Utrecht University
Department of Physics and Astronomy

Master Thesis

Charm baryon production at central rapidity in proton-proton collisions at centre-of-mass energy of $\sqrt{s} = 13$ TeV

Author
Floris M. Jonkman
3943259

Supervisors
Dr. Panos Christakoglou
Dr. Davide Caffarri

January 20, 2020

Abstract

By colliding heavy-ions accelerated close to the speed of light the building blocks of ordinary matter, quarks and gluons, are not in their usual state by being confined inside hadrons. Instead they form a hot and dense medium called the Quark-Gluon Plasma. This primordial medium is of fundamental importance as it existed a few microseconds after the Big Bang. Up until now, proton-proton collisions have been used as reference for lead-lead collisions, since it was not believed that small systems could create the necessary conditions of a Quark-Gluon Plasma. However, recent experimental results in high multiplicity proton-proton collisions resemble the ones that are usually attributed to the creation of a Quark-Gluon Plasma in heavy-ion collisions. These observations led to the question whether the underlying physical origin between these experimental results is the same across all collision systems.

This project measured the p_{T}-differential corrected yield of prompt Λ_c^+ charmed baryons with the ALICE detector at the Large Hadron Collider (LHC) in 1.5 billion proton-proton collisions at $\sqrt{s} = 13$ TeV at midrapidity. Charm quarks are produced from initial hard scattering processes that can be calculated by the theory of the strong interaction, Quantum Chromodynamics. Therefore, the charm quark can test calculations made with one of the three pillars of the Standard Model of Particle Physics. The Λ_c^+ is reconstructed in the hadronic decay mode $\Lambda_c^+ \to pK_S^0$. Due to its low production rate and hard to reconstruct decay topology, a sophisticated machine learning technique has been used to extract a signal from the dominating background. This study lays grounds to perform the measurement as function of event activity, allowing to make a first step in shedding light whether a Quark-Gluon Plasma is also created in collision between two partons.

The measured values of the p_{T}-differential corrected yield of prompt Λ_c^+ baryons are presented and are in agreement with other hadronic decay mode $\Lambda_c^+ \to pK\pi$ of the same system and energy. The results are compared with the expectations obtained from Monte Carlo event generators SOFTQCD and HARDQCD which did not reproduce data. The measured values of the Λ_c^+/D^0 ratio, which is sensitive to the c-quark hadronisation mechanism, and in particular to the production of baryons, are presented and compared with Monte Carlo tune MODE2 which is in agreement with data.

Contents

Abstract **2**

1 Theory **5**

1.1 The Quark-Gluon Plasma . . . 5

1.2 Heavy-ion collisions as a tool for the QGP . . . 6

1.2.1 Collective effects . . . 7

1.2.2 Heavy flavour probes . . . 7

1.3 Heavy flavour measurements in proton-proton collisions . . . 8

2 Experimental Setup **11**

2.1 The Large Hadron Collider . . . 11

2.2 A Large Ion Collider Experiment . . . 12

2.2.1 Inner Tracking System . . . 13

2.2.2 Time Projection Chamber . . . 13

2.2.3 Time-of-Flight . . . 14

2.2.4 VZERO . . . 15

3 Analysis **16**

3.1 Data and Monte Carlo Sample . . . 16

3.2 Multiplicity calibration . . . 17

3.3 Reconstruction . . . 19

3.4 Cut strategies . . . 20

3.4.1 Topological cuts . . . 20

3.4.2 Particle identification . . . 21

3.5 Machine learning optimization . . . 21

3.6 Decision tree . . . 22

3.6.1 Boosted Decision Tree . . . 23

3.6.2 Training the BDT . . . 23

3.7 BDT Configuration . . . 24

3.7.1 BDT Performance . . . 25

3.8 Signal extraction . . . 28

3.8.1 Finding optimal BDT response . . . 28

3.8.2 Fit results . . . 30

3.9 Corrections and Efficiencies . . . 33

3.10 Systematic Uncertainties . . . 35

3.10.1 Yield extraction . . . 35

3.10.2 Cut variation 36
3.10.3 Particle identification 36
3.10.4 Monte Carlo transverse-momentum shape 37
3.10.5 Feed-down subtraction 37
3.10.6 Tracking efficiency 37
3.10.7 Total systematic uncertainties 39

4 Results **40**

5 Conclusion and Outlook **45**

Acknowledgements **46**

Chapter 1

Theory

The search for the question what our universe is made of and how its constituents interact with each other has lead us to the Standard Model of Particle Physics. The Standard Model is a relativistic quantum field theory which describes three of the four know fundamental forces (electromagnetic, weak and strong force) in the universe, together with classifying the known elementary particles. The Standard Model currently gives the best description about the building blocks of our universe, consequently scientists never found significant deviation to it. Quantum Chromodynamics, the theory describing the strong interaction in the Standard Model predicts a medium, which can be created under extreme conditions, in which particles do not behave as they would in ordinary matter. This medium is known as the Quark-Gluon Plasma. This chapter gives a theoretical background about this medium, why this medium is interesting for the future of physics and why the study of heavy flavour particles such as the Λ_c baryon in proton-proton collisions can contribute to this.

1.1 The Quark-Gluon Plasma

One of the fundamental questions that one could try to answer is what the states of matter are under conditions of extreme density and temperature. Calculations using QCD on the lattice [2] predict that at sufficiently high temperatures and energy densities a phase transition occurs to a new state of matter in which the quarks and gluons are deconfined thus can move trough the medium without being bound to their hadronic structure. This medium is called the Quark-Gluon Plasma (QGP).

One way to intuitively try to understand the states of Quantum chromodynamics (QCD) matter is by looking at the phase diagram of QCD, an example of this is given in Figure 1.1. This figure shows the QCD phases as a function of temperature and baryon chemical potential. The baryon chemical potential measures the imbalance between baryons and antibaryons or in other words the energy that needs to be spend to add a unit of baryon number in a system. At around the proton mass one encounters the normal nuclear matter. Increasing the baryon density we get to the ultra-dense neutrons stars, which are predicted to have a QGP in their core. On the other hand, systems with extremely low values of μ_B but with large temperatures ($T > 155$ MeV) are also expected to consist of a Quark-Gluon Plasma. The high temperature and low chemical potential can be probed experimentally using heavy-ion collision at the Large Hadron Collider (LHC) and Relativistic Heavy Ion Collider (RHIC). One of the reasons that makes the study of the QGP of fundamental importance is that a few microseconds after the Big Bang

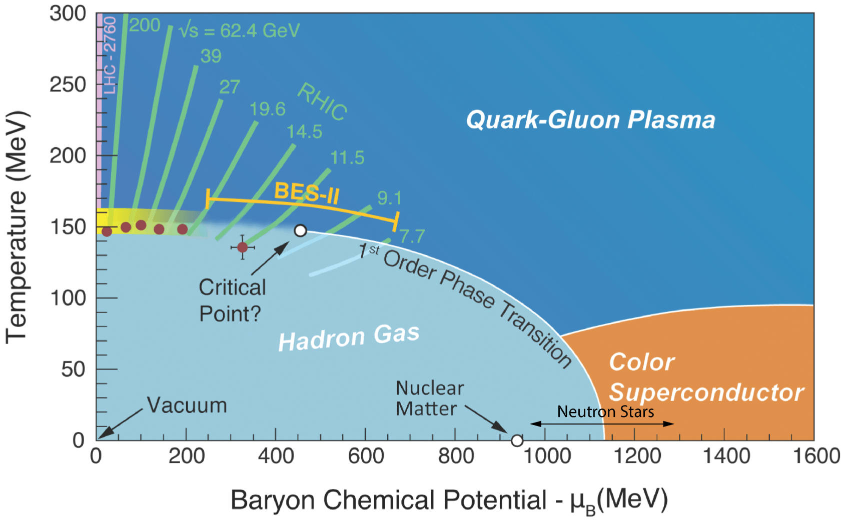


Figure 1.1: Illustration of the QCD phase diagram calculated by lattice QCD calculations [1].

an extremely hot QGP is expected to have existed.

1.2 Heavy-ion collisions as a tool for the QGP

In order to investigate experimentally the strong phase transition and understand the properties of the QGP, the necessary extreme conditions predicted by lattice QCD calculations need to be recreated in the laboratory. At the LHC this is done by colliding lead ions at large energies. These ions that have a high number of nucleons are accelerated close to the speed of light, consequently the usually round objects are Lorentz contracted to flat discs. Eventually the lead ions collide into one another resulting in a region with a very high energy density because of their speed and compact volume.

Describing the evolution of the collision and the QGP has been illustrated in Figure 1.2, and marks the time of the collision as $t = 0$. The phase right after the collision is known as the pre-equilibrium phase. In this phase partons with large momentum and mass (beauty and charm) are produced. On the time scale of the full evolution this phase is smaller than the formation of the QGP ($t < \tau_0$), and is in the order of less than 1.0 fm/c. The hot and dense matter consequently thermalizes at ($t = \tau_0$), the point where the QGP is formed. Partons move through this dense and hot medium and lose energy by emitting gluons or interacting with other partons. The temperature in the medium creates thermal pressure, which causes the QGP to expand and boosts the particles in the medium. This effect, which is felt collectively by all the constituents of the medium, is called radial flow. As a consequence of the expansion the temperature drops rapidly, until it reaches a critical value indicated as T_C after which the chemical freeze-out will start. This happens approximately 10 fm/c after which the QGP starts to hadronize.

After this moment, partons will be confined inside hadrons in a process characterized as hadronisation. From that point quarks cannot move as quasi-free particles and must be bound with other quarks to form a colour singlet state such as a two quarks system called mesons or three quarks system called baryons. Eventually, when the chemical freeze-out temperature T_{ch} has been reached the QGP has hadronized into hadron gas. The hadrons are now confined and the only form of energy transfer is due to (in)ellastic collisions. Lastly, the expansion of the hadron gas reaches the kinetic free out temperature T_{fo} and the

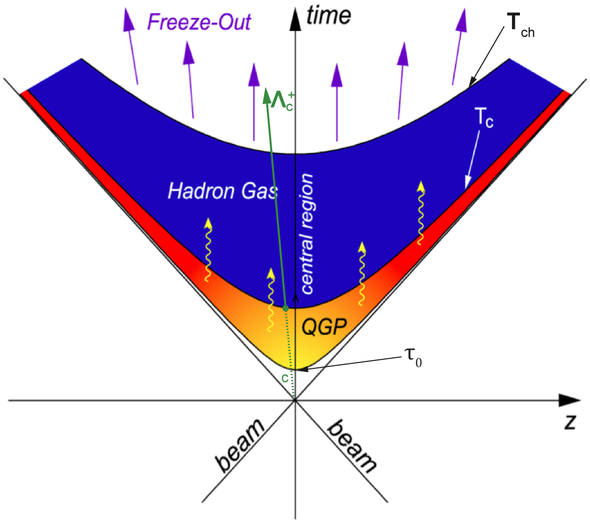


Figure 1.2: Space-time evolution of the Quark-Gluon Plasma [3]. The c-quark is produced in the pre-equilibrium phase ($t < \tau_0$) and experience the full evolution of the Quark-Gluon Plasma.

only method of energy transfer can be by decay. Eventually the particles or their decay products will be measured by the experimental apparatus.

1.2.1 Collective effects

In the presence of a hot and dense deconfined medium that expands explosively, the momentum spectra of the constituents are modified accordingly. In particular, the partons feel a boost which is characterised by a common velocity field (usually called radial flow) which acts additively to their thermal motion.

The effect of radial flow causes the momentum distribution of a particle to become harder, shifted more to higher momentum, as the relativistic momentum is defined by $p = m\gamma v$. Since, the momentum is proportional to the mass of the particle, particles with a higher mass get a harder momentum distribution. The effect of this is clearly visible in Figure 1.3. It is seen that the spectra become becomes harder in more central collision (lower percentages) as we expect a larger QGP droplet to be formed, as it has a larger interaction zone. This effect is even more pronounced for heavy particles (protons) compared to lighter ones (pions), due to their mass.

1.2.2 Heavy flavour probes

Quarks are a very good probe of the QGP since they hold color charge and can thus interact with the constituents of the medium. Their production time is inverse proportional to the mass of the quarks. This categorizes the resulting produced particles in two categories: 'light flavour' and 'heavy flavour' particles. Light flavour up, down and strange quarks predominantly give rise to light flavour hadrons such as pions, kaons and lambda's and can be formed during the QGP phase. These light flavour particles measure global properties of the QGP by testing the hydrodynamic expansion of the medium based by models that attempt to describe its evolution. On the other hand, 'heavy flavour' beauty and charm quarks form eventually heavy flavour hadrons such as the Λ_c-baryon and D^0 meson and can only be formed at the very early stages of a heavy-ion collision thus experience the full evolution of the medium, see the heavy flavour c-quark in the illustration of Figure 1.2. The heavy c-quark is predominantly

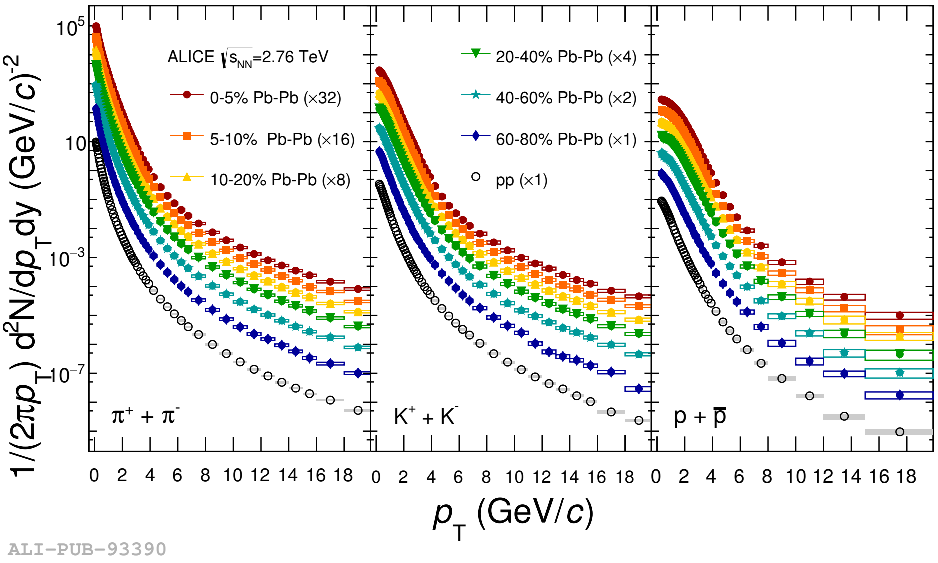


Figure 1.3: Transverse momentum spectra of charged pions (left panel), kaons (middle panel), and (anti)protons (right panel) measured in lead-lead and proton-proton collisions. The spectra have been scaled by the factors listed in the legend for clarity [4].

produced in the pre-equilibrium phase, after which it passes through the entire medium interacting with its constituents. Eventually it hadronizes into charm hadrons. One such case is the Λ_c^+ baryon, the main focus of this project, of which the decay products can be detected.

The study of heavy flavour quarks measurement also provides unique insight into the hadronisation process, which is believed to take place through two mechanism: recombination and fragmentation. This is another effect next to radial flow that affects the momentum distribution of particles. How these two mechanisms affect the transverse momentum (p_T) spectrum can be seen in Figure 1.4. Fragmentation is dominant for high momentum particles and occurs when partons fragment by breaking their energy in smaller 'pieces', consequently producing lower momentum particles. Recombination is thought to be the dominant process of low to intermediate (e.g. between 2 and 8 GeV/c) momentum particles and usually involves the combination of low momentum partons which leads to the formation of a higher momentum meson and baryon consisting of a combination of two- and three-quarks, respectively.

Lastly, heavy flavour quarks can be used to test perturbative QCD calculation. Heavy flavour quarks are produced mainly via hard parton scattering, a process occuring via strong interaction and involving large momentum transfers compared to the QCD scale $Q \gg \Lambda_\mathrm{QCD}$.

1.3 Heavy flavour measurements in proton-proton collisions

The study of heavy flavour particles in proton-proton collisions has always been used a as reference to test perturbative QCD calculations, as naively one does not expect a QGP to be formed in collisions of such small systems. However in recent studies clues for similar QGP medium effects have been observed in proton-proton collisions, which has a lower multiplicity (number of produced particles) compared to lead-lead collisions [6].

A recent study examined the multiplicity dependent production of light flavour particles in proton-proton

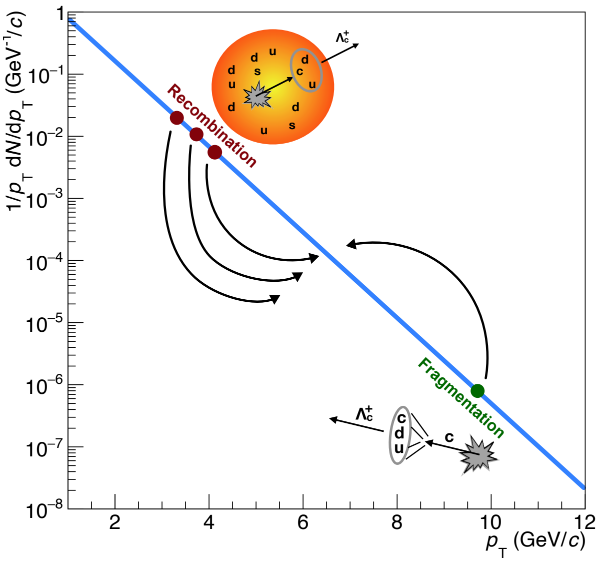


Figure 1.4: Illustration of effect of recombination and fragmentation on the transvere momentum distribution (Illustrations from [5]).

collisions at $\sqrt{s} = 7$ TeV [7], and compared transverse momentum with spectra obtained in larger systems such as proton-lead and lead-lead. First they examined the particle production as function of transverse momentum for different multiplicities in the proton-proton systems for different baryons and mesons. In all cases, a hardening of the transverse momentum is observed with increasing multiplicity. The same study reported the multiplicity dependent comparison of the baryon-over-meson ratio versus transverse momentum for the two lightest strange particles of its kind Λ/K^0_S, see the top graph of Figure 1.5. The ratio shows for all systems a depletion at low momentum and an enhancement at intermediate values. The lower panels of Figure 1.5 shows the same Λ/K^0_s ratio, but as a function of multiplicity. This ratio is reported at three characteristic values of transverse momentum, low (left), intermediate (middle) and high p_{T}. The plots indicate that the ratio scales as a function of multiplicity in all three collision systems which could hint at a common underlying physics mechanism.

The light flavour baryon-over-meson used to be an observable to illustrate the radial flow and recombination effects in lead-lead collisions, but as qualitatively similar effects have been observed in high multiplicity proton-proton, proton-lead and lead-lead collisions the question could be raised whether in high multiplicity proton-proton collisions a QGP droplet is formed. Are the effects for the different systems coming from the same underlying physics?

The momentum distribution of Λ_c as function of multiplicity in order to form the heavy flavour baryon-over-meson ratio Λ_c/D^0 will allow us to test if the observed similarities in the light flavour baryon-over-meson ration can be extended in the heavy flavour sector. This project makes a step forward in this direction by focusing on the production of heavy flavour Λ_c baryon. Finally, the study of the heavy flavour particles can be used to test perturbative QCD calculations.

