ETP-KA/2020-21

Karlsruher Institut für Technologie

Triple-Differential Measurement of the Dijet Cross Section at $\sqrt{s} = 13\,\text{TeV}$ with the CMS Detector

Zur Erlangung des akademischen Grades eines
DOKTORS DER NATURWISSENSCHAFTEN
(Dr. rer. nat.)

von der KIT-Fakultät für Physik des
Karlsruher Instituts für Technologie (KIT)
angenommene

DISSERTATION

von

MSc. Daniel Săvoiu

Referent:	Prof. Dr. Günter Quast
Korreferent:	Priv.-Doz. Dr. Klaus Rabbertz

Tag der mündlichen Prüfung: 10. Juli 2020

Abstract

In light of the large quantity of data collected during the second operational run of the Large Hadron Collider (LHC) at CERN, which has made it possible to perform measurements at unprecedented energies with a high degree of statistical precision, the necessity of understanding and constraining the systematic effects on such measurements has become increasingly important. Precision measurements of jet observables in proton-induced collisions have proven instrumental in constraining the parton distribution functions (PDFs) which describe the internal structure of protons, and which remain one of the largest sources of uncertainty in many analyses performed at hadron colliders.

This thesis presents the first triple-differential measurement of the dijet production cross section performed in proton-proton collisions at a center-of-mass energy of 13 TeV, based on a data sample of 35.9 fb^{-1} recorded by the CMS experiment at the Large Hadron Collider at CERN. The cross section is measured using anti-k_{T} jets with radius parameters of $R = 0.4$ and $R = 0.8$ as a function of the dijet rapidity separation y^*, the total boost of the dijet system y_{b}, and either the average transverse momentum $\langle p_{\text{T}} \rangle_{1,2}$ or the invariant mass m_{jj} of the dijet system as the third variable. This choice of rapidity variables exploits the topology of the dijet system to achieve an increased sensitivity to the proton PDFs.

After accounting for detector-induced systematic effects in a three-dimensional unfolding procedure, the measured spectra are compared to fixed-order theory predictions at next-to-next-to-leading order accuracy in perturbative quantum chromodynamics, obtained using several recent PDF sets. While the data are observed to be described by the theory within the experimental and theoretical uncertainties across a large portion of the phase space, potentially significant deviations are observed in areas where a heightened sensitivity to the PDFs is expected.

Acknowledgements

First of all, I would like to express my sincere thanks to Prof. Günter Quast for offering me the opportunity to contribute to an interesting and active field of research. His attentive guidance, as well as his continued support and encouragement have been invaluable to me. I would especially like to thank him for the research period at CERN, which has been an immensely enriching experience.

I would also like to thank Dr. Klaus Rabbertz for the excellent relationship on both a professional and personal level, and for the insight and expertise he has shared with me over the course of our collaboration. I am particularly grateful for his diligence in advancing the progress of the theory calculations that have constituted a valuable input to this work.

To all my colleagues at the Institute of Experimental Particle Physics, I extend my thanks and appreciation for the pleasant and supportive work environment. Their openness to dialogue has sparked countless interesting discussions on both serious and lighthearted topics and has made my time there very enjoyable. I would also like to thank Denise Müller, Sebastian Wozniewski and Stefan Wunsch for the pleasant time spent together at CERN and the many productive conversations.

For their constant readiness to help with technical matters and to share a portion of their vast knowledge of computing with me, I would like to thank Florian von Cube, Christoph Heidecker, Matthias Schnepf, and all others involved in the development of support of the computing infrastructure at the institute. Many thanks also to Maximilian Horzela, Matthias Schnepf and Sebastian Wozniewski for proofreading.

Finally, to my parents, whose love and unconditional support has been a source of great reassurance and helped me to persevere, I express my most profound gratitude.

Contents

1 Introduction

The question of the fundamental nature of reality has been a source of fascination for philosophers and scientists alike, who have long engaged in observation and deductive reasoning in order to explain the principles behind physical phenomena. In particular, the study of the subatomic world in high-energy particle collisions has proved successful at probing these phenomena in unprecedented detail, establishing the Standard Model of particle physics as one of the most accurate and thoroughly tested scientific theories.

This achievement would not have been possible without the wealth of experimental data recorded by collider experiments during the last decades at ever-increasing energies. At the Large Hadron Collider (LHC), the largest and most powerful particle accelerator built to date, protons are brought to collision at center-of-mass energies in the TeV range. In these collisions, the protons disintegrate, giving rise to hundreds of new particles, which are then recorded with the help of particle detectors.

Between 2015 and 2018, over 150 inverse femtobarns of data from proton collisions at a center-of-mass energy of 13 TeV were collected by the CMS experiment at the LHC. The large amount of data, which corresponds to roughly five times that collected at 7 and 8 TeV during the earlier data taking periods, highlights the importance of gaining a precise understanding of the experimental and theoretical uncertainties, which are becoming the limiting factor in many analyses. At the same time, it provides an excellent premise for the precision measurements required to further constrain the underlying systematic uncertainties.

One of the largest sources of uncertainty at the LHC is due to the composite nature of protons, which are bound systems of quarks and gluons held together by the strong interaction. Despite the well-established description of the strong interaction given by the theory of quantum chromodynamics (QCD), the proton structure, as described by the parton distribution functions (PDFs), cannot be derived from first principles in the

framework of perturbative QCD, and must instead be determined from the comparison of experimental data to predictions obtained from theoretical models.

Jet observables in hadron-induced processes are particularly well-suited for this purpose, with the abundance of jets at hadron colliders making precise differential measurements of the jet cross section possible across a wide kinematic range. This is further supported by recent progress made in calculating theory predictions for these processes at next-to-next-to-leading order (NNLO) in perturbative QCD [1–3].

The analysis presented in this thesis concerns a triple-differential measurement of the dijet production cross section as a function of three variables: the dijet rapidity separation y^*, the total boost of the dijet system y_b, and a third kinematic variable, chosen to be either the average transverse momentum of the two jets $\langle p_\mathrm{T} \rangle_{1,2}$ or their invariant mass m_jj.

This choice of rapidity variables is particularly advantageous for PDF determinations, since it enables a better separation of the two main effects that contribute to the overall cross section: while y^* is representative of the scattering angle in the dijet rest frame, and thus mainly determined by the scattering matrix element, y_b is determined by the fractions x of the scattering parton momenta, and is thus sensitive to the PDFs.

The enhanced constraining power of such triple-differential measurements was explored in pioneering studies [4, 5] using the CMS 8 TeV data. In a systematic comparison of jet data sets from multiple experiments to NNLO theory predictions, a recent analysis [6] revealed the measurement to have a significant contribution to the PDF determination. This underscores the renewed interest in such measurements at higher energies and with larger data sets, extending the accessible phase space and enhancing the statistical precision.

In the following chapters, the theoretical and experimental aspects relevant for the measurement are introduced. An overview of the theoretical background is given in chapter 2, introducing the Standard Model of particle physics and the basic concepts of QCD. This is complemented by a discussion of the theoretical tools used to model scattering processes at hadron colliders, with a particular emphasis placed on jets.

Chapter 3 presents the experimental setup, providing a description of the Large Hadron Collider and the CMS detector. In chapter 4, the experimental techniques employed at CMS for the reconstruction and calibration of jets are described.

The main cross section measurement is presented in chapter 5, providing a detailed description of the analyzed data samples, the selection of events and the reconstruction of the triple-differential dijet spectrum. A three-dimensional unfolding procedure is used to correct the measurement for detector effects, and the systematic uncertainties on the measurement are studied. The unfolded cross sections are compared to simulations and fixed-order perturbative QCD calculations at NNLO.

2 Theoretical foundations

The endeavor to describe the fundamental nature of matter and provide an explanation for physical phenomena in a unified theory has been the driving force behind many scientific advances in the field of particle physics. Developed over the course of many years in the latter part of the 20$^{\text{th}}$ century, the Standard Model of particle physics provides an understanding of these phenomena in terms of elementary particles and the fundamental interactions between them: the *strong*, *weak* and *electromagnetic* interactions.

Of these, the strong interaction described by the theory of quantum chromodynamics (QCD) is perhaps the most intriguing, describing hadrons such as the protons and neutrons in atomic nuclei as composite systems comprised of color-charged quarks and gluons, which cannot be observed in isolation. When separated from their parent hadrons in high-energy interactions, these particles fragment and recombine to a multitude of hadrons, concentrated in bundled streams called *jets*, whose experimental observation was instrumental in developing an understanding of the strong interaction.

In this chapter, the theoretical concepts developed as part of the Standard Model are introduced, beginning with a brief overview of the particles and interactions in section 2.1. Section 2.2 reviews the quantum field theory of *quantum chromodynamics* (QCD), starting with the Lagrangian formulation and the perturbative framework of calculating predictions for observables. The particular properties of QCD such as asymptotic freedom and confinement are discussed, touching upon the topics of renormalization and the running of the strong coupling. In addition, the parton model of hadrons is discussed, introducing the formalism of collinear factorization for the calculation of observables in hadron collisions, and the description of the proton structure in terms of *parton distribution functions* (PDFs). More detailed introductions on these topics can be found in references [7–10]. Finally, section 2.3 discusses jets as essential tools for the study of QCD and introduces the aspects of Monte Carlo event generation relevant for the modeling of jet observables.

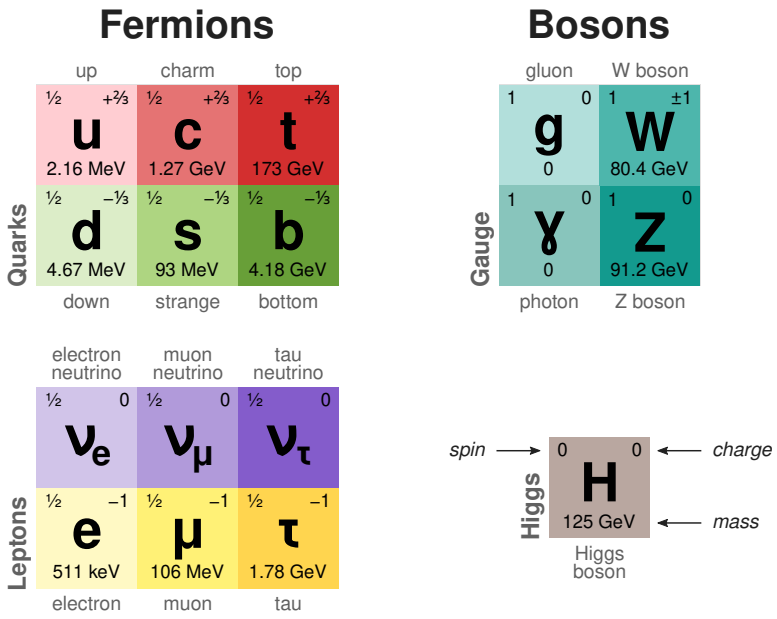

Figure 2.1 – Overview of the particles of the Standard Model.

2.1 The Standard Model

The Standard Model (SM) of particle physics is a unified theoretical model that describes the elementary particles and their interactions based on the fundamental principles established by quantum field theory (QFT). In this framework, particles emerge as localized excitations of underlying quantum fields, which extend throughout all spacetime. The properties of these particles and their interactions arise from the fundamental symmetries of these fields. With the notable exception of *gravity*, the SM provides a description of the remaining known fundamental forces: the *electromagnetic*, *strong* and *weak* interactions.

Particles in the SM are divided into categories based on their quantum numbers, which characterize their participation in the fundamental interactions. The *fermions*, which have half-integer spin, are the fundamental constituents of matter and are further subdivided into the strongly-interacting *quarks* and the *leptons*, which do not participate in the strong interaction. They are arranged into three *generations* or *families*, with particles within each generation distinguished by their electric charge, and the masses of particles generally increasing with each generation.

The particles with full-integer spins are called *bosons* and comprise the four mediator particles associated with the fundamental interactions called the *gauge bosons*, and the scalar *Higgs boson*. For every particle there is a corresponding antiparticle with identical properties, but opposite electric charge, with some particles being their own antiparticles. A representation of the Standard Model particles can be seen in figure 2.1.

As in classical field theory, the dynamic evolution of fields is described by equations of motion derived via the Euler-Lagrange formalism from a postulated *Lagrangian density*. For a relativistic, noninteracting quantum field theory, the fermionic fields with a spin of 1/2 are subject to the Dirac equation, while scalar (spin-0) fields obey the Klein-Gordon equation. Interactions between the fields are introduced through the principle of *local gauge invariance*. Requiring that the SM Lagrangian remain invariant under local gauge transformations leads to the introduction of gauge fields, which couple to the others, resulting in a full interacting theory.

The mathematical structure of the symmetries associated with the fundamental interactions in the SM are described by the composite Lie group denoted SU(3) × SU(2) × U(1). The SU(3) component describes the strong interaction, which affects particles that carry color charge. The theory of these interactions is called *quantum chromodynamics* (QCD) and is described in more detail the following section.

The remaining two components describe the unified electroweak interaction, which is associated with an SU(2)×U(1) symmetry. However, directly imposing such a symmetry would require the electroweak gauge bosons to be massless, which is contrary to the observation of the large W and Z boson masses.

Instead, these bosons acquire mass dynamically through the *Brout–Englert–Higgs mechanism* [11, 12] by coupling to an underlying scalar SU(2) field called the *Higgs field*. Due to the nonzero vacuum expectation value of this field, the electroweak symmetry is spontaneously broken, resulting in three massive vector bosons, corresponding to three of the four degrees of freedom of the Higgs field. The fourth vector boson, the photon, does not gain a mass, resulting in the $\mathrm{U}(1)_{\mathrm{em}}$ symmetry associated with the electromagnetic interaction. Instead, the remaining degree of freedom gives rise to a real scalar field, which manifests itself as the physical *Higgs boson*, whose discovery at the LHC in 2012 [13, 14] provided experimental confirmation of the last remaining particle predicted by the Standard Model.

2.2 Quantum chromodynamics (QCD)

The interaction of color-charged particles such as quarks and gluons by means of the strong force is described by the theory of *quantum chromodynamics* (QCD). Mathematically, QCD is a gauge theory whose interaction structure is described by the nonabelian Lie group SU(N_{C}), where $N_{\mathrm{C}} = 3$ refers to the number of color charges. Summing over repeated indices, the QCD Lagrangian can be written as:

$$\mathcal{L} = \sum_q \bar{\psi}_{q,a} \left(\mathrm{i}\gamma^\mu \partial_\mu \delta_{ab} - g_{\mathrm{s}} \gamma^\mu t^C_{ab} A^C_\mu - m_q \delta_{ab} \right) \psi_{q,b} - \frac{1}{4} F^A_{\mu\nu} F^{A\mu\nu} \tag{2.1}$$

In the above, the quark fields are represented by the spinors $\psi_{q,a}$, with the sum running over each of the six quark flavors q. The indices a and b run from 1 to N_{C} and refer to

the distinct color states occupied by quarks. The γ_μ denote the Dirac matrices. Gluons are represented by the vector fields A_μ^C, where C runs from 1 to $N_\mathrm{C}^2 - 1 = 8$, and t_{ab}^C are the 8 generators of the SU(3) group. The quantity g_s is the *coupling constant*, which is a dimensionless parameter that determines the strength of QCD interactions, and δ_{ab} is the Kronecker symbol.

The gluon field strength tensor $F_{\mu\nu}^A$ is given by

$$F_{\mu\nu}^A = \partial_\mu A_\nu^A - \partial_\nu A_\mu^A - g_\mathrm{s}\, f^{ABC}\, A_\mu^B A_\nu^C\,, \tag{2.2}$$

with f^{ABC} denoting the structure constants of the group, which determine the relation between the group generators and their commutators:

$$[t^A, t^B] = \mathrm{i} f^{ABC}\, t^C \tag{2.3}$$

Since SU(3) is a nonabelian group, the above commutators do not vanish and the associated structure constants are nonzero. This results in a net color charge for gluons, which in turn leads to an interaction of the gluon field with itself. This distinguishes QCD from other, abelian quantum field theories such as quantum electrodynamics (QED), where the gauge field quanta – the photons – are electrically neutral.

The implications of color-charged field quanta for the behavior of strongly interacting systems can be understood in terms of a polarization of the vacuum in the presence of a color-charged particle. As in QED, where a bare electric charge is screened by the creation of virtual electron–positron pairs from the vacuum, virtual quark–antiquark pairs lead to a screening of color charge. However, since gluons carry color charge themselves, the additional polarization contribution from gluons leads to an antiscreening effect, thereby enhancing the effective charge.

The magnitude of this effect depends on the energy scale of the interaction. At high energies, which correspond to short distance scales, the screening of color charges predominates, meaning that quarks behave essentially as free particles. This is called *asymptotic freedom*.

Conversely, as the energy scale decreases (or, equivalently, at high distances), antiscreening becomes the dominant contribution, leading to an increase in the effective charge. This is related to a second notable property of QCD called *color confinement*, which prevents particles that carry color charge from existing as individual free particles. Instead, quarks and gluons are confined to color-neutral bound states called *hadrons*. When bound quarks are separated from each other, the energy density of the mediating field increases with distance so that it becomes energetically more favorable for additional color-charged particles to be created from the vacuum. In the context of high-energy scattering of hadrons, the fragmentation of quarks and gluons and subsequent recombination to color-neutral objects produces showers of hadronic particles called *jets*.

