# Pentaquark Search in $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ Decays and LHCb Open Data Release

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
**Mindaugas Šarpis**
aus
Vilnius

Bonn, 2023

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen
Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

| | |
|---|---|
| Gutachter: | Prof. Sebastian Neubert |
| Gutachter: | Prof. Florian Bernlochner |
| Tag der Promotion: | 21 December 2023 |
| Erscheinungsjahr: | 2024 |

Tėvams Ramonai ir Viktorui

# Abstract

This PhD thesis reports results of the physics analysis focused on the search of pentaquarks in $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ decays. Pentaquarks were first reported in 2015 by the LHCb collaboration in $J/\psi p$ system. After collecting more data, further results followed in 2019 and 2022. Observation or non-observation of $P_c^+$ in $\Lambda_c^+ \overline{D}^{*0}$ system would help to deepen the understanding of these exotic states. The data collected by the LHCb detector at CERN in the years 2015-2018 from proton-proton collisions at center of mass energies of 13 TeV is analyzed. This data amounts to 5.7 $\text{fb}^{-1}$ of integrated luminosity. After reconstructing the missing momentum of the soft $\pi^0$ or a $\gamma$ from the decay of a $D^{*0}$ meson, the decay dynamics are studied and the search for the pentaquark states in the $\Lambda_c^+ \overline{D}^{*0}$ system is performed. No signal is observed and the `CLs` upper limits on the fit fractions for each of the studied pentaquark states are established at 95% confidence level:

$$f_{P_c(4312)^+} < 0.52\%$$

$$f_{P_c(4440)^+} < 0.65\%$$

$$f_{P_c(4457)^+} < 0.54\%$$

The Extended Cone Closure method developed to reconstruct the missing momentum in a 3-body decay could be used for further analyses with missing neutrals. Part of this thesis is dedicated to reporting the work done in preparation for the first LHCb Open Data release. A subset of the LHCb data from the years 2011 - 2012 were made publicly available. The data sets are described with `Open Data Records` and are available on CERN Open Data Portal. An involved automated workflow was established for data curation and documentation.

# Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own, except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Mindaugas Šarpis, 2023*)

# Acknowledgments

When I started this PhD, I thought, I roughly knew how it was going to go. Just a normal PhD with LHCb (if there is such a thing). Well, this was further away from normal than I could have ever imagined. Six months into my PhD, while still in Heidelberg, a global pandemic started. These were the times of great uncertainty, many of the foundations, holding the society, were under stress. With the pandemic starting to ease towards the end of my time in Bonn, a war in Europe broke out. Needless to say, the minds would often veer off to these vast complex societal issues unfolding before our eyes. The support of my family, friends and colleagues was thus more important than ever. First, I would like to thank my supervisor Sebastian. Your trust in me, coming from a small country and a different field of physics, has opened me many doors and will open many more. I am very grateful for the opportunity to join you on the scientific endeavors. I also want to thank all my colleagues, many of whom became friends from Heidelberg, Bonn and the LHCb collaboration. Nicole, Marian, Jascha, thank you for your patience in overseeing my first steps in particle physics. Klaas, thank you for your continued help all around. Adam, thank you for opening a whole new dimension to the lock-downs during the pandemic. Hannah, Elli, Piet, Kai, thank you for the interesting discussions and listening to my Lithuania-stories. I was very lucky that on this path I was not alone. I want to thank my beautiful wife Justė for not only being patient with my physics struggles, but also for enabling and encouraging me to follow this path when it got hard. As I said in the very beginning, *"Net ir šaltą fiziko širdį galima sušildyti, reikia tik labai nepaprastos daktarės."*. I want to thank my son Augustas, who made the last year of the PhD even more exciting than it already was. Finally, I would like to thank my brother Gediminas, my sister Viktorija and my parents Ramona and Viktoras, whom I dedicate this thesis to.

# Contents

# List of Figures

# List of Tables

# Thesis Overview

Following a short introduction and description of the analysis strategy, the first chapter of this thesis provides an overview of the accelerator complex at CERN as well as a deeper look at the LHCb detector. Each sub-detector is briefly introduced as well as data management workflow at LHCb. The second chapter focuses on the theoretical background relevant to this analysis as well as the motivation behind the pentaquark search in the $\Lambda_c^+ \overline{D}^{*0}$ system. Statistical basis for the analysis and specifically the `CLs` limit setting procedure is explained. The fourth chapter is dedicated to the physics analysis itself, discussing data samples, various corrections applied, a method to reconstructing missing momenta in the decays and the results of the limit setting procedure. A lot of focus was given to analysis preservation and reproducability during this work. The fifth chapter introduces computational tools and environments used to achieve this. In the sixth chapter, the principles of making experimental data available to the public are discussed and the results of the first LHCb Open Data release are reported.

# Introduction

In the quest to understand the inner workings of the Universe, scientists have been looking into its smallest constituents. Protons and neutrons as well as electrons are now widely known as the particles which make up atoms. Naturally to gain a deeper understanding of the origins of matter a better understanding of these building block was needed. A big advancement came with the discovery of quarks. These particles were independently proposed in the 1960s by Gell-Mann and Zweig in what later became known as the quark model. It is now accepted that protons and neutrons both have 3 valence quarks (uud) for the proton and (ddu) for the neutron. The quarks within nucleons are bound by the strong force. Particles containing an odd number of quarks are called baryons. Another group of particles - mesons have an even number of quarks. Even with the very first introduction of the quark model, particles containing more than 2 or 3 quarks were hypothesized. They are states of four, five or more quarks, named tetraquarks, pentaquarks, etc. Investigating such states provides a good probe of *Quantum Chromo Dynamics* (*QCD*), the theory explaining how quarks bind together to form hadrons and how these bound states interact via the strong interaction. After decades of inconclusive or conflicting results from different experiments first confirmed discovery of a pentaquark ($P_c^+$) state was reported by the LHCb Collaboration in the year 2015. This discovery was made in the decay $\Lambda_b^0 \to J/\psi p K^-$ in the $J/\psi p$ system. In 2015 two pentaquark states were discovered, namely $P_c^+(4380)$ and $P_c^+(4450)$. Later, in 2019 with more data becoming available, it was realized that the second state is actually two distinct states $P_c(4440)^+$ and $P_c(4457)^+$. These pentaquarks are also predicted to decay into $\Lambda_c^+ \overline{D}^{*0}$, thus, a pentaquark search in this system and observation or non-observation of pentaquarks will help to improve the understanding of these resonances and their production mechanisms which are predicted to be different from the mechanisms of production of conventional baryons. LHCb provides a unique and brilliant data set for studies of exotic hadrons, including pentaquarks.

# Analysis strategy

The analysis focuses on the candidates selected for the decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^{-1}$. The raw data from the LHCb detector is first centrally written out into various *stripping streams*. After selecting the relevant stream, a brief preselection on the data is performed to optimize the file size. Later, a full offline selection is performed involving a number of kinematic and fiducial cuts as well as using `Boosted Decision Trees` (`BDTs`) as a means of selecting the candidates according to particle identification (`PID`) variables. Parts of the selection are adopted from previous analyses of this channel [20] where the so called `D-from-B BDTs` [21] were constructed and optimized. Along with the data samples used in the analysis, Monte Carlo (MC) samples are also put through the same selection procedure. Fits are performed to the invariant mass spectrum of $\Lambda_c^+ \overline{D}^0 K^-$ in order to extract the yields of signal and background components. The region in the $\Lambda_c^+ \overline{D}^0 K^-$ mass with partially reconstructed $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ decay candidates is selected and `Extended Cone Closure` method is applied on the candidates to reconstruct the missing 4-momentum of a $\pi^0$ or $\gamma$ from the decay of $D^*$ meson. Efficiency corrected and background subtracted Dalitz plot $\mathrm{m}^2(\Lambda_c^+ \overline{D}^{*0})$- $\mathrm{m}^2(\overline{D}^{*0} K^-)$ is studied for the first time. A model independent study is performed using Legendre Moments and a CLs limit setting procedure is performed to obtain upper limits on the yields of three possible pentaquark states - $P_c(4312)^+$, $P_c(4440)^+$ and $P_c(4457)^+$. The limits are set at 95% confidence level.

[1] Charge conjugation is implied throughout this thesis.

# Chapter 1

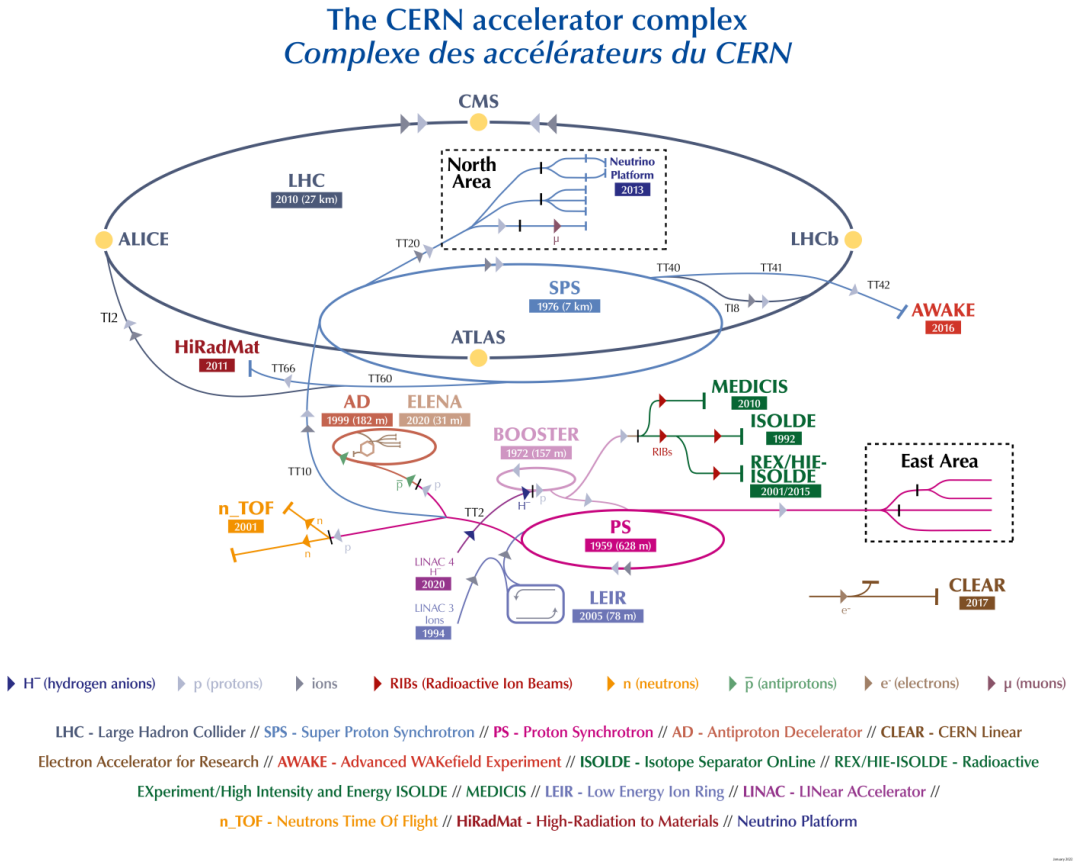
# The LHCb Experiment

In this chapter, an overview of the accelerator complex at CERN and the LHCb experiment is given. A short introduction and some of the information on the Large Hadron Collider (LHC) is provided. The Large Hadron Collider Beauty (LHCb) experiment and its sub-detectors are described in greater detail. LHCb's data management system is also introduced.

## 1.1 Accelerator Complex at CERN

In the year 1954, with the formation of CERN (European Organization for Nuclear Research) a small village called Meyrin was chosen as a place to host the largest science laboratory ever build. Since then, CERN expanded across the Swiss-French border and became the home to the largest complex of particle accelerators in the world. With time, whenever a bigger accelerator was built, the former one would serve the purpose of a pre-acceleration stage. This way, instead of simply shutting down once world leading, complex and expensive particle accelerators they were repurposed to work in an ever growing ecosystem of various stages of accelerators at CERN. Starting from linear accelerators `LINAC`, multiple booster stages are connected to incrementally ramp up the energy of proton or ion beams. At different stages, the beams are steered towards smaller fixed target experiments or injected into a higher-energy acceleration stage. Eventually, the particles are injected into the Large Hadron Collider where 4 big experiments LHCb, ATLAS, CMS and ALICE are located. A schematic diagram of the current (2022) view [1] of the particle accelerator complex at CERN is given in Fig. 1.1.

**Figure 1.1** *Accelerator complex at CERN with multiple* `LINAC` *linear accelerators for initial beam acceleration, booster stages for ramping up the energy of the beam and the LHC as a final stage where the highest center-of-mass energies of 13 TeV are achieved. Multiple fixed target and collider experiments are situated throughout the complex. Reproduced from [1]*



## 1.2 The Large Hadron Collider

The last addition to the CERN accelerator complex was the Large Hadron Collider (LHC) [22]. It was built in the tunnel $\sim$ 100m under the ground surface. The circumference of the LHC is 26.7 kilometers. This accelerator replaced the Large Electron Positron Collider (LEP) which the tunnel was initially built for. Protons (or ions) in the LHC are gradually accelerated to higher and higher energies up to center-of-mass energies $\sqrt{s} = 13$ TeV. To maintain the circular trajectory of the particle beams inside the LHC, a number of bender magnets are used. Since the particles do $\sim$ 11000 round trips per second in the LHC, the bending power of these magnets has to be large. The electromagnets create magnetic field of 8 Tesla. Only when superconductivity is achieved in the magnets, the immense current of 11000 amperes can be supplied to them to create this enormous magnetic field. To do this, they need to be cooled to 1.9 K using complex liquid helium cooling system. Moreover, after a full revolution in the accelerator, a particle needs to arrive to the acceleration point at the right time to be further accelerated. This makes synchronization of acceleration and circulating phases of the particle beam a real engineering challenge. Since the particles are injected at a lower energy of 400 GeV and are accelerated with each revolution inside the LHC, the bending power of the magnets has to adapt depending on the beam conditions. Magnets are continuously ramped up in power until the nominal energy of the beam is reached.

The LHC was designed to achieve center of mass energies $\sqrt{s}$ of 14 TeV [23]. This was done incrementally - 7/8 TeV in Run1 (years 2011,2012), 13 TeV in Run2 (2016-2018), and up to 14 TeV in Run3. The particles are injected in bunches. The number of bunches varies depending on the filling scheme. In Run2 $\sim$2500 bunches of protons would be circulating in the accelerator at the same time. The bunch spacing in is designed to be 25 ns. There are $\sim 1.15 \times 10^{11}$ protons per bunch. After the injection, bunches are accelerated until they reach nominal energy. Then the bunches are brought together at different interaction points (at the experimental caverns where the physics experiments are located). Data from collisions is recorded by ATLAS, CMS, ALICE and LHCb experiments. After some time the population of particles in the bunches is depleted and the beam is dumped to allow for a new fill.

## 1.3 The LHCb Detector

The LHCb Detector [24] is different in geometry to the other experiments at CERN, e.g. ATLAS, CMS and ALICE. While the first two are general purpose detectors, LHCb is constructed such that it detects particles in the forward direction. Many of these are the decay products of $B$ mesons - particles which have a bottom (or beauty) quark as one of their constituents. Since its construction, the LHCb detector has undergone a few stages of upgrades. The data used in this analysis is from the years 2015-2018 (Run2 of the detector). For this reason, descriptions of the detector and its sub-detectors are mainly those of Run2. In the time of writing detector is being commissioned for Run3, where large portions of the detector were heavily upgraded or completely replaced. LHCb provides excellent vertex resolution using VELO detector, described later in this chapter, Section 1.3.1. Physics analyses are aided by great particle momentum and mass resolution as well as great particle identification capabilities (PID) provided by LHCb tracking detectors and RICH (Cherenkov detectors) respectively [25]. In short, LHCb detector would be characterized as a single arm forward mass spectrometer. A schematic diagram of the entire detector is provided in Fig. 1.2. The LHCb is covering a pseudo-rapidity region of $2 < \eta < 5$. The pseudo-rapidity $\eta$ is a spacial coordinate defined as

$$\eta = -ln\left(tan\left(\frac{\theta}{2}\right)\right) ,$$

where $\theta$ is the polar angle measured from the positive beam axis.

The main physics goal of the experiment is to look for *Beyond Standard Model* (BSM) heavy flavor physics, focusing on CP violation and rare decays. There is also a rich hadronic spectroscopy program. During Run1 and Run2 of data taking, LHCb has collected the data equivalent to 9 $\text{fb}^{-1}$ of integrated luminosity. LHCb is accredited with recent discoveries of various excited meson and baryon states. These include observations of five new narrow $\Omega_c^0$ states [26], observation of new $\Xi_c^0$ baryons [27] as well as the first observation of a doubly charmed baryon $\Xi_{cc}^{++}$ [28]. New discoveries of tetraquarks [29], [30] and pentaquarks [15], [16] were also reported by the LHCb collaboration.

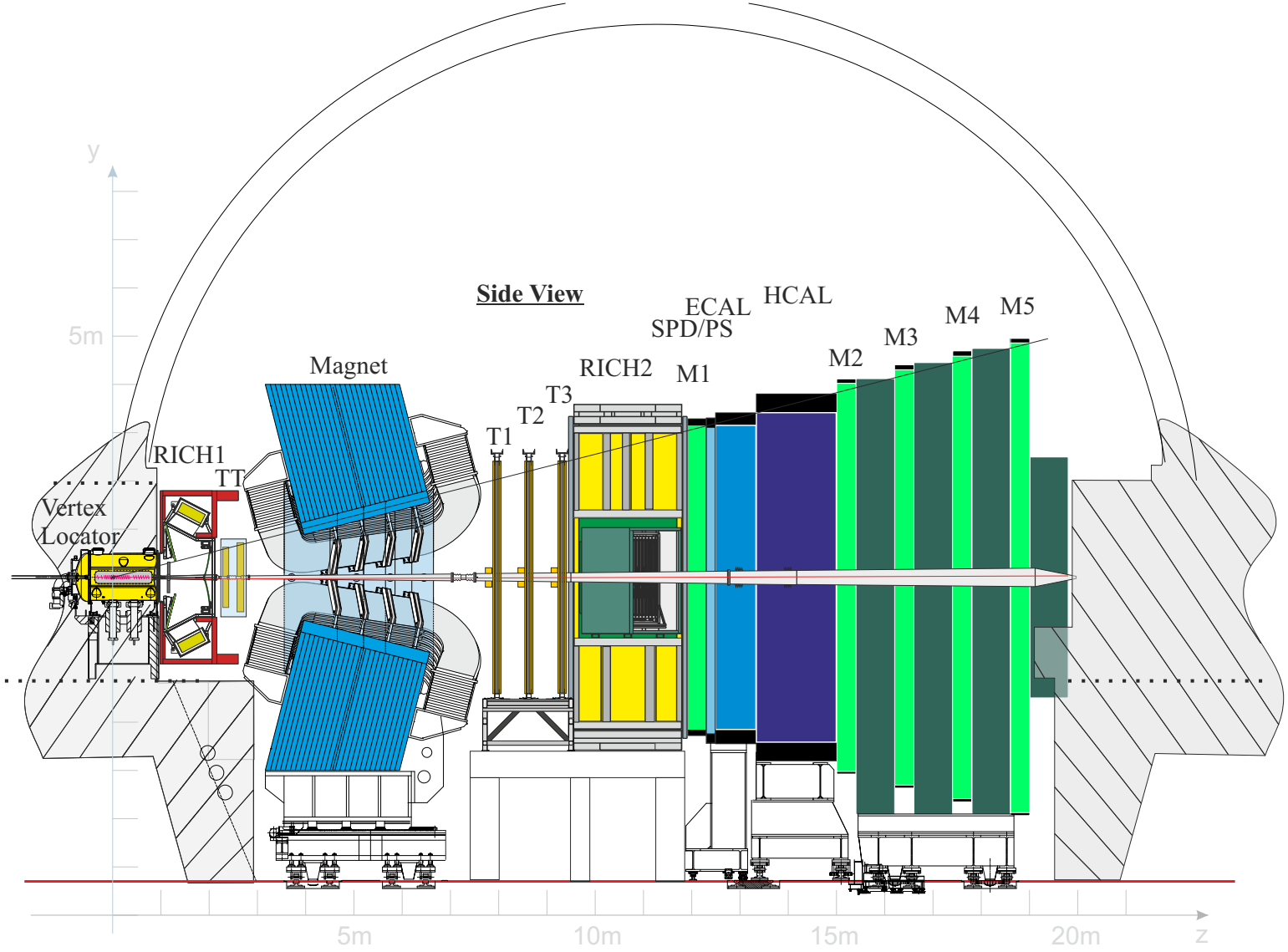





**Figure 1.2** *A schematic diagram of the LHCb detector. Starting from the bottom, where the interaction point is, various sub-detectors are situated upstream. Different particles are detected or their parameters recorded at different points in the detector. Reproduced from [2].*

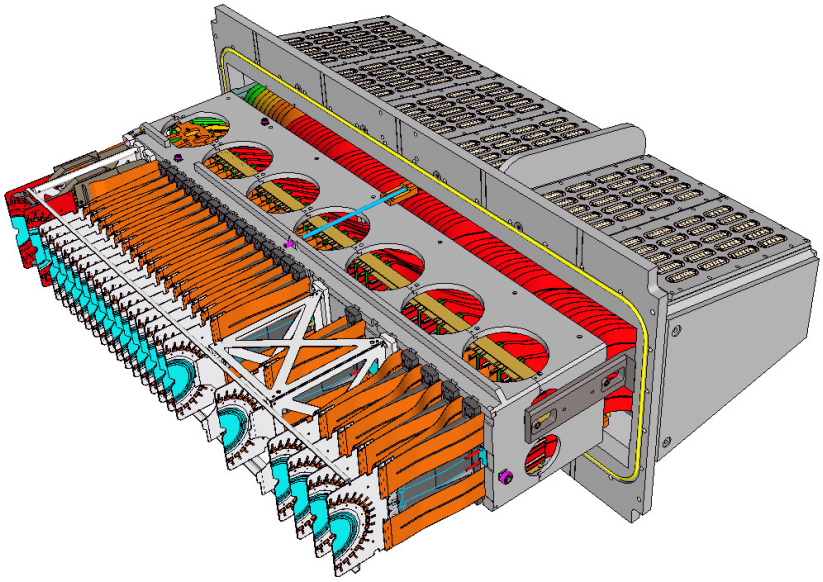


**Figure 1.3** *One half of the VErtex LOcator (VELO). VELO is used for accurate determination of primary and secondary vertices and is instrumental to the physics program at LHCb.*

### 1.3.1 VErtex LOcator (VELO)

The Vertex Locator [31] is an integral part of LHCb detector. This is a silicon strip detector with the main goal of providing accurate vertex information for particle decays. To reconstruct primary and secondary vertices of the decays at high resolution, VELO needs to be moved very close to the particle beam in the accelerator. The issue arises from the fact that at injection of the proton (ion) beam into the LHC, the radius of the beam is large and quite strongly fluctuating. Only after the beam is cleaned and squeezed (focused) and *stable beam* status is declared, the VELO moves into its position and the physics data collection can begin. To do this, VELO is split into two parts of multiple sensors which are then mechanically brought together around the beam. If the VELO sensors were exposed to the beam directly, there is would be a high probability that radiation damage would occur. VELO operates in very extreme environments. Its proximity to the beam pipe means that all the sensors and electronics need to be able to withstand high doses of radiation during the entire life cycle of the detector $\mathcal{O}$ (10 years). When designing and constructing the detector, radiation hardness was one of the main points of focus.

### 1.3.2 Ring Imaging Cherenkov Detectors (RICH)

Accurate identification of pions and kaons as well as other particles originating from B hadron decays is of most importance in LHCb. For this reason, two *Ring Imaging Cherenkov Detectors* (RICH1 and RICH2) are employed [32], [33]. The former is located upstream of LHCb dipole magnet (described later in this chapter, Section 1.3.3) and the latter is located downstream of the magnet. The reason for such positioning of the detectors is because the RICH1 is intended for identification of low momentum particles in the entire LHCb acceptance range, while the RICH2 is intended for identification of high momentum particles in the range where they are observed. The schematic diagram of both detectors is provided in Fig. 1.4. The need to use two detectors arises from the fact that particle identification needs to be handled in a broad momentum range (from 1 GeV [1] to 100 GeV) while also spanning the entire LHCb acceptance range. Cherenkov angles are computed by registering the Cherenkov light emitted in the radiators of RICH detectors. Optical system involving spherical and flat mirrors is then used to channel this light outside of the LHCb acceptance where it is detected by the Hybrid Photon Detectors (HPD) [4]. The optical system is designed such that the components can still be laid out in the limited space available. The tilt of the spherical mirrors is chosen such that the image formed on the flat mirrors is reflected. This way the flat mirrors can be positioned outside of the acceptance of LHCb tracking detectors. Geometry is further constrained by magnetic shielding. The highest allowed magnetic field strength is 3 mT to avoid deforming of the sub-detectors and their supporting structure. Since spherical mirrors are located within LHCb acceptance, charged particles and photons traverse the material. To reduce the material budget, carbon fiber reinforced polymer (CFRP) substrate is used and the supports of the mirrors are placed outside LHCb acceptance. This results in a reduction of radiation length $X_0$ from about 8% to less than 2%. The Cherenkov angle is defined as

$$cos(\theta_c) = \frac{c}{nv}\,,$$

where $c$ is the speed of light in the vacuum, $n$ is the refractive index of the medium and $v$ is the speed of a charged particle.

[1] The use of natural units $\hbar = c = 1$ is implied here and throughout the thesis.

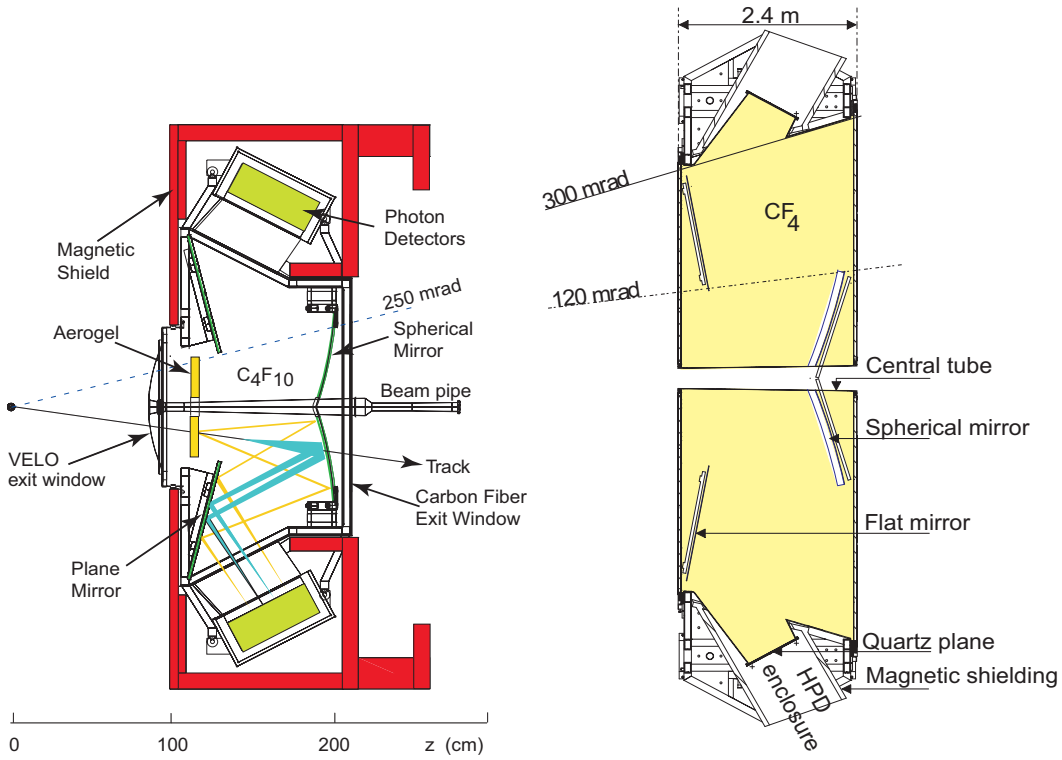



**Figure 1.4** *A schematic diagram showing the layout of the Ring Imaging Cherenkov Detectors (RICH1 and RICH2). Their geometry is such that particles in the entire LHCb acceptance are subjected to particle identification procedure. Reproduced from [3].*

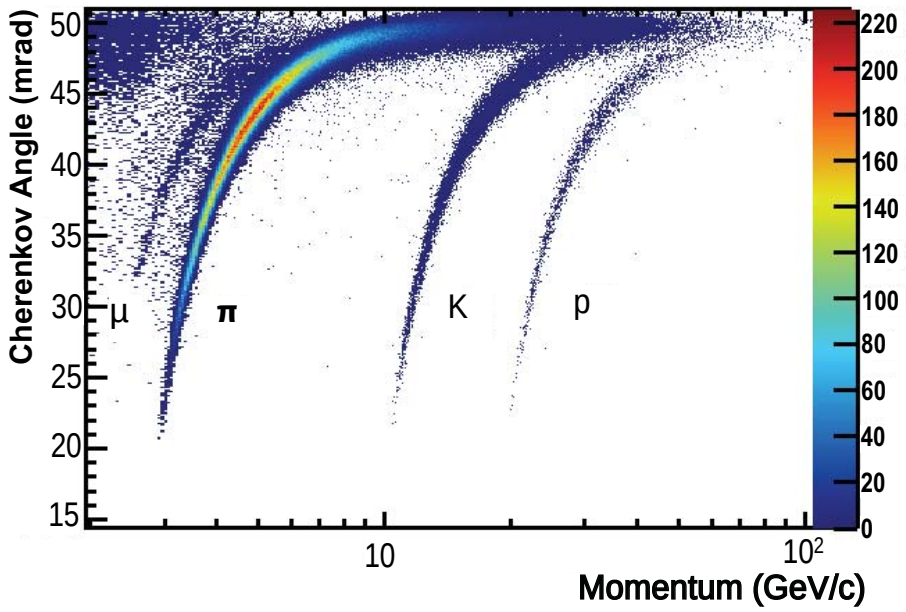



**Figure 1.5** *Reconstructed Cherenkov angle as a function of the track momentum in the $C_4F_{10}$ radiator. Reproduced from [3].*

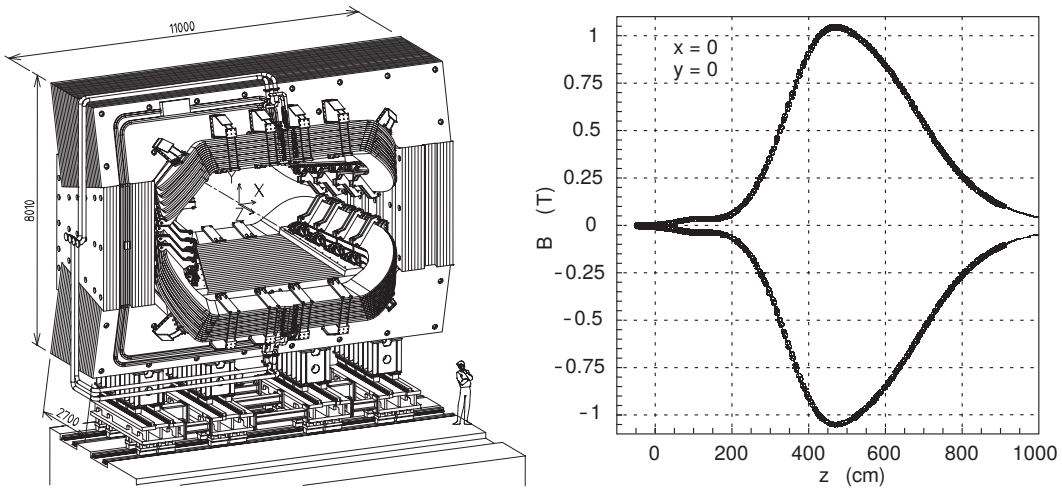



**Figure 1.6** *A schematic diagram of the LHCb dipole magnet with a person next to it for an indication of scale on the left. The magnetic field profile in the z-y plane on the right. Reproduced from [4].*

### 1.3.3 Dipole Magnet

An integral part of the LHCb detector is its dipole magnet. It is a *warm iron yoke* magnet, providing a magnetic field of about 4 Tm [34]. The magnet is used to bend the trajectories of charged particles traversing through the detector. The radius of curvature of the reconstructed track provides the information on the momentum of the particle. A schematic diagram of the magnet as well as its magnetic field profile is given in figure Fig. 1.6. To reduce the systematic effects of operating magnet at a given polarity, which is especially important in CP violation studies, the polarity of the magnet is regularly reversed. The window of a magnet is designed such that the magnetic field is applied in the entire angular acceptance region of the detector while at the same time minimizing its effect on VELO, RICH and tracking stations nearby. The magnet also has an effect on the proton beam in the LHC beam pipe, this effect is compensated further in the path of the beam.

### 1.3.4 Tracking Stations

Different types of tracks are recorded by the LHCb tracking system. They are: VELO tracks - where particle trajectory is recorded by the VELO sub-detector, but no signatures in any of the remaining tracking stations are present, upstream tracks - where particle is recorded both in VELO and the Upstream Tracker but leaves the detector acceptance before reaching the downstream sub-detectors,

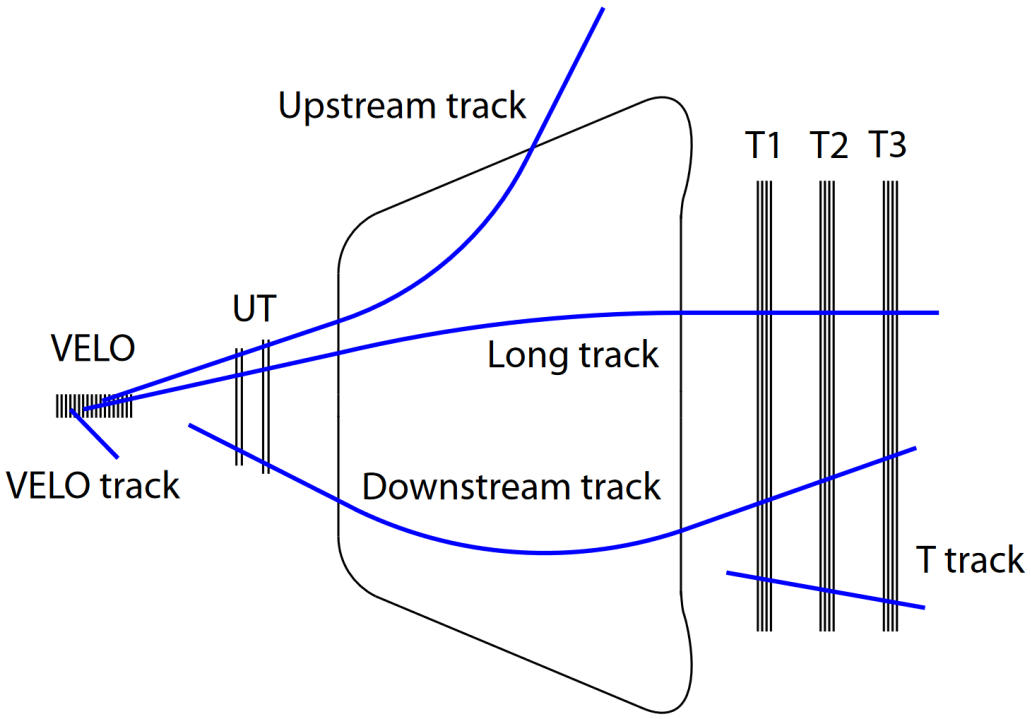



**Figure 1.7** *A diagram showing the different types of particle tracks in LHCb. These tracks are recorded by dedicated components of the tracking system.*

downstream tracks - where no signatures in VELO are recorded, T tracks - which are only recorded by the outer tracker stations and long tracks - where a particle traversing the detector leaves a signature in each of the elements of the tracking system.

The tracking system in LHCb is composed of two main parts. The inner tracker `IT` [35] and the outer tracker `OT` [36]. The components of these systems can be identified in the detector schematic diagram in Fig. 1.2 as `TT` (*Tracker Turicensis*) and tracking stations `T1` through `T3`. The sensors in the individual detection layers of the tracking stations for both `TT` and `OT` are arranged in an alternating pattern of orientations from vertical to slightly inclined in both directions. This is done to increase the momentum resolution of the reconstructed tracks. While `TT` covers the entire LHCb acceptance, the `IT` only covers a small region around the beam pipe where the particle flux is largest. Since the tracking stations `T1` through `T3` are located downstream of the magnet, the particle trajectories here are strongly diverging. A large area needs to be covered to still detect particles in the detector acceptance. In order to save costs and cover large area, the rest of

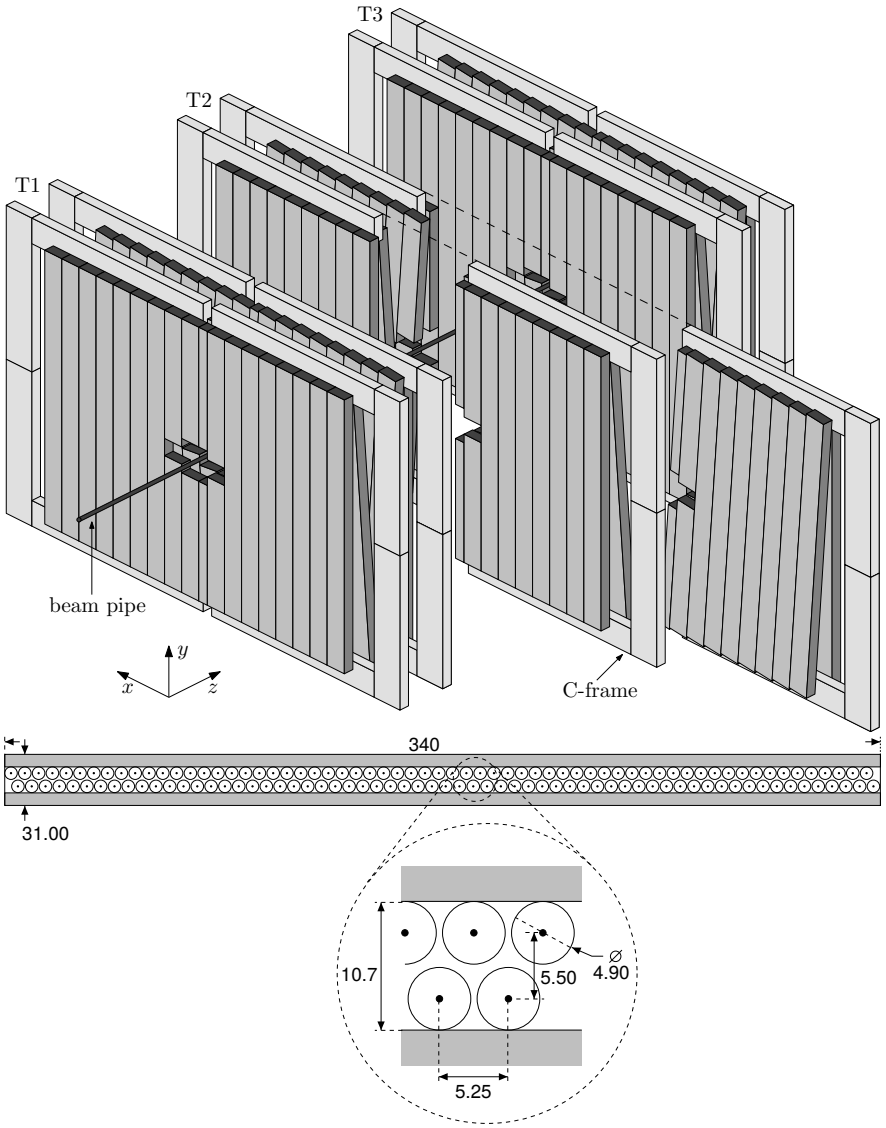



**Figure 1.8** *A schematic diagram showing the general layout of the OT stations as well as a cross section of one layer. Reproduced from [5].*

the OT is a drift tube gas detector with a total active area of 597 cm x 485 cm. Hit position in the OT is determined by measuring the arrival time of the signals from the detector readout. The schematic diagram of the OT and the cross section of an individual layer of drift tubes is shown in Fig. 1.8.

### 1.3.5 Calorimeter System

The LHCb calorimeter system serves the purpose of additional particle identification and fast transverse momentum measurements. The main goal of this system is to differentiate electrons, photons and hadrons. The signals read

out in the calorimeters provide fast transverse energy $E_T$ measurements to the LHCb trigger system. The calorimeter system is described in detail in [6]. A schematic diagram showing the layout of the calorimeter system is provided in Fig. 1.9.

The Scintillating Pad Detector (SPD) and the Pre-shower Detector (PS) are located downstream of RICH2 and are a part of LHCb calorimeter system. Both of these detectors consist of scintillating pads (cells) with a scintillating fiber coil grooved inside for better light collection. Multi-anode photo multipliers are used for the readout of optical signal. Information is registered, depending on the comparison of the energy deposited in the cell with a threshold. This helps with distinguishing charged particles from the neutral ones and, in association with the energy measured in the corresponding pre-shower detector cells, helps with photon and electron identification.

The Electromagnetic Calorimeter (ECAL) in LHCb follows a *shashlik* geometry. An alternating pattern of absorber material (lead plates) and scintillator is used to register electromagnetic showers arising from material interaction of charged particles. In ECAL, scintillating light is delivered by wavelength shifting fibers and registered by individual photo multiplier tubes.

The Hadronic Calorimeter (HCAL) is used to measure the energies of hadronic showers. The HCAL also has a layered structure of scintillator and absorber material [37].

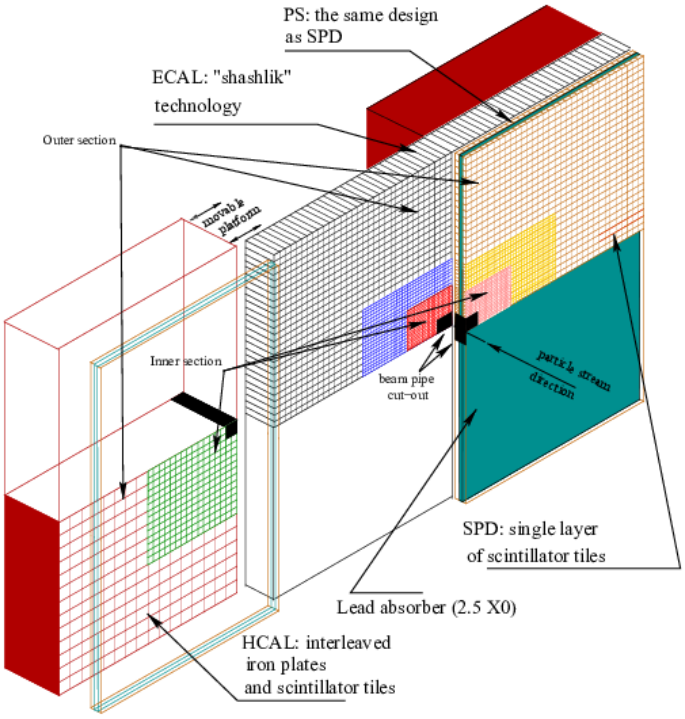



**Figure 1.9** *Layout of the LHCb calorimeter system. Different layers are designed to produce and records electromagnetic or hadronic showers. There are also Pre-shower and Scintillating Pad detectors. Reproduced from [6].*

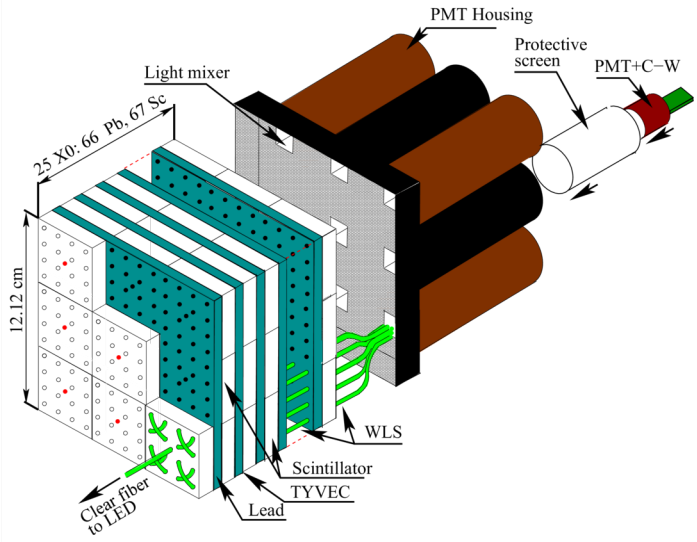



**Figure 1.10** *A single cell of the Electromagnetic Calorimeter (ECAL). The layout if following a shashlik geometry - alternating layers of scintillator and absorber material are reminiscent of a skewer of meat. Reproduced from [6].*

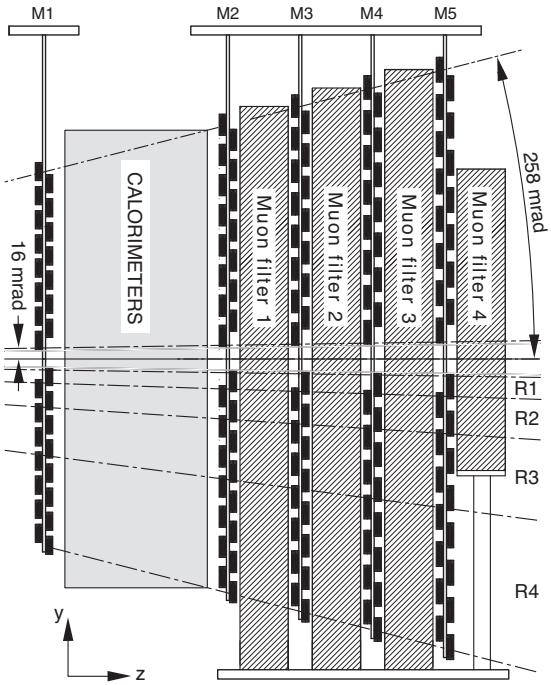



**Figure 1.11** *A schematic diagram of the muon system. Reproduced from [7].*

### 1.3.6 Muon Detectors

The sub-detector system situated most downstream of the collision point is designed to detect muons. As the majority of muons traverse the other sub-detectors with no loss, a dedicated system is used. It is described in detail in [7]. As shown in Fig. 1.11, the muon system is comprised of five stations, one of which is placed upstream of the calorimeters and the other four - downstream. The stations are separated by 80 cm thick layers of iron. Muons with momentum larger than 6 GeV will traverse the entire muon system and leave a signature in all of the five muon stations. Detection of muons in each of the stations is performed by multi-wire proportional chambers (MWPC). The very first station M1 is also equipped with triple gas electron multiplier (triple GEM) in the region with highest rates of traversing particles. In Run1 and Run2 of the LHC the muon system was part of the H0 hardware trigger, being able to distinguish muons from other particles and improve $p_T$ measurements.

### 1.3.7 Summary of Sub-detector Parameters

Some of the parameters for the LHC and the different sub-detectors of the are provided in Table 1.1.

## LHC

| Parameter | Value |
|---|---|
| Type of accelerator | Synchrotron |
| Circumference | 27 km |
| Beam energy | 13 TeV |
| Number of bunches colliding at IP8 (LHCb) | 2500 |
| Bunch spacing | 25 ns |
| Bunch intensity | $1.2 \times 10^{11}$ ppb |
| Emittance | 2.2 $\mu$m |
| Crossing angle | 300 - 260 $\mu$rad |
| Peak luminosity | $2 \times 10^{34}$ cm$^{-2}s^{-1}$ |
| Peak pileup | 60 |

## LHCb

| Parameter | Value |
|---|---|
| Angular acceptance | $2 < \eta < 5$ |
| Readout rate | 40 MHz |
| Weight | 5600 t |
| Dimensions (L x H x W) | 21 m x 10 m x 13 m |

## VELO

| Parameter | Value |
|---|---|
| Type of sensor | Silicon strip |
| Transverse resolution | 13 $\mu$m |
| Longitudinal resolution | 71 $\mu$m |
| IP resolution | 15 $\mu$m |
| Decay time resolution | $\sim$ 50 fs |
| No. modules | 42 |
| No. strips per sensor | 2048 |
| No. channels | 172 000 |
| Readout rate | 1 MHz analogue |
| Data rate | 150 Gbit/s |
| Operating temperature | -10°C |

## Magnet

| Parameter | Value |
|---|---|
| Type of magnet | Warm yoke dipole |
| Bending power | $\sim$ 4 Tm |

## RICH

| Parameter | Value |
|---|---|
| Type of sensor | Cherenkov detector |
| Radiator Material | $C_4F_{10}$ (RICH1) |
| | $CF_4$ (RICH2) |
| Momentum acceptance | 2 GeV - 60 GeV (RICH1) |
| | 15 GeV - 100 GeV (RICH2) |
| Optical layout | Vertical (RICH1) |
| | Horizontal (RICH2) |
| Mirror radius of curvature | 2700 mm |
| Beam pipe clearance | 10 mm |

## Tracking System

| Parameter | Value |
|---|---|
| Spacial single hit resolution | $\sim 53$ $\mu$m (TT and IT) |
| | 205 $\mu$m |
| Track reconstruction efficiency | $\sim 95\%$ |
| Momentum resolution | 0.5% - 0.8% |

## Calorimeter System

| Parameter | Value |
|---|---|
| Number of channels | 2 x 6016 (SPD/PS) |
| | 6016 (ECAL) |
| | 1488 (HCAL) |
| XY dimensions | 6.2 m x 7.6 m (SPD/PS) |
| | 6.3 m x 7.8 m (ECAL) |
| | 6.8 m x 8.4 m (HCAL) |
| Depth in z direction | 180 mm (SPD/PS) |
| | 835 mm (ECAL) |
| | 1655 mm (HCAL) |
| Radiation length | 2.5 $X_0$ (SPD/PS) |
| | 25 $X_0$ ECAL |
| Nuclear interaction length | 0.1 $\lambda_{int}$ (SPD/PS) |
| | 1.1 $\lambda_{int}$ (ECAL) |
| | 5.6 $\lambda_{int}$ (HCAL) |

| Muon system | |
|---|---|
| **Parameter** | **Value** |
| No. detector stations | 5 |
| Total area coverage | 435 $m^2$ |
| No. chambers | 1380 |
| No. channels | 122 000 |
| Iron absorber thickness | 80 cm |
| Efficiency | $\sim$ 99% |
| Resolution | 4 x 10 $mm^2$ - 150 x 180 $mm^2$ |

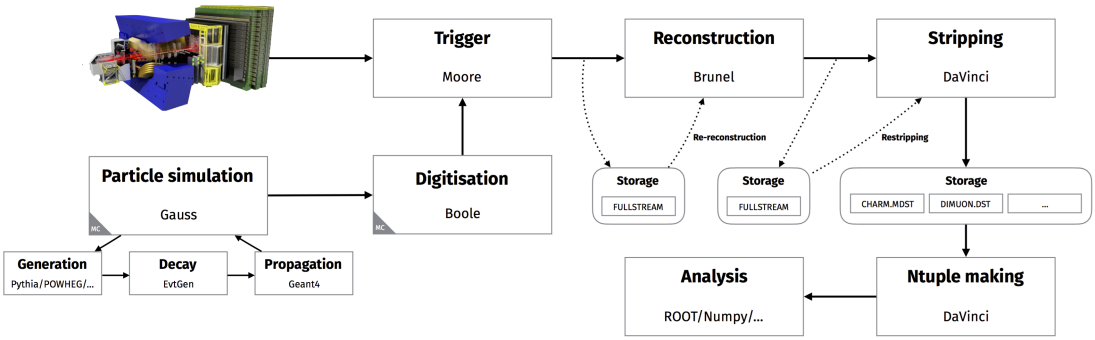
**Table 1.1** *The summary of the technical specifications of different sub-detectors within LHCb.*

## 1.4 Data Processing at the LHCb Experiment

The data flow in the LHCb experiment in described in detail in [38]. The LHCb detector output is passed through multiple stages of software based filters. The trigger is based on MOORE framework, followed by reconstruction algorithms implemented in BRUNEL and by stripping and filtering steps written within DAVINCI framework. Monte Carlo simulation workflow is also used as an input to the trigger where the detector material interactions, resolution and physical particle decay processes mimic the output of the LHCb detector very closely. The schematic data flow chart is given in Fig. 1.12.

LHCb is storing vast amounts of data, for Run1 and Run2 combined the amount of data stored is about 10 Petabytes. This data is staged on disk for quick and easy access. Legacy data is stored on tape for archiving. Upon request, the data from tape storage can be temporarily staged on disk again. Raw detector output in terms of data rate is about 1TB/s. Recording and storing such amounts of data is of course not feasible. For this reason, a number of filtering steps are taken to reduce the amount of data to a more manageable figure. Only good quality data, satisfying a number of requirements is maintained for the next filtering step and eventually stored in appropriate storage media. The logic of selecting the events for the next step in the data flow or discarding them is called `Trigger`. In Run1, the LHCb had two types of `Trigger` stages - the hardware `L0 Trigger`. The unwanted data in `L0` stage would be discarded by a FPGA circuit at a very

**Figure 1.12** *LHCb data flow. Starting from the digital output or the simulation of the detector multiple data filtering steps are undertaken to select good quality event of physical interest. The data size is also reduced greatly with each step so only the relevant and needed data is saved to storage.*




fast rate. This was followed by two further software based stages, the `High Level Trigger` `HLT1` and `HLT2`, where the event would be partially reconstructed and evaluated for its quality. This is shown in Fig. 1.13. For comparison, Run2 trigger scheme is also given.

The `Trigger` step in the data flow is followed by `Reconstruction`. Here, a full event topology is reconstructed, using input from all different sub-detectors. Particle tracks are fitted and mass hypotheses assigned. The quality of fitted tracks and vertices is evaluated. This is then followed by a `Stripping` step, where reconstructed events are categorized in broad groups according to the physics involved called stripping streams. The `Stripping` is a result of an optimization between storage and computing resources. If everything was recorded into one stripping stream, because of no repetition in storing the same events multiple times the storage space requirements would be minimized, but for every analysis this entire data set would need to be read in. Computing for a single analysis would take a lot more time. On the other hand, if the stripping streams were too finely defined, a lot more storage would be needed and the same events would be stored multiple times in different streams. For this reason, the stripping streams are still quite broadly defined, categorized by a broad physics use case. The list of stripping streams is provided in Table 1.2.

The output of the stripping step is the files in `DST` (`Data Summary Tape`) or `MDST` (`micro Data Summary Tape`) format. Internally in LHCb a specific `stripping`

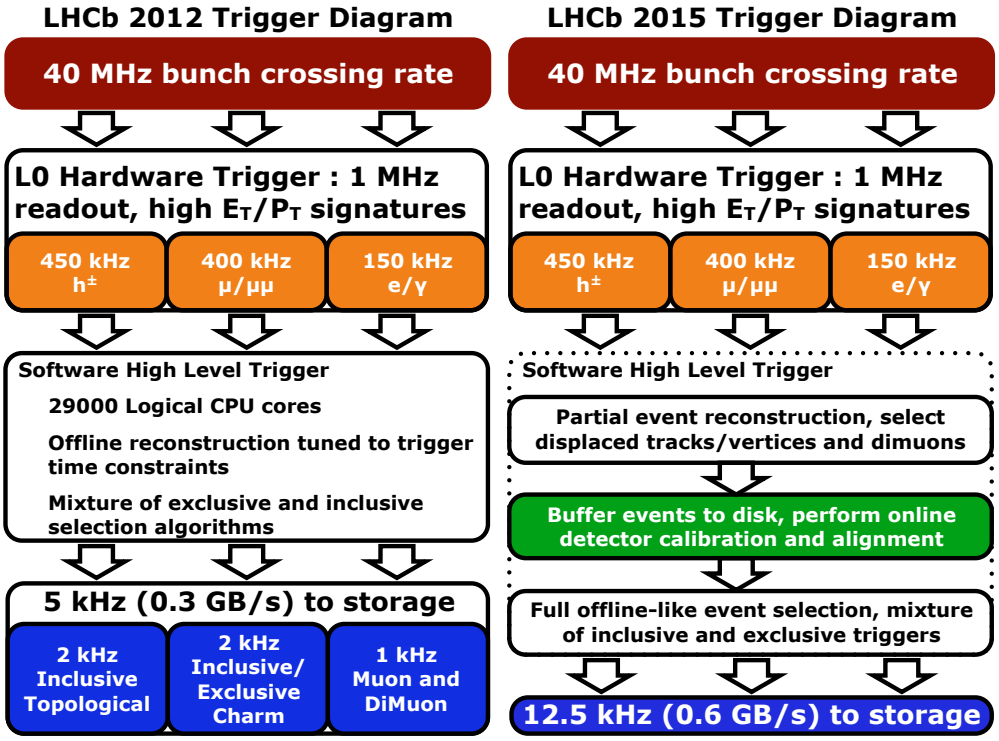



**Figure 1.13** *Comparison of LHCb Run1 and Run2 Trigger schemes. Each step reduces the data volume. Eventually only the needed events are saved in storage. Reproduced from [8].*

| Stream name |
|---|
| Bhadron |
| BhadronCompleteEvent |
| Calibration |
| Charm |
| CharmCompleteEvent |
| CharmToBeSwum |
| Dimuon |
| EW |
| Leptonic |
| Minibias |
| PID |
| Radiative |
| Semileptonic |

**Table 1.2** *LHCb data stripping streams. A broad categorization of physics data collected by the LHCb detector.*

line is run to filter the data further. The stripping line is another collection of requirements (cuts) which are not only used to reduce the size, but also to select physical processes relevant to a specific physics data analysis. There are hundreds of stripping lines already written.

