Department of Physics and Astronomy
Heidelberg University

Bachelor Thesis in Physics
submitted by

Philip Quicker

born in Mannheim (Germany)

2025

Feasibility study of the prompt and non-prompt $J/\psi \to e^+e^-$ analysis in pp collisions at $\sqrt{s} = 13.6$ TeV with ALICE in Run 3

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

Abstract

The study of the production of charmonia, such as the J/ψ, is crucial to the investigation of the quark-gluon plasma and the hadronization processes occurring inside of it. In order to measure the various effects taking place within this exotic state of matter produced in heavy-ion collisions, it is of utmost importance to precisely reconstruct the decays of prompt and non-prompt J/ψ. The former are produced in the initial hard scattering of partons, while the latter originate from the decay of beauty hadrons, resulting in different interactions with the quark-gluon plasma, thus a separation of both types is essential. Naturally, proton-proton collisions should also be considered, serving as a point of reference for heavy-ion collisions.
In this thesis, the feasibility of the analysis of prompt and non-prompt J/ψ in proton-proton collisions at a center of mass energy of $\sqrt{s} = 13.6$ TeV with the ALICE detector in Run 3 via the $J/\psi \to e^+e^-$ decay channel is studied. The reconstructed candidates from Run 3 data collected in 2022 are split into four transverse momentum intervals between 0 GeV/c and 12 GeV/c. For the separation of the two types of J/ψ and background, Boosted Decision Trees (BDT), implemented using the XGBoost algorithm, are employed as so-called "multiclassifiers". Lastly, the yield of prompt and non-prompt J/ψ, as well as the associated significances are computed through fits of the signal peaks.
Significances between 110.01 and 179.47 are observed for the prompt signal, while lower significances ranging from 52.19 to 78.28 are determined for non-prompt J/ψ. Application of the non-prompt selections reveals an enhanced amount of entries for mass values lower than the J/ψ mass, indicating that the models are not fully capable of separating background and non-prompt signal. Thus, a full analysis of both prompt and non-prompt J/ψ can only be deemed feasible, if this residual background problem can ultimately be resolved.

Zusammenfassung

Die Messung der Produktion von Charmonia, wie dem J/ψ, spielen eine wichtige Rolle für die Untersuchung des Quark-Gluon-Plasmas und der Hadronisierungsprozesse, welche darin stattfinden. Um die verschiedenen Effekte, welche in diesem in Schwerionenkollisionen erzeugten Zustand der Materie stattfinden, zu messen, ist es von äußerst großer Bedeutung die Zerfälle von prompt und non-prompt J/ψ zu rekonstruieren. Erstere werden in harter Streuung von Partonen erzeugt, während letztere aus Zerfällen von Beauty-Hadronen stammen, was zu unterschiedlichen Interaktionen mit dem Quark-Gluon-Plasma führt, weshalb die Trennung der beide Typen essenziell ist. Selbstverständlich sollten Proton-Proton-Kollisionen ebenfalls betrachtet werden, da sie Vergleichswerte für Schwerionenkollisionen zur Verfügung stellen.
In dieser Arbeit wird die Machbarkeit einer Analyse von prompt und non-prompt J/ψ für Proton-Proton-Kollisionen bei einer Schwerpunktsenergie von $\sqrt{s} = 13.6$ TeV mit dem ALICE Detektor in Run 3 für den Zerfallskanal $J/\psi \to e^+e^-$ untersucht. Die rekonstruierten Kandidaten aus den in 2022 gesammelten Run 3 Daten werden in vier Intervalle des transversalen Impulses zwischen 0 GeV/c und 12 GeV/c aufgeteilt. Für die Trennung der beiden Typen von J/ψ und des Hintergrunds werden Boosted Decision Trees (BDT) eingesetzt, welche durch den XGBoost Algorithmus implementiert wurden und als sogenannte Multiclassifier eingesetzt werden. Abschließend werden die Erträge, sowie die zugehörigen Signifikanzen der Signal-Peaks für prompt und non-prompt J/ψ bestimmt.
Die berechnetten Signifikanzen für prompt J/ψ liegen zwischen 110.01 und 179.47, während für non-prompt J/ψ niedrigere Signifikanzen im Bereich zwischen 52.19 und 78.28 vorliegen. Die Anwendung der BDT-Selektion für non-prompt J/ψ zeigt, dass eine ungewöhnlich hohe Menge von Einträgen für Massen unterhalb der J/ψ-Masse vorliegt, was nahelegt, dass die trainierten Modelle nich in der Lage sind, Hintergrund und non-prompt Signal vollständig zu trennen. Somit kann eine volle Analyse von prompt und non-prompt J/ψ nur als machbar bezeichnet werden, wenn dieses Problem des übrig bleibenden Hintergrundes endgültig gelöst werden kann.

Contents

1 Introduction

1.1 Motivation

The breakthroughs of high-energy physics have fundamentally changed our understanding of nature by furthering the comprehension of the composition of matter and the interactions that govern it. Research of high-energy hadronic collisions has made it possible to understand strongly interacting matter and establish the associated theory of Quantum Chromodynamics (QCD). Moreover, collisions of heavy ions allow us to explore the quark-gluon plasma (QGP), an exotic state of matter characterized by extremely high temperatures and pressures, which is thought to have been present shortly after the birth of our universe. One of the experiments dedicated to the research of QCD and the QGP is the ALICE experiment, which is one of the four major experiments at the Large Hadron Collider (LHC) at CERN [1].

Measuring the production of heavy-flavor hadrons, which contain charm (c) and beauty (b) quarks, in proton-proton (pp) collisions serves as a test of perturbative QCD, while also providing an important reference to production measurements in heavy-ion collisions, which is essential for the research of the QGP in heavy-ion collisions. Heavy-flavor hadrons are suitable probes of this rare state of matter, since charm (c) and beauty (b) quarks are exclusively produced in the initial hard scattering of the collision due to their large masses. Therefore, these quarks experience the evolution of the QGP from start to finish, thus being subject to momentum and energy exchange with the medium. Additional effects observed in heavy-ion collisions are the suppression and regeneration of so-called "quarkonia". The J/ψ meson, which is part of this class of particles, is of particular interest for the research of the QGP. It not only experiences these effects itself, but it is also the decay product of heavier hadrons which experience the aforementioned interactions with the medium. Therefore, precise measurements of the production of J/ψ in pp collisions are of extraordinary importance, in order to fully understand the evolution of the QGP in heavy-ion collisions, as well as its effect on the production of J/ψ.

In order to reach unprecedented precision and to consequently further the current understanding of the QGP, the ALICE experiment has undergone major upgrades during the Long Shutdown 2 (LS2). Using Run 3 data taken with the upgraded detector, this thesis explores the potential of machine learning methods as a tool to reconstruct prompt and non-prompt $J/\psi \to e^+e^-$ decays in pp collisions.

1.2 The Standard Model of particle physics

The Standard Model (SM) forms the fundament of the field of particle physics and describes the fundamental constituents of the universe: namely the elementary particles and the interactions between them. While the Standard Model describes all current experimental data remarkably well, it does not account for all effects, as seen for example with neutrino oscillations. It should be noted that the Standard Model includes the electromagnetic interaction, the weak interaction and the strong interaction, but not gravity, thus it only accounts for three of the four fundamental interactions [2].

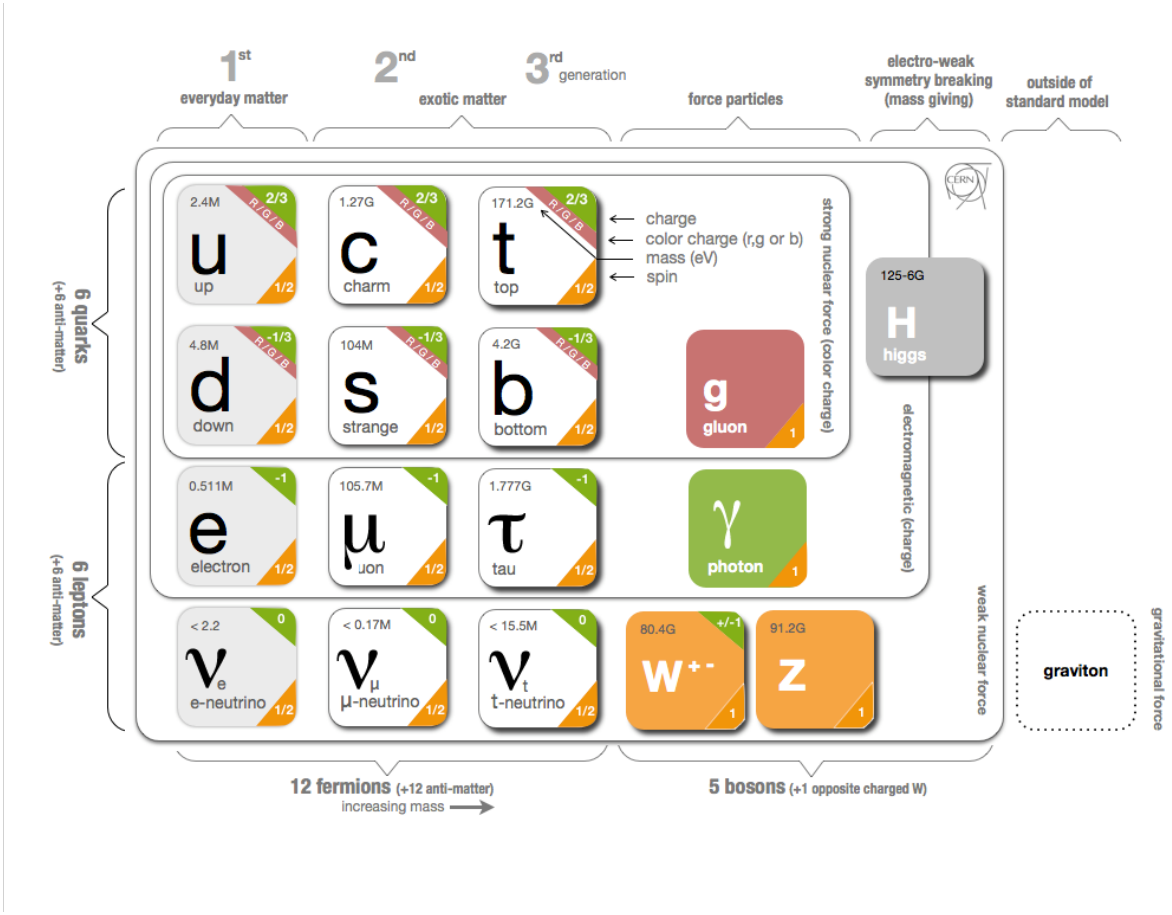

Figure 1: The elementary particles and bosons of the Standard Model of particle physics [3].

Matter in our universe is made up of twelve fundamental spin $\frac{1}{2}$ particles (fermions), of which six are quarks and six are leptons, with both types divided into three generations (see fig. 1). The quarks have a charge of either $\frac{2}{3}e$ (u, c, t) or $-\frac{1}{3}e$ (d, s, b), carry color charge and differ in mass and flavor, while interacting strongly, electromagnetically and weakly. Due to their ability to interact via the strong interaction, they form color-neutral bound states called hadrons (mesons and baryons). The top quark is the only exception, since its lifetime is shorter than the hadronization time. Leptons on the other hand can be divided into charged and uncharged leptons (neutrinos). Electrons (e^-), muons (μ^-) and taus (τ^-) interact electromagnetically and weakly, whereas their respective neutrinos (ν_e, ν_μ, ν_τ) only interact weakly, since they have

no electrical charge. Each fermion has an antiparticle, which is charge conjugated form of the particle. The three fundamental interactions described by the SM are mediated by the gauge bosons, which are spin 1 particles. Strong interactions are mediated by the eight gluons (g) carrying color charge, while photons (γ) mediate the electromagnetic interaction and the $W^{\pm}$ and Z bosons mediate the weak interaction. Lastly, the Higgs boson (H), a scalar boson with spin 0, gives rise to the masses of the quarks, charged leptons and the mediators of the weak interaction via the Higgs mechanism. With the discovery of the Higgs boson by the ATLAS and CMS experiments at the Large Hadron Collider (LHC) the SM was completed in 2012, validating the theoretical concepts of the SM [2].

1.3 The Large Hadron Collider and heavy-ion collisions

The Large Hadron Collider (LHC) is a synchrotron with a circumference of 26.7 km, making it the world's largest particle accelerator. It is part of the CERN accelerator complex (see fig. 2), of which large parts are situated on Swiss territory, while the LHC itself and all experiments except ATLAS also stretch far into French territory. After the Large Electron-Positron collider was dismantled when operation was finished, the underground circular tunnel at a depth of roughly 100 m, which was housing it, was repurposed for the LHC. The LHC is capable of accelerating bunches of protons as well as heavy ions and colliding them at four interaction points along the beam line, where the four main experiments ATLAS, CMS, LHCb and ALICE are located. While ATLAS and CMS are general purpose detectors, which are specialized for the discovery the Higgs boson, LHCb is dedicated to beauty physics and ALICE was constructed to investigate the QGP. For proton-proton (pp) collisions in Run 3 the LHC can achieve a center of mass energy of $\sqrt{s} = 13.6$ TeV after the upgrades made during the Long Shutdown 2, whereas center of mass energies per nucleon pair of $\sqrt{s_{NN}} = 5.36$ TeV can be reached in lead-lead (Pb-Pb). However, a series of particle accelerators (see fig. 2) is first required to produce the beams of protons or heavy ions at lower energies, from which the LHC can accelerate them to their final energies. Proton beams are produced from negative negative hydrogen ions (H^-), which are stripped of their electrons after the first steps of acceleration. Lead ions, on the other hand, originate from evaporated lead that is injected into a plasma chamber, where the lead atoms are ionized. They then are accelerated and further stripped of the remaining electrons to create Pb^{82+} ions, which are then injected into the LHC as a beam with up to 1248 bunches [4, 5, 6].

When ultra-relativistic heavy ions collide, a state of matter called quark-gluon plasma (QGP) is formed, which is researched by the ALICE experiment. It is a phase of matter, where the density and temperature of the medium are very high and the strongly interacting gluons and quarks are deconfined and therefore not bound in hadrons. Heavy ions take the form of Lorentz contracted discs at ultra-relativistic velocities, which are present at LHC energies. These discs consist of many nucleons and carry large amounts of energy, leading to an enormous number of quarks, gluons and sea quarks contained by them. Due to the high concentration of quarks, gluons and energy during the collision, the conditions for the QGP to form are fulfilled. From the initial state, the QGP is formed and after a short period of evolution as an ideal relativistic

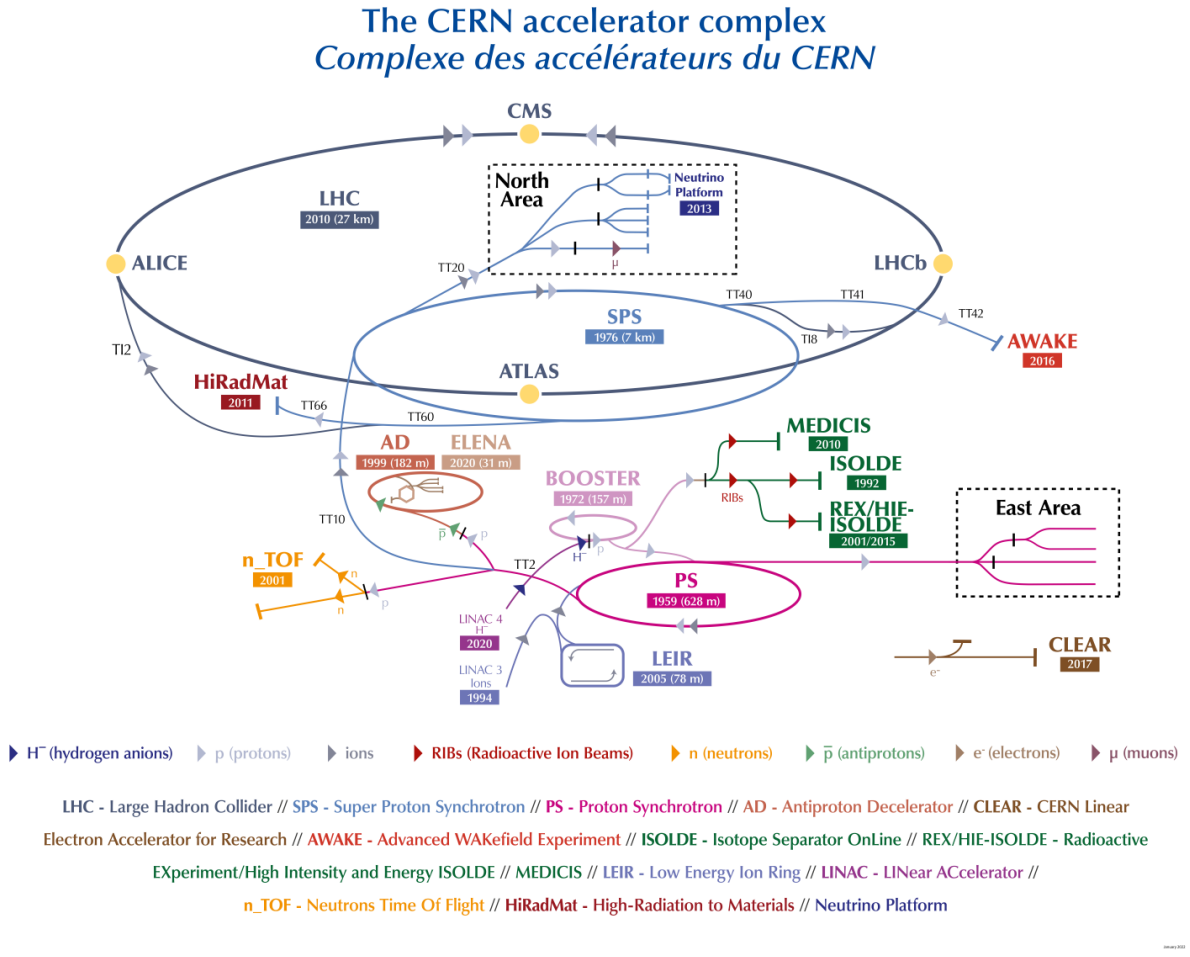

Figure 2: The CERN accelerator complex [7].

fluid, hadronization occurs inside the medium (see fig. 3). Lastly the freeze-out of the hadrons takes place, from which hadrons with high momenta remain, that can then be detected inside the ALICE detector. The quark-gluon plasma is a topic of interest, since it is believed that the QGP was formed very shortly after the Big Bang, after which the matter cooled down and hadronized. On another note, heavy ions are required in order to study the QGP with a hadron collider, although there are some indications that it might be formed in pp collisions as well [1].

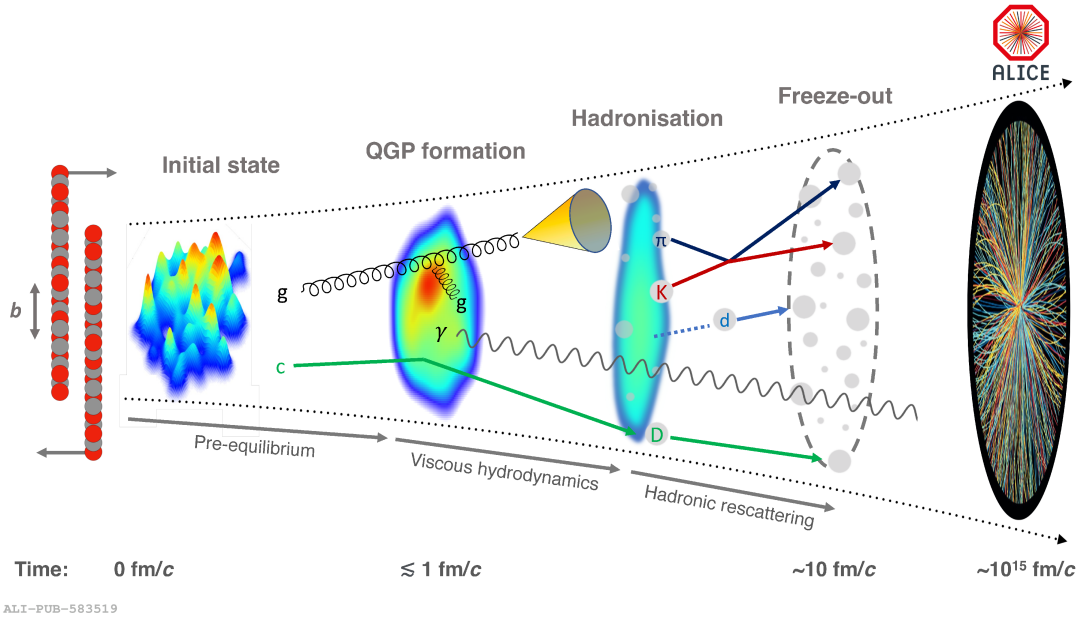

Figure 3: The evolution of a heavy-ion collision at LHC energies [1].

1.4 Prompt and non-prompt J/ψ

One particle of notable interest in Pb-Pb collisions and the research of the QGP is the J/ψ meson. It is an electrically uncharged vector meson consisting of a charm quark and a charm antiquark ($c\bar{c}$), which is categorized as "quarkonium" and more specifically "charmonium". When $c\bar{c}$ pairs are created in the initial collision of two lead nuclei, they are likely to separate inside the QGP due to the high temperatures of the QGP that give rise to so-called "color-charge screening". Due to the high abundance of quarks in the QGP, it is more likely for the charm quarks to recombine with lighter quarks than with anticharm quarks. This suppression effect of the production rate of the J/ψ has been observed at SPS and RHIC. However, the LHC operates at larger center of mass energies of $\sqrt{s_{NN}} = 5.36$ TeV per nucleon pair which leads to a regeneration effect through statistical recombination, that counteracts the suppression effect. The increase in the production rate of J/ψ at higher energies, is due to many more $c\bar{c}$ pairs being produced initially and therefore the probability is higher that different c quarks and $\bar{c}$ quarks recombine after some scattering inside the QGP. Naturally, these effects also take place for other quarkonia, which consist of $c\bar{c}$ or $b\bar{b}$ pairs. Therefore, one way to study the suppression and regeneration effects as properties of the QGP is to measure the production rates of these quarkonia [1].

When measuring J/ψ, a distinction between prompt J/ψ and non-prompt J/ψ has to be

made. Prompt J/ψ originate from the initial hard scattering of the collision or the regeneration process, while non-prompt J/ψ are decay products of heavier B hadrons like B mesons (see fig. 4). These heavier hadrons are only produced in the initial hard scattering, thus they travel through the QGP and are subject to energy and momentum exchange with the medium during it's entire evolution. Therefore prompt J/ψ and non-prompt J/ψ as decay daughters are well suited candidates to probe the QGP. In order to understand the J/ψ in heavy-ion collisions, the production of J/ψ in pp collisions as a reference is studied. The latest measurements of the differential cross sections of prompt and non-prompt J/ψ via the $J/\psi \to e^+e^-$ decay channel (see fig. 5) are shown in fig. 6. Measurements of the cross sections might be improved in future analysis by employing Boosted Decision Trees (BDT) in order to differentiate between prompt J/ψ and non-prompt J/ψ. The reconstruction of prompt and non-prompt J/ψ in the $J/\psi \to e^+e^-$ decay channel in pp collisions with the ALICE detector through the use of Boosted Decision Trees is carried out in this thesis. [1].