2.2.1 Observables in perturbative QCD

To make predictions of physical observables, it is necessary to solve the equations of motion obtained from the Lagrangian. However, for the full interacting theory, these cannot be solved analytically, requiring instead a perturbative approach. Observables in perturbative QCD (pQCD) are expressed as a power series in the strong coupling constant α_s, which is defined in terms of the coupling g_s in the Lagrangian:

$$\alpha_s = \frac{g_s^2}{4\pi} \tag{2.4}$$

The pQCD prediction of an observable X, at a fixed order N in α_s, is given by the N-th partial sum of the perturbative expansion,

$$X^{(N)} = \sum_{k=0}^{N} c_k \, \alpha_s^k \,, \tag{2.5}$$

and is referred to as being at *leading order* (LO) if N is the smallest integer for which the perturbative coefficient c_N is nonzero. Higher-order predictions are said to be at *next-to-leading* order (NLO), *next-to-next-to-leading* order (NNLO), and so on.

Among the most common observables are scattering cross sections, defined in terms of the initial and final states of the scattering process. In the quantum formulation of scattering theory, these are modeled via the S-matrix, which codifies the transition amplitudes between these states. The transition probability between a given initial and final state is given by the square of the corresponding S-matrix element. For a given initial state, and the overall cross section is obtained by Fermi's golden rule by integrating the transition probability over the space of available final states (commonly referred to as the *phase space*).

When final states are described in terms of a characteristic variable (or set of variables) x, several types of cross section can be distinguished. They are said to be *inclusive* in x if the phase space integration is carried out over all possible final states with respect to x. Conversely, if x is restricted to particular values, the corresponding cross sections are said to be *exclusive* in x. For *differential* cross sections with respect to x, the integration over x is not carried out, expressing instead the differential rate of change in the cross section as a function of an infinitesimal change $\mathrm{d}x$.

An essential component of cross section predictions is the calculation of the S-matrix elements. In the perturbative framework, the contributions to the S-matrix can be derived from the interaction Lagrangian through a set of prescriptions known as *Feynman rules*, which lead to a representation of these contributions as so-called *Feynman diagrams*. These are constructed out of basic elements such as *external legs*, which correspond to the initial and final state particles, *propagators*, which represent the exchange of virtual particles, and *vertices*, which represent the particle interactions.

Feynman diagrams embody a calculation prescription for the scattering amplitude of perturbative contributions, essentially involving the integration over internal degrees of freedom such as virtual particle momenta while enforcing momentum conservation at the vertices, each of which contributes to the amplitude with a factor corresponding to the Lagrangian coupling g. Due to the latter fact, the diagrams are organized naturally by the order in perturbation theory, which is determined based on the number of vertices.

2.2.2 Renormalization

Beyond leading order, loops may appear in the perturbative calculations, for instance due to virtual particles being emitted and reabsorbed by propagators or external legs. A particularity of loops are the unconstrained momenta, which need to be integrated over the entire spectrum, resulting in so-called *ultraviolet divergences*. To obtain finite predictions at higher orders in perturbation theory, the divergences must be removed in a procedure known as *renormalization*.

The premise of renormalization lies in the conceptualization of the quantities in the Lagrangian, such as masses or couplings, as unobservable or *bare* quantities, which result in the infinities discussed before when handled in the context of perturbation theory. To avoid this, the Lagrangian is first rewritten in terms of observable or *renormalized* quantities, additionally introducing so-called *counterterms* to recover the original bare Lagrangian. These are defined in such a way as to reflect the divergent contributions at all orders in perturbation theory, which can then be subtracted order by order to obtain finite predictions.

An important part of this process consists in expressing divergent amplitudes as the limits of finite expressions with respect to an additional parameter called the *regulator*, allowing them to be manipulated algebraically in a consistent way. In the context of UV divergences, one possible regulator is given by the introduction of a physical cutoff scale Λ as the upper bound of the amplitude integration, recovering the divergence in the limit $\Lambda \to \infty$. A more widely-used approach called *dimensional regularization* instead makes use of the fact that divergent amplitudes in four-dimensional spacetime generally converge for higher dimensionalities $4 + \varepsilon$, allowing the divergence to be regulated by taking the limit $\varepsilon \to 0$.

The subtraction is then performed by removing the dependence of the predictions on the explicit regulator according to a specific prescription or *subtraction scheme*. A variety of such schemes exist, of which the most common in the context of QCD is the *modified minimal subtraction* scheme, also referred to as the "MS-bar" scheme ($\overline{\text{MS}}$) [15–17].

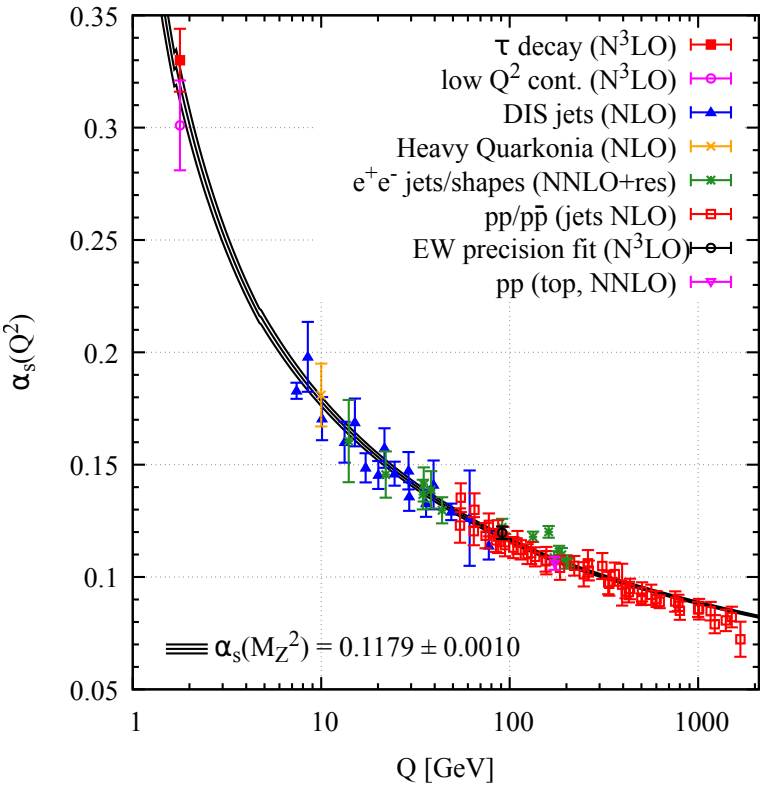

Figure 2.2 – Running of the strong coupling α_s, as determined in a fit to experimental results across a wide range of energy scales. Taken from [10]

Running coupling

A side effect of the renormalization procedure is the introduction of an additional energy scale to the problem, called the *renormalization scale* μ_R. In dimensional regularization, it arises from the need to redefine the coupling in a higher-dimensional theory to ensure it remains dimensionless while preserving the overall mass dimension of the Lagrangian.

As a result, all renormalized quantities such as the particle masses or the coupling constant, become dependent on the choice of the renormalization scale. However, since this is an arbitrary parameter, physical observables cannot depend on its value. This requirement imposes a restriction on the dependence of the coupling constant α_s on the energy scale of an interaction.

Introducing a formal dependence of an observable X at an interaction scale Q on the renormalization scale μ_R, this is expressed via the *renormalization group equation* (RGE):

$$\mu_R^2 \frac{\mathrm{d}}{\mathrm{d}\mu_R^2} X\left(\frac{Q^2}{\mu_R^2}, \alpha_s\right) = \mu_R^2 \frac{\partial}{\partial \mu_R^2} X + \mu_R^2 \frac{\partial \alpha_s}{\partial \mu_R^2} \frac{\partial}{\partial \alpha_s} X = 0 \tag{2.6}$$

The two terms above can be represented more compactly by expressing the renormalization scale dependence of X as a function of the logarithm τ with respect to the interaction scale Q, and introducing the *beta function* $\beta(\alpha_s)$, which encodes the dependence of α_s itself on the renormalization scale:

$$\tau = \log \frac{Q^2}{\mu_R^2}, \qquad \beta(\alpha_s) = \mu_R^2 \frac{\partial \alpha_s}{\partial \mu_R^2} \tag{2.7}$$

The RGE then becomes

$$\frac{\partial}{\partial \tau} X = \beta(\alpha_s) \frac{\partial}{\partial \alpha_s} X, \tag{2.8}$$

indicating that the energy scale dependence of X is defined entirely by the renormalization scale dependence of α_s encoded by the beta function. The equation is solved by introducing a running coupling $\alpha_s(Q)$, given by:

$$\beta\left(\alpha_s(Q)\right) = \frac{\partial \alpha_s(Q)}{\partial \tau} \tag{2.9}$$

As outlined above, QCD is asymptotically free at high energies, leading to an overall negative beta function. The calculation of $\beta(\alpha_s)$ can be done in perturbation theory, by expanding it as a perturbative series in α_s with the beta-function coefficients b_i, starting with the 1-loop contribution at α_s^2:

$$\beta(\alpha_s) = -\left(b_0 \alpha_s^2 + b_1 \alpha_s^3 + b_2 \alpha_s^4 + \ldots\right) \tag{2.10}$$

The value of the coefficients depends on the renormalization scheme. In the $\overline{\text{MS}}$ scheme, the first two are given by [10]:

$$b_0 = \frac{11 C_A - 4 n_F T_R}{12\pi} \tag{2.11}$$

$$b_1 = \frac{17 C_A^2 - n_F T_R \left(10 C_A + 6 C_F\right)}{24\pi^2} \tag{2.12}$$

In the above, n_F represents the number of active light quark flavors. Additional factors appear that are related to the SU(3) algebra, with $T_R = 1/2$ referring to the normalization of the t matrices of SU(3), and C_F and C_A denoting color factors associated with the emission of gluons from a quark or another gluon, respectively. They are given as a function of the number of colors N_C as:

$$C_A = N_C = 3, \qquad C_F = \frac{N_C^2 - 1}{2 N_C} = 4/3 \tag{2.13}$$

2.2.3 Parton distribution functions (PDFs)

Historically, measurements of scattering processes at hadron colliders have been instrumental in achieving an understanding of the strong interaction. Measurements of differential cross sections in deep inelastic scattering (DIS) have revealed that hadrons such as the proton are not fundamental particles, but are instead composed of point-like constituents called *partons*. This is consistent with the modern understanding of hadrons as bound states of QCD composed of quarks and gluons.

This introduces an additional difficulty when calculating theoretical predictions for observables at hadron colliders. Since the hadrons that appear in the initial state are composite objects, the parton-level predictions obtained in the framework of perturbative QCD must be complemented by an additional model of the hadron structure. However, this structure cannot be modeled using perturbative methods, since at the low energy scale of hadrons the coupling constant becomes large.

Instead, an approach called *collinear factorization* is used to recover the predictive power of perturbative calculations for hadron scattering. It consists of separating the low-energy interactions that characterize the hadron structure from the hard scattering process by introducing an energy threshold called the *factorization scale* μ_F. This allows the latter to be handled in perturbative QCD on the parton level, while the former is modeled using nonperturbative methods.

This is achieved by describing the low-energy part in terms of the probability of constituent partons to participate in the scattering. The probability is parametrized in terms of the energy scale Q^2 of the interaction and the Lorentz-invariant *Bjorken* x variable, which indicates the fraction of the total hadron momentum carried by a constituent parton. In deep inelastic scattering of electrons and protons, the energy scale Q^2 of the interaction is given by the magnitude of the four-momentum q_μ exchanged between the electron and the constituent parton: $Q^2 = -q_\mu q^\mu$. Denoting the proton four-momentum as p_μ, the *Bjorken* x variable is given by:

$$x = \frac{Q^2}{2 p_\mu q^\mu} \tag{2.14}$$

The resulting probability densities as a function of x are called *parton distribution functions* (PDFs). They are not given by the perturbative theory and must instead be determined by other means such as the comparison of theory predictions to experimental data. The constraining of PDFs from experimental data is an active area of research, with several sets of PDFs being published by specialized groups employing different data sets, theory calculations and methodologies [18–21].

Figure 2.3 shows an example of PDF parametrizations, as obtained by the NNPDF3.1 analysis [18]. The contribution of each parton flavor is parametrized separately: The distribution of the valence *up* and *down* quarks peak at high values of x, indicating that

these quarks carry a large fraction of the total proton momentum, while other quark flavors – the so-called *sea quarks* – are predominantly found at low x. However, the dominant contribution at low x is from gluons.

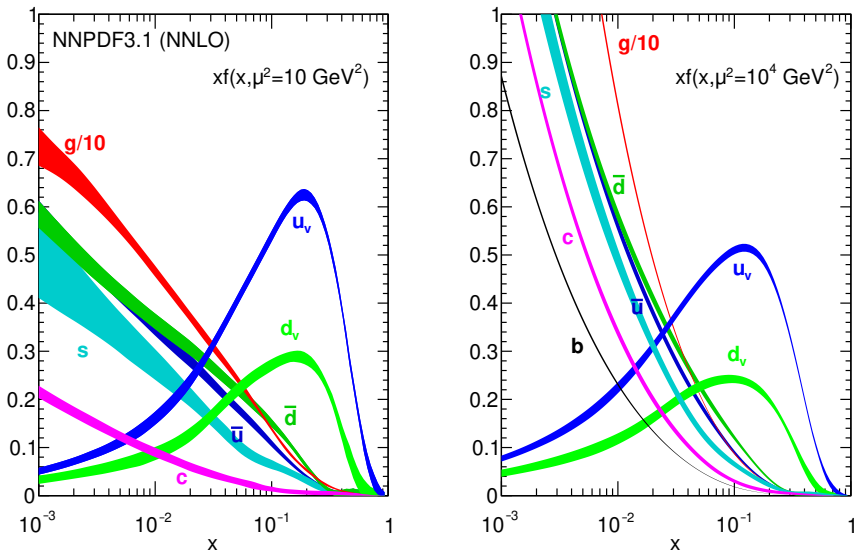

Figure 2.3 – Parton distribution functions (PDFs) obtained by the NNPDF3.1 analysis for the different parton flavors. Shown are the PDF parametrizations as a function of the momentum fraction x at two energy scales, $Q^2 = 10\,\mathrm{GeV}^2$ (*left*) and $Q^2 = 10^4\,\mathrm{GeV}^2$ (*right*), showing the dominant contribution from gluons and sea quarks at low x, and from the valence *up* and *down* quarks at higher momentum fractions. Taken from [18].

In proton-proton collisions, where the initial state is given by two hadrons, the cross section for the production of a final state X can be factorized at each order n in perturbation theory as a double convolution of the partonic cross section $\hat{\sigma}^{(n)}_{ij\to X}$ and the proton PDFs f_i and f_j, where i and j denote the constituent parton flavors. The cross section prediction at order N in α_s is thus given by:

$$\sigma^{(N)}_{\mathrm{pp}\to X} = \sum_{n=0}^{N} \alpha_\mathrm{s}^n\left(\mu_\mathrm{R}^2\right) \sum_{i,j} \int_0^1 \mathrm{d}x_1 \int_0^1 \mathrm{d}x_2\, f_i\left(x_1, \mu_\mathrm{F}^2\right) f_j\left(x_2, \mu_\mathrm{F}^2\right) \hat{\sigma}^{(n)}_{ij\to X}\left(x_1 x_2 s, \mu_\mathrm{R}^2, \mu_\mathrm{F}^2\right) \quad (2.15)$$

DGLAP evolution equations

In contrast to the PDF dependence on x, dependence on the energy scale Q^2 is constrained by the choice of the factorization scale μ_F. This is similar to the dependence of α_s on the energy scale induced by the renormalization procedure discussed in section 2.2.2: If perturbative calculations of hadron-level observables were possible *ab initio*,

the result would necessarily be independent of the arbitrary factorization scale μ_F.

This gives rise to a set of relations for the dependence of PDFs on the energy scale, which are analogous to the renormalization group equation for the energy dependence of α_s. These are called the DGLAP[1] evolution equations [22–24]. and are derived for each order in perturbation theory. At leading-order, the DGLAP evolution equations of the PDF $f_i(x, Q^2)$ for a particular parton flavor i are given by [10]:

$$\mu_\mathrm{F}^2 \frac{\partial_i f(x, \mu_\mathrm{F}^2)}{\partial \mu_\mathrm{F}^2} = \sum_j \frac{\alpha_\mathrm{s}(\mu_\mathrm{F}^2)}{2\pi} \int_x^1 \frac{\mathrm{d}z}{z} P_{i \leftarrow j}(z)\, f_j\left(\frac{x}{z}, \mu_\mathrm{F}^2\right), \tag{2.16}$$

In the above, $P_{i \leftarrow j}(z)$ are the Altarelli-Parisi splitting kernels, which indicate the probability for a parton of flavor j to emit a parton of flavor i carrying an energy fraction z of the original parton.

2.3 Jets

The observation of jets in high-energy hadron collisions was essential in developing the modern understanding of hadrons as bound states composed of particles carrying color charge. Jets are sprays of collimated particles that are produced as a result of color-charged particles in the final state of the scattering process. Due to the confining nature of QCD, these cannot exist as free particles and instead recombine to form an abundant stream of hadronic particles: a *jet*. Consequently, the measurement of jet-related observables provides an experimental handle on QCD final states, making jets an essential tool for studying the strong interaction in both the perturbative and nonperturbative regimes.

