Rijksuniversiteit Groningen

Bachelor Thesis

Predicting the origin of the first hits measured by the Vertex Locator for $B_{(c)}^{+} \rightarrow \tau + \nu_{\tau}$

Author:
Frederiek de Bruine

Supervisor:
dr. K.A.M de Bruyn

Abstract

At the LHCb experiment at CERN, proton beams are collided to study the decays of hadrons that include a beauty quark. To measure these short-lived particles before they decay, one of the detectors at the LHCb experiment, the Vertex Locator (VELO), is placed at only a 5.1 mm distance from the proton-proton collision point to measure the location of charged particles that hit the detector. A way to test the Standard Model is to measure the branching fractions of the $B_c^+ \to \tau(\to \pi^+\pi^+\pi^-\overline{\nu}_\tau)\nu_\tau$ and $B^+ \to \tau(\to \pi^+\pi^+\pi^-\overline{\nu}_\tau)\nu_\tau$ decays. To possibly improve the separation of background decays from these signal decays, the goal of this research is to determine the $B_{(c)}^+$-to-τ ratio of first hits measured at the VELO. The origin of the first hits is determined by comparing the location of the first hits with the location of the decay vertex of the $B_{(c)}^+$ particle for data simulated with the RapidSim package. For the B_c^+-decay this resulted in 90.6% of first hits to be classified as τ-hits and 9.4% as B_c^+ hits and for the B^+-decay in 30.9% as τ hits and 69.1% as B^+ hits. Relatively more B_c^+ particles than B^+ particles decay before they reach the VELO, which can be explained by the differences in the lifetimes of the particle. Since the decay vertex of the $B_{(c)}^+$ particle cannot be determined experimentally, it is also studied whether it is possible to predict the origin of the first VELO hits with only the observed parameters. By implementing a Multivariate Analysis, it is shown that this is possible for both the B_c^+ and B^+ decay.

Contents

1 Introduction

In particle physics, a framework has been developed that describes the most fundamental building blocks of the universe. This framework, the Standard Model (SM), is excellent at describing how elementary particles interact, but there are also limitations to the model [2]. Probing these limitations can lead to a better understanding of particle physics and can lead to physics beyond the SM.

The SM of particle physics describes what is known about elementary particles and the fundamental forces that govern them, excluding gravity [2]. The SM divides elementary particles into two groups, fermions and bosons. Fermions are matter particles, while bosons are particles that carry a force. The fermions can further be divided into leptons and quarks based on the interactions that they participate in. Furthermore, the group of leptons consists of three generations of charged particles, electrons, muons and tauons, and of the corresponding neutral neutrinos. The charged leptons interact through the electromagnetic and the weak force, and the neutrinos only interact through the weak force.

One of the possible limitations of the SM arises when considering the properties of charged leptons. The SM predicts that the charged leptons interact almost identically, the only difference being their mass. This principle is called lepton flavour universality (LFU). Until recently, measurements were in agreement with LFU. However, new research performed at the LHCb, Belle and BaBar experiments shows small deviations from the SM prediction [3] [4] [5]. These deviations have not reached the five-σ significance that particle physicists need to indicate that LFU is likely violated and therefore physicists are still looking for additional measurements that could confirm LFU violation. A method to study whether LFU holds is to compare the branching fractions of decays that only differ by lepton flavour. The branching fractions can be measured by performing high-energy collider experiments, such as the Large Hadron Collider beauty experiment (LHCb).

The LHCb experiment is an experiment at CERN where the beauty quark is studied. This experiment focuses on searching for new physics by studying the decays of particles containing a beauty or a charm quark. At the LHCb there are proton-proton collisions that produce many

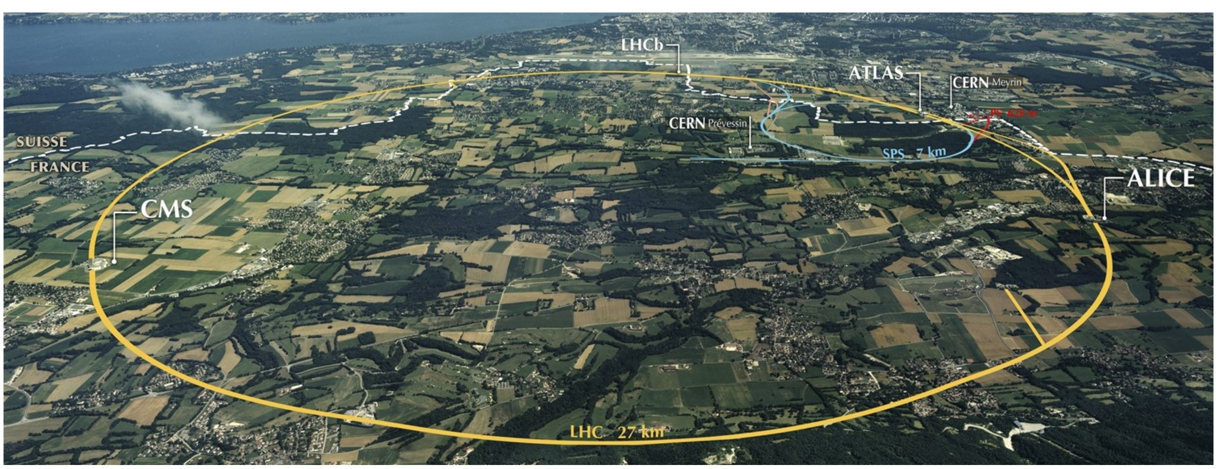


Figure 1: An illustration of the Large Hadron Collider (LHC), drawn on top of the area where the LHC is located underground, showing the four major experiments: ATLAS, CMS, ALICE and LHCb [1].

different particles, including the for this thesis relevant particles: B_c^+ and B^+ mesons. The B_c^+ and B^+ are short-lived particles consisting of two quarks, an anti-beauty quark and respectively a charm and an up quark. The decay modes that will be focused on in this thesis, called the signal decays, are $B_c^+ \to \tau(\to \pi^+\pi^+\pi^-\overline{\nu}_\tau)\nu_\tau$ and $B^+ \to \tau(\to \pi^+\pi^+\pi^-\overline{\nu}_\tau)\nu_\tau$. By measuring their branching fractions, i.e. the fraction of particles that decay through this particular decay mode, SM predictions can be tested. This could contribute to answering the question of whether LFU is violated.

To be able to distinguish the signal decays from any background decays and to reconstruct the decays, the LHCb consists of different detectors that measure the characteristics of the particles. One of the detectors that is part of the LHCb experiment is a vertex locator (VELO). It consists of 52 modules and is able to measure the locations of the charged particles that hit the different modules of the detector. The special feature of the VELO is its proximity to the proton-proton collision point. This makes it possible for short-lived particles, such as the B_c^+ and B^+ to reach the detector. Measuring the location of the B_c^+ and B^+ particles helps identify the type of decay. However, since the VELO cannot distinguish between charged particles, the hits in the VELO that originate from the signal decay can either be from the $B_{(c)}^+$ or the τ particle.

This thesis aims at predicting the ratio of $B_{(c)}^+$ to τ first hits in the VELO, using data from the simulation package RapidSim. Additionally, a Multivariate Analysis (MVA) will be performed to study whether it is possible to predict the origin ($B_{(c)}^+$ or τ) of the first hits in the VELO based on observed parameters from the detectors in LHCb.

Being able to predict the origin of the first hits in the VELO could help separate the B_c^+ and B^+ signal decays from the background decays. This is critical for measuring the branching fractions of these decays.

2 Theory

2.1 The Standard Model

The Standard Model of particle physics (SM) describes very accurately how the elementary particles interact. However, there are several phenomena that the SM cannot explain, such as the asymmetry between matter and anti-matter and how the fundamental force of gravity would fit into the theory. Therefore, (potential) deviations from SM predictions are thoroughly researched, such as LFU violation. To understand what the SM entails, a brief explanation of the SM will be given.