Accessing, storing, moving and copying petabytes of data is a huge computational challenge. DIRAC Interware [39] is used to facilitate all the computing tasks. DIRAC is a software framework enabling the interaction between a large community of users with various distributed computing resources. It is an open source project written in Python [40] programming language. LHCbDIRAC [41] is an extension to DIRAC framework designed to work specifically for LHCb. DIRAC was first developed at LHCb and the main maintainers of DIRAC are members of LHCb collaboration. DIRAC is used for all distributed computing needs within LHCb, such as grid computing, private clusters, batch computing, clouds, storage, data management and catalogs. One of the data catalogs available with LHCbDIRAC via the web interface is DIRAC Bookkeeping. In this catalog, all of the data collected with LHCb can be browsed and the metadata accompanying the data files can be explored. A desired data set can be browsed via its Bookkeeping Path - the identifying entity for a number of data files from a specific data taking run and with specific detector conditions.

# Chapter 2

# Theoretical Background and Motivation

In this chapter, a brief overview of the theory topics from particle physics, relevant to parts of the analysis described in this thesis is given. First, the *Standard Model* (SM) of fundamental particles and interactions is introduced. Then a more detailed look is taken into *Quantum Chromo Dynamics* (QCD) - a field of particle physics governing the interactions of quarks and hadrons. An overview of the decay kinematics, as well as Dalitz plot analysis is given. The theory of Legendre Moments is presented which is relevant to Section 4.14. This chapter is not intended to be a complete description of the theory topics involved, but rather an introduction to the main concepts, with references provided to further reading as well as some historical outline of how the field was developed.

## 2.1 The Standard Model of Elementary Particles and Interactions

The twentieth century will be remembered for, among other things, the advent of the field of elementary particle physics. Starting from ray physics where experiments were performed on cathode rays, a more in depth theoretical understanding was needed to explain various observed processes. Some of them, completely unseen before. Following the work of now-famous physicists Niels Bohr, Erwin Schrödinger, Werner Heisenberg, Max Born, Paul Dirac and others, the field of quantum mechanics was born. The most complete mathematical

description of the phenomena occurring between elementary particles was called the *Standard Model* of elementary particles and interactions [42, 43]. It is already clear that the *Standard Model* is not complete, but it is the best description of natural processes at a fundamental scale devised by the humanity to date. The *Standard Model* is a Lorentz Invariant Field Theory.

Three of the four known forces are included in the *Standard Model*. The so called strong, weak and electromagnetic interactions are described [44, 45]. Their underlying symmetry groups are $SU(3)$, $SU(2)$ and $U(1)$. The fourth force, gravity, is the oldest force know to humanity, but the efforts to find a description for it which would fit into the *Standard Model* have so far been futile. The difficulty arises due to its very small strength compared to the other forces. In elementary particle interactions, the effects of gravity are negligible. Finding a description of gravity which would work at the scale of fundamental particles is one of the biggest goals of modern science. In the *Standard Model* the forces are propagated (or mediated) by force-carrier particles called *gauge bosons*. The interactions of the *bosons* with matter particles (called *fermions*) are responsible for all of the processes which exist in nature. Fermions are further classified into *quarks* and *leptons*. Both, quarks and leptons are arranged into 3 groups called *generations*. Particles from different generations are very much alike, the only notable difference is their mass, which is increasing from *generation I* to *generation III*. The *bosons* can be separated into *gauge bosons*, namely $W^+$, $W^-$, $Z$, $\gamma$ and the gluon, and the only *scalar boson*, the *Higgs*. The particles of the *Standard Model* and some of their parameters are shown in figure Fig. 2.1.

The force in the *Standard Model* which is familiar to most is the electromagnetic force. It is responsible for all electric and magnetic phenomena, from the current flowing in electric circuits to the materials being magnetized and attracting one another. All of the effects of light are also a result of the electromagnetic interaction. The force carrier of electromagnetism is the photon. Another force, which is responsible for radioactive processes, is the weak force. It is propagated by massive $W^+$, $W^-$ and $Z$ bosons. The fact that these particles are so massive is the cause of small interaction strength characteristic to the weak force. The strongest force in the *Standard Model* is fittingly named - the strong force. It is propagated by particles called gluons. The self interaction of gluons is responsible for the immense strength of the strong force as well as other very interesting effects.

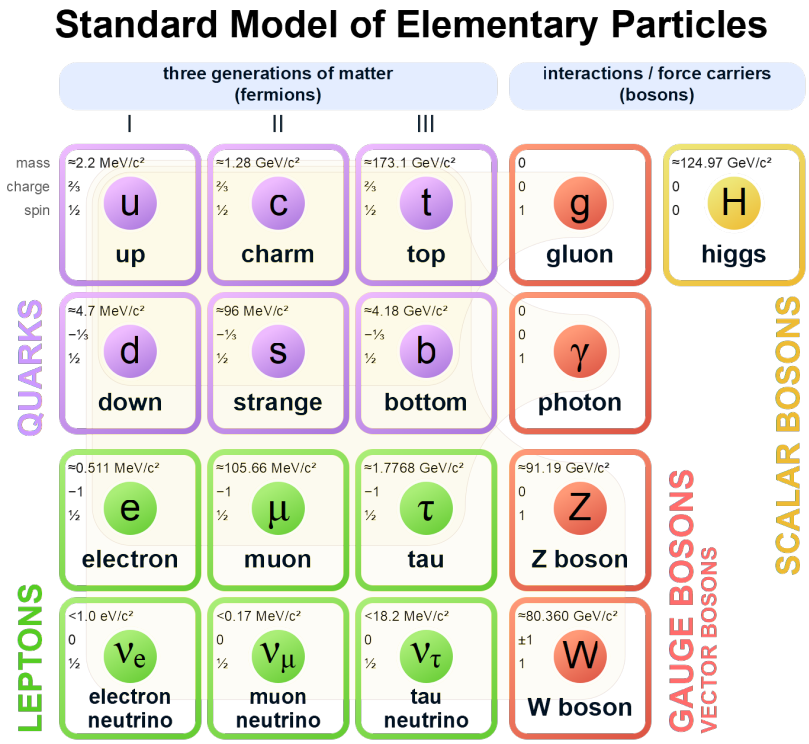



**Figure 2.1** *The matter and force carrier particles in the Standard Model of Elementary Particle Physics.*

The building blocks of matter susceptible the strong force are called quarks. They are bound together into mesons - made of a quark and an anti-quark and baryons - made of three quarks. The residual interactions of the strong force also bind protons and neutrons together in a nucleus of an atom.

## 2.2 Introduction to Quantum Field Theories

The afore-mentioned forces are mathematically formulated as so-called quantum field theories (or QFTs). The three field theories making up the standard model are quantum electrodynamics, electroweak theory and quantum chromodynamics. [46–49]. A very important mathematical construct, called the *Lagrangian* is a functional which can be used to obtain the equations of motion. *Lagrangians* are used in the mathematical description of the Standard Model. The idea of *Lagrangians* spans wider than just particle physics. If the equations of motion are known for an arbitrary system, its state can be determined at any point of time or with any given parameter values. The difference in particle physics from a classical

system is that the Lagrange function is replaced by Lagrangian density and the coordinates are replaced by continuous fields. The *Standard Model Lagrangian* is an extensive function containing many different terms. These terms can be broken down into the free part and the interaction part of the *Lagrangian*. The free part describes the dynamics of fields in the absence of interactions and typically contain the terms involving the fields and their derivatives but no higher powers of a single field or products between different fields. The interaction part of the *Lagrangian* describes the interplay between different fields. The strength of interaction is given by a coupling constant. The *Lagrangian* can be expressed as

$$\mathcal{L}_{SM} = \mathcal{L}_{gauge} + \mathcal{L}_{fermion} + \mathcal{L}_{Yukawa} + \mathcal{L}_{Higgs}\,,$$

where $\mathcal{L}_{gauge}$ terms describe the interactions and propagation of spin 1 *bosons* - $\gamma$, $W^+$, $W^-$, $Z$ and gluons. $\mathcal{L}_{fermion}$ or *matter kinetic terms* describe the propagation of *quarks* and *leptons* in a presence of a *boson* field. $\mathcal{L}_{Yukawa}$ describes the interactions of the Higgs *boson* with the fermions and are also responsible for the emergence of their mass. $\mathcal{L}_{Higgs}$ terms describe the process of spontaneous symmetry breaking through self-interactions of the Higs field.

## 2.3 Quantum Chromo Dynamics

In this section, a short introduction to *Quantum Chromo Dynamics* (QCD) is given. QCD is a field theory which describes how quarks interact with each other via the strong force. The birth of the strong force can be thought of to coincide with the discovery of the neutron [50]. It became clear that there must be a force within the nuclei of the atoms which is strong enough to counteract the electromagnetic repulsion between positively charged protons and also bind the neutral nucleons together. The strong force is a part of the *Standard Model* with the underlying $SU(3)_c$ symmetry group. This group is used to describe matter particles called quarks, they are the *fundamental representations of the* $SU(3)$ *group*, which are the building blocks of matter bound by the strong force. The propagators of the strong force are 8 fields of gluons. The idea of quarks first came about in 1964 to two scientists simultaneously, Murray Gell-Mann [51] and George Zweig [52]. They proposed that nucleons are not elementary, but in fact are constructed of smaller elementary objects called quarks. Zweig, in his paper put forward the name *aces* but Gell-Mann's term *quarks* was somehow better

accepted. The original quark model had three varieties of quarks, namely $u^{+2/3}$, $d^{-1/3}$ and $s^{-1/3}$ [9]. These are now referred to as light quarks, as later the heavier $c^{+2/3}$, $t^{+2/3}$ and $b^{-1/3}$ quarks were discovered.

### 2.3.1 The Color Charge

QCD is a generalization of a quark model proposed by Gell-Mann. In the naive quark model, bound states of three quarks of the same flavor (like (sss) state $\Omega^-$) were proposed. But the existence of a state like this would violate the Pauli Exclusion Principle. Another *hidden* quantum number of the quarks was introduced - a type of charge within the strong interaction. This solved the issue of the same-flavor quarks combining into to form a baryon. Different values of this charge allow the quarks to coexist in the same state. Three different values of this charge add together to produce a neutral object. Combining a particle which carries specific value of this charge with the particle which carries the opposite value of it also produces a neutral object. The name *color charge* became widely used because the behavior of charge in the strong interaction is similar to how people perceive colors. The values of the charge were thus dubbed "red", "green" and "blue". The values opposite to these were the "anti-colors" - "cyan", "magenta" and "yellow". Possible combinations of colors charges to form baryons and mesons are illustrated in Fig. 2.2.

### 2.3.2 Asymptotic Freedom and Confinement

Unlike the propagators of *Quantum Electro Dynamics* (QED), the gluons of QCD themselves carry the color charge. For this reason, there is also triple and quadruple gauge vertices of QCD which makes the dynamics within QCD vastly different from those of QED. This is illustrated in Fig. 2.3. In other words, gluons can self-interact and form gluon loops. From the nature of the strong interaction arises some very important notions of the dynamics of quarks and nucleons. Two important characteristics of the strong force are asymptotic freedom and confinement [9]. It is clear that the strong force must be scale (or energy) dependent. At small distances, the gluons bind the quarks into color neutral objects. On the other hand, at low energies - larger distances, the residual of this force is propagated by pions - they are themselves bound states of two quarks, but are light enough to adhere to an approximate symmetry - so called chiral

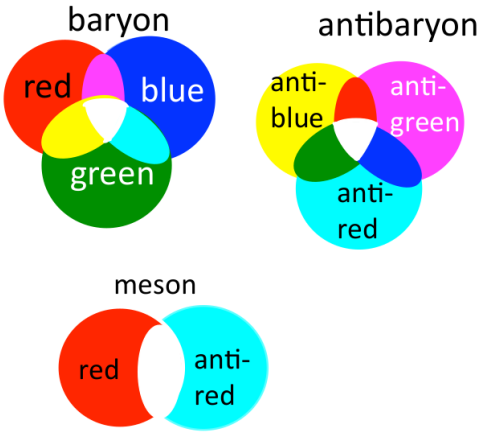



**Figure 2.2** *Color charges combining into color-neutral objects - mesons and baryons. Reproduced from [9].*

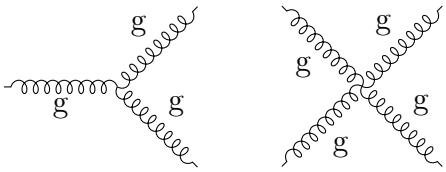



**Figure 2.3** *Diagram showing the 3-gauge and 4-gauge vertices of the QCD. Gluon self-interaction gives rise to important phenomena in the QCD such as asymptotic freedom, confinement, gluon jets, pomeron production [10].*

symmetry. At large distances the potential between two nucleons takes the form of Yukawa potential [53]. This is of paramount importance to the inner workings of the Universe, if the strong force was weaker or the energy gap between vacuum and the pion would be large, the atomic nuclei might not be able to form or their behavior would be vastly different.

s

The coupling constant of the strong force $\alpha_s$ is scale dependent. Unlike the QED potential which decreases in magnitude when moving away from the source, the strong force gets stronger by creating a gluon flux tube. For this reason, isolated quarks can never be observed. They are confined in color neutral objects. If enough energy is put into the system, rather than separating a quark from within a hadron, a quark - anti-quark pair is created from the vacuum and a new meson is emitted. This is illustrated in the Fig. 2.4. For very large energies, this process can happen over and over again and rather than having a single quark propagating in space, a jet of mesons and hadrons is created. This is also what

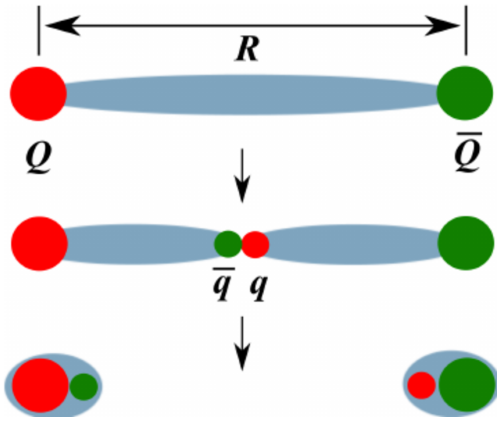



**Figure 2.4** *Color string breaking in QCD via formation of a gluon flux tube. Reproduced from [11].*

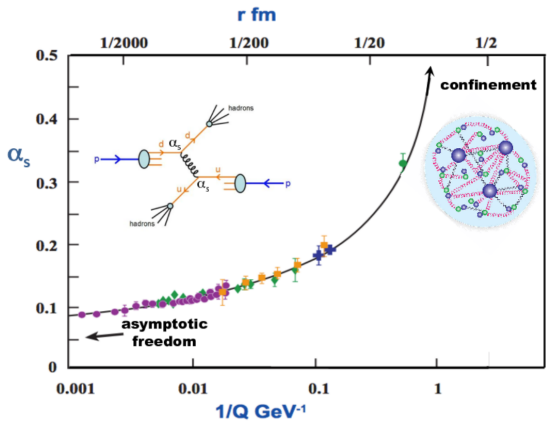



**Figure 2.5** *Strong force coupling constant $\alpha_s$ dependence on r. Reproduced from [12].*

happens at the Large Hadron Collider, when high energy partons react with one another. This behavior, at scales comparable to the size of a hadron makes QCD predictions very difficult. On the other hand, for very small distances, quarks are said to become *asymptotically free*. Within a proton or any other hadron, the quarks inhibit a certain region of space slightly separated from each other. A schematic diagram of these regimes is shown in Fig. 2.5.

### 2.3.3 The Eight-fold Way

The understanding of the spectrum of ground state baryons and mesons came with the formulation of the so-called *Eight-fold Way* by Gell-Mann and Ne'eman [54, 55]. The ground state hadrons can be arranged in flavor supermultiplets. Some of the states were predicted this way before they were experimentally

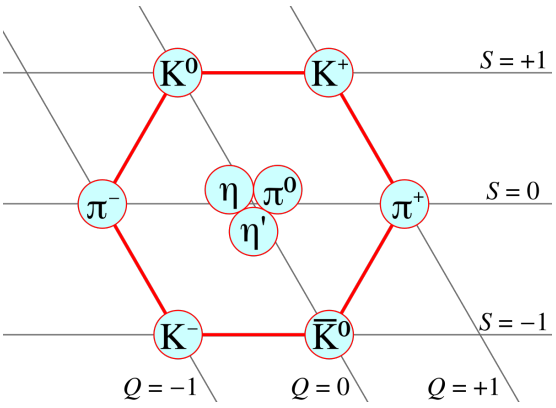



**Figure 2.6** *The ground state meson nonet.*

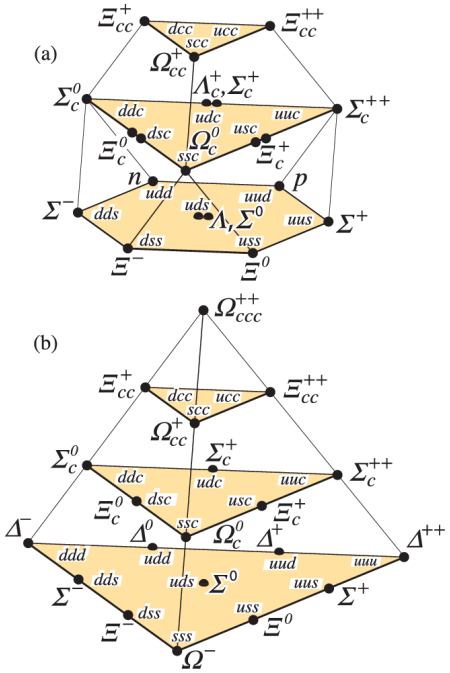



**Figure 2.7** *The ground state baryon multiplets. Reproduced from [13].*

observed. The multiplets are illustrated in Fig. 2.6 and Fig. 2.7 for mesons and baryons respectively. The spectrum of excited states of the QCD is very difficult to accurately predict, virtual gluon loop corrections become dominant, the coupling constant $\alpha_s$ is of order 1 for low-momentum-transfer interactions / distances close to the size of a hadron. In this region, the QCD regime is highly non-perturbative.

## 2.4 Exotic Hadrons

Already with the introduction of the quark model, states beyond the conventional mesons and baryons were proposed [56–60]. As long as the combined object is

color-neutral, states made of four, five, six and more quarks are not theoretically ruled out. There are also states proposed, consisting solely of gluons, called *glueballs* and hybrid states, where among the valence quarks, one or more valence gluons are present. Nowadays, such states are called the exotic hadrons. A search for them was paved with various controversies. For decades no such states were found.

The first exotic state was found back in 2003 by Belle experiment [61]. In later years it was confirmed by other experiments, including LHCb [62]. It is the $\chi_{c1}(3872)$ and is now mostly accepted as a tetraquark state. The first solid experimental discovery of a pentaquark state came from the LHCb in 2015 [15]. Since then, with more data becoming available, a plethora of tetraquark and pentaquark states were discovered, opening the door to very exciting new chapter in hadronic spectroscopy studies. Some of more prominent experimental results are highlighted later in this chapter, in Section 2.5 and Section 2.6. At the time of writing, no *glueball* states were experimentally confirmed. There are a few candidates for hybrid states, namely, the $\pi_1(1400)$ and $\pi_1(1600)$, they are both observed in $\eta\pi$ channels by COMPASS experiment [63]. The nature of tetraquark and pentaquark states is still not fully understood and there is a huge effort by the experiment and theory communities to construct the spectrum of such states and write down the theoretical models which would make sound predictions and agree with experiment. A summary of discovered exotic hadrons at the LHC is given in Fig. 2.8.

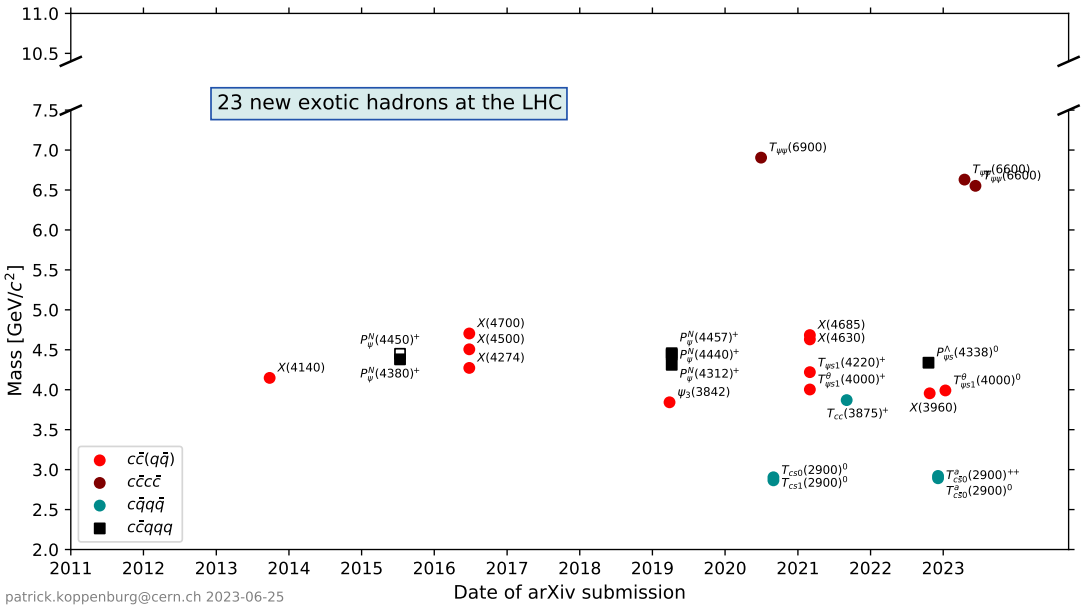



**Figure 2.8** *Summary of exotic states discovered at the LHC since the start of data taking in 2011. The onset of many experimental results coincide with the end of Run2 of the LHC in the beginning of 2019. Reproduced from [14].*

## 2.5 Experimental Results for Tetraquark States

The first experimental result for an exotic hadron came in the form of the discovery of $\chi_{c1}(3872)$ [61] charmonium-like state in $B^{\pm} \to K^{\pm}\pi^{+}\pi^{-}J/\psi$ decays by Belle collaboration. Since then, many tetraquark candidates were observed and discovered. Among the so-called $XYZ$ states: the $X_0(2900)$ and $X_1(2900)$ [64, 65] with the quark content of $(\bar{c}du\bar{s})$ in $B^{+} \to D^{+}D^{-}K^{+}$ decays, $Z_c(3900)$ [66–70] in $e^{-}e^{+} \to \pi^{+}\pi^{-}J/\psi$, $e^{-}e^{+} \to \pi D^{0}D^{*0}$ decays at Belle and BESIII, $Z_c(4020)$ [71, 72] in $e^{+}e^{-} \to \pi^{0}\pi^{-}h_c$ and $e^{+}e^{-} \to (D^{*}\bar{D}^{*})^{\pm}\pi^{\pm}$ decays, $Z_c(4050)$ [73] in $\overline{B}^{0} \to K^{-}\pi^{+}\ \chi_{c1}$ decays, $X(4100)$ [74] in $B^{0} \to \eta_c(1S)K^{+}\pi^{-}$ decays by LHCb, $Z_c(4200)$ [75] in $\overline{B}^{0} \to J/\psi K^{-}\pi^{+}$ decays by Belle, $Z_c(4430)$ [76–80] in B $\to$ K $\pi^{\pm}\psi'$ decays by LHCb and Belle and $R_{c0}(4240)$ [79] all with the quark content $(c\bar{c}ud)$. With the quark content $(c\bar{c}u\bar{s})$, the states $Z_{cs}(3985)$ [81, 82] in the decay $e^{+}e^{-} \to K^{+}(D_s^{-}D^{*0} + D_s^{*-}D^{0})$ at BESIII and, along with $Z_{cs}(4220)$ and $Z_{cs}(4000)$ states in the decay $B^{+} \to J/\psi\ \phi\ K^{+}$ by LHCb. Also reported are the $\chi_c$ or X states $\chi_{c1}(3872)$, $\chi_{c1}(4140)$ [83, 84], $\chi_{c1}(4274)$, $\chi_{c0}(4500)$, $\chi_{c0}(4700)$ [80], X(4630), X(4685) [82], X(4740) [85] all with the quark content $(c\bar{c}s\bar{s})$. Very exciting new discoveries are those of purely charm tetraquark state $(c\bar{c}c\bar{c})$ X(6900) [86], doubly charmed tetraquark $(cc\bar{u}\bar{d})$ $T_{cc}^{+}$ [87] and bottomonium $(b\bar{b}u\bar{d})$ states $Z_b(10610)$ and $Z_b(10650)$ [88].

## 2.6 Experimental Results for Pentaquark States

The first standing discovery of a pentaquark state was made in the decay $\Lambda_b^0 \to J/\psi p K^-$ in the $J/\psi p$ system. In 2015, two pentaquark states were discovered, namely $P_c^+(4380)$ and $P_c^+(4450)$. Both with the quark content of $(c\bar{c}uud)$. Sometimes called *hidden charm pentaquarks*. Later, in 2019, with more data becoming available the second state was resolved into two distinct states $P_c(4440)^+$ and $P_c(4457)^+$ [16]. A Breit-Wigner maximum likelihood fit to the invariant mass spectrum of $J/\psi p$ candidates is shown in Fig. 2.10. A comparison of the Dalitz plots from these two analyses is shown in Fig. 2.9. The main properties of the three discovered pentaquark states are given in Table 2.1. The evidence for the first pentaquark with strangeness, with the quark content $(c\bar{c}uds)$ was given in the analysis of $\Xi_b^- \to J/\psi \Lambda K^-$ decays [89]. In 2022, the first pentaquark with strangeness was discovered in the decay $B^- \to J/\psi \Lambda \overline{p}$ in the $J/\psi$ $\Lambda$ system [90]. Recently, evidence for $P_\psi^{N+}$ $(c\bar{c}uud)$ pentaquark state was found in the decay $B_s^0 \to J/\psi p K^-$ [91, 92]. Pentaquarks are also predicted to decay to $\Lambda_c^+ \overline{D}^{*0}$, thus, a pentaquark search in this system and observation or non-observation of pentaquark states will help us to improve the understanding of these resonances and their production mechanisms which are predicted to be different from the mechanisms of production of conventional baryons and mesons.

**Table 2.1** *Summary of the properties of the three pentaquark states discovered by LHCb. Mass, width and the yield of pentaquark candidates obtained by a maximum likelihood fit to data are reported. Reproduced from [16].*

| State | M [MeV] | $\Gamma$ [MeV] | $N_{P_c^+}$ |
|---|---|---|---|
| $P_c(4312)^+$ | $4311.9 \pm 0.7^{+6.8}_{-0.6}$ | $9.8 \pm 2.7^{+\ 3.7}_{-\ 4.5}$ | $657^{+153}_{-128}$ |
| $P_c(4440)^+$ | $4440.3 \pm 1.3^{+4.1}_{-4.7}$ | $20.6 \pm 4.9^{+\ 8.7}_{-10.1}$ | $836^{+256}_{-191}$ |
| $P_c(4457)^+$ | $4457.3 \pm 0.6^{+4.1}_{-1.6}$ | $6.4 \pm 2.0^{+\ 5.7}_{-\ 1.9}$ | $795^{+150}_{-140}$ |

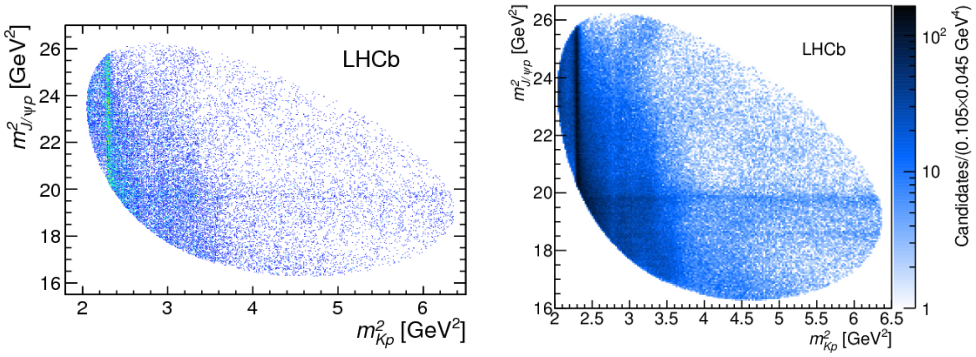



**Figure 2.9** *Comparison of two Dalitz plots for the decay $\Lambda_b^0 \to J/\psi p K^-$. LHCb Run1 data is used for the one on the left. This is the first observation of $P_c^+$ state (appearing here as a thin horizontal band) (reproduced from [15]). The one on the right is constructed using LHCb Run2 data sample (about 2 times larger than the first one). Here The horizontal bands are more pronounced. Reproduced from [16].*

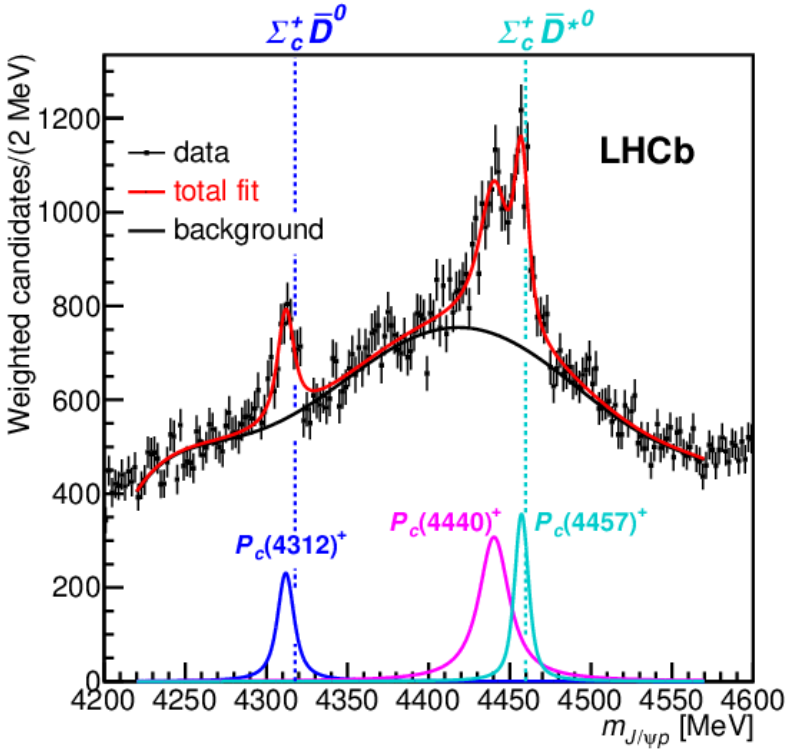



**Figure 2.10** *Fit to the $cos\theta_{P_c}$ - weighted $m(J/\psi p)$ distribution with three Breit-Wigner amplitudes and a sixth-order polynomial background. This fit is used to determine the central values of the masses and widths of the $P_c^+$ states. The mass thresholds for the $\Sigma_c^+ \bar{D}^0$ and $\Sigma_c^+ \bar{D}^{*0}$ final states are superimposed. Reproduced from [16].*

## 2.7 Theory Predictions for Pentaquarks

The combined quark content in the $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ channel (see the quark diagram in Fig. 2.12) is the same as the one where pentaquarks were first discovered (sometimes called the *discovery channel*) - $\Lambda_b^0 \to J/\psi p K^-$ (quark diagram in Fig. 2.13). Having the same quark content of the final state particles might already suggest that $P_c^+$ could decay through $\Lambda_b^0 \to P_c^+[\to \Lambda_c^+ \overline{D}^{*0}]K^-$. This can be looked at as a different configuration of the quarks forming a hadronic molecule or a tightly bound state. An important observation is that the discovered pentaquark states manifest themselves as excesses in the number of candidates in the invariant mass spectrum very close to meson-baryon thresholds. This is visible in $J/\psi p$ mass projection of [16] given here in Fig. 2.10. $P_c(4312)^+$ and $P_c(4457)^+$ lie 10 MeV and 5 MeV below the $\overline{D}\ \Sigma_c^+$ and $\overline{D}^*\ \Sigma_c^+$ meson-baryon thresholds, respectively.

Theory predictions (see Table 2.2) and above mentioned logic makes the analysis of channel $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ warranted. Different theoretical models discussing $P_c^+$ production and dynamics in $\Lambda_c^+ \overline{D}^{*0}$ system can be found in references [93–102]. For example, in [93], the effective range expansion and compositeness relations are used to compute the properties of the $P_c^+$ states decaying to $\Sigma_c^+ D^{*0}$ and $\Lambda_c^+ \overline{D}^{*0}$. The input from the experiment is used. The branching fraction predicted is similar to that of the $J/\psi p$ channel, but the model has large uncertainties and is composition dependent. In [97], the partial width of the $P_c(4312)^+$ as a $\overline{D}\ \Sigma_c^+$ molecule and the $P_c(4440)^+$ and $P_c(4457)^+$ as the $\overline{D}^*\ \Sigma_c^+$ molecules in $\Lambda_c^+ \overline{D}^{*0}$ system are predicted to be up to 3800 times larger than those for the $J/\psi p$ system. This would suggest that if the description is complete, the $P_c^+$ states would be easily observable in the $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ decay in the LHCb Run2 data sample. On the other hand, with the chiral unitary approach in [102], the predicted partial widths for the $\Lambda_c^+ \overline{D}^{*0}$ system for all three studied $P_c^+$ states are very small. According to this model, the $P_c^+$ signal in the m($\Lambda_c^+ \overline{D}^{*0}$) invariant mass spectrum would be very minimal, likely below the combinatorial background. Very sizable signal is predicted in $\eta_c N$ system. In general, most model seem to favor the hadronic molecule picture, where a meson and a baryon maintain some residual interaction after the decay (likely, through vector meson exchange). This should then be visible in the invariant mass spectrum of $\Lambda_c^+ \overline{D}^{*0}$.

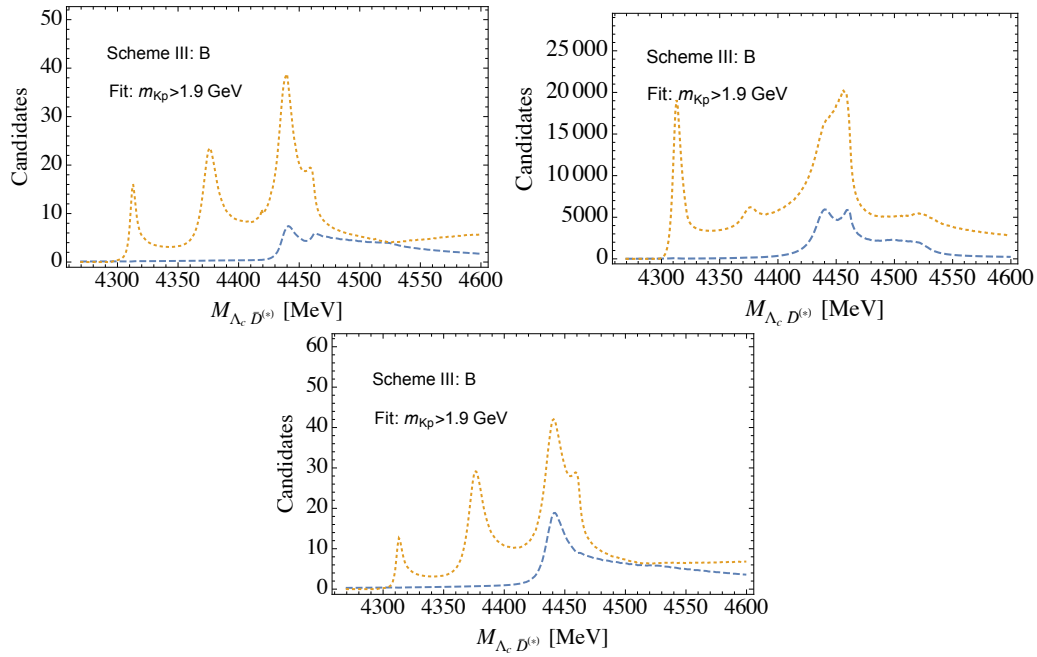



**Figure 2.11** *Predictions for the line shapes in the $\Lambda_c^+ \overline{D}^{*0}$ channel based on the fit results of scheme III. The yellow line indicates the $\Lambda_c^+ \overline{D}^{*0}$ channel, while the blue line indicates the $\Lambda_c^+ \overline{D}^0$ channel. Reproduced from [17].*

Making sound predictions on the branching ratio or other parameters related to the $P_c^+$ states is complicated by the fact that no solid experimental measurements of the quantum numbers of these states have been done so far. The shape of these signals is also difficult to predict. In this analysis, the $P_c^+$ states are assumed to be narrow Breit-Wigner like resonances without coupled-channel or interference effects between them. In reality, their proximity to meson-baryon thresholds may result in a broadened, asymetric shape for each of the studied states. A very nice work is laid out in [17]. Three different schemes for obtaining the shapes are constructed. This is because the inelastic parameters are very badly constrained by the current data. Scheme I - with pure contact interactions between the elastic, i.e., $\Sigma_c^* \overline{D}^*$, and inelastic channels and without the $\Lambda_c^+ \overline{D}^{*0}$ interactions, scheme II - where the one-pion exchange (OPE) is added to the scheme I, and scheme III - where the $\Lambda_c^+ \overline{D}^{*0}$ interactions are included in addition. The line shapes obtained in scheme III are shown in Fig. 2.11.

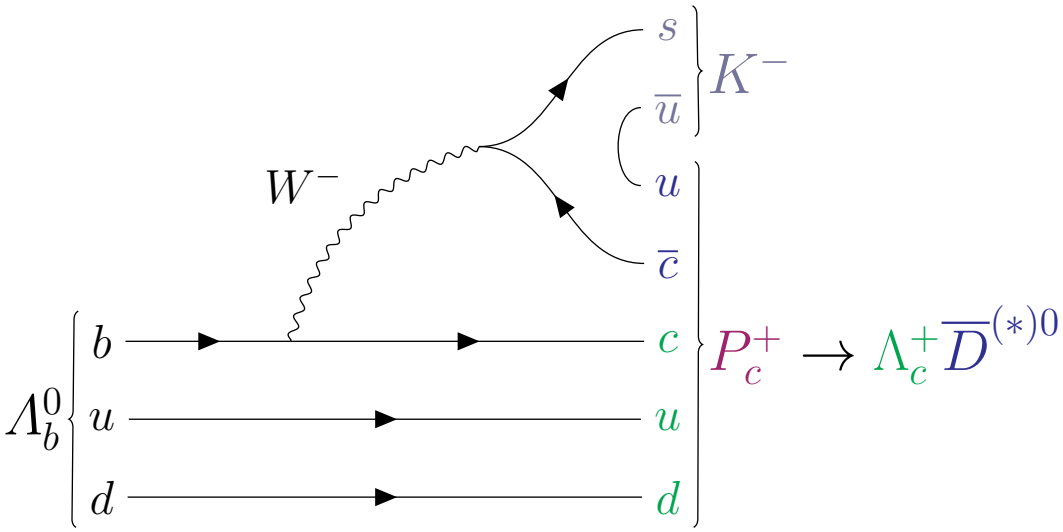



**Figure 2.12** *Quark diagram of decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$, where $\Lambda_b^0$ weakly decays via a $b \to c$ transition. A $u\bar{u}$ pair is created in the hadronization process. The $(c\bar{c}uud)$ quarks could form a pentaquark as either a tightly bound state or a $\Lambda_c^+ \overline{D}^{*0}$ hadronic molecule. This type of state would be considered a open charm pentaquark.*

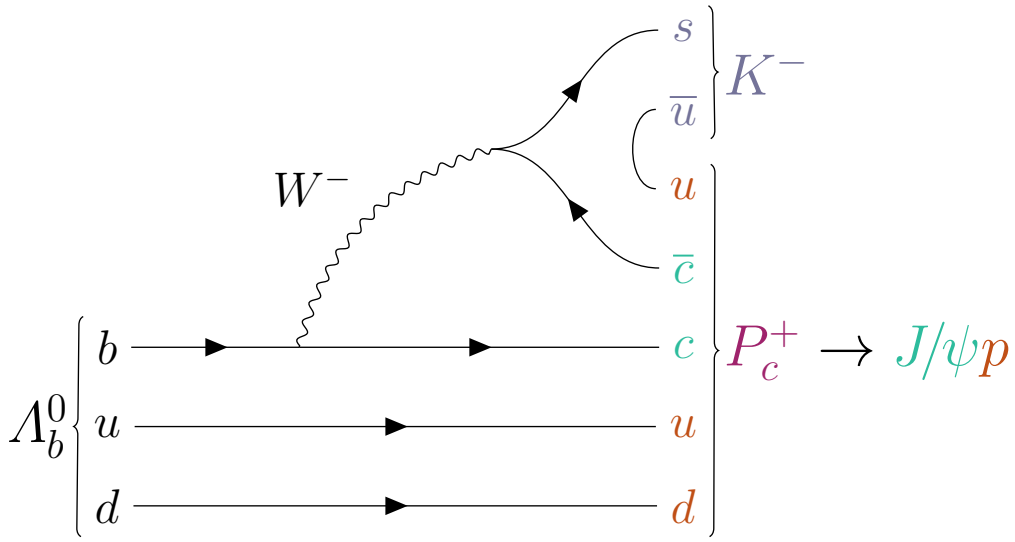



**Figure 2.13** *Quark diagram of decay $\Lambda_b^0 \to J/\psi p K^-$, where $\Lambda_b^0$ weakly decays via a $b \to c$ transition. A $u\bar{u}$ pair is created in the hadronization process. The $(c\bar{c}uud)$ quarks could form a pentaquark as either a tightly bound state or a $\Lambda_c^+ \overline{D}^{*0}$ hadronic molecule. This type of state would be considered a hidden charm pentaquark.*

**Table 2.2** *Predictions for $P_c^+$ decays to $\Lambda_c^+ \overline{D}^{*0}$. If the prediction depends on the model parameters, the lowest predicted ratio of partial widths, given a set of model parameters is cited.*

| Predictions for $\Lambda_c^+ \overline{D}^*$ | | |
|---|---|---|
| Model & reference | $\frac{\mathcal{B}(P_c \to \Lambda_c^+ \overline{D}^{(*)})}{\mathcal{B}(P_c \to J/\psi N)}$ | |
| **$P_c(4312)^+$ decays** | | |
| Effective-range-expansion and compositeness [93] | $\sim 1$ | Composition dependent, huge uncertainties |
| Coupled channel unitary [94] | 0.54 | $J^P = \frac{1}{2}^-$ $\Sigma_c^+ \overline{D}$ molecule; HQSS, dynamics from local hidden gauge symmetry; partial width from couplings |
| Extended chromomagnetic model [95] | 1 | Tightly bound $J^P = \frac{1}{2}^-$ pentaquark; Setting $\frac{\Gamma(\Lambda_c^+ \overline{D}^*)}{\Gamma(J/\psi p)} = 1$ since open/closed charm treated separately |
| Molecules in HQSS [96] | yes | |
| Effective Lagrangian (rescattering) [97] | 107 | $J^P = \frac{1}{2}^-$ $\Sigma_c^+ \overline{D}$ molecule; cut-off, form-factor dependent |
| Fierz rearrangement [99] | $\geq 0.69$ | $J^P = \frac{1}{2}^-$ $\Sigma_c^+ \overline{D}$ molecule |
| Chiral constituent quark model [100] | $8^{+6}_{-5}$ | Genuine pentaquark; Binding through quark re-arrangement |
| Chiral unitary [102] | 0.03 | $J^P = \frac{1}{2}^-$ $\Sigma_c^+ \overline{D}$ molecule; HQSS; Coupled channels; pion exchange |
| **$P_c(4440)^+$ decays** | | |
| Effective-range-expansion and compositeness [93] | $\sim 0.5$ | Composition dependent, huge uncertainties |
| Coupled channel unitary [94] | 0.23 | $J^P = \frac{3}{2}^-$ $\Sigma_c^+ \overline{D}^*$ molecule; HQSS, dynamics from local hidden gauge symmetry; partial width from couplings |
| Extended chromomagnetic model [95] | 1 | Tightly bound $J^P = \frac{3}{2}^-$ pentaquark; Setting $\frac{\Gamma(\Lambda_c^+ \overline{D}^*)}{\Gamma(J/\psi p)} = 1$ since open/closed charm treated separately |
| Molecules in HQSS [96] | yes | If $P_c(4440)$ is $P_{c3}$ $J^P = \frac{3}{2}^-$ |
| Effective Lagrangian (rescattering) [97] | 21 | $J^P = \frac{1}{2}^-$ $\Sigma_c^+ \overline{D}^*$ molecule; cut-off, form-factor dependent |
| Fierz rearrangement [99] | $\geq 0.41$ | $J^P = \frac{1}{2}^-$ $\Sigma_c^+ \overline{D}^*$ molecule; $J^P$ of $P_c(4440)$ and $P_c(4457)$ could be swapped. |
| Chiral unitary [102] | 0.006 | $J^P = \frac{1}{2}^-$ $\Sigma_c^+ \overline{D}^*$ molecule; HQSS; Coupled channels; pion exchange |
| **$P_c(4457)^+$ decays** | | |
| Effective-range-expansion and compositeness [93] | $\sim 0.2$ | Composition dependent, huge uncertainties |
| Coupled channel unitary [94] | 0.16 | $J^P = \frac{1}{2}^-$ $\Sigma_c^+ \overline{D}^*$ molecule; HQSS, dynamics from local hidden gauge symmetry; partial width from couplings |
| Extended chromomagnetic model [95] | 1 | Tightly bound $J^P = \frac{3}{2}^-$ pentaquark; Setting $\frac{\Gamma(\Lambda_c^+ \overline{D}^*)}{\Gamma(J/\psi p)} = 1$ since open/closed charm treated separately |
| Molecules in HQSS [96] | yes | If $P_c(4457)$ is $P_{c1}$ $J^P = \frac{1}{2}^-$; $\frac{\Gamma(\Lambda_c^+ \overline{D}^*)}{\Gamma(\Lambda_c^+ \overline{D})} = \frac{4}{3}$ |
| Effective Lagrangian (rescattering) [97] | 11.5 | $J^P = \frac{3}{2}^-$ $\Sigma_c^+ \overline{D}^*$ molecule; cut-off, form-factor dependent |
| Fierz rearrangement [99] | $\geq 0.35$ | $J^P = \frac{1}{2}^-$ $\Sigma_c^+ \overline{D}^*$ molecule; $J^P$ of $P_c(4440)$ and $P_c(4457)$ could be swapped. |
| Chiral unitary [102] | 0.08 | $J^P = \frac{3}{2}^-$ $\Sigma_c^+ \overline{D}^*$ molecule; HQSS; Coupled channels; pion exchange |

## 2.8 Legendre Moments

A model independent approach has been taken to probe the m($\Lambda_c^+ \overline{D}^{*0}$) invariant mass spectrum for the possible existence exotic states. Known as the *Method of Moments*, it was first used by the BABAR collaboration in 2009 [103] and then further refined and improved by the LHCb collaboration in 2015 while looking for $Z(4430)^-$ tetraquark state [79]. This procedure was also used for a separate model independent study to verify the results from the amplitude analysis of $\Lambda_b^0 \to J/\psi p K^-$ decays in the search of pentaquarks in $J/\psi p$ system [15]. The notion of expanding the mass variable in Legendre Polynomials stems from their connection to the *Spherical Harmonics* [104]. These functions, called *Spherical Harmonics* appear in the Eigenvalue equations for Angular Momentum operators $\hat{\underline{L}}^2$ and $\hat{\underline{L}}_z$

$$\hat{\underline{L}}^2 Y_\ell^m(\theta,\phi) = \ell(\ell+1)\hbar^2 Y_\ell^m(\theta,\phi) \tag{2.1}$$

$$\hat{\underline{L}}_z Y_\ell^m(\theta,\phi) = m\hbar Y_\ell^m(\theta,\phi) \tag{2.2}$$

where, $\ell$ and $m$ are the eigenvalues associated with $\hat{\underline{L}}^2$ and $\hat{\underline{L}}_z$, respectively, and $Y_\ell^m(\theta,\phi)$ are the spherical harmonics. Spherical harmonics can be expressed as follows:

$$Y_\ell^m(\theta,\phi) = \frac{(-1)^\ell}{2^\ell \ell!}\sqrt{\frac{(2\ell-1)(\ell+m)!}{4\pi(l-m)!}} e^{im\phi}(sin\theta)^{-m}\frac{d^{\ell-m}}{d(cos\theta)^{\ell-m}}(sin\theta)^{2\ell} \tag{2.3}$$

It is important to note that the expression for *Spherical Harmonics* can be written down in many different ways and a convenient way, gathering all terms with sines and cosines is given in [105]:

$$Y_\ell^m(\theta,\phi) = (-1)^m\sqrt{\frac{(\ell-m)!(2\ell+1)}{4\pi(l+m)!}} e^{im\phi} P_\ell^m(cos\theta) \tag{2.4}$$

where the term before $e^{im\phi}$ is a pre-factor, $e^{im\phi}$ itself is a phase factor and the term $P_\ell^m(cos\theta)$ is a function of the cosine of the angle $\theta$ which is the helicity angle

defined in Fig. 2.14. The functions $P_\ell^m(cos\theta)$ are called the *Associated Legendre Polynomials*. The *Legendre Polynomials* which are used to obtain the *associated legendre polynomials* can be obtained using the following generating formula:

$$P_n(cos\theta) = \sqrt{\frac{2n+1}{2}} \times 2^n \sum_{r=0}^{n} (cos\theta)^r \binom{n}{r} \binom{\frac{n+r-1}{2}}{n}$$

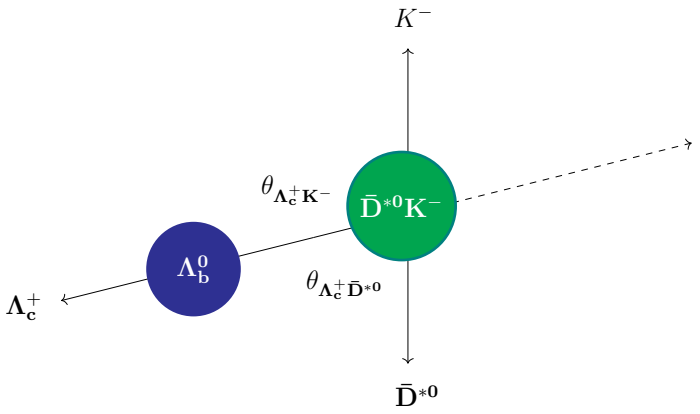
The pre-factor in this formula ensures the domain of the function is from -1 to 1. The first nine orders of *Legendre polynomials* with $cos\theta$ as an argument are given below:

$$\begin{aligned}
P_0(\cos(\theta)) &= 1 \\
P_1(\cos(\theta)) &= \cos(\theta) \\
P_2(\cos(\theta)) &= \frac{1}{2}(3\cos^2(\theta) - 1) \\
P_3(\cos(\theta)) &= \frac{1}{2}(5\cos^3(\theta) - 3\cos(\theta)) \\
P_4(\cos(\theta)) &= \frac{1}{8}(35\cos^4(\theta) - 30\cos^2(\theta) + 3) \\
P_5(\cos(\theta)) &= \frac{1}{8}(63\cos^5(\theta) - 70\cos^3(\theta) + 15\cos(\theta)) \\
P_6(\cos(\theta)) &= \frac{1}{16}(231\cos^6(\theta) - 315\cos^4(\theta) + 105\cos^2(\theta) - 5) \\
P_7(\cos(\theta)) &= \frac{1}{16}(429\cos^7(\theta) - 693\cos^5(\theta) + 315\cos^3(\theta) - 35\cos(\theta)) \\
P_8(\cos(\theta)) &= \frac{1}{128}(6435\cos^8(\theta) - 12012\cos^6(\theta) + 6930\cos^4(\theta) - 1260\cos^2(\theta) + 35).
\end{aligned}$$

The distribution of the helicity angle in a certain region of invariant mass or the combination of particles (in this case, $D^{*0}K^-$) can be expressed as a series of basis functions. If the chosen basis functions are the *Legendre Polynomials*, the coefficients of this expansion contain all of the information on the angular structure of the system and characterize the spin of the contributing resonances. These coefficients are called the *Legendre moments*. Where a k-th unnormalized moment is:

$$\langle Y_k^{U,j} \rangle = \sum_{l=1}^{N_j^{Data}} w_l P_k(\cos\theta)$$

**Figure 2.14** *Sketch of the helicity angles in $\overline{D}^{*0}K^-$ rest frame. First, the decay of $\Lambda_b^0$ is shown in $\Lambda_b^0$ rest frame, where $\Lambda_c^+$ and $\overline{D}^{*0}K^-$ decay back-to-back. The helicity angle is that between the direction vector of $\overline{D}^{*0}K^-$ in $\Lambda_b^0$ rest frame and $\overline{D}^{*0}$ or $K^-$ direction vector in $\overline{D}^{*0}K^-$ rest frame.*




Here, $w_l$ are efficiency and background subtraction weights (described in Section 4.12). $P_k$ is the k-th order *Legendre polynomial* and the sum is over all the candidates in a certain interval of $\overline{D}^{*0}K^-$ invariant mass spectrum. How this method is applied is better described in the model independent analysis Section 4.14 of the analysis Chapter 4.

## 2.9 Decay Kinematics

In particle physics, one of the pivotal concepts when it comes to decay kinematics and especially the limitations of what types of decays can happen is the Einsteins energy - momentum relation.

$$E^2 = (mc^2)^2 + (\mathbf{p}c)^2 \tag{2.5}$$

The expression above is actually an extension to the original mass - energy relation introduced by Einstein. In this case, motion of a particle is also taken into account. This was introduced in 1928 by Paul Dirac. In natural units, where $\hbar = c = 1$, the above equation becomes:

$$p^2 \equiv E^2 - |\mathbf{p}|^2 = m^2. \tag{2.6}$$

Whenever dealing with particle decays, it is important to keep in mind that the decay daughters each carry a part of the momentum of the decaying particle. The mass of decaying particle as well as its momentum determines the kinematically allowed values of the masses and momenta of the decay daughters. This, through the law of energy conservation already introduces a set of limits on what kinds of decays are kinematically allowed. There are of course other considerations, such as various other conservation laws of different quantum numbers - charge, angular momentum, etc. A diagram of a three-body decay is provided in Fig. 2.15.

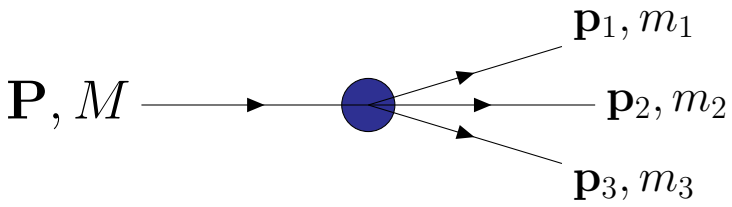



**Figure 2.15** *Diagram of a three-body decay.*

Examining such decays and the distribution of the momenta of the daughters further, mass and momentum of a combination of the decay daughters can be defined, namely, $\mathbf{p}_{ij}$ and $m_{ij}$. In the rest frame of the decaying particle:

$$\begin{aligned} \mathbf{p}_{ij} &= \mathbf{p}_i + \mathbf{p}_j \\ m_{ij}^2 &= \mathbf{p}_{ij}^2 \\ m_{12}^2 + m_{23}^2 + m_{13}^2 &= M^2 + m_1^2 + m_2^2 + m_3^2 \\ m_{12}^2 &= (\mathbf{P} - \mathbf{p}_3)^2. \end{aligned}$$

Here, $m_{ij}^2$ is a Lorentz invariant quantity. The phase space of a decay is also flat in squared-mass variables. This allows to construct and investigate a parameter space of $m_{ij}^2$ and $m_{jk}^2$, where $_{ijk}$ are the indices referring to a certain selection of the daughters of the three body decay. This parameter space is called a Dalitz space. It is a powerful and intuitive tool used to study the decay dynamics in a multi-body decay. As illustrated in Fig. 2.16, the kinematically allowed limits, which depend on the masses of the decaying particle and decay daughters restrict the allowed Dalitz space.

As well as displaying a region of kinematically allowed values for events coming from a three body decay, if the decay is happening through an intermediate

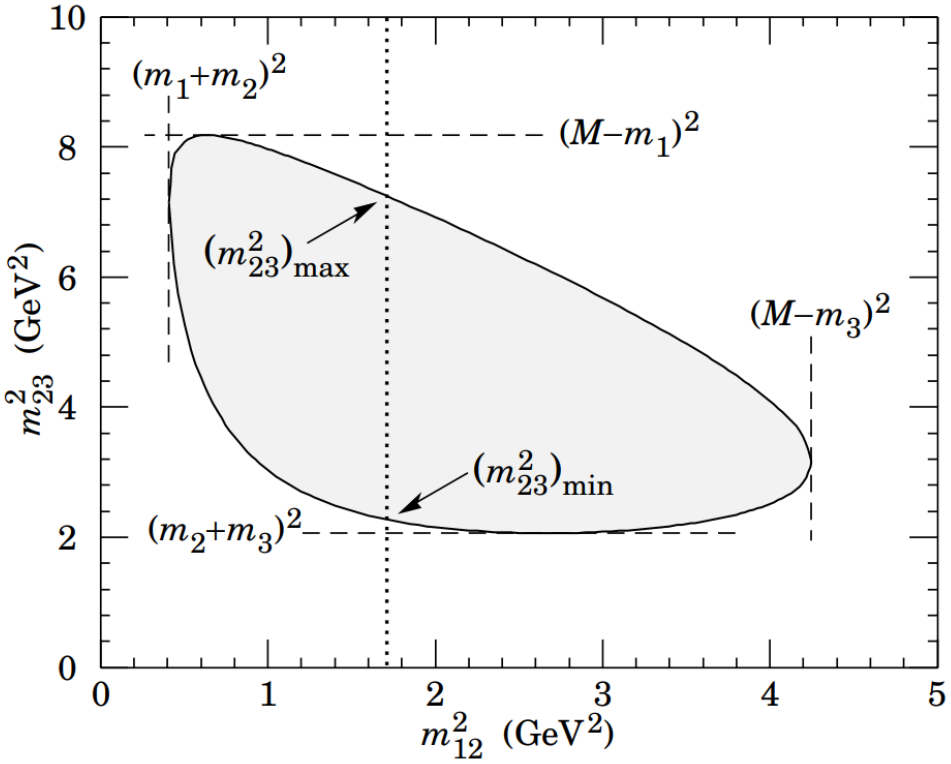



**Figure 2.16** *Dalitz plot for a three body final state. 4-momentum conservation restricts the events to the shaded region. Reproduced from [18].*

resonance - an enhancement in the cross section, this would be visible in the Dalitz plot. Resonances appear as regions (or bands) of increased density of events. The angular structure of the decay can also be inferred. If the bands are not uniform, but have gaps (nodes) in them, the intermediate resonance through which the particle is decaying carries a certain value of angular momentum.

