DISSERTATION

submitted to the

Combined Faculty of Mathematics, Engineering and Natural Sciences

of Heidelberg University, Germany

for the degree of

Doctor of Natural Sciences

Put forward by

Federica Zanone

born in Gattinara (VC), Italy

Oral examination: 03.07.2025

Measurement of prompt production of Ξ_c^0 in proton-proton collisions at $\sqrt{s} = 13.6$ TeV with ALICE

Referees:

Dr. Kai Schweda

Prof. Dr. Ulrich Uwer

Abstract

At LHC energies, charmed-baryons are copiously produced. The recent upgrade of the ALICE experiment for Run 3 improved the vertexing capabilities and allows to operate at an increased interaction rate and in continuous readout mode, thus allowing for collecting a significantly larger dataset compared to Run 2. This opens a new dimension of precise charmed-baryon measurements in several decay channels, helping to shed light on the mechanisms responsible for the production and decay of these particles. This topic became of crucial interest after recent studies by ALICE challenged the assumption of the universality of charm hadronisation processes across different collision systems. Moreover, the decay of charmed-baryons is still poorly understood and measurements of corresponding branching ratios pose a challenge to all models.

This thesis presents the measurement of the prompt cross section of the Ξ_c^0 baryon using a dataset collected by the ALICE experiment. The weak decay $\Xi_c^0 \to \Xi^- \pi^+ + \mathrm{c.c.}$ is reconstructed using minimum bias proton-proton collisions data corresponding to 0.8 pb^{-1} of integrated luminosity recorded at a centre-of-mass-energy of 13.6 TeV in 2022. The measurement is performed at midrapidity and in the transverse momentum interval $1 < p_\mathrm{T} < 12$ GeV/c. The Run 3 ALICE analysis framework has been optimized to process large-size datasets and a new strategy to produce realistic charm-enriched Monte Carlo has been developed and implemented. The simulation used in this analysis is anchored to the corresponding data taking conditions and constrained by real data in a data-driven mode. The analysis is validated with a closure test. However, a $\sim 40\%$ p_T-independent discrepancy is observed between the Run 3 measurement and the corresponding Run 2 result, pointing to unresolved issues in data reconstruction and/or Monte Carlo production.

In Run 3, besides permanently storing a small fraction of minimum bias datasets, ALICE relies on the implementation of software triggers to investigate the huge amount of proton-proton collisions delivered by the LHC, while at the same time applying an efficient data storage procedure. This allows to investigate an integrated luminosity which is more than three orders of magnitude larger compared to Run 2. The ALICE software triggers have already processed an integrated luminosity of 82 pb^{-1}, and an additional sample corresponding to more than 100 pb^{-1} is expected to be analysed by the end of Run 3. Within this work, a dedicated trigger has been developed and applied to proton-proton data collected in 2023 and 2024 to reconstruct Cabibbo-suppressed Ω_c^0 decays. This software trigger is dedicated to the reconstruction of the singly-suppressed mode $\Omega_c^0 \to \Xi^- \pi^+ + \mathrm{c.c.}$ and of the doubly-suppressed decay $\Omega_c^0 \to \Xi^- \mathrm{K}^+ + \mathrm{c.c.}$ and it will allow for relative branching fraction measurements in the coming years.

Zusammenfassung

Bei LHC-Energien werden Baryonen mit Charm-Quarks in großer Zahl produziert. Das Upgrade des ALICE-Experiments für Run 3 ermöglicht einen Betrieb mit erhöhter Kollisionsrate in einem kontinuierlichen Auslesemodus, so dass im Vergleich zu Run 2 ein wesentlich größerer Datensatz gesammelt werden kann. Damit eröffnet sich eine neue Dimension präziser Messungen von Baryonen mit Charm-Quarks in verschiedenen Zerfallskanälen. Diese Messungen tragen zu einem tieferem Verständnis der Mechanismen, die für die Produktion und den Zerfall dieser Teilchen bei. Dies ist von großem Interesse, nachdem aktuelle ALICE-Studien die Annahmen einer Universalität der Hadronisierungsprozesse in verschiedenen Kollisionssystemen in Frage gestellt haben. Darüber hinaus ist der Zerfall von Baryonen mit Charm-Quarks noch nicht vollständig verstanden und Messungen der entsprechenden Verzweigungsverhältnisse stellen eine Herausforderung für alle Modelle dar.

In dieser Dissertation wird die Messung des prompten Wirkungsquerschnitts des Ξ_c^0-Baryons anhand eines vom ALICE-Experiment aufgenommenen Datensatzes vorgestellt. Der schwache Zerfall $\Xi_c^0 \rightarrow \Xi^- \, \pi^+ \, + \, \mathrm{c.c.}$ wird anhand von *minimum bias* Proton-Proton-Kollisionsdaten rekonstruiert, die 0.8 pb^{-1} integrierter Luminosität entsprechen und im Jahr 2022 bei einer Schwerpunktsenergie von 13.6 TeV aufgezeichnet wurden. Die Messung wird bei zentraler Rapidität und im Transversalimpulsintervall $1 < p_{\mathrm{T}} < 12$ GeV/c durchgeführt. Das Analyse Framework für Run 3 wurde für die Verarbeitung großer Datensätze optimiert und eine neue Strategie zur Erstellung realistischer, mit Charm angereicherter Monte-Carlo-Simulationen entwickelt und umgesetzt. Die in dieser Analyse verwendete Simulation berücksichtigt die jeweiligen Bedingungen der Datenaufzeichnung und wird in einem datengetriebenen Ansatz durch reale Daten eingeschränkt. Die Analyse wird mit einem Closure Test validiert. Unabhängig vom Transversalimpuls wird jedoch eine Diskrepanz von ca. 40% zwischen der Messung in Run 3 und dem entsprechenden Ergebnis in Run 2 beobachtet, was auf offene Fragen bei der Datenrekonstruktion und/oder der Monte-Carlo-Produktion hinweist.

Neben der permanenten Speicherung eines kleinen Teils der minimum bias Daten stützt sich ALICE in Run 3 auf die Implementierung von Software-Triggern, um die riesige Menge an Proton-Proton Kollisionen zu untersuchen, die der LHC liefert, und gleichzeitig ein effizientes Datenspeicherverfahren anzuwenden. Dies erlaubt die Untersuchung einer um viele Größenordnungen höheren integrierten Luminosität als in Run 2. Im Rahmen dieser Arbeit wurde ein spezieller Trigger entwickelt und auf Proton-Proton Daten, die 2023 und 2024 aufgenommen wurden, angewendet, um Cabibbo-unterdrückte Ω_c^0-Zerfälle zu rekonstruieren. Dieser Software-Trigger dient der Rekonstruktion des einfach-unterdrückten Zerfalls $\Omega_c^0 \rightarrow \Xi^- \, \pi^+ \, + \, \mathrm{c.c.}$ und des doppelt-unterdrückten Zerfalls $\Omega_c^0 \rightarrow \Xi^- \, \mathrm{K}^+ \, + \, \mathrm{c.c.}$ und wird in den kommenden Jahren eine Messung der relativen Verzweigungsverhältnisse ermöglichen.

Contents

Chapter 1

Introduction

According to the Big Bang model, space-time was created ~ 14 billion years ago. Back then our universe was concentrated in an extremely small, hot, and dense region of space. Then it expanded, becoming cooler and less dense. During the first microseconds of its life, its density was too high to allow for hadron formation: quarks and gluons existed in a deconfined state, the quark-gluon plasma. When the energy density and the temperature reached the critical values of $\epsilon_{\mathrm{c}} \sim 1\ \mathrm{GeV/fm^3}$ and $T_{\mathrm{c}} \sim 160\ \mathrm{MeV}$, colour charges combined to form colour-neutral objects of $\sim 1\ \mathrm{fm}$ in size, the hadrons. This transition is called *hadronisation*. Then, few minutes after the Big Bang, the temperature fell below ~ 100 keV, allowing for the formation of atomic nuclei (primordial nucleosynthesis). At this stage, the universe was still ionized, and therefore opaque to electromagnetic radiation. Only $\sim 300\,000$ years after the Big Bang the temperature became low enough (~ 3000 K) such that electrons and ions could combine to form atoms.

Understanding the hadronisation process is crucial to improve our view of the early stages of the universe and to explain how fundamental particles manifest as the matter we can detect. Prior to the phase transition from quark-gluon plasma to hadronic matter, the universe was essentially structureless, consisting of a hot, dense soup of fundamental particles. It was only after hadronisation that structures began to emerge, first in the form of hadrons, and later through the formation of atoms, molecules, and eventually complex matter. However, the opacity of the universe for times prior to the decoupling of the electromagnetic radiation makes it impossible to study the phenomena that occurred before the formation of atoms. Nevertheless, it is possible to recreate such processes in high-energy collisions at particle accelerators. The theory that has been developed to describe the interactions between quarks and gluons, which are responsible for the creation of hadrons, is quantum chromodynamics (QCD). Due to the nature of such interactions, it is impossible to calculate hadronisation from first-principles QCD. Despite decades of research, hadronisation remains incompletely understood and this makes it an active and important area of research in particle physics.

The structure of the thesis is as follows: Chapter 2 introduces the general concepts of QCD and the classification of charmed-baryons, describes the modelling of the open charm production process, presents the different hadronisation models and relevant experimental results. A general introduction to the Large Hadron Collider and to the ALICE detector, as well as the description of the relevant data reconstruction techniques, are given in Chapter 3. Chapter 4 introduces the ALICE analysis framework and presents the software developments implemented to perform charmed-baryon analyses. In Chapter 5, the measurement of the Ξ_c^0 prompt cross section is described. First, the preselections and the machine learning algorithm applied in the analysis are presented. Then, the signal extraction, the acceptance-times-efficiency correction and the prompt-fraction estimation are discussed. Afterwards, the systematic uncertainties are described. Finally, the results are presented, including a comparison to the previously published result, the outcome of the closure test and the comparison to theoretical models. Conclusions and outlook are given in Chapter 6.

Chapter 2

Open charm production

2.1 Quantum Chromodynamics

Quantum chromodynamics (QCD) is the quantum field theory of the strong interaction. Such interaction is mediated by eight massless gluons corresponding to the eight generators of the SU(3) local gauge symmetry of the theory. The conserved charge associated to QCD is the *colour*, with red, green and blue ($r,\ g,\ b$) being the labels of the SU(3) colour space orthogonal states. While quarks carry a colour charge, anti-quarks carry an anti-colour charge ($\bar{r},\ \bar{g},\ \bar{b}$). Since the theory predicts the existence of vertices where a gluon connects quarks of different colours, in order for the colour charge to be conserved at the interaction vertex, gluons must carry simultaneously both a colour and an anti-colour charge. Only particles carrying (anti-)colour charge, namely quarks and gluons themselves, couple to gluons.

The coupling constant of QCD is α_s and its value evolves with the energy scale Q^2 of the considered process. The *running* of α_s accounts for the contribution to a given QCD process coming from Feynman diagrams containing quark and/or gluon loops. Such higher-order corrections are absorbed in the definition of an effective strong coupling strength, $\alpha_s(Q^2)$. While colour-neutral virtual $\mathrm{q\bar{q}}$ pairs from fermionic loops screen the interacting colour charges, coloured virtual gluons from bosonic loops lead to an *anti-screening* effect, and the number of colour charges and quark flavours determines which component dominates. For a theory predicting the existence of three colour charges and six quark flavours, the net effect of the presence of virtual particle clouds is an increase in the colour field, hence a dominance of the anti-screening phenomenon. For this reason, decreasing the distance from a quark results in a weakening of the effective charge arising from the anti-screening effect of the surrounding gluons, and vice versa. QCD is therefore an *asymptotically free* theory, with the effective coupling constant α_s decreasing at increasingly high values of momentum transfer and diverging at small Q^2.

A consequence of such behaviour of α_s is the phenomenon of *colour confinement*, according to which coloured objects can not propagate as free particles over long distances but are found confined into colour-neutral bound states, namely the hadrons. This hypothesis explains the lack of evidence for free quarks in processes with small Q^2 values. The experimental observation of colour confinement is ascribed to gluon-gluon self-interactions, as they give rise to attractive forces between the virtual gluons exchanged between two strongly interacting quarks. This has the result of squeezing the corresponding colour field into a tube and, when the two interacting quarks are at relatively large distances, the energy density in the tube is constant. As a consequence, the energy stored in the field must be proportional to the distance r between the two quarks. Due to the linear dependence between the quark separation and the energy stored in the exchanged colour field, to free a quark, or equivalently to separate the two quarks to infinity, an infinite amount of energy would be required. The evolution of α_s with the process energy scale is experimentally well established [1]. The coupling constant is $\alpha_s \sim \mathcal{O}(1)$ at $|Q| \sim 1$ GeV, which is the same scale $\Lambda_{\rm QCD}$ at which hadrons are observed. Hence, the confinement radius is $\propto \frac{1}{\Lambda_{\rm QCD}} \sim$ fm, which implies that in such low-energy processes quarks and gluons can not travel distances exceeding this size. From the asymptotic freedom it follows that perturbative QCD is applicable only at high-energy regimes, for $Q^2 \gtrsim 1~{\rm GeV}^2/c^2$. On the other hand, the phenomenon of colour confinement is observed at large distances or, equivalently, small Q^2 values, therefore the non-perturbative regime applies to the discussion of the hadron formation process. This short introduction to the QCD aspects relevant for this work is mainly summarised from [2].

2.2 Charmed-baryons classification

In accordance with the naming conventions [3], the charmed-baryon base symbol is determined by its isospin I and charm-strangeness $C + S$ quantum numbers. The charmed-baryons Λ_c, Σ_c, $\Xi_{c,cc}$ and $\Omega_{c,cc,ccc}$ correspond to the quantum numbers $I(C+S)$ 0(1), 1(1), 1/2(2) and 0(3), respectively. Therefore, the Λ is an isospin singlet with one "heavy" (s or c) quark, the Σ a triplet with one heavy quark, the Ξ a doublet with two heavy quarks and the Ω an isospin singlet containing three heavy quarks. Unlike the charm quark, the strange quark is usually not considered heavy, this lexical choice is dictated by the need to illustrate the baryon naming convention and is to be regarded as valid only in the context of the sentence above.

According to the Quark Model [3], the ${\rm SU}(4)_{flavour}$ representation is useful for bookkeeping purposes to classify the various charmed-baryons, even if it does not reflect an observed symmetry due to the large mass of the charm quark. Based on such description, charmed-baryons can be represented in a three-dimensional space with coordinates $(I_z,\ Y,\ C)$, where I_z is the third component of the baryon isospin, C is the charm-flavour quantum number and Y is the hypercharge. The latter is in general defined as $Y = \mathcal{B} + S - \frac{C-B+T}{3}$, where $\mathcal{B}$ is the baryon number and S, C, B, T indicate respectively the strangeness, charm, bottomness and topness content of the baryon. However, the lifetime of the top quark is of the order of $\sim 10^{-25}$ s, while the hadronisation process requires $\sim 10^{-23}$ s. Therefore, the top quark lifetime is too short for it to hadronise. The charmed-baryons representation in the $(I_z,\, Y,\, C)$ space is depicted in Fig. 2.1. Given four flavours, namely u, d, s and c, sixty-four configurations of three quarks are possible.

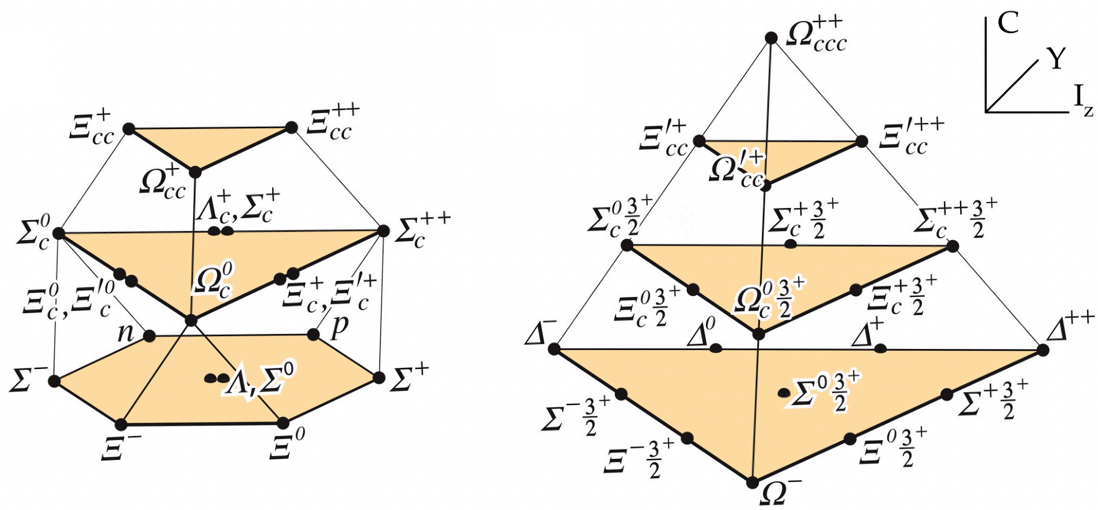


Figure 2.1: Multiplets of ground state baryons made of u, d, s and c quarks in the $(I_z,\, Y,\, C)$ space [3]. The $J = 1/2$ multiplet is shown on the left side, the $J = 3/2$ multiplet on the right side.

In this thesis, only singly-charmed-baryons ($C = 1$) are discussed. These hadrons can be described as a bound objects composed of a charm quark and a light diquark whose content is limited to three possible flavours (u, d, s). The $\mathrm{SU}(3)_{flavour}$ tensor product of the two light quarks gives $\mathbf{3}\otimes\mathbf{3} = \mathbf{6}+\bar{\mathbf{3}}$, with the multiplet $\mathbf{6}$ being symmetric under the interchange of the two light quark flavours and the multiplet $\bar{\mathbf{3}}$ being antisymmetric under this interchange. For ground state baryons, i.e. states with zero angular momentum, the overall antisymmetry of the hadron wave function implies that the light diquark has to be symmetric under the exchange of spin and flavour. The possibilities are therefore limited to either a symmetric spin-1 diquark belonging to the symmetric flavour-$\mathbf{6}$ multiplet or to an antisymmetric spin-0 diquark belonging to the antisymmetric flavour-$\bar{\mathbf{3}}$ multiplet. The spin-0 $\bar{\mathbf{3}}$ combines with the charm quark to create the $J = 1/2$ ground state baryons while the spin-1 $\mathbf{6}$ combines to form either $J = 1/2$ or $J = 3/2$ ground states. The structure in the $(I_z,\, Y)$ space of the $J = 1/2$ singly-charmed-baryon ground states is represented in Fig. 2.2.

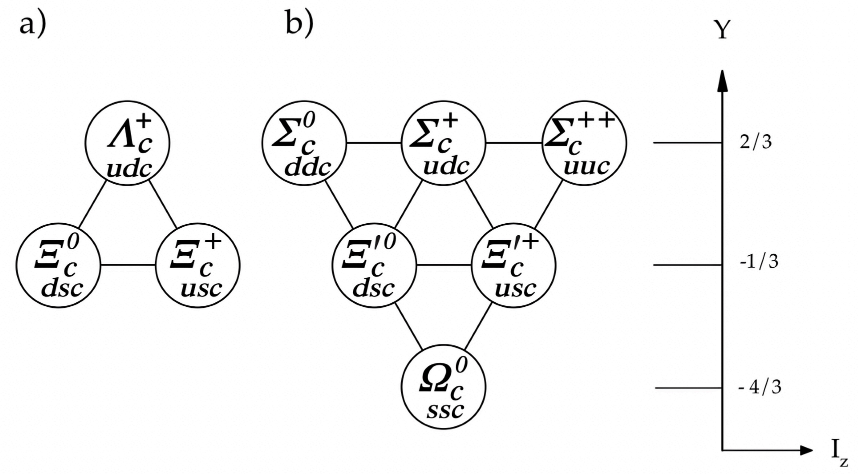


Figure 2.2: Schematic representation of ground state $J = 1/2$ singly-charmed-baryons corresponding to the spin-0 flavour-$\bar{\mathbf{3}}$ diquark (a) and spin-1 flavour-$\mathbf{6}$ diquark (b) in the (I_z, Y) space. The structure of the ground state $J = 3/2$ representation containing the spin-1 flavour-$\mathbf{6}$ diquark is identical to the one in (b) [3].

2.3 Factorisation approach

The study of hadrons presenting among their constituent quarks charm or anti-charm is an effective probe to test QCD calculations based on the factorisation theorem and to investigate the hadronisation mechanisms. Moreover, measurements performed in pp collisions allow to establish a benchmark for Pb–Pb collisions and thus to infer properties of the quark-gluon plasma (QGP) produced in heavy-ion collisions.

Due of the large charm quark mass $m_c = (1.2730 \pm 0.0046)$ GeV/c^2 [4], charmed-hadrons are always the result of the hadronisation of a c or $\bar{c}$ quark produced prior to hadronisation, either in the initial hard-scatterings or via gluon splitting in parton showers. Such processes can involve large momentum transfers ($Q^2 > (2m_c)^2$), unlike the soft interactions occurring at hadronisation time. The hadronisation process determines the relative abundances of the various charmed-hadron species and their kinematics, but does not affect the total charmed-hadrons production. The latter matches the production rate of the parent charm quark, therefore charmed-hadrons are considered a calibrated probe of the hadronisation process [5]. The theoretical description of prompt charmed-hadron production in pp collisions relies on the factorisation approach in terms of the squared momentum transfer Q^2 [6] according to the formula

$$\sigma_{\mathrm{AB}\to\mathrm{H_c}} = \mathrm{PDF}(x_a, Q^2)\mathrm{PDF}(x_b, Q^2) \otimes \sigma_{ab\to c\bar{c}}(x_a, x_b, Q^2) \otimes D_{c\to\mathrm{H_c}}(z, Q^2). \tag{2.1}$$

In Eq. 2.1, the cross section describing the production of the charmed-hadron H_c in proton-proton collisions $\sigma_{\mathrm{AB}\to\mathrm{H}_c}$ is computed as the convolution of the parton distribution functions (PDFs) of the incoming protons, the partonic cross section describing the production of the charm–anti-charm pair $\sigma_{ab\to c\bar{c}}$, and the fragmentation function $D_{c\to\mathrm{H}_c}$. Due to the large mass of the charm quark $m_c >> \Lambda_{\mathrm{QCD}}$, the large momentum transfer required to produce the $c\bar{c}$ pair ensures that the $c\bar{c}$ production cross section can be computed as a perturbative series in powers of the strong coupling constant α_s down to zero momentum. On the other hand, the PDFs and the fragmentation functions describe non-perturbative processes, therefore they are derived through parametrisation of experimental results. The PDFs give the probability of finding a parton a in the incoming proton A as a function of the fraction x of the hadron momentum carried by the parton itself for a process with squared momentum transfer Q^2. The PDFs are extracted from deep inelastic scattering measurements and their evolution with Q^2 is described by the DGLAP equations [7]. The fragmentation functions describe the probability that the quark c gives rise to the hadron H_c carrying a fraction z of the parent quark momentum. These functions have been parametrised for more than four decades from measurements performed in $\mathrm{e}^+\mathrm{e}^-$ and ep collisions and, under the assumption that they are "universal" based on the time scale separation between the $c\bar{c}$ pair production and the hadronisation process, they have been applied to all the colliding systems [5]. Such assumption has recently been challenged by ALICE measurements, as discussed in Section 2.4.

Measurements of charmed-hadron p_{T}-differential cross sections allow both to test the factorisation approach itself and to validate the ingredient used in the prediction. Additionally, such measurements are used to measure the total $c\bar{c}$ production cross section by summing the the cross sections of the measured ground states spin-0 charmed-mesons and spin-1/2 charmed-baryons [8]. They also allow to trace back to the fragmentation fractions of the different charmed-hadrons by dividing the p_{T}-integrated cross section of each measured hadron species by the total charm cross section [8].

The ALICE detector is well suited for the study of charmed-hadrons thanks to its high-precision tracking and particle identification capabilities. At midrapidity, the low material budget characterizing its tracking system together with the intensity of the solenoidal magnetic field applied allow to focus on the low and moderate transverse momentum regions, making this experiment design complementary to the other LHC detectors. This kinematic range is crucial for hadronisation studies, as it will be discussed in Section 2.4.

2.4 Charm hadronisation

Hadronisation is the process by which partons, namely quarks and gluons, combine to form colour-neutral hadrons. In the early universe, this transition is supposed to have taken place $\sim 10\ \mu$s after the Big Bang [9]. In conditions of energy density typical of ordinary matter, the direct observation of free colour charges is prevented by the phenomenon of colour confinement, but the hadronisation process is accessible in laboratory in high-energy collisions at particles accelerators.

With reference to Eq. 2.1, the component that is sensitive to hadronisation is the fragmentation function. Therefore, measurements of hadron-to-hadron prompt production cross section ratios are an effective tool to probe charm hadronisation, as the contributions from the PDFs and the charm production cross section substantially cancel out in the ratio.

The production cross section of several charmed-hadron species have been measured at various collision energies and across different colliding systems. Electron-positron collisions provide a clean environment, with no initial hadronic state, and the centre-of-mass energy of the partonic system is known and coincides with the $\sqrt{s}$ of the colliding leptons, thus allowing for the measurement of the fragmentation functions. The integral of the fragmentation function over z gives the corresponding fragmentation fraction (FF). If the fragmentation functions universality holds, also the FFs do not depend on the colliding system.

The fragmentation fractions of different charmed-hadron species have been measured in $\mathrm{e^+e^-}$ collisions both at B-factories ($\sqrt{s} \sim 10.5$ GeV/c^2) and in Z-boson decays at LEP ($\sqrt{s} \sim 90$ GeV/c^2), as well as in $\mathrm{e^\pm p}$ collisions at HERA. The FFs did not exhibit significant differences, as shown in Fig. 2.3, thus supporting the hypothesis of universality of the fragmentation fractions [5].

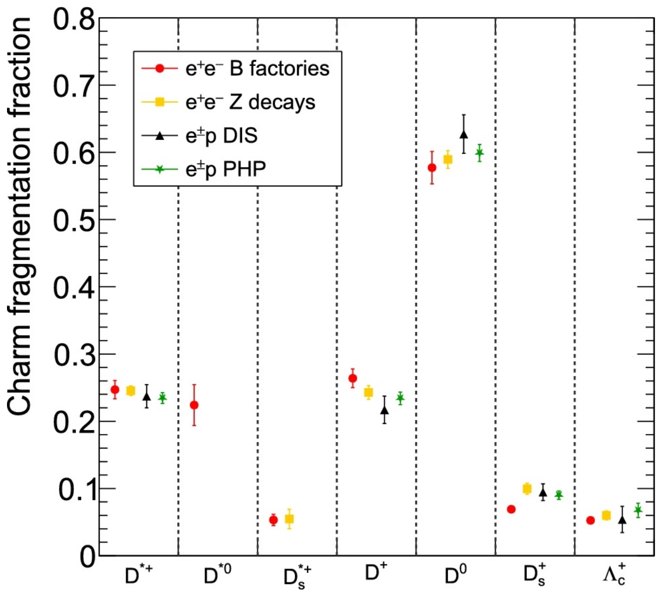


Figure 2.3: Fragmentation fractions for different charmed-hadron species measured in $\mathrm{e^+e^-}$ and $\mathrm{e^\pm p}$ collisions [5].

The D-meson measurements performed on the first data collected in pp collisions at the LHC confirmed once more this hypothesis. Multiple experiments measured the prompt ratio $\mathrm{D_s^+/(D^0 + D^+)}$, representing the FF of charm quarks to strange mesons f_s divided by the FF to non-strange mesons $f_u + f_d$. As shown in Fig. 2.4, the results extracted by the various collaborations were compatible within the uncertainties [5].

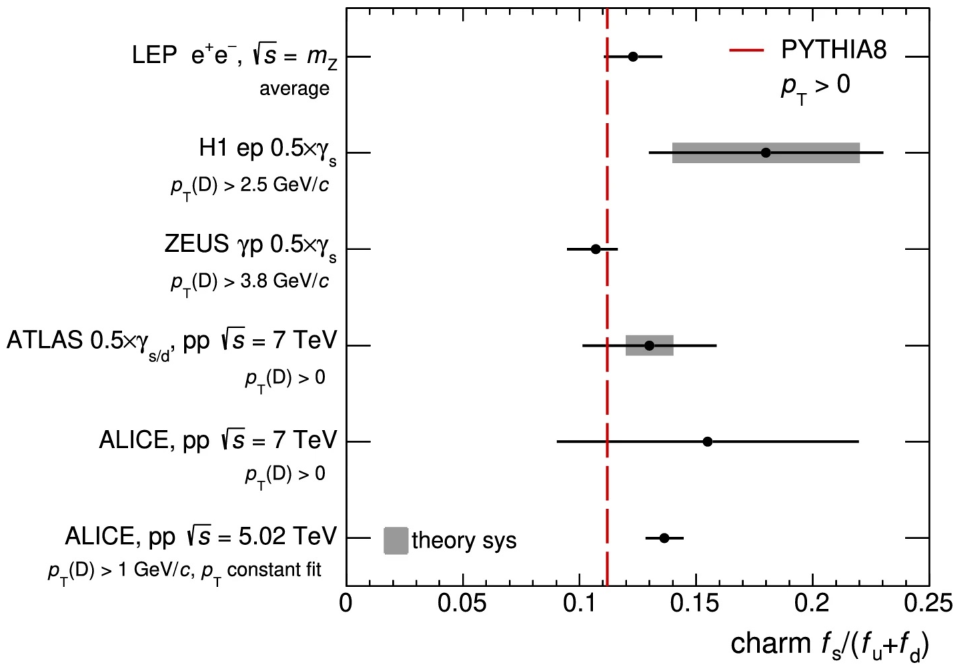


Figure 2.4: Measurement of prompt $D_s^+/(D^0 + D^+)$ ratio in different colliding systems [5].

However, when the measurement of charmed-baryons became accessible at the LHC in pp collisions, the corresponding FF results showed significant deviations from the values extracted from e^+e^- and $e^{\pm}p$ data, challenging the assumption of universality of the hadronisation process across collision systems [10]. In hadronic collisions at the LHC energies, an increase of the Λ_c^+ fragmentation fraction with respect to e^+e^- and $e^{\pm}p$ collisions was observed, as reported in Fig. 2.5. The concomitant decrease of the D-mesons FFs accounts for the conservation of the total $c\bar{c}$ production. More recent ALICE measurements at higher collision energies confirmed these observations and proved that the FFs do not exhibit significant energy dependence within the uncertainties [11], as shown in Fig. 2.5. Moreover, the baryon enhancement has also been observed in the beauty sector by the LHCb experiment [12].

These results suggest that different hadronisation mechanisms are at work depending on the colliding system and that the number of constituent quarks plays a role in the hadron formation process. One is instinctively tempted to relate this observation to the parton density of the initial state. While leptonic collisions can be regarded as a "vacuum-like" reference, hadronic collisions involve composite objects with multiple partons that can interact simultaneously, thus creating more complex colour field topologies. To validate this assumption, meson-to-meson, baryon-to-baryon, and baryon-to-

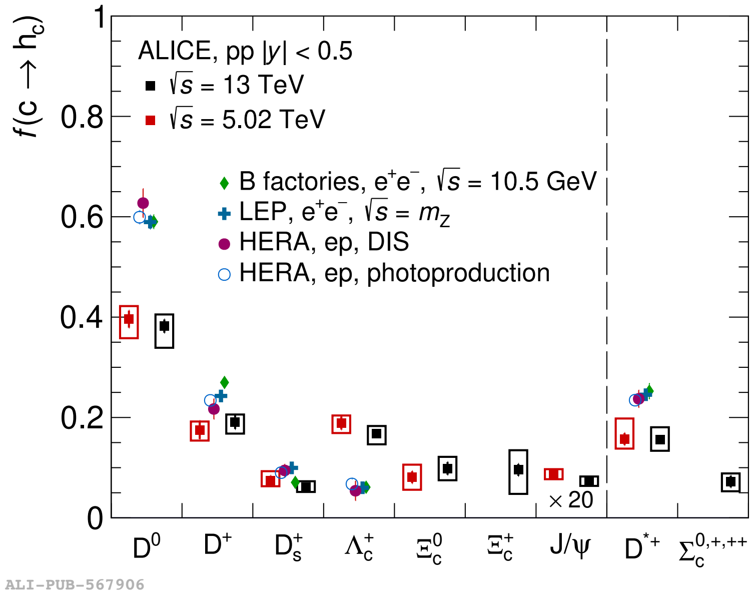


Figure 2.5: Fragmentation fractions for different charmed-hadron species measured in e^+e^-, $e^\pm p$ and pp collisions [11].

meson ratios are measured in different hadronic colliding systems, namely pp, p–Pb and Pb–Pb collisions. The multiplicity dependence of these ratios is also studied, as it can be considered a measurable proxy for the "hadronic activity" of the event, which in turn is connected to the parton density, and allows to bridge hadronic collision systems of different size [13]. Experimental results for these observables are reported in Section 2.4.2. The hadronisation theoretical models whose predictions are compared to the result presented in this work are discussed in Section 2.4.1.

2.4.1 Hadronisation models

There is a variety of models that aim at describing the hadronisation process. Most of them, namely the Catania model, the Quark re-Combination Mechanism (QCM) and Pythia, describe the hadron production starting from the partonic stage, but they differ for the colour neutralisation procedure implemented. On the other hand, the Statistical Hadronisation Model (SHM) predicts the abundances of different hadron species based on statistical weights, without dealing with the microscopic mechanism responsible for the transition from the partonic phase to the hadronic one.

Pythia

In Pythia, hadronisation is implemented on the basis of the Lund string fragmentation model [14]. In accordance with the linear dependence of the confining potential on the separation between the quarks, it describes the colour confinement field between a colour and an anti-colour charge in an overall singlet state as a string with constant tension and characterised by relativistic dynamics. In a collision process, the two quarks move apart and the string connecting them is stretched until it becomes energetically more favourable to break it. At the site of the break, a light-flavoured (u, d, s) (di)quark–anti-(di)quark pair is created and two smaller string systems are thus produced. This iterative fragmentation process stops when there is no longer sufficient energy to break the string, and it ends with the production of the final-state hadrons [5]. This process is depicted in Fig. 2.6.

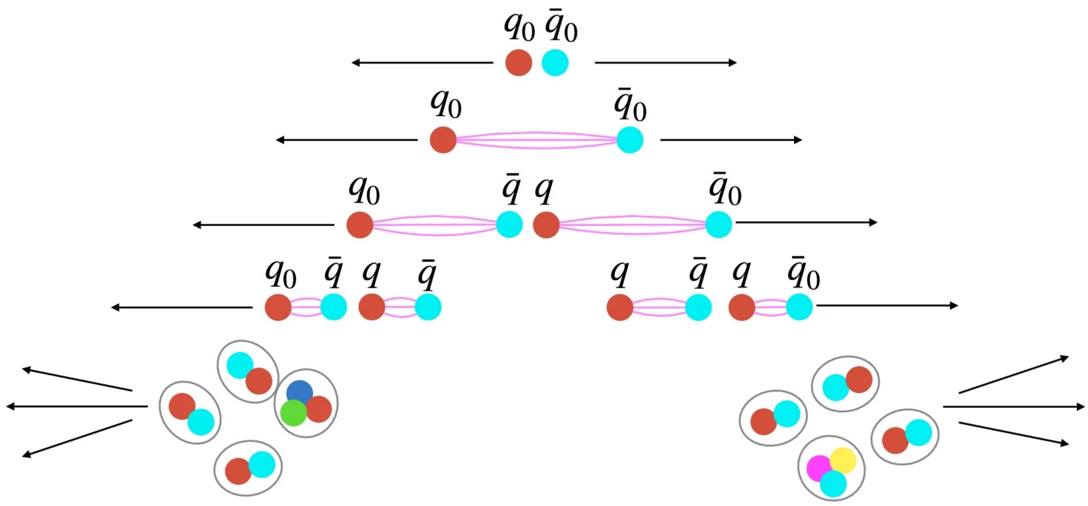
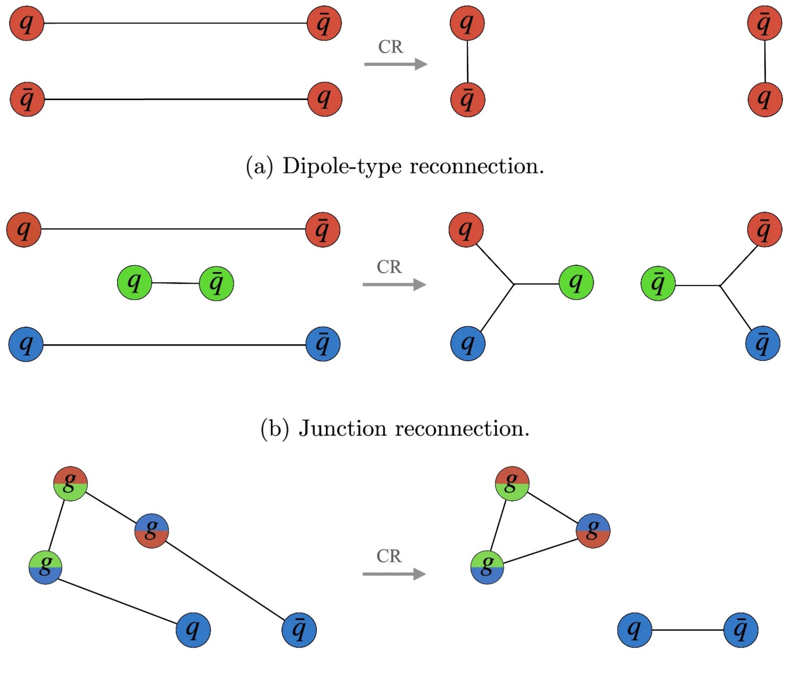


Figure 2.6: Schematic representation of the string breaking process in the Lund model [5].

The non-perturbative process of string breaking is modelled via a quantum tunnelling process resulting in a suppression factor $\propto \exp\left(\frac{-\pi\, m_{\perp q}^2}{k}\right)$, with $m_{\perp q}$ being the quark transverse mass and $k \sim 1$ GeV/fm the string tension inferred from hadron spectroscopy. As a consequence, it is very unlikely to produce both heavy quarks and high-p_{T} quarks via string breaking, reflecting the fact that such partons can only be produced by perturbative mechanisms. Once that the p_{T} and the flavour of the quarks created at the break site have been selected, the hadron species to be created, that has to be compatible with the flavour of the current endpoint and of the quarks produced in the string breakup, is determined from data, via tuning [5].

The **Pythia Monash tune** [15] is tuned on $\mathrm{e^+e^-}$ and $\mathrm{e^{\pm}p}$ measurements and implements string formation in the leading-colour (LC) limit. In this approximation, the number of colours N_C is taken to infinity $N_C \to \infty$. As a consequence, in a given event each colour is uniquely matched to a single anti-colour, thus resulting in unambiguous string configurations. Given these settings, only dipole string configurations are allowed, with the two endpoints being a colour and the corresponding anti-colour charge. In such framework, charmed-baryon production can only occur via diquark–anti-diquark pair creation next to a charm-flavoured string endpoint [5].

Pythia Monash is suitable to describe the processes that take place in leptonic collisions, but it can not reflect the complexity of hadronic collisions, which can be characterised by multiparton interactions (MPI) [16]. In this case, the LC limit does not hold and the limited number of colours in QCD $N_C = 3$ has to be reintroduced. This implies that multiple different partons can in principle be colour-connected as each colour is no more unique. To select between which quarks the confining potential arises, a string length minimisation criterion is used. As a consequence, strings can be formed between partons not necessarily produced in the same hard scattering and can involve different MPI [5]. This framework is implemented in the **Pythia Colour Reconnection** (CR) model [17]. In such context, the colour-neutrality can be achieved through three different string configurations: dipole reconnections, that involve a colour and corresponding anti-colour charge, the junction reconnection, that implements the colour-neutral combination of the three charges r, g, b, and the gluon-loop. The junction and the gluon-loop topologies arise from beyond leading colour effects (BLC). These three configurations are depicted in Fig. 2.7. The additional mechanism of charmed-baryons formation via the junction reconnection, which carries a non-zero baryon number, results in an enhancement of the baryon yield [5].



(a) Dipole-type reconnection.

(b) Junction reconnection.

(c) Gluon-loop formation.

Figure 2.7: Schematic representation of the string topologies allowed in Pythia Monash (on the left) and Pythia CR (on the right) [5].

The Catania model

The Catania model [18] implements the hadronisation process both via fragmentation and coalescence. While in the fragmentation process an energetic parton gives rise to lower momentum hadrons, the coalescence mechanism assumes that partons close to each other in the phase space can recombine in a hadron with higher momentum with respect to the initial quarks. Given that the bulk of the quarks is produced at low p_{T}, it is more likely to find partons close in phase space in this kinematic region. Consequently, coalescence is expected to dominate at low p_{T} and fragmentation at high p_{T}. A schematic representation of the two hadronisation processes is depicted in Fig. 2.8. The Catania model assumes that charm quarks that do not hadronise via coalescence are converted into hadrons via fragmentation with a p_{T}-dependent probability $P_{\mathrm{frag}}(p_{\mathrm{T}}) = 1 - P_{\mathrm{coal}}(p_{\mathrm{T}})$.

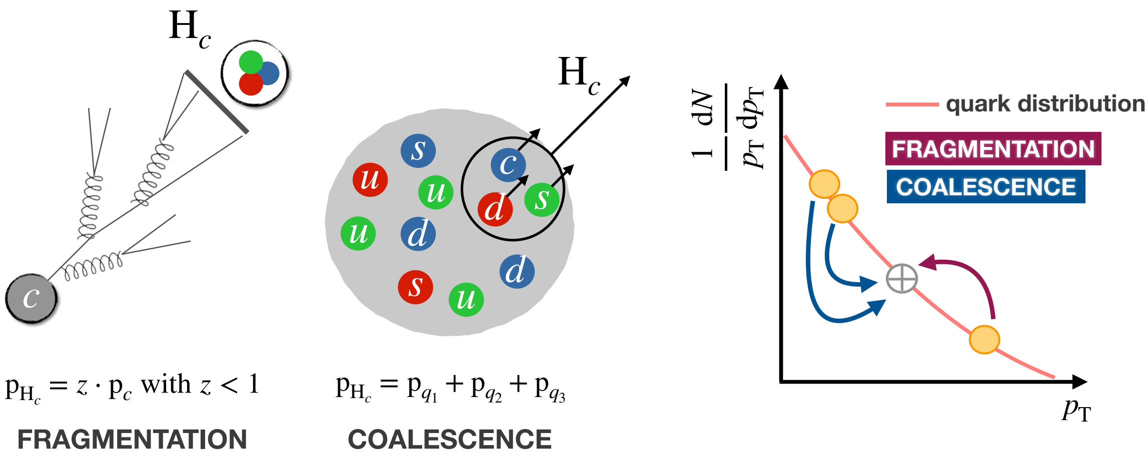


Figure 2.8: Schematic representation of the fragmentation and coalescence hadronisation mechanisms that result in the production of the charmed-hadron H_c.

The hadron p_{T}-spectrum is affected differently depending on the hadronisation mechanism involved, with coalescence pushing the average hadron transverse momentum to larger values with respect to fragmentation. The predicted P_{T} spectrum of a hadron with N_q constituent quarks and formed via coalescence is described by

$$\frac{\mathrm{d}N_{\mathrm{H}}}{\mathrm{d}y\mathrm{d}^2P_{\mathrm{T}}} = g_{\mathrm{H}} \cdot \int \prod_{i=1}^{N_q} \frac{\mathrm{d}^3p_i}{(2\pi)^3E_i} \cdot p_i \cdot \mathrm{d}\sigma_i \cdot f_{q_i}(x_i, p_i) \cdot f_{\mathrm{H}}(x_1, ..., x_{N_q}, p_1, ..., p_{N_q}) \cdot \delta^{(2)}(P_{\mathrm{T}} - \sum_1 p_{\mathrm{T,\,i}}). \quad (2.2)$$

In Eq. 2.2, g_{H} is a statistical factor indicating the probability that two (or three) random quarks have quantum numbers matching the one of the colourless hadron, the integral is evaluated over the phase space, $\mathrm{d}\sigma_i$ denotes an element of a space-like hypersurface, f_{q_i} is the phase space distribution of the i-th (anti-)quark and p_i is its transverse momentum, f_{H} is the Wigner function describing the spatial and momentum distribution of quarks inside the hadron, and the δ accounts for momentum conservation. The integral is usually evaluated with Monte Carlo methods.

On the other hand, the hadron momentum spectra from the charm quark fragmentation are given by

$$\frac{\mathrm{d}N_{\mathrm{had}}}{\mathrm{d}y\mathrm{d}^2P_{\mathrm{T}}} = \sum \int \mathrm{d}z \frac{\mathrm{d}N_{\mathrm{fragm}}}{\mathrm{d}y\mathrm{d}^2P_{\mathrm{T}}} \frac{D_{\mathrm{had/c}}(z, Q^2)}{z^2}, \tag{2.3}$$

where $D_{\mathrm{had/c}}(z, Q^2)$ is the fragmentation function, $z = p_{\mathrm{had}}/p_{\mathrm{charm}}$ is the momentum fraction of the heavy quark transferred to the final heavy hadron and Q^2 is is the momentum scale for the fragmentation process.

The model assumes the presence of a flowing thermalized medium of light (u, d, s) quarks, with the production of quark-gluon plasma occurring both in pp and p–Pb collisions. As for pp collisions, the charm quark spectrum is produced with FONLL calculations.

The Catania model was initially developed to describe AA collisions, where the high energy densities achievable and the formation of the quark-gluon plasma, which represents a large reservoir of partons, were expected to favour the coalescence mechanism as a hadron formation process, whereas in small colliding systems the predicted dominant process was the creation of quarks from the vacuum, and so the fragmentation. Indications for a hadronisation mechanism sensitive to the presence of partons in the initial state already came from measurements of charm production in pion-nucleon collisions, in particular from the observation that the production of charmed-hadrons sharing valence quarks with beam hadrons was favoured (leading particle effect) [19]. Nowadays the Catania model is also used to interpret measurements performed in pp collisions and it gives the best description of the available experimental data.

The Quark re-Combination Mechanism

According to this model [20], hadronisation occurs via coalescence at all momenta. The charmed-hadron formation process is described as the combination of a charm quark with equal-velocity co-moving light quarks, neglecting the spatial properties of the system. The quark distribution functions employed in this framework are extracted from a fit to π, K, and D-mesons experimental data. Differently from the Catania model, where the temperature of the medium determines the thermal distributions of the light quarks, the concept of medium temperature does not play a direct role [5].

The Statistical Hadronisation Model

The Statistical Hadronisation Model (SHM) [21] predicts the hadron abundances without modelling the complex hadronisation process, but simply based on statistical weights governed by the hadron mass and the hadronisation temperature $T_{\mathrm{H}} \sim 160$ MeV. The thermal hadron densities n_i are predicted according to the equation

$$n_i = \frac{d_i}{2\,\pi^2} m_i^2\, T_{\mathrm{H}} K_2 \left(\frac{m_i}{T_{\mathrm{H}}}\right), \tag{2.4}$$

where m_i is the hadron mass, d_i is the spin-isospin degeneracy, and K_2 is the modified Bessel function of second order. Charm quarks are produced in the initial stages of the collision with a multiplicity larger than the one expected from thermal production at T_{H}. To account for this excess, a fugacity factor is introduced [5].

In order to provide a good description of the Λ_c^+/D^0 ratio measured by ALICE [22], the feed-down from a largely augmented set of unobserved charmed-baryon states beyond the ones currently listed in the PDG [4] has to be taken into account. The missing states are predicted with the Relativistic Quark Model (RQM) [23], which is based on a relativistic description of a bound object composed of a charm quark and a light diquark. In this model, both orbital and radial baryon excitations are considered. They can only occur between the heavy quark and the light diquark, whereas the latter is always taken in the ground (scalar or axial-vector) state. Such a scheme reduces the number of possible excited baryon states with respect to a simple three-quark picture and nicely accommodates the available experimental data. Heavy baryons up to the radial excitation quantum number $n_r = 5$ and orbital excitation quantum number $L = 5$ are considered. The model calculations predict mass values reaching ~ 4 GeV/c^2 for the heaviest excited states. For observed excited charmed-hadrons states, the branching ratios to the ground states as available from the PDG are used, while equal weights are assumed for decay channels without quoted branching ratios. While the PDG currently lists six Λ_c states, three Σ_c, eight Ξ_c and two Ω_c, the RQM predicts the existence of extra eighteen Λ_c states, forty-two Σ_c, sixty-two Ξ_c and thirty-four Ω_c.

2.4.2 Measurements of baryon-to-meson ratios

The measurement of baryon-to-meson ratios in different colliding systems is an effective tool to investigate the dependence of the hadronisation process on the underlying event and on the number of constituent quarks. Since the predicted charmed-hadron p_{T} spectra depend on the hadronisation mechanism at work, studying these ratios as a function of the hadron transverse momentum allows to validate the hadronisation models.

In pp collisions, ALICE measured the Λ_c^+/D^0 [22], $\Sigma_c^{0,++}/\mathrm{D}^0$ [24], $\Xi_c^{0,+}/\mathrm{D}^0$ [11] and Ω_c^0/D^0 [25] ratios as a function of the hadron transverse momentum. The results for non-strange baryons are shown in Fig. 2.9. These ratios show a clear p_{T} dependence, with larger baryon production at low and intermediate transverse momentum. The Λ_c^+/D^0 ratio is larger than the average ratio measured at LEP in e^+e^- collisions, amounting to ~ 0.11 [5], across the whole investigated p_T range. The theoretical prediction driven by charm quark fragmentation processes in e^+e^- collisions, namely Pythia Monash, underestimates the data, while Pythia CR, the Catania model and the prediction by the SHM coupled with the RQM successfully describe the data. These results support the scenario of charm quark hadronisation in pp collisions via additional mechanisms to those in leptonic collisions.

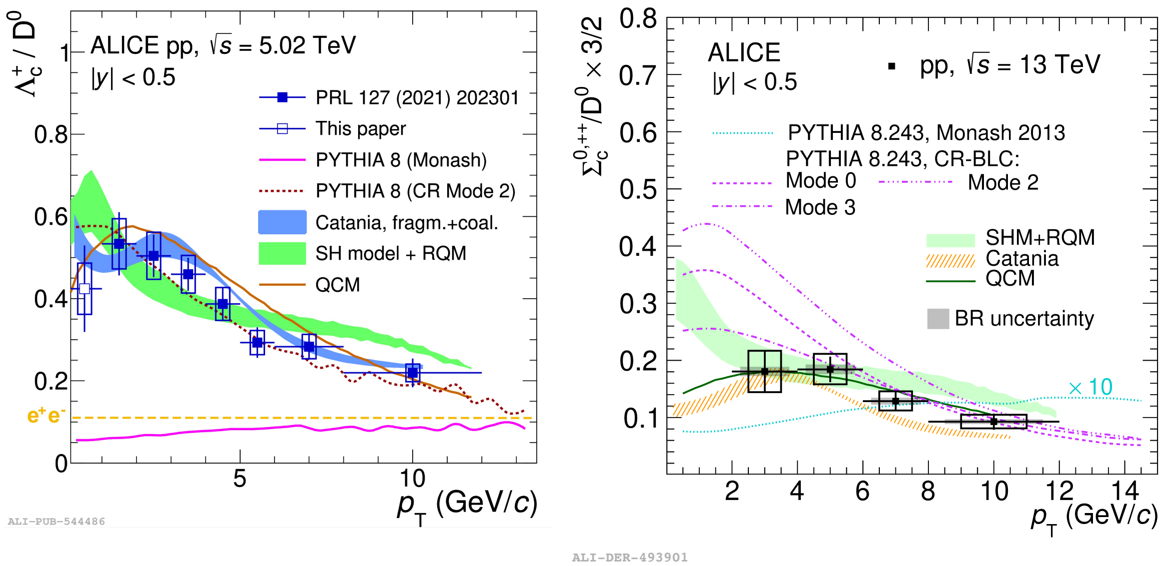


Figure 2.9: Measurement of the Λ_c^+/D^0 ratio (on the left) and $\Sigma_c^{0,++}/D^0$ ratio (on the right) as a function of the hadron transverse momentum performed by ALICE in pp collisions. The results are compared to theoretical predictions, the average Λ_c^+/D^0 ratio measured in leptonic collisions is also reported [22] [24].

The results for $\Xi_c^{0,+}$ and Ω_c^0 are reported in Fig. 2.10. The baryon-to-meson ratios exhibit an enhancement with respect to the e^+e^- baseline, namely the Pythia Monash prediction, also for the strange-charmed sector. The result for the $\Xi_c^{0,+}$ shows a p_T dependence, with a larger increase of the ratio at low and intermediate transverse momentum. The theoretical models that successfully described the Λ_c^+/D^0 ratio, i.e. Catania, SHM+RQM and Pythia CR, provide a poor description of this measurement and underestimate the result. This observation raises the question of whether there is a larger enhancement of strange-charmed baryons. The theoretical prediction closest to the $\Xi_c^{0,+}/D^0$ data is based on the Catania model, thus suggesting that the mechanism of coalescence could occur also in small hadronic colliding systems. Measurements of the doubly-strange charmed-baryon Ω_c^0 could be crucial to improve our understanding of the role of strangeness in the hadronisation process, but the large uncertainty on the Ω_c^0 branching ratio does not allow to draw a firm conclusion on the p_T dependence and limits the effectiveness of the comparison to theoretical models. However, the theoretical predictions based on Pythia and QCM underestimate the data significantly, while Catania is again consistent with the measurement.

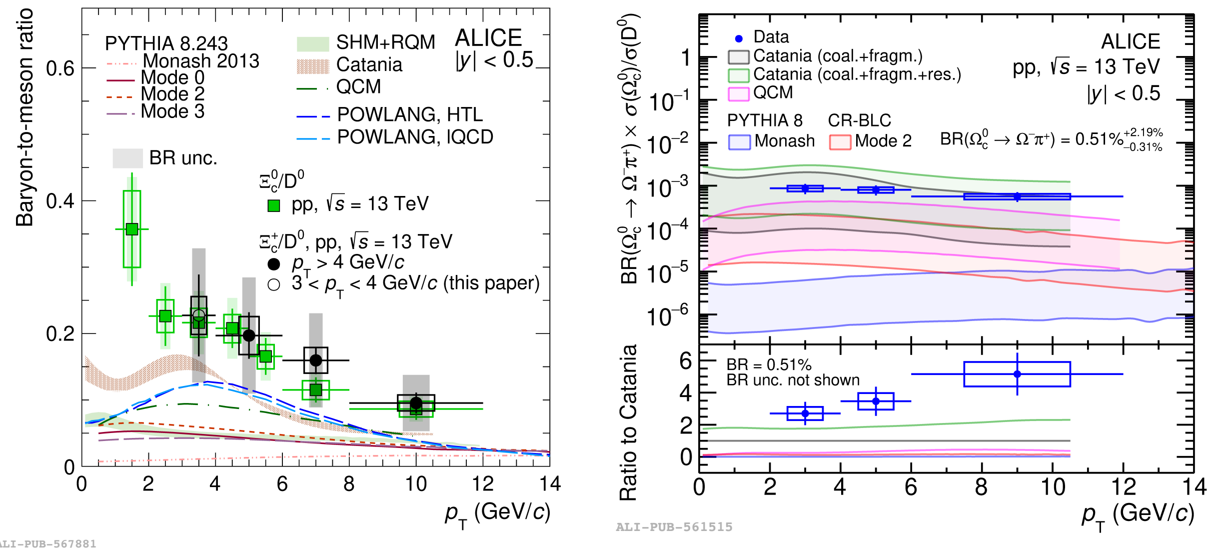


Figure 2.10: Measurement of the $\Xi_c^{0,+}/\mathrm{D}^0$ ratio (on the left) and Ω_c^+/D^0 ratio (on the right) as a function of the hadron transverse momentum performed by ALICE in pp collisions. The results are compared to theoretical predictions [11] [25].

The Λ_c^+/D^0 and Ξ_c^0/D^0 ratios have also been measured in p–Pb collisions [22] [26], the results are shown in Fig. 2.11. A similar magnitude in the enhancement of Λ_c^+/D^0 in pp and p–Pb with respect to $\mathrm{e}^+\mathrm{e}^-$ is observed and the shift of the distribution peak towards higher p_T could be attributed to radial flow. The QCM successfully describes the magnitude of the ratio and predicts the hardening of the Λ_c^+ spectrum in p–Pb collisions. As for the Ξ_c^0/D^0 measurement, higher precision is needed to draw a firm conclusion on the multiplicity dependence, but the comparison to the QCM prediction shows that this model underestimates the ratio also in p–Pb collisions.

Finally, the Λ_c^+/D^0 ratio has been measured in Pb–Pb collisions [27]. In this system, the quark-gluon plasma (QGP), a colour-deconfined state of matter, is formed. For many years physicists suggested that in the QGP hadrons may form via recombination of existing quarks [28], thus affecting hadron abundances and their momentum distributions. The Λ_c^+/D^0 result, reported in Fig. 2.12, shows that the ratio increases from pp to mid-central and central Pb–Pb collisions at intermediate transverse momentum. The distribution peak increases in magnitude and shifts towards higher p_T values. Such behaviour could be the result of the modification of the p_T spectra due to recombination, possibly reinforced by radial flow, with the charmed-hadrons partly inheriting the flow of light quarks. Theoretical models describing hadronisation as proceeding primarily via coalescence successfully describe the measurement [27].

ALICE also investigated the multiplicity dependence of both the Λ_c^+/D^0 and $\Xi_c^{0,+}/\mathrm{D}^0$ ratios [29] [30]. The corresponding results, reported in Fig. 2.13, show that there is a significant multiplicity-

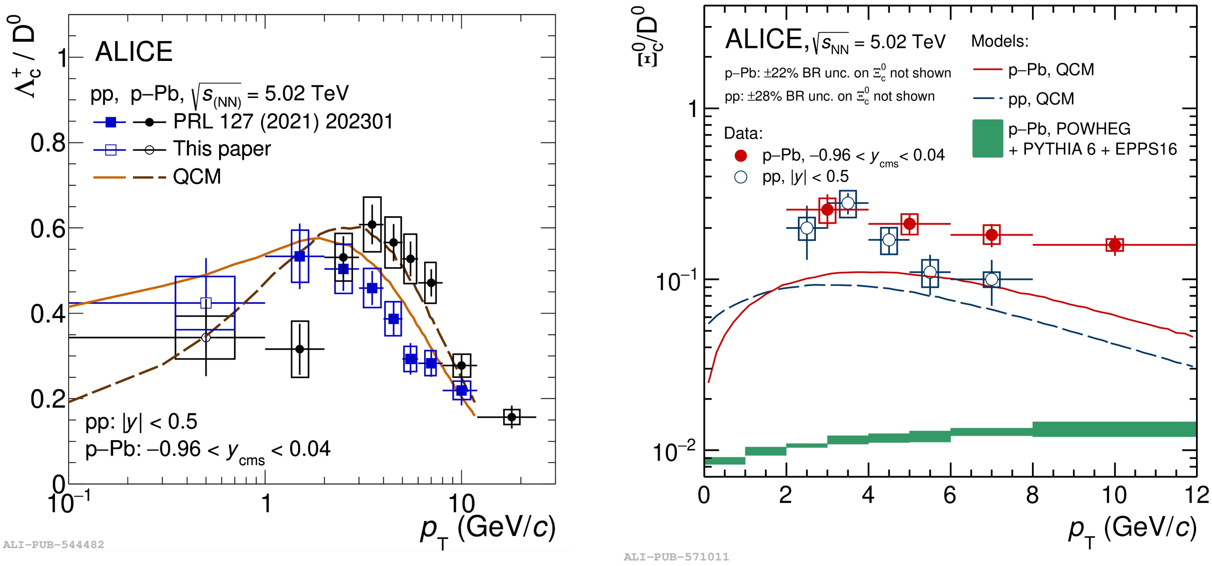


Figure 2.11: Measurement of the Λ_c^+/D^0 ratio (on the left) and Ξ_c^0/D^0 ratio (on the right) as a function of the hadron transverse momentum performed by ALICE in p–Pb collisions. The results are compared to theoretical predictions [22] [26].

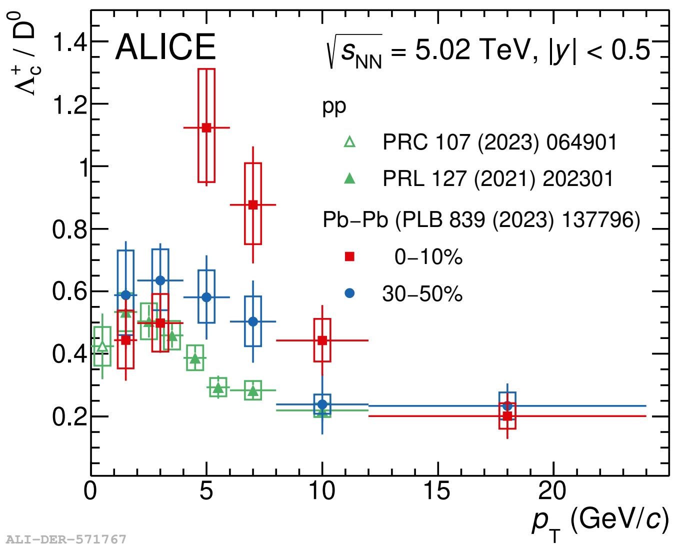


Figure 2.12: Measurement of the Λ_c^+/D^0 ratio as a function of the hadron transverse momentum performed by ALICE in Pb–Pb collisions [27].

dependent enhancement of the Λ_c^+/D^0 ratio from the lowest to the highest multiplicity class, with the ratio measured in the lowest multiplicity class still being higher than the average value measured in $\mathrm{e}^+\mathrm{e}^-$ collisions at LEP. Pythia Monash fails to describe the ratio, while Pythia CR catches the trend of the measurement but not its magnitude. As for the $\Xi_c^{0,+}/\mathrm{D}^0$ ratio, there is no strong multiplicity dependence within the uncertainties, higher precision is needed to draw firm conclusions.

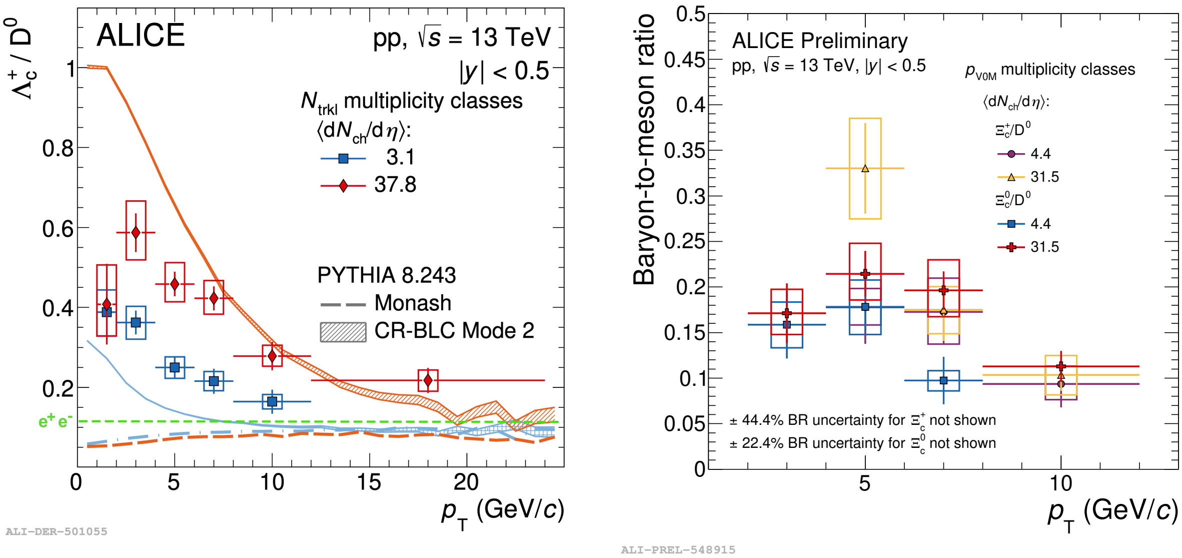


Figure 2.13: Measurement of the Λ_c^+/D^0 ratio (on the left) and $\Xi_c^{0,+}/D^0$ ratio (on the right) as a function of the hadron transverse momentum for different multiplicity classes performed by ALICE in pp collisions [29] [31].

Meson-to-meson ratios

ALICE measured the cross sections of multiple D-meson species and extracted the corresponding ratios [11]. These observables do not show any significant dependence on the p_T of the hadron within the uncertainties, as shown for instance in Fig. 2.14. Moreover, no appreciable dependence on the collision energy is observed within the current experimental uncertainties. This can be interpreted as indication of common charm quark fragmentation functions to the considered D-meson species, independently of the strange quark content and on the collision energy. The ALICE measurement of $D_s{}^+/D^0$ as a function of p_T for different multiplicity classes [29] indicates no significant multiplicity dependence of the ratio, as shown in Fig. 2.15.

Baryon-to-baryon ratios

The measurement of the Ξ_c^0/Λ_c^+ and $\Xi_c^{0,+}/\Sigma_c^{0,+,++}$ ratios in pp collisions performed by ALICE [32] showed no significant p_T dependence. As reported in Fig. 2.16, all the models underestimate the measurements with the exception of the Pythia Monash prediction for the Ξ_c/Σ_c ratio, but this is due to the fact that such model underestimates the cross sections of these baryons by a similar amount. This could be either accidental or it may indicate the removal of a similar suppression mechanism affecting both Ξ_c and Σ_c production in e^+e^- collisions [5]. The Ξ_c^0/Λ_c^+ ratio measured by ALICE [30] does not show any multiplicity dependence within the uncertainties and all the Pythia predictions underestimate the data, as shown in Fig. 2.17.