Particle physics phenomena are explained in the SM by the properties and interactions of the elementary particles, which are particles that are not composed of other particles and are considered point-like [2]. These elementary particles are divided into groups, conveniently summarized in Figure 2. Particles with half-integer spin, fermions, are divided into two groups: leptons and quarks. The group of integer spin particles, bosons, consist of gauge bosons and the Higgs boson. The fermions are the building blocks for all matter in the Universe, while the bosons are the particles that 'carry' forces.

Both the quark and the leptons consist of three generations made out of 2 particles. For the leptons, each generation consists of a charged lepton and a neutral lepton of the same flavour. The charged leptons come in three flavours: electron (e^-), muon (μ) and tauon (τ). The neutral leptons - neutrinos - come in the same three flavours. Neutrinos only interact through the weak force and are because of this very difficult to detect, while charged leptons can also interact electromagnetically. The group of quarks consists of six different flavours, up (u), down (d), charm (c), strange (s), top (t) and bottom (b), also known as the beauty quark. Quarks do not appear on their own, but will always be found together with other quarks to form a 'neutral colour' combination.

The gauge bosons are force carriers, making it possible for the fermions to interact with

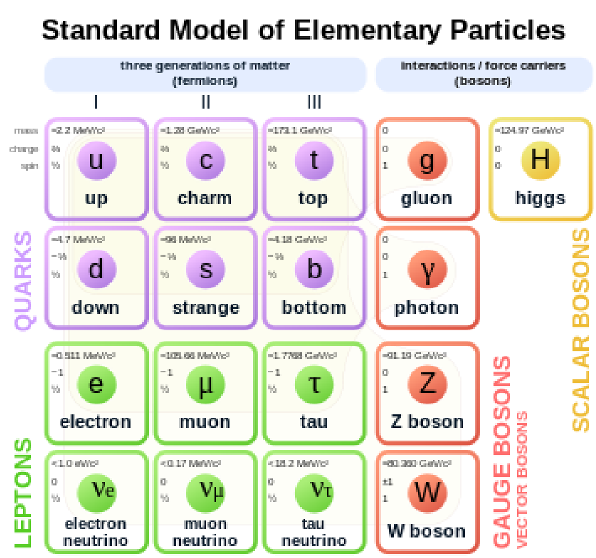


Figure 2: The Standard Model, showing all elementary particles grouped into bosons (force carriers) and three generations of fermions (matter) [6].

each other. The massless neutral photons are the carriers of the electromagnetic force. All particles that have an electric charge can interact electromagnetically, which makes it easy to detect charged particles. The strong force is mediated by gluons, which are also massless neutral particles. The weak force is mediated by the W and Z bosons, which do have mass.

2.2 $B^+_{(c)}$ decay

This thesis focuses on the decay of the B_c^+ and the B^+ mesons. Mesons are particles that consist of a quark and an anti-quark. The B_c^+ and B^+ particles both contain an anti-beauty quark ($\bar{b}$) and respectively a charm quark (c) and an up quark (u). The mean lifetime of the B_c^+ is very short: $0.510 \pm 0.009 \times 10^{-12}$s [7]. The mass of B_c^+ is $6274.47 \pm 0.27 \pm 0.17$ MeV, where the first error is a statistical error and the second a systematic error [8]. The lifetime of the B^+ is approximately three times longer: $1.638 \pm 0.004 \times 10^{-12}$ s and the mass is: 5279.25 ± 0.26 MeV [7].

The decay modes that are used in the analysis of this thesis are the following: $B_c^+ \rightarrow \tau(\rightarrow \pi^+\pi^+\pi^-\overline{\nu}_\tau) + \nu_\tau$ and $B^+ \rightarrow \tau(\rightarrow \pi^+\pi^+\pi^-\overline{\nu}_\tau) + \nu_\tau$. In both decays, the B-meson decays into a charged lepton (τ) and the corresponding neutral lepton (ν_τ).

The branching fraction of the signal B_c^+ decay mode has not been determined experimentally yet. However, there does exist a SM prediction of the branching fraction, which is [9]:

$$\mathcal{BR}(B_c^+ \rightarrow \tau^+ + \nu_\tau)^{SM} = (1.95 \pm 0.09) \times 10^{-2} \tag{1}$$

This prediction of the SM can be tested when physicists succeed at measuring the B_c^+ signal decay at the LHCb.

The branching fraction of the $B^+ \rightarrow \tau + \nu_\tau$ has been measured already [7]:

$$\mathcal{BR}(B^+ \rightarrow \tau + \bar{\nu_\tau}) = (1.09 \pm 0.24) \times 10^{-4} \tag{2}$$

However, this measurement has a large uncertainty and could potentially be improved.

In both decays, the decay mode is chosen in which the τ decays into three pions and a neutrino: $\tau \rightarrow \pi^+\pi^+\pi^-\nu_\tau$. Neutrinos do not interact electromagnetically, which makes it very difficult to detect them; at the LHCb experiment, it is not possible to directly detect neutrinos. This makes it more difficult to measure the decay. Additionally, the $B^+_{(c)}$ and τ particles are also difficult to measure because of the lifetime of the particles. Similar to the B-mesons, the τ is a very unstable particle, with a mean lifetime of $(0.2903 \pm 0.00050 \times 10^{-12}$ s [7]. This means that both the $B^+_{(c)}$ particle and the τ particles decay shortly after they are created. Therefore, in order to reconstruct the decay, these particles need to be detected shortly after they are created. Additionally, the measurements of the three pions (which are the decay products of the τ particle) help identify the decay.

It is necessary to measure different properties of the signal decay, because at the LHCb, not only $B^+_{(c)}$ are produced in the proton-proton collisions. There are many different particles measured by the detectors of the LHCb. These background decays need to be distinguished from the signal decay when you want to measure the branching fraction of the signal decay. This is difficult when the observed properties of the background decay are similar to that of the signal decay, e.g. when three pions are produced that have similar momenta as the pions created in the signal decay. Background decays that do not contain three pions can likely be separated with the data from the detectors at the LHCb experiment. Thus, the more dangerous backgrounds are those containing three pions in their final state.

The most dangerous background decays are the one where a B^+ decays directly into three pions and an additional particle, such as the D-meson [10]. This decay is hard to distinguish from the signal decays since the pions will likely have similar momenta as the pions from the signal decay. Another (less) dangerous background decay is when a B^+ decays into a $\tau\nu_\tau$ and an additional particle. When in the analysis of the data the reconstruction of the additional particle fails, the decay can be incorrectly identified as the signal decay. Furthermore, the B^+ can include a D-meson, which in turn decays into $\tau(\rightarrow \pi^+\pi^+\pi^-\overline{\nu}_\tau)\nu_\tau$. This decay can be mistaken for the signal decay, but the final pions are likely to have less energy compared to the signal decay since part of the energy is transferred to the D-mesons mass. Lastly, the decay of a D-meson into $\tau(\rightarrow \pi^+\pi^+\pi^-\overline{\nu}_\tau)\nu_\tau$ can also be a dangerous background decay because of the presence of three pions, but again the momenta of the pions are likely to be different from pions in the signal decay.

3 LHCb experiment

3.1 LHC

The LHCb is an experiment at CERN, the European Organization for Nuclear Research. It is one of the major experiments in the LHC, which is the most powerful and largest particle collider, having a circumference of 27 kilometres [11]. Before particle beams enter the LHC, they pass multiple accelerators to gain speed. At the LHC particle beams of protons can reach energies up to 6.8 TeV [11]. Two particle beams are accelerated in opposite directions, such that they can be collided at different experiments. In the LHC primarily protons are accelerated, but ions such as lead ions can also be accelerated.

3.2 Setup

The goal of the LHCb experiment is to test SM predictions and possibly discover new physics. This is done by researching the decays of hadrons consisting of a beauty or charm quark. In 2020 the LHCb experiment was upgraded, improving most of the detectors [13]. The upgrade makes it possible to acquire larger data sets, which makes it possible to test the SM more precisely.