An important prerequisite for this is the connection between the parton-level calculations performed at a fixed order in perturbation theory and the experimental observation of hadronic jets, accounting for effects such as the fragmentation of final-partons through the emission of additional QCD radiation and the final recombination to hadrons at low energies. These aspects are modeled with the help of Monte Carlo event generation techniques, which are discussed section 2.3.1.

A further important component for such studies is a rigorous definition of jets, which can be applied both in the context of experimental observations and when performing theoretical calculations of observable quantities. In particular, requirements must be imposed on the jet definitions to ensure that the observables they describe can be measured in a clean and unambiguous way, and that the corresponding theoretical predictions are not sensitive to effects such as divergent amplitudes that are often encountered in perturbation theory. This is discussed in section 2.3.2.

[1] The equations are named for the authors Y. Dokshitzer, V. Gribov, L. Lipatov, G. Altarelli and G. Parisi.

2.3.1 Monte Carlo event generation

Fixed-order calculations in perturbative QCD provide predictions for well-defined partonic final states of hard scattering processes. Since the final-state partons are not directly observable, a connection to the observed hadronic jets must be established.

The tool of choice for this are Monte Carlo event generators, in which scattering events are first simulated at the parton level by sampling the distributions obtained from fixed-order calculations. The partonic final state is then evolved to the hadron level through a series of both perturbative and nonperturbative techniques, which are outlined below. More detailed introductions to this topic can be found in references [25, 26].

Parton shower

The color-charged partons that take part in the hard scattering process are accelerated to high energies and consequently emit radiation in the form of gluons. These also carry color charge, leading to further emissions and resulting in a cascade of color-charged particles referred to as a *parton shower*. These high-multiplicity states cannot be described directly by the fixed-order calculation, and are instead modeled using Monte Carlo techniques.

The parton shower modeling consists in simulating a series of independent emissions starting from partons involved in the hard process. These emissions are ordered in terms of a shower evolution variable such as the parton virtuality, its transverse momentum or the emission angle with respect to the parent parton. The probability of branching at a particular value of the evolution variable is calculated based on the Altarelli-Parisi splitting kernels discussed in section 2.2.3.

In essence, the parton shower approximates higher-order corrections in perturbation theory by resumming the logarithmically divergent contributions from real emissions of soft and collinear partons to all orders. When using the parton shower technique in connection with fixed-order calculations, it is important to avoid any double counting that may arise. Since the parton shower contains a component modeled with the help of perturbative calculations, the fixed-order predictions used for the parton-level event generation must be matched to the parton shower by subtracting the perturbative component. Furthermore, since high parton multiplicities can be accounted for by both fixed-order calculations at higher orders and parton showers, a prescription to separate the two contributions is required to prevent double counting.

Hadronization

At sufficiently low energies, the quarks and gluons resulting from the fragmentation recombine to form color-neutral hadrons. This process cannot be modeled perturbatively

on account of the low-energy regime, prompting the need for phenomenological models of hadron formation. Two such models are widely implemented in modern Monte Carlo generators:

Lund string model [27, 28] In this model, the notion of color flux tubes between quarks is exploited to explain hadronization. The flux tubes are modeled as overall colorless strings spanned between quark–antiquark pairs moving away from one another, with gluons represented by kinks in these strings. As the energy density between two partons increases linearly with the distance, it becomes increasingly likely to produce new quark–antiquark pairs, which is modeled by breaking the string using a probabilistic approach, preserving the overall color neutrality of both resulting strings. This continues until the available energy is no longer sufficient to produce new particles, at which point the existing ones are combined to form hadrons. This hadronization model is implemented in the Pythia Monte Carlo event generator [29].

Cluster fragmentation model [30, 31] This approach consists in the identification of color-singlet clusters which arise naturally due to a property of QCD called *preconfinement* [32]. These clusters are independently decayed to hadrons according to their mass. Low-mass clusters give rise to hadrons directly, while clusters with intermediate masses are decayed assuming a two-body phase space. High-mass clusters are optionally separated into two lighter ones beforehand. Variants of this hadronization model are implemented in the Herwig [33] and Sherpa [34] Monte Carlo event generators.

Underlying event

A final component for achieving an improved description of hadron scattering involves the so-called *underlying event* (UE), which is the result of interactions outside of the main hard process. These include mostly low-energy (soft) contributions, originating, for example, from the so-called *beam remnants*, which consist of the state left behind after the partons that participate in the hard scattering are ejected from the colliding hadrons. Another possibility are hard scattering interactions between additional pairs of beam-hadron partons, known as *multiple parton interactions* (MPI).

Monte Carlo generators model both the soft and hard components of the UE using a variety of techniques, which are reviewed in detail in references [26, 35]. To achieve a good description of experimental data, the parameters of these techniques are determined by comparing to measurements of UE-sensitive observables, resulting in a so-called event generator *tune* [36].

2.3.2 Clustering algorithms

Jets are composite objects that consist of stable particles propagating within an approximately cone-like region determined by the kinematics of the originating parton. To identify the particles belonging to jets in a systematic way, a variety of *clustering algorithms* are employed that operate on the collection of particles based on their individual kinematic properties. These provide a prescription for jet identification that can be applied on the level of partons or hadrons with kinematics sampled from theoretical distributions, as well as on stable particles reconstructed in a scattering experiment. At hadron colliders, the variables most commonly used to describe the particle kinematics are the transverse momentum p_T, the rapidity y of the longitudinal Lorentz boost with respect to the laboratory frame, and the azimuthal angle ϕ.

An important property of jet algorithms is their stability to contributions from additional particles. In particular, jet configurations are required to be stable with respect to the collinear splitting of particles or the addition of soft particles, which must not result in additional jets. This is known as *collinear and infrared safety* [37, 38], and is particularly important for observables that are exclusive in jet multiplicity.

Jet algorithms can be classified into two broad categories. *Cone-based* algorithms aim to identify stable particle configurations in circular regions of the (y, ϕ) plane with a fixed radius R. In contrast, *sequential recombination* algorithms reconstruct jets by clustering together particles based on a distance metric.

In the latter approach, the distance between two particles is defined in terms of the particle transverse momenta and the Euclidean separation in (y, ϕ) space, with the relative contribution of each being determined by two parameters, p and R. In addition, a distance between each particle and the beam is introduced. The distance between two particles i and j with transverse momenta $k_{\mathrm{T},i}$ and $k_{\mathrm{T},j}$ and a (y, ϕ) separation of ΔR_{ij}, as well as the distance of particle i to the beam is given by:

$$d_{ij} = \min\left(k_{\mathrm{T},i}^{2p}, k_{\mathrm{T},j}^{2p}\right) \frac{\Delta R_{ij}^2}{R^2}, \quad \text{with } \Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2 \tag{2.17}$$

$$d_{i\mathrm{B}} = k_{\mathrm{T},i}^{2p} \tag{2.18}$$

Particles are then clustered sequentially by evaluating the smallest overall distance d_{ij} or $d_{i\mathrm{B}}$. If the smallest overall distance is between two particles, these are clustered together and their four-momenta are summed to yield a new composite particle, replacing the original ones. Otherwise, if the smallest overall distance is between a particle and the beam, the particle is identified as a jet and removed from the particle collection. This is repeated until all particles have been clustered.

The radius parameter R controls the geometric extent of the particles clustered into a jet and represents the minimum (y, ϕ) distance between two jets. The kinematic parameter p determines the contribution of the particle momentum to the distance metric and can

be used to prioritize the clustering of low or high-momentum particles. The most commonly used choices for p distinguish three types of sequential clustering algorithms:

k_T algorithm [39, 40] Setting $p = 1$ leads to an equal contribution from the kinematic and geometric variables to the distance. Soft particles result in the smallest distances and are clustered first, inverting the branching structure of QCD parton showers used to model the perturbative evolution of partonic final states down to the final hadronization scale at which jets are measured (see section 2.3.1) [41].

Cambridge/Aachen algorithm [42, 43] A value of $p = 0$ describes a clustering procedure that only takes the spatial separation of particles into account. In particular, angular relationships between particles are preserved, providing a connection to angular-ordered parton showers, making this jet algorithm popular for jet substructure studies [44].

anti-k_T algorithm [45] For $p = -1$, the particle momenta contribute reciprocally to the distance metric, leading to high-energy particles being clustered first. This is particularly advantageous in the context of high-luminosity hadron collisions, where events contain a small number of hard particles reflecting the products of the hard scattering interaction, superimposed over a background of soft emissions coming from unrelated scattering activity (pileup) or the underlying event.

3 The Large Hadron Collider and the CMS experiment

Particle colliders have long been established as the experimental tool of choice for gaining insights into the fundamental physical laws governing subatomic processes. The experimental premise here involves the acceleration of particles to energies equivalent to many thousands of times their rest energy and allowing them to collide in a controlled environment. The original particles involved in the collision disintegrate, producing instead an abundance of new particles, in line with the principle of matter–energy equivalence. By identifying and measuring the properties of these particles with the help of particle detectors, it becomes possible to make quantitative statements about the fundamental physical processes by which they were produced.

At the Large Hadron Collider (LHC), protons are accelerated to high center-of-mass energies and brought to collision inside particle detectors. One of the four main detectors located at the LHC is the CMS detector (short for *Compact Muon Solenoid*). It is a general-purpose detector designed to provide as much information as possible about the particles produced in a collision event across a wide range of kinematic configurations, as well as an accurate and precise measurement of their properties.

This section gives a brief overview of the LHC accelerator, as well as a description of the CMS detector, highlighting the purpose and functionality of its main subdetector systems, and the technology used for triggering and data acquisition. The information presented here summarizes the one provided in reference [46], where a more in-depth description can be found.

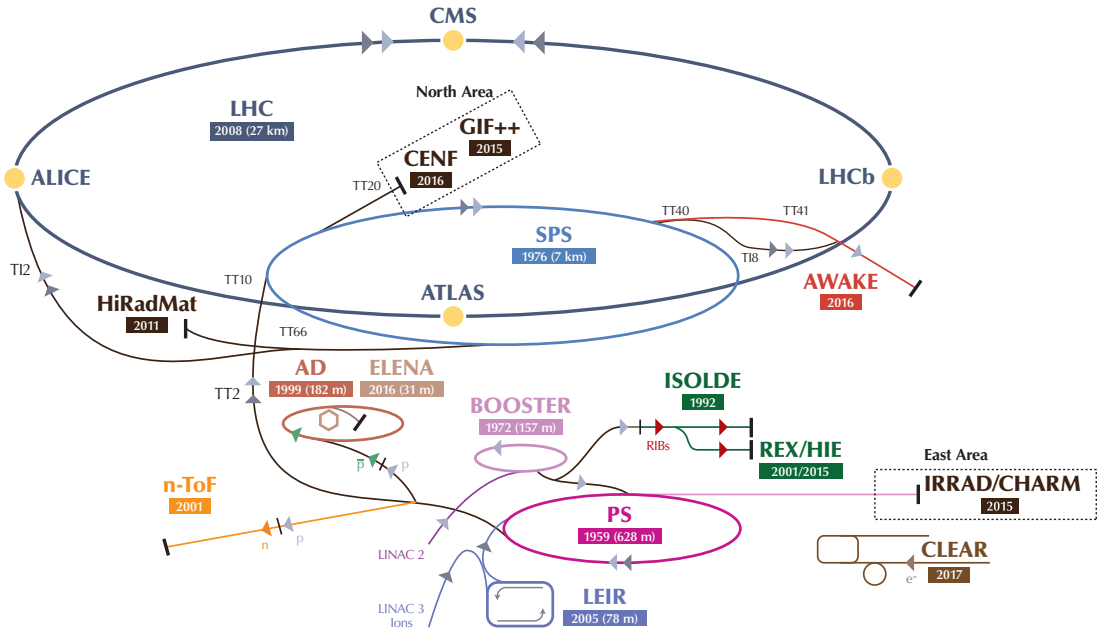

Figure 3.1 – The CERN accelerator complex. Taken from [47].

3.1 The Large Hadron Collider

Located at the European Organization for Nuclear Research (CERN) near Geneva in Switzerland, the Large Hadron Collider (LHC) is a machine designed to accelerate protons (and to a lesser extent heavy ions) to high energies and bring them to collision. It is the largest machine of its kind, both in terms of its size and the peak center-of-mass energy of the colliding protons, which during the most recent operational run reached a nominal value of 13 TeV.

The most prominent feature of the LHC is its main storage ring, which extends throughout a circular tunnel with a circumference of 27 kilometers. Two proton beams are circulated in opposite directions along the ring, being kept on the circular trajectory with the help of superconducting magnets able to generate field strengths of the order of 7 tesla.

In 2015, the LHC started its second operational run, with the center-of-mass energy in proton collisions being increased from 8 TeV to 13 TeV. Further optimizations were made with the goal of increasing the instantaneous luminosity, reaching the design value of $10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ in 2016.

Protons at the LHC are accelerated to their end energies in several stages, which can be seen in figure 3.1. The acceleration chain starts with a linear accelerator (LINAC 2), in which protons obtained from the ionization of hydrogen gas are accelerated by a series

of radio-frequency electromagnetic cavities. This is followed by a series of four circular accelerators: the Proton Synchrotron Booster (PSB), the Proton Synchrotron (PS), the Super Proton Synchrotron (SPS), and finally the main LHC storage ring. In each stage, the energy of the protons is increased, reaching values of 50 MeV after the LINAC 2 stage, 1.4 GeV after the PSB, 26 GeV after the PS, and 450 GeV after the SPS.

During this process, protons are grouped into so-called *bunches*, each of which contains an approximate 10^{11} protons. The bunches are injected into the main storage ring, where they are accumulated to form the two proton beams until the number of stored bunches reaches the ca. 3000 required to attain the intended luminosity.

Once this happens, the proton energy is increased further, reaching the target value of 6.5 TeV per proton per beam, and thus 13 TeV with respect to the center of mass of two protons circulating in opposite directions. After reaching this energy, the beams are adjusted, stabilized and finally intersected at so-called *interaction points*, located at the centers of particle detectors where the collision products are recorded.

There are four main detectors installed along the LHC beam line. Two of these, ATLAS [48] and CMS [46], are general-purpose detectors designed to collect data suited for a broad range of analysis cases. This is complemented by two special-purpose detectors: ALICE [49], which focuses on ion collisions and studies of quark–gluon plasma, and LHCb [50], with the primary focus on b-quark physics and CP violation.

3.2 The CMS detector

The Compact Muon Solenoid (CMS) is a general-purpose particle detector located at point 5 of the LHC tunnel, on the side of the ring directly opposite the main CERN campus. Situated in a cavern about 100 meters underground, the detector occupies a cylindrical volume of approximately 22 meters in length and 15 meters in diameter centered around the nominal interaction point of the colliding protons. As the name suggests, it has a highly compact design. With a total weight surpassing 12 000 tonnes, it is approximately twice as heavy as the comparable ATLAS detector, while occupying less than a third of the volume.

Within this volume, the various subsystems that make up the detector are arranged in a succession of layers. Closest to the beam line is the *inner tracking system*, which consists of a pixel and a silicon strip tracking detector. This is followed by two layers containing the *electromagnetic* and *hadronic calorimeter* systems. Immediately outside the calorimeters is the superconducting solenoid magnet, which generates a high magnetic field in its interior, reaching magnetic flux densities of up to 3.8 tesla. The final layer consists of an iron yoke, which accounts for the bulk of the detector weight and serves to return the magnetic flux along the outer detector. Interspersed among the flux return plates are detector modules destined for the detection of muons. In addition to the modules

containing active detector material, a significant amount of space is taken up by cables, readout electronics, cooling and the support structure.

Figure 3.2 shows a transverse cross-sectional view of the detector. In the following, a more detailed discussion of each subsystem is given.

Coordinate system

By convention, the reference coordinate system at CMS is a Cartesian coordinate system centered at the nominal interaction point, with the x-axis oriented toward the center of the LHC ring, the y-axis pointing upward and the z-axis pointing in the circulation direction of the counter-clockwise proton beam. While this choice of coordinates provides a good reference, it is not particularly well-suited to the cylindrical detector geometry, which is described using a set of derived variables.

The spatial configuration in the dimensions transverse to the beam line is described in terms of a radial coordinate r indicating the distance to the z-axis, and an azimuthal angle ϕ. In longitudinal cross sections along the z-axis, spatial directions are indicated using the *pseudorapidity* η, which is related to the polar angle θ with respect to the positive z-axis by:

$$\eta = -\ln \tan\left(\frac{\theta}{2}\right) \tag{3.1}$$

Thus, the transverse plane passing through the origin point corresponds to $\eta = 0$, while the positive and negative longitudinal directions along the beam line correspond to $\eta = +\infty$ and $\eta = -\infty$.

The use of pseudorapidity instead of the polar angle is motivated by physical considerations. At velocities close to the speed of light, for which the mass of a particle becomes negligible compared to its total energy, this quantity represents a good approximation of the *rapidity* of its longitudinal motion, defined as the parameter of its Lorentz boost in the z direction with respect to the laboratory frame. This makes differences in rapidity invariant under longitudinal Lorentz boosts, which is particularly advantageous for measurements at hadron colliders, where the rest frame of the initial-state parton pair has a nonvanishing boost in the longitudinal direction which cannot be controlled experimentally.

