BACHELOR'S THESIS

Feasibility of $B_c^+ \to \tau^+\nu_\tau$ search without VELO hits at LHCb

Author:
Lucía Peña García
S3637212

First Supervisor:
Mick Mulder
Second Supervisor
Julia Even

Faculty of Science and Engineering
University of Groningen
Groningen, The Netherlands
August 30, 2024

Abstract

Lepton Flavour Universality (LFU) posits that the couplings of all charged leptons to the gauge bosons will be identical when corrected for their deferring masses. An observed deviation from LFU could signal at physics Beyond the Standard Model (BSM). At the LHCb group at the University of Groningen, the possibility of detecting de decay $B_c^+ \to \tau^+\nu_\tau$ and measuring its branching fraction is being studied to probe LFU. However, this decay is particularly challenging to measure due to the presence of two neutrinos in the final state and the absence of a detectable B_c^+ decay vertex. For that reason, this research team has developed a method of examining the $B_c^+ \to \tau^+\nu_\tau$ decay which relies on hits left directly by B_c^+ mesons in the VELO detector for analysis. While this method has been shown to be feasible, it significantly reduces the amount of data available for analysis. This thesis explores the feasibility of an alternative approach which does not rely on VELO hit data, increasing the available data, but reducing the number of observables used to separate between the signal decay and background decays.

The study utilised data obtained with the lightweight simulation program RapidSim to train a Boosted Decision Tree classifier model on several observables from signal and background decays. The trained model was applied to a large dataset that emulated the experimentally expected data with respect to signal and background detection rates at the LHCb experiment. The classifier output was then combined with a mass variable called the corrected mass, which compensates for energy lost to the neutrinos, to create a two-dimensional histogram. To consider the method described in this thesis feasible would have involved identifying a region with a significant signal-to-background ratio in this two-dimensional histogram. However, the study found no such region, rendering this method unsuitable.

Contents

1 Introduction

The Standard Model (SM) currently serves as the primary framework for particle physics. It classifies the fundamental particles that make up matter and explains their properties, and provides a theoretical framework that describes how the electromagnetic, weak, and strong forces operate and how they interact with the fundamental particles [1]. Although it is a very successful framework, it has caveats and limitations. Notably, it does not accommodate for the force of gravity, which is instead described by the theory of general relativity [2].

In addition to the omission of gravity, the SM assumes lepton flavour universality (LFU). This is the concept that all charged leptons (i.e. electrons, muons and taus) are identical in every way except for their mass, implying that their couplings to other particles should not depend on the type, or flavour, of lepton they are. This means that the strength of their interactions with the weak force bosons should be identical when normalised for their mass differences [3]. A breach of LFU would indicate a departure from the SM's predictions, suggesting the existence of physics Beyond the Standard Model (BSM).

Experiments to further study whether LFU holds are currently ongoing at the LHCb experiment; the main method used by LHCb to probe LFU is to measure the branching fractions of each type of b-hadron into final states that contain a lepton pair[1]. Specifically, by measuring the different branching fractions of b-hadrons with decays of the form $\bar{b}q \to \ell^+\nu_\ell$ (where ℓ can be e, μ, τ), one can measure for a specific quark q the branching fraction into the three different lepton pairs in the final state. It is then possible to study whether the mass difference of the charged leptons is the sole reason for varying branching fractions or if LFU is violated. Similarly, for a specific lepton pair final state $\ell^+\nu_\ell$, one can measure the different branching fractions of a b-hadron with a specific quark q.

In order to measure the branching fractions and the properties of different decays, the Vertex Locator (VELO) subdetector of LHCb plays a crucial part. It is used to determine the points where protons collide and create decay products in the LHCb detector. This point is where, among other decay products, the B mesons used in this study are created. The B mesons, however, are not usually detected directly. One of the mesons of interest and the focus of this thesis is the B_c^+ meson, which has not yet been observed in its decay to any final state of the form $\ell + \nu_l$ (like $\tau + \nu_\tau$). To detect this decay and study its branching fraction to test LFU, a method has been devised by the LHCb group at the University of Groningen whereby hits left directly by the B_c^+ in the VELO detector are used to

[1] The final state can contain hadrons alongside the lepton pair.

determine its approximate flight direction. This approach successfully separates the B_c^+ meson from backgrounds in RapidSim simulations by comparing their flight directions [4]. As the likelihood of a B meson reaching the VELO modules before decaying is very low, the amount of data is significantly reduced. This process, known as VELO hit filtering, involves selecting only the simulated decays that left a VELO hit for analysis, while discarding the remaining decays. **The purpose of this thesis will be to determine if the method described is truly necessary to detect $\mathbf{B_c^+}$ decays, by attempting to separate the $\mathbf{B_c^+}$ signal from the existing backgrounds at LHCb without making use of the $\mathbf{B_c^+}$ hit in the VELO detector.** This approach will benefit from much larger datasets, but will be hindered by having less information with which to make a separation between signal and background.

2 The Standard Model

The Standard Model of particle physics is a theory that aims to describe the elementary particles that make up all matter and antimatter in the universe, and describe their interactions via the fundamental forces of nature. It categorises particles into two main groups: half-integer spin particles that make up matter, called fermions, and integer spin particles that mediate the fundamental forces, called bosons.

Matter particles, or fermions, are divided into two separate classes: quarks and leptons. Quarks form composite particles with integer charges, such as mesons, which are made up of a quark-antiquark pair; and baryons, which are made up of three quarks. Leptons have integer charges and can be further separated into charged leptons and neutral leptons, called neutrinos[2]. Fermions are categorised into three generations, each generation contains a pair of quarks and a pair of leptons, as illustrated in Fig. 1. Each successive generation has heavier particles than the previous one; for example, the bottom quark has the same charge and spin as the down quark but is much heavier.

[2] Throughout the remainder of this thesis, it is assumed that matter and antimatter are considered together.

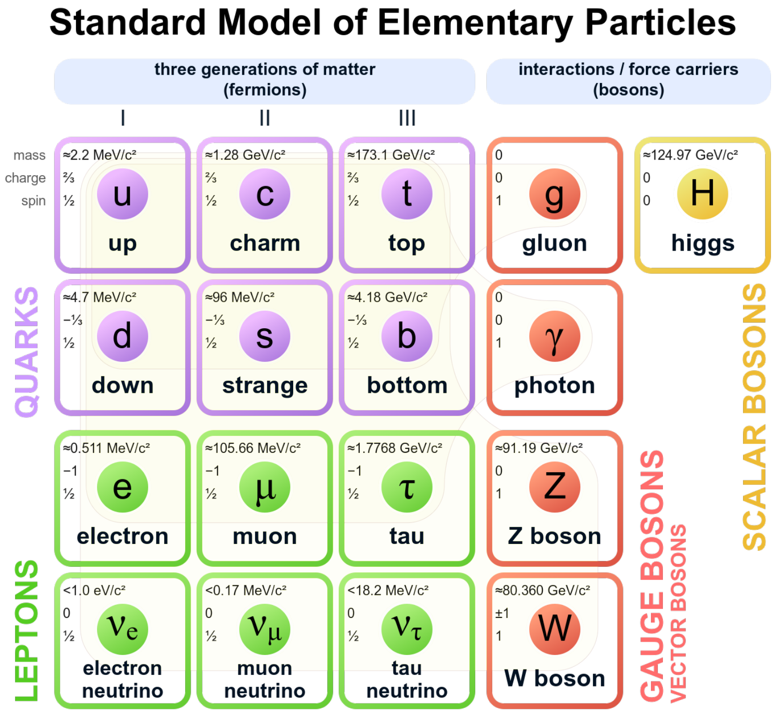

Figure 1: The Standard Model of Particle Physics, taken from [5]

.

Bosons are divided into two classes: particles with spin-0, called scalar bosons, and particles with positive integer spin, called gauge bosons. The only scalar boson in the SM is the Higgs boson, which couples to particles, including bosons, to give them their mass. Gauge bosons are referred to as force carriers, since fermions can interact with one another by exchanging these gauge bosons. The electromagnetic force interacts with charged particles and is mediated by the emission and absorption of photons, which are massless particles with an infinite range. The strong force interacts with quarks, which carry colour charge, and is mediated through the exchange of gluons. Gluons are massless particles like the photon, but carry a colour charge and can therefore also self-interact. The strong interaction can be described by quantum chromodynamics (QCD). The weak force is mediated via the massive $W^{\pm}$ and Z^{0} bosons, which interact with all matter. The strong and weak forces are short-ranged, unlike the electromagnetic force. Gravity and the theoretical graviton are not included in the SM.

