Study of double parton scattering using four-jet scenarios in proton-proton collisions at $\sqrt{s} = 7$ TeV with the CMS experiment at the Large Hadron Collider

Dissertation

zur Erlangung des Doktorgrades
an der Fakultät für Mathematik,
Informatik und Naturwissenschaften
Fachbereich Physik
der Universität Hamburg

Proefschrift voorgelegd tot het behalen van de graad van
doctor in de wetenschappen aan de Universiteit Antwerpen

vorgelegt von
te verdedigen door

PAOLO GUNNELLINI

aus
uit

PERUGIA, ITALIEN

Hamburg
2014

Gutachter/in der Dissertation: Prof. Dr. Johannes Haller
Prof. Dr. Pierre Van Mechelen
Prof. Dr. Mark Strikman

Gutachter/in der Disputation: Prof. Dr. Johannes Haller
Dr. Hannes Jung
Prof. Dr. Pierre Van Mechelen
Prof. Dr. Nick Van Remortel

Datum der Disputation: 20. November 2014

Vorsitzender des Prüfungsausschusses: Prof. Dr. Dieter Horns

Vorsitzende des Promotionsausschusses: Prof. Dr. Daniela Pfannkuche

Dekan des Fachbereichs Physik: Prof. Dr. Heinrich Graener

The story repeats itself...

I would like to dedicate this thesis to my loving parents, who supported me each step of the way, my brother Marco, my sister-in-law Letizia, my niece Maria and my nephews, Pietro and the last comer Michele.

Cerca la felicitá nelle piccole cose e, vedrai, che non è poi così difficile trovarla.

Abstract

Measurements of the differential cross sections for multijet scenarios in proton-proton collisions are presented as a function of the transverse momentum p_{T} and pseudorapidity η, together with the correlations in azimuthal angle and the p_{T} balance among the jets. Two different scenarios are separately studied; in the first one an exclusive four-jet final state is selected in $|\eta| < 4.7$, by requiring two hard jets with $p_{\mathrm{T}} > 50$ GeV each, together with two jets of $p_{\mathrm{T}} > 20$ GeV each. No other jets with $p_{\mathrm{T}} > 20$ GeV are allowed in the selected events. In the second one at least four jets with $p_{\mathrm{T}} > 20$ GeV are required: two of the four selected jets are asked to be originated by a b-quark in $|\eta| < 2.4$, while no requests on the flavour of the other two jets, which are selected within $|\eta| < 4.7$ are applied. The data sample was collected in 2010 at a center-of-mass energy of 7 TeV with the CMS detector at the LHC, with an integrated luminosity of 36 pb^{-1}. The total cross section is measured to be $\sigma(\mathrm{pp} \rightarrow 4\mathrm{j}+\mathrm{X}) = 330 \pm 5$ (stat.)± 45 (syst.) and $\sigma(\mathrm{pp} \rightarrow 2\mathrm{b}+2\mathrm{j}+\mathrm{X}) = 67.2 \pm 2.2$ (stat.) $\pm$ 22.5 (syst.) for, respectively, the two selected multijet scenarios. It is found that fixed-order matrix element calculations including parton showers describe the measured differential cross sections in some regions of phase space only, and that adding contributions from double parton scattering brings the Monte Carlo predictions closer to the data. A new method of extraction of double parton scattering contributions from an experimental measurement is introduced for the first time: it is applied to W+dijet measurements and to both multijet channels. Values of σ_{eff} are measured to be $19.0^{+4.6}_{-3.0}$ mb and $23.2^{+3.3}_{-2.5}$ mb for the two examined selections. These values are consistent between each other and compatible with measurements based on different physics channels at 7 TeV.

Samenvatting

Metingen van de werkzame doorsneden voor multijet scenario's in proton-proton botsingen worden als functie van de transversale impuls p_{T} en pseudorapidity η voorgesteld, samen met de correlaties in azimuthale hoek en de p_{T} balans tussen de jets. Twee verschillende selecties worden afzonderlijk toegepast; de eerste selecteert een exclusieve vier-jet eindtoestand in $|\eta| < 4.7$, door de aanwezigheid van twee harde jets met $p_{\mathrm{T}} > 50$ GeV en twee jets met $p_{\mathrm{T}} > 20$ GeV te eisen. Extra jets met $p_{\mathrm{T}} > 20$ GeV zijn niet toegestaan in de geselecteerde botsingen. De tweede selectie methode vereist ten minste vier jets met $p_{\mathrm{T}} > 20$ GeV, waarvan twee van de vier geselecteerde jets moeten ontstaan zijn vanuit een b-quark in $|\eta| < 2.4$, terwijl er geen flavour voorwaarden zijn voor de andere twee jets die in het gebied $|\eta| < 4.7$ worden geselecteerd. De data werd genomen in 2010 met de CMS detector bij de LHC, en heeft een massamiddelpuntsenergie van 7 TeV, en een geïntegreerde luminositeit van 36 pb^{-1}. De gemeten totale werkzame doorsneden zijn $\sigma(\mathrm{pp} \to 4\mathrm{j}+\mathrm{X}) = 330 \pm 5$ (stat.) $\pm\, 45$ (syst.) en $\sigma(\mathrm{pp} \to 2\mathrm{b}+2\mathrm{j}+\mathrm{X}) = 67.2 \pm 2.2$ (stat.) $\pm\, 22.5$ (syst.) voor, respectievelijk, de twee geselecteerde multijet scenario's. Het blijkt dat fixed-order matrix element berekeningen, inclusief parton showers, de gemeten werkzame doorsneden slechts in bepaalde gebieden van de faseruimte beschrijven, en dat het toevoegen van bijdragen afkomstig van tweevoudige parton verstrooiingen de Monte Carlo voorspellingen dichter bij de data brengen. Daarom wordt voor de eerste keer een nieuwe methode geïntroduceerd die de bijdrage van tweevoudige parton verstrooiingen in een experimentele meting kan bepalen. Deze is gevalideerd met bestaande W + dijet metingen, en daarna toegepast op de hiervoor beschreven multijet kanalen: de waarden van σ_{eff} zijn respectievelijk $19.0^{+4.6}_{-3.0}$ mb en $23.2^{+3.3}_{-2.5}$ mb. Deze resultaten zijn consistent met elkaar, en in overeenstemming met metingen gebaseerd op andere selectie kanalen bij eenzelfde massamiddelpuntsenergie van 7 TeV.

Zusammenfassung

Messungen der differentiellen Querschnitte für Multijet-Szenarien in Proton-Proton Kollisionen werden als Funktion des Transversalimpulses p_{T} und der Pseudorapidität η der Jets dargestellt. Außerdem werden Korrelationen in azimuthalem Winkel und dem p_{T}-Gleichgewicht zwischen den Jets gemessen. Zwei verschiedene Szenarien werden untersucht. Als Erstes wird ein Vier-Jet Endzustand in $|\eta| < 4.7$ gewählt, zwei harte Jets mit jeweils $p_{\mathrm{T}} > 50$ GeV und zwei Jets mit $p_{\mathrm{T}} > 20$ GeV. Keine anderen Jets mit $p_{\mathrm{T}} > 20$ GeV sind erlaubt. Das zweite Szenario besteht aus mindestens vier Jets mit $p_{\mathrm{T}} > 20$ GeV, wobei zwei aus einem b-Quark in $|\eta| < 2.4$ kommen müssen und keine Anforderungen an die anderen beiden Jets gestellt werden. Die Daten wurden in 2010 bei einer Schwerpunktsenergie von 7 TeV und einer integrierten Luminosität von 36 pb^{-1} mit dem CMS Detektor am LHC gesammelt. Die totalen Wirtungsquerschnitte für die jeweiligen Multijet-Szenarien sind $\sigma(\mathrm{pp} \to \mathrm{4j{+}X}) = 330 \pm 5$ (stat.) $\pm$ 45 (syst.) und $\sigma(\mathrm{pp} \to \mathrm{2b{+}2j{+}X}) = 67.2 \pm 2.2$ (stat.) $\pm$ 22.5 (syst.). Es wird gezeigt, dass Matrixelementberechnungen in fester Ordnung in α_{S} zusammen mit Parton Shower Simulationen die gemessenen differentiellen Querschnitte nur in einigen Bereichen des Phasenraums beschreiben und dass die Hinzunahme eines Beitrags von Double-Parton-Scattering die Monte-Carlo-Vorhersage näher an die Daten bringt. Zum ersten Mal wird eine neue Methode eingeführt, die den Beitrag des Double-Parton-Scatterings aus einer experimentellen Messung extrahiert. Diese Methode wird angewandt auf W+dijet Messungen und auf beide Multijet Kanäle. Werte von σ_{eff} werden als $19.0^{+4.6}_{-3.0}$ mb und $23.2^{+3.3}_{-2.5}$ mb gemessen. Diese Werte sind untereinander konsistent und kompatibel mit Messungen basierend auf verschiedenen Physikkanälen mit 7 TeV.

Declaration on oath

I hereby declare, on oath, that I have written the present dissertation by my own and have not used other than the acknowledged resources and aids.

Hamburg, 2014
Paolo Gunnellini

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Hamburg, 2014
Paolo Gunnellini

List of publications

The work described in this thesis is based on the following publications:

- CMS Collaboration, *Studies of 4-jet production in proton-proton collisions at* $\sqrt{s} = 7\ TeV$, CMS-PAS-FSQ-12-013;
- CMS Collaboration, *Measurement of four-jet production in proton-proton collisions at* $\sqrt{s} = 7\ TeV$, Phys.Rev. D89 (2014) 092010, CERN-PH-EP-2013-229;
- P.Cipriano, S.Dooling, A.Grebenyuk, P.Gunnellini, F.Hautmann, H.Jung, P.Katsas, *Higgs boson as a gluon trigger*, Phys.Rev.D88(2013)9, 097501, DESY-13-139;
- S.Dooling, P.Gunnellini, F.Hautmann, H.Jung, *Longitudinal momentum shifts, showering, and nonperturbative corrections in matched next-to-leading-order shower event generators*, Phys.Rev.D87(2013)9, 094009, DESY-12-166;
- CMS Collaboration, *Underlying Event Tunes and Double Parton Scattering*, CMS-PAS-GEN-14-001;

Contents

List of Figures

List of Tables

Introduction

Our knowledge of the particle physics world is strongly connected to the formulation of the Standard Model (SM) [1]. The SM is a very successful theory which is able to describe a wide class of phenomenona undergone by elementary particles, by including a consistent picture of the interactions experienced by them. One of the considered interactions is called "strong interaction" and, as the name might suggest, it is the one with the largest intensity. Only coloured particles, namely quarks and gluons, participate to this interaction. Hence, strong interactions dominate the phenomenology of proton-proton collisions, which are investigated at the Large Hadron Collider (LHC), in Geneva. A general event at the LHC can be interpreted as a hard collision between two partons, one for each of the interacting protons, and softer scatterings among the other proton components. A peculiar and interesting scenario is, in particular, the occurrence of two hard scatterings within the same collision, driven by two different pairs of partons. This is what is generally called "Double Parton Scattering" (DPS). The relevance of DPS is rather high since, on one hand, it may constitute a relevant background in searches for signals of new phenomena, and on the other hand, it might have an impact on the development of partonic models of hadrons. Both theoretical and experimental investigations of DPS have a long history in the literature [2, 3, 4]: a large number of phenomenological studies, evaluating DPS contributions to various high energy processes, and several measurements of different final states at previous experiments have been carried out. For instance, final states with four jets have been studied at the AFS and CDF collaborations [5, 6], while the photon+three jets channel has been measured at the D0 experiment [7].

Events at the LHC offer a unique opportunity to study DPS signals. In fact, the high center-of-mass energy of the proton-proton collisions is able to scan partons carrying a low fraction of longitudinal momentum; in this particular regime, the parton densities are large and the probability to have more than one hard partonic interaction becomes non-negligible. So far, only the W+dijet channel has been measured at the LHC with DPS goals by the CMS and ATLAS collaborations [8, 9]. Nevertheless, due to its high cross section, jet production is certainly the most abundant and intuitive process to consider at the LHC: particularly relevant for DPS studies is a channel with four jets. Such final state might originate both from a DPS event, where the two pairs of jets are produced by two distinct interactions, or from a single parton scattering (SPS), where two jets come from the hard scattering and are accompanied by two additional jets, arising from the evolution of the interacting partons. Even though the SPS background is generally much higher than the DPS signal, techniques of how to disentangle the two contributions have been investigated. They are based on the measurements of correlation observables between the jets in the final state, which test their relative configuration in the transverse plane. A final state arising from SPS tends to have a strongly correlated configuration in azimuthal angle and a balance in transverse momentum between the two jet systems. In contrast, DPS events generally have uncorrelated topologies for jet pairs. Such measurements suffer from experimental complications: one of them is related to a correct association of pairs of the jets in the final state. It is indeed crucial to properly link the objects that come from the same hard interaction. For this purpose, the selection of jet pairs with different flavours can help to remove combinatorial background of jets. Another experimental difficulty is represented by possible pile-up contamination. However, the use of low pile-up data samples, recorded in 2010, allows to minimize its contribution.

From the measurement of the forementioned correlation observables, it is then possible to evaluate the amount of DPS which contributes to the specific physics channel. This contribution is in the simplest models quantified by the parameter σ_{eff}. So far, several methods have been developed for a measurement of σ_{eff} but all of them are based on the same concept: they fit the measured distributions with templates, built for SPS and DPS, and give as an output their corresponding fractions for an optimal description of the data. Even though this method has given stable and reliable results, recent discussions have highlighted the difficulty of obtaining a proper definition of the background template: hence, a different approach seems to be needed for σ_{eff} measurements.

This thesis enters this state of art and investigates the study of double parton scattering (DPS) signals in multijet final states, measured in proton-proton collisions at 7 TeV, with the CMS detector. The CMS detector is able to reconstruct jets with an excellent resolution and in a wide phase space. The goal of this measurement is to study possible DPS contributions in two separate four-jet scenarios, never measured at the LHC so far: in the first one, jets are selected in pairs with different p_{T} thresholds, while in the second one, the jet pairs are required to have a different flavour and are then associated according to it. The measurements of the correlation observables between the jet pairs are accordingly used to extract values of σ_{eff} for both scenarios. A new method is introduced, based on a global fit of the data without any separation of background and signal components.

The thesis is organized as follows: in Chapter 1 a brief overview of the Standard Model of particle physics is discussed, with a special focus on strong interactions and their mathematical and phenomenological treatment. In Chapter 2 the components of a hadronic collision, mainly ruled and affected by strong interactions, are described; Multiple Parton Interactions (MPI) and DPS are defined and illustrated from a theoretical and experimental point of view. In Chapter 3, a description of the Large Hadron Collider and the Compact Muon Solenoid is provided: this helps to understand the origin of the data used for the physics analysis. Chapter 4 deals with details of the detector simulation, implemented in Monte Carlo event generators: by using them, results of a sensitivity study are also shown for jet correlation observables, in order to examine the discriminating power for a DPS signal. In Chapter 5, details about the event reconstruction and the physics objects performed with the CMS detector are treated: special attention is given to the reconstruction of jets, which are the main entities of this thesis. In Chapter 6 the physics selections applied to the analyzed multijet scenarios are explained and comparisons between data and simulation at the detector level are shown for the measured observables. A full treatment of the detector effects affecting the measurements is given in Chapter 7, while the procedure of data unfolding is described and results of it are shown in Chapter 8: this needs to be performed in order to correct the selected data for all detector effects. Chapter 9 deals with the study of the systematical uncertainties to assign to the measurements, while in Chapter 10 a full description of the measured total and differential cross sections is given, along with the physics outcomes obtained by comparisons with various theoretical predictions. In Chapter 11 details of the new method used for extraction of DPS signal are explained: this method is then applied to the measured multijet scenarios and results are shown and interpreted in the light of measurements performed at different experiments and with different final states. An overview of the whole work, together with future outlooks, is presented in the conclusive Chapter.

Additionally, supplementary work, performed during the PhD time, is presented in the five appendices which are added at the end of this manuscript. In Appendix A, an attempt of calibration for the CASTOR forward calorimeter is described through the dijet balance method, while Appendix B deals with a phenomenological study related to the possibility of identifying quark- and gluon-initiated processes in proton-proton collisions by measuring Drell-Yan and Higgs final states. In Appendix C, studies of non-perturbative corrections are shown for inclusive jet p_T spectra: in particular, the impact of the single UE components for these variables is isolated and evaluated. In Appendix D, results on new tunes performed with the PYTHIA8 event generator are shown: they can describe the energy dependence of UE data and, hence, may become very important for measurements performed at higher collision energies in future phases of LHC. In Appendix E, technical details are given about the determination of σ_eff within Monte Carlo event generators and additional material is provided about comparisons of various predictions to the correlation observables in the two measured channels.

The work is based on various publications: measurements of four-jet observables have become public with preliminary results[1] in May 2013 [10] and published in May 2014 [11] in Phys.Rev.D. The measurement performed in the channel with two b- and two additional jets is ready to seek for approval within the CMS collaboration and it will be soon available for review. The measurement of σ_eff in the four-jet channel is published as preliminary result in [12]. The work, described in Appendix B–D, is included in the following publications: two phenomenological papers submitted to Phys.Rev.D [13, 14] and a preliminary result approved by CMS [12] in the generator group.

---

[1] The preliminary results have a reduced phase space, by selecting jets only in the central region, while the published results have an extended jet selection, covering the full pseudorapidity coverage available in CMS.

Chapter 1

The Standard Model of particle physics

The theory that helps to explain and understand most of the phenomena related to particle physics is the Standard Model (SM) of the elementary particles. This theory is the result of centuries of ideas and research which started from basic questions that were asked back in the time of the ancient Greece. At that time, these questions were mixed with philosophical arguments related to the explanation of what was seen with bare eyes. In particular, already Democritus, in the 9th century B.C., raised the discussion about the basic elements of nature, asking himself which were the smallest constituents of the matter surrounding him, and starting the philosophical school of reductionism. He conjectured the existence of tiny objects, that he called "atoms" (from the Greek "indivisible") [15], as elementary components of all matter. This is the proof of the human interest in the basic structure of nature, already in ancient times, and of the attempt of reducing it to simple elements. Later, systematic research was conducted in this direction, making use of the "scientific method" [16] introduced by G. Galilei in the 15th century. After the discovery of the electron by Thompson [17] in 1898, models of the atomic structure started to be conceived: one of the most famous was the "plum pudding" model [18], created by Thompson himself, but falsified by the "golden foil" experiment [19], performed by Geiger and Marsden in 1909 and interpreted by Rutherford in 1911. It showed clearly that the atom is essentially composed of vacuum with a concentrated positive nucleus. With the advent of quantum mechanics in the 1920s, the whole picture experienced a small revolution: the classical concepts of position and velocity were revisited in the light of the Heisenberg principle and the wave-particle duality of matter [20]. With these basic principles, several breakthroughs followed very quickly and the 20th century was an incredibly successful time in terms of particle discoveries: from the neutron, discovered by Chadwick [21] in 1932, to the quarks (the last one of which was discovered in 1995), leptons, and gauge bosons, which have been interpreted as mediators of the interactions between the elementary particles. All these pieces constitute the building blocks of the SM. The SM was formulated in 1970-73 [1] as a renormalizable Yang-Mills theory, based on the concept of quantum field, with a SU(3)×SU(2)×U(1) gauge symmetry [22]. It gave an elegant framework of elementary particle physics, consistent with quantum mechanics and special relativity [20]. Besides the accurate explanation of the phenomena already known, its astonishing success derived from the fact that particles and mechanisms hypothesized at that time, were discovered and confirmed at experiments later. Nevertheless, this theory is somehow incomplete, since some missing pieces, like, for instance, gravity, need to be introduced in order to achieve a coherent picture of particles and interactions. Some experimental observations, like neutrino oscillations, or the existence of dark matter and dark energy, along with the matter-antimatter asymmetry, are still unresolved in the SM framework [23].

The mathematical formulation of the SM follows the lagrangian approach, namely it is defined by a function which describes the whole dynamics of the considered system. The forementioned SU(3)×SU(2)×U(1) represents the symmetry group of the SM lagrangian. A symmetry group defines under which transformations a theory is invariant. The symmetries of a theory are important because they are connected to conserved physical quantities by Noether's theorem: it states that a conservation law is associated to a physical system for each continuous symmetry of its action [24][1]. For the SM, the three pieces of the symmetry group are related to respectively, the conservation of colour, weak isospin and electric charge. This means that only processes that satisfy these conservation laws are allowed within the SM. The fundamental constituents of the SM are the fermions and the interactions which they experience. Each interaction is associated to a carrier. Fermions and carriers are currently considered elementary particles, namely particles whose internal structure is not measurable.

1.1 The particles in the SM

A fermion is a particle that has a half-integer spin according to the laws of quantum mechanics. In contrast, a boson is a particle that has an integer spin. Fermions and bosons are not necessarily elementary particles as they might also be composite ones. The particles of the SM are described by wave functions, changing in space-time, and have the nature of quantum fields. A quantum field is generally defined as a physical quantity which propagates and extends to every point in space-time and is in this system characterized by quantized numbers. The wave functions for fermions and bosons have different features and respect different symmetries. The fermionic wave function, i.e. the wave function of a system of several identical fermions, is antisymmetric under the exchange of two particles, while the bosonic wave function is symmetric if two bosons of the system are swapped. The space-time evolution of the fermions is described by the Dirac equation [25]: it has been demonstrated that this equation follows the laws of special relativity and admits two solutions, one with positive energy and one with negative energy. While it was straightforward to associate the solution with positive energy with the matter particles, the solutions at negative energy were less intuitive: they were interpreted as antiparticles, i.e. particles with all quantum numbers (charge, leptonic and baryonic number, etc.) inverted with respect to the corresponding particles but with the same properties and features. Thus, all fermions have to be considered as being accompanied by their corresponding antiparticles[2]. The space-time evolution of a boson follows the Klein-Gordon equation [25].

The SM recognizes two different types of fermions: quarks and leptons. For each of them, three groups, generally called "generations", are identified. For the quarks, the up-type generation is positively charged, while the down-type one is negatively charged. The first generation, composed

---

[1] The action is the lagrangian integrated over the time.

[2] The convention to indicate antiparticles is to take the name of the corresponding particle and to add a bar on top: for instance, $u \to \bar{u}$, to be read "anti-u" or "u bar".

by the up-quark (u) and the down-quark (d), is the lightest one (mass $m_{\mathrm{u,d}} \sim$ 5-15 MeV/c^2) and the first to be observed in 1969 [26]. The second generation comprises the strange-quark (s) and the charm-quark (c), while the bottom-quark (b) and the top-quark (t) are included in the third generation. The existence of the strange quark ($m_{\mathrm{s}} \sim$ 200 MeV/c^2) was deduced from observations of strange particles in cosmic rays, bottom ($m_{\mathrm{b}} \sim$ 4 GeV/c^2) and charm ($m_{\mathrm{c}} \sim$ 1 GeV/c^2) evidences appeared in electron-positron collisions. The top quark, instead, due to its mass equal to about 175 GeV/c^2, much higher than all the other quarks, was discovered only in 1995 in proton-antiproton ($p\bar{p}$) collisions. The quarks may combine in composite particles: the resulting particle of a group of three quarks (antiquarks) is called “baryon” (antibaryon), while if a quark (antiquark) is coupled to an antiquark (quark), one speaks about “mesons”. Baryons and mesons are grouped into the category of “hadrons”. An example of a baryon is the proton, composed by the combination of two u-quarks and one d-quark, while the family of kaons, with a s-quark and either a $\bar{d}$- or a $\bar{u}$-quark, is an example for mesons.

The three generations of leptons are formed by a negatively charged particle and a neutrally charged one. The former are respectively the electron (e^-), the muon (μ^-) and the tau (τ^-), while the latter are the corresponding neutrinos, indicated with the letter ν. The range of masses is also very wide for leptons: while the electron has a mass of 0.5 MeV/c^2, μ and τ own respectively a mass of 105 MeV/c^2 and 1.8 GeV/c^2. Separate mention goes to the neutrino masses: while in the SM they are massless, there are experimental observations that invalidate this hypothesis and suggest a non-zero but very small mass for all the neutrino flavours, of the order of 1-10 eV. Figure 1.1 shows a summary of the particles, included in the SM, divided in the different classes: quarks, leptons and gauge bosons. The Higgs boson is also indicated (see Section 2.3).

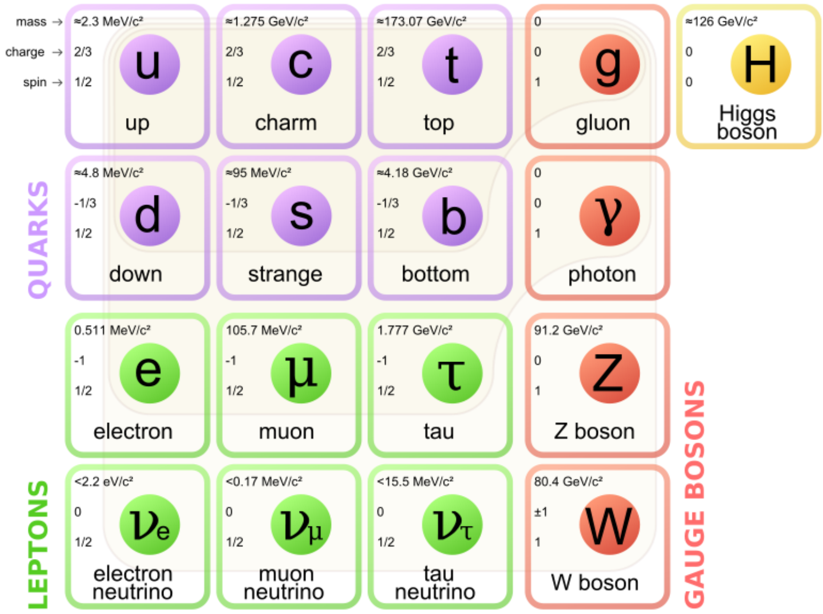


Figure 1.1: Sketch of the particle zoo included in the Standard Model [27]. Quarks, leptons and gauge bosons are represented, respectively, in blue, green and red. The Higgs boson is drawn in yellow. Values of mass, spin and electric charge are indicated for each particle.

1.2 The interactions in the SM

The interactions are described in the SM framework as an exchange of gauge bosons between interacting particles. In the SM, three types of interactions are defined and each of them is characterized by different gauge bosons which mediate it[1]. The gauge bosons carry energy and momentum and they are identified by specific quantum numbers. The possibility for a SM particle to participate in a certain interaction is determined by symmetry properties and by its quantum numbers. During the interaction, the exchanged bosons might appear as virtual states, namely they might not have the same mass as the corresponding real particle. This can be also seen as a violation of the energy-momentum conservation, but still allowed according to the Heisenberg's uncertainty principle ($\Delta E \Delta t \geq \hbar$) if the virtual state lives for a very short time. The longer the virtual particle exists, the closer its mass is to that of the real particle. Every interaction is characterized by a dimensionless coupling, which is proportional to the strength of the force exerted by the interaction, and by an interaction range which represents the maximum distance that the virtual state of the exchanged boson is able to travel within the bounds set by the uncertainty principle. The range is directly related to the mass of the exchanged boson.

Every interaction can be described by Feynman diagrams, which give a compact mathematical expression of the physics process and help to visualize it in a pictorial view. They consist of lines, that represent particles moving in the space-time, and vertices which illustrate the interactions. The opened lines are the particles in the initial and final state while the closed ones are the virtual particles that are created and absorbed during the interaction. For any vertex, the running coupling describes the strength of the interaction. An example of Feynman diagram is shown in Figure 1.2.

In the following, all the interactions of the SM are briefly described. Note that the gravitational interaction is not included in the SM, because a quantum gravity theory still encounters theoretical problems [28]. Since it is anyway much weaker than the others and does not affect any of the results of this thesis, it will not be further treated.

1.2.1 The electroweak interaction

The electroweak interaction groups the electromagnetic and the weak interactions within the same force. In the following Sections, the two are first described separately, then the formalism of their unification is treated.

1.2.1.1 The electromagnetic interaction

The electromagnetic interaction is described within the theory of the Quantum Electrodynamics (QED) [1]. It is based on a symmetry group $\mathrm{U}(1)_{\mathrm{EM}}$ and is represented in the quantum formalism by the lagrangian:

$$L = \bar{\psi}(i\hbar c \gamma_\mu D^\mu - mc^2)\psi - \frac{1}{4\mu_0} F_{\mu\nu} F^{\mu\nu} \tag{1.1}$$

---

[1] This is the reason why gauge bosons are generally referred to as "mediators" of a specific interaction.

where ψ and $\bar{\psi}$ are the quantum fields of the interacting particles, D_μ is the covariant derivative and $F_{\mu\nu}$ is the electromagnetic tensor. The mediator of the interaction is the photon (γ). It is a neutral and massless particle of spin 1. Only particles which carry an electric charge are coupled to it. This is the reason why the electromagnetic interaction does not provide the photon self-coupling, i.e. the photon cannot interact with itself. The coupling strength of the electromagnetic interaction results to be:

$$\alpha_{\mathrm{EM}} = \frac{e^2}{4\pi\epsilon_0\hbar c} \approx \frac{1}{137} \tag{1.2}$$

where e is the electric charge of the electron, and ϵ_0 is the vacuum permittivity. Since α_{EM} is smaller than 1, it is possible to treat the processes inside QED within a perturbative framework. Due to the fact that the photon is massless, the interaction range is infinite and its intensity decreases for increasing distances between the interacting particles. Another interesting feature of the electromagnetic coupling is that α_{EM} is different depending on the momentum scale at which the interaction occurs: it is smaller at big distances, corresponding to low momenta, while it increases as long as the interaction scale gets larger. Every charged particle is, in fact, surrounded by a cloud of virtual γ and e^+e^- pairs which are continuously emitted and absorbed through quantum processes; this implies that, in a classical picture, the net charge of the particle seen by an external test charge varies depending on the distance between the two. This effect is named "charge screening" and it traduces into the so-called "running coupling" α_{EM}.

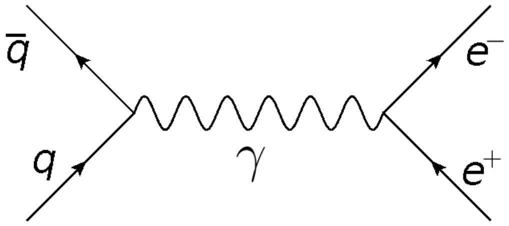


Figure 1.2: Feynman diagram of an electromagnetic interaction between a quark and an antiquark in the initial state: they annihilate into a virtual photon which then produces an electron-positron pair.

In the daily life, the electromagnetic interaction may be thought as responsible for the atomic and molecular structure of the matter, the electromagnetic wave emission, the electric and magnetic currents, and many other phenomena.

1.2.1.2 The weak interaction

The mediators of the weak interaction form a set of three gauge bosons: one with a neutral electric charge (Z^0) and two with a non-zero electric charge (W^+ and W^-), positive and negative, respectively. Differently from the photon, they are massive: the W bosons have a mass of 80.4 GeV/c^2, while the Z boson has a mass of 91.2 GeV/c^2 [29]. Due to the large mass of the mediators, the range of the weak interaction is extremely small ($\sim 10^{-16}$-10^{-17} m); in particular, the extreme smallness of the range of the weak interaction induced Enrico Fermi in the early 1940s to elaborate a theory of contact interaction with no range for weak processes. The ideas of this preliminary contact interaction were then implemented in a quantum field picture, and evolved in the theory that is currently accepted. It is based on a SU(2) symmetry and the corresponding lagrangian, relative to the interaction part, can be written as:

$$L = \frac{g}{2}\gamma^{\mu}(c_{\mathrm{V}} - c_{\mathrm{A}}\gamma^{5}) \tag{1.3}$$

where g is the weak coupling, and c_{V}, c_{A} are coefficients which rule the size of the vectorial and axial part of the interaction.

The quantum number that governs the weakly interacting particles is the weak isospin. Its third component (called T_3) becomes relevant because it is conserved by the weak interaction, i.e. its value for the initial and the one for the final state coincide. Every left-handed[1] fermion has a non-zero T_3: the up-type fermions have $T_3 = 1/2$, while the down-type ones have $T_3 = -1/2$. Right-handed particles have, instead, a null T_3[2]. The T_3 value determines the strength of the weak coupling, which is therefore different for left-handed and right-handed fermions. The strength of the weak interaction is, however, very small as the name "weak" might suggest: it is about 4-5 orders smaller than α_{EM}.

Since the mediators have different electric charges, different types of processes are identified: those mediated by a W boson, where the particles interact via a charged current, and those mediated by a Z boson, which is a neutral current. The two are represented in the Feynman diagrams of figure 1.3.

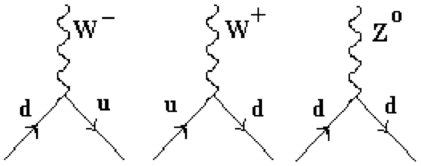


Figure 1.3: Feynman diagrams of weak interactions: the two charged currents occurring via an exchange of a W boson, and the neutral weak current, through a Z boson exchange, are shown.

The two different weak currents present some differences: while the neutral current couples only to quarks of the same flavour, quarks of different generations can interact via charged current. Both of them behave in the same manner in lepton interactions, where they couple only to particles inside the same generation. This phenomenology is explained inside the quark sector by the mixing mechanism, according to which the weak eigenstates of the quarks are different from their mass eigenstates. This means that, when a quark is produced as a mass eigenstate in a certain generation, it may decay via weak interaction into a quark of another family, because its weak eigenstate is a superposition of quarks of all three generations. Formally, this mechanism is achieved by the Cabibbo-Kobayashi-Maskawa (CKM) matrix (V_{CKM}), which specifically gives the contents of the weak eigenstates in terms of the mass ones. Between them, this relation holds:

$$\begin{pmatrix} d' \\ s' \\ t' \end{pmatrix} = V_{\mathrm{CKM}} \begin{pmatrix} d \\ s \\ t \end{pmatrix} = \begin{pmatrix} V_{\mathrm{ud}} & V_{\mathrm{us}} & V_{\mathrm{ub}} \\ V_{\mathrm{cd}} & V_{\mathrm{cs}} & V_{\mathrm{cb}} \\ V_{\mathrm{td}} & V_{\mathrm{ts}} & V_{\mathrm{tb}} \end{pmatrix} \begin{pmatrix} d \\ s \\ t \end{pmatrix} \tag{1.4}$$

---

[1] The concept of left- and right-handed particles concerns the relative direction of spin and of momentum of a particle: if they are opposite to each other, the particle is referred as left-handed, while if they point to the same direction, it is called right-handed. This feature is closely related to the concepts of helicity and chirality.

[2] The weak isospin values are symmetric for the corresponding antifermions.

The CKM matrix is a 3×3 unitary complex matrix with small off-diagonal coefficients. This translates in a higher probability for a charged current to change the quark flavour within the same generation and in a non-zero probability to access the other generations. Another important feature of the weak interaction is the fact that it does not conserve either the parity (P) symmetry nor the charge-parity (CP) symmetry. These violations imply that right-handed and left-handed particles behave differently under weak interaction, as well as particles and antiparticles.

Even though it is quite difficult to find examples of weak interactions in the daily life, it is however very important: it explains, for instance, the β-radioactivity and the thermonuclear processes inside the sun.

1.2.1.3 The electroweak unification and the Higgs boson

The electroweak unification is the formulation of electromagnetic and weak interactions within the same framework. This has been introduced by Glashow by noticing that the two forces share the main features and the apparent differences basically come from the different masses of the mediators. In particular, the weak neutral currents and the electromagnetic interactions mediated by the photon, also neutral, are basically the same, with the only difference of the mass of the respective carriers; the photon is massless, while the Z boson has a bigger mass. Hence, a "new" electroweak interaction can be defined with a gauge symmetry U(1)×SU(2), which determines the conservation of the electric charge and the weak isospin. In this unified formulation, each fermion generation presents two chirality multiplets, a left-handed and a right-handed one. The gauge bosons are divided in a triplet W^{μ}, with two electrically charged and one neutral component, and a singlet B^{μ} with zero electric charge. To describe the physical states of the neutral component, the symmetry needs to be broken, due to the different mass of the two physical states, γ and Z^0. Hence, two of the electroweak fields, B^0 and W^0, do not correspond to the physical states but they are only a mixture of them (γ and Z^0). In a matrix formulation, it is written as:

$$\begin{pmatrix} |\gamma\rangle \\ |Z^0\rangle \end{pmatrix} = \begin{pmatrix} \cos\theta_{\mathrm{W}} & \sin\theta_{\mathrm{W}} \\ -\sin\theta_{\mathrm{W}} & \cos\theta_{\mathrm{W}} \end{pmatrix} \begin{pmatrix} |B^0\rangle \\ |W^0\rangle \end{pmatrix} \tag{1.5}$$

where θ_{W} is the weak mixing angle or Weinberg angle. Its value has been obtained by measuring neutral and charged current events in different channels and is equal to 0.231 [29]. The coupling of the gauge bosons to the different multiplets is different: in particular, the charged components, W^+ and W^-, couple only left-handed multiplets, γ couple equally both left-handed and right-handed multiplets, while Z^0 pairs the two multiplets with different coupling constants.

In this picture, though, no mass terms are allowed in the lagrangian, to satisfy gauge invariance, a crucial ingredient for the consistency of the model. In this case, all particles, gauge bosons and fermions, should be massless. However, from experimental measurements it is well established that this is not the case. This apparent inconsistency is solved by a spontaneous symmetry breaking mechanism, elaborated by Englert, Brout and Higgs in the 1960s [30, 31]. This model implements the possibility for particles to have mass, by introducing a new field ϕ (the Higgs field), and an associated potential $V(\phi)$. The peculiarity of $V(\phi)$ is that it has a sombrero-hat shape, which determines the spontaneous symmetry breaking by choosing a vacuum state. The direction of the symmetry breaking is the one which gives mass to the Z^0 boson and leaves γ massless. By postulating the coupling between ϕ and the fermions, it is also possible to confer them a non-zero mass and save the gauge invariance. As a result of the broken symmetry, a new scalar field h appears in the lagrangian and this gives rise to a new massive particle, the Higgs boson. The research for this new state has a two-decade history [32] and it had a breakthrough on 4th July 2012 with the presentation of the evidence of the boson at the LHC, as measured by the CMS [33] and ATLAS [34] collaborations. Further measurements and analyses are ongoing, in order to verify whether it corresponds actually to the SM Higgs boson or to some other new particles. The current state of art is a general compatibility with the SM predictions. Indeed, more precise future measurements might show discrepancies, opening up different scenarios.

1.2.2 The strong interaction

The theory that explains and describes strong interactions in the SM is the Quantum ChromoDynamics (QCD) [35]. It is based on a $\mathrm{SU}(3)_{\mathrm{C}}$ symmetry and its lagrangian can be written as:

$$L = \bar{\psi}(i\hbar c\gamma_{\mu}D^{\mu} - mc^2)\psi - \frac{1}{4}G^{\mathrm{a}}_{\mu\nu}G^{\mu\nu}_{\mathrm{a}} \tag{1.6}$$

where $\mathrm{G}^{\mu\nu}_{\mathrm{a}}$ is the gluonic field tensor, which differs from the electromagnetic one for additional non-linear terms. The mediators of the strong interaction are the gluons, which are massless particles with spin equal to 1. In strong interactions, only particles which carry a colour charge are coupled to the gluons: this is the quantum number of the QCD symmetry. There are three types of colours: green (g), red (r) and blue (b) and their corresponding anticolours. Each type of quark and gluon has a non-zero colour charge and takes part to the strong interaction. In particular, a quark carries one of the three colours, while a gluon has a colour and anticolour charge: there are eight different types of gluons, depending on their colour-anticolour combination. The combination of the three different colours (anti-colours) or a pair of the same colour-anticolour corresponds to a neutral (white) colour. Figure 1.4 shows the first order Feynman diagrams of a QCD interaction: the number of represented diagrams has to be multiplied by all the possible colour combinations. Note that the colour is conserved by the strong interaction, i.e. the colour before and after the interaction remains the same.

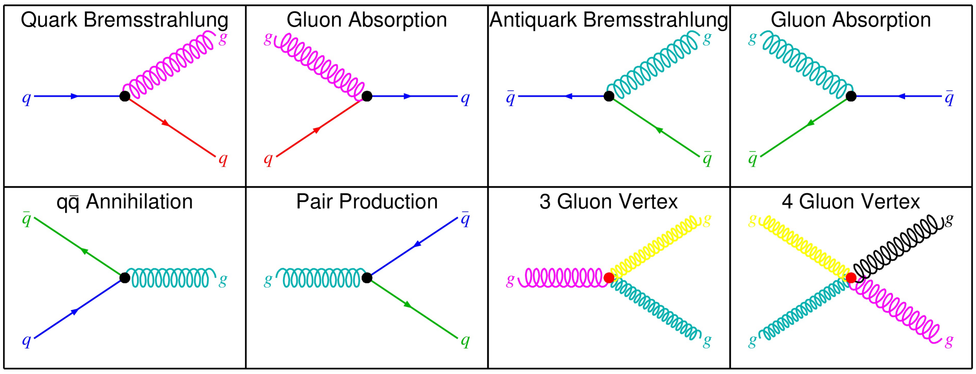


Figure 1.4: Feynman diagrams of splitting diagrams under strong interactions with quarks and gluons in the initial state [36]: a quark can emit a gluon or split into two gluons, while a gluon can emit a quark, split into two quarks or into two gluons. A four-gluon vertex is also allowed in QCD. No specific meaning is associated with the different colours of the lines in the diagrams.

Even though the gluon is massless, the range of the strong interaction is very small ($\sim 10^{-15}$ m, of the order of the size of the proton). This is due to the fact that the gluon can interact with itself, by emitting further gluons. This effect determines that the colour field does not spread out in space when coloured particles move apart, but it remains confined inside the created gluon cloud. This is the so-called antiscreening, in contrast with the screening effect seen in QED. The antiscreening effect has interesting consequences: at high energy scales, or low distances, the interaction becomes very weak and the quarks can be treated as free particles, while at very low energy scales, or high distances, the quarks perceive a very high colour field which prevents them to escape from colour-neutral objects. The QCD behaviour at high energies is called "asymptotic freedom", while the one at low energies is referred to as "confinement". These phenomena can be followed by the evolution of the strong running coupling which depends on the scale of the interaction: unlike the electromagnetic interaction, the strong coupling is very low at high energies, while it dramatically grows up when going to low energies. In particular, while in the first high-energy regime, the inequality $\alpha_S \ll 1$ is valid and perturbative calculations can be analytically performed, for scales ~ 3 GeV, α_S becomes greater than 1 and non-perturbative phenomenology studies need to be used for understanding and modelling these low-energy mechanisms.

1.2.2.1 QCD quark parton models

After the observation in the 1960s that the protons were not point-like objects and the introduction of the quarks, QCD research started to grow in importance from both an experimental and a theoretical point of view. In particular, a systematic study of the proton structure was conducted in order to understand how the inner partons are bound together and how they determine the macroscopic properties of the proton. The first attempt was done by Feynman, Bjorken and Paschos in 1969, when the Quark Parton Model (QPM) was developed to explain the results of the SLAC experiment [26] in electron-proton (*ep*) collisions. In QPM, the proton is only composed by three loosely bound quarks and an *ep* collision can be seen like a simple scattering of the electron off a quark. This model seemed to work initially very well but failed to describe the dependence of the cross section on the scale of the exchanged photon. The QPM predicts, in fact, no dependence on the transverse momentum exchanged in the scattering process, while a violation of the scaling was observed at EMC and HERA where measurements over several orders of magnitude in scales have been performed. This observation could be explained by the existence of the gluon, interpreted as the binding force of the quarks inside the proton. In order to explain *ep* results, further processes needed to be included in the model: the partons of the proton are allowed to interact between each other via the exchange of gluons and the gluons themselves can also further interact. The new model was called the Improved Quark Parton Model (iQMP). In the iQMP, the violation of the scaling observed in the data is naturally described by these new processes: a photon interacting with a quark at a certain scale probes the proton with a certain resolution. If the photon probes the same quark at a higher scale, the quark might have radiated a gluon not visible at the previous one and the photon effectively interacts with a quark carrying less momentum. This fact can be seen with a change of the internal structure of the proton. One of the main points of the iQMP is that the proton can now be represented like a set of three quarks, called "valence quarks", which determine its macroscopic quantities (like spin, charge, etc.) surrounded by a huge amount of gluons and quarks, called "gluons and quarks of the sea", which constitute the partons continuously emitted and absorbed by the valence quarks via the strong interaction. Note that in the old

QPM model, gluons and quarks of the sea were not considered at all. All possible scenarios for a photon-proton scattering in the iQMP model are pictured in Figure 1.5, where only diagrams with one splitting are represented.

The evolution of the partons, namely the fact that quarks and gluons can emit additional partons, is treated by the so-called “splitting functions” and results in a dependence of the cross section on the scale. The QCD semi-hard interactions experienced by quarks and gluons are treated analytically under the name of “parton evolution” or “parton shower” (PS). Different approaches to perform the parton evolution have been developed, depending on which kinematic region of the phase space is probed and several evolution equations are available for the calculations. Important variables, on which the equations depend, are Q^2, that indicates the scale at which the QCD evolution process occurs, and x, which corresponds to the fraction of longitudinal momentum of the proton carried by the considered parton, going from 0 to 1. In case of a scattering process, at scales where the perturbative approach can not be longer applied, the result is a set of coloured partons, with a relative low k_T and very close in space.

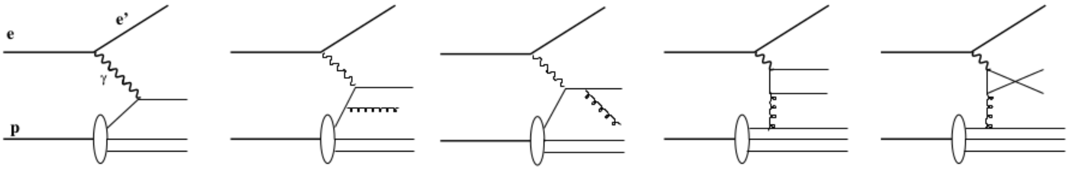


Figure 1.5: Possible contributions (at the first order) for an electron-proton scattering [37]: a quark in the proton may interact with the photon without further emission, or it may emit a gluon, before or after the interaction. The photon can also interact with a quark emitted by a gluon of the sea in the proton. All these diagrams summarize the quark and gluon QCD evolution at the first order in α_S.

1.2.2.2 The DGLAP evolution equation

At large Q^2 and not too large x, the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) equations [38, 39] are good tools to calculate the evolution of the density functions of quarks, $q_\mathrm{i}(x, Q^2)$, and gluons, $g_\mathrm{i}(x, Q^2)$, from a starting scale Q^0. The DGLAP equation can be written through a matrix formulation:

$$Q^2 \frac{\partial}{\partial Q^2} \begin{pmatrix} q_\mathrm{i}(x, Q^2) \\ g(x, Q^2) \end{pmatrix} = \frac{\alpha_\mathrm{S}}{2\pi} \sum_{q_\mathrm{j}, \bar{q}_\mathrm{j}} \int_x^1 \frac{dx'}{x'} \begin{pmatrix} P_{q_i,q_j}(\frac{x}{x'}, \alpha_\mathrm{S}) & P_{q_i,g}(\frac{x}{x'}, \alpha_\mathrm{S}) \\ P_{g,q_j}(\frac{x}{x'}, \alpha_\mathrm{S}) & P_{g,g}(\frac{x}{x'}, \alpha_\mathrm{S}) \end{pmatrix} \begin{pmatrix} q_\mathrm{j}(x', Q^2) \\ g(x', Q^2) \end{pmatrix} \tag{1.7}$$

where x/x', usually indicated with z, is the energy fraction of the initial parton with respect to the emitted one, $P_\mathrm{ij}(z)$ is the splitting function, which describes the probability of finding a parton of type i splitting in a parton of type j with an energy fraction z, with respect to the parent parton. The equations treat the partons as massless and assume that they move in the same direction as the proton. The $P_\mathrm{ij}(z)$ functions can be expressed also as perturbative expansions in α_S:

$$P(z, \alpha_\mathrm{S}) = P^{(0)}(z) + \frac{\alpha_\mathrm{S}}{2\pi} P^{(1)}(z) + \frac{\alpha_\mathrm{S}^2}{(2\pi)^2} P^{(2)}(z) + ... \tag{1.8}$$

They have been calculated up to order α_S^2 [40]. In $O(\alpha_S)$, for the processes shown in Figure 1.4, the $P_{ij}(z)$ functions are given by:

$$P^0_{qq}(z) = \frac{4}{3} \cdot \frac{1+z^2}{1-z} \tag{1.9}$$

$$P^0_{gq}(z) = \frac{4}{3} \cdot \frac{1+(1-z)^2}{z} \tag{1.10}$$

$$P^0_{qg}(z) = \frac{1}{2} \cdot [z^2 + (1-z)^2] \tag{1.11}$$

$$P^0_{gg}(z) = 6\left(\frac{z}{1-z} + \frac{1-z}{z}\right) + z(1-z) \tag{1.12}$$

As it can be noticed by these formulas, two divergences arise for values of z→0 and z→1; these correspond to the so-called infrared and ultraviolet divergences, respectively. What is physically happening is an infinite growth of the cross section for collinear (at small angle) and soft (at low p_T) gluon emissions. These theoretical inconsistencies are solved by resumming terms at all orders in α_S and by setting limits in the integrals. This is the approach implemented in the treatment of the PS in event generators (see Chapter 2), .

Usually, a Leading Logarithmic Approximation (LLA) for Equations 1.8-1.12 is used: it takes into account only the leading term of the splitting functions $P^0(z)$. In the LLA regime, the DGLAP equation resums all terms of the type $(\alpha_S \ln(Q^2/Q_0^2))^n$, where n denotes the order of the expansion of the cross section. It can be shown that the leading logarithmic terms correspond to an evolution, in which the parton virtuality is strongly ordered in transverse momentum:

$$Q^2 \gg |k^2_{T,n}| ... \gg |k^2_{T,i+1}| \gg |k^2_{T,i}| ... \gg |k^2_{T,0}| \tag{1.13}$$

where $k_{T,i}$ is the four-momentum of the parton i. This chain of inequalities includes also the fact that $x_{i+1} < x_i$, since momentum conservation needs to be fulfilled at each splitting. Equation 1.13 means that a parton, evolving from a hard scale Q^2 according to the DGLAP equation, may emit only successive partons with a lower transverse momentum. It is notable that due to the fact that DGLAP evolution resums $(\alpha_S \ln(Q^2/Q_0^2))^n$ terms, it will only be meaningful at large values of Q^2, where these terms are dominating. The DGLAP evolution gives a good description of the data in this region but it is expected to fail when higher-order contributions to the splitting functions become relevant and in different regimes, e.g. in the low-x region.

1.2.2.3 The BFKL evolution equation

Going to the low-x region, terms of the type $\alpha_S \ln(1/x)$ start to become important and need to be considered for the parton evolution. The low-x region is resolved for increasing energies and the gluon contribution is the dominant one for $x \leq 10^{-3}$. In this regime, the DGLAP approximation breaks down and needs to be improved by the Balitsky-Fadin-Kuraev-Lipatov (BFKL) evolution scheme [41, 42], which resums terms of the type $(\alpha_S \ln(1/x))^n$. Hence, this approach is valid only at small x. The DGLAP ordering in k_T is here replaced by an ordering in the longitudinal momentum fraction of the parton propagators, according to the relation:

$$x_0^2 \gg x_1^2 \gg ... \gg x_i^2 \gg x_{i+1}^2 \gg ... \gg x_n^2 \tag{1.14}$$

This implies that the emitted partons take a large fraction of the momentum of the original one in successive splittings. Furthermore, this ordering does not set any restriction in k_{T}^2 of the emitted partons and their virtualities can take any kinematically allowed value even down to small scales. Hence, the collinear approach is no longer valid, since the transverse momentum components cannot be neglected with respect to the longitudinal ones, and the virtuality of the partons needs to be considered. Consequently, a k_{T} factorization of the cross section needs to be used, with unintegrated parton density functions (uPDF), which have an explicit dependence on k_{T}, and virtual masses assigned to the partons, treated off-mass shell. The BFKL evolution of the uPDF $F(x, k_{\mathrm{T}}^2)$ can be written as:

$$\frac{dF(x, k_{\mathrm{T}}^2)}{d\ln(1/x)} = \int_0^\infty dk' F(x, k_{\mathrm{T}}'^2) K(k_{\mathrm{T}}^2, k_{\mathrm{T}}'^2) \tag{1.15}$$

where the function K is the BFKL splitting kernel.

1.2.2.4 The CCFM evolution equation

The Ciafaloni-Catani-Fiorani-Marchesini (CCFM) evolution equation [43, 44] tries to combine both the previous approaches by resumming terms of the types $(\alpha_{\mathrm{S}} \ln(1/x))^n$ and $(\alpha_{\mathrm{S}} \ln(Q^2/Q_0^2))^n$. Thus, it is valid at small x and large Q^2 values. It resembles the BFKL equation at the former ones, while it is more DGLAP-like at the latter values. The CCFM evolution equation introduces strong angular ordering of the parton emissions, ruled by the relation:

$$\Xi \gg \xi_{\mathrm{n}} \gg ... \gg \xi_{\mathrm{i}} \gg \xi_{\mathrm{i+1}} \gg \xi_0 \tag{1.16}$$

where the maximum angle Ξ allowed in the emission is given by the hard scattering process. The full CCFM splitting function can be written as:

$$\bar{P}_{\mathrm{g}}(z_{\mathrm{i}}, \bar{q}_{\mathrm{i}}, k_{\mathrm{Ti}}^2)) = \frac{\bar{\alpha}_{\mathrm{S}}(\bar{q}_{\mathrm{i}}^2(1 - z_{\mathrm{i}}^2))}{1 - z_{\mathrm{i}}} + \frac{\bar{\alpha}_{\mathrm{S}}(k_{\mathrm{Ti}}^2)}{z_i} \Delta_{\mathrm{ns}}(z_{\mathrm{i}}, \bar{q}_{\mathrm{i}}^2, k_{\mathrm{Ti}}^2) \tag{1.17}$$

where $\bar{\alpha}_{\mathrm{S}}=3\alpha_{\mathrm{S}}/\pi$, Δ_{ns} is the non-Sudakov factor[1], z_{i} is the ratio of the energy fractions in the branching (i-1)→i and k_{Ti} is the transverse momentum of the parton in the initial state of the cascade. This approach is very important in a k_{T}-dependent framework in both low- and large-x regimes.

1.2.2.5 Hadronization

The name “hadronization” generally refers to the process, experienced by coloured partons resulting from the shower, of rearrangement into a set of colour-singlet hadrons. The produced hadrons may further decay. The hadronization process is ruled by QCD and starts at a scale $Q^2 \sim 1$ GeV^2 where the perturbative regime cannot be applied. Different models have been developed to explain this transition from partons to hadrons and have been implemented in event generators. Since analytical calculations cannot be easily performed in a non-perturbative region, they have been constrained and optimized through data comparisons [45]. Two of the main and widely recognized hadronization models are:

---

[1] These factors, which implement the virtual corrections of real emissions due to the 1/z part of the splitting functions, are called non-Sudakov form factors, in contrast with the Sudakov form factors, which, instead, use emissions expressed by 1/(1-z) terms (see Chapter 2).

- the "string" model (known also as the "Lund" model);
- the "cluster" model.

String models started to appear in literature in the late 1970s and early 1980s, until the concept of the "Lund" model [46] has been formulated in 1983. This is based on the idea of a linear confinement between quarks moving apart, which is expressed at large distances through a potential of the form[1]:

$$V(r) = \kappa r \tag{1.18}$$

being r, the distance between the quarks, considered pointlike, and $\kappa \sim 1$ GeV/fm^2, the tension of the string, namely the energy per unit length. The quarks are imagined to be connected by a colour flux tube that carries more and more energy as long as the distance between them increases. When this rapidly growing potential energy reaches the order of hadron masses, it becomes energetically favourable for the string field to break and to create a quark-antiquark pair. These new particles decrease the energy of the string because they screen the colour charge of the initial quarks; the initial strings are, in fact, splitted in two separate "shorter" pieces. These connect the new quark-antiquark pair with the original quarks. This process ends when only ordinary hadrons remain in the chain and no energy is left out. A sketch of the whole chain of string creation under the strong potential is represented in Figure 1.6.

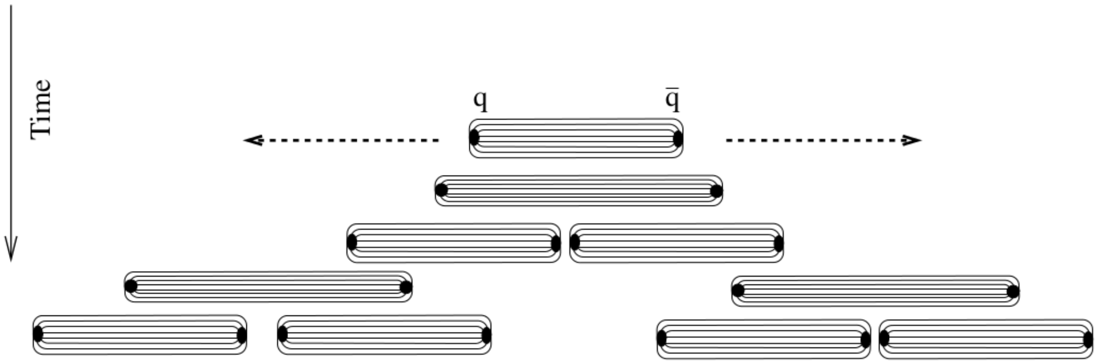


Figure 1.6: Sketch of the string mechanism for hadronization [47]: from an initial parton pair, additional pairs are created from the vacuum under the effect of the strong potential. New strings between the initial and the created objects are produced and the string formation continues with the same mechanism until only light hadrons are left.

Baryon production can be also included in this model, by allowing string breaks that produce pairs of "diquarks" which may then combine with a single quark to form the baryons. The function which governs the string break is the fragmentation function $f(z)$; it gives the probability of a produced hadron to have a fraction z of the momentum of the original string. It is defined as:

$$f(z) \propto \frac{(1-z)^{\mathrm{a}}}{z} e^{(-bm_\perp^2/z)} \tag{1.19}$$

where a and b are the parameters of the model and $m_\perp = \sqrt{m^2 + p_\perp^2}$ is the hadron transverse mass.

---

[1] The dependence of the potential includes also a Coulomb term proportional to 1/r for small distances, but it can be neglected at large distances

The cluster model [48] for the hadronization is based on the property of the preconfinement of QCD. This implies that at scales much below the scale of the hard process ($q \ll Q$), the partons produced in the shower are clustered in colourless groups with an invariant mass distribution, which is independent on the scale and on the process occurring in the hard scattering. The mass distribution falls rapidly for high values and depends only on the scale reached by the partons and on the QCD scale Λ. It is then possible to start the hadronization from these clusters, seen as proto-hadrons, that decay into a pair of colour-singlet hadrons. In the simplest models, a low scale for the cluster mass (~ 1 GeV) is set and each cluster is assumed to decay isotropically in its own rest frame, with branching ratios for the produced hadrons given simply by the density of states. Hence, the production of heavy mesons and baryons are also included in this mechanism. They account for the multiplicities of the various types of hadrons as well as for the hadronic energy and transverse momentum distributions observed in leptonic colliders, without the introduction of any specific fragmentation function. The simplest way for colour-singlet clusters to be produced is through splitting of gluons into $q\bar{q}$ pairs that then combine into singlets. A final state with partons associated in clusters is shown in Figure 1.7.

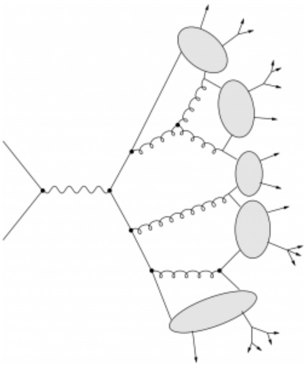


Figure 1.7: Sketch of the cluster mechanism for hadronization [35]: colourless clusters are created by partons at relatively low scales. They decay further into colour singlets until the mass of the clusters is lower than the mass of the hadrons.

In more advanced models, a higher cut-off scale is set and an anisotropic multihadron decay scheme is used [35], but these approaches lie already between the string and the cluster models.

At the end of the hadronization process, sequential decays of the unstable hadrons occur. The amplitudes and the widths of these decays are listed in [29] and their knowledge and precision are crucial to interpret a wide range of measurements dealing with QCD.

1.2.2.6 Definition of jet

After the hadronization process, the system is left with a bunch of hadrons, collimated towards the direction of the initial parton that originated them. It becomes then very practical to define an object that is able to group these particles into a single physics entity that can be approximately identified with the original parton. This procedure is very useful both for an experimental approach, where signals in near parts of the detector are combined in a single output, and from a theoretical point of view, where the multitude of hadrons is substituted by quantities less sensitive to the details of the hadronization process, that can be directly compared to the experimental data. This object is called a “jet”. Algorithms have been developed to define criteria to group and cluster hadrons into jets and two different classes of algorithms can be identified [49]:

- Iterative cone algorithms;
- Sequential recombination algorithms.

The iterative cone algorithms [50] are based on the definition of cones of a certain width R, in the θ-ϕ space. The relative distance of all the particles is iteratively calculated and compared with R. If the distance between the considered hadrons is smaller than R, they are clustered together in a jet and the jet direction is updated with the directions of the clustered particles, otherwise they initiate two different jets. The iteration stops when all final hadrons have been considered and stable jets are found. In case of two or more overlapping jets, particles tend to be associated to the jet that is closest to their direction.

The sequential recombination algorithms are based on the definition of a distance measure, according to which two objects (particles or pseudo-jets consisting of several particles) are combined into a new object or a new jet is created. The goodness and reliability of a jet algorithm are evaluated according to several criteria:

- it has to be simple to implement in an experimental analysis and a theoretical calculation;
- it has to be defined at any order of perturbation theory;
- it has to be infrared- and collinear-safe; infrared and collinear safety deals with the fact that, if one modifies an event by a collinear or a soft emission, the set of hard jets found in that event and their properties should remain unchanged;
- it has to be not strongly affected by the contribution of the spectator quarks and hadron remnants;
- it has to be not strongly sensitive to the details and the mechanisms of the hadronization process;

In the analyses described in this thesis, the anti-k_{T} algorithm [51] is used. This has been chosen because it fulfills the forementioned conditions and is invariant under boosts along the beam direction. This feature is important since it allows the algorithm to be used in hadronic collisions in the laboratory frame, where the initial boost of the interacting partons is not known. Indeed, quantities in the center of mass frame are obtained under a longitudinal boost from the laboratory one. The algorithm works through the following steps:

- for all the final state particles h_{i}, here referred to as objects, the resolution variables d_{ij} and d_{iB} are evaluated for all possible pairs of objects h_{i}-h_{j} and for any object h_{i} and the beam h_{i}-B, respectively. The resolution variables are defined as:

$$\begin{aligned} d_{\mathrm{ij}} &= \min(p_{\mathrm{T_i}}^{2p}, p_{\mathrm{T_j}}^{2p})\frac{\Delta_{\mathrm{ij}}^2}{R^2} \\ d_{\mathrm{iB}} &= p_{\mathrm{T_i}}^{2p} \end{aligned} \tag{1.20}$$

where $\Delta^2_{\rm ij}=(y_{\rm i}-y_{\rm j})^2+(\phi_{\rm i}-\phi_{\rm j})^2$ and $p_{\rm T_i}$, $y_{\rm i}$, $\phi_{\rm i}$ are respectively the transverse momentum, rapidity and azimuthal angle of particle i. R is a constant that measures the cone width used for the clustering procedure and p is taken equal to -1[1] in the anti-$k_{\rm T}$ algorithm;

- The smallest value between $d_{\rm ij}$ and $d_{\rm kB}$ is called $d_{\rm min}$. If $d_{\rm min}$ is the $d_{\rm ij}$ variable, then the two objects $h_{\rm i}$ and $h_{\rm j}$ are combined into a new object $h_{\rm k}$, whose kinematical quantities are calculated through the $p_{\rm T}$-weighted average of the quantities of the two initial objects; in a compact form:

$$\begin{aligned} \eta_{\rm k} &= \frac{p_{\rm T,i}\cdot\eta_{\rm i}+p_{\rm T,j}\cdot\eta_{\rm j}}{p_{\rm T,k}} \\ \phi_{\rm k} &= \frac{p_{\rm T,i}\cdot\phi_{\rm i}+p_{\rm T,j}\cdot\phi_{\rm j}}{p_{\rm T,k}} \\ p_{\rm T,k} &= p_{\rm T,i}+p_{\rm T,j} \end{aligned} \tag{1.21}$$

  The new object, obtained in this way, replaces the initial particles.

- If $d_{\rm min}$ is the $d_{\rm iB}$ variable, then the object $h_{\rm i}$ is considered as a final jet and is removed from the objects which are further processed by the algorithm;

- The algorithm starts again with the remaining objects until no object is left.

The performance of the anti-$k_{\rm T}$ algorithm is discussed in [51]. It tends to cluster particles starting from the ones at highest $p_{\rm T}$, and to produce jets preferably with circular area, as shown in Figure 1.8. Minor effects on the jet observables are observed in presence of soft particles. Values of R are chosen usually between 0.4 and 1. Effects of the choice of different R are discussed in Appendix C.

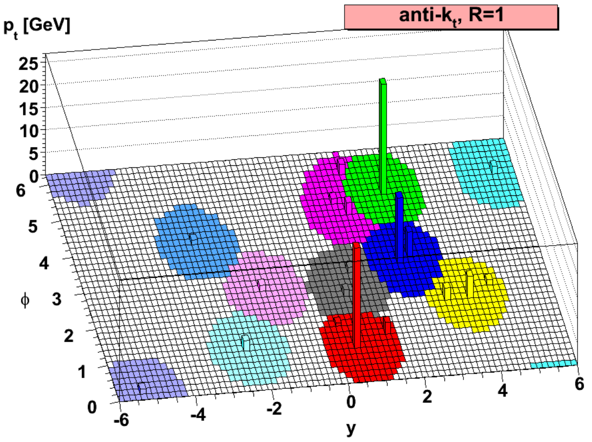


Figure 1.8: Study at the parton level, showing the active areas in the y-ϕ space, for jets clustered with the anti-k_T algorithm with R = 1 [51]. Towers represent the jet p_T. Note that the jet areas tend to be circular; in case of overlapping jets, the one with highest p_T tends to "eat" area from the softer jets.

Additional techniques have been developed to further improve the performance of these algorithms for specific purposes, like study of boosted topologies originated by resonance decays. A

---

[1] The other sequential recombination algorithms differs from the anti-$k_{\rm T}$ only in the exponent: p=1 for the $k_{\rm T}$ algorithm [52] and p=0 for the Cambridge-Aachen algorithm [53].

detailed description of methods and possible applications is described in [54], however they have not been used in this thesis work.

1.2.2.7 Flavoured jets: production of b-jets

The theory of QCD is able to explain a huge range of phenomena in jet physics and the study of jets is crucial for a detailed understanding of the mechanisms of strong interactions: a jet contains, indeed, information about production and evolution of the original parton and includes features related to the hadronization. In particular, jets are classified according to the flavour of the parton which originates them: if a jet is initiated by an up, down or strange quark, or by a gluon, it is called "light-jet", while if the jet derives from the evolution of a charm or a bottom quark, it takes the name of "heavy-flavour jet". Note that a top quark does not originate any jet since, due to its very small lifetime, it decays before hadronizing; thus, "top-jets" do not exist in nature. Discrimination of light and heavy-flavour jets is very important in order to separate and compare features of partons of different flavours, and to understand their dynamics and study their evolution.

For the work described in this thesis, heavy-flavour jets are very important, in particular b-jets. A jet is called a b-jet if it is originated from a b-quark. In a particular approximation, used in most event generators, where the transverse momenta of the initial partons are neglected[1], a b-quark can be produced in different ways at hadron colliders[2]; in particular three process categories at the first order are defined:

- flavour creation (FC): a $b\bar{b}$ quark pair is produced by a gluon pair or a quark-antiquark in the initial state, via a 2→2 process;
- flavour excitation (FEX): a 2→3 process where a $b\bar{b}$ quark pair is emitted by a gluon of the initial state. One of the two emitted quarks interacts then further with a parton of the other proton;
- gluon splitting (GS): a $b\bar{b}$ quark pair is emitted by a gluon of the final state. The b-quarks are mainly emitted collinearly and the overall interaction is a 2→3 process.

The Feynman diagrams of the three different processes are represented in Figure 1.9(top). After the production, the b-quarks are subjected to the parton evolution before forming B-hadrons, namely hadrons which contain a b-quark, accompanied by quarks of light flavour. These B-hadrons have a very peculiar feature: their lifetime is quite high ($\sim$ 5 ps) and this property can be used experimentally to separate b-jets from light-jets[3]. The relative contribution of the different processes to the b-quark (and consequently to the b-jet) production depends on the collision energy and on the colliding hadrons: for increasing energies, the higher-order diagrams become more and more relevant, while for $p\bar{p}$ collisions, the FC process remains the most important production process. In proton-proton (pp) collisions at 7 TeV, FC contributes only for the 10-20% of the total b-jet production, while the higher-order diagrams share the remaining 80%. Figure 1.9(bottom) shows the relative percentage as a function of the b-jet p_{T} for pp collisions at 7 TeV. While the FEX constitutes the main production process and its rate is rather constant as a function of p_{T}, the GS increases when going at higher p_{T} up to 40% for 120 GeV. FC processes are the less relevant

---

[1] This is referred to as "collinear factorization" and it is treated in more detail in Chapter 2.

[2] A hadronic collision is described in full detail in Chapter 2.

[3] Details of the experimental techniques will be given in Chapter 5.

diagrams for the b-jet production and from an initial contribution of 20% at $p_{\mathrm{T}} \sim 20$ GeV, their influence decreases down to $\sim 15\%$ for $p_{\mathrm{T}} \sim 120$ GeV.

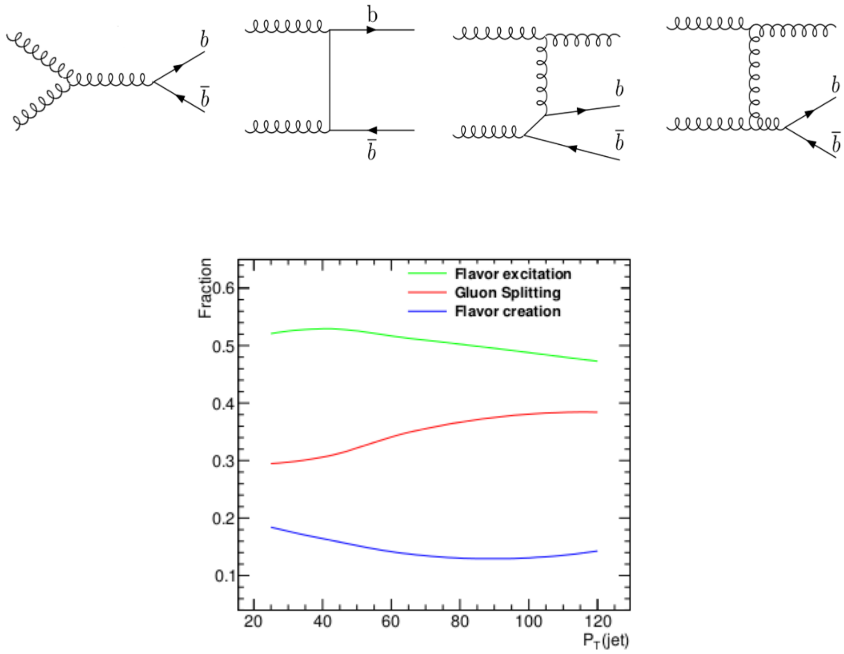


Figure 1.9: (Top) Feynman diagrams for b-quark pair production at the LHC: from left to right, the first two are called flavour creation (FC), then flavour excitation (FEX) and gluon splitting (GS). (Bottom) Percentage of each process on the total b-quark pair production as a function of the p_T of the b-jets [55].

The study of heavy flavour sector is particularly important for precision measurements [56], like CP violation or rare decays. From the point of view of strong interactions, measurements relative to b-jets might give interesting insights on the content of b-quarks inside the proton, on details of b-hadronization and on the understanding of the higher-order processes (i.e. FEX and GS).

1.3 Possible scenarios beyond the Standard Model

Even though the SM appears to give a very satisfying description of the particle physics and a well consistent model for the particle world, some issues are still unsolved within its formulation. These can be thought as divided in two branches: the ones connected to experimental observations which cannot be explained by the SM, and the ones related to theoretical incoherences or inconsistencies. Several measurements, mainly from astrophysical sources, fall into the first category. In the following list, the main ones are addressed:

- Neutrino oscillations [57];
- Existence of dark matter [58];
- Existence of dark energy [59].

A different kind of topics belong to the second class of problems. One of these is the so-called “hierarchy problem” [60], related to the huge difference (almost 32 orders of magnitude) between the electroweak scale and the Planck scale, which makes the theory unnatural. To this topic, in particular, a “fine-tuning problem” [60] is related, when looking at the radiative corrections to the Higgs mass: very large contributions come from loop diagrams, bringing the Higgs mass up to a very high value, far from the measured value. These corrections might cancel out, giving a value of the Higgs mass, similar to the experimentally measured one, only by choosing particular values of the theory parameters, which again seem to be a non-natural and non-probable solution.

Several speculations have been promoted in order to solve all these issues and they are collected under the name of “New Physics”. The most fascinating and natural way to extend the SM and to solve part of the forementioned problems is supersymmetry (SUSY) [61]. This theory introduces, in its minimal formulation, a further symmetry that connects fermions and bosons: in particular, each fermion is accompanied by a bosonic supersymmetric partner, and for each boson, a fermionic partner is present. Supersymmetric partners are in general called “sparticles”. This doubled particle multiplicity has implications related to the fine-tuning problem, where the big radiative corrections are naturally (but only partly) cancelled out thanks to loops originated by sparticles which add contributions with an opposite sign with respect to the ordinary particles. Moreover, as visible in Figure 1.10, within the SUSY model, the running couplings tend to converge to a common value at certain high scales and this might be interpreted as an important step towards a unified theory, which is able to interpret all interactions within the same framework. The SUSY theory should be a broken symmetry since obviously there is no perfect match between ordinary and supersymmetric particles and different mechanisms of spontaneous SUSY breaking have been conjectured. The SUSY model would provide also a candidate for dark matter: the neutralino, recognized as the lightest supersymmetric particle. Worth to mention is also the fact that SUSY would automatically imply the existence of multiple Higgs bosons; the boson, discovered by CMS and ATLAS, might thus be the lightest component of a Higgs multiplet. Many results have been obtained by CMS and ATLAS and, since no SUSY has been observed so far, various limits are available for the sparticle masses. Different theories have been also formulated in order to solve these problems of the SM: one of these, which is currently under debate, is the “extra-dimensions” model [62], which conjectures the existence of a higher number of dimensions in which the particle fields can propagate.

In the neutrino sector, a theory of the oscillations has been elaborated in a similar manner as the one for the quarks [57] and various experiments are running for the measurement of the mixing coefficients, while astrophysical models, which include a cosmological constant, try to address a possible candidate for dark energy.

Indeed, even more exotic models are discussed in the literature [64] and, in general, experimental research is ongoing in several directions to get indications and hints of New Physics as well as to validate or reject some of the proposed theories. In addition, a quantum theory of gravity, which is not yet included in the SM, is also under formulation. The final aim is, indeed, to build up a Grand Unified Theory (GUT), able to describe and explain all possible phenomena observed in nature, with an as small as possible number of free parameters. This is indeed the direction entered by the “string theory” [65], with a very fascinating model of particles seen as vibrating strings.

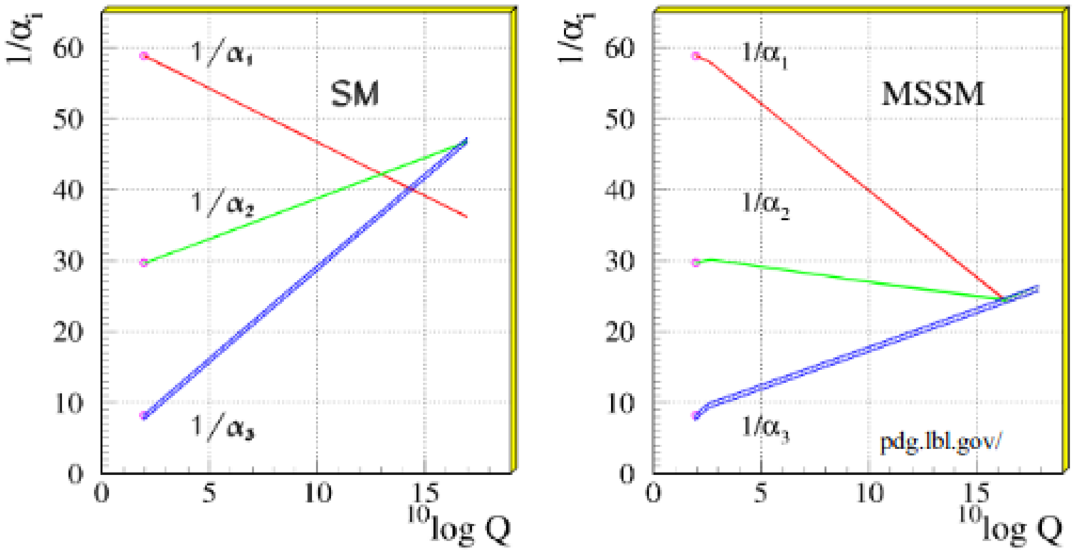


Figure 1.10: Running of the couplings for each interaction as a function of the energy scale to which the particles are probed [63]. The trend of the couplings are shown within the SM (left) and in the minimal SM extension which includes SUSY (right). Blue, green and red curves show the QCD, weak and electromagnetic couplings.

Chapter 2

A hadronic collision

As explained in Chapter 1, the theory of QCD is able to explain a wide collection of phenomena in particle physics, initiated by strong interactions between partons and consisting in final states of hadronic jets. Many QCD predictions concerning deep inelastic scattering of electron beams colliding on protons and jet physics give a good agreement over a large number of measurements. It becomes thus crucial to study how partons interact between each other at high energies, namely to study collisions between partons. Since partons are not free in nature due to the confinement property, the most natural way to make partons collide is to use beams of particles, for instance protons, which contain quarks and gluons (see Chapter 1). In order to give a satisfactory view of a *pp* collision, or more generally of a hadronic collision, the simple picture of only two interacting partons is not sufficient. This is due to the fact that the colliding hadrons are composite objects: the two partons, one for each colliding hadron, that take part in the interaction, are accompanied by other quarks and gluons that rearrange themselves after the collision in colourless objects under the effect of QCD. In addition, the possibility that additional partons of the two colliding hadrons may interact, is not excluded. The complexity of a hadronic collision can be represented, for simplicity, by distinguishing two different elements:

- the Hard Scattering which represents the hardest part of the collision, where two partons in the initial state interact by exchanging a high transverse momentum (see Section 2.2);

- the Underlying Event (UE) which describes the whole remaining part of a hadronic collision, namely the interaction and the evolution of the partons that do not take part in the hard scattering (see Section 2.4).

In this chapter, these two components are discussed. Particular attention is given to how they are combined in the whole picture of a hadronic collision, and how they are implemented in simulations. The focus is then drawn to the main topic of this thesis, namely the Multiple Parton Interactions: phenomenology and experimental results are described from the literature and an introduction to the physics interest of the experimental measurement, described in the following chapters, is provided.

2.1 Formalism of a hadronic cross section: the factorization theorem

One of the most important quantities to evaluate in a hadronic collision is indeed the cross section σ of a certain process: it represents the probability that the two hadrons interact and give rise to that specific process. For any cross section calculation, the fact that the hadrons are composite is treated by the theory with an idea, which combines the two components of the collision (hard scattering and UE). It is assumed that the hard scattering process occurs in a timescale much shorter than the one characterizing the interactions of the other "spectator" partons. In a pictorial view, the hard probe takes a snapshot of the projectile hadron structure at the instant of the interaction, with the other partons, which do not participate, seen as "frozen". This is the basic idea of the factorization theorem [66], that states the independence between short-distance hard processes, included in the hard scattering, and softer components, contained in the UE. Mathematically, this translates into the factorization of the hard scattering and the UE parts. Thus, it is possible to write any hadron-hadron cross section as a product of two non-perturbative process-independent parton density functions (PDF), one for each of the colliding protons, and a perturbative parton-parton cross section[1]. In a compact formula, this can be expressed as:

$$d\sigma_{\mathrm{h1h2}} = \sum_{\mathrm{i,j}} \int_0^1 dx_1 \int_0^1 dx_2 \sum_{\mathrm{f}} \int d\Phi_{\mathrm{f}} f_{\mathrm{i/h_1}}(x_{\mathrm{i}},\mu_{\mathrm{F}}^2) f_{\mathrm{j/h_2}}(x_{\mathrm{j}},\mu_{\mathrm{F}}^2) \frac{d\sigma_{\mathrm{ij}\to\mathrm{f}}}{dx_{\mathrm{i}}\, dx_{\mathrm{f}}\, d\Phi_{\mathrm{f}}} \tag{2.1}$$

where $x_{\mathrm{i,j}}$ is the fraction of longitudinal momentum of the colliding hadrons carried by the parton (i,j), i,j are indices that run over all possible parton types in the incoming hadrons, while f stands for all possible partonic final states with Lorentz-invariant phase space Φ_{f}. The parton density function, $f_{\mathrm{i/h_1}}$, parametrizes the distributions of the partons inside each hadron: it represents specifically the probability of finding a parton of flavour i at a certain longitudinal momentum fraction x_{i}, when the hadron, that contains it, is probed at a certain scale μ_F. The PDF are non-perturbative functions which are not calculable a priori. However, they can be extracted from fits to the data, mainly from *ep* collisions, and various PDF sets are available for each parton flavour going down to very low scales. Then, from each initial scale, it is possible to evaluate the PDF evolution to any other scale through the evolution equations defined in Chapter 1. Examples of these sets are CTEQ6L1 [67], MRST [68] and HERAPDF [69]. Figure 2.1 shows some examples of currently used PDF sets; the variation of the gluonic distribution function is also shown at different scales. It is shown that the valence quarks, namely the quarks that determine the macroscopic properties of a hadron, are mostly present inside the proton at high x values (> 0.01), while the gluons are almost absent. In the low-x region, instead, the gluon densities increase rapidly and they dominate over the quarks by several orders of magnitude. Antiquark and heavy quark PDF are very small compared to gluons and light quarks at all scales.

In the partonic cross section only the hard scattering part is considered, where one parton of a colliding hadron interacts with another parton of the other hadron and it is calculable within the perturbation theory as a series in the running coupling α_{S}. A more detailed description is provided in Section 2.3. The separation line between these two components, namely the partonic cross sec-

---

[1] In a lepton-hadron collision, this formalism is still valid but only one parton density function is obviously included.

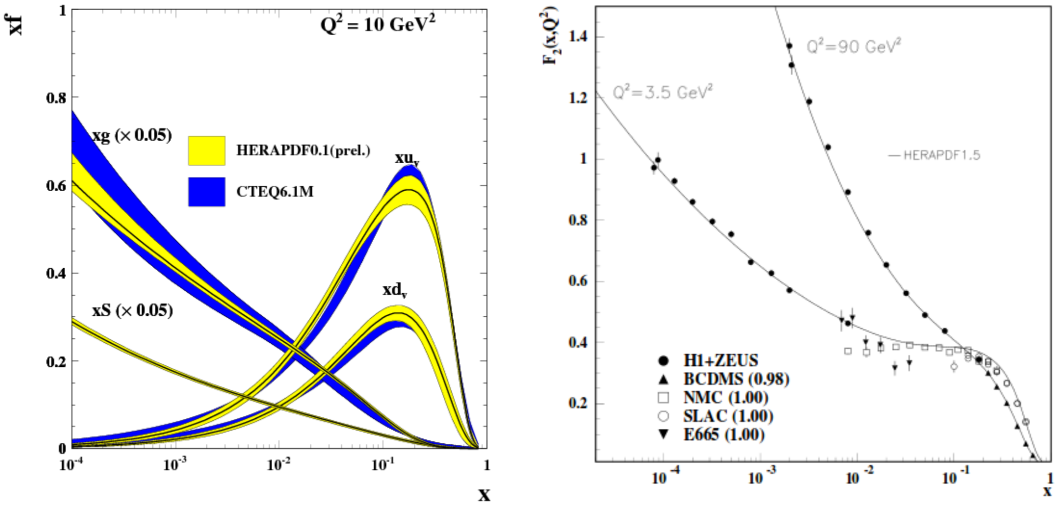


Figure 2.1: (Left) Parton distribution functions obtained by the HERAPDF and CTEQ collaborations at a scale Q^2 = 10 GeV^2 with uncertainty representing the 68% confidence level [70]. (Right) Gluon parton distribution functions measured at two different scales, Q^2 = 3.5 GeV^2 and Q^2 = 90 GeV^2, by the HERAPDF collaboration with deep inelastic scattering data [29].

tion and the PDF, is the so-called "factorization scale" (μ_{f}). The value of the factorization scale is arbitrary but a common choice is to take μ_{f} equal to the hard scattering scale [71]. However, no universal value can be identified and it is always important to evaluate the effects of changes in μ_{f} for the calculated cross section values.

A further assumption that is generally done inside the formalism of the factorization theorem is that the transverse momenta of the incoming partons, involved in the hard interaction, are neglected; this is called "collinear approximation" and it is motivated by the fact that at high energy scales, transverse momenta of the partons in the initial state can be neglected. This kind of approach is able to describe a wide range of data. However, it starts to fail when gluons at low-x values are probed: in fact, low-x values translate into longitudinal momenta which are no longer larger than the transverse momenta of these partons. In this region, a dependence on the partonic transverse momentum needs to be explicitly expressed in both the partonic cross section and in the PDF. In particular, in this case, uPDF need to be used for calculations in the low-x regime.

2.2 Hard scattering and matrix element

In the cross section formulation, the matrix element (ME) is the mathematical expression which represents the hard scattering occurring during the collision. The hard scattering might be thought of as an interaction between two partons in the initial state, which produces a final state with two or more partons. Considering an arbitrary final state F and a state f with only two final partons, the generic total cross section, defined in Equation 2.1, might be expanded, by including explicitly all orders, as:

$$\sigma_{\mathrm{F}} = \sum_{\mathrm{k}=0}^{+\infty} \int d\Phi_{\mathrm{f+k}} | \sum_{\mathrm{l}=0}^{\infty} M^{(\mathrm{l})}_{\mathrm{f+k}}|^2 \tag{2.2}$$

Note that the terms relative to the proton PDF have been dropped for compactness and the focus is drawn to the hard scattering part. $\Phi_{\mathrm{f+k}}$ is the phase space for a final state f with k additional partons (legs), while $M^{(l)}_{\mathrm{f+k}}$ is the ME, namely the amplitude, for producing the same final state f with k legs and with l additional loops. A leg is a real correction to the total cross section, while a loop is called a virtual correction, because it does not produce any particle in the final state but only enters in the calculations (see Figure 2.2). By choosing certain values for legs and loops, namely for k and l respectively, different processes are included in the analytical calculation of the cross section: in particular, it is generally said that the calculation is performed at a fixed order in α_{S}. The order in α_{S} refers to the number of QCD vertices included in the calculation of $M^{(l)}_{\mathrm{f+k}}$. Real and virtual corrections increase the order in α_{S} of, respectively, one and two units. Increasing the number k means that a higher number of real partons is included in the final state, while with the number l, loop virtual corrections are taken into account into the cross section calculation, without increasing the number of partons in the final state. Depending on different choices on k and l, various cases can be identified for the final state f:

- k=0, l=0: production of the final state f at Leading Order (LO);
- k=n, l=0: production of $f + n$ jets at Leading Order (LO);
- $k + l \leq n$: production of f at N$^{\mathrm{n}}$LO which includes N$^{\mathrm{n-1}}$LO for $f + 1$ jet, N$^{\mathrm{n-2}}$LO for $f + 2$ jets and so on until LO for $f + n$ jets.

It is, for instance, referred to as LO 2→2 processes if k=0 and l=0, LO 2→3 processes if k=1 and l=0, and NLO 2→2 processes if k+l $\leq$ 1.

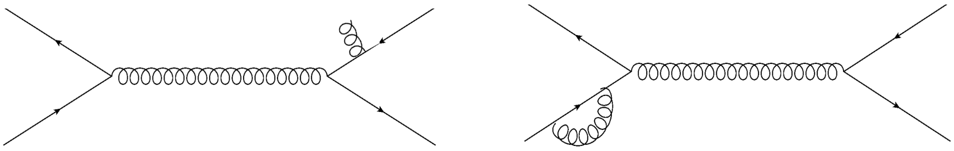


Figure 2.2: Example of QCD diagrams of two initial quarks which annihilate into a gluon which further splits into two final quarks [72]: in the left plot, a real gluon is emitted by one of the final quarks, while in the right plot an initial quark emits and reabsorbs a virtual gluon. The left figure is an example of real correction to a 2→2 process, while the figure on the right includes a virtual correction for the same process.

Since the cross section is an infinite sum in α_{S}, by choosing finite values for k and l, higher order contributions are left out from the matrix element calculation; however, they are not ignored but the *leading log* contributions are absorbed into the parametrization of the PDF and its evolution starting from a certain scale, given by μ_{f}. QCD calculations at LO normally suffer from large higher order corrections and scale uncertainties and the more terms are included into the matrix element calculation, the smaller the dependence on the choice of the factorization and renormalization[1] scales is. On the other hand, the degree of complication in the calculations increases dramatically with any further order that is added. A naive way to improve LO calculations is to correct them with an overall factor corresponding to the ratio of cross sections at NLO and LO for a given final state. This is the so-called "K-factor", defined as:

$$K = \frac{\sigma_{\mathrm{F}}^{\mathrm{NLO}}}{\sigma_{\mathrm{F}}^{\mathrm{LO}}} \tag{2.3}$$

---

[1] The renormalization scale is an arbitrary parameter introduced in the theory to treat divergences appearing in loop diagrams.

By multiplying it to the LO cross section, virtual corrections are aimed to be corrected for and K-factors for several processes have been calculated [73]. A problem of this procedure is that often NLO calculations result in a shape change of the studied observable and this effect is not taken into account by the K-factor.

2.2.1 Matching between matrix element and parton shower

Since an analytical calculation of the ME at all orders in α_S is not doable, the missing diagrams which are not included, are computed by the PS. The issue of combining ME calculations with the PS is addressed by the "matching procedure". An important role of a matching procedure is to remove the possibility of double counting terms present in both the matrix element and the PS expansion. There are many methods to perform that. After the production of a $2 \rightarrow 2+n$ process in the ME, the PS is invoked for each of the final legs. The most used matching schemes [74] are the CKKW [75] and the MLM [76]: the former is based on a shower veto, the so-called "truncated shower", which generates additional partons only below the lowest scale of the ME, while the latter works with an event rejection, where an event is discarded if a parton is emitted above the scale of the partons in the ME. Additional ways have been also recently developed, mainly to perform the matching of NLO ME with PS: these are the POWHEG method [77], based on the generation of the first hard emission which sets the upper scale for the rest of the shower, and the MC@NLO method [78], where some part of the shower is subtracted from the total cross section to avoid double counting. These methods are widely used in Monte Carlo event generators (see Section 2.5), to get predictions for higher-order ME or NLO ME, interfaced with PS and UE.

2.3 Definition of the Underlying Event and Multiple Parton Interactions

The UE is a very important component in a hadronic collision, including most of the occurring partonic interactions; it can be easily represented as everything that happens in the collision, but the hard scattering. This general definition stems from the fact that it includes all the different elements which underlie (as the name suggests) the primary hard interaction. In particular, different subphenomena are grouped under the name of the UE:

- Initial- and final-state radiation which includes the emission of additional particles by partons in the initial or in the final state, namely before or after the hard scattering;
- Beam-beam remnants (BBR) which group the colour interactions undergone by the spectator partons in each of the two colliding hadrons;
- Multiple parton interactions (MPI) which contain the whole of additional interactions between partons of the colliding hadrons, occurring together with the hard scattering.

The contribution provided by these components to the collision events is generally softer than the hard scattering, but still calculable within the perturbative framework. A sketch of a hadronic collision as a whole is shown in Figure 2.3 where the different components are represented in various colours.

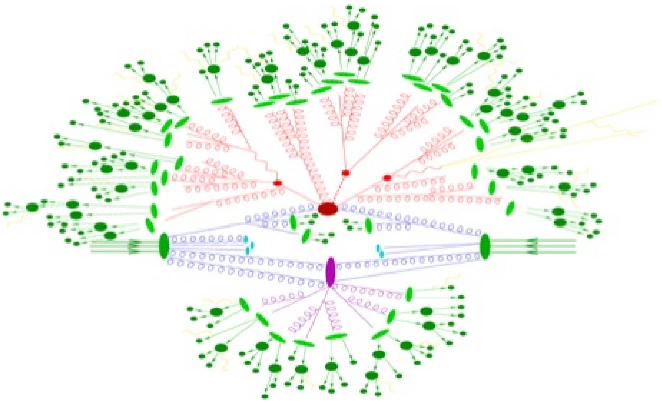


Figure 2.3: Sketch of a hadronic collision with all the contributions included [79]: the hard scattering is represented in dark red, the initial- and the final-state radiation is pictured respectively in blue and light red, a MPI event in violet and the beam remnants are sketched in cyan. At the very end of the lines, the hadronization products are shown in light and dark green.

The initial- and final-state radiation comprises quarks and gluons that are emitted by the partons that take part in the hard scattering respectively before and after the interaction. As seen in Section 2.2, if initial- and final-state radiation is included in the ME calculation by adding additional legs, they have to be considered as a part of the hard scattering; this case mainly addresses n-objects final states, with $n > 2$, at large p_{T}. However, in general, for 2→2 processes, initial- and final-state radiation needs to be considered under the UE classification, and plays a key role in both soft and hard UE components. The effect and the impact of initial- and final-state radiation has been studied at leptonic and hadronic colliders. It is possible to disentangle the corresponding contributions: for the study of final-state radiation, it is important to select a leptonic initial state with a "coloured" final state (like in [80], for quark production in electron-positron collisions). A hadronic initial state with a leptonic final state (like in [81], for Drell-Yan production in pp collisions) is necessary for measurements related to initial-state radiation.

The BBR are the outcome of the rearrangement of the "spectator partons" of the two colliding protons. The spectator partons are the ones that do not participate in the hard scattering. Since they carry colour charge, they need to form coloured objects after the hard scattering. The main contribution of the BBR is concentrated very closely towards the initial directions of the incoming protons and it takes most of the energy of the collision energy. Hence, it is often detected only in the part of the detectors close to the beam pipe. Beam-beam remnants mainly contribute to the soft part of the UE.

The MPI describe the possibility to have more than one distinct and simultaneous parton interaction inside the same hadronic collision. The MPI are mainly due to the compositeness of the incident hadrons that may be thought of as bunches of partons whose wave functions overlap with each other. In presence of a hard partonic collision between the two hadrons, there is a high probability that the spectator partons experience a hard interaction as well; the scale of a MPI is, indeed, lower than the scale of the hard scattering but large enough to have a relevant and considerable contribution to the final state. MPI predominantly contribute to the soft and semi-hard UE components but some of these additional interactions, in a lesser extent, can reach a high scale which gives also a hard component to the UE (see Section 2.5).

Historically, the existence of the MPI has been conjectured from the observation that the total parton-parton cross section, σ_{hard}, tends to rapidly increase at low values of exchanged transverse momentum. In fact, if σ_{hard} is represented as an integral of a generic partonic cross section, $\hat{\sigma}$, over the exchanged transverse momentum between the two partons, from an initial cut-off to the maximum value allowed by energy-momentum conservation, it can be written as:

$$\sigma_{\mathrm{hard}}(p_{\mathrm{T_{min}}}) \propto \int_{p^2_{\mathrm{T_{min}}}}^{\mathrm{s}/4} \frac{d\hat{\sigma}}{dp^2_{\mathrm{T}}} dp^2_{\mathrm{T}} \propto \frac{1}{p^2_{\mathrm{T_{min}}}} \tag{2.4}$$

where $p_{\mathrm{T_{min}}}$ is the threshold of the exchanged transverse momentum between the two partons. It is obtained that, at $p_{\mathrm{T_{min}}}$ values around 3-5 GeV which is still in the perturbative regime, σ_{hard} exceeds the non-diffractive cross section σ_{nd}, measured to be about 70 mb at 7 TeV [82, 83]. Formulated in this way, this might seem indeed a paradox; it would result that the total *pp* cross section, at a relatively high scale, is lower than the cross section of two of their constituents. The possibility of having MPI solves this issue. If more than one partonic interaction in a hadron-hadron collision is assumed, the consistency is actually restored. In fact, σ_{nd} indicates the hadron-hadron cross section, conceived as a whole of multiple interactions where partons from the different colliding hadrons might simultaneously interact between each other. In this new formulation, σ_{hard} is not an event but a jet quantity, since partons are not free but only contained in hadrons; σ_{hard} becomes meaningful to give an estimation of the number of MPI through the relation:

$$\langle n \rangle = \frac{\sigma_{\mathrm{hard}}(p_{\mathrm{T}})}{\sigma_{\mathrm{nd}}} \tag{2.5}$$

where n follows a Poisson distribution. However, σ_{hard} still diverges for $p_{\mathrm{T}} \rightarrow 0$. This would mean that the number of MPI becomes rapidly infinite when going to very low exchanged p_{T}. This problem is solved by introducing effectively a p_{T} cut-off in the formula for σ_{hard}, motivated by the fact that, since the hadrons are colour-neutral objects and the parton wavelength increases at lower p_{T}, the partons at relatively low p_{T} ($\sim$ 1 GeV) can no longer resolve the individual colour charges and the coupling is decreased in this region. This matter is equivalent to include a limit in the gluon density function, in order to tame its increase at low-x values: this mechanism goes under the name of “gluon saturation” [84].

The effects of MPI naturally tend to increase for higher energies, where the low-x region becomes accessible and where the parton densities are very high. This is why they have been measured at past and present colliders and they gain more and more importance and attention at future ones [85].

2.3.1 Measurements of UE and MPI at colliders

The concept of MPI has been successfully applied to describe the properties of soft collisions, as well as the details of jet final states. A wide collection of measurements that exhibited sensitivity to MPI is available at several hadron colliders and at various energies. In particular, the presence of MPI tends to increase the general event activity, in terms of number of particles and energy. One of the most popular ways to measure and to estimate the MPI contribution is the measurement of the UE “a-la-Rick-Field” [86]. This method has been introduced for the first time in the CDF collaboration and then applied to similar measurements at other colliders. It consists of measuring the event content in terms of charged particles in different regions of the phase space relative to the direction of the hardest objects in the event. In particular, the hard object (a jet in case of measurements of UE in hadronic events, a dimuon pair in case of Drell-Yan events) identifies a direction ϕ_{hard} in the transverse plane. The transverse plane is then divided into four regions, as shown in figure 2.4(left), which are commonly referred to as:

- toward region: $|\phi - \phi_{\mathrm{hard}}| < \pi/3$
- transverse region: $\pi/3 < |\phi - \phi_{\mathrm{hard}}| < 2\pi/3$
- away region: $|\phi - \phi_{\mathrm{hard}}| > 2\pi/3$

where ϕ is the azimuthal angle considered in the transverse plane. In these regions, the observables which are generally measured are the multiplicity and the p_{T} sum of the charged particles in every event[1]. Measurements of these observables are performed as a function of the p_{T} of the hard object: this gives a dependence on the scale of the scattering that may be easily thought of as the centrality of the collision between the two hadrons. Assuming that the direction of the hardest objects may be identified in good approximation with the direction of the hard scattering, every region in the transverse plane is affected by different components of the collision. The toward region contains the products of the hard scattering and most probably the contribution of the initial- and final-state hard radiation. The transverse region contains the products of the MPI and the BBR along with additional soft radiation. The away region measures the area of the recoiling object, namely the opposite region with respect to the leading object. A more sophisticated approach has been also recently considered for this type of measurements by considering separately the two transverse regions in terms of event content [87][2]. This approach, explained more extensively in Appendix D, helps to further separate the components that contribute to the content of the transverse region.

Hadronic event measurements have been performed as a function of the leading jet p_{T} and of the leading track p_{T} by the CDF [88], the CMS [89] and the ATLAS [90] collaborations for different energies. The curves show always a rapid rise at low p_{T} and a plateaux region at higher p_{T} starting from 5-10 GeV, depending on the collision energy. All these measurements have shown the importance of the implementation of the MPI in the event generators to give a good and consistent description of the data. Predictions obtained without MPI significantly fail to reproduce the measurements. An example of this curve, which represents the charged particle multiplicity as a function of the hard scale, measured by CMS at 7 TeV, is shown in Figure 2.4(right).

Similar measurements have been performed for different processes occurring at the hard scattering, for instance with Drell-Yan events [81] and top-antitop quark pair production [91][3]; a very good description of the data is only achieved with models which include MPI. Other evidences of MPI come also from a wide collection of different measurements, related to jets [93], particle event contents [94] and energy flow [95]. Thanks to all these observations and measurements, the concept of MPI is well established and well received in the scientific community.

---

[1] These observables are chosen because they are experimentally accessible down to low p_{T} and easy to measure with a tracking system.

[2] By separating them in the region with higher and lower activity, it is possible to disentangle even further the UE components.

[3] Also a UE measurement in $b\bar{b}$ pair events is planned in the CMS collaboration and studies at the generator level have been performed [92].

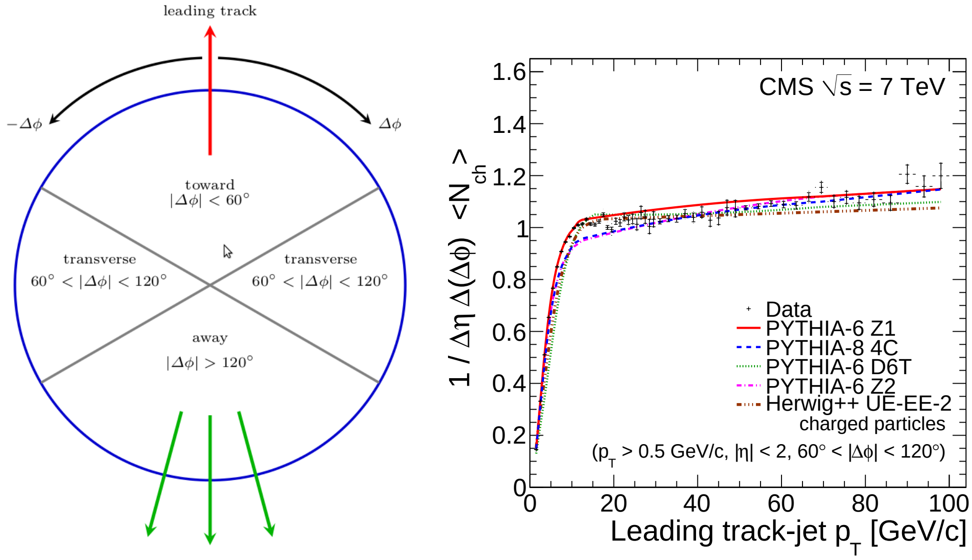


Figure 2.4: (Left) Transverse plane division used for measurements of UE contribution [90]: the direction and the scale of the hard scattering are defined by the leading object in the event and the xy-plane is thus divided in three regions, depending on their azimuthal angle with respect to that object. These regions are called toward, transverse and away. (Right) Example of UE measurement performed with the method "a-la-Rick-Field" [89]: charged particle multiplicity, as a function of the p_T *of the leading track, measured at the CMS experiment at 7 TeV, in the transverse region. The measured curve is characterized by a rising part at low scales and a plateaux region for increasing ones.*

2.4 Double Parton Scattering

In a framework where the MPI are so important in order to explain such a big variety of measurements, even additional scattering at a larger scale may accompany the first hard interaction. In fact, nothing prevents a MPI to reach a scale comparable to the one of the hard scattering. If this occurs, with one additional hard scattering, the literature generally refers to a Double Parton Scattering (DPS)[1]. The understanding of DPS is particularly relevant for the estimation of backgrounds in measurements of specific or rare physics channels, especially the ones involving jets in the final state. A good comprehension of the DPS dynamics is also of great help for the general phenomenology, with increasing importance for higher collision energies, and for the development of partonic models of hadrons. Likewise a MPI measurement which can be performed by looking at average observables, like charged particle multiplicity or p_T event content, a detection of a DPS event is not possible in an event-by-event basis; however, it is possible to identify specific channels and region of the phase space where its contribution is enhanced. The possibility to produce high-p_T physics objects from DPS came up together with the concept of MPI and several phenomenological studies have been performed by scanning different channels and defining sensitive observables for its discrimination. Already in 1982 [96], a first attempt of DPS formalism for a multiquark scattering treatment was discussed and first ideas related to double Drell-Yan production [2] have been developed. Studies of sensitivity to DPS for various physics channels are widely documented in the literature: in pp collisions, four-jet final states [3, 97, 98], W-boson production associated with jets [4], di-boson [99] and double J/ψ [100] final states have shown promising outcomes for specific regions of the phase space. The capability to disentangle a DPS signal from the background is

---

[1] Events with a higher number of additional interactions might be also considered and they are thus called "Triple, Quartic, .. Parton Scattering" but their contribution becomes very small. In this work, only the actual DPS will be treated.

discussed in proton-lead [101, 102] and $p\bar{p}$ [103] collisions as well.

2.4.1 Theory of Double Parton Scattering

Double Parton Scattering phenomenology is based on the general expression for the cross section, introduced in Equation 2.1, adapted for multiple scattering at the matrix element level. The new parton distributions include information of multiple parton structure of the proton and refer to the probability distributions of finding two partons with given longitudinal momentum fraction at a given scale and with a certain separation distance in the transverse plane; hence, these new PDF are called double parton distribution functions. In this formalism, $\sigma^{\mathrm{DPS}}_{(\mathrm{A,B})}$ is defined as the cross section of a DPS event in hadronic collisions, where two independent processes A and B occur simultaneously. It can be expressed as:

$$\begin{aligned}\sigma^{\mathrm{DPS}}_{\mathrm{A,B}} = \frac{m}{2}\ \Sigma_{\mathrm{i,j,k,l}} \int & dx_1 dx_2 dx'_1 dx'_2 d^2b\ \Gamma_{\mathrm{i,j}}(x_1, x_2, b; t_1, t_2) \\ & \Gamma_{\mathrm{k,l}}(x'_1, x'_2, b; t_1, t_2)\ \hat{\sigma}^A_{\mathrm{i,j}}(x_1, x'_1, t_1)\ \hat{\sigma}^B_{\mathrm{j,k}}(x_2, x'_2, t_2)\end{aligned} \tag{2.6}$$

where $x_{1,2}$ ($x'_{1,2}$) are the longitudinal momentum fractions of the two partons of the first (second) proton, $t_{1,2}$ are the scales characteristic of the subprocesses A and B, and b is the distance parameter between the two partons. The two $\Gamma(x_1, x_2, b; t_1, t_2)$ represent the double parton distribution functions and the two $\hat{\sigma}$ are the inclusive partonic cross sections of the two processes. The quantity m is a symmetry factor that is equal to 1 if A = B, and 2 otherwise. This definition for $\sigma^{\mathrm{DPS}}_{\mathrm{A,B}}$ is process- and energy-independent. If the dependence on the transverse distance factorizes with the rest of the parton distribution, like:

$$\Gamma_{\mathrm{k,l}}(x'_1, x'_2, b; t_1, t_2) = D_{i,j}(x_1, x_2, t_1, t_2) \cdot F(b), \tag{2.7}$$

the equation becomes:

$$\begin{aligned}\sigma^{\mathrm{DPS}}_{\mathrm{A,B}} = \frac{m}{2\sigma_{\mathrm{eff}}}\ \Sigma_{\mathrm{i,j,k,l}} \int & dx_1 dx_2 dx'_1 dx'_2\ D_{\mathrm{i,j}}(x_1, x_2; t_1, t_2) \\ & D_{\mathrm{k,l}}(x'_1, x'_2; t_1, t_2)\ \hat{\sigma}^A_{\mathrm{i,j}}(x_1, x'_1, t_1)\ \hat{\sigma}^B_{\mathrm{j,k}}(x_2, x'_2, t_2)\end{aligned} \tag{2.8}$$

with $\sigma_{\mathrm{eff}} = \left[\int d^2b(F(b))^2\right]^{-1}$. Note that $F(b)$ represents the dependence of the DPS cross section on the relative parton position inside the proton and it is thus referred to as "matter distribution". To further simplify the formalism, the double parton distribution functions can be reduced to two independent single ones, namely $D_{\mathrm{i,j}} = D_{\mathrm{i}} \cdot D_{\mathrm{j}}$, under the assumption of complete uncorrelation between partons of the same protons, and the DPS cross section can be finally expressed in the simple form:

$$\sigma^{\mathrm{DPS}}_{(\mathrm{A,B})} = \frac{m}{2} \frac{\sigma_{\mathrm{A}}\sigma_{\mathrm{B}}}{\sigma_{\mathrm{eff}}} \tag{2.9}$$

With these approximations, σ_{eff} carries information about the size of the partonic core of the proton and might be thought of as related to the transverse size of the proton. Its dimensions are the ones of an area (i.e. a cross section). In the most trivial approximation with no correlations between partons and assuming a uniform matter distribution, σ_{eff} should be equal to the total inelastic cross section ($\sim$ 70 mb).

It has been demonstrated in [104], that this formula remains valid only if inclusive scenarios are considered, namely selections where no vetoes are applied for additional objects in the final state and where a number of hard scatterings greater than two is allowed. If this is the case, then the resulting σ^{DPS} presents an explicit dependence on the impact parameter b and the formula needs to be modified [104].

However, under discussion is the possibility to improve the DPS picture and to leave out some of the currently adopted approximations. Indeed, it is questionable that hard scatterings in the same event may occur independently and without any correlation. Interesting conjectures have been published about the transverse impact parameter dependence in the double parton distribution functions [105, 106, 107] and spin correlations between the partons inside the same hadron [108, 109, 110, 111]. Even though these theories are still at a speculative stage, they predict significant consequences in both the DPS cross section values and the event topology[1] and they could give many insights on the understanding of the hadron structure. The final goal is to find a consistent picture that would be able to explain multiple interactions at all possible scales, from soft ones to the ones with hard scatterings.