The LHCb experiment consists of multiple detectors, tracking the particles and measuring different properties of the particles. The detectors are spatially oriented in such a way that they mainly detect forward-going particles such as the B-mesons [13]. The angle (θ) relative to the beam axis that the detectors cover is expressed in a pseudorapidity range:

$$\eta \equiv -\ln[\tan(\theta/2)] \tag{3}$$

The pseudorapidity range of the LHCb is $2 < \eta < 5$, which means that only particles that travel within this range will be detected.

Figure 3 shows all of the detectors of the LHCb experiment. The tracking system of the LHCb consists of the Vertex Locator (VELO) that is placed around the interaction region, the upstream tracker (UT) placed before the magnet and the scintillating fibre tracker (SciFi Tracker) placed

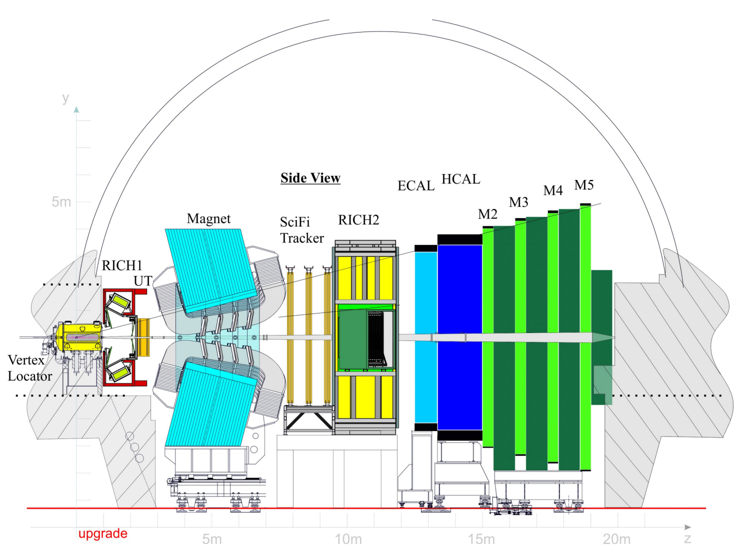


Figure 3: The setup of the upgraded LHCb experiment, including the vertex locator [12].

after the magnet [13]. The tracking system can only measure charged particles. The VELO consists of multiple modules of pixel silicon detectors. This detector will be discussed in more detail in section 3.3. The UT is made out of silicon strips and its goal is to provide tracking information that can be combined with the VELO hits. The SciFi Tracker consists of three scintillating planes that are made out of fibres. Since the SciFi Tracker is located downstream of the magnet, its measurements can be combined with the data from the trackers upstream of the magnet to provide information on the momentum of the particles by measuring the curvature of the particles. Additionally, there are detectors to identify the type of the particles that are measured. These are two Cherenkov detectors (RICH1 and RICH2), two types of calorimeters (ECAL and HCAL) and four muon chambers.

3.3 Vertex Locator

The VELO consists of 52 modules of silicon detectors with square pixels. The modules are L-shaped, as can be seen in Figure 4 which shows half of the modules. These modules are placed around the beam axis to form a square as shown in the right picture of Figure 5. The inner radius is decreased to 5.1 mm after the upgrade of the LHCb experiment. A smaller inner radius increases the chance of $B^+_{(c)}$ particles to reach the VELO since these particles have a short lifetime and are otherwise very likely to decay before reaching the VELO.

The modules are distributed along the z-axis according to the schematic shown in Figure 5. Most of the modules have a separation of 25 mm, except for 16 modules that are mostly further away from the proton-proton (pp) collision point. The two halves of the VELO have a relative shift of 12.5 mm on the z-axis from each other to make it possible for the modules to overlap. This is to detect as many particles as possible.

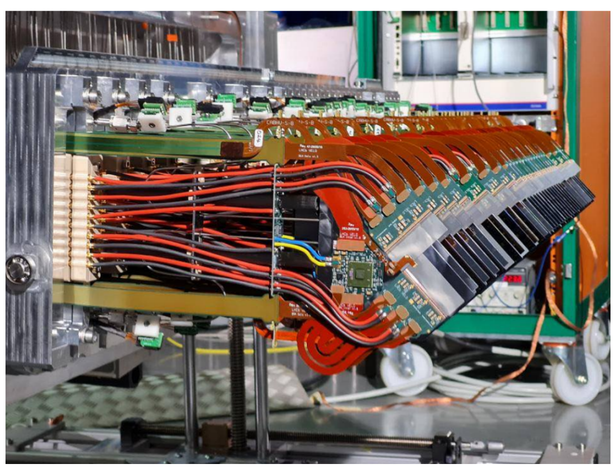


Figure 4: A picture of half of the upgraded VELO modules [13]. Combined with the other half, the modules can enclose the particle beam, with the minimum distance between the detector and beam axis being 5.1 mm.

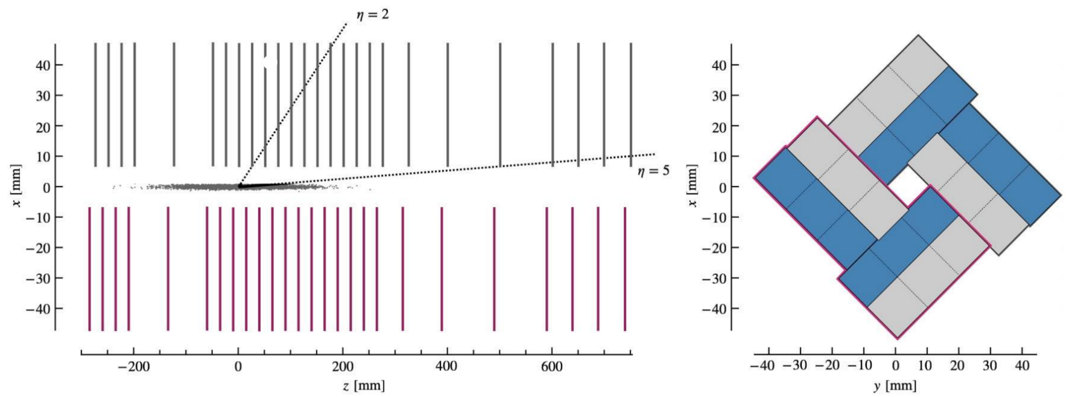


Figure 5: A schematic of the front and side view of the VELO modules, showing the pseudorapidity acceptance range of the LHCb and the interaction region [13].

4 Analysis Strategy

During the pp collisions at the LHCb experiment, many particles are created of which only a few of them are the B-mesons that are studied in this research. To measure the branching fraction of the B_c^+ and B^+ signal decays, it needs to be possible to distinguish these decays from background decays. To make it easier to separate signal from background decays, the data is filtered to reduce the amount of data that needs to be analysed. Specifically, for the $B_c^+ \rightarrow \tau + \nu_\tau$ decay, the data is filtered based on the presence of a VELO hit [10]. Only events where there is at least one VELO hit are saved for further analysis. Another selection requirement is the presence of three pions.

In this analysis, instead of experimental data from the LHCb experiment, simulated data will be used. The following sections will explain how the data is simulated, the filtering process and how the first hits in the VELO are defined.

4.1 RapidSim

The package that will be used for simulating the data is RapidSim, which can simulate millions of signal and background decays in seconds [14]. RapidSim generates data much faster than other simulations because it does not provide a full-detector simulation. Instead of simulating the underlying event of a decay (the pp collision) and then providing the detector response for that event, it only generates certain parameters for the selected decay, such as the momentum distributions and decay vertices of the decay. For this study, RapidSim approximates the properties of the signal decays accurately enough to estimate the origin of the first hits in the VELO. To account for the resolution of the detectors, a 'smear' is added to the momentum of the particles. This smear makes the momentum of the simulated particles less precise.

4.2 Observables

RapidSim can generate certain parameters for the chosen signal decays. The parameters that are used in the analysis will be explained in the following paragraph.