The signal decay for this study, namely $B_c^+ \to \tau^+\nu_\tau$, has a meson composed of a bottom and a charm quark, that decays via the weak force into a tau and a tau neutrino, as is seen in Fig 2. In this decay, leptons of the third generation interact with the $W^\pm$ boson and this interaction should produce a specific branching fraction, since, according to the LFU theory, all leptons interact identically with other particles when corrected for their different masses. The purpose of studying the branching fraction of $B_c^+ \to \tau^+\nu_\tau$ is to check if the experimental value corresponds to the value derived from theory or if it deviates significantly from theory, implying physics beyond the SM.

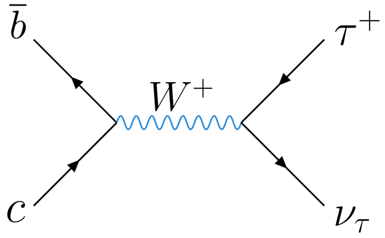

Figure 2: Feynman diagram of the $B_c^+ \to \tau^+\nu_\tau$ decay, adapted from [4].

3 Detection at LHCb

In order to detect $B_c^+ \to \tau^+\nu_\tau$ decays and measure their branching fraction, they must first be produced. The B_c^+ meson is one of the heaviest mesons, composed of the second and third heaviest quarks; this means it requires very high energies to be produced with enough energy to be studied. These energies are found in the Large Hadron Collider (LHC) at CERN, a proton-proton collider designed to reach a Center of Mass (CoM) energy of 14 TeV [6]. One of the several experiments at CERN is the Large Hadron Collider beauty experiment, or LHCb, which is designed to search for BSM physics by studying the decays of beauty hadrons, which are hadrons which contain a beauty/bottom quark.

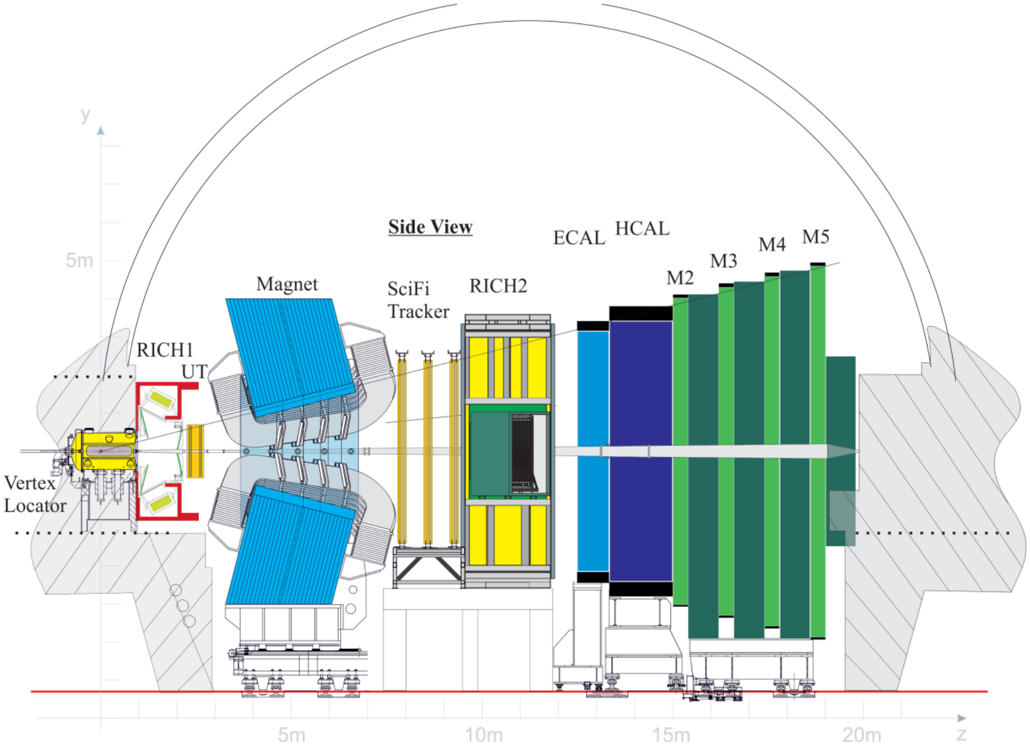

Figure 3: Structural outline of the LHCb detector. A cross-section showing LHCb's main subdetectors. At LHCb, proton beams travelling in opposite directions along the z-direction collide at the interaction point, located inside the VELO subdetector [7].

The subdetectors in the LHCb experiment, seen in Fig. 3, are each designed for one of two primary purposes: particle tracking or particle identification. The VELO and tracking stations (UT, SciFi Tracker) are utilised for particle tracking.

For particle identification, the RICH1, RICH2, the calorimeters ECAL and HCAL, and the muon stations (M2-M5) are employed. These are discussed in further detail below.

Inside the VELO detector, silicon pixel sensors surround the proton-proton (pp) collisions of the LHCb experiment. The detector's main purpose is to reconstruct the exact location of each of the pp-collisions in an event in 3D space. It is composed of an array of many individual modules, with each module including a set of silicon pixel sensors and associated electronics. These are arranged perpendicularly to the beam, spanning a length of approximately one meter [8]. Each VELO pixel module consists of square pixels arranged in an "L" shape. Half of the modules sit to the left of the proton beam and the other half sit to the right of it, effectively forming a "LT" shape around the LHC beam[3]. This ensures almost full coverage of decay products. When charged particles interact with the silicon pixel detectors in a module, they leave a hit, providing precise location information about the interaction points in three-dimensional space. Fig. 4 shows a representation of how the VELO hits are used to reconstruct the location of a pp-collision.

The Upstream Tracker (UT) is a silicon strip detector positioned before (or upstream of) the LHCb bending magnet. It comprises four layers of silicon microstrip detectors. It extends the tracking coverage before particles enter the bending magnet region. The SciFi tracker is a scintillating fibre tracker, responsible for measuring the trajectories of charged particles as they pass through the fibres with very high precision. This tracking information helps determine particle momenta and identify particle decays, since the tracks of particles such as pions originating from the same decay can be reconstructed to identify a point of origin.

RICH1 and RICH2 are Ring Imaging CHerenkov detectors, they each cover a different specific momentum range for which they detect and analyse the Cherenkov rings produced by charged particles. This information helps distinguish particle types based on their velocities. The two calorimeters, ECAL (Electromagnetic Calorimeter) and HCAL (Hadronic Calorimeter) are responsible for measuring the energies of the decay products. The ECAL identifies and measures the energies of particles involved in electromagnetic processes and decays, and is complimented by the HCAL, which provides information on the energy carried away by hadrons. The muon detector consists of four muon stations (M2-M5) of rectangular shape whose purpose is to detect and track muons - which due to being minimum ionizing particles that also do not have strong interactions, can traverse through other detectors. The muon stations are arranged in alternate layers with 80cm-thick iron absorbers.

[3] For a visual representation of the "LT" configuration, refer to Fig. A1.

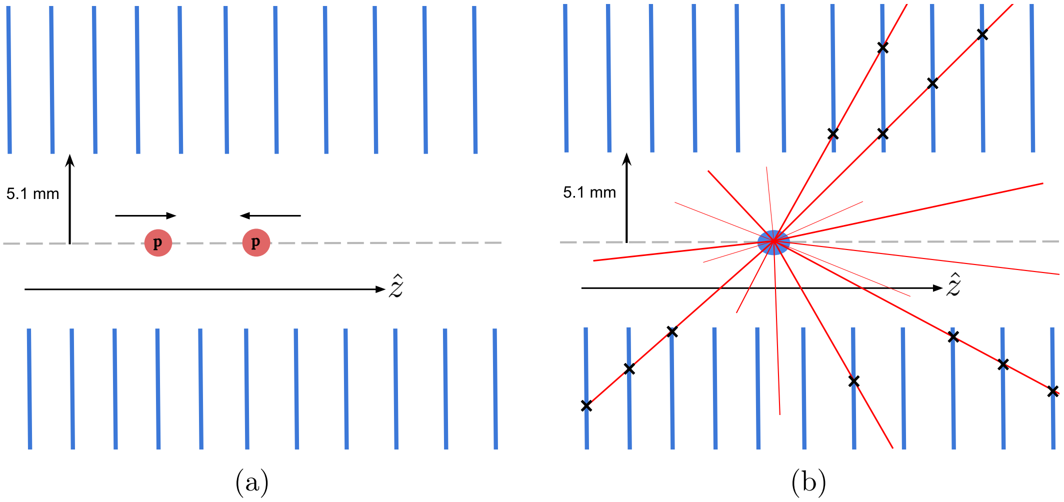

Figure 4: Cross-sectional outline of the functioning of VELO, where the blue vertical strips represent the VELO modules. (a) Protons travelling in opposite directions along the z-direction, moments before collision inside the VELO detector. (b) The particle shower after the pp-collision. Charged decay products that interact with the pixel detectors in VELO leave 'hits', the paths of these products are then reconstructed using these hits to accurately reconstruct the location of a pp-collision.