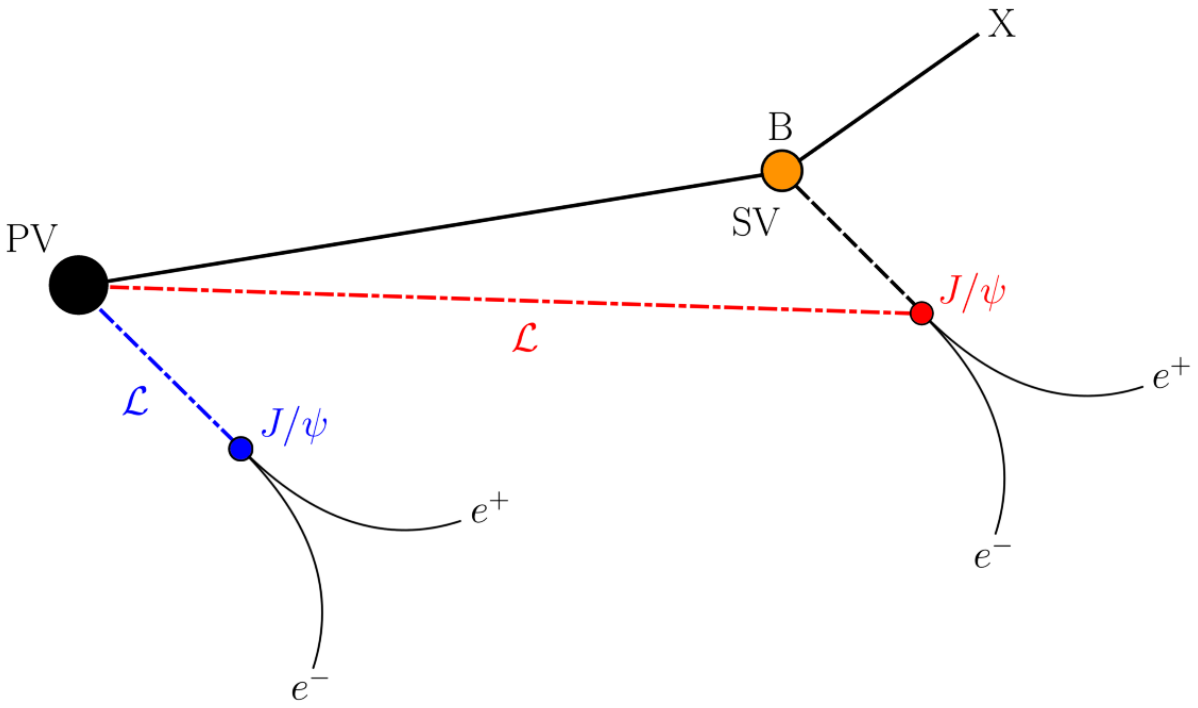

Figure 4: Visualization of the decay topology of prompt J/ψ and non-prompt J/ψ. PV stands for primary vertex, where particles produced in the collision originate from. At a SV (secondary vertex) particles decay and produce multiple lighter particles. Prompt and non-prompt J/ψ do not necessarily originate from the same PV, but for illustrative purposes they do in this figure. The decay product X depends on the type of beauty (B) hadron that decays into the non-prompt J/ψ. $\mathcal{L}$ is the decay length of the J/ψ, which is an important quantity for the differentiation between prompt and non-prompt J/ψ.

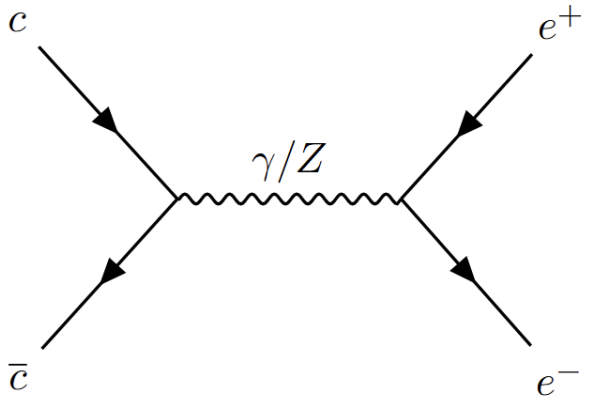

Figure 5: Feynman diagram for the $J/\psi \to e^+e^-$ decay.

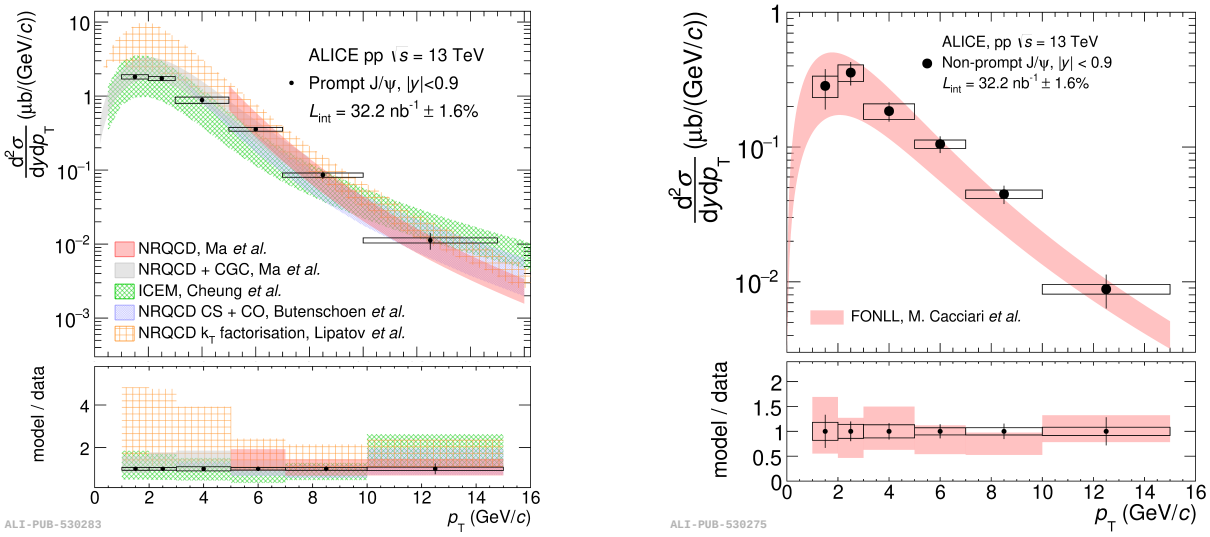

Figure 6: Measurements of the p_T-differential cross section of prompt J/ψ (left) and non-prompt J/ψ (right) at midrapidity ($|y| < 0.9$) in pp collisions at $\sqrt{s} = 13$ TeV through the dielectron decay channel. Additionally, predictions of the differential cross section from model calculations are shown [8].

2 The ALICE experiment

2.1 Overview of the ALICE detector

ALICE (A Large Ion Collider Experiment) is a heavy-ion detector at the Large Hadron Collider (LHC) at CERN, with a focus on the strong interaction of matter, as well as the quark-gluon plasma. It allows the extensive study of hadrons, electrons, muons and photons produced in heavy-ion collisions, particularly Pb-Pb collisions. Since the detector was upgraded during the Long Shutdown 2 and exclusively Run 3 data is used in this thesis, the description will focus on the Run 3 version of the detector, ALICE 2 (fig. 7).

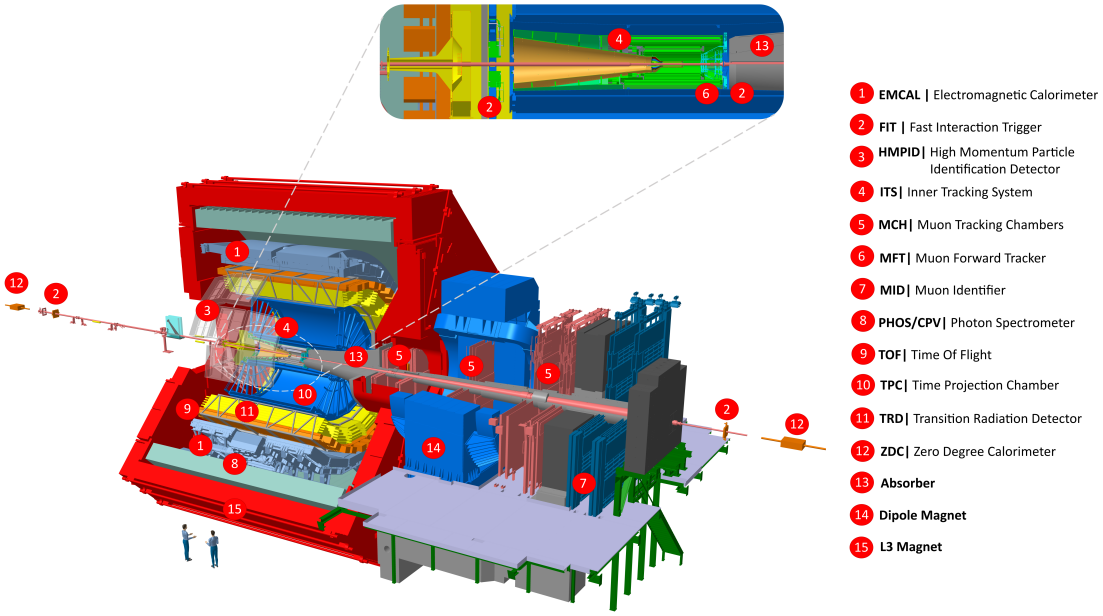

Figure 7: Schematic view of the ALICE detector during Run 3 [9].

The dimensions of ALICE are $16 \times 16 \times 26$ m^3 with a total weight of approximately 10000 t. It is comprised of a central barrel part, which measures hadrons, electrons and photons, as well as a forward muon spectrometer. A 0.5 T magnetic field is provided by a large solenoid magnet surrounding the central part, which was repurposed from the L3 experiment at LEP

The innermost system of the central barrel is the upgraded Inner Tracking System (ITS2), which consists of seven layers of ALPIDE (ALICE Pixel Detector) monolithic active pixel sensors (MAPS). Its main function is the reconstruction of particle trajectories, as well as the primary vertex (PV) and secondary vertices (SV) of heavy-flavor and strange particle decays with a high spatial and momentum resolution. Furthermore it is used to improve the resolution for particles reconstructed by the Time Projection Chamber (TPC)

Moving outwards, the ITS is followed by the Time Projection Chamber (TPC), which extends from 0.85 m to 2.5 m in radial direction with a length of 5 m along the beam. It measures the mean energy loss of particles $\frac{dE}{dx}$ via ionization of the gas in the TPC and can be used for tracking and particle identification. Since the ITS and TPC are especially important for the reconstruction of J/ψ decays, their structure and functionality will be further elaborated upon in section 2.2 and section 2.3

The subsequent Transition Radiation Detector (TRD) provides electron identification in the central barrel. It consists of six layers of gas chambers, each containing a foam/fibre radiator and a Xe-CO_2 gas mixture. Above 1 GeV/c the transition radiation from electrons passing through a radiator combined with the information from the specific energy loss in the TPC can be utilized to differentiate electrons from pions. Below this momentum threshold, the specific energy loss measurement in the TPC suffices to identify electrons

The TRD is followed by the Time-of-Flight detector (TOF); a large array of Multi-gap Resistive-Plate Chamber (MRPC) detectors and which further enables the identification of hadrons over a wide momentum range and electrons at low momentum. However, not all detectors cover the full azimuthal range and one of them is the ElectroMagnetic Calorimeter (EMCal), which is comprised of alternating layers of lead and scintillators. Unlike the EMCal, the PHOton Spectrometer (PHOS) only covers a small range of the acceptance in the central barrel. It is a high-resolution high-granularity electromagnetic calorimeter specialized for the detection of photons and consists of scintillating lead tungstate ($PbWO_4$) crystals with avalanche photodiode (APD) photodetectors and preamplifiers. Another detector that not spanning the whole azimuthal range is the High Momentum Particle Identification Detector (HMPID), which is a ring-imaging Cherenkov detector with liquid perfluorohexane (C_6F_{14}) radiators and adds hadron identification capabilities at large transverse momenta that can not be provided by the energy-loss measurement in the TPC

Additionally, the forward pseudorapidity range $-4.0 < \eta < -2.5$ is covered by muon detectors, which identify muons and remove hadrons by utilizing a system of absorbers. Multiwire proportional chambers (muon tracking chambers, MCH) and resistive plate chambers (muon identifier, MID) are used for the main muon detector stations. In ALICE 2, the Muon Forward Tracker (MFT), was added. It is comprised of tracking stations with ALPIDE silicon pixel sensors that are installed in front of the muon absorber to increase pointing resolution and mass resolution for the detection of secondary charmonia and muons from B-meson decays. Additionally, a number of trigger systems, like the Fast Interaction Trigger (FIT), and other detectors, like the Zero-Degree Calorimeters (ZDC), are utilized. For the online and offline reconstruction and the physics analysis in Run 3 the new common software framework O^2 was developed [9, 10, 11].

The reference coordinate system used in ALICE is a right handed system, where the x-axis points horizontally towards the center of the LHC, the y-axis points vertically upwards and the z-axis points along the beam line, away from the muon arm, with the origin of the coordinate system being the nominal interaction point [9]. A spherical coordinate system is also often used, where the azimuthal angle ϕ lies in the x-y-plane, the polar angle θ lies in the y-z-plane and the

r-axis points in radial direction. To describe the acceptance for a portion of the detector, most of the time the pseudorapidity η is used instead of the polar angle θ. The pseudorapidity η is defined as follows:

$$\eta = -\ln \tan \frac{\theta}{2}. \tag{1}$$

2.2 Inner Tracking System

The Inner Tracking System (ITS) constitutes the innermost detector layers in the central barrel of the ALICE detector. Its main purpose is the reconstruction of particle trajectories, the primary vertex (PV) and secondary vertices (SV) of heavy-flavor and strange particle decays with a high spatial and momentum resolution for tracks and high spatial resolution for vertices. Additionally, it can provide improvements to the momentum and angle resolution of the reconstruction of particles by the TPC. Therefore, the ITS contributes to, in principle, all physics topics that the ALICE experiment addresses [9, 10].

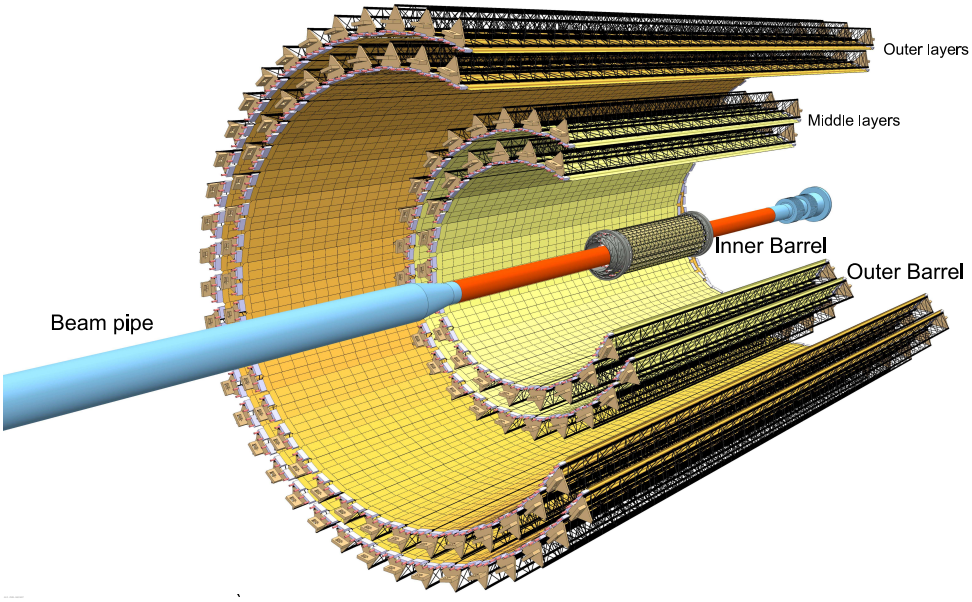

Figure 8: Schematic layout of the ITS2 [9].

During the LS2, substantial upgrades were made to the ALICE detector, most notably the original ITS was replaced by the new Inner Tracking System (ITS2). Due to its reduced distance to the interaction point, which was made possible by new beam pipe, and better position resolution compared to the first ITS, it provides better pointing resolution, while also enabling it to handle a higher interaction rate of 50 kHz for high hit densities in Pb-Pb collisions and 1 MHz in pp collisions. The ITS2, as seen in fig. 8, consists of seven layers of ALPIDE (ALICE Pixel Detector) Monolithic Active Pixel Sensors (MAPS), making it the largest-scale application of these sensors in any high-energy physics experiment. It is structured into the inner barrel (IB), which is made of the three innermost layers, and the outer barrel OB, which consists of two double layers. Each radial position of the layers, which can be found in table 1, was optimized

in order to achieve best performance regarding p_T resolution, pointing resolution and tracking efficiency for Pb-Pb collisions with their high track-density environment. In total the sensors cover a surface area of around 10 m^2 and 12.5 billion pixels with digital readout. The ITS covers the pseudorapidity range of $|\eta| < 1.22$ for the region where 90% of collisions take place, which translates to interaction vertices located in the range of approximately ± 10 cm around the nominal interaction point along the beam axis [9, 12].

Table 1: Main layout parameters of the new ITS2. A HIC (Hybrid Integrated Circuit) is an assembly of polyimide Flexible Printed Circuit on which pixel chips and some passive components are bounded. A stave is the basic detector unit, on which the pixel detectors, electronics and cooling are mounted [9].

Layer no.	Average radius (mm)	Stave length (mm)	No. of staves	No. of HICs/stave	Total no. of chips
0	23	271	12	1	108
1	31	271	16	1	144
2	39	271	20	1	180
3	196	844	24	8	2688
4	245	844	30	8	3360
5	344	1478	42	14	8232
6	393	1478	48	14	9408

The ITS2 encloses the new beam pipe with a central beryllium section, where the outer radius was reduced from 28 mm to 18 mm compared to Run 1 and Run 2. Furthermore, the innermost detector layer was moved closer, from 39 mm to 22.4 mm to interaction point and the material budget was reduced to $0.36\%X_0$ per layer for the innermost layers and limited to $1.10\%X_0$ per layer for the outer layers. Most importantly, the pixel size of the silicon pixel sensors was reduced to $29.24\mu m \times 26.88\mu m$, while the number of layers in the inner barrel was increased from two to three. A comparison of the main detector parameters of the ITS1 and ITS2 can be found in table 2 [9, 12].

Table 2: Comparison of the main detector parameters of the ITS1 and ITS2 [9].

	ITS1	**ITS2**
Technology	Hybrid pixel, strip, drift	MAPS
No. of Layers	6	7
Radius	$39 - 430$ mm	$22 - 395$ mm
Rapidity coverage	$\|\eta\| \leq 0.9$	$\|\eta\| \leq 1.3$
Material budget/layer	$1.14\%X_0$	inner barrel: $0.36\%X_0$ outer barrel: $1.10\%X_0$
Pixel size	425 μm $\times$ 50 μm (only the two innermost layers)	27 μm $\times$ 29 μm (all seven layers)
Spatial resolution ($r\varphi \times z$)	12 μm $\times$ 100 μm	5 μm $\times$ 5 μm
Readout	Analogue (drift, strip), Digital (Pixel)	Digital
Max rate (Pb-Pb)	1 kHz	50 kHz

2.3 Time Projection Chamber

The Time Projection Chamber (TPC) represents the main tracking detector of the central barrel. It is optimized to provide, with additional information from other detectors, charged-particle momentum measurements with good separation of tracks and particle identification over a wide momentum range. This is done through the measurement of the mean energy loss $\frac{dE}{dx}$ by ionization of the gas in the TPC and the momentum of each charged particle traversing the detector gas [9, 10, 13].

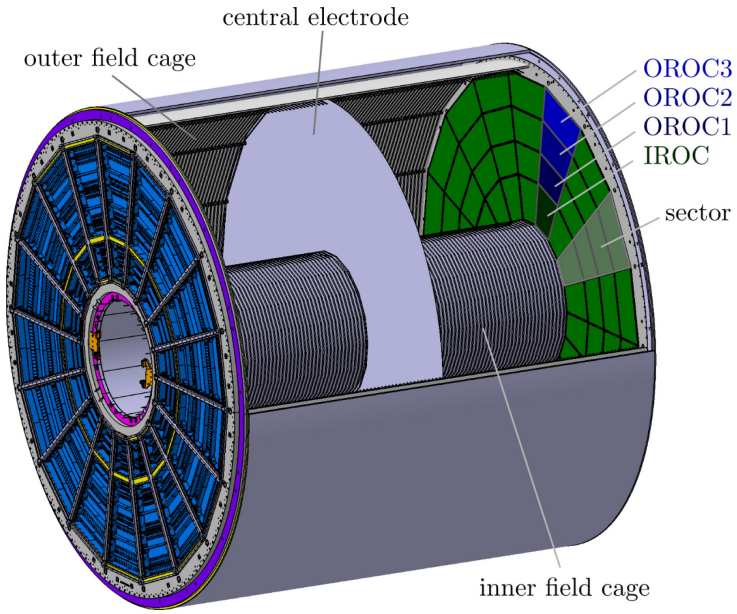

Figure 9: Schematic view of the TPC [14].

The layout of the TPC is visualized in fig. 9. It is cylindrical and ranges from 0.85m to 2.5m in radial direction over a length of 5m, giving it a total active volume of 88 m^3. Thus, it covers a symmetric pseudorapidity interval around midrapidity ($|\eta| < 0.9$) covering the full azimuth. The detector's field cage has a high-voltage electrode in the center that divides the active volume into halves and causes the free electrons created by ionization of the gas to drift towards the endplates. Both endplates are subdivided into 18 azimuthal sectors, which each house one inner (IROC) and one outer readout chamber (OROC). The TPC contains a gas mixture of Ne-CO_2-N_2 (90-10-5), which is advantageous since Neon offers a higher ion mobility compared to Argon-based mixtures, reducing the magnitude of space-charge distortions by a factor of almost two [9, 10].

In Run 1 and Run 2 the readout chambers of the TPC used multiwire proportional chambers (MWPC), which required active ion gating to minimize space-charge distortions. Since this requires triggered readout, which is not compatible with the goals of the upgraded ALICE detector, the new upgraded TPC has to be read out continuously. Simultaneously, the excellent performance has to be maintained and the space-charge distortions have to be kept at a tolerable level, despite the high collision rate and the missing active ion gate. For this purpose, Gas Electron Multipliers (GEMs) have been installed in the upgraded TPC, as they can be arranged in stacks to create layers of amplification stages that can be tuned and suppress the ion backflow

to the required level by blocking the path of back-drifting ions [9].

Using the measurements of specific energy loss, momentum and charge of particles, it is possible to identify the particles (PID) by utilizing the parameterized Bethe-Bloch formula

$$f(\beta\gamma) = \frac{P_1}{\beta^{P_4}} \left(P_2 - \beta^{P_4} - \ln\left(P_3 + \frac{1}{(\beta\gamma)^{P_5}} \right) \right) \tag{2}$$

where β is the particle velocity relative to the speed of light, γ is the Lorentz factor and P_1-P_5 are fit parameters. A visualization of the measured energy loss versus particle momentum and corresponding parameterized Bethe-Bloch fits for different particle species can be seen in fig. 10. At low momenta ($p \lesssim 1$ GeV/c), it is mostly possible to identify the particles on a track-by-track basis, whereas statistical methods like multi-Gaussian fits have to be applied in order to separate particles at higher momenta. In any case, identification also relies on PID information from other detectors [13].

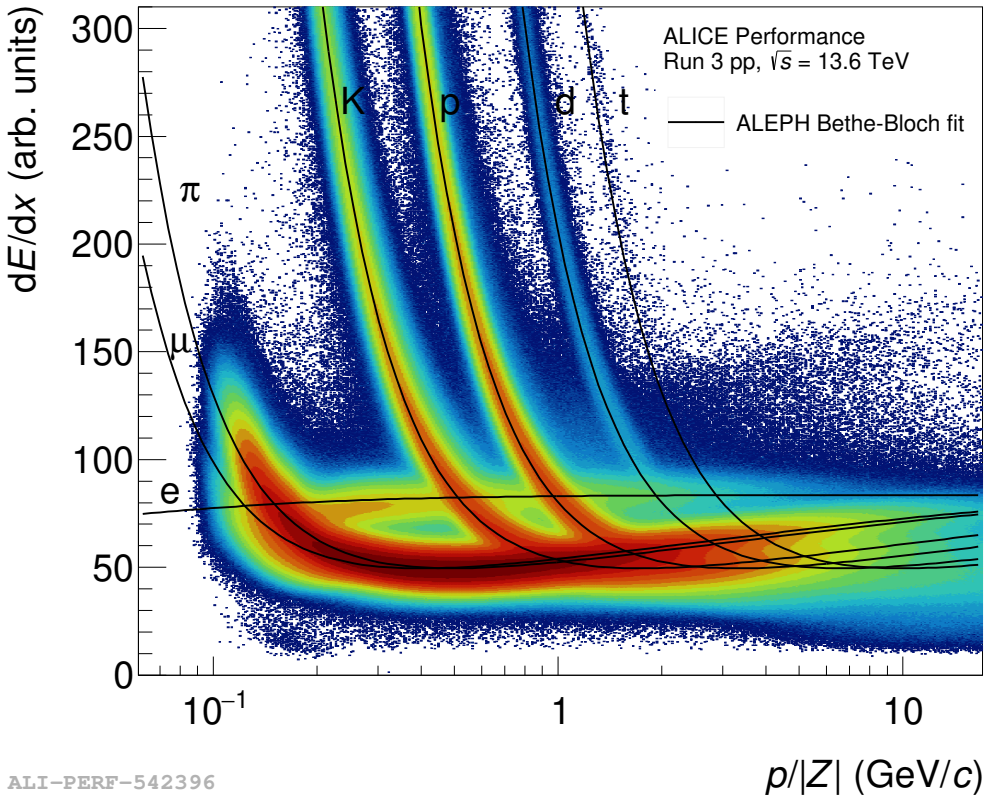

Figure 10: Specific energy loss as a function of momentum in the TPC for different particle species in pp collisions at $\sqrt{s} = 13.6$ TeV in Run 3. Warmer colors represent a higher concentration in track counts, while the black lines show the parameterized Bethe-Bloch fits for different particle species [15].

