Department of Physics and Astronomy
University of Heidelberg

Bachelor Thesis in Physics
submitted by

Christian Kleiber

born in Heidelberg (Germany)

March 2023

Feasibility study of the non-prompt $\Lambda_c^+ \rightarrow pK^-\pi^+$ analysis in p–Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV with ALICE

This Bachelor Thesis has been carried out by Christian Kleiber at the
Physikalisches Institut of the Heidelberg University
under the supervision of
Prof. Dr. Silvia Masciocchi

Abstract

In this thesis, the feasibility of an analysis of the non-prompt Λ_c^+ baryon in the proton-lead collision system at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV at midrapidity is studied for the decay channel $\Lambda_c^+ \rightarrow \mathrm{pK}^-\pi^+$. This is done by measuring the yield of the non-prompt Λ_c^+ in data collected during Run 2 in 2016 and determining the significance of the signal peaks. The reconstructed candidates have a transverse momentum in the range of $2 < p_{\mathrm{T}} < 12$ GeV/c. The signal extraction is optimised with Boosted Decision Trees, implemented with the XGBoost algorithm, where the selections on the candidates are structurally determined.

Conducting a full non-prompt analysis allows an indirect investigation of beauty hadrons and the beauty hadronisation. Furthermore, Cold Nuclear Matter (CNM) effects can be investigated, helping with the disentanglement of CNM effects and the final state effects in the Quark–Gluon Plasma (QGP) in more complex collision systems like Pb–Pb.

The investigation found significances ranging from 4.1 to 6.8 for the non-prompt signal peaks at non-prompt fractions between 73% and 80%. This indicates that a full analysis of the non-prompt $\Lambda_c^+ \rightarrow \mathrm{pK}^-\pi^+$ channel is feasible.

Zusammenfassung

In dieser Arbeit wird die Machbarkeit einer Analyse des non-prompt Λ_c^+ Baryons im Proton-Blei Kollisionssystem bei $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV und Midrapidität für den Zerfallskanal $\Lambda_c^+ \to \mathrm{pK}^-\pi^+$ untersucht. Hierfür wird der non-prompt Λ_c^+ Ertrag in Daten aus Run 2 aus 2016 gemessen und die Signifikanzen der Signalpeaks bestimmt. Die rekonstruierten Kandidaten besitzen einen transversalen Impuls im Bereich $2 < p_{\mathrm{T}} < 12$ GeV/c. Das Extrahieren des Signals wird mit Hilfe von Boosted Decision Trees durchgeführt, welche durch den XGBoost Algorithmus implementiert werden. Die Auswahl der Signalkandidaten wird anschließend strukturiert ermittelt.

Das Durchführen einer erweiterten non-prompt Analyse erlaubt die indirekte Untersuchung der Beauty-Hadronen und der Beauty-Hadronisierung. Weiterhin können sogenannte Cold Nuclear Matter (CNM) Effekte untersucht werden, was schließlich zu einer Entwirrung von CNM Effekten und der finalen Effekte im Quark-Gluon-Plasma (QGP) in komplexeren Kollisionssystemen wie Pb–Pb beitragen wird.

Diese Arbeit fand die ermittelten Signifikanzen im Bereich zwischen 4.1 und 6.8, wobei die non-prompt Anteile zwischen 73% und 80% liegen. Dies weist darauf hin, dass eine erweiterte non-prompt Analyse für den Zerfallskanal $\Lambda_c^+ \to \mathrm{pK}^-\pi^+$ realisierbar ist.

Contents

1 Introduction

> "The first principles of the universe are atoms and empty space; everything else is merely thought to exist."
>
> - Democritus (c. 460 BC - c. 370 BC),
> trans. by Robert Drew Hicks 1925 [1]

We have come a long way since the first attempts of Greek Atomism, where the atom derived its name from the Greek words *atomos*/*atomon*, 'indivisible' [2]. Now, over two millennia later, we know that the atom could not stay true to this origin. The atom is not indivisible, but rather a composite of objects with constituents that in turn have their own substructure. With particle physics as an established field of physics in general, we know about protons, electrons, quarks, neutrinos, and so on. A lot of today's knowledge comes from collider experiments, where particles can be exposed to very high energy densities and temperatures, in some specific cases even simulating the conditions right after the Big Bang.

The collider used to perform the measurements in this thesis is the Large Hadron Collider (LHC), located at CERN in Geneva, Switzerland. At the LHC, heavy-ion collisions can generate conditions of extreme temperatures, where the field theory of the strong interaction, Quantum Chromodynamics (QCD), predicts the creation of a colour-deconfined state, the Quark–Gluon Plasma (QGP). In this state, quarks can be considered as free objects, while in ordinary matter they are only able to exist in colour-neutral confined states. After a short time of the order of $\sim 10^{-23}$ seconds, the QGP cools down and hadrons start to form [3]. The produced hadrons (or their decay products) are then observed with the ALICE detector.

1.1 The Goal of this Thesis

The hadronisation mechanisms of heavy quarks are still open questions, especially for the beauty quark. Having comparably large masses ($m_{\mathrm{c}} \approx 1.27$ GeV/c^2 [4] and $m_{\mathrm{b}} \approx 4.18$ GeV/c^2 [4]), heavy quarks are predominantly produced in hard scattering processes, e.g. the initial collision, where the momentum transfer Q^2 exceeds the $4m_{\mathrm{c/b}}^2$ production threshold [5, 6]. They then undergo the complete evolution of the collision, until they hadronise into the particles, whose decay products will then pass through the detector.

According to the QCD factorisation approach, the heavy-flavour hadron production cross sections can be calculated as convolutions of the parton distribution functions, the parton hard-scattering cross sections and the fragmentation functions [5]. This can be written compactly as [7]

$$\frac{\mathrm{d}\sigma}{\mathrm{d}p_{\mathrm{T}}}^{pPb\rightarrow H_q} = f_i(x_1, Q^2) f_j(x_2, Q^2) \cdot \frac{\mathrm{d}\sigma^q}{\mathrm{d}p_{\mathrm{T}}} \cdot D_{q\rightarrow H_q}. \tag{1.1}$$

Here, the parton distribution functions f_i and f_j show the probability of finding a parton of a certain type in a hadron H to carry a momentum fraction x. They cannot be calculated theoretically and need to be measured, usually in deep inelastic scattering processes ($\mathrm{e^-p} \rightarrow \mathrm{e^-X}$) [8]. Here, x is the fraction of hadron momentum carried by the parton. This entity is also called Bjorken-x and is for deep inelastic scattering ($\mathrm{e^-p} \rightarrow \mathrm{e^-X}$) defined as $x = Q^2/2M\nu$, where Q^2 is the squared 4-momentum transferred in the scattering process, M is the mass of the nucleon and ν is the energy loss of the scattering electrons [8]. The parton hard-scattering cross sections $\frac{\mathrm{d}\sigma^q}{\mathrm{d}p_{\mathrm{T}}}$ describe the probability of the creation of the parton q and are calculable with the methods of perturbative Quantum Chromodynamics (pQCD) [6, 7]. Here, they are the production cross sections of $\mathrm{c\bar{c}}$ (or $\mathrm{b\bar{b}}$) pairs. Lastly, the fragmentation functions $D_{q\rightarrow H_q}$, tuned on electron-positron and electron-proton collisions, express the probability of a quark q to hadronise into a specific hadron H_q [5, 9]. They were assumed to be universal for different collision system [10].

An observable for the hadronisation is the baryon-to-meson yield ratio, like the $\Lambda_{\mathrm{c}}^+/\mathrm{D}^0$ yield ratio [10]. According to the aforementioned QCD factorisation approach, the parton distribution functions and the parton hard-scattering cross sections in Eq. 1.1 are the same for all charm (and respectively beauty) hadrons, thus cancelling entirely in ratios, leaving only the fragmentation functions to govern the production [10]. In previous pp collision measurements, it was found that other meson-to-meson yield ratios (D^+ and $\mathrm{D}_{\mathrm{s}}^+$ to D^0) are within their uncertainties independent of their transverse momentum and consistent with the model predictions using the fragmentation functions from $\mathrm{e^+e^-}$ and $\mathrm{e^-p}$ collisions [10]. However, the charmed baryon-to-meson ratios for Λ_{c}^+, $\Xi_{\mathrm{c}}^{0,+}$, Ω_{c}^0 and $\Sigma_{\mathrm{c}}^{0,++}$ deviate significantly from their respective $\mathrm{e^+e^-}$ and $\mathrm{e^-p}$ collision measurements [10]. Therefore, the assumption that the hadronisation processes do not depend on the collision system (i.e. are universal) is challenged.

While looking at charm hadrons is possible in Run 2 data, directly doing the same for beauty is not practicable, due to a lack of statistical precision. However, looking into

the beauty sector indirectly via the decay of a beauty hadron into a charm hadron was realised. Previously, this was done on the decay channel of $\Lambda_c^+ \rightarrow pK_s^0$ with the subsequent decay of $K_s^0 \rightarrow \pi^+\pi^-$. Fig. 1.1 shows the measurements of the cross section via this decay channel for different p_T. Considering the scaling of the y-axis, the statistical uncertainties were found to be improvable. This is where the scope of this thesis starts.

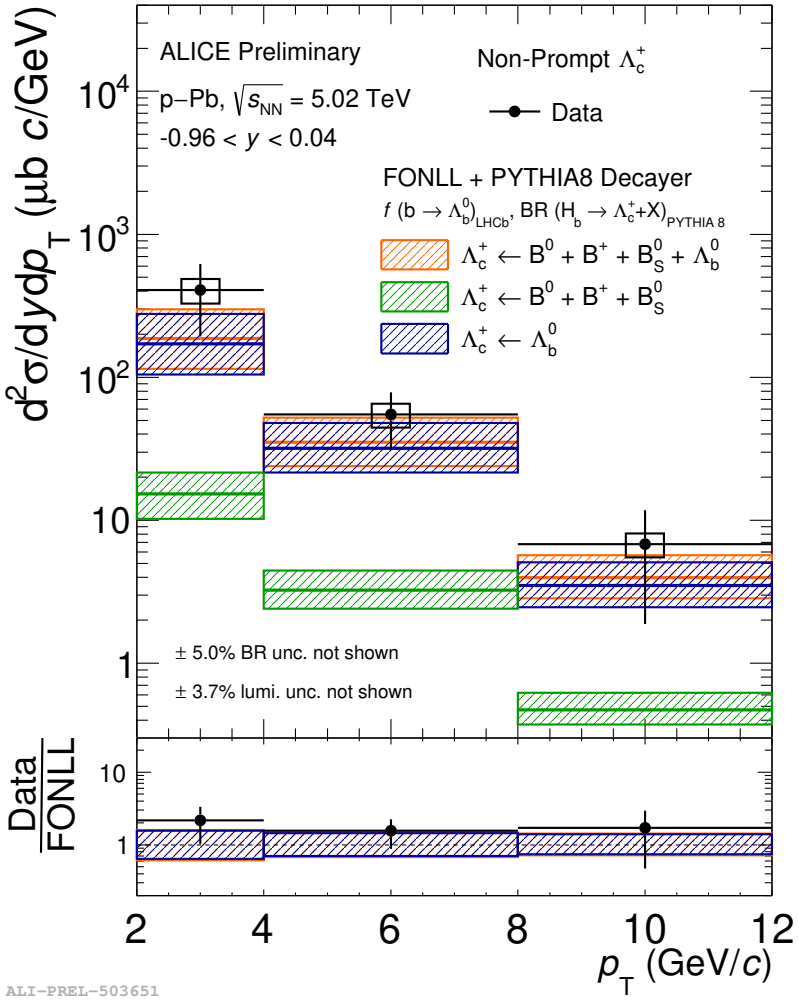

Figure 1.1: Measurements and predictions of the NP Λ_c^+ cross section in p–Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV measured via the $\Lambda_c^+ \rightarrow pK_s^0$ decay channel. The different colours show the origin of the NP particles in the predictions. Image source: [11].

This thesis focuses on the Λ_c^+ baryon in the decay channel $\Lambda_c^+ \rightarrow pK^-\pi^+$. A Feynman diagram of this channel is shown in Fig. 1.2. The baryon consists of three valence quarks with the flavours up, down and charm and has a mass of $m_{\Lambda_c^+} = (2286.46 \pm 0.14)$

MeV/c^2 [4]. With a lifetime of $\tau_{\Lambda_c^+} \approx 2.015 \times 10^{-13}$ s, the mean proper decay length is $c\tau_{\Lambda_c^+} \approx 60.0\,\mu$m [4]. Although this thesis also covers the anti-particle $\overline{\Lambda}_c^-$, for readability, only the Λ_c^+ is highlighted.

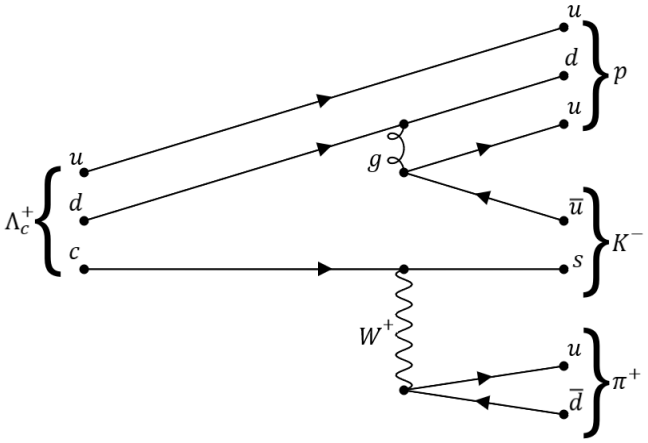

Figure 1.2: Feynman diagram of the $\Lambda_c^+ \to pK^-\pi^+$ decay channel.

Generally, there are two classes of Λ_c^+: prompt and Non-Prompt (NP, also called Feed-Down, FD). Prompt means that the particle has been created at the Primary Vertex (PV), while the NP particles are decay products of other particles and are therefore not originating from the PV, but rather a Secondary Vertex (SV). Therefore, the NP charm particles allow the indirect access to the prompt beauty hadrons, as mentioned earlier. A visualisation is shown in Fig. 1.3. There are several possible particles which can decay into a NP Λ_c^+, as Fig. 1.1 shows. In the figure, the predictions of the cross sections of the NP Λ_c^+ are shown for the Λ_b^0 baryon and three B mesons as origin particles. It shows that the majority of NP Λ_c^+ originate from a Λ_b^0 decay. The mean proper decay length of the Λ_b^0 can be calculated to be $c\tau_{\Lambda_b^0} = 441.0\,\mu$m [4]. Considering the decay lengths of these two particles, it can be assumed that all of them will decay while still being in the vacuum tube, before even entering any kind of detector [12]. Looking further at the decay products of the Λ_c^+, the proton is stable, while the kaon and pion have mean proper decay lengths of $\sim$3.7 m and $\sim$7.8 m [4]. This means, that these are the particles which will likely be passing through most of the relevant detectors for this analysis.

The advantage of the $\Lambda_c^+ \to pK^-\pi^+$ channel over the $\Lambda_c^+ \to pK_s^0$ channel lies in their branching ratios. While the branching ratio for the $\Lambda_c^+ \to pK_s^0$ channel is $(1.59 \pm 0.08)\%$ [4], it has to be multiplied by the subsequent $K_s^0 \to \pi^+\pi^-$ branching ratio, which is $(69.20 \pm 0.05)\%$, resulting in a total ratio of about 1.1%. The branching ratio for the $\Lambda_c^+ \to pK^-\pi^+$ channel is found to be $(6.28 \pm 0.32)\%$ [4]. This means, that there is an

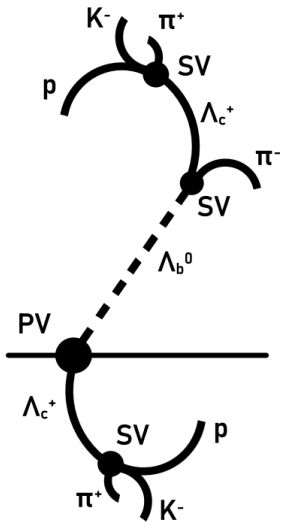

Figure 1.3: Visualisation of an example for a NP) Λ_c^+ particle track from a Λ_b^0 (top) and a prompt Λ_c^+ particle track (bottom). PV and SV are abbreviations for Primary Vertex and Secondary Vertex, respectively. The PV is the point of initial collision where the proton and lead beams collide, while there are several SVs. These are the vertices where other particles which originate from decays are created. Naturally, the top and bottom part do not originate from the same PV, but rather serve only illustrative purposes.

expected factor of at least 5 between the expected yields for these two channels. Looking at the $\Lambda_c^+ \rightarrow pK^-\pi^+$ channel may therefore reduce the statistical uncertainties in the cross sections. However, the challenge here lies at the large combinatorial background of this channel, due to the decay into three decay products. Therefore, it is unclear whether the combination of a larger branching ratio and also a larger background component make this analysis feasible.

This thesis has the goal of measuring the yield of the non-prompt $\Lambda_c^+ \rightarrow pK^-\pi^+$ and assess the feasibility of a full analysis by determining the significances of the yields. If the analysis proves to be feasible, then it can be continued and used to add statistics to the analysis of the $\Lambda_c^+ \rightarrow pK_s^0$ channel. The prompt $\Lambda_c^+ \rightarrow pK^-\pi^+$ yield will also be analysed, since this can eventually be used as a validation when comparing it to the published values [13].

1.2 The Proton-Lead Collision System

Generally, the main goal of the heavy-ion collisions at ALICE is the examination of the QGP. However, besides the QGP, there are other initial state effects which are entan-

gled with the final state effects. This creates the need to differentiate between effects coming from the actual QGP colour-deconfinement and others, such as the shadowing and Cronin effects, labelled as Cold Nuclear Matter (CNM) effects, originating from the presence of a nucleus in the collision [14]. While heavy-flavour production in pp collisions provide a general reference for studies of heavy-ion collisions and a powerful test of pQCD, p–Pb collisions are intermediate states where traditionally no QGP is expected (although some theoretical models also predict the formation of a mini-QGP for smaller collision systems [15, 16]), but CNM effects have to be accounted for [5]. Therefore, the investigation of the intermediate p–Pb collision system may help to disentangle the QGP and CNM effects in more complex collision systems, permitting a deeper understanding of the QGP. Hence, this work has been realised with measurements of p–Pb collisions.

For deviations between the pp collision baseline and larger collision systems, e.g. p–Pb, the ratios of observables are typically considered, such as the nuclear modification factor

$$R_{\mathrm{pPb}}(y, p_{\mathrm{T}}) = \frac{\mathrm{d}^2\sigma_{\mathrm{pPb}}/\mathrm{d}y\mathrm{d}p_{\mathrm{T}}}{A \cdot \mathrm{d}^2\sigma_{\mathrm{pp}}/\mathrm{d}y\mathrm{d}p_{\mathrm{T}}}, \tag{1.2}$$

where besides the mass number A (here for lead $A = 208$), the differential cross sections for the two collision systems are needed [6]. If there were no modifications in p–Pb with respect to a simple superposition of pp collisions, this ratio would be unity [6].

The aforementioned CNM effects consist mainly of two parts, namely the shadowing effect and the Cronin effect [6]. Starting with the shadowing effect, the ratio $R_{\mathrm{i}}^{\mathrm{A}}$ is defined as [17]

$$R_{\mathrm{i}}^{\mathrm{A}} = \frac{f_{\mathrm{i/A}}(x, Q^2)}{f_{\mathrm{i}}(x, Q^2)}. \tag{1.3}$$

This ratio compares the parton distribution functions f_{i} for a nucleon bound in a nucleus (numerator) to the distribution of a free nucleon (denominator), where index i represents the parton species, i.e. valence quark, sea quark or gluon [17]. At LHC energies, i.e. the low x region, the shadowing effect occurs, where the parton densities in the bound nucleons decrease in comparison with the free nucleons, therefore leading to a nuclear modification factor smaller than unity [5]. From this effect alone, it would be expected that the charm production for p–Pb collisions is suppressed in comparison to pp collisions [6].

The Cronin effect, also Cronin enhancement, describes the modified production of heavy-flavour particles [18]. Through multiple elastic collisions of partons within their initial target particle before the actual hard scattering process, the initial transverse momentum of the partons is increased [18]. This results in a shift of the transverse mo-

mentum spectrum towards higher p_{T} values and therefore leads to an increased nuclear modification factor in Eq. 1.2 [6].

2 The ALICE Detector

ALICE (A Large Ion Collider Experiment) is one of the four main experiments located along the 26.7 km tunnel of the Large Hadron Collider (LHC) at CERN in Geneva, Switzerland [19]. With the goal of investigating the QGP, the ALICE detector was optimised for high-energy heavy-ion collisions [20]. Bunches of particles like protons and lead nuclei are accelerated to energies up to several TeV and brought to collisions at the interaction point in the center of the ALICE detector. After the collision, particles, produced in very high multiplicities, will travel through the detector and deposit energy. This energy is then measured and processed by the different detector components, which allows further analysis.