3.2.1 Inner tracking system

Tracking is an essential part of particle reconstruction that consists of measuring the trajectories of charged particles emerging from the proton collisions. In addition to providing information about the precise position at which a proton collision occurred, tracking is essential in measuring the particle momentum, which can be inferred from the curvature of the particle trajectories within the strong magnetic field at the center of the detector.

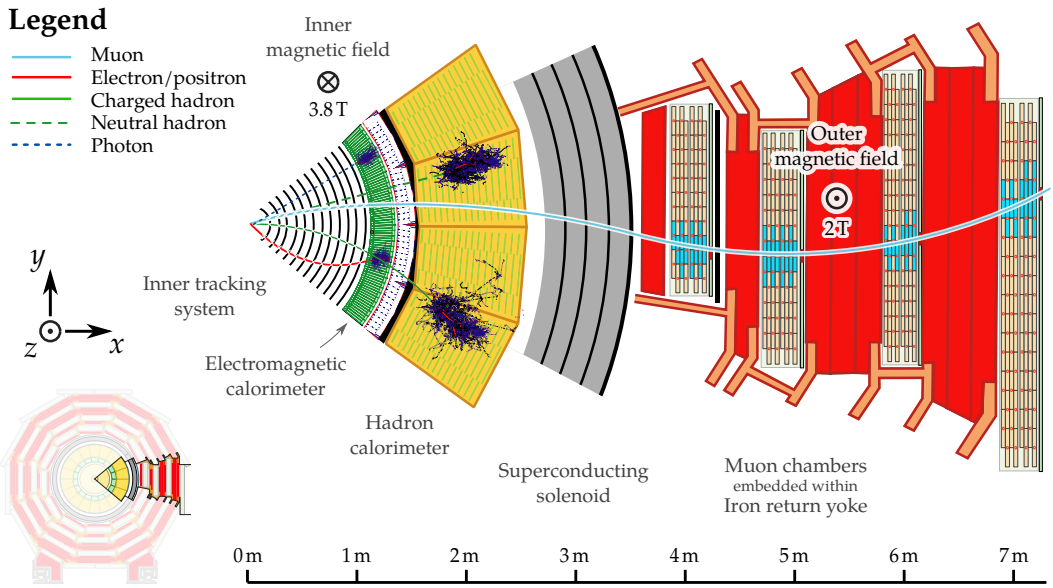

Figure 3.2 – Schematic transverse section through the CMS detector, showing the different subsystems and the typical detection signatures of particles. Adapted from [51].

The CMS tracking system consists of two subdetector systems which both make use of silicon-based semiconductor tracking technology. The innermost subdetector consists of an array of pixel sensors, providing the high resolution needed for an accurate reconstruction in the high-particle flux environment of the LHC. In the outer regions, where the particle flux is reduced, tracking information is provided by silicon strip detector modules.

Pixel detector The pixel detector modules consist of sensors with a cell size of $100 \times 150\,\mu m^2$ and are arranged in three cylindrical shells mounted at distances of $r = 4.1, 7.3$ and $10.2\,cm$ from the beam line, providing coverage of the barrel region of the detector within $|\eta| < 2.5$. This is complemented in the more forward regions by two additional modules per quadrant, which are mounted vertically at $z = \pm 34.5\,cm$ and $\pm 46.5\,cm$, covering distances between 6 and 15 cm from the beam line.

This configuration ensures high-resolution coverage of the interaction region, being able to reconstruct at least three high-precision points on the trajectory of most charged particles within pixel coverage. This is instrumental not only in reconstructing the full trajectories, but also in the identification of primary and secondary interaction vertices and their association to one or more particles.

Silicon strip tracker The outer layers of the tracking system consist of silicon strip modules, covering approximately the same pseudorapidity region of $|\eta| < 2.5$ as the pixel sensors, but extending outward to distances of $r = 1.1\,m$ from the beam line. The modules are arranged both in cylindrical configurations with strips running

parallel to the beam axis and along endcap rings with radial strips running perpendicular to it.

When traversed by a charged particle, each module delivers a one-dimensional measurement of its position along the direction perpendicular to the strips. For a module mounted at a distance r in the barrel region, this yields a value of the $r\phi$ coordinate of the particle, while for a module with radial strips mounted at fixed z along the endcap, this results in a measurement of the ϕ coordinate. To provide a second coordinate, a number of barrel layers and endcap modules are equipped with two sets of modules, with the second being mounted at an angle of 100 mrad with respect to the first. This enables the measurement of the position along the long side of the strip, namely z in the barrel and r in the endcaps.

3.2.2 Electromagnetic calorimeter

While the momentum information provided by the tracking system is essential for particle reconstruction, a measurement of the particle energy is required to fully reconstruct the associated four-momenta. Furthermore, tracking does not provide information on neutral particles, which must be detected by additional means. Both of these requirements are addressed by the calorimeter system, which is designed to absorb the majority of particles and provide a measurement of their energy.

Immediately outside the tracking system lies the electromagnetic calorimeter (ECAL). It is a homogeneous calorimeter composed of over 70 000 lead tungstate ($PbWO_4$) crystals, which function as scintillators used to measure the energy of light, electromagnetically interacting particles such as electrons, positrons and photons.

When these high-energy particles interact with the crystal material, they initiate a chain reaction known as an *electromagnetic shower*. In this process, electrons and positrons emit bremsstrahlung photons by scattering on atomic nuclei, while photons give rise to further electrons and positrons via pair production. Below the pair production threshold, shower photons deposit their energy in the material via the Compton and photoelectric effects. As the resulting excited atomic electrons relax back to the ground state, scintillation light is emitted, with the number of photons being proportional to the total energy deposited. An energy measurement is then made by registering the number of scintillation photons with the help of photodetectors.

The use of lead tungstate as a scintillation material is motivated by its high density of $8.3\,\mathrm{g\,cm^{-3}}$ and consequently short radiation length of 0.89 cm. Since this is the characteristic length scale of energy deposition, lower values allow for a compact calorimeter design, while ensuring that all of the particle energy is absorbed. Furthermore, lead tungstate has a comparatively small Molière radius, which is a measure of the transverse spread of electromagnetic showers within the material. With a value of 2.2 cm, this makes it possible to lay out the scintillator crystals with a high granularity while providing adequate lateral shower containment in each crystal.

The ECAL is composed of several subsystems. The largest of these is the ECAL barrel (EB), which covers the central detector region. This is complemented by the two ECAL endcaps (EE), extending the range to cover the forward areas of the detector. In addition, a preshower detector (ES) is mounted in front of each endcap.

ECAL barrel (EB) The ECAL barrel covers the pseudorapidity region $|\eta| < 1.479$. This range is subdivided into 170 regions, each with 360 crystals covering all directions in ϕ. These have a length of 230 mm, corresponding to 25.8 interaction lengths, and are mounted with their main axis oriented towards the nominal interaction point. The crystal shape is tapered and corresponds to a truncated pyramid, with the cross section increasing from 22 mm^2 at the inner face to 26 mm^2. At the far end of each crystal, silicon avalanche photodiodes are installed for reading out the scintillation signals.

ECAL endcaps (EE) In the endcaps, crystals with a similar size to those in the barrel are mounted on a rectangular x–y grid, covering the approximately ring-shaped area within $1.479 < |\eta| < 3.0$. The photodetectors used here are vacuum phototriodes.

ECAL preshower (ES) Covering most of the ECAL endcaps are preshower detectors. These are sampling calorimeters with a total depth of 20 cm consisting of two layers of lead radiators followed by silicon strip sensors. Their purpose is to provide additional information to better distinguish individual high-energy photons from collimated photon pairs resulting from neutral pion decay. In addition, they provide an improved position determination of electrons and photons and are also useful for distinguishing electrons from minimum ionizing particles.

3.2.3 Hadron calorimeter

While the electromagnetic calorimeter (ECAL) is effective in absorbing light, electromagnetically interacting particles, heavier decay products such as hadrons generally do not deposit their entire energy in the ECAL and thus are able to escape this detection layer without being absorbed. Since hadrons make up a significant fraction of decay products and are the main component of particle jets, their containment is essential for measuring the jet energy. To this end, the ECAL is complemented by an additional hadron calorimeter system (HCAL).

High-energy hadrons traversing matter undergo inelastic collisions with the atomic nuclei. This creates a cascade of secondary particles that consists mainly of pions and nucleons ejected from the nucleus, initiating a chain reaction called a *hadronic shower*.

The length scale of hadronic showers is characterized by the material-specific *nuclear interaction length*, which is typically larger than the radiation length determining the extent of electromagnetic showers. The containment of hadronic showers thus requires more

material, with about 9 interaction lengths being necessary to contain the full shower [52].

In addition to the hadronic component, the secondary pions created in hadronic cascades decay to highly energetic photon pairs, giving rise to secondary electromagnetic showers. This makes it possible to obtain an estimate of the shower energy by sampling the electromagnetic component of showers with scintillation detectors in a manner analogous to electromagnetic calorimeters.

Situated immediately outside the ECAL, the HCAL consists of four calorimeter subsystems, covering different portions of the detector. The two largest are the HCAL barrel (HB), which occupies the space between the ECAL barrel and the superconducting solenoid magnet, and the two HCAL endcaps (HE), located outside the ECAL endcaps. The HB is complemented by an outer calorimeter system (HO) that adds 1–2 additional detection layers outside the magnet coil. Finally, two hadronic forward calorimeters (HF) are placed in close proximity to the beam line at either end of the detector.

HCAL barrel (HB) The HCAL barrel covers the pseudorapidity region $|\eta| < 1.3$ and is organized into longitudinally aligned units with a trapezoidal cross section called *wedges*. Eighteen such units cover the complete ϕ range separately in the positive and negative η regions, resulting in a total of 36 wedges. Each wedge functions as a sampling calorimeter module, containing 14 alternating layers of brass absorber plates and plastic scintillators, which provide the active detection medium. The structure is reinforced by steel plates at the near and far ends of a wedge, which also function as absorbers.

The scintillation light in the active layers is collected by optical fibers and brought to an array of hybrid photodiodes (HPD) for readout. Signals are read out in terms of a grid of so-called calorimeter *towers*, which are defined by partitioning the detector volume into regions in terms of pseudorapidity and azimuthal angle. In HB, each wedge has 16×4 towers, covering equally sized intervals in pseudorapidity and azimuthal angle, respectively, which is equivalent to a granularity of 0.087 in both variables.

To obtain energy measurements, the signals registered in each tower are summed for the total projective depth of the calorimeter, which amounts to 5.39 nuclear interaction lengths at 90° to the beam line, increasing to 10.3 at the outer limit of the HB. The material of the ECAL barrel mounted in front of the HB corresponds to an additional 1.1 interaction lengths.

HCAL outer calorimeter (HO) In order to contain the hadronic showers in the central part of the calorimeter where the projective depth of the material is smaller, an outer calorimeter system is installed outside the superconducting solenoid. Covering the pseudorapidity region $|\eta| < 1.3$, it makes use of the material in the magnet coil and the iron return yoke as an absorber, measuring the energy of the final shower portion using a series of scintillator layers installed inside the first layer

of the iron return yoke. The innermost layer is complemented by a second layer, which is placed directly outside the magnet coil. The resulting configuration ensures that the projective depth of the entire calorimeter system in the central region remains above 11.8 interaction lengths, which is sufficient for shower containment.

HCAL endcaps (HE) The pseudorapidity region $1.3 < |\eta| < 3$ is covered by the HCAL endcaps, which are mounted directly behind the ECAL endcaps. These consist of layers of brass absorbers bolted together to form a compact structure, minimizing the gap between the endcap and the HB. The total calorimeter length, including the ECAL endcaps and preshower detectors, corresponds to about 10 interaction lengths, ensuring an adequate containment of hadronic showers.

The active medium consists of plastic scintillator trays, which are inserted transversely into the absorber structure. Like in the HB, the readout is organized into towers, with the outermost set of HB towers overlapping with the innermost ones in HE. The tower granularity in both η and ϕ is 0.087 within $|\eta| < 1.6$ and 0.17 beyond.

HCAL forward calorimeter (HF) The two forward calorimeters are installed close to the beam line, covering the pseudorapidity region $2.8 < |\eta| < 5.2$. Due to the very high radiation levels here, the calorimeter is encased in layers of concrete, steel and polyethylene shielding. Steel plates are used as an absorber material, incorporating radiation-resistant quartz fibers as the active detection medium. These collect the Cherenkov light emitted by charged shower particles, guiding it to external photomultiplier tubes for readout. As in the other components, the HF readout is organized into $\eta \times \phi$-towers, achieving a granularity of 0.175 in both variables.

3.2.4 Superconducting solenoid

The strong magnetic field at the center of the CMS detector is generated by a solenoidal electromagnet with a length of 12.5 m and a free bore diameter of 6 m, surrounding the inner tracking and calorimeter systems. In order to reach the field magnetic induction of 3.8 tesla needed for an accurate measurement of particle momenta, the main conductor is made of a superconducting niobium-titanium (NbTi) alloy and wound around the inner volume in a total of four layers, amounting to a total conductor length of 45.4 km. At the full operating current of 19 kA, the magnet stores 2.6 GJ of energy.

For stability and structural reasons, the superconducting material in the winding is co-extruded with an outer envelope of pure aluminium and reinforced with high-strength aluminium alloy, resulting in a total mass of 220 tonnes. During operation, this mass is cooled down to a temperature of 4.7 K using liquid helium as a cooling agent. This is well below the 9.25 K critical temperature of the NbTi conductor, thus ensuring that it remains superconducting.

An iron structure weighing 10 000 t called the *return yoke* encloses the solenoid and serves to return the magnetic flux along the outer region of the detector. It has a compartmen-

talized structure, allowing the detector modules of the muon system to be placed within it and thus contributing to overall compact design.

3.2.5 Muon system

While most other particles are fully absorbed by the calorimeters in the inner part of the detector, this is not the case for muons, which penetrate through to the outer layers largely unhindered. This is due to two factors. First, the lifetime of muons is sufficiently long to prevent them from decaying on the distance scale of the detector. In addition, since muons are leptons with a comparatively high mass, they deposit energy in material mainly through ionization and do not initiate electromagnetic (or hadronic) showers. This makes them *minimum ionizing particles*, allowing them to pass through the calorimeter system without being absorbed.

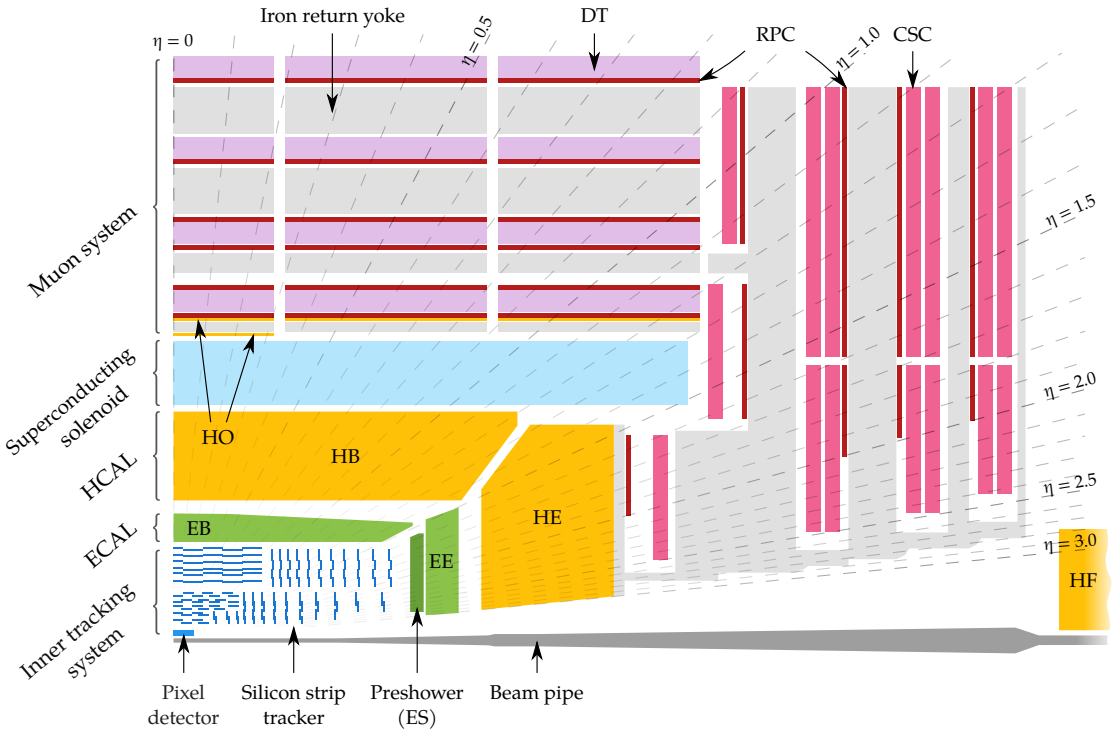

Figure 3.3 – Schematic longitudinal section through a quadrant of the CMS detector, showing the pseudorapidity coverage of the different subsystems. The muon detectors are embedded within the iron return yoke in the outer parts of the detector. Adapted from [46].

Muon detectors in CMS are thus installed in the outer detector regions, making use of the compartments within the iron return yoke around the superconducting solenoid (see figure 3.3). Their purpose is to reconstruct the trajectories of muons in the outer

detector, thus resulting in an overall improved muon momentum resolution when correlated with information from the inner tracking detector. In contrast to the inner tracking system, however, the muon system modules all rely on gas ionization as the primary detection mechanism.