In order to distinguish between different particle species by utilizing the TPC measurement, the discriminating variable $n_{\sigma^i_{\mathrm{TPC}}}$ is used. It describes the deviation of the measured signal S_{TPC} (energy loss $\frac{dE}{dx}$) from the expected signal $\langle S^i_{\mathrm{TPC}} \rangle$ predicted by eq. (2) for a particle of species i in units of the TPC resolution σ^i_{TPC}. Thus, it is calculated as follows [16]:

$$n_{\sigma^i_{\mathrm{TPC}}} = \frac{S_{\mathrm{TPC}} - \langle S^i_{\mathrm{TPC}} \rangle}{\sigma^i_{\mathrm{TPC}}}. \tag{3}$$

2.4 Central barrel tracking

For the track reconstruction of the particles passing through the detector, mainly the capabilities of the ITS2 and the upgraded TPC, but also the TRD and TOF are utilized. Unlike in Run 1 and Run 2, the track reconstruction for the ITS2 is done completely independently from the TPC.

The first step is the vertexing, i.e. the reconstruction of a PV, which is done by combining the hits from every layer into a preliminary track and prolonging it to the inside of the beam pipe. A PV can then be identified as the point in space, where the maximum number of tracks meet. Afterwards the track finding and track fitting is applied to reconstruct the tracks through the ITS. Simultaneously, the TPC starts with Cluster Finding, where the signals in the detector are assigned positions and errors in both $r\phi$ and z direction to form so-called clusters. Then the Track Finding, Track Merging and Track Fitting is carried out for the TPC tracks by using a Kalman Filter. The overall track reconstruction is done in an inward-outward-inward scheme. In the first inward step the reconstruction starts in the outermost part of the TPC and moves inward until it reaches the innermost part. Next the tracks from the standalone ITS tracking and the TPC tracking are matched and the reconstruction is redone from the innermost layer of the ITS outwards to the outermost part of the TPC. From there it is prolonged into the TRD. Finally, the reconstruction is then repeated inwards from the outermost part of the TRD to the innermost layer of the ITS, which improves the reconstructed tracks even further and allows for an even more precise determination of not only PVs, but also SVs. Due to the high resolution made possible by the upgraded tracking systems, it is possible to determine the position of SVs with even greater spatial resolution than during Run 1 and 2, by finding tracks with a distance of closest approach (DCA) to the PV above a certain threshold. This is essential for the study of short-lived heavy-flavor hadrons, which decay before reaching any detector.

On another note, the central barrel tracking is revolutionized by leveraging the potential of hardware accelerators (GPUs) in Run 3. By running the track reconstruction for the TPC, ITS and TRD on GPUs it is possible to do more synchronous (previously "online") event reconstruction, which is necessary for calibration, data compression, as well as online quality control and was unattainable in this scope in Run 1 and 2 [17, 13, 18].

2.5 Bremsstrahlung

Whenever charged particles pass through matter, they are subject to the emission of bremsstrahlung. It is the electromagnetic radiation that is produced when a charged particle is decelerated by the electric field of an atomic nucleus. Due to its low mass, electrons and positrons are affected significantly stronger than other charged particles. The energy loss per distance traveled through the medium from bremsstrahlung for electrons is given by

$$\left(\frac{dE}{dx}\right)_{rad} = \frac{4n_a Z^2 \alpha^3 \hbar^2 c^2 E}{m_e^2 c^4} \cdot \ln \frac{a(E)}{Z^{1/3}}, \tag{4}$$

where $\alpha = e^2/(4\pi\varepsilon_0\hbar c)$ is the fine-structure constant, E is the energy of the electron, n_a is

the atom density of the medium, Z is the number of protons in the atomic nucleus and $a(E)$ is a numerical factor, which indicates at what impact parameter the incoming electron or positron is still close enough to the nucleus to be deflected. Therefore, the energy loss from bremsstrahlung increases slightly more than linearly with the energy of the electrons and outweighs the energy loss from ionization at high momenta. If one disregards the energy dependence of the factor $a(E)$, eq. (4) can be integrated to compute the electron energy as a function of distance traveled:

$$E(x) = E(0) \cdot e^{-x/X_0}. \tag{5}$$

The length X_0, at which the energy of the electron has decreased to $1/e$ of its original value, is defined as

$$X_0 = \left(\frac{4 n_a Z^2 \alpha^3 \hbar^2 c^2}{m_e^2 c^4} \cdot \ln \frac{a(E)}{Z^{1/3}} \right)^{-1}. \tag{6}$$

The cross section for bremsstrahlung with positrons is in general lower than with electrons, but this effect is negligible for the momenta relevant here. Similarly to bremsstrahlung, the high momentum electrons and positrons are also deflected by the strong magnetic field of the solenoid magnet, leading to the emission of synchrotron radiation. When using the $J/\psi \to e^+e^-$ decay channel, the effects of bremsstrahlung are clearly visible as the invariant mass distribution of the J/ψ reconstructed from the e^+e^- pairs has a tail on the left side. This tail indicates that for a portion of the e^+e^- pairs, at least one of the leptons loses energy through bremsstrahlung and the produced photons are not taken into consideration when reconstructing the J/ψ mass [19, 20].

3 Analysis tools

3.1 KF Particle package

In high-energy particle physics, experiments need to cope with very high track densities, while only rare signals may be of interest. Therefore, high accuracy and high speed are required for the reconstruction of events, in order to find these rare signals efficiently in the large amounts of data. The reconstruction of events involves finding and fitting particle tracks, aligning the detectors, as well as determining the PV and SVs of events from the reconstructed tracks. Fit algorithms like the Kalman filter are utilized to carry out this process. The Kalman filter is a recursive fit algorithm for the analysis of linear discrete dynamic systems described by a state vector, which contains a set of parameters. It provides an optimal estimation of the particle track parameters in order to achieve the highest accuracy. Even nonlinearities can be taken into account, as long as the model describing the system can be linearized beforehand. In high-energy physics, the Kalman filter is used to fit the tracks of charged particles, where trajectories are affected by energy loss through ionization and excitation in the material of the detector, as well as multiple scattering and inhomogenities of the magnetic field inside the detector. It takes great effort to account for such effects using the least squares method when fitting a particle trajectory, since new parameters have to be introduced and fitted for each effect. On the other hand, the Kalman filter is able to handle these nonlinearities, due to the fact that the discrete measurements in the detectors allow for a linearization of the particle track segments, while aforementioned effects in the detector material can be added in the neighborhood of each measurement. Therefore, it is perfectly suited for the challenges of tracking particles with utmost precision in high-energy physics experiments [21, 22, 23].

The KF Particle package is a software package, which uses the Kalman filter method for the reconstruction of decay chains and short-lived particles. It was developed by the CBM collaboration for the vertex reconstruction in high-energy experiments. Since the KF Particle package uses the Kalman filtering algorithm, it provides an optimal estimation of the state vector $\mathbf{r}$ of a particle with the corresponding covariance matrix $\mathbf{C}$, which are iteratively updated after each propagation to the next measurement. The state vector is defined as follows:

$$\mathbf{r} = (x, y, z, p_x, p_y, p_z, E, s)^T. \tag{7}$$

In the KF Particle package, particles are parametrized with their spatial coordinates (x, y, z), their momentum components (p_x, p_y, p_z) and their energy E. Additionally, the parameter $s = \frac{L}{p}$ is included, where L is the distance between the production and decay vertex of the particles in the laboratory frame, while p is the total momentum of the particle. After the estimate optimization of the state vector and its covariance matrix is completed, physical properties of the particle such as mass, lifetime and decay length can be calculated with low computing effort. Additional parameters of the reconstructed tracks, for example the signal that the particle leaves in the ALICE TPC due to energy loss, are also provided by the KF Particle package. Some of these quantities provided by the KF Particle package are used as input features for the training

of BDTs, to differentiate between background and prompt or non-prompt J/ψ [21, 22, 23].

3.2 Boosted Decision Trees

The importance of Machine Learning (ML) for modern particle physics as an analysis tool has grown over the years. Many analyses now utilize ML to handle billions of events, which allows for an even more precise extraction of signal compared to rectangular cuts. In this thesis, Boosted Decision Trees (BDT) are employed to precisely classify background, prompt J/ψ and non-prompt J/ψ. For this task, the XGBoost algorithm and the hipe4ml Python package were used [24, 25].

Decision trees, which are non-parametric supervised learning algorithms used for classification and regression tasks, are the basic units of a BDT. The structure of such a decision tree can be seen in fig. 11 [26].

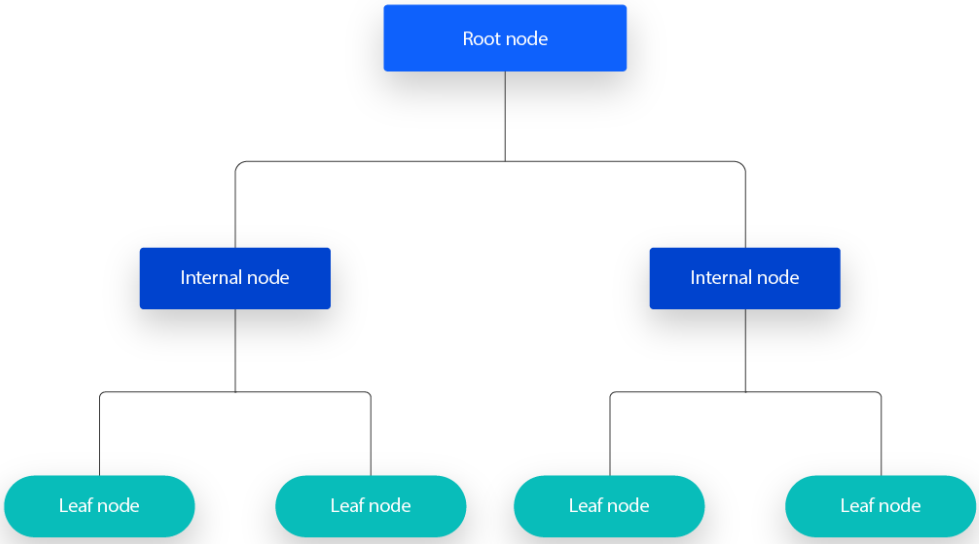

Figure 11: Structure of a simple decision tree [26].

A decision tree forms a hierarchical tree structure with a root node at the top, branching into internal nodes and lastly, leaf nodes. To train a model, a set of candidates where each candidate holds a set of parameters and should be assigned to one unambiguous class, is considered. The first decision is made at the root node, where all candidates start. If, for example, the decay length of a candidate is higher than a certain threshold, it will follow the first branch to an internal node, while it will follow the second otherwise. This procedure is then repeated until a leaf node is reached. At each leaf node, candidates of predominantly one class should be collected, since it is essential to achieve the highest possible purity for each leaf node. Decision trees can be chosen to have higher depth and complexity, but at a certain point not enough data falls into each subtree, which can lower the purity of leaf nodes. This is called overfitting and it manifests in data not used for the training reaching significantly lower scores than the training data. The reason for this is that the model learns statistical fluctuations instead of the desired properties

when the trees have high depth or complexity. In the XGBoost algorithm, classification and regression trees (CART) are employed. These CARTs are a slightly modified version of decision trees, where the leaf nodes do not represent a given class. Instead they assign a score to each candidate, which leaves more room for the interpretation of tree outputs. In order to increase performance without increasing the complexity of the trees, an ensemble of trees is be used, which can yield better performance than any single decision tree could [24, 26, 27].

XGBoost employs so-called "boosting", where low complexity decision trees (weak learners) are combined into one strong learner . This sequential learning process is iterative, meaning that each tree is constructed in such a way that the previous weak learners with their errors are considered. In fig. 12, a visualization of this boosted decision tree model is depicted [24, 27].

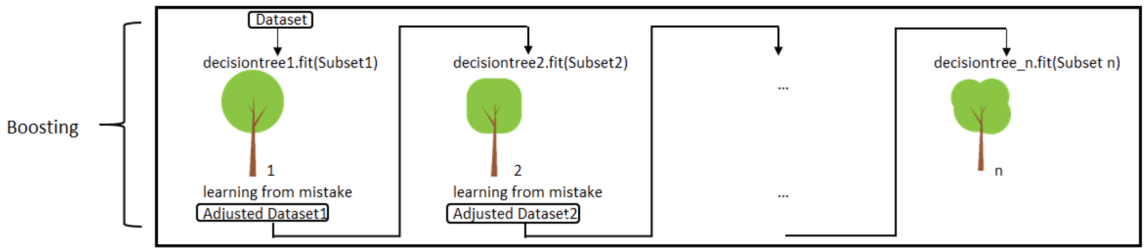

Figure 12: Scheme of tree boosting, where each tree learns from errors of the previous weak learners [28].

The XGBoost algorithm in the context of supervised learning tasks uses a set of features x_i of the training data set to make a prediction $\hat{y}_i$ of the target value y_i. For the given use case of classifying heavy-flavor hadrons, the prediction value is interpreted as the probability for a candidate to belong to a given class. BDT models are made up of many different parameters, such as the features which will be used to make a decision, but also so-called hyperparameters, which are optimize for the training, like the maximum tree depth. During the training of a model, parameters will be varied to fit the training data, while hyperparameters are determined beforehand. Hyperparameters like the maximum tree depth and learning rate are optimized using the Optuna package included in the hipe4ml Python package, ultimately improving the predictions [24, 29].

BDT models can be used for binary classification, where the classes are background and signal. However, since prompt and non-prompt J/ψ should be differentiated, multiclass classification is utilized, with background, prompt and non-prompt as the classes. In order to carry out multiclass classification, it is separated into multiple binary classifications, where two different approaches, One-vs-One (OvO) and One-vs-Rest (OvR), can be used. For the OvO approach, one BDT is trained for every combination of two classes. The classes compete against each other one by one, resulting in the pairs background vs. prompt, background vs. non-prompt and prompt vs. non-prompt. Thus, three models are trained in total and when the model is applied, a majority vote decides the final model prediction and an output score from zero to one for each class. On the other hand, the OvR approach trains three models again, one for each class. However, in this case, the classes always train against the rest of the classes, i.e. background vs. rest, prompt vs rest and non-prompt vs. rest. The result is again an output score from zero to one for each class,

which states the likeliness of the candidate to belong to the specific class. At a later stage of the analysis, selections for the output scores must be determined, to classify the data with minimal losses. For this analysis, the OvR approach is chosen. In this specific use case, the performance difference between the approaches is minimal [30].

4 Analysis

4.1 Preselections

This analysis of the decay channel $J/\psi \to e^+e^-$ uses data measured in pp collisions at a center of mass energy of $\sqrt{s} = 13.6$ TeV with the ALICE detector during Run 3 at midrapidity range ($|y| < 0.9$). To carry out the analysis in this thesis, a data set with reconstructed candidates and a Monte Carlo (MC) simulation set for both prompt and non-prompt J/ψ were provided. Specifically, 22pass7 data and monte carlo simulations are used, where 22pass7 is an expression for the seventh reconstruction pass of the raw data from 2022. Before the data is further analyzed, it should be noted that a number of preselections for the tracks were applied to the dataset prior to this analysis. The preselection cuts are listed in table 3 and will be explained briefly. In order to restrict the data to regions of the detector where the track reconstruction works best, i.e. the center of the detector, the event cut $|V_z| < 10$ cm was chosen, where V_z is the position of the PV in z-direction of an event.

Table 3: List of preselections applied to the reconstructed tracks.

Selection conditions
$p_T > 0.7$ GeV/c
$\|\eta\| < 0.9$
ITSibAny = true
$\chi^2_{ITS} < 5.0$
TPCncls > 60
$\|DCA_{xy}\| < 1.5$ cm
$\|DCA_z\| < 1.5$ cm
$\|n_{\sigma^{e^-}_{\mathrm{TPC}}}\| < 4.0$
$n_{\sigma^{\pi}_{\mathrm{TPC}}} > 2.5$
$n_{\sigma^{p}_{\mathrm{TPC}}} > 2.5$

- p_T: This is the transverse momentum of each decay particle, which can in principle be any charged particle of which the momentum is measured, e.g. a pion, kaon, proton or electron. It is sensible to use only tracks above a certain momentum threshold to exclude irrelevant background.
- η: The pseudorapidity range is limited to the coverage of the TPC, since the PID and tracking capabilities of it are used.
- : This cut requires the fit of a track via the ITS to pass the given threshold. Additionally, the tracks of the particles have to originate from the same vertex.
- ITSibAny: This variable indicates whether at least one hit was measured by the innermost three layers of the ITS.
- χ^2_{ITS}: This cut requires the fit of a track via the ITS to pass the given threshold. Additionally, the tracks of the particles have to originate from the same vertex.

- TPCncls: In order to ensure that only tracks with a certain amount of measurements by the TPC are used, in order to obtain high track quality, it is required that the TPC measured a certain amount of clusters for the particle.
- DCA_{xy} and DCA_z: The distance of closest approach (DCA) of a particle to its associated PV needs to be below a certain threshold. Above this threshold, it is very unlikely that particles stem from a J/ψ.
- $n_{\sigma^{e^-}_{\mathrm{TPC}}}$, $n_{\sigma^{\pi}_{\mathrm{TPC}}}$ and $n_{\sigma^{p}_{\mathrm{TPC}}}$: These PID quantities were explained in section 2.3. They describe the distance of the energy loss measurement to the energy loss band of the specified particle. Therefore, they are used to exclude most of the background, since no electrons will be present in the regions beyond the used cuts.

These preselection are necessary in order to filter out any background candidates, which would be easily classified, since the BDT training should focus on candidates more difficult to classify to achieve the best results.

4.2 Signal extraction with rectangular cuts

Before making use of machine learning via BDTs, the signal extraction achieved with only the rectangular cuts of the preselection is determined for later use as a point of reference. Since the goal, ultimately, is to measure the p_T-differential cross section of the J/ψ, the data is divided into four transverse momentum intervals ($0-2$ GeV/c, $2-4$ GeV/c, $4-6$ GeV/c and $6-12$ GeV/c) of the J/ψ. In order to reconstruct the J/ψ signal, the invariant mass distributions for these intervals have to be fit with an adequate fit function that represents the shape of the invariant mass distribution of the J/ψ and the background. The final fit function consists of a Crystal Ball function to accurately describe the signal shape and a second degree polynomial for the background. An asymmetrical tail provided by the Crystal Ball function takes into account both the detector resolution, as well as the loss in invariant mass resolution due to bremsstrahlung. The Crystal Ball function consists of a Gaussian core with a power-law tail to the left side and is defined as follows:

$$f(x;\alpha,n,\mu,\sigma) = N \cdot \begin{cases} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), & \text{for } \frac{x-\mu}{\sigma} > -\alpha \\ A \cdot \left(B - \frac{x-\mu}{\sigma}\right)^{-n}, & \text{for } \frac{x-\mu}{\sigma} \leq -\alpha \end{cases}, \tag{8}$$

where

$$A = \left(\frac{n}{|\alpha|}\right)^n \cdot \exp\left(-\frac{|\alpha|^2}{2}\right),$$
$$B = \frac{n}{|\alpha|} - |\alpha|,$$

with the normalization factor N, the mean of the Gaussian μ and the standard deviation of the Gaussian σ. The parameter α specifies the position of the transition between the Gaussian and the power-law tail, while n specifies how fast the tail falls off. However, before the fitting is carried out, a like sign invariant mass distribution is created via the combination of two reconstructed

particles with the same charge (like e^+e^+ or e^-e^-) and is subsequently subtracted from the unlike sign distribution from two particles with opposite charge (like e^+e^-). By subtracting the combinatorial background like this, the background fitted by the polynomial is lower and more flat, improving the quality of the overall fit. It should be noted that the like sign background had to be scaled up to represent the combinatorial background. Afterwards, the remaining mass distribution is fitted by the aforementioned combination of a Crystal Ball function and a second order polynomial. The invariant mass distributions of unlike sign data and like sign data, as well as the fits of the residual distributions are shown in fig. 13 for every p_T interval. From the fits, the number of J/ψ candidates in the signal region mass range of $2.7 < m_{J/\psi} < 3.2$ GeV/c^2 is calculated via the fit of the Crystal Ball function, as well as the number of background candidates in the same region, with both the residual and combinatorial background included. Additionally, the signal-to-background ratio $\frac{S}{B}$ and the significance $\frac{S}{\sqrt{S+2B}}$ are computed. The same procedure is carried out with additional PID cuts of $n_{\sigma^{\pi}_{\mathrm{TPC}}} > 4.0$ and $n_{\sigma^{p}_{\mathrm{TPC}}} > 4.0$ and the resulting invariant mass fits are shown in fig. 14. These specific cuts were chosen, since they are later also utilized for the selection of features for the BDT training. Applying such selections may be beneficial to increase the signal-to-background ratio and significance, due to the reduction of background, though the yield of J/ψ is also drastically reduced. In the case of the $6-12$ GeV/c p_T range, the significance is lower than before, demonstrating that rectangular cuts have to be applied with great care, especially in a range where only a low amount of signal is available. The signal extraction carried out here does not differentiate between prompt and non-prompt J/ψ. Separation of these two categories via rectangular cuts would lead to a further decrease in signal events. Therefore, the application of BDTs for the purpose of further decreasing background and separating prompt and non-prompt J/ψ is tested.

4.3 Machine learning training

4.3.1 Training candidate selection

Boosted decision trees are well suited for the separation of background, prompt J/ψ and non-prompt J/ψ, due to their ability to find and use patterns and correlations in the data for the classification. ML models however, have to be trained on a training data set. For the background data, the sidebands of the J/ψ invariant mass distribution in data are chosen, which in this case includes the ranges $1.2 < m_{candidate} < 2.2$ GeV/c^2 and $3.2 < m_{candidate} < 4.0$ GeV/c^2. These ranges are chosen since the training background data set should include as few J/ψ signal as possbile, while also matching the properties of the background that lies within the signal region ($2.2 < m_{candidate} < 3.2$ GeV/c^2). The signal data set is made up of MC simulated prompt and non-prompt J/ψ. Additionally, it should be noted that a model is trained for each p_T interval, since features can behave differently and have varying importance in the classification process, depending on the transverse momentum of the J/ψ. The total number of candidates used in the training for each class and each transverse momentum interval is shown in table 4. For the number of candidates used in the training of the model all available prompt and non-prompt candidates are considered, while for the background the number equals to the total amount of

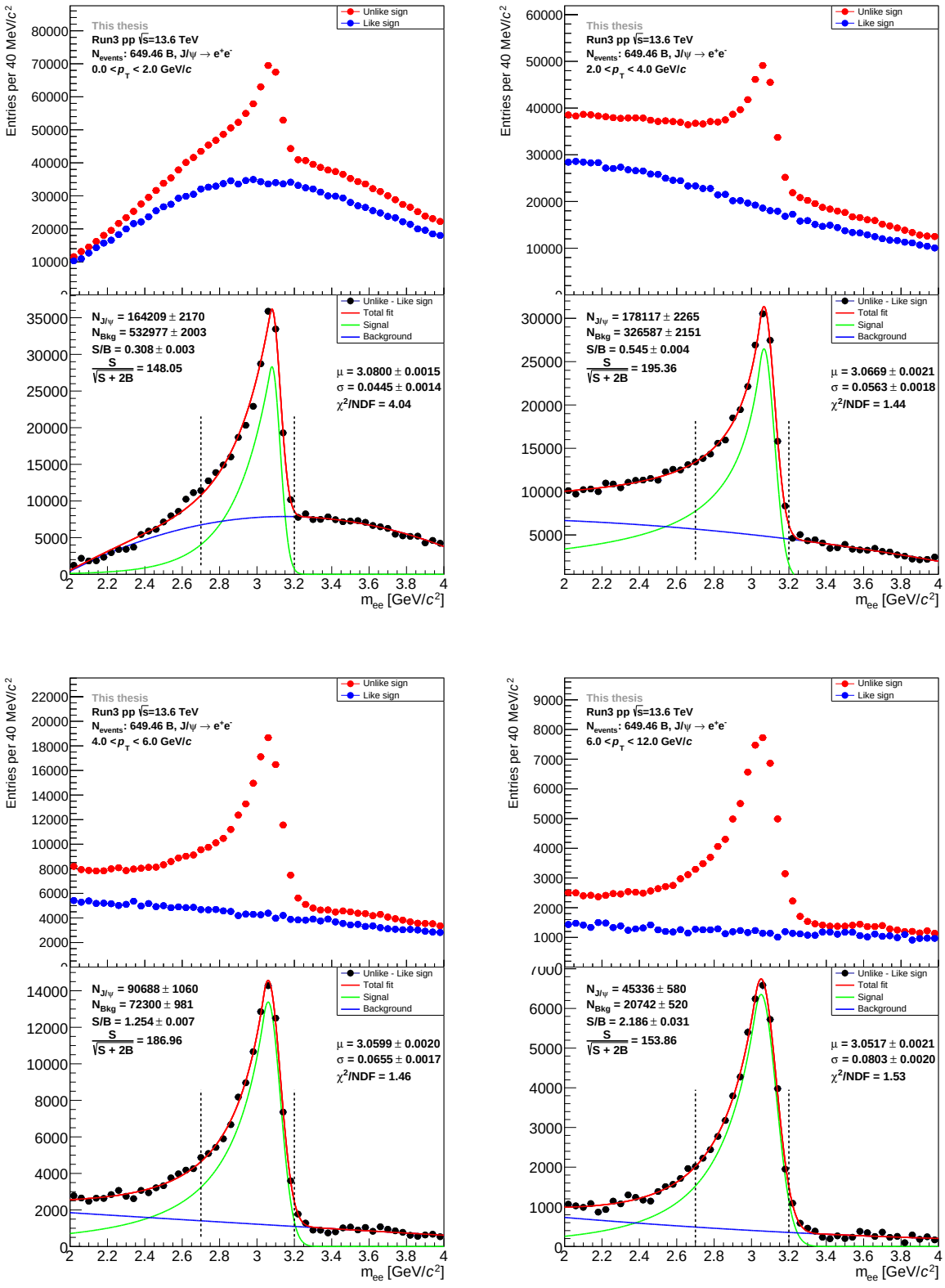

Figure 13: Invariant mass distribution fits for inclusive $J/\psi \to e^+e^-$ for all p_T ranges with no additional PID cuts applied and combinatorial background subtracted.