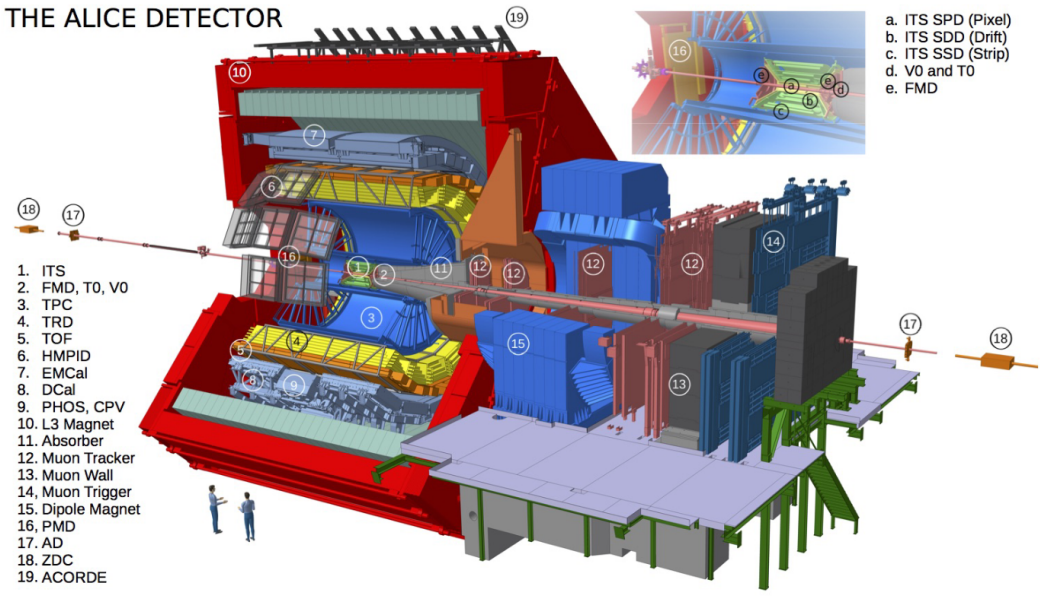

Figure 2.1: Run 2 configuration of the ALICE Detector. Image source: [21].

Fig. 2.1 shows a schematic setup of the $16 \times 16 \times 26$ m^3 detector setup during Run 2 from 2015 to 2018 [20, 22]. The detector consists of several sub-detectors, some of which are located within the L3 magnet, inherited from the L3 experiment at the previous LEP (Large Electron-Positron Collider) at CERN. The magnet is coloured red in Fig. 2.1 and provides a magnetic field with a strength of $B = 0.5$ T in the direction parallel to the beam line [20]. The sub-detectors located within the magnet operate at midrapidity ($|\eta| < 0.9$) are called central-barrel detectors [20]. They contain most notably the ITS, TPC and TOF detector. The ITS and TPC are the detector's main systems for the tracking of charged particles, while the TOF detector provides PID information for charged par-

ticles at intermediate momenta [20]. Further information on these systems is provided in the next sections of this chapter.

ALICE uses a right-handed coordinate system, which has its origin at the LHC Interaction Point 2 (IP**2**) (centre of the central barrel detectors) [20]. The x-axis is defined as the horizontal pointing towards the center of the LHC ring, the y-axis points vertically upwards and the z-axis is parallel to the beam line [20]. In addition to this Cartesian system, spherical coordinates are also used. Generally, the x-y-plane is used to define the transverse direction, e.g. for transverse momentum of the particles and the Lorentz-invariant pseudorapidity η is used instead of the spherical polar angle θ [22], where η is defined as [23]

$$\eta = -\ln \tan \frac{\theta}{2}. \tag{2.1}$$

2.1 Inner Tracking System

The Inner Tracking System (ITS) is a cylindrical detector in the central barrel of the ALICE detector, oriented along the z-axis. It is the detector closest to the beam line and faces track densities of up to 50 tracks/cm^2 in heavy-ion collisions [24]. The ITS is one of the main tracking detectors of ALICE and provides crucial information for the preliminary determination of the Primary Vertex [20]. It is also able to provide PID information through measurements of the specific energy loss $\mathrm{d}E/\mathrm{d}x$ of especially low momenta (< 100 MeV/c) particles [24]. Since this thesis only analyses particles with momenta of at least 300 MeV/c, a range where the PID capabilities of the TPC and TOF detectors are sufficient, no ITS PID information is used. Thus, the description of PID in the ITS is omitted here.

In total, the detector consists of six layers, with the two innermost (starting at the smallest possible radius $r = 4$ cm, closest to the beam line) consisting of Silicon Pixel Detectors (SPD), the two middle layers of Silicon Drift Detectors (SDD) and the two outermost of Silicon Strip Detectors (SSD) [12]. This layout is visualised in Fig. 2.2, also showing the total diameter of 87.2 cm of the outermost layer.

To determine a preliminary PV, only the two innermost layers (SPD) are used [20]. The SPD provide a high spacial resolution of 12 µm in transverse direction and 100 µm in z-direction [25]. The signals in the two layers can be combined to pairs of clusters, so-called tracklets, which can be prolonged into the vacuum tube [20]. Looking at all the possible prolonged tracklets, the point in space where a maximum number of them meet

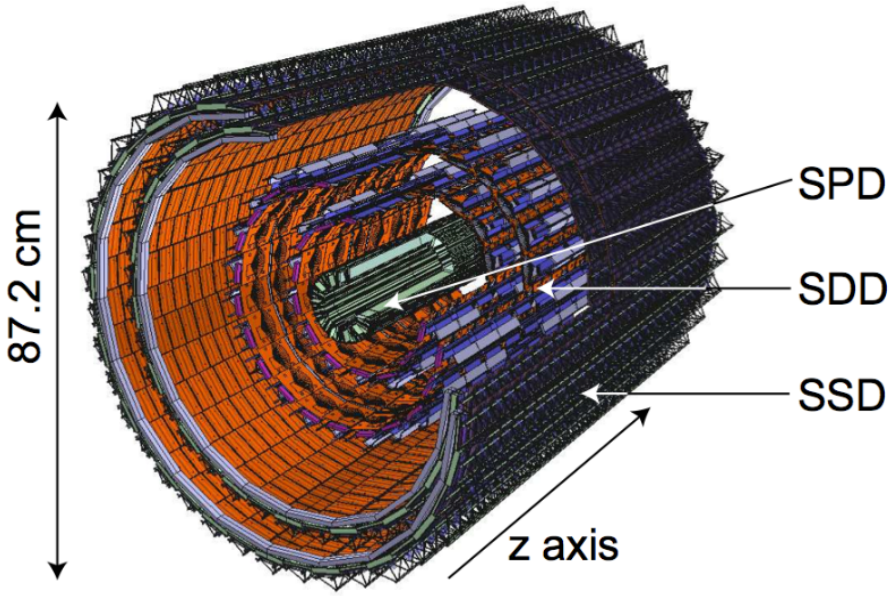

Figure 2.2: Layout of the Run 2 ITS. Image source: [12].

is identified as a preliminary PV [20]. For further track reconstruction the signals in the ITS are combined with the TPC measurements in an inward-outward-inward scheme, starting at the outer TPC [20]. With this reconstruction, the primary vertex can be determined with a resolution of $60-75$ µm for a 1 GeV/c charged particle in a collision of protons and/or lead nuclei [20]. Also, due to the high resolutions, secondary vertices can be identified by looking for tracks, whose distance of closest approach to the PV is above a certain threshold [20]. This is important for the investigation of short-lived heavy-flavour hadrons, which decay before even entering any detector, like already mentioned in Chapter 1 for the Λ_c^+ and Λ_b^0.

2.2 Time Projection Chamber

The Time Projection Chamber (TPC) surrounds the ITS and is build to cope with the high multiplicity environment of ALICE, providing information for PID, track reconstruction and final vertex determination [20]. With the given magnetic field of the L3 magnet, it covers a wide range of transverse momentum from 100 MeV/c up to 100 GeV/c with a good momentum resolution [25].

Like visualised in a 3D sketch in Fig. 2.3, the TPC is a hollow cylinder with a length of 500 cm and an inner and outer radius of 85 cm and 250 cm, respectively [23]. The inside is divided into two parts by a central high-voltage electrode and contains a total

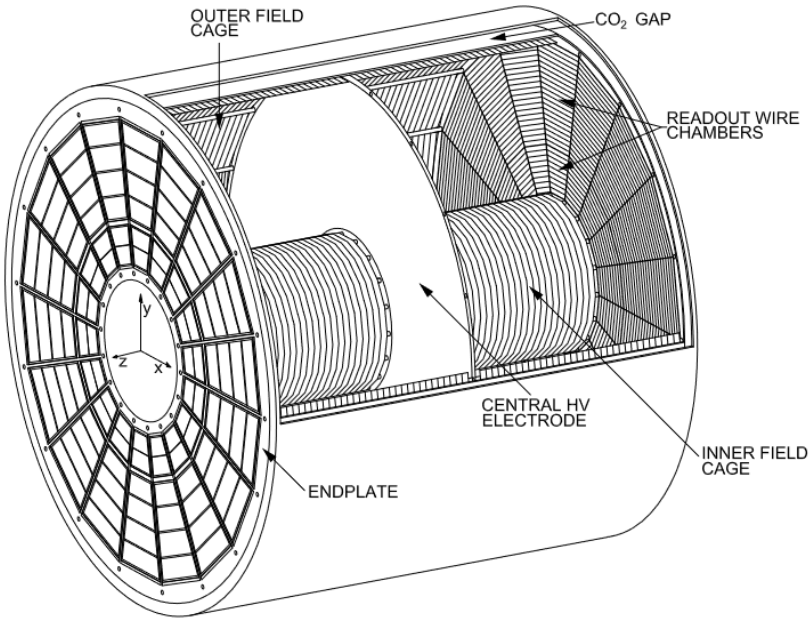

Figure 2.3: 3D sketch of the TPC. Image source: [23].

volume of about 90 m^3 [23]. The volume contains a gas mixture of Ne/CO_2/N_2 (briefly during Run 2: Ar/CO_2/N_2 [22]) at atmospheric pressure, serving with a large radiation length compared to the dimensions of the TPC [23]. The readout electronics are located at the two endplates and can be radially separated into two readout chambers at a radius of about 133 cm [25]. The reason for this is the change in track density along the radius, which allows optimisation of the geometry of the readout pads in the chambers [25]. Furthermore, both parts are then further divided into 18 trapezoidal sections in azimuthal direction, covering $20°$ each [23].

Generally, the functionality of the TPC is based upon the fact that charged particles traversing through the chamber ionise the gas mixture, creating free electrons. These electrons then drift, due to the electric field, towards the endplates, where they induce an avalanche in the Multi-Wire Proportional Chamber (MWPC) and thus creating a signal in the readout pads [23].

As already mentioned in the previous section, the track reconstruction starts at the outer TPC. To reconstruct the 3D trajectory, the location of pads which received a signal are used to identify the track position in $r\phi$ direction, while the z-positions are calculated with the drift velocity and drift time of the electrons in the gas measured against a temporal reference, e.g. time of collision [25]. Assigning position and errors to signals

in the detectors makes them so-called clusters [22].

To start the track reconstruction with the aforementioned inward-outward-inward scheme, two induced clusters at the outer TPC are combined with the preliminary PV to build so-called seeds, i.e. first possible track candidates [20]. These seeds allow the identification of further clusters lying in the TPC which may also belong to the same particle. Adding more and more clusters and recalculating the seeds after each added cluster leads eventually to the ITS. After all possible clusters for this stage were identified, the outward stage begins. Here, the track gets refitted to the clusters from the inside out and the track is prolonged into the detectors beyond the TPC [20]. Clusters found in these detectors will not modify the kinematics of the trajectory, but rather allow PID through identification of clusters in e.g. the TOF detector [20]. Lastly, the second inward phase starts again by fitting the clusters from the outer TPC to the clusters in the ITS [20]. This time, the kinematic parameters of the track are determined and the final PV can be estimated similarly to the preliminary PV, but with the extended trajectories [20].

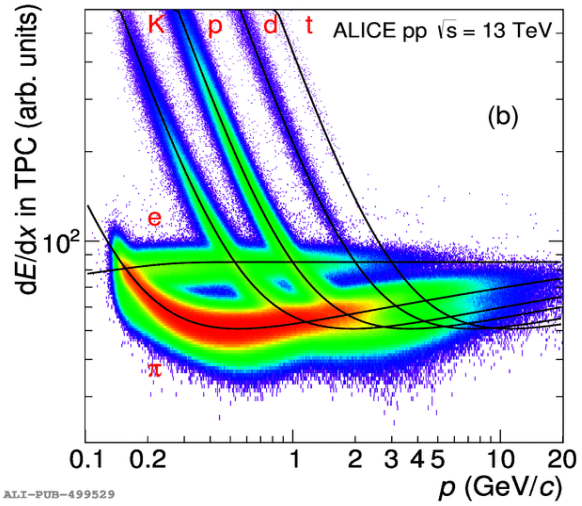

Figure 2.4: Example of specific energy loss over momentum in the TPC for p-p collisions at $\sqrt{s} = 13$ TeV in Run 2. Warmer colours represent higher track counts and the lines indicate the fit solutions for different particles. Image source: [26].

With measurements of charge, momentum and specific energy loss of the particle trajectories, PID can be carried out by using the parameterised Bethe-Bloch formula,

$$f(\beta\gamma) = \frac{P_1}{\beta^{P_4}}\left(P_2 - \beta^{P_4} - \ln\left(P_3 + \frac{1}{(\beta\gamma)^{P_5}}\right)\right), \tag{2.2}$$

with β as the particle velocity, γ as the Lorentz factor and the fit parameters $P_1 - P_5$

[20]. Drawing the $\mathrm{d}E/\mathrm{d}x$ measurements as a function of momentum as shown in Fig. 2.4 (for a different collision system) shows distinct curves originating from different particle species. Overlaid in the plot is the parameterisation from Eq. 2.2. For low momenta ($\lesssim 1$ GeV/c), particles can be separated precisely (see in Fig. 2.4 clearly separated curves), while particles with higher momenta need to be separated using statistical methods (overlapping curves in the figure) and eventually additional information given by other detectors [20].

For this analysis, in order to separate particle species, the variable $n_{\sigma^i_{\mathrm{TPC}}}$ is used. It represents the deviation of the measured signal S_{TPC} (in case of the TPC, this signal is the specific energy loss) from the expected signal $\langle S^i_{\mathrm{TPC}} \rangle$ calculated with Eq. 2.2 for a particle type i in units of the resolution σ^i_{TPC} [7]. Therefore, it can be calculated as [7]

$$n_{\sigma^i_{\mathrm{TPC}}} = \frac{S_{\mathrm{TPC}} - \langle S^i_{\mathrm{TPC}} \rangle}{\sigma^i_{\mathrm{TPC}}}. \tag{2.3}$$

2.3 Time Of Flight Detector

The Time Of Flight (TOF) detector is used to provide additional PID information for charged particles. It focuses on the intermediate momentum range, also because particles need at least 0.3 GeV/c to even reach the detector in the given magnetic field [20, 27].

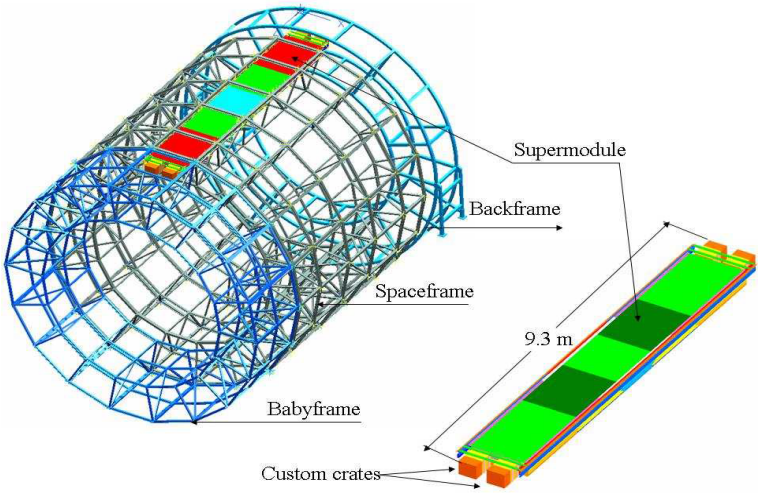

Figure 2.5: Layout of the TOF detector. Image source: [27].

The detector is located at a radial distance of 370 cm to 399 cm to the beam axis and consists of 18 segments attached to a frame like shown in Fig. 2.5 [25]. Each segment,

also called supermodule, consists of 5 smaller modules, which all use Multigap Resistive Plate Chambers (MRPC) technology [27]. The double-stacked MRPCs used in the TOF design are composed of two stacks (therefore double-stacked) of resistive plates, which are fixed at equal distances between them [28]. This allows gas to be filled in between the resistive plates [28].

Particles passing through the system ionise the gas, and because of the high and uniform electric field applied, the induced electrons start avalanching instantly [27]. This process will happen in each of the gas filled gaps between the plates [27]. After the avalanches have been registered at the pick-up electrodes, the total signal is constructed by taking the sum of all avalanches [22].

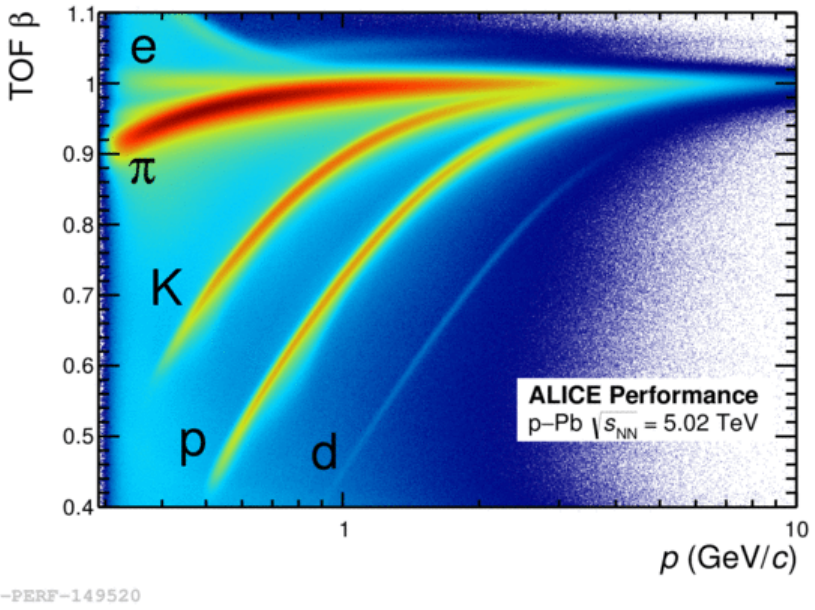

Figure 2.6: Example of a diagram of the measured velocity $\beta = \frac{v}{c}$ over the momentum of a particle for p–Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV. Warmer colours indicate higher track densities. Image source: [29].

The PID performed in the TOF detector uses the measured time of flight t over the track length L to calculate the velocity v [27]. The track length is known through track reconstruction explained in the previous section, while the time of flight is measured as the time of arrival in the TOF detector in reference to the start time provided by the T0 Cherenkov detector (contained in number 2 in Fig. 2.1) [20]. Knowing the momentum p, also through track reconstruction, as well as the track length and time of flight allows

the calculation of the mass m via

$$m^2 = \frac{p^2}{c^2}\left(\frac{c^2t^2}{L^2} - 1\right), \tag{2.4}$$

which provides the PID information [27]. Similar as for the TPC, a $n_{\sigma^i_{\text{TOF}}}$ can be defined by taking the difference of the measurement t and the expected time of flight t^i_{exp} for a certain particle species i and divide it by the time resolution σ_{TOF} of the detector,

$$n_{\sigma^i_{\text{TOF}}} = \frac{t - t^i_{\text{exp}}}{\sigma_{\text{TOF}}}, \tag{2.5}$$

to get the deviation analogous to Eq. 2.3 [27]. Plotting the dimensionless velocity β (i.e. velocity v over speed of light in vacuum c) in a diagram over the particle momentum, distinct bands for each particle species can again be seen. An example of this is shown in Fig. 2.6 for p–Pb collisions at 5.02 TeV.

