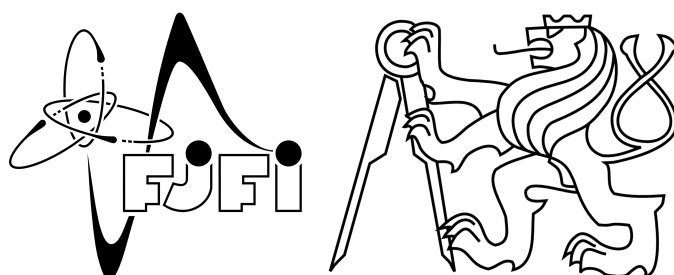


Czech Technical University in Prague
Faculty of Nuclear Sciences and Physical Engineering

Department of Physics
Nuclear and Particle Physics
Supervisor: Ing. Katarína Křížková Gajdošová, Ph.D.



Master Thesis

Λ_c^+ Production in pp Collisions at 13.6 TeV

Prague 2026

Bc. Karla Žertová

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Žertová** Jméno: **Karla** Osobní číslo: **502580**
Fakulta/ústav: **Fakulta jaderná a fyzikálně inženýrská**
Zadávající katedra/ústav: **Katedra fyziky**
Studijní program: **Jaderná a částicová fyzika**

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Produkce Λ_c baryonu v pp srážkách při energii 13,6 TeV

Název diplomové práce anglicky:

Λ_c production in pp collisions at 13.6 TeV

Jméno a pracoviště vedoucí(ho) diplomové práce:

Ing. Katarína Křížková Gajdošová, Ph.D. katedra fyziky FJFI

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) diplomové práce:

Datum zadání diplomové práce: **30.10.2025**

Termín odevzdání diplomové práce: **08.05.2026**

Platnost zadání diplomové práce: **30.10.2027**

Digitálně podepsal(a)
Martin Štefaňák
Datum: 04.11.2025
08:48:38
ID: 250650

podpis vedoucí(ho) ústavu/katedry

Digitálně podepsal(a)
Milan Krbálek
Datum: 05.11.2025
07:26:44
ID: 40556

podpis proděkana(ky) z pověření děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Diplomantka bere na vědomí, že je povinna vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

05.11.2025

Datum převzetí zadání

Bc. Žertová Karla

Podpis studentky

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Žertová** Jméno: **Karla** Osobní číslo: **502580**
Fakulta/ústav: **Fakulta jaderná a fyzikálně inženýrská**
Zadávající katedra/ústav: **Katedra fyziky**
Studijní program: **Jaderná a částicová fyzika**

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Produkce Λ_c baryonu v pp srážkách při energii 13,6 TeV

Název diplomové práce anglicky:

Λ_c production in pp collisions at 13.6 TeV

Pokyny pro vypracování:

1. Aktuální měření půvabných hadronů v pp srážkách na LHC.
2. Rekonstrukce půvabného baryonu Λ_c v pp srážkách při energii 13,6 TeV pomocí softvéru O2.
3. Optimalizace výběrových kritérií pro selekci Λ_c baryonu s využitím strojového učení.
4. Stanovení hrubého výtěžku Λ_c baryonu v pp srážkách při energii 13,6 TeV.

Seznam doporučené literatury:

- [1] ALICE Collaboration, Eur. Phys. J. C 84 (2024) 813.
- [2] A. Andronic et al., Eur.Phys.J.C 76 (2016) 3, 107.
- [3] ALICE Collaboration, Phys.Rev.Lett. 128 (2022) 1, 012001.
- [4] ALICE Collaboration, Phys.Lett.B 829 (2022) 137065.
- [5] Chen and Guestrin, arXiv: 1603.02754.

DECLARATION

I, the undersigned

Student's surname, given name(s): Žertová Karla
Personal number: 502580
Programme name: Nuclear and Particle Physics

hereby declare that the Master's Thesis titled

Lambda_c production in pp collisions at 13.6 TeV

is my own independent work and that I have declared all sources of information used in accordance with the Methodological Guideline for adhering to ethical principles when elaborating an academic final thesis and the Framework Rules for the use of artificial intelligence at CTU for study and pedagogical purposes in Bachelor and continuing Masters studies.

I declare that I have used artificial intelligence tools during the preparation and writing of the final thesis.
I've verified the generated content. I confirm that I am aware that I am fully responsible for the content of the final thesis.

In Prague on 04.05.2026

Bc. Karla Žertová

.....
student's signature

Acknowledgements

I would like to sincerely thank my supervisor, Ing. Katarína Gajdošová Křížková, Ph.D., for all her help during my studies and for the time she devoted to consultations and discussions of my work. My deepest thanks go to my husband Filip for his infinite patience with me and my work, for always being there for me, and for all his love. I am also grateful to my best friend Lída for making the years we spent together at university truly unforgettable. Last but not least, I want to express my gratitude to my family, who have always supported me in doing what I love, namely my brothers Adam and Max, my father Zdeněk, and my mum Alice, whose name made the choice of the ALICE Collaboration an obvious one.

Bc. Karla Žertová

<i>Název práce:</i>	Produkce Λ_c^+ baryonu v pp srážkách při energii 13,6 TeV
<i>Autor:</i>	Bc. Karla Žertová České vysoké učení technické v Praze, Fakulta jaderná a fyzikálně inženýrská, Katedra fyziky
<i>Studijní program:</i>	Jaderná a částicová fyzika
<i>Vedoucí práce:</i>	Ing. Katarína Krížková Gajdošová, Ph.D. České vysoké učení technické v Praze, Fakulta jaderná a fyzikálně inženýrská, Katedra fyziky
<i>Akademický rok:</i>	2025/2026
<i>Počet stran:</i>	94
<i>Klíčová slova:</i>	Produkce těžkých kvarků, hadronizace půvabného kvarku, baryon Λ_c^+ , selekce pomocí strojového učení, Boosted Decision Tree

Abstrakt

Tato diplomová práce se zabývá rekonstrukcí půvabného baryonu Λ_c^+ v rozpadovém kanálu $\Lambda_c^+ \rightarrow pK^-\pi^+$ s využitím Run 3 proton–protonových srážek zaznamenaných experimentem ALICE při energii $\sqrt{s} = 13,6$ TeV. Analýza byla provedena v rozsahu příčné hybnosti $1 < p_T < 24$ GeV/c. Pro zlepšení oddělení kandidátů Λ_c^+ od kombinatorického pozadí byla vytvořena, natrénována a následně aplikována selekce založená na strojovém učení s algoritmem Boosted Decision Tree. Vybraní kandidáti byli analyzováni pomocí rozdělení invariantní hmotnosti v jednotlivých intervalech p_T , ze kterých byly extrahovány surové výtěžky. Získaná spektra invariantní hmotnosti ukazují viditelný signál Λ_c^+ , což naznačuje, že použitá selekce pomocí strojového učení je účinná a vhodná pro další rozvoj této analýzy.

<i>Title:</i>	Λ_c^+ Production in pp Collisions at 13.6 TeV
<i>Author:</i>	Bc. Karla Žertová Czech Technical University in Prague, Faculty of Nuclear Sciences and Physical Engineering, Department of Physics
<i>Degree program:</i>	Nuclear and Particle Physics
<i>Supervisor:</i>	Ing. Katarína Krížková Gajdošová, Ph.D. České vysoké učení technické v Praze, Fakulta jaderná a fyzikálně inženýrská, Katedra fyziky
<i>Academic year:</i>	2025/2026
<i>Number of pages:</i>	94
<i>Keywords:</i>	Heavy-flavour production, charm-quark hadronisation, Λ_c^+ baryon, machine-learning selection, Boosted Decision Tree

Abstract

This thesis presents the reconstruction of the charm baryon Λ_c^+ in the decay channel $\Lambda_c^+ \rightarrow pK^-\pi^+$ using Run 3 proton–proton collision data recorded by the ALICE experiment at $\sqrt{s} = 13.6$ TeV. The analysis was performed in the transverse-momentum range $1 < p_T < 24$ GeV/ c . A machine-learning based selection using a Boosted Decision Tree algorithm was created, trained, and then applied to improve the separation of Λ_c^+ candidates from the combinatorial background. The selected candidates were analysed through invariant-mass distributions in individual p_T intervals, from which the raw yields were extracted. The obtained invariant-mass spectra show a visible Λ_c^+ signal, indicating that the machine-learning selection is efficient and suitable for further development of this analysis.

Contents

Introduction	10
1 Theoretical Background	12
1.1 Quantum Chromodynamics (QCD)	12
1.2 Study of QGP	13
1.2.1 Formation of the Quark–Gluon Plasma	13
1.2.2 The QCD Phase Diagram	14
1.2.3 Space–Time Evolution of Heavy-Ion Collisions	15
1.3 Prompt and Non-Prompt Hadrons	20
1.4 Observables in Heavy-Ion Collisions	20
1.4.1 Production Cross Sections	20
1.4.2 Yield Ratio	21
1.4.3 Fragmentation Fraction	22
1.4.4 Nuclear Modification Factor	22
2 Current Measurements of Charm Hadrons	24
2.1 Production Cross Section	24
2.2 Baryon-to-meson Ratio	26
2.3 Fragmentation Fraction	29
3 Charm-Hadron Reconstruction in ALICE	33
3.1 The ALICE Experiment	33
3.2 O^2 Analysis Framework	34
3.3 Data Structure	35
3.4 Heavy-flavour Analysis in O^2	36
3.5 Hyperloop	38

4	Machine Learning	39
4.1	Boosted Decision Trees (BDTs)	41
4.2	MLHEP Framework	42
4.3	Application of the Trained Model	44
5	Λ_c^+ Reconstruction and Selection	46
5.1	Datasets	46
5.2	Selection Variables	47
5.2.1	Rectangular Pre-selection	48
5.2.2	Training Variables	48
5.3	BDT-based Candidate Selection	59
5.3.1	Training and Test Sets	60
5.3.2	Training ML Model	60
5.3.3	Performance Evaluation	62
5.3.4	Working Point Optimization	65
6	Analysis	69
6.1	Invariant-mass Fit Procedure	70
6.2	Signal Significance	72
6.3	Signal-to-background Ratio	72
6.4	Mass Peak Stability	73
6.5	Raw Yield Extraction	74
6.6	Corrections	76
6.6.1	Acceptance Times Efficiency Correction	76
6.6.2	Feed-down Correction	77
6.6.3	Cross-section Normalisation	78
6.7	Systematic Uncertainties	79
	References	84

Introduction

The study of heavy-flavour particles is an important part of high-energy nuclear physics, because charm and beauty quarks provide information about both the hard partonic processes and the later stages of hadron formation. In particular, charm hadrons are useful probes of the production and hadronisation of charm quarks in different collision systems. Measurements performed in recent years have shown that charm-baryon production in hadronic collisions is not fully described by models based on fragmentation fractions from leptonic collisions. This makes the reconstruction of charm baryons, such as the Λ_c^+ , an important topic for testing our understanding of charm-quark hadronisation.

This thesis focuses on the reconstruction of the Λ_c^+ baryon in proton–proton collisions at $\sqrt{s} = 13.6$ TeV recorded by the ALICE experiment during LHC Run 3. The studied decay channel is

$$\Lambda_c^+ \rightarrow pK^-\pi^+.$$

The reconstruction of this particle is experimentally challenging because of its short lifetime, three-particle decay topology, and large combinatorial background. For this reason, a machine-learning based selection is used in this work to improve the separation of signal candidates from background candidates.

The main goal of the thesis is to train and apply a Boosted Decision Tree model for Λ_c^+ candidate selection and to study the reconstructed signal in different transverse-momentum intervals. The performance of the trained model is evaluated using training and testing distributions, score distributions, efficiency-related plots, and the final invariant-mass spectra. After the selection, the raw yield is extracted from invariant-mass fits and studied together with the signal significance, signal-to-background ratio, and mass-peak stability. The analysis represents a first step towards a full production measurement, for which acceptance and efficiency corrections, feed-down corrections, and systematic uncertainties would still have to be evaluated.

The thesis is organised as follows. Chapter 1 gives the theoretical motivation for the analysis. It introduces QCD, the quark–gluon plasma, the role of heavy-flavour particles, and the observables commonly used in heavy-ion and heavy-flavour measurements. Chapter 2 summarises current measurements of charm-hadron production, with emphasis on charm-baryon production, baryon-to-meson ratios, and fragmentation fractions. These results motivate why Λ_c^+ reconstruction is relevant for understanding charm-quark hadronisation.

Chapter 3 describes the experimental and software environment used in the analysis. It introduces the ALICE detector, the ALICE analysis framework, the structure of Run 3 data, the heavy-flavour reconstruction workflow, and the use of Hyperloop. Chapter 4 explains the machine-learning methods used in this work, especially Boosted Decision Trees and the MachineLearningHEP framework. Chapter 5 then presents the practical reconstruction and selection of Λ_c^+ candidates, including the datasets, pre-selection, training samples, input variables, model performance, and working-point optimisation.

Chapter 6 contains the final analysis of the selected candidates. It describes the invariant-mass fitting procedure, the extraction of the raw yield, and the study of the main performance quantities as a function of p_T . The correction procedure and systematic uncertainties are also discussed as necessary future steps towards a corrected production measurement. The thesis ends with a summary of the obtained results and an outlook for further improvements of the analysis.

Chapter 1

Theoretical Background

This chapter introduces the theoretical framework and experimental context relevant for this thesis, which focuses on heavy-flavour production in ultra-relativistic collisions. We begin with Quantum Chromodynamics (QCD) and its limitations, motivate the study of quark–gluon plasma (QGP) and finally describe the key observables used to probe the medium.

1.1 Quantum Chromodynamics (QCD)

The strong interaction is one of the four fundamental forces of nature and is responsible for binding quarks into hadrons and nucleons into atomic nuclei. Its theoretical description is given by Quantum Chromodynamics (QCD), which constitutes one of the fundamental pillars of the Standard Model of particle physics [1].

QCD is a non-Abelian gauge field theory based on the local symmetry group $SU(3)_c$ [2, 1]. In this context, the term non-Abelian means that gluons, which are the mediators of the strong interaction, carry colour charge and can therefore interact with each other. The group $SU(3)_c$ describes the symmetry connected with the three possible colour charges of quarks, usually labelled as red, green, and blue. Since $SU(3)$ has eight generators, the corresponding gauge theory contains eight gluon fields, which mediate the strong interaction between coloured particles. The requirement that this symmetry is local means that the theory must remain unchanged under colour rotations that can be different at each point in space and time. This local symmetry requirement leads to the introduction of the gluon fields. A distinctive property of QCD is asymptotic freedom, arising from the non-Abelian structure of the theory [2]. As a consequence, the strong coupling constant α_s decreases with increasing momentum transfer Q^2 , meaning that quarks and gluons interact often but weakly at short distances and high energies [1].

The scale dependence of α_s naturally separates QCD phenomena into two regimes [1]. At large momentum transfer, corresponding to high values of Q^2 , the strong coupling α_s becomes small and perturbative QCD (pQCD) can be applied. The quantity Q^2 denotes the squared four-momentum transfer of the interaction. In this regime,

physical observables can be calculated using a perturbative expansion in powers of α_s and therefore the processes within are well understood [3]. In contrast, at low momentum transfer (small Q^2), the coupling becomes large and the perturbative approach breaks down. In the non-perturbative regime of QCD, the strong interaction cannot be fully described using analytical calculations from first principles. Theoretical information in this regime can be obtained from lattice QCD, where the theory is solved numerically on a discrete space-time lattice [1]. Lattice QCD provides important results on the thermodynamic properties of strongly interacting matter, such as the equation of state and the transition from hadronic matter to the deconfined phase. However, it does not directly describe the full space-time evolution of a heavy-ion collision [4]. Therefore, experimental studies are essential for understanding the behaviour of strongly interacting matter under extreme conditions [4].

The non-perturbative regime governs essential features of the strong interaction, such as confinement and hadronisation. Confinement is a fundamental property of QCD stating that colour-charged particles, such as quarks and gluons, cannot be observed as isolated free particles and must ultimately form colour-neutral hadronic states [1]. During hadronisation, the coloured partons produced in the collision evolve into colour-neutral particles, such as mesons and baryons, which can then be measured in the detector. This process can occur through fragmentation, where a high-energy parton produces a shower of hadrons, or through recombination, where nearby quarks combine directly to form a hadron [5]. A particularly important system in this context is the quark–gluon plasma (QGP), a state of matter in which quarks and gluons are no longer confined inside individual hadrons. Such a medium can be created for a very short time in ultra-relativistic heavy-ion collisions [4].

In order to connect theoretical predictions with measurable observables, a combination of lattice QCD calculations, phenomenological models, and experimental measurements is needed. Lattice QCD provides first-principles theoretical input on the thermodynamic properties of strongly interacting matter, while phenomenological models describe the formation, expansion, cooling, and hadronisation of the QGP [4]. These models contain several assumptions and parameters, which must be constrained by experimental data. By comparing model predictions with measurements, it is possible to test different descriptions of the medium and to extract information about the properties of QCD matter under extreme conditions.

1.2 Study of QGP

1.2.1 Formation of the Quark–Gluon Plasma

The study of strongly interacting QCD matter can be performed by creating systems of quarks and gluons with large energy density. If such a system is heated to temperatures above approximately 150–160 MeV, it can undergo a transition to the quark–gluon plasma (QGP) [1]. At LHC energies, this occurs at close to zero net baryon density, corresponding to a baryochemical potential μ_B close to zero. The

net baryon density is defined as the difference between baryon and antibaryon densities at midrapidity [6]. Such conditions are achieved in ultra-relativistic heavy-ion collisions at large accelerator facilities such as the LHC [4]. Similar conditions existed approximately 10^{-5} s after the Big Bang, making heavy-ion collisions a unique opportunity to recreate and study an early stage of the Universe [4].

The study of QGP addresses fundamental questions about the behaviour of strongly interacting matter under extreme conditions [4]. It provides insight into the evolution of the early Universe and offers a laboratory for testing QCD in a regime where collective and non-perturbative phenomena dominate. In this regime, the behaviour of strongly interacting matter cannot be described only in terms of independent particle interactions, because the produced system evolves collectively and its properties are governed by long-range, many-body effects. As the simplest known form of complex quantum matter, QGP enables the refinement of theoretical models, the experimental validation of QCD-based calculations, and the extraction of medium parameters that cannot be computed directly from first principles. A primary role in the study of QGP is played by lattice QCD [1].

1.2.2 The QCD Phase Diagram

The QCD phase diagram shown in Fig. 1.1 presents the phases of strongly interacting matter as a function of the temperature in MeV and the baryochemical potential μ_B . Below the transition line, partons (quarks and gluons) are confined inside hadrons, forming ordinary matter. Above this line lies the QGP phase. The green line and band indicate the μ_B region accessible to Lattice QCD calculations. Along this line lies the crossover temperature, where the transition between hadronic matter and QGP takes place smoothly. The open points represent experimental results for the chemical freeze-out temperature, at which the relative abundances of produced hadrons become fixed. The diagram also marks the position of ordinary nuclear matter, corresponding to the conditions under which protons and neutrons are bound in stable nuclei [1].

At the LHC, the phase diagram is probed close to $\mu_B = 0$, because particles and antiparticles are produced in nearly equal amounts at LHC energies [4]. The most recent measurements show that the critical temperature under these conditions is around $T_c \approx 155$ MeV [1]. Other experiments study the crossover region at higher μ_B using lower collision energies, where the collision is less transparent. Such experiments include the STAR experiment at the RHIC accelerator in the USA and the CBM experiment at FAIR in Germany. Their goal is to understand how strongly interacting matter behaves under different conditions and also to search for the critical point, where the smooth crossover transition is expected to change into a first-order phase transition, meaning a sharp non-continuous change between phases.

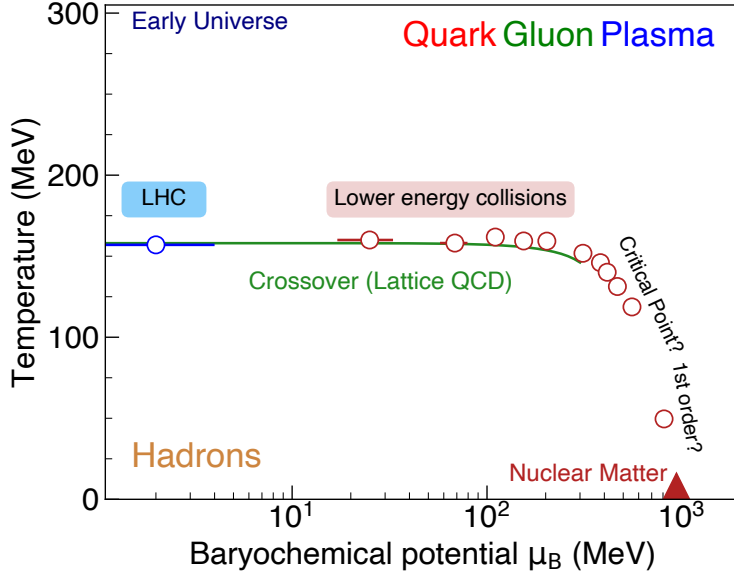


Figure 1.1: A schematic representation of the QCD phase diagram. The region accessible at LHC energies is indicated. Above the transition line lies the QGP phase, while below it hadronic matter dominates [4].

1.2.3 Space–Time Evolution of Heavy-Ion Collisions

The evolution of a heavy-ion collision proceeds through several stages, illustrated schematically in Fig. 1.2. It begins with the initial state, where excited partons transfer energy in the overlap region of the collision, defined by the impact parameter b . The impact parameter represents the distance between the centres of the colliding nuclei. For small values of b , the overlap region between the two nuclei is large. As a result, many nucleons participate in the collision, and the number of binary nucleon–nucleon interactions is also large. These quantities are denoted as N_{part} and N_{coll} , respectively, where N_{part} is the number of nucleons that undergo at least one inelastic interaction and N_{coll} is the total number of individual nucleon–nucleon collisions. In this stage of inelastic nucleon–nucleon interactions, a variety of QCD processes occur, involving up to large Q^2 momentum transfer. Softer interactions, corresponding to lower values of Q^2 , contribute to the initial matter distribution in the overlap region. Event-by-event fluctuations in the positions of the nucleons and in the deposited energy lead to an irregular initial density profile of the produced medium. At the same early stage, hard processes with large Q^2 produce heavy quarks and high-energy partons, which later form jets. Since these probes are produced before the QGP is fully formed and then propagate through the medium, they provide information about its properties [4].

At the next stage of smaller momentum transfer, the system evolves into a pre-equilibrated parton gas. This is followed by the creation of even softer partons, which enable the formation of the QGP phase. The QGP behaves as a strongly coupled, low-viscosity liquid [4], meaning that its constituents interact strongly and the medium can develop collective motion. This collective motion is driven by pressure gradients formed in the early stages of the collision and is commonly described in

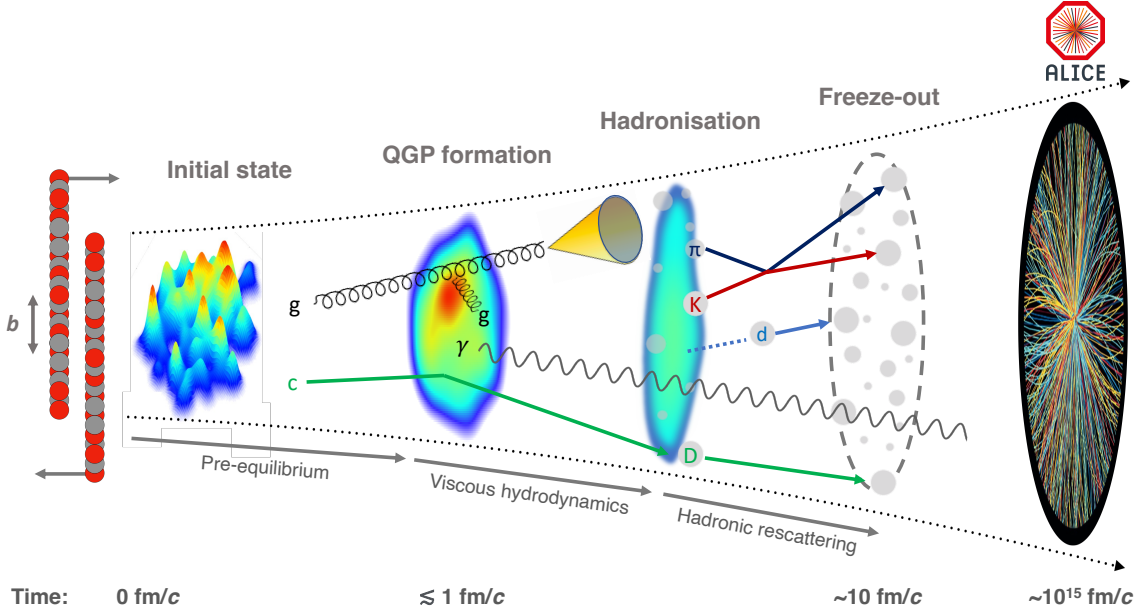


Figure 1.2: Spacetime diagram of the evolution of a heavy-ion collision, showing the stages leading to QGP formation and hadronisation. The impact parameter b corresponds to the distance between the centres of the colliding nuclei and is closely related to the number of participating nucleons [4].

terms of hydrodynamic flow, illustrated in Fig. 1.3. Radial flow develops because the pressure is larger in the centre of the medium than near its outer regions. This pressure gradient pushes the medium outward and leads to collective expansion. The rate of this expansion is influenced by the viscosity of the QGP, which describes how strongly the medium resists deformation during its evolution. Anisotropic flow arises when the pressure gradients are not the same in all directions. This directional dependence originates from spatial anisotropies in the initial collision geometry, for example in non-central heavy-ion collisions [4].

In addition to this collective bulk evolution, heavy quarks and high-momentum partons produced in the earliest hard scatterings propagate through the QGP. As they interact with the quarks and gluons in the medium, they lose part of their energy. These energy-loss effects can therefore be used to study how transparent or opaque the QGP is to energetic probes. This energy loss occurs mainly via two mechanisms: collisional energy loss, resulting from elastic scattering with medium constituents, and radiative energy loss, where gluons are emitted during the propagation of the parton. The average energy loss per unit length can be expressed as

$$\frac{dE}{dx}, \quad (1.1)$$

which depends on the properties of the medium and on the mass and energy of the propagating parton.

Due to the so-called dead-cone effect, heavy quarks are expected to lose less energy than light quarks and gluons [7]. The dead-cone effect describes the suppression of

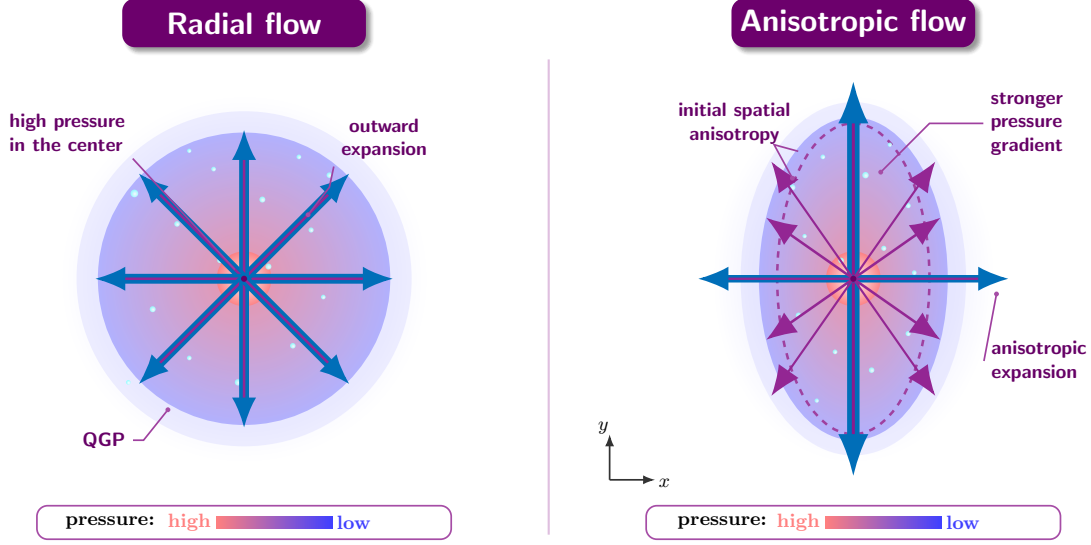


Figure 1.3: Schematic illustration of the development of anisotropic flow in a non-central heavy-ion collision. The initial spatial asymmetry is transformed into momentum-space anisotropy via the hydrodynamic response of the QGP.

gluon radiation at small angles with respect to the direction of a heavy quark. This suppression occurs inside an angular cone approximately given by

$$\theta < \frac{m}{E}, \quad (1.2)$$

where m and E are the mass and energy of the heavy quark. Since gluon radiation is dominated by soft emissions at small angles, the suppression of radiation inside this cone removes an important part of the radiation phase space. As a result, heavier quarks lose less energy through gluon radiation than light quarks. This effect is illustrated in Fig. 1.4. Gluon radiation is suppressed inside the dead cone and is therefore emitted mainly at angles larger than the dead-cone angle defined in Eq. 1.2. Since this angle depends on the ratio between the quark mass and energy, the dead cone becomes wider as the charm quark loses energy.