2.4.2 Measurements of Double Parton Scattering at colliders

Evidence for DPS events has been observed in various channels and at different energies. Final states involving four jets have been measured by the AFS collaboration at the CERN ISR [5], as well as γ + 3 jets by the D0 [7] and CDF [112] collaborations in $p\bar{p}$ collisions. More recently, the ATLAS [9] and CMS [8] collaborations have contributed with the W+dijet channel and the D0 collaboration with the γ+b/c jet+2 jets [113]. Furthermore, the double J/ψ production measurement performed by the CMS [114] and the LHCb [100] experiments may also be used to improve the understanding of DPS. Conclusions of this collection of measurements have been that the addition of a DPS signal in the simulation helps to get a better agreement with the data. All of them, except the double J/ψ measurement in CMS, have provided an estimate of σ_{eff} in the theoretical framework of Equation 2.9, along with distributions that are sensitive to the DPS signal. The DPS-sensitive observables use information of the topology of the event and the configuration of the physics objects in the final state (see Chapter 4). A technique used for the extraction of the σ_{eff} in CMS and ATLAS, is the so-called "template method" and it is based on a fitting procedure to DPS-sensitive observables. Two templates are considered in the fit: one for the background and one for the DPS signal. From the relative fraction of the two templates that fit the observed results best, a determination of σ_{eff} can be performed. The application of this method will be treated in Chapter 11.

From the experimental point of view, it becomes important to provide measurements of observables sensitive to DPS and, from these data, to extract a value of σ_{eff} by following the definition provided in Equation 2.9. Even if the approximations used in that case are too simplicistic, an extraction of σ_{eff} with these assumptions for different channels and different energies is meaningful: if results of σ_{eff} are different for the various channels, it would be a clear indication that this simplified model needs to be improved, e.g. by introducing correlations between the partons. The

---

[1] For instance, in particular scenarios with large spin correlations between the initial-state partons, a modulation on the relative azimuthal angle of the final-state partons appears, while the inclusion of colour correlation adds extra colour factors to the DPS cross section which would change its absolute contribution.

current situation of the measured σ_{eff} values for different channels and as a function of the collision energy, $\sqrt{s}$, is represented in Figure 2.5. Uncertainties, mainly of systematic nature related to the extraction method, are rather big and they do not yet allow a conclusive statement about the channel and energy dependence. The value of σ_{eff} seems to be around 15-20 mb at $\sqrt{s} = 7$ TeV.

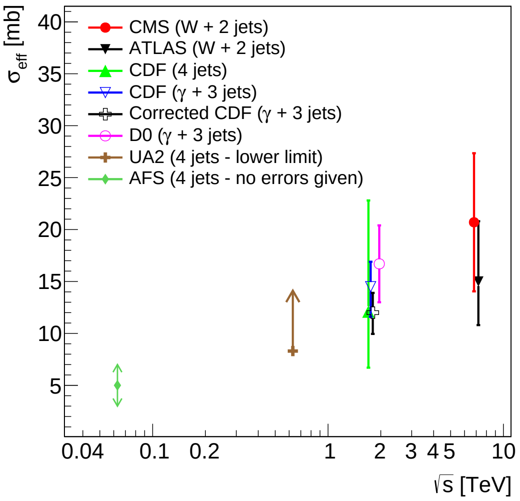


Figure 2.5: Measured σ_{eff} values as a function of the collision energy [8]: data points come from measurements at different experiments. The physics channel used for the measurement is also indicated in the plot for each point.

The work, described in this thesis, contributes in this research topic, by measuring DPS-sensitive observables for two channels involving jets in the final state: one with four jets, and one with two b-jets and two additional ones. A new approach for σ_{eff} extraction is described and applied for the first time to the measured channels. In the following Section, details of these two channels are described, with a particular focus on the physics interest related to DPS.

2.4.3 DPS in four-jet channels

The physics channels which are described in this thesis are related to a four-jet scenario. Four jets might be produced by two different processes occurring in the collision: a Single Parton Scattering (SPS), represented in Figure 2.6 with a pictorial view and with the specific Feynman diagrams, and a DPS, shown in the same way in Figure 2.7. In SPS events, the four jets are produced by a single chain, with all jets coming from the hard scattering; in DPS, instead, two different chains emit the two pairs of jets in the final state. These two production mechanisms translate into different configurations and topologies in the final state, with a higher correlation among the jets for SPS. Thus, by associating the jets into two pairs and investigating particular observables which are sensitive to the relative position of the jets, it is possible to separate the two different contributions. Quantitative studies about how to separate a DPS signal from a SPS background are presented in Chapter 4. In particular, two scenarios with four jets in the final state have been studied: one with exclusively four jets and one with two b- and two other jets. The possibility of tagging heavy-flavoured jets is very important for separating SPS and DPS, because it helps the association in pairs of the jets in the final state.

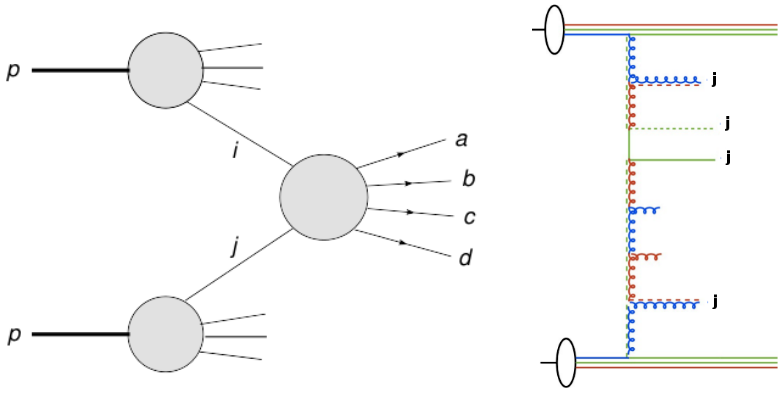


Figure 2.6: A Single Parton Scattering event in a pp collision, with four jets in the final state. In the left plot, a sketch of the active partons (i,j) producing the four objects, (a,b,c,d), in a single chain, is represented, while in the right plot, the Feynman diagram of the process is shown.

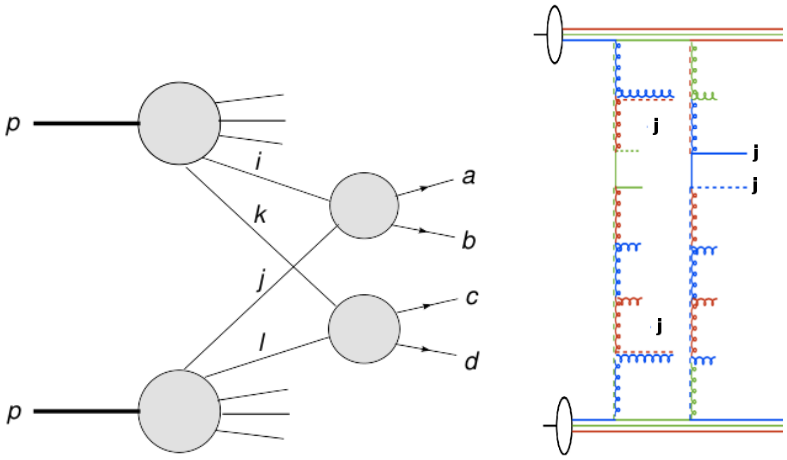


Figure 2.7: A Double Parton Scattering event in a pp collision, with four jets in the final state. In the left plot, a sketch of the active partons (i,j) and (k,l) producing the four objects, (a,b) and (c,d), in two different chains, is represented, while in the right plot, the Feynman diagram of the process is shown.

2.5 Monte Carlo event generators

The goal of a Monte Carlo (MC) event generator is to simulate everything that happens in a hadron-hadron collision. This includes the production of a hard scattering, the implementation of the PS related to the evolution of the partons and the UE. MC generators differ mainly in the implemented ME for the hard scattering process and in the UE simulation, namely how PS and MPI are produced. While in LO generators only 2→2 processes are calculated and the additional emitted partons are implemented through the PS, there are several cases where higher-order ME are implemented at the Born level. By including additional emissions in the ME, the exact kinematics and all the interference and helicity parton structures are taken into account. On the other hand, to calculate higher-order diagrams and loop corrections is a very time consuming process and a compromise needs to be done. The corrections which are not included at the matrix element level are estimated through the PS. The PS is simulated in the current generators by defining a parton probability of no-branching which can be defined as:

$$P_{\mathrm{a}}(t) = \Gamma_{\mathrm{a}}(t)e^{\left(-\int_{t_0}^{t} dt'\Gamma_{\mathrm{a}}(t')\right)} \tag{2.10}$$

The factor $e^{\left(-\int_{t_0}^{t} dt'\Gamma_{\mathrm{a}}(t')\right)}$ is the so-called Sudakov form factor which expresses the probability for a parton a at an initial scale t_0 to evolve without any branching to a successive scale t. The quantity $P_a(\mathrm{t})$ indicates the probability for the parton of evolving without any branching from t_0 to t, and of splitting at exactly the scale t. The function $\Gamma_{\mathrm{a}}(t')$ contains in its definition the parton evolution. In an event generator, the evolution starts at the hard scattering and proceeds backwards to lower scales. All the partons at the end of the evolution experience the hadronization process.

The matching between the matrix element and the PS can be performed in different ways (see Section 2.2.1).

In this thesis, the following event generators have been used for predictions to be compared to the measurements[1]:

- Pythia6 and Pythia8 [115, 116]: a LO 2→2 ME, matched with a DGLAP evolution at an approximate Next-to-Leading-Log level (NLL) for the simulation of the PS is generated; in this approximation, terms up to $P^0(z)$ are resummed to all orders of α_{S} of the evolution, with colour coherence and energy-momentum conservation constraints[2]. The PS is ordered in transverse momentum of the emitted partons: this means that partons with decreasing p_{T} are successively emitted. Initial-state radiation is generated with a backward evolution: this means that first, the hard scattering is calculated and then the incoming partons are evolved backwards down to a low scale. The hadronization is performed with the string model and the UE is simulated with specific sets of parameters which regulate colour reconnection and MPI contribution, determined by fits to measurements at colliders;

- Herwig++ 2.5 [117]: like in Pythia, a LO 2→2 ME is generated; the PS is simulated by using DGLAP evolution, where the shower is angular-ordered: this means that the radiation

---

[1] A brief description of the main features used is given here for every generator. More information can be found in the quoted references.

[2] This is why it is not just LLA but approximate NNL.

is generated coherently inside the splitting cones of the partons. The hadronization follows the cluster model and the UE is tuned to data;

- MADGRAPH 5 [118, 119]: a LO 2→2+n ME, where n is the number of additional partons included in the analytical calculation of the hard scattering is generated. It does not simulate any PS or UE contributions, which are provided by interfacing it with either PYTHIA or HERWIG++, generally through the MLM method [76];

- POWHEG [77, 120]: based on the POWHEG box [121], a NLO 2→2 ME, with an additional parton included in the matrix element, is generated. Again this generator does not include PS and UE which need to be borrowed by interfacing POWHEG with PYTHIA or HERWIG++. The ME-PS matching is performed with the POWHEG method;

- SHERPA [122]: similar to MADGRAPH, a LO 2→2+n ME is generated, but PS and UE simulation is also included. The PS follows the DGLAP evolution and it is matched to the ME with CKKM scheme. The type of hadronization used in the generator is the cluster model. Specific parameters are used for the UE in order to best describe available UE data.

A schematic summary of the matrix element used in the different generators in given in Table 2.1.

Monte Carlo event generator	Matrix element	Parton Shower	Hadronization
PYTHIA	LO 2→2	DGLAP	string
HERWIG++	LO 2→2	DGLAP	cluster
MADGRAPH	LO 2→2+n	-	-
POWHEG	NLO 2→2 + 1 hard jet	-	-
SHERPA	LO 2→2+n	DGLAP	cluster

Table 2.1: Summary of the event generators used in this thesis for comparison with the measurements: the matrix element, the parton evolution and the hadronization models used by each of them are also listed.

Note that only PYTHIA, HERWIG++ and SHERPA have their own UE generation, comprising PS, MPI and BBR. The MADGRAPH and POWHEG ME need to be interfaced to one of them for a full hadron-hadron collision simulation.

Since the work described in this thesis deals directly with MPI, in the following some current models of MPI are presented.

2.5.1 Models of Multiple Parton Interaction

In MC generators, the MPI have a key role for the description of the UE and Minimum Bias (MB) data, recorded in hadronic collisions. A big effort has been made to improve their modelling and further constrain the existing theories through comparison with the available measurements[1]. On one hand, this is important to better understand the internal proton structure, as well as the MPI mechanisms; on the other hand, the study of the contribution of MPI and DPS is fundamental to make predictions for any physics channel. In the following, the MPI models implemented in the most used MC generators are described.

---

[1] A wider treatment is given in Appendix D, where results of new UE tunes are described.

2.5.1.1 MPI in Pythia

The possibility that several parton pairs undergo simultaneous interactions is implemented in the PYTHIA event generator. The first necessary ingredient is the regularization of the partonic QCD cross section, $\hat{\sigma}$, which is divergent for $p_{\rm T} \to 0$. This is achieved by including the parameter $p_{\rm T_0}$ in the denominator, which removes the divergence, by modifying the dependence of $\hat{\sigma}$ on $p_{\rm T}$ in this way:

$$\frac{d\hat{\sigma}}{dp_{\rm T}^2} \propto \frac{\alpha_{\rm S}^2(p_{\rm T}^2)}{p_{\rm T}^4} \to \frac{\alpha_{\rm S}^2(p_{\rm T_0}^2 + p_{\rm T}^2)}{(p_{\rm T_0}^2 + p_{\rm T}^2)^2} \tag{2.11}$$

The parameter $p_{\rm T_0}$ is set to be energy-dependent and is controlled by two additional parameters, $p_{\rm T_0}^{\rm ref}$ and $E_{\rm pow}$, in the generator. The energy dependence is expressed by the formula:

$$p_{\rm T_0} = p_{\rm T_0}^{\rm ref} \cdot \left(\frac{E}{E_{\rm ref}}\right)^{E_{\rm pow}} \tag{2.12}$$

The value of the energy reference, $E_{\rm ref}$, is generally set in most of the tunes to 1800 GeV for historical reasons. This regularization is included in all MPI models implemented in PYTHIA. The basic idea of all models is that the average number of interactions per event is given by Equation 2.5. The simultaneous interactions are assumed to be independent of each other and without any dependence on the relative position of the partons inside the hadrons. The PYTHIA event generator implements it by defining:

$$f(x_{\rm T}) = \frac{1}{\sigma_{\rm nd}(s)} \frac{d\sigma}{dx_{\rm T}} \tag{2.13}$$

that represents the probability to have a parton-parton interaction at a given $x_{\rm T}$=$2p_{\rm T}/E_{\rm cm}$, given that the two hadrons undergo an inelastic collision. From this definition, a natural ordering in the hardness $x_{\rm T}$ of the scatterings is automatically set; the probability to have the hardest interaction at a scale $x_{\rm T_1}$ is obtained by:

$$P(x_{\rm T_1}) = f(x_{\rm T_1}) e^{\left\{-\int_{x_{T_1}}^{1} f(x'_T)dx'_{\rm T}\right\}} \tag{2.14}$$

that includes the probability to have an interaction at $x_{\rm T_1}$ and not to have any interaction at scales larger than that value[1]. For the ith scattering, the probability is obtained by iteration:

$$P(x_{{\rm T}_i}) = \frac{f(x_{{\rm T}_i})}{(i-1)!} \left(\int_{x_{{\rm T}_i}}^{1} f(x'_T)dx'_{\rm T}\right)^{\rm i-1} \cdot e^{\left\{-\int_{x_{T_1}}^{1} f(x'_T)dx'_T\right\}} \tag{2.15}$$

The chain of interactions terminates when $x_{{\rm T}_i}$ becomes smaller than $p_{\rm T_{min}}/E_{\rm cm}$.
In this mechanism, energy-momentum conservation becomes necessary and needs to be introduced. This is ensured by rescaling the parton distributions after every interaction by a factor:

$$x'_{\rm i} = \frac{1}{1 - \sum_{j=1}^{i=1} x_{\rm j}} \tag{2.16}$$

This takes into account the energy already used in previous interactions and it decreases the amount of phase space available for further interactions, until no energy is left. For the hard scattering,

---

[1] This formalism uses the same concept of the Sudakov form factor in the parton evolution framework.

namely the first interaction of the chain, the ordinary parton distributions are used.

Each MPI is associated with its set of initial- (ISR) and final-state (FSR) radiation as well as the partons participating in the hard scattering. In order to combine these different components, the previous formalism is implemented in an interleaved evolution between MPI, ISR and hard scattering process, following the equation given below:

$$\frac{dP}{dp_{\mathrm{T}}} = \left(\frac{dP_{\mathrm{MPI}}}{dp_{\mathrm{T}}} + \sum \frac{dP_{\mathrm{ISR}}}{dp_{\mathrm{T}}} \right) \cdot e^{\left(- \int_{p_{\mathrm{T}}}^{p_{\mathrm{T}_{\mathrm{i-1}}}} \left(\frac{dP_{\mathrm{MPI}}}{dp'_{\mathrm{T}}} + \sum \frac{dP_{\mathrm{ISR}}}{dp'_{\mathrm{T}}} \right) dp'_{\mathrm{T}} \right)} \tag{2.17}$$

where $p_{\mathrm{T_i}}$ is the transverse momentum of the ith interaction or parton branching (for a pictorial view, see Figure 2.8(left)). Note the similarity of this equation with Equation 2.10; it differs from that, only because initial-state radiation is also included in the formalism[1]. The FSR is included separately to the outgoing partons. While for the hard scattering all the processes can be generated, the additional interactions are limited only to QCD events, both initiated by quarks or gluons.

To take into account the possible different centralities of the collision, a dependence on the varying impact parameter is included. A small value of b corresponds to a large overlap between the colliding hadrons and consequently to a large probability of MPI, while large values for b refer to a smaller probability for parton-parton interactions. A double gaussian function is used for the overlap distribution:

$$\rho(r) \propto \frac{1-\beta}{a_1^3} e^{\left\{ -\frac{r^2}{a_1^2} \right\}} + \frac{\beta}{a_2^3} e^{\left\{ -\frac{r^2}{a_2^2} \right\}} \tag{2.18}$$

This corresponds to a distribution with a small core region, of radius a_2 and containing a fraction β of the total hadronic matter, embedded in a larger hadron of radius a_1. The choice of a double Gaussian function is motivated by comparisons with data of MB and UE hadronic activity [115]. Colour reconnection has been also included: this implements the possibility for colour strings coming from different interactions to be connected and exchange colour information; this idea is motivated by the fact that MPI tend to create lots of colour strings that most probably overlap in physical space and would be naturally connected.

These general features are implemented in the Pythia6 MPI model and very similarly in the Sherpa one. In Pythia8, a systematic improvement has been carried out by adding more sophisticated mechanisms, to get a more realistic picture of the MPI dynamics:

- the colour reconnection is performed by giving each system a probability to reconnect with a harder system defined as:

  $$P = \frac{p^2_{\mathrm{T_{Rec}}}}{(p^2_{\mathrm{T_{Rec}}} + p^2_{\mathrm{T}})} \text{ with } p^2_{\mathrm{T_{Rec}}} = R \cdot p_{\mathrm{T_0}} \tag{2.19}$$

  where R is a parameter and $p_{\mathrm{T_0}}$ is the same quantity defined in Equation 2.12;

- the rescattering [123]: this concept allows events where an outgoing state from one scattering becomes the initial state of another scattering, as represented in Figure 2.8(right);

---

[1] In particular, if the hard scattering occurs at a scale $p_{\mathrm{T_1}}$, either a secondary hard interaction or a hard emission from an initial parton take place at the scale $p_{\mathrm{T_2}} < p_{\mathrm{T_1}}$.

- a richer range of generated MPI processes (photon, J/ ψ, Drell-Yan production..);
- a wider set of overlap matter distribution functions, ranging from gaussian and double gaussian to an exponential low-x dependent function;
- a new interleaved evolution which includes also FSR together with ISR and MPI [124], based on the equation:

$$\frac{dP}{dp_{\mathrm{T}}} = \left(\frac{dP_{\mathrm{MPI}}}{dp_{\mathrm{T}}} + \sum \frac{dP_{\mathrm{ISR}}}{dp_{\mathrm{T}}} + \sum \frac{dP_{\mathrm{FSR}}}{dp_{\mathrm{T}}}\right) \cdot e^{\left(-\int_{p_{\mathrm{T}}}^{p_{\mathrm{T}_{\text{i-1}}}} \left(\frac{dP_{\mathrm{MPI}}}{dp'_{\mathrm{T}}} + \sum \frac{dP_{\mathrm{ISR}}}{dp'_{\mathrm{T}}} + \sum \frac{dP_{\mathrm{FSR}}}{dp'_{\mathrm{T}}}\right) dp'_{\mathrm{T}}\right)} \quad (2.20)$$

Furthermore, Pythia8 adds the possibility to generate manually two hard scatterings in the same event. Thanks to all these features, the MPI model in Pythia8, is believed to be the most complete. Hence, the Pythia8 event generator has been chosen for the preliminary sensitivity studies for the analyses presented in Chapter 4.

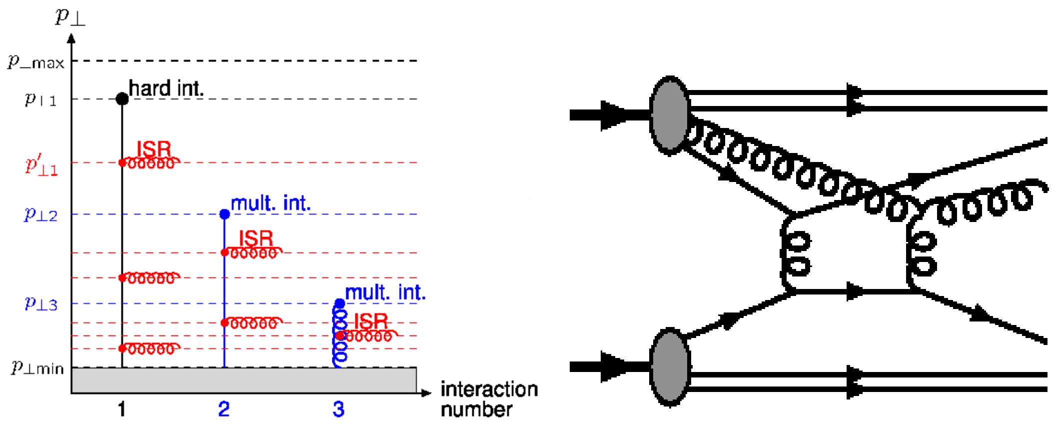


Figure 2.8: (Left) Sketch of the interleaving evolution in Pythia6 [125]: the highest scale of the collision is set by the hard scattering. Then, hard emission and MPI are generated at successively decreasing scales. The formalism in Pythia8 is the same but it additionally considers also the FSR in the evolution. (Right) Example of a rescattering diagram [123]: a parton of one proton interacts with a parton of the other proton; from this interaction, two new partons are produced and one of which interacts again successively with another parton of the other proton.

2.5.1.2 MPI in Herwig++

Also Herwig++ includes MPI [126] in order to simulate the UE. It is assumed that at fixed impact parameter b, individual scatterings are independent (only the momentum conservation is required, but no further constraints) and that the distribution of partons in the hadrons, factorizes with respect to the b and the x dependence. In this framework, the average number of hard interactions is given for hadronic collisions by:

$$n(b, s) = A(b, \mu^2)\sigma^{\mathrm{inc}}(s, p_{\mathrm{T}}^{\mathrm{min}}) \quad (2.21)$$

The function $A(b, \mu^2)$ describes the spatial overlap of the two colliding hadrons and the parameter μ is interpreted as the inverse radius of the proton; $p_{\mathrm{T_{min}}}$ sets the lower scale for allowed hard MPI. Below this threshold, soft non-perturbative MPI are also generated within the same model but they make use of a parametrization from the optical theorem in order to restore unitarity.

As in Pythia, this simple model is enriched by the introduction of colour reconnection that may cause non-trivial changes to the colour topology of the final state. In particular, parton jets emerging from different interactions are colour-connected if they are located closely in phase space. Inside the cluster hadronization model, the distance between two partons, namely the length of the colour connection, is related to the invariant mass of the cluster:

$$\lambda = \sum_{\mathrm{i}=1}^{\mathrm{N_i}} m_{\mathrm{i}}^2 \tag{2.22}$$

where the sum is performed over all the final clusters. The goal of CR is to reduce λ for every cluster. This mechanism is carried out in Herwig++, by iterating over all possible cluster pairs and by finding the best association among the partons in the clusters, according to the minimum colour length. If the invariant mass of a new cluster is lower than the one of the previous cluster, a colour swap is applied with a certain probability. This model has been shown to improve the description of several observables, like the charged particle multiplicity [127].

Chapter 3

The CERN Large Hadron Collider and the Compact Muon Solenoid experiment

In this chapter the Large Hadron Collider (LHC) and the Compact Muon Solenoid (CMS) experiment are described; in particular, the parts of the detector relevant for the analyses of this thesis are treated in detail.

3.1 The Large Hadron Collider

The LHC [128], operating at CERN[1] near Geneva in Switzerland, is a double superconducting storage ring operated in collider mode. It is installed in the 27 km tunnel which formerly hosted the Large Electron-Positron (LEP) collider [129].
The elements which define a collider may be summarized as follows [130]:

- the accelerated and colliding particles;
- the center of mass energy of the collision;
- the luminosity;
- the type of collider.

Depending on the physics that is addressed, hadrons (proton-proton or proton-antiproton) or leptons (electron-positron) can be chosen to be accelerated and brought to collision. Also lepton-hadron (electron-proton or positron-proton) colliders have been used, like at the HERA machine (DESY, Hamburg) [131]. The only requirements for the particles in a collider are that they need to be charged and stable (or with a long enough lifetime), in order not to decay during the acceleration process.

The particles are then accelerated up to the collision energy through radiofrequencies [130]. By increasing the center-of-mass energy, heavier objects can be produced. The first colliders had a

---

[1] from French, Conseil Européen pour la Recherche Nucléaire

center-of-mass energy of a few MeV, while the TeV scale is reached with the LHC.

The luminosity is the parameter which expresses the number of particles in the beam at the collision point, per unit of transverse area and time; it is an indicator of the flux of the accelerated particles, defined as:

$$L = \frac{N_b^2 n_b f_{rev} \gamma_r}{4\pi \epsilon_n \beta^*} F \tag{3.1}$$

where $f_{\mathrm{rev}} = \mathrm{c/r}$ is the revolution frequency, equal to 11.25 kHz for the LHC and $\gamma_{\mathrm{r}} = 1/\sqrt{1 - v^2/c^2}$ is the relativistic gamma factor. With ϵ_{n}, the normalized transverse beam emittance is indicated, while β^* represents the focus of the beam and F is the geometric reduction factor due to the beam crossing angle at the interaction point. The values n_{b} and N_{b} refer to the specifics of the proton beam: protons are bundled in small bunches which circulate along the acceleration ring until they reach the nominal energy before the collision; n_{b} is the number of bunches, while N_{b} is the number of protons in each bunch.

With the type of a collider, its shape is meant, namely linear or circular. A circular collider gives the opportunity to the particles to turn several times through the ring before reaching the final energy; it has to face, though, the challenge of bending the particles, in order to maintain them inside the beam pipe. This is achieved thanks to the magnetic field; the higher the energy of the particles, the larger the magnetic field has to be for a fixed radius of the orbit.

The LHC has been designed to produce *pp* collisions at a center-of-mass energy of 14 TeV with a peak instantanous luminosity $L = 10^{34}$ cm^{-2} s^{-1}, a nominal bunch spacing $T_{\mathrm{b}} = 25$ ns and a number of bunches equal to 2808, 7.5 metres distant from each other. This brings a fraction f of bunches in the ring of the order of 2808 × 7.5 m / 27 km = 0.78. The expected collision rate is 40 MHz. These parameters have not been pushed to the design values in the first three years of operation but they have been kept lower for a safe operation and in order to avoid magnet quenches, as the one happened in September 2008 [132]. Table 3.1 lists the accelerator parameters, comparing the designed and the effective ones, updated with the values in June 2012.

Parameters	Units	Nominal Value	Effective Value
number of bunches	n_{b}	2808	1374
bunch spacing	T_{b} [ns]	25	50
protons per bunch	N_{b} [$\times 10^{11}$]	1.15	1.5
norm.tr.emittance	ϵ_{n} [μm]	3.75	2.6
r.m.s. bunch length	σ_{S}[cm]	7.55	> 9
β at IP	β^* [m]	0.55	0.6
full crossing angle	θ_{c} [μrad]	285	290
luminosity lifetime	τ_{L} [h]	15	20
peak luminosity	L_{peak} [$\times 10^{34}$ cm^{-2}s^{-1}]	1.0	~ 0.68
events/crossing	-	25	35

Table 3.1: List of the main parameters of accelerator machine and proton beams at the LHC. Design values and effective ones, updated for June 2012, are compared. Units for each parameter are also indicated.

The instantanous luminosity L is proportional to the rate dN/dt of a certain process of cross section σ:

$$L = \frac{dN}{dt} \cdot \frac{1}{\sigma} \tag{3.2}$$

The integrated luminosity is the integral of the instantanous luminosity over time. It is usually expressed in units of barn^{-1} to give a direct indication of the number of produced events for a process. For instance, an integrated luminosity of 30 pb^{-1} means that 30 events of a process with cross section equal to 1 pb are produced.

The instantanous luminosity is also connected to the number of overlapping events in the same collision, namely the pile-up events. Since the colliding bunches are composed by a very high number of protons, it is likely that more than one pp interaction occurs within the same collision. This relation can be exploited to have a rough idea of the number of pile-up events for a given instantanous luminosity:

$$N_{\mathrm{PU}} = L \cdot \sigma_{\mathrm{T}} \frac{T_{\mathrm{b}}}{f} \tag{3.3}$$

where T_{b} is the proton bunch spacing (see Table 3.1) and σ_{T} is the total pp cross section for the considered energy, equal to approximately 70 mb for a collision energy of 7 TeV. In 2010 data taking, N_{PU} was around 2-3, while in 2011 and 2012, it increased up to 20-25.

The proton beams colliding in the LHC are accelerated and injected in stages [133]. Each proton beam is produced from gaseous hydrogen which is injected into a cavity where a strong electric field breaks them up into protons and electrons. Protons are collected and accelerated up to 100 keV through a radiofrequency quadrupole which provides the first focusing and a further energy kick up to 750 keV. This beam is sent to a linear accelerator, Linac 2, that provides a 50 MeV beam. Then, the first transfer line drives the beam to the Proton Synchrotron (PS) booster, a small four-ring synchrotron with 25 m radius. The PS booster is used to raise the proton energy up to 1.4 GeV, then the obtained beam is accelerated again up to 25 GeV by the PS, a 100 m radius synchrotron. The final bunch structure of LHC beams is thus achieved: 81 bunches with 50 ns time spacing are extracted at the end of the cycle. Triplets of 81 bunches are transferred to the Super Proton Synchrotron (SPS), a 1 km radius proton synchrotron, where protons are accelerated up to an energy of 450 GeV and injected into the LHC, where they reach the final energy. The whole injection chain is shown in Figure 3.1. Accelerating protons up to 7 TeV in a circumference of $\approx$ 27 km radius requires a bending magnetic field of 8.4 Tesla, furnished by 1232 superconductive dipoles working at 1.9 K. A photo of the tunnel, taken inside the underground LHC cavern, is shown in Figure 3.2.

Using four collision points along the ring, six experiments are installed along the accelerator:

- ALICE (A Large Ion Collider Experiment) [136], which is a dedicated heavy-ion detector aiming for studying strongly interacting matter at very high energy densities, where the formation of a new phase of matter, the quark-gluon plasma, is expected. The detector itself is composed by a tracking system, a time projection chamber (TPC) and a transition radiation detector; they ensure high detection efficiency in a high multiplicity environment;
- ATLAS (A Toroidal LHC ApparatuS) [137], that is a multipurpose detector, aiming for searches of New Physics and precise measurements in the SM, primarily in the Higgs sector.

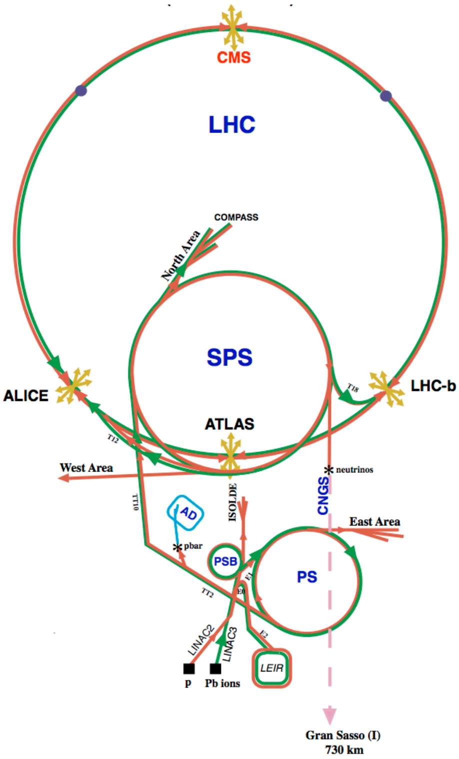


Figure 3.1: Sketch of the whole injection chain which feeds the LHC [134]: the different smaller accelerators are shown from the production point of the accelerated protons up to the access in the main ring. The position of the several experiments which use the beams is also shown in the picture.

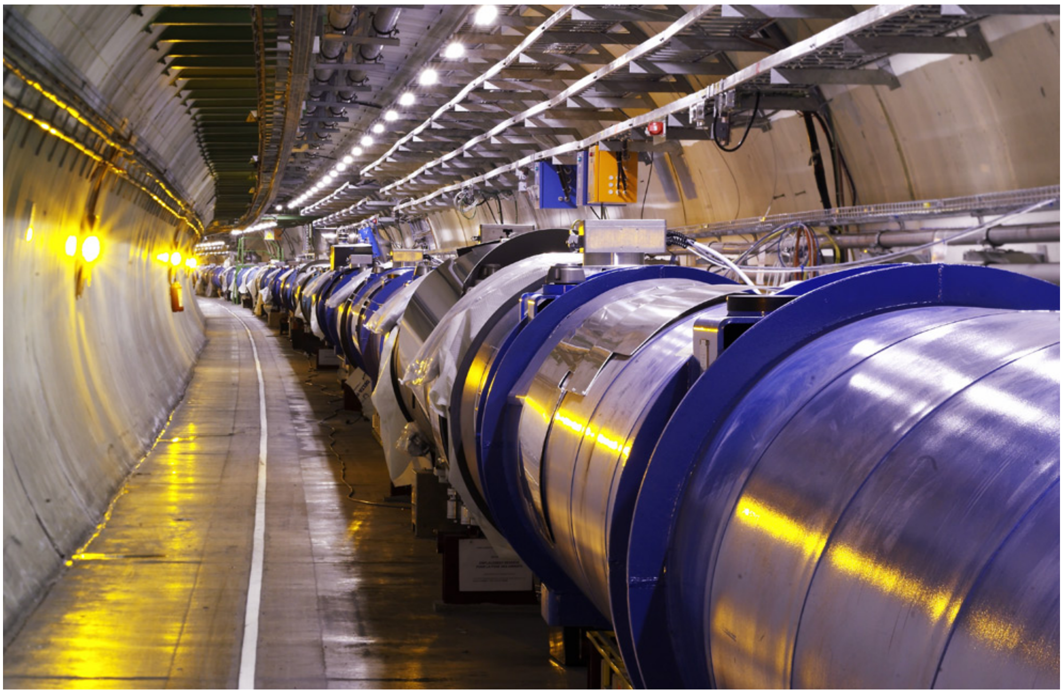


Figure 3.2: Real photo of the the superconducting magnets surrounding the beam pipe which hosts the LHC colliding protons [135]. The picture was taken before the start of the accelerating operations.

Its structure comprises a tracking and a calorimetry system, immersed in a toroidal magnetic field of 2 T intensity, and an external muon detector. It is the biggest experiment at the LHC, with a length of 44 m and a diameter of 25 m;

- CMS (Compact Muon Solenoid) [138], which is the other multipurpose detector at the LHC, with smaller dimensions and different technologies with respect to ATLAS; a detailed description is provided in Section 3.2;

- LHCb (Large Hadron Collider beauty) [139], that is specialized on measurements in the heavy flavour sector, particularly focusing on rare decays of charm and bottom hadrons and the parameters of the CP violation, in searches for New Physics and insights on the primordial matter-antimatter asymmetry; it is a single-arm spectrometer detecting particles going into the forward direction[1] with a very precise tracking system, provided by a magnetic field of 4 T for the measurement of the momentum of the charged particles, and an electromagnetic and hadronic calorimetric structure;

- LHCf (Large Hadron Collider forward) [140], which addresses the problem of understanding the development of atmospheric showers induced by very high energy cosmic rays, by measuring the neutral-particle production cross section; it is the smallest experiment at the LHC, installed near the ATLAS experiment region very close to the beam pipe, consisting of two very forward imaging calorimeters;

- TOTEM (ToTal Elastic and diffractive cross section Measurement) [141], which, together with a wide program in diffractive physics, wants to measure the total pp cross section through the optical theorem, by measuring elastic scatterings; built in the very forward region close to the CMS detector, it is made of Roman Pots and two telescopes (a Cathode Strip Chamber and a Gas Electron Multiplier) for the measurement of the momentum of the charged particles.

3.2 The Compact Muon Solenoid detector

The CMS experiment is a general purpose detector housed at interaction point 5 of the LHC. It is designed with a 4 T solenoidal magnetic field provided by the largest superconducting solenoid ever built. The detector has a cylindrical symmetry, where the center coincides with the proton collision point, with a big barrel region covering the central part and two endcaps that close the structure on both sides. CMS is symmetric in the radial direction around the beam pipe and also symmetric along the beam pipe from the center of the detector. The choice of the magnetic field configuration [142] for the measurement of the momentum of the charged particles drove the whole detector layout. The solenoid that produces the magnetic field is 13 metres long with a 5.9 metres inner diameter. Inside this, the inner tracker and the calorimeters are located, while outside four muon stations in the barrel part and four in the endcap part consisting of several layers each, are installed. The overall dimensions of the CMS detector are a length of 21.6 m, a diameter of 14.6 m and a total weight of 12500 tons. The structure of CMS is shown in Figure 3.3.

---

[1] The symmetric backward detector has not been built because of space and economic reasons.

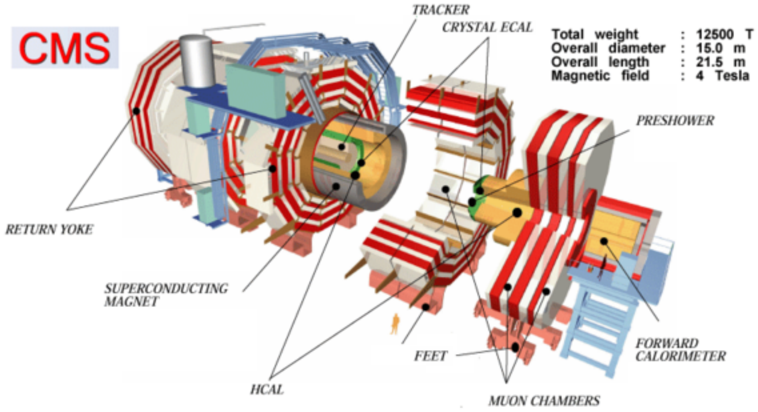


Figure 3.3: Sketch of the CMS detector at the LHC [143] in an "unmounted" configuration, in order to better show its different parts. All the subdetectors are displayed and the dimensions of the whole detector are also indicated.

3.2.1 Definition of the experimental coordinate system

The CMS coordinate system is oriented such that the x-axis points towards south with respect to the center of the LHC ring, the y-axis points vertically upward and the z-axis towards the direction of the beam to the west. The azimuthal angle ϕ is measured from the x-axis in the xy plane and the radial coordinate in this plane is denoted by r. The polar angle θ is defined in the rz plane with respect to the beam pipe. It is sometimes preferred to use a quantity, called pseudorapidity, because, for massless particles, differences between pseudorapidities are Lorentz-invariant under boost along the z-direction. It is defined as:

$$\eta = -\ln \tan(\theta/2) \tag{3.4}$$

The relation between θ and η is illustrated in Figure 3.4, which links together values of the two quantities in the rz plane.

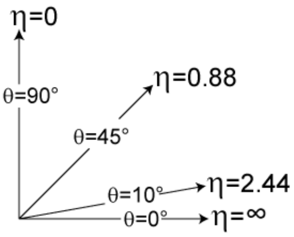


Figure 3.4: Corresponding values of η and θ in the longitudinal plane [144]: the horizontal axis is the z-axis, while the vertical one is any direction in the xy plane.

The particle production can also be assumed as constant per unit of pseudorapidity. The momentum transverse to the beam direction, denoted by p_{T}, is computed from the x- and y-components, while the transverse energy is defined as $E_{\mathrm{T}} = E{\cdot}\sin\theta$. In hadron colliders, the transverse quantities become important: in fact, in the transverse plane, the sum of all momenta and energies should be equal to 0, assuming that the incoming protons have no transverse component at the moment of the interaction. In the described analyses, the p_{T}, ϕ and η quantities will be considered to identify and select the physics objects. The physics objects, selected approximately with $|\eta| < 2.5$ are referred to as “central”, while the ones in $|\eta| > 2.5$ are called “forward”.

3.3 Data taking and luminosity in phase I of data taking

The integrated luminosity in CMS is based on signals measured in the forward part of the hadronic calorimeter. Two methods for extracting a real-time relative instantaneous luminosity are used. A “zero counting” method in which the average fraction of empty towers is used to infer the mean number of interactions per bunch crossing. A second method uses the linear relation between the average transverse energy per tower and the luminosity. The different algorithms agree between each other at the 5% level. In the different years of data taking, a total integrated luminosity of about 29.4 fb^{-1} has been recorded by CMS. The increase of integrated luminosity and of the peak luminosity in the different periods are shown in Figure 3.5. Note the rapid increase of both quantities from 2010 to the following years.

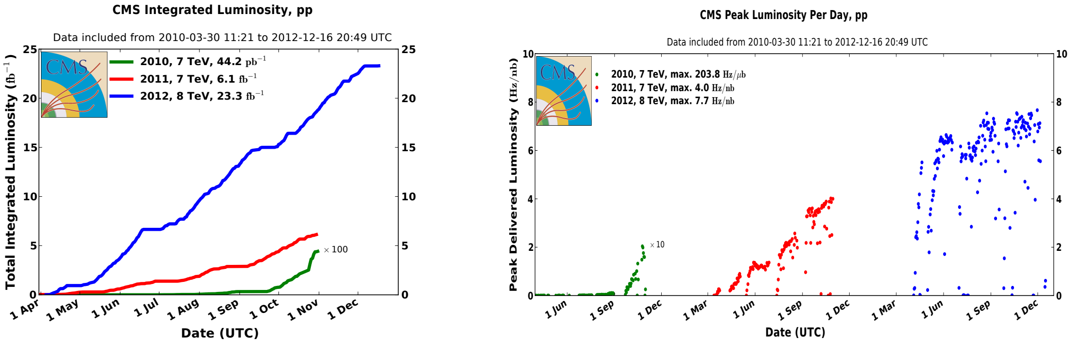


Figure 3.5: Summary of the integrated (left) and daily peak (right) luminosity for the three different years of the data taking at the LHC Phase I [145]. Note the big increase in peak luminosity from 2010 to 2012.

In the current year (2014) and successively in the coming ones, an upgrade program has been planned for both LHC and CMS. The schedule is shown in Figure 3.6. The overall goal of the upgrade is to increase the integrated luminosity accumulated by the experiments, rising up the instantanous luminosity and the energy, and maintaining the ability to collect good quality data. Higher luminosity, together with greater energy, extends the discovery reach of the experiments. The integrated luminosity, which is planned to be delivered to the experiments, is $\sim$200 fb^{-1} until 2018 and $\sim$500 fb^{-1} at the end of 2022 [146].

For LHC, the upgrade [147] will bring a reduction of the colliding bunch spacing of the protons down to 25 ns, equal to the design value, and an energy of 13 TeV to be increased later on to 14 TeV. This is referred to as “Phase I upgrade”. These goals will be reached after the shutdown in 2013-2014. In 2017, LHC will very likely bring an increase of two times the design luminosity and even bigger after the long shutdown in 2018 for the successive upgrade phase.

The CMS experiment is dealing with these new settings of the accelerator by replacing several parts of the detector [148]. The increased number of overlapping events in the same bunch crossing, leading to a higher flux of particles hitting the detectors, and the problem of fastness of the response will be the main challenges for the upgrade.

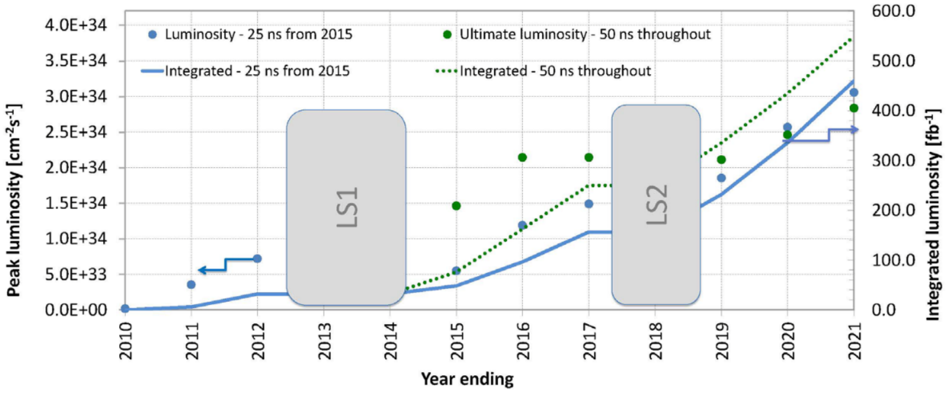


Figure 3.6: Upgrade phases of the LHC accelerator as shown by F.Zimmermann in [146]. Two scenarios are considered: one with bunch spacing of 25 ns and lower peak luminosity and one with bunch spacing equal to 50 ns and higher peak luminosity. The first one is the most likely to happen.

3.4 The tracking system

The core of CMS is a tracker, designed to provide a precise and efficient measurement of the trajectories of charged particles emerging from LHC collisions, and a reconstruction of secondary vertices. The tracker [149], with its 2.5 m diameter and 5.8 m length, surrounds the interaction region and is fully immersed in the magnetic field. In order to obtain reliable trajectory identification and attribution to the correct bunch crossing, the CMS tracker needs high granularity and fast response because of the large number of tracks to be processed at the full LHC luminosity with several overlapping interactions per bunch crossing. It is composed of both silicon pixel and strip detectors, as shown in Figure 3.7, with an active surface of about 200 m^2 and it assures efficiencies higher than 95% on the reconstruction of the trajectories of particles with $p_{\mathrm{T}} > 1$ GeV in the angle coverage range of $|\eta| < 2.5$. Heavy quark flavours can be produced in many of the interesting channels and so, a precise measurement of secondary vertices is also needed. Moreover, the tracker completes the functionalities of the electromagnetic calorimeter and of the muon system to identify electrons and muons, and it is crucial for a good jet reconstruction. Finally, tracker information is used in the High Level Trigger to help in reducing the event collection rate to an acceptable amount of storage.

The pixel detector is the most inner part of the detector and consists of 1440 pixel modules arranged in three barrel layers (BPIX) at radii of 4.4, 7.3 and 10.2 cm of length of 53 cm, and four disks (FPIX), two at each side of the barrel, at 34.5 and 46.5 cm from the interaction point. Each pixel occupies a surface of 0.100×0.150 mm^2. The tracker design includes about 66 million pixels (48 in the barrel and 18 in the endcap), covering a total area of about 1 m^2 and measuring at least three high precision points on each charged particle trajectory in $|\eta| < 2.5$. The resolution measured for a single pixel module is about 13 μm along the x direction, and about 30 μm along the y direction [150]. The resolution achieved along the longitudinal direction is around 20 μm.

The silicon strip tracker covers an intermediate radial region ($20 < r < 116$ cm) where the particle flux is lower than in the pixel region. This enables the use of detectors with a bigger area: each strip has a cell size of 10 cm $\times$ 80 μm for the inner part and of 25 cm $\times$ 180 μm for the outer part. The total area of the strip tracker is about 198 m^2, read out by 9.3 million channels.

The barrel tracker is divided in two parts: a Tracker Inner Barrel (TIB) and a Tracker Outer Barrel (TOB); the endcap tracker has equally two components: a Tracker Inner Disk (TID) and a Tracker End Cap (TEC). In the barrel part, the strips are aligned along the z-axis, while they are perpendicular to the beam pipe in the endcap. The complementary dimension, respectively z and r for barrel and endcap, is measured by using two layers for each strip, tilted of 100 milliradians with respect to each other, in order to have a three-dimensional configuration. The TIB tracking system delivers up to 4 $r\phi$ measurements on trajectories of charged particles and achieves a single-point resolution of between 23-34 μm in the ϕ direction and 230 μm in z. The TOB surrounds the TIB/TID and provides up to 6 $r\phi$ and the resolution is 53 μm in the four innermost layers, and 35 μm in the outermost two ones, depending on the strip pitch. Finally, the TEC encloses the previous subdetectors and its strips are radially oriented. This design ensures at least 9 hits in the silicon strip tracker in the full acceptance region and at least four of them are two-dimensional.

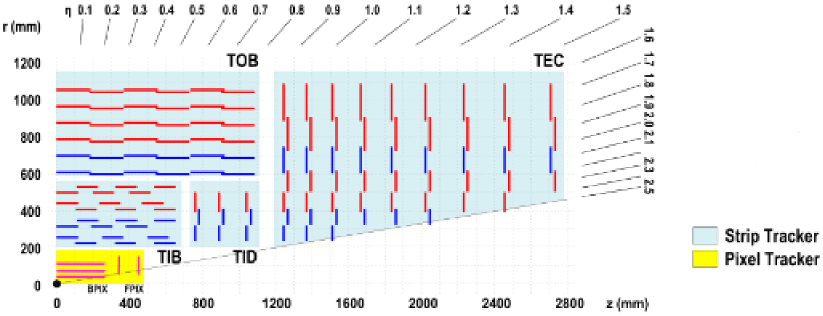


Figure 3.7: Sketch of the tracking system in CMS [151]: the different pixel and strip modules are shown in the longitudinal plane for the barrel and endcap regions.

The whole structure of the CMS tracking system is pictured in Figure 3.7 which shows both pixel and strip subdetectors. The performance of the whole tracking system has been studied for charged particles at different transverse momenta: the results are described in [152] for muons and pions simulated with particle guns. In Figure 3.8, p_{T} and spatial resolution is shown for muons of different transverse momenta as a function of η, when only the tracking system is used. The relative p_{T} resolution increases for higher energies and increasing pseudorapidities. In the most central region, it is of the order of 0.5-1% for p_{T} up to 10 GeV and around 1-2% for p_{T} equal to 100 GeV. When going to the edge of the tracking system ($\eta \sim 2$) the resolution increases respectively up to 2% and 7-8%, respectively for p_{T} equal to 10 and 100 GeV. Resolutions of impact parameter and longitudinal coordinate are also shown. The impact parameter is measured by the tracker with a resolution of around 100, 20 and 10 μm for muons with p_{T} of 1, 10 and 100 GeV. While at high transverse momenta, the resolution does not change in the whole acceptance, it increases of a factor of 2 for lower p_{T} when going from $\eta \sim 0$ to 2–2.5. The longitudinal resolution is more η dependent: at high p_{T}, it ranges between 20 and 40 μm while for muons with p_{T} equal to 1 GeV, it goes from 100 μm in the central part of the tracker up to values of 1 mm for $\eta \sim$ 2–2.5.

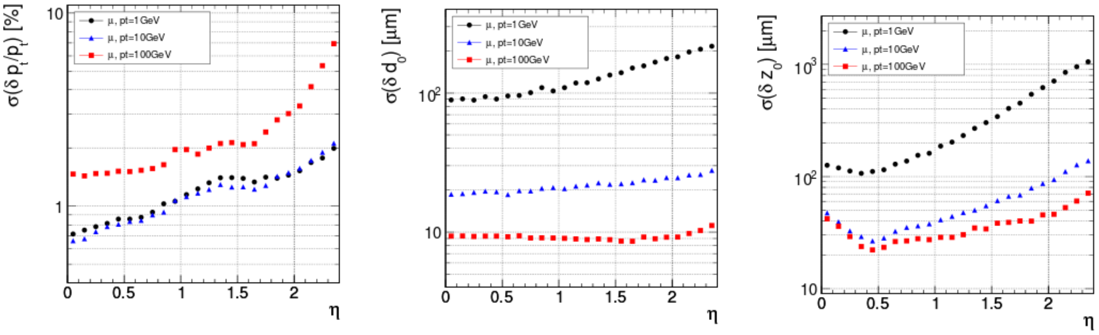


Figure 3.8: Resolution measured in simulated events for muons at energies of 1, 10, 100 GeV, as a function of the particle pseudorapidity η [138]: resolution in p_T (left), in impact parameter d_0 in xy plane (center) and in longitudinal position z_0 (right) are shown for reconstruction performed with the whole tracking system. Events are simulated through particle guns.

3.5 The calorimetry system

The identification of electron, photons and hadrons relies on an accurate calorimetry system in CMS. It consists of a destructive measurement of the energy of a particle. A distinction is made between electromagnetic and hadronic calorimetry.

Electromagnetic calorimetry is based on the production of a shower initiated by a photon or an electron inside an absorber; in particular, the photons tend to be converted in electron-positron pairs via pair production, while electron and positrons undergo a bremstrahlung process with the emission of a photon. These successive conversions continue until the photons fall below the pair production threshold and other energy losses of electrons start to dominate. In the simplest model [130], the energy of the initial particle is shared in equal parts between the particles of the final state and the average amount of matter that the particle needs to traverse in order to produce the forementioned emissions, is called “radiation length” X_0. The radiation length is also defined as the average distance needed for an electron to reduce its initial energy by a factor 1/e. An electromagnetic shower also develops in the transverse plane. Its transverse shape can be approximated by a cylinder whose width is quantified by the “Moliere radius”: it represents the radius of the cylinder containing an average percentage of 90% of the total energy deposition of the shower, and is specific for any material.

The hadron shower, instead, is produced by inelastic scatterings of hadrons in the detector material. Incident hadrons release energy through nuclear excitation and hadron production, producing many other additional particles. In the simplest model [130], it can be assumed that all the final particles are pions, in form of charged ones, $\pi^{\pm}$, for two thirds and neutral ones, π^0, for the remainder. The latter ones decay further in two photons, whose signal constitutes the electromagnetic part of the hadronic shower. The quantity, corresponding to the radiation length for hadron showers, is the interaction length, λ_t, defined similarly as the distance which a hadron needs to traverse in a material before losing a factor $1/e$ of its energy. Generally, λ_t is bigger than X_0 for a given material. Also the space development of the two shower types is different: a hadronic one tends to be wider and longer, while the electromagnetic one has generally a more compact shape.

In the CMS detector, two different detectors carry out the detection of the two types of showers: the electromagnetic calorimeter, described in Section 3.5.1, and the hadronic calorimeter, described in Section 3.5.2. The former one is placed in front of the latter, because, as previously mentioned, the interaction length is greater than the radiation length; thus, the hadrons passing through the electromagnetic calorimeter lose only a small fraction of their energy ($\sim 20\%$) in it, and they are then fully measured in the successive hadronic calorimeter.

3.5.1 The electromagnetic calorimeter

The CMS electromagnetic calorimeter [153] is a homogeneous calorimeter composed by 61200 lead tungstate ($PbWO_4$) crystals in the barrel region and 7324 ones in the endcaps. This material was chosen because of its high density (8.28 g/cm^3), its short radiation length (0.89 cm) and small Moliere radius (2.2 cm), its fast response time and good radiation tolerance. Signal from the scintillation light, produced by electrons and positrons of the shower, is transmitted through total internal reflection and is detected by avalanche photodiodes in the barrel region and vacuum phototriodes in the endcaps. The barrel section has an inner radius of 129 cm and its structure is organized with 36° "supermodules", each of them covering a $|\eta| < 1.479$ region; a supermodule is a collection of four modules, equipped with five pairs of crystals each. Every crystal covers 0.0174 in both $\Delta\phi$ and $\Delta\eta$ (corresponding to 1° in θ) angular region and has a length of 230 mm corresponding to 25.8 X_0. The endcaps are at a distance of 314 cm from the interaction point and close the barrel part on both sides; they cover a pseudorapidity range of $1.479 < |\eta| < 3.0$ and are contained inside two semi-circular aluminium plates with basic units of 5×5 crystals. The endcaps are also equipped with a preshower sampling calorimeter in front of the whole system, composed of lead radiators and silicon strip detectors in order to identify neutral pions in the forward region and to have a better determination of the position for electrons and photons. An overview of the ECAL subdetector is sketched in Figure 3.9.

The energy resolution measured during calibration [154] is parametrized by the function:

$$\left(\frac{\sigma_E}{E}\right)^2 = \left(\frac{2.8\%}{\sqrt{E}}\right)^2 + \left(\frac{0.12\%}{E}\right)^2 + (0.30\%)^2 \tag{3.5}$$

The first term refers to the stochastic contribution due to fluctuations in the lateral shower development and in the energy released in the preshower; the second term quantifies the noise, due to electronics, digitization and pile-up, and finally the third term is a constant due to calibration errors, energy leakage or non-uniformity in the light collection.

3.5.2 The hadronic calorimeter

The CMS hadronic calorimeter [155] (HCAL) is a sampling calorimeter, relevant for measuring the hadronic jet energy and for providing information used for photon and lepton identification; its structure is not totally contained inside the magnet coil because of the small space left empty between the solenoid and the ECAL. The hadron calorimetry system is, therefore, organized in four parts: an inner hadron barrel (HB), an outer detector (HO), an endcap part (HE) and a forward calorimeter (HF).

The hadron barrel part consists of 36 wedges covering the pseudorapidity region $|\eta| < 1.3$, segmented into four azimuthal sectors each, and made out of 14 flat brass absorber layers, enclosed between two steel plates. An additional segmentation in pseudorapidity thanks to plastic scintillators provides an overall division in $\Delta\phi \times \Delta\eta = 0.087 \times 0.087$ angular regions. Due to the limited space between the ECAL and the solenoid, the hadronic interaction lengths λ_t ranges from only $5.82\lambda_t$ at the center ($\eta \sim 0$) to $10.6\lambda_t$ at the edges ($|\eta| \sim 1.3$). However, hadrons traversing HB have already passed ECAL which provides an additional $1.1\lambda_t$ of material.

The hadron outer detector contains scintillators with the same angular segmentation and lies outside the solenoid. The solenoid is used as absorber and the thickness of the scintillators depend on the angle, resulting in $1.4\lambda_t/\theta$. This is achieved by adding one layer of scintillator in the more forward part and two layers of scintillators in the more central part of the calorimeter. The HO covers the region $|\eta| < 1.26$ and works as a tail catcher, sampling the energy from penetrating hadron showers leaking through the back part of the barrel calorimeter; the HO information serves to improve the energy resolution, by increasing the total thickness of the calorimeter to $11.8\lambda_t$.

The hadron endcaps consist of 18 towers in η with segmentation in ϕ of about 5° for the lower pseudorapidities inside the range $1.3 < |\eta| < 1.6$ and of 10° for the higher ones, inside $1.6 < |\eta| < 3.0$. The HE includes 18 layers made of alternating 79 mm brass plate and 9 mm scintillator. A drawing of HCAL is shown in Figure 3.9, for both barrel and endcap parts.

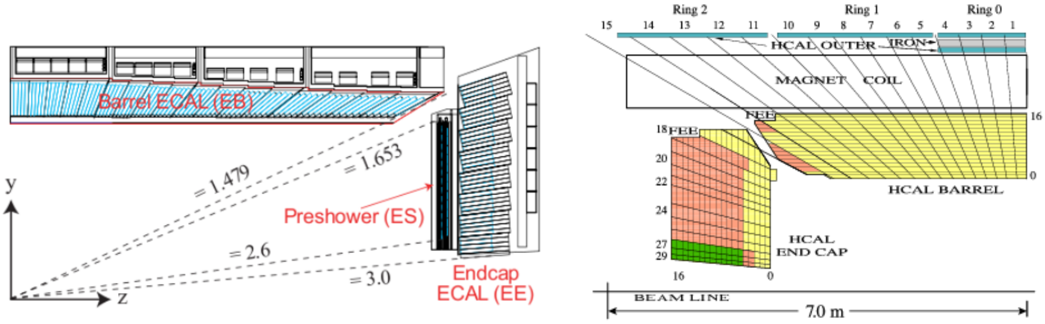


Figure 3.9: Sketch of ECAL (left) [138] and HCAL (right) [156] in barrel and endcap regions of the CMS detector: they are represented in the longitudinal plane.

The performance in terms of timing and synchronization is described in [157]. The energy resolution, determined by test beam data [156], is parametrized for single pions by the function:

$$\frac{\sigma_E}{E} = \frac{84.7\%}{\sqrt{E}} + 7.4\% \tag{3.6}$$

where the first term includes the effects of leakage and sampling fluctuations, while inhomogeneities and shower leakages contribute to the second one. The jet energy resolution is dominated by fluctuations inherent to the jet physics, while the instrumental effects are less relevant. However, the response and resolution of the CMS calorimetry system depends on both ECAL and HCAL, since most particles start showering in the ECAL. The ECAL and the HCAL fraction of the energy deposited in each calorimeter varies not linearly with energy and, as a result, the raw energy measurements require substantial corrections.

Finally, the hadron forward calorimeter assures a coverage up to $\eta = 5$ and, because of the high flux of particles in this region, is provided with a sandwich of different layers of steel as absorber and quartz fibre as active material; indeed, this design leads to narrower and shorter showers and allows to detect both electromagnetic and hadronic showers. The absorber-fiber layers are arranged in 900 towers that run parallel to the beam line, at a distance of 11.2 metres from the interaction point. The signal originates from Cherenkov light emitted in the quartz fibres, which is then channeled towards photomultipliers that produce the electric signal. An outline of the subdetector is shown in Figure 3.10. The performance of HF is described in [158], together with the whole calibration and compensation procedure. The energy resolution can be parametrized as:

$$\frac{\sigma_E}{E} = \frac{a}{\sqrt{E}} + b \tag{3.7}$$

where values of the two parameters are listed in [158] for different particles and different fit assumptions. In general, the coefficient a is around 200% for electromagnetic particles and 300% for hadrons, while b is around 10% for both types.

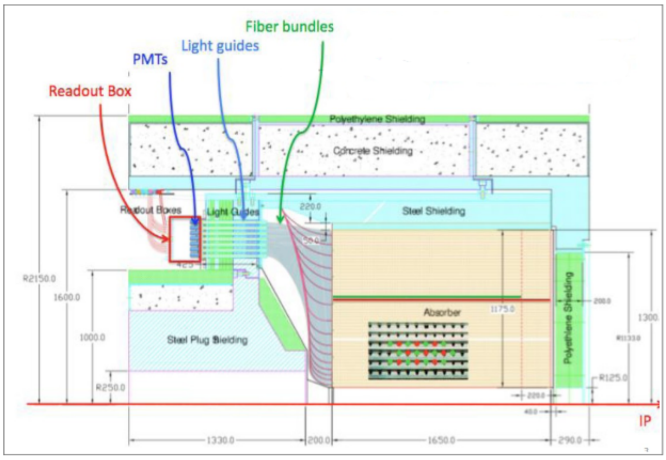


Figure 3.10: Sketch of the HF subdetector in the CMS detector [158]: the whole structure for particle detection and signal readout is pictured.

The calorimetry system is completed by CASTOR, an electromagnetic and hadronic calorimeter installed in the very forward region: in Appendix A, more details about the detector are given.

3.6 The magnet

A magnetic field is an important component for an experiment at colliders. By using the curvature of a charged particle in a magnetic field, a measurement of its transverse momentum is possible. The CMS magnet [142], which provides a magnetic field of 4 T, is a superconducting solenoid, 220 tons heavy and 3.9 radiation lengths thick. The field is closed by a 10000 tons iron return yoke made by five barrels and two endcaps of three layers each. The yoke is instrumented with four layers of muon stations and the coil is cooled down to 4.8 °K by a helium refrigeration plant; the whole structure is kept isolated by two pumping stations providing vacuum on the 40 m^3 of the cryostat volume. Such a strong magnetic field enables a very compact layout and an efficient muon detection.

3.7 The muon detectors

The CMS muon system [159] forms the outer part of the CMS layout; this is because the muons are able to travel through the whole solenoid with minimal energy loss inside the inner detectors. The muon system is composed of three types of gaseous detectors, located inside the empty volumes of the iron yoke and therefore arranged in barrel and endcap sections. In the barrel region where the muon flux is quite low, standard drift chambers with rectangular cells are used; they are arranged in four stations inside the return yoke and cover the region of $|\eta| < 1.2$. Since the muon and background flux is higher in the forward region, the choice for muon detectors fell upon cathode strip chambers (CSC) because of their fast response time, fine segmentation and radiation tolerance. Each endcap is equipped with four stations of CSCs that cover in total the region of $0.9 < |\eta| < 2.4$. They are arranged in concentric rings, three in the innermost stations and two in the last one. In total, the muon system contains about 25000 m^2 of active detection planes and nearly one million electronic channels.

For the muon reconstruction, the tracking system is used in addition to the muon detectors. The reconstruction performance has been measured in [160]: the identification efficiency for muons with a transverse momentum of more than a few GeV is greater than 95% in all detector regions, while the misidentification rate lies only between 0.1% and 1%, depending on the selection. For muons with p_{T} between 20 and 100 GeV, the relative p_{T} resolution is between 1.3% and 2% in the barrel and slightly bigger than 6% in the endcaps. Even for high-energetic muons with $p_{\mathrm{T}} > 1$ TeV, the resolution is still greater than 10%.

3.8 The trigger

The task of a trigger system is to select rare events inside a huge multiplicity of non interesting interactions, and to suppress background as efficiently as possible. High bunch crossing rates and high values of the luminosity at the LHC correspond to a total of 10^9 events per second to be recorded by CMS. This large amount of data is impossible to store and process in an inclusive and complete way with the current technology of data transferring and saving. Therefore, a dramatic rate reduction has to be achieved. Fortunately, interesting events are rare (with a frequency of about 1 Hz) and hence, it is possible with an efficient trigger system to retain as much signal as possible and reject background events. In case the condition of rarity of the examined process is not fulfilled, e.g. for Minimum Bias samples or events with jets at low p_{T}, a prescaling is applied: this procedure consists of storing only a fraction of events of the same type. The events that are effectively recorded are probabilistically chosen, e.g. the first event out of ten is recorded while the others are rejected. The decision of recording or dropping an event has to be performed very quickly and it is based on signals of certain physics objects inside the detector.

CMS achieves this condition in two steps: the Level 1 (L1) Trigger [161] and the High Level

Trigger (HLT) [162, 163]. The Level 1 trigger is based on custom and programmable electronics (FPGA, ASICs and LUTs), while the HLT is a software system implemented on a ~ 1000 processor farm. The overall trigger is designed to reduce the rate at least 106 times. The maximum allowed output rate for L1 is 100 kHz. It uses rough information from coarse segmentation of calorimeters and muon detectors and keeps data in a pipeline until the acceptance/rejection decision is made. HLT exploits the full amount of collected data for each bunch crossing accepted by L1 Trigger and is capable of complex calculations such as the offline ones. Configuration and operation of the trigger components are handled by a software system, called Trigger Supervisor. Currently, the transmission of data from the L1 to the HLT is handled with optical links. The size of each event is about 1 MB and the total rate of data to be passed to HLT is ~ 100 TB/s.

The L1 Trigger involves the calorimetry and muon systems[1], as well as some correlation of information between the two. The L1 decision is based on the presence of particle candidates such as photons, electrons, muons and jets above set E_{T} or p_{T} thresholds. It also employs sums of $E_{\mathrm{T}}^{\mathrm{miss}}$ and E_{T}. The total allowed latency time for the L1 Trigger is 3.2 μs.

All events that pass the L1 Trigger are sent to a computer farm (Event Filter), that performs physics selections, using faster versions of the offline reconstruction software, to filter events and to achieve the required output rate. The HLT is able to reduce the rate of recorded events down to 100 Hz and only these events are stored and processed by the Data Acquisition (DAQ). The whole trigger chain is outlined in Figure 3.11, where the different trigger operations are shown, together with hardware and software parts used in each step and rate of events, until the data storage.

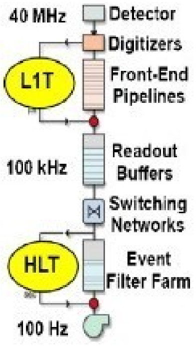


Figure 3.11: Sketch of the different steps of the CMS trigger system [162]: thanks to the two different trigger levels, it reduces the rate of events from an initial value of 40 MHz down to 100 Hz, which is the effective quantity of data recorded and stored for further analysis.

3.8.1 Jet triggers

In CMS, triggers specifically dedicated to the jet physics are used for QCD studies. At L1, they use mainly information from the calorimeter, looking for an energy cluster and a high energy deposit [165]. In particular, they evaluate transverse energy sums (in ECAL + HCAL), computed in 4×4 trigger towers, except in the HF region where this quantity is measured in the whole trigger tower

---

[1] For a high luminosity phase, the possibility to use tracker information in the L1 has been investigated [164]; with the current layout, this is not feasible because of the slowness of the response.

itself. If this deposit is greater than a certain threshold, the event is selected at L1 and it is passed to the HLT. Then, a simple and fast iterative cone algorithm is used [163]. The algorithm uses all the calorimeter towers and has two parameters:

- the size of the cluster cone R (generally set to 0.5 or 0.7);
- the seed threshold applied to each tower (generally set to 2 GeV).

Then, the output of the algorithm, using the tower signals, is converted into units of energy and if it exceeds a jet threshold, the event is selected and further processed. More specific and demanding triggers are available for multijet scenarios. Since the rate of jet events is quite high, the jet triggers are generally highly prescaled; the prescale depends on the p_T threshold and on the luminosity; for 2010 data, for instance, it is about 30 for trigger jets of 15 GeV and about 8 for jets of 30 GeV.

3.9 Data Quality Monitoring

Data quality monitoring (DQM) [166, 167] is the standardised chain to certify the recorded data for physics analyses. The purpose of the DQM is to discover and to trace errors due to the detector hardware or the data reconstruction software occurring during the data taking in order to have a high operation efficiency. The DQM applications receive event data and produce histograms to monitor elements and check the quality of the results for each subdetector, specifying run number and number of events. By looking at those histograms and comparing them with a reference, it is possible to check the detector performance for each data run and, in case, to pin-point and identify anomalies. The DQM procedure is organized in shifts of two types [167]: online and offline shifts. The former ones aim to spot problems during the run from simple and fast detector responses, while in the latter ones, more elaborated information from reconstruction algorithms and improved calibration and alignment, are added to the online ones and the global data certification is produced. The online ones take place every day for 24 hours during the detector operation at the CERN CMS control room, where the experiment is located while the offline ones are carried out at CERN during day time and supported by Fermilab and DESY with remote shifts. Shifters change every 8 hours and are well instructed by web interfaced tutorials. The software, used in the CMS DQM, is a run registry database; the role of the shifter is to fill into the database basic run information about the examined data set, adding any pertinent observations relative to it based on the available histograms. For each subdetector, a certification flag “good” or “bad” can be used in order to judge if those data can be used for physics analyses. If a “bad” flag is set for certain data, it is advisable to inform the subdetector group about the ongoing problem or ask for clarification. The final combined quality results, which include the information from the online and offline shifters, is then communicated to the detector and physics object groups for confirmation and final certification. At the end of the DQM process, a list of good data sets is produced, ready to be used and filtered for any physics analyses. The whole certification procedure takes a time that varies between days and weeks.

Chapter 4

Event simulation

In this Chapter, a description of how an event in particle physics is simulated is provided. The event simulation includes both the physics of the generated process, arising from a defined initial beam state and producing a bunch of particles in the final state, and the interactions of the final particles with the detector. The whole machinery uses the generation of random numbers through a Monte Carlo (MC) method, and it is generally divided in different steps. First of all, the physics process is generated by an event generator, which implements a ME for the hard scattering, interfaced with an UE simulation. The event generators, used in this thesis, start from a pp initial state and have been presented in Chapter 2. This generation calculates the cross section of the hadronic process, takes into account scattering amplitudes and particle decays, treats consistently beam remnants and MPI, and produces a collection of particles which constitute the final state. The way how the event generator implements all this information is ruled by random numbers (see Section 4.1) which decide, for instance, which hard process or specific decay occurs, how the particles of the initial and final state evolve and which are their kinematic variables. Indeed, the physics needed for the generation is given as an input to the generator itself, such as, for instance, the PDF set, the hard scattering, the branching probabilities, the hadronization mechanism and the MPI model.

The second step of the simulation is the interaction of the physical particles of the final state with the detector. This procedure requires a sophisticated and complex simulation of the detector material and of the behaviour of the particles in it. Particles may split via bremsstrahlung, electromagnetic or hadronic showers, may change direction due to multiple scattering or just decay inside the material. All these effects need to be taken into account in the detector simulation and all the particles, additionally created, need to be considered and treated in the same way. After that, the signal, produced by particles crossing the detector is also simulated and this is given in output with the same format as for the real data. Finally, the simulated detector response is used for the physics reconstruction which, by associating signals from different subdetectors or different parts of the same subdetector, creates the objects used for the analysis. It is important that the reconstruction algorithms, processed in the simulation, are the same as the ones used for the data, in order to have a direct comparison of the two. The whole detector simulation is performed by using the Geant4 software, described in Section 4.2.2.

The event simulation definitely plays a key role in a physics analysis: it is needed in order to correct the data, by understanding and removing the detector effects, and to produce measurements at the so-called "generator level" or "stable particle level". The "generator level" or "stable particle level" is defined as the collection of stable particles from the full ME+PS generation without the simulation of the interaction of these particles with the detector components. It is basically the "truth", without any effect introduced by an experimental device. By having a sample of generated events with the full detector simulation, it is indeed possible to build a detector response matrix, apply that to the measured data and produce, as output, detector-independent results[1]. Data at the stable particle level are easier to compare to any model, without going through the very time consuming detector simulation. Indeed, comparisons of predictions obtained with different and various models at the stable particle level are crucial for the interpretation of the data: they help to understand the measurements, exclude or confirm theoretical hypotheses, or have evidence for predicted effects. Furthermore, the possibility for event generators to switch off specific parts of the simulation gives chance to understand the different pieces which constitute a model.

4.1 The Monte Carlo method

A key role in event generation is played by the MC method. It uses the production of random numbers through a mathematical algorithm to perform different tasks, useful in particle physics. Through random numbers, decisions are taken by the simulation programs about, for instance, which processes are generated in the collision, if particles decay or not and in which channel, how partons evolve and many more other actions. A program which implements the MC method to get physics predictions, is called hereafter "Monte Carlo generator", while the predictions themselves are referred to as "Monte Carlo predictions". Indeed, the physics of the generated particles, namely their evolution, their decay modes and their couplings, needs to be given as input to a MC generator. The randomness of the MC numbers guarantees the probabilistic behaviour of the produced processes. To get predictions based on the input physics, a MC generator must be able to accomplish different functions. They are, for instance, the generation of numbers according to a specified probability distribution or the calculation of integrals.

The starting point of any MC program is the generation of a random number R_{i}, uniformly distributed between 0 and 1. It is, however, possible to translate this flat distribution, into an arbitrary function $f(x)$, and two main methods can be used:

- the inverse transform method [168]: given the function $f(x)$, it consists of inverting the integral equation $R = \int_0^x f(t)dt$. From this operation, one gets the function $g(R) = x$. By generating R between 0 and 1 and applying the function $g(R)$, the final set of random numbers is distributed like the initial function $f(x)$. This method is only applicable if the probability function is analytically integrable;

- the acceptance-rejection method [168]: given the function $f(x)$ and the maximum value of the function f_{max} in a given range [a,b], this method works by comparing the value of the function $f(x_{\mathrm{i}})$, where x_{i} is defined as $x_{\mathrm{i}} = a + (b-a)R_{\mathrm{i}}$, with the generated number $R_{\mathrm{i}} \cdot f_{\mathrm{max}}$. If $R_{\mathrm{i}} \cdot f_{\mathrm{max}} < f(x_{\mathrm{i}})$, the generated number R is accepted, otherwise is rejected. At the end of the generation, again, a set of random numbers R, distributed like $f(x)$, is obtained. Note that also for this method, the function $f(x)$ is required not to be divergent in the range [a,b].

---

[1] This is the unfolding procedure, described in Chapter 8.

Generation of random numbers according to any probability distribution, is indeed very crucial because physics is described by various mathematical functions, generally different from flat ones and even very complicated.

Furthermore, MC programs are able to estimate the value of integrals by using a statistical method [37]. From the definition of expectation value applied to an integral of a function $f(x)$ between a and b, it results that the value of the integral is:

$$I = (b-a)\frac{1}{N}\sum_{i=1}^{N} f(x_\mathrm{i}) \tag{4.1}$$

where $f(x_\mathrm{i})$ is the value of the function, evaluated in the point x_i, randomly generated, and N is the total number of generated numbers. To the value of the integral, a variance, depending on the number of generated numbers, is associated:

$$V = (b-a)^2\frac{1}{N}\left(\frac{1}{N}\sum_{i=1}^{N} f(x_\mathrm{i})^2 - \left(\frac{\sum_{i=1}^{N} f(x_\mathrm{i})}{N}\right)^2\right) \tag{4.2}$$

This method gives a probabilistic value of the integral and of its uncertainty and is indeed essential, for instance, for cross section calculations or parton evolution computations. The higher is the number of generated numbers, the more accurate and precise is the result of the integration. It becomes particularly important when calculating multi-dimensional integrals, where an analytical evaluation would be very complicated and not always possible.

Note that the numbers, generated by a MC program, are not exactly random by definition, since they follow a specific algorithm, but they can reach a high degree of randomness, which is possible to check through several statistical tests. The algorithms, which are generally used, are the basic congruential algorithm [168], based on a recursive definition starting from a seed and making use of three parameters, the modulus, the multiplier and the increment, or the RANLUX [169] one, which uses a more elaborate function with the mantissa operation. These two are implemented in libraries, like CERNLIB [170] and ROOT [171] and the user can choose which one to apply for the generation, depending on the required complexity of the random series. More complicated algorithms are also available, in order to reach a higher degree of randomness of the generated numbers. Since the generation of this type of random numbers, generated by MC programs, is ruled by a defined algorithm, they are thus called “pseudo-random”. Truly random numbers could be only generated by physical processes, like radioactive decays, atmospheric or thermal noise, or other quantum phenomena.

4.2 CMS detector simulation

The simulation of the CMS detector and the MC production for the different analyses are organized in different steps and done centrally inside the CMS software, called CMSSW [172]; a specific group inside the collaboration provides each analyzer with the needed samples. The separated operations of the sample generation serve to divide the production in jobs shorter in time, and also to produce different samples which can be used for different purposes. Three different steps are performed and identified with the following abbreviations:

- GEN-SIM: it produces the generator level MC event with all the particles in the final state and their corresponding energy depositions in the detector (the SIM hits), simulated with Geant4;
- RAW-DIGI: the SIM hits are converted into detector response; this is the raw information, coming out from the detector after any internal activity. Trigger information is also included in this step;
- RECO-AOD: the detector response is processed and reconstructed objects, like tracks, vertices, jets, electrons and muons, are produced; the AOD (Analysis Object Data) format is a subset of the RECO, which contains all the information needed for an analysis and is smaller in size.

Note that the output of each step serves as input for the following one, until the final RECO samples are produced. A physics analysis is generally performed by using the AOD samples which contain all the relevant information of reconstructed objects at the detector level. However, the samples produced at the end of the intermediate steps are also very important: a GEN-SIM sample is useful in case of comparison with data corrected to stable particle level, while with a RAW-DIGI sample, studies about details of detector reconstruction or noise can be performed. For the storage and the processing of the samples, CMS uses a tiered structure [173]: MC samples are generally stored in Tier-2 centers, which can be accessed by the user for processing.

For each step of the generation, configuration files, in python language [174], are available and easy to create through the CMSDRIVER command. Then, jobs need to be submitted externally to the GRID [175] and the final samples are returned in the local storage. A good exercise, during this work of thesis, has also been to simulate some samples privately: this helped to get a closer look at the different MC generators and understand which information is stored and needed at each generation step.

4.2.1 Pile-up simulation

In order to have a consistent and reliable picture of a collision at the LHC, pile-up interactions need to be simulated as well. Standard event generators, like PYTHIA, have the possibility to generate more than one interaction during the same collision and to overlap them. The additional interactions, which are generated on top of the primary hardest one, are mainly MB events. The larger is the number of generated pile-up events, the longer is the computation time. In a real physics analysis in CMS, pile-up events are added to the MC samples by randomly overlapping MB events, simulated separately. This procedure allows the inclusion of a high number of additional interactions without increasing dramatically the time needed for the simulation. Generally, pile-up events are added on top of the hard scattering according to a flat distribution up to a maximum value which can go up to 40. It is then a task of the analyzer to match the pile-up distributions in the considered MC and data samples.

The pile-up events are added at the RAW-DIGI step of the generation. Hence, objects coming from pile-up events do not have generator level information but only the one at the reconstruction level. In low pile-up runs, as the ones used in this thesis, this property can be used to estimate the pile-up contribution. In fact, measured detector jets, which do not have a corresponding object at the generator level, can be identified as pile-up jets and events with jets with this feature can be then subtracted. This strategy of pile-up removal is performed inside the unfolding procedure (see Chapter 8).

In high pile-up environments, more sophisticated strategies need to be applied, but they will not be described in this work. Interesting references are [176, 177].

4.2.2 The Geant4 software

The Geant4 (GEometry ANd Tracking) package [178] is a platform for the simulation of particles crossing detector materials. It includes several tools which are interfaced to the user for the definition of the specific detector used in the experiment:

- Geometry which includes the position of the active detector, together with passive support structures, dead material and technical constituents, such as cooling pipes and beam tubes;
- Tracking which simulates the crossing of the particles through the matter. It implies the knowledge for the software of all the possible interactions and particle decays;
- Detector Response which simulates how a real detector would respond to an injected particle; it takes into account quantities like detection efficiency, absorption and transparency effects, and propagation trajectories. This tool gives on output the signal expected from the detector, in the same format as the output available in real data;
- Run manager which registers the details of simulated runs and events, storing them as if it is set in a real experiment;
- Visualization which allows the 3D visualization of an event, based on a graphical interface.

Every particle is propagated along its direction inside the detector for a defined length, the step of the simulation, in 3D space: at every step, a decision based on a random number generation is taken for interaction/no interaction with the material. In case of no interaction, the particle stays the same and it is propagated to the next step, while in case of interaction, the particle may emit secondary particles, change direction, release energy or all of the previous actions and it is propagated to the next step in the modified state. The step of the simulation should be as large as the average interaction path in the material; the smaller it is set, the longer the processing time of the simulation results. The length of the step changes with respect to the physics processes which the propagated particle is subjected to, inside the crossed material.

At the end of the simulation, the detector responses are used in order to apply the same reconstruction algorithm as in the data, so that a consistent comparison between the two is possible. In case deviations are observed between data and simulation for specific quantities, the latter is adapted to the former, namely the simulation is tuned to the data. By using this software, excellent predictions are obtained for most of the measured physics channels in CMS, whose geometry and response are very well understood and modelled.

A sophisticated simulation has, however, to pay the price of a long processing time: indeed, the more complicated the detector geometry is, the longer is the time needed to process an event. In CMS, a standard processing time is of the order of a minute for each event, depending on the process; hence, large computing resources are dedicated to it. Sometimes, mainly in simulation studies, a "fast simulation" [179] is used for the detector simulation: it uses a parametrization of the detector response and it is faster but less accurate. However, in this work, a full Geant4 simulation has been used for the MC sample generation.

Besides particle physics, the Geant4 software is widely used in nuclear studies and has applications in accelerator physics, astrophysics and medicine.

4.3 Monte Carlo predictions for the four-jet analyses

By using various MC event generators, different predictions have been obtained at the stable particle level, for the analyses presented in this thesis. These studies have been very useful in order to test models and search for improvements in four-jet scenarios, and to understand the different behaviours of details of the simulation. As already described in Chapter 2, event generators make use of a combination of ME+PS+UE. Some generators are able to produce all these components, while some others calculate only the ME and need to be interfaced to other generators, which provide the remaining parts (PS+UE). The simulation of PS and UE relies on the use of a tune[1], namely a set of parameters which optimizes the description of measurements sensitive to PS and UE. The definition of a tune has been so far performed only for generators, like PYTHIA, HERWIG and SHERPA, which simulate all the components. These tunes are then also applied to the interface of the same generator with other generators, like POWHEG or MADGRAPH, which calculate only the matrix element. Some discussion is still open whether this procedure is correct, namely whether a tune is independent of the ME, or, instead, specific tunes should be obtained for the different ME generators used. Results on this topic are presented in Chapter 10.

Predictions used for the comparison with four-jet measurements are considered. Two different selections have been considered in these studies: one involves jets without any flavour tagging, while in the other one, the presence of heavy-flavour jets is required. Details of the two selections are further described in Chapter 6. MC generators and corresponding tunes used in the simulation are listed in the following and summarized in table 4.1:

- PYTHIA6 Tune Z2*: Tune Z2* [180] has been obtained inside the CMS collaboration by refining the description of UE data with a slightly higher MPI contribution with respect to the tune Z2 [181]. It uses the CTEQ6L1 PDF set [68]. Samples for these predictions have been obtained by generating 2→2 processes with exchanged transverse momentum of the outgoing partons of the ME, $\hat{p}_{\mathrm{T}}$, greater than 45 GeV and 15 GeV for the two analyses, described in the thesis;
- PYTHIA8 Tune 4C: Tune 4C [124] uses the CTEQ6L1 PDF set and it is tuned to CDF UE and dijet data at $\sqrt{s} = 1.8$ TeV and early LHC data at 7 TeV; the generated processes refer to 2→2 QCD diagrams with $\hat{p}_{\mathrm{T}} > 45$ GeV and 15 GeV for, respectively, the two selections;

---

[1] For how to perform a tune and results about an energy-dependent tune performed in CMS, see Appendix D.

- Herwig++ Tune UE_EE_3: UE_EE_3 [117, 182] is a UE tune, performed with MRST-2008LO** PDF set [183], which achieves a good energy dependence description and a careful choice of colour reconnection parameters [127]. As done for the other 2→2 QCD processes, the $\hat{p}_{\mathrm{T}}$ was required to be larger than 45 GeV and 15 GeV for the two different selections;
- Powheg+Pythia6 Tune Z2* and Powheg+Pythia6 Tune Z2': Z2*, obtained originally with Pythia6, is propagated to Powheg; Tune Z2' has been obtained by reducing the phase space of the hard radiation in the tune Z2*. This tune has been validated for a four-jet scenario and used for the first time in the analyses described in this thesis. A systematic study of comparisons and validation is described in Chapter 10. The sample relies on the generation of a NLO dijet ME with a real hard emission included, produced with a $\hat{p}_{\mathrm{T}} > 15$ GeV;
- Madgraph+Pythia6 Tune Z2*: Z2* is also used in the interface with Madgraph; in this analysis, the ME includes up to four partons in the final state on the basis of Leading Order (LO) ME calculations, by generating 2→2, 2→3 and 2→4 diagrams, matched with PS. The ME/PS matching scale is taken to be 10 GeV, within the MLM scheme [76]. The validation of tune Z2* is also shown in Chapter 10. A binned generation has been performed for the Madgraph predictions: the four generated bins are organized as a function of the scalar p_{T} sum of the partons in the final state, H_{T};
- Sherpa 1.4: for Sherpa, the tune documented in [184] has been used with the CTEQ6L1 PDF set; this has been obtained by tuning UE data. The Sherpa predictions are based on 2→2 and 2→3 QCD processes, matched with PS, and generated with $\hat{p}_{\mathrm{T}}$, greater than 45 GeV and 15 GeV for the two analyses.

Generator	PDF	Tune
Pythia6	CTEQ6L1	Z2*
Pythia8	CTEQ6L1	4C
Herwig++	MRST2008**	UE_EE_3
Powheg+Pythia6	CT10	Z2*
Powheg+Pythia6	CT10	Z2'
Madgraph+Pythia6	CTEQ6L1	Z2*
Sherpa	CTEQ6L1	UE

Table 4.1: List of event generators, along with the PDF implemented in the calculation of the matrix element and the tune for the UE simulation, used for obtaining predictions of four-jet scenarios.

4.4 Sensitivity studies at generator level

Before going to the data measurement, preliminary studies have been performed at the generator level in order to check the sensitivity of four-jet scenarios to DPS. In particular, correlation observables, defined through kinematical observables of jets in the final state, which show different behaviours for single chain processes and double parton scatterings, are studied. From this study, it has been possible to estimate the size of the DPS contribution in current generators, the regions of the phase space where to expect a DPS signal, and which correlation observables are most sensitive to it. Different behaviours are expected between SPS and DPS events in the event topology: if the four jets are produced through the same process chain in the SPS, a high correlation between the objects of the final state is obtained and this is reflected in their relative configuration in the transverse plane. The direction of the hard jets, for example, is randomized by the emission of the additional two jets within the same chain and their initial p_T balance is ruined. Instead, jet pairs coming from DPS events, namely from two independent scatterings, tend to be uncorrelated and their initial back-to-back configurations are less subject to smearing effects coming from additional hard radiation: they are expected to exhibit a more balanced configuration in p_T and azimuthal angle.

Thus, the final aim is to separate and disentangle the SPS and DPS contributions by looking at the measurements of the correlation observables. A separation between the two contributions on an event-by-event basis is not doable, since it is not possible to identify whether a single event has been produced by a single or a double chain process. What is indeed possible, is the discrimination of the two production channels by looking at correlation observables sensitive to DPS: results should be a combination of different fractions of SPS and DPS contributions. For the study of the correlation observables, the PYTHIA8 event generator has been considered, with the tune 4C; distributions with HERWIG++ have shown a very similar behaviour and are not presented here.

This study has been performed by using the RIVET machinery [185] (see Appendix D for a detailed description). Results are described in the following. Quantitative values are provided for the sensitivity of the correlation observables to the DPS, in terms of shape comparisons and absolute cross sections. Two different selections have been set at the generator level: one with exactly four jets in the final state within $|\eta| < 4.7$, with two hard jets with $p_T > 50$ GeV and two soft jets with $p_T > 20$ GeV, and one with at least four jets in the final state with $p_T > 20$ GeV, of which two b-jets in $|\eta| < 2.4$ and two other jets in $|\eta| < 4.7$ are selected[1]. A summary of the phase space selections is given in Tables 4.2 and 4.3.

With four-jets in the final state, well defined physics objects can be associated in order to define the correlation observables. For the first selection, the hard jets are grouped together to the "hard-jet" pair, while the soft jets in the "soft-jet" pair. For the second selection, the two b-jets form the b-pair, while the light-pair[2] is composed by the two other jets. In this way, correlation observables can be defined and the ones considered in these analyses have been:

- the azimuthal angular differences between the jets belonging to the same jet pair, and the quadrature sum of these two quantities:

$$\Delta\phi^{\text{pair}_1} = |\phi^{\text{object}_1} - \phi^{\text{object}_2}| \tag{4.3}$$

$$\Delta\phi^{\text{pair}_2} = |\phi^{\text{object}_3} - \phi^{\text{object}_4}| \tag{4.4}$$

$$S_\phi = \sqrt{(\Delta\phi^{\text{pair}_1})^2 + (\Delta\phi^{\text{pair}_2})^2} \tag{4.5}$$

- the balance in transverse momentum of the two associated jets, and the quadrature sum of

---

[1] The choice of such pseudorapidity ranges is mainly driven by detector-based reasons. The possibility to detect jets at more forward pseudorapidities, down to η = -6.6, has been attempted in the CMS collaboration, by using the CASTOR calorimeter, in order to be sensitive to partons colliding at lower x-values inside the proton. Work related to a jet calibration in CASTOR is described in Appendix A.

[2] In the following, observables which refer to the two additional jets are labelled with the tag "light", in order to distinguish them from the b-jet observables.

these two quantities:

$$\Delta^{\text{rel}}_{\text{pair}_1} p_{\text{T}} = \frac{|\vec{p}_{\text{T}}^{\,\text{object}_1} + \vec{p}_{\text{T}}^{\,\text{object}_2}|}{|\vec{p}_{\text{T}}^{\,\text{object}_1}| + |\vec{p}_{\text{T}}^{\,\text{object}_2}|} \tag{4.6}$$

$$\Delta^{\text{rel}}_{\text{pair}_2} p_{\text{T}} = \frac{|\vec{p}_{\text{T}}^{\,\text{object}_1} + \vec{p}_{\text{T}}^{\,\text{object}_2}|}{|\vec{p}_{\text{T}}^{\,\text{object}_1}| + |\vec{p}_{\text{T}}^{\,\text{object}_2}|} \tag{4.7}$$

$$S'_{p_{\text{T}}} = \sqrt{(\Delta^{\text{rel}}_{\text{pair}_1} p_{\text{T}})^2 + (\Delta^{\text{rel}}_{\text{pair}_2} p_{\text{T}})^2} \tag{4.8}$$

- the pseudorapidity differences between the jets belonging to the same pair:

$$\Delta\eta^{\text{pair}_1} = |\eta^{\text{object}_1} - \eta^{\text{object}_2}| \tag{4.9}$$

$$\Delta\eta^{\text{pair}_2} = |\eta^{\text{object}_3} - \eta^{\text{object}_4}| \tag{4.10}$$

- the azimuthal angle ΔS between the two dijet pairs, defined as:

$$\Delta\text{S} = \arccos\left(\frac{\vec{p}_{\text{T}}(\text{object}_1, \text{object}_2) \cdot \vec{p}_{\text{T}}(\text{object}_3, \text{object}_4)}{|\vec{p}_{\text{T}}(\text{object}_1, \text{object}_2)| \cdot |\vec{p}_{\text{T}}(\text{object}_3, \text{object}_4)|}\right) \tag{4.11}$$

where object_1 (object_2) and object_3 (object_4) are respectively the leading (subleading) jets of the first and second jet pairs and pair_1 (pair_2) indicates the first (second) jet pair.

Different samples have been generated with PYTHIA8, using the possibility to switch off and on specific parts of the simulation:

- SPS: a sample where the contribution of the MPI is switched off (PartonLevel:MI = off);
- DPS-NO PS: a sample where two hard scatterings are forced to occur in the same *pp* collision and the contribution of the PS is switched off (PartonLevel:PS = off);
- DPS: a sample where two hard dijet scatterings are forced to occur in the same *pp* collision ('SecondHard:generate = on', 'SecondHard:TwoJets = on');
- Heavy flavour production: a sample where *b*-quark pairs are generated only through FC processes (see Chapter 1);
- Inclusive: a nominal QCD sample with all the parts of the simulation switched on; all processes producing *b*-quarks (FC, FEX and GS) are included in the generation.

For the samples where the four-jet selection has been applied, the first hard scattering has been generated with a $\hat{p}_{\text{T}}$ higher than 45 GeV and a secondary hard scattering with a $\hat{p}_{\text{T}}$ larger than 15 GeV. In the two *b*- and two other jet selection, first and second hard scatterings have been generated with a $\hat{p}_{\text{T}}$ larger than 15 GeV. The choices of these values maximize the efficiency of the generation and cover the whole phase space of the two selections. For the *b*-jet selection, a separate study has been performed for a Heavy Flavour *b*-production (HardQCD:hardbbbar) and an inclusive QCD production (HardQCD:all), in order to study the behaviour of the different *b*-jet production diagrams. The behaviour of correlation observables relative to soft and hard jets and to *b* and light jets is very similar between each other. Results for $\Delta\eta$ observables showed a negligible sensitivity to DPS contributions; this is the reason why they are not presented here. Hence, results are presented for a selection of observables, namely ΔS, $\Delta^{\text{rel}}_{\text{soft}} p_{\text{T}}$ (for the first selection) and $\Delta^{\text{rel}}_{\text{light}} p_{\text{T}}$

(for the second selection), and $\Delta\phi^{\text{soft}}$ (for the first selection) and $\Delta\phi^{\text{light}}$ (for the second selection).

Exactly four jets
Two hard jets: $p_{\text{T}} > 50$ GeV in $\|\eta\| < 4.7$
Two soft jets: $p_{\text{T}} > 20$ GeV in $\|\eta\| < 4.7$

Table 4.2: Selection criteria applied to MC in the four-jet analysis.

At least four jets	
At least two b-jets: $p_{\text{T}} > 20$ GeV	$\|\eta\| < 2.4$
At least two additional jets[1]: $p_{\text{T}} > 20$ GeV	$\|\eta\| < 4.7$

Table 4.3: Selection criteria applied to MC in the two b- and two other jet analysis.

Figure 4.1 shows the normalized differential cross sections of the correlation observables for the SPS and the DPS-NO PS samples, using the four-jet selection. This comparison helps to understand the regions of the phase space filled by "pure" SPS events, without any MPI contribution, and "pure" DPS events, without any contribution from additional hard radiation. All the observables show a very different behaviour. For $\Delta\phi^{\text{soft}}$ and $\Delta^{\text{rel}}_{\text{soft}}p_{\text{T}}$, SPS produces a broad shape over the whole phase space, while DPS exhibits a peaked distribution at low values of $\Delta^{\text{rel}}_{\text{soft}}p_{\text{T}}$ and values of $\Delta\phi^{\text{soft}}$ close to π. This is because the jets of the same pair tend to be in a back-to-back configuration in DPS, well balanced in azimuthal angle and transverse momentum, while they have a higher correlation when they are produced by a single chain process. This feature is also confirmed for ΔS cross sections, where SPS has a distribution peaked at π, while DPS is much flatter. ΔS uses information from both jet pairs. The reason of this behaviour is again because the jet pairs produced via SPS are highly correlated in their configuration in the transverse plane, while pairs originated by two independent scatterings are uncorrelated and randomly distributed in ΔS. From this comparison, it appears that all these correlation observables, defined by the selected jet pairs, present a high discriminating power to disentangle SPS and DPS events, just from the jet configuration in the transverse plane. In particular, DPS events tend to contribute at low values of ΔS and $\Delta^{\text{rel}}_{\text{soft}}p_{\text{T}}$ and values of $\Delta\phi^{\text{soft}}$ close to π. This feature is observed also for two b- and two other jets in the final state. However, the samples examined in this comparison are not very realistic: in particular, PS has been switched off for the DPS but it is indeed an important component for such sample. Nevertheless, this preliminary study helps to understand the different event configurations and the effects of the different pieces of the UE simulation.

In Figure 4.2, a more realistic scenario is considered for the four-jet selection: the normalized cross sections of the same observables are compared among the SPS, the Inclusive and the DPS samples. The PS is now activated for the DPS sample. Small differences appear between SPS and Inclusive samples in all correlation observables; this is an indication that the contribution of MPI to this channel, that is already implemented in the simulation, is quite small. Differences are in the regions of phase space where a DPS signal is expected. Bigger differences appear between the DPS and the other two samples. DPS samples exhibit much broader distributions than the ones

---

[1] A further requirement for the flavour of the additional jets has been set: only light and charm jets are selected while b-jets are not considered for the analysis. The impact of these excluded events has been evaluated and consists of 4–5% of the total cross section.

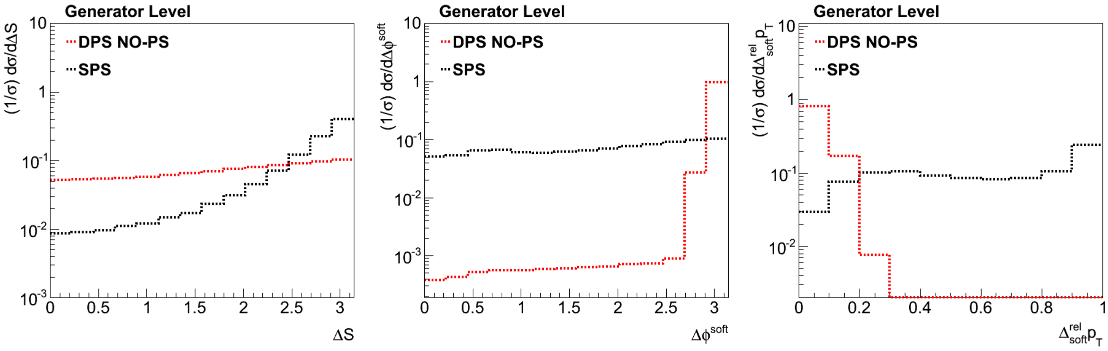


Figure 4.1: Differential cross section, normalized to the total number of events, as a function of ΔS (left), $\Delta\phi^{soft}$ (center), and $\Delta^{rel}_{soft}p_T$ (right). Distributions are shown for two QCD samples when the exclusive four-jet scenario is selected: one where MPI are switched off (SPS) and another one where two hard scatterings are generated and not interfaced with the PS (DPS). All samples are generated with Pythia*8.*

shown in Figure 4.1, under the smearing effect of the PS. However, shapes obtained for DPS and SPS events are still very different and able to discriminate the two processes. Differences between DPS and Inclusive samples are up to 30-40%.

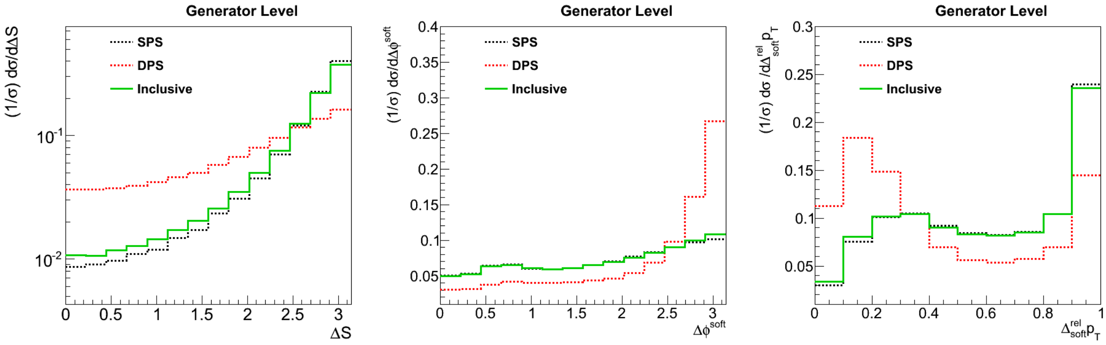


Figure 4.2: Differential cross section, normalized to the total number of events, as a function of ΔS (left), $\Delta\phi^{soft}$ (center), and $\Delta^{rel}_{soft}p_T$ (right): distributions are shown for three QCD samples when the exclusive four-jet scenario is selected: the nominal sample (Inclusive), generated with tune 4C, a sample where MPI are switched off (SPS), and a sample where two hard scatterings are generated and are then interfaced with the PS and hadronization (DPS). All samples are generated with Pythia*8.*

This conclusion holds for a selection with two b- and two other jets, as well. Figure 4.3 and 4.4 show the normalized differential cross sections as a function of the correlation observables for samples with the first hard scattering generated, respectively, via Heavy Flavour and QCD inclusive diagrams. These two generations help to understand the contribution of the different diagrams producing b-jets in the final state. A Heavy Flavour production shows large differences between the three different samples, while the shapes start to be closer for an inclusive QCD generation. This is mainly due to the additional FEX and GS diagrams introduced by the QCD production and not present in a Heavy Flavour scenario, which may produce collinear jets in the final state. In particular, in DPS events, they fill the phase space at high values of $\Delta^{\mathrm{rel}}_{\mathrm{light}}p_{\mathrm{T}}$ and at low values of $\Delta\phi^{\mathrm{light}}$ where, instead, SPS events are mainly contributing. However, even though the shapes of the different samples tend to resemble and the emission of collinear jets contaminates the peculiar back-to-back configuration of jets in DPS events, relevant differences are still present between SPS and DPS distributions and the sensitivity of the correlation observables to DPS events is still large for this selection, as well.

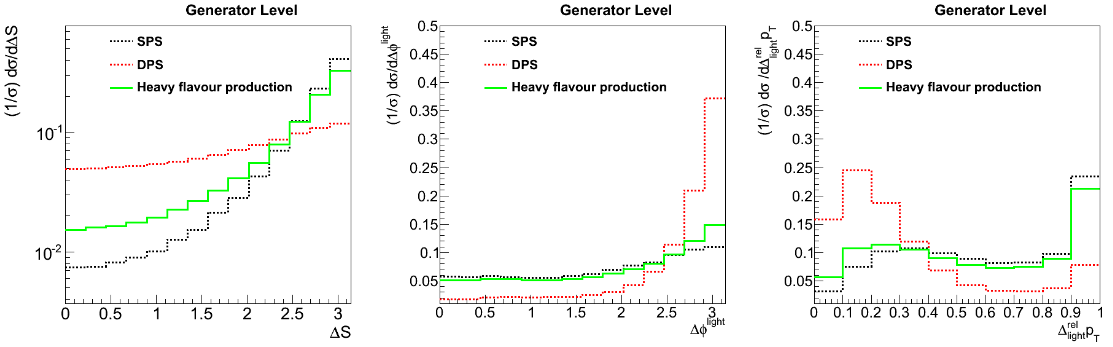


Figure 4.3: Differential cross section, normalized to the total number of events, as a function of ΔS *(left),* $\Delta\phi^{light}$ *(center), and* $\Delta^{rel}_{light}p_T$ *(right): distributions are shown for three samples, generated by a heavy flavour ME, when two b- and two other jets are selected: the nominal sample (Inclusive), generated with tune 4C, a sample where MPI are switched off (SPS), and a sample where two hard scatterings are generated and are then interfaced with the PS and hadronization (DPS). All samples are generated with* PYTHIA*8.*

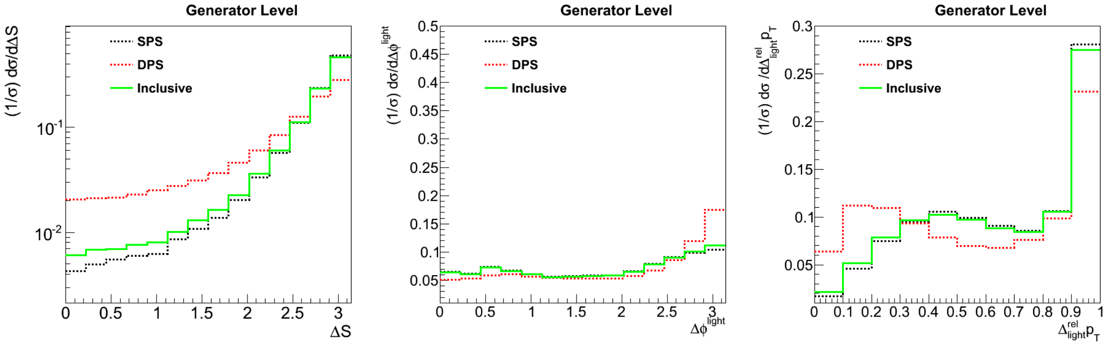


Figure 4.4: Differential cross section, normalized to the total number of events, as a function of ΔS *(left),* $\Delta\phi^{light}$ *(center), and* $\Delta^{rel}_{light}p_T$ *(right): distributions are shown for three samples, generated by an inclusive QCD production, when two b- and two other jets are selected: the nominal sample (Inclusive), generated with tune 4C, a sample where MPI are switched off (SPS), and a sample where two hard scatterings are generated and are then interfaced with the PS and hadronization (DPS). All samples are generated with* PYTHIA*8.*

A final comparison has been performed among the three previous samples by considering absolute cross sections instead of normalized ones. To calculate the absolute cross sections, the information from PYTHIA8 has been used for the SPS and the Inclusive samples. The cross section of DPS has been extracted by rescaling the sample through σ_{AB} as defined in Equation 2.9, assuming a value of σ_{eff} equal to 20 mb, as measured by CMS [8]. For the four jet selection, indicating with A and B respectively the first and the second hard scattering processes, σ_{A} has been set to 12.91 μb, while σ_{B} to 340 μb. For the two b- and two other jet selection, a value of 340 μb has been fixed for both processes. In both selections, the symmetry factor has been used in the definition of σ_{eff}.

Even with cross sections at LO, this study helps to get an overview of the absolute DPS contribution in the considered channels. Figure 4.5 shows the results for both selections. In the four-jet selection, a very small contribution comes from DPS compared to the overwhelming background of SPS: the DPS signal in the region where it is expected to be important, is of the order of 5-10% of the inclusive sample. In the two b- and two other jet selection, instead, a much higher DPS contribution is expected with respect to SPS: in particular, in some bins of ΔS, the DPS cross section is 60–80% with respect to SPS. For the other correlation observables, a significant DPS contribution is also observed of up to 40% for $\Delta^{\mathrm{rel}}_{\mathrm{light}} p_{\mathrm{T}}$ and 20% for $\Delta\phi^{\mathrm{light}}$. However, in both selections, ΔS appears to be the most discriminating observable: a reason is that this is the only variable which considers the whole event topology and not only information from one of the two jet pairs.

An important remark needs to be made about the sensitivity of the two selections: in general, the two b- and two other jet selection appears to be more sensitive to DPS events. This is mainly due to the differences in the phase space used in the two analyses: the selection of four jets at the same threshold strongly suppresses further emission of radiation within the same chain. This choice is possible only for the two b- and two other jet scenario, because the information related to the different jet flavour can be easily used for the jet association in pairs. In the four-jet analysis, flavour information is not available and different p_{T} thresholds need to be applied in order to distinguish hard and soft jets and associate them in pairs. The price to pay for applying different p_{T} thresholds, is an increase of the background coming from SPS events and obviously, the sensitivity to the DPS signal is lower in presence of a higher background. In addition, different selections have been tried in the attempt to maximize the DPS sensitivity: for instance, for the four-jet selection, the requirement for exclusivity has been removed by allowing an arbitrary number of jet higher than 4 or the transverse momentum of the soft jets has been reduced only to the range between 20 and 50 GeV. It has been observed that the DPS sensitivity did not change with the restriction of the p_{T} range but significantly decreased when allowing a higher number of jets: hence, the selection on Table 4.2 has been applied since it maximizes both DPS sensitivity and selected statistics. Similar studies have been performed for the b-jet selection, where the cuts listed in Table 4.3 lead to a high discriminating power and large number of selected events.

In conclusion, the studies performed for two different physics channels with jets in the final state have shown that it is possible to define observables, which are able to disentangle the DPS contribution from the SPS one. These observables use only kinematical information from the jets of the final state and make use of the configuration and topology in the transverse plane. Correlated jet pairs are expected to be produced by SPS events, while DPS diagrams tend to generate jet pairs in a back-to-back configuration. An important role has been observed for PS, which tends to smear out the back-to-back configuration arising from DPS events, and for GS and FEX diagrams in the b-jet production, which, in the way how they are simulated in the generators, may produce a correlated configuration of the jets in DPS events, as well. However, SPS and DPS events show a different behaviour for normalized differential cross sections as a function of the correlation observables. Differences between the shapes produced by SPS and DPS events are of the order of 30–40% for both selections in some regions of the phase space: low values of ΔS and $\Delta^{\mathrm{rel}} p_{\mathrm{T}}$ and values of $\Delta\phi$ close to π, where back-to-back topologies of jet pairs mainly contribute. These differences are also observed for absolute cross sections: they are of up to 5–10%

for the four-jet selection and of up to 30–40% for the two b- and two other jet selection. These results give confidence that the measurements of these correlation observables and comparison with different models might help to discriminate the different production processes and estimate the DPS contribution.