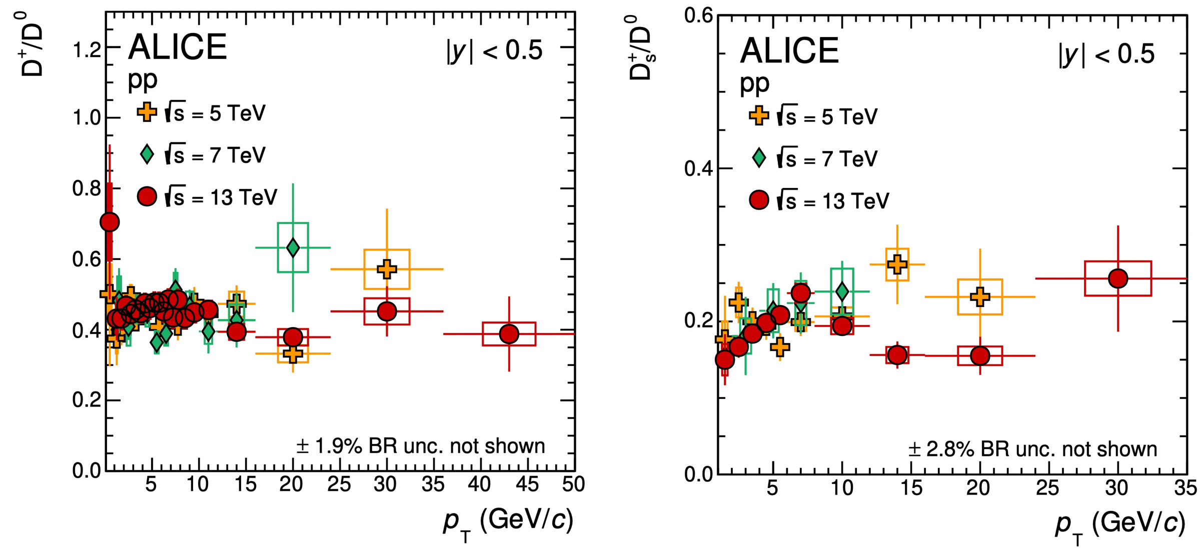


Figure 2.14: Measurement of cross section ratios of different D-meson species as a function of p_{T} performed by ALICE in pp collisions [11].

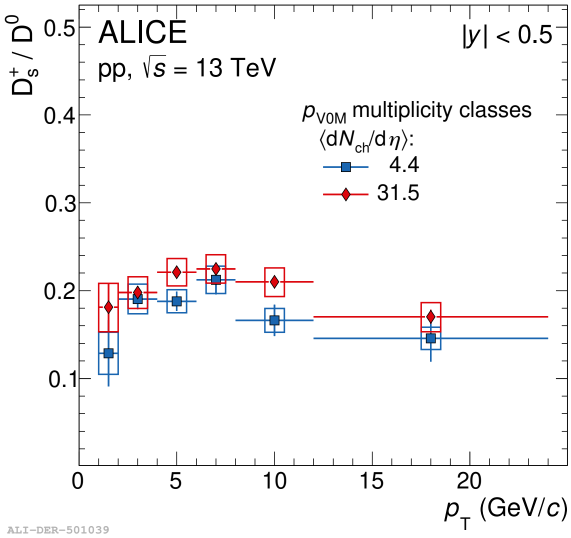


Figure 2.15: Measurement of $\mathrm{D_s}^+/\mathrm{D}^0$ ratio as a function of p_{T} for different multiplicity classes performed by ALICE in pp collisions [29].

Rapidity puzzle

The results reported above discussed the transverse momentum and multiplicity dependence. All the ALICE measurements presented here have been performed at midrapidity, while the forward and backward rapidity regions have been investigated by the LHCb experiment. The comparison between the ALICE and LHCb results indicates a possible rapidity dependence of the baryon-to-meson ratio for both the Λ_c^+ and $\Xi_c^{0,+}$ baryons in multiple hadronic colliding systems [5]. As an example, Fig. 2.18 reports the ALICE [26] and LHCb [33] measurements of the $\Xi_c^{0,+}/\mathrm{D}^0$ ratio and it shows that

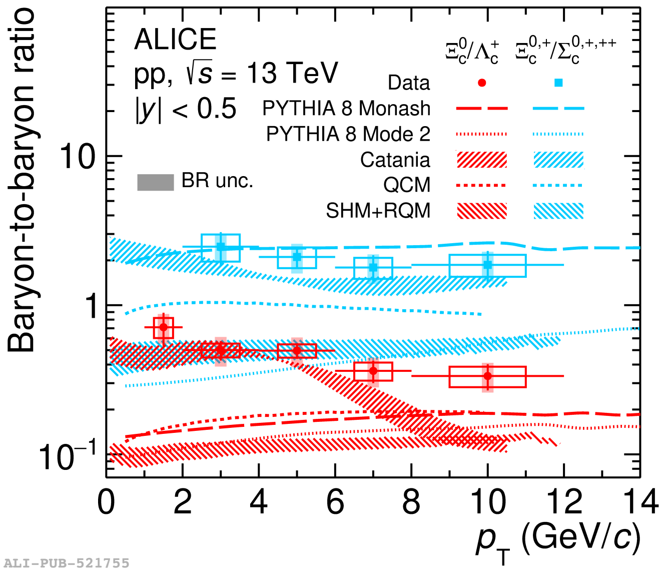


Figure 2.16: Measurement of the Ξ_c^0/Λ_c^+ and $\Xi_c^{0,+}/\Sigma_c^{0,+,++}$ ratios as a function of the hadron transverse momentum performed by ALICE in pp collisions [32].

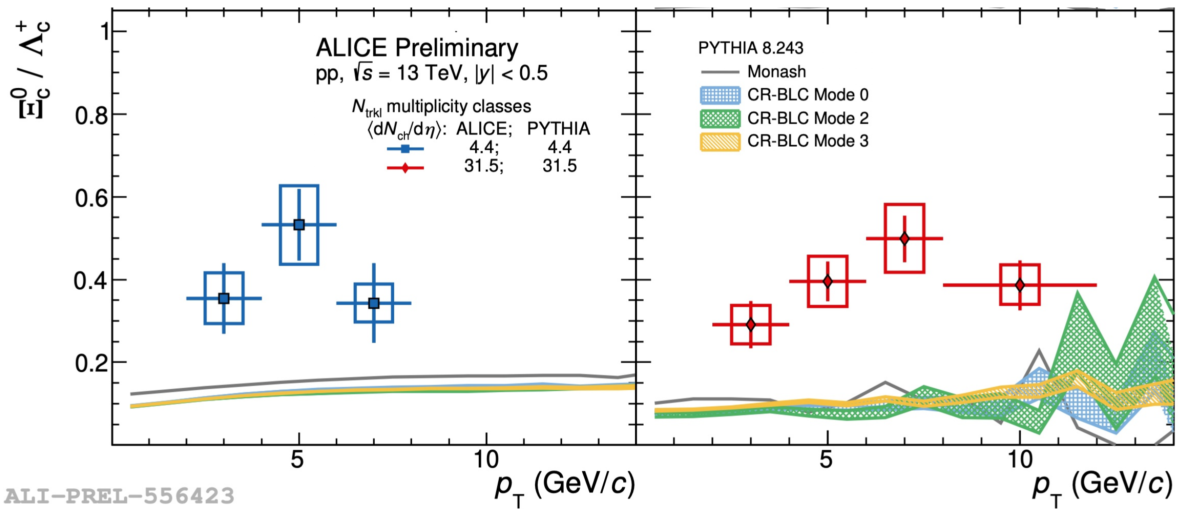


Figure 2.17: Measurement of the Ξ_c^0/Λ_c^+ ratio as a function of the hadron transverse momentum for different multiplicity classes performed by ALICE in pp collisions [31].

larger values are observed at midrapidity. On the other hand, the results for the baryon-to-baryon ratio $\Xi_c^{0,+}/\Lambda_c^+$ are compatible within uncertainties in the two rapidity intervals. However, either because of the large uncertainties, or due to inconsistencies in the accessed p_T intervals or in the definition and ranges used for multiplicity and centrality by the two collaborations, currently it is not possible to draw any conclusion.

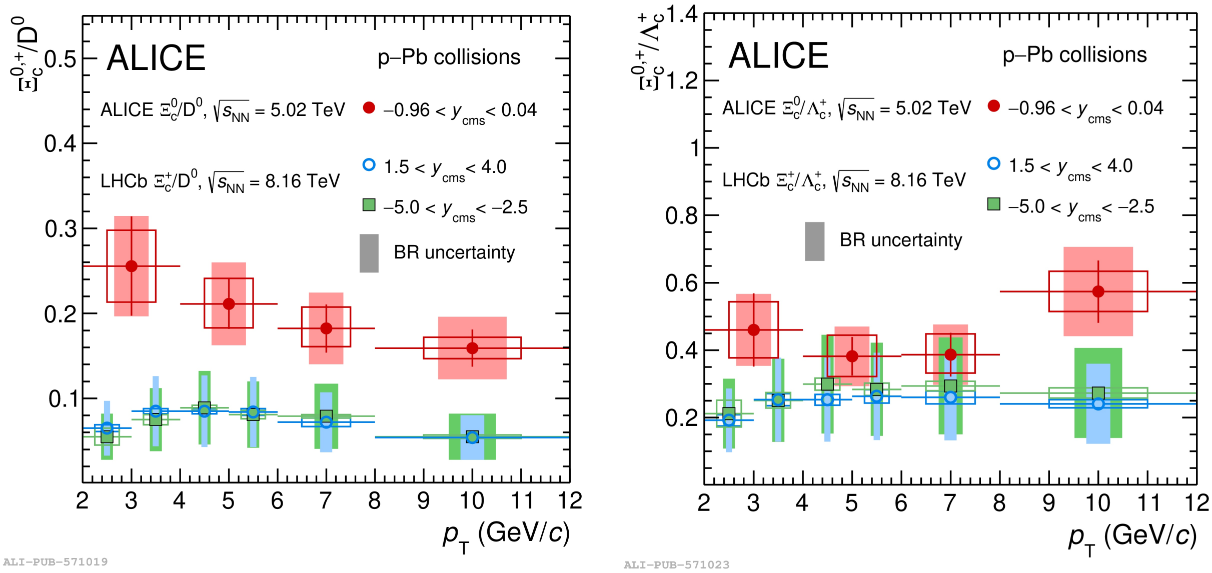


Figure 2.18: Comparison of ALICE and LHCb measurements in p–Pb collisions of $\Xi_c^{0,+}/D^0$ ratio (on the left) and $\Xi_c^{0,+}/\Lambda_c^+$ ratio (on the right) as a function of the hadron transverse momentum [26]

Emerging picture

The experimental results for production cross section ratios of various charmed-hadron species across different colliding systems provided evidence of the fact that charmed-baryons are more abundantly produced in pp collisions compared to e^+e^- collisions, thus implying the breaking of the fragmentation universality paradigm. The measurements proved that the hadronisation process depends on the underlying event, as the presence of multiparton interactions in hadronic collisions translates in a larger variety of possible colour-reconnection topologies that can even span across different MPI. According to the theoretical model currently providing the best description of the experimental results, coalescence emerges progressively as hadron formation process with increasing number of partons produced in the collision, probably even in small hadronic colliding systems. This is justified by the fact that such hadron formation process requires an overlap of the parton wave functions in the phase space, and higher partonic densities imply larger overlap probabilities. Due to the decreasing trend of the momentum distribution of the quarks produced in the collision, the probability of having two partons close in phase space is larger at low p_T, therefore coalescence is expected to dominate in this kinematic region. At high p_T the baryon-to-meson ratios seem to approach universal values, with the recovery of fragmentation as a suitable description of hadronisation also for baryons. The number of the constituent quarks of the hadron seems to favour one hadronisation process over another, with a dependency on the hadron p_T. While the string fragmentation mechanism is particularly efficient at any p_T for creating quark–anti-quark pairs, and thus for mesons production, the increased proximity in phase space of partons produced in the collision which is typical of the low momentum region could make coalescence more effective to produce baryons at low p_{T}. The fact that different hadron formation mechanisms are expected to dominate in different regions of the baryon p_{T}, and that they induce different redistributions of the charm quark momentum, introduce a p_{T} dependence of the baryon-to-meson ratios. As hadrons with the same number of constituent quarks are equally affected by the partonic density of the surrounding environment at hadronisation, no p_{T} dependence is observed in the meson-to-meson and baryon-to-baryon ratios. Despite the more limited number of free parameters with respect to the Catania model, statistical hadronisation also emerges as an effective model to describe some of the measured baryon-to-meson ratios in pp collisions, on condition that the feed-down from the largely augmented set of unobserved charmed-baryon states predicted by the RQM is taken into account.

Finally, the role of strangeness in the hadronisation process is not fully understood, as a larger enhancement from $\mathrm{e^+e^-}$ to pp is observed for strange-charmed-baryons with respect to Λ_c^+, whereas no significant difference is found for the $\mathrm{D_s^+/D^0}$ ratio.

The increased interaction rate and continuous readout mode adopted by ALICE in Run 3 allow to collect a significantly larger dataset compared to Run 2. This opens a new dimension of precise charmed-baryon measurements in several decay channels, thus helping to investigate the mechanisms responsible for the production of these particles.

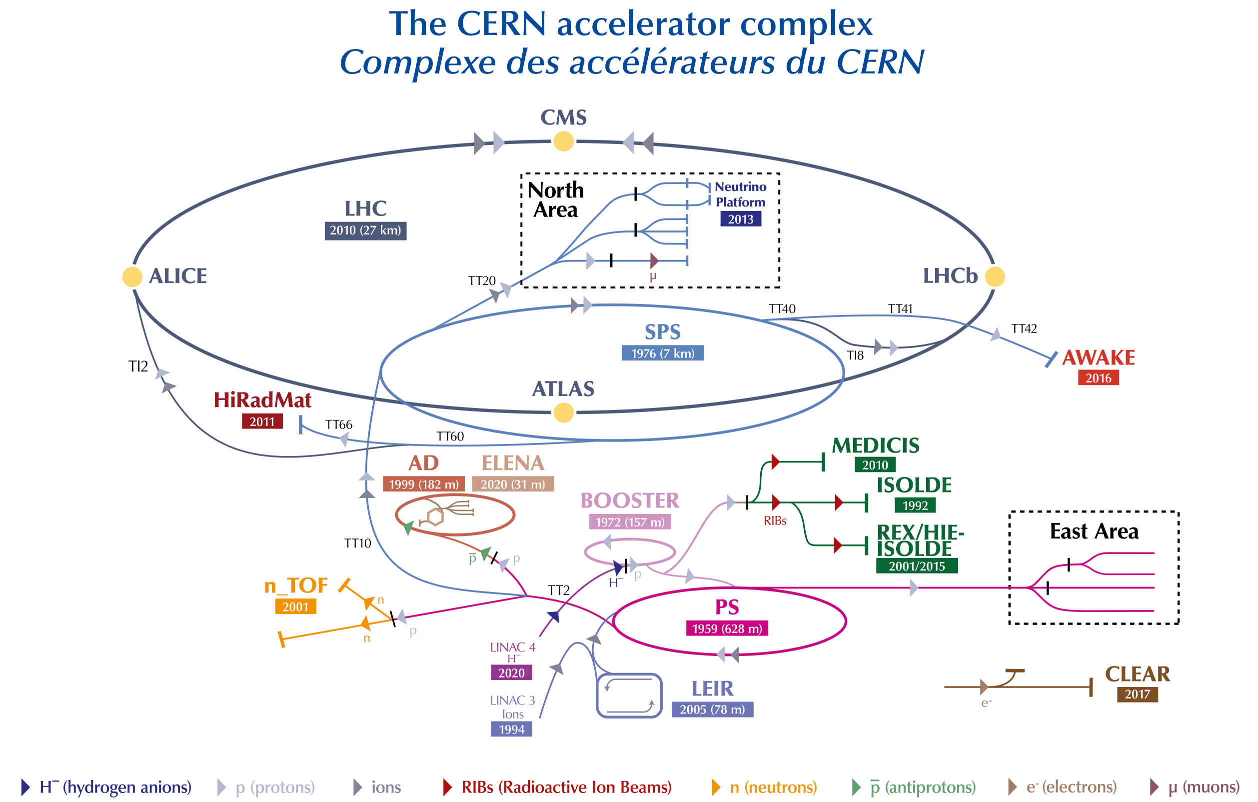
Chapter 3

The ALICE experiment at the LHC

The analysis presented in this thesis is based on the data collected by ALICE (A Large Ion Collider Experiment), one of the four major experiments at the CERN Large Hadron Collider (LHC).
This chapter is meant to introduce the sophisticated infrastructure used to deliver high-energy collisions and to reconstruct and analyse the particles produced in the interactions. In particular, Section 3.1 introduces the acceleration chain, Section 3.2 gives an overview of the ALICE experimental apparatus, with a focus on the relevant detectors used for heavy-flavour measurements, and Section 3.3 outlines the data reconstruction techniques used in ALICE.

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [34] situated at CERN, the European Organization for Nuclear Research, is currently the largest and most powerful hadron accelerator in the world. The LHC consists of a 27-kilometres ring of superconducting magnets with several accelerating structures to boost the energy of the particles along the way. The whole apparatus is designed to accelerate protons and ions at a maximum centre-of-mass-energy of 14 TeV for proton-proton and ~ 5.5 TeV per nucleon pair for AA (nucleus-nucleus) collisions. The structure is placed tens of meters underground across the border between France and Switzerland.
Thanks to various accelerating steps, particles can reach ultrarelativistic energies inside pipes kept at ultra-high vacuum. Such projectiles are guided around the accelerator ring by a strong magnetic field maintained by superconducting electromagnets, which require temperatures colder than the outer space ($\sim -271^{\circ}$ C). Such temperatures can be achieved thanks to a cryogenic system based on the distribution of liquid helium.
The proton acceleration process consists of various steps, with the LHC being the last machine of the chain. First of all, the protons are extracted from a gas bottle and accelerated in a duoplasmatron, which produces H^{-}. The negative hydrogen ions are accelerated by the Linear accelerator 4 (Linac



[redacted]

Figure 3.1: Schematic view of the CERN accelerator complex [35]

4) up to 160 MeV, then the ions are stripped of their two electrons and injected into the Proton Synchrotron Booster (PSB), which pushes the beam up to 2 GeV. After that, the protons are injected into the Proton Synchrotron (PS), which accelerates them to 26 GeV. The following machine is the the Super Proton Synchrotron (SPS), that pushes the beam to 450 GeV. Finally, the protons are injected into the LHC pipes, where they are accelerated up to the record energy of 6.8 TeV per beam. Besides delivering pp collisions, the CERN accelerator infrastructure can also operate ion beams, such as lead and xenon. At the Ion Accelerator Complex, high-purity ^{208}Pb is provided by an electron-cyclotron resonance source that heats solid lead which then vaporizes and is subsequently ionized. The complex is composed of a linear accelerator (Linac3) and the Low Energy Ion Ring (LEIR) synchrotron. The subsequent accelerating machines are the same operating proton beams, starting from the PS. Detailed information regarding the different acceleration steps and facilities can be found in Ref. [36].

During Run 1 of the LHC (2009–2013), the maximum possible collision energy was $\sqrt{s} = 8$ TeV for pp collisions and $\sqrt{s_{\mathrm{NN}}} = 2.76$ TeV for Pb–Pb collisions. Thanks to the improvements made during the first Long Shutdown (2013–2015), during Run 2 (2015-2018) the acceleration system allowed to reach energies of $\sqrt{s} = 13$ TeV for pp collisions and $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV for Pb–Pb collisions. Moreover, in 2017, the machine accelerated and collided for the very first time Xenon ions (Xe–Xe), reaching a centre-of-mass energy per colliding nucleon pair of 5.44 TeV. Further interventions upgraded the LHC during Long Shutdown 2 (LS2, 2018-2022), thus allowing to push once more the maximum achievable energy. Now Run 3 is ongoing (2022-2026) and the machine is delivering pp collisions at $\sqrt{s} = 13.6$ TeV and Pb–Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.36$ TeV. During Run 3, the LHC will also operate for the first time oxygen ion beams for Oxygen–Oxygen (O–O) and proton–Oxygen (p–O) collisions. Along the LHC ring, two beams are circulating in opposite directions and are brought into collision in four interaction points. At each of them, an experiment - which is made up of several particle detectors - is installed: ALICE (A Large Ion Collider Experiment), CMS (Compact Muon Solenoid), ATLAS (A Toroidal LHC ApparatuS), and LHCb (Large Hadron Collider beauty). In the following Section, the ALICE experiment is discussed in more detail.

3.2 The ALICE experiment

ALICE has been conceived and constructed as a heavy-ion experiment dedicated to the investigation of quark-gluon plasma properties. However, it also deals with proton-proton collisions in order to both obtain reference data for QGP-related analyses and shed light on open issues in elementary particle physics.

The apparatus was designed in order to satisfy the requirements of tracking particles from less than 100 MeV/c up to about 100 GeV/c, identifying them in a wide momentum range, and reconstructing short-lived particles in a high multiplicity environment (up to 8000 charged particles per rapidity unit at midrapidity). Moreover, during the LHC Long Shutdown 2 (LS2), ALICE underwent major upgrades that allow to take data in continuous readout mode for collision rates that reach 500 (50) kHz when the LHC delivers pp (Pb–Pb) collisions. Higher pp collision rates have also been tested, but a severe degradation in performance has been observed when approaching 1 MHz. As a result, extremely large datasets are now available, thus allowing to improve the statistical precision of measurements and to access rarer probes.

This experiment consists of several particle detectors that can be grouped in three main blocks:

- central barrel detectors: their pseudorapidity coverage is $|\eta| < 0.9$, with most of them covering the full azimuth. These detectors are embedded into a solenoid which provides a magnetic field of 0.5 T along the beam axis. They are used for vertexing, tracking and particle identification (PID), as well as for jet reconstruction. The central barrel detectors are the Inner Tracking System (ITS), the Time Projection Chamber (TPC), the Transition Radiation Detector (TRD), the Time Of Flight detector (TOF), the High Momentum Particle IDentification ring imaging Cherenkov detector (HMPID), the Photon Spectrometer (PHOS), and the Electromagnetic Calorimeter (EMCal).

- muon spectrometer: it covers the forward rapidity region ($2.5 < \eta < 4$) and is used for muon tracking and triggering.
- additional forward detectors, namely the Fast Interaction Trigger (FIT) and the Zero-Degree Calorimeter (ZDC): they are mainly used for interaction triggering, event characterization and luminosity measurements.

In Fig.3.2 the outline of the ALICE apparatus is reported.

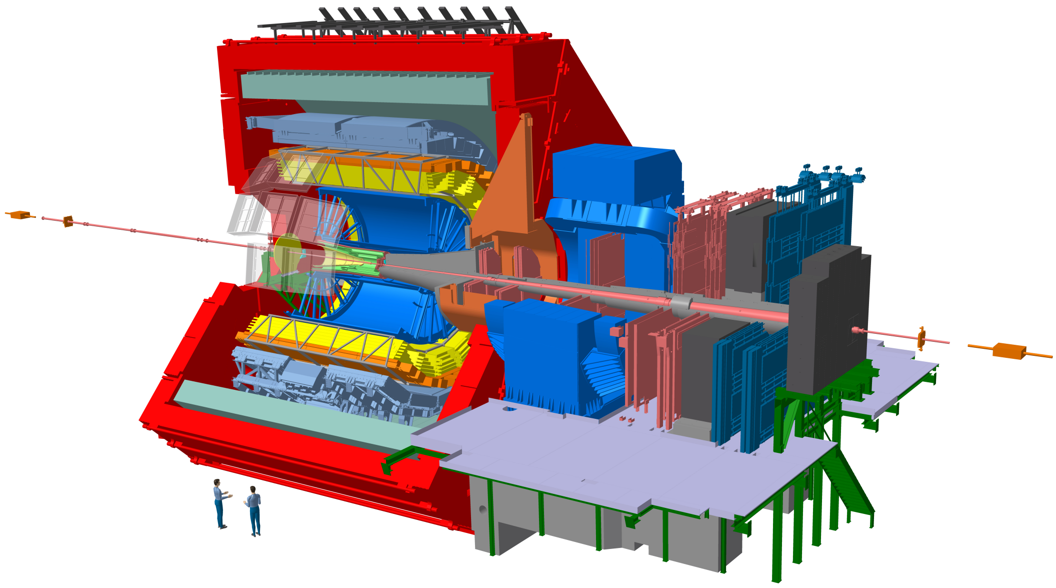


Figure 3.2: ALICE experimental apparatus in Run 3 [31].

In the following sections more details about the ALICE detectors relevant for the analysis presented in this thesis are given.

3.2.1 Inner Tracking System (ITS)

The ITS is a silicon detector mainly dedicated to tracking and vertexing purposes. It is the most central detector surrounding the beryllium beam pipe, which has an outer radius of 19 mm. It is depicted in bright green in Fig. 3.2, while a more detailed sketch and a picture are reported in Fig. 3.3.

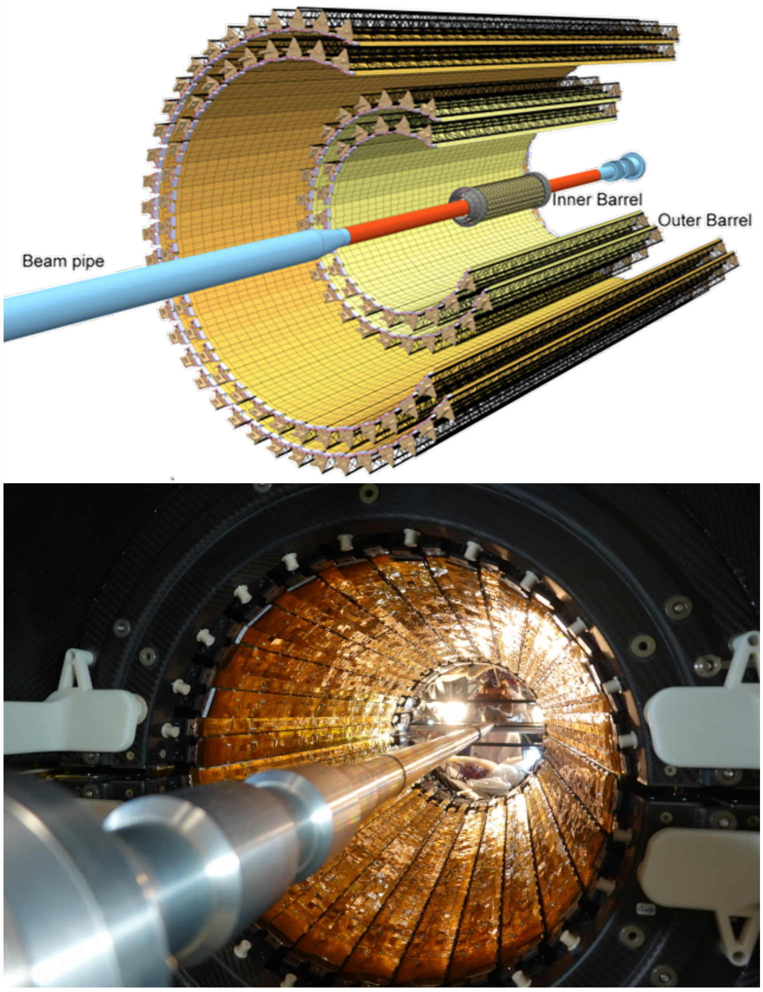


Figure 3.3: ALICE Inner Tracking System [37].

The ITS has a pseudorapidity coverage of $|\eta| < 1.3$ and full azimuthal coverage. This detector consists of seven cylindrical layers, with a radius varying from 22 to 395 mm. The three innermost ones define the inner barrel, while the outer barrel is composed of two double layers. The radial position of each layer is optimised to achieve the best possible performance in terms of pointing resolution, p_{T} resolution and tracking efficiency. It uses the Monolithic Active Pixel Sensor technology with a pixel size of 27 x 29 $\mu\mathrm{m}$, thus allowing to reach a spatial resolution of 5 $\mu\mathrm{m}$ in both the longitudinal and transverse direction. It has an active area of $\sim 10\ \mathrm{m}^2$, segmented in 12.5 billion pixels with digital readout.

The ITS improves the determination of the collision primary vertex initially derived from FIT measurements, allowing to reach a resolution better than $\sim$ 25 $\mu\mathrm{m}$ in both transverse and longitudinal directions. Moreover, the information provided by this detector improves the momentum and pointing resolution for particles reconstructed in the TPC. For instance, the tracks considered in the analysis presented in this thesis are all determined on the basis of both TPC and ITS measurements, with the first detector being more relevant to determine the momentum and the second one driving the determination of the track pointing. The ITS provides crucial information to reconstruct secondary vertices originating from the decay of heavy-flavour hadrons. In particular, the detector upgrades carried out during the LS2 allowed to improve of a factor 2 the impact parameter resolution in the transverse plane for 2-prong and 3-prong decaying heavy-flavour particles, as shown in Fig. 3.4 for D^+ mesons. Finally, the ITS allows to track low-momentum particles ($p < 100$ MeV/c) and to identify them on the basis of their cluster size. The latter is defined as the number of pixels turned on by a single charged particle and it depends both on the particle specific energy-loss $\mathrm{d}E/\mathrm{d}x$ and on the track inclination. The effect of the second factor can be corrected for, so that it is possible to trace back to the particle species thanks to its specific energy-loss via ionisation. However, in this analysis only TPC and TOF measurements are used to perform PID. It is relevant to mention the ITS per-layer material budget, which amounts to 0.36% X_0 for the inner barrel and to 1.10% X_0 for the outer barrel. This, together with the excellent pointing resolution of the detector, allows to reconstruct heavy-flavour particles down to 0 p_T. More details about the ITS can be found in [37] and [38].

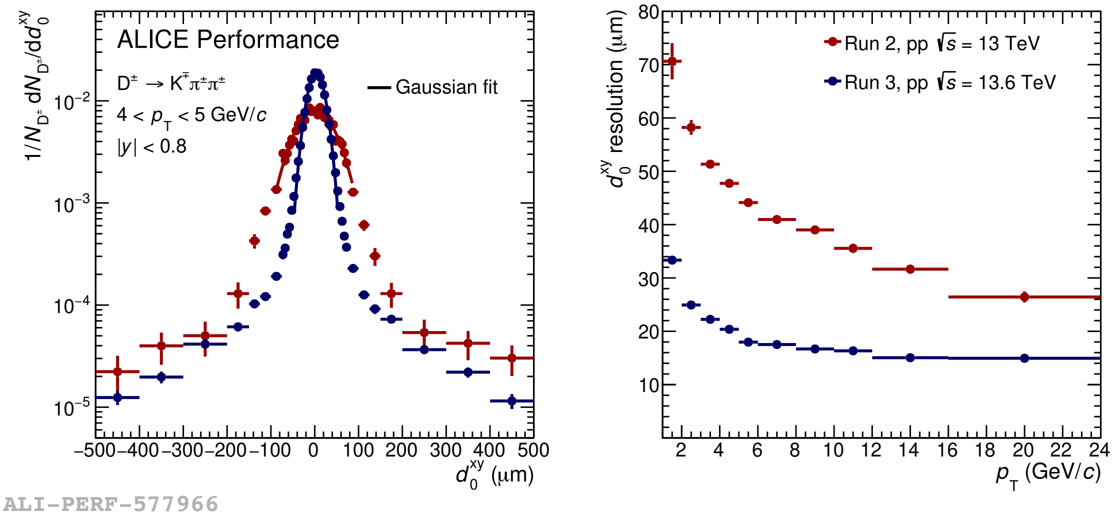


Figure 3.4: Comparison of D^+ meson impact parameter distribution (on the left) and resolution (on the right) in pp collisions in Run 3 with respect to Run 2 performance [31].

3.2.2 Time Projection Chamber (TPC)

Surrounding the ITS is the main tracking device in the central barrel, the TPC. It is a gaseous detector with an inner radius of ~ 85 cm, an outer radius of ~ 250 cm, and a total length of about 5 m. It is depicted in blue in Fig. 3.2, while a more detailed sketch is reported in Fig. 3.5. The TPC has a pseudorapidity coverage of $|\eta| < 0.9$ and full azimuthal coverage. The drift volume is filled with 88 m^3 of a gas mixture composed of Ne, $\mathrm{CO_2}$, and $\mathrm{N_2}$ (90-10-5). A central electrode kept at –100 kV and dividing the drift volume in two halves, together with the field cage, provide a uniform

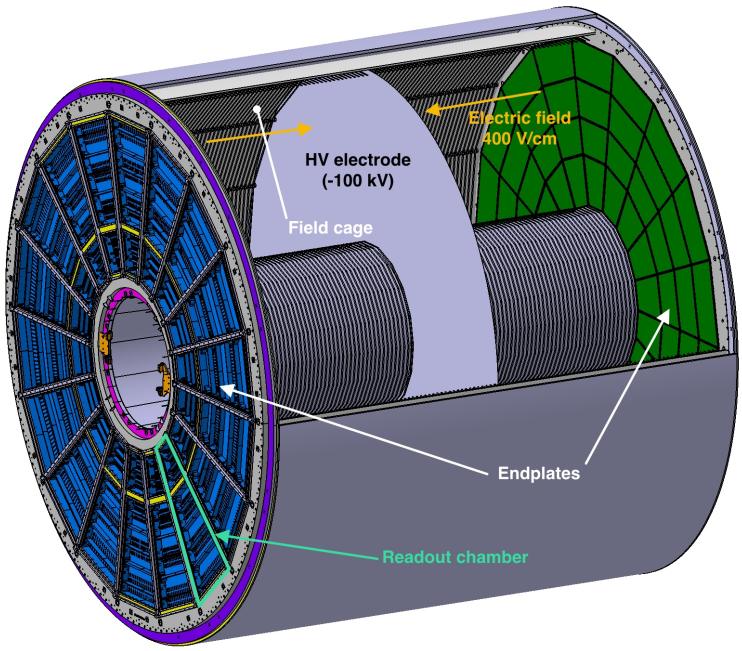


Figure 3.5: Schematic view of the TPC detector [39].

drift field of 400 V/cm oriented along the beam axis. Atoms of the gas in the detector are ionised by interactions with the passing charged particles, and the electrons that are freed then drift to the end-caps of the detector due to the applied electric field. The operational parameters listed above result in the drift velocities of $v_{\mathrm{electron}} \sim 2.6$ cm/μs and $v_{\mathrm{ion}} \sim 1.2$ cm/ms. Consequently, given the detector size, the maximum drift time is $\sim 100\ \mu\mathrm{s}$ for the electrons and of $\sim 200\ \mathrm{ms}$ for the ions. Two end-plates subdivided in 18 azimuthal sectors host the readout chambers. These are based on the Gas Electron Multipliers (GEMs) technology, the reason being that this amplification technique provides sufficient ion blocking without an active ion gate, as it was done in Run 2. The ions produced at the amplification stage at the end-plates tend to flow back into the drift volume towards the central electrode, and distort the electric field. As a consequence, the drift motion of the electrons produced via ionisation is not uniform anymore, thus leading to deviations in the measured spatial points that can reach ~ 10 cm. This phenomenon is known as space-charge distortions. Keeping the ion-induced space-charge distortions at a tolerable level translates into an upper limit of 2% for the fractional ion backflow, i.e. the fraction of ions coming from the amplification of a single primary electron that go back to the drift volume. Such requirement is defined at the operational gas gain of 2000. This is achieved by stacking the GEMs and choosing a hole pattern that avoids holes alignment in subsequent layers, as well as by optimising both the GEMs gain share and their transfer fields. However, at data reconstruction time it is anyway needed to deploy dedicated calibrations to correct for the effect of the leftover space-charge distortions. This phenomenon represents one of the main challenges now that ALICE is taking data in continuous readout mode: as a reference, given the maximum ions drift time, when the machine is operating at 50 kHz there are $\sim 10^4$ collisions partially contributing to the space-charge distribution.

The TPC nominal spatial resolution is $\sim 200\ \mu\mathrm{m}$. This detector determines charged particles momentum based on the measurement of their track curvature and provides particle identification from the low-momentum region ($\sim 100\ \mathrm{MeV}/c$) up to few tens of GeV/c, depending on the particle species to be separated. PID is performed by means of the specific energy-loss (dE/dx) technique, with a resolution of $\sim 5\%$ for isolated tracks. Particle identification is made possible thanks to the proportionality of the energy-loss per unit distance of the original particle to the deposited charge per cluster at the end-plate. More details about the TPC can be found in [39].

3.2.3 Time Of Flight system (TOF)

The TOF detector is a large array of Multi-gap Resistive-Plate Chambers (MRPC) with a gas mixture of Freon and SF_6. As a charged particle crosses a chamber, it ionises the gas and induces a series of electrons avalanches inside the different gaps. Their timing is read out by the detector and this allows to measure, with a precision better than 10^{-10} s, the time that a particle takes to travel from the interaction vertex to the detector. This information, together with the precise collision time provided by the FIT detector, allows to trace back to the particle speed and therefore determine its species. Its PID range is approximately 0.2-2.5 GeV/c, depending on the particle species to be separated. The TOF system is located at a radial distance of 3.7 m away from the nominal collision point, it has a pseudorapidity coverage of $|\eta| < 0.9$ and full azimuthal coverage. This detector is depicted in orange in Fig. 3.2, more details about the TOF can be found in [40].

3.2.4 Fast Interaction Trigger (FIT)

The Fast Interaction Trigger consists of different subsystems positioned at different locations along the beam line: the FT0, the FV0 and the Forward Diffractive Detector (FDD). The first system is the relevant one for this work. The FIT operates as interaction trigger and luminometer, and as a whole provides data for event characterisation, including the temporal measurement used to give a first estimate of the collision vertex location along the beam axis. A schematic view of the FIT is reported in Fig. 3.6.

The **FT0** consists of two arrays of Cherenkov radiators coupled to photomultipliers, the FT0-A, located at ~ 3.3 m away from the nominal interaction point along the beam axis and providing a pseudorapidity coverage of $3.5 < \eta < 4.9$, and the FT0-C, located at ~ 84 cm away from the nominal interaction point along the beam axis and with a pseudorapidity coverage of $-3.3 < \eta < -2.1$. The FT0 system determines the timing of the collision with a resolution of 25 ps, thus providing also information for TOF-based PID. Moreover, given the position of the detectors and of the primary in-

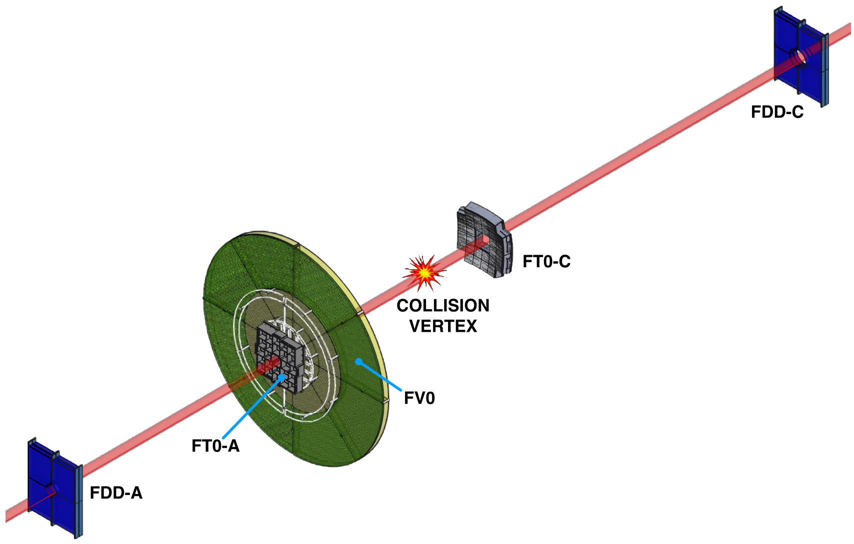


Figure 3.6: Schematic representation of the FIT detectors, with the nominal collision point shown in red and yellow [31].

teraction vertex, the timing measurements performed by the two systems allow to reject background events induced by particles coming from directions other than the signal region, such as beam-gas interactions. The counts of signal coincidences in the FT0-A and FT0-C also allow for luminosity estimation both online, during data taking, and offline, at analysis level. The latter is evaluated as $\mathcal{L}_{\rm int} = N_{\rm TVX}/\sigma_{\rm TVX}$, where $N_{\rm TVX}$ is the number of FT0 trigger counts and $\sigma_{\rm TVX} = 59.4$ mb is the associated visible cross section. Such $\sigma_{\rm TVX}$ is a first estimate provided by the luminosity experts. It is computed as the total Pythia inelastic cross section in pp collisions at $\sqrt{s} = 13.6$ TeV, which amounts to $\sigma_{\rm PYTHIA} = 78.6$ mb, scaled by the TVX efficiency $\epsilon_{\rm TVX} = (75.6 \pm 1.1)\%$. Such efficiency is extracted from a general purpose Monte Carlo simulation by dividing the number of TVX triggers $N_{\rm TVX}$, namely the FT0 coincidences, by the total number of generated events. A more precise value extracted with a more rigorous method will be provided by the luminosity experts in the future. If further event selection criteria are applied, the corresponding scaling factors are taken into account when computing the integrated luminosity analysed. The FT0 system is depicted in grey in Fig. 3.6. More details about the FIT can be found in [37].

3.3 Data reconstruction

In this section, the data reconstruction techniques relevant to the analysis presented in this work are presented.

3.3.1 Event building

Since the beginning of Run 3, ALICE is taking data in continuous readout mode. This strategy requires to define a unit of information, the so-called Time Frame (TF) [41]. This is a time window of configurable length and it represents the minimal processing unit at all stages of the data processing, with each TF being processed independently. The event building is based on the assembly of data recorded during a TF and its length is common to all detectors. In order to minimise the number of events whose data are spread across two consecutive TF, the duration of the time frame is chosen long with respect to the TPC electron drift time ($\sim 100~\mu$s). This reduces the information loss originating from the spread of TPC tracks belonging to the same event among multiple TFs. Consequently, a value of at least 10 ms is foreseen, such that the corresponding loss of statistics is below 1%. The boundaries of the TFs are communicated to the detectors readout electronics via the transmission of non-physics heartbeat triggers. However, each detector splits its data differently to accommodate for its readout rate and segmentation. Therefore, the first stage of data processing consists in assembling the TFs. This operation is carried out by the First Layer Processors (FLPs), a farm of around 200 compute nodes. At this step, a first reduction of the data volume is also done, with the dedicated software, namely O^2, performing local data processing such as cluster and tracklet (track segment) finding. The second level of data aggregation is done by the Event Processing Nodes (EPNs) and consists of global data processing. Examples of these processes are tracks reconstruction and their association to a collision primary vertex. At this stage events are built, with the different interactions falling in the same TF being disentangled. This step also allows for further reduction in the data volume. The output of the EPNs, which is referred to as "raw data", is then sent to storage. The entire data processing described above takes place online. Offline, at a later stage and after the required calibrations are produced, the data reconstruction is carried out and the so-called AO2D files, which are the analysis input, are produced.

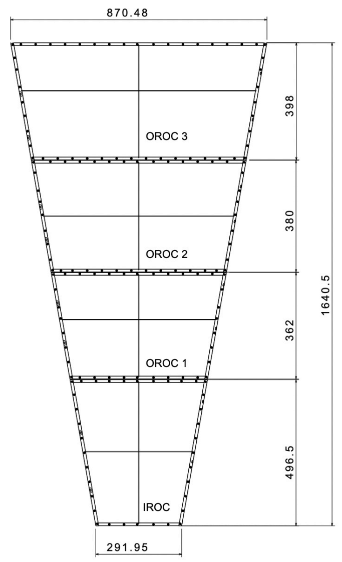
The data stream from the detectors to the FLPs amounts to ~ 4 TB/s, independently on the colliding system. Then, from the FLPs to the EPNs the flow is reduced to ~ 150 GB/s for pp collisions at the interaction rate of 500 kHz and to ~ 700 GB/s for Pb–Pb collisions at 50 kHz. Then, from the EPNs to storage, the data stream amounts to ~ 30 GB/s in the first case and to ~ 140 GB/s in the second one. As the continuous readout mode allows to inspect a huge amount of collisions, the storage space is not enough to permanently store all the raw data. Therefore, every year the whole collected pp raw data is regularly reconstructed and inspected with a set of software triggers. Then, only the raw data corresponding to bunch crossing windows centred on "interesting" events are kept. More details about bunch crossings and analysis object associative logic are given in Section 3.3.2. Moreover, depending on the disk space availability, every year a subset of the collected pp raw data is also permanently stored. Further details about this strategy are given in Section 4.3.

3.3.2 Tracking

Depending on the particle under reconstruction, different strategies and selections are used. As for the analysis presented in this thesis, two methods are applied, one is specific for the pion $\pi \leftarrow \Xi_c^0$ and the other one for the cascade and the V0, namely the Ξ and the Λ.

As for the **pion** coming directly from the charmed-baryon decay, only tracks reconstructed by matching TPC and ITS tracks are considered, these are the so-called global tracks. ITS tracks are defined with four or more hits in the detector. Additionally, at least one hit is required to belong to the ITS inner barrel to ensure that the track originates close to the primary vertex, as the Ξ_c^0 $c\tau$ is of the order of $\sim 45\ \mu\mathrm{m}$. Moreover, the maximum acceptable value for the χ^2 over the number of found ITS clusters is set to 36. As for the TPC tracking requirements, a minimum of 70 clusters is required, the number of TPC crossed rows has to be equal or greater than 70 and the minimum value of the ratio of the number of crossed rows over the number of findable clusters is set to 0.8. Additionally, the maximum acceptable value of the χ^2 over the number of found clusters is set to 4. All these criteria ensure the quality of the track reconstruction and have been validated by the experts. To clarify the TPC selections, the geometry of a TPC readout chamber is depicted in Fig. 3.7. Each chamber is divided in four sectors, namely one Inner ReadOut Chamber (IROC) and three Outer ReadOut Chambers (OROCs). Each of them is further divided in multiple regions, with each region being segmented in tens of rows, and each row being further divided into pads. In total, a single readout chamber consists of 152 rows. A more detailed description is reported in [39], while the explanation of the TPC tracking algorithm can be found in [42].

In Run 3, track-to-collision association is no longer unambiguous because of the continuous readout operation mode. This issue is addressed by deploying a dedicated workflow that reassociates the $\pi \leftarrow \Xi_c^0$ track to all the time-compatible collisions, namely the `track-to-collision-associator` [43]. To understand how this task works, the definition of bunch crossing is needed: the protons accelerated by the LHC are not continuously distributed along the beam, but they are packed into bunches separated in time by at least ~ 25 ns. Each bunch is ~ 9 cm long, has a transverse radius of the order of few $\mu\mathrm{m}$, and contains $\sim 10^{11}$ protons [4]. A "bunch crossing" (BC) is literally a collision of two of such structures and therefore each proton-proton collision - and consequently every track produced in the interaction - will be associated to a BC. Indeed, in Run 3, the associative logic of analysis objects follows a hierarchical structure: given a bunch crossing, multiple pp collisions are associated to it and for each of them sets of analysis objects – such as tracks – can be accessed. A schematic representation of this structure is reported in Fig. 3.8. Due to the high interaction rate, a track time window can overlap in time with multiple collisions. For instance, the ITS has a default



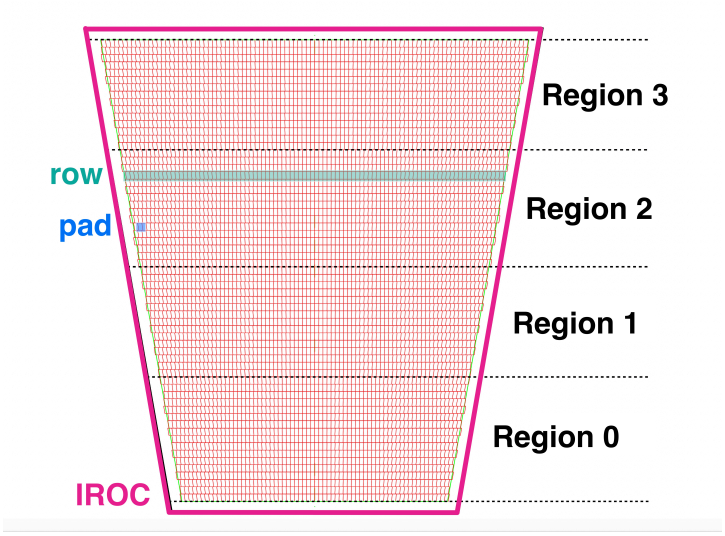


Figure 3.7: Schematic representation of a TPC readout chamber (left plot) and of an IROC pad plane layout (right plot), from [39]. The dimensions are in mm.

track time window of 198 BC, that corresponds to $\sim 4.95\ \mu$s. With an interaction rate of 500 kHz, there is approximately one collision every 2 μs, meaning that an ITS track readout window overlaps with 2.5 collisions on average.

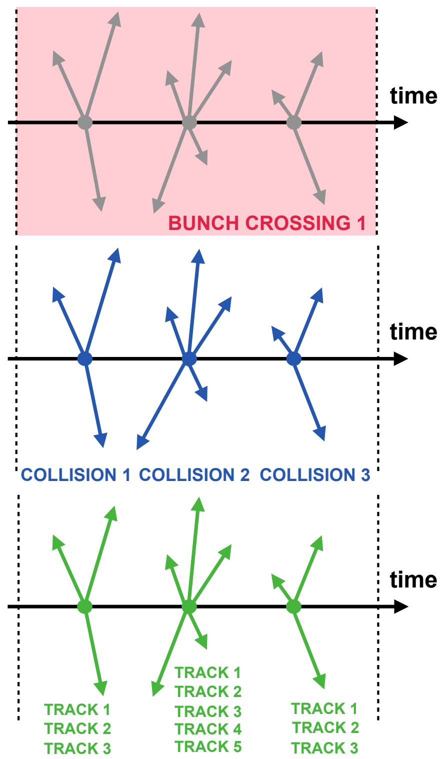


Figure 3.8: Representation of analysis object hierarchical associative logic.

For each event, the `track-to-collision-associator` implements a loop over all the global tracks and, for each track, a loop over the collisions, with the ones lying too far in time being discarded by a hard cut. Let $\mathrm{BC}_{\mathrm{track}}$ be the BC in the middle of the readout window the track belongs to and $\mathrm{BC}_{\mathrm{collision}}$ the BC associated to the collision. As a first step, the task associates to each track all the collisions such that $|\mathrm{BC}_{\mathrm{track}} - \mathrm{BC}_{\mathrm{collision}}| < \mathrm{n}\sigma_{\mathrm{BC}} + \Delta t_{\mathrm{c}}/\Delta t_{BS}$, where $\mathrm{n}\sigma_{\mathrm{BC}}$ is the maximum BC window to be considered, Δt_{c} is a time margin added to account for a not perfectly calibrated TPC, and Δt_{BS} is the bunch crossing spacing (25 ns). Tracks that were not initially assigned to any collision are also associated with this criterion. The values of these variables have been carefully tuned by the experts in order to avoid eliminating possible time-compatible collisions with the track and, at the same time, to discard collisions that are too far in time from the track, avoiding wasting precious computing resources. According to the final settings, the time margin is set to 500 ns, σ_{BC} to 60, and n to 4. Subsequently, a tighter cut on the time difference between the track and the collision is applied according to the criterion $\Delta_{\mathrm{time}} = |t_{\mathrm{track}} - t_{\mathrm{collision}}| < t^*$. Here, t_{track} is either the time associated to the track based on the extrapolation of both TPC and ITS measurements or the collision time itself if the track is primary, while $t_{\mathrm{collision}}$ is the collision time computed as the average of the time associated to all the primary tracks. These time measurements have Gaussian uncertainties σ_{track} and $\sigma_{\mathrm{collision}}$, respectively. The threshold time t^* is computed as $t^* = \sigma_{\mathrm{track}} + \mathrm{n}\sigma_{\mathrm{collision}} + \Delta t_{\mathrm{c}}$. The given track is finally reassociated to all the collisions fulfilling the described requirement. As a result, a track that was initially assigned to only one collision - which is not even the most likely one, by default - can be associated to multiple ones. At analysis level, this has the drawback of increasing the background and, additionally, duplicates the signal if all the daughter tracks undergo the time-based reassociation. However, such duplicates can be discarded at analysis level by implementing topological selections. As for the analysis presented in this thesis, no signal duplication occurs as the cascade, as well as its daughters, are associated to only one collision at AO2D production time. The effect of the time-based track-to-collision reassociation is depicted in Fig. 3.9.

The **cascade** is reconstructed at AO2D production time such that at analysis level both the Ξ and its daughters, together with the collision that produced them, are already identified. Unlike the $\pi \leftarrow \Xi_c^0$ track, neither the Ξ nor its daughters undergo a time-based track-to-collision reassociation, therefore they are associated to only one event. To reconstruct the cascades and the V0s, both ITS-TPC matched tracks and afterburned tracks are used. To clarify these strategies, let's consider the decay reconstructed in this thesis $\Xi \rightarrow \Lambda\,\pi \rightarrow (p\,\pi)\,\pi$. The pion and proton tracks coming from the Λ decay can be either built by matching an ITS track to a TPC track to create a matched track, or by matching a TPC track to unused ITS clusters to create an afterburned track. The latter strategy is of particular relevance for the reconstruction displaced tracks, i.e. weak decays. An

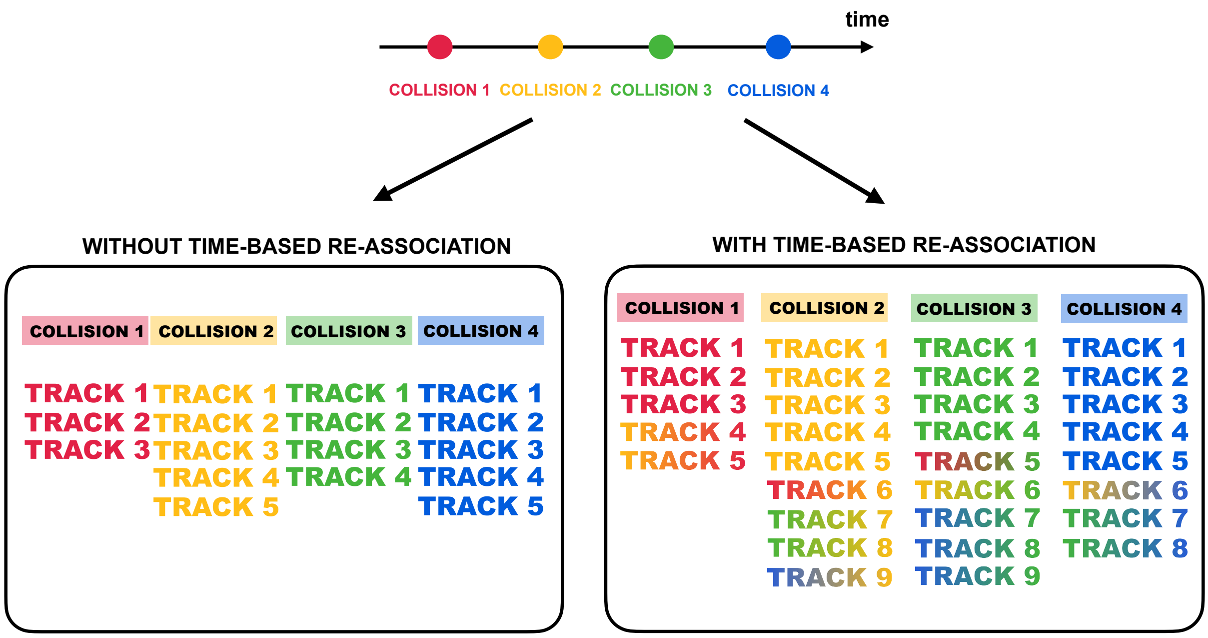


Figure 3.9: Representation of the effect of time-based track-to-collision reassociation, tracks initially belonging to only one collisions may be duplicated to the neighbouring ones. Each colour identifies a collision and tracks reported in multiple colours are associated to multiple events.

ITS track can be defined with four or more ITS hits, while to build an afterburned track only two ITS clusters are needed. A schematic representation of the different types of tracks is depicted in Fig. 3.10. Subsequently, the π and p are combined to create the Λ. At the following step, the track corresponding to the pion $\pi \leftarrow \Xi$ is built with the same approach and then, if loose selection criteria are met, it is combined with the Λ to create the cascade. Further detector signals, such as the TOF measurement, improve the track timing information. Finally, the cascade, together with its daughters, is assigned to the collision they are most likely to belong to, namely the one whose primary vertex yields the best value of cascade cosine of pointing angle. This criterion, based on the decay topology, is expected to allow the cascade-to-collision association with the least ambiguity.

3.3.3 Particle identification

In this work, particle identification (PID) is performed based on TPC and TOF information using a $n\sigma$ criterion.

The **TPC** provides PID on the basis of specific energy-loss measurements. As a charged particle crosses the drift volume, it ionises the gas molecules and the thus produced electrons are collected at the end-plate. Such deposited charge is converted into the particle energy-loss per unit distance $\mathrm{d}E/\mathrm{d}x$, that in turn can be described by the Bethe-Bloch formula. As the energy loss of charged particles in

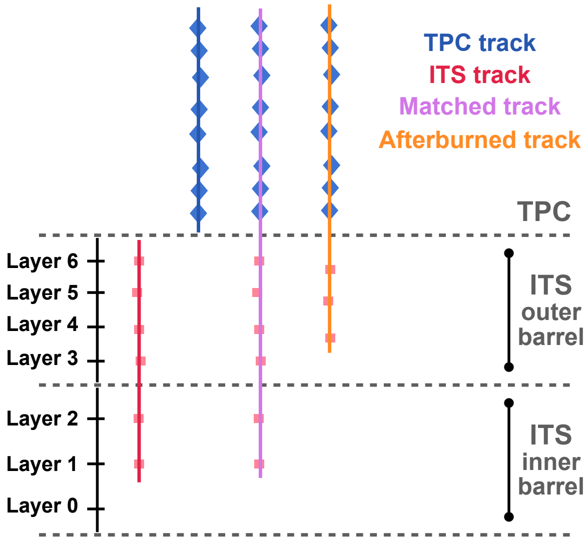


Figure 3.10: Schematic representation of the different types of daughter tracks reconstructed by the cascade finding algorithm.

a material is a statistical process, the energy-loss signal in the detector is described by a statistical distribution around the Bethe–Bloch expectation. This probability follows a Landau distribution, which is characterised by a long tail in correspondence of large values of deposited energy. If, to assign a $< \mathrm{d}E/\mathrm{d}x >$ value to a track, the average of the deposited charge along the segments in which the readout chamber is divided were used, large fluctuations would be observed. In order to get rid of the effects of the Landau tail, a truncated mean is used instead. This strategy allows to retrieve a reliable $< \mathrm{d}E/\mathrm{d}x >$ to be assigned to the track and the final distribution of the thus obtained $< \mathrm{d}E/\mathrm{d}x >$ values can be approximated with a Gaussian function. To perform PID, the $\mathrm{d}E/\mathrm{d}x$ assigned to a track is compared to a theoretical expectation. The latter is computed starting from the ALEPH parametrisation of the Bethe-Bloch formula [44], with the optimal parameters being extracted from fits to data. In particular, clean samples of multiple particle species are used to perform a simultaneous fit. As a first result, a prediction of the mean $\mathrm{d}E/\mathrm{d}x$ signal based purely on the measured momentum and charge of the particles is produced. Then, further corrections accounting for variables affecting the ionisation measurement, such as the track pseudorapidity and the drift length, as well as for detector effects, like the TPC occupancy, are applied with a neural network approach. Finally, the analysers are provided with the momentum-dependent mean and σ parameters of the Gaussian distributions of $\mathrm{d}E/\mathrm{d}x$ measurements for different particle species, with the mean representing the theoretical expectation. At analysis level, the distance in units of resolution (σ) between the $< \mathrm{d}E/\mathrm{d}x >$ value assigned to a given track and the thus predicted corrected theoretical expectation (mean) for a specific particle species is the discriminating variable used to perform PID. An example of $\mathrm{d}E/\mathrm{d}x$ measurements as a function of the particles momentum and charge, together with the ALEPH prediction, is shown in Fig. 3.11.

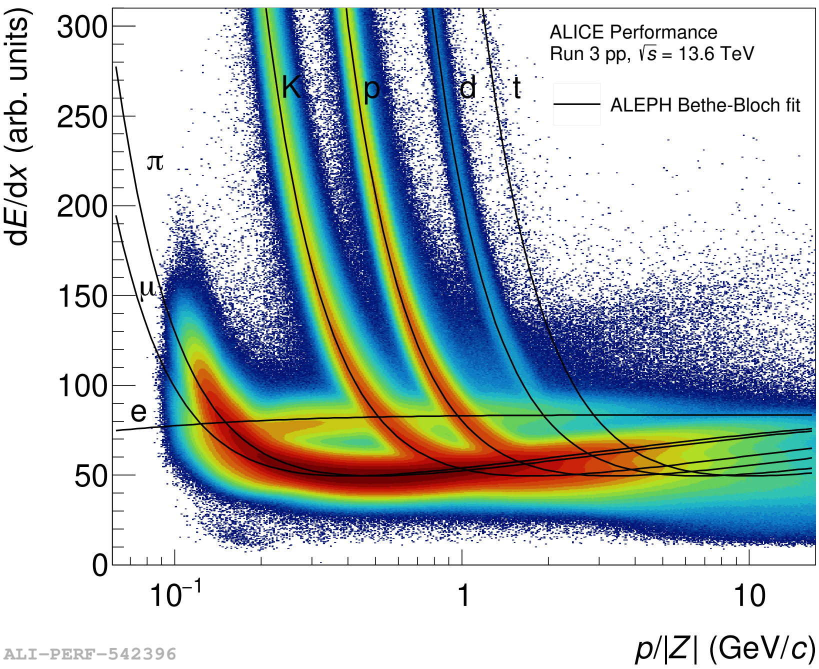


Figure 3.11: Example of Run 3 TPC $\mathrm{d}E/\mathrm{d}x$ measurements as a function of the particle momentum and charge [31]. The ALEPH prediction is also reported.

The **TOF** provides the velocity of a charged particle, or equivalently its β, by measuring its time-of-flight over a given distance along the track trajectory. The time-of-flight is computed as the difference between the reconstructed time of the TOF hit and the collision time measured by the FIT. Particle identification is carried out by comparing the measured time-of-flight to the expected one for a given particle species [45]. The particle mass m is connected to its $\beta\gamma$ according to the formula $m = \frac{p}{\beta\gamma c}$, where p is the track momentum, which is measured by other detectors, and γ is the Lorentz factor. A Gaussian distribution is assumed for the detector response. At analysis level, the distance in units of resolution (σ) between the measured time value and the expected one for a given particle species is the discriminating variable used to perform PID. The σ value used accounts for the intrinsic TOF time resolution, the uncertainty on the collision time and the uncertainty due to the tracking and reconstruction. An example of β measurements as a function of the particle momentum is shown in Fig. 3.12.

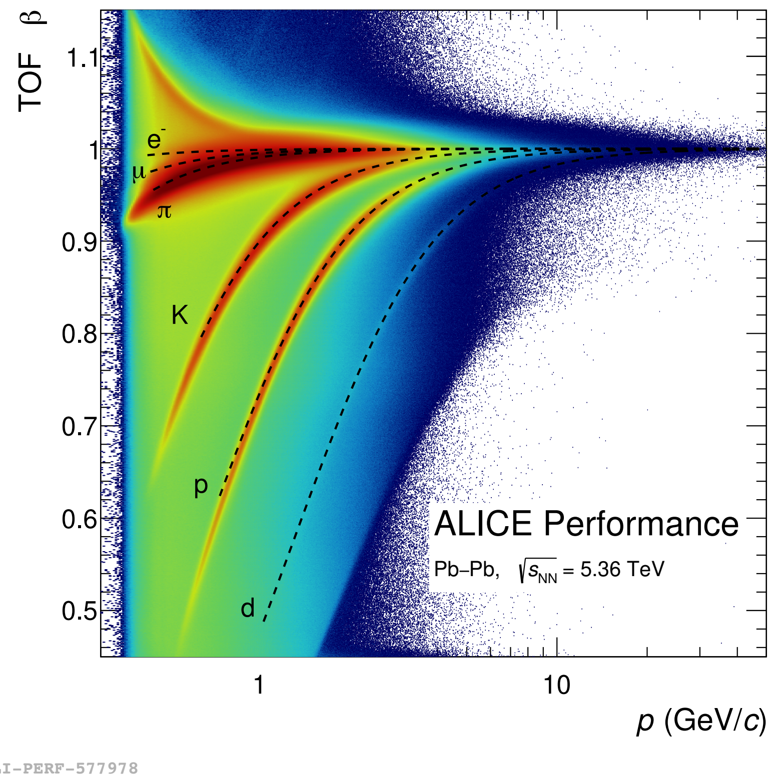


Figure 3.12: Example of Run 3 TOF β measurements as a function of the particle momentum [31]. Theoretical predictions for the different particles species are also shown.

Chapter 4

Software development

4.1 ALICE analysis framework

In 2022, the beginning of Run 3 data taking period, the ALICE experiment started collecting data with an upgraded detector. To cope with the enormous data flow streamed by the new system, which operates in continuous readout mode, a new framework for both online and offline data processing, O^2, was developed. It combines all the functionalities needed in a HEP experiment: detector read-out, event building, data recording, detector calibration, data reconstruction, physics simulation - handled in O2DPG - and analysis - implemented in O2Physics.

The latter is an analysis framework based on arrow tables, i.e. a columnar memory format for flat and hierarchical data optimized for bulk operations, supporting highly modular and extensible data representation. It allows for vectorized optimization of analytical data processing, thus unlocking enormous processing speed. In this setting, data is organized in split - but linked - tables, with ROOT files being still the I/O back end of the framework. These are the key features that make O2Physics a suitable software for Run 3 analyses: the framework employed during Run 2, namely AliPhysics, was relying instead on a different data model, based on the usage of containers with object instances, and it is not complying to the requirements that arose in Run 3.

The choice of this structure is driven by the need of processing a huge volume of data: during Run 3 ALICE has already collected a pp collisions sample corresponding to an integrated luminosity larger than 80 pb^{-1}, while the entire minimum bias sample from Run 2 amounts to $\sim 30\ nb^{-1}$. The data stream from the detectors to the online system reaches almost 4 TB/s independently on the colliding system before zero-suppression is deployed, with the TPC being the major contributor. In pp collisions, when the machine is operating at a rate of 500 kHz, the data rate to storage is reduced to ~ 30 GB/s thanks to partial event reconstruction and data compression. With few PB of data corresponding to $\sim 1\ pb^{-1}$ it is clear that the amount of information that has to be processed at analysis

time turns out to be incomparably large with respect to Run 2, hence the need for a new analysis paradigm.
Within O2Physics, an analysis is structured as a workflow, a collection of tasks that manipulate, process, and produce tables. Each analysis is based on a set of core tasks that implement basic operations, such as track propagation to the primary interaction vertex and event selection, plus a set of tasks that depends on the type of analysis. Given that the different tasks can be factorized, the analysis submission system, *hyperloop*, manages different analyses running on the same dataset at the same time, so that the common parts can be run only once for all the workflows, thus allowing for saving a substantial amount of computing resources. In the following section, the framework dedicated to the reconstruction of heavy-flavour decays to cascade-bachelor pairs is discussed.