For this purpose, an electric field is generated inside a volume of inert gas by applying a high voltage between an anode and a cathode. When high-energy muons traverse the volume, they ionize the gas atoms, releasing electrons which are then accelerated towards the anode. This is detected as an electric signal, making it possible to pinpoint the muon location.

Three distinct detector technologies are employed in the muon system. *Drift tubes* (DT) are used as the primary detector type in the barrel region, where the magnetic field is uniform and the overall muon rate is low. In the end caps, the magnetic field becomes nonuniform and the overall muon rate increases, prompting the use of the more resilient *cathode strip chambers* (CSC) as primary detectors. Both these detector types are complemented by additional modules composed of *resistive plate chambers* (RPC), serving mainly as an input to the trigger, on account of their speed. These components are outlined below.

Drift tubes (DT) This technology is used in the pseudorapidity range $|\eta| < 1.2$. DT modules are arranged around the beam line in twelve sectors, each comprising four stations at different radial distances. The individual DTs consist of a rectangular volume filled with a gas mixture of 80% Ar and 10% CO_2. The anode consists of a wire running through the center of the tube, while electrodes affixed to the tube walls function as cathodes. Electrons dislodged by high-energy muons drift towards the anode wire, triggering avalanches in close proximity to it. The muon position can then be inferred from timing measurements on the resulting electric pulse.

The tubes are oriented parallel to the beam line, thus providing a measurement of the muon tracks in the r–ϕ magnetic bending plane. A complementary measurement of the z coordinate is provided by additional layers oriented perpendicular to the beam line.

Cathode strip chambers (CSC) Detector modules of this type are installed in the end caps of the iron return yoke, covering the pseudorapidity range $0.9 < |\eta| < 2.4$, in partial overlap with the range covered by DTs. The individual CSC modules are trapezoidal in shape and are arranged radially around the beam line, each covering an azimuthal angle of 10°. The modules are filled with a gas mixture of 40% Ar, 50% CO_2 and 10% CF_4, providing the ionization medium.

Each module consists of alternating layers of cathode panels with metallic strips, laid out radially at fixed $\Delta\phi$ intervals, and anode wires, running azimuthally at right angles to the strips. This structure enables a precise tracking of the position of incoming muons in both r and ϕ by correlating the readout signals from cathode strips and anode wires.

Resistive plate chambers (RPC) In order to provide a fast detection signal, RPCs complement the main detection modules (DTs and CSCs) in all areas of the detector. Their response times are significantly lower than the 25 ns between proton bunch crossings, making them suitable for triggering. The modules consist of resistive plates made of Bakelite separated by a 2 mm gap, which contains a gas mixture composed mostly of R134a ($C_2H_2F_3$). The ionization signal is picked up by metallic strips mounted outside the resistive plates, providing information on the muon location.

3.3 Trigger system

The proton bunches circulating at the LHC are crossed at intervals of 25 ns, which corresponds to a frequency of 40 MHz. Multiple individual proton collisions occur during each bunch crossing, resulting in a very high data rate. Since it is not feasible to read out, process and store data from every collision event, a reduction of the rate by many orders of magnitude is necessary.

This is the purpose of the CMS *trigger system*. Its role in the data acquisition process is to identify the collision events which are considered useful for physics analysis or detector calibration. By only writing out these events and discarding the rest, a reduction in the event rate by approximately five orders of magnitude is achieved.

To meet the stringent time requirements imposed by the high rates, the trigger system is divided into two tiers. A hardware-based *Level-One* trigger (L1) first identifies potentially interesting events based on a fast but coarse readout of the detector. These are then forwarded to the *High-Level Trigger* (HLT), which performs a more detailed analysis implemented in software. In this way, an event rate of ca. 100 kHz after the L1 trigger, and finally down to ca. 1 kHz after the HLT is achieved.

3.3.1 Level-One trigger (L1)

The L1 trigger is the starting point of the event processing chain and consists of purpose-built, programmable electronics. These process the raw data coming from the different detector subsystems in order to perform a coarse-grained reconstruction of an event and decide whether it should be kept based on this information.

As shown in figure 3.4, the trigger is organized into a hierarchy of interconnected components. At the lowest level are *Trigger Primitive Generators*, which reconstruct local features in individual detector components, such as energy deposits in calorimeter towers, and track segments or hit patterns in muon system modules.

Next are *Regional Triggers*, which combine trigger primitive information in larger regions of the detector to produce *trigger objects*. These can be electron or photon candidates, or

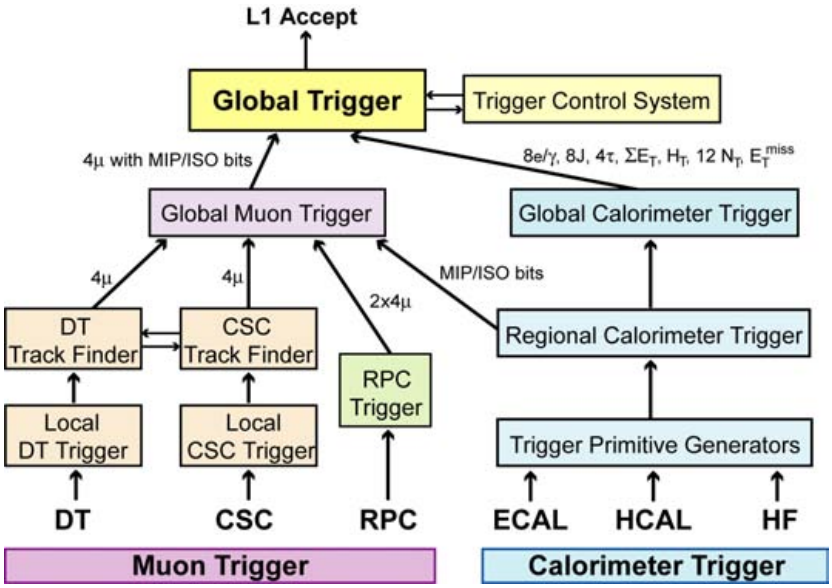

Figure 3.4 – Architecture of the L1 trigger. Information from the calorimeters and muon subdetectors is processed by a hierarchical structure of modules to reconstruct a simplified event representation in the form of trigger objects. The final decision is taken by the Global Trigger based on the trigger object properties. Taken from [46].

regional transverse energy sums in case of the regional calorimeter triggers, or muon candidates for the regional muon triggers. Candidates are ranked and sorted based on energy and quality criteria.

Subsequently, global triggers combine information across regions to produce a complete representation of the event content. The *Global Calorimeter Trigger* (GCT) reconstructs jets using a simplified jet finder algorithm, and computes the total and missing transverse energy. In addition, it selects the highest-rank electron or photon candidates. In the *Global Muon Trigger* (GMT), redundant information from the three muon subsystems is correlated in order to filter out spurious muon candidates, thereby lowering the trigger rate and increasing the efficiency.

The highest-level component is the overall *Global Trigger*, which evaluates the trigger objects provided by the GCT and GMT and takes the decision to either reject the event or accept it for further processing. For this purpose, up to 128 programmable logic pathways operating in parallel on trigger objects can be configured. The pathway decisions are combined by a logical OR function, yielding the final trigger decision. In case of a positive decision, a so-called *L1 Accept* signal (L1A) is issued, which instructs the detector subsystems to send event information to the Data Acquisition system (DAQ).

The decision criteria for each pathway are conceived in such a way as to maintain a manageable event rate. However, since this depends strongly on the instantaneous lu-

minosity, an additional mechanism of controlling the rate called *trigger prescaling* is employed. By configuring an integer prescaling factor k for each pathway, only one in k events meeting the criteria will be triggered, resulting in a k-fold reduction of the rate.

A table of prescaling factors is precomputed for each trigger pathway based on its expected event rate across a wide range of instantaneous luminosities. The factor applied is then adjusted accordingly during data taking, regularly switching to a new set of prescaling factors as the instantaneous luminosity decreases during a run. This enables an optimal use of trigger bandwidth, avoiding potential deadtime or trigger idling due to the event rate being to high or too low, respectively.

An important additional component is the *Trigger Control System* (TCS), which monitors the status of the subdetectors and the DAQ and ensures that an L1A signal is only issued when both systems are in a ready state. In addition, so-called *trigger rules* are implemented in TCS to ensure that L1A signals are not sent too frequently, which could overwhelm the DAQ. The trigger rules allow no more than one L1A signal to be issued in an interval of 75 ns (equivalent to 3 bunch crossings). In addition, extended rules allow no more than 2, 3 or 4 L1A signals per intervals of 0.625, 2.5 and 6 µs (25, 100, 240 bunch crossings), respectively [53].

3.3.2 High-Level Trigger (HLT)

Events which are accepted by the L1 trigger are forwarded to the High-Level Trigger (HLT) for a more detailed analysis. Since the timing requirements at this level are relaxed compared to the L1 trigger, the HLT is implemented entirely in software and is thus capable of running more complex reconstruction algorithms to identify interesting events.

Processing is done on a distributed system with more than 13000 CPU cores [54] called a *filter farm*, allowing events to be evaluated in a highly parallelized fashion. In this way, the event rate is further reduced to a maximum of 1 kHz after the HLT.

The software deployed on the HLT is an optimized version of the CMS software framework (CMSSW) [55], which is also used for the offline analysis of events, although the criteria applied at the HLT-level are often simplified in the interest of processing speed. The trigger criteria are organized into *trigger paths*, each of which delivers a binary accept/reject response for an event based on a series of reconstruction and filtering steps. As for the L1 trigger, prescaling factors can be configured separately for individual HLT paths and adjusted during data taking to optimize the final event rate. Events accepted by at least one HLT path are forwarded to the data storage and are available for offline analysis.

4 Jet measurement at CMS

Jets are produced abundantly at hadron colliders in processes mediated by the strong interaction. They originate from quarks and gluons in the final state of the hard scattering process, which hadronize to form a stream of particles propelled in the same general direction as the original quark or gluon. The connection between jets and partonic final states makes them a valuable experimental handle on these processes, which can be exploited for probing QCD across a wide kinematic range. This chapter provides an overview of the experimental techniques used at CMS to reconstruct and measure jets.

Experimentally, jets are characterized by localized energy depositions in the calorimeters, with a high number of associated particle tracks in the inner tracking detector. While it is possible to reconstruct jets only from calorimeter energies, a better description can be achieved by combining information from all subdetectors, which is done at CMS using an approach called *particle flow* reconstruction. This provides well-defined particle candidates that can be clustered to yield jets with a significantly improved resolution. Section 4.1 gives an overview of particle-flow reconstruction.

In addition, the composite nature of jets makes them susceptible to other factors such as pileup or the detector response, which prompts the need for a dedicated calibration procedure, which is described in section 4.2.

Finally, the strategies employed to identify jets on the trigger level and the associated efficiencies are discussed in section 4.3.

4.1 Particle-flow reconstruction

Particles created in high-energy proton collisions propagate through the CMS detector, producing characteristic signals in the various subdetectors. The raw data collected in this way consists of comparatively low-level information, such as individual points on the trajectories of charged particles, or the magnitude of energy depositions in the calorimeters. To analyze this data in a meaningful way, it is first necessary to reconstruct high-level objects such as particles and jets, as well as other quantities of interest for physics analysis.

At CMS, this is done using an approach called *particle flow* (PF), which combines information from all subdetectors to identify and measure individual particles. This exploits the strengths of each subdetector, leading to an improved event description.

In this section, a brief overview of the PF approach is given, starting with the reconstruction of basic elements in the individual subdetectors, and outlining the strategies used to obtain well-defined particle candidates by connecting these elements together. A more detailed description can be found in references [56, 57].

4.1.1 Basic elements

The first step in the particle-flow approach consists in reconstructing the basic elements specific to each subdetector: tracks left by the charged particles in the inner tracking system and the outer detector layers, and clusters of energy depositions in the electromagnetic and hadronic calorimeters.

Tracks The trajectories of charged particles are reconstructed from hits in the inner tracker layers using a combinatorial track finding algorithm [58]. Starting from a small number of hits in consecutive layers of the pixel detector called a *seed*, the initial trajectory is extended into the tracker volume, collecting hits from additional layers along the way. The track parameters such as the initial direction, the originating vertex, and the transverse momentum are determined in a least-squares fit. Finally, the resulting tracks are filtered based on quality requirements, such as requiring a minimum number of hits or rejecting tracks originating from a vertex too far from the interaction region.

In the muon chambers, the particle flux is significantly reduced compared to the inner detector, largely due to the high probability of absorption of all other detectable particles in the calorimeter layers. This allows muon tracks to be reconstructed with a high efficiency and a comparatively low misidentification rate. Similar reconstruction techniques are employed as for the inner tracker, starting from hits in the DTs and CSCs as track seeds and gathering hits from other subdetectors along the initial seed direction to reconstruct track segments.

Calorimeter clusters Energy deposited in the calorimeters is resolved with a granularity given by the cellular structure of each subdetector. In the ECAL, a cell consists of a single lead tungstate crystal, while the HCAL granularity corresponds to the calorimeter towers, corresponding to an array of 5×5 ECAL crystals.

Calorimeter clusters are reconstructed by first identifying all cells with energy depositions larger than those of the neighboring cells above a predetermined threshold, called *cluster seeds*. These are extended to form *topological clusters* by sequentially incorporating neighboring cells with energies above the noise threshold. The final positions and energies of calorimeter clusters are determined in a maximum likelihood fit, assuming Gaussian profiles of energy depositions around the seeds within a topological cluster.

4.1.2 Particle-flow candidates

After the basic elements have been identified, the particle content of the event is reconstructed in the form of so-called *particle-flow candidates* (PF candidates). This is done by first linking together potentially related elements based on their relative locations in the detector, and subsequently analyzing the sets of linked elements to determine their compatibility with different particle hypotheses.

Particle tracks are extrapolated to the calorimeter regions and linked to clusters there based on the distance in (η, ϕ) space (or the xy plane for the end caps). Similarly, tracks reconstructed in the inner detector are linked to track segments in the muon chambers if their extrapolated positions are compatible, and calorimeter clusters in the ECAL and HCAL are linked together if they contain overlapping cells. Ambiguous links are resolved by choosing the one with the smallest distance.

Elements connected by a link, either directly or through an intermediate related element, form a so-called *PF block*. The final PF candidates are obtained by looking for signatures specific to each particle type inside the individual PF blocks, ensuring the unique association of basic elements to PF candidates by removing the elements associated to successfully identified particles from PF blocks. The reconstruction of PF candidates in each PF block proceeds in a predefined order, starting with muons, followed by electrons and isolated photons, and finally hadrons and nonisolated photons, as described in the following:

Muons In a first step, muons are identified by correlating the information of matched tracks in the inner and outer detector layers, in association with calorimeter clusters compatible with minimum ionizing particles. The muon momentum is obtained exclusively from the inner track at values below 200 GeV, for which the tracker provides a sufficiently high momentum resolution. At higher values, a least-squares fit is performed including points on the outer track segments to obtain the global muon momentum.

Electrons and isolated photons Next, electrons are reconstructed from tracks linked to ECAL clusters. An important property exploited for distinguishing between electrons and charged hadrons is the high probability of the former to emit soft bremsstrahlung photons in the longitudinal magnetic field. To take this into account, ECAL clusters spread over an extended region in ϕ are grouped into so-called *superclusters*. In addition, tracks with a lower goodness-of-fit are refitted using an alternative method that better accounts for progressive energy losses due to bremsstrahlung. The radiated photons are collected by extrapolating the tangents of electron tracks to the ECAL, and the energy of associated clusters in these regions is counted towards the electron energy.

Electron candidates are then discriminated from charged hadrons using a multivariate approach involving fourteen variables related to the linked clusters and tracks. In the same step, isolated photons are identified from ECAL superclusters without an associated particle track, provided the energy depositions in the associated ECAL and HCAL clusters are consistent with an electromagnetic shower.

Hadrons and nonisolated photons Finally, the remaining elements are processed to reconstruct charged and neutral hadrons, as well as nonisolated photons. Within the tracker acceptance, calorimeter clusters without associated tracks are interpreted as photons, while hadrons are given precedence outside this region in the presence of an HCAL cluster. Charged hadrons are identified by calorimeter clusters linked to particle tracks. By comparing the calibrated calorimetric energy with the track energy, the presence of overlapping neutral hadrons is inferred.

4.1.3 Missing transverse momentum

Missing transverse momentum (MET[1]), is defined as the negative sum of the transverse momenta of all particles produced in a scattering interaction. In the idealized case of a perfectly calibrated detector, and where no particles escape without being detected, this quantity is expected to be zero, on account of the negligible transverse momentum of the colliding protons.

In practice, however, a finite amount of missing transverse momentum is observed for every reconstructed event. This can either have a physical origin – stemming, for instance, from the production of undetectable particles such as neutrinos –, or it can be traced back to experimental causes such as detector noise, imperfections in the detector calibration or alignment, gaps in the detector acceptance or the finite detector resolution. The former is referred to as *genuine* or *true MET*, while distortions due to the latter are called *anomalous* or *fake MET*.

[1]The abbreviation is derived from the *missing transverse energy*, which corresponds to the magnitude of the four-vector associated to the missing transverse momentum.