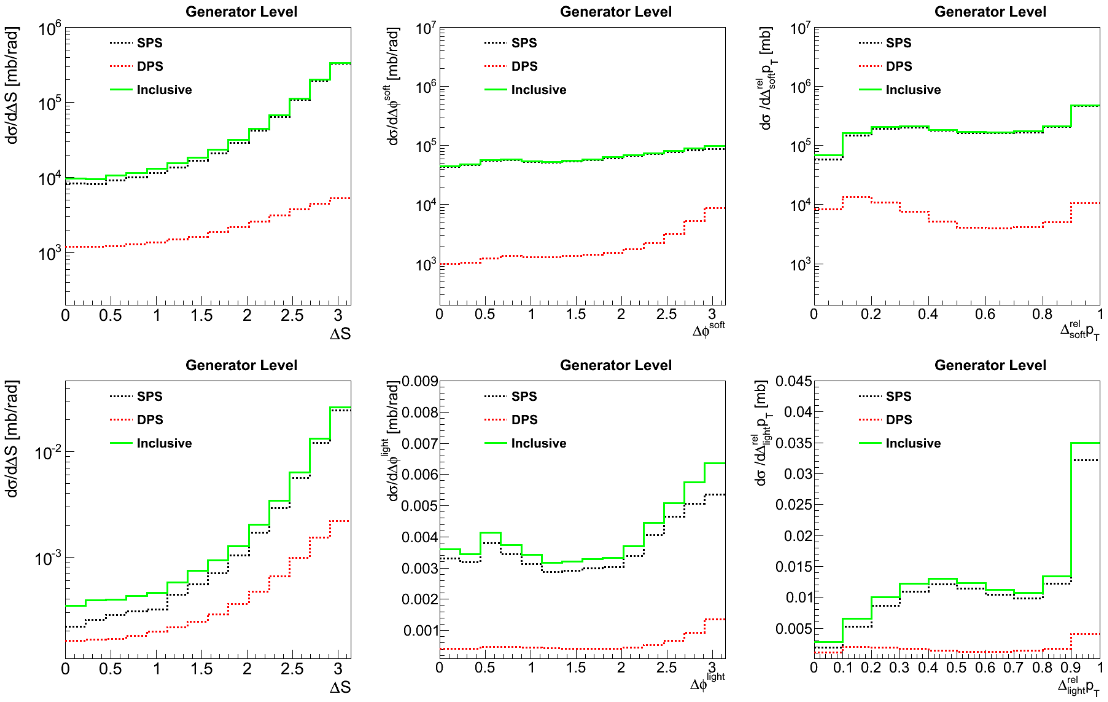


Figure 4.5: Absolute differential cross section as a function of ΔS (left), $\Delta\phi^{light}$ (center), and $\Delta^{rel}_{light}p_T$ (right), when four-light jets (top) and two b- and two other jets (bottom) are selected: distributions are shown for three samples, generated by an inclusive QCD production: the nominal sample (Inclusive), generated with tune 4C, a sample where MPI are switched off (SPS), and a sample where two hard scatterings are generated and are then interfaced with the PS and hadronization (DPS). All samples are generated with PYTHIA*8.*

Chapter 5

Event reconstruction

In this Chapter, the event reconstruction of data measured with the CMS experiment is described in full detail. From the huge amount of recorded data, it is important to extract compact information small in size, related to particles which have crossed the detector and generated the signal. In fact, for a physics analysis, the detailed detector response is not actually needed, it only has to be processed in order to extract physics objects to which the selection cuts are applied. Hence, the aim of the event reconstruction is to build well calibrated physics objects. The CMS experiment has developed a technique, called Particle-Flow (PF), which is able to identify and reconstruct individually each particle in every event, by combining information from all the subdetectors. In CMS, the PF algorithm relies on an excellent tracking efficiency in the high magnetic field and a very fine calorimeter granularity. This type of reconstruction leads to an improved performance for the detection of all physics objects and it has been used in the analyses described in this thesis. In this chapter, after a description of all objects which can be measured in CMS, a specific focus is drawn on jets. Details of the reconstruction, the energy correction and the heavy-flavour discrimination are here also described.

5.1 Physics objects in CMS

Outputs from the detector can be translated into several physics objects, depending on the type of signal, left after the particles have crossed the detector. In particular, it is important to synchronize and combine information from the different subdetectors in order to truly discriminate which particles have been actually produced during the hadron-hadron collision. In Figure 5.1, a sketch of the CMS detector is provided with focus on the signal produced by the various particles crossing different subdetectors. It can be seen that muons (light blue curve) are the most penetrating particles produced in a collision, able to cross all the subdetectors and to reach the most external layers. A photon (dashed blue line) is just seen as energy deposit in the electromagnetic calorimeter, while an electron (red curve) has additionally a track in the tracking system. Hadrons mainly consist of energy in the hadronic calorimeter; the neutral ones (dashed green line) have no associated tracks, while the charged ones (solid green line) have corresponding hits in the tracker.

These features are used by the event reconstruction to build the physics objects. The list of the reconstructed physics objects in CMS is given below, with a very brief description of their detection:

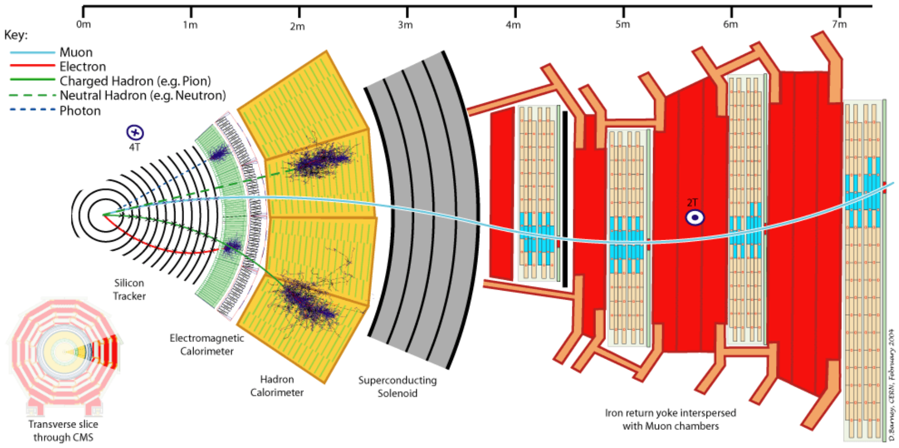


Figure 5.1: Sketch of a slice in the transverse plane of the CMS detector [186]: all the subdetectors are drawn, along with the trajectories of particles hitting the detector. The flight of a muon, an electron, a charged and a neutral hadron, and a photon is represented in the detector, with a visualization of the signals released in the crossed subdetectors.

- Jet: a jet is seen through a highly-collimated energy deposit in the calorimeters and a collection of tracks in the tracker in the same direction. Different techniques have been developed in CMS for a reliable and well calibrated jet reconstruction and they are described in Section 5.1.1;

- Muon (μ): a muon can be with high probability detected because it is the only particle whose energy is not completely absorbed by the calorimeter system and can reach the muon system. Its reconstruction makes use of a combination of hits in the muon chambers and in the internal tracker, and may use also the energy in the calorimeters;

- Electron (e): electrons can be detected by searching for signals in the inner tracker and corresponding clusters in the electromagnetic calorimeter. Quality criteria are then applied to reject fake jets or converted photons;

- Heavy flavour jet: the CMS detector is able to discriminate jets of different flavour, by identifying hadrons from the fragmentation of heavy-flavour quarks. This feature is particularly used for b-jets: their identification is based on the detection of a displaced secondary vertex from a long lived B-hadron decay, on the measurement of the jet mass or on the presence of high p_{T} leptons inside the jet cone (see Section 5.1.2);

- Tau (τ): the detector signature of τ leptons, decaying hadronically, is a collimated jet with low multiplicity (up to three charged hadrons) and constituents isolated from other particles; the reconstruction algorithm is quite complex and uses energy clusters in the calorimeters, together with tracker information, in particular a signal in one tracker strip. Quality and isolation criteria help to increase the τ tagging efficiency, by rejecting QCD jets, which might mimic a signal; τ particles that decay leptonically are not considered in CMS because of the overwhelming background coming from prompt muon and electron production;

- Photon (γ): a photon appears in CMS as an amount of energy in ECAL. A prompt photon is a photon which is produced at the primary vertex, and not emitted, for example, via electron conversion or decay of other particles, in general. The detection of a prompt photon requires the presence of a wide energy deposit in many ECAL towers, due to its shower; this energy spread is collected in the so-called "electromagnetic supercluster". The supercluster has to be not matched to any tracks in the tracker and isolated with respect to other energy signals in the calorimeter; furthermore, an upper threshold is also set to the energy present in the hadronic calorimeter in the photon direction;

- Missing transverse energy ($E_{\mathrm{T}}^{\mathrm{miss}}$): since the initial state of hadronic collisions at LHC has no transverse components, $E_{\mathrm{T}}^{\mathrm{miss}}$ should be equal to 0 because of conservation laws, in a completely hermetic and ideal experiment, namely an experiment which is able to measure every particle with infinite resolution. In real life, this is not true and the presence of neutrinos and detector resolution effects contribute to give a certain amount of $E_{\mathrm{T}}^{\mathrm{miss}}$. To measure it, all the energy deposits in the calorimeter acceptance are measured: in particular, $E_{\mathrm{T}}^{\mathrm{miss}} = -\sum_i E_{\mathrm{T}}^{\mathrm{i}}$, where i refers to each energy cluster in the event.

5.1.1 Jet reconstruction

Jets in CMS appear as energy deposits in both ECAL and HCAL in the same region, together with tracks pointing to the same direction. Hence, the PF jet algorithm works by extrapolating the tracks through the calorimeters: if they fall within the boundaries of one or several clusters, the clusters are associated to the track. The set of a cluster and a track constitutes a charged hadron; it might also be possible that a track is associated to more than one cluster. As tracks and clusters are associated, they are removed from the rest of the algorithm. The identification of electrons and muons inside or outside a jet is also possible. Once all tracks are treated, the remaining clusters in ECAL are treated as photons, and the ones in HCAL are identified as neutral hadrons.

Other types of reconstruction are also used in CMS and are listed below:

- Calo-Jets: jets are obtained only by clustering the ECAL and HCAL deposits;
- Jet Plus Track (JPT): jets are reconstructed by energy deposits in ECAL and HCAL but the calorimetric energy value is corrected by using the transverse momentum measured in the tracker for the charged particles in the jet; this reconstruction algorithm differs from the PF technique, since information from the different subdetectors is just merged and not combined for a detailed particle identification, as done in the PF;
- Track-Jets: jets reconstructed from tracks of charged particles, independent of calorimetric information.

Since PF jets use the totality of the available information from the subdetectors, while the others are reconstructed with only a part of it (namely the one measured either with the tracker or with the calorimeters), the performance of the PF algorithm is much better [187]. Resolution measurements have been performed for PF and Calo-Jets in the simulation: after matching the corresponding jets at generator and detector level, the differences in p_{T}, η and ϕ have been evaluated between the two. As shown in Figure 5.2, PF jets achieve a p_{T} resolution up to three times better than Calo-Jets, mostly in the low p_{T} region, and much better η and ϕ resolution also for jets with p_{T} up to 100 GeV. In particular, a PF jet has a p_{T} resolution of 15, 10 and 5% respectively at 20, 100 and 500 GeV, very similar in the barrel and in the endcap regions. The η and ϕ resolutions stay at values of 0.02-0.03 over the whole phase space. The matching and mismatching efficiencies have been also measured for PF and Calo-Jets from the simulation: this implies the measurement of jets at detector level which are associated to generator level jets, in order to see how many measured jets are effectively produced by a real jet at the generator level. The matching is performed by looking for a generator level jet around the detector jet through a cone algorithm, defined by $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2} = 0.1$. When the two jets are found within this cone, the jets are matched, otherwise not. Figure 5.3 shows the matching and mismatching efficiencies as a function of jet p_{T} in barrel and endcap regions. While for jets with $p_{\mathrm{T}} >$ 60-80 GeV, the matching for PF jets is almost 100%, at lower values, mismatching effects start to appear, greater in the endcap region. The performance of Calo-Jets is much worse than for PF jets in the low p_{T} region, reason why PF jets are preferred for physics analyses.

Considering the described jet performance, measurements of PF jets are considered reliable for p_{T} values down to 20 GeV in the full η acceptance range. In case the p_{T} threshold needs to be decreased, a choice of Track-Jets would be more appropriate, since also the tracks of very small p_{T}, which do not hit the calorimeters due to the curvature in the magnetic field, would be considered and would improve the reconstruction performance.

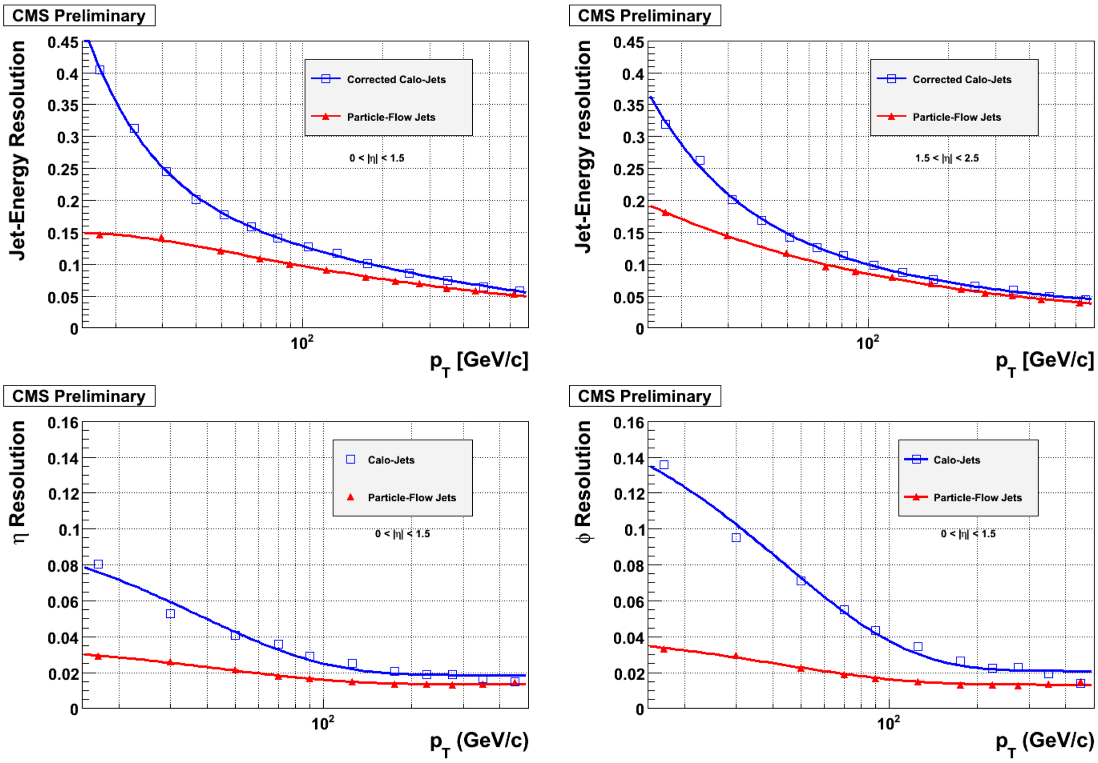


Figure 5.2: Jet resolution as a function of the jet p_T (top) in the barrel (left) and endcap (right) regions and as a function of the jet η (bottom left) and ϕ (bottom right) [187]: results for PF jets and CALO jets are shown in the plots.

Spectra for jets are generally measured in a differential way: this means that the cross section is measured as a function of jet observables. The most common ones are jet p_{T}, η and ϕ. The cross section as a function of p_{T} is rapidly decreasing for increasing p_{T}: the decrease can be parametrized as a power law with an exponent between -3 and -4. A flat distribution is observed as a function of jet ϕ, due to the perfect symmetry of the collision in the transverse plane, while the cross section as a function of η is rather flat (for $p_T > 20$ GeV) in the more central region ($|\eta| < 3$) but starts to decrease for higher pseudorapidities, due to kinematical effects. Note that jets coming from pile-up events contribute mainly in the low p_T region up to 20-40 GeV and in the central region [176].

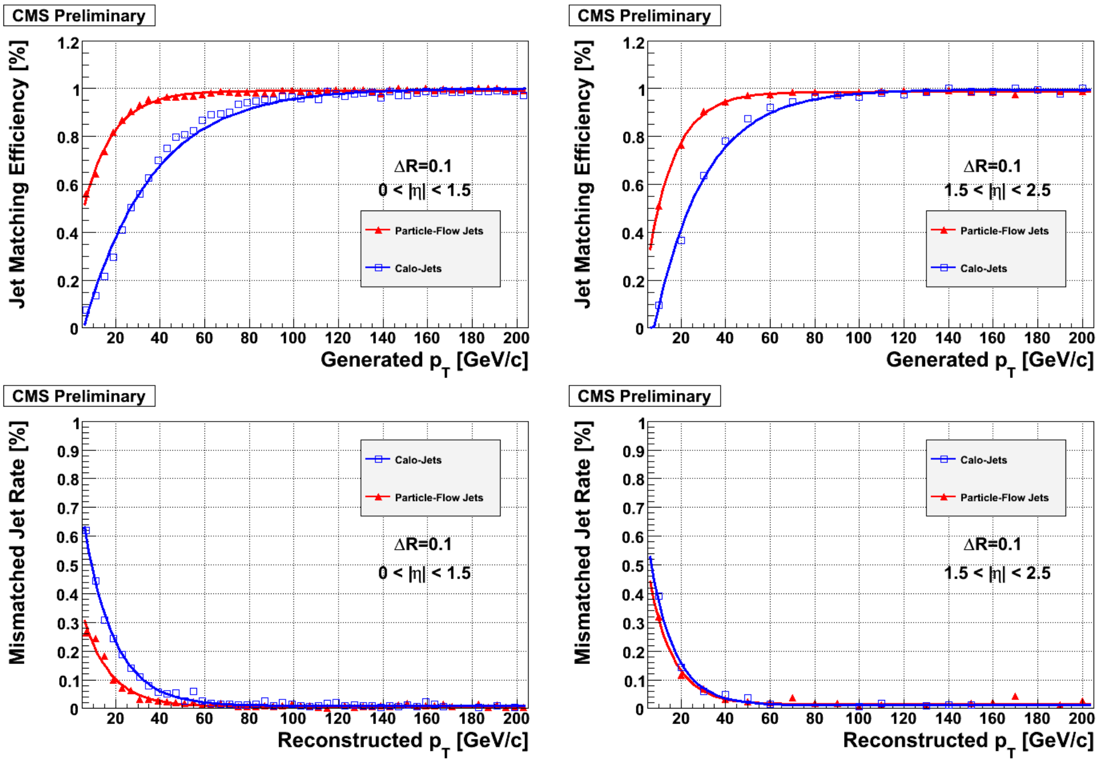


Figure 5.3: Jet matching (top) and mismatching (bottom) efficiency as a function of the jet p_T in the barrel (left) and endcap (right) regions [187]: results for PF jets and CALO jets are shown in the plots. The matching between jets at detector and generator level has been performed with a η-ϕ cone algorithm with width ΔR = 0.1.

5.1.1.1 Jet clustering algorithms

As soon as all the particles have been correctly reconstructed and identified, they need to be grouped in order to be clustered in a jet. As described in Chapter 2, different algorithms are available to do so. In CMS, the anti-k_T algorithm is generally used with a cone width of 0.5 or 0.7. In the analyses of this thesis, a value of 0.5 has been chosen. It has been found that it is a good compromise between the amount of energy of the jet itself, which stays inside the cone and the amount of external energy, coming from pile-up or overlap with other jets, which are accidentally included in it. It is evident that too small cone widths would exclude some of the particles coming from the shower of the partons, while a too big cone width would include more external energy.

After clustering, it is interesting also to investigate which particles constitute the jet itself. These studies have been performed by CMS [188, 189] and results are shown in Figure 5.4. In a very simplicistic approximation, a jet might be thought as composed at the end, only by pions: in this picture, because of charge symmetry, it would result that two thirds of the components would be charged pions and one third neutral pions. The reality is a bit more complicated, since heavier mesons and hadrons should be accounted for, as well as particle decays. Hence, a jet is mainly composed by charged hadrons ($\sim$ 66%), photons ($\sim$ 20-25%, originating from π^0 decays), neutral hadrons ($\sim$ 8-10%) and electrons and muons ($\sim$ 1%, arising from hadron decays). The jet composition does not change much as a function of the jet p_T, while the measured η constituents are different because, in the forward region, CMS can only use calorimeter information, and particle identification is not possible without the tracker measurements. Thus, in the forward region, the main part is composed by hadronic energy deposits in HF, while a very little amount of energy is identified as an electromagnetic component from HF signals. The discrimination between hadronic and electromagnetic energy deposits in HF, relies on the detection of the different profiles of the showers, produced inside the detector.

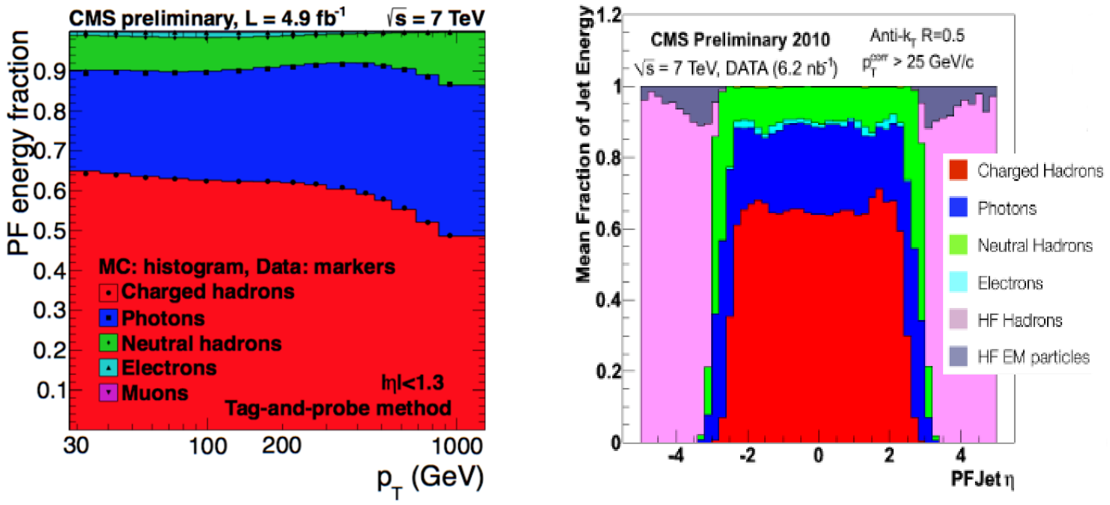


Figure 5.4: Particle composition of anti-k_T PF jets clustered with ΔR = 0.5, as a function of p_T (left) [188] and η (right) [189].

5.1.1.2 Jet energy correction

The energy of a PF jet, obtained after clustering, is not yet optimal for many reasons: for instance, the calorimeter response to particles is not linear or there might be pile-up contamination or other detector-related effects which need to be taken into account. This is the reason why a set of jet energy corrections is applied to properly map the measured jet energy as precisely as possible to the true value of the initial parton which originates the jet. These jet energy corrections are organized in a factorized approach, where each level of correction is independent of each other and takes care of a different effect. In CMS, several correction levels are available and used in physics analyses: they are applied sequentially (the output of each step is the input to the next) and with fixed order, and consist of a scaling of the jet four momentum with a factor which depends on various jet related quantities.

In the analyses, described in this thesis, three levels of corrections are used:

- Level 1 (L1): the goal of this correction is to remove the energy coming from pile-up events. The estimation of the pile-up contribution is performed in an average way: the average energy deposited in the calorimeter is evaluated per unit area in pile-up events and then removed from the energy inside the area of the measured jet [190];

- Level 2 (L2): this corrects for the non-uniformity of the detector response as a function of pseudorapidity; at the end of this correction, the jet response is flat in η. Scale factors have been measured by using the p_{T} balance in dijet events, where one jet is measured in the central region ($|\eta| < 1.3$) as a reference, and the other one in a different η region up to the very forward one ($|\eta| < 5.2$) [191];
- Level 3 (L3): this level makes the jet energy response flat in p_{T} by correcting for the non-uniformity of the calorimeter detection as a function of the jet transverse momentum. The corresponding scale factors have been measured through a data-driven method by exploiting the p_{T} balance between a well measured and calibrated object (Z or γ) and a jet [192].

These corrections need to be applied to both simulation and data. An additional jet correction is applied to the data in order to take into account small remaining discrepancies between data and MC, up to 10% depending on η. This additional correction is called the “residual” correction and it helps to get an optimal description of the data by the simulation. A schematic sketch which summarizes all the correction levels applied, is represented in Figure 5.5. The corresponding factors are given officially by the CMS collaboration for each level.

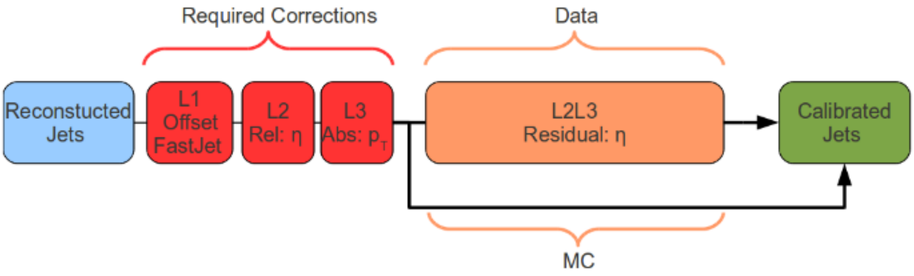


Figure 5.5: Sketch of the jet energy correction levels at CMS for data and simulation [193].

Quality jet criteria are also applied in order to reject fake jets, arising from detector noise or misreconstructed particles. For the analyses described in the thesis, a tight selection has been applied to the jets and this implies requirements in the neutral and charged hadron fraction as follows:

- for jets in $|\eta| > 2.4$
    - neutral hadron fraction < 0.90
    - neutral electromagnetic fraction < 0.90
    - number of constituents > 0
- for jets in $|\eta| < 2.4$, since they can profit of the presence of the tracker, additional cuts are placed for:
    - charged hadron fraction > 0
    - charged electromagnetic fraction < 0.99

- charged multiplicity > 0

This selection requires that a non-zero percentage of hadronic and electromagnetic components is present, as well as a certain number of charged and neutral constituents. This selection achieves an excellent selection efficiency and a negligible fake rate, respectively of about 99.9% and $< 0.5\%$ for jets with $p_{\mathrm{T}} > 20$ GeV [194].

5.1.2 B-jet tagging

For many analyses performed in CMS, as well as the ones described in this thesis, the identification and the detection of b-jets in the final state become crucial. The algorithm which is implemented at the detector level in order to discriminate b-jets from the multitude of light jets, is known as "b-tagging". Good and reliable b-tagging is of particular interest for all analyses involving top quarks, which decay into b-quarks, the Higgs boson decaying b-hadronically or for any b-jet cross section measurement. The b-tagging algorithm is based on the properties of the b-quarks: as already mentioned in Chapter 1, when a b-quark is produced in the hard scattering or emitted through hard radiation, it immediately hadronizes and travels a macroscopic average mean path before the decay. The lifetime τ_{B} of B hadrons is of the order of 1 ps (10^{-12} s), and the average distance, before it decays, is:

$$\lambda_{\mathrm{B}} = c\tau_{\mathrm{B}}\gamma \sim 500\gamma \ \mu m \sim 2 \text{ mm} \tag{5.1}$$

assuming a $B^{\pm}$ ($\mathrm{m}_{\mathrm{B}^{\pm}} \sim 5.2$ GeV) as B-hadron, with an energy of the order of 20 GeV. This distance is measurable thanks to the resolution provided by the CMS pixel tracker. A B-hadron decay produces, on average, five charged particles. An event, where a B-hadron is produced at high energy, appears at the detector level as a collection of tracks, mostly pointing to the interaction point, namely to the primary vertex corresponding to the hard scattering, but with a subset of them coming from a different point, corresponding to the decay vertex of the B-hadron, hence called secondary vertex. An event of this type is shown in Figure 5.6. In the picture, well visible are the tracks produced by light quarks in the hard scattering and pointing to the primary vertex, and the tracks coming from the secondary vertex generated by the decay of the B-hadron after travelling a distance L_{XY}. An important quantity is also the impact parameter d_0, defined for each track detected in the event. The impact parameter is defined as the distance between the track and the primary vertex at the point of closest approach. Tracks produced by the B-hadron decay have a large impact parameter, while this is smaller for tracks produced by the hard scattering.

Note that a secondary vertex produced by the decay of a B-hadron looks differently from a vertex created by a pile-up interaction: while the former presents a spatial displacement in the transverse plane with respect to the primary vertex, the latter lies exactly along the beam line and the displacement from the primary interaction is only along the longitudinal direction.

The CMS experiment has developed several b-tagging algorithms [196, 197, 198, 199]. They are based on the forementioned b-quark and B-hadron properties and translate into requests at the detector level. Two quantities are important to define the performance of a b-tagging algorithm:

- b-tagging efficiency: it is the efficiency of tagging a true b-jet;
- mistag fraction: it is the fraction of non b-jets which are misidentified as b-jet and tagged as such.

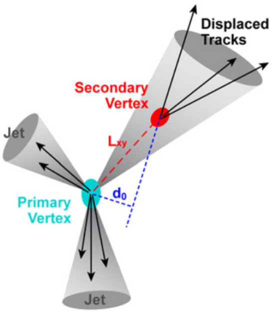


Figure 5.6: Sketch of a hard scattering with production of a b-quark [195]: two light jets are originated from the primary vertex (cyan area), together with a b-quark which, after hadronization, travels a length L_{XY} *(red dashed line) in the detector. The decay of the* B*-hadron produces a secondary vertex (red area) where tracks are produced from. The displaced tracks, pictured in the figure, are the input to the b-jet reconstruction.*

The main algorithms available, which can be used in a CMS analysis are the following:

- "Track Counting" (TC): it is the simplest algorithm, which calculates the impact parameter significance, namely the quantity $d_0/\sigma_{\rm IP}$, being $\sigma_{\rm IP}$ the spatial resolution in the measurement of the impact parameter, for all the tracks, which are then ordered by decreasing significance. The b-tag discriminator is defined as the significance of the $N^{\rm th}$ track, with $N = 2$ for a high efficiency discriminator (TCHE), or $N = 3$ for a high purity one (TCHP);

- "Soft Muon" or "Soft Electron": this algorithm uses the fact that semi-leptonic B-hadron decays often produce leptons with high relative $p_{\rm T}$ with respect to the jet axis. This information, together with the impact parameter significance, is included in a neural network whose output defines the discriminator;

- "Simple Secondary Vertex" (SSV): the discriminator for this algorithm is built through the reconstruction of the decay vertex by using a complex vertex finder and the evaluation of the decay length significance, which is the ratio of the decay length to its estimated uncertainty;

- "Combined Secondary Vertex" (CSV): this algorithm uses all the variables related to decay length significance, secondary vertex and jet kinematics. These variables are combined using a likelihood ratio technique to compute the b-tag discriminator.

For every algorithm, official working points are provided [200], which are values corresponding to cuts to be applied on the specific discriminator. A working point is defined by the level of achieved mistag rate. The different working points with the corresponding performance are listed below:

- Loose: it is the discriminator value which guarantees a 10% mistag rate and b-tagging efficiency of the order of 80% for the CSV algorithm at 7 TeV;

- Medium: for this, the mistag rate is 1% and the b-tagging efficiency decreases to values between 60–65% for the CSV algorithm at 7 TeV;

- Tight: it ensures a very low mistag rate, equal to 0.1%, with a b-tagging efficiency between 45–50% for the CSV algorithm at 7 TeV.

In the analyses described in this thesis, the CSV algorithm has been used with a loose working point: this ensures a very reliable and efficient selection of b-jets in the final state and maximizes the statistical accuracy available for the measurement with an acceptable purity.

5.1.3 Definition of the jet flavour

A further issue, which needs to be addressed in the simulation, deals with the definition of the jet flavour. In fact, a jet contains inside its cone many partons, which generally have different flavours. The point is then to define rules in order to assign a flavour to every jet. The aim of this procedure is to identify the flavour of the parton from which the considered jet has been originated. The assigned flavour should indeed reflect the flavour of the parton which mostly determines the kinematic and the internal structure of the jet. This is performed by using MC truth information of the partons inside the jet cone (in particular, the status code of the particles) but there is no unambiguous way to do so. Hence, in CMS, different definitions have been developed. They are described in [201] and listed in the following:

- Algorithmic definition: the flavour of the heaviest parton is assigned as the flavour of the jet. In case a b- or a c- quark are present inside the cone, then respectively the b- or the c-flavour is then chosen. In case partons of both flavours are inside the cone, the b-flavour is assigned, due to its higher mass;
- Energetic definition: the flavour of a considered jet is determined by the flavour of the most energetic parton, contained in the cone;
- Physics definition: the jets are matched to the “initial” parton from the primary physics process; secondary jets from gluon splitting are thus labelled as gluon jets.

The choice of a definition algorithm, with respect to another, depends on the specific goals of the analysis. For instance, for top analyses, the physics definition is preferred, since it suppresses the background from b-jets coming from gluon splitting, while in QCD analyses or channels where the b- or c-jet production is relevant, the other algorithms should be used. Hence, in the analyses described in this thesis, the algorithmic definition has been used for b-jets both for detector and generator levels. Thus, in the following chapters, when true flavour of jets is mentioned, the jet flavour obtained with the algorithmic definition is meant.

Chapter 6

Event selection

In this Chapter, the selection of events at detector level is described in full detail. In particular, two different analyses are considered:

- four-jet (**4j**), which deals with the selection of an exclusive scenario with four jets in the final state;
- two *b*- and two other jet (**2b2j**), which selects an inclusive scenario with at least two *b*- and at least two other jets in the final state.

The two analyses are referred to with these names, hereafter, and the details are described in the following Sections.

6.1 Data samples

Data recorded with the CMS detector have been used for the two analyses. They correspond to the data of the first year of data taking, 2010. The instantaneous luminosity of the accelerator was relatively low, with a maximum of about 10^{32} cm^{-2}s^{-1}, and the mean value of pile-up interactions varied between 1.1 and 1.8 throughout the whole period. The understanding of the pile-up is crucial for jet analyses. In fact, additional jets may easily arise from pile-up events and might determine the selection or rejection of some events, depending on the applied selection criteria, or alter measurements of jet correlations. Hence, a correct and accurate modelling of the pile-up is very important in the simulation, in order to identify, remove and correct for the events with selected jets coming from overlapping interactions.

In the whole collection of data recorded in 2010, small subsamples have been considered for the two analyses and they are listed in Table 6.1 and 6.2, respectively for the 4j and the 2b2j channels. For the first one, the whole amount of 2010 data, corresponding to an integrated luminosity of 36 pb^{-1}, has been analyzed; for the second one, the examined sample has been reduced, down to an integrated luminosity of 3 pb^{-1}, because of an asymmetric behaviour of the low jet p_{T} trigger, observed as a function of η for part of the data samples [202]. Only good data runs, listed in official CMS JavaSpcript Object Notation (JSON) files, are considered for the analyses. A JSON file contains the list of good runs, relative to each specific period of data taking, and it is the output of the DQM procedure (see Chapter 3).

A physics analysis is usually performed in several steps. Starting from very large data files, a preliminary filter is applied in order to save and store only the information useful for the specific selection: this operation produces the so-called Ntuples which can be analyzed much faster than the complete data files. The creation of Ntuples is performed through submission of GRID [173] jobs implemented in the CMSSW software. The code used for the creation of the Ntuples has been built by merging and synchronizing the usual CMS code used for jet analyses with the one used for the selection of flavoured (bottom and charm) jets. After this step, the whole analysis is performed by further filtering and selecting the content of the Ntuples inside the ROOT [171] software.

Data sample	Run range	Trigger	Integrated luminosity (pb^{-1})
JETMET	141950-144114	HLT_Jet30U	0.192895
/JetMET/Run2010A-	141950-144114	HLT_Jet50U	2.896
Apr21ReRecov1/AOD			
JETMETTAU	135821-141887	HLT_Jet30U	0.117223
/JetMETTau/Run2010A	135821-141887	HLT_Jet50U	0.278789
-Apr21ReReco-v1/AOD			
JET	146240-149711	HLT_Jet30U	0.026783
/Jet/Run2010B	146240-149711	HLT_Jet50U	0.239874
-Apr21ReRecov1/AOD			

Table 6.1: List of data samples analyzed for the 4j channel. The path in the official CMS storage, the run range, the triggers and the integrated luminosity are also provided for each subsample.

Data sample	Run range	Trigger	Integrated luminosity (pb^{-1})
JETMETTAU	135821-141887	HLT_Jet15U	0.013927
/JetMETTau/Run2010A			
-Apr21ReReco-v1/AOD			
JETMET	141950-144114	HLT_Jet30U	0.192895
/JetMET/Run2010A-	141950-144114	HLT_Jet50U	2.896
-Apr21ReReco-v1/AOD			

Table 6.2: List of data samples analyzed for the 2b2j channel. The path in the official CMS storage, the run range, the triggers and the integrated luminosity are also provided for each subsample.

6.2 Trigger

As explained in Chapter 3, in order to reduce the amount of recorded events to a sustainable rate, CMS implements a trigger system organized in two levels. Since the analyses, described in this thesis, deal with jets in the final state, jet triggers are used to select events in data. A jet at L1 trigger is identified by a coarse energy cluster in both hadronic and electromagnetic calorimeters, corresponding to a certain amount of transverse energy. A more elaborate but still very fast algorithm, the “jet finder”, is then implemented on this cluster but with a finer segmentation in order to select the raw object for the HLT trigger: the algorithm makes use of a cone size in order to cluster in a primitive jet the calorimeter towers whose energy is larger than the seed threshold. If the primitive HLT jet has an energy above the threshold set by the trigger, the event is selected and the collection of recorded data is saved and stored in streams. The acceptance of the jet triggers extends up to the full coverage of the hadronic calorimeter, $|\eta| \sim 5.2$. The jet triggers used for the analyses are identified by the name "HLT_JetXU", where X stands for the energy threshold, expressed in GeV, set for the HLT jets. For this work, triggers with values of X equal to 15, 30 and 50 have been used. The L1 and HLT thresholds for each of the triggers are listed in Table 6.3. For instance, an event is selected by the HLT_Jet15U trigger, in case a calorimeter cluster exceeds the energy of 6 GeV in $|\eta| < 5.2$ and the primitive jet, clustered with the jet finder algorithm, has an energy greater than 15 GeV. Note that the primitive jet needs a more accurate and complicated reconstruction (see Chapter 5) with additional corrections, before being used for any analysis.

In order to select events for the four-jet analyses, the exclusive division method [203] has been used. This consists of dividing the phase space in independent regions as a function of the leading jet p_{T}; in each region, only one trigger is used and every region has no overlap with the others, in order to avoid any double counting. The division has been organized for the four-jet analyses in the following way:

- $20 \leq p_{\mathrm{T}}^{\mathrm{leading}} < 50$ GeV → HLT_Jet15U
- $50 \leq p_{\mathrm{T}}^{\mathrm{leading}} < 140$ GeV → HLT_Jet30U
- $p_{\mathrm{T}}^{\mathrm{leading}} \geq 140$ GeV → HLT_Jet50U

where the specified triggers are the ones used in each region. Note that for the 4j analysis, the first trigger region is not considered. The trigger regions are schematically represented in Figure 6.1 for the 4j and the 2b2j analyses, which differ from each other on the leading jet p_{T} threshold. The choice of these regions with the corresponding triggers is the result of a compromise between sufficiently high statistics for each of them and quite high and well understood trigger efficiency (see Section 6.2.1).

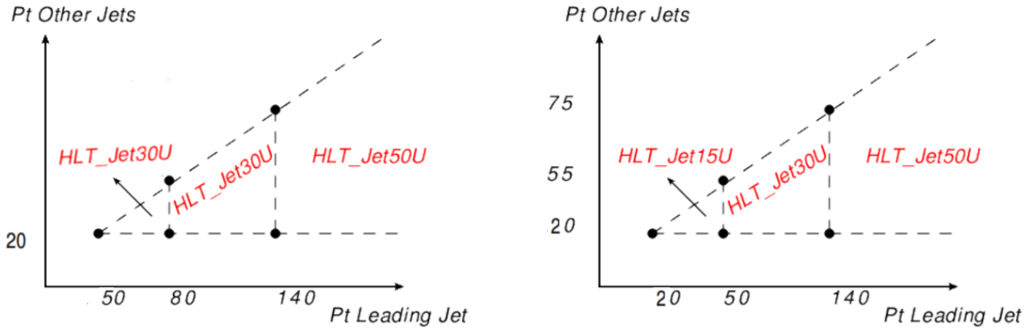


Figure 6.1: Sketch of the trigger regions used for the exclusive division method, defined by the leading jet in the events for the 4j (left) and the 2b2j (right) analyses: for each region, the used trigger is specified along with the p_T threshold of the leading jet, shown in the x-axis.

6.2.1 Trigger efficiency correction

When requiring the presence of a trigger signal, some interesting events might be left out in case the trigger itself has not fired. This is referred to as "trigger inefficiency". Trigger inefficiencies may be due to dead regions of the subdetector, trigger objects close to the thresholds in transverse momentum or pseudorapidity which are not well reconstructed, or temporary problems during data runs. A correction needs to be applied in the regions of the phase space where the triggers are not fully efficient. The trigger efficiency has been measured in data in two different ways which have given compatible results. The two methods are:

- Cross section ratios: differential jet cross sections, as a function of η and p_{T}, are measured separately when the trigger under examination, $\mathrm{trig}_{\mathrm{i}}$, and a reference trigger, $\mathrm{trig}_{\mathrm{ref}}$, have fired. The reference trigger needs to be fully efficient, in the considered region of the phase space and it is normally a jet trigger with lower p_{T} threshold or a MB trigger. The ratio of the two differential cross sections constitutes the measured trigger efficiency, ϵ_{trig}, as defined by the equation:

  $$\epsilon_{\mathrm{trig}} = \frac{\left(\frac{d\sigma}{dO}\right)^{\mathrm{trig}_{\mathrm{i}}}}{\left(\frac{d\sigma}{dO}\right)^{\mathrm{trig}_{\mathrm{ref}}}} \tag{6.1}$$

  with O, any kinematical jet observable (η, p_{T}, etc.). For instance, in order to measure the efficiency of HLT_Jet50U, HLT_Jet30U has been used, while for HLT_Jet30U, cross sections measured with HLT_Jet15U have been compared.

- Trigger emulation: the trigger decision is emulated in data by using the trigger elements of a reference trigger, $\mathrm{trig}_{\mathrm{ref}}$. In order to reproduce the trigger decision, two objects are needed, one for each of the two trigger levels (see Chapter 3): they are referred to as “L1” and “HLT” objects. For jet triggers, the L1 object consists of a broad energy deposit in HCAL and ECAL by using a coarse segmentation. Information obtained with the full calorimetric granularity is added to the L1 object to produce the HLT one. In the assumption that the reference trigger is fully efficient, in the considered region of the phase space, the emulation method is expressed by the following equation:

  $$\epsilon_{\mathrm{trig}_{\mathrm{i}}} = \frac{\mathrm{InclusiveRecoJet\_}O(\mathrm{trig}_{\mathrm{ref}} + \mathrm{L1Object\_}p_{\mathrm{T}} > Z + \mathrm{HLTObject\_}p_{\mathrm{T}} > Y)}{\mathrm{InclusiveRecoJet\_}O(\mathrm{trig}_{\mathrm{ref}})} \tag{6.2}$$

  where Y indicates the p_{T} threshold of $\mathrm{trig}_{\mathrm{i}}$, and with Z, the threshold of the L1 object is identified. The quantity O is again any observable for which the trigger efficiency has to be measured. The denominator corresponds to the number of events for which the emulator trigger path $\mathrm{trig}_{\mathrm{ref}}$ has fired. The numerator is the number of events for which $\mathrm{trig}_{\mathrm{ref}}$ has fired and the p_{T} of the HLTObject corresponding to the trigger path $\mathrm{trig}_{\mathrm{i}}$ is $> Y$. For example, in order to obtain the turn on curve for HLT_Jet30U, the HLT path threshold, used for HLT_Jet15U, is chosen: the p_{T} cut on L1Object corresponding to this trigger path is 20 GeV. The complete list of measured triggers with the corresponding reference triggers are listed in Table 6.3, along with the values of the L1 and HLT thresholds.

The second method is preferred for efficiency measurements because it achieves a higher statistical accuracy for four-jet scenarios and it does not need any luminosity information of the triggers, like when evaluating cross section ratios. Hence, in the following only results of the trigger emulation method are considered and used. The trigger emulation method has been also used in previous CMS analyses [93, 204]. In order to identify the regions of the phase space where a correction needs to be applied, the efficiency, as a function of the leading jet p_{T} selected in $|\eta| < 4.7$, has been measured for the three triggers under study. The results are shown in Figure 6.2, for inclusive jet scenarios, as a function of the leading jet p_{T}, and different data samples. The trigger efficiencies show a turn-on curve, with a rising part, where the trigger is partly inefficient, until a plateaux region, corresponding to the region of full efficiency of the trigger. From these results, the jet p_{T} threshold, from which each trigger starts to become fully efficient, can be identified and is listed in Table 6.3. Moreover, as shown in Figure 6.2, trigger efficiencies, measured in data samples corresponding to different periods of data taking, are compatible among each other.

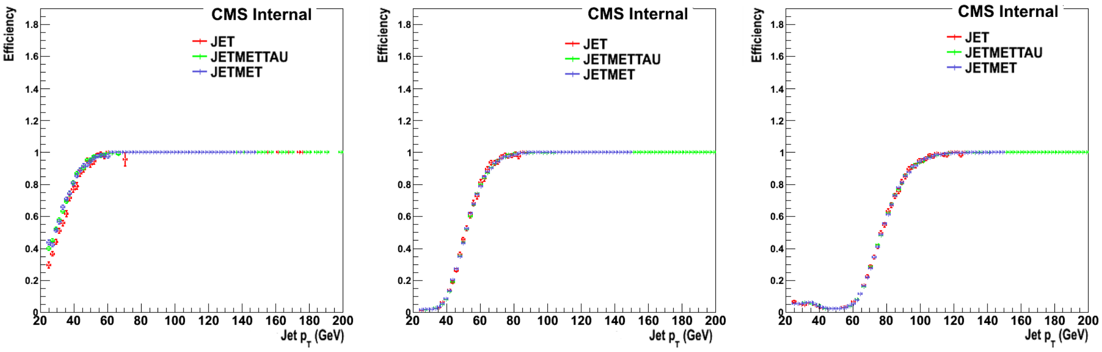


Figure 6.2: Trigger efficiency measurement as a function of the leading jet p_T selected in $|\eta| < 4.7$ for the HLT_Jet30U (center) and HLT_Jet50U (right) triggers: the turn-on curves are measured for different data samples recorded in 2010.

Trigger	Reference Trigger	L1 Threshold	HLT Threshold	Full efficiency threshold
HLT_Jet15U	HLT_L1Jet6U	6 GeV	15 GeV	50 GeV
HLT_Jet30U	HLT_Jet15U	20 GeV	30 GeV	80 GeV
HLT_Jet50U	HLT_Jet30U	30 GeV	50 GeV	140 GeV

Table 6.3: List of the triggers used in the four-jet analyses with corresponding reference triggers, and p_T threshold at L1 and HLT. The p_T threshold corresponding to the starting point of full efficiency is also specified for each trigger.

For the interest of these analyses, a trigger efficiency correction is needed for the following triggers and regions:

- HLT_Jet30U → $50 <$ leading jet $p_{\mathrm{T}} < 80$ GeV

for the 4j analysis, while for the 2b2j analysis:

- HLT_Jet15U → $20 <$ leading jet $p_{\mathrm{T}} < 50$ GeV
- HLT_Jet30U → $50 <$ leading jet $p_{\mathrm{T}} < 80$ GeV

The emulation method is applied in the regions where the triggers are not fully efficient for the two selections, 4j and 2b2j, listed in Tables 6.6 and 6.7. A p_{T}-η dependent efficiency correction is applied as a function of the selected leading jet observables, as described in the following. The curves are measured separately as a function of the leading jet p_{T} and leading jet η. The curve as a function of p_{T} is fitted with a polynomial function: a 3-degree polynomial is used for the HLT_Jet15U trigger and a 4-degree polynomial is used for HLT_Jet30U. They can well reproduce the rising part of the turn-on curves. Weights corresponding to the p_{T}-dependent fit are applied to the same events and the efficiency curve as a function of η is again measured. In order to correct for the η dependence, bin-by-bin weights are applied: after this additional correction, the efficiency curve is flat at 1 by definition. By applying the p_{T}- and η-dependent correction to the curve as a function of p_{T}, a similar behaviour, flat at unity, is observed; this shows that the method, used for the removal of trigger inefficiencies, is not affected by any bias. Figures 6.3 and 6.4 show the turn-on curves as a function of p_T and η, before any correction (left), after the p_T-dependent (center) and the p_T-η dependent (right) corrections, for, respectively, HLT_Jet15U and HLT_Jet30U, in case of the 2b2j selection. Note that Figures 6.3(top and bottom right) and 6.4(top and bottom right), representing the curves after full correction, show a flat distribution around 1 over the whole phase space; this is the proof of the goodness of the method which allows to achieve full efficiency after correction.

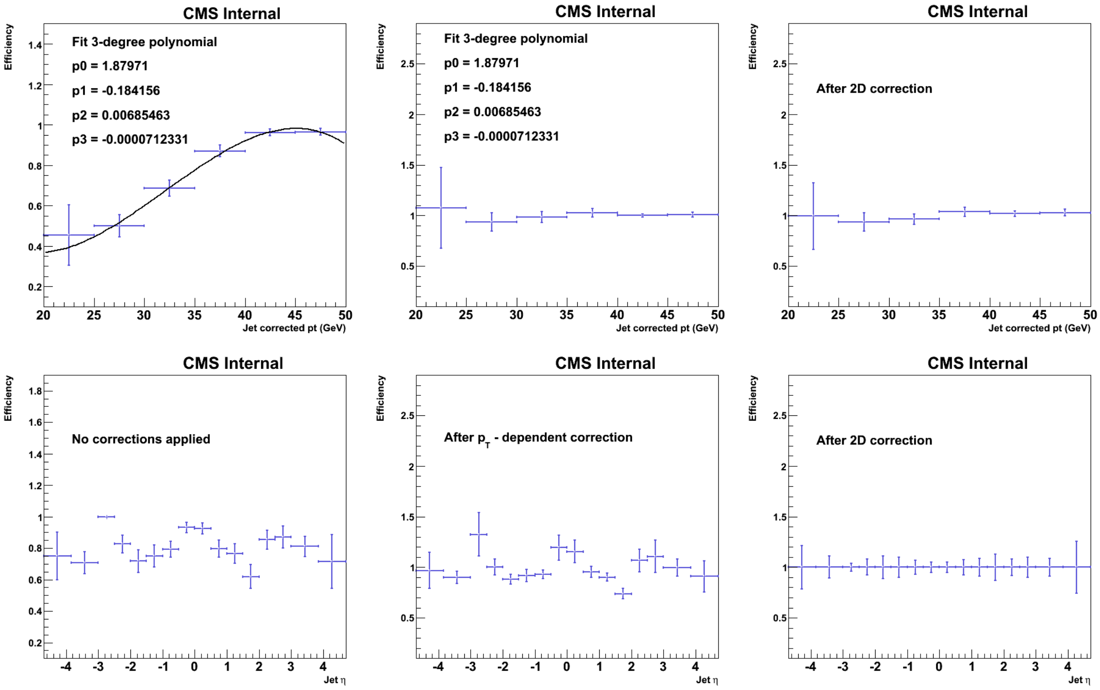


Figure 6.3: Trigger efficiency measurement and correction for the HLT_Jet15U trigger as a function of leading jet p_T (top row) and η (bottom row): the trigger efficiency curves are shown before any correction (left), after the p_T-dependent correction (center) and after the p_T- and η-dependent correction (right).

The trigger efficiency correction applied to the 4j selection, results in the same conclusion; its effect is anyway less relevant, because the inefficient region is smaller than the one for the 2b2j analysis. The trigger efficiency correction is hereafter applied to the measured data distributions of the considered four-jet scenarios.

6.3 Monte Carlo generated samples

In order to interpret and correct the measurements, a reliable detector simulation has to be used. For this analysis, two different samples, produced by the central CMS generator group, have been generated with full detector simulation. The two samples use a ME which simulates QCD events with $\hat{p}_T > 15$ GeV, generated with the PYTHIA6 and the HERWIG++ event generators. The samples apply a p_T weighting of the ME to generate a flat distribution as a function of $\hat{p}_T$, in order to obtain a larger statistical accuracy in the region of high transverse momenta. The UE simulation is provided by two of the most up-to-date tunes available: the PYTHIA6 sample uses the tune Z2* [180], while the one generated with HERWIG is simulated with the tune UE_EE_3 [205]. They have been considered for both analyses, with the unique difference that for the 2b2j analysis the

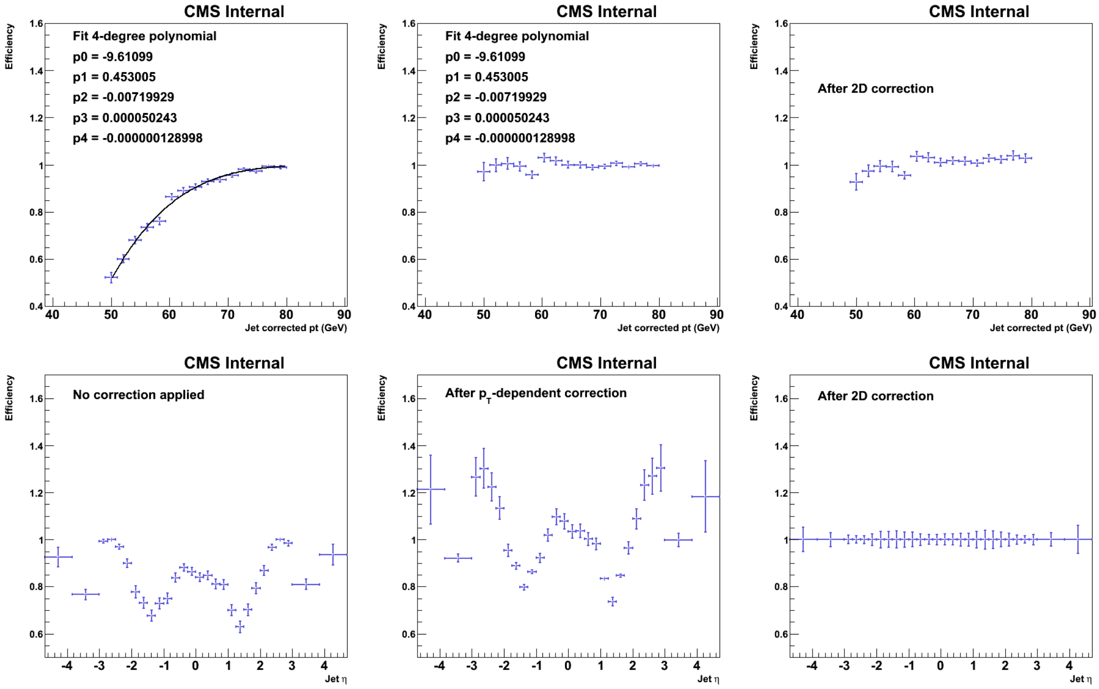


Figure 6.4: Trigger efficiency measurement and correction for the HLT_Jet30U trigger as a function of leading jet p_T (top row) and η (bottom row): the trigger efficiency curves are shown before any correction (left), after the p_T-dependent correction (center) and after the p_T- and η-dependent correction (right).

amount of simulated events has been increased by a factor of six, in order to reach a sufficient statistical accuracy. The details of the MC samples are listed for the two analyses in Tables 6.4 and 6.5. A simulation of the pile-up is included in the MC samples.

Ntuples, produced in the same way as for data, have been created from the MC samples. In addition, they include information of jets and particles at the generator level, referred to as "MC truth information". Particularly important is the true flavour of the jets, namely the flavour of the parton which determines the properties of a jet[1]: this variable is relevant for the analysis with b-jets. No trigger information is used in the MC, since the jet trigger simulation, especially at low p_{T}, is not reliable. Particular attention has been given to match beam properties, as well as detector responses and calibrations, in data and MC. This has been done through the use of specific global tags, when creating the Ntuples, which are used to process data and MC by using the same event reconstruction.

MC sample	Number of Events	Cross section (pb)
PYTHIA6 Tune Z2*	9982400	$2.21 \cdot 10^{10}$
HERWIG++ Tune UE_EE_3	9971200	$2.31 \cdot 10^{10}$

Table 6.4: List of Monte Carlo samples used for the 4j channel. The number of generated events and the total cross section are also provided for each subsample.

---

[1] For the definition of the flavour of a jet, see Chapter 5

MC sample	Number of Events	Cross section (pb)
PYTHIA6 Tune Z2*	59896000	$2.21 \cdot 10^{10}$
HERWIG++ Tune UE_EE_3	59379200	$2.31 \cdot 10^{10}$

Table 6.5: List of Monte Carlo samples used for the 2b2j channel. The number of generated events and the total cross section are also provided for each subsample.

6.3.1 Pile-up treatment

The number of pile-up events, implemented in the simulation, is not the same as the one measured in data. In particular, the mean value of overlapping interactions is much higher in the MC samples[1]: it is around 8-9 pile-up events, while it is below 2 for the data. Hence, the MC pile-up distributions are reweighted to the ones in data. In CMS, there are several procedures to perform that: the most used method [206] is the one which reweights each MC sample according to the instantaneous luminosity recorded by the detector in the data as a function of the number of pile-up events in the simulation. This is mainly addressed to high pile-up scenarios, recorded in 2011 and 2012. For the 2010 data, the instantaneous luminosity is not available, since it was not recorded, and an alternative method needs to be applied. This is called the "iterative method" and it has been used in the two analyses described here.

The iterative method consists of comparing the distributions of good reconstructed primary vertices obtained in data and MC. A primary vertex is identified by a collection of tracks, measured in the tracker with a good fit quality between the hits and compatible with the beam line. The tracks are clustered according to the z-coordinate of their point of closest approach to the beam axis. A primary vertex candidate is obtained through a three-dimensional fit. Primary vertices are retained only if their z-coordinate stays at a distance less than 24 cm from the beam spot. Further quality requirements are then also applied:

- the number of degrees of freedom (NdF) of the fit is required to be greater than 4: NdF is related to the free parameters of the fit and the number of associated tracks;
- fake vertices are discarded: they may be produced by weak decays, secondary interactions with the detector material, or by tracks coming from the beamspot or with poor momentum resolution.

In an ideal case, a good reconstructed primary vertex corresponds to a pp interaction. Since a pile-up event is separated in space and independent of the other interactions occurring in the same collision, one could think, in a first approximation, that the number of pile-up events is equal to the number of reconstructed vertices. Unfortunately, many detector effects spoil this identity: inefficiencies (a true vertex is not reconstructed), resolution issues (two vertices are too close to be resolved separately) and fake reconstructions (a primary vertex not corresponding to a pile-up event is reconstructed as such, because of track misidentification) determine the decrease or increase of the number of reconstructed vertices. This is why an exact correspondence between the number of primary vertices and pile-up interactions, generally, does not hold.

---

[1] In fact, the number of pile-up events in the MC is set to a very high value, in order to have the possibility to match it to a wide range of pile-up scenarios in data.

The bin-by-bin ratios of the primary vertex distributions measured in data and MC are considered for the iterative method. They are applied as weights to the true number of pile-up interactions[1] in the simulation. Provided that exactly each pile-up interaction creates a separated primary vertex, the described procedure would give a perfect agreement between the primary vertex distributions in data and MC after the application of the weights. This is in fact not true, because of the forementioned effects and implies the fact that the reweighting procedure in the simulation as a function of the true number of pile-up interactions needs to be repeated (hence, the name "iterative") several times. In both analyses, the values of the weights start to converge to 1 already after the third iteration and a number of iterations equal to five has been found to be optimal. The normalized cross sections of the good reconstructed primary vertices are shown in Figure 6.5 for data and MC. A ratio equal to 1 along the whole spectrum is observed for distribution obtained in data and MC after reweighting, showing a good agreement between them.

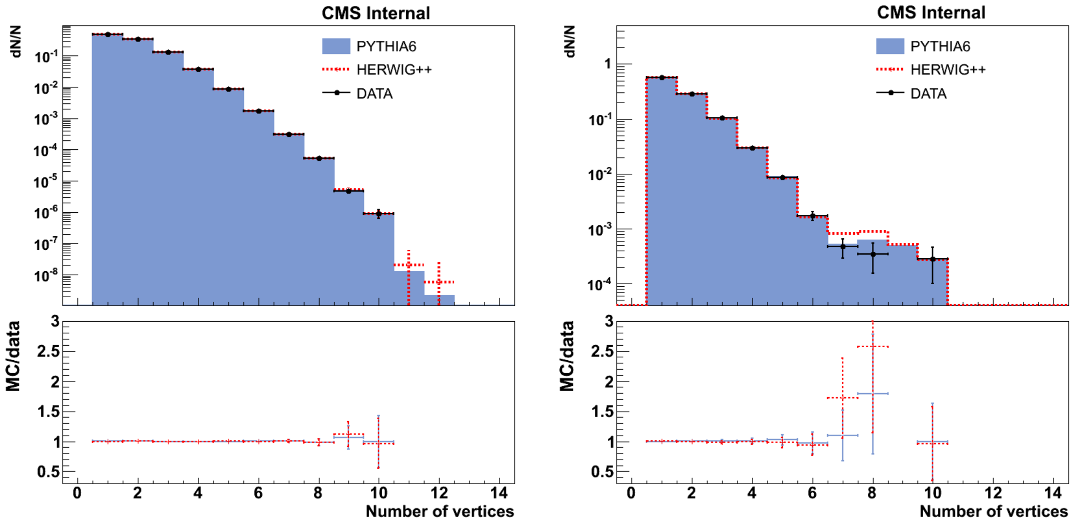


Figure 6.5: Good reconstructed primary vertex measured in data and in the simulation after applying the weights of the iterative method for the 4j (left) and the 2b2j (right) analysis. A number of five iterations has been applied for both MC samples. The lower panel shows the ratios of the theoretical predictions to data.

The final weights, obtained in this way for the pile-up reweighting, have been applied to the simulation in all comparisons at the detector level, which are presented in the following Sections.

6.4 Jet selection

Inclusive jet distributions are measured at detector level and compared to the simulation after applying the pile-up reweighting procedure. This is important in order to understand whether the available Monte Carlo samples are able to reproduce jet data, before going to study more complex scenarios with four jets selected. First of all, the leading jet is selected with $p_T > 20$ GeV in $|\eta| < 4.7$ and compared to the simulation. Results are shown in Figure 6.6 for the absolute differential cross section as a function of p_T and the normalized differential cross sections as a function of jet η and jet ϕ. Note that a trigger efficiency correction has not been applied to the considered spectra,

---

[1] In MC, the true number of pile-up events, namely how many interactions, overlapped to the hard scattering, have been effectively simulated, is indeed available, while in data it is not.

because they refer to inclusive jet measurements and not to four-jet scenarios, as investigated in Section 6.2.1. As expected, the $p_{\rm T}$ spectrum is rapidly decreasing for increasing values and a difference of almost four orders of magnitude between $p_{\rm T} \sim 50$ GeV and $p_{\rm T} \sim 200$ GeV is observed for the cross section. The simulation is able to reproduce the high $p_{\rm T}$ region ($p_{\rm T} > 80$ GeV), while discrepancies are observed for lower $p_{\rm T}$. This is due to trigger inefficiencies in the data. The differential cross sections as a function of η and ϕ are presented after normalizing the bin contents to the total number of events; this removes the normalization problem due to trigger inefficiencies. The η distribution shows a flat behaviour in $|\eta| < 3.5$, and a rapid decrease for higher η values, due to kinematical reasons: when going to the more forward region, jets must have an increasingly higher energy to fulfill the $p_{\rm T}$ requirement and, hence, their cross section is smaller than the one in the central region. The ϕ spectrum is flat over the whole phase space. While the simulation is able to reproduce very well the shape of the ϕ distribution, a slight disagreement is observed when describing the η spectrum. This is again due to a modulation of the trigger efficiency as a function of jet pseudorapidity.

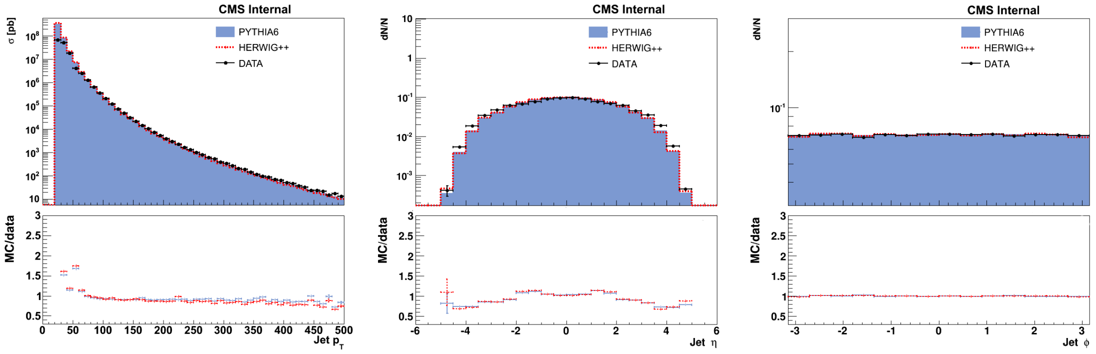


*Figure 6.6: Comparison of data with simulations from MC event generators (*Pythia6 *and* Herwig++*) for the transverse momentum of the leading jet $p_{\rm T}$ (left), η (center), ϕ (right): absolute cross sections are shown for $p_{\rm T}$, while normalized cross sections are shown as a function of η and ϕ. Only statistical uncertainties are shown for the data. Data and simulation are not normalized to the bin width. The lower panel shows the ratios of the theoretical predictions to data.*

After these preliminary distributions, where it is seen that detector level jet data are reasonably well understood and reproduced by the simulation in the regions of full trigger efficiency, the specific selection for the 4j analysis has been set. At least one good primary vertex has been requested for the triggered events and exactly four jets are required in $|\eta| < 4.7$: two jets must have $p_{\rm T} > 50$ GeV, while the threshold for the other two jets is decreased down to 20 GeV. The first two jets are associated and labelled as "hard-jet pair", while the other two jets constitute the "soft-jet pair". Jets are reconstructed by using the PF information and clustered with the anti-$k_{\rm T}$ algorithm, by using a cone size R = 0.5. The tight jet selection is also applied to all jets, in order to remove non-physical jets, arising from detector noise. Exactly the same selection is applied to data and MC for a consistent comparison.

6.4.1 The b-jet selection

Jet distributions are measured at detector level for leading b-tagged jets with $p_{\rm T} > 20$ GeV in $|\eta| < 2.4$ and compared to the simulation. The b-tag discriminant, used for the selection, is the CSV, described in Chapter 5. The restriction in pseudorapidity is due to the absence of the tracker at forward rapidities, which makes the definition of a b-tag discriminant not applicable. Similarly to the inclusive leading jet distributions, the absolute differential cross section as a function of p_{T} and the normalized differential cross sections as a function of jet η and ϕ are measured for leading b-tagged jets and shown in Figure 6.7. The p_{T} cross section distribution is very similar to the one of inclusive leading jets: it is rapidly decreasing over more than four orders of magnitude from $p_{\mathrm{T}} \sim 20$ GeV down to $p_{\mathrm{T}} \sim 300$ GeV and the comparison with the simulation shows again a good agreement for $p_{\mathrm{T}} \sim 80$ GeV. Trigger inefficiencies are responsible for the bad description of the low p_{T} region. Note that the trigger efficiency correction shown in Section 6.2.1 is not applied at this level since it refers to a four-jet scenario and not to inclusive jet distributions. The normalized differential cross section as a function of the b-tagged jet ϕ, is flat over the whole phase space and well reproduced by the Monte Carlo predictions. The η distributions increase when going to the more forward region, due to the b-tag performance (see Chapter 7) and the simulation is able to reproduce this feature, with a slight modulation due to the trigger inefficiency which affects b-tagged jets, as well.

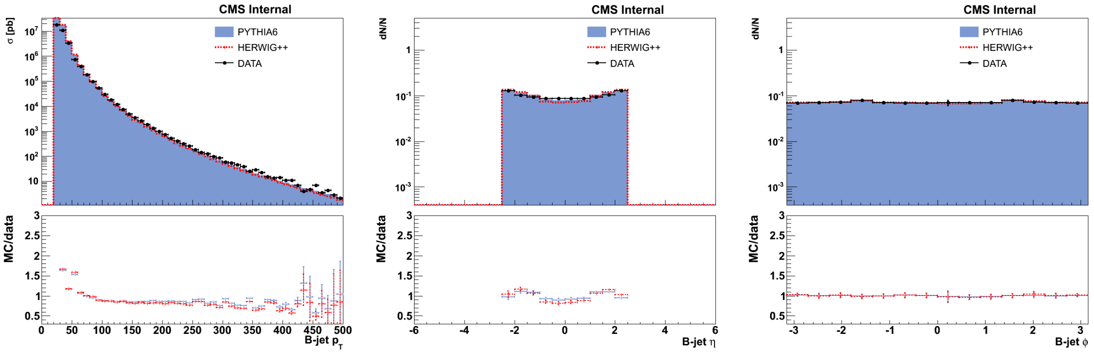


*Figure 6.7: Comparison of data with simulations from MC event generators (*Pythia6 *and* Herwig++*) for the transverse momentum of the leading tagged jet p_T (left), η (center), ϕ (right): absolute cross sections are shown for p_T, while normalized cross sections are shown as a function of η and ϕ. Only statistical uncertainties are shown for the data. Data and simulation are not normalized to the bin width. The lower panel shows the ratios of the theoretical predictions to data.*

It is also interesting to check the level of agreement between data and simulation in the specific b-tag discriminants: the normalized differential cross sections for the quantities Combined Secondary Vertex (CSV), Secondary Vertex (SV) and Track Counting High Efficiency (TCHE), are measured for leading jets with $p_{\mathrm{T}} > 20$ GeV and $|\eta| < 2.4$ and are shown in Figure 6.8. A slight discrepancy is observed for TCHE and CSV, while the SV discriminant exhibits a larger deviation at high values. These differences show a non optimal match between b-tag discriminant spectra measured in data and MC. The procedure of how to correct for this point, by applying scale factors in the simulation, is explained in Chapter 7.

The distributions show again that the behaviour of b-tagged jets is also rather well understood by the simulation. The selection set for the 2b2j analysis requires the presence of at least one good primary vertex. Inside this sample, at least four jets are required to have a $p_{\mathrm{T}} > 20$ GeV: two jets must fulfill the loose b-tag condition in $|\eta| < 2.4$, while no b-tag requirement for the other two jets is required. The acceptance for the non-tagged jets extends to the forward region of the detector and goes up to $|\eta| < 4.7$. The two leading jets are selected for each category; thus, for instance, in case three jets are b-tagged, the two with highest p_{T} are selected. Different selections have also been studied, in order to find the optimal requirements for a highest b-jet purity:

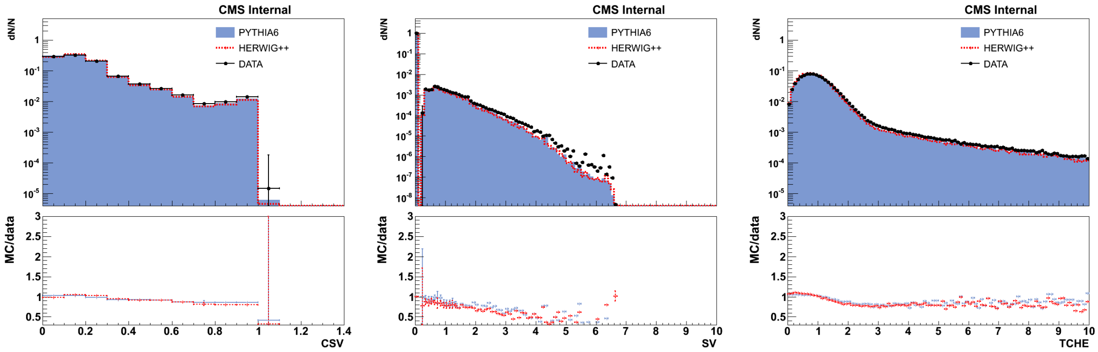


*Figure 6.8: Comparison of data with simulations from MC event generators (*Pythia*6 and* Herwig*++) for various b-tag discriminants: CSV (left), SV (center), TCHE (right). Normalized cross sections are shown as a function of the discriminants. Only statistical uncertainties are shown for the data. Data and simulation are not normalized to the bin width. The lower panel shows the ratios of the theoretical predictions to data.*

- events are selected only if the two tagged jets are the ones with the highest p_{T}, otherwise they are rejected;
- tagged jets are chosen according to the value of the CSV discriminant: instead of choosing tagged jets with highest p_{T}, the two tagged jets with the highest CSV value are selected.

No differences in purity, namely the percentage of tagged jets which correspond to true b-jets at the generator level (see Section 6.7), are observed if any of these two selections are applied. Hence, the selection has been kept as explained before, because it maximises the achieved statistical accuracy in data and gives an unambiguous definition at the stable particle level.

The two b-tagged jets are associated and labelled as "b-jet pair", while the other two additional jets constitute the "light-jet pair". Similarly as for the 4j analysis, jets are reconstructed by using the PF information, and clustered with the anti-k_{T} algorithm, by using a cone size R = 0.5. The tight jet selection is also applied to all jets, in order to remove non-physical jets, arising from detector noise. Exactly the same selection is applied to data and MC for a consistent comparison.

6.5 Selection summary

In summary, two different selections have been applied for the two analyses, the 4j and the 2b2j channels, and the selection requirements are summarized respectively, in Table 6.6 and 6.7.

At least one good reconstructed primary vertex		
Quality criteria applied to the vertex		
Exactly four jets		
Tight selection applied to the jets		
Two hard jets	$p_{\mathrm{T}} > 50$ GeV	$\lvert\eta\rvert < 4.7$
Two soft jets	$p_{\mathrm{T}} > 20$ GeV	$\lvert\eta\rvert < 4.7$

Table 6.6: Selection criteria applied to data and MC in the 4j analysis.

By applying these analysis cuts separately, events at detector level are selected and further analyzed. Tables 6.8 and 6.9 list the selected event counts, for each data sample and each trigger

	At least one good reconstructed primary vertex	
	Quality criteria applied to the vertex	
	At least four jets	
	Tight selection applied to the jets	
Two tagged jets	$p_{\mathrm{T}} > 20$ GeV	$\lvert\eta\rvert < 2.4$
	CSV discriminant > 0.244	
Two jets	$p_{\mathrm{T}} > 20$ GeV	$\lvert\eta\rvert < 4.7$

Table 6.7: Selection criteria applied to data and MC in the 2b2j analysis.

region, in the different steps of the analysis for the 4j selection. Tables 6.10 and 6.11 list the selected event counts for the 2b2j selection. In the tables, different quantities are represented:

- Total number of events: number of events present in the data/MC samples;
- Trigger region (only for data);
- Good vertex requirement: number of events with at least one good reconstructed vertex;
- Number of events in the trigger region: number of events with at least one jet in $|\eta| < 4.7$ with $p_{\mathrm{T}} > 50$ (20) GeV, for the 4j (2b2j) analysis;
- Number of selected events: number of events after applying the whole chain of analysis cuts.

For the 4j analysis, about 0.2M events are selected, of which 0.13M in the most populated data sample. A number of 1.3M events are selected from MC; thus, a sample with a number of selected events bigger than a factor of ten is available in simulation with respect to data. For the 2b2j analysis, about 65000 events are selected in data and more than three million in MC.

DATA Sample	Total Number of Events	Trigger region	Good Vertex requirement	Number of events in trigger region	Number of Selected Events
JETMET	20350952	50-80	6730757	1168570	32149
	20350952	80-140	6730757	232628	31033
	20350952	>140	6730757	283420	72372
JETMETTAU	4887422	50-80	1728630	722490	20039
	4887422	80-140	1728630	143640	18850
	4887422	>140	1728630	27606	7052
JET2010B	11234922	50-80	408361	165592	4832
	11234922	80-140	408361	33016	4424
	11234922	>140	408361	24902	6312

Table 6.8: Number of events passing the applied selections for each data sample and each trigger region in the 4j analysis. Under the column "Number of triggered events" are the events with at least one jet above 50 GeV in $|\eta| < 4.7$.

Monte Carlo Sample	Total Number of Events	Good Vertex requirement	Number of Triggered Events	Number of Selected Events
Pythia6	9982400	7128229	7127080	1310106
Herwig++	9971200	6732445	6731530	1300296

Table 6.9: Number of events passing the applied selections for each MC sample. Under the column "Number of triggered events" are the events with at least one jet above 50 GeV in $|\eta| < 4.7$.

DATA Sample	Total Number of Events	Trigger region	Good Vertex requirement	Number of events in trigger region	Number of Selected Events
JETMETTAU	4887422	20-50	1728630	615234	3964
JETMET	20350952	50-140	6730757	1716886	36910
JETMET	20350952	>140	6730757	301015	25445

Table 6.10: Number of events passing the applied selections for each trigger region in the 2b2j analysis. Under the column "Number of triggered events" are the events with at least one jet above 20 GeV in $|\eta| < 4.7$.

Monte Carlo Sample	Total Number of Events	Good Vertex requirement	Number of Triggered Events	Number of Selected Events
Pythia6	59893271	54909643	54900746	3057866
Herwig++	59607080	54564533	54555620	2862452

Table 6.11: Number of events passing the applied selections for each MC sample in the 2b2j analysis. Under the column "Number of triggered events" are the events with at least one jet above 20 GeV in $|\eta| < 4.7$.

6.6 Data-MC comparisons at detector level

In order to check the reliability of the detector simulation, the agreement with the measurements needs to be checked. In order to trust the simulation for data correction and estimation of uncertainties, data need to be described as good as possible by the simulation. In case of bad agreement in some regions of the phase space, best is to apply some corrections to improve the description or choose different samples. After trigger efficiency correction and selection criteria, distributions at detector level of data and MC are compared. The weights related to the pile-up reweighting are applied to the MC samples. Results are described separately for the two analyses.

6.6.1 Four-jet (4j) analysis

Absolute cross sections are measured at the detector level for the 4j analysis, differentially as a function of p_{T} and η of single jets, respectively in Figure 6.9 and 6.10. The p_{T} spectra of all jets are rapidly decreasing for increasing transverse momenta. Leading and subleading jets decrease over three orders of magnitude from 50 to 300 GeV, while the slope of the soft jets is steeper: it falls over five orders of magnitude in the same p_{T} range, from 50 to 300 GeV. This fast decrease is understandable, since they correspond to jets mainly coming from hard radiation, in presence of other jets with higher p_{T}: additional jet radiation tends, indeed, to fill lower p_{T} regions. The differential cross sections as a function of η show again a different behaviour between hard and soft jets. The hard jets, leading and subleading, are mostly present in the region of $|\eta| < 3$; the distributions fall down over two orders of magnitude in the most forward region ($|\eta| > 4$). The soft jets show, instead, a rather flat distribution in the considered pseudorapidity range, with a difference of less than a factor of 10 between central and forward region. This main difference reflects the different p_{T} thresholds applied for the two types of jets.

Comparisons between data and available simulation are also shown for the measured cross sections. The considered MC samples are able to describe reasonably well the falling p_{T} spectra for all jets: differences mostly appear in the low p_{T} region and they are overall less than 20% with respect to the data. Note that samples generated with PYTHIA6 and HERWIG++, even though they use a LO ME interfaced with PS and UE, offer a very good description of a scenario with high jet multiplicity. This can be explained, on one hand, by the fact that a DGLAP-like scenario is selected, with two hard jets and two softer jets which are produced in LO MC generators via PS, and, on the other hand, by the reliable UE simulation in the samples, provided by well tested tunes. Comparisons of cross sections between data and simulation are also shown for the pseudorapidity distributions: while, in the central region, the description, provided by the simulation, is reasonably good with overall discrepancies below 20% with respect to the data, bigger differences, of the order of 30-80%, appear when looking at the hard jets in the most forward region, $|\eta| > 3$. Further investigations have been performed to understand this behaviour. First of all, this forward region is affected by large uncertainties in the jet energy scale, up to 60%. Secondly, a strong generator dependence is observed for jets in this region: differences up to 30% are observed for cross sections obtained with different generators. A slight difference in the η spectra of the hard jets is observed in the region of $|\eta| \sim$ 2-2.5 between predictions obtained with PYTHIA6 and HERWIG++: further investigations have shown that this is mainly due to the different parton shower model, implemented in the two generators.

Correlation observables have been also measured at detector level and compared to MC predictions and the level of agreement is very good, as shown for the p_T and η spectra. This gives confidence that these two samples, generated with PYTHIA6 and HERWIG++, are reliable and can be used for the data correction at the particle level. This procedure is described in Chapter 8.

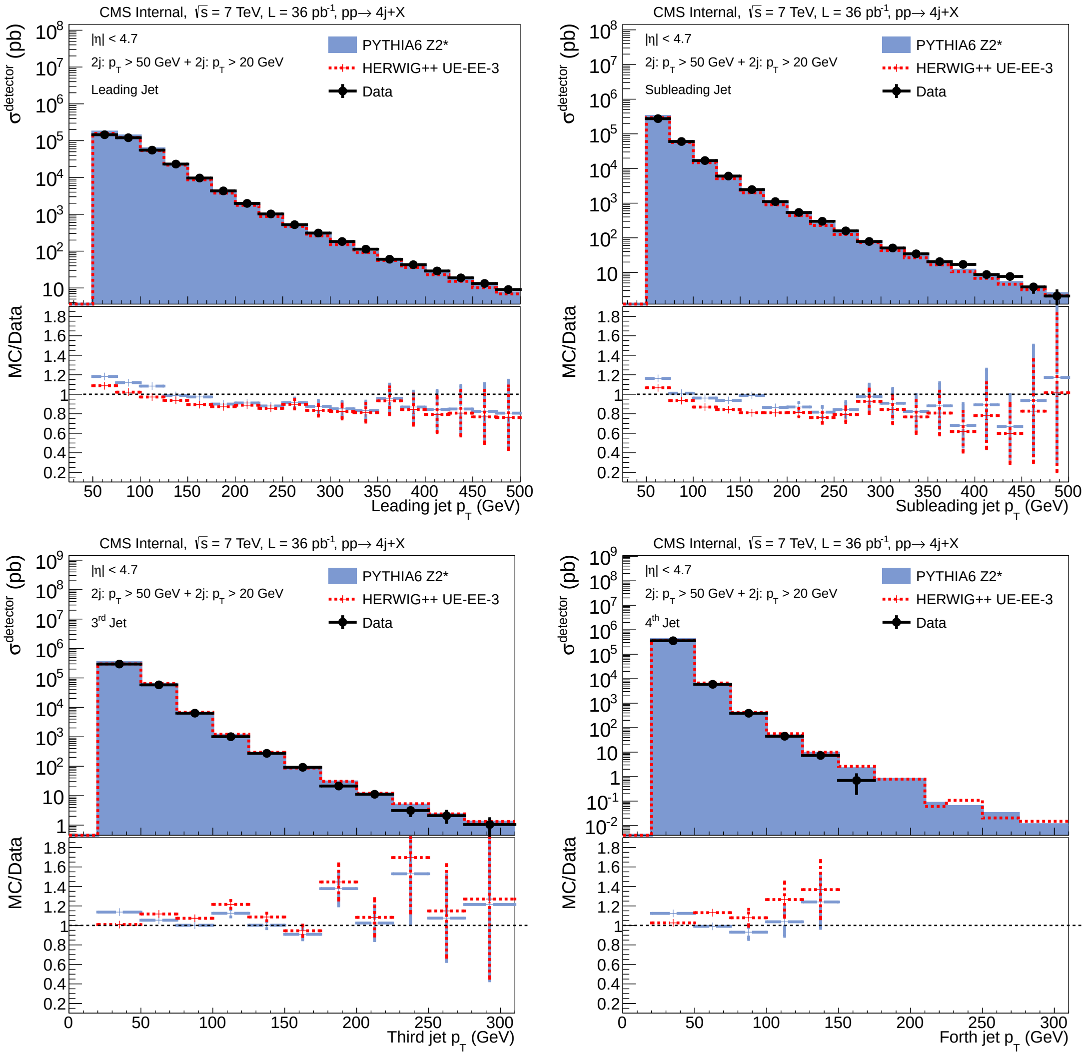


*Figure 6.9: Comparison of data with simulations from MC event generators (*PYTHIA*6 and* HERWIG*++) for the transverse momentum of the leading (top left), subleading (top right), 3^{rd} (bottom left) and 4^{th} (bottom right) jets. Only statistical uncertainties are shown for the data. Data and simulation are not normalized to the bin width. The lower panel shows the ratios of the theoretical predictions to data.*

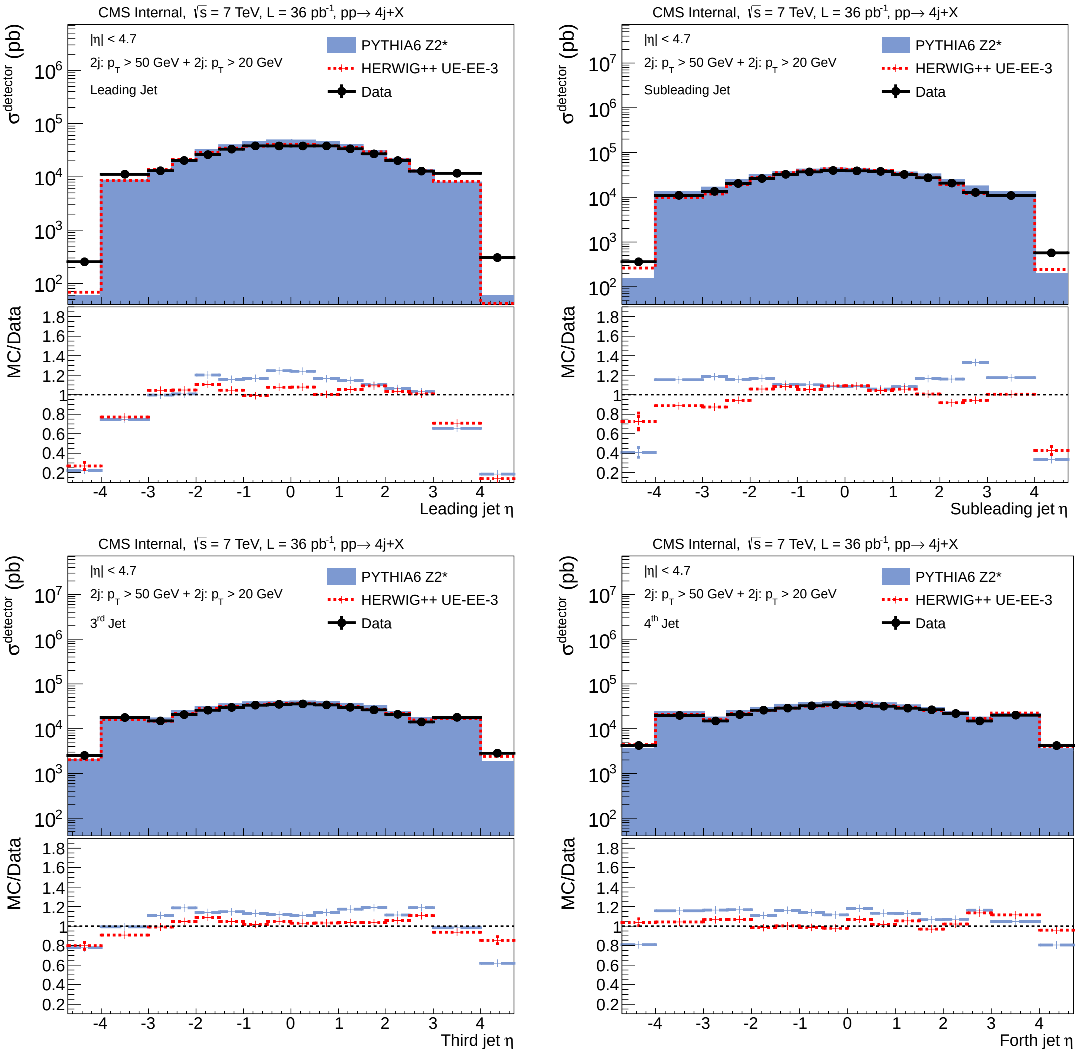


*Figure 6.10: Comparison of data with simulations from MC event generators (*Pythia*6 and *Herwig*++) for the pseudorapidity of the leading (top left), subleading (top right), 3rd (bottom left) and 4th (bottom right) jets. Only statistical uncertainties are shown for the data. Data and simulation are not normalized to the bin width. The lower panel shows the ratios of the theoretical predictions to data.*

6.6.2 Two b- and two other jet (2b2j) analysis

Differential cross sections are also measured for the 2b2j scenario. In this case, distributions normalized to the total number of events are considered to evaluate the level of agreement between data and MC. This is motivated by the fact that already in differential inclusive b-jet cross sections [207], a non-negligible disagreement in the normalization is observed for data and LO generators, like PYTHIA6. A better agreement with the measurement is achieved by generators implementing a NLO ME, but their predictions are not available at the detector level for the considered analyses. The disagreement for LO generators is not expected to disappear in scenarios with a higher jet multiplicity, like the considered 2b2j one. By normalizing the distributions to the total number of events, these normalization problems, due to missing higher-order diagrams and missing virtual corrections, are largely removed and a more fair comparison is provided.

Figure 6.11 shows the normalized differential cross sections as a function of jet p_{T}, measured at detector level. They exhibit a decreasing distribution when going to large p_{T} values. From 30–50 GeV down to 400 GeV, the difference in cross section is of about five orders of magnitude for the leading jets, and six orders of magnitude for the subleading ones. The decrease of cross section in the first bin is just an artifact of the binning, because it contains only jets with p_{T} between 20 and 25 GeV. Figure 6.12 displays the normalized differential cross sections as a function of jet η, measured at detector level. The pseudorapidity distributions of the selected b-tagged jets increase when going to higher η with maximum for values around 2 and minimum in the very central region ($\eta \sim 0$): this is an effect of the performance of the specific b-tag algorithm applied, in particular of the loose working point (see Chapter 7). For both the other jets, the η cross sections are rather flat in the region of $|\eta| < 3.8$ with a maximum for $\eta \sim 0$.

The measurements are also compared to the predictions provided by PYTHIA6 and HERWIG++. They show that, while the η distributions are quite well described by both predictions, with only small discrepancies in the forward region for the additional jets of up to 20%, the p_{T} distributions are not well reproduced. In particular, both event generators predict a softer spectrum for all jets with respect to the data, underestimating the high-p_{T} region of about 30–40%. The slight disagreement observed for the η distributions may be explained again by the jet energy scale uncertainty, which plays a relevant role in the forward region. The discrepancies, arising from the comparisons of the p_{T} spectra, need further studies to be solved and understood; relevant effects might come from well known differences in the b-tag performance in data and MC, or from the physics models implemented in the specific generators, PYTHIA6 and HERWIG++, which generate only LO 2→2 ME. Studies of possible effects have been carried out for this analysis and the solution adopted to improve the jet p_{T} description is described in Chapter 7.

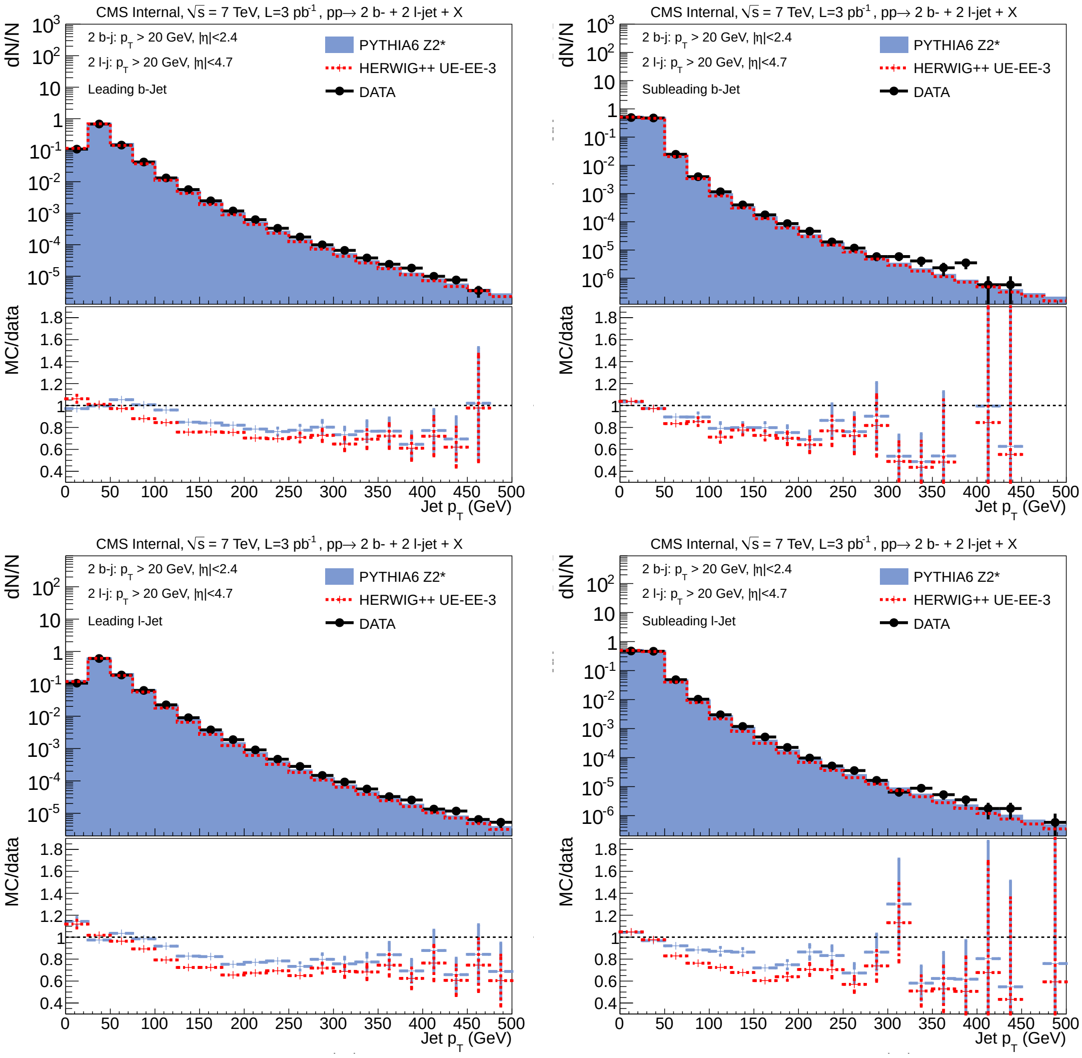


*Figure 6.11: Comparison of data with simulations from MC event generators (*Pythia*6 and* Herwig++*) for the transverse momentum of the leading (top left), subleading (top right) tagged jets, and leading (bottom left) and subleading (bottom right) additional jets. Only statistical uncertainties are shown for the data. Data and simulation are not normalized to the bin width. The lower panel shows the ratios of the theoretical predictions to data.*

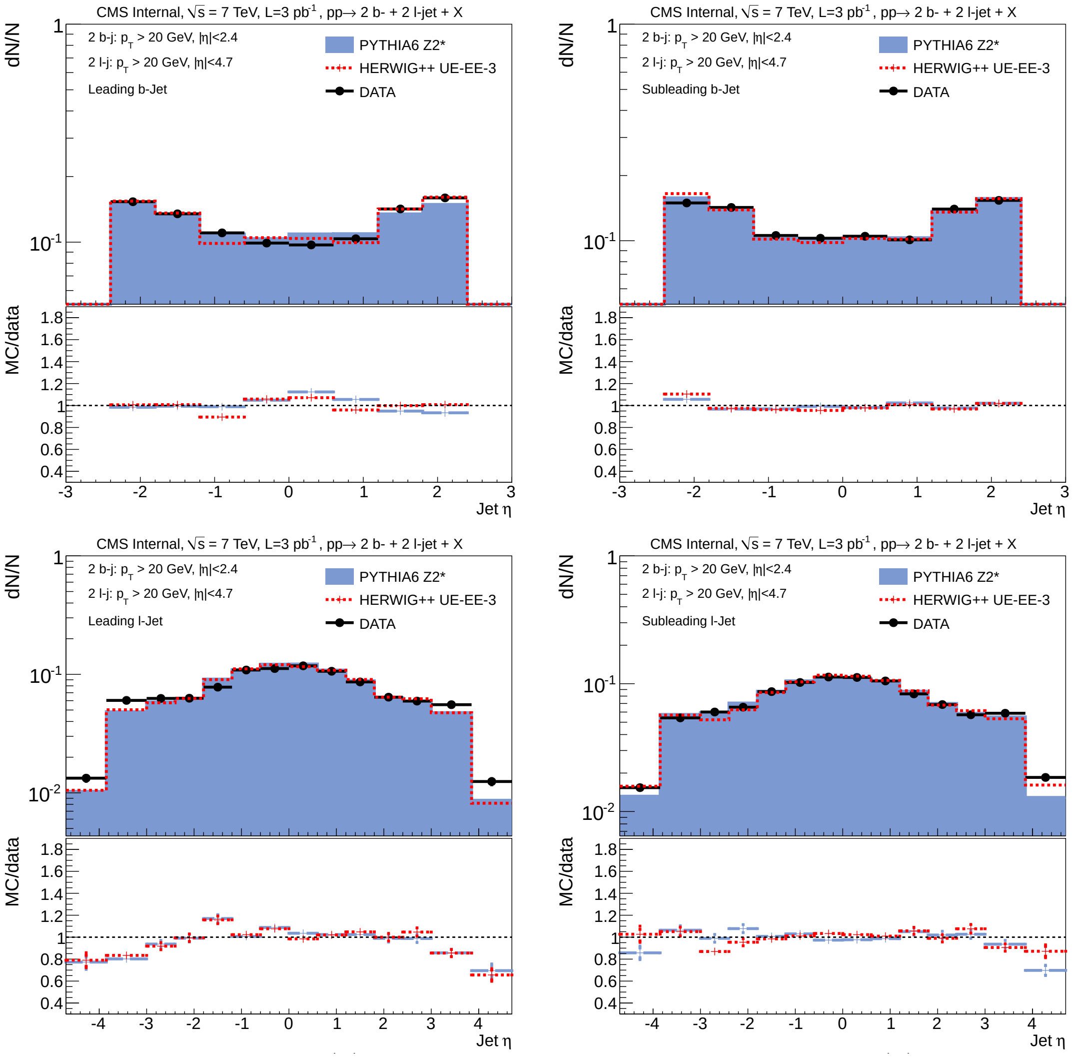


*Figure 6.12: Comparison of data with simulations from MC event generators (*Pythia6 *and* Herwig++*) for the pseudorapidity of the leading (top left), subleading (top right) tagged jets, and leading (bottom left) and subleading (bottom right) additional jets. Only statistical uncertainties are shown for the data. Data and simulation are not normalized to the bin width. The lower panel shows the ratios of the theoretical predictions to data.*

Chapter 7

Study of detector effects

In this chapter, all possible experimental effects are described in full detail. The investigation of the detector effects has been performed in different steps:

- study of the experimental resolution from simulated events: this helps to choose an appropriate binning for the measured histograms (Section 7.1);
- investigation of purity, stability, background and acceptance for the understanding of the migration effects (Section 7.2);
- study of the response matrices, which connect detector and generator level quantities (Section 7.3);
- analysis of the *b*-tag performance in data and simulation (Section 7.4).

Note that effects due to detector trigger, along with the related corrections to be applied to the data, have been already described in Chapter 6. All the studies about the detector effects are shown for the 2b2j analysis, and similar conclusions, where not explicitly said, can be drawn for the 4j scenario.

7.1 Resolution effects

Measurements must, in general, always deal with resolution effects. An ideal experimental measurement would be able to determine exactly and in an infinitely precise way the value of a physical quantity: suppose, for instance, that a particle crosses an ideal calorimeter which is designed for the measurement of particle energies. If the traversing particle has an energy E, the output of the measurement would return exactly the value E. In addition, E would be also the value of energy measured by the calorimeter for any other particle of the same type, crossing the detector with the same energy. In this case, one can say that this detector has an infinite resolution, namely it measures an experimental quantity with an infinite precision. Nevertheless, real detectors differ from this ideal picture: in particular, they are affected by many effects, such as detector noise, calibration uncertainties, non-linearity of the response and many others, which determine a casual shift of the measured quantity from its true value. This translates into a finite value of the detector resolution. Suppose, for instance, that the same previous particle of energy E crosses a real detector: the measured energy might differ from the value E under the forementioned effects. In addition, in case the measurement is repeated several times for different crossing particles, all with the same energy E, different values would be detected among the different measurements. In most cases, the response of a set of measurements of this type would be a gaussian distribution centered around the true value of the measured quantity (for a well calibrated detector). The width of this gaussian distribution[1] is generally interpreted as detector resolution, which is an indicator of how precise the detector is able to measure a given physical observable. The higher the resolution, the narrower the distribution is and the more precise and accurate the detector performs.

The measurement of the detector resolution for the interesting observables is crucial for the determination of the histogram binning. In fact, the resolution represents how much the measured value differs from the true one; in binned histograms, special care needs to be given to the fact that measured quantities may migrate from one bin to another with respect to their true value. This effect might complicate the correction procedure of the data and one should try to avoid these migrations. The solution is a choice of bin widths which are at least two-three times larger than the detector resolution in that particular bin. The bin widths may also differ between each other for the same observable.

The resolution has been measured for all observables. The sample generated with Pythia6 has been used where events at detector and generator level are selected. In case the same event is selected in both levels, an additional requirement is applied, by matching the corresponding selected jets. This assures that only corresponding jets are coupled at the two levels, in the studied events. This is the way to study how a true generator level four-jet scenario is reconstructed at the detector level and how the measured quantities are degraded under the effect of the jet reconstruction. The matching has been performed in different ways for the various quantities in order not to introduce any bias in the measurement of the resolution; in particular, a cone of width equal to $R = \sqrt{(q_1^{\mathrm{gen}} - q_1^{\mathrm{det}})^2 + (q_2^{\mathrm{gen}} - q_2^{\mathrm{det}})^2}$[2], is used for the matching:

- $\Delta\eta$, jet η (η-based observables): $q_1 = \phi$, $q_2 = p_{\mathrm{T}}$, $R_\phi = 0.1$, $R_{p_{\mathrm{T}}} = 0.02$;
- $\Delta\phi$ (ϕ-based observables): $q_1 = \eta$, $q_2 = p_{\mathrm{T}}$, $R_\eta = 0.1$, $R_{p_{\mathrm{T}}} = 0.02$;
- ΔS, $\Delta^{\mathrm{rel}} p_{\mathrm{T}}$ and jet p_{T} (p_{T}-based observables): $q_1 = \eta$, $q_2 = \phi$, $R = 0.3$;

The matching quantities, q_1 and q_2, need to be different from the measured ones: for instance, for measurements of p_{T}-based observables, the matching should be applied in the η-ϕ space, for η-based observables, a ϕ-p_{T} is more appropriate and so on. The matching needs, in fact, to be performed in a phase space which does not include the specific measured quantity. In case this is not applied, a strong dependence of the resolution values is observed on the matching widths. Values of resolution for different matching widths have been checked: the cone aperture, whose measurement is presented here, provides a very stable resolution response. It has been observed that for higher widths, the resolution tends to increase because the correlation between the corresponding selected jets starts to be lost, while for lower widths, similar values of the resolution are obtained. In this last case, the only difference is that some statistics is lost for the selected events,

---

[1] In case of a non-gaussian distribution, the measurement of the detector resolution is still doable but is more complicated.

[2] If q_2 is p_{T}, the cone width is set separately for the two matching variables in this way: $R_{q_1} = (q_1^{\mathrm{gen}} - q_1^{\mathrm{det}})$ and $R_{p_{\mathrm{T}}} = (p_{\mathrm{T}}^{\mathrm{gen}} - p_{\mathrm{T}}^{\mathrm{det}})/(p_{\mathrm{T}}^{\mathrm{gen}})$.

indicating that too tight matching criteria are applied.

The resolution is measured from the distributions of the difference of the quantities measured at generator and detector level for the matched events: for a sufficient number of events, these distributions are gaussian. The resolution is then obtained by taking the width of the gaussian distributions in every bin of the considered observable. Only for the measurement of jet p_T, the relative resolution, instead of the absolute one, is measured, namely the difference of transverse momenta measured at generator and detector level for the matched events, divided by the p_T at generator level. This is motivated by the fact that the measured p_T range extends from 20 up to 500 GeV, for more than an order of magnitude, and the relative resolution is a better estimation of the detector performance, rather than the absolute one.

For the 2b2j analysis, two resolution measurements have been performed for two different selections at detector level: the first one (labelled as "No MC Truth") where the b-jets are selected through the CSV-tagging algorithm, the second one (labelled as "MC Truth") where the MC truth information about the jet flavour is used for the b-jets in the final state, namely only reconstructed b-jets are truly b-jets. This separate study has been done to disentangle the effects of the b-tag on the resolution measurement. Figures 7.1 and 7.2 show the resolution measurements, respectively for the correlation observables and the jet p_T spectra. The use of MC truth information improves the $\Delta\phi$ resolution; this is expected, since it removes possible smearing effects due to the b-tag procedure. For the $\Delta^{\mathrm{rel}}p_\mathrm{T}$ and the jet p_T measurements, the resolution does not change much if the MC truth information is used or not. This is due to the fact that, by applying a matching in η and ϕ, for which the detector resolution is very good, the possibility of associating different jets at detector and generator level is already removed without any MC truth information.

The resolution for $\Delta\phi$ and $\Delta^{\mathrm{rel}}p_\mathrm{T}$ is between 0.02 and 0.1, when the MC truth information is applied, with maximum values in the low region of $\Delta\phi^{\mathrm{light}}$. Low variations of the resolution are observed for $\Delta^{\mathrm{rel}}p_\mathrm{T}$, over the whole range. The relative resolution for the jet p_T is decreasing for increasing transverse momenta, going from 0.17 at 20–50 GeV down to 0.07 for 400–500 GeV.

Considering the results obtained for the resolution, a bin width at least bigger than two times the corresponding resolution, has been chosen for the measured histograms. This choice minimizes the migration effects due to detector reconstruction but still allows to get detailed information, differentially in the measured observable. This feature holds for both the 4j and the 2b2j analyses. The histograms, shown hereafter, use the binning derived from resolution determinations.