4.1.1 Heavy-flavour framework

In the heavy-flavour (HF) framework, a dedicated event and track selection is performed in a common task for all the reconstructed decay channels. The information produced at this first step together with the content of the input files created at reconstruction time, the AO2Ds (Analysis Object Data), is then fed into a *decay vertex finder* to produce derived data. This task combines tracks, computes decay vertices and applies loose selections to the decay topologies. More details about derived data are given in Section 4.2, as for now it is sufficient to think of it as a new table whose content allows to access directly the identity of the daughters of all the reconstructed heavy-flavour particles, i.e. their indexes, in the original AO2D files.
Heavy-flavour decays can exhibit multiple topologies: besides the two and three prong channels, HF particles can also decay to V0s, i.e. short-lived neutral particles that in turn weakly decay to a pair of charged tracks forming a characteristic "V" shape in tracking detectors, as well as to cascades, namely short-lived charged particles that weakly decay producing a chain of successive decays that resembles, precisely, a "cascade". The remaining track(s) originating from the decay vertex of the HF hadron in case of channels including either a V0 or a cascade is referred to as "bachelor". The derived data production for HF decays to two prong, three prong, and V0s plus bachelor was already available in the framework. For this thesis, the production of derived data for charmed-baryon decays to cascade plus bachelor has been implemented for the following decay channels:

- $\Xi^0_c \rightarrow \Xi^- \, \pi^+ \rightarrow (\Lambda\pi^-)\pi^+ \rightarrow ((p\pi^-)\pi^-)\pi^+ \;\; + \;\; c.c$
- $\Omega^0_c \rightarrow \Xi^- \, \pi^+ \rightarrow (\Lambda\pi^-)\pi^+ \rightarrow ((p\pi^-)\pi^-)\pi^+ \;\; + \;\; c.c$
- $\Omega^0_c \rightarrow \Omega^- \, \pi^+ \rightarrow (\Lambda K^-)\pi^+ \rightarrow ((p\pi^-)K^-)\pi^+ \;\; + \;\; c.c$
- $\Omega^0_c \rightarrow \Omega^- \, K^+ \rightarrow (\Lambda K^-)K^+ \rightarrow ((p\pi^-)K^-)K^+ \;\; + \;\; c.c$

- $\Xi_c^+ \rightarrow \Xi^- \pi^+ \pi^+ \rightarrow (\Lambda\pi^-)\pi^+\pi^+ \rightarrow ((p\pi^-)\pi^-)\pi^+\pi^+ \quad + \quad c.c$

The data processing carried out in the following part of the analysis is based on the information produced by the decay vertex finder. The final part of the workflow consists of a set of dedicated files specific for the decay channel under investigation, their structure can be schematized as follows:

- candidate creator, implementing the reconstruction of the heavy-flavour decay vertex and computing the relevant topological and kinematic variables
- candidate selector, applying specific selections on the variables computed at the previous step and on daughters PID information
- tree creator, that allows to save to a ROOT TTree the information needed for the offline analysis

In Fig. 4.1, a schematic representation of the structure of the workflow implementing the reconstruction of heavy-flavour baryons decaying to a cascade-bachelor pair is reported.
This whole workflow - from derived data to TTree production - has been implemented for this thesis for all the Ξ_c^0 and Ω_c^0 decays listed above.

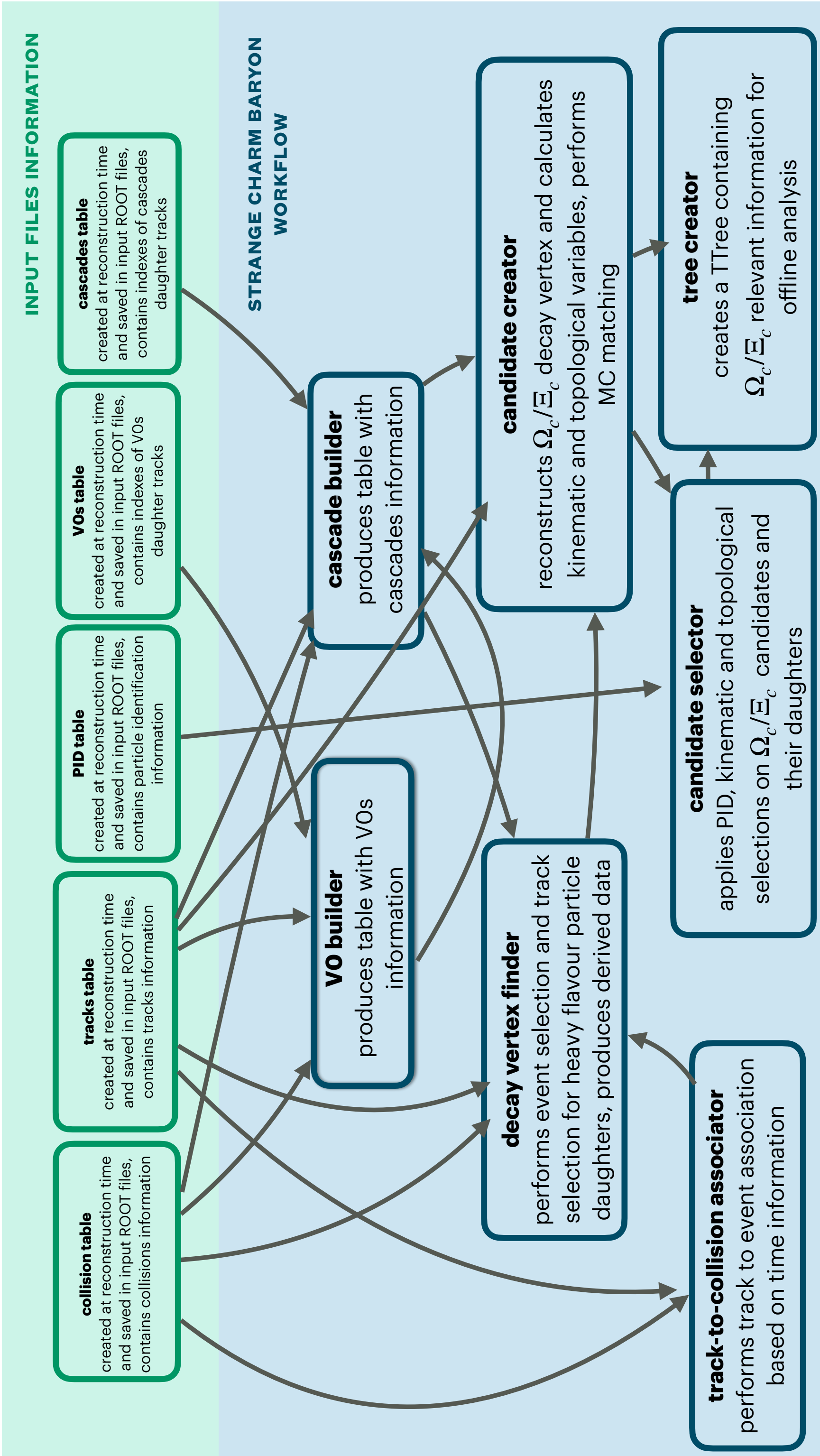


Figure 4.1: Simplified analysis workflow structure for heavy-flavour baryons decaying to cascade plus bachelor.

4.2 Derived data

The data format employed in O2Physics is based on the usage of flat tables, and derived data are essentially nothing else than tables storing indexes information and providing access to the original parent files. To better understand why such data format is needed, it is necessary to highlight the main challenge that ALICE analyses are facing in Run 3: the requirement of processing a huge amount of data. This results in the usage of an enormous amount of computing resources, especially in terms of CPU time. As a consequence, the world-wide computing grid for analyses submission (hyperloop) restricts the computing resources available per user and, when such limits are not met, the system prevents the analysers from submitting their workflow.

The reconstruction of particles that are not directly tracked by the experiment, such as neutral and/or short-lived particles, relies on the combinatorics of objects representing their decay products, which essentially translates in implementing `for` loops over the list of candidate daughter tracks. As a consequence, if the `Tracks` table stored in the original files contains n elements, the reconstruction of a two prong decay, such as the decay of a V0, will require a number of iterations of the order of n^2. If then we would like to reconstruct a cascade, we will need to iterate over the V0s built at the previous step combining them again with another track, and this will turn out in a number of iterations of the order of n^3, and so on. All the decay chains listed above will therefore require $\sim n^4$ iterations. This reconstruction process is extremely expensive in terms of CPU time - actually, it is one of the most consuming ones - and this is where derived data comes into play.

Given that there are multiple charmed-baryon decays to cascade-bachelor, it is possible to take into account all the possible mother particle, cascade and bachelor species and reconstruct them in the same loop, with a flag that keeps track of the channel-specific selections that each combination fulfils. After reconstructing the charmed-baryon, the indexes of its daughters in the original input files are stored together with the corresponding collision index and a bit mask (i.e. the channel-specific flag) in a new table, the derived data table. This ensures that at a later stage of the analysis, it will still be possible to retrieve the information about the daughters and the event from the parent file. The step of derived data production also includes the implementation of event and track selections beforehand. This reduces the number of combinations and excludes candidates that would be discarded at analysis time, thus saving resources. The process described above is implemented in the decay vertex finder task.

When the analysers will execute their workflow for charmed-baryon reconstruction, they will run the analysis on the derived table produced by the decay vertex finder rather than on the original `Tracks` table. Now the advantage of derived data becomes clear: a consistent part of various analyses - the tracks combinatorics - is factorized in the derived data production and multiple analysers can profit from the execution of a single task, i.e. the decay vertex finder, instead of implementing their own version of the cascade-bachelor combinatorics.

The resources needed for different configurations of the $\Xi_c^0 \rightarrow \Xi^- \pi^+ + c.c.$ analysis are listed in Table 4.1 to directly compare the full workflow run on regular input files to the same analysis performed on derived data. The throughput increased from slightly less than 0.5 MB/s/core to 1.7 MB/s/core when switching from standard AO2D files to derived data as analysis input, and the total processing time (wall time) is reduced from more than 1 year to slightly more than 111 days. The throughput is independent on the dataset size, includes both CPU time and I/O time, with the latter being a negligible contribution to the total value. Its increase is due to the fact that the analysis on derived data requires to include less tasks in the workflow, such as the decay vertex finder. As for the wall time, this is real time to be "redistributed" over multiple CPUs.

All the workflows reconstructing charmed-baryons decays to cascade-bachelor can exploit the same derived data production: it is sufficient to check on the channel-specific bit of the flag initialized by the decay vertex finder, depending on decay of interest, to select suitable daughter pairs to be further processed. For this thesis, the derived data production for all the channels listed in Section 4.1.1 has been implemented.

Process	Analyis on parent input files	Analysis on derived data
Wall time	1y 32 d	111d 5h
Throughput	0.5 MB/s/core	1.7 MB/s/core

Table 4.1: Comparison of computing grid resources needed for charmed-baryon analysis on parent input files and on derived data.

4.3 Software trigger

One of the major advances of the ALICE experiment with respect to Run 2 is the capability to operate in continuous readout mode. This translates into an enormous data stream to storage, of the order of tens of GB/s when the machine is delivering pp collisions at a rate of 500 kHz, and the current storage capabilities are not sufficient to allow for the permanent storage of all the collected pp data. To address this issue, the ALICE strategy relies on the deployment of a set of offline software triggers: this approach involves to regularly reconstruct with a physics-ready calibration all the collected pp data within few weeks and then to run an asynchronous event selection. This is done by the Central Event Filter Processor (CEFP), a series of analysis tasks that tags every reconstructed collision containing an event of interest. The raw data corresponding to bunch crossing windows centred on "interesting" events are permanently stored, while everything else is deleted in order to free disk space. The availability of raw data ensures that once improved calibrations or a new software version are ready, a new data reconstruction can be carried out, providing the analyser with state-of-the-art version of data collected during the whole Run 3.

Within this thesis, profiting from this strategy, a dedicated software trigger to search for Cabibbo-suppressed decays of strange charmed-baryons has been implemented:

- $\Omega_c^0 \to \Xi^- \, \pi^+ \to (\Lambda\pi^-)\pi^+ \to ((p\pi^-)\pi^-)\pi^+ \;\; + \;\; c.c$
- $\Omega_c^0 \to \Xi^- \, K^+ \to (\Lambda\pi^-)K^+ \to ((p\pi^-)\pi^-)K^+ \;\; + \;\; c.c$

The corresponding datasets will be used to perform branching ratio fractions measurements of the Ω_c^0 in the coming years. These results will provide a tool to validate theoretical predictions and to shed light on the tension between existing measurements published by BELLE [46] and LHCb [47].

The Run 3 software trigger campaign has an integrated luminosity target of 200 pb^{-1}, the $\mathcal{L}_{int}$ collected during every year of Run 3 is shown in Fig. 4.2. The Ω_c^0 dedicated trigger has been deployed starting from 2023 and the study of these rare decay channels will greatly benefit from the availability of such a large data sample.

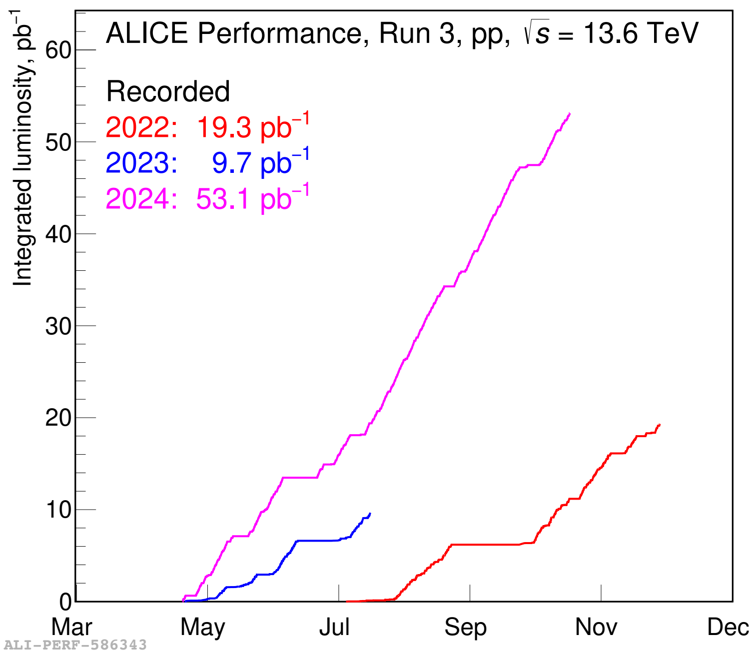


Figure 4.2: Integrated luminosity collected in pp collisions during Run 3 in 2022, 2023, and 2024 [31].

The requirement that the set of heavy-flavour triggers have to fulfil is to reach a total selectivity not higher than $\sim 5 \cdot 10^{-5}$, meaning that it is not possible to select more than ~ 5 collisions out of 100 000. The trigger-specific selections have been tuned in order to meet this limit. However, when the selectivity turned out to be too high, a down-scaling factor was applied to the corresponding trigger. This is the case for the rare charmed-baryon trigger: only one out of two events fulfilling the selections is kept. In Fig. 4.3 the selectivity for all the heavy-flavour triggers on 2023 data is shown, with the last bin (`kCharmBarToXiBach`) corresponding to the Ω_c^0 trigger.

The charmed-baryon dedicated trigger essentially consists in the implementation of a per-collision

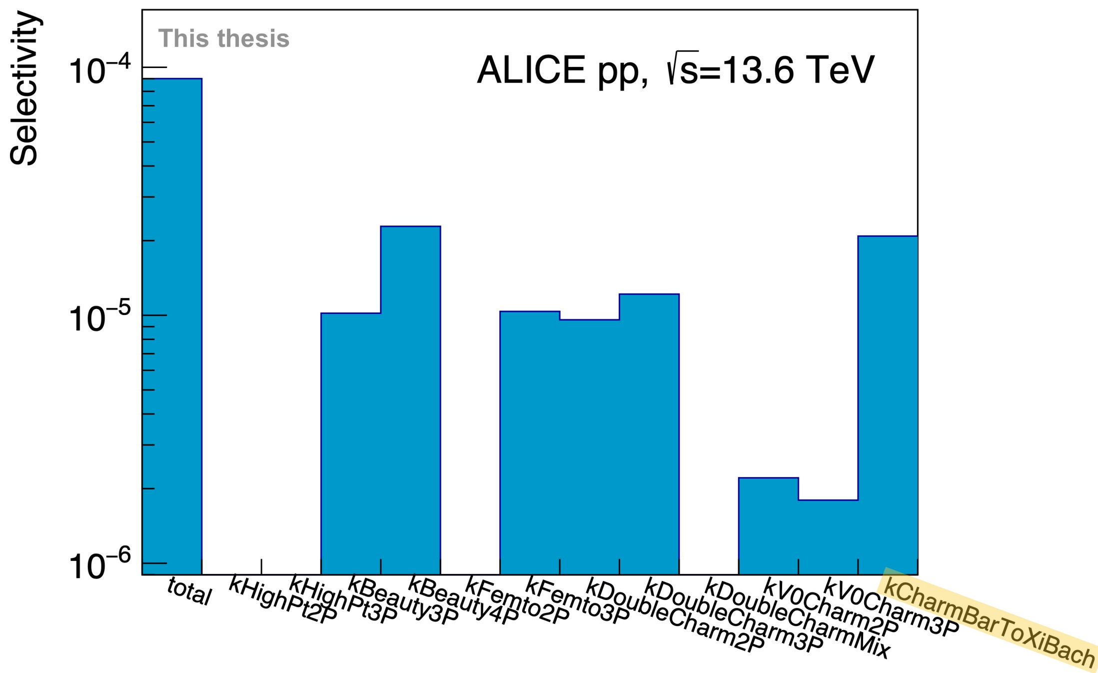


Figure 4.3: Heavy-flavour trigger selectivity on 2023 data. The Ω_c^0 dedicated trigger is shown in the last bin of the histogram, the downsampling factors are not shown.

cascade-track pair combinatorics and a selection on their invariant mass. The trigger tags the collisions containing at least one charmed-baryon candidate with $p_{\mathrm{T}} > 5$ GeV$/c$ fulfilling the selections listed in Table 4.2. As for the candidate bachelor pion $\pi \leftarrow \Xi_c^0$, only tracks built as matched ITS-TPC tracks have been used. The p_{T} cuts have been tuned based on the output of a Pythia toy simulation where the charmed-baryon is forced to decay to the desired channel. These studies are shown in Fig. 4.4 for the $\Omega_c^0 \to \Xi\pi$ mode: 10^7 decays are generated, with the input Ω_c^0 p_{T} distribution being uniform and ranging from 0 to 50 GeV/c. The majority of the mother momentum is carried by the heaviest daughter, i.e. the Ξ. Given that the software trigger selects Ω_c^0 candidates with $p_{\mathrm{T}} > 5$ GeV/c, the daughter kinematic cuts have been chosen as $p_{\mathrm{T}\,\Xi} \geq 2$ GeV/c and $p_{\mathrm{T}\,\pi\leftarrow\Xi} \geq$ 0.2 GeV/c, with a softer selection on the lighter daughter, namely the pion.

The events selected by the software trigger fulfil the following selections:

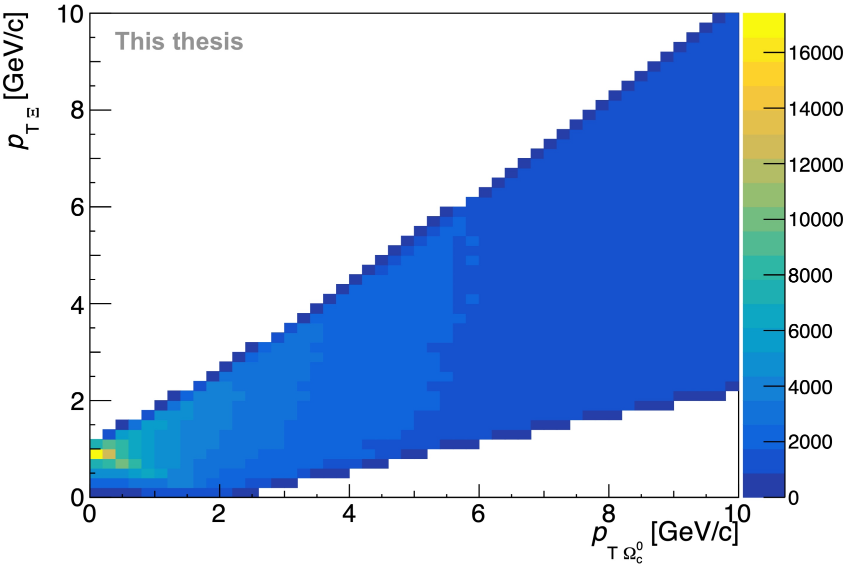
- TVX trigger: the time difference between the signals coming from the FT0A and FT0C detectors is acceptable. This requirement allows to remove background signals induced by particles coming from directions other than the signal region, i.e. the nominal collision point.

- The coordinate along the beam axis (z) of the reconstructed primary vertex (PV) is within 11 cm from the nominal collision point, i.e. -11 cm < z_{PV} < 11 cm. This is a loose requirement that allows to have a uniform tracking acceptance for all the selected events. This cut is normally set to 10 cm at analysis level, but in order to have a margin to estimate the impact of the selection on z_{PV} the cut value has been chosen slightly looser. Moreover, the triggered data can undergo new reconstruction processes and this can result in variations of the reconstructed primary vertex position. A looser selection on this variable allows to avoid biases in case of new reconstructions and to not discard events with a PV that would fulfil the analysis event selections at a later stage.

- Events close to the Time Frame (TF) borders and events close to the ITS Read-Out Frame (ITSROF) are discarded. This allows to remove collisions whose information is spread over two neighbouring minimal processing units and consequently can't be properly reconstructed with the algorithm currently used. In pp collisions, the fraction of events affected by the TF borders cut is of the order of a few percent, with small variations depending on the dataset. As for the ITSROF borders cut, $\sim 15\%$ of the events are impacted.

Variable	**Criterion**
R_{Ξ}	$\geq$ 0.6 cm
R_{Λ}	$\geq$ 1.2 cm
$\cos\theta_{\Xi}$	$\geq$ 0.99
$\cos\theta_{\Lambda}$	$\geq$ 0.99
|DCA| of Ξ daughters tracks	$\leq$ 1 cm
|DCA| of Λ daughters tracks	$\leq$ 1 cm
$\lvert\text{DCA}\rvert_{xy}$ Ξ to PV	$\leq$ 0.3 cm
$\lvert\eta\rvert$ of Ξ and Λ daughter tracks	$\leq$ 1
$p_{\mathrm{T}\ \pi\leftarrow\Xi}$	$\geq$ 0.2 GeV/c
$p_{\mathrm{T}\ \Xi}$	$\geq$ 2 GeV/c
$\lvert m_{\Xi} - m_{\Xi_{PDG}}\rvert$	$\leq$ 0.01 GeV/c
$\lvert m_{\Lambda} - m_{\Lambda_{PDG}}\rvert$	$\leq$ 0.01 GeV/c
$\lvert\eta\rvert_{\pi\leftarrow\Xi_c^0}$	$\leq$ 0.8
$p_{\mathrm{T}\ \pi\leftarrow\Xi_c^0}$	$\geq$ 0.5 GeV/c
$\lvert\text{DCA}\rvert_{xy\ \pi\leftarrow\Xi_c^0}$ to PV for $p_{\mathrm{T}\ \pi} \leq 2$ GeV/c	$\leq$ 0.2 cm
$\lvert\text{DCA}\rvert_{xy\ \pi\leftarrow\Xi_c^0}$ to PV for $p_{\mathrm{T}\ \pi} > 2$ GeV/c	$\leq$ 10 cm
$\lvert\text{DCA}\rvert_{z\ \pi\leftarrow\Xi_c^0}$ to PV	$\leq$ 2 cm
PID nσ cut of final state tracks	3σ
Number of found TPC clusters of final state tracks	$\geq$ 70
Number of TPC crossed rows of final state tracks	$\geq$ 70
Ratio crossed rows over findable clusters of final state tracks	$\geq$ 0.8
TPC χ^2 / number of clusters of final state tracks	$\leq$ 4
Number of found ITS clusters of $\pi \leftarrow \Xi_c^0$ track	$\geq$ 4
Number of found ITS inner barrel clusters of $\pi \leftarrow \Xi_c^0$ track	$\geq$ 1
ITS χ^2 / number of clusters of $\pi \leftarrow \Xi_c^0$ track	$\leq$ 36

Table 4.2: Ω_c^0 dedicated trigger selections for cascade and charm bachelor. The PID $n\sigma$ cut is implemented as a logic OR of the TPC and TOF information. More details about the variables definitions are provided in Appendix A.



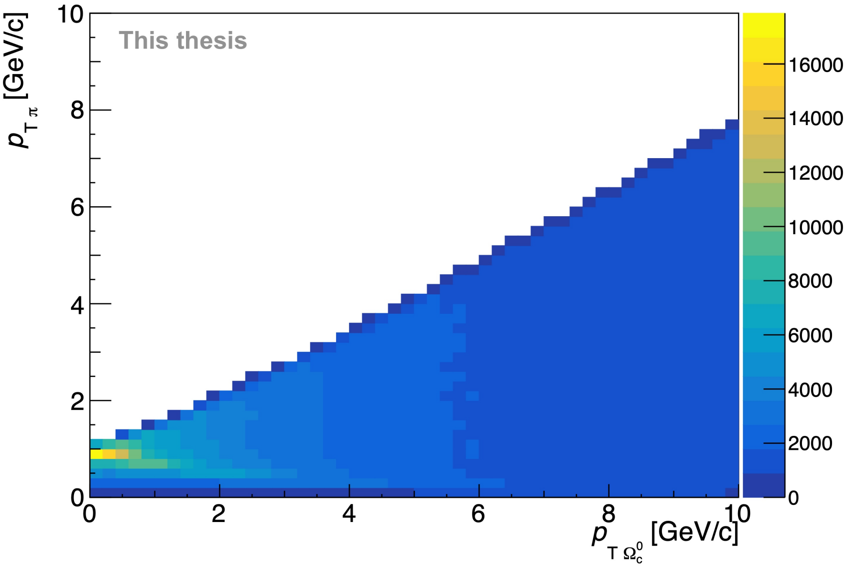


Figure 4.4: Transverse momentum distributions of Ξ and π coming from Ω_c^0 decays, output of Pythia toy simulation of 10^7 decays.

4.4 Simulation studies

The continuous operation mode requires the development of Monte Carlo simulation methods that realistically describe the data taking conditions. As a consequence, new strategies are needed with respect to Run 2, where single events could be "isolated" as ALICE was taking data in triggered mode. Different types of Monte Carlo simulations have been taken into account and are illustrated in Fig. 4.5:

- general purpose: the particle yields reflect the Pythia cross sections
- injected: the desired particle is injected on top of a Pythia minimum bias event
- triggered: Pythia events are generated until the desired signal is produced and only such events are kept
- gap-triggered: similar to the triggered Monte Carlo, the difference is that in between two triggered events a given number of minimum bias events are generated and kept (the "gap")

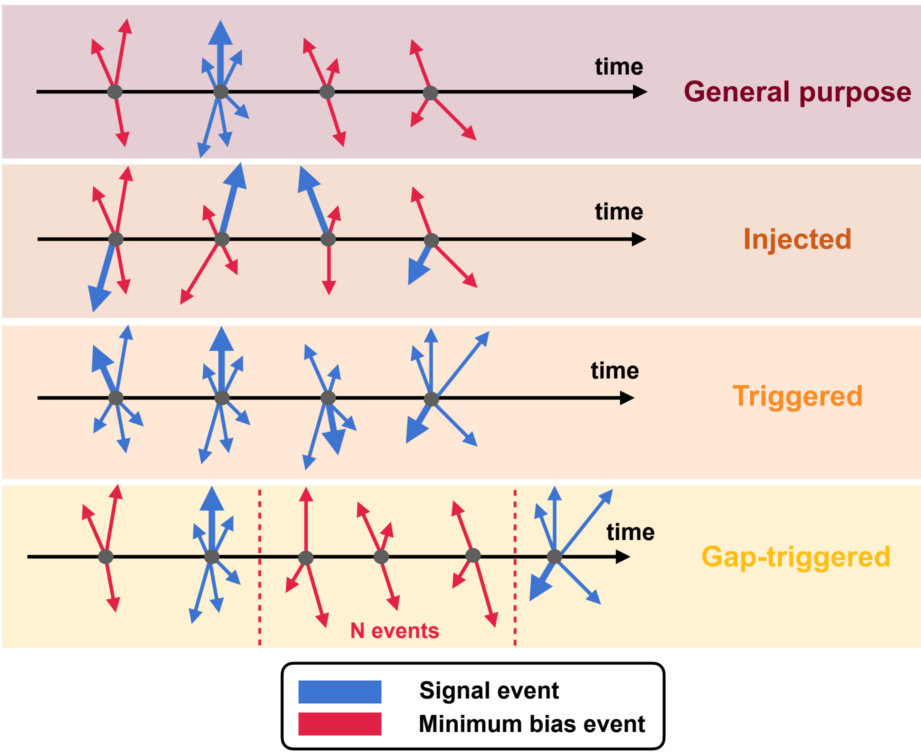


Figure 4.5: Different Monte Carlo simulation strategies. Minimum bias events are shown in red and signal events in blue.

The general purpose approach is not suitable for rare particle studies, since the simulation would have to run for a prohibitively long time before accumulating a sufficiently large signal sample. In injected

Monte Carlo, the multiplicity of the event is not realistic as it does not necessarily match the characteristics of events where charm production occurs. Triggered Monte Carlo also proved to be suboptimal, as the presence of the heavy quark affects the reconstruction of the event, worsening it and reducing the charmed-hadron reconstruction efficiency itself. The reason for these efficiency losses is that the presence of slightly displaced secondary tracks coming from extremely short-lived particles, such as charmed-hadrons, leads to the reconstruction of fake "split" primary vertices. The presence of beauty also leads to similar effects, as illustrated in Fig. 4.6. Typical $c\tau$ values are $\mathcal{O}(10^2)$ μm and $\mathcal{O}(10^1)$ μm for beauty-hadrons and charmed-baryons, respectively [4]. The $c\tau$ characteristic of charmed-mesons is also $\mathcal{O}(10^2)$ μm, but these particles still exhibit shorter mean lifetimes compared to beauty-hadrons [4]. The larger displacement of non-prompt charmed-hadron and beauty-hadron decay vertices makes the situation even worse. What we expect is that the more the heavy quark decay vertex is displaced, the more likely it is to reconstruct fake split vertices. Keeping in mind that to each collision corresponds a single primary vertex, the ratio of reconstructed collisions to generated collisions was computed for general purpose, charm-enriched, and beauty-enriched Monte Carlo simulations. It was verified that the ratio is higher for beauty-enriched productions, slightly lower for the charm-enriched ones, and even lower for general purpose simulations, as expected. However, it has been observed that introducing a sufficiently large gap of minimum bias events in

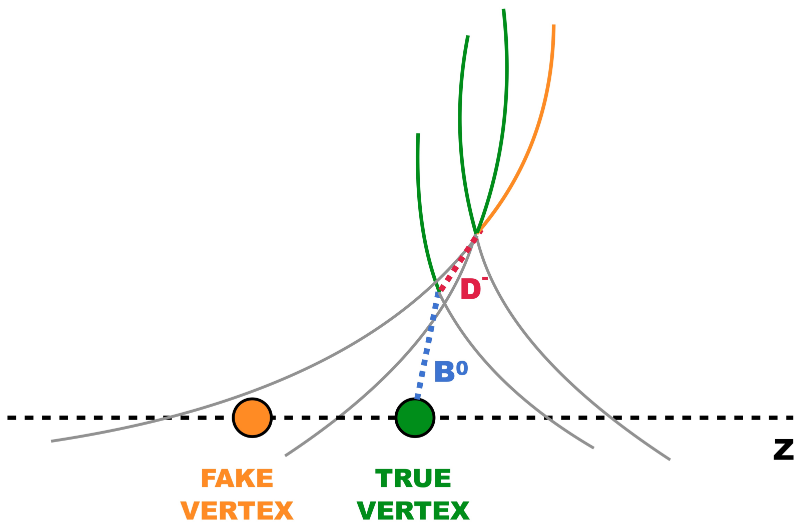


Figure 4.6: Schematic representation of the reconstruction of fake split primary vertices in charm and beauty enriched Monte Carlo productions. The green tracks are correctly assigned to the true collision vertex, while the orange track is assigned to the fake collision vertex.

between every pair of triggered events can help to mitigate the problem: it improves the primary vertex reconstruction and allows to partially recover the reconstruction efficiency for heavy-flavour particles. This is exactly the approach of gap-triggered simulations, which proved to be the most realistic strategy to reproduce the data taking conditions. Moreover, to ensure that tracks are associated to the correct collision and to further partially recover the charm reconstruction efficiency, tracks themselves are also reassociated to all the time-compatible primary vertices by a dedicated task at analysis level. As already discussed in Section 3.3.2, this has the downside of duplicating signals for two prong and three prong decays but does not affect decays to cascade-bachelor pairs. Each cascade is associated to only one collision at reconstruction time, so the time-based reassociation only has the effect of increasing the combinatorial background.

Nevertheless, the gap-trigger strategy requires dedicated studies to choose a proper gap size. The goal is to find a balance between the computing resources needed to run the simulation and the mitigation of the effect described above: a larger gap would be more realistic and would diminish the impact of the heavy quark presence on the event reconstruction, but also comes at the cost of a larger-size simulation and higher CPU consumption.

In order to optimize the settings for the analysis presented in this thesis, three centralized anchored proton-proton Monte Carlo productions for the decay channel $\Xi_c^0 \to \Xi\,\pi$ have been used to study how the charmed-baryon reconstruction efficiency varies with the gap size. In particular, the gap sizes 3, 5, and 8 have been tested, with a gap size of 3 corresponding to 2 minimum bias events in between two consecutive triggered events. We expect to see that the mitigation effect tends to saturate at increasingly larger gaps.

The results of this study are shown in Fig. 4.7 and Fig. 4.8. Thanks to the time-based track-to-collisions reassociation, the observed differences between the investigated gap configurations are smaller than $\sim 10\%$. At reconstruction level, the performance corresponding to the three different sizes is compatible within the uncertainties in almost all the p_{T} bins. At the selection step and in the low/intermediate transverse momentum region, the outcome of this study suggests that the larger gap sizes of 5 and 8 result in a better performance compared to the gap 3 configuration, but uncertainties are large. Larger Monte Carlo productions would be needed to improve the statistical precision and pinpoint significant variations in the efficiencies for the different configurations, thus allowing to draw a definitive conclusion.

Given that no extra resources were available to extend these test productions, the gap 5 setting has been chosen as default for charmed-baryons-enriched Monte Carlo simulations. This configuration is a compromise between the need to increase the gap size as much as possible and the limited computing resources.

An extended centralized production anchored to a specific data reconstruction was then requested in order to perform the Ξ_c^0 analysis presented in this thesis. A total of ~ 250 million events with a gap 5 configuration were simulated, which resulted in collecting ~ 50 million of events containing the desired signal. This production ran for roughly one month at 10k CPU and the output file size turned out to be ~ 15 TB. The particles requested by the trigger of this production were set to Ξ_c^0, Ξ_c^+, and Ω_c^0, with the three baryons having an equal share of the total number of signal events.

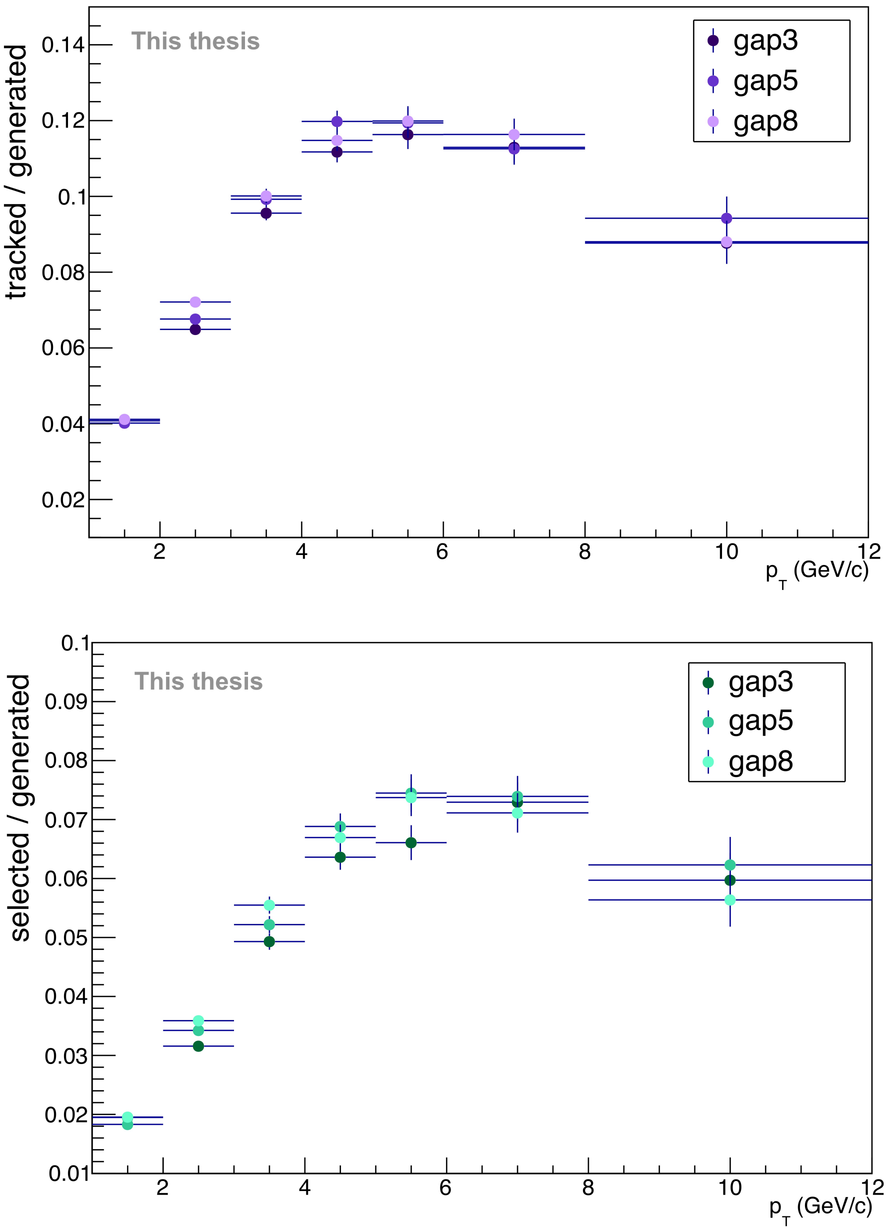


Figure 4.7: Comparison of the Ξ_c^0 acceptance-times-efficiency as a function of the Ξ_c^0 transverse momentum for different proton-proton Monte Carlo gap settings. The tracking $A \cdot \epsilon$ is shown in the top plot and the selection $A \cdot \epsilon$ is reported in the bottom plot. Generated Ξ_c^0 are within $|y|<0.5$ and decay to the desired channel, tracked Ξ_c^0 are the reconstructed Ξ_c^0, selected Ξ_c^0 are the reconstructed Ξ_c^0 passing the analysis-level selections listed in Section 5.4.

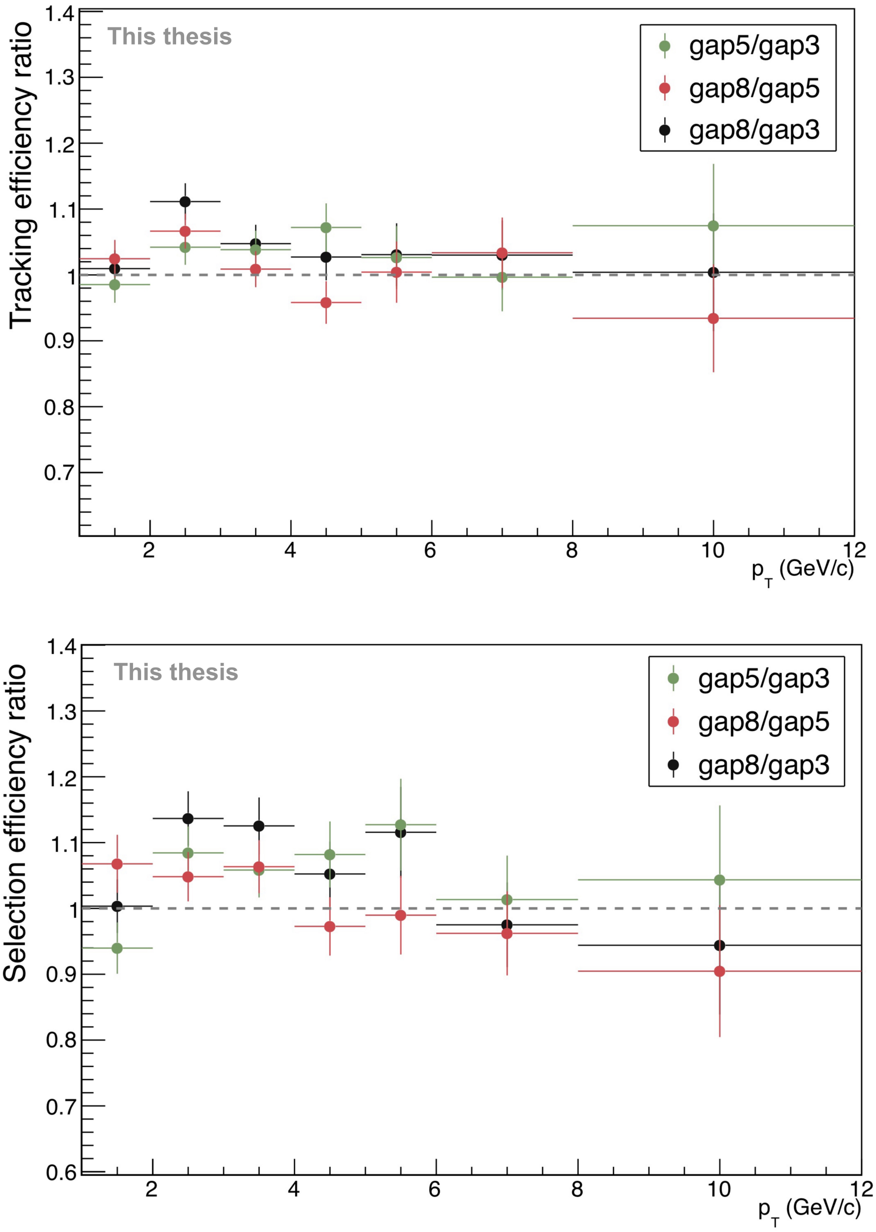


Figure 4.8: Ratios of the Ξ_c^0 acceptance-times-efficiency as a function of the Ξ_c^0 transverse momentum for different proton-proton Monte Carlo gap settings. The top plot shows the results for the tracking $A \cdot \epsilon$, while the bottom plot reports the results for the selection $A \cdot \epsilon$.

Chapter 5

Data analysis

5.1 Analysis Strategy

The analysis presented in this thesis aims at measuring the prompt Ξ_c^0 cross section using the hadronic weak decay channel $\Xi_c^0 \to \Xi^- \pi^+ \to (\Lambda\,\pi^-)\,\pi^+ \to ((p\,\pi^-)\,\pi^-)\,\pi^+ \ \ + \ \ c.c.$

Each decay of the chain is reconstructed by implementing the combinatorics of the candidate daughter tracks and the corresponding decay vertex is reconstructed using the DCAFitter algorithm. More details about this method are given in Section 5.1.1.

The Ξ_c^0 analysis is affected by the issue of a poor signal-over-background ratio. This is partly due to the fact that it aims at extracting a rare signal, but also to the presence of a large combinatorial background originating from the three steps of track combinatorics needed to build the Ξ_c^0, Ξ and Λ. This challenge is met by excellent vertexing and particle identification capabilities. Moreover, the availability of a large data sample allows to improve the significance of the signal extraction. The Boosted Decision Tree (BDT) algorithm is applied to process the reconstructed candidates and reject background events, thus allowing to increase the statistical significance of the signal extraction and to improve its stability. This machine learning method is discussed in Section 5.1.2.

5.1.1 DCAFitter

Secondary-vertex reconstruction algorithms are implemented in the `DCAFitterN` class in the AliceO2 software [48]. This code determines the short-lived particle decay vertex as the point at the distance of closest approach (DCA) between the trajectories of the daughter particles.

As a first step, the candidate daughter tracks are propagated to their point of closest approach (PCA) taking into account the presence of material budget, which is parametrized in a table stored on the Condition and Calibration Data Base (CCDB) and is accessible at analysis time. When passing through material, charged particles lose a small amount of energy due to ionisation and consequently their momentum decreases and their track curvature changes as they propagate through the detector. The DCAFitter algorithm implements this phenomenon to provide the most realistic description possible. Then, user-defined cuts on the maximum acceptable distance of the tracks at their PCA in the longitudinal and transverse directions are used to either accept or discard the pair of tracks and the corresponding vertex.

The decay point is first reconstructed via DCA minimization of the daughter tracks at their PCA, which means that the vertex is computed as a simple average of the daughter track positions propagated to their point of closest approach. At the next step, the decay vertex is recomputed as a weighted average taking into account the uncertainties of the daughter tracks.

The mother particle is reconstructed by propagating the daughter tracks to the decay vertex: their momenta are transported to this point and the mother three-momentum vector is computed as their sum, while its invariant mass is calculated using the updated momenta of the daughters and the masses values corresponding to their PID hypotheses. For the analysis presented in this thesis, the DCAFitter is used to reconstruct the Ξ_c^0, Ξ, and Λ decay vertices.

5.1.2 Machine Learning

Machine learning (ML) techniques are widely used in high-energy particle physics to address classification and regression problems. The ML algorithm used in the analysis presented in this thesis is a supervised binary classifier.

In the training process, a dataset is fed as input to the ML model so that it can learn how to assign each element to a class depending on its feature values. Given a training dataset D consisting of n instances of candidates, with each of them corresponding to a set of m features $\mathrm{x_i} = \{x_1, x_2, ..., x_m\}$, a machine learning classifier is referred to as "supervised" when the true value y_i identifying the class each $\mathrm{x_i}$ belongs to is known. For a binary classifier $y_i \in \{0, 1\}$.

Assuming there exists a function $f(\mathrm{x}) = y$ mapping x to y, the training goal is to find an approximation $\hat{f}(\mathrm{x})$ of this function given D. This is achieved by solving a minimization problem where a specific function is minimised, the so-called *regularized objective*. Such function is given by the sum of a loss function $L(y, \hat{f}(x))$ and a regularisation term Ω. The function L quantifies the accuracy of the model prediction for the object x while Ω penalises the complexity of the model. This second term is needed to prevent the model from overfitting, i.e. fitting the training data more than wanted. This would result in learning not only general data structures typical of given classes, but also not representative patterns present only in the training data. As a more complex model is more likely to overfit, a regularisation term is needed.

After the training process, the algorithm undergoes the test step, which allows to evaluate how well the model can generalize once it is applied to a set of (labelled) candidates that were not utilized during the training. This gives an estimate of how well the model will perform when it will be used for the classification task in the analysis. Additionally, the testing also allows to check whether the model is overfitting. In this case, a large discrepancy of the model performance on the training set with respect to the test set would be observed, with a worse result on the latter. If this situation occurs, the model complexity needs to be reduced. On the other hand, if the classifier has learnt the structures of the data and can properly generalize, the model performance on the two sets should be similar.

In this analysis, the *hipe4ML* library [49] is used. Its name stands for "Heavy-Ion Physics Environment for Machine Learning" and it provides helper functions for the Python ML toolkit *scikit-learn*.

Boosted Decision Tree

The supervised machine learning technique applied in the analysis presented in this thesis is based on the Decision Tree classifier.

Each data point to be classified is represented by a set of features whose distribution depends on the class to which the candidate belongs. During the training process, the decision tree "learns" these distributions and the correlations in the feature space such that it can implement a series of consecutive rectangular cuts with a hierarchical structure that resembles, precisely, a tree. A schematic representation of a decision tree is shown in Fig. 5.1.

At each splitting step, the variable to be used in the selection and the cut value itself are chosen and optimized such that the best separation between the classes is achieved. Each decision node produces two branches, one with the candidates failing the criterion and one with the candidates passing it. This procedure is repeated iteratively until a stopping condition is reached. The outcome of the terminal decision nodes corresponds to the tree leaves. The leaves are either classified as signal if the signal candidates are dominating the branch or as background if the opposite is the case. The presence of a stopping condition is needed to prevent the model from overfitting.

Single decision trees are well established, but they are also known to be unstable. However, a robust model can be built by combining a large number of trees to form an ensemble. This is exactly the strategy of the machine learning technique used in this analysis, namely the Boosted Decision Tree (BDT) algorithm. It consists in building sequentially a series of single decision trees, referred to as "weak learners", with each new tree giving more importance to the observations in the dataset that were misclassified by the previous models. This is achieved thanks to the usage of weights. To control the structure of the model and its complexity, a set of hyperparameters can be initialized by the analyser and then tuned. As a result, the BDT will be a stronger learner with a lower bias compared to each single tree of the algorithm. A schematic representation of the BDT model is shown in Fig. 5.2. The output of the model (BDT score) is a value ranging from 0 to 1 that represents the

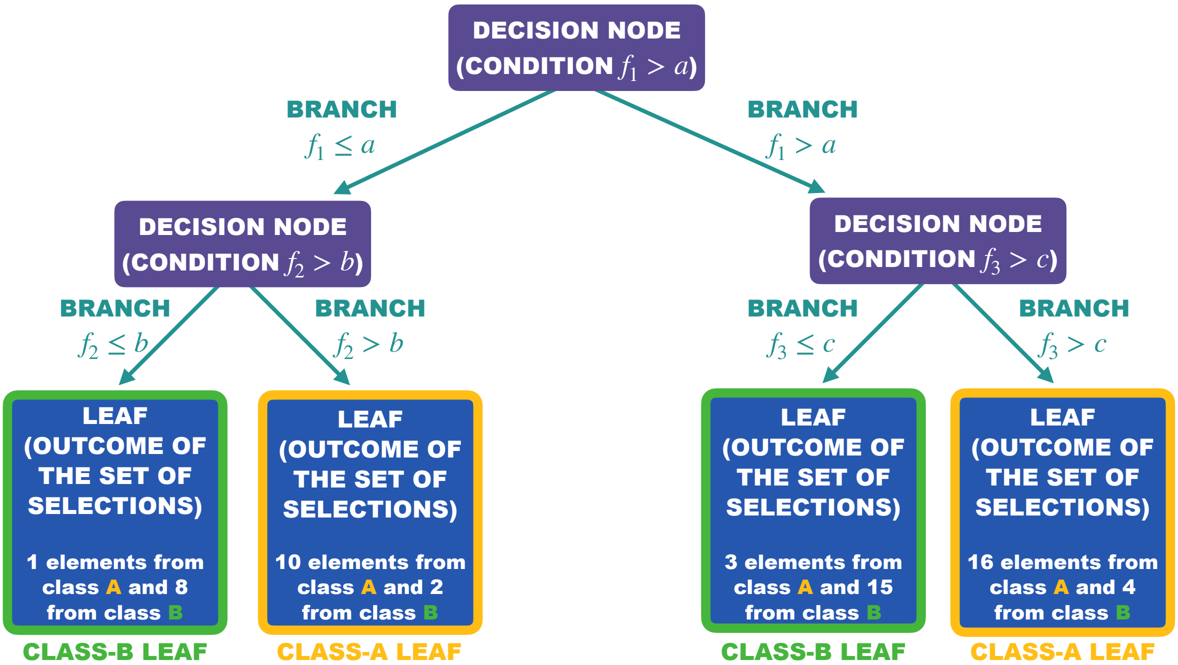


Figure 5.1: Schematic representation of a decision tree.

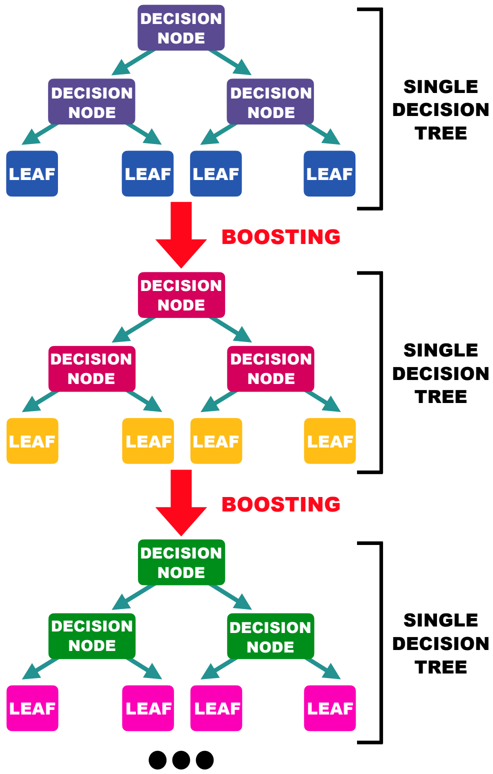


Figure 5.2: Schematic representation of the BDT algorithm.

probability that the the given candidate belongs to a specific class. For binary classification problems, it follows that 1–BDT score represents the probability that the candidate belongs to the remaining class. In this work, the BDT classifies the reconstructed candidates as either signal or background. After having processed the full dataset with the trained BDT model, the analyser selects a value of BDT score where to cut in order to discard background candidates at analysis level.

XGBoost

The library XGBoost [50][51], that implements a gradient boosting algorithm, is used in the analysis presented in this thesis. The tree ensemble model uses K additive functions to predict the output for a given dataset containing n examples of m features $\mathcal{D} = \{(\mathrm{x_i}, y_i)\}$ $(|\mathcal{D}| = n,\ \mathrm{x_i} \in \mathbb{R}^m, y_i \in \mathbb{R})$

$$\hat{y}_i = \Phi(\mathrm{x_i}) = \sum_{k=1}^{K} f_k(\mathrm{x_i}), \quad f_k \in \mathcal{F} \tag{5.1}$$

where

$$\mathcal{F} = \{f(\mathrm{x}) = w_{q(\mathrm{x})}\}(q\ :\ \mathbb{R}^m \to T,\ w \in \mathbb{R}^T) \tag{5.2}$$

is the space of regression trees. Here q represents the structure of each tree, the decision rule mapping an instance of the dataset to the corresponding leaf index, T is the number of leaves in the tree and f_k corresponds to an independent tree structure q and leaf weights w (with w_i being the weight on the i-th leaf). A continuous score is associated to each tree leaf and the final prediction for a certain instance $\mathrm{x_i}$ is given by summing up the score in the corresponding leaf of each tree of the ensemble. The regularized objective of this model is

$$\mathcal{L}(\Phi) = \sum_i l(y_i, \hat{y}_i) + \sum_k \Omega(f_k) \tag{5.3}$$

where

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda||w||^2. \tag{5.4}$$

Here l is a differentiable convex loss function that measures the difference between the model prediction $\hat{y}_i$ and the true target value y_i, while Ω is the regularization term.

The model is trained in an additive way. At the t-th iteration, we add the function f_t that most improves the prediction of the $(t-1)$-th iteration (see Eq. 5.1), so the objective to be minimised becomes

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(\mathrm{x_i})) + \Omega(f_t) \tag{5.5}$$

where $\hat{y}_i^{(t)}$ is prediction of the i-th instance. Then, a second-order approximation is used to simply the objective optimization

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n}[l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(\mathrm{x_i}) + \frac{1}{2}h_i f_t^2(\mathrm{x_i})] + \Omega(f_t) \tag{5.6}$$

where $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$.
Removing the terms that do not have any dependence on f_t, we are left with the objective

$$\tilde{\mathcal{L}}^{(t)} \simeq \sum_{i=1}^{n}[g_i f_t(\mathrm{x_i}) + \frac{1}{2}h_i f_t^2(\mathrm{x_i})] + \Omega(f_t). \tag{5.7}$$

We now focus on the determination of the weights of leaf j, so we define $\mathcal{I}_j = \{i|q(\mathrm{x}_i) = j\}$. This means that $\mathcal{I}_j$ is the set of instances that end up in leaf j according to the decision rule q. Expanding the regularization term in Eq. 5.7 with its definition in Eq. 5.4, we obtain

$$\tilde{\mathcal{L}}^{(t)} \simeq \sum_{i=1}^{n}[g_i f_t(\mathrm{x_i}) + \frac{1}{2}h_i f_t^2(\mathrm{x_i})] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2 = \sum_{j=1}^{T}[(\sum_{i\in\mathcal{I}_j} g_i)w_j + \frac{1}{2}(\sum_{i\in\mathcal{I}_j} h_i + \lambda)w_j^2] + \gamma T. \tag{5.8}$$

Here, the definition of f_t reported in Eq. 5.2 has been used.
For a fixed tree structure q the optimal weight w_j^* of leaf j is computed by imposing $\frac{\partial\tilde{\mathcal{L}}^{(t)}}{\partial w_j} = 0$. This gives

$$w_j^* = -\frac{\sum_{i\in\mathcal{I}_j} g_i}{\sum_{i\in\mathcal{I}_j} h_i + \lambda}. \tag{5.9}$$

Using the definition of w_j^* reported in Eq. 5.9 inside Eq. 5.8 we obtain the scoring function used to measure the quality of a tree structure q

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in\mathcal{I}_j} g_i)^2}{\sum_{i\in\mathcal{I}_j} h_i + \lambda} + \gamma T. \tag{5.10}$$

The algorithm starts from a single leaf and then iteratively adds branches, evaluating the loss reduction after each split. The formula used for this purpose can be derived from Eq. 5.10: if $\mathcal{I}_L$ and $\mathcal{I}_R$ are the instances sets in the left and right node coming from the splitting of node $\mathcal{I}$, we have $\mathcal{I} = \mathcal{I}_L \cup \mathcal{I}_R$ and the loss reduction after the split will be

$$\mathcal{L}_{split} = \frac{1}{2}\left[\frac{(\sum_{i\in\mathcal{I}_L} g_i)^2}{\sum_{i\in\mathcal{I}_L} h_i + \lambda} + \frac{(\sum_{i\in\mathcal{I}_R} g_i)^2}{\sum_{i\in\mathcal{I}_R} h_i + \lambda} - \frac{(\sum_{i\in\mathcal{I}} g_i)^2}{\sum_{i\in\mathcal{I}} h_i + \lambda}\right] - \gamma. \tag{5.11}$$

Scanning all the possible splitting points, especially for continuous features, would be extremely resource-demanding. As an alternative, for each feature, the algorithm proposes a series of candidate splitting points according to percentiles of the feature distribution.

As additional measures to avoid overfitting, two techniques are applied: shrinkage and feature subsampling. The first one requires to scale the newly added weights by a factor η after each step of tree boosting in order to reduce the influence of the individual weak learner and leave space for future trees to improve the performance of the ensemble. The second technique consists in sampling, for each new weak learner, a given fraction ρ of the input features to train the single decision tree.

Optuna

The XGBoost algorithm includes a set of parameters that allow to control the BDT structure. These are the so-called hyperparameters:

- the **maximum depth**, which represents the maximum number of nodes along the longest path from the initial node to the last leaf of the single tree
- the **shrinkage step size** η described above
- the **number of estimators**, that controls the number of trees in the ensemble
- the **minimum child weight**, which controls the stopping condition to end the splitting of a tree. If the tree splitting step results in a leaf node with the sum of instance weights lower than this value, then the building process will give up further partitioning
- the **subsample ratio**, representing the fraction of the training dataset randomly sampled at each boosting iteration to be used to build the tree
- the **column subsample ratio**, representing the fraction of training features randomly sampled prior to growing each tree and used to train the single weak learner (see the ρ parameter described above)

In this analysis, the hyperparameters are optimised with a Bayesian approach using the dedicated framework *Optuna* [52]. Bayesian optimization is an iterative method that requires to evaluate the model performance for different hyperparameter values while keeping track of past evaluation results. Such results are used to build a probabilistic model (called "surrogate") that maps the hyperparameters to a probability of a score on the objective function $P(\text{score} \mid \text{hyperparameters})$. The surrogate is updated after each evaluation of the objective function with a new set of hyperparameters. The Bayesian approach consists in finding the best set of hyperparameters by selecting the values close to the hyperparameter configurations that have previously given good results, and are therefore the most promising ones according to the surrogate.

If the hyperparameters search would be carried out until the model performs optimally on the whole training data, we may end up with a non-generalisable model (overfitting). To avoid this eventuality and to ensure that each set of hyperparameters is evaluated robustly, the k-fold cross validation (CV) method is used. Given a fixed set of hyperparameters sampled on the basis of the surrogate, this technique requires to randomly divide the training dataset into k non overlapping subsets (folds), use $k-1$ of them for model training and the remaining one for validation. This procedure is repeated k times, each time permuting the subsets. The final performance measure corresponding to the set of hyperparameters used is a score calculated as the average of the scores computed at each validation step in the loop. Optuna collects the averaged score and uses it to decide what hyperparameters to try next. This procedure is repeated for a number of times `ntrials` that is configured by the analyser.

5.2 Data sample and Monte Carlo simulation

The analysis presented in this thesis uses a sample of minimum bias data of proton-proton collisions at a centre-of-mass energy of $\sqrt{s} = 13.6$ TeV recorded by the ALICE experiment at the LHC in the year 2022. This dataset contains about 60 billion collisions collected in continuous readout mode. The overall data sample corresponds to an integrated luminosity of $\sim 0.8\ \mathrm{pb}^{-1}$, almost 30 times larger than the pp minimum bias dataset collected by ALICE during the whole Run 2 at $\sqrt{s} = 13$ TeV and used to perform the previous measurement of the prompt Ξ_c^0 cross section.

Multiple charm- and beauty-enriched Monte Carlo (MC) simulations anchored to the corresponding data taking conditions are used to train the Boosted Decision Tree model and to perform the acceptance-times-efficiency correction. These simulations are produced using the gap-triggered production scheme, with a gap size of 5. More details about this strategy are given in Section 4.4. The particle generator chosen is Pythia 8 Colour Reconnection Mode 2, which enables the formation of the junction topology at hadronisation time and therefore allows to increase the baryon production and save computing resources. This Pythia tune predicts a p_{T} spectrum which is known to underestimate the existing Ξ_c^0 measurement, but its shape gives a quite good description of the cross section dependence on the transverse momentum. In total ~ 2.4 billion events are generated and almost $\sim 9 \cdot 10^6$ prompt $\Xi_c^0 + \overline{\Xi}_c^0$ are produced in the midrapidity interval $|y| < 0.8$ within the p_{T} region of interest $1 < p_{\mathrm{T}} < 12$ GeV/c and decay to the channel investigated in this analysis. The simulations are constrained by real data in a data-driven mode as for what concerns the impact parameter resolution (or distance of closest approach, DCA) and the particle identification (PID) performance. In particular, the difference in DCA resolution between data and MC is parametrised, the parametrization is stored on the CCDB and then directly accessed when the analysis workflow is executed on MC files. Parametrisations of the Bethe-Bloch curves obtained using the ALEPH parametrisation and the neural network approach are employed to describe the energy loss in the TPC (see Section 3.3.3). The

$\mathrm{d}E/\mathrm{d}x$ distributions are obtained by sampling from a Gaussian distribution with mean equal to the value of the Bethe-Bloch parametrisation for a given track momentum and for a given particle species hypothesis, and width equal to the corresponding expected resolution. TOF PID post-calibrations are used to obtain $\mathrm{n}\sigma$ distributions matching the performance observed in real data. A smearing of the MC p_{T} resolution is done at analysis level using a constant factor ξ with no p_{T} dependence. The value chosen for ξ is validated by the analyser by checking against real data the MC signal peak widths. Currently, the constraints on the DCA and p_{T} resolution are applied only to primary tracks and to tracks directly originating from heavy-flavour decays. These data-driven constraints are needed because the comparison between real data and the simulation showed that the MC does not properly reproduce such variables, but they are both used in the preselections and processed by the BDT. Derived data are produced both for real data and MC starting from the original AO2D files and the analysis is then performed on the output of this processing step.

5.3 Event selection

The collisions analysed in this thesis, now referred to as ”events”, must satisfy a set of selections:

- TVX trigger: the time difference between the signals coming from the FT0A and FT0C detectors is acceptable. The coincidence of these two timing signals is the minimum bias condition. This requirement allows to remove background signals induced by particles coming from directions other than the signal region, i.e. the nominal collision point. Examples of such background are events associated to cosmic rays and beam-gas interactions.

- The coordinate along the beam axis (z) of the reconstructed primary vertex is within 10 cm from the nominal collision point, i.e. -10 cm < z_{PV} < 10 cm. This allows to have a uniform tracking acceptance for all the selected events.

- Events close to the Time Frame (TF) borders and events close to the ITS Read-Out Frame (ITSROF) are discarded. This allows to remove collisions whose information is spread over two neighbouring minimal processing units and consequently can not be properly reconstructed.

The event selection requested for data and MC is the same except for the TF and ITSROF borders selection: the first is not used when processing the simulation because at production time events are not generated close to the TF edges, while the effect of the latter is not properly reproduced in the MC, so the selection itself is not applied.

5.4 Preselections

Given that the analyses of rare charmed-baryons are affected by the issue of a poor signal-over-background ratio, the usage of rectangular cuts is needed to partly reduce the background even before the deployment of machine learning algorithm.

In Table 5.1, the preselections applied on Ξ, Λ, and their daughters are listed. The preselections applied on the bachelor pion coming directly from the charmed-baryon decay are reported in Table 5.2. As for the candidate $\pi \leftarrow \Xi_c^0$, only tracks built as matched ITS-TPC tracks have been used. The Ξ_c^0 cuts are listed in Table 5.3.

The four tracks used to reconstruct the Ξ_c^0 candidate, i.e. the proton and the three pion tracks, have to fulfil a set of requirements that ensure the quality of the tracking within the TPC. These selections are listed in Table 5.4. As for the track of the pion coming directly from the Ξ_c^0 decay, there are further requirements concerning the ITS. These are listed in Table 5.5.