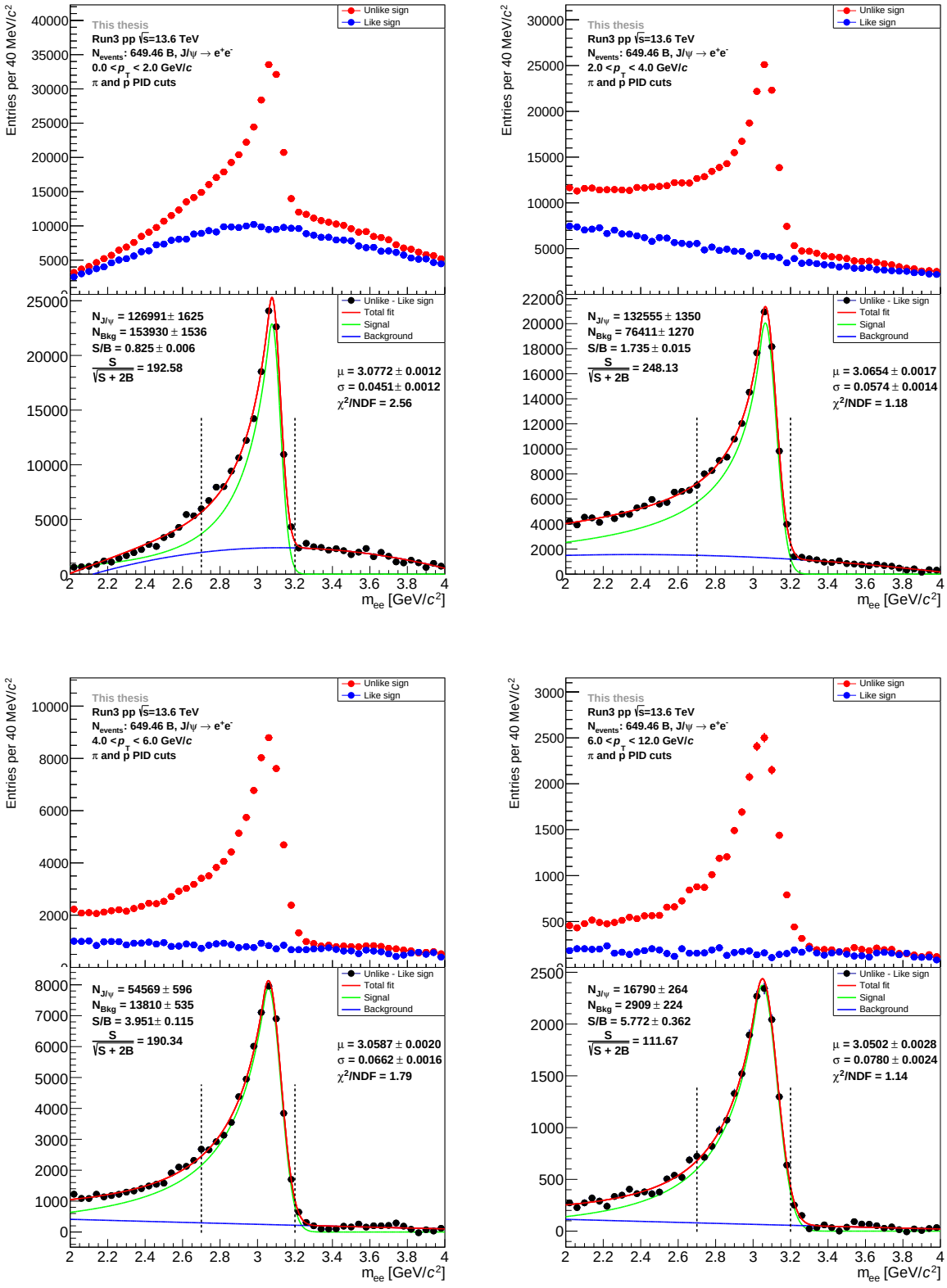

Figure 14: Invariant mass distribution fits for inclusive $J/\psi \to e^+e^-$ for all p_T ranges with $n_{\sigma^{\pi}_{\mathrm{TPC}}} > 4.0$ and $n_{\sigma^{p}_{\mathrm{TPC}}} > 4.0$ cuts applied and combinatorial background subtracted.

signal candidates (Bkg = P + NP). In case there is not enough background data to fulfill this condition, as in the case of the $6-12$ GeV/c p_T range, all available background candidates are used. Using an equal amount of candidates for each class was tested and found to yield slightly worse results. Only 50% of the chosen candidates are used for training, while the other 50% are used for testing and validation.

Table 4: Number of candidates for each p_T interval and each class, used in the BDT training process.

p_T [GeV/c]	$0-2$	$2-4$	$4-6$	$6-12$
Background	505201	498096	280534	127613
Prompt	437855	429855	243665	186320
Non-prompt	67346	68241	36869	24794

4.3.2 Feature selection

In order to train a model, it is necessary to select features, which allow the BDT to differentiate between the classes. It is, in principle, possible to use every available quantity of the candidates as an input feature, but this increases the complexity of the model, possibly resulting in overfitting. Hence, only the most impactful features are used to train the BDTs. The chosen features should not include the invariant mass and the transverse momentum, since it would bias the classification towards certain values of mass and momentum. However, fitting the mass and measuring the p_T-differential cross section correctly, requires a model that is unbiased towards these quantities. For the selection of the features, their distributions in the prompt and non-prompt MC data, the sideband background data, as well as in the signal region data with $n_{\sigma^{\pi}_{\mathrm{TPC}}} > 4.0$ and $n_{\sigma^{p}_{\mathrm{TPC}}} > 4.0$ PID cuts applied, are examined (see fig. 15).

Some features are better suited to differentiate between background and signal, while others help to separate prompt and non-prompt J/ψ. Both types of features are necessary for an efficient and meaningful classification. As an example for the first type, the quantity "TPCNSigmaEl" ($n_{\sigma^{e^-}_{\mathrm{TPC}}}$) is chosen (see fig. 16), since parts of the distribution for the background differ significantly from both MC distributions. It should be noted that the MC distributions are slightly shifted to the right due to calibration problems in this data set, but it was found that the usage of this feature improves the model nevertheless.

For the second type of feature, the quantity "DecayLengthOverErrorKFGeo" is utilized, due to the difference in the distributions of prompt and non-prompt J/ψ, enabling a differentiation between the two classes. The difference in the distributions becomes more pronounced for rising transverse momentum. In total, seven features have been chosen for model training, of which the remaining distributions for every transverse momentum interval are shown in fig. 24 to fig. 27 in the appendix. The features are the following:

- **TPCNSigmaEl1/2:** This is the $n_{\sigma^{e^-}_{\mathrm{TPC}}}$ PID variable for electrons provided by the TPC energy loss measurement, as explained in section 2.3.
- **DecayLengthOverErrKFGeo:** This is the decay length of the reconstructed candidate

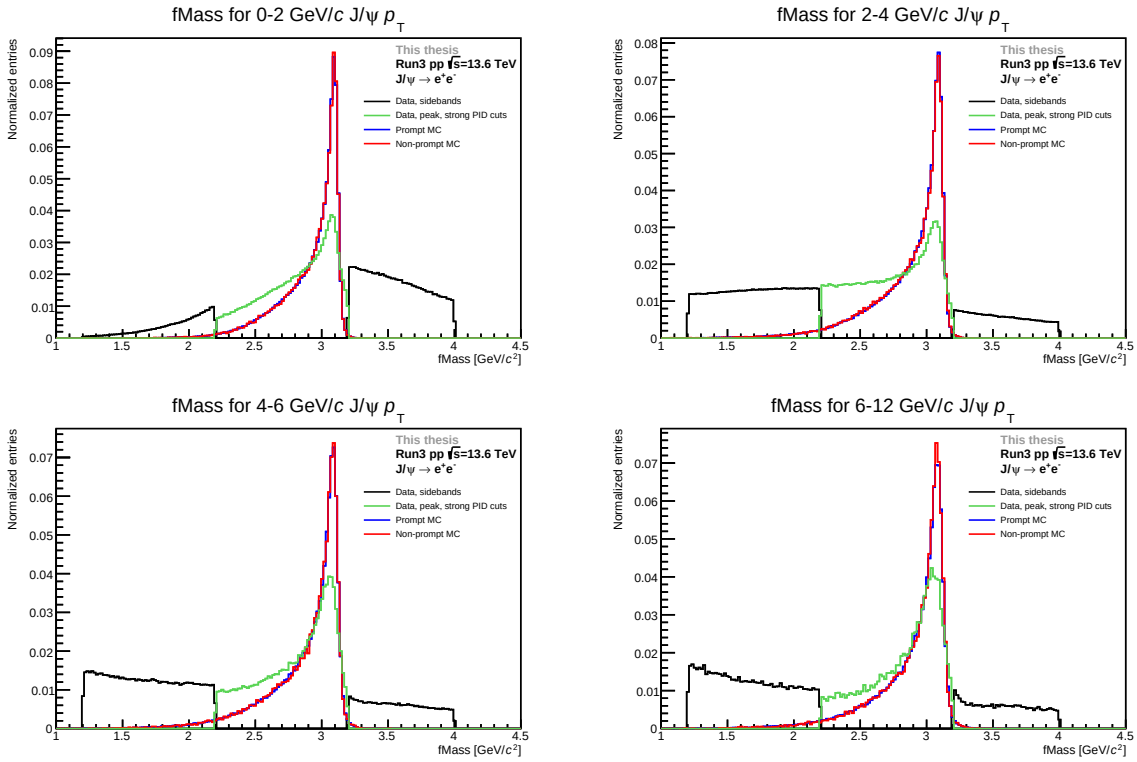

Figure 15: Distribution of the invariant mass for data in the sideband regions, prompt MC, non-prompt MC and data in the peak region with $n_{\sigma^{\pi}_{\mathrm{TPC}}} > 4.0$ and $n_{\sigma^{p}_{\mathrm{TPC}}} > 4.0$ cuts applied in all p_T ranges.

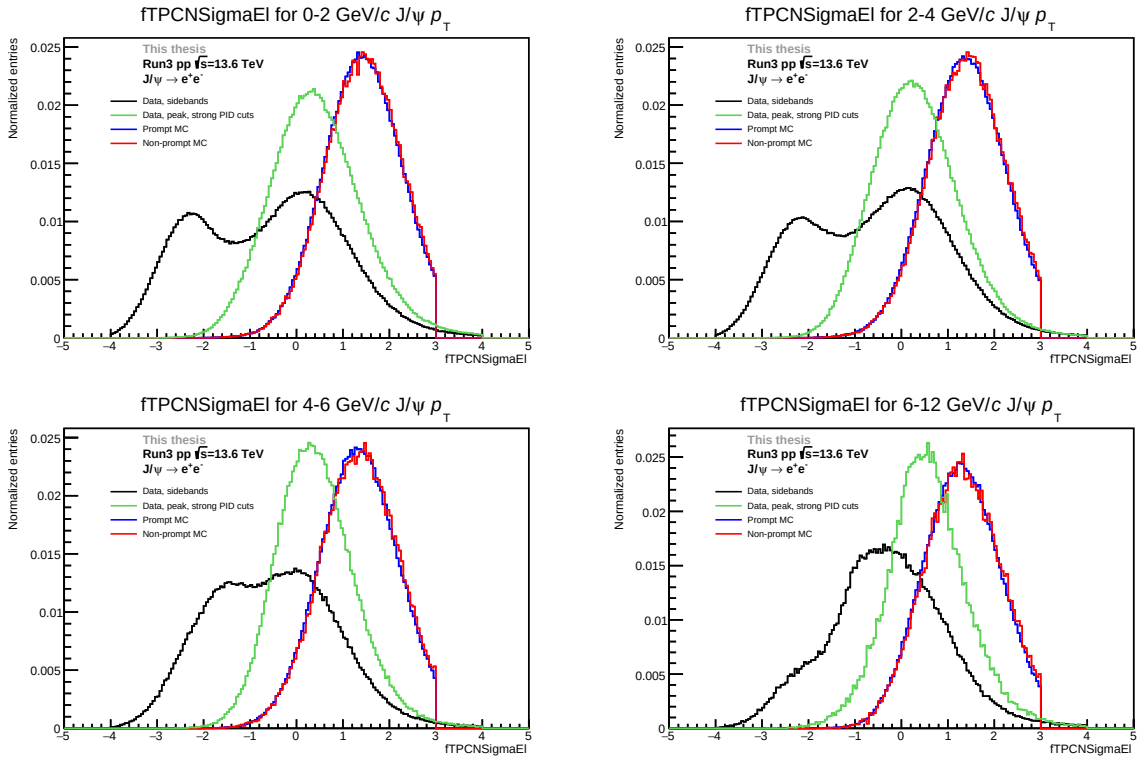

Figure 16: Distribution of TPCNSigmaEl for data in the sideband regions, prompt MC, non-prompt MC and data in the peak region with $n_{\sigma^{\pi}_{\mathrm{TPC}}} > 4.0$ and $n_{\sigma^{p}_{\mathrm{TPC}}} > 4.0$ cuts applied in all p_T ranges.

divided by its error. The decay length is defined as the distance between the PV and the decay vertex of the J/ψ.

- **PseudoproperDecayTimeKFGeo:** This is the decay time calculated from the projection of the decay length in the x-y-plane. The formula is:

$$\tau = \frac{m}{p_T} \mathcal{L}_{xy}. \tag{9}$$

- **DCAxyzBetweenTrksKF:** This is the distance of closest approach between the two tracks associated with the candidate in three dimensions.

- **Chi2OverNDFKFGeo:** This describes the χ^2/NDF of a geometrical fit of the candidate track.

- **CosPAKFGeo:** This is the cosine of the pointing angle, which is the angle between the momentum of the reconstructed J/ψ candidate and the line that connects its decay vertex with the primary vertex.

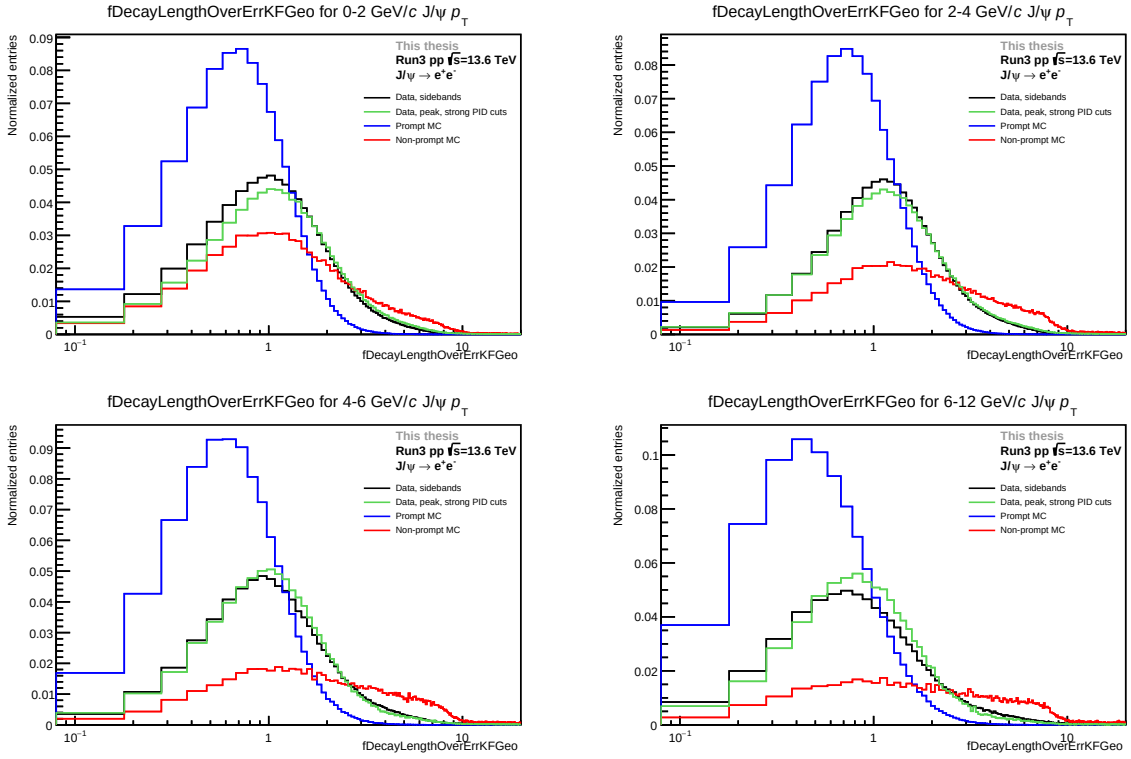

Figure 17: Distribution of DecayLengthOverErrKFGeo for data in the sideband regions, prompt MC, non-prompt MC and data in the peak region with $n_{\sigma^{\pi}_{\mathrm{TPC}}} > 4.0$ and $n_{\sigma^{p}_{\mathrm{TPC}}} > 4.0$ cuts applied in all p_T ranges.

Additionally, it is essential to check the correlations of the features, since correlations only present in a specific class can beneficial to the classification. On the other hand, correlations between the features used for training and the invariant mass as well as the momentum have

to be avoided, since it would distort the results. The correlation matrices for all three classes for the $2-4$ GeV/c p_T interval model are shown in fig. 18. For all transverse momentum intervals, no significant levels of correlation are found. Although there is some correlation between "DCAxyzBetweenTrksKF" and "Chi2OverNDFKFGeo", it was found that including both features improves the performance of the models. It is also insightful to take a look at the importance of the different features, which is shown in fig. 19 for the $2-4$ GeV/c p_T interval. In the figure, the features are sorted by relevance, which is quantified by the so-called "mean SHAP (SHapley Additive exPlanations) values". These values indicate the average impact of the respective feature on the model, while the different colors show how important the feature is for the classification to a specific class.

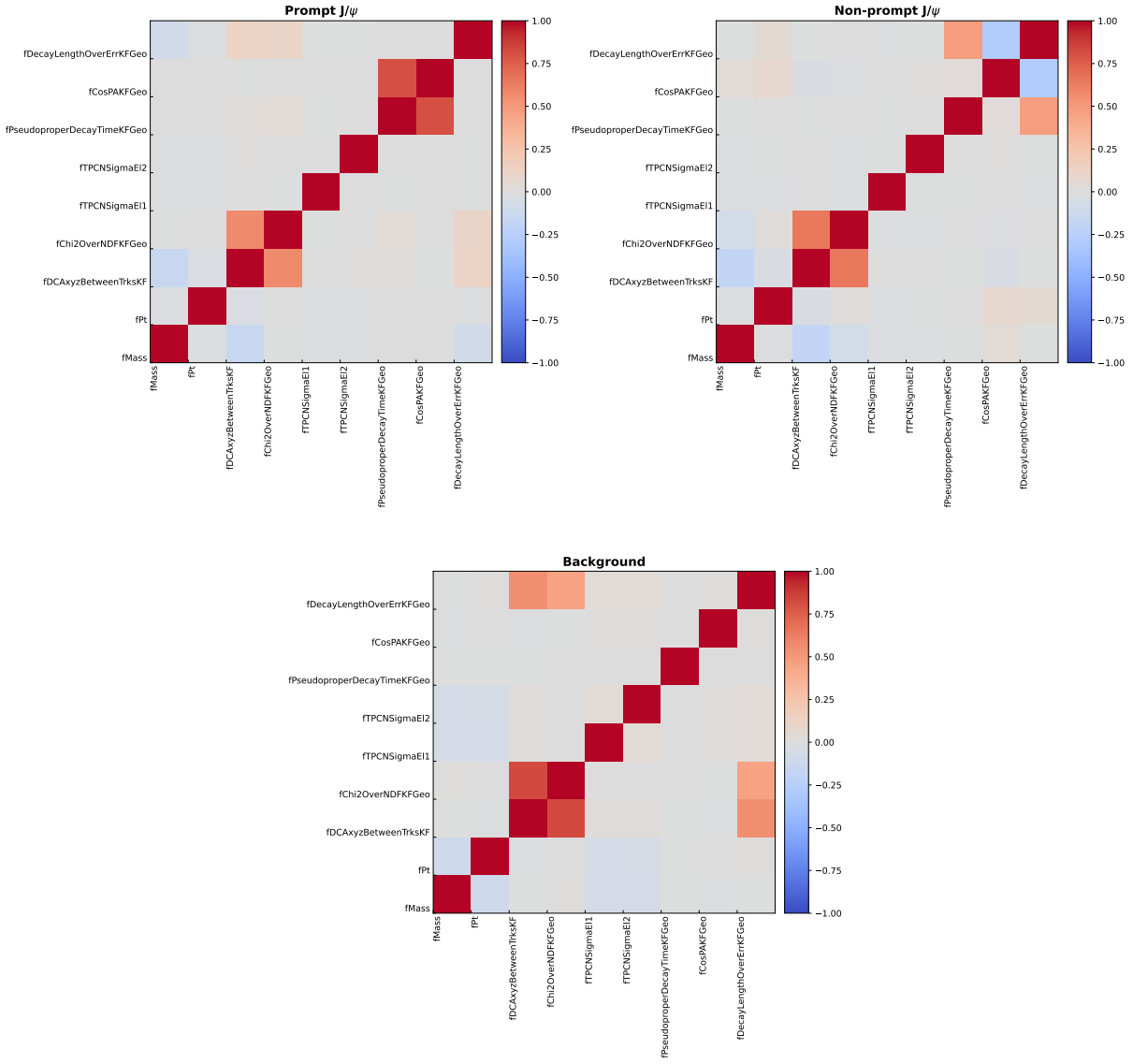

Figure 18: Correlation matrices of the used features, invariant mass m and transverse momentum p_T for the three classes in the $2 < p_T < 4$ GeV/c range.