The capability of differentiating between two particles with different masses depends on the time difference of the two particles and on the time resolution σ_{TOF} (For Run 2 $\sim$56 ps [30]) of the detector. The time difference can be calculated with [27]

$$t_1 - t_2 = \frac{L}{2c}\left(\frac{m_1^2c^2 - m_2^2c^2}{p^2}\right). \tag{2.6}$$

Modifying Eq. 2.5 yields

$$n_\sigma = \frac{t_1 - t_2}{\sigma_{\text{TOF}}}, \tag{2.7}$$

for the difference in units of the detector resolution [27]. Therefore, the separation power decreases with increasing momentum and between species with more similar masses. This can also be observed in Fig. 2.6. For example, with a track length of 3.7 m ([27]) and the time resolution given above, the TOF detector provides a 3σ separation for pions and kaons below $\sim$2.7 GeV/c and for protons and kaons below $\sim$4.8 GeV/s.

2.4 Combined Particle Identification

The PID information from the TPC and the TOF detectors can either be used separately, or they can be combined into a single $n_{\sigma^i_{Comb}}$ variable, which allows the consideration of both detectors at the same time. It is defined as [7]

$$n_{\sigma^i_{Comb}} = \begin{cases} |n_{\sigma^i_{\text{TPC}}}|, & \text{if signal only in TPC} \\ |n_{\sigma^i_{\text{TOF}}}|, & \text{if signal only in TOF} \\ \frac{1}{\sqrt{2}}\sqrt{(n_{\sigma^i_{\text{TPC}}})^2 + (n_{\sigma^i_{\text{TOF}}})^2}, & \text{if signals in both TPC and TOF} \end{cases} . \quad (2.8)$$

This has the advantage of combining two detectors with complementary detection techniques. If one detector cannot provide good enough separation power in a specific case on its own, the other detector might give enough additional information to have a sufficient separation again.

3 Analysis Methods

Many analyses in modern particle physics depend heavily on Machine Learning (ML) as an analysis tool. It allows the convenient handling of billions of events and more precise signal classification than manual selection optimisation. Machine Learning is an umbrella term for several different algorithms, but for this thesis, only Boosted Decision Trees (BDT) are relevant. This chapter focuses on the explanation of BDTs and the relevant concepts around it. The practical implementation of the ML has been achieved via the XGBoost algorithm [31] and the hipe4ml python package [32].

3.1 Boosted Decision Trees

3.1.1 Decision Trees

The basic unit of a BDT is a decision tree. A decision tree is an algorithm based on supervised learning and can be used for classification and regression problems [33].

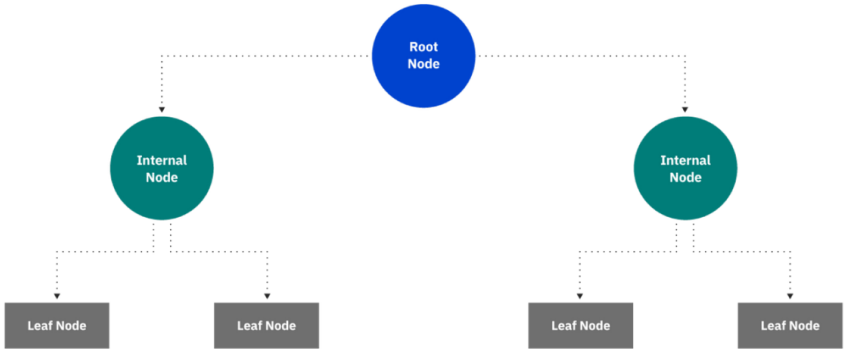

Figure 3.1: The basic structure of a decision tree. Image source: [33].

The basic structure of a simple decision tree can be seen in Fig. 3.1. It consists of a hierarchical tree structure, starting with a root node at the top, followed by branches into internal and eventually leaf nodes [33]. For the purposes of training the model, there is a set of candidates, each with a set of parameters and one unambiguous class to which it should be assigned. All of the candidates start at the root node, where the first decision will be made, e.g. if a particle's momentum is higher than a certain threshold, the candidate will follow the first branch; if it is below, it will follow the second branch. At the end of each branch is a new internal node, where this procedure will be repeated

until it will eventually end at a leaf node. The goal is to achieve a purity as high as possible for each leaf node, i.e. each leaf node collects mostly candidates of only one class. The complexity and depth of a tree can be increased as much as desired, but at a certain complexity it is increasingly difficult to maintain the purity of the leaf nodes, since too little data falling in each sub-tree results in overfitting [33]. When a model is overfitted, it may show perfect results for the data which the tree has been trained on, but will score much worse on similar data that was not used for the training [33].

For the purpose of XGBoost, a slightly modified version of decision trees is used: Classification and Regression Tree (CART) [34]. The modification to the standard decision tree lies in the leaf nodes, where one leaf node does not automatically represent a certain class, but it assigns a certain score to the candidates, which allows more room for interpretation of the tree output [34].

If the result of a single tree is not satisfactory, but increasing complexity decreases performance, an ensemble of trees might be helpful. Several smaller decision trees may be combined in a single superior model, which then provides a better performance than any given single tree. Here, the method used to create the ensemble is called boosting.

3.1.2 Boosting

In general, boosting is the process of combining several weak learners, i.e. decision trees with low complexity, into one strong learner via sequential learning [35]. It is an iterative process, where each tree is built with consideration of the previous weak learners and most importantly their errors [35]. A visualisation of a boosted model is shown in Fig. 3.2. The algorithm used as an implementation of Boosted Decision Trees is called XGBoost (Extreme Gradient Boosting).

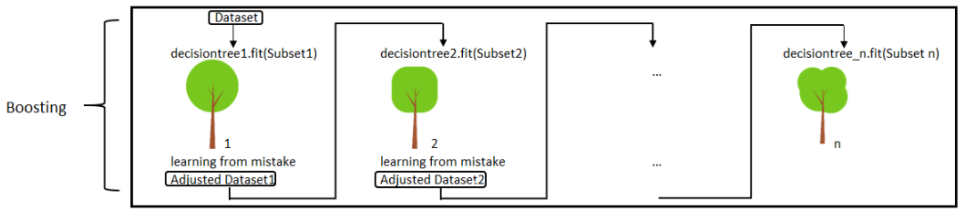

Figure 3.2: Visualization of decision tree boosting with each tree learning from the errors of the previous weak learners. Image source: [36].

3.1.2.1 Extreme Gradient Boosting

The algorithm is used for supervised learning tasks, which give a training data set with multiple features $\mathbf{x_i}$, which will be used to make a prediction $\hat{y}_i$ of the target value y_i [34]. The prediction value can have different interpretations, from representing a probability to ranking the outputs [34], but for heavy-flavour physics, the expected output represents a probability for a reconstructed candidate to belong to certain class. A model consists of several different parameters, like e.g. the depth of a tree or even the feature on which a decision will be made, which will be summarised under the variable $\boldsymbol{\theta}$ [34].

When a model is being trained, some parameters are the variable components that will be adjusted to fit the training data [34]. Others, so-called hyperparameters, are not adjusted during the training, but rather before. An objective function quantifies how well the model fits the data [34],

$$obj(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \Omega(\boldsymbol{\theta}). \tag{3.1}$$

This is a sum of the training loss function L, which measures the predictiveness of the model and the regularization term Ω, which represents the complexity of the model [34]. In general, when training a model, the goal is to minimise objective function. Considering Eq. 3.1, a good model needs to compromise between effectively classifying the data set, while remaining as simple as possible.

As already mentioned, the optimization is done iteratively, which is the general concept of gradient boosting. The first basic weak learner is added and makes predictions on the training data, by taking the average of the true target values of all candidates in a single leaf node [37]. Since the true target values of the training set are known, the difference between prediction and target, the so-called residual, can be calculated [37]. The next step of the optimization is to find the next weak learner predicting these residuals [37]. By trying to predict the error of the previous tree, and continuing this process for the entire model, the final output, the sum of all individual predictions (possibly scaled with a learning rate, regulating the influence of individual learners) will converge towards the target value.

XGBoost also optimises the way each weak learner is created. Each tree starts off with only the start node. Each possible branching into two leaves with the available features will then be tested for their overall information gain [34], i.e. does the added value of this branch in the shape of function loss optimization outweigh the added complexity? If this is the case for the most optimised splitting, the branch is created and the process

repeated for each node, however if the answer is no and no split increases the gain, then the branching will not be accepted and the process for this node is finished [34]. This process is known as pruning.

3.1.2.2 Hyperparameter Optimisation and Multiclass Classification

The implementation of XGBoost for this thesis has been achieved via the Heavy-Ion Physics Environment for Machine Learning (hipe4ml) Python package. The usage of this package also offers the possibility of optimizing hyperparameters (through the Optuna package [38]), such as number of estimators, maximum tree depth and learning rate, which improves the overall performance of the training and predictions [32]. To avoid overfitting during this process, the hyperparameters need to be tested on different samples of the data, but since there is usually only a limited number of training data available, the cross validation method is applied [32]. For this, the entire data set is divided in n fractions (called folds), where one fold is used as a test set, while the model is trained on the other $n-1$ folds [32]. After permuting through all possibilities, the average of the results is calculated as a final value for the hyperparameters used in this version of the model [32].

The BDTs model explained above is only able to perform binary classification, e.g. signal vs. background. However, for an analysis of non-prompt Λ_c^+, another class needs to be accounted for, resulting in a so-called multiclass classification of prompt, NP and background. There are two possible methods for splitting this multiclass classification problem into several binary classifications: these are known as One-vs-One (OvO) and One-vs-Rest (OvR).

The OvO approach trains one BDT for each combination of classes [39]. So for the three classes mentioned above, there are three possible combinations (prompt vs. NP, prompt vs. background and NP vs. background) and therefore three separate models trained. When applying, each model will make a prediction and a majority vote will determine the final result. Therefore, the two models containing the comparison between the true target class and another incorrect class needs to assign the candidate correctly to make the third comparison between the two incorrect classes irrelevant due to the majority vote. In case there is a three-way tie, i.e. all models assign the candidate to a different class, the candidate cannot be assigned to only one class, and should therefore not be used in further analysis (It is also possible to chose the vote with the highest output scores, if all candidates should be used).

The OvR approach trains one model for each class, where it will compete against all

other classes combined [39]. So again for the three classes, there are three possibilities: prompt vs. rest (NP and background), NP vs. rest (prompt and background) and background vs. rest (prompt and NP). With this method, the end result will be a score for the candidate for each class of zero to one, determining how likely it is that this candidate belongs to the class. A selection must then be made on the output scores when classifying the data. A successive analysis is performed to determine the selection thresholds for each class, known as the working point determination (Section 4.3). In this analysis, the OvR approach is chosen.

4 Analysis

This analysis focuses on the Λ_c^+ particle via the decay channel $\Lambda_c^+ \rightarrow pK^-\pi^+$. The data was measured for p–Pb collisions with the ALICE detector during Run 2 in 2016 at midrapidity range ($-0.96 < y < 0.04$). The energy for the collisions was $\sqrt{s_{NN}} = 5.02$ TeV. For this thesis, a data set with reconstructions and a Monte Carlo (MC) set with simulations were provided. The candidates given in the data set can be assigned one of three possible classes: either prompt candidate, NP candidate, or combinatorial background. This creates the need for classification in the data set. Therefore, the first step is the application of preselections, which are used to sort out easily classifiable candidates, followed by a more complex classification via BDTs. The preselection is applied to filter the candidates so that the ML training focuses on selections that could not be easily done by hand. The preselections were already applied to the data set prior to the work in this thesis, but are still briefly discussed here for completeness. Afterwards, the ML model training is explained with a description of the final models used for the classification. There are three transverse momentum intervals (2–4 GeV/c, 4–8 GeV/c and 8–12 GeV/c) used to train three models. After all of the available data has been given classification scores by the models, the working points, i.e. selections in the machine learning outputs which signify the difference between each of the three classes, need to be determined structurally. Lastly, fits to the invariant mass distributions of the selected candidates are used to conduct the signal extraction.

4.1 Preselections

As already mentioned, the preselections were already applied prior to this analysis, but are still listed for completeness in Tab. 4.1. The features will be briefly explained here.

- **dca**:
  The abbreviation stands for distance of closest approach. It is measured from the Λ_c^+ track to the reconstructed PV.

- **d_len**:
  This is the decay length of the reconstructed candidate, i.e. distance between the PV and the SV of the Λ_c^+ decay. It is expected that the NP candidates have a larger decay length on average, since their production origin deviates from the PV by the decay length of a prompt beauty hadron. A visualisation is drawn in Fig. 4.3.

Feature	Selection
dca	< 0.05 cm
d_len	> 0.005 cm
cos_p	> 0.8
sigma_vert	< 0.04 cm
pt_prongX	> 0.3 GeV/c

Table 4.1: Preselections applied to the reconstructed candidates.

- **cos_p**:
  This is the cosine of the so-called pointing angle, i.e. the angle between the momentum of a particle and the sum of the momenta of its decay products. It is expected to be close to unity.

- **sigma_vert**:
  This feature describes the resolution of the reconstructed SV via the three decay particles, also called prongs. It is the sum of the quadratic distances of closest approaches of all three prongs.

- **pt_prongX**:
  This variable describes the transverse momentum of the decay particle X, i.e. either proton, pion or kaon. In this case, the transverse momentum for all prongs was chosen to be above 0.3 GeV/c.

4.2 Machine Learning Training

For further separation between signal and background and also prompt and Non-Prompt candidates, rectangular selections by hand are not viable, due to the complexity of the data. Therefore, Boosted Decision Trees are used, which are able to use correlations of multiple features for classification. The concept has been described in the previous Chapter 3. For each of the aforementioned transverse momentum intervals a separate model is trained. The training data consists of simulated Monte Carlo (MC) data for the prompt and NP classes and a fraction of real data for the background. The background data is selected to exclude the invariant mass range $2.24 < m_{\text{candidate}} < 2.33$ GeV/c^2 for the training, since this contains the vast majority of real signal candidates. These selections are calculated via the MC data peak mean value and $\pm 5\sigma$ ranges.

p_T [GeV/c]	$2-4$	$4-8$	$8-12$
Prompt	107718	182350	43440
Non-Prompt	203164	339223	79670
Background	310882	421573	84817

Table 4.2: Number of candidates used in the training process for each class and each p_T interval.

In Tab. 4.2, the number of candidates used for the training is shown. For the training of the first two intervals, the number of background candidates has been chosen to be equal to the signal (combined prompt and NP) candidates. In the third interval, the ratio of prompt to NP to background is roughly 1:2:2, making it approximately a 3:2 signal to background ratio. This last ratio could have been adjusted to fit the 1:1 ratio and is due to the original ML attempts and the way the training data was prepared. However, the results of the models which are presented in the following sections were already sufficient enough and no important improvement was to be expected from increasing the training statistics. Therefore, the original ratio was kept. For each of these numbers in Tab. 4.2, only 80% are used for the actual training, while 20% are used for testing and validation. Overall, the fraction of background training candidates compared to the total number of candidates in the respective momentum intervals remained below 1.25% for all three models.

4.2.1 Training Variables

Using all available features for the training of the models is not optimal, since this adds unnecessary complexity which may result in overfitting. Therefore, only the ten most relevant features shall be used. Importantly, the invariant mass and transverse momenta of candidates will be excluded from the training feature space, since the model may learn to exploit the selections used to create the training data, resulting in biased classification. To determine the best set of features, a model for the momentum range $4 < p_{\mathrm{T}} < 8$ GeV/c is trained with all available features (except redundant PID variables). They are then evaluated on their average influence on the final output. Additionally, the degree of correlation between the features of the data has to be checked. Differences in correlations between the three classes are particularly useful to perform classification. Importantly, however, correlations between the training variables and the target observables, namely invariant mass and transverse momentum, should be avoided as much as possible. A

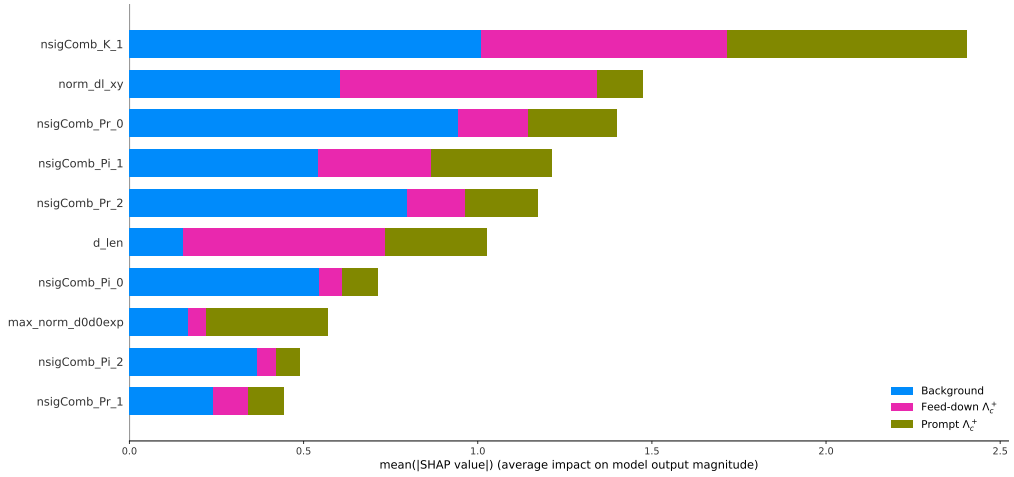

Figure 4.1: Quantification of the average impact of each of the ten most relevant features on the model for each class. This is for the model for $4 < p_{\mathrm{T}} < 8$ GeV/c.

selection of these kind of correlations risks the artificial enhancing or reducing of the signal, and thus biasing the results.

After the feature importance and correlations have been accounted for, the importance of the ten most relevant features in a model trained with only these as inputs is shown in Fig. 4.1 for the interval $4 < p_{\mathrm{T}} < 8$ GeV/c. The figure shows on the y-axis the training features sorted by relevance. The relevance is quantified by the mean SHAP (Shapley Additive Explanations) values along the x-axis, which in essence show the average impact on the model output of the respective feature [40]. The different colours show the relevance of the feature concerning the classification of a candidate to a certain class. The correlation matrices for each class for this model are shown in Fig. 4.2, where it shows no significant correlation (i.e. no strong red or blue colour) between the features and mass or momentum. However, other (strong) correlations do exist, which enable the model to classify efficiently. These features have also been used for the other two momentum intervals and the feature importance and correlations can be found in the appendix ($2 < p_{\mathrm{T}} < 4$ GeV/c: feature importance in Fig. 7.1 and correlations in Fig. 7.3; $8 < p_{\mathrm{T}} < 12$ GeV/c: feature importance in Fig. 7.2 and correlations in Fig. 7.4). Arguably, it is possible that these best features for one model are not necessarily the best for the other models too. However, differences between the feature importance over the models may also appear due to statistical fluctuations. Since a systematic study of the feature influence on the different models is not feasible with only three models and the

models show satisfying results with the same given features, no further investigations have been carried out.

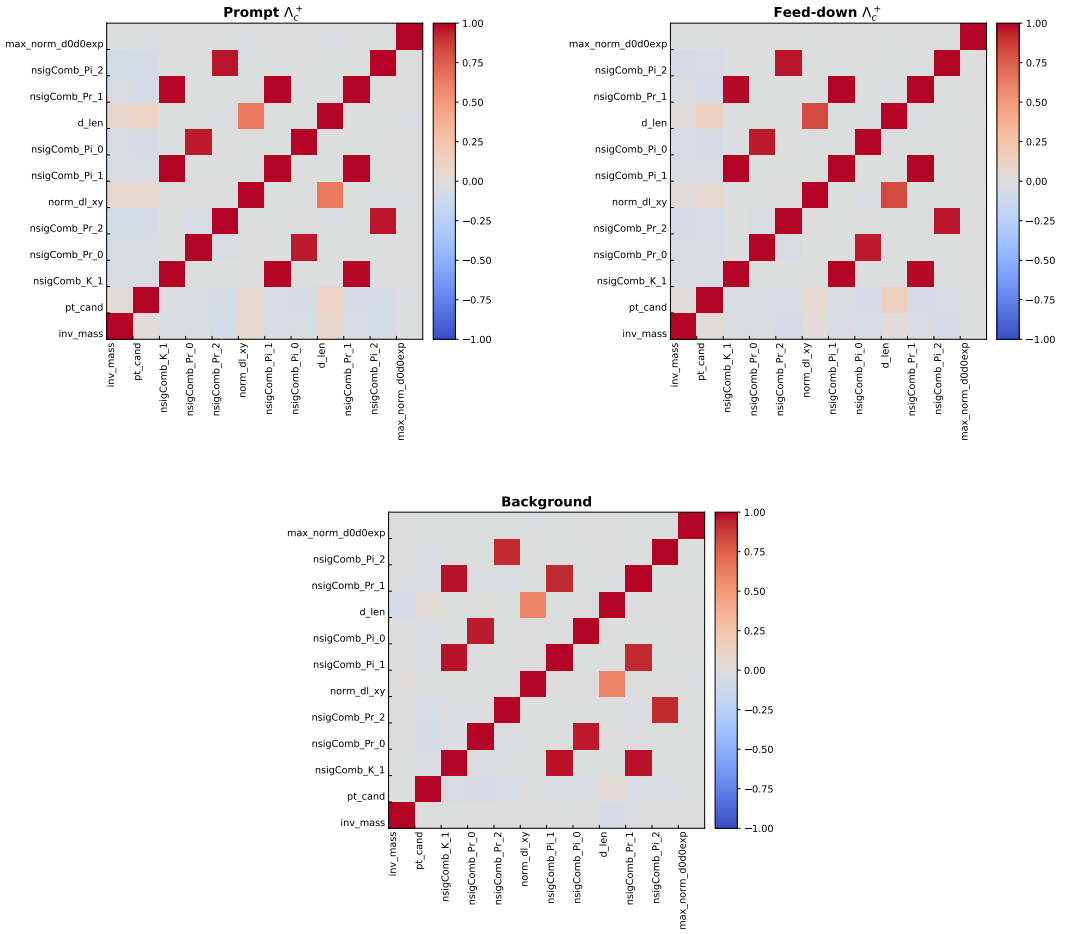

Figure 4.2: Correlation matrices of the ten relevant features with the additional parameters transverse momentum p_{T} and invariant mass m for $4 < p_{\mathrm{T}} < 8$ GeV/c.