The parameters that can be inferred experimentally are shown in bold in Figure 6. The primary vertex (PV) is the vertex where the $B_{(c)}^+$ originates from. This is at the point of the pp collision. The PV can be determined in the LHCb experiment by analysing the tracks of the particles that are created during the collision. The secondary vertex (SV) where the $B_{(c)}^+$ decays into a τ and a ν_τ cannot be reconstructed with experimental data. This is because there is no information on the direction of the neutrino since neutrinos do not interact with the detectors. For the same reason the angle between the ν_τ and τ^+ (θ_{true}) can also not be determined. θ_{true} can be approximated by the angle (θ_{corr}) between the direction of the PV to the first hit in the VELO and the combined direction of the three pions. Section 4.3 will explain how this VELO hit is generated. The tertiary vertex (TV) is where the τ decays into three pions and an anti-neutrino. This vertex can be reconstructed by the tracks of the three pions. Once the PV and the TV are reconstructed, the flight distance (FD) and transverse flight distance ($\text{FD}_\perp$) can be calculated, see Figure 6. Additionally, the impact parameter can be determined, which is the shortest distance between the PV and the combined direction of the 3π-system. Furthermore, an important parameter that helps identify the parent particle of the decay (in this case the B_c^+ or B^+) is the corrected mass. Commonly, the invariant mass is used to identify the parent particle of the decay ($B_{(c)}^+$ for the signal decays). However, for the signal decays information about the invariant mass of the B_c^+ and B^+ is lost because the momentum transferred to the neutrinos in the decay cannot be reconstructed. Therefore, the corrected mass is defined to retrieve some of this lost information which helps identify the parent particle. The corrected mass is defined as:

$$m_{corr} = \sqrt{m_{3\pi}^2 + |\vec{p}_\perp(3\pi)|^2} + |\vec{p}_\perp(3\pi)| \quad (4)$$

where $m_{3\pi}$ is the invariant mass: $m_{3\pi} = \sqrt{E_{3\pi}^2 + \vec{p_{3\pi}}^2}$ and $\vec{p}_\perp(3\pi)$ the momentum of the 3π-system perpendicular to the flight direction of the $B_{(c)}^+$, which is approximated by the direction between the PV and the first hit in the VELO.

4.3 VELO hits

The generated data by RapidSim needs to be filtered based on the presence of a VELO hit. However, RapidSim does not simulate the detector response for the simulated decay. Therefore, to generate the location and amount of VELO hits a Python script is implemented. The script also filters the simulated data based on the presence of a VELO hit. This script is written by Maria Dominica Galati. The classification of first hits ($B_{(c)}$ or τ) is added to the Python script in this research.

When calculating the amount and location of the VELO hits, the script uses a simplified version of the geometry of the VELO. In the simplified model, the VELO modules are distributed evenly along the positive z-axis, with a separation of 25 mm. This differs from the actual setup (see Figure 5) in multiple ways. In the actual setup, the modules are not distributed evenly; most modules have a 25 mm separation, but the modules further from the collision point have a larger separation. The modules further from the pp collision point are less likely to register the first hits in the VELO. Therefore, the effect of this difference in the setup on the location of the

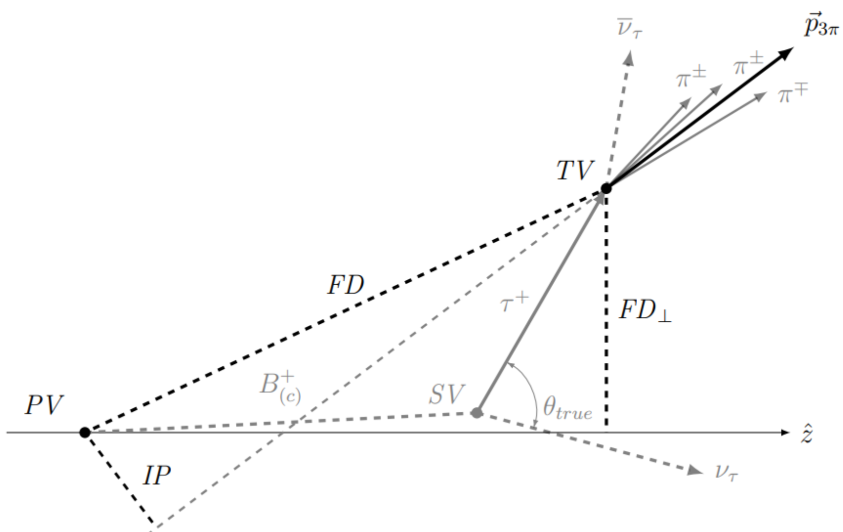
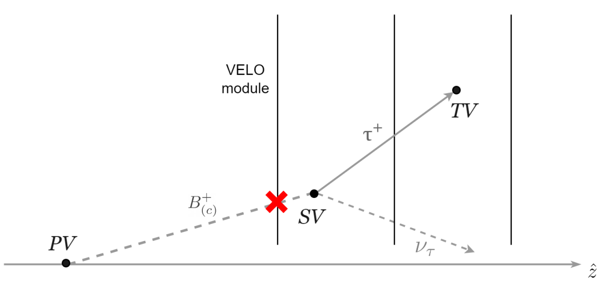


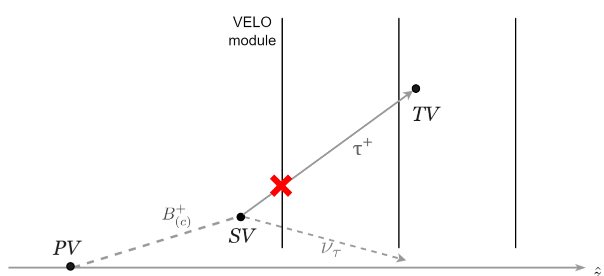
Figure 6: A schematic showing the parameters of the $B_c^+ \to \tau(\to \pi^+\pi^+\pi^-\overline{\nu}_{\tau^+}) + \nu_\tau$ decay. Indicating the observables (PV, TV, $\vec{p_{3\pi}}$, FD, $FD_\perp$ and IP) that can be inferred experimentally in bold. θ_{true} and the SV cannot be inferred.

first hits is likely to be small. Additionally, it is not taken into account that the z-position of the modules on the positive x-axis is shifted by 12.5 mm with respect to the modules on the negative x-axis. This is also not expected to affect the results of the location of the first hit much since the PV is randomly distributed between the modules, and thus a shift in the z-direction of the modules averages out. Furthermore, the length of the VELO is modelled to be 1 m, which means the model only accounts for 41 modules instead of the 52 modules that are present in the actual VELO. However, a small part of these 52 modules are placed upstream of the pp collision point, and only the downstream modules are considered for the hits of the signal decays. Thus, the implementation of 41 modules in the model is a reasonable approximation of the actual number of VELO modules that are considered when reconstructing the signal decay.

The VELO hits are generated by first considering the distance between the PV and the SV. The z-position of the PV is randomly generated to be between two modules, to account for the fact that the actual pp collision point can also be anywhere between two modules. Using the simplified geometry of the VELO, it is calculated how many modules lie between these two vertices. Then the same process is repeated for the distance between the SV and the TV. To determine the location of the first hit in the VELO, the location of the first cross-section between a module and the PV-SV direction or SV-TV direction is calculated. The origin of the first hit is determined based on the location of the first hit on the z-axis with respect to the SV. Figure 7 shows the two cases in which a first hit is either defined as a first hit originating from a $B_{(c)}$ particle or a τ particle. When a $B_{(c)}$ particle traversing from the PV to the SV encounters a VELO module, the first hit is associated with a $B_{(c)}$ particle. In this case, the z-coordinate of the first hit has a lower value than the z-coordinate of the SV. While if the $B_{(c)}$ particle already decays before reaching a VELO module but the τ particle does reach a module, the first hit originates from the τ particle. When neither of the two types of particles reaches the VELO, the corresponding event is filtered out by the Python script. To simplify the analysis, the analysis is only based on the z-coordinate of the first hit and SV. This method could fail if the *tau* particle is emitted in the negative z-direction. However, this is unlikely to happen since most of the particles in the reconstructed signal decays travel in the positive z-direction since the $B_{(c)}$ particles are mostly produced at small angles with respect to the beam axis.



(a) First hit associated with the $B^+_{(c)}$



(b) First hit associated with the τ

Figure 7: A schematic of the two cases in which a first hit in the VELO originates from either $B^+_{(c)}$ particle or a τ particle.