# Chapter 3

# Statistics Background

From the statistical perspective, particle physics experiments at CERN can be treated as counting experiments. Over the entire period when the experiment is running, data is accumulated in discrete increments - candidates for specific particles are selected from collision events and recorded into digital storage media. During physics analyses, the distributions for various parameters related to these particle candidates can be obtained and some conclusions on the underlying physical phenomena can be drawn. The question then becomes: how likely is it that by running the same experiment again, the obtained distributions would be the same or similar. The outcome of a counting experiment in particle physics is described by the Poisson model:

$$P(N|\mu) = \frac{\mu^N e^{-\mu}}{N!}$$

Here, N observations of a random process are measured in a fixed time interval, $\mu$ events are expected on average.

It is important to note that unlike for some other experiments, *time interval* for a particle physics experiment usually refers to the entire lifespan of the experiment or at least the effective data taking period, which can be years or decades. For this , redoing a particle physics experiment is not feasible and statistical inference is used to draw observations from the data obtained. In statistics terms, most of the particle physics experiments follow a so-called *frequentist* paradigm, as opposed to *Bayesian*, which is more popular in other fields. The frequentist approach only

deals with probabilities assigned to experimental outcomes - the observed data, but not the theoretical models. It is only a convention that a discovery in particle physics is an observation, for which a probability to occur randomly is less than $2.87 \cdot 10^{-7}$, which is equivalent to $5\sigma$ fluctuation from the unit Gaussian. [106]

## 3.1 Foundations of Statistics

### 3.1.1 Likelihood

A concept of likelihood is instrumental in understanding statistical tests. As explained before, particle physics experiments are hardly repeatable, and in the vast majority of cases there is only one singular data sample. For this reason, in the statistical context of a particle physics experiment one often deals with likelihoods rather than probabilities. This is where confusion can occur. These two concepts should not be mistaken for one another. Probability measures a chance for future events occurring. In this case, a chance to obtain the same value of a certain observable (like the invariant mass of a particle). It is important to note that these obtained samples of the observables are thought of as distributions drawn from a continuous random vector. For this reason, the probability to observe any specific value of the observable is 0. This is where likelihood comes into play. Likelihoods only make sense in the context of a model which is used to parametrize the distributions of observables. Such a statistical model (or just - model) contains a number of parameters. In most cases, there is one or more parameters of interest (obtaining the values for which are the main focus of the physics analysis) and other parameters - nuisance parameters. In conclusion, likelihood measures how well a statistical model describes the observed data. Unlike probability, the values of a likelihood is not fixed between 0 and 1. A likelihood function can also be defined, it is a function of a parameter which gives the value of the likelihood for a fixed data sample. This is where the relation to probability lies - the value of the likelihood function is equal to the probability of the data sample to be observed at a fixed value of the parameter. Likelihoods and probabilities can be thought of as logical inverses on each other. In mathematical notation it is written as:

$$\mathcal{L}(\theta; \xi) = f_{\Xi}(\xi; \theta)$$

Here, $\theta$ is the parameter of a model, $\xi$ is a sample - realization of a random vector $\Xi$.

Some very important notions to remember are:

- The likelihood describes the extent to which the given sample provides support to a particular value of a parameter $\theta$.
- An arbitrary value of a likelihood is meaningless.
- The relative value, compared to a different value of a $\theta$ parameter or a parameter of a different model on the same data sample, is informative. [1]

The fact that only comparing likelihoods to one another makes sense, points to the idea of a likelihood ratio. It is one of the most important concepts in statistical tests as will be described later.

$$\frac{L(\theta_0;\xi)}{L(\theta_1;\xi)}$$

In a sample of size n the likelihood function takes the form of a product, because the joint density of a set of independent variables is equal to the product of their marginal densities.

$$L(\theta) = \prod_{i=1}^{n} f_i(\xi_i;\theta)$$

## 3.2 Hypothesis Testing

### 3.2.1 Test Statistic

A test statistic is a quantity calculated from the sample of data. Its value can be used to estimate how probable was the observed result with respect to some null hypothesis. In the absence of systematics, the most powerful test statistic that one

[1] This is also the logic of *minimization* where different values of parameters are used in a model and likelihoods estimated. The minimum value of a negative log likelihood function then points to the best fitting model for the data sample.

can construct is the likelihood ratio. This is a consequence of *Neyman-Pearson* lemma,

$$Q(\mu) = \frac{L(H_1)}{L(H_0)} = \frac{L(s(\mu) + b)}{L(b)}\,.$$

Here the null hypothesis is $H_0$ and an alternative hypothesis - $H_1$. In particle physics, as well as other fields, the role of the two hypotheses is usually given to a *background only* - $b$ hypothesis, where no signal is present, and a signal plus background hypothesis - $s + b$, where a relevant signal is present. A test statistic can be used to infer how compatible the alternative hypothesis is with the null (in physics usually *background only*) hypothesis. The likelihood value which goes into computing the test statistic is the maximum likelihood estimate for a specific model *pdf*.

Furthermore, Wilks Theorem states that as the sample size approaches $\infty$ the distribution of the test statistic $-2ln(\Lambda)$ asymptotically approaches the $\chi^2$ distribution under the null hypothesis $H_0$. So a good test statistic to choose is:

$$t = -2ln\frac{L(H_1)}{L(H_0)}$$

It is important to remember that the test statistic obtained from the data sample might be a result of years of data taking. To infer anything about the physics of the process one needs a *pdf* of the test statistic. This is where "Toy MC Samples" come in - there is a need to "replicate" the distribution obtained by collecting data many times over to see how likely it is to get the test statistic which was observed.

### 3.2.2 The p-value

Once enough toys are obtained, a distribution of the test statistic `t` can be drawn. A control region is then defined by choosing a confidence level. If the test statistic observed in data lies within this control region, one can say that the alternative hypothesis can't be ruled out up to this chosen confidence level. The p-value is a probability of obtaining as or more extreme result for the test statistic

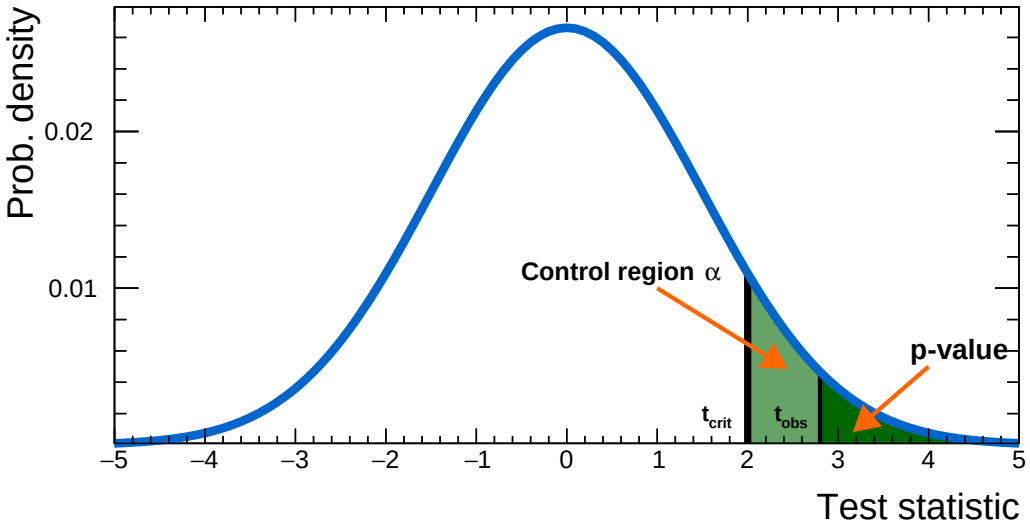



**Figure 3.1** *Control region and p-value shown on the distribution of test statistic t.*

by redoing the experiment (or by generating random toy samples from the original distribution) [107]. Mathematically, the p-value can be expressed as an integral

$$p = \int_{t_{obs}}^{\infty} g(t; H_0) dt \,,$$

here, $t_{obs}$ is the test statistic obtained from data, and $g(t)$ is the function of the test statistic obtained by generating multiple toy experiments. The illustration of how p-value is defined as an area of a distribution of test the test statistic is shown in Fig. 3.1

Computing the p-value at multiple values of the parameter of interest is called the *p-value scan*. This procedure is common and used in many analyses. Observing a large difference in the p-value at a certain value of the parameter can already indicate the presence of some physics phenomena which would make the alternative hypothesis difficult to rule out. Such techniques are used for example in deducing mass of a particle. The issue occurs if the statistics is limited and a background model (usually described by $H_0$ - null hypothesis) is complex. In this case, such an analysis becomes very sensitive to fluctuations of the background model. In the end it becomes troublesome to draw any conclusions via statistical inference because even though $H_1$ alternative hypothesis (usually signal + background) can

be established with some chosen precision, at the same time the background - $H_0$ hypothesis can not be ruled out. This is where *CLs* technique comes into play.

### 3.2.3 CLs Technique

The notion of using a modified p-value was first put forward by Birnbaum in [108]. This deals with the issue of `Type I` and `Type II` errors in statistics. `Type I` error is referred to as *False Positive* this is the probability to reject null hypothesis when the null hypothesis is true. `Type II` error, referred to as *False negative* is the probability to accept null hypothesis when the alternative hypothesis is true. Taking these errors into account, it becomes important not only what support can be put on the alternative hypothesis, but also with what certainty the null hypothesis can be rejected. If `Type II` error is denoted by $\beta$, the *POWER* of statistical test is defined as $1 - \beta$ [109]. The p-value is modified by taking into account the power of the statistical test. In other words, the sensitivity to the background model is accounted for in computing the p-value for signal. Schematically, on the distribution of the test statistic obtained by generating pseudo-experiments under null and alternative hypotheses, the p-values used in CLs computation are shown in figure Fig. 3.2. CLs is then defined as [110]:

$$\text{CLs} = \frac{p_{s+b}}{1 - p_b}$$

CLs is very powerful in cases where the experiment lacks sensitivity. It prevents exclusion or discovery of signal which the experiment is not sensitive to [111]. In this analysis, the signal region is sparsely populated with events, also, the local efficiency in the region which is probed for signal is low. That means the data set involved large statistical uncertainties after efficiency correction (effective density of events is increased at the expense of statistical power). For these reasons CLs method suits this analysis really well, and a reported result is a CLs Upper Limit on fit fractions of the $P_c^+$ signal in $\Lambda_c^+ \overline{D}^{*0}$.

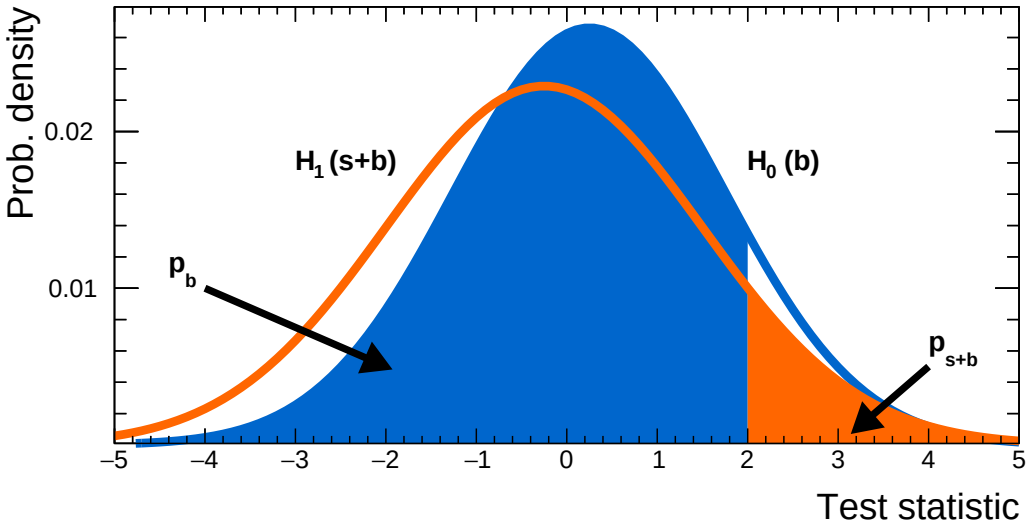
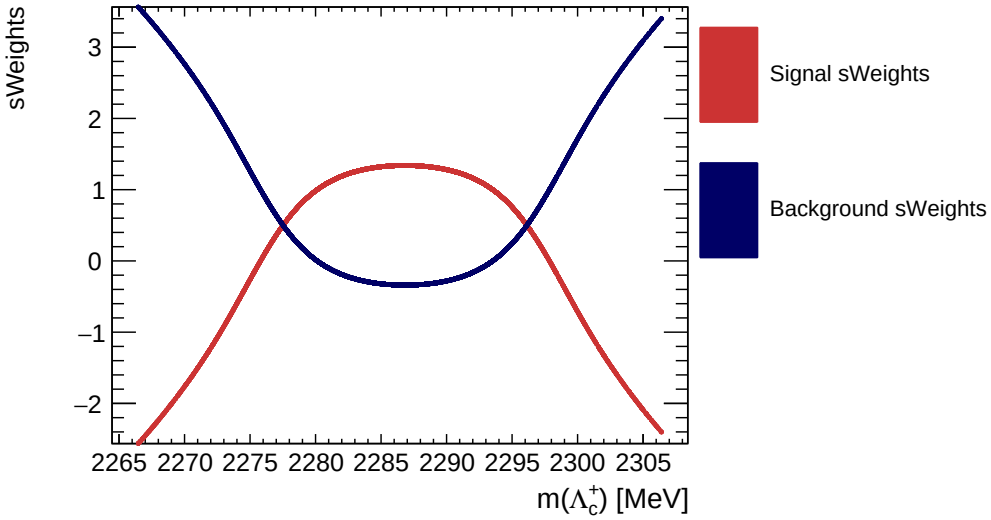



**Figure 3.2** *The p-values for background and signal + background model highlighted. They are inputs to calculating the value of CLs. The orange shape indicates the probability to obtain as or more extreme test statistic than a certain observed value. The blue shape shows the probability to reject the background hypothesis.*

## 3.3 *sPlot* Technique

In this analysis, combinatorial background in the data sample is subtracted by defining a set of weights called `sWeights`. These weights are constructed by applying a statistical method called *sPlot* [112]. This is a data driven statistical approach which uses different variables in the same data set. One such variable is established as a *discriminating variable*, another - a *control variable*. If the two variables are not correlated, different components can be unfolded in the *discriminating variable* and then identified in the control variable by applying `sWeights`. Different sources of events in the *discriminating variable* are unfolded by doing a maximum likelihood fit. In this analysis, the two sources of events (or the components which are unfolded) are the combinatorial background and $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ or $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ signal. The maximum likelihood fit is performed to the $\Lambda_c^+$ baryon invariant mass spectrum on a m($\Lambda_c^+$) variable without mass constraints by the `Decay Tree Fitter` (DTF) (as described in the Section 4.6). For each data point weights are computed, which characterize it as either background or signal. The *pdfs* for signal and background are used in computation of the `sWeights`. The `sWeights` are in fact the differences between these probabilities, ensuring that the sum of the weights is equal to the number

**Figure 3.3** *Distribution of signal and background* `sWeights` *plotted in* $m(\Lambda_c^+)$ *inv. mass. Signal* `sWeights` *peak around the known mass of a* $\Lambda_c^+$ *baryon while background* `sWeights` *are more pronounced in the side bands. This is an indication that* sPlot *technique worked correctly.*




of signal events in the sample. This set of weights can then be used on other variables of interest (`p, pt` and Dalitz variables).

The `sWeights` are mathematically defined as

$$ {}_s\mathcal{P}(y_e) = \frac{\sum_{j=1}^{N_s} \mathbf{V}_{nj} f_j(y_e)}{\sum_{k=1}^{N_s} N_k f_k(y_e)}, $$

here, $y_e$ is the discriminating variable used in the maximum likelihood fit, $N_s$ is the number of different components being unfolded (in this analysis $N_s = 2$ - signal and combinatorial background). $N_i$ are yields for different sources of events in the discriminating variable obtained from the fit to data. $f$ are the probability density functions for each source of events (signal and background in this case). Here $\mathbf{V}_{nj}$ is a covariance matrix and it acts as a correction term if the variable $y_e$ is not totally discriminating.

# Chapter 4

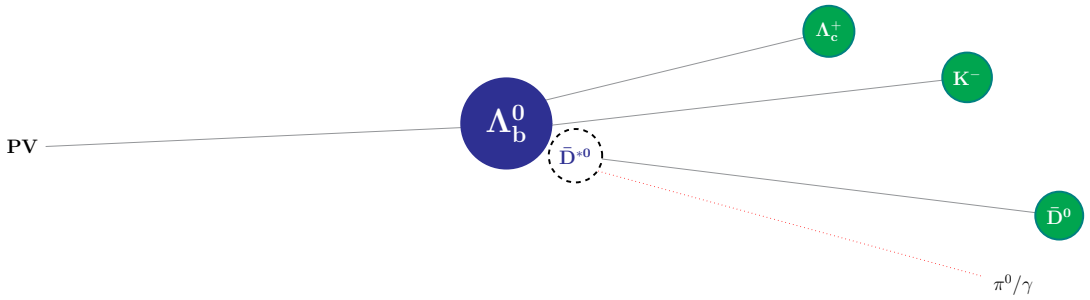
# Pentaquark Search in $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ Decays

## 4.1 $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\ \pi^0/\gamma) K^-$ Decay Chain

Since the decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ is partially reconstructed in the LHCb detector, candidates from this channel end up in the same data set as the ones for decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$. This is illustrated in Fig. 4.1. Two decay channels can be distinguished by how the candidates appear in the $\Lambda_b^0$ invariant mass spectrum. This is described in detail in Section 4.4. For this reason, in data acquisition routines, the decay chain is described in the same way for the decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ and $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$. In the latter, the $\overline{D}^{*0}$ candidate is not reconstructed. There is also no vertex or time of flight information because $\overline{D}^{*0}$ does not propagate in the detector and doesn't leave a `VeLo` track (described in section 1.3.1). No displaced vertex information is available for $\overline{D}^{*0}$ candidates. The lower limit for the mean lifetime of this resonance is set to $3.1 \times 10^{-22}$ seconds. Moreover $\overline{D}^{*0}$ decays predominantly into $\overline{D}^0 \pi^0$ (or $\overline{D}^0\ \gamma$) [113]. The branching fractions are provided in Table 4.15. The efficiency of $\pi^0$ and $\gamma$ reconstruction in this channel would be too low to investigate the Dalitz plot. Another approach on how to handle partially reconstructed decays was searched for during this work. The decay chain being studied in this analysis is:

$$\mathbf{\Lambda_b^0 \to \Lambda_c^+[\to pK^-\pi^+]\overline{D}^{*0}[\to \overline{D}^0[\to K^+\pi^-]\pi^0(or\ \gamma)]K^-}.$$

**Figure 4.1** *Sketch of a $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\ \pi^0/\gamma)K^-$ decay. Blue-dashed contour indicates that $\overline{D}^{*0}$ does not propagate in the detector and does not leave the track. Red dashed line indicates that $\pi^0$ or $\gamma$ from the decay is not reconstructed.*




## 4.2 Data Samples

### 4.2.1 $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ Decay Data Samples

This analysis is based on data collected by the LHCb experiment at CERN in the years of operation from 2015 to 2018. This amounts to about 6 $\text{fb}^{-1}$ of integrated luminosity from proton-proton collisions at 13 TeV center of mass energy. The breakdown of the data samples, categorized by year and LHCb magnet polarity is given in Table 4.1. These data samples are created using LHCb DaVinci software to pre-filter candidates for $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ decay. Already at this stage, optimized data handling pipeline reduces the size of data sample from petabytes to around 10 gigabytes without decreasing signal to noise ratio.

**Table 4.1** *$\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ Data Samples used in the analysis. Number of events and samples sizes are those after trigger selection, stripping selection and rough preselection to optimize file size.*

| **Data** $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ | | Event Type **90000000** | |
|---|---|---|---|
| **Year** | **Polarity** | **Number of Candidates** | **File Size** |
| 2015 | MagDown | 74508 | 290 MB |
| | MagUp | 48686 | 189 MB |
| 2016 | MagDown | 435423 | 1647 MB |
| | MagUp | 394517 | 1493 MB |
| 2017 | MagDown | 502374 | 1889 MB |
| | MagUp | 485929 | 1826 MB |
| 2018 | MagDown | 269020 | 1027 MB |
| | MagUp | 287703 | 1099 MB |
| **Total** | | **2498160** | **9460 MB** |

## 4.3 Monte Carlo Samples

Along with the data collected in Run2 of the LHCb experiment, this analysis uses a number of Monte Carlo Samples. Detector simulation is an important part of all the physics analyses at LHCb. It is a part of LHCb data flow which is described in Section 1.4. The need for these samples arises from having to describe the complicated shapes, governed by kinematics of the $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ decay used in the fit to the part of the $\Lambda_b^0$ invariant mass spectrum where the decays are partially reconstructed. These shapes are described in Section 4.7. Also, the Monte Carlo (MC) samples are used for efficiency correction and investigation of the resolution effects of both the detector and the Extended Cone Closure method, which is used to reconstruct the missing momentum of the $\overline{D}^{*0}$ candidate in the decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\ \pi^0/\gamma) K^-$. $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ MC samples were used to investigate the effects on the Dalitz plot by applying cuts based on the *Boosted Decision Tree* (BDT) response. `BDTs` were used for $\Lambda_c^+$ and $\overline{D}^0$ identification. They are desribed in the Section 4.5.2. The breakdowns of the MC samples, categorized by year and LHCb magnet polarity are given in Tables 4.2, 4.3, 4.4 and 4.5 for the decays $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$, $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$, $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0 \gamma) K^-$ and $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0 \pi^0) K^-$, respectively. `Events Produced` here are all of the events obtained in the MC generation step, `Events Selected` are the events passing a pre-selection as described in Section 4.4. There is also a `RapidSim` [114] particle-gun MC sample used for the decay $\Lambda_b^0 \to \Sigma_c^+(\to \Lambda_c^+ \pi^0) \overline{D}^0 K^-$. This sample is signal only, since the shape is difficult to estimate. $\Sigma_c^+$ component is strongly anti-correlated with $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0 \pi^0) K^-$ component, but it is color suppressed. This sample is only used to aid the fitting procedure described in Section 4.7. Each Monte Carlo sample was put through the same selection procedure as data samples. In the MC samples, some extra information about how the candidates were produced and what true decay was constructed and propagated through a simulation of the LHCb detector to obtain the correct response, is available. This information can be used to only select signal in the MC samples. To obtain clean signal shapes from the Monte Carlo samples, the candidates within these samples are `Truth-matched`. This involves only leaving the candidates with `Background Categories` which contain signal events. Also the `TRUE ID` assigned to each candidate is checked to match that of the particle it is identified as. `Background Categories` and `TRUE ID` act as data labels which are only available for the simulated samples. `TRUE ID` labels are following the EvtGen definition of particle IDs. EvtGen is a very widely used Monte Carlo

**Table 4.2** *$\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ MC Samples used in the analysis. Number of events and samples sizes are those after trigger selection, stripping selection and rough preselection to optimize file size.*

| **MC** $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ | | | Event Type **15396000** | |
|---|---|---|---|---|
| **Year** | **Polarity** | **Events produced** | **Events selected** | **File Size** |
| 2015 | MagDown | 303'089 | 2935 | 96 MB |
| | MagUp | 300'382 | 3341 | 107 MB |
| 2016 | MagDown | 751'190 | 14972 | 436 MB |
| | MagUp | 752'785 | 14965 | 434 MB |
| 2017 | MagDown | 1'001'161 | 16258 | 437 MB |
| | MagUp | 1'018'793 | 16222 | 438 MB |
| 2018 | MagDown | 1'251'516 | 16562 | 539 MB |
| | MagUp | 1'260'517 | 26007 | 826 MB |
| **Total** | | | **85255** | **2487 MB** |

**Table 4.3** *$\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} (\to \overline{D}^0\ \pi^0/\gamma) K^-$ MC Samples used in the analysis.*

| **MC** $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ | | | Event Type **15496220** | |
|---|---|---|---|---|
| **Year** | **Polarity** | **Events produced** | **Events selected** | **File Size** |
| 2015 | MagDown | 334'998 | 14550 | 195 MB |
| | MagUp | 318'950 | 13892 | 186 MB |
| 2016 | MagDown | 2'010'998 | 100077 | 1214 MB |
| | MagUp | 2'015'317 | 99677 | 1216 MB |
| 2017 | MagDown | 2'023'783 | 107675 | 1251 MB |
| | MagUp | 2'008'700 | 105961 | 1238 MB |
| 2018 | MagDown | 2'004'999 | 87735 | 1165 MB |
| | MagUp | 2'000'926 | 85205 | 1152 MB |
| **Total** | | | **614772** | **7617 MB** |

generator for B-Physics [115].

### 4.3.1 Monte Carlo Background Categories

In the simulated samples a number of `Background Categories` are defined using the logic highlighted in Fig. 4.2. The background in there categories is intended to mimic the sources of background which may be present in the real data collected by the LHCb detector. In the MC samples, these different background contributions can be distinguished from one another according to the `Background Category` (`BKGCAT`) identifier assigned to each candidate. A summary of conditions for assigning a certain background category to a candidate as well a list of categories is given in Tables 4.7 and 4.6.

**Table 4.4** *$\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$ MC Samples used in the analysis.*

| **MC** $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$ | | | Event Type **15396200** | |
|---|---|---|---|---|
| **Year** | **Polarity** | **Events produced** | **Events selected** | **File Size** |
| 2015 | MagDown | 252'957 | 3682 | 132 MB |
| | MagUp | 256'361 | 3689 | 132 MB |
| 2016 | MagDown | 1'001'636 | 16845 | 528 MB |
| | MagUp | 1'001'707 | 16721 | 527 MB |
| 2017 | MagDown | 1'001'695 | 22571 | 552 MB |
| | MagUp | 1'001'926 | 22315 | 551 MB |
| 2018 | MagDown | 2'005'805 | 38025 | 1074 MB |
| **Total** | | | **123848** | **3496 MB** |

**Table 4.5** *$\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ MC Samples used in the analysis.*

| **MC** $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ | | | Event Type **15396400** | |
|---|---|---|---|---|
| **Year** | **Polarity** | **Events produced** | **Events selected** | **File Size** |
| 2015 | MagDown | 255'227 | 4246 | 134 MB |
| | MagUp | 252'990 | 4188 | 132 MB |
| 2016 | MagDown | 1'001'531 | 18749 | 531 MB |
| | MagUp | 1'000'405 | 18446 | 529 MB |
| 2017 | MagDown | 1'005'973 | 25074 | 559 MB |
| | MagUp | 1'002'279 | 25136 | 557 MB |
| 2018 | MagDown | 2'023'569 | 42608 | 1093 MB |
| **Total** | | | **138447** | **3535 MB** |

| **BKGCAT Number** | **BKGCAT Name** | **Conditions** |
|---|---|---|
| 0 | Signal | !G !K !L A B C D E |
| 10 | Quasi-signal | !G !K !L A B C D !E |
| 20 | Phys. back. (full rec.) | !G !K !L A B C !D |
| 30 | Reflection (miss-ID) | !G !K !L A B !C |
| 40 | Phys. back. (part. rec.) | !G !K !L A !B !(C F) |
| 50 | Low-mass background | !G !K !L A !B C F |
| 60 | Ghost | G |
| 63 | Clone | !G K |
| 66 | Hierarchy | !G !K L |
| 70 | FromPV | !G !K !L M |
| 80 | AllFromSamePV | !G !K !L N |
| 100 | Pileup/FromDifferentPV | !G !K !L !A H |
| 110 | bb event | !G !K !L !A !H I |
| 120 | cc event | !G !K !L !A !H !I J |
| 130 | light-flavour event | !G !K !L !A !H !I !J |

**Table 4.6** *Definitions of Monte Carlo `Background Categories`*

| Condition | Explanation |
|---|---|
| A | All final-state particles used to form the candidate are matched to decay products of the same true MC particle (not necessarily the signal). |
| B | All final-state MC particles originating from the true MC particle defined in A are matched to particles used to form the candidate (except photons generated by PHOTOS); in case the decay descriptor of the true MC particle defined in A corresponds to an inclusive decay, "all final-state MC particles" only include the particles required in this (semi-)inclusive decay. |
| C | All final-state particles used to form the candidate are correctly identified, i.e. have been assigned their correct (true) mass. |
| D | The true MC particle defined in A is a signal decay according to the decay descriptor, or the head of a decay chain which differs from the signal decay chain only by the presence or absence of intermediate resonances but has other otherwise the same head, same final state particles and same topology. |
| E | The true MC particle defined in A is a signal decay according to the decay descriptor, and all intermediate states of this decay are correctly reconstructed (as listed in the decay descriptor). |
| F | The true MC particle defined in A has a mass which does not exceed the mass of the head of the decay descriptor by more than 100 MeV (tunable parameter). |
| G | At least one final-state particle used to form the candidate is a ghost. |
| H | Final-state particles used to form the candidate are matched to true particles from at least two different collisions (pileup). |
| I | At least one final-state particles used to form the candidate is matched to a true decay product of a b-hadron (following mother-daughter relationships all the way through). |
| J | At least one final-state particles used to form the candidate is matched to a true decay product of a c-hadron (following mother-daughter relationships all the way through). |
| K | At least two final state daughters are matched to the same MCParticle. |
| L | At least one final state daughter is matched to an MCParticle which is the MCMother of an MCParticle matched to another final state daughter. |
| M | At least one final state daughter is associated to an MCParticle from the primary vertex. |
| N | Every final state daughter is associated to an MCParticle from the same primary vertex. |

**Table 4.7** *Conditions used to deduce MC Background Categories.*

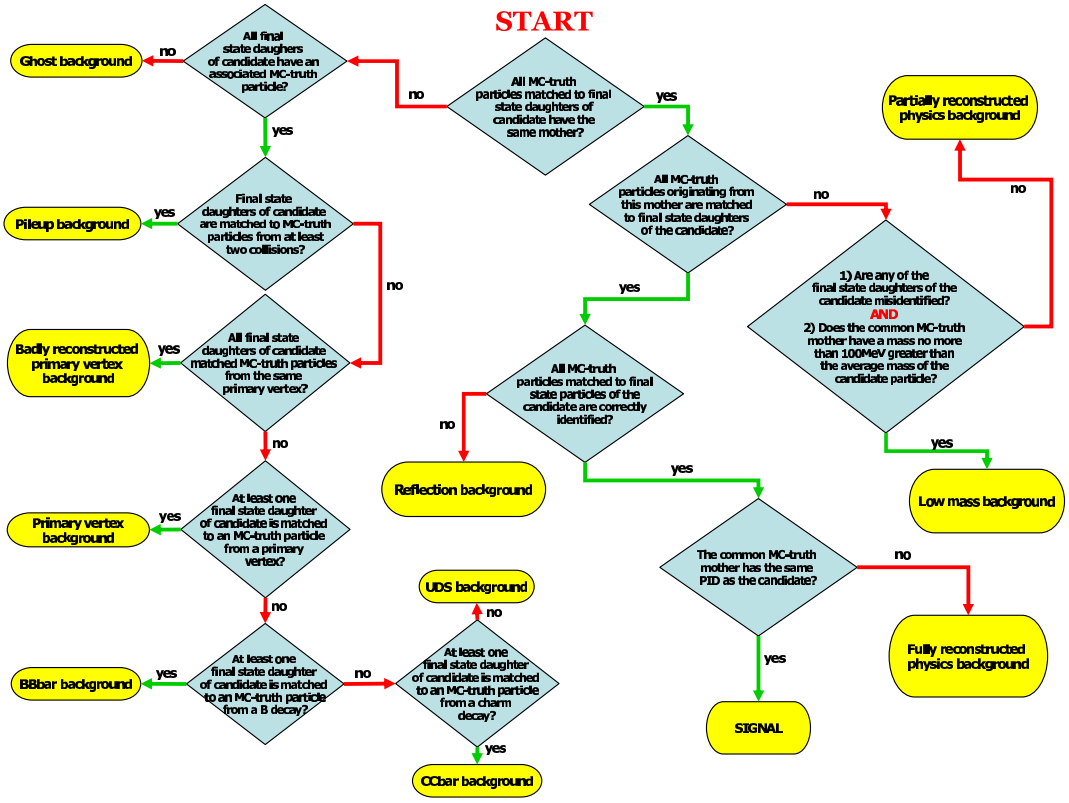



**Figure 4.2** *A flowchart highlighting the logic of assigning a specific* `Background Category` *for each candidate in the sample of simulated events.*

| Variable | Value |
|---|---|
| Lb_BKGCAT | 40 |
| Lb_BKGCAT | 20 |
| Lc_BKGCAT | 0 |
| Lc_BKGCAT | 10 |
| D0_BKGCAT | 0 |
| p_TRUEID | ± 2212 |
| Lc_K_TRUEID | ± 321 |
| Lc_pi_TRUEID | ± 211 |
| D0_K_TRUEID | ± 321 |
| D0_pi_TRUEID | ± 211 |
| K_TRUEID | ± 321 |
| Lc_TRUEID | ±4122 |
| D0_TRUEID | ± 421 |
| Lb_TRUEID | ±5122 |

**Table 4.8** *Background Categories and IDs used for truth-matching MC candidates.*

In this analysis, it is important to obtain the most accurate shapes in the invariant mass spectra of a particle or a combination of particles. This is why only the `Background Categories` which contain the signal for the decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ are selected. It is important to note that the usual `Background Category` indicating signal (`BKGCAT_0`) is not present in the MC samples, where the generated decay is $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} (\to \overline{D}^0\ \pi^0/\gamma) K^-$, because, just like in the real data, these samples are selected in such a way that $\overline{D}^{*0}$ candidate is not fully reconstructed but has a missing $\pi^0$ or $\gamma$. For this reason, `Background Categories` 40 and 20, which are normally *Physics Background*, constitute the signal. As an example, the distributions of different `Background Categories` in $\Lambda_b^0$, $\Lambda_c^+$ and $\overline{D}^0$ invariant mass spectra for $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ are provided in Fig. 4.3. In the $\Lambda_b^0$ mass spectrum (top of the aforementioned figures), there is no contribution from `BKGCAT 0` - signal. The number of candidates in other categories for this sample is minimal. The MC samples are in general generated as *signal MC* with a small background component. All the combinatorial background contributions are removed by selecting only `BKGCAT 40` and `BKGCAT 20`. Background categories for $\overline{D}^0$ and $\Lambda_c^+$ candidates include `BKGCAT 0` (and `BKGCAT 10` in the case of $\Lambda_c^+$) because here a fully reconstructed signal of $\Lambda_c^+ \to pK^-\pi^+$ and $\overline{D}^0 \to K^+\pi^-$ is produced. Small contributions from backgrounds in these two channels are again removed via selection. A list of `Background Categories` and `TRUEID` variables used to select signal in the MC samples is provided in Table 4.8.

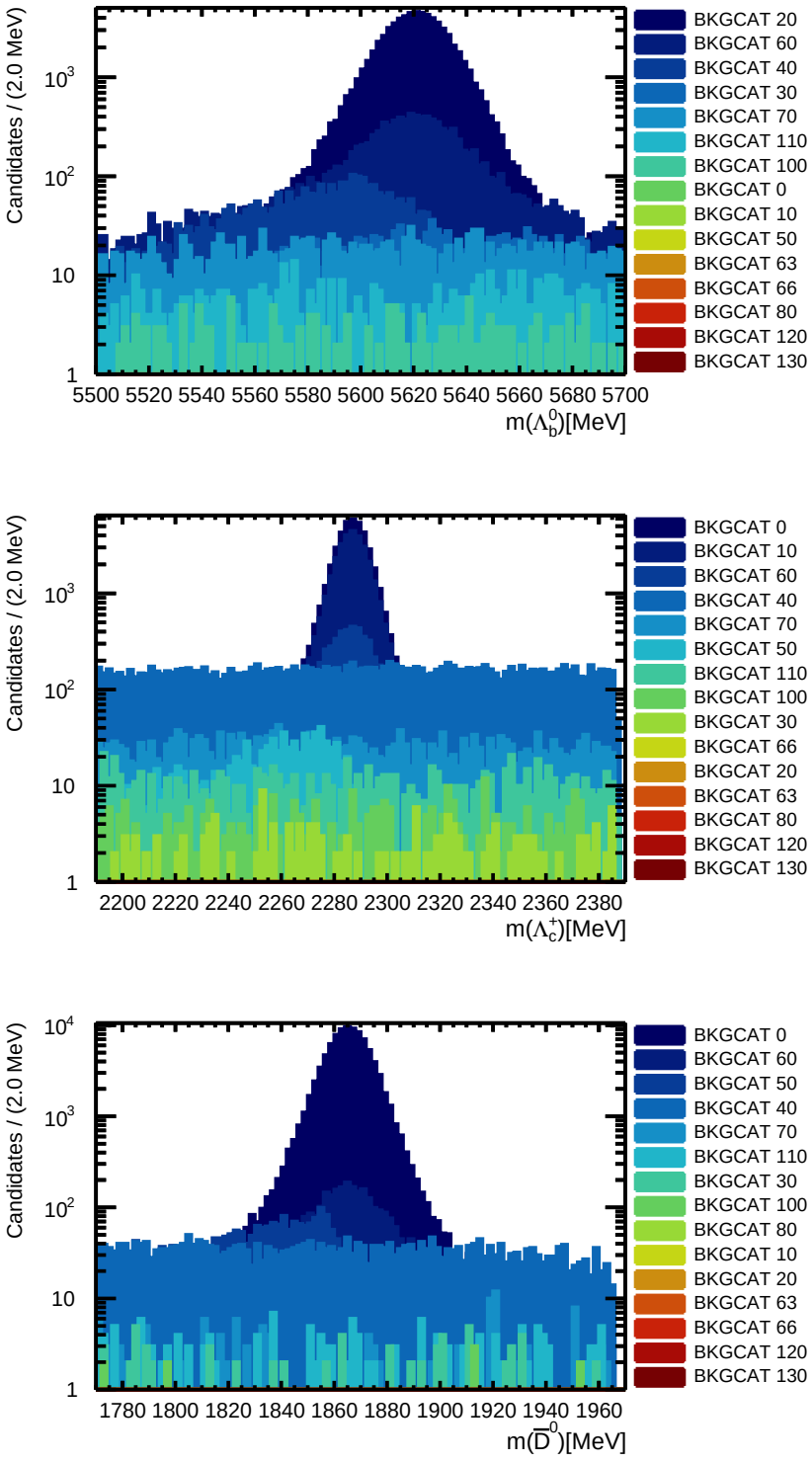



**Figure 4.3** *Invariant mass spectra of $\Lambda_b^0$, $\Lambda_c^+$ and $\overline{D}^0$ candidates in the $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ (Event type **15396000**) MC sample with breakdown to different `Background Categories` (`BKGCAT`).*

## 4.4 Candidate Selection

After multiple filtering stages, starting with real time filtering of detector output in `L0 Hardware Trigger`, building of physics candidates and tracks and getting rid of outliers in the offline stages of the `Trigger` as described in the Detector Chapter 1, extra stages of filtering are undertaken - the so-called offline selection. These steps already deal with a data sample where entries are various physical, fiducial and detector-related parameters of specific particles in the decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$. At this stage, the single entries are still treated as `Candidates` of being a certain particle. Offline selection deals with achieving best signal to noise ratio and obtaining the spectra for each of the particles in the decay with highest achievable resolution. The selection is performed by imposing a number of requirements - defining a number of `rectangular cuts` or conditions for each of the parameters in a high-dimensional parameter space. Some of the parameters are complicated derivatives of multiple more fundamental variables. Finding the optimal selection is a big challenge in a physics analysis like this. It involves multiple automatic or manual steps to probe the parameter space until a `working point` is determined. In any physics analysis, a working point can be chosen differently at any time. Choosing a working point in this analysis was a process spanning over 5 years of careful trial and error work and inspecting how this affects the spectra of observables. After the complete selection procedure $\sim 27000$ candidates for the decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ are maintained. A good understanding of the channel from previous analyses [20] allows for a tight selection. Some of the cuts can be already applied at the DaVinci level (described in LHCb Data Flow Section 1.4 of the Detector Chapter of this thesis) - while writing out the tuples (files containing arrays of required variables) in `.root` format. This pre-selection shrunk size of the data sample from `500GB` to `6.5GB`. Such optimization allows for a lot faster data handling and saves computation and data filtering time without loss of signal candidates, maintaining optimal signal to background ratio. Various selection stages are described in this section.

| **L0 trigger** | |
|---|---|
| L0Hadron \|\| L0Muon \|\| L0Dimuon \|\| L0Photon \|\| L0Electron \|\| L0DiElectron | |
| **HLT1** | |
| Good quality primary vertices | Charged tracks with $p_T > 500$ MeV |
| **HLT2** | |
| Hlt2Topo.*Decision \|\| Hlt2IncPhi.*Decision | |

**Table 4.9** *Summary of the* `Trigger` *selection requirements.*

### 4.4.1 Trigger Selection

The data in this analysis was collected during the Run2 of the LHCb. This means, that part of the `trigger` system was the, so called, `L0 Hardware Trigger`. Events first need to pass the `L0Physics` hardware trigger decision. Then, two software trigger stages further select the candidates depending on the quality of the primary vertices in the event, generic $p_T$ requirements and a set of topological conditions. This is also described in Section 1.4. This is a very rough pre-selection dealing more with the quality of candidates and the outliers in the data set. Its main purpose is to reduce the event rate from 40 MHz to the rate at which the events could be written to storage. The `L0 Trigger` selects the events with high transverse momentum ($p_T$) particles or with high energy deposits in the calorimeter. To optimize the real-time use of computing resources, the `HLT1` performs a partial event reconstruction, requiring good quality vertices and tracks with $p_T > 500$ MeV. This allows to filter on events with secondary vertices. It ensures that $K^-$, $\pi^+$ and $p$ candidates do not come directly from the primary vertex, but rather through a decay of a mother particle. In this case, the $\Lambda_b^0$ baryon. In the `HLT2` two `Trigger Lines` are used. The Topological Trigger Line (`HLT2Topo`) and the $\phi$ Inclusive Trigger Line (`IncPhi`). The first one is based on multivariate classification algorithms with discretized input variables [116]. This is used to select the n-body B hadron decays. The second `Trigger Line` filters the detached $\phi$ mesons decaying into a pair of kaons with large transverse momentum. The topology of such events is similar in detector signatures to the decays $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ studied in this analysis. Summary of trigger stages can be found in Table 4.9.

### 4.4.2 Stripping Selection

The events in this analysis come from the `BHADRON Stripping Stream`. During the filtering stage called `Stripping`, a number of requrements on the values of different variables are imposed. A `Stripping Line` is defined as a set of these requirements (also called *cuts*). The name of the stripping line used in this analysis is `X2LcD0KD02KPiBeauty2CharmLine`. This line was specifically written to select the events relevant to the analyses of $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ decay. Cuts applied in the stripping selection stage are provided in Table 4.10. The variables here are mainly used to make sure the event is reconstructed with good quality tracks, vertices and momentum requirements. Variables like `VCHI2` and `TRCHI2` provide information on the quality of a track or a vertex, separation of tracks can be inferred from distance-of-closest-approach (e.g. `ACUTDOCA`) variables. Momentum reqirements can be imposed on individual momenta of the candidates and tracks (e.g. `P`, `PT`) or combination of multiple particles (e.g. `ASUM(...PT)`). There are also some neural-network based variables used for individual particle identification (e.g. `PIDk`, `PIDp`). At this stage no mass cuts on the individual particles are applied, because the candidates are not well defined yet. Similarly to the `Trigger` selection, more basic properties like tracks, momenta and fiducial variables are employed in the stripping selection.

| Variable | Cut | Applies to | Description |
|---|---|---|---|
| `CloneDist` | $> 5000$ | all tracks | Clone track rejection using the Kullback-Liebler Distance [117]. |
| `TRCHI2DOF` | $< 3$ | all tracks | Good track quality using track $\chi^2$/nDoF from Kalman-Filter track fit [118], [119] (cut already applied in reconstruction). |
| `nLongTracks` | $< 500$ | global | Number of Long tracks in the event. |
| `TRGHP` | $< 0.4$ | all tracks | Neural-net based fake track rejection [120]. |
| `P` | $> 1000$ MeV | all tracks | Reject low momentum tracks. |
| `PT` | $> 100$ MeV | all tracks | Reject tracks with low transverse momentum. |
| `MIPCHI2DV(PV)` | $> 4.0$ | all tracks | Reject tracks from any primary vertex in the event using the impact parameter $\chi^2$ computed by a vertex fit [121]. |
| `ASUM(PT)` | $> 1800$ MeV | $\overline{D}^0/\Lambda_c^+$ comb. | Reject soft backgrounds *e.g.* from decays with missing neutrals. |

| | | | |
|---|---|---|---|
| PT | $>$ 500 MeV | at least one $\overline{D}^0/\Lambda_c^+$ daughter | Require leading track. |
| P | $>$ 5000 MeV | | |
| ACUTDOCA | $<$ 0.5 mm | two-particle combinations | Distance of closest approach for all two-particle combinations. Remove combinatorial background. |
| BPVVDCHI2 | $>$ 36 | $\overline{D}^0/\Lambda_c^+$ | Reject prompt charm decays by using the geometrical $\chi^2$ defined in [122]. |
| BPVDIRA | $>$ 0 | $\overline{D}^0/\Lambda_c^+$ | Charm decay vertex downstream of primary vertex. Uses cosine of angle between $\overline{D}^0/\Lambda_c^+$ momentum and vector from PV to decay vertex. |
| VCHI2/VDOF | $<$ 10 | $\overline{D}^0/\ \Lambda_c^+/\Lambda_b^0$ | Select good decay vertex quality using $\chi^2$/nDoF computed by a vertex fit [121]. |
| PIDp | $>$ -10 | $p_{\Lambda_c^+}$ | Loose particle ID using $\Delta\mathcal{L}(p-\pi)$ [123]. |
| PIDk | $>$ -10 | $K^-_{\Lambda_c^+}$ | Loose particle ID using $\Delta\mathcal{L}(K-\pi)$ [123]. |
| PIDk | $<$ 20 | $\pi^+_{\Lambda_c^+}$ | Loose particle ID using $\Delta\mathcal{L}(K-\pi)$ [123]. |
| ASUM(...PT) | $>$ 5000 MeV | $\Lambda_b^0$ comb. | Sum of transverse momenta of all stable $\Lambda_b^0$ daughters including soft photons. |
| AM | 5200 - 9000 MeV | $\Lambda_b^0$ comb. | Invariant mass of $\Lambda_b^0$ daughters. |
| PT | $>$ 1700 MeV | at least one $\Lambda_b^0$ daughter | Require leading track. |
| P | $>$ 10000 MeV | | |
| MIPCHI2DV(PV) | $>$ 16 | | |
| MIPDV(PV) | $>$ 0.1 mm | | |
| BPVLTIME | $>$ 0.2 ps | $\Lambda_b^0$ | Reject combinatorial background from PV by using life-time computed by a dedicated fit [124]. |
| BPVIPCHI2 | $<$ 25 | $\Lambda_b^0$ | Select candidate produced in PV using the impact parameter $\chi^2$ w.r.t the best PV [121]. |
| BPVDIRA | $>$ 0.999 | $\Lambda_b^0$ | Select candidate pointing back to its best PV. |

**Table 4.10** *Stripping selection variables, their cut values and explanation. Reproduced from [20].*

### 4.4.3 Pre-selection in Tuple Production

Data and Monte Carlo samples are available via the LHCbDIRAC [41] interware. These samples are provided in `.DST` or `.MDST` format. They are generally large in size because they are unfiltered. The data on DIRAC is intended to serve as an entry point for various LHCb physics analyses. For different analyses the same sample might be used in different ways. A piece of LHCb software called `DaVinci` is used to output files in `.root` format - called tuples. These files have columns (or branches) of various physics, statistics and detector related parameters. When writing out these files, it is a good practice to apply some pre-filters already in `DaVinci`. When pre-filters are applied, the resulting files tend to be smaller in size. This section provides a number of cuts applied in `DaVinci` pre-filter stage with their values and some information about the reasoning for a specific cut. In general, the parameters where these pre-selection cuts are applied are quite low-level and the imposed restrictions are loose. They are only used to filter out the outliers and reduce the sample size by reducing the number of events for which no good quality vertex or track was fitted. Pre-selection requirements are chosen such that applying them should not reduce the signal to background ratio. There are very gentle cuts on multivariate analysis response variables (described in Section 4.5.2). These cuts remove the events which are very strongly background-like. Variables and cut values as well as the explanation are provided in Table 4.11. Unlike in the previous stages of selection, the requirements here, even if loose, deal with reconstructed candidates in the decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$. The variables again include distance of closest approach (`DOCA`), momenta of the candidates, the impact parameter `IP` (closest distance from the candidate track to the primary vertex) and the quality of the `IP` fit.

| Variable | Cut Value | Applies to | Information |
|---|---|---|---|
| Lc_BDT | $> -0.5$ | $\Lambda_c^+$ | BDT output for $\Lambda_c^+$ candidate. |
| D0_BDT | $> -0.5$ | $\overline{D}^0$ | BDT output for $\overline{D}^0$ candidate. |
| DOCACHI2_12 | $< 16$ | $\Lambda_c^+ \overline{D}^0$ | Quality of Distance of Closest Approach between $\Lambda_c^+$ and $\overline{D}^0$ candidates. |
| DOCA_12 | $< 0.2$ mm | $\Lambda_c^+ \overline{D}^0$ | Distance of Closest Approach between $\Lambda_c^+$ and $\overline{D}^0$ candidates. |
| DOCACHI2_13 | $< 20$ | $\Lambda_c^+ K^-$ | Quality of Distance of Closest Approach between $\Lambda_c^+$ and $K^-$ candidates. |
| DOCA_13 | $< 0.4$ mm | $\Lambda_c^+ K^-$ | Distance of Closest Approach between $\Lambda_c^+$ and $K^-$ candidates. |
| DOCACHI2_23 | $< 25$ | $\overline{D}^0 K^-$ | Quality of Distance of Closest Approach between $\overline{D}^0$ and $K^-$ candidates. |
| DOCA_23 | $< 0.5$ mm | $\Lambda_c^+ K^-$ | Distance of Closest Approach between $\Lambda_c^+$ and $K^-$ candidates. |
| M | $> 5.1$ GeV | $\Lambda_b^0$ | Mass of the $\Lambda_b^0$ candidate. |
| M | $< 6.2$ GeV | $\Lambda_b^0$ | Mass of the $\Lambda_b^0$ candidate. |
| CHI2VXNDF | $< 32$ | $\Lambda_b^0$ | Quality of the vertex fit per degree of freedom. |
| BPVVDZ | $> 0.2$ mm | $\Lambda_b^0$ | Distance from the end vertex of the particle to the related primary vertex. |
| PT | $> 4$ GeV | $\Lambda_b^0$ | Transverse momentum of a particle. |
| P | $> 32$ GeV | $\Lambda_b^0$ | Magnitude of the momentum of a particle. |
| IP_1 | $< 0.2$ mm | $\Lambda_c^+$ | Impact parameter of the $\Lambda_c^+$ candidate. |
| IPCHI2_1 | $< 16$ | $\Lambda_c^+$ | Quality of the impact parameter of the $\Lambda_c^+$ candidate. |
| IP_2 | $< 0.3$ mm | $\overline{D}^0$ | Impact parameter of the $\overline{D}^0$ candidate. |
| IPCHI2_2 | $< 16$ | $\overline{D}^0$ | Quality of the impact parameter of the $\overline{D}^0$ candidate. |
| IP_3 | $< 0.5$ mm | $K^-$ | Impact parameter of the $K^-$ candidate. |
| IPCHI2_3 | $< 16$ | $K^-$ | Quality of the impact parameter of the $K^-$ candidate. |

**Table 4.11** *Pre-selection variables, their cut values and explanation.*

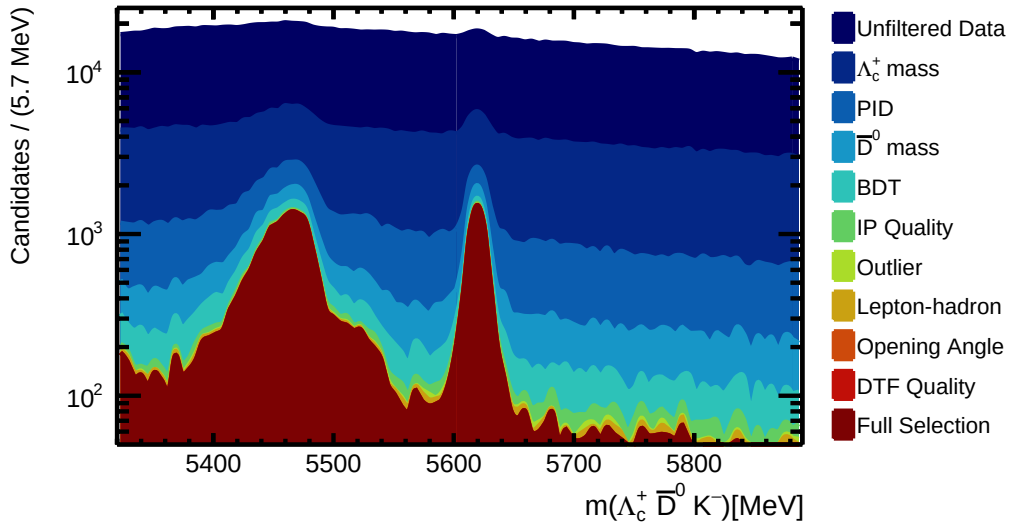
### 4.4.4 Further Filtering

Further offline selection on Data and MC samples has been performed. The variables to which the selection is applied, their values and the explanation is provided in Table 4.12. At this stage of the selection, a number of cuts is applied to select the $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ candidates used for physics analysis. The cuts here are fiducial, kinematic and PID related. Here the different entries to the selection procedure are treated as $\Lambda_b^0$, $\overline{D}^0$, $K^-$ and $\Lambda_c^+$ signal candidates. A correct mass window is selected by imposing the requirements on the reconstructed candidate mass. Some sidebands are retained in order to correctly parametrize the combinatorial background. This is explained in Section 4.10. The opening angles `oa` between the different combinations of these candidates are used to suppress the contributions from outliers. Comparative cuts to neural-network based `ProbNN` variables are applied to veto the misidentified particles. A kaon could be misidentified as a proton or a pion. In very rare cases an electron or a muon could be misidentified as a kaon, these cases are also filtered out. The final cuts are also applied to the so called `DTF` variables (described in the Section 4.6). Well discriminating variables are the response variables of a multivariate classifier, the, so called, `BDT` variables (described in the Section 4.5.2). A cumulative effect of adding different selection steps is illustrated in Fig. 4.4. The y-axis here is logarithmic. The background is greatly reduced by applying these selection steps. already selecting the correct mass window reduces the background by an order of magnitude. Further steps not only reduce the overall background but improve signal to background ratio. The `DTF Quality, Opening Angle, Lepton-hadron and Outlier Removal` cuts deal with very fine cleaning of the signal. The `DTF` cut is not visible in the figure because it only removes $\sim 900$ candidates if applied to the entire data set. It is still included for completeness. Also to show that the `DTF` does not converge in a very small fraction of cases. This is described in detail in Section 4.6. The imposed requirements in this final selection step often need to be changed to evaluate systematic effects or when selecting the working point for the analysis. Previous steps ensure that applying the final selection requirements can be done in a quick and simple manner, only dealing with relevant parameters and only employing the releavnt sub-sample of data.

| Variable | Cut Value | Information |
|---|---|---|
| Lb_1cdDTF_M | $> 5320$ | Inv. mass of a $\Lambda_b^0$ candidate with DTF |
| Lb_1cdDTF_M | $< 5900$ | Inv. mass of a $\Lambda_b^0$ candidate with DTF |
| D0_M | $> 1844.451$ | Inv. mass of a $\overline{D}^0$ candidate |
| D0_M | $< 1883.2981$ | Inv. mass of a $\overline{D}^0$ candidate |
| Lc_M | $> 2266.46$ | Inv. mass of a $\Lambda_c^+$ candidate |
| Lc_M | $< 2306.46$ | Inv. mass of a $\Lambda_c^+$ candidate |
| Lb_D0_K_DOCA | $< 0.25$ | Distance of closest approach |
| Lb_Lc_K_DOCA | $< 0.25$ | Distance of closest approach |
| Lb_K_IP | $< 0.25$ | Impact parameter |
| K_ProbNNe | $<$ K_ProbNNk | Neural network response K id'ed as e |
| K_ProbNNmu | $<$ K_ProbNNk | Neural network response K id'ed as $\mu$ |
| K_ProbNNk | $> 0.12$ | Neural network response K id'ed as K |
| Lb_BPVIPCHI2 | $< 8$ | $\chi^2$ - vertex quality for the $\Lambda_b^0$ |
| Lc_BDT | $> -0.15$ | BDT output for $\Lambda_c^+$ candidate |
| D0_BDT | $> -0.3$ | BDT output for $\overline{D}^0$ candidate |
| Lb_dDTF_DTF_CHI2 | $> 0$ | $\chi^2$ - goodness of fit for the DTF |
| Lb_1cdDTF_DTF_CHI2 | $> 0$ | $\chi^2$ - goodness of fit for the DTF |
| p_Lc_K_oa | $> 0.0002$ | Opening angle cut, for outlier removal |
| p_Lc_pi_oa | $> 0.0002$ | Opening angle cut, for outlier removal |
| p_D0_pi_oa | $> 0.0002$ | Opening angle cut, for outlier removal |
| p_D0_K_oa | $> 0.0002$ | Opening angle cut, for outlier removal |
| p_K_oa | $> 0.0002$ | Opening angle cut, for outlier removal |
| Lc_K_Lc_pi_oa | $> 0.0002$ | Opening angle cut, for outlier removal |
| Lc_K_D0_pi_oa | $> 0.0002$ | Opening angle cut, for outlier removal |
| Lc_K_D0_K_oa | $> 0.0002$ | Opening angle cut, for outlier removal |
| Lc_K_K_oa | $> 0.0002$ | Opening angle cut, for outlier removal |
| Lc_pi_D0_K_oa | $> 0.0002$ | Opening angle cut, for outlier removal |
| Lc_pi_D0_pi_oa | $> 0.0002$ | Opening angle cut, for outlier removal |
| Lc_pi_K_oa | $> 0.0002$ | Opening angle cut, for outlier removal |
| D0_K_D0_pi_oa | $> 0.0002$ | Opening angle cut, for outlier removal |
| D0_K_K_oa | $> 0.0002$ | Opening angle cut, for outlier removal |
| D0_pi_K_oa | $> 0.0002$ | Opening angle cut, for outlier removal |

**Table 4.12** *Variables used for further selection, their values and explanation. Here* `pi` *and* `K` *are pre-pended with the name of their mother particle (*`Lc` *or* `Lb`*). For example,* `Lc_K` *is a Kaon which originated in a $\Lambda_c^+$ decay. This is to avoid confusion between $\overline{D}^0$ decay products, $\Lambda_c^+$ decay products and the bachelor $K^-$. There is no prefix for the proton (p) because there is only the proton from $\Lambda_c^+$ decay in the final state.* `oa` *here is the opening angle between particles and is set to* 0.0002 *which is* 3.5 *mrad* $\sim 0.2$ *degrees (*0.45% *slope).*

**Figure 4.4** *Effect of different filtering conditions (cuts) applied during the offline selection procedure to the $\Lambda_c^+ \overline{D}^0 K^-$ inv. mass spectrum. Shown on a logarithmic scale.*



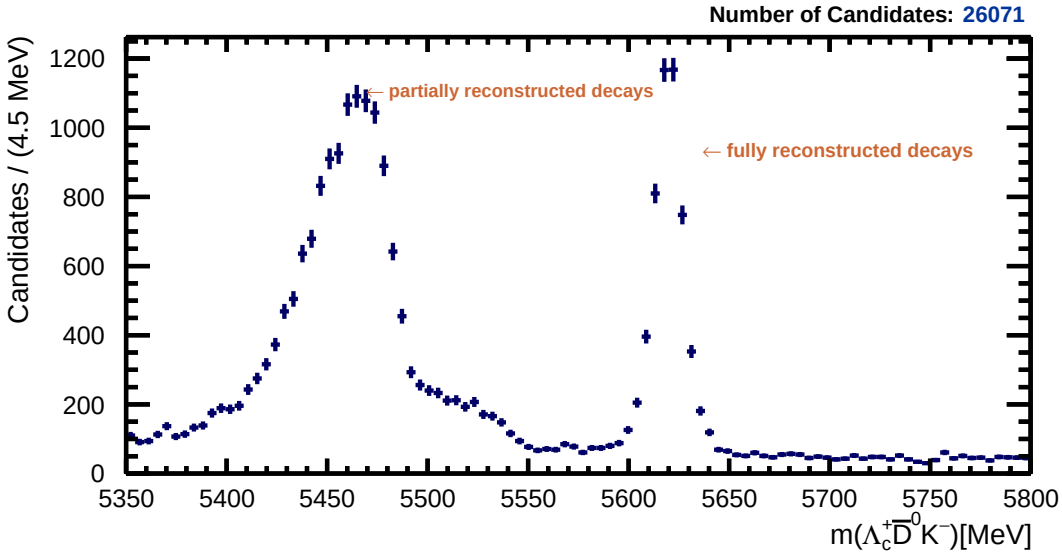




**Figure 4.5** *$\Lambda_c^+ \overline{D}^0 K^-$ inv. mass spectrum after complete candidate selection. It is visible that even before any background subtraction routine, combinatorial background is already strongly suppressed by the optimized selection procedure.*

## 4.5 Monte Carlo Correction

It is known that the distributions of some of the variables in the MC samples generated for LHCb do not exactly match those of the data. It is difficult to accurately simulate the variables related to the track-multiplicity. Also the kinematic variables like the magnitude of the momentum `Lb_P` and the transverse momentum `Lb_PT` are not matched exactly even with the best available tuning of PYTHIA [125]. The spectra of the b-hadron momenta produced by PYTHIA from proton-proton collisions are not completely accurate. To obtain a more accurate representation of the invariant mass spectra, the MC samples need to undergo a correction procedure. The shapes in the invariant mass spectra are determined by the decay kinematics. The variables related to candidate momenta are later used to construct the efficiency map and resolution graphs (described in Section 4.11 and Section 4.13). A data driven Machine Learning approach is chosen to correct the longitudinal and the transverse momentum of a $\Lambda_b^0$ candidate (`Lb_P`), `Lb_PT`, the number of tracks in an event (`nTracks`) and number of Scintillating Pad Detector hits (`hSPD`). The SPD detector is described in Section 1.3.5 of the Detector Chapter 1 of this thesis. A folding reweighting procedure is performed by using `hep_ml` [126] `Gradient-boosted decision tree` `GBDT` routine.