4 Decay Analysis and Simulation

4.1 The Signal Decay

The point where a pp-collision takes place is called the Primary Vertex (PV). It is here where the B_c^+ meson, along with many other B_c^+ candidate particles, is produced. The produced B_c^+ meson will briefly travel away from the PV, decaying into a τ^+ and a ν_τ, as seen in Fig. 5. The point of this decay is the Secondary Vertex (SV). Since the B_c^+ usually decays before reaching the VELO, the tau lepton does not decay immediately, and because the LHCb detector cannot detect the neutrino, it is not possible to accurately reconstruct the location of the SV. This means there is no available information regarding the direction of the B_c^+ meson.

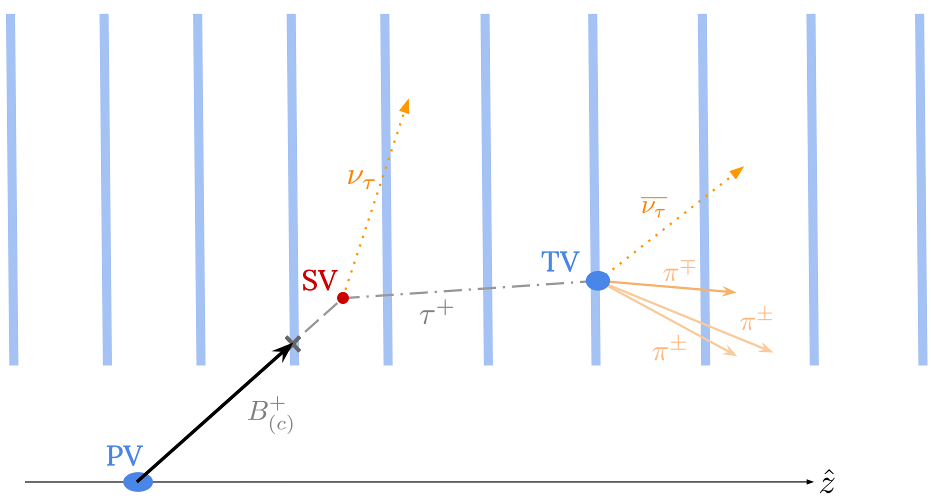

Figure 5: Diagram describing the signal decay with a B_c^+ hit left in the VELO detector, displaying how the direction of the B_c^+ meson, depicted with a black arrow, is inferred.

After travelling away from the SV, the τ^+ decays into three charged pions ($\pi^\pm\pi^\pm\pi^\mp$) and a tau antineutrino ($\overline{\nu_\tau}$) at a point called the Tertiary (or Tau) Vertex (TV). This second neutrino, like the first, cannot be reconstructed by LHCb. However, the three pions do leave reconstructible tracks. By calculating the point of intersection of these tracks, one can obtain an accurate measure of the location of the TV in 3D space. The fact that two neutrinos are emitted in this decay makes accurate reconstruction challenging since their momentum is lost information, preventing the accurate reconstruction of the momentum of the decaying particles and allowing only estimates of their directions.

In previous studies, a hit from the B_c^+ or τ^+ in the VELO modules was used to provide additional information on the direction of either particle.

In studies where VELO hit filtering takes place, this hit is assumed to be left by the B_c^+ meson, rather than the tau, providing a value for the approximate direction of the meson: PV → hit. This assumption is made because the VELO, whose purpose is to detect tracks of any charged particles in order to measure the position of the pp-collisions precisely, is not equipped to distinguish between different types of charged particles, so it is not possible to know whether it was the tau or the B_c^+ that left the hit.

4.2 RapidSim Simulations

This study has not been conducted using real collision data collected by the LHCb experiment; instead, data provided by the lightweight simulation software RapidSim was used [9]. The reason for this is that the study of a possible analysis of B_c^+ started relatively recently[4], and before analysing real LHCb data, it is necessary to ensure that the proposed study is feasible. Therefore, RapidSim has been used as an early-stage analysis for the feasibility of measuring the decay and its properties, and it has yielded positive results when filtering data based on a hit in the VELO detector. Should the method described in this thesis to search for $B_c^+ \to \tau^+\nu_\tau$ prove feasible, the next step would be to move the study to a comprehensive simulation program where more methods used to identify and measure the signal would be perfected. Only if this more thorough analysis continued yielding promising results would the methods be tested on real LHCb experiment collision data.

RapidSim simulates the pp-collisions that take place inside the VELO detector, giving the decay products of the collisions a set of kinematic values which would replicate their experimental values. This is possible because RapidSim contains a set of pre-established parameters derived from the theoretical properties of each particle. In addition to their properties, RapidSim also reproduces the behaviour of the decay products of the collision, allowing them to decay further and ultimately giving values for all the particles involved in a decay chain.

[4] 2018, through correspondence with M. Mulder.

All the kinematic properties and decay vertices are stored for each possible decay channel of the pp-collisions at the reproduced energies of the LHC, provided they fulfil some requirements. Namely, since the only decay mode for the signal that is being considered is $B_c^+ \to \tau^+(\to \pi^+\pi^+\pi^-\overline{\nu_\tau})\nu_\tau$, the backgrounds must also decay into three signal-like pions. Additionally, these three pions must have been created with a trajectory that places their path within the geometric limitations of the LHCb detector. RapidSim can simulate particles travelling in a wider array of directions than can be detected by the LHCb. Since the LHCb detector is not spherically symmetric around the point of pp-collisions but instead built like a forward spectrometer, this can result in particles flying outside of it rather than through it. This means the events that fulfil the selection requirements are those which decay into three signal-like pions that then fly into the LHCb detector. The stored data is then arranged into several large datasets, each containing data for either the signal or a single type of background decay mode. This results in distributions of the kinematic and geometric properties of the different decay modes. Then, by performing an offline analysis of the different properties in these distributions, one can attempt to distinguish signals from backgrounds.

In analyses that depend on VELO detector hits, an additional condition for offline analysis would be imposed: the B_c^+ candidate of each event must leave at least one hit in the VELO detector. This is an extremely rare occurrence for B_c^+ mesons so it greatly reduces the number of data available for offline analysis. Since this thesis is testing the feasibility of signal identification without the use of VELO hits, this part of the selection is skipped, allowing for much larger datasets for both signal and backgrounds. It is important to note that RapidSim is a lightweight software, and does not account for uncertainties that would be present experimentally when storing data. For example, values for the momentum of a pion in 3D will have no margins of error. To make up for the lack of error in individual entries RapidSim creates a smearing effect that affects the values of each entry, such that the peaks for the distributions of each observable are widened.

4.3 Description of Different Backgrounds

A summary of the different backgrounds is given to provide context for the separation process and highlight the specific challenges encountered when distinguishing between signal and background. There are 25 different types of background decays at play, among which $B^+ \to \tau^+\nu_\tau$ is a decay of particular interest, where B^+ is composed of a bottom and an up quark. The remaining backgrounds that have passed the offline selection requirements can be divided into four main groups, namely

- **B→DD :** These are backgrounds where the B meson or Λ_b baryon decays into two different D mesons, the lightest group of particles containing charm quarks. One of these D mesons later decays into a tau, which decays into three charged pions. This is the most diverse group of backgrounds, containing 12 different types of decays of which nine originate from neutral b-hadrons, however, the yields of decays in this group are relatively small when compared to other backgrounds.