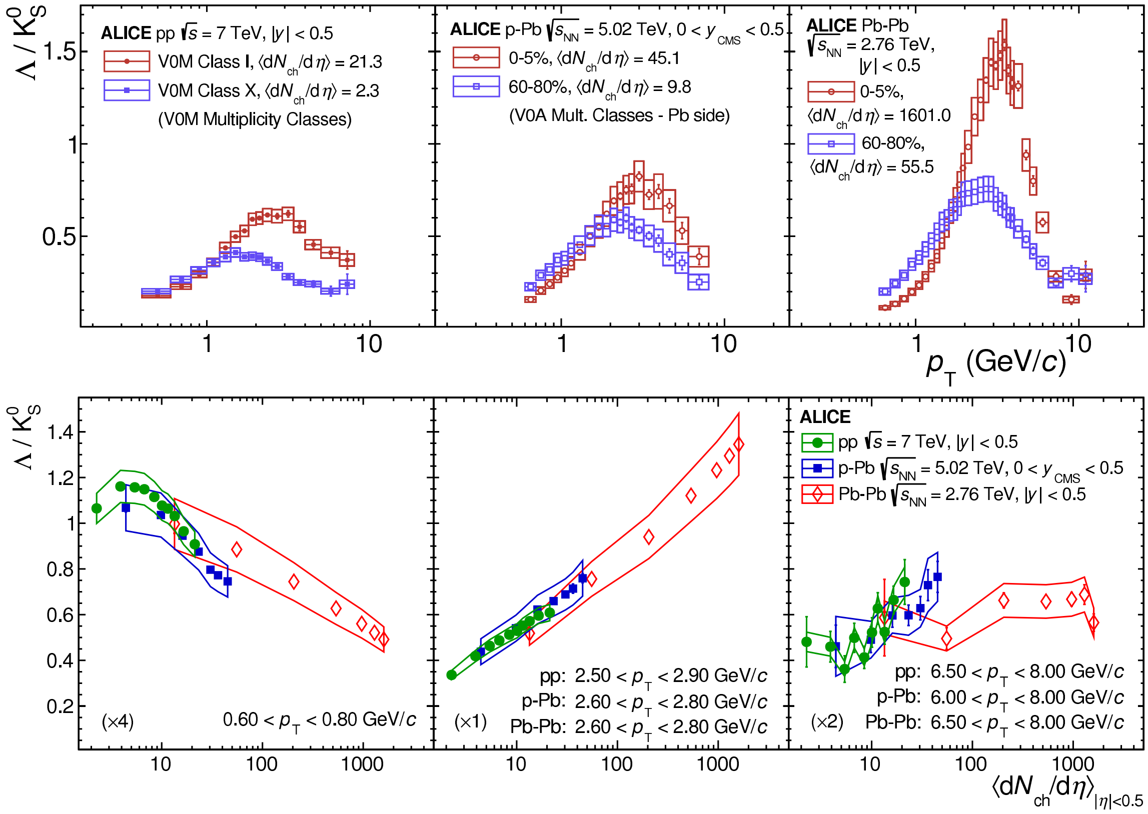


Figure 1.5: Top: Transverse momentum dependence of Λ/K_S^0 ratio in proton-proton, proton-lead and lead-lead collisions for high- (red) and low-multiplicity (blue) classes. Bottom: Multiplicity dependence of Λ/K_S^0 ratio in proton-proton, proton-lead and lead-lead collisions at low (left), intermediate (middle) and high (right) transverse momentum. [7].

Chapter 2

Experimental Setup

Theoretical particle physicists came up with admirable ideas how the nature of universe acted. These theories remain merely ideas if they are not confronted with experimental measurements. In many cases scientists need an experimental setup larger than any other man made structure ever built. This is why in 1954 the *Conseil Européen pour la Recherche Nucléaire* (CERN) was founded as an international organization that combines the intellectual power and resources to increase our scientific achievements in particle physics. CERN helped to prove new physics with the use of collider experiments and have made incredible discoveries thus far. CERN holds the Large Hadron Collider (LHC) which is the world's largest and most powerful particle collider and largest machine in the world.

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is a circular particle accelerator built in a 27 kilometers long tunnel located 100 meters under the ground at the French-Swiss border. The circular ring contains two beam pipes incased with super conducting magnets in order to keep the high energy particles in orbit. The LHC will be able to reach a centre-of-mass energy of $\sqrt{s} = 14$ TeV for proton-proton collisions and $\sqrt{s_{NN}} = 5.5$ TeV for lead-lead collisions in the upcoming runs that start in 2021 . The left of Figure 2.2 gives a schematic overview of the CERN complex and the LHC. For proton-proton collisions hydrogen atoms are accelerated up to 50 MeV in a linear accelerator (LINAC2) and are injected in the Proton Synchrotron Booster (PSB) which accelerates them to an energy of 1.5 GeV. Subsequently, the Proton Synchrotron (PS) brings the energy up to 25 GeV. Finally the Super Proton Synchrotron (SPS) brings it up to 450 GeV before being injected in the LHC. The LHC eventually accelerated the protons to an energy of 13 TeV. The protons travel in bunches of 1.15×10^{11} particles through the 27 kilometers long tunnel separated by 25 nanoseconds from each other. The protons at last have four points where they can collide, ATLAS and CMS, which mainly focus on research on the Higgs boson, LHCb, which does measurements on CP-violation and ALICE, designed for heavy-ion collisions. This analysis used data collected the ALICE collaboration, hence only this detector will be described.

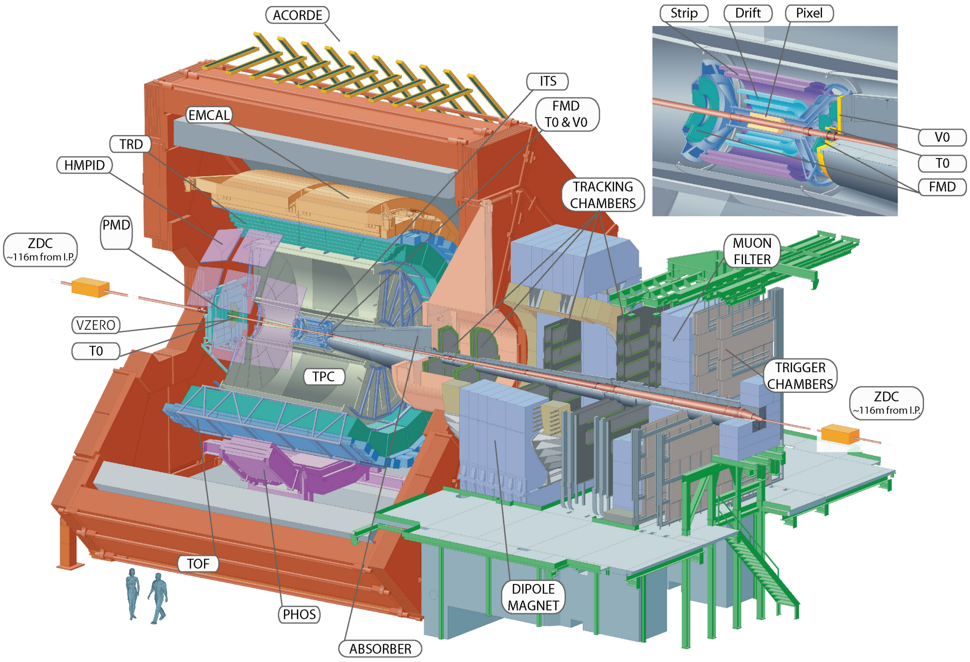


Figure 2.1: Schematic cross section of the ALICE detector [8].

2.2 A Large Ion Collider Experiment

A Large Ion Collider Experiment also known as ALICE is a 16 meters tall, 16 meters wide and 26 meters long detector and weights approximately 10.000 tons. An schematic overview of the ALICE detector is shown in Figure 2.1. The experiment consists of 18 detectors, each with its own technology to measure trajectory, mass, charge, velocity and energy of the particles that transverse their sensitive areas. The central part of ALICE detects hardrons, electrons, and photons while the forward part consists of a muon spectrometer. By definition the beam is aligned with the z-axis. The particle kinematics are expressed in terms of the transverse momentum p_{T}, the magnitude of the projection of the three momentum $\vec{p}$ in the xy-plane. The azimuthal angle ϕ lays in the xy-plane, and the polar angle θ perpendicular to this in the yz-plane. The central part of the detector covers a polar region of $\theta = 45°$ to $\theta = 135°$. Particle physicist rather talk about the pseudorapidity, which is defined as

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right) = \frac{1}{2}\ln\left(\frac{|\vec{p}| + p_{\mathrm{L}}}{|\vec{p}| - p_{\mathrm{L}}}\right), \tag{2.1}$$

with p_{L} the component of $\vec{p}$ along the beam z-axis. The ALICE detector consequently covers a pseudorapidity range of $|\eta| < 0.9$. From the inside out, the detector consists of a Inner Tracking System (ITS), a Time-Projection Chamber (TPC), Transition Radiation detectors (TRD), three particle identification arrays of Time-of-Flight (TOF), Ring Imaging Cherenkov (HMPID) and two electromagnetic calorimeters (PHOS and EMCal). Only HMPID, PHOS and EMCal do not cover the full azimthal angle. The ITS consists of three subdetectors, the Silicon Pixel Detector (SPD), the Silicon Drift Detector (SDD) and the Silicon Strip Detector (SSD). The forward part, which covers a small 2° ($\eta = 4.0$) to 9° ($\eta = 2.5$) angle consists of absorbers, a large dipole magnet and fourteen planes of tracking and triggering chambers. At low angles smaller detectors ZDC, PMD, FMD, T0 and VZERO are located. These detectors are used for event triggering and characterization. A full report about the detector can be found in [9], this section will only describe the most important detectors used for this analysis.

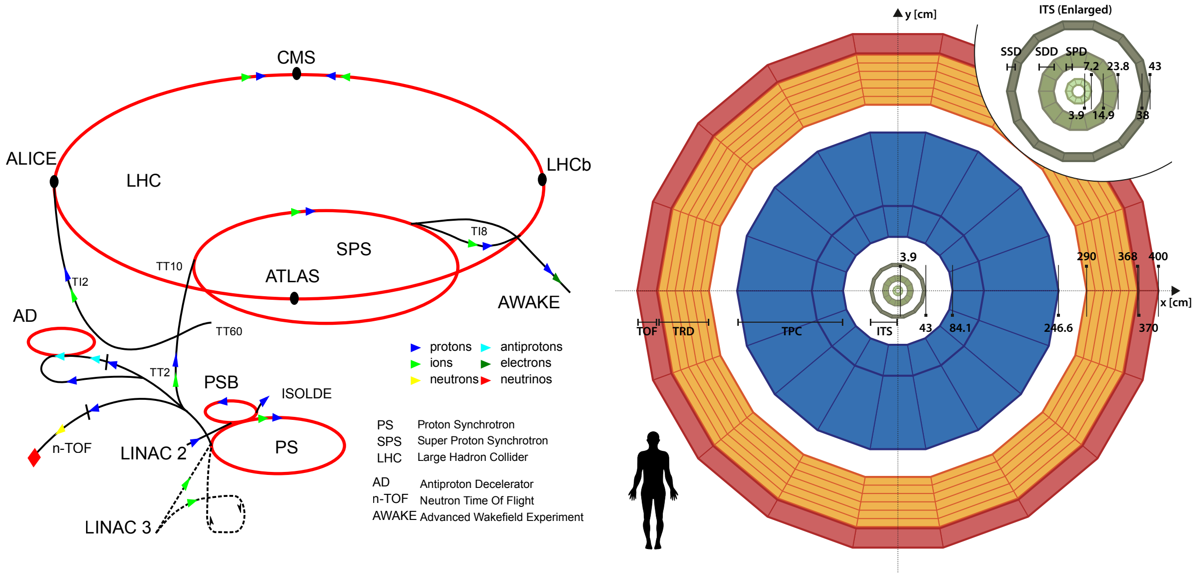


Figure 2.2: Left: Schematic overview of the CERN complex and the LHC. Right: Schematic view of the projection the inner part of the ALICE detector on the xy-plane. The units correspond to the begin and end distrance of the corresponding detector from the beam axis in centimeters.

2.2.1 Inner Tracking System

The Inner Tracking System (ITS, Figure 2.2 right side), is the most central detector of ALICE. It is a six-layer silicon detector positioned closest to the collision point. The main goal of the ITS is to determine the primary collision point (primary vertex) and to reconstruct the trajectories of the particles (tracks). The two most inner layers, the Silicon Pixel Detector (SPD), are positioned at 3.9 and 7.6 centimeters around the beam and it contains 1200 readout chips. The SPD has a key role in determining the position of the primary vertex Z_{vtx} and measuring the impact parameter of the secondary tracks, the distance of the particle to the primary vertex, originating from weak decays of strange, charm and beauty hadrons. The SPD provides also the multiplicity of charged particles produced in the collision. One tracklet is reconstructed by the two hits in the two SPD layers, together with the primary vertex [10]. The two middle layers, the Silicon Drift Detectors (SDD), are positioned at 14.9 and 23.8 centimeters around the beam. The SDD provides the energy loss information and give a excellent spatial resolution [11]. The two outermost layers of the ITS, the Silicon Strip Detector (SSD), are positioned at 38 and 42 centimeters respectively. The SSD is crucial for tracking the particles, and connecting the tracks from the Time Projection Chamber (TPC) to the ITS. It also contributes to the particle identification by measuring the energy loss of the particles [12].

2.2.2 Time Projection Chamber

The Time Projection Chamber (TPC, Figure 2.2 right side) is a 88 cubic meter cylinder filled with a gas mixture of Ne-CO_2 (90%:10%) covering the full azimuthal angle and pseudorapidity range of $|\eta| < 0.9$. The TPC covers a detection radius of 84.1 up to 246.6 centimeters from the collision point and has a length of 500 centimeters. Charged particles that transverse the TPC ionize the gas, this ionization causes electrons to liberate from the gas. These liberated electrons drift towards the two end plates at which an electric field is applied. The drift time combined with the location where the electron hits the end plate,

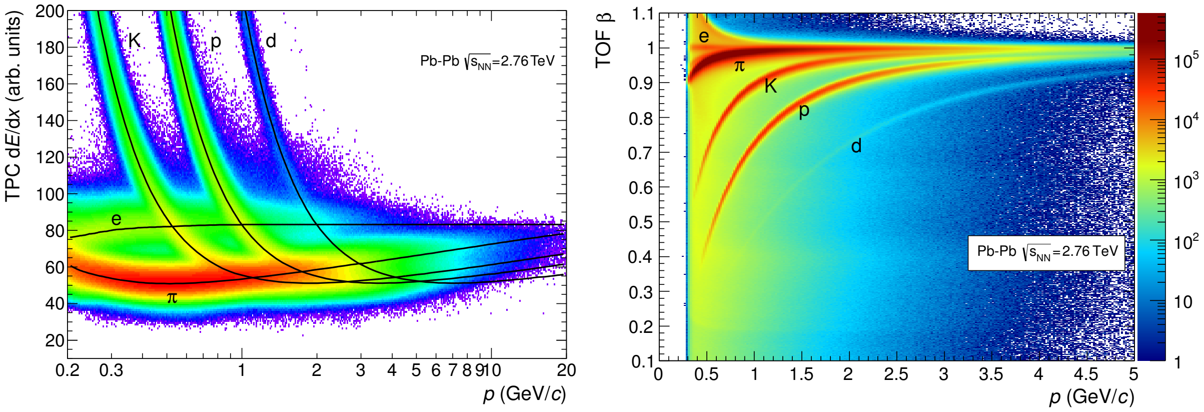


Figure 2.3: Left: Specific energy loss (dE/dx) in the TPC versus particle momentum in Pb-Pb collisions at $\sqrt{s_{NN}} = 2.76$ TeV. The lines show the parametrizations of the expected mean energy loss of the Bethe-Bloch formula [15]. Right: Distribution of $\beta = v/c$ as measured by the TOF detector as a function of momentum for particles reaching the TOF in Pb-Pb interactions [15].

gives a three dimensional trajectory of the particle. The TPC provides next to track finding, momentum measurements and particle identification (PID) at transverse momenta $0.1 < p_{\mathrm{T}} < 100$ GeV/c [13]. As ALICE probes observables using hadronic decay channels, determining the type of hadron of a track enhances the signal. Fortunately, many observables are either mass or flavour dependent. In the TPC particle identification is done by measuring the specific energy loss dE/dx of particles, see the left of Figure 2.3. The Bethe-Bloch formula for different particles is fitted. The detector energy loss, the Bethe-Bloch fit and the resolution of the detector give the commonly used definition of the number of sigma for TPC, $n\sigma_{\mathrm{TPC}}$, as the deviation of the measured energy loss to the expected energy loss of a certain particle expressed in terms of the detector resolution [14].

2.2.3 Time-of-Flight

The Time-of-Flight detector (TOF) covers a detection radius of 370 up to 400 centimeters from the central beam and has a a length of 741 cm. It consists of 18 sectors in the ϕ-direction which are divided in five modules along beam direction. The modules contain a total of 1638 detector elements (MRPC strips), covering an area of 160 m^2. The MRPC is a stack of glass plates filled with tetra-fluoro-ethane ($C_2H_2F_4$). A charged particle transverse the strip ionizes the gas and an electric field amplifies the ionization. This amplified signal is read out. The TOF measures the difference between time difference with a resolution of 80 ps. The particle identification of the TOF detector relies on the comparison between the time of the track from the primary vertex up to the TOF detector, the time in the TOF detector, and the expected time under a given mass hypothesis of the particle. Therefore the TOF detector relies on the tracking, mass hypothesis and momentum of the particle. The commonly $n\sigma_{\mathrm{TOF}}$ is defined as the sum of these three times divided by the total uncertainty of all three [16]. The velocity distribution $\beta = v/c$ measured in TOF dectector as function of momentum p is shown on the right of Figure 2.3. The background is due to TOF mismatched tracks, where the reconstructed TPC track has been matched with an incorrect TOF hit.

2.2.4 VZERO

The VZERO detector is compossed of two arrays, VZERO-A and VZERO-C, located at opposite sides of the interaction point, which cover a pseudorapidity range of $-3.7 < \eta < 1.7$ and $2.8 < \eta < 5.1$ respectively. VZERO-A is located 330 and VZERO-C 90 centimeters from the interaction point. Each array is segmented into four rings, and each ring is divided in to eight segments. The segments are made of a plastic scintillator, which radiate low-energy photons when transversing it. This radiations is converted into current and amplified. The current is proportional to the number of charged particles transversing the VZERO segment [17]. The VZERO is used for triggering and centrality and multiplicity determination, as well as for event plane determination.

Chapter 3

Analysis

In this chapter the details of Λ_c production in pp collisions at $\sqrt{s} = 13$ TeV will be presented. This data sample is large enough to allow for the extraction of the yield of Λ_c particles as a function of the event activity, reflected by the multiplicity. The Λ_c has two main hadronic decay channels, $\Lambda_c^+ \rightarrow pK^-\pi^+$ with branching ratio of $6.28 \pm 0.32\%$ and $\Lambda_c^+ \rightarrow pK_S^0$ with a branching ratio of $1.59 \pm 0.08\%$ [18], where $K_S^0 \rightarrow \pi^+\pi^-$ has a branching ratio of $69.20 \pm 0.05\%$. This project focuses on the later. In what follows, the sample and the selection criteria at the event level will be described, followed by a discussion on how one fully reconstructs the Λ_c baryon from its decay products. The chapter also discusses the innovative technique used to select the Λ_c candidates using a boosted decision tree as well as the corrections that need to be applied in order to get a fully corrected result.