Due to changes in the feature distributions with respect to momentum, the feature importances

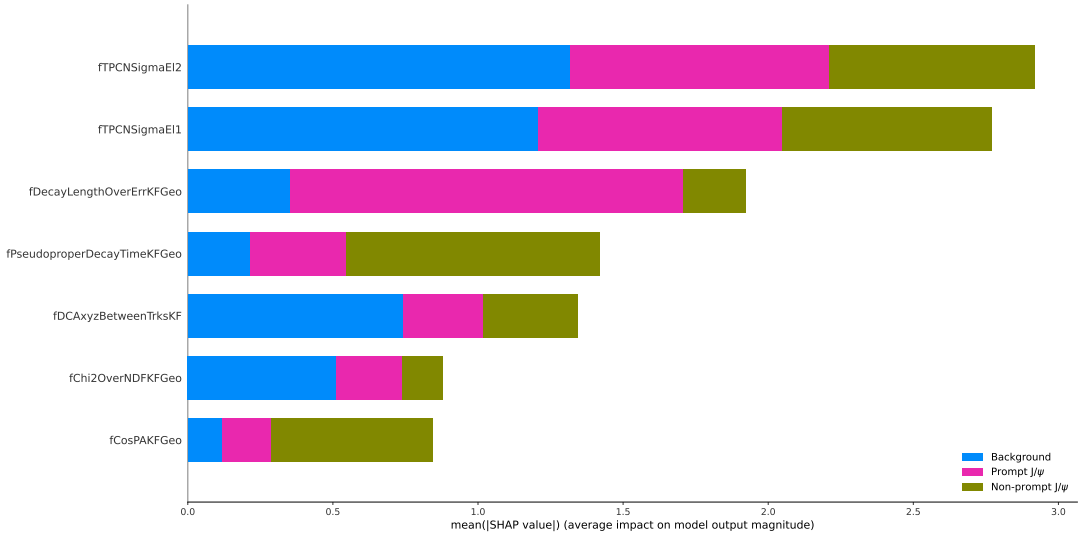

Figure 19: Impact of the used features on the model for each class measured by the mean SHAP value. These are the feature importances of the model for the $2 < p_T < 4$ GeV/c range.

and correlations vary between the different transverse momentum intervals, but they do not show any unwanted behavior in any case. Feature correlations for the remaining intervals are shown in fig. 32, fig. 33 and fig. 34, while importances are shown in fig. 35, fig. 36 and fig. 37.

4.3.3 Hyperparameter optimization

After the selection of the features used to train the models, the hyperparameter optimizations are carried out using the Optuna module. In essence, this means that the predefined ranges for the parameters are scanned iteratively, where each iteration considers the previous evaluations to choose the next set of values for the hyperparameters. They are optimized to optimize performance, while reducing complexity without signficant performance loss. The optimized hyperparameters are shown in table 5 for the four models associated with transverse momentum intervals.

The first hyperparameter is the maximum depth of a single weak learner, for which values between 1 and 4 were tested. In three of the four models the upper limit of the range is reached. However, allowing higher values for the maximum depth can lead to overfitting, which should be avoided to retain optimal performance of the models. The rate at which the model is adapted to the data is parameterized by the learning rate. For this parameter, a range from 0.01 to 0.1 was tested. High learning rates achieve faster adaptation, which can lead to the model not converging to the optimum, while low learning rates are slower, but more thorough. The number of weak learners in a BDT model is determined by the number of estimators, which should be high, since a low tree depth and therefore trees with low complexity are used. A range from 20 to 1500 was tested for this hyperparameter. For half of the models, the number of estimators is close to the upper boundary, so higher values could be tested to possibly yield a better performance, but this was deemed sufficient for now. The hyperparameter minimum child weight is important for the so-called pruning process, which optimizes the depth of each

weak learner, so unnecessary complexity is avoided. This parameter quantifies the minimum sum of weights for a daughter node of a decision tree to not be removed by the pruning process. Low values lead to a large number of tree partitions, while large values lead to fewer partitions. The test range for this parameter was between 1 and 10. Since always using the whole data sample can lead to overfitting, subsamples which are randomly chosen are used. The size of these subsamples for every step of boosting as a fraction of the entirety of the data is described by the subsample hyperparameter, where a test range from 0.8 to 1.0 was chosen. Finally, the hyperparameter column sample by tree is the fraction of randomly selected features used to train each decision tree. As for the fraction of the data samples, always using every feature to generate each tree, instead of randomly choosing a fraction of the features, can lead to overfitting. For this last hyperparameter, the range between 0.8 and 1.0 was tested [24].

Table 5: Optimized hyperparameters for each p_T interval and therefore each BDT model.

p_T [GeV/c]	$0-2$	$2-4$	$4-6$	$6-12$
Maximum Depth	3	4	4	4
Learning Rate	0.092	0.053	0.052	0.039
No. of Estimators	782	1092	1495	1461
Minimum child weight	1	4	6	3
Subsample	0.934	0.901	0.969	0.849
Col. Sample by Tree	0.833	0.961	0.900	0.865

4.3.4 Trained models

After the optimization of the hyperparameters, the models for each interval of transverse momentum are trained on the candidates (see table 4) with the seven chosen features (see fig. 35). In order to check the performance of a model for each class, the Receiver Operating Characteristic (ROC) curves are considererd. The ROC curves describe the true positive rate as a function of the false positive rate for each class. True positive candidates are correctly classified candidates, e.g. a candidate from the non-prompt MC data set is correctly assesed to belong to the non-prompt class. On the other hand, a false positive candidate is incorrectly classified, which for this example means that a candidate from the prompt MC or background data set is incorrectly classified to the non-prompt class. However, the model only assigns scores to the candidates, which quantify the likelihood of candidates to be part of a certain class, so it does not classify them unambiguously. Therefore, the true positive rates and true negative rates are dependent on the selection of the scores.

A ROC curve is constructed by applying the entire range of possible selections and plotting the resulting true positive rates as a function of the true negative rates. When the selection is set as high as possible, both the true positive rate and the false positive rate are 0, since all candidates from a given data set are not assigned correctly, while no candidates from the other two data sets are incorrectly assigned to that class. If, however, the lowest score and every value higher than that is selected, both a true positive rate of 1 and a false positive rate of 1 are observed, because then every background, prompt, and non-prompt candidate is assigned to that

specific class. Good models produce ROC curves that show a steep rise to a true positive value of 1 with respect to the false positive rate, hence selecting the correct class of the candidate effectively, while rejecting most of the candidates from the remaining classes. Therefore, a simple way to assess the quality of the model is to calculate the Area Under Curve (AUC) of a ROC curve, where a perfect classifier yields an AUC value of 1. For the $2 - 4$ GeV/c p_T model, the ROC curves of each class and their respective AUC values are shown in fig. 20.

In order to check for possible overfitting, plotting the ROC curves and AUC values for both the training and test data sets is necessary. For strong overfitting, the ROC curves of the training and test data sets will strongly deviate from each other, resulting in a lower AUC score for the test data. Curves that lie very close to each other in all regions on the other hand, are a sign for little to no overfitting being present and therefore lead to similar AUC scores for both ROC curves. The AUC values of every model for each class, as well as the average values are shown in table 6, while the ROC curves for the remaining three p_T intervals can be seen in fig. 38, fig. 39 and fig. 40. None of the trained models show strong overfitting.

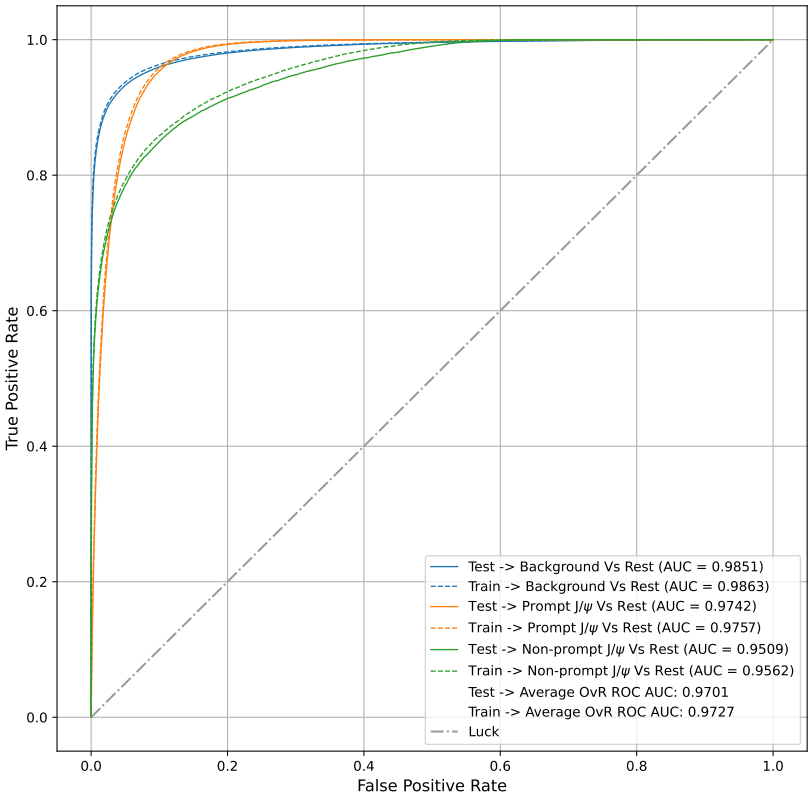

Figure 20: ROC curves and their AUC values of the model for the $2 < p_T < 4$ GeV/c range.

Finally, the results of the classifier model are shown in fig. 21. In these histograms, the probabilities of all candidates to belong to the given class is visualized for background, prompt and

Table 6: AUC scores of each class for each p_T interval, as well as the average scores.

p_T [GeV/c]	$0-2$	$2-4$	$4-6$	$6-12$
Background Test	0.9786	0.9851	0.9818	0.9745
Background Training	0.9796	0.9863	0.9850	0.9788
Prompt Test	0.9603	0.9742	0.9756	0.9711
Prompt Training	0.9622	0.9757	0.9794	0.9757
Non-prompt Test	0.9113	0.9509	0.9573	0.9557
Non-prompt Training	0.9158	0.9562	0.9671	0.9686
Average Test	0.9501	0.9701	0.9716	0.9671
Average Training	0.9525	0.9727	0.9772	0.9744

non-prompt candidates and both the training and test data set. The distributions for the training and test data sets do not deviate significantly, which leads to the conclusion that practically no overfitting is present. Looking at the BDT output for background, given the large range of the y-axis, a clear separation of real background from prompt and non-prompt J/ψ is visible. For the prompt output, most real prompt J/ψ candidates get assigned high probabilities and most background candidates receive low values, while a small fraction of non-prompt candidates are incorrectly associated. Lastly, background and prompt candidates achieve low values for the non-prompt BDT output and non-prompt candidates have high probabilites, but a small portion of the non-prompt candidates gets low values assigned. Nevertheless, the model is capable of separating candidates from the three classes efficiently and only a small fraction of the candidates that are not of interest should remain after the cuts for a specific class are applied. The same is true for all other models, for which the BDT output plots are shown in fig. 41, fig. 42 and fig. 43.

4.4 Selection of working points

The BDTs are now applied to all available data, hence every reconstructed candidate is assigned an output value for each of the three classes. In order to efficiently extract prompt or non-prompt J/ψ signal, working points have to be chosen, i.e. optimal selections for the outputs have to be made. For the background outputs, the selection value is an upper bound, since only prompt and non-prompt candidates are of interest and background should be excluded. In contrast, selecting candidates with the prompt or non-prompt output requires a lower bound, since either prompt or non-prompt candidates should be chosen. Setting loose working points leads to high efficiency of the signal extraction with a low purity, where for example the fraction of non-prompt J/ψ is high, while only prompt are of interest. Although high purity can be achieved by choosing the working points very strictly, this approach results in low efficiency for signal selection. Therefore, a balance between efficiency and signal purity needs to be found by finding an appropriate working point. In principle, this optimal working point is found using a full working point determination, where the significance and fraction of prompt and non-prompt signal is optimized for an efficient and pure extraction of either prompt or non-prompt J/ψ. However, this process is not carried out here. Instead, a working point for the prompt or non-prompt output is first chosen and then the significance $\frac{S}{\sqrt{S+2B}}$ of the signal S is maximized. For every model, the working point for the

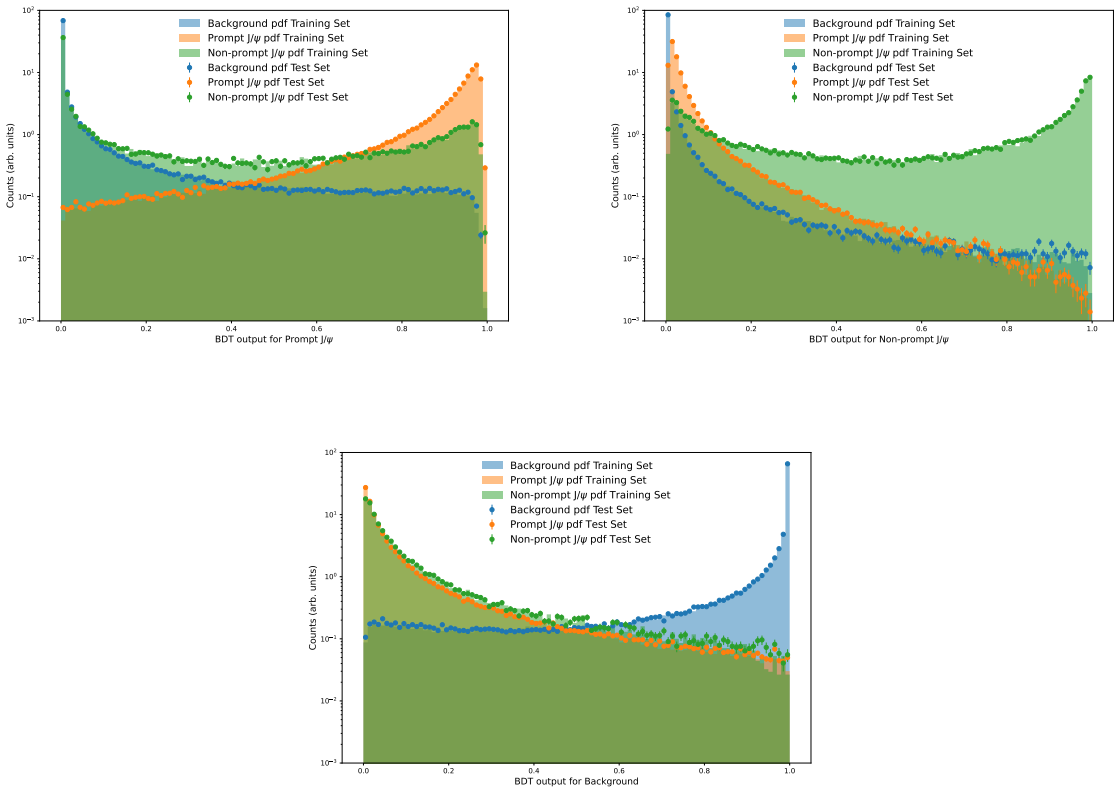

Figure 21: BDT outputs for the three classes with training and test data in the $2 < p_T < 4$ GeV/c range.

extraction of prompt signal is chosen to be at the crossing point of the histograms of prompt and non-prompt candidates in the prompt BDT output plot. These selections are chosen to achieve a high fraction of prompt candidates, as well as a low fraction of non-prompt candidates, although exact numbers for these fractions are unknown. This may be in contradiction to the statement, that efficieny should not be given up for high purity, but it is a good method to examine whether the separate extraction of prompt and non-prompt J/ψ leads to any problems. In order to calculate the significance, a similar fitting procedure as in section 4.2 is employed, while simultaneously a large portion the candidates is already excluded due to the prompt BDT output cut applied. The important difference is that the parameters α and n of the Crystal Ball function are fixed to the values retrieved from fits of the prompt and non-prompt MC invariant mass distributions, to match the shape of the power-law tail. By varying the background selection using the background BDT outputs, the significance can then be maximized. The process is analogously applied to select the working points for non-prompt signal, where the fraction of non-prompt candidates is now chosen to be high, while the fraction of prompt candidates is low. These procedures yield the working points and significances shown in table 7 and table 8.

Table 7: Working points selected for each p_T interval for the prompt signal, as well as the significances gained from fits.

p_T [GeV/c]	$0-2$	$2-4$	$4-6$	$6-12$
Prompt selection	0.72	0.70	0.66	0.65
Background selection	0.3	0.3	0.35	0.35
Significance	130.33	179.47	146.64	110.01

Table 8: Working points selected for each p_T interval for the non-prompt signal, as well as the significances gained from fits.

p_T [GeV/c]	$0-2$	$2-4$	$4-6$	$6-12$
Non-prompt selection	0.12	0.12	0.11	0.10
Background selection	0.9	0.9	0.9	0.9
Significance	68.43	78.28	63.49	52.19

4.5 Results

After selecting the woring points, the cuts on the BDT outputs are applied to the data. The resulting invariant mass distributions of the four transeverse momentum intervals are then fit with the same procedure as in section 4.3.4, which is shown in fig. 22 for prompt and in fig. 23 for non-prompt J/ψ signal. As before, the like sign invariant mass distribution is first subtracted from the unlike sign distribution and the residual distribution is then fit with the combination of a Crystal Ball function for the signal and a second order polynomial for the remaining background. The parameters α and n of the Crystal Ball function are fixed to the values gained from fitting the MC mass distributions. It is especially important to fix these quantities in the case of the non-prompt signal, since the fit otherwise does not converge correctly, due to the unexpectedly large amount of entries on the left side of the signal peak. The total fit is shown in red, while the

signal fit and the background fit are plotted in green and blue, respectively. Additionally, the number of J/ψ (S) as well as the amount of background (B) in the signal region ($2.7 < m_{ee} < 3.2$ GeV/c^2) is computed, from which the signal-to-background ratio $\frac{S}{B}$ and significance $\frac{S}{\sqrt{S+2B}}$ are calculated. In table 9, a summary of all calculated quantities, as well as the values for the mean and width of the peak retrieved from the fits is shown for all p_T intervals and both prompt and non-prompt signal.

Table 9: Summary of quantities calculated by the fits for prompt and non-prompt J/ψ.

p_T [GeV/c]	$0-2$	$2-4$	$4-6$	$6-12$
P Signal S	37856 ± 528	47389 ± 409	25676 ± 246	13748 ± 164
P Background B	23257 ± 466	11166 ± 329	2492 ± 180	935 ± 110
P S/B	1.628 ± 0.020	4.244 ± 0.096	10.303 ± 0.659	14.710 ± 1.581
P Significance	130.33	179.47	146.64	110.01
P Mean [GeV/c^2]	3.0955 ± 0.0010	3.0878 ± 0.0007	3.0792 ± 0.0010	3.0650 ± 0.0015
P Width [GeV/c^2]	0.0329 ± 0.0007	0.0401 ± 0.0006	0.0534 ± 0.0010	0.0672 ± 0.0015
NP Signal S	15769 ± 534	13649 ± 316	6779 ± 159	4227 ± 117
NP Background B	18667 ± 501	8377 ± 279	2311 ± 128	1166 ± 91
NP S/B	0.845 ± 0.015	1.629 ± 0.028	2.934 ± 0.107	3.624 ± 0.201
NP Significance	68.43	78.28	63.49	52.19
NP Mean [GeV/c^2]	3.0746 ± 0.0034	3.0751 ± 0.0023	3.0690 ± 0.0029	3.0572 ± 0.0039
NP Width [GeV/c^2]	0.0549 ± 0.0035	0.0515 ± 0.0023	0.0623 ± 0.0027	0.0823 ± 0.0038

For the prompt signal, the number of J/ψ first increases from the lowest p_T interval to the next and then decreases for the remaining intervals, which matches the trend in fig. 6. On the other hand, the number of non-prompt J/ψ decreases with increasing p_T. This is not compatible with the expectation (see fig. 6), since a lower value for the lowest p_T interval is expected. The signal-to-background ratio for prompt and non-prompt J/ψ increases with rising transverse momentum. Furthermore, the significances are overall high, but decrease for rising p_T after an initial increase from the lowest p_T interval. However, the mean μ of the fitted functions in general shifts to lower values and the width σ increases for higher p_T intervals, due to the peak shifting and widening in the mass distribution, which the fit function matches. Particles with higher momenta show tracks with less curvature, leading to worse momentum resolution, which could be an explanation for the shifting and widening. The PDG reference value for the mean is $m_{J/\psi} = (3096.900 \pm 0.006)$ MeV/c^2, for which only the prompt value from the lowest transverse momentum interval is in a 3σ range [31].

The quality of the residual mass distributions and fits should also be addressed. In the case of the prompt signal fits, the background is significantly reduced compared to the invariant mass distributions before the application of ML. This trend continues with respect to the peak height with rising transverse momentum. Therefore, the invariant mass distributions foremostly show the invariant mass distribution of the J/ψ with the bremsstrahlung tail. As a result, the signal fit lies close to the total fit, hence it can be concluded that the reduction of background works well for prompt signal with the specific working points applied. However, it should be noted that the number of J/ψ calculated from these fits is significantly lower than before the usage of the

BDTs. The reason for this is twofold: First, the numbers from the first fits included both prompt and non-prompt J/ψ, and second, the chosen working points are most likely cutting away a significant amount of real J/ψ as they are chosen quite harshly. Nevertheless, the resulting fits are overall satisfactory.

On the other hand, the invariant mass distributions for non-prompt selections indicate some problems with the BDT models or data. While the background overall is reduced with respect to the peak heights compared to before the models were applied, there is also an increased amount of entries to the left side of the mass peak. This effect is present for all p_T intervals, but is dampened with increasing p_T. Since the background decreases with rising p_T for the prompt cuts, it is very likely that the increased amount of entries here also stems from background that is not filtered out by the BDT model. Further increasing the harshness of the non-prompt and background selection does not lead to any meaningful reduction of the increased background on the left side. The effect is also observed in the J/ψ sidebands, where the distributions of the training features after application of the non-prompt selections reveal that the remaining background might resemble the non-prompt J/ψ to such a degree, that it is incorrectly classified as such. It is ultimately unknown what causes this high remaining background, but a fraction of it might stem from correlated background from semileptonic decays of charm and beauty hadrons [32, 33].

Due to the distorted shape of the non-prompt invariant mass distributions, fitting the Crystal Ball function with polynomial background leads to signal shapes, which differ severly from the prompt fits. Therefore, it is very likely that the values calculated from the fits do not reflect reality correctly. The cause of the effect has to be found and eliminated, in order to correctly measure the non-prompt J/ψ and all associated values via the usage of BDT models in a future analysis.

In this thesis, the real fractions of prompt and non-prompt J/ψ in the prompt and non-prompt signals are not determined and have to be further studied in a future analysis. Considering the high significances and the stringently selected working points to achieve a high fraction of either prompt or non-prompt J/ψ, the analysis of prompt and non-prompt J/ψ in the $J/\psi \to e^+e^-$ decay channel through the usage of BDTs can be feasible in a future anlaysis, as long as the mentioned unexpected residual background in the non-prompt invariant mass distribution can be explained and reduced to an acceptable level. As this thesis shows very promising results for the prompt J/ψ, an anlysis of only prompt J/ψ could be feasible. However, a point of contention is the usage of BDT models that can not fully distinguish between background and non-prompt J/ψ for reasons that have yet to be investigated.