- **B→D$\boldsymbol{\tau\nu}$:** The B meson in these backgrounds decays simultaneously into a D meson, a tau and a tau neutrino. The tau lepton then decays the same way as in the signal decay and its pions are detected. This is the second most diverse group and contains eight different types of decays, six of which originate from neutral B mesons.

- **D→ $\boldsymbol{\tau\nu}$:** These are the backgrounds which originate from a D meson, which decays into a tau and a tau neutrino. The tau is followed by three charged pions, in a similar decay structure to the signal. There are only two types of decays in the D cocktail, however, one of the main challenges they pose during the separation process is the exceedingly large number of these backgrounds that are produced in the pp-collisions. However, in analyses that rely on a VELO hit, these backgrounds are very strongly rejected by the VELO hit requirement.

- **B→D3$\boldsymbol{\pi}$:** These are backgrounds with no intermediate tau lepton. The B-meson decays directly into three pions and a D-meson, making them the simplest group to separate from the signal.

5 Separation Variables

Since every B_c^+ candidate that belongs to a background decay will have different properties than the B_c^+, their average decay will look different than that of the signal. For example, they might be longer-lived particles, giving larger average values for the distance between the PV and the TV (also called the flight distance (FD)); or they might decay into other particles at the SV besides the tau, resulting in tau leptons with a lower average momentum spectrum. This section will elaborate on the physical observables used to differentiate between the signal and backgrounds. It will also rely heavily on comparison between the methods for obtaining values for observables in this study and in studies which benefit from reconstructed VELO hits.

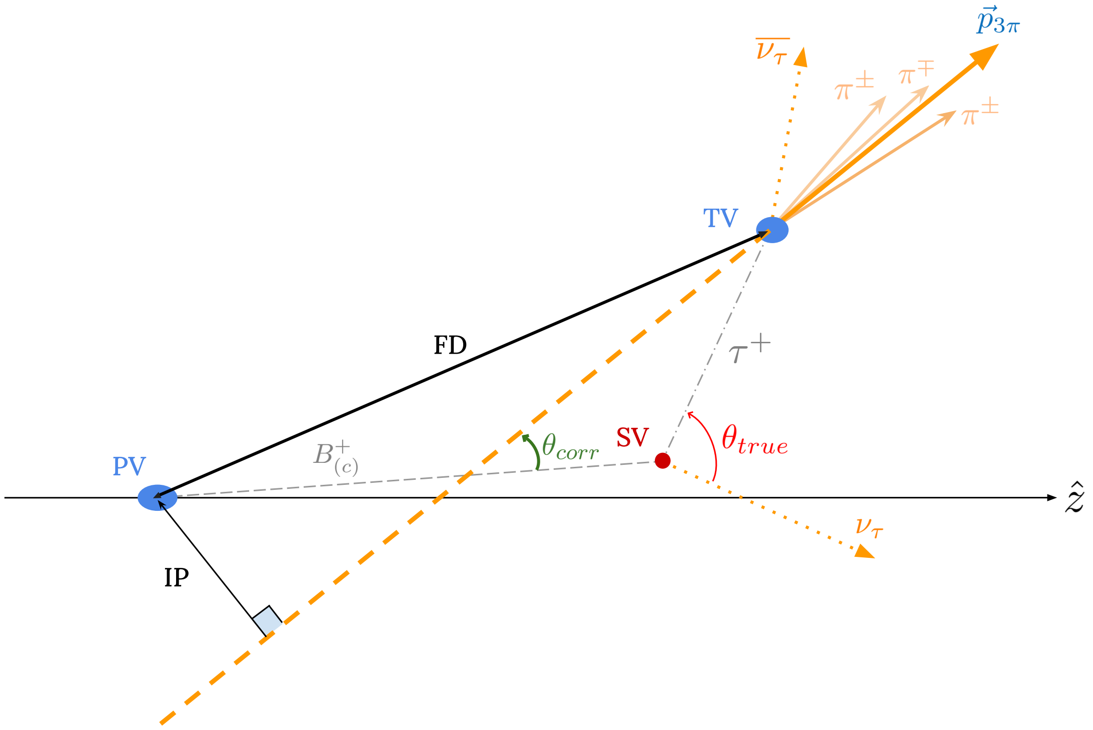

Figure 6: Diagram displaying the decay with some of the observables used in this study, such as Flight Distance (FD), Impact Parameter (IP) and the total pion momentum ($\vec{p}_{3\pi}$); as well as some variables that cannot be used due to lack of information on the location of the SV and the direction of the B_c^+ meson, such as θ_{true} and θ_{corr}.

5.1 Flight Distance and Impact Parameter

As already mentioned, the FD is the distance between the pp-collision and the point where the three signal-like pions are created. It is a composite variable, meaning that several different factors affect its final value. The main factor is the lifetime of the B_c^+ candidate particle, as the longer this particle lives, the farther it will travel. This is useful information because the B_c^+ meson, being the heaviest meson of all the pp-collision products that pass the selection, is expected to have the shortest lifetime. Another factor that affects the FD is the number of intermediate particles between the PV and the TV, as the larger the number of intermediate particles, the larger the FD will be. Finally, the FD also depends on the momentum of the particles, since the larger their momentum, the longer they travel before decaying. The FD is meant to provide a rough estimate of the lifetime of the B candidate, which is why it is one of our variables of interest.

The IP is the distance of closest approach of a certain track to the PV. If a particle originated directly from the PV and its direction could be tracked, the IP would have a value near zero. Consequently, particles with larger IP values are likely not to have originated directly from the PV, but instead from an intermediate particle with a marginally long lifetime, such as a tau lepton. In this study, the momenta of the three pions were summed and the impact parameter of this combined track was taken. The longer the lifetime of the intermediate particle(s), the farther away from the PV the pions are created and so it is less likely for the IP to have values close to zero. i.e. the distribution of this variable is heavily dependent on the lifetime of the intermediate particles. Since the lifetime of the B_c^+ meson is shorter than that of the background B_c^+ candidates, the signal distribution is predicted to have lower distribution values than the backgrounds.

Since both the FD and the IP are only concerned with the initial and final states of the decay, the method of analysis of this study does not vary from studies with VELO hit filtration.

5.2 Invariant Mass

The rest mass and the invariant mass are related, but distinct concepts in particle physics. The rest mass is a property that relates to individual particles, such as a single tau lepton. It is the mass of the particle when it is at rest and it is a property that remains constant, regardless of the particle's velocity, in any reference frame. For a single particle, the invariant mass would be the same as its rest mass; however, the invariant mass is a property that also applies to a system of particles, such as three pions and a neutrino. It is a property that is calculated using the total energy and momentum, properties which notably remain constant during particle decay, of the system. It is called invariant mass because it has the same value in all inertial reference frames. For a system of particles, the invariant mass is given by

$$m_{inv}^2 = (\Sigma E_i)^2 - (\Sigma \vec{p_i})^2 \tag{5.1}$$

where E_i and $\vec{p_i}$ are the energy and momentum of each particle in the system. When a tau lepton decays into three pions and a neutrino, the invariant mass of the decay products must be equal to the rest mass of the tau lepton; however, information on the neutrino's momentum and energy are lost in the measurement process. This results in a system of three pions where the value for the invariant mass will not sharply peak at the tau lepton's rest mass, m_τ, but instead it will result in a distribution of values for the corrected mass of the three pions, given by

$$m_{3\pi}^2 = (E_{3\pi})^2 - (\vec{p}_{3\pi})^2 \tag{5.2}$$

where $E_{3\pi}$ and $\vec{p}_{3\pi}$ are the total energy and momentum of the three-pion system. This distribution of values for $m_{3\pi}$ will range from being equal to $3m_\pi$, in a case where the three pions are at rest and the neutrino possesses all the momentum after the decay, to being equal to the tau rest mass, in a case where the massless neutrino is at rest and all the momentum has been transferred to the three pions. The invariant mass squared distribution is then described as $m_{3\pi} \in [3m_\pi, m_\tau]$.

The invariant mass squared proves to be a good separation tool for the B→D3π group. Since these backgrounds decay into a system containing three pions directly from the B_c^+ candidate and not from an intermediate tau, their invariant mass distribution is instead constrained by a much larger maximum value, the mass of the B_c^+ candidate. It is for this exact reason, however, that the invariant mass is not a good separation variable for all the other background groups, since the system containing three pions originates from a tau lepton, resulting in invariant mass distributions that look effectively identical to the signal's.