3.1 Data and Monte Carlo Sample

The data used for this analysis consists of 1.5 billion events (collisions), collected during the LHC 2016, 2017 and 2018 run of proton-proton collisions at centre-of-mass energy of $\sqrt{s} = 13$ TeV, corresponding to an integrated luminosity of 27.1 ± 1.4 nb^{-1}. Data are divided into runs. A run is a period of a couple of hours during which data is collected. In between runs the particle beam is dumped and the detector gets recalibrated. An overview of the dataset is shown in Table 3.1. A subset of the data has also been used to train the machine learning optimization model, more about this in Section 3.7. The chosen periods for training were LHC16h, LHC16k, LHC17m and LHC18p. In this analysis we analyze minimum bias events. These events require at least one hit in VZERO-A and VZERO-C. An offline event selection protocol relying on the timing of fast detectors was applied to reject background events coming from the interaction of particles with the beam pipe materials or beam-gas interactions. Only events within interaction vertex of $|Z_{\mathrm{vtx}}| < 10$ cm were selected. A dedicated algorithm to detect multiple interaction vertices based on the tracklets (i.e. pairs of reconstructed hits in the SPD) was used to reduce the pile-up contribution. A pile-up occurs when multiple collisions are stored in the same event. Tracks of the collision are rejected if their corresponding second interaction vertex is found with at least 5 tracklets.

To simulate the particle production event generators are used based on hadronisation models and elementary processes. These models depend on experimental results and are mainly based on the theoretical knowledge about particle physics. There are different particle generators (models) who all accurately try to model the underlying physics. For this analysis Monte Carlo productions are used generated by

Period(s)	Type	System	$\sqrt{s}$ (TeV)	$N_{\text{events}}(\times 10^6)$	Used for	Anchord to
LHC16	Data	pp	13	422	Signal extraction	
LHC17	Data	pp	13	566	Signal extraction	
LHC18	Data	pp	13	494	Signal extraction	
LHC16kh	Data	pp	13	184	Training BDT	
LHC17m	Data	pp	13	89	Training BDT	
LHC18p	Data	pp	13	57	Training BDT	
LHC19h4c2	MC	pp	13	16	Training BDT	LHC16
LHC19h4b2	MC	pp	13	27	Training BDT	LHC17
LHC19h4a2	MC	pp	13	27	Training BDT	LHC18
LHC17h8a	MC	pp	13	48	Efficiencies	LHC16deghjop
LHC18f4a	MC	pp	13	23	Efficiencies	LHC16kl
LHC18l4a	MC	pp	13	99	Efficiencies	LHC17
LHC18l4b	MC	pp	13	93	Efficiencies	LHC18

Table 3.1: Overview of used data and Monte Carlo samples. The number of events N_{events} is the amount after the event selections. The 'Anchord to' column refers to which data production the Monte Carlo sample is linked. For the 'Used for' column refers to the different sections: 'Signal extraction' Section 3.8, 'Training BDT' Section 3.7 and 'Efficiencies' Section 3.9.

PYTHIA [19].

The Monte Carlo productions are used for two types of purposes. First of all, a $\Lambda_c^+ \to K_S^0 p$ dedicated Monte Carlo is used to enhance the signal for our machine learning optimization model. This is needed as the Λ_c^+ production is low in general purpose Monte Carlo, on average $\sim 1/2000$ events. For this reason, a simulation is used where they enhance Λ_c^+ production, on average approximately 1 per 2.3 events. Secondly, Monte Carlo simulations are used to calculate the detector and methodology efficiencies, Section 3.9. A overview of the used Monte Carlo sample is shown in Table 3.1.

3.2 Multiplicity calibration

As stated before, the goal is to study the production of Λ_c as function of multiplicity in pp collisions at $\sqrt{s} = 13$ TeV. Multiplicity is the measurement of the number of particles that have been produced in the collision. The multiplicity can be measured in different detectors in ALICE: SPD, VZERO-A and VZERO-C that all cover different rapidity regions. In this analysis we took the number of SPD tracklets in the interval $|\eta| < 1$, $N_{\text{tracklets}}$, as multiplicity estimator. An SPD tracklet is obtained by joining space points on the two SPD layers, together with the primary vertex. This multiplicity estimator is the same as used in previous studies performed for prompt D-meson production [20]. The uncorrected distribution of the number of tracklets, $N_{\text{tracklets}}$, for the data periods LHC16h, LHC16k, LHC17m and LHC18p are shown in Figure 3.1. For clarity only these periods are shown, but the subsequent calibration procedure is applied on all data periods used in this analysis.

The mean number of tracklets, $N_{\text{tracklets}}$, as function of position of the interaction vertex along the beam line, Z_{vtx}, is shown on the right of Figure 3.1. The two graphs in the figure clearly show that the SPD acceptance decreases over time and is dependent on the position of the vertex. At 40 tracklets the

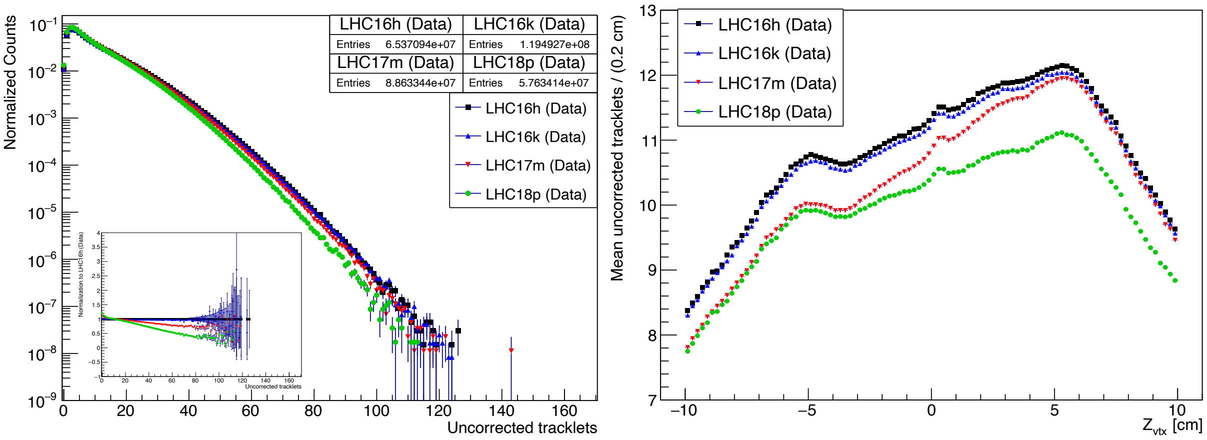


Figure 3.1: Left: Uncorrected number of tracklets from SPD for different data periods, the small figure shows the same normalized to period LHC16h. Right: Mean number of uncorrected number of tracklets from SPD as function of the postion of the interaction vertex for different data periods.

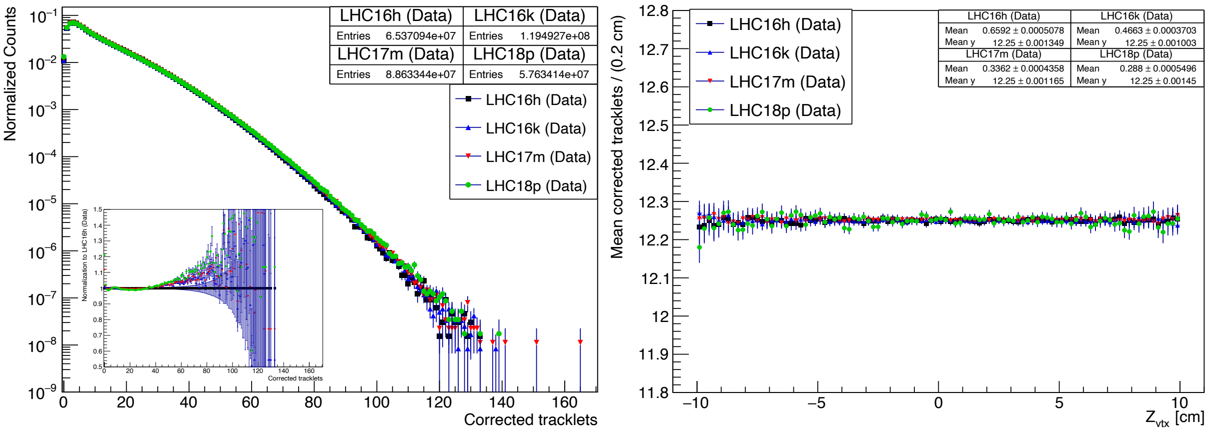


Figure 3.2: Left: corrected number of tracklets from SPD for different data periods, the small figure shows the same normalized to period LHC16h. Right: Mean number of corrected number of tracklets from SPD as function of the position of the interaction vertex for different data periods.

difference in the number of counts between LHC16h and LHC18p is 25%, and at 60 tracklets already 50%. As Λ_c^+ production as function of multiplicity is to be analyzed, the multiplicities for different data samples need to be corrected in order to compare them accordingly. The measured number of tracklets, N_{raw} is corrected event-by-event to equalize the average number of tracklets among all the periods and to correct for the Z_{vtx} dependence. This was done according to the following formula,

$$N_{\text{corr}} = \frac{\langle N_{\text{ref}} \rangle}{\langle N_{\text{period}}(Z_{\text{vtx}}) \rangle} N_{\text{raw}}, \tag{3.1}$$

with N_{corr} the corrected number of tracklets, $\langle N_{\text{ref}} \rangle$ the mean number of tracklets in the reference Z_{vtx} point and $\langle N_{\text{period}}(Z_{\text{vtx}}) \rangle$ the mean number of tracklets for events with vertex Z_{vtx} for a given period. In other words, $\langle N_{\text{ref}} \rangle$ is a reference value taken from the mean number of tracklets of one of the periods, this is the value to which all the periods will be calibrated. By a rule of thumb this is usually chosen as the highest mean number of tracklets at a given Z_{vtx} for all the periods, hence $\langle N_{\text{ref}} \rangle = 12.25$ at $Z_{\text{vtx}} = 5.55$ cm is chosen from LHC16h. In the correction procedure, the Poisson statistics is applied to

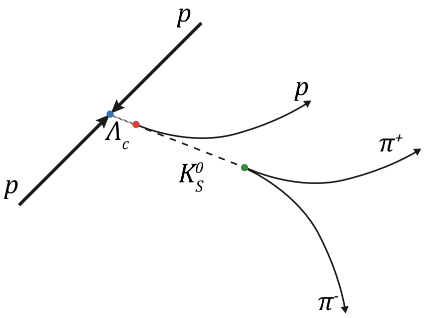


Figure 3.3: Decay topology of Λ_c^+. The distances are not to scale.

get an integer value for N_{corr}.

The corrected distribution of the number of tracklets, N_{corr}, is shown in Figure 3.2. The mean number of tracklets, $N_{\text{tracklets}}$, as a function of position of the interaction vertex along the beam line, Z_{vtx}, is shown on the right. It is seen that up to 40 tracklets, all periods agree within 5%. Above 40, this difference seems to increase with increasing multiplicity. To goal of this project was to analyze four different multiplicity ranges: multiplicity integrated range, $1 \leq N_{\text{tracklets}} \leq 9$, $10 \leq N_{\text{tracklets}} \leq 29$, $30 \leq N_{\text{tracklets}} \leq 59$, but due to the lack of time only the multiplicity integrated range has been analyzed.

3.3 Reconstruction

The hadronic decay channel $\Lambda_c^+ \to pK_S^0$ with branching ratio of $6.28 \pm 0.32\%$ was used for this analysis, see Figure 3.3. The lifetime of the Λ_c is $c\tau = 60$ μm and has a mass of 2286.46 ± 0.14 MeV [18]. The reconstruction of the Λ_c is obtained in two phases. The first phase consists of the K_S^0 reconstruction via its decay topology into two charged pions with branching ratio of $1.59 \pm 0.08\%$ [18]. This decay has a distinctive V-shape because two particles of opposite charged arise from a (invisible) neutral track, see Figure 3.3. Particles with this characteristic decay shape are called V0's. Two other particles fall also into this category, $\Lambda \to p\pi^-$, and its anti-particle. The second phase is combining the reconstructed proton (called bachelor) together with the K_S^0, to form a Λ_c-candidate.

All the used reconstructed bachelor tracks must have been refitted in the ITS and TPC. The tracks must have a minimum of 70 crossed rows in TPC with a minimum of ratio of 0.8 to the number findable clusters and at least one hit in either of the two SPD layers. Only tracks are selected within a pseudorapidity range of $|\eta| < 0.8$ with transverse momentum $p_{\text{T}} > 0.3$ GeV/c. The kink-daughter tracks are rejected, which are tracks where suddenly a slight 'kink' appears in their trajectories without obvious reason.

The V0 (K_S^0, Λ, $\bar{\Lambda}$) candidates were identified by applying selections on the characteristics of their decay tracks and of their weak decay topology. The candidate must have transverse momentum of $p_{\text{T}} > 0.3$ GeV/c, a minimum transverse impact parameter to the primary vertex d_0 of 0.05 cm and a maximum distance of closest approach (DCA) to the secondary vertex between the two daughter tracks of 1.5 cm. The V0 requires a minimum (maximum) transverse decay radius of 0.2 (100) cm and a minimum cosine of the V0 pointing angle of 0.99 [21]. The V0 candidates reconstructed on-the-fly, i.e. during the tracking of particles by the algorithm of the corresponding detector, were rejected. The ones reconstructed by the so-called offline finder, that is invoked after the entire tracking iteration is finished, are the considered candidates.

All the daughter tracks of the K^0_S candidate must have been refitted in the TPC and require a minimum of 70 crossed rows with a minimum of ratio of 0.8 to the number findable clusters in the TPC. Only daughter tracks within a pseudorapidity range of $|\eta| < 0.8$ with a transverse momentum $p_{\mathrm{T}} > 0.1$ GeV/c were selected.

Additionally, a fiducial acceptance cut on Λ_c candidates has been applied,

$$y_{\mathrm{fid}}(p_{\mathrm{T}}(\Lambda_c)) = \begin{cases} -\frac{0.2}{15}p^2_{\mathrm{T}}(\Lambda_c) + \frac{1.9}{15}p^2_{\mathrm{T}}(\Lambda_c) + 0.5, & \text{if } p_{\mathrm{T}}(\Lambda_c) < 5\mathrm{GeV}/c \\ 0.8, & \text{if } p_{\mathrm{T}}(\Lambda_c) \geq 5\mathrm{GeV}/c, \end{cases} \tag{3.2}$$

as the acceptance in rapidity for these Λ_c cuts drops to zero for $|y| > 0.5$ at low p_{T} and $|y| > 0.8$ at $p_{\mathrm{T}} > 5$ GeV/c, as described in [22].

3.4 Cut strategies

The now obtained bachelor tracks, K^0_S candidates, pion daughter tracks, and Λ_c candidates are still not refined enough. The signal of real Λ_c's is neglible to the combinatorial background. This is due to the low production per event and the short lifetime of the Λ_c. This background can be reduced using dedicated cuts, such that they enhance the 'signal'. Previously, merely 'standard' optimized topological cuts were used to obtain this. With contemporary technique's making use of machine learning algorithms for classification improvements can be made in the signal over background ratio. The used technique is in this analysis a Multivariate Analysis (MVA) [23], more about this in Section 3.5. A subgoal of this analysis is to see if this MVA technique can make an improvement in the reconstruction of the Λ_c. Therefore, three different analysis strategies have been setup. The first reference strategy (Standard, 1) uses standard optimized topological cuts together with optimized particle identification (PID), the second (StandardMVA, 2) uses the same topological cut as particle identification as the first but as addition makes use of a MVA, lastly the third (PrefilteringMVA, 3) uses loosened topological cuts and loosened PID cuts together with a MVA, relying solely on the MVA. The motivation for this is to see at first if a MVA can improve an already refined signal (Standard vs. StandardMVA); if not giving any optimized cuts and only relying MVA already gives a considerable result (PrefilteringMVA).

3.4.1 Topological cuts

Two different topological cuts configurations have been applied. The 6 used $p_{\mathrm{T}}(\Lambda_c)$ ranges are [1-2, 2-4, 4-6, 6-8, 8-12, 12-24] GeV/c, as these ranges are also used in the $\Lambda_c \rightarrow pK\pi$ analysis. A full overview of the topological cuts is shown in Table 3.2. The variables in this table are defined as follows: $p_{\mathrm{T}}(\Lambda_c)$ is the transverse momentum of the Λ_c, this variable is divided into six ranges; $m(\Lambda_c \rightarrow pK^0_S) = |m(K^0_S, p) - m(\Lambda_c^{\mathrm{PDG}})|$ is the invariant mass of Λ_c under the assumption that the V0 is a K^0_S and the bachelor a proton minus the Particle Data Group (PDG) mass of Λ_c; $|m(K^0_S) - m(K^{0,\mathrm{PDG}}_S)|$ is the invariant mass of the V0 under the assumption that it is a K^0_S minus the PDG mass; $|m(\Lambda) - m(\Lambda^{\mathrm{PDG}})|$ is the invariant mass of the V0 under the assumption that it is a Λ minus the PDG mass; $p_{\mathrm{T}}(p)$ is the transverse momentum of the bachelor; $p_{\mathrm{T}}(\pi^+, \pi^-)$ is the transverse momentum of the negative and positive daughter of the V0; $p_{\mathrm{T}}(\mathrm{V0})$ is the transverse momentum of the V0; DCA(V0,p) is the distance of closest approach between the V0 and the bachelor track; DCA(π^+, π^-) is the distance of closest approach of the positive and negative daughter of the V0; cos(PA) is the cosine of the pointing angle, defined as the angle between the sum of the three momentum of the two daughters and the line between primary

Cat.	$p_{\mathrm{T}}(\Lambda_c)$ GeV/c^2	$m(\Lambda_c \to pK^0_S)$ GeV/c^2	$m(K^0_S)$ GeV/c^2	$m(\Lambda)$ GeV/c^2	$p_{\mathrm{T}}(p)$ GeV/c	$p_{\mathrm{T}}(\pi^+, \pi^-)$ GeV/c	p_{T}(V0) GeV/c
1, 2	[1-2]	< 0.25	< 0.008	$0.005 < m < 0.05$	> 0.4	> 0.25	> 0.4
	[2-4]	< 0.25	< 0.008	$0.005 < m < 0.05$	> 0.5	> 0.30	> 0.6
	[4-6]	< 0.25	< 0.012	$0.005 < m < 0.05$	> 0.7	> 0.30	> 1.2
	[6-8]	< 0.25	< 0.012	$0.005 < m < 0.05$	> 1.3	> 0.40	> 1.5
	[8-12]	< 0.25	< 0.015	$0.005 < m < 0.05$	> 2.0	> 0.40	> 1.9
	[12-24]	< 0.25	< 0.020	$0.005 < m < 0.05$	> 2.5	> 0.40	> 2.5
3	[1-2]	< 0.20	< 0.030	$m < 0.05$	> 0.0	> 0.0	> 0.4
	[2-4]	< 0.20	< 0.030	$m < 0.05$	> 0.0	> 0.0	> 0.6
	[4-6]	< 0.20	< 0.030	$m < 0.05$	> 0.0	> 0.0	> 1.2
	[6-8]	< 0.20	< 0.030	$m < 0.05$	> 0.0	> 0.0	> 1.5
	[8-12]	< 0.20	< 0.030	$m < 0.05$	> 0.0	> 0.0	> 1.9
	[12-24]	< 0.20	< 0.030	$m < 0.05$	> 0.0	> 0.0	> 2.5
Cat.	$p_{\mathrm{T}}(\Lambda_c)$ GeV/c^2	DCA(V0,p) cm	DCA(π^+, π^-) nσ	cos(PA)	$d_0(p)$ cm	d_0(V0) cm	$m(\mathrm{V0} \to ee)$ GeV/c^2
1, 2	[1-2]	< 1000	< 1.5	> 0.99	< 0.04	< 999	> 0.1
	[2-4]	< 1000	< 1.5	> 0.99	< 0.06	< 999	> 0.1
	[4-6]	< 1000	< 1.5	> 0.99	< 0.08	< 999	> 0.1
	[6-8]	< 1000	< 1.5	> 0.99	< 0.09	< 999	> 0.1
	[8-12]	< 1000	< 1.5	> 0.99	< 0.10	< 999	> 0.1
	[12-24]	< 1000	< 1.5	> 0.99	< 0.20	< 999	> 0.1
3	[1-2]	< 1000	< 0.8	> 0.997	< 3.00	< 1.5	> 0.0
	[2-4]	< 1000	< 0.8	> 0.997	< 3.00	< 1.5	> 0.0
	[4-6]	< 1000	< 0.8	> 0.997	< 3.00	< 1.5	> 0.0
	[6-8]	< 1000	< 0.8	> 0.997	< 3.00	< 1.5	> 0.0
	[8-12]	< 1000	< 0.8	> 0.997	< 3.00	< 1.5	> 0.0
	[12-24]	< 1000	< 0.8	> 0.997	< 3.00	< 1.5	> 0.0

Table 3.2: Topological cuts for the different cut strategies; Standard, 1; StandardMVA, 2; Prefiltering-MVA, 3.

and secondary vertex; $d_0(p)$ is the impact parameter of the bachelor; d_0(V0) is the impact parameter of the V0; $m(\mathrm{V0} \to ee)$ the invariant mass of the V0 under the assumption that the two daughter are an electron-positron pair.