In the particle-flow approach, the missing transverse momentum is estimated by taking the negative vector sum of transverse momenta of all PF candidates, which is referred to as the *PF MET*. Since this quantity is only a rough estimate of the true MET, it is refined by applying further corrections, as described in section 4.2.5.

4.1.4 Jet reconstruction

The particle-flow approach described in the previous sections results in a collection of PF candidates with well-defined four-momenta which can be used as an input to jet clustering algorithms. Jets obtained in this manner are referred to as *detector-level* or *reconstruction-level* jets, since the kinematic properties of the clustered objects are ultimately reconstructed from detector signals and are therefore subject to detector-induced biases and resolution effects.

This is contrasted to the concept of *particle-level* jets, which are defined as the result of applying clustering algorithms directly to the stable particles produced in high-energy collisions. While inaccessible in actual measurements, particle-level jets are a useful notion in the context of simulations, providing a reference for studying the detector response.

Standard jet reconstruction at CMS is done using the anti-k_T algorithm [45]. PF candidates are clustered for two values of the radius parameter, $R = 0.4$ and $R = 0.8$, yielding two separate jet collections. In simulated samples, particle-level jets are additionally clustered from the stable decay products of particles produced in the hard process, with the exception of neutrinos. The clustering algorithm is implemented in the FastJet package [59], interfaced to the CMS software framework CMSSW [55].

4.1.5 Pileup mitigation

In the high-luminosity environment of the LHC, multiple proton collisions known collectively as *pileup* occur during the same bunch crossing as the hard scattering, giving rise to additional tracks and energy depositions in the calorimeters. With more than 20 pileup interactions happening on average during each bunch crossing, this affects the reconstruction of jets emerging from the hard process, resulting in an additional contribution to the reconstructed transverse momentum of jets, which has to be compensated.

The strategy employed to address this involves two complementary approaches: the targeted subtraction of pileup contributions from charged hadrons using track and vertex information, and an overall technique in which the per-jet pileup contribution is estimated based on the jet area and the diffuse energy density in the event.

Charged hadron subtraction

The excellent position resolution of the tracker allows pileup interactions to be inferred from the additional *primary vertices* identified by the track reconstruction. These are determined by grouping together tracks that originate at nearby longitudinal positions along the beam line, and performing a least-squares fit to determine the vertex position [60]. The vertex with the highest p_T sum of associated tracks is assumed to correspond to the hard scattering, while the other vertices are identified with pileup interactions.

In the particle-flow approach, jet clustering is run indiscriminately on PF candidates originating from both the hard scattering and pileup vertices, causing a degradation of the jet energy measurement. To mitigate this, a pileup removal algorithm called *charged hadron subtraction* (CHS) [61] is applied during offline jet reconstruction to exclude all charged hadrons associated to pileup vertices from the jet clustering.

This approach successfully suppresses a large part of the pileup contribution. However, it does not account for contributions from neutral and nonhadronic pileup particles, and is not applicable outside the tracker coverage. To estimate the remaining pileup contributions, a more sophisticated approach involving jet areas is used and a corresponding correction is applied as part of the jet calibration workflow (see section 4.2).

Jet area

An alternative approach to pileup mitigation is based on the notion of *jet area*, which is a measure of the susceptibility of jets to contributions from diffuse sources [62]. The pileup contribution to the transverse momentum of jets can then be factorized in terms of the jet area and the overall level of diffuse energy in the event [63].

Jet areas are specific to the jet clustering algorithm and are determined by clustering the original jet constituents together with a set of infinitely soft "ghost" particles, distributed evenly in rapidity and azimuth to represent a diffuse energy background. The jet area is proportional to the number of ghost particles clustered as part of a jet. For the anti-k_T algorithm with a radius parameter R, the resulting areas for well-separated jets are roughly circular in (η, ϕ) space, with an area of approximately πR^2, while for jets in close proximity, this results in a crescent-shaped area for the lower-energy jet.

To estimate the diffuse energy present in an event, the accessible (η, ϕ) range is divided into a grid of equally-sized cells. The p_T density is determined as the ratio of the total transverse momentum of PF candidates in each cell to the cell area. The median across all cells is taken as a robust estimator of the overall diffuse p_T density ρ:

$$\rho = \text{median} \left\{ \frac{p_T^{\text{cell}}}{A_{\text{cell}}} \right\} \tag{4.1}$$

4.2 Jet energy calibration

An essential part of jet measurements consists in relating the kinematic properties of reconstructed jets with those of stable particles emerging from the fragmentation of partons produced in the hard scattering process. Among these, the jet energy is particularly susceptible to contributions from unrelated sources such as pileup or the underlying event, and to the detector response relating the true and measured energies of jets and their constituents.

While the jet energy measurement is significantly improved as a consequence of the particle-flow approach, the jet energy scale requires further calibration to account for these effects. At CMS, this is achieved in a multi-stage approach, in which correction factors are derived with the aid of simulations and data-driven techniques, and applied successively to the raw jet four-momenta to obtain the corrected jet energy.

Figure 4.1 shows an overview of the jet calibration workflow, highlighting the different stages in which correction factors are derived and applied to jets in recorded data and Monte Carlo simulation. The first stage involves the correction of the energy offset caused by particles originating from pileup interactions. This is followed by the correction of the detector response, which connects the energy of jets clustered from PF candidates after the detector simulation to that of the particle-level jets, as determined directly from the simulated event record. Finally, additional corrections are derived for recorded data in order to account for residual differences between the simulated and actual detector response.

The derivation of correction factors for these steps is outlined in the following. The described methods were originally developed for the analysis of CMS data taken during Run 1 of the LHC at center-of-mass energies of 7 and 8 TeV, and adapted to 13 TeV data. A more detailed description of the methods is given in references [64] and [65].

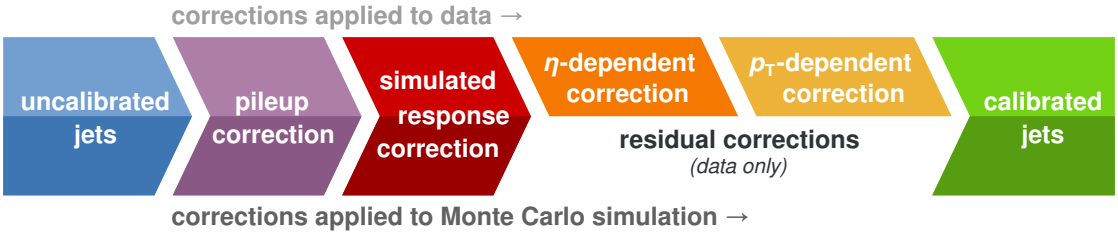

Figure 4.1 – Overview of the jet energy calibration workflow. Shifts in the jet energy due to pileup and the simulated detector response are estimated with the aid of Monte Carlo simulations and corrected accordingly in both simulated and recorded events. Residual differences between data and simulation are accounted for using dedicated corrections applied only in data. Adapted from [64].

4.2.1 Pileup offset correction

The first step in the jet calibration workflow aims to compensate for additional contributions to the jet energy from pileup. The magnitude of this shift, referred to as the *pileup offset*, is estimated in a simulation-based approach by comparing the reconstructed jet p_T in hard-scattering events with and without an admixture of pileup interactions.

The correction is derived as a multiplicative factor, expressed as a function of the raw (uncalibrated) jet p_T, the jet pseudorapidity η, the jet area A and the median p_T density of the event ρ:

$$c_{\text{pileup}}(p_T^{\text{raw}}, \eta, A, \rho) = 1 - \frac{\left[\rho_0(\eta) + \rho \cdot \beta(\eta)\left(1 + \gamma(\eta) \cdot \log\left(p_T^{\text{raw}}\right)\right)\right] \cdot A}{p_T^{\text{raw}}} \tag{4.2}$$

The numerator in the above expression corresponds to the absolute pileup offset. The parameters $\rho_0(\eta)$, $\beta(\eta)$ and $\gamma(\eta)$ account for the nonuniform response to pileup in different pseudorapidity regions and are determined from parametrizations of the pileup offset, performed separately in each bin as a function of the remaining correction variables.

Residual differences between data and simulation are acocunted for using an alternative method of estimating the pileup offset, known as the *random cone* method. Using events from pileup-only simulations and so-called *zero-bias* data, where no hard scattering interactions are present, the offset is determined using randomly placed cones in (η, ϕ) and summing the transverse momenta of PF candidates inside the cone. A scale factor, obtained separately in each pseudorapidity region from the ratio of the average measured and simulated random cone offsets, is then multiplied to the corresponding correction parameters $\rho_0(\eta)$ and $\beta(\eta)$. The average pileup offset per pileup interaction, determined as a function of η using the random-cone method, as well as the data/simulation scale factor can be seen in figure 4.2.

4.2.2 Simulated response correction

After the pileup correction has been applied, the jet energy is corrected for the detector response, which connects the energy of particle-level jets to that of jets clustered from PF candidates. The response is estimated from simulated event samples, which contain a full detector simulation based on the Geant4 toolkit [67], allowing detector effects to be studied in detail.

To determine the detector response, jets in simulated samples are first matched to their particle-level counterparts based on the distance ΔR in (η, ϕ) space. The response is then estimated in bins of η and p_T from the ratio of the average transverse momenta of the

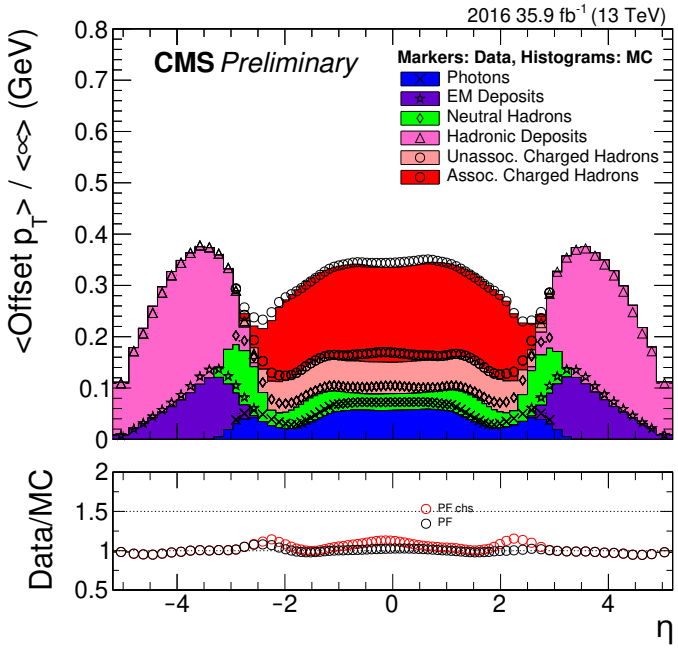

Figure 4.2 – Average pileup contribution to the jet transverse momentum per pileup interaction, shown for data (markers) and simulation (colored areas) for each type of particle or calorimeter deposit. The charged hadron contribution removed from jets by the CHS pileup subtraction method is shown separately. The remaining contributions are compensated by the pileup offset correction to the jet energy. Taken from [66].

matched reconstructed and particle-level jets:

$$R_{\text{ptcl}} = \frac{\langle p_{\text{T}}^{\text{reco}} \rangle}{\langle p_{\text{T}}^{\text{ptcl}} \rangle} \tag{4.3}$$

Figure 4.3 shows the simulated jet response as a function of η in the different p_{T} regimes. As can be seen, the varying detector composition in each pseudorapidity slice leads to significant differences in the response behavior, which are compensated by the simulated response correction.

4.2.3 Residual corrections

The derivation of the corrections outlined above relies heavily on simulated samples. In order to account for any remaining differences observed between simulation and data, a further set of so-called *residual corrections* is derived in recorded data using a series of *in situ* calibration techniques.

The residual correction is obtained in two stages. First, relative calibration factors are obtained as a function of the jet pseudorapidity η with respect to the energy of jets in the central region at $|\eta| < 1.3$. Subsequently, the absolute jet energy scale is measured in the central region, using jets that are transversely balanced against different types of well-measured reference objects. This approach takes advantage of the high reconstruction accuracy in the central region to calibrate jets across the entire pseudorapidity range.

For the relative calibration stage, dijet events are selected for which the highest-p_T jet is reconstructed within $|\eta| < 1.3$, and the response is determined by comparing the reference jet p_T to that of the second jet as a function of the latter's pseudorapidity. The imbalance resulting from the presence of additional jets is compensated by measuring the response as a function of the ratio of the third jet p_T to the average dijet transverse momentum, and extrapolating to zero.

To determine the absolute jet energy scale, events are selected in which jets are transversely balanced against well-measured objects such as photons or Z bosons reconstructed from electron and muon pairs. As for the relative calibration, the response is measured in bins of the additional jet activity α, defined as the ratio of the second-leading jet p_T to that of the reference object, and extrapolated to the ideal topology at $\alpha = 0$.

The response is determined using two complementary methods, which are sensitive to different biases. In the p_T *balance* method, the response is defined directly as the ratio of the jet p_T to be calibrated and that of the reference object:

$$R_{p_T} = \frac{p_T^{\text{jet}}}{p_T^{\text{ref}}} \tag{4.4}$$

While this definition is effective in describing local influences on the jet reconstruction, it is very sensitive to the additional jet activity, leading to large uncertainties due to the α extrapolation. To avoid this, a second response definition is introduced based on the fact that, for a well-calibrated jet energy scale, the missing transverse momentum in Z+jets and γ+jets events is expected to be zero. The projection of the missing transverse momentum in the event onto the p_T of the reference object can therefore be interpreted as a measure of the miscalibration of the jet energy scale, leading to the so-called *MPF response* definition:

$$R_{\text{MPF}} = 1 + \frac{\mathbf{p}_T^{\text{miss}} \cdot \mathbf{p}_T^{\text{ref}}}{\left(p_T^{\text{ref}}\right)^2} \tag{4.5}$$

The absolute calibration factors for the jet energy scale are then determined as a function of p_T in a global fit to all measurements. At high p_T, the response is additionally constrained using multijet events by balancing the recoil of the multijet system against a high-p_T central jet. The resulting response after applying both the relative and the absolute the calibration factors is shown on the right in figure 4.3.

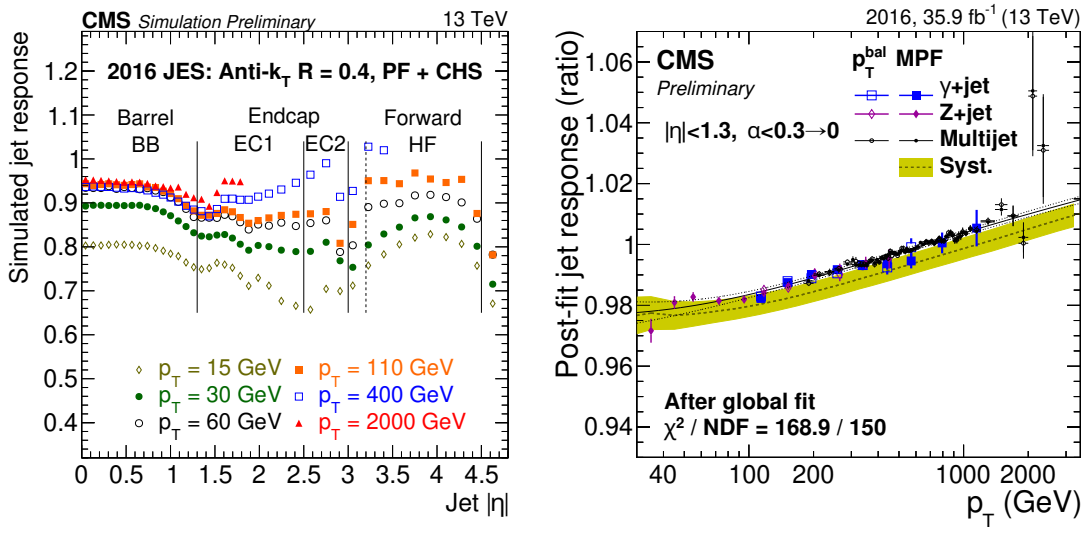

Figure 4.3 – Simulated jet response (*left*), shown as a function of the absolute jet pseudorapidity $|\eta|$ for selected values of the jet transverse momentum. Residual differences between data and simulation are accounted for in a global fit of the absolute jet energy scale, determined by balancing the jets against well-measured objects within $|\eta| < 1.3$ (*right*). Taken from [66].

4.2.4 Jet energy resolution

The measurement of the jet energy is limited by the finite detector resolution, which manifests as a random spread of the reconstructed jet p_T with respect to the particle-level jet. This effect is of particular importance when performing a differential measurement of the steeply falling jet p_T spectrum, which is systematically shifted towards higher values as a result of the convolution with the detector resolution.

The jet p_T resolution is estimated using simulated samples by examining the distribution of the particle-level response, which is defined as the p_T ratio of matched reconstructed and particle-level jets. The distribution is fitted with a Gaussian function in different p_T and pseudorapidity regions, and the width parameter σ is taken as an estimate of the resolution.

To ensure that the resolution is modeled accurately to reflect the actual measurement conditions, scale factors are derived by comparing recorded and simulated events. Alternative estimates of the jet p_T resolution are obtained from the distribution of the p_T asymmetry of the two leading jets in dijet events, and the p_T balance distribution in γ+jets and Z+jets events. As in the case of residual corrections, additional jet activity is extrapolated to zero to obtain the final resolution estimates. The scale factor is determined in bins of $|\eta|$ as the ratio of the resolution in data and simulation.