High-energy partons can initiate parton showers, in which the original parton radiates further quarks and gluons. Experimentally, a jet is reconstructed as a collimated group of final-state particles, interpreted as the observable remnant of an energetic parton. The exact set of particles included in a jet depends on the chosen jet reconstruction algorithm and its parameters. In Fig. 1.2, this process is indicated by the yellow cone. These jets, together with heavy quarks, represent hard probes of the medium because they are produced at the earliest stages of the collision and then propagate through the evolving QGP.

When an energetic parton travels through the QGP, it can interact with the medium and lose part of its energy. This phenomenon is known as jet quenching. It leads to a modification of the final reconstructed jet, which can appear with lower energy, a broader angular structure, or a modified particle composition compared to a jet produced in vacuum or in pp collisions. Jet quenching is therefore an important

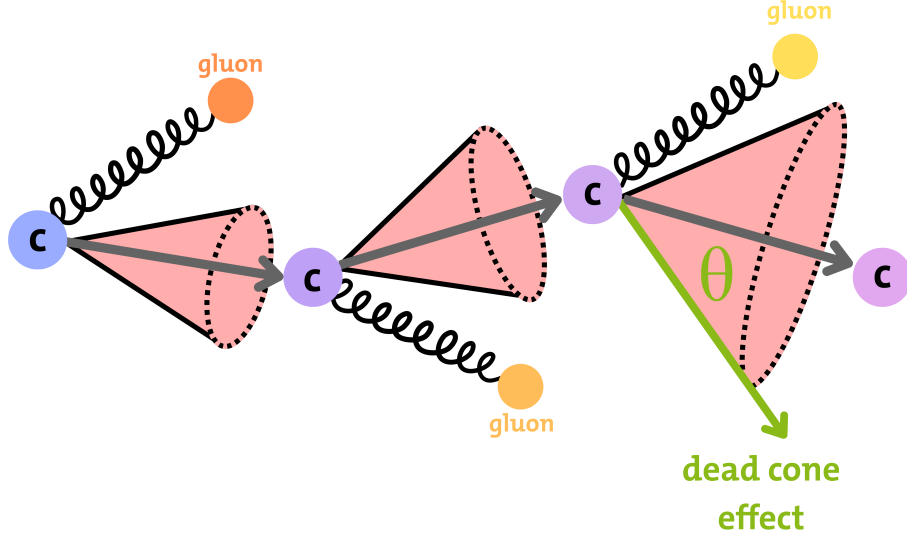


Figure 1.4: Schematic illustration of the dead-cone effect for a charm quark. Gluon radiation is suppressed within a cone around the quark direction, defined by the dead-cone angle θ .

signature of parton energy loss in the QGP and provides information about the density and transport properties of the medium. A schematic illustration of this effect is shown in Fig. 1.5. A hard scattering produces two approximately back-to-back energetic partons. The parton that crosses only a short path through the QGP is only weakly modified, while the parton that travels through a longer path in the medium loses more energy and forms a quenched jet.

Once the QGP cools below the transition temperature T_c , the hot and dense matter hadronises. At LHC energies, the transition from the QGP to hadronic matter is expected to be a smooth crossover rather than a sharp phase transition. As a result, different regions of the system hadronise at different times. The hard partons in jets fragment and hadronise in the same way as in elementary collisions, but partons with lower momenta can combine into hadrons via coalescence, meaning they combine with partons that share similar position and momentum.

As the hadrons materialise out of partons, primary pions (π), kaons (K), and protons (p) are created. They may still have enough energy to undergo inelastic interactions, which can create or destroy some particle species. This phase ends with chemical freeze-out at temperature T_{ch} , fixing the number of particles in each species. This can be followed by elastic scattering, which is halted at the kinetic freeze-out temperature T_{kin} . At this point, in addition to the particle species, also the momenta become fixed. These particles then travel towards the ALICE detector, where they are measured $\sim 10^{15}$ fm/ c after the initial collision.

The most commonly used collision system for heavy-ion studies at the LHC is lead-lead (Pb-Pb) collisions, where particle production takes place in the presence of a strongly interacting medium. After hadronisation, the produced hadrons can be described using statistical hadronisation models, which are connected to the non-perturbative regime of QCD [4]. In addition, the LHC also provides small sys-

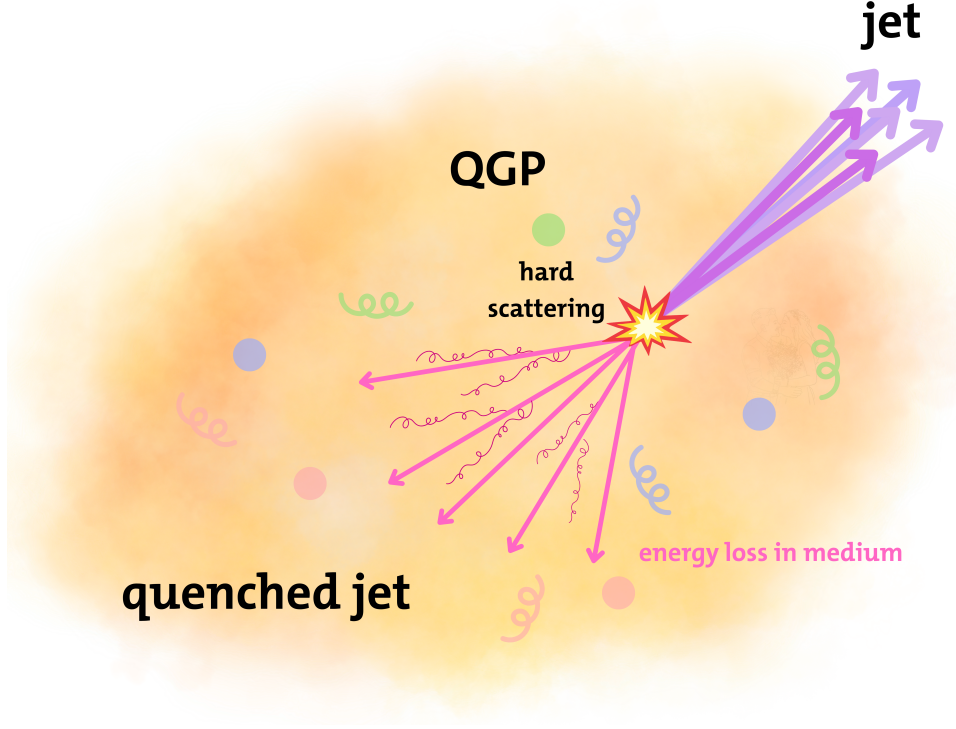


Figure 1.5: Schematic illustration of jet quenching in a heavy-ion collision.

tem collisions, such as proton–proton (pp) and proton–lead (p–Pb) collisions. In these smaller systems, the formation of a large QGP medium is not expected in the same way as in Pb–Pb collisions, and they are therefore often used as reference systems [8].

Particle production in pp collisions differs from that in heavy-ion collisions because the colliding system is much smaller and no large, hot, and dense QGP medium is expected to be formed. Therefore, pp collisions provide a reference for particle production in elementary hadronic interactions, without strong final-state effects from an extended medium. By comparing particle production in heavy-ion collisions with the corresponding pp reference, scaled by the number of binary nucleon–nucleon collisions, one can study whether the observed yields follow a simple superposition of independent pp-like interactions or whether they are modified by the presence of the QGP. In addition, p–Pb collisions play a complementary role, providing a reference system in which cold-nuclear-matter (CNM) effects are expected, but without the formation of a large QGP medium [9, 10]. CNM effects include, for example, modifications of the parton distribution functions inside nuclei compared to free protons, multiple scattering of incoming partons in the nucleus, the Cronin effect, parton energy loss in cold nuclear matter, and possible absorption or breakup of the produced particles in the nuclear environment. The Cronin effect refers to an enhancement of particle production at intermediate transverse momentum, caused by multiple scatterings before the hard interaction [9].

These collision systems are useful for distinguishing general effects of hadronic and nuclear collisions from effects caused by the additional QGP medium [4]. For this

reason, when an observable is measured in Pb–Pb collisions, corresponding measurements in pp and p–Pb collisions are needed as a comparison. If a certain effect is seen mainly in Pb–Pb collisions, but not in the reference systems, it can be interpreted as an indication of the presence and influence of the QGP [4].

1.3 Prompt and Non-Prompt Hadrons

In heavy-flavour measurements, it is important to distinguish between prompt and non-prompt charm hadrons. Prompt charm hadrons are produced directly from charm quarks created in the initial hard partonic scatterings, or from the strong decays of excited charm states [4]. They therefore carry information about charm-quark production, hadronisation, and the interaction of charm quarks with the medium. Examples of prompt charm hadrons are prompt D^0 , D^+ , D_s^+ mesons and prompt Λ_c^+ baryons.

Non-prompt charm hadrons, on the other hand, are not produced directly from charm quarks created in the initial hard scattering. Instead, they originate from weak decays of beauty hadrons, because beauty quarks can decay into charm quarks [4]. For example, a non-prompt Λ_c^+ can be produced as a decay product of a beauty baryon, such as $\Lambda_b^0 \rightarrow \Lambda_c^+ + X$. Similarly, non-prompt D^0 mesons can originate from decays of beauty mesons, such as $B \rightarrow D^0 + X$.

Since beauty hadrons have their own lifetimes, non-prompt charm hadrons are typically produced further away from the primary collision vertex than prompt charm hadrons. This difference in displacement is one of the main experimental handles used to separate prompt and non-prompt contributions [10]. This classification is also used later in the analysis, where prompt Λ_c^+ , non-prompt Λ_c^+ , and combinatorial background are treated as separate categories [10].

1.4 Observables in Heavy-Ion Collisions

The properties of the quark–gluon plasma (QGP) can be accessed experimentally through a wide variety of observables. These observables probe different stages of the collision evolution and different aspects of the interaction of partons with the produced medium. The following sections introduce several key observables commonly used in heavy-flavour physics. These include production cross sections, yield ratios, fragmentation fractions and the nuclear modification factor.

1.4.1 Production Cross Sections

Production cross sections represent one of the most fundamental observables in high-energy collisions. They quantify the probability that a given process occurs and provide direct information about the production mechanisms of heavy quarks [11]. Charm quarks originate in the perturbative QCD regime. Therefore their initial

production can be calculated, allowing their subsequent modification by the medium to be quantified, by comparing the computed and measured production [8].

Production of hard probes in hadronic collisions can be described using the QCD factorisation framework [12]:

$$d\sigma_{AB\rightarrow h} = f_{i/A}(x_i, Q^2) \otimes f_{j/B}(x_j, Q^2) \otimes d\sigma_{ij\rightarrow q\bar{q}}(x_i x_j, Q^2) \otimes D_{q\rightarrow h}(z, Q^2), \quad (1.3)$$

where $d\sigma_{AB\rightarrow h}$ is the differential cross section for producing a hadron h in the collision of A and B , $f_{i/A}$ and $f_{j/B}$ are the parton distribution functions, $d\sigma_{ij\rightarrow q\bar{q}}$ is the hard-scattering cross section, and $D_{q\rightarrow h}(z, Q^2)$ is the fragmentation function, which must be determined experimentally [11]. It describes the successive stages of parton production and hadronisation.

Experimentally, both total and differential cross sections are studied. While the total cross section gives the overall production rate, differential cross sections provide information about the kinematic distributions of the produced particles. In particular, the differential production cross section can be written as

$$\frac{d\sigma}{dp_T} = \frac{1}{2\Delta p_T} \frac{N_{\text{raw}} f_{\text{prompt}}}{(\text{Acc} \times \varepsilon) \text{BR} \mathcal{L}_{\text{int}}}, \quad (1.4)$$

where N_{raw} is the extracted raw yield, f_{prompt} is the prompt fraction, $(\text{Acc} \times \varepsilon)$ is the acceptance-times-efficiency correction, BR is the branching ratio of the decay channel, and $\mathcal{L}_{\text{int}}$ is the integrated luminosity. The factor $1/2$ accounts for the averaging over particle and antiparticle yields.

Measurements of charm-hadron production cross sections in pp collisions have been reported by the ALICE experiment at several centre-of-mass energies [11, 8]. These results allow stringent tests of perturbative QCD predictions [11]. In heavy-ion collisions the pp and p-Pb cross section serves as a baseline for identifying medium-induced modifications of particle production [13].

1.4.2 Yield Ratio

Yield ratios of heavy-flavour hadrons complement absolute cross-section measurements by isolating hadronisation effects while reducing sensitivity to parton production uncertainties. Since the charm-quark production cross section is common for both hadrons, it cancels in the ratio together with several experimental systematic uncertainties. The observable therefore becomes primarily sensitive to the ratio of fragmentation fractions and provides a direct probe of the charm-quark hadronisation process [4]. This makes the ratios such as

$$\frac{\Lambda_c^+}{D^0}, \quad \frac{D_s^+}{D^0}$$

especially sensitive probes of the hadronisation process.

1.4.3 Fragmentation Fraction

Fragmentation fractions describe the probability that a charm quark hadronises into a specific charm-hadron species [5]. They are usually denoted as $f(c \rightarrow h_c)$, where h_c represents a charm hadron. They are closely related to the fragmentation function $D_{c \rightarrow h_c}(z, Q^2)$, which describes the distribution of the momentum fraction z carried by the produced hadron with respect to the original charm quark. In this sense, the fragmentation fraction corresponds to the integrated probability for producing a given hadron species and can be written schematically as

$$f(c \rightarrow h_c) \sim \int_0^1 D_{c \rightarrow h_c}(z, Q^2) dz. \quad (1.5)$$

Experimentally, fragmentation fractions are often estimated from measured production cross sections as

$$f(c \rightarrow h_c) = \frac{\sigma(\text{pp} \rightarrow h_c)}{\sum_i \sigma(\text{pp} \rightarrow h_{c,i})}, \quad (1.6)$$

where the denominator includes the production cross sections of all considered charm-hadron species [5].

Historically, fragmentation fractions were mainly obtained from measurements in relatively clean collision systems, such as e^+e^- and e -p collisions [1]. Since in these systems hadronisation occurs in a simpler environment than in hadronic collisions, these measurements were commonly used as an input for theoretical calculations. In many calculations, it was assumed that fragmentation fractions are universal, meaning that the probability for a charm quark to hadronise into a given charm-hadron species is independent of the collision system [1, 5]. However, recent measurements in pp and p-Pb collisions show that this assumption does not hold fully in hadronic environments [5, 10].

1.4.4 Nuclear Modification Factor

One of the most important observables used to quantify medium effects in heavy-ion collisions is the nuclear modification factor. In nucleus–nucleus collisions, it is denoted as R_{AA} . It compares particle production in nucleus–nucleus collisions with that in proton–proton collisions scaled by the number of binary nucleon–nucleon interactions [13]. It can be studied separately for prompt and non-prompt heavy-flavour particles, which allows one to compare the in-medium behaviour of charm and beauty production.

It is defined as

$$R_{AA}(p_T) = \frac{1}{\langle T_{AA} \rangle} \frac{dN_{AA}/dp_T}{d\sigma_{pp}/dp_T}, \quad (1.7)$$

where dN_{AA}/dp_T is the p_T -differential yield in nucleus–nucleus collisions, $d\sigma_{pp}/dp_T$ is the corresponding pp cross section, and $\langle T_{AA} \rangle$ is the nuclear overlap function related to the number of binary nucleon–nucleon collisions.

A value of $R_{AA} < 1$ indicates suppression of particle production in heavy-ion collisions, which is commonly interpreted as a consequence of energy loss of high-energy partons propagating through the QGP medium [13]. The same idea can also be applied to p-Pb collisions, where deviations of R_{pPb} from unity can indicate cold nuclear matter effects.

Chapter 2

Current Measurements of Charm Hadrons

A large number of new measurements of charm-hadron production have recently been reported by different experimental collaborations and research groups. In this work, however, we focus primarily on results obtained by the ALICE experiment. In particular, we concentrate on the currently active topic of the universality of fragmentation fractions. These fractions were previously assumed to be universal across different collision systems, but recent measurements indicate that this assumption may not hold. To understand this question, we will first examine the production cross section of charm hadrons, how it is calculated, and what information it provides about the charm-production process. Next, we will study the baryon-to-meson ratio, which offers additional insight into charm-quark hadronisation. Finally, we will present the most recent measurements of the fragmentation fraction and discuss their implications for the understanding of charm-hadron production.

Charm-hadron production is a central component of heavy-flavour measurements in high-energy collisions. Heavy flavour refers to hadrons containing second- and third-generation quarks, primarily charm and bottom, as the top quark does not hadronize. Charm hadrons contain at least one charm quark or antiquark and include mesons (D^0 , D^+ , D_s^+) and baryons (Λ_c^+ , Ξ_c , Ω_c). Owing to their finite lifetimes, they decay at displaced secondary vertices, making precise tracking and vertex reconstruction essential, as achieved in detectors such as ALICE.

2.1 Production Cross Section

The production cross section is a fundamental observable used to quantify how often a given particle is produced in a collision. In heavy-flavour measurements, it provides information about the production of charm hadrons in hadronic collisions. As described in more detail in Sec. 1.4.1, its value can be predicted using theoretical calculations together with phenomenological models, and it can also be measured experimentally. By comparing the predicted and measured production cross sections, it is possible to test how well the theoretical description and the modelling of particle

production describe the data.

Recently, in Ref. [10, 5], new measurements of charm-hadron production in ALICE experiment have been conducted, in both pp and p-Pb collisions. For the first time, it was possible to reconstruct baryons down to low p_T , where the majority of charm hadrons are produced, which provides new statistics for studying charm-hadron production. In the following, we discuss the production of charm mesons and baryons separately and then consider the total charm production cross section.

The reconstruction of charm mesons is experimentally simpler than that of charm baryons. This is mainly because charm baryons have a more complex decay topology, shorter lifetimes, and lower production rates, which makes their signal more difficult to separate from the combinatorial background. In Ref. [5], the production cross sections of the charm mesons D^0 , D^+ , D_s^+ , and D^{*+} were measured at midrapidity in pp collisions at $\sqrt{s} = 13$ TeV and compared with previous measurements at $\sqrt{s} = 5.02$ TeV and 7 TeV. The measured production of these charm mesons is, within uncertainties, described by perturbative QCD calculations combined with fragmentation functions extracted from e^+e^- collisions. This means that, for charm mesons, production in hadronic collisions appears to be compatible with the production picture obtained from leptonic collisions.

The situation becomes more interesting for charm baryons. In Ref. [10], ALICE performed the challenging reconstruction of Λ_c^+ baryons in pp and p-Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV. An important aspect of this paper is the introduction of the lowest- p_T reconstruction of Λ_c^+ baryons, which is experimentally difficult because of the large combinatorial background and the short lifetime of the Λ_c^+ . By combining this low- p_T measurement with the previously measured Λ_c^+ production at higher transverse momenta, extending up to $p_T = 10$ GeV/ c , a more complete set of measurements over a broad p_T range is obtained. This is important because it improves the determination of the full charm-hadron production, especially in the baryon sector, which is experimentally difficult to measure. At the same time, it allows a more detailed comparison of charm-baryon production in hadronic collisions with the production picture known from leptonic collisions.

Firstly, this comparison shows that the measured Λ_c^+ production is not fully described by calculations based on the assumed universality of charm-quark fragmentation. In contrast to charm mesons, whose production is generally described within uncertainties by calculations using fragmentation functions derived from e^+e^- collisions, the measured Λ_c^+ production is significantly higher than expected from such calculations. This indicates that the production of charm baryons in hadronic collisions is enhanced with respect to the conventional hadronisation picture. The disagreement suggests that charm-quark hadronisation in hadronic collisions may differ from that in leptonic collisions, particularly in the baryon sector. This enhancement of charm-baryon production is therefore an important result, and possible explanations will be discussed in the following section.

Secondly, improved measurements of charm baryons can be used together with charm-meson measurements to obtain the total charm-quark production cross section, a quantity that is otherwise usually accessed only through theoretical cal-

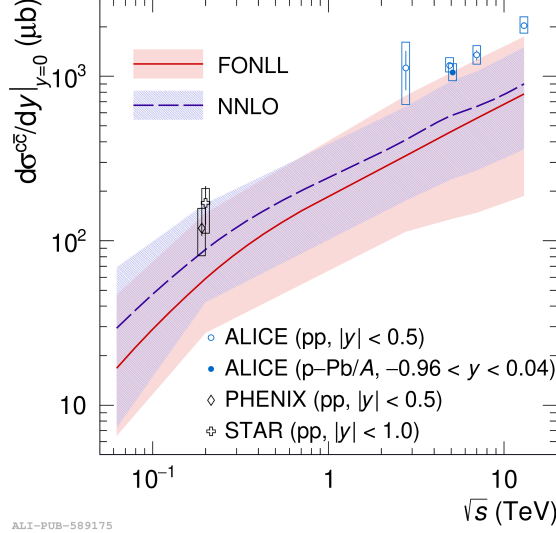


Figure 2.1: Total $c\bar{c}$ production cross section per unit of rapidity and per nucleon in pp [5] and p-Pb collisions, compared with measurements in pp collisions from PHENIX and STAR. The measurements are compared with FONLL and NNLO pQCD calculations [14].

culations. In this procedure, the contribution of each measured charm hadron is included, while the missing contribution from charm hadrons that were not measured directly has to be estimated. These missing contributions mainly correspond to charm-hadron states or decay modes that are experimentally difficult to access, for example because their production rates are small, their branching ratios are poorly known, their decay topologies are too complex, or their signal cannot be separated with sufficient precision from the combinatorial background. The total charm production cross section was obtained for both pp and p-Pb collisions. The results are shown in Fig. 2.1, where the pp measurements from [5] and the p-Pb measurements from [14] are presented as a function of the collision energy. The measurements are compared with pQCD calculations. Since the measured values lie close to the upper edge of the theoretical predictions, they provide important constraints on pQCD calculations of charm production.

2.2 Baryon-to-meson Ratio

To further study the baryon production enhancement over the predictions, we need to clearly understand where the increase in baryon production comes from. To make sure that the discrepancy does not originate from the $c\bar{c}$ production itself, it is useful to study the baryon-to-meson ratio. Since the charm-quark production cross section is common for both hadrons, it cancels in the ratio together with several experimental systematic uncertainties. The observable therefore becomes primarily sensitive to the ratio of the fragmentation fractions. Therefore, the baryon-to-meson ratio provides a direct probe of the charm-quark hadronisation process.

The charm baryon-to-meson ratio can be seen in Fig. 2.2, where the Λ_c^+/D^0 ratio is shown as a function of p_T in pp collisions at $\sqrt{s} = 5.02$ TeV from Ref. [10] and at $\sqrt{s} = 13.6$ TeV from Ref. [15], respectively. The measurements are compared with PYTHIA 8 with the Monash tune, which uses fragmentation fractions obtained from e^+e^- collisions. In both cases, the model significantly underestimates the measured Λ_c^+/D^0 yield ratio, especially at low and intermediate p_T . This indicates that fragmentation fractions measured in e^+e^- collisions are not sufficient to describe charm-hadron production in hadronic collisions. The data are reproduced only by models that predict an enhancement of baryon production through additional mechanisms. These include colour reconnection beyond the leading-colour approximation, feed-down from unobserved resonant charm-baryon states, or quark coalescence.

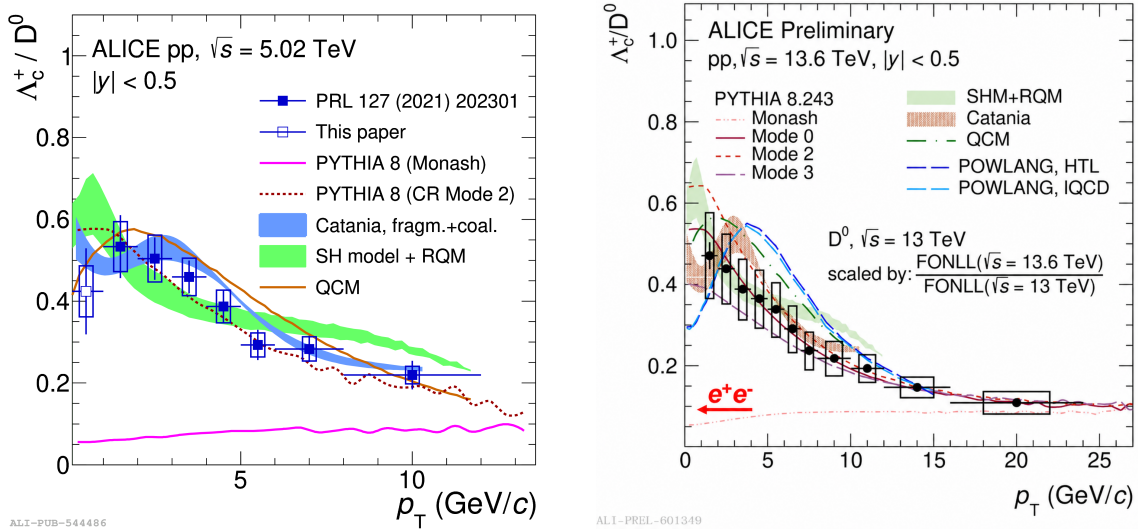


Figure 2.2: Λ_c^+/D^0 ratio as a function of p_T in pp collisions at $\sqrt{s} = 5.02$ TeV [10] and preliminary results at $\sqrt{s} = 13.6$ TeV [15], including comparisons with models.

Colour reconnection beyond the leading-colour approximation is a mechanism, in which partons produced in the collision are connected by colour strings, which later break and form hadrons. In the leading-colour approximation, only the simplest colour connections are considered. However, in a real hadronic collision, the colour structure can be more complicated. Before hadronisation, the colour strings may therefore reconnect in a different way. Some of these new colour configurations can favour the formation of baryons, because they bring together a charm quark and nearby light quarks in a colour topology suitable for producing a charm baryon such as Λ_c^+ .

Another possible explanation is feed-down from unobserved excited charm-baryon states. In this case, the collision does not produce only ground-state charm baryons, such as Λ_c^+ , but also heavier excited charm-baryon resonances. These excited states are difficult to measure directly because they may have small production rates or decay channels that are experimentally challenging to reconstruct. However, after being produced, they can decay into the ground-state charm baryons that are measured in the detector. As a result, the observed Λ_c^+ yield can be larger than expected if these additional sources are not fully included in the theoretical prediction.

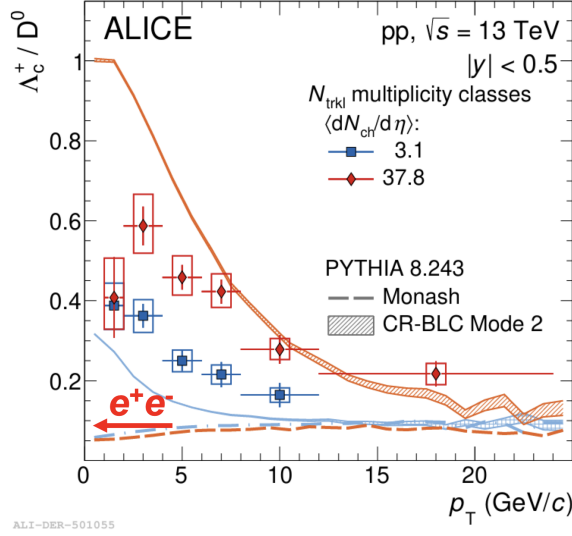


Figure 2.3: Λ_c^+/D^0 ratio as a function of p_T in pp collisions at $\sqrt{s} = 13$ TeV for two multiplicity classes [16].

A third mechanism is quark coalescence, also called recombination. In the usual fragmentation picture, a charm quark hadronises rather independently after radiating gluons, with additional quark–antiquark pairs produced during the non-perturbative hadronisation process. In the coalescence picture, instead, hadrons can be formed by the direct combination of quarks that are close to each other in position and momentum space. This mechanism is especially relevant in environments with a high density of produced quarks, such as heavy-ion collisions. In such an environment, a charm quark has a larger probability to find nearby light quarks and combine with them. Since a charm baryon is formed from one charm quark and two light quarks, this naturally increases the probability of charm-baryon production.

Although pp collisions are much smaller systems than heavy-ion collisions, they can still have a higher final-state particle density than the much simpler e^+e^- collision system. Therefore, mechanisms connected to the local density of produced quarks may also play a role in pp collisions, especially in high-multiplicity events. For this reason, it is useful to study charm-baryon production as a function of multiplicity in pp collisions and to compare low- and high-multiplicity events. Such a comparison is shown in Fig. 2.3, where the dependence of the baryon-to-meson ratio on p_T is presented for two different multiplicity intervals. Although the ratio varies across the individual p_T bins, the overall mean ratio remains compatible between the two multiplicity classes.

Together, these effects provide possible explanations for the larger charm-baryon production observed in experimental data compared to predictions based on fragmentation functions derived from e^+e^- collisions.