Variable	**Criterion**
R_Ξ	$\geq$ 0.6 cm
R_Λ	$\geq$ 1.2 cm
$\cos\theta_\Lambda$	$\geq$ 0.97
$\cos\theta_\Xi$	$\geq$ 0.97
|DCA| of Ξ daughters tracks	$\leq$ 1 cm
|DCA| of Λ daughters tracks	$\leq$ 1 cm
$\|\mathrm{DCA}\|_{xy}$ Λ to PV	$\geq$ 0.02 cm
$\|\mathrm{DCA}\|_{xy}$ Ξ to PV	< 2 cm
$\|\mathrm{DCA}\|_{z}$ Ξ to PV	$\leq$ 10 cm
$\|\mathrm{DCA}\|_{xy}$ $\pi \leftarrow \Xi$ to PV	$\geq$ 0.04 cm
$\|\mathrm{DCA}\|_{xy}$ Λ daughters tracks to PV	$\geq$ 0.06 cm
$\|m_\Xi - m_{\Xi_{PDG}}\|$	< 0.01 GeV/c
$\|m_\Lambda - m_{\Lambda_{PDG}}\|$	< 0.008 GeV/c
$\|m_{K^0_S} - m_{K^0_{S\,PDG}}\|$	> 0.01 GeV/c
$\|\eta\|$ of Ξ and Λ daughter tracks	$\leq$ 1
$p_{\mathrm{T}\,\pi\leftarrow\Xi}$	$\geq$ 0.15 GeV/c
$p_{\mathrm{T}\,\Xi}$	> 1.0 GeV/c

Table 5.1: Preselection criteria for Ξ and Λ. These cuts have been validated by the ALICE light-flavour physics analysers. More details about variables definitions are provided in Appendix A.

Variable	**Criterion**
$\|\eta\|$	$\leq$ 0.8
p_T	> 0.5 GeV/c
$\|\mathrm{DCA}\|_{xy}$ to PV	$\leq$ 10 cm
$\|\mathrm{DCA}\|_{z}$ to PV	$\leq$ 10 cm

Table 5.2: Preselection criteria for $\pi \leftarrow \Xi_c^0$.

Variable	**Criterion**
$\|\eta\|$	≤ 0.8
p_{T}	> 1 GeV/c
p_{T}	< 12 GeV/c
|DCA| of Ξ_c^0 daughters tracks	≤ 2 cm

Table 5.3: Preselection criteria for Ξ_c^0.

Variable	**Criterion**
Number of found clusters	≥ 70
Number of crossed rows	≥ 70
Ratio crossed rows over findable clusters	≥ 0.8
χ^2 / number of clusters	≤ 4

Table 5.4: TPC tracks preselections.

Variable	**Criterion**
Number of found clusters	≥ 4
Number of found clusters inner barrel	≥ 1
χ^2 / number of clusters	≤ 36

Table 5.5: ITS tracks preselections.

The particle identification for the final state particles, i.e. the three pions and the proton, is performed by combining the TPC and TOF information. At the first step, the candidate selector task running on hyperloop discards a given track only if both the TPC and TOF signals are more than 4σ away from the expected signal of the desired particle species. At a later stage, a tighter criterion consisting of a 3σ cut on the TPC signal is used to further skim the TTree produced by the tree creator task.
The preselections listed above result in the acceptance-times-efficiency values shown in Fig. 5.3. A drop is observed at high p_{T} in the preselection efficiency. It will be discussed in Section 5.7, when the total $A \cdot \epsilon$ is presented.

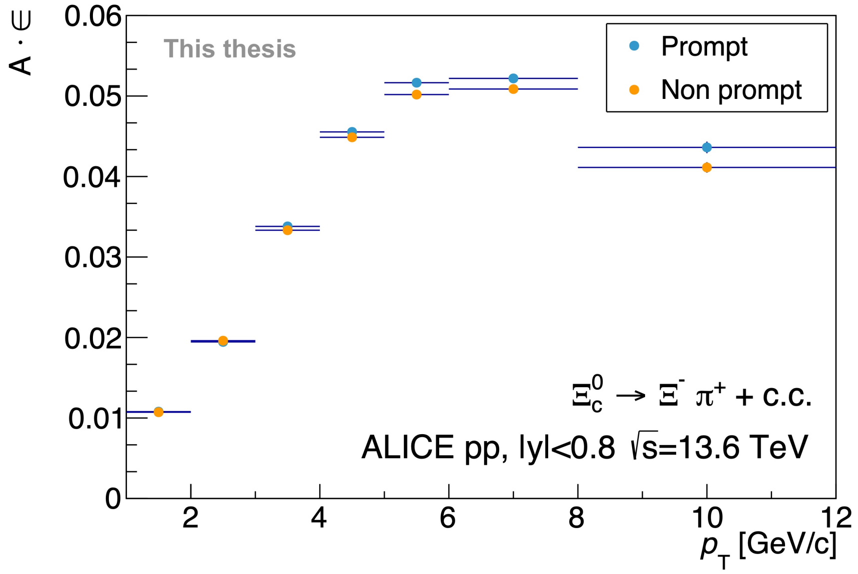


Figure 5.3: Preselection acceptance-times-efficiency for prompt and non prompt Ξ_c^0. The values are averaged for Ξ_c^0 and $\overline{\Xi}_c^0$.

5.5 Boosted Decision Tree

5.5.1 BDT training

In this analysis, a BDT algorithm is used to perform a binary classification of the Ξ_c^0 candidates. The classes considered are signal (S) and background (B). The BDT score computed by the model represents the probability that a given candidate belongs to the signal class.

To train the model, prompt signal candidates are taken from the MC simulations and background candidates from real data, more precisely from the sidebands corresponding to the invariant-mass regions $2.00 < m_{\mathrm{INV}} < 2.15$ GeV/c^2 and $2.77 < m_{\mathrm{INV}} < 2.92$ GeV/c^2. This ensures that no true Ξ_c^0 ($m = (2.47044 \pm 0.00028)$ GeV/c^2, from Particle Data Group, PDG) or Ω_c^0 ($m = (2.6952 \pm 0.0017)$ GeV/c^2, from PDG) - which can also decay to the same channel as the Ξ_c^0 - is misidentified as background during the training. The criterion used to select such invariant-mass regions assumes that the signal is Gaussian-distributed and that it is safe to look for background candidates at least 2.5σ away from the signal invariant mass. The value of σ, the width of the signal peak, is extracted from MC and is found to be $<$ 0.03 GeV/c^2 (see Section 5.6). The signal invariant-mass distribution is assumed to be Gaussian even if the measured variable is the track curvature, which is proportional to $1/p_{\mathrm{T}}$. Therefore, to be precise, it is the inverse of the transverse momentum to be Gaussian-distributed. However, the validity of this approximation is confirmed by the χ^2 values of the Gaussian fits performed on MC signal $m_{\rm inv}$ distributions (see Section I). The width values of such distributions reflect the experimental resolution rather than the intrinsic width that depends on the particle lifetime. The reason being, weakly-decaying particles are characterised by long lifetimes τ because of the large gauge-boson mass. As a result, their intrinsic width $\Gamma = \frac{\hbar}{\tau}$ is extremely small, by far smaller than the experimental resolution. For instance, given the Ξ_c^0 lifetime $\tau = (1.504 \pm 0.028) \cdot 10^{-13}$ s [4], its intrinsic width corresponds to fractions of eV.

In order to exploit the dependence of the topological and kinematic training variables on the charmed-baryon $p_{\rm T}$, multiple BDT models are implemented and trained, one for each $p_{\rm T}$ bin used to extract the raw yield and measure the cross section. The only exception is represented by the last bin at high $p_{\rm T}$: the lack of signal candidates in the MC simulation in this kinematic region requires to train a single model to cover the transverse momentum interval $5 < p_{\rm T} < 12$ GeV/c.

The number of signal and background candidates used for the BDT training and testing for each $p_{\rm T}$ bin is reported in Table 5.6. With the aim of not biasing the model by providing an unbalanced number of candidates belonging to the different classes, the number of instances from each class is limited to the one of the less represented class, i.e. the number of B candidates is chosen equal to that of the S candidates. Once more the exception is represented by the transverse momentum region where there is a lack of signal, namely the last two bins at high $p_{\rm T}$. In these cases, the number of B candidates is chosen to be double the number of S candidates, so that a reasonable number of instances can be used to perform the training and testing of the models. A fraction of the dataset equal to 70% is used for the BDT training, the remaining 30% for testing.

$p_{\mathrm{T}\,\Xi_c^0}$ **[GeV/c]**	$1 < p_{\rm T} < 2$	$2 < p_{\rm T} < 3$	$3 < p_{\rm T} < 4$	$4 < p_{\rm T} < 5$	$5 < p_{\rm T} < 12$
Signal candidates	30 333	36 415	34 420	23 448	27 176
Background candidates	30 333	36 415	34 420	46 896	54 352

Table 5.6: Number of signal and background candidates used for training and testing.

The normalised distributions of the variables characterizing the Ξ_c^0 candidates belonging to the transverse momentum region $1 < p_{\rm T} < 2$ GeV/c are reported in Fig. 5.4. The background class is shown in light blue and the signal one in yellow. Especially for some of them, such as the DCA of the daughter tracks and the impact parameters, the distributions of the same variable differ a lot depending on the class. The same plots for the remaining $p_{\rm T}$ intervals are reported in Appendix B. The variables names are described in Appendix A. The subset of features selected to train the BDT models is listed in Table 5.7.

The correlations between the features are studied by computing the Pearson correlation coefficient ρ for each variable pair. In order not to introduce distortions in the signal invariant-mass distributions, features correlated with the signal invariant mass $m_{\Xi\pi}$ are not used for the model training. Fig.

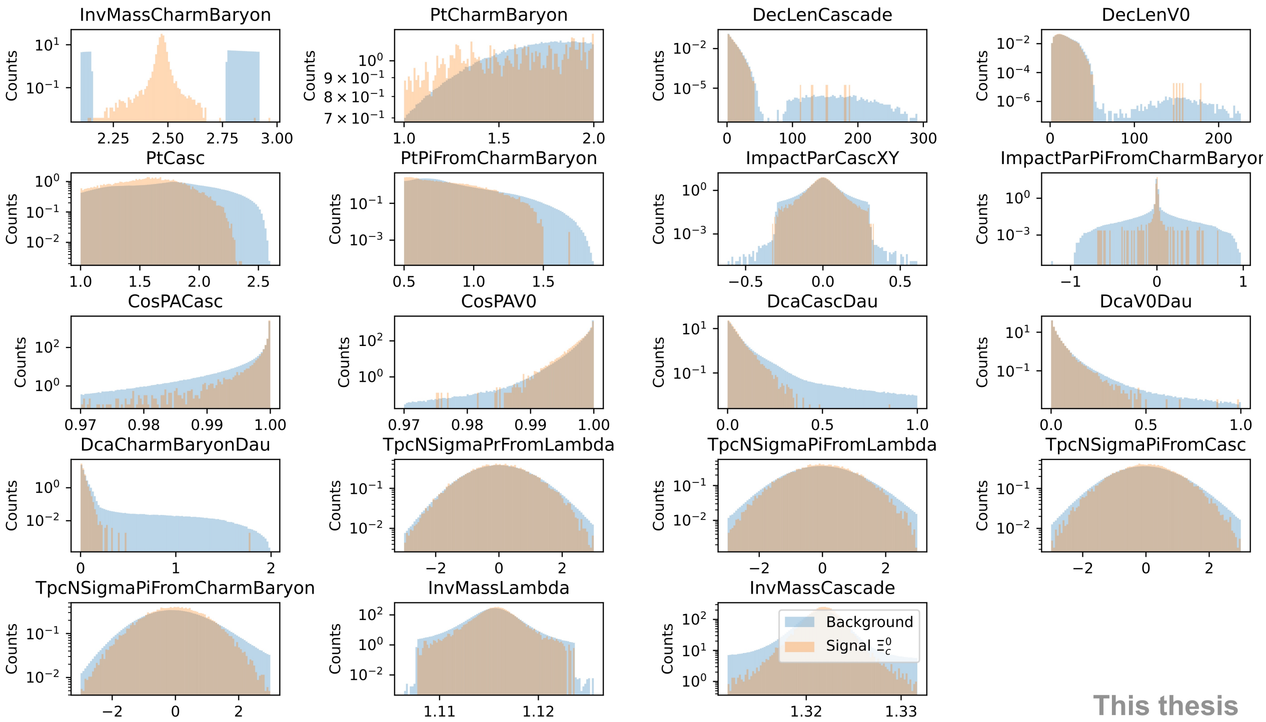


Figure 5.4: Feature normalised distributions for $1 < p_{\mathrm{T}\,\Xi_c^0} < 2$ GeV/c in data and MC.

Training features
$\boldsymbol{\Xi}$
Cosine of pointing angle DCA of daughters tracks Decay length Impact parameter on the transverse plane Invariant mass TPC nσ $\pi \leftarrow \Xi$
$\boldsymbol{\Lambda}$
Cosine of pointing angle DCA of daughters tracks Decay length Invariant mass TPC nσ $\pi \leftarrow \Lambda$ TPC nσ p $\leftarrow \Lambda$
$\boldsymbol{\Xi_c^0}$
DCA of daughters tracks Impact parameter on the transverse plane $\pi \leftarrow \Xi_c^0$ Transverse momentum $\pi \leftarrow \Xi_c^0$ TPC nσ $\pi \leftarrow \Xi_c^0$

Table 5.7: Features used to train the BDT models.

5.5 shows the correlation matrix for the signal class. From this plot is it possible to affirm that the training features considered do not exhibit any strong correlation with $m_{\Xi\pi}$. Again, the same plot corresponding to the remaining p_{T} intervals are reported in Appendix C.

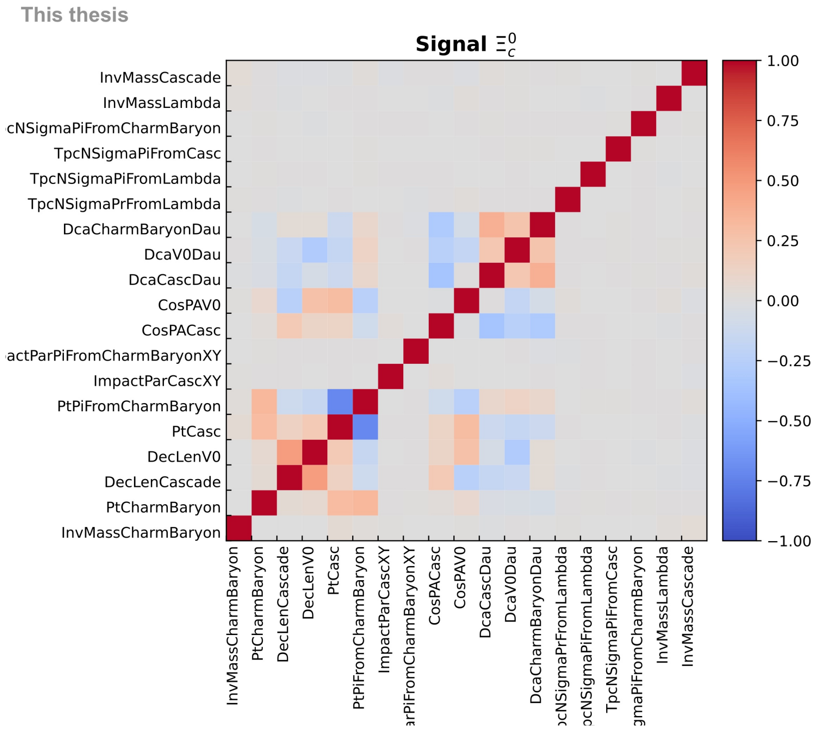


Figure 5.5: Signal class feature correlation matrix for $1 < p_{\mathrm{T}\,\Xi_c^0} < 2$ GeV/c.

The normalised BDT score distributions are reported in Fig. 5.6 for the p_{T} interval $1 < p_{\mathrm{T}\,\Xi_c^0} < 2$ GeV/c. The analogous distributions for the remaining transverse momentum intervals are shown in Appendix D. The results for the signal and background class are shown in red and blue, respectively. The filled distributions are the result of the training, the dots of the testing. These two distributions follow a similar trend, which is interpreted as a sign of good model performance, without over- or under-fitting. This is true for all the BDT models trained in this analysis. In particular, at high p_{T}, the background distribution becomes more peaked for low BDT scores: this is due to the fact that large Lorentz-boosted Ξ_c^0 at high transverse momenta can be more easily separated from the combinatorial background thanks to their more displaced decay topology. For an ideal classifier, the signal distribution would be peaking at a value of BDT score equal to 1 and the background one at 0, which would allow perfect separation.

The hyperparameters selected for each BDT model using the Bayesian optimisation are reported in Table 5.8. A 5-fold cross validation method is used, `ntrials` is set to 25 and the performance metric chosen to evaluate the model is the ROC AUC. More details about this metric are given below, while the hyperparameters definition and the corresponding optimisation method are discussed in Section

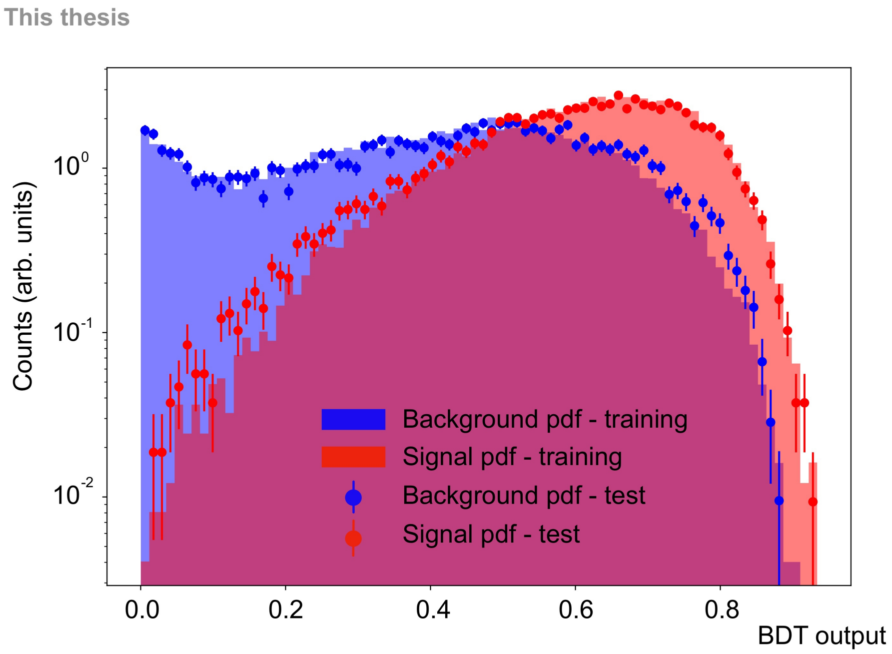


Figure 5.6: BDT score distribution for $1 < p_{\mathrm{T}\,\Xi_c^0} < 2$ GeV/c.

5.1.2.

$p_{\mathrm{T}\,\Xi_c^0}$ [GeV/c]	$1 < p_\mathrm{T} < 2$	$2 < p_\mathrm{T} < 3$	$3 < p_\mathrm{T} < 4$	$4 < p_\mathrm{T} < 5$	$5 < p_\mathrm{T} < 12$
Maximum depth	3	5	2	4	4
Shrinkage step size	0.031	0.017	0.091	0.022	0.026
Number of estimators	1397	559	971	1023	1282
Minimum child weight	2	3	6	6	5
Subsample ratio	0.35	0.38	0.63	0.87	0.95
Column subsample ratio	0.87	0.75	0.51	0.92	0.81

Table 5.8: Hyperparameter values of the BDT models.

The subset of variables selected for the training is reported in Fig. 5.7. This plot is referred to the BDT model corresponding to the transverse momentum region $1 < p_\mathrm{T} < 2$ GeV/c and it ranks the features depending on their discriminating power, with the ones at the top leading to the largest gain in model performance. Depending on how often a variable is used in the building process of the BDT, a feature importance is assigned to it, representing the feature average impact on the model output. Multiple tests for all the p_T bins have been carried out, trying to decrease the number of training features by removing those that had the least impact on the classification task. The results showed that reducing the number of input variables leads to a worsening of the BDT performance and that the set of training features listed in Fig. 5.7 is the one that allows to obtain the highest separation power for all the p_T intervals. The analogous information corresponding to the remaining p_T intervals is reported in Appendix E.

As for the transverse momentum bin 5< p_{T}<12 GeV/c, the p_{T} of the pion coming directly from the Ξ_c^0 decay is kept as a training variable because it has the largest discriminating power (see Fig. E.4), even if it shows a weak correlation with the signal invariant mass $m_{\Xi\pi}$ (see C.4).

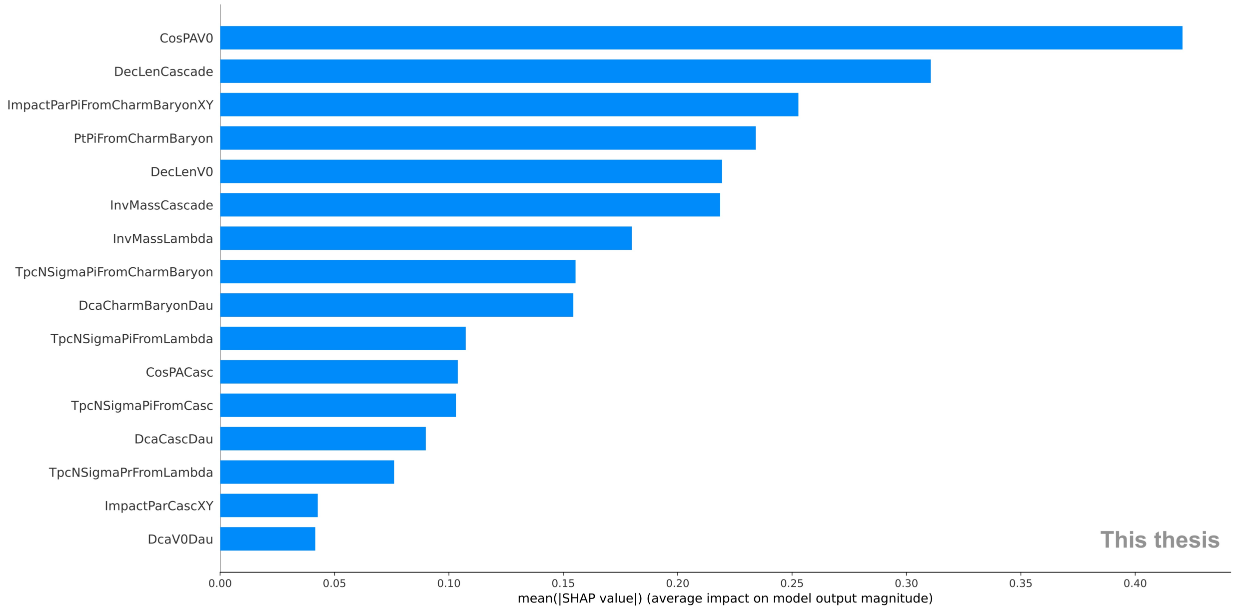


Figure 5.7: Ranking of training features based on their impact on the classification for the BDT model corresponding to $1 < p_{\mathrm{T}\,\Xi_c^0} < 2$ GeV/c.

The BDT performance is estimated on the basis of the Receiver Operating Characteristic curve (ROC). This plot illustrates the performance of a binary classifier by reporting the true positive rate (TPR) as a function of the false positive rate (FPR) for the signal class. The TPR is the fraction of correctly classified instances and corresponds to the model *efficiency*, the FPR is the fraction of wrongly classified instances and corresponds to 1-*purity*. The ROC curve of the BDT model corresponding to the p_{T} interval $1 < p_{\mathrm{T}} < 2$ GeV/c is shown in Fig. 5.8. The same plots for the remaining p_{T} intervals are reported in Appendix F. The blue dashed line corresponds to the model performance on the training set, the solid line on the test set. The Area Under the Curve (AUC) gives a global estimation of the model performance and it can be interpreted as the probability that the classifier assigns to a randomly chosen signal instance a larger BDT score with respect to a randomly chosen background instance. A large discrepancy between the train and test ROC curves or, similarly, between the train and test AUC, would be interpreted as a sign of overfitting. The performance of a random classifier corresponds to the grey line: in this case, the TPR is always equal to the FPR. A model whose ROC curve lies below this line is misinterpreting the data, while the most desirable behaviour corresponds to a point in the top left corner of the plot. However, in reality, an extremely high efficiency comes at the cost of a very low purity and vice versa, so one needs to find a balance between these two extremes. In this analysis, all the ROC curves indicate good model performances, as already considered given the previously mentioned observables. The AUC values increase with the $p_{\mathrm{T}\,\Xi_c^0}$, indicating that the corresponding models perform better. This was already observed when discussing the BDT score distributions.

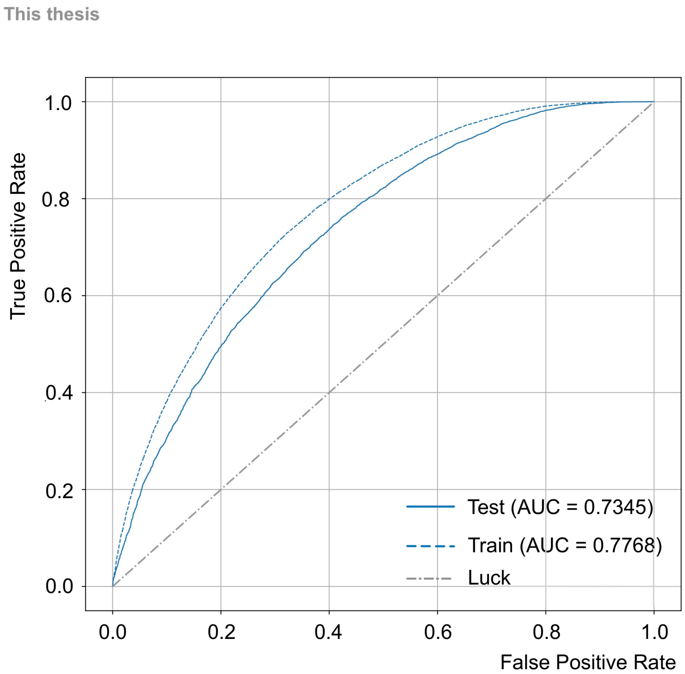


Figure 5.8: ROC curve of the BDT model corresponding to the p_{T} interval $1 < p_{\mathrm{T}\,\Xi_c^0} < 2$ GeV/c.

The last indicator of the model performance taken into account in this thesis is the learning curve. It represents the Root Mean Square Error (RMSE), quantifying the deviation of the model prediction with respect to the truth, as a function of the training set size, i.e. the number of candidates that the model "has seen". Consequently, such curve carries the information about how well the model is *learning* based on its *experience*. To assess the quality of the model performance, the learning curve on the training set is compared to the one on the test set: the desirable behaviour is a convergence of the two curves at increasingly larger instance set size. For small numbers of candidates, the fitting on the training set becomes trivial, but the corresponding trained model can not generalise well. Therefore, the RMSE on the training set is expected to be small, while on the test set large. The comparison of the two learning curves also allows to spot signs of overfitting. In this case, for large values of instance set size, the two curves would diverge, meaning that the error on the training set would become lower and lower while the RMSE on the test set would become increasingly large.
The learning curves of the BDT model corresponding to the p_{T} interval $1 < p_{\mathrm{T}} < 2$ GeV/c are depicted in Fig. 5.9, the analogous plots corresponding to the remaining p_{T} intervals are reported in Appendix G. The red curve corresponds to the train set, the blue one to the test set. Also this indicator

suggests that the model performance is good for all the p_{T} bins, without signs of overfitting.

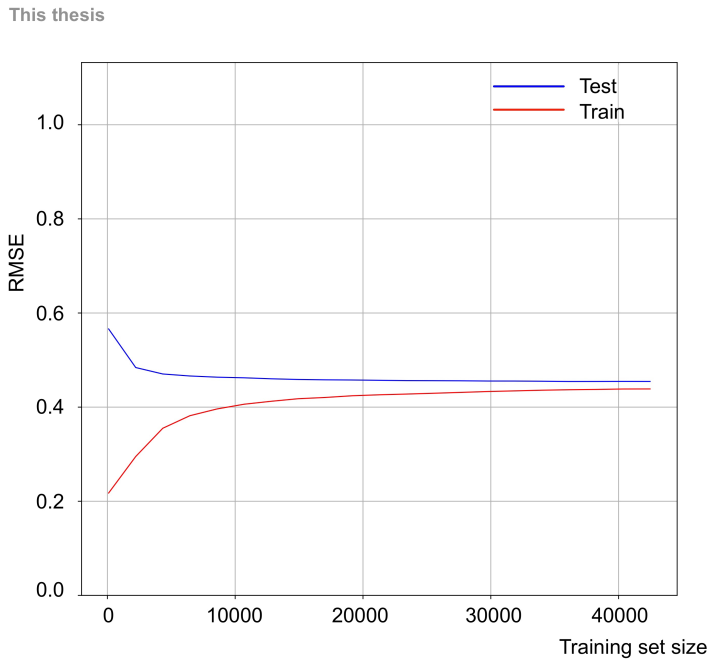


Figure 5.9: Learning curves of the BDT model corresponding to $1 < p_{\mathrm{T}\,\Xi_c^0} < 2$ GeV/c.

5.5.2 Working point determination

After training and testing, the BDT models are used to process the whole analysis dataset. As a result, each Ξ_c^0 candidate is associated to a BDT score. In order to discard candidates that are more likely to belong to the background class, a threshold value of the BDT score, namely the working point, has to be selected. In this analysis, the choice is made on the basis of qualitative considerations on both the BDT score distribution and the BDT efficiency, trying to reject as much background as possible while still preserving the signal. This is equivalent to limiting the choice to BDT score values corresponding to high efficiencies. Requiring tighter BDT score thresholds would correspond to selecting minimum BDT scores closer to 1, and this would results in lower efficiencies and higher purities. The impact of the working point choice on the stability of the result is estimated by assigning a systematic uncertainty, more details are given in Section 5.9.2. After having applied the preselections listed in Section 5.4 and the BDT selections corresponding to the chosen working point, the prompt BDT efficiency is evaluated as the ratio of the correctly classified reconstructed prompt signal candidates fulfilling the preselections over the total number of reconstructed prompt signal candidates fulfilling the preselections and processed by the model. To calculate the BDT efficiency, the same MC simulation used for training and testing is used: this choice is not optimal, as the BDT performance has been optimized exactly on the training set, but the lack of signal candidates in the available production left no other choice. However, the model performance on the training set and on the test set proved to be similar, therefore this choice is not expected to have a large impact on the analysis. This procedure is repeated for every BDT model corresponding to each p_{T} bin used for the signal extraction.

In Fig. 5.10 the prompt BDT efficiency as a function of the BDT score threshold is shown for the transverse momentum interval $1 < p_{\mathrm{T}} < 2$ GeV/c, while the corresponding BDT score distribution is depicted in Fig. 5.6. In this case, the BDT threshold has been set to 0.35, which means that all the Ξ_c^0 candidates with a BDT score lower than this value are discarded. The corresponding BDT efficiency is 92% with a statistical uncertainty of 1% due to the finite size of the MC sample used. The analogous plots for the remaining p_{T} bins are shown in Appendix H. The selected BDT score thresholds and the corresponding prompt BDT efficiencies are reported in Table 5.9.

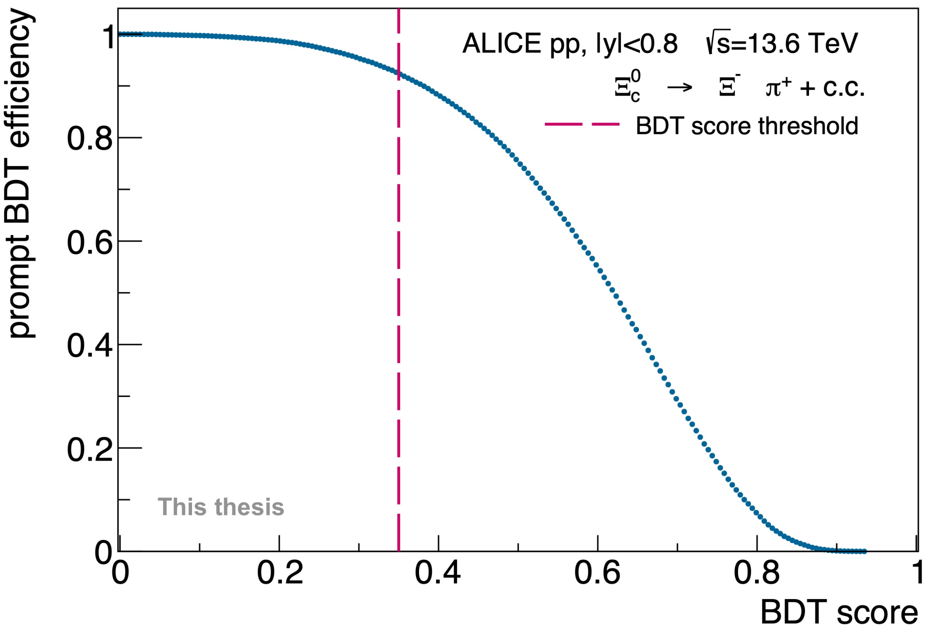


Figure 5.10: Prompt BDT efficiency for the transverse momentum interval $1 < p_{\mathrm{T}\,\Xi_c^0} < 2$ GeV/c as a function of the BDT score threshold. The magenta lines shows the selected BDT score threshold.

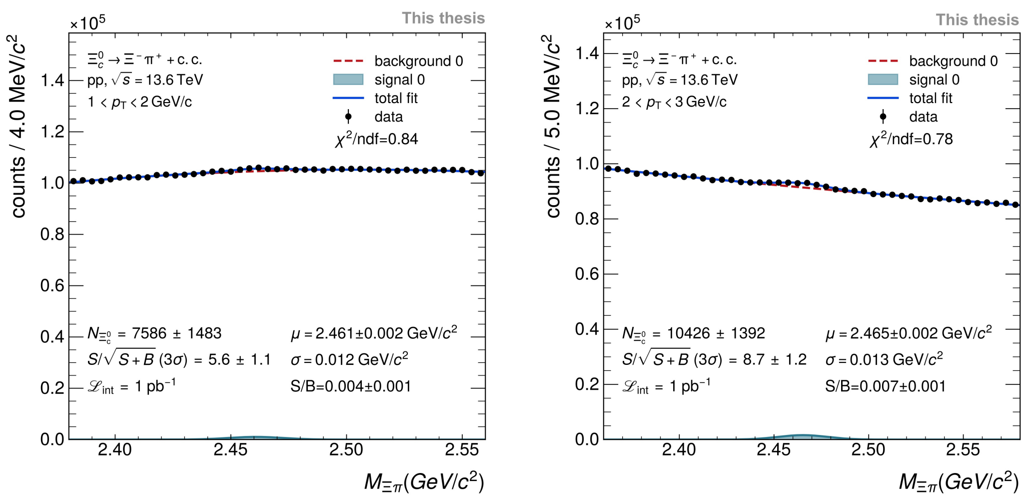
$p_{\mathrm{T}\,\Xi_c^0}$ [GeV/c]	$1 < p_{\mathrm{T}} < 2$	$2 < p_{\mathrm{T}} < 3$	$3 < p_{\mathrm{T}} < 4$	$4 < p_{\mathrm{T}} < 5$	$5 < p_{\mathrm{T}} < 6$	$6 < p_{\mathrm{T}} < 8$	$8 < p_{\mathrm{T}} < 12$
BDT working point	0.35	0.35	0.35	0.25	0.25	0.25	0.25
Prompt BDT efficiency	$(92.3 \pm 0.1)\%$	$(91.4 \pm 0.1)\%$	$(89.3 \pm 0.2)\%$	$(85.8 \pm 0.2)\%$	$(84.3 \pm 0.3)\%$	$(90.5 \pm 0.3)\%$	$(95.3 \pm 0.4)\%$

Table 5.9: Selected BDT working points and corresponding prompt BDT efficiency.

5.6 Signal extraction

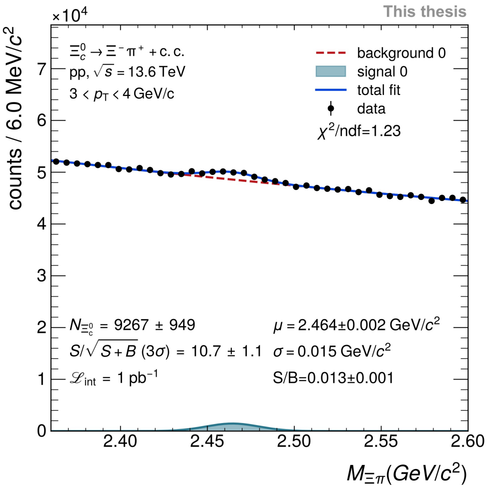
The p_{T}-differential raw yields are extracted by performing maximum-likelihood fits to the invariant-mass spectra corresponding to the various Ξ_c^0 transverse momentum intervals. These spectra report the invariant mass $m_{\Xi\,\pi}$ of the Ξ_c^0 candidates for which the BDT score is higher than the selected working point. In order to facilitate the comparison of the results and to validate the Run 3 analysis, the p_{T} binning chosen to extract the raw yields is the same as the one used in the Ξ_c^0 analysis performed on Run 2 data.

The FlareFly Python library [53] has been used to perform the fits. It allows to manipulate probability density functions (PDFs) to perform multi-fits with several signal and/or background PDFs. The total fit function, which is given by the sum of the signal PDF(s) and background PDF(s), is normalised to 1 and a fit parameter $\texttt{frac}_i$ is assigned to each single PDF. This parameter represents the fraction of the total PDF represented by the i-th specific PDF. In case there is only one signal function and one background function, only one `frac` parameter is needed to give a complete description. In this analysis, the signal is described with a Gaussian function and the background with a polynomial of second order. The raw yield is computed as the product of the number of histogram entries times the signal PDF fit parameter `frac`. The Ξ_c^0 mass corresponds to the mean fit parameter of the Gaussian PDF, the width to its σ. To improve the stability of the signal extraction, the width of the Gaussian function is fixed to the value extracted by performing a Gaussian fit to a MC invariant-mass spectrum of reconstructed signal candidates corresponding to the same Ξ_c^0 p_{T} interval. Fig. 5.11 shows the fits to real data invariant-mass spectra. The dashed red line corresponds to the background PDF, the filled gray distribution represents the signal PDF, and the solid blue line shows the total PDF. In Fig. 5.12 the corresponding residuals are reported. The MC fits can be found in Appendix I.

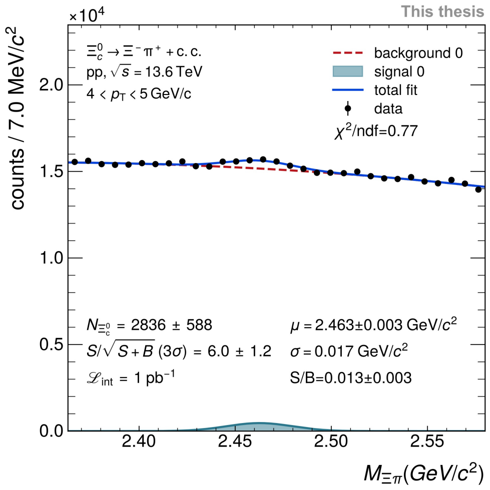


(a) $1 < p_{\mathrm{T}\,\Xi_c^0} < 2$ GeV/c

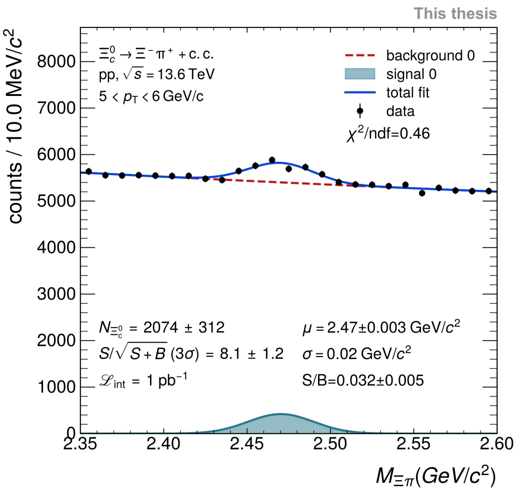
(b) $2 < p_{\mathrm{T}\,\Xi_c^0} < 3$ GeV/c



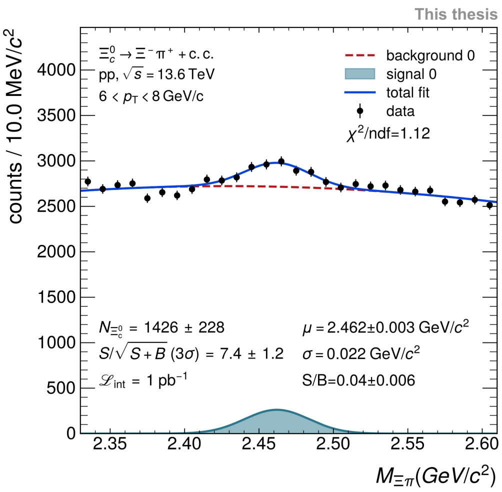
(c) $3 < p_{\mathrm{T}\,\Xi_c^0} < 4$ GeV/c



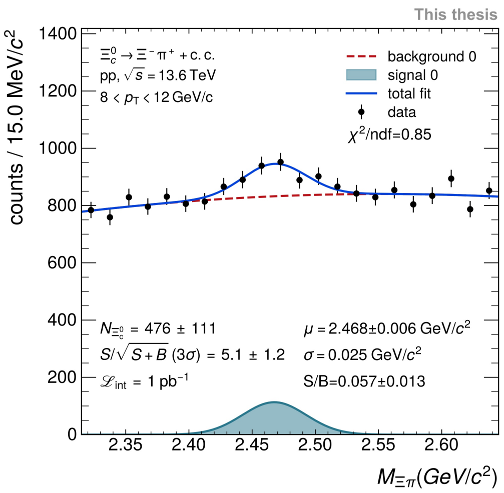
(d) $4 < p_{\mathrm{T}\,\Xi_c^0} < 5$ GeV/c



(e) $5 < p_{\mathrm{T}\,\Xi_c^0} < 6$ GeV/c

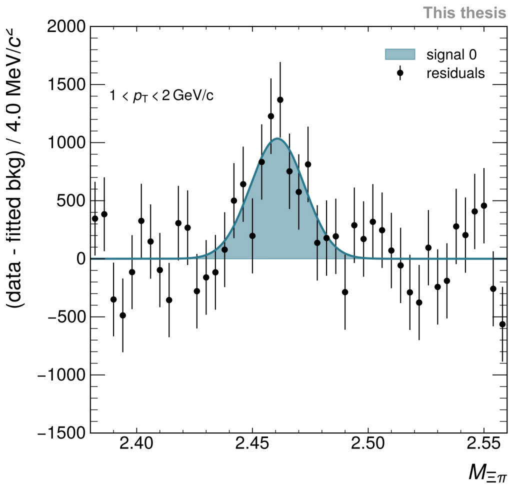


(f) $6 < p_{\mathrm{T}\,\Xi_c^0} < 8$ GeV/c

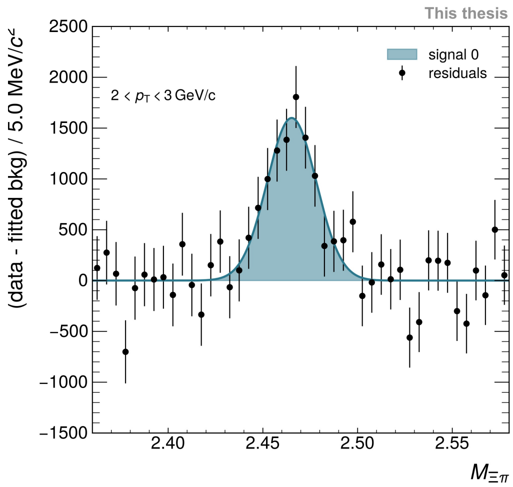


(g) $8 < p_{\mathrm{T}\,\Xi_c^0} < 12$ GeV/c

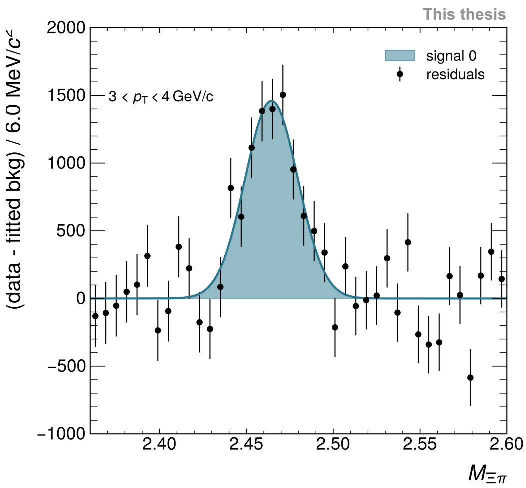
Figure 5.11: Fits to invariant-mass spectra from real data.



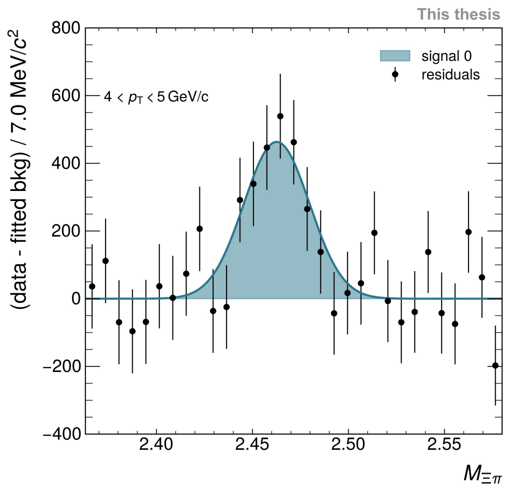
(a) $1 < p_{\mathrm{T}\,\Xi_c^0} < 2$ GeV/c



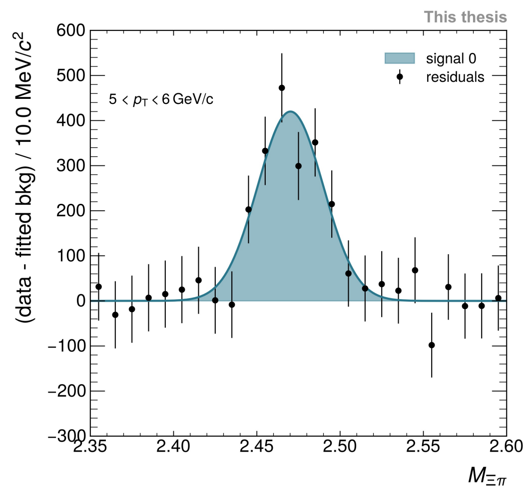
(b) $2 < p_{\mathrm{T}\,\Xi_c^0} < 3$ GeV/c



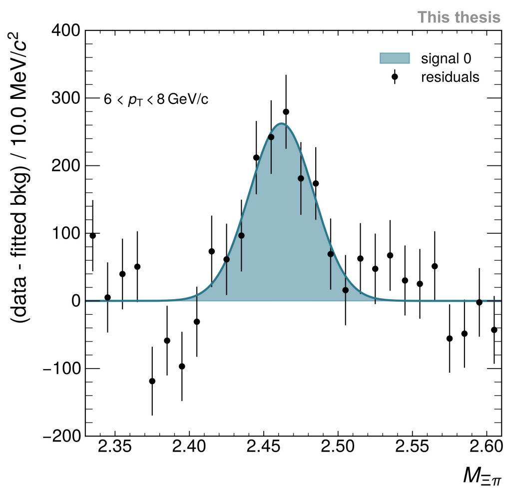
(c) $3 < p_{\mathrm{T}\,\Xi_c^0} < 4$ GeV/c



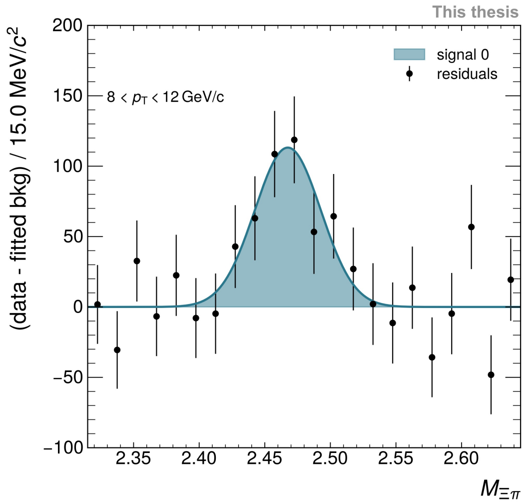
(d) $4 < p_{\mathrm{T}\,\Xi_c^0} < 5$ GeV/c



(e) $5 < p_{\mathrm{T}\,\Xi_c^0} < 6$ GeV/c



(f) $6 < p_{\mathrm{T}\,\Xi_c^0} < 8$ GeV/c



(g) $8 < p_{\mathrm{T}\,\Xi_c^0} < 12$ GeV/c

Figure 5.12: Residuals of fits to invariant-mass spectra from real data.

The results of the fits are studied as a function of the charmed-baryon p_T. The Ξ_c^0 mass is represented by the Gaussian PDF mean parameter. It is underestimated in real data and overestimated in MC (see the MC spectra fits in Appendix I) with respect to the Particle Data Group value $m_\mathrm{PDG} = (2.47044 \pm 0.00028)$ GeV/c. This comparison is reported in Fig. 5.13. A similar behaviour was already observed in Run 2.

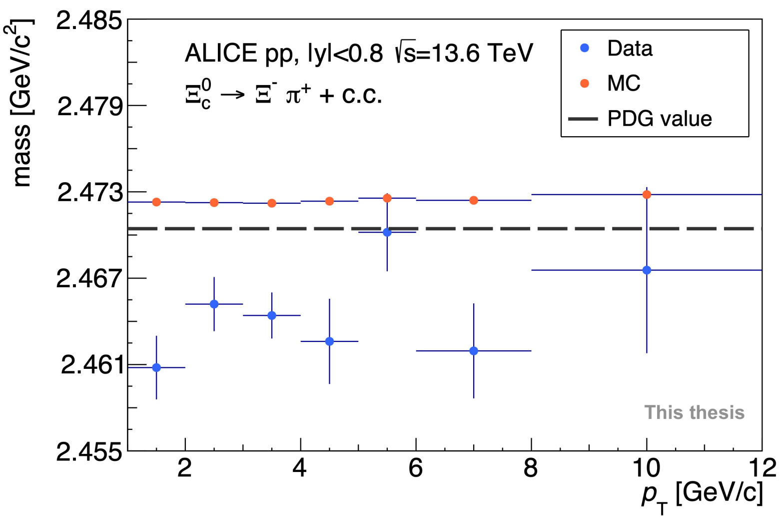


Figure 5.13: Mass of the Ξ_c^0 as a function of the charmed-baryon transverse momentum.

A dedicated study has been performed for the Ξ_c^0 peak width, that corresponds to the σ parameter of the Gaussian PDF. The factor used to smear the MC p_T resolution mentioned in Section 5.2 has been set to 1.5 and validated by comparing the MC signal peak width to the σ extracted by fitting real data without constraining any parameter. This comparison is reported in Fig. 5.14. Given that the width values are compatible within the uncertainties in almost all the p_{T} bins, the smearing factor is considered validated. The real data invariant-mass spectra fits without parameter constraints are reported in Appendix J together with their residuals.

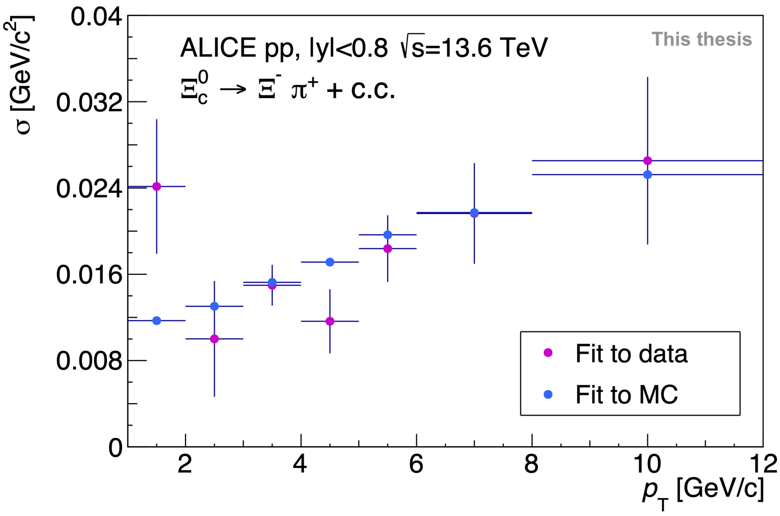


Figure 5.14: Width of the Ξ_c^0 as a function of the charmed-baryon transverse momentum.

Fig. 5.15 shows the raw yield trend, while Fig. 5.16 and 5.17 report, respectively, the significance and the signal-over-background ratio. The first one is defined as $\frac{S}{\sqrt{S+B}}$, the second one as $\frac{S}{B}$, where S and B correspond to the extracted signal and background in a 3σ invariant-mass window around the mean of the Gaussian PDF, respectively. A significance smaller than 3 is interpreted to not significantly describe a signal peak, while higher values suggest a signal peak on top of the background spectrum. In this analysis, the significance larger than 3 in all the p_{T} bins, with the largest value observed at intermediate transverse momentum, for $3 < p_{\mathrm{T}} < 4$ GeV/c. The signal-over-background ratio increases as a function of p_{T} as a result of the faster decrease at increasingly high p_{T} of the combinatorial background with respect to the signal, combined with the higher effectiveness of the selections applied. Downward fluctuations are observed for the $4 < p_{\mathrm{T}} < 5$ GeV/c bin in the raw yield, significance and signal-over-background distributions. This does not seem to depend on the corresponding σ_{MC} value used in the fit, as the $\sigma(p_{\mathrm{T}})$ trend for MC values (see Fig. 5.14) does not show any discontinuity for the corresponding data point. Such results are rather attributed to a statistical fluctuation.

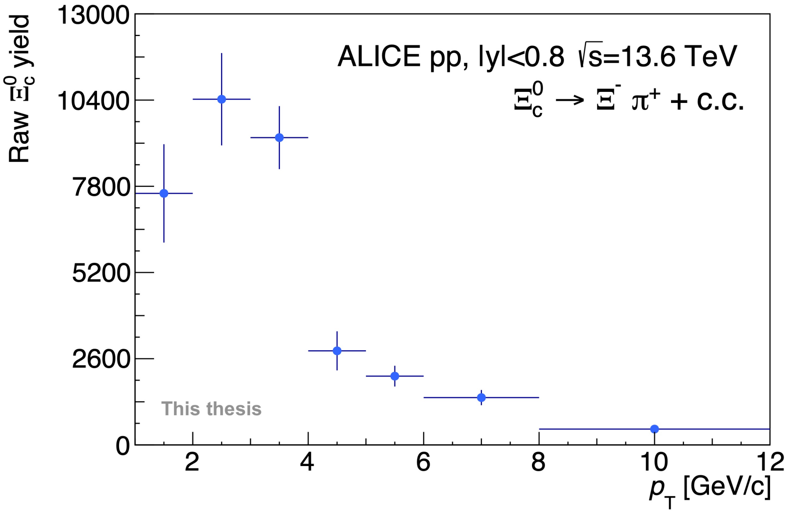


Figure 5.15: Raw Ξ_c^0 yield as a function of the charmed-baryon transverse momentum.

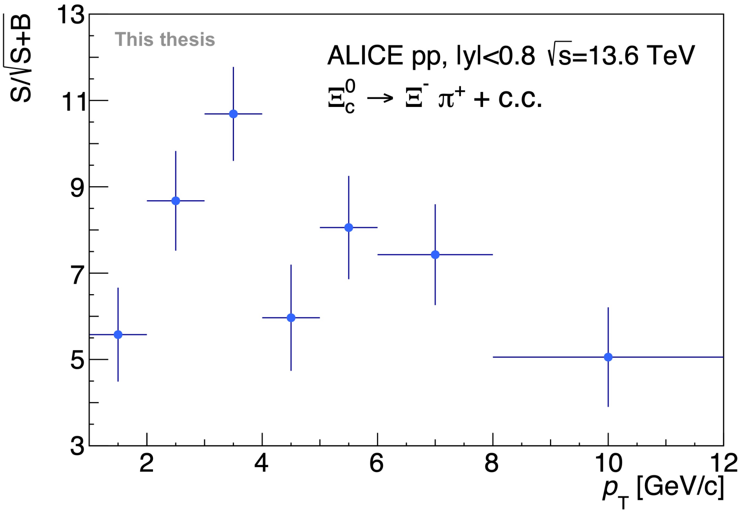


Figure 5.16: Significance of the Ξ_c^0 as a function of the charmed-baryon transverse momentum.

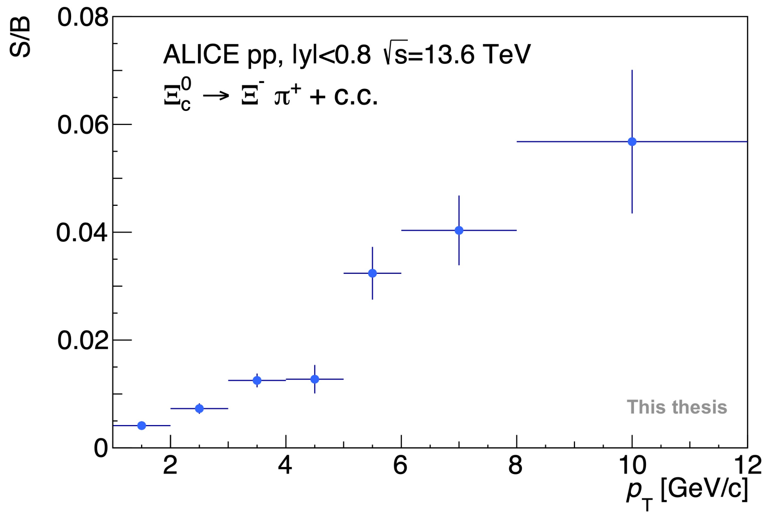


Figure 5.17: Signal-over-background ratio of the Ξ_c^0 as a function of the charmed-baryon transverse momentum.

5.7 Acceptance-times-efficiency correction

The extracted raw yields need to be corrected for the limited detector acceptance, as well as for the reconstruction and selection efficiency. To do so, both the preselections efficiency, which includes the reconstruction efficiency, and the BDT efficiency must be taken into account. The total $A \cdot \epsilon$ corresponds to their product and is computed as the ratio of the reconstructed signal candidates fulfilling the preselections and being correctly classified by the BDT ($N_{\mathrm{R,BDT}}$) over the total number of generated Ξ_c^0 and $\overline{\Xi}_c^0$ with $|y| < 0.8$ and decaying to the desired channel (N_{G}). This translates into

$$(A \cdot \epsilon)_{\mathrm{TOT}} = (A \cdot \epsilon)_{\mathrm{presel}} \cdot (A \cdot \epsilon)_{\mathrm{BDT}} = \frac{N_{\mathrm{R,presel}}}{N_{\mathrm{G}}} \cdot \frac{N_{\mathrm{R,BDT}}}{N_{\mathrm{R,presel}}} = \frac{N_{\mathrm{R,BDT}}}{N_{\mathrm{G}}}, \tag{5.12}$$

where $N_{\mathrm{R,presel}}$ is the number of reconstructed signal candidates fulfilling the preselections.

The prompt and non-prompt components are considered separately. The $A \cdot \epsilon$ is averaged for Ξ_c^0 and $\overline{\Xi}_c^0$, with the larger absorption for anti-protons with respect to protons being properly reproduced in the simulation.

The resulting acceptance-times-efficiency spectrum, reported in Fig. 5.18, shows that lower efficiency values are obtained in the low p_{T} region. Given that such kinematic range is characterised by both higher combinatorial background and smaller Lorentz-boost, the selections needed to remove enough background candidates to allow for signal extraction result in a higher signal loss. The efficiency drop at high p_{T} can be understood in terms of cascades reconstruction requirements. As a matter of fact, for all the tracks coming from the Ξ decay, namely the two charged pions and the proton, both TPC and ITS information is required, as discussed in Section 3.3.2. First of all, one has to keep in mind the particles $c\tau$, with τ being the particle mean lifetime: it corresponds to ~ 45 μm for the Ξ_c^0, ~ 4.9 cm for the Ξ^- and ~ 7.8 cm for the Λ. Given that the ITS external radius is ~ 34 cm, for high-p_{T} Ξ_c^0, and therefore high-p_{T} Ξ^- and Λ, the cascade and V0 $c\tau$ together with the Lorentz boost in the transverse direction will likely result in a lack of ITS information for the daughters tracks, especially for those coming from the Λ. As a consequence, the corresponding charmed-mother particle will be lost. The measurement of $\Xi^{\pm}$ production at $\sqrt{s} = 13.6$ TeV is ongoing within the ALICE collaboration, and the current estimate of the cascade $A \cdot \epsilon$, reported in Fig. 5.20, also exhibits a decreasing trend at high $p_{\mathrm{T}\,\Xi}$. The observed $p_{\mathrm{T}\,\Xi}$ trend is the result of both reconstruction requirements and analysis level selections, and is consistent with the considerations reported above.

Because of the BDT selections, there is a slight decrease in the $A \cdot \epsilon$ with respect to the preselection step (see the preselection efficiency in Fig. 5.3 and the BDT efficiency in Table 5.9). The effect of training the BDT models only with prompt signal candidates results in a larger drop of the non-prompt efficiency. This is due to the fact that the non-prompt signal exhibits specific distributions for multiple training features and the BDT selections are optimised to select the prompt component.

The training feature with the highest separation power when performing a multi-classification task for prompt Ξ_c^0, non-prompt Ξ_c^0, and background is the impact parameter of the pion $\pi \leftarrow \Xi_c^0$ track with respect to the primary vertex, as shown in Fig. 5.19 for the p_{T} bin $4 < p_{\mathrm{T}\,\Xi_c^0} < 6$ GeV/c. The same result is observed both at low and high p_{T}, as reported in Appendix K. Even if the BDT training carried out to produce this plot has a limited reliability because of the low number of prompt and non-prompt candidates available, this observation is in line with the expectations, as non-prompt charmed-hadrons decays exhibit a more displaced charm decay vertex with respect to prompt decays, and consequently the corresponding impact parameter of the pion $\pi \leftarrow \Xi_c^0$ track will be larger. Moreover, the ITS upgrade carried out during LS2 allows to reach an excellent impact parameter resolution, as discussed in Section 3.2.1. Therefore, besides being a variable exhibiting a characteristic behaviour in non-prompt decays, the impact parameter of the pion $\pi \leftarrow \Xi_c^0$ is also measured with an excellent resolution.

The corrected Ξ_c^0 yield $N_{\Xi_c^0}$ is computed as

$$N_{\Xi_c^0}(p_{\mathrm{T}}) = \frac{N_{\Xi_c^0,\,\mathrm{raw}}(p_{\mathrm{T}})}{(A\cdot\epsilon)_{\mathrm{TOT}}(p_{\mathrm{T}})} \tag{5.13}$$

where $N_{\Xi_c^0,\,\mathrm{raw}}$ is the raw Ξ_c^0 yield extracted by fitting the invariant-mass spectra and $(A\cdot\epsilon)_{\mathrm{TOT}}$ is the total prompt acceptance-times-efficiency reported in Fig. 5.18. The corrected yield thus obtained is shown in Fig. 5.21.

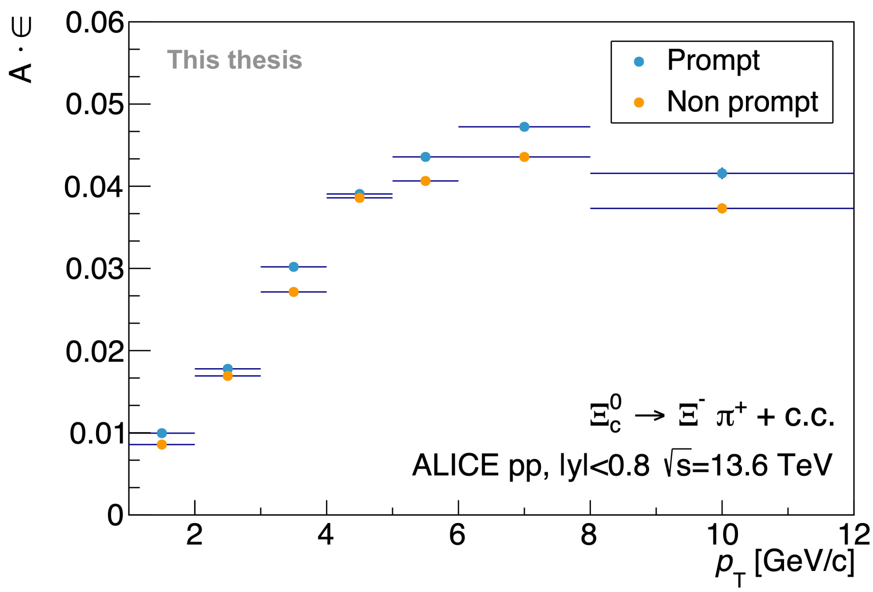
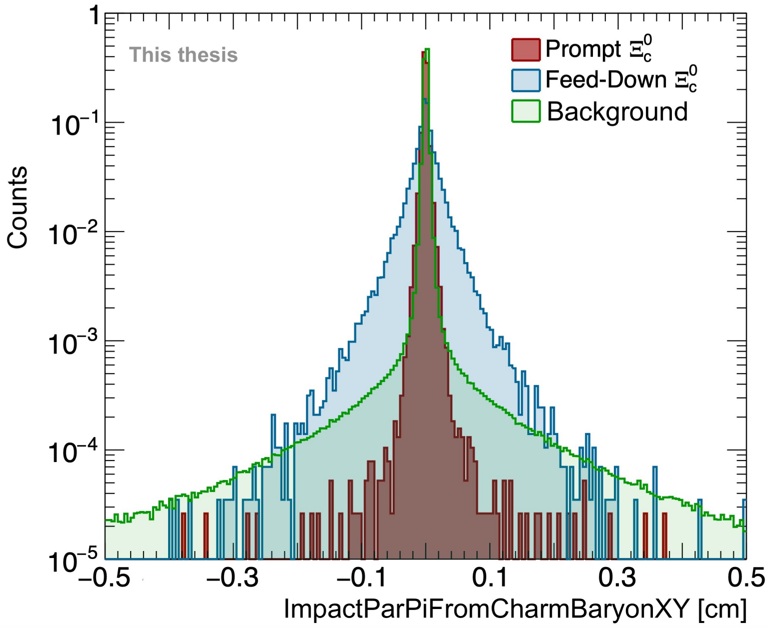
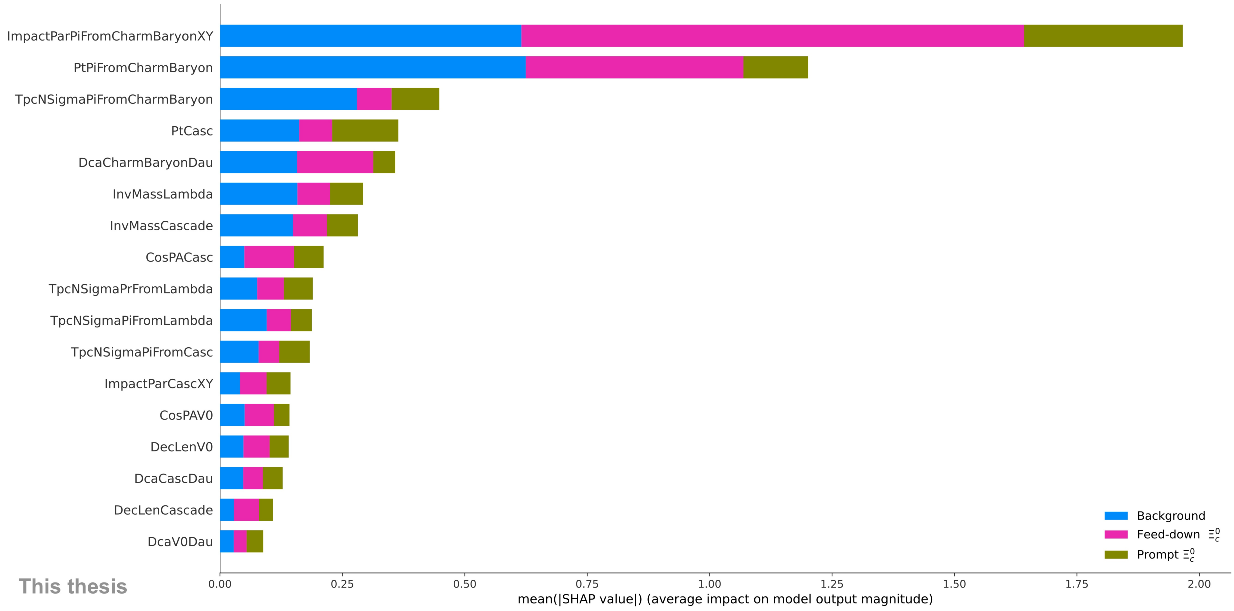


Figure 5.18: Total acceptance-times-efficiency for prompt and non prompt Ξ_c^0 as a function of the charmed-baryon transverse momentum.