7.2 Purity, stability, background and acceptance

After determining the bin widths, the response at generator and detector level is studied. Several occurrences may arise for a measured observable: the measurement at detector and generator level stays in the same bin or the measurement corresponds to different bins at the two levels. It may also happen that an event is selected in only one of two levels. If the measurement at generator and detector level does not remain in the same bin, one speaks of "migration effects". There are two types of migration effects: the migrations "within the phase space" are the ones where the events are selected in both levels but they fill different bins in the histograms at detector and generator

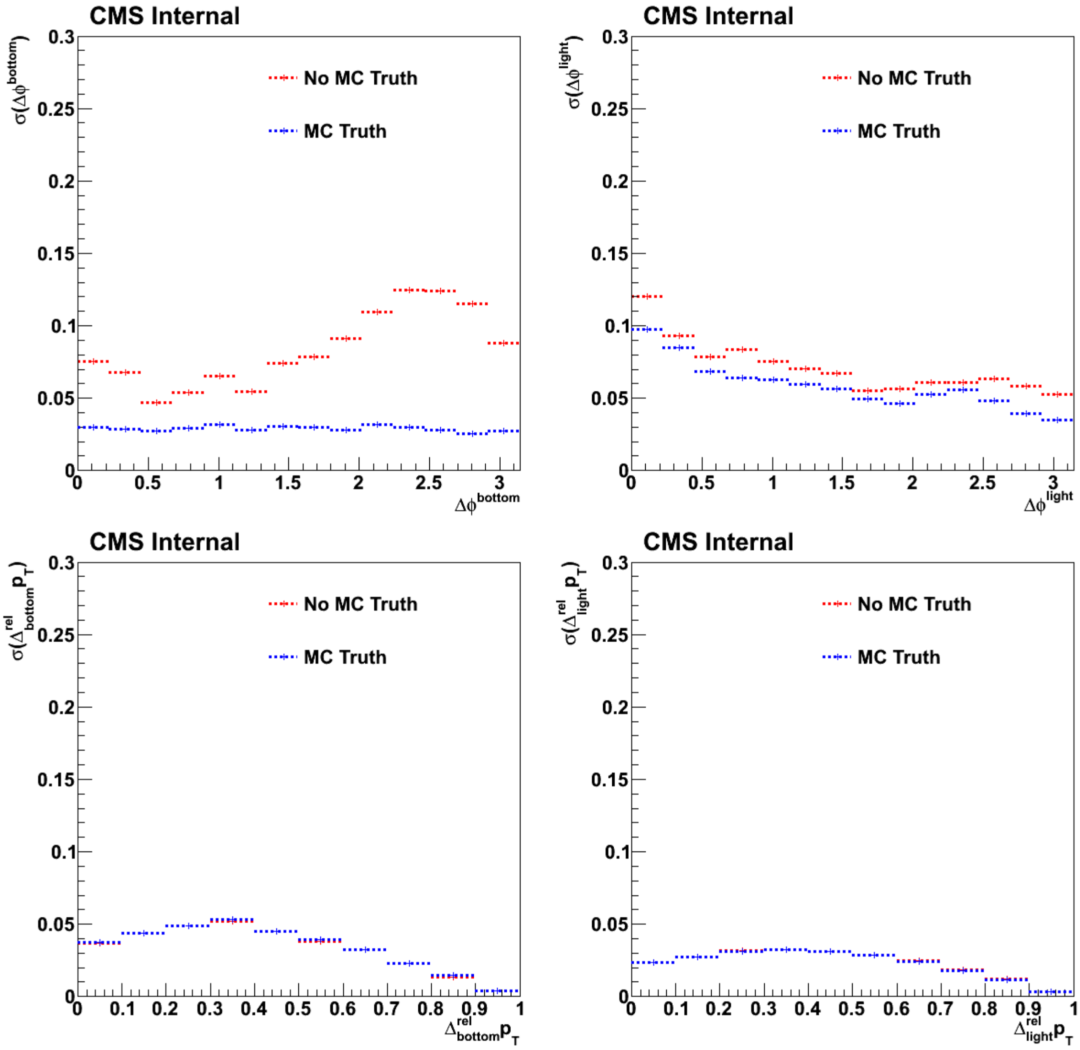


Figure 7.1: Resolution determined for correlation observables in the 2b2j analysis: $\Delta\phi^{bottom}$ *(top left),* $\Delta\phi^{light}$ *(top right),* $\Delta^{rel}_{bottom}p_T$ *(bottom left) and* $\Delta^{rel}_{light}p_T$ *(bottom right). The resolutions are shown when events are selected at the detector level with b-tag discriminants or with MC truth information. Resolutions are determined by using the* PYTHIA*6 sample. Events at detector and generator level are matched with a cone algorithm.*

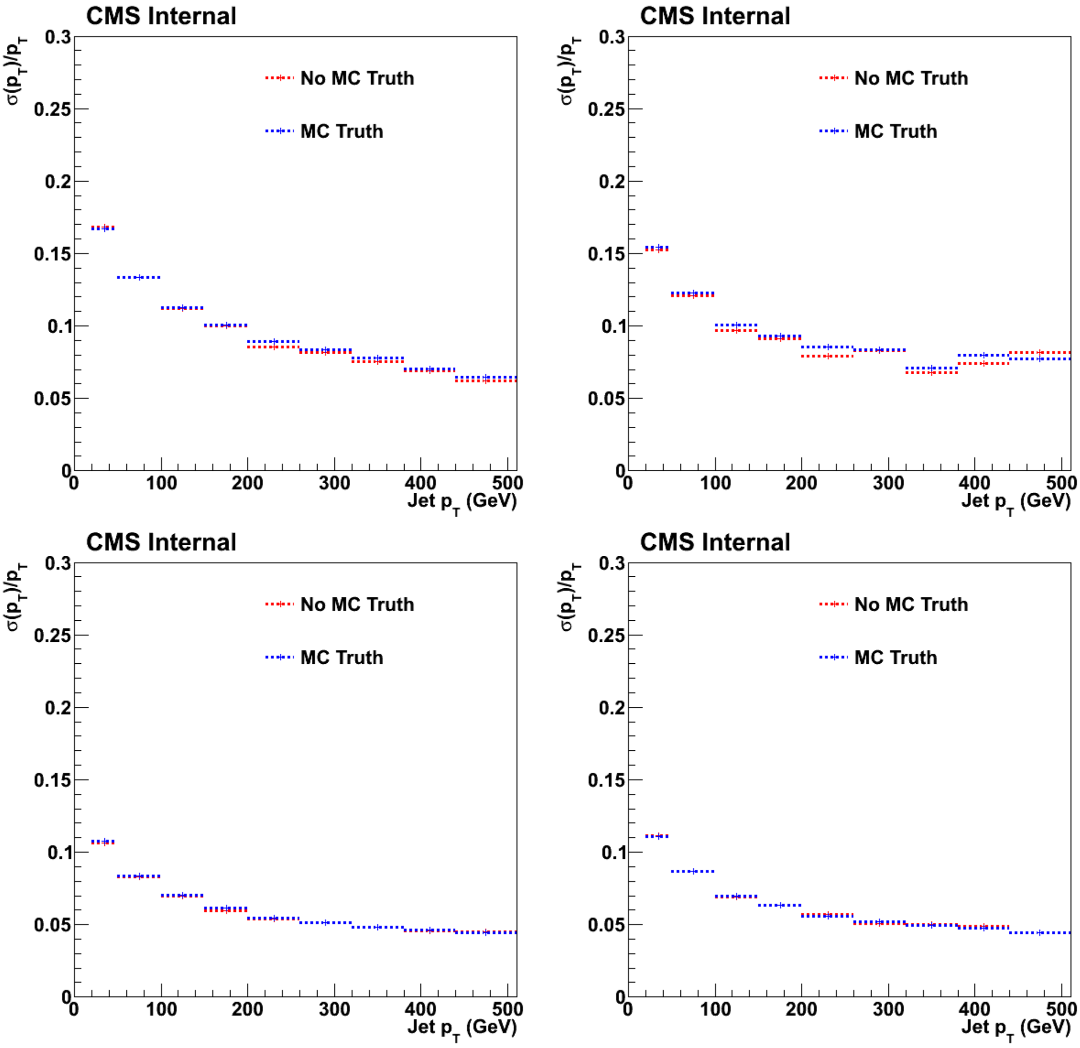


Figure 7.2: Resolution determined for p_T spectra in the 2b2j analysis: leading b-tagged jet (top left), subleading b-tagged jet (top right), leading additional jet (bottom left), subleading additional jet (bottom right). The resolutions are shown when events are selected at the detector level with b-tag discriminants or with MC truth information. Resolutions are determined by using the PYTHIA6 *sample. Events at detector and generator level are matched with a cone algorithm.*

level while the migrations “into or out of the phase space” are the ones where the events are selected in only one of the two levels. Studies on migration effects can be performed by measuring purity, stability, background and acceptance [208]. The first two measure the behaviour of the migrations within the phase space, while with the second two, the migrations into or out of the phase space are studied. They are defined as follows:

- the purity represents the percentage of events in a certain bin i at the detector level *det*, which are also selected at the generator level *gen* and belong to the same bin; in a compact formula, it can be written as:

$$P_{\mathrm{i}}^{\mathrm{MC}} = \frac{N^{\mathrm{MC}}_{\text{both select}}(E^{\mathrm{MC}}_{\mathrm{det}} \in bin\, i \wedge E^{\mathrm{MC}}_{\mathrm{gen}} \in bin\, i)}{N^{\mathrm{MC}}_{\text{both select}}(E^{\mathrm{MC}}_{\mathrm{det}} \in bin\, i)}; \tag{7.1}$$

- the stability quantifies the percentage of events in a certain bin i at the generator level *gen*, which are also selected at the detector level *det* and belong to the same bin; in the same way, it can be written as:

$$S_{\mathrm{i}}^{\mathrm{MC}} = \frac{N^{\mathrm{MC}}_{\text{both select}}(E^{\mathrm{MC}}_{\mathrm{det}} \in bin\, i \wedge E^{\mathrm{MC}}_{\mathrm{gen}} \in bin\, i)}{N^{\mathrm{MC}}_{\text{both select}}(E^{\mathrm{MC}}_{\mathrm{gen}} \in bin\, i)}; \tag{7.2}$$

- the background measures the percentage of events in a certain bin i at the detector level *det*, which are also selected also at the generator level *gen*; this translates into the following definition:

$$B_{\mathrm{i}}^{\mathrm{MC}} = 1 - \frac{N^{\mathrm{MC}}_{\text{both select}}(E^{\mathrm{MC}}_{\mathrm{det}} \in bin\, i)}{N^{\mathrm{MC}}_{\text{select}}(E^{\mathrm{MC}}_{\mathrm{det}} \in bin\, i)}; \tag{7.3}$$

- the acceptance measures the percentage of events in a certain bin i at the generator level *gen*, which are also selected also at the detector level *det*; it can be written as:

$$A_{\mathrm{i}}^{\mathrm{MC}} = \frac{N^{\mathrm{MC}}_{\text{both select}}(E^{\mathrm{MC}}_{\mathrm{gen}} \in bin\, i)}{N^{\mathrm{MC}}_{\text{select}}(E^{\mathrm{MC}}_{\mathrm{gen}} \in bin\, i)} \tag{7.4}$$

where N represents a generic number of events, the subscripts *both select* and *select* indicate events selected, respectively, at both levels and at only one of them. These quantities have been determined by using the PYTHIA6 sample. As done for the resolution determination of the 2b2j analysis, two selections have been separately applied at the detector level: one which applies the b-tag CSV algorithm, in order to select the tagged jets, and one which uses the MC truth information. These separate studies help to understand the effect of the specific b-tag algorithm, applied at the detector level. Figure 7.3 shows the results of the migration effects for the correlation observables when the CSV algorithm has been used for the b-jet selection. It can be seen that acceptance and background are rather flat around the whole ranges and respectively, of the order of 70–80% and 80–90%. In words, this means that a percentage of 70–80% of true events with two b- and two additional jets, is efficiently selected at detector level, but that 80–90% of the events selected at the detector level do not correspond to a real event with the correct number and flavour of jets at the generator level. In addition, it can be observed that purity and stability for observables related to the additional jets, are between values of 50 and 70%, while they tend to deviate from each other for the b-jet observables: purity is between 50–75%, while stability has values between 25–80%. Bigger fluctuations are observed for $\Delta^{\mathrm{rel}}_{\mathrm{bottom}} p_{\mathrm{T}}$ with respect to the ones obtained for $\Delta\phi^{\mathrm{bottom}}$: low $\Delta^{\mathrm{rel}}_{\mathrm{bottom}} p_{\mathrm{T}}$ values have smaller purity and stability which increase for higher $\Delta^{\mathrm{rel}}_{\mathrm{bottom}} p_{\mathrm{T}}$. Interesting is to note that these values significantly change when the MC truth information is used for the selection at the detector level: studies under these conditions isolate the detector effects from the ones related to the b-tag. Figure 7.4 shows the distributions measured by using MC truth information. Background and acceptance result to be very stable along the phase space with values of respectively 20% and 90%. Purity and stability are now very similar between each other for light- and b-jet based observables: values between 60–90% are obtained for both. A stable behaviour is observed for $\Delta\phi$, while growing values of purity and stability are measured for increasing $\Delta^{\rm rel}p_{\rm T}$. This is understandable by the fact that migrations are higher for low $\Delta^{\rm rel}p_{\rm T}$, corresponding to a back-to-back jet configuration: in such a scenario, indeed, small changes in jet $p_{\rm T}$, due to detector resolution, might destroy more easily the balance and induce larger migrations with respect to a collinear topology, like at high values of $\Delta^{\rm rel}p_{\rm T}$. This effect reflects on the distribution of purity and stability observed for $\Delta^{\rm rel}p_{\rm T}$. The big change in values observed among the studies performed with and without the use of MC truth information is also reasonable, due to the fact that the loose working point has been chosen for the b-tag CSV algorithm: this assures a very high b-jet selection efficiency but it implies a high background of fake jets (for quantitative values, see Section 7.4).

Measurements of purity, stability, background and acceptance, performed in the 4j scenario, resemble the ones obtained for the 2b2j scenario when MC truth information is used. The observed values in this last study on migration effects have shown a low background level and quite high degree of acceptance, purity and stability. This suggests that the detector is well understood and a good correlation between detector and generator level is obtained.

It is also interesting to have a look at migration effects for the ΔS observable, which is the most complex quantity because it uses information of all jets in the final state, not only of jets belonging to the same pair. Figure 7.5 shows migration effects in case the b-tag algorithm or the MC truth information is applied at the detector level. The same behaviour as for the other correlation observables is observed for background and acceptance: without MC truth information, the background is around high values (85–90%) and the acceptance stays at 70–75%. With the use of MC truth information, background drops down to 20%, while the acceptance increases up to 80–85%. In all cases, both of them have flat distributions along the measured range. For purity and stability, the behaviour is very similar between each other but very different along the phase space: at low values of ΔS, up to 2–2.5, they are around 10–20% before increasing to 60–80% at high values of ΔS. This trend does not change neither when the MC truth information is used. Note that the first three bins have a larger bin width, due to the lower resolution at low ΔS values; even with wider bins, purity and stability still remain below 50–60%. This means that this variable is affected by large migration effects at low ΔS values, not depending on the b-tag algorithm but only on the detector reconstruction. This might be understood by the complexity of the definition of this observable, which strongly depends on the jet $p_{\rm T}$ resolution. The migration effects are taken into account and corrected for, through the unfolding, which aims to remove any detector effect in order to bring the data to the stable particle level. The unfolding procedure is described in Chapter 8.

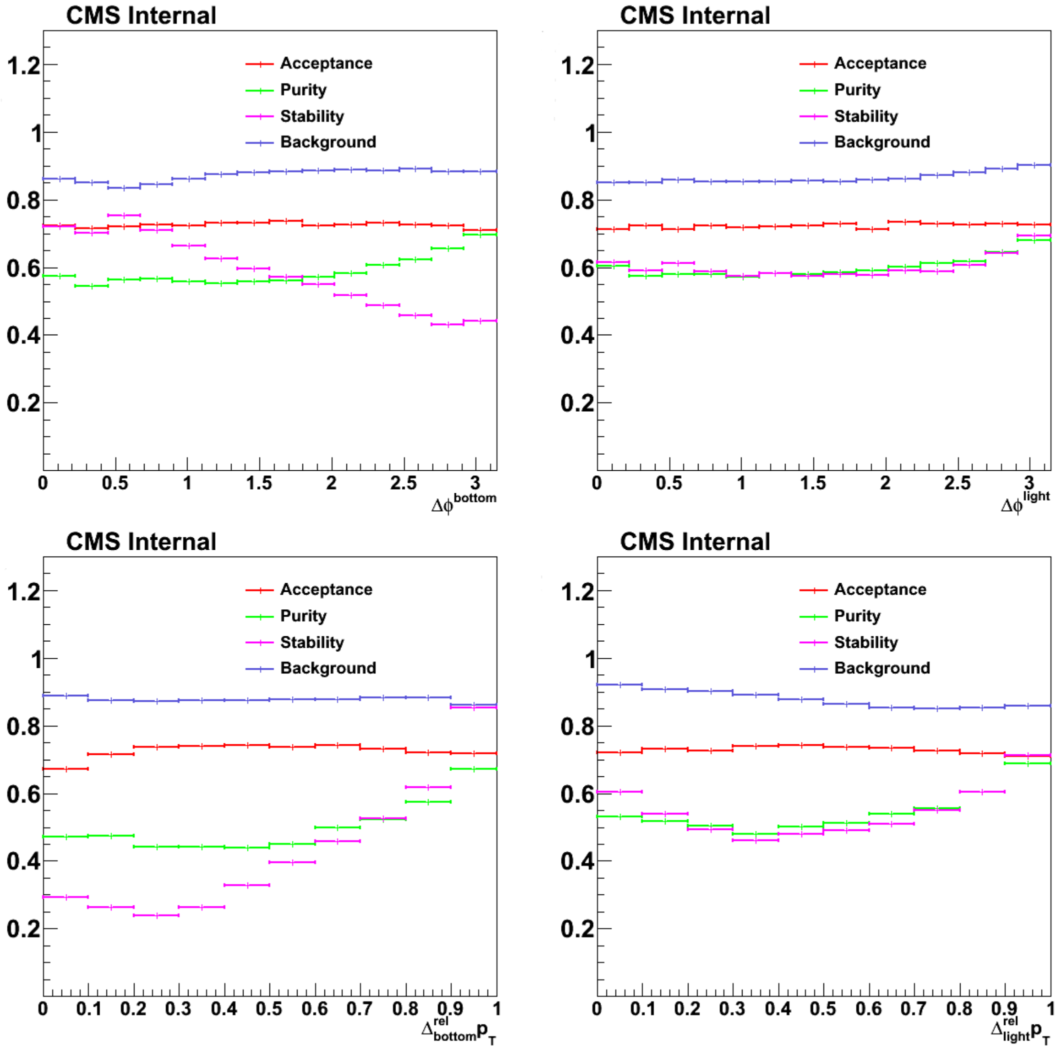


Figure 7.3: Purity, stability, background and acceptance determined for correlation observables in the 2b2j analysis: $\Delta\phi^{bottom}$ *(top left),* $\Delta\phi^{light}$ *(top right),* $\Delta^{rel}_{bottom}p_T$ *(bottom left) and* $\Delta^{rel}_{light}p_T$ *(bottom right). No MC truth information has been used for these measurements. Quantities are measured by using the* PYTHIA*6 sample. No matching has been applied to events at detector and generator level.*

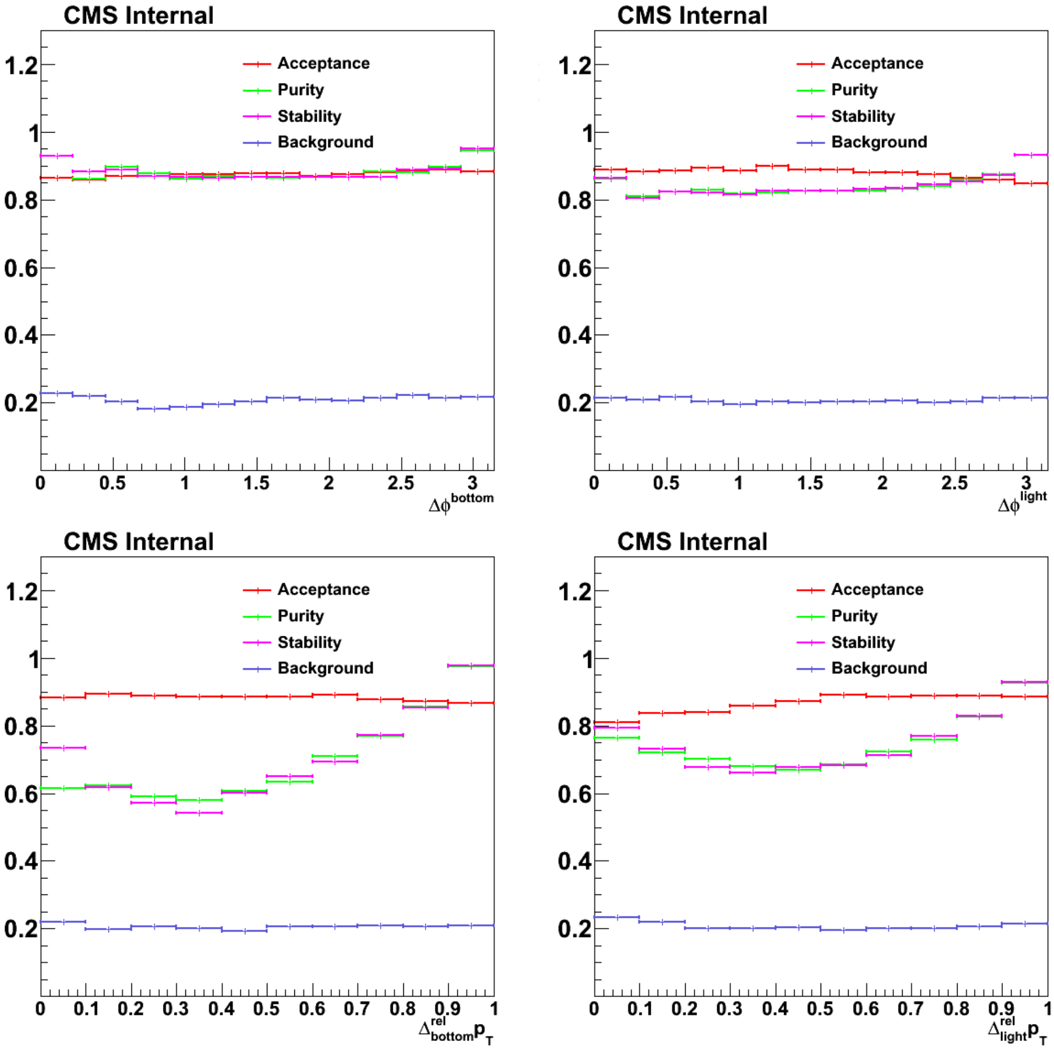


Figure 7.4: Purity, stability, background and acceptance determined for correlation observables in the 2b2j analysis: $\Delta\phi^{bottom}$ (top left), $\Delta\phi^{light}$ (top right), $\Delta^{rel}_{bottom}p_T$ (bottom left) and $\Delta^{rel}_{light}p_T$ (bottom right). MC truth information has been used for these measurements. Quantities are measured by using the PYTHIA*6 sample. No matching has been applied to events at detector and generator level.*

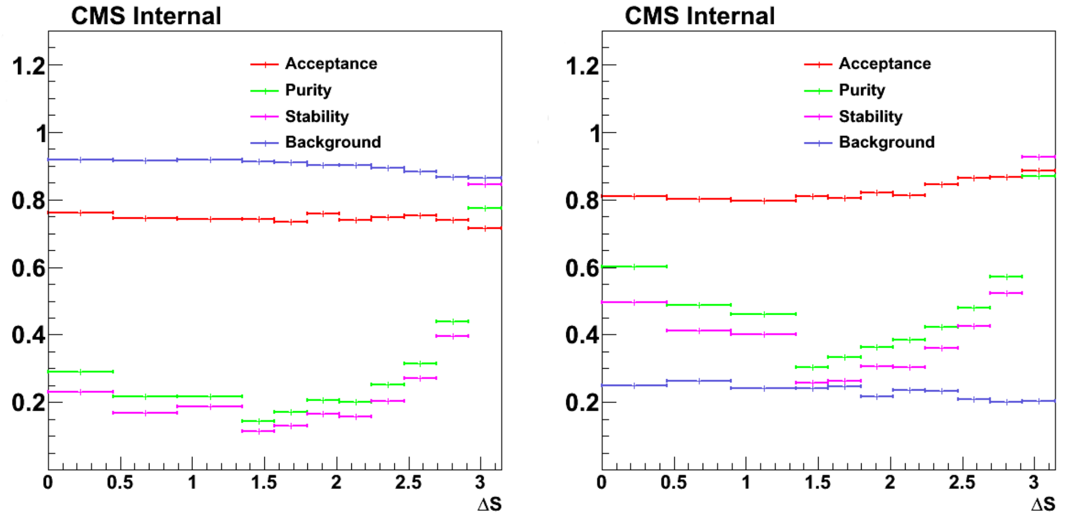


Figure 7.5: Purity, stability, background and acceptance measured for ΔS in the 2b2j analysis. Measurements are shown when MC truth information has not been used (left) and when they have been used (right). Quantities are measured by using the PYTHIA*6 sample. No matching has been applied to events at detector and generator level.*

7.3 Response matrices

Migration effects and correlations between generator and detector level can be also studied and displayed by using the response matrices. They are two-dimensional histograms which represent the same variable at the generator and detector level, for events passing both selections. Response matrices are the necessary inputs for the data unfolding. Figures 7.6 and 7.7 represent the response matrices for some correlation observables and jet p_{T} spectra, for the 2b2j analysis. A perfectly diagonal response matrix means that a complete correlation exists between detector and generator level, namely that every observable, measured in a certain bin at the generator level, is reconstructed in the same bin at the detector level. This situation might appear either, in case of a very precise detector, with high resolution, or with the use of large bins. Indeed, the more diagonal the response matrix is, the better is for the unfolding procedure. The response matrices have been obtained without the use of MC truth information at detector level and with no matching applied for the jets at the two levels. For the correlation observables, the response matrices appear to be rather peaked along the diagonal with low migration effects. Slightly higher migrations are observed at low values of $\Delta^{\mathrm{rel}}p_{\mathrm{T}}$, for back-to-back jet topologies. Jet p_{T} spectra also present a well defined diagonal shape. Note that the p_{T} response matrices are represented in logarithmic scale in the z-axis in order to show more clearly the contents of all bins.

In general, the conclusions extracted from the response matrices confirm what has been observed in the measurements of purity, stability, background and acceptance and they still hold in the 4j scenario.

7.4 B-tag performance

The performance of the b-tag algorithm, which is applied in the analysis, needs also to be studied. The investigation has been carried out in different steps:

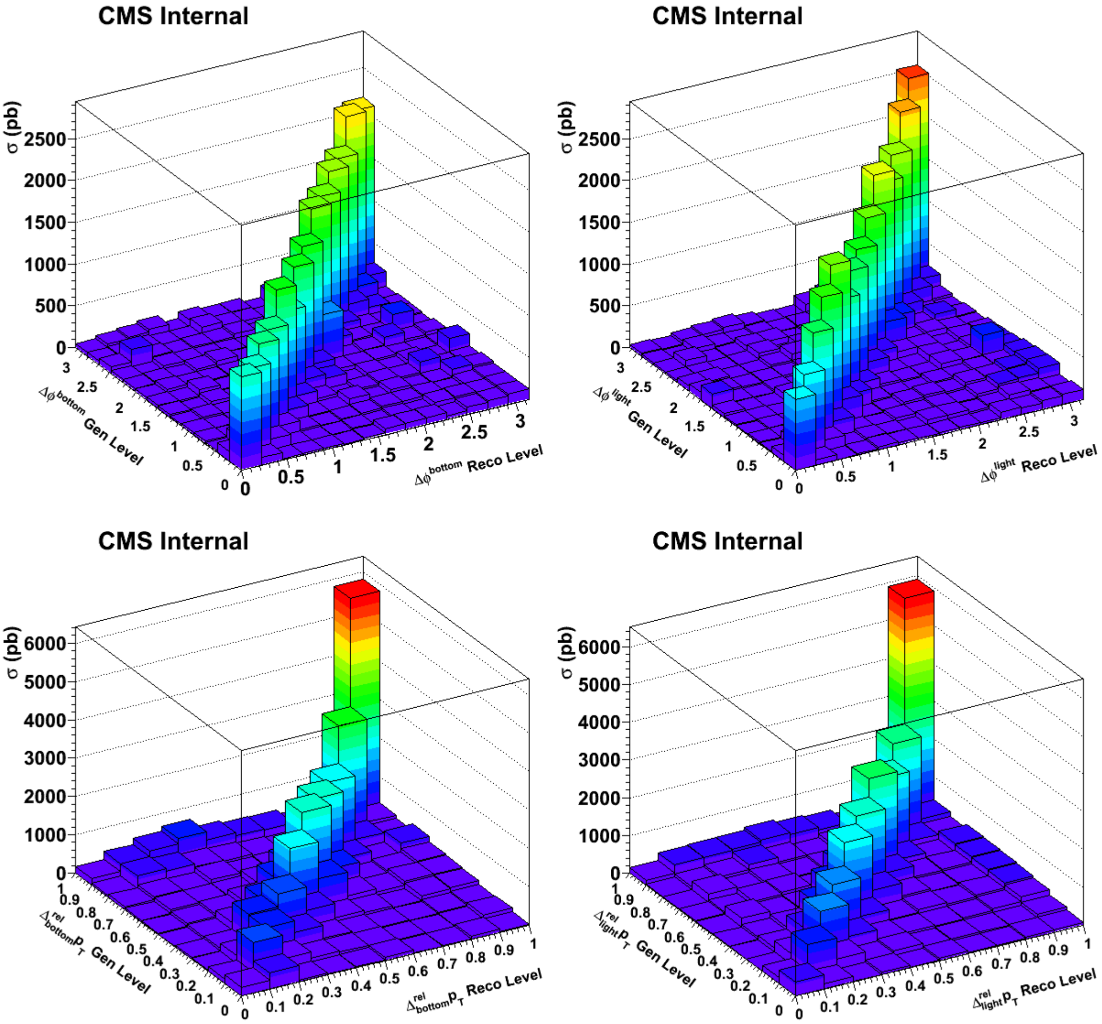


Figure 7.6: Two-Dimensional response matrix measured for correlation observables in the 2b2j analysis: $\Delta\phi^{bottom}$ *(top left),* $\Delta\phi^{light}$ *(top right),* $\Delta^{rel}_{bottom}p_T$ *(bottom left) and* $\Delta^{rel}_{light}p_T$ *(bottom right). Detector and generator level quantities in events, selected in both levels, are represented, respectively, in the x- and y-axis. MC truth information has not been used for these measurements. Quantities are measured by using the* PYTHIA*6 sample. No matching has been applied to events at detector and generator level. The different colours help to display the different heights of the bin contents.*

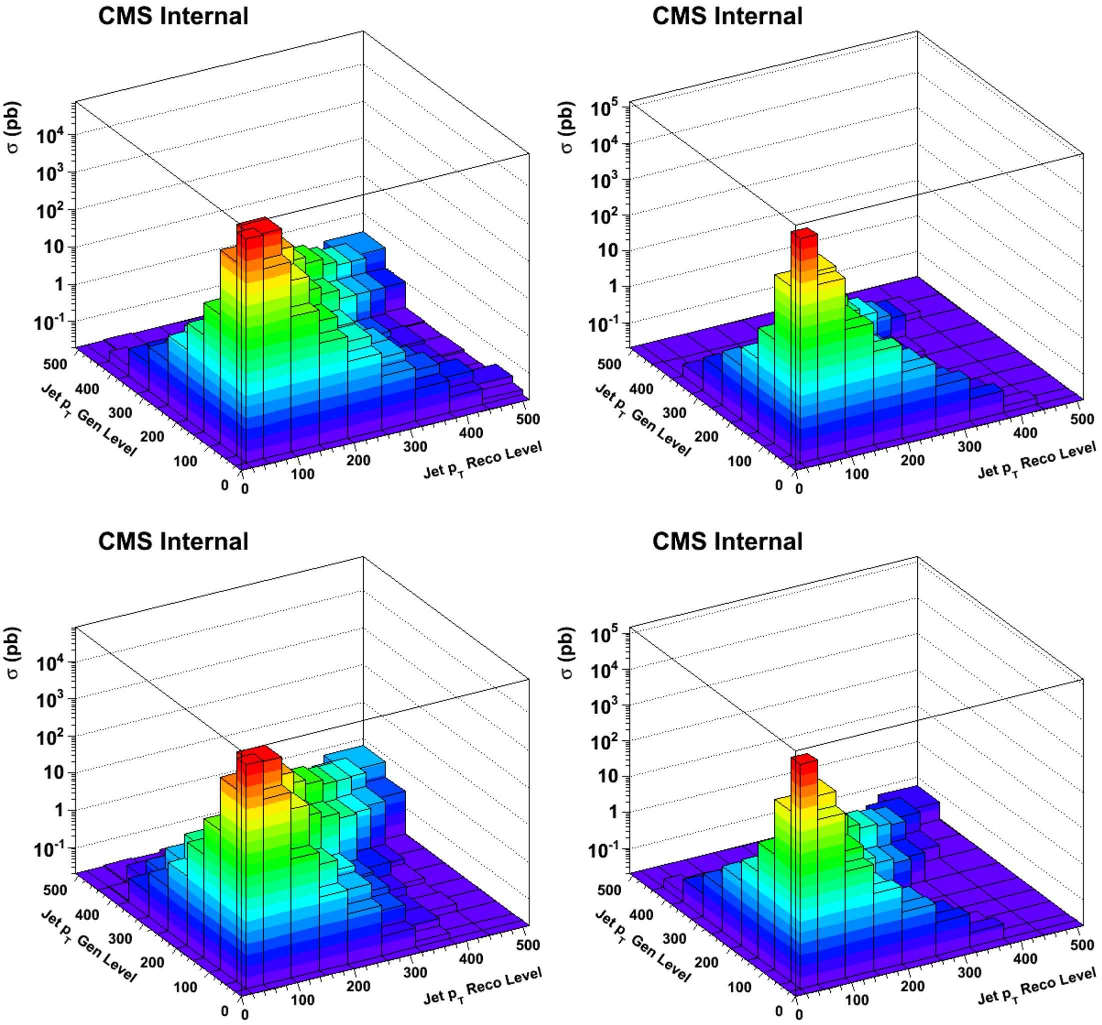


Figure 7.7: Two-Dimensional response matrix measured for jet p_T spectra in the 2b2j analysis: leading and subleading b-tagged jets (top) and leading and subleading additional jets (bottom). Detector and generator level quantities in events, selected in both levels, are represented, respectively, in the x- and y-axis. MC truth information has not been used for these measurements. Quantities are measured by using the PYTHIA*6 sample. No matching has been applied to events at detector and generator level. The different colours help to display the different heights of the bin contents.*

- study of the discriminating power of the b-tag algorithm in the simulation (Section 7.4.1): it shows the distributions of the discriminant variable used in the 2b2j analysis for different jet flavours and investigates the percentage of true b-jets selected after applying the b-tag;
- study of the b-tag performance in the data (Section 7.4.2): since it has been observed [197, 199] that the b-tag performance is different between data and MC, scale factors need to be applied to the simulation in order to correct for this effect. The flavour composition of the sample selected in the 2b2j scenario has been also investigated to compare the obtained purities in MC and in data; this study has been performed by using MC truth information in the simulation and a template fit method in the data;
- determination of the scale factors in a multijet scenario, like the one selected in the 2b2j analysis (Section 7.4.3).

7.4.1 Discriminant power of the b-tag algorithm

As described in Chapter 5, the CSV algorithm is used with the loose working point for the b-tag. This assures a highly efficient selection of b-jets in the final state, but also a mistag rate of about 10% for charm and light jets. These quantities have been measured for the analysis of this thesis, by using the MC sample generated with PYTHIA6. First of all, the CSV discriminant has been studied for different jet flavours. Absolute and normalized cross sections are shown as a function of the CSV output, in Figure 7.8 for bottom, charm, light (which includes up, down and strange) and gluon jets with $p_T > 20$ GeV in $|\eta| < 2.4$. It can be seen that light and gluon jets have a much larger cross section than the others (more than one and two orders of magnitude bigger with respect to, respectively, charm and bottom jets) but the shapes of the CSV discriminant are very different between the various flavours. While light and gluon jets mainly fill the region at low values, charm and bottom jets progressively shift towards higher values. Note the very pronounced peak at 1, which appears for bottom jets and which shows that the CSV output has a very high discriminating power for selection of heavy-flavour jets against light ones. The value of the working point for the loose selection is set to 0.244 [200].

By applying this working point, the percentages of tagged and mistagged jets in the same phase space have been extracted and results are shown in Figure 7.9(top), as a function of jet p_T and η. Different curves are drawn here for various selections:

- Tag: true b-jets with CSV discriminant > 0.244;
- Mistag: true charm or light jets with CSV discriminant > 0.244;
- Miss: true b-jets with CSV discriminant < 0.244;
- Light good ID: true charm or light jets with CSV discriminant < 0.244;

While the distributions are quite flat as a function of p_T, with lower performance for the low region between 20 and 50 GeV, the curves exhibit a strong dependence on jet η. The tag fraction is around 80%, while the mistag rate is about 10% with a small increase up to 20% for the low p_T region. This is in agreement with the definition of the loose working point of the b-tag algorithm (see Chapter 5). By looking at the η dependence, instead, a flat behaviour is observed for the tag fraction, while the identification of light jets dramatically decreases when going to the more

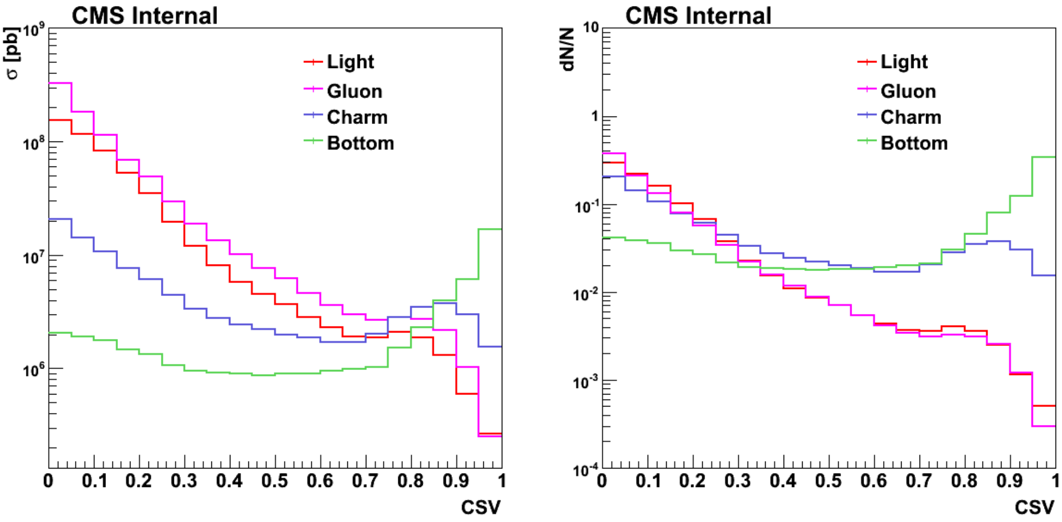


Figure 7.8: Absolute (left) and normalized (right) cross section as a function of the CSV discriminant for each jet flavour.

forward region; the mistag rate for $|\eta| > 2$ is of the order of 35%. This translates into the observation that the tagged jets in this η region have a higher contamination from light and gluon jets.

Finally, an additional study has been carried out about the true flavour of tagged jets, with $p_T > 20$ GeV in $|\eta| < 2.4$. Results are shown in Figure 7.9(bottom), as a function of p_T and η. The overall picture is that most of the tagged jets are light and gluon jets (around 70%); charm and bottom jets share equally the remaining 30%. The percentage of light jets is higher in the p_T region between 20 and 50 GeV, up to 85%, while it is stable at 60% for higher p_T. A similar behaviour is observed as a function of η: in the central region, the fraction of light tagged jets is 60%, while it increases for increasing pseudorapidities up to 85%. This study was necessary to understand the behaviour of the b-tag algorithm in the available simulation and in different regions of the phase space, besides keeping under control any detector effect during the data correction.

7.4.2 Studies of the scale factors

A similar study needs to be performed in a data-driven way, in order to check if the performance observed in the simulation is compatible with data. However, this study in data is much more complicated since the MC truth information related to the true flavour of the jets is not available. The performance of the b-tag algorithms in data has been measured in CMS and is documented in [197, 199], for two different periods of data taking. They show that small differences arise when comparing results for data and simulation, which need to be corrected for. The corrections are performed through the application of Scale Factors (SF), which depend on p_T, η, and flavour of the jet, as well as on the b-tag algorithm and on the specific working point. They have been determined by taking the ratio of the b-tagging efficiencies for each jet flavour, measured in data and MC; they consist of weights to be applied in the simulation in inclusive events with tagged jets and they only account for matching the b-tag performance in data and MC.

For the 2010 data, used in this analysis, the SF have not been directly measured for the CSV algorithm; nevertheless, by observing that the performance has not changed for the other

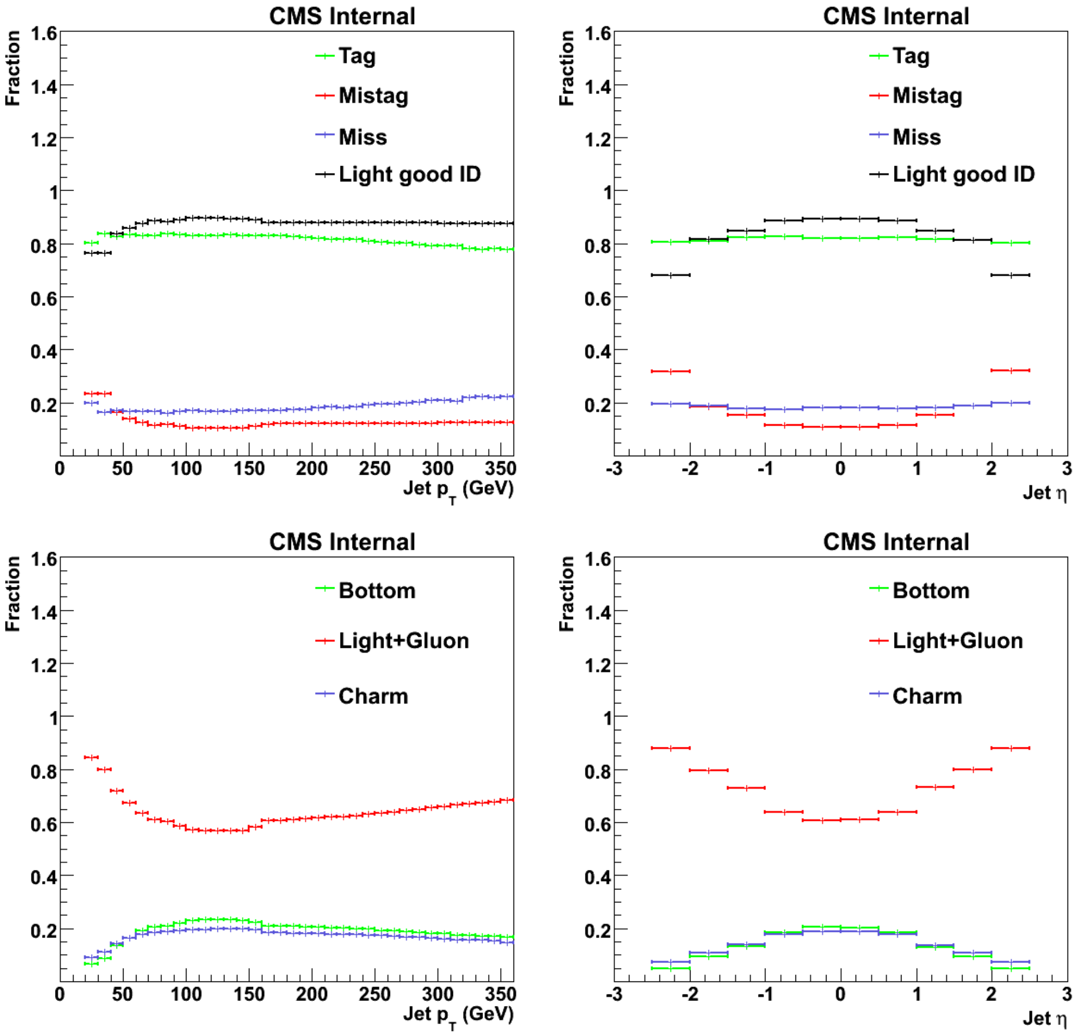


Figure 7.9: (Top) Fraction of tag, mistag, miss and good identified jets as a function of p_T *(left) and* η *(right). (Bottom) Fraction of jets of different flavours as a function of* p_T *(left) and* η *(right).*

discriminants [198] from results obtained with data recorded in 2011 [197], it is appropriate to use the SF measured with that sample, with some minor variants [209]:

- the uncertainty is increased to 10% (versus the 2011 uncertainty of around 5%) in a conservative approach;
- when measuring jets with $p_{\mathrm{T}} < 30$ GeV[1], the SF are kept equal to the one for $p_{\mathrm{T}} = 30$ GeV, assuming a small variation of the b-tag performance in the p_{T} range 20–30 GeV for all flavours.

However, the SF are tested and well-performing only for Pythia6 MC samples: for other models, the SF are no longer valid[2]. The SF, measured for the CSV algorithm, when using the loose working point, are given by the following equations:

$$SF_{\mathrm{b}} = 0.60 \cdot \left(\frac{1 + 0.29 \cdot x}{1 + 0.17 \cdot x} \right); \tag{7.5}$$

$$SF_{\mathrm{c}} = 0.60 \cdot \left(\frac{1 + 0.29 \cdot x}{1 + 0.17 \cdot x} \right); \tag{7.6}$$

$$SF_{\mathrm{l}} = 1.08 + 1.75 \cdot 10^{-4} \cdot x - 8.63 \cdot 10^{-7} \cdot x^2 + 3.28 \cdot 10^{-10} \cdot x^3 \quad \text{for jets in } |\eta| < 0.5 \tag{7.7}$$

$$SF_{\mathrm{l}} = 1.08 + 3.24 \cdot 10^{-4} \cdot x - 1.30 \cdot 10^{-6} \cdot x^2 + 8.51 \cdot 10^{-10} \cdot x^3 \quad \text{for jets in } 0.5 < |\eta| < 1.0 \tag{7.8}$$

$$SF_{\mathrm{l}} = 1.08 + 4.75 \cdot 10^{-4} \cdot x - 1.44 \cdot 10^{-6} \cdot x^2 + 1.13 \cdot 10^{-9} \cdot x^3 \quad \text{for jets in } 1.0 < |\eta| < 1.5 \tag{7.9}$$

$$SF_{\mathrm{l}} = 1.06 + 1.74 \cdot 10^{-4} \cdot x - 5.29 \cdot 10^{-7} \cdot x^2 + 5.56 \cdot 10^{-10} \cdot x^3 \quad \text{for jets in } 1.5 < |\eta| < 2.4 \tag{7.10}$$

where x is the jet p_{T} expressed in GeV. Note that the SF for charm and bottom have no dependence on η, as the curves observed in Figure 7.9 show, while SF for light jets are divided in four η bins to take into account the dependence on the pseudorapidity. An attempt of cross check of the reliability of the SF for this analysis has been performed with a template method. The template method aims for the determination of the flavour composition of the sample selected in the data. The CSV discriminants of leading and subleading tagged jets have been considered in data after the 2b2j selection. Templates of CSV discriminants, separately for jets of different flavours[3], have been built with the Pythia6 sample, after applying the previous SF. The effect of the application of the SF is that each tagged jet which fills a generic histogram contributes with a different weight

---

[1] Note that a jet p_{T} threshold of 30 GeV has been applied for the measurement of the SF in [197]. No corrections have been measured for lower jet p_{T}.

[2] In this analysis, they are anyway applied to the Herwig++ sample but only for comparison, and this sample has not been used for the final correction of the data.

[3] Light and gluon jets are considered within the same template because the corresponding shapes of the CSV discriminant are very similar between each other (see Figure 7.8).

according to its p_{T} η and flavour. Then, the distribution in data is fitted with three templates for bottom, charm and light+gluon jets[1]. In output, three numbers representing the percentage of each flavour which best fits the data are given; the sum of the these fractions returns 1. The fitted distributions are represented in Figure 7.10(top), for leading and subleading tagged jets; the templates for each flavour are scaled according to the corresponding output fraction. These fractions are then compared to the purities measured in the MC sample, given by the true flavours of the selected jets. The results are shown in Figure 7.10(bottom); the bottom and charm fractions are around 15–18%, while most of selected jets, 65–70%, are light. Note the very good compatibility between results obtained from data and simulation. A QCD sample has an uncertainty of about 20% [198] on describing properly the purity in data and the obtained results are well in agreement with the simulation. From these studies, it might be concluded that the official SF are reliable for a 2b2j scenario and they can be applied for the correction of the detector effects due to the b-tag algorithm.

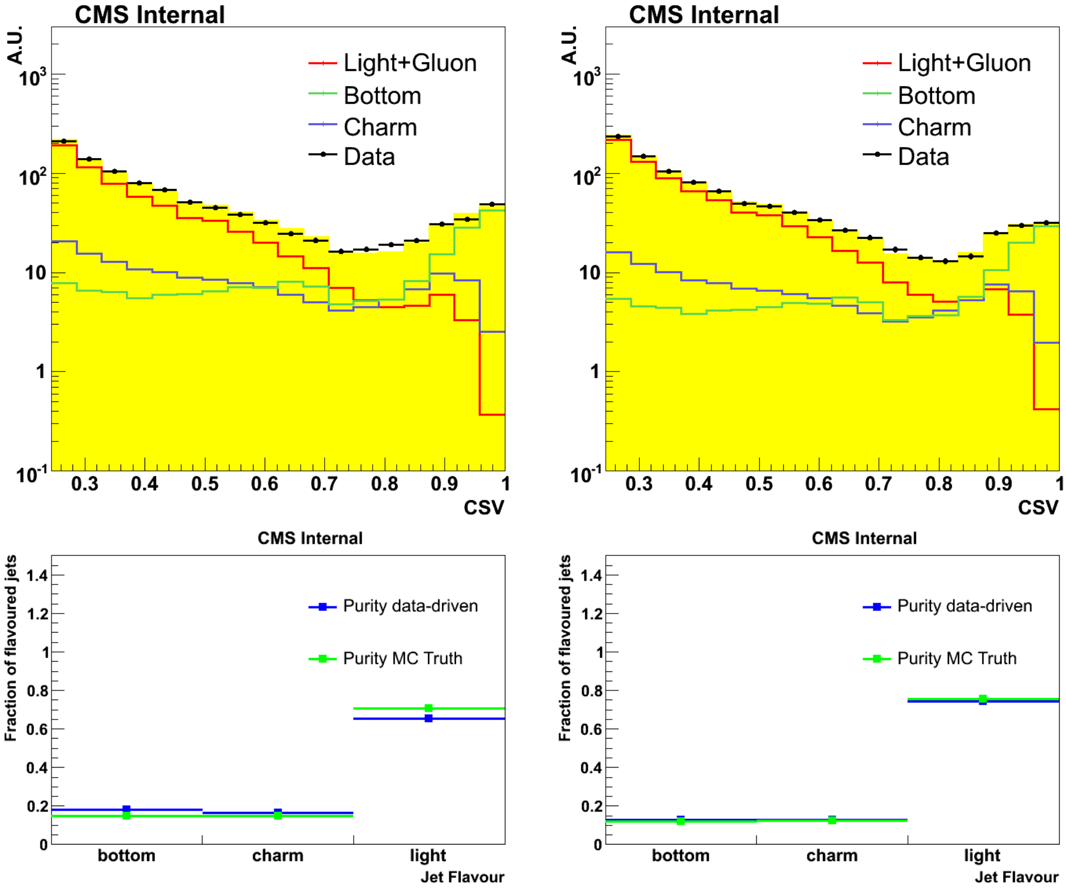


Figure 7.10: (Top) Distributions of CSV discriminant in data fitted with templates of different jet flavours for leading (left) and subleading (right) tagged jets: in yellow, the fitted distributions are shown, while in black are the data points. The templates for each flavour are scaled to the corresponding flavour fractions obtained from the fit. (Bottom) Fractions of the different jet flavours obtained from data, with the template method, and in the simulation from MC truth information, for leading (left) and subleading (right) tagged jets.

---

[1] The fit has been performed with the TFractionFitter tool, implemented in ROOT [171]

7.4.3 Applying the Scale Factors to multijet scenarios

The SF, calculated in the previous Section, refer to inclusive selections of tagged jets: in these cases, they need to be applied to every tagged jet, measured in the selected events. Specifically, for inclusive jet distributions, all tagged jets in the simulation are reweighted through a SF chosen from Equations 7.5-7.10 according to their p_{T}, η and flavour. In CMS, there are several approaches [210] of how to propagate and apply the SF in multijet scenarios; these include, for instance, the possibility to have more than one tagged jet, to have multiple working points, or both tagged- and non-tagged jets in the same selected final state. For the multijet selection used in the analysis of this thesis, a combination of SF for tagged and non-tagged jets is required and a weight for each selected event is applied to the MC samples. The inputs for the calculation of the weights are the SF, provided by Equations 7.5-7.10 as a function of p_{T}, η and flavour, the kinematical quantities (p_{T}, η) of each jet selected in the final state and the b-tag algorithm efficiency for the MC sample. While the SF have been measured in other analyses [197, 199] and are available for each period of data taking, the b-tag efficiency needs to be determined for the considered MC samples. In particular, it is important that the b-tag efficiencies are matched with the SF of Equations 7.5-7.10: they need to be obtained separately for the different jet flavours (bottom, charm and light), differentially in p_{T} and η (only for light jets) with the same binning used for the SF. The efficiency is determined by taking the bin-by-bin ratio of the inclusive cross section of jets of a certain flavour whose b-tag discriminant is above the working point, divided by the inclusive cross section, measured for these jets. For this analysis, this determination has been carried out for the PYTHIA6 and the HERWIG++ samples. Results for the two generators are shown in Figure 7.11, as a function of jet p_{T}: they show that the highest efficiency is achieved for true b-jets over the whole range, and it is around 75–85% depending on the jet p_{T}. The efficiency increases from 20 GeV and reaches a maximum at around 100 GeV, before again decreasing for higher transverse momenta: this behaviour of the efficiency curve has already been observed in [207]. At low p_{T}, the efficiency rises because the B-hadron lifetime increases, consequently giving a better discrimination of the secondary vertex. Going to higher p_{T}, the jets become boosted toward their direction and their tracks start to be more collinear; in this configuration, tracking inefficiencies rise in the dense core and they imply a decrease of the secondary vertex reconstruction, and hence of the b-tag efficiency. The mistag for charm jets is higher at low transverse momenta, around values of 45–50%, and decreasing down to 35%–40% from $p_{\mathrm{T}} > 70$ GeV: the higher mistag values at lower p_{T} are mainly due to tuning effects of the CSV discriminant, which has been optimized at high transverse momenta [211]. Contamination from light jets is also studied in different pseudorapidity ranges: the p_{T} dependence is very similar to the c-jets in each η range. The mistag efficiency increases when going to more forward pseudorapidities: in the most central region ($|\eta| < 0.5$), it is between 5 and 15% while for the most forward η bin ($1.5 < |\eta| < 2.4$) it increases up to 20–35%. This feature is seen in the shape in Figure 6.12(top) for η of the tagged jets, due to higher contamination of light jets at forward rapidities. The values and behaviour of the efficiency curves are very similar when measured with PYTHIA6 or HERWIG++.

With the SF in Equation 7.5-7.10 and the measured b-tag efficiencies, it is then possible to calculate the event weights w to apply to MC events; the weights are calculated under the assumption that the tagging efficiencies of the different selected jets, both tagged and non-tagged, factorize between each other. The whole procedure is documented in [212] and it is based on the following

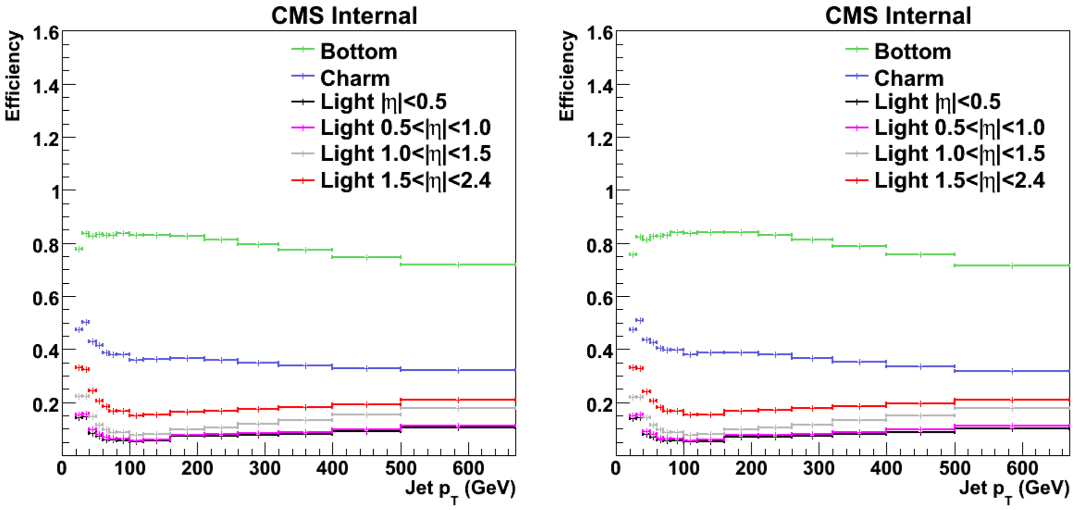


Figure 7.11: B-tag efficiencies as a function of p_T, measured with Pythia*6 (left) and* Herwig*++ (right) for different jet flavours: efficiencies for light jets are measured in four different η bins, according to the different SF available for them.*

quantities:

$$P(MC^{\text{gen}}) = \prod_{\text{i=tagged}} \epsilon_{\text{i}} \prod_{\text{j=not-tagged}} (1 - \epsilon_{\text{j}}) \tag{7.11}$$

$$P(MC^{\text{det}}) = \prod_{\text{i=tagged}} SF_{\text{i}}\epsilon_{\text{i}} \prod_{\text{j=not-tagged}} (1 - SF_{\text{j}}\epsilon_{\text{j}}) \tag{7.12}$$

$$w = \frac{P(MC^{\text{det}})}{P(MC^{\text{gen}})} \tag{7.13}$$

where i and j represent, respectively, the tagged and the non tagged jets required and selected in every event; the quantities ϵ and SF are evaluated according to the p_{T} and η of the considered jet. The weights w are then applied to MC simulations and the normalized differential cross sections of the kinematical jet variables, measured in the data, are compared to the new predictions. Results are shown in 7.12 as a function of each jet p_{T}. The comparisons between data and simulation still show some discrepancies in the high p_{T} region of the order of 30–40% for all jet spectra. It has been observed [213] that SF for single jets or multijet scenarios have an impact only in the absolute cross section measurements, while the shapes of the cross sections themselves do not significantly change. The overall effect on absolute cross sections as a function of jet p_{T} is a shift below of the order of 5–8%.

Since a not yet satisfying agreement is achieved after the application of the SF, a further correction of the MC is attempted for a better description of the data. Since it is observed that the p_{T} spectrum is not well described and it is known that a LO 2→2 ME, as implemented in the considered generators, is not sufficient to describe a multijet scenario, a further reweighting procedure has been applied. It assumes that the leading tagged jet p_{T} is equal to the transverse momentum, $\hat{p}_{\text{T}}$, of the partons of the hard scattering[1]. With this hypothesis, a reweighting factor,

---

[1] This assumption is indeed not very precise and sophisticated, since it neglects effects like parton evolution, additional hard emissions, etc., but it is a good starting point since it connects a parton level quantity with a detector level one. Furthermore, such reweighting procedure, as implemented here, may be applied iteratively if after the first iteration, a not satisfying description is obtained.

equal to the inverse of the ratios of the lower panel of Figure 7.12(top left), is calculated and is applied in the simulation according to the $\hat{p}_{\mathrm{T}}$ of the hard scattering. The weights quantify the discrepancy between data and MC on the leading tagged jet. Indeed, the two different reweighting factors are applied for PYTHIA6 and HERWIG++. Qualitatively, what it attempts to do, is to correct the p_{T} spectra of the MC to the ones in the data. Since the ratios are below 1 at high p_{T}, the net effect is to make the simulated p_{T} spectrum harder, going towards the direction of the data. The same factors are applied to all p_{T} spectra, along with all the other measured variables. Results after reweighting are shown in Figure 7.13 and 7.14, as a function of the jet p_{T} and jet η, respectively. The description provided by the simulation significantly improves after reweighting: the predicted spectra of the tagged jets show discrepancies less than 5%. The spectra of the additional jets are also very close to the data, with little differences up to 15%. The normalized differential cross sections as a function of the jet η are well predicted by the simulation, with differences due to the effects described in Chapter 6, in the forward region. All the other measured correlation observables are nicely reproduced by the available MC samples. This level of agreement is satisfying to give a reliable correction of the data at the stable particle level. This is performed with the unfolding procedure and is described in Chapter 8.

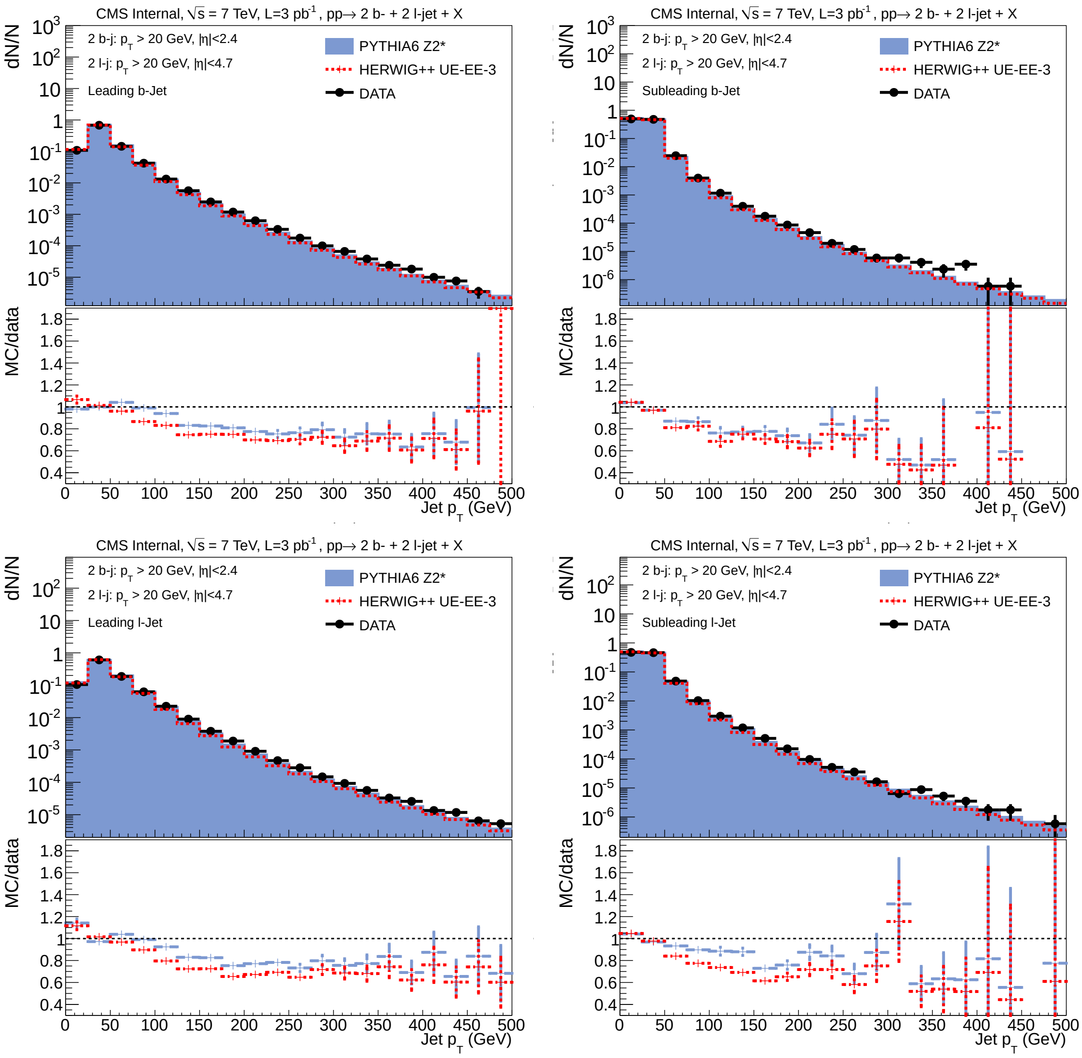


*Figure 7.12: Comparison of data with simulations from MC event generators (*Pythia6 *and* Herwig++*) for the transverse momentum of the leading (top left), subleading (top right) tagged jets, and leading (bottom left) and subleading (bottom right) additional jets. The b-tag scale factors are applied to the simulation in order to match the b-tag performance with the data. Only statistical uncertainties are shown for the data. Data and simulation are not normalized to the bin width. The lower panel shows the ratios of the theoretical predictions to data.*

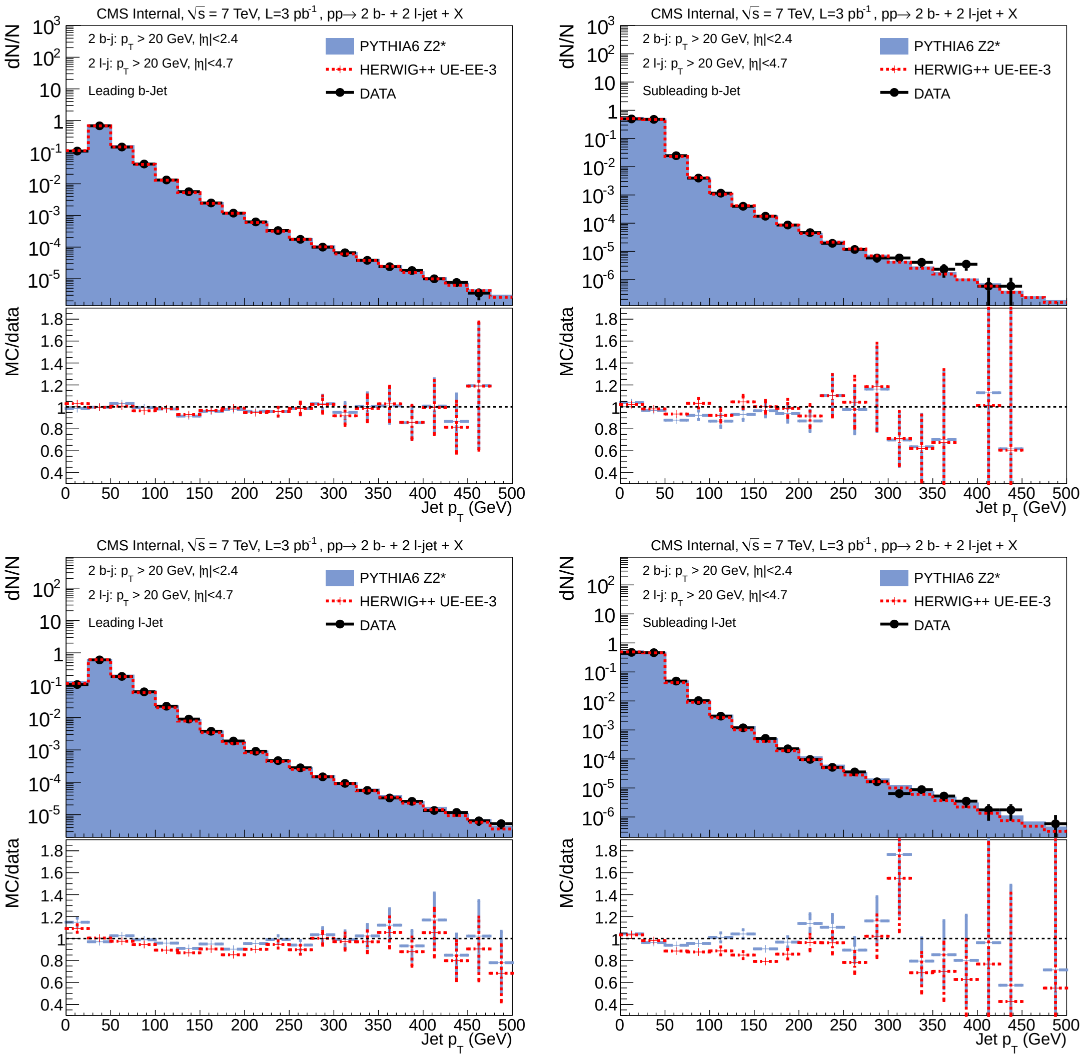


*Figure 7.13: Comparison of data with simulations from MC event generators (*Pythia*6 and* Herwig++*) for the transverse momentum of the leading (top left), subleading (top right) tagged jets, and leading (bottom left) and subleading (bottom right) additional jets. The b-tag scale factors are applied to the simulation in order to match the b-tag performance with the data. The reweighting based on the $\hat{p}_T$ of the hard scattering is also applied to the simulation. Only statistical uncertainties are shown for the data. Data and simulation are not normalized to the bin width. The lower panel shows the ratios of the theoretical predictions to data.*

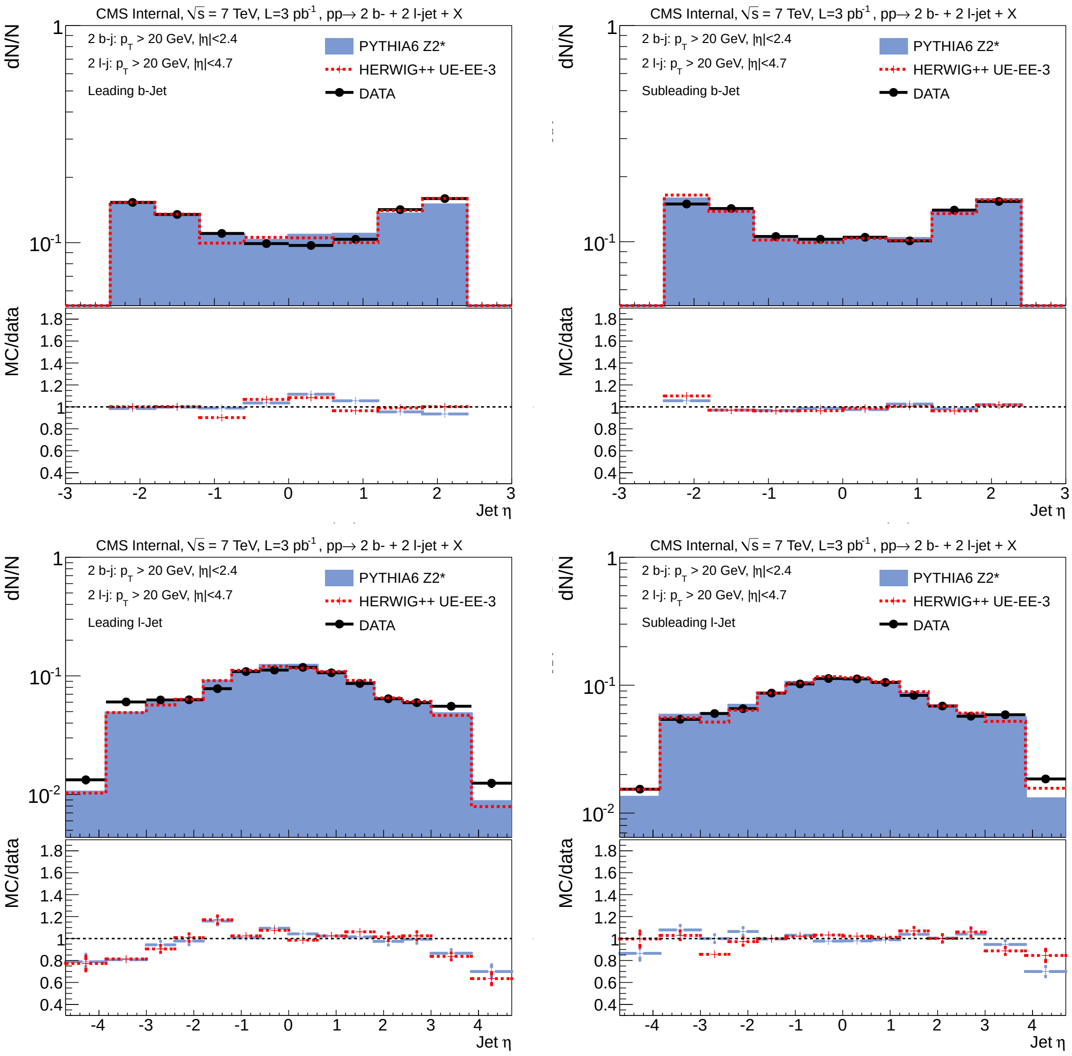


*Figure 7.14: Comparison of data with simulations from MC event generators (*Pythia6 *and* Herwig++*) for the pseudorapidity of the leading (top left), subleading (top right) tagged jets, and leading (bottom left) and subleading (bottom right) additional jets. The b-tag scale factors are applied to the simulation in order to match the b-tag performance with the data. The reweighting based on the $\hat{p}_T$ of the hard scattering is also applied to the simulation. Only statistical uncertainties are shown for the data. Data and simulation are not normalized to the bin width. The lower panel shows the ratios of the theoretical predictions to data.*

Chapter 8

Data unfolding

Comparisons between data and theoretical predictions need to be performed in a consistent way: either, detector effects are added to any model predictions and curves are compared at detector level, or data are corrected to the stable particle level, where detector effects are removed; any event generator which includes parton evolution, hadronization and UE simulation, produces predictions at this level. The first approach is inconvenient, because it implies that any model needs to be interfaced to the specific detector simulation, which might be complicated and is usually private within collaborations. The second method is, instead, much more flexible and straightforward: by correcting the data for the ensemble of the detector effects, it is possible to use them for comparisons with predictions from any model. This is why the latter analysis strategy is preferred: the procedure of correcting the data to the stable particle level is generally called "unfolding". The name refers to the fact that the data are interpreted as a true "level" folded with the detector response: the unfolding specifically aims for eliminating this last component (see Section 8.1). The detector components include resolution, efficiency and acceptance effects, as described in full detail in Chapter 7.

The unfolding procedure is determined by the two levels, detector and generator, and various methods can be used for the correction. The detector level coincides with the set of analysis cuts applied to physics objects selected by using exclusively detector information, while the stable particle level contains the selection of "true" objects before detector interface. For each of the two analyses considered in this thesis, the two levels have been defined. The detector level requirements are listed in Tables 6.6 and 6.7, respectively for the 4j and the 2b2j channels. The generator level definitions are specified in Tables 8.1 and 8.2, for the two selections. The latter is implemented, in particular, in the corresponding RIVET [185] plugin used for obtaining the theoretical predictions.

Exactly four jets clustered with anti-k_T algorithm Jets clustered with neutral and charged stable particles A stable particle is defined as a particle with $c\tau > 10$ mm		
Two hard-jets	$p_\mathrm{T} > 50$ GeV	$\lvert\eta\rvert < 4.7$
Two soft jets	$p_\mathrm{T} > 20$ GeV	$\lvert\eta\rvert < 4.7$

Table 8.1: Particle level definition applied to data and MC in the 4j analysis.

At least four jets clustered with anti-k_{T} algorithm		
Jets clustered with neutral and charged stable particles		
A stable particle is defined as a particle with $c\tau > 10$ mm		
At least two jets initiated by a primary or secondary b-quark		
Two b-jets	$p_{\mathrm{T}} > 20$ GeV	$\lvert\eta\rvert < 2.4$
Two additional jets[1]	$p_{\mathrm{T}} > 20$ GeV	$\lvert\eta\rvert < 4.7$

Table 8.2: Particle level definition criteria applied to data and MC in the 2b2j analysis.

Unfolding is performed by using MC samples which include full detector simulation and by considering the response matrices shown in Chapter 7. In order to obtain a proper correction to the stable particle level, the MC simulation needs to reproduce well the behaviour of the measured distributions: this requirement is fulfilled by the two considered samples, generated with Pythia6 and Herwig++. By studying the same event before and after detector simulation, it is indeed possible to study the migration effects and apply them to the measured quantities. This enables to obtain the "true" level of the data. The entire picture is schematically represented in Figure 8.1.

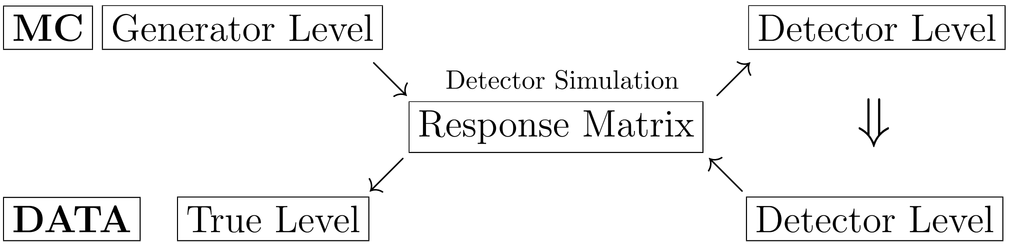


Figure 8.1: Representation of a generic method of unfolding: the response matrix is obtained from the simulation where both generator and detector level distributions are available. The detector level is obtained by interfacing the distributions at the generator level with a full detector simulation. The response matrix is then applied to the output from data in order to get the true distributions, corresponding to the measured underlying physics.

This chapter is dedicated to the explanation of the possible unfolding methods which are generally used for an analysis in particle physics. The results are then shown for the multijet analyses.

8.1 The different unfolding methods

The unfolding procedure corresponds to the solution of a matrix equation which connects detector and stable particle level. The fact that it is a multi-dimensional equation, rather than a single-dimensional one, stems from the fact that one generally considers binned distributions, measured differentially as a function of a specific observable. The number of rows, n, and columns, m, of the matrix depends on the number of bins chosen for the histograms. The matrix equation translates into a system of n linear equations, corresponding to an equation for each bin of the detector level distribution. Calling $\hat{A}$ the response matrix, $\vec{x}$ the vector of true distribution and $\vec{b}$ the vector of

---

[1] A further requirement about the flavour of the additional jets has been set: only light and charm jets are selected at the stable particle level while b-jets are not considered for this analysis. The impact of these excluded events has been evaluated and consists of 4–5% of the total cross section for this scenario.

the measured observable, one can write[1]:

$$\hat{A}\vec{x} = \vec{b} \tag{8.2}$$

The unknown in the unfolding problem is the vector $\vec{x}$; the vector $\vec{b}$ is the output of the measurement and the matrix $\hat{A}$ is determined by MC simulation. In order to extract the unknown, different methods are available. The most intuitive way is the analytical inversion of the response matrix: by multiplying Equation 8.2 from the right side by $\hat{A}^{-1}$, one gets $\vec{x} = \hat{A}^{-1}\vec{b}$. This method has to face two rather serious problems: the first one depends on the fact that there are no mathematical constraints which guarantee that a response matrix must admit an inverse but it might also be singular. In this case, the inversion method does not lead to any meaningful solution. Furthermore, if the inverse of the matrix exists, a second additional problem also contributes: since the response matrix is determined by MC simulation, some bins might be affected by statistical fluctuations, due to limited statistics. These issues affect the method resulting in unstable solutions [214]. Hence, different approaches have been developed, which are easy to implement and applicable in any scenario; they go from simple bin-by-bin corrections, to more sophisticated procedures, like the D'Agostini or the SVD methods. All of them are presented in the following.

8.1.1 Correction Factor method

The Correction Factor Method (CFM) [215] uses a bin-by-bin correction and is rather simplistic. In fact, it takes into account only the diagonal elements of the response matrix. Thus, migration effects within the phase space are considered without treating the correlations between adjacent bins. Miss and fake events are respectively added and subtracted to the total number of detected events, to account for migrations into or out of the phase space. The assumption made by the CFM method holds as long as the response matrix is diagonal, or at least the off-diagonal coefficients are much smaller than the ones along the diagonal. The bin-by-bin correction uses factors, which depend only on the fraction of events that are selected in a specific bin at detector and generator levels. In the case of absence of background, the solution becomes straightforward and the unfolded value for each bin i can be defined as:

$$x_{\mathrm{i}} = \frac{x_{\mathrm{i}}^{\mathrm{MC}}}{b_{\mathrm{i}}^{\mathrm{MC}}} \cdot b_{\mathrm{i}} = C_{\mathrm{i}} \cdot b_{\mathrm{i}} \tag{8.3}$$

where $x_{\mathrm{i}}^{\mathrm{MC}}$ and $b_{\mathrm{i}}^{\mathrm{MC}}$ are the binned distributions, respectively, measured at detector and generator level in the considered MC sample. Note that the same binning needs to be set for detector and generator level distributions. The ratios between the two numbers are the correction factors to be applied to the considered bin i. The treatment of the uncertainties is also very straightforward; a value equal to $C_{\mathrm{i}}\sqrt{b_{\mathrm{i}}}$ is assigned to the unfolded value x_{i}, assuming a poissonian distribution for the bin contents of the histograms. An additional underlying assumption is also that the correction factors must be determined for each bin with large statistics in the MC samples: this enables to

---

[1] In the continuous space, Equation 8.2 becomes an integral equation which can be written, with a similar notation, as:

$$B(y) = \int_{y_0^{\mathrm{true}}}^{y_{\mathrm{n}_x}^{\mathrm{true}}} dy^{\mathrm{true}} A(y, y^{\mathrm{true}}) X(y^{\mathrm{true}}) \tag{8.1}$$

The detector level quantity is the convolution of the real generator level distribution with the detector effects, parametrized by a response matrix which depends on the true value of the considered observable, y^{true}, and on the measured one, y. The integral formulation is not considered in the following, since we always deal with binned distributions take them without uncertainty.

For the presented analyses, it has already been observed that some of the measured observables are affected by large migration effects inside the phase space; in presence of a slight disagreement between data and MC in the distributions at the detector level, this makes the use of CFM quite inappropriate and unreliable. Hence, this method has only been used as an initial cross check for the presented analyses and other methods have been preferred.

8.1.2 D'Agostini unfolding method

The D'Agostini method [216] has been elaborated in 1995 and it is based on a Bayesian approach of the unfolding problem. Besides many other features, a peculiarity of the D'Agostini method is the possibility of treating successfully multidimensional problems. The method is based on Bayes' theorem, which rules the conditional probability of a certain effect to happen from a given cause. This is the reason why the D'Agostini method is also referred to as "Bayesian" method. One can write the number of events $\hat{n}(C_{\mathrm{i}})$[1] produced by a specific cause (namely, by a specific physics process), as a product of the total number of measured events $\hat{n}(E_{\mathrm{j}})$, that are the events which produce a specific effect E_{j}, and the conditional probability of the cause C_{i}, to result in the given effect, E_{j}. It can be written in a compact way, as the following:

$$\hat{n}(C_{\mathrm{i}}) = \frac{1}{\epsilon_{\mathrm{i}}} \sum_{\mathrm{i}=1}^{\mathrm{n_E}} n(E_{\mathrm{j}}) P(C_{\mathrm{i}}|E_{\mathrm{j}}) \tag{8.4}$$

where ϵ_{i} refers to the selection efficiency and it can be interpreted as the probability of C_{i} to produce a final state E_{j} which fulfills the analysis cuts. Note that $\hat{n}(C_{\mathrm{i}})$ and $\hat{n}(E_{\mathrm{j}})$ may refer to values of bin contents in case of a binned distribution. The Bayes' theorem is invoked now in order to determine the quantity $P(C_{\mathrm{i}}|E_{\mathrm{j}})$[2], which is unknown. In fact, one can write:

$$\sum_{\mathrm{i}=1}^{\mathrm{n_c}} P(C_{\mathrm{i}}|E_{\mathrm{j}}) = 1 \tag{8.5}$$

$$P(C_{\mathrm{i}}|E_{\mathrm{j}}) = \frac{P(E_{\mathrm{j}}|C_{\mathrm{i}}) P_0(C_{\mathrm{i}})}{\sum_{\mathrm{l}=1}^{\mathrm{n_c}} P(E_{\mathrm{j}}|C_{\mathrm{l}}) P_0(C_{\mathrm{l}})} \tag{8.6}$$

where the $P(E_{\mathrm{j}}|C_{\mathrm{i}})$ is the conditional probability to obtain the effect E_{j} from the given cause C_{i} and $P_0(C_{\mathrm{i}})$ is the so-called "prior" distribution, which indicates the initial probability of the cause. The denominator is just a normalization factor. The condition, expressed by Equation 8.5, represents the fact that each effect must come from one or more of the considered causes. The quantity $P(E_{\mathrm{j}}|C_{\mathrm{i}})$ is obtained with MC simulation and specifically corresponds to the forementioned response matrix, while the prior $P_0(C_{\mathrm{i}})$ needs to be initially chosen (hence, "prior" like "a priori"), based on the knowledge of the considered cause; in the worst case, of total lack of knowledge of the cause, a flat distribution might be also chosen. However, generally, the distribution observed at the generator level in the simulation is set as initial prior; this might not necessarily be the true distribution but it is definitely more physical than a flat spectrum. The power of the D'Agostini method is that this procedure can be iterated by including the obtained distribution as new input

---

[1] Here, it might seem that the notation has changed but the value $\hat{n}(C_{\mathrm{i}})$ is just the sum of the number of entries of the previously defined vector $\vec{x}$. The new notation is convenient for expressing the concepts of cause and effect.

[2] Note that $P(C_{\mathrm{i}}|E_{\mathrm{j}})$ contains information of the background and the fake rates, which are thus taken into account in the unfolding procedure.

for the prior; the more iterations are performed, the higher the knowledge of the physical process is. The iterations can stop when the output results are stable; this is quantified by the χ^2 between two successive iterations in this way:

$$\chi^2 = \sum_{\mathrm{i}=1}^{\mathrm{N}_{bins}} \left(\frac{n_{\mathrm{i}}^{\mathrm{after}}(C) - n_{\mathrm{i}}^{\mathrm{before}}(C)}{\sqrt{n_{\mathrm{i}}^{\mathrm{before}}(C)}} \right)^2 \tag{8.7}$$

The condition of stability is generally set to $\chi^2/N_{\mathrm{bins}} < 1$[1]. Hence, the number of iterations needs to be determined separately for any considered observable. The treatment of the uncertainties is also within the formalism of the Bayes' theorem: a detailed description can be found in [216]. Thanks to its good capability and handiness, the D'Agostini method constitutes the reference for most of the unfolding exercises in high energy physics and it has been used for the considered multijet analyses, as well.

8.1.3 SVD unfolding method

Another method, presently used for unfolding, is the Singular Value Decomposition [217] (SVD). This uses the homonymous method [218], in order to obtain a factorization of the response matrix through orthogonal matrices and isolate its singular values. In a compact formula, the response matrix can be thus expressed as:

$$\hat{A} = USV^{\mathrm{T}} \tag{8.8}$$

where U and V are orthogonal matrices and S is a diagonal matrix with non-negative elements. The matrices U and V are called "singular matrices" of the initial matrix A. This method is generally used for solving linear equations; hence, it is particularly useful for unfolding purposes, where the whole issue is to solve a system of linear equations. In fact, such decomposition of the response matrix allows to transform Equation 8.2 into the following:

$$USV^{\mathrm{T}}\vec{x} = \vec{b} \;\rightarrow\; SV^{\mathrm{T}}\vec{x} = U^{\mathrm{T}}\vec{b} \tag{8.9}$$

where, after multiplying from the right side by U^{T}, the orthogonality of the matrix U ($U^{\mathrm{T}}U = I$) has been used. Remembering that the matrix S is diagonal and calling z_{i} and d_{i}, respectively, the elements of the vectors $V^{\mathrm{T}}x$ and $U^{\mathrm{T}}b$, the final unfolded values can be expressed as:

$$s_{\mathrm{i}}z_{\mathrm{i}} = d_{\mathrm{i}} \;\rightarrow\; z_{\mathrm{i}} = \frac{d_{\mathrm{i}}}{s_{\mathrm{i}}} \;\rightarrow\; x_{\mathrm{i}} = \mathrm{V}_{\mathrm{ij}}\frac{d_{\mathrm{j}}}{s_{\mathrm{j}}} \tag{8.10}$$

The whole issue of inverting a matrix reduces, however, to the problem of finding the singular matrices of the detector response. At this point, two additional complications need to be tackled: the first one comes from the fact that the errors of each bin of the vector $\vec{b}$ are generally not equal to each other, while the second one is driven by the possibility to have small values for some of the singular values s_{i}. Both of these occurrences result in an unstable behaviour for the unfolded results: in particular, they would cause a high sensitivity of the z_{i} values to statistical fluctuations

---

[1] The convergence of the χ^2 does not, in fact, guarantee the reliability of the obtained results. Indeed, even with small χ^2 variations, the distributions might change substantially or might converge to the wrong solution. In this thesis, this possibility has been studied by considering the folding-back procedure (see Section 8.4). In this way, it has been verified that the number of iterations, given by the condition of stability, is also able to reproduce the detector level distributions after including the migration effects to the unfolded results.

of the MC sample in the response matrix. This needs to be avoided and the SVD method faces these issues in the following way [217]:

1. it applies a bin-by-bin rescaling of Equation 8.2 according to the errors in $\vec{b}$, in order to get the same significance for all linear equations;
2. it introduces a regularization term in the formalism, in order to suppress possible oscillating solutions for small values of s_{i}.

The final equation, seen as a minimization problem, can be parametrized as:

$$(\mathrm{A}\vec{w}-\vec{b})^{\mathrm{T}}(\mathrm{A}\vec{w}-b)+\tau\cdot(\mathrm{C}\vec{w})^{\mathrm{T}}\mathrm{C}\vec{w}=\min \tag{8.11}$$

where the product $\mathrm{A}w$ includes the reweighting of point (1) and the second term adds the regularization condition: the matrix C defines the “a priori” condition on the solution while the τ parameter determines how important the regularization condition is, by giving a relative weight to it. Normally, the matrix C is based on the belief that the unfolded distributions should be smooth, with small bin-to-bin variations. The parameter τ is, generally, problem-dependent and needs to be extracted through some closure tests for each measured observable. With this new formalism and after some linear algebra [217], Equation 8.10 is modified into:

$$x_{\mathrm{i}}=\mathrm{V}_{\mathrm{ij}}\frac{d_{\mathrm{j}}}{s_{\mathrm{j}}}\cdot\frac{s_{\mathrm{j}}^{2}}{s_{\mathrm{j}}^{2}+\tau} \tag{8.12}$$

Note that for large $s_{\mathrm{j}}\gg\tau$, the multiplying factor is 1 and no suppression is introduced, while for small s_{j}, the new term works as a low-pass filter: its impact depends on the specific value chosen for τ.

It has been shown [217] that this method gives reliable results for unfolding with a proper treatment of uncertainties and correlation effects. Hence, it is, currently, one of the reference methods; in the presented analyses, it has been used as an important cross-check for the D’Agostini results.

8.1.4 Software implementation

All presented methods are implemented in the RooUnfold software [219] and have been used for the multijet analyses within this framework. The classes RooUnfoldBinByBin, RooUnfoldBayes and RooUnfoldSvd are defined in the software and perform the unfolding respectively with the CFM, D’Agostini and SVD method. The RooUnfoldBinByBin class needs as input only the response matrix and the measured distribution at the detector level, while RooUnfoldBayes requires the two previous elements and the number of iterations to perform for the considered observable; the initial prior distribution is taken from the distribution at the generator level of the simulation and it is updated at each successive iteration. For the RooUnfoldSvd class, the response matrix, the data distribution and the regularization parameter need to be provided. The response matrix is implemented for all methods through the class RooUnfoldResponse: it has the advantage of including fake and miss events without defining additional histograms. Response matrices used for the CFM or the SVD methods need to be built with the same number of bins for detector and generator level, while the D’Agostini method does not necessarily require this condition.

8.2 Cross closure tests

Different cross closure tests have been tried in order to check the performance of the unfolding procedure in various observables and with different settings. A cross closure test is the process of unfolding MC-based detector level distributions which can be compared to the corresponding generator level; in this way, the agreement between the unfolded and the true level can be investigated and the performance maximized, by choosing the optimal value of number of iterations for the D'Agostini method or of the regularization parameter for the SVD. Thus, in the following, two different cross closure tests have been examined, before going to unfold the data themselves:

- performance of different unfolding methods;

- dependence on the specific MC sample, used for the unfolding.

8.2.1 Testing different unfolding methods

The first cross closure test aims for studying the effects of different methods used for unfolding. The MC sample, simulated with PYTHIA6, is considered: generator and detector level distributions, as well as the response matrix, have been obtained with this generator. Then, the observed spectra are unfolded with the three previous methods. Results are shown in Figure 8.2 for a subsample of observables, selected in the 2b2j channel: two p_{T} spectra, leading and subleading b-tagged jets, and two jet correlation observables, $\Delta\phi^{\mathrm{light}}$ and $\Delta^{\mathrm{rel}}_{\mathrm{light}}p_{\mathrm{T}}$. Curves for generator and detector levels are shown together with the Bayes-, CFM- and SVD-unfolded spectra. The number of iterations for the D'Agostini method and the regularization parameter for the SVD method have been optimized in order to achieve the best agreement between unfolded and true distributions. The agreement between the two levels is very good, with differences smaller than 5%. The conclusions of this study are very similar for observables measured in the 4j channel. This shows that the machinery of unfolding is properly set for the three methods and gives a preliminary estimation of the settings to use for D'Agostini and SVD methods.

8.2.2 Testing different Monte Carlo samples

In order to test the performance of the unfolding for different MC samples, the generators PYTHIA6 and HERWIG++ have been used. In particular, detector level distributions, as obtained in PYTHIA6, are unfolded with the response matrix produced with the HERWIG++ sample. The distributions unfolded with the D'Agostini method are then compared with the true ones to check their agreement. Results are shown in Figure 8.3 in the 2b2j channel. Since the inputs of response matrix and histograms to unfold are not the same, differences are visible. Between unfolded and true distributions, differences of up to 40% are observed: these might depend on different migration effects between the two samples or different shapes between the two input distributions. In the 4j channel, only differences up to 15–20% have been observed. Effects of these differences are, however, taken into account for the final measurement in the systematical uncertainty related to the model dependence (see Chapter 9).

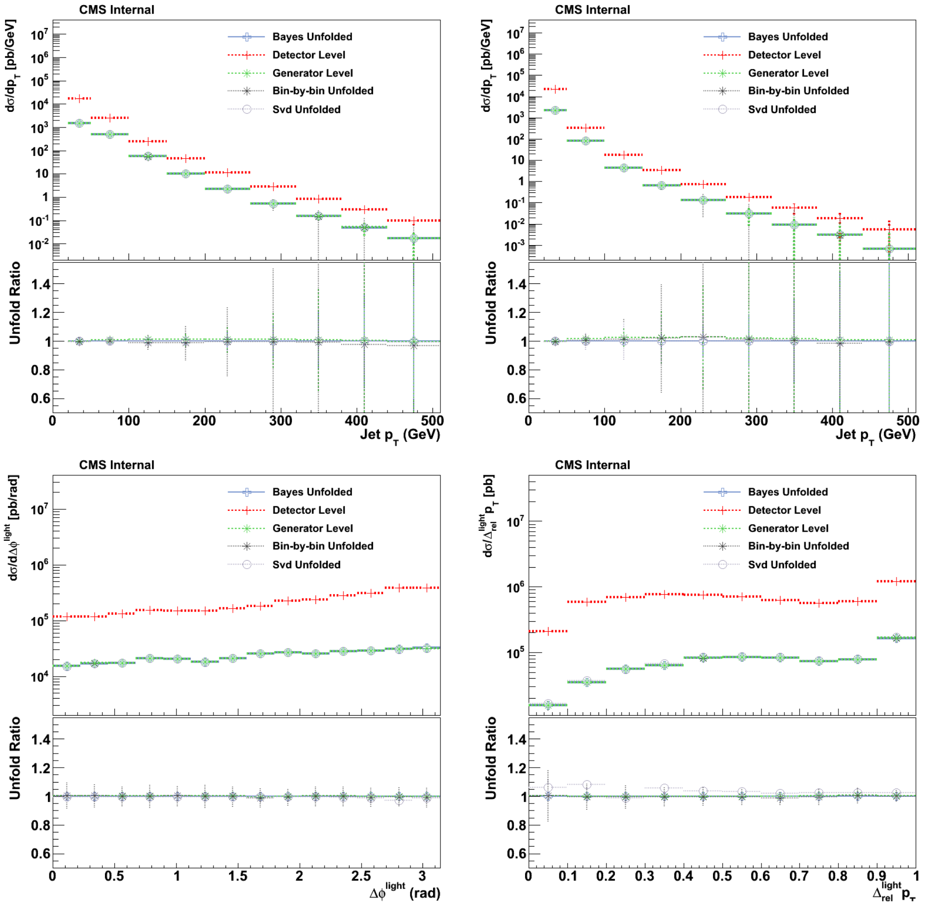


Figure 8.2: Closure test of the unfolding procedure performed with the Pythia*6 sample for the p_T spectra of leading (top left) and subleading (top right) b-tagged jets and the correlation observables $\Delta\phi^{light}$ (bottom left) and $\Delta^{rel}_{light}p_T$ (bottom right). Detector level distributions are unfolded with the response matrix obtained with the same* Pythia*6 sample and with different unfolding methods: Bayesian, SVD and bin-by-bin methods. The true generator level distributions are also shown as obtained in the* Pythia*6 sample. The lower panel shows the ratio of all the curves to the distributions unfolded with the Bayesian method.*

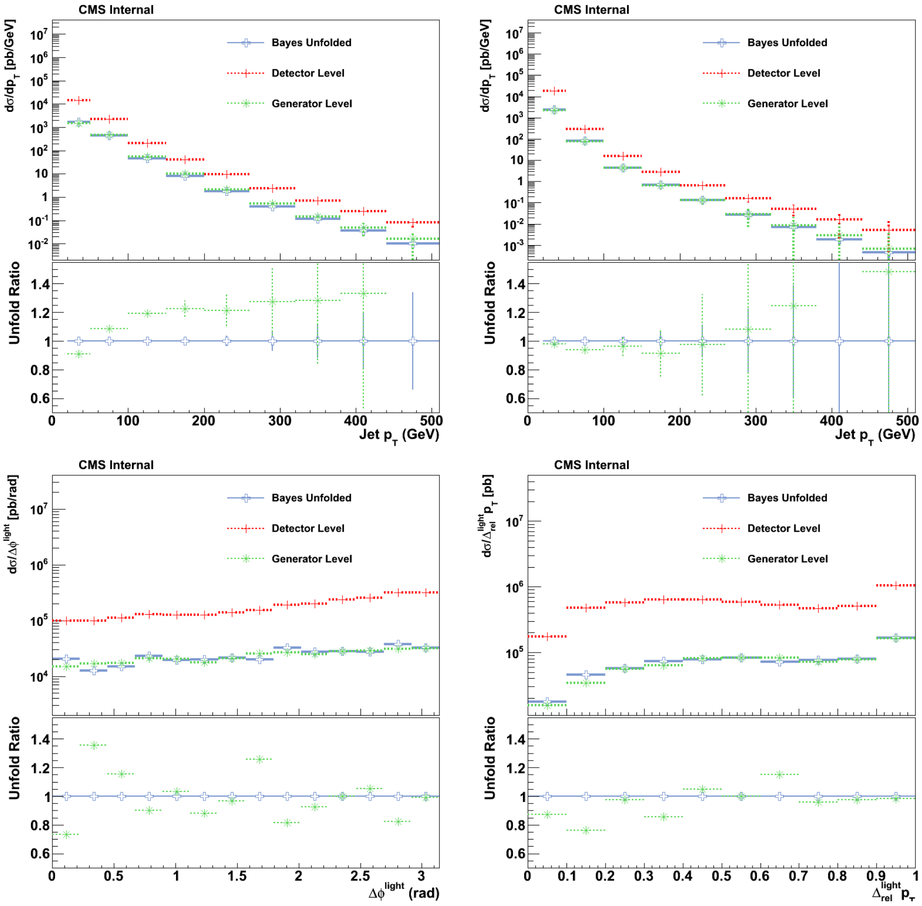


Figure 8.3: Closure test of the unfolding procedure performed with the Pythia*6 sample for the p_T spectra of leading (top left) and subleading (top right) b-tagged jets and the correlation observables $\Delta\phi^{light}$ (bottom left) and $\Delta^{rel}_{light}p_T$ (bottom right). Detector level distributions are unfolded with the Bayesian method by using the response matrix obtained with the* Herwig++ *sample. The true generator level distributions are also shown as obtained in the* Pythia*6 sample. The lower panel shows the ratio of all the curves to the distributions unfolded with the Bayesian method.*

8.3 Unfolding the data

After the previous cross closure tests, the unfolding of the real measurements can take place. The only difference with respect to the previous tests is that, in this case, a "true" level is obviously not available for comparisons. The strategy of data unfolding is different for the two measured physics channels. In particular:

- for the 4j selection, data are unfolded separately with both Pythia6 and Herwig++ with the D'Agostini method; then, the final data points are taken as the bin-by-bin average between the two obtained distributions, while the systematical uncertainties, due to the model dependence (see Chapter 9), is evaluated through the semi-difference between them;
- for the 2b2j selection, data are unfolded, again with the D'Agostini method, but only with the Pythia6 sample. The Herwig++ sample is only used for the determination of the model dependence uncertainty (see Chapter 9): in this case, the whole difference (and not half of the difference) of the two unfolded distributions is taken.

Data are unfolded with the response matrix obtained with Pythia6 with different unfolding methods: results are shown for the 2b2j channel in Figure 8.4, for different observables: leading and subleading b-tagged jet p_{T}, $\Delta\phi^{\mathrm{light}}$ and $\Delta^{\mathrm{rel}}_{\mathrm{light}} p_{\mathrm{T}}$. While the distributions obtained with the CFM are quite far from the outputs produced with the D'Agostini method, with differences even higher than 100%, results of the SVD method get closer to them: discrepancies do not go above 40% for the p_{T} spectra and above 5–10% for the correlation observables. The instability of the results obtained with the CFM method is understandable since large migration effects, which are not properly treated in CFM, are present for these observables, as shown in Chapter 7. Differences between the SVD and D'Agostini outputs may be explained by the specific feature of the p_{T} spectra to have a rapidly falling distribution, which is particularly complicated to treat within the unfolding procedure with large migration effects. It has been observed that a relatively big variation of the unfolded results is obtained when the Bayesian and SVD inputs, namely number of iterations and regularization parameter, are changed. This variation is much smaller for the other measured distributions. Differences between the Bayes- and SVD-unfolded results, observed in the 4j channel, where less sizable migration effects are present, are much smaller and go only up to 5–10%. The method used to verify the reliability of the different unfolding procedures is described in Section 8.4 and helps to decide which results to trust.

Data have been also unfolded with different models with the D'Agostini method: samples generated with Pythia6 and Herwig++ have been considered. The unfolding method has been also applied with Pythia6 without applying the $\hat{p}_{\mathrm{T}}$ reweighting: this gives an idea of how big the effect of the reweighting is on the final unfolded distributions. Results are shown in Figure 8.5 in the 2b2j channel. Results obtained with Pythia6 and Herwig++ are very close to each other, compatible within 10–15% for all considered variables. Results obtained with the Pythia6 sample without applying the reweighting show a very good agreement for the correlation observables, while larger differences appear for the p_{T} spectra. This is an expected effect: for the variables where the reweighting does not determine a change in shape, like the correlation observables, the unfolded results are consistent with the ones obtained with the weighted sample. Instead, for observables where a change in shape is obtained, like for the p_{T} spectra, the unfolded results start to be different. One should anyway remember that the description of the unreweighted sample is very

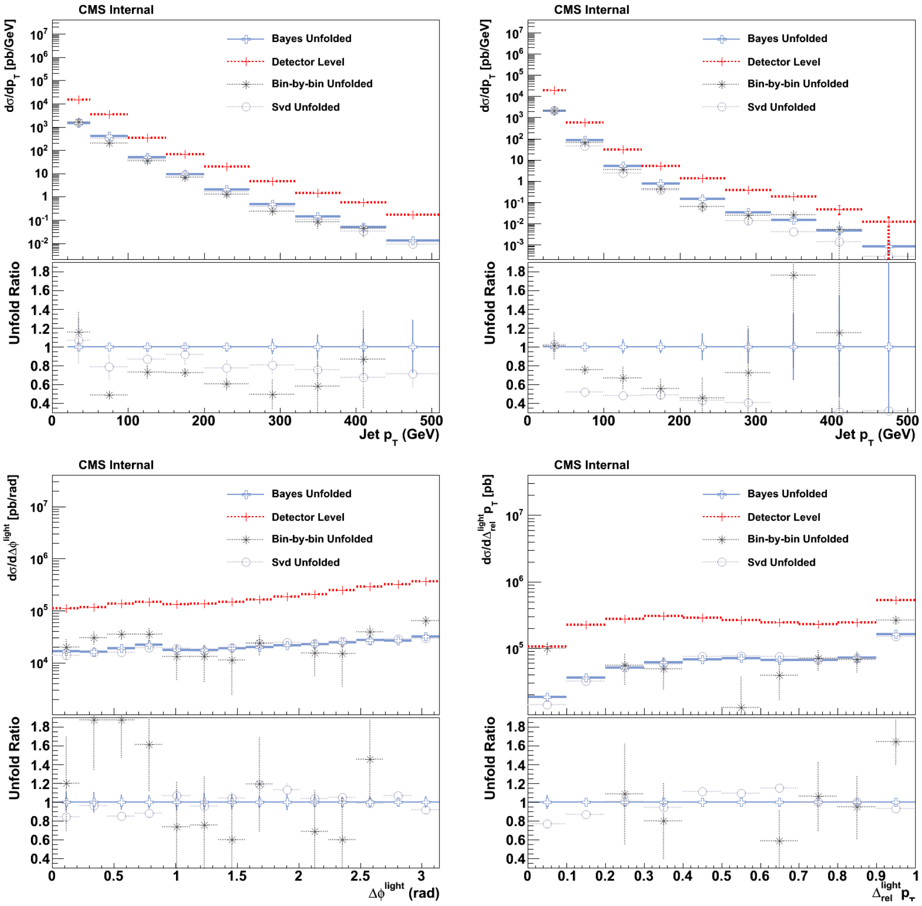


Figure 8.4: Normalized distributions of data obtained with different unfolding procedures for the p_T spectra of leading (top left) and subleading (top right) b-tagged jets and the correlation observables $\Delta\phi^{light}$ (bottom left) and $\Delta^{rel}_{light} p_T$ (bottom right). Detector level data distributions are unfolded with the response matrix obtained with the Pythia6 *sample and with different unfolding methods: Bayesian, SVD and bin-by-bin methods. The lower panel shows the ratio of all the curves to the distributions unfolded with* Pythia6 *with the Bayesian method.*

poor for the p_{T} spectra (see Figure 7.12): hence, unfolded results, when using this sample, can not be trusted.

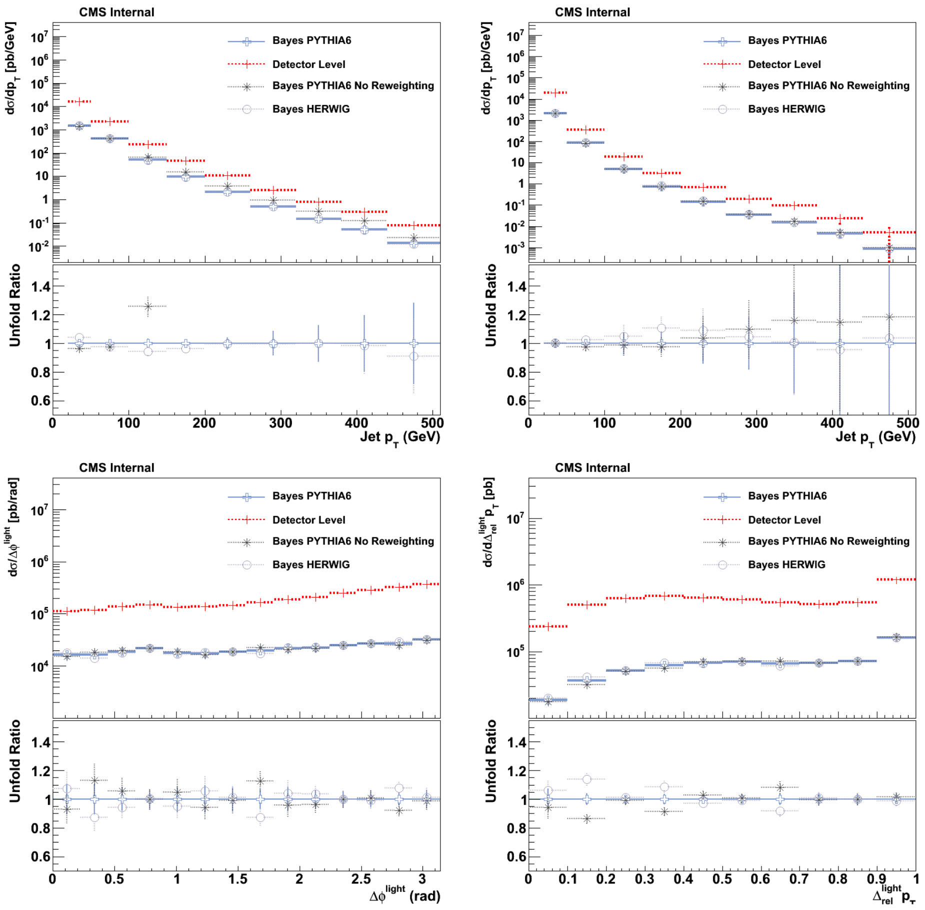


Figure 8.5: Absolute cross section distributions of data obtained with different unfolding procedures for the p_T spectra of leading (top left) and subleading (top right) b-tagged jets and the correlation observables $\Delta\phi^{light}$ (bottom left) and $\Delta^{rel}_{light} p_T$ (bottom right). Detector level data distributions are unfolded with the response matrix obtained with the PYTHIA*6 and* HERWIG++ *samples, as well as with the* PYTHIA*6 sample, when the reweighting procedure has not been applied. The lower panel shows the ratio of all the curves to the distributions unfolded with* PYTHIA*6. with reweighting*

8.4 Folding the distributions back

An important check of the unfolded results is the attempt of folding the obtained distributions back to the detector level, by using MC simulation. It is in fact possible, by using the response matrix, to apply acceptance, efficiency and migrations to the unfolded distributions and manually build the detector effects. In this way, a schematic but reliable detector simulation can be added on top of the “true” level and the folded spectra can be compared to the measured data distributions.

Since no related class is implemented in ROOUNFOLD, the folding exercise has been “manually” set; it is, in particular, performed by applying the following formula:

$$N^{\mathrm{i}}_{\mathrm{fold}} = \sum_{\mathrm{j}=1}^{N_{\mathrm{bins}}} \frac{P^{\mathrm{ij}} \cdot N^{\mathrm{j}}_{\mathrm{unfold}} \cdot (1 - P^{\mathrm{j}}(\mathrm{Miss}))}{1 - P^{\mathrm{i}}(\mathrm{Fake})} \tag{8.13}$$

where $P(\mathrm{Miss})$ and $P(\mathrm{Fake})$ are respectively the probability of a miss and a fake event, $N^{\mathrm{j}}_{\mathrm{unfold}}$ is the content of the distribution unfolded with the PYTHIA6 sample in bin j. $N^{\mathrm{i}}_{\mathrm{fold}}$ is the content of the folded distribution in bin i and P^{ji} is the probability of an event unfolded in bin j to be detected in bin i. Note that $P(\mathrm{Miss})$ and $P(\mathrm{Fake})$ have different indices in Equation 8.13: this is due to the fact that the probability of a fake event is affected by the detector level bin, while only the generator level bin determines the probability of a miss event. This equation has been applied to Bayes-unfolded distributions, obtained with different number of iterations. The goodness of the unfolding has been evaluated by looking at the values of the reduced χ^2 between the folded and the detector level distributions. The χ^2 definition takes into account uncertainties from both spectra and it considers them as independent; in a compact way, it can be defined as:

$$\chi^2 = \sum_{\mathrm{i}=1}^{\mathrm{Nbins}} \left(\frac{N^{\mathrm{i}}_{\mathrm{det}} - N^{\mathrm{i}}_{\mathrm{fold}}}{\sqrt{\sigma^{i\,2}_{\mathrm{det}} + \sigma^{i\,2}_{\mathrm{fold}}}} \right)^2 \tag{8.14}$$

For most of the considered distributions, it has been observed that the χ^2 distribution as a function of the number of iterations decreases down to a minimum and then slightly increases again. The expected behaviour is, instead, a monotonic decrease for increasing number of iterations. This might be due to the specific χ^2 definition, which neglects any correlation between the uncertainties. However, the optimal number of iterations has been chosen as the one which minimizes the χ^2. A summary of the obtained results is provided in Table 8.3 for all measured observables in the 2b2j channel. It shows for each observable the optimal number of iterations, the χ^2, the number of degrees of freedom, defined for each histogram as N_{bins}-1, and the reduced χ^2. Values of the obtained number of iterations result to be between one and three and they are, for most of the observables, equal to the ones determined in the cross closure test: in some few cases, they differ from that number by a maximum of one iteration. It is very remarkable that the reduced χ^2 values are all around 1, meaning a very good agreement between the folded distributions and the detector level spectra, measured in the data. The highest reduced χ^2 value is 1.95, observed for ΔS. Direct comparisons between the two considered distributions are shown in Figure 8.6 for some measured observables: leading and subleading b-tagged jet, $\Delta\phi^{\mathrm{light}}$ and $\Delta^{\mathrm{rel}}_{\mathrm{light}} p_{\mathrm{T}}$. As expected and quantified by the χ^2 values, the two curves are very close to each other and overlapping in most of the bins.

This manifests very good performance of the D'Agostini unfolding procedure. The same test performed with distributions unfolded with the SVD method led to much higher values of the minimum of the reduced χ^2, up to about 6 for the jet p_{T} spectra. This is an indication that the SVD unfolding is ill-behaving for the 2b2j selection. However, the SVD unfolding method has shown good agreement with the detector level distributions, when the folding-back test has been performed in the 4j scenario. For the 2b2j analysis, the SVD method has been dropped out and the data points obtained with the D'Agostini method by using the number of iterations listed in Table 8.3, are considered for the final measurement.