A similar measurement has been conducted to explore the beauty sector and to study the universality of observed baryon-to-meson enhancement across flavours. The Ref. [17] measures the dependence of different baryon-to-meson ratios on p_T , which can be seen in Fig. 2.4, where the ratios of non-prompt Λ_c^+/D^0 and prompt

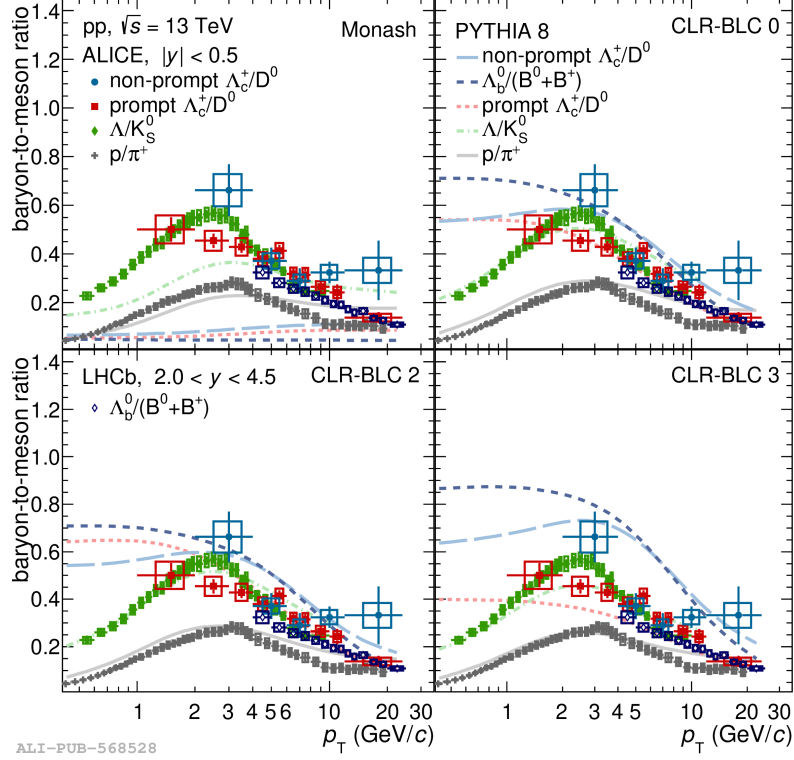


Figure 2.4: The ratios non-prompt Λ_c^+/D^0 , prompt Λ_c^+/D^0 , Λ/K_S^0 , and p/π^+ measured in pp collisions at $\sqrt{s} = 13$ TeV at midrapidity ($|y| < 0.5$) [17].

Λ_c^+/D^0 , Λ/K_S^0 , p/π^+ are shown. This measurement therefore covers beauty, charm, strange, and even light flavours. The measurement was performed at midrapidity in pp collisions at $\sqrt{s} = 13$ TeV. The baryon-to-meson ratios of the heavy-flavour hadrons show a similar p_T dependence as in the previous measurement. The ratios compared to the models are significantly higher than predicted by the model Monash tune, containing fragmentation fraction from e^+e^- collisions. Models that include additional baryon-production mechanisms are again in better agreement with the measured data, but more precise measurements are still needed. The similar behaviour observed across different flavour sectors suggests that the baryon enhancement may originate from a universal hadronisation mechanism rather than a process specific only to charm quarks.

2.3 Fragmentation Fraction

Lastly, we consider the fragmentation fraction, which describes the probability that a charm quark forms a given charm-hadron species. As discussed above, the comparison of charm production in leptonic and hadronic collisions shows that these fractions may not be universal in all collision systems. Such a modification would indicate that charm hadronisation in hadronic collisions is not purely vacuum-like, but can be influenced by the surrounding hadronic environment.

Fragmentation fractions were previously known mainly from e^+e^- and e -p collisions.

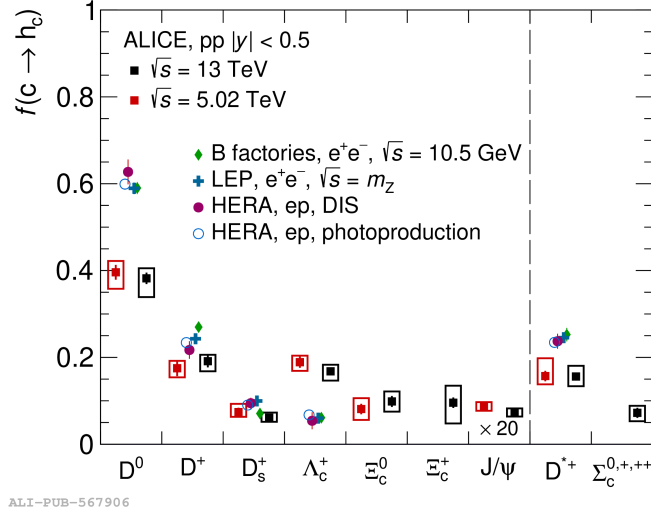


Figure 2.5: Fragmentation fraction for different charm hadrons with new ALICE measurements at energies $\sqrt{s} = 5.02$ and 13 TeV in pp collisions. The results are compared with measurements from leptonic collision systems [5].

New measurements in pp and p-Pb collisions at $\sqrt{s} = 5.02$ and 13 TeV therefore make it possible to determine them directly in hadronic systems. The method used to extract them from experimental data is described in Sec. 1.4.3. This became possible thanks to improved tracking performance and reconstruction techniques, especially for charm baryons. To determine the fragmentation fraction for a given hadron species, its production is compared with the total charm-hadron production, which requires measuring as many charm-hadron species as possible.

Such measurements have been performed and are shown in Fig. 2.5 and Fig. 2.6. Fig. 2.5 presents the fragmentation fraction for different charm hadrons, including new measurements by the ALICE experiment at $\sqrt{s} = 5.02$ and 13 TeV in pp collisions. The results are compared with measurements from e^+e^- and e -p collision systems and show that the fragmentation fractions differ for several charm hadrons. In particular, hadronic collision systems show an enhancement of charm-baryon production accompanied by a corresponding reduction in charm-meson production relative to e^+e^- and e -p collisions. This observation indicates a violation of the universality of fragmentation fractions across different collision systems. Furthermore, the comparison of the two collision energies suggests that the fragmentation fraction does not depend significantly on the collision energy, implying that the observed differences arise primarily from the type of collision system rather than from the energy itself [5].

The second figure, Fig. 2.6, shows a similar comparison, but now between pp and p-Pb collisions at the same energy of $\sqrt{s} = 5.02$ TeV. The measured values are compatible within uncertainties. This indicates that the fragmentation fraction is consistent between these two hadronic systems. The difference observed with respect to leptonic collisions is therefore likely related to the presence of hadronic matter in the collision environment, even though the p-Pb system has a larger system size and a higher charged-particle multiplicity density. This observation provides an

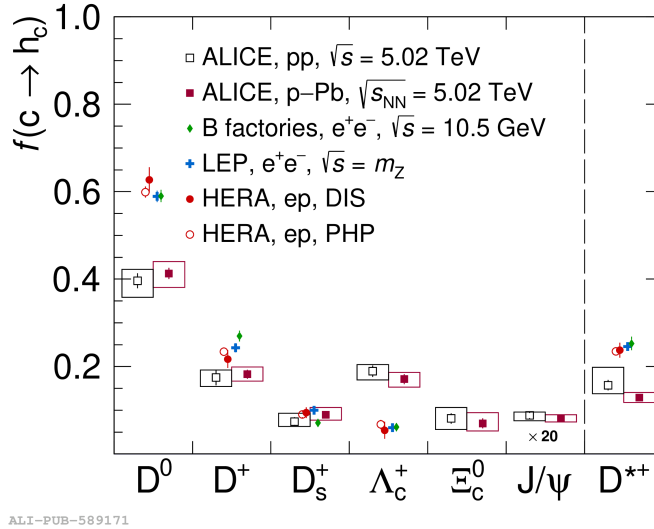


Figure 2.6: Fragmentation fraction for different charm hadrons with ALICE measurements at $\sqrt{s} = 5.02$ TeV in pp and p-Pb collisions [14].

important constraint regarding the origin of the fragmentation-fraction modification.

Recent ALICE preliminary measurements also extend these studies to the non-prompt fraction of charm hadrons, as shown in Figs. 2.7 and 2.8. The non-prompt fraction describes the fraction of reconstructed charm hadrons that originate from beauty-hadron decays. For the D^0 meson, shown in Fig. 2.7, the Run 3 measurement at $\sqrt{s} = 13.6$ TeV shows improved precision and a more detailed p_T dependence with respect to previous Run 2 results. The data appear to lie systematically below the PYTHIA 8 Monash prediction, but no large disagreement is observed within the current uncertainties. This suggests that the change in the charm-hadron fragmentation fractions in hadronic collisions is likely driven mainly by the enhanced production of charm baryons, rather than by a strong suppression of charm-meson production. For the Λ_c^+ baryon, shown in Fig. 2.8, the experimental uncertainties are still rather large, but the measurement appears to follow the same general trend observed in prompt charm-baryon production. In particular, the data tend to be higher than the PYTHIA 8 Monash tune, which is based on fragmentation fractions obtained from e^+e^- collisions. This again suggests that charm-baryon production in hadronic collisions may require a different hadronisation description than the one based on fragmentation fractions from leptonic collisions.

These measurements will help improve the modelling of hadronisation and enable more precise theoretical predictions, since the fragmentation fraction represents one of the fundamental ingredients of the total cross section for hadronic production. They provide an important step toward further studies of the quark-gluon plasma in Pb-Pb collisions. Measuring the fragmentation fraction in heavy-ion collisions will help quantify modifications of hadronisation driven by the formation of a quark-gluon plasma and will allow one to distinguish between effects originating from hadronic interactions and those caused by hot deconfined matter.

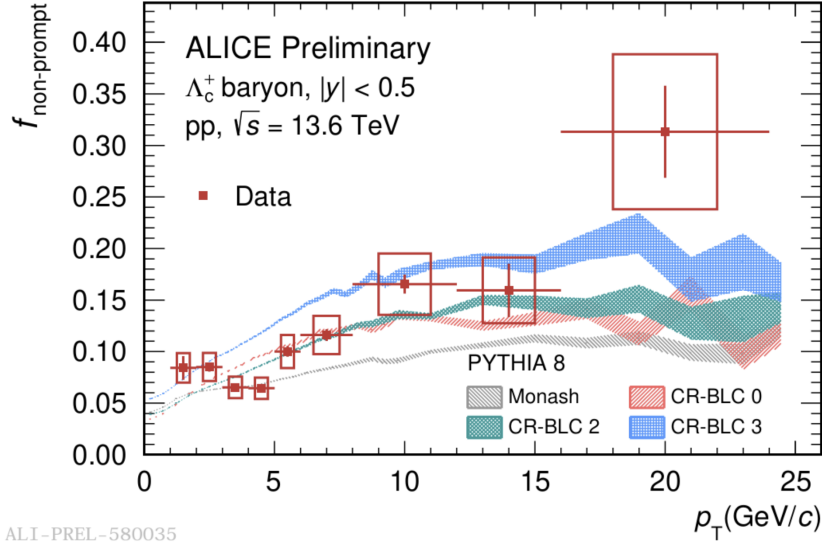


Figure 2.7: Non-prompt fraction of D^0 mesons measured by ALICE in pp collisions at $\sqrt{s} = 13.6$ TeV [16].

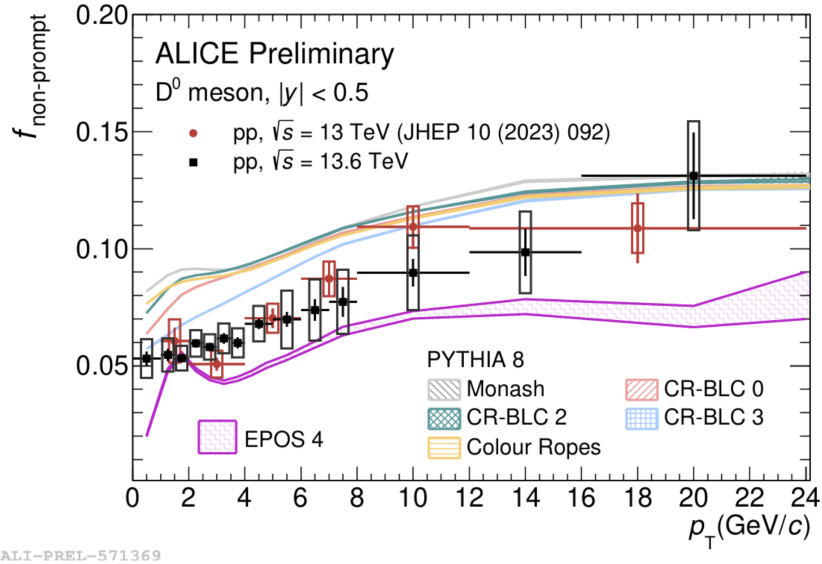


Figure 2.8: Non-prompt fraction of Λ_c^+ baryons measured by ALICE in pp collisions at $\sqrt{s} = 13.6$ TeV [16].

Chapter 3

Charm-Hadron Reconstruction in ALICE

In this chapter, the analysis procedure used to obtain charm-hadron measurements with the ALICE detector is introduced. First, the ALICE experiment is briefly described, together with the detector systems most relevant for this work. Then, the ALICE analysis framework is presented, including its main structural features and data organization. Next, the heavy-flavour analysis workflow is discussed, from track selection to charm-hadron candidate reconstruction and final analysis. Finally, the structure of the data used in this work is described, with emphasis on derived datasets and their role in efficient data processing.

3.1 The ALICE Experiment

The ALICE experiment is the only LHC detector specifically designed to study the properties of QGP. Its physics program focuses on measuring particle production over a wide transverse momentum (p_T) range, enabling access to multiple observables sensitive to different stages of the collision evolution. The detector is centred around a central barrel providing robust particle identification up to high transverse momenta ($p_T \approx 20 \text{ GeV}/c$), combined with a low momentum cut-off ($p_T \approx 0.15 \text{ GeV}/c$) and excellent primary and secondary vertex reconstruction capabilities. These features are essential for reconstructing heavy-flavour hadrons, which act as sensitive probes of the QGP medium [4]. A schematic overview of the ALICE detector after the LS2 (Long Shutdown 2) upgrades is shown in Fig. 3.1.

The two most important detectors for this analysis are the Time Projection Chamber (TPC) and the Time-Of-Flight system (TOF). The TPC, labelled as 15 in Fig. 3.1, is the main device in the ALICE central barrel for tracking charged particles and particle identification. It is a large cylindrical volume filled with a Ne/CO₂ gas mixture at atmospheric pressure, where ionisation electrons drift over distances of more than 2.5 m towards the end-plates equipped with proportional chambers and a readout system. This allows precise measurement of particle momentum using the curvature of tracks in the solenoidal magnetic field of 0.5 T, together with

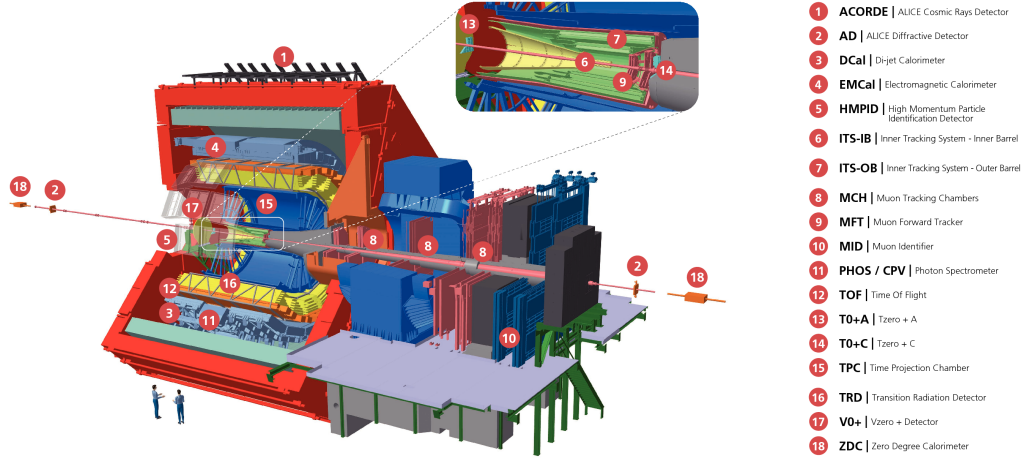


Figure 3.1: Schematic view of the ALICE detector for LHC Run 3 after the upgrades [18].

particle identification through specific energy loss (dE/dx). The detector covers a wide pseudorapidity range with excellent spatial and two-track resolution [19].

The TOF detector, labelled as 12 in Fig. 3.1, on the other hand, is built from Multigap Resistive Plate Chambers (MRPCs). Its double-stack strip design achieves a time resolution of around 85 ps. This performance enables clean identification of pions, kaons, and protons up to momenta of several GeV/c , complementing the tracking and energy-loss measurements of the inner detectors [20].

3.2 O^2 Analysis Framework

The ALICE experiment started its Run 3 data-taking period in 2022, after Long Shutdown 2 (LS2). During LS2, several major detector upgrades were implemented, including continuous readout and the ability to handle much higher data rates. These changes required a new software framework capable of processing the much larger Run 3 data volume. For this reason, the previous AliPhysics framework was replaced by the Offline–Online (O^2) software. The O^2 framework is written primarily in C++ and is designed for high-throughput reconstruction and analysis, including the reconstruction of rare probes such as charm hadrons.

One of the main structural changes introduced in O^2 is the use of an Apache Arrow-based data model. While AliPhysics followed a more object-oriented approach, O^2 stores the data in columnar tables. This means that different quantities, such as collision information, track parameters, and particle-identification information, are stored in separate but connected tables. The framework also uses a more track-centric structure, where tracks are the central objects and are linked to the collisions to which they belong. Only the most essential quantities are stored directly, while derived quantities are computed when needed. This reduces the size of the stored data and improves memory usage and compression efficiency.

The reconstructed data are organised into time frames, each of which can contain

multiple collisions. Within these time frames, tracks are associated with their corresponding collisions. The output is stored in ROOT format in Analysis Object Data files, commonly referred to as AO2D files [21]. These files contain the tables needed for later physics analysis and form the starting point for most O² analysis workflows.

Analyses in O² are built as modular workflows composed of individual tasks. Each task performs a specific part of the analysis chain, such as event selection, track selection, particle identification, candidate reconstruction, or the final physics analysis. For charm-hadron reconstruction, candidates are formed by combining selected tracks according to the expected daughter particles of a given decay channel. The individual tasks are connected into a pipeline, which makes the workflow flexible, configurable, and reproducible. A detailed description of the O² analysis structure and available utilities is given in Ref. [21].

O² analysis workflows can be executed either centrally through the Hyperloop train system or locally. Running an analysis *locally* means executing the workflow outside Hyperloop, for example on a personal computer or on a local computing cluster, using downloaded input AO2D files. Local running is mainly used for testing, debugging and smaller-scale studies.

3.3 Data Structure

As described above, the data used in the analysis are stored in table-like structures, where each table contains a well-defined set of physical quantities. These input data are not raw detector signals, but already reconstructed data products. The detector signals have already been processed into physics objects such as collisions, tracks, vertices, and particle-identification information. In the analysis workflow, the term reconstruction therefore refers mainly to the reconstruction of specific particle candidates, such as charm hadrons, from these already available objects. The data are typically stored in the AO2D format and form the starting point for most physics analyses.

In order to extract additional information from these data, one can run analysis tasks within the workflow. These tasks act as processing modules that read existing tables and produce new ones, enriching the dataset with derived quantities such as selection flags, reconstructed candidates, or calibration corrections. However, many analyses within a given working group require similar derived quantities. Running the same processing repeatedly on large datasets would be computationally inefficient.

For this reason, so-called *derived data* are produced. These are datasets obtained from the original AO2D by applying a predefined sequence of tasks, resulting in additional tables that can be reused by multiple analyses. Derived data are typically produced centrally within a working group and stored for common use, significantly reducing the need for repeated large-scale processing. At the same time, users can also create their own custom derived datasets by running specific workflows tailored to their analysis needs.

It is important to note that derived data do not always fully replace the original

dataset. In many cases, they contain only a subset of the information (or newly created tables), and access to the parent AO2D data may still be required to retrieve additional quantities not included in the derived tables.

Most common derived datasets are produced after a skimming step. In the heavy-flavour workflow, this means that instead of working with all reconstructed tracks and all possible track combinations, the analysis starts from a reduced set of pre-selected candidates. The skim creator identifies track combinations that are loosely compatible with heavy-flavour decays and stores only the essential information, namely the indices of the daughter tracks together with basic flags indicating the decay hypothesis. This approach significantly reduces the data volume and computational cost, while still preserving almost all potentially interesting candidates.

The resulting skimmed derived data serve as an efficient input for subsequent analysis steps, such as detailed candidate reconstruction, multivariate analysis training, or final signal extraction. For example, derived datasets produced for Λ_c^+ reconstruction can be directly used for machine learning training or for further refinement of selection criteria, while other derived workflows may provide inputs for calibration, quality assurance, or systematic studies.

3.4 Heavy-flavour Analysis in O²

Charm-hadron reconstruction in this thesis follows the Heavy-Flavour (HF) analysis workflow developed within the ALICE collaboration. The HF Physics Working Group is responsible for implementing and maintaining reconstruction strategies for heavy-flavour particles within the O² framework.

Run 3 brings significant improvements for HF analyses, mainly due to the upgraded Inner Tracking System (ITS2), which consists of seven layers and provides precise vertexing down to very low transverse momentum. In addition, continuous readout allows ALICE to record much larger data samples, which is especially important for rare processes such as Λ_c^+ production. At the same time, these improvements also introduce new challenges, including larger background, low signal-to-background ratios, and higher computational demands. This makes an efficient and carefully optimised reconstruction workflow essential.

The HF analysis workflow consists of several connected stages, illustrated in Fig. 3.2. Since charm hadrons decay before reaching the detector, they cannot be measured directly. Instead, they are reconstructed from their decay products. For example, in the decay channel $\Lambda_c^+ \rightarrow pK^-\pi^+$, the analysis searches for suitable combinations of three reconstructed tracks that are compatible with the proton, kaon, and pion hypotheses. These tracks are then used to reconstruct a secondary vertex, which represents the possible decay point of the charm hadron.

The first step of the workflow is the track index skim creator. This task applies a basic event selection, for example by rejecting unsuitable collisions, and a basic track selection, keeping only tracks with sufficient reconstruction quality and within the required kinematic acceptance. Its purpose is to reduce the input data size

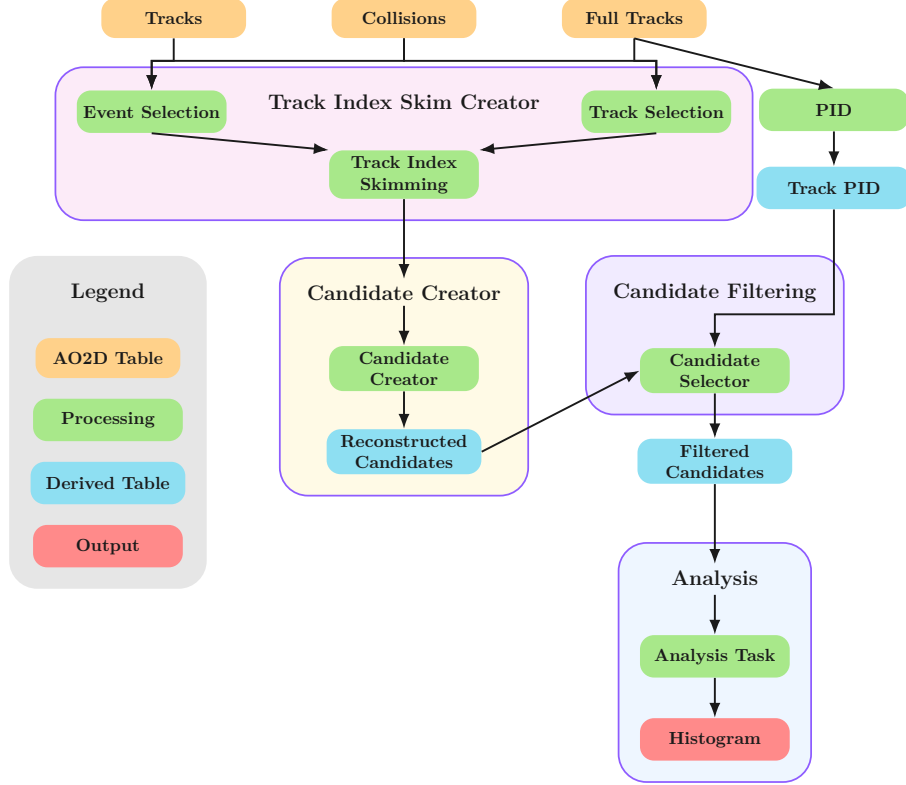


Figure 3.2: Scheme of the PWGHF analysis framework.

and keep only collisions and tracks that can be useful for further charm-hadron reconstruction. The task also prepares indices of selected tracks, which are later used to build possible charm-hadron candidates.

In the next step, the candidate creator combines the selected tracks according to the chosen decay-channel hypothesis. For each accepted track combination, it reconstructs a possible secondary vertex and computes the corresponding topological and kinematic quantities. These include, for example, the decay length, pointing angle, impact parameters, and invariant mass. The output of this step is a table of reconstructed charm-hadron candidates.

The candidate selector then applies particle-identification information and additional selection criteria to these candidates. These criteria are based on quantities that help distinguish real charm-hadron decays from random combinations of tracks, such as the quality of the reconstructed secondary vertex, the displacement from the primary vertex, and the compatibility of the daughter tracks with the expected particle species. Instead of directly removing all candidates, the selector usually stores the result of these checks in selection flags. These flags indicate whether a candidate satisfies a given set of requirements and can then be used in the final analysis step. Finally, the analysis task processes the selected candidates and produces the histograms and observables needed for the physics analysis, such as invariant-mass distributions in different p_T intervals[22].

3.5 Hyperloop

Hyperloop is a web-based train system used within the ALICE O2 analysis framework. It provides a practical interface for configuring and running analysis workflows on available datasets. In Hyperloop, an analysis workflow is organised into so-called wagons, where each wagon represents a particular O2 workflow together with its configuration. A configuration file contains the settings used by the workflow, such as selection cuts, task parameters, histogram binning, input and output options, and paths to external objects needed by the analysis. Through the interface, the user can choose the dataset on which the workflow should be tested or executed, modify the configurable parameters of the analysis tasks, and inspect the output and basic performance information of the run.

This is especially useful for analyses based on large Run 3 datasets, where running the full workflow locally would not be practical. Hyperloop allows the workflow to be submitted to the ALICE analysis train system and processed using the available computing resources. The same analysis can therefore be tested first on smaller samples and later executed on larger datasets in a controlled and reproducible way. Different configurations of the same workflow can also be compared using subwagons, which makes Hyperloop useful for testing several analysis settings without changing the main workflow structure.

In heavy-flavour analyses, Hyperloop can be used to run reconstruction, filtering, and analysis workflows on selected data samples and to produce derived tables for later analysis steps. The configurable task parameters can include event-, track-, and candidate-selection criteria, histogram binning, and, when needed, the application of machine-learning models during candidate selection. These models or other calibration and configuration objects can be accessed through the Conditions and Calibration Database (CCDB). The CCDB is a central database used in O2 to store and retrieve time-dependent objects needed during reconstruction and analysis, such as calibration constants, detector conditions, and machine-learning models. By storing these objects centrally, the same workflow can use the correct version of the required inputs in a reproducible way.

The resulting derived tables and histograms can then be used for further studies, such as checking reconstructed candidates, optimising the selection, and performing invariant-mass analyses [23].

Chapter 4

Machine Learning

In this chapter, I would like to provide an overview of what machine learning is, how it works, and how it can make our lives easier in the field of HF candidate reconstruction.

Machine learning is a set of methods through which patterns and information can be extracted from data. It is commonly divided into two main approaches: supervised learning and unsupervised learning [24]. In a typical learning process, the algorithm is provided with input data described by a set of variables, called features. In the case of supervised learning, the data are also assigned to known output categories, called labels. Based on the input features, the model learns how to assign each input object to one of the possible categories, often with a certain probability. During training, it is not the input variables themselves that are optimised, but rather the internal parameters of the model that determine how the decision is made.

Supervised learning is a type of learning in which the algorithm receives both the input data and the corresponding true labels. When the model makes a prediction for a given data point, this prediction can be directly compared to the known label. The disagreement between the prediction and the true label is quantified during training and used to adjust the model parameters. In practice, the available labelled dataset is typically divided into two subsets: one used for training the model and the other for testing its performance. This helps to verify that the model does not simply memorise the training data, but instead learns a more general method for distinguishing between categories [24]. Supervised learning is therefore particularly powerful in cases where reliable labels are available, and it is commonly used in particle physics applications.

On the other hand, if such well-labelled data are not available, unsupervised learning methods can be used. In this case, the algorithm is provided only with input data and attempts to identify underlying structures or patterns without access to true labels. Rather than learning from a known target, the model searches for similarities, clusters, or deviations from the dominant structure of the data. This approach is particularly useful for tasks such as anomaly detection. Since unsupervised learning is not directly used in this work, it will not be discussed in further detail.