(a) Normalised distribution of $\pi \leftarrow \Xi_c^0$ track impact parameter with respect to the primary vertex for the different training classes.



(b) Ranking of training features based on their impact on the candidates classification for the BDT model.

Figure 5.19: Result of a BDT multi-classification with prompt signal, non-prompt signal, and background classes for the $p_{\mathrm{T}\,\Xi_c^0}$ interval $4 < p_{\mathrm{T}\,\Xi_c^0} < 6$ GeV/c.

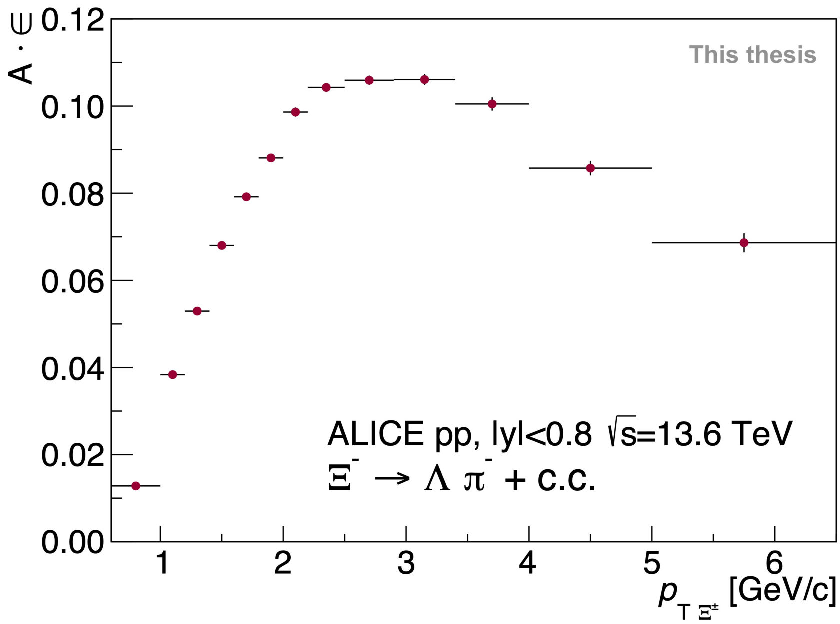


Figure 5.20: Current estimate of $\Xi^{\pm}$ acceptance-times-efficiency, from an ongoing ALICE analysis [30].

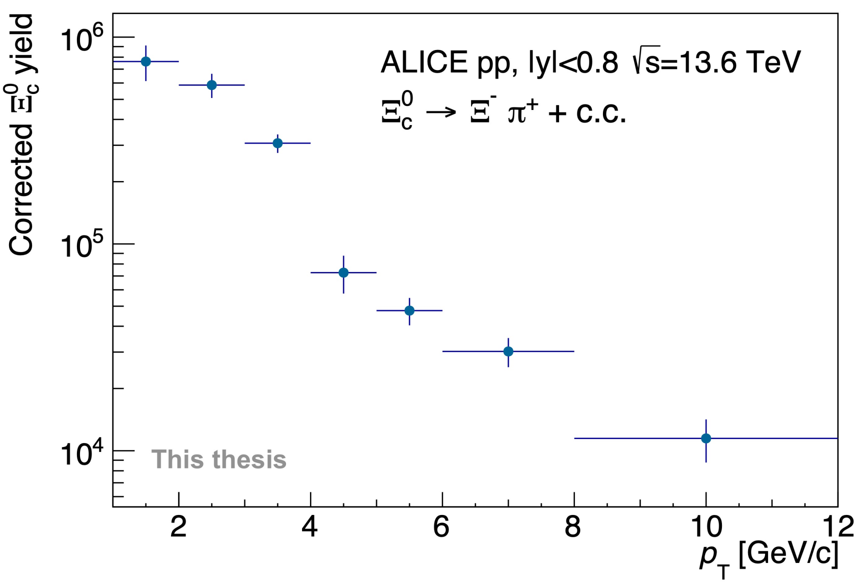


Figure 5.21: Corrected Ξ_c^0 yield as a function of the charmed-baryon transverse momentum.

5.8 Prompt-fraction estimation

As no distinction between the prompt and non-prompt component is made when reconstructing the Ξ_c^0, in order to calculate the prompt cross section, the measured corrected yield $N_{\Xi_c^0}$ has to be scaled for the prompt fraction f_{prompt}, i.e. the ratio between the number of prompt Ξ_c^0 and the total number of Ξ_c^0

$$N_{\Xi_c^0,\,\mathrm{prompt}} = N_{\Xi_c^0} \cdot f_{\mathrm{prompt}}. \tag{5.14}$$

In this analysis, a theory-driven feed-down subtraction method is used to remove the contributions to the Ξ_c^0 yield from beauty-hadron decays. This method involves calculating the prompt fraction as

$$\begin{aligned} f_{\mathrm{prompt}} &= 1 - \frac{N^{\mathrm{feed-down}}_{\Xi_c^0,\,\mathrm{raw}}}{N_{\Xi_c^0,\,\mathrm{raw}}} \\ &= 1 - \left(\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_{\mathrm{T}}\mathrm{d}y}\right)^{\mathrm{theory}}_{b\to\mathrm{X}\to\Xi_c^0} \cdot \frac{(A\cdot\epsilon)_{\mathrm{tot,\,non-prompt}}\cdot\Delta y\cdot\Delta p_{\mathrm{T}}\cdot 2\cdot\mathrm{BR}\cdot\mathcal{L}_{\mathrm{int}}}{N_{\Xi_c^0,\,\mathrm{raw}}}, \end{aligned} \tag{5.15}$$

where $\left(\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_{\mathrm{T}}\mathrm{d}y}\right)^{\mathrm{theory}}_{b\to\mathrm{X}\to\Xi_c^0}$ is the theoretical prediction for the non-prompt Ξ_c^0 cross section and $(A\cdot\epsilon)_{\mathrm{tot,\,non-prompt}}$ is the total non-prompt acceptance-times-efficiency. The latter is computed as the product of the non-prompt preselection efficiency times the non-prompt BDT efficiency and is shown in Fig. 5.18. The terms Δp_{T} and Δy correspond to the width of the transverse momentum bin and rapidity range considered, the factor 2 accounts for the contribution from the anti-particle, $\mathcal{L}_{\mathrm{int}}$ is the integrated luminosity analysed, and $N_{\Xi_c^0,\,\mathrm{raw}}$ is the measured raw yield reported in Fig. 5.15. The term BR is computed as the product of all the branching ratios of the decays of the chain $\mathrm{BR} = \mathrm{BR}_{\Xi_c^0\to\Xi\,\pi}\cdot\mathrm{BR}_{\Xi\to\Lambda\,\pi}\cdot\mathrm{BR}_{\Lambda\to p\,\pi}$, with the values taken from the PDG.

Given that there are no theoretical predictions and/or measurements of the non-prompt Ξ_c^0 cross section, the term $\left(\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_{\mathrm{T}}\mathrm{d}y}\right)^{\mathrm{theory}}_{b\to\mathrm{X}\to\Xi_c}$ is computed starting from the FONLL prediction for beauty quark production and the results obtained by LHCb for beauty-hadron fragmentation fractions. Let $\mathcal{F}$ be a function describing both the beauty quark fragmentation and the corresponding beauty-hadron decay, the theoretical prediction for non-prompt Ξ_c^0 cross section can be written as

$$\begin{aligned} \left(\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_{\mathrm{T}}\mathrm{d}y}\right)^{\mathrm{theory}}_{b\to\mathrm{X}\to\Xi_c} &= \left(\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_{\mathrm{T}}\mathrm{d}y}\right)^{\mathrm{FONLL}}_{b} \cdot\mathcal{F}(b\to\mathrm{X}\to\Xi_c) \\ &= \left(\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_{\mathrm{T}}\mathrm{d}y}\right)^{\mathrm{FONLL}}_{b} \cdot\mathcal{F}(b\to\mathrm{X}\to\Lambda_c)\cdot\frac{\mathcal{F}(b\to\mathrm{X}\to\Xi_c)}{\mathcal{F}(b\to\mathrm{X}\to\Lambda_c)} \\ &= \left(\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_{\mathrm{T}}\mathrm{d}y}\right)^{\mathrm{FONLL}}_{b\to\mathrm{X}\to\Lambda_c} \cdot\frac{\mathcal{F}(b\to\mathrm{X}\to\Xi_c)}{\mathcal{F}(b\to\mathrm{X}\to\Lambda_c)} \\ &= \left(\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_{\mathrm{T}}\mathrm{d}y}\right)^{\mathrm{FONLL}}_{b\to\mathrm{X}\to\Lambda_c} \cdot\frac{\sum_{\mathrm{H}_b} b\to\mathrm{H}_b\to\Xi_c}{\sum_{\mathrm{H}_b} b\to\mathrm{H}_b\to\Lambda_c}, \end{aligned} \tag{5.16}$$

where $\left(\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_\mathrm{T}\mathrm{d}y}\right)_b^{\mathrm{FONLL}}$ is the theoretical prediction from fixed-order-next-to-leading-log (FONLL) calculations for the beauty quark production cross section. Then, to estimate the theoretical prediction for $\left(\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_\mathrm{T}\mathrm{d}y}\right)_{b\to\mathrm{X}\to\Lambda_c}^{\mathrm{FONLL}}$, the LHCb measurements [54] of fragmentation fractions are used and the decay of the beauty-hadron X is taken care of by Pythia. The term $\frac{\sum_{\mathrm{H}_b} b\to\mathrm{H}_b\to\Xi_c}{\sum_{\mathrm{H}_b} b\to\mathrm{H}_b\to\Lambda_c}$ corresponds to the ratio of non-prompt Ξ_c^0 over non-prompt Λ_c^+. According to the ALICE measurement [55], the prompt and non-prompt Λ_c^+/D^0 ratios are compatible, so it is reasonable to assume that the same holds for the prompt and non-prompt Ξ_c^0/Λ_c^+ ratios. Therefore, one can write

$$\frac{\sum_{\mathrm{H}_b} b\to\mathrm{H}_b\to\Xi_c}{\sum_{\mathrm{H}_b} b\to\mathrm{H}_b\to\Lambda_c} \approx \frac{c\to\Xi_c}{c\to\Lambda_c}. \tag{5.17}$$

It follows that the Ξ_c^0 feed-down can be estimated as

$$N_{\Xi_c^0,\,\mathrm{raw}}^{\mathrm{feed-down}} = \left(\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_\mathrm{T}\mathrm{d}y}\right)_{b\to\mathrm{X}\to\Lambda_c^+}^{\mathrm{FONLL}} \cdot \frac{c\to\Xi_c^0}{c\to\Lambda_c^+} \cdot (A\cdot\epsilon)_{\mathrm{tot,\,non-prompt}} \cdot \Delta y \cdot \Delta p_\mathrm{T} \cdot 2 \cdot \mathrm{BR} \cdot \mathcal{L}_{\mathrm{int}}. \tag{5.18}$$

The FONLL prediction for non-prompt Λ_c^+ production used in this analysis is shown in Fig. 5.22. The systematic errors reported in this plot take into account the uncertainties on the FONLL theoretical prediction as well as the uncertainties of the LHCb fragmentation fraction measurements.

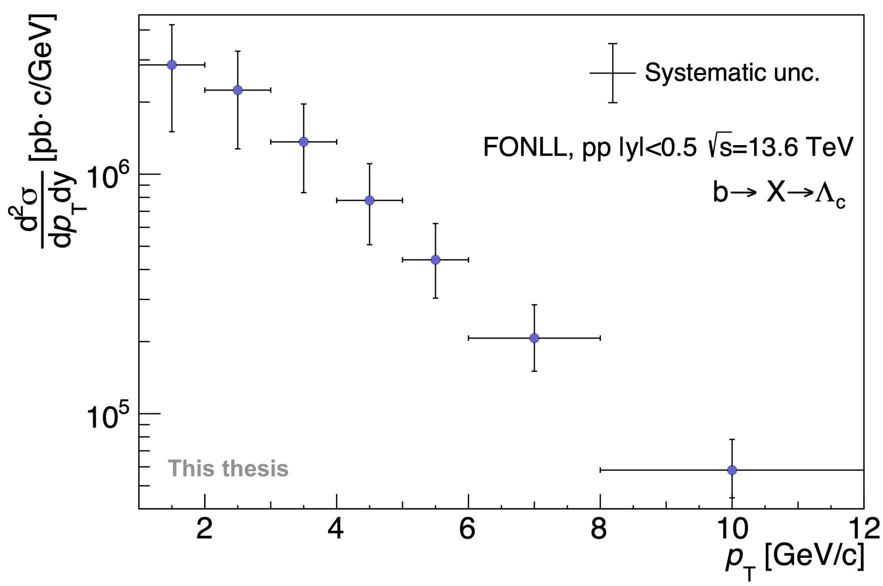


Figure 5.22: Theoretical prediction for non-prompt Λ_c^+ production based on FONLL calculations, LHCb measurements of fragmentation fractions and Pythia decays.

The charm fragmentation fractions ratio $\frac{c\to\Xi_c^0}{c\to\Lambda_c^+}$ is reported in Fig. 5.23. It is computed as the ratio of the Ξ_c^0 and Λ_c^+ prompt cross sections measured by ALICE [32]. This procedure is justified in Section

2.4. The uncertainties reported in the plot are obtained by propagating on the ratio the sum in quadrature of the statistical and systematic uncertainties for each cross section measurement.

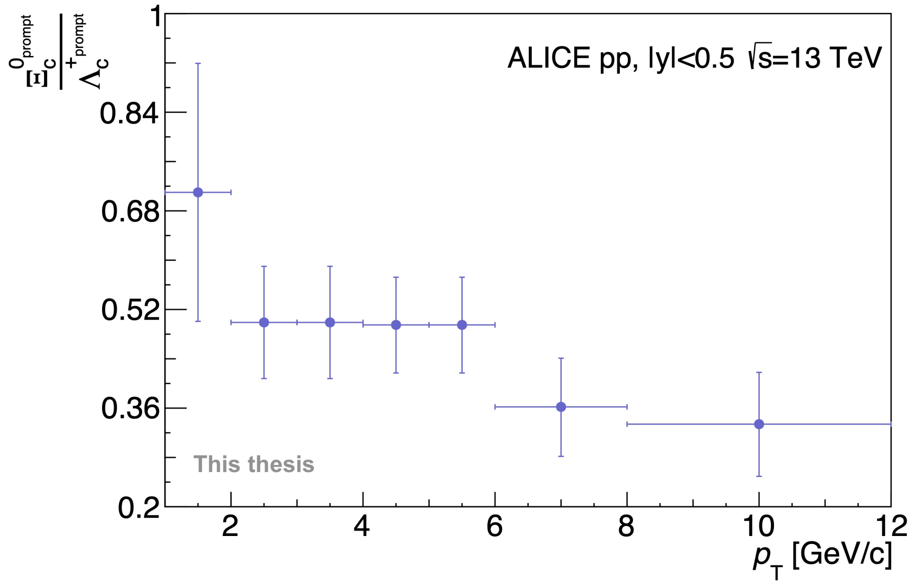


Figure 5.23: Ratio of Ξ_c^0 and Λ_c^+ prompt cross sections measured by ALICE in pp collisions [32].

The Ξ_c^0 feed-down thus calculated is reported in Fig. 5.24. The error bars shown report the propagation of the uncertainties on the FONLL prediction and on the cross sections ratio previously described. Finally, the $p_{\rm T}$-differential prompt fraction computed accordingly to the calculations reported above is shown in Fig. 5.25. As expected, the non-prompt component accounts only for a minor fraction of the Ξ_c^0 yield. The error bars reported in Fig. 5.25 correspond the total systematic uncertainty on $f_{\rm prompt}$ and are obtained by propagating the systematics on $N^{\rm feed-down}_{\Xi_c^0,\,{\rm raw}}$. The statistical uncertainties on $(A \cdot \epsilon)_{\rm non-prompt}$ are negligible, while the raw yield ones together with the uncertainties on $\rm BR$ and $\mathcal{L}_{\rm int}$ are taken into account at a later stage, when computing the cross section.

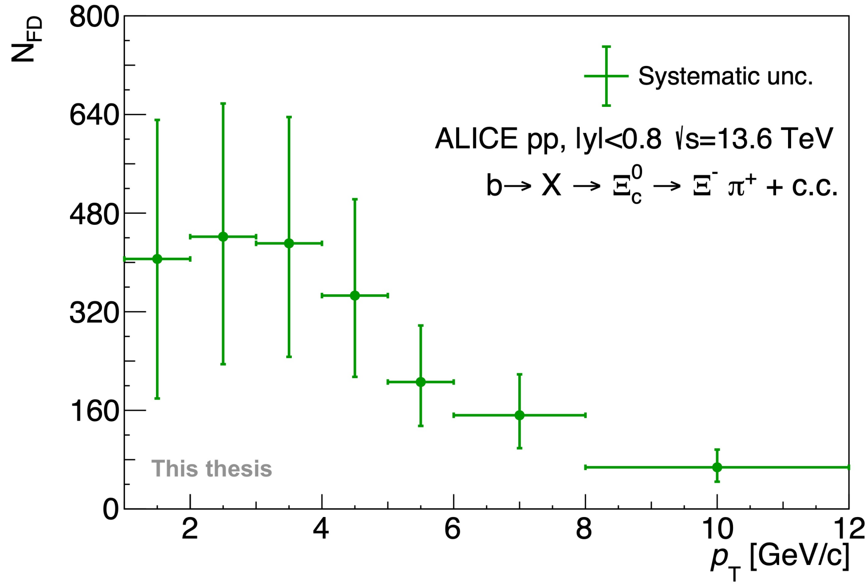


Figure 5.24: Non-prompt Ξ_c^0 yield estimated with the theory-driven feed-down subtraction method.

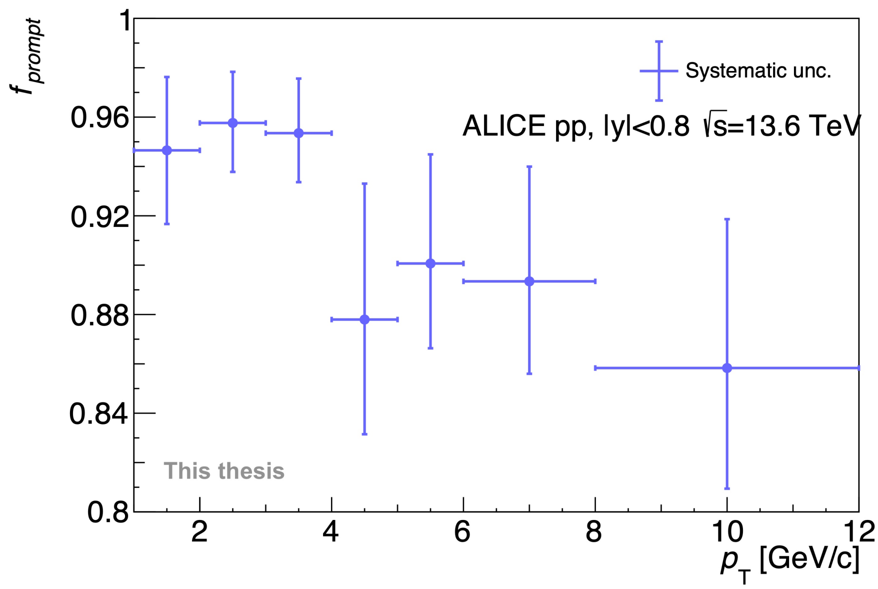


Figure 5.25: Result of Ξ_c^0 prompt-fraction estimation.

5.9 Systematic uncertainties

The different steps in the analysis procedure introduce systematic uncertainties that have to be taken into account in the measurement of the cross section. The sources of such uncertainties are the raw yield extraction settings, the not perfectly realistic description in the MC of the features considered in the BDT selections and in the track quality preselections, thus introducing a sensitivity of the final result to the BDT working point choice and to the track quality cuts, the choice of p_{T} shape taken as input in the MC simulation and the non perfectly realistic description of the TPC-ITS matching efficiency in the MC simulation. The (asymmetric) systematic uncertainty associated to the prompt-fraction estimation and presented in Section 5.8 is also included. The predominant contributions come from the raw yield extraction, the BDT working point choice, and the ITS-TPC matching efficiency. At low and high p_{T}, also the tracking preselections of the Ξ daughters give a relevant contribution to the total systematic uncertainty. As the BDT selections are tighter with respect to the preselections discussed in Section 5.4 and a systematic uncertainty is associated to the choice for the BDT working point, there is no further uncertainty associated to the preselections. Only the contribution of the tracking quality preselections is considered separately as the corresponding variables are not processed by the BDT. The relative systematic uncertainties are reported in Table 5.10. The final systematic is computed as the sum in quadrature of all the uncertainties listed, as their sources are considered uncorrelated.

$p_{\mathrm{T}\,\Xi_c^0}$ [GeV/c]	1< p_{T} <2	2< p_{T} <3	3< p_{T} <4	4< p_{T} <5	5< p_{T} <6	6< p_{T} <8	8< p_{T} <12
Raw yield extraction	8%	7%	8%	18%	9%	14%	13%
BDT working point	13%	13%	13%	13%	13%	13%	13%
$\pi \leftarrow \Xi_c^0$ tracking preselections	5%	5%	5%	5%	5%	5%	5%
Ξ daughters tracking preselections	16%	15%	3%	3%	2%	8%	9%
ITS-TPC matching efficiency	12%	12%	12%	12%	12%	12%	12%
Prompt fraction	+3% -3%	+2% -2%	+2% -2%	+5% -6%	+4% -5%	+4% -5%	+6% -7%
MC p_{T} shape	1%	3%	2%	1%	1%	1%	8%
Total systematic	+26% -26%	+25% -25%	+20% -20%	+26% -27%	+21% -21%	+25% -25%	+26% -26%

Table 5.10: Systematic uncertainties.

The uncertainties on the integrated luminosity and on the branching ratio are taken into account separately. According to the indication provided by the ALICE Physics Coordination, the luminosity uncertainty amounts to 10%. The precise value used in this analysis is $\mathcal{L}_{\mathrm{int}} = (0.793 \pm 0.079)\ \mathrm{pb}^{-1}$. In the cross section measurement, the product of the branching ratios corresponding to each decay of the chain is considered. Therefore, the total BR relative uncertainty is computed as the sum in quadrature of the relative uncertainties on all the BR as reported in the PDG. The values used are listed in Table 5.11 and the total relative uncertainty on the BR amounts to 19%.

Decay channel	**Branching ratio**
$\Xi_c^0 \rightarrow \Xi^- \, \pi^+$	$(1.43 \pm 0.27)\%$
$\Xi^- \rightarrow \Lambda \, \pi^-$	$(99.887 \pm 0.035)\%$
$\Lambda \rightarrow p \, \pi^-$	$(64.1 \pm 0.5)\%$

Table 5.11: Branching ratios and corresponding uncertainties as reported in the PDG.

5.9.1 Raw yield extraction

The systematic uncertainty associated to the raw yield extraction is estimated using a multi-trial fit procedure that implements multiple combinations of the fit settings. Across the trials, the background fit function is changed between a second- and a third-order polynomial, the signal shape is always set to a Gaussian function with no constraint on the mean parameter. However, its width value is either free or fixed, with the values considered being the MC value σ_{MC} and $\sigma_{\mathrm{MC}} \pm 10\%$. Variations of both the lower and upper limits of the fit range as well as of the invariant-mass bin width are also implemented. For each fit with a reduced χ^2 no greater than 3 the raw yield is extracted. Then, for each p_{T} bin, the systematic uncertainty is calculated as the sum in quadrature of the RMS of the raw yield distribution and the shift of its mean with respect to the raw yield central value measured as described in Section 5.6.

Fig. 5.26 shows the results of the multi-trial approach for the raw yield extraction in the bin $1 < p_{\mathrm{T}} < 2$ GeV/c. The analogous plots for the remaining p_{T} bins are reported in Appendix L. On the right side, the extracted raw yield is shown as a function of the fit trial number, while on the left side the raw yield distribution is reported. The red line indicates the raw yield central value, while in blue is depicted the raw yield distribution extracted as described above. If the signal extraction is stable, the red line is expected to lie well within the blue distribution. This is the case for all the p_{T} bins. As an additional check, the raw yield is also measured with a bin-counting method considering both a 5σ and a 3σ interval. The associated distributions are shown in orange and green and the corresponding results are in general similar to those obtained with the standard method. The transverse momentum interval $4 < p_{\mathrm{T}} < 5$ GeV/c (see Fig. L.3) exhibits a larger systematic uncertainty due to the larger raw yield distribution. Moreover, both the bin counting distributions have larger mean values compared to the blue distribution. These observations are ascribed to the statistical downward fluctuation discussed in Section 5.6, which makes the extracted raw yields more drastically affected by the fit settings. The p_{T} bin $6 < p_{\mathrm{T}} < 8$ GeV/c (see Fig. L.5) also shows a particularly large systematic uncertainty. This is due to the presence of outliers at high yield values, thus resulting in a large RMS. In order to neglect the contribution of these instances, a lower uncertainty is assigned to this bin. In this case, the value of the systematic is chosen in line with the 5σ bin counting result, as the corresponding distribution is not affected by any outlier and shows a good agreement with the blue distribution.

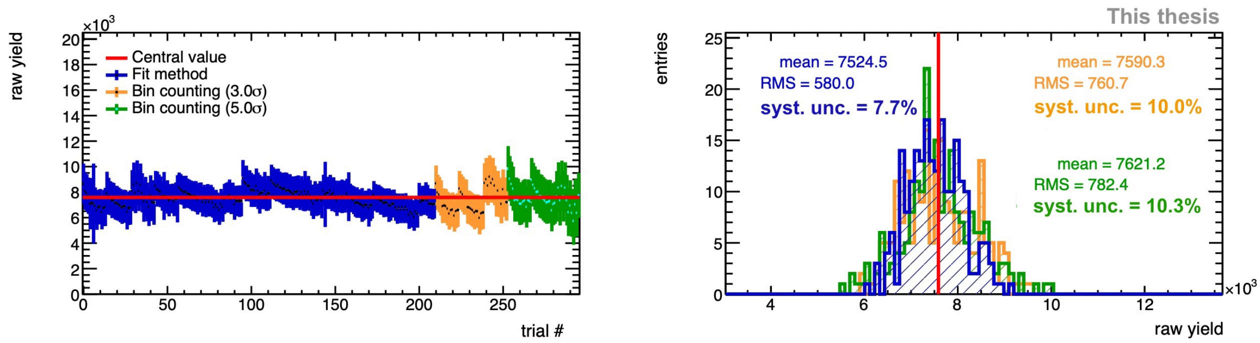


Figure 5.26: Result of multi-trial approach for raw yield extraction in the transverse momentum interval $1 < p_{\mathrm{T}\,\Xi_c^0} < 2$ GeV/c. On the left side the raw yield is shown as a function of the trial number, on the right side the raw yield distributions are reported.

The exact uncertainty values estimated with the procedure described above are reported in Table 5.12, while the final systematic uncertainty assigned to each $p_{\mathrm{T}\,\Xi_c^0}$ bin is listed in Table 5.10. In general, slightly larger systematic uncertainties are observed at increasingly high p_{T}. This is due to the decrease of statistics at high transverse momenta resulting in larger statistical uncertainties and larger impact of data points fluctuations, and therefore in a less stable fit. The improvement of the signal-over-background ratio at high p_{T} discussed in Section 5.6 can not compensate for this effect.

$p_{\mathrm{T}\,\Xi_c^0}$ **[GeV/c]**	$1 < p_{\mathrm{T}} < 2$	$2 < p_{\mathrm{T}} < 3$	$3 < p_{\mathrm{T}} < 4$	$4 < p_{\mathrm{T}} < 5$	$5 < p_{\mathrm{T}} < 6$	$6 < p_{\mathrm{T}} < 8$	$8 < p_{\mathrm{T}} < 12$
Syst. unc.	8%	7%	8%	18%	9%	27%	13%

Table 5.12: Systematic uncertainties associated to raw yield extraction computed according to the described procedure.

5.9.2 BDT selections

The BDT selections corresponding to the chosen working point introduce a systematic uncertainty that originates from potential differences between the training features in MC compared to real data. This would result in a biased BDT efficiency and consequently a wrong correction of the raw yield.

To estimate the impact of this possible mismatch, a working point scan corresponding to a BDT efficiency variation of $\pm 20\%$ with respect to the central value is performed. For each working point, the raw yield is extracted and only fits with a significance larger than 3 are considered. Then, the corresponding total prompt acceptance-times-efficiency is calculated in order to retrieve the corrected yield. Finally, the systematic uncertainty is computed as the sum in quadrature of the RMS of the corrected yield distribution and the shift of its mean with respect to the corrected yield central value. Fig. 5.27 shows the corrected yield distribution (top) and the trend of the corrected yield as a function of the BDT working point (bottom) for $1 < p_{\mathrm{T}} < 2$ GeV/c. The red dashed line indicates the central corrected yield value. The analogous plots for the remaining transverse momentum intervals are reported in Appendix M. For most of the p_{T} bins, the red line is quite centred with respect to the corrected yield distribution, as expected. However, in the plots corresponding to the transverse momentum intervals $4 < p_{\mathrm{T}} < 5$ (Fig. M.3) and $6 < p_{\mathrm{T}} < 8$ GeV/c (Fig. M.5) the central value lies in the left tail of the distribution. As for the $4 < p_{\mathrm{T}} < 5$ GeV/c transverse momentum interval, this is interpreted as the result of the statistical downward fluctuation previously discussed and the consequent instability of the signal extraction that introduce a larger shift of the central value with respect to the mean of the corrected yield distribution. However, what is of greater concern is the fact that a positive correlation between the corrected yield and the BDT cut is observed. As shown in the lower plots of Fig. 5.27 and Fig. M.1-M.6, the corrected yield decrease at increasingly tighter cuts on the BDT score. This is a symptom of the simulation not accurately describing the data, thus introducing a mismatch between the real efficiency and its estimate based on the MC. As a consequence, the corrected yield distributions are extremely wide and and exhibit large RMS values. Most of them are quite far from a Gaussian-like shape, as it would be expected in case a healthy simulation was used. This issue seems to be the cause of the large systematic uncertainty found for the p_{T} bin $6 < p_{\mathrm{T}} < 8$ GeV/c, at least partly. As a matter of fact, in the bottom plot reported in Fig. M.5 there is a particularly strong dependence of the corrected yield on the BDT cut. This problem could be addressed by producing more realistic simulations, for instance improving calibrations to reproduce the DCA and p_{T} resolution in the MC and extending the data-driven tuning to deep secondary tracks. The exact uncertainty values estimated with this procedure are reported in Table 5.13. No clear p_{T} dependence of the systematic uncertainties is observed and, also given that the same efficiency variation has been considered across all the p_{T} intervals, there is no reason to expect large variations in adjacent p_{T} bins. Hence, the values are smoothened: the final systematic uncertainty is computed as the average of the values reported in Table 5.13 and is evenly assigned to all the p_{T} bins. This choice is also supported by the fact that the BDT cuts applied to extract the central values are quite loose and similar. Moreover, the Monte Carlo appears to be not healthy and demonstrated limited reliability. The final uncertainties associated to the BDT working point choice are reported in Table

5.10.

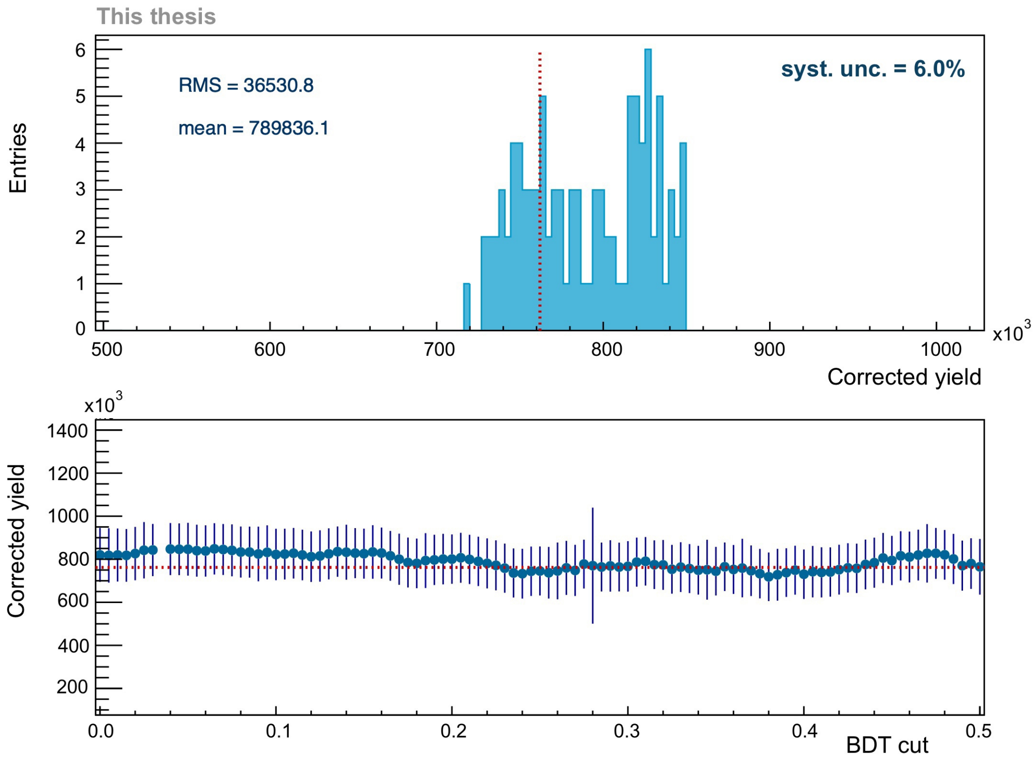


Figure 5.27: Systematic uncertainty estimation associated to BDT selections for the transverse momentum interval $1 < p_{\mathrm{T}\,\Xi_c^0} < 2$ GeV/c. In the top plot the corrected yield distribution is shown, while the bottom plot reports the corrected yield as a function of the BDT score cut.

$p_{\mathrm{T}\,\Xi_c^0}$ **[GeV/c]**	$1< p_\mathrm{T} <2$	$2< p_\mathrm{T} <3$	$3< p_\mathrm{T} <4$	$4< p_\mathrm{T} <5$	$5< p_\mathrm{T} <6$	$6< p_\mathrm{T} <8$	$8< p_\mathrm{T} <12$
Syst. unc.	6%	9%	6%	20%	12%	27%	7%

Table 5.13: Systematic uncertainties associated to BDT selections computed according to the described procedure.

5.9.3 Tracking quality preselections

The systematic uncertainty associated to the track quality requirements is estimated by varying the selections on TPC- and/or ITS-related variables. The pion $\pi \leftarrow \Xi_c^0$ track and the three Ξ daughter tracks, namely the two pions and the proton, are considered separately and two distinct systematic uncertainties are extracted.

As for the pion coming directly from the Ξ_c^0 decay, tracking variables both TPC- and ITS-related are considered. The corresponding systematic uncertainty includes the effect of the selections on the number of ITS clusters and TPC crossed rows. The minimum number of ITS clusters requested when calculating the central Ξ_c^0 yield value is 4, then tighter selections requiring at least 5 and 7 clusters are implemented. As for the TPC selections, the central Ξ_c^0 yield value is computed requiring π tracks with at least 70 crossed rows, then tighter selections requestinga minimum of 100 and 120 crossed rows are applied. For each of these selections, the raw yield is extracted and the corresponding total prompt acceptance-times-efficiency is calculated to compute the corrected yield. The final systematic uncertainty corresponds to the maximum variation of the corrected yield with respect to the central value. The raw yields extracted for all the $\pi \leftarrow \Xi_c^0$ tracking preselection configurations and the corresponding total prompt acceptance-times-efficiency values are reported in Appendix N. The corrected yields are depicted in Fig. 5.28, while Fig. 5.29 shows the variations of the corrected Ξ_c^0 yield with respect to the central value for all the $p_{\rm T}$ bins. The exact uncertainties estimated with this procedure are reported in Table 5.14. A large systematic uncertainty is found for the transverse momentum interval $4 < p_{\rm T} < 5$ GeV/c. As shown in Fig. 5.29, this is the result associated to the requirement of having a cluster in all the seven ITS layers, which can be considered an extreme variation of the corresponding selection. Therefore, a smaller systematic uncertainty is assigned to this bin, in line with the results found for the other transverse momentum intervals. The largest value found among the remaining $p_{\rm T}$ bins amounts to 5%, with minor by-by-bin fluctuations, so this uncertainty is assigned evenly to all the $p_{\rm T}$ intervals. The final uncertainties associated to the pion $\pi \leftarrow \Xi_c^0$ tracking preselections are reported in Table 5.10.

$p_{\rm T\,\Xi_c^0}$ [GeV/c]	1< $p_{\rm T}$<2	2< $p_{\rm T}$<3	3< $p_{\rm T}$<4	4< $p_{\rm T}$<5	5< $p_{\rm T}$<6	6< $p_{\rm T}$<8	8< $p_{\rm T}$<12
Syst. unc.	5%	3%	4%	10%	4%	5%	4%

Table 5.14: Systematic uncertainties associated to the pion $\pi \leftarrow \Xi_c^0$ tracking preselections computed according to the described procedure.

As for the systematic uncertainty arising from the tracking preselections associated to the three Ξ daughter tracks, only variations of the number of TPC crossed rows are considered. Since there is no constraint on ITS-related variables at analysis level, there are no variations of the corresponding selections included in the systematic uncertainty estimation. However, a detailed study carried out centrally by the ALICE Data Preparation Group would be needed to investigate the impact of the settings of the cascade reconstruction algorithm and its reproduction in the MC simulation. For the time being, such study is not available. The central Ξ_c^0 yield value is computed requiring tracks with at least 70 crossed rows, then tighter selections requesting a minimum of 100 and 120 crossed rows are applied. The same method used to estimate the systematic uncertainty on the $\pi \leftarrow \Xi_c^0$ tracking preselection is used. The raw yields extracted for all the Ξ daughters tracking preselection configurations together with the corresponding total prompt acceptance-times-efficiency values are reported in Appendix N. The corrected yields are depicted in Fig. 5.30, while Fig. 5.31 shows

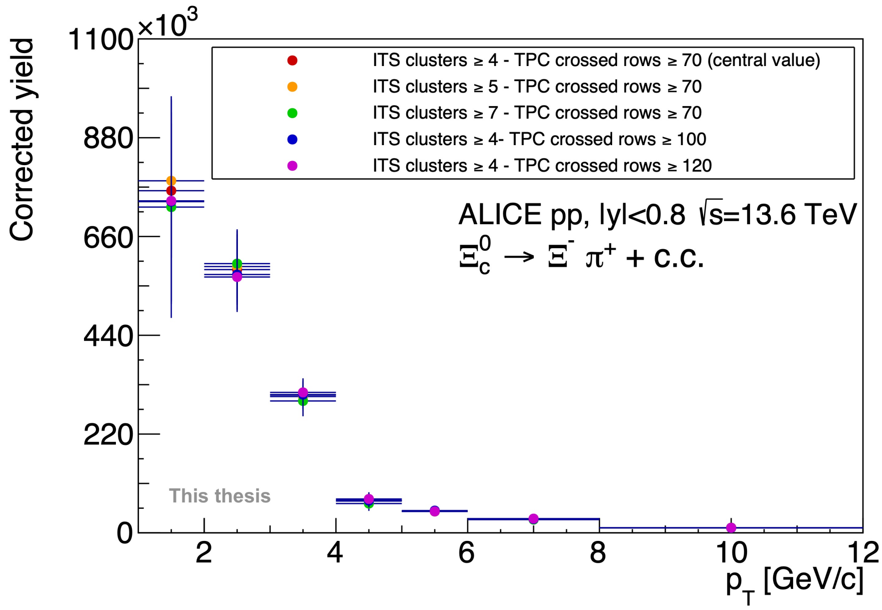


Figure 5.28: Corrected Ξ_c^0 yields for all the pion $\pi \leftarrow \Xi_c^0$ tracking preselection configurations.

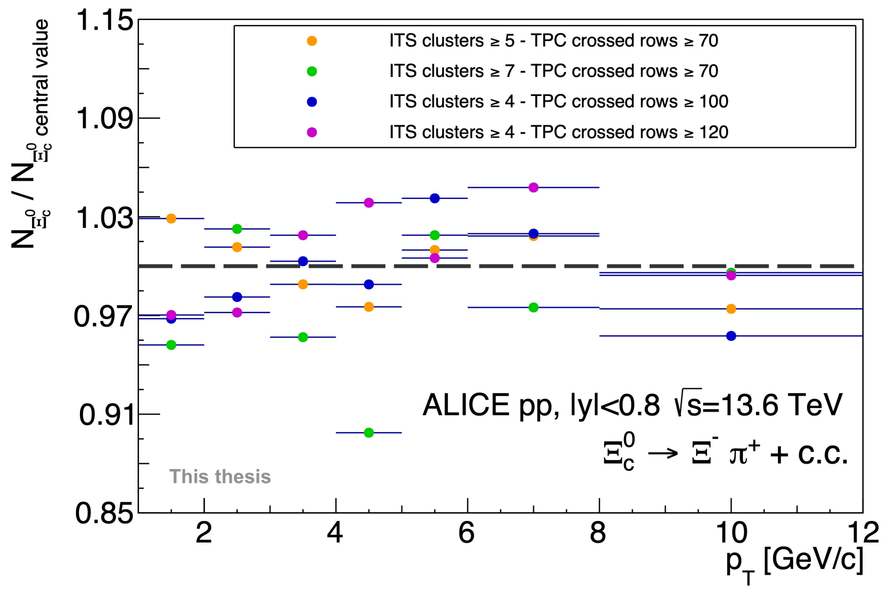


Figure 5.29: Ratio of corrected Ξ_c^0 yields corresponding to the variations of the pion $\pi \leftarrow \Xi_c^0$ tracking preselections over the corrected Ξ_c^0 yield central value. The uncertainties on the ratio are neglected.

the variations of the corrected Ξ_c^0 yield with respect to the central value for all the p_{T} bins. The uncertainties estimated with this procedure are reported in Table 5.15.

$p_{\mathrm{T}\,\Xi_c^0}$ [GeV/c]	1< p_{T}<2	2< p_{T}<3	3< p_{T}<4	4< p_{T}<5	5< p_{T}<6	6< p_{T}<8	8< p_{T}<12
Syst. unc.	16%	15%	3%	3%	2%	8%	9%

Table 5.15: Systematic uncertainties associated to Ξ daughters tracking preselections.

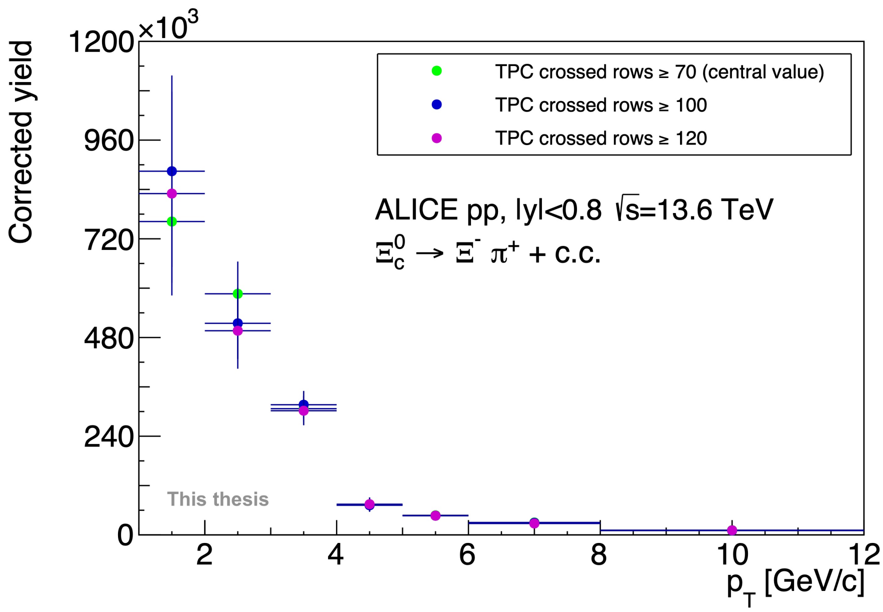


Figure 5.30: Corrected Ξ_c^0 yields corresponding to the variations of the Ξ daughters tracking preselection configurations.

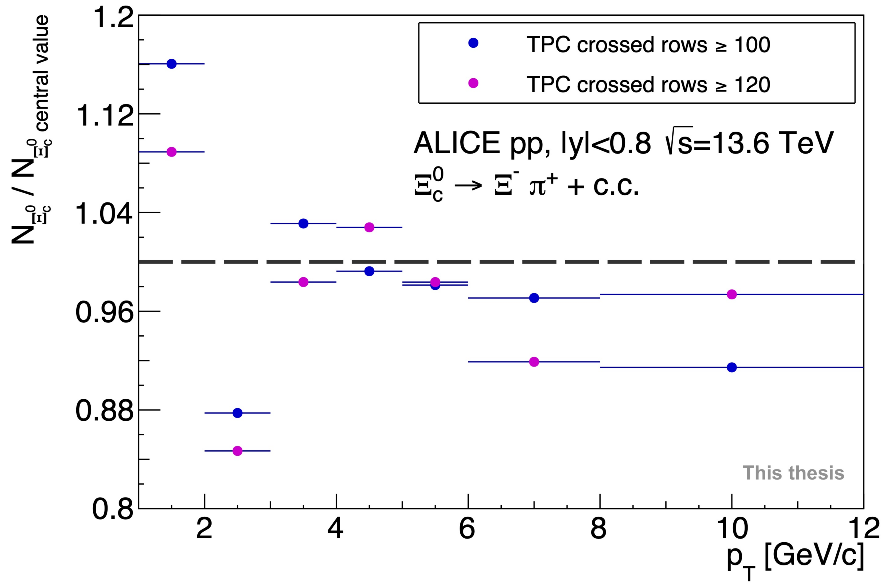


Figure 5.31: Ratio of corrected Ξ_c^0 yields corresponding to the variations of the Ξ daughters tracking preselections over the corrected Ξ_c^0 yield central value. The uncertainties on the ratio are neglected.

5.9.4 Monte Carlo p_{T} shape

The choice for the MC p_{T} input shape, namely the p_{T} distribution of the Ξ_c^0 generated in the simulation, introduces a systematic uncertainty as it may differ from the real Ξ_c^0 transverse momentum spectrum and therefore result in a biased efficiency shape. The MC p_{T} spectrum used in this analysis is predicted by Pythia Colour Reconnection Mode 2. To estimate the systematic uncertainty, the p_{T} spectrum predicted by the Quark re-Combination Mechanism (QCM) model is considered to extract a new $A \cdot \epsilon(p_{\mathrm{T}})$ shape with a reweighting procedure. As a matter of fact, the comparison of the QCM prediction to the Run 2 Ξ_c^0 measurement shows a better agreement than the Pythia one, but due to the lack of resources the MC generator availability is currently limited to Pythia. As a first step, both the Pythia and CR p_{T} spectra are normalized. The comparison of the two predictions is shown in Fig. 5.32. Then, their ratio is computed. A much finer binning with respect to the one chosen to measure the cross section is used in order to better describe the p_{T} dependence. This allows to extract a set of p_{T}-dependent weights, as reported in Fig. 5.33. Afterwards both the reconstructed and generated Ξ_c^0 p_{T} spectra are reweighted and the corresponding total prompt $A \cdot \epsilon$ is calculated. The distribution is then rebinned to retrieve the larger binning used to measure the cross section. This allows to compare the two $A \cdot \epsilon$ spectra, as shown in Fig. 5.34. The relative systematic uncertainty is then computed as the efficiency variation over the central efficiency value. The result is reported in Fig. 5.35. The impact of varying the MC p_{T} spectrum is very small and amounts to a few percent, with minor bin-by-bin fluctuations. However, a larger systematic uncertainty is observed at high p_{T}. This can be explained in terms of low number of entries in the p_{T} spectrum at high transverse momentum and low weight values for $p_{\mathrm{T}} > 8$ GeV/c due to the softening of the spectrum when switching from Pythia to the QCM prediction. Other models are currently not available for the collision energy $\sqrt{s} = 13.6$ TeV, but if the reweighting procedure was based on a spectrum exhibiting larger changes in shape with respect to the Pythia CR prediction, the corresponding systematic uncertainty would be larger. In Table 5.16 the systematic uncertainties associated to the MC p_{T} spectrum are reported.

$p_{\mathrm{T}\,\Xi_c^0}$ [GeV/c]	1< p_{T}<2	2< p_{T}<3	3< p_{T}<4	4< p_{T}<5	5< p_{T}<6	6< p_{T}<8	8< p_{T}<12
Syst. unc.	1%	3%	2%	1%	1%	1%	8%

Table 5.16: Systematic uncertainties associated to MC input p_{T} shape.

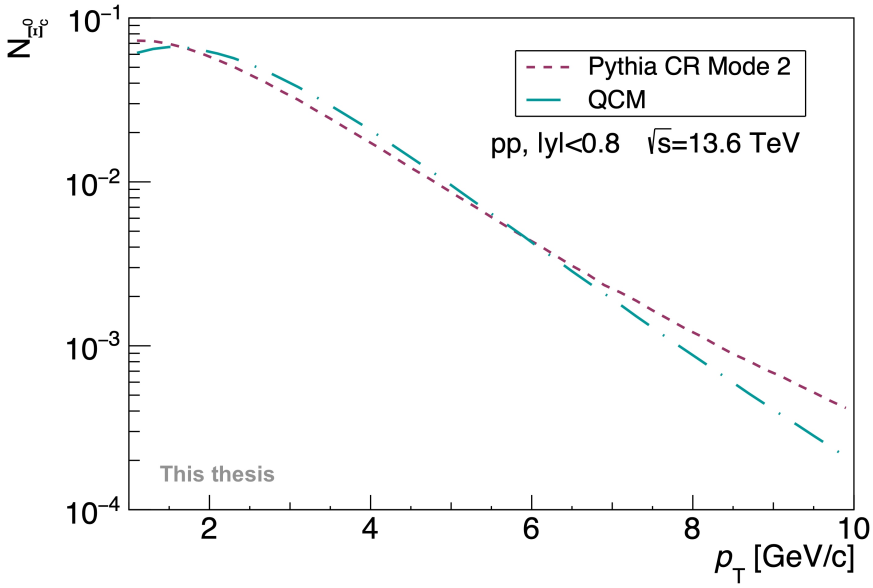


Figure 5.32: Comparison of Ξ_c^0 p_{T} spectrum predictions by QCM and Pythia CR Mode 2.

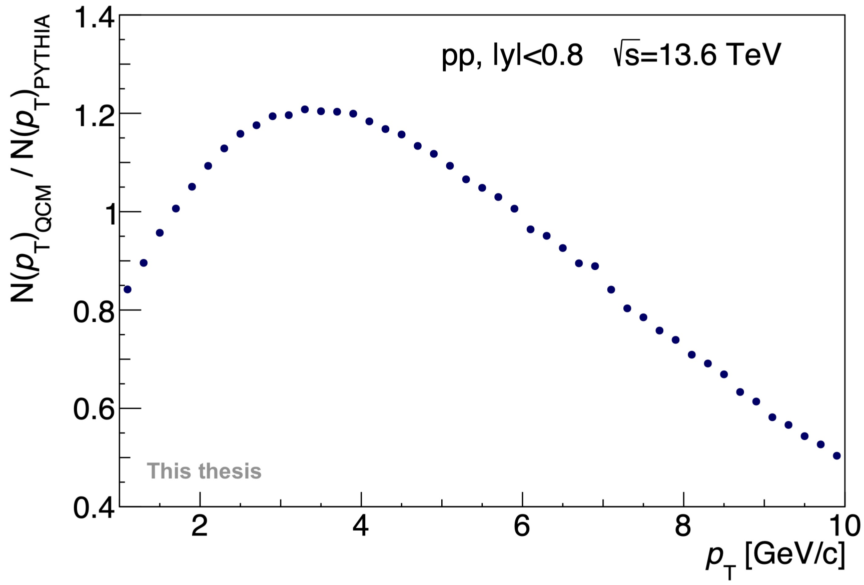


Figure 5.33: Ratio of p_{T} spectrum predictions by QCM and Pythia CR Mode 2.

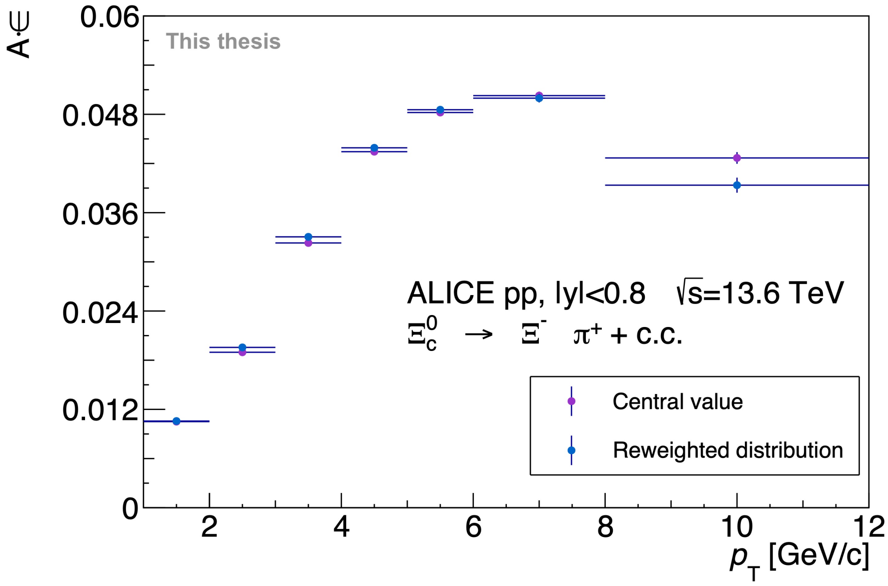


Figure 5.34: Comparison of total prompt acceptance-times-efficiency distributions.

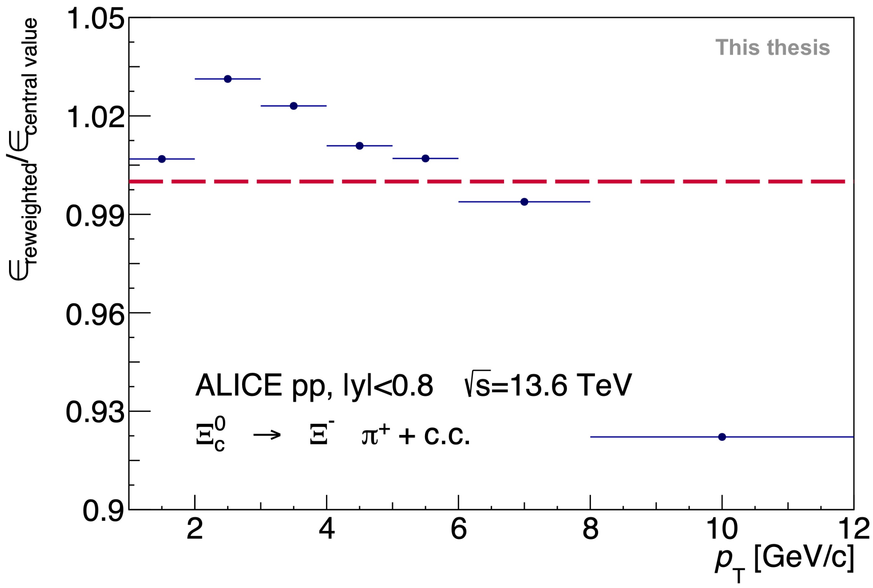


Figure 5.35: Systematic uncertainty associated to MC input $p_{\mathrm{T}\,\Xi_c^0}$ shape. The uncertainties on the ratio are neglected.

5.9.5 ITS-TPC matching efficiency

A further source of systematic uncertainty is the potential mismatch between the reproduction of the TPC-ITS matching efficiency in MC with respect to real data. A study carried out centrally by the ALICE Data Preparation Group estimates the ITS-TPC matching efficiency using a data-driven tag-and-probe method that exploits the D-meson three-body decay $\mathrm{D}^+ \rightarrow \mathrm{K}^- \, \pi^+ \, \pi^+$. It involves using two daughter tracks as "tag" to reconstruct the decay by applying on such objects topological, PID and track quality selections, and the third daughter track as a "probe", checking whether it fulfils either TPC+ITS or ITS-only selections. The D-meson yield is then extracted for the two set of requirements, and the ratio R of the yield corresponding to probes fulfilling TPC+ITS requirements over the yield corresponding to probes fulfilling ITS-only requirements is computed. This is done separately for data and MC and then the ratio $R_{\mathrm{data}}/R_{\mathrm{MC}}$ is calculated. The value thus obtained quantifies the difference in TPC-ITS matching in MC with respect to real data. It has been observed that there is a larger fraction of ITS tracks matched to TPC tracks in MC compared to data and that this difference is $\sim 3\%$, with no p_{T} dependence. Such value corresponds to a per-track systematic uncertainty due to a different ITS-TPC matching efficiency in MC with respect to real data.

Let $\alpha = 3\%$ be the per-track uncertainty provided by the study described above. The final state of the Ξ_c^0 decay corresponds to three pions and one proton, so four tracks in total. Therefore, for the Ξ_c^0 decay, the total systematic uncertainty associated to the ITS-TPC matching efficiency description in the MC will be $4\alpha = 12\%$, as the uncertainties are considered correlated. As α does not exhibit any p_{T} dependence, the corresponding Ξ_c^0 systematic will also be p_{T}-independent. This result is reported in Table 5.17.

$p_{\mathrm{T}\,\Xi_c^0}$ **[GeV/c]**	$1< p_{\mathrm{T}}<2$	$2< p_{\mathrm{T}}<3$	$3< p_{\mathrm{T}}<4$	$4< p_{\mathrm{T}}<5$	$5< p_{\mathrm{T}}<6$	$6< p_{\mathrm{T}}<8$	$8< p_{\mathrm{T}}<12$
Syst. unc.	12%	12%	12%	12%	12%	12%	12%

Table 5.17: Systematic uncertainties associated to ITS-TPC matching efficiency reproduction in MC.

This is only a first raw estimate of the impact of the potential differences in the ITS-TPC matching efficiency between data and MC and further studies are needed to properly address decays involving cascades. The D-meson daughters are long tracks that cross a large portion of the ITS, and consequently they are almost comparable to primary tracks. This description does not hold for the two pions and the proton coming from the cascade decay, as they have a much smaller portion lying within the ITS volume. As a consequence, the per-track systematic may be different for the pion $\pi \leftarrow \Xi_c^0$, that can be considered equivalent to the D-meson daughters, and for the Ξ daughters. This open point will be addressed in the near future by a dedicated centralised study.

5.10 Results

5.10.1 Prompt production cross section

In this thesis, the measurement of the Ξ_c^0 prompt production cross section in pp collisions at $\sqrt{s} = 13.6$ TeV with the ALICE detector is presented. The measurement is performed at midrapidity $|y| < 0.8$ and in the transverse momentum interval $1 < p_{\rm T} < 12$ GeV/c. The Ξ_c^0 and its corresponding anti-particle are reconstructed in the hadronic weak decay channel $\Xi_c^0 \to \Xi^- \, \pi^+ \to (\Lambda \, \pi^-) \, \pi^+ \to ((p\pi^-)\,\pi^-)\,\pi^+ + c.c.$ using a data sample that corresponds to an integrated luminosity of ~ 0.8 pb^{-1}. The $p_{\rm T}$-differential prompt cross section is computed as

$$\frac{\mathrm{d}^2\sigma^{\Xi_c^0}}{\mathrm{d}p_{\rm T}\mathrm{d}y} = \frac{1}{\mathrm{BR}} \cdot \frac{1}{2 \cdot \Delta y \cdot \Delta p_{\rm T}} \cdot \frac{N_{\Xi_c^0,\,\mathrm{raw}}(p_{\rm T}) \cdot f_{\mathrm{prompt}}(p_{\rm T})}{(A \cdot \epsilon)_{\mathrm{TOT,\,prompt}}(p_{\rm T})} \cdot \frac{1}{\mathcal{L}_{\mathrm{int}}}. \tag{5.19}$$

The $p_{\rm T}$-dependent raw yield including both the particle and anti-particle $N_{\Xi_c^0,\,\mathrm{raw}}$ is multiplied for the prompt fraction $f_{\mathrm{prompt}}(p_{\rm T})$ to retrieve the prompt raw yield. The thus obtained value is divided by the total prompt acceptance-times-efficiency $(A \cdot \epsilon)_{\mathrm{TOT,\,prompt}}(p_{\rm T})$ to correct for the limited detector geometrical acceptance and for the reconstruction, preselection and BDT efficiencies. For each $p_{\rm T}$ bin, the corrected prompt yield is scaled for the width of the transverse momentum interval $\Delta p_{\rm T}$. Under the assumption that the Ξ_c^0 rapidity distribution is uniform in the range $|y| < 0.8$, the result is normalised by the width of the rapidity interval $\Delta y = 1.6$ and therefore will be referred to one rapidity unit. Assuming identical yields for particles and anti-particles, a factor 2 is included in the denominator to account for the anti-particle yield. The integrated luminosity analysed $\mathcal{L}_{\mathrm{int}} = (0.793 \pm 0.079)$ pb^{-1}, as well as the total branching ratio, are also to be taken into account in the normalisation factors. The total branching ratio is computed as the product of the individual branching ratios corresponding to all the decays of the chain and is found to be $\mathrm{BR} = (0.92 \pm 0.17)\%$. The resulting production cross section is reported in Fig. 5.36. The vertical error bars and the boxes represent statistical and systematic uncertainties, respectively, with the statistical ones including the raw yields and $A \cdot \epsilon$ uncertainties. The branching ratio uncertainty is shown as a plain-shaded box, while the luminosity uncertainty as a pattern-shaded box. These conventions apply to all results reported in this Section. The data point corresponding to the transverse momentum range $4 < p_{\rm T} < 5$ GeV/c is affected by a downward fluctuation with respect to the trend of the other measurements. This is due to the statistical fluctuation already observed in Section 5.6 and discussed when evaluating the systematic uncertainties.