Measured Observable	χ^2	NdF	Reduced χ^2	Optimal iter. number
S'_{p_T}	10.26	9	1.14	3
S_ϕ	4.97	13	0.38	2
$\Delta\phi^{\mathrm{bottom}}$	7.45	13	1.05	1
$\Delta\phi^{\mathrm{light}}$	2.91	13	0.22	3
$\Delta\mathrm{S}$	19.48	10	1.95	3
$\Delta\eta^{\mathrm{bottom}}$	15.85	11	1.44	2
$\Delta\eta^{\mathrm{light}}$	17.86	15	1.19	3
$\Delta^{\mathrm{rel}}_{\mathrm{bottom}}p_\mathrm{T}$	7.40	9	0.82	3
$\Delta^{\mathrm{rel}}_{\mathrm{light}}p_\mathrm{T}$	7.30	9	0.81	2
1^{st} jet p_T	3.98	9	0.44	2
2^{nd} jet p_T	6.87	9	0.76	2
3^{rd} jet p_T	5.42	9	0.60	2
4^{th} jet p_T	4.32	9	0.48	3
1^{st} jet η	7.8	7	1.11	3
2^{nd} jet η	1.09	7	0.16	3
3^{rd} jet η	4.10	13	0.31	2
4^{th} jet η	2.86	13	0.22	2

Table 8.3: Comparison between the folded and the detector level distributions for each measured observable. The χ^2, the number of degrees of freedom, the reduced χ^2 and the number of iterations used in the unfolding are listed in the Table.

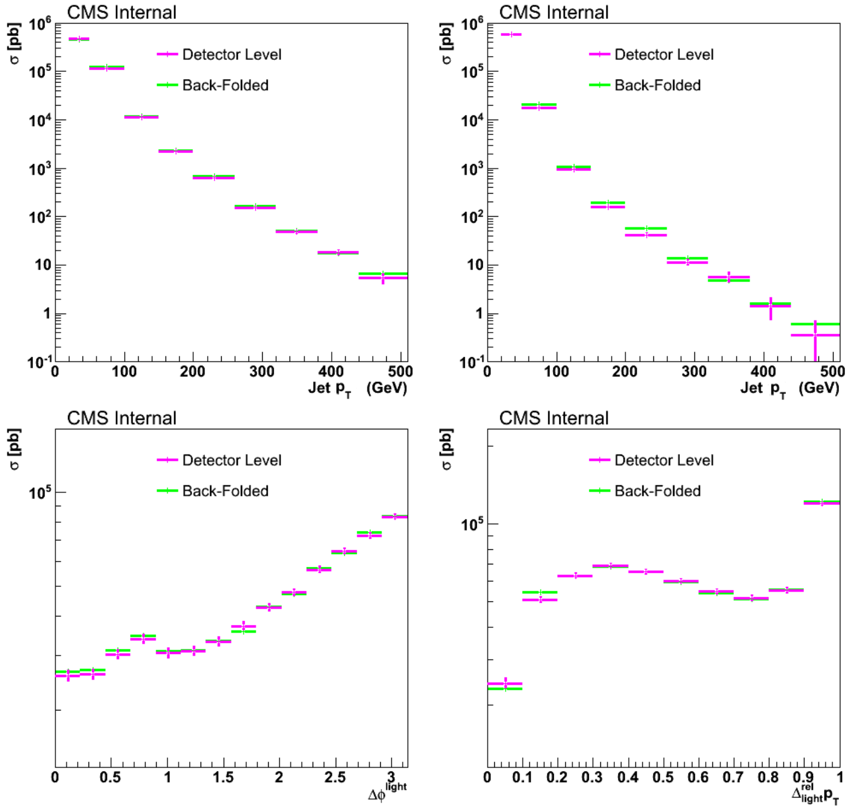


Figure 8.6: Absolute cross section distributions of data obtained after folding back to the detector level the unfolded measurements; they are for the p_T spectra of leading (top left) and subleading (top right) b-tagged jets and the correlation observables $\Delta\phi^{light}$ (bottom left) and $\Delta^{rel}_{light}p_T$ (bottom right). The unfolded data distributions are folded back with the response matrix obtained with the Pythia6. *The folded distributions are compared to the detector level ones, as measured in the data.*

Chapter 9

Systematic uncertainties

For a physics analysis, it becomes crucial to study the uncertainties which affect the measurements, in order to extract meaningful conclusions. Besides the statistical ones, which only deal with the number of selected events, the systematic uncertainties are also very important: they have different origins, going from known detector effects, to the specific choice of the models used for the correction of the data. Since differential cross sections are measured, uncertainties are evaluated such that a value can be assigned for each bin of the measured observable. The uncertainties have been estimated for absolute cross sections and normalized distributions: the latter results to be smaller than the former, because they include only migration effects within the phase space, while the ones into and out of the phase space are excluded. The uncertainties of the normalized distributions represent only the amount of shape variation under a specific effect, while the uncertainties of the absolute cross sections include also information about the variation of the selected number of events. In this Chapter, all the uncertainty sources are described and results are shown for each measured observable in the two analyses.

9.1 Jet energy scale

One of the major uncertainties in jet measurements is the jet energy scale. As explained in Chapter 5, the energy deposited in the detector which is reconstructed as a jet needs to be corrected to obtain a reliable measurement, in order to match as precise as possible the kinematical quantities of the object, which originated it. The corrections which are organized in three sequential levels (see Chapter 5), carry a measurement uncertainty. Since the two analyses, described in the thesis, are based on the selection of a large number of jets with a relatively low p_{T}, it is expected that this is the major systematic uncertainty. The uncertainty on the jet energy scale for the data recorded in 2010 is of about 2–5% in the central region ($|\eta| < 2$) and it increases up to 3–7% in the forward region ($|\eta| \sim 4$). It is higher for lower p_{T}, of the order of 4% at 20 GeV, and it decreases at larger p_{T}, to the order of 2% and 1.5%, respectively at 50 and 200 GeV [192].

The effect of the jet energy scale has been evaluated by varying up and down the energy of all jets by the uncertainty; the observables obtained with these changes are then compared to the nominal distributions. Nominal distributions are the ones measured when the nominal values of the jet energy scale have been used. The differences between the nominal distributions and the ones obtained with the modified jet scale reflect the effect of the jet energy scale. The values of these differences, taken bin-by-bin, are referred to as the jet energy scale uncertainties. When applying the jet energy scale uncertainties, symmetric differences are obtained in the up and down directions, with respect to the nominal distributions. For the 4j analysis, they are 13–15% for the absolute cross sections and around 2–3% for the normalized distributions. For the 2b2j analysis, they are 30-35% for the absolute cross sections and less than 3-8% for the shape distributions of the correlation observables; the shapes of jet p_{T} spectra are affected by the jet energy scale by about 10–15%.

9.2 Jet energy resolution

Together with the jet energy scale, another important detector effect is the energy resolution. The detector response in any measured quantity, as shown in section 5, is not exactly corresponding to the true value of the measured physical quantity but it results in a gaussian distribution around it. The wider the distribution is, the less accurate the measurement is. The width of this distribution is called resolution. While the accurate angular resolutions in η and in ϕ in CMS have a negligible effect for the described measurement, the resolution in transverse momentum, which is equivalent to the one in energy, is more relevant and its effect needs to be taken into account. In particular, it is important that the resolution measured in data matches the one observed in the simulation; it has been observed that this can be achieved by applying a correction to the jet p_{T} in the simulation at the detector level. This correction depends on the jet η and p_{T}. These correction factors are officially provided by the CMS Collaboration and they are listed in [220]: they range between 1.06 and 1.17 with higher values when going to larger pseudorapidities. In order to correct for the resolution mismatch, the jet p_{T} in the simulation has been smeared out around its true value by matching every jet at the detector level to the closest one at the generator level. The match is performed through an angular cone with an aperture $\Delta\mathrm{R} = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} = 0.3$. This value of cone width is reliable enough for the matching due to the high resolution achieved by the PF jets in the angular variables. If the matching algorithm fails, due to fluctuations or detector inefficiencies, the jet p_{T} is not corrected but left as it is. The smearing procedure can be summarized by the formula:

$$p_{\mathrm{T}}^{\mathrm{smeared}} = p_{\mathrm{T}}^{\mathrm{true}} \pm a \cdot (p_{\mathrm{T}}^{\mathrm{true}} - p_{\mathrm{T}}^{\mathrm{det\ level}}) \tag{9.1}$$

where a is the correction factor taken from [220], and $p_{\mathrm{T}}^{\mathrm{true}}$ and $p_{\mathrm{T}}^{\mathrm{det\ level}}$ are the transverse momenta of the matched jets, respectively at generator and detector level. The correction factors a have been measured in Z-boson and dijet events [192], for each period of data taking. The uncertainty due to this effect depends on the uncertainty of the correction factors themselves. The values for each bin are taken as uncertainties. The results show a contribution which ranges between 3 and 15% for both absolute cross sections and normalized distributions, and for both analyses. A large effect is observed for jet p_{T} spectra and ΔS, which are the variables that depend most on the accuracy of the p_{T} measurement, while ϕ- and η-based quantities present a less relevant contribution from the jet resolution.

9.3 Model dependence

The correction of the data to the stable particle level relies on particular models used for the unfolding (see Chapter 8). The uncertainty associated with the choice of these models is referred to as “model dependence”. The models, which have been used for the two analyses, are PYTHIA6 and HERWIG++. The model dependence has been evaluated by unfolding the data with the two models and by measuring the differences between them. For the 4j analysis, the effect is very similar for absolute and normalized cross sections: it ranges from 2% up to 6%. For the 2b2j analysis, the relevance of the model dependence is higher and it reaches values of 15–20 % for absolute cross sections and for normalized distributions. Note that, while the model dependence for the 4j analysis has been evaluated by taking half of the differences of the unfolded results obtained with the PYTHIA6 and HERWIG++ samples, for the 2b2j analysis, the whole difference between the two samples has been taken as uncertainty: this is due to the fact that the HERWIG++ samples cannot be used for the data correction, due to missing SF (see Chapter 7). This is one of the reasons why the model dependence uncertainties for the 2b2j analysis are higher than the ones measured in the 4j analysis.

9.4 B-tag scale factors

The SF used to match the b-tagging performance in data and simulation are also affected by an uncertainty. As mentioned in Chapter 7, the uncertainty for the SF is of 10%, independently on jet p_{T}, η and flavour[1]. The effect of this uncertainty has been evaluated by shifting up and down the SF in the selected events by that amount. The distributions obtained in this way are compared to the ones measured with the nominal values of the SF. The quoted uncertainty consists of the maximum bin-by-bin difference between the two new samples and the nominal ones. The uncertainty is found to be of about 15% for the cross section measurements and of 1–3% for the normalized distributions in the 2b2j analysis.

9.5 Other minor uncertainties

Minor uncertainties are related to the trigger efficiency correction and the pile-up reweighting.

The trigger efficiency correction introduces a systematic uncertainty due to the statistical uncertainty in the measurement of the turn-on curve. In order to give an estimation of that effect, all the bins in the turn-on curve of Figures 6.3(top right) and 6.4 (bottom right), have been varied up and down by the statistical uncertainty and the considered observables have been measured with the new settings. The differences between them and the distributions obtained when using the nominal values of the trigger efficiency have been extracted and the quoted uncertainty has been chosen to be the maximum deviation between the up/down settings and the nominal values. These effects are of the order of 3-5% for the cross section measurements and of 1% for the normalized distributions in the 2b2j analysis. For the 4j analysis, this contribution turned out to be negligible, of less than 1% for both cross section and normalized distributions. The difference between the two analyses comes from the different phase space (in the 4j analysis, jets of higher p_{T} are required) and from the different triggers which are used.

The pile-up reweighting procedure also introduces a systematic effect for the measurements depending on how well the primary vertex distribution in data matches the one in simulation. Since an iterative procedure has been applied, the uncertainty is estimated by the difference between the results obtained with the nominal reweighting procedure and the ones obtained with a fewer

---

[1] This large uncertainty reflects the fact that data-driven measurements of the tag and mistag performance are available only for 2011 data and the SF for 2010 data are just translated from those, with a higher uncertainty.

number of iterations (four instead of five). This effect contributes with a negligible uncertainty ($< 0.1\%$) for both absolute cross sections and normalized distributions in the two analyses.

In addition to the systematic sources, an uncertainty coming from the limited number of selected events also contributes to the measurement and needs to be assigned to the distributions: this is the statistical uncertainty. This has been calculated by considering the contents of each bin as quantities following a poissonian distribution[1] and assigning thus the poissonian standard deviation [168] to them. For the 4j analysis, the statistical uncertainty adds a 1% contribution to the measurements, while for the 2b2j analysis, it reaches average values of 2.5%. This uncertainty becomes the main contribution only for jet p_{T} spectra, starting from values of 350–400 GeV. In the 4j analysis, the p_{T} spectrum of the soft jets stops at 200–300 GeV because of lack of events at larger values.

9.6 Total uncertainty

The previous uncertainties are combined at the end in order to get the total systematic uncertainty. For 2010 data, an additional uncertainty of 4% [221] due to the luminosity is added for the absolute cross section measurements: this value has been taken for both analyses and included in the total uncertainty. The combination of the uncertainties has been evaluated by summing in quadrature the single contributions, assuming absence of correlation among the different sources. In Figures 9.1 and 9.2, the various uncertainties, together with their combination, are plotted for the absolute cross sections, as a function of the jet p_{T} spectra and the correlation observables, measured in the 2b2j analysis. The same quantities are represented in Figures 9.3 and 9.4 for the normalized distributions. The uncertainties for two adjacent bins are merged by taking their average and then assigning to them the resulting value: this has been done in order to reduce statistical fluctuations in the simulation, mainly observed in the evaluation of the model dependence.

In the 2b2j analysis, the main contribution of the uncertainty to the cross section measurements comes from the jet energy scale, while it is the model dependence for the normalized distributions. This still holds in the 4j analysis: for the normalized cross sections, the jet energy scale and model dependence uncertainty add a contribution of the same amount.

9.7 Summary of uncertainties for two analyses

Tables 9.1 and 9.2 summarize the effects for each observable in the 4j analysis[2], while Tables 9.3 and 9.4 concern the 2b2j analysis. In particular, the total uncertainties are listed as taken from a weighted average of all bins of each differential distribution.

---

[1] Note that for a large number of entries, a poissonian distribution is equivalent to a gaussian distribution, according to the central limit theorem.

[2] Some of the observables in the 4j analysis are not included in the final paper [11], but only presented in a limited phase space as preliminary results [10]; the uncertainties quoted here are the ones, obtained when considering the full phase space for the four jets.

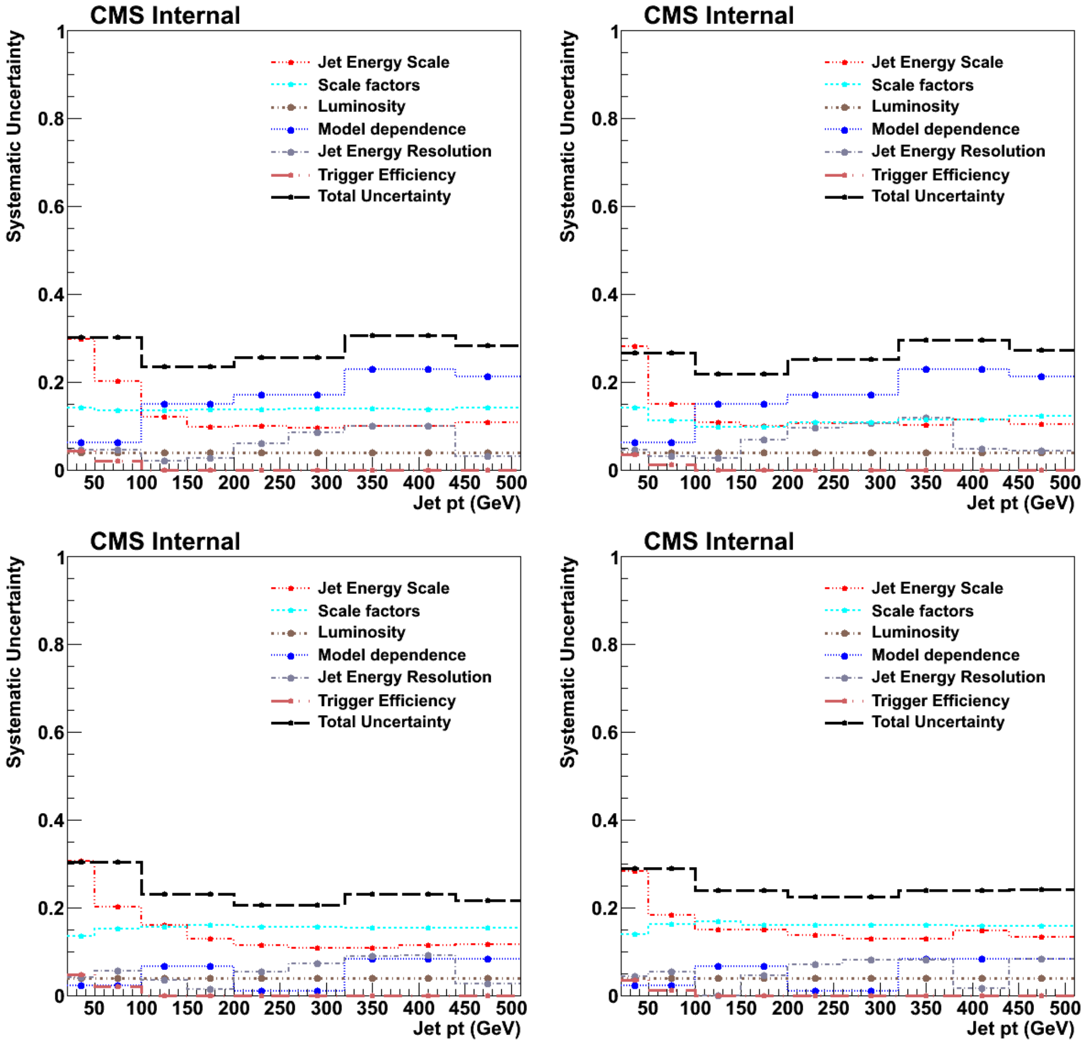


Figure 9.1: Systematic uncertainties for the absolute cross sections of the jet p_T spectra in the 2b2j analysis: leading b-tagged jet (top left), subleading b-tagged jet (top right), leading additional jet (bottom left), subleading additional jet (bottom right). The total uncertainty is obtained by summing up in quadrature the single contributions.

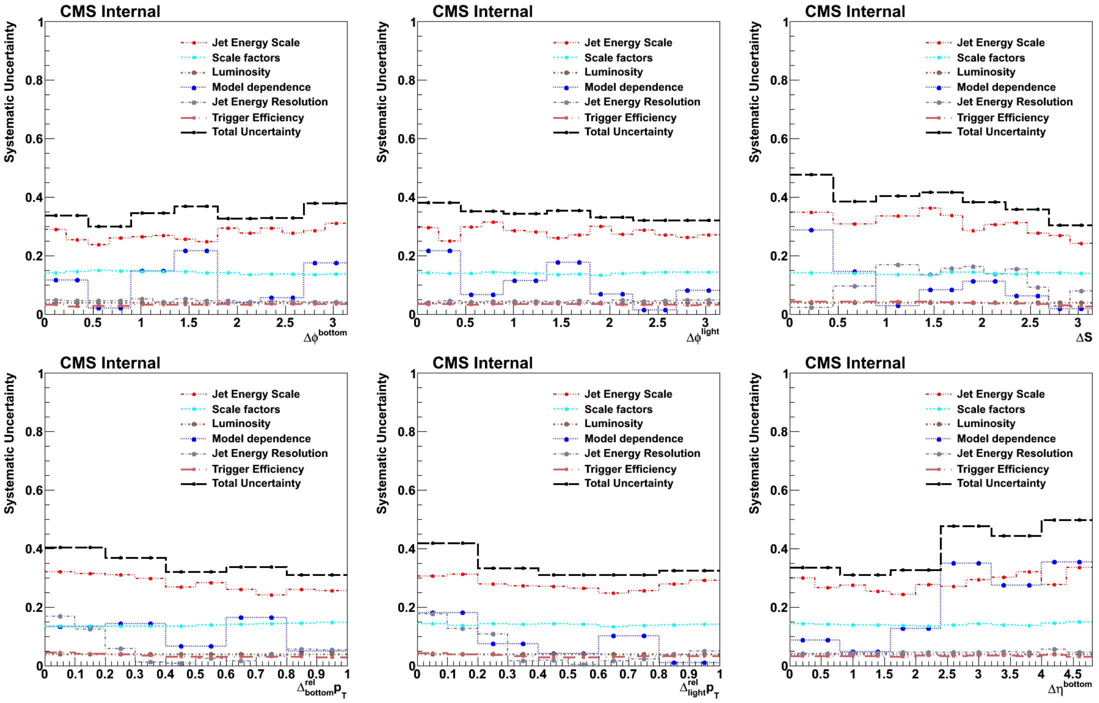


Figure 9.2: Systematic uncertainties for the absolute cross sections of the correlation observables in the 2b2j analysis: $\Delta\phi^{bottom}$ (top left), $\Delta\phi^{light}$ (top center), ΔS (top right), $\Delta^{rel}_{bottom}p_T$ (bottom left), $\Delta^{rel}_{light}p_T$ (bottom center) and $\Delta\eta^{bottom}$ (bottom right). The total uncertainty is obtained by summing up in quadrature the single contributions.

Measured observable	Model	JES	JER	PU	Total
Hard jet p_{T}	2%	13%	1%	0.1%	15%
Soft jet p_{T}	3%	13%	1%	0.1%	15%
Jet $\lvert\eta\rvert \leq 3$	2%	13%	1%	0.1%	15%
Jet $\lvert\eta\rvert > 3$	10%	27%	5%	0.1%	30%
$\Delta\phi^{\mathrm{hard}}$	6%	15%	1%	0.1%	17%
$\Delta\phi^{\mathrm{soft}}$	3%	14%	2%	0.1%	15%
$\Delta\eta^{\mathrm{hard}}$	3%	13%	1%	0.1%	15%
$\Delta\eta^{\mathrm{soft}}$	3%	13%	2%	0.1%	15%
$\Delta^{\mathrm{rel}}_{\mathrm{hard}}p_{\mathrm{T}}$	3%	13%	1%	0.1%	15%
$\Delta^{\mathrm{rel}}_{\mathrm{soft}}p_{\mathrm{T}}$	3%	13%	1%	0.1%	14%
$\Delta\mathrm{S}$	5%	15%	4%	0.1%	16%
S_{ϕ}	4%	14%	2%	0.1%	15%
$S'_{p_{\mathrm{T}}}$	4%	14%	2%	0.1%	15%

Table 9.1: Systematic uncertainties affecting the absolute cross section distributions for each measured observable in the 4j analysis: every systematic source is specified and the value reflects the average over all the bins of the observable. The 4% uncertainty from the luminosity measurement is included in the total uncertainty. This is obtained by summing the individual uncertainties in quadrature.

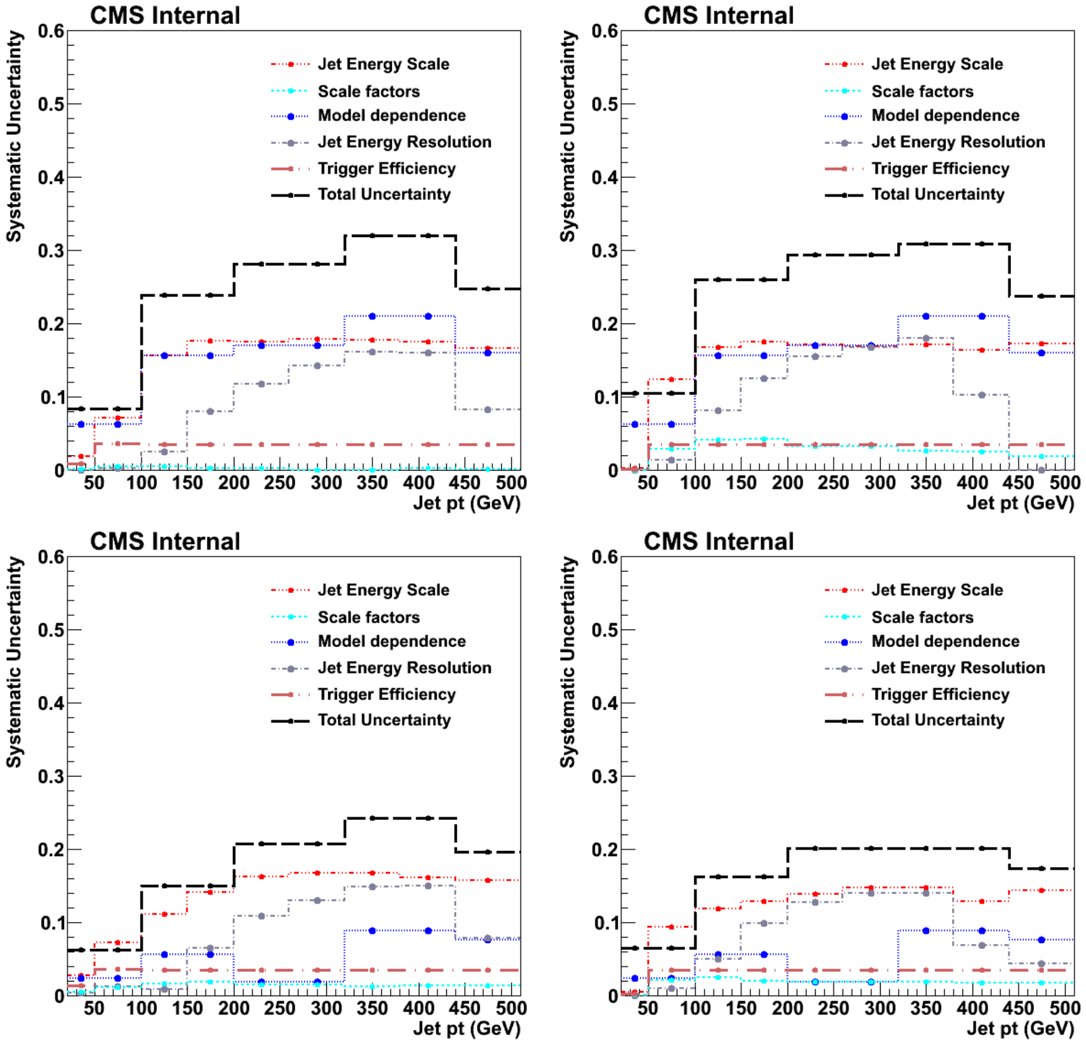


Figure 9.3: Systematic uncertainties for the normalized cross sections of the jet p_T spectra in the 2b2j analysis: leading b-tagged jet (top left), subleading b-tagged jet (top right), leading additional jet (bottom left), subleading additional jet (bottom right). The total uncertainty is obtained by summing up in quadrature the single contributions.

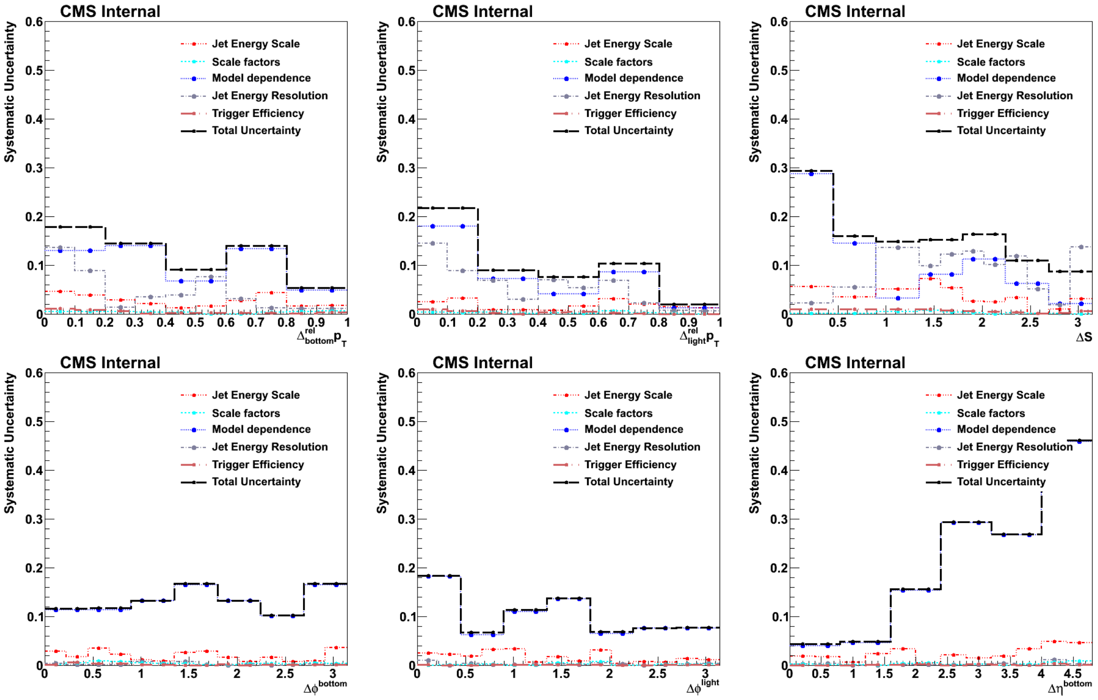


Figure 9.4: Systematic uncertainties for the normalized cross sections of the correlation observables in the 2b2j analysis: $\Delta\phi^{bottom}$ *(top left),* $\Delta\phi^{light}$ *(top center),* ΔS *(top right),* $\Delta^{rel}_{bottom}p_T$ *(bottom left),* $\Delta^{rel}_{light}p_T$ *(bottom center) and* $\Delta\eta^{bottom}$ *(bottom right). The total uncertainty is obtained by summing up in quadrature the single contributions.*

Measured observable	Model	JES	JER	PU	Total
Hard jet p_{T}	5%	3%	2%	0.1%	6%
Soft jet p_{T}	5%	5%	3%	0.1%	7%
Jet $\lvert\eta\rvert \leq 3$	3%	2%	1%	0.1%	4%
Jet $\lvert\eta\rvert > 3$	10%	15%	2%	0.1%	18%
$\Delta\phi^{\mathrm{hard}}$	6%	3%	0.5%	0.1%	7%
$\Delta\phi^{\mathrm{soft}}$	3%	3%	2%	0.1%	5%
$\Delta\eta^{\mathrm{hard}}$	3%	2%	0.5%	0.1%	4%
$\Delta\eta^{\mathrm{soft}}$	3%	2%	1%	0.1%	4%
$\Delta^{\mathrm{rel}}_{\mathrm{hard}}p_{\mathrm{T}}$	3%	2%	0.5%	0.1%	4%
$\Delta^{\mathrm{rel}}_{\mathrm{soft}}p_{\mathrm{T}}$	3%	3%	2%	0.1%	5%
$\Delta\mathrm{S}$	4%	3%	3%	0.1%	5%
S_{ϕ}	3%	3%	1%	0.1%	5%
$S'_{p_{\mathrm{T}}}$	3%	3%	1%	0.1%	5%

Table 9.2: Systematic uncertainties affecting the normalized cross section distributions for each measured observable in the 4j analysis: every systematic source is specified and the value reflects the average over all the bins of the observable. The 4% uncertainty from the luminosity measurement is included in the total uncertainty. This is obtained by summing the individual uncertainties in quadrature.

Measured Observable	Model	JES	JER	PU	SF	Trigger efficiency	Total
b-tagged jet p_T	20%	30%	4%	< 0.1%	15%	6%	30%
light-jet p_T	10%	30%	4%	< 0.1%	15%	6%	30%
Jet $\lvert\eta\rvert \leq 3$	10%	25%	4%	< 0.1%	15%	5%	30%
Jet $\lvert\eta\rvert > 3$	20%	35%	4%	< 0.1%	15%	5%	43%
$\Delta\phi^{\mathrm{bottom}}$	12%	35%	4%	< 0.1%	15%	6%	41%
$\Delta\phi^{\mathrm{light}}$	12%	30%	4%	< 0.1%	15%	6%	36%
$\Delta\eta^{\mathrm{bottom}}$	18%	30%	4%	< 0.1%	15%	6%	39%
$\Delta\eta^{\mathrm{light}}$	15%	35%	4%	< 0.1%	15%	6%	42%
$\Delta^{\mathrm{rel}}_{\mathrm{bottom}} p_\mathrm{T}$	15%	30%	4%	< 0.1%	15%	6%	38%
$\Delta^{\mathrm{rel}}_{\mathrm{light}} p_\mathrm{T}$	5%	30%	4%	< 0.1%	15%	6%	35%
$\Delta\mathrm{S}$	12%	35%	4%	< 0.1%	15%	6%	41%
S_ϕ	12%	35%	4%	< 0.1%	15%	6%	41%
S'_{p_T}	12%	35%	4%	< 0.1%	18%	6%	42%

Table 9.3: Systematic uncertainties affecting the absolute cross section distributions for each measured observable in the 2b2j analysis: every systematic source is specified and the value reflects the average over all the bins of the observable. The 4% uncertainty from the luminosity measurement is included in the total uncertainty. This is obtained by summing the individual uncertainties in quadrature.

Measured Observable	Model	JES	JER	PU	SF	Trigger efficiency	Total
b-tagged jet p_T	20%	15%	10%	< 0.1%	2%	5%	27%
light-jet p_T	10%	15%	10%	< 0.1%	2%	5%	21%
Jet $\lvert\eta\rvert \leq 3$	5%	8%	1%	< 0.1%	1%	1%	10%
Jet $\lvert\eta\rvert > 3$	10%	15%	1%	< 0.1%	2%	1%	30%
$\Delta\phi^{\mathrm{bottom}}$	13%	5%	1%	< 0.1%	2%	1%	14%
$\Delta\phi^{\mathrm{light}}$	13%	5%	1%	< 0.1%	2%	1%	14%
$\Delta\eta^{\mathrm{bottom}}$	15%	5%	1%	< 0.1%	2%	1%	16%
$\Delta\eta^{\mathrm{light}}$	15%	5%	1%	< 0.1%	2%	1%	16%
$\Delta^{\mathrm{rel}}_{\mathrm{bottom}} p_\mathrm{T}$	9%	5%	7%	< 0.1%	2%	1%	12%
$\Delta^{\mathrm{rel}}_{\mathrm{light}} p_\mathrm{T}$	13%	5%	7%	< 0.1%	2%	1%	15%
$\Delta\mathrm{S}$	20%	5%	10%	< 0.1%	2%	1%	23%
S_ϕ	10%	5%	1%	< 0.1%	2%	1%	11%
S'_{p_T}	10%	5%	1%	< 0.1%	2%	1%	11%

Table 9.4: Systematic uncertainties affecting the normalized distributions for each measured observable in the 2b2j analysis: every systematic source is specified and the value reflects the average over all the bins of the observable. The total uncertainty is obtained by summing the individual uncertainties in quadrature.

Chapter 10

Cross section measurement of the 4j and 2b2j channels and comparisons with predictions

After the data unfolding, absolute and normalized differential cross sections are available at the stable particle level. For each measured bin i, they are defined as given below:

$$\frac{d\sigma^{\mathrm{i}}}{\Delta O} = \frac{N^{\mathrm{i}}}{L \cdot \Delta O^{\mathrm{i}} \cdot C^{\mathrm{i}}_{\mathrm{unfold}}} \tag{10.1}$$

$$\frac{1}{\sigma}\frac{d\sigma^{\mathrm{i}}}{\Delta O} = \frac{1}{N^{\mathrm{total}}}\frac{N^{\mathrm{i}}}{\Delta O^{\mathrm{i}} \cdot C^{\mathrm{i}}_{\mathrm{unfold}}} \tag{10.2}$$

where L is the integrated luminosity, C^{i} are the correction factors determined from the unfolding and N^{i} is the number of events measured in each bin at the detector level. The variable ΔO^{i} indicates the bin width used in the specific histograms: both cross sections are divided by this quantity in order to produce distributions which are independent on the choice of the binning. The measurement of the normalized differential cross sections needs an additional input with respect to the absolute ones: this is the total number of events, N^{total}, selected in the considered phase space. Absolute distributions are measured in units of cross section, while normalized ones are fractions of events. The particle level has been defined for the 4j analysis in Table 8.1, while the corresponding definition for the 2b2j scenario is listed in Table 8.2. Predictions for both channels have been obtained by using the corresponding RIVET plugins (see Appendix D).

10.1 Studies of the UE simulation in Monte Carlo samples

Before showing the measurements and the results of the two four-jet channels, a validation of the UE part of the MC generators used for comparison needs to be performed. This is necessary because the data are unfolded back to the stable particle level and hence, any ME calculation, used for comparison, which produces results at parton level, needs a simulation of the underlying event. This, as seen in Chapter 4, is provided either by standard MC generators, like PYTHIA, which are interfaced to ME calculations like in the case of POWHEG or MADGRAPH, or by the same generator which produces both ME and UE like Sherpa, Pythia or Herwig. Comparing ME results, namely results at parton level, with measurements at stable particle level might lead to wrong conclusions due to the inconsistency of the final states. While deep investigations have been performed for standard MC generators like Pythia and Herwig and a large literarure is available about the most up-to-date UE tunes, no such studies have been carried out for other generators, like Powheg, Madgraph and Sherpa. Since predictions obtained with these generators are very relevant for the interpretation of the four-jet measurements, it is important to understand their performance. This is done by comparing specific predictions against inclusive measurements to check their ability to describe basic distributions, relative to particle event activity or jet observables. If the distributions are described, then it is reasonable to produce predictions with the same tunes of the UE simulation for the four-jet analyses and the comparisons become meaningful.

The validation of Powheg, Madgraph and Sherpa[1] is presented in the following and it is the first reference which addresses this issue; it has been performed by considering inclusive jet cross sections in different pseudorapidity ranges and UE data concerning hadronic activity in the transverse region, both measured by CMS [89, 204]. These measurements refer to a center-of-mass energy of 7 TeV. Since Powheg and Madgraph are interfaced with the UE simulation provided by Pythia6, the tune dependence has been studied, namely predictions have been generated for different Pythia6 UE settings. The ones, considered in this study, have been:

- Tune Z2 and Z2* [180], which are currently two of the most up-to-date tunes for Pythia6 and the standard ones used in CMS[2];
- Tune P11 [45], which uses different values of UE and fragmentation parameters;
- Tune AMBT [222], which is a tune performed by the ATLAS collaboration.

In addition, predictions have been also produced for different values of the parameter $PARP(82)$ [115], which sets the value of the MPI regulator in Pythia6, according to Equation 2.12. A higher value reduces the MPI contribution, while a lower value increases it. For Sherpa, only one tune for each considered PDF is available and recommended by the authors; hence, no tune dependence has been studied for it. For all generators, investigations about the role of MPI have also been investigated, by making predictions with and without the simulation of the MPI.

Figure 10.1 shows the comparisons between inclusive jet cross sections in two pseudorapidity ranges ($|\eta| < 0.5$ and $1.5 < |\eta| < 2.0$) in the p_{T} range between 19 GeV and 2 TeV and predictions of Powheg and Madgraph interfaced with different tunes of Pythia6. It can be noticed that almost all predictions are compatible with the measurements: tunes Z2, Z2* and P11 are performing very well for both Powheg and Madgraph while tune AMBT, when interfaced to Powheg, tends to overshoot the low p_{T} region, for $p_{\mathrm{T}} < 50$ GeV. All tunes predict, however, very similar results for $p_{\mathrm{T}} > 50$ GeV and they describe well the measurements. Only predictions obtained with Madgraph tend to slightly overestimate the very high p_{T} region going towards more forward pseudorapidities. Figure 10.2 shows the comparisons between the same data and predictions obtained with Powheg,

---

[1] The same ME, considered in the 4j comparisons, have been used for the validation: a NLO dijet ME with a hard emission for Powheg, a LO 2→4 ME for Madgraph and a LO 2→3 ME for Sherpa.

[2] CMS has recently released a new tune, CUETP6S1 [12], which improves the description of UE data at different collision energies; though, this tune has not been considered in these studies because it was not available at the time of the publication.

Madgraph and Sherpa with different settings of the MPI. For Powheg and Madgraph ME, the tune Z2 has been considered in the interface with Pythia6. Comparisons for Powheg show that the MPI are relevant for jet p_T up to 50–80 GeV: for lower values, the jet cross section decreases up to a factor of 0.7–0.8, when the MPI contribution is progressively reduced. For higher p_T values, all predictions are very close to each other. Predictions obtained with Madgraph result in the same conclusions: MPI are relevant only for jet p_T up to 80 GeV. In this case, though, predictions without the simulation of MPI significantly undershoot the data by about 40% in the low p_T region. Comparisons with Sherpa have been obtained with the nominal tune with and without the simulation of MPI. When MPI are switched on, the data are well reproduced by the predictions in both considered η ranges, while, when they are switched off, a significant disagreement of around 50% appears for $p_T < 100$ GeV. Note that the large fluctuations at $p_T \sim 100$ GeV and 400–500 GeV are due to statistical limitations of the samples: in particular at $p_T \sim 100$ GeV, two different samples have been merged in order to cover also the higher p_T region.

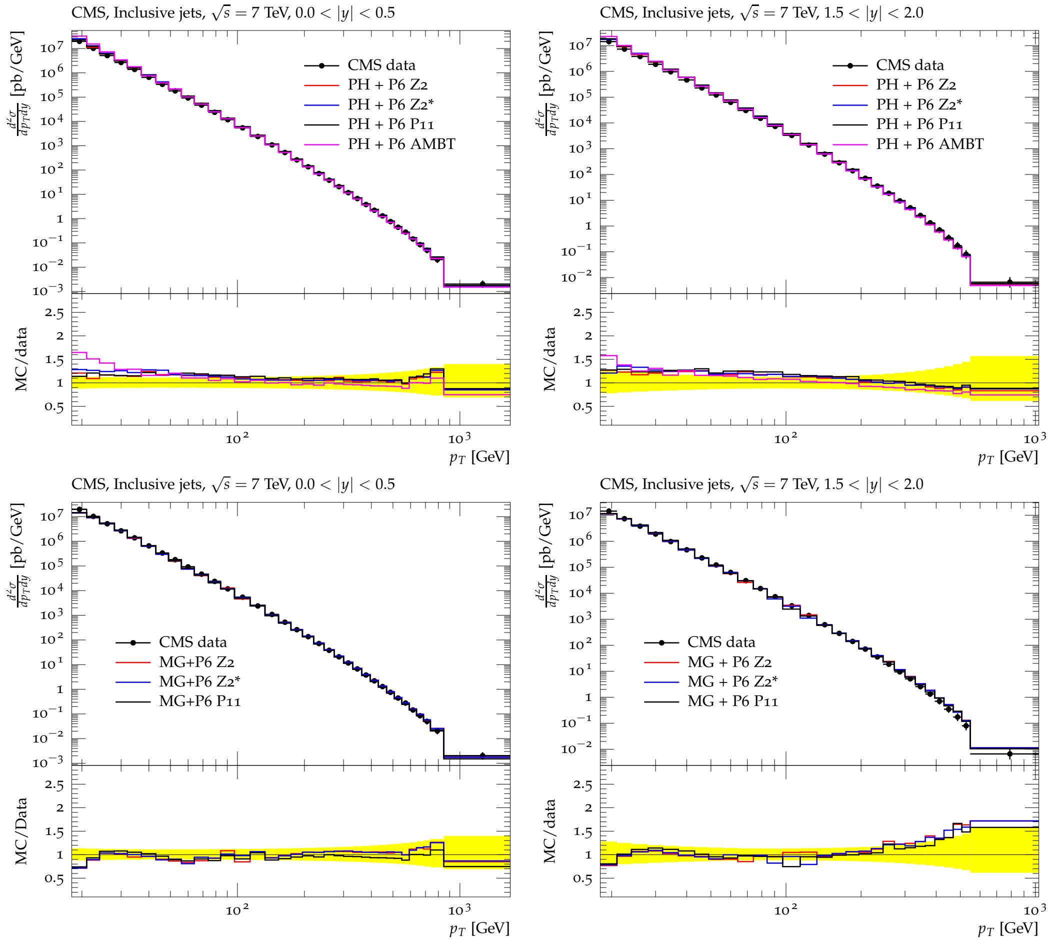


Figure 10.1: Comparisons of predictions obtained with Powheg *interfaced with* Pythia*6 (top) and* Madgraph *(bottom) with inclusive jet cross section data, measured in* $|\eta| < 0.5$ *(left) and* $1.5 < |\eta| < 2.0$ *(right). The* Powheg *predictions are shown when the ME is interfaced with different* Pythia*6 tunes: Z2, Z2*, P11 and AMBT. The* Madgraph *predictions are shown for the* Pythia*6 tunes Z2, Z2* and P11. The lower panels show the ratios between MC predictions and data.*

In Figure 10.3, predictions from the same MC event generators and tunes are compared to UE measurements. The charged particle multiplicity and p_T sum in the transverse region, as a

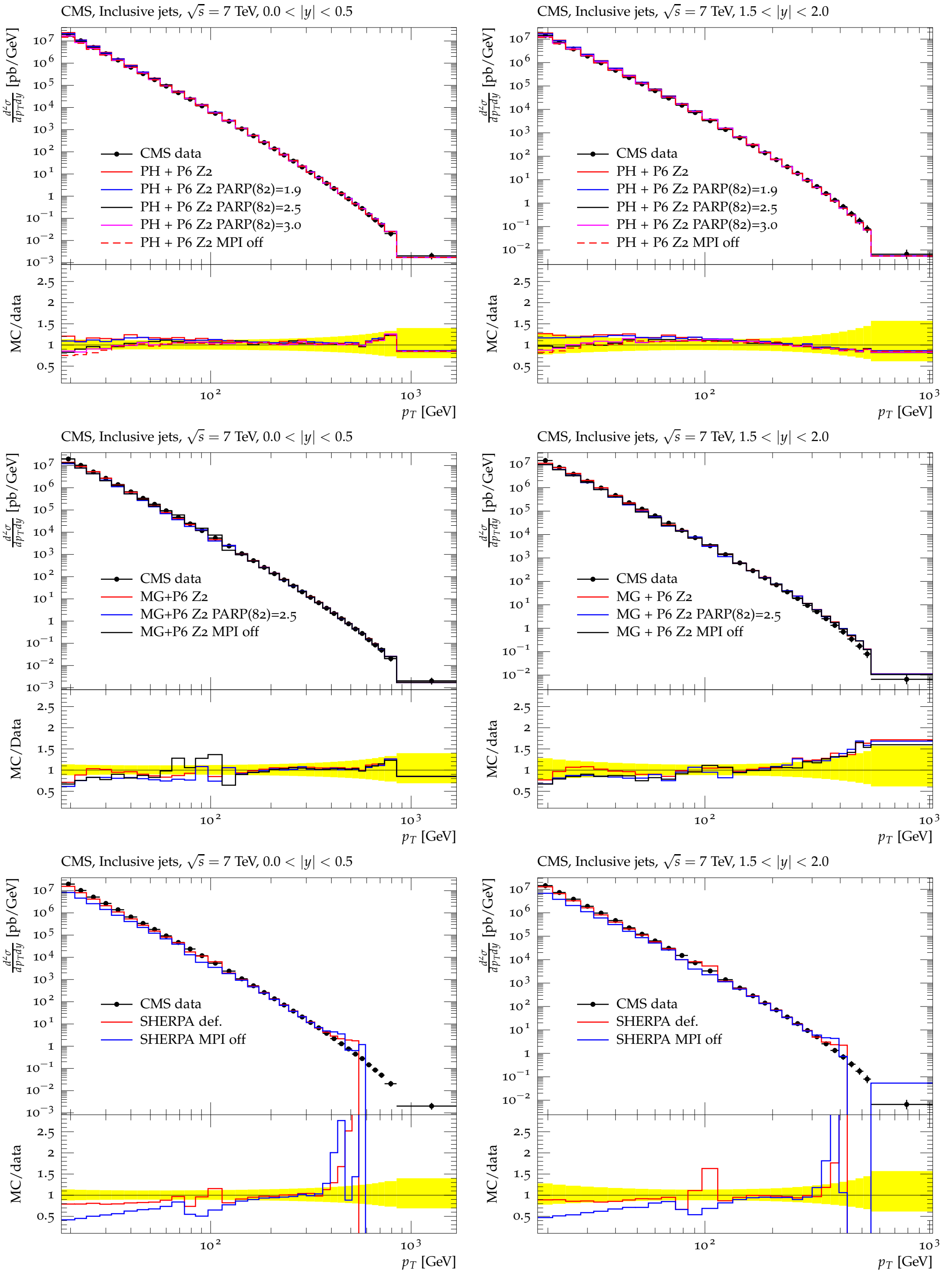


Figure 10.2: Comparisons of predictions obtained with Powheg *interfaced with* Pythia*6 (top) and* Madgraph *(middle) and* Sherpa *(bottom) with inclusive jet cross section data, measured in $|\eta| < 0.5$ (left) and $1.5 < |\eta| < 2.0$ (right). The* Powheg *predictions are shown when the ME is interfaced with different* Pythia*6 tune Z2 for different values of the MPI regulator: the default one equal to 1.821, 1.9, 2.5, 3.0 and when the MPI are switched off. The* Madgraph *predictions are shown for the* Pythia*6 tune Z2 for different values of the MPI regulator: the default one equal to 1.821, 2.5 and when the MPI are switched off. For* Sherpa*, predictions are shown as obtained with the default tune, and with the default tune without the contribution of MPI. The lower panels show the ratios between MC predictions and data.*

function of the leading track jet p_T are shown. Again, all considered tunes describe the data very well over the whole phase space. A small difference is observed only for $p_\mathrm{T} < 5$ GeV: this is just an artifact of the event generation, where the lower $\hat{p}_\mathrm{T}$ threshold has been set to 5 GeV. In Figure 10.4 the role of the MPI is further considered, by comparing of the same UE data with different MPI settings for Powheg, Madgraph and Sherpa. Here, striking differences between predictions with and without MPI appear. Predictions without MPI, or with a reduced contribution of them, when a higher value is set for $PARP(82)$, significantly fail to describe the data of up a factor of 10. The default Z2 tune describes the data very well for Powheg and Madgraph. Predictions obtained with Sherpa and the nominal tune, instead, overshoot by 10–20% the plateaux region at $p_\mathrm{T} > 10$ GeV.

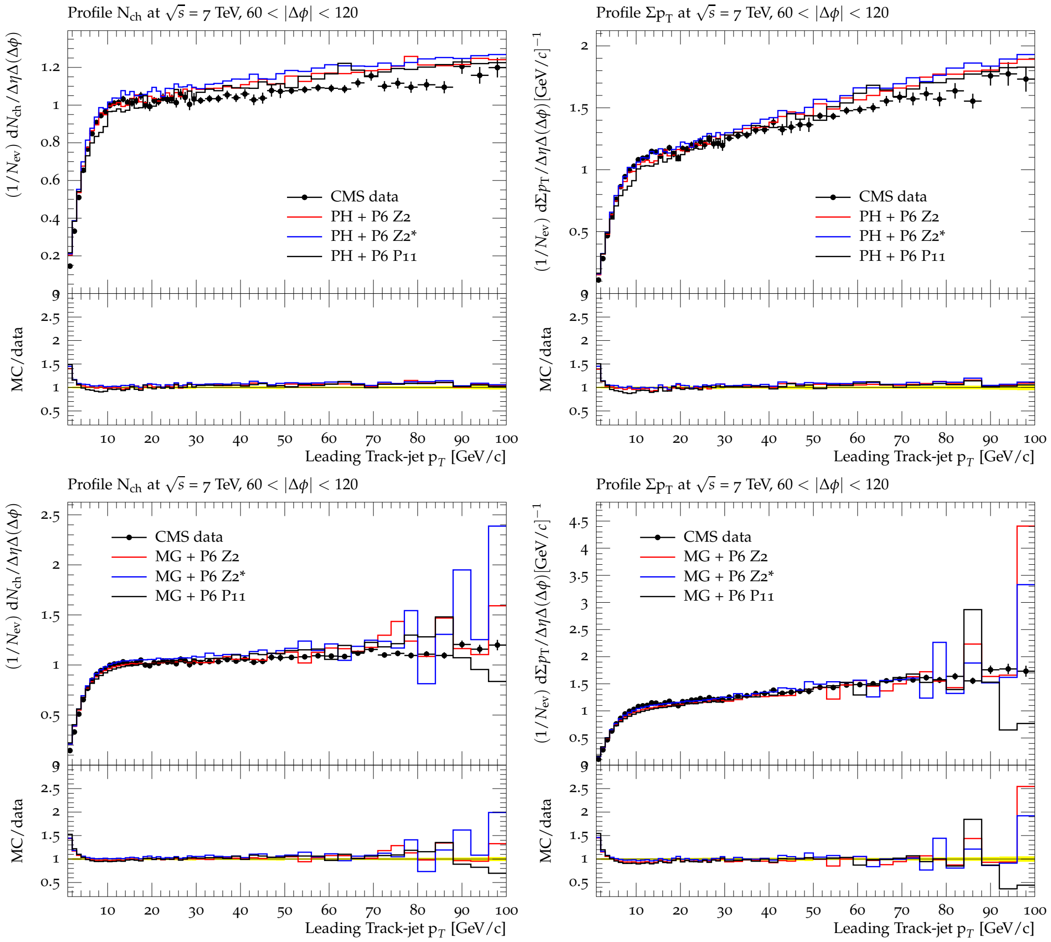


Figure 10.3: Comparisons of predictions obtained with Powheg *interfaced with* Pythia*6 (top) and* Madgraph *(bottom) with UE data of charged particle multiplicity (left) and p_T sum (right), measured in the transverse region. The* Powheg *predictions are shown when the ME is interfaced with different* Pythia*6 tunes: Z2, Z2* and P11. The* Madgraph *predictions are shown for the* Pythia*6 tunes Z2, Z2* and P11. The lower panels show the ratios between MC predictions and data.*

The validation studies, shown in this Section, suggest interesting insights. On one hand, the need for MPI is clear for all considered generators. This indicates that loop corrections or higher multiplicities of the final state included in the ME, do not replace the contribution of additional partonic interactions, which need to be simulated with a careful choice of parameters. On the other

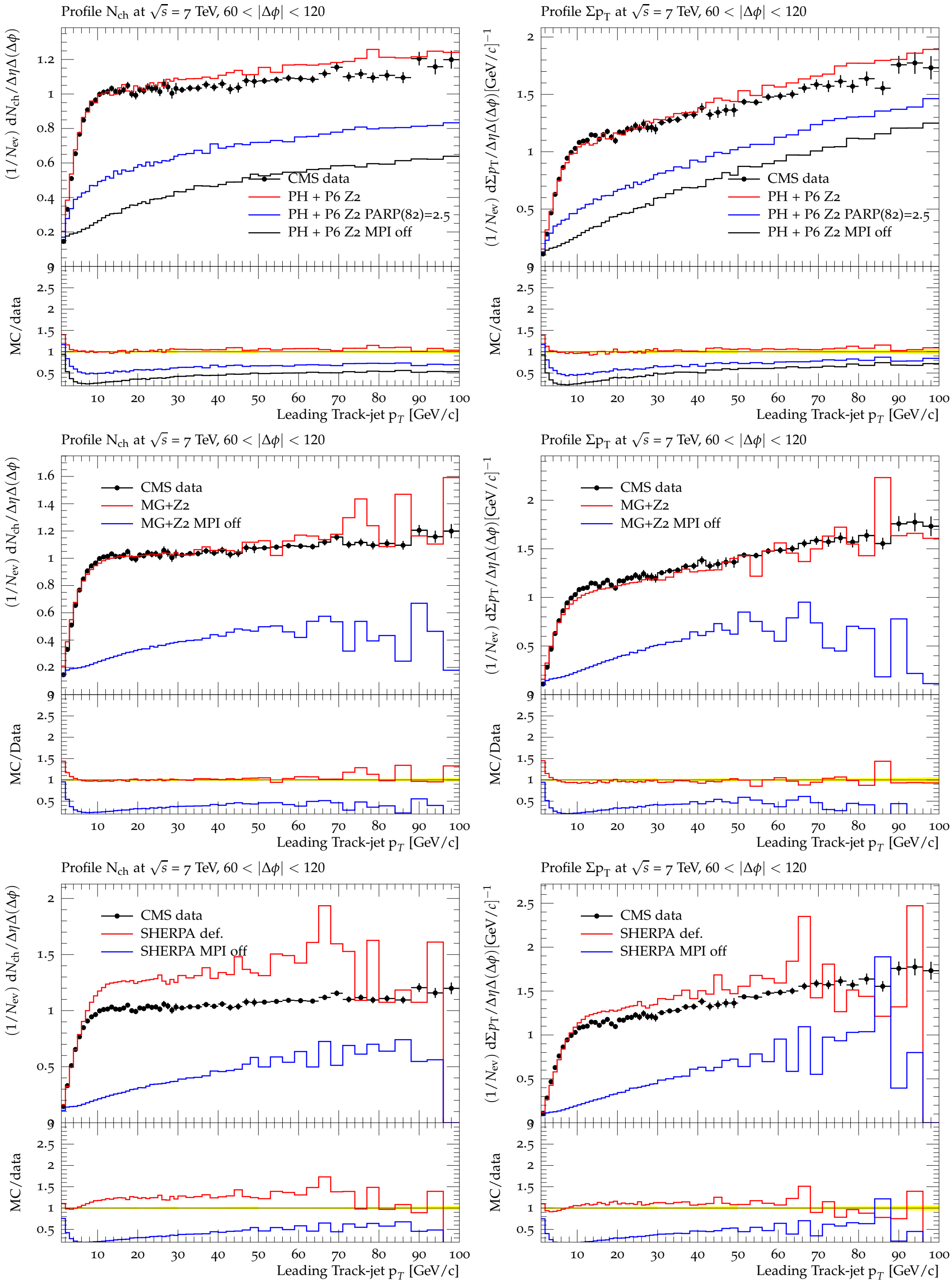


Figure 10.4: Comparisons of predictions obtained with Powheg *interfaced with* Pythia6 *(top) and* Madgraph *(middle) and* Sherpa *(bottom) with UE data of charged particle multiplicity (left) and p_T sum (right), measured in the transverse region. The* Powheg *predictions are shown when the ME is interfaced with different* Pythia6 *tune Z2 for different values of the MPI regulator: the default one equal to 1.821, 2.5 and when the MPI are switched off. The* Madgraph *predictions are shown for the* Pythia6 *tune Z2 with and without the contribution of MPI. For* Sherpa*, predictions are shown as obtained with the default tune, and with the default tune without the contribution of MPI. The lower panels show the ratios between MC predictions and data.*

hand, the compatibility of the results obtained with Powheg and Madgraph shows that tunes determined with standard LO MC generators, such as Pythia, remain meaningful also for other generators and can be propagated to different ME for the description of jet data[1]. This study shows that there is no need to retune the UE simualtion of every considered ME and to choose different sets of parameters for each of them. Having different tunes for every ME would be a very uninteresting scenario: summed to the tedious and time consuming tuning work, it would mean that the UE simulation is strongly dependent on the hard scattering and on at which level this is calculated. On the contrary, the hard scattering and UE simulation can be considered as independent.

At this stage, the tunes used for Powheg and Madgraph can be considered validated for description of jet and UE observables and they are ready to be compared to four-jet measurements. The disagreement observed for the Sherpa predictions in comparisons with UE data might indicate a too high MPI contribution, implemented in the tune and has to be taken into account in the interpretation of the 4j measurement. For Madgraph, the tune Z2*, very similar to the investigated Z2 but with a more refined energy dependence for MPI [180], is considered for the interface with Pythia6 in the following Sections.

An additional study has been performed with Powheg. It has been noticed that the phase space allowed for the PS in most of Pythia6 tunes is quite large: emissions of hard partons are allowed via PS up to a p_{T} equal to four times the scale of the hard scattering. This is regulated in Pythia6 by the parameters $PARP(67)$ and $PARP(71)$ [115], which set the multiplicative factor to be applied to the hard scattering scale to obtain the maximum p_{T} reached by hard emissions. The two parameters separate the scales, respectively for initial- and final-state PS. In tune Z2, like in most of the tunes, they are both set to a value of four.

The feature of having a large phase space for the PS is important when Pythia6 standalone is used. In this way, additional hard emissions, which are not included in the ME, can occur. When Powheg+Pythia6 is considered, this aspect becomes problematic: in fact, a hard emission is already included in the ME and an extra large phase space for the PS is not needed. In particular, this might have a big impact in the considered 4j scenario, where, in a generation with Powheg+Pythia6, the fourth jet must come from the PS, while the other three are produced at the ME level. A negligible effect is instead expected for other jet observables with a lower jet multiplicity. Hence, a tune has been "manually" modified by reducing the maximum scale from a value of 4 to a value of 1 times the scale of the hard scattering. Technically, this is set in the configuration file by adding the following lines:

$$PARP(67) = 1$$

$$PARP(71) = 1$$

In this way, hard emissions through PS are only limited to a scale which is equal to the scale of the hard scattering. The other parameters have been kept equal to the ones implemented in tune Z2*. The new tune built with the two new settings for the PS has been labelled as "tune Z2' ".

---

[1] Note that this conclusion seems not to be absolutely valid for every observable: studies related to Drell-Yan p_{T} [223] have shown that standalone Pythia6 predictions reproduce the data better than predictions obtained with Powheg+Pythia6; in this case, a retuning effort [224] is ongoing in order to fix this inconsistency.

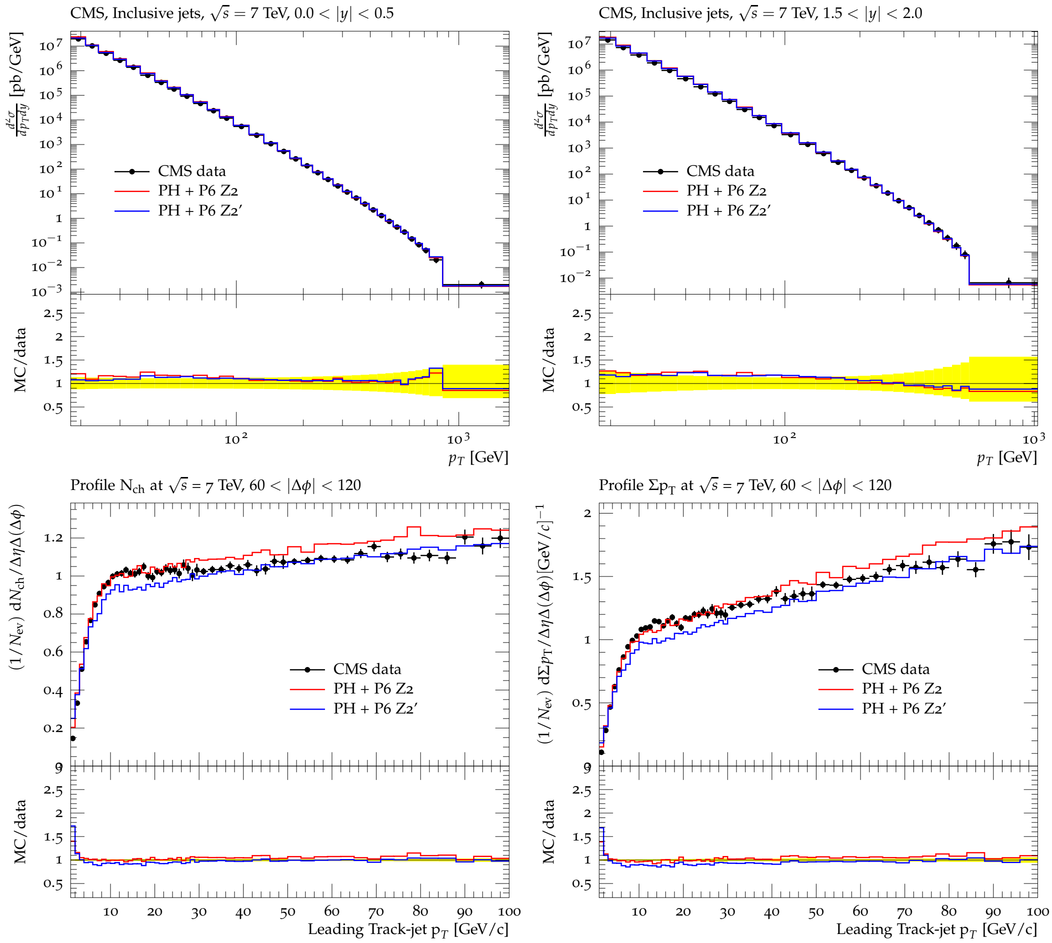


Figure 10.5: Comparisons of predictions obtained with Powheg *interfaced with* Pythia*6 with inclusive jet cross section data (top), measured in* $|\eta| < 0.5$ *(left) and* $1.5 < |\eta| < 2.0$ *(right) and with UE data (bottom) of charged particle multiplicity (left) and* p_T *sum (right), measured in the transverse region. The* Powheg *predictions are shown when the ME is interfaced with different* Pythia*6 tunes, Z2 and Z2'. The lower panels show the ratios between MC predictions and data. The discrepancies at low* p_T *in the bottom plots are due to the* $\hat{p}_T$ *threshold set to 5 GeV in the generated samples.*

Comparisons with the usual jet and UE data have been performed for predictions of the new tune Z2'. Figure 10.5 shows them, together with the predictions obtained with Powheg interfaced with the nominal tune Z2. From the plots, as expected, one does not see big differences between the two Powheg predictions: both of them give a very good description of the data and the PS does not play a big role in inclusive jet cross sections and UE observables. Conversely, when looking at 4j measurements, relevant differences appear. Figure 10.6 shows the absolute differential cross section as a function of the leading and subleading jet p_T[1]: it can be noticed that, while predictions from Powheg+Pythia6 tune Z2' are very close to the data, the predicted cross section is too large at low p_T, when the nominal tune Z2 is used. This is also observed in the third and fourth jet p_T spectra. This is exactly the expected effect: with a wide phase space for the PS, the contribution of low p_T jets becomes too high, due to an excessive production via parton evolution. This issue is, thus, solved by limiting the PS phase space, whose reduction brings the predictions closer to the data.

Hence, from these observations, the new tune Z2' has been considered in the interface with Powheg, for the understanding of the 4j measurement; this is considered more reliable and suitable than the nominal tune Z2 for multijet scenarios.

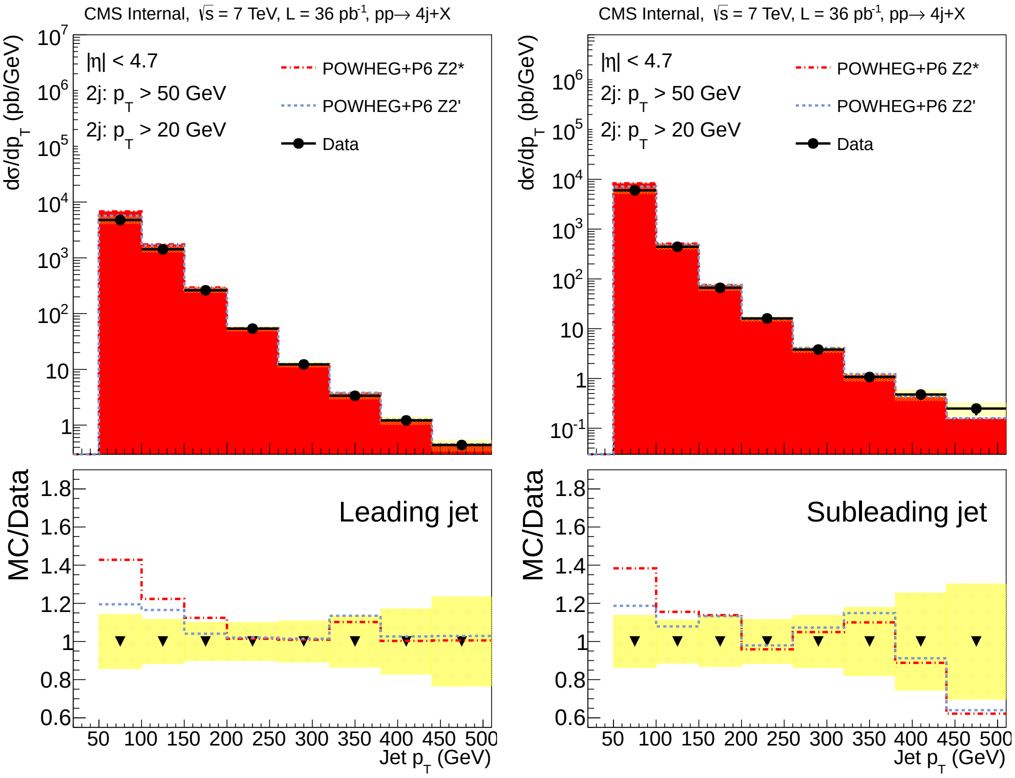


Figure 10.6: Comparisons of predictions obtained with Powheg *interfaced with* Pythia*6 with* p_T *spectra, leading (left) and subleading (right) jet, as measured in the 4j channel. The* Powheg *predictions are shown when the ME is interfaced with different* Pythia*6 tunes, Z2 and Z2'. The lower panels show the ratios between MC predictions and data.*

---

[1] The results of the 4j measurements will be discussed in detail in Section 10.2.

10.2 4j selection

In this Section, results of absolute and normalized differential cross sections, measured in the exclusive 4j scenario, are presented. The results are published in Phys.Rev.D [11]. The total integrated cross section for four jets in the final state, in the considered phase space has been measured to:

$$\sigma(\mathrm{pp} \to 4\mathrm{j}+\mathrm{X}) = 330 \pm 5 \text{ (stat.)} \pm 45 \text{ (syst.) nb.}$$

This value is compared in Table 10.1 with the theoretical predictions, described in Chapter 4. The predictions have been obtained at the stable particle level and show quite different results: PYTHIA8 tune 4C gives a value for the cross section which is larger than the measured one, while the MADGRAPH generator, interfaced with PYTHIA6, tune Z2*, instead, predicts a too low value compared to the data. The best description of the data is provided by the predictions obtained with HERWIG++ and SHERPA. The prediction of POWHEG, producing a NLO dijet ME and interfaced with PYTHIA6, tune Z2', including MPI, is a bit larger but still in agreement with the measurement. Note that no estimation of uncertainties related to the choice of PDF or renormalization and factorization scale has been calculated for any of these predictions: this procedure would imply not only a change of the scales but also a variation of the UE tunes. This is not crucial for the measurement itself and, hence, data have been only compared to predictions using specific settings. In order to understand the different predictions and the effect of the UE, predictions have been also generated with different settings. For instance, it has been found that the differences between the predictions obtained with MADGRAPH and SHERPA are due to the different contributions coming from MPI in the UE tunes, while results without MPI agree with each other. Note that MADGRAPH implements a LO 2→4 ME while SHERPA generates a LO 2→3 ME, producing the last selected jet through the parton shower: the previous result shows that the PS in SHERPA is able to correctly reproduce additional hard emissions, while the key point is the contribution of MPI.

Sample	PDF	Cross section (nb)
PYTHIA8, tune 4C [124]	CTEQ6L1 [68]	423
HERWIG++, tune UE-EE-3 [117, 127]	MRST2008LO** [183]	343
MADGRAPH + PYTHIA6, tune Z2* [180]	CTEQ6L1 [68]	234
SHERPA tune [122]	CTEQ6L1 [68]	293
POWHEG + PYTHIA6, tune Z2'	CT10 [67]	378
Data	-	330 ± 5 (stat) ± 45 (syst)

Table 10.1: Cross sections for MC predictions and measured data for pp → 4j+X: the four jets are selected within $|\eta| < 4.7$, and with $p_T > 50$ GeV for the two leading jets and $p_T > 20$ GeV for the other jets.

The cross sections have been measured differentially as a function of p_T and η of each of the four jets and are presented in Figure 10.7. The cross sections are falling with increasing p_T for all the jets in the final state. For the hard jets, which have $p_\mathrm{T} > 50$ GeV, the cross section decreases by two orders of magnitude for p_T between 50 and 200 GeV. For the soft jets with $p_\mathrm{T} > 20$ GeV, the cross section decreases over five orders of magnitude for the same p_T range. The behaviour of the cross section as a function of η exhibits some differences among hard and soft jets. In particular, since hard jets are mainly contained in the most central region, the cross section drops very rapidly when going to forward pseudorapidities, with two orders of magnitude of difference between $|\eta| \sim 0$ and $|\eta| > 4.0$. Instead, the distributions of the soft jets are flatter, with cross sections dropping by only about an order of magnitude between central ($|\eta| \sim 0$) and forward region ($|\eta| \sim 4.7$).

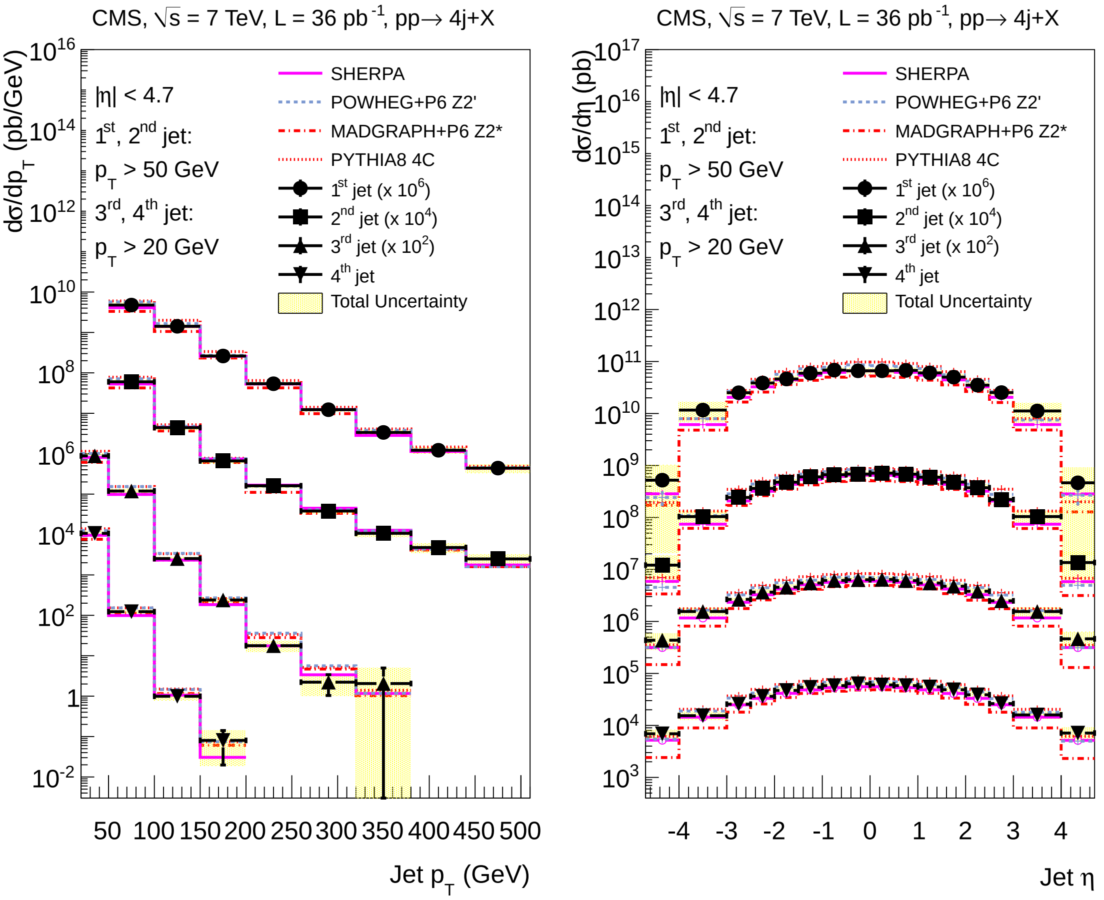


Figure 10.7: Differential cross sections as a function of the jet transverse momenta p_T (left) and pseudorapidity η (right) compared to predictions of Powheg, Madgraph, Sherpa, *and* Pythia8. *Scale factors of 10^6, 10^4 and 10^2 are applied to the measurement of the leading, subleading and third jet, respectively. The yellow band represents the total uncertainty, including the statistical and systematic components added in quadrature. Results are published in [11].*

The measured differential cross sections are also compared to theoretical predictions. Ratios between the predictions and the measurements are presented in Figure 10.8. All predictions, except HERWIG++, are in agreement with the measurement for the leading and subleading jets at large transverse momenta $p_{\mathrm{T}} \sim 300$ GeV (Figure 10.8(left)). However, predictions start to differ at smaller p_{T}: while PYTHIA8 and MADGRAPH deviate significantly from the data up to 30%, POWHEG and SHERPA are in agreement with the measurement for the leading and subleading jets. The soft jets are not very well described by all predictions: POWHEG and PYTHIA8 are about 10–30% above the measurement, while the predictions from MADGRAPH are below the data in the first p_{T} bin. The SHERPA generator offers the best agreement with the data. The HERWIG++ generator, even though it correctly predicts the value of the total cross section (Table 10.1), achieves a good agreement at small p_{T} but a worse description at large p_{T}. The distributions of the soft jets are also not optimally reproduced by HERWIG++, which overshoots the data points by around 20% in the p_{T} region between 50 and 150 GeV.

The differential cross sections as a function of η are shown as ratios between data and theoretical predictions in Figure 10.8(right). Distributions of leading and subleading jets are described by SHERPA and HERWIG++ within the systematic uncertainties. The predictions obtained with MADGRAPH underestimate the measurements, because of the different total cross sections, shown in Table 10.1, but they predict the correct shape of the distributions. The POWHEG and PYTHIA8 generators tend to be below the measurement at large η. The measurements of the soft jets are difficult to be reproduced for ranges of $|\eta| > 3$, which is described only by SHERPA and HERWIG++. Experimental issues, encountered in the forward region, have been already discussed in Chapter 6.

The normalized differential cross sections have been measured as a function of the correlation observables, defined in Chapter 4. The distribution as a function of $\Delta\phi^{\mathrm{soft}}$ is shown in Figure 10.9(top left). It exhibits a maximum at $\Delta\phi^{\mathrm{soft}} \sim \pi$ and it has a tail down to low $\Delta\phi^{\mathrm{soft}}$ values; in particular, it falls by less than an order of magnitude towards $\Delta\phi \sim 0$. Large $\Delta\phi$ translates into a back-to-back jet configuration and highly correlated jet pairs, while at small $\Delta\phi$ the jets are less correlated. The local maximum, visible at values around $\Delta\phi \sim 0.5$–0.8, is due to the jet distance parameter of 0.5, chosen for the anti-k_{T} jet algorithm. Jets, which originate from collinear parton emissions with an angular separation of less than 0.5 units in the η-ϕ space, are, by definition, merged by the clustering algorithm. In Figure 10.9(top right), the balance in transverse momentum between the soft jets, $\Delta^{\mathrm{rel}}_{\mathrm{soft}}p_{\mathrm{T}}$, is shown. It exhibits a maximum around 1, and it covers about an order of magnitude over the whole spectrum. Increasing values of $\Delta^{\mathrm{rel}}_{\mathrm{soft}}p_{\mathrm{T}}$ indicate unbalanced configurations of the soft jets. In particular, the highest bins are expected to be filled by jets coming from radiation of the initial- or final-state of the hard pair of jets. The cross section as a function of the azimuthal angle between the planes of the two dijet systems, $\Delta\mathrm{S}$, is shown in Figure 10.9(bottom). The distribution reaches the maximum at π and falls over almost two orders of magnitude over the entire phase space. At high $\Delta\mathrm{S}$ values, the dijet systems are correlated; the correlation progressively reduces for decreasing $\Delta\mathrm{S}$ values.

In Figure 10.9, the normalized differential cross sections as a function of the correlation observables are also compared to theoretical predictions. In addition to the previous ones, the distributions obtained by POWHEG+PYTHIA6, tune Z2' with MPI switched off, are also shown in the plots. The cross section as a function of $\Delta\phi^{\mathrm{soft}}$ is well described by all predictions; also the POWHEG prediction without MPI is able to reproduce the behaviour. The normalized differential

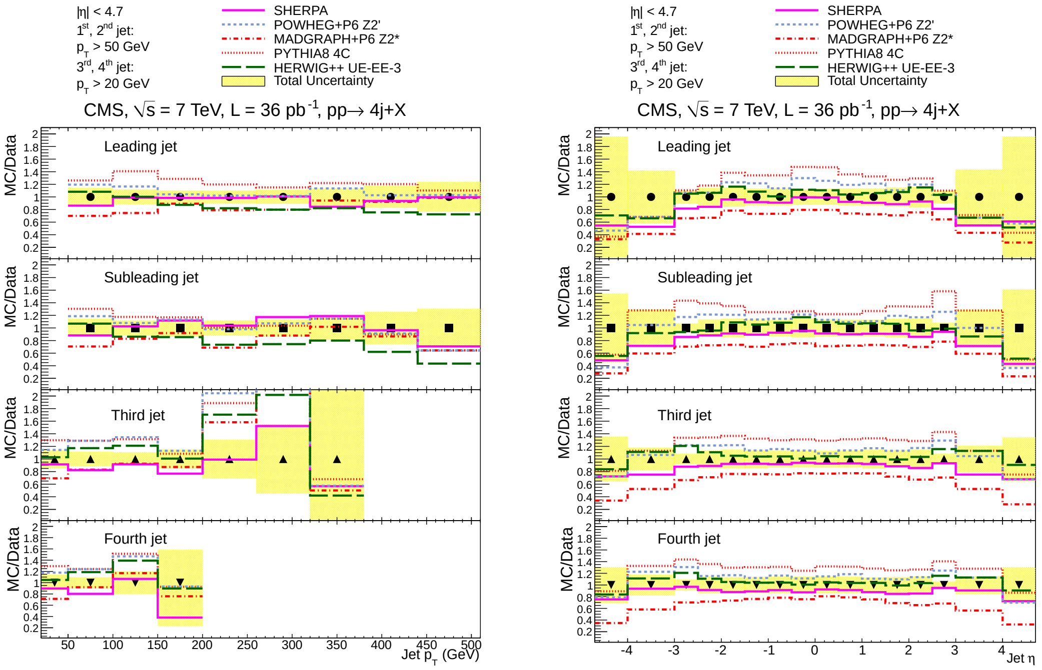


Figure 10.8: Ratios of predictions of Powheg*,* Madgraph*,* Sherpa*,* Pythia*8 and* Herwig++ *to data as a function of the jet transverse momenta p_T (left) and pseudorapidity η (right) for each specific jet. The yellow band represents the total uncertainty, including the statistical and systematic components added in quadrature. Results are published in [11].*

cross section as a function of $\Delta^{\rm rel}_{\rm soft} p_{\rm T}$ is reasonably described by all predictions except in the very first bins, where significant differences appear. The prediction of Powheg without MPI underestimates the measurements and shows clearly the need of additional contributions in this region. None of the predictions correctly describe the normalized differential cross section as a function of ΔS. In the range ΔS < 2.5, predictions from Sherpa are above the data while all other predictions are significantly below the measurement, of about 20–30%. This means that all predictions, except Sherpa, tend to predict a more correlated scenario for the jet pairs than observed in the measurement. The prediction from Powheg without MPI is several standard deviations away from the measurement at small ΔS: this is the phase space where a DPS signal is expected. Since all the predictions, except Sherpa, undershoot this region, it might indicate a too low DPS contribution in the UE simulation. In Sherpa, instead, whose predictions are above the data for ΔS < 2.5, a reduction of the MPI contribution seems to be needed to improve the description of the data: this observation is consistent with the UE results, shown in Figure 10.4(bottom).

The results, described in this Section, can be complemented by the measurements of similar observables in the 2b2j channel, which add additional information to the considered multijet scenarios; results of the 2b2j channel are presented in the following Section.

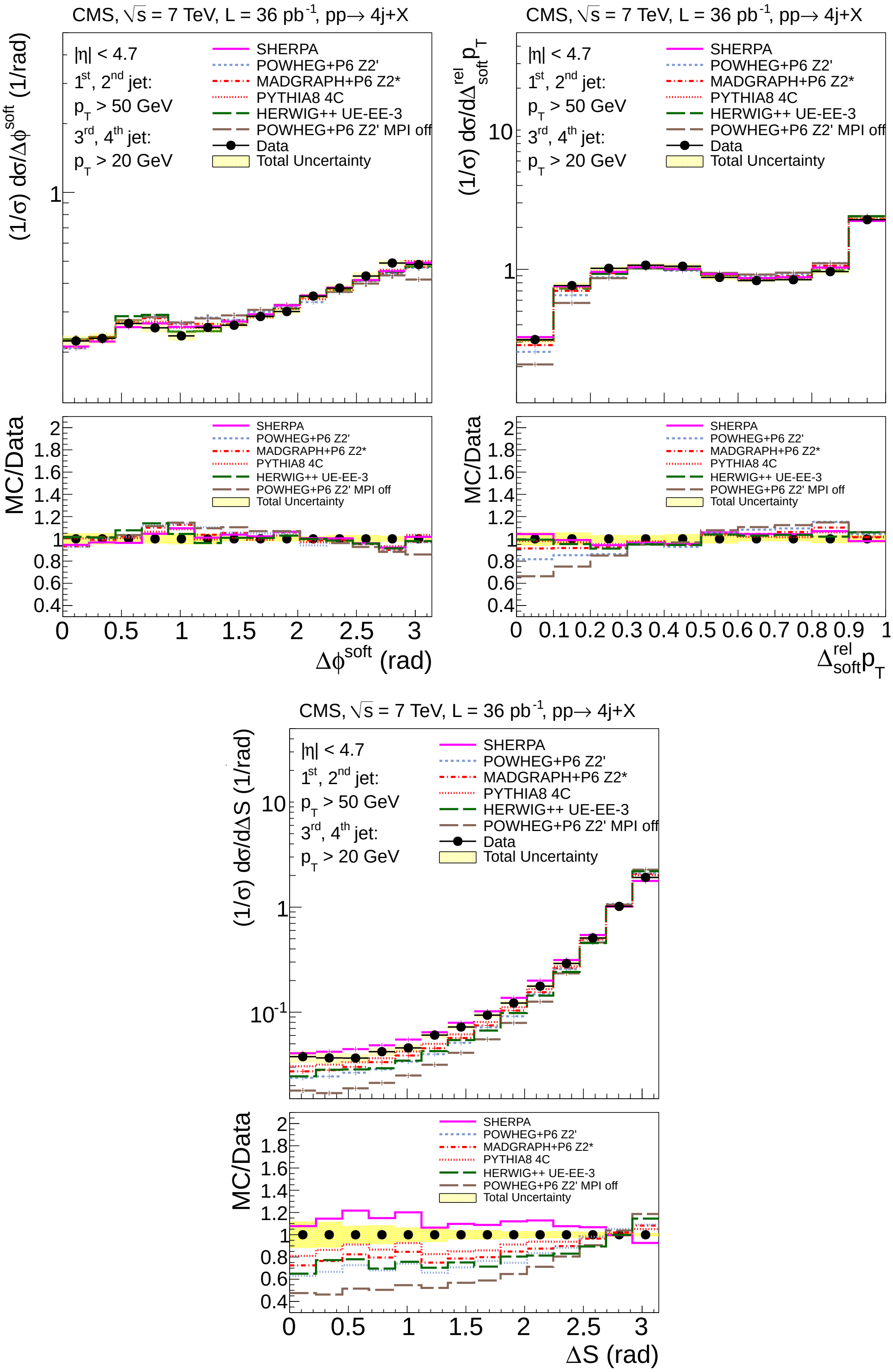


Figure 10.9: Normalized differential cross sections as a function of the difference in azimuthal angle $\Delta\phi^{soft}$ (top left), $\Delta^{rel}_{soft}p_T$ (top right), and ΔS (bottom) compared to the predictions of Powheg, Madgraph, Sherpa, Pythia*8 and* Herwig++*. A comparison with the* Powheg *predictions interfaced with the parton shower* Pythia*6 tune Z2' without MPI is also shown. The lower panel shows the ratios of the predictions to the data. The yellow band represents the total uncertainty, including the statistical and systematic components added in quadrature. Results are published in [11].*

10.3 2b2j selection

Measurements of absolute and normalized differential cross sections for the channel with 2b2j in the final state add additional information about the heavy flavour sector. The total cross section is measured and compared to various predictions provided by different generators, Pythia6, Pythia8, Herwig++ with a LO 2→2 ME, Powheg, using a NLO 2→2 ME, and Madgraph, which implements a LO 2→4 ME. The predictions are further subdivided in distinct groups:

- Nominal tunes: samples where the tunes for the UE simulation have been used without modifications;
- MPI off: samples in which the MPI contribution is switched off;
- Heavy Flavour (HF): samples in which the *b*-quarks are generated with a ME which includes their mass and not in a massless regime, like in the nominal tunes. Note that in a LO 2→2 ME with a HF production, only the pair creation diagrams are generated, while the gluon splitting and the flavour excitation processes are not included.

The measured value of the cross section for the selected 2b2j final state is:

$$\sigma(\mathrm{pp} \to 2\mathrm{b}+2\mathrm{j}+\mathrm{X}) = 67.2 \pm 2.2\ (\mathrm{stat.}) \pm 22.5\ (\mathrm{syst.})\ \mathrm{nb}.$$

The predictions are shown in Table 10.2. Similar conclusions as in the 4j analysis are observed. While Pythia8, Pythia6 and Powheg+Pythia6 Z2* are above the data, the best agreement is achieved by Herwig++, Madgraph+Pythia6, and, even though with a slightly worse level of compatibility, by Powheg+Pythia6 Z2'. Differences observed in the predictions obtained with Powheg when two different tunes, namely Z2* and Z2', are used, show the strong sensitivity to the parameters of the UE simulation. The Powheg generator, with a HF ME interfaced with Pythia6 tune Z2', is also in good agreement with the measured value. The cross section predicted by a HF production generated with Pythia8 is instead much lower than the measured one: this indicates the importance of flavour excitation and gluon splitting diagrams at lowest order, not included in this calculation. This is also observed when the HF generation is set in Madgraph. The prediction obtained with Madgraph without the contribution of MPI is below the measured value. Predictions from Sherpa could not be obtained for a QCD multijet production with *b*-jets, because of the long computing time.

Additionally, predictions have been also obtained with Pythia8 for an “exclusive and high threshold” scenario with *b*-jets, in order to mimic the phase space selection applied in the 4j scenario[1]. The “exclusive and high threshold” selection refers to the following analysis cuts:

- exactly four jets in $|\eta| < 4.7$
- two *b*-jets with $p_{\mathrm{T}} > 20$ GeV
- two additional jets with $p_{\mathrm{T}} > 20$ GeV

The value of the cross section obtained with Pythia8 for this scenario is 4.57 nb, while for the same selection without *b*-jets it was 423 nb (see Table 10.1). This means that requesting a *b*-jet

---

[1] Note that this could not be performed in the data because the phase space available for the selection of *b*-jets is limited to $|\eta| < 2.4$, and it is not then possible to extend it up to the forward region.

Sample	Cross section (nb)
PYTHIA6 tune Z2*	121.31
HERWIG++ tune UE-EE-3	69.67
PYTHIA8 tune 4C	132.74
PYTHIA8 tune 4C MPI off	72.85
PYTHIA8 tune 4C Heavy Flavour	8.35
POWHEG+PYTHIA6 tune Z2'	99.79
POWHEG+PYTHIA6 tune Z2' MPI off	56.83
POWHEG+PYTHIA6 tune Z2*	144.02
POWHEG+PYTHIA6 tune Z2* Heavy Flavour	85.51
MADGRAPH+PYTHIA6 tune Z2* Heavy Flavour	29.14
MADGRAPH+PYTHIA6 tune Z2*	58.10
MADGRAPH+PYTHIA6 tune Z2* Heavy Flavour MPI off	18.20
MADGRAPH+PYTHIA6 tune Z2* MPI off	37.08
Data	67.2 ± 2.2 (stat.) ± 22.5 (syst.)

Table 10.2: Comparison of measured and predicted cross sections in the 2b2j channel.

pair suppresses the cross section by a factor of about 100. This value is very close to the ratio between inclusive b-jet and inclusive jet cross section, measured in [207].
The differential absolute and normalized cross sections have been also measured as a function of jet p_T and η, along with jet correlation observables in azimuthal angle, pseudorapidity and p_T balance. Results are compared to the predictions of different event generators at the stable particle level. The LO generators, PYTHIA6, HERWIG++ and PYTHIA8 and the NLO dijet ME generated with POWHEG interfaced with the parton shower provided by PYTHIA6 are considered. Predictions with two different tunes are used for POWHEG: the standard Z2* and the newly introduced Z2', where the phase space for the hard radiation is reduced with respect to the nominal tune[1].

In Figure 10.10, p_T and η spectra are shown. The jet transverse momenta are all rapidly decreasing over five–six orders of magnitude up to 500 GeV. No significant differences are observed between jets of different flavour. Rather flat distributions are observed for the b-jet differential η measurements, selected in the central region: a factor of 2 is obtained between the cross sections at $|\eta| < 0.5$ and at the edges of the acceptance region ($2.0 < |\eta| < 2.4$). The cross section for the additional jets, whose acceptance region goes up to $|\eta| = 4.7$, decreases by an order of magnitude between central and forward pseudorapidities but are similar in $|\eta| < 2.4$.

Figure 10.11 shows the ratios between the theoretical predictions and the data for p_T and η distributions of all jets. Predictions from POWHEG+PYTHIA6 interfaced with the tune Z2*, PYTHIA8 and PYTHIA6 overshoot the p_T spectra by factors of 1.5–2.5: the largest differences are mainly observed in the low p_T region, below 200 GeV. The use of the new tune Z2' for the POWHEG ME helps to improve the description of the measurements but it is not optimal. In particular, the leading additional jet is still overshot by a factor 1.5 over the entire phase space. The HERWIG++ event generator predicts a too low contribution at high p_T for all jets. These effects have already been seen in the 4j analysis. The η spectra are well reproduced in the central region, taking into account that some of the predictions fail in normalization with respect to the data. In the forward

---

[1] Predictions obtained with MADGRAPH are not included in the comparisons of differential cross sections because of a low number of selected events available for jet $p_T > 250$ GeV.

region, instead, the data are described only by Pythia and Herwig. Powheg underestimates the jet content in the region of $|\eta| > 3$. Figure 10.12 helps to further understand the results, by comparing the ratios of the normalized differential cross sections between data and simulation, as a function of p_{T} and η. Interesting is to note that Powheg+Pythia6 tune Z2' is able to describe the p_{T} normalized distributions over the whole phase space, while Herwig++, even though it correctly predicts the value of the total cross section, is the worst in describing the high p_{T} region of all jets. The normalized cross sections as a function of η are well reproduced by all predictions for the b-jets while the additional jets are underestimated by Powheg in the forward region, independently on the tune used for the UE simulation.

Figure 10.13 shows the correlation observables between the selected jets. The $\Delta\phi$ variables are very similar for the differently flavoured jets. The distribution is rather flat with a little increase towards high values ($\sim \pi$) corresponding to a back-to-back configuration for the jet pairs. This behaviour of $\Delta\phi$ for both jet pairs is very different from a similar distribution obtained in dijet events [93], which shows a large correlation between the jets with a peak at values $\sim \pi$. The difference observed in the 2b2j channel is due to the requirement of additional jets, which introduces a stronger decorrelation for the selected objects. A small increase for values around $\Delta\phi \sim 0.5$–0.8 is observed for both distributions. This is due to collinear jet emission which starts to be resolved at $\Delta\phi > 0.5$ because of the width of the jet clustering cone. The $\Delta^{\mathrm{rel}} p_{\mathrm{T}}$ variables are also very similar between the two jet pairs. Both of them increase towards values around 1, which correspond to correlated jet configurations. Differences between observables related to the jet pairs start to appear for $\Delta\eta$: this is mainly due to the different η acceptance of the jet pairs. Bottom jets are preferably very close in η and the cross section decreases rapidly from $\Delta\eta \sim 1.5$ until the edges. The $\Delta\eta$ relative to the additional jets presents a broader distribution around the maximum and the cross section starts to decrease only at $\Delta\eta > 3$. The ΔS observable, which is the most DPS-sensitive one, has a falling distribution from correlated configurations at high values, down to uncorrelated jet topologies at low values. The differential cross sections of the correlations observables have been also compared to MC predictions. In general, the best description is achieved by Pythia8 and Herwig++ for all the distributions in Figure 10.13. Note that for low values of the $\Delta^{\mathrm{rel}} p_{\mathrm{T}}$-based observables, the predictions from Powheg are far below the data. In these regions, the data are in agreement with the predicitons from Herwig++ and Pythia8. The other predictions undershoot the data. The $\Delta\eta$ observable are well described by Herwig++. The Powheg+Pythia6 Z2' predictions do not describe the high $\Delta\eta$ region where the data are underestimated by an order of magnitude. ΔS is not very well described by any prediction: in particular, all of them do not follow the decrease at low values. Predictions obtained with Pythia8 tune 4C without the contribution of MPI have been also added for comparison. While they reproduce quite well the shape of observables less sensititve to DPS, like $\Delta\phi$ or $\Delta\eta$, with a level of agreement very similar to the predictions of the nominal tune 4C, they start to fail in the description of ΔS and $\Delta^{\mathrm{rel}} p_{\mathrm{T}}$ which have been proven to be more sensitive to signals from hard MPI. In particular, the region at low values of $\Delta^{\mathrm{rel}} p_{\mathrm{T}}$ is underestimated by about 20–30% and the same behaviour is observed for the whole spectrum of ΔS, where at low values the measurement is undershot by 50%. These regions are the ones where a DPS signal is expected. The comparisons of predictions with and without MPI might be an indication for the need of a higher contribution of additional hard partonic interactions in the simulation.

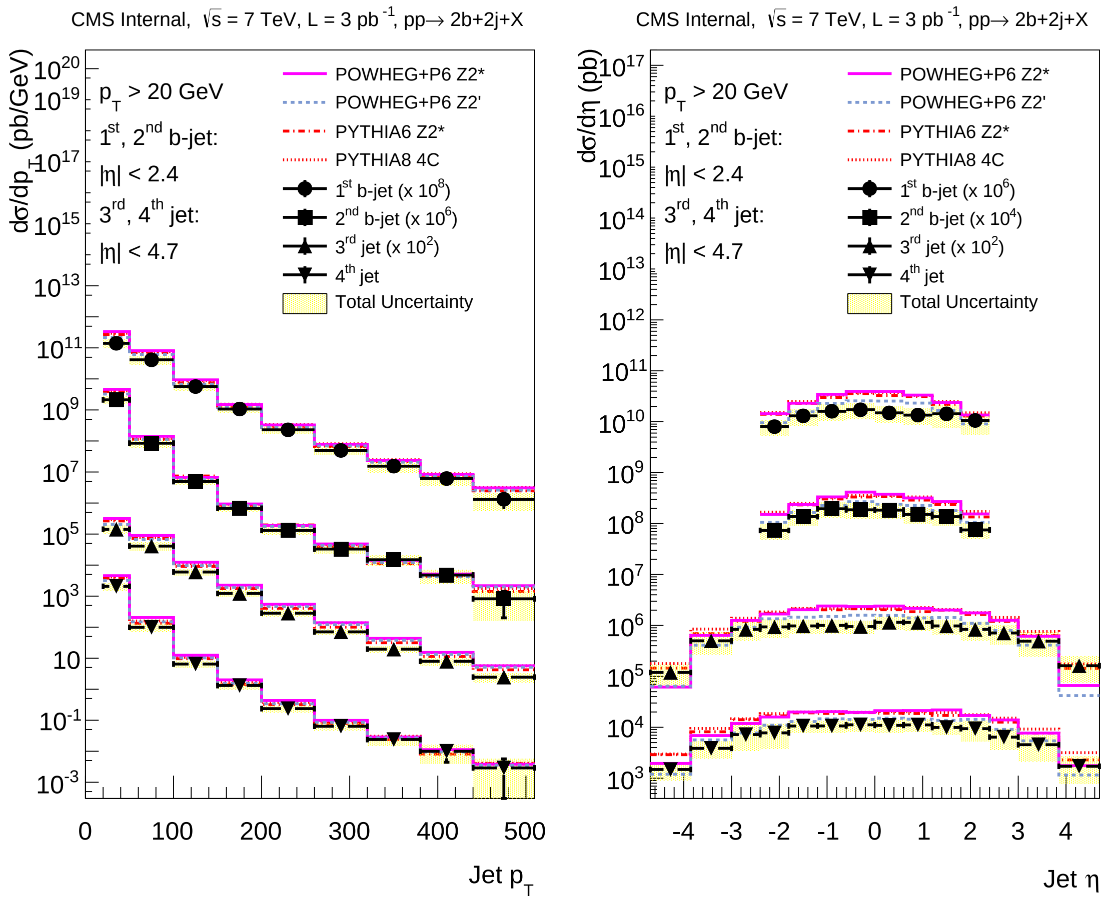


Figure 10.10: Differential cross sections as a function of the jet transverse momenta p_T (left) and pseudorapidity η (right) compared to predictions of POWHEG, PYTHIA*6, and* PYTHIA*8. Scale factors of 10^8, 10^6 and 10^2 are applied to the measurement of transverse momentum of the leading, subleading and third jet, respectively. Scale factors of 10^6, 10^4 and 10^2 are applied to the measurement of pseudorapidities of the leading, subleading and third jet, respectively. The yellow band represents the total uncertainty, including the statistical and systematic components added in quadrature.*

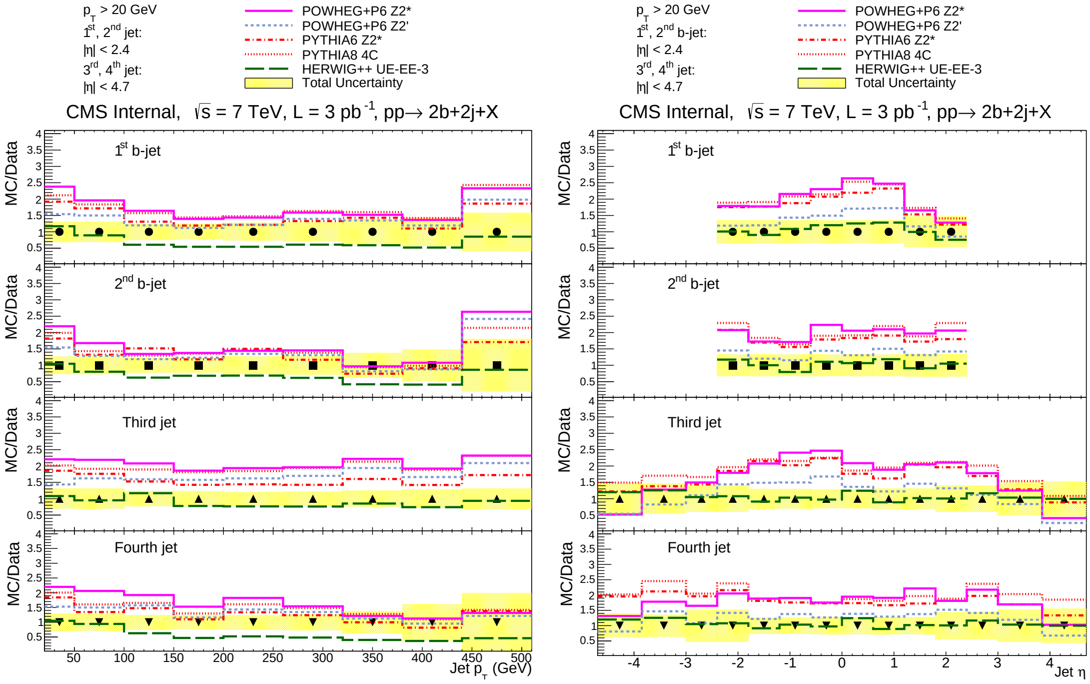


Figure 10.11: Ratios of predictions of Powheg, Pythia*6*, Pythia*8 and* Herwig*++ to data for absolute differential cross sections as a function of the jet transverse momenta p_T (left) and pseudorapidity η (right) for each specific jet. The yellow band represents the total uncertainty, including the statistical and systematic components added in quadrature.*

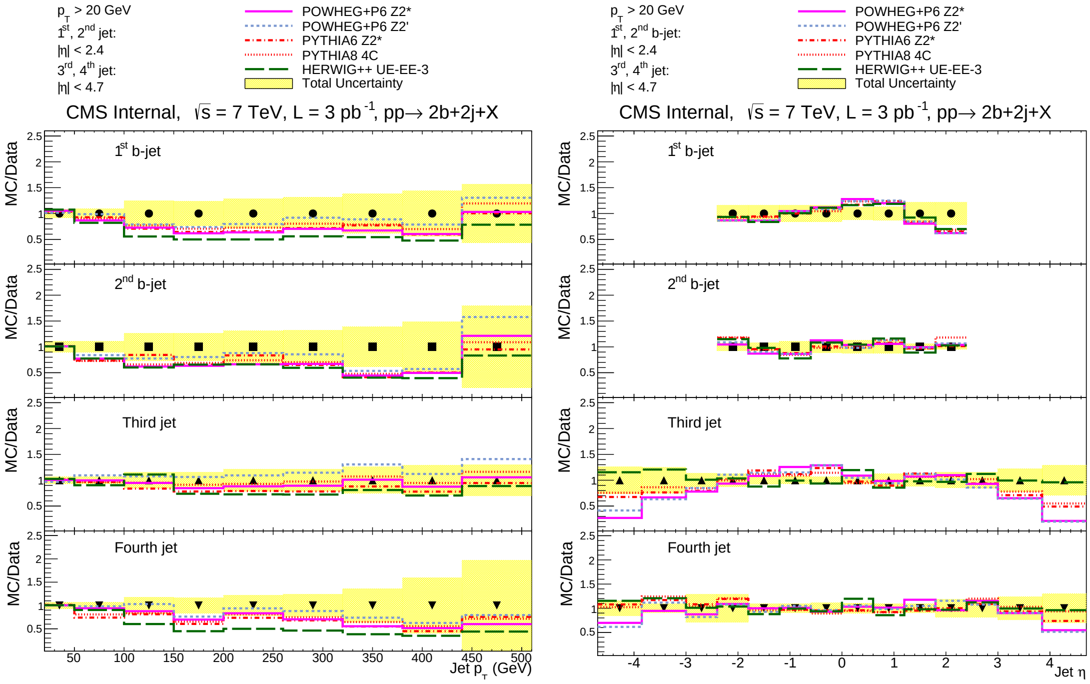


Figure 10.12: Ratios of predictions of Powheg, Pythia*6*, Pythia*8 and* Herwig*++ to data for normalized differential cross sections as a function of the jet transverse momenta p_T (left) and pseudorapidity η (right) for each specific jet. The yellow band represents the total uncertainty, including the statistical and systematic components added in quadrature.*

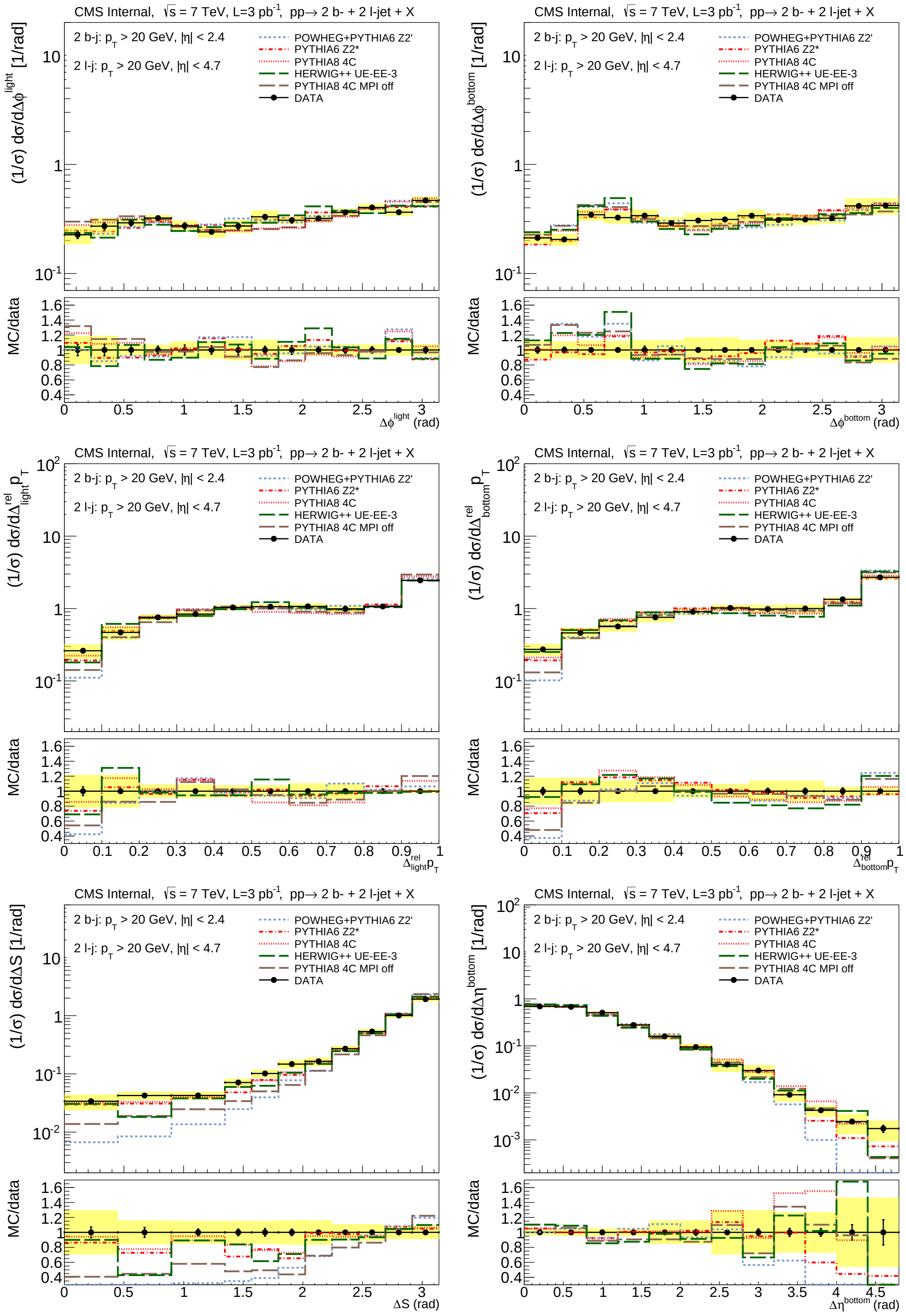


Figure 10.13: Normalized differential cross sections as a function of the differences in azimuthal angle $\Delta\phi^{light}$ (top left) and $\Delta\phi^{bottom}$(top right), of the p_T balance, $\Delta^{rel}_{light}p_T$ (middle left) and $\Delta^{rel}_{bottom}p_T$ (middle right), of ΔS (bottom left) and of the difference in pseudorapidity, $\Delta\eta^{bottom}$ (bottom right) compared to the predictions of POWHEG, PYTHIA*6* PYTHIA*8 and* HERWIG++. *The lower panel shows the ratios of the predictions to the data. The yellow band represents the total uncertainty, including the statistical and systematic components added in quadrature.*

10.4 Summary of the results

Measurements of four-jet final states have been performed for the first time at the CMS experiment in proton-proton collisions at a center of mass energy of 7 TeV. It uses the full detector coverage in pseudorapidity, with jets measured up to $|\eta| = 4.7$, and $p_{\mathrm{T}} > 20$ GeV. The performance, reached by the detector for such challenging physics objects in a very populated environment is remarkable: in particular, it also allows measurements of correlations between the jets in the final state with a very high precision. This goal has been achieved also for selections requiring jets originating from b-quarks: the discrimination power of the algorithm, developed in CMS and able to separate jets of different flavours, is also considerable.

