aoife Bharucha Martino Borsato Meril Reboud. Private communication. 2023.

[59] B. Efron. "Bootstrap Methods: Another Look at the Jackknife". In: *The Annals of Statistics* 7.1 (1979), pp. 1–26. DOI: 10.1214/aos/1176344552. URL: https://doi.org/10.1214/aos/1176344552.

[60] Giovanni Punzi. *Sensitivity of searches for new signals and its optimization.* 2003. arXiv: physics/0308063 [physics.data-an].

[61] LHCb collaboration. *The LHCb upgrade I.* 2023. arXiv: 2305.10515 [hep-ex].

Erklärung:

Ich versichere, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den 01. Dezember 2023 Tom Wolf