The MC samples are reweighted such that the variables mentioned above best match the corresponding distributions on the `sWeighted` data. The combinatorial background in the data is statistically subtracted as described in Section 4.10. For this, the complete data sample of Run2 is used. The MC samples are also combined into complete Run2 sample before the reweighting. A Gradient Boosted Decision Tree (GBDT) reweighting routine is used [127]. A data sample is split into training and testing sub-samples. The distribution in data, for each of the variables concerned, is called the target distribution and the MC distribution is the original distribution to be reweighted. To aid the process and apply the `sWeights` on the data as the weights of the target distribution, a *KFolding* procedure is performed. The data set is divided into $k$ samples and the alorithm is trained $k$ times. Each time, the number of samples used in the training is *k-1*. The training procedure itself involves imposing requirements on different variables in the data set. Some of these requirements are applied sequentially, in a tree-like pattern. Then the classifier result is compared. This is done many times until the required level of training is achieved. The, so called, hyper parameters used in the reweighting procedure are provided in Table 4.13.

| Hyper parameter | Value |
|---|---|
| Number of Estimators | 200 |
| Learning Rate | 0.05 |
| Max Dept | 20 |
| Min Samples per leaf | 1000 |
| Split Fraction | 0.6 |
| Number of Folds | 10 |

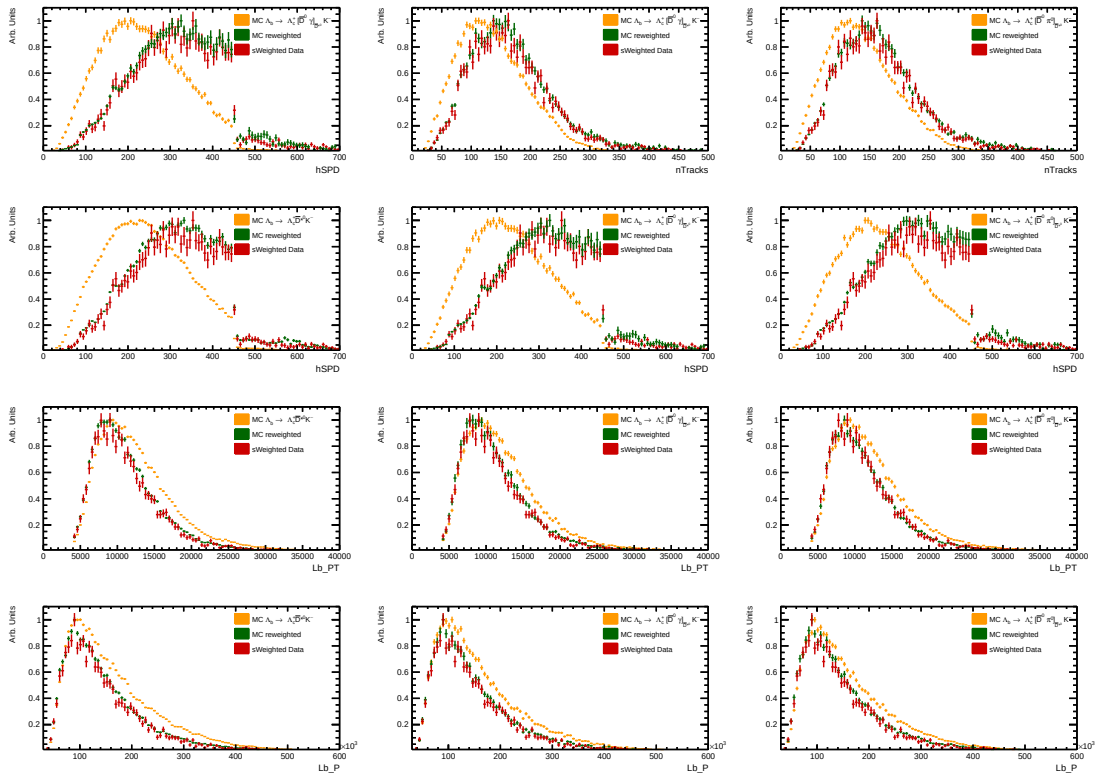
**Table 4.13** *Hyper parameters used for GBDT based kinematic reweighting.*

The comparison of the original MC distributions, the background subtracted (`sWeighted`) data distributions (see Section 4.10) and the reweighted MC distributions is shown in Fig. 4.6. The MC samples here are for the decays $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$, $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} (\to \overline{D}^0 \gamma) K^-$ and $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} (\to \overline{D}^0 \pi^0) K^-$, respectively. Before the reweighting the MC samples are put through the same selection procedure as the data, except the background is subtracted by choosing the appropriate `Background Categories` and `TRUE ID` values in a step called `Truthmatching` (see Section 4.4). Correction weights are computed for each of the three MC samples separately, the data sample they are compared against is always the same - combinatorial background subtracted complete Run2 data sample. It is visible that the momenta variables `Lb_P` and `Lb_PT` as well as the number of tracks variable `nTracks` do not differ much between the MC and the data. The `nTracks` and `hSPD` variables are related to the track multiplicity which is highly fluctuating during the operation of the LHC. This is why it is very difficult to obtain a good description for these variables in the simulation. The largest discrepancy between the MC and data distributions is in the `hSPD` variable. This is because on data, there is a selection requirement imposed on the number of long tracks in the event (see Section 4.4). If the number of long tracks is suppressed, number of SPD hits variable is truncated. Because the MC fails to correctly simulate this variable, the requirements are very loose in the MC and the variable is usually reweighted to match the data in later steps of the analysis, as it is done here.

### 4.5.1 PID Variables

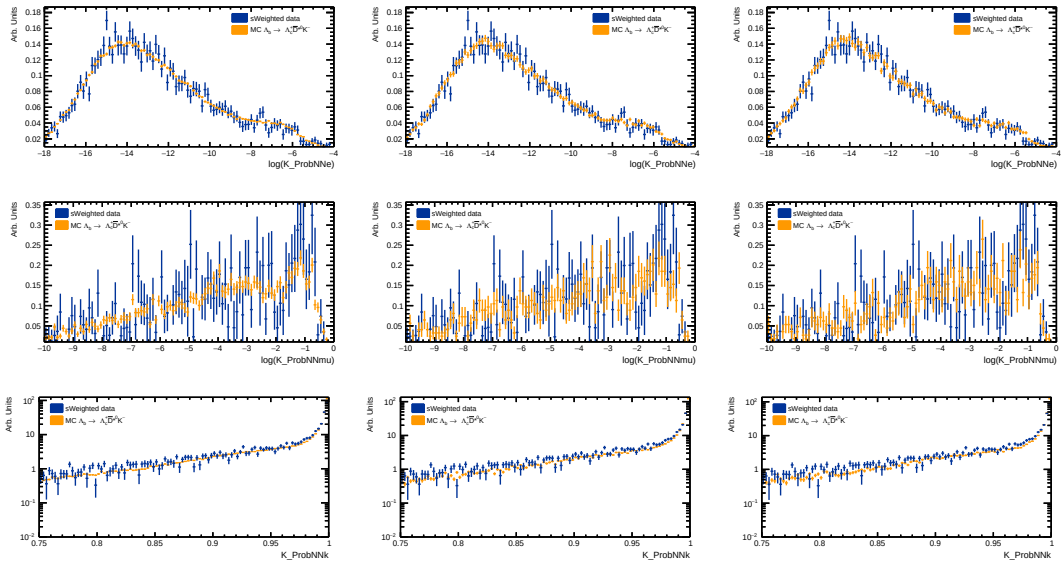
Particle Identification (PID) variables which using neural networks (`XProbNNY`) to asign a probability for a particle of being of certain type, are not always recreated well during the Monte Carlo generation stage. Because of this, there can be a discrepancy between the data and MC samples in these variables. It may be important to correct for this to obtain the closest agreement between MC and

**Figure 4.6** *Effects of Monte Carlo correction for 3 different MC samples on the variables used in the* `GBDT` *reweighting procedure.*




data. This way, the efficiency of the selection can be evaluated more accurately, since the applied cuts correspond to the same proportion of sample selected. In this analysis, only those PID variables which already have good data - MC agreement are used so such correction is not necessary. Since for the fine selection in this analysis only Kaon PID variables (`K_ProbNNe, K_ProbNNmu, K_ProbNNk`) are used, after an investigation, it was found that MC PID Correction step can be omitted as MC - Data agreement in `K_ProbNN_X` variables is the best of all PID variables and the differences are very minor. This can be seen in Fig. 4.7. The comparison to background subtracted data is shown for all three variables used in the selection and for each one of the three MC samples separately. The variables displayed here are used for comparative conditions in the selection procedure to suppress misidentification backgrounds where a kaon is identified as an electron or a muon (as described in Section 4.4). The `K_ProbNNk` cut is used for bachelor kaon identification. The agreement in this variable between data and MC is good. Using these variables without correction allows to use the same selection step for Data and MC which makes the analysis code and results clearer.

**Figure 4.7** *Comparison of PID Variables between data and MC.*



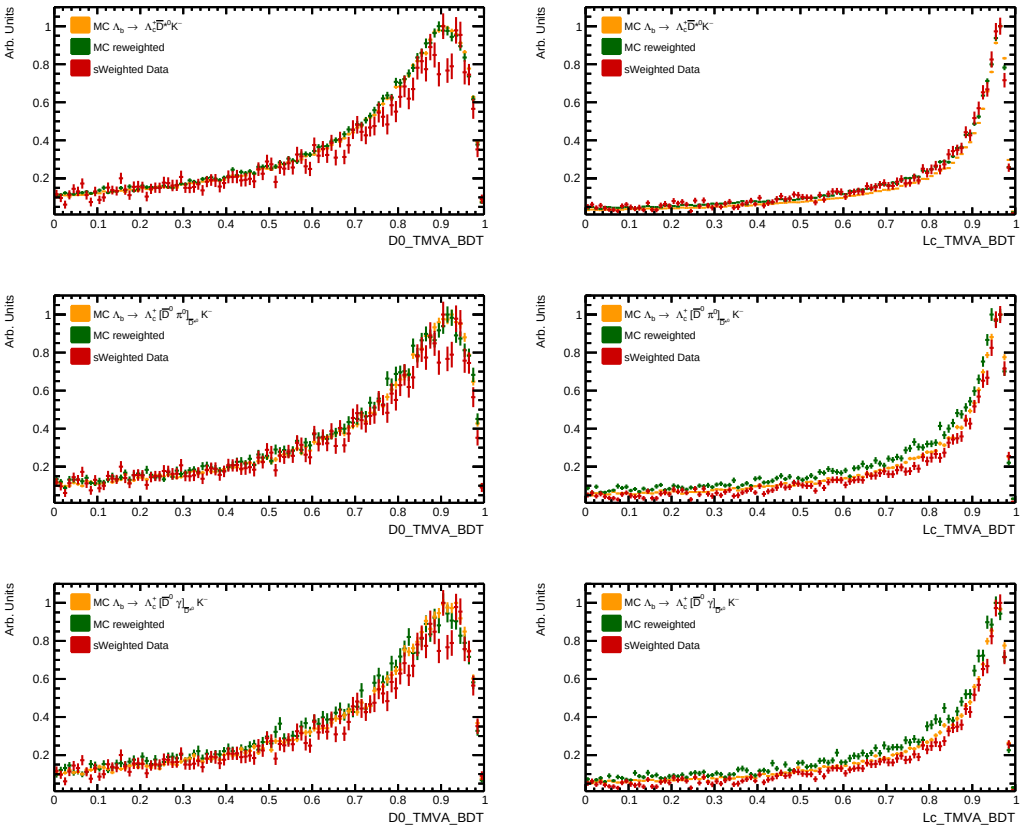

## 4.5.2 D-from-B BDT Variables

Two very powerful variables used in the selection and preselection in DaVinci are the PID-like variables for identifying $\Lambda_c^+$ and $\overline{D}^0$ candidates. B-hadron decays to the final states with charm quarks predominantly result from a $b \to \overline{c}s$ transition. During the hadronisation process the two charm quarks end up in different hadrons. Experimentaly, an important feature of such decays is the large number of charged particles in the final state that needs to be reconstructed. In the case of the decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$, there are 6 such particles. Since the decay products of $\Lambda_c^+$ and $\overline{D}^0$ are mainly $p$, $K^-$, $\pi^-$, a large combinatorial background is present in the data set. Using the fiducial and topological requirements alone is not enough to determine with high confidence if a final state particle came from a decay of a hadron or from the interaction point. For this purpose the particle identification information of the final state hadrons is used. Moreover, the correlations between the variables for different hadrons can be taken into account. For example, in a $\Lambda_c^+ \to pK^-\pi^+$ decays, if a proton is identified with high significance, the requirements on $K^-$ and $\pi^+$ could be relaxed. Conversely, if the kinematic region of a proton is such that the PID is not performing well, the requirements on the other particles can be tightened. Such correlations are taken into account when computing the BDT response variables.

These variables, called D-from-B BDTs [21] are constructed for the decay modes of the charmed hadrons $\Lambda_c^+ \to pK^-\pi^+$ and $D^0 \to K^-\pi^+$. In the (multivariate analysis) MVA training, the control channels are used, where the signal of a $\Lambda_c^+$ and a $D^0$ is very clean. Namely, the $\Lambda_b^0 \to \Lambda_c^+\pi^-$, for the $\Lambda_c^+$ mode and the $B^- \to D^0\pi^-$, for the $D^0$ mode. In a purely data driven procedure the BDTs are trained using the gradient boosting method withib the TMVA [128] framework. It is important to use a reference sample of real data and not the simulation here, because, as explained before in Section 4.5, the MC-data agreement is not absolute. Also, MC in the LHCb is mainly, so called, signal MC, where the combinatorial background is minimally reproduced. A very important task for the BDT variables is to corerctly identify the candidates which come from a true hadron decay and are not part of the combinatorial background. This can only be done by using a reference sample of the data also collected with the same detector conditions. Also, for this reason, the BDTs are trained for each year of the data taking separately, to make sure the slight changes in the detector conditions are taken into account.

These variables span the range from -1 to 1. Events with with values of BDT closer to -1 are more *background like* and events with the values of the BDT closer to 1 are more *signal like*. The value of $-0.5$ used as a cut in the pre-selection in DaVinci for both $\Lambda_c^+$ and $D^0$ BDT variables removes the majority of events which are characterized to a very high probability to be background. This helps to reduce the size of the data samples greatly without losing sensitivity and maintaining good signal to background ratio. In further selection, values Lc_BDT $> -0.15$ and D0_BDT $> -0.3$ are chosen after a selection optimization procedure implemented in $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ branching fraction analysis [20]. A grid scan through the parameter space of $\Lambda_b^0$ candidates is performed to select a desired working point for the BDTs. At each point in the scan, fits to $\Lambda_c^+$ invariant mass spectrum are performed to obtain best signal to background ratio. The distributions of these two variables in the range from 0 to 1 for the data and MC after selection is provided in Fig. 4.8. Further information can be found in the analysis note [129].

**Figure 4.8** *Comparison of BDT variables between background subtracted data and MC. The comparison is made for the three MC samples used in the analyses. First row - $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ MC (Event type **15496220**). Second row - $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} (\to \overline{D}^0 \pi^0) K^-$ MC (Event type **15396400**). Third row - $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} (\to \overline{D}^0 \gamma) K^-$ MC (Event type **15396200**).*



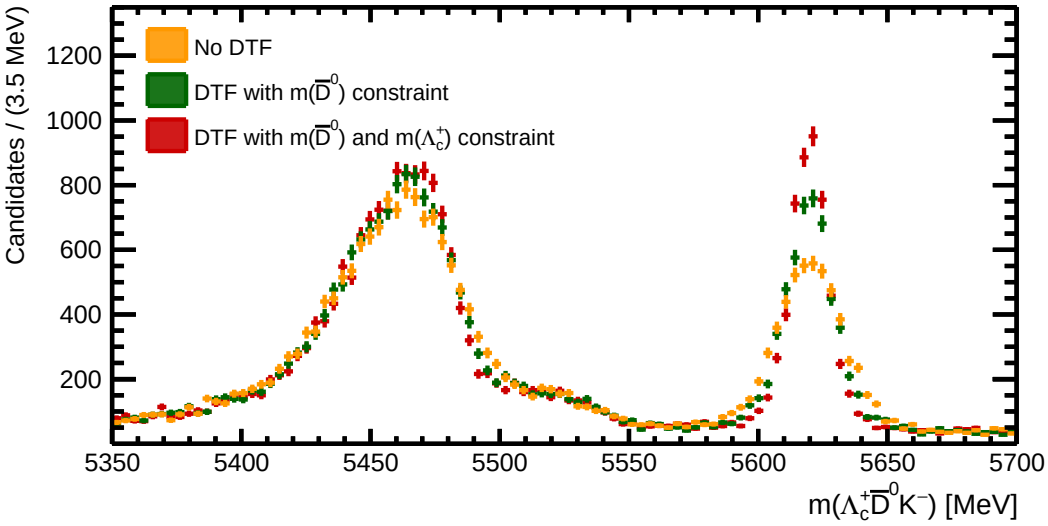
## 4.6 $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ Decay Tree Kinematic Refit

A software implementation of the `Kalman Filter`, called the `Decay Tree Fitter` or `DTF` [130] is used to perform a kinematic refit of the $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ decay tree. This is done withing the DaVinci framework. The main idea of such kinematic refit is to reoptimize the measured momenta by applying some kinematic constraints. These constraints are mainly the mass hypotheses for the candidates in the decay and constraints on the vertex location. An important thing to note is that performing this kinematic refit does not transform variables or distributions in any way, but rather introduces a number of additional recalculated variables. These new variables are the `X,Y,Z,E` components of the 4-momentum for each candidate in the decay and the mass.

Since a neutral object is missing in the reconstruction of the decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$, the constraints for the kinematic refit are different from the conventional case, where the whole decay chain is fully reconstructed. The masses of $\overline{D}^0$ and $\Lambda_c^+$ candidates are constrained to their world average values (published by the Particle Data Group [18]), but there are no vertex constraints as the combined direction vector, obtained from the final state particles reconstructed in the LHCb detector, does not point to the primary vertex. A kinematic refit in this analysis is very important because the increased mass resolution in the $\Lambda_c^+ \overline{D}^0 K^-$ invariant mass spectrum allows for a more accurate reconstruction of the missing $\overline{D}^{*0}$ momentum. The effect of using the `DTF` variables as inputs to the `ECC` method is described in Section 4.8. Especially, the `DTF` with a $\overline{D}^0$ mass constraint yields a larger number of $\overline{D}^{*0}$ candidates reconstructed with the `ECC` method as well as an improved mass resolution in the $\Lambda_c^+ \overline{D}^0 K^-$ invariant mass spectrum.

There are two different cases of kinematic refit implemented in this analysis. First, with only the $\overline{D}^0$ mass constrained and another, with both the $\overline{D}^0$ and the $\Lambda_c^+$ masses constrained. There are no vertex constraints applied in both cases. This can be used for comparison and was instrumental in studying the reconstruction of missing momentum of the $\overline{D}^{*0}$ candidate. From Fig. 4.9 one can see that applying the m($\overline{D}^0$) constraint in the `DTF` improves the mass resolution with respect to the case where no `DTF` is used. Applying the m($\Lambda_c^+$) constraint improves the resolution further. For this reason, the `DTF` with both the m($\Lambda_c^+$)

**Figure 4.9** *Effects on the $\Lambda_c^+\overline{D}^0K^-$ inv. mass spectrum of applying different constraints in the* `Decay Tree Fitter`*. The resolution is improved by imposing mass constraints.*

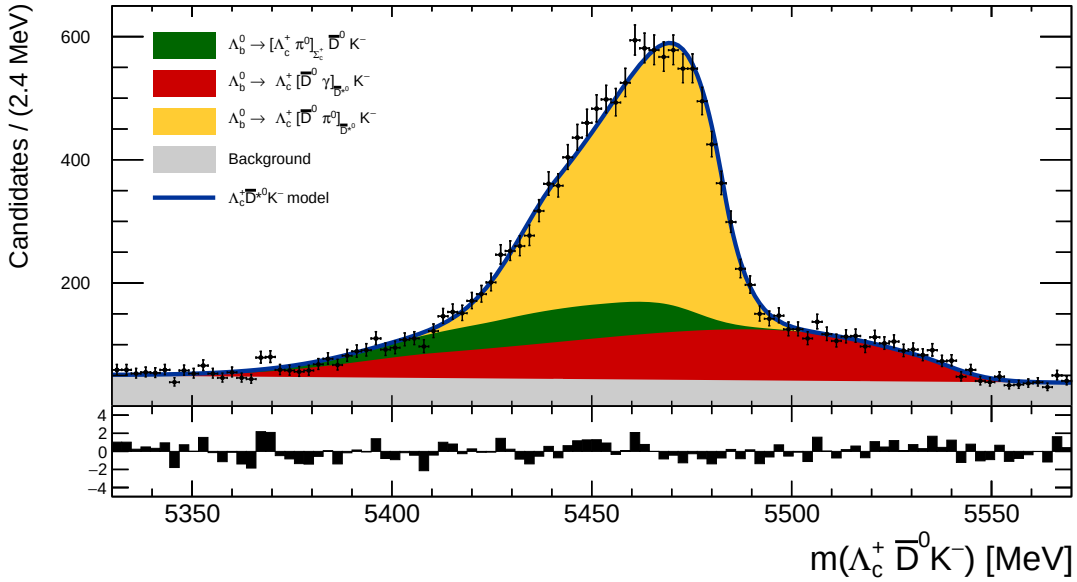



and the m($\overline{D}^0$) constraints was chosen to be used throughout the analysis.

## 4.7 Fit to $\Lambda_c^+\overline{D}^0K^-$ Invariant Mass Spectrum

A maximum likelihood fit is performed to the distribution of the candidates in the $\Lambda_c^+\overline{D}^0K^-$ invariant mass spectrum. The mass is that, recalculated by the `Decay Tree Fitter`, as described in Section 4.6. Where m($\Lambda_c^+$) and m($\overline{D}^0$) are both fixed to their world average values, but no vertex constraints are applied. Only the region where the $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\ \pi^0/\gamma)K^-$ decays are present is selected. Three different shapes, corresponding to three different contributions to the population of the observed candidates, are fitted with the help of MC samples. The MC samples here are those for the decays $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$, $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$ and $\Lambda_b^0 \to \Sigma_c^+(\to \Lambda_c^+\pi^0)\overline{D}^0K^-$. These samples mimic only the signal for a relevant component. It was thus important to parametrize the shapes from the MCs correctly. These shapes are in their own right obtained by a separate fit to $\Lambda_b^0$ invariant mass spectrum of each of the MC data sets separately. The results of these fits are provided in Fig. 4.12. The final fit to data is then performed in order to obtain a fraction of candidates from the decay $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$ under the $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ signal peak. This

**Figure 4.10** *Fit to data in the $\Lambda_c^+ \overline{D}^0 K^-$ inv. mass spectrum.*




is referred to as *gamma contamination* in the analysis. Understanding what fraction of $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$ component is present in the region where $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ decays are reconstructed is important in constructing the signal model for the limit setting procedure later (described in Section 4.15).

It is visible that the shapes of these partially reconstructed signal components are non-Gaussian. They are strongly asymmetric - *leaning* towards the right. Also the side-bands for each component are of different shapes. A `Bukin` probability density function [131] was employed as it was specifically constructed for fitting distributions with asymmetric peaks. The `Bukin` pdf is defined with 6 parameters, controlling the center of the peak, its width and shape of each of the side-bands separately. The fit to $\Lambda_b^0 \to \Sigma_c^+(\to \Lambda_c^+\pi^0)\overline{D}^0 K^-$ MC is done on the events generated within `RapidSim` [114] framework. `RapidSim` allows the generation of simluated samples of particle decays. By defining the decay tree and including possible intermediate resonances, a sample is generated containing such events. There are some limitations to `Rapid Sim`. Only one resonance can be included and the background contributions, as well as the detector effects, are not fully simulated. The actual shape for this component is very difficult to deduce from data or simulation, because the component is strongly anti correlated with $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ component. Kernel density estimation was used to parametrize the shape of $\Lambda_b^0 \to \Sigma_c^+(\to \Lambda_c^+\pi^0)\overline{D}^0 K^-$ component.

| | |
|---|---|
| $N_{\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0 \pi^0) K^-}$ | $7512 \pm 162$ |
| $N_{\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0 \gamma) K^-}$ | $4071 \pm 175$ |
| $N_{\Lambda_b^0 \to \Sigma_c^+(\to \Lambda_c^+ \pi^0) \overline{D}^0 K^-}$ | $1637 \pm 178$ |
| $N_{background}$ | $4353 \pm 131$ |

**Table 4.14** *Yields extracted from fit to $\Lambda_b^0$ inv. mass.*

**Table 4.15** *$\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ branching fractions [18].*

| | $D^{*0}$ Decay Mode | Fraction ($\Gamma_i/\Gamma$) |
|---|---|---|
| $\Gamma_1$ | $D^0\pi^0$ | $(64.7 \pm 0.9)\%$ |
| $\Gamma_2$ | $D^0\gamma$ | $(35.3 \pm 0.9)\%$ |

The decay $\Lambda_b^0 \to \Sigma_c^+(\to \Lambda_c^+\pi^0)\overline{D}^0 K^-$, through a $\Sigma_c^+$ resonance, is largely suppressed at the lowest order. The suppression comes from the difference in the color and the phase space factors as compared to decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$. The two quark diagrams are provided in Fig. 4.11 for comparison. The color suppression arises because of a possible decay configuration, where the $W^-$ boson is emitted internally. In the standard external tree level diagram (on the left), the $s\bar{c}$ pair is emmited externaly. In the hadronization process, the $u\overline{u}$ pair can be created from the vacuum in any colors. On the other hand, when there is an internal $W^-$ emission, the $\bar{c}$ and $s$ have to match the color of the $u$ coming from the $\Lambda_b^0$. The decay is suppressed by a color factor which is $1/N_c^2 = 1/9$. The contributions to the amplitude from the strong decays are not taken into account. This factor could be differnt because of these contributions. The yield for $\Lambda_b^0 \to \Sigma_c^+(\to \Lambda_c^+\pi^0)\overline{D}^0 K^-$ contribution obtained from the fit suggests that the factor is close to 1/9. Also, in the standard external tree level diagram, the $u$ and the $d$ are both spectators. The (ud) diquark here has isospin $I = 0$, because it is coming from a $\Lambda_b^0$ and not a $\Sigma_b$ decay. There is no change in the isospin when transitioning the the final state. In the case of the internal $W^-$ emission, the $u$ in the baryon in the final state can come in a different isospin alignment. For this reason, the emisison of a $\Sigma_c^+$ is now possibe. The $\Sigma_c^+$ carries the isospin $I = 1$.

For the final fit to data, the shapes obtained from the fits to MC are fixed and only their fit fraction is allowed to vary. The two dominating components are the decays involving $\gamma$ and $\pi^0$. A $\Lambda_b^0$ candidate in this region decays through $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ with the probability of $\sim 65\%$ and through $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$ with the probability of $\sim 35\%$. It is important to note that

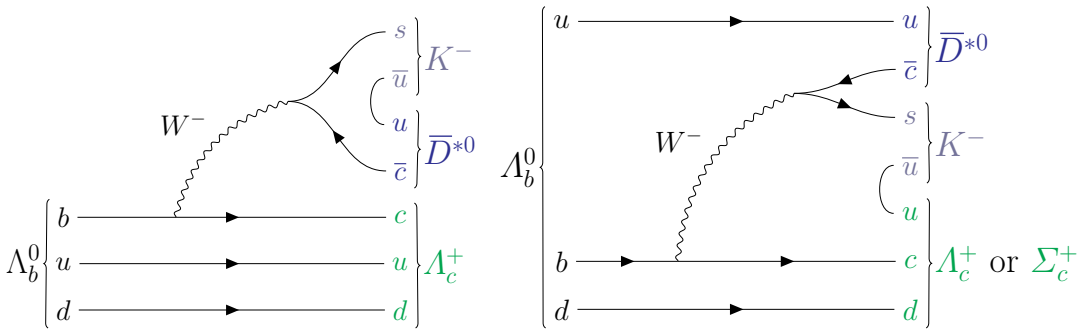
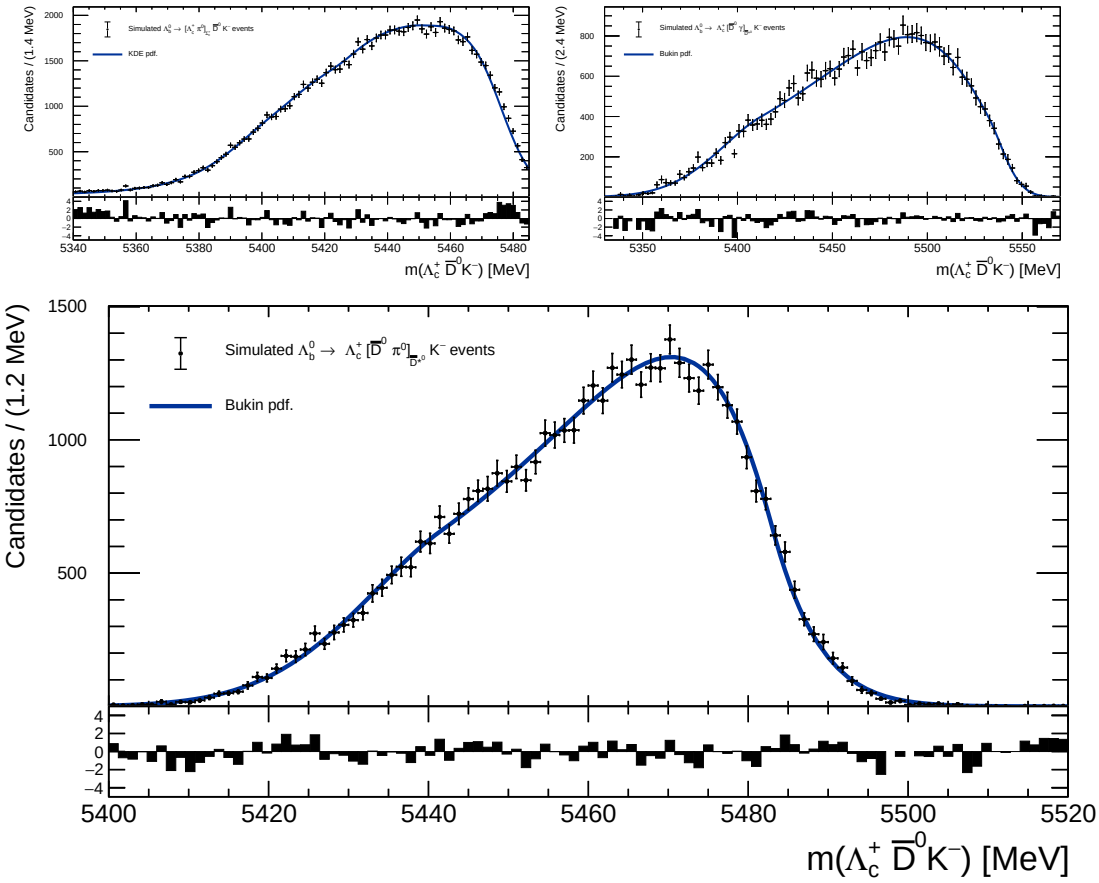



**Figure 4.11** *Comparison of* $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ *and* $\Lambda_b^0 \to \Sigma_c^+ (\to \Lambda_c^+ \pi^0) \overline{D}^0 K^-$ *decays. The latter one is color suppressed because of an internal* $W^-$ *emission.*

the two dominating contributions are also strongly anti-correlated. They share the same phase space. For this reason, when reconstructing the full $\overline{D}^{*0}$ 4-momentum an assumption is made of which particle was missing. This is described in greater detail in Section 4.8. The extracted yields are provided in Table 4.14. The ratio $\frac{N_\gamma}{N_\gamma + N_{\pi^0}}$ is 0.355 and $\frac{N_{\pi^0}}{N_\gamma + N_{\pi^0}}$ is 0.644 which agrees well with the $\overline{D}^{*0}$ branching ratio quoted in the PDG [18]. This confirms that the chosen parametrization for the MC shapes, and the correction procedures applied to MC reproduce the signal for these components well. The fit to data converges well and the residuals are minimal. Another important property extracted from this if is the fraction of $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} (\to \overline{D}^0 \gamma) K^-$ events under $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} (\to \overline{D}^0 \pi^0) K^-$. This is a contamination to our primary $ECC_{\pi^0}$ hypothesis. It's estimated to be $\sim 23\%$. The effects of this are described in Section 4.8, this is also taken into account when describing the signal model for a limit setting procedure in Section 4.15.

**Figure 4.12** *Fit to $\Lambda_c^+\overline{D}^0K^-$ inv. mass spectra of $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$, $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ and $\Lambda_b^0 \to \Sigma_c^+(\to \Lambda_c^+\pi^0)\overline{D}^0K^-$ MC samples*



## 4.8 Reconstruction of the Missing $\overline{D}^{*0}$ Momentum: Extended Cone Closure Method (ECC)

The $\pi^0$ and $\gamma$ candidates from the decay $\overline{D}^{*0} \to \overline{D}^0\pi^0/\gamma$ are reconstructed in the LHCb with very low efficiency. For this reason, the $\Lambda_b^0 \to \Lambda_c^+\overline{D}^0K^-$ decay candidates are selected, where the $\overline{D}^{*0}$ is partially reconstructed. The momentum of the $\pi^0$ or the $\gamma$ is missing. This is visible in the invariant mass spectrum of the $\Lambda_c^+\overline{D}^0K^-$ candidates. A number of candidates appear in a different mass range, below the $\Lambda_b^0$ mass of 5619 MeV. To be able to investigate the Dalitz plot of $\mathrm{m}^2(\Lambda_c^+\overline{D}^{*0})$- $\mathrm{m}^2(\overline{D}^{*0}K^-)$, the full 4-momentum of the $\overline{D}^{*0}$ candidate needs to be reconstructed. This is achieved by using a kinematic over-constraint method. Methods similar to this are used in semileptonic and neutrino physics. In hadronic spectroscopy analysis, such a method was introduced by Johns [132]. In this analysis the method is adapted for three-body decays, thus, it is called the `Extended Cone Closure` (ECC). In the ECC method, a number of mass and kinematic constraints are introduced to reconstruct the 4-momentum of a particle in several steps.

First the magnitude of the $\overline{D}^{*0}$ 4-momentum is analytically deduced, then the two angles which characterize the direction of travel of this particle are computed. The constraints are provided in the Table 4.16. One of the angles $\theta$ can be computed analytically, but the other, the polar angle $\phi$, is deduced numerically. This is achieved by scanning the space of possible directions of travel of a $\Lambda_b^0$ candidate around a cone and requiring that the 4-momentum of a reconstructed $\Lambda_b^0$ candidate best matches the vector pointing to the Primary Vertex (PV). The schematic diagram of the two angles (computed in $\overline{D}^0$ rest frame) is provided in Fig. 4.13. A detailed derivation is provided later in this section.

**Table 4.16** *Constraint applied in the Extended Cone Closure (ECC) method.*

| ECC Method constraints | |
|---|---|
| $m_{\Lambda_b^0}$ | 5619.6 MeV |
| $m_{\overline{D}^{*0}}$ | 2006.8 MeV |
| $m_X$ | 0 or 135 MeV |
| F. Dir. | Points to PV |

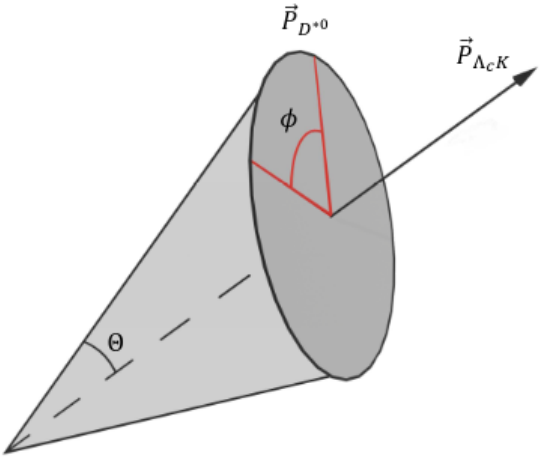



**Figure 4.13** *Diagram indicating two angles between a $\overline{D}^{*0}$ 3-momentum vector and a 3-momentum vector of a combined $\Lambda_c^+ K^-$ system. These angles are used in reconstructing the missing momentum of a $\overline{D}^{*0}$ candidate*

Partially reconstructed events inhabit an incorrect region of the Dalitz plot. Every point is shifted toward the lower values of the invariant mass due to the missing momentum of the $\overline{D}^{*0}$ candidate. To ilustrate this, the region of the $\Lambda_c^+ \overline{D}^0 K^-$ invariant mass spectrum is selected where the decays are only partially reconstructed ($5400 < \mathrm{m}(\Lambda_c^+ \overline{D}^0 K^-) < 5500$). Two Dalitz plots are created. First, with the $\mathrm{m}^2(\Lambda_c^+ \overline{D}^0)$ and $\mathrm{m}^2(\overline{D}^0 K^-)$ variables in the case, where the $\overline{D}^{*0}$ momentum is not reconstructed. The second one, with the variables with a reconstructed momentum, namely the $\mathrm{m}^2(\Lambda_c^+ \overline{D}^{*0})$ and $\mathrm{m}^2(\overline{D}^{*0} K^-)$. A contour of the kinematically allowed region for $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\ \pi^0/\gamma) K^-$ decays is also overlaid in the figures. Partially reconstructed case is shown in Fig. 4.14 and fully reconstructed case in Fig. 4.15. It is visible that reconstructed candidates now appear in the correct region within the Dalitz plane. Some structures can be identified in the partially reconstructed case, but they appear in the incorrect region. Also, resonances are smeared out in the dalitz plane.

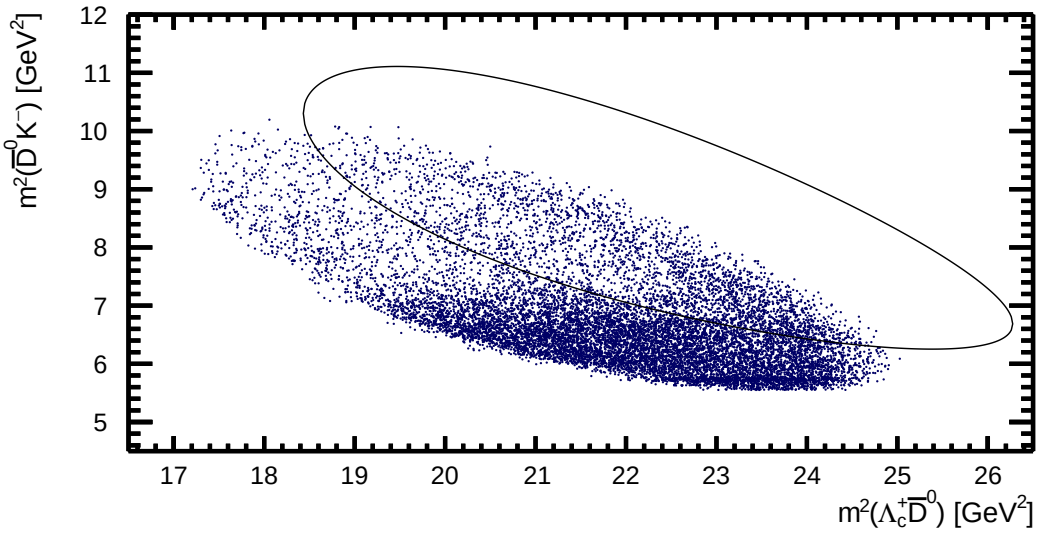



**Figure 4.14** *$m^2(\Lambda_c^+\overline{D}^{*0})$- $m^2(\overline{D}^{*0}K^-)$ Dalitz plot where decay $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\ \pi^0/\gamma)K^-$ is partially reconstructed. The missing momentum of a $\pi^0$ or $\gamma$ candidate means that the reconstructed candidates of the decay occupy an incorrect region in the Dalitz plane. The correct kinematically allowed region for the $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\ \pi^0/\gamma)K^-$ decays is indicated by the black contour.*

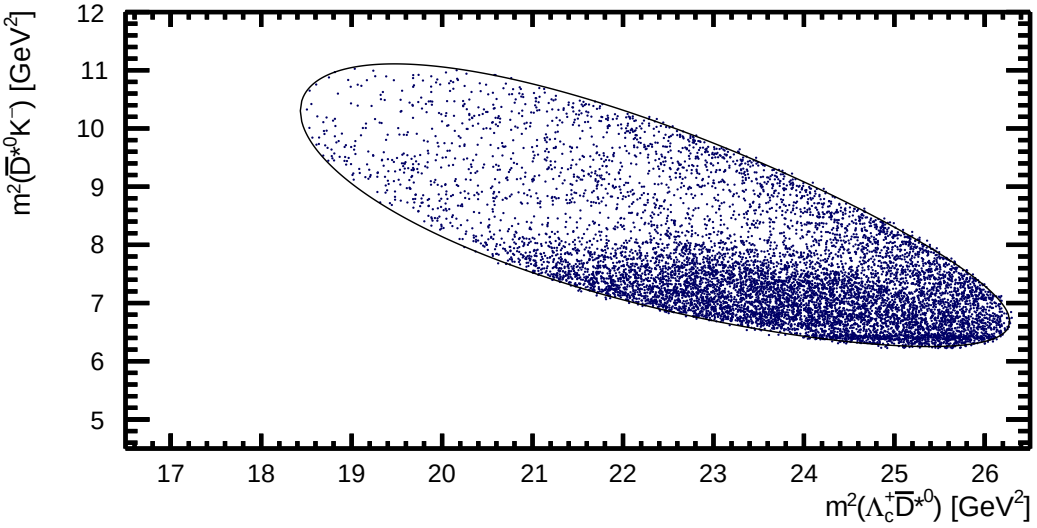



**Figure 4.15** *$m^2(\Lambda_c^+\overline{D}^{*0})$- $m^2(\overline{D}^{*0}K^-)$ Dalitz plot where decay $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\ \pi^0/\gamma)K^-$ is fully reconstructed using Extended Cone Closure (ECC) method. The correct kinematically allowed region for the $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\ \pi^0/\gamma)K^-$ is indicated by the black contour. It is visible that the candidates of $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\ \pi^0/\gamma)K^-$ are now appearing within the correct region in the Dalitz plane.*

### 4.8.1 Derivation of the ECC: Reconstruction of the Magnitude of $\overline{D}^{*0}$ momentum

One may start with applying the law of energy conservation to the decay $A \to BX$. Where X is an unknown missing particle, $A$ is a particle with some missing momentum and $B$ is fully reconstructed and known. In the case of this analysis, $A \equiv \overline{D}^{*0}$, $B \equiv \overline{D}^{0}$ and X may be a $\pi^0$ or a $\gamma$, but, as discussed before, predominantly it will be a $\pi^0$.

$$E_{\overline{D}^{*0}} = E_{\overline{D}^{0}} + E_X. \tag{4.1}$$

Following the energy - momentum relation, the following expression is obtained for the energy of a $\overline{D}^{*0}$. It is expressed here in the rest frame of a $\overline{D}^{0}$ candidate. This makes the equations shorter by dropping some of the cross terms.

$$\begin{aligned} E_{\overline{D}^{*0}} &= m_{\overline{D}^{0}} + \sqrt{p_X^2 + m_X^2} \\ &= m_{\overline{D}^{0}} + \sqrt{p_{\overline{D}^{*0}}^2 + m_X^2}. \end{aligned} \tag{4.2}$$

In the $\overline{D}^{0}$ rest frame, the $\overline{D}^{*0}$ and the $\gamma$ or the $\pi^0$ decay back-to-back. Also, the momenta of the two particles are equal, $p_{\overline{D}^{*0}} = p_X$. The mass of the $\overline{D}^{0}$ candidate is fixed and equal to the rest mass. The two unknowns which are left are the $p_{\overline{D}^{*0}}$ and $m_X$. Here, a hypothesis needs to be established. This is where the first assumption (or a constraint) comes in. $m_X \equiv m_\gamma = 0$ MeV or $m_X \equiv m_{\pi^0} = 135$ MeV. Namely, the momentum of a $\overline{D}^{*0}$ is reconstructed either with a missing $\gamma$ or a missing $\pi^0$. These two hypotheses were called $ECC_\gamma$ and $ECC_{\pi^0}$. It is always important to refer to a certain hypothesis when discussing results where the $\overline{D}^{*0}$ momentum is reconstructed.

By using energy - momentum relation on the $\overline{D}^{*0}$ and the expression obtained in Eq. (4.2), one can arrive at an analytical expression for the magnitude of the $\overline{D}^{*0}$ momentum $p_{\overline{D}^{*0}}$.

$$
\begin{aligned}
m^2_{\overline{D}^{*0}} &= E^2_{\overline{D}^{*0}} - p^2_{\overline{D}^{*0}} \\
&= \left(m_{\overline{D}^0} + \sqrt{p^2_{\overline{D}^{*0}} + m^2_X}\right)^2 - p^2_{\overline{D}^{*0}} \\
&= m^2_{\overline{D}^0} + 2 \cdot m_{\overline{D}^0} \cdot \sqrt{p^2_{\overline{D}^{*0}} + m^2_X} + m^2_X + \cancel{p^2_{\overline{D}^{*0}}} - \cancel{p^2_{\overline{D}^{*0}}} \\
2 \cdot m_{\overline{D}^0} \cdot \sqrt{m^2_X + p^2_{\overline{D}^{*0}}} &= m^2_{\overline{D}^{*0}} - m^2_{\overline{D}^0} - m^2_X \\
\sqrt{m^2_X + p^2_{\overline{D}^{*0}}} &= \frac{m^2_{\overline{D}^{*0}} - m^2_{D^0} - m^2_X}{2 \cdot m_{D^0}} \\
m^2_X + p^2_{\overline{D}^{*0}} &= \left(\frac{m^2_{\overline{D}^{*0}} - m^2_{D^0} - m^2_X}{2 \cdot m_{D^0}}\right)^2 \\
p^2_{\overline{D}^{*0}} &= \left(\frac{m^2_{\overline{D}^{*0}} - m^2_{D^0} - m^2_X}{2 \cdot m_{D^0}}\right)^2 - m^2_X
\end{aligned}
\tag{4.3}
$$

$$
p_{\overline{D}^{*0}} = \sqrt{\left(\frac{m^2_{\overline{D}^{*0}} - m^2_{\overline{D}^0} - m^2_X}{2 \cdot m_{\overline{D}^0}}\right)^2 - m^2_X}. \tag{4.4}
$$

Here, all of the terms are by now established except for $m_{\overline{D}^{*0}}$. The world average value is taken from the PDG [133]. It is 2007 MeV.

It is important to note here that unlike in some of the implementations of similar methods in semileptonic or neutrino physics analyses, $\overline{D}^{*0}$ has to be constrained. The mass of the particle for which the 4-momentum is partially reconstructed ($\overline{D}^{*0}$ in this case) is used to determine the magnitude of $\overline{D}^{*0}$ 4-momentum. It is exactly in this place here where putting in a different missing particle (e.g. $\Sigma_c^+$) would yield a different value of the magnitude of the 4-momentum. This is why the Extended Cone Closure method is applicable to multiple analyses with partially reconstructed decays.

### 4.8.2 Derivation of the ECC: Computing the Polar Angle $\theta$

Now that the magnitude of the $\overline{D}^{*0}$ momentum has been computed, the flight direction of a $\overline{D}^{*0}$ candidate needs to be recovered. It is assumed that in the data

the $\overline{D}^{*0}$, the $\Lambda_c^+$ and the $K^-$ candidates are all the decay daughters of a $\Lambda_b^0$ baryon from the decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$. Since $\Lambda_c^+$ and $K^-$ are fully reconstructed, in the calculations, they are combined into a single object - the $\Lambda_c^+ K^-$ system. Treating the $\Lambda_c^+ K^-$ system as one object, the expression for the angle of separation between $\overline{D}^{*0}$ candidate and $\Lambda_c^+ K^-$ system can be computed.

$$\begin{aligned} m^2_{\Lambda_b^0} &= (E_{\overline{D}^{*0}} + E_{\Lambda_c^+ K^-})^2 - (\vec{p}_{\overline{D}^{*0}} + \vec{p}_{\Lambda_c^+ K^-})^2 \\ &= \underline{E^2_{\overline{D}^{*0}}} + 2 \cdot E_{\overline{D}^{*0}} \cdot E_{\Lambda_c^+ K^-} + \underline{\underline{E^2_{\Lambda_c^+ K^-}}} - \underline{p^2_{\overline{D}^{*0}}} - 2 \cdot p_{\overline{D}^{*0}} \cdot p_{\Lambda_c^+ K^-} \cdot \cos\theta - \underline{\underline{p^2_{\Lambda_c^+ K^-}}} \\ &= \underline{m^2_{\overline{D}^{*0}}} + \underline{\underline{m^2_{\Lambda_c^+ K^-}}} + 2 \cdot E_{\overline{D}^{*0}} \cdot E_{\Lambda_c^+ K^-} - 2 \cdot p_{\overline{D}^{*0}} \cdot p_{\Lambda_c^+ K^-} \cdot \cos\theta \end{aligned}$$

$$2 \cdot p_{\overline{D}^{*0}} \cdot p_{\Lambda_c^+ K^-} \cdot \cos\theta = m^2_{\overline{D}^{*0}} + m^2_{\Lambda_c^+ K^-} + 2 \cdot E_{\overline{D}^{*0}} \cdot E_{\Lambda_c^+ K^-} - m^2_{\Lambda_b^0}$$

$$\cos\theta = \frac{m^2_{\overline{D}^{*0}} + m^2_{\Lambda_c^+ K^-} + 2 \cdot E_{\overline{D}^{*0}} \cdot E_{\Lambda_c^+ K^-} - m^2_{\Lambda_b^0}}{2 \cdot p_{\overline{D}^{*0}} \cdot p_{\Lambda_c^+ K^-}} \tag{4.5}$$

Here, another variable is dealt with as a constraint, namely, the mass of a decaying mother particle, the $\Lambda_b^0$. It has to be fixed here to the world average value of 5619 MeV. It is here where the drawback of the ECC method comes in. In the case of this analysis, combining $\Lambda_c^+ \overline{D}^{*0} K^-$ candidates and plotting the invariant mass spectrum would result in a Delta function at 5619 MeV. This is because, $\overline{D}^{*0}$ and $\Lambda_b^0$ masses are constrained in the method, and the magnitude of the momentum of the $\overline{D}^{*0}$ candidate is such that after the reconstruction, the correct mass of the $\Lambda_b^0$ is obtained.

### 4.8.3 Derivation of the ECC: Obtaining the Azimuthal Angle $\phi$

No analytic expression for the azimuthal angle $\phi$ was found. A numerical technique is employed to find the best matching angle $\phi$. A new direction vector of the momentum of the $\Lambda_b^0$ candidate is constructed. First, it is not known what the

true direction of the $P_{\overline{D}^{*0}}$ vector is. The $P_{\overline{D}^{*0}}$ is matched with the $P_{\Lambda_c^+ K^-}$ vector. This vector then undergoes a number of rotations to match it best to the flight direction of the $\Lambda_b^0$ candidate. The true flight direction is computed separately, using the coordinates of the primary vertex and the $\Lambda_b^0$ decay vertex (the $\Lambda_c^+$ origin vertex). The information on the position of vertices is provided by the VELO sub-detector in the LHCb. It is described in more detail in Section 1.3.1. A very important aspect of the VELO is its very good vertex resolution $\mathcal{O}(10)$ $\mu$m. The two rotations are highlighted in Fig. 4.13. The first rotation is by the polar angle $\theta$, computed previously. This ensures the correct angular separation between $\overline{D}^{*0}$ and $\Lambda_c^+ K^-$ momenta vectors. At this stage, though, the direction of the $\overline{D}^{*0}$ candidate is still not fully determined. It is only known that the $\overline{D}^{*0}$ vector lies somewhere on a cone, extended around the $\Lambda_c^+ K^-$ direction vector. The azimuthal angle $\phi$ is then computed by incrementally rotating the $\overline{D}^{*0}$ vector (with 1000 increments). The agreement between the fully reconstructed $\Lambda_b^0$ direction vector and the direction vector computed from $\Lambda_b^0$ origin and end vertices is checked with each increment. The value where the agreement is best is then retained. The full $\overline{D}^{*0}$ 4-momentum is then reconstructed using the magnitude and 2 angles $\theta$ and $\phi$.