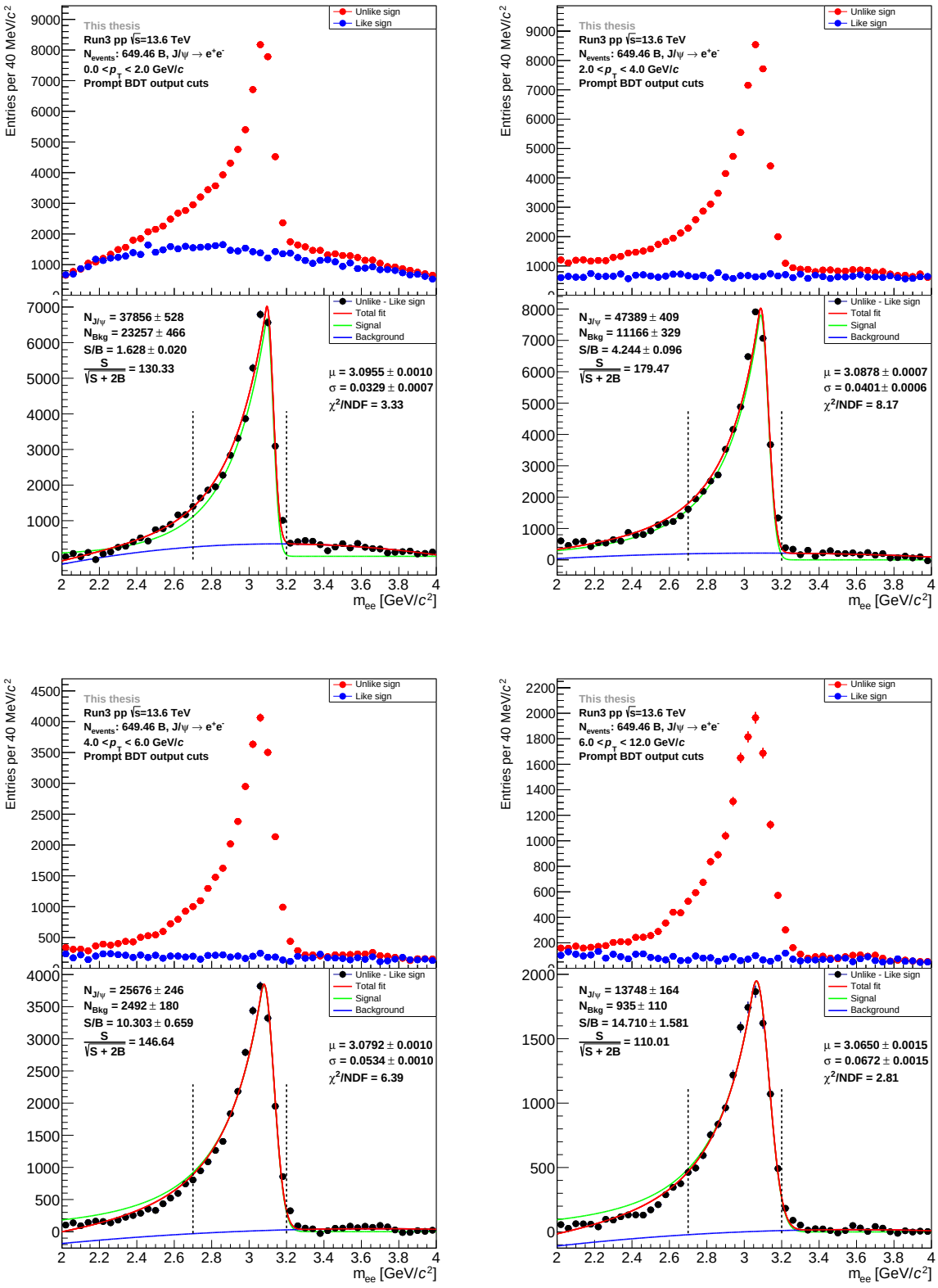

Figure 22: Invariant mass distribution fits for prompt $J/\psi \to e^+e^-$ for all p_T ranges with combinatorial background subtracted.

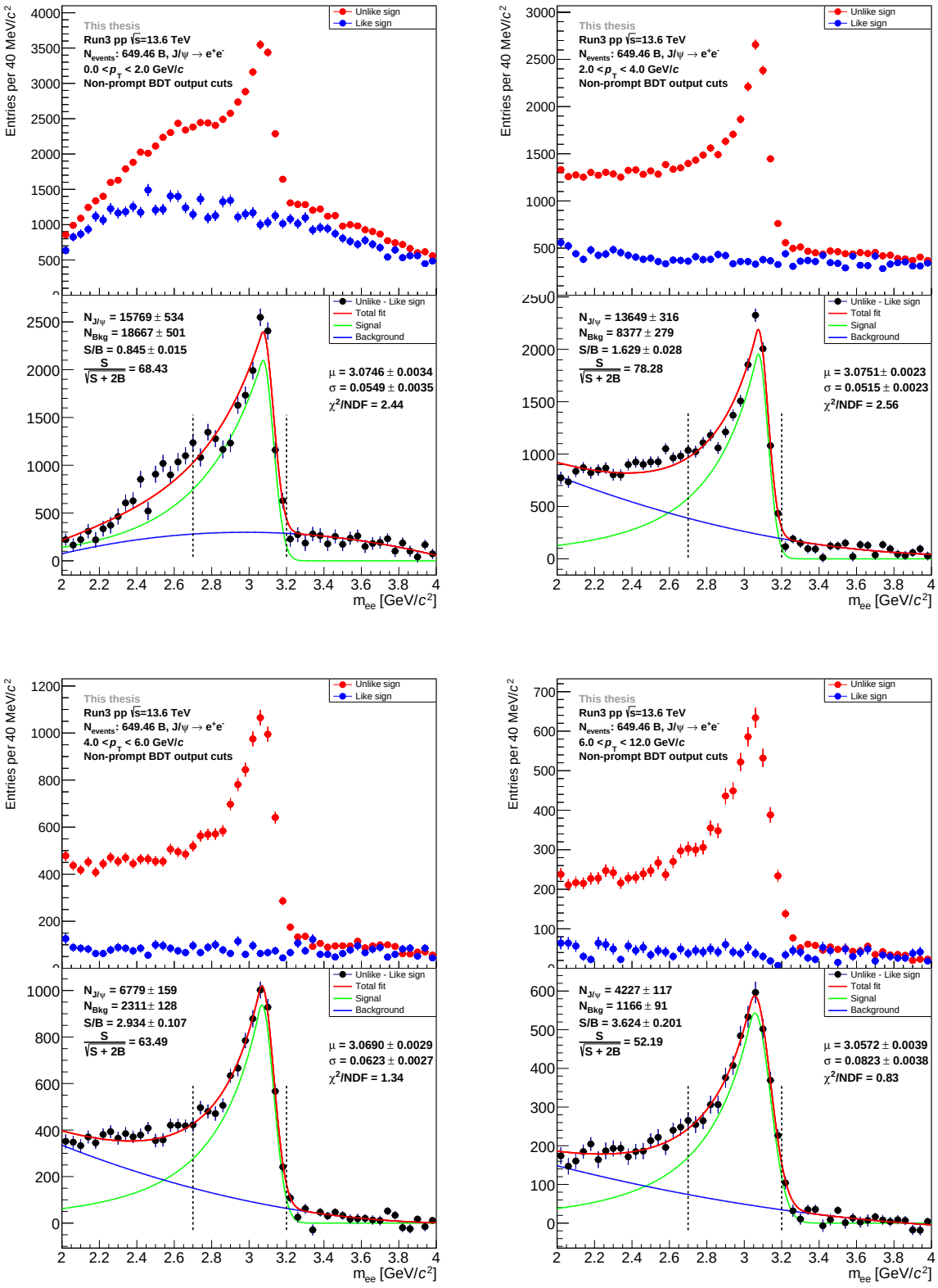

Figure 23: Invariant mass distribution fits for non-prompt $J/\psi \to e^+e^-$ for all p_T ranges with combinatorial background subtracted.

5 Conclusion and outlook

In this thesis, the feasibility of the prompt and non-prompt J/ψ analysis in pp collisions at $\sqrt{s} = 13.6$ TeV with ALICE in Run 3 at midrapidity via the $J/\psi \to e^+e^-$ decay channel has been investigated. The data set of reconstructed J/ψ candidates is split into four intervals of the transverse momentum in the $0 < p_T < 12$ GeV/c range. Suitable features are selected for the differentiation of background, prompt and non-prompt candidates, by comparing the feature distributions for the background data with the distributions for prompt and non-prompt MC data. Using these features, a multiclass BDT model is trained for each separate transverse momentum interval in order to classify the reconstructed candidates into the categories background, prompt and non-prompt. Working points are chosen for the BDT outputs on the basis of either a high prompt or non-prompt fraction and an optimized significance, in order to investigate possible classification problems. The selections are applied to the data and the signal peaks of the invariant mass distributions are fitted for both prompt and non-prompt signal. Exact fractions of prompt and non-prompt signal have not been not calculated, but selections are optimized to ensure high fractions for the respective type of interest, in order to facilitate the judgement of the model capabilities. Significances between 110.01 and 179.47 are found for prompt J/ψ, while for non-prompt J/ψ lower significances between 52.19 and 78.28 are observed. The invariant mass distributions for non-prompt signal show an increased amount of background to the left of the mass peak, of which the origin is not fully understood. Therefore, the analysis of prompt and non-prompt J/ψ using the $J/\psi \to e^+e^-$ decay channel can be deemed feasible, as long as the problem of residual background for non-prompt signal can be resolved. Thus, further investigation is necessary to definitively settle the question of feasibility.

Expanding on the studies of this thesis, a full working point determination should be carried out in order to find the fractions of prompt and non-prompt candidates and the significance of the signal. Furthermore, the exact prompt and non-prompt fractions of the (non-)prompt signal need to be determined in a separate calculation. Additionally, remaining prompt candidates in non-prompt signal need to be removed, while non-prompt candidates are to be subtracted from the prompt signal. In the selection processes, a number of candidates is excluded, which also needs to be accounted for in the calculation of the differential cross section. Due to the fact that a fraction of real prompt and non-prompt J/ψ are not included in the extracted signals, efficiency corrections have to be applied. These efficiency corrections include preselection efficiencies and BDT efficiencies, but also consider the detector acceptance. In order to correctly assess the results, systematic uncertainties also need to be considered. Different bin widths, fit ranges or background fit functions can lead to slightly different signal fits, thus varying the number of extracted J/ψ. Systematic uncertainties also result from the selection of working points, which can be estimated through variation of the chosen strictness.

The successful continuation of this analysis will simultaneously allow for an application of these analysis methods in heavy-ion collisions, which can lead to improvements in the measured J/ψ and beauty hadron production cross sections in Pb-Pb collisions. This will ultimately contribute to an improved understanding of the QGP and its evolution, as well as the hadronization of

beauty quarks. On the other hand, the analysis of a variety of other charm and beauty hadron decays will be required, to fully grasp the mechanisms behind hadronization and the evolution of the QGP.

6 Appendix

6.1 Feature distributions

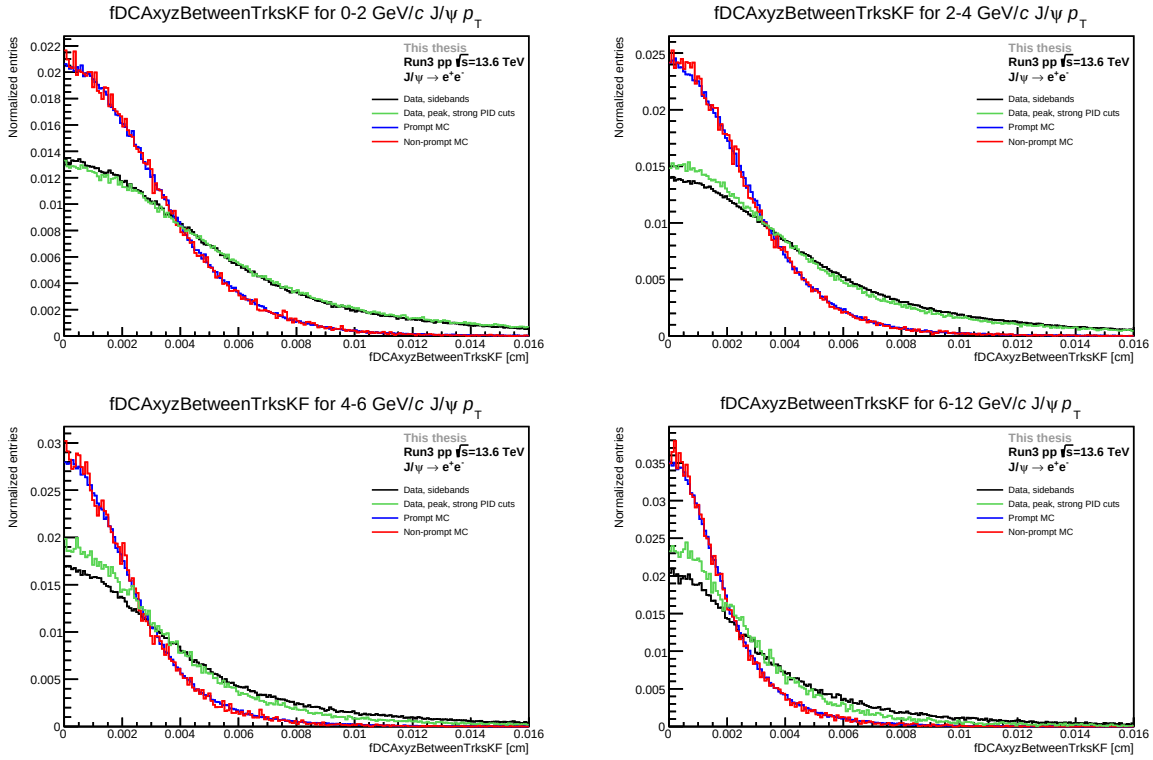

Figure 24: Distribution of DCAxyzBetweenTrksKF for data in the sideband regions, prompt MC, non-prompt MC and data in the peak region with $n_{\sigma^{\pi}_{\mathrm{TPC}}} > 4.0$ and $n_{\sigma^{p}_{\mathrm{TPC}}} > 4.0$ cuts applied in all p_T ranges.

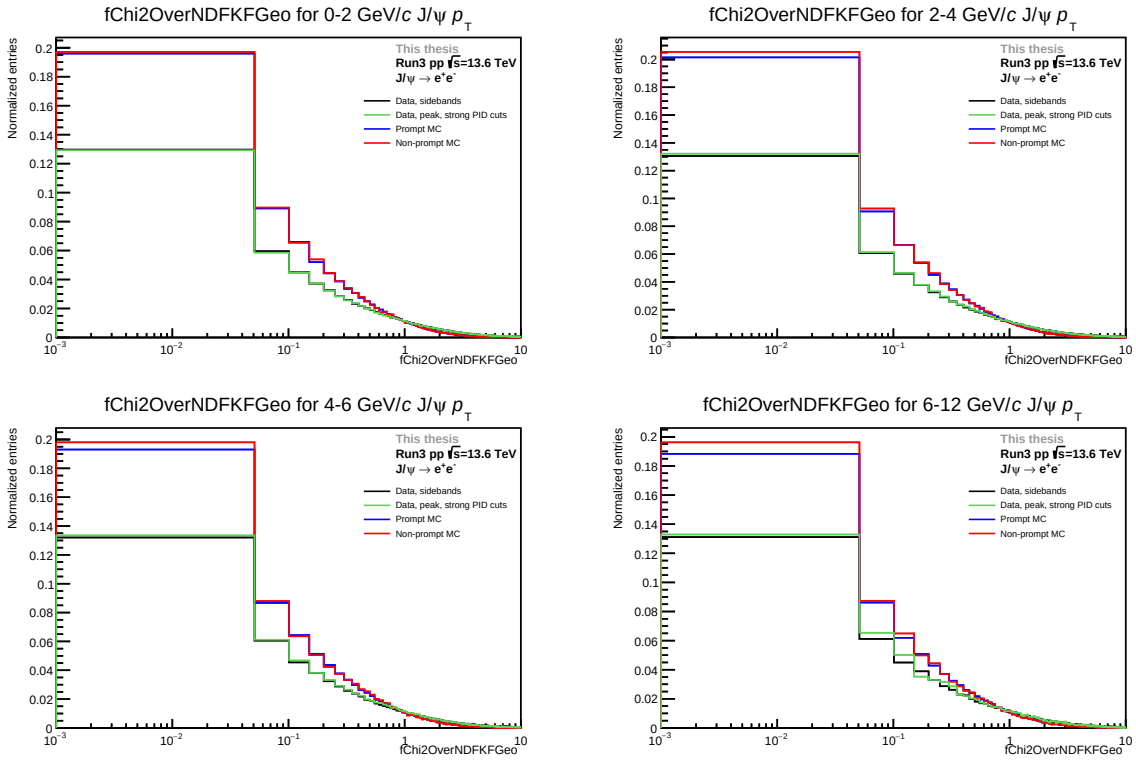

Figure 25: Distribution of Chi2OverNDFKFGeo for data in the sideband regions, prompt MC, non-prompt MC and data in the peak region with $n_{\sigma^{\pi}_{\mathrm{TPC}}} > 4.0$ and $n_{\sigma^{p}_{\mathrm{TPC}}} > 4.0$ cuts applied in all p_T ranges.

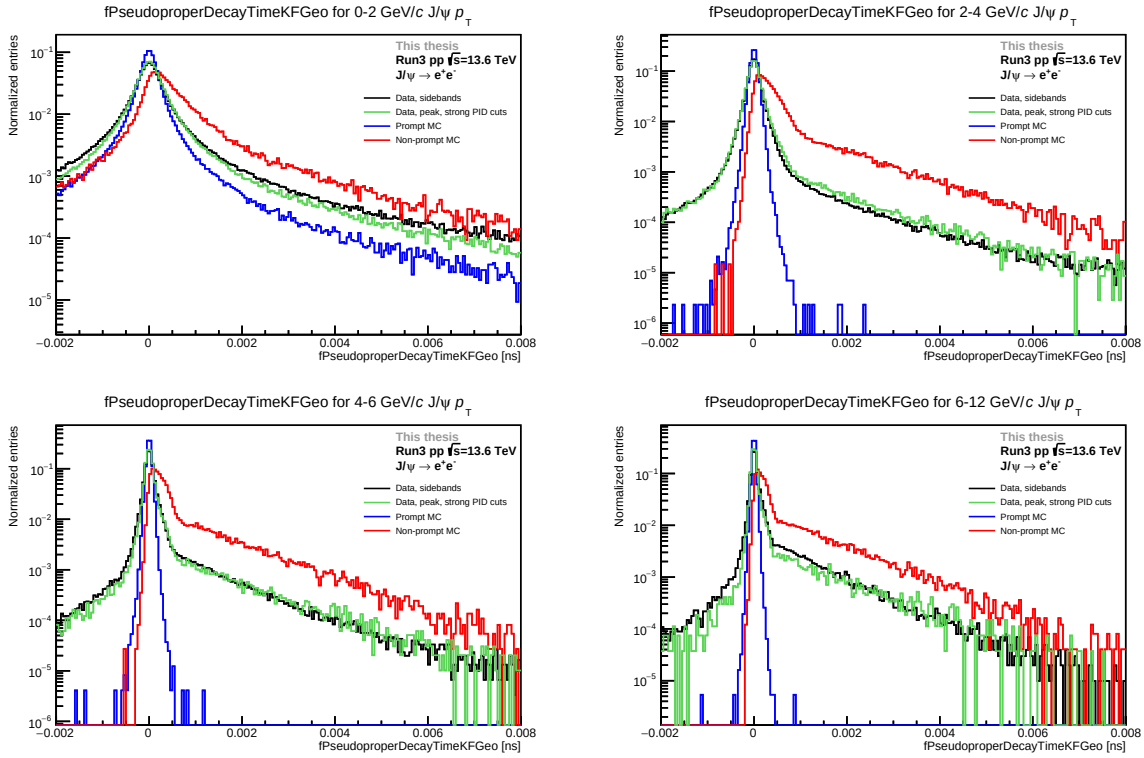

Figure 26: Distribution of PseudoproperDecayTimeKFGeo for data in the sideband regions, prompt MC, non-prompt MC and data in the peak region with $n_{\sigma^{\pi}_{\mathrm{TPC}}} > 4.0$ and $n_{\sigma^{p}_{\mathrm{TPC}}} > 4.0$ cuts applied in all p_T ranges.

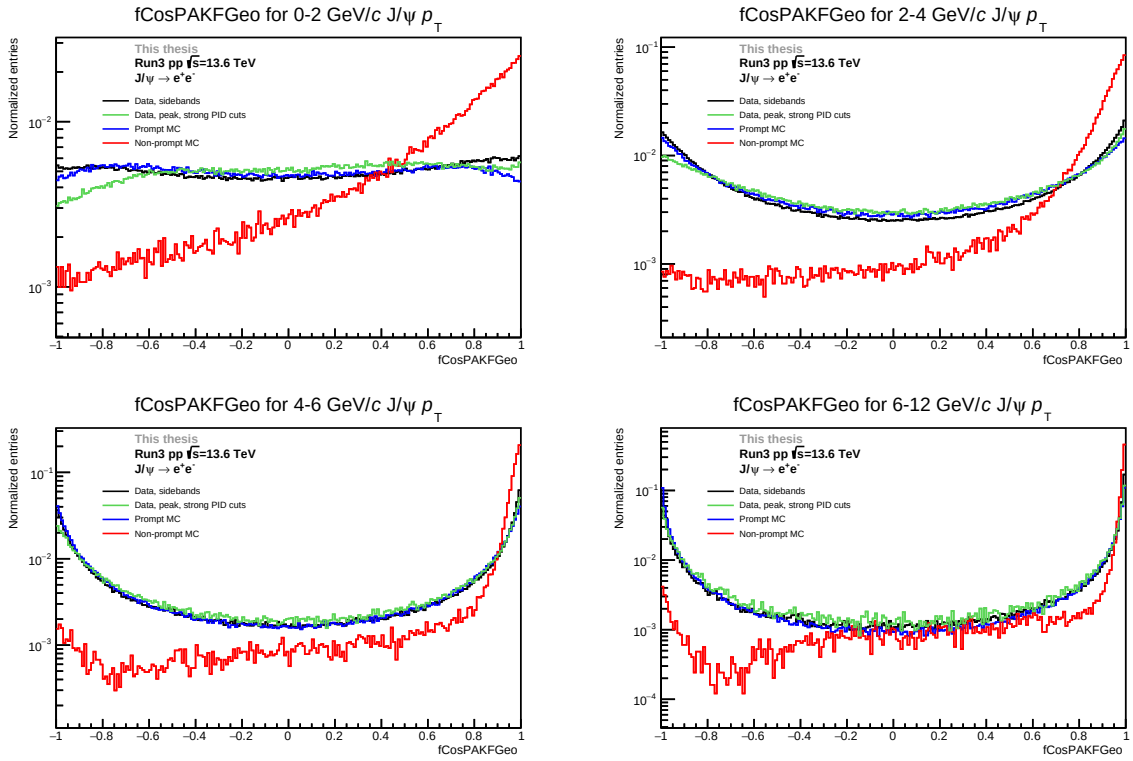

Figure 27: Distribution of CosPAKFGeo for data in the sideband regions, prompt MC, non-prompt MC and data in the peak region with $n_{\sigma^{\pi}_{\mathrm{TPC}}} > 4.0$ and $n_{\sigma^{p}_{\mathrm{TPC}}} > 4.0$ cuts applied in all p_T ranges.

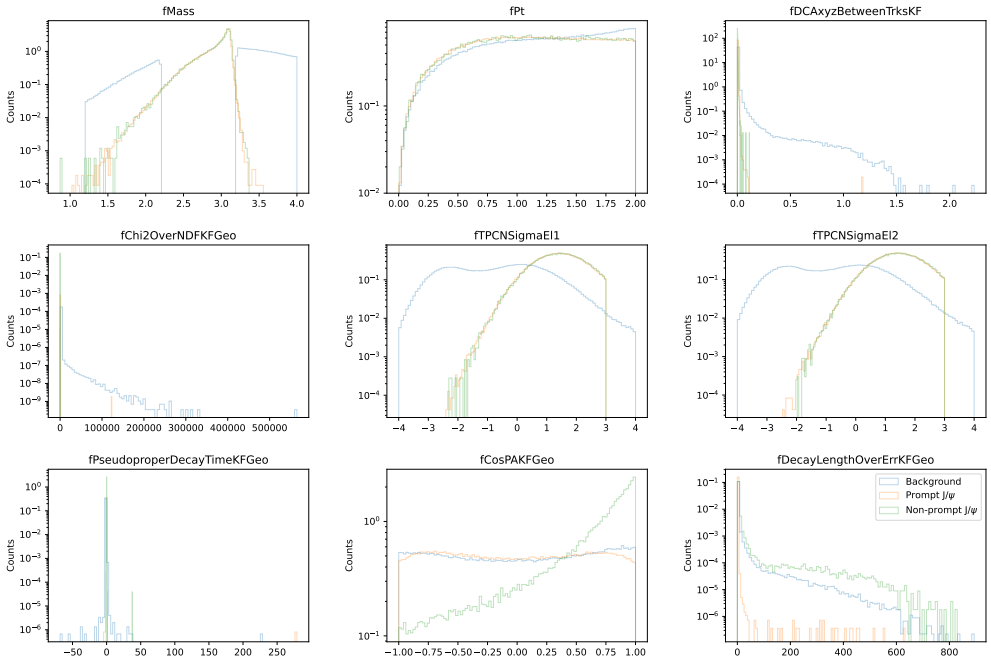

Figure 28: Distributions of the used features, invariant mass m and transverse momentum p_T in the $0 < p_T < 2$ GeV/c range.