3.4.2 Particle identification

Two different particle identification strategies are applied. The analysis reyling on the Standard(MVA) makes use of the Bayesian PID method. This uses the individual PID response of each detector and converts it into a probability that each particle is of a given species. The PrefilteringMVA has the most loose PID, since it requires for the proton a 3σ cut in TPC if detected or 3σ cut in TOF if detected. Protons that are not detected in the TPC and TOF therefore can also be accepted.

3.5 Machine learning optimization

A machine learning optimization is used the further enhance the signal. The used ROOT software comes with the Multivariate Analysis (MVA) package. For this analysis the same classification machine learning algorithm is chosen which was also used in the Higgs Boson Discovery, also known as a Boosted Decision Tree (BDT). This section gives a brief theoretical overview about this algorithm. Before diving into the algorithm the foundation on which it is based needs to be explained, the decision tree.

#	cos(PA)	d_0(V0) [cm]	DCA [cm]	S/B
1	0.998	0.09	0.13	S
2	0.995	0.40	0.20	B
3	0.997	0.13	0.21	S
4	0.994	0.21	0.16	S
5	0.994	0.21	0.40	B
6	0.993	0.26	0.15	B
7	0.999	0.11	0.30	S
8	0.992	0.12	0.16	B

Table 3.3: Eight Λ_c candidates either background (B, combinatoric-Λ_c) or signal (S, real-Λ_c), with their kinematic variables. cos(PA) is the cosine of the pointing angle of the V0, d_0(V0) is the impact parameter of the V0 and DCA is the distance of closest approach of the positive and negative daughter of the V0.

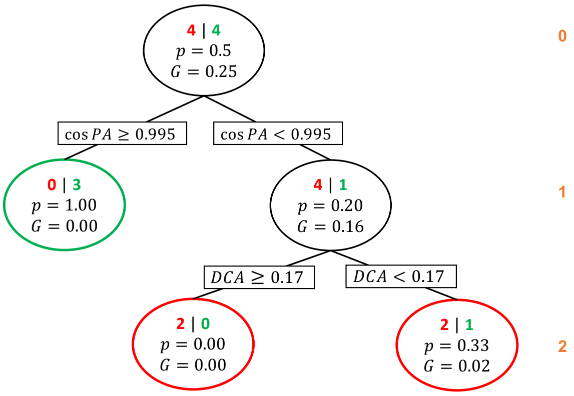


Figure 3.4: Simple decision tree for the candidates in Table 3.3. The green values represent signal and the red background candidates. This decision tree has depth of $D_{\max} = 2$. The tree has three end nodes.

3.6 Decision tree

A decision tree is a decision support tool with a flowchart-like structure that uses conditional control statements to classify cases as either background or signal. The splits in each node are chosen in such way that they maximize reduction of purity for that node. To explain this in more detail, we take as example a tree which is trained on Λ_c candidates. Lets assume we have N Λ_c-candidates in our training sample from which we know if they are background (combinatoric-Λ_c) or signal (real-Λ_c), see Table 3.3. The goal is to make a decision tree which we can be used to determine an unidentified Λ_c candidate as signal or background. A decision tree is 'grown' in such way that it makes cuts which maximizes the purity of the consecutive node. The purity at each node is defined as,

$$p = \frac{N_S}{N_S + N_B}, \tag{3.3}$$

where N_S and N_B denote the number of signal and background candidates in that node respectively. In the first node, before making any cut, we have an purity of $p = 0.5$ for the candidates in Table 3.3. With the purity the Gini-index of a node is computed,

$$G = p(1-p). \tag{3.4}$$

The decision tree cuts the candidates on the variable which maximizes the reduction of purity,

$$\underset{x \in \text{var}}{\operatorname{argmax}} \Delta G(x) = G - \left(\frac{N_L}{N} G_L + \frac{N_R}{N} G_R \right), \tag{3.5}$$

with G the purity in current node, N_L and N_R the number of candidates going the left and right consecutive node respectively, G_L and G_R the corresponding Gini-index of those left and right nodes. The cut which maximizes the impurity reduction is applied which creates two new nodes. This method is repetitively applied until all nodes are purely background or signal ($p = 1$ or $p = 0$).

The problem with repetitively applying equation 3.5 until all nodes are pure is that the tree is overtrained on the data. Hence parameters are introduced to the tree to prevent this from happening. One parameter is the maximum depth of the tree, $D_{\max}$, this prevents the tree from growing larger than a certain level. Level 0 is the level of the starting node. Another parameter is the minimum node size, $S_{\min}$, this is the minimum size a node must have before allowed to split. Figure 3.4 shows a simple decision tree for the candidates in Table 3.3. This tree can be used to classify if a new Λ_c candidate is either background or signal. Consider a real Λ_c candidate which has cos(PA) = 0.994 for the V0 and DCA = 0.13 cm, this candidate would then be classified as a background candidate using the decision tree in Figure 3.4. The problem with a single decision tree is that it is sensitive to the training sample. Therefore a boosting algorithm is applied to reduce this.

3.6.1 Boosted Decision Tree

The idea behind a Boosted Decision Tree (BDT) [23] is that for a given training set, a whole forest of decision trees is generated, each trained on a subset of the complete training sample. The trees are successively trained after each other, giving the misclassified candidates a higher weight. This final forest of decision trees, also called the BDT, gives an average classification for the new Λ_c candidate. The average of all results gives the 'BDT response' of the BDT $\bar{y}_{\text{BDT}} \in [-1, +1]$ and represents how much the BDT classifies a unidentified Λ_c candidate as background or signal, with pure background at $\bar{y} = -1$ and pure signal at $\bar{y} = +1$, e.g. if all of the trees in the forest classifies the candidate as signal the BDT output is $\bar{y}_{\text{BDT}} = 1.0$.

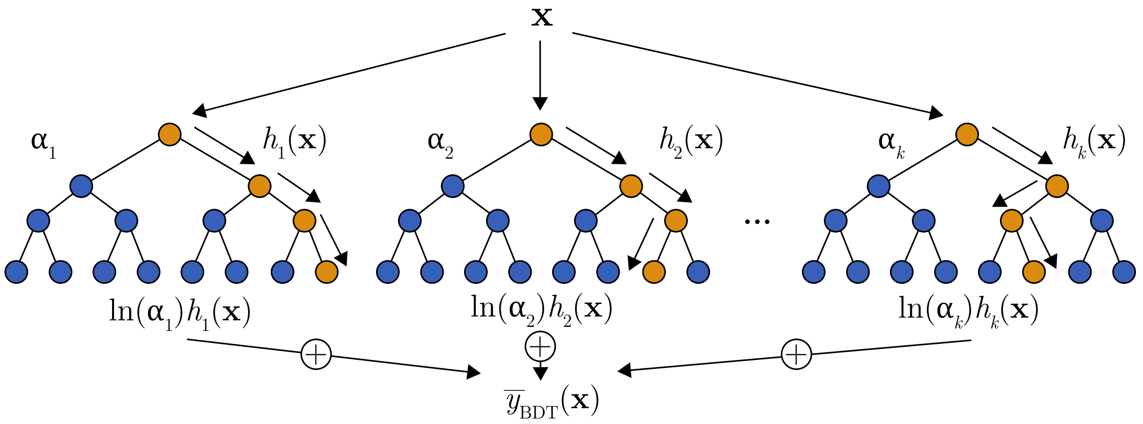
3.6.2 Training the BDT

Let's assume we want to train a BDT of N_{T} trees. Firstly, all the candidates in the training sample get the same weight $w^i_{k=0} = 1/N$, with N the number of candidates in the training sample, k the tree number and i the candidate number. The first tree is grown on the training sample. Successively, all the candidates that are incorrect classified provide the misclassification rate $e_{k=1}$ of the first tree,

$$e_{k=1} = \frac{N_{\text{mis}}}{N}, \tag{3.6}$$

where N_{mis} determines all the candidates in all the end-nodes that are misclassified. Next, the previous weights of all the misidentified candidates are multiplied by their boost weight α_k to obtain their new

Figure 3.5: Schematic overview of the forest created by Boosted Decision Tree algorithm.



weight,

$$w^i_{k+1} = w^i_k \alpha_k = w^i_k \left(\frac{1 - e_k}{e_k} \right)^\beta , \tag{3.7}$$

with β the learning rate. This parameters determines how quickly the subsequent trees learn from their mistakes. These rescaled weights are eventually renormalized to get the weights for the next tree. This procedure is performed until the number of trees in the forest has been reached.

We define the result of an individual classifier (tree) as $h_k(\mathbf{x})$, with $\mathbf{x}$ the tuple of variables. $h_k(\mathbf{x})$ could either be $+1$ for signal or -1 for background, as a single tree can only classify a candidate as signal or background. The BDT output $\bar{y}_{\mathrm{BDT}}$ for the whole forest of trees is defined as,

$$\bar{y}_{\mathrm{BDT}}(\mathbf{x}) = \frac{1}{N_{\mathrm{T}}} \sum_{k=1}^{N_{\mathrm{T}}} \ln (\alpha_k) \, h_k(\mathbf{x}). \tag{3.8}$$

A schematic overview of the BDT is shown in Figure 3.5. The now trained model (BDT) is ready for classifying Λ_c candidates. The trained model is applied on a testing sample from which is known if the candidates are signal or background in order to check the preciseness of our model. The quality of the model is checked if the model is not overtrained on the training sample. The model that increases the significance $\mathcal{S} = \frac{N_S}{\sqrt{N_S + N_B}}$ and seems not to be overtrained or biased is chosen as model to be applied on the full data set of unidentified Λ_c candidates, to enhance real Λ_c's.

3.7 BDT Configuration

The optimization was performed using a Boosted Decision Tree from the ROOT TMVA-Toolkit for Multivariate Data Analysis [23]. In order to build the training sample, signal candidates were taken from Monte Carlo simulation considering only prompt Λ_c candidates generated with PYTHIA, full overview see Table 3.1. A prompt Λ_c is a candidate that is directly formed in the pp collision originating from the primary vertex. The opposite is a feed-down candidate coming from a Λ_b decaying into Λ_c and a π, thus does not directly originate from the primary vertex. The background candidates were taken from the 7-12σ side bands from data for each $p_{\mathrm{T}}(\Lambda_c)$ (from now determined as p_{T}) range. The σ for each p_{T} range is extracted from the invariant mass fit of solely Monte Carlo. The number of signal and background candidates for each year are taken as there contribution to in real data, i.e. the runs in 2017 has the most number of events hence this has the highest number of signal and background events in the training sample. In this way the model is trained on the same ratio of signal and background events as represented in real data.

Table 3.4: Boosted Decision Tree configuration together with trained variables

Parameter	Value	Trained variables	
Number of trees	850	$m(K^0_S)$	$n\sigma_{\mathrm{TOF}}(p)$
Number of grid steps	20	d_0(bach.)	$n\sigma_{\mathrm{TPC}}(p)$
Maximum depth	3	d_0(V0)	$n\sigma_{\mathrm{TPC}}(\pi)$
Boosting	Adaboost, $\beta = 0.5$	$c\tau(K^0_S)$	$n\sigma_{\mathrm{TPC}}(K)$
Minimum node size	5.0%	cos(PA)	
Normalization mode	EqualNumEvents	d'_0	

The last transverse momentum interval $12 < p_{\mathrm{T}} < 24$ GeV/c has a large width, hence the distribution of variables could be different for candidates around 12 GeV/c compared to 24 GeV/c. Moreover, the momentum spectrum decreases exponentially, so this interval would be more trained on the lower range of it. For this reason, this interval in addition has been trained the interval $12 < p_{\mathrm{T}} < 16$ GeV/c and $16 < p_{\mathrm{T}} < 24$ GeV/c, after which the results are eventually merged. The models are trained for every MVA cut strategy per p_{T} range integrated over multiplicity. The reasons for this are: firstly, given the current Monte Carlo statisics available, it was not possible to perform an optimization in bins of multiplicity. In addition, it is overall preferable to perform an unique selection as a function of multiplicity in order to limit the effect of possible MC/data differences that can emerge in events at higher multiplicities. The configuration of the BDT is summarized in Table 3.4, in case a parameter is not explicitly mentioned, the default value was used. To be noted that most values reported in Table 3.4 correspond to the default settings which should deliver, in recent TMVA versions, a very good performance. Tests running the classifier with different settings were performed in previous analyses without a significant gain in performance [24]. Around 70% of the sample is used for training, with a maximum of 500.000 candidates, while the other 30% is used for testing.

The BDT has been trained on the set of 10 variables shown in Table 3.4. The variables are defined as followed: $m(K^0_S)$ is the invariant mass of the V0 under the assumption that it is a K^0_S; d_0(bach.) is the impact parameter of the bachelor; d_0(V0) is the impact parameter of the V0; $c\tau(K^0_S) = l_{\mathrm{V0}} \times m(K^0_S)/p(\mathrm{V0})$ is the proper decay length of the K^0_S with l_{V0} the decay distance of the V0, $m(K^0_S) = 0.497$ GeV/c^2 the PDG mass of K^0_S; cos(PA) is the cosine of the pointing angle of the V0; d'_0 is signed impact parameter of the bachelor track with respect to the primary vertex; p(bach.) is the momentum of the bachelor; $n\sigma_{\mathrm{TOF}}(p)$ is the number of sigma in the TOF assuming the bachelor is a proton; $n\sigma_{\mathrm{TPC}}(p)$ is the number of sigma in the TPC assuming the bachelor is a proton; $n\sigma_{\mathrm{TOF}}(\pi)$ is the number of sigma in the TPC assuming the bachelor is a pion; $n\sigma_{\mathrm{TPC}}(p)$ is the number of sigma in the TPC assuming the bachelor is a kaon. These variables were selected because they already have been studied in a previous analysis.

3.7.1 BDT Performance

The training variables and invariant mass distribution of Λ_c for $2 < p_{\mathrm{T}} < 4$ GeV/c for background and signal candidates are shown in Figure 3.6. Superimposed in green (brown) are shown the left (right) background distributions. From the figure it directly can be seen that already a big improvement can be made by cutting away the background of $n\sigma_{\mathrm{TPC}}(p)$. Sequentially, it is not surprisingly that this variable came out as the most important variable in the MVA algorithm. The left and right background separation does not show a discrepancy between each other, hence an one to one left and right background input to reduce overtraining on either of one is not needed.

The top two graphs of Figure 3.7 shows first of all the correlations between the input variables, as measured by the TMVA package, separately for the signal and the background candidates for PrefilteringMVA $2 < p_{\mathrm{T}} < 4$ GeV/c. The graphs shows that there is a high correlation between all the identifications of the bachelor in the TPC. Nevertheless, in Λ_c Pb-Pb analysis is shown that this is in general not a problem since it has been shown that the BDT analysis and performance is not affected by correlations between input variables [24]. Secondly, the middle left graph of Figure 3.7 shows the comparison between training and test sample distribution for both signal and background candidates. The test sample clearly shows a similar result as the training sample for both background and signal, thus we can conclude that our model has not been overtrained on the training sample. Lastly, the middle right graph of Figure

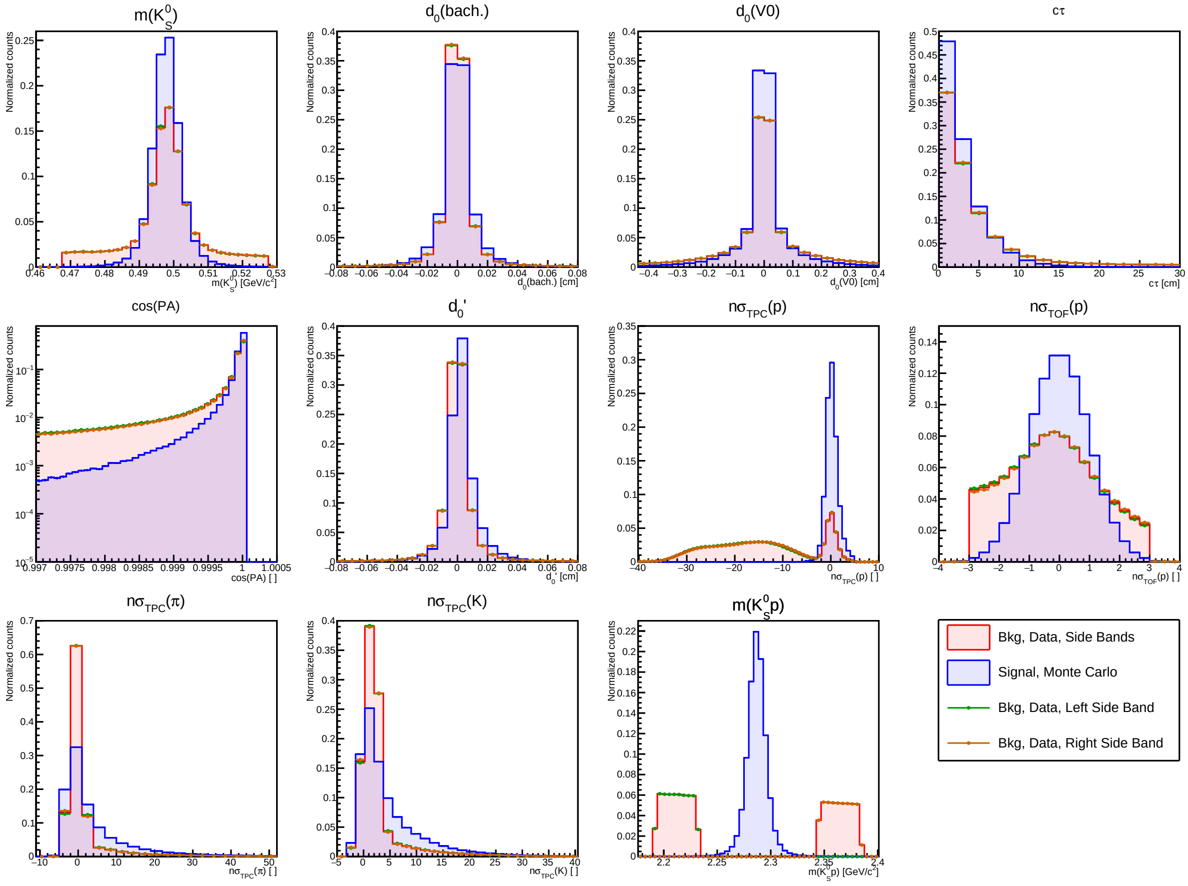


Figure 3.6: BDT input variables and invariant mass distributions of Λ_c for $2 < p_{\mathrm{T}} < 4$ GeV/c for PrefilteringMVA strategy. The filled red denote the background sample, while the filled blue the signal sample. Superimposed is the left (green markers) and right side band (brown markers) of the background candidates. The y-axis shows the normalized counts for each distribution. Note that cos(PA) has a logarithmic y-axis.