Before continuing, the features used in the models need to be explained:

- **nsigComb_X_n**:
  All features with a name similar to this are PID variables. The X represents the particle species, so for the decay channel inspected in this thesis either proton (Pr), pion (Pi) or kaon (K). The n is representative for the number of the track which has been used to reconstruct the candidate. It can either be 0, 1 or 2, where the even numbers are tracks that show a curvature typical for a positively charged particle and the odd number shows negatively charged particle behaviour (or oppositely, in

case of the $\overline{\Lambda}_c^-$). The feature contains the combined PID information as explained in Section 2.4.

- **d_len**:
  This is the decay length of the reconstructed candidate, it was already explained in Section 4.1. A visualisation is shown in Fig. 4.3

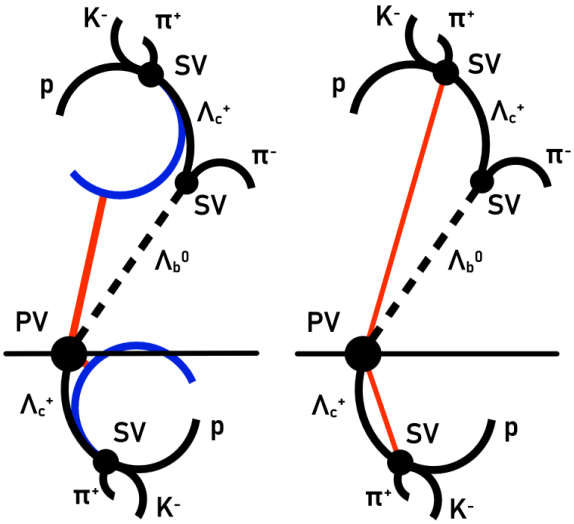

Figure 4.3: Again, for clarification, the top and bottom part do not actually originate from the same PV, but this is for illustrative purposes. Left: The red lines visualise examples of the impact parameter for non-prompt (above the PV) and prompt (below the PV) Λ_c^+ decay protons. The tracks of the protons get extrapolated beyond their origin (blue lines) and the distances of closest approaches are calculated. Right: The red lines visualise the reconstructed decay length for non-prompt (above the PV) and prompt (below the PV) Λ_c^+.

- **norm_dl_xy**:
  This is the normalised decay length in the x-y plane. The reconstructed track gets projected on the xy plane in the ALICE coordinate system and the decay length obtained is divided by its error.

- **max_norm_d0d0exp**:
  This feature describes the largest difference in measured and expected impact parameter of any of the three decay products of the Λ_c^+. Here, the measured impact parameter is the distance of closest approach between a continued trajectory of a decay particle and the PV. An example sketch can be seen in Fig. 4.3. The expected impact parameter is the sine of the pointing angle and multiplying it with the decay length of the initial particle.

The distribution of the candidate features in the momentum range $4 < p_{\mathrm{T}} < 8$ GeV/c for each class can be seen in Fig. 4.4. The background is generated using real data and cutting out the signal area (which can be seen in the first subplot), while the prompt and NP candidates are simulated MC data. These plots show that differences in the distributions also help the discrimination between the three classes. While the distributions are nicely visible for the three non-PID features, the PID distributions are distorted by a few outliers. It is important to note, that the values located in the high negatives (around -1000) are not real measurements, but actually candidates which actually have no PID information at all. An example of a zoomed in version of a PID distribution is shown in Fig. 4.5. The distributions for the other two momentum intervals can also be found in the appendix ($2 < p_{\mathrm{T}} < 4$ GeV/c in Fig. 7.5 and $8 < p_{\mathrm{T}} < 12$ GeV/c in Fig. 7.6).

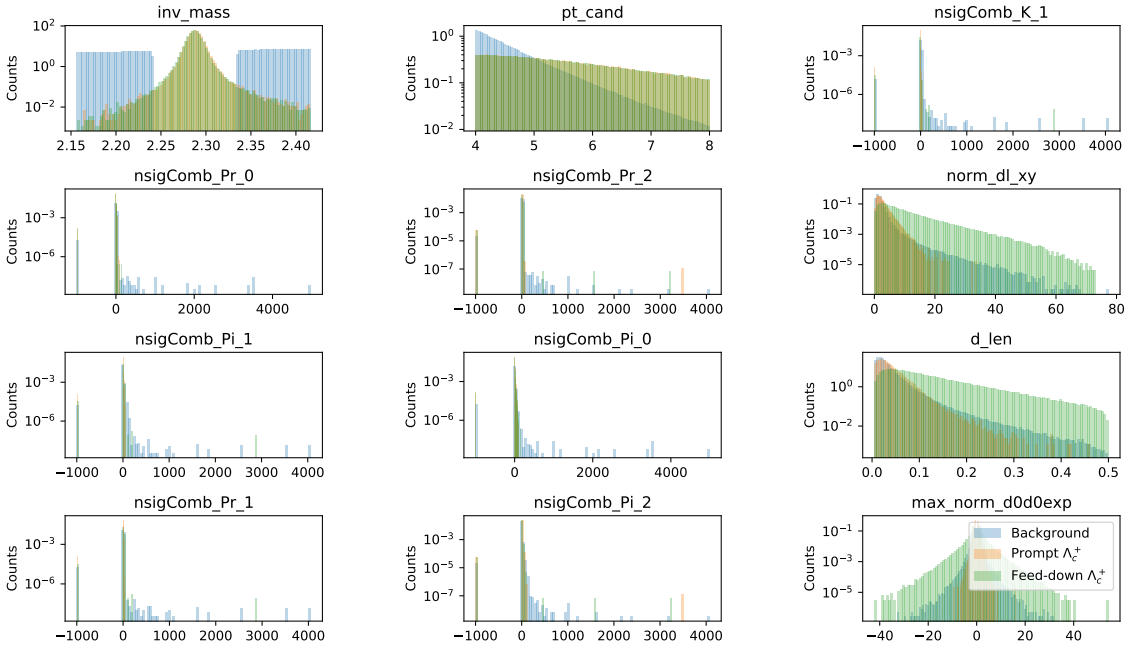

Figure 4.4: Distributions of the ten relevant features with additional parameters transverse momentum p_{T} and invariant mass m for $4 < p_{\mathrm{T}} < 8$ GeV/c.

4.2.2 Hyperparameter Optimisation

After the set of features is chosen, the hyperparameter optimisation is performed with a Bayesian optimisation approach via the Optuna package. In this approach, a given range of parameters are scanned iteratively while considering previous evaluations for choosing the next set of parameters [6]. To avoid statistical fluctuations while testing,

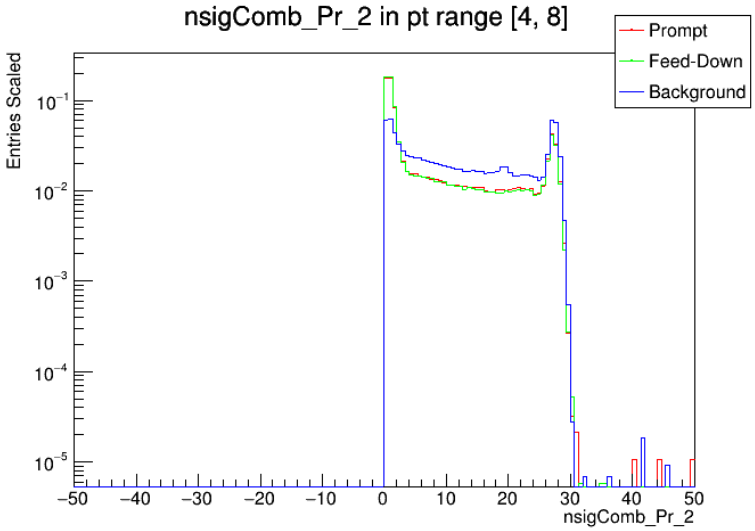

Figure 4.5: Zoomed in distribution of the nsigComb_Pr_2 feature for $4 < p_{\mathrm{T}} < 8$ GeV/c.

the cross validation method which was explained in Section 3.1.2.2 is used. Tab. 4.3 shows the optimal hyperparameters for each model.

p_T [GeV/c]	$2-4$	$4-8$	$8-12$
Max. Depth	3.9	4.0	3.8
Learning Rate	0.027	0.057	0.034
No. of Estimators	1927.4	1925.9	1924.4
Min. Child Weight	5.5	5.4	6.4
Subsample	0.98	0.91	0.84
Col. Sample by Tree	0.91	0.91	0.86

Table 4.3: Optimised hyperparameters for each model for the three p_{T} intervals.

The first parameter in Tab. 4.3, Max. Depth, is the maximum depth of a single weak learner in the BDT. For the optimisation, the range between 1 and 4 was tested. All three parameters are close to the upper limit of the tested range, which in general is not optimal, but it was found that a larger depth increases the overfitting of the ensemble, so 4 was chosen as a maximum upper limit.

The learning rate was already briefly mentioned in Chapter 3. It can be described as the rate at which the model adapts to the training data. A smaller learning rate results in a more careful and slower approach, while a larger learning rate results in faster

adaptation, but makes the model prone to overshooting the optimum. The tested range was from 0.01 to 0.1.

The number of estimators is the amount of weak learners in an ensemble. Since each weak learner only has a low tree depth, many of these low-complexity trees are combined for the ensemble. A range between 800 and 2000 was tested. Arguably, the parameters are close to the upper limit here as well. However, no tests were done with a higher upper limit, so if a better performance would be needed, these tests could yield some further improvements.

The fourth parameter in the table, Min. Child Weight, is a hyperparameter associated with the pruning process. It describes the minimum sum of weights in a child node for it to survive the pruning, therefore a larger value indicates less tree partitions [34]. The parameter range was set to be between 0.2 and 7.

The subsample parameter describes the percentage of available training data used in a single boosting step. By avoiding the use of all data all the time and rather randomly choosing subsamples of it, overfitting can be reduced [34]. The given range was from 0.8 to 1.

Lastly, Col. Sample by Tree is the fraction of available features used to generate a weak learner, where the features are chosen randomly according to the set fraction [34]. The reason for this is the same as for the subsample parameter and the set range for optimisation is also identical.

4.2.3 Trained Models

Eventually, three models with the optimised hyperparameters in Tab. 4.3 are trained with the features evaluated in Fig. 4.1 on the amount of data stated in Tab. 4.2. To control the model performance, the ROC curve is evaluated, where ROC stands for Receiver Operating Characteristics. The curve shows the true positive rate as a function of the false positive rate, e.g. looking at only the submodel which evaluates the prompt vs. rest classification, the positive label is given to candidates in the test set of the training data which are actually prompt. A candidate is a true positive, when the model classifies a prompt candidate correctly as prompt. Analogously, if the model assigns the prompt class to any other candidate that is either NP or background, it is called a false positive. Since the model itself does not actually assign classes to the candidates, but rather scores of how likely the candidates belong to the class, the true positive and false negative rates depend on where the selection in these scores is set. If the selection is set as low as possible, i.e. every candidate is classified as prompt, the true positive rate is 1, since

all prompt candidates are classified correctly as prompt. However, the false positive rate is 1 too, since actually every candidate is classified as prompt. On the other end of the spectrum, setting the selection as high as possible results in a true positive rate of 0, since all prompt candidates are incorrectly assigned. Yet, the false positive rate is 0 as well, since none of the non-prompt and background candidates are classified as prompt too. Changing the discrimination selection and drawing the respective true and false positive rates in a diagram will reveal the ROC curve. The worst model possible is random guessing, which will show a linear dependency in the diagram. In contrast, a good model will increase its true positive rate rapidly for small steps in the false positive rate in the beginning, indicating the capability to discriminate between prompt and the other classes effectively with a high true positive and low false positive rate. A way to quantify the quality of the model is the Area Under Curve (AUC) of the ROC curve. A random model will have an AUC score of 0.5, while a perfect discriminator will score 1. If a model scores lower than 0.5, it is objectively worse than random guessing, but it can actually be used to be better that random guessing, by just inverting the outputs. This will then yield an AUC score of better than 0.5.

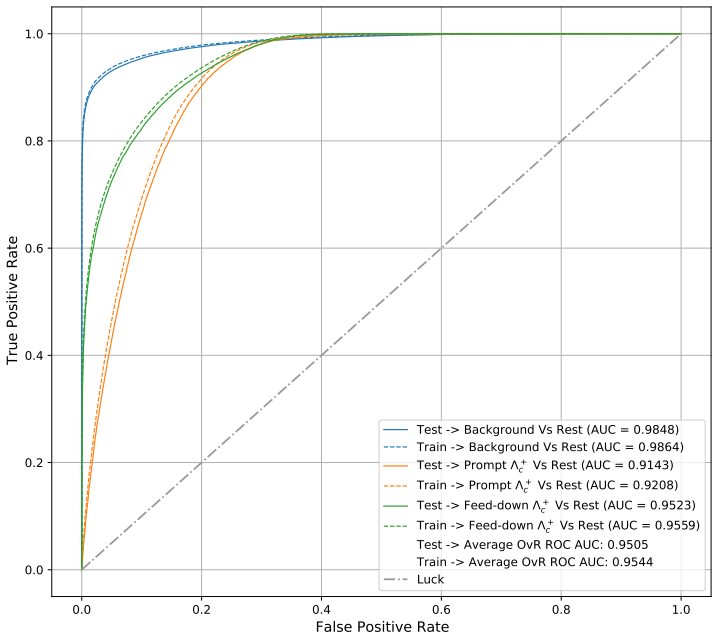

Figure 4.6: The ROC curves with the respective AUC values for the model for $4 < p_{\mathrm{T}} < 8$ GeV/c.

Furthermore, plotting the ROC curves and AUC scores for both training and test data enables the observation of overfitting. In case of low or no overfitting, both curves should be close together, while for a case of strong overfitting, the curves can deviate strongly, usually showing a better score for the training data.

Fig. 4.6 shows the ROC curve and the AUC scores for the model for $4 < p_{\mathrm{T}} < 8$ GeV/c. The plots for the other two intervals can be found in the appendix in Fig. 7.7 and Fig. 7.8. These show that the models are working efficiently and have very high scores, like seen in the respective figures or the summary in Tab. 4.4. Furthermore, the models show no serious signs of overfitting.

p_T [GeV/c]	$2-4$	$4-8$	$8-12$
Prompt Test	0.9247	0.9143	0.9037
Prompt Training	0.9297	0.9208	0.9199
NP Test	0.9541	0.9523	0.9440
NP Training	0.9571	0.9559	0.9524
Background Test	0.9927	0.9848	0.9840
Background Training	0.9937	0.9864	0.9876
Average Test	0.9572	0.9505	0.9439
Average Training	0.9602	0.9544	0.9533

Table 4.4: Summary of the AUC score of each model for each class with additional average score per model.

The results of the model can be seen in Fig. 4.7, where the test and training data have been evaluated and visualised in the distributions. Another validation that can be concluded here is that the distribution of the test and training data are not deviating much from one another, also validating that the model is not overfitted. It shows, that the model is able to determine background and NP very accurately, but cannot show such certainty for the prompt classification, where almost no candidates are scored with a probability above 0.8.

4.3 Working Point Determination

After confirming the quality of the BDTs, they can be applied on all available data to assign an output score to every reconstructed candidate. Afterwards, working points,

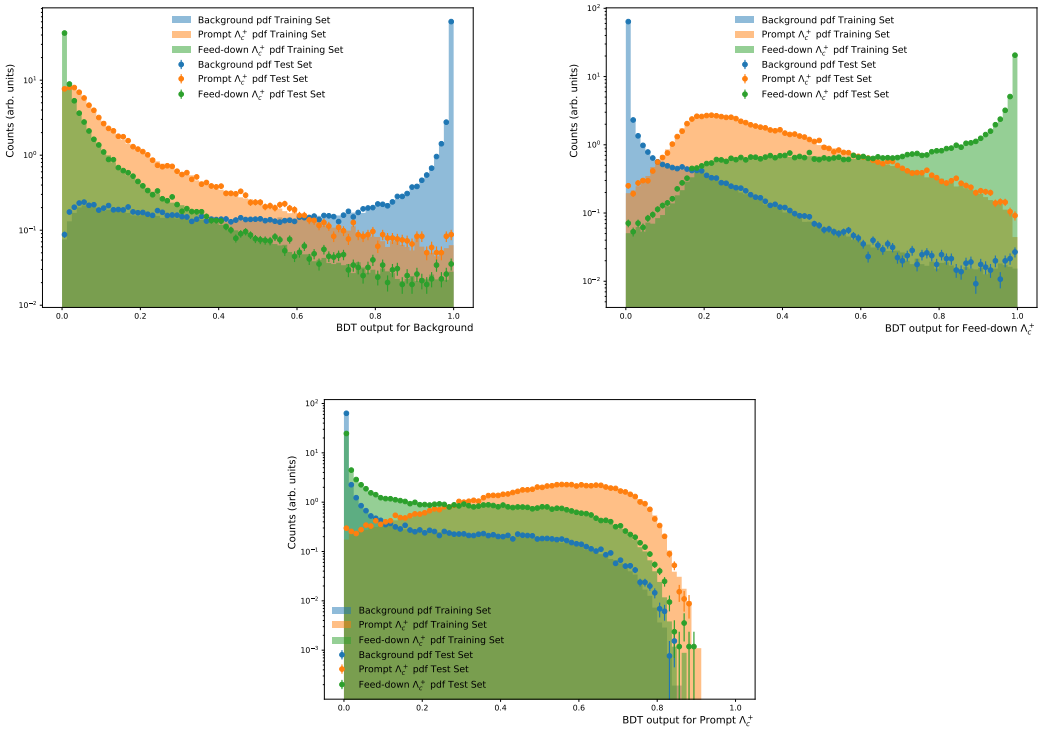

Figure 4.7: ML outputs for the training and test data for each class in $4 < p_{\mathrm{T}} < 8$ GeV/c.

i.e. selections in the outputs, have to be chosen to classify the candidates. In general, the background selections are set as upper bounds, i.e. only candidates below the chosen selection are valid, while selections in the prompt and NP outputs are lower bounds, therefore only candidates with an output above the selection are eligible. On one hand, if the working points are set too strictly, then the resulting signal shows a high purity, but the efficiency is very low. On the other hand, if the selections are set too loose, the efficiency is high, but the purity is very low. Therefore, a compromise needs to be found.

The compromise is obtained via the working point determination, where the pseudosignificance

$$S = \frac{s}{\sqrt{s+b}} \tag{4.1}$$

is calculated with the pseudosignal s and the real background b as functions of the BDT outputs. A pseudosignal is used, because the real significance of the signal may be subject to statistical fluctuations, which should be avoided when choosing the working point. Therefore, it is also recommended to not look at the real data for this process, otherwise a human bias might be induced.