The use of supervised learning can be illustrated with the example of training on

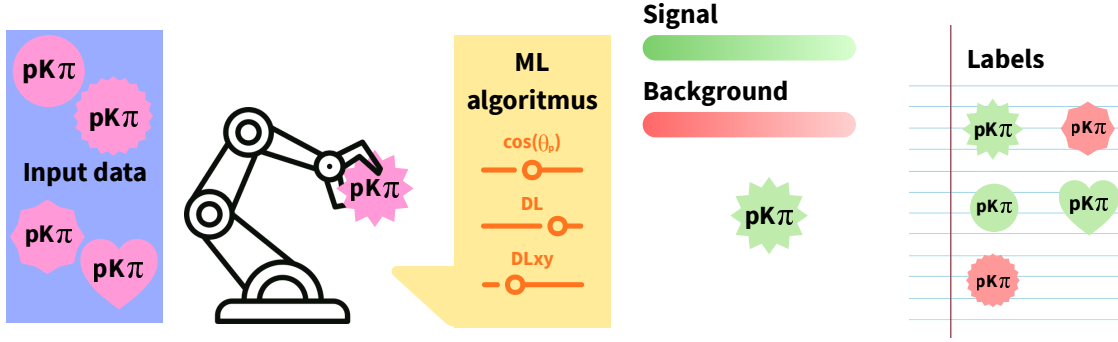


Figure 4.1: Visualisation of supervised learning in the case of an ML task used to distinguish signal and background in Λ_c^+ candidate reconstruction.

Λ_c^+ candidates [10]. A schematic illustration is shown in Fig. 4.1. Each candidate is reconstructed from a combination of three daughter particles: proton, kaon, and pion (p , K , π). The algorithm processes each candidate by evaluating a set of input features, such as the decay length, cosine of the pointing angle, and other PID and topological variables. Based on these features, the model assigns a score that quantifies how likely it is that the candidate corresponds to signal or background in a binary classification task. Examples of algorithms used for this purpose include Boosted Decision Trees and deep neural networks [10].

The predicted score, typically taking values between 0 and 1, is compared to the true label of the candidate during training. By repeating this procedure over the full training dataset, the internal parameters of the model are gradually adjusted so that the separation between signal and background candidates improves.

In the context of this work, the machine-learning task is formulated as a classification problem, where the goal is to assign a label to each input candidate. Depending on the number of possible outcomes, one distinguishes between binary classification, where candidates are assigned to two classes, such as signal and background, and multi-class classification, where more categories are considered. The model may either assign a discrete class label or provide a probability, or score, reflecting the likelihood of belonging to each class.

Within supervised learning, the quality of the predictions is quantified by a loss function. A commonly used loss function in classification tasks is the cross-entropy loss, defined as

$$\mathcal{L} = -y \log(\hat{y}) - (1 - y) \log(1 - \hat{y}), \quad (4.1)$$

where y denotes the true label and $\hat{y}$ the predicted probability. This loss function strongly penalises confidently incorrect predictions. In heavy-flavour candidate reconstruction, where a large combinatorial background is typically present, this is important because the model should assign high confidence only to candidates that are likely to correspond to true signal, while remaining conservative in ambiguous cases.

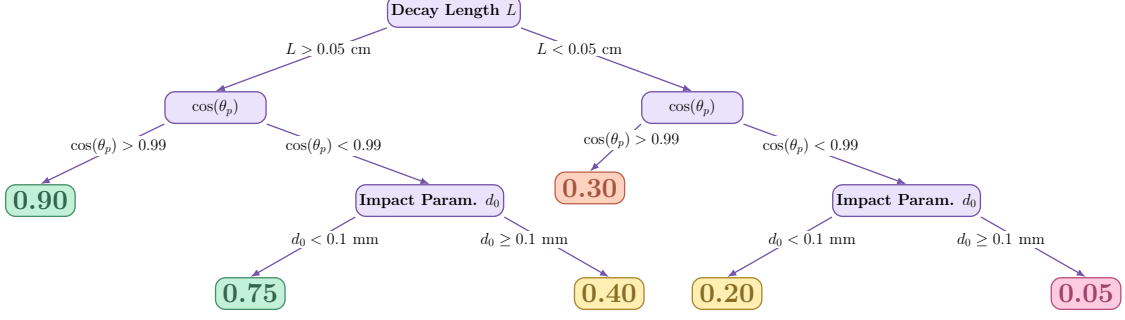


Figure 4.2: Visualization of one decision tree used to train a model to identify signal particle candidates.

4.1 Boosted Decision Trees (BDTs)

For different machine-learning tasks, understood here as different types of ML problems, different algorithms can be employed. In particle physics, and in heavy-flavour candidate reconstruction in particular, one of the most commonly used approaches is Boosted Decision Trees (BDTs) [5, 24]. A BDTs can be understood as an ensemble model, where multiple decision trees are combined into a single predictive function of the form

$$f(\mathbf{x}) = \sum_{m=1}^M \alpha_m T_m(\mathbf{x}), \quad (4.2)$$

where $\mathbf{x}$ represents the vector of input features (such as decay length, cosine of the pointing angle, etc.), T_m denotes the m -th decision tree, M is the total number of trees, and α_m are weights associated with each tree, typically controlled by a learning rate parameter.

Each individual decision tree partitions the feature space into a set of disjoint regions and can be written as

$$T(\mathbf{x}) = \sum_{j=1}^J c_j \mathbb{I}(\mathbf{x} \in R_j), \quad (4.3)$$

where the feature space is divided into J regions R_j , and the indicator function $\mathbb{I}(\mathbf{x} \in R_j)$ assigns an output value c_j to each region. These values correspond to scores that quantify how likely candidates falling into a given region are to belong to a particular class (e.g. signal or background). A visual representation of a single decision tree is shown in Fig. 4.2, where one can observe a sequence of decision nodes based on different input features, leading to terminal nodes (leaves) that provide the final output values [5].

In the boosting procedure, trees are not trained independently. Instead, they are added sequentially, with each new tree trained to correct the errors of the previous ensemble. More precisely, at each iteration the algorithm evaluates the discrepancy between the predicted values and the true labels using a loss function, and the subsequent tree is trained to approximate the negative gradient of this loss. In this way, the model gradually improves its predictive performance by focusing on candidates that were previously misclassified [24].

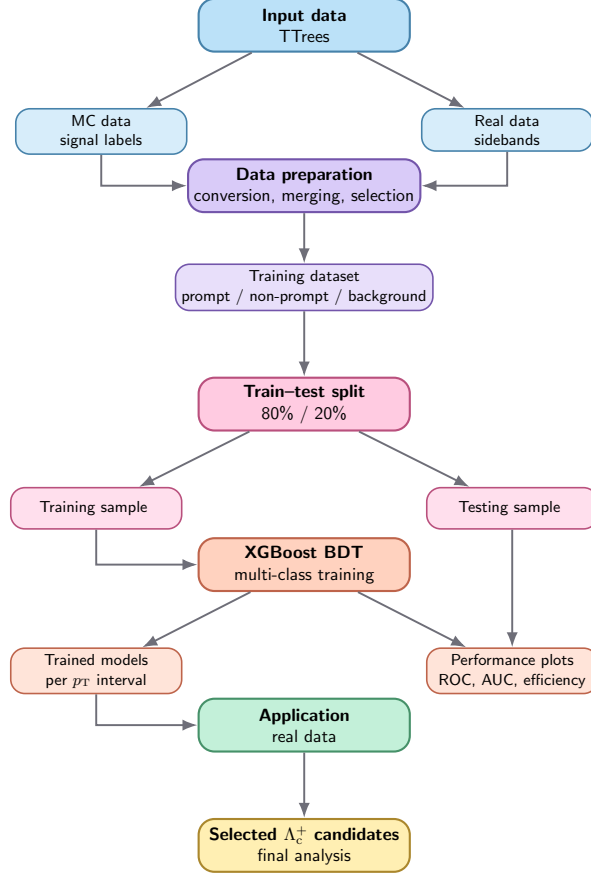


Figure 4.3: Simplified workflow of the MLHEP framework used for training and applying the machine-learning model.

Through this iterative procedure, the ensemble gradually improves its ability to distinguish between signal and background. The result of the training procedure is a trained model composed of a sequence of decision trees, where each tree is defined by its splitting criteria and corresponding output values.

4.2 MLHEP Framework

Within the particle physics community, there has been a growing interest in applying machine-learning techniques to particle reconstruction [24]. As a result, several groups within the ALICE collaboration have developed dedicated frameworks tailored to their specific analyses. One example is the MLHEP (Machine Learning for High Energy Physics) framework, developed within the Heavy-Flavour (HF) working group [25]. This framework is designed for the reconstruction of heavy-flavour hadrons and jets in pp collisions using machine-learning techniques. The general workflow of the MLHEP framework, from input data preparation to the application of the trained model, is schematically illustrated in Fig. 4.3.

The MLHEP framework formulates the problem as a multi-class classification task,

where each reconstructed candidate is assigned to one of three categories: prompt signal, non-prompt signal, or combinatorial background. Prompt candidates originate directly from the collision primary vertex, while non-prompt candidates originate from secondary decays, typically from beauty-hadron decays. The third category corresponds to combinatorial background, formed by random combinations of tracks that accidentally satisfy the reconstruction criteria. Even when the main interest lies in prompt particles, including all three classes in the training improves the model’s ability to distinguish real charm-hadron decays from background and from other signal components.

The framework is designed for fast and parallel data processing and machine-learning optimisation using modern computing resources. It is implemented in Python and relies on data handling through Pandas. For the machine-learning part, it employs algorithms from XGBoost, which is an efficient implementation of gradient-boosted decision trees [25]. In this approach, an ensemble of shallow decision trees is constructed sequentially, where each new tree is trained to improve the performance of the previous ensemble. The framework is organised into tasks, for example for data conversion, merging, model training, and subsequent analysis.

In the case of Λ_c^+ reconstruction, the framework operates on input data stored in the form of TTree tables, both for real data and Monte Carlo (MC) simulations. These tables contain topological, kinematic, and particle-identification information for each candidate and its daughter particles. In MC simulations, candidates are associated with a true label, indicating whether they correspond to prompt signal, non-prompt signal, or background. Ideally, the MC samples should be simulated under conditions matching those of the real data, in order to ensure consistency in detector response and reconstruction performance.

From these inputs, the framework constructs the datasets used for training and evaluation. The data are first converted into a Python-readable format, typically `.pkl`, for both MC and real data. Signal candidates are selected from MC simulations using their true labels. In contrast, the combinatorial background is usually taken from real data rather than from MC simulations. This is done to reduce possible biases caused by an imperfect simulation of the combinatorial background. If the background were taken only from MC, the model could learn features specific to the simulation rather than to real data, which would reduce its predictive power when applied to experimental data.

To select a background sample from real data, candidates are taken from the sidebands of the invariant-mass distribution, outside the signal region. In the Λ_c^+ analysis, this corresponds to selecting candidates outside the mass interval

$$2.22 < m_{\text{inv}}(\Lambda_c^+) < 2.35 \text{ GeV}/c^2.$$

This ensures that the selected candidates predominantly represent combinatorial background. Since the amount of combinatorial background is usually much larger than the number of signal candidates, only a fraction of the available sideband candidates is used, typically of the order of a few to ten percent. The prepared samples are constructed separately for each transverse-momentum interval, because the candidate properties and the optimal selection criteria can vary significantly

with p_T .

Once the labelled samples are prepared, they are divided into training and testing subsets, typically in an 80:20 ratio. The training subset is used to optimise the model parameters, while the testing subset is reserved for performance evaluation. During this split, the relative fraction of prompt, non-prompt, and background candidates is preserved in both subsets. This is important because the testing dataset should represent the same classification problem as the training dataset. A significant difference between the model response on the training and testing samples may indicate overtraining or biases in the trained model, while small differences are expected due to statistical fluctuations and the finite size of the samples.

Within the training configuration, one also defines the set of input features, meaning the variables used by the model to distinguish between the different candidate classes. These variables are usually chosen from topological, kinematic, and particle-identification quantities, based on their discriminating power and their availability in the dataset.

The training is performed using the XGBoost implementation of boosted decision trees [25], resulting in a trained model and a set of performance plots. These include, for example, the distributions of the BDT output scores for the different classes, ROC curves with the corresponding AUC values, feature-importance plots, and efficiency plots used to evaluate the model response and support the choice of the working point. The framework trains separate models for different intervals of transverse momentum p_T , so that a dedicated trained model is obtained for each predefined p_T interval. Each trained model encodes the learned decision structure, including the splitting criteria and output values for all input features. These models can then be integrated into the analysis workflow and applied to real data for the reconstruction of Λ_c^+ candidates.

In addition to the trained models, the framework provides a set of evaluation plots, such as efficiency curves and classifier-response distributions. These are obtained by applying the trained model to the testing dataset, which was not used during training, providing an unbiased estimate of the model performance. The evaluation allows one to assess how effectively the model distinguishes between prompt, non-prompt, and background candidates, as well as to identify the most relevant input variables. Based on these results, the feature selection and model configuration can be further refined, or the trained model can be directly used in the final analysis.

4.3 Application of the Trained Model

After the model is trained, there are several ways to apply it in the data analysis. One of them, in the case where the model was trained using the MLHEP framework, is to perform the full reconstruction directly within the same framework [25]. This is done by running an additional task, usually referred to as **analysis**. In this case, the framework does not only apply the trained model, but also performs the candidate selection and the final reconstruction procedure itself. As an output, it

already provides the invariant-mass distributions together with the fitted signal and background contributions and the corresponding statistical quantities.

By default, this analysis is carried out on the same real or Monte Carlo datasets that were used in the previous steps of the workflow. However, the trained model can also be applied to a different dataset simply by changing the path to the input data. In that case, it is not necessary to repeat the full machine-learning training procedure, and the already trained model can be reused directly for the candidate reconstruction [25].

Another possibility is to apply the trained model within the standard ALICE Run 3 workflow, either on Hyperloop or in a local O² analysis using the `candidate-selector` task [23, 21]. In this setup, the model is incorporated into the usual reconstruction chain and used during the candidate-selection stage. For Hyperloop processing, the trained models corresponding to the individual p_T intervals must be uploaded to the CCDB [23], whereas for local analysis they are stored locally and loaded from the corresponding paths. In both cases, the model paths are specified in the configuration file and the relevant machine-learning options are enabled.

For the analysis task to be able to read and evaluate the trained model correctly, the model has to be stored in a format compatible with the O² framework. In this analysis, the trained model is used in the ONNX format. ONNX, the Open Neural Network Exchange format, is a standard format for representing machine-learning models, which allows models trained in different frameworks to be exported and used in a common way. If the model is trained within the MLHEP framework, it can be exported to a format suitable for use in O² [25]. In other cases, an additional conversion step may be needed to ensure that the model can be correctly loaded and evaluated by the analysis task.

Chapter 5

Λ_c^+ Reconstruction and Selection

In this analysis, the Λ_c^+ baryon is reconstructed via its charged hadronic decay channel $\Lambda_c^+ \rightarrow pK^-\pi^+$. The Λ_c^+ has a measured mass of $2286.46 \pm 0.14 \text{ MeV}/c^2$ [1] and the branching ratio of this channel is $\mathcal{B} = (6.28 \pm 0.32)\%$ [26]. This is the dominant hadronic decay mode of the Λ_c^+ ; the branching ratio gives the fraction of all Λ_c^+ decays that proceed via this final state. The primary objective is to select prompt Λ_c^+ baryons, which originate directly from the primary vertex (PV), i.e. the proton–proton interaction point [27]. Heavy-flavour candidates are built from track triplets with charge combinations corresponding to the daughter particles (p , K^- , π^+). The resulting sample comprises prompt signal, non-prompt signal, and combinatorial background.

5.1 Datasets

The data used in this analysis include both reconstructed real data and simulated Monte Carlo (MC) samples. The real-data sample corresponds to the `HF_LHC23_pass4_Thin_small_2P3PDstar` production chain. The prefix `HF` indicates that heavy-flavour-specific tables were produced. `LHC23` identifies the 2023 data-taking period, during which proton–proton collisions were recorded at a centre-of-mass energy of $\sqrt{s} = 13.6 \text{ TeV}$ as part of the LHC Run 3 campaign. The label `pass4` denotes the reconstruction pass, where updated calibrations and algorithms generally yield higher-quality data than earlier passes. The `Thin` suffix means that tables unnecessary for this analysis were removed, significantly reducing the dataset size. The suffix `small` denotes a reduced subset of the full dataset. In this case, the original dataset size was approximately 1.5 PB, while the reduced `small` sample had a size of about 30 TB, corresponding to roughly 2% of the full dataset. The dataset used in this thesis was further reduced through a skimming procedure, producing derived data optimised for heavy-flavour candidate reconstruction. This smaller sample is typically used to validate the analysis procedure, check that the workflow runs correctly, and verify that the processing is efficient before requesting the analysis to be run on the full statistics. Finally, `2P3PDstar` indicates that only 2-prong, 3-prong and D^* candidates were stored; this analysis uses only the 3-prong candidates.

The Monte Carlo sample carries the production tag `HF_LHC24h1c_A11`. Here, `HF` signals that the sample is enriched in heavy-flavour events containing $c\bar{c}$ and $b\bar{b}$ pairs. `LHC24h1c` identifies the simulation campaign: the events were generated in 2024 using detector conditions and geometry matched to the 2023 data and reconstructed with the same `pass4` settings. The MC is therefore anchored to the same run conditions as the real data described above.

Using real and simulated data anchored to identical run conditions guarantees that the collision parameters, detector settings and reconstruction algorithms are consistent between the two samples. This consistency helps prevent biases between signal and background candidates during machine-learning training.

Both datasets were further processed to reconstruct Λ_c^+ candidates. The relevant kinematic and topological information was stored in analysis tables, and for the MC sample candidates were labelled as prompt, non-prompt, or combinatorial background. The derived data were converted from `ROOT` files into pickle format for subsequent reading with Python scripts. During this step, candidates were divided into ten p_T intervals. The bin boundaries were chosen to provide sufficient statistics for a stable signal extraction while still allowing the selection strategy to be optimised separately in each momentum region [27].

5.2 Selection Variables

The selection of reconstructed Λ_c^+ candidates is performed in two steps. The first step is a rectangular selection, also referred to as a cut-based selection. In this approach, each variable is treated independently and a lower or upper threshold is applied to it. Only candidates satisfying all required conditions are kept. Rectangular selections are simple, transparent, and useful for rejecting candidates that clearly do not satisfy the expected topological or kinematic properties of the reconstructed decay.

The second step is a machine-learning based selection. In this approach, several variables are used simultaneously and the classifier learns how to separate signal-like candidates from background-like candidates using their combined multidimensional information. In contrast to rectangular selections, the machine-learning method can also exploit correlations between variables, which is especially important for complex decay topologies.

For the reconstruction of short-lived charm baryons such as the Λ_c^+ , the rectangular selection alone is not sufficient to obtain a clearly separated signal. The Λ_c^+ decay topology is complex and the discriminating information is distributed over several correlated variables. For this reason, the rectangular selection is used only as a loose pre-selection to remove the most obvious background-like candidates, and not as the main selection approach of the analysis. The final candidate selection is instead performed using a machine-learning based method. The invariant-mass distribution obtained using only the rectangular selection is shown in the Appendix in Fig. 6.7, illustrating the need for the additional machine-learning selection.

5.2.1 Rectangular Pre-selection

Before any sophisticated selection is applied to the reconstructed candidates described above, a set of loose rectangular cuts is imposed, shown in Tab. 5.2.1. A rectangular cut places a lower and upper bound on each individual variable, thereby defining a rectangular acceptance region in the multidimensional parameter space. The meaning of each variable is described in the following section, since the same variables are also used for the machine-learning selection. This pre-selection efficiently rejects the most obvious and strongest combinatorial background and reduces the size of the data sample [28].

p_T interval [GeV/ c]	p_T^p	p_T^K	p_T^π	χ_{PCA}^2	Decay length	CPA	Decay length _{xy}
0 – 1	0.3	0.3	0.3	20	0.006	0.80	0
1 – 4	0.4	0.3	0.3	20	0.010	0.85	0
4 – 100	0.4	0.3	0.3	20	0.020	0.90	0

Table 5.1: Pre-selection cuts used in different p_T intervals.

5.2.2 Training Variables

Within the datasets described in 5.1, each Λ_c^+ candidate is described by a set of variables. These variables can generally be divided into three categories: topological, kinematic, and particle-identification (PID) variables. In the training procedure, only the topological and PID variables were used.

To understand these variables, it is useful to first describe the decay topology of the Λ_c^+ baryon. The two beams collide in the region of the interaction point (IP), where the actual proton–proton interaction is reconstructed as the primary vertex (PV). A prompt Λ_c^+ baryon is produced directly at the PV. Due to its finite lifetime, it travels a short distance before decaying into its daughter particles. The position of this decay is called the secondary vertex (SV), which is the common origin of the daughter tracks. In this work, the decay channel

$$\Lambda_c^+ \rightarrow pK^-\pi^+$$

is studied. The proton, kaon, and pion then propagate through the detector, where they are measured by the different detector subsystems.

The distributions shown in the following sections are obtained after a basic preselection of candidates. Therefore, the background sample does not consist of completely random track combinations, but only of combinations that already satisfy loose candidate-selection requirements and can therefore resemble real Λ_c^+ decays.

Besides prompt Λ_c^+ production, several other processes contribute to the reconstructed candidate sample. A large number of unrelated tracks can accidentally form combinations with a topology similar to the signal. These candidates form the combinatorial background. In addition, beauty hadrons are produced in proton–proton collisions and can subsequently decay into Λ_c^+ baryons. These Λ_c^+ baryons

are referred to as non-prompt. Although they are less abundant, they can resemble prompt candidates, since the decay lengths involved are still relatively short.

Topological variables describe the geometry of the decay. They encode how the daughter tracks are arranged with respect to the PV, the SV, and the reconstructed momentum direction of the Λ_c^+ candidate. Since the Λ_c^+ baryon has a short lifetime, the small displacement of the decay vertex and the spatial configuration of the daughter tracks provide important information for separating signal candidates from background.

The topological variables used in this analysis are the impact parameters of the prong particles (d_0), the cosine of the pointing angle (CPA), the PCA fit quality (χ_{PCA}^2), the decay length (DL), and the decay length in the transverse plane (DL_{xy}). Each variable is discussed together with its definition, its distribution for prompt, non-prompt, and background candidates, and its dependence on p_T . For each variable, an example distribution is shown for the interval $3 < p_T < 4$ GeV/ c , while the corresponding distributions for the remaining p_T intervals are presented in the Appendix. The following subsection is based on the Monte Carlo sample introduced at the beginning of this chapter. This sample is used because it provides the generated-particle information needed to distinguish prompt Λ_c^+ candidates, non-prompt Λ_c^+ candidates, and combinatorial background.

Decay Length (DL/DL_{xy})

The Λ_c^+ baryon has a mean proper lifetime of approximately

$$c\tau \approx 60 \text{ } \mu\text{m}. \quad (5.1)$$

Due to relativistic time dilation, particles produced with finite momentum in the laboratory frame travel a measurable distance before decaying. As a result, one of the key signatures of charm-hadron reconstruction is the presence of a displaced secondary vertex (SV) separated from the primary vertex (PV) [10]. The decay length (DL) is defined as the three-dimensional distance between the PV and the reconstructed SV. Introducing the flight vector

$$\vec{L} = \vec{r}_{\text{SV}} - \vec{r}_{\text{PV}} = (x_{\text{SV}} - x_{\text{PV}}, y_{\text{SV}} - y_{\text{PV}}, z_{\text{SV}} - z_{\text{PV}}), \quad (5.2)$$

the decay length is given by

$$L = |\vec{L}|. \quad (5.3)$$

Here, $\vec{r}_{\text{PV}}$ and $\vec{r}_{\text{SV}}$ denote the position vectors of the primary and secondary vertices, respectively. A schematic illustration of the decay topology is shown in Fig. 5.1.

In addition to the full three-dimensional decay length, one often uses its projection onto the transverse plane,

$$\vec{L}_{xy} = (x_{\text{SV}} - x_{\text{PV}}, y_{\text{SV}} - y_{\text{PV}}, 0), \quad (5.4)$$

$$L_{xy} = |\vec{L}_{xy}|. \quad (5.5)$$

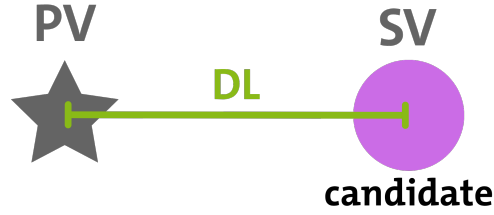


Figure 5.1: Schematic representation of the decay length. The Λ_c^+ baryon is produced at the primary vertex (PV), travels a finite distance, and decays at the secondary vertex (SV). The vector connecting the two vertices defines the flight vector $\vec{L}$, whose magnitude corresponds to the decay length DL .

In many analyses, signed decay lengths are used,

$$L_{\text{signed}} = \frac{\vec{L} \cdot \vec{p}}{|\vec{p}|}, \quad (5.6)$$

where $\vec{p}$ is the reconstructed candidate momentum. This quantity distinguishes candidates whose reconstructed flight direction is consistent with the momentum vector from those pointing in the opposite direction, which are typically associated with misreconstructed or combinatorial background candidates.

Both L and L_{xy} provide complementary information. The full decay length L uses the complete three-dimensional geometry of the event, while L_{xy} depends only on the transverse plane and is therefore insensitive to boosts along the beam (z) direction. Since the transverse plane is generally reconstructed with better precision than the longitudinal direction, L_{xy} often provides improved discrimination power. It is also commonly used in the determination of proper decay time.

The distributions of the decay length (DL) and transverse decay length (DL_{xy}) for prompt, non-prompt, and combinatorial background candidates are shown in Fig. 5.2. For the full three-dimensional decay length DL , prompt Λ_c^+ candidates peak around ~ 0.02 cm. This is consistent with the expected average flight distance of a Λ_c^+ baryon in the interval $3 < p_T < 4$ GeV/ c , where the Lorentz boost extends the decay length to roughly $200 \mu\text{m}$. Non-prompt candidates peak at larger values, since they originate from displaced beauty-hadron decays before the Λ_c^+ itself is produced, resulting in an additional separation from the PV. In contrast, the combinatorial background peaks closer to zero, as many randomly combined tracks originate directly from the primary vertex.

The transverse decay length DL_{xy} exhibits peaks for all candidate classes close to zero, since only the displacement in the transverse plane is considered and the projected flight distance is therefore smaller than the full three-dimensional value. At the same time, the prompt signal (blue and purple bars in the figure) is more clearly separated from the background than in DL . This is because the transverse plane is measured more precisely than the beam direction, so removing the noisier z component sharpens the signal distribution. The background distribution is somewhat

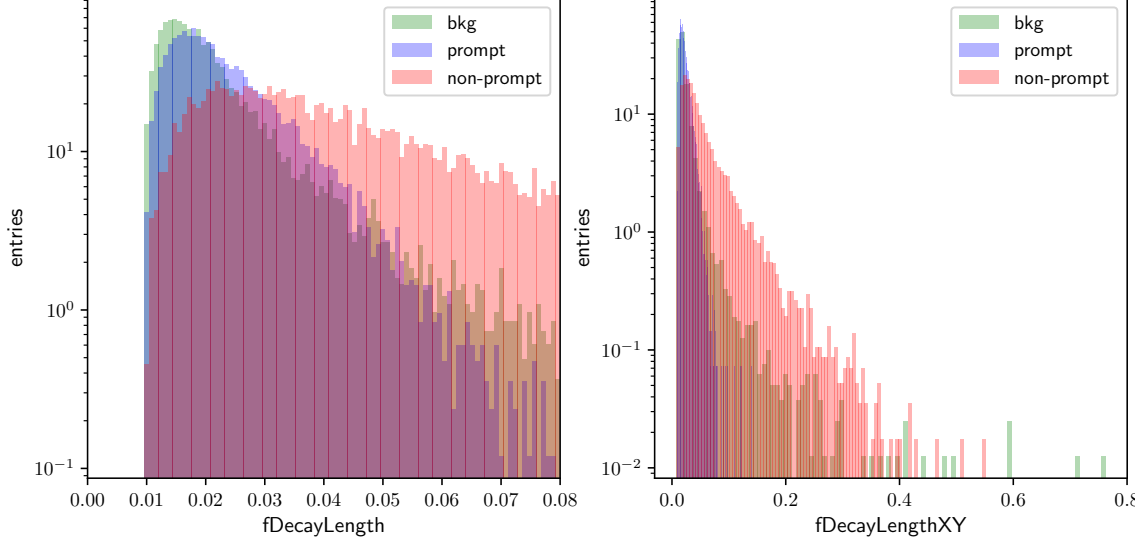


Figure 5.2: Distribution of the decay length (DL) and transverse decay length (DL_{xy}) for prompt, non-prompt, and combinatorial background candidates for the example interval $3 < p_T < 4$ GeV/ c .

broadener due to random track combinations, while the broadest distribution is again observed for non-prompt candidates, which remain systematically displaced also in the transverse plane.

Both the measured decay length L and its transverse projection L_{xy} increase on average with increasing transverse momentum p_T due to relativistic time dilation. Consequently, higher- p_T Λ_c^+ baryons travel farther before decaying, leading to broader distributions shifted toward larger values.