3.7 shows the performance of the model, the area under this 'ROC-curve' (in this case 0.876) gives a quantitative outcome of how good the model can classify signal from background candidates. In the ideal case this area is equal to 1, and the curve shows a rectangular shape. This p_{T} range shows an adequate results, so this model has been accepted.

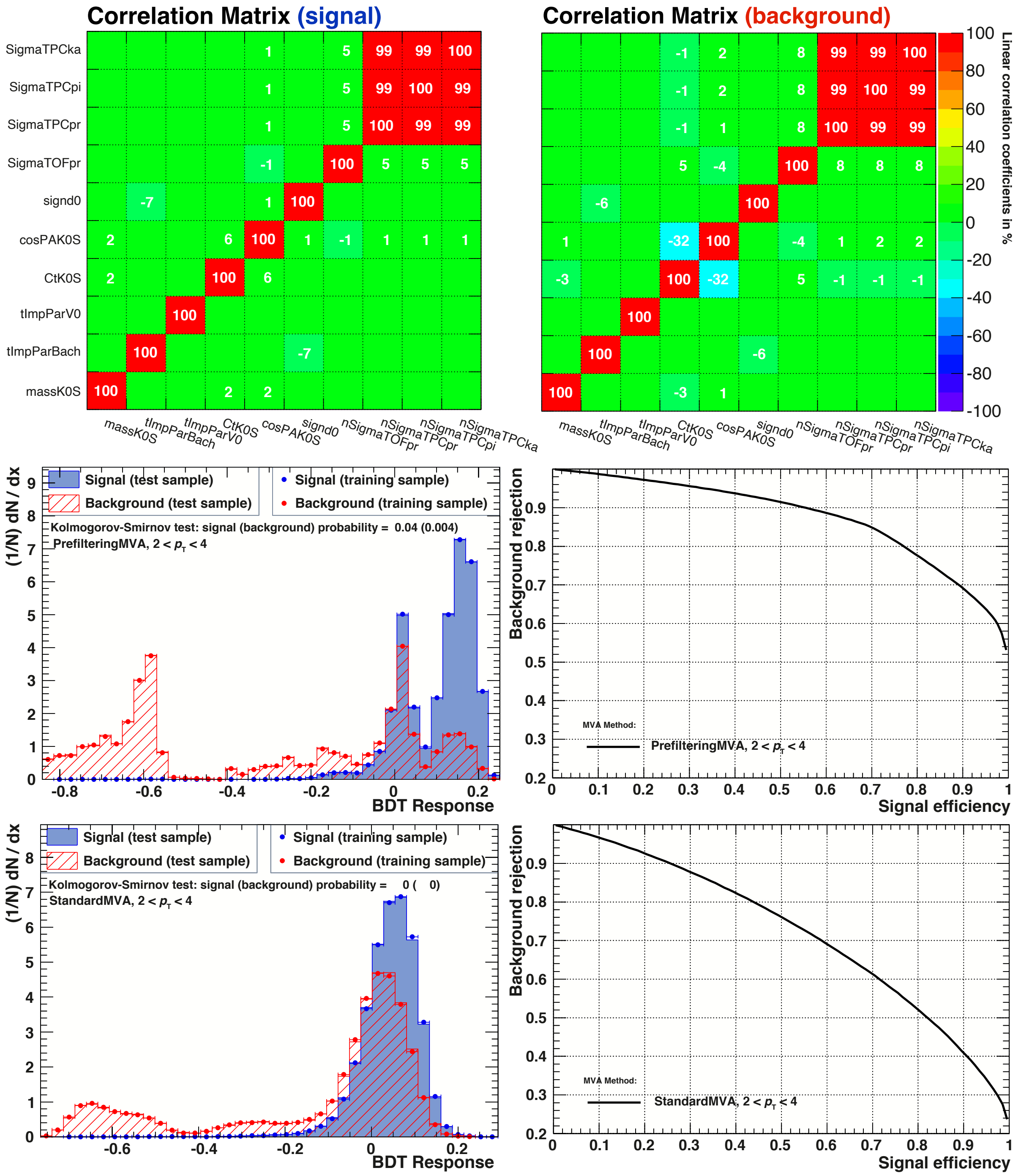


Figure 3.7: Top left (right) shows the correlation between the signal (background) candidates for PrefilteringMVA $2 < p_T < 4$ GeV/c. Middle left shows a BDT response $\bar{y}_{\mathrm{BDT}}$ comparison between the training and test sample distribution for signal and background for PrefilteringMVA $2 < p_T < 4$ GeV/c. Middle right shows the corresponding ROC-curve and has an area of 0.876. Bottom left shows a BDT response $\bar{y}_{\mathrm{BDT}}$ comparison between the training and test sample distribution for signal and background for StandardMVA $2 < p_T < 4$ GeV/c. Bottom right shows the corresponding ROC-curve and has an area of 0.723.

The same performance results are shown in the bottom two graphs of Figure 3.7 for the StandardMVA strategy for $2 < p_{\mathrm{T}} < 4$ GeV/c. This model has an ROC-curve with an area of 0.723, and is therefore at first sight compared to the PrefiltergMVA model worse in classifying background from signal. But as the StandardMVA configuration has already a stricter set of cuts, this conclusion cannot been drawn. If one want to make qualitative statements between cuts plus MVA strategies, we need to look at the output of the invariant mass fits. This same procedure of training and classifying is done for all two MVA cut strategies and p_{T} ranges.

3.8 Signal extraction

The raw yield extraction of the Λ_c is usually done by fitting the invariant mass distributions of the Λ_c-candidates. This still holds for the 'Standard' analysis, but as the MVA analyses has stored the BDT response together with the invariant mass of all the candidates, a dedicated selection on the BDT response needs to be made in order extract the raw yield. For both the non-MVA as the MVA analyses a fitting function has been chosen consisting of polynomial of the second for the background together with a Gaussian describing the signal, whose width was fixed to the value obtained in the Monte Carlo.

3.8.1 Finding optimal BDT response

The MVA analyses currently did not have dedicated algorithm for determining the optimal BDT response, $\bar{y}_{\mathrm{BDT}}$. Therefore a tool is build which visualized what the outcome would be for different BDT responses, this tool will briefly been explained, see Figure 3.8. First recall that $\bar{y}_{\mathrm{BDT}}$ shows quantitatively how much the model classifies Λ_c-candidates as background ($\bar{y}_{\mathrm{BDT}} = -1$) or signal ($\bar{y}_{\mathrm{BDT}} = +1$). The goal is to find a value for $\bar{y}_{\mathrm{BDT}}$, which seems to have a stable fit, has a width compared to Monte Carlo and has a large significance. In order to do so with small steps $\Delta\bar{y}_{\mathrm{BDT}} = 0.0002$ is iterated over the BDT response $\bar{y}_{\mathrm{BDT}}$, while for each step all the candidates that have $\bar{y}_{\mathrm{BDT}}$ greater or equal to that value are selected, after which a fit is made through the invariant mass distribution. Please see graph 9 of Figure 3.8, this graph shows the BDT response of all the Λ_c candidates valued by the model for StandardMVA $2 < p_{\mathrm{T}} < 4$ GeV/c, together with the minimum ($\bar{y}_{\mathrm{BDT}} = -0.2$) and maximum ($\bar{y}_{\mathrm{BDT}} = 0.1$) BDT response between which is iterated visualized by the red dotted lines. In order to prevent the tool from selecting statistical fluctuation the fit results are grouped in sets of ten from which the mean output of their fit parameters is computed. The result of their mean fit parameters are eventually shown by the tool, which can be seen in panel 1-7.

First the best fit parameters (rebinning, fit ranges) have been chosen by running the tool for different parameter settings. After the best fit parameters were found the tool is ran with sigma fixed to Monte Carlo and with sigma unconstrained. The unconstrained sigma is used as reference for the fixed sigma. We first look for an BDT response where the unconstrained sigma is in the region of Monte Carlo, in case of the figure around BDT bin 5000 ($\bar{y}_{\mathrm{BDT}} = 0$). Next we would like to maximize the significance, avoiding the extreme BDT response cuts where too much signal is cut away, as this would decrease the TMVA efficiency drastically. Usually a minimum TMVA Efficiency of 0.65 is accepted, which in this case is around BDT bin 5100 ($\bar{y}_{\mathrm{BDT}} = 0.2$). Lastly, the best fit is taken favoring lower BDT reponses over higher. Successively, we found that the best and most stable fit was at BDT bin 4949 ($\bar{y}_{\mathrm{BDT}} = -0.0104$), see Figure 3.9. This same procedure is done for all MVA cut strategies and p_{T} ranges.

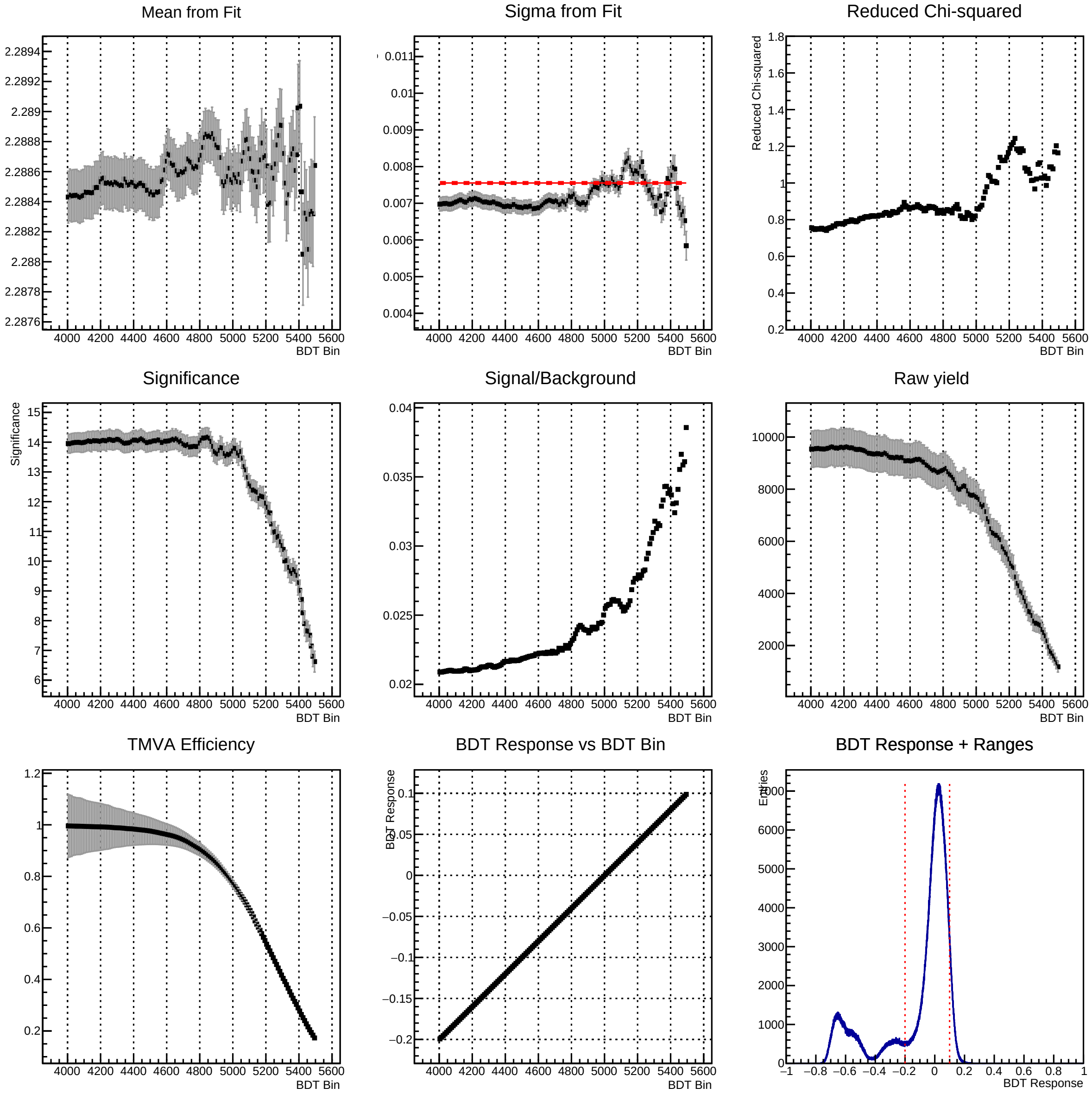


Figure 3.8: Visual output of the tool for determining the optimal BDT response $\bar{y}_{\mathrm{BDT}}$ for StandardMVA $2 < p_{\mathrm{T}} < 4$ GeV/c fitted with sigma unconstrained. The x-axis of graph 1-7 is the BDT bin which is one to one related to the BDT response, see graph 7, BDT bin 4000 ($\bar{y}_{\mathrm{BDT}} = -0.2$) and BDT bin 5500 ($\bar{y}_{\mathrm{BDT}} = 0.1$). Graph 1-2 show the mean and sigma from fit, superimposed the values from Monte Carlo in red. Graph 3 shows the reduced chi-squared, graph 4 the significance $\mathcal{S} = \frac{N_S}{\sqrt{N_S+N_B}}$ within 3σ under the Gaussian with N_S and N_B the number of signal and background candidates, graph 5 shows the signal over background ratio N_S/N_B within 3σ under the Gaussian, graph 6 shows the raw yield within 3σ under the Gaussian, graph 7 shows the TMVA Efficiency defined as the by Monte Carlo classified real Λ_c candidates selected by the BDT response divided by the total number of classified real Λ_c candidates, graph 9 shows the BDT response distribution for all the Λ_c-candidates together with the minimum and maximum BDT response between which is iterated. Each point represents the mean of 10 successive BDT response steps in order to prevent statistical outliers.

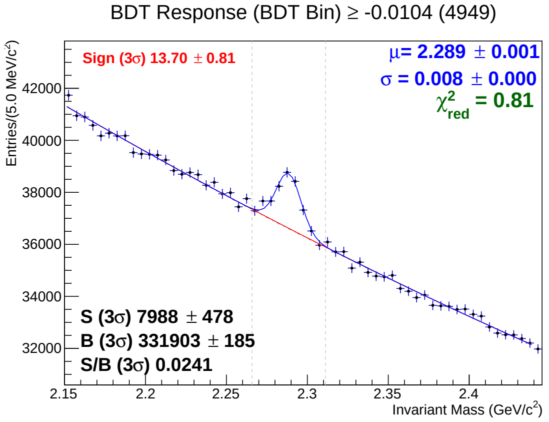


Figure 3.9: Invariant mass fit of Λ_c for the optimal BDT response for StandardMVA $2 < p_{\mathrm{T}} < 4$ GeV/c. A optimal BDT response of $\bar{y}_{\mathrm{BDT}} \geq -0.0104$ has been chosen. The red (blue) lines represent the background (plus signal) fit. The results of the Gaussian fit is reported, together with the significance and the signal over background ratio. The number of signal and background candidates is obtained within a 3σ range (grey dotted line) around the mean value of the Gaussian fit.

3.8.2 Fit results

Figure 3.10 shows the invariant-mass distributions together with the fit for different p_{T} intervals of the Standard, StandardMVA and PrefilteringMVA strategy. A total overview of the mean, sigma (unconstrained), raw-yield, signal over background ratio and significance of all fits is shown Figure 3.10.

These results show that only in the StandardMVA analysis the highest p_{T} interval could reach a significance above three, even worse in the PrefilteringMVA analysis no fit could be made in this interval, even though it has been trained separately in $12 < p_{\mathrm{T}} < 16$ GeV/c and $16 < p_{\mathrm{T}} < 24$ GeV/c. On the other hand, PrefilteringMVA is the only strategy of which an unconstrained sigma does not significantly deviates from Monte Carlo, as the other two strategies deviate more than three standard deviations in $6 < p_{\mathrm{T}} < 8$ GeV/c. The StandardMVA analysis improves compared to Standard in the signal over background ratio drastically with factor 1.5 in low p_{T} and 2.5 in intermediate/high p_{T}. The StandardMVA analysis improves compared to Standard in significance with factor 1.3 in low p_{T} and 1.7 in intermediate/high p_{T}. The significance increases in PrefilteringMVA compared to Standard in all p_{T} except the $2 < p_{\mathrm{T}} < 4$ GeV/c range the significance.

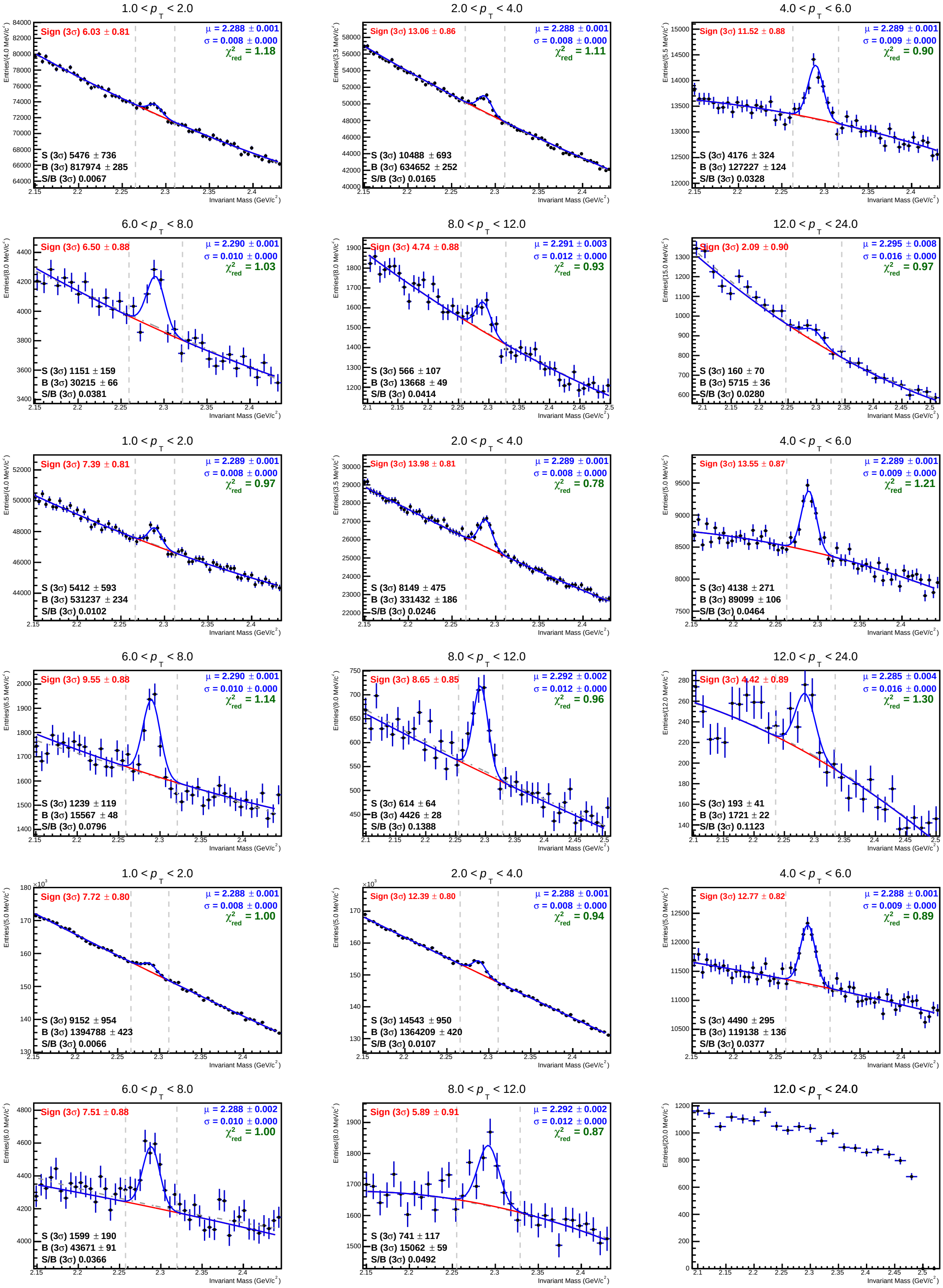


Figure 3.10: Invariant mass distributions of Λ_c candidates for $1 < p_{\mathrm{T}} < 2$ GeV/c, $2 < p_{\mathrm{T}} < 4$ GeV/c, $4 < p_{\mathrm{T}} < 6$ GeV/c, $6 < p_{\mathrm{T}} < 8$ GeV/c, $8 < p_{\mathrm{T}} < 12$ GeV/c, $12 < p_{\mathrm{T}} < 24$ GeV/c for Standard (top six) and StandardMVA (middle six) and PrefilteringMVA (bottom six) strategy. The red (blue) lines represent the background (plus signal) fit. The results of the Gaussian fit is reported, together with the significance and the signal over background ratio. The number of signal and background candidates is obtained within a 3σ range (grey dotted line) around the mean value of the Gaussian fit.