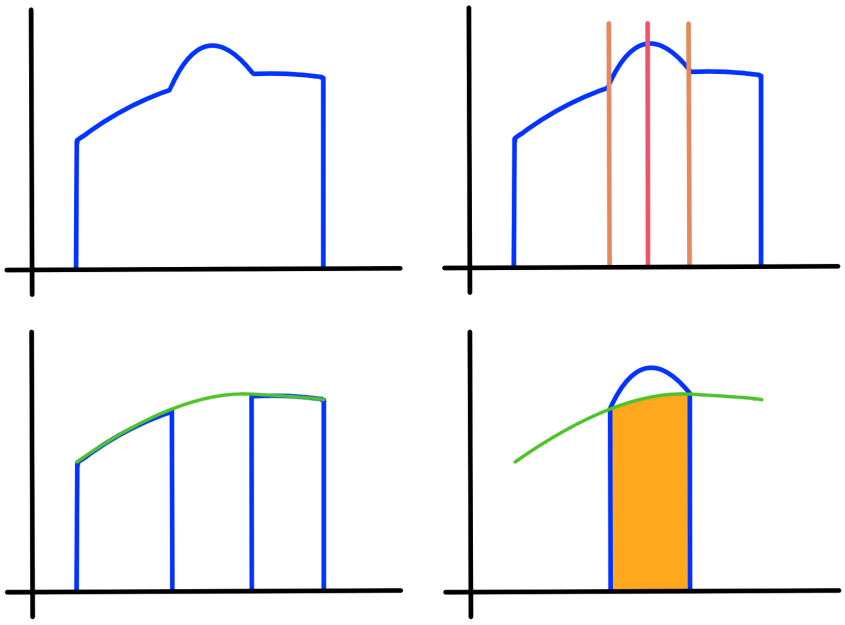

Figure 4.8: A visual representation of the process of determining the background for the pseudosignificance calculation. In the top left sketch, the invariant mass distribution is shown for a certain working point. The top right sketch shows the usage of the mean and standard deviation of the MC data to selection out the signal area. In the bottom left picture, a fit to the background side bands is performed, which is then in the bottom right picture used to estimate the amount of background in the signal area.

To calculate the pseudosignificance in Eq. 4.1, two variables are needed: background and pseudosignal. To determine the background component, the invariant mass histogram is generated for a fraction of the data with the desired BDT output selections. Afterwards, MC data is used to determine the expected mean and standard deviation of the signal peak, which is used to cut out the 3σ signal area in the invariant mass histogram. The resulting background side bands are used to fit a second-order polynomial. With the resulting fit, the behaviour of the background distribution within the signal peak area can be approximated and the number of background candidates within the 3σ range can be estimated by scaling the number up to the full data set. A sketch of this process can be seen in Fig. 4.8.

To calculate the pseudosignal, the formula for the cross section [6]

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}p_\mathrm{T}\mathrm{d}y} = \frac{1}{2} \cdot \frac{f_\mathrm{prompt}(p_\mathrm{T}) \cdot N^{\Lambda_\mathrm{c}^+}_\mathrm{raw}(p_\mathrm{T})}{\Delta y \cdot \Delta p_\mathrm{T} \cdot (\mathrm{Acc} \times \epsilon)_\mathrm{prompt}(p_\mathrm{T}) \cdot \mathrm{BR} \cdot \mathcal{L}_\mathrm{int}} \tag{4.2}$$

can be rearranged for the $N^{\Lambda_\mathrm{c}^+}_\mathrm{raw}(p_\mathrm{T})$ variable, which stands for the raw yield, i.e. without

any corrections:

$$N_{\text{raw}}^{\Lambda_c^+}(p_T) = 2 \cdot \frac{\mathrm{d}^2\sigma}{\mathrm{d}p_T \mathrm{d}y} \cdot \frac{\Delta y \cdot \Delta p_T \cdot (\text{Acc} \times \epsilon)_{\text{prompt}}(p_T) \cdot \text{BR} \cdot \mathcal{L}_{\text{int}}}{f_{\text{prompt}}(p_T)}. \tag{4.3}$$

In these equations, there are many variables that need some explanation. First, Δp_T is the transverse momentum range for each of the three intervals and Δy is the observed rapidity range, which is given through detector limitations as $\Delta y = 1.6$. A factor of 2 appears in this formula, because both particles and antiparticles need to be considered. BR is the branching ratio for the desired decay of $\Lambda_c^+ \rightarrow \text{pK}^-\pi^+$. As mentioned in Chapter 1, this is given as $(6.28 \pm 0.32)\%$. The acceptance times efficiency term $(\text{Acc} \times \epsilon)_{\text{prompt}}$ describes the total efficiency for prompt Λ_c^+, i.e. taking detector acceptance, preselection efficiency and BDT efficiency into account. The acceptance and preselection efficiency are calculated with the MC data, where the number of candidates that pass through the acceptance range and preselections is divided by the total number of generated candidates. This is independent of the BDT and therefore constant for each respective p_T interval. The BDT efficiency is calculated in the same way, but is dependent on the selections in the BDT outputs and therefore needs to be calculated again for every selection. The integrated luminosity $\mathcal{L}_{\text{int}}$ can be calculated by dividing the total number of analysed events N_{events} ($\approx 627 \cdot 10^6$) by the minimum bias cross section of p–Pb collisions $\sigma_{\text{pPb,mb}}$ ($= 2.093$ pb [41]). This yields $\mathcal{L}_{\text{int}} \approx 0.3$ nb^{-1}. The differential cross section is obtained via FONLL calculations in pp collisions at 5.02 TeV, scaled by the lead mass number. FONLL (Fixed-Order plus Next-to-Leading-Log) is an implementation to improve calculations concerning the transverse momentum spectrum for heavy flavour particles systematically [42]. Lastly, the fraction of prompt candidates f_{prompt} can be calculated with the FONLL cross sections and acceptance times efficiencies calculated earlier.

With Eq. 4.3 and the method of background determination described earlier, the pseudosignificance can be calculated for two selections in the three ML outputs for certain ranges. The calculations for $4 < p_T < 8$ GeV/c are shown in Fig. 4.9 for prompt and Fig. 4.10 for NP candidates. Most relevant in these figures are the top left and the two bottom plots. The top left shows the pseudosignificances calculated for the respective selections in the ML outputs. The lower plots show the fraction of prompt (bottom left) and NP (bottom right) candidates in the signal. The other plots show intermediate steps for the calculation of the pseudosignificance, such as the expected signal and background numbers. Choosing the selection for the prompt signal, attention should be paid to a high pseudosignificance at a high fraction of prompt, while the NP signal analogously should

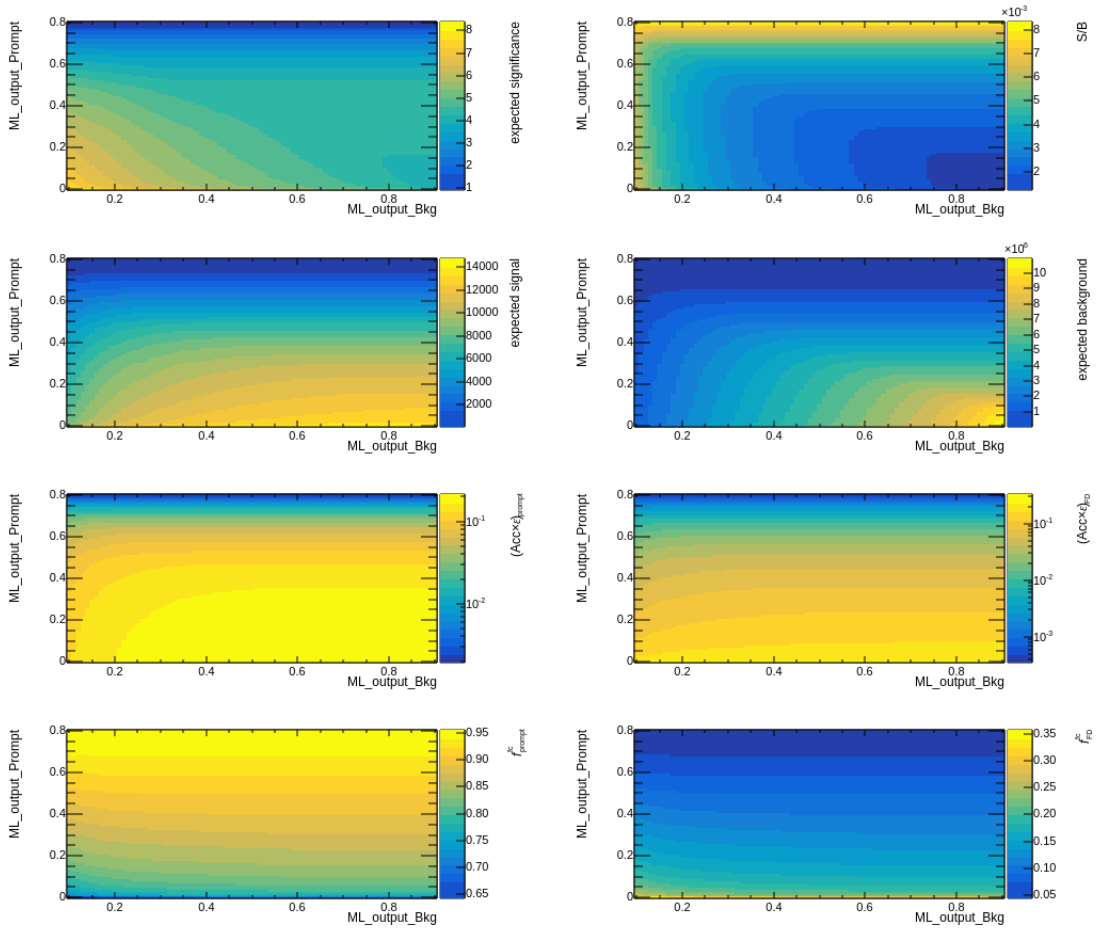

Figure 4.9: Working point calculations as functions of the ML output selections for prompt candidates in $4 < p_{\mathrm{T}} < 8$ GeV/c.

have a low prompt and therefore a high NP fraction. The selections chosen for the signal extraction can be found in Tab. 4.5 and Tab. 4.6. The tables also feature the expected significances, as well as the expected fractions of prompt or NP.

The calculations for the other intervals are included in the appendix in Fig. 7.11 and Fig. 7.13 for prompt and Fig. 7.12 and Fig. 7.14 for NP.

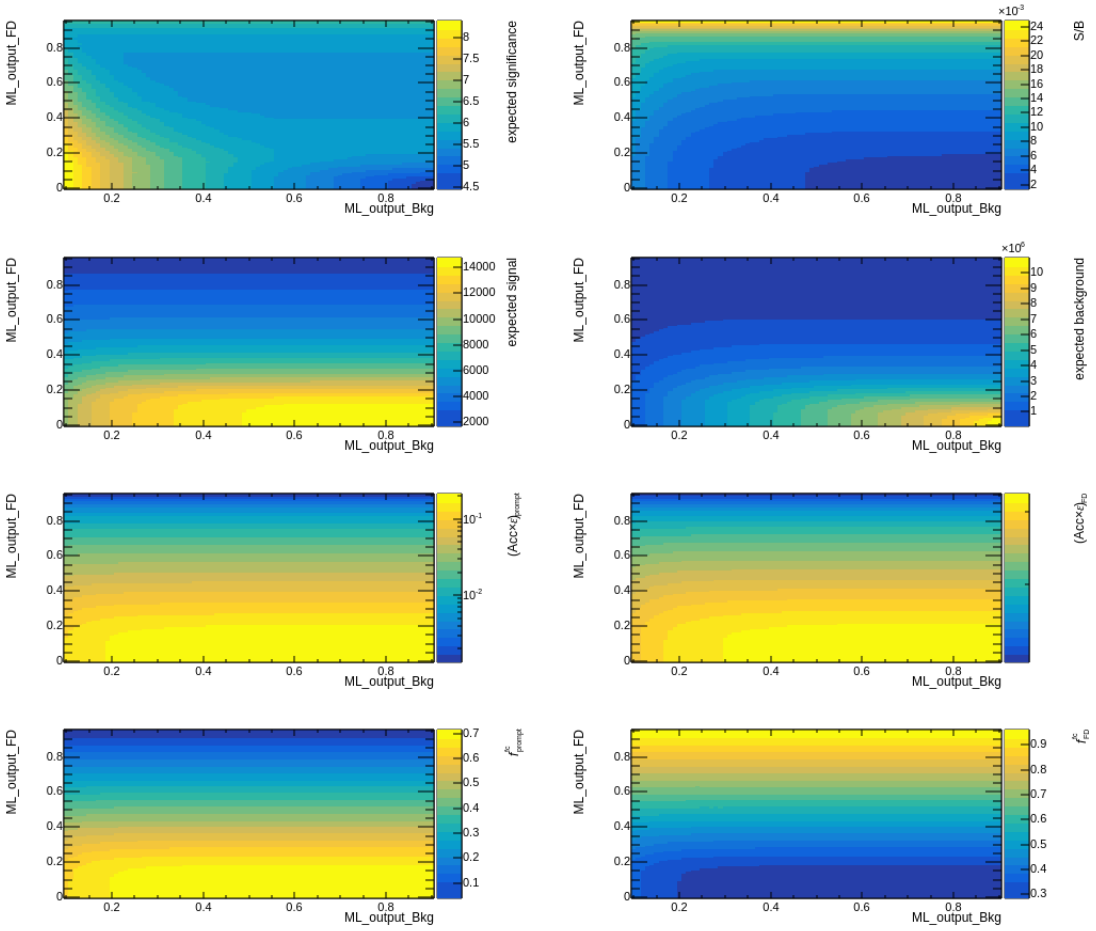

Figure 4.10: Working point calculations as functions of the ML output selections for NP candidates in $4 < p_{\mathrm{T}} < 8$ GeV/c.

p_T [GeV/c]	$2-4$	$4-8$	$8-12$
Bkg. Selection	0.08	0.12	0.20
Prompt Selection	0.00	0.00	0.00
Prompt Fraction	0.73	0.66	0.69
Exp. Significance	3.2	8.1	9.0

Table 4.5: Summary of the working points chosen for each p_{T} for the prompt signal. Additionally, the expected prompt fraction is listed, as well as the expected significance from the calculations.

p_T [GeV/c]	$2-4$	$4-8$	$8-12$
Bkg. Selection	0.11	0.10	0.20
NP Selection	0.74	0.75	0.70
NP Fraction	0.73	0.80	0.75
Exp. Significance	2.1	6.2	6.5

Table 4.6: Summary of the working points chosen for each p_{T} for the NP signal. Additionally, the expected NP fraction is listed, as well as the expected significance from the calculations.

5 Results

After the working points have been selected, they can be applied and the resulting invariant mass spectra can be fit. The resulting fits can be seen in Fig. 5.1 for NP and in Fig. 5.2 for prompt, as well as their respective background subtracted residuals.

In these fits, a Gaussian distribution was chosen for the signal peak, while a second order polynomial was chosen as a background fit. In the fits, the red line shows the background fit, while the blue line shows the total fit function. Since there is no physical explanation and expectation for the background, a polynomial is satisfying as a background fit function. For the NP fit, the width of the Gaussian was fixed by giving the respective values of the prompt fits as inputs, to improve the stability of the fit due to higher statistics in the prompt peak.

p_T [GeV/c]	$2-4$	$4-8$	$8-12$
NP Significance from WP	2.1	6.2	6.5
NP Significance form Fit	4.1	6.8	6.8
P Significance from WP	3.1	8.1	9.0
P Significance from Fit	6.1	13.8	12.0

Table 5.1: Comparison of the expected significances via the working point calculations and the significances gained from the fits. For clarification, P is an abbreviation of prompt, while NP is the usual non-prompt acronym and WP is the acronym for working point.

In the figures, the significances of the respective signal peaks are also listed. In Tab. 5.1, they are compared with the expectations gained from the working point calculations. It shows good alignment for the NP significances. However, for the prompt, some stronger deviations are found. These strong deviations in the prompt cases and the fact that all working point calculations are below the fit values may be cause of statistical fluctuation, but this is rather unlikely. This suggests that perhaps the MC simulations used in the working point calculations are not accurate descriptions of the real data.

When applying the fit, a third order polynomial was also tested as a background function. This resulted in only minor, if any, improvements, namely only additional 0.1 in the significances found in the lowest momentum interval for prompt and NP. Weighing up the minor improvements and the additional complexity of the function, it was decided that the simpler model (second order) is preferred and kept to produce the final results.

p_T [GeV/c]	$2-4$	$4-8$	$8-12$
NP Signal S	5466 ± 1064	3345 ± 399	734 ± 100
NP Background B	1763327 ± 510	237962 ± 208	10746 ± 56
NP S/B	0.0031	0.0141	0.0683
NP Mean [GeV/c^2]	2.289 ± 0.001	2.289 ± 0.001	2.289 ± 0.001
NP Width [GeV/c^2]	0.005 ± 0.000	0.006 ± 0.000	0.009 ± 0.000
P Signal S	20052 ± 3334	17405 ± 1343	3227 ± 287
P Background B	10812985 ± 1263	1580442 ± 535	69230 ± 143
P S/B	0.0019	0.0110	0.0466
P Mean [GeV/c^2]	2.289 ± 0.001	2.289 ± 0.001	2.290 ± 0.001
P Width [GeV/c^2]	0.005 ± 0.001	0.006 ± 0.001	0.009 ± 0.001

Table 5.2: Summary of other attributes calculated by the fit. Note, the width for the NP was fixed as the values gained by the prompt fit, therefore the uncertainties are not actually 0.000.

Besides the significance, the panels in the figures also show other attributes, namely signal and background counts (S and B), the signal-to-background ratio S/B (within the 3σ range of the peak) and the mean and width of the peak. A summary of them can be found in Tab. 5.2. Looking at these values, it can be seen that for prompt and NP both signal and background counts are decreasing with increasing momenta. However, the signal-to-background ratio is increasing with increasing momenta. The mean of all peaks except one was found to be (2.289 ± 0.001) GeV/c^2, with the only exception being the 8-12 GeV/c prompt interval at (2.290 ± 0.001) GeV/c^2. Therefore, these values include the reference ($m_{\Lambda_c^+} = (2286.46 \pm 0.14)$ MeV/c^2 [4]) within their 3σ ranges, with the mentioned exception including the reference at 4σ. However, the uncertainties in the table are only statistical, therefore it is likely that by eventually including systematic uncertainties, the values will coincide within the 3σ range in all cases. Lastly, the prompt signal peak width was found to be increasing from 0.005 GeV/c^2 to 0.009 GeV/c^2 with increasing momentum.

The actual fractions of (non-)prompt candidates in the (non-)prompt signals are not measured in this thesis and are part of the extended analysis. However, they were estimated via simulations in the working point calculations and were also listed in Tab. 4.6 and Tab. 4.5. For prompt, they range between 66% and 73% and for NP between 73% and 80%. With reference to the significances, it is shown that the majority of signal candidates used to calculate these significances are actually (non-)prompt candidates, which

is important for testing the feasibility. Considering the significances and fractions of NP candidates found in this study, the feasibility of a full NP analysis in the $\Lambda_c^+ \to pK^-\pi^+$ channel can be confirmed.

p_T [GeV/c]	$2-4$	$4-8$	$8-12$
NP Significance $\Lambda_c^+ \to pK^-\pi^+$	4.1	6.8	6.8
NP Significance $\Lambda_c^+ \to pK_s^0$	3.6	4.8	3.2
P Significance $\Lambda_c^+ \to pK^-\pi^+$	6.1	13.8	12.0
P Significance $\Lambda_c^+ \to pK_s^0$	8.9	10.3	6.2
NP Fraction $\Lambda_c^+ \to pK^-\pi^+$	0.73	0.80	0.75
NP Fraction $\Lambda_c^+ \to pK_s^0$	0.41	0.44	0.41
P Fraction $\Lambda_c^+ \to pK^-\pi^+$	0.73	0.66	0.69
P Fraction $\Lambda_c^+ \to pK_s^0$	0.91	0.91	0.83

Table 5.3: Comparison of the significances and fraction of (non-prompt) candidates of the analyses of the $\Lambda_c^+ \to pK_s^0$ and $\Lambda_c^+ \to pK^-\pi^+$ channels. The values for the fraction for this analysis are estimations from the working point determination, while for the $\Lambda_c^+ \to pK_s^0$ channel, they are the measured fractions after the application of the working points, without further corrections or optimisations. The values for the $\Lambda_c^+ \to pK_s^0$ were provided in personal communications.

Additionally, the results found in this analysis can be compared to the results of an analysis of the $\Lambda_c^+ \to pK_s^0$ channel introduced in Chapter 1. An overview of that can be found in Tab. 5.3. Comparing the significances for the NP cases, it is found that they are higher for the $\Lambda_c^+ \to pK^-\pi^+$ channel, with increases between 0.5 and 3.6. Also, the estimated fractions of non-prompt candidates in this analysis are found to be higher than the measured values for the $\Lambda_c^+ \to pK_s^0$ channel, at values of 73% to 80% compared with 41% to 45%. This means, that the significance was improved with respect to the $\Lambda_c^+ \to pK_s^0$ analysis, while also improving the fraction of the desired non-prompt candidates. This shows, that although the $\Lambda_c^+ \to pK^-\pi^+$ channel has a larger combinatorial background than the $\Lambda_c^+ \to pK_s^0$ channel, the classification and isolation of the non-prompt candidates was still very successful, making the increase in statistics (due to the branching ratios explained in Chapter 1) noticeable.