To adjust the p_T resolution to that in data, the four-momentum of each jet is smeared by

a random multiplicative factor c_{JER}, derived as follows:

$$c_{\mathrm{JER}} = 1 + \mathcal{N}(0, \sigma_{\mathrm{JER}}) \sqrt{\max(s_{\mathrm{JER}}{}^2 - 1, 0)} \tag{4.6}$$

In the above equation, $\mathcal{N}(0, \sigma_{\mathrm{JER}})$ indicates a random number sampled from a normal distribution with a mean of zero and a width parameter corresponding to the relative jet p_{T} resolution, as obtained from simulation. This number is then scaled further by an expression involving the simulation-to-data scale factor s_{JER}. The maximum function prevents the resolution from being degraded if the scale factor should drop below unity. Nevertheless, this is not observed to be the case. The scale factors obtained can be seen in figure 4.4.

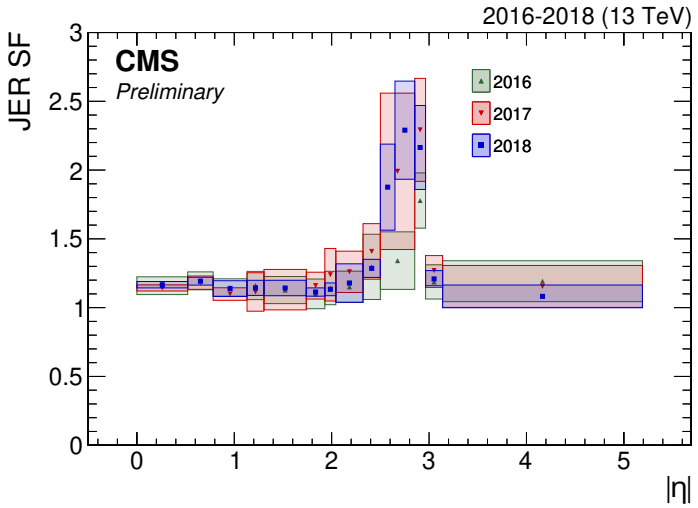

Figure 4.4 – Simulation-to-data scale factors for the jet energy resolution (JER). Shown are the scale factors obtained for the data taking periods 2016–2018 in bins of the absolute jet pseudorapidity $|\eta|$. Taken from [66].

4.2.5 Missing transverse momentum correction

The corrections presented in the previous sections account for systematic biases affecting the jet energy on a jet-by-jet basis. This introduces a discrepancy between the underlying PF candidates clustered as part of the jet and the final corrected jet transverse momentum. Since the missing transverse momentum (MET) in an event is calculated based on the PF candidates, the corrections must be propagated to the MET to ensure consistency.

This is done by applying the so-called *type-I* MET correction. For every jet above a threshold of 15 GeV, the difference between the raw (uncalibrated) PF jet and the fully

corrected jet is added back to the raw PF MET. To avoid anisotropies caused by the subtraction of pileup clustered into jets, which could result in a MET bias, the jet pileup offset estimated with the random cone method, denoted $\mathbf{O}_{\mathrm{RC}}$, is resubtracted. The *type-I*-corrected MET is then given by:

$$\mathbf{p}_{\mathrm{T}}^{\mathrm{miss,corr}} = \mathbf{p}_{\mathrm{T}}^{\mathrm{miss,raw}} + \sum_i \mathbf{p}_{\mathrm{T},i}^{\mathrm{jet,raw}} - \sum_i \mathbf{p}_{\mathrm{T},i}^{\mathrm{jet,corr}} - \sum_i \mathbf{O}_{\mathrm{RC},i} \tag{4.7}$$

4.3 Jet triggers

The presence of jets in events is detected by algorithms implemented in both the L1 trigger and the high-level trigger (HLT).

At the L1 trigger level, a full jet reconstruction cannot be performed due to the stringent time constraints. Instead, jet-like objects are reconstructed from regions of high activity in the electromagnetic and hadronic calorimeters, providing only a rough estimate of the jet energy and location. Several jet-related L1 pathways are implemented, which deliver a positive trigger decision if jet-like objects with transverse momenta p_{T} surpassing preset thresholds are found.

At the HLT, two different categories of jet trigger paths are implemented. The simpler, *single-jet* trigger paths identify the jet with the largest transverse momentum p_{T} and accept the event if this value lies above a preconfigured threshold. In addition, a set of dedicated *dijet trigger paths* exists, for which the trigger decision is computed based on the average transverse momentum $\langle p_{\mathrm{T}} \rangle_{1,2}$ of the two highest-p_{T} jets.

To take advantage of the L1 trigger decisions, HLT jet paths are seeded by a suitable L1 jet trigger pathway, meaning that events which were not accepted by this pathway are also rejected by the HLT. The L1 pathway is chosen so that its p_{T} threshold is sufficiently below that of the HLT path and thus corresponds to looser event selection criteria. If no such path is available, so-called *zero bias* pathways are used as L1 seeds, which deliver random trigger decisions that are not based on the event content.

For events passing the L1 seed requirement, jets are clustered from particle-flow candidates using the anti-k_{T} algorithm. For single-jet triggers, the clustering is done for both of the standard radius parameter values used in CMS, $R = 0.4$ and $R = 0.8$, with a separate set of HLT paths for each radius, while for the dijet trigger paths, only jets with $R = 0.4$ are considered.

To meet the time requirements imposed on the trigger, only a simplified version of the track reconstruction is run at the trigger level compared to the full offline reconstruction. Due to the limited momentum resolution of the simplified reconstruction, the trigger efficiency exhibits a characteristic "turn-on" behavior as a function of the offline-reconstructed p_{T} (or $\langle p_{\mathrm{T}} \rangle_{1,2}$), reaching full efficiency only at values much larger than the nominal trigger threshold. Figure 4.5 shows an example of typical trigger efficiency curves, determined in 2016 data for a representative selection of single-jet trigger paths.

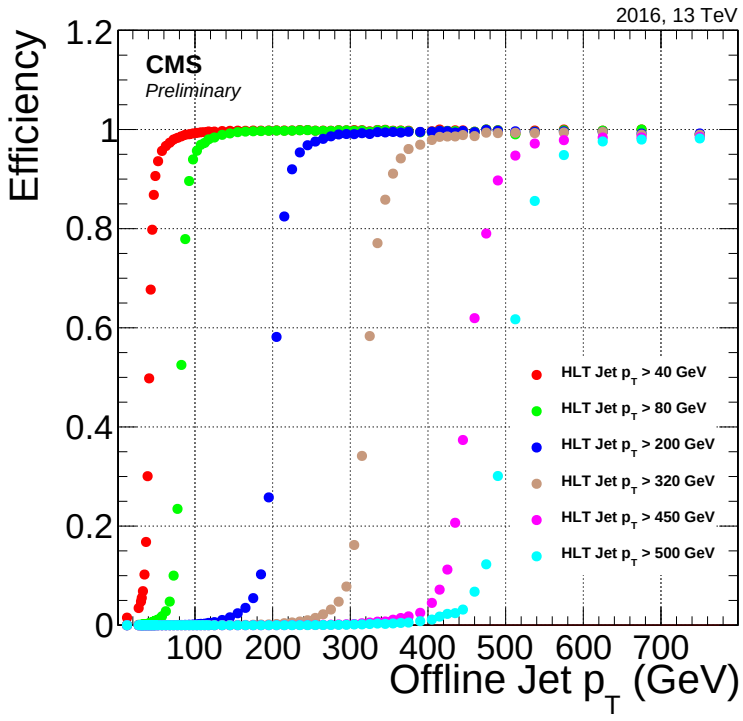

Figure 4.5 – Illustration of typical efficiency turn-on curves, determined for a representative selection of single-jet triggers paths from a subset of the 2016 data. The efficiency of each trigger path increases as a function of the offline-reconstructed jet p_{T}, reaching values of ca. 50% at the nominal trigger threshold value and full efficiency only at significantly larger values. Taken from [68].

4.3.1 L1 trigger prefiring

Data taken in 2016 is affected by a further source of trigger inefficiency which is due to an association of L1 trigger ECAL primitives to the incorrect bunch crossing (BX), leading to a so-called *prefiring effect*.

The effect is determined by a negative shift in the absolute ECAL timing, which is known to be caused by the progressive loss of ECAL crystal transparency as a result of radiation damage over time. The shift has an impact on the timing of L1 trigger primitives, causing them to be incorrectly associated to the preceding BX, hence *pre*firing.

To counteract this, a time offset is implemented in the L1 trigger, whose value is optimized to reduce the prefiring rate. However, during 2016, the accumulation of ECAL radiation damage caused a timing shift for which the time offset was not adequately adjusted for data taking. This resulted in a dramatic increase in prefiring rates in a portion of the endcap region of the ECAL ($2 < |\eta| < 3$).

The inefficiency due to prefiring itself results from the L1 trigger rules put in place to

avoid an overflow of the readout buffers (cf. section 3.3.1). One such rule prevents more than one event from being read out within three consecutive bunch crossings. If this occurs, the detector is only read out for the first triggered bunch crossing, while any others are vetoed.

When prefiring occurs, this causes the detector readout to be triggered during the early BX, which was mistakenly associated to the prefired L1 primitive, while the correct event associated to the in-time BX is discarded. In addition, the mistimed detector readout likely leads to an event which contains no interesting information and is thus discarded by the HLT or during the offline analysis.

This inefficiency is not accounted for in simulation and its estimation is not directly possible in data using standard techniques: since all affected events are discarded during data taking, there is no way to obtain a reference sample in the offline analysis. However, it is possible to identify a subset of so-called *unprefireable* events, for which a mistimed detector readout is prevented by the L1 trigger rules.

This is the case when a readout is triggered for events that are exactly three bunch crossings apart, denoted here BX-3 and BX0 for simplicity. Assuming that prefiring occurs for the later event in BX0, the readout would be triggered incorrectly for BX-1, while BX0 would be vetoed. However, if a readout is triggered for BX-3, a readout for BX-1 would be forbidden by the trigger rules. This ensures that BX0 is correctly read out despite the prefiring.

Using the set of unprefireable events, an estimate of the prefiring probability can be obtained by determining the fraction of events for which a prefired L1 primitive exists in the preceding bunch crossing. Due to the relative rarity of unprefireable events, this approach is impractical for use by individual analyses. Instead, the prefiring efficiency is estimated centrally using a sufficiently large analysis-independent sample and provided as a function of the transverse momentum p_T and the pseudorapidity η of reconstructed objects which are likely to cause prefiring. Figure 4.6 shows the prefiring efficiency measured for jets in the affected pseudorapidity regions.

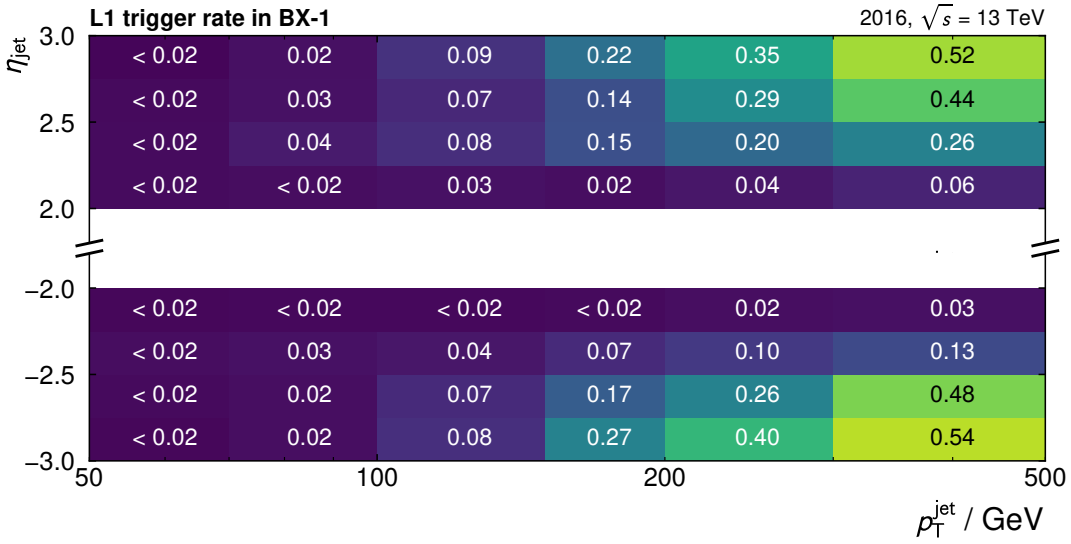

Figure 4.6 – Prefiring probability, estimated in a special set of unprefireable events as the fraction of events with a positive L1 trigger decision misattributed to the previous bunch crossing BX-1. The misattributed trigger decision is due to calorimeter deposits at pseudorapidities $2 \leq |\eta| \leq 3$ and is therefore associated with jets in these regions. Shown are the fractions of prefired events in bins of the jet pseudorapidity and transverse momentum.

5 Triple-differential cross section measurement

Jet observables at hadron colliders provide an excellent premise for precision measurements, which are essential for constraining theoretical parameters such as PDFs and reducing the theory-induced uncertainties in many analyses. The production of dijet pairs is particularly well-suited for this purpose. Having one of the highest cross sections at hadron colliders, a high degree of statistical precision can be reached. In addition, the configuration of the dijet system can be exploited to improve the PDF sensitivity.

This chapter presents a triple-differential measurement of the dijet cross section performed using proton collision data recorded at CMS during 2016 at a center of mass energy of 13 TeV. The data samples analyzed are outlined in section 5.1.

Dijet events are identified by applying a set of selection criteria, which are described in section 5.2. The differential cross section is then measured as a function of three observables describing the kinematics of the dijet system. Two measurements are performed in parallel: one for the variables y^*, y_b, and the average dijet transverse momentum $\langle p_T \rangle_{1,2}$, and another using the dijet invariant mass m_{jj} as an alternative observable. The measurement phase space and binning are detailed in section 5.3.

Section 5.4 presents a comparison of the main kinematic variables in recorded data and samples obtained from Monte Carlo simulation. The latter are used to study to gauge the impact of the detector response and resolution on the measurement, which are corrected for using a three-dimensional unfolding procedure, described in section 5.5.

The main uncertainties affecting the measurement are discussed in section 5.6. Finally, the unfolded cross sections are compared to fixed-order calculations at next-to-leading (NLO) and next-to-next-to-leading order (NNLO) accuracy in perturbative QCD (section 5.7).

5.1 Data samples

The measurement of the triple-differential dijet cross section presented here is based on proton-proton collision data collected at a center-of-mass energy of 13 TeV by the CMS detector in 2016. The total data set corresponds to an integrated luminosity of 35.9 fb^{-1}.

Events recorded by CMS are grouped into so-called primary datasets depending on the trigger path by which they were accepted. The `JetHT` primary dataset contains all the events accepted by jet-related trigger paths, and is therefore chosen for this measurement.

During data taking, events are organized into consecutively numbered *runs*, during which the detector conditions remain largely unchanged. The length of a run is not predetermined and can vary between a few seconds and several hours, depending on the necessity to make adjustments to the detector parameters. Runs are further subdivided into *luminosity blocks*, which correspond to a data-taking interval of roughly 25 seconds and during which the instantaneous luminosity is approximately constant.

Following data taking, each run and luminosity block is verified to ensure the well-functioning of all detector components in a process called *data certification*. This analysis uses only certified runs and luminosity blocks, a list of which is centrally provided by the CMS collaboration.

In order to better gauge the impact of changing detector conditions over time, data collected in 2016 is divided chronologically into seven data taking periods labeled alphabetically from B to H, each covering a range of consecutive runs. The data samples are cataloged along with the associated metadata in the CMS Data Aggregation System (DAS) [69]. An overview of the data taking periods, together with the corresponding DAS datasets and the corresponding integrated luminosities can be seen table 5.1.

Simulated samples

To facilitate a better understanding of the data, a selection of simulated samples is provided by the CMS collaboration for a wide range of fundamental processes. These are obtained using Monte Carlo (MC) event generators and include not only a simulation of the hard scattering process, but also incorporate contributions from pileup and a full simulation of the detector. This allows them to be compared directly to recorded data and provides a means to gauge the impact of experimental and theoretical effects on the measurement.

The main process of interest in this analysis is QCD jet production, which is simulated using the Pythia 8 event generator [29]. The hard scattering amplitude for this process is calculated at leading order (LO) in perturbation theory, with additional higher-order contributions from QCD radiation being modeled using the parton shower approach.

Table 5.1 – Overview of proton collision datasets recorded by CMS in 2016. Events are separated into seven data-taking periods, each of which covers a range of consecutive runs and corresponds to a particular integrated luminosity.

Period	Dataset(s)	Luminosity / fb^{-1}	Run range
B	`/JetHT/Run2016B-17Jul2018_ver1-v1/MINIAOD`, `/JetHT/Run2016B-17Jul2018_ver2-v2/MINIAOD`	5.7	272007–275376
C	`/JetHT/Run2016C-17Jul2018-v1/MINIAOD`	2.6	275657–276283
D	`/JetHT/Run2016D-17Jul2018-v1/MINIAOD`	4.3	276315–276811
E	`/JetHT/Run2016E-17Jul2018-v1/MINIAOD`	4.0	276831–277420
F	`/JetHT/Run2016F-17Jul2018-v1/MINIAOD`	3.1	277772–278808
G	`/JetHT/Run2016G-17Jul2018-v1/MINIAOD`	7.5	278820–280385
H	`/JetHT/Run2016H-17Jul2018-v1/MINIAOD`	8.6	280919–284044

The simulation of the underlying event is configured using the CUETP8M1 tune [36]. The hadronization of final-state partons is carried out using the Lund string model.