Scanning around the cone in such a way could reconstruct the 4-momentum of a particle with a 2-fold ambiguity. The azimuthal angle $\phi$ could in some cases be deduced at a certain value, but the value on the opposite side of the cone, 180 °apart could also work. The $\Lambda_b^0$ mass constraint ensures that the momentum is reconstructed with no ambiguity. This constraint is also needed in computing the $\cos(\theta)$ - the polar angle of separation, seen in Eq. (4.5). Note that this equation holds in the $\overline{D}^0$ rest frame, the $\Lambda_b^0$ mass here already has to be constrained. Also, the energy of the missing particle $E_X$ indirectly appears in this equation through Eq. (4.2). To show, that for every minimization step there is only one minimum and that the angle $\phi$ is deduced unambiguously, the value of this angle is inspected in the full range from 0 to 360 °. This is shown in figure Fig. 4.16.

$$1 - TFDV(\Lambda_b^0) \cdot AFDV(\Lambda_b^0)\,,$$

where `TFDV` is `Test Flight Direction Vector`, obtained by computing the unit direction vector of a now fully reconstructed $\Lambda_b^0$ candidate in the lab frame after each rotation of the $\overline{D}^{*0}$ flight direction vector around a cone. And `AFDV` is `Actual`

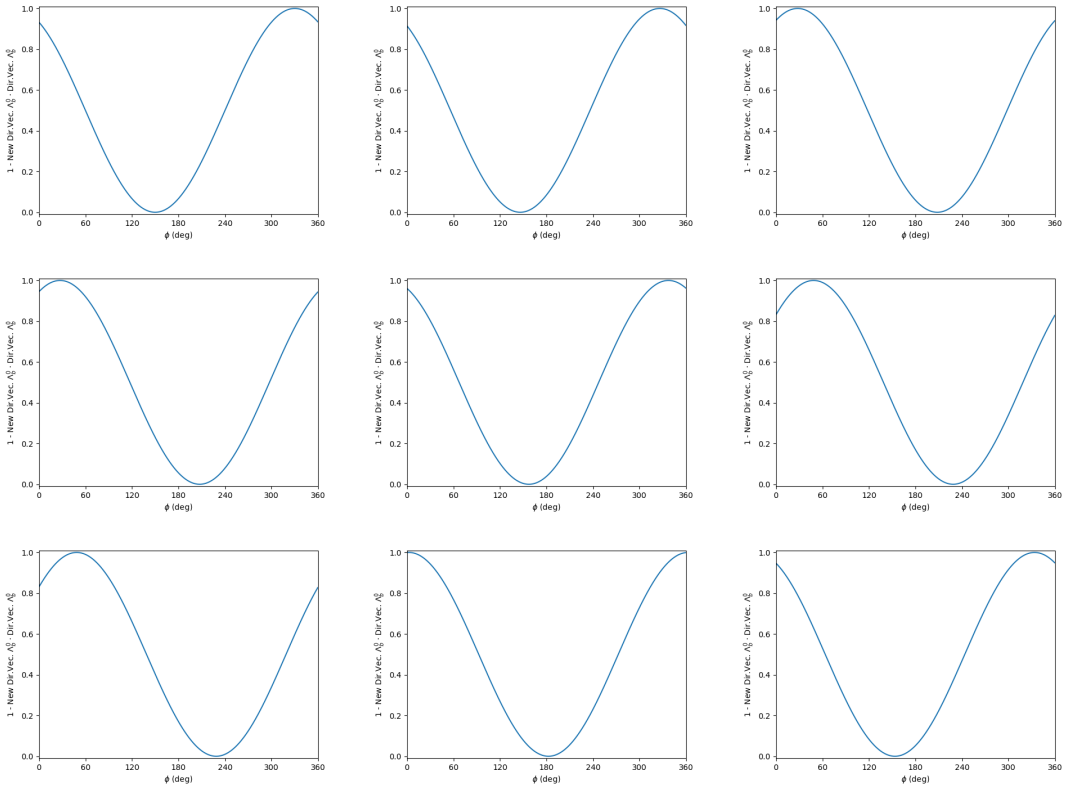


**Figure 4.16** *Scans of the azimuthal angle $\phi$. A number of successfully reconstructed $\overline{D}^{*0}$ candidates are selected at random. The scan in the angle is performed to ensure there is only one minimum, no 2-fold ambiguity in reconstruction of the flight direction of a $\overline{D}^{*0}$.*

`Flight Direction Vector`, obtained from end-vertex information of $\Lambda_b^0$ and $\Lambda_c^+$ candidates recorded by VELO detector.

### 4.8.4 ECC-sensitive Regions in $\Lambda_c^+ \overline{D}^0 K^-$ Invariant Mass Spectrum

The Extended Cone Closure method is applied to all of the candidates in the selected m($\Lambda_c^+ \overline{D}^0 K^-$) invariant mass range. The selection is described in Section 4.4. Depending on the hypothesis (the missing particle with which the $\overline{D}^{*0}$ 4-momentum is reconstructed), the method is successful on a larger or smaller fraction of the candidates. This can be seen in Fig. 4.17. These regions align with the kinematically allowed regions for the decays $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} (\to \overline{D}^0 \gamma) K^-$ or $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} (\to \overline{D}^0 \pi^0) K^-$ to happen. It can be seen that the region where the $\pi^0$ candidate was assumed to be missing, is contained within the region where the $\gamma$ was missing. This arises because of a zero rest mass of a photon, whereby

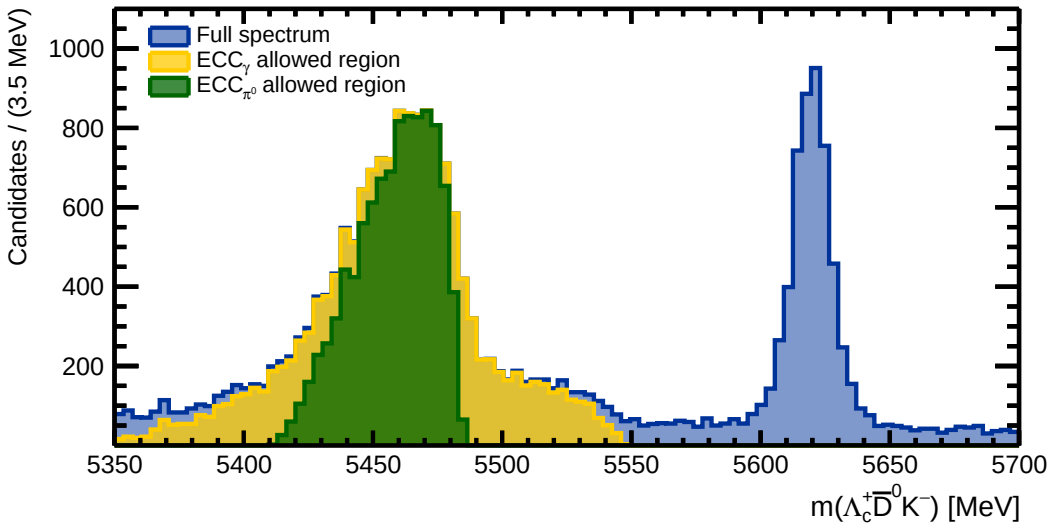



**Figure 4.17** *ECC-sensitive regions of the $\Lambda_c^+\overline{D}^0K^-$ inv. mass spectrum.*

the decay kinematics allow the decay $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$ to happen in a wider range in the m($\Lambda_c^+\overline{D}^0K^-$) invariant mass spectrum than that for the $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ decays. It is visible that when the m($\Lambda_c^+\overline{D}^0K^-$) $> 5550$ none of the candidates are successfully reconstructed anymore. This means that the candidates beyond the ECC acceptance could not have decayed through $\overline{D}^{*0}$ resonance. It is particularly well illustrated by examining the right hand side of the $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ ECC reconstruction region. A sharp cut-off is visible. This is because if the invariant mass in the $\Lambda_c^+\overline{D}^0K^-$ system is higher than a certain value (m($\Lambda_b^0$)- $m(\pi^0)$) the candidates there could not have decayed through $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$, because the difference in mass is less than the rest mass of a $\pi^0$. The mass region involving the missing $\gamma$ is decided by the rest mass of a reconstructed $\overline{D}^{*0}$ candidate. Here it is important to remember that the branching fraction of the $\overline{D}^{*0}$ to final state involving a $\pi^0$ is about 65% and to the one involving a $\gamma$ - about 35%. For this reason, the main focus of the analysis is $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$, the other channel is also included in most results as a crosscheck.

After careful investigation into the waysto disentangle the two contributions, no way was found to separate the $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ from the $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$ as well as the small component with the $\Lambda_b^0 \to \Sigma_c^+(\to \Lambda_c^+\pi^0)\overline{D}^0K^-$ events. An investigation of how this impacts the Dalitz plot was performed and the findings were that if a candidate is reconstructed with a

wrong hypothesis, the point in the Dalitz plot is symmetrically smeared out (the candidates are reconstructed with worse resolution). This is explained later in Section 4.8.6. This means that no sharp structures are introduced and Dalitz plot can be examined as normal. This is also one of the reasons why it was chosen to work on LHCb Run2 Data - using Run1 data would introduce another 20% of candidates again reconstructed with a worse resolution, this would not have a positive effect on the final results but would introduce additional complexity.

### 4.8.5 ECC with DTF Variables

While developing the ECC method for this analysis, a few different cases of reconstructing the $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ decays were investigated. Namely, the effects of using a Decay Tree Fitter (see Section 4.6) with $\overline{D}^0$ mass constraint or with $\overline{D}^0$ and $\Lambda_c^+$ mass constraints. Note, that unlike in the usual case, where the vertex constraints are also applied, they can not be applied here. This is because the partially reconstructed $\Lambda_b^0$ candidate direction vector does not point to the primary vertex. It was noticed that using both $\Lambda_c^+$ and $\overline{D}^0$ mass constraints for the DTF not only improves the mass resolution but also the results in a higher rate of successfully reconstructed $\overline{D}^{*0}$ candidates by the ECC. Since the known mass of a $\overline{D}^{*0}$ and the mass of a missing particle is used as the constraints to the ECC method, it makes sense to re-fit the decay tree with the mass of a $\overline{D}^0$ candidate fixed to the known value using the `DecayTreeFitter` in `DaVinci`. It can be seen that without this constraint, the mass resolution after ECC is worse and there is a large number of not reconstructed candidates. This is visible by the not populated region in the $\overline{D}^0$ invariant mass spectrum. The candidates with the $\overline{D}^0$ invariant mass values higher that 1860 MeV are not reconstructed. This can be seen in Fig. 4.18.

The reason for the drop in the number of reconstructed candidates is again related to the kinematically allowed values, in this case - of the $\overline{D}^0$ mass. For the valuos of the $\overline{D}^0$ invariant mass larger than the (m($\overline{D}^{*0}$)- m($\overline{D}^0$)), the decays can not happen through $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} (\to \overline{D}^0 \pi^0) K^-$, because this difference is then smaller than the rest mass of the $\pi^0$. This effect was given a name the $\pi^0$ *cut*. It is also the reason why the shape of the $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} (\to \overline{D}^0 \pi^0) K^-$ component (visible in Fig. 4.10) has a sharp drop off on the right. This inefficiency was solved by applying the DTF and recalculating the 4-momenta of $\overline{D}^0$ candidates with the

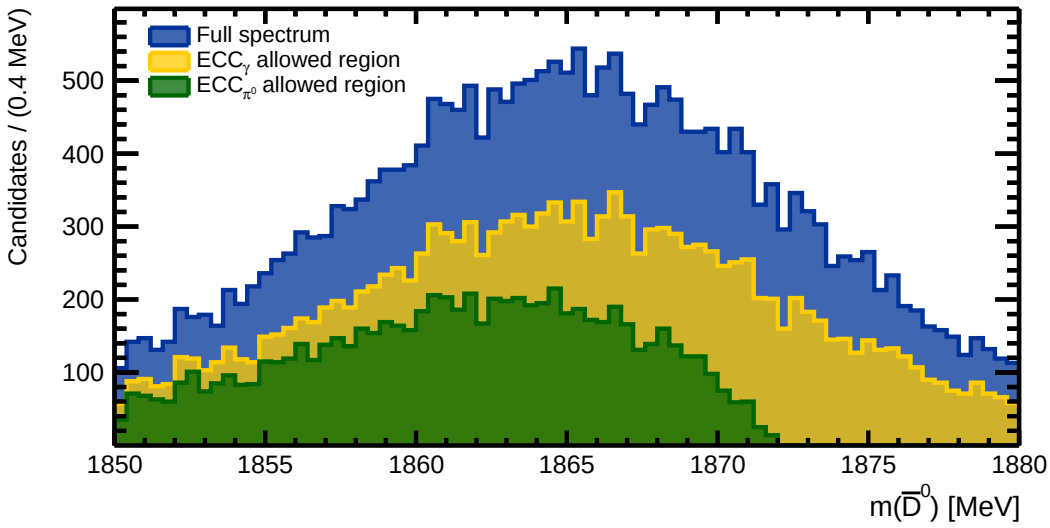



**Figure 4.18** *$\overline{D}^0$ invariant mass spectrum with overlaid components for all selected candidates of decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ (including both fully and partially reconstructed decays ) - full range, candidates which pass $ECC_\gamma$ and candidates which pass $ECC_{\pi^0}$ conditions.*

$\overline{D}^0$ mass fixed to that of the PDG value.

### 4.8.6 Resolution of the Extended Cone Closure Method

The resolution of the ECC method in the m($\Lambda_c^+ \overline{D}^{*0}$) and m($\overline{D}^{*0} K^-$) invariant mass spectra is computed using the information from the Monte Carlo simulation. The MC samples used here are those for the decays $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0 \pi^0) K^-$ and $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0 \gamma) K^-$. The local resolution is obtained by comparing the reconstructed m($\Lambda_c^+ \overline{D}^{*0}$) and m($\overline{D}^{*0} K^-$) distributions to the true ones. The `MC Truth` variables are available with the simulation and provide information on the true values of various parameters in a decay. In this case, the `X,Y,Z,E` components of the 4-momenta of the particles are used.

The resolution in m($\Lambda_c^+ \overline{D}^{*0}$) for the $ECC_{\pi^0}$ hypothesis ranges from 3 to 6 MeV, and for the $ECC_\gamma$ hypothesis from 10 to 25 MeV, depending on the mass region. In m($\overline{D}^{*0} K^-$), the resolution ranges from 2 to 10 MeV in the $ECC_{\pi^0}$ case and from 5 to 32 MeV in the $ECC_\gamma$ case. It can be seen that the $\overline{D}^{*0}$ candidates from the $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0 \pi^0) K^-$ dacays are reconstructed at a higher resolution. Again, this is because of the non-zero $\pi^0$ rest mass. This

is illustrated for m($\Lambda_c^+\overline{D}^{*0}$) variable for both $ECC_{\pi^0}$ and $ECC_\gamma$ hypotheses in Figures 4.19 and 4.20 respectively and for m($\overline{D}^{*0}K^-$) variable for both $ECC_{\pi^0}$ and $ECC_\gamma$ hypotheses in Figures 4.21 and 4.22, respectively.

Since $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ is the dominant process in the decay $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}K^-$ and the resolution is better, the $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ component is chosen as the main focus of the analysis. The $\sim 35\%$ of the decays $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$ which are then reconstructed incorrectly do not cause a big negative effect on the Dalitz plot. They are just reconstructed at a lower resolution. The $\sigma$ is computed at different points of the m($\Lambda_c^+\overline{D}^{*0}$) and the m($\overline{D}^{*0}K^-$) invariant mass spectrum. Spline interpolation is used to obtain the shape of a resolution function. This study was primarily done to investigate how the isolated resolution of ECC method changes in different regions of the m($\Lambda_c^+\overline{D}^{*0}$) and m($\overline{D}^{*0}K^-$) invariant mass spectra. A separate study is performed including all the resolution effects, detector and ECC combined. It is described in Section 4.13.

### 4.8.7 ECC Resolution with Incorrect Hypothesis Assignment

A study is also performed to investigate the effects incorrect ECC hypothesis assignment. The method was applied with the $ECC_{\pi^0}$ hypothesis on the $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$ MC sample and the $ECC_\gamma$ hypothesis on the $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ sample. It is concluded that assigning the $ECC_\gamma$ or $ECC_{\pi^0}$ hypothesis incorrectly does not reduce too much. It symmetrically smears out the reconstructed momenta around some central value, but this effect is smaller than the overall detector resolution and thus does not interfere with the analysis. This can be seen in Fig. 4.23 and Fig. 4.23.

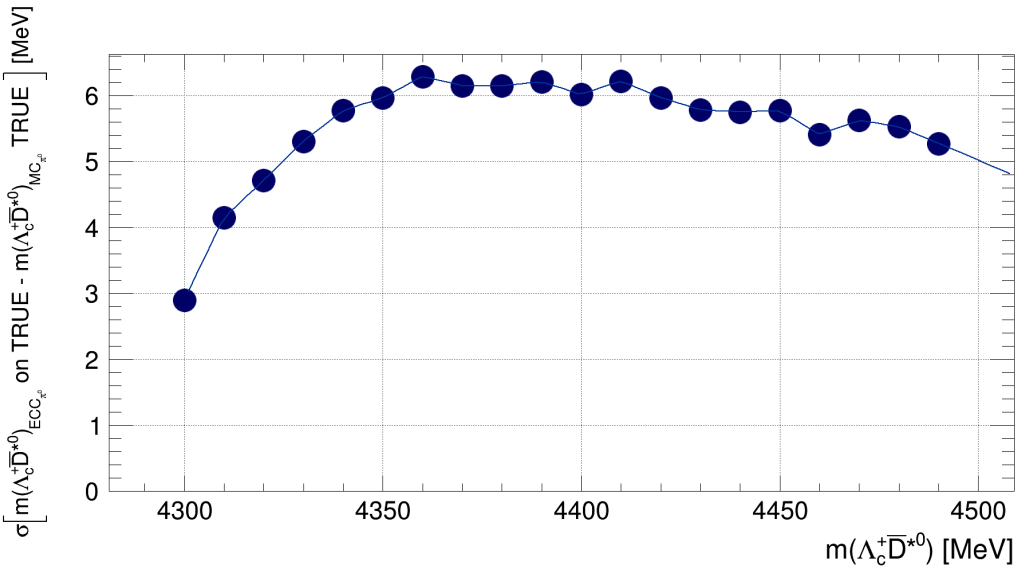



**Figure 4.19** *Resolution of the ECC method in $m(\Lambda_c^+\overline{D}^{*0})$ obtained by performing ECC on the MC Truth variables and comparing to the case where $\overline{D}^{*0}$ 4-momentum was deduced from $P_{\Lambda_b^0} - P_{\Lambda_c^+} - P_K$. Here $ECC_{\pi^0}$ is applied to MC $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ `TRUTH` variables.*

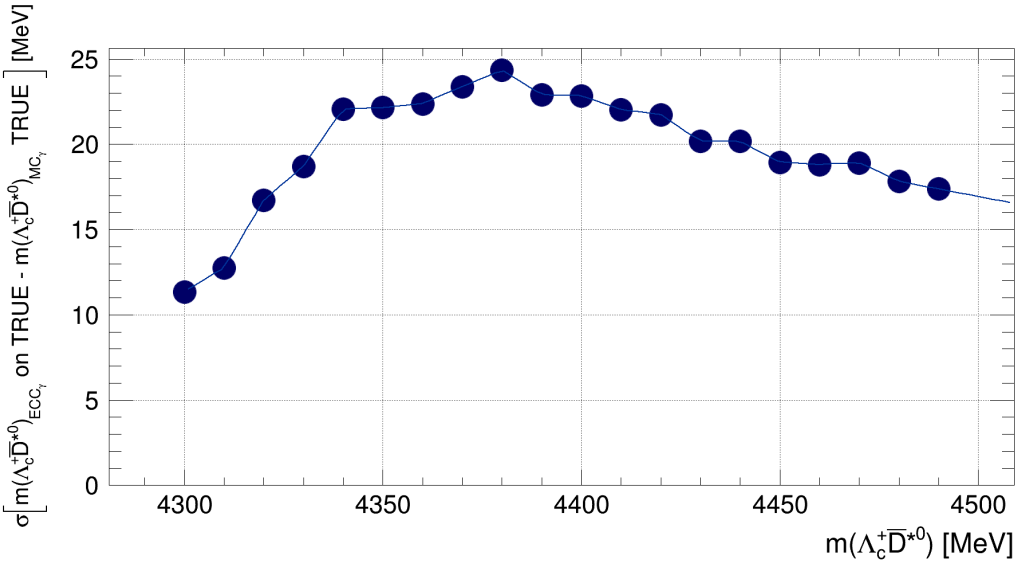



**Figure 4.20** *Resolution of the ECC method in $m(\Lambda_c^+\overline{D}^{*0})$ obtained by performing ECC on the MC Truth variables and comparing to the case where $\overline{D}^{*0}$ 4-momentum was deduced from $P_{\Lambda_b^0} - P_{\Lambda_c^+} - P_K$. Here $ECC_{\gamma}$ is applied to MC $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$ `TRUTH` variables.*

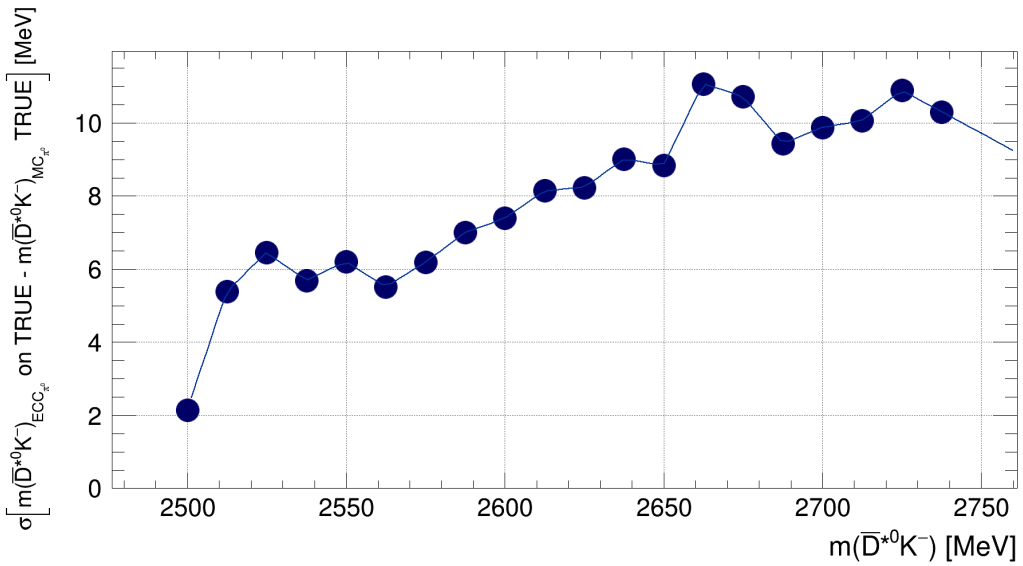



**Figure 4.21** *Resolution of the ECC method in $m(\overline{D}^{*0}K^-)$ obtained by performing ECC on the MC Truth variables and comparing to the case where $\overline{D}^{*0}$ 4-momentum is deduced from $P_{\Lambda_b^0} - P_{\Lambda_c^+} - P_K$. Here $ECC_{\pi^0}$ is applied to MC $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$ `TRUTH` variables.*

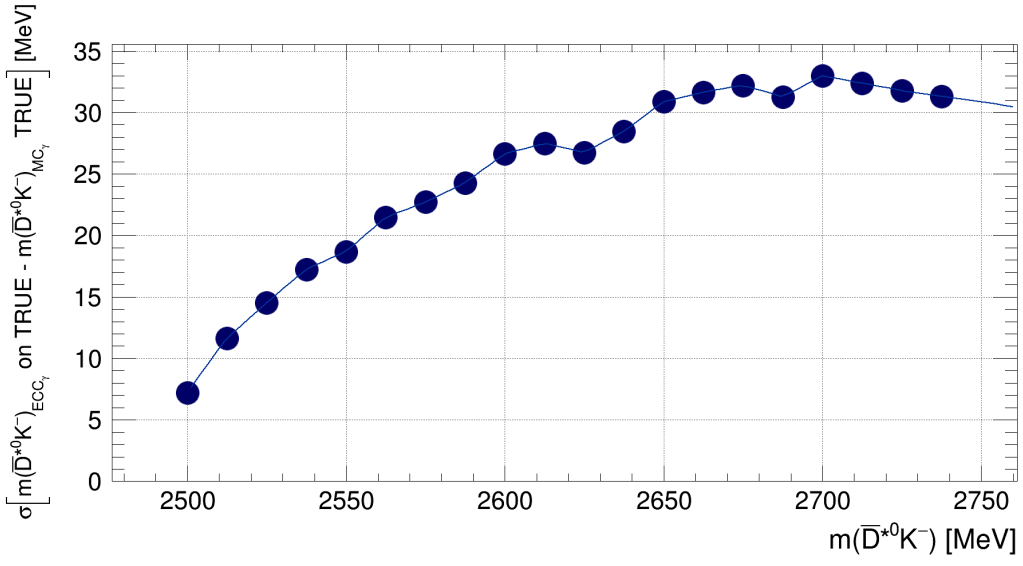



**Figure 4.22** *Resolution of the ECC method in $m(\overline{D}^{*0}K^-)$ obtained by performing ECC on the MC Truth variables and comparing to the case where $\overline{D}^{*0}$ 4-momentum is deduced from $P_{\Lambda_b^0} - P_{\Lambda_c^+} - P_K$. Here $ECC_\gamma$ is applied to MC $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$ `TRUTH` variables.*

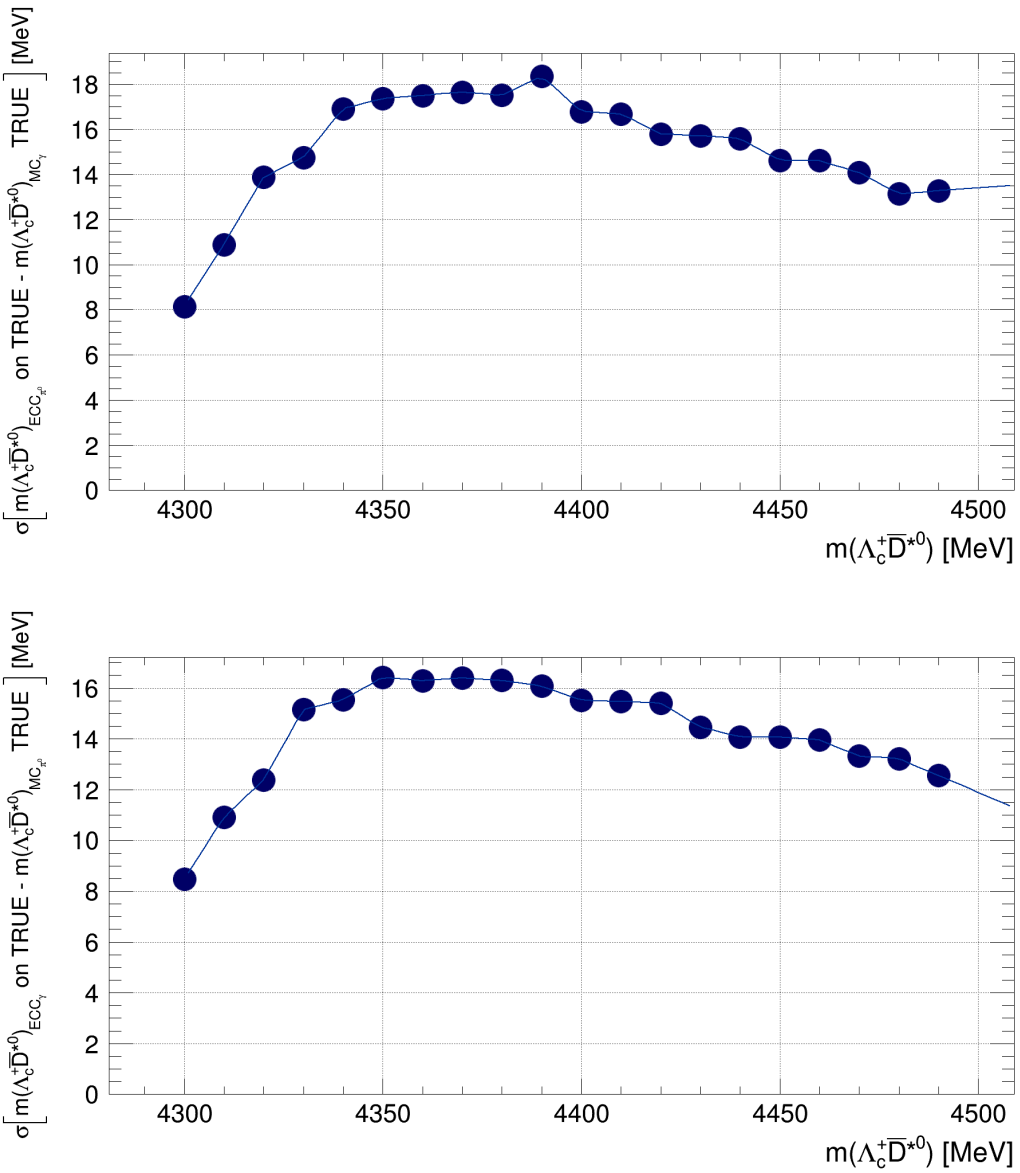



**Figure 4.23** *Resolution of the ECC method in $m(\Lambda_c^+ \overline{D}^{*0})$ obtained by performing ECC on the MC Truth variables for $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ under $ECC_\gamma$ hypothesis and vice versa and compared to the case where $\overline{D}^{*0}$ 4-momentum is deduced from via $P_{\Lambda_b^0} - P_{\Lambda_c^+} - P_K$*

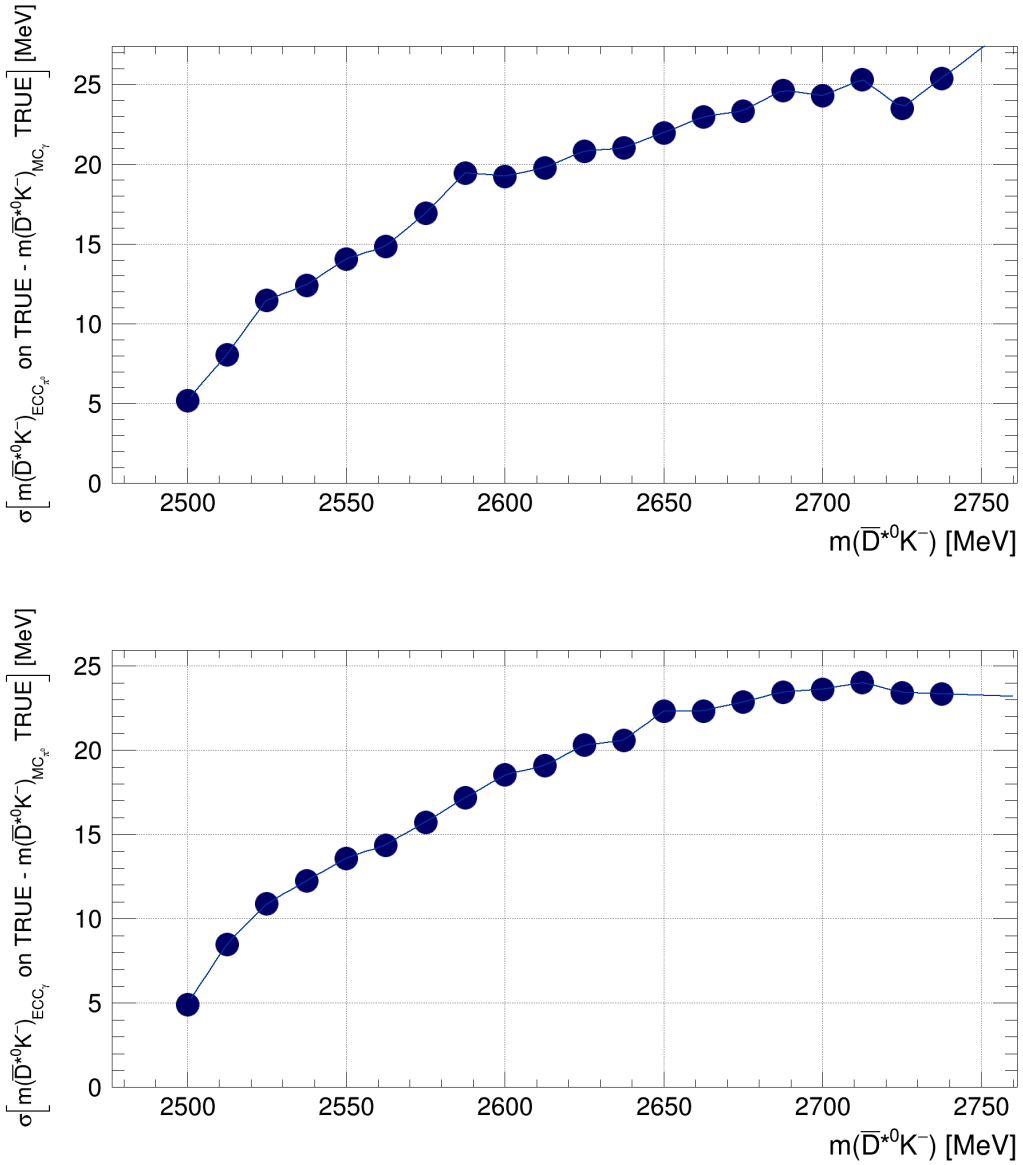



**Figure 4.24** *Resolution of the ECC method in $m(\overline{D}^{*0}K^-)$ obtained by performing ECC on the MC Truth variables for $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ under $ECC_\gamma$ hypothesis and vice versa and compared to the case where $\overline{D}^{*0}$ 4-momentum is deduced from via $P_{\Lambda_b^0} - P_{\Lambda_c^+} - P_K$*

## 4.9 $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ Decay with ECC Applied

### 4.9.1 Simulated $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ Decay with $P_c^+$ States Included

Before the $\overline{D}^{*0}$ 4-momentum is fully reconstructed, a Dalitz plot for the decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ can not be constructed. Some theory predictions already hint that the $P_c^+$ states ($P_c(4312)^+$, $P_c(4440)^+$, $P_c(4457)^+$) should be identifiable in LHCb Run2 data. A simulated sample containing all three of these states is used to illustrate, how these resonances would appear in the $\Lambda_c^+ \overline{D}^{*0}$- $\overline{D}^{*0} K^-$ Dalitz plot. In the simulation a Breit-Wigner line shape is used for all three $P_c^+$ states. In reality, interference effects, proximity to meson-baryon thresholds and the decay kinematics can cause the shapes to be complicated. A $\Lambda_c^+ \overline{D}^{*0}$ spectrum obtained with these more complete models predicting the shapes of $P_c^+$ states is provided in [17].

The $P_c^+$ states are predicted to be narrow. $P_c(4440)^+$ and $P_c(4457)^+$ are close in mass and could appear as a single smeared out band. The side bands of these shapes could also overlap. The $P_c(4312)^+$ should appear at the very edge of the phase space. In some theory predictions, the predicted yield is higher for the $P_c(4312)^+$ than for the other two studied states. This can be seen in Table 2.2 in Section 2.4. An important check to do here is to look for other, known narrow states and to make sure that the ECC method does not disfigure these shapes in the Dalitz plot. The known resonances are not included in the simulation for clarity. The Dalitz-like plot of these simulated events and the projection to the m($\Lambda_c^+ \overline{D}^{*0}$) invariant mass spectrum are provided in Fig. 4.25 and Fig. 4.26, respectively. Examining the projection and comparing it to the observed m($\Lambda_c^+ \overline{D}^{*0}$) invariant mass spectrum, show in Fig. 4.29 it becomes clear that the observed $\Lambda_c^+ \overline{D}^{*0}$ distribution is not phase space like. Also, no visible $P_c^+$ signatures are observed in the regions indicated by the dashed lines in Fig. 4.26.

**Figure 4.25** *Dalitz-like plot $m(\Lambda_c^+\overline{D}^{*0})$- $m(\overline{D}^{*0}K^-)$ constructed with simulated events. The simulation is that of the phase space determined by the decay kinematics. Three $P_c^+$ resonances which are the main focus of this analysis are included in the MC generation step. This is used to highlight an approximate location of where these states might appear. As well as show their position in relation to each other and the edge of the Dalitz plot.*

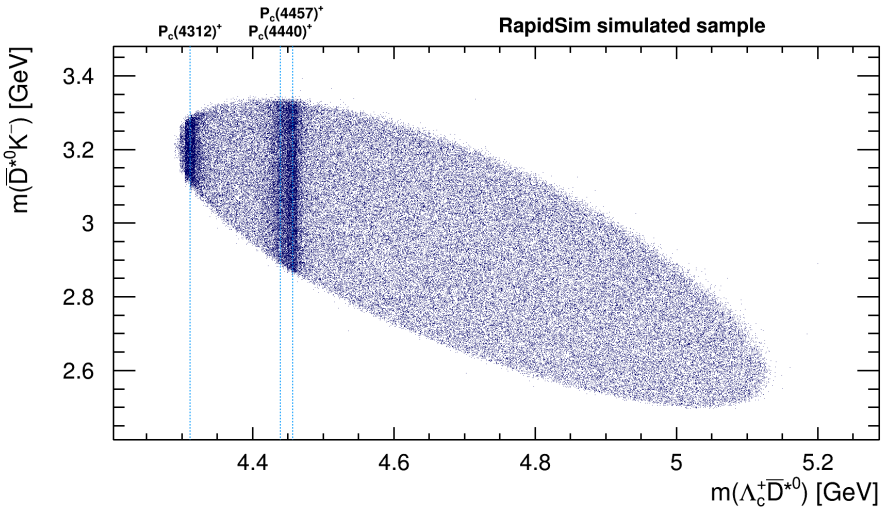
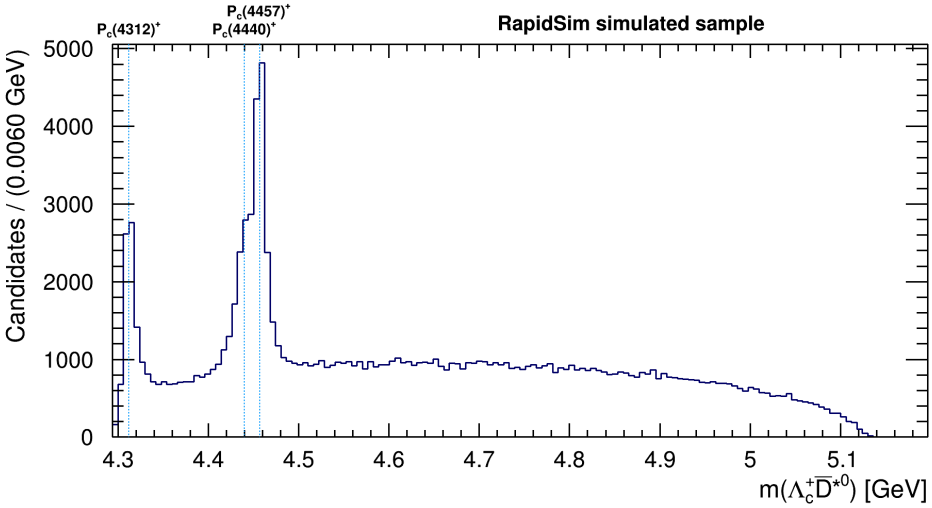



**Figure 4.26** *$m(\Lambda_c^+\overline{D}^{*0})$ invariant mass spectrum of simulated events with $P_c^+$ resonances included. This is used to illustrate the possible location of where an excess in events might be expected. The $P_c(4440)^+$ and $P_c(4457)^+$ here overlap each other and resemble a single wider resonance. The shape of the $P_c(4312)^+$ is impacted by its proximity to the edge of the phase space.*



### 4.9.2 $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ Dalitz Plot and Projections

$\mathrm{m}^2(\Lambda_c^+ \overline{D}^{*0})$ against $\mathrm{m}^2(\overline{D}^{*0} K^-)$ Dalitz plot is constructed from efficiency corrected and combinatorial background subtracted data. In Figures 4.27 and 4.28 it is visible that the region where pentaquarks are expected to be observed is sparsely populated. The efficiency in that region is also lower (as deduced from MC and described in Section 4.11). On the other hand, a narrow resonance $D_{s1}^*(2536)$ is clearly visible in the correct place. It is even possible to see the angular structure - two maxima, characteristic of how a spin 1 resonance appears in a Dalitz plot. Another resonance which is observed is $D_{s1}^*(2700)$ - a wide resonance with $\Gamma \sim 122$ MeV. What is already obvious is that no clearly visible excess is observed where all or any of the investigated pentaquarks could be. The Dalitz projections for both hypotheses are provided in Figures 4.29 and 4.30.

It is important to note, that one of the constraints introduced with the Extended Cone Closure method before, was that of the mass of the $\Lambda_b^0$ candidate. This means that after the reconstructing full 4-momentum of the $\overline{D}^{*0}$ candidate, the invariant mass spectrum of $\Lambda_c^+ \overline{D}^{*0} K^-$ is not accessible. The spectra of the combinations of the decay daughters $\Lambda_c^+ \overline{D}^{*0}$, $\overline{D}^{*0} K^-$, and $\Lambda_c^+ K^-$ can be investigated separately. Dalitz plots can also be constructed for all three different combinations of the decay daughters.

**Figure 4.27** *Dalitz plot $m^2(\Lambda_c^+ \overline{D}^{*0})$ against $m^2(\overline{D}^{*0} K^-)$. Location of where pentaquarks would appear in the Dalitz plot are indicated by vertical lines. Two known resonances in $\overline{D}^{*0} K^-$ system are indicated by horizontal lines. Combinatorial background is subtracted by applying sPlot technique and efficiency is corrected using simulated sample of the decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$. The $\overline{D}^{*0}$ candidates here are reconstructed under $ECC_{\pi^0}$ hypothesis. The contour shows the kinematically allowed region for the decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$. It is visible that the candidates populate the correct region in the Dalitz plane.*

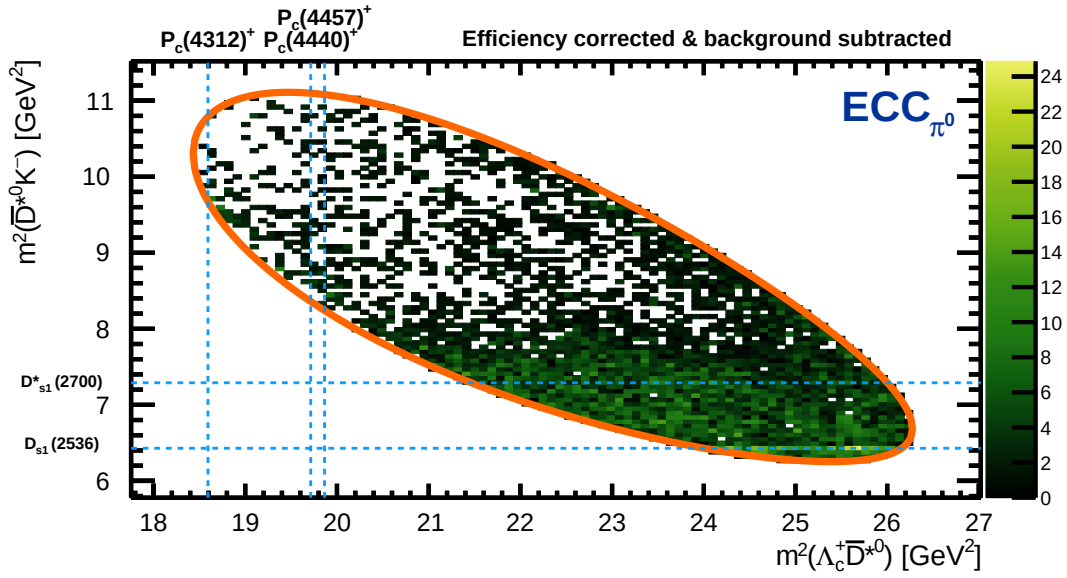


**Figure 4.28** *Dalitz plot $m^2(\Lambda_c^+\overline{D}^{*0})$ against $m^2(\overline{D}^{*0}K^-)$. Location of where pentaquarks would appear in the Dalitz plot are indicated by vertical lines. Two known resonances in $\overline{D}^{*0}K^-$ system are indicated by horizontal lines. Combinatorial background is subtracted by applying sPlot technique and efficiency is corrected using simulated sample of the decay $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}K^-$. The $\overline{D}^{*0}$ candidates here are reconstructed under $ECC_\gamma$ hypothesis. The contour shows the kinematically allowed region for the decay $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}K^-$. It is visible that the candidates populate the correct region in the Dalitz plane.*

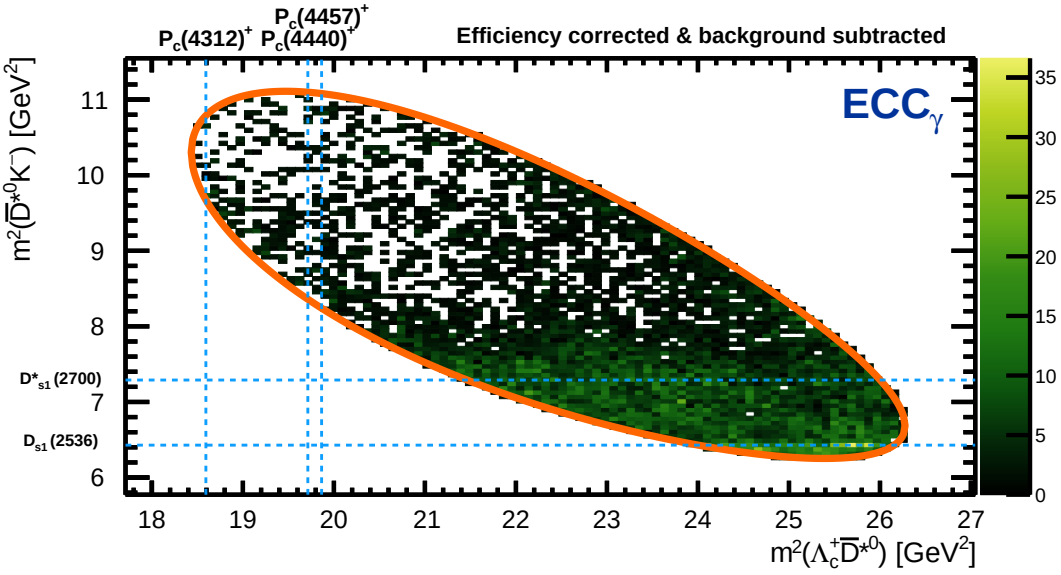


**Figure 4.29** *Dalitz projections where* $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ *candidates were reconstructed under* $ECC_{\pi^0}$ *hypothesis.*

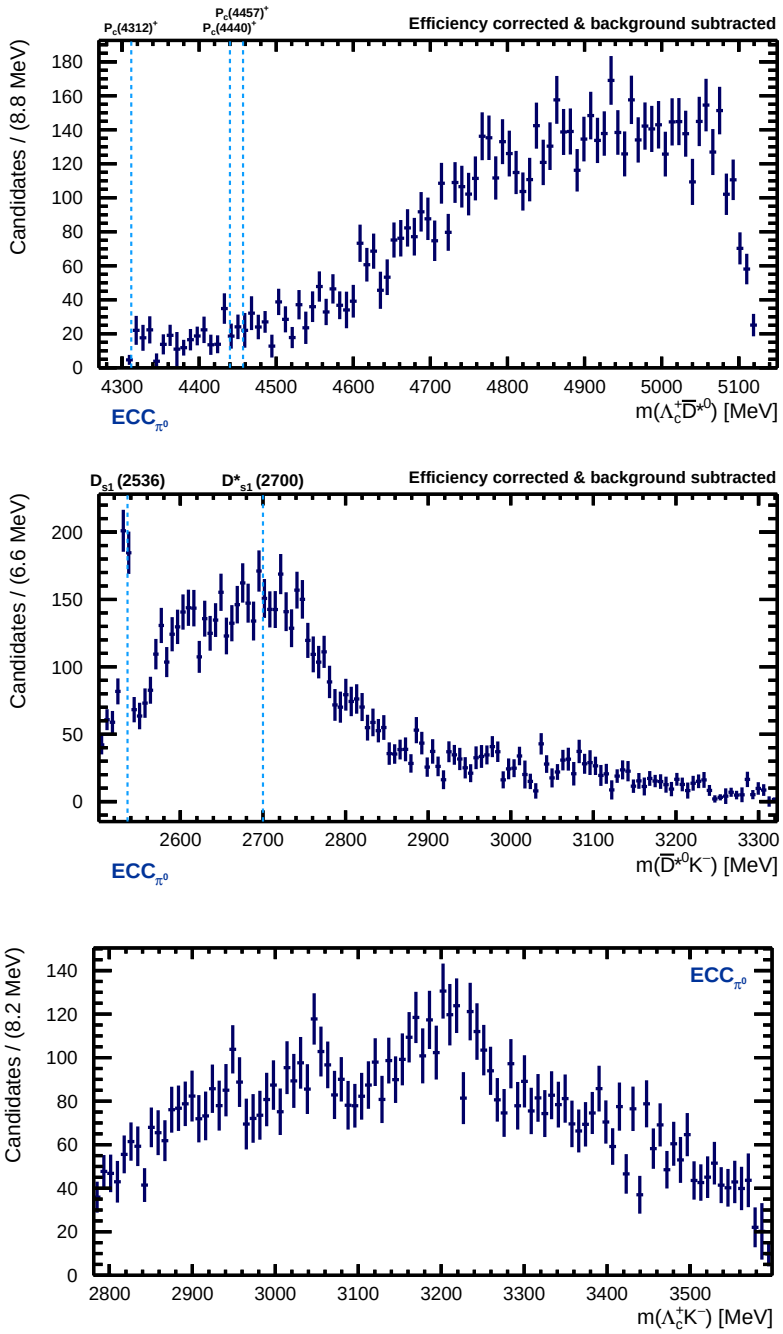
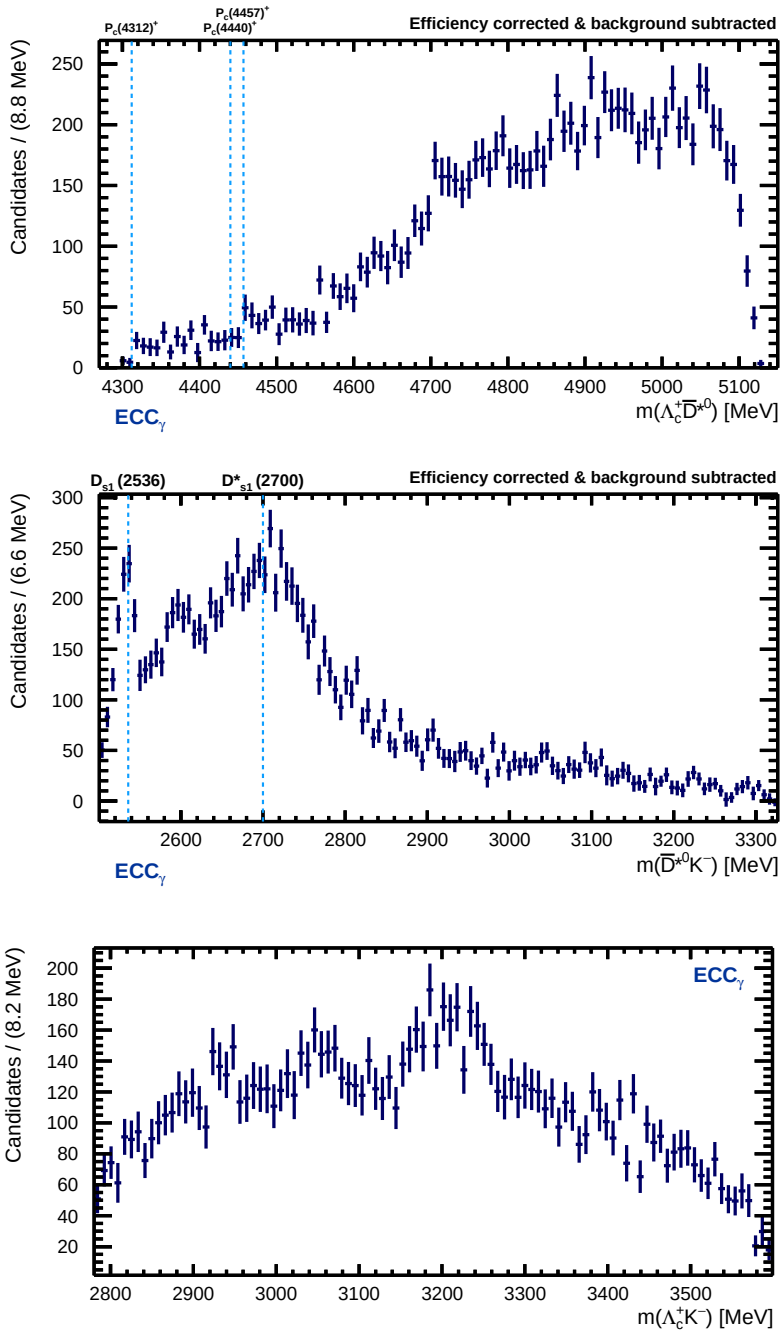


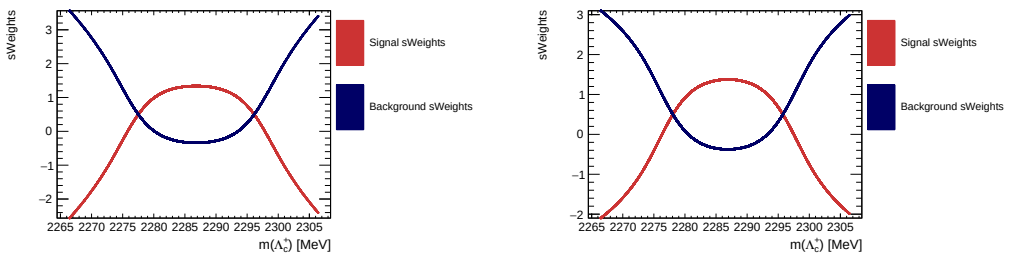
**Figure 4.30** *Dalitz projections where* $\Lambda_b^0 \rightarrow \Lambda_c^+ \overline{D}^{*0} K^-$ *candidates were reconstructed under* $ECC_\gamma$ *hypothesis.*



## 4.10 Background Subtraction

The combinatorial background is already strongly reduced because of the finely tuned selection. Some of the other methods of background subtraction (e.g. using the $\Lambda_b^0$ side-bands) are not available in this analysis, because m($\Lambda_b^0$) is one of the constraints for the Extended Cone Closure method (see Section 4.8). To further suppress the combinatorial background, the invariant mass spectrum of $\Lambda_c^+$ candidates is used, but contrary to the other uses throughout the analysis - without kinematic refit (DTF m($\Lambda_c^+$) constraint) (described in Section 4.6). The m($\overline{D}^0$) constraint is still applied. The component unfolding *sPlot* [112] technique is performed on the fit to data. The resulting distributions of signal and background weights are shown in Fig. 4.31 for both hypotheses. It is clear that in both cases the signal `sWeights` peak around the known mass of the $\Lambda_c^+$ baryon and the background `sWeights` are more pronounced in the side bands. This is an indication that the *sPlot* technique worked correctly. It is important to note, that after computing the `sWeights`, no further cuts should be applied to the data set. This is a purely statistical method, using the data sample itself to unfold two (or more) different components. As described in Section 3.3 of the Statistical Foundations Chapter 3. If the weights are computed and then the sample is further filtered, the resulting propability density functions become not normalized. For this reason, the *sPlot* is done after having already applied the ECC method on the sample. This also means that the `sWeights` are specific to the hypothesis chosen ($ECC_{\pi^0}$ or $ECC_\gamma$) and that the range in $\Lambda_c^+\overline{D}^0K^-$ invariant mass spectrum is limited to that of partially reconstructed $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}K^-$ decays (this is determined by the validity on ECC in a certain region as described in Section 4.8).