This result is compared to the prompt Ξ_c^0 cross section measured by ALICE at $\sqrt{s} = 13$ TeV [32] using the minimum bias proton-proton dataset collected in Run 2. The Run 3 data sample is almost thirty times larger than the previous one, that amounts to 30 nb^{-1}. The difference in collision energy

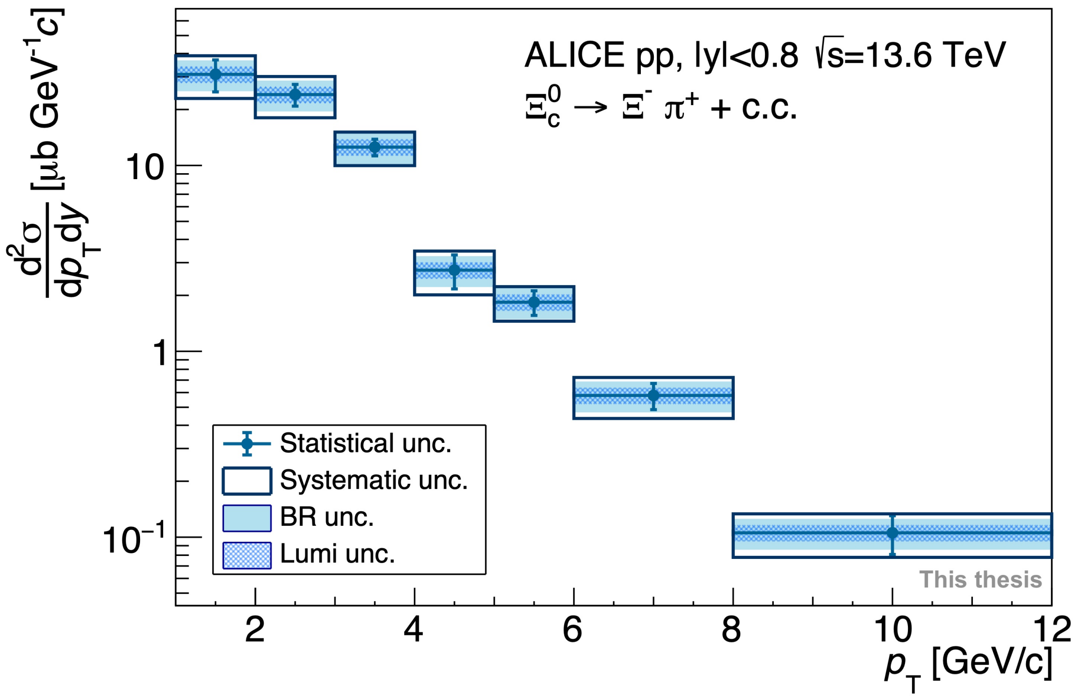


Figure 5.36: Prompt Ξ_c^0 production cross section.

is small enough not to induce significant differences, at least compared to the current uncertainties. If we consider for instance the FONLL prediction [56] for the D^0 meson, the differences in the $\frac{\mathrm{d}\sigma}{\mathrm{d}p_\mathrm{T}}$ for these two collision energies is of the order of a few percent, with larger deviations found for higher $p_{\mathrm{T\,D^0}}$ values, as it can be observed in Fig. 5.39. Given that the uncertainties on the two Ξ_c^0 cross section measurements are larger than such difference, they are expected to be in agreement. The comparison is reported in Fig. 5.37 and it shows that the measurement presented in this thesis underestimates the published ALICE result. In Fig. 5.38, the ratio of the new measurement over the Run 2 result is reported. Here the statistical, systematic and luminosity uncertainties are summed in quadrature separately for each cross section result, then the thus obtained values are propagated on the ratio. The discrepancy between the two measurements does not exhibit a p_T dependence and amounts to $\sim 40\%$. Slightly larger differences are observed for the p_T intervals $1 < p_\mathrm{T} < 2$ GeV/c and $4 < p_\mathrm{T} < 5$ GeV/c. This can be explained by considering that the low p_T region is known to be particularly delicate as the corresponding efficiency is extremely low, therefore even small differences between data and MC result in a considerable mismatch between the raw yield and the $A \cdot \epsilon$. As for the p_T bin $4 < p_\mathrm{T} < 5$ GeV/c, this is ascribed to the statistical fluctuation already discussed. To validate the analysis procedure used in this thesis, a closure test is performed. Further details about this procedure can be found in Section 5.10.2.

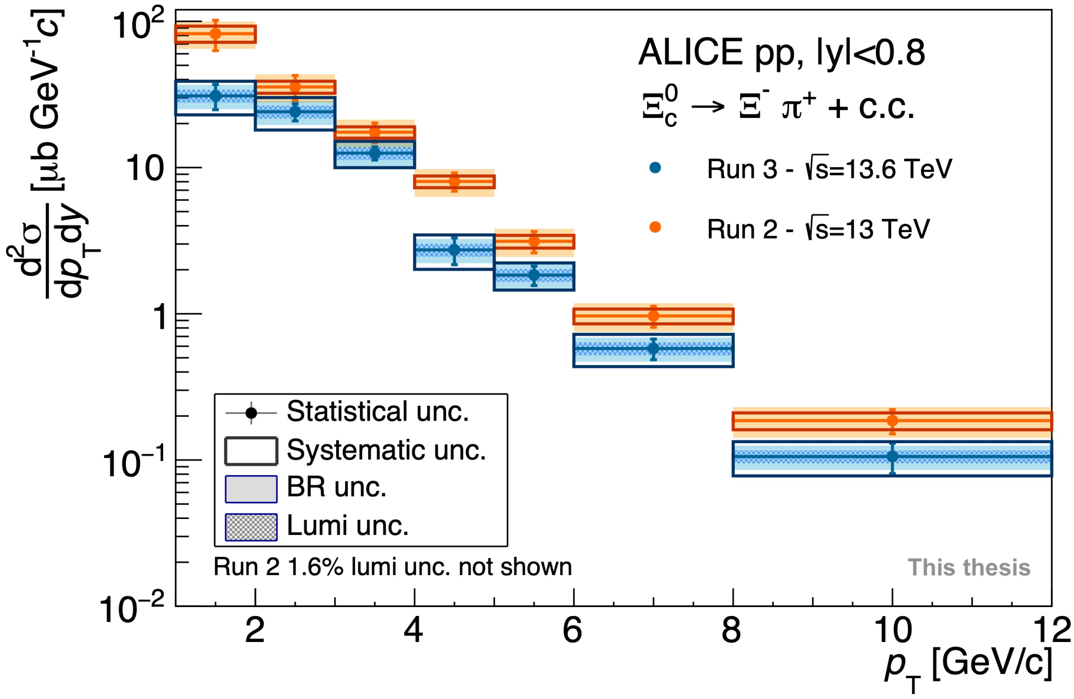


Figure 5.37: Comparison of the prompt Ξ_c^0 production cross section measured in this analysis with the corresponding ALICE measurement in pp collisions at $\sqrt{s} = 13$ TeV.

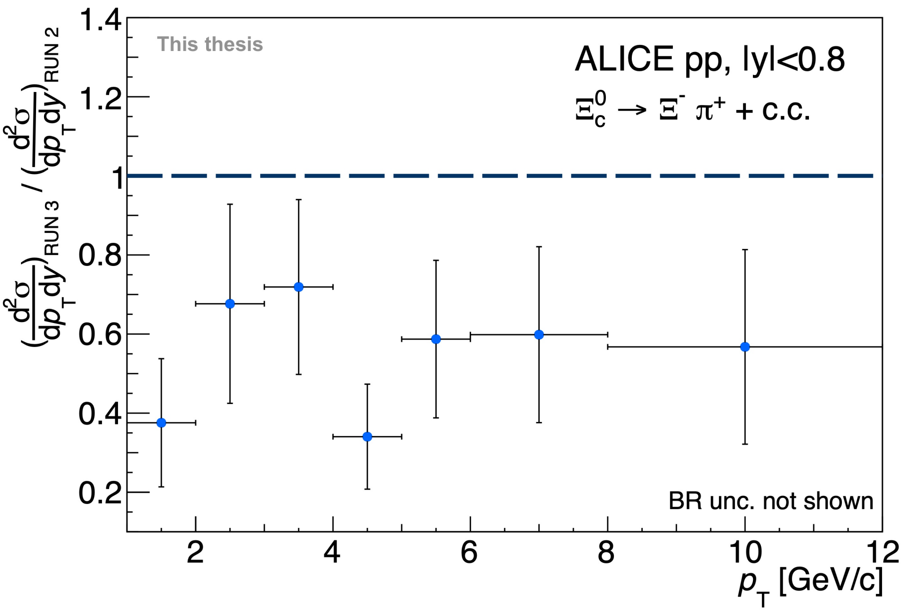


Figure 5.38: Ratio of the prompt Ξ_c^0 production cross section measured in this analysis over the corresponding ALICE measurement in pp collisions at $\sqrt{s} = 13$ TeV.

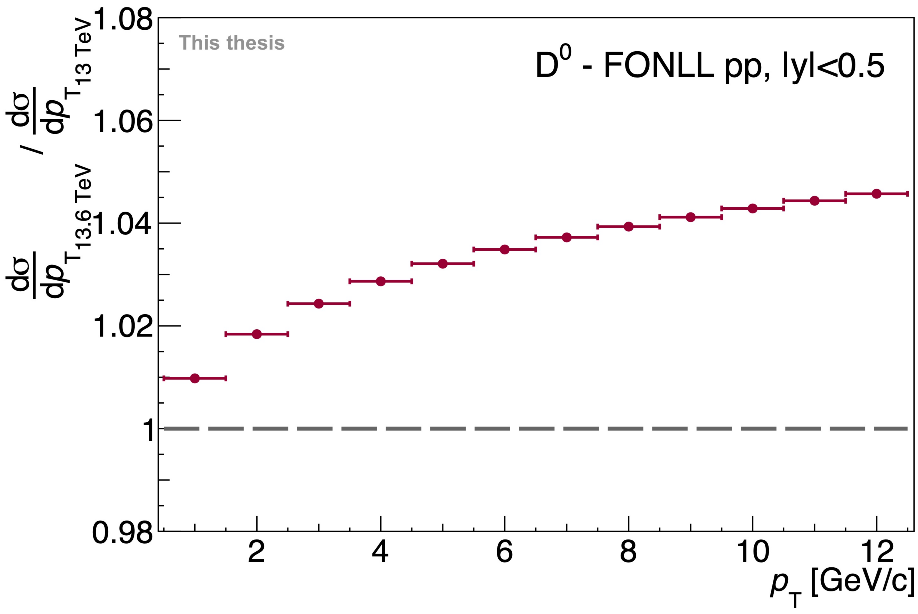


Figure 5.39: Ratio of FONLL predictions for D^0 production cross section at $\sqrt{s} = 13.6$ TeV and $\sqrt{s} = 13$ TeV as a function of the hadron transverse momentum. The FONLL uncertainties are not considered here as most of their sources cancel out.

5.10.2 Closure test

Given the large discrepancy observed with respect to the published measurement at a slightly lower collision energy, a closure test is performed on the MC in order to validate the analysis procedure. This method does not allow to check on the consistency of the estimation of f_{prompt}. However the non-prompt fraction is too small to be responsible for the observed discrepancy and affects the final measurement only to a minor extent. The purpose of the test is instead to validate the raw yield correction for the acceptance-times-efficiency. To do so, the same analysis procedure described in Sections 5.3, 5.4, 5.5, 5.6, 5.7 is followed. The dataset used is a subset of the MC, where all the background and prompt signal candidates are merged and treated as a real dataset. The same event selection, preselections and BDT selections implemented in the analysis are applied. Then, the "raw yield" is extracted by fitting the invariant-mass spectra, describing the signal and background distributions with a double Gaussian function and a polynomial of second order, respectively. Afterwards the obtained yields are corrected for the total prompt $A \cdot \epsilon$ extracted from the very same dataset. Finally, the corrected yields are compared to the number of prompt Ξ_c^0 generated within the preselected rapidity range $|y| < 0.8$ and decaying to the desired channel from the same MC simulation. If the analysis procedure is consistent, their ratio is compatible with 1. The charge-conjugate particles

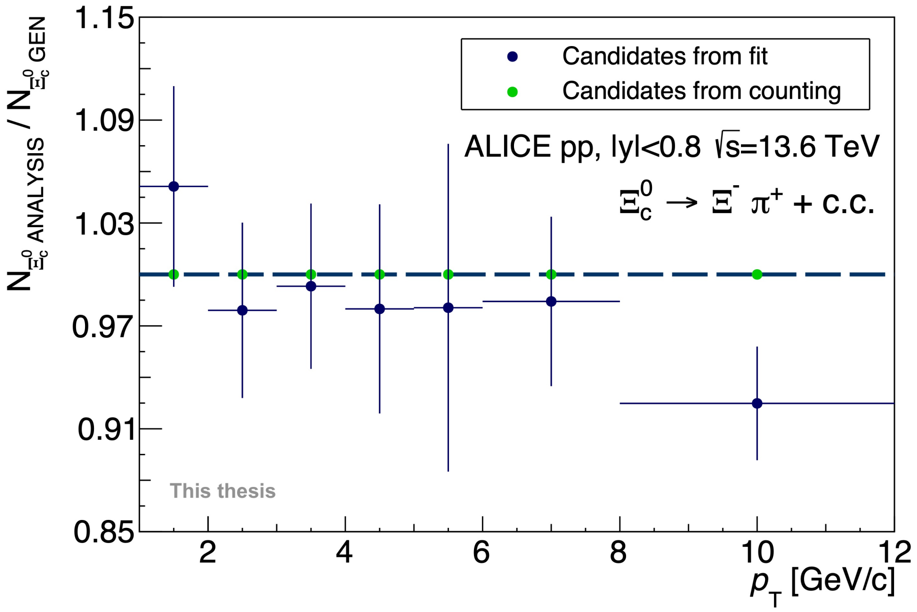


Figure 5.41: Ratio of the corrected yield extracted from the MC using the analysis procedure over the number of generated prompt Ξ_c^0.

An interesting observation that arises from the closure test concerns the signal shape in the fitting procedure. It turned out that modelling the signal shape with a double Gaussian function instead of a single-Gaussian results in a $\sim 10\%$ increase of the Ξ_c^0 yield, the reason being that the signal exhibits non-Gaussian tails. This result is reported in Fig. 5.42 and 5.43. The first plot shows the comparison between the raw yields extracted with the two different signal shapes, the second one reports their ratio. In order for the closure test to give a positive result, the double Gaussian function has to be used in the fits. The origin of the non-Gaussian tails is the invariant-mass resolution dependence on the p_{T} resolution of the daughters, which in turn depends on the daughter p_{T} itself. Consequently, each Ξ_c^0 candidate whose N daughters have a specific p_{T}, exhibits a specific invariant-mass resolution σ^*. Given that the different Ξ_c^0 invariant-mass spectra are filled depending on the charmed-baryon p_{T}, without making any distinction between the daughters p_{T} and, additionally, the $p_{\mathrm{T}\,\Xi_c^0}$ bins are relatively large, for each $p_{\mathrm{T}\,\Xi_c^0}$ bin the signal shape is given by a convolution of all the single Gaussian functions corresponding to each Ξ_c^0 entry, with each Gaussian shape having its specific width σ^*. The resulting non-Gaussian distribution of the signal is better described by a double Gaussian function with respect to a single Gaussian. This conclusion is supported by the reduced χ^2 values of the MC fits of the signal invariant-mass spectra reported in Appendix I (single Gaussian function) and O (double Gaussian function), the latter being systematically lower. This effect was not observed in Run 2 because of the dataset size: with a smaller sample available, the data points fluctuations are always included.

Taking into account only prompt signal allows to factorize out the calculation of the prompt fraction. In the case of a closure test performed on a charm- and/or beauty-enriched MC, such as the one used in this analysis, the FONLL prediction for the beauty production cross section used to estimate f_{prompt} does not reflect the heavy quark yields, therefore the theory-driven feed-down subtraction method can not be validated.

The results of the closure test are reported in Fig. 5.40 and 5.41. The first one shows the comparison between the number of generated prompt Ξ_c^0 and the corrected yield extracted from the MC using the analysis procedure, while the second one reports their ratio. The two yields are compatible within uncertainties, and so is their ratio with respect to unity. In Fig. 5.41, the purple points correspond to the ratio computed using as numerator the yield extracted by fitting the invariant-mass spectra, the green points are a result of an extra check. They correspond to the ratio computed using, as numerator, the counts of prompt signal candidates fulfilling all the selections and then corrected for the $A \cdot \epsilon$. Even if this is the result of a trivial calculation, it confirms once more that the analysis is healthy, as this ratio is exactly equal to 1 for all the p_{T} bins. Consequently, it is possible to conclude that the differences between the purple points and unity are simply due to the fit procedure.

The result of the closure test validated the analysis, therefore the discrepancy observed with respect to Run 2 can not be attributed to a mistake in the analysis code.

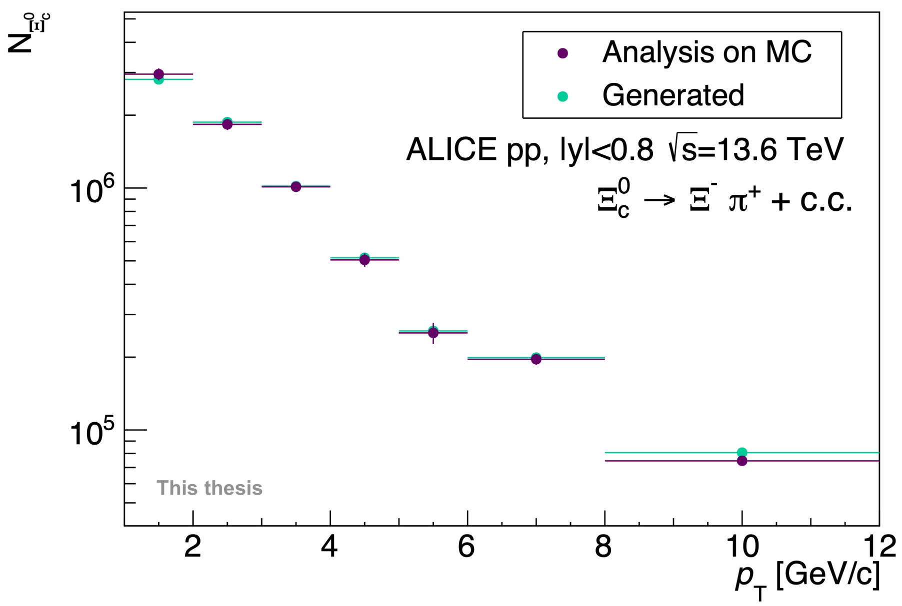


Figure 5.40: Comparison of the number of generated prompt Ξ_c^0 and the corrected yield extracted from the MC using the analysis procedure.

hid the non-Gaussian shape. As it is possible to see in Fig. 5.42, the uncertainty on the raw yield increases when the double Gaussian shape is used. This is due to the fact that the double Gaussian function has two additional parameters with respect to the simple Gaussian, and in the MC fits none of them is constrained.

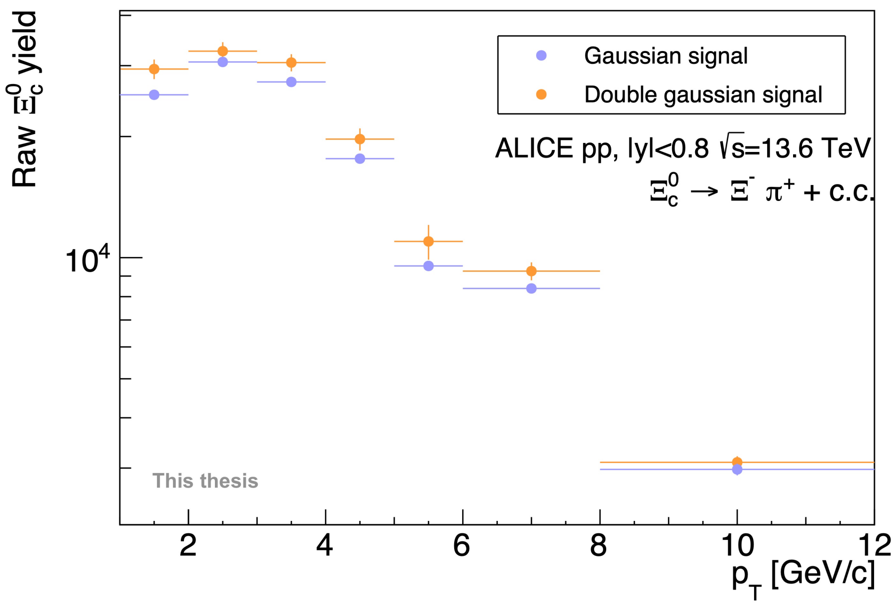


Figure 5.42: Comparison of the raw yields extracted from the MC using two different signal shapes, namely a double Gaussian function and a Gaussian shape.

The impact of the choice of signal shape in the fitting procedure to real data is estimated by performing new fits. This time, a double Gaussian function is used to model the signal. The fit parameters describing the width of the two Gaussian functions σ_1 and σ_2, as well as the fraction of the signal represented by the first Gaussian, are fixed to the corresponding values extracted by performing double Gaussian fits to MC signal. The results of the maximum-likelihood fitting procedure are reported in Appendix O, while Fig. 5.44 shows the comparison between the raw yields extracted with different signal shapes. In this case, the two raw yield sets are compatible within their uncertainties. This is due to the fact that the large combinatorial background does not allow the fit to be sensitive to the real signal shape, with the contribution of the long non-Gaussian tails being merged in the background. The reduced χ^2 values of the fits of the invariant-mass spectra from real data are also very similar for the two signal functions.

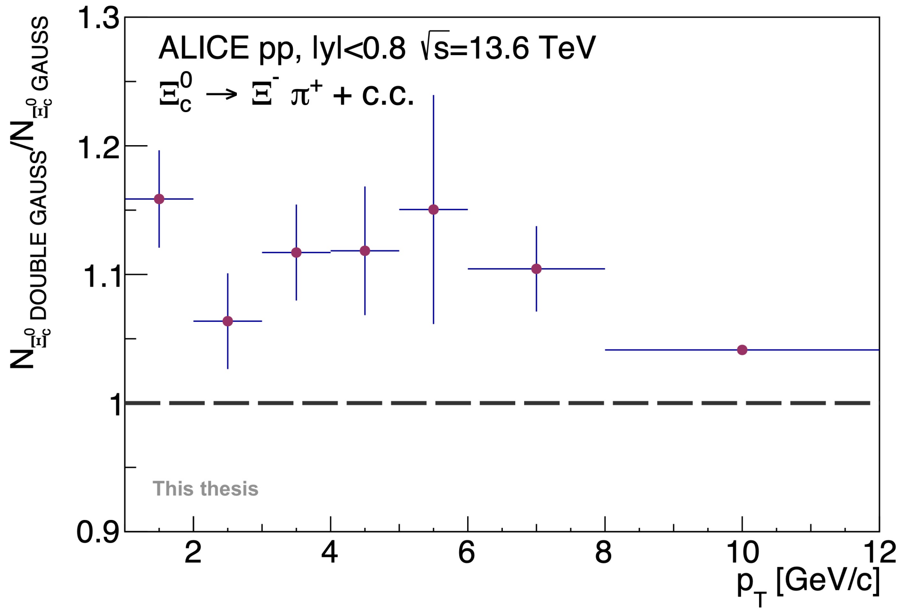


Figure 5.43: Ratio of the raw yields extracted from the MC using a double Gaussian function over the raw yield extracted with a Gaussian shape.

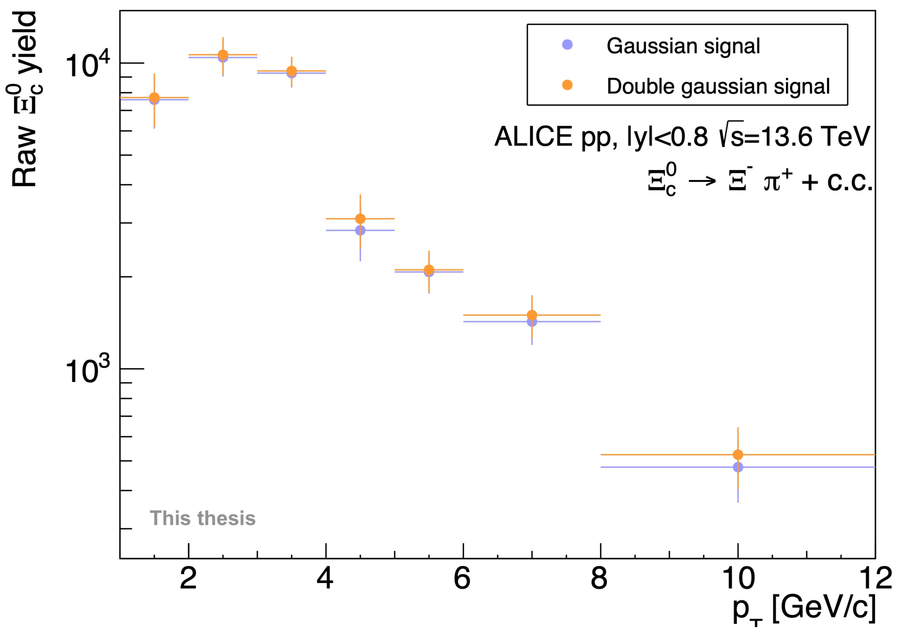


Figure 5.44: Comparison of the raw yields extracted from real data using two different signal shapes, namely a double Gaussian function and a Gaussian shape.

5.10.3 Comparison to theoretical predictions

The results for the prompt Ξ_c^0 production cross section are compared to theoretical predictions in Fig. 5.45. The orange data points correspond to the Run 2 result, the blue ones to the measurement presented in this thesis. The theoretical predictions from Pythia Monash (dark blue line), Pythia Color Reconnection Mode 2 (light blue line), QCM Model (green line), Statistical Hadronisation Model (SHM, pink band) and Catania Model (orange band) are considered. All these predictions have been computed for a collision energy $\sqrt{s} = 13$ TeV, but given the considerations reported above, the comparison can still be considered valid. Moreover, as for the time being, no calculation is available for the collision energy $\sqrt{s} = 13.6$ TeV. While the predictions by Pythia and QCM estimate directly the cross section, only predictions for the ratio $\frac{\Xi_c^0}{\mathrm{D}^0}$ have been produced according to the Catania and SHM models. Therefore, the corresponding results shown here are extracted by scaling the ratio theoretical prediction with a Tsallis fit of the D^0 cross section measured by ALICE in Run 2 at $\sqrt{s} = 13$ TeV [11]. To test the different hadronisation models, it would be necessary to compute the p_{T}-dependent ratio $\frac{\Xi_c^0}{\mathrm{D}^0}$. However the D-meson cross section measurement performed on Run 3 data is not yet finalized. Moreover, firstly the discrepancy observed for the Ξ_c^0 cross section with respect to the published result has to be understood. What can be concluded anyway is that at low p_{T} there seems to be a mechanism other than fragmentation contributing to the Ξ_c^0 production: the Pythia predictions, that only implement hadronisation via fragmentation, significantly underestimate both Run 2 and Run 3 measurements. The CR model gets closer to the data with respect to the Monash tune, and gives similar results to the SHM coupled with the RQM. The Catania prediction gives the best description of the data, followed by the QCM. Both this models implement hadronisation (also) via coalescence. On the other hand, at intermediate-high transverse momentum, it is more difficult to draw firm conclusions, especially given the disagreement between the two measurements.

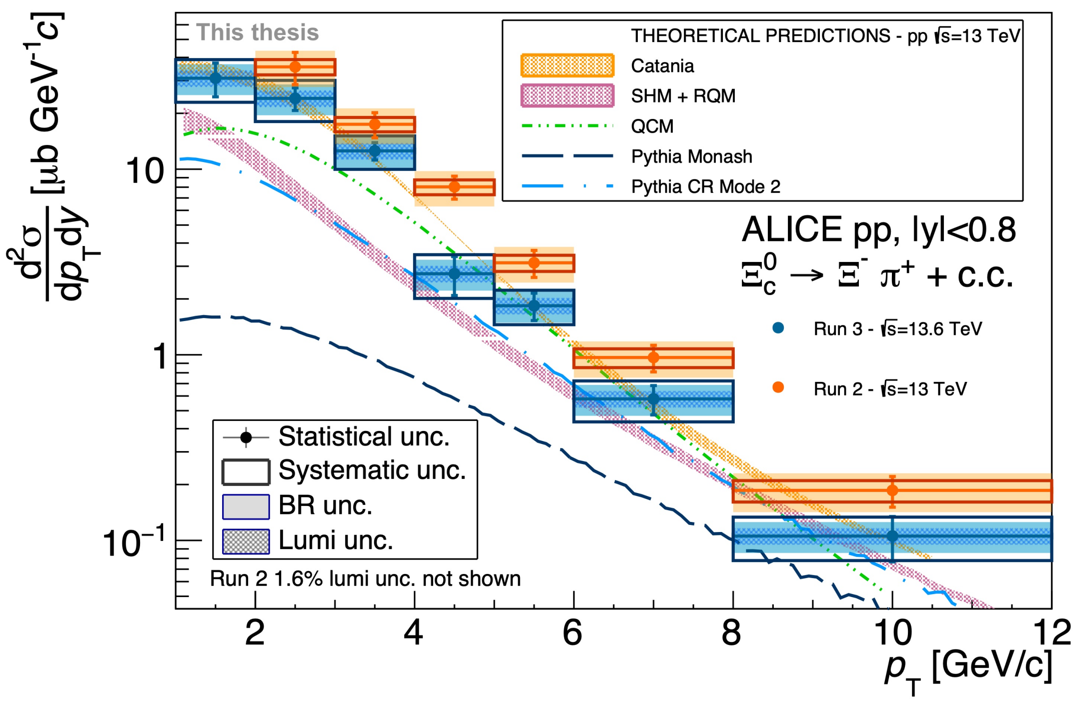


Figure 5.45: Comparison of Run 2 and Run 3 Ξ_c^0 prompt production cross section measurements to theoretical predictions.

Chapter 6

Conclusions and outlook

In this thesis, the measurement of the prompt Ξ_c^0 cross section in pp collisions at $\sqrt{s} = 13.6$ TeV with the ALICE detector is presented. The charmed-strange-baryon is reconstructed in the decay channel $\Xi_c^0 \rightarrow \Xi^- \pi^+ + \mathrm{c.c.}$ using minimum bias data corresponding to 0.8 pb^{-1}. The analysis is performed in the midrapidity interval $|y| < 0.8$ and in the transverse momentum region $1 < p_\mathrm{T} < 12$ GeV/c.
Measurements of charmed-baryon production in pp collisions provide essential tests of pQCD calculations and hadronisation models. Charmed-hadron production cross sections are also a crucial ingredient to extract the total $c\bar{c}$ cross section as well as charm fragmentation fractions. Besides this, they are fundamental to setup a benchmark for Pb–Pb collisions measurements.
The result presented is the first measurement of a fully corrected charmed-baryon spectrum performed on Run 3 data with ALICE. The data sample analysed is thirty times larger than the one used in Run 2, thus allowing to extract an integrated raw yield of more than 30k Ξ_c^0. This has been made possible by the major upgrades that the experiment underwent during the Long Shutdown 2, thus making it feasible to take data in continuous readout mode. The challenge of handling such large datasets is met by using an innovative dedicated software framework (O^2) and deploying a tailored analysis submission system (hyperloop), as well as by developing a new data format that is optimised for computing resource efficiency. Moreover, specific MC simulation methods have been developed to provide the best possible representation of the experimental conditions.
The measurement presented in this thesis underestimates the previous ALICE result extracted at $\sqrt{s} = 13$ TeV, with a $\sim 40\%$ p_T-independent discrepancy. Such deviation can not be due to the small difference in the collision energy. Moreover, as the cross section increases with $\sqrt{s}$, the effect of the difference in collision energy would contribute in the opposite direction, leading to an increase of the Run 3 result with respect to the Run 2 measurement. A closure test validated the analysis procedure, therefore the observed gap must be due to some other cause. Part of this discrepancy can be understood by considering that the $\Xi^\pm$ yield measured by ALICE using the same data sample also

underestimates the Run 2 measurement, with a $\sim 15\%$ difference that exhibits no p_{T} dependence. The closure test performed in this thesis proved that the Ξ_c^0 signal is characterised by non-Gaussian tails that can not be properly described when fitting real data invariant-mass spectra, thus leading to a $\sim 10\%$ decrease in the measured raw yield. A precise estimate of the integrated luminosity is still to be provided by the ALICE Physics Coordination and a correction of $\mathcal{L}_{\mathrm{int}}$ would actually result in an overall shift of the data points. A new reconstruction of the raw data used in this analysis might also allow for a better agreement between the two measurements, as improved TPC space-charge distortions calibrations will soon be delivered and hints of a residual ITS-TPC misalignment in the current data reconstruction have been observed. Recent $\Xi^- \to \Lambda\,\pi^-$ measurements performed on the dataset collected in 2024, for which a more refined reconstruction has been carried out, already show a smaller discrepancy with respect to Run 2 results. Further refinements could also be achieved by improving the MC production: currently, the smearing of both the DCA and p_{T} resolution are implemented only for primary tracks and for the daughters of heavy-flavour particles, but not for the cascade daughters. Extending the resolution smearing to deep secondary tracks could help to obtain a more realistic description of the data. This analysis would also benefit from a larger MC production, as the lack of Ξ_c^0 candidates prevented the acceptance-times-efficiency correction based on a dataset different from the one used for the boosted decision tree training. All these ingredients could help to improve the agreement between the current measurement and the published one, however the main assumption is that such discrepancy could be due to a misrepresentation of the cascade-finding efficiency. The reason for this consideration is that all the production measurements involving cascades and/or V0s underestimate Run 2 results. This is true not only for direct measurements of these particles, but also for other heavy-flavour analyses measuring charmed-baryons decaying to either $\Xi^{\pm}$ or $\Omega^{\pm}$. Such HF measurements exhibit discrepancies with respect to Run 2 results similar to the one observed in this analysis. On the other hand, measurements of primary particles yields, such as pion, proton and kaon spectra, are in agreement with the published results. Detailed studies addressing the performance of the cascades reconstruction algorithm, and especially its reproduction in the MC simulations, are still incomplete. The replication of the cascades reconstruction strategy described in 3.3.2 is challenging, and it is not yet excluded that the fraction of afterburned tracks in the simulation does not match what is found in real data. As the cascades finding and their association to collisions take place when raw data are processed to produce the AO2D files, which are the analysis input, it is not possible to change the algorithm settings at analysis level. Another potential issue is related to the cascade-to-collision association and, again, its reproduction in the MC. The cascade-finding algorithm mentioned before associates each cascade to a single collision, in particular to the one whose primary vertex results in the highest value of the cosine of pointing angle. As the Ξ_c^0 decay vertex is very close to the collision primary vertex due to the extremely short lifetime of the charmed-baryon, this criterion should still be suitable to handle the secondary Ξ. However, the continuous data taking operation mode complicates the association of tracks to collisions, and it could be that the MC is actually too ideal and does not reproduce the loss of charmed-baryon reconstruction efficiency due to a wrong cascade-to-collision association. What could happen is that both the cascade and the pion coming from the Ξ_c^0 decay are actually associated to the same collision, but not to the correct one. Therefore, it could be that topological selections involving the position of the primary vertex discard the corresponding Ξ_c^0. Analysers do not have access to this information and this effect is still to be quantified. Detailed studies addressing this potential issues are a top priority and will soon be carried out within the ALICE collaboration.

Once the observed discrepancy is fully understood, this measurement could be performed on a much larger minimum bias dataset amounting to an integrated luminosity of $\sim 10\ \mathrm{pb}^{-1}$. Such an increased-size data sample could help to heal the fluctuation observed in this analysis for the transverse momentum bin $4 < p_{\mathrm{T}\,\Xi_c^0} < 5$ GeV/c. Moreover, it will allow to increase the granularity of the measurement as well as to extend the transverse momentum reach, both at high and at low p_{T}. In particular, the low transverse momentum region is crucial for hadronisation studies, as the coalescence mechanism is expected to become more and more dominant as the p_{T} decreases. The new cross section measurement will be fundamental to compute the total $c\bar{c}$ production cross section at the top energy $\sqrt{s} = 13.6$ TeV and it will be used to extract the charm fragmentation fractions, too. Additionally, the availability of such a large dataset will give access to the measurement of the non-prompt Ξ_c^0 production cross section at midrapidity.

The development of software triggers that have been implemented within this thesis work will allow to observe Cabibbo-suppressed decays of the Ω_c^0 baryon, namely $\Omega_c^0 \rightarrow \Xi^- \pi^+ + \mathrm{c.c}$ (singly-Cabibbo-suppressed) and $\Omega_c^0 \rightarrow \Xi^- \mathrm{K}^+ + \mathrm{c.c}$ (doubly-Cabibbo-suppressed). The ALICE software triggers have already processed an integrated luminosity of 82 pb^{-1}, and an additional sample corresponding to more than 100 pb^{-1} is expected to be analysed by the end of Run 3. Such measurements will be used to evaluate Ω_c^0 branching ratio fractions in the coming years. Presently, only theoretical predictions are available for the Ω_c^0 absolute branching ratios. While the effect of Cabibbo suppression is clearly identifiable with the CKM matrix elements, the effects of the kinematics of the decay, including spin-spin rearrangement processes, are highly challenging to be computed. They also affect the decay amplitude and calculating the form factors aims at describing the contribution of such processes, allowing to account for the fact that the charm quark does not decay freely but is surrounded by other quarks that map into a final state hadron. Measuring the Ω_c^0 yields in different decay channels will allow to extract relative branching ratios, providing a tool to validate theoretical models, shed light on tension between existing measurements from different experiments, and overcome the lack of experimental results for the doubly-Cabibbo-suppressed channel.

Appendix A

Variables description

More details about variables definitions and conventions used in this thesis are provided here:

- the collision primary vertex is referred to as PV
- the XY direction identifies the transverse plane, i.e. the plane perpendicular to the beam axis
- the pointing angle is defined as the angle between a particle's reconstructed momentum and the direction connecting its production point and decay vertex
- DCA stands for distance of closest approach, it can be either referred to a track and the PV or to a pair of tracks. In the first case it is the track impact parameter, in the latter the minimum distance between the tracks
- the variable $n\sigma_{\mathrm{TPC}}$ represents the distance between the $\mathrm{d}E/\mathrm{d}x$ signal associated to a track and the expected value for a certain PID hypothesis in units of σ, i.e. resolution
- R identifies the radial distance on the XY plane of a particle decay vertex with respect to the collision PV
- η is the particle pseudorapidity
- PDG stands for Particle Data Group [4], the review of particle physics results that lists the most updated and precise values available

Variable name	**Description**
InvMassCharmBaryon	Ξ_c^0 invariant mass
PtCharmBaryon	Ξ_c^0 transverse momentum
DecLenCascade	Ξ decay length
DecLenV0	Λ decay length
PtCasc	Ξ transverse momentum
PtPiFromCharmBaryon	$\pi \leftarrow \Xi_c^0$ transverse momentum
ImpactParCascXY	Ξ DCA to the PV in the XY direction
ImpactParPiFromCharmBaryon	$\pi \leftarrow \Xi_c^0$ DCA to the PV in the XY direction
CosPACasc	Ξ cosine of pointing angle
CosPAV0	Λ cosine of pointing angle
DcaCascDau	DCA between Ξ daughter tracks
DcaV0Dau	DCA between Λ daughter tracks
DcacharmBaryonDau	DCA between Ξ_c^0 daughter tracks
TpcNSigmaPrFromLambda	$p \leftarrow \Lambda\ n\sigma_{\mathrm{TPC}}$
TpcNSigmaPiFromLambda	$\pi \leftarrow \Lambda\ n\sigma_{\mathrm{TPC}}$
TpcNSigmaPiFromCasc	$\pi \leftarrow \Xi\ n\sigma_{\mathrm{TPC}}$
TpcNSigmaPiFromCharmBaryon	$\pi \leftarrow \Xi_c^0\ n\sigma_{\mathrm{TPC}}$
InvMassLambda	Ξ invariant mass
InvMassCascade	Λ invariant mass

Table A.1: Features corresponding to a Ξ_c^0 candidate.

Appendix B

Feature distributions

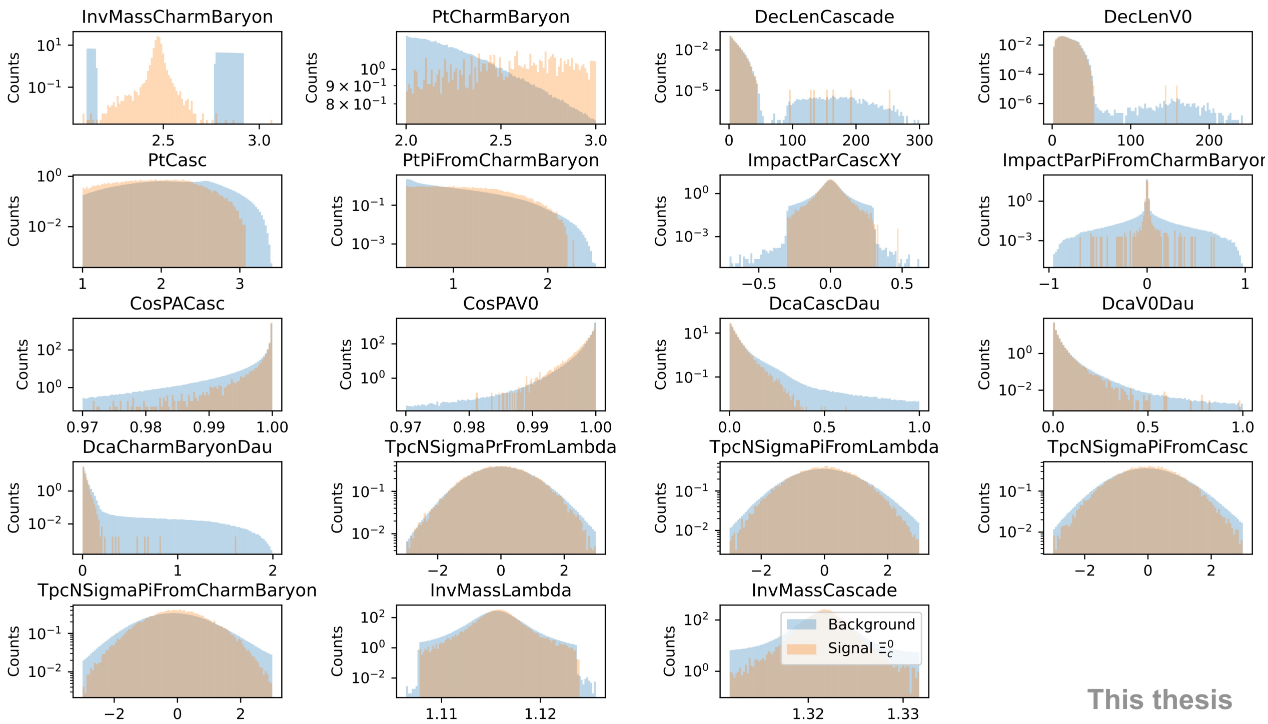


Figure B.1: Feature normalised distributions for $2 < p_{\mathrm{T}\,\Xi_c^0} < 3$ GeV/c in data and MC.

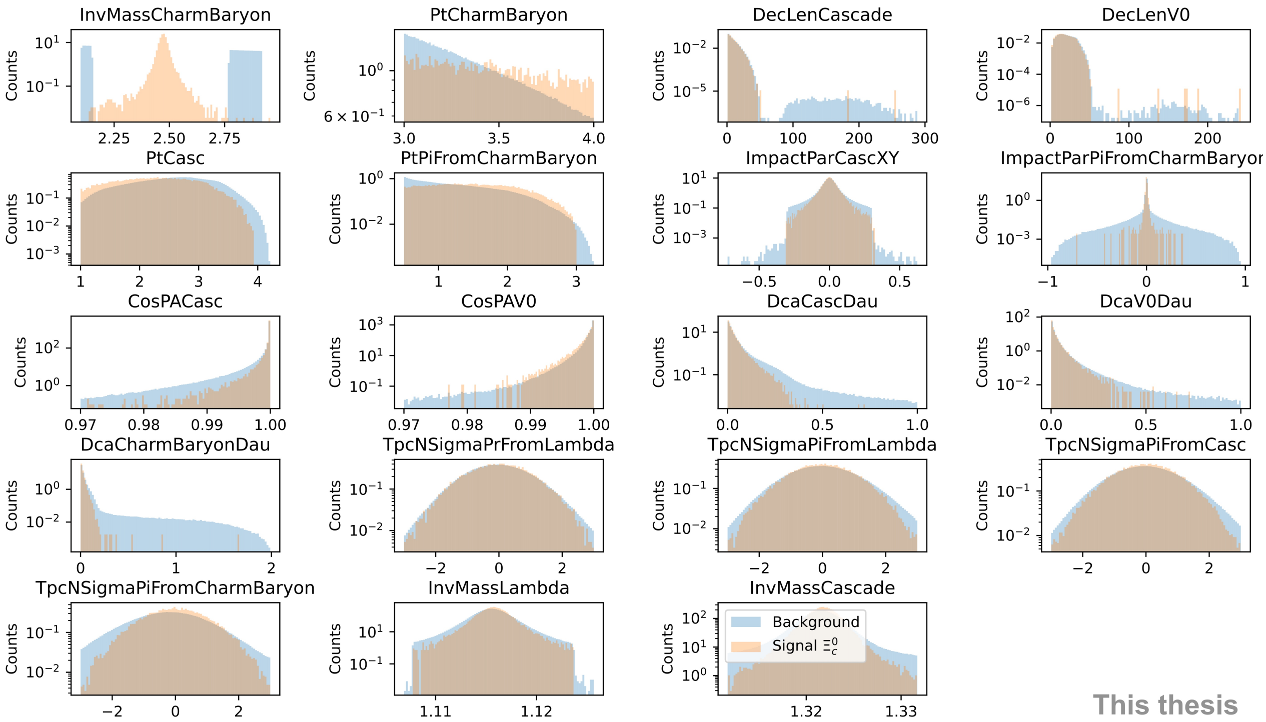


Figure B.2: Feature normalised distributions for $3 < p_{\mathrm{T}\,\Xi_c^0} < 4$ GeV/c in data and MC.

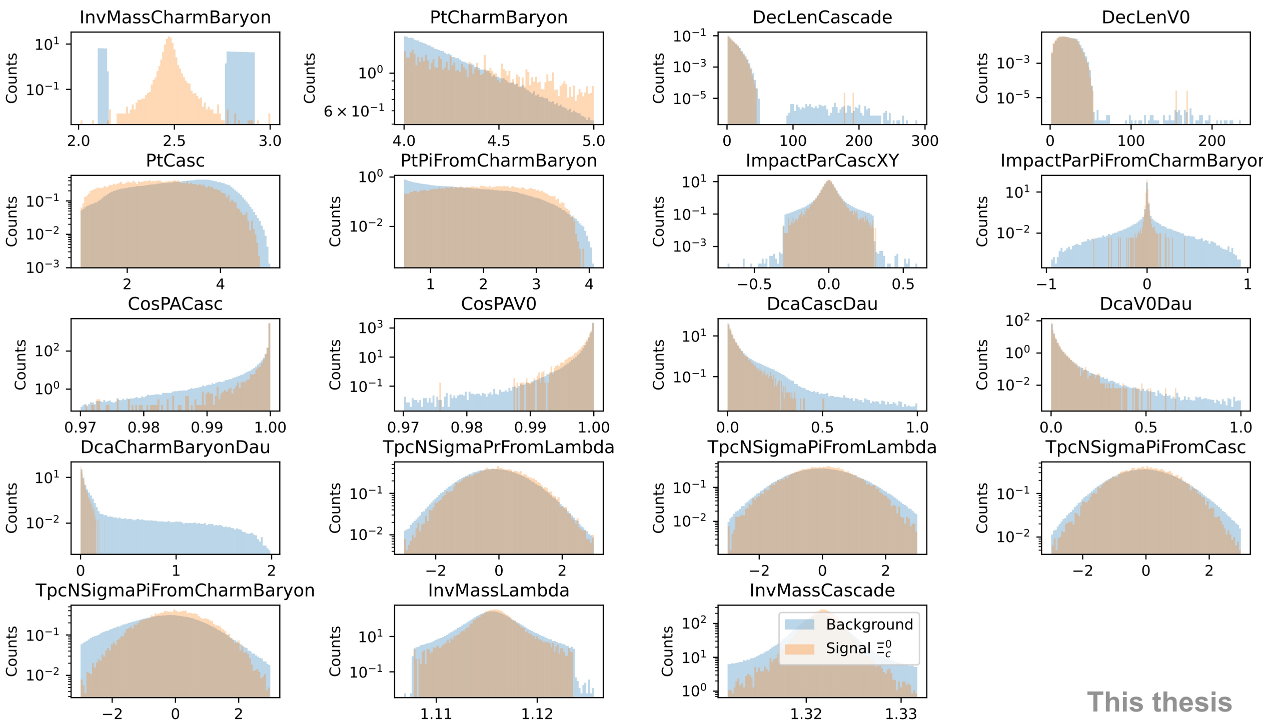


Figure B.3: Feature normalised distributions for $4 < p_{\mathrm{T}\,\Xi_c^0} < 5$ GeV/c in data and MC.

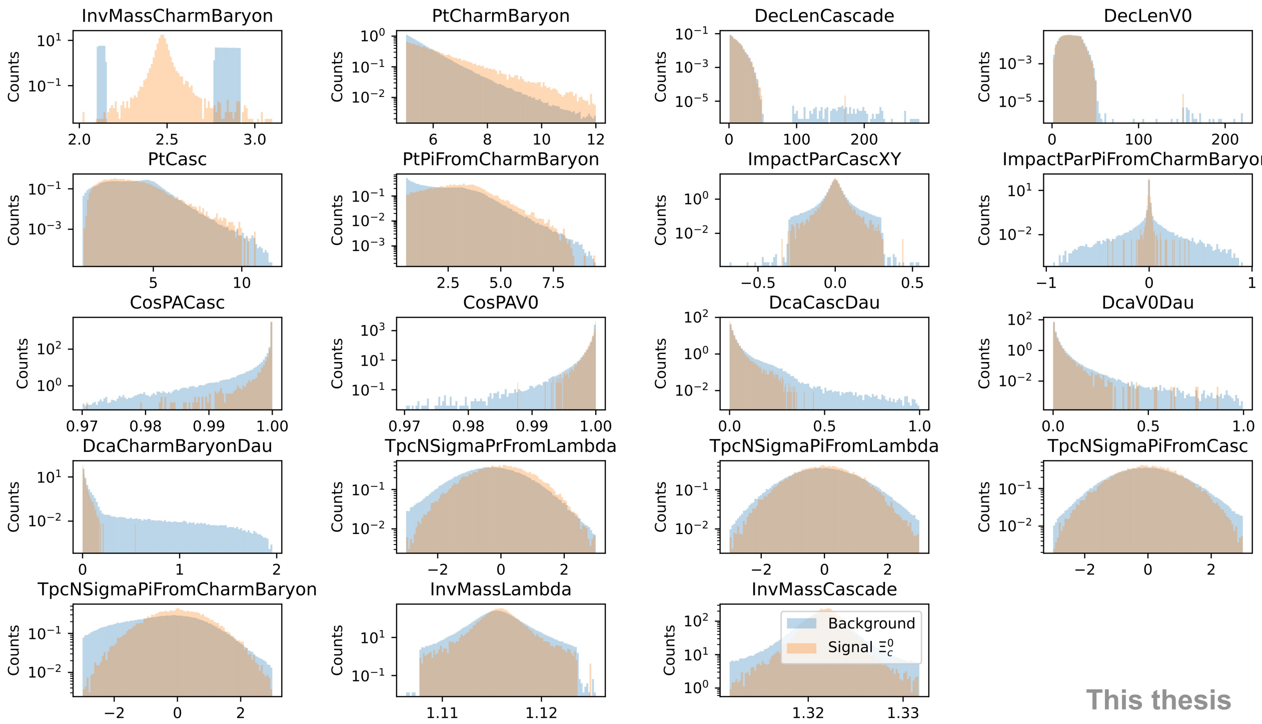


Figure B.4: Feature normalised distributions for $5 < p_{\mathrm{T}\,\Xi_c^0} < 12$ GeV/c in data and MC.

Appendix C

Feature correlation matrices for the signal class

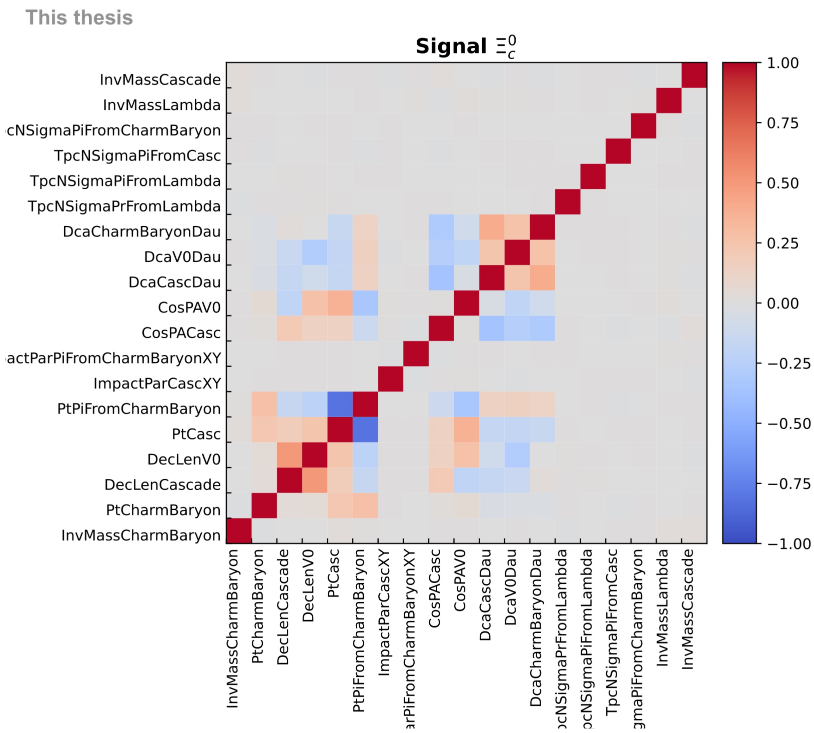


Figure C.1: Signal class feature correlation matrix for $2 < p_{\mathrm{T}\,\Xi_c^0} < 3$ GeV/c.

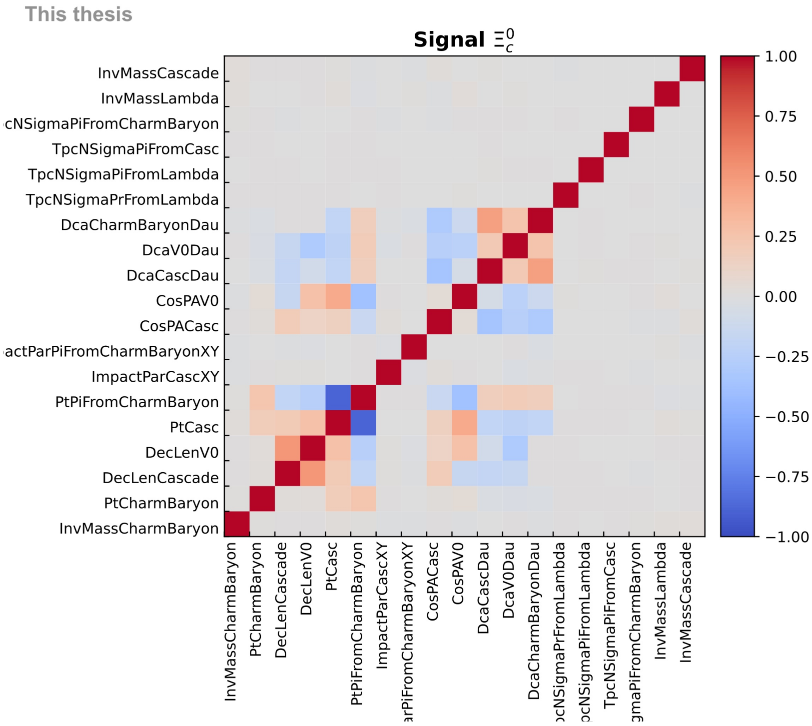


Figure C.2: Signal class feature correlation matrix for $3 < p_{\mathrm{T}\,\Xi_c^0} < 4$ GeV/c.

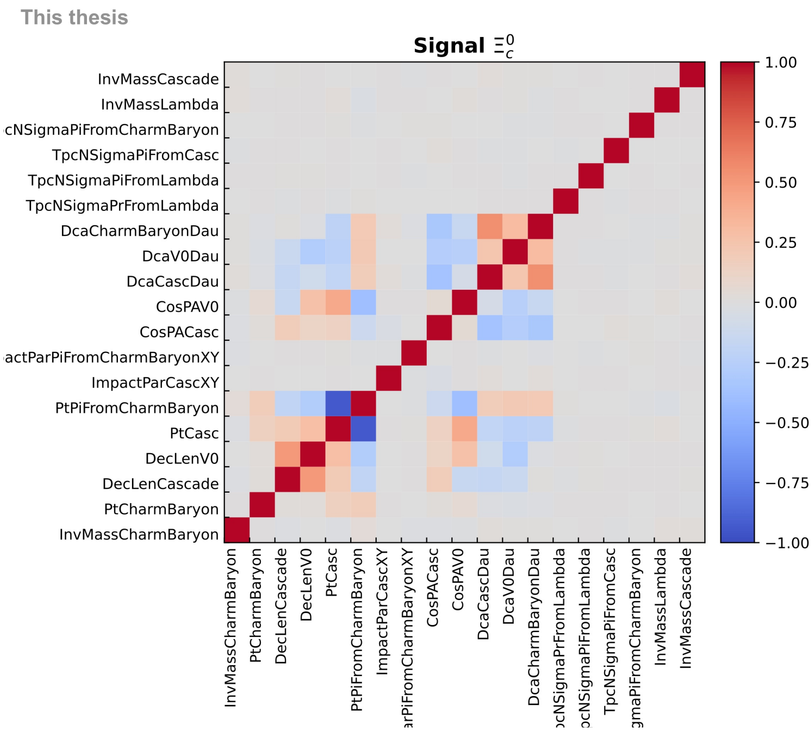


Figure C.3: Signal class feature correlation matrix for $4 < p_{\mathrm{T}\,\Xi_c^0} < 5$ GeV/c.

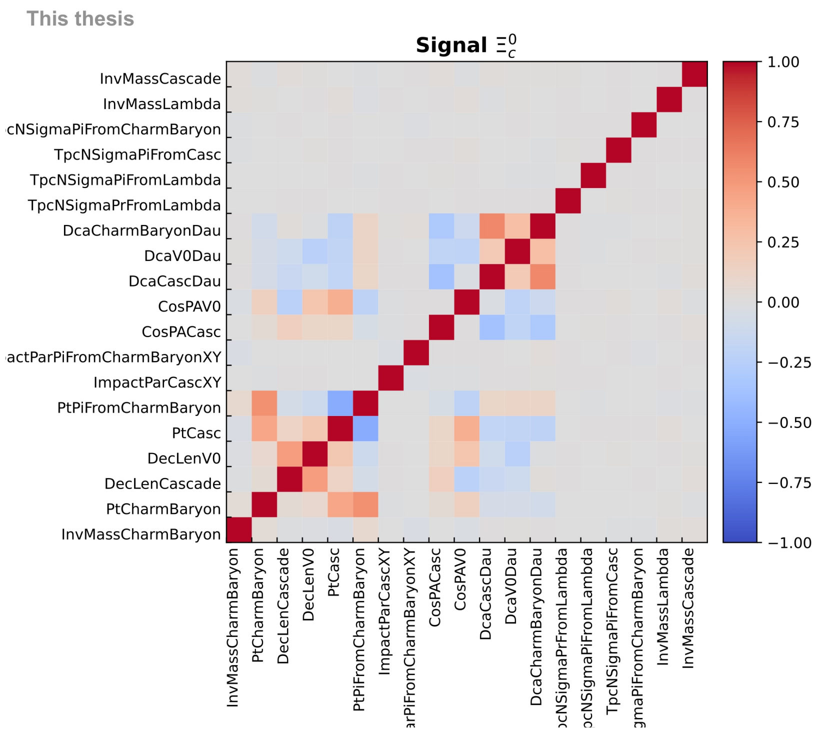
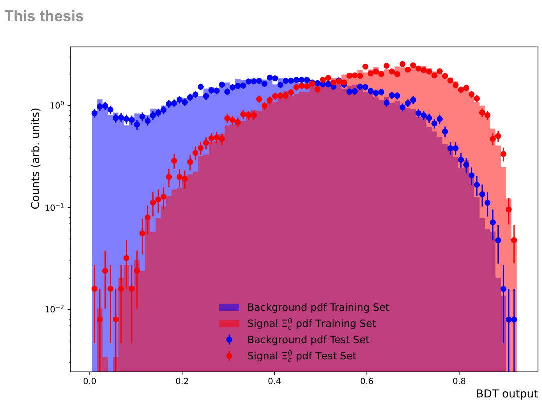


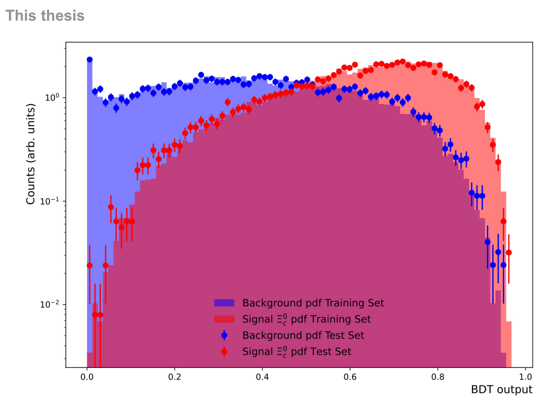
Figure C.4: Signal class feature correlation matrix for $5 < p_{\mathrm{T}\,\Xi_c^0} < 12$ GeV/c.

Appendix D

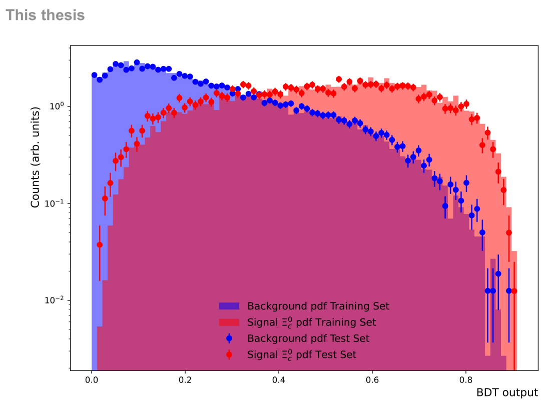
BDT score distributions



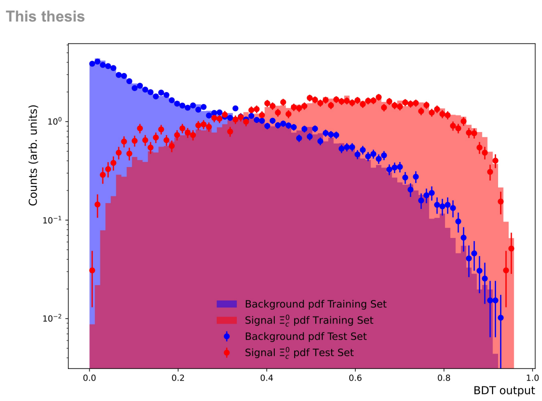
(a) $2 < p_{\mathrm{T}\,\Xi_c^0} < 3$ GeV/c



(b) $3 < p_{\mathrm{T}\,\Xi_c^0} < 4$ GeV/c



(c) $4 < p_{\mathrm{T}\,\Xi_c^0} < 5$ GeV/c



(d) $5 < p_{\mathrm{T}\,\Xi_c^0} < 12$ GeV/c

Figure D.1: BDT score distributions.

Appendix E

Training features importance

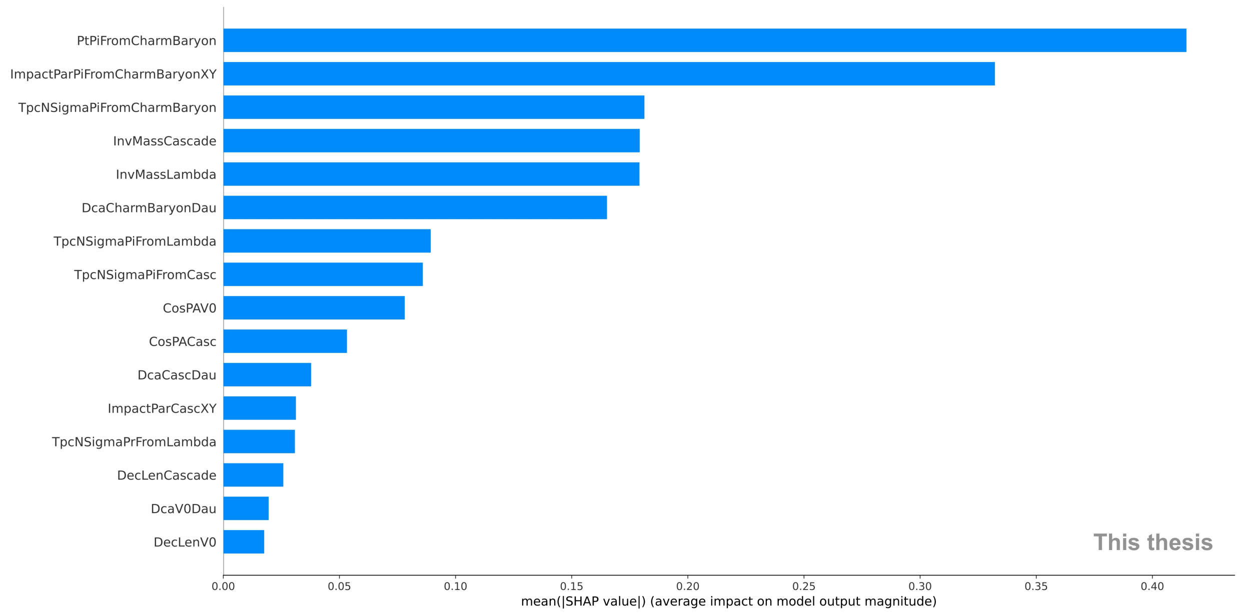


Figure E.1: Ranking of training features based on their impact on the classification for the BDT model corresponding to $2 < p_{\mathrm{T}\,\Xi_c^0} < 3$ GeV/c.

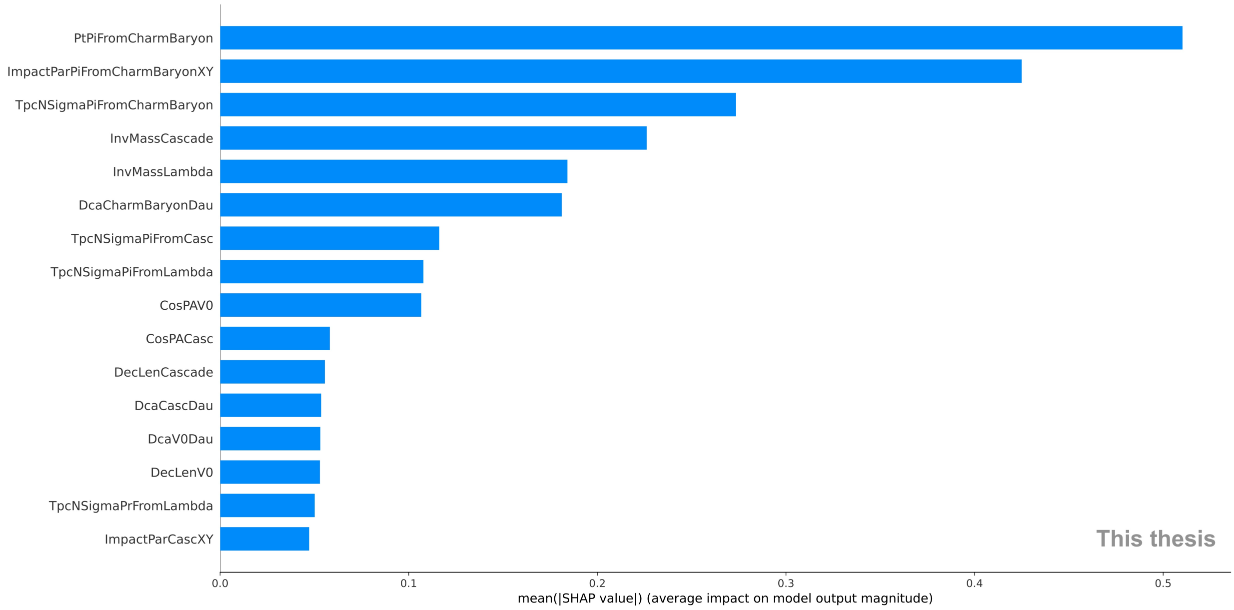


Figure E.2: Ranking of training features based on their impact on the classification for the BDT model corresponding to $3 < p_{\mathrm{T}\,\Xi_c^0} < 4$ GeV/c.