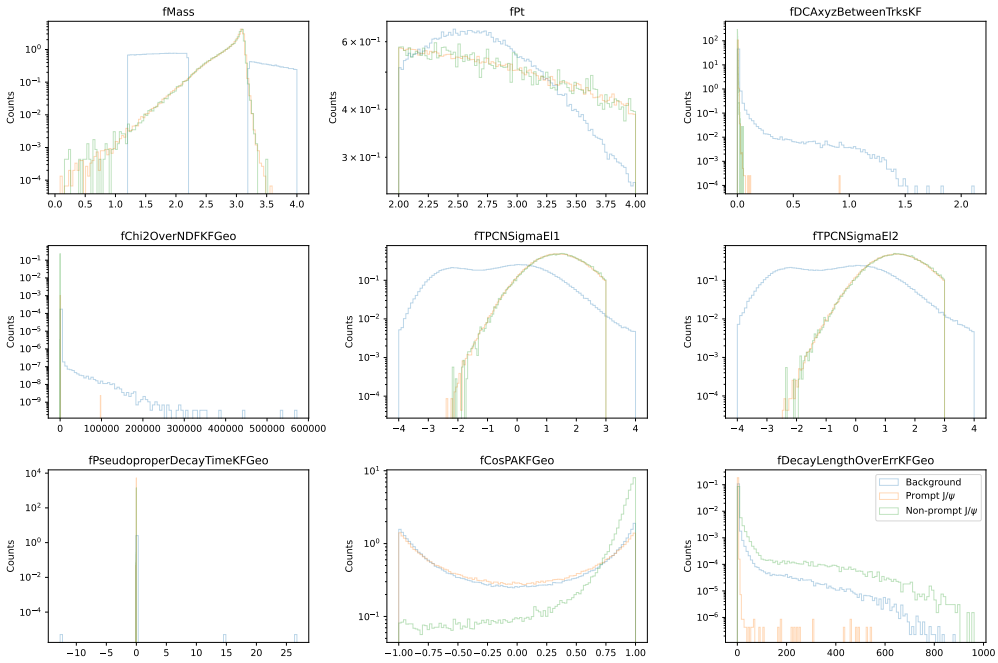

Figure 29: Distributions of the used features, invariant mass m and transverse momentum p_T in the $2 < p_T < 4$ GeV/c range.

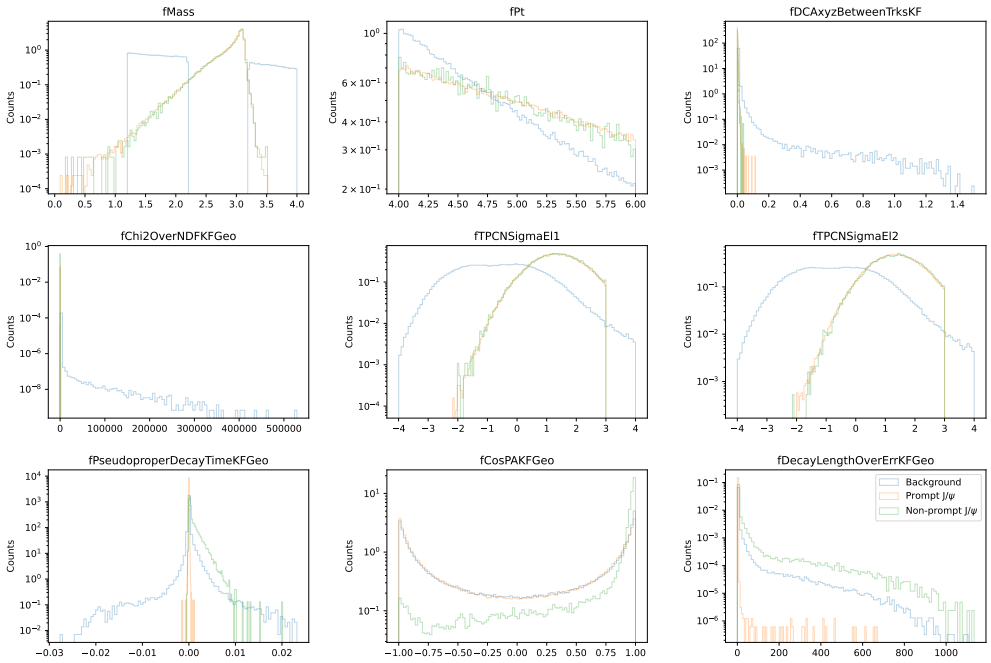

Figure 30: Distributions of the used features, invariant mass m and transverse momentum p_T in the $4 < p_T < 6$ GeV/c range.

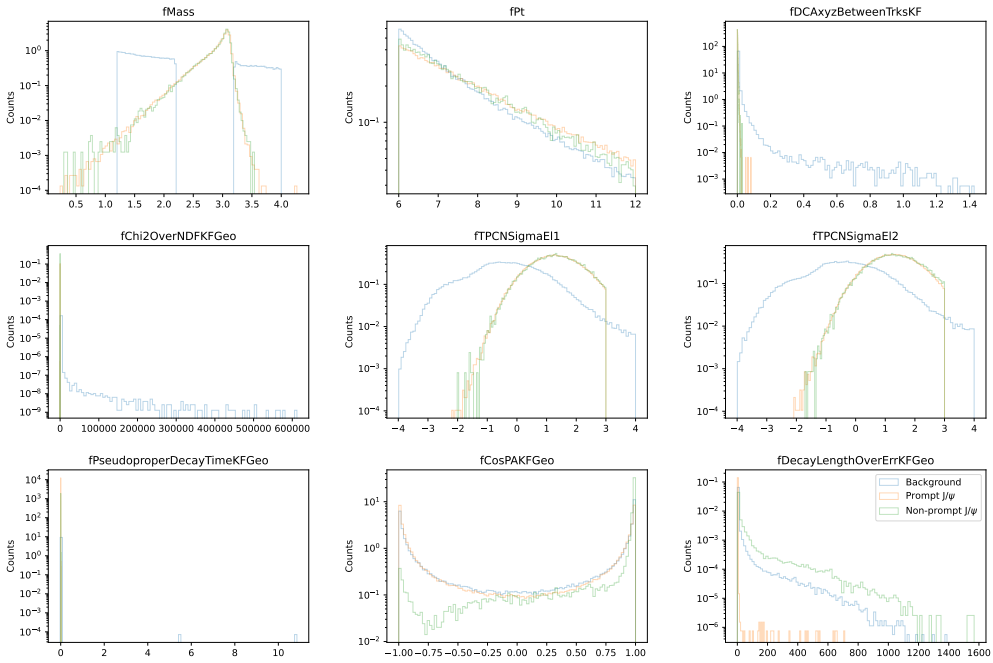

Figure 31: Distributions of the used features, invariant mass m and transverse momentum p_T in the $6 < p_T < 12$ GeV/c range.

6.2 Correlation matrices

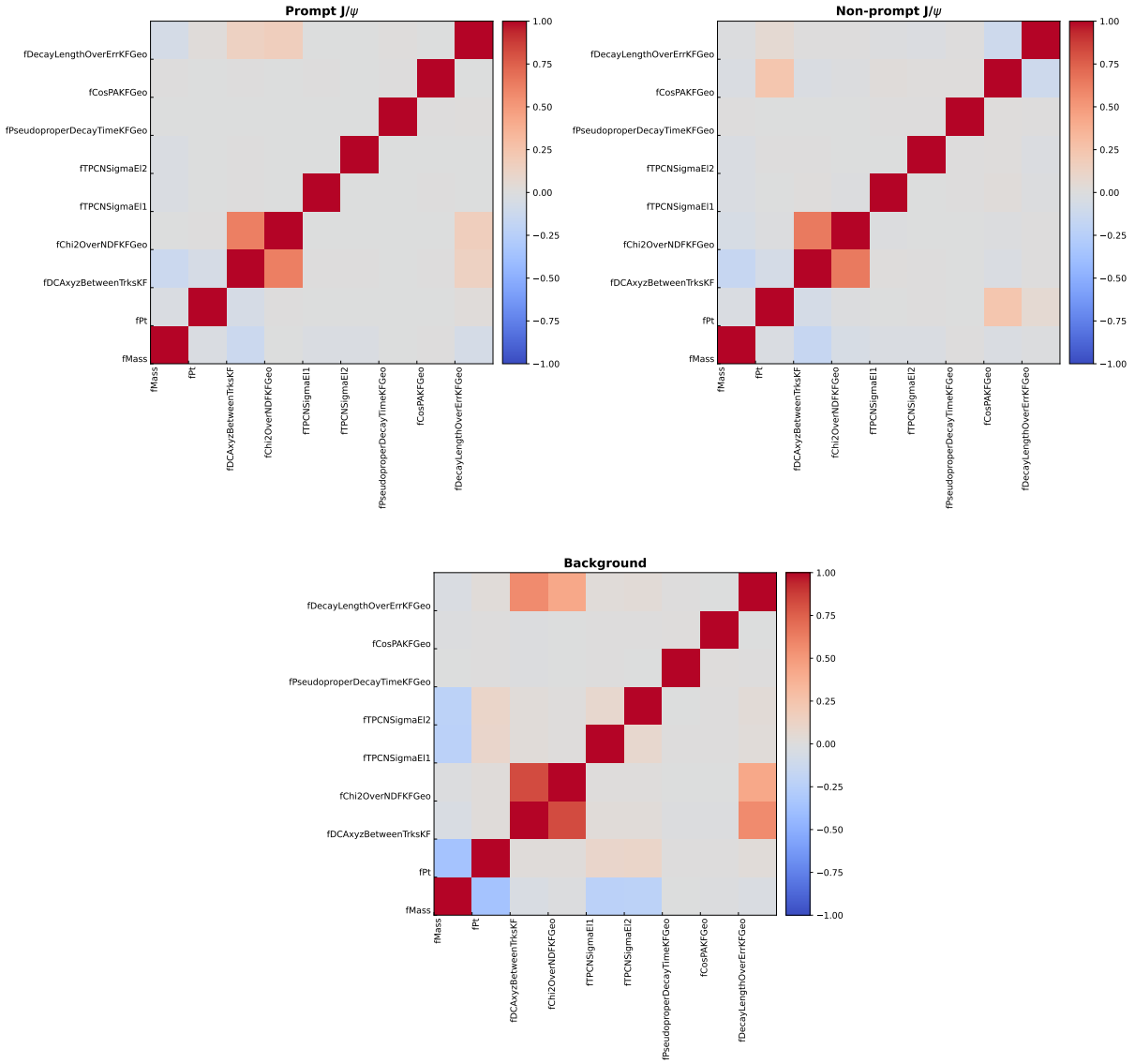

Figure 32: Correlation matrices of the used features, invariant mass m and transverse momentum p_T for the three classes in the $0 < p_T < 2$ GeV/c range.

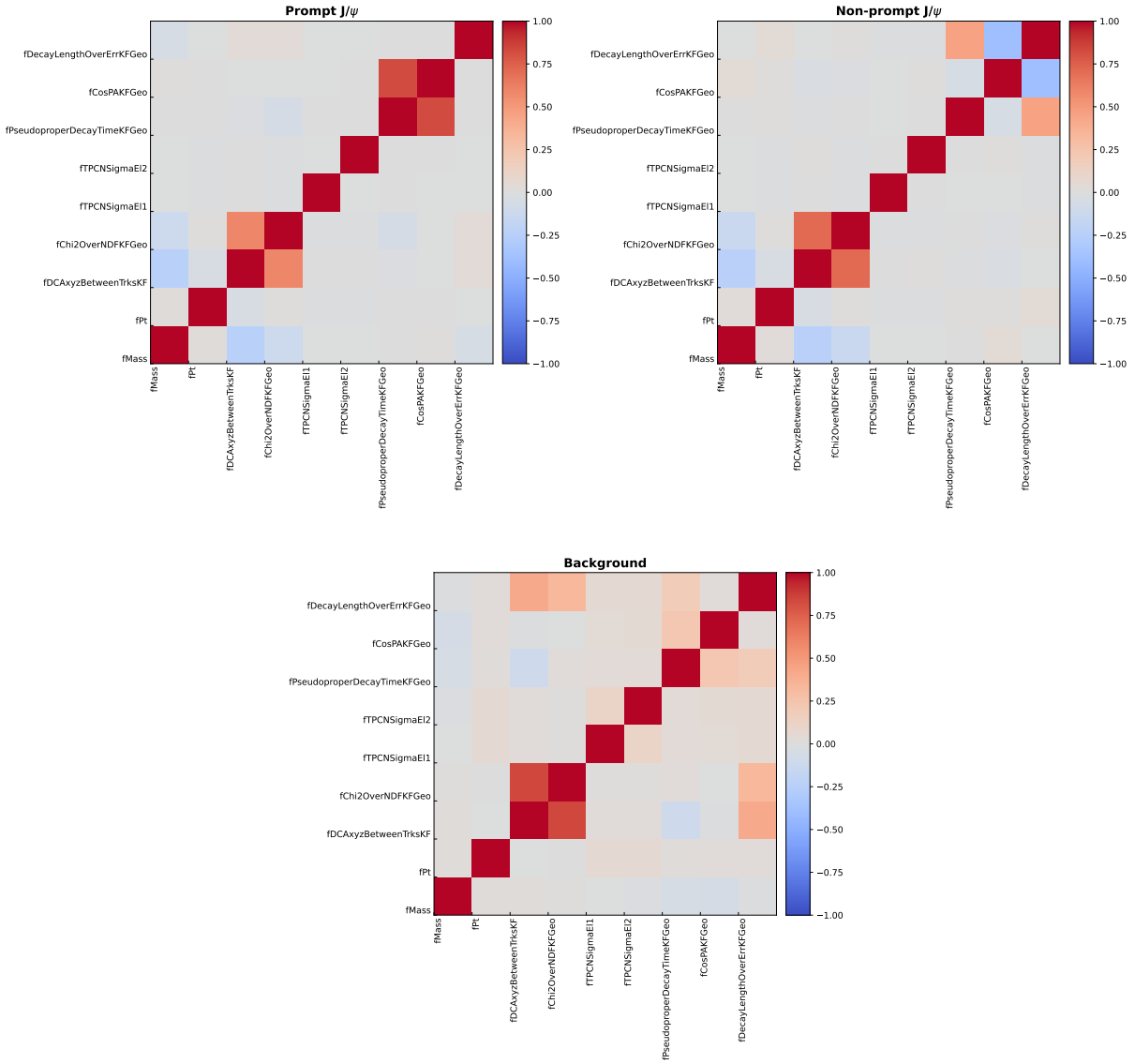

Figure 33: Correlation matrices of the used features, invariant mass m and transverse momentum p_T for the three classes in the $4 < p_T < 6$ GeV/c range.

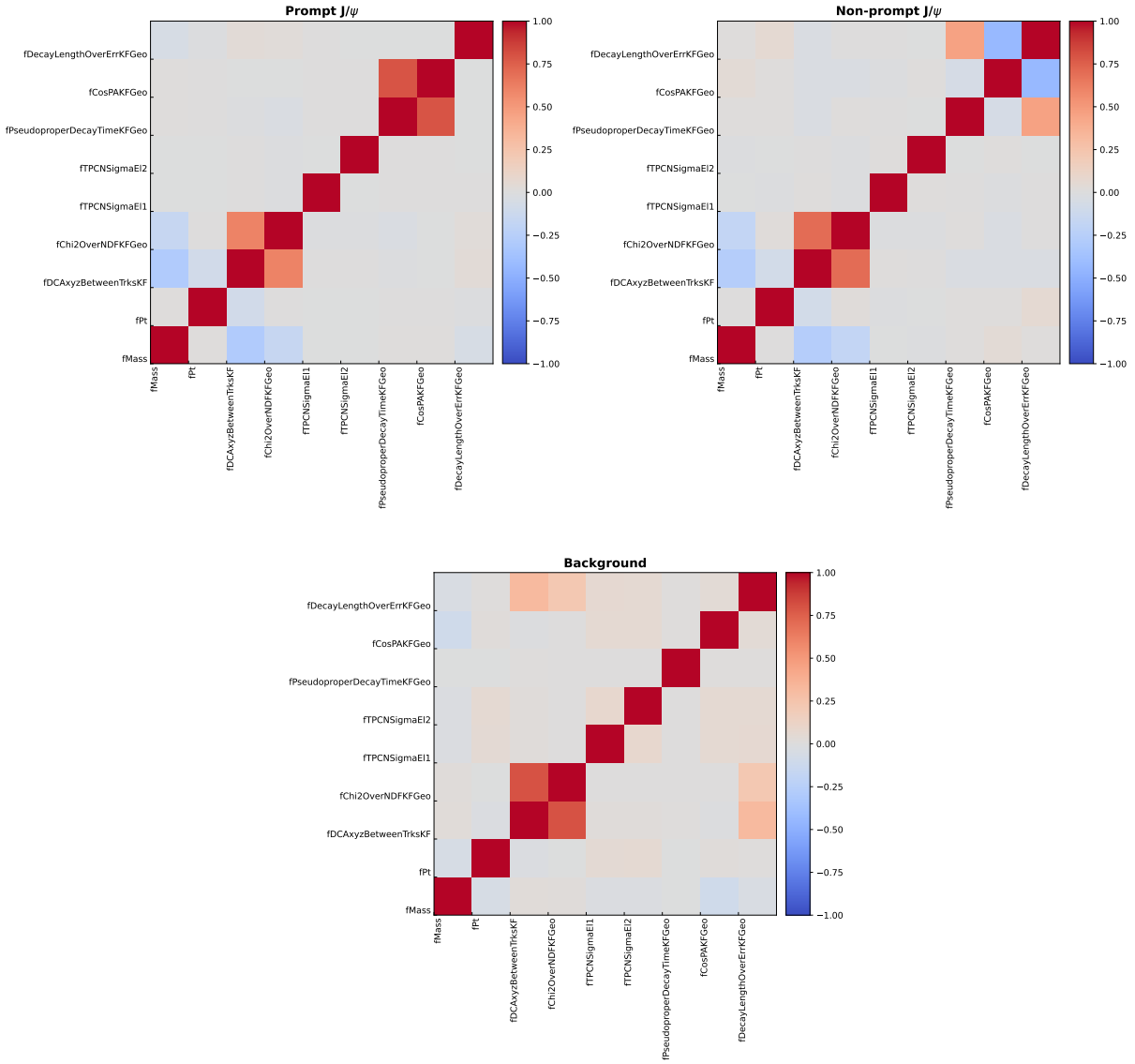

Figure 34: Correlation matrices of the used features, invariant mass m and transverse momentum p_T for the three classes in the $6 < p_T < 12$ GeV/c range.

6.3 Feature importance

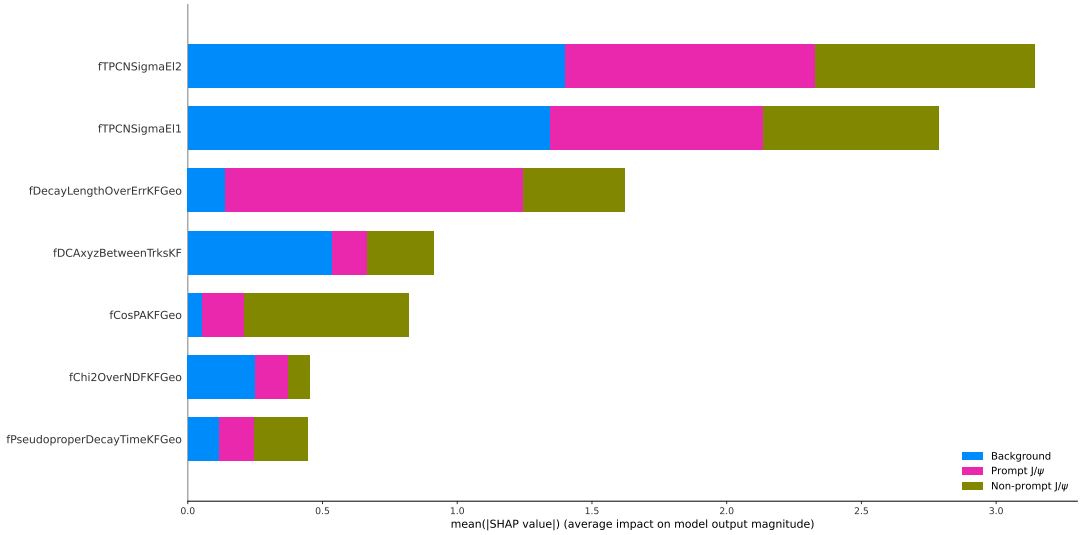

Figure 35: Impact of the used features on the model for each class measured by the mean SHAP value. These are the feature importances of the model for the $0 < p_T < 2$ GeV/c range.

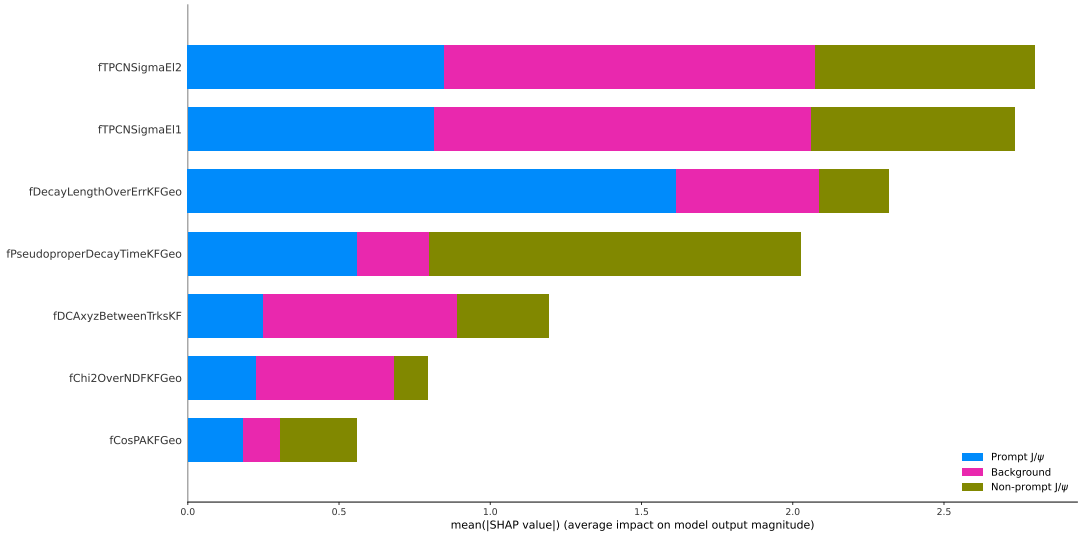

Figure 36: Impact of the used features on the model for each class measured by the mean SHAP value. These are the feature importances of the model for the $4 < p_T < 6$ GeV/c range.

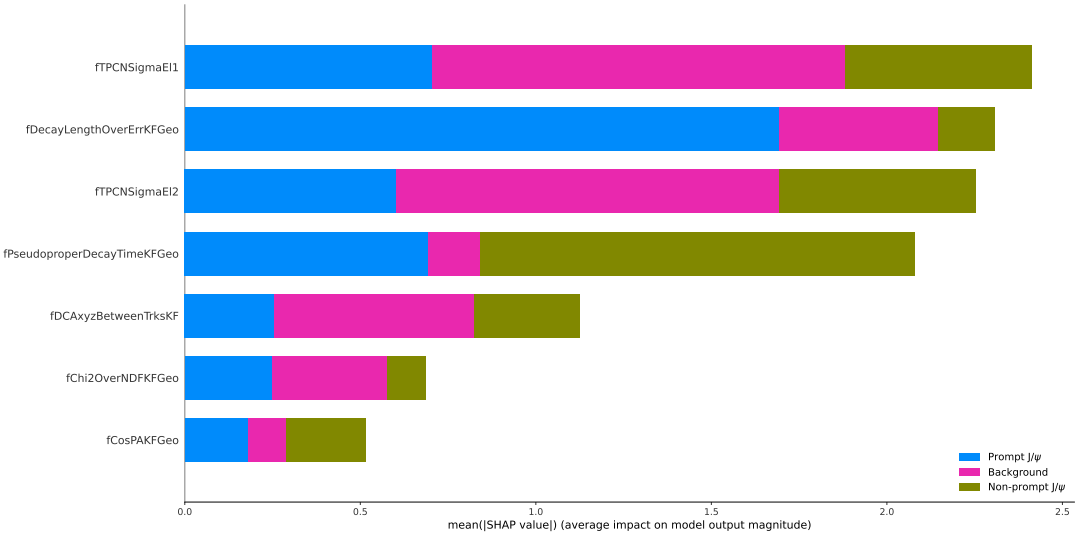

Figure 37: Impact of the used features on the model for each class measured by the mean SHAP value. These are the feature importances of the model for the $6 < p_T < 12$ GeV/c range.