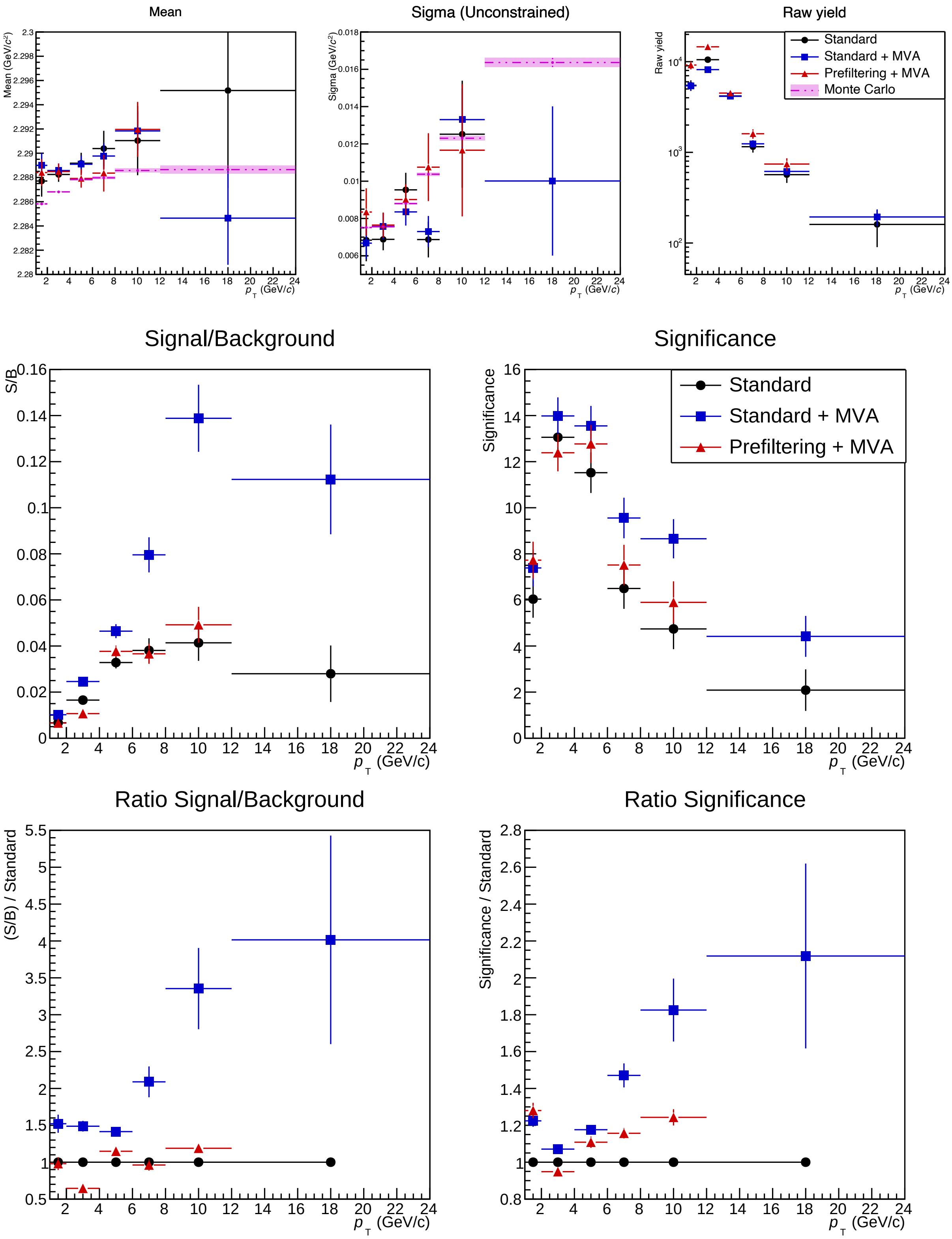


Figure 3.11: Top three: Mean (left) and raw-yield (right) from fit with the width of the peak fixed to Monte Carlo, as in Figure 3.10. The middle graph shows the sigma from fit if the width of the peak is unconstrained. The pink dotted line shows the mean and sigma from Monte Carlo. Bottom four: Signal over background ratio and significance of different strategies: Standard (black), StandardMVA (blue) and PrefilteringMVA (red). The ratios are divided by the Standard analysis.

3.9 Corrections and Efficiencies

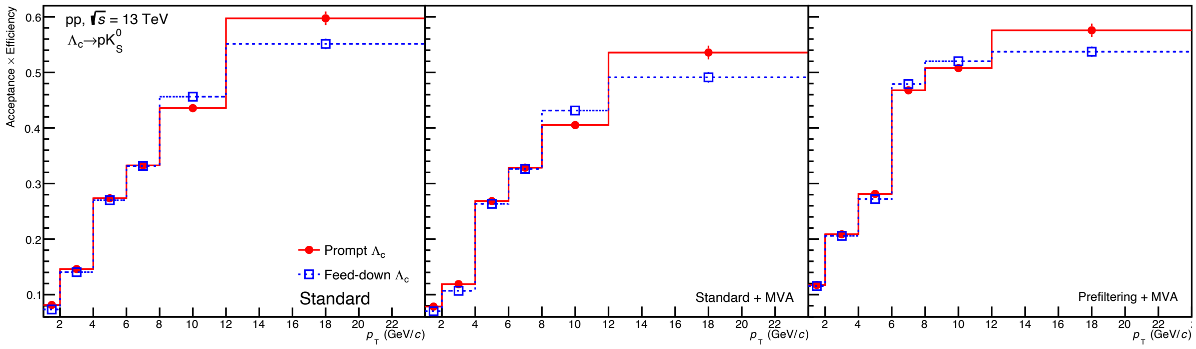
The outcome of each strategy is dependent on the performance of ALICE and selection criteria that were made. As the underlying physics is the same and as all the different analyses come from the same data set, corrections need to adjust for this. This section will give a overview of all the corrections that had to be considered in order to compute p_{T}-differential corrected yield per event. The p_{T}-differential corrected yield per event for $|y| < 0.5$ of prompt Λ_c^+ was obtained using the following formula,

$$\frac{1}{N_{\mathrm{ev}}}\frac{\mathrm{d}N^{\Lambda_c^+}}{\mathrm{d}p_{\mathrm{T}}}\bigg|_{|y|<0.5} = \frac{1}{2\Delta p_{\mathrm{T}} c_{\Delta y}}\frac{1}{N_{\mathrm{ev}}\mathrm{BR}}\frac{N^{\Lambda_c^{\mathrm{Raw}}}(p_{\mathrm{T}})|_{y<y_{\mathrm{fid}}}\cdot f_{\mathrm{prompt}}}{(\mathrm{Acc}\times\epsilon)(p_{\mathrm{T}})^{|y|<0.5}_{\mathrm{prompt}}}, \tag{3.9}$$

where $N^{\Lambda_c^{\mathrm{Raw}}}$ is the raw-yield (both particles and antiparticles) in a given p_{T} interval with width Δp_{T}, the factor 2 is to account for the fact that $N^{\Lambda_c^{\mathrm{Raw}}}$ contains both particle and antiparticles, f_{prompt} is the fraction of the raw-yield from prompt Λ_c, $(\mathrm{Acc}\times\epsilon)$ is the product of the acceptance and efficiency of prompt Λ_c in a given p_{T} interval for $|y| < 0.5$, N_{ev} is the number of analyzed events, $\mathrm{BR} = 0.0110$ is the branching ratio for the total decay mode and $c_{\Delta y}$ is the correction factor for the rapidity coverage.

The correction for the detector acceptance and reconstruction efficiency $(\mathrm{Acc}\times\epsilon)$ was obtained using the Monte Carlo periods listed under 'Efficiencies' in Table 3.1. The efficiencies are calculated in two steps, the cut efficiency and the MVA efficiency. The cut efficiency is the number of reconstructed Λ_c baryons after topological cuts and PID selection over the number of particle generated in the acceptance. The MVA efficiency is computed by applying to the Monte Carlo sample the same cuts, PID, MVA model and optimal BDT response for p_{T} as used in extracting the raw-yield, from which the ratio of the number of accepted prompt (feed-down) Λ_c's over the total number of prompt (feed-down) Λ_c's is computed. This efficiency depends solely on the cut in the BDT response. The resulting total efficiency is the product of the MVA and cut efficiency. The acceptance is the number of Λ_c baryons generated in the acceptance over generated in the limited acceptance. Figure 3.12 shows the resulting product of acceptance and efficiency for prompt and feed-down Λ_c's. As observed, the product of acceptance and efficiency increases with the transverse momentum, a few percentile at low p_{T} going up to 40-60% at high p_{T}. The relative error for prompt (feed-down) varies from 1% (1%) at low p_{T} up to 3% (2%) at high p_{T}. The efficiencies of StandardMVA are overal 10% less with a maximum of 20% at $2 < p_{\mathrm{T}} < 4$ GeV/c compared to Standard. The PrefilteringMVA efficiencies are overall 35% higher with an exception of 1% higher at $4 < p_{\mathrm{T}} < 6$ GeV/c and 1% lower at $16 < p_{\mathrm{T}} < 24$ GeV/c.

Figure 3.12: Product of acceptance and efficiency for Λ_c in pp collisions at $\sqrt{s} = 13$ TeV as function of p_{T} for Standard, StandardMVA and PrefilteringMVA. The solid red lines correspond to the prompt Λ_c, while the dotted blue lines represent Λ_c baryons originating from beauty-hadron decays.



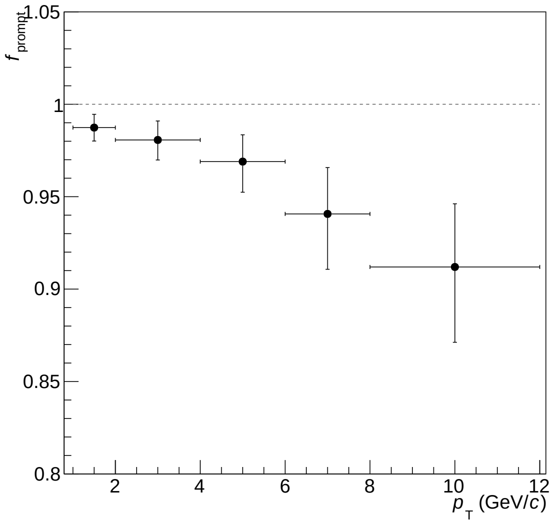


Figure 3.13: Fraction of prompt Λ_c baryons with asymmetric systematic uncertainties.

As the efficiencies and raw-yield for all the different strategies have been computed, the correction in Equation 3.9 can be applied to obtain the corrected yield of Λ_c for all strategies. All the strategies would ideally show the same outcome, hence the most reliable strategy must be chosen to become the final results of which also the systematics uncertainties will be calculated.

The strategy to continue with will be PrefilteringMVA, as this strategy this strategy only does not significantly differs from Monte Carlo in $6 < p_{\mathrm{T}} < 8$ GeV/c. A discrepancy would result in a large systematic uncertainty in the raw-yield extraction. As this strategy could not be fitted in the highest transverse momentum interval, it will be left out of the analysis. The reason that StandardMVA is not chosen although has a higher signal over background ratio, significance and has a significant result in $12 < p_{\mathrm{T}} < 24$ GeV/c is because this strategy might be statistically biased as it is based on optimized topological cuts plus a MVA.

To obtain the factor f_{prompt}, i.e. the fraction of prompt Λ_c in the raw-yield, the production cross section of Λ_c from Λ_b decays was estimated using the beauty hadron p_{T} shape from FONLL calculation [25]. The so called N_{b} method was used in order to compute the fraction of prompt Λ_c, using,

$$f_{\mathrm{prompt}} = 1 - \frac{N^{\Lambda_c,\text{feed-down}}}{N^{\Lambda_c}} = 1 - \frac{(\mathrm{Acc} \times \epsilon)_{\text{feed-down}}\, c_{\Delta y}\, \Delta p_{\mathrm{T}}\, \mathrm{BR}\, \mathscr{L}_{\mathrm{int}}}{N^{\Lambda_c}/2} \times \left(\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_{\mathrm{T}}\, \mathrm{d}y}\right)^{\mathrm{FONLL}}_{\text{feed-down}}, \quad (3.10)$$

where $N^{\Lambda_c}/2$ is the raw-yield with a of factor two for the particle and antiparticle correction, $(\mathrm{Acc} \times \epsilon)_{\text{feed-down}}$ the product of acceptance and efficiency for feed-down Λ_c's, $\mathscr{L}_{\mathrm{int}} = N_{\mathrm{ev}}/\sigma_{\mathrm{int}}$ the integrated luminosity, this is the number of analyzed events (1.5 billion) divided by the total inelastic cross-section measured as 57.8 ± 2.3 mb for LHC16, LHC17 and LHC18 pp $\sqrt{s} = 13$ TeV data. The last term the differential cross-section for feed-down Λ_c from the FONLL calculation. The result of f_{prompt} is shown in Figure 3.13.

3.10 Systematic Uncertainties

The sources of systematic uncertainties considered in this analysis are the following: (1) yield extraction, (2) cut variation, (3) particle identification, (4) Monte Carlo p_{T}-shape of generated Λ_c, (5) feed-down subtraction, (6) tracking efficiency. A summary of the systematic uncertainties is shown in Table 3.5 and in Figure 3.17.

3.10.1 Yield extraction

The systematic error on the raw-yield extraction was estimated in each p_{T} interval by fitting the invariant-mass distributions repetitively under different approaches. The following approaches were considered: (i) the background function, either parabolic, linear or exponential, (ii) the lower and upper limit of the fit range were varied, (iii) varying the width of the invariant-mass histogram, and (iv) counting the entries within 2.5, 3, 3.5 times the width after subtracting the background. For all the possible combinations the fit was performed under different assumptions on the Gaussian width of the Λ_c, namely: (a) fixing the Gaussian width to the value obtained in Monte Carlo, (b) leaving the width as free parameter. Only cases where the fit had a reduced chi-squared smaller than two, significance above three, and relative error above 0.5 were considered. Cases with raw-yield far off the central value were in more depth investigated, fits that did not represent the data were not considered. The final systematic uncertainty was eventually estimated as the root-mean-square (RMS) of the signal yield from the obtained trials.

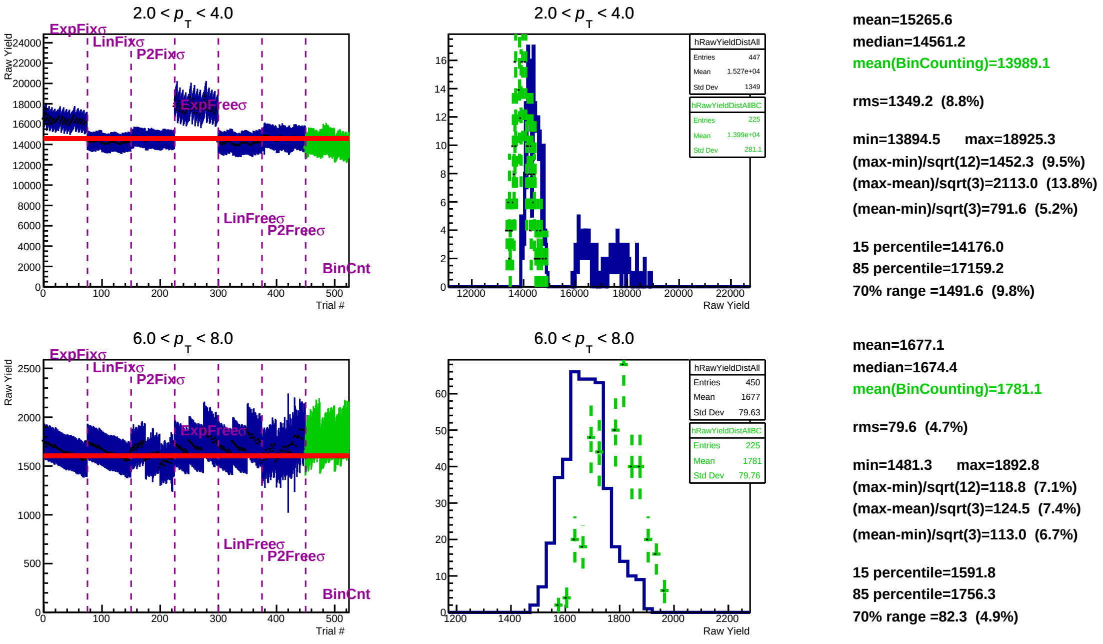


Figure 3.14: Results of multi-trial yield extraction for multiplicity integrated $2 < p_{\mathrm{T}} < 4$ GeV/c and $8 < p_{\mathrm{T}} < 12$ GeV/c. The red line denotes the yield of the central point.

The results of this multi-trial yield extraction for $2 < p_{\mathrm{T}} < 4$ GeV/c and $6 < p_{\mathrm{T}} < 8$ GeV/c is shown in Figure 3.14. The result for the $6 < p_{\mathrm{T}} < 8$ GeV/c shows a clear fluctuation around the central value and the bin counting is in agreement with the other trials, hence the outcome of the RMS is estimated at 5%. In the case of the $2 < p_{\mathrm{T}} < 4$ GeV/c interval the trials with the exponential function showed a clear discrepancy with the central value and the other trials. Consequently, these trials are removed and the RMS is estimated at 7%.

Since the systematic error in raw-yield extraction is expected to be reasonably smooth as a function of p_{T}, additional smoothening of the systematic uncertainties is performed. The final results are shown in Table 3.5 and Figure 3.17.