Impact Parameter (d_0)

The impact parameter is represented by three separate variables, one for each charged daughter track of the Λ_c^+ candidate. The impact parameter is defined as the distance of closest approach of a prong track to the primary vertex, i.e. the shortest distance between the PV and the extrapolated track trajectory. A schematic illustration of the impact parameter is shown in Fig. 5.3.

This variable provides information about the geometrical origin of the particle. If a particle originates directly from the primary vertex (PV), its trajectory tends to pass close to the PV, resulting in a small impact parameter. In contrast, particles produced in displaced decays generally have larger impact parameters, since their trajectories do not have to intersect the PV.

However, for true prompt Λ_c^+ decays, the impact parameters of the daughter particles can still remain very small, even though the daughters originate from a secondary vertex. This is because the Λ_c^+ has a short lifetime and its decay products are emitted in a narrow forward cone, following the motion of the mother particle. As a result, the daughter trajectories often pass very close to the PV, leading to a narrow

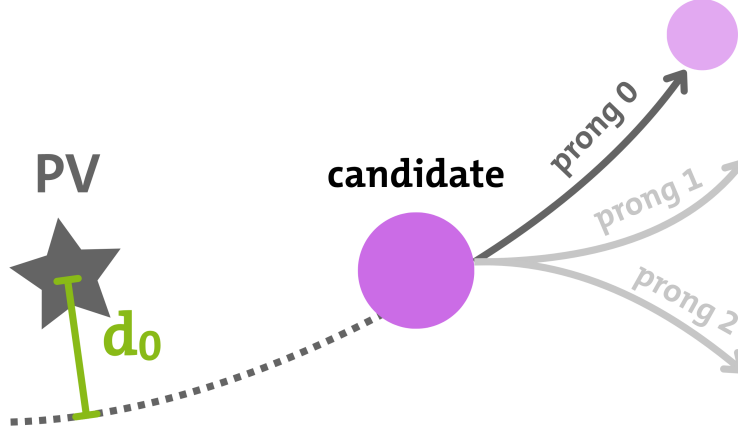


Figure 5.3: Schematic representation of the impact parameter d_0 as an example for prong 0.

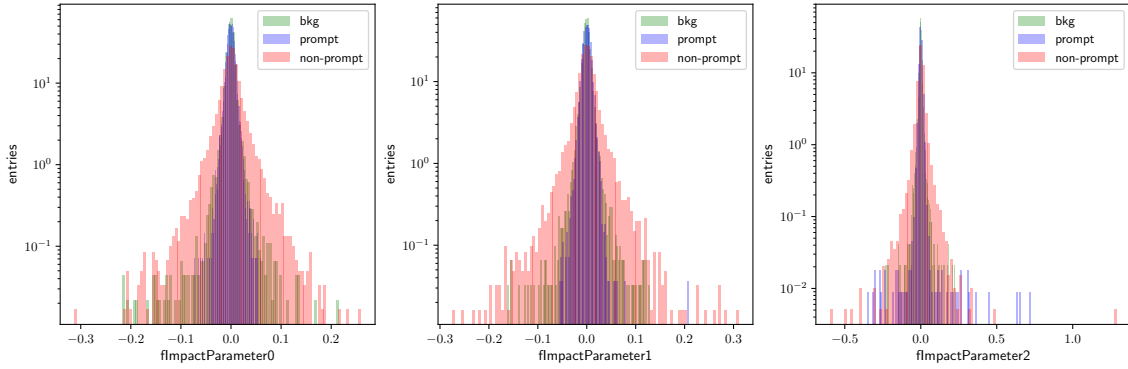


Figure 5.4: Distribution of impact parameters for prompt, non-prompt, and combinatorial background candidates, shown for the example p_T interval $3 - 4 \text{ GeV}/c$.

distribution of impact parameter values.

This behaviour can be observed in Fig. 5.4, shown here for the example interval $3 < p_T < 4 \text{ GeV}/c$. The combinatorial background is mainly concentrated around zero, since a large fraction of the background tracks originates from the PV. At the same time, it exhibits a somewhat broader distribution, as it also contains contributions from other displaced decays and random combinations not related to the signal. The prompt signal is even more narrowly peaked around zero, reflecting the strong forward collimation of the daughter particles. The non-prompt candidates show the broadest distribution of impact parameters, as they originate from decays further displaced from the primary collision.

The forward production of daughter particles is a consequence of the Lorentz boost of the Λ_c^+ . With increasing momentum, the emission becomes more collimated, leading to smaller opening angles. Therefore, at lower p_T a broader distribution of both signal and background of impact parameters is expected.

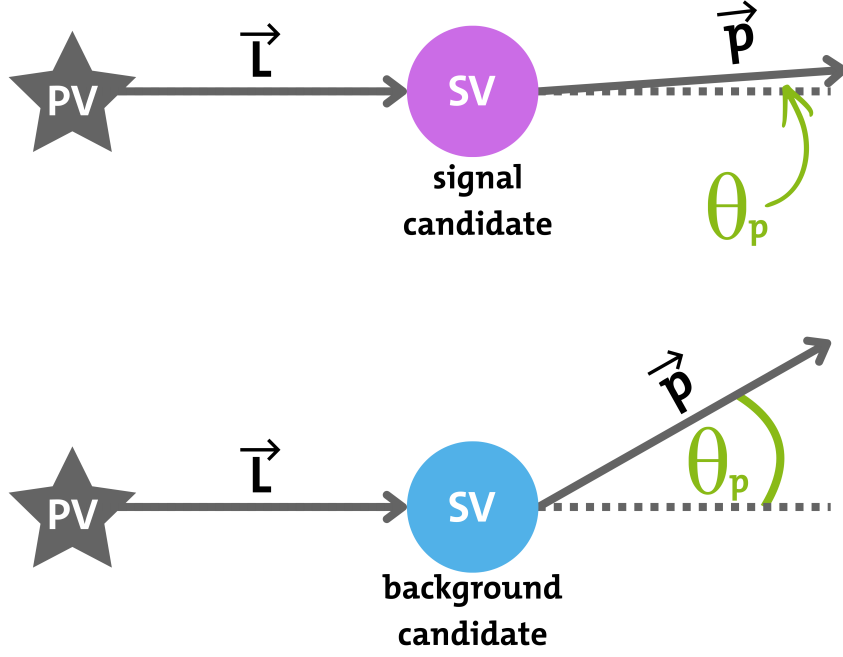


Figure 5.5: Schematic representation of the pointing angle, defined as the angle between the reconstructed candidate momentum vector and the line connecting the primary and secondary vertices.

Cosine of Pointing Angle (CPA)

The cosine of the pointing angle (CPA) is one of the most important topological selection variables in heavy-flavour candidate reconstruction. It describes the angle between the reconstructed momentum vector $\vec{p}$ of the candidate and its flight line $\vec{L}$, defined within the decay length. The CPA is calculated as

$$\text{CPA} = \cos(\theta_p) = \frac{\vec{L} \cdot \vec{p}}{|\vec{L}| |\vec{p}|}. \quad (5.7)$$

A schematic representation of the CPA is shown in Fig. 5.5. Physically, this variable quantifies how well the reconstructed momentum of the candidate is aligned with the measured decay direction from the PV to the SV. For an ideal signal candidate, one expects

$$\cos(\theta_p) = 1, \quad (5.8)$$

since the particle momentum should point in the same direction as its flight path. Deviations from unity arise due to finite detector resolution, such as uncertainties in the reconstructed vertex positions and track momenta, or because the candidate does not originate from a prompt Λ_c^+ decay, as is often the case for combinatorial background or non-prompt candidates.

The distributions of the CPA for prompt, non-prompt, and combinatorial background candidates are shown in Fig. 5.6. Prompt Λ_c^+ candidates are strongly peaked

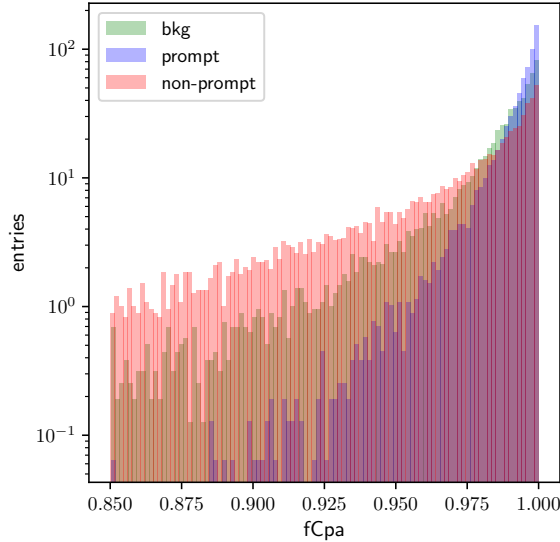


Figure 5.6: Distribution of the cosine of the pointing angle (CPA) for prompt, non-prompt, and combinatorial background candidates for the example interval $3 < p_T < 4$ GeV/ c .

near unity, as expected for particles produced directly at the primary vertex and correctly reconstructed. The combinatorial background distribution is significantly broader and extends over a wider range of values, reflecting the random combination of unrelated tracks. Nevertheless, many background candidates still populate the region close to one, since a large fraction of tracks originate near the PV. Non-prompt candidates, produced in decays of beauty hadrons, typically exhibit an even broader distribution. Since they originate from a displaced beauty-hadron decay vertex rather than directly from the PV, their reconstructed flight direction with respect to the PV is less perfectly aligned.

The dependence of the CPA on transverse momentum p_T is strong. With increasing p_T , the CPA distributions of all candidate categories shift closer to unity. This is mainly caused by the improved precision of both momentum and secondary-vertex reconstruction at higher momentum. Due to relativistic time dilation, higher-momentum charm hadrons travel a larger average distance before decaying, resulting in a more clearly displaced secondary vertex and a more accurately determined flight line with less noise from other events.

As a consequence, true alignment and misalignment can be resolved more precisely at higher p_T , which generally improves the separation power between signal and background candidates.

PCA Fit Quality (χ^2_{PCA})

The PCA fit quality, denoted as χ^2_{PCA} , is a topological variable that quantifies how well the reconstructed daughter-particle trajectories are compatible with the hypothesis that they originate from a common decay vertex. In the case of the decay channel $\Lambda_c^+ \rightarrow pK^-\pi^+$, the three daughter tracks are expected to emerge from a

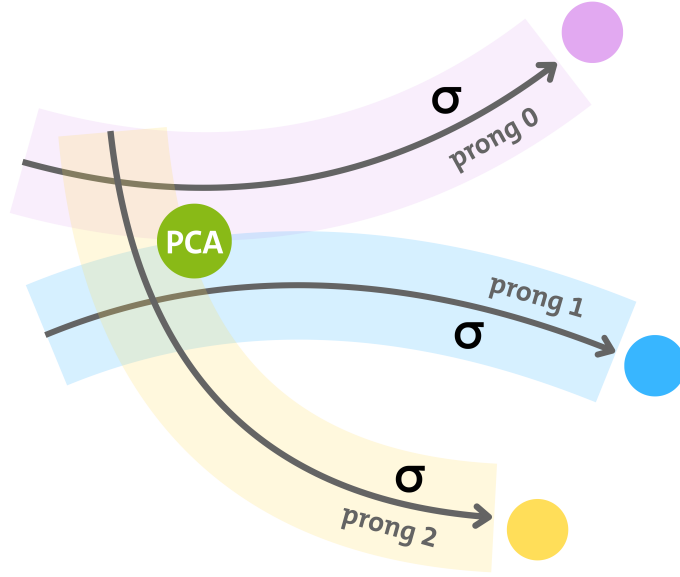


Figure 5.7: Schematic illustration of the point of closest approach (PCA) vertex fit for two daughter-track candidates. The reconstructed trajectories do not intersect exactly due to detector resolution. Instead, the PCA is defined as the point that minimises the distances of closest approach (DCA) to the reconstructed track paths.

single point in space, namely the secondary vertex (SV).

To understand this variable, it is first useful to define the point of closest approach (PCA), illustrated schematically in Fig. 5.7. Due to finite detector resolution, reconstructed charged-particle trajectories do not usually intersect exactly at one point. Instead, one determines the spatial point for which the distances to all daughter trajectories are simultaneously minimised. This point is taken as the best estimate of the decay vertex.

The corresponding distances between the fitted vertex position and the daughter trajectories are commonly referred to as distances of closest approach (DCA). However, the absolute DCA values alone are not sufficient, since their interpretation depends strongly on the detector resolution. A larger DCA may still be consistent with a genuine decay candidate if the measurement uncertainty is also large.

For this reason, the vertex quality is evaluated through a χ^2 quantity, which effectively compares the residual distances to their corresponding uncertainties. In simplified form, this can be written as

$$\chi_{\text{PCA}}^2 \approx \sum_{i=p, K^-, \pi^+} \left(\frac{\text{DCA}_i}{\sigma_{\text{DCA},i}} \right)^2, \quad (5.9)$$

where the sum runs over the daughter tracks and $\sigma_{\text{DCA},i}$ denotes the corresponding uncertainty estimate. In practice, the vertex fit is more sophisticated: the full track-parameter covariance matrices are used, including correlations between parameters, and the fit simultaneously adjusts the vertex position to minimise the total χ^2 . Once the minimum is found, the corresponding point is identified as the reconstructed sec-

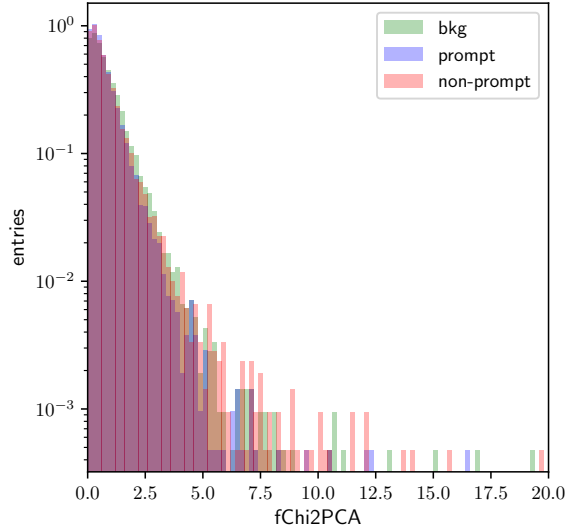


Figure 5.8: Distribution of the PCA fit quality variable χ_{PCA}^2 for prompt, non-prompt, and combinatorial background candidates for the example interval $3 < p_T < 4 \text{ GeV}/c$.

ondary vertex. For an ideal signal candidate, whose daughter particles truly originate from a common decay point, one expects low values of χ_{PCA}^2 .

The distributions of χ_{PCA}^2 for prompt, non-prompt, and combinatorial background candidates are shown in Fig. 5.8. Signal candidates peak at low values, reflecting a good common-vertex hypothesis. The combinatorial background is expected to populate somewhat larger values, since randomly associated tracks do not naturally originate from the same decay point and therefore lead to a poorer fit. Nevertheless, because the fitting procedure always searches for the best possible common point, many background candidates can still obtain relatively small values.

Non-prompt candidates, originating from beauty-hadron decays, also correspond to genuine displaced decay vertices and therefore often exhibit similarly small values. As a result, the three distributions partially overlap, indicating that χ_{PCA}^2 alone has limited discrimination power. Its main usefulness arises when combined with other topological or PID variables in multivariate analyses.

The dependence of χ_{PCA}^2 on transverse momentum p_T is non-trivial. With increasing p_T , the daughter particles become more collimated due to the Lorentz boost of the parent hadron. Their trajectories therefore become more parallel, which can reduce the precision with which the vertex position is determined along the flight direction. This tends to worsen the fit quality and broaden the distribution.

At the same time, higher-momentum tracks are measured more precisely and undergo less multiple scattering, which improves the track-parameter resolution. For non-prompt candidates, the larger displacement of the decay vertex from the primary vertex often helps the reconstruction further. The observed p_T dependence of χ_{PCA}^2 is therefore determined by the interplay between improved tracking precision and increasing geometric collimation of the daughter tracks.

Particle Identification Variables

Particle identification (PID) variables are observables used to determine whether a daughter track corresponds, in the case of the Λ_c^+ reconstruction, to a proton, kaon, or pion. This is an essential step, since many random track triplets can be formed in high-multiplicity collisions, and without proper PID a wrong particle species could be assigned to one of the daughter tracks. Different particles interact with the detector material differently according to their mass, charge, momentum, and velocity. In ALICE, the two main detectors used for PID are the Time Projection Chamber (TPC) and the Time-Of-Flight (TOF) detector.

The TPC serves as the primary tracking detector and also provides PID information. It measures the specific energy loss dE/dx of a charged particle as it ionizes the detector gas while traversing the chamber. The measured energy loss is then compared with the expected theoretical value for each particle hypothesis at the corresponding momentum. The output variable is the $n\sigma$ separation, defined as

$$n\sigma_{\text{TPC}}^i = \frac{(dE/dx)_{\text{meas}} - (dE/dx)_{\text{theory}}^i}{\sigma_{dE/dx}}, \quad (5.10)$$

where $\sigma_{dE/dx}$ is the experimental resolution of the detector. This variable expresses how many standard deviations the measured signal differs from the expected response for a given particle hypothesis. Therefore, a track consistent with the proton hypothesis would have $|n\sigma_{\text{TPC}}^p| \approx 0$, while the pion or kaon hypotheses would generally give values farther from zero.

For the TPC, excellent separation is achieved at low momentum, where particles are located in the non-relativistic region of the Bethe–Bloch curve shown in Fig. 5.9. Slower particles spend more time interacting with the detector gas and therefore lose more energy through ionization. As a result, the species-dependent energy-loss bands are well separated below approximately 0.5 GeV/ c . Around 1 GeV/ c , particles approach the minimum-ionizing region, where the separation becomes weakest. At higher momentum, a relativistic rise appears, but the separation power remains reduced compared with the low-momentum region.

PID can also be performed using the TOF detector, which measures the time needed for a particle to travel from the primary vertex to the TOF surface. Combined with the reconstructed track length, this allows the particle velocity to be determined as $\beta = v/c$. The corresponding PID separation variable is defined as

$$n\sigma_{\text{TOF}}^i = \frac{\beta_{\text{meas}} - \beta_{\text{theory}}^i}{\sigma_{\beta}}, \quad (5.11)$$

where σ_{β} is the TOF resolution.

For the TOF detector, the best performance is typically achieved in the intermediate momentum range, approximately 0.5–4 GeV/ c , where particles reach the detector with good track quality while still having measurable velocity differences. In this

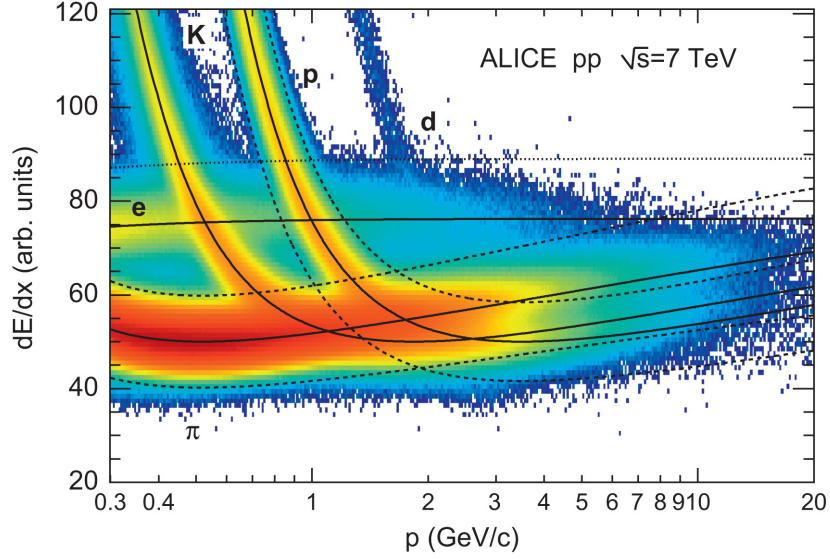


Figure 5.9: Specific energy loss dE/dx measured in the ALICE TPC as a function of particle momentum. The characteristic Bethe–Bloch bands for pions, kaons, protons, electrons, and deuterons are visible. Adapted from Ref. [29].

region, typical values are $\beta_p \approx 0.85$, $\beta_K \approx 0.95$, and $\beta_\pi \approx 0.99$. At high momentum, all particles become highly relativistic with $\beta \rightarrow 1$, making their time-of-flight differences very small and reducing the separation power.

In this analysis, combined information from the TPC and TOF detectors is used, defined as

$$n\sigma_{\text{comb}}^{p,K,\pi} = \begin{cases} |n\sigma_{\text{TPC}}^{p,K,\pi}|, & \text{track available only in TPC,} \\ |n\sigma_{\text{TOF}}^{p,K,\pi}|, & \text{track available only in TOF,} \\ \frac{1}{\sqrt{2}} \sqrt{\left(n\sigma_{\text{TPC}}^{p,K,\pi}\right)^2 + \left(n\sigma_{\text{TOF}}^{p,K,\pi}\right)^2}, & \text{track available in both detectors.} \end{cases} \quad (5.12)$$

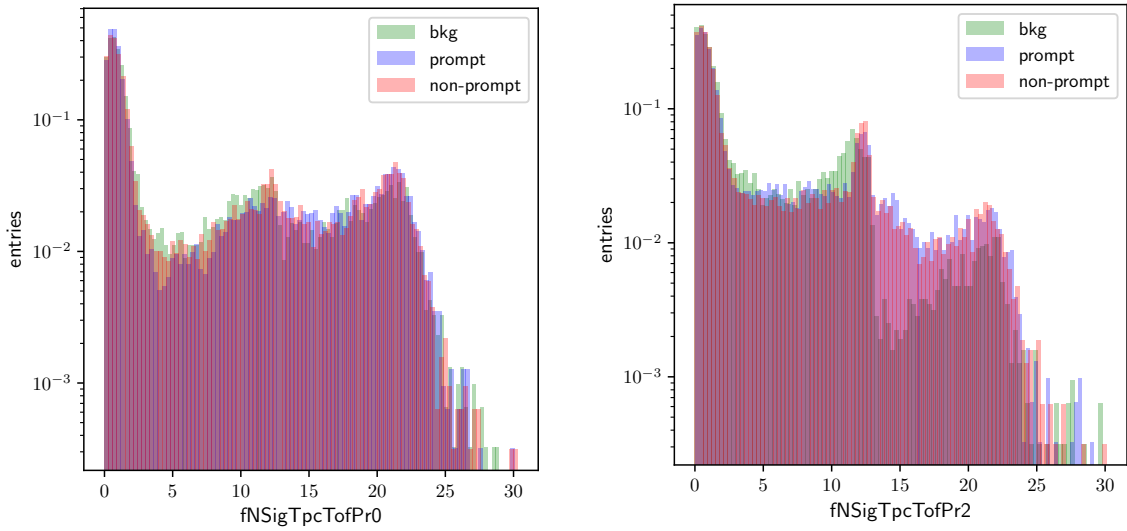
This combined variable reduces the total number of features used in the machine-learning training. Some tracks, especially depending on the momentum range, do not have valid TOF information because they do not reach the detector or cannot be reliably matched. In such cases, only the TPC information is used.

When both detectors are combined, the low-momentum region is dominated by the TPC response, the intermediate region provides the strongest overall PID thanks to the complementarity of both detectors, and the high-momentum region again relies mainly on the TPC, although with reduced total performance.

In total, five PID variables are used. Among the three daughter tracks, two are positively charged and may correspond either to a proton or a pion, while the negatively charged daughter can correspond to the kaon in the decay channel $\Lambda_c^+ \rightarrow pK^-\pi^+$. The corresponding PID distributions are shown in Fig. 5.10 and 5.11.

For all particle hypotheses, a peak close to zero is visible for prompt and non-prompt candidates, since both classes contain genuine Λ_c^+ decays with the correct daughter species. The combinatorial background also shows an enhancement near zero, because pions, kaons, and protons are abundantly produced in collisions. Wider tails arise when a particle is tested against an incorrect mass hypothesis.

In the proton distributions, additional structures around $n\sigma \approx 10$ and 20 can be observed. These structures are connected to the overlap of the TPC particle-identification responses of different particle species in certain momentum regions. As illustrated in Fig. 5.9, the Bethe–Bloch bands of pions, kaons, protons, and other particles are not fully separated over the whole momentum range. Therefore, tracks belonging to other particle species can appear at larger $n\sigma$ values when evaluated under the proton hypothesis.



(a) Combined PID response for the proton hypothesis of daughter track 0.

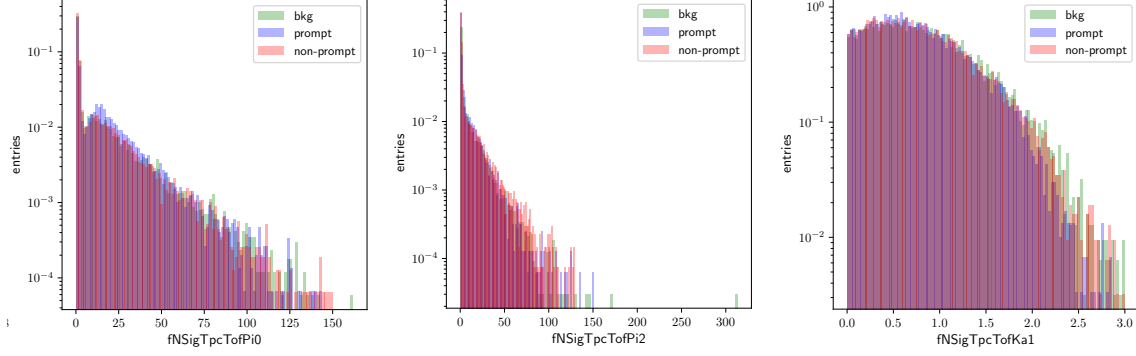
(b) Combined PID response for the proton hypothesis of daughter track 2.

Figure 5.10: Distributions of the combined PID variable for the proton hypothesis for positively charged daughter tracks. Prompt, non-prompt, and combinatorial background candidates are shown separately.

The PID performance evolves strongly with particle momentum, because the discriminating power of the TPC and TOF detectors is momentum dependent.

5.3 BDT-based Candidate Selection

To select Λ_c^+ candidates, a boosted decision tree (BDT) model implemented in the XGBoost library is employed [30]. The model performs a multiclass classification separating prompt signal, non-prompt signal, and combinatorial background. For each candidate, the trained classifier outputs three probability scores in the range $[0, 1]$, each representing the likelihood that the candidate belongs to the corresponding class.



(a) Pion hypothesis of daughter track 0. (b) Pion hypothesis of daughter track 2. (c) Kaon hypothesis of daughter track 1.

Figure 5.11: Distributions of the combined PID variable for pion and kaon hypotheses of the daughter tracks.

After applying the trained model to the candidate sample, only those with a prompt-signal probability exceeding the chosen working point (WP) are retained; the remainder are rejected. The WP is determined by scanning the expected significance as a function of the probability threshold, separately in each p_T interval. This procedure is described in more detail below.

Because the data are divided into ten p_T bins, a dedicated BDTs model is trained and optimised independently for each interval. The training and performance evaluation were carried out locally using the MLHEP analysis framework.

5.3.1 Training and Test Sets

The preparation of the training and testing datasets follows the general MLHEP procedure described in Sec. 4.2. In this thesis, the reconstructed Λ_c^+ candidates were divided into three classes: prompt signal, non-prompt signal, and combinatorial background. The prompt and non-prompt candidates were taken from MC simulations using their true labels, while the combinatorial background was extracted from real-data invariant-mass sidebands, outside the signal region defined by

$$2.22 < m_{\text{inv}}(\Lambda_c^+) < 2.35 \text{ GeV}/c^2.$$

Only 10% of the available background candidates were used, due to the much larger amount of combinatorial background in data. The labelled samples were prepared separately for each p_T interval and then split into training and testing subsets in an 80:20 ratio, while preserving the relative fraction of all three classes in both subsets.

5.3.2 Training ML Model

After the training and testing datasets were prepared and converted into a Python-readable format, the BDTs models were trained using the MachineLearningHEP framework. The training was performed locally on a computing cluster, separately

for each p_T interval. For every interval, the model was trained using the input variables described in Sec. 5.2.2. Since the analysis is divided into several p_T bins, this procedure resulted in one trained model for each interval.

The model was trained as a multi-class classifier, separating prompt Λ_c^+ candidates, non-prompt Λ_c^+ candidates, and combinatorial background. During the training, the classifier learns how to assign each candidate a probability of belonging to each of these three classes. The final selection is then based on the probability assigned to the prompt-signal class.

The candidate selection does not rely only on the individual values of the input variables, but also benefits from correlations and anti-correlations between them. This is one of the main advantages of a multivariate method compared to a simple cut-based selection. In Fig. 5.12, the correlation matrices for prompt, non-prompt, and combinatorial-background candidates are shown for the example interval $3 < p_T < 4$ GeV/ c . These matrices illustrate that the different candidate classes populate the multidimensional variable space in different ways.

Variables carrying similar physical information, such as those related to the decay length, are naturally strongly correlated. At the same time, some correlations differ noticeably between signal and background candidates. For example, the relation between the decay length, χ_{PCA}^2 , and the cosine of the pointing angle (CPA) provides important discriminating information. Genuine Λ_c^+ candidates are expected to have a more consistent decay topology, with a decay direction better aligned with the reconstructed momentum, while random background combinations usually show less consistent topological behaviour.