4.4 Multivariate Analysis

The method described in the previous section uses the location of the SV with respect to the location of the first hit to determine the origin of the first hit. However, the location of the SV cannot be reconstructed with observed parameters. Therefore, a multivariate analysis (MVA) will be implemented to predict the $B^+_{(c)}$ t0 τ ratio of first hits in the VELO based on only the observables as described in section 4.2. The MVA is a statistical method that uses the data of multiple parameters to predict the origin of the first hits [9].

Being able to predict the origin of the first hits can make it easier to separate the signal decay from the background decays because the relevant background decays have different parent particles (e.g. D^+ or B^0), which results in a different parent particle-to-τ ratio of first hits. Thus, the prediction of the origin of the first hits adds a variable that can be used to distinguish the signal decay from background decays. Additionally, certain parameters that are important in identifying the decay are dependent on the direction between the PV and the first hit. When it is possible to identify the $B^+_{(c)}$ first hits, for example, the distribution of the corrected mass can more correctly approximate the invariant mass of the B^+_c particle.

Since the first hits are already classified according to the method in the previous section, a supervised learning method can be used. The algorithm that will be used for the MVA is the gradient boosting classifier. Specifically, the package from the scikit learn library is used [15]. This algorithm classifies the hits by implementing boosted decision trees.

The data that is generated by RapidSim for the signal decays of B^+_c and B^+ will be split into 50% train data and 50% test data. The train and test data consists of already classified first hits. The gradient boosting classifier algorithm will be trained with the train data and with the test data the performance of the algorithm can be checked. This is to check whether the algorithm not only works for the specific data set it is trained on but also on other data.

The MVA will not predict the class of every hit correctly; an event that is incorrectly labeled as a $B^+_{(c)}$ hit is called a false positive (FP) and when the event is incorrectly labeled as a τ hit it is called a false negative (FN) [16]. A correctly classified event is called a true positive (TP) in the case of a $B^+_{(c)}$ hit or a true negative (TN) in the case of a τ hit. The fraction of events that is correctly identified as a $B^+_{(c)}$ hit is called the true positive rate (TPR) and the fraction that is incorrectly identified as a $B^+_{(c)}$ hit is called the false positive rate (FPR). These fractions will be used to test the performance of the MVA by plotting the TPR against the FPR. The resulting curve is called the Receiver Operating Characteristics (ROC) curve. The area under the ROC curve (AUC) quantifies the performance of the MVA. The AUC-score ranges from zero to one, where one means that all events are classified correctly and zero that none of the events are correctly classified.

5 Results

In this chapter the analysis of the first hits in the VELO using data from the simulation package RapidSim will be discussed. Firstly, the results will be presented in which the origin of the first hits (either $B^+_{(c)}$ or τ) is determined based on the location of the first hit with respect to the decay vertex of the $B^+_{(c)}$ particle. Secondly, it will be discussed how successful the gradient boosting classifier algorithm is in predicting the origin of the first hits.

5.1 $B^+_{(c)}$ to τ ratio of first VELO hits

The data generated by RapidSim is analyzed according to the method described in section 4.3. The classification of first VELO hits for the $B_c^+ \to \tau(\to \pi^+\pi^+\pi^-\overline{\nu}_\tau)\nu_\tau$ decay results in 90.6% of the first hits associated with the τ and 9.4% of the first hits associated with the B_c^+, as shown in Table 1. Thus, for the B_c^+ signal decay in the case where there is at least one VELO hit, more than 90% of the B_c^+ particles decay before reaching a VELO module. This affects the distributions of the observables that are based on the location of the first hit. It was assumed that the direction between the PV and the first hit could approximate the B_c^+ direction, but when most of the first hits do not originate from the B_c^+, this approximation might not hold. The distributions of the observables are discussed in more detail in the next section.

For the $B^+ \to \tau(\to \pi^+\pi^+\pi^-\overline{\nu}_\tau)\nu_\tau$ decay 69.1% of the hits originate from B^+ particles and 30.9% of τ particles. Compared to the B_c^+ signal decay, relatively more B^+ particles reach the VELO than B_c^+ particles. This can be explained by the lifetime of the B^+ and B_c^+ particles; the lifetime of the B^+ is approximately three times longer than the lifetime of the B_c^+ particle [7]. Therefore, the B_c^+ particles are more likely to decay before reaching a VELO module than the B^+ particles.

Decay	Number of events	τ first hits	$B^+_{(c)}$ first hits
$B_c^+ \to \tau\nu_\tau$	405770	90.6%	9.4%
$B^+ \to \tau\nu_\tau$	2079369	69.1%	30.9%

Table 1: A table with the total number of decays with at least one VELO hit generated by RapidSim for the $B^+_{(c)}$ signal decays and the percentages of first hits in the VELO that originate from τ and $B^+_{(c)}$ particles.

5.2 Distributions of observables

With the classification of the first hits of the VELO, the data of the observables can also be split into two categories: events corresponding with a $B^+_{(c)}$ first hit in the VELO and events corresponding with a τ first hit. For certain observables the separation into the two categories results in distinctively different distributions. Beginning with the signal decay of the B_c^+, one of these observables is the corrected mass. Figure 8 shows that for the corrected mass the events corresponding with first hits originating from τ particles show a peak at a mass around 1.6 GeV/c^2. This is close to the invariant mass of the τ, which is 1.78 GeV/c^2. This can be explained by the dependence of the corrected mass on the direction from the PV to the first hit. This is supposed to approximate the direction of the B_c^+ particle, but in the case of a first hit originating from a τ particle, the direction between the PV and the first hit is likely to approximate the τ

direction better than the B_c^+ direction. Therefore, the distribution corresponding with a τ first hit peaks at the invariant mass of the τ particle. The events corresponding with a B_c^+ first hit do not show such a peak, but compared to the events associated with a τ first hit, the distribution is shifted towards higher values of the corrected mass, which is in accordance with the invariant mass of B_c^+ around 6.27 GeV/c^2.

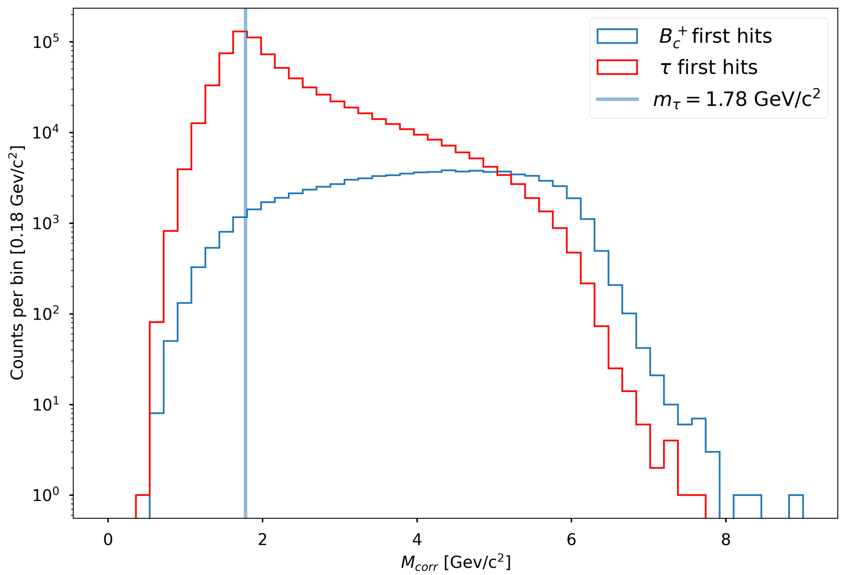


Figure 8: The counts per bin for the corrected mass of the $B_c^+ \to \tau(\to \pi^+\pi^+\pi^-\overline{\nu}_\tau)\nu_\tau$ decay, with the data separated according to whether an event corresponds to a B_c^+ or a τ first hit in the VELO. The vertical line shows the invariant mass of the τ particle [7]. The bin width is 0.16 GeV/c^2.