3.10.2 Cut variation

Systematic uncertainties on the selection on Λ_c can arise from certain imperfection between data and simulation in the distributions and resolution of variables on which is cut. As we expect the Prefiltering-MVA strategy to dependent mostly on the BDT response and not on the initial topological and PID cuts, the cut variation systematic uncertainty can be estimated varying the BDT responses around the selected BDT response for the central point. Two stricter points of $\bar{y}_{\mathrm{BDT,Central}} + 0.01$ and $\bar{y}_{\mathrm{BDT,Central}} + 0.02$ and two looser of $\bar{y}_{\mathrm{BDT,Central}} - 0.01$ and $\bar{y}_{\mathrm{BDT,Central}} - 0.02$ were chosen.

The raw-yield has been extracted using the same fit parameters as the central value, to limit fluctuations due to fit performance. The ratio of the corrected yield with the central value is shown on the left of Figure 3.15. Independently of each other the four different cuts fluctuate around 1, thus no p_{T} is observed in these cut variations. Though an asymmetry is observed, more points lay beneath the central point then above, therefore an asymmetric systematic uncertainty is estimated of 4% above and 8% below.

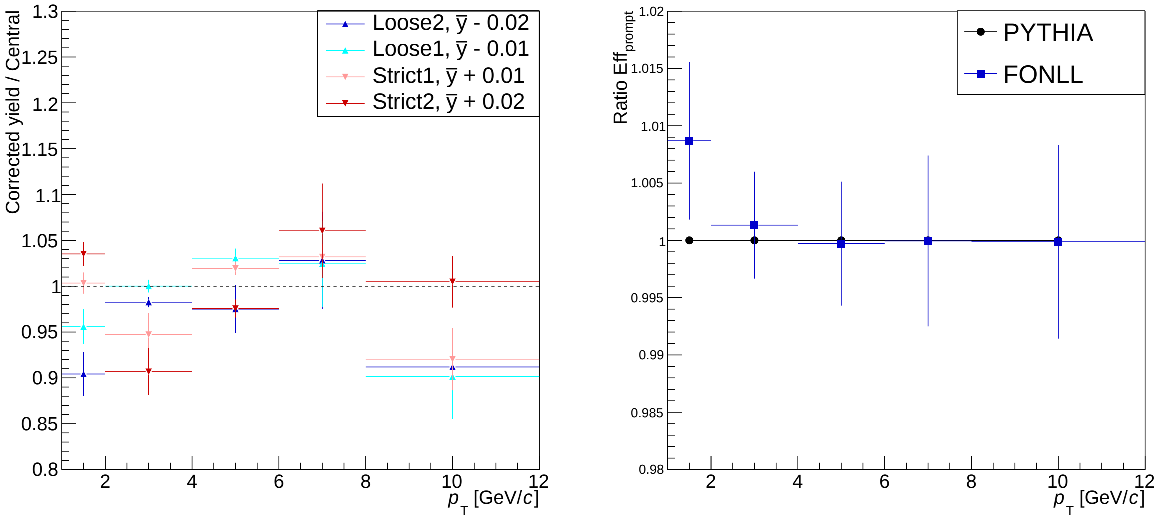


Figure 3.15: Left: Ratio of the corrected yield with central point varying the Λ_c selection cuts. Right: Systematic uncertainty due to Monte Carlo p_{T}-shape, ratio of prompt Λ_c efficiency for PYTHIA and FONLL.

3.10.3 Particle identification

The effect of the particle identification is usually estimated doing the same analysis without using any PID, as the Λ_c baryon is hard to reconstruct without any PID the other option is to vary the number of sigma in TOF and TPC of which eventually corrected yield ratio with the central value is calculated. In this analysis there was not enough time to compute these ratios, therefore the results have been taken from the other decay channel $\Lambda_c \rightarrow pK\pi$, which also required the reconstruction of the proton. The analysis estimated a 5% systematic on the particle identification for all p_{T}.

3.10.4 Monte Carlo transverse-momentum shape

The efficiencies computed with Monte Carlo are dependent on the generated p_{T} distribution of the Λ_c baryons in PYTHIA. The effect of this shape from PYTHIA was estimated from the relative variation in the Monte Carlo efficiencies obtained after using p_{T} shapes from FONLL. The uncertainty due to these selection was estimated by of ratio of the prompt Λ_c efficiency generated by PYTHIA over generated by FONLL, the results can been seen on the right of Figure 3.15. The ratio lays within uncertainties for all transverse momenta, except the first, hence only for the first interval a systematic uncertainty has been estimated of 1%.

3.10.5 Feed-down subtraction

The contribution to the systematic uncertainties coming from the feed-down correction f_{prompt} in Equation 3.9 was calculated with the theoretical uncertainties of the FONLL predictions of $\Lambda_b \rightarrow \Lambda_c$. The uncertainty in the FONLL prediction of the Λ_b production arise by varying the b-quark mass, the perturbative normalization scales or the factor of the fraction of beauty quarks decaying into a Λ_c baryon. The resulting asymmetric errors on the value of f_{prompt} are shown in Figure 3.13 and summarized in Table 3.5.

3.10.6 Tracking efficiency

The systematic uncertainty on the track reconstruction efficiency arise from two different effects: the quality of the tracks and the quality of propagation from TPC to ITS. These effects were estimated using the following two tests: one is the effect of different track selections on the corrected yield of the Λ_c, the other is the comparison of the TPC-ITS track matching efficiency in data and simulations.

The effect of the different track selections was estimated using the following set of cuts, for which all the corrected yield is computed. Only one cut at a time was varied compared to the central value. The following three track variations were tested: (i) additional cut on the number TPC crossed rows $> 120 - (5/p_{\mathrm{T}})$, (ii) number of TPC clusters $> 0.65\times$ number of TPC crossed rows, (iii) ratio of crossed rows over finable clusters in TPC > 0.9.

The ratio of the corrected yield for the different track cuts with the central point is shown on the top of Figure 3.16. Based on the variation of the distributions, a systematic uncertainty of 4% op to 8% was estimated. This corresponds to a 1.3% - 2.7% uncertainty per track, as the Λ_c is reconstructed from a three-body decay.

The second part of the tracking efficiency is the propagation from TPC to ITS matching efficiency. This TPC-ITS track matching efficiency is defined as the fraction of track with clusters is both ITS and TPC over the total number of track with clusters in the TPC. The three main reason for which a deviation could arise is: (i) there is mismatch between a track in the TPC and the ITS, (ii) as we require at least one hit in the SPD for the bachelor track, the track could still be present in the TPC and (iii) a secondary which is not rejected and not reconstructed in the ITS, but is reconstructed in the TPC. The systematic uncertainty arises from discrepancies between data and MC, and is particle dependent. Since the pion tracks of K^0_S does not have ITS requirements, this not taken into consideration. A transverse momentum dependent uncertainty varying from 1.8% - 3% for the proton was estimated, taken from the other decay channel $\Lambda_c \rightarrow pK\pi$.

A Monte Carlo simulation was used to propagate the uncertainty at the track level to Λ_c baryon level, accounting for the proton kinematic in the hadron p_{T} range of our analysis, see figure 3.16. The mean for each $p_{\mathrm{T}}(\Lambda_c)$ is taken from the $p_{\mathrm{T}}(p)$ versus $p_{\mathrm{T}}(\Lambda_c)$ distribution adding the corresponding uncertainty of the proton in quadrature to the uncertainty of the track quality computed above in order to get the final systematic uncertainty for the track efficiency.

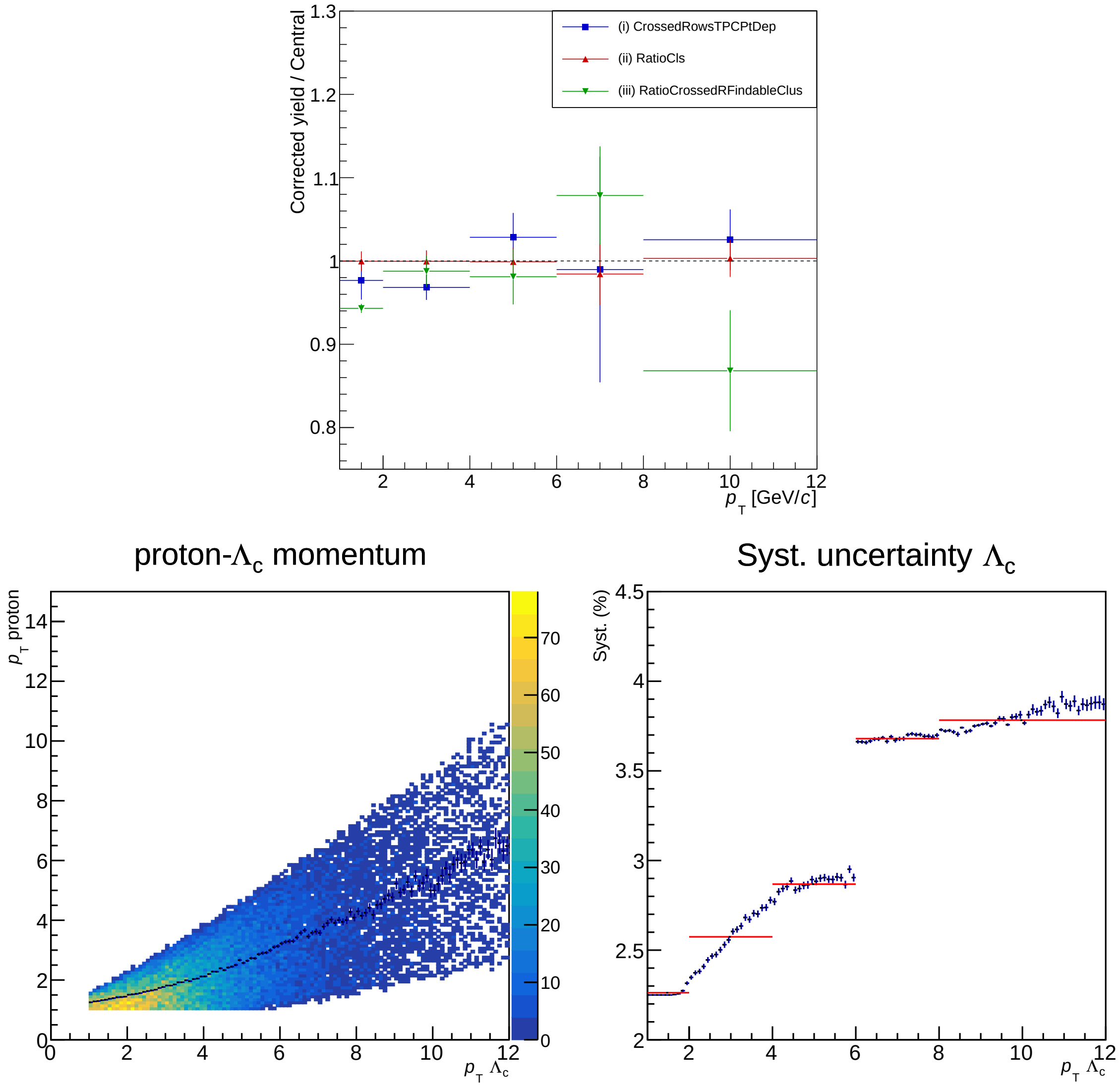


Figure 3.16: Top: Ratio of the central point corrected yield with the corrected yield calculated after varying the quality track cuts. Bottom left: Distribution of proton momentum versus Λ_c momentum, superimposed the mean for each Λ_c momentum bin. Bottom right: Right: Total systematic uncertainty for tracking efficiency.

3.10.7 Total systematic uncertainties

Table 3.5 and Figure 3.17 show the total systematic uncertainty discussed in this section. The various contributions are added in quadrature. In addition to the discussed uncertainties a 5% is assigned due to normalization and a 5.1% uncertainty for the the branching ratio.

p_{T} (GeV/c)	1 – 2	2 – 4	4 – 6	6 – 8	8 – 12
Yield extraction (%)	14	7	4	5	6
Cut variation (%)	$^{+4}_{-8}$	$^{+4}_{-8}$	$^{+4}_{-8}$	$^{+4}_{-8}$	$^{+4}_{-8}$
PID Efficiency (%)	5	5	5	5	5
MC p_{T} shape (%)	1	neg.	neg.	neg.	neg.
Feed-down (%)	$^{+1}_{-1}$	$^{+1}_{-1}$	$^{+2}_{-2}$	$^{+3}_{-3}$	$^{+4}_{-5}$
Tracking efficiency (%)	2	3	3	4	4
Total systematic uncertainty (%)	$^{+19}_{-20}$	$^{+14}_{-16}$	$^{+13}_{-15}$	$^{+14}_{-16}$	$^{+14}_{-16}$

Table 3.5: Systematic uncertainties (in percentages) evaluated for the Λ_c baryon.

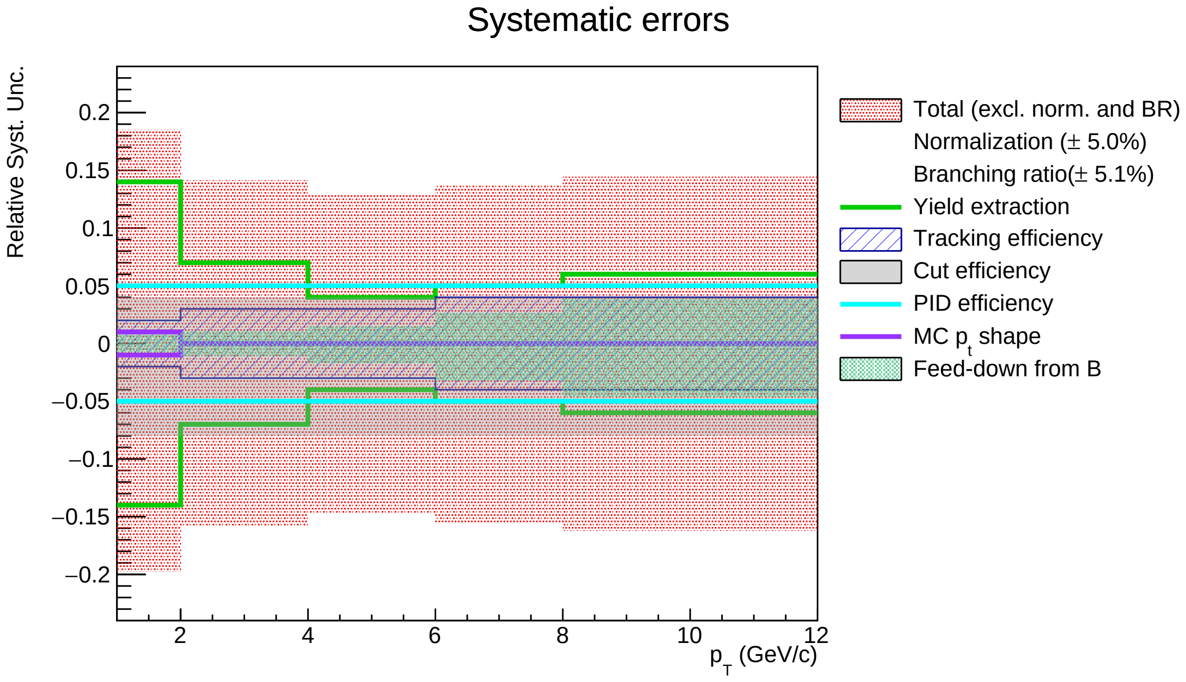


Figure 3.17: Relative systematic uncertainty evaluated for the Λ_c baryon. The different contributions are added in quadrature.

Chapter 4

Results

The p_{T}-differential corrected yield per event of prompt Λ_c^+ baryons integrated over multiplicity in $|y| < 0.5$ in pp collisions $\sqrt{s} = 13$ TeV as measured in the $\Lambda_c^+ \to pK_S^0$ decay channel is shown in Figure 4.1. In this and following figures the experimental results are represented by the black markers and is placed at the centre of the p_{T} inverval, the horizontal bars spans the width of the p_{T} interval, the vertical error bar is the statistical uncertainty, the blank box is the (asymmetric) systematic uncertainty and the filled box is the asymmetric systematic uncertainty coming from beauty feed-down.

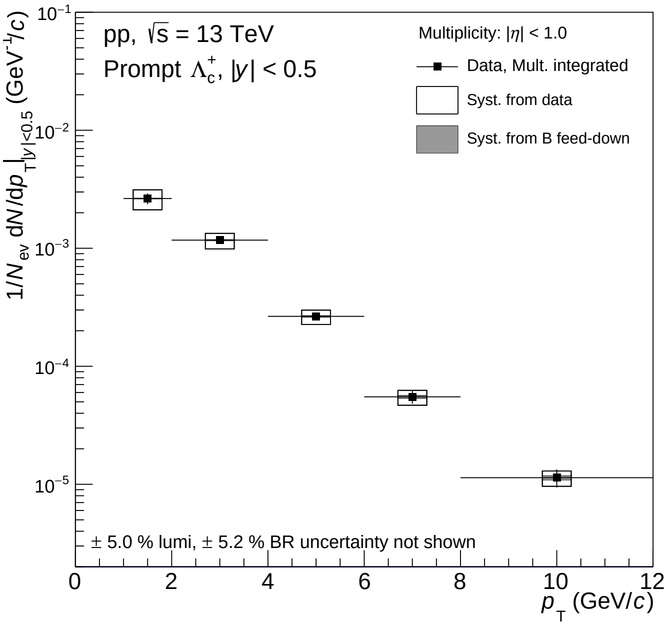


Figure 4.1: p_{T}-differential corrected yield per event of prompt Λ_c^+ baryons integrated over multiplicity in $|y| < 0.5$ in pp collisions $\sqrt{s} = 13$ TeV as measured in the $\Lambda_c^+ \to pK_S^0$ decay channel. The statistical uncertainties are shown as error bars, the systematic uncertainties from data as blank boxes and the systematic uncertainties from feed-down as filled boxes.

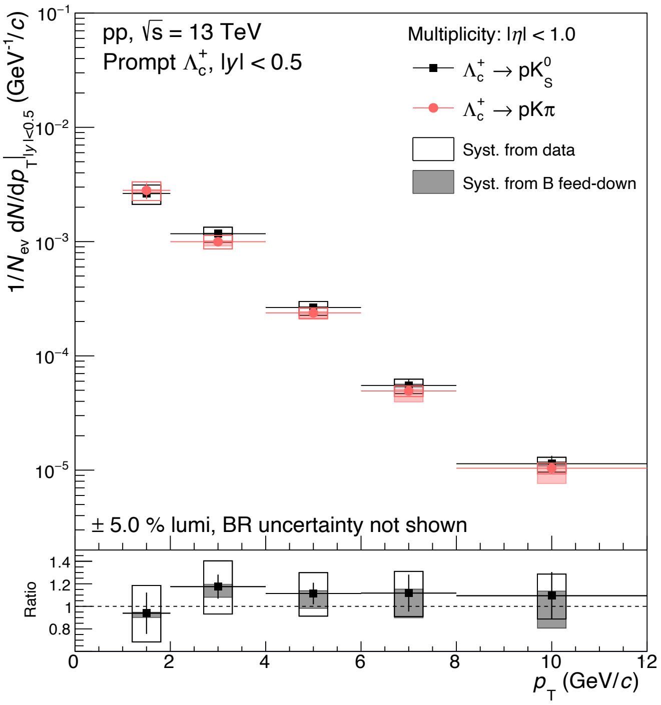


Figure 4.2: p_{T}-differential corrected yield per event of prompt Λ_c^+ baryons integrated over multiplicity in $|y| < 0.5$ in pp collisions $\sqrt{s} = 13$ TeV as measured in the $\Lambda_c^+ \to pK_S^0$ (black) and $\Lambda_c^+ \to pK\pi$ (red) decay channel, together with their ratio. The statistical uncertainties are shown as error bars, the systematic uncertainties from data as blank boxes and the systematic uncertainties from feed-down as filled boxes.