Although correlations between variables can improve the separation power of the model, the choice of input variables still has to be checked carefully. Including too many weakly useful or redundant variables can increase the complexity of the model and may lead to overtraining. This would appear as a visible difference between the model performance on the training and testing samples. Therefore, the importance of the input variables was also studied.

The feature importance was evaluated using SHAP values. SHAP, standing for SHapley Additive exPlanations, is a method based on cooperative game theory, where each variable is treated as a contributor to the final model decision. For each candidate, the SHAP value quantifies how much a given variable shifts the model output with respect to the average prediction. By averaging the absolute SHAP values over many candidates, one obtains a ranking of the variables according to their influence on the model response. Variables with larger mean absolute SHAP values therefore have a stronger impact on the classification.

The SHAP-based feature-importance ranking is shown in Fig. 5.13 for the transverse-momentum interval $3 < p_T < 4$ GeV/ c . This ranking was used as an additional check that the selected variables provide meaningful information for the separation of prompt Λ_c^+ candidates from non-prompt candidates and combinatorial background. For this example interval, the largest contributions are obtained from the decay-length related variables, in particular the decay length, the cosine of the pointing angle, and the decay length in the transverse plane. This indicates that the recon-

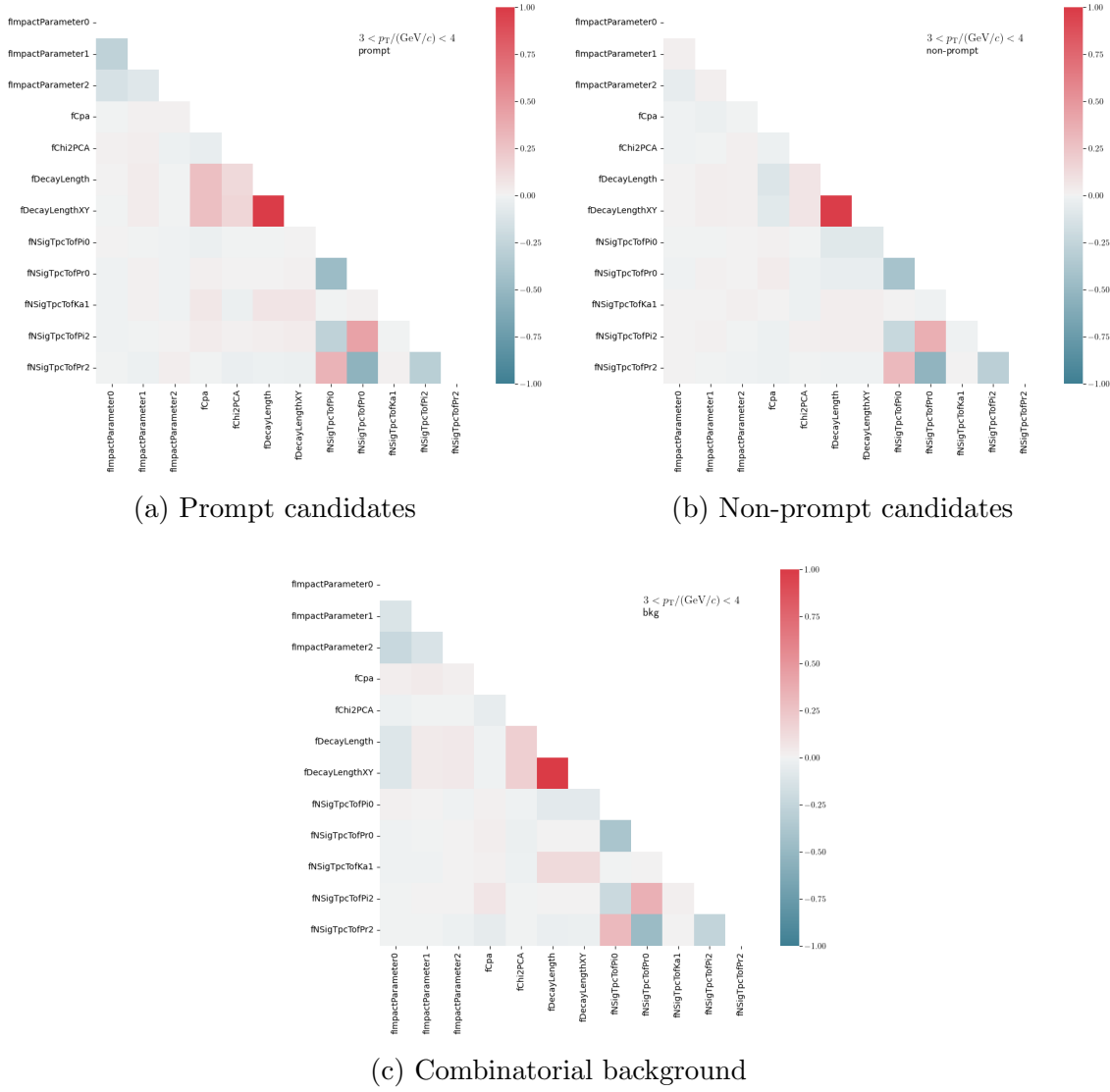


Figure 5.12: Correlation matrices of the input variables for prompt, non-prompt, and combinatorial-background Λ_c^+ candidates, shown for the example p_T interval $3 < p_T < 4$ GeV/ c .

structured decay topology provides the strongest information for the classification. The PCA quality, the impact parameters of the daughter tracks, and the PID variables based on the TPC $n\sigma$ response also contribute to the model decision, although with smaller importance. The trained model therefore combines both topological and PID information, with the dominant separation power in this interval coming mainly from variables describing the secondary-vertex geometry.

5.3.3 Performance Evaluation

After the training of the model is completed, the trained classifier is applied to both the training and independent test datasets. Using the obtained output scores, several performance distributions can be studied in order to evaluate the classification

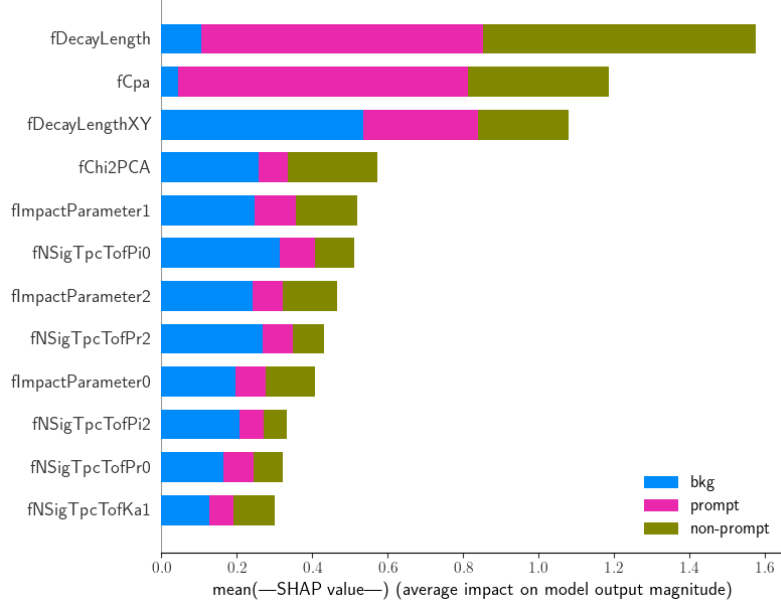


Figure 5.13: Feature-importance ranking based on the mean absolute SHAP values for the $\Lambda_c^+ \rightarrow p K^- \pi^+$ candidate selection in the transverse-momentum interval $3 < p_T < 4$ GeV/ c .

quality and possible overtraining effects. These checks are not the final physics result, but they are an important validation step before the model is applied to real data.

Firstly, the distributions of the ML output scores are shown in Fig. 5.14. The figure contains the three classifier responses corresponding to the probability of a candidate being classified as prompt Λ_c^+ , non-prompt Λ_c^+ , or combinatorial background. For each panel, candidates belonging to the respective true class tend to peak close to unity, while candidates from the other classes are shifted towards zero. This behaviour is observed consistently for both the training sample, shown as filled histograms, and the test sample, shown as markers. The agreement between the training and test distributions indicates that the model generalises well and does not show significant overtraining. If the model were strongly overtrained, a visible disagreement between the training and test distributions would be expected.

It can also be seen that the non-prompt candidates are more clearly separated from the prompt and background classes. This is most likely due to their displaced decay topology, which provides strong discriminating variables such as larger decay length and impact parameters. In contrast, the separation between prompt candidates and combinatorial background is less pronounced, suggesting that these two classes share more similar topological features. Further improvement could therefore be achieved by increasing the training statistics, refining the background sample, or introducing additional discriminating variables into the training procedure.

Next, the performance of the trained model must be evaluated. This is done using the ROC AUC metric, where ROC stands for Receiver Operating Characteristic and AUC denotes the Area Under the Curve. The ROC curve is constructed by plotting

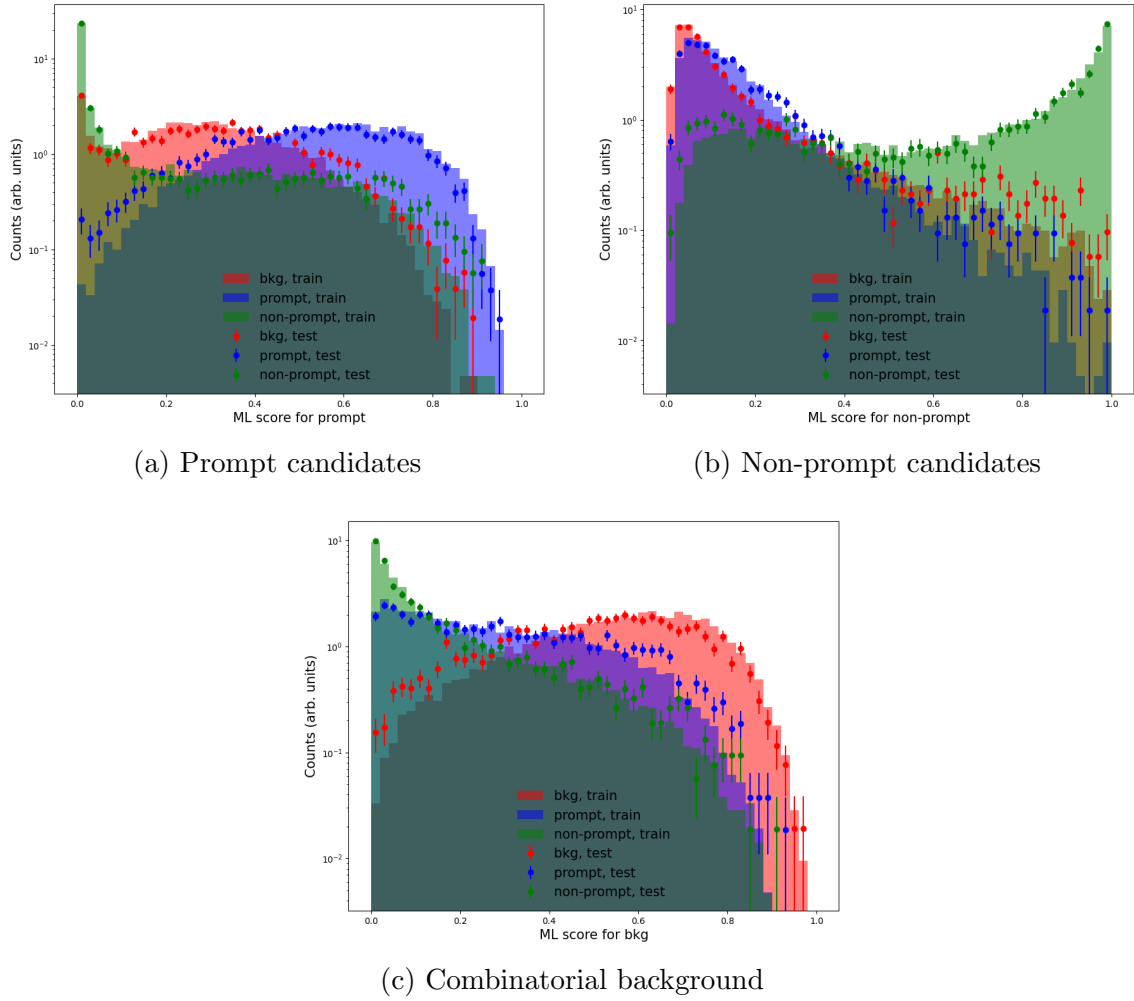


Figure 5.14: ML output score distributions for the three-class XGBoost classifier in the transverse momentum interval $3 < p_T < 4$ GeV/ c . Filled histograms correspond to the training sample, while markers represent the independent test sample. Additional figures for the other p_T intervals are provided in the Appendix in Figs. 7–9.

the true positive rate (TPR), also referred to as signal efficiency, as a function of the false positive rate (FPR), corresponding to the background selection efficiency, for different threshold values applied to the model output.

The true positive rate describes how many candidates of a given class are correctly identified by the model out of all true candidates belonging to that class, while the false positive rate quantifies how many candidates from other classes are incorrectly assigned to the probed category. These quantities are defined as

$$\text{TPR} = \frac{TP}{TP + FN}, \quad \text{FPR} = \frac{FP}{FP + TN},$$

where TP , FP , TN , and FN represent the numbers of true positives, false positives, true negatives, and false negatives, respectively.

The AUC value ranges from 0.5 to 1.0. A value of 0.5 corresponds to a random clas-

sifier with no discriminating power, while a value of 1.0 represents perfect separation between the two compared classes.

The ROC AUC results for the trained model in the interval $3 < p_T < 4$ GeV/ c are shown in Fig. 5.15. Since this is a multiclass classifier with three categories (prompt, non-prompt, and combinatorial background), the evaluation is performed using the One-vs-One (OvO) approach. This requires considering all pairwise combinations of classes in both directions. For each pair, one class is first treated as the first class and the other as the second class and then the assignment is reversed. This results in six ROC curves in total. For each pair, curves are shown separately for the training and testing datasets.

In the ideal case, all curves would rise sharply towards the upper-left corner, indicating high signal efficiency together with low false positive rate. Although the present classifier does not achieve perfect separation, all curves lie significantly above the diagonal behaviour expected from a random model, confirming meaningful discriminating power.

The weakest separation is observed for the prompt versus background pair, which is also the most relevant case for the present analysis. Ideally, this separation should be stronger, since the main goal of the selection is to suppress combinatorial background while keeping prompt Λ_c^+ candidates. Instead, the best performance is obtained for the prompt versus non-prompt and non-prompt versus background comparisons. This behaviour is not fully expected and may indicate that the model mainly learns variables connected to the displacement of non-prompt candidates, while the separation of prompt signal from background remains less efficient.

In addition, the testing curves are systematically below the corresponding training curves, especially for the prompt-background comparison. This may indicate mild overtraining or limited generalisation of the model. One possible reason is the limited size of the training sample, which can make the classifier more sensitive to statistical fluctuations in the training dataset rather than to robust physical differences between prompt signal and combinatorial background. Therefore, it would be useful to train the model on a larger dataset, which could reduce the sensitivity to statistical fluctuations and improve the agreement between the training and testing ROC curves.

5.3.4 Working Point Optimization

After the training procedure, the final step is to determine the optimal selection threshold applied to the BDTs output. In the present analysis, a cut is imposed on the maximum predicted probability for the combinatorial background class. Candidates with a large background score are rejected, while those more likely to correspond to signal are retained. The choice of this threshold is commonly referred to as the *working point* (WP).

The optimal working point should provide the best compromise between high sample purity and sufficient signal efficiency. In practice, this means maximising the expected statistical significance while avoiding an unnecessarily strong loss of sig-

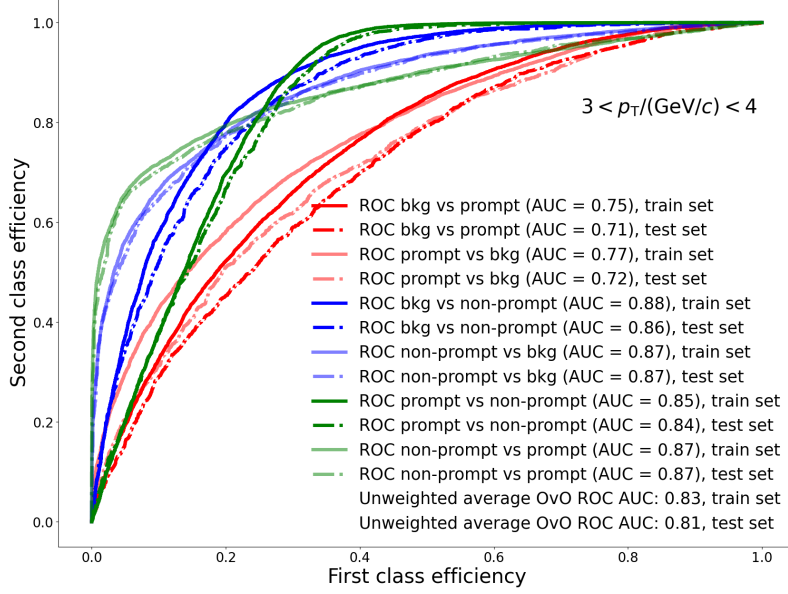


Figure 5.15: One-vs-One ROC curves for the XGBoost classifier in the interval $3 < p_T < 4$ GeV/c. The three classes considered are prompt Λ_c^+ , non-prompt Λ_c^+ , and combinatorial background. Solid lines correspond to the training dataset, while dashed lines represent the testing dataset.

nal candidates. For this purpose, both significance and efficiency-based performance curves are considered.

The statistical significance is defined as

$$Z = \frac{S}{\sqrt{S+B}}, \quad (5.13)$$

where S is the expected signal yield and B is the expected background yield after the applied BDTs selection.

The resulting dependence of the expected significance on the background threshold is shown in Fig. 5.16. In the ideal case, the working point would correspond to the global maximum of the curve. However, at very strict thresholds the number of surviving candidates becomes small, leading to larger statistical fluctuations and increased uncertainties. This behaviour is visible in the form of local jumps and enlarged error bars. Therefore, the absolute maximum should be interpreted with caution.

For this reason, the significance scan is complemented by the Precision–Recall versus threshold plot shown in Fig. 5.17. This figure illustrates how the classifier quality evolves as the applied background-score threshold is varied. Precision is defined as

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (5.14)$$

and quantifies the purity of the selected sample, i.e. the fraction of selected candi-

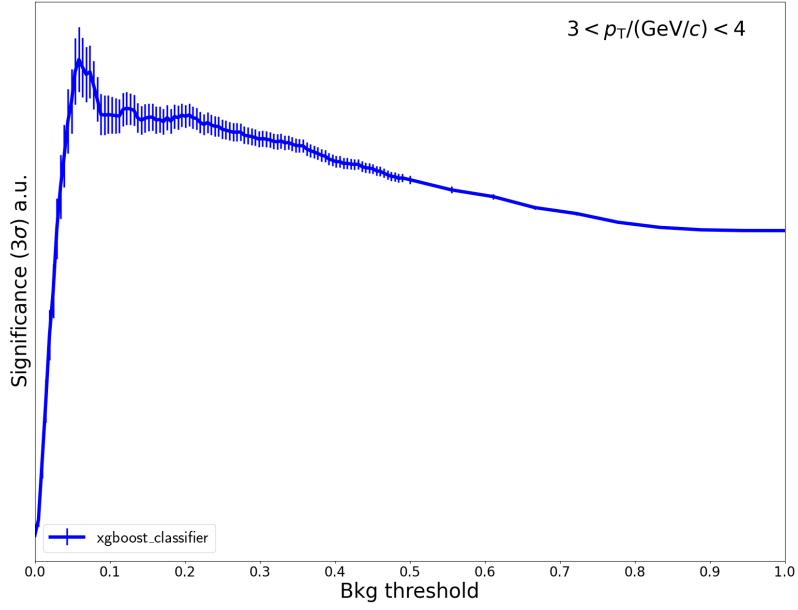


Figure 5.16: Expected statistical significance as a function of the applied background-score threshold for the interval $3 < p_T < 4$ GeV/ c . Additional figures for the other p_T intervals are provided in the Appendix in Fig. 10.

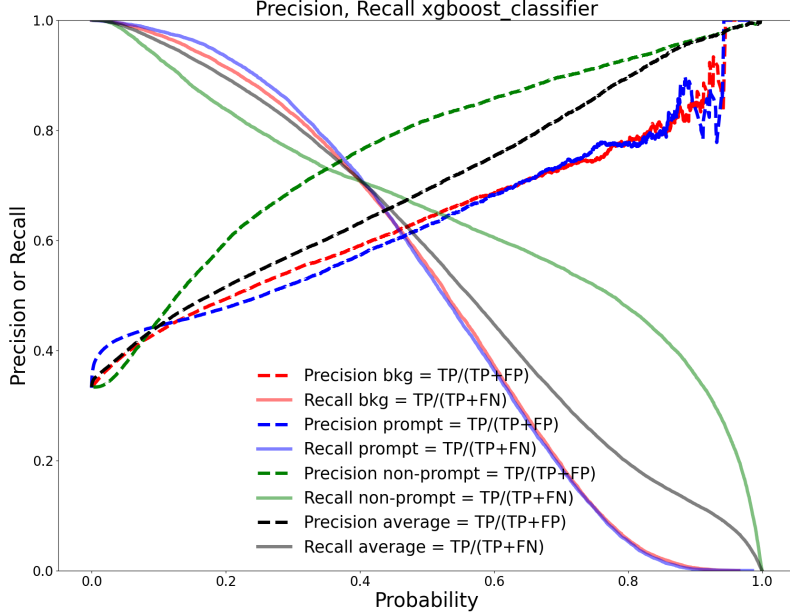


Figure 5.17: Precision and recall as functions of the applied classifier threshold for the XGBoost model in the interval $3 < p_T < 4$ GeV/ c . Precision describes the purity of the selected sample, while recall corresponds to the retained fraction of true candidates. Additional figures for the other p_T intervals are provided in the Appendix in Fig. 11.

dates that truly belong to the given class. Recall is defined as

$$\text{Recall} = \frac{TP}{TP + FN}, \quad (5.15)$$

and corresponds to the selection efficiency, i.e. the fraction of true signal candidates retained after the cut.

In the present optimisation, the probability shown on the horizontal axis is used as the scanned cut value and is equivalent to the background threshold in the significance figure. For each value of this cut, the corresponding precision and recall are evaluated. The recall of the prompt Λ_c^+ class is of particular importance, since it reflects how much signal is preserved after the selection. While a tighter cut generally improves the purity of the selected sample, it simultaneously reduces the signal efficiency. The final working point is therefore chosen in a region where the significance remains close to its maximum, while the prompt-candidate recall is still sufficiently high. In this way, a balance is achieved between background suppression and signal retention. The chosen optimal working point for each p_T bin is reported in Tab. 5.2.

p_T	[1, 2)	[2, 3)	[3, 4)	[4, 5)	[5, 6)	[6, 7)	[7, 8)	[8, 10)	[10, 12)	[12, 24)
Limit	25%	13%	6%	15%	14%	20%	32%	29%	24%	31%

Table 5.2: Significance and recall limits for different p_T intervals.

Chapter 6

Analysis

After the reconstruction of Λ_c^+ candidates and after applying the machine-learning based selection described in the previous chapter, the next step is to extract and study the reconstructed signal. In this chapter, the selected Run 3 data sample is analysed in order to obtain the raw yield of Λ_c^+ baryons in individual p_T intervals and to evaluate the performance of the applied selection strategy.

The trained model was applied to a different real-data sample than the one used during the training stage. The path to the used trained model stored on CCDB can be seen in the Appendix in Tab. 1. The final application of the trained model was performed on the independent dataset `HF_LHC24_pass1_sampled_2P3PDstar`, corresponding to proton–proton collisions at $\sqrt{s} = 13.6$ TeV recorded in 2024. This dataset contains approximately 4×10^9 analysed events and was centrally produced for heavy-flavour studies involving two-prong, three-prong, and D^* candidates. The term *sampled* indicates that the dataset contains only a selected fraction of events from the full production, taken across the available runs. In this way, the sample provides a representative subset of the larger dataset. Since the final application was performed on an independent real-data sample, this also provides an important check that the trained model can be transferred beyond the dataset used in the training procedure.

The Λ_c^+ baryon was reconstructed through the hadronic decay channel

$$\Lambda_c^+ \rightarrow pK^-\pi^+, \quad (6.1)$$

in the transverse-momentum range

$$1 < p_T < 24 \text{ GeV}/c. \quad (6.2)$$

The data were divided into separate p_T intervals shown in Tab. 5.2, and the invariant-mass distribution of selected candidates was studied independently in each of them.

The analysis is based on the fits of invariant-mass distributions, from which the signal contribution and the remaining combinatorial background are estimated. Together with the raw yield, the signal significance, signal-to-background ratio, and stability of the reconstructed mass peak are studied.

6.1 Invariant-mass Fit Procedure

The standard method used for the reconstruction of short-lived hadrons in collider experiments is based on the invariant mass of their decay daughters. For the decay channel $\Lambda_c^+ \rightarrow pK^-\pi^+$ all selected daughter-track combinations are formed, and their invariant mass is calculated by assigning to each track the particle mass expected for the studied decay channel.

If genuine Λ_c^+ baryons are present in the analysed data sample, their decays into the selected daughter particles introduce an additional contribution on top of the random combinatorial background. While background track combinations form a smooth invariant-mass distribution, daughter tracks originating from a real Λ_c^+ decay reconstruct preferentially around the mass of the parent particle. As a result, a localized peak is expected near the value

$$m_{\Lambda_c^+} = (2286.46 \pm 0.14) \text{ MeV}/c^2 [1]. \quad (6.3)$$

The measured invariant-mass distribution can therefore be described as the sum of two contributions: a signal component associated with genuine Λ_c^+ decays and a smooth background component originating from random track combinations. To extract the signal yield, each invariant-mass distribution was fitted using a combined function

$$f(m) = G(m) + B(m), \quad (6.4)$$

where $G(m)$ is a Gaussian function describing the signal peak,

$$G(m) = A \exp \left[-\frac{(m - \mu)^2}{2\sigma^2} \right], \quad (6.5)$$

and $B(m)$ is a second-order polynomial describing the combinatorial background:

$$B(m) = a + bm + cm^2. \quad (6.6)$$

The Gaussian mean μ represents the reconstructed peak position, σ describes the width of the reconstructed mass peak and the parameter A represents the height of the Gaussian peak. One way to determine the raw signal yield is to integrate the Gaussian component, so that the yield corresponds to the area under the signal peak.

Fits were performed independently in each p_T interval after applying the optimized machine-learning working-point selection. The resulting invariant-mass distributions in the analysed p_T intervals are shown in Fig. 6.1. In all intervals, a localized structure is visible close to the expected Λ_c^+ mass. The signal peak is most clearly visible in the intermediate- p_T intervals, where the separation between signal and background is most efficient and the available statistics are still sufficient.

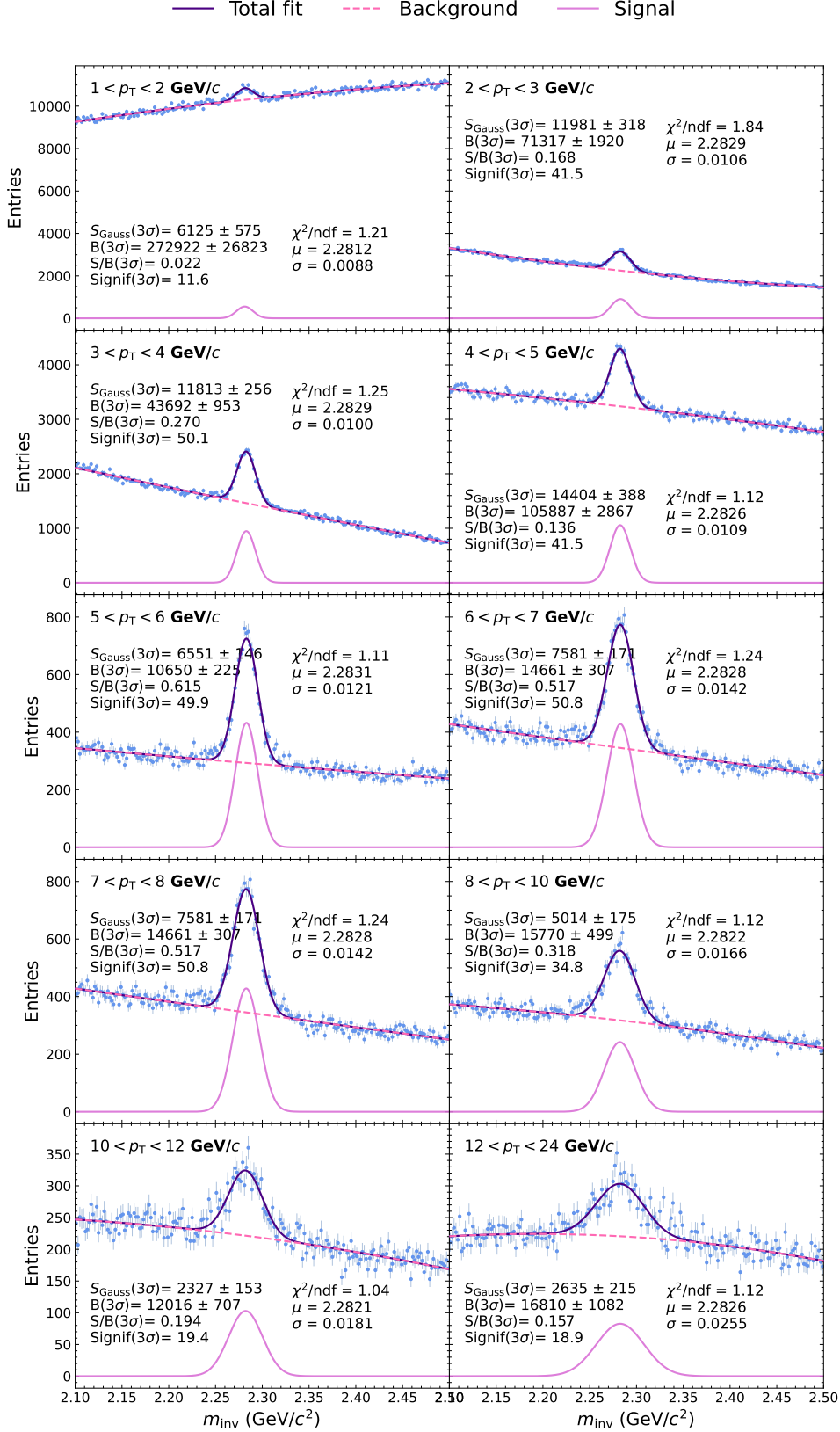


Figure 6.1: Invariant-mass distributions of Λ_c^+ candidates in the analysed p_T intervals after applying the selected cuts on the background BDTs score. The signal contribution is described by a Gaussian function, while the background is fitted with a second-order polynomial.