Comparing the values for the prompt cases of this thesis to the $\Lambda_c^+ \to pK_s^0$ analysis shows a significance decrease of 2.9 in the 2-4 GeV/c interval, while showing an in-

crease of 3.5 and 5.8 in the 4-8 GeV/c and 8-12 GeV/c intervals, respectively. This thesis found prompt fraction estimated between 66% and 73%, while the $\Lambda_c^+ \rightarrow pK_s^0$ analysis measured values between 83% and 91%. However, here it has to be considered that the optimisation of the prompt identification was not the focus of this work. The selections for the prompt classification in Tab. 4.5 actually show no selections in the prompt ML output, making this rather a binary classification of background vs. signal. As the fractions show, a majority of these signal candidates are then actually prompt, making this classification sufficient for the purpose of this thesis.

For further context of the use of this work, Fig. 5.3 shows the nuclear modification factor of Eq. 1.2 for prompt and NP $\Lambda_c^+ \rightarrow pK_s^0$ over the transverse momentum. It clearly shows deviations of the prompt class from unity, while the NP values are compatible with unity, within their large statistical uncertainties. Now, adding the statistical significance of the decay channel analysed in this thesis could decrease the uncertainties in the NP case, allowing a better interpretation of the behaviour of the NP Λ_c^+. Understanding this NP behaviour will then give indirect insight in the beauty hadron sector concerning the initial state effects.

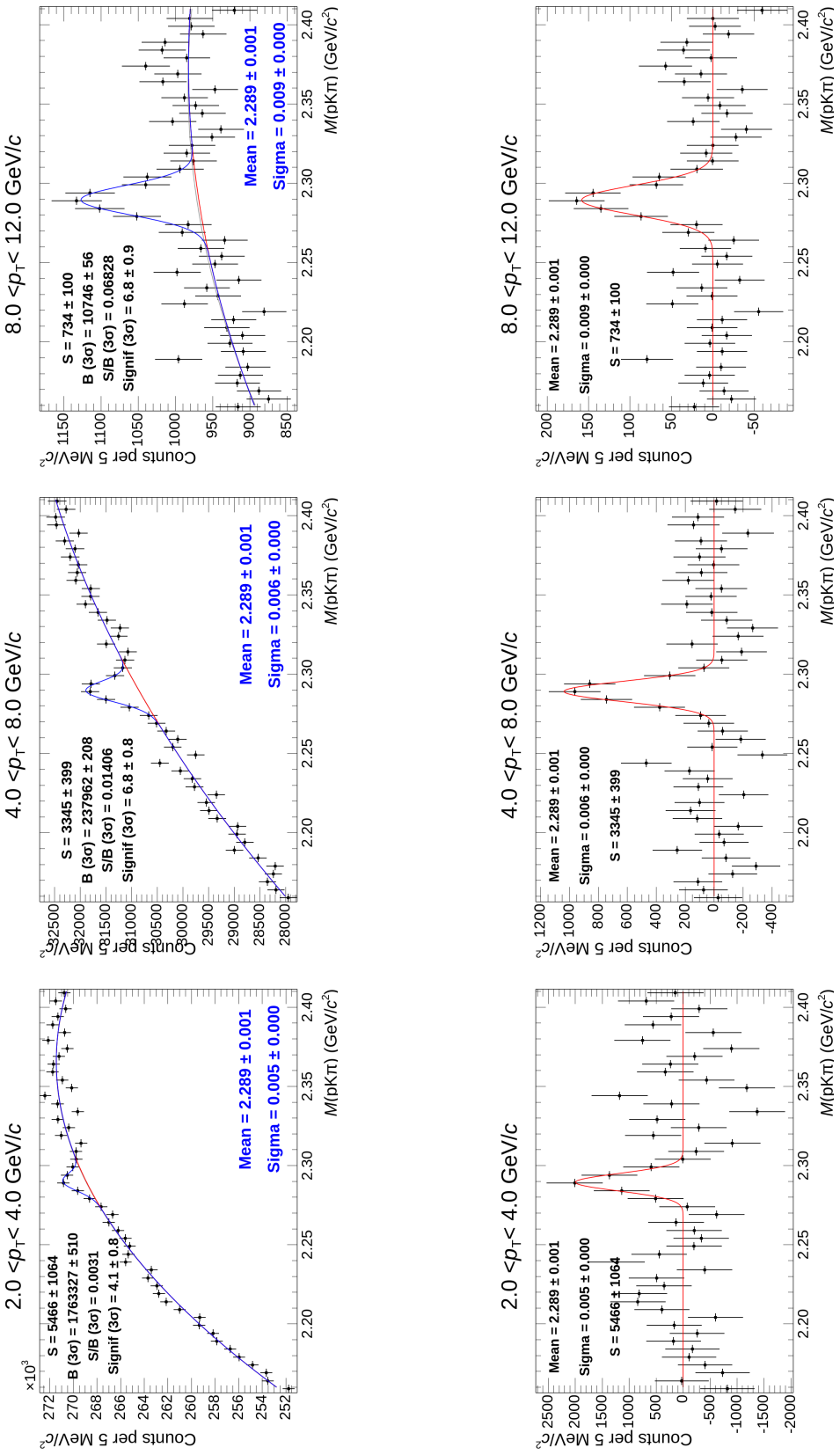

Figure 5.1: Invariant mass fits and background subtracted residuals for NP $\Lambda_c^+ \rightarrow pK^-\pi^+$ for all three p_T ranges.

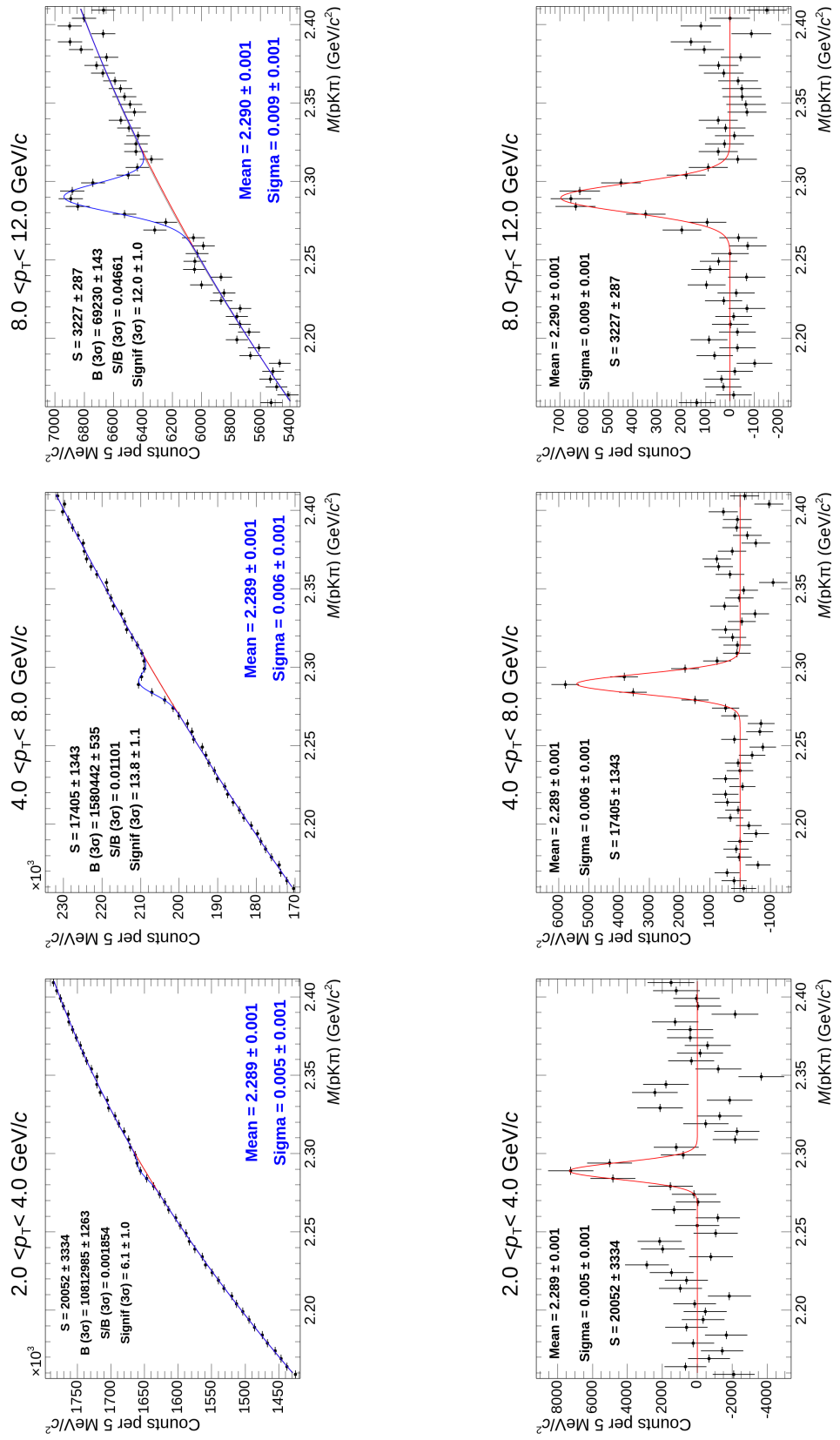

Figure 5.2: Invariant mass fits and background subtracted residuals for prompt $\Lambda_c^+ \to pK^-\pi^+$ for all three p_T ranges.

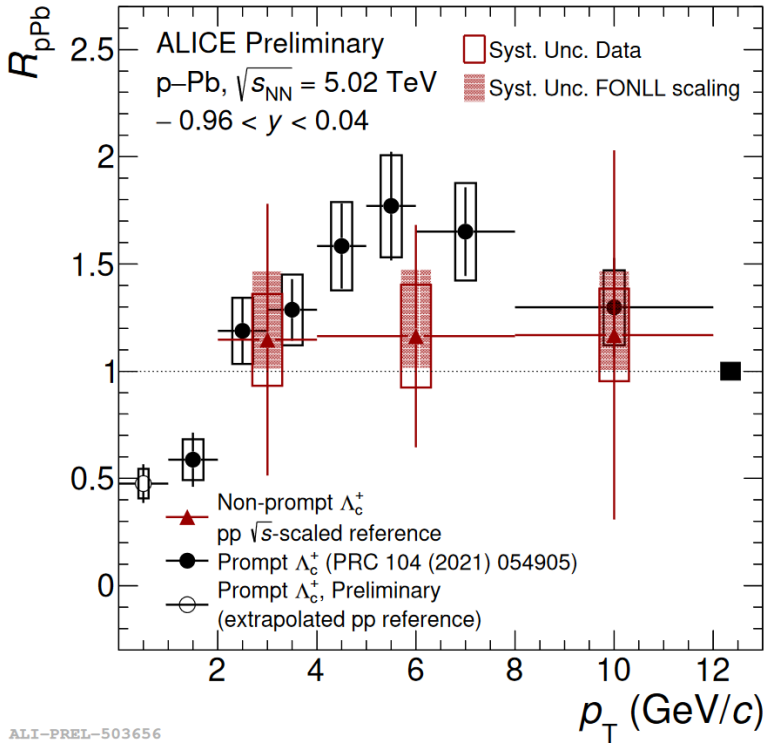

Figure 5.3: Comparison of the nuclear modification factor R_{pPb} of prompt and NP Λ_c^+ at $\sqrt{s_{NN}} = 5.02$ TeV in the transverse momentum interval $2 < p_T < 12$ GeV/c. These are the results of an analysis of the $\Lambda_c^+ \rightarrow pK_s^0$ decay channel. Image source: [43]

6 Conclusion and Outlook

In this thesis, the feasibility of a Λ_c^+ analysis for NP cases in the decay channel $\Lambda_c^+ \rightarrow pK^-\pi^+$ for Run 2 data at midrapidity and an energy of $\sqrt{s_{NN}} = 5.02$ TeV was investigated in p–Pb collisions. The data which was analysed consisted of reconstructed Λ_c^+ candidates in the p_T range from 2 GeV/c to 12 GeV/c, which was split in three intervals. Since this decay channel has a large amount of combinatorial background candidates, a BDT was trained with MC signal simulations and real background from a fraction of the data to perform the classification into prompt, NP and background for each p_T interval separately. Optimal pairs of ML outputs were selected based on pseudosignificances, as explained in Section 4.3. The final results of this thesis are the extracted signal peaks and their significances in Fig. 5.2 for prompt and Fig. 5.1 for NP Λ_c^+. The prompt signals show significances between 6.1 and 13.8, while more importantly the NP signals show significances between 4.1 and 6.8. For NP, the estimated fraction of NP candidates in the signal is between 73% and 80%, so the majority of the signal actually consists of NP candidates. Therefore, the feasibility of the NP $\Lambda_c^+ \rightarrow pK^-\pi^+$ analysis can be confirmed. Hence, further continuation of the cross section calculations are reasonable.

Comparisons with the analysis of the $\Lambda_c^+ \rightarrow pK_s^0$ channel in Tab. 5.3 show for the NP cases, that the NP signal fraction is increased in this thesis by a factor of around 1.8, while also increasing the significances by values between 0.5 and 3.6. This suggests that the increased NP statistics in the $\Lambda_c^+ \rightarrow pK^-\pi^+$ could improve the uncertainties of previous $\Lambda_c^+ \rightarrow pK_s^0$ results.

The continuation of the analysis contains most notably the efficiency corrections of the extracted signals, as well as the subtraction of NP candidates from the prompt signal (and vice versa) and the estimation of systematic uncertainties. The efficiency corrections take the detector acceptance and the preselection and BDT efficiencies into account. This needs to be done, since the extracted signals do not contain every real (non-)prompt candidate. Some candidates have been filtered out in earlier processes, however to eventually calculate the final corrected cross section, these candidates have to be considered. Afterwards, the signals also contain many candidates which are falsely classified, therefore the fraction of (non-)prompt candidates has to be subtracted from the signal. Lastly, no systematic uncertainties were evaluated in this thesis. However, to portray the uncertainties of the final results correctly, systematic uncertainties have to be considered. In the scope of this thesis, systematic uncertainties for the BDT output selections and yield extraction would need to be considered. The ML selections can be varied to gener-

ate looser and tighter selections, whose results will then offer the possibility to estimate uncertainties. For the yield extraction, variations in e.g. the background fit, bin width and signal range can all result in slightly different yields.

Finally, the results of a continuation of this analysis may allow further investigation of the hadronisation mechanisms in the p–Pb collision system, by providing more statistical significance to the NP Λ_c^+ values in Fig. 5.3. The indirect approach via the NP will eventually contribute to the understanding of the beauty hadronisation. However, it is likely that this analysis alone will not be sufficient to clarify the NP behaviour and the investigation will rely on the statistics eventually provided by Run 3. All in all, this thesis may not give a complete piece of the puzzle of the heavy quark hadronisation mechanism, but it can at least give a hint as to where the next piece might be.

7 Appendix

7.1 Feature Importance

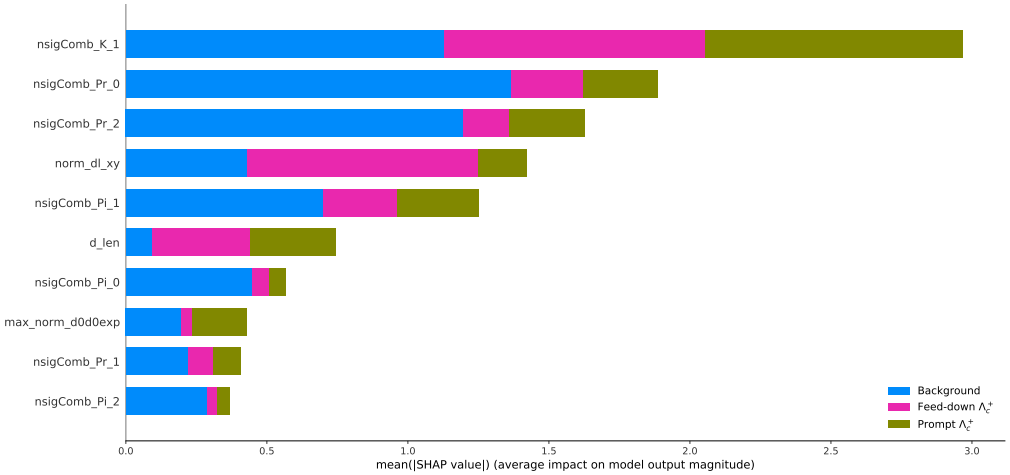

Figure 7.1: Quantification of the average impact of each of the ten most relevant features on the model for each class. This is for the model for $2 < p_{\mathrm{T}} < 4$ GeV/c.

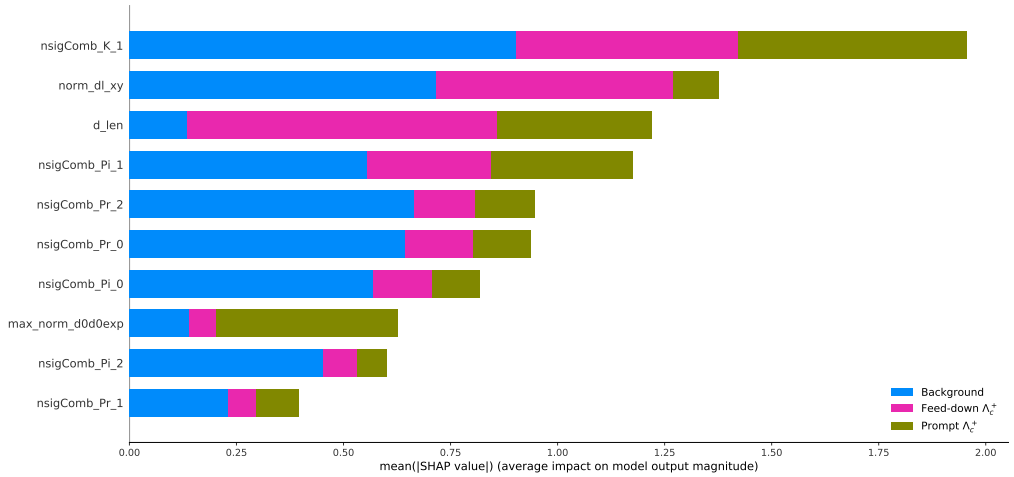

Figure 7.2: Quantification of the average impact of each of the ten most relevant features on the model for each class. This is for the model for $8 < p_{\mathrm{T}} < 12$ GeV/c.

7.2 Correlation Matrices

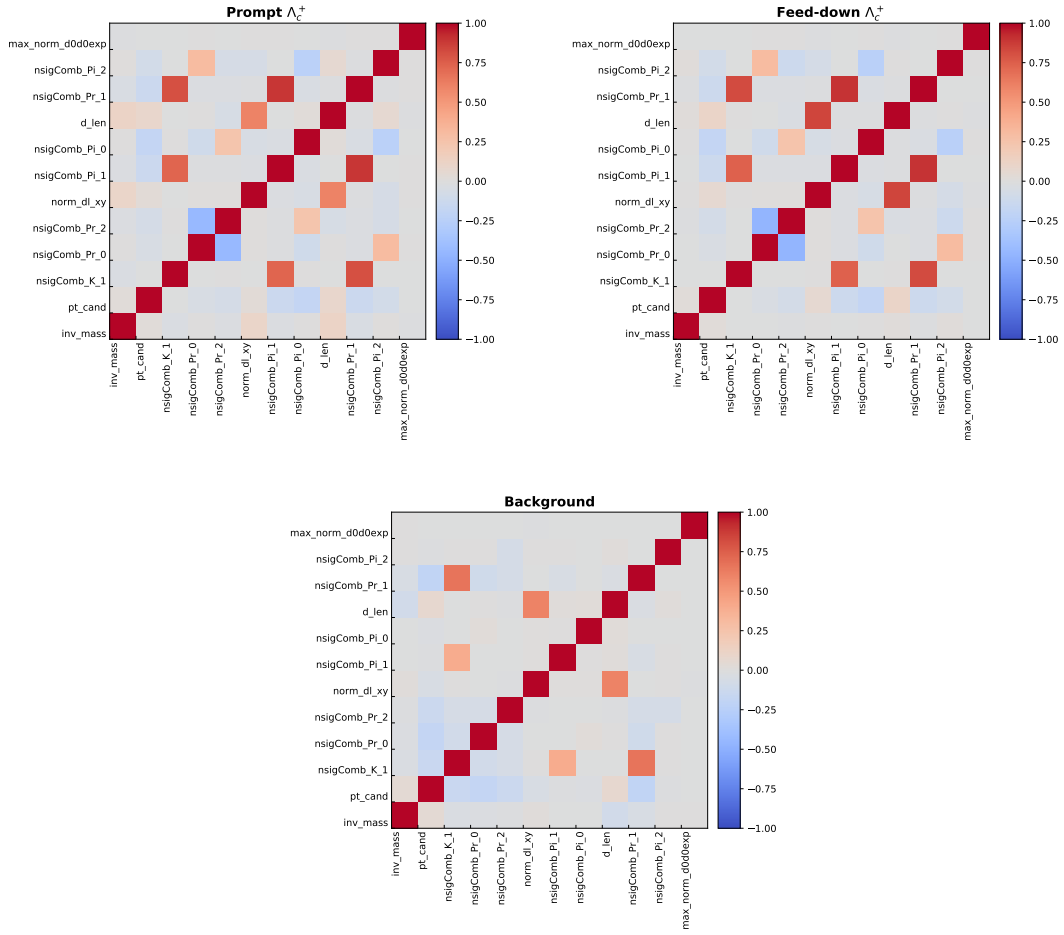

Figure 7.3: Correlation matrices of the ten relevant features with the additional parameters transverse momentum p_{T} and invariant mass m for $2 < p_{\mathrm{T}} < 4$ GeV/c.