In order to validate these results, the data points of Figure 4.1 are compared with the p_{T}-differential corrected yield per event of prompt Λ_c^+ baryons for the other decay channel $\Lambda_c^+ \to pK\pi$. The comparison is shown in Figure 4.2 with the results from this analysis represented by the black markers and the ones from the three-prong decay denoted as red circles. The ratio is defined as the p_{T}-differential corrected yield per event of $\Lambda_c^+ \to pK_S^0$ over $\Lambda_c^+ \to pK\pi$, taking into account the correlation between the statistical and systematic uncertainties. As not all uncertainties of the $\Lambda_c^+ \to pK\pi$ decay channel were separately available (only the statistical, systematic uncertainty without feed-down and feed-down systematic uncertainty were provided) the following assumptions were made calculating the final uncertainties: the statistical uncertainties is calculated as uncorrelated, because they are reconstructed from (mostly) different tracks; the systematic uncertainty from feed-down is calculated as correlated, as it has been computed using the same calculations for the beauty quark becoming the Λ_c baryon; the systematic uncertainty of the others is calculated as uncorrelated as it dominated by the uncorrelated yield extraction and cut variation systematic uncertainties.

All the points, except the first p_{T} interval are systematically above the results for the other decay channel, but are within systematic uncertainties. The results of this analysis are on average 9% higher than the other decay channel. The measured p_{T}-differential corrected yield per event is thus within uncertainties with the other decay channel, however being systematically higher.

Figure 4.3 shows the result of the p_{T}-differential corrected yield per event of prompt Λ_c^+ baryons in $|y| < 0.5$ integrated over multiplicity in pp collisions $\sqrt{s} = 13$ TeV as measured in the $\Lambda_c^+ \to pK_S^0$ decay channel compared with different event generators.

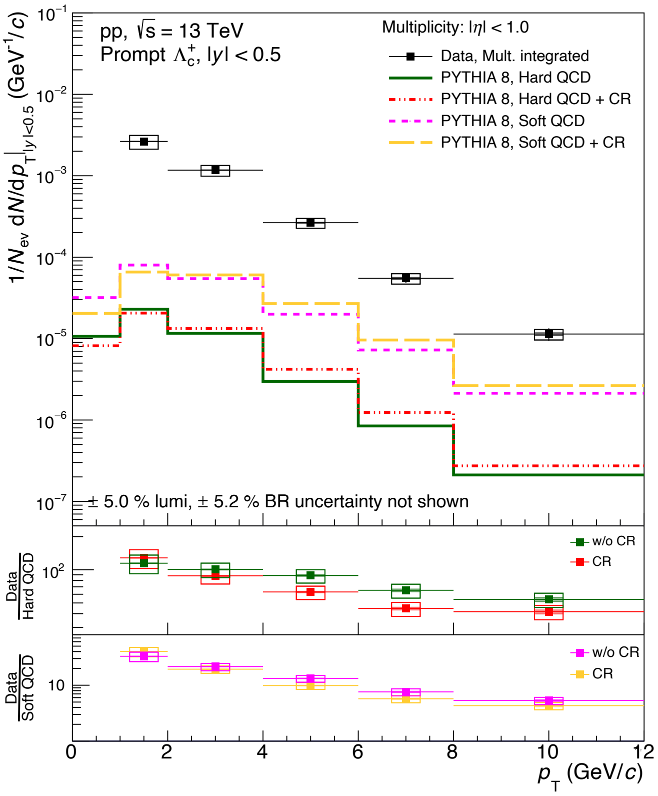


Figure 4.3: p_{T}-differential corrected yield per event of prompt Λ_c^+ baryons integrated over multiplicity in $|y| < 0.5$ in pp collisions $\sqrt{s} = 13$ TeV as measured in the decay channel $\Lambda_c^+ \to pK_S^0$. The measurements are compared with different event generators (see text for details).

The data are compared with expectations from the PYTHIA event generator, having two main processes: what is called SOFTQCD and HARDQCD within the jargon of the model. Within the framework of SOFTQCD, PYTHIA attempts to describe minimum bias physics and bulk observables through processes that involve multi-parton interactions. The processes described are mainly relying on phenomenological approaches that attempt to describe bulk observables in collisions between two protons. On the other hand, within the HARDQCD framework PYTHIA describes the formation of jets and heavy flavour partons through initial stage hard processes. In addition for each of this category of processes, we attempted to see the influence of the mechanism of colour reconnection (CR) to the production of charm baryons. The colour reconnection model in PYTHIA is applied prior to the hadronisation, and takes leading order colour strings and transforms a different colour based on three principles: (i) colour rules from QCD, (ii) space-time causal contact between strings and (iii) a measurement if possible reconnection is actually favoured. This addition should result into a baryon enhancement.

The ratio plots show the experimental corrected yield over the particular event generator, with and without colour reconnection. The green line and green markers denote HARDQCD without colour reconnection, the red dashed line and red markers denote HARDQCD with colour reconnection, the purple dashed line and purple markers denote SOFTQCD without colour reconnection, the yellow dashed line and yellow markers denote SOFTQCD with colour reconnection.

All curves underestimate the production of prompt Λ_c, HARDQCD by two orders of magnitude, SOFTQCD by one. There is clearly no significant difference between the curves with and without colour reconnection. Contrary to my initial expectation the models with SOFTQCD are closer to data.

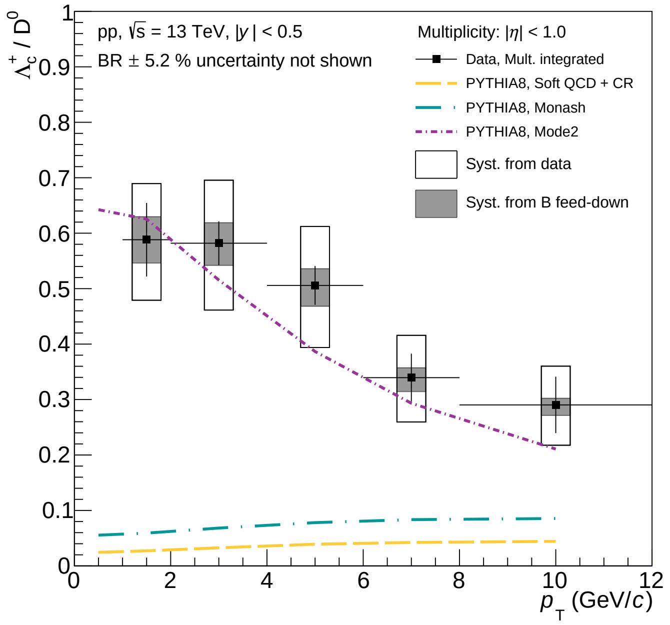


Figure 4.4: p_T-differential Λ_c^+/D^0 ratio in $|y| < 0.5$ in pp collisions $\sqrt{s} = 13$ TeV together with different tunes of PYTHIA8: SOFTQCD + CR [26], MONASH [26] and MODE2 [27]. The statistical uncertainties are shown as error bars, the systematic uncertainties from data as blank boxes and the systematic uncertainties from feed-down as filled boxes.

Figure 4.4 shows the p_T-differential Λ_c^+/D^0 ratio in $|y| < 0.5$ in pp collisions $\sqrt{s} = 13$ TeV. The D^0 corrected yield has been reconstructed in the same system and energy integrated over multiplicity with the decay channel $D^0 \to K^-\pi^+$ with branching ratio $3.89 \pm 0.04\%$. The following assumption were made calculating uncertainties: the statistical uncertainties are calculated as uncorrelated, as it is reconstructed from different tracks; the uncorrelated systematic uncertainties are Monte Carlo p_T-shape, raw-yield, branching ratio and cut variation, as these uncertainties are dependent on the studied particle type and used cuts; the correlated systematic uncertainties are particle identification, tracking efficiency, as these are related to the detector efficiencies; again the systematic uncertainty from beauty feed-down is correlated, as has been computed using the same calculations; the systematic uncertainties from normalization cancels out.

The four different models in Figure 4.3 showed all within statistical uncertainties the same qualitative trend as a function of p_T for the Λ_c/D^0 ratio, hence only the result of PYTHIA8, SOFTQCD + CR is shown. In addition to this, PYTHIA tunes MODE2 [26] and MONASH 2013 [27] are shown. As stated in [26] MODE2 is currently the best tune for colour reconnection. The enhanced colour reconnection mechanism of MODE2 increases the baryon-to-meson ratio in the charm sector and is in agreement with data for all p_T within the current level of the uncertainties. It is probably interesting to point out that this agreement is on the lower side. Clearly the other two models do not reproduce the data, either in magnitude or the p_T trend. Furthermore, the colour reconnection in SOFTQCD which should enhance baryon-to-meson ratio does not show the expected effect.

The charm baryon-to-meson ratio in pp collisions shows significant enhancement with respect to electron-positron calculation, which is flat at 0.12 [28, 29]. This suggest different underlying physics in hadronic collision compared to electron-positron annihilation. Colour reconnection and recombination are proposed as possible explanations. A more in depth study on the multiplicity dependent charm baryon-to-meson ratio can give more insight in the QGP like medium effects which were seen in the light flavour baryon-to-meson ratios.

Chapter 5

Conclusion and Outlook

The p_{T}-differential corrected yield of prompt Λ_c^+ charmed baryons in the hadronic decay mode $\Lambda_c^+ \to pK_S^0$ is measured with the ALICE detector at the Large Hadron Collider (LHC) in minimum bias proton-proton collisions at $\sqrt{s} = 13$ TeV at midrapidity in the transverse momentum range $1 < p_{\mathrm{T}} < 12$ GeV/c. A machine learning optimization algorithm has been used in order to identify the signal of the low produced Λ_c baryons which is covered under a dominating background. The machine learning optimization algorithm showed improvement in the signal extraction, and is favoured in future analyses.

The results were in agreement with corrected yield of prompt Λ_c^+ charmed baryons in $\Lambda_c^+ \to pK\pi$, though on average systematically higher. The prediction from SOFTQCD and HARDQCD event generators underestimated the measured prompt Λ_c^+ production, on average by 1 and 2 orders of magnitude respectively. As my initial expectation expected HARDQCD to be closer to data than SOFTQCD, a more in depth analysis on the different tunes of the generators is recommended.

We also measured p_{T}-differential the baryon-to-meson ratio Λ_c^+/D^0 and compared it to different pp event generators. All tunes of PYTHIA8 examined underestimate the ratio with the exception of MODE2 that is able to describe the data within the current level of uncertainties.

This project started with the plan to study more multiplicity ranges than only multiplicity integrated, but due to the lack of time this could not be done. A study on the different multiplicity ranges, especially the high and low multiplicity, will give us more insight in the multiplicity effect of the charmed baryon-to-meson ratio and can validate and improve the already preliminary results in the other hadronic decay channel. When continuing with the other multiplicities an improvement must be made in signal extraction in the highest transverse momentum interval, which was left out in this analysis.

Recently the PWG-HF group started using machine learning techniques to improve the signal extraction. Within the group multiple machine learning techniques were used, but none of them has done an extensive study on their technique, and drawn conclusions about which technique is best for this purpose. I would suggest to do an extensive study on the different machine learning techniques, and show qualitatively what the leading technique is. Currently to many people are playing around with their own technique in which they trust. The ALICE collaboration has to bundle the knowledge, spend some time in understanding the known techniques and should come with a final plan which is clear for everyone in the collaboration. It is 2020, machine learning is a technique of the future, in a couple of years no scientific field cannot make use of it.

Acknowledgements

The past year I have spend most of my time doing research for this thesis. Looking back at this year I cannot imagine how time consuming the tasks the first month were which I now can do in a couple of days. The past year I have gained a lot of skills, experienced to be part of a huge collaboration and tasted what it is like to be a scientist.

All the work could not be achieved without my supervisors Panos and Davide. Their help and support made it possible to complete this project. Panos, your expertise and experience helped me a lot getting on the right track, I really would like to thank you for your commitment to me. Davide, you were always there to answer my questions. The time and effort you have spend in helping me and the amount of questions you have answered can not be described in words. You were the best daily supervisor I could imagine, and a great person too.

I also want thank all the people that worked I with, all the people in Utrecht but especially the ones at Nikhef. I will miss the discussions during coffee and lunch breaks. A special thanks to my all time best office-mates Davide and Zhanna. You have two great personalities, and I wish you both a prosperous life and career.

I would like to thank my family and girlfriend. Starting a year long master thesis two weeks after the loss my father has been a tough and challenging project. I could never have made this without your support.

Last and definitely no least, I would like to thank my father, without you I would never have studied physics and be so interested in natural sciences. I miss the long discussion that we could have about natural science related topics, the universe and about the stars in it. You would have been extremely proud of what I have achieved, I miss you.

Bibliography

[1] B. Mohanty, "Exploring the QCD phase diagram through high energy nuclear collisions: An overview," *PoS* **CPOD2013** (2013) 001, `arXiv:1308.3328 [nucl-ex]`.

[2] S. Borsanyi, G. Endrodi, Z. Fodor, A. Jakovac, S. D. Katz, S. Krieg, C. Ratti, and K. K. Szabo, "The QCD equation of state with dynamical quarks," *JHEP* **11** (2010) 077, `arXiv:1007.2580 [hep-lat]`.

[3] P. Rosnet, "Quark-Gluon Plasma: from accelerator experiments to early Universe," in *11th Rencontres du Vietnam: Cosmology: 50 years after CMB discovery Quy Nhon, Vietnam, August 16-22, 2015.* 2015. `arXiv:1510.04200 [hep-ph]`.

[4] **ALICE** Collaboration, J. Adam *et al.*, "Centrality dependence of the nuclear modification factor of charged pions, kaons, and protons in Pb-Pb collisions at $\sqrt{s_{\rm NN}} = 2.76$ TeV," *Phys. Rev.* **C93** no. 3, (2016) 034913, `arXiv:1506.07287 [nucl-ex]`.

[5] L. Vermunt, "Summary of one year at cern: Heavy-flavour hadronisation in pp and pb-pb." Sap meeting, 2019.

[6] J. L. Nagle and W. A. Zajc, "Small System Collectivity in Relativistic Hadronic and Nuclear Collisions," *Ann. Rev. Nucl. Part. Sci.* **68** (2018) 211–235, `arXiv:1801.03477 [nucl-ex]`.

[7] **ALICE** Collaboration, S. Acharya *et al.*, "Multiplicity dependence of light-flavor hadron production in pp collisions at $\sqrt{s} = 7$ TeV," *Phys. Rev.* **C99** no. 2, (2019) 024906, `arXiv:1807.11321 [nucl-ex]`.

[8] **ALICE** Collaboration, P. Dupieux, B. Joly, F. Jouve, S. Manen, and R. Vandaele, "Upgrade of the ALICE muon trigger electronics," *JINST* **9** (2014) C09013.

[9] **ALICE** Collaboration, K. Aamodt *et al.*, "The ALICE experiment at the CERN LHC," *JINST* **3** (2008) S08002.

[10] **ALICE SPD** Collaboration, S. Moretto, "The silicon pixel detector for ALICE experiment," *Italian Phys. Soc. Proc.* **96** (2008) 263–265.

[11] S. Beole *et al.*, "The ALICE silicon drift detectors: Production and assembly," *Nucl. Instrum. Meth.* **A582** (2007) 733–738.

[12] **ALICE** Collaboration, G. Contin, "The ALICE Silicon Strip Detector performance during the first LHC data taking," *Conf. Proc.* **C100901** (2010) 257–260, `arXiv:1101.2776 [physics.ins-det]`.

[13] **ALICE TPC** Collaboration, C. Lippmann, "The Time Projection Chamber for the ALICE Experiment," `arXiv:0809.5133 [nucl-ex]`.

[14] J. Wiechula, "Everything you wanted to know about the tpc but were afraid to ask," 3, 2016. `https://alice-analysis.web.cern.ch/sites/alice-analysis.web.cern.ch/files/documents/Analysis/JensJD.pdf`. Universitt Tbingen.

[15] **ALICE** Collaboration, B. B. Abelev *et al.*, "Performance of the ALICE Experiment at the CERN LHC," *Int. J. Mod. Phys.* **A29** (2014) 1430044, `arXiv:1402.4476 [nucl-ex]`.

[16] **ALICE** Collaboration, J. Adam *et al.*, "Determination of the event collision time with the ALICE detector at the LHC," *Eur. Phys. J. Plus* **132** no. 2, (2017) 99, `arXiv:1610.03055 [physics.ins-det]`.

[17] **ALICE** Collaboration, E. Abbas *et al.*, "Performance of the ALICE VZERO system," *JINST* **8** (2013) P10016, `arXiv:1306.3130 [nucl-ex]`.

[18] **Particle Data Group** Collaboration, M. Tanabashi *et al.*, "Review of Particle Physics," *Phys. Rev.* **D98** no. 3, (2018) 030001.

[19] T. Sjostrand, S. Mrenna, and P. Z. Skands, "A Brief Introduction to PYTHIA 8.1," *Comput. Phys. Commun.* **178** (2008) 852–867, `arXiv:0710.3820 [hep-ph]`.

[20] **ALICE** Collaboration, J. Adam *et al.*, "Measurement of charm and beauty production at central rapidity versus charged-particle multiplicity in proton-proton collisions at $\sqrt{s} = 7$ TeV," *JHEP* **09** (2015) 148, `arXiv:1505.00664 [nucl-ex]`.

[21] **ALICE** Collaboration, K. Aamodt *et al.*, "Strange particle production in proton-proton collisions at $\sqrt{s} = 0.9$ TeV with ALICE at the LHC," *Eur. Phys. J.* **C71** (2011) 1594, `arXiv:1012.3257 [hep-ex]`.

[22] **ALICE** Collaboration, S. Acharya *et al.*, "Measurement of D-meson production at mid-rapidity in pp collisions at $\sqrt{s} = 7$ TeV," *Eur. Phys. J.* **C77** no. 8, (2017) 550, `arXiv:1702.00766 [hep-ex]`.

[23] A. Hocker *et al.*, "TMVA - Toolkit for Multivariate Data Analysis," `arXiv:physics/0703039 [physics.data-an]`.

[24] **ALICE** Collaboration, S. Acharya *et al.*, "Λ_c^+ production in pp collisions at $\sqrt{s} = 7$ TeV and in p-Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV," *JHEP* **04** (2018) 108, `arXiv:1712.09581 [nucl-ex]`.

[25] M. Cacciari, M. Greco, and P. Nason, "The p_{T} spectrum in heavy flavor hadroproduction," *JHEP* **05** (1998) 007, `arXiv:hep-ph/9803400 [hep-ph]`.

[26] J. R. Christiansen and P. Z. Skands, "String Formation Beyond Leading Colour," *JHEP* **08** (2015) 003, `arXiv:1505.01681 [hep-ph]`.

[27] P. Skands, S. Carrazza, and J. Rojo, "Tuning PYTHIA 8.1: the Monash 2013 Tune," *Eur. Phys. J.* **C74** no. 8, (2014) 3024, `arXiv:1404.5630 [hep-ph]`.

[28] **CLEO** Collaboration, P. Avery *et al.*, "Inclusive production of the charmed baryon Lambda(c) from e^+ e^- annihilations at $\sqrt{s} = 10.55$ GeV," *Phys. Rev.* **D43** (1991) 3599–3610.

[29] G. S. Abrams *et al.*, "Observation of Charmed Baryon Production in e^+ e^- Annihilation," *Phys. Rev. Lett.* **44** (1980) 10.