6.2 Signal Significance

To quantify the quality of the reconstructed signal, the statistical significance was evaluated using Eq. (5.13). It describes how clearly the signal can be distinguished from combinatorial background. It is one of the most important quantities when optimizing heavy-flavour analyses, since the signal must not only be large, but also sufficiently separated from the background.

The significance obtained from the invariant-mass distributions is shown in Fig. 6.2 as a function of transverse momentum. The expected behaviour is that the significance is relatively low in the lowest- p_T intervals, then increases rapidly as the quality of the reconstructed tracks and the separation power improve, and finally decreases again at high p_T due to the limited number of reconstructed candidates. This trend can be seen in the figure. At low transverse momentum, the significance is smaller because the reconstruction of low- p_T tracks is more challenging. The daughter tracks have smaller momenta and are more affected by detector resolution effects, multiple scattering, and less precise reconstruction of the decay topology. As a result, the separation between genuine Λ_c^+ candidates and random track combinations is weaker, and the combinatorial background is less efficiently suppressed.

With increasing p_T , the track reconstruction becomes more precise and the decay topology is better defined. The PID in this region provides the strongest separation between signal and background candidates, together with over all good topology variables quality. This leads to a strong increase of the significance and to the best performance in the intermediate- p_T region, where the balance between signal yield and background suppression is most favourable. At higher transverse momentum, the topological separation of the candidates become the strongest together with very good daughter track measurement. However, the production yield of Λ_c^+ baryons decreases with increasing p_T , which reduces the available statistics. As a consequence, the significance starts to decrease again in the highest- p_T intervals despite the improved purity of the selected sample.

One exception to this otherwise smooth trend is observed in the $4 < p_T < 5$ GeV/ c interval, where the significance is lower than in the neighbouring bins. This behaviour may be related to the chosen BDTs working point in this interval, but it could also indicate a less optimal response of the trained model in this particular kinematic region.

6.3 Signal-to-background Ratio

In addition to the statistical significance, the signal-to-background ratio was studied as a function of transverse momentum. It is defined as

$$\frac{S}{B}, \tag{6.7}$$

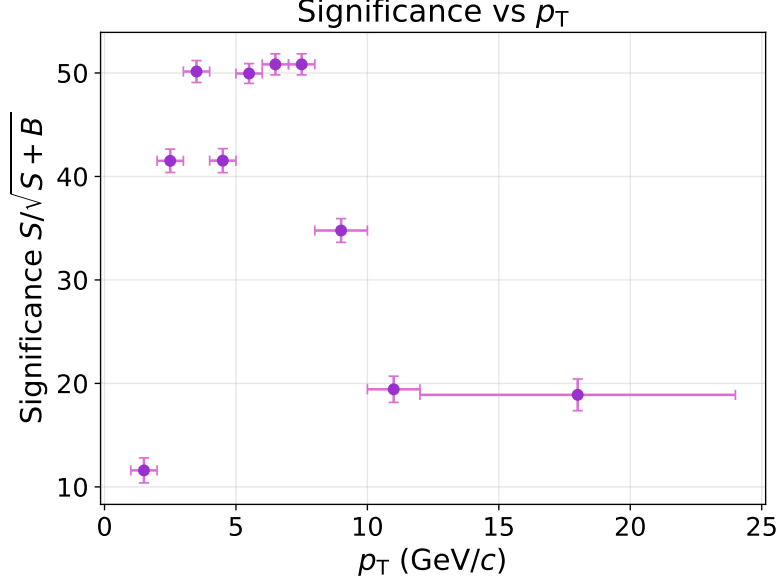


Figure 6.2: Significance of the reconstructed Λ_c^+ signal as a function of p_T after applying the optimized machine-learning selection. The significance increases from low p_T towards the intermediate region, where the track quality, PID performance, and signal yield provide the most favourable conditions. At higher p_T , the significance decreases due to limited statistics.

where S is the extracted signal yield and B is the background contribution in the signal region. While the significance describes how clearly the signal can be distinguished from statistical background fluctuations, the S/B ratio gives direct information about the purity of the selected candidate sample.

The obtained S/B values are shown in Fig. 6.3. The highest values are observed in the intermediate- p_T region, where the overall candidate separation is most efficient. In this interval, the daughter tracks are sufficiently energetic to provide a cleaner decay topology, while particle-identification performance remains strong. At higher transverse momentum, the S/B ratio remains better than in the lowest- p_T bins, indicating that the physical separation between signal and background candidates is still relatively good. In contrast, the lowest- p_T region exhibits the worse purity. This is caused by the combination of large combinatorial background and weaker separation power. In addition, the reconstructed candidates are more affected by the dense low-momentum track environment, which further increases the background contribution.

6.4 Mass Peak Stability

An important check of the reconstruction quality is the stability of the fitted mass peak position as a function of transverse momentum. If the reconstruction is well conducted, the extracted mean mass μ should remain compatible with the world-average value shown in Eq. (6.3) within uncertainties for all p_T intervals. The fitted

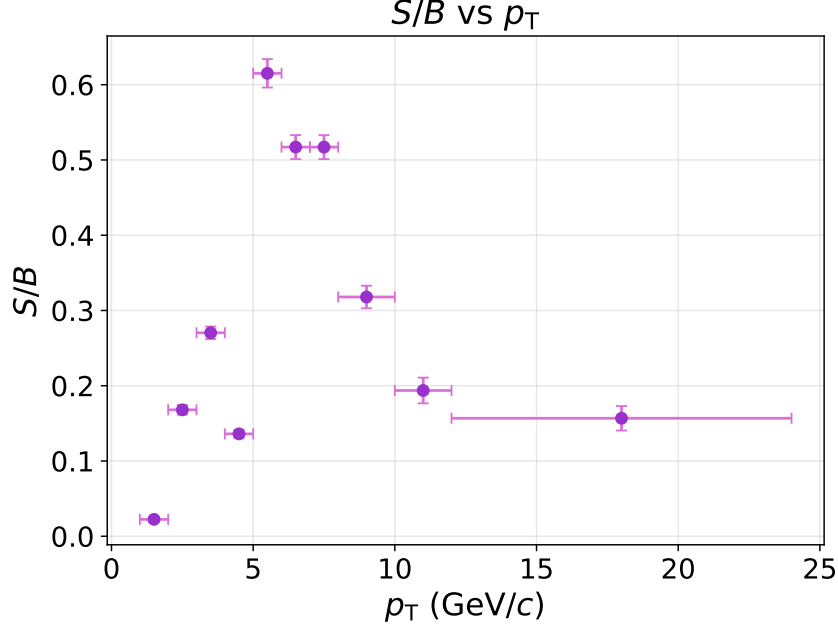


Figure 6.3: Signal-to-background ratio as a function of transverse momentum for selected Λ_c^+ candidates after the optimized machine-learning selection.

mean mass values obtained from the Gaussian component of the invariant-mass fits are shown in Fig. 6.4. The values remain approximately stable as a function of p_T and do not show a strong systematic trend with increasing transverse momentum. From the values shown in the figure, the average reconstructed mass was determined to be

$$\langle m_{\Lambda_c^+} \rangle = (2.28 \pm 0.05) \text{ GeV}/c^2. \quad (6.8)$$

This value is compatible with the expected Λ_c^+ mass shown in Eq. (6.3) within the obtained uncertainty. The stability of the peak position supports the interpretation that the selected candidates correspond to the targeted decay channel and indicates that no strong momentum-dependent bias is introduced by the machine-learning selection.

6.5 Raw Yield Extraction

The main output of the present analysis is the raw yield of reconstructed Λ_c^+ baryons in each p_T interval. In this work, the raw yield is obtained from the fitted signal component of the invariant-mass distribution. After fitting the distribution with a Gaussian function for the signal and a second-order polynomial for the combinatorial background, the raw yield is calculated as the integral of the Gaussian signal contribution within the signal region.

The signal region is defined as $\mu \pm 3\sigma$ around the reconstructed Λ_c^+ peak, where μ and σ are the mean and width of the fitted Gaussian. The raw yield can therefore

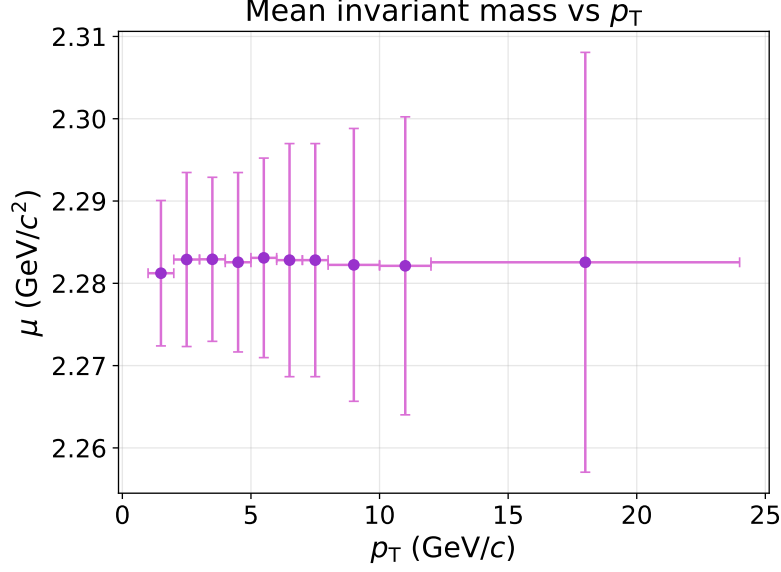


Figure 6.4: Mean invariant mass of the reconstructed Λ_c^+ candidates as a function of p_T , obtained from the Gaussian component of the invariant-mass fits. The points are approximately stable across the analysed p_T intervals and give an average reconstructed mass of (2.28 ± 0.05) GeV/ c^2 .

be written schematically as

$$N_{\text{raw}} = \int_{\mu-3\sigma}^{\mu+3\sigma} G(m) dm, \quad (6.9)$$

where $G(m)$ is the Gaussian signal function obtained from the fit. In this way, the extracted raw yield corresponds to the fitted signal area in the selected mass interval, rather than to a direct bin counting of the background-subtracted histogram. Therefore, the raw yield is less sensitive to statistical fluctuations in individual mass bins and to the exact binning of the histogram.

The extracted raw yield therefore represents the number of candidates remaining after subtracting the estimated combinatorial background from the measured invariant-mass distribution. At this stage, it does not yet correspond to the physical production yield, since acceptance, efficiency, branching-ratio, prompt/feed-down, and normalization corrections still need to be applied.

The raw yield as a function of transverse momentum is shown in Fig. 6.5. The distribution reflects both the physical production of Λ_c^+ baryons and the reconstruction performance in different p_T intervals. The yield is smaller in the lowest- p_T interval, where track reconstruction and candidate selection are more challenging. It then increases towards the intermediate- p_T region, where the reconstruction quality is better. At higher transverse momentum, the raw yield decreases again, due to the production of Λ_c^+ baryons becoming less abundant.

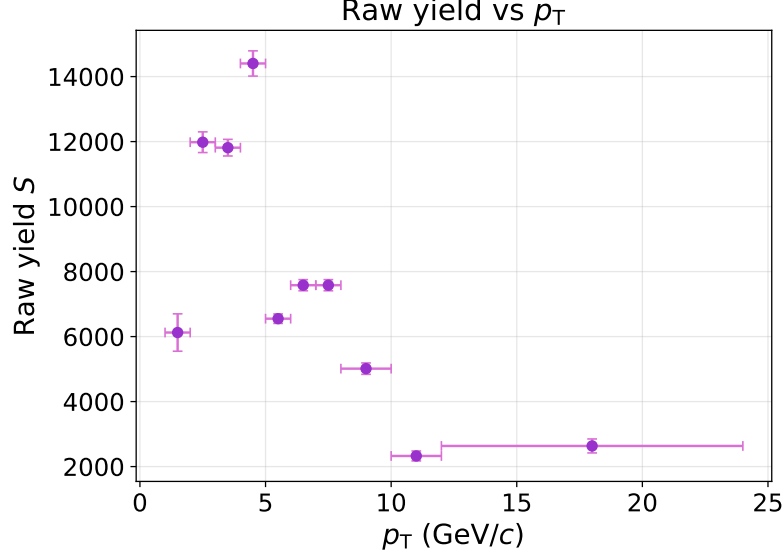


Figure 6.5: Raw yield of reconstructed Λ_c^+ candidates as a function of p_T .

6.6 Corrections

After the raw yield of Λ_c^+ baryons has been extracted from the invariant-mass fits, it still cannot be directly interpreted as the physical production yield. The raw yield represents only the number of reconstructed candidates that survived the full analysis procedure. It is therefore affected by the limited geometrical acceptance of the detector, by the reconstruction efficiency, by the selection efficiency, and by the efficiency of the machine-learning based selection. These effects should be corrected for before a physical yield or cross section can be obtained. The transition from the measured raw yield to the physical production cross section is performed using Eq. (1.4). The raw yield N_{raw} is corrected for detector acceptance and reconstruction efficiency by the factor $(\text{Acc} \times \varepsilon)$, for the prompt fraction by f_{prompt} , and for the branching ratio of the reconstructed decay channel by BR. The result is then normalised by the analysed integrated luminosity $\mathcal{L}_{\text{int}}$ and by the width of the p_T interval.

In the present thesis, the main focus is on the reconstruction of the Λ_c^+ signal, the application of the trained machine-learning model, and the extraction of raw yields in individual p_T intervals. The full correction procedure has not yet been performed. Therefore, the following sections describe the correction steps that should be applied in a future continuation of the analysis in order to obtain the corrected yield or production cross section.

6.6.1 Acceptance Times Efficiency Correction

The acceptance times efficiency correction, usually denoted as $(\text{Acc} \times \varepsilon)$, accounts for the fact that not all produced particles are reconstructed and selected by the analysis. The acceptance describes the fraction of produced particles whose decay

products fall into the geometrical and kinematic region covered by the detector. The efficiency describes the probability that such particles are reconstructed, pass the selection criteria, and survive the machine-learning based selection.

The correction factor should be obtained from Monte Carlo simulations. In the planned continuation of this analysis, the same reconstruction and analysis procedure as used for real data should be applied to the simulated sample. This means that the Λ_c^+ candidates should be reconstructed, the trained machine-learning model should be applied, the same working point should be used, and the reconstructed raw yield should be extracted in each p_T interval. The number of reconstructed and selected candidates should then be compared with the number of generated Λ_c^+ baryons in the simulation.

For a given p_T interval, the correction factor can be written as

$$(\text{Acc} \times \varepsilon)(p_T) = \frac{N_{\text{reco}}(p_T)}{N_{\text{gen}}(p_T)}, \quad (6.10)$$

where $N_{\text{gen}}(p_T)$ is the number of generated Λ_c^+ baryons in the Monte Carlo sample and $N_{\text{reco}}(p_T)$ is the number of reconstructed Λ_c^+ candidates that pass the full analysis selection in the same p_T interval.

After obtaining the correction factor, the corrected raw yield should be calculated as

$$N_{\text{corr}}(p_T) = \frac{N_{\text{raw}}(p_T)}{(\text{Acc} \times \varepsilon)(p_T)}, \quad (6.11)$$

where $N_{\text{raw}}(p_T)$ is the raw yield extracted from the invariant-mass fit in real data.

6.6.2 Feed-down Correction

The raw yield extracted from data contains both prompt and non-prompt Λ_c^+ baryons. However, in this measurement the goal is to obtain the prompt Λ_c^+ production yield. Therefore, the contribution from non-prompt Λ_c^+ baryons originating from beauty-hadron decays has to be accounted for.

Feed-down refers to Λ_c^+ baryons that are not produced directly in the charm-quark hadronisation process, but instead come from decays of beauty hadrons. These candidates are reconstructed in the same invariant-mass region as prompt Λ_c^+ baryons and therefore contribute to the raw signal yield. Since they have a different physical origin, they have to be treated separately.

The raw yield can be written schematically as

$$N_{\text{raw}} = N_{\text{raw}}^{\text{prompt}} + N_{\text{raw}}^{\text{feed-down}}. \quad (6.12)$$

In the present analysis, the machine-learning model was trained with separate classes

for prompt Λ_c^+ baryons, non-prompt Λ_c^+ baryons, and combinatorial background. Therefore, the treatment of the feed-down contribution differs from an approach where prompt and non-prompt candidates are included in a single signal class. After applying the trained model and the selected working points, the relative prompt and non-prompt contributions can be evaluated separately in each p_T interval.

This separation can be expressed through the prompt fraction

$$f_{\text{prompt}}(p_T) = \frac{N_{\text{raw}}^{\text{prompt}}(p_T)}{N_{\text{raw}}^{\text{prompt}}(p_T) + N_{\text{raw}}^{\text{feed-down}}(p_T)}. \quad (6.13)$$

The prompt fraction represents the part of the measured raw signal yield that is attributed to prompt Λ_c^+ baryons after the ML-based selection. It is then used to remove the estimated feed-down contribution from the raw yield.

The prompt corrected yield can therefore be written as

$$N_{\text{corr}}^{\text{prompt}}(p_T) = \frac{f_{\text{prompt}}(p_T) \cdot N_{\text{raw}}(p_T)}{(\text{Acc} \times \varepsilon)_{\text{prompt}}(p_T)}, \quad (6.14)$$

where $(\text{Acc} \times \varepsilon)_{\text{prompt}}$ is the acceptance times efficiency correction for prompt Λ_c^+ baryons. The value of f_{prompt} is generally p_T dependent, because the relative contribution of non-prompt Λ_c^+ baryons and the corresponding reconstruction and selection efficiencies change with transverse momentum [31].

6.6.3 Cross-section Normalisation

To convert the corrected raw yield into a p_T -differential production cross section, the result has to be normalised by the branching ratio, the integrated luminosity, and the width of the p_T interval. The branching ratio accounts for the fact that only the decay channel $\Lambda_c^+ \rightarrow pK^-\pi^+$ and its charge-conjugate decay are reconstructed. The value used for this channel is

$$\mathcal{B}(\Lambda_c^+ \rightarrow pK^-\pi^+) = 6.28\%. \quad (6.15)$$

The integrated luminosity $\mathcal{L}_{\text{int}}$ represents the size of the analysed data sample and has to correspond to the same dataset from which the raw yields were extracted. The factor Δp_T is the width of the transverse-momentum interval and converts the yield in a finite bin into a differential quantity. The factor $1/2$ is included because Λ_c^+ and $\overline{\Lambda}_c^+$ candidates are reconstructed together, while the cross section is quoted as the average of the two charge states.

Combining these factors, the p_T -differential production cross section is obtained as

$$\frac{d\sigma}{dp_T} = \frac{1}{2\Delta p_T} \frac{N_{\text{raw}}(p_T) f_{\text{prompt}}(p_T)}{(\text{Acc} \times \varepsilon)_{\text{prompt}}(p_T) \text{BR} \mathcal{L}_{\text{int}}}. \quad (6.16)$$

Here, N_{raw} is the extracted raw yield, f_{prompt} corrects for the non-prompt feed-down contribution, $(\text{Acc} \times \varepsilon)_{\text{prompt}}$ corrects for acceptance, reconstruction, and selection efficiency, BR is the branching ratio, and $\mathcal{L}_{\text{int}}$ is the integrated luminosity.

6.7 Systematic Uncertainties

Systematic uncertainties describe possible shifts of the measured result caused by choices made in the analysis procedure, by detector effects, or by assumptions used in the correction. Unlike statistical uncertainties, they do not originate from the finite number of measured candidates, but from the method used to extract and correct the signal.

In a complete measurement, one important source of systematic uncertainty would come from the raw-yield extraction. The raw yield is obtained by fitting the invariant-mass distribution in each p_{T} interval. The result can depend on the chosen fit range, the function used to describe the background, the treatment of the signal width and mean, and the histogram binning. To estimate this uncertainty, the raw-yield extraction should be repeated with several reasonable variations of the fit procedure. The spread of the obtained raw yields should then be used as a systematic uncertainty.

Another source of uncertainty is related to the machine-learning selection. The final result can depend on the chosen working point of the trained model. Therefore, the analysis should be repeated for different BDT-score thresholds around the nominal working point. If the corrected yield remains stable within uncertainties, the selection can be considered robust. The observed variation should be included as a systematic uncertainty associated with the machine-learning selection.

The acceptance times efficiency correction also introduces a systematic uncertainty. This uncertainty comes from possible differences between Monte Carlo simulations and real data. The simulated sample should reproduce the detector conditions, the reconstruction performance, the multiplicity distribution, and the kinematic distributions of the particles. If these quantities are not perfectly described, the calculated efficiency correction may be biased. This effect can be studied by varying the Monte Carlo reweighting procedure, comparing different simulated samples, or changing the selection criteria used in the correction.

For the prompt yield, an additional systematic uncertainty comes from the feed-down correction. Since the non-prompt contribution is estimated using theoretical calculations and simulations, it depends on the assumed beauty-production cross section, the decay kinematics, and the efficiency of non-prompt Λ_c^+ reconstruction. Variations of these assumptions should be used to estimate the corresponding uncertainty.

The final systematic uncertainty should be obtained by combining the relevant contributions. Each contribution should be evaluated separately, and the total uncertainty is usually calculated as the quadratic sum of the individual independent sources:

$$\sigma_{\text{syst}}^{\text{tot}} = \sqrt{\sum_i (\sigma_{\text{syst}}^i)^2}. \quad (6.17)$$

This procedure would ensure that the final corrected yield does not depend strongly on one specific technical choice of the analysis. The systematic uncertainty therefore gives an estimate of how robust the measured result is with respect to the assumptions and variations of the full analysis procedure [31].

Conclusion and Outlook

Conclusion

In this work, the reconstruction of the Λ_c^+ baryon in the decay channel

$$\Lambda_c^+ \rightarrow pK^-\pi^+$$

was performed using Run 3 proton–proton collision data at $\sqrt{s} = 13.6$ TeV. The main goal of the analysis was to reconstruct the Λ_c^+ signal after applying a machine-learning based selection and to study the raw signal extraction in individual p_T intervals. The analysis therefore represents an important intermediate step towards a full measurement of Λ_c^+ production in Run 3 data.

The signal was selected using a Boosted Decision Tree model trained to separate prompt Λ_c^+ baryons, non-prompt Λ_c^+ baryons, and combinatorial background candidates. The prompt and non-prompt signal candidates were taken from Monte Carlo simulations, while the background candidates were taken from real-data sidebands. This approach was chosen in order to make the description of the combinatorial background closer to the conditions in real data. After training, the model was applied to an independent real-data sample. This is an important test of the model performance, since a trained classifier should be able to generalise to data that were not used during training. The obtained invariant-mass distributions and reconstructed signal show that this approach works reasonably well in this analysis. The use of a multivariate selection allowed several topological and PID-related variables to be combined in a single classifier response, which improved the separation between signal-like and background-like candidates compared to a simple rectangular-cut selection.

The performance of the machine-learning selection was found to depend on the transverse-momentum interval. The best behaviour was observed in the intermediate- p_T region, where the analysis benefits from sufficient statistics, good PID separation, and favourable reconstruction conditions. In this region, the agreement between the training and testing distributions was generally good, as shown in the Appendix in Fig. 7. This indicates that the trained model provides a stable description of the selected candidates and does not show strong signs of overtraining in the most important p_T intervals. At low and high p_T , the agreement was less optimal. These regions are more challenging because low transverse momentum is affected by the more difficult reconstruction of low-momentum daughter tracks, while high transverse momentum is mainly limited by the available statistics. They would therefore

benefit from larger training samples, further optimisation of the input variables, and a more detailed tuning of the working points.

After applying the selected working points shown in Tab. 5.2, the invariant-mass distributions showed a visible peak around the expected Λ_c^+ mass, as shown in Fig. 6.1. The reconstructed mean mass was compatible with the known Λ_c^+ mass within uncertainties, and its stability as a function of p_T was studied in Fig. 6.4. This confirms that the reconstructed candidates are consistent with the expected signal and that the selected sample can be used for raw-yield extraction.

The raw yield was extracted in individual p_T intervals from fits to the invariant-mass distributions. The signal contribution was described by a Gaussian function, while the combinatorial background was described by a second-order polynomial. The raw yield was then obtained from the fitted signal component in the region around the reconstructed peak. The dependence of the raw yield and signal significance on p_T , shown in Fig. 6.5 and Fig. 6.2, follows the expected behaviour. The yield is largest in the region where the available statistics and reconstruction performance are most favourable, while it decreases towards higher transverse momentum due to the limited number of reconstructed candidates. At low transverse momentum, the extraction is mainly limited by the more challenging reconstruction of low-momentum daughter tracks.

A similar Λ_c^+ reconstruction and production analysis is being performed within the ALICE Collaboration. The results obtained in this work are well compatible with the expected behaviour. However, a direct quantitative comparison should be performed only after the full correction procedure is applied. The results presented here are based on raw yields, which are not yet corrected for detector acceptance, reconstruction efficiency, selection efficiency, and branching ratio. In addition, the machine-learning model used in this work was trained independently, using a specific choice of signal and background samples, and was applied to a different real-data sample than the one used in the reference analysis. These differences can affect the selected candidate sample and the resulting raw yields. Therefore, the present comparison should be understood as a validation of the reconstruction and signal-extraction strategy.

The importance of this analysis lies in its connection to the broader heavy-flavour physics programme of ALICE. Measurements of charm-baryon production provide important information about charm-quark hadronisation. In particular, the production of Λ_c^+ baryons relative to charm mesons is sensitive to the hadronisation mechanism, namely whether charm quarks hadronise mainly through fragmentation, as expected in leptonic collisions, or whether additional mechanisms such as recombination also contribute. Precise measurements of Λ_c^+ production in proton–proton collisions are therefore needed as a reference for studies in proton–nucleus and nucleus–nucleus collisions. The Run 3 proton–proton data sample at $\sqrt{s} = 13.6$ TeV provides a new opportunity to improve this reference measurement. Compared to previous data-taking periods, Run 3 offers larger statistics and benefits from the upgraded reconstruction framework, which can improve the precision of charm-hadron measurements.

Outlook

The present analysis represents a first step towards a full production measurement of Λ_c^+ baryons in Run 3 proton–proton collisions at $\sqrt{s} = 13.6$ TeV. The results presented in this work are based on raw yields and therefore cannot yet be interpreted as corrected physical yields. Several additional steps are needed in order to complete the measurement.

- **Efficiency correction.** The acceptance and efficiency correction, $(\text{Acc} \times \varepsilon)$, has to be determined from Monte Carlo simulations by applying the same trained BDTs model as in real data. The number of reconstructed and selected Λ_c^+ candidates should then be compared with the number of generated Λ_c^+ baryons in each p_T interval.
- **Prompt and non-prompt separation.** The non-prompt contribution from beauty-hadron decays has to be estimated in order to determine the prompt fraction f_{prompt} . Since the ML model separates prompt and non-prompt Λ_c^+ candidates, this information can be used to evaluate their relative contributions after the selected working points are applied.
- **Cross-section normalisation.** The corrected yields have to be normalised by the branching ratio, the integrated luminosity $\mathcal{L}_{\text{int}}$, and the p_T -bin width Δp_T . The factor $1/2$ also has to be included because Λ_c^+ and $\bar{\Lambda}_c^+$ candidates are reconstructed together, while the cross section is quoted as the average of the two charge states.
- **Systematic uncertainties.** Systematic uncertainties should be evaluated, especially those related to the raw-yield extraction, fit range and background function, BDTs working-point selection, data–MC differences, and the feed-down correction.
- **Further optimisation of the ML selection.** The BDTs performance could be improved with larger training samples and further tuning of the working points. This is especially relevant at low p_T , where track reconstruction is more challenging, and at high p_T , where the analysis is limited by statistics.
- **Extension to corrected production measurements.** With the full correction procedure and larger Run 3 statistics, the analysis can be extended towards a corrected p_T -differential Λ_c^+ production cross-section measurement at $\sqrt{s} = 13.6$ TeV.