5.3 Corrected Mass

Since for the groups of backgrounds where the three charged pions decay from a tau, the invariant mass is not an effective separation tool, another mass variable is used in an effort to differentiate between the different backgrounds and the signal. This variable is the corrected mass, given as

$$m_{corr} = \sqrt{m_{3\pi}^2 + |\vec{p}_\perp(3\pi)|^2} + \vec{p}_\perp(3\pi). \tag{5.3}$$

In studies implementing VELO hits, $\vec{p}_\perp(3\pi)$ would be the total three-momentum component of the three pions perpendicular to the flight direction of the B_c^+ candidate. Since the particles inside the LHCb are ultrarelativistic and there is no available information on the $\overline{\nu}_\tau$, in studies with VELO hits it is assumed that the tau direction is equal to the direction of the 3π combination; Fig. 6 is then a diagram with exaggerated opening angles for visual aid. Implementing a similar logic, this thesis also approximates the direction of the B_c^+ candidate to be equal to the direction of the flight distance, under the understanding that in practice they are mostly very close to each other. For this reason, $\vec{p}_\perp(3\pi)$ in this study is approximated by taking the perpendicular component of the 3π momentum with respect to the direction of the FD.

The variable m_{corr} makes use of the perpendicular component of the three-pion combination because this value provides an estimate of the energy lost not only to the neutrinos during the decay, but also to any other massive particles that could have been produced alongside the three pions.

5.4 Pion Momentum

Finally, the magnitude of the total momentum of the 3π combination, $p_{3\pi}$, as well as its momentum transverse to the beam direction, p_T, are used as separation variables. Generally speaking, larger tau momenta lead to higher total pion momenta. This is an important property, since taus produced alongside other particles, or from an intermediate particle instead of directly from the B_c^+ candidate, will have different momentum spectra, specifically with lower average momenta. Complementing this, the transverse momentum also provides a basis for separation; since particles with higher velocities will decay into products with smaller opening angles, providing the tau and therefore the pions with smaller transverse momenta, while particles with similar speeds but larger masses will yield decay products with larger opening angles.

6 Data Analysis

Neither of the separation variables described in the previous chapter provides sufficient discrimination of signal from backgrounds by itself. For this reason, a multivariate analysis (MVA) is used, taking in the five observables of FD, IP, the invariate mass squared, and the total and transverse momentum of the three pions. The separation obtained from the MVA is then plotted against the corrected mass, forming a 2-D plot. This plot will reveal if there are any areas where the ratio of signal to background is large enough to consider it an area of successful separation.

6.1 Multivariate Analysis

Multivariate analysis is a method used to analyse data that involves processing multiple variables simultaneously. By considering how these variables relate to each other, instead of seeing them as independently occurring from one another and by identifying patterns in the data, the MVA creates a single composite score that distinguishes between signal and background more effectively. To classify a single entry as signal or background, this study makes use of a gradient boosting classifier, namely, a boosted decision tree (BDT), which is a machine learning algorithm belonging to the scikit-learn package [10].

Using a BDT involves creating a dataset that contains entries of both signal and background events, and their corresponding values for a defined set of variables which represent the observables of interest. The machine learning algorithm will then work through this dataframe, identifying underlying patterns and relationships within the data to train a model. This (trained) model is then the output of a machine learning algorithm, applied to a dataset.

To train the model to classify between a signal and a background entry, one variable must be added: the MVA Identifier. This variable has a value of 1 for the signal and 0 for all of the backgrounds. Since the data originates from simulation software, there are different datasets for the signal and each background. An example of what these datasets look like is given in Table A1. As further preparation, all the individual data frames (signal and background) are merged into one large dataframe for training and testing the BDT model. In this combined dataframe, called the Learning Dataset, half of the entries belong to the signal, and the other half to the backgrounds. The number of entries for each background type maintains the ratio of detected backgrounds to each other, which would be observed in an experimental setting where signal and background are mixed [11]; the only intentional deviation from real data is that the dataset consists of 50% signal and 50% background.

Once the BDT model has been trained and tested, it will be run on a second dataset, called the Application Dataset. This is made entirely of new entries, such that no data is repeated. The Application Dataset differs from the Learning Dataset by having signal entries that emulate the signal-to-background ratio expected in an experimental setting. The Application Dataset emulates a dataset obtained from real, experimentally obtained data.

6.1.1 Training and Testing a BDT Model

To obtain a functional BDT model, it must first go through a training phase, followed by a testing phase. The training phase uses 80% of the entries (rows) from the Learning Dataset, selected at random. This subset is referred to as the Training Dataset. The remaining 20% is used for testing and is known as the Testing Dataset.

Training the BDT model involves using the columns of the Training Dataset that contain the separation variables and associating each row's values to the corresponding value in the MVA Identifier column, standing at 1 for signal and 0 for background. In other words, for each entry, the columns with separation variables are taken as the independent variables for the model and the MVA Identifier is taken as the dependent variable that the model will later aim to predict.

While testing, which utilises the Testing Dataset, the trained model reads the values of the independent variables and outputs a prediction of the dependent variable, called y_prob, for each entry. This prediction is a probability value ranging from 0 to 1. For each entry in the Testing Dataset, y_prob is the model's attempt to guess the value of the MVA Identifier based on the independent variables. While training, the model can read the value of the MVA Identifier and associates these values. During testing, it predicts the MVA Identifier without seeing its actual value, outputting y_prob instead. A value of 0 represents a 0% chance of the entry, given its values for its independent variables, being a signal; 0.5 represents a 50%, and 1 represents a 100% chance that the entry belongs to a signal decay.

The values for y_prob obtained from the Training Dataset, where approximately half of the entries belong to the signal, can be seen in Fig. 7. The spike of background decays between y_prob values of 0.0 and 0.1 belong to the B$\rightarrow$D3π group, since in this group the 3 pions decay directly from the B meson, resulting in their invariant mass distributions to be drastically different from the signal's.

This allows for a very high separation efficiency[5]. Other than this, it is apparent that the MVA results in some separation of signal and background, since the background distribution peaks at approximately 0.3 while the signal distribution peaks around 0.85. Additionally, there is a small section at the highest scored values of y_prob with significantly more signal than background. However, the overall values for signal and background overlap heavily.

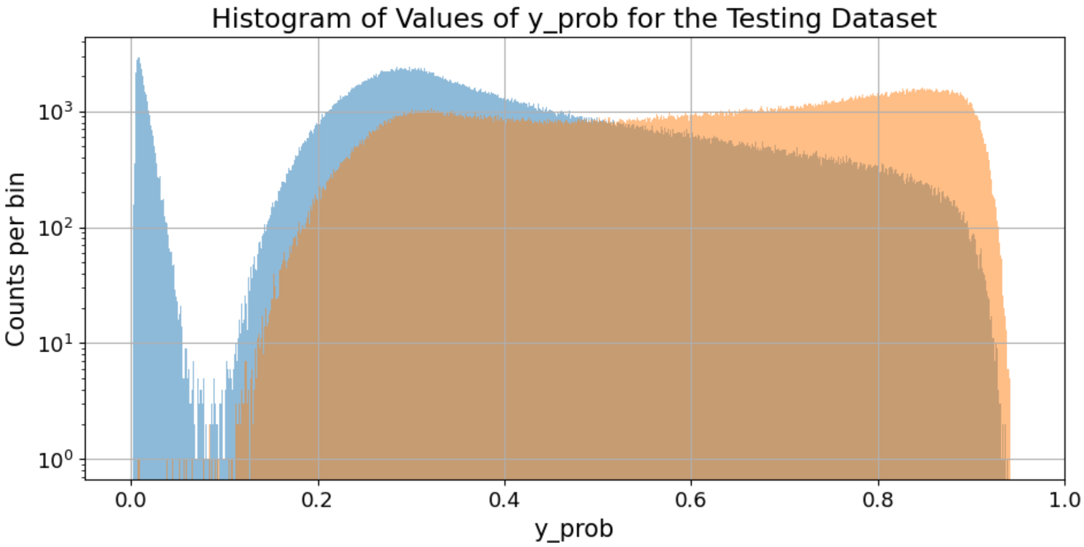

Figure 7: Histogram representing the values of y_prob obtained by applying the BDT to the Testing Dataset. The background is shown in blue, including a sharp spike at low values of y_prob, and the signal is shown in orange. The specific counts per bin are arbitrary, as the size of the Leaning Dataset dataset was chosen arbitrarily.