6.4 ROC curves

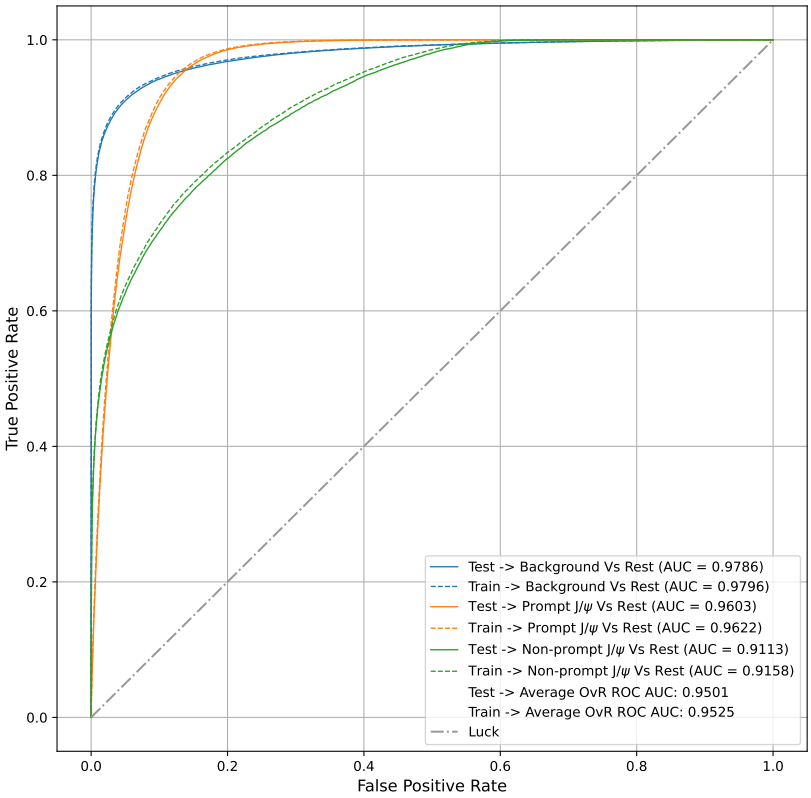

Figure 38: ROC curves and their AUC values of the model for the $0 < p_T < 2$ GeV/c range.

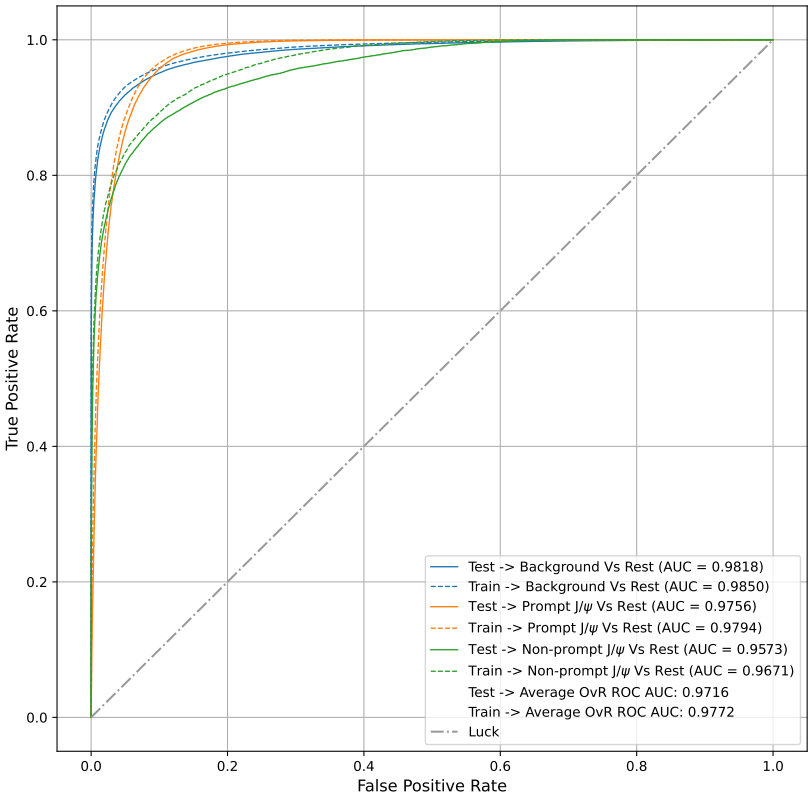

Figure 39: ROC curves and their AUC values of the model for the $4 < p_T < 6$ GeV/c range.

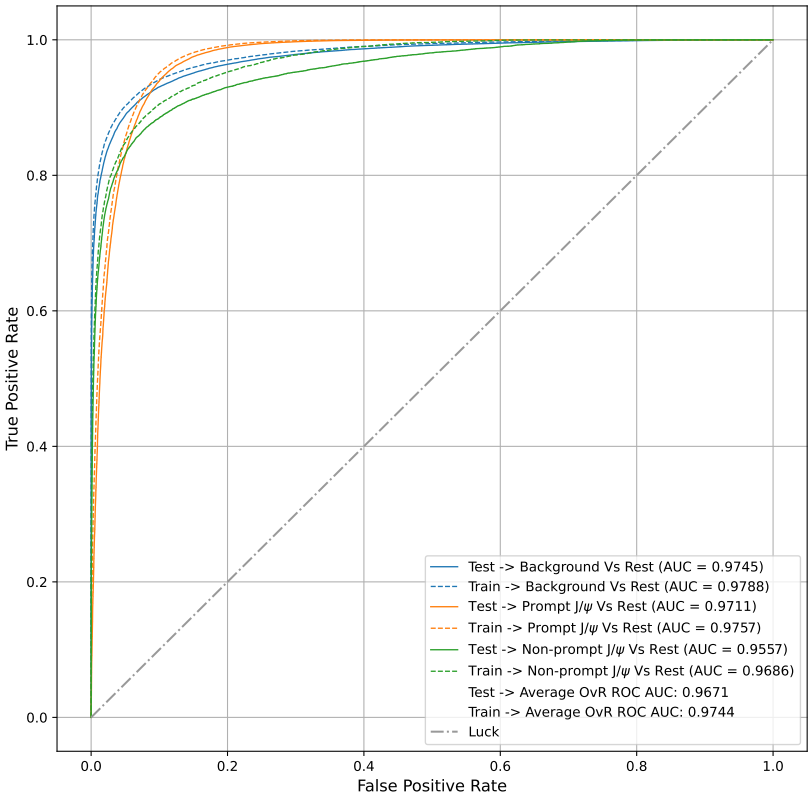

Figure 40: ROC curves and their AUC values of the model for the $6 < p_T < 12$ GeV/c range.

6.5 BDT outputs

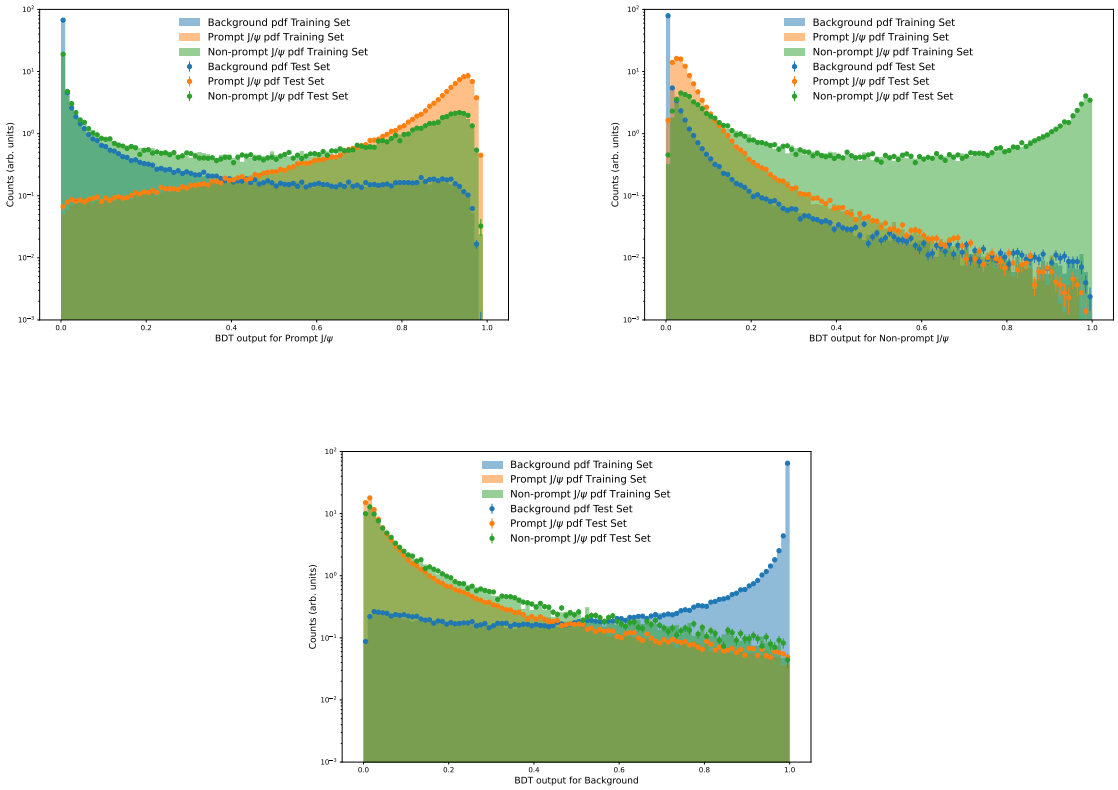

Figure 41: BDT outputs for the three classes with training and test data in the $0 < p_T < 2$ GeV/c range.

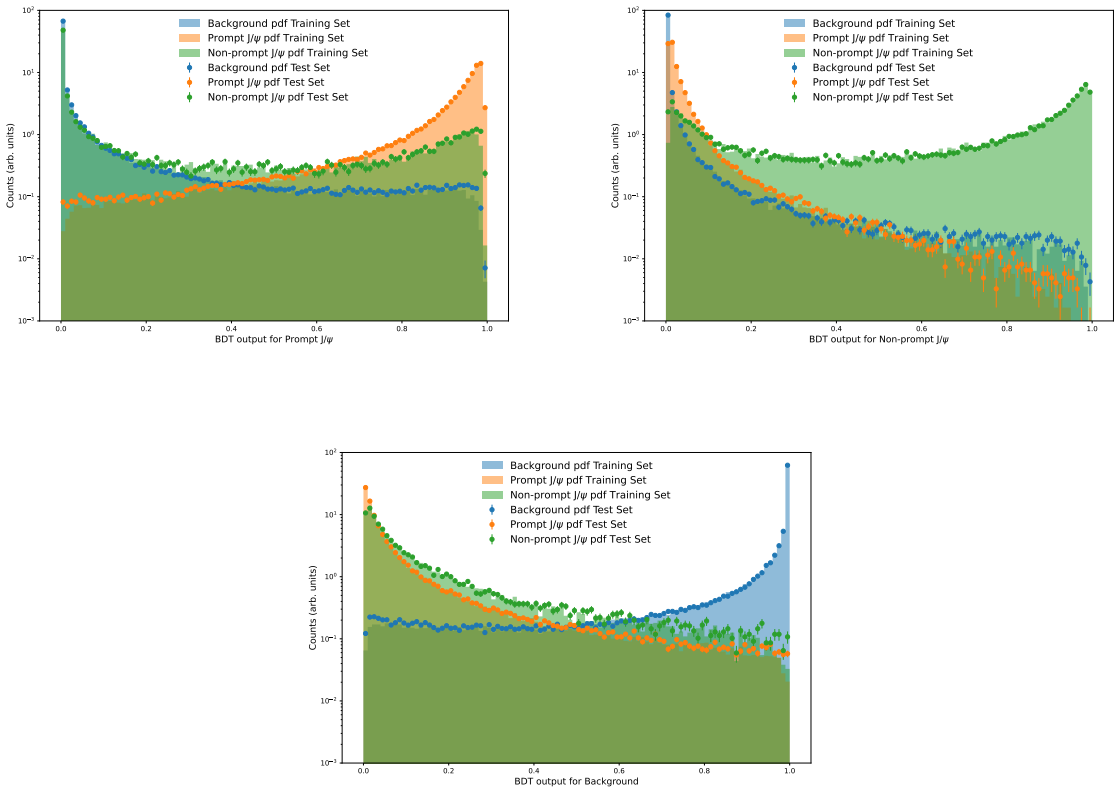

Figure 42: BDT outputs for the three classes with training and test data in the $4 < p_T < 6$ GeV/c range.

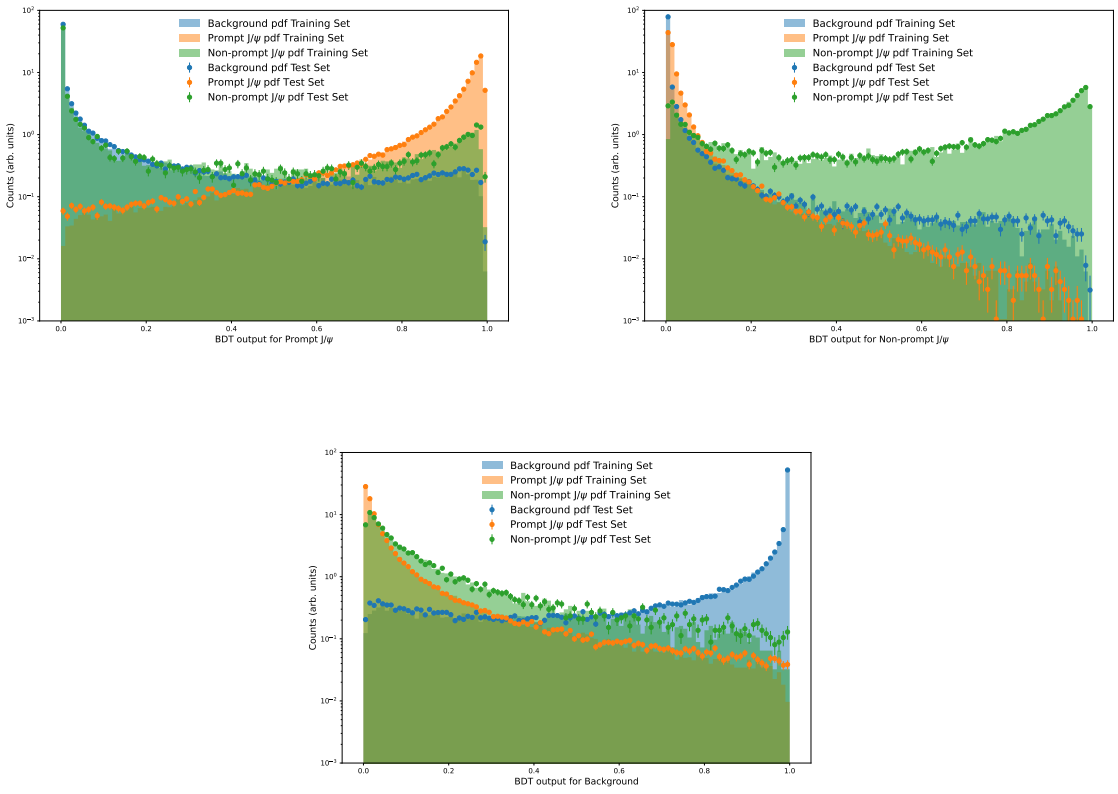

Figure 43: BDT outputs for the three classes with training and test data in the $6 < p_T < 12$ GeV/c range.

List of Acronyms

ALICE	A Large Ion Collider Experiment
ALPIDE	ALICE Pixel Detector
APD	Avalanche Photodiode
AUC	Area Under Curve
BDT	Boosted Decision Tree
CART	Classification and regression trees
CBM	Compressed Baryonic Matter
CERN	Conseil Européen pour la Recherche Nucléaire (European Organization for Nuclear Research)
CMS	Compact Muon Solenoid
DCA	Distance of Closest Approach
EMCal	ElectroMagnetic Calorimeter
FIT	Fast Interaction Trigger
GEM	Gas Electron Multiplier
GPU	Graphics Processing Unit
HIC	Hybrid Integrated Circuit
hipe4ml	Minimal heavy ion physics environment for Machine Learning
HMPID	High Momentum Particle Identification Detector
IB	Inner Barrel
IROC	Inner Readout Chamber
ITS	Inner Tracking System
LEP	Large Electron-Positron Collider
LHC	Large Hadron Collider
LS2	Long Shutdown 2
MAPS	Monolithic Active Pixel Sensors
MC	Monte Carlo
MCH	Muon tracking Chambers
MFT	Muon Forward Tracker
MID	Muon Identifier
ML	Machine Learning
MRPC	Multi-gap Resistive-Plate Chamber
MWPC	Multiwire Proportional Chamber
NP	Non-prompt
O^2	Software framework used for online and offline reconstruction and physics analysis in Run 3

OB	Outer Barrel
OROC	Outer Readout Chamber
OvO	One-vs-One
OvR	One-vs-Rest
P	Prompt
Pb-Pb	Lead-Lead
PDG	Particle Data Group
PHOS	PHOton Spectrometer
PID	Particle Identification
pp	proton-proton
PV	Primary Vertex
QCD	Quantum Chromodynamics
QGP	Quark-Gluon Plasma
RHIC	Relativistic Heavy Ion Collider
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SM	Standard Model of particle physics
SPS	Super Proton Synchrotron
SV	Secondary Vertex
TOF	Time-of-Flight detector
TPC	Time Projection Chamber
TRD	Transition Radiation Detector
XGBoost	eXtreme gradient Boosting
ZDC	Zero-Degree Calorimeters

References

[1] S. Acharya et al. „The ALICE experiment: a journey through QCD“. In: *The European Physical Journal C* 84.8 (Aug. 2024). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-024-12935-y. URL: http://dx.doi.org/10.1140/epjc/s10052-024-12935-y.

[2] Mark Thomson. *Modern Particle Physics*. Cambridge, United Kingdom: Cambridge University Press, 2013.

[3] Andrew Purcell. „Go on a particle quest at the first CERN webfest. Le premier webfest du CERN se lance à la conquête des particules“. In: 35/2012 (2012), p. 10. URL: https://cds.cern.ch/record/1473657.

[4] Piotr Traczyk. *The LHC lead-ion collision run starts*. Sept. 2023. URL: https://home.cern/news/news/experiments/lhc-lead-ion-collision-run-starts.

[5] CERN webpage. *Accelerator upgrades during LS2*. 2022. URL: https://home.cern/press/2022/accelerator-upgrades-during-ls2.

[6] Rende Steerenberg. *Accelerator Report: Getting lead ions ready for physics*. Sept. 2023. URL: https://home.cern/news/news/accelerators/accelerator-report-getting-lead-ions-ready-physics.

[7] Ewa Lopienska. „The CERN accelerator complex, layout in 2022. Complexe des accélérateurs du CERN en janvier 2022“. In: (2022). General Photo. URL: https://cds.cern.ch/record/2800984.

[8] S. Acharya et al. „Prompt and non-prompt J/ψ production cross sections at midrapidity in proton-proton collisions at $\sqrt{s} = 5.02$ and 13 TeV“. In: *Journal of High Energy Physics* 2022.3 (Mar. 2022). ISSN: 1029-8479. DOI: 10.1007/jhep03(2022)190. URL: http://dx.doi.org/10.1007/JHEP03(2022)190.

[9] ALICE Collaboration. *ALICE upgrades during the LHC Long Shutdown 2*. 2023. arXiv: 2302.01238 [physics.ins-det].

[10] The ALICE Collaboration et al. „The ALICE experiment at the CERN LHC“. In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08002. DOI: 10.1088/1748-0221/3/08/S08002. URL: https://dx.doi.org/10.1088/1748-0221/3/08/S08002.

[11] In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 881 (Feb. 2018), pp. 88–127. ISSN: 0168-9002. DOI: 10.1016/j.nima.2017.09.028. URL: http://dx.doi.org/10.1016/j.nima.2017.09.028.

[12] F. Reidt. „Upgrade of the ALICE ITS detector“. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1032 (2022), p. 166632. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2022.166632. URL: https://www.sciencedirect.com/science/article/pii/S0168900222002042.

[13] „Performance of the ALICE experiment at the CERN LHC“. In: *International Journal of Modern Physics A* 29.24 (2014), p. 1430044. DOI: 10.1142/S0217751X14300440. eprint: https://doi.org/10.1142/S0217751X14300440. URL: https://doi.org/10.1142/S0217751X14300440.

[14] Philip Hauer. „The upgraded ALICE TPC“. In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1039 (2022), p. 167023. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2022.167023. URL: https://www.sciencedirect.com/science/article/pii/S016890022200448X.

[15] ALICE Performance Figures. *ALI-PERF-542396*. URL: https://alice-figure.web.cern.ch/node/26843.

[16] Simon Groß-Bölting. „Measurement of the Λ_c^+ production in proton-proton collisions for $\Lambda_c^+ \to pK_s$ at $\sqrt{s} = 5.02$TeV with the ALICE detector“. Bachelor's Thesis. Heidelberg, Germany: Physikalisches Institut of the University of Heidelberg, Nov. 2021. URL: https://www.physi.uni-heidelberg.de/Publications/BachelorThesis_Simon_Gross-Boelting.pdf.

[17] David Rohr et al. *Track Reconstruction in the ALICE TPC using GPUs for LHC Run 3*. 2018. arXiv: 1811.11481 [physics.ins-det].

[18] 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. URL: https://www.physi.uni-heidelberg.de/Publications/MasterThesis_CarolinaReetz_wo.pdf.

[19] Wolfgang Demtröder. *Experimentalphysik 4 Kern-, Teilchen- und Astrophysik*. Kaiserslautern, Germany: Springer Spektrum, 2017.

[20] I. J. Feng, R. H. Pratt, and H. K. Tseng. „Positron bremsstrahlung“. In: *Phys. Rev. A* 24 (3 Sept. 1981), pp. 1358–1363. DOI: 10.1103/PhysRevA.24.1358. URL: https://link.aps.org/doi/10.1103/PhysRevA.24.1358.

[21] Sergey Gorbunov. „On-line reconstruction algorithms for the CBM and ALICE experiments“. PhD thesis. Goethe U., Frankfurt (main), Frankfurt U., 2013.

[22] Maksym Zyzak. „Online selection of short-lived particles on many-core computer architectures in the CBM experiment at FAIR“. doctoralthesis. Universitätsbibliothek Johann Christian Senckenberg, 2016, p. 165.

[23] Phil Lennart Stahlhut. „Performance test of the KF Particle package for open heavy-flavour baryon reconstruction with ALICE“. Bachelor's Thesis. Heidelberg, Germany: Physikalisches Institut of the University of Heidelberg, June 2023. URL: https://www.physi.uni-heidelberg.de/Publications/BachelorThesis_PhilStalhut.pdf.

[24] *XGBoost Documentation*. URL: https://xgboost.readthedocs.io/en/stable/index.html.

[25] *hipe4ml*. URL: https://github.com/hipe4ml/hipe4ml.

[26] IBM. *What is a decision tree?* URL: https://www.ibm.com/think/topics/decision-trees.

[27] Christian Kleiber. „Feasability study of the non-prompt $\Lambda_c^+ \to pK^-\pi^+$ analysis in p-Pb collisions at $\sqrt{s_{NN}} = 5.02$TeV with ALICE“. Bachelor's Thesis. Heidelberg, Germany: Physikalisches Institut of the University of Heidelberg, Mar. 2023. URL: https://www.physi.uni-heidelberg.de/Publications/BachelorThesis_Christian_Kleiber.pdf.

[28] Gaurov. *An Introduction to Gradient Boosting Decision Trees.* URL: https://www.machinelearningplus.com/machine-learning/an-introduction-to-gradient-boosting-decision-trees/.

[29] *Optuna - An open source hyperparameter optimization framework to automate hyperparameter search.* URL: https://optuna.org/.

[30] Jason Brownlee. *One-vs-Rest and One-vs-One for Multi-Class Classification.* Apr. 2021. URL: https://machinelearningmastery.com/one-vs-rest-and-one-vs-one-for-multi-class-classification/.

[31] S. Navas et al. „Review of Particle Physics“. In: *Phys. Rev. D* 110 (3 Aug. 2024), p. 030001. DOI: 10.1103/PhysRevD.110.030001. URL: https://link.aps.org/doi/10.1103/PhysRevD.110.030001.

[32] S. Acharya et al. „Dielectron and heavy-quark production in inelastic and high-multiplicity proton–proton collisions at $\sqrt{s} = 13$ TeV“. In: *Physics Letters B* 788 (Jan. 2019), pp. 505–518. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2018.11.009. URL: http://dx.doi.org/10.1016/j.physletb.2018.11.009.

[33] S. Acharya et al. „Inclusive J/ψ production at midrapidity in pp collisions at $\sqrt{s} = 13$ TeV“. In: *The European Physical Journal C* 81.12 (Dec. 2021). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-021-09873-4. URL: http://dx.doi.org/10.1140/epjc/s10052-021-09873-4.

Decleration of Authorship

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

Mannheim, den 22.03.2025.