**Figure 4.31** *Distribution of signal and background* `sWeights` *graphed in m($\Lambda_c^+$) inv. mass. Signal* `sWeights` *peak around the known mass of a $\Lambda_c^+$ baryon and background* `sWeights` *are more pronounced in the side bands. This is an indication that sPlot technique worked correctly.*



The signal shape in the m($\Lambda_c^+$) spectrum is obtained from the MC, but in order to correct the MC, the `sWeights` have to be already applied to the data sample to use it as a target distribution as described in Section 4.5. To avoid such cyclic behavior, different MC samples are used for interim and final *sPlot* procedure. To obtain the MC shape which matches data with missing momentum best the $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ MC is used. It represents best the partially reconstructed signal observed in data. The correct fraction of $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ to $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$ component is preserved in this MC sample. The data sample with weights applied for combinatorial background subtraction is then used as a target distribution to kinematically re-weight the MC samples where a missing particle is known (either $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$ or $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$). The *sPlot* technique is performed again to obtain the final data sets with a specific hypothesis fixed. Here all the cuts are already applied and a specific MC sample with either only $\pi^0$ or $\gamma$ missing (isolated signal of either $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ or $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$) is used to obtain the shape of m($\Lambda_c^+$) signal. The MC samples in both cases are those with reconstructed $\overline{D}^{*0}$ momentum as well. The fit is performed using a Double Gaussian shape for the signal peak in m($\Lambda_c^+$) inv. mass spectrum and a linear background. This is shown in Fig. 4.33 for $ECC_{\pi^0}$ hypothesis and Fig. 4.35 for $ECC_\gamma$ hypothesis. The parameters of the Double Gaussian are fixed from a fit to the same spectrum on a sample of simulated events (depending on the ECC hypothesis - either MC for $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$ or that for $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$). This is shown in Figures 4.36 and 4.34 for the former and latter MC sample respectively, the ECC hypothesis applied is that of the decay in the MC sample ($ECC_{\pi^0}$ for $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ and $ECC_\gamma$ for $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$).

**Figure 4.32** *$\Lambda_c^+ \overline{D}^0 K^-$ invariant mass spectrum with and without subtracted background.*

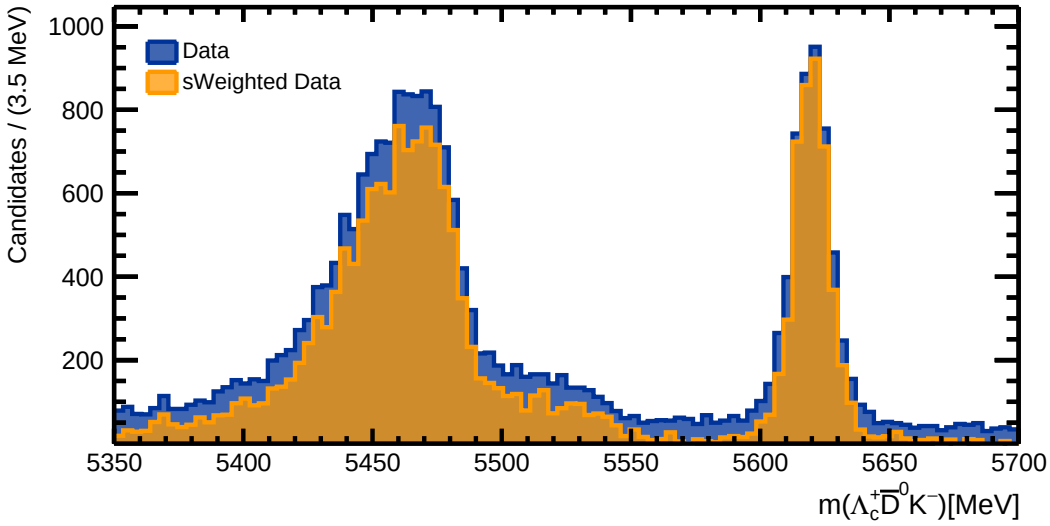



**Figure 4.33** *Fit to $m(\Lambda_c^+)$ inv. mass spectrum on data for $ECC_{\pi^0}$ hypothesis.*

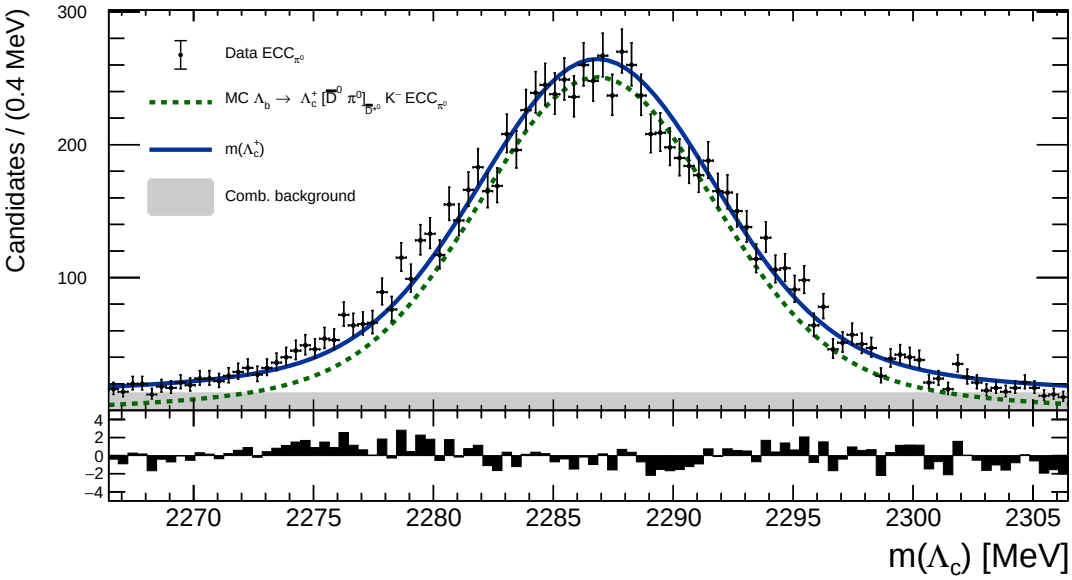


**Figure 4.34** *Fit to $m(\Lambda_c^+)$ inv. mass spectrum on $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ MC for $ECC_{\pi^0}$ hypothesis.*

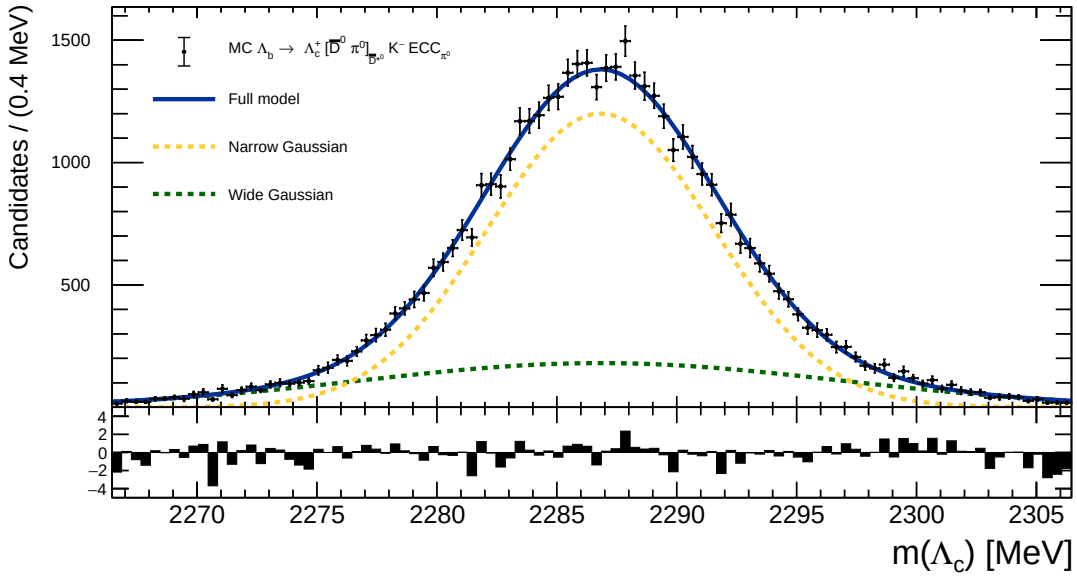



**Figure 4.35** *Fit to $m(\Lambda_c^+)$ inv. mass spectrum on data for $ECC_\gamma$ hypothesis.*

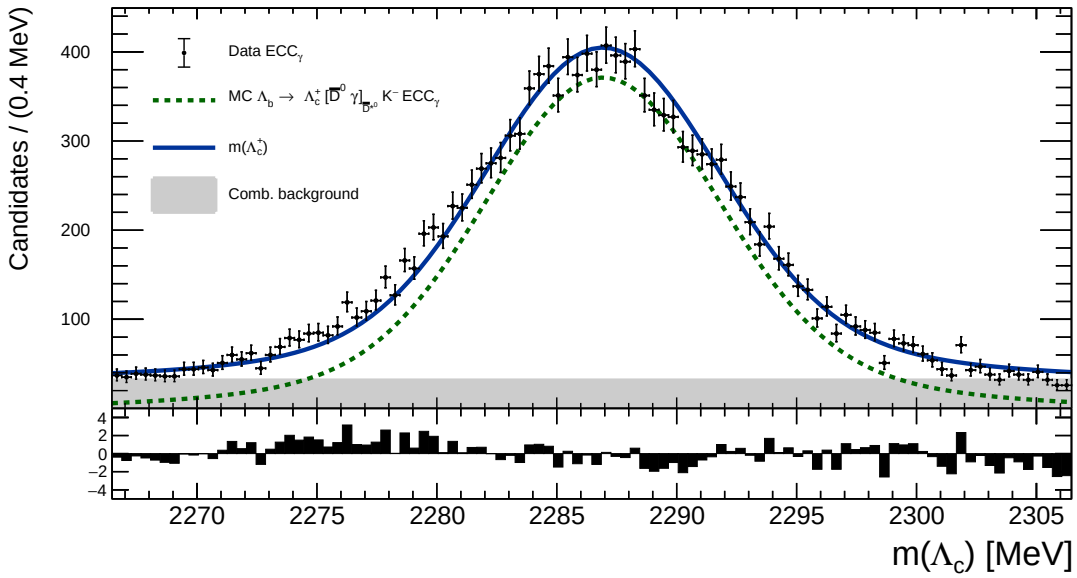
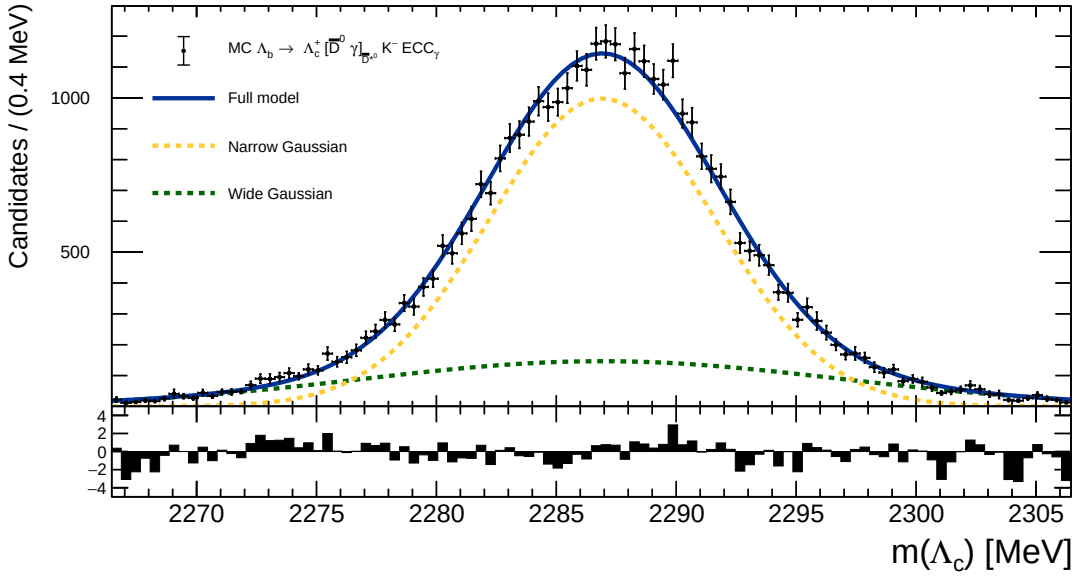


**Figure 4.36** *Fit to $m(\Lambda_c^+)$ inv. mass spectrum on $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$ MC for $ECC_\gamma$ hypothesis.*




# 4.11 Efficiencies

## 4.11.1 Efficiency Map

For this analysis, the efficiencies have to be treated in the context of the $m^2(\Lambda_c^+\overline{D}^{*0})$- $m^2(\overline{D}^{*0}K^-)$ Dalitz plot. The absolute value of the efficiency is not as important as investigating how the efficiencies change across the Dalitz plot. Efficiency map is acquired by taking the $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ phase space MC sample and applying the same offline selection and momentum reconstruction as that applied to the data. The MC here is truth-matched during the selection to make the signal acquisition as close to that of the data as possible. The truth matching involves selecting the appropriate background categories and particle IDs all available with the simulation. This is described in Section 4.4. The efficiency map is then constructed by binning the MC sample in $m^2(\Lambda_c^+\overline{D}^{*0})$ and $m^2(\overline{D}^{*0}K^-)$ and normalizing this Dalitz plot such that the average bin content in it corresponds to unity. Any values below or above 1 now mean that local efficiency in that region will be corrected by applying an efficiency weight. In the regions where the efficiency is higher than 1, the local population of candidates is reduced. This ensures that the efficiency correction acts uniformly on the Dalitz plot. Normalizing to the maximum value would scale every point up in the low

efficiency regions but keep the high efficiency regions unchanged.

To avoid the edge effects, having in mind that the binning can not be too fine since the data sample is limited ($\mathcal{O}(10^4)$ candidates), the efficiency map used in the analysis is actually the square efficiency map where instead of using $\mathrm{m}^2(\overline{D}^{*0}K^-)$, the cosine of the helicity angle of the $\Lambda_c^+\overline{D}^{*0}$ is used. The relevant helicity angle is sketched in Section 2.8. This square Dalitz plot can then be subdivided into a number of square bins and edge effects are avoided. The conventional Dalitz plot and the respective square Dalitz plot is shown in Fig. 4.38 for $ECC_{\pi^0}$ hypothesis and Fig. 4.39 for the $ECC_\gamma$ hypothesis. The efficiency correction on data is performed on the per-event basis. For every event in data a respective bin in the acquired efficiency map is selected and the bin height is saved to later be used in a weight for efficiency. N.B. Events have to be weighted by $\frac{1}{\epsilon}$.

### 4.11.2 Selection Efficiencies

To isolate and investigate the effects of the different selection requirements on the efficiency these requirements are established one by one and the square Dalitz plot of the decays $\Lambda_b^0 \to \Lambda_c^+\overline{D}^0K^-$ is displayed. Since these are the fully reconstructed decays a Dalitz plot can be constructed directly. The aim is to see if some of the cuts introduce sharp structures or shape the Dalitz plot in any way. In other words, if the efficiency of a specific cut is not flat over the Dalitz plot. It is visible from this study that isolated requirements act on the population of candidates in a uniform way.

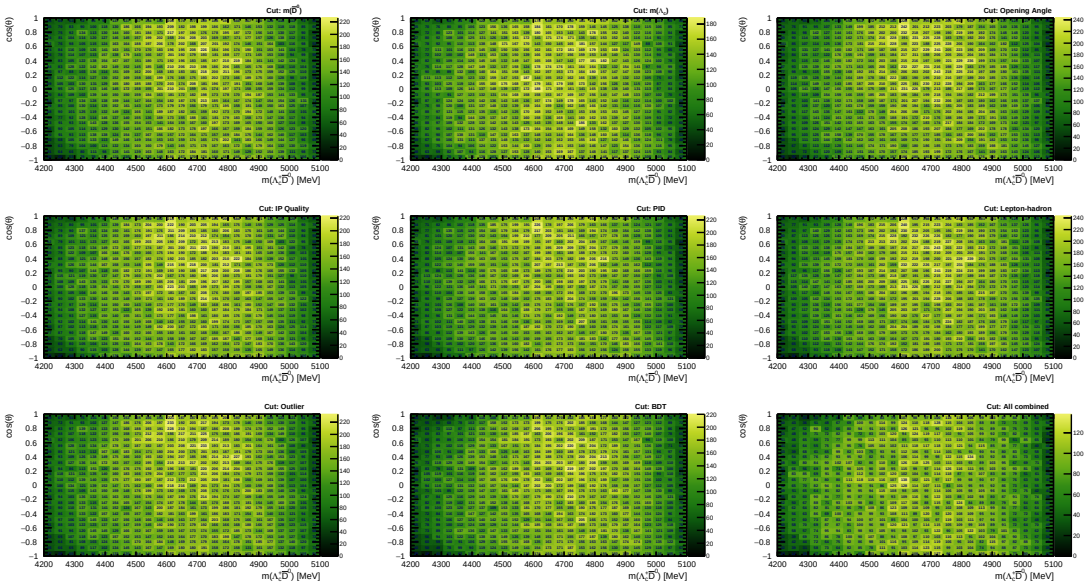



**Figure 4.37** *Effect of different cuts on the $m(\Lambda_c^+\overline{D}^0)$- $cos(\theta)$ Dalitz plot. It is visible that applying cuts one by one doesn't introduce any sharp structures to the square Dalitz plot, which is when constructed from a phase space MC sample has the information of reconstruction and selection efficiencies.*

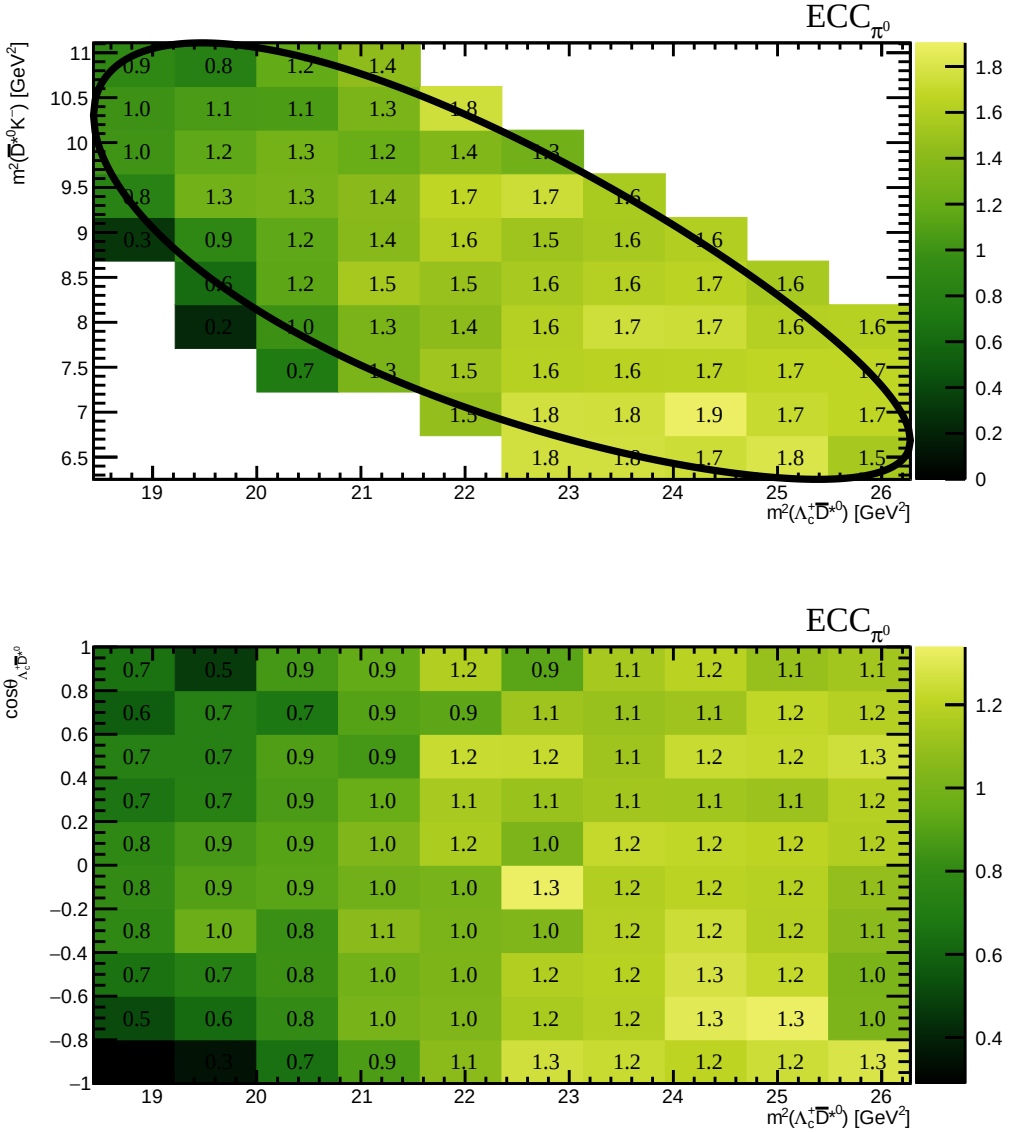



**Figure 4.38** *Efficiency maps for $ECC_{\pi^0}$ hypothesis.*

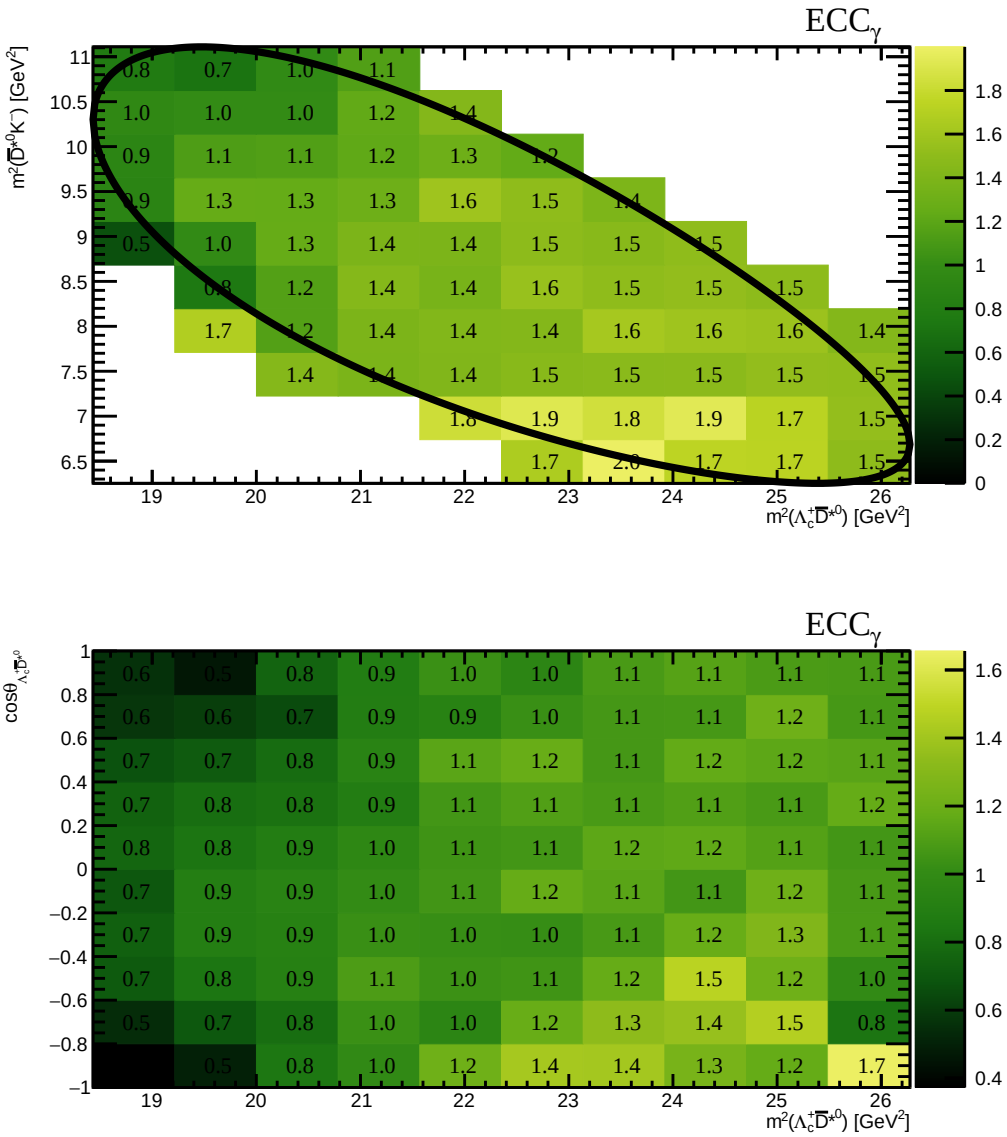



**Figure 4.39** *Efficiency maps for $ECC_\gamma$ hypothesis.*

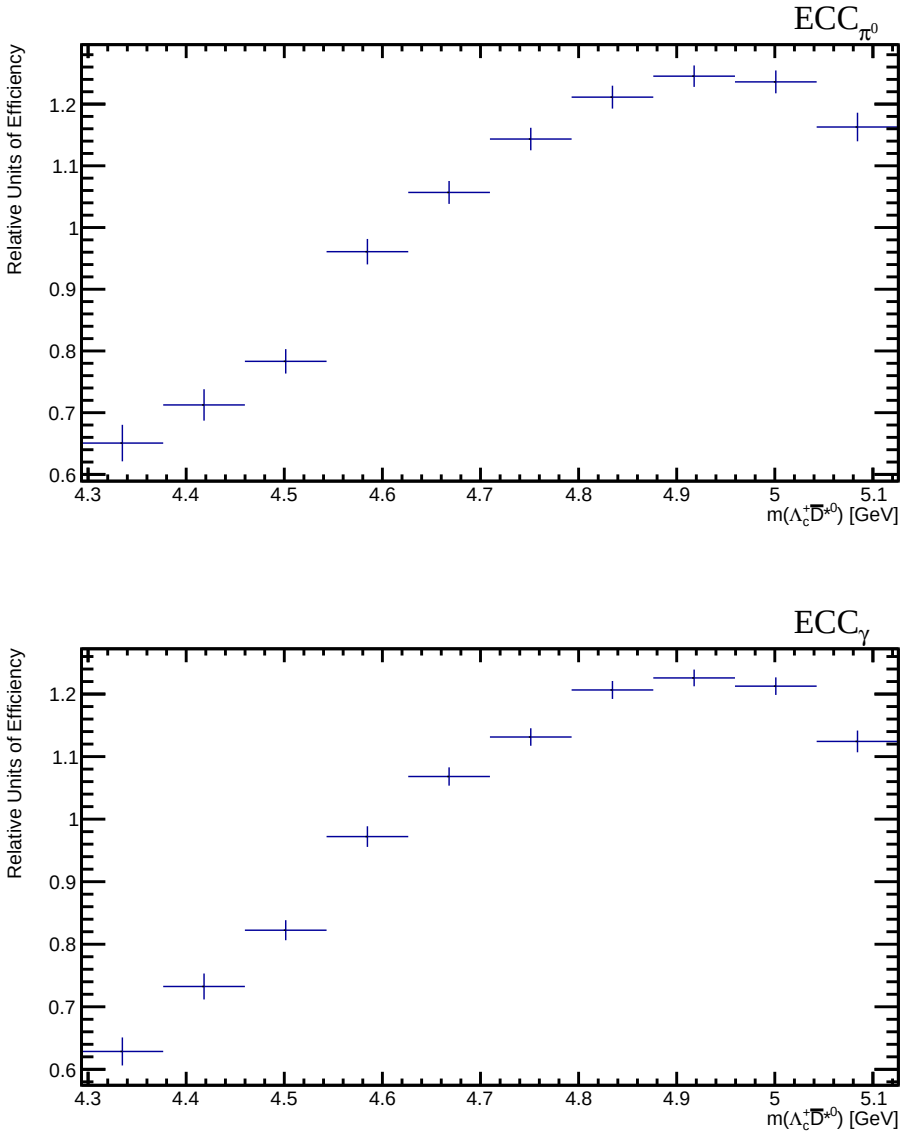



**Figure 4.40** *Efficiencies in $m(\Lambda_c^+ \overline{D}^{*0})$ for both $ECC_{\pi^0}$ and $ECC_\gamma$ hypothesis.*

## 4.12 Combined Effects of Reweighting the Data

Starting from the unfiltered data obtained by the LHCb detector, some systematic effects of the detector acceptance and varying sensitivity to particles of different momenta have to be corrected. For this purpose, the MC samples are used. They undergo the same selection procedure as the data samples. The simulation has the true information of the decay parameteres as well as parameters where the detector effects are taken into account. This is used to create an efficiency map and assign per-event weights to the data called `efficiency weights`. They are computed as $1/\epsilon$ where $\epsilon$ is the selection efficiency for the specific data point. The procedure is described in detail in Section 4.11. Another source of correction (a column of weights) is the `sWeights` described in Section 4.10. These weights are obtained by unfolding the signal and background components and deducing the most signal-like or background-like candidates. Under the assumption that the efficiency and `sWeights` are not correlated, the combined set of corrections for data samples then becomes:

$$\text{Correction Weights} = \text{EffWeights} \cdot \text{sWeights}$$

Effects of the background subtraction (applying the `sWeights`) and the efficiency correction (applying the efficiency weights) are illustrated in Figures 4.41, 4.42 for the m($\overline{D}^{*0}K^-$) invariant mass spectrum and Figures 4.43 and 4.44 for the m($\Lambda_c^+\overline{D}^{*0}$) invariant mass spectrum. The combined effect in both the m($\Lambda_c^+\overline{D}^{*0}$) and the m($\overline{D}^{*0}K^-$) invariant mass spectra are mainly, the reduced combinatorial background and the change of apparent local density of candidates. The removed contribution from the combinatorial background is indicated by the red distribution in the figures. The efficiency correction reduces the apparent population of the events in the regions where the density of candidates is high. In the regions where the density is low, the apparent population of the candidates is increased at the expense of a larger statistical error. It is important to remember that applying the efficiency correction weights does not really increase the number of events in the data sample, but acts as a tradeoff between the apparent number of events and statistical error. The efficiency corrected spectrum is more representative of the physical spectrum obtained by the detector.

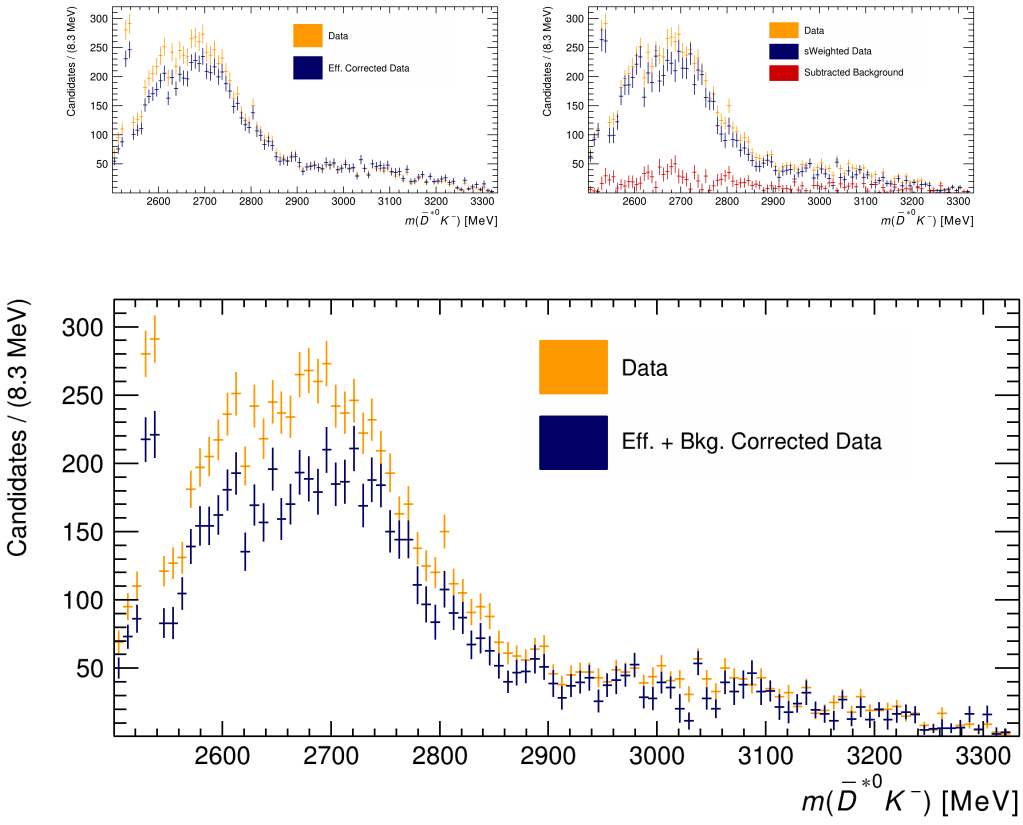



**Figure 4.41** *Effect of applying sWeights and EffWeights on the* $m(\overline{D}^{*0}K^-)$ *spectrum for* $ECC_{\pi^0}$ *hypothesis. The above plots show effect of applying sWeights and efficiency weights separately. The bottom plot shows the combined effect of both weights used to correct the data sample.*

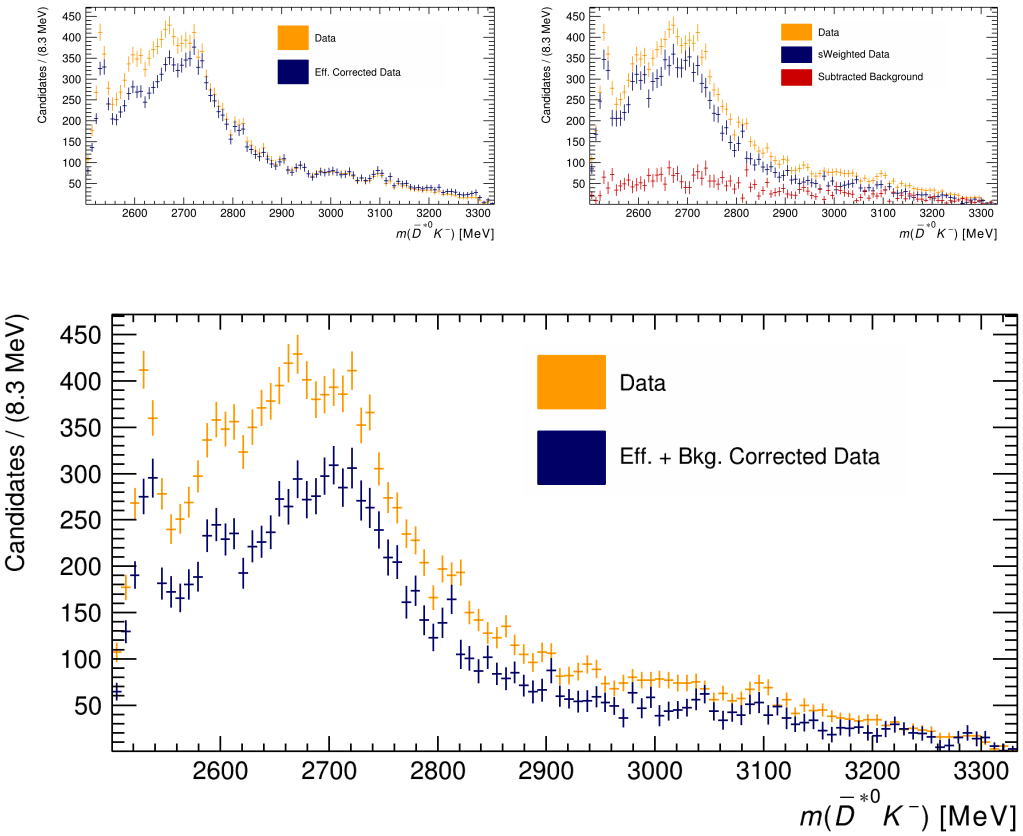



**Figure 4.42** *Effect of applying sWeights and EffWeights on the $m(\overline{D}^{*0}K^-)$ spectrum for $ECC_\gamma$ hypothesis. The above plots show effect of applying sWeights and efficiency weights separately. The bottom plot shows the combined effect of both weights used to correct the data sample.*

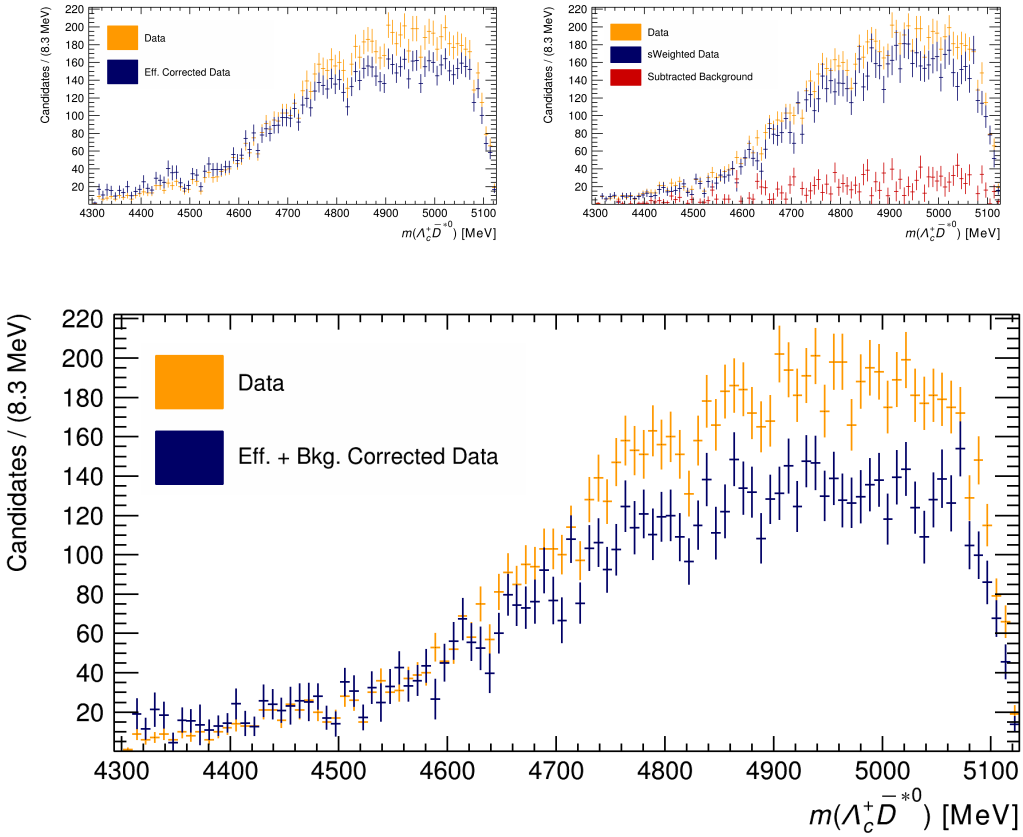
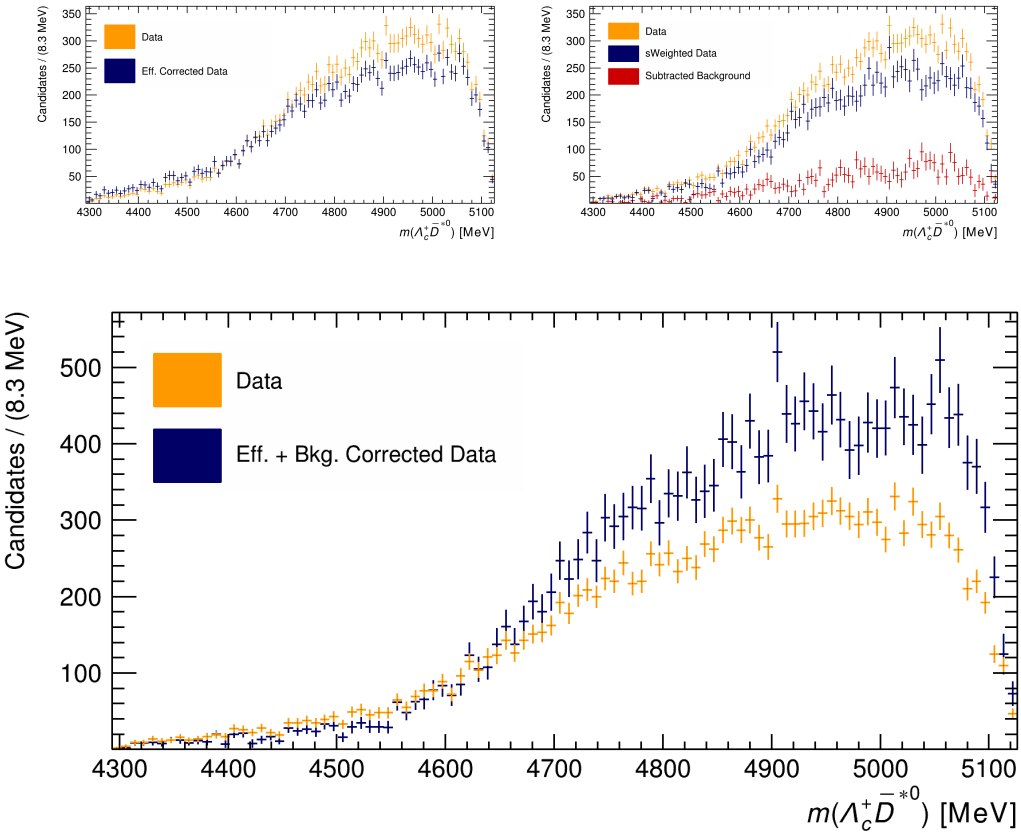



**Figure 4.43** *Effect of applying sWeights and EffWeights on the $\Lambda_c^+\overline{D}^{*0}$ spectrum for $ECC_{\pi^0}$ hypothesis. The above plots show effect of applying sWeights and efficiency weights separately. The bottom plot shows the combined effect of both weights used to correct the data sample.*

**Figure 4.44** *Effect of applying sWeights and EffWeights on the $\Lambda_c^+ \overline{D}^{*0}$ spectrum for $ECC_\gamma$ hypothesis. The above plots show effect of applying sWeights and efficiency weights separately. The bottom plot shows the combined effect of both weights used to correct the data sample.*



## 4.13 Resolution in the m($\Lambda_c^+ \overline{D}^{*0}$) nvarian mass spectrum

Detector response (or the resolution) for a specific parameter is obtained using the MC samples for the decays $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ and $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$. The resolutions in this analysis are computed for the m($\Lambda_c^+ \overline{D}^{*0}$) invariant mass spectrum. The values of the reconstructed variable are compared to true values at different points, when plotted against the variable of interest. The standard deviation in this computed parameter is taken as the combined resolution of the detector and ECC method. The higher the difference between reconstructed and true variables, the worse the resolution. Studies on the intrinsic resolution of the ECC method were also performed. They are described in Section 4.8. In this analysis the resolution is obtained from $\Lambda_c^+ \overline{D}^{*0}$ inv. mass spectrum in the MC reconstructed via the ECC method and `MCDecayTreeTuple` true branches ($\Lambda_b^0 - \Lambda_c^+ - K^-$). The local point spread function is obtained in the region of m($\Lambda_c^+ \overline{D}^{*0}$) where one of the three studied pentaquarks ($P_c(4312)^+$, $P_c(4440)^+$ and $P_c(4457)^+$) are expected to be observed. This is indicated in Fig. 4.45.

The resolution is evaluated for both $ECC_{\pi^0}$ and $ECC_\gamma$ hypotheses on the $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ (in the nominal case). This way the signal model is constructed with both, the correct ($ECC_{\pi^0}$) (see Fig. 4.46) and incorrect ($ECC_\gamma$) (see Fig. 4.47) hypotheses included. It is worth noting, that when the incorrect hypothesis is applied (in this case, the $ECC_\gamma$), a bias is observed in the point spread function. This is expected when reconstructing the missing momentum with a wrong mass assumption. This distribution is well described by a Double Gaussian function. A "cocktail" of the two obtained shapes is produced where the fraction of $\gamma$ component to the $\pi^0$ component is that from the fit to the $\Lambda_b^0$ invariant mass spectrum in the region of interest. This is described in Section 4.7. Since the fraction of $\gamma$ events is $\sim 23\%$, the combined point spread function is a mixture of $\pi^0$ and $\gamma$ components with the ratio of them fixed. These point spread functions for the three studied pentaquarks are shown in Fig. 4.48. The nominal signal model is chosen to be the Breit-Wigner shape, convoluted with the obtained point spread function. This way, the resolution effects are taken into account.

**Figure 4.45** *Regions where the point spread functions are evaluated to obtain the local resolution.*

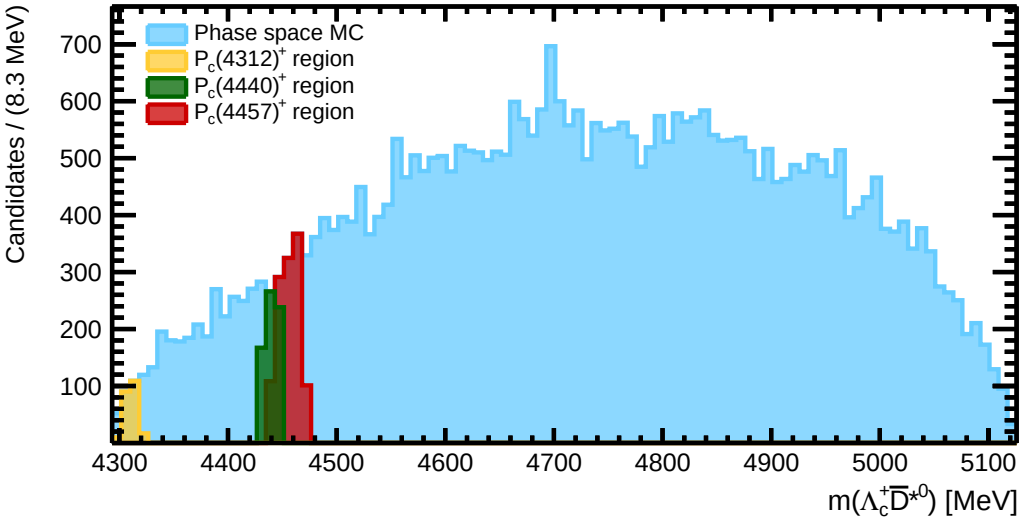



**Figure 4.46** *Point spread functions of $P_c(4312)^+$, $P_c(4440)^+$ and $P_c(4457)^+$ $ECC_{\pi^0}$ MC $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$.*

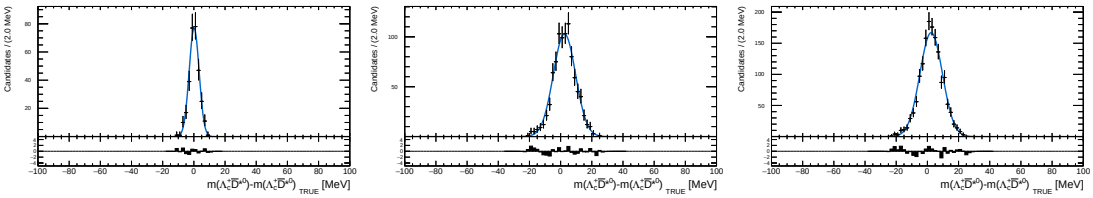


**Figure 4.47** *Point spread functions of $P_c(4312)^+$, $P_c(4440)^+$ and $P_c(4457)^+$ $ECC_{\gamma}$ MC $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$.*

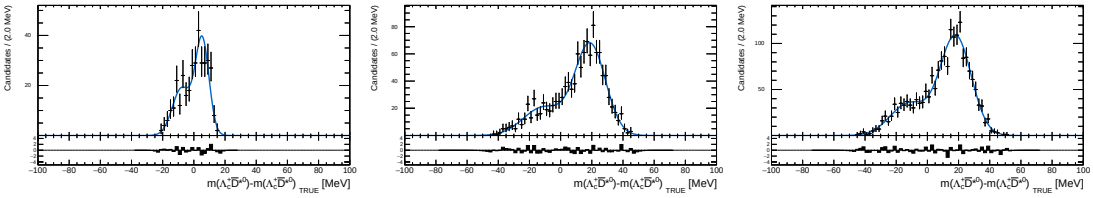


**Figure 4.48** *Combined point spread functions of $P_c(4312)^+$, $P_c(4440)^+$ and $P_c(4457)^+$ $ECC_\gamma$ MC $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$.*

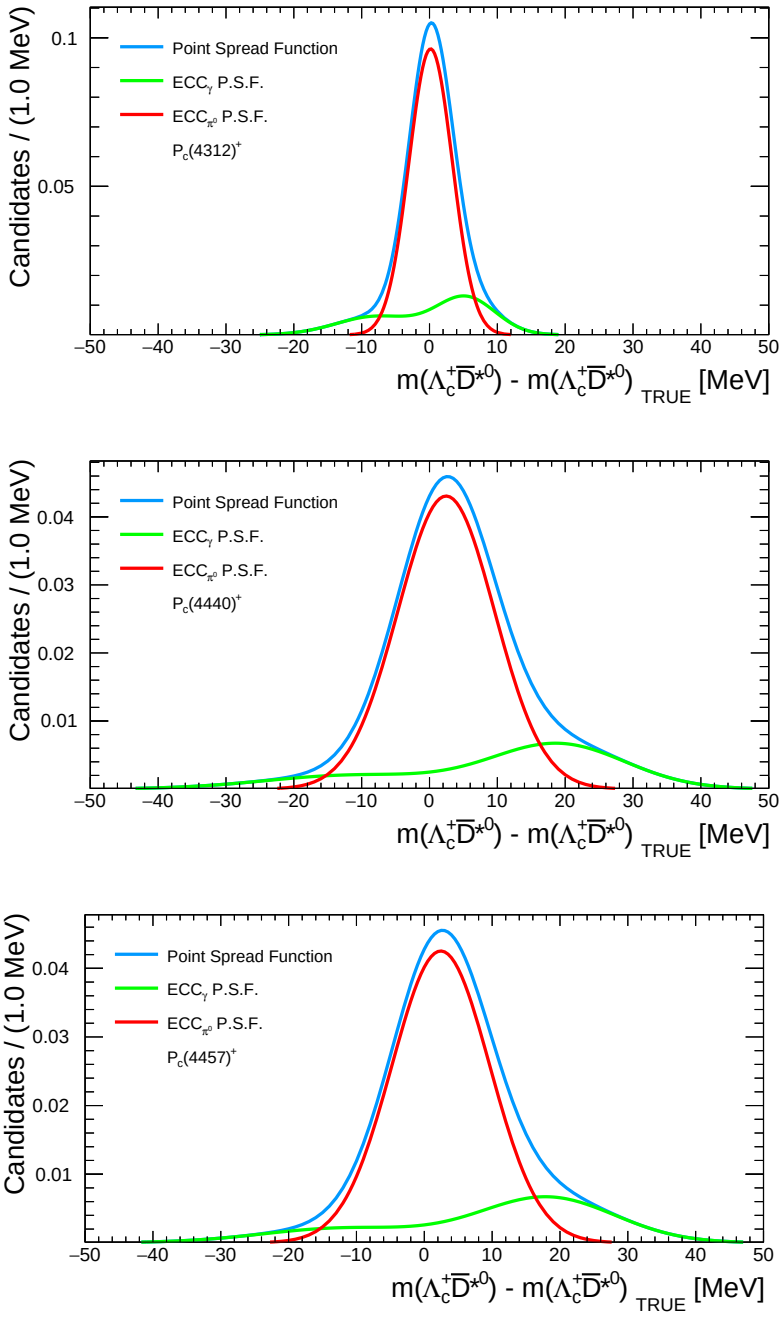
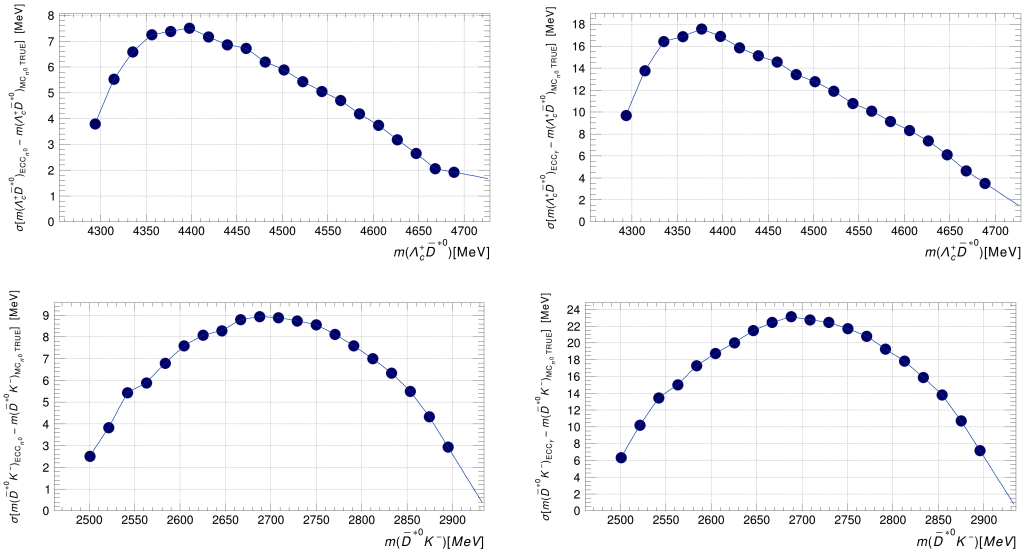


**Figure 4.49** *Resolution in $m(\Lambda_c^+\overline{D}^{*0})$ and $m(\overline{D}^{*0}K^-)$ for $ECC_{\pi^0}$ and $ECC_\gamma$ hypotheses.*



## 4.13.1 Resolution Scans in $\mathbf{m}(\Lambda_c^+\overline{D}^{*0})$ and $\mathbf{m}(\overline{D}^{*0}K^-)$ Spectra

Treating resolutions as Gaussian, scans through the $\mathrm{m}(\Lambda_c^+\overline{D}^{*0})$ and $\mathrm{m}(\overline{D}^{*0}K^-)$ spectra are performed to observe how the resolution varies across the phase space. The resolution in this context is defined as the standard deviation (width of a Gaussian) at a specific point (a narrow region) in the parameter of interest. This is extracted using `StdDev` method provided within `ROOT Data Frame` library. The resolution dependence on $\mathrm{m}(\Lambda_c^+\overline{D}^{*0})$ and $\mathrm{m}(\overline{D}^{*0}K^-)$ variables is shown in Fig. 4.49. It is visible that for the $P_c(4312)^+$, the resolution is better because of the vicinity of this supposed state to the edge of a Dalitz plot where a spread in momenta of the particles is limited by the decay kinematics. For the states $P_c(4440)^+$ and $P_c(4457)^+$ resolution is about 10 MeV for $ECC_{\pi^0}$ hypothesis. As expected and visible in the Dalitz projections, resolution in the $ECC_\gamma$ case is worse by about a factor of 2.

## 4.14 Spectral Analysis Using Legendre Moments

A model independent approach taken in this analysis involves the Legendre moments, described in Section 2.8 of the Theory Chapter 2. A representation of the Dalitz plot is constructed directly from the data sampe. An assumption is made that all of the resonances in the decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ appear in one of the 2-particle systems. In this analysis - the resonances are assumed to only appear in the $\overline{D}^{*0} K^-$ system. The helicity angle is used to obtain the angular distributions for each bin in the m($\overline{D}^{*0} K^-$) invariant mass spectrum. Using the m($\overline{D}^{*0} K^-$) and the helicity angle as inputs, the Legendre moments are computed. These moments are the coefficients of the Legendre series expansion. A desired maximum order $k_{max}$ is established and the series is truncated at this order. The Legendre weights are defined as the sum of the Legendre moments up $k_{max}$. A phase space MC sample is taken and these weights are applied on the per-event basis. A representation of the data is obtained by examining the m($\overline{D}^{*0} K^-$) invariant mass spectrum. The statistical fluctuations in this spectrum are filtered out because of the truncation of the Legendre series.

With the Legendre weights applied, the m($\Lambda_c^+ \overline{D}^{*0}$) invariant mass spectrum is investigated. At this point, any large deviations in the data from the projected distribution would already indicate the existence of an exotic component or at least a large statistical effect in one of the two remaining projections - the $\Lambda_c^+ \overline{D}^{*0}$ or the $\Lambda_c^+ K^-$. This way, without performing a full amplitude analysis it is possible to know whether the known resonances are enough to describe the data. If there were an indication of the existence of some exotic states with a large deviation from the null hypothesis, a full amplitude analysis would need to be performed with additional exotic states introduced in the amplitude model. The $\overline{D}^{*0} K^-$ variable is shown in Fig. 4.50. The binning is chosen such that the angular distribution in each interval of the m$^2$($\overline{D}^{*0} K^-$) variable is still rich enough in statistics. This way, the moments describe the angular distribution more accurately in a specific region of the invariant mass spectrum.