Additionally, the distribution of the impact parameter shows a clear distinction between the two categories (B_c^+ or τ first hits in the VELO) as well, as shown in Figure 9. The distribution of the IP is concentrated at small values for the events associated with τ first hits. A small IP value indicates that there is almost no difference between the PV-TV direction and the direction of the 3π system. A distribution that is concentrated around smaller IP values is more likely when the combined lifetime of the B_c^+ and τ is short. Since as a result of a short combined lifetime, the particles travel a shorter distance and thus have on average a smaller IP value. For a longer combined lifetime, the IP's distribution would be more spread out. This corresponds with the results since for a τ first hit the B_c^+ particle has decayed before reaching the VELO (unlike for a B_c^+ hit), which indicates a shorter lifetime of the B_c^+ and thus a shorter combined lifetime.

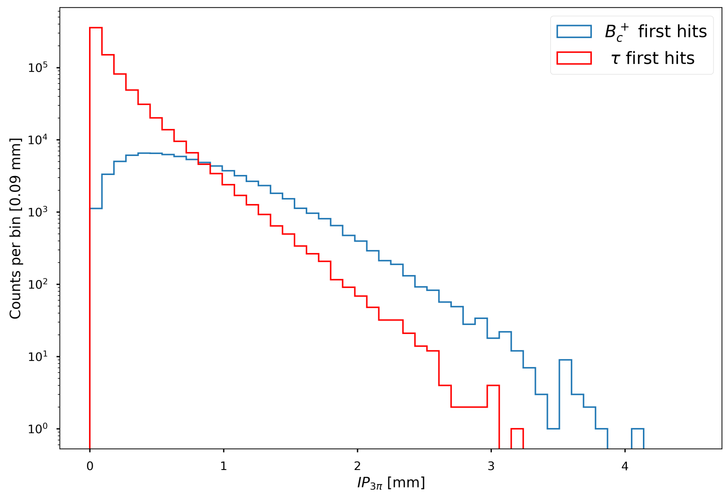


Figure 9: The counts per bin for the impact parameter of the $B_c^+ \to \tau(\to \pi^+\pi^+\pi^-\nu_\tau)\overline{\nu}_\tau$ decay with the data separated according to whether an event corresponds to a B_c^+ or a τ first hit in the VELO. The bin width is 0.24 mm.

The transverse flight distance is one of the other observables that shows a difference between the two distributions. Figure 10 shows the distribution of the $FD_\perp$ separated into two categories. In contrast to the other figures, the counts per bin are normalized and the y-axis has a linear scale instead of a logarithmic scale to highlight the differences between the two distributions. The distribution that corresponds with first hits originating from B_c^+ particles is slightly shifted towards higher values of the $FD_\perp$. The $FD_\perp$ is the transverse distance between the PV and the TV. This is expected to be higher when the combined lifetime of the B_c^+ and τ particles is longer. The combined lifetime is likely longer for events corresponding with a B_c^+ first hit since the B_c^+ has not decayed yet before reaching the VELO and thus likely had a longer lifetime than for B_c^+ particles that do decay before reaching the VELO.

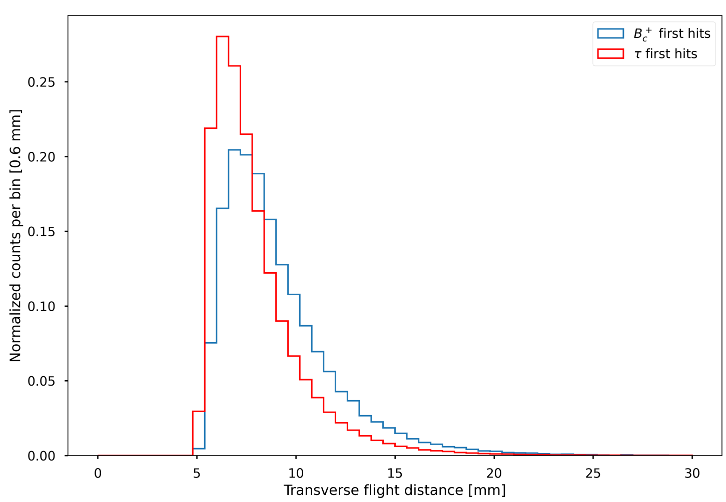


Figure 10: The normalized counts per bin versus the transverse flight distance in mm for the $B_c^+ \to \tau(\to \pi^+\pi^+\pi^-\nu_\tau)\overline{\nu}_\tau$ decay, with the data separated according to whether an event corresponds to a B_c^+ or a τ first hit in the VELO. The bin width is 0.6 mm.

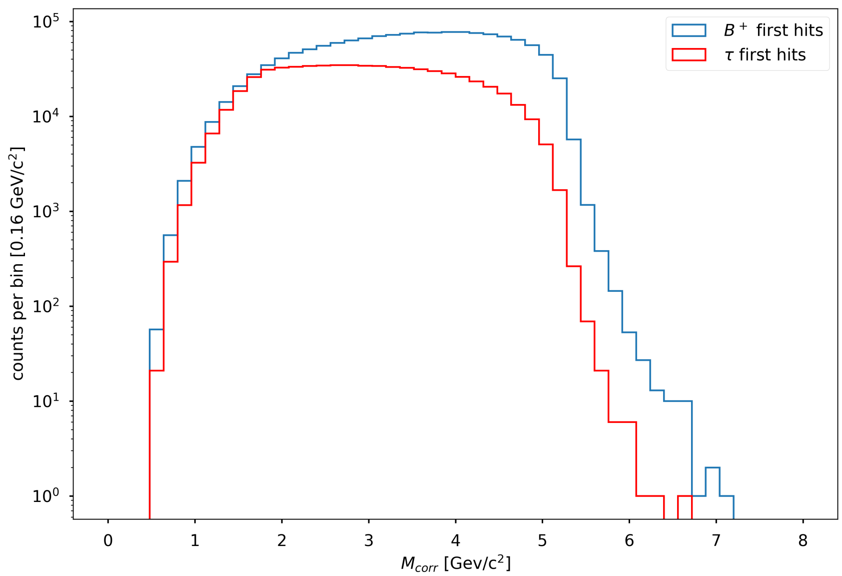


Figure 11: The counts per bin for the corrected mass in the $B^+ \to \tau(\to \pi^+\pi^+\pi^-\overline{\nu}_\tau)\nu_\tau$ decay, with the data separated according to whether an event corresponds to a B^+ or a τ first hit in the VELO. The bin width is 0.16 GeV/c^2.

Similarly, in the case of the B^+ signal decay, the distributions of the observables are separated based on whether the events correspond to a first hit originating from a B^+ or a τ. This separation resulted in distributions that were more similar to each other than for the signal decay of B_c^+. For example, in Figure 11, which shows the corrected mass, there is no peak visible for the distribution associated with first hits originating from τ particles. The distributions do differ slightly, in which the distribution associated with B^+ first hits is shifted more towards higher values of the corrected mass and the distribution associated with τ first hits is shifted towards lower values. This is in accordance with the different invariant masses of the B^+ and τ, which are approximately 5.3 GeV/c^2 and 1.8 GeV/c^2[7].

The distributions of the IP do show a similar difference between the two categories as for the B_c^+ signal decay. Similarly to the distribution of the IP for the B_c^+ signal decay, the distribution corresponding to τ first hits is more concentrated around the lower values, while the distribution corresponding to B^+ first hits is more spread out.