In addition to the excellent detector output, the measurements can also profit from a large amount and variety of available predictions for the understanding of the data. From comparisons between data and predictions obtained with various ME and UE simulations, it has been possible to extract results about the sensitivity of the measurements to these components. Further interesting comparisons might also be to predictions of a NLO 2→4 ME, interfaced with PS, as implemented in the NJet [225] or BlackHat [226] event generators, or of a 2→2 ME, interfaced with a parton evolution following the CCFM equation (see Chapter 1), as used in Cascade [227].

The results, shown for the 4j and the 2b2j channel, have given very interesting insights for multijet scenarios and can be summarized in the following points:

- most of the considered theoretical predictions are able to describe very well the jet p_{T} and η absolute differential cross sections, in a wide phase space; a reasonably good description is also provided by generators like Pythia or Herwig, even though they use LO 2→2 ME interfaced with PS. This is not surprising for the 4j selection where the jets are separated in pairs at different scales (50 and 20 GeV); this is a typical DGLAP scenario, which is expected to be well reproduced by event generators based on the simulation of a hard ME interfaced with a DGLAP evolution. The hard jets come from the hard scattering and the softer ones are produced via PS. When the hard p_{T} threshold is decreased to the same value of the soft one, like in the 2b2j measurement, generators have more difficulties to reproduce the data; in particular, the description in the low p_{T} region, where larger contributions are expected from softer radiation and MPI becomes more critical;

- the multijet scenario in the heavy flavour sector can be also well described by different theoretical predictions, even though with a lower level of agreement than in final states with four jets without any flavour requirements. It has been demonstrated that 2→2 processes are not sufficient to describe the b-jet production at the LHC and higher-order diagrams clearly need to be included to take into account emissions of hard b-jets. These can be simulated either by contributions in PS or by including additional hard partons in the ME; the predictions of these two approaches deviate quite significantly between each other in terms of absolute cross sections. The shapes of single jet p_{T} and η are well reproduced by most of the considered models, with a slight improvement observed for predictions which simulate hard parton emissions within the ME formalism. Predictions which treat b-quarks in a massive or a massless regime, within the ME formalism are both in agreement with the data. This is mainly due to large experimental uncertainties, which the measurements are affected from;

- jet measurements performed in the forward region have led to interesting conclusions as well, in both considered multijet channels: all considered models are not optimal in reproducing data at $|\eta| > 3$, showing systematically lower predictions. This effect increases when looking at hard jets, with $p_{\mathrm{T}} > 50$ GeV, and for ME generators, like Powheg and Madgraph, where the PS is provided by an external interface with standard generators, like Pythia. This shows that there is plenty of room for improvement in the description and the understanding of data in this region. At present, comparisons between various predictions and UE data have already shown the same trend for observables measured in the forward region [12]: in particular, tunes, which are well performing in the description of measurements in the central region, ceases to do so at forward pseudorapidities. However, no systematic studies have been performed about validation of tunes used for Powheg, Madgraph and Sherpa, as described in Section 10.1, against observables measured in the forward region;

- the wide variety of considered predictions has shown that the correct description of the data is closely related to a correct admixture of ME and PS contributions: four-jet measurements appear to be quite sensitive to choices of tunes and ME generators. In particular, for a Powheg ME, it has been shown that a significant role is played by the phase space allowed for the PS. It is, however, crucial to determine a meaningful UE simulation for specific ME calculations, in the sense that they must be able to reproduce basic UE- and jet-related data. Only in this way, conclusions for a more complicated and elaborate selection can be extracted. Four-jet measurements have given a great opportunity for this studies, performed with different ME generators, namely Powheg, Madgraph and Sherpa; they have shown that the UE simulation used for LO 2→2 ME generators is still well performing when interfaced with higher-order ME. This is a very promising indication of a unified view of the UE, which remains independent on how the hard scattering is treated. The studies of four-jet measurements led specifically to some outcomes about theoretical models. On one hand, the need of NLO calculations: the description of the data significantly improves when using, for instance, a NLO Powheg ME, interfaced with Pythia, with respect to the level of agreement achieved by Pythia standalone. On the other hand, the reliability of the PS simulation, implemented in the MC generators within a DGLAP approach, can well reproduce the results of hard emissions included directly in the ME (for instance, by looking at the compatibility between predictions of 2→3 ME and 2→4 ME, both interfaced to PS). The role of additional parton interactions is also a central component: an increased MPI contribution, implemented in the tune used for Sherpa, plays a key role in the description of the low p_{T} region, while Madgraph, using a Pythia6 tune with a relatively low amount of MPI, fails in that phase space. Note that the description of UE data is not optimal for Sherpa, though, indicating a too large contribution from the MPI of around 10–20%, confirmed by the predictions for the ΔS observable. The interplay among all described elements is, however, very delicate and complicated and further multijet measurements may indeed improve the general overview and help to understand and constrain models and parameters, both in the light and heavy flavour sectors;

- measurements of correlation observables in both selections have shown the clear need of hard MPI, namely DPS contributions, where two or more hard scatterings occur within the same hadronic collisions; the addition of this component in MC event generators seems to improve the description of the data but the amount of DPS, currently included in the tunes, does not yet give to a satisfactory agreement for the most sensitive variables. The measured correlation observables help to estimate how big the DPS component, which contributes to each considered channel, should be.

Chapter 11

Extraction of the DPS contribution

With the results of the 4j and 2b2j scenarios, the collection of measurements in DPS-sensitive channels is largely enriched and additional interactions occurring in proton-proton collisions may be further investigated. This can be done by extracting from the measurements the value of σ_{eff}, namely the amount of DPS which contributes to the channels. This Chapter is dedicated to the determination of σ_{eff}, based on the distributions presented in Chapter 10. In Section 11.1, the methods used in past analyses for σ_{eff} measurements are presented and their strengths and weaknesses are discussed; a different MC-based method, introduced for the new physics channels, is then described in full detail. In Section 11.2, the new method is applied to the W+dijet channel using the CMS measurement at 7 TeV and the result is compared to the value of σ_{eff} obtained with another independent method. In Section 11.3 and 11.4, results are presented for the 4j and the 2b2j scenarios. Finally, in Section 11.5 and 11.6, an attempt of including the new DPS-based findings in a more general framework is presented, and the results are interpreted.

In Appendix E, additional material relative to the determination of σ_{eff} with the new method is provided: operative and technical details of the procedure of extracting σ_{eff} from different MC event generators are described and supplementary comparisons are shown for the two multijet channels.

11.1 Methods for extraction of σ_{eff}

In the literature, different methods have been used for the extraction of the DPS contribution, which translates into a value of σ_{eff}: all of them are based on measurements of differential cross sections as a function of observables which might be sensitive to a DPS signal, like variables expressing correlations in the transverse plane among the physics objects of the final state. From these measurements, then, templates for background and signal are built and their relative fractions are determined through a fitting procedure: the results correspond to the combination of background and signal which best describe the DPS-sensitive observables. The background template includes events coming from SPS events, while in the signal one, DPS processes are collected. With the fraction of DPS signal contained in the data, σ_{eff} can be obtained with a simple formula which also accounts for selection efficiencies and cross section values of the selected processes [8, 9]. This approach of σ_{eff} determination is generally called the "template method". So far, in all experiments aiming to measure DPS contribution, the template method was adopted, without any attempt of a different conceptual approach. The differences between the various measurements have been just the definition of background and signal templates. For instance, the D0 and CDF measurements mainly rely on data-driven signal definitions [7, 112, 113]; DPS-like events are selected from two independent collisions recorded in data, carefully corrected for possible contamination from pile-up. The physics objects of the final state are then selected pair-wise in the two different events, mimicking, thus, the occurrence of a DPS in the same collision. This definition is obviously based on the assumption for a DPS event of having two independent interactions in colliding protons. The background template, instead, relies on events generated on a MC basis, by switching off the simulation of MPI. The CMS and ATLAS collaborations refined this method by comparing results obtained from additional template definitions. For instance, for the signal definition, DPS events are also evaluated from MC simulations performed with PYTHIA8 when two hard scatterings are forced to occur (see Chapter 2). The SPS background, instead, is built again only based on simulation, by generating a higher-order ME (like, for instance, a 2→3 or 2→4 one) and setting an upper scale for the MPI, not simply switching them off: this is a better solution, since a scenario without MPI is, indeed, too unrealistic. In particular, the upper scale depends on the p_{T} threshold of the physics objects of the final state, which are expected to come from the second hard scattering: the PYTHIA8 and the HERWIG6+JIMMY event generators have been used for the background definition, respectively in the CMS and ATLAS measurements.

The criteria for the definition of background and signal templates can be summarized as follows [228]:

(a) use more than one MC generator in order to properly estimate the systematic uncertainty;

(b) make sure that signal and background cover the full phase space: this translates into the fact that the background template is not a sample with MPI switched off, but rather a sample with a second interaction below a given scale;

(c) select an inclusive scenario, rather than an exclusive one, namely set a physics selection which allows any number of additional interactions, and not only two;

(d) investigate the dependence of the background template on the generator used for its definition: it might happen that the inclusion of a higher number of partons in the ME fills a similar region of the phase space as the DPS signal; if this is the case, a choice of a ME which guarantees stability of the obtained results has to be taken;

The template method has, however, aspects which complicate its applicability: these are both operative and conceptual and mostly related to the construction of the templates. First of all, the signal and background definition: the definitions adopted by CMS and ATLAS seem to be more appropriate than the ones used in CDF and D0[1], but they can be pursued only by the use of a few MC generators (and only PYTHIA8 among the most recent ones). Only few of them have control of the MPI at the parton level and allow the selection of events with a certain scale of additional interactions. This feature complicates the estimation of the model dependence, namely of the evaluation of the uncertainty coming from the particular choice of the generator. This is due

---

[1] This is because the background definitions adopted by CDF and D0 do not fulfill (b).

to the fact that the same background definition can not be used by all MC generators. The current model uncertainty estimated in CMS and ATLAS relies on background templates where the MPI are switched off, generated for instance with PYTHIA6, giving large uncertainties, which are the dominant source of the systematic effects[1]. A similar problem appears for the signal template, whose effects are however much smaller.

Another important issue additionally arises from the background definition: it is related to a "philosophical" problem of considering some events in the background template rather than in the signal one. Imagine, for instance, to select a pair of jets with $p_{\mathrm{T}} > 20$ GeV in $|\eta| < 2.0$, expected to come from a DPS event, together with the W system. The CMS and ATLAS background sample has been defined such that any event with partons coming from hard MPI with $p_{\mathrm{T}} > 20$ GeV in $|\eta| < 2.0$ is removed. Therefore, in this way, events with partons coming from hard MPI with $p_{\mathrm{T}} > 20$ GeV but in $|\eta| > 2.0$ are considered part of the background template. An unambiguous choice has to be taken for the templates but the effect of classifying events as background instead as signal needs to be evaluated.

The solution and the correct investigation of these issues is certainly possible within the template method, by evaluating all the corresponding effects of generator and phase space dependence, but it might become indeed lengthy and tedious. Because of the forementioned technical complications, for the first time inside the CMS collaboration, a different approach is proposed in this thesis and applied to DPS-sensitive measurements in order to bypass these issues and find a solution for them. The new method relies on inclusive tunes of predictions simulated with MC generators. In particular, it similarly uses measured differential cross sections, as a function of DPS-sensitive correlation observables, but it does not try to separate background and signal contributions: it rather fits the variables inclusively. In this way, no problems from the choice of the templates affect the measurement. The output of this method is the value of σ_{eff} which gives the best description of the fitted data, within a given model. Note that the procedure can be tested for any model implemented in a MC generator which has tunable parameters. A simple estimation of the model dependence of the measurement is thus feasible. An important point of the method is also that predictions obtained with the new parameter settings can be tested on the measured DPS-sensitive observables, in order to check if the agreement with the data effectively improves after the fit.

The method is performed by running the RIVET and PROFESSOR machinery, the same used for UE tunes and described in Appendix D; hence, the new method has been labelled as the "tuning method". A complete tune of a measurement is really fast to perform. Starting from reference tunes, the tuning method varies the UE part of the simulation (see Appendix D) and returns the set of parameters which best fit the correlation observables. Note that the uncertainties on the data points are considered uncorrelated in the fit. The σ_{eff} value is determined by the resulting set of parameters as well as its uncertainty. The uncertainty of the σ_{eff} measurement is obtained from the parameters given by the eigentunes (see Appendix D). The eigentunes are determined by allowing a variation of the minimized function equal to the χ^2 obtained for the best tune[2]; this choice corresponds to a conservative approach and alternatives might prefer less sizable variations,

---

[1] In CDF and D0 measurements, the adopted model uncertainty has been based on background templates simulated with different event generators but all with MPI switched off.

[2] These are the default settings of the PROFESSOR eigentunes: the result of the choice of this large χ^2 variation is a tune whose error bands are roughly the same as the uncertainties of the fitted data points.

like $\Delta\chi^2 = 1$. In this case, the resulting uncertainties would be smaller. For each set of the new parameters, a value of σ_{eff} is determined. In total, for each eigentune, an upper and a lower value are obtained; since the eigentunes are built towards orthogonal directions, the extracted σ_{eff} values are considered independent and the total uncertainty is calculated through the following formulas:

$$\Delta\sigma^{+}_{\text{eff}} = \sqrt{\sum_{i=1}^{N_{\text{P}}}[X^i(S+) - X(S)]^2} \tag{11.1}$$

$$\Delta\sigma^{-}_{\text{eff}} = \sqrt{\sum_{i=1}^{N_{\text{P}}}[X^i(S-) - X(S)]^2} \tag{11.2}$$

where N_{P} is the number of eigentunes, $X^{\text{i}}(S+)$ $(X^{\text{i}}(S-))$ is the σ_{eff} value obtained for each eigentune in the up- (down-) direction and $X(S)$ is the σ_{eff} value of the nominal tune. This approach follows the one used for the evaluation of the PDF uncertainty, described in [68].

The value of σ_{eff} depends mostly on the non-diffractive cross section and on the amount of overlap between the colliding protons. In MC generators, the overlap is regulated by the distribution of matter inside the protons, which is parametrized as a function of the impact parameter b (see Chapter 2). In PYTHIA8, different functions $f(b)$ can be selected as overlap matter distributions:

- single gaussian, with an impact parameter dependence of this type:

  $$f(b) \propto \exp(-b^2) \tag{11.3}$$

  with no free parameters;

- double gaussian, following the mathematical relation of Equation 2.18, with “coreRadius” representing the radius of the small core proton region containing a fraction “coreFraction” of the total hadronic matter;

- negative exponential, implementing a function of the form:

  $$f(b) \propto \exp(-b^{\text{expPow}}) \tag{11.4}$$

  with EXPPOW being a tunable parameter expressing the decrease of the overlap function;

- x-dependent single gaussian, extending the first case to:

  $$f(b) \propto \exp\left(\frac{-b^2}{[a_0(1 + a_1 \cdot \log(1/x))]^2}\right) \tag{11.5}$$

  with a_0, set to 0.18 by default, and a_1, tunable parameter expressing the amount of x-dependence part.

The DPS model in PYTHIA8 is based on the general assumption that the simultaneous hard interactions occurring in the same collision are independent of each other. In particular, σ_{eff} is defined as[1]:

$$\sigma_{\text{eff}} = \frac{\sigma_{\text{ND}}}{f_{\text{EF}}} \tag{11.6}$$

---

[1] It can be proven [229] that this definition follows the same assumptions of Equation 2.9

where σ_{ND} is the non-diffractive cross section and f_{EF} is the enhancement factor.
The enhancement factor depends on how central the proton collisions occur: for smaller b, they tend to be more central and with higher activity, while higher impact parameters lead to more peripheral collisions. For central collisions, the enhancement factor tends to be large, translating to a lower value of σ_{eff} and a higher DPS contribution. Conversely, for peripheral interactions, one gets small enhancement factors, high values of σ_{eff} and a little DPS contribution.
The non-diffractive cross section is fixed in PYTHIA8 to the measured value at each energy; this introduces a dependence of f_{EF} on the MPI regulator p_{T}^{0}, although weaker than the overlap matter distribution, and on the collision energy. In particular, since σ_{ND} is kept constant but is dependent on p_{T}^{0} in the model, a lower choice of p_{T}^{0} reduces the proton size, translating in a larger overlap, and hence a lower value for σ_{eff}. Conversely, a higher p_{T}^{0} implies more diffused protons and higher values of σ_{eff}. No strong dependence on the colour reconnection is observed for the σ_{eff} value but it is anyway important for tuning purposes since it defines the interplay between the two hard scatterings. For the interested reader, the whole description of the machinery used for the σ_{eff} determination is given in Appendix E. Note that tuning the impact parameter dependence of the matter distribution function from DPS-sensitive data translates on constraints on the gluonic content of the proton. The impact parameter dependence of the gluonic distribution function has been also extracted by using the measurements of exclusive vector meson production and J/ψ photoproduction at HERA [230].

The new method has been applied to the W+dijet channel measured at the CMS experiment: a measurement of σ_{eff}, extracted from the template method, is available for comparison. By using the same data, the compatibility between the values obtained with the two different methods is checked. Results are presented in Section 11.2. The new method is then applied to the two analyses presented in this thesis and values of σ_{eff} for the two jet channels at 7 TeV are measured. The 4j and 2b2j analyses are the first ones which have used the new extraction method before any other σ_{eff} determination.

11.2 Extraction of σ_{eff} from W+dijet measurements

The tuning method is used to obtain σ_{eff} from the W+dijet scenario. In this channel an independent measurement of σ_{eff} [8], obtained with the template method, is available for cross check, based on two correlation observables, ΔS and $\Delta^{\mathrm{rel}}p_{\mathrm{T}}$. The variable ΔS is the azimuthal angle between the W and the dijet planes, while $\Delta^{\mathrm{rel}}p_{\mathrm{T}}$ represents the normalized p_{T} balance between the two selected jets. The tuning method has used exactly the same data points as the template method to extract the value of σ_{eff}. The MADGRAPH event generator, interfaced with the UE simulation provided by PYTHIA8, has been considered for the tune: it implements a ME with up to six partons in the final state. Indeed, by looking at the results in [8], this is the most suitable generator for describing the data: it has been shown that PYTHIA8 standalone is not able to describe successfully the measurements, due to missing higher orders in the ME. Using MADGRAPH, samples have been generated with different UE parameters in PYTHIA8. Two different tunes have been performed, which use the CTEQ6L1 PDF set and a negative exponential overlap distribution function:

- "CDPSTP8S1-Wj": PYTHIA8 "partial" tune where only the exponent of the negative exponential overlap matter distribution function has been varied; the other UE parameters have been kept equal to the ones of the reference tune 4C;

- "CDPSTP8S2-Wj": PYTHIA8 "full" tune where all the UE parameters have been varied.

The parameter ranges are the same as the ones used for the UE tunes and described in Appendix D. The envelopes of the predictions, important for checking the pertinence of the chosen parameter ranges (see Appendix D), are shown in Figure 11.1.

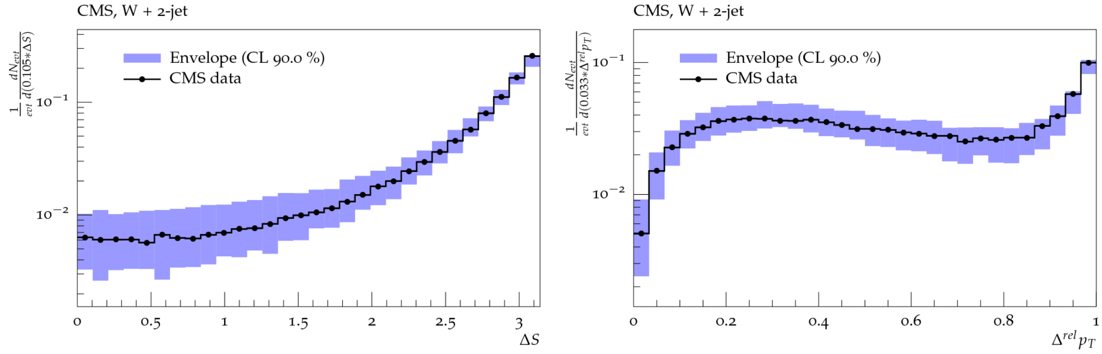


Figure 11.1: Envelopes of the MC samples are shown on the normalized distributions of the correlation observables ΔS (left) and $\Delta^{rel}p_T$ (right) for the W+dijet channel, together with the measured CMS data.

After the extraction of the new tunes, predictions are compared to the data. Figure 11.2 provides a picture of the new results: normalized cross sections as a function of the correlation observables obtained with tune 4C without MPI, tune 4C and the new tune CDPSTP8S1-Wj[1] are considered. From the comparisons with the data, it turns out that MPI are necessary to obtain a good description of the data and that the new tunes are able to reproduce the measurements. The description provided by the tune 4C is slightly worse than the new tunes in terms of reduced χ^2 (see Equation D.1). The small fluctuations which appear in some bins are due to limited statistics of the MADGRAPH samples.

In order to quantify the DPS contribution which is predicted in the new tunes, values of σ_{eff} are extracted through PYTHIA8, as explained in Appendix E, and are shown in Table 11.1. The new tunes predict σ_{eff} values around 25–26 mb, which are compatible with the value implemented in the old tune 4C and also with the published result of 20.7 mb. Note the quite large uncertainties affecting the σ_{eff} values measured in the new tunes, due mostly to the relatively large experimental uncertainties and the large χ^2 variation (equal to the absolute χ^2 of the best tune) set for the determination of the eigentunes. The uncertainty, quoted for σ_{eff} of the new tunes, includes the statistical and systematic uncertainties measured for the normalized cross sections. The agreement between σ_{eff} measurements, obtained with the two different methods, gives confidence in the values obtained for σ_{eff} with the tuning method.

---

[1] The performance for the other new tune CDPSTP8S2-Wj is very similar to CDPSTP8S1-Wj.

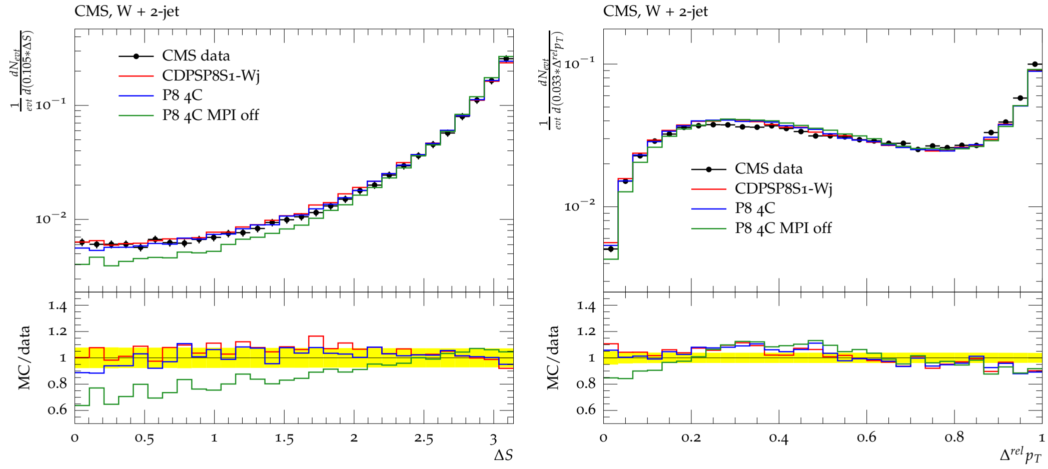


Figure 11.2: Comparisons of the correlation observables ΔS (left) and $\Delta^{rel}p_T$ (right) measured in the W+dijet channel compared with Madgraph *(MG) interfaced with* Pythia*8 Tune 4C, Tune 4C with no MPI, and the new* Pythia*8 partial tune (overlap only). Also shows the ratio of the tunes with the data. Results are published in [12].*

Pythia Tune	Pythia σ_{eff} value (mb)
Tune 4C	30.3
CDPSTP8S1-Wj	$25.9^{+2.4}_{-2.9}$
CDPSTP8S2-Wj	$25.8^{+8.2}_{-4.2}$
CMS published result [8]	20.7 ± 0.8 (stat.) ± 6.6 (syst.)

Table 11.1: Values of σ_{eff} obtained for each Madgraph+Pythia*8 tune. The values of the old tune 4C, and the new DPS-based tunes, CDPSTP8S1-Wj and CDPSTP8S2-Wj are compared. The uncertainties are obtained from the MC eigentunes. The value obtained with the template method and published by the CMS collaboration is listed in the last row.*

11.3 4j selection

The tuning method can be used for various other physics channels. The first one which is presented in this work is the 4j channel. The value of σ_{eff} has been obtained by fitting the predictions from Pythia8 and Pythia6 to the measurements of the correlation observables. With Pythia8, σ_{eff} has been determined by using different overlap matter distribution functions, such as negative exponential, double-gaussian and x-dependent single gaussian[1] (see Section 11.1). Only the double gaussian function has been set in the fit performed with Pythia6, since the others are not implemented in the generator.

The choice of these generators has been motivated by the fact that their predictions describe quite well the measured normalized distributions of the correlation observables (see Figure 10.9), already without any specific tuning; this might be an indication that the physics needed for describing the 4j scenario is already implemented in these generators, even though they produce a 2→2 LO ME. The determination of σ_{eff} with higher-order ME generators, like Madgraph and Powheg, might also be meaningful. Attention needs to be paid to the possibility that the contribution of missing higher orders in Pythia might be hidden by an overestimation of the UE components. This might indeed constitute a problem and might lead to biased results for σ_{eff}. Nevertheless, tunes with these generators could not be directly extracted because of technical reasons[2]. However, the tunes obtained with the LO ME generators have been propagated to Madgraph and Powheg ME to check their performance.

The following parameter settings have been investigated:

- "CDPSTP8S1-4j": Pythia8 tune where only the exponent of the negative exponential overlap matter distribution function has been varied; the other UE parameters have been kept fixed to the ones of the reference tune 4C;
- "CDPSTP8S2-4j": Pythia8 tune where all the UE parameters have been varied with a negative exponential overlap matter distribution function;
- "CDPSTP8S3-4j": Pythia8 tune performed with a double gaussian overlap matter distribution function;
- "CDPSTP8S4-4j": Pythia8 tune performed with a x-dependent single gaussian overlap matter distribution function, expressed by Equation 11.5;
- "CDPSTP6S1-4j": Pythia6 tune performed with a double gaussian overlap matter distribution function.

These different settings allow to study the dependence of the values of σ_{eff} on the different choices of overlap distribution functions and generators. Tunes extracted with various overlap matter distribution functions may differ in the amount of the soft MPI component, for instance

---

[1] The value of σ_{eff} for the x-dependent matter distribution function is calculated as the σ_{eff} average for different x values of the partons which take part in the first hard scattering. The machinery for additional hard scatterings is implemented later in the Pythia8 simulation phase.

[2] The binned generation available for Madgraph events is not suitable for tuning purposes, while the interface between Powheg and Pythia8 is not validated for inclusive jet events.

if different p_{T}^{0} values are obtained. This may alter migration effects for both SPS and DPS contributions and might translate into different values of σ_{eff}. In the previous ATLAS and CMS measurements [8, 9], no corresponding studies have been performed for different overlap distribution functions. The fit of the 4j channel has been based on two of the three variables measured on this channel: ΔS and $\Delta^{\mathrm{rel}}_{\mathrm{soft}} p_{\mathrm{T}}$. This choice is motivated by the preliminary studies, described in Chapter 4, which show a higher sensitivity to DPS contributions for these variables than for ϕ-based observables. The observables ΔS and $\Delta^{\mathrm{rel}} p_{\mathrm{T}}$ are the same variables which have been used for the tunes of the W+dijet channel: this gives a consistent picture of the two channels. The third measured observable, $\Delta\phi^{\mathrm{soft}}$, has been used to cross check if the predictions of the new tunes are still compatible with the data points. For each of the tunes, a value of σ_{eff} has been extracted.

The envelopes obtained with the simulated MC predictions are shown in Figure 11.3; the plots refer to the CDPSTP8S1-4j tune. The data points are well covered in each bin by the blue bands of the MC envelopes; a similar level of description is obtained for the other tunes.

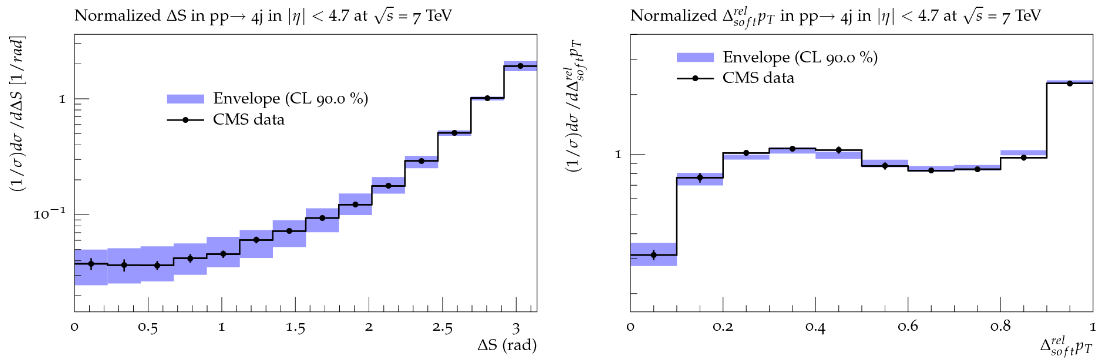


Figure 11.3: Envelopes of the MC samples are shown for the correlation observables ΔS (left) and $\Delta^{rel} p_T$ (right) for the 4j channel, together with the measured CMS data.

Predictions obtained with CDPSTP8S1-4j are shown in Figure 11.4 for ΔS and $\Delta^{\mathrm{rel}}_{\mathrm{soft}} p_{\mathrm{T}}$, together with the ones obtained with tune 4C and with tune 4C and MPI off. The predictions obtained without MPI are far away from the data. When the MPI simulation is switched on but kept at a relatively low amount, as implemented in tune 4C, the predictions get closer to the data but the agreement is not yet optimal. With a higher MPI contribution as obtained from the new tunes, the data are very well described. The variation between the predictions for the tunes with and without simulated MPI, is larger for ΔS than for $\Delta^{\mathrm{rel}}_{\mathrm{soft}} p_{\mathrm{T}}$. However, in order to describe 4j correlation observables, a higher DPS contribution is needed, than the one implemented in standard UE-based tunes. A similar level of agreement in the description of the correlation observables is achieved by predictions obtained with the other DPS-based tunes. Further comparisons are shown in Appendix E.

In order to quantify the DPS contribution, predicted in the tunes, the value of σ_{eff} has been extracted for each of them. Results are shown for different parameter settings in Table 11.2, along with the overlap matter model implemented in them. The parameter uncertainties coming from the eigentunes have also been taking into account. While predictions using UE-based tunes, like Pythia8 4C and Pythia6 Z2*, tend towards higher values of σ_{eff}, around 30-33 mb, the new tunes favour lower σ_{eff}, covering the range between 16.5–23.1. Lower σ_{eff} values translate into a larger

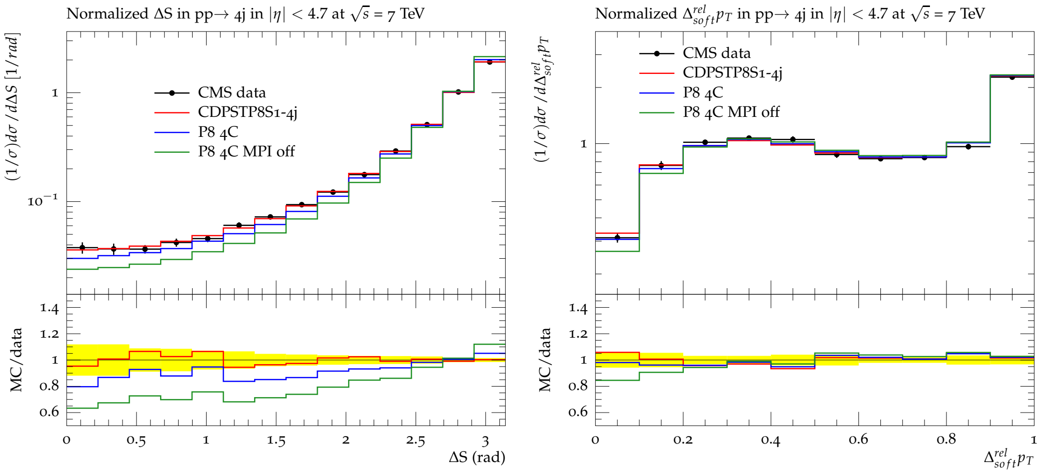


Figure 11.4: Comparisons of predictions of the new tune for ΔS (left column) and $\Delta^{rel}_{soft} p_T$ (right column), measured in the 4j channel; predictions obtained with the tune 4C without MPI simulated and with the tunes 4C and CDPSTP8S1-4j are compared to the data. The lower panels show the ratios between MC predictions and data. Results are published in [12].

DPS contribution, according to Equation 2.9. In order to achieve a better description of the 4j correlation observables, the DPS contribution needs to be increased. This conclusion holds for each tested overlap matter distribution, even though to different extent.

Pythia Tune	Matter distribution	Pythia σ_{eff} value (mb)
Tune 4C	Negative exponential	30.3
Tune Z2*	Double Gaussian	33.0
CDPSTP8S1-4j	Negative exponential	$21.3^{+1.2}_{-1.6}$
CDPSTP8S2-4j	Negative exponential	$19.0^{+4.7}_{-3.0}$
CDPSTP8S3-4j	Double Gaussian	$23.1^{+3.9}_{-4.0}$
CDPSTP8S4-4j	x-dependent single gaussian	$16.3^{+6.1}_{-3.7}$
CDPSTP6S1-4j	Double Gaussian	$16.5^{+4.0}_{-3.2}$

Table 11.2: Values of σ_{eff} obtained for each Pythia*8 and* Pythia*6 tune. The values of the old tunes 4C and Z2*, and the new DPS-based tunes, CDPSTP8S1-4j, CDPSTP8S2-4j, CDPSTP8S3-4j, CDPSTP8S4-4j and CDPSTP6S1-4j are compared. The uncertainties are derived from the* PROFESSOR *eigentunes and express the value of σ_{eff} corresponding to a $\Delta\chi^2$ of the fit equal to the chi^2 obtained for the best tune.*

11.4 2b2j selection

The tuning method has been also applied for the determination of σ_{eff} in the 2b2j channel, by considering the correlation observables presented in Chapter 10. Only the Pythia8 generator has been considered for the tune at this stage. The considered parameters are the same as used for the 4j scenario, as well as the tuning range. Nevertheless, for the 2b2j channel, only the normalized differential cross section as a function of ΔS has been considered for the extraction of σ_{eff}, while the $\Delta^{\text{rel}}_{\text{soft}} p_{\text{T}}$ variable has been excluded because of some difficulties observed for the Pythia8 predictions to describe its shape (see Appendix E). The normalized differential cross section as a function of $\Delta^{\text{rel}}_{\text{light}} p_{\text{T}}$, as well as of the other correlation observables, has been used to cross check the performance of the obtained tunes. Four different tunes have been extracted for the 2b2j channel; they use the same settings as done in the 4j scenario and are listed in the following:

- “CDPSTP8S1-2b2j”;
- “CDPSTP8S2-2b2j”;
- “CDPSTP8S3-2b2j”;
- “CDPSTP8S4-2b2j”.

Comparisons of predictions obtained from old and new tunes are shown for ΔS in Figure 11.5. In Figure 11.5, predictions are shown for tune 4C, CDPSTP8S1-2b2j and tune 4C with MPI switched off: while the shape obtained with the tune 4C without the simulation of the MPI does not reproduce the data, tune 4C is able to follow better the trend of the measurements. The best description is, however, provided by the predictions of the new tune CDPSTP8S1-2b2j, which are compatible with the data within uncertainties in each measured bin. The performance of the other tunes in describing the shape of ΔS is very similar to the one achieved by CDPSTP8S1-2b2j. Comparisons between the measurements of all correlation observables and the predictions of the new tunes are shown in Appendix E.

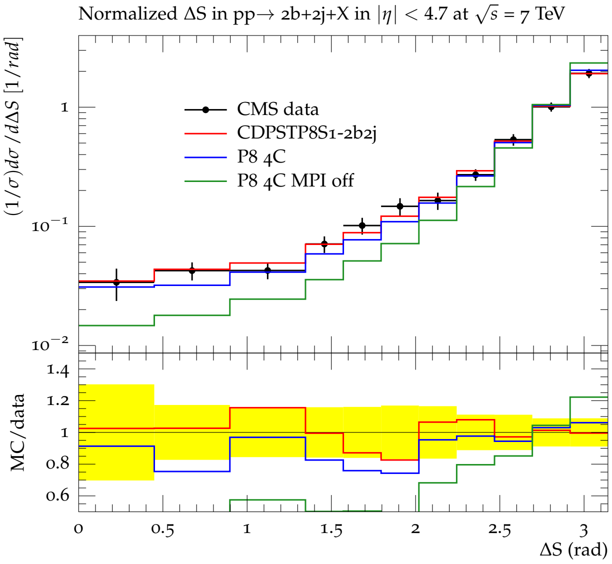


Figure 11.5: Comparisons of predictions obtained with the new DPS-based tune with the correlation observable ΔS, measured in the 2b2j channel; in the left plot, predictions obtained with the tune 4C without MPI simulated and with the tunes 4C and CDPSTP8S1-2b2j are compared to the data. The lower panel shows the ratios between MC predictions and data.

The final step is to extract the σ_{eff} values predicted by the new tunes. Results are reported in Table 11.3: the values of σ_{eff} obtained by DPS-based fits are lower compared to the one implemented in the old tune 4C. The values of σ_{eff} range between 23 and 26 mb for the negative exponential and double gaussian overlap matter distribution functions. When a x-dependent overlap function is used, σ_{eff} decreases down to 14.2 mb; this indicates that lower values of σ_{eff} are preferred when correlations between the impact parameter and the parton momentum fraction are included in the overlap function. However, in all new tunes, a contribution of DPS higher than the one predicted by the UE-based tunes needs to be implemented, in order to get a better agreement with the measurements.

PYTHIA8 Tune	Matter distribution	PYTHIA8 σ_{eff} value (mb)
Tune 4C	Negative exponential	30.3
CDPSTP8S1-2b2j	Negative exponential	$25.2^{+4.1}_{-2.9}$
CDPSTP8S2-2b2j	Negative exponential	$23.2^{+3.3}_{-2.5}$
CDPSTP8S3-2b2j	Double Gaussian	$26.1^{+1.9}_{-4.4}$
CDPSTP8S4-2b2j	x-dependent single gaussian	$14.2^{+1.8}_{-1.3}$

Table 11.3: Values of σ_{eff} obtained for each PYTHIA*8 tune. The values of the old Tune 4C, and the new DPS-based tunes, CDPSTP8S1-2b2j, CDPSTP8S2-2b2j, CDPSTP8S3-2b2j and CDPSTP8S4-2b2j are compared. The uncertainties are derived from the* PROFESSOR *eigentunes and express the value of σ_{eff} corresponding to a $\Delta\chi^2$ of the fit equal to the chi^2 obtained for the best tune.*

In summary, the tuning method has produced the first determination of σ_{eff} in a 4j and in a 2b2j scenario for *pp* collisions at 7 TeV. For more details, in Appendix E an extensive set of comparisons is given for all the fits, extracted with the different settings.

11.5 Interpretation of the results

The determination of σ_{eff} from fits to DPS-sensitive observables becomes important to understand and describe softer and harder MPI within the same framework. The final goal is, indeed, to have a tune which is able to reproduce UE and inclusive data, like hadronic activities and charged particle multiplicities, and at the same time to predict the correct σ_{eff} value and have a good description of DPS signals. Studies towards this direction are presented in the following: the DPS-based CDPSTP8S2-4j and the UE-based CUETP8S1-CTEQ6L1 and CUETP8S1-HERAPDF1.5LO tunes, described in Appendix D, are compared to a set of observables, different from the ones used for the corresponding fits. Predictions obtained with the CDPSTP8S2-4j tune are compared to UE observables at 7 TeV measured by ATLAS [90], namely charged particle multiplicities and p_{T} sum in transverse, toward and away regions, as a function of the p_{T} of the leading charged particle in each event. Figure 11.6 shows the results: the solid line shows the central value, while the band corresponds to the envelope of the parameter uncertainties. While the plateaux regions of all UE spectra are slightly underestimated by the predictions of the central values but are compatible within the tune uncertainties, larger discrepancies appear in the very first part of the distributions, for p_{T} between 1 and 4 GeV. The rising part of the spectrum is underestimated by 20–40% for all three regions. In particular, the disagreement is larger for observables measured in the transverse region, which are expected to be more sensitive to soft MPI. It appears that DPS-based tunes are not able to reproduce optimally measurements which are mostly affected by MPI of a softer scale.

In order to investigate further this aspect, predictions of UE-based tunes are in turn compared to the normalized differential cross sections, measured in the 4j channel. Additionally, the compatibility of σ_{eff} values predicted by UE- and DPS-based tunes is examined. Figure 11.7 shows the comparisons between predictions of DPS- and UE-based tunes and 4j correlation observables. While $\Delta^{\text{rel}}_{\text{soft}}p_{\text{T}}$ is well described by all curves, the predictions of UE-based tunes fail in the de-

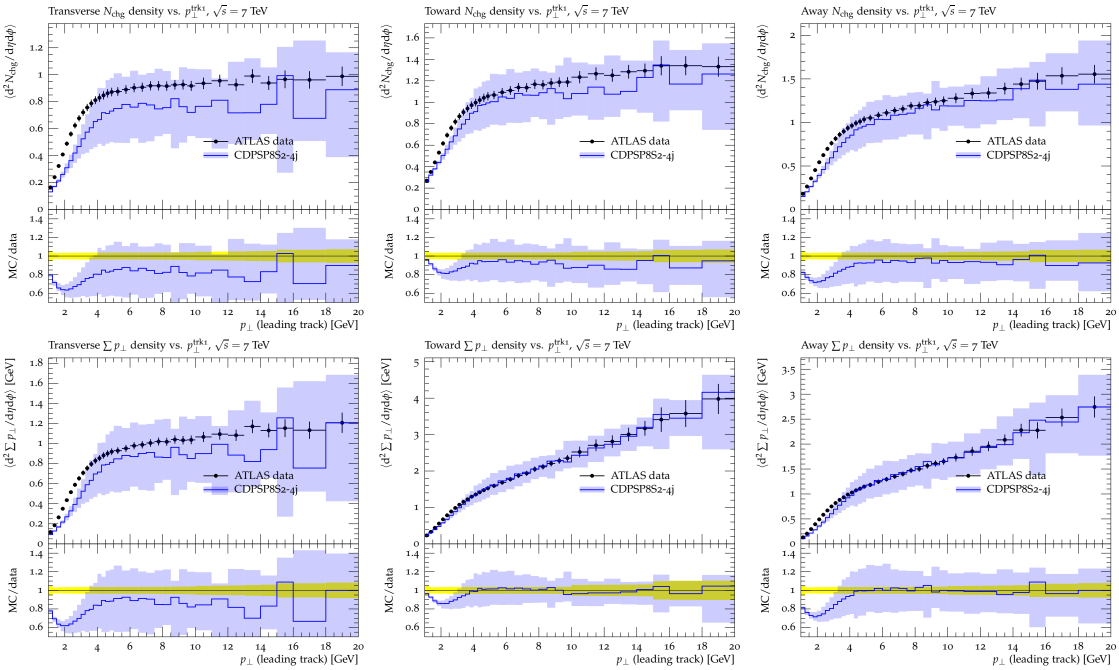


Figure 11.6: ATLAS data on the charged particle multiplicity N_{ch} (top) and p_T sum (bottom) measured in the transverse (left), toward (center) and away (right) regions compared with CDPSTS2-4j. Also shows the ratio of the tunes with the data and the uncertainties of the predictions based on the PROFESSOR *eigentune. Results are published in [12].*

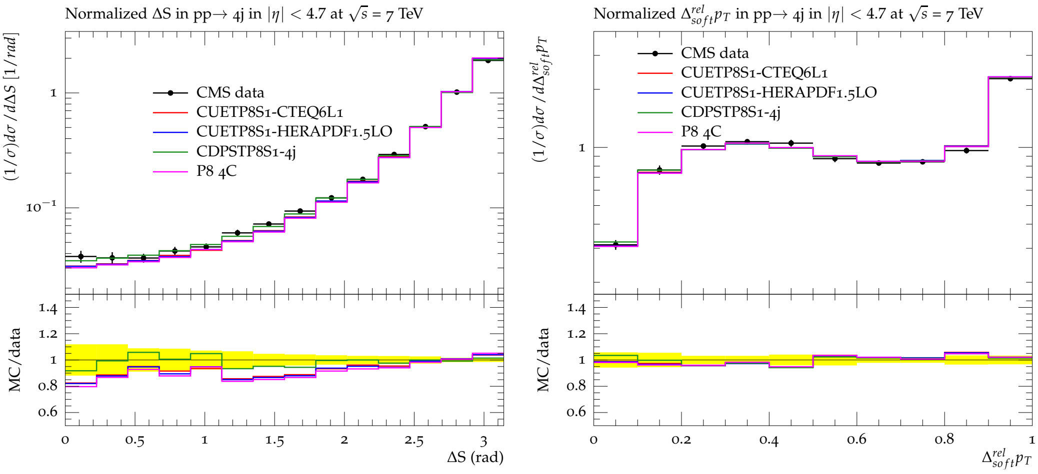


Figure 11.7: Comparisons of predictions obtained with the new UE-based tunes with the correlation observables ΔS (left) and $\Delta^{rel}_{soft}p_T$ (right), measured in the 4j channel. Predictions obtained with CUETP8S1-CTEQ6L1 and CUETP8S1-HERAPDF1.5LO, together with the tune 4C and the CDPSTP8S2-4j, are compared to the data. The lower panels show the ratios between MC predictions and data.

scription of ΔS, where the region at low values is underestimated by 10–20%. Table 11.4 shows the $\sigma_{\rm eff}$ values obtained from the different tunes: a tension between UE- and DPS-based tunes appear. While, as seen in Section 11.3, CDPSTP8S1-4j and CDPSTP8S2-4j tunes give a $\sigma_{\rm eff}$ value between 19 and 21 mb, CUETP8S1-CTEQ6L1, CUETP8S1-HERAPDF1.5LO, the old 4C and the new Monash tune [231] seem to be systematically larger, leading to values between 27 and 30 mb. Measurements of UE- and DPS-based tunes are not compatible within uncertainties.

The overview given by this study shows an incompatibility in the description of the whole spectrum of MPI, from the softest to the hardest ones. At this stage, predictions which describe very well UE data and their energy dependence are not able to reproduce at the same level DPS-sensitive data, while predictions extracted from hard MPI signals, fail to describe UE data.

PYTHIA8 Tune	PYTHIA8 $\sigma_{\rm eff}$ value (mb)
Tune 4C	30.3
CUETP8S1-CTEQ6L1	$27.8^{+1.2}_{-1.3}$
CUETP8S1-HERAPDF1.5LO	$29.1^{+2.3}_{-2.0}$
CDPSTP8S1-4j	$21.3^{+1.2}_{-1.6}$
CDPSTP8S2-4j	$19.0^{+4.7}_{-3.0}$
Monash Tune	27.4

Table 11.4: Values of $\sigma_{\rm eff}$ obtained for each PYTHIA8 tune. The values of the old Tune 4C, the new UE tunes, CUETP8S1-CTEQ6L1 and CUETP8S1-HERAPDF1.5LO, and the new DPS tunes, CUETP8S1-4j and CDPSTP8S2-4j, are compared. The uncertainties are derived from the PROFESSOR eigentunes and express the value of $\sigma_{\rm eff}$ corresponding to a $\Delta\chi^2$ of the fit equal to the χ^2 obtained for the best tune. The value of $\sigma_{\rm eff}$ predicted by the Monash tune is also listed.

11.6 Summary of $\sigma_{\rm eff}$ measurements

With the new $\sigma_{\rm eff}$ measurements in the 4j and 2b2j scenarios, the DPS picture becomes richer and more interesting. By looking at measurements performed at different energies and physics channels, one might speculate about possible dependencies or correlations. Table 11.5 shows the values of $\sigma_{\rm eff}$ published by the different collaborations, over almost thirty years.

Figure 11.8 shows the most significant $\sigma_{\rm eff}$ values as a function of the center-of-mass energy, at which they have been extracted, covering almost two orders of magnitude. Note that for some of the measurements, like at 63 GeV and 630 GeV, the uncertainties are not provided. From the AFS measurement in the 4j channel at an energy of 63 GeV, in 1987, until the most recent findings by CMS and D0, many points have been added to the list and comparisons among the different measurements can be performed. The values of $\sigma_{\rm eff}$ systematically increase from 10–15 mb at 1–2 TeV towards slightly higher numbers at 7 TeV, around 15–23 mb. However, it is indeed too early to state something conclusive about the energy dependence, due to the still large uncertainties affecting all the measurements. This is also valid for any observation related to channel dependence: no clear behaviour may be conjectured for $\sigma_{\rm eff}$ values measured in different final states. The trend observed from DPS measurements in jet and boson events is, however, very intriguing. If one considers the measurements of $\sigma_{\rm eff}$ obtained with the tuning method in the

W+dijet and 4j channels, a value of 25.8 mb is obtained for the first one, while in the second one a value of 19.0 mb is obtained. This seems to be in agreement with UE data measured in the same types of events. The charged particle multiplicities, as a function of the leading jet p_T in hadronic events and as a function of the dimuon pair p_T in Drell-Yan events, have been measured by CMS [81, 232] in the transverse region and are shown in Figure 11.9. It may be observed that at the same p_T of the hard object, e.g. 10 GeV, the hadronic activity is systematically higher in jet events than in Drell-Yan events. This goes towards the same direction of the σ_{eff} values extracted in the two channels: the lower σ_{eff} of the 4j channel implies, in fact, a higher activity than the one measured in W+dijet channel, which exhibits a larger value of σ_{eff}. If one assumes that the MPI contribution is independent on the scale of the hard scattering, this behaviour might be qualitatively interpreted as a different transverse dimension between quarks and gluons in the protons: in particular, gluons might have smaller transverse width.

Experiment	Collision Energy	Year	Channel	σ_{eff} value (mb)
AFS [5]	63 GeV	1987	four jets	5.0
UA2 [233]	630 GeV	1991	four jets	> 8.3
D0 [113]	1.96 TeV	2014	γ+jets	$12.7 \pm 0.2 \pm 1.3$
D0 [113]	1.96 TeV	2014	γ+b/c jets	$14.6 \pm 0.6 \pm 3.2$
CDF [112]	1.96 TeV	1997	γ+jets	$14.5 \pm 1.7 \pm 2.0$
CDF [6]	1.96 TeV	1993	four jets	$12.1\ ^{+10.7}_{-5.4}$
ATLAS [9]	7 TeV	2012	W-jet	$15 \pm 3\ ^{+3}_{-5}$
CMS [8]	7 TeV	2013	W-jet	$20.7 \pm 0.8 \pm 6.5$
CMS[1]	7 TeV	2014	W-jet	$25.8^{+8.2}_{-4.2}$
CMS[2]	7 TeV	2014	four jets	$19.0^{+4.6}_{-3.0}$
CMS[3]	7 TeV	2014	2b2j	$23.2^{+3.3}_{-2.5}$
ATLAS [234]	7 TeV	2013	four jets	$16.0^{+0.5}_{-0.8}\ ^{+1.9}_{-3.5}$

Table 11.5: Summary of σ_{eff} values measured at different collision energies and in various physics channels. For the values measured in the 4j and 2b2j channels, only the values obtained with CDPSTP8S2 tunes are listed.

In this picture, gluonic initial states would tend to produce more central collisions than quark initial states. Since jet events are produced mostly by gluonic initial states, while Drell-Yan and W-boson events are mainly driven by quark initial states, the larger hadronic activity, measured for the former type of events in both UE and DPS data, might be thus explained. A further investigation may consider measurements of DPS and UE observables from interactions measured at the same energy scale, in order to have a more direct comparison of hard and soft MPI contributions.

It is, therefore, very interesting to continue on the path already entered by these measurements of DPS-sensitive observables. On one hand, the new LHC phase might give the possibility to add new measurements to the picture at a even higher collision energy; on the other hand, the experimental challenge to lower the systematic uncertainties needs further studies and improvements. The two new values of σ_{eff} measured in four-jet scenarios will both play a key role to understand the energy and the channel dependence, as well as to investigate the possibility of introducing parton correlations in the DPS modelling.

---

[1] This thesis.

[2] This thesis.

[3] This thesis.

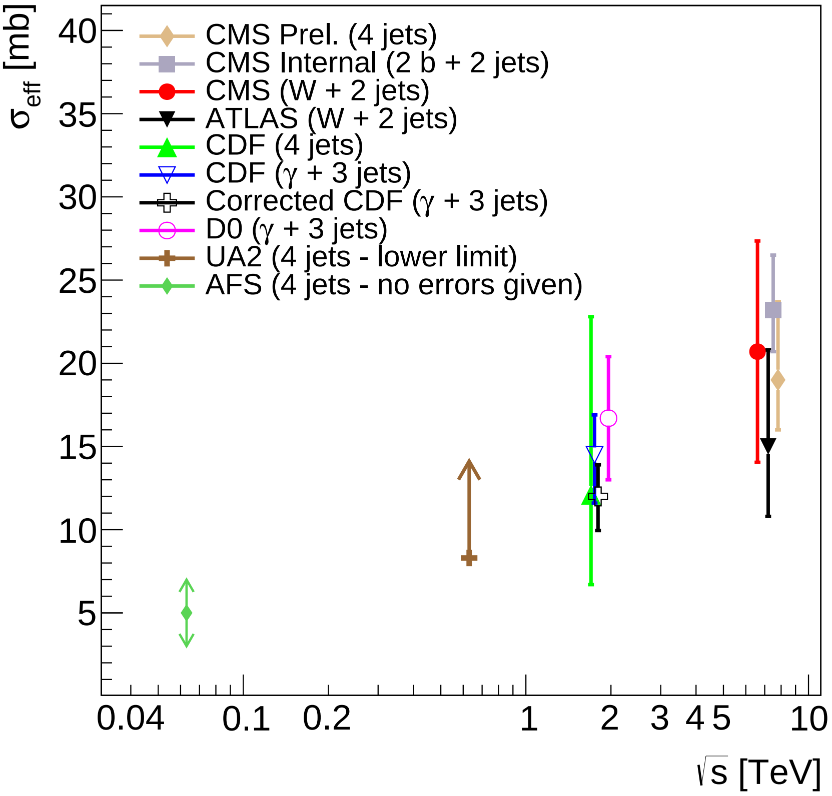


Figure 11.8: Center-of-mass energy dependence of σ_{eff} measured by different experiments using various processes. These measurements used different approaches for the extraction of the DPS contribution.

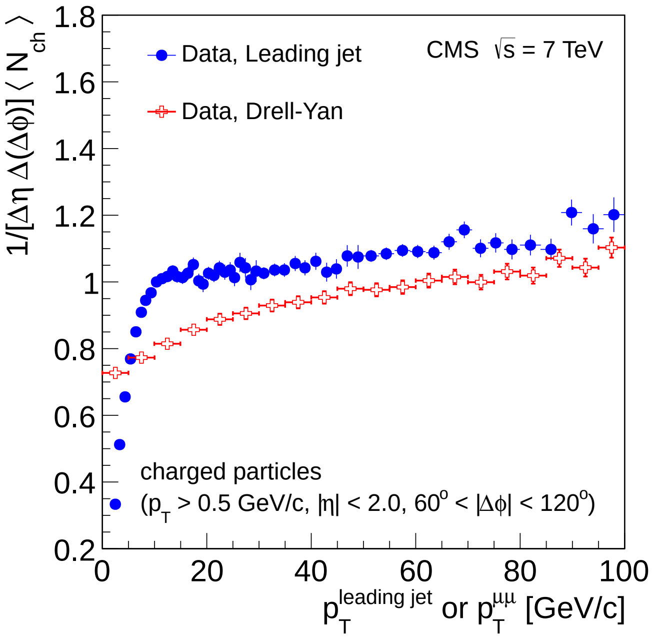


Figure 11.9: UE activity in Drell-Yan and hadronic events measured by CMS [81] in the transverse region: the charged particle multiplicity is shown as a function of the p_T of the muon pair in Drell-Yan events and of the leading jet p_T in hadronic events.

Chapter 12

Summary and conclusions

In this thesis, measurements of multijet channels in proton-proton collisions at 7 TeV, performed with the Compact Muon Solenoid at the Large Hadron Collider are presented. These measurements are relevant for improving the understanding of the theory of Quantum Chromodynamics, in the light and heavy flavour sector, and for investigating possible signals of Double Parton Scattering (DPS). Two scenarios have been considered; the first one requires exactly four jets, selected in pairs at different scales. In the second channel, differently flavoured jets are considered: two jets originated by a b-quark are selected in the central region, together with two other additional jets in the entire detector acceptance. Jets are associated in pairs: in the former analysis, two pairs are built from the jets selected at the same scale, while in the latter one, jets are combined according to their flavour. For both channels, absolute and normalized differential cross sections have been measured as a function of transverse momentum and pseudorapidity of the single selected jets, as well as of the correlation observables, expressing the balance of the jet pairs in the transverse plane. It is observed that the two jet pairs are preferably correlated but a significant fraction of events exhibits an uncorrelated topology between the jet systems, corresponding to a back-to-back configuration in the transverse plane for the jets of the same pair. These events are particularly interesting for a possible signal produced by DPS, to be separated from events of Single Parton Scattering (SPS).

Measurements have been compared to various theoretical predictions; several Monte Carlo event generators, producing matrix elements at different order and with different number of partons in the final states, have been considered, interfaced with a simulation of the parton shower and the multiple parton interactions (MPI). After validating suitable tunes for the simulation of the underlying event in generators like Powheg, Madgraph and Sherpa, important conclusions have been extracted from the comparisons to the multijet measurements; fixed-order matrix element calculations, which include parton shower and underlying event (UE) simulation, are able to give an overall good description of the differential cross sections. Nevertheless, the description gets worse in some regions of the phase space, driven by a significant dependence on the simulation of the MPI.

Furthermore, the addition of contributions from DPS in the considered models helps to bring the predictions closer to the measured correlation observables. A measurement of the DPS contribution, quantified by the parameter σ_{eff}, included in the Pythia MPI model has been performed, based on the measured correlation observables. This measurement relies on a new method, based on tuning parameters which are sensitive to DPS, within the framework of event generators, in order to get the best description of the data. First of all, this method has been validated with previous DPS-sensitive measurements and has shown solid and consistent results. Secondly, it has been applied to the multijet measurements: it has led to $\sigma_{\rm eff}$ values around 20–25 mb for both multijet analyses. The obtained $\sigma_{\rm eff}$ values are compatible with previous measurements of the same quantity, performed in different physics channels, at different collision energies and with different methods. The experimental uncertainties are rather large, mainly coming from jet energy scale.

Besides the investigation of DPS and hard MPI components, new PYTHIA8 tunes, extracted from UE data, have given new light to the softer part of the MPI, as well: they are able to give a better description of the energy dependence of UE observables. Considering that the development of PYTHIA6 will be stopped to leave room to PYTHIA8, they will constitute one of the most reliable and appropriate tools for predictions at a higher collision energy, as foreseen for the next LHC phase in 2015.

12.1 Outlook

Measurements of multijet scenarios, described in this thesis, constitute an important baseline for future studies and investigations, along with the new $\sigma_{\rm eff}$ extraction method introduced with these analyses. The theoretical predictions already give a clear and unambiguous picture of the physics behind it. Comparisons with predictions with a Next-To-Leading-Order 2→4 matrix element, interfaced with parton shower and UE simulation, are also important, in order to have a correct treatment of the additional jets present in the final state. Event generators with these features are available and they seem to be able to treat $p_{\rm T}$ thresholds properly and with stable results down to low values, similar to the ones used in these multijet analyses.

From this work, the experimental determination of $\sigma_{\rm eff}$ turns out to be rather enriched: the new extraction method gives a solid and reliable alternative to older methods. On one hand, the method being simple and intuitive may allow to determine $\sigma_{\rm eff}$ from the previous and already published measurements, without any restriction given by collision energy or measured observable. On the other hand, it opens up further applications for ongoing and future analyses. A large gain may be achieved with analyses at 13 TeV, foreseen for the future LHC phase. The determination of $\sigma_{\rm eff}$ with the new method is possible with any available model and generator. Particularly interesting is to investigate the dependence of the DPS contribution on the different assumptions made on parton correlations inside the proton. Results in case of inclusion of x-dependence in the matter distribution function have already been obtained within the PYTHIA8 model, and other more complicated correlations, including spin or colour effects, can be also examined, when event generators which incorporate them will be available.

The total uncertainty which affects the current $\sigma_{\rm eff}$ values is rather large, coming from the experimental measurements. Specific efforts are necessary to decrease the jet energy scale which drives the systematic uncertainty in jet measurements, or to find new physics channels with a higher sensitivity to DPS and a smaller background contamination. It might also be useful to identify new correlation observables with a greater discriminating power. Same-sign WW or ZZ final states might be the best channels to be used for a background-cleaned DPS measurement.

Furthermore, as shown by these analyses, higher DPS sensitivity is achieved by observables, like ΔS, which take into account information of the entire final state and not only of a subset of objects. These improvements would allow better determinations of energy and channel σ_{eff} dependence.

A parallel goal is also to obtain a good description of both soft and hard MPI components, by building tunes which are able to predict the correct value of σ_{eff} and describe UE observables. This work has shown that this is not yet achieved by the new UE- or DPS-based tunes. This might lead to two different research lines: on one hand, the tension in the description of hard- and soft-scale processes within the same framework may indicate some deficiency of the whole model and a completely different approach may be attempted. In particular, a very interesting perspective is the release of the currently used collinear factorization approximation in the formalism of a hadronic collision [235]: in this picture, MPI appear naturally inside the model, without the addition of *ad hoc* terms and components, and DPS is consequently interpreted as an additional contribution to a general hadronic scattering, not as a physical process by itself.

On the other hand, a refinement of the current model may be also considered. In particular, including a x-dependent overlap matter distribution function [236] seems to bring an improvement towards a better description of soft and hard MPI components and helps to unify hard- and soft-scale processes in a more consistent picture. A dependence on x in the proton matter distribution has been suggested from the observation of the rise of the total proton-proton cross section with the collision energy [237], and theoretical studies [238, 239] similarly suggest that wave functions of low-x partons are wider than the ones at high-x. However, further investigations need to be performed before a more definitive conclusion is drawn. A falsification of the current concept of MPI in favour of other approaches is admittedly rather difficult: in fact, current models with tuned MPI are able to describe a wide range of measurements. An important point of other approaches might be whether they are able to reproduce the same measurements in a more “natural” way with less tuning and have a consistent picture for the whole scenario of a hadronic collision. In the case of validity of the present MPI view, models need to be improved: they do not yet well reproduce measurements of long-range particle correlations and ridge effects [240] as well as hard diffraction data [241].

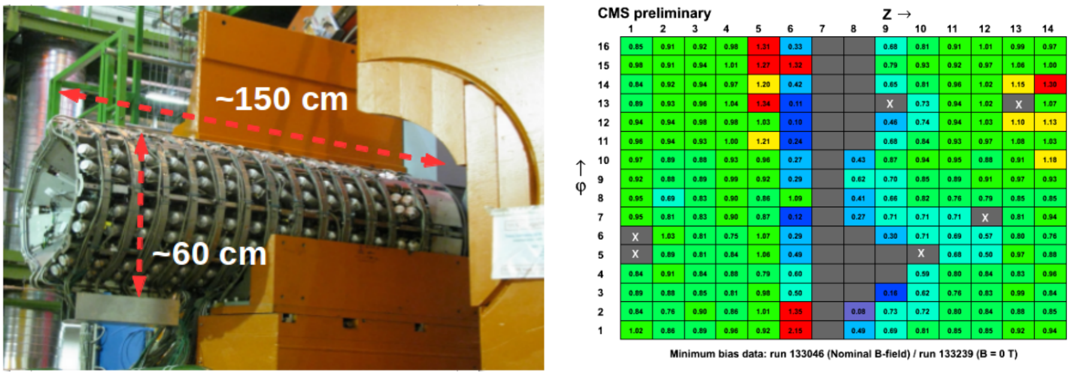
Finally, specifically related to the heavy flavour sector, the measurements performed with b-jets in the multijet final state may serve as a testing ground for predictions using unintegrated PDFs; these predictions, in fact, suffer from double counting problems in the light flavour sector, while requirements of heavy flavour jets remove these issues. It is indeed interesting to check the compatibility of these predictions with the data, which offer a great opportunity to study a non-standard approach of hadronic collisions.

Appendix A

The CASTOR calorimeter at the CMS experiment

The CASTOR calorimeter at the CMS experiment is an electromagnetic/hadronic calorimeter which covers the very forward region of the detector (-6.6 $< \eta <$ -5.2). It is indeed the most forward detector at the CMS experiment. Note that it covers only the negative side of pseudorapidities, mainly due to budget and space constraints in the detector cavern. The name CASTOR stands for Centauro And STrange Object Research [242], because it originally aims for the detection of CENTAURO events, observed in high-energy collisions in cosmic-ray physics and characterized by an anomalous hadron content. This would appear at the LHC, as a high flux of hadrons boosted towards very forward rapidities. The CASTOR subdetector is a sampling calorimeter, made of quartz plates embedded in tungsten absorbers. The calorimeter is segmented in 16 ϕ-sectors in the plane transverse to the beam direction and 14 z-modules along the direction of the beam, for a total of 224 cells. A tower in CASTOR is composed by all cells in a given sector. The first two modules have half the depth of the others and serve to detect electromagnetic showers. The full calorimeter has a depth of 10.5 interaction lengths. This structure, shown in Figure A.1(left), allows to detect cascades of particles crossing the detector by collecting Cherenkov light. This kind of light is released when particles pass through a certain material with a velocity greater than the speed of light in the material itself. Coherent light is then emitted inside a cone, whose aperture depends only on the refractive index of the material. In CASTOR, the traversing particles emit photons in the quartz layers, while they are decelerated by the more dense tungsten absorbers, which create a high number of additional secondary particles. These particles form what is called a hadronic or electromagnetic cascade, depending on which particles have entered the layer: a hadronic cascade is produced by incident hadrons and it consists of further particles, mainly pions, while electrons or photons generate the electromagnetic shower (consisting of further electrons and photons) via the pair production and bremstrahlung mechanisms. The particles of the shower are crucial because they produce further Cherenkov light in the quartz layers. The emitted light is then driven via internal reflection to the edges of the quartz layers where it is detected and converted into an electrical signal by using fine-mesh Hamamatsu R5505 photomultiplier tubes. They allow operation under up to 0.5 T magnetic field if the field direction is within ±45° with respect to the photomultiplier axis. The photons are converted to photoelectrons which constitute after amplification the electric signal at the CASTOR outputs. The performance of the CASTOR calorimeter has been studied in a test beam environment [243].

The main challenge in the operation of CASTOR is the very special location at about $z = -14.3$ m from the interaction point, close to the beam pipe and surrounded by massive shields. This requires a very compact form of the detector, with one of the consequences being that the 224 photomultiplier tubes (PMTs) are mounted directly on the detector, less than 30 cm away from the LHC beam. The PMTs are thus exposed to high-radiation levels and strong fringe magnetic fields. In particular, the complicated magnetic field configuration at the location of CASTOR is caused by the fact, that the massive shields that surround CASTOR meet in proximity of its center (around module 7), producing an air gap of 40 mm between them. The absolute value of the magnetic field flux measured at this region does not exceed 0.2 T but the direction of the field varies strongly. This results in totally suppressed responses of the PMTs located around the gap in the shielding (see Figure A.1(right)).



	1	2	3	4	5	6	7	8	9	10	11	12	13	14
16	0.85	0.91	0.92	0.98	1.31	0.33			0.68	0.81	0.91	1.01	0.99	0.97
15	0.98	0.91	0.94	1.01	1.27	1.32			0.79	0.93	0.92	0.97	1.06	1.00
14	0.84	0.92	0.94	0.97	1.20	0.42			0.65	0.81	0.96	1.02	1.15	1.30
13	0.89	0.93	0.96	1.04	1.34	0.11			X	0.73	0.94	1.02	X	1.07
12	0.94	0.94	0.98	0.98	1.03	0.10			0.46	0.74	0.94	1.03	1.10	1.13
11	0.96	0.94	0.93	1.00	1.21	0.24			0.68	0.84	0.93	0.97	1.08	1.03
10	0.97	0.89	0.88	0.93	0.96	0.27		0.43	0.87	0.94	0.95	0.88	0.91	1.18
9	0.92	0.88	0.89	0.99	0.92	0.29		0.62	0.70	0.85	0.89	0.91	0.97	0.93
8	0.95	0.69	0.83	0.90	0.86	1.09		0.41	0.66	0.82	0.76	0.79	0.85	0.85
7	0.95	0.81	0.83	0.90	0.87	0.12		0.27	0.71	0.71	0.71	X	0.81	0.94
6	X	1.03	0.81	0.75	1.07	0.29			0.30	0.71	0.69	0.57	0.80	0.76
5	X	0.89	0.81	0.84	1.06	0.49				X	0.68	0.50	0.97	0.88
4	0.84	0.91	0.84	0.88	0.79	0.60				0.59	0.80	0.84	0.83	0.96
3	0.89	0.88	0.85	0.81	0.98	0.50			0.16	0.62	0.76	0.83	0.99	0.84
2	0.84	0.76	0.90	0.86	1.01	1.35		0.08	0.73	0.72	0.80	0.84	0.88	0.85
1	1.02	0.86	0.89	0.96	0.92	2.15		0.49	0.69	0.81	0.85	0.85	0.92	0.94

Minimum bias data: run 133046 (Nominal B-field) / run 133239 (B = 0 T)

Figure A.1: (Left) Picture of CASTOR before the inclusion in the CMS cavern. Transverse and longitudinal dimensions are also indicated with red dashed lines. (Right) Map (ϕ vs z) of the ratio S_i (B = 3.8 T)/S_i (B = 0 T) of the average response of each channel i of CASTOR with and without magnetic field. The grey colour in the central region indicates a ratio close to 0, namely a high inefficiency in presence of magnetic field, while the crossed channels have been observed to be inactive regardless of the magnetic field.

Another consequence of the strong remnant fields in the forward region of the CMS detector is that the CASTOR calorimeter slightly shifts when the CMS solenoid is switched on. The largest shift is found to be approximately 12 mm. This results in some ϕ sectors moving to more central rapidities, covering the range between $-6.3 < \eta < -5.13$. Currently, a strong effort is being made to understand the shifts for every data taking period and to implement it in the software simulation. Despite these challenges, interesting and nice analyses have been performed with a very relevant contribution of the CASTOR detector: the measurement of the UE in the very forward region in pp collisions [180] and the measurement of the energy density in lead-lead collisions [244] are only two examples of those and they gave important insights on the understanding of MPI and heavy-ion collisions.

A.1 Looking at jets in CASTOR

The electric signals coming from the photomultipliers have to be interpreted in terms of energy released by the particles which crossed the detector and further of the total energy of those. The procedure of finding the relation between the detector output and a measured physical quantity is commonly known under the name of “calibration”. Actually, there are two types of calibration:

- Absolute calibration, which studies the detector response without any dependence on the η position of the measured object; it aims to convert the electric output signal into an energy deposition in GeV;
- Relative calibration, which studies the dependence of the detector response as a function of η of the measured object; since CASTOR has no η segmentation, the relative calibration only aims to equalize the electric signals of the different PMTs.

It is important to perform a calibration for each measured physics object; the most important ones for the CASTOR detector are electrons and jets. The CASTOR calibration has been already performed but, for both electrons and jets, it is only based on simulated events. The final aim is indeed to have a data-driven calibration using measured physical processes. Such calibration would open the possibility of using jets measured in CASTOR; specifically, it would allow the extension of the phase space to the very forward pseudorapidities for the four-jet measurements, described in this thesis. In the following, the work related to an attempt of absolute jet calibration is described. Work on a data-driven way for electron calibration is also ongoing.

The current absolute jet calibration is described in [245]. First of all, in order to remove differences in the performance of the photomultipliers, the response of individual CASTOR cells has been equalized using a sample of beam halo muon events. Then, an absolute calibration of 0.015 GeV/fC, with an uncertainty of ±30%, has been obtained from a MC-based extrapolation of the η dependence of the energy density per unit of pseudorapidity measured in the HF calorimeter to the CASTOR acceptance [95]. Even though this result was found to be consistent with test beam measurements, an attempt to perform a data-driven absolute calibration has been tried. In particular, it has relied on the “dijet balance method” and it is described in the following. It consists of requiring events with two balanced objects: one, the “probe”, reconstructed in CASTOR and the other one, the “tag”, selected in a well calibrated region of the CMS detector, i.e. in $|\eta| < 2.5$. Specifically, the aim of the absolute jet calibration is to find a functional form which relates the charge of photoelectrons of the sum of photomultiplier outputs, to the jet p_{T}. Jets in CASTOR are identified by energy released in both the electromagnetic and hadronic regions. Jets in the acceptance of CASTOR are reconstructed by using an anti-k_{T} jet clustering algorithm with a radius parameter R = 0.7. A conventional jet algorithm usually operates in 3 dimensions, the pseudorapidity η, the azimuthal angle ϕ and the transverse energy, E_{T}. Since CASTOR has no segmentation in η, the polar angle of the seeds for the jet algorithm is kept fixed and the jet algorithm is operated in two dimensions. Jets are reconstructed for fixed η = -5.9, corresponding to the central value of the pseudorapidity range covered by the calorimeter. Jet identification (JetID) properties are available to characterize jets in CASTOR and they are:

- E_{em}: energy deposit in the electromagnetic section;
- E_{had}: energy deposit in the hadronic section;

- f_{em}=$E_{\mathrm{em}}/(E_{\mathrm{em}}+E_{\mathrm{had}})$: ratio between the energy in the electromagnetic section and the total energy;
- width of the jet, estimated as $\sigma^2 = \sum_{\mathrm{tower}}(\phi_{\mathrm{tower}}\text{-}\phi_{\mathrm{tot}})^2 E_{\mathrm{tower}}/E_{\mathrm{tot}}$;
- depth of the jet, estimated as $\langle z\rangle = \sum_{\mathrm{cell}} z_{\mathrm{cell}}E_{\mathrm{cell}}/E_{\mathrm{tot}}$;
- f_{hot}=$E_{\mathrm{max}}/(E_{\mathrm{em}}+E_{\mathrm{had}})$: ratio between the energy in the hottest cell and the total energy;
- variance of the depth, $\sigma_{\mathrm{z}}^2 = \sum_{\mathrm{cell}}(z_{\mathrm{cell}}\text{-}\langle z\rangle)^2 E_{\mathrm{cell}}/E_{\mathrm{tot}}$;

Jet properties have been investigated and the effects of dead channels have been analyzed with a MC simulation. This study relies on MB events generated with Pythia6 tune D6T. Figures A.2 and A.3 show the distributions, normalized to the total number of events, of the jet properties, for two different cases where the dead channels have been excluded and where all the channels have been used for the reconstruction. The differences between the two curves give an idea of the effect of the dead channels. Jets with $p_{\mathrm{T}} > 1$ GeV in the CASTOR acceptance have been selected.

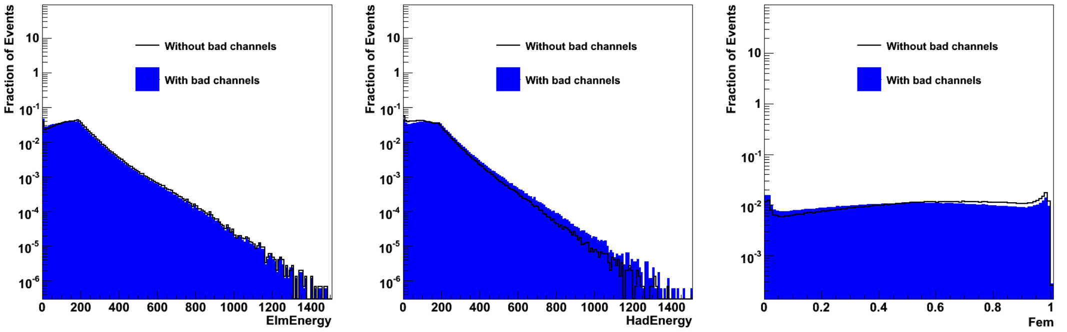


Figure A.2: Properties measured for CASTOR jets in simulation of MB events, when bad channels are included and excluded from the reconstruction: E_{em} (left), E_{had} (center) and f_{em} (right). Distributions are normalized to the total number of events.

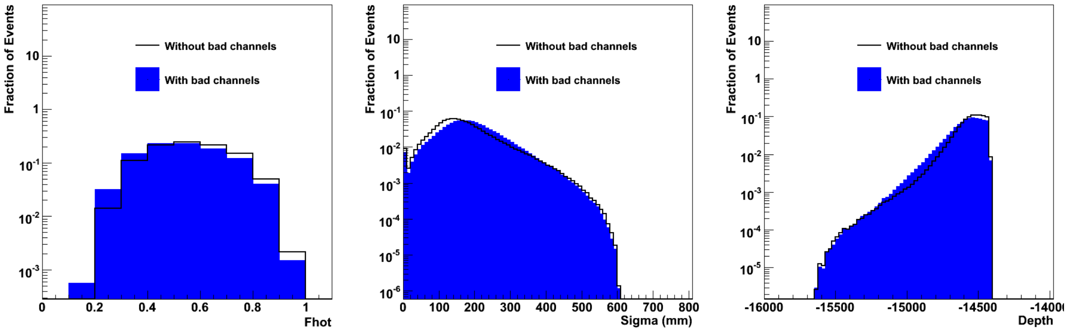


Figure A.3: Properties measured for CASTOR jets in simulation of MB events, when bad channels are included and excluded from the reconstruction: f_{hot} (left), σ_z^2 (center) and $\langle z\rangle$ (right). Distributions are normalized to the total number of events.

The reconstruction of jets, where the dead channels have been excluded, shows peculiar features: the content of hadronic energy is systematically lower than the reconstruction with all channels included, while the electromagnetic energy does not vary much, since the dead channels are mostly in the hadronic part of the calorimeter. Consequently f_{em} tends to be higher for the reconstruction without dead channels. The quantity f_{hot} shifts also towards higher values when the dead channels are excluded. Concerning the jet shape in the calorimeter, objects reconstructed without dead channels appear to be less wide (lower σ_{z}^2) and penetrating (smaller $\langle z \rangle$).

In order to perform the absolute jet calibration, the following Minimum Bias samples have been used:

- /MinimumBias/Run2010A-Apr21ReReco-v1/AOD
- /MinimumBias/Run2010B-Apr21ReReco-v1/AOD

Technical trigger bits have been required for data in order to assure that they originate from interactions arising from colliding bunches. The triggers require a consistent timing with a LHC bunch crossing and a coincidence of trigger signals in each of the two beam scintillation counters. Events originating from beam halo have been also rejected. In order to exclude beam scraping events, in events with more than 10 tracks, the fraction of high-purity tracks with respect to the total number of tracks was required to be at least 25%. At least one good primary vertex is required to be reconstructed from at least five tracks with $|z| < 15$ cm. In addition, events have been removed if they contained an invalid HCAL Trigger Readout (HTR) flag, issued during reconstruction. Such events occur rarely when, in case of a trigger burst, the HTR buffer overflows and as a consequence only some fraction of the 224 readout channels are properly sent by the Data Concentration Cards (DCC) to the DAQ. The occurrence of an invalid HTR flag determines that blocks of digis and rechits are missing in the event.

The dijet balance method has been performed by selecting only events fulfilling the following requirements:

- presence of at least a jet in $|\eta| < 2.5$ with $p_{\mathrm{T}} > 5$ GeV
- in case of the presence of an additional jet in the same phase space, $p_{\mathrm{T}}^{\mathrm{sublead}} < p_{\mathrm{T}}^{\mathrm{lead}} \cdot 0.2$;
- presence of a jet in CASTOR with $p_{\mathrm{T}} > 5$ GeV
- $\Delta\phi = |\phi^{\mathrm{CENTRAL}} - \phi^{\mathrm{CASTOR}}| > \pi$-1

The requests in $\Delta\phi$ and in $p_{\mathrm{T}}^{\mathrm{sublead}}$ ensure that the jets are well balanced and no further jets with similar p_{T} are present. Jets in the central detector ($|\eta| < 2.5$) have been reconstructed with the PF information (see Chapter 5). Balance between jets is evaluated through the variable B:

$$B = 2 \cdot \frac{F \cdot p_{\mathrm{T}}^{\mathrm{CASTOR}}(fC) - p_{\mathrm{T}}^{\mathrm{CENTRAL}}(GeV)}{F \cdot p_{\mathrm{T}}^{\mathrm{CASTOR}}(fC) + p_{\mathrm{T}}^{\mathrm{CENTRAL}}(GeV)} \tag{A.1}$$

where $p_{\mathrm{T}}^{\mathrm{CASTOR}}$ is the CASTOR p_{T} response measured in fC, $p_{\mathrm{T}}^{\mathrm{CENTRAL}}$ is the p_{T} of the central jet measured in GeV and F is the calibration factor which converts the CASTOR output in fC into a p_{T} measurement in GeV. By definition, B is a quantity whose range is between -1 and 1. A value of B close to 0 means a balanced configuration between the two jets and a good calibration, while values far from 0 indicate a not optimal calibration. Distributions of the quantity B have been measured for different values of F. In order to extract the values of B, a gaussian fit of the distributions has been performed and the mean value has been taken. Figure A.4(left) shows the B

values obtained for different F choices; the plotted uncertainty for each factor is the one obtained from the gaussian fit for the mean value parameter. The value of F which give the optimal balanced configuration is:

$$F_O^{p_\mathrm{T}} = 0.018 \pm 0.001 \mathrm{GeV/fC} \tag{A.2}$$

where the quoted uncertainty takes into account the fact that the condition $B = 0$ is fulfilled by several values of F within the fit uncertainty. Figure A.4(right) shows the balance distribution obtained in data when $F_O^{p_\mathrm{T}}$ is used. A gaussian shape, centered in 0, is well visible and the values of the parameters are shown as obtained from the fit.

The impact of this calibration factor is measured for inclusive distributions, shown in Figure A.5. Normalized differential cross sections are measured as a function of jet p_T and energy and compared to different MC samples. Jets are measured up to 2.5 TeV in energy: they correspond to jets up to 10 GeV in p_T, due to kinematics of the forward region. The agreement between data and simulation appears to be rather good, considering the fact that systematic uncertainties are not included yet. Attempts of evaluation of systematic effects are described in the following Section.

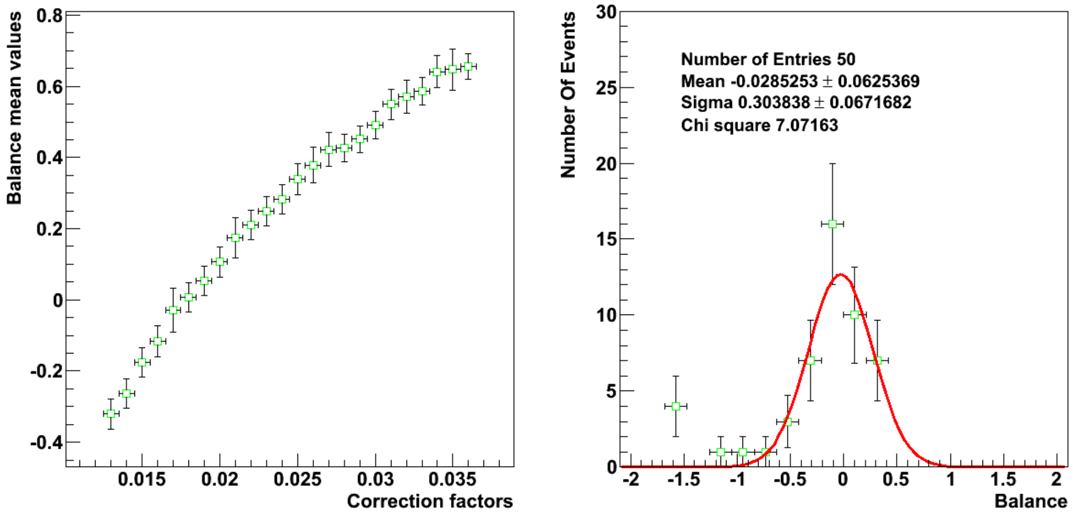


Figure A.4: (Left) Balance mean values obtained for different correction factors. The mean values are obtained from gaussian fits extracted by the balance distributions. The error bands represent the uncertainty given by the fit. (Right) Balance distribution obtained for F = 0.018; it is the correction factor which gives a mean value of the balance compatible with 0.

A.2 Systematic checks

In order to verify the robustness of the obtained results, different systematic checks have been performed:

- the balance measurement has been repeated with track-jets selected in the central region ($|\eta| < 1.9$), instead of PF jets; by using only information from the tracker, track-jets are well calibrated down to low p_T ($\sim$ 300 MeV) and reach a very good p_T resolution;
- the effect of the intercalibration uncertainty has been evaluated by varying the CASTOR channel response; for each channel, a new intercalibration factor has been chosen randomly,

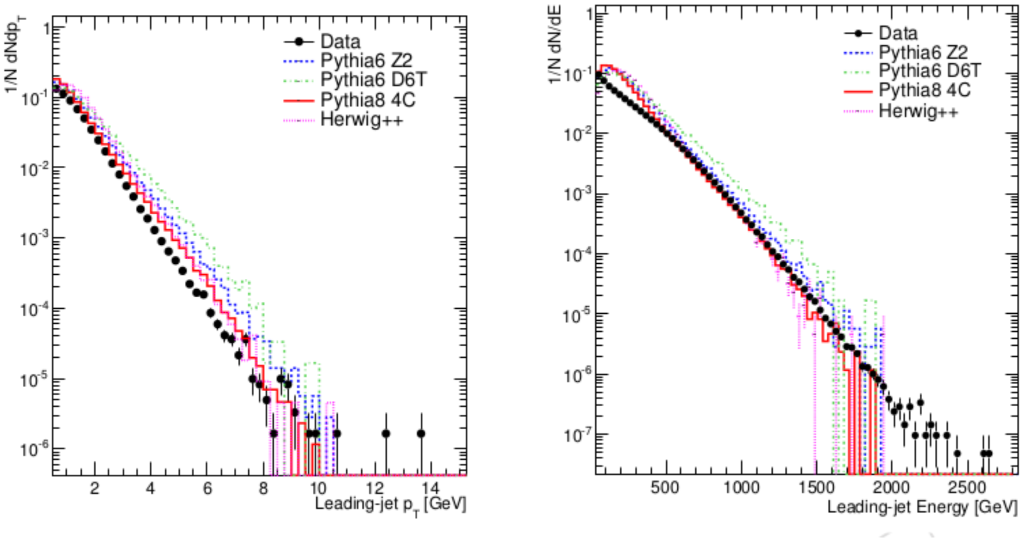


Figure A.5: Differential cross sections as a function of p_T (left) and E (right) of jets measured in CASTOR, normalized to the total number of events. A calibration factor equal to 0.018 is used to correct the data.

generated with a Gaussian distribution centered at the nominal value and with a width equal to the uncertainty of the intercalibration factors, as measured in [245];

- for the measured factor F, the dijet method has been evaluated by imposing asymmetric thresholds for the selected jets; this helps to understand whether the previously measured balance is effectively a physical quantity or an artifact of the symmetric thresholds and the rapidly decreasing jet p_{T} distributions.

While the first two checks have given very promising results, with balance factors very close to $\mathrm{F}_O^{p_{\mathrm{T}}}$, some complications have appeared when asymmetric thresholds have been imposed. In particular, a degradation of the balance is observed, with a systematic shift of the distribution with increasing differences in the p_{T} thresholds. In order to understand the behaviour of B without any detector effect, a study at the generator level has been performed. Hereafter, MC truth jets have been used and, hence, no balance factor has been applied.

Figure A.6 shows the balance obtained by setting different jet p_{T} thresholds with the PYTHIA6 sample. While the CASTOR jet p_{T} threshold is held at 5 GeV, the central jet p_{T} threshold is progressively increased to 5, 6, 7 and 8 GeV. This is referred to as “Selection test 1”. From a well balanced structure obtained with symmetric thresholds, the peak at 0 disappears with asymmetric ones (Figure A.6(left)). This is a clear indication that the balance, achieved with symmetric thresholds and previously measured, is an artifact of the falling jet p_{T} spectrum. The number of soft jets (~ 1-2 GeV) in the event spoils the balance and needs to be considered in the p_{T} evaluation. A further check has been performed in the simulation with a more sophisticated balance definition, referred to as “Selection test 2”. In particular, all the jets outside the CASTOR acceptance ($|\eta| < 5.2$) with $p_{\mathrm{T}} > 300$ MeV are added to the central jet p_{T}. The balance is then defined as:

$$B = 2 \cdot \frac{p_{\mathrm{T}}^{\mathrm{CASTOR}} - p_{\mathrm{T}}^{\mathrm{CENTRAL}} - \sum_{j>300\ \mathrm{MeV}} p_{\mathrm{T}}^{\mathrm{j}}}{p_{\mathrm{T}}^{\mathrm{CASTOR}} + p_{\mathrm{T}}^{\mathrm{CENTRAL}} + \sum_{j>300\ \mathrm{MeV}} p_{\mathrm{T}}^{\mathrm{j}}} \tag{A.3}$$

where $j > 300$ MeV is every jet in the event in $|\eta| < 5.2$. All the quantities are measured in GeV. This ensures a more inclusive balance definition where also the soft jets are considered. The jets have to fulfill:

- presence of a jet in $|\eta| < 2.5$ with $p_\mathrm{T} > p_\mathrm{T}^\mathrm{CENTRAL}$, with $p_\mathrm{T}^\mathrm{CENTRAL}$ set to 2, 3, 4 and 5 GeV for different measurements;
- presence of a jet in CASTOR with $p_\mathrm{T} > 2$ GeV

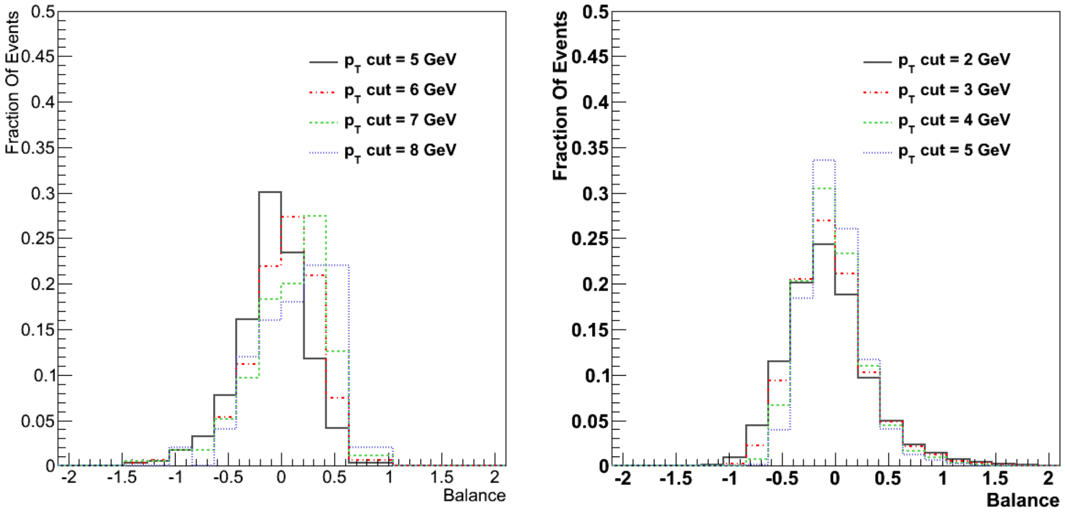


Figure A.6: Balance distributions at generator level obtained for different p_T thresholds applied to the central jet for the two different selection cuts: Selection Test 1 (left) and Selection Test 2 (right).

The result of the obtained balance, when applying the "Selection test 2", is shown in Figure A.6(right). With this new definition, the balance is found to be very stable: with increasing thresholds, the number of selected events obviously decreases but the peak structure does not shift. It has been checked that any set of other looser selections, e.g. when considering only jets with 300 MeV in $|\eta| < 2.5$, leads again to a degradation of the balance with changing thresholds. On the other hand, the new selection cannot be used at the detector level for absolute jet calibration because of the absence of a tracker system in the forward region ($2.5 < |\eta| < 5.2$) that would allow the detection of the low p_T jets, so important for the balance measurement.

A.3 Conclusions

A first look at jets in CASTOR has been carried out, by studying the jet ID properties and the effect of the dead channels. An attempt of absolute calibration has been performed for jets in CASTOR with the use of the dijet balance method. It has been shown that soft jets with p_T down to 300 MeV cannot be ignored in the balance measurement and need to be considered in the whole region outside CASTOR. This introduces a detector issue, since CMS is not able to detect such low p_T jets in the detector region of $2.5 < |\eta| < 5.2$. Currently, more work is ongoing in the CASTOR group to find a compromise between a stable balance measurement from a physics point of view and the detector possibilities. This prevents the current possibility to extend the phase space of the four-jet measurements to the very forward pseudorapidities covered by CASTOR. Part of the work described in this appendix is described in [246], as a CMS internal note.

Appendix B

Higgs as a gluon trigger

In this Chapter, the idea of using the detection of the Higgs boson as a gluon trigger, is described. The work is documented in [14] and the whole paper is reported at the end of this Appendix.

The discovery of the Higgs boson announced by the two collaborations, ATLAS [34] and CMS [33] in 2012, is indeed a fundamental piece of information that further confirms and validates the theory of the SM as basic theory of the elementary particle physics. In particular, the mechanism of the origin of the mass of the particles, the appearance of divergences in cross section calculations and many other phenomena are now explained by the introduction of the Higgs boson in the particle field collection. Despite the fact that it is surely not the last piece to be added in the SM, since many observations still need to be explained, this discovery is a breakthrough in the history of particle physics. Currently, research effort is dedicated to identify the nature of the new boson; in particular, physicists want to understand whether it is really the SM Higgs or it has other characteristics, typical, for instance, of a SuperSymmetric Higgs boson. In addition, different research branches and phenomenological studies have considered the use of the properties of the Higgs boson for different purposes, for instance, DPS measurements [99] or extraction of gluonic PDF.

A non-conventional idea is also to try to deepen the understanding of the QCD mechanisms by profiting from Higgs boson physics. The Higgs boson offers a very peculiar case whose main features can be summarized in a few points:

- the Higgs boson carries a neutral electric charge and has a mass of around 125 GeV/c^2;
- its main production diagram is a gluon-gluon fusion, where two gluons interact in a loop of heavy quarks which then creates the Higgs boson;
- it is a colour-neutral particle;
- it carries spin 0.

These characteristics make the study of Higgs boson observables very interesting if compared, for instance, to a Z boson production. The Z boson has a similar scale, equivalent to its mass of about 91 GeV/c^2, but its production is initiated by a quark-antiquark initial state. Hence, the measurement of the Higgs and the Z boson decay products gives a direct comparison of colour-neutral final states, that sit approximately at the same scale, induced by respectively gluon and quark interactions. The observables that might be studied are, for instance, the boson p_T spectrum, the UE associated activity or the correlation in the transverse plane between the boson direction and the direction of an additional jet (see Section 2). Since this thesis focuses on measurements sensitive to MPI contributions, results and insights coming from UE activity in the two boson-induced processes are described in the following.

The measurement of UE observables, like charged particle multiplicity or charged p_T sum, may be thought for Higgs and Drell-Yan[1] production in the same way as described in Chapter 2. The direction of the hard scattering is defined by the boson and the transverse plane is divided in the usual three regions: toward, transverse and away according to the azimuthal angle with respect to the boson direction. Charged particles are counted with $p_T > 0.5$ GeV in $|\eta| < 2.0$ and UE observables are measured separately in the different regions. A phenomenological study at the generator level has been performed: MC samples generated with POWHEG at 7 TeV, interfaced with the UE simulation provided by PYTHIA6 Tune Z2*, are used. The two samples are generated for inclusive Higgs and Drell-Yan at NLO in a given mass interval. Only the boson mass region, between 115 and 135 GeV/c^2, is considered for both analyses. Only gluon-gluon fusion diagrams are considered for Higgs production. The two bosons are set to stable, in order not to have a contamination of the UE activity, from the decay products.

In figures B.1 and B.2, the charged particle multiplicity and p_T sum are shown in the three different regions of the transverse plane, as a function of the boson p_T.

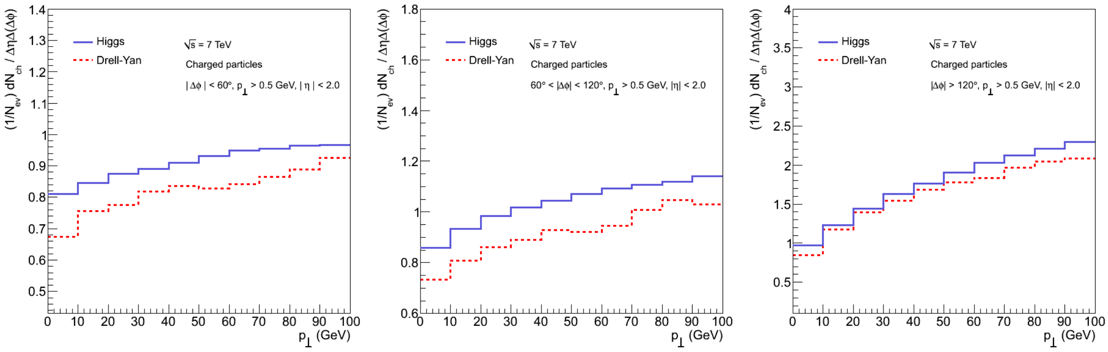


Figure B.1: Charged particle multiplicity measured for Higgs and Z boson production in the toward (left), transverse (center) and away (right) regions: samples have been generated at 7 TeV with a POWHEG ME, interfaced with the PS and UE provided by PYTHIA6 Tune Z2. Only the boson mass region, between 115 and 135 GeV/c^2, is considered for both analyses.*

Different behaviours are observed for the two processes; in particular, a higher activity is expected for Higgs-induced processes. The difference with respect to the UE in Drell-Yan processes might be quantified in a 20-30% increase in charged particle multiplicity and p_T sum. This is mainly due to the different Sudakov factors in the splitting functions (see Chapter 2) for gluons and quarks: gluons, which constitute the initial state for a Higgs boson production, have a bigger Sudakov splitting factor (with a multiplicative factor equal to 6 for gluon-gluon splittings), while

---

[1] A Drell-Yan process is the production of a virtual γ or Z boson, through $q\bar{q}$ annihilation.

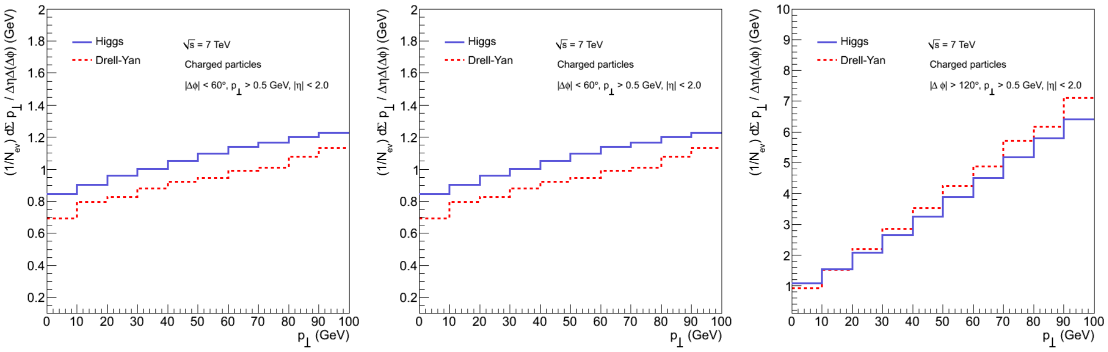


Figure B.2: Charged particle p_T sum measured for Higgs and Z boson production in the toward (left), transverse (center) and away (right) regions: samples have been generated at 7 TeV with a POWHEG *ME, interfaced with the PS and UE provided by* PYTHIA*6 Tune Z2*. Only the boson mass region, between 115 and 135 GeV/c^2, is considered for both analyses.*

quarks and antiquarks, whose mutual interactions produce Drell-Yan events, have a smaller one (the multiplicative factor is equal to 4/3 for quark-quark transitions). Therefore a gluon has a higher probability to split into further partons than a quark and, thus, a larger activity in Higgs processes is predicted. Similar conclusions can be drawn for a center of mass energy of 14 TeV [14]. At any considered energy, high statistical accuracy needs to be achieved to obtain a significant number of Higgs candidates and this program is in the plan of the high luminosity LHC phase. Note that UE activity in Drell-Yan events has been already measured by CMS in the first LHC phase with 2011 data [81]. Despite the fact that a high luminosity brings a very high number of pile-up events (for the Run II, about 100 overlapped interactions are expected at a peak luminosity of about 2-3·10^{34} cm^{-2}s^{-1}), the study of the UE in Higgs and Drell-Yan processes remains feasible if differences between the two measurements are considered. This is due to the fact that pile-up interactions are independent of each other and their contribution to the considered activity sums up. This has been shown in the study at the generator level documented in [247]. Hence, this measurement is very promising for a high luminosity phase and a physics plane has been proposed for the LHC run II in 2015.

PHYSICAL REVIEW D **88,** 097501 (2013)

Higgs boson as a gluon trigger

P. Cipriano,[1] S. Dooling,[1] A. Grebenyuk,[1] P. Gunnellini,[1] F. Hautmann,[2,3,4] H. Jung,[1,5] and P. Katsas[1]
[1]*Deutsches Elektronen Synchrotron, D-22603 Hamburg, Germany*
[2]*Department of Physics and Astronomy, University of Sussex, Brighton BN1 9QH, United Kingdom*
[3]*Physics Department, University of Oxford, Oxford OX1 3NP, United Kingdom*
[4]*Rutherford Appleton Laboratory, Chilton OX11 0QX, United Kingdom*
[5]*Elementaire Deeltjes Fysica, Universiteit Antwerpen, B-2020 Antwerp, Belgium*

(Received 30 October 2013; published 26 November 2013)

In the forthcoming high-luminosity phase at the LHC, many of the most interesting measurements for precision QCD studies are hampered by conditions of large pileup, particularly at not very high transverse momenta. We study observables based on measuring ratios of color-singlet currents via Higgs boson and Drell-Yan production, which may be accessed also at large pileup, and used for an experimental program on QCD physics of gluon fusion processes in the LHC high-luminosity runs. We present results of Monte Carlo calculations for a few specific examples.

DOI: 10.1103/PhysRevD.88.097501 PACS numbers: 14.80.Bn, 12.38.−t

I. INTRODUCTION

The observation of the Higgs boson by the ATLAS and CMS experiments [1] marks the beginning of a revolutionary era in high-energy physics. It affects profoundly the paradigms by which we define the limits of our knowledge on the nature of interactions of elementary particles. This observation gives us confidence in the physical picture of fundamental interactions encoded by the Standard Model (SM) Lagrangian and provides us with guidance in the search for its generalizations.

The electroweak sector of the SM and the nature of electroweak symmetry breaking will be explored in detail in the coming years of operation of the LHC by measuring properties of the observed boson [2]. In this paper, we remark that the observation of the Higgs boson opens up the possibility of a rich experimental program in the strong-interaction sector of the SM as well. In particular, we propose that a program of QCD measurements at high luminosity can be carried out at the LHC by using the Higgs boson as a trigger, focusing on QCD gluonic processes at high mass scales.

Classic collider probes of QCD in e^+e^- annihilation, deep inelastic ep scattering, and Drell-Yan production (DY) all involve color-singlet currents which couple to quarks. With the Higgs, for the first time, LHC experiments will probe QCD by a color-singlet current which, in the heavy top limit, couples to gluons. The physics of gluon fusion processes can be explored from a new perspective compared to experimental investigations over the past three decades. As illustrated below, we propose measuring systematically differences of differential distributions for Higgs and Drell-Yan final states. This comparison allows one to access experimentally distinctive QCD features of gluon fusion physics.

In the next high-intensity phase at the LHC, one faces high pileup conditions leading to large numbers of overlaid events. In these conditions, many of the most interesting measurements for precision QCD studies, particularly for not very high transverse momenta, become extremely difficult—see e.g. Refs. [2,3]. Here we argue that by studying the differences of Higgs and Drell-Yan for masses around 125 GeV, the effects of pileup largely drop out. This offers the possibility of a program of QCD measurements of great physics interest in the high-luminosity runs of the LHC.

In this paper, we illustrate this by Monte Carlo simulation for three specific examples: the ratio of Higgs vs Drell-Yan $p_\perp$ spectra; the structure of the associated underlying event and charged-particle multiplicities; and the scattering angle in the center-of-mass reference frame. These involve QCD physics both at high transverse momenta and at low transverse momenta, and allow one to study both high-x and low-x physical effects.

We contrast the distinctive features of the Higgs trigger with other LHC short-distance probes such as jets, heavy flavors, and vector boson pairs which either couple perturbatively to color-octet and color-triplet sources on an equal footing, or imply final-state color-charged particles, or both.

We leave to detailed phenomenological investigations the study of the optimal channels to be used to access gluon fusion and suppress Higgs production by vector boson fusion and quark annihilation; of the luminosity requirements for reaching sufficient statistics; and of the different treatment of pileup for different channels.

Very recently, the ATLAS Collaboration has presented its first measurements of Higgs differential cross sections based on the 2012 data set in the diphoton decay channel [4].

II. HIGGS VS DRELL-YAN

Consider first transverse momentum spectra for Higgs bosons and for Drell-Yan (DY) pairs in the invariant mass range 115 GeV $< M <$ 135 GeV. Transverse momentum spectra, comparing Higgs and Z bosons, were examined

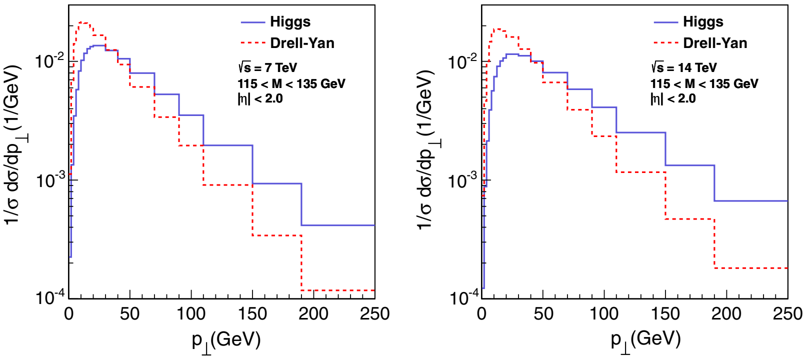


FIG. 1 (color online). Normalized transverse momentum spectra for Higgs bosons and for Drell-Yan pairs.

early on in Ref. [5]. The transverse momentum spectra can be described by QCD factorization in the form

$$d\sigma/dp_\perp = \int H \otimes S \otimes J_1 \otimes J_2, \qquad (1)$$

decomposing the cross section into hard (H), soft (S), and collinear-to-initial-states (J_1, J_2) contributions—see e.g. Ref. [6] for analysis of how this decomposition arises. In Fig. 1, we show the result of Monte Carlo simulations for the $p_\perp$ spectra in the central region based on the next-to-leading-order (NLO) POWHEG [7] event generator interfaced with PYTHIA [8] shower, at $\sqrt{s} = 7$ TeV and $\sqrt{s} = 14$ TeV. In Fig. 2, we plot the ratio of the Higgs and DY spectra at invariant mass 115 GeV $< M <$ 135 GeV.

The $p_\perp \ll M$ region of the spectrum measures infrared aspects of the cross section in Eq. (1); i.e., (i) the ratio of the gluon vs quark Sudakov form factor [factor S in Eq. (1)], and (ii) the evolution of the collinear-to-initial-states functions [factors J_1, J_2 in Eq. (1)]. In particular, gluon polarization terms $p_\perp^\mu p_\perp^\nu$ in gluon fusion, related to eikonal polarizations at high energy [9], give rise to distinctive radiation patterns from initial-state functions in the Higgs case—see e.g. Ref. [10]. The $p_\perp \geq M$ region measures the ultraviolet function H in Eq. (1) and the features of hard jets recoiling against the Higgs or DY pair. In particular, the leading-jet contribution to the measured ratio depends on the $p_\perp$ distribution for the spin-1 vs spin-1/2 exchange and on the corresponding color emission probabilities. Further aspects on jet recoil are discussed below in the context of angular distributions.

In the large pileup environment of the high-luminosity LHC runs, one has to deal with the contribution of large numbers of overlaid events. However, this contribution cancels in the comparison of Higgs to DY spectra at fixed invariant mass. Using this comparison, one can go to low $p_\perp$ and access QCD effects in this region experimentally also at high pileup.

Measurements on gluon fusion which can be performed using the Higgs trigger open a new experimental area. They may also be relevant to interpreting data for other, more complex processes, e.g. processes that depend on both quark and gluon channels on an equal footing, or involve color-charged particles in the final state.

One such example is given by top quark production. This is often studied as a process sensitive to gluonic initial states at the LHC. For instance, the top quark $p_\perp$ spectrum [11] receives contributions at low $p_\perp$ from the gluon Sudakov form factor and gluonic initial-state recoil analogous to those discussed above. However, since the final

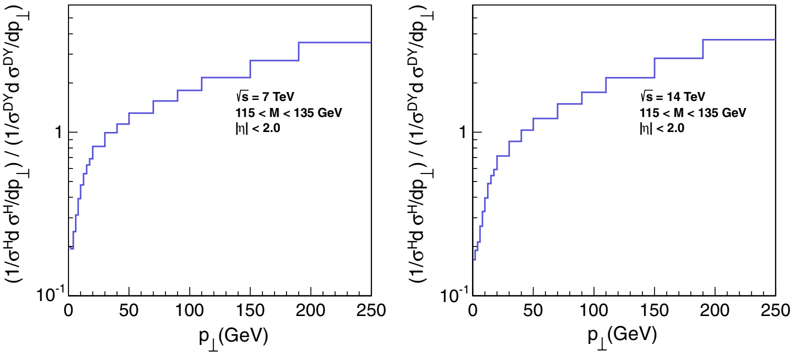


FIG. 2 (color online). Ratio of Higgs to DY spectra vs $p_\perp$.

state is not a color-singlet current, the analysis of the $p_\perp$ spectrum is made more complex by final soft color emission. The Higgs case serves to single out the initial-state contributions, including gluon polarization effects.

For observables more exclusive than the cross section in Eq. (1), e.g. measuring the associated jets, full QCD factorization formulas are still lacking. For parton shower event generators, inclusive measurements are still useful to control methods [12,13] for merging parton showers and matrix elements. Higgs vs DY studies similar to those considered above can be done, for instance, in boson + jet states, now fixing, in addition to invariant mass, the jet transverse momentum or rapidity.

III. UNDERLYING EVENTS

The structure of underlying events and color flows associated with Higgs boson final states was investigated long ago [14] as a possible method to analyze $gg \to H$ and $WW \to H$ production mechanisms. In the case of vector boson final states it was pointed out [15] that the treatment of parton showers, and in particular of the recoils in the shower, is essential for a proper description of W/Z spectra. This affects the amount of multiparton interactions [16] needed to describe the events [15,17]. Analogous effects may be investigated for gluonic showers [18,19] in the case of events associated with Higgs final states.