The projections on the $\Lambda_c^+ \overline{D}^{*0}$ invariant mass spectrum of the truncated series added up to the order $k_{max} = 4$, $k_{max} = 9$ and $k_{max} = 100$ are shown in the Fig. 4.51. Chosing the maximum order is motivated by the fact that adding the Legendre moments up to the order $k_{max}$ is enough to describe the resonances

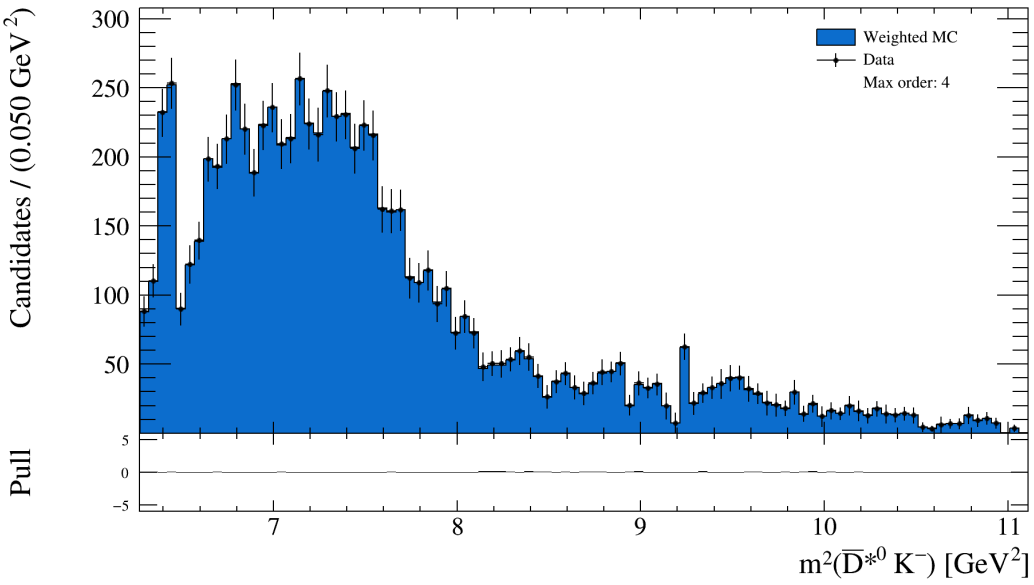



**Figure 4.50** *Phase space MC re-weighted to match the data in the $m^2(\overline{D}^{*0}K^-)$ spectrum.*

with the orbital angular momentum $l = k_{max}/2$. The angular structure of the resonances in the $\overline{D}^{*0}K^-$ system should be well reproduced by adding the first 4 moments if we assume that there are no contributions from resonances with spin larger than 1. The order can be chosen to be arbitrarily high. In this analysis the series of orders $k_{max} = 4$, $k_{max} = 9$, $k_{max} = 100$ and others were investigated. Taking the series to such a high order means that rather than correctly reproducing the angular distributions of the resonances in the $\overline{D}^{*0}K^-$ system, any statistical fluctuations in the angle are captured. This defeats the purpose of trying to project the obtained shapes to a different axis. It is visible that by chosing the order 4, the data is described well already. Fig. 4.51 also shows the angular disctributions for the three cases of $k_{max}$. The same interval in the $\Lambda_c^+\overline{D}^{*0}$ invariant mass is taken to show the difference of parametrizing the angular distribution with different orders of moments. It is seen that while $k_{max} = 4$ already matches the angular distribution well, at $k_{max} = 9$ some effects similar to overfitting can be seen towards the edges of the distribution. At $k_{max} = 100$ all the statistical fluctuations are matched. High frequency pattern can be seen in the angular distribution. The fact that m$(\Lambda_c^+\overline{D}^{*0})$ invariant mass spectrum can be well described by using a low order Legendre series expansion in $\overline{D}^{*0}K^-$ system is a strong indication that no $P_c^+$ resonances are present in the given data set. For this reason, a limit setting procedure is performed to obtain the limit on the fit fraction of $P_c^+$ contribution to m$(\Lambda_c^+\overline{D}^{*0})$ ivariant mass spectrum. This is described in Section 4.15.

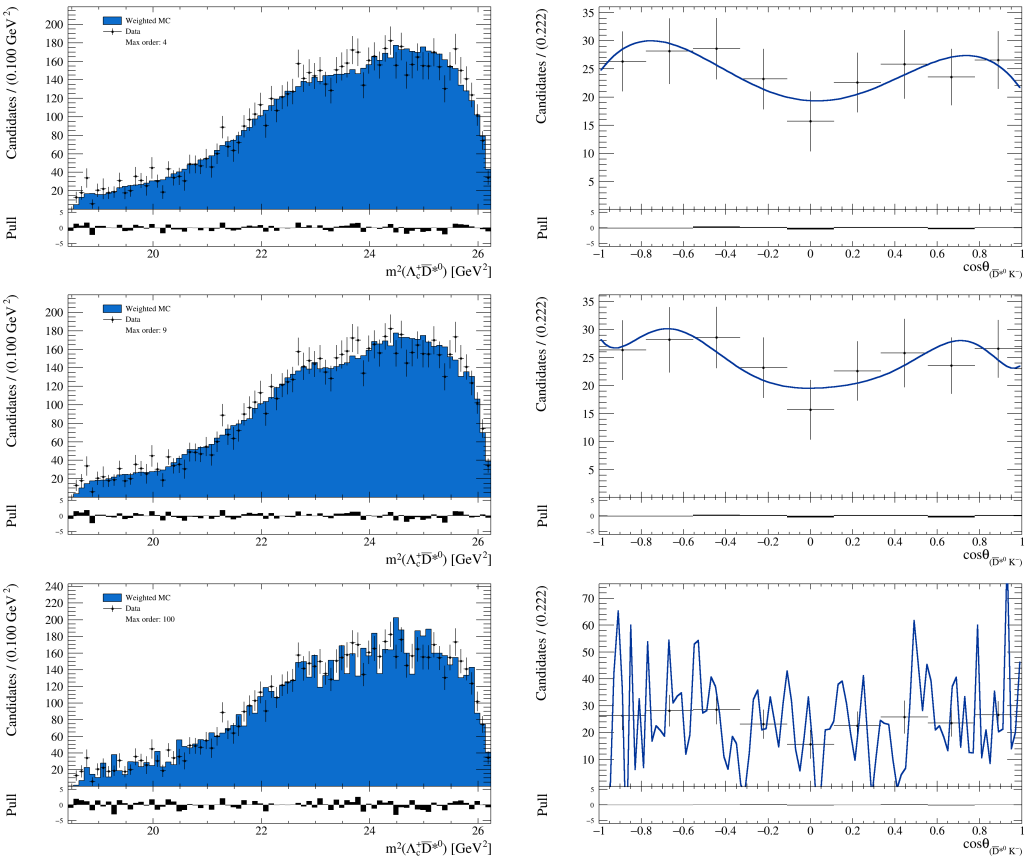



**Figure 4.51** *Projections of the reweighted MC (right) and the angular distribution in a narrow region of the $m(\overline{D}^{*0}K^-)$ (left) with the Legendre series truncated at the order $k_{max} = 4$, $k_{max} = 9$, $k_{max} = 100$.*

It is important to note that this approach is using mild assumptions about the maximum spin of the possible exotic states present. Also, the decaying particle ($\Lambda_b^0$ in this case) is assumed to be unpolarized. The method does not require any constraints or prior knowledge on the number of the supposed exotic states, their line shapes, masses or widths. If the states are close to each other or to a mass threshold of a 2-particle system, their shapes might be difficult to predict due to the interference effects. Yet again, with this model independent approach, any deviation in the data due to a resonance or even a non-resonant contribution being present would be possible to observe. On the other hand, while this model independent approach is valid for probing the phase space for exotic resonances and, to some extent, can even be used to pinpoint their location in the m($\Lambda_c^+ \overline{D}^{*0}$) spectrum, it can not be used to make a determination of the properties of the said resonances. Quantum numbers, width and other parameters need to be obtained through an amplitude analysis. Since these supposed exotic states can interfere with each other and the conventional resonances present in the decay, their shapes can be different or they might be shifted in mass from the actual value.

Another important notion when using this method is the fact that if a deviation from the obtained representation is observed, there is no way to be sure that the exotic resonances are only present in the projection of interest ($\Lambda_c^+ \overline{D}^{*0}$). Deviations from data could be caused by a resonance (or multiple resonances) in the other projection (m($\Lambda_c^+ K^-$)). Indeed, the resonances in $\Lambda_c^+ K^-$ system are present. They are $\Xi_c$ (2923), $\Xi_c$ (2930), $\Xi_c$ (2970). An indication of where these resonances could appear is provided in Fig. 4.52.

It can be seen, that the contribution from $\Xi_c$ resonances is very small. In fact, the spectrum is dominated by the reflection of the $D_{s1}(2536)$ and the $D_{s1}(2700)$ resonances in the $\overline{D}^{*0} K^-$ system. Moreover, the $\Xi_c$ resonances appear far away from the region where $P_c^+$ states are expected to be. In the m$^2$($\Lambda_c^+ \overline{D}^{*0}$)-m$^2$($\overline{D}^{*0} K^-$) Dalitz plot they appear on the diagonal but are difficult to observe, since the dalitz plot is dominated by the contributions from the $\overline{D}^{*0} K^-$ system. An assumption is made that no interference or feed down is present from the $\Xi_c$ resonances in the region where any one of the three pentaquarks being studied might appear. For completeness, the truncated Legendre series is also projected on the $\Lambda_c^+ K^-$ invariant mass spectrum at different orders. This is illustrated in Fig. 4.53. It is visible here that at low orders ($k_{max} = 4$ and $k_{max} = 9$) the

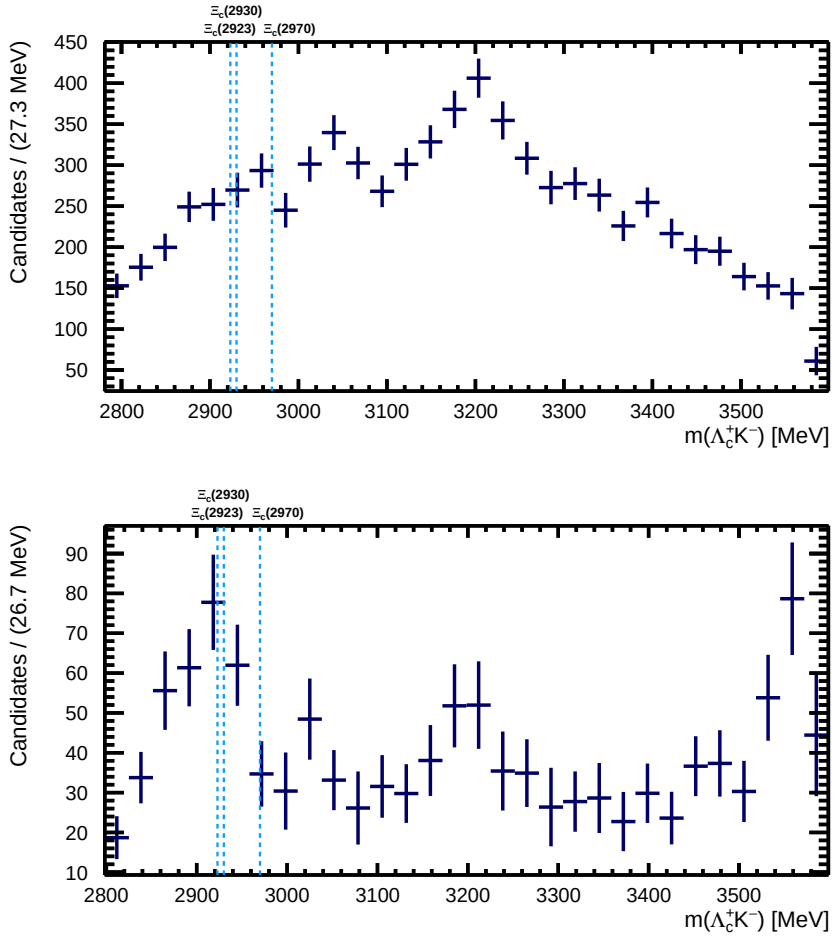



**Figure 4.52** *$\Lambda_c^+ K^-$ invariant mass spectrum. On the left, no cuts are applied and the sizable reflection from the $D_{s1}(2536)$ resonance is visible as a two-peaked structure in the centre. On the right, a cut is applied where $m(\overline{D}^{*0}K^-) > 2910$ MeV. This makes sure that both $D_{s1}(2536)$ and $D_{s1}(2700)$ resonances in the $\overline{D}^{*0}K^-$ system are strongly suppressed. Contribution from the $\Xi_c$ resonances is very small.*

projection does not match the data. This is mainly because of a large reflection from the $\overline{D}^{*0}K^-$ system but also the presence of the $\Xi_c$ resonances. Two other Dalitz plots, namely $\Lambda_c^+\overline{D}^{*0}$- $\Lambda_c^+K^-$ and $\overline{D}^{*0}K^-$- $\Lambda_c^+K^-$ are provided in Fig. 4.54 and Fig. 4.55, respectively.

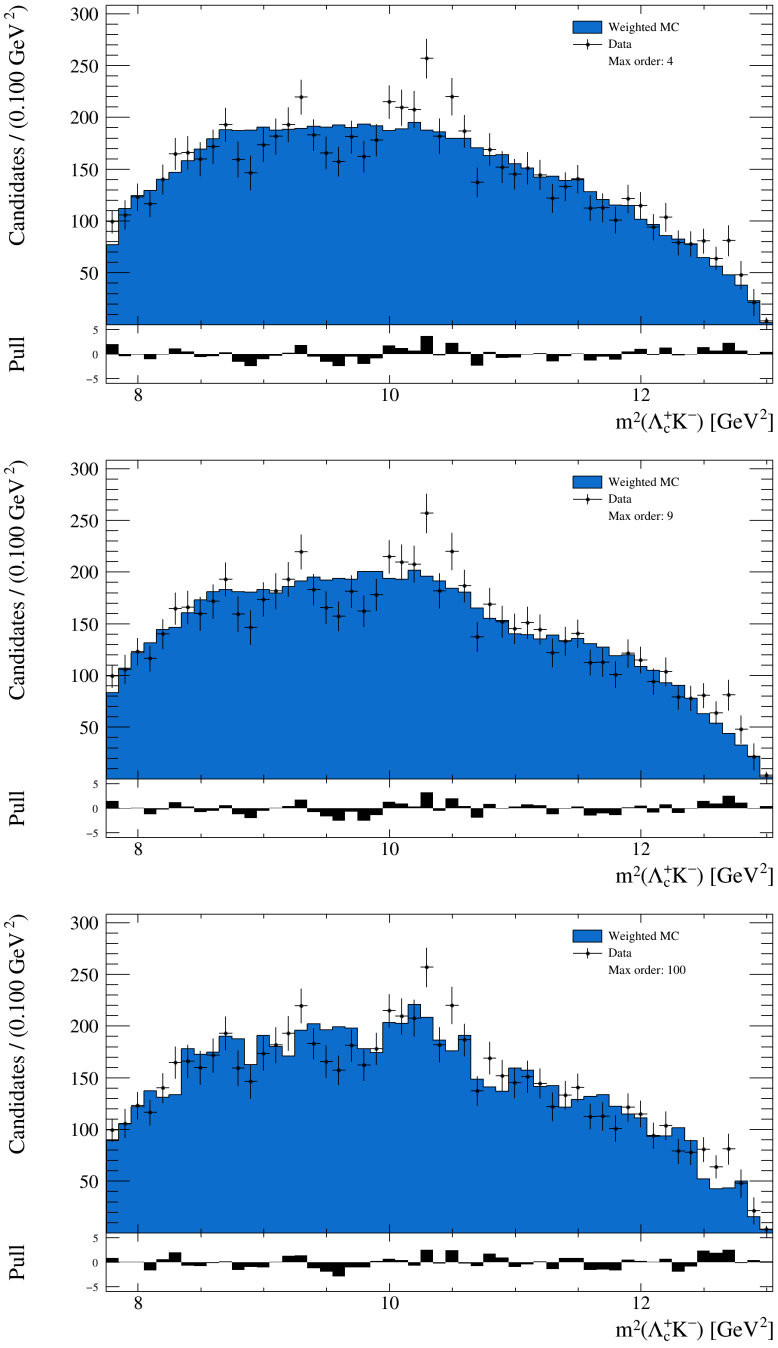



**Figure 4.53** *Projection of reweighted MC on the $m(\Lambda_c^+ K^-)$ invariant mass spectrum with the Legendre series truncated at the order $k_{max} = 4$, $k_{max} = 9$, $k_{max} = 100$.*

**Figure 4.54** *Dalitz plot $m^2(\Lambda_c^+ \overline{D}^{*0})$ against $m^2(\Lambda_c^+ K^-)$. $ECC_{\pi^0}$ hypothesis.*

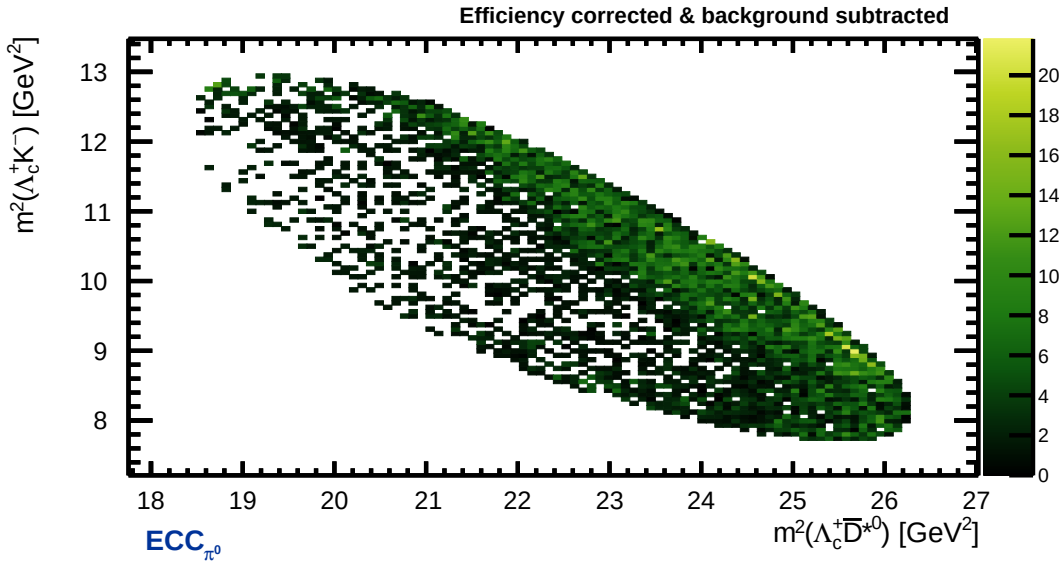
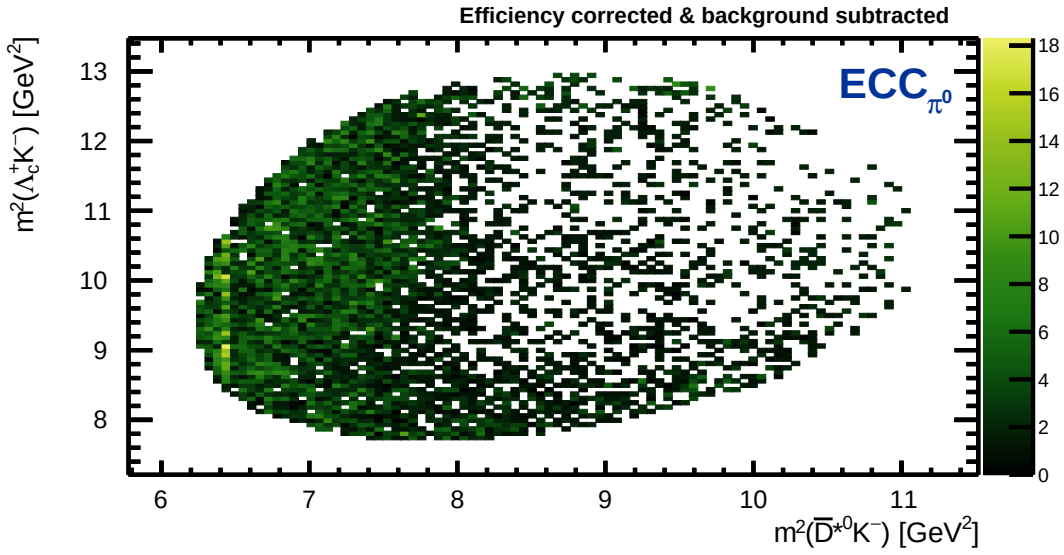



**Figure 4.55** *Dalitz plot $m^2(\overline{D}^{*0} K^-)$ against $m^2(\Lambda_c^+ K^-)$. $ECC_{\pi^0}$ hypothesis.*



## 4.15 CLs Limit Setting

This section describes the procedure of setting the CLs upper limits [111]. The statistical description of this method is given in the Section 3.2.3 of the Statistical Background Chapter. The limits are set on the fit fractions of the pentaquark signal in the $\Lambda_c^+\overline{D}^{*0}$ invariant mass spectrum. In this procedure, a number of fits is performed on the generated pseudo-experiment data samples (toy samples). The values of the parameter of interest (in this case, the fit fraction for a specific pentaquark state) are scanned up to a desired value. The CLs result (the modified p-value) is plotted against the parameter of interest. The upper limit is then defined as the value of the parameter of interest which would be established with a desired confidence level. In this case, the confidence level is chosen to be 95%. The reported result is that for the $ECC_{\pi^0}$ hypothesis - decay $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ since it is the dominant component in the data sample.

Choosing $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ also means that there is a smaller fraction of contamination from the other decay channel with $\overline{D}^{*0}$ (namely the decay $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\gamma)K^-$) as well as the decay $\Lambda_b^0 \to \Sigma_c^+(\to \Lambda_c^+\pi^0)\overline{D}^0K^-$ contribution. As discussed in Sections 4.8 and 4.7, the first source of contamination smears out $\sim 20\%$ of events symmetrically in the Dalitz plot $\mathrm{m}^2(\Lambda_c^+\overline{D}^{*0})$- $\mathrm{m}^2(\overline{D}^{*0}K^-)$. This manifests as a resolution effect and is small on the upper limit value. During this analysis, studies on setting the upper limit with the $ECC_\gamma$ hypothesis showed that at the same confidence level the limit is generally higher than that for the $ECC_{\pi^0}$ hypothesis. It is important to note that in the case of $ECC_\gamma$ 65% of the events are reconstructed at worse resolution. This is discussed in Section 4.8.6. The $\Lambda_b^0 \to \Sigma_c^+(\to \Lambda_c^+\pi^0)\overline{D}^0K^-$ component acts as a type of background of uniformly distributed events in the Dalitz plot. The main focus on this component was to check if it does not introduce any sharp structures into the Dalitz plot when reconstructed with ECC, which could be mistaken for $P_c^+$ resonances. It was determined that even though this component can not be disentangled from the dominant component, the $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$, it does not introduce any sharp structures or impede the limit setting procedure.

### 4.15.1 Fit Model to m($\Lambda_c^+ \overline{D}^{*0}$) Invariant Mass Spectrum

For the background model, *Kernel Density Estimation* is used to obtain the shape of $\Lambda_c^+ \overline{D}^{*0}$. The shape is obtained by taking the truncated series of Legendre Moments up to a specific order ($k_{max} = 4$ in this case). This is described in Section 4.14. The $P_c^+$ signals are introduced, centered around the masses of $P_c(4312)^+$, $P_c(4440)^+$ and $P_c(4457)^+$. The widths of these shapes are fixed to the observed values, highlighted in Table 2.1. The Breit-Wigner shape is used to described the $P_c^+$ resonances. The use of the Breit-Wigner here is understandably a simplified approach. This means that the limits have to be understood as valid for narrow $P_c^+$ resonances which do not interfere with each other. In reality, various efects could ammount to a much more complicated shape of the combined system of the 3 $P_c^+$ states. With more data becoming available in the future studies using such more complicated models would be a good extension to this analysis.

The point spread functions, described in the section on resolutions 4.13 is convolved with the Breit-Wigner shape. This procedure is valid for searches of narrow pentaquark states, thus using local resolution in m($\Lambda_c^+ \overline{D}^{*0}$) is a good conservative way of setting upper limits without too much model dependent input. The nominal fits are provided in Figures 4.56, 4.57, 4.58 for $P_c(4312)^+$, $P_c(4440)^+$, $P_c(4457)^+$ respectively. Since the statistics in the region of interest, where the pentaquarks are predicted is limited, the fits return non-zero values for $P_c^+$ yields, but the errors on these obtained parameter values are large. Then, as explained before, the same fit is rerun at different values of the introduced $P_c^+$ fractions. The CLs value is extracted.

### 4.15.2 Upper Limits for $P_c^+$ States

During the CLs limit setting procedure (described in Section 3.2.3 of the Statistics Background Chapter 3), a test statistic has to be chosen. Three of the most used test statistics are the Simple Likelihood Ratio (SLR) (used at LEP), the Ratio of Profiled Likelihoos (RPL) (used at the Tevatron) and the Profile Likelihood Ratio (PLR) used at the (LHC). The difference between the them is how the parameters of the model are reevaluated at each point for the parameter of interest. This is called *profiling*. Thus, the test statistic is called the Profile

**Figure 4.56** *Nominal fit to data for $P_c(4312)^+$. $ECC_{\pi^0}$ hypothesis.*

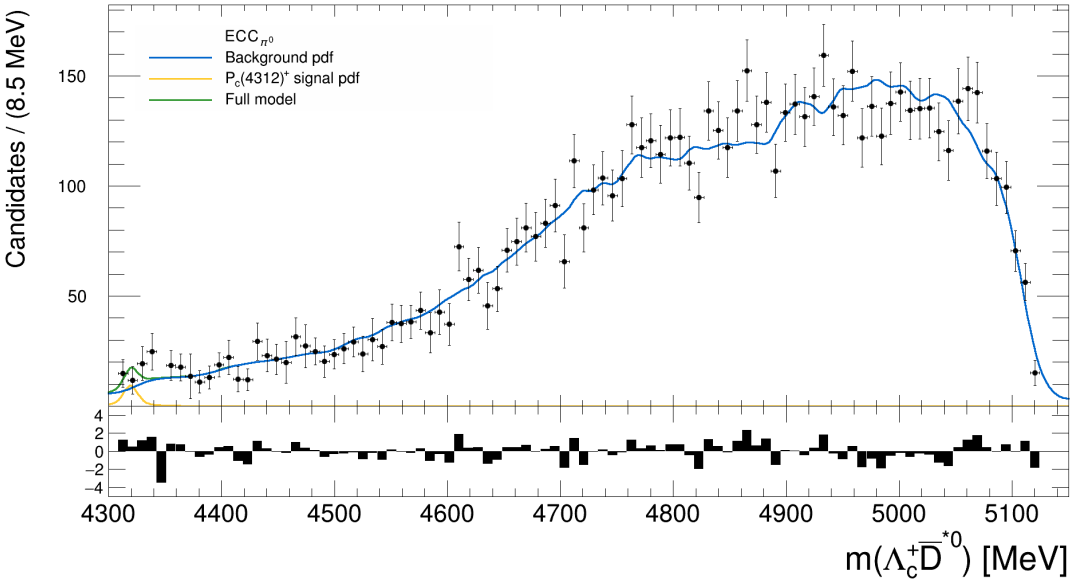



**Figure 4.57** *Nominal fit to data for $P_c(4440)^+$. $ECC_{\pi^0}$ hypothesis.*

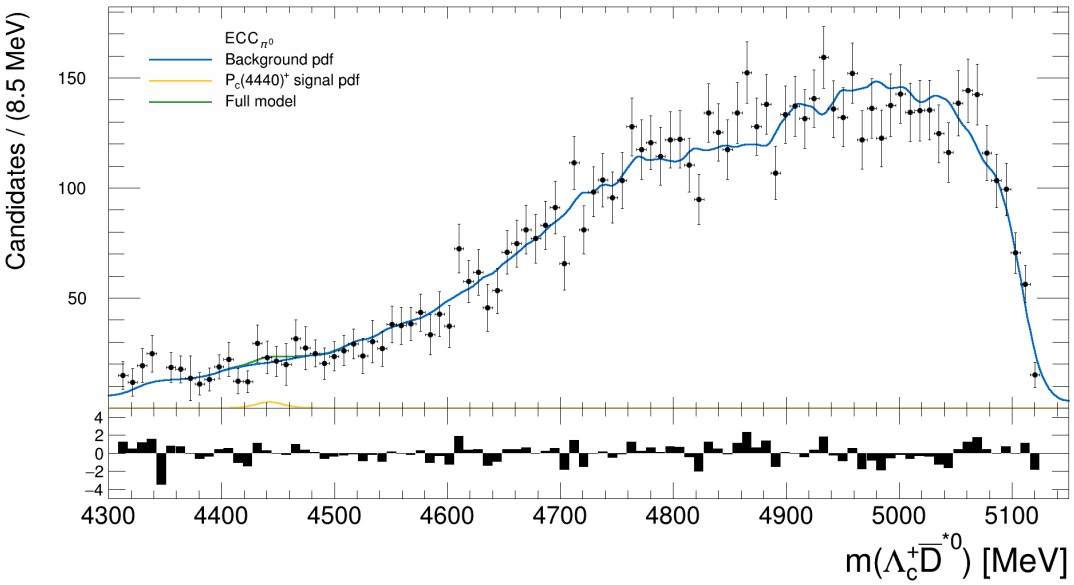
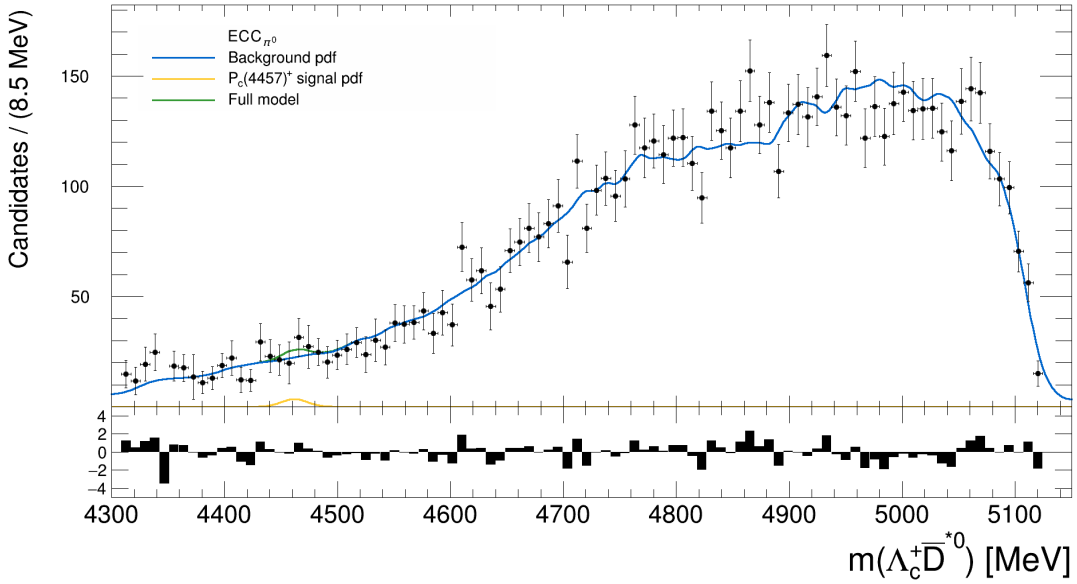


**Figure 4.58** *Nominal fit to data for $P_c(4457)^+$. $ECC_{\pi^0}$ hypothesis.*




| Experiment | Test Statistic used |
|---|---|
| LEP | $Q_\mu = -2\,ln\frac{\mathcal{L}(data\|\mu,\tilde{\theta})}{\mathcal{L}(data\|0,\tilde{\theta})}$ |
| Tevatron | $Q_\mu = -2\,ln\frac{\mathcal{L}(data\|\mu,\hat{\theta}_\mu)}{\mathcal{L}(data\|0,\hat{\theta}_0)}$ |
| LHC | $Q_\mu = -2\,ln\frac{\mathcal{L}(data\|\mu,\hat{\theta}_\mu)}{\mathcal{L}(data\|\hat{\mu},\hat{\theta})}$ |

**Table 4.17** *Test statistic used in different experiments.*

Likelihood Ratio or the Ratio of Profiled Likelihoods. For the Simple Likelihood Ratio, on the other hand, the change in likelihood is evaluated when probing for different signal fractions but the background model always stays the same and is not reevaluated. In this analysis it was noted, that the choice of a test statistic does not greatly impact the obtained value of the upper limit at a set confidence level. This shows the robustness of the models and the procedure. When using PLR, one gains in sensitivity in the regions where the signal is very small. The drawback is then the computing time. If the number of pseudo-experiments is large $\mathcal{O}(1000)$, the process can take days or weeks on a workstation computer. SLR upper limits are set in a matter of hours and a much larger number of pseudo-experiments allows for a more even contour of the predicted CLs value.

The result of the CLs limit setting procedure is the upper limit on the fit fraction of the $P_c^+$ signal component in the m($\Lambda_c^+\overline{D}^{*0}$) invariant mass spectrum. This is

expressed in % of the entire fitted population and is lower than 1% in this data sample. The values for the upper limits on $P_c^+$ fit fractions for the three different $P_c^+$ states are provided in Section 4.15.2. The test statistic used here is the Profile Likelihood Ratio and the confidence level is set to 95%.

**Table 4.18** *CLs upper limits results summary on $f_{P_c^+}$. Test statistic: Profile Likelihood Ratio. Hypothesis: $ECC_{\pi^0}$*

| Pentaquark | $f_{P_c^+}^{ECC_{\pi^0}}$ |
|---|---|
| $P_c(4312)^+$ | 0.52 % |
| $P_c(4440)^+$ | 0.65 % |
| $P_c(4457)^+$ | 0.54 % |

The change of the CLs value can be seen in the, so called, *Brasic plot*. Regardless of the chosen test statistic, when the introduced $P_c^+$ signal is 0, the Likelihood ratio is 1. For the first point in the parameter of interest, the fit fraction ($f_{P_c^+}^{ECC_{\pi^0}}$) the background model is compared to itself. When higher and higher signal is introduced, the likelihood ratio yealds smaller and smaller values. The behavior here is slightly different in the SLR case in Fig. 4.59 and the PRL case in Fig. 4.60. This is because, when the statistics is limited and the background can not be rejected with high confidence, profiling the Likelihood for both signal and backgound model leaves the CLs value unchanged. Only when a sufficienty high signal is introduced, the PRL and SLR behavior becomes similar. PRL is commonly used in the LHC experiments because this way background fluctuations are taken into account.

**Figure 4.59** *CLs upper limits for $P_c(4312)^+$, $P_c(4440)^+$ and $P_c(4457)^+$. Simple Likelihood Ratio chosen as the test statistic. $ECC_{\pi^0}$ hypothesis.*

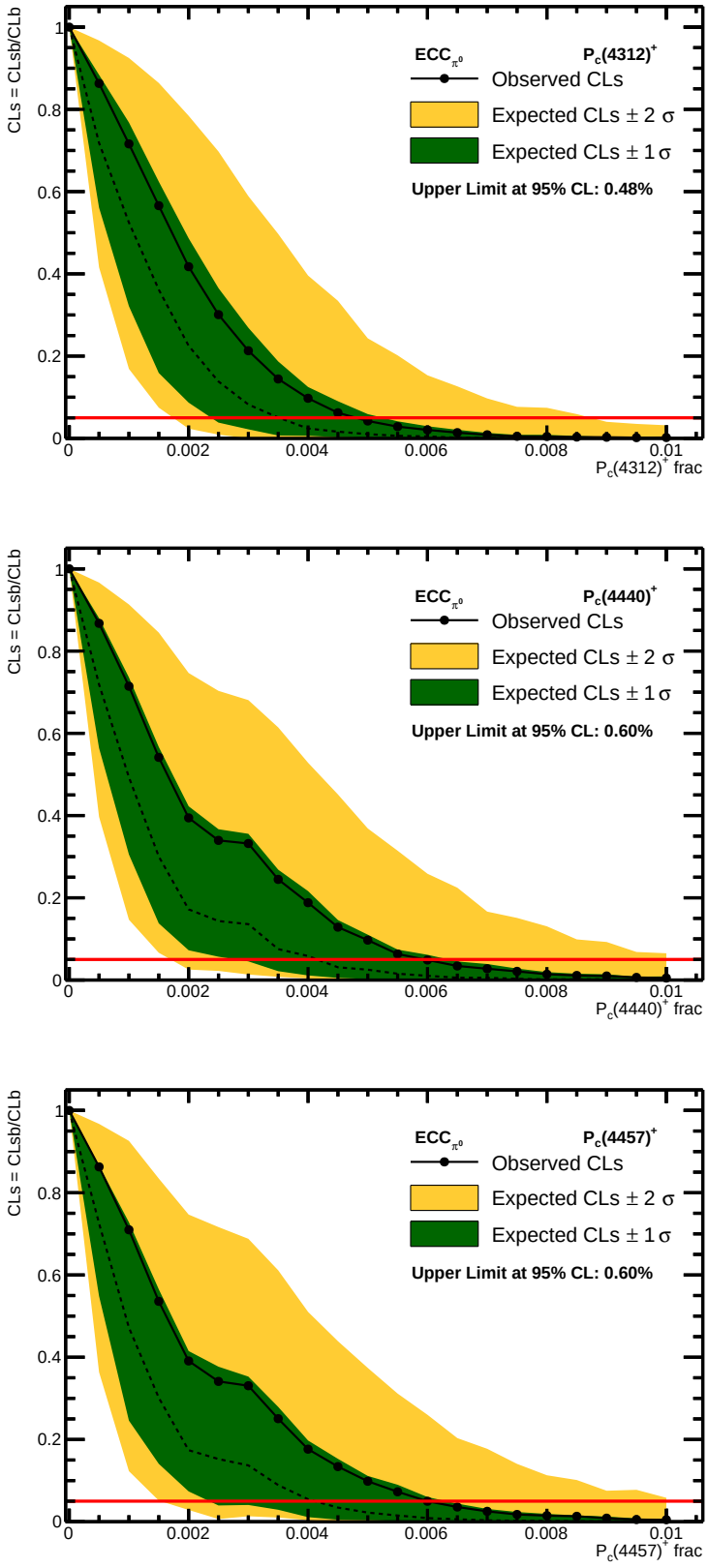
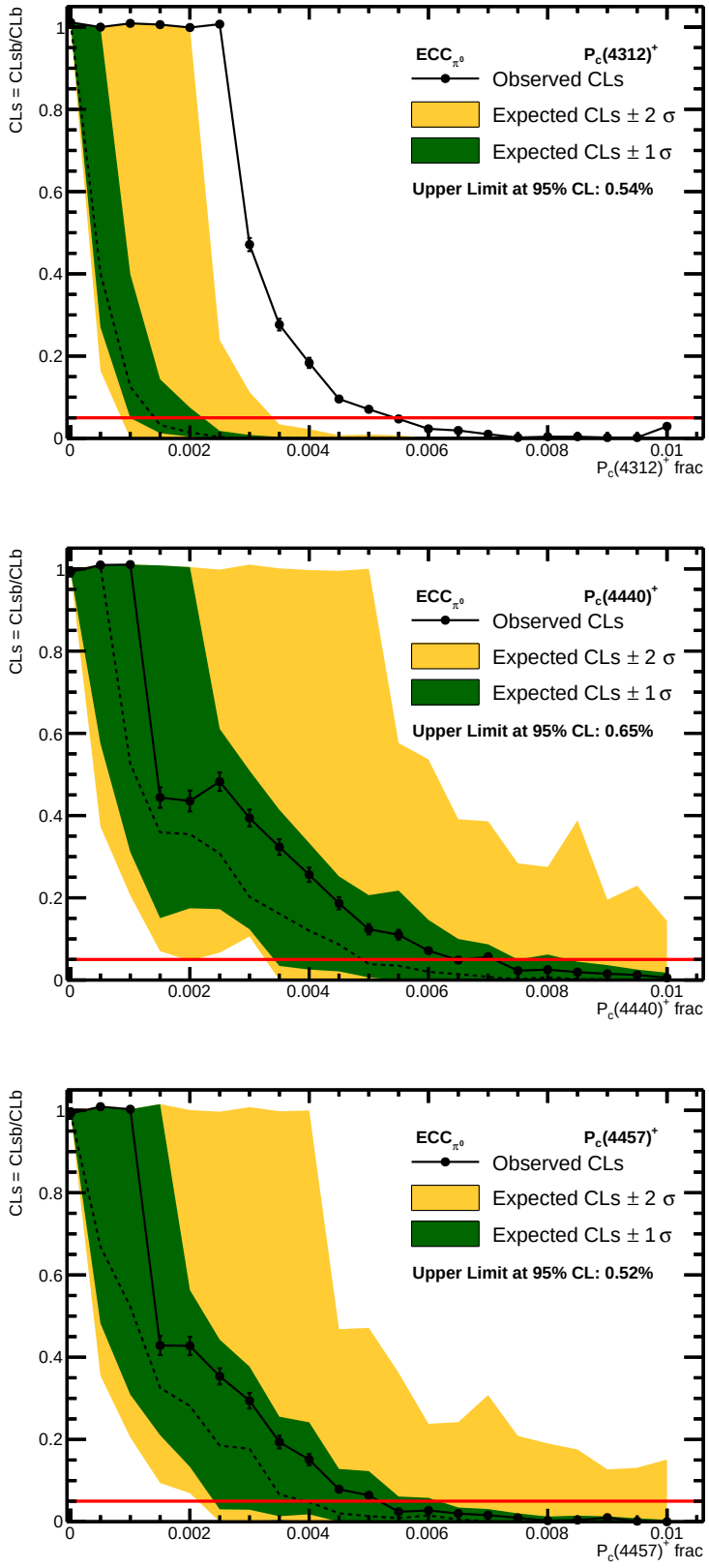


**Figure 4.60** *CLs upper limits for $P_c(4312)^+$, $P_c(4440)^+$ and $P_c(4457)^+$. Profile Likelihood Ratio chosen as the test statistic. $ECC_{\pi^0}$ hypothesis.*



## 4.16 Systematic Checks

A number of systematic checks were performed in the analysis. These checks were mostly related to choice of binning in resolution graph and efficiency map, also different $H_0$ background models were studied as an input to CLs upper limit setting procedure. More over, two different test statistics were investigated. A summary of these checks and assigned systematic uncertainties is provided in this section.

A summary of systematic checks and total systematic uncertainty for three different $P_c^+$ states is provided in Table 4.19. It is clear that for $P_c(4312)^+$ systematic uncertainties are larger than for the other two. The main contribution to the total systematic uncertainty comes from the $H_0$ background model, there is a significant difference in the yields when parametrising the entire $\Lambda_c^+ \overline{D}^{*0}$ spectrum using Kernel Density Estimation or by using `RooHistPdf` and using an exponential model in the region where the pentaquark $P_c(4312)^+$ is expected. Since $P_c(4312)^+$ is very close to the edge of the phase space and the number of candidates in that region is very limited, the $H_0$ background model has a large impact on the observed yield or fit fraction.

The systematic uncertainties are propagated back to the fit model as a Gaussian constraint centered at $\mu = 1$, the width $\sigma$ is set to the systematic uncertainty for a specific $P_c^+$ state.

| **Variable** | $\mathbf{P_c(4312)^+}$ | $\mathbf{P_c(4440)^+}$ | $\mathbf{P_c(4457)^+}$ |
|---|---|---|---|
| $H_0$ Background model | 10.00% | 3.41% | 1.13% |
| Resolution graph binning | 7.12% | 3.50% | 2.73% |
| Efficiency map binning | 6.72% | 4.56% | 6.29% |
| CLs test statistic | 0.98% | 0.00% | 0.54% |
| **Total** | 14.03 % | 6.68 % | 6.97 % |

**Table 4.19** *Summary of systematic checks.*

### 4.16.1 Efficiency Map Binning

To correct for efficiency effects, an efficiency map in m($\Lambda_c^+ \overline{D}^{*0}$) and the helicity angle of $\Lambda_c^+ \overline{D}^{*0}$ is used (as defined in Section 2.8 of the Theory Chapter 2). This map is constructed by taking a large MC sample and putting the candidates through the same selections and ECC reconstruction procedure as that of data. The efficiency map is normalized such that the mean bin content corresponds to unity. With this in mind, it becomes important how finely the efficiency map is binned, since the data sample is much smaller than that of the MC, the number of bins is varied in both m($\Lambda_c^+ \overline{D}^{*0}$) and $\theta$ variables. The effect of this on the upper limit is then evaluated. All other parameters of the analysis are kept constant.

The number of bins for the efficiency map is changed to 10, 12, 15, 20, 25, 30. The working point in this variable for the analysis is 10, since the data sample is limited and should not be binned too finely. Extreme values of 5 and 7 were also investigated but at that stage phase space is binned too broadly and systematic uncertainty increases. For each of these values, difference from the nominal value (at the working point) is calculated. Then the standard deviation of this set of differences is computed. The systematic uncertainty for three different $P_c^+$ states is reported as $\sigma/N$ where N is the nominal yield.

The obtained systematic uncertainties coming from efficiency are:

$$\sigma^{eff}_{P_c(4312)^+} = 6.72\%$$
$$\sigma^{eff}_{P_c(4440)^+} = 4.56\%$$
$$\sigma^{eff}_{P_c(4457)^+} = 6.29\%$$

### 4.16.2 Resolution Graph Binning

The resolution is evaluated from MC as described in Section 4.13. Resolution in m($\Lambda_c^+ \overline{D}^{*0}$) is one of the inputs to the signal model for pentaquark shapes used in the limit setting procedure. This is why the resolution is computed for each bin of m($\Lambda_c^+ \overline{D}^{*0}$) with a chosen binning scheme. Then a graph is constructed interpolating the obtained values of the resolution. An important check is to see how binning in m($\Lambda_c^+ \overline{D}^{*0}$) changes the obtained values of resolution and eventually the final result. While the change in the resolution in the extremes of this check

was $\sim 50\%$ (for incorrect ECC hypothesis and taking the largest change), the change in the upper limit is not as pronounced so binning in the resolution doesn't introduce a large systematic uncertainty. The selected working point for resolution binning is 20 bins in m($\Lambda_c^+ \overline{D}^{*0}$). The resolution binning was varied selecting one of the values: 5, 7, 10, 12, 15, 20, 25, 30, 35, 40. The systematic uncertainty is reported in the same way as that of the binning in the efficiency map, taking the standard deviation of a set of differences and dividing by the nominal value.

$$\sigma^{res}_{P_c(4312)^+} = 7.12\%$$

$$\sigma^{res}_{P_c(4440)^+} = 3.50\%$$

$$\sigma^{res}_{P_c(4457)^+} = 2.73\%$$

### 4.16.3 $H_0$ Background Models

Different $H_0$ background models were studied: `KDE` - background shape parameterized by Kernel Density Estimation, `hist` - background shape approximated directly as a `RooHistPdf`. This binned background model is obtained directly by reweighting the phase space MC sample with weights obtained from the moments analysis described in Sections 2.8 and 4.14. This is motivated by the idea that in the process of computing Legendre moments, one of the Dalitz projections ($\overline{D}^{*0}K^-$) is binned and moments are obtained in each bin of m($\overline{D}^{*0}K^-$) separately. Some issues arise when introducing a Gaussian signal on top of the binned background pdf. Since binning can not be too fine because of the limited data sample (especially around the region of interest at $\sim$ 4300 - 4500 MeV) a Gaussian signal spans only a few bins and artifacts while fitting the combined model can arise reducing the accuracy of such model. A study with exponential background model where the fit was performed only in the region where $P_c^+$ states can be observed (18.5 - 20.5 GeV$^2$ in m$^2$($\Lambda_c^+ \overline{D}^{*0}$) inv. mass spectrum). This study was instrumental in highlighting the model sensitivity of the $P_c(4312)^+$ state, since it is so close to the edge of the phase space and also in the region with very limited statistics. The nominal $H_0$ background model is `KDE` as it spans the entire $\Lambda_c^+ \overline{D}^{*0}$ spectrum but is not binned like `hist` model. Exponential model is locally accurate, but suffers from very high statistical uncertainties because of low population of candidates in the region of interest. It is more used as a cross check.

The systematic uncertainties related to the choice of the $H_0$ background model

are found to be:

$$\sigma^{bkg}_{P_c(4312)^+} = 10.00\%$$
$$\sigma^{bkg}_{P_c(4440)^+} = 3.41\%$$
$$\sigma^{bkg}_{P_c(4457)^+} = 1.13\%$$

**Test statistic**

As part of the systematic studies as well as a general investigation in the analysis, CLs upper limits are computed separately using Profile Likelihood Ratio (PRL) and Simple Likelihood Ratio (SLR) as the test statistic. The systematic uncertainties related to this choice are found to be minimal.

$$\sigma^{t}_{P_c(4312)^+} = 0.98\%$$
$$\sigma^{t}_{P_c(4440)^+} = 0.00\%$$
$$\sigma^{t}_{P_c(4457)^+} = 0.54\%$$

## 4.17 Summary

In this analysis, the decay channel $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ was investigated in order to look for pentaquarks $P_c(4312)^+$, $P_c(4440)^+$, $P_c(4457)^+$ in the $\Lambda_c^+ \overline{D}^{*0}$ system. In order to be able to construct the Dalitz plot of the $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ decay, the missing momentum of the $\overline{D}^{*0}$ candidate first had to be reconstructed. This missing momentum arises from the fact that the $\pi^0$ or the $\gamma$ are missing from the decay and the selected data sample is actually that of the decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$. The reconstruction efficiency for the $\gamma$ or the $\pi^0$ in LHCb for an analysis of this type would be too low. A kinematic over-constraint method (Extended Cone Closure (ECC)) was used to reconstruct the missing momentum. By using this method, the decay $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^{*0} K^-$ is reconstructed under a hypothesis of a missing $\gamma$ ($ECC_\gamma$) or missing $\pi^0$ ($ECC_{\pi^0}$). By inspecting the Dalitz plot $\mathrm{m}^2(\Lambda_c^+ \overline{D}^{*0})$- $\mathrm{m}^2(\overline{D}^{*0} K^-)$ and the $\mathrm{m}(\Lambda_c^+ \overline{D}^{*0})$ mass projection, no clear pentaquark signal was observed.

Model independent analysis using Legendre moments was performed in order to probe for possible presence of the exotic resonances and to get the background shape of the $\mathrm{m}^2(\Lambda_c^+\overline{D}^{*0})$ invariant mass spectrum. Then, a CLs limit setting procedure was performed for different $H_0$ background models, both of the ECC hypotheses and three different pentaquarks. The reported upper limits are those for $ECC_{\pi^0}$ hypothesis since $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\pi^0)K^-$ contribution amounts to $\sim 65\%$ of the candidates in the $\Lambda_b^0 \to \Lambda_c^+\overline{D}^{*0}(\to \overline{D}^0\ \pi^0/\gamma)K^-$ decay. Effects of incorrectly choosing a hypothesis were investigated and described. It was found that incorrect hypothesis makes the resolution worse in the Dalitz plot but does so in a symmetric manner. This just means that the $\sim 20\%$ of events have wider point spread function in the Dalitz plot. This is accounted for in treatment of resolution effects. The CLs upper limits are set at 95% confidence level and reported to be:

$$f_{P_c(4312)^+} \leq 0.52\%$$

$$f_{P_c(4440)^+} \leq 0.65\%$$

$$f_{P_c(4457)^+} \leq 0.54\%$$

These limits are expressed as the fit fraction of the $P_c^+$ signal in the $\mathrm{m}(\Lambda_c^+\overline{D}^{*0})$ invariant mass spectrum. The fact that no clear signal of pentaquarks is visible in $\Lambda_c^+\overline{D}^{*0}$ system is a valuable input to the theory community. The upper limits set in this analysis can help to probe the validity of various theoretical models predicting different pentaquark yields im $\Lambda_c^+\overline{D}^{*0}$ system. For example, models predicting a triangle singularity, due to rescattering effects also predict much larger yelds than the ones observed in this channel. Models like [97] predicting sizable yields in $\Lambda_c^+\overline{D}^{*0}$ channel are thus are put under stress. On the other hand, there are models (e.g. [102] using the chiral unitary approach) where no large $P_c^+$ contribution to the $\Lambda_c^+\overline{D}^{*0}$ spectrum is predicted. It is important to note that this analysis is valid as a search for narrow Breit-Wigner-like pentaquark states. With more data becoming available in Run3 of the LHC this analysis could be rerun in order to see if any signal of $P_c^+$ states is visible or to further constrain the upper limits set here.

# Chapter 5

# Analysis tools

A number of computational tools used for this analysis are introduced in this chapter. Some of the main features and the capabilities relevant to this and future physics analyses are highlighted. Since this analysis was structured with preservation of analysis workflows and reproducability in mind, the described tools used in combination provide a possible solution for creating such analyses.

For data preselection, DaVinci framework is used which is available as part of the freely distributed LHCb software. Another piece of LHCb software which is mainly used for the second part of the thesis (Open Data Release) is `LHCb DIRAC`. For the technical parts of the analysis dealing primarily in data handling, computations of physical variables and statistical methods the aim was to use most up-to-date modularized tools (or develop such tools) in order to make the analysis easily understandable, repeatable and extendable.

The analysis is primarily written in `Python` programming language, the scientific data analysis framework used is `ROOT`, or more accurately `PyROOT` - implementation of ROOT framework in `Python`. Modern approaches such as `RDataFrame` are used for swift and parallelizable handling of data (uncorrelated data entries are treated as arrays, `RDataFrame` allows applying a set of transformations or actions on entries of such arrays in parallel).

Another big focus of the analysis, with the aim to become one of the very first LHCb analyses where modern analysis preservation techniques would be well

demonstrated was creating analysis pipeline, where all the steps of the analysis could be performed from start to finish. Using inputs from shared remote storage (so that there is no need to download or keep large unfiltered data files ($\sim$ 1TB) locally) and finishing up with print-quality plots and figures as well as relevant data files ($\sim$ 10MB). To achieve this, the analysis is all set up in a `snakemake` workflow engine. Continuous integration tools (`CI`) from `Git` are used to check the validity of results when rerunning the analysis. The analysis was also checked and proven to work in `REANA` framework. Since it was one of the very first LHCb analyses ran of `REANA`, the tools and back end of `REANA` was also updated to make it more usable for future LHCb analyses.

## 5.1 Python Programming Language


**Figure 5.1** *Logo of the Python programming language.*

`Python` is a general purpose programming language used for many different tasks, including those of data analysis [40]. The choice to write (or rewrite) most of the analysis code in `Python` was guided by the exponential growth in the popularity of this language for many different use cases, ranging from physics related, data analysis specific areas to machine learning, reporting and integration with other programming tools.

## 5.2 Analysis Packaging in Python

Isolated pieces of software written in `Python` are called Packages and are all listed in Python Package Index (PyPI). Among thousands of these packages, which are all freely available there are PyHEP packages, especially created by and for physicists and computer scientists in the High Energy Physics community. When working on this analysis, a separate Python package called `lcdstk` was also created, hosting a number of tools related to computing physical variables of particles (e.g. Dalitz variables), creating histograms or 2D scatter plots, running parameter estimation and minimization tasks and creating print-quality reports. Making analysis specific tools available as a package allows for much quicker workflow,

implementation of new analysis steps, repeating or replicating the current steps as well as making an analysis a lot more portable. An analysis can be easily rerun on any platform by creating a well defined environment with `lcdstk` package, all the dependencies are installed automatically.

## 5.3 ROOT Data Analysis Framework


**Figure 5.2** *Logo of ROOT data analysis framework.*

ROOT was created and is widely used in many fields of science where large samples of data need to be studied. It was written to be fast (written in `C++`). Many of mathematical, statistical tools as well as common visualization techniques used in science are readily available within the ROOT data analysis framework. Recently, ROOT was heavily updated to adhere with the current paradigms of data analysis, allowing parallel computations using multiprocessing capabilities of current computers. A big advancement in ROOT was the introduction of ROOT data frame - a modern high-level interface for analysis of data stored in various common formats. Calculations are expressed in terms of a type-safe functional chain of actions and transformations. The use of so called `lazy` actions allow to set up complete parts of analysis workflow. The actions are `booked`, checked for integrity and registered with `RDataFrame`. The execution of code only takes place when all the lazy actions are finished. This creates an optimized way of reading data, performing computations and saving results. Inputs are only read once, and actions are performed in a logical manner, without repeating any unneeded operations. Many-fold decrease in computation time for various analysis steps (e.g. data selection) is possible using `RDataFrame` as compared to some older ways of performing these tasks.

## 5.4 Snakemake Workflow Engine


**Figure 5.3** *Logo of Snakemake workflow engine.*

The Snakemake workflow management system is a tool to create reproducible and scalable data analyses. Workflows are described via a human readable, Python based language. They can be seamlessly scaled to server, cluster, grid and cloud environments, without the need to modify the workflow definition. Finally, Snakemake workflows can entail a description of required software, which will be automatically deployed to any execution environment [134]. Snakemake uses acyclic logic for workflow execution. Isolated steps of the analysis are performed in sequence (the inputs or computation might be using multiprocessing capabilities). The steps which are not dependent of each other can be also run in parallel. Different steps are defined as `rules` of snakemake workflow. A rule graph for the analysis is shown in Fig. 5.4. Similar graph to this, but now showing all the different inputs to various analysis steps - the Directed Acyclic Graph (DAG) is shown in Fig. 5.5.

In the aim to use this analysis as an example of analysis preservation, the snakemake reports were also investigated and used for parts of the analysis. This is a tool which allows to define and categorize analysis results and save them in `html` format. These report files can be made available on a dedicated website or a server. They are also accessible as artifacts of the GitLab CI. Not only the final results are provided in the report, but computation times of a specific run of the analysis are provided for each step separately. This is a great tool which is very useful during the analysis to check for bottlenecks in the computation and accumulate interim and final results in one place.


**Figure 5.4** *The rule graph indicating the relations of analysis steps and their execution sequence defined in* ***Snakemake*** *workflow engine.*

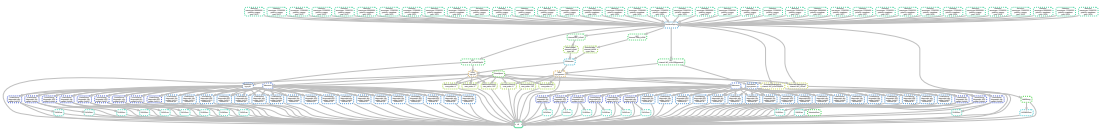


**Figure 5.5** *The Directed Acyclic Graph indicating the relations of analysis steps and their execution sequence defined in* ***Snakemake*** *workflow engine. Separate inputs to different analysis steps are shown here. It is clear that analysis is highly parallelized.*

## 5.5 GitLab CI


**Figure 5.6** *Logo of GitLab Continuous Integration.*

`GitLab` infrastructure at CERN allows the user to follow good version control practices for the physics analysis. Additions to the software related to the analysis can be committed to `remote` - a secure server. This helps with both analysis development and preservation as well as distributing the analysis and its results. A great tool used within this `GitLab` infrastructure is `Continuous Integration` (CI). This way, with every `commit` to the remote analysis repository a number of tests are run to check the integrity of the analysis and save interim results which are then available via web interface for quick reference. CI helps with quality control of different contributions to the analysis. Since some physics analyses, including this one become quite complex with a number of interconnected steps depending on one another it is certainly a good practice to continuously check if the analysis still runs and if the results are as expected. With Gitlab CI, this is done in an automated manner on the server side.