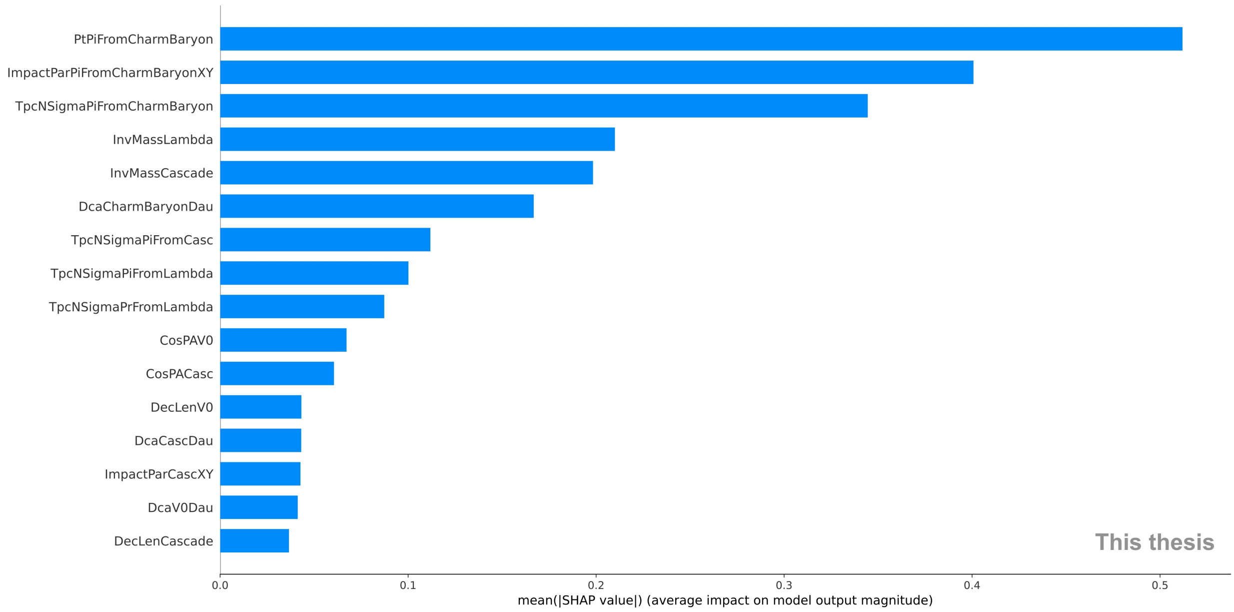


Figure E.3: Ranking of training features based on their impact on the classification for the BDT model corresponding to $4 < p_{\mathrm{T}\,\Xi_c^0} < 5$ GeV/c.

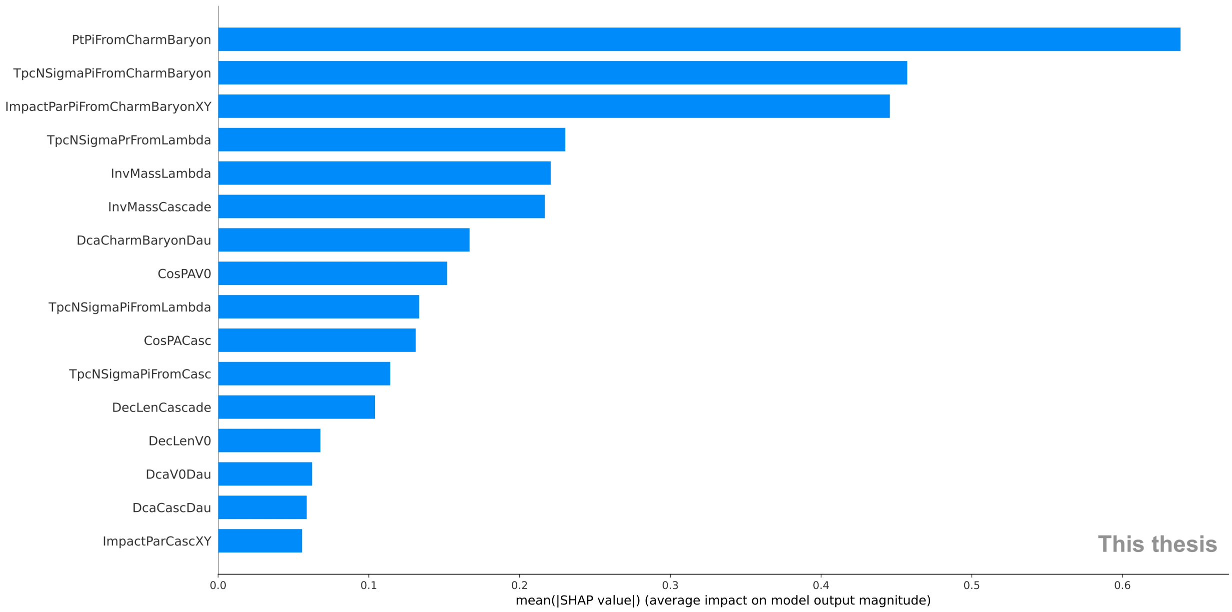


Figure E.4: Ranking of training features based on their impact on the classification for the BDT model corresponding to $5 < p_{\mathrm{T}\,\Xi_c^0} < 12$ GeV/c.

Appendix F

ROC curves

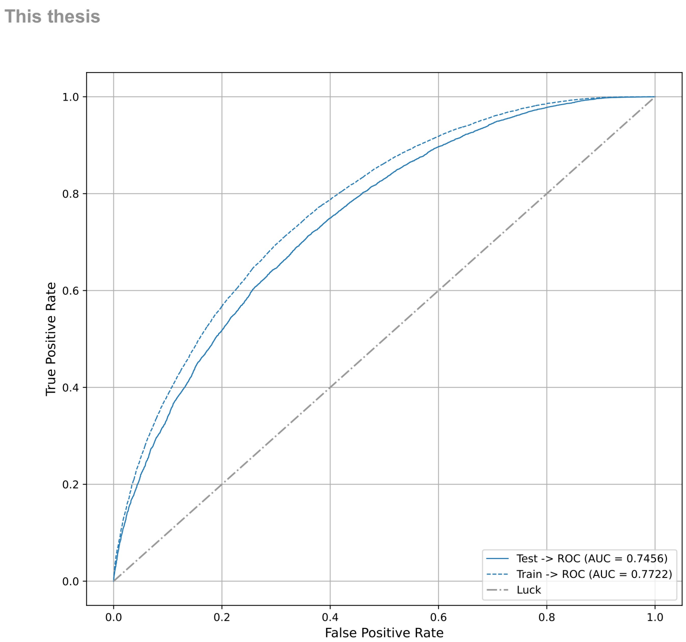


Figure F.1: ROC curve of the BDT model corresponding to the p_{T} interval $2 < p_{\mathrm{T}\,\Xi_c^0} < 3$ GeV/c.

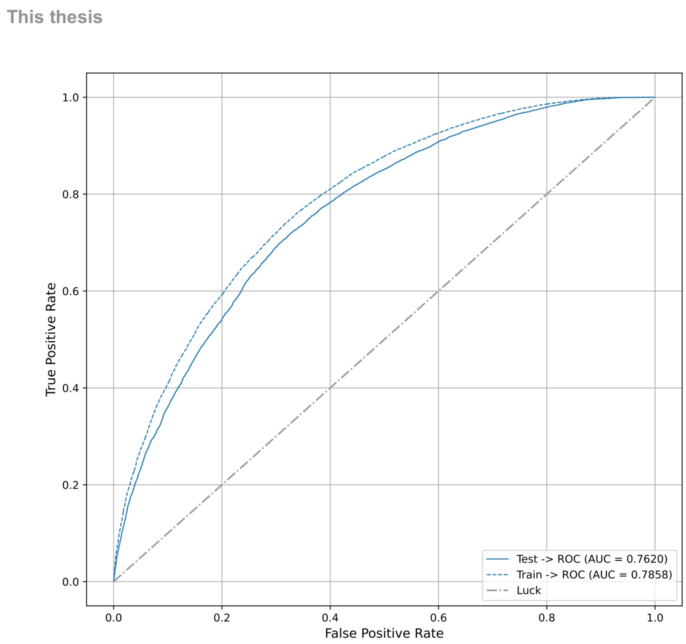


Figure F.2: ROC curve of the BDT model corresponding to the p_{T} interval $3 < p_{\mathrm{T}\,\Xi_c^0} < 4$ GeV/c.

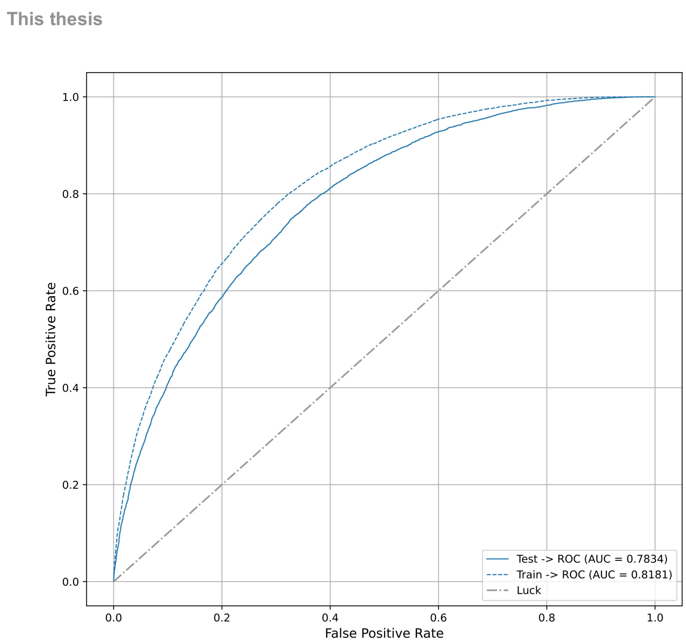


Figure F.3: ROC curve of the BDT model corresponding to the p_{T} interval $4 < p_{\mathrm{T}\,\Xi_c^0} < 5$ GeV/c.

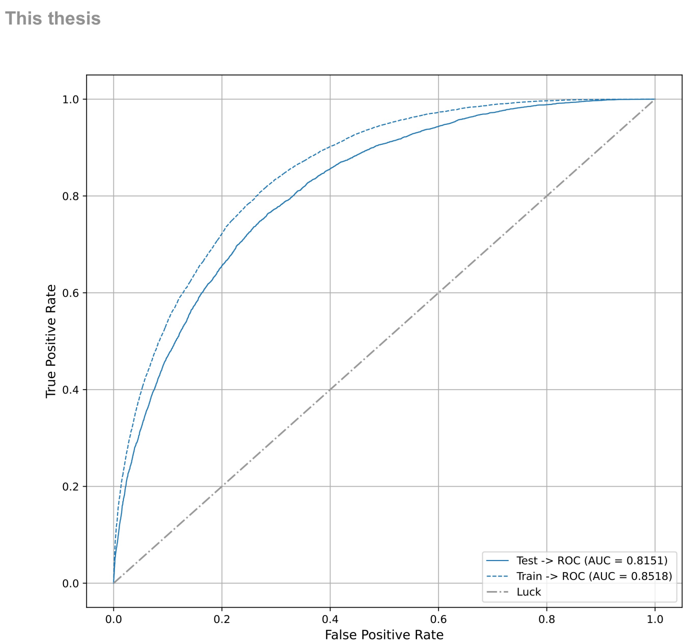


Figure F.4: ROC curve of the BDT model corresponding to the p_{T} interval $5 < p_{\mathrm{T}\,\Xi_c^0} < 12$ GeV/c.

Appendix G

Learning curves

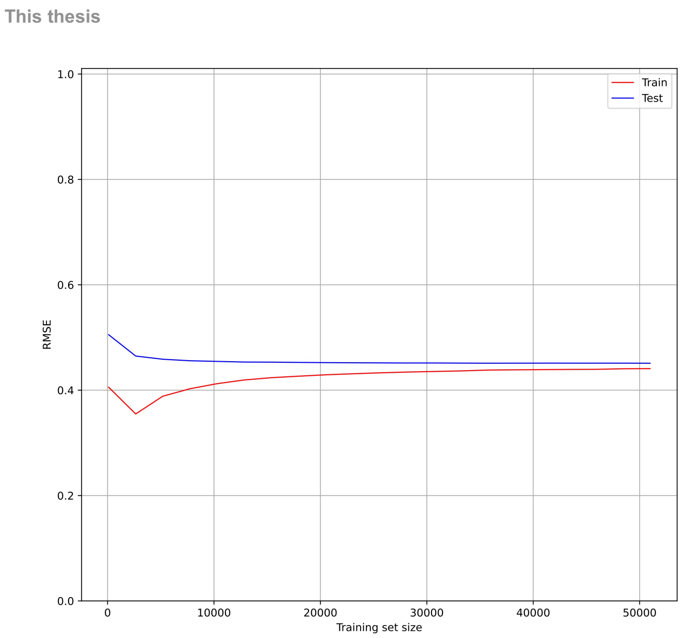


Figure G.1: Learning curves of the BDT model corresponding to $2 < p_{\mathrm{T}\,\Xi_c^0} < 3$ GeV/c.

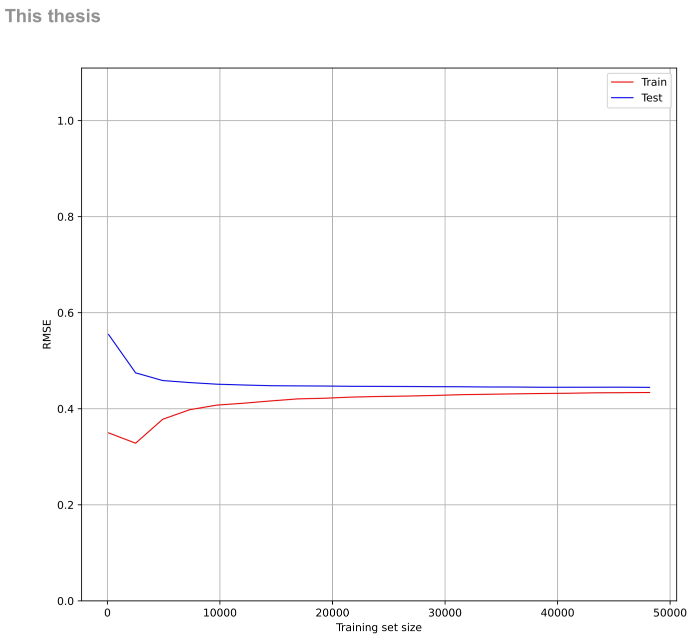


Figure G.2: Learning curves of the BDT model corresponding to $3 < p_{\mathrm{T}\,\Xi_c^0} < 4$ GeV/c.

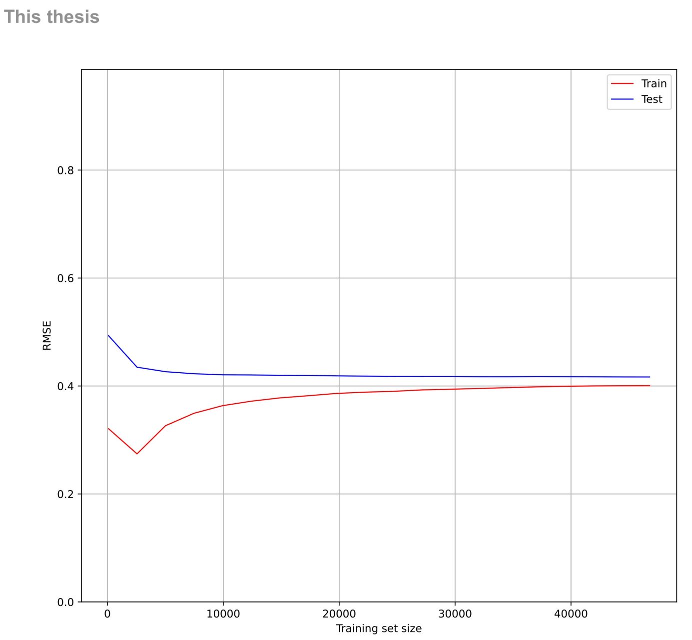


Figure G.3: Learning curves of the BDT model corresponding to $4 < p_{\mathrm{T}\,\Xi_c^0} < 5$ GeV/c.

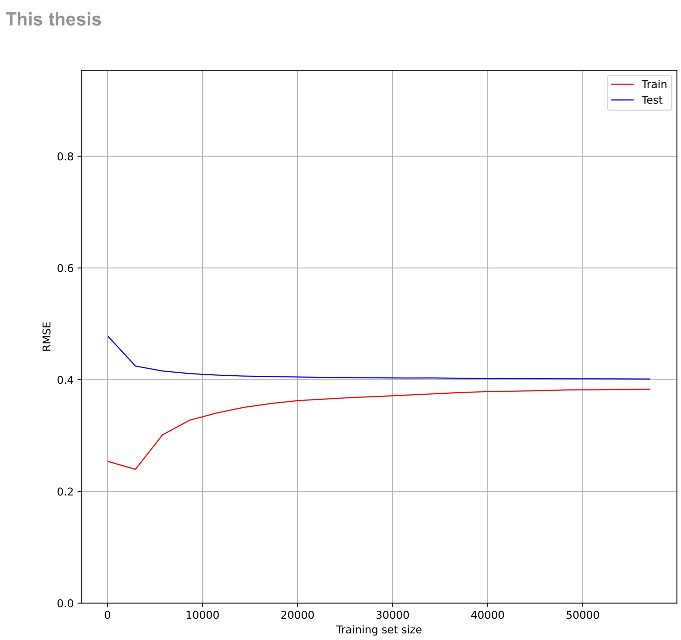


Figure G.4: Learning curves of the BDT model corresponding to $5 < p_{\mathrm{T}\ \Xi_c^0} < 12$ GeV/c.

Appendix H

BDT working point

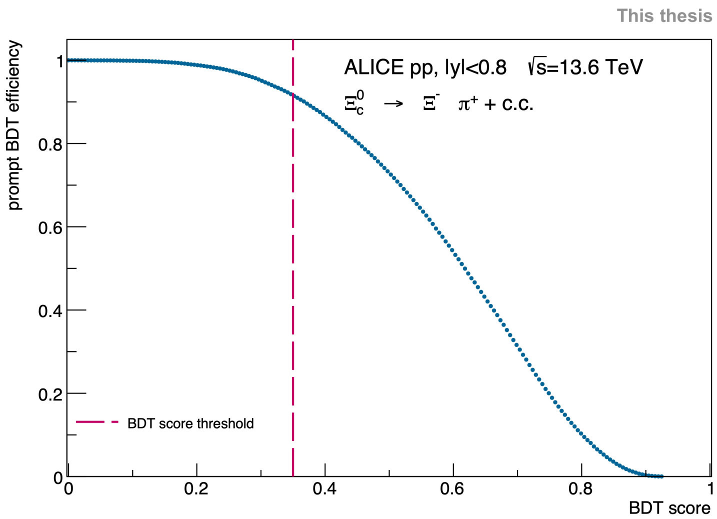


Figure H.1: Prompt BDT efficiency for the transverse momentum interval $2 < p_{\mathrm{T}\,\Xi_c^0} < 3$ GeV/c as a function of the BDT score threshold.

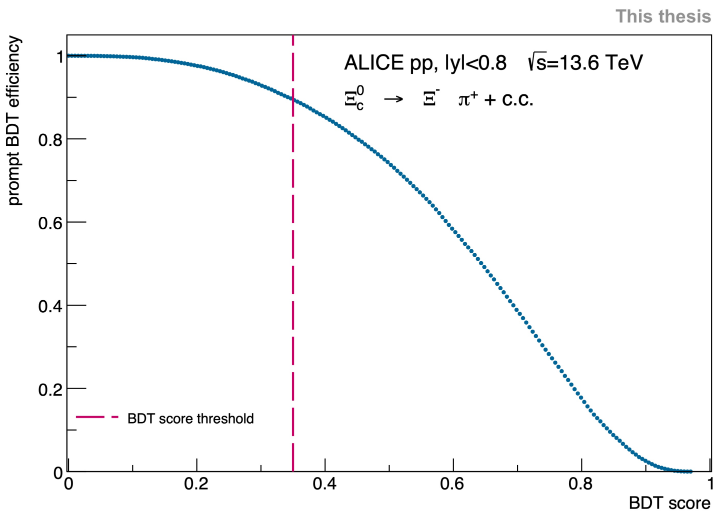


Figure H.2: Prompt BDT efficiency for the transverse momentum interval $3 < p_{\mathrm{T}\,\Xi_c^0} < 4$ GeV/c as a function of the BDT score threshold.

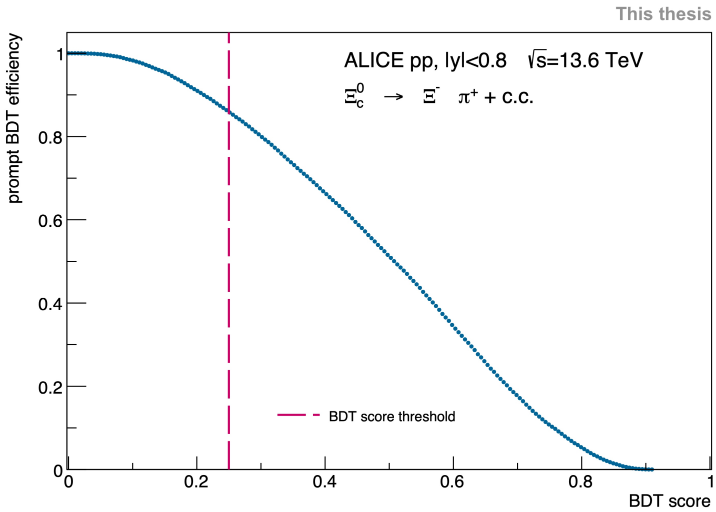


Figure H.3: Prompt BDT efficiency for the transverse momentum interval $4 < p_{\mathrm{T}\,\Xi_c^0} < 5$ GeV/c as a function of the BDT score threshold.

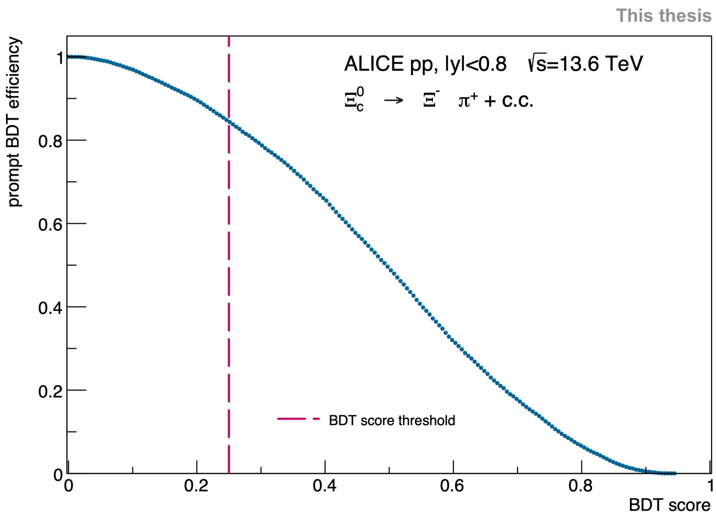


Figure H.4: Prompt BDT efficiency for the transverse momentum interval $5 < p_{\mathrm{T}\,\Xi_c^0} < 6$ GeV/c as a function of the BDT score threshold.

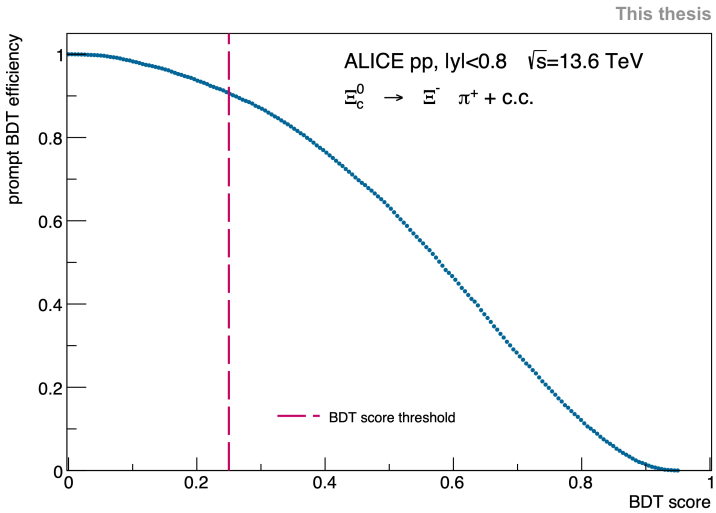


Figure H.5: Prompt BDT efficiency for the transverse momentum interval $6 < p_{\mathrm{T}\,\Xi_c^0} < 8$ GeV/c as a function of the BDT score threshold.

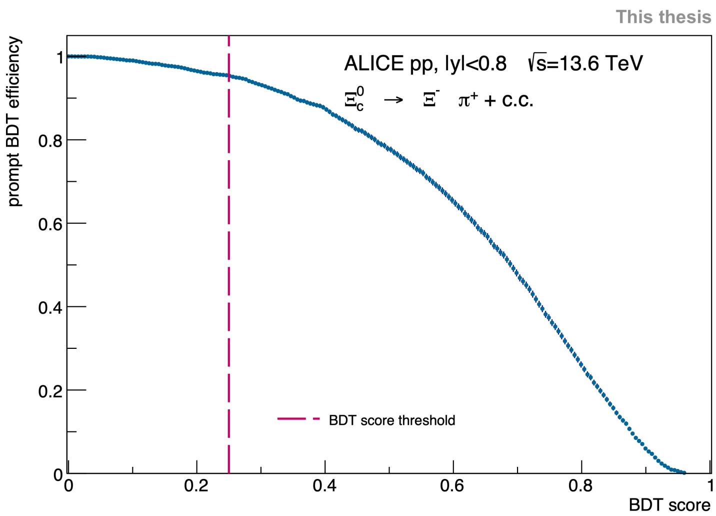
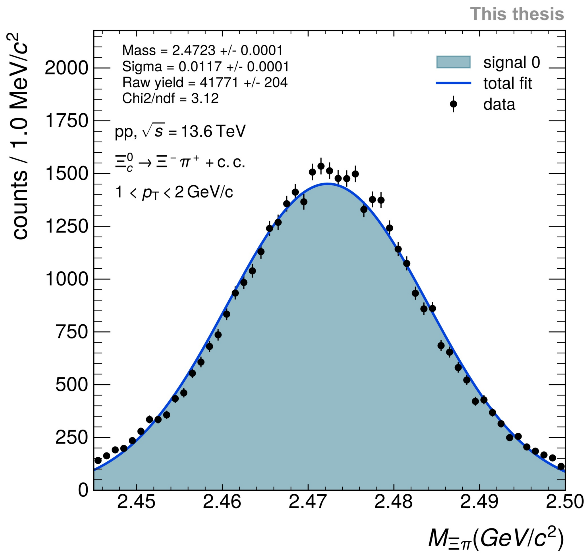


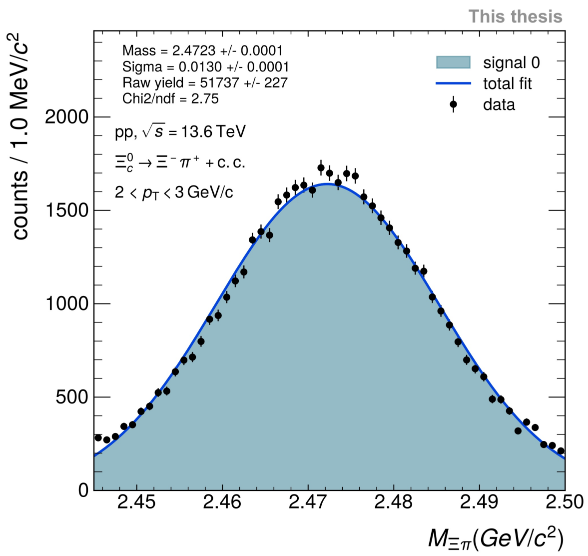
Figure H.6: Prompt BDT efficiency for the transverse momentum interval $8 < p_{\mathrm{T}\,\Xi_c^0} < 12$ GeV/c as a function of the BDT score threshold.

Appendix I

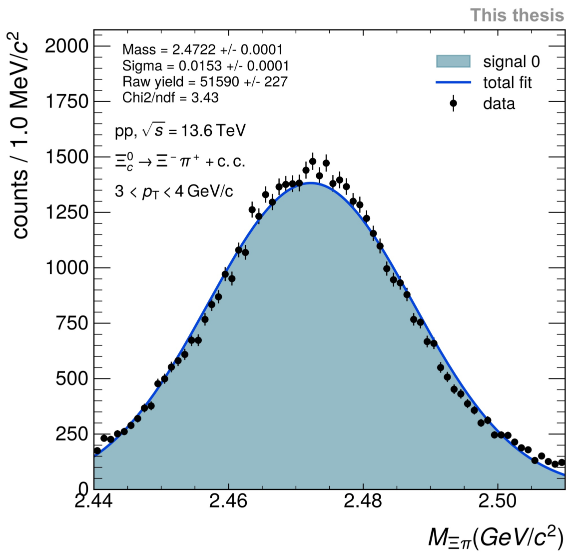
Monte Carlo invariant-mass spectra fits



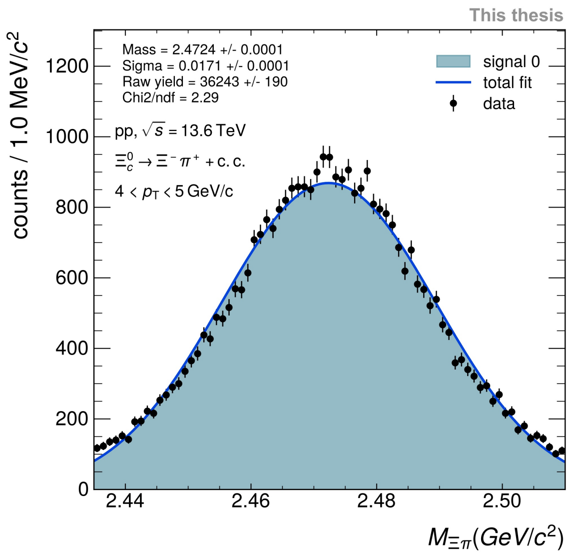
(a) $1 < p_{\mathrm{T}\,\Xi_c^0} < 2$ GeV/c



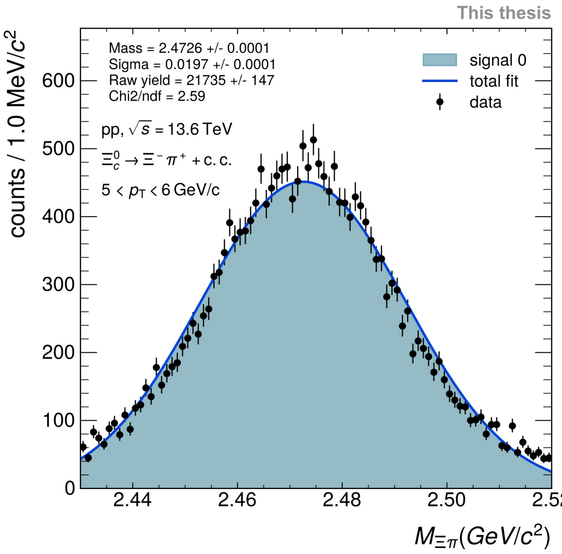
(b) $2 < p_{\mathrm{T}\,\Xi_c^0} < 3$ GeV/c



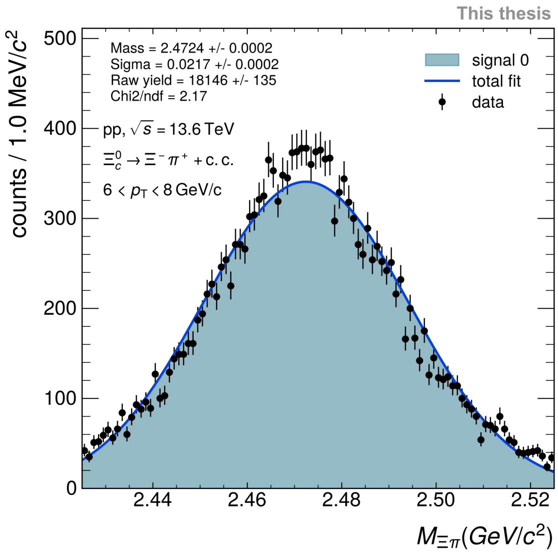
(c) $3 < p_{\mathrm{T}\,\Xi_c^0} < 4$ GeV/c



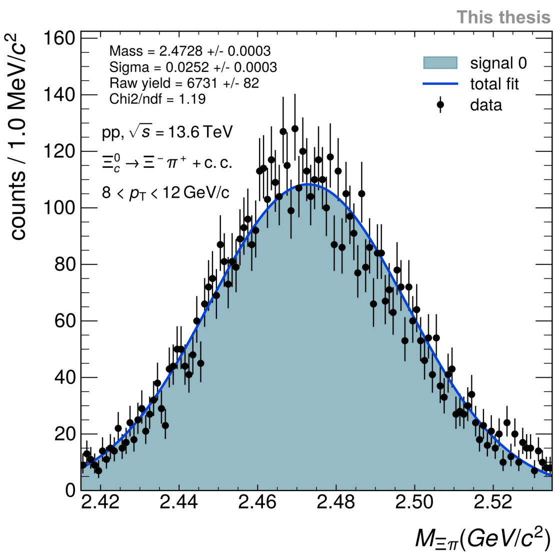
(d) $4 < p_{\mathrm{T}\,\Xi_c^0} < 5$ GeV/c



(e) $5 < p_{T\,\Xi_c^0} < 6$ GeV/c



(f) $6 < p_{T\,\Xi_c^0} < 8$ GeV/c

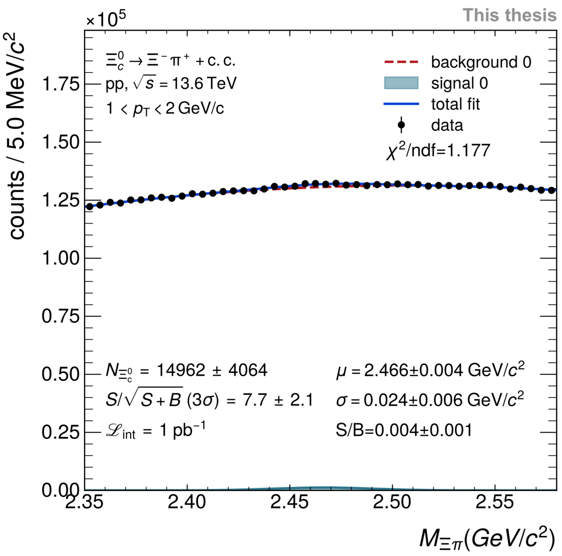


(g) $8 < p_{T\,\Xi_c^0} < 12$ GeV/c

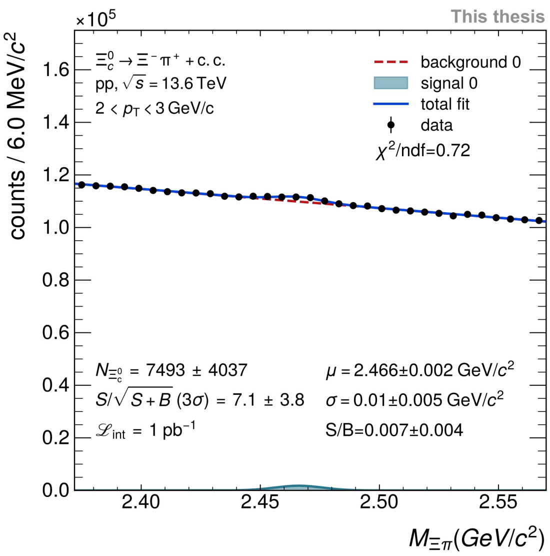
Figure I.1: Gaussian fit to MC signal invariant-mass spectra to extract the Ξ_c^0 peak width.

Appendix J

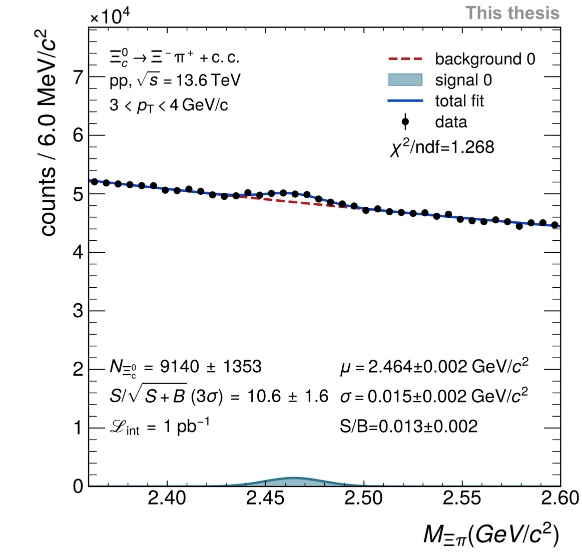
Real data invariant-mass spectra fits with free parameters



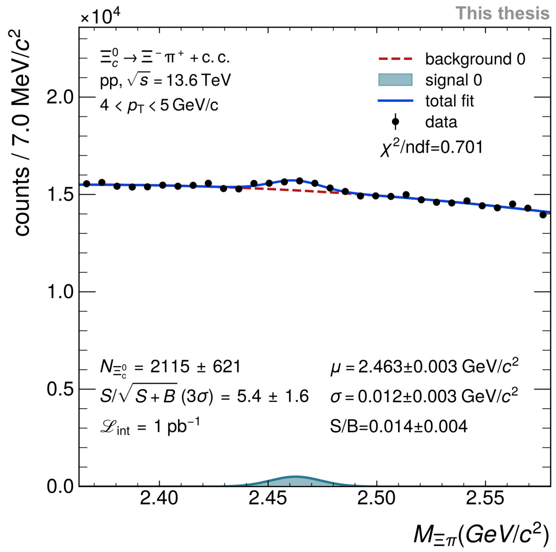
(a) $1 < p_{\mathrm{T}\,\Xi_c^0} < 2$ GeV/c



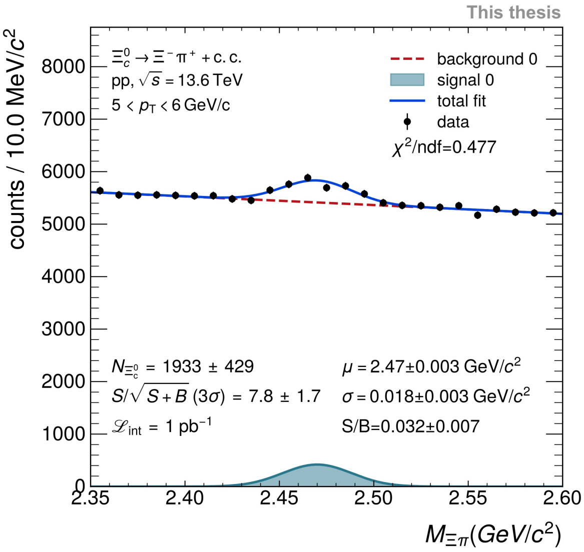
(b) $2 < p_{\mathrm{T}\,\Xi_c^0} < 3$ GeV/c



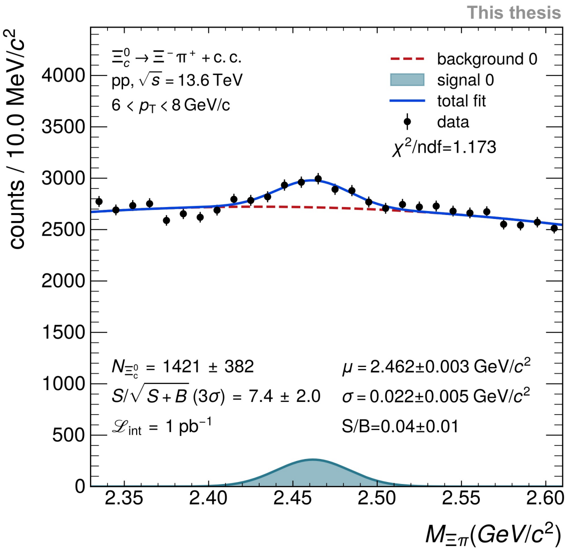
(c) $3 < p_{\mathrm{T}\,\Xi_c^0} < 4$ GeV/c



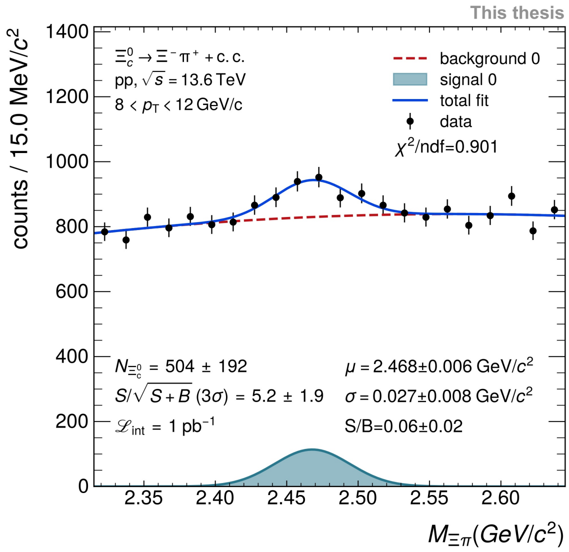
(d) $4 < p_{\mathrm{T}\,\Xi_c^0} < 5$ GeV/c



(e) $5 < p_{\mathrm{T}\,\Xi_c^0} < 6$ GeV/c

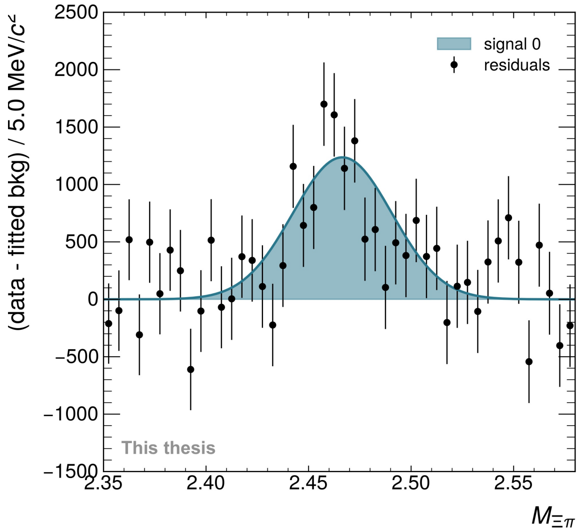


(f) $6 < p_{\mathrm{T}\,\Xi_c^0} < 8$ GeV/c

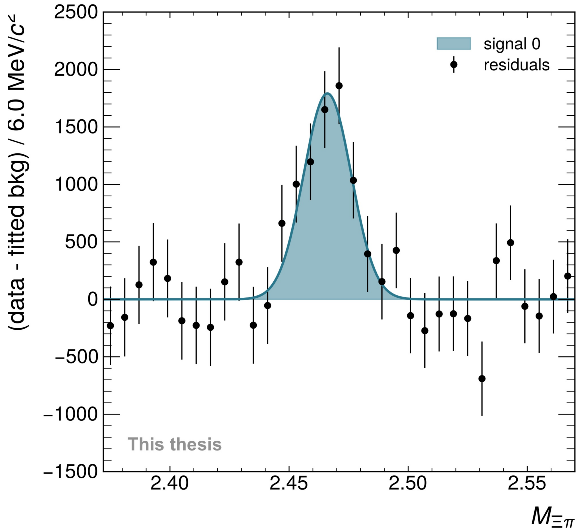


(g) $8 < p_{\mathrm{T}\,\Xi_c^0} < 12$ GeV/c

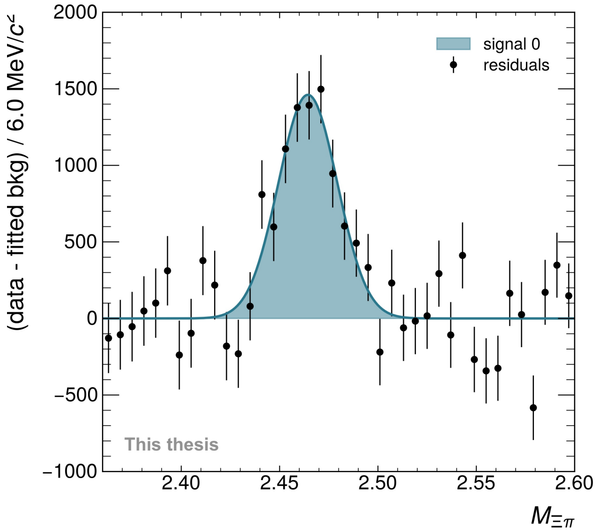
Figure J.1: Fit to real data invariant-mass spectra with no constraints on the width parameter.



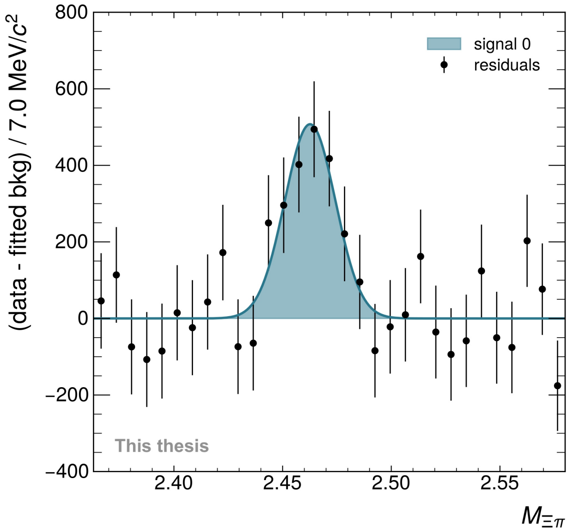
(a) $1 < p_{\mathrm{T}\,\Xi_c^0} < 2$ GeV/c



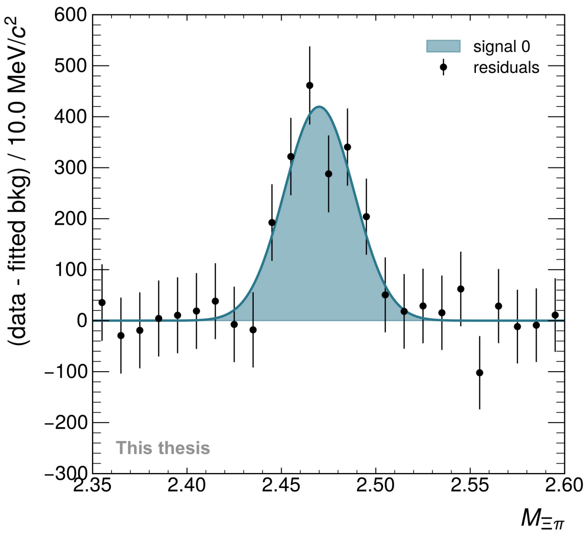
(b) $2 < p_{\mathrm{T}\,\Xi_c^0} < 3$ GeV/c



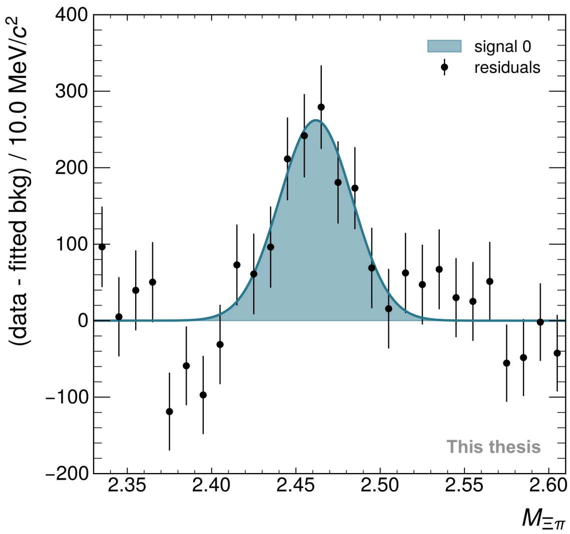
(c) $3 < p_{\mathrm{T}\,\Xi_c^0} < 4$ GeV/c



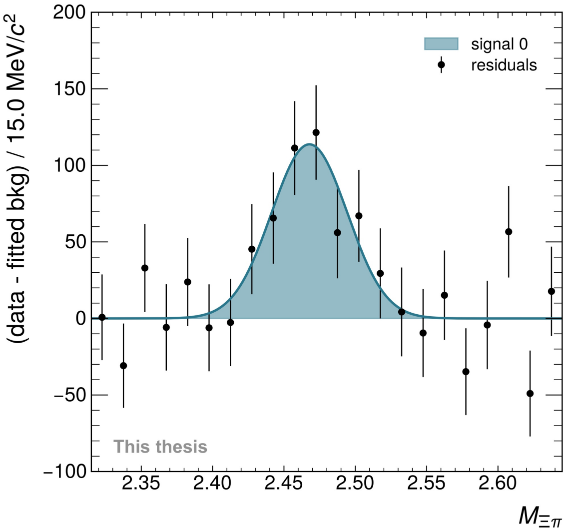
(d) $4 < p_{\mathrm{T}\,\Xi_c^0} < 5$ GeV/c



(e) $5 < p_{\mathrm{T}\,\Xi_c^0} < 6$ GeV/c



(f) $6 < p_{\mathrm{T}\,\Xi_c^0} < 8$ GeV/c

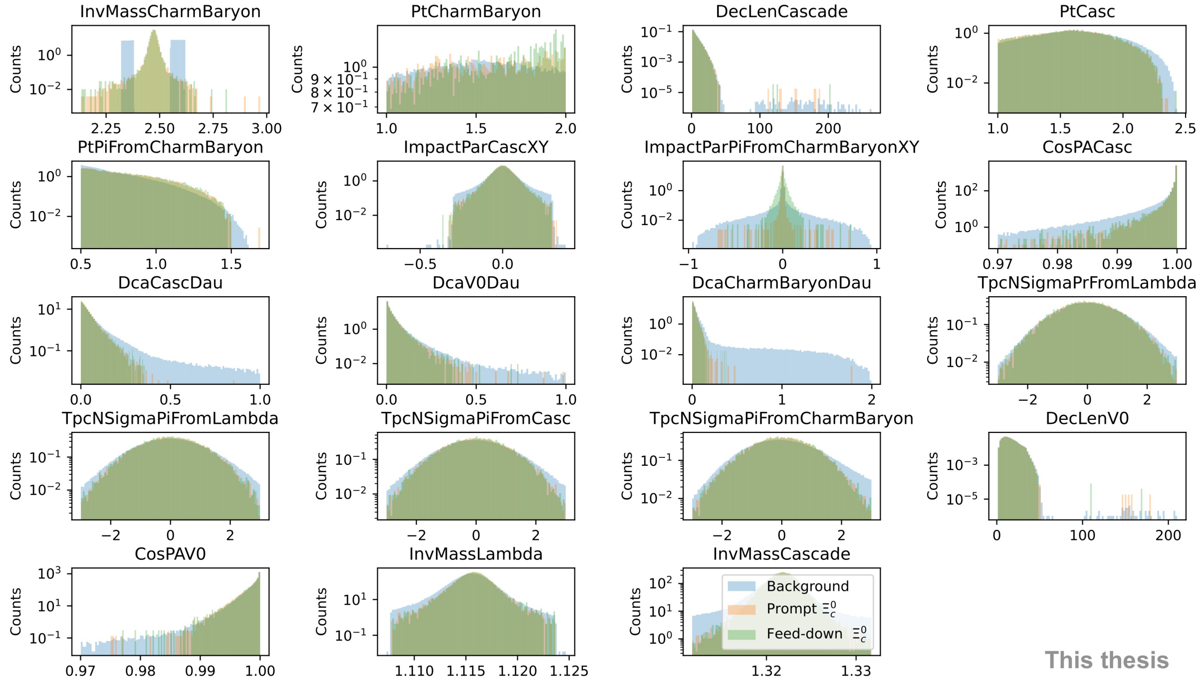


(g) $8 < p_{\mathrm{T}\,\Xi_c^0} < 12$ GeV/c

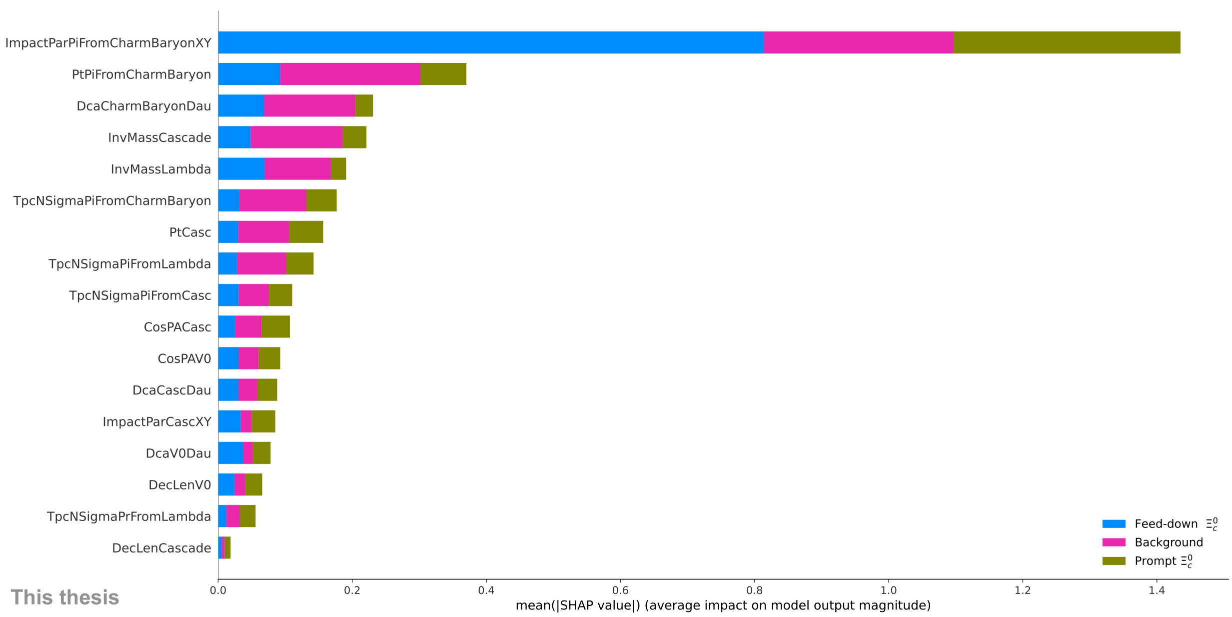
Figure J.2: Residuals of fit to real data invariant-mass spectra with no constraints on the width parameter.

Appendix K

BDT multi-classification with non-prompt signal

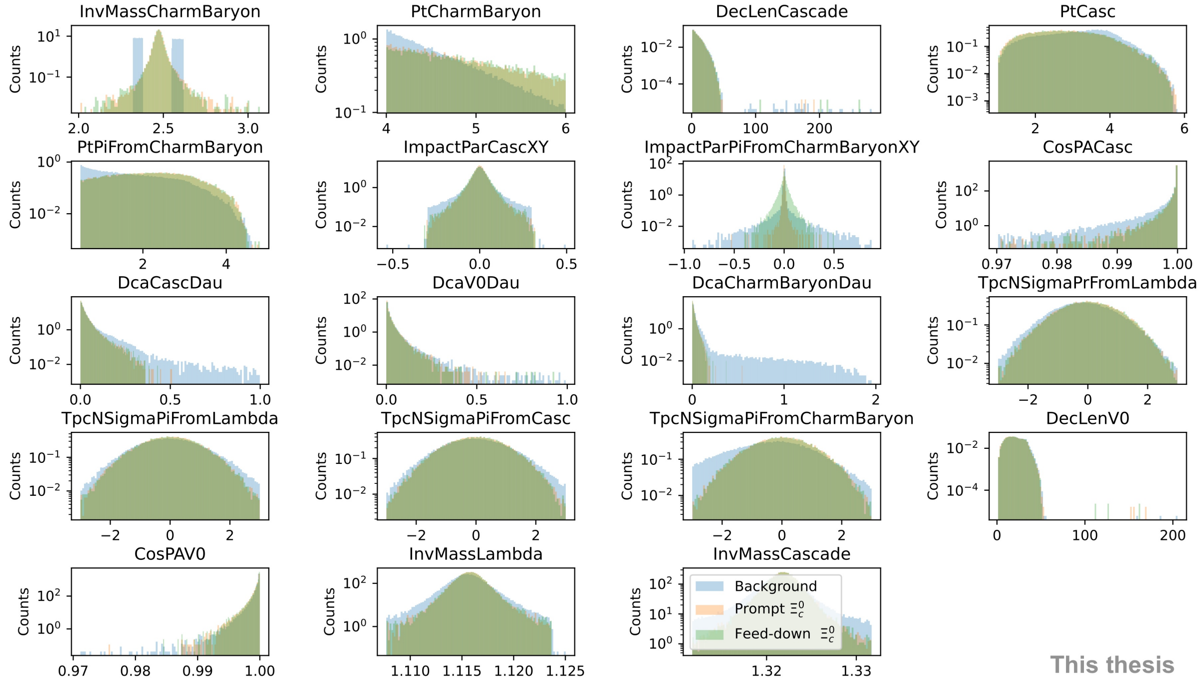


(a) Feature normalised distributions in data and MC.

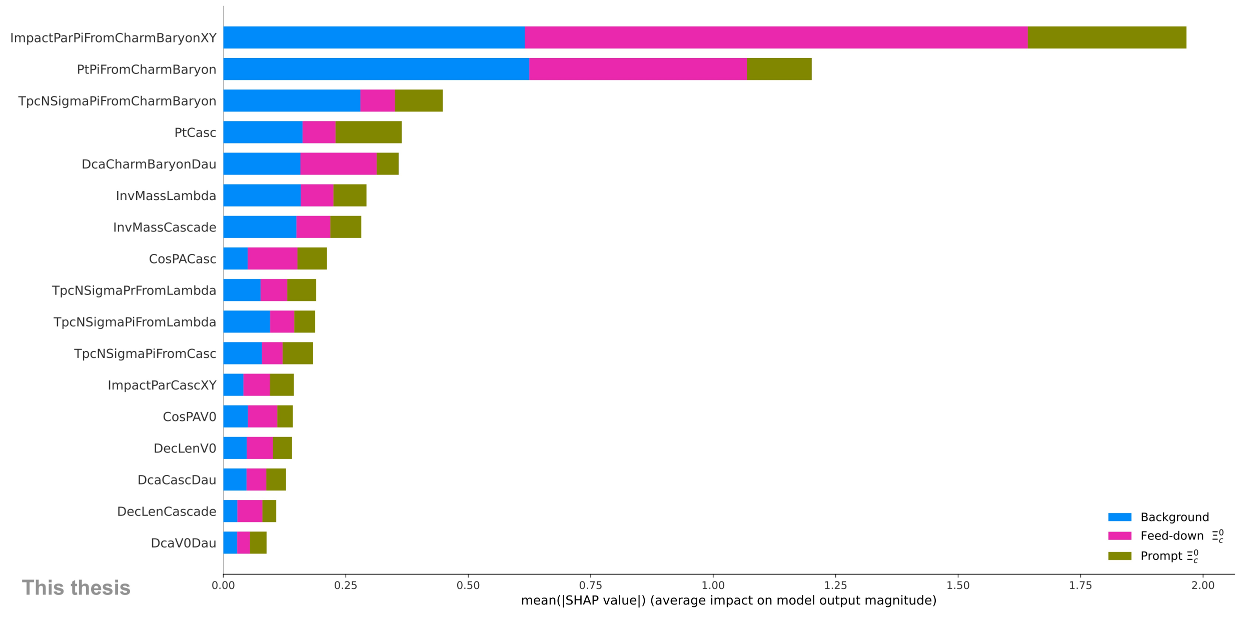


(b) Ranking of training features based on their impact on the candidates classification for the BDT model.

Figure K.1: Result of a BDT multi-classification with prompt signal, non-prompt signal, and background classes for the $p_{\mathrm{T}\,\Xi_c^0}$ interval $1 < p_{\mathrm{T}\,\Xi_c^0} < 2$ GeV/c.

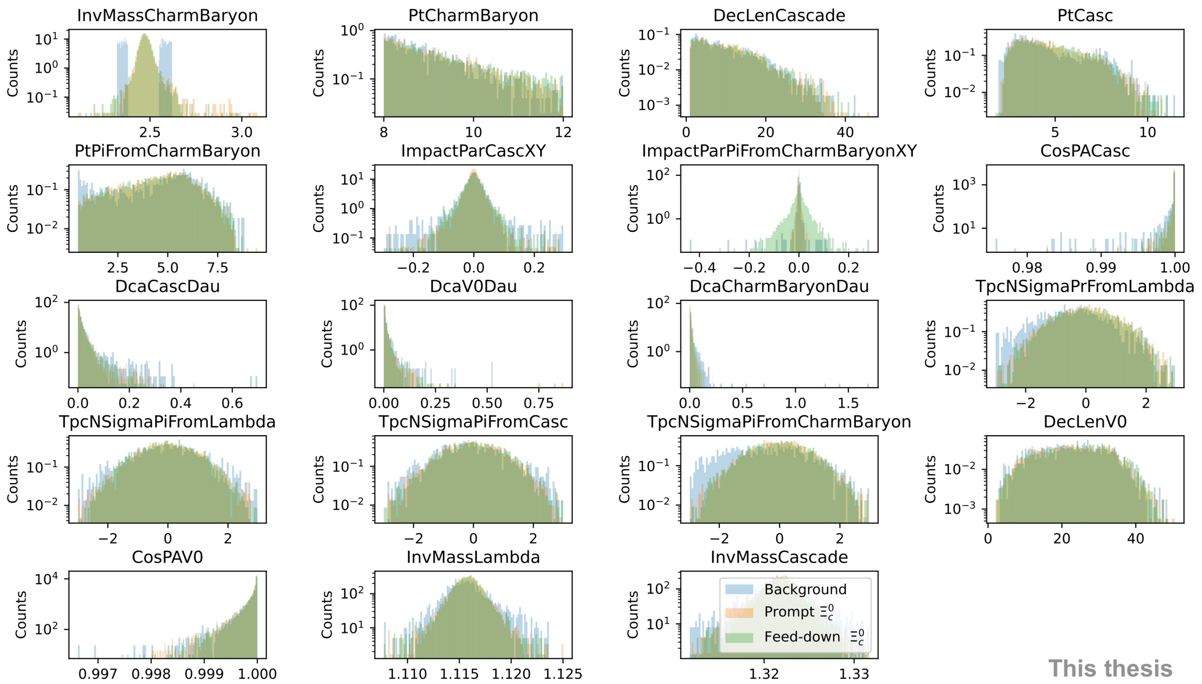


(a) Feature normalised distributions in data and MC.

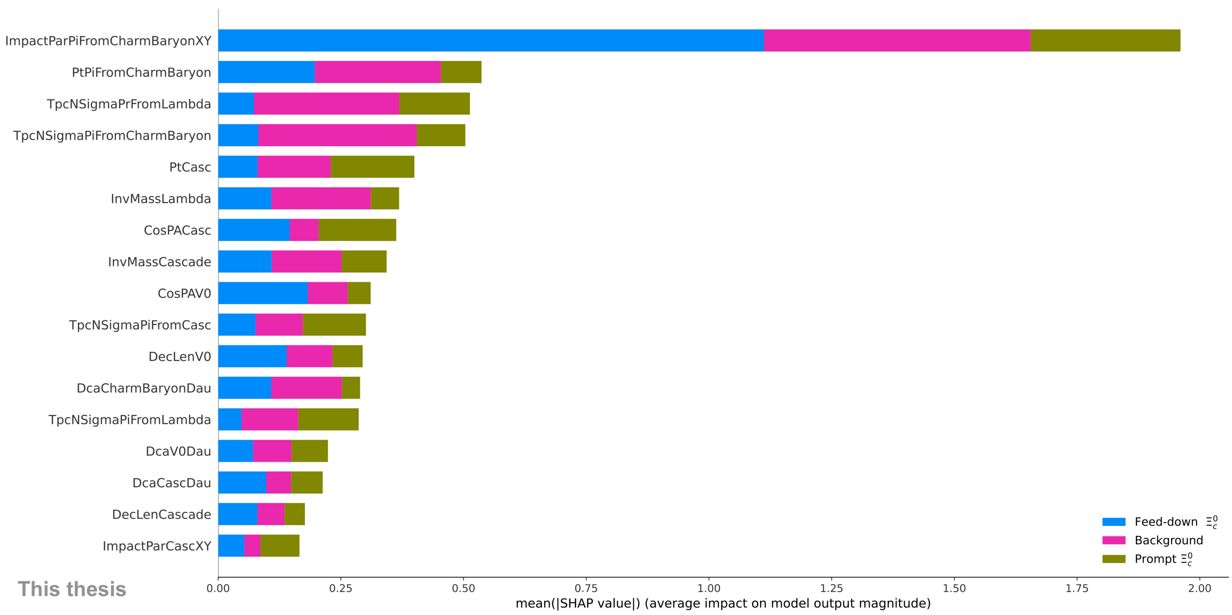


(b) Ranking of training features based on their impact on the candidates classification for the BDT model.

Figure K.2: Result of a BDT multi-classification with prompt signal, non-prompt signal, and background classes for the $p_{\mathrm{T}\,\Xi_c^0}$ interval $4 < p_{\mathrm{T}\,\Xi_c^0} < 6$ GeV/c.



(a) Feature normalised distributions in data and MC.