We follow the treatment [20] of underlying events in the azimuthal plane, with the directions of the Higgs momentum and the DY-pair momentum, respectively, defining the origin in the azimuthal plane. In Figs. 3 and 4, we show the result of NLO POWHEG + PYTHIA Monte Carlo calculations for charged-particle multiplicities associated with Higgs and DY. (Analogous calculations can be usefully performed for multiplicities of minijets defined e.g. as in Ref. [21].) We plot the average multiplicity vs Higgs and DY $p_\perp$ (Fig. 3) and the multiplicity distribution (Fig. 4) in the transverse region of the azimuthal plane ($60° < |\Delta\Phi| < 120°$).

The distributions in the Higgs case are dominated by higher multiplicities from gluon cascades.

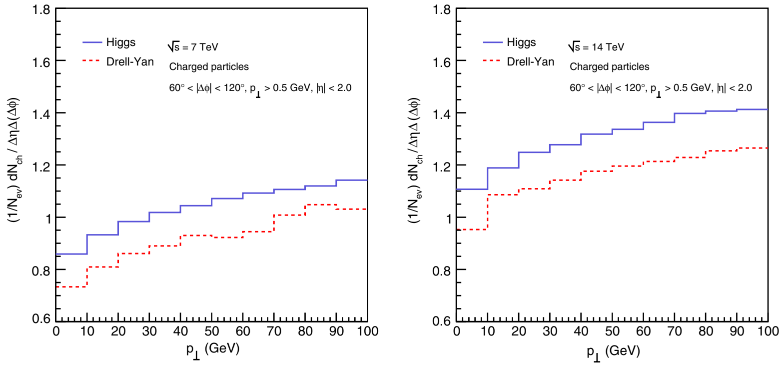


FIG. 3 (color online). Normalized charged-particle average multiplicity in the transverse region of the azimuthal plane vs the Higgs transverse momentum (solid blue line) and the DY transverse momentum (red dashed line).

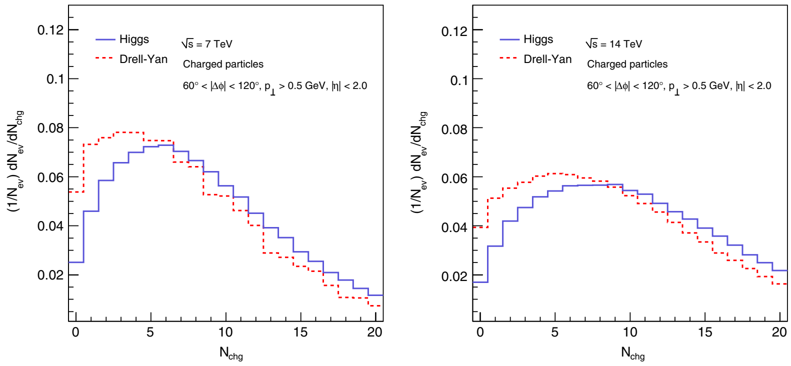


FIG. 4 (color online). Charged-particle multiplicity distribution in the transverse region of the azimuthal plane in the Higgs (solid blue line) and Drell-Yan (red dashed line) cases.

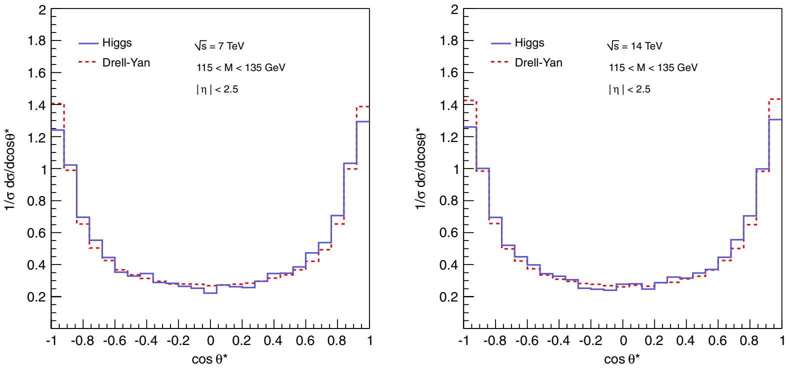


FIG. 5 (color online). Angular distribution in the center-of-mass scattering angle.

Similarly to the case of the previous section, the effects of a large number of overlaid events due to pileup will be reduced if one measures the difference between Higgs and DY underlying event distributions.

IV. ANGULAR DISTRIBUTIONS

Besides soft radiation from underlying events, we consider Higgs vs DY distributions in the case of hard radiation accompanying the heavy bosons, for example boson + jet [22]. For Higgs production, the angular distribution in the scattering angle θ^* of the boson-jet center-of-mass frame is characterized by the scalar coupling to gluons partially canceling the small-angle Coulomb singularity $d\theta^{*2}/\theta^{*4}$ from gluon scattering—see e.g. Ref. [9]. The Drell-Yan θ^* distribution is determined by spin-$1/2$ exchange. Owing to the cancellation from the scalar coupling to gluons, the angular distributions have the same small-angle asymptotics in the Higgs and DY cases, despite the two processes occurring via spin-1 and spin-$1/2$ exchange. The $\theta^* \to 0$ behavior thus tests the Higgs spin at the level of the production cross section.

In Fig. 5, we consider one-jet production associated with Higgs and Z bosons, and show the differential distributions in $\cos\theta^*$, for jet $p_\perp > 20$ GeV and boson-jet invariant mass m such that 200 GeV $< m < 500$ GeV. The rise for increasing $\cos\theta^*$ reflects the mechanism described above. This large $\cos\theta^*$ power counting is the basic reason why the difference between Higgs and DY in the low-$p_\perp$ regions of Figs. 1 and 2 gives a measurement of higher-loop radiative contributions. Further effects from higher-order color emission may be analyzed via angular correlations in the boson-jet azimuthal plane in the laboratory frame.

In summary, this paper points out that a program of QCD measurements can be carried out in the high-luminosity phase at the LHC, using the Higgs boson as a gluon trigger. By measuring systematically differences between Higgs and Drell-Yan differential distributions for masses around 125 GeV, the effects of pileup largely cancel. Such measurements allow one to access experimentally, for the first time, gluon fusion processes at high mass scales via a color-singlet current. Detailed studies are warranted to investigate quantitatively the reduction of pileup contributions in different channels, the optimal Higgs channels to access gluon fusion by suppressing vector boson fusion and quark annihilation, and the required Higgs statistics. The observables discussed in this paper illustrate that this program spans a broad range of physics issues on strong interactions, from soft gluon dynamics showing up in the ratio of Higgs to DY low-$p_\perp$ spectra, to underlying events and multiple parton interactions associated with gluonic showers, to hard-QCD contributions in large-$p_\perp$ spectra and angular distributions for boson + jet production. These angular distributions in particular may be used to test the spin of the Higgs at the level of production processes.

ACKNOWLEDGMENTS

F. H. gratefully acknowledges the hospitality and support of the Terascale Physics Helmholtz Alliance and DESY.

---

[1] G. Aad *et al.* (ATLAS Collaboration), Phys. Lett. B **716**, 1 (2012); S. Chatrchyan *et al.* (CMS Collaboration), Phys. Lett. B **716**, 30 (2012).
[2] ATLAS and CMS collaborations, submission to European Strategy for Particle Physics, Krakow, 2012.
[3] CMS Collaboration, "QCD at the Extremes", in European Strategy for Particle Physics, Krakow, 2012.
[4] ATLAS Collaboration, Report No. ATLAS-CONF-2013-072.
[5] I. Hinchliffe and S. F. Novaes, Phys. Rev. D **38**, 3475 (1988).

[6] J.C. Collins and F. Hautmann, Phys. Lett. B **472**, 129 (2000).
[7] S. Frixione, P. Nason, and C. Oleari, J. High Energy Phys. 11 (2007) 070.
[8] T. Sjöstrand, S. Mrenna, and P. Skands, J. High Energy Phys. 05 (2006) 026.
[9] F. Hautmann, Phys. Lett. B **535**, 159 (2002).
[10] S. Mantry and F. Petriello, Phys. Rev. D **81**, 093007 (2010); **83**, 053007 (2011).
[11] CMS Collaboration, Report No. CMS-PAS-TOP-12-028; CMS Collaboration, Eur. Phys. J. C **73**, 2339 (2013); ATLAS Collabortaion, Report No. ATLAS-CONF-2012-149.
[12] S. Plätzer, arXiv:1307.0774; J. High Energy Phys. 08 (2013) 114.
[13] L. Lönnblad and S. Prestel, Prog. Part. Nucl. Phys. **70**, 1 (2013); arXiv:1211.7228.
[14] Yu. L. Dokshitzer, V. A. Khoze, and T. Sjöstrand, Phys. Lett. B **274**, 116 (1992).
[15] P. Z. Skands, Phys. Rev. D **82**, 074018 (2010).
[16] T. Sjöstrand and M. van Zijl, Phys. Rev. D **36**, 2019 (1987).
[17] R. D. Field, arXiv:1010.3558.
[18] L. Lönnblad and M. Sjödahl, J. High Energy Phys. 02 (2004) 042; 05 (2005) 038.
[19] M. Deak *et al.*, arXiv:1006.5401.
[20] S. Chatrchyan *et al.* (CMS Collaboration), J. High Energy Phys. 09 (2011) 109.
[21] M. Deak, F. Hautmann, H. Jung, and K. Kutak, Eur. Phys. J. C **72**, 1982 (2012).
[22] R.K. Ellis, W.J. Stirling, and B.R. Webber, *QCD and Collider Physics* (Cambridge University Press, Cambridge, 1996).

Appendix C

Perturbative and nonperturbative corrections in shower event generators

As already described throughout the whole thesis, jet physics is fundamental and essential to study and understand details of strong interactions in the SM. In particular, on one hand, very precise jet measurements have been performed in a large p_{T} range and in separate regions of the η acceptance, giving the opportunity for testing the accuracy of different predictions. On the other hand, more and more sophisticated calculations going beyond LO have been released for multijet scenarios and those experimental measurements offer the most natural testing ground for their validity. For a fair comparison between the two, a MC simulation of perturbative and nonperturbative effects, added on top of a ME computation, is needed in order to connect the calculation at the parton level with the final jet observables. A direct comparison among jet distributions and parton spectra is indeed inconsistent. So far, these corrections, for example applied to NLO ME calculations, are usually obtained by considering nonperturbative effects simulated with LO generators (mostly PYTHIA and HERWIG). Nonperturbative effects are usually obtained by ratios of observables with different parts of the simulation switched on; in particular, they include contributions from MPI and hadronization components and they are thus calculated by the ratio of a sample where the whole UE simulation is switched on and a sample where MPI and hadronization are switched off. These ratios are then multiplied bin-wise to the ME calculations for a given observable and then, the so-corrected theory predictions are compared to the data.

Perturbative corrections, which are related to the PS contribution, are generally neglected. This approach may bring to several inconsistencies: first of all, since nonperturbative corrections are calculated by using LO event generators and then applied to, for instance, NLO ME, like FASTNLO [248], they might not be strictly correct because the additional hard radiation is treated differently by the two calculations. Secondly, perturbative corrections introduced by the PS might be relevant especially for less central final states. This is, for instance, speculable in [249], where the agreement between data and NLO calculations progressively worsens when going to more forward pseudorapidities.

This study aims to discuss these points and proposes an alternative way to evaluate corrections to be applied to a NLO ME calculation. The new method provides a factorized correction that includes perturbative and nonperturbative effects, both evaluated with a NLO ME, interfaced with PS and an UE simulation. It is, in particular, shown that nonperturbative effects are slightly different if evaluated by starting from a LO or a NLO ME. Furthermore, studies about the effects of the single parts of the UE simulation on the longitudinal momentum fraction x carried by the partons are performed. It is shown that when adding PS, MPI and hadronization to the partons participating in the hard scattering, a systematic shift of the x quantity is produced. This effect, studied for different hard processes and pseudorapidity ranges, is found to reach a highly significant level in the forward region and, indeed, introduces again an inconsistency in MC event generators: this is related to the fact that a value of x is used for the hard scattering, which is different from the actual x value of the partons after evolution[1]. Hence, a speculation on the use of k_{T}-dependent PDFs, is made, in order to solve this incompatibility.

In Section C.1, the part of the work related to the perturbative and nonperturbative corrections obtained with LO generators is shown in full detail. In particular, the new approach is tested and the corrections expected for every component of the PS and of the UE are considered and studied in different pseudorapidity ranges and for different cone widths to check their influences. At the end of Section C.1, the paper which documents the whole new approach as published in PRD [13], is reported.

C.1 Perturbative and nonperturbative effects in jet p_{T} spectra

Currently, the standard nonperturbative effects are evaluated by using a LO generator (e.g. Pythia). The ratio between a nominal event generation using a well performing tune and a sample with hadronization and MPI effects switched off is taken as correction. The perturbative effects are instead ignored. This approach has been used in several jet measurements [204, 249]. Note that the NP corrections, so defined, which are extracted from a pure LO event generator, are used to correct any available NLO calculation to bring it to the jet level for direct comparison. In a compact formulation, the NP correction factors can be defined as:

$$C_0^{\mathrm{NP}} = \frac{N_{\mathrm{LO}}^{\mathrm{PS+HAD+MPI}}}{N_{\mathrm{LO}}^{\mathrm{PS}}} \tag{C.1}$$

where in the superscript, the components of the simulated UE are listed and in the subscript the order of the ME is specified. In [204], these factors are evaluated with different tunes, generators and PDF sets, and the envelope resulting from them is considered as theoretical uncertainty of the correction factors.

The proposed new method takes into account two types of corrections, instead:

---

[1] In a MC event generator, the hard scattering is generated first and then a backward evolution is generated for the partons in the initial state.

- perturbative effects, related to initial and final state radiation;
- nonperturbative effects, related to hadronization and MPI components.

Since it is currently possible to match NLO calculations, like e.g. Powheg, with PS, it is important that these corrections are evaluated with a NLO ME. This removes possible inconsistencies in the corrections, due to different treatment of hard emissions in LO and NLO ME. The new correction factors might be defined in a compact way as:

$$C^{\text{NLO}} = C^{\text{NP}} \cdot C^{\text{PS}} \tag{C.2}$$

where C^{NP} refers to the nonperturbative corrections and is equal to:

$$C^{\text{NP}} = \frac{N_{\text{NLO}}^{\text{PS+HAD+MPI}}}{N_{\text{NLO}}^{\text{PS}}} \tag{C.3}$$

and C^{PS} includes the perturbative corrections and is defined as:

$$C^{\text{PS}} = \frac{N_{\text{NLO}}^{\text{PS}}}{N_{\text{NLO}}} \tag{C.4}$$

Note that in Equation C.2, the two effects are assumed to factorize in the final correction. The syntax of Equations C.3 and C.4 is the same as Equation C.1: the superscripts list the parts of the UE simulated while the subscript indicates the order of the ME. Specifically, while $N_{\text{NLO}}^{\text{PS+HAD+MPI}}$ represents a sample where all the UE components are simulated with a NLO ME, $N_{\text{NLO}}^{\text{PS}}$ implements only the parton shower, leaving at the end of the generation only coloured particles (not hadronized) and N_{NLO} returns only the partons of the ME, without any further evolution.

Before going to study the new corrections for a NLO ME, it is however important to evaluate them with a LO ME. This is not particularly relevant for jet measurements, because any LO ME generator is able to fully produce hadronic final states with all the UE components, but it is interesting in order to interpret the NLO corrections and understand the nature of analogies and differences between them. The part of obtaining the corrections with LO ME is described in the following in full detail.

The Pythia6 event generator has been used as LO ME generator. The UE components are simulated according to the parameters set in the Z2 tune. The two separate correction factors have been studied in different pseudorapidity ranges (five regions from 0 to 2.5, in η steps of 0.5) and in a p_{T} range spanning from 19 GeV to 2 TeV. The chosen phase space region reflects the measurement performed at the CMS experiment [204]. Different clustering cone apertures, ΔR, for jets reconstructed with the anti-k_{T} algorithm have been also considered and compared: these have been set to 0.5 and 0.7, values used in the jet reconstruction in CMS. Figure C.1 shows the correction factors for jet p_{T} spectra in increasing η ranges for a cone width of 0.5 and a center-of-mass energy of the collisions equal to 7 TeV. The correction factors are the ones defined in Equations C.3 and C.4, with the only difference that a LO ME is used. Fits to the points are also drawn to give a continuous description of the curves. What might be concluded from the curves is very interesting; $C_{\text{LO}}^{\text{NP}}$ is rapidly decreasing for increasing p_{T}, starting from an initial

30% at $p_{\mathrm{T}} \sim 20$ GeV and going towards an asymptotic behaviour at 1 for high p_{T}. This trend is very similar in all the η regions and correction factors very close to 1 are already observed for $p_{\mathrm{T}} > 40$ GeV. The behaviour of $C^{\mathrm{PS}}_{\mathrm{LO}}$ is quite different. In particular, in the most central regions ($|\eta| < 1.0$), it is flat and very close to 1, while going to more forward regions, the $C^{\mathrm{PS}}_{\mathrm{LO}}$ factors start to deviate from the unity, especially at high p_{T}, reaching differences of about 40%. In Figure C.2, the same correction factors are measured for a cone width equal to 0.7, in order to see the effect of an increased clustering cone. In this case, $C^{\mathrm{NP}}_{\mathrm{LO}}$ have higher values at low p_{T}, starting from 3 at $p_{\mathrm{T}} \sim 20$ GeV, but they exhibit again the asymptotic behaviour, having factors close to 1 for $p_{\mathrm{T}} > 150$ GeV. The $C^{\mathrm{PS}}_{\mathrm{LO}}$ factors show a similar trend as observed for smaller cone widths: while in the most central regions, the corrections are very close to 1 and decreasing for increasing p_{T}, at more forward regions, $C^{\mathrm{PS}}_{\mathrm{LO}}$ start to deviate from the unity and the main differences are observed for high p_{T}. The results for $C^{\mathrm{NP}}_{\mathrm{LO}}$ confirm the behaviour and the values, used in [204] and [250], respectively for a cone width of 0.5 and 0.7. Note that in those measurements, the $C^{\mathrm{PS}}_{\mathrm{LO}}$ corrections have been neglected in the whole phase space but, as shown here, they might play a relevant role, especially in the high η regions. In [13], it is also shown that the behaviours of $C^{\mathrm{NP}}_{\mathrm{LO}}$ and of $C^{\mathrm{NP}}_{\mathrm{NLO}}$, respectively evaluated with the PYTHIA6 LO ME and with the POWHEG NLO dijet ME, interfaced with PYTHIA6, are very similar but they differ of about 20% in the low p_{T} region, due to the different upper scales set for the MPI. In particular, the corrections for a LO ME are bigger because the MPI scale is higher. The behaviour of $C^{\mathrm{PS}}_{\mathrm{LO}}$ and $C^{\mathrm{PS}}_{\mathrm{NLO}}$ is also comparable.

Additionally, the single components that compose these correction factors have been also studied in order to understand the relative contributions of UE parts. In particular, different elements of the UE have been simulated and compared to the parton level of the LO ME. Figure C.3 shows the results for $\Delta R = 0.5$ for the usual η ranges. Various curves are drawn in the plots:

- $\sigma_{\mathrm{LO+PYTHIA(MPI+HAD+PS)}}$: LO ME with PS and full UE generated;
- $\sigma_{\mathrm{LO+PYTHIA(MPI+PS)}}$: LO ME with PS and MPI generated and hadronization switched off;
- $\sigma_{\mathrm{LO+PYTHIA(PS)}}$: LO ME with PS generated and MPI and hadronization switched off;
- $\sigma_{\mathrm{LO+PYTHIA(ISR)}}$: LO ME with only ISR generated;
- $\sigma_{\mathrm{LO+PYTHIA(FSR)}}$: LO ME with only FSR generated.

All these samples as a function of the jet p_{T}, are normalized to the cross section of the ME with no further evolution simulated. The FSR (blue line) tends to decrease the cross section with ratios below 1, while the ISR (red line) brings the cross section up with correction factors between 2 and 1.2 in the whole phase space. Note that the combined effect of the two, represented in the PS curve (black line) and equivalent to the previously discussed $C^{\mathrm{PS}}_{\mathrm{LO}}$, is not the product of the single ISR and FSR contributions. In particular, they result to be very interconnected in the parton evolution implemented in the MC generators. The effects of introducing MPI (green line) can be evaluated by the difference between the black and the green curve; it is sizeable only for low p_{T} jets, up to 50-60 GeV. The same conclusion is valid for the effect of the hadronization (violet line). The behaviour of the curves at more forward regions confirms the previous observations: at high jet p_{T}, correction factors below 1 start to appear, driven by PS effects.

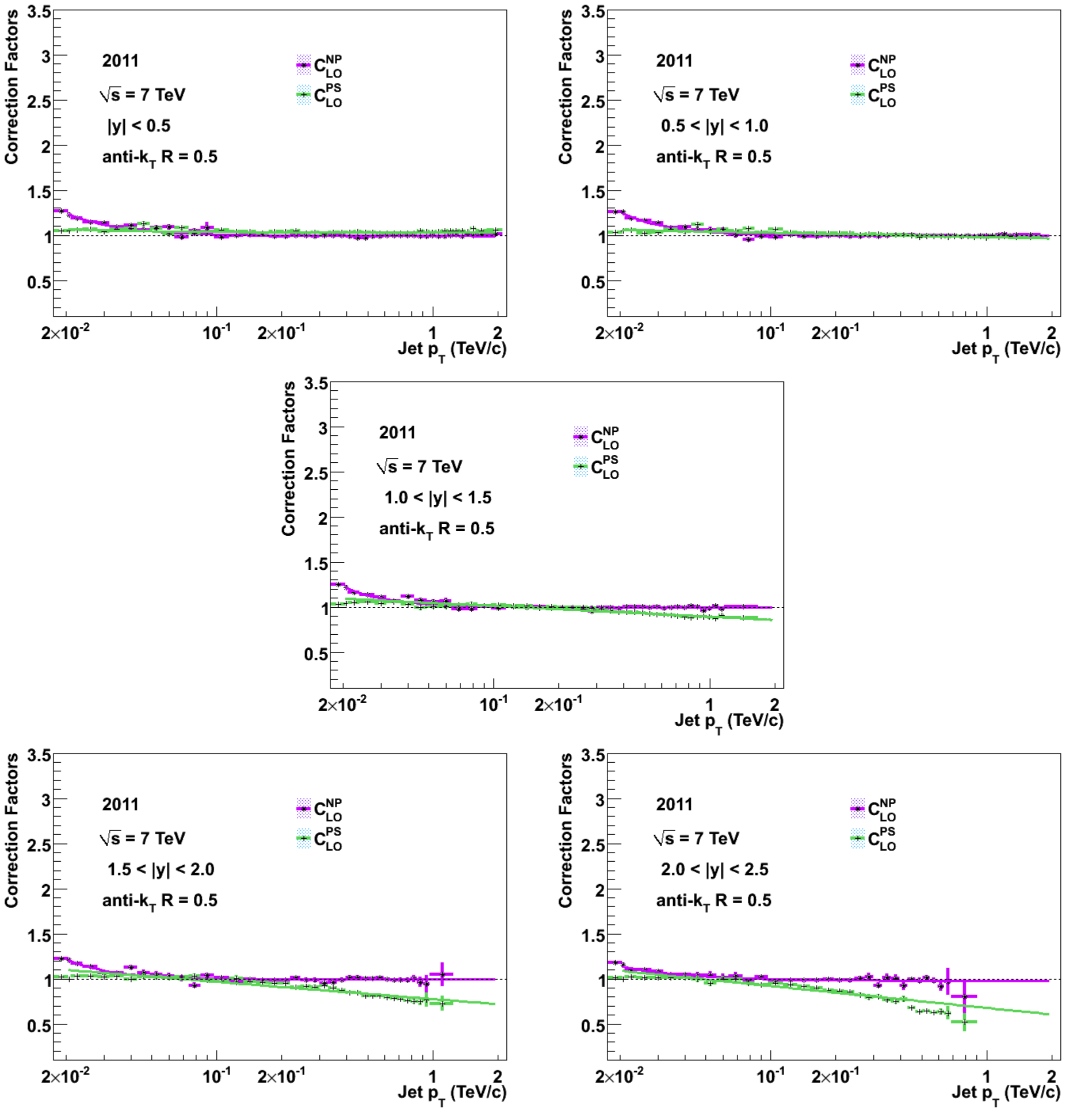


Figure C.1: Nonperturbative and perturbative correction factors measured for inclusive jet distributions as a function of jet p_T in the range 19-2000 GeV and in five bins of pseudorapidity, obtained with the Pythia*6 event generator. Bins in η go from 0 up to 2.5 in steps of 0.5. Jets are reconstructed with the anti-k_T algorithm with $R = 0.5$. Nonperturbative correction factors are obtained from the ratio between jet distributions obtained with the nominal* Pythia*6 Z2 sample and the* Pythia*6 Z2 sample where MPI and hadronization are switched off. Perturbative correction factors are obtained from the ratio between jet distributions obtained with the* Pythia*6 Z2 sample where MPI and hadronization are switched off and the* Pythia*6 Z2 sample where MPI, hadronization and PS are switched off.*

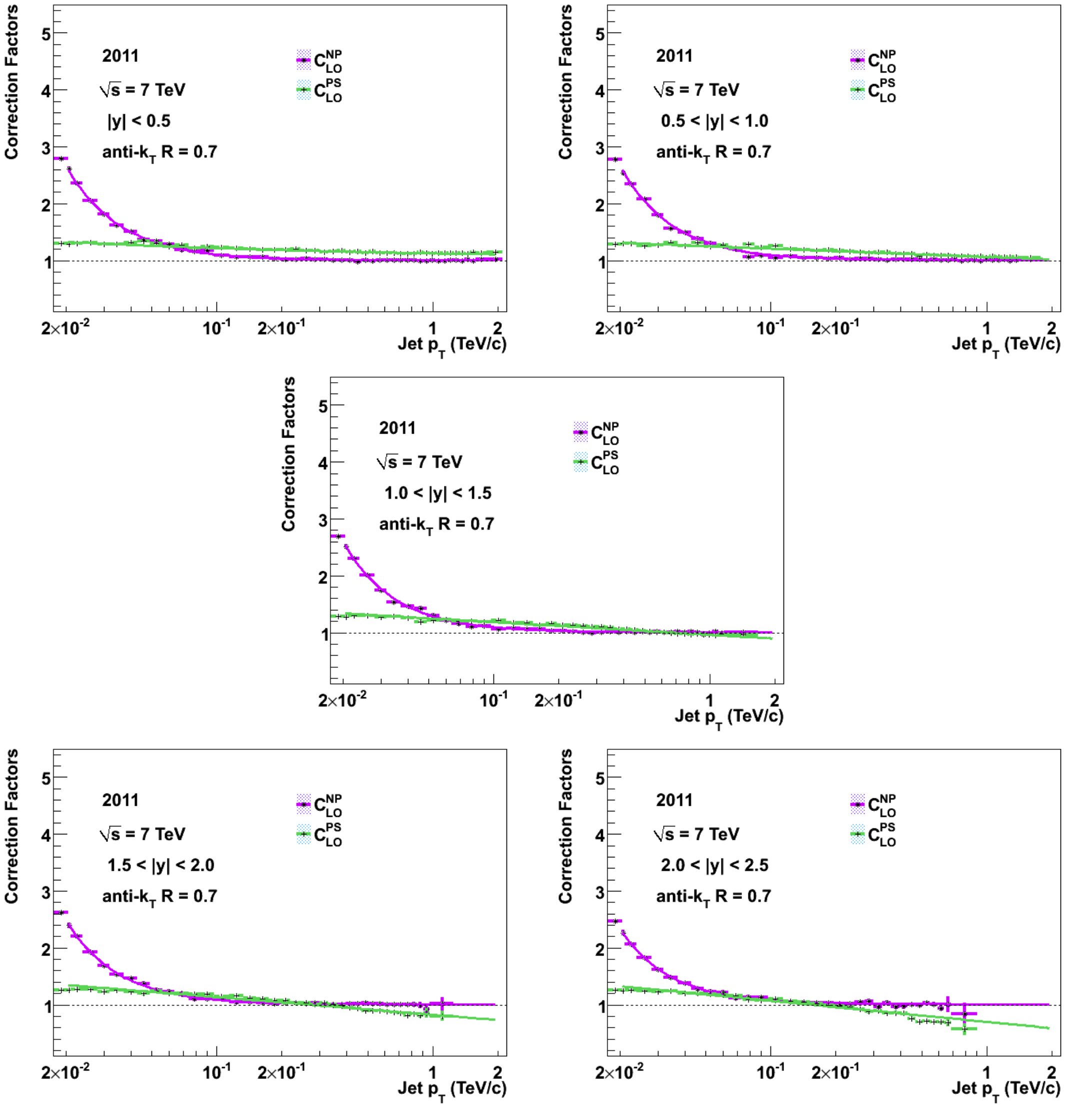


Figure C.2: Nonperturbative and perturbative correction factors measured for inclusive jet distributions as a function of jet p_T in the range 19-2000 GeV and in five bins of pseudorapidity, obtained with the PYTHIA*6 event generator. Bins in η go from 0 up to 2.5 in steps of 0.5. Jets are reconstructed with the anti-k_T algorithm with R = 0.7. Nonperturbative correction factors are obtained from the ratio between jet distributions obtained with the nominal* PYTHIA*6 Z2 sample and the* PYTHIA*6 Z2 sample where MPI and hadronization are switched off. Perturbative correction factors are obtained from the ratio between jet distributions obtained with the* PYTHIA*6 Z2 sample where MPI and hadronization are switched off and the* PYTHIA*6 Z2 sample where MPI, hadronization and PS are switched off.*

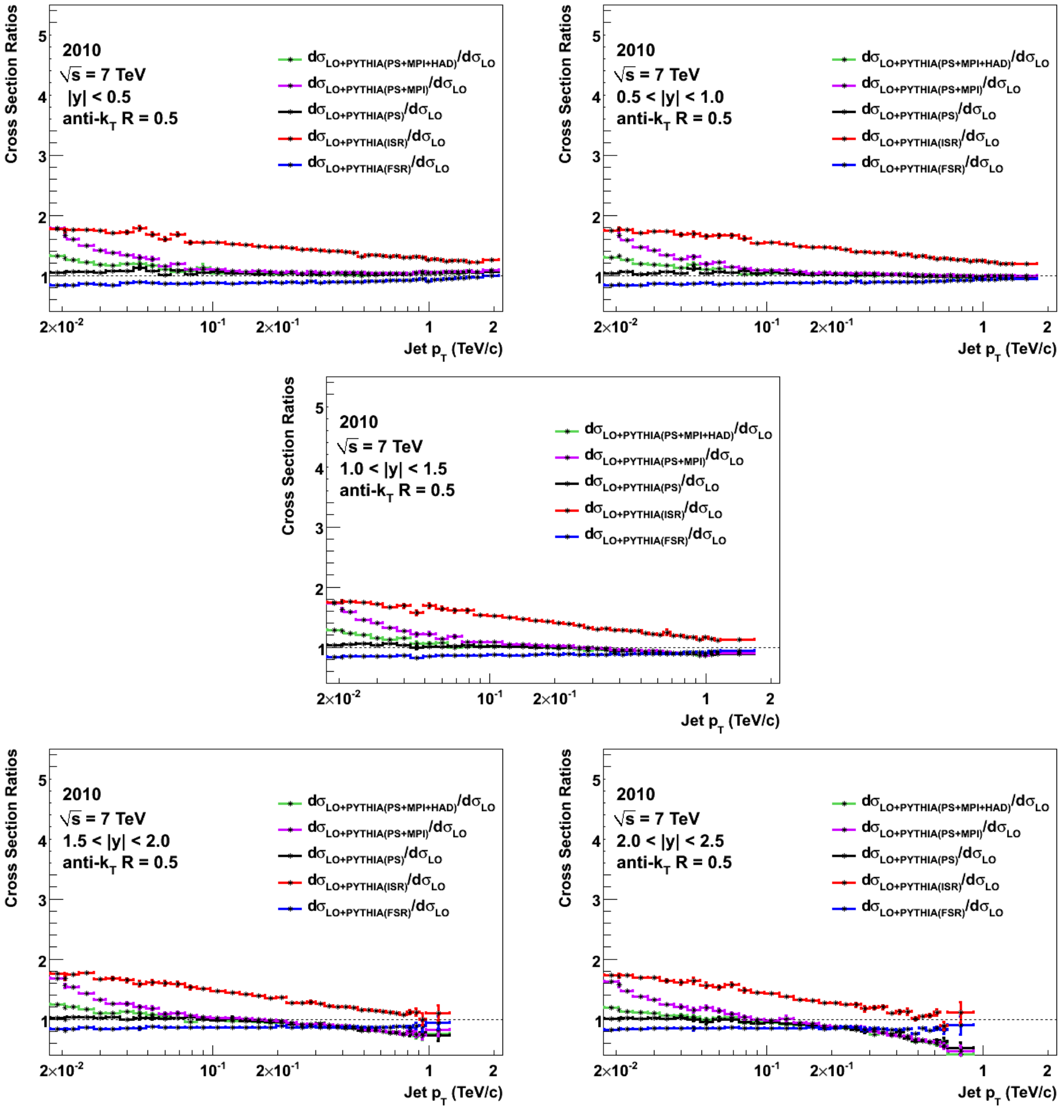


Figure C.3: Correction factors representing the contribution of the different parts of the UE: ISR, FSR, PS, MPI and hadronization. They are measured for inclusive jet distributions as a function of jet p_T in the range 19-2000 GeV and in five bins of pseudorapidity, obtained with the Pythia*6 event generator. Bins in η go from 0 up to 2.5 in steps of 0.5. Jets are reconstructed with the anti-k_T algorithm with $R = 0.5$. The correction factors represent the ratios of the different samples with different pieces of the UE simulation switched off and the sample where they are all switched off.*

In Figure C.4, the same curves are shown for $\Delta R = 0.7$. Very similar conclusions are drawn for all the effects and the different η ranges. A higher contribution of MPI and hadronization is present for $p_T < 150$ GeV, where the correction factors go up to 3-3.5.

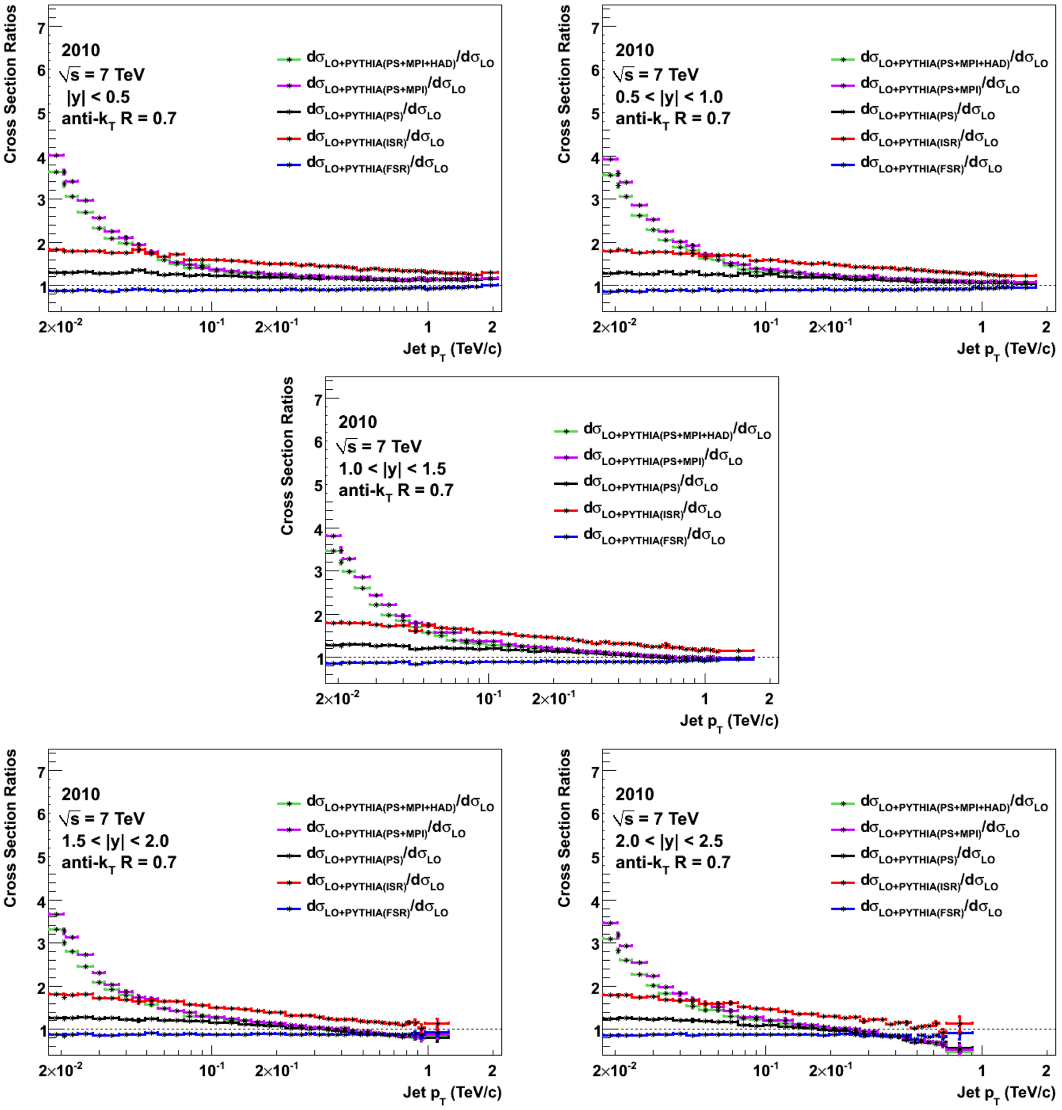


Figure C.4: Correction factors representing the contribution of the different parts of the UE: ISR, FSR, PS, MPI and hadronization. They are measured for inclusive jet distributions as a function of jet p_T in the range 19-2000 GeV and in five bins of pseudorapidity, obtained with the Pythia6 *event generator. Bins in η go from 0 up to 2.5 in steps of 0.5. Jets are reconstructed with the anti-k_T algorithm with R = 0.7. The correction factors represent the ratios of the different samples with different pieces of the UE simulation switched off and the sample where they are all switched off.*

The same study has been performed by using the Herwig++ generator to evaluate effects of different MPI and hadronization models and at a center-of-mass energy of 8 TeV. The trends of the curves are similar to the ones obtained for Pythia6; slightly different correction factors are obtained but they differ at most 20-30% from the ones described in this section. The results are extensively described in [251]. The new approach has been used in CMS for the first time on the measurement of inclusive differential jet distributions at 8 TeV [252].

PHYSICAL REVIEW D **87,** 094009 (2013)

Longitudinal momentum shifts, showering, and nonperturbative corrections in matched next-to-leading-order shower event generators

S. Dooling,[1] P. Gunnellini,[1] F. Hautmann,[2,3] and H. Jung[1,4]
[1]*Deutsches Elektronen Synchrotron, D-22603 Hamburg, Germany*
[2]*Theoretical Physics Department, University of Oxford, Oxford OX1 3NP, United Kingdom*
[3]*Physics & Astronomy, University of Sussex, Brighton BN1 9QH, United Kingdom*
[4]*Elementaire Deeltjes Fysica, Universiteit Antwerpen, B 2020 Antwerpen, Belgium*
(Received 27 December 2012; published 10 May 2013)

Comparisons of experimental data with theoretical predictions for collider processes containing hadronic jets rely on shower Monte Carlo event generators to include corrections to perturbative calculations from hadronization, parton showering, and multiple parton collisions. We examine current treatments of these corrections and propose alternative methods to take into account nonperturbative effects and parton showering in the context of next-to-leading-order event generators. We point out sizable parton-showering corrections to jet transverse energy spectra at high rapidity and discuss kinematic shifts in longitudinal momentum distributions from initial state showering in the case both of jet production and of heavy mass production at the Large Hadron Collider.

DOI: 10.1103/PhysRevD.87.094009 PACS numbers: 12.38.−t

I. INTRODUCTION

Phenomenological analyses of collider processes involving the production of hadronic jets rely on event simulation by parton shower Monte Carlo generators [1,2]. The subject of this paper concerns two different, common uses of shower Monte Carlo generators: one in which they are combined with hard scattering matrix elements via a matching scheme, e.g., at the next-to-leading order (NLO) [3,4] in perturbative QCD, and another in which they are used to obtain corrections to perturbative calculations due to hadronization, showering, and multiple parton interactions (see e.g., [5,6]), with such correction factors then being applied to determine realistic predictions, which can be compared with experimental data.

We begin in Sec. II by considering methods to evaluate the nonperturbative (NP) corrections to jet cross sections using shower event generators. We also estimate the corrections that arise from the initial state and final state parton showers and observe that they are sizeable (beyond NLO) in jet transverse energy spectra over the full range of rapidity. We propose a decomposition of the corrections to be applied to fixed NLO calculations, consisting of a truly NP contribution supplemented with a contribution coming from all order resummation via parton showers.

Next, in Sec. III we investigate kinematic aspects of parton showers associated with combining the approximation of collinear, on-shell partons with energy-momentum conservation. The main effect is an event-by-event shift in longitudinal momentum distributions whose size depends on the observable and on the phase space region, and increases with increasing rapidities. We illustrate this by numerical Monte Carlo results in different phase space regions for four specific examples of jet, heavy-quark, electroweak gauge boson, and Higgs boson production. First results on kinematic shifts have been presented in [7].

The approach of this work may be helpful to analyze corrections to finite-order perturbative calculations for jet observables from parton-showering and nonperturbative dynamics. These encompass both final state fragmentation effects and initial state contributions associated with collinearity approximations. Dynamical high-energy effects on jet final states, distinct from the ones discussed in this paper, have been emphasized in [8–10] due to noncollinear contributions to parton branching processes. We note that both these results and the results in this paper stress the phenomenological relevance of more complete descriptions of QCD parton cascades in terms of transverse momentum dependent parton fragmentation and parton density functions [11–14]. Concluding comments on the results of this work are given in Sec. IV.

II. MONTE CARLO NONPERTURBATIVE CORRECTION FACTORS

In this section we consider methods to evaluate NP and parton shower correction factors. To be definite, we refer to the case of inclusive production of single jets at the LHC [15]. In order to compare theory with experimental data corrected to stable particle level, Refs. [5,6] supplement NLO perturbative calculations with NP corrections estimated from Monte Carlo event generators. Using leading-order Monte Carlo (LO-MC) generators [1,2], the correction factors K_0 are schematically obtained by [5,6]

$$K_0^{\rm NP} = N_{\text{LO-MC}}^{(\rm ps+mpi+had)} / N_{\text{LO-MC}}^{(\rm ps)}, \qquad (1)$$

where (ps + mpi + had) and (ps) mean, respectively, a simulation including parton showers, multiparton interactions, and hadronization, and a simulation including

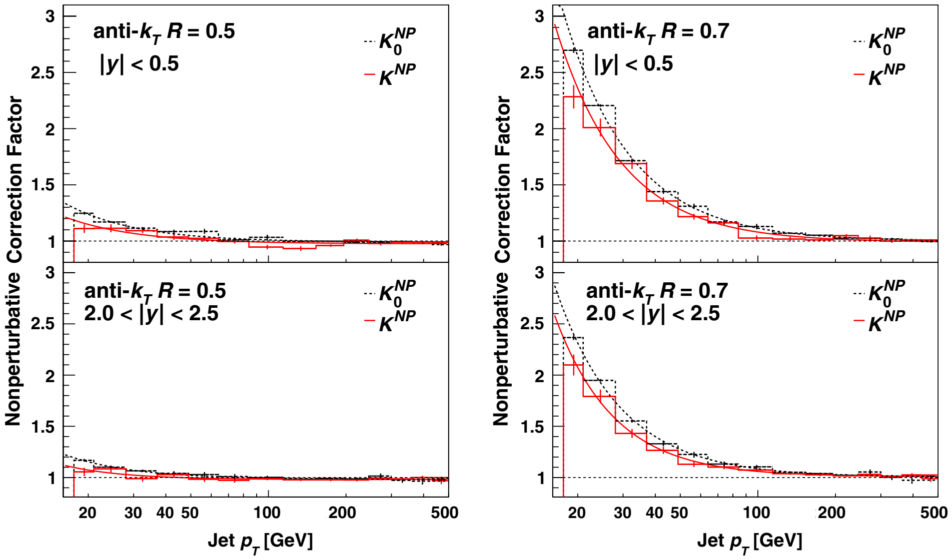


FIG. 1 (color online). The NP correction factors to jet transverse momentum distributions obtained from Eqs. (1) and (2), using PYTHIA and POWHEG respectively, for $|y| < 0.5$ and $2 < |y| < 2.5$. Left: $R = 0.5$. Right: $R = 0.7$.

only parton showers in addition to the LO hard process. Having only LO + PS event generators available, this is the most obvious way to estimate NP corrections to be applied to NLO parton-level calculations. However, when these corrections are combined with NLO parton-level results, a potential inconsistency arises because the radiative correction from the first gluon emission is treated at different levels of accuracy in the two parts of the calculation.

We here suggest that an alternative method that avoids this is to use NLO Monte Carlo (NLO-MC) generators to determine the correction. In this case one can consistently assign correction factors to be applied to NLO calculations. Moreover, this method allows one to study separately

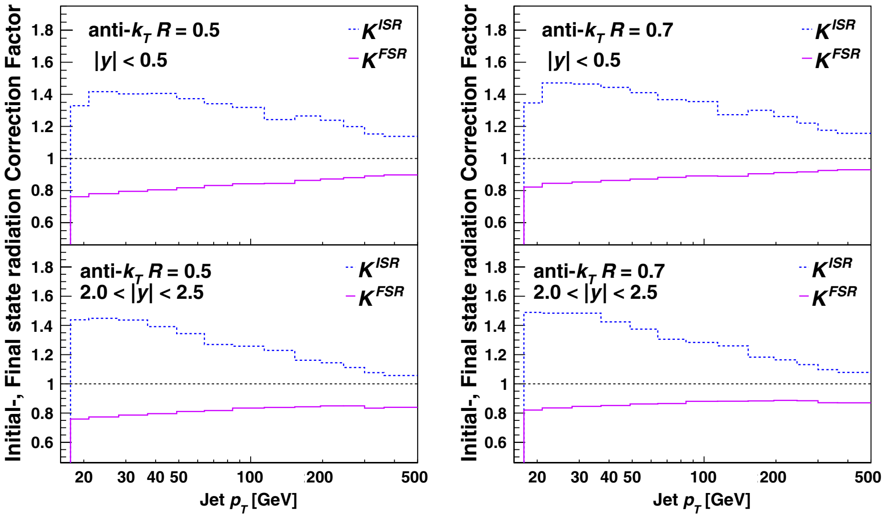


FIG. 2 (color online). The initial and final state parton shower correction factor to jet transverse momentum distributions, obtained from Eq. (3) using POWHEG for $|y| < 0.5$ and $2 < |y| < 2.5$. Left: $R = 0.5$. Right: $R = 0.7$.

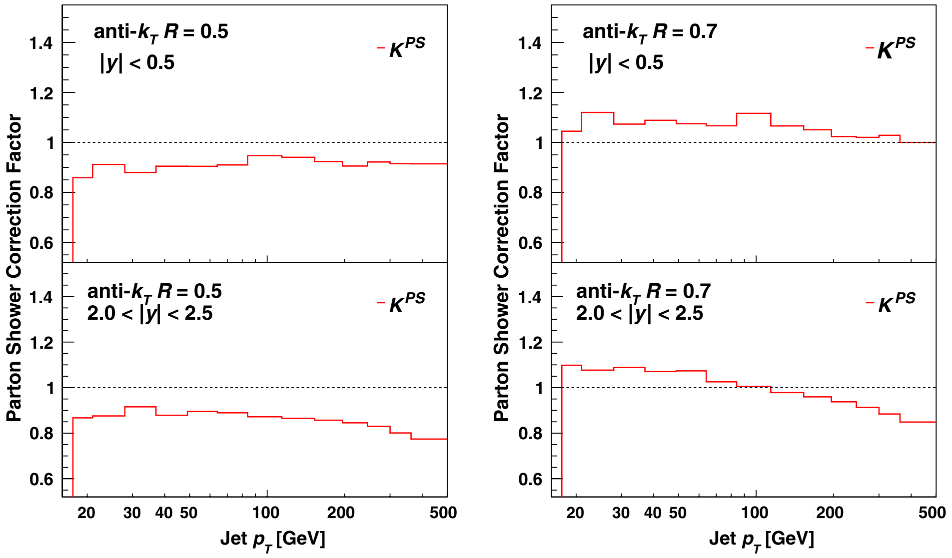


FIG. 3 (color online). The parton shower correction factor to jet transverse momentum distributions, obtained from Eq. (3) using POWHEG for $|y| < 0.5$ and $2 < |y| < 2.5$. Left: $R = 0.5$. Right: $R = 0.7$.

correction factors to the fixed-order calculation due to parton-showering effects. To this end, we introduce the correction factors K^{NP} and K^{PS} as

$$K^{\mathrm{NP}} = N^{(\mathrm{ps+mpi+had})}_{\mathrm{NLO\text{-}MC}} / N^{(\mathrm{ps})}_{\mathrm{NLO\text{-}MC}}, \tag{2}$$

$$K^{\mathrm{PS}} = N^{(\mathrm{ps})}_{\mathrm{NLO\text{-}MC}} / N^{(0)}_{\mathrm{NLO\text{-}MC}}, \tag{3}$$

where the denominator in Eq. (3) is defined by switching off all components beyond NLO in the Monte Carlo simulation. The difference between the correction factors in Eqs. (1) and (2) comes primarily from the way in which the multiple parton interaction (MPI) contribution is matched to the NLO calculation. MPI processes have typical transverse momentum scales smaller than the scale of the hard process, which may be defined as the average transverse momentum of the hard partons. This, however, is different in LO and NLO calculations, giving rise to non-negligible numerical differences, which we will show below. The correction factor in Eq. (3), on the other hand, is new. It singles out contributions due to parton showering. This correction factor has not been considered in earlier analyses. We show below its numerical significance. We anticipate that taking properly into account these showering corrections can be relevant in fits for parton distribution functions (pdfs) using inclusive jet data.

In Fig. 1 we compute results for the NP correction factors in Eqs. (1) and (2) to jet transverse momentum distributions. We define jets using the anti-k_T algorithm [16] with jet size $R = 0.5$ and $R = 0.7$. We plot the results versus the jet transverse momentum p_T for different regions in the jet rapidity y. We show K^{NP} as obtained using the NLO event generator POWHEG [17] and compare it to the result obtained at leading order from PYTHIA [2] (tune Z2 [18] and CTEQ6L1 pdfs [19]). The curves in Fig. 1 illustrate the differences coming from the definition of the hard process.

In Figs. 2 and 3 we compute the corrections from parton shower K^{PS} as obtained from Eq. (3) as a function of the jet p_T for different values of R and different rapidities y. Figure 2 shows the contributions coming from initial state and final state parton showers separately. We note that the initial and final state showers are so interconnected that the combined effect is nontrivial and cannot be obtained by simply adding the two results. In general the effect from parton shower is largest at large $|y|$, where the initial state

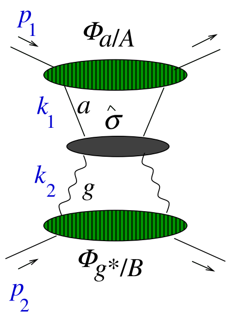


FIG. 4 (color online). Factorized structure of the jet cross section at high rapidity.

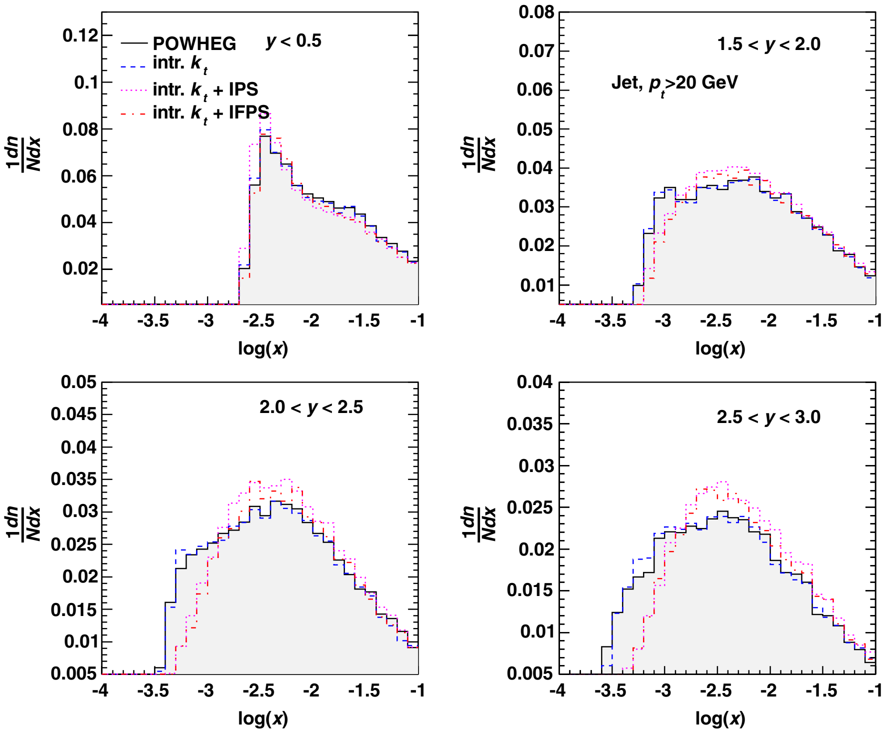


FIG. 5 (color online). Distributions in the parton longitudinal momentum fraction x before (POWHEG) and after parton showering (POWHEG + PS), for inclusive jet production at different rapidities for jets with $p_T > 18$ GeV obtained by the anti-k_t jet algorithm [16] with $R = 0.5$. Shown is the effect of intrinsic k_t, initial (IPS) and initial+final state (IFPS) parton shower.

parton shower is mainly contributing at low p_T, while the final state parton shower is contributing significantly over the whole p_T range. In particular, note in Fig. 3 that, while at central rapidity the combined shower correction is rather flat in p_T, at higher rapidity this is no longer flat and for large p_T it may even dip below the correction from the purely final state shower reported in Fig. 2. This suggests that migration effects become relevant not only in p_T but also in y.

While the NP corrections studied in Fig. 1 become vanishingly small at sufficiently large p_T, the showering correction in Figs. 2 and 3 gives finite effects also for large p_T. Since, as shown by our results, the size of this effect does depend on the value of rapidity y, this will influence the shape of jet distributions and the comparisons of theory predictions with experimental data. In particular, if the showering correction factor is not consistently taken into account, besides the NP corrections, this may affect the determination of parton distribution functions from data sets including jets.

Note that in [5,6] NP correction factors K_0 are applied to the NLO calculation [20], and the data comparison shows that the NLO calculation agrees with data at central rapidities, while increasing deviations are seen with increasing rapidity at large transverse momentum p_T [5]. A second comparison is performed in [5] with NLO-matched POWHEG calculations [17], showing large differences in the high rapidity region between results obtained by interfacing POWHEG with different shower models [1,2] and different model tunes [18,21].[1] Motivated by this observation, in the next section we consider more closely the kinematics of the initial state parton shower at high rapidity.

III. INITIAL STATE SHOWERING AND KINEMATIC SHIFTS

Let us recall the physical picture [10] of jet production at high rapidity (Fig. 4) based on QCD high-energy factorization [23]. Take the incoming momenta p_1 and p_2 in Fig. 4 in the plus and minus lightcone directions, defined, for any four-vector v^μ, as $v^\pm = (v^0 \pm v^3)/\sqrt{2}$. Let us parametrize the exchanged momenta k_1 and k_2 in terms of purely transverse four-vectors $k_{\perp 1}$ and $k_{\perp 2}$ and longitudinal (light cone) momentum fractions x_i (collinear) and $\bar{x}_i$

---

[1] Further discussion of parton showering effects on high-rapidity jets may be found in [22].

(anticollinear) as $k_1 = x_1 p_1 + k_{\perp 1} + \bar{x}_1 p_2$ and $k_2 = x_2 p_2 + k_{\perp 2} + \bar{x}_2 p_1$. To single-logarithmic accuracy in the jet rapidity and the jet transverse momentum, we may approximate k_1 and k_2 using strong ordering in the longitudinal momenta and get [10]

$$k_1 \simeq x_1 p_1, \qquad k_2 \simeq x_2 p_2 + k_{\perp 2}, \qquad x_1 \gg x_2. \quad (4)$$

The physical picture corresponding to the factorization [10,23] consists of the scattering of a highly off-shell, low-x parton off a nearly on-shell, high-x parton. The calculations [10,22] embody this picture through the longitudinal and transverse momentum dependences of both perturbative and nonperturbative components of the jet cross section, denoted, respectively, by $\hat{\sigma}$ and Φ in Fig. 4. In what follows, however, we will not use the specific content of these calculations, but we will simply use the underlying physical picture as a guidance to examine kinematic effects of collinear approximations.

In the light of this picture, let us consider the NLO-matched shower Monte Carlo calculations, following [7]. In the Monte Carlo event generator first the hard subprocess events with full four-momentum assignments for the external lines are generated. In particular, the momenta $k_j^{(0)}$ ($j = 1, 2$) of the partons initiating the hard scatter are on shell, and are taken to be fully collinear with the incoming state momenta p_j,

$$k_j^{(0)} = x_j p_j \quad (j = 1, 2). \quad (5)$$

Next the showering algorithm is applied, and complete final states are generated including additional QCD radiation from the initial state and final state parton cascades. As a result of QCD showering, the momenta k_j are no longer exactly collinear,

$$k_j \neq x_j p_j \quad (j = 1, 2). \quad (6)$$

Their transverse momentum is to be compensated by a change in the kinematics of the hard scattering subprocess. By energy-momentum conservation, however, this implies a reshuffling, event by event, in the longitudinal momentum fractions x_j of the partons scattering off each other in the hard subprocess. The size of the shift in x_j depends on the emitted transverse momenta.

Let us now focus on jets measured in the rapidity range $y < 2.5$ [6] and examine the effect of the kinematical shift in the longitudinal momentum fractions. To this end we

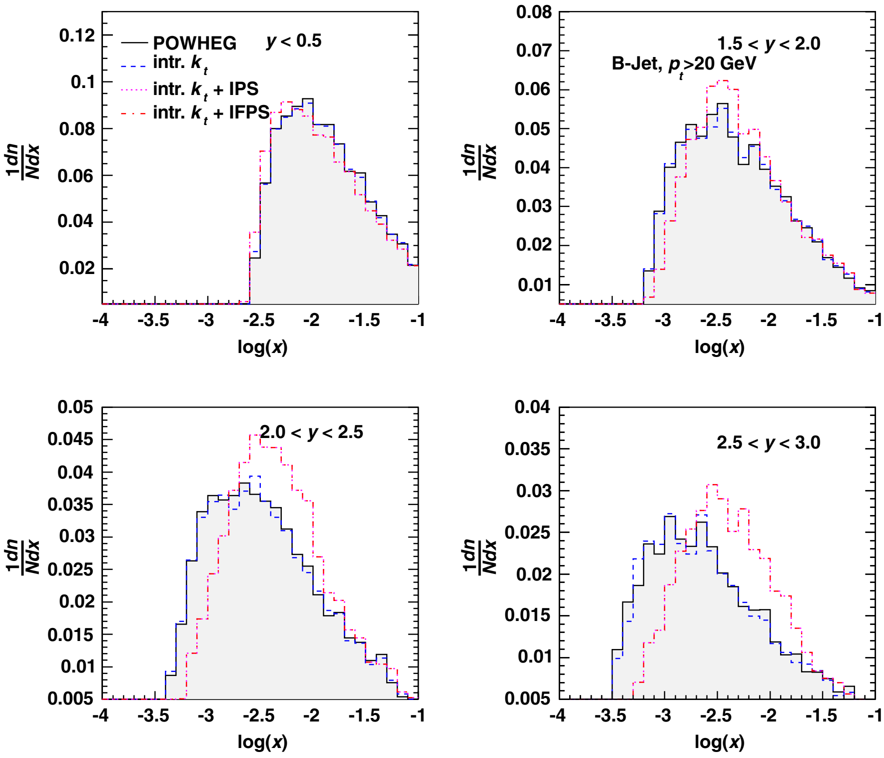


FIG. 6 (color online). Production of b-jets: distribution in the parton longitudinal momentum fraction x, before and after parton showering, for different rapidity regions. Shown is the effect of intrinsic k_t, IPS and IFPS parton shower.

compute the distribution in x_j from POWHEG before parton showering and after parton showering [7]. Figure 5 shows the distribution for one of the x_j partons. We plot the result before showering (POWHEG) and the results of successively including intrinsic k_t, initial state parton shower, and initial + final state parton showers. The results are obtained using the PYTHIA parton shower (tune Z2 [18] and CTEQ6L1 pdfs [19]). This does not include multiple parton interaction and hadronization effects. Using the definition of light cone momentum fractions given at the beginning of this section, the kinematic variable x is computed as $x = (E + p^z)/(2E_{\mathrm{beam}})$, where E and p^z are the energy and z component of momentum of parton j, and E_{beam} is the energy of the hadron beam. The momentum fraction x is first calculated for the partons given by POWHEG before shower and then calculated from the PYTHIA event record after shower.

We see from Fig. 5 that the kinematical reshuffling in the longitudinal momentum fraction is negligible for central rapidities but becomes significant for $y > 1.5$. This effect characterizes the highly asymmetric parton kinematics, which becomes important for the first time at the LHC in significant regions of phase space [10]. Since the perturbative weight for each event is determined by the initial POWHEG simulation, predictions of matched NLO-shower calculations for observables sensitive to this asymmetric region can be affected significantly by the kinematical shift as shown in Fig. 5. Similarly, since the momentum reshuffling is done after the evaluation of the parton distribution functions, the kinematical shift can affect predictions also through the pdfs. It will be of interest to examine the impact of this phase space region on total cross sections as well.

Let us next consider the case of bottom-flavor jet production [24,25]. The LHC measurements [24,25] are reasonably described by NLO-matched shower generators MC@NLO [26] and POWHEG [27] at central rapidities, and they are below these predictions at large rapidity and large p_T. In Fig. 6 we consider B-jets in different rapidity regions [24] and plot the gluon x distribution from POWHEG before parton showering and after including various components of the parton shower generator, similarly to what is done above for Fig. 5. We use the PYTHIA parton shower (tune Z2 [18], here including hadronization to identify the B-jet). We observe a similar shift in longitudinal momentum with increasing rapidity as in the inclusive jet case.

In Fig. 7 we consider Drell-Yan (DY) production in the mass range $16 < m_{\mathrm{DY}} < 166$ GeV and perform a similar study to what is done above for jets. In this case too we find that the effects of the kinematical reshuffling in x evaluated from POWHEG become non-negligible away from the

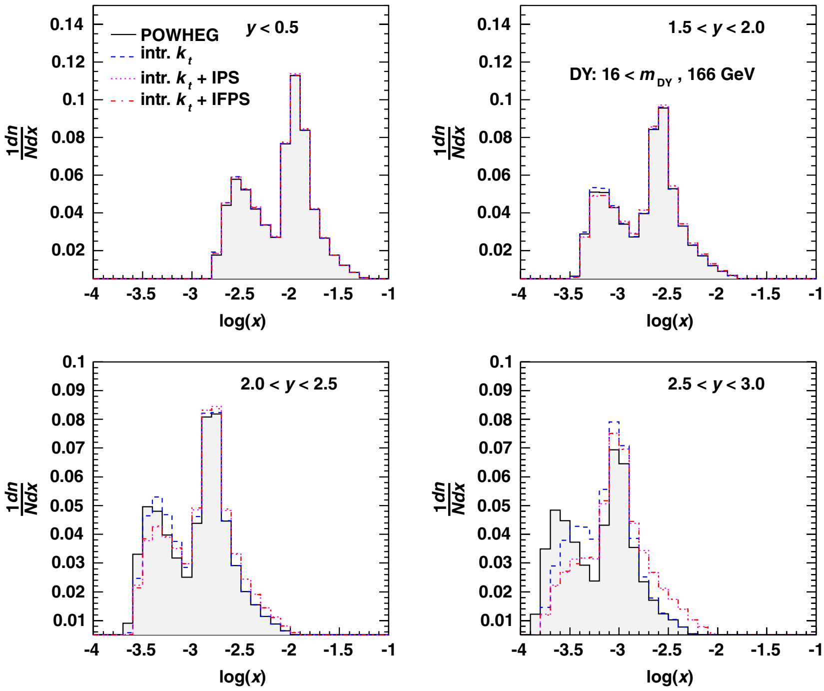


FIG. 7 (color online). Drell-Yan production with $16 < m_{\mathrm{DY}} < 166$ GeV: distribution in the parton longitudinal momentum fraction x before and after showering. Shown is the effect of intrinsic k_t, IPS and IFPS parton shower.

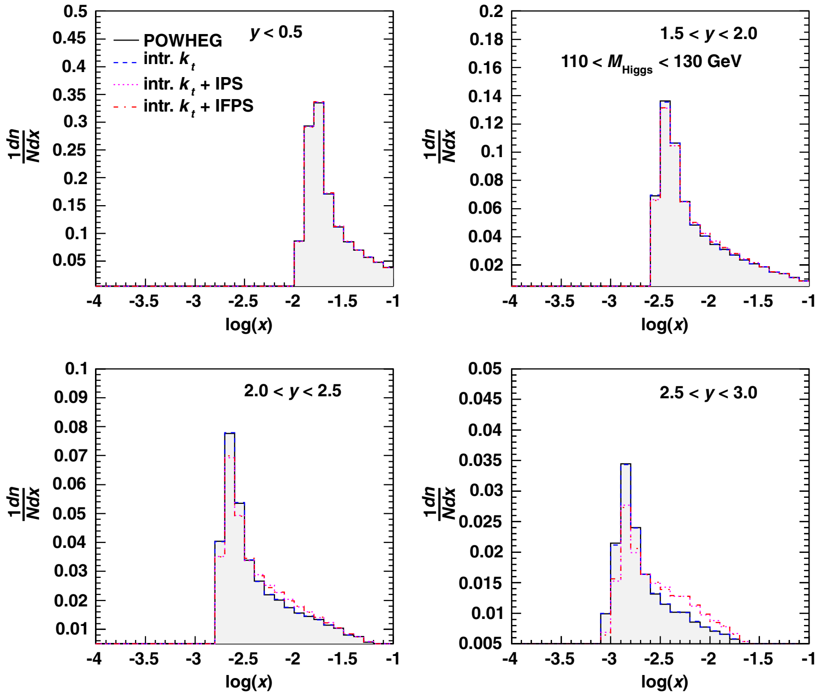


FIG. 8 (color online). Higgs boson production with $110 < m_{\mathrm{Higgs}} < 130$ GeV: distribution in the parton longitudinal momentum fraction x before and after showering. Shown is the effect of intrinsic k_t, IPS and IFPS parton shower.

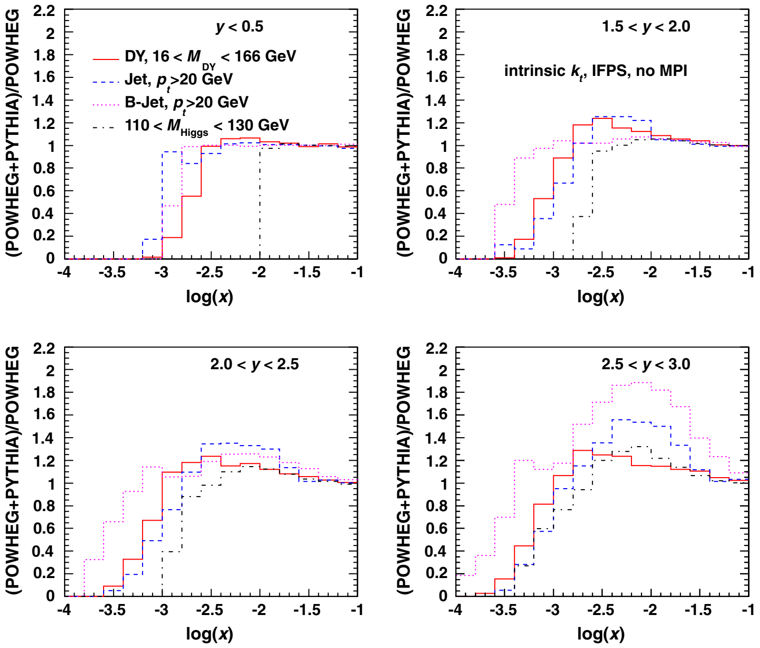


FIG. 9 (color online). Ratio of the cross sections obtained with POWHEG after and before inclusion of initial + final state parton shower and intrinsic k_t for the different processes.

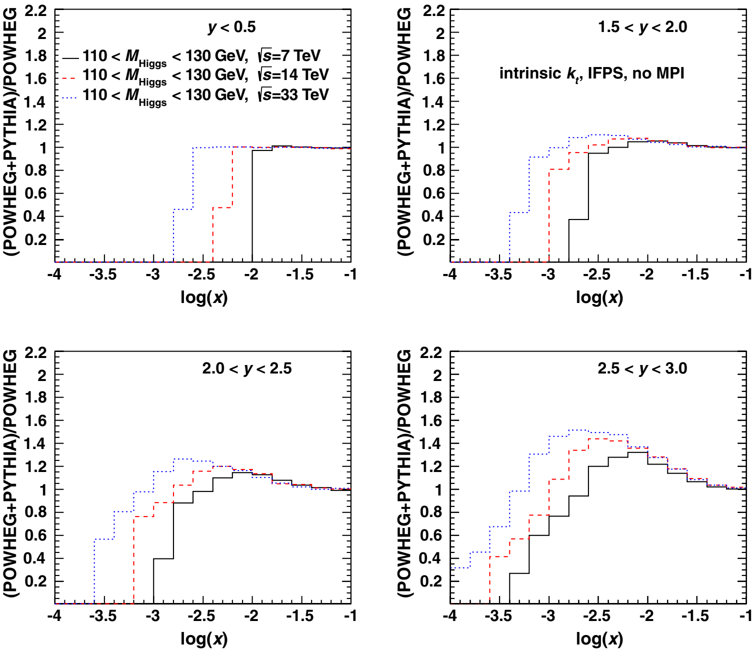


FIG. 10 (color online). Ratio of the cross sections obtained with POWHEG after and before inclusion of initial + final state parton shower and intrinsic k_t for Higgs production at different energies: $\sqrt{s} = 7$, 14, 33 TeV.

central rapidity region. The double peak structure in Fig. 7 comes from the continuum DY production in addition to Z_0 production. It will be of interest to investigate the kinematic reshuffling effect along with the forward Drell-Yan enhancements discussed in [28].

Finally we consider Higgs boson production in Fig. 8 for $110 < m_{\text{Higgs}} < 130$ GeV. We observe a smaller effect at $\sqrt{s} = 7$ GeV than in the previous cases since the x range is limited by the Higgs mass.

Figure 9 summarizes the results in Figs. 5–8 for the ratio of the cross section obtained by POWHEG after inclusion of parton showering to the cross section before parton showering, plotted for different processes. In Fig. 10 we plot this ratio for Higgs boson production at different $\sqrt{s}$ energies of 7, 14, and 33 GeV.

The longitudinal momentum shifts from parton showering computed in this section measure effects from QCD radiation beyond perturbative fixed-order calculations and provide a significant contribution to the correction factors in Sec. II. They affect initial state showers and need to be consistently taken into account in calculations that are used to determine parton density functions. The origin of the kinematical shifts lies with the approximation of collinearity [7] on the partonic states to which the branching algorithms describing showers are applied. Although for explicit calculations we have used a particular NLO-shower matching scheme (POWHEG), the effect is common to any calculation matching NLO with collinear showers. In calculations using integrated parton density functions the correction factors studied in this paper have to be applied after the evaluation of the cross section (and, as remarked on earlier, this may induce systematic inconsistencies if these corrections are not taken into account properly). On the other hand, this is avoided in approaches using transverse momentum dependent pdfs [11–14,28] from the beginning (transverse momentum dependent pdfs or unintegrated pdfs), as is done for example in the CASCADE event generator [29].

IV. CONCLUSIONS

Theoretical predictions for high-energy collider processes containing hadronic jets require supplementing finite-order perturbative calculations with parton showering and nonperturbative corrections. In this paper we have studied methods to treat parton-showering and nonperturbative corrections in the context of matched NLO-shower event generators.

We have pointed out potential inconsistencies in current approaches that on the one hand apply NP correction factors from leading-order Monte Carlo generators to NLO parton-level predictions and on the other hand fail to include showering corrections. We have proposed methods to address these deficiencies by using consistently available NLO Monte Carlo tools. We have shown that the differences in the predictions for jet cross sections induced by the modified approach we propose are significant in regions of phase space that are explored with hard probes for the first time at the LHC. In particular, the nonperturbative correction factor K^{NP} introduced in Sec. II gives non-negligible differences at low to intermediate jet p_T, and the showering correction factor K^{PS} of Sec. II gives significant effects over the whole p_T range and is largest at large jet rapidities y.

Because of this y and p_T dependence, taking properly into account NP and showering correction factors changes the shape of jet distributions and affects significantly the comparison of theory predictions with experimental data. The numerical results we have presented show effects as large as 50% in regions of y and p_T phase space relevant to jet measurements at the LHC. The showering correction factor K^{PS}, in particular, can affect the determination of parton distribution functions from fits to experimental data sets comprising inclusive jet measurements.

We have investigated in closer detail the sources of the showering correction from initial state and final state effects. We have observed that the main initial state showering effect comes from kinematical shifts in longitudinal momentum distributions [7] due to combining collinearity approximations with the Monte Carlo implementation of energy-momentum conservation constraints. We have examined the longitudinal shifts for specific processes in Sec. III. This effect is largest for inclusive jets and b-flavor jets at the LHC in the higher rapidity bins. We have extended the study of longitudinal shifts [7] to the case of Drell-Yan pair production by analyzing the Drell-Yan mass region $16 < m_{\mathrm{DY}} < 166$ GeV and found that the shifts are non-negligible for Drell-Yan production at forward rapidities $y \geq 2$. We have also examined the case of Higgs boson production for $110 < m_{\mathrm{Higgs}} < 130$ GeV and found that the shifts are non-negligible at large rapidities at $\sqrt{s} = 7$ GeV and become more and more important at higher center-of-mass energies.

It will be interesting to study the impact of the effects discussed in this work on phenomenological analyses of LHC final states involving hadronic jets. We expect these effects to also influence determinations of parton distributions. Longitudinal momentum shifts can be avoided in formulations that keep track of noncollinear (i.e., transverse and/or anticollinear) momentum components from the beginning using unintegrated initial state distributions [12,13], also at parton shower level [29,30]. It will be interesting to investigate to what extent this can be exploited to construct approaches in which nonperturbative contributions such as multiple parton interactions, finite transverse momenta, and hadronization are consistently incorporated into parton branching event generators.

ACKNOWLEDGMENTS

We are grateful to Torbjörn Sjöstrand for many discussions concerning hadronization corrections and multiparton interactions in PYTHIA. We are also grateful for many discussions and clarifications on POWHEG to Simone Alioli.

---

[1] G. Corcella, I. G. Knowles, G. Marchesini, S. Moretti, K. Odagiri, P. Richardson, M. H. Seymour, and B. R. Webber, J. High Energy Phys. 01 (2001) 010; G. Corcella *et al.*, arXiv:hep-ph/0210213.

[2] T. Sjöstrand, S. Mrenna, and P. Skands, J. High Energy Phys. 05 (2006) 026.

[3] P. Nason and B. R. Webber, Annu. Rev. Nucl. Part. Sci. **62**, 187 (2012).

[4] S. Höche, SLAC Report No. SLAC-PUB-14498, 2011; S. Höche and M. Schönherr, Phys. Rev. D **86**, 094042 (2012).

[5] G. Aad *et al.* (ATLAS Collaboration), Phys. Rev. D **86**, 014022 (2012).

[6] S. Chatrchyan *et al.* (CMS Collaboration), Phys. Rev. Lett. **107**, 132001 (2011).

[7] F. Hautmann and H. Jung, Eur. Phys. J. C **72**, 2254 (2012).

[8] G. Marchesini and B. R. Webber, Nucl. Phys. **B386**, 215 (1992).

[9] F. Hautmann and H. Jung, J. High Energy Phys. 10 (2008) 113.

[10] M. Deak, F. Hautmann, H. Jung, and K. Kutak, J. High Energy Phys. 09 (2009) 121; arXiv:0908.1870; Eur. Phys. J. C **72**, 1982 (2012).

[11] J. C. Collins, *Foundations of Perturbative QCD* (Cambridge University Press, Cambridge, England, 2011), p. 624.

[12] E. Avsar, arXiv:1203.1916; Int. J. Mod. Phys. Conf. Ser. **04**, 74 (2011).

[13] F. Hautmann, Acta Phys. Pol. B **40**, 2139 (2009); Phys. Lett. B **655**, 26 (2007); F. Hautmann and H. Jung, Nucl. Phys. B, Proc. Suppl. **184**, 64 (2008); AIP Conf. Proc. **1056**, 79 (2008).

[14] P. J. Mulders, Pramana **72**, 83 (2009); P. J. Mulders and T. C. Rogers, arXiv:1102.4569.

[15] K. Rabbertz, ISMD 2012, Kielce, 2012.

[16] M. Cacciari, G. Salam, and G. Soyez, J. High Energy Phys. 04 (2008) 063.

[17] S. Alioli, K. Hamilton, P. Nason, C. Oleari, and E. Re, J. High Energy Phys. 04 (2011) 081.

[18] R. D. Field, arXiv:1010.3558.

[19] J. Pumplin, D. R. Stump, J. Huston, H.-L. Lai, P. Nadolsky, and W.-K. Tung, J. High Energy Phys. 07 (2002) 012.

[20] Z. Nagy, Phys. Rev. D **68**, 094002 (2003).

[21] P. Z. Skands, Phys. Rev. D **82**, 074018 (2010).

[22] M. Deak *et al.*, arXiv:1206.7090; arXiv:1012.6037.

[23] S. Catani, M. Ciafaloni, and F. Hautmann, Phys. Lett. B **242**, 97 (1990); Nucl. Phys. **B366**, 135 (1991); Phys. Lett. B **307**, 147 (1993); Phys. Lett. B **315**, 157 (1993); Nucl. Phys. **B427**, 475 (1994).

[24] S. Chatrchyan *et al.* (CMS Collaboration), J. High Energy Phys. 04 (2012) 084.

[25] G. Aad *et al.* (ATLAS Collaboration), Eur. Phys. J. C **71**, 1846 (2011).

[26] S. Frixione, P. Nason, and B. R. Webber, J. High Energy Phys. 08 (2003) 007.

[27] S. Frixione, P. Nason, and G. Ridolfi, J. High Energy Phys. 09 (2007) 126.

[28] F. Hautmann, M. Hentschinski, and H. Jung, Nucl. Phys. **B865**, 54 (2012); arXiv:1205.6358; arXiv:1209.6305.

[29] H. Jung *et al.*, Eur. Phys. J. C **70**, 1237 (2010).

[30] S. Jadach and M. Skrzypek, Acta Phys. Pol. B **40**, 2071 (2009); S. Jadach, M. Jezabek, A. Kusina, W. Placzek, and M. Skrzypek, Acta Phys. Pol. B **43**, 2067 (2012).

Appendix D

New Underlying Event tunes

A precise and accurate description of inclusive measurements by predictions of MC event generators is necessary for any field in particle physics. In particular, for the next LHC phase, it becomes crucial to have reliable predictions at 13–14 TeV, still unexplored energy collisions and about two times bigger than the ones tested so far. This upcoming step, starting in the year 2015, has brought up in the physics community the need for a better tune of the phenomenological part of the hadronic collisions, implemented in the MC models. This is not only related to hadronization parameters, but also to the simulation concerning MPI and colour reconnection in the UE part. The term "tune" is commonly used to identify a set of parameters, implemented in a MC event generator for a given model. A tune is generally obtained by optimising the description of a subsample of measurements, which are sensitive to the considered parameters. A complete description of the most up-to-date tunes is provided in Chapter 5; in particular, up to now in CMS, for none of the generators, including PYTHIA6 and PYTHIA8, tunes have been obtained by investigating the energy dependence of specific observables: Tune Z2* in PYTHIA6 has been derived from inclusive UE data at 0.9 and 7 TeV, while Tune 4C in PYTHIA8 has been extracted from UE observables measured mainly at 7 TeV. Recently, a new tune performed by Peter Skands et al. [231] has been released for PYTHIA8, with the NNPDF2.3QED PDF set [253], which uses a different parametrization of the gluon density at low x. For the tune both hadronization and UE parameters have been included on a broad collection of data, and an overall very good description of a wide range of observables in the central and forward region is achieved. This is the reason why it has been set as default tune for the newer PYTHIA8 versions. However, it is still not optimal in describing data at the lower collision energies.

An attempt performed by CMS to tune UE observables measured at different collision energies is described in this Chapter. Measurements at the CDF and CMS experiments have been used with the aim for a better description of the energy dependence of the UE contribution in hadronic collisions. Two event generators have been considered: PYTHIA6 which has been the reference in CMS measurements so far, and PYTHIA8. For PYTHIA8, in particular, two PDF sets are considered: CTEQ6L1 [183], which is currently used by the tune 4C, and the HERAPDF1.5LO [254], which has been recently extracted from HERA data.

In Sections D.1 and D.2 the measurements and the software used for the tune are described.

In Section D.3 the results of the tuning for the Pythia8 event generator are discussed, while in Section D.4 the predictions obtained with the new sets of parameters on other observables, not considered for the tune, are examined. In Section D.5 the predictions of the new tunes at 13 TeV are shown. The whole collection of results has been made public by the CMS collaboration [12] and the new tunes for Pythia6 and Pythia8 are starting to be used.

D.1 The measurements

The measurements that have been used to perform the new tunes include CDF [255] and CMS [232] data at four different collision energies: 300, 900 and 1960 GeV for $p\bar{p}$ collisions and 7000 GeV for pp collisions. The most interesting feature of these measurements is that they extend the usual analysis strategy described in Chapter 2, in order to better disentangle the different UE contributions. In particular, they further divide the two transverse regions into separate measurements. The transverse region with higher activity is called "TransMAX", while the one with lower activity is labelled as "TransMIN". Separate measurements are also performed for "TransDIF", which is defined as the difference between the activity in "TransMAX" and "TransMIN", and "TransAV", which is instead the average between the two. Note that only two of the four measurements are independent of each other.

The following observables are investigated: charged particle multiplicity and p_{T} sum as a function of the hardest scale in each event, namely the p_{T} of the leading charged particle. The direction of the hard object defines the direction of the hard scattering and the transverse plane is divided accordingly, as defined in Chapter 2. For these measurements, charged particles in $|\eta| < 0.8$ and with $p_{\mathrm{T}} > 0.5$ GeV have been counted. The definition of "TransMIN" and "TransMAX" occurs in every event separately for multiplicity and p_{T} sum: it may thus happen that the "TransMIN" ("TransMAX") for the charged particle multiplicity is the "TransMAX" ("TransMIN") for the p_{T} sum, depending on the possibility to have a region with low (high) multiplicity but high (low) p_{T} of the particles. This does not represent an inconsistency but rather gives an unambiguous definition of the two regions. The information, which this kind of measurement adds to the usual strategy, deals with the possibility of separating the PS components from the MPI and BBR contributions. While the objects coming from the hard scattering fall in the "Toward" and "Away" regions, in events with large initial- or final-state radiation, the "TransMAX" region might possibly contain a third jet, accompanied by MPI and BBR, while the "TransMIN" results to be affected only by MPI and BBR. In addition, the measurements in the "TransDIF" are only affected by the PS component, since MPI and BBR cancel out in the difference. A sketch of the situation is represented in Figure D.1, where "TransMAX" and "TransMIN" are shown in different colours. By performing four separate measurements in the transverse regions, a better tune can be extracted with a wide range of observables, sensitive to the different UE components. For the tune described in this Chapter, measurements of charged particle multiplicity and p_{T} sum in "TransMIN" and "TransMAX" have been considered at the four different energies; observables in "TransDIF" and "TransAV" have been used for cross checks.

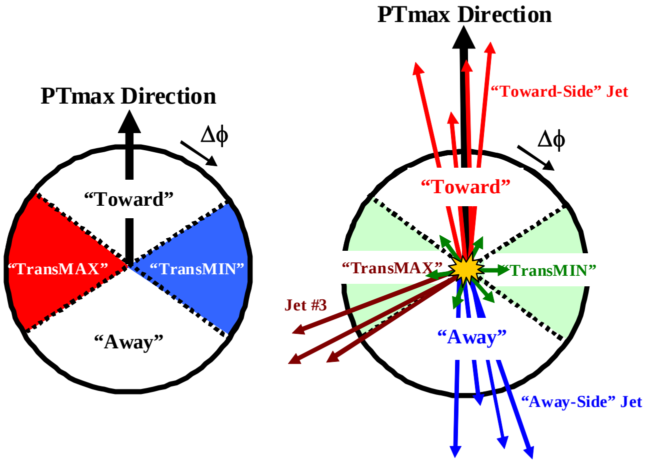


Figure D.1: Sketch of the transverse plane, as used in the new UE measurements performed at the CDF and CMS experiments: the transverse regions result to be divided in two subregions, "TransMIN" and "TransMAX", depending on their event content in terms of charged particle multiplicity and p_T *sum. While "TransMAX" might possibly contain an additional jet produced by hard radiation, only MPI and BBR contribute to "TransMIN".*

D.2 The software

Tuning parameters in a MC event generator is a very important matter since, in the available models, it is unavoidable that some of the parameters used in the simulation must be determined by looking for the best agreement between the predictions and the data. This function is generally accomplished by the tuning procedure and might be performed in several ways; one is the manual method, which consists of producing predictions by changing parameters "by hand" and searching for an improvement of the agreement with the data points in a bin-by-bin basis. Another similar method is the so-called brute-force method: it produces predictions with different choices of parameters, with a phase space divided in intervals. These methods have been used to model the first hadronization parameters in the very early stages of electron-positron colliders at the TASSO [256] and ALEPH [257] experiments. Even though they might be successful and well performing, these methods are certainly highly inefficient: the number of samples which need to be generated is very large if a detailed scan of the generator response has to be produced and, especially with the manual method, the correlations among the tuned parameters are very difficult to investigate. Hence, a different method has been chosen for the extraction of the new tunes, which largely decreases the time computation and is able to investigate the parameter correlations: this is the so-called parametrization-based procedure. What it basically adds to the previous methods, is the parametrization of the generator behaviour which is then used for a minimization of a goodness of fit function. This approach is the one implemented in the PROFESSOR software [258]: in particular, it automatizes the whole procedure by running parallel jobs with a random choice of the parameters. These predictions are obtained for a set of observables O, chosen for the tune, with specific UE parameters p, which are left free to vary within a defined range of values. The examined observables should indeed be sensitive to the selected parameters, otherwise a risk of obtaining inaccurate values arises. Different weights w_0 can also be applied to the observables in order to set the relative importance of each of them: this procedure is quite arbitrary, though, and needs particular investigation and care.

With the various generated predictions, a grid in the parameter space is built; the software performs then an interpolation of the bin values, for the considered observables, according to a polynomial function. A third-order polynomial function is generally used because it is a good compromise between the required number of simulated samples and the fitting performance; it has been checked that the degree of the polynomial does not influence the performance of a good interpolation. The only requirement to fulfill, in order to obtain a reliable parametrization, is the oversampling condition, such that the number of generated samples for different choices of parameters is at least two times larger than the lower limit specified in the Table in [259]. This lower limit depends on the order of the polynomial and the number of the tuned parameters. The obtained function $f^{\mathrm{b}}(p)$ models the MC response of each bin b as a function of the vector of the parameters p. The final step is the minimization of the χ^2 function given by this formula:

$$\chi^2(p) = \sum_{O} w_0 \sum_{\mathrm{b}\in O} \frac{(f^{\mathrm{b}}(p) - R_{\mathrm{b}})^2}{\Delta_{\mathrm{b}}^2} \tag{D.1}$$

where R_{b} is the data value for each bin b and Δ_{b} expresses the total bin uncertainty of the data. An important recommendation is also that the statistics of each generated sample should be such that for each bin the MC uncertainty is much smaller than the experimental one. This is because the statistical error in the MC samples is neglected in the χ^2 definition. Note that the experimental uncertainties are assumed to be uncorrelated between data points. The minimization procedure gives in output the values of the parameters which are able to best fit the considered data.

Furthermore, the uncertainty for each of the tuned parameters is also calculable by means of the eigentunes [260]; they produce a collection of deviation tunes, representing a set of uncertainties relative to the best tune. Eigentunes are created by using the covariance matrix in order to determine independent directions in the parameter space: along these directions the eigentunes are defined by the parameters obtained by allowing a specific variation of the χ^2. Possible values for the variation $\Delta\chi^2$ may be 1, the reduced χ^2 of the best tune or its absolute χ^2.

This whole procedure is executed by PROFESSOR, in a very user-friendly environment in simple successive steps. It works along with the RIVET package [185], which produces the whole set of MC predictions with the different choice of parameters which are then fed to the interpolation and minimization procedure. The RIVET software is a tool for producing predictions obtained with an input model implemented in a MC event generator and it is particularly useful for exploration and validation of any new MC settings. It contains a library of validated and published analyses, but it also allows the user to build its own analysis and get predictions for any selection and observable (see Chapter 4).

In the automatized way, described in this Section, the completion of a tune becomes a matter of two- or three-day timescale, which is the time needed to run the predictions. Examples of tunes, using the PROFESSOR machinery are documented in [222, 258].

D.3 Tune procedure and Pythia8 settings

In this Section, details and results of the new tunes performed with PYTHIA8 are described. Two tunes have been extracted: they use different PDF sets, CTEQ6L1 and HERAPDF1.5LO. The new tunes are, thus, respectively called CUETP8S1-CTEQ6L1 and CUETP8S1-HERAPDF1.5LO[1]. Parameters in the PYTHIA8 generator relative to the UE quantify the amount of MPI and colour reconnection included in the simulation. Specifically they are:

- MULTIPLEINTERACTIONS:PT0REF: it quantifies the amount of MPI during the collision: in particular, it sets the lower scale of the MPI contribution (see $p_{\mathrm{T0}}^{\mathrm{ref}}$ in Equation 2.12), together with the parameter MULTIPLEINTERACTIONS:ECMPOW;
- MULTIPLEINTERACTIONS:ECMPOW: it is also related to the amount of MPI at a given colliding energy; in particular, it is the positive exponent of the energy dependence for the MPI lower scale (see E_{pow} in Equation 2.12);
- MULTIPLEINTERACTIONS:EXPPOW: it expresses the amount of overlap between the two protons; in particular, if the overlap function is set to a negative exponential distribution, EXPPOW represents the exponent. For increasing exponents, the overlap is smaller, while smaller exponents translate to a higher overlap between the colliding protons;
- BEAMREMNANTS:RECONNECTRANGE: it quantifies the amount of colour reconnection which is performed in the simulation between the hard scattering and the UE system (see parameter R in Equation 2.19).

These parameters have been varied for the tuning, while the others, sensitive to hadronization and initial- and final-state radiation, have been kept fixed to the previous tune 4C [124]. Note that tune 4C and the new tunes use an exponential matter overlap function between the two colliding protons. In a tuning procedure, the choice of the parameter ranges becomes very important; in particular, it should guarantee a wide variation of the predictions but also be narrow enough to have a good sampling of the parameter grid with a reasonably small number of generated samples. The range of variation of the tuned parameters in PYTHIA8 is listed in Table D.1. The various intervals have been taken over from previous tuning attempts [261]. The simplest way to check if the tuning ranges are properly defined and chosen is to investigate the envelopes: they consist of bands representing the ensemble of all generated MC predictions for a given observable, generated with random choices of parameters inside the tuning range. The envelopes are centered around the mean value of the predictions on a bin-by-bin basis and their width depends on the standard deviation of the MC predictions with respect to the mean value. If the data points are well inside the envelopes, the tuning range can be considered appropriate, otherwise needs to be extended.

For the description of the energy dependence of UE data, only charged particle multiplicity and p_{T} sum in the "TransMIN" and "TransMAX" have been considered for the four different collision energies[2]. The analyses are included as plugins in the RIVET package under the name "CDF_Tevatron" and "CMS_FSQ_12_020", respectively for the measurements performed at the CDF and CMS experiments. Minimum Bias (MB) events have been produced for each energy with

---

[1] The first part of the name stands for *C*ms *U*nderlying *E*vent *T*une *P*ythia 8 *S*et 1, while the second part refers to the PDF set.

[2] A weight w_0 equal to 1 has been applied to each observable.

Parameter	Tuning Range
MultipleInteractions:pT0Ref	1.0-3.0
MultipleInteractions:ecmPow	0.0-0.4
MultipleInteractions:expPow	0.4-10.0
BeamRemnants:reconnectRange	0.0-9.0

Table D.1: Tuning ranges of the parameters used in the construction of the PROFESSOR *grid.*

200 different choices of parameters in the simulation settings: the envelopes obtained from the MC predictions are shown in Figure D.2 for the charged particle multiplicity in the "TransMIN" region for four different energies (900, 1960 and 7000 GeV) when using the two different PDF sets. A MB event simulates a generic inelastic pp or $p\bar{p}$ collision with relatively low exchanged p_{T}. The envelopes show that the data points are well inside the blue bands at all energies and for each observable: this is a good indication that the chosen parameter range is sufficient. The tune has been then extracted for the two different PDF sets by interpolating the parameter grid and minimizing the χ^2 function in Equation D.1. A difficulty has been observed when tuning data at all collision energies: data at 300 GeV are particularly difficult to be described within the same tune together with data at the other energies. Since the content of MPI at this low energy is quite small, it has been preferred to exclude these data from the tuning and leave further investigations for later. Hence, only data at 900, 1960 and 7000 GeV have been considered for the tuning in the following.

Results of the new tunes are shown in Table D.2, compared to the parameters implemented in the old tune 4C: it can be seen that the exponent of the matter function tends to decrease in the new tunes, indicating a higher overlap between the two colliding protons, while the regulator p_T^0, as defined in Equation 2.12, increases. Parameters for CUETP8S1-CTEQ6L1 and CUETP8S1-HERAPDF1.5LO tunes are very similar between each other, except the one related to colour reconnection, which depends on the parton dynamics at small longitudinal momenta, which is different for the two PDF sets.

Parameter	CUETP8S1-CTEQ6L1	CUETP8S1-HERAPDF1.5LO	4C
MultipleInteractions:pT0Ref	2.100620	2.000072	2.085
MultipleInteractions:ecmPow	0.2105712	0.2498802	0.19
MultipleInteractions:expPow	1.608895	1.690506	2.0
BeamRemnants:reconnectRange	3.312569	6.096364	1.5

Table D.2: Parameters obtained for the new tunes of PYTHIA*8, performed with CTEQ6L1 and HERA-PDF1.5LO PDF sets. Values of the parameters as implemented in the old tune 4C are listed in the last column.*

Comparisons of predictions of old and new tunes are then investigated in order to check if the agreement with the data effectively improves with the new parameter settings. Charged particle multiplicities in the "transMIN" and "transMAX" regions are shown in Figure D.3 for the three tuned collision energies. Charged particle p_{T} sums are shown in the same regions in Figure D.4. From the comparisons, it can be seen that the agreement between data and simulation significantly improves with the new tunes: while the rising part and the plateaux region of the UE spectra are underestimated by the predictions of the tune 4C at 900 and 1960 GeV for both charged particle

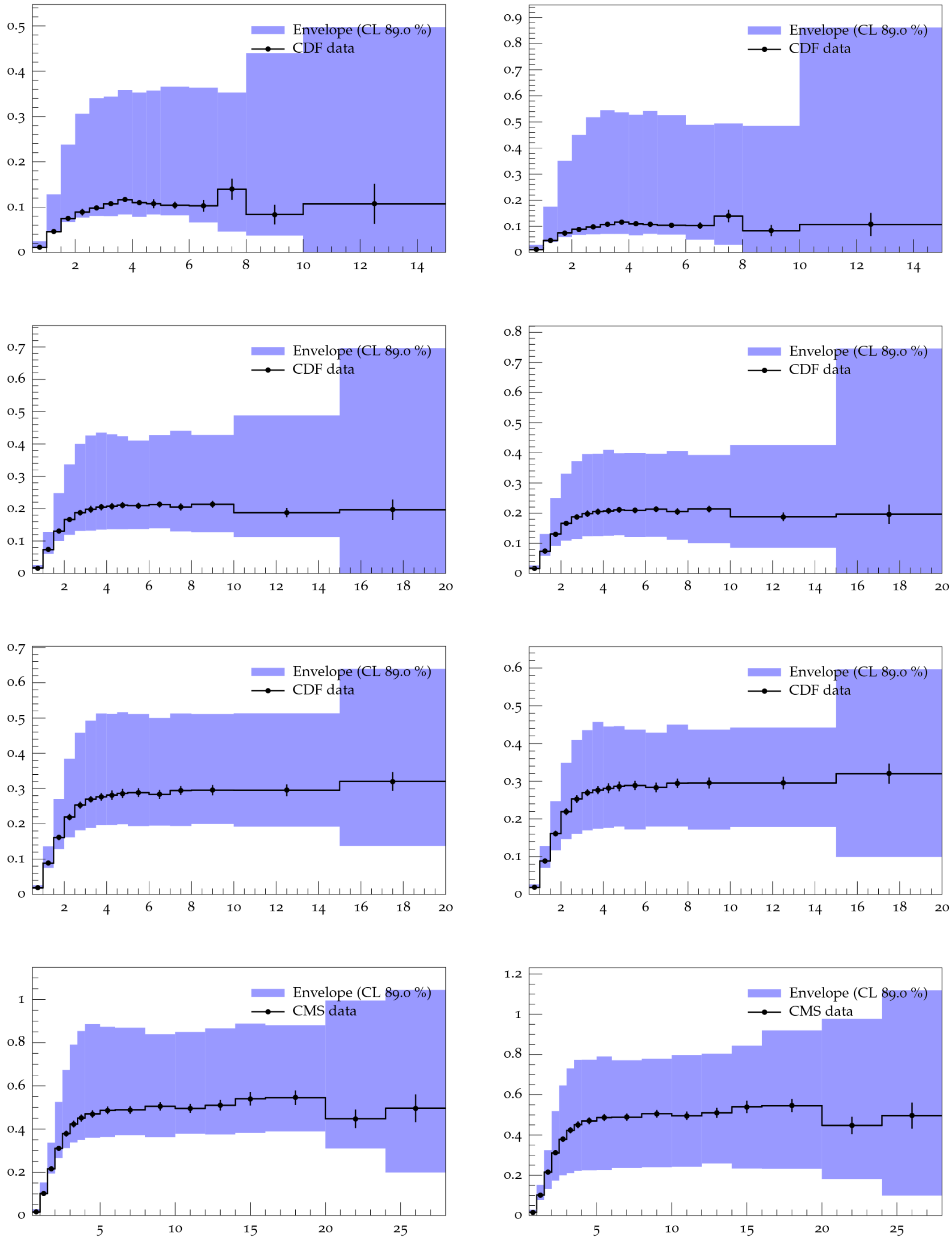


Figure D.2: Envelopes of the grid MC samples generated for the tunes, performed with PYTHIA*8 using the CTEQ6L1 (left) and the HERAPDF1.5LO (right) sets, on the charged particle multiplicity density measured in the "transMIN" region for energy collisions equal to, from top to bottom, 300, 900, 1960 and 7000 GeV.*

multiplicity and p_{T} sum, the data points are very well described by the predictions of the new tunes in the whole phase space. At 7 TeV, predictions of tune 4C describe the data reasonably well and a very high level of agreement is also achieved by the new tunes.

Comparisons at 300 GeV have been also investigated and are shown in Figure D.5 for charged particle multiplicity and p_{T} sum in the "transMIN" and "transMAX" regions. Data at this energy have not been considered for the tuning. It can be seen that the predictions of the new tunes are above the data and they overestimate the plateaux region by 40% in "transMIN" and by 20% in "transMAX"; the predictions of the old tune 4C are closer to the data but the description is not optimal. This inconsistency observed at low collision energies has been further investigated: data at 300 GeV seem to prefer a higher p_T^0 which reduces the MPI contribution contained in the new tunes. A tune performed with PYTHIA6, described in [12], which uses a double gaussian matter overlap, is able to reproduce UE data at all collision energies without any problem at 300 GeV. A tune, which uses a double gaussian overlap function, extracted with PYTHIA8, has not shown, though, the same feature of correctly describing the UE data at 300 GeV; the understanding of the different behaviours between PYTHIA8 and PYTHIA6 is still under investigation. However, the new tunes, CUETP8S1-CTEQ6L1 and CUETP8S1-HERAPDF1.5LO, are considered very reliable thanks to the robustness of the results obtained at all the other energies. Comparisons with observables in the "transDIF" and "transAV" regions are shown in [262] for the new tunes: the level of agreement with the data is very high, similar to the one achieved in the "transMIN" and "transMAX" regions.

D.4 Validation plots

After checking the predictions of the new tunes, it is also very important to investigate if the new tunes are able to reproduce more inclusive data, related to charged particles, jets and energy event content, in different pseudorapidity regions and at different collision energies. A set of predictions have been produced for a wide list of measurements, included in the RIVET package; they are listed in the following:

- Pseudorapidities at three energies, charged multiplicity at 7 TeV in ALICE (ALICE_2010_S8625980 [263]);
- UE and MB measurement in ATLAS and CMS (ATLAS_2010_S8918562 [264], ATLAS_2010_S8894728 [90], CMS_2011_S9120041 [89] and CMS_2012_PAS_FWD_11_003 [265]);
- Measurement of track jet properties at 7 TeV in CMS and ATLAS (ATLAS_2011_I919017 [266] and CMS_2011_S8957746 [267]);
- K_0^{S} and Λ production at 0.9 and 7 TeV with ATLAS and CMS (ATLAS_2011_I944826 [268] and CMS_2011_S8978280 [269])
- Pseudorapidity dependence of the total transverse energy at 7 TeV in ATLAS (ATLAS_2012_I1183818 [270])
- Measurement of the NSD charged particle multiplicity at 0.9, 2.36, and 7 TeV with CMS (CMS_2011_S8884919 [271])

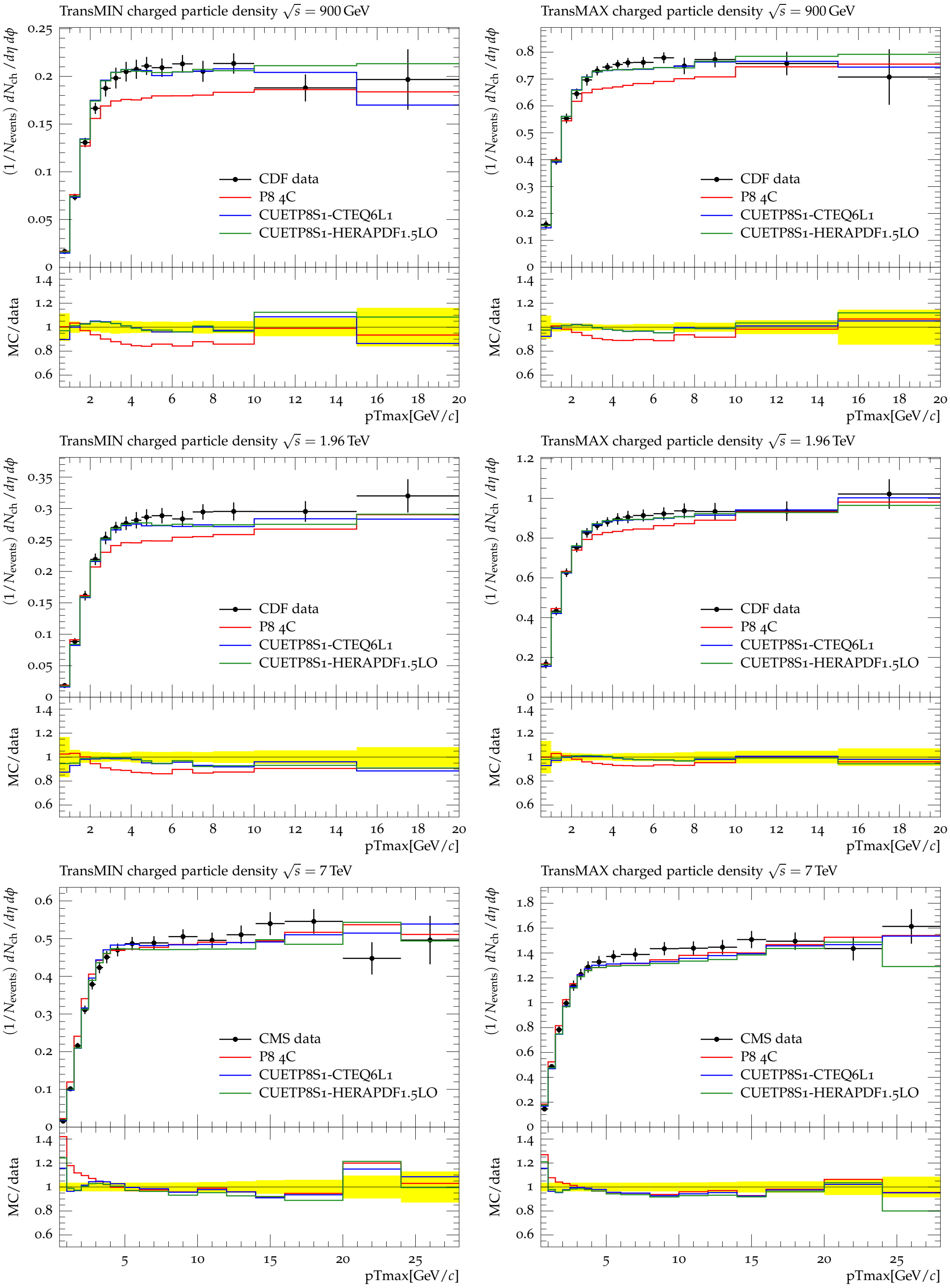


Figure D.3: CDF and CMS data for respectively $p\bar{p}$ and pp collisions at 900 (top row), 1960 (center) and 7000 (bottom row) GeV: density of charged particles with $p_T > 0.5$ GeV/c and $|\eta| < 0.8$ in the "transMIN" (left column) and "transMAX" (right column) regions as defined by the leading charged particle, as a function of p_T^{max}. The data are compared with PYTHIA*8 Tune 4C and the two new CMS* PYTHIA*8 tunes using CTEQ6L1 and the HERAPDF1.5LO. Results are published in [12].*

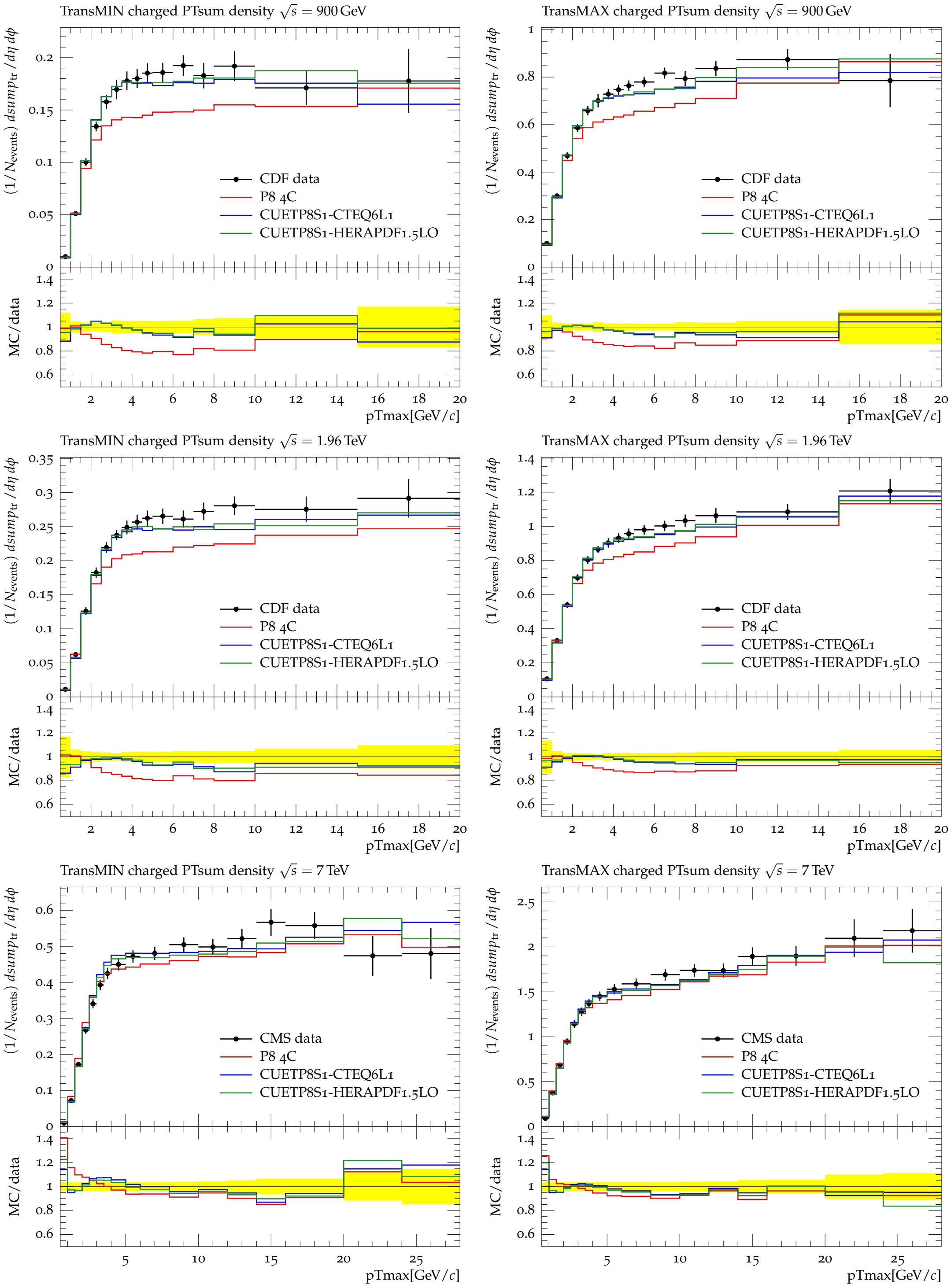


Figure D.4: CDF and CMS data for respectively $p\bar{p}$ and pp collisions at 900 (top row), 1960 (center) and 7000 (bottom row) GeV: p_T sum density of charged particles with $p_T > 0.5$ GeV/c and $|\eta| < 0.8$ in the "transMIN" (left column) and "transMAX" (right column) regions as defined by the leading charged particle, as a function of p_T^{max}. The data are compared with Pythia*8 Tune 4C and the two new CMS* Pythia*8 tunes using CTEQ6L1 and the HERAPDF1.5LO. Results are published in [12].*

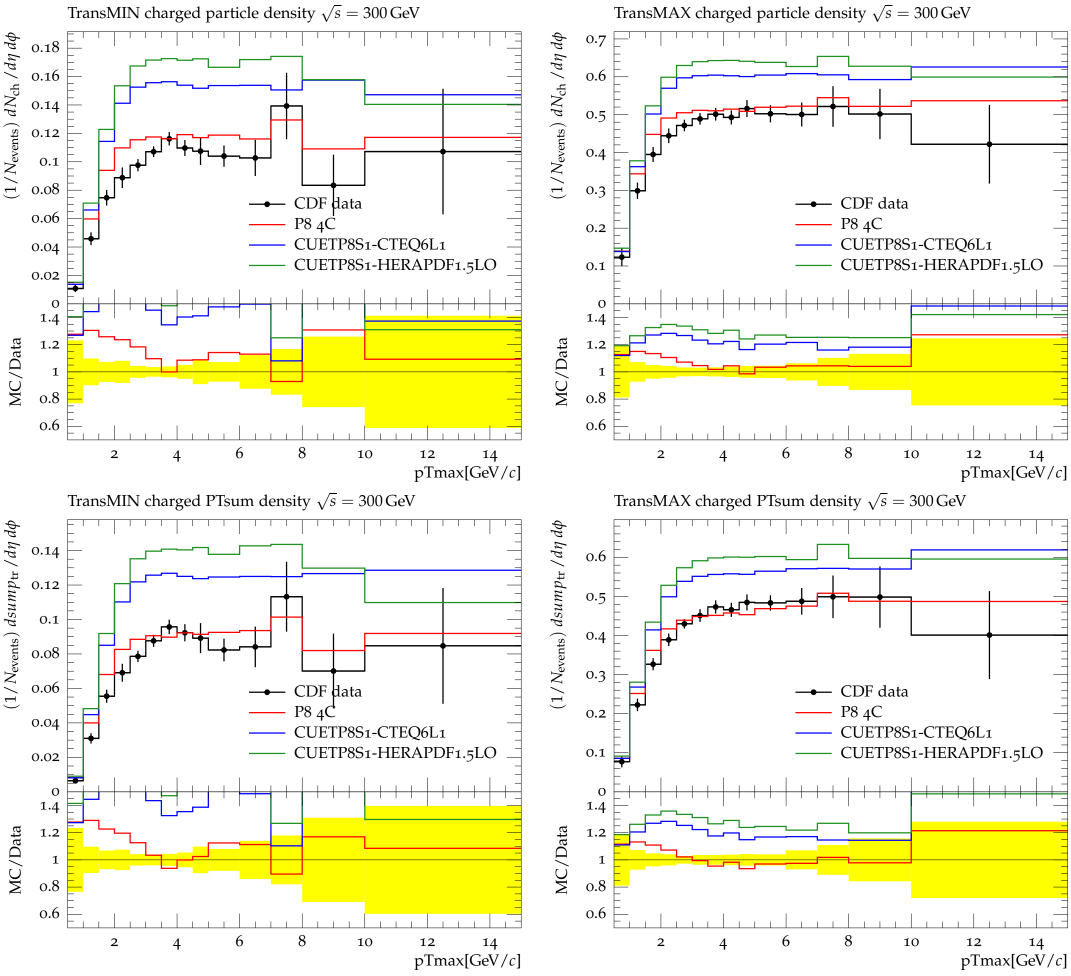


Figure D.5: CDF data for $p\bar{p}$ collisions at 300 GeV: multiplicity (top row) and p_T sum (bottom row) density of charged particles with $p_T > 0.5$ GeV/c and $|\eta| < 0.8$ in the “transMIN” (left column) and “transMAX” (right column) regions as defined by the leading charged particle, as a function of p_T^{max}. The data are compared with Pythia8 Tune 4C and the two new CMS Pythia8 tunes using CTEQ6L1 and the HERAPDF1.5LO.