6.1.2 The ROC Curve

The BDT model is a binary classifier between signal (labelled as 1) and background (labelled as 0). A method to visualise the performance of binary classifiers is by plotting a Receiver Operating Characteristic curve, or ROC curve. A ROC curve judges the separation ability of a model by comparing the model's prediction (i.e. the values of y_prob) to the true statuses of the data (i.e. if they belong to signal or background).

The ROC curve of the trained model in this study can be seen in Fig. 8, where AUC stands for Area Under the Curve. The AUC of a perfect separation

[5] An example can be found in Fig. A2

of signal from the background would equal one, in which case, a histogram of values of y_prob would display no overlap between background and signal. In the case that the model is entirely unable to distinguish signal from background, the AUC would equal 0.5. In this graph, the False Positive Rate (FPR), plotted on the horizontal axis, measures the proportion of background incorrectly identified as signal by the model. Conversely, the True Positive Rate (TPR) measures the proportion of signal correctly identified by the model. Since the BDT model has an AUC of 0.76, it demonstrates the capability to distinguish between signal and background, but with limited accuracy. This is not ideal, since the proportion of signal to background in the Applied Dataset is vastly lower than in the Learning Dataset, which means that accuracy is crucial to providing a successful separation of signal and background in the Applied Dataset.

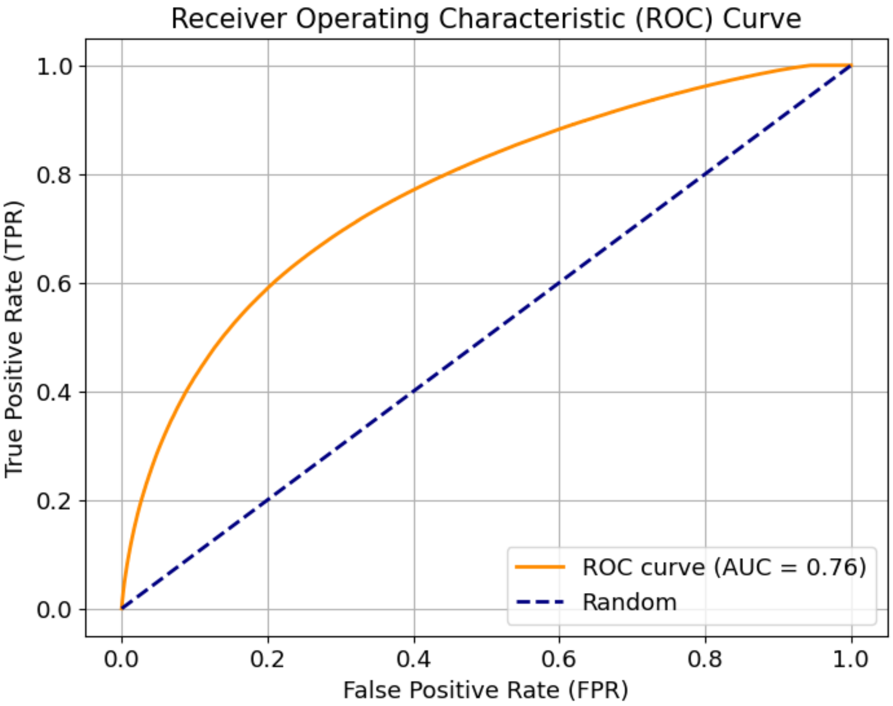

Figure 8: ROC curve for the trained BDT classifier model, where the TPR on the y-axis is plotted against the FPR on the horizontal axis.

It is important to note that the Gradient Boosting model is configured according to certain parameters such as the tolerance, learning rate, number of estimators, and maximum depth of each tree in the model. These parameters can be optimized to maximize the separation power of the model. The number of estimators was manually optimised, training the model for 50, 100, 200, and 250 estimators, and finding maximal separation once 200 estimators were implemented. However, the remaining parameters in the model were kept to their default values, i.e. they were not optimised. This means it is possible that the classifier model was not separating signal from background optimally, and further parameter tuning could potentially improve its performance.

6.1.3 Implementing the Trained BDT Model

After the model has been trained, tested, and assessed, it is implemented on the Application Dataset. A new histogram of y_prob for signal and background is created with this dataset. If in the new histogram of values of y_prob there is a visible value of y_prob where a cut in the data can be made (such that most of the background is eliminated and the remaining data displays a satisfactory signal-to-background ratio) this would result in a successful separation of signal and background. Such a result would indicate the search for the signal decay without VELO hits at the LHCb experiment is feasible. As shown in Fig. 9 the signal is completely overlapped by much higher levels of background. Accordingly, there is no value of y_prob where the data can be effectively separated from the background.

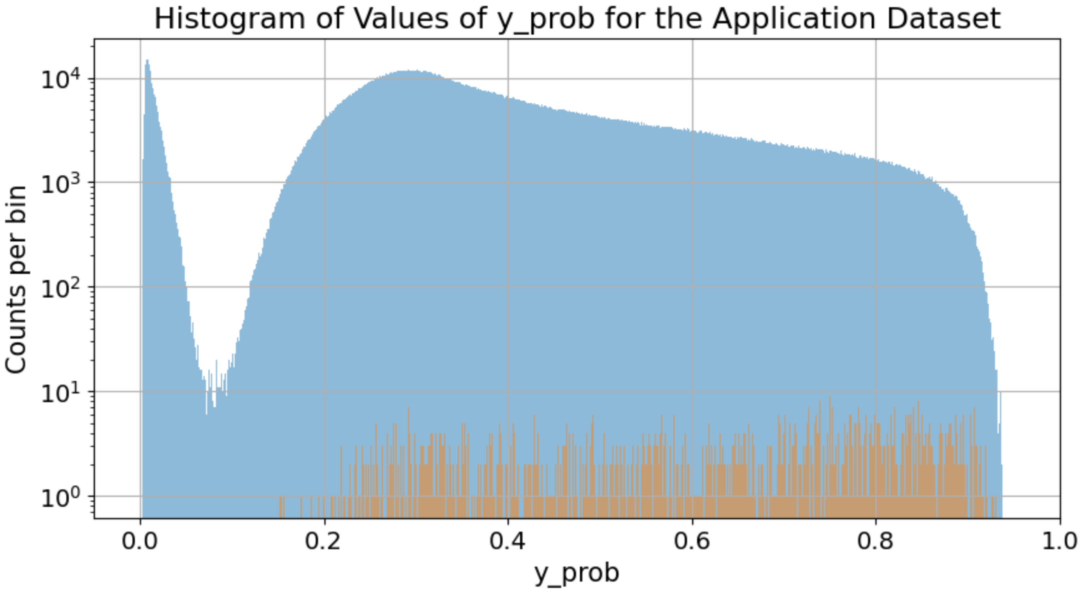

Figure 9: Histogram representing the values of y_prob obtained by applying the trained model on a dataset that emulates experimentally obtained data. The background is seen in blue and the much smaller signal is seen in orange.

6.2 Corrected Mass and Two-Dimensional Plot

6.2.1 Corrected Mass Distribution

Since the signal can not be effectively separated from the background by using only the trained model, the values of y_prob shown in Fig. 9 are then plotted against the values obtained for the corrected mass variable distributions. This two-dimensional plot is then used to attempt to find an area of separation between signal and background. Details on this method are given in Section 6.2.2. It has been previously mentioned that the corrected mass is a mass variable that tries to correct for the missing momentum of the three pions lost to the neutrinos and, in the case of some backgrounds, other massive particles produced in the decays. It is the clearest separation variable. Fig. 10 is a histogram showing the different corrected mass distributions for the signal and every type of background. For the corrected mass variable, each background group mentioned in Section 4.3 has very similar distributions, such that in Fig. 10 they can be expressed with a representative decay from each group.