(b) Ranking of training features based on their impact on the candidates classification for the BDT model.

Figure K.3: Result of a BDT multi-classification with prompt signal, non-prompt signal, and background classes for the $p_{\mathrm{T}\,\Xi_c^0}$ interval $8 < p_{\mathrm{T}\,\Xi_c^0} < 12$ GeV/c.

Appendix L

Systematic uncertainty on raw yield extraction

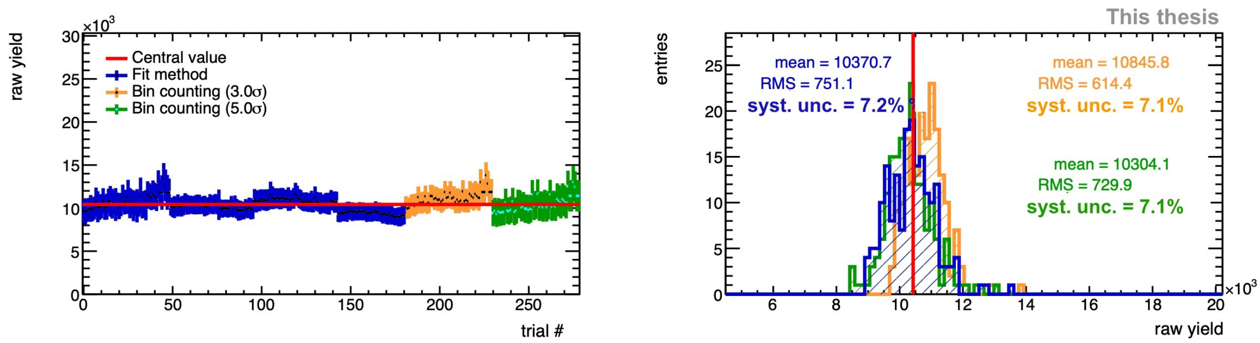


Figure L.1: Result of multi-trial approach for raw yield extraction in the transverse momentum interval $2 < p_{\mathrm{T}\,\Xi_c^0} < 3$ GeV/c.

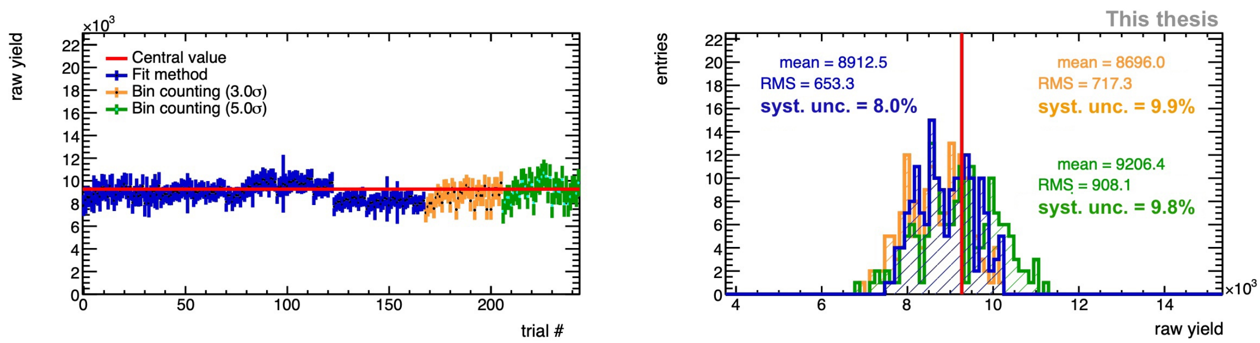


Figure L.2: Result of multi-trial approach for raw yield extraction in the transverse momentum interval $3 < p_{\mathrm{T}\,\Xi_c^0} < 4$ GeV/c.

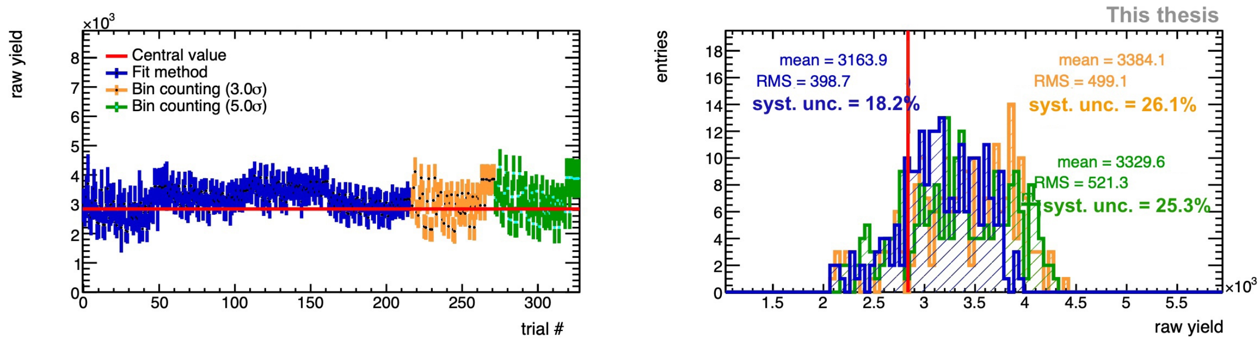


Figure L.3: Result of multi-trial approach for raw yield extraction in the transverse momentum interval $4 < p_{\mathrm{T}\,\Xi_c^0} < 5$ GeV/c.

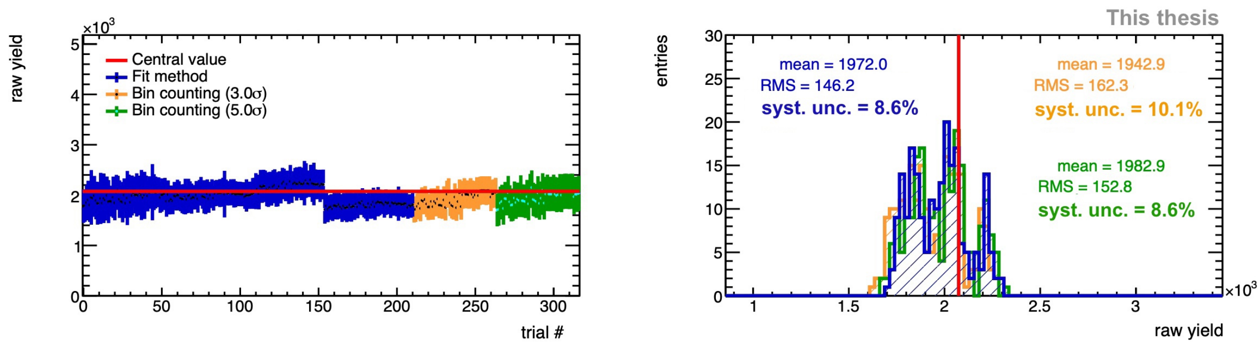


Figure L.4: Result of multi-trial approach for raw yield extraction in the transverse momentum interval $5 < p_{\mathrm{T}\,\Xi_c^0} < 6$ GeV/c.

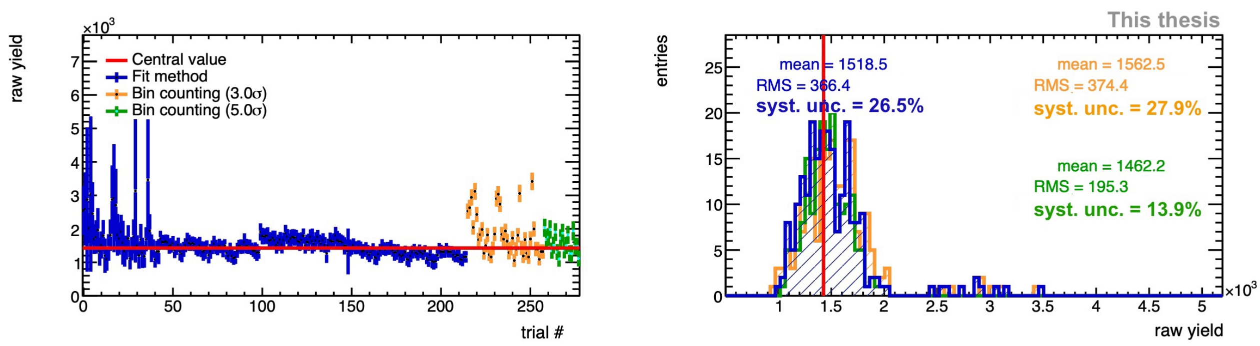


Figure L.5: Result of multi-trial approach for raw yield extraction in the transverse momentum interval $6 < p_{\mathrm{T}\,\Xi_c^0} < 8$ GeV/c.

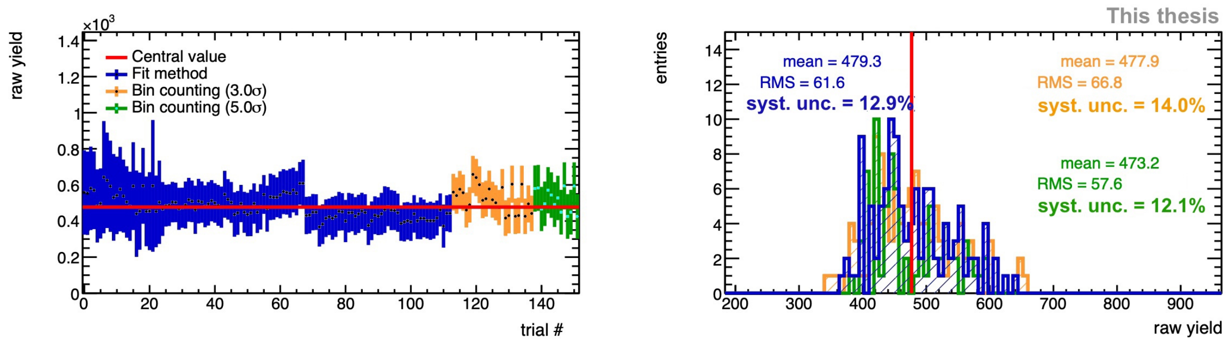


Figure L.6: Result of multi-trial approach for raw yield extraction in the transverse momentum interval $8 < p_{\mathrm{T}\,\Xi_c^0} < 12$ GeV/c.

Appendix M

Systematic uncertainty on BDT selections

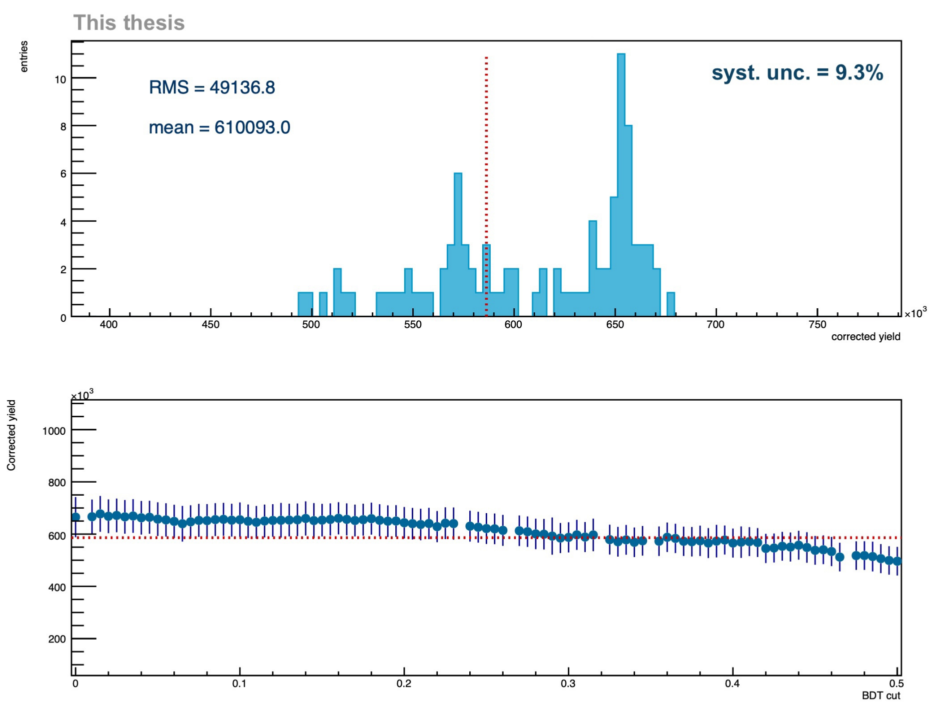


Figure M.1: Systematic uncertainty estimation associated to BDT selections for the p_{T} interval $2 < p_{\mathrm{T}\,\Xi_c^0} < 3$ GeV/c.

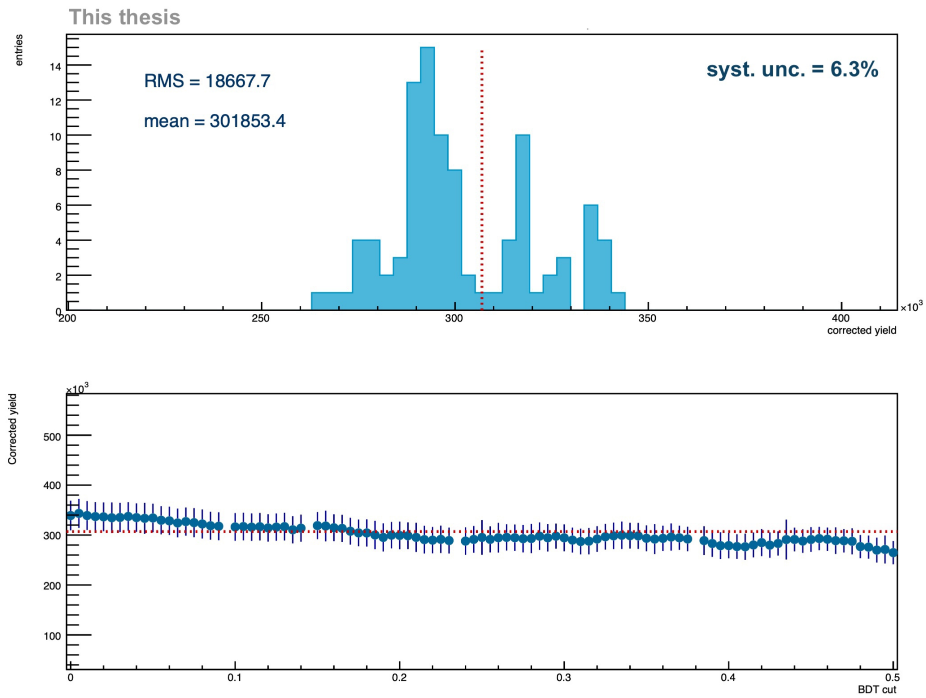


Figure M.2: Systematic uncertainty estimation associated to BDT selections for the p_{T} interval $3 < p_{\mathrm{T}\,\Xi_c^0} < 4$ GeV/c.

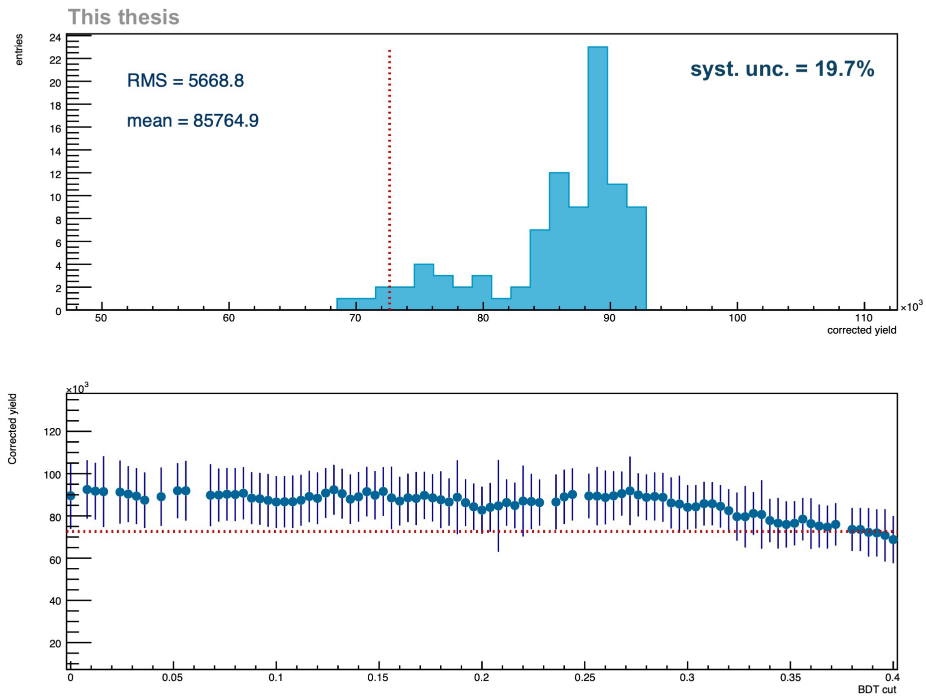


Figure M.3: Systematic uncertainty estimation associated to BDT selections for the p_{T} interval $4 < p_{\mathrm{T}\,\Xi_c^0} < 5$ GeV/c.

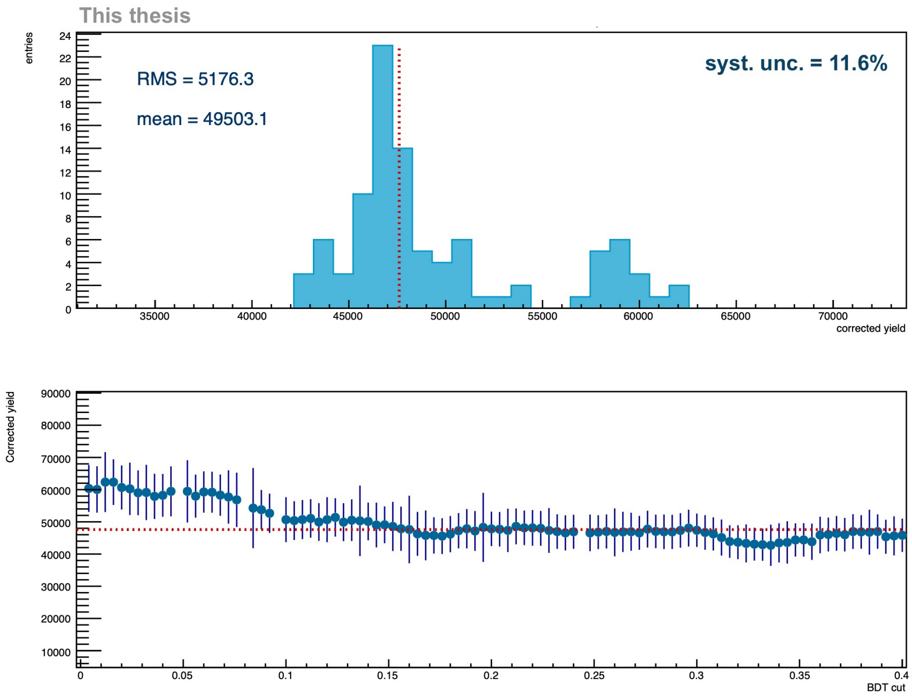


Figure M.4: Systematic uncertainty estimation associated to BDT selections for the p_{T} interval $5 < p_{\mathrm{T}\,\Xi_c^0} < 6$ GeV/c.

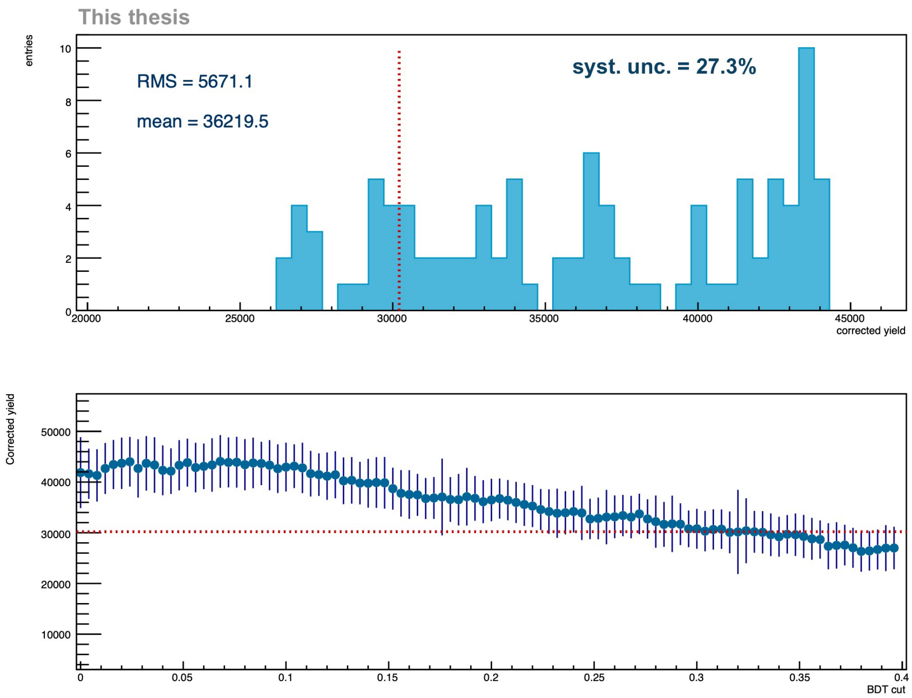


Figure M.5: Systematic uncertainty estimation associated to BDT selections for the p_{T} interval $6 < p_{\mathrm{T}\,\Xi_c^0} < 8$ GeV/c.

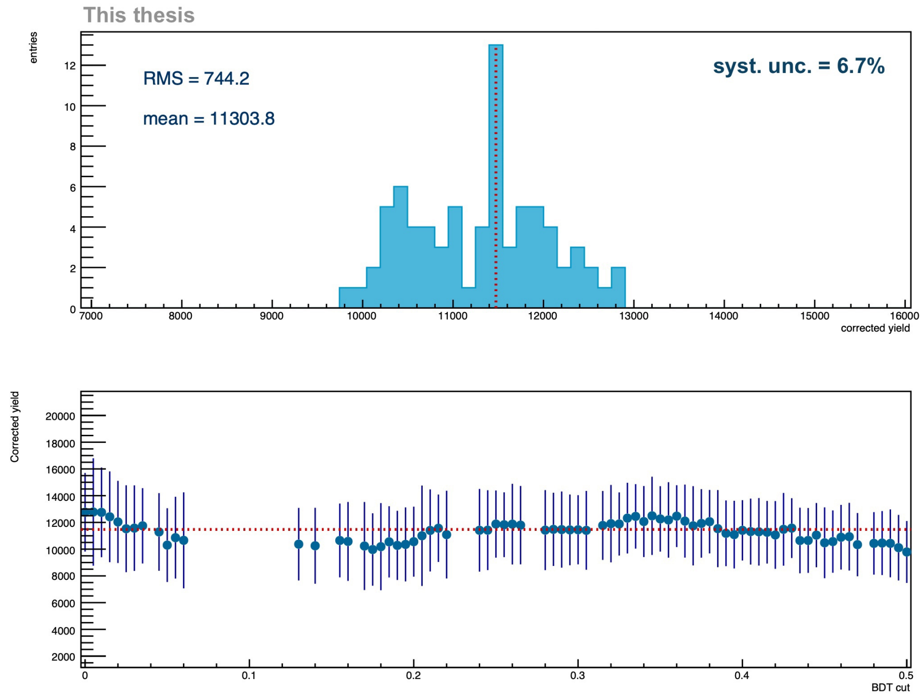


Figure M.6: Systematic uncertainty estimation associated to BDT selections for the p_{T} interval $8 < p_{\mathrm{T}\,\Xi_c^0} < 12$ GeV/c.

Appendix N

Systematic uncertainty on tracking preselections

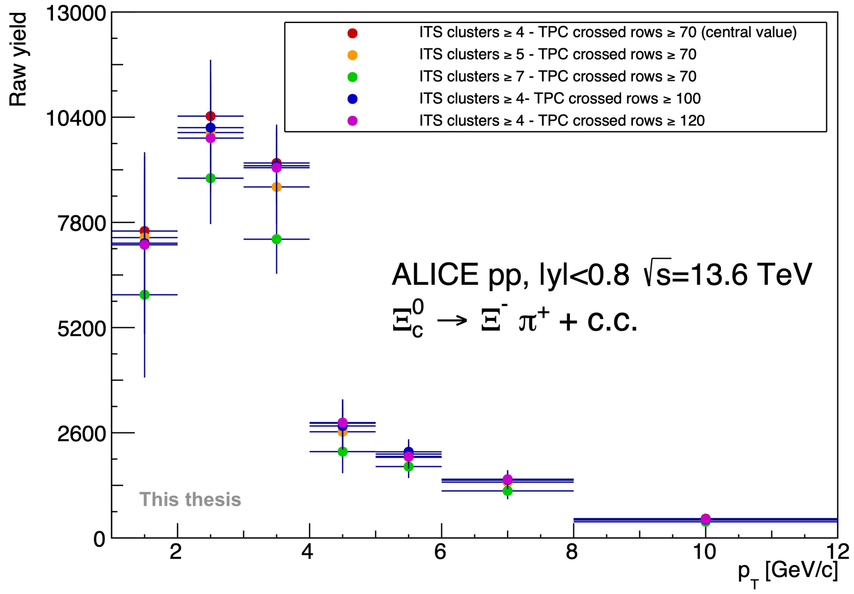


Figure N.1: Raw Ξ_c^0 yields for all the pion $\pi \leftarrow \Xi_c^0$ tracking preselection configurations.

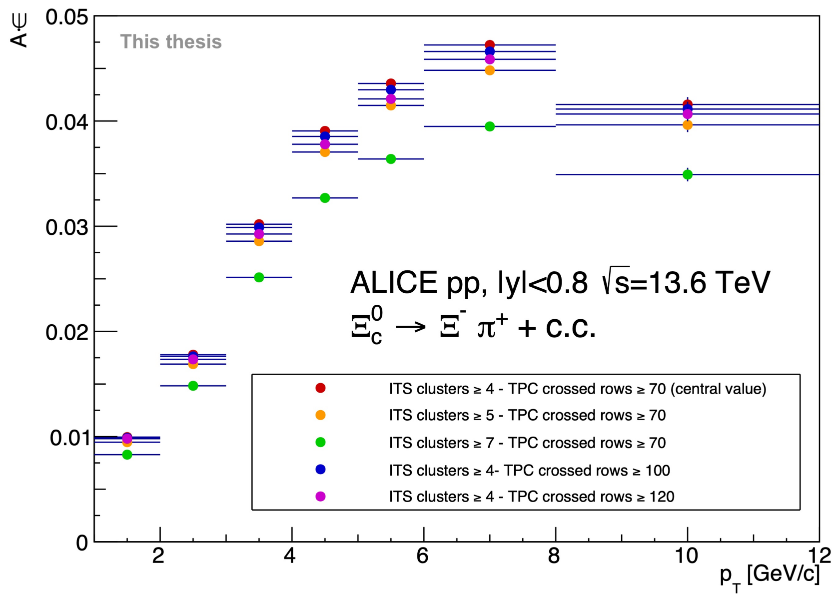


Figure N.2: Prompt total Ξ_c^0 acceptance-times-efficiency for all the pion $\pi \leftarrow \Xi_c^0$ tracking preselection configurations.

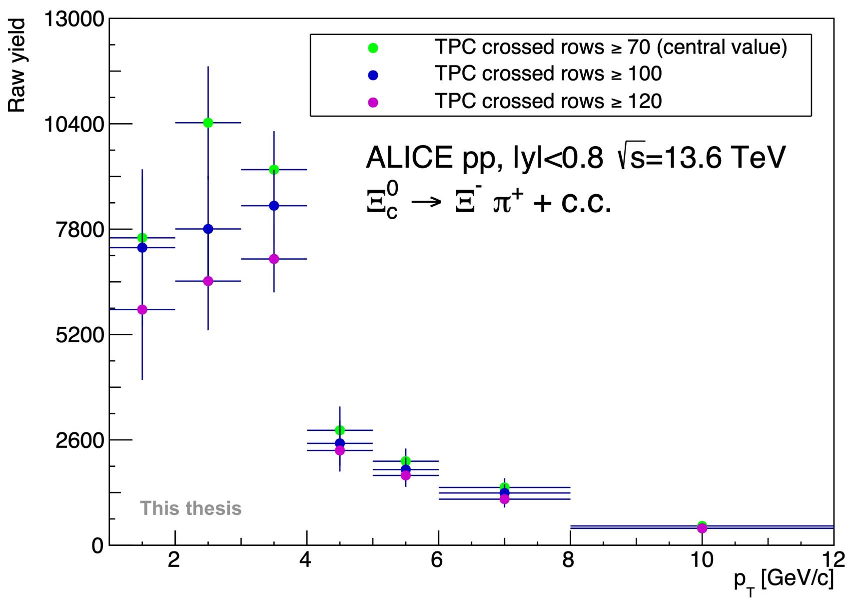


Figure N.3: Raw Ξ_c^0 yields for all the tracking preselection configurations of the Ξ daughters.

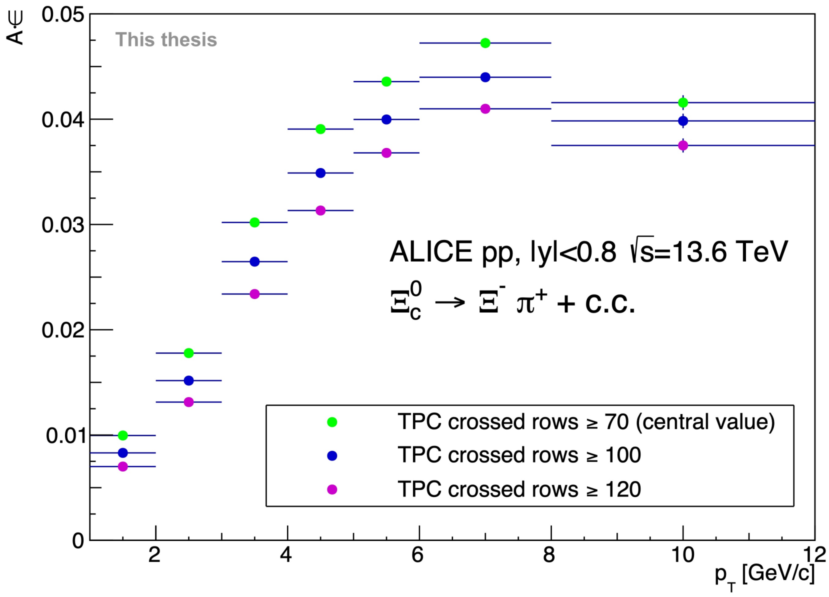
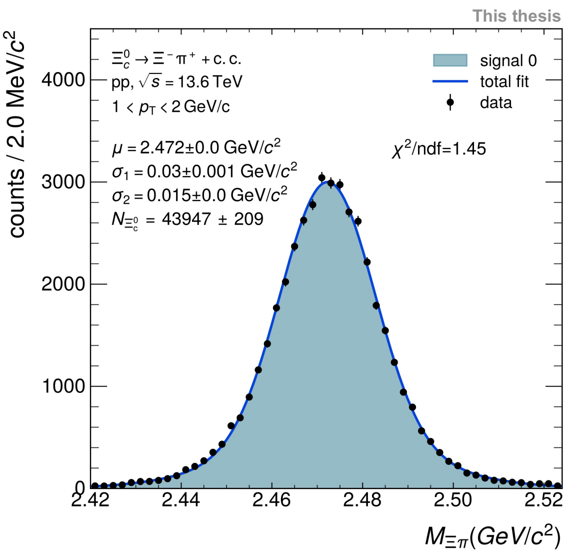


Figure N.4: Prompt total Ξ_c^0 acceptance-times-efficiency for all the tracking preselection configurations of the Ξ daughters.

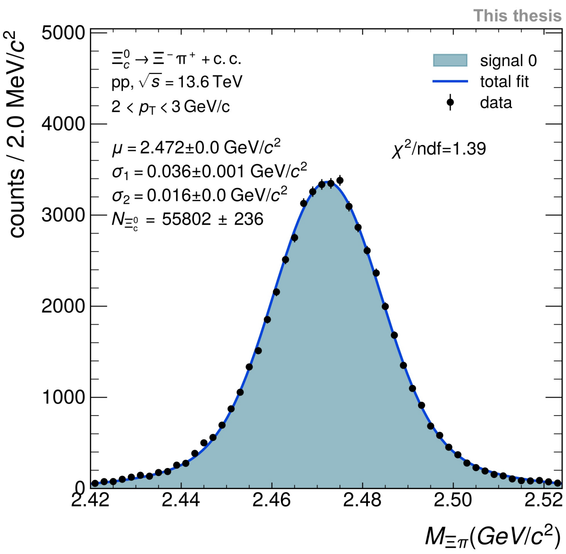
Appendix O

Double Gaussian fits

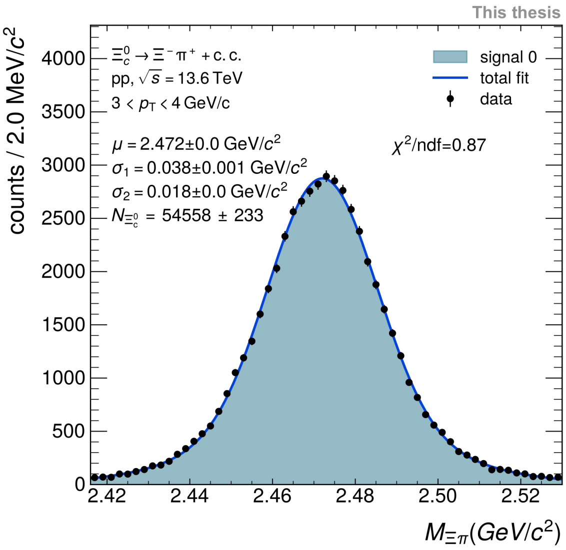
O.1 Monte Carlo invariant-mass spectra fits



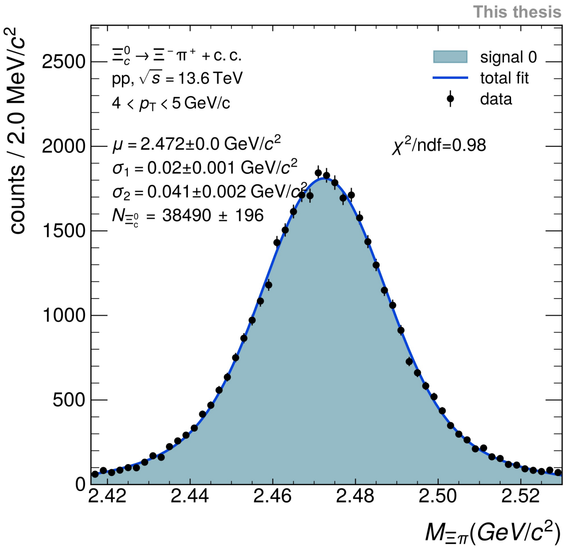
(a) $1 < p_{\mathrm{T}\,\Xi_c^0} < 2$ GeV/c



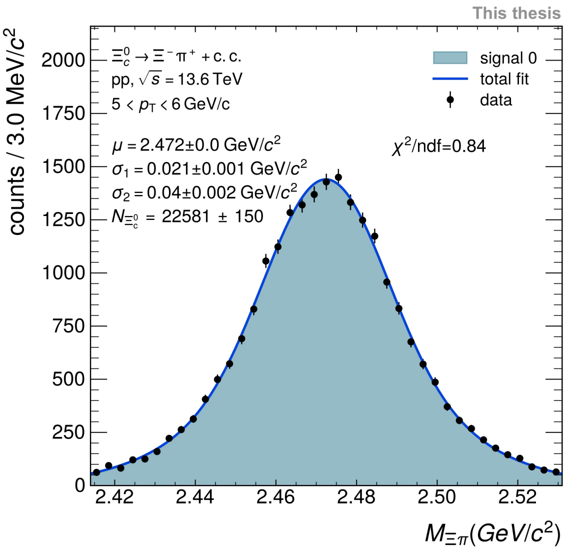
(b) $2 < p_{\mathrm{T}\,\Xi_c^0} < 3$ GeV/c



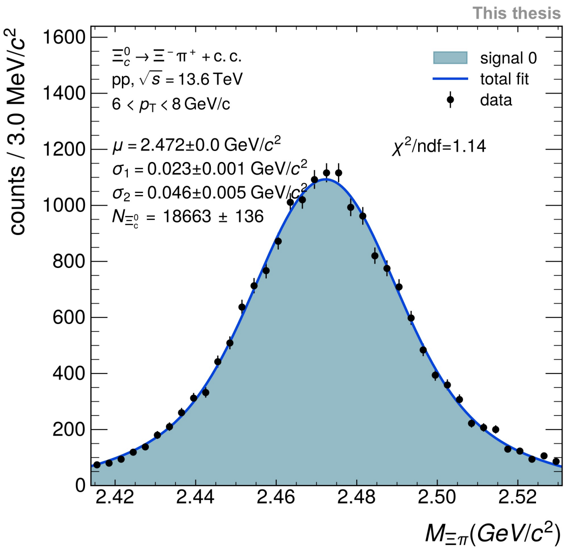
(c) $3 < p_{\mathrm{T}\,\Xi_c^0} < 4$ GeV/c



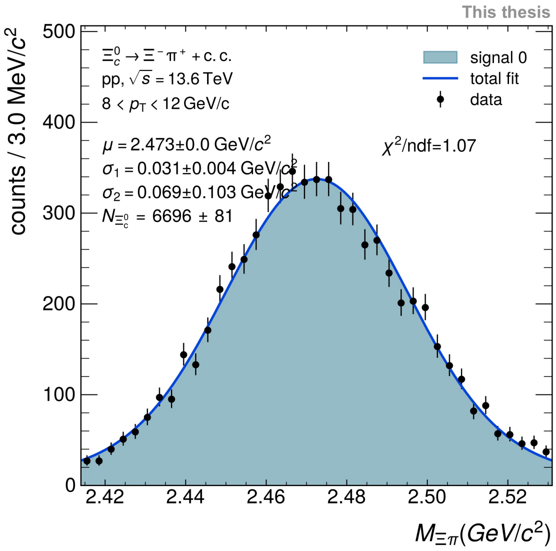
(d) $4 < p_{\mathrm{T}\,\Xi_c^0} < 5$ GeV/c



(e) $5 < p_{\mathrm{T}\,\Xi_c^0} < 6$ GeV/c



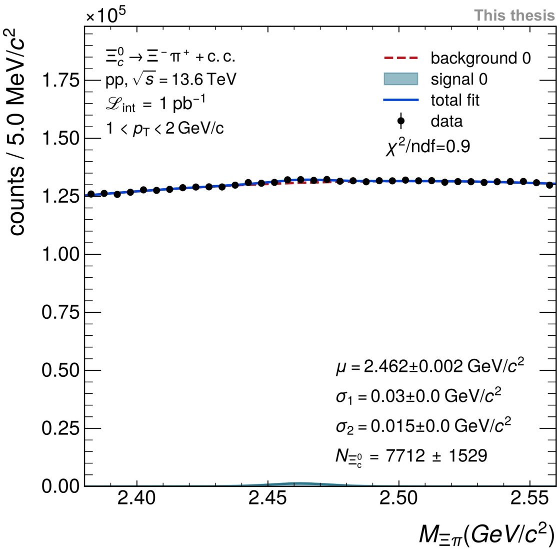
(f) $6 < p_{\mathrm{T}\,\Xi_c^0} < 8$ GeV/c



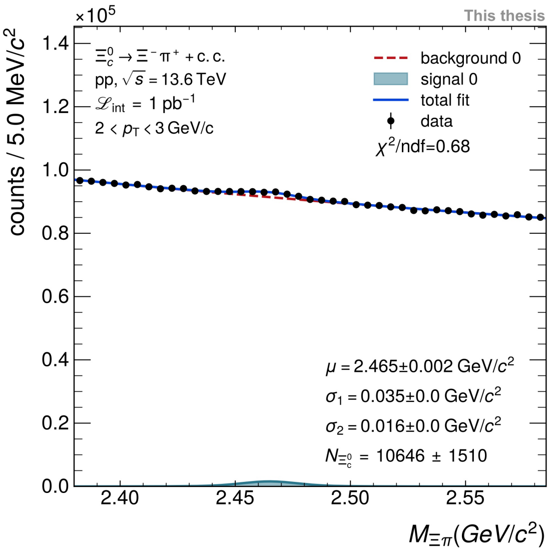
(g) $8 < p_{T\ \Xi_c^0} < 12$ GeV/c

Figure O.1: Double Gaussian fit to MC signal invariant-mass spectra to extract the Ξ_c^0 fit parameters σ_1, σ_2, and the fraction of signal represented by the first Gaussian shape.

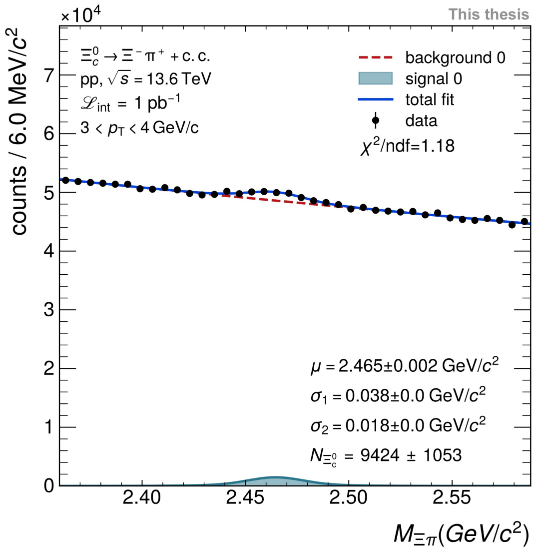
O.2 Real data invariant-mass spectra fits



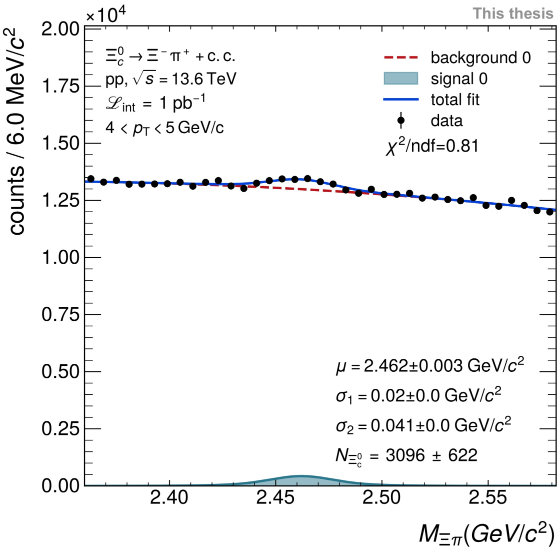
(a) $1 < p_{T\ \Xi_c^0} < 2$ GeV/c



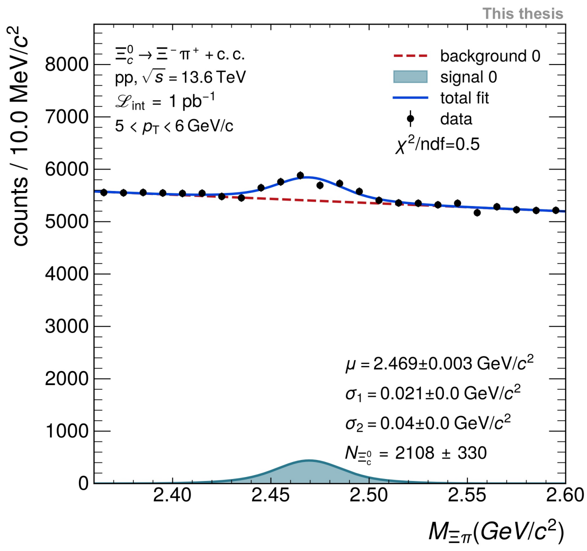
(b) $2 < p_{T\ \Xi_c^0} < 3$ GeV/c



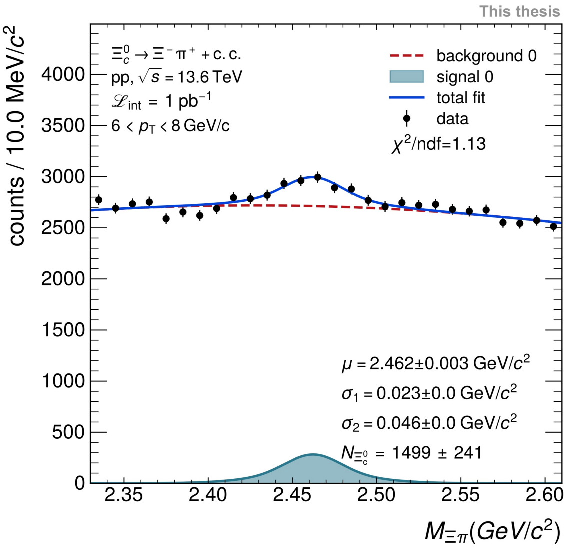
(c) $3 < p_{\mathrm{T}\,\Xi_c^0} < 4$ GeV/c



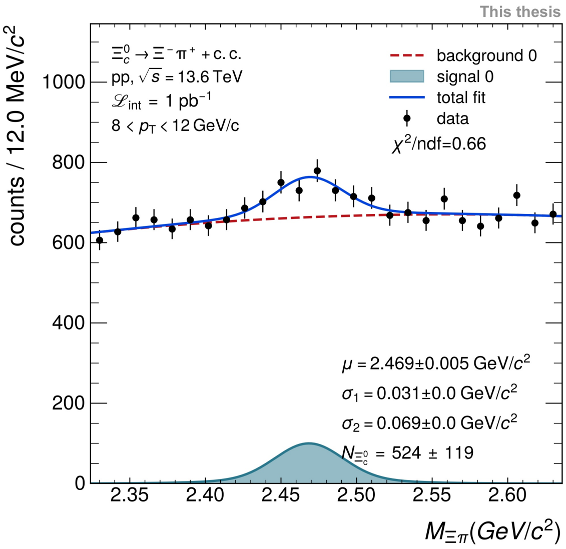
(d) $4 < p_{\mathrm{T}\,\Xi_c^0} < 5$ GeV/c



(e) $5 < p_{\mathrm{T}\,\Xi_c^0} < 6$ GeV/c



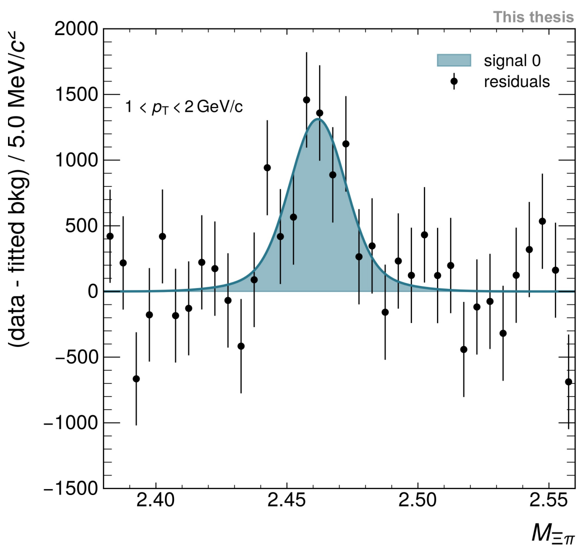
(f) $6 < p_{\mathrm{T}\,\Xi_c^0} < 8$ GeV/c



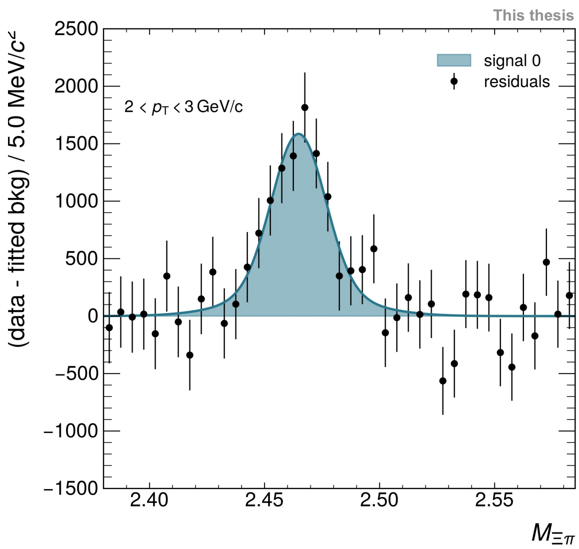
(g) $8 < p_{\mathrm{T}\,\Xi_c^0} < 12$ GeV/c

Figure O.2: Double Gaussian fit to real data invariant-mass spectra.

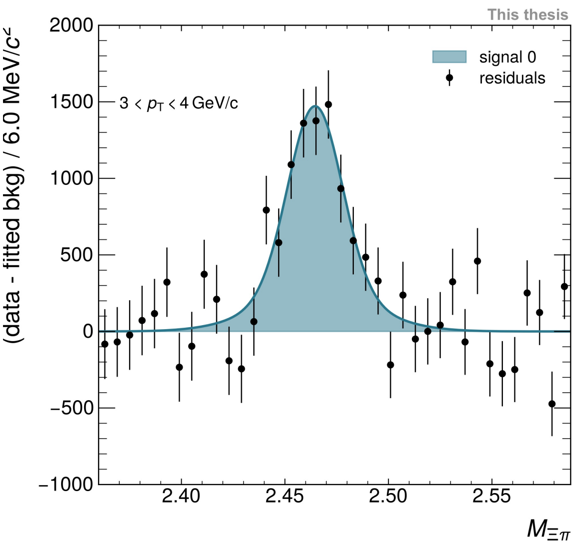
O.3 Residuals of real data invariant-mass spectra fits



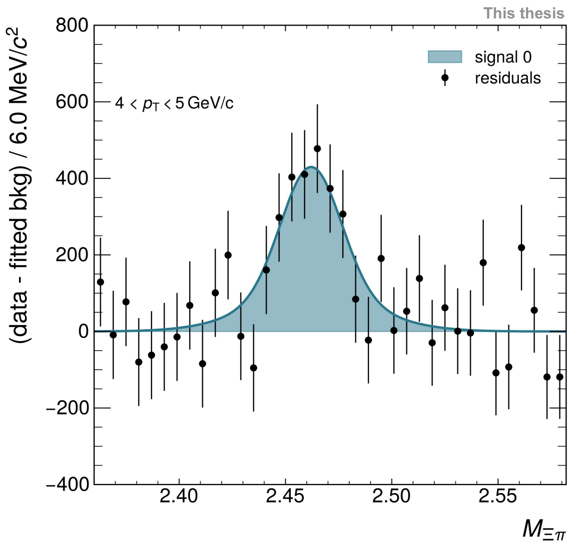
(a) $1 < p_{\mathrm{T}\,\Xi_c^0} < 2$ GeV/c



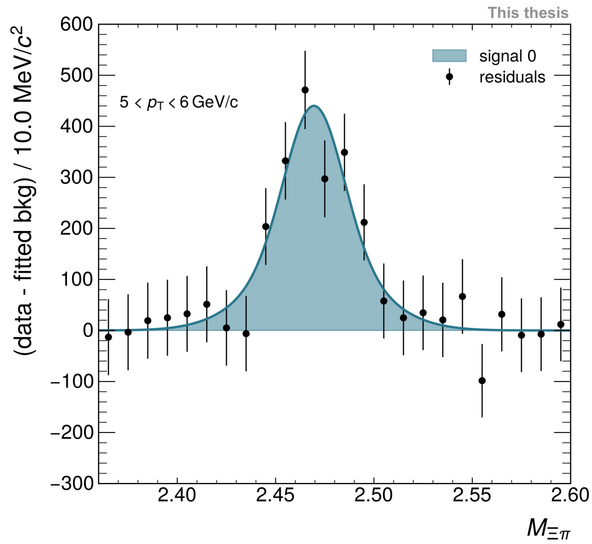
(b) $2 < p_{\mathrm{T}\,\Xi_c^0} < 3$ GeV/c



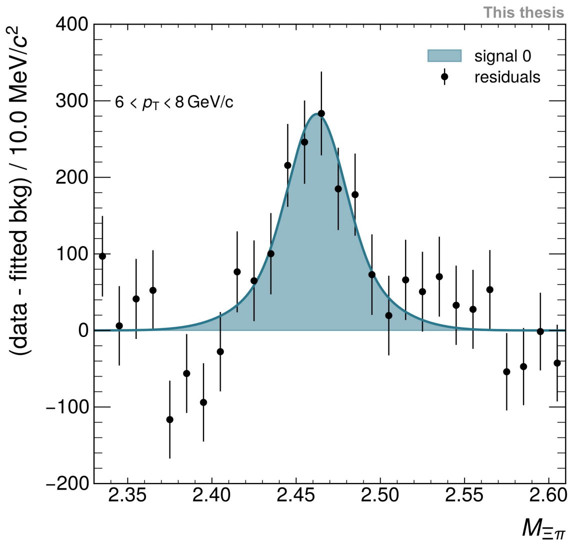
(c) $3 < p_{\mathrm{T}\,\Xi_c^0} < 4$ GeV/c



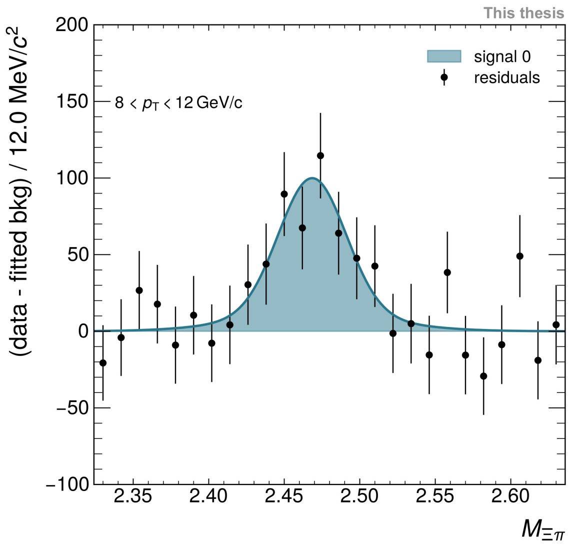
(d) $4 < p_{\mathrm{T}\,\Xi_c^0} < 5$ GeV/c



(e) $5 < p_{\mathrm{T}\,\Xi_c^0} < 6$ GeV/c



(f) $6 < p_{\mathrm{T}\,\Xi_c^0} < 8$ GeV/c



(g) $8 < p_{\mathrm{T}\ \Xi_c^0} < 12$ GeV/c

Figure O.3: Residuals of double Gaussian fit to real data invariant-mass spectra.

Bibliography

[1] Particle Data Group collaboration. Review of Particle Physics. *Phys. Rev. D*, 110, 2024. doi: https://link.aps.org/doi/10.1103/PhysRevD.110.030001.

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

[3] Particle Data Group collaboration. Review of Particle Physics. *Progress of Theoretical and Experimental Physics*, 2022. doi: https://doi.org/10.1093/ptep/ptac097.

[4] Particle Data Group. URL https://pdg.lbl.gov/.

[5] J. Altmann A. Dubla V. Greco A. Rossi and P. Skands. Towards the understanding of heavy quarks hadronization: from leptonic to heavy-ion collisions. *The European Physical Journal C*, 85, 2025. doi: http://dx.doi.org/10.1140/epjc/s10052-024-13641-5.

[6] ALICE collaboration. The ALICE experiment: a journey through QCD. *The European Physical Journal C*, 84, 2024. doi: http://dx.doi.org/10.1140/epjc/s10052-024-12935-y.

[7] Davison E. Soper. Parton distribution functions. *Nuclear Physics B - Proceedings Supplements*, 53, 1997. doi: http://dx.doi.org/10.1016/S0920-5632(96)00600-7.

[8] ALICE collaboration. Charm-quark fragmentation fractions and production cross section at midrapidity in pp collisions at the LHC. *Physical Review D*, 105, 2022. doi: http://dx.doi.org/10.1103/PhysRevD.105.L011103.

[9] D.J. Schwarz. The first second of the Universe. *Annalen der Physik*, 2003. doi: http://dx.doi.org/10.1002/andp.20035150402.

[10] ALICE collaboration. Charm-quark fragmentation fractions and production cross section at midrapidity in pp collisions at the LHC. *Phys. Rev. D*, 2022. doi: https://link.aps.org/doi/10.1103/PhysRevD.105.L011103.

[11] ALICE collaboration. Charm production and fragmentation fractions at midrapidity in pp collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 2023. doi: https://doi.org/10.1007/JHEP12(2023)086.

[12] LHCb collaboration. Enhanced production of Λ_b^0 baryons in high-multiplicity pp collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. Lett.*, 132, 2024. doi: https://link.aps.org/doi/10.1103/PhysRevLett.132.081901.

[13] A. Rossi. Hadronization mechanisms (via heavy-flavour hadrons): Experiment. *Proceedings of Hard Probes 2023 conference*, 2023. doi: https://doi.org/10.48550/arXiv.2308.10202.

[14] S. Ferreres-Solé and T. Sjöstrand. The space–time structure of hadronization in the Lund model. *The European Physical Journal C*, 2018. doi: http://dx.doi.org/10.1140/epjc/s10052-018-6459-8.

[15] P. Skands S. Carrazza and J. Rojo. Tuning PYTHIA 8.1: the Monash 2013 tune. *The European Physical Journal C*, 2014. doi: http://dx.doi.org/10.1140/epjc/s10052-014-3024-y.

[16] P. Bartalini et al. Multi-Parton Interactions at the LHC. 2011. doi: https://arxiv.org/abs/1111.0469.

[17] J. Christiansen and P. Skands. String formation beyond leading colour. *Journal of High Energy Physics*, 2015. doi: https://doi.org/10.1007/JHEP08(2015)003.

[18] V. Minissale S. Plumari and V. Greco. Charm hadrons in pp collisions at LHC energy within a coalescence plus fragmentation approach. *Physics Letters B*, 821, 2021. doi: http://dx.doi.org/10.1016/j.physletb.2021.136622.

[19] WA82 collaboration. Study of D^+ and D Feynman's x distributions in π^--nucleus interactions at the SPS. *Physics Letters B*, 305, 1993. doi: https://journals.scholarsportal.info/details?uri=/03702693/v305i0004_l/402_sodadfdiiats.xml.

[20] J. Song at al. New feature of low p_T charm quark hadronization in pp collisions at $\sqrt{s} = 7$ TeV. *The European Physical Journal C*, 78, 2018. doi: https://doi.org/10.1140/epjc/s10052-018-5817-x.

[21] M. He and R. Rapp. Charm-baryon production in proton-proton collisions. *Physics Letters B*, 795, 2019. doi: https://doi.org/10.1016/j.physletb.2019.06.004.

[22] ALICE collaboration. First measurement of Λ_c^+ production down to $p_T = 0$ in pp and p-Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV. *Phys. Rev. C*, 107, 2023. doi: https://link.aps.org/doi/10.1103/PhysRevC.107.064901.

[23] D. Ebert R. N. Faustov and V. O. Galkin. Spectroscopy and Regge trajectories of heavy baryons in the relativistic quark-diquark picture. *Phys. Rev. D*, 84, 2011. doi: https://link.aps.org/doi/10.1103/PhysRevD.84.014025.

[24] ALICE collaboration. Measurement of prompt D^0, Λ_c^+, and $\Sigma_c^{0,++}(2455)$ production in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. Lett.*, 128, 2022. doi: https://link.aps.org/doi/10.1103/PhysRevLett.128.012001.

[25] ALICE collaboration. First measurement of Ω_c^0 production in pp collisions at $\sqrt{s} = 13$ TeV. *Physics Letters B*, 846, 2023. doi: https://doi.org/10.1016/j.physletb.2022.137625.

[26] ALICE collaboration. Measurement of the production cross section of prompt Ξ_c^0 baryons in p–Pb collisions at $\sqrt{s_{\rm NN}} = 5.02$ TeV. *The European Physical Journal C*, 85, 2025. doi: http://dx.doi.org/10.1140/epjc/s10052-024-13531-w.

[27] ALICE collaboration. Constraining hadronization mechanisms with Λ_c^+/D^0 production ratios in Pb–Pb collisions at $\sqrt{s_{\rm NN}}$=5.02 TeV. *Physics Letters B*, 839, 2023. doi: https://doi.org/10.1016/j.physletb.2023.137796.

[28] V. Greco C.M. Ko and R. Rapp. Quark coalescence for charmed mesons in ultrarelativistic heavy-ion collisions. *Physics Letters B*, 595, 2004. doi: http://dx.doi.org/10.1016/j.physletb.2004.06.064.

[29] ALICE collaboration. Observation of a multiplicity dependence in the $p_{\rm T}$-differential charm baryon-to-meson ratios in proton–proton collisions at $\sqrt{s}$=13 TeV. *Physics Letters B*, 829, 2022. doi: https://doi.org/10.1016/j.physletb.2022.137065.

[30] ALICE Analysis Notes. URL https://alice-notes.web.cern.ch/.

[31] ALICE Figures Repository. URL https://alice-figure.web.cern.ch/.

[32] ALICE collaboration. Measurement of the cross sections of Ξ_c^0 and Ξ_c^+ baryons and of the branching-fraction ratio $\mathrm{BR}(\Xi_c^0 \to \Xi^- e^+ \nu_e)/\mathrm{BR}(\Xi_c^0 \to \Xi^- \pi^+ \nu_e)$ in pp collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. Lett.*, 127, 2021. doi: https://doi.org/10.1103/PhysRevLett.127.272001.

[33] LHCb collaboration. Measurement of Ξ_c^+ production in $\mathrm{p-Pb}$ collisions at $\sqrt{s_{\rm NN}} = 8.16$ TeV at LHCb. *Physical Review C*, 109, 2024. doi: http://dx.doi.org/10.1103/PhysRevC.109.044901.

[34] P. Bryant and L. Evans. LHC Machine. *JINST*, 3, 2008. doi: 10.1088/1748-0221/3/08/S08001.

[35] E. Lopienska. The CERN accelerator complex, layout in 2022. Complexe des accelerateurs du CERN en janvier 2022. *General Photo*, 2022.

[36] M. Benedikt P. Collier V. Mertens J. Poole and K. Schindl. LHC Design Report. *CERN Yellow Reports: Monographs*, v.3 : the LHC Injector Chain, 2004. doi: 10.5170/CERN-2004-003-V-3.

[37] ALICE collaboration. ALICE upgrades during the LHC Long Shutdown 2. *JINST*, 19, 2024. doi: https://doi.org/10.1088/1748-0221/19/05/P05062.

[38] ALICE collaboration. Technical Design Report for the Upgrade of the ALICE Inner Tracking System. *Journal of Physics G: Nuclear and Particle Physics*, 41, 2014. doi: https://dx.doi.org/10.1088/0954-3899/41/8/087002.

[39] ALICE collaboration. The upgrade of the ALICE TPC with GEMs and continuous readout. *JINST*, 16, 2021. doi: https://doi.org/10.1088/1748-0221/16/03/p03022.

[40] ALICE collaboration. Technical Design report of the ALICE Time of Flight System (TOF), 2000. URL https://cds.cern.ch/record/430132/files/alice_toftdr.pdf.

[41] ALICE collaboration. Upgrade of the Readout Trigger System Technical Design Report, 2013. URL http://cds.cern.ch/record/1603472/files/ALICE-TDR-015.pdf?version=6.

[42] D. Rohr et al. for the ALICE Collaboration. Track Reconstruction in the ALICE TPC using GPUs for LHC Run 3. *Proceedings to Connecting The Dots Workshop, Seattle, 2018*. doi: https://doi.org/10.48550/arXiv.1811.11481.

[43] Track to collision association in ALICE. URL https://aliceo2group.github.io/analysis-framework/docs/advanced-specifics/trackToCollAssoc.html.

[44] W. Blum W. Riegler and L. Rolandi. *Particle Detection with Drift Chambers (2nd ed.)*. Springer, 2008.

[45] A. Alici for the ALICE collaboration. Particle identification with the ALICE Time-Of-Flight detector at the LHC. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 766, 2014. doi: https://doi.org/10.1016/j.nima.2014.05.059.

[46] BELLE collaboration. Evidence for the singly Cabibbo-suppressed decay $\Omega_c^0 \to \Xi^-\pi^+$ and search for $\Omega_c^0 \to \Xi^- K^+$ and $\Omega^- K^+$ decays at Belle. *Journal of High Energy Physics*, 2023. doi: http://dx.doi.org/10.1007/JHEP01(2023)055.

[47] LHCb collaboration. Observation of Cabibbo-suppressed two-body hadronic decays and precision mass measurement of the Ω_c^0 baryon. *Physical Review Letters*, 2024. doi: http://dx.doi.org/10.1103/PhysRevLett.132.081802.

[48] ALICE Collaboration. DCAFitterN from AliceO2 software. URL https://github.com/AliceO2Group/AliceO2/blob/dev/Common/DCAFitter/include/DCAFitter/DCAFitterN.h.

[49] L. Barioglio et al. Minimal heavy-ion physics environment for Machine Learning (hipe4ML). URL https://zenodo.org/records/7014886.

[50] XGBoost developers. XGBoost documentation, . URL https://xgboost.readthedocs.io/en/stable/index.html#.

[51] T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '16*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

[52] Optuna developers. Optuna documentation, . URL https://optuna.readthedocs.io/en/stable/index.html#.

[53] FlareFly developers. Flarefly code and documentation, . URL https://github.com/flarefly/flarefly.

[54] LHCb collaboration. Measurement of b hadron fractions in 13 TeV pp collisions. *Phys. Rev. D*, 100, 2019. doi: https://doi.org/10.1103/PhysRevD.100.031102.

[55] ALICE collaboration. Study of flavor dependence of the baryon-to-meson ratio in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. D*, 108, 2023. doi: https://doi.org/10.1103/PhysRevD.108.112003.

[56] M. Cacciari M. Greco and P. Nason. The p_{T} spectrum in heavy-flavour hadroproduction. *JHEP*, 1998. doi: https://doi.org/10.1088/1126-6708/1998/05/007.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Dr. Kai Schweda, for his constant support and the countless opportunities he has provided throughout my doctoral studies. His expertise, enthusiasm and encouragement have been instrumental in shaping both my research and my growth.
I am profoundly thankful to my conveners at CERN, Dr. Fabrizio Grosa and Dr. Fabio Colamaria, who were always eager to help and to answer my questions. Their expertise and insightful suggestions provided invaluable learning opportunities and improved the quality of my research.
Special thanks go to the TPC team for initiating me into the world of hardware projects. Their competence, guidance, and patience provided a stimulating learning environment that significantly broadened my technical skills and experimental perspective.
I would like to express my heartfelt thanks to my loved ones for their emotional support and for all the joyful moments we shared during these years. Most importantly, I owe my deepest gratitude to my family, whose unwavering support and belief in my abilities have been my foundation throughout this journey.