- Charged particle transverse momentum and pseudorapidity spectra from proton-proton collisions at 7000 GeV in CMS (CMS_QCD_10_024 [272])
- Strange particle production in underlying events in proton–proton collisions at 7 TeV in CMS (CMS_2012_PAS_QCD_11_010 [273])
- Forward energy flow at 7 TeV with CMS (CMS_2011_S9215166 [95])
- Forward dN/dη at 7 TeV with TOTEM (TOTEM_2012_I1115294 [237])

A small selection of the comparisons is shown in Figure D.6, D.7 and D.8. Figure D.6 shows the forward energy flow at 7 TeV in MB and dijet events measured by CMS, and the charged particle multiplicity as a function of pseudorapidity in the central and forward region, measured by ALICE and TOTEM. In Figure D.7, a comparison for the standard UE observables measured by ATLAS is shown, while Figure D.8 shows an overview of measurements at different energies: the forward energy flow in MB and dijet events and the charged particle multiplicity measured at 900 GeV and ratios of energy deposits in dijet and MB events in the very forward region at 900, 2360 and 7000 GeV. A more complete overview is provided in [262]. Predictions obtained with old and new tunes implemented in PYTHIA6 and PYTHIA8 are shown in the plots. An overall good picture is offered by the new tunes: they achieve a high level of agreement for all the observables measured in the central region, at all collision energies: for some measurements, for instance charged particle multiplicities, a significant improvement is observed. In the forward region, still, the description is not optimal for both energy flow and charged particle multiplicity: in particular, the data are underestimated by the predictions of the new tunes. This disagreement seems to be due to the gluon content at low x in the proton: in fact, a nice description is observed for predictions obtained with the Monash tune [231], which uses a PDF set that is different at very small x. However, the Monash tune is not always performing better than the new CMS UE tunes, especially for measurements performed at collision energies lower than 7 TeV. This gives confidence that the CUETP8S1-CTEQ6L1 and CUETP8S1-HERAPDF1.5LO tunes may be very meaningful for predictions at collision energies of 13 and 14 TeV and become relevant for the future LHC phase.

D.5 Predictions at 13 TeV

In this Section, predictions at 13 TeV are described for old and new tunes. They are shown for the same observables measured at the other energies in Figure D.9. The obtained predictions are very encouraging: the curves obtained with the new tunes, both for PYTHIA6 and PYTHIA8, are very close to each other, showing a high stability in predicting the event contents. Some discrepancies are observed with respect to predictions obtained with the old tunes: in particular, they are slightly higher in the rising part of the spectrum.

From these results, it is fair to say that the new CMS UE tunes, which describe the energy dependence of UE observables, are ready to become the reference for measurements at 13 TeV in the early stages of the new LHC phase: measurements of UE variables, charged particle multiplicities and energy flow at 13 TeV have been set at high priority in the CMS collaboration and they are expected to be released within the first months of next year.

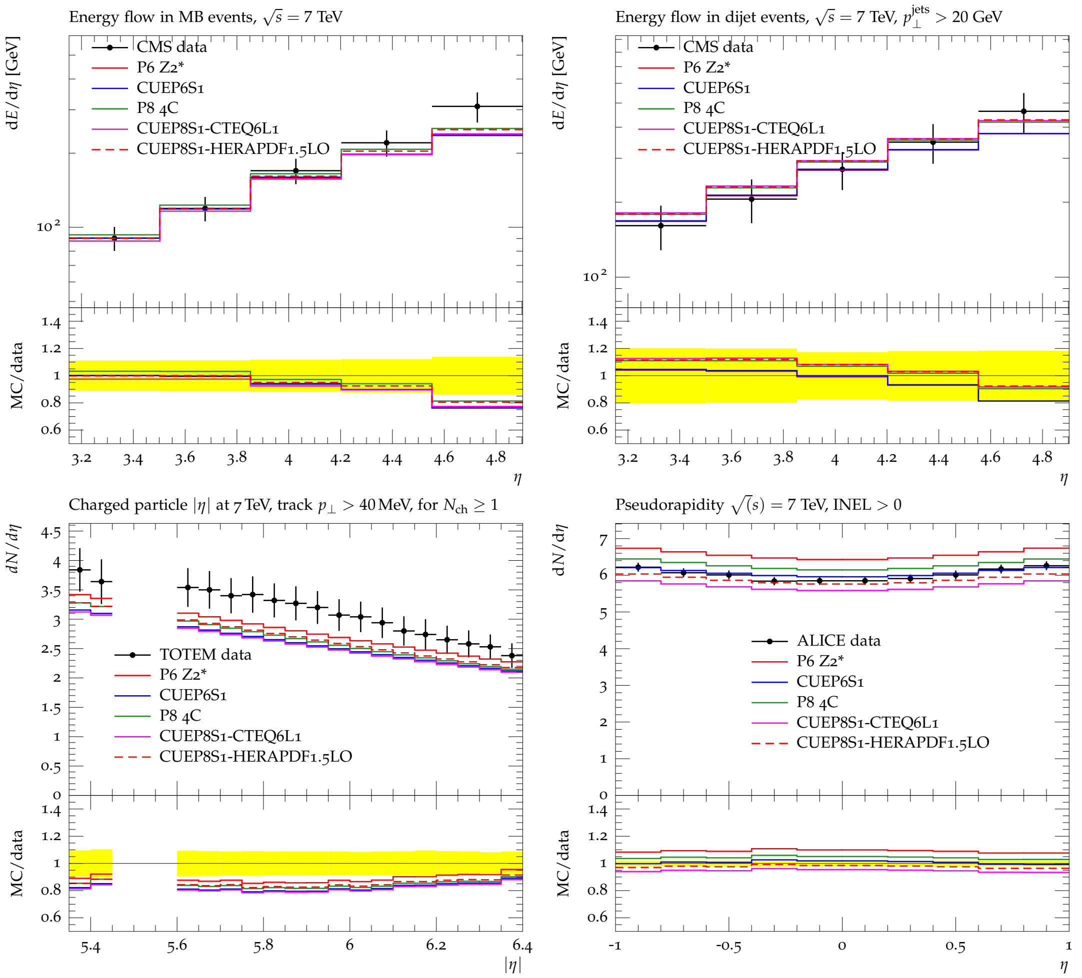


Figure D.6: CMS data on the forward energy flow in MB (top left) and dijet (top right) events, ALICE data and TOTEM data on charged particle pseudorapidity in, respectively, the central (bottom left) and forward (bottom right) region. The data are compared with Pythia*6 tune Z2*,* Pythia*8 tune 4C, the new* Pythia*6 tune, and the two new CMS* Pythia*8 tunes using CTEQ6L1 and the HERAPDF1.5LO. Also shows the ratio of the tunes with the data. Results are published in [12].*

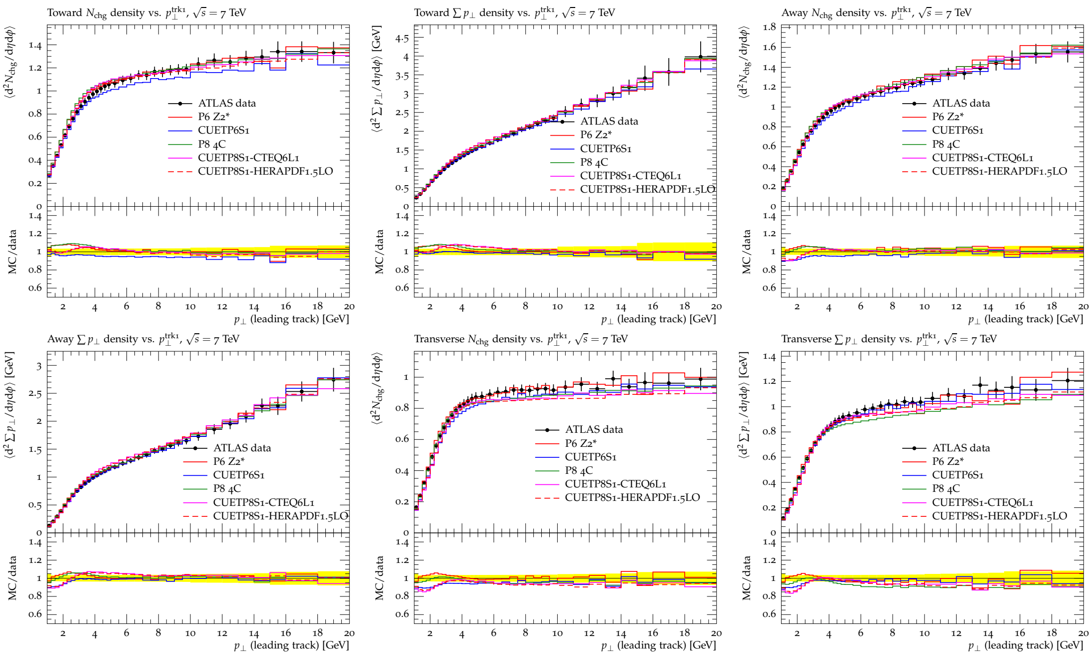


Figure D.7: ATLAS data on the charged particle multiplicity N_{ch} (top) and p_T sum (bottom) measured in the transverse (left), toward (center) and away (right) regions. The data are compared with PYTHIA*6 tune Z2*,* PYTHIA*8 tune 4C, the new* PYTHIA*6 tune, and the two new CMS* PYTHIA*8 tunes using CTEQ6L1 and the HERAPDF1.5LO. Also shows the ratio of the tunes with the data. Results are published in [12].*

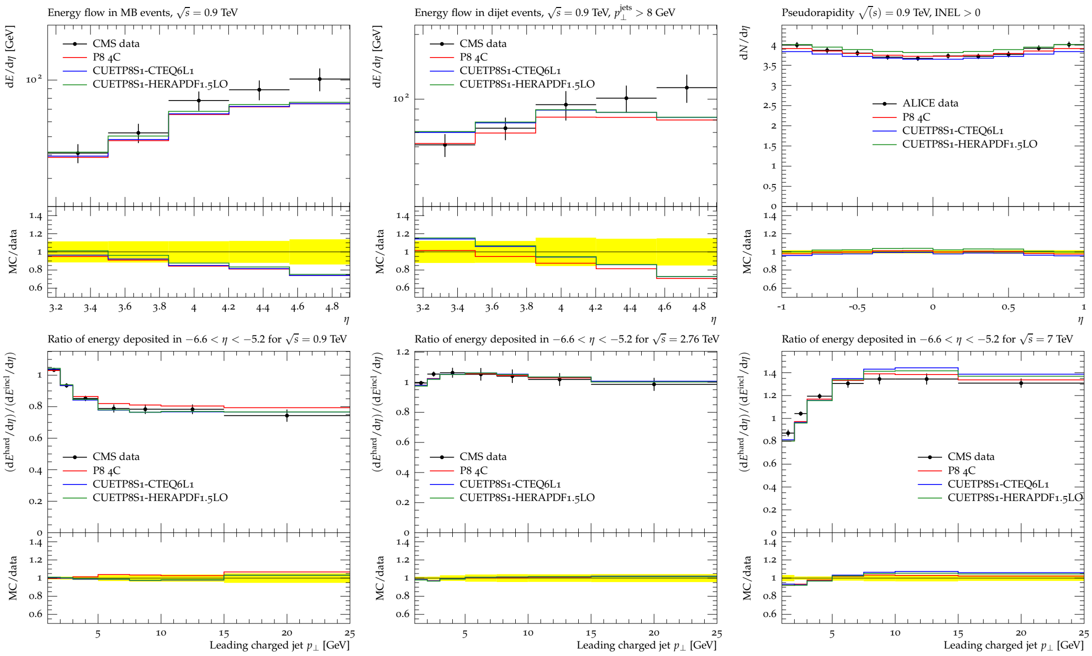


Figure D.8: (Top row) CMS and ALICE data on energy flow in MB (left) and dijet (center) events and on charged particle multiplicity as a function of pseudorapidity, at 900 GeV. (Bottom row) CMS data on the ratios of energy deposited in dijet and MB events in the very forward region, measured at different collision energies: 900 GeV (left), 2.76 (center) and 7 TeV (right). The data are compared with PYTHIA*8 tune 4C, the two new CMS* PYTHIA*8 tunes using CTEQ6L1 and the HERAPDF1.5LO. The lower panel shows the ratios between each MC prediction and the data.*

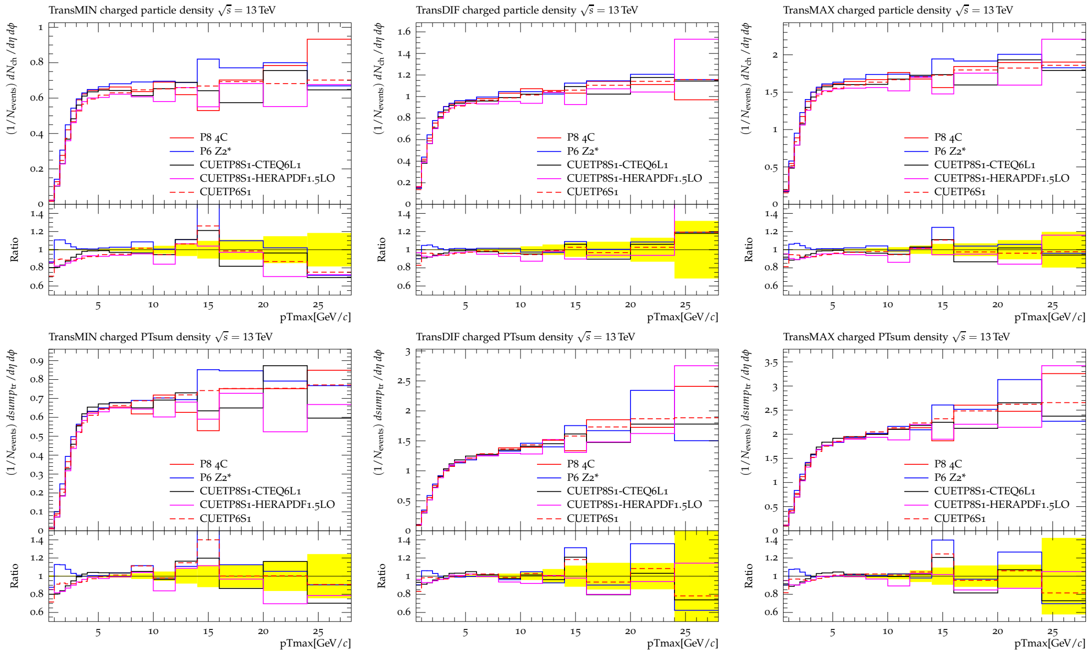


Figure D.9: Predictions for pp collisions at 13 TeV for Pythia8 Tune 4C, Pythia6 tune Z2, the two new CMS Pythia8 tunes using CTEQ6L1 and the HERAPDF1.5LO, and the the new CMS Pythia6 tune: charged particle density (top plots) and p_T sum density (bottom plots) for charged particles with $p_T > 0.5$ GeV/c and $|\eta| < 0.8$ in the “transMIN” (left), “transDIF” (center) and the “transMAX” (right) regions as defined by the leading charged particle, as a function of p_T^{max}. Also shown are the ratios of the new CMS tunes to Tune 4C. Results are published in [12].

Appendix E

Technical details of the tuning method and additional comparisons

In this Appendix, a deeper view of the fits performed for the measurement of the DPS contribution is provided. Technical details about the tuning procedure and the determination of σ_{eff} with MC event generators are given. The results of the parameters obtained for each fit are listed for both 4j and 2b2j channels. Additional comparisons are also shown: they include predictions obtained with all new tunes for the whole collection of correlation observables, measured in the two scenarios.

E.1 Parameter ranges used for the DPS-based tunes

In this Section, the parameter ranges used for the DPS-based tunes are listed for all settings relative to the different choice of overlap matter distribution function. The parameters which are considered for the tuning are the same as used in [12] and are related to the amount of MPI, of proton overlap and of colour reconnection. As explained in Appendix D, the parameter ranges are essential for constructing the grid built by the PROFESSOR machinery and for performing the χ^2 minimization. Tables E.1-E.3 list the tuned parameters and their ranges for, respectively, the PYTHIA8 tunes using the negative exponential, the double gaussian and the x-dependent single gaussian overlap matter distribution functions. All of them refer to tunes extracted from both the 4j and the 2b2j channels. For the 4j measurements, an additional tune has been performed with the PYTHIA6 generator using a double gaussian overlap matter distribution function; the tuned parameters[1] and their ranges are shown in Table E.4.

---

[1] In total, six parameters are tuned in PYTHIA6, in contrast to the double gaussian case in PYTHIA8 which has only five. The difference stems from the fact that PYTHIA6 uses two parameters for defining the quantity of colour reconnection, instead of one as in PYTHIA8.

CDPSTP8S1 - CDPSTP8S2	
Parameter	Tuning Range
MultipleInteractions:pT0Ref	1.0-3.0
MultipleInteractions:ecmPow	0.0-0.4
MultipleInteractions:expPow	0.4-10.0
BeamRemnants:reconnectRange	0.0-9.0

Table E.1: Variation ranges of UE parameters used for fits performed with PYTHIA*8 and implementing a negative exponential matter overlap distribution function.*

CDPSTP8S3	
Parameter	Tuning Range
MultipleInteractions:pT0Ref	1.0–3.0
MultipleInteractions:ecmPow	0.1–0.4
MultipleInteractions:coreRadius	0.2–0.8
MultipleInteractions:coreFraction	0.4–0.95
BeamRemnants:reconnectRange	0.0–9.0

Table E.2: Variation ranges of UE parameters used for fits performed with PYTHIA*8 and implementing a double gaussian matter overlap distribution function.*

CDPSTP8S4	
Parameter	Tuning Range
MultipleInteractions:pT0Ref	1.0–3.0
MultipleInteractions:ecmPow	0.1–0.4
MultipleInteractions:a1	0.0–1.0
BeamRemnants:reconnectRange	0.0–9.0

Table E.3: Variation ranges of UE parameters used for fits performed with PYTHIA*8 and implementing a x-dependent matter overlap distribution function.*

CDPSTP6S1	
Parameter	Tuning Range
PARP(82)	1.6–2.2
PARP(77)	0.25–1.2
PARP(78)	0.2–0.6
PARP(90)	0.18–0.28
PARP(83)	0.1–1.0
PARP(84)	0.1–1.0

Table E.4: Variation ranges of UE parameters used for fits performed with PYTHIA*6 and implementing a double gaussian matter overlap distribution function.*

For each fit performed from measurements in the 4j channel, 400 samples, with 10^6 QCD events each, have been generated to build the grid in the parameter space for the minimization. For the tuning procedure used in the 2b2j channel, 20 million QCD events have been generated for 100 different choices of UE parameters. A $\hat{p}_{\mathrm{T}}$ of 45 and 15 GeV has been set in the MC production for, respectively, the 4j and 2b2j analyses, in order to increase the selection efficiency of the considered scenario.

E.2 Determination of σ_{eff} from MC generators

The fits based on the correlation observables of DPS-sensitive channels and described in Chapter 11 give in output a set of parameters relative to the UE simulation. From these parameters, the value of the predicted σ_{eff} can be easily determined within the machinery of a MC event generator. In Pythia8, after setting the values of the parameters in the configuration file, a generation with two hard scatterings inside the same collision needs to be produced: this is accomplished by including the command:

```
'SecondHard : generate = on'
```

which forces the event at the second hard scale in the collision within the interleaved formalism in Pythia8 (see Chapter 2) to be a MPI. Such generation activates the calculation of f_{EF} (see Equation 11.6), needed for the determination of σ_{eff}. The factor f_{EF} is obtained through Monte Carlo generation; hence, a sufficient number of generated events[1] is necessary for getting reliable values. The generator output of f_{EF} and σ_{ND} looks like the following:

```
[..] an impact-parameter enhancement factor of 1.774e+00
```

```
[..] using a sigma (nonDiffractive) of 5.091e+01 mb [..]
```

and it is given at the end of each run. The two values can be then applied to Equation 11.6 to get σ_{eff}.

Through the same method, the estimation of σ_{eff} is also possible with the Pythia6 generator but is a bit more complicated, since it needs to use again the Pythia8 output. The procedure has been provided by the Pythia8 authors. First of all, in the Pythia8 configuration file, the following command:

```
SigmaDiffractive:dampen = off
```

needs to be set, in order to switch off the dampening of the diffractive cross section, which is not implemented in Pythia6. Then, a fine tuning of p_{T}^{0} needs to be applied in order to match the values of the total QCD cross sections in the outputs of Pythia6 and Pythia8. The σ_{eff} value is thus determined with the new settings from the Pythia8 output, as described before. Estimation of σ_{eff} is also possible with Herwig++ in a similar way, but since a final measurement is not yet

---

[1] In the results presented in the following sections, 20000 generated events have been considered sufficient for this purpose.

available, it is not further treated.

E.3 Additional comparisons of the 4j selection

In this Section, the parameters obtained for each fit of the 4j channel are listed and additional comparisons are provided. Results of the minimization are listed in Table E.5 for CDPSTP8S1-4j and CDPSTP8S2-4j tunes which use the negative exponential overlap matter distribution function, together with the parameters of tune 4C. The parameters, which are accompanied by a star in CDPSTP8S1-4j, have been kept fixed to the values of tune 4C and not considered for the fit. From the comparisons of the parameters of the new tunes and the old tune 4C, it may be noticed that the exponent of the matter distribution function decreases, indicating a higher overlap between the colliding protons and translating into a larger contribution of DPS implemented in the tune. Furthermore, the CDPSTP8S2-4j tune predicts a higher value of p_{T}^{0} and of the colour reconnection parameter. The parameters of the other new tunes, CDPSTP8S3-4j and CDPSTP8S4-4j, are shown respectively in Tables E.6 and E.7. In PYTHIA8, no reference tunes are available for these settings.

The parameters of the new PYTHIA6 tune, CDPSTP6S1-4j, are shown in Table E.8, compared to the ones of the reference tune Z2*. The PYTHIA6 DPS-based tune tends to have again a higher overlap between the protons, expressed by the higher fraction (PARP(84)) and the lower radius (PARP(83)) of the core. The value of p_{T}^{0} is higher than the one of the old tune Z2*, while values of colour reconnection parameters are similar between new and old tunes.

Parameter	CDPSTP8S1-4j	CDPSTP8S2-4j	4C
MultipleInteractions:pT0Ref	2.085*	2.125405	2.085
MultipleInteractions:ecmPow	0.19*	0.3450478	0.19
MultipleInteractions:expPow	1.160	0.691404	2.0
BeamRemnants:reconnectRange	1.5*	6.525605	1.5
Reduced χ^2	0.751	0.428	-

Table E.5: Parameters obtained for the new PYTHIA8 DPS-based tunes, CDPSTP8S1-4j and CDPSTP8S2-4j, performed with the CTEQ6L1 PDF set, with a negative exponential overlap matter distribution. Parameters of the tune CDPSTP8S1-4j accompanied by a star, are the ones which have been kept fixed at the values set in the reference tune 4C. Values of the parameters as implemented in the old tune 4C are listed in the fourth column. The χ^2 obtained in the fit is also listed for each new tune, where available.

For each tune, the χ^2, obtained in the minimization, is also listed: for all PYTHIA8 tunes, values around 0.4–0.7 are obtained, while for the new PYTHIA6 tune, a higher value of the fit χ^2 is observed, up to 1.8. The low values of the χ^2 obtained with PYTHIA8 may be explained by the fact that the fit is performed by accounting both statistical and systematic uncertainties of the data, which are quite large compared to the level of agreement achieved by the new tunes. Further studies have been carried out to check the stability of the obtained values of σ_{eff} for fits performed without including the systematic uncertainties in the data[1]. The motivation for this is that sys-

---

[1] As seen in Chapter 8, the systematic effects are the driven sources of uncertainties for the 4j measurements. The statistical ones are of the order of 1%.

Parameter	CDPSTP8S3-4j
MultipleInteractions:pT0Ref	1.714327
MultipleInteractions:ecmPow	0.3921473
MultipleInteractions:coreRadius	0.3020031
MultipleInteractions:coreFraction	0.4353391
BeamRemnants:reconnectRange	7.502975
Reduced χ^2	0.441

Table E.6: Parameters obtained for the new DPS-based tune of Pythia*8, performed with the CTEQ6L1 PDF set, with a double gaussian overlap matter distribution. The χ^2 obtained in the fit is also listed for the new tune.*

Parameter	CDPSTP8S4-4j
MultipleInteractions:pT0Ref	1.989806
MultipleInteractions:ecmPow	0.2424875
MultipleInteractions:a1	0.2869012
BeamRemnants:reconnectRange	5.782312
Reduced χ^2	0.434

Table E.7: Parameters obtained for the new DPS-based tunes of Pythia*8, performed with CTEQ6L1 PDF set, with a single gaussian x-dependent overlap matter distribution. Parameters have been tuned to the ΔS and $\Delta^{rel}_{soft}p_T$ observables measured in the 4j channel. The χ^2 obtained in the fit is also listed for the new tune.*

Parameter	CDPSTP6S1-4j	Z2*
PARP(77)	1.475120	1.016
PARP(78)	0.3749465	0.538
PARP(83)	0.6284394	0.356
PARP(84)	0.3324244	0.651
PARP(90)	0.2833249	0.227
PARP(82)	1.483450	1.921
Reduced χ^2	1.78	

Table E.8: Parameters obtained for the new DPS-based tunes of Pythia*6, performed with CTEQ6L1 PDF set, with a double gaussian overlap matter distribution. Parameters have been tuned to the ΔS and $\Delta^{rel}_{soft}p_T$ observables measured in the 4j channel. The χ^2 obtained in the fit is also listed for the new tune.*

tematic uncertainties may lead to results which predict different shapes with respect to the central data points but are still compatible within uncertainties with them. This would indeed indicate an instability of the results. Values of σ_{eff} obtained with the two different procedures are shown in Table E.9 for the new Pythia8 tunes. When the systematic uncertainties are not considered in the fit, the values of the reduced χ^2 increase up to a maximum of 1.7: this is a reasonable and expected effect and the goodness of the new fits is still rather high. What is even more interesting is the high stability of the σ_{eff} values with the two types of fits. This conclusion further confirms the reliability of the results, obtained for the new tunes.

Tune	Reduced χ^2 with systematics	σ_{eff} with systematics	Reduced χ^2 no systematics	σ_{eff} no systematics
CDPSTP8S1-4j	0.751	$21.3^{+1.2}_{-1.6}$	1.74	20.3
CDPSTP8S2-4j	0.428	$19.0^{+4.7}_{-3.0}$	1.34	19.2
CDPSTP8S3-4j	0.441	$23.1^{+3.9}_{-4.0}$	0.988	23.4
CDPSTP8S4-4j	0.434	$16.3^{+6.1}_{-3.7}$	1.39	16.0

Table E.9: Values of σ_{eff} and fit χ^2 obtained for each Pythia*8 tune, when the experimental systematic uncertainty has been included and removed from the fitted data.*

Besides the comparisons shown in Figure 11.4, predictions from the other tunes are also compared to the measurements. Results from CDPSTP8S2-4j are shown in Figure E.1(top): the predictions of this tune propagated to a higher-order ME simulated with Madgraph are also shown. Both of them, the standard Pythia8 CDPSTP8S2-4j and the Madgraph+CDPSTP8S2-4j predictions, are equally able to reproduce very well the data: in particular, predictions obtained with Madgraph+CDPSTP8S2-4j are for every bin consistent with the data points. This suggests that a higher DPS contribution in the UE simulation is needed for describing 4j data, independently on the ME used in the generator; hence, the obtained tunes can be propagated to different generators. Figure E.1(middle) shows a summary of the predictions of the new Pythia8 tunes: all of them describe optimally both correlation observables. Finally, results from CDPSTP6S1-4j are shown in Figure E.1(bottom) when different ME are used: the 2→2 LO ME, simulated with Pythia6, the 2→4 LO ME, provided by Madgraph, and the 2→2 NLO ME with a hard emission, generated with Powheg. Even though the level of agreement obtained with the new Pythia6 tune is slightly worse than the one achieved by any of the new Pythia8 tunes, the predictions obtained with CDPSTP6S1-4j and different ME are consistent with each other, showing again the independence between the UE and the ME simulation used in the generator for the description of the correlation observables.

In Figure E.2, comparisons are shown between data and predictions obtained with different tunes, for the $\Delta\phi^{\mathrm{soft}}$ variable. All new tunes are able to describe very well the measurements. A very good description is also provided by predictions obtained with a Powheg or Madgraph ME interfaced with the UE simulation of a DPS-based tune. Note that also predictions obtained without the simulation of MPI are in agreement with the measurements: this indicates a very low sensitivity of the $\Delta\phi^{\mathrm{soft}}$ variable to the DPS contribution.

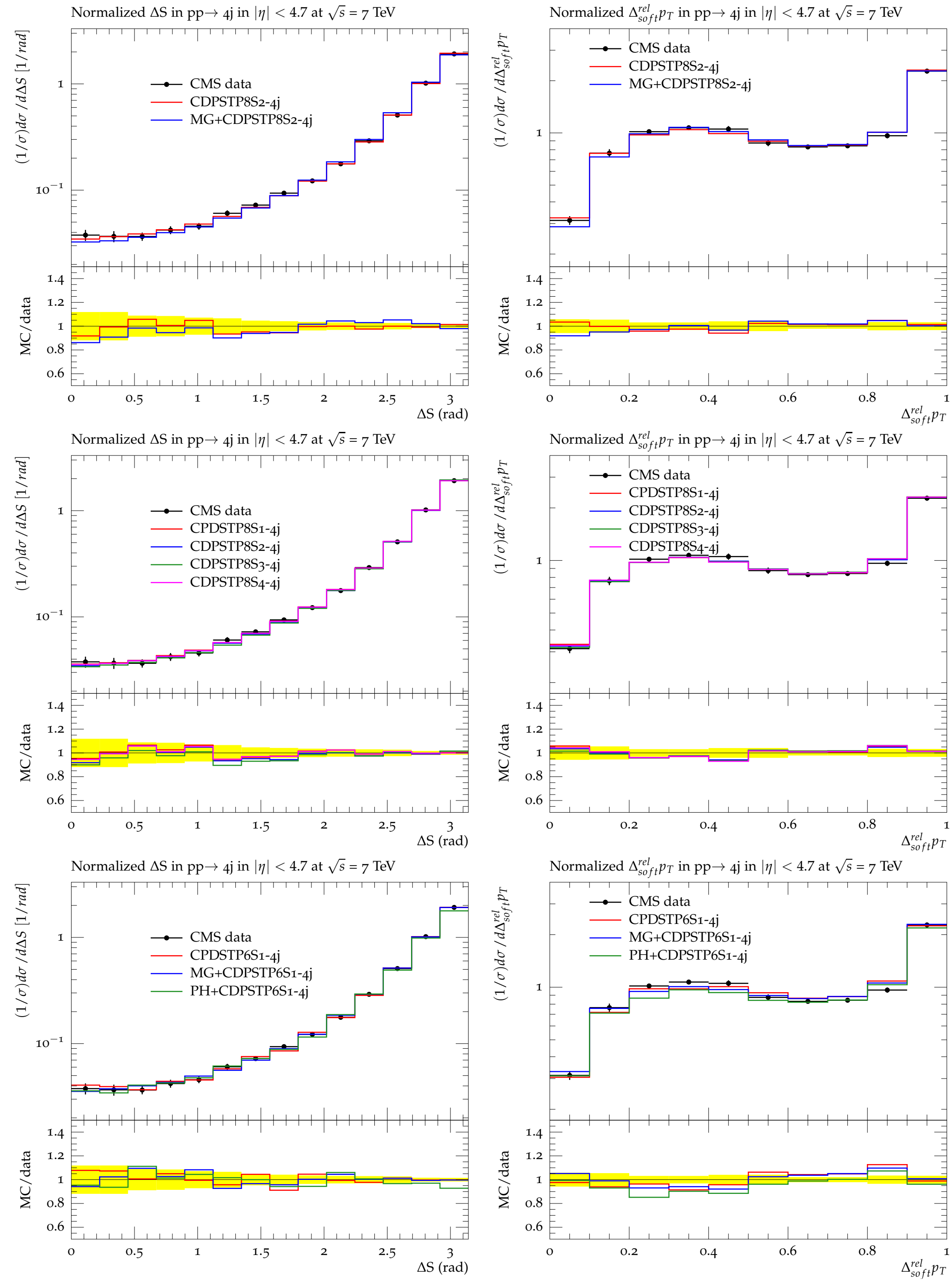


Figure E.1: Comparisons of predictions of the various tunes for ΔS (left column) and $\Delta^{rel}_{soft}p_T$ (right column), measured in the 4j channel; in the top plots, predictions obtained with the tune the CDPSTP8S2-4j interfaced with the ME simulated by Pythia*8 and by* Madgraph *are shown, in the middle plots all new DPS-based tunes are compared to the data, while in the bottom plots, the performance of the new tune CDPSTP6S1-4j are displayed, when the ME is simulated with* Pythia*6,* Madgraph *and* Powheg*. The lower panels show the ratios between MC predictions and data.*

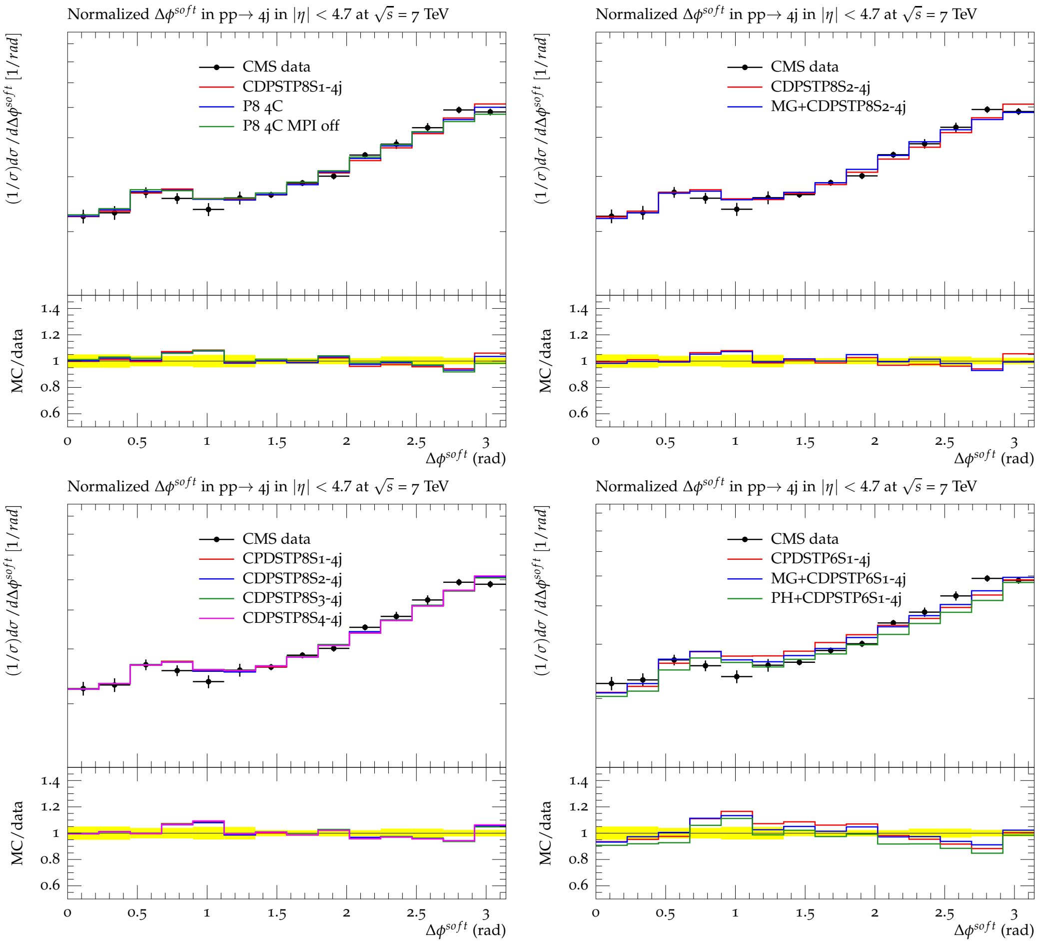


Figure E.2: Comparisons of predictions of various tunes for $\Delta\phi^{soft}$, measured in the 4j channel; in the top left plot, predictions obtained with tune 4C and MPI off, tune 4C and CDPSTP8S1-4j are compared to the data, while in the top right plot, the CDPSTP8S2-4j is considered, interfaced with the ME simulated by Pythia*8 and by* Madgraph*. The bottom left plot shows comparisons of predictions obtained with all new DPS-based tunes with the data, while the bottom right plot displays the performance of the new tune CDPSTP6S1-4j, when the ME is simulated with* Pythia*6,* Madgraph *and* Powheg*. The lower panels show the ratios between MC predictions and data.*

E.4 Additional comparisons for the 2b2j channel

In this Section, the parameters obtained for each tune of the 2b2j channel are listed and results of Chapter 11 are extended by comparisons to a larger set of measured correlation observables. Results of the parameters of the tunes using a negative exponential overlap matter distribution are shown in Table E.10, compared with the ones of tune 4C. The new values show the need for a higher DPS contribution. Furthermore, in CDPSTP8S2-2b2j, the tuned value of p_{T}^{0} is lower than the one in the reference tune 4C: this effect has already been observed in the 4j channel. Table E.11 shows the results for the tune CDPSTP8S3-2b2j, which uses a double gaussian matter distribution function, while in Table E.12, the values of the parameters for CDPSTP8S4-2b2j with a x-dependent overlap are listed. For each extracted tune, the values of the reduced χ^2 are also shown: they range between 0.14 and 0.38. As seen for the 4j channel, these low values of χ^2 come from the large systematic uncertainties in the measurements.

Parameter	CDPSTP8S1-2b2j	CDPSTP8S2-2b2j	4C
MultipleInteractions:pT0Ref	2.085*	2.169204	2.085
MultipleInteractions:ecmPow	0.19*	0.3231304	0.19
MultipleInteractions:expPow	1.461717	0.8882267	2.0
BeamRemnants:reconnectRange	1.5*	8.684605	1.5
Reduced χ^2	0.358	0.381	-

Table E.10: Parameters obtained for the new DPS-based tunes of Pythia*8, performed with the CTEQ6L1 PDF set, with a negative exponential overlap matter distribution. Parameters have been tuned to the ΔS observable measured in the 2b2j channel. Results of a partial tune where only the exponent of the overlap matter distribution has been changed are shown in the second column, while the results of the full tune where all the UE parameters have been varied, are listed in the third column. The parameters of the partial tune accompanied by a star, are the ones which have been kept fixed at the values set in the reference tune 4C. Values of the parameters as implemented in the old tune 4C are listed in the fourth column. The χ^2 obtained in the fit is also listed for each new tune.*

Parameter	CDPSTP8S3-2b2j
MultipleInteractions:pT0Ref	1.946435
MultipleInteractions:ecmPow	0.1315723
MultipleInteractions:coreRadius	0.7575266
MultipleInteractions:coreFraction	0.5652114
Reduced χ^2	0.142

Table E.11: Parameters obtained for the new DPS-based tune of Pythia*8, performed with the CTEQ6L1 PDF set, with a double gaussian overlap matter distribution. Parameters have been tuned to the ΔS observable measured in the 2b2j channel. The χ^2 obtained in the fit is also listed for the new tune.*

Figure E.3 shows the comparisons on ΔS between the data and the predictions of all new tunes. All of them are able to reproduce the data very well over the whole phase space. The new tunes are also tested on correlation observables not used in the fit and results are shown in Figures E.4 and E.5. The $\Delta\phi^{\mathrm{light}}$ and $\Delta\phi^{\mathrm{bottom}}$ variables are quite well described by the predictions of the new tunes: from the comparisons between predictions of tunes with and without MPI, they appear not to be sensitive to a DPS signal. Instead, $\Delta^{\mathrm{rel}}_{\mathrm{light}}p_{\mathrm{T}}$ and $\Delta^{\mathrm{rel}}_{\mathrm{bottom}}p_{\mathrm{T}}$ exhibit a higher contribution from MPI: predictions of tune 4C with MPI switched off fail to describe the measurement, especially at low values of $\Delta^{\mathrm{rel}}p_{\mathrm{T}}$, where the DPS signal is expected. The measured normalized cross sections,

Parameter	CDPSTP8S4-2b2j
MultipleInteractions:pT0Ref	1.983348
MultipleInteractions:ecmPow	0.2890126
MultipleInteractions:a1	0.9185178
BeamRemnants:reconnectRange	5.703317
Reduced χ^2	0.27

Table E.12: Parameters obtained for the new DPS-based tunes of Pythia*8, performed with CTEQ6L1 PDF set, with a single gaussian x-dependent overlap matter distribution. Parameters have been tuned to the ΔS observable measured in the 2b2j channel. The χ^2 obtained in the fit is also listed for the new tune.*

as a function of $\Delta^{\mathrm{rel}} p_{\mathrm{T}}$ of both jet pairs, are difficult to be reproduced by the new tunes: in particular, all of them overestimate the lower part of the spectrum, at values between 0.2 and 0.4, and underestimate the region at values between 0.5 and 0.8. This is the reason why it has been decided to exclude the measurements of $\Delta^{\mathrm{rel}}_{\mathrm{light}} p_{\mathrm{T}}$ from the tuning procedure, contrary to the 4j channel. It has been observed that strong modifications of the obtained tunes appear in case $\Delta^{\mathrm{rel}}_{\mathrm{light}} p_{\mathrm{T}}$ is fitted together with ΔS; this is because the tune tries to fit the "critical" regions of $\Delta^{\mathrm{rel}}_{\mathrm{light}} p_{\mathrm{T}}$, with the result of worsening the description of the other observables.

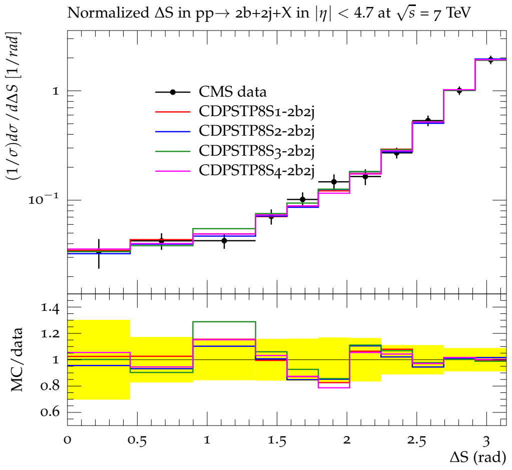


Figure E.3: Comparisons of predictions obtained with all new DPS-based tunes with the correlation observable ΔS, measured in the 2b2j channel. The lower panel shows the ratios between MC predictions and data.

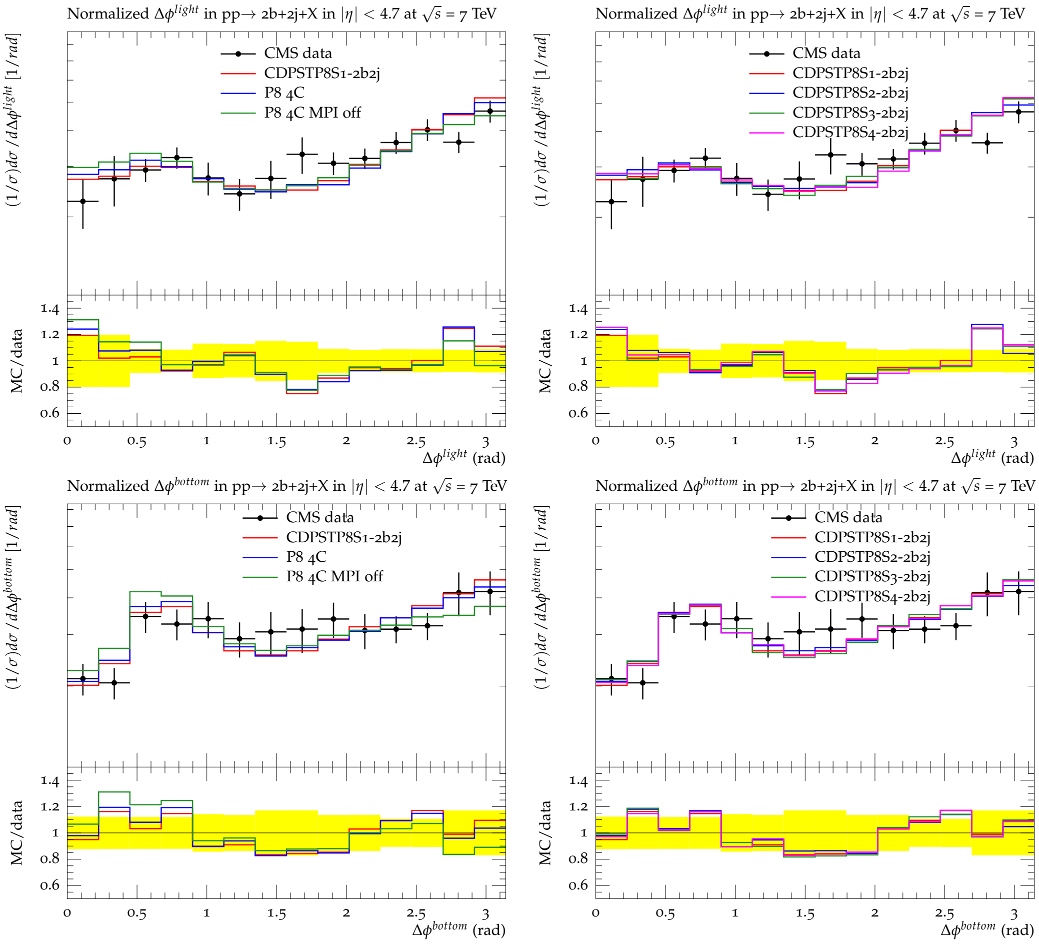


Figure E.4: Comparisons of predictions obtained with the new DPS-based tunes with the correlation observables $\Delta\phi^{light}$ (top), $\Delta\phi^{bottom}$ (bottom) measured in the 2b2j channel; in the left plots, predictions obtained with the tune 4C without MPI simulated and with the tunes 4C and CDPSTP8S1-2b2j are compared to the data, while in the right ones, all the new tunes, CDPSTP8-2b2j, are considered. The lower panels show the ratios between MC predictions and data.

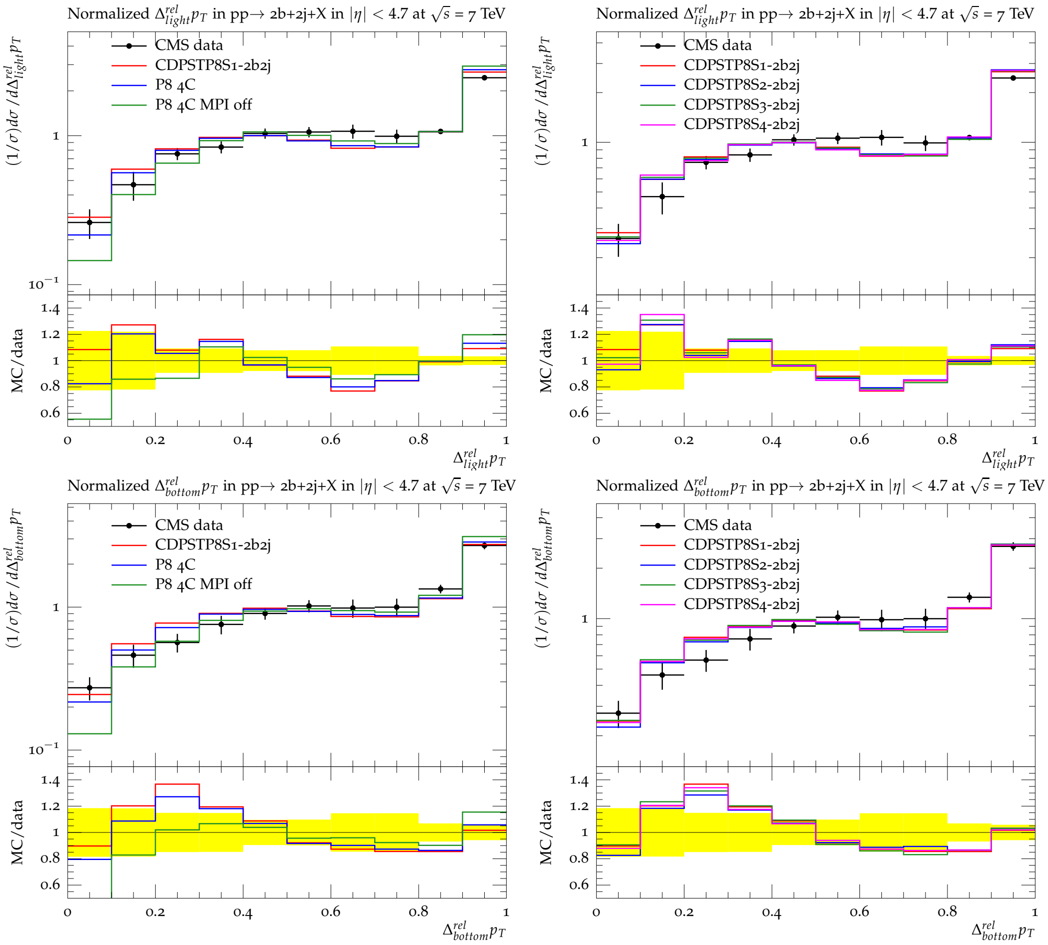


Figure E.5: Comparisons of predictions obtained with the new DPS-based tunes with the correlation observables $\Delta^{rel}_{light}p_T$ (top) and $\Delta^{rel}_{bottom}p_T$ (bottom) measured in the 2b2j channel; in the left plots, predictions obtained with the tune 4C without MPI simulated and with the tunes 4C and CDPSTP8S1-2b2j are compared to the data, while in the right ones, all the new tunes, CDPSTP8-2b2j, are considered. The lower panels show the ratios between MC predictions and data.

Acknowledgements

Finally, I would like to express my gratitude to the people whom I had alongside during this time and without whom, I'm sure, this thesis would have not been written.

First of all, I would like to thank Prof. Dr. Hannes Jung for teaching me how to do research in a very professional and systematic way; thanks for giving me the opportunity to work in this exciting discipline and to develop my own physics interests and research lines. He has conveyed me a deep passion towards this field of science and taught me the right approach to investigation and knowledge. I really appreciate how he leads his working group and I'm very proud and happy to be a part of it.

I would also like to thank my second advisor Prof. Dr. Pierre Van Mechelen for accepting me in the CMS group at the University of Antwerp and leading my research during my staying there: I really benefitted from fruitful discussions with him, by getting precious insights for my analysis and significant improvements for my thesis during the correction phase. Special thanks also to Prof. Dr. Johannes Haller, Prof. Dr. Nick Van Remortel and Prof. Dr. Dieter Horns for promptly accepting to be referees of my thesis and disputation. Thanks to Sarah Van Mierlo for her huge patience and kindness in taking care of the bureaucratic part of my joint PhD and the effort dedicated for making it possible.

Thanks to Samantha Dooling and Pedro Cipriano, my nice and familiar officemates: I have been sharing happy and less joyful moments with them, always finding good friends to speak with, as well as important sources of precious physics advice. It was a pleasure for me to spend my daily life with them. Thanks to Dr. Albert Knutsson for the help in the analysis and the very valuable discussions about his baby, RIVET: I had big profit from his programming skills and physics expertise. Thanks to Dr. Anastasia Grebenyuk for the great time in Hamburg and Antwerp spent together, as well as for the huge help with many details of my thesis. Thanks to Dr. Panagiotis Katsas, former supervisor of my summer student program and very crucial guide afterwards, during my PhD: my path in DESY would not have been possible without him. Thanks to Prof. Dr. Francesco Hautmann for the nice and interesting discussions about many physics aspects I treated in my thesis: I am really happy to have profitted from his expertise which helped me a lot on proceeding in the analyses. Thanks to Dr. Tomasz Fruboes, Juan Grados, Martin Van der Driesch, Dr. Hans Van Haevermaet, Dr. Benoit Roland and Dr. Vardan Khachatryan for creating a stimulating and productive environment inside the CMS QCD group at DESY and sharing their knowledge with me during the time of the thesis: it was really great to have such nice and well prepared colleagues and friends and to get always great solutions to any physics problem or doubt that I had to face. Specifically, thanks to Hans for translating the abstract in Dutch and to Benoit and Vardan for very helpful comments of some chapters of the thesis. Thanks to Gregor Mittag for helping me with the German title and abstract. Thanks to Merjin Van der Klundert, Alex Van Spilbeeck, Tom Cornelis and the rest of the Antwerp crew: the friendly environment helped me a lot during the delicate months of writing the thesis. Separate mention is reserved for the hilarious time in the crazy office U.202.

Thanks to Alvin and Ivan for the great time together at the discovery of Hamburg from the early days until now. The transition between Perugia and Hamburg would not have been so great and smooth without their smiles and jokes. Thanks to Alessandra, Francesco, Francesca, Luigi, Marco B., Marco F., Martina, Matteo, Ninetta, Sara and Valentina for having always time for coffees, breaks, nice chats and evenings. I am very glad to have shared and lived my "italianity" with them. Thanks to Marco T. for his sincere and warm friendship. I have been always impressed by his bright words, full of light and faith: he has really something

special inside (and I know what it is) and I had the fortune to have a taste of that through him. Thanks to Stefano for being such a selfless and considerate friend; I would like to say "thanks" to him, as many times as he said "sorry" to me! Thanks to Peter for being such a solar friend: the hand shaking and German hugs were my ways to communicate that. Thanks to Annika, Andrei, Artur, Carmen, Halil, Jan, Jasone, Maikel, Martin S., Martin V., Nezhla, Özgur, Somnath, Tigran, Wolf, the whole DESY football group and the German and Spanish classes: time in DESY and in Hamburg has been really an enjoyable experience thanks to their friendship and the time spent with them.

Thanks to the little Italy in Belgium, "Italiani ad Anversa", for the nice time in Antwerp: great trips, parties and dinners in Italian style. All of them made my staying very pleasant and full of nice memories. Particular acknowledgement to Luca, mate of great experiences at the discovery of Flanders.
Big thanks to the "Pacchetto Fisici on the world": during these three years, the wonderful reunions all over Europe have been unforgettable and gorgeous moments. It was always like the old times and if the wise man is right, "after 8 years, friends will be forever in your life". Well, we have reached that point :) .
Thanks to all my (unfortunately far) friends in Perugia: they always gave me an energetic recharge during my holidays in Italy. Special thought to Celine, Leandro, Chiara, Giacomo and Michael (even though, lazy as they are, they have never visited me anywhere abroad!).

Thanks to my wonderful parents: they have always supported me with warm love, bearing patiently my complaints and joying with me for any little achievement. I'm very grateful to them for managing to be constantly present even from afar, by having always time for me, by giving me hearty advice and precious help and by understanding my mood and my feelings from the very first word I pronounce.
Thanks to my brother Marco and his family, Letizia, Pietro, Maria and Michele: I have continuously felt their presence and encouragement, strictly necessary for me to go ahead and do my best. They still represent my model of family and life I would like to achieve and towards which my eyes are directed.

And last but not least. Thanks to Jennifer, my anchor and lifemate: she has kindly taken care of me during stressful periods and enjoyed with me joyful moments. Staying with her is always the best choice of spending my time. I thank her for being as she is and for letting the best part of me come out. Das Größte für mich.

Queste persone, che mi hanno accompagnato durante tutta questa fase della vita sono state per me fondamentali ed essenziali, il meglio che potessi avere a disposizione: senza di loro, i momenti che ho vissuto non sarebbero stato cosí belli ed emozionanti; che lo vogliano o no, tutti loro rimarranno per sempre nel mio cuore. Grazie!

Bibliography

[1] Halzen F. and Martin A.D. *Quarks and leptons: an introductory course in modern particle physics.* (ISBN-9780471887416), 1984. 1, 5, 8

[2] Halzen F. et al. *Evidence for Multiple Parton Interactions From the Observation of Multi - Muon Events in Drell-Yan Experiments.* *Phys.Lett.*, B188:375–378, 1987. doi: 10.1016/0370-2693(87)91400-6. 1, 33

[3] Berger E. et al. *Characteristics and Estimates of Double Parton Scattering at the Large Hadron Collider.* *Phys.Rev.*, D81:014014, 2010. doi: 10.1103/PhysRevD.81.014014. 1, 33

[4] Godbole R.M. et al. *Double parton scattering contribution to W+jets.* *Z.Phys.*, C47:69, 1990. doi: 10.1007/BF01551914. 1, 33

[5] AFS Collaboration. *Double parton scattering in pp collisions at* $\sqrt{s}$ *= 63 GeV.* *Z. Phys.*, C34:163, 1987. 1, 35, 191

[6] Abe F. et al. *Study of four jet events and evidence for double parton interactions in* $p\bar{p}$ *collisions at* $\sqrt{s} = 1.8$ *TeV.* *Phys.Rev.*, D47:4857–4871, 1993. doi: 10.1103/PhysRevD.47.4857. 1, 191

[7] D0 Collaboration. *Double parton interactions in photon+3 jet events in* $p\bar{p}$ *collisions at* $\sqrt{s}$ *= 1.96 TeV.* *Phys.Rev.*, D82, 2010. 1, 35, 178

[8] Chatrchyan S. et al. *Study of double parton scattering using W + 2-jet events in proton-proton collisions at* $\sqrt{s}$ *= 7 TeV.* *JHEP*, 1403:032, 2014. doi: 10.1007/JHEP03(2014)032. 1, 35, 36, 72, 177, 181, 183, 185, 191

[9] Aad G. et al. *Measurement of hard double-parton interactions in* $W(\to l\nu)+$ *2 jet events at* $\sqrt{s}$*=7 TeV with the ATLAS detector.* *New J.Phys.*, 15:033038, 2013. doi: 10.1088/1367-2630/15/3/033038. 1, 35, 177, 185, 191

[10] CMS Collaboration. *Measurement of the 4-jet production at the CMS experiment.* (CMS-PAS-FSQ-12-013), 2013. 3, 146

[11] Chatrchyan S. et al. *Measurement of four-jet production in proton-proton collisions at* $\sqrt{s}$ *= 7 TeV.* *Phys.Rev.*, D89:092010, 2014. doi: 10.1103/PhysRevD.89.092010. 3, 146, 162, 163, 165, 166

[12] CMS Collaboration. *Underlying Event Tunes and Double Parton Scattering.* (CMS-PAS-GEN-14-001), 2014. 3, 154, 174, 183, 186, 189, 236, 242, 243, 244, 247, 248, 249, 251

[13] Dooling S. et al. *Longitudinal momentum shifts, showering, and nonperturbative corrections in matched next-to-leading-order shower event generators. Phys.Rev.*, D87(9):094009, 2013. doi: 10.1103/PhysRevD.87.094009. 3, 216, 218

[14] Cipriano P. et al. *Higgs as a gluon trigger. Phys.Rev.*, D88:097501, 2013. doi: 10.1103/PhysRevD.88.097501. 3, 207, 209

[15] Taylor C.C.W. *The atomists, Leucippus and Democritus: fragments, a text and translation with a commentary.* University of Toronto Press Incorporated, 1999. ISBN 0-8020-4390-9 pp.157-158. 5

[16] Wilson E. *An Introduction to Scientific Research.* McGraw-Hill, 1952. 5

[17] Thomson J.J. *Cathode rays. Phil.Mag.*, 44:293–316, 1897. doi: 10.1080/14786449708621070. 5

[18] Dahl P. *Flash of the Cathode Rays: A History of J.J.Thomson's Electron.* IOP Publishing Ltd, 1997. 5

[19] Geiger H. and Marsden E. *On a Diffuse Reflection of the α particles. Proc.R.Soc.Lond.A*, 82, 1909. doi: 10.1098/rspa.1909.0054. 5

[20] Messiah A. *Quantum Mechanics.* Dover Publications, 1999. 5

[21] Chadwick J. *Possible Existence of a Neutron. Nature*, 129:312, 1932. doi: 10.1038/129312a0. 5

[22] Peskin M.E. and Schroeder D.V. *An Introduction to quantum field theory.* (ISBN-9780201503975), 1995. 5

[23] Bettini A. *Introduction to elementary particle physics.* (ISBN-9781107406094), 2008. 6

[24] Goldstein H. *Classical mechanics.* Number ISBN-13:978-0201657029. Zanichelli Italia, 2005. 6

[25] Sakurai J. and Napolitano J. *Modern quantum physics.* (ISBN-9780805382914), 2011. 6

[26] Bloom E.D. et al. *High-Energy Inelastic e p Scattering at 6-Degrees and 10-Degrees. Phys.Rev.Lett.*, 23:930–934, 1969. doi: 10.1103/PhysRevLett.23.930. 7, 13

[27] PBSNOVA/Fermilab/OfficeofScience/USDeptofEnergy. *http://en.wikipedia.org/wiki/Standard_Model.* Standard Model image, modified after the Higgs discovery, 2012. 7

[28] Strominger A. *Five Problems in Quantum Gravity. Nucl.Phys.Proc.Suppl.*, 192-193:119–125, 2009. doi: 10.1016/j.nuclphysbps.2009.07.049. 8

[29] Beringer J. et al. *Review of Particle Physics (RPP). Phys.Rev.*, D86:010001, 2012. doi: 10.1103/PhysRevD.86.010001. 9, 11, 18, 27

[30] Englert F. and Brout R. *Broken Symmetry and the Mass of Gauge Vector Mesons. Phys.Rev.Lett.*, 13:321–323, 1964. doi: 10.1103/PhysRevLett.13.321. 11

[31] Higgs P.W. *Broken Symmetries and the Masses of Gauge Bosons. Phys.Rev.Lett.*, 13:508–509, 1964. doi: 10.1103/PhysRevLett.13.508. 11

[32] Horvth D. *Twenty years of searching for the Higgs boson: Exclusion at LEP, discovery at LHC*. *Mod.Phys.Lett.*, A29:1430004, 2014. doi: 10.1142/S0217732314300043. 12

[33] Chatrchyan S. et al. *Observation of a new boson with mass near 125 GeV in pp collisions at* $\sqrt{s} = 7$ *and 8 TeV*. *JHEP*, 1306:081, 2013. doi: 10.1007/JHEP06(2013)081. 12, 207

[34] Aad G. et al. *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*. *Phys.Lett.*, B716:1–29, 2012. doi: 10.1016/j.physletb.2012.08.020. 12, 207

[35] Ellis R.K. et al. *QCD and collider physics*. *Camb.Monogr.Part.Phys.Nucl.Phys.Cosmol.*, 8: 1–435, 1996. 12, 18

[36] Brock I. *http://brock.physik.uni-bonn.de/∼brock/feynman/ktp_ws1011/chapter14/qgluon_vertex.jpg*. Collection of images of Feynman diagrams, drawn with MN_fit software, developed by the University of Bonn, 2011. 12

[37] Jung H. *QCD and Monte Carlo*. Lecture write-up for QCD and Monte Carlo lectures, 2011. 14, 63

[38] Altarelli G. and Parisi G. *Asymptotic Freedom in Parton Language*. *Nucl.Phys.*, B126:298, 1977. doi: 10.1016/0550-3213(77)90384-4. 14

[39] Gribov V.N. and Lipatov L.N. *Deep inelastic ep scattering in perturbation theory*. *Sov.J.Nucl.Phys.*, 15:438–450, 1972. 14

[40] Moch S. et al. *The Three loop splitting functions in QCD: The Nonsinglet case*. *Nucl.Phys.*, B688:101–134, 2004. doi: 10.1016/j.nuclphysb.2004.03.030. 15

[41] Kuraev E.A. et al. *Multi-Reggeon Processes in the Yang-Mills Theory*. *Sov.Phys.JETP*, 44: 443–450, 1976. 15

[42] Kuraev E.A. et al. *The Pomeranchuk Singularity in Nonabelian Gauge Theories*. *Sov.Phys.JETP*, 45:199–204, 1977. 15

[43] Catani S. et al. *QCD Coherence in Initial State Radiation*. *Phys.Lett.*, B234:339, 1990. doi: 10.1016/0370-2693(90)91938-8. 16

[44] Ciafaloni M. *Coherence Effects in Initial Jets at Small* q^2 */ s*. *Nucl.Phys.*, B296:49, 1988. doi: 10.1016/0550-3213(88)90380-X. 16

[45] Skands P. *Tuning Monte Carlo Generators: The Perugia Tunes*. *Phys.Rev.*, D82:074018, 2010. doi: 10.1103/PhysRevD.82.074018. 16, 154

[46] Andersson B. *The Lund model*. Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology. Cambridge University Press, 1998. ISBN 9780521420945. 17

[47] Osman S. *Multiple parton interactions in deep inelastic ep-scattering at HERA*. 2008. doi: 10.3204/DESY-THESIS-2008-048. 17

[48] Webber B.R. *A QCD Model for Jet Fragmentation Including Soft Gluon Interference*. *Nucl.Phys.*, B238:492, 1984. doi: 10.1016/0550-3213(84)90333-X. 18

[49] Salam G.P. *Towards Jetography*. *Eur.Phys.J.*, C67:637–686, 2010. doi: 10.1140/epjc/s10052-010-1314-6. 19

[50] Salam G.P. and Soyez G. *A Practical Seedless Infrared-Safe Cone jet algorithm*. *JHEP*, 0705:086, 2007. doi: 10.1088/1126-6708/2007/05/086. 19

[51] Cacciari M. et al. The anti-k_t jet clustering algorithm. *JHEP*, 04:063, 2008. doi: 10.1088/1126-6708/2008/04/063. 19, 20

[52] Ellis S.D. and Soper D.E. *Successive combination jet algorithm for hadron collisions*. *Phys.Rev.*, D48:3160–3166, 1993. doi: 10.1103/PhysRevD.48.3160. 20

[53] Dokshitzer Y.L. et al. *Better jet clustering algorithms*. *JHEP*, 9708:001, 1997. doi: 10.1088/1126-6708/1997/08/001. 20

[54] Ellis S.D. et al. *Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches*. *Phys.Rev.*, D81:094023, 2010. doi: 10.1103/PhysRevD.81.094023. 21

[55] CMS Collaboration. *Measurement of the inclusive jet and b-jet production in pp collisions at* $\sqrt{s}=7$ *TeV using Particle Flow*. Internal note about b-jet inclusive cross section measurement, 2010. 22

[56] Quinn H.R. *B Physics and CP violation*. (SLAC-PUB-10415):47–84, 2004. 22

[57] Giunti C. and Kim C.W. *Fundamentals of Neutrino Physics and Astrophysics*. (ISBN-9780198508717), 2007. 22, 23

[58] Trimble V. *Existence and Nature of Dark Matter in the Universe*. *Ann.Rev.Astron.Astrophys.*, 25:425–472, 1987. doi: 10.1146/annurev.aa.25.090187.002233. 22

[59] Peebles P.J.E. and Ratra B. *The Cosmological constant and dark energy*. *Rev.Mod.Phys.*, 75:559–606, 2003. doi: 10.1103/RevModPhys.75.559. 22

[60] 't Hooft G. *Naturalness, chiral symmetry, and spontaneous chiral symmetry breaking*. *NATO Adv.Study Inst.Ser.B Phys.*, 59:135, 1980. 22, 23

[61] Martin S.P. *A Supersymmetry primer*. 1997. hep-ph/9709356. 23

[62] Arkani-Hamed N. et al. *The Hierarchy problem and new dimensions at a millimeter*. *Phys.Lett.*, B429:263–272, 1998. doi: 10.1016/S0370-2693(98)00466-3. 23

[63] Reeb D. *Quantum Gravity Effects through Running of Newton's Constant*. 2009. hep-th/0901.2963. 24

[64] Pomarol A. *Beyond the Standard Model*. 2012. hep-ph/1202.1391. 23

[65] Susskind T. *Video lectures about String Theory and M-Theory*. Stanford University Youtube Channel, March, 2011. 23

[66] Curci G. et al. *Evolution of Parton Densities Beyond Leading Order: The Nonsinglet Case*. *Nucl.Phys.*, B175:27, 1980. doi: 10.1016/0550-3213(80)90003-6. 26

[67] Lai H. et al. *New parton distributions for collider physics*. *Phys.Rev.*, D82:074024, 2010. doi: 10.1103/PhysRevD.82.074024. 26, 162

[68] Pumplin J. et al. *New generation of parton distributions with uncertainties from global QCD analysis*. *JHEP*, 0207:012, 2002. doi: 10.1088/1126-6708/2002/07/012. 26, 66, 162, 180

[69] Aaron F.D. et al. *Combined Measurement and QCD Analysis of the Inclusive* $e^{\pm}p$ *Scattering Cross Sections at HERA*. *JHEP*, 1001:109, 2010. doi: 10.1007/JHEP01(2010)109. 26

[70] Dittmar M. et al. *Parton Distributions*. 2009. hep-ph/0901.2504. 27

[71] Skands P. *Introduction to QCD*. 2012. hep-ph/1207.2389. 27

[72] Ruiz R. *http://home.fnal.gov/ rruiz/FeynLib/*. Feynman diagrams images, produced with the FeynLib and JaxoDraw software, 2011. 28

[73] Andersen J.R. et al. *The SM and NLO Multileg Working Group: Summary report*. pages 21–189, 2010. 29

[74] Hoeche S. et al. *Matching parton showers and matrix elements*. 2006. hep-ph/0602031. 29

[75] Lönnblad L. *Correcting the color dipole cascade model with fixed order matrix elements*. *JHEP*, 0205:046, 2002. doi: 10.1088/1126-6708/2002/05/046. 29

[76] Alwall J. et al. *Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions*. *Eur.Phys.J.*, C53:473–500, 2008. doi: 10.1140/epjc/s10052-007-0490-5. 29, 39, 67

[77] Frixione S. et al. *Matching NLO QCD computations with parton shower simulations: the POWHEG method*. *JHEP*, 07:070, 2007. doi: 10.1088/1126-6708/2007/11/070. 29, 39

[78] Nason P. and Webber B. *Next-to-Leading-Order Event Generators*. *Ann.Rev.Nucl.Part.Sci.*, 62:187–213, 2012. doi: 10.1146/annurev-nucl-102711-094928. 29

[79] Krauss F. *QCD and Monte Carlo tools*. Talk at the CERN-Fermilab Hadron Collider Physics Summer Schools, 2007. 30

[80] Ackerstaff K. et al. *Measurements of flavor dependent fragmentation functions in Z0 → q anti-q events*. *Eur.Phys.J.*, C7:369–381, 1999. doi: 10.1007/s100529901067. 30

[81] Chatrchyan S. et al. *Measurement of the underlying event in the Drell-Yan process in proton-proton collisions at* $\sqrt{s} = 7$ *TeV*. *Eur.Phys.J.*, C72:2080, 2012. doi: 10.1140/epjc/s10052-012-2080-4. 30, 32, 191, 192, 209

[82] Chatrchyan S. et al. *Measurement of the inelastic proton-proton cross section at* $\sqrt{s} = 7$ *TeV*. *Phys.Lett.*, B722:5–27, 2013. doi: 10.1016/j.physletb.2013.03.024. 31

[83] Aad G. et al. *Measurement of the Inelastic Proton-Proton Cross-Section at* $\sqrt{s} = 7$ *TeV with the ATLAS Detector*. *Nature Commun.*, 2:463, 2011. doi: 10.1038/ncomms1472. 31

[84] Kovner A. and Wiedemann U.A. *Taming the BFKL intercept via gluon saturation*. *Nucl.Phys.*, A715:871–874, 2003. doi: 10.1016/S0375-9474(02)01529-4. 31

[85] CTEQ workshop. *From 8 to 14 TeV: What's needed and why?* http://http://tigger.uic.edu/∼varelas/cteq_workshop_2013/, 2013. 31

[86] Acosta D. et al. *The underlying event in hard interactions at the Tevatron $\bar{p}p$ collider.* *Phys.Rev.*, D70:072002, 2004. doi: 10.1103/PhysRevD.70.072002. 31

[87] Field R. et al. *Defining Maximum and Minimum Transverse Regions to Study the Underlying Event in Hard Scattering Processes.* CDF Note CDF/ANAL/MIN_BIAS/CDFR/5626, 2001. 32

[88] Aaltonen T. et al. *Studying the Underlying Event in Drell-Yan and High Transverse Momentum Jet Production at the Tevatron.* *Phys.Rev.*, D82:034001, 2010. doi: 10.1103/PhysRevD.82.034001. 32

[89] Chatrchyan S. et al. *Measurement of the Underlying Event Activity at the LHC with $\sqrt{s} = 7$ TeV and Comparison with $\sqrt{s} = 0.9$ TeV.* *JHEP*, 1109:109, 2011. doi: 10.1007/JHEP09(2011)109. 32, 33, 154, 242

[90] Aad G. et al. *Measurement of underlying event characteristics using charged particles in pp collisions at $\sqrt{s} = 900 GeV$ and 7 TeV with the ATLAS detector.* *Phys.Rev.*, D83:112001, 2011. doi: 10.1103/PhysRevD.83.112001. 32, 33, 188, 242

[91] CMS Collaboration. *Study of the underlying event, b-quark fragmentation and hadronization properties in $t\bar{t}$ events.* (CMS-PAS-TOP-13-007), 2013. 32

[92] Delabat Y. *Monte Carlo study of the underlying event in b-quark pair production with the CMS detector at the LHC.* DESY Summer Student 2013 Report, September, 2013. 32

[93] CMS Collaboration. *Measurement of azimuthal correlations between forward and central jets in proton proton collisions at $\sqrt{s}$=7 TeV.* (CMS-PAS-FSQ-12-008), 2014. 32, 88, 169

[94] Chatrchyan S. et al. *Measurement of pseudorapidity distributions of charged particles in proton-proton collisions at $\sqrt{s} = 8$ TeV by the CMS and TOTEM experiments.* 2014. 32

[95] Chatrchyan S. et al. *Measurement of energy flow at large pseudorapidities in pp collisions at $\sqrt{s} = 0.9$ and 7 TeV.* *JHEP*, 1111:148, 2011. doi: 10.1007/JHEP11(2011)148,10.1007/JHEP02(2012)055. 32, 199, 246

[96] Paver N. and Treleani D. *Multi - Quark Scattering and Large p_T Jet Production in Hadronic Collisions.* *Nuovo Cim.*, A70:215, 1982. doi: 10.1007/BF02814035. 33

[97] Humpert B. and Odorico R. *Multiparton scattering and QCD radiation as sources of four jet events.* *Phys.Lett.*, B154:211, 1985. doi: 10.1016/0370-2693(85)90587-8. 33

[98] Ametller L. et al. *Possible signature of multiple parton interactions in collider four jet events.* *Phys.Lett.*, B169:289, 1986. doi: 10.1016/0370-2693(86)90668-4. 33

[99] Hussein M.Y. *Double parton scattering in associate Higgs boson production with bottom quarks at hadron colliders.* 2007. hep-ph/0710.0203. 33, 207

[100] Kom C.H. et al. *Pair Production of J/psi as a Probe of Double Parton Scattering at LHCb.* *Phys.Rev.Lett.*, 107:082002, 2011. doi: 10.1103/PhysRevLett.107.082002. 33, 35

[101] Calucci G. et al. *Double Parton Scatterings in High-Energy Proton-Nucleus Collisions and Partonic Correlations.* 2013. hep-ph/1309.6201. 34

[102] Blok B. et al. *Hard four-jet production in pA collisions*. *Eur.Phys.J.*, C73(6):2433, 2013. doi: 10.1140/epjc/s10052-013-2433-7. 34

[103] Bahr M. et al. *Extracting sigma_effective from the CDF gamma+3jets measurement*. *JHEP*, 1303:129, 2013. doi: 10.1007/JHEP03(2013)129. 34

[104] Seymour M.H. and Siodmok A. *Extracting sigma_effective from the LHCb double-charm measurement*. 2013. hep-ph/1308.6749. 35

[105] Gaunt J.R. and Stirling W.J. *Double Parton Distributions Incorporating Perturbative QCD Evolution and Momentum and Quark Number Sum Rules*. *JHEP*, 1003:005, 2010. doi: 10.1007/JHEP03(2010)005. 35

[106] Blok B. et al. *pQCD physics of multiparton interactions*. *Eur.Phys.J.*, C72:1963, 2012. doi: 10.1140/epjc/s10052-012-1963-8. 35

[107] Strikman M. *Transverse Nucleon Structure and Multiparton Interactions*. *Acta Phys.Polon.*, B42:2607–2630, 2011. doi: 10.5506/APhysPolB.42.2607. 35

[108] Diehl M. et al. *Correlations in double parton distributions: effects of evolution*. *JHEP*, 1405: 118, 2014. doi: 10.1007/JHEP05(2014)118. 35

[109] Manohar A.V. and Waalewijn W.J. *A QCD Analysis of Double Parton Scattering: Color Correlations, Interference Effects and Evolution*. *Phys.Rev.*, D85:114009, 2012. doi: 10.1103/PhysRevD.85.114009. 35

[110] Diehl M. *Multiple hard scattering and parton correlations in the proton*. 2014. hep-ph/1411.0847. 35

[111] Diehl M. and Schäfer A. *Theoretical considerations on multiparton interactions in QCD*. *Phys.Lett.*, B698:389–402, 2011. doi: 10.1016/j.physletb.2011.03.024. 35

[112] Abe F. et al. *Double parton scattering in $\bar{p}p$ collisions at $\sqrt{s} = 1.8\,TeV$*. *Phys.Rev.*, D56: 3811, 1997. doi: 10.1103/PhysRevD.56.3811. 35, 178, 191

[113] Abazov V.M. et al. *Double parton interactions in photon + 3 jet and photon + b/c jet + 2 jet events in ppbar collisions at $\sqrt{s}$=1.96 TeV*. *Phys.Rev.*, D89:072006, 2014. doi: 10.1103/PhysRevD.89.072006. 35, 178, 191

[114] CMS Collaboration. *Measurement of the Prompt Double J/psi Production Cross Section in pp Collisions at $\sqrt{s} = 7$ TeV*. (CMS-PAS-BPH-11-021), 2013. 35

[115] Sjöstrand T. et al. *PYTHIA 6.4 Physics and Manual*. *JHEP*, 0605:026, 2006. doi: 10.1088/1126-6708/2006/05/026. 38, 41, 154, 159

[116] Sjöstrand T. et al. *A Brief Introduction to PYTHIA 8.1*. *Comput.Phys.Commun.*, 178: 852–867, 2008. doi: 10.1016/j.cpc.2008.01.036. 38

[117] Gieseke S. et al. *Herwig++ 2.5 Release Note*. 2011. hep-ph/1102.1672. 38, 67, 162

[118] Maltoni F. and Stelzer T. *MadEvent: automatic event generation with MadGraph*. *JHEP*, 02:027, 2003. doi: 10.1088/1126-6708/2003/02/027. 39

[119] Alwall J. et al. *MadGraph 5: going beyond*. *JHEP*, 06:128, 2011. doi: 10.1007/JHEP06(2011)128. 39

[120] Nason P. *A new method for combining NLO QCD with shower Monte Carlo algorithms*. *JHEP*, 04:040, 2004. doi: 10.1088/1126-6708/2004/11/040. 39

[121] Oleari C. *The POWHEG-BOX*. *Nucl. Phys. Proc. Suppl.*, 205-206:36, 2010. doi: 10.1016/j.nuclphysbps.2010.08.016. 39

[122] Gleisberg T. et al. *Event generation with SHERPA 1.1*. *JHEP*, 02:007, 2009. doi: 10.1088/1126-6708/2009/02/007. 39, 162

[123] Corke R. and Sjöstrand T. *Multiparton Interactions and Rescattering*. *JHEP*, 01:035, 2010. doi: 10.1007/JHEP01(2010)035. 41, 42

[124] Corke R. and Sjöstrand T. *Interleaved parton showers and tuning prospects*. *JHEP*, 03:032, 2011. doi: 10.1007/JHEP03(2011)032. 42, 66, 162, 239

[125] Sjöstrand T. *New $p_\perp$-ordered showers and Interleaved Multiple Interactions*. Talk at the ATLAS week, 2005. 42

[126] Gieseke S. et al. *Multiple Partonic Interactions in Herwig++*. *Acta Phys.Polon.Supp.*, 6:613–620, 2013. doi: 10.5506/APhysPolBSupp.6.613. 42

[127] Gieseke S. et al. *Colour reconnections in Herwig++*. *Eur.Phys.J.*, C72:2225, 2012. doi: 10.1140/epjc/s10052-012-2225-5. 43, 67, 162

[128] Evans L. and Bryant P. *LHC Machine*. *JINST*, 3:S08001, 2008. doi: 10.1088/1748-0221/3/08/S08001. 45

[129] Myers S. and Picasso E. *The Design, construction and commissioning of the CERN Large Electron Positron collider*. *Contemp.Phys.*, 31:387–403, 1990. doi: 10.1080/00107519008213789. 45

[130] Fernow R.C. *Introduction to Experimental Particle Physics*. (ISBN-13-9780521301701), 1983. 45, 54

[131] Abramowicz H. and Caldwell A. *HERA collider physics*. *Rev.Mod.Phys.*, 71:1275–1410, 1999. doi: 10.1103/RevModPhys.71.1275. 45

[132] CERN press office. *Incident in LHC sector 3-4*. http://press.web.cern.ch/press-releases/2008/09/incident-lhc-sector-3-4, 2008. 46

[133] Bruning O.S. et al. *LHC Design Report. 1. The LHC Main Ring*. (CERN-2004-003-V-1, CERN-2004-003), 2004. 47

[134] CMS Collaboration. *http://cms.web.cern.ch/content/cms-photo-book-and-brochures*. Poster about CERN, 2010. 48

[135] Brice M. *http://www.crviewer.com/targets/090218/lhc10.jpg*. Maximilien Brice © CERN photos, 2008. 48

[136] Aamodt K. et al. *The ALICE experiment at the CERN LHC*. *JINST*, 3:S08002, 2008. doi: 10.1088/1748-0221/3/08/S08002. 47

[137] Aad G. et al. *The ATLAS Experiment at the CERN Large Hadron Collider*. *JINST*, 3: S08003, 2008. doi: 10.1088/1748-0221/3/08/S08003. 47

[138] Chatrchyan S. et al. *The CMS experiment at the CERN LHC*. *JINST*, 3:S08004, 2008. doi: 10.1088/1748-0221/3/08/S08004. 49, 54, 56

[139] Alves A.Jr. et al. *The LHCb Detector at the LHC*. *JINST*, 3:S08005, 2008. doi: 10.1088/1748-0221/3/08/S08005. 49

[140] Adriani O. et al. *The LHCf detector at the CERN Large Hadron Collider*. *JINST*, 3:S08006, 2008. doi: 10.1088/1748-0221/3/08/S08006. 49

[141] Anelli G. et al. *The TOTEM experiment at the CERN Large Hadron Collider*. *JINST*, 3: S08007, 2008. doi: 10.1088/1748-0221/3/08/S08007. 49

[142] Acquistapace G. et al. *CMS, the magnet project: Technical design report*. (CERN-LHCC-97-10), 1997. 49, 57

[143] CMS Collaboration. *http://ippog.web.cern.ch/resources/2011/cms-collaboration-poster-2010-version*. Poster about CMS detector and collaboration, 2010. 50

[144] Wikipedia. *http://en.wikipedia.org/wiki/Pseudorapidity_mediaviewer/File:Pseudorapidity2.png*. Wikipedia webpage about "pseudorapidity", 2007. 50

[145] CMS Collaboration. *https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults*. Public CMS luminosity information, 2007. 51

[146] Zimmermann F. *LHC: The machine*. Presentation at SLAC Summer Institute 2012, 26th July, 2012. 51, 52

[147] Mangano M.L. *The super-LHC*. 2009. CERN-PH-TH-2009-164. 51

[148] CMS Collaboration. *Technical proposal for the upgrade of the CMS detector through 2020*. Technical Report CERN-LHCC-2011-006. LHCC-P-004, CERN, Geneva, Jun 2011. 51

[149] CMS Collaboration. *CMS, tracker technical design report*. (CERN-LHCC-98-06, CMS-TDR-5), 1998. 52

[150] Khachatryan V. et al. *CMS Tracking Performance Results from early LHC Operation*. *Eur.Phys.J.*, C70:1165–1192, 2010. doi: 10.1140/epjc/s10052-010-1491-3. 52

[151] Chatrchyan S. et al. *Alignment of the CMS Silicon Tracker during Commissioning with Cosmic Rays*. *JINST*, 5:T03009, 2010. doi: 10.1088/1748-0221/5/03/T03009. 53

[152] Adam W. et al. *Track reconstruction in the CMS tracker*. (CERN-CMS-NOTE-2006-041, CMS-NOTE-2006-041), 2005. 53

[153] CMS Collaboration. *CMS: The electromagnetic calorimeter. Technical design report*. (CERN-LHCC-97-33, CMS-TDR-4), 1997. 55

[154] Chatrchyan S. et al. *Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at $\sqrt{s} = 7$ TeV.* *JINST*, 8:P09009, 2013. doi: 10.1088/1748-0221/8/09/P09009. 55

[155] CMS Collaboration. *The CMS hadron calorimeter project: Technical Design Report.* Technical Design Report CMS. CERN, Geneva, 1997. 55

[156] Chatrchyan S. et al. *Performance of the CMS Hadron Calorimeter with Cosmic Ray Muons and LHC Beam Data.* *JINST*, 5:T03012, 2010. doi: 10.1088/1748-0221/5/03/T03012. 56

[157] Chatrchyan S. et al. *Performance of CMS Hadron Calorimeter Timing and Synchronization using Test Beam, Cosmic Ray, and LHC Beam Data.* *JINST*, 5:T03013, 2010. doi: 10.1088/1748-0221/5/03/T03013. 56

[158] Bayatian G. et al. *Design, performance and calibration of the CMS forward calorimeter wedges.* *Eur.Phys.J.*, C53:139–166, 2008. doi: 10.1140/epjc/s10052-007-0459-4. 57

[159] CMS Collaboration. *The CMS muon project: Technical Design Report.* Technical Design Report CMS. CERN, Geneva, 1997. 58

[160] Chatrchyan S. et al. *Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV.* *JINST*, 7:P10002, 2012. doi: 10.1088/1748-0221/7/10/P10002. 58

[161] CMS Collaboration. *CMS TriDAS project: Technical Design Report, Volume 1: The Trigger Systems.* Technical Design Report CMS. 58

[162] Sphicas P. *CMS: The TriDAS project. Technical design report, Vol. 2: Data acquisition and high-level trigger.* (CERN-LHCC-2002-026), 2002. 59

[163] Adam W. et al. *The CMS high level trigger.* *Eur.Phys.J.*, C46:605–667, 2006. doi: 10.1140/epjc/s2006-02495-8. 59, 60

[164] Biasini M. et al. *Use of associative memories for L1 triggering in LHC environment.* *JINST*, 7:C08009, 2012. doi: 10.1088/1748-0221/7/08/C08009. 59

[165] Smith W.H. et al. *CMS regional calorimeter trigger jet logic.* pages 238–242, 2001. C01-09-10.12. 59

[166] Tuura L. et al. *CMS data quality monitoring: Systems and experiences.* *J.Phys.Conf.Ser.*, 219:072020, 2010. doi: 10.1088/1742-6596/219/7/072020. 60

[167] CMS Collaboration. *https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookDQM.* CMS DQM workbook twiki page, 2012. 60

[168] D.Cowan. *Statistical data analysis.* Oxford, UK: Clarendon, 1998. 62, 63, 146

[169] James F. *RANLUX: A FORTRAN implementation of the high quality pseudorandom number generator of Luscher.* *Comput.Phys.Commun.*, 79:111–114, 1994. doi: 10.1016/0010-4655(94)90233-X. 63

[170] CERN Program Library. *http://wwwasd.web.cern.ch/wwwasd/cernlib.html.* CERNLIB webpage, 1999. 63

[171] ROOT Collaboration. *ROOT user's guide.* http://root.cern.ch/drupal/content/users-guide, 1995. 63, 86, 121

[172] CMS Collaboration. *https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBook.* CMSSW Workbook Offline Twiki Page, 2008. 63

[173] Worldwide LHC Computing Grid Collaboration. *Understanding for Collaboration in the Deployment and Exploitation of the Worldwide LHC Computing Grid.* CERN-C-RRB-2005-01/Rev. 1, Version 29 March 2011. 64, 86

[174] CMS Collaboration. *https://github.com/cms-sw/cmssw/tree/CMSSW_7_0_X/Configuration/Generator/python.* Git repository for Monte Carlo configuration files, 2013. 64

[175] WLCG Collaboration. *http://wlcg.web.cern.ch.* GRID at CERN webpage, 2008. 64

[176] CMS Collaboration. *Pileup Jet Identification.* (CMS-PAS-JME-13-005), 2013. 65, 79

[177] Soyez G. et al. *Pileup subtraction for jet shapes.* *Phys.Rev.Lett.*, 110(16):162001, 2013. doi: 10.1103/PhysRevLett.110.162001. 65

[178] Agostinelli S. et al. *GEANT4: A Simulation toolkit.* *Nucl.Instrum.Meth.*, A506:250–303, 2003. doi: 10.1016/S0168-9002(03)01368-8. 65

[179] Jain S. *CMS Fast Simulation: A tool for physics searches at the LHC.* *IEEE Nucl.Sci.Symp.Conf.Rec.*, 2010:317–321, 2010. doi: 10.1109/NSSMIC.2010.5873772. 66

[180] Chatrchyan S. et al. *Study of the underlying event at forward rapidity in pp collisions at* $\sqrt{s}$ *= 0.9, 2.76, and 7 TeV.* *JHEP*, 1304:072, 2013. doi: 10.1007/JHEP04(2013)072. 66, 90, 154, 159, 162, 198

[181] Field R. *Min-Bias and the Underlying Event at the LHC.* 2012. hep-ph/1202.0901. 66

[182] HERWIG++ Collaboration. *https://herwig.hepforge.org/trac/wiki/MB_UE_tunes.* Hepforge Herwig++ twiki page, 2010. 67

[183] Sherstnev A. and Thorne R.S. *Parton Distributions for LO Generators.* *Eur.Phys.J.*, C55: 553–575, 2008. doi: 10.1140/epjc/s10052-008-0610-x. 67, 162, 235

[184] SHERPA Collaboration. *https://sherpa.hepforge.org/trac/wiki/SherpaDownloads/Sherpa-1.3.1.* Hepforge Sherpa twiki page, 2010. 67

[185] Buckley A. et al. *Rivet user manual.* *Comput.Phys.Commun.*, 184:2803–2819, 2013. doi: 10.1016/j.cpc.2013.05.021. 68, 129, 238

[186] CMS Collaboration. *How CMS detects particles.* CMS Document 5581-v1, 2012. 76

[187] CMS Collaboration. *Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET.* (CMS-PAS-PFT-09-001), 2009. 77, 78, 79

[188] CMS Collaboration. *Jet Energy Scale performance in 2011.* CMS Internal Note, DP-2012/006, May, 2012. 79, 80

[189] CMS Collaboration. *Commissioning of the Particle-Flow reconstruction in Minimum-Bias and Jet Events from pp Collisions at 7 TeV.* (CMS-PAS-PFT-10-002), 2010. 79, 80

[190] CMS Collaboration. *Offset Energy Correction for Cone Jets.* (CMS-PAS-JME-09-003), 2009. 81

[191] CMS Collaboration. *Determination of the Relative Jet Energy Scale at CMS from Dijet Balance.* (CMS-PAS-JME-08-003), 2009. 81

[192] Chatrchyan S. et al. *Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS. JINST*, 6:P11002, 2011. doi: 10.1088/1748-0221/6/11/P11002. 81, 143, 144

[193] H. Kirschenmann. Determination of the jet energy scale in CMS. *J.Phys.Conf.Ser.*, 404: 012013, 2012. doi: 10.1088/1742-6596/404/1/012013. 81

[194] CMS Collaboration. *Particle Flow Jet Identification Criteria.* CMS Internal Note, AN-2010/003, February, 2010. 82

[195] Galtieri A.B. et al. *Precision measurements of the top quark mass from the Tevatron in the pre-LHC era. Rept.Prog.Phys.*, 75:056201, 2012. doi: 10.1088/0034-4885/75/5/056201. 83

[196] CMS Collaboration. *https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideBTagging-Algorithms.* CMS BPOG workbook twiki page, 2010. 82

[197] CMS Collaboration. *b-Jet Identification in the CMS Experiment.* (CMS-PAS-BTV-11-004), 2012. 82, 117, 118, 120, 122

[198] CMS Collaboration. *Commissioning of b-jet identification with pp collisions at $\sqrt{s} = 7$ TeV.* (CMS-PAS-BTV-10-001), 2010. 82, 120, 121

[199] Chatrchyan S. et al. *Identification of b-quark jets with the CMS experiment. JINST*, 8: P04013, 2013. doi: 10.1088/1748-0221/8/04/P04013. 82, 117, 118, 122

[200] CMS Collaboration. *https://twiki.cern.ch/twiki/bin/viewauth/CMS/BTagPerformanceOP.* CMS BPOG workbook twiki page, 2010. 83, 117

[201] CMS Collaboration. *https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideBTagMCTools.* B-tag MC tools twiki page, 2011. 84

[202] Gunnellini P. *https://indico.cern.ch/event/232850/contribution/2/material/slides/0.pdf.* Internal CMS Presentation about progress in the analysis of two b- and two other-jet final states, 31 October 2013. 85

[203] Lendermann V. et al. *Combining Triggers in HEP Data Analysis. Nucl.Instrum.Meth.*, A604: 707–718, 2009. doi: 10.1016/j.nima.2009.03.173. 87

[204] Chatrchyan S. et al. *Measurement of the Inclusive Jet Cross Section in pp Collisions at $\sqrt{s} = 7$ TeV. Phys.Rev.Lett.*, 107:132001, 2011. doi: 10.1103/PhysRevLett.107.132001. 88, 154, 216, 217, 218

[205] Bahr M. et al. *Herwig++ 2.3 Release Note.* 2008. hep-ph/0812.0529. 90

[206] CMS Collaboration. *https://twiki.cern.ch/twiki/bin/viewauth/CMS/Pileup_2011_Reweighting.* Pile-up reweighting twiki page, 2011. 92

[207] Chatrchyan S. et al. *Inclusive b-jet production in pp collisions at* $\sqrt{s} = 7$ *TeV*. *JHEP*, 1204: 084, 2012. doi: 10.1007/JHEP04(2012)084. 102, 122, 168

[208] Van Mechelen P. *Going beyond the detector level.* Slides at the Forward Jets Meeting, CMS Internal, 23 April, 2010. 110

[209] *Private Conversation with Luca Scodellaro.* 2014. 120

[210] CMS Collaboration. *https://twiki.cern.ch/twiki/bin/viewauth/CMS/BTagSFMethods.* Internal twiki page for Scale Factors recommendation, 2012. 122

[211] CMS Collaboration. *Mistag rate on b-tagging in 2011 data.* Internal note about mistagging rates for the different algorithms, 2011. 122

[212] CMS Collaboration. *https://twiki.cern.ch/twiki/bin/viewauth/CMS/BtagPOG.* Internal CMS twiki page for b-tag scale factor methods, 2012. 122

[213] CMS Collaboration. *Studies of 2 b-jet + 2 jet production in proton-proton collisions at 7 TeV.* Internal note about the analysis FSQ-13-010, 2013. 123

[214] Blobel V. *Unfolding Methods in High-energy Physics Experiments.* (DESY-84-118), 1984. 131

[215] Choudalakis G. *Unfolding in ATLAS.* 2011. hep-ex/1104.2962. 131

[216] D'Agostini G. *A Multidimensional unfolding method based on Bayes' theorem.* *Nucl.Instrum.Meth.*, A362:487–498, 1995. doi: 10.1016/0168-9002(95)00274-X. 132, 133

[217] Hocker A. and Kartvelishvili V. *SVD approach to data unfolding.* *Nucl.Instrum.Meth.*, A372: 469–481, 1996. doi: 10.1016/0168-9002(95)01478-0. 133, 134

[218] Trefethen L.N. and Bau D. *Numerical Linear Algebra.* Philadelphia: Society for Industrial and Applied Mathematics, 1997. ISBN 0-89871-361-7. 133

[219] Adye T. *Unfolding algorithms and tests using RooUnfold.* pages 313–318, 2011. arXiv/1105.1160. 134

[220] CMS Collaboration. *https://twiki.cern.ch/twiki/bin/viewauth/CMS/JetResolution.* Internal CMS twiki page for jet energy resolution correction, 2010. 144

[221] CMS Collaboration. *Measurement of CMS Luminosity.* (CMS-PAS-EWK-10-004), 2010. 146

[222] ATLAS Collaboration. *ATLAS tunes of PYTHIA 6 and Pythia 8 for MC11.* (ATL-PHYS-PUB-2011-009, ATL-COM-PHYS-2011-744), 2011. 154, 238

[223] Chatrchyan S. et al. *Measurement of the Rapidity and Transverse Momentum Distributions of Z Bosons in pp Collisions at* $\sqrt{s} = 7$ *TeV*. *Phys.Rev.*, D85:032002, 2012. doi: 10.1103/PhysRevD.85.032002. 159

[224] ATLAS Collaboration. *Example ATLAS tunes of Pythia8, Pythia6 and Powheg to an observable sensitive to Z boson transverse momentum.* Technical Report ATL-PHYS-PUB-2013-017, CERN, Geneva, Nov 2013. 159

[225] Badger S. et al. *Numerical evaluation of virtual corrections to multi-jet production in massless QCD*. *Comput.Phys.Commun.*, 184:1981–1998, 2013. doi: 10.1016/j.cpc.2013.03.018. 173

[226] Berger C.F. et al. *An Automated Implementation of On-Shell Methods for One-Loop Amplitudes*. *Phys.Rev.*, D78:036003, 2008. doi: 10.1103/PhysRevD.78.036003. 173

[227] Jung H. et al. *The CCFM Monte Carlo generator CASCADE version 2.2.03*. *Eur.Phys.J.*, C70:1237–1249, 2010. doi: 10.1140/epjc/s10052-010-1507-z. 173

[228] Bartalini P. *MPI and DPS studies at CMS*. Talk at the DIS Conference 2013, 23 April 2013. 178

[229] Sjöstrand T. *http://home.thep.lu.se/ torbjorn/pythia81html/Welcome.html*. PYTHIA 8 online manual, 2011. 180

[230] Frankfurt L. et al. *Transverse nucleon structure and diagnostics of hard parton-parton processes at LHC*. *Phys.Rev.*, D83:054012, 2011. doi: 10.1103/PhysRevD.83.054012. 181

[231] Skands P. et al. *Tuning PYTHIA 8.1: the Monash 2013 Tune*. 2014. hep-ph/1404.5630. 190, 235, 246

[232] CMS Collaboration. *Measurement of the Underlying Event Activity at the LHC at 7 TeV and Comparison with 0.9 TeV*. (CMS-PAS-FSQ-12-020), 2012. 191, 236

[233] Alitti J. et al. *A Study of multi-jet events at the CERN $\bar{p}p$ collider and a search for double parton scattering*. *Phys.Lett.*, B268:145–154, 1991. doi: 10.1016/0370-2693(91)90937-L. 191

[234] Sadeh I. *Double parton scattering in four-jet events in pp collisions at 7 TeV with the ATLAS experiment at the LHC*. 2013. hep-ex/1308.0587. 191

[235] Manohar A.V. and Waalewijn W.J. *What is Double Parton Scattering?* *Phys.Lett.*, B713: 196–201, 2012. doi: 10.1016/j.physletb.2012.05.044. 195

[236] Gunnellini P. *https://indico.cern.ch/event/320101/contribution/2/material/slides/0.pdf*. Internal CMS Presentation about 4CX-based tune at the physics comparisons and generator tunes meeting, 25 June 2014. 195

[237] Antchev G. et al. *First measurement of the total proton-proton cross section at the LHC energy of $\sqrt{s}$ =7 TeV*. *Europhys.Lett.*, 96:21002, 2011. doi: 10.1209/0295-5075/96/21002. 195, 246

[238] Henzi R. and Valin P. *Towards a blacker, edgier and larger proton*. *Phys.Lett.*, B132:443–448, 1983. doi: 10.1016/0370-2693(83)90344-1. 195

[239] Bourrely C. et al. *Impact Picture Expectations for Very High-Energy Elastic pp and $p\bar{p}$ Scattering*. *Nucl.Phys.*, B247:15, 1984. doi: 10.1016/0550-3213(84)90369-9. 195

[240] Khachatryan V. et al. *Observation of Long-Range Near-Side Angular Correlations in Proton-Proton Collisions at the LHC*. *JHEP*, 1009:091, 2010. doi: 10.1007/JHEP09(2010)091. 195

[241] Chatrchyan S. et al. *Observation of a diffractive contribution to dijet production in proton-proton collisions at $\sqrt{s}=7$ TeV*. *Phys.Rev.*, D87:012006, 2013. doi: 10.1103/PhysRevD.87.012006. 195

[242] Gladysz-Dziadus E. et al. *CASTOR: Centauro and strange object research in nucleus-nucleus collisions at LHC.* 2002. hep-ex/0209008. 197

[243] Gottlicher P. *Design and test beam studies for the CASTOR calorimeter of the CMS experiment.* *Nucl.Instrum.Meth.*, A623:225–227, 2010. doi: 10.1016/j.nima.2010.02.203. 198

[244] CMS Collaboration. *Measurement of the pseudorapidity and centrality dependence of the very forward energy density in PbPb collisions at* $\sqrt{s}$=*2.76 TeV.* (CMS-PAS-HIN-12-006), 2012. 198

[245] CMS Collaboration. *Calibration of CASTOR: Status and Prospects.* CMS Internal Note, 2011. 199, 203

[246] CMS Collaboration. *Measurement of very forward jets in CASTOR.* CMS Internal Note, 2012. 205

[247] Van Haevermaet H. *Higgs as a gluon trigger.* DIS2014 Conference, Warsaw, 2014. 209

[248] Britzger D. et al. *New features in version 2 of the fastNLO project.* pages 217–221, 2012. doi: 10.3204/DESY-PROC-2012-02/165. 215

[249] Aad G. et al. *Measurement of multi-jet cross sections in proton-proton collisions at a 7 TeV center-of-mass energy.* *Eur.Phys.J.*, C71:1763, 2011. doi: 10.1140/epjc/s10052-011-1763-6. 215, 216

[250] Chatrchyan S. et al. *Measurements of differential jet cross sections in proton-proton collisions at* $\sqrt{s} = 7$ *TeV with the CMS detector.* *Phys.Rev.*, D87(11):112002, 2013. doi: 10.1103/PhysRevD.87.112002. 218

[251] CMS Collaboration. *Jet cross sections with POWHEG*, May, 2012. CMS Internal Note, DP-201. 223

[252] CMS Collaboration. *Measurement of jet cross sections at low transverse momentum in proton-proton collisions at 8 TeV.* (CMS-PAS-FSQ-12-031), 2013. 223

[253] Ball R.D. et al. *Parton distributions with QED corrections.* *Nucl.Phys.*, B877(2):290–320, 2013. doi: 10.1016/j.nuclphysb.2013.10.010. 235

[254] H1 and ZEUS Collaborations. *HERAPDF1.5LO PDF Set with Experimental Uncertainties.* H1prelim-13-141, ZEUS-prel-13-003, October,2013. 235

[255] Field R. *The Tevatron Energy Scan: Findings & Surprises.* published in the proceedings of the International Symposium on Multiparticle Dynamics, Chicago Illinois, September 17,2013. 236

[256] Althoff M. et al. *Determination of* α_s *in First and Second Order QCD From* e^+e^- *Annihilation Into Hadrons.* *Z.Phys.*, C26:157, 1984. 237

[257] Buskulic D. et al. *Properties of hadronic Z decays and test of QCD generators.* *Z.Phys.*, C55: 209–234, 1992. doi: 10.1007/BF01482583. 237

[258] Buckley A. et al. *Systematic event generator tuning for the LHC.* *Eur. Phys. J. C*, 65:331, 2010. doi: 10.1140/epjc/s10052-009-1196-7. 237, 238

[259] Professor Collaboration. *http://professor.hepforge.org/docs.sphinx/scripts/profruncombs.html.* Professor twiki page for the oversampling recommendation, 2011. 238

[260] Professor Collaboration. *http://professor.hepforge.org/docs.sphinx/scripts/prof-tune.html.* Professor twiki page for the eigentune method, 2011. 238

[261] Alderweireldt S. *https://indico.cern.ch/event/260761/contribution/3/material/slides/0.pdf.* Internal CMS Presentation about 4C* tune at the physics comparisons and generator tunes meeting, 3 July 2013. 239

[262] CMS Collaboration. *Underlying Event tunes and Double Parton Scattering.* Internal Note AN-14-072, 2014. 242, 246

[263] Aamodt K. et al. *Charged-particle multiplicity measurement in proton-proton collisions at* $\sqrt{s}$ *= 7 TeV with ALICE at LHC.* *Eur.Phys.J.*, C68:345–354, 2010. doi: 10.1140/epjc/s10052-010-1350-2. 242

[264] Aad G. et al. *Charged-particle multiplicities in pp interactions measured with the ATLAS detector at the LHC.* *New J.Phys.*, 13:053033, 2011. doi: 10.1088/1367-2630/13/5/053033. 242

[265] CMS Collaboration. *Study of the Underlying Event at Forward Rapidity in Proton-Proton Collisions at the LHC.* (CMS-PAS-FWD-11-003), 2012. 242

[266] Aad G. et al. *Properties of jets measured from tracks in proton-proton collisions at center-of-mass energy* $\sqrt{s} = 7$ *TeV with the ATLAS detector.* *Phys.Rev.*, D84:054001, 2011. doi: 10.1103/PhysRevD.84.054001. 242

[267] Khachatryan V. et al. *First Measurement of Hadronic Event Shapes in pp Collisions at* $\sqrt{s} = 7$ *TeV.* *Phys.Lett.*, B699:48–67, 2011. doi: 10.1016/j.physletb.2011.03.060. 242

[268] Aad G. et al. *Kshort and* Λ *production in pp interactions at* $\sqrt{s} = 0.9$ *and 7 TeV measured with the ATLAS detector at the LHC.* *Phys.Rev.*, D85:012001, 2012. doi: 10.1103/PhysRevD.85.012001. 242

[269] Khachatryan V. et al. *Strange Particle Production in pp Collisions at* $\sqrt{s} = 0.9$ *and 7 TeV.* *JHEP*, 1105:064, 2011. doi: 10.1007/JHEP05(2011)064. 242

[270] Aad G. et al. *Measurements of the pseudorapidity dependence of the total transverse energy in proton-proton collisions at* $\sqrt{s} = 7$ *TeV with ATLAS.* *JHEP*, 1211:033, 2012. doi: 10.1007/JHEP11(2012)033. 242

[271] Khachatryan V. et al. *Charged particle multiplicities in pp interactions at* $\sqrt{s} = 0.9$*, 2.36, and 7 TeV.* *JHEP*, 1101:079, 2011. doi: 10.1007/JHEP01(2011)079. 242

[272] CMS Collaboration. *Pseudorapidity distributions of charged particles in pp collisions at* $\sqrt{s}$ *= 7 TeV with at least one central charged particles.* (CMS-PAS-QCD-10-024), 2011. 246

[273] CMS Collaboration. *Measurement of Strange Particle Production in Underlying Events in proton-proton collisions at* $\sqrt{s}$ *= 7 TeV.* (CMS-PAS-QCD-11-010), 2011. 246