After these steps, the analysis would provide a more complete physics result. In particular, it could contribute to an improved understanding of charm-baryon production in Run 3 data and demonstrate the potential of machine-learning based candidate selection in heavy-flavour analyses.

Bibliography

- [1] S. Navas *et al.*, “Review of particle physics,” *Phys. Rev. D*, vol. 110, no. 3, p. 030001, 2024.
- [2] M. E. Peskin and D. V. Schroeder, *An Introduction to Quantum Field Theory*. Reading, Massachusetts: Addison-Wesley, 1995.
- [3] J. C. Collins *et al.*, “Factorization of hard processes in qcd,” *Adv. Ser. Direct. High Energy Phys.*, vol. 5, pp. 1–91, 1989.
- [4] S. Acharya *et al.*, “The ALICE experiment: a journey through QCD,” *Eur. Phys. J. C*, vol. 84, no. 8, p. 813, 2024.
- [5] S. Acharya *et al.*, “Charm production and fragmentation fractions at midrapidity in pp collisions at $\sqrt{s} = 13$ tev,” *JHEP*, vol. 12, p. 086, 2023.
- [6] J. I. Kapusta and C. Gale, *Finite-Temperature Field Theory: Principles and Applications*. Cambridge University Press, 2006.
- [7] S. Acharya *et al.*, “Direct observation of the dead-cone effect in quantum chromodynamics,” *Nature*, vol. 605, no. 7910, pp. 440–446, 2022.
- [8] S. Acharya *et al.*, “Measurement of d -meson production at mid-rapidity in pp collisions at $\sqrt{s} = 5.02$ tev,” *Eur. Phys. J. C*, vol. 77, no. 550, p. 550, 2017.
- [9] S. Acharya *et al.*, “Measurement of prompt d -meson production in p -pb collisions at $\sqrt{s_{NN}} = 5.02$ tev,” *Phys. Rev. Lett.*, vol. 113, no. 232301, p. 232301, 2014.
- [10] S. Acharya *et al.*, “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*, vol. 107, no. 6, p. 064901, 2023.
- [11] S. Acharya *et al.*, “Measurement of charm production at central rapidity in proton–proton collisions at $\sqrt{s} = 7$ tev,” *J. High Energy Phys.*, vol. 01, p. 128, 2012.
- [12] J. C. Collins, *Foundations of Perturbative QCD*. Cambridge University Press, 2011.
- [13] “Suppression of high transverse momentum d mesons in central pb–pb collisions at $\sqrt{s_{NN}} = 2.76$ tev,” *J. High Energy Phys.*, vol. 09, p. 148, 2015.

- [14] S. Acharya *et al.*, “Charm fragmentation fractions and $c\bar{c}$ cross section in p - p b collisions at $\sqrt{s_{NN}} = 5.02$ tev,” *Eur. Phys. J. C*, vol. 84, no. 12, p. 1286, 2024.
- [15] L. Dello Stritto, “Studying charm and beauty quark production and hadronisation with charm hadron measurements with alice,” 4 2025. ALICE.
- [16] S. Acharya *et al.*, “Observation of a multiplicity dependence in the p_t -differential charm baryon-to-meson ratios in proton-proton collisions at $\sqrt{s} = 13$ tev,” *Phys. Lett. B*, vol. 829, p. 137065, 2022.
- [17] S. Acharya *et al.*, “Study of flavor dependence of the baryon-to-meson ratio in proton-proton collisions at $\sqrt{s} = 13$ tev,” *Phys. Rev. D*, vol. 108, no. 11, p. 112003, 2023.
- [18] CERN, “Alice detector overview,” 2017.
- [19] J. Alme *et al.*, “The alice tpc, a large 3-dimensional tracking device with fast readout for ultra-high multiplicity events,” *Nucl. Instrum. Methods Phys. Res. A*, vol. 622, pp. 316–367, 2010.
- [20] F. Carnesecchi, “Performance of the alice time-of-flight detector at the lhc,” *JINST*, vol. 14, no. 06, p. C06023, 2019.
- [21] AliceO2Group, “Getting started · alice o2 documentation,” 3 2024.
- [22] V. Kucera, “Introduction to hf framework in o2 and validation framework.” HF O2 software hackathon, 2021.
- [23] ALICE O2 Analysis Framework, “Hyperloop user documentation.” <https://aliceo2group.github.io/analysis-framework/docs/hyperloop/userdocumentation.html>. Accessed: 2026-05-03.
- [24] K. Albertsson *et al.*, “Machine learning in high energy physics community white paper,” *J. Phys. Conf. Ser.*, vol. 1085, no. 2, p. 022008, 2018.
- [25] ALICE Collaboration, “MachineLearningHEP – machine learning for high energy physics in the alice experiment,” 2023. GitHub repository.
- [26] S. Navas *et al.*, “Review of particle physics,” *Phys. Rev. D*, vol. 110, no. 3, p. 030001, 2024.
- [27] K. Aamodt *et al.*, “The ALICE experiment at the CERN LHC,” *JINST*, vol. 3, p. S08002, 2008.
- [28] L. D. Stritto and M. Karwowska, “Measurement of the Λ_c^+ cross section (integrated and vs multiplicity) in pp collisions at $\sqrt{s} = 13.6$ tev,” alice analysis note, CERN, Geneva, 2025.
- [29] W. Yu, “Particle identification of the alice tpc via dE/dx ,” *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 706, pp. 55–58, 2013. TRDs for the Third Millenium.

- [30] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16, pp. 785–794, ACM, 2016.
- [31] S. Acharya *et al.*, “ D_s^+/D^0 and Λ_c^+/D^0 relative abundances at mid-rapidity as a function of multiplicity in proton-proton collisions at $\sqrt{s} = 13$ TeV,” *J. High Energy Phys.*, vol. 2021, no. 3, pp. 1–24, 2021.

Appendix

Invariant-mass spectra after rectangular pre-selection

As discussed in Sec. 5.2, the rectangular selection is used only as a loose pre-selection and not as the main selection approach of the analysis. The invariant-mass distributions obtained after applying only the rectangular pre-selection are shown in Fig. 6. These spectra illustrate that, due to the large combinatorial background and the complex three-body decay topology of the Λ_c^+ baryon, an additional machine-learning based selection is needed to improve the signal extraction.

Technical information about the trained BDT models

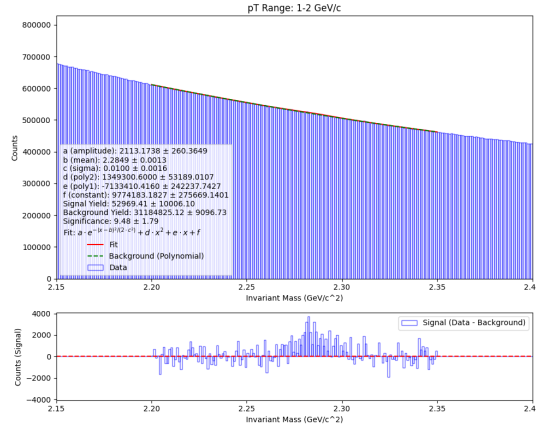
The BDT models used for the machine-learning based candidate selection were trained separately in each p_T interval, as described in Sec. 6. After training, the models were stored in the ALICE CCDB so that they could be used later in the O² analysis workflow. The corresponding CCDB paths are listed in Tab. 1. This information is included to document the exact model locations and to make the analysis setup reproducible.

p_T interval	CCDB path
1 – 2 GeV/ c	Users/k/kzertova/BDT_training_Lc/22_4_26/pt12
2 – 3 GeV/ c	Users/k/kzertova/BDT_training_Lc/22_4_26/pt23
3 – 4 GeV/ c	Users/k/kzertova/BDT_training_Lc/22_4_26/pt34
4 – 5 GeV/ c	Users/k/kzertova/BDT_training_Lc/22_4_26/pt45
5 – 6 GeV/ c	Users/k/kzertova/BDT_training_Lc/22_4_26/pt56
6 – 7 GeV/ c	Users/k/kzertova/BDT_training_Lc/22_4_26/pt67
7 – 8 GeV/ c	Users/k/kzertova/BDT_training_Lc/22_4_26/pt78
8 – 10 GeV/ c	Users/k/kzertova/BDT_training_Lc/22_4_26/pt810
10 – 12 GeV/ c	Users/k/kzertova/BDT_training_Lc/22_4_26/pt1012
12 – 24 GeV/ c	Users/k/kzertova/BDT_training_Lc/22_4_26/pt1224

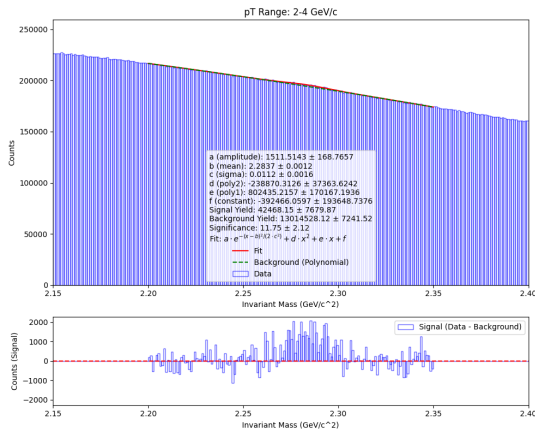
Table 1: CCDB paths of the trained BDT models used for the Λ_c^+ candidate selection in the individual p_T intervals.

Additional BDT training and performance distributions

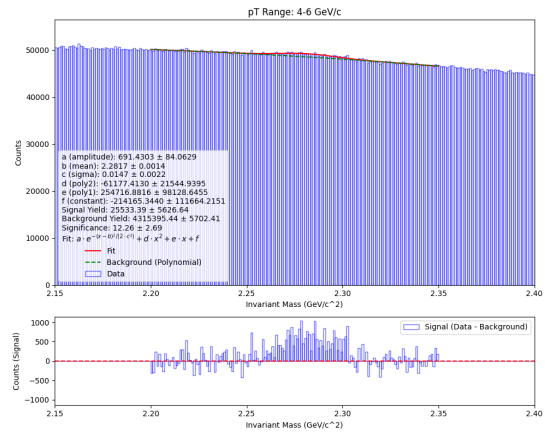
This appendix collects additional plots related to the training and validation of the BDT models. In the main text, only representative examples are shown in order to keep the discussion readable. The figures below show the same quantities for all p_T intervals used in the analysis. They complement the discussion of the training procedure, variable importance, performance evaluation, and working-point optimisation in Sec. 5.3.



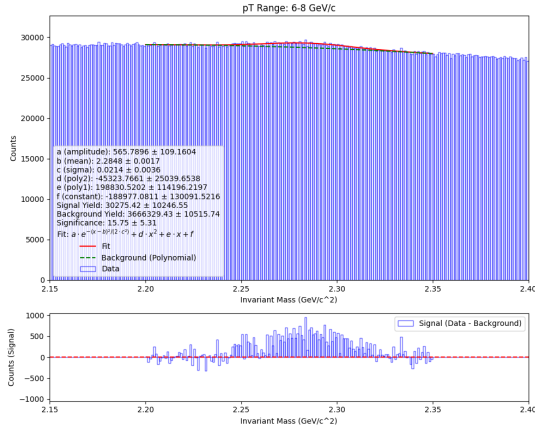
(a) 1 – 2 GeV/c



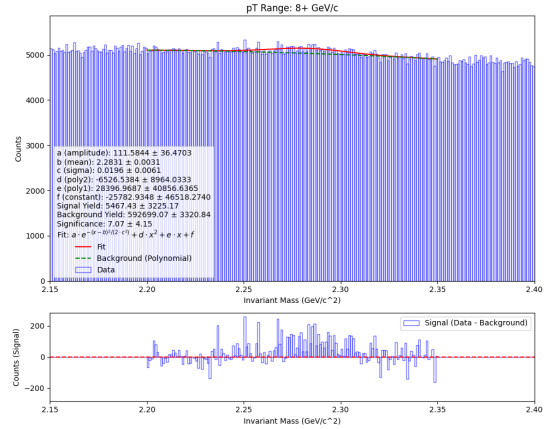
(b) 2 – 4 GeV/c



(c) 4 – 6 GeV/c



(d) 6 – 8 GeV/c



(e) $p_T > 8$ GeV/c

Figure 6: Λ_c invariant-mass spectra in pp collisions at $\sqrt{s} = 13.6$ TeV after applying only the rectangular pre-selection.

Model-output score distributions

The model-output score distributions provide a direct check of the classifier response for each class. They are used to verify whether candidates belonging to a given true class tend to obtain a high score for the corresponding classifier output, and whether the training and testing samples show compatible behaviour. The three sets of score distributions are shown separately for the background, prompt, and non-prompt classes in Figs. 7–9.

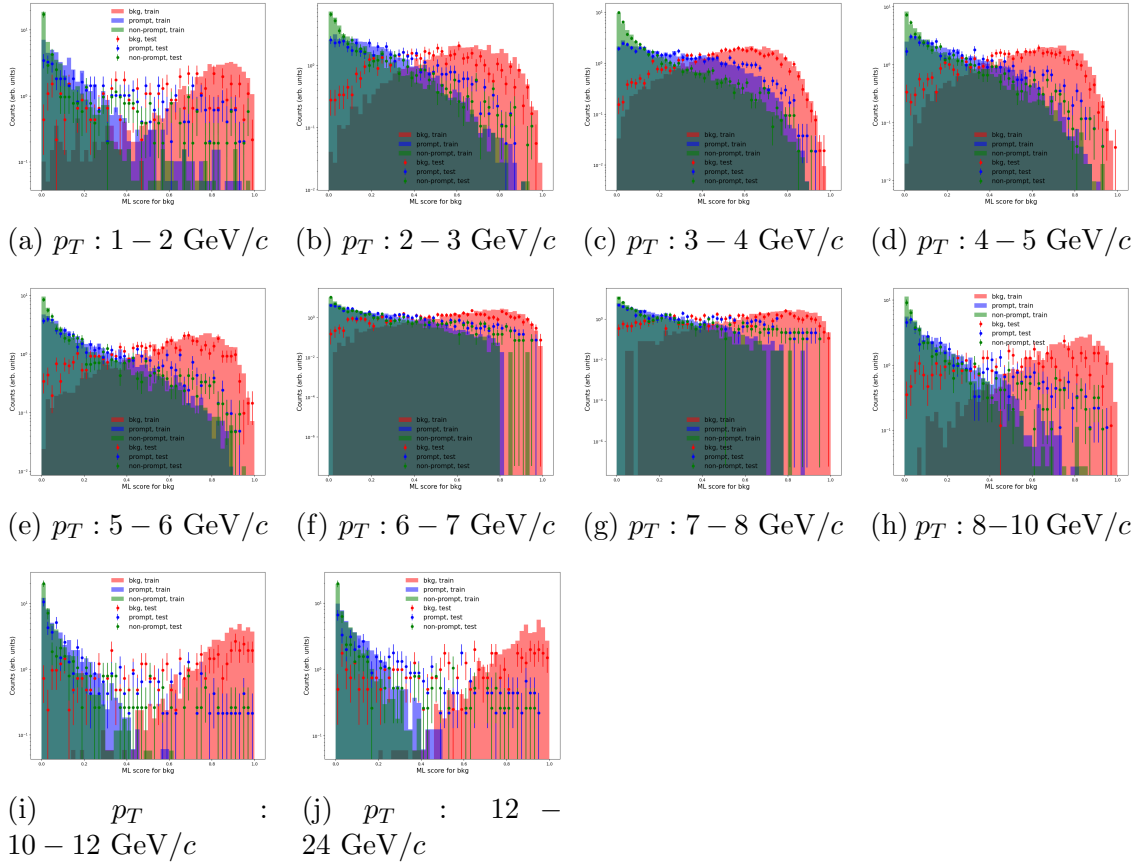


Figure 7: Distributions of the machine-learning output score for the combinatorial-background class in the individual p_T intervals.

Precision–recall and efficiency curves

The precision–recall curves and efficiency plots are used during the working-point optimisation discussed in Sec. 5.3. The probability shown on the horizontal axis corresponds to the scanned cut value. These plots therefore show how the classification purity, prompt-signal retention, and model significance change when the applied probability cut is varied.

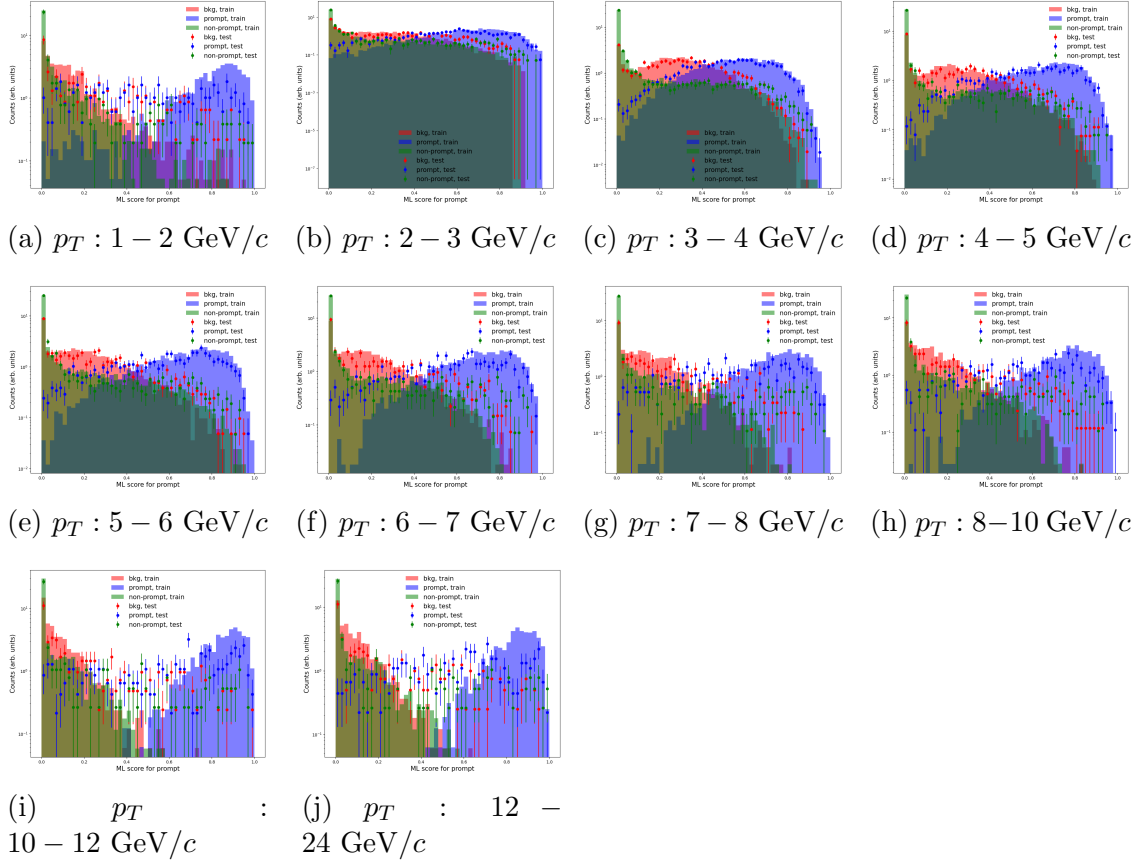


Figure 8: Distributions of the machine-learning output score for the prompt Λ_c^+ class in the individual p_T intervals.

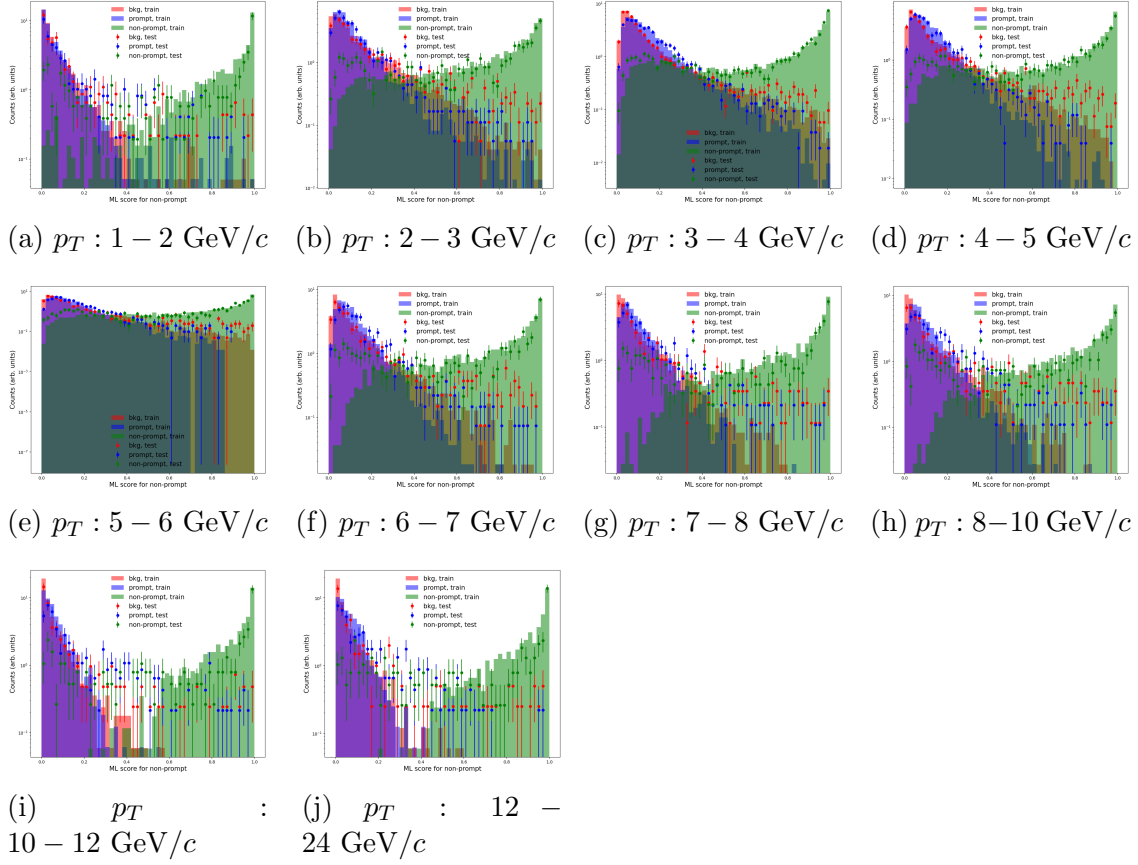


Figure 9: Distributions of the machine-learning output score for the non-prompt Λ_c^+ class in the individual p_T intervals.

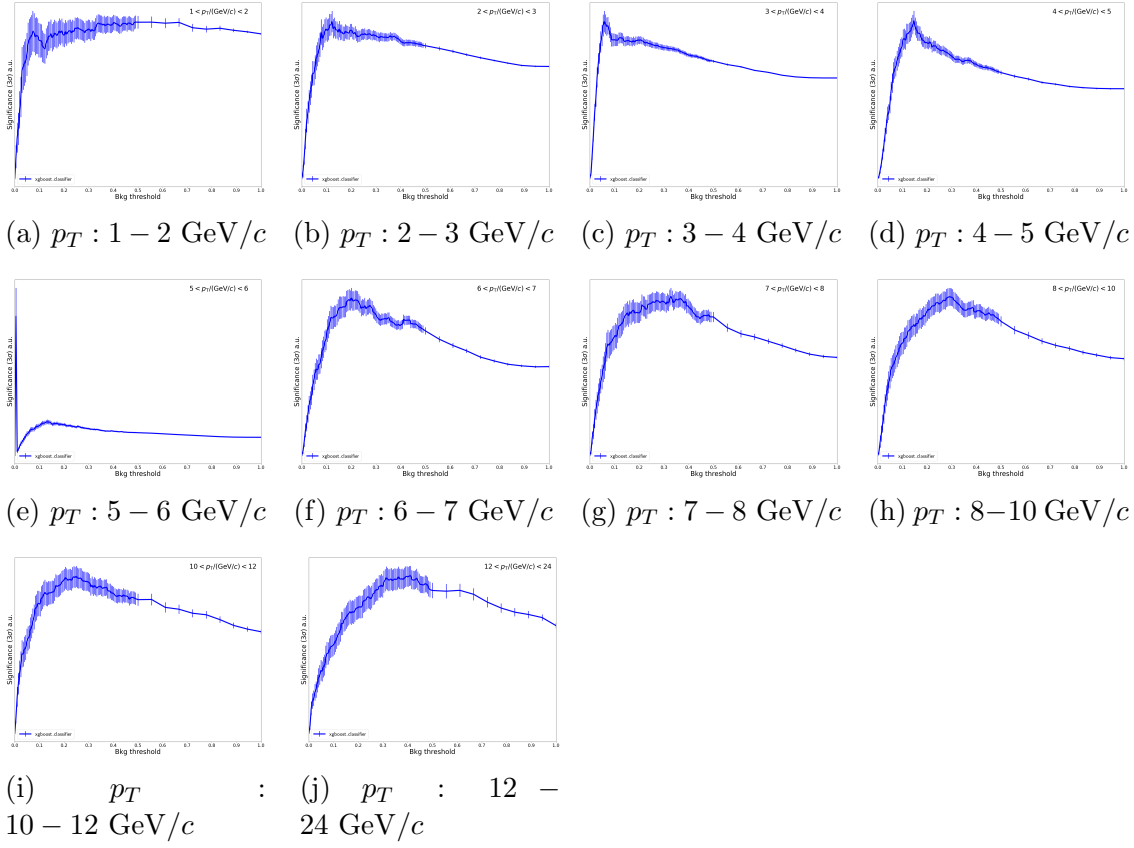


Figure 10: Model significance as a function of the background-score threshold for the trained BDT classifier in the individual p_T intervals.

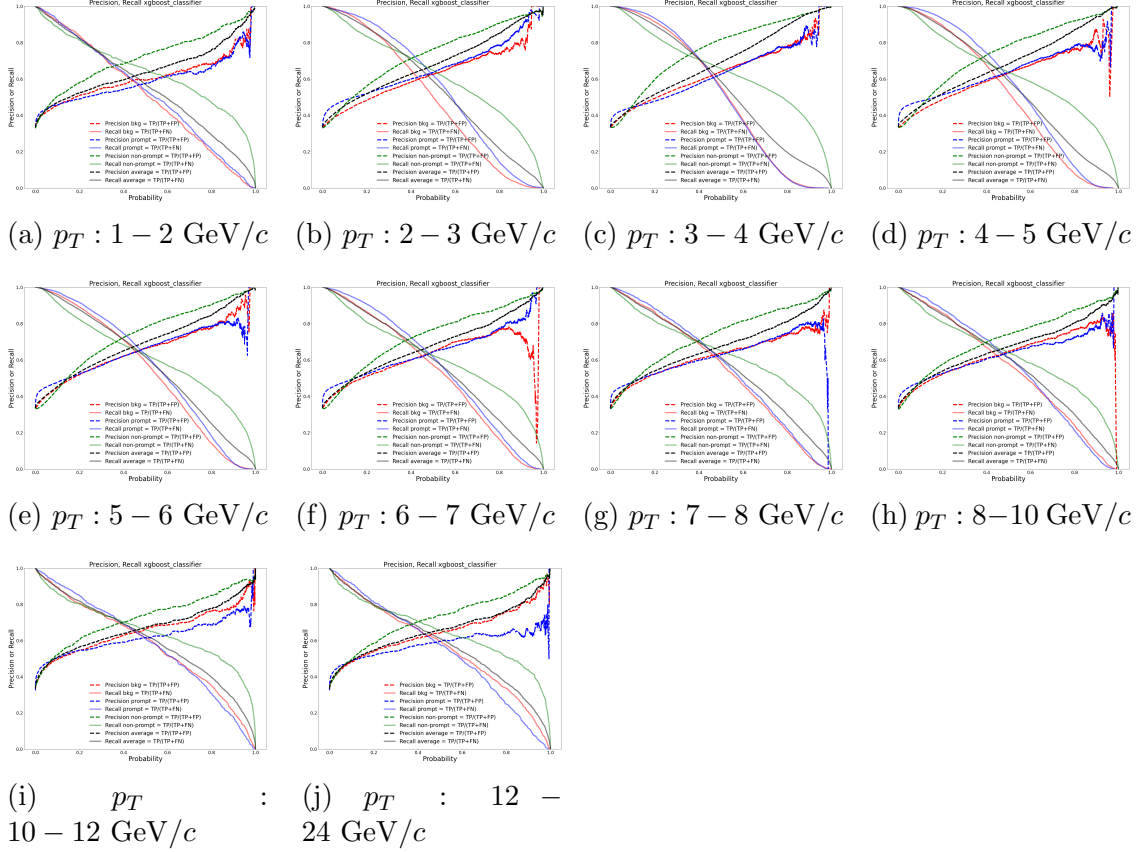


Figure 11: Precision and recall curves of the trained BDT classifier in the individual p_T intervals. The curves are shown separately for the background, prompt Λ_c^+ , and non-prompt Λ_c^+ classes as functions of the classifier output probability.