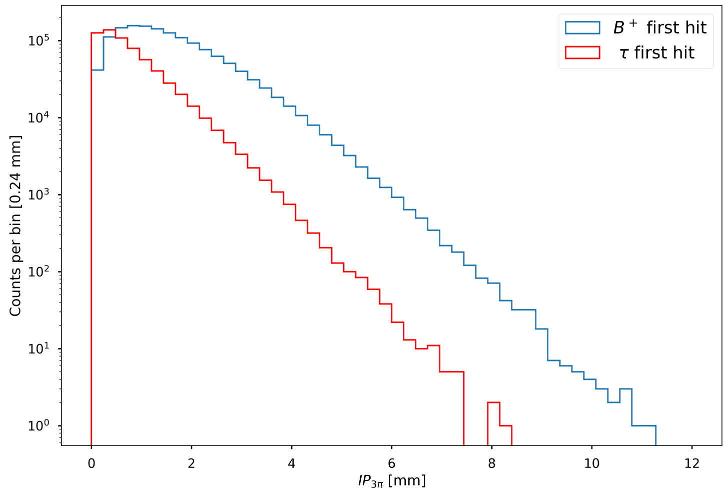


Figure 12: The counts per bin for the impact parameter of the $B^+ \to \tau(\to \pi^+\pi^+\pi^-\nu_\tau)\overline{\nu}_\tau$ decay, with the data separated according to whether an event corresponds to a B^+ or a τ first hit in the VELO. The bin width is 0.24 mm.

5.3 MVA

An MVA is implemented to determine if it is possible to predict the origin of the first hits of the VELO based on only the observables. For this, the same data set is used as in the previous section. Figures 8, 9 and 10 show that for the signal decay of B_c^+ there exist observables which can be used to distinguish events with a B_c^+ first hit from events with a τ first hit. The MVA is trained with these observables (m_{corr}, $IP_{3\pi}$ and $FD_\perp$) together with the following observables: $p_{3\pi}$, $p_{3\pi}^\perp$, FD, θ_{corr}, n_{hits}, m_{miss}^2 and $m_{3\pi}$.

The MVA gives each event a score between 0 and 1, where the closer the score is to 0 the more probable it is that the event corresponds with a τ first hit, while a score closer to 1 means that the event is more probable to correspond with a $B_{(c)}^+$ first hit. In the MVA, all events that have a score higher than 0.5 are predicted to be a $B_{(c)}^+$ hit and all events with a score lower than 0.5 are predicted to be a τ hit.

For the B_c^+ signal decay, this results in 7.5 % of the first hits being predicted to originate from B_c^+ particles and 92.5 % of the hits to originate from τ particles. This deviates by approximately 2 percentage points from the results in the previous section (Table 1). The results for the B^+ signal decay are that 71.2 % of the hits are predicted to originate from B^+ particles and 28.8% from τ particles. This result differs from the results in the previous section by 2 percentage points as well.

The prediction of the MVA results in a similar $B_{(c)}^+$-to-τ ratio of first hits as the results from the previous section, which already indicates that the MVA can predict the origin of the first hits. To quantify the performance of the MVA, the ROC curves are determined with the prediction scores of the MVA. The area under the ROC curve (AUC) shows how successful the MVA is at predicting the origin of the first hits.

Figure 13 shows for the B_c^+ decay the ROC curves for both the train and the test data. The left plot has a linear scale of the x-axis and the right plot a logarithmic scale. The AUC scores for the train and test data are respectively 0.965 and 0.963. The high AUC score for the test data shows that the MVA is successful in predicting the origin of the first VELO hits. In comparison, the AUC scores of the B^+ signal decay are significantly lower: 0.912 for the train data and 0.911 for the test data, as can be seen in Figure 14. However, these scores do indicate that the MVA can predict the origin of the first hits correctly for most of the events. Furthermore, the AUC-scores of the test and train data are very similar, indicating that the MVA is likely not overtrained.

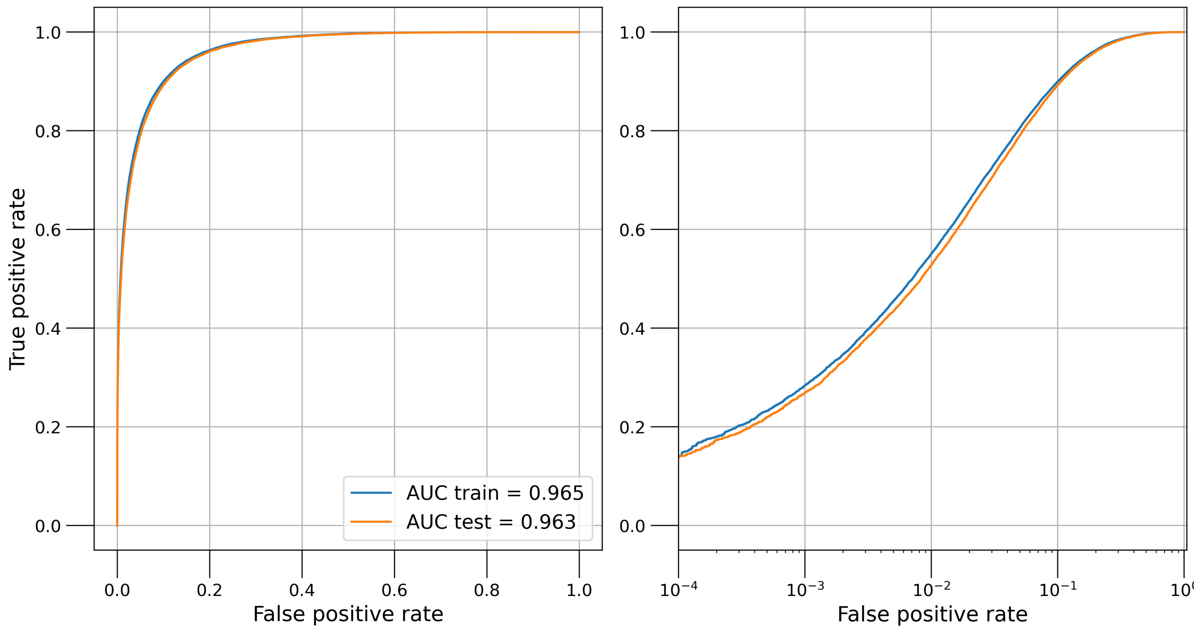


Figure 13: ROC curve for the train and test data of the $B_c^+ \to \tau(\to \pi^+\pi^+\pi^-\overline{\nu}_\tau)\nu_\tau$ decay, where the left figure has a linear scale for the x-axis and the right figure a logarithmic scale.

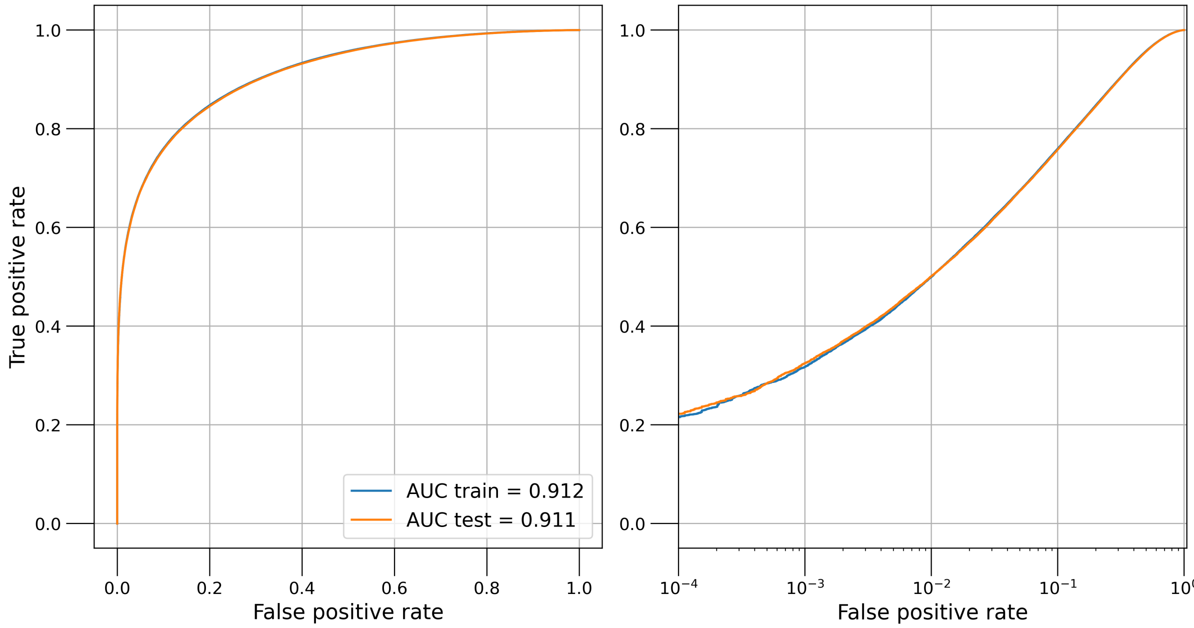


Figure 14: ROC curve for the train and test data of the $B^+ \to \tau(\to \pi^+\pi^+\pi^-\overline{\nu}_\tau)\nu_\tau$ decay, where the left figure has a linear scale for the x-axis and the right figure a logarithmic scale.