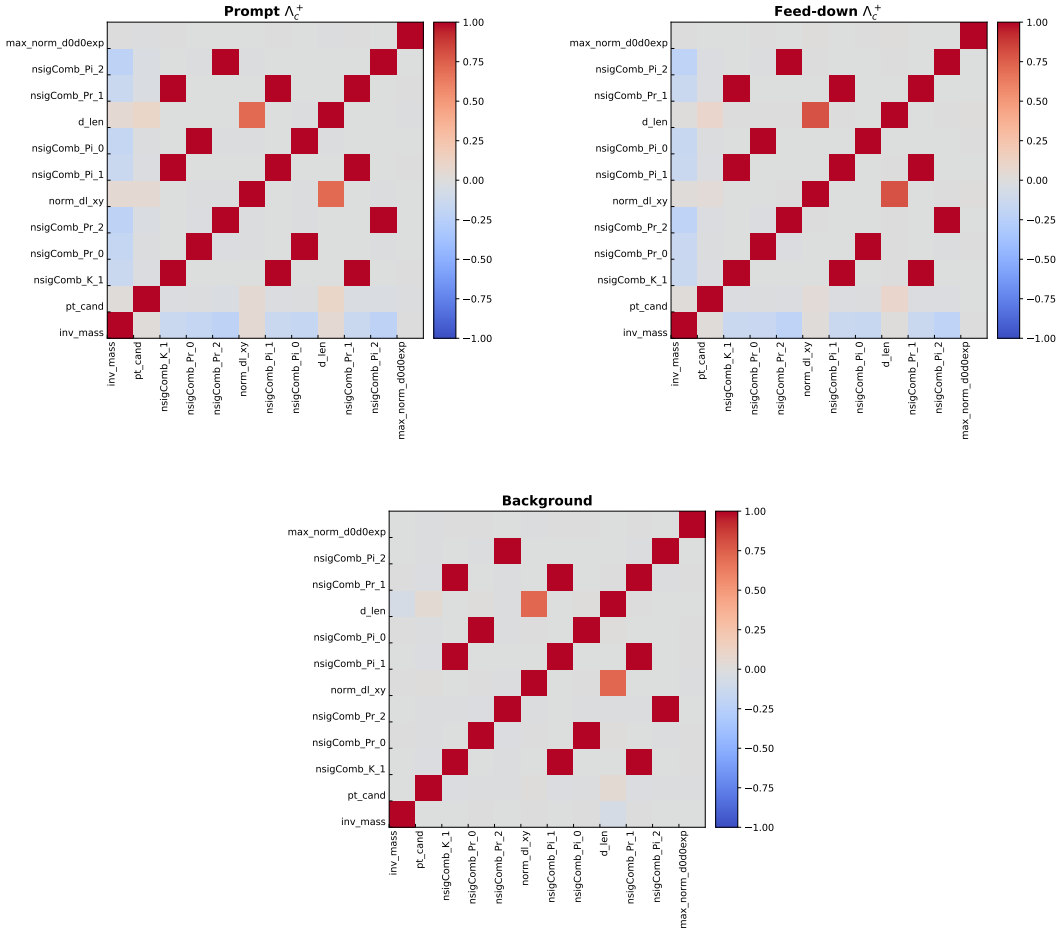

Figure 7.4: Correlation matrices of the ten relevant features with the additional parameters transverse momentum p_{T} and invariant mass m for $8 < p_{\mathrm{T}} < 12$ GeV/c.

7.3 Feature Distributions

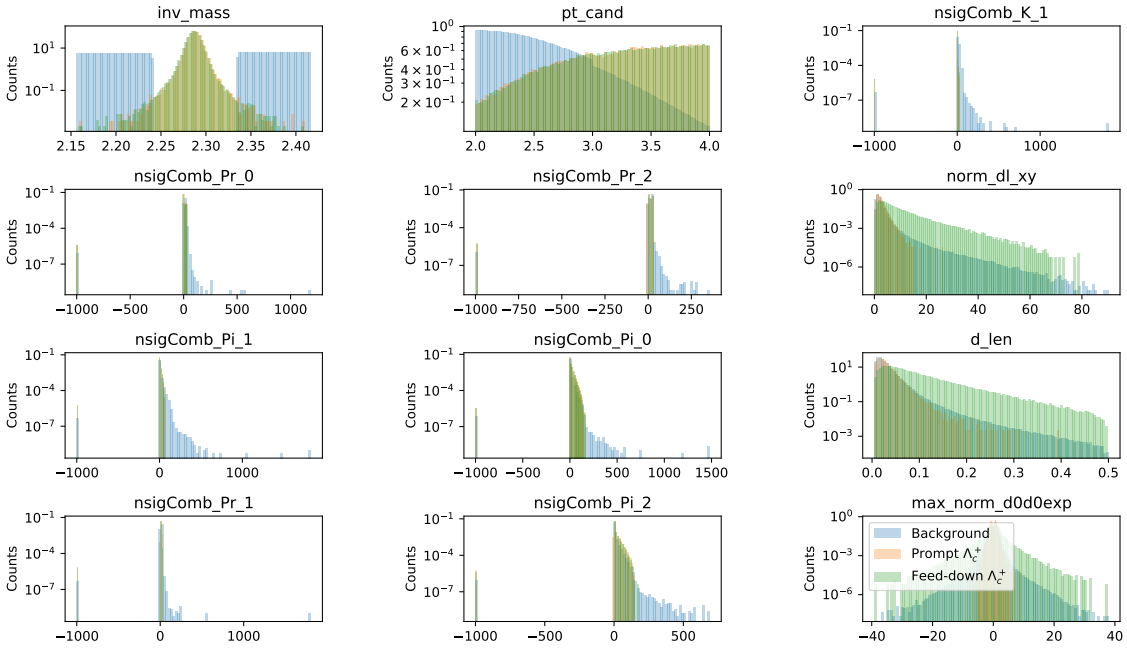

Figure 7.5: Distributions of the ten relevant features with additional parameters transverse momentum p_{T} and invariant mass m for $2 < p_{\mathrm{T}} < 4$ GeV/c.

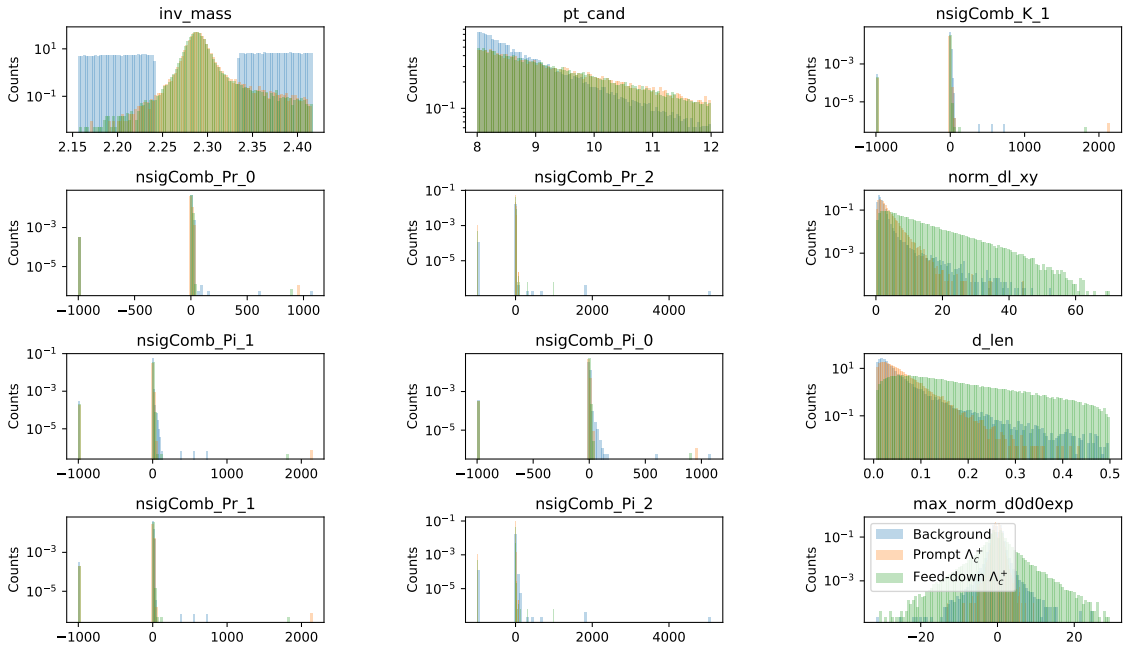

Figure 7.6: Distributions of the ten relevant features with additional parameters transverse momentum p_{T} and invariant mass m for $8 < p_{\mathrm{T}} < 12$ GeV/c.

7.4 ROC Curves

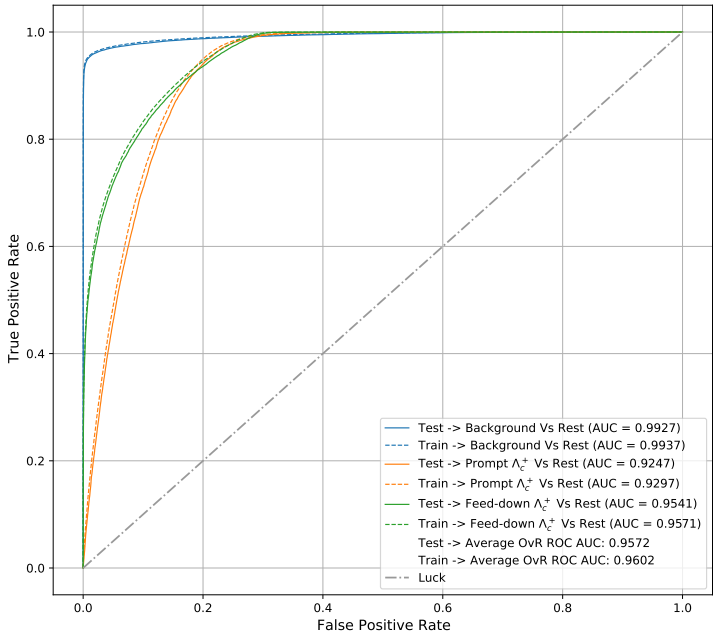

Figure 7.7: The ROC curves with the respective AUC values for the model for $2 < p_{\mathrm{T}} < 4$ GeV/c.

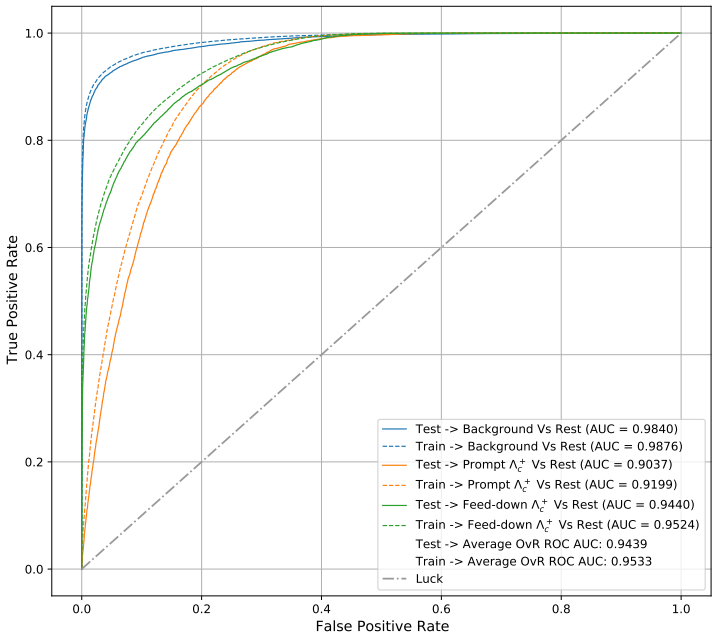

Figure 7.8: The ROC curves with the respective AUC values for the model for $8 < p_{\mathrm{T}} < 12$ GeV/c.

7.5 Machine Learning Outputs

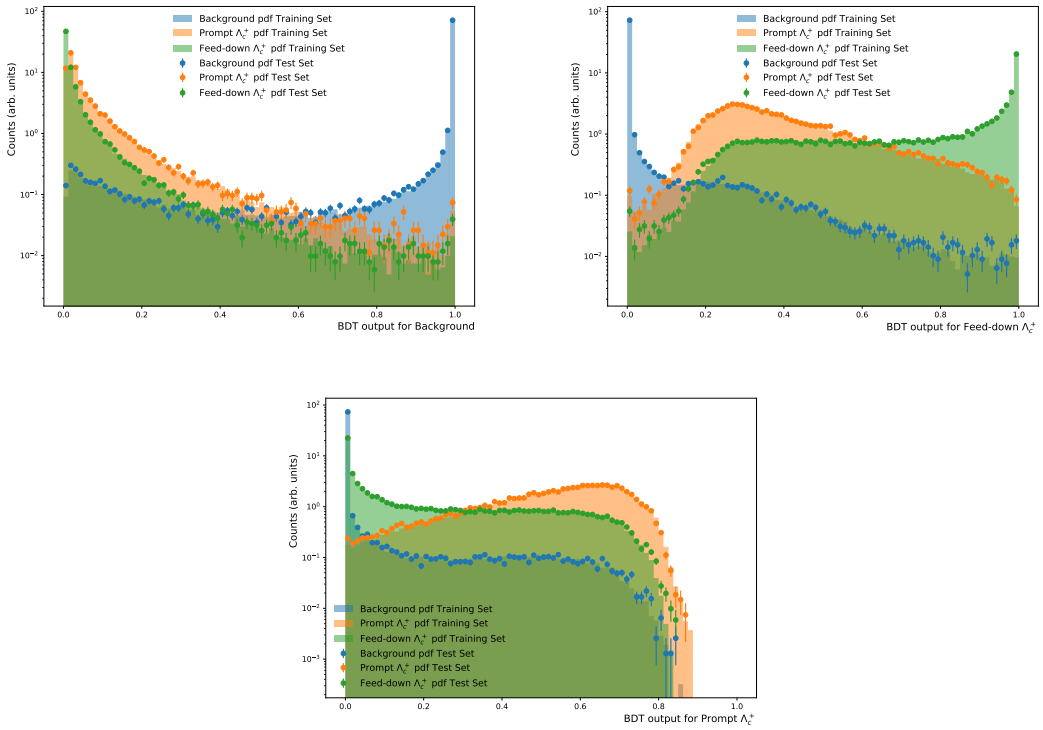

Figure 7.9: ML outputs for the training and test data for each class in $2 < p_{\mathrm{T}} < 4$ GeV/c.

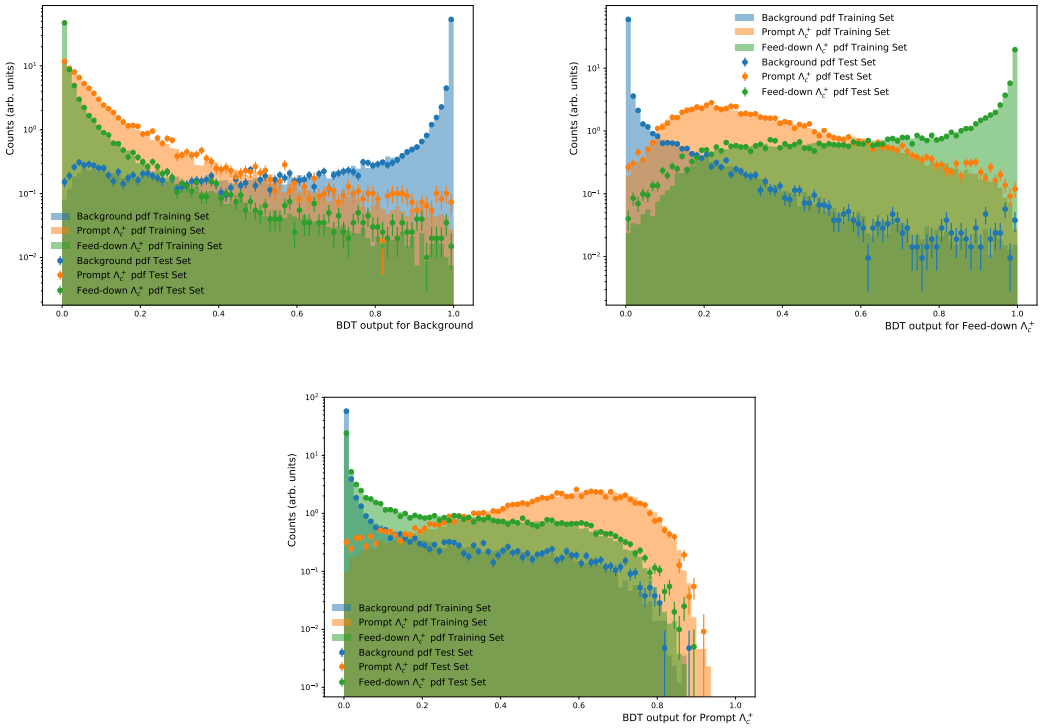

Figure 7.10: ML outputs for the training and test data for each class in $8 < p_{\mathrm{T}} < 12$ GeV/c.

7.6 Working Point Calculations

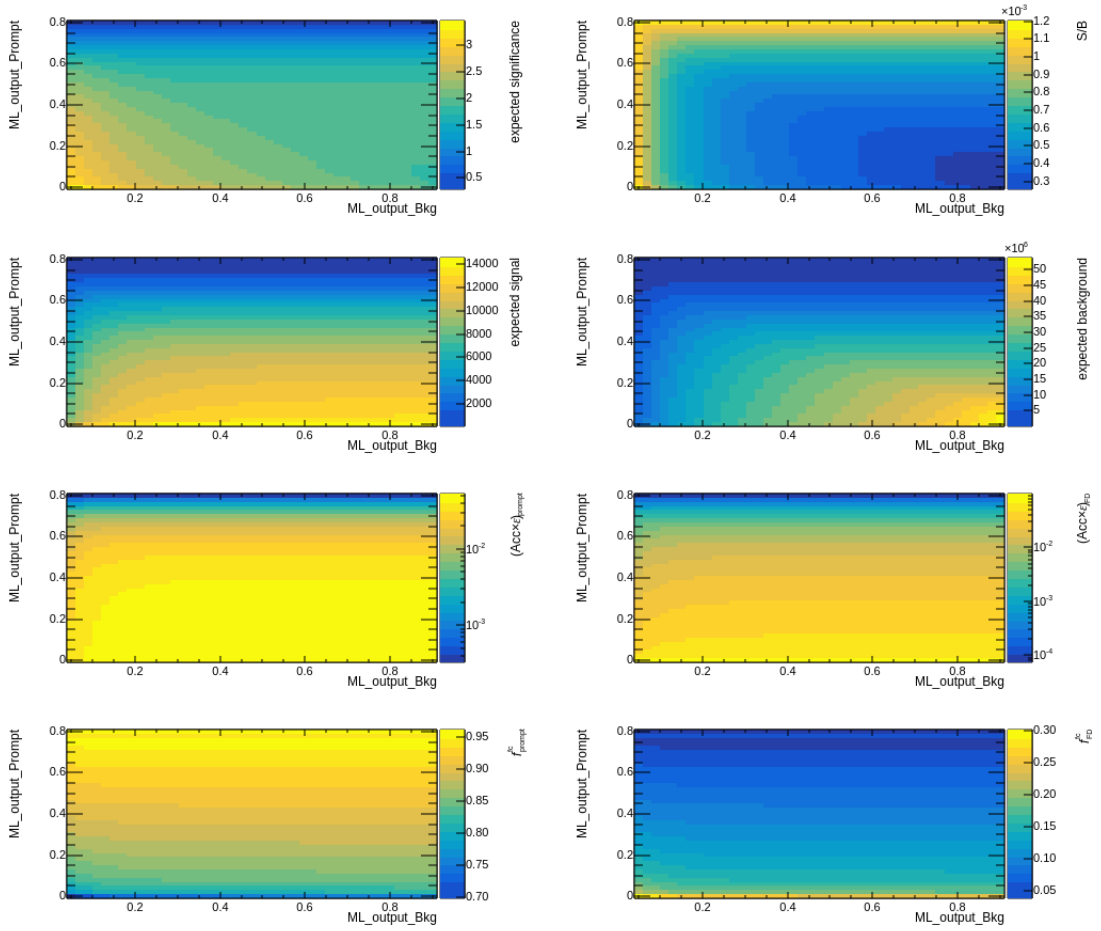

Figure 7.11: Working point calculations as functions of the ML output selections for prompt candidates in $2 < p_{\mathrm{T}} < 4$ GeV/c.