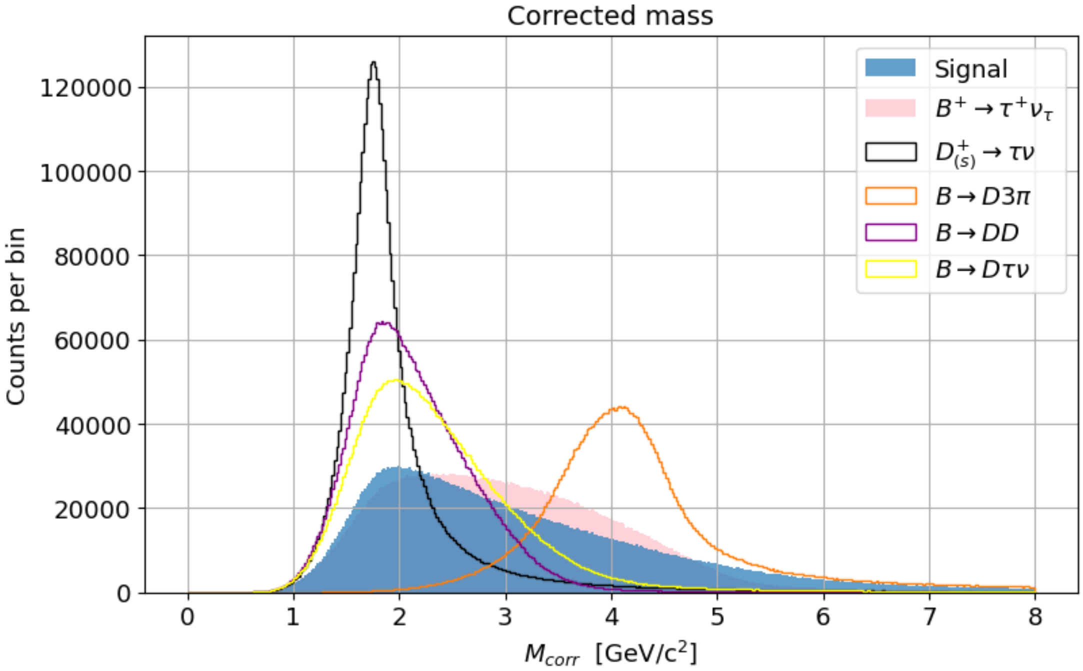

Figure 10: Histogram showing the corrected mass distributions of the signal and backgrounds. Each distribution has been plotted with a normalised yield to allow for a more meaningful comparison.

6.2.2 Two-Dimensional Plot for Signal Identification

Neither the MVA or the corrected mass variable can single-handedly separate signal from background. However, they both give a histogram with a small amount of separation between signal and background. The last step to separate signal from background is plotting the histogram seen in Fig. 9 against the histogram for the logged corrected mass values of the Application Dataset.

Plotting one histogram against another creates a two-dimensional plot. This plot is seen in Fig. 11, where the peak at around 1.4 for the logged corrected mass for the 2D histogram for background events is indicative of the leftmost peak in Fig. 9. In this Figure, an area where one can identify pixels for signal events in a significantly larger frequency than background events would signify that there is a range of values of y_prob and corrected mass that can successfully separate the signal from the backgrounds.

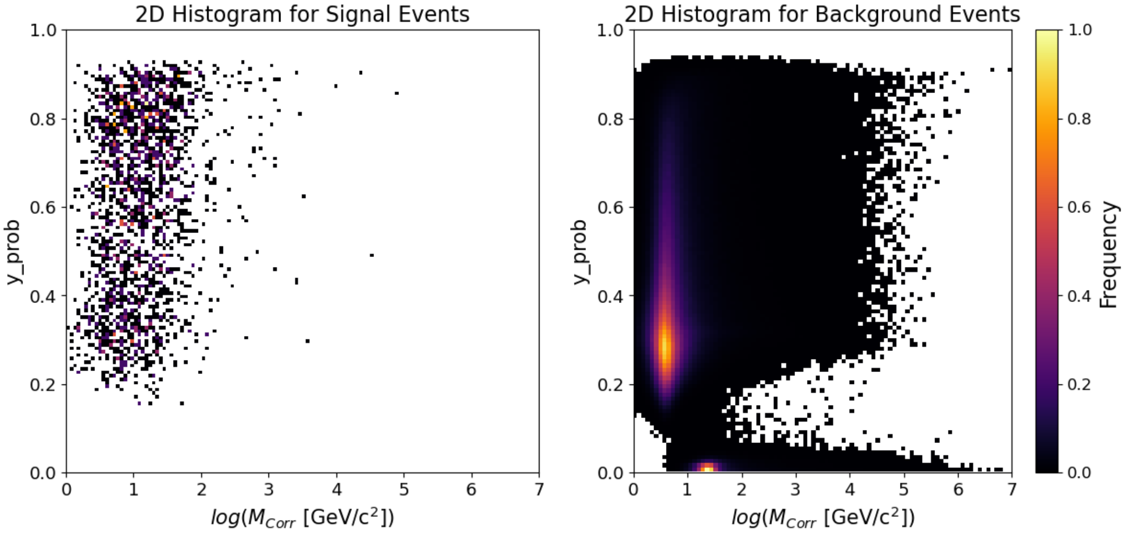

Figure 11: 2D histograms with an independently normalised z-axis such that the bin with the most entries is given a value of 1.0. Plotted are the values of y_prob given by the BDT against the respective value for logged corrected mass for each entry in the Applied Dataset. The histograms are separated between signal and background for visual aid.

At first glance, there is no area where one can distinctly identify signal from background. Therefore, a zoomed-in version of Fig 11 is made, namely for values of y_prob above 0.9, where the background significantly decreases. This results in Fig. 12, where an area of effective separation between signal and background is still not seen. This means that even by combining the separation power of the BDT model and the corrected mass, there is no region in which one can observe a satisfactory ratio of signal to background events. This means that it is not possible to separate signal from background using the method described in this study.

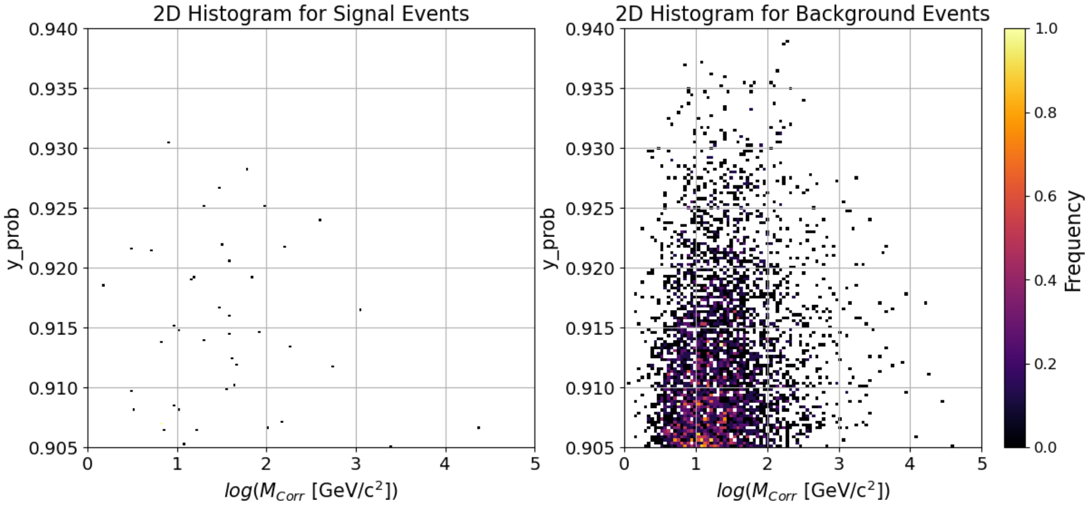

Figure 12: Zoomed-in 2D histograms of the results displayed in Fig.11 for values of y_prob between 0.905 and 0.940, and values for the logged corrected mass between 0 and 5. In these histograms, 50 signal events and 5277 background events have been plotted. For comparison, the Application Dataset contained 2060 signal events and 3,763,767 background events.

7 Conclusion

The objective of the research into the $B_c \to \tau\nu_\tau$ decay channel performed by the LHCb Groningen research team is to measure its branching fraction and compare it to the prediction value established by relying on lepton flavour universality. A deviation from the LFU prediction could hint at physics beyond the SM. To achieve this, the research team has developed a method that only selects for analysis the decays that directly interact, or leave a 'hit' in the VELO detector at LHCb by the B_c^+ meson or other viable B_c^+ meson candidates. This hit provides information on the approximate direction of the B_c^+ meson and the background B_c^+ candidates, which aids in the separation of signal from background events.

This method suffers from an important limitation. Namely, filtering all the decays to analyse only those which have left at least one VELO hit entails a drastic reduction in usable data. This is because such direct hits in the VELO are very rare due to the extremely short lifespans of the B_c^+ mesons. Accordingly, the purpose of this thesis was to evaluate the feasibility of studying this decay channel without relying on VELO hit information.