6 Conclusion

This thesis aimed at determining the $B^+_{(c)}$-to-τ ratio of first hits in the VELO for the $B^+_{(c)} \to \tau(\to \pi^+\pi^+\pi^-\overline{\nu}_\tau)\nu_\tau$ decay. This ratio was determined with simulated data, both by using the location of the SV, which cannot be determined experimentally, and by using parameters that can be determined experimentally.

Determining the origin of the first hit based on its location with respect to the location of the decay vertex of the B_c^+ particle resulted in 90.6% of first hits to be classified as τ hits and 9.4% as B_c^+ hits. Thus, most B_c^+ particles decay before reaching a VELO module. This was expected based on the short mean lifetime of the B_c^+ particle.

For the B^+ signal decay, 69.1% of the first hits originate from B^+ particles and 30.9% from τ particles. The difference between the B_c^+ and the B^+ decays can be explained by the different lifetimes of the particles. The B^+ particle has a mean lifetime that is approximately three times longer than the B_c^+ particle. This increases the chances of the B^+ particles reaching the VELO before they decay in comparison to the B_c^+ particle. Therefore, this results in a higher percentage of B^+ first hits compared to B_c^+ hits.

To study the effect of the origin of the first hit on the distributions of observables, their distributions were separated based on whether an event corresponds to either a $B^+_{(c)}$ first hit or a τ first hit. For the B_c^+ decay, this resulted in a major difference between the two categories of distributions for multiple observables. For example, for the corrected mass, the distribution that corresponds with τ first hits shows a peak at the invariant mass of the τ particle, while the distribution corresponding with B_c^+ first hits is shifted more towards the invariant mass of the B_c^+ particle. This shows that direction from the PV to the first hit depends on which particle causes the hit; in the case of a τ first hit, it approximates the direction of the τ particle and vice versa for the B_c^+ particle.

For the B^+ signal decay, separating the distributions of the observables into two categories (B^+ first hits and τ first hits) resulted in distributions that were more similar to each other than for the distributions of the B_c^+ decay. For instance, the corrected mass distribution that corresponds with τ first hits did not show a peak at the invariant mass of the τ particle. Only a slight shift towards the invariant mass of the τ particle in comparison to the distribution corresponding to B^+ first hits was visible. Therefore, it was expected that predicting the origin of the first hits based on only observed parameters would be more successful for the B_c^+ decay than for the B^+ decay.

The MVA that was implemented to predict the origin of the first hits performed indeed better for the B_c^+ decay, as shown by the AUC-scores. The AUC-score of the test data for the B_c^+ decay is 0.963 and 0.911 for the B^+ decay. Although the AUC-score of the B_c^+ decay is higher than that of the B^+ decay, both scores are relatively close to a AUC-score of 1. Therefore the MVA is successful in predicting the origin for most of the first hits for both the signal decays.

A limitation of this research is the implementation of a simplified model of the geometry of the VELO. This is not expected to have a major influence on the location of the first hits, but it might influence the number of hits, which is one of the parameters used in the MVA. Implementing a more accurate model of the VELO could improve the accuracy of the location and number of hits.

Furthermore, the motivation for predicting the origin of the first hits is that it could improve the separation of background decays from the signal decay. However, before it can be tested whether predicting the origin of the first hits improves the signal from background separation, the MVA needs to be implemented on all relevant background decays. The analysis of the performance of the MVA for the B_c^+ and B^+ signal decays shows that it differs per decay in how accurately the origin of the first hits can be predicted. This research showed that it is possible to predict the origin of the first hits for the $B_c^+ \to \tau\nu_\tau$ and $B^+ \to \tau\nu_\tau$ decays, but future research would need to show whether this is also the case for the relevant background decays.

7 References

[1] M. Brice, "Aerial View of the CERN taken in 2008.," 2008. [Online]. Available: https://cds.cern.ch/record/1295244.

[2] B. R. Martin and G. Shaw, *Particle physics*. John Wiley & Sons, 2017.

[3] R. Aaij, C. A. Beteta, T. Ackernley, B. Adeva, M. Adinolfi, and H., "Test of lepton universality in beauty-quark decays," *Nature Physics*, vol. 18, pp. 277–282, 3 Mar. 2022, ISSN: 1745-2473. DOI: 10.1038/s41567-021-01478-8.

[4] S. Choudhury, S. Sandilya, K. Trabelsi, *et al.*, "Test of lepton flavor universality and search for lepton flavor violation in b→ k decays," *Journal of high energy physics*, vol. 2021, no. 3, pp. 1–24, 2021.

[5] J. Lees, V. Poireau, V. Tisserand, *et al.*, "Precision measurement of the ratio b (Υ (3 s)→ τ+ τ-)/b (Υ (3 s)→ μ+ μ-)," *Physical Review Letters*, vol. 125, no. 24, p. 241 801, 2020.

[6] MissMJ, *Wikipedia, standard model of elementary particles*. [Online]. Available: https://en.wikipedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg.

[7] R. L. Workman *et al.*, "Review of Particle Physics," *PTEP*, vol. 2022, p. 083C01, 2022. DOI: 10.1093/ptep/ptac097.

[8] R. Aaij, C. Abellán Beteta, T. Ackernley, *et al.*, "Precision measurement of the B_c^+ meson mass," *Journal of High Energy Physics*, vol. 2020, no. 7, pp. 1–21, 2020.

[9] Y. Amhis, M. Hartmann, C. Helsens, D. Hill, and O. Sumensari, "Prospects for $B_c^+ \to \tau+ \nu_\tau$ at fcc-ee," *Journal of High Energy Physics*, vol. 2021, no. 12, pp. 1–35, 2021.

[10] J. R. D. Jong, "Feasibility study of the branching fraction measurements of $B_c^+ \to \tau(\to \pi^+\pi^+\pi^-\nu_\tau)\bar{\nu_\tau}$ and $B^+ \to \tau(\to \pi^+\pi^+\pi^-\nu_\tau)\bar{\nu_\tau}$ at lhcb," 2022, Master Thesis on feasibility of measure the branching fraction. Explains about which backgrounds are most dangerous and how to distinguish signal from background. Has info on how VELO is used to cut down on background and to track B-meson.

[11] CERN, *Facts and figures about the lhc*. [Online]. Available: https://home.cern/resources/faqs/facts-and-figures-about-lhc.

[12] R. Nobrega, A. Barbosa, I. Bediaga, *et al.*, "Lhcb trigger system : Technical design report," Jan. 2003.

[13] R. Aaij, A. S. W. Abdelmotteleb, C. A. Beteta, and e. a. F. Abudinén, "The lhcb upgrade i," 2023. arXiv: 2305.10515 [hep-ex].

[14] G. A. Cowan, D. C. Craik, and M. D. Needham, "RapidSim: an application for the fast simulation of heavy-quark hadron decays," *Comput. Phys. Commun.*, vol. 214, pp. 239–246, 2017. DOI: 10.1016/j.cpc.2017.01.029. arXiv: 1612.07489 [hep-ex].

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] W. J. Krzanowski and D. J. Hand, *ROC curves for continuous data*. Crc Press, 2009.

8 Acknowledgements

I want to thank everyone at the LHCb group for the great time I had during my Bachelor's Research Project.

Specifically, I want to thank my supervisor Kristof de Bruyn for your help and feedback during my project. MD, for providing me with the Phython scripts and for your patience when answering all my questions. Andrej and Daniele for helping me with my computer problems. Mick for the useful suggestions during the weekly meetings. And all the bachelor students at VSI for the support during my thesis.