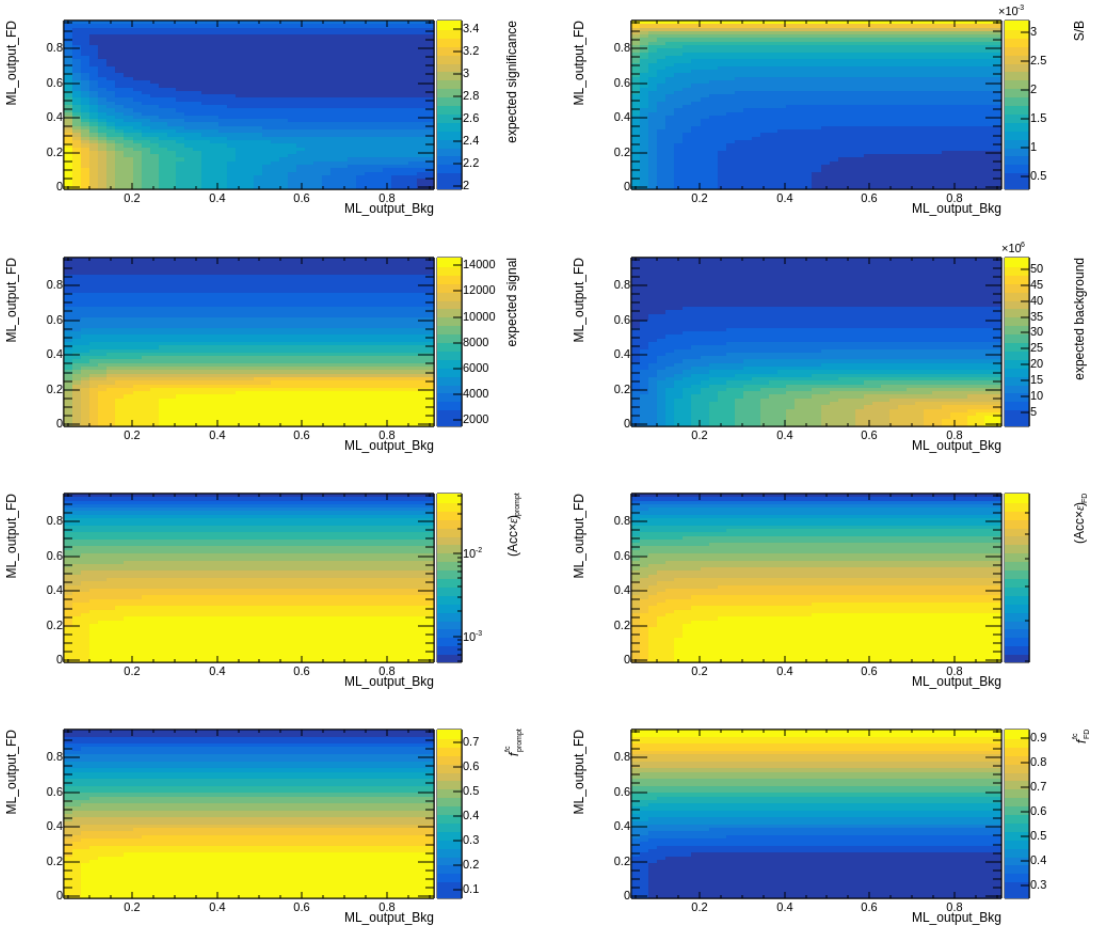

Figure 7.12: Working point calculations as functions of the ML output selections for NP candidates in $2 < p_{\mathrm{T}} < 4$ GeV/c.

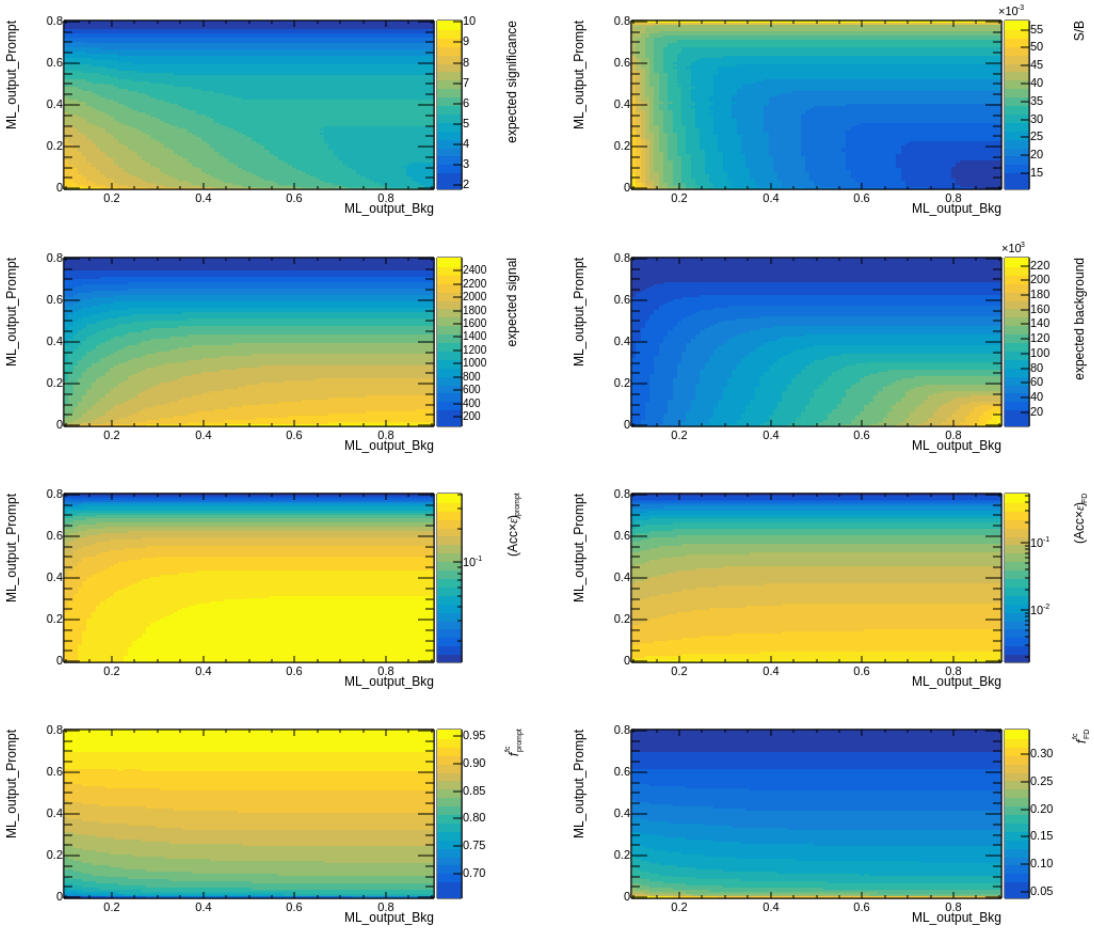

Figure 7.13: Working point calculations as functions of the ML output selections for prompt candidates in $8 < p_{\mathrm{T}} < 12$ GeV/c.

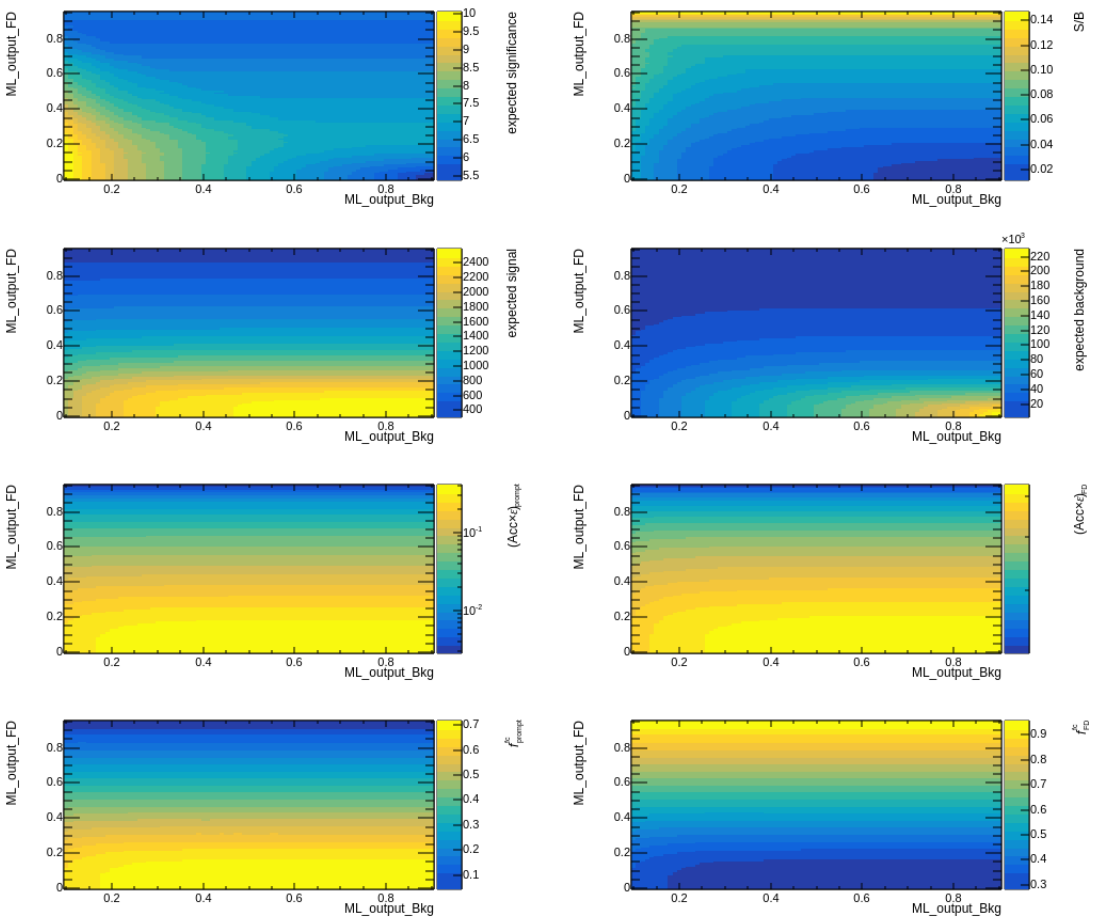

Figure 7.14: Working point calculations as functions of the ML output selections for NP candidates in $8 < p_{\mathrm{T}} < 12$ GeV/c.

List of Acronyms

ALICE	A Large Ion Collider Experiment
AUC	Area Under Curve
BDT	Boosted Decision Tree
CART	Classification and Regression Tree
CERN	Conseil Européen pour la Recherche Nucléaire (European Organization for Nuclear Research)
CNM	Cold Nuclear Matter
FD	Feed-Down
FONLL	Fixed-Order plus Next-to-Leading-Log
hipe4ml	Heavy-Ion Physics Environment for Machine Learning
IP2	Interaction Point 2
ITS	Inner Tracking System
LEP	Large Electron-Positron Collider
LHC	Large Hadron Collider
MC	Monte Carlo
ML	Machine Learning
MRPC	Multigap Resistive Plate Chambers
MWPC	Multi-Wire Proportional Chamber
NP	Non-Prompt
OvO	One-vs-One
OvR	One-vs-Rest
PID	Particle Identification
pQCD	perturbative Quantum Chromodynamics
PV	Primary Vertex
QCD	Quantum Chromodynamics
QGP	Quark–Gluon Plasma
ROC	Receiver Operating Characteristics
SDD	Silicon Drift Detectors
SHAP	Shapley Additive Explanations
SPD	Silicon Pixel Detectors
SSD	Silicon Strip Detectors

SV Secondary Vertex
TOF Time Of Flight
TPC Time Projection Chamber
XGBoost Extreme Gradient Boosting

Bibliography

[1] Wikiquote, *Democritus*. July 2022. URL: https://en.wikiquote.org/wiki/Democritus (visited on 12/23/2022).

[2] Sylvia Berryman, *Democritus*. Dec. 2016. URL: https://plato.stanford.edu/entries/democritus/ (visited on 12/23/2022).

[3] Johann Rafelski and Berndt Müller, *Strangeness Production in the Quark-Gluon Plasma*. Phys. Rev. Lett. **48** (16 Apr. 1982), 1066–1069. DOI: 10.1103/PhysRevLett.48.1066.

[4] R. L. Workman *et al.*, *Review of Particle Physics*. PTEP **2022** (2022), 083C01. DOI: 10.1093/ptep/ptac097.

[5] The ALICE Collaboration, *Λ_c^+ production in pp and in p-Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV*. Physical Review C **104**.5 (Nov. 2021). DOI: 10.1103/physrevc.104.054905.

[6] Annalena Kalteyer, *Reconstruction of Λ_c^+ in p-Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV with the ALICE detector*. Master's Thesis. Heidelberg, Germany: Physikalisches Institut of the University of Heidelberg, May 2021.

[7] Simon Groß-Bölting, *Measurement of the Λ_c^+ production in proton-proton collisions for $\Lambda_c^+ \rightarrow pK_s$ at $\sqrt{s} = 5.02$ TeV with the ALICE detector*. Bachelor's Thesis. Heidelberg, Germany: Physikalisches Institut of the University of Heidelberg, Nov. 2021.

[8] Mark Thomson, *Modern Particle Physics*. Cambridge University Press, 2013.

[9] V.G. Kartvelishvili, A.K. Likhoded, and V.A. Petrov, *On the fragmentation functions of heavy quarks into hadrons*. Physics Letters B **78**.5 (1978), 615–617. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(78)90653-6.

[10] ALICE Collaboration, *First measurement of Ξ_c^+ production down to $p_T = 0$ in pp and p-Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV*. 2022. DOI: 10.48550/ARXIV.2211.14032.

[11] The ALICE Collaboration, *Production cross section of non-prompt Λ_c^+ in p–Pb collisions at 5.02 TeV*. URL: https://alice-figure.web.cern.ch/node/21670 (visited on 03/03/2023).

[12] Domenico Colella, *ALICE ITS: The run 1 to run 2 transition and recent operational experience*. Proceedings of 24th International Workshop on Vertex Detectors — PoS(VERTEX2015) (June 2015). DOI: 10.22323/1.254.0003.

[13] The ALICE Collaboration, *Λ_c^+ production in pp and p–Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV*. Physical Review C **104**.5 (Nov. 2021). DOI: 10.1103/physrevc.104.054905.

[14] R. Vogt, *Cold Nuclear Matter Effects on Open and Hidden Heavy Flavor Production at the LHC*. 2015. DOI: 10.48550/ARXIV.1508.01286.

[15] The ALICE Collaboration, *First measurement of Λ_c^+ production down to $p_T = 0$ in pp and p-Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV*. 2022. DOI: 10.48550/ARXIV.2211.14032.

[16] Jun Song, Hai-hong Li, and Feng-lan Shao, *New feature of low p_T charm quark hadronization in pp collisions at $\sqrt{s} = 7$ TeV*. The European Physical Journal C **78**.4 (Apr. 2018). DOI: 10.1140/epjc/s10052-018-5817-x.

[17] K.J. Eskola, V.J. Kolhinen, and C.A. Salgado, *The scale dependent nuclear effects in parton distributions for practical applications*. The European Physical Journal C **9** (June 1999). DOI: 10.1007/s100529900005.

[18] B. Z. Kopeliovich *et al.*, *Cronin Effect in Hadron Production Off Nuclei*. Phys. Rev. Lett. **88** (23 May 2002), 232303. DOI: 10.1103/PhysRevLett.88.232303.

[19] Lyndon Evans and Philip Bryant, *LHC Machine*. Journal of Instrumentation **3**.08 (Aug. 2008). DOI: 10.1088/1748-0221/3/08/s08001.

[20] ALICE Collaboration, *Performance of the ALICE experiment at the CERN LHC*. International Journal of Modern Physics A **29**.24 (Sept. 2014), 1430044. DOI: 10.1142/s0217751x14300440.

[21] Elena Botta, *Particle identification performance at ALICE*. Tech. rep. 2017. arXiv: 1709.00288 [nucl-ex].

[22] Carolina Reetz, *Measurement of Ξ_c^+ in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ALICE detector*. Master's Thesis. Heidelberg, Germany: Physikalisches Institut of the University of Heidelberg, Aug. 2022.

[23] J. Alme *et al.*, *The ALICE TPC, a large 3-dimensional tracking device with fast readout for ultra-high multiplicity events*. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **622**.1 (Oct. 2010), 316–367. DOI: 10.1016/j.nima.2010.04.042.

[24] ALICE Collaboration, *Technical Design Report of the Inner Tracking System (ITS)*. Tech. rep. Geneva, Switzerland: CERN, June 1999.

[25] The ALICE Collaboration *et al.*, *The ALICE experiment at the CERN LHC*. Journal of Instrumentation **3**.08 (Aug. 2008), S08002. DOI: 10.1088/1748-0221/3/08/S08002.

[26] The ALICE Collaboration, *Production of light-flavor hadrons in pp collisions at $\sqrt{s} = 7$ TeV and $\sqrt{s} = 13$ TeV*. The European Physical Journal C **81**.3 (Mar. 2021). DOI: 10.1140/epjc/s10052-020-08690-5.

[27] Roberto Preghenella, *The Time-Of-Flight detector of ALICE at LHC: construction, test and commissioning with cosmic rays*. PhD thesis. Bologna, Italy: Università di Bologna, 2009.

[28] A. Akindinov *et al.*, *The MRPC detector for the ALICE Time Of Flight System: Final Design and Performances*. Nuclear Physics B - Proceedings Supplements **158** (Aug. 2006), 60–65. DOI: 10.1016/j.nuclphysbps.2006.07.035.

[29] The ALICE Collaboration, *TOF Beta vs Momentum performance in Run2 p-Pb at 5.02 TeV (LHC16q,t)*. 2018. URL: https://alice-figure.web.cern.ch/node/13313 (visited on 01/14/2023).

[30] Nicolò Jacazio, *PID performance of the ALICE-TOF detector at Run 2*. PoS **LHCP2018** (2018), 232. DOI: 10.22323/1.321.0232.

[31] Tianqi Chen and Carlos Guestrin, *XGBoost*. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, Aug. 2016. DOI: 10.1145/2939672.2939785.

[32] *hipe4ml*. 2022. URL: https://colab.research.google.com/github/hipe4ml/hipe4ml/blob/master/tutorials/hipe4ml_tutorial_multiclass.ipynb (visited on 01/03/2023).

[33] IBM, *Learn the pros and cons of using decision trees for data mining and knowledge discovery tasks*. URL: https://www.ibm.com/topics/decision-trees (visited on 01/03/2023).

[34] XGBoost Developers, *XGBoost Documentation*. 2022. URL: https://xgboost.readthedocs.io/en/stable/index.html (visited on 01/18/2023).

[35] Rukshan Pramoditha, *Introduction to Boosted Trees*. Oct. 2021. URL: https://towardsdatascience.com/introduction-to-boosted-trees-2692b6653b53 (visited on 01/03/2023).

[36] Gaurav, *An Introduction to Gradient Boosting Decision Trees*. June 2021. URL: https://www.machinelearningplus.com/machine-learning/an-introduction-to-gradient-boosting-decision-trees/ (visited on 01/04/2023).

[37] Anshul Saini, *Gradient Boosting Algorithm: A Complete Guide for Beginners*. Sept. 2021. URL: https://www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/ (visited on 01/05/2023).

[38] *Optuna - A Hyperparameter optimization framework*. 2023. URL: https://optuna.org/ (visited on 01/14/2023).

[39] Jason Brownlee, *One-vs-Rest and One-vs-One for Multi-Class Classification*. Apr. 2020. URL: https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/ (visited on 01/05/2023).

[40] Christoph Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Munich, Germany: Independently Published, 2022.

[41] The ALICE Collaboration, *Measurement of visible cross sections in proton-lead collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV in van der Meer scans with the ALICE detector*. Journal of Instrumentation 11 (Nov. 2014), P11003–P11003. DOI: 10.1088/1748-0221/9/11/p11003.

[42] Matteo Cacciari, Mario Greco, and Paolo Nason, *The p_{T} spectrum in heavy-flavour hadroproduction*. Journal of High Energy Physics **1998**.05 (May 1998), 007–007. DOI: 10.1088/1126-6708/1998/05/007.

[43] The ALICE Collaboration, *Nuclear modification factor R_{pPb} of non-prompt Λ_{c}^{+} in p–Pb collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV*. URL: https://alice-figure.web.cern.ch/node/21673 (visited on 03/03/2023).

Declaration of Authorship

Ich versichere, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Heidelberg, den 13.03.2023