Using data generated by the lightweight simulation software RapidSim, this study analysed various decay observables, and a selection procedure was developed to distinguish between signal and background events. The majority of these observables were used as input variables in a multivariate analysis - namely, the flight distance between the pp-collision and the vertex where the three pions are created; the impact parameter of the three pions; the invariate mass squared; and the total and transverse momentum of the three pions with respect to the beam axis. The multivariate analysis entailed training a boosted decision tree classifier model with these input variables, and the output of this model was then plotted against the corrected mass (a separation variable that accounts for energy lost to undetectable neutrinos and, for backgrounds, other massive particles) to create a two-dimensional histogram.

Identifying a region in this two-dimensional histogram with a significant signal-to-background ratio, within a dataset containing the expected number of signal and background events in real data, would have validated this separation method, making the study of the $B_c^+ \to \tau\nu_\tau$ branching fraction feasible. However, the study found no region that met this criterion, rendering the proposed method ineffective for this purpose. For this reason, this approach to studying the $B_c^+ \to \tau\nu_\tau$ decay channel is considered not feasible.

8 References

[1] B. R. Martin and G. Shaw, *Particle Physics*. Wiley, 2017, ISBN: 978-1-118-91190-7.

[2] J. Ellis, *Limits of the Standard Model*, 2002. arXiv: hep-ph/0211168 [hep-ph]. [Online]. Available: https://arxiv.org/abs/hep-ph/0211168.

[3] K. Müller and on behalf of the LHCb Collaboration, "Tests of Lepton Flavour Universality at LHCb," *Journal of Physics: Conference Series*, vol. 1271, 2019. DOI: 10.1088/1742-6596/1271/1/012009.

[4] J. de Jong, "Feasibility study of the branching fraction measurements of Bc -> tau nu and Bu -> tau nu at LHCb," M.S. thesis, University of Groningen, 2022.

[5] Cush, *Standard Model of Elementary Particles*, Sep. 2019. [Online]. Available: https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg.

[6] L. Evans and P. Bryant, "LHC Machine," *Journal of Instrumentation*, vol. 3, no. 08, Aug. 2008. DOI: 10.1088/1748-0221/3/08/S08001.

[7] LHCb Collaboration, "LHCb Tracker Upgrade Technical Design Report," Tech. Rep., 2014. [Online]. Available: https://cds.cern.ch/record/1647400.

[8] LHCb Collaboration, R. Aaij, A. S. W. Abdelmotteleb, *et al.*, *The LHCb upgrade I*, 2023. arXiv: 2305.10515 [hep-ex]. [Online]. Available: https://arxiv.org/abs/2305.10515.

[9] G. Cowan, D. Craik, and M. Needham, "RapidSim: An application for the fast simulation of heavy-quark hadron decays," *Computer Physics Communications*, vol. 214, May 2017, ISSN: 0010-4655. DOI: 10.1016/j.cpc.2017.01.029.

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

[11] M. D. Galati, "B2taunu feasibility paper," Document in preparation.

A Appendix

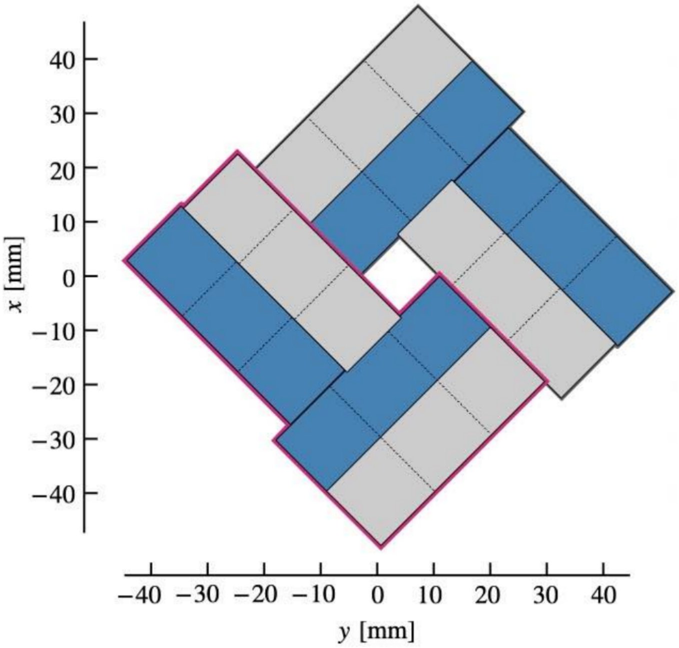

Figure A1: Diagram showing the layout of the two L-shaped modules around the z-axis in the closed VELO configuration, taken from [8]. The proton beam is aligned along the z-axis at the (0,0) coordinate.

Table A1: Signal MVA Dataset

	log(FD[mm])	log(IP[mm])	$m_{3\pi}^2$[GeV/c^2]	log($\vec{p}_{3\pi}$[GeV/c])	log(p_T[GeV/c])	MVA Identifier
0	2.655	-1.178	1.590	4.101	1.611	1.0
1	2.380	-1.239	0.616	4.133	0.657	1.0
2	0.274	-0.589	0.834	2.069	-0.066	1.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮
3800000	1.446	-0.564	0.505	1.911	-0.355	1.0

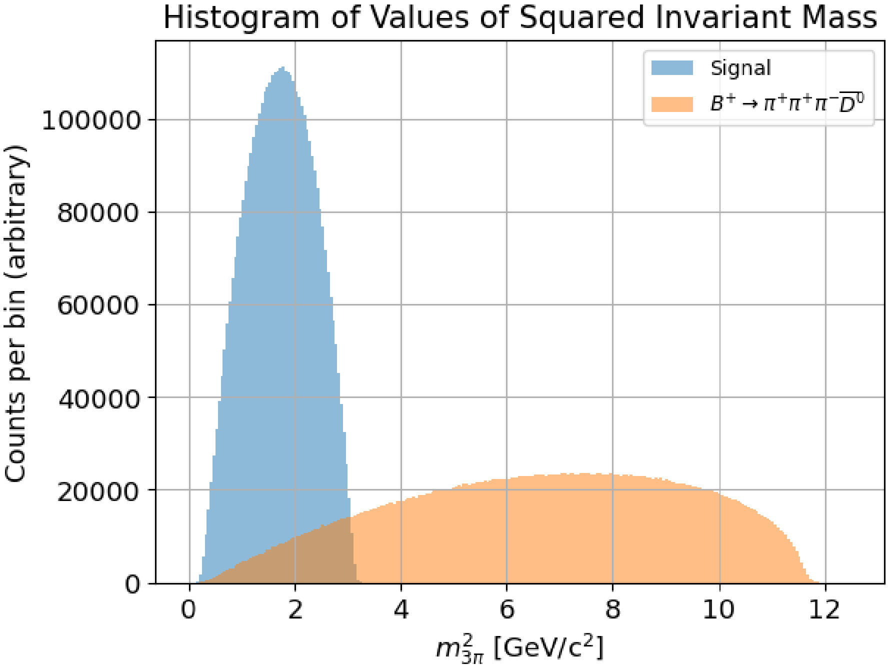

Figure A2: Histogram showing the distribution of values of the squared invariant mass for the signal decay and the $B^+ \rightarrow \pi^+\pi^+\pi^-\overline{D}^0$ decay, from the B→D3π group.

Acknowledgements

I want to extend my deepest gratitude to my supervisor, Mick, for always cheering me on and believing in me, even when I struggled to do so. Your unwavering support has been monumental to me, you have given me the confidence to keep pushing towards the end. Thank you truly.

I want to thank all the incredible friends that I have made throughout my BSc. You have been the light of my life. You have all kept my life beautiful and fun and I am incredibly grateful to all of you. Your love and support have helped me more than I can ever repay. You have truly been the wind in my sails.

Finalmente, quiero dar mil gracias a mi familia. Me siento la persona más afortunada del mundo por ser parte de la familia Cebolleta. Me inspiráis todos los días. Gracias mil por haber podido contar siempre con vuestro apoyo y vuestro amor.

Gracias en especial a mi madre, a mi padre y a mi abuelita Rosita. Os quiero muchísimo. Gracias por quererme tanto y apoyarme de forma incondicional, y por recordarme, una y tantas veces, que en la vida hay que ser feliz.