## 5.6 REANA

**Figure 5.7** *Logo of REANA reproducible analysis framework.*

`REANA` is a reproducible research data analysis platform. This is another great tool (still in development at the time of writing) focused on analysis preservation. In physics data analysis it is important not to only preserve the final results, but also the entire workflow, including the computational environment which was used to obtain these results. After finishing the analysis it might need to be rerun on a larger data sample or some of the interim results might become necessary for another analysis related to the former. This is usually done by reaching out to analysis proponents in the hopes that they still have analysis code available somewhere. `REANA` would allow for authenticated users to access all of the code which was run in a specific container and is checked to work. Steps of the analysis or the entire analysis could be rerun with slightly different parameters if needed. Most importantly, since `REANA` is based on code running in containers, legacy analyses which were proven to work on REANA should be still valid years after submitting them.

# Chapter 6

# First LHCb Open Data Release

As part of the commitment by the University of Bonn to the LHCb collaboration, a subset of the LHCb data was prepared and released to the public. This was done in accordance to the LHCb Open Data Policy. Preparing and releasing of the LHCb data was an involved task. Since this was the first such release, a lot of decisions had to be made on how to present the data and how to make it accessible and understandable to an external user. Keeping in mind that the Open Data user would not have access to the internal LHCb computing infrastructure and other resources, the data needed to be copied to the public-facing storage and curated with additional information, hosted on the CERN Open Data Portal (described in Section 6.4). A workflow for incrementally releasing more or different data sets with their additional information and documentation was created. These efforts are described in this chapter.

## 6.1 Introduction to Open Data and FAIR principles

It is in the interest of many parties, including the academia, industry, funding agencies and publishers, to have a well defined and efficient infrastructure of the management of scientific data. The lack of unified criteria or guidelines for obtaining, storing and sharing data from different ongoing or decommissioned science experiments introduces more unneeded complexity and impedes the scientific progress. *The FAIR Guiding Principles for scientific data management and stewardship* [135] were put forward in order to address these issues. *FAIR* here

is an acronym standing for *Findable, Accessible, Interoperable, Reusable*. *Findable* data should have persistent identifiers, rich metadata. The data sets should be well indexed and searchable. *Accessible* data sets should be retrievable through standard protocols. Metadata should be accessible even if the data is not publicly available. *Interoperable* data means the appropriate language and formats are used in data management. Qualified references to other relevant data sets should be provided. *Reusable* data should have correct and appropriate licenses, detailed provenance information and meet the standards set by the community. The main goal of these principles is to make the data discoverable by both the individual and the machine and to enable the efficient reuse of the data.

Some of the data sets stay relevant long after the data collection and the decommissioning of the experiment used to collect these data. Data from the Large Electron-Positron Collider (LEP) [136] will be the largest $e^-e^+$ collision data for about 40 years before the Future Circular Collider FCC-ee [137] may be launched. BaBar collaboration [138] acquired a unique data set of $e^+e^- \to \Upsilon(3S)$ data. JADE [139] data is still analyzed, 30 years after its collection. The LHCb data set is also expected to be useful and relevant long into the future. This makes preservation of the legacy data an important step in the life span of any physics experiment. Even when the data infrastructure of the experiment itself is not in-place anymore, the data set should be accessible to researchers from around the world. As described before, these data should be self sufficient in documentation because users of such data may not have any of the resources from the original infrastructure available to them.

*Data Preservation and Long Term Analysis in High Energy Physics (DPHEP) Study Group* [140] have identified 4 different Levels of complexity to the high energy physics data. The *physics data* here has the meaning of the data sets from the main physics programmes at the LHC. Namely, the proton-proton and heavy-ion collision data. Level 1 data - *Published Results*. Peer-reviewed publications. This is the main output from the experiments at CERN. As well as the research articles and letters being available themselves, some accompanying information is provided. This information may come in the form of figures, likelihoods and tables. Level 2 data - *Outreach and Education*. Dedicated subsets of data. These are formatted and selected to best suit the educational needs. How much of Level 2 data is released and when it is released is determined by each experiment separately.

Light-weight data formats are chosen, the environments used to analyze this data may also be defined. Level 3 data - *Reconstructed Events*. The data with the Level of detail useful for algorithmic, performance and physics studies. Individual data sets are curated by accompanying their metadata. Documentation on how to use Level 3 data is also provided. Level 4 data - *Raw Data*. Direct output of a detector. It is not practically possible to make the raw data from the experiments available outside the collaborations. The required knowledge is very specific to each experiment. The complexity of the data, metadata and the software is too high for any meaningful external use. The computing resources necessary to handle Level 4 data are vast and the volume of stored data is enormous. For these reasons, direct access to Level 4 data is not available even to the individuals within the collaborations.

## 6.2 CERN Open Data Policy

LHCb and other experiments at CERN have made a decision to make the physics data, collected during the lifespan of an experiment, available to the public. This is reflected in the *CERN Open Data Policy* [141] and again described and reaffirmed in the *European Strategy for Particle Physics* [142]. Making data available responsibly (applying *FAIR* standards, as described before), at different levels of abstraction and at different points in time, allows the maximum realization of the scientific potential of experimental collaborations at CERN. This also aids in the fulfillment of the collective moral and fiduciary responsibility to the member states and the broader global scientific community. It also helps to create a legacy for the experiments after their end-of-life, described in Section 6.1.

## 6.3 LHCb Open Data Policy

The LHCb Open Data policy states that Level 3 data sets (as described in Section 6.1) from LHCb should be released just as they appear internally. This is the output of the *Stripping* step as described in the section on LHCb data flow 1.4 in the Detector Chapter 1. The data is released after an initial embargo period. 50% of the data collected in a single run is released after 5 years after end-of-run and 100% - after 10 years after the end-of-run. This is intended to allow the internal users, whom themselves contributed to the data collection and

preparation activities, as well as other tasks within the experimental collaboration, to explore the data first. Subsets of data can still be exceptionally withheld upon request from the proponents if there are any individual analyses ongoing. Since Level 3 data, in most cases, needs to undergo calibration and correction procedures, the LHCb should provide tools for such activities (`PIDCalib` [143], `TrackCalib`, etc.), but no further support to the Open Data community may be given. If the LHCb Open Data set is used for publication, it should be referenced accordingly. It should also be clear that this publication is external to the LHCb and did not undergo the review process by the collaboration. Members of the LHCb collaboration should not use and publish papers using the Open Data sets, since they have the access to the complete LHCb data set. The LHCb publications are thus subject to the internal review procedure. It is of upmost importance that releasing the data to the public as a whole or through some computing service (as described in the Section 6.8) should be safe. Careful considerations have to be taken that malicious code injection or other means of wrong-doing from the outside to the LHCb computing infrastructure become impossible.

## 6.4 CERN Open Data Portal and Storage

External users, who want to work with the data collected by the experiments at CERN can do so by accessing it via the CERN Open Data portal at *opendata.cern.ch*. Data from the four large experiments at CERN - LHCb, ATLAS, CMS and ALICE as well as other experiments - Opera and Phoenix is being made accessible through the the Open Data Portal. The aim of this portal is to provide an entry point to using the data for both education and research purposes. Level 3 data (as described in Section 6.1), which is being released, is intended for research professionals, graduate students and university staff. There are also data sets and pages intended for high school students, for initiatives like *Particle Physics Masterclass* and others. CERN Open Data Portal is a centralized system, solely working with participating experiments in order to create a better and clearer interface for the outside user. The first scientific paper using data accessible via CERN Open Data Portal came out in the year 2017 [144]. The portal is continuously updated following the feedback from both Open Data users and the experimental collaborations. An intuitive and easy to use interface was developed. The welcome page of the CERN Open Data portal is shown in Fig. 6.1.

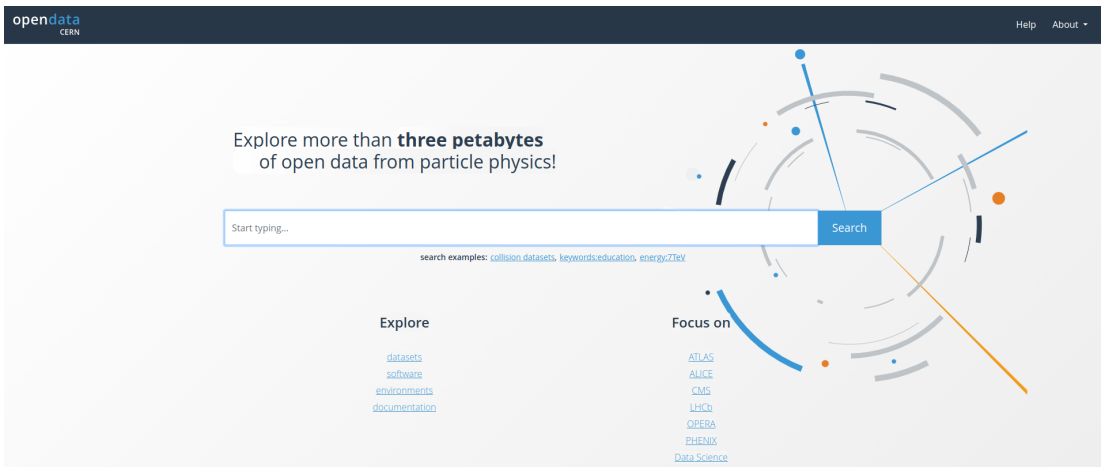



**Figure 6.1** *The welcome page of the CERN Open Data Portal. A specific experiment can be selected from the list or a portal-wide search of records, documentation pages and glossary items can be performed.*

The Open Data Portal hosts various pages and records listing the relevant metadata for each data set, providing documentation about the experiments themselves and the usage of Open Data as well as pages with additional information. The main construct on the Open Data Portal of the metadata accompanying a data set, which was released, is an `Open Data Record`. In these records as much as possible of the relevant information about the data is provided. There are multiple fields which are used as identifiers for the data. These are: the name of the data set, the type of the data set, the experiment which collected the data, the collision energy, the accelerator where the experiment was based, the type of collision, the year of data taking, the number of events, the number of files, the total size of the data set, the production steps, the detector conditions, the trigger conditions, the instructions on how to use the data, the indexes of files and logical file names.

The data files are stored in a dedicated instance of advanced storage system created by CERN, called `eos` [145]. This is an *organic* file storage system built within the `XRootD` framework [146]. Since it is designed to handle vast amounts of data, the system is made such that it demonstrates many of the features of a living organism. Namely resilience, self-healing, adaptive accommodation and constant renewal. `eos` is now widely used across many experiments and working groups at CERN. Some of the characteristics of this system are provided in Table 6.1. The files are stored on an instance of `eos` called `eospublic`, available at *http://eospublic.cern.ch/*. All of the files listed by the Open Data Portal

| Parameter | Value |
|---|---|
| Storage Volume | 780 Petabytes |
| Number of Hard Disks | 60 000 |
| Number of files | 8 000 000 000 |
| Number of clients | 30 000 |

**Table 6.1** *Characteristics of* `eos` *- organic file storage system developed at CERN.*

are categorized and stored on `eospublic`. This storage does not require any authentication and the user could browse and download files directly from there as well as the Open Data Portal.

## 6.5 Open Data Curation - MetadataWriter

LHCb has an extensive Level 3 data set with hundreds of thousands of individual files. Run1 Level 3 data - the output of the stripping step in the LHCb data flow (as described in the data flow Section 1.4 of the Detector Chapter 1) already amounts to about 1 Petabyte of data. With the initial Open Data release, only 20% of these data are made available . Mainly because of storage considerations. This still means that 70 000 of individual files and 28 unique data sets have to be handled.

An automated approach to obtaining, copying and curating files selected for the Open Data release was needed. One of the challenges of the first LHCb Open Data release was bridging the gap between the `LHCbDIRAC` interface, which is accessible only to the members of the LHCb collaboration, with appropriate authentication and the CERN Open Data Portal, where data sets and documentation are in the public domain, accessible to everybody either via web-interface or machine readable API. A software workflow was created which allows the person responsible for releasing the Open Data to access `DIRAC`, write out a list of files which are intended to be released and obtain the metadata accompanying the files. For this part of the Open Data release workflow, a dedicated piece of software was designed, called the `MetadataWriter`.

After authenticating with appropriate CERN and LHCb credentials, the `MetadataWriter` accesses the data catalogs within `DIRAC` and writes out information for a given `Bookkeeping path` (described in Section 1.4 of the

Detector Chapter 1). From the metadata obtained, an `Open Data Record` (6.4) is created automatically, which can be directly uploaded to the CERN Open Data portal. There is a lot of functionality built in to the `MetadataWriter` which helps the person responsible for current and upcoming Open Data releases to inspect, prepare and troubleshoot Open Data Records created for a specific data set. The design of this workflow is such that releasing more data sets and adding new `stripping streams` and `stripping versions` (Section 1.4) with the upcoming Open Data releases should require as little manual intervention as possible. The `MetadataWriter` outputs various files used to complete the Open Data release, these include: `Open Data Records` for each released data set, an index of individual files used in the file staging step, lists of `Logical File Names` (`LFNs`) used for prevalence as this is how files are listed internally in LHCb and information on the combined size of the files for each released data set.

## 6.6 Open Data Preparation

The Open Data Release workflow is such that after being written out by the `MetadataWriter`, the relevant files can be uploaded to Open Data Portal directly. They are then rendered as the `Open Data Records`, shown in Figures 6.3 and 6.4. Also, the documentation pages can be expanded this way, example of which is shown in Figures 6.5 and 6.6. Making of these records and developing `MetadataWriter` was done in close collaboration with the CERN IT team which is responsible for the maintenance of CERN Open Data Portal. A `JSON` schema used in the Open Data Portal was expanded with the LHCb specific fields. LHCb specific facets were added to the selection page in the Open Data portal. These include items like magnet polarity which is very important in categorizing LHCb data as it is switched regularly to deal with the possible systematic effects of running the experiment with a given magnet polarity. The `Stripping stream` and `stripping version` is also something specific to LHCb. The facets are shown in figure 6.2. Within the documentation pages, a drop-down menu is enabled to provide additional information on various LHCb specific entries. This is illustrated in Fig. 6.7.

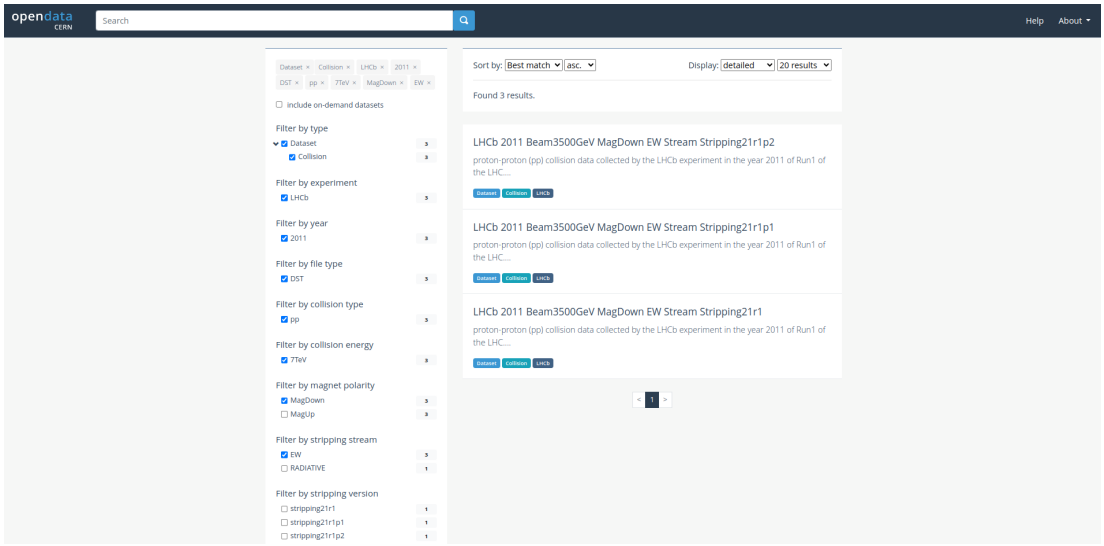



**Figure 6.2** *LHCb specific facets e.g. magnet polarity and stripping version shown in the Open Data Portal.*

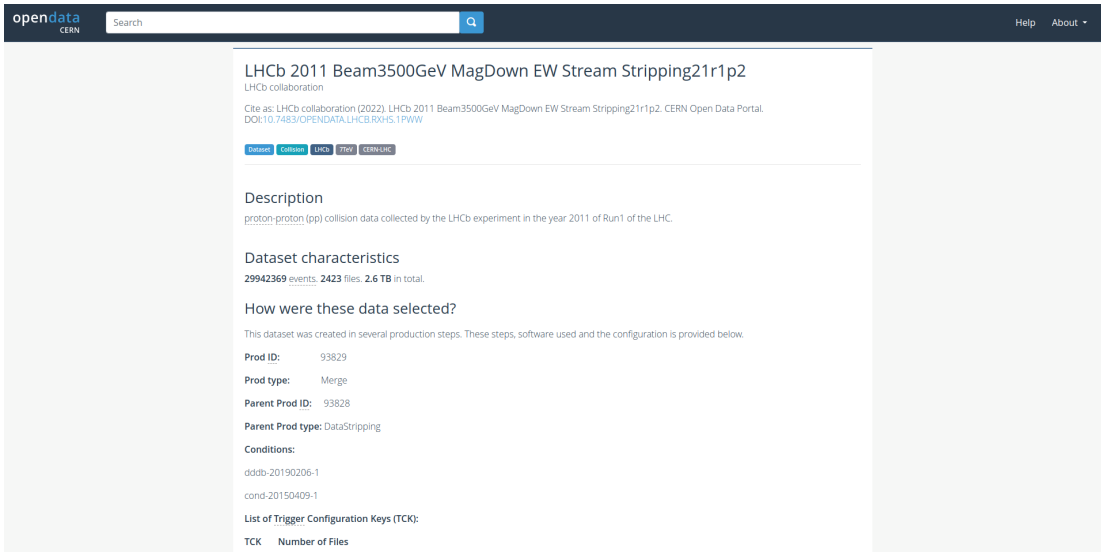



**Figure 6.3** *An example of LHCb Open Data Record as is appears in the CERN Open Data Portal. Information on the file size, production steps, number of events, etc. is provided as the metadata for a selected data set.*

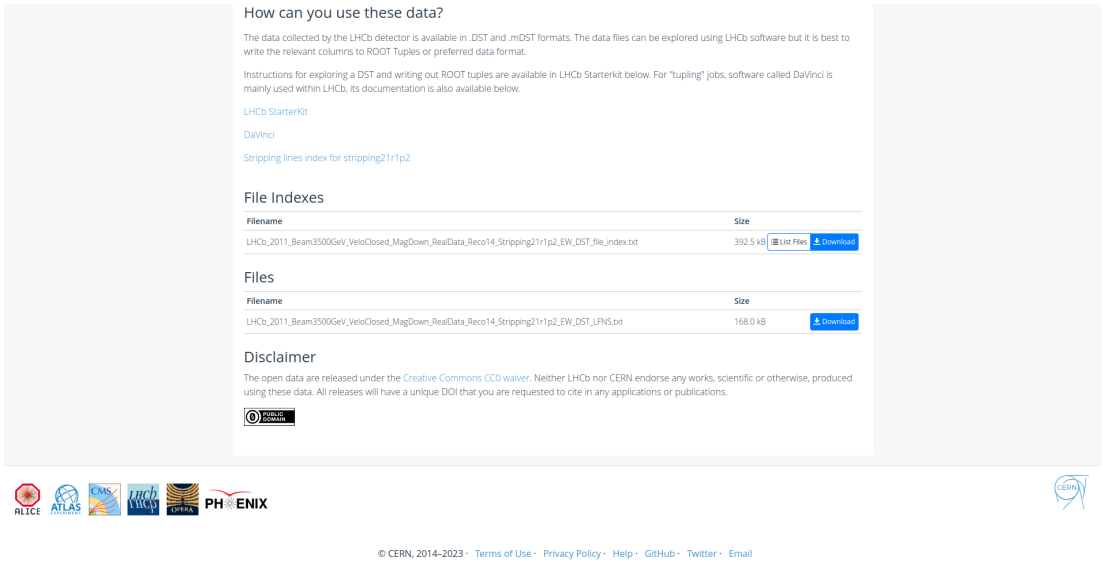



**Figure 6.4** *An example of LHCb Open Data Record as is appears in the CERN Open Data Portal. Instructions on the usage of Open Data files as well as file indexes and lists of files are available with each Open Data Record.*

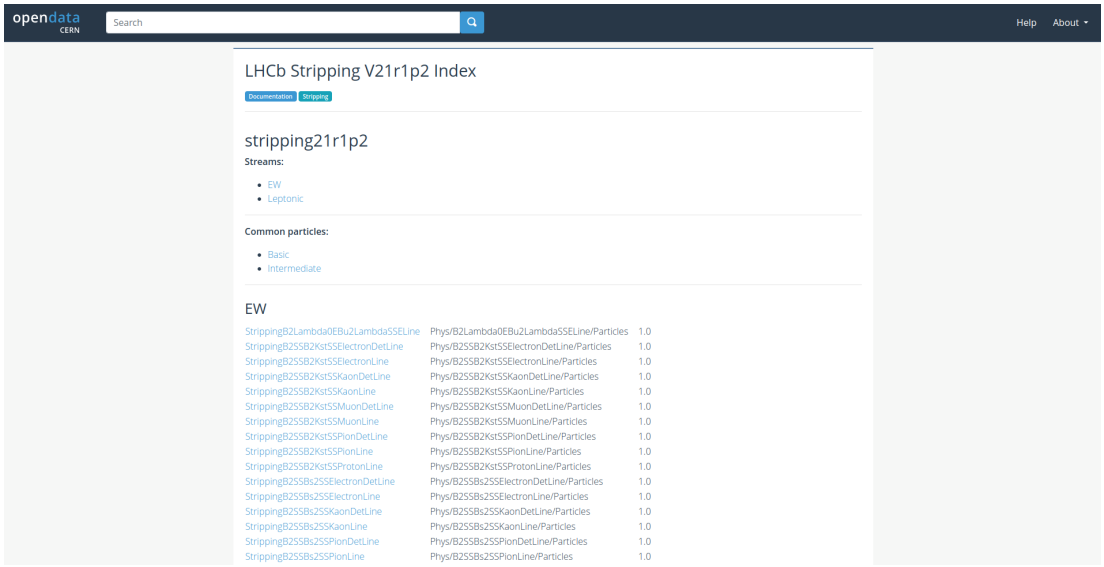



**Figure 6.5** *Information from LHCb Stripping Project copied and provided in the format of CERN Open Data Portal. An interactive list of stripping streams and stripping lines is provided. It is accessible directly via portal-wide search or from a relevant record of a specific detest.*

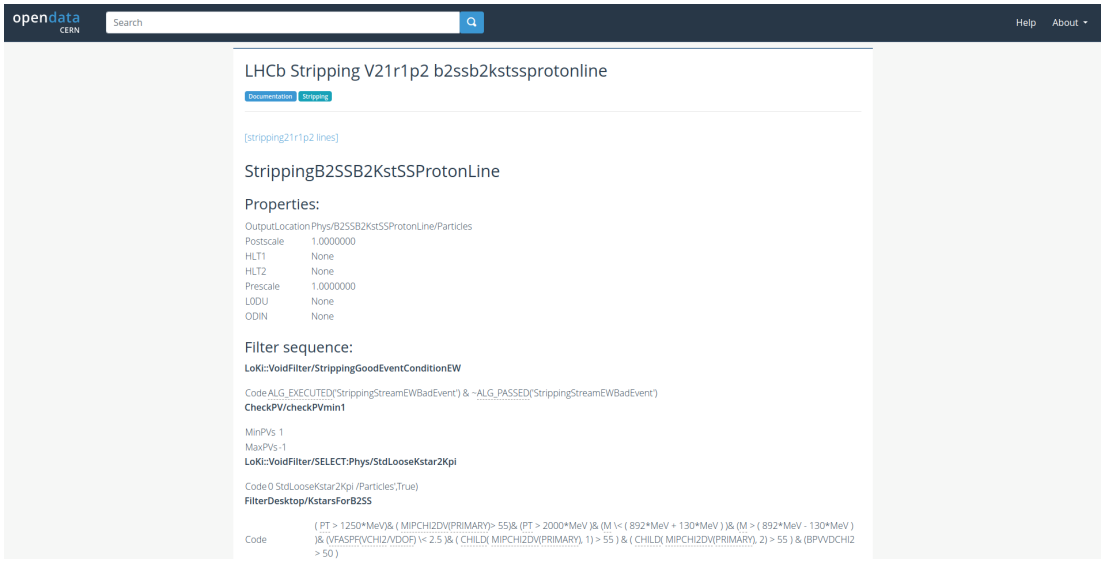



**Figure 6.6** *A specific stripping line documentation page can be opened and inspected. Detailed descriptions on how the stripping line is defined, are available. The user can directly chose an existing stripping line for their analysis or define a new one with similar conditions.*

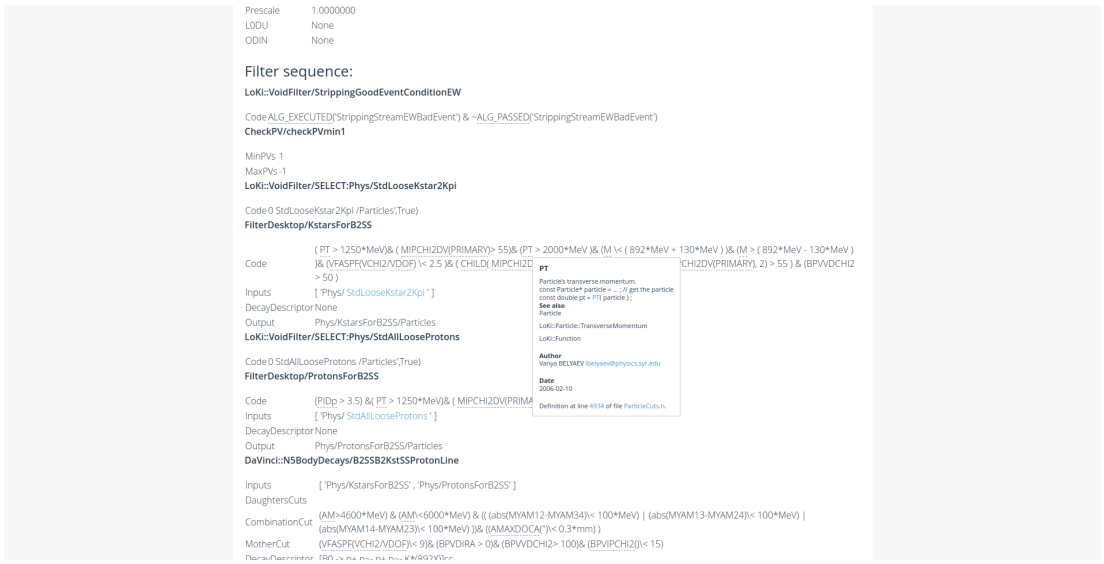



**Figure 6.7** *LHCb specific functors or terms are curated by a drop-down context menu. This menu can be opened on-the-fly by simply hovering the cursor over it. This allows the open data user to navigate and work with a plethora of LHCb specific objects and concepts much easier.*

## 6.7 Open Data Release

With the first release of LHCb Open Data, 200 TB of data were copied to the publicly accessible storage `eospublic`. For every data set released, an index of files was created. Lists of `Logical File Names` were also made available, as the `LFN` is an unchanging identifier of LHCb data. LHCb `stripping` pages are copied and converted to the Open Data Portal format. They are provided as documentation pages. The glossary in the Open Data Portal is enriched with 960 LHCb specific terms. With this proof-of-concept release, only 20% of the Run1 data of LHCb was released. In the near future, the rest of Run1 data should be released as well as the initial portions of Run2 data. The `MetadataWriter` was instrumental in making 70 000 individual files available to the public in an automated and predictable manner, where all of the files are indexed according to their categories within LHCb data management infrastructure. These files each belong to one of the 28 released data sets. The metadata for each data set is accessible on the Open Data portal. The released stripping streams and stripping versions are provided in Table 6.2.

| **Stripping Stream name** |
|---|
| `Radiative` |
| `EW` |
| `Leptonic` |

| **Stripping version** |
|---|
| Stripping 21 |
| Stripping 21r0p1 |
| Stripping 21r0p2 |
| Stripping 21r1 |
| Stripping 21r1p1 |
| Stripping 21r1p2 |

**Table 6.2** *LHCb data stripping streams and stripping versions released with the first Open Data Release.*

## 6.8 Ntuple Wizard

LHCb is also working on alternative services directed at the Open Data users. Some of the resources available internally to collaboration might be made available to aid with the Open Data analyses. One such service is called the `Ntuple wizard` [19]. It is an application which provides a user interface to the Open Data community where the user can define various filters and tools which need to be applied to a chosen data set in order to obtain a much smaller subset of data in a convenient format which can directly be used for physics analysis. In LHCb this is done centrally. The `NTuple wizard` should allow the Open Data community to use some of the data handling capabilities of LHCb. A request submission system would be used. The user will interact with a clear graphical interface dealing with physics related objects and define how the data should be filtered. For safety reasons, so as not to make malicious code injection possible through the Ntuple wizard application, only a YAML configuration file with defined parameter values and selected tools is passed to the next step. After a review of a responsible party from the LHCb side, the provided configuration file is used to filter the data using the LHCb Analysis Productions system [147]. The full architecture of the Ntuple wizard is shown in Fig. 6.8.

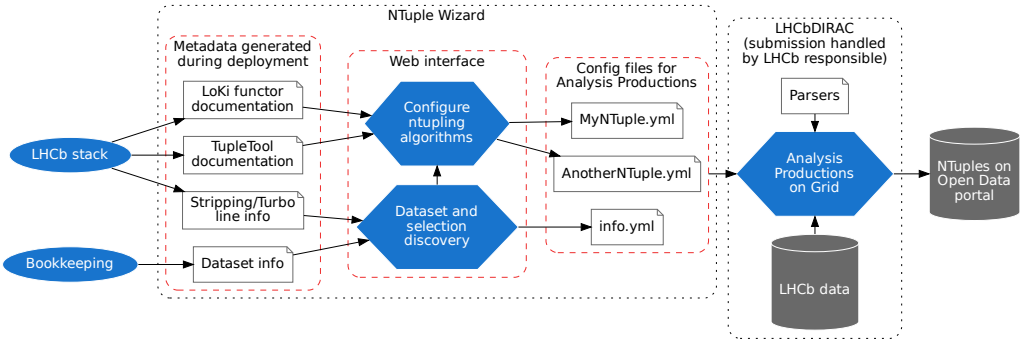



**Figure 6.8** *Architecture of the `Ntuple wizard`. Various stages of the configuration file creation and data filtering workflow are shown. The user would interact via a dedicated application. There is no direct access to LHCb computing infrastructure which makes this approach safe. Reproduced from [19].*

A custom filtered data set will be made available on the Open Data Portal. This way, the Open Data can be made available in a shorter time scale since individual data files do not have to be copied to a dedicated public storage. Also,

Open Data users can make use of the existing LHCb computing infrastructure. Working with large data sets would require a sizable local computing capabilities. Using the `Ntuple wizard`, the output should be small enough in size so that most of computers could handle physics analysis of the data.

## 6.9 Conclusions and Prospects

There is already a track record for the benefits of releasing Open Data to the public. Some of the first data sets to be released were those of the CMS experiment at CERN. Challenges and suggestions arising from CMS Open Data release were published in 2021 [148]. A number of publications using CMS Open Data externally to the collaboration stands as an example of how such initiatives can be adopted by third parties. Scientific papers on jet substructure studies [144, 149], search for new particles [150, 151] and Standard Model analyses [152] are among those where CMS open data was used in a physics analysis.

Further releases of LHCb Open Data are planned already. First, releasing the entire Run1 data set, followed by the incremental releases of Run2 data. The releases are incremental partly because for each release CERN needs to make large volumes of dedicated storage available. This also yields a large financial commitment. In order to expedite the release of Open Data, new ways of sharing data are explored by the LHCb collaboration. One of which is the `Ntuple wizard`, described in section 6.8. There are already groups external to the LHCb starting analyses with the released Open Data. With continued Open Data efforts a number of publications is expected by the third parties. Clarity and good management of experimental data acts as a legacy of successes of the LHCb collaboration.

# Bibliography


[1] F. Landua, *The CERN accelerator complex layout in 2022. Complexe des accélérateurs du CERN en janvier 2022*, 2022.

[2] LHCb collaboration, *LHCb reoptimized detector design and performance: Technical Design Report*, CERN-LHCC-2003-030.

[3] R. Calabrese *et al.*, *Performance of the LHCb RICH detectors during LHC run 2*, Journal of Instrumentation **17** (2022) .

[4] LHCb collaboration, *The LHCb Detector at the LHC*, Journal of Instrumentation **3** (2008) .

[5] R. Arink *et al.*, *Performance of the LHCb Outer Tracker*, JINST **9** (2014) P01002, arXiv:1311.3893.

[6] C. Abellan Beteta *et al.*, *Calibration and performance of the LHCb calorimeters in Run 1 and 2 at the LHC*, arXiv:2008.11556, submitted to JINST.

[7] LHCb collaboration, *LHCb muon system: Technical Design Report*, CERN-LHCC-2001-010.

[8] G. Dujany and B. Storaci, *Real-time alignment and calibration of the LHCb Detector in Run II*, J. Phys. Conf. Ser. **664** (2015) .

[9] S. L. Olsen, T. Skwarnicki, and D. Zieminska, *Nonstandard heavy mesons and baryons: Experimental evidence*, Rev. Mod. Phys. **90** (2018) .

[10] S. Kagiyama, *Pomeron Structure and Multiple-Production Mechanisms*, Progress of Theoretical Physics **51** (1974).

[11] M. N. Chernodub, *QCD string breaking in strong magnetic field*, Modern Physics Letters A (2014) .

[12] A. Usachov, *Study of charmonium production using decays to hadronic final states with the lhcb experiment*, 2019.

[13] E. Ortiz-Pacheco, R. Bijker, and C. Fernández-Ramírez, *Hidden charm pentaquarks: mass spectrum, magnetic moments and photocouplings*, Journal of Physics G: Nuclear and Particle Physics (2019) .

[14] LHCb collaboration, P. Koppenburg, *List of hadrons observed at the LHC*, 2021. LHCb-FIGURE-2021-001.

[15] LHCb collaboration, *Observation of $J/\psi p$ resonances consistent with pentaquark states in $\Lambda_b^0 \to J/\psi p K^-$ decays*, Phys. Rev. Lett. **115** (2015) , arXiv:1507.03414.

[16] LHCb collaboration, *Observation of a narrow pentaquark state, $P_c(4312)^+$, and of two-peak structure of the $P_c(4450)^+$*, Phys. Rev. Lett. **122** (2019) , arXiv:1904.03947.

[17] M. Du *et al.*, *Revisiting the nature of the Pc pentaquarks*, Journal of High Energy Physics **2021** (2021) .

[18] Particle Data Group, *Review of particle physics*, Prog. Theor. Exp. Phys. **2022** (2022) .

[19] C. A. Aidala *et al.*, *Ntuple wizard: An application to access large-scale open data from LHCb*, Computing and Software for Big Science **7** (2023) .

[20] M. Stahl, *First observation of the decay $\Lambda_b^0 \to \Lambda_c^+ \bar{D}^{(*)0} K^-$ in preparation of a pentaquark search in the $\Lambda_c^+ \bar{D}^{(*)0}$ system at the LHCb experiment*, 2018. CERN-THESIS-2018-176.

[21] M. Stahl *et al.*, *Identification of non-prompt charm hadrons*, 2018. LHCb-INT-2018-033. CERN-LHCb-INT-2018-033.

[22] O. S. Brüning *et al.*, *LHC Design Report*, 2004. 10.5170/CERN-2004-003-V-1.

[23] L. Evans and P. Bryant, *LHC Machine*, JINST **3** (2008) .

[24] LHCb collaboration, *The LHCb detector at the LHC*, JINST **3** (2008) .

[25] LHCb collaboration, *LHCb detector performance*, Int. J. Mod. Phys. **A30** (2015) , arXiv:1412.6352.

[26] LHCb collaboration, *Observation of five new narrow $\Omega_c^0$ states decaying to $\Xi_c^+ K^-$*, Phys. Rev. Lett. **118** (2017) , arXiv:1703.04639.

[27] LHCb collaboration, *Observation of new $\Xi_c^0$ baryons decaying to $\Lambda_c^+ K^-$*, Phys. Rev. Lett. **124** (2020) , arXiv:2003.13649.

[28] LHCb collaboration, *First observation of the doubly charmed baryon decay $\Xi_{cc}^{++} \to \Xi_c^+ \pi^+$*, Phys. Rev. Lett. **121** (2018) , arXiv:1807.01919.

[29] LHCb collaboration, *Observation of an exotic narrow doubly charmed tetraquark*, Nature Physics **18** (2022) , arXiv:2109.01038.

[30] LHCb collaboration, *Study of the doubly charmed tetraquark $T_{cc}^+$*, Nature Communications **13** (2022) , arXiv:2109.01056.

[31] LHCb collaboration, *LHCb VELO (VErtex LOcator): Technical Design Report*, CERN-LHCC-2001-011.

[32] LHCb collaboration, *LHCb RICH: Technical Design Report*, CERN-LHCC-2000-037.

[33] M. Adinolfi *et al.*, *Performance of the LHCb RICH detector at the LHC*, Eur. Phys. J. **C73** (2013) 2431, arXiv:1211.6759.

[34] LHCb collaboration, *LHCb magnet: Technical Design Report*, CERN-LHCC-2000-007.

[35] LHCb collaboration, *LHCb inner tracker: Technical Design Report*, CERN-LHCC-2002-029.

[36] LHCb collaboration, *LHCb outer tracker: Technical Design Report*, CERN-LHCC-2001-024.

[37] LHCb collaboration, *LHCb calorimeters: Technical Design Report*, CERN-LHCC-2000-036.

[38] LHCb collaboration, *LHCb computing: Technical Design Report*, CERN-LHCC-2005-019.

[39] A. Tsaregorodtsev *et al.*, *DIRAC3: The new generation of the LHCb grid software*, J. Phys. Conf. Ser. **219** (2010) .

[40] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*, CreateSpace, Scotts Valley, CA, 2009. 1441412697.

[41] LHCb collaboration, *LHCbDIRAC*, 2018. doi: 10.5281/zenodo.1451768.

[42] M. Thomson, *Modern Particle Physics*, Cambridge University Press, 2013.

[43] B. R. Martin and G. Shaw, *Particle Physics*, Manchester Physics Series, Wiley, 2017.

[44] S. L. Glashow, *Partial Symmetries of Weak Interactions*, Nucl. Phys. **22** (1961) .

[45] A. Salam and J. C. Ward, *Electromagnetic and weak interactions*, Physics Letters **13** (1964) .

[46] M. E. Peskin and D. V. Schroeder, *An Introduction to Quantum Field Theory*, Westview Press, 1995.

[47] S. Weinberg, *The Quantum Theory of Fields, Volume 1: Foundations*, Cambridge University Press, 1995.

[48] A. Zee, *Quantum Field Theory in a Nutshell*, Princeton University Press, 2nd ed., 2010.

[49] M. D. Schwartz, *Quantum Field Theory and the Standard Model*, Cambridge University Press, 2014.

[50] J. Chadwick, *The existence of a neutron*, Proc. R. Soc. Lond **A136** (1932), 2053-9150.

[51] M. Gell-Mann, *A Schematic Model of Baryons and Mesons*, Phys. Lett. **8** (1964) .

[52] G. Zweig, *An $SU_3$ model for strong interaction symmetry and its breaking; Version 2*, 1964. CERN-TH-412.

[53] F. Gross *et al.*, *50 years of quantum chromodynamics*, 2022. arXiv:2212.11107.

[54] Y. Ne'eman, *Derivation of strong interactions from a gauge invariance*, Nuclear Physics **26** (1961) .

[55] M. Gell-Mann, *Symmetries of baryons and mesons*, Phys. Rev. **125** (1962) .

[56] R. J. Jaffe, *Multiquark hadrons. I. Phenomenology of $Q^2\overline{Q}^2$ mesons*, Phys. Rev. D **15** (1977) .

[57] R. L. Jaffe, *Multiquark hadrons. II. Methods*, Phys. Rev. D **15** (1977) .

[58] G. C. Rossi and G. Veneziano, *A Possible Description of Baryon Dynamics in Dual and Gauge Theories*, Nucl. Phys. B **123** (1977) .

[59] R. L. Jaffe, $Q^2\overline{q}^2$ *resonances in the baryon-antibaryon system*, Phys. Rev. D **17** (1978) .

[60] H. J. Lipkin, *New possibilities for exotic hadrons — anticharmed strange baryons*, Physics Letters B **195** (1987) .

[61] Belle Collaboration, *Observation of a Narrow Charmoniumlike State in Exclusive* $B^{\pm} \to K^{\pm}\pi^{+}\pi^{-}J/\psi$ *Decays*, Phys. Rev. Lett. **91** (2003) .

[62] LHCb collaboration, R. Aaij *et al.*, *Determination of the* $X(3872)$ *meson quantum numbers*, Phys. Rev. Lett. **110** (2013) , arXiv:1302.6269.

[63] JPAC, A. Rodas *et al.*, *Determination of the pole position of the lightest hybrid meson candidate*, Phys. Rev. Lett. **122** (2019), no. 4 , arXiv:1810.04171.

[64] LHCb Collaboration, *Model-independent study of structure in* $B^{+} \to D^{+}D^{-}K^{+}$ *decays*, Phys. Rev. Lett. **125** (2020) .

[65] LHCb Collaboration, *Amplitude analysis of the* $B^{+} \to D^{+}D^{-}K^{+}$ *decay*, Phys. Rev. D **102** (2020) .

[66] BESIII Collaboration, *Observation of a Charged Charmoniumlike Structure in* $e^{+}e^{-} \to \pi^{+}\pi^{-}J/\psi$ *at* $\sqrt{s}$=4.26 GeV, Phys. Rev. Lett. **110** (2013) .

[67] Belle Collaboration, *Study of* $e^{+}e^{-} \to \pi^{+}\pi^{-}J/\psi$ *and Observation of a Charged Charmoniumlike State at Belle*, Phys. Rev. Lett. **110** (2013) .

[68] BESIII Collaboration, *Observation of a Charged* $(D\overline{D}^{*})^{\pm}$ *Mass Peak in* $e^{+}e^{-} \to \pi D\overline{D}^{*}$ *at* $\sqrt{s} = 4.26$ GeV, Phys. Rev. Lett. **112** (2014) .

[69] BESIII Collaboration, *Observation of* $Z_c(3900)^0$ *in* $e^{+}e^{-} \to \pi^{0}\pi^{0}J/\psi$, Phys. Rev. Lett. **115** (2015) .

[70] BESIII Collaboration, *Observation of a Neutral Structure near the* $D\overline{D}^{*}$ *Mass Threshold in* $e^{+}e^{-} \to (D\overline{D}^{*})^{0}\pi^{0}$ *at* $\sqrt{s} = 4.226$ *and 4.257 GeV*, Phys. Rev. Lett. **115** (2015) .

[71] BESIII Collaboration, *Observation of a Charged Charmoniumlike Structure in* $e^+e^- \to (D^*\overline{D}^*)^{\pm}\pi^{\mp}$ *at* $\sqrt{s} = 4.26$ GeV, Phys. Rev. Lett. **112** (2014) .

[72] BESIII Collaboration, *Observation of a Charged Charmoniumlike Structure* $Z_c(4020)$ *and Search for the* $Z_c(3900)$ *in* $e^+e^- \to \pi^+\pi^- h_c$, Phys. Rev. Lett. **111** (2013) .

[73] Belle Collaboration, *Observation of two resonancelike structures in the* $\pi^+\chi_{c1}$ *mass distribution in exclusive* $\overline{B}^0 \to K^-\pi^+\chi_{c1}$ *decays*, Phys. Rev. D **78** (2008) .

[74] LHCb collaboration, *Evidence for a* $\eta_c(1S)\pi^-$ *resonance in* $B^0 \to \eta_c(1S)K^+\pi^-$ *decays*, Eur. Phys. J. **C78** (2018) 1019, arXiv:1809.07416.

[75] Belle Collaboration, *Observation of a new charged charmoniumlike state in* $\overline{B}^0 \to J/\psi K^-\pi^+$ *decays*, Phys. Rev. D **90** (2014) .

[76] Belle Collaboration, *Observation of a Resonancelike Structure in the* $\pi^{+-}\psi'$ *Mass Distribution in Exclusive* $B \to K\pi^{+-}\psi'$ *Decays*, Phys. Rev. Lett. **100** (2008) .

[77] Belle Collaboration, *Experimental constraints on the spin and parity of the* $Z(4430)^+$, Phys. Rev. D **88** (2013) .

[78] LHCb collaboration, *Observation of the resonant character of the* $Z(4430)^-$ *state*, Phys. Rev. Lett. **112** (2014) , arXiv:1404.1903.

[79] LHCb collaboration, *Model-independent confirmation of the* $Z(4430)^-$ *state*, Phys. Rev. **D92** (2015) , arXiv:1510.01951.

[80] LHCb Collaboration, *Observation of* $J/\psi\phi$ *Structures Consistent with Exotic States from Amplitude Analysis of* $B^+ \to J/\psi\phi K^+$ *Decays*, Phys. Rev. Lett. **118** (2017) .

[81] BESIII Collaboration, *Observation of a Near-Threshold Structure in the* $K^+$ *Recoil-Mass Spectra in* $e^+e^- \to K^+(D_s^- D^{*0} + D_s^{*-} D^0)$, Phys. Rev. Lett. **126** (2021) .

[82] LHCb collaboration, *Observation of new resonances decaying to* $J/\psi K^+$ *and* $J/\psi\phi$ , Phys. Rev. Lett. **127** (2021) , arXiv:2103.01803.

[83] CDF Collaboration, *Evidence for a Narrow Near-Threshold Structure in the $J/\psi\phi$ Mass Spectrum in $B^+ \to J/\psi\phi K^+$ Decays*, Phys. Rev. Lett. **102** (2009) .

[84] D0 Collaboration, *Search for the $X(4140)$ state in $B^+ \to J/\psi\phi K^+$ decays with the D0 detector*, Phys. Rev. D **89** (2014) .

[85] LHCb collaboration, *Study of $B_s^0 \to J/\psi\pi^+\pi^-K^+K^-$ decays*, JHEP **02** (2021) , arXiv:2011.01867.

[86] LHCb collaboration, *Observation of structure in the $J/\psi$-pair mass spectrum*, Science Bulletin **65** (2020) .

[87] LHCb collaboration, *Study of the doubly charmed tetraquark $T_{cc}^+$*, Nature Communications **13** (2022) .

[88] Belle Collaboration, *Observation of two charged bottomoniumlike resonances in $\Upsilon(5s)$ decays*, Phys. Rev. Lett. **108** (2012) .

[89] LHCb collaboration, *Evidence of a $J/\psi\Lambda$ structure and observation of excited $\Xi^-$ states in the $\Xi_b^- \to J/\psi\Lambda K^-$ decay*, Science Bulletin **66** (2021) , arXiv:2012.10380.

[90] LHCb Collaboration, *Observation of a $J/\psi\Lambda$ Resonance Consistent with a Strange Pentaquark Candidate in $B^- \to J/\psi\Lambda\overline{p}$ Decays*, Phys. Rev. Lett. **131** (2023) .

[91] LHCb Collaboration, *Observation of $B_{(s)}^0 \to J/\psi p\overline{p}$ Decays and Precision Measurements of the $B_{(s)}^0$ Masses*, Phys. Rev. Lett. **122** .

[92] LHCb collaboration, *Evidence for a new structure in the $J/\psi p$ and $J/\psi\overline{p}$ systems in $B_s^0 \to J/\psi p\overline{p}$ decays*, Phys. Rev. Lett. **128** (2022) , arXiv:2108.04720.

[93] Z. Guo and J. A. Oller, *Anatomy of the newly observed hidden-charm pentaquark states: $P_c(4312)$, $P_c(4440)$ and $P_c(4457)$*, Phys. Lett. B **793** (2019) , arXiv:1904.00851.

[94] C. W. Xiao, J. Nieves, and E. Oset, *Heavy quark spin symmetric molecular states from $\bar{D}^{(*)}\Sigma_c^{(*)}$ and other coupled channels in the light of the recent LHCb pentaquarks*, Phys. Rev. D **100** (2019), no. 1 , arXiv:1904.01296.

[95] X. Weng, X. Chen, W. Deng, and S. Zhu, *Hidden-charm pentaquarks and $P_c$ states*, Phys. Rev. D **100** (2019), no. 1 , arXiv:1904.09891.

[96] M. B. Voloshin, *Some decay properties of hidden-charm pentaquarks as baryon-meson molecules*, Phys. Rev. D **100** (2019) , arXiv:1907.01476.

[97] Y. Lin and B. Zou, *Strong decays of the latest LHCb pentaquark candidates in hadronic molecule pictures*, Phys. Rev. D **100** (2019) , arXiv:1908.05309.

[98] A. N. Semenova, V. V. Anisovich, and A. V. Sarantsev, *New narrow LHCb pentaquarks as lowest antiquark-diquark-diquark systems*, Eur. Phys. J. A **56** (2020) , arXiv:1911.11994.

[99] C. Hua-Xing, *Decay properties of $P_c$ states through the Fierz rearrangement*, The European Physical Journal. C, Particles and Fields. **80** (2020), 14346044.

[100] Y. Dong, P. Shen, F. Huang, and Z. Zhang, *Selected strong decays of pentaquark State $P_c(4312)$ in a chiral constituent quark model*, Eur. Phys. J. C **80** (2020) , arXiv:2002.08051.

[101] Z. Wang, H. Wang, and Q. Xin, *Hadronic coupling constants of the lowest hidden-charm pentaquark state, using qcd sum rules with rigorous quark-hadron duality*, Chinese Physics C **45** (2021) .

[102] C. W. Xiao, J. X. Lu, J. J. Wu, and L. S. Geng, *How to reveal the nature of three or more pentaquark states*, Physical Review D **102** (2020) .

[103] BABAR collaboration, *Search for the $Z(4430)^-$ at BABAR*, Phys. Rev. D **79** (2009) .

[104] D. J. Griffiths and D. F. Schroeter, *Introduction to Quantum Mechanics*, Cambridge University Press, 2018. 9781107189638.

[105] F. W. J. Olver *et al.*, *NIST Digital Library of Mathematical Functions*, http://dlmf.nist.gov/, 2022.

[106] K. Cranmer, *Practical Statistics for the LHC*, 2015. arXiv:1503.07622.

[107] E. Gross, *LHC Statistics for Pedestrians*, 2008. 10.5170/CERN-2008-001.205.

[108] A. Birnbaum, *On the foundations of statistical inference*, Journal of the American Statistical Association **57** (1962), 01621459.

[109] E. Gross, *Practical Statistics for High Energy Physics*, CERN Yellow Rep. School Proc. **3** (2018) .

[110] L. Lista, *Practical statistics for particle physicists*, CERN Yellow Reports: School Proceedings (2017) .

[111] A. L. Read, *Presentation of search results: The $CL_s$ technique*, J. Phys. **G28** (2002) .

[112] M. Pivk and F. R. Le Diberder, *sPlot: A statistical tool to unfold data distributions*, Nucl. Instrum. Meth. **A555** (2005) , arXiv:physics/0402083.

[113] B. Aubert *et al.*, *Measurement of the branching ratios $\Gamma(D_s^{*+} \to D_s^+\pi^0)/\Gamma(D_s^{*+} \to D_s^+\gamma)$ and $\Gamma(D^{*0} \to D^0\pi^0)/\Gamma(D^{*0} \to D^0\gamma)$*, Physical Review D **72** (2005) .

[114] 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. **214** (2017) , arXiv:1612.07489.

[115] A. Ryd and D. Lange, *EvtGen A Monte Carlo Generator for B-Physics*, 2004.

[116] V. V. Gligorov and M. Williams, *Efficient, reliable and fast high-level triggering using a bonsai boosted decision tree*, JINST **8** (2013) , arXiv:1210.6861.

[117] M. Needham, *Clone Track Identification using the Kullback-Liebler Distance*, tech. rep., CERN, 2008. LHCb-2008-002. CERN-LHCb-2008-002. LPHE-2008-002.

[118] R. Hierk, M. Merk, M. Needham, and R. Van der Eijk, *Performance of the LHCb 00 track fitting software*, tech. rep., CERN, 2000. LHCb-2000-086.

[119] E. Bos *et al.*, *The Trajectory Model for Track Fitting and Alignment*, tech. rep., CERN, 2007. LHCb-2007-008. CERN-LHCb-2007-008.

[120] M. De Cian, S. Farry, P. Seyfert, and S. Stahl, *Fast neural-net based fake track rejection in the LHCb reconstruction*, LHCb-PUB-2017-011.

[121] *LoKi::KalmanFilter in LHCb Software Documentation*, https://lhcbdoc.web.cern.ch.

[122] *LoKi::Geometry in LHCb Software Documentation*, `https://lhcbdoc.web.cern.ch`.

[123] LHCb RICH Group, M. Adinolfi *et al.*, *Performance of the LHCb RICH detector at the LHC*, Eur. Phys. J. **C73** (2013) , `arXiv:1211.6759`.

[124] *PropertimeFitter in LHCb Software Documentation*, `https://lhcbdoc.web.cern.ch`.

[125] T. Sjöstrand, S. Mrenna, and P. Skands, *PYTHIA 6.4 physics and manual*, JHEP **05** (2006) , `arXiv:hep-ph/0603175`.

[126] A. Rogozhnikov, *hep_ml repository*, `https://github.com/arogozhnikov/hep_ml`.

[127] A. Rogozhnikov, *Reweighting with Boosted Decision Trees*, J. Phys. Conf. Ser. **762** (2016) , `arXiv:1608.05806`.

[128] H. Voss, A. Hoecker, J. Stelzer, and F. Tegenfeldt, *TMVA - Toolkit for Multivariate Data Analysis with ROOT*, PoS **ACAT** (2007) .

[129] A. Morris, S. Neubert, M. Sarpis, and M. Stahl, *Measurement of the relative branching fraction of $\Lambda_b^0 \to \Lambda_c^+ \overline{D}^0 K^-$ and $\Lambda_b^0 \to \Lambda_c^+ D_s^{*-}$ decays*, LHCb-ANA-2023-003.

[130] W. D. Hulsbergen, *Decay chain fitting with a Kalman filter*, Nucl. Instrum. Meth. **A552** (2005) , `arXiv:physics/0503191`.

[131] A. D. Bukin, *Fitting function for asymmetric peaks*, `arXiv:0711.4449`.

[132] W. E. Johns, *Measurements of the semileptonic decay of the neutral charmed meson $d^0 \to k^- \mu^+$ muon-neutrino*, 1995. 10.2172/1423664.

[133] Particle Data Group, *Review of particle physics*, Prog. Theor. Exp. Phys. **2020** (2020) .

[134] F. Mölder *et al.*, *Sustainable data analysis with snakemake*, F1000Research **10** (2021) .

[135] M. D. Wilkinson *et al.*, *The FAIR Guiding Principles for scientific data management and stewardship*, Scientific data **3** (2016).

[136] R. Bailey, C. Benvenuti, S. Myers, and D. Treille, *The LEP collider*, Comptes Rendus Physique **3** (2002), no. 9 1107.

[137] I. Agapov *et al.*, *Future Circular Lepton Collider FCC-ee: Overview and Status*, 2022. arXiv:2203.08310.

[138] V. Poireau, *A selection of recent results from the BaBar experiment*, 2012. arXiv:1205.2201.

[139] S. Bethke and A. Wagner, *The JADE experiment at the PETRA $e^+e^-$ collider: history, achievements and revival*, The European Physical Journal H **47** (2022) .

[140] Z. Akopov *et al.*, *Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics*, 2012. arXiv:1205.4667.

[141] *CERN Open Data Policy for the LHC Experiments*, tech. rep., CERN, 2020. 10.17181/CERN.QXNK.8L2G.

[142] European Strategy Group, *2020 Update of the European Strategy for Particle Physics*, 2020. 10.17181/ESU2020.

[143] L. Anderlini *et al.*, *The PIDCalib package*, LHCb-PUB-2016-021.

[144] A. Tripathee *et al.*, *Jet substructure studies with CMS open data*, Physical Review D **96** (2017) .

[145] E. A. Sindrilaru, A. J. Peters, G. M. Adde, and D. Duellmann, *EOS developments*, J. Phys. : Conf. Ser. **898** (2017) .

[146] S. de Witt and A. Lahiff, *Quantifying XRootD Scalability and Overheads*, Journal of Physics: Conference Series **513** (2014) .

[147] M. Ferrillo, *New generation offline software for the LHCb upgrade I*, 2022. Poster-2021-1056.

[148] K. Lassila-Perini, C. Lange, E. C. Jarrin, and M. Bellis, *Using CMS open data in research – challenges and directions*, EPJ Web of Conferences **251** (2021) .

[149] A. Larkoski *et al.*, *Exposing the QCD splitting function with CMS open data*, Physical Review Letters **119** (2017) .

[150] C. Cesarotti *et al.*, *Searching in CMS open data for dimuon resonances with substantial transverse momentum*, Physical Review D **100** (2019) .

[151] C. G. Lester and M. Schott, *Testing non-standard sources of parity violation in jets at the LHC, trialled with CMS open data*, Journal of High Energy Physics (2019) .

[152] H. An, Z. Hu, Z. Liu, and D. Yang, *Exploring Uncharted Soft Displaced Vertices in Open Data*, 2021. arXiv:2107.11405.