To simulate pileup, a random number of additional so-called *minimum-bias* interactions are simulated and the resulting particles are appended to the products of the hard scattering. The mean number of pileup interactions expected is determined from the overall minimum-bias cross section and the instantaneous luminosity, and the exact number is drawn from a Poisson distribution with a corresponding expectation value.

The interaction of the generated particles with the detector is simulated using the Geant4 toolkit [67]. The simulated signals obtained in this way are processed employing the same reconstruction methods as for recorded data.

Stitching

Differential jet cross sections depend strongly on the p_T of the jets, decreasing exponentially with p_T over several orders of magnitude within the experimentally accessible phase space. This makes direct MC sampling across the entire p_T range inefficient, leading to oversampling at low p_T where the cross section is highest, and a sparsely populated high-p_T region.

To avoid this, the phase space is first partitioned into several ranges, chosen such that the cross section decrease in each range is small compared to its average value, and an adequate number of events are then sampled in each range.

The Pythia 8 QCD samples used here are sampled in bins of the generator variable $\hat{p}_\text{T}$, which represents the characteristic transverse momentum of the hard process. The samples comprise a total of 14 regions, commonly referred to as *slices*, which cover the $\hat{p}_\text{T}$ region above 30 GeV.

The final spectrum is obtained as a weighted distribution, where all events from a slice are assigned a *stitching weight* based on the total cross section calculated for the corresponding $\hat{p}_T$ range. This is equivalent to normalizing the slices to the same equivalent luminosity. Denoting the number of events in a $\hat{p}_T$ slice as $N_{\text{total}}(\hat{p}_T)$, and the total cross section there as $\sigma_{\text{total}}(\hat{p}_T)$, the stitching weight is given by:

$$w_{\text{stitching}}(\hat{p}_T) = \sigma_{\text{total}}(\hat{p}_T)/N_{\text{total}}(\hat{p}_T) \tag{5.1}$$

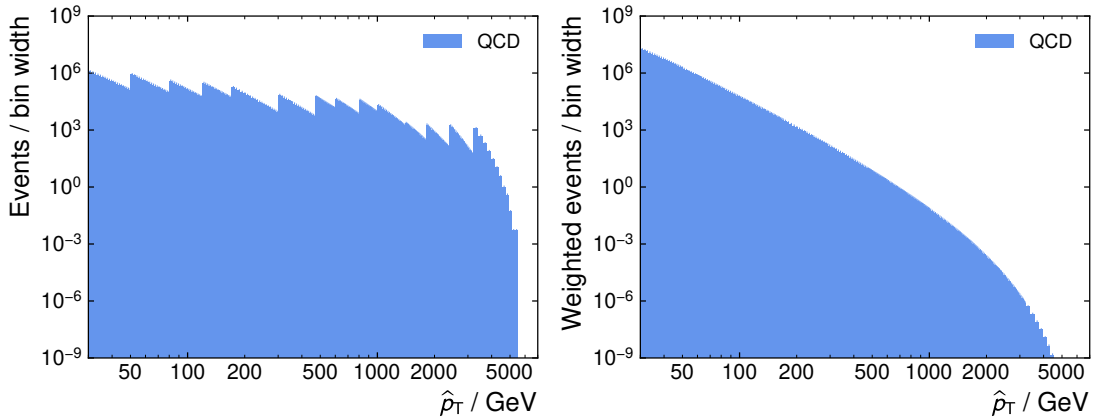

Figure 5.1 – Distribution of the generator variable $\hat{p}_T$ before (*left*) and after (*right*) applying stitching weights to normalize each slice to the same equivalent luminosity. The weighting produces a spectrum with a characteristic steeply falling shape and eliminates the discontinuities between $\hat{p}_T$ slices.

An overview of the samples, together with the associated cross sections can be found in table A.1 in section A.1 of the appendix.

Background processes

In addition to QCD jet production, a number of unrelated background processes can also result in dijet pairs. This is mainly due to electroweak vector boson production, where jets may arise from initial and final state QCD radiation, or the production of top quark pairs, which decay to lighter quarks before hadronization, resulting in secondary jets.

The cross section for these processes is orders of magnitude below that of QCD jet production, so that their contribution to the main measurement can be neglected. Nevertheless, this is checked explicitly using additional MC samples for three main background processes: W boson production (denoted "W($\to \ell\nu$) + jets"), production of a Z boson decaying to neutrinos ("Z($\to \nu\bar{\nu}$) + jets"), and top quark pair production ("$t\bar{t}$ + jets").

An overview of the explicit background samples used can be found in table A.2 in section A.1 of the appendix.

All the background MC samples contain events generated based on LO matrix element calculations performed with the MadGraph5_aMC@NLO software package [70]. This is interfaced to Pythia 8, which applies the parton shower algorithm, matching it to the matrix element calculation in the MLM scheme.

5.2 Event selection

Events contained in the recorded and simulated data sets are filtered by applying specific selection criteria. The purpose of the selection is to obtain a representative, high-quality sample of QCD dijet events for performing the cross section measurement.

The first selection of recorded data happens during data taking, where events containing jets with transverse momenta above predefined thresholds are identified by the trigger system and stored for offline analysis. These events are then reprocessed following data taking, applying the latest calibrations and corrections.

While a simulation of the trigger is also available for simulated data sets, this is not taken into account for the event selection, such that all simulated events are treated under the assumption of a fully efficient, unprescaled trigger.

The reprocessed event sample then undergoes an offline selection, where analysis-specific criteria are applied. The analysis phase space is defined by applying cuts to the kinematics of the two leading jets, taking into account the detector acceptance and the reconstruction efficiency.

The event sample is further refined by applying a series of quality cuts meant to reduce the impact of background, detector noise and other sources of bias. These selection steps are described in detail in the following.

5.2.1 Trigger selection

Events entering the analysis are first selected based on the trigger decision. As mentioned in section 4.3, a number of trigger paths that are specifically designed for jet measurements are defined in the HLT configuration. Since the cross section decreases exponentially across the experimentally accessible p_{T} range, multiple prescaled trigger paths with different thresholds are deployed to ensure that all of it is covered.

For a differential cross section measurement across the entire spectrum, the trigger paths must be combined, taking the reduction in event numbers due to prescaling into account. This is achieved by determining an *effective luminosity* for each trigger path, which corresponds to the integrated luminosity necessary for an equivalent unprescaled path to result in the same overall number of accepted events.

Since the prescaling factor is adjusted continually during data taking to optimize the trigger bandwidth, the rate of overall accepted events varies with time. To obtain the total effective luminosity, the integrated luminosity measured for each luminosity block is scaled down by the prescaling factors configured during data taking for the HLT path and the associated L1 seeds, and summed across all luminosity blocks contributing to the measurement.

In this analysis, only events triggered by one of 9 available dijet HLT paths are considered. This set of triggers constitutes a natural choice for a dijet analysis, since they are optimized for the exact jet topology that is being studied. Moreover, in a majority of cases, they provide higher effective luminosities than the single-jet trigger paths with corresponding thresholds. An explicit list of the trigger paths, together with their corresponding p_T thresholds and effective luminosities is given in table 5.2.

A further aspect relevant to the trigger selection is the trigger efficiency. As discussed in section 4.3, triggers exhibit a "turn-on" behavior, causing a sharp decrease in the number of events accepted by a trigger path at $\langle p_\mathrm{T} \rangle_{1,2}$ values below the trigger threshold.

In general terms, an estimate of the efficiency of a selection can be obtained by evaluating the proportion of *true positives* in a sample. To result in an unbiased estimate, the preselection criteria for putting together this sample have to be independent of the selection criteria.

In the specific case of dijet triggers, an event is considered a *true positive* if the average transverse momentum of the two leading jets is above the trigger threshold both when calculated from the final calibrated jets ("true") and the corresponding trigger objects ("positive").

The method used for estimating the trigger efficiency in this analysis is known as *trigger emulation* (also called *bootstrapping*), and is applicable to trigger paths that are part of a set with increasing thresholds, as described in section 4.3.

In this case, the efficiency of a trigger path T is determined from a sample of events triggered by the path with the next-lowest threshold within the same set. For these events, the decision of the trigger T is emulated, and the fraction of events accepted by the trigger emulation are taken as an estimate of the efficiency of T.

More rigorously, for a sequence of n trigger paths (T_k), $k \in \{1, 2, \dots, n\}$, belonging to the same set and arranged in ascending order of their HLT thresholds, let $\{T_k\}$ denote the set of events accepted by the k-th trigger path. Furthermore, let $\{T_k^{\text{emulated}}\}$ denote the set of events for which the emulation of the decision of the k-th trigger path has resulted in the event being accepted.

Then, using N to mean set cardinality, an estimate for the efficiency of T_k is given by:

$$\hat{\varepsilon}(T_k) = \frac{N(\{T_{k-1}\} \cap \{T_k^{\text{emulated}}\})}{N(\{T_{k-1}\})} \tag{5.2}$$

Table 5.2 – Dijet trigger paths deployed in 2016. Each entry indicates the name of the trigger path as it appears in the trigger menu, the effective luminosity for the entire 2016 data period, as well as the $\langle p_T \rangle_{1,2}$ (p_T) thresholds for trigger objects used for determining the HLT (L1) trigger decisions.

Trigger path	**Trigger thresholds** / GeV		**Eff. luminosity** / pb^{-1}
	HLT ($\langle p_T \rangle_{1,2}$)	L1 (p_T)	
`HLT_DiPFJetAve40`	40	–	0.1
`HLT_DiPFJetAve60`	60	–	1.7
`HLT_DiPFJetAve80`	80	60	4.2
`HLT_DiPFJetAve140`	140	90	27.6
`HLT_DiPFJetAve200`	200	120	138
`HLT_DiPFJetAve260`	260	170	523
`HLT_DiPFJetAve320`	320	170	2969
`HLT_DiPFJetAve400`	400	170	9026
`HLT_DiPFJetAve500`	500	170	29309

Compared to other strategies, the trigger emulation approach has the advantage of exploiting trigger-level information directly: Since the preselection of events is done on the basis of jet triggers, the kinematic properties of jets reconstructed at the trigger level are available and can be used to model the trigger behavior more accurately. This comes at the expense of requiring a fully efficient jet trigger path for event preselection, meaning that the efficiency of the lowest-threshold trigger path in a sequence cannot be estimated using this method. Furthermore, the trigger paths used for preselection must be fully efficient in the phase space region used for the efficiency measurement, which would otherwise result in a biased estimate. Thus, the thresholds of successive triggers must be sufficiently far apart to ensure that there is no overlap of the turn-on regions, which is observed to be the case for all trigger paths considered.

The emulation of a trigger path consists of two steps. First, the HLT and L1 trigger objects corresponding to the two leading jets are identified. For jets clustered with a radius parameter R, these are defined as the trigger objects closest to the respective jets in (η, ϕ)-space within a cone of $\Delta R < R/2$.

An HLT match is always required to exist for both leading jets. For HLT paths that are seeded by an L1 trigger pathway, an L1 match is required as well. If a required match is not found for either one of the two leading jets, the event is considered to have been rejected by the trigger path emulation.

If all required trigger object matches are found, the decision of the trigger path is emulated by comparing the average momentum of the matched HLT (and, if applicable, L1) trigger objects to the configured threshold(s). Events yielding values above the threshold(s) are accepted, the others are rejected.

Figure 5.2 shows the trigger efficiency curves obtained using this method as a function of the average jet transverse momentum. The efficiencies shown here are derived based on an inclusive sample where no offline event selection has been applied. In the final analysis, the trigger efficiency is studied individually in each measurement bin, as discussed in section 5.3.2.

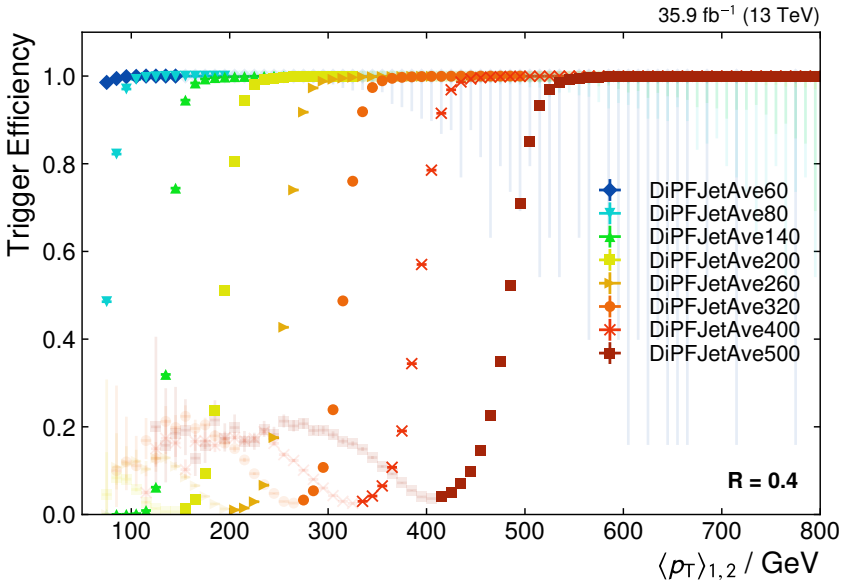

Figure 5.2 – Efficiencies of the dijet trigger paths obtained using the emulation method. The curves display the expected turn-on behavior as a function of the average jet transverse momentum around the nominal trigger threshold. The emulation method provides reliable efficiency estimates only in regions where the preceding trigger path is fully efficient. Outside these regions, the efficiency estimates are masked for better readability. The error bars correspond to Clopper-Pearson confidence intervals.

5.2.2 Jet phase space selection

After the trigger selection, a further set of criteria is applied based on the kinematic properties of the two p_T-leading jets. Events are required to contain at least two jets, with the p_T of the leading and subleading jets being larger than 100 GeV and 50 GeV, respectively. In addition, the absolute rapidity of both jets is required to be less than 2.5.

These criteria serve the primary purpose of restricting the measurement to a kinematic region where the reconstruction efficiency is high. In addition, they provide a well-defined phase space for the calculation of theory predictions, ensuring that a direct comparison to the measurement can be made.

5.2.3 Quality criteria

A further set of selection criteria is imposed in order to reduce the impact of detector signals that do not originate from collision products. These can be caused by a variety of sources, the main ones being calorimeter noise and the so-called *beam halo*, which consists of particles produced by the interaction of the proton beams with the accelerator material.

Since these signals disturb the transverse energy balance of events, a reliable method for identifying these cases consists in examining the missing transverse momentum (MET). Events with a particularly high MET are screened additionally for effects consistent with noise or beam halo particles, for which dedicated algorithms have been developed within CMS, and excluded from the measurement.

In addition, high-MET events are less likely to originate from pure QCD interactions, since these do not give rise to the undetectable particles associated with genuine MET (e.g. neutrinos). A variable commonly used to quantify this is the ratio of the missing transverse energy to the total scalar sum of transverse energies of all particles.

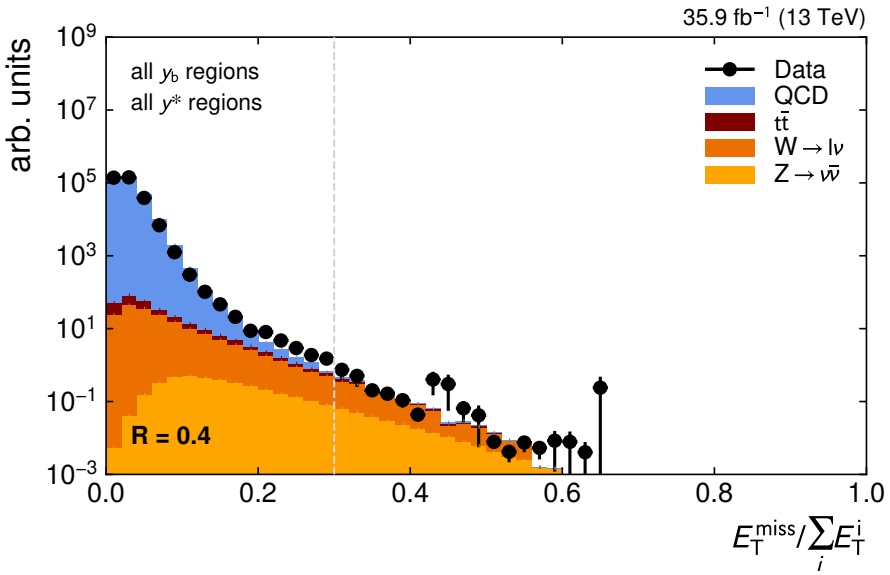

Figure 5.3 – Distribution of the missing transverse energy, as a ratio to the scalar sum of particle transverse energies. QCD events occur predominantly at low values, while higher regions are affected by detector noise and background. Events above a value of 0.3, indicated by the dashed line, are rejected from the final analysis.

Figure 5.3 shows the contribution from QCD and select non-QCD background processes, as given by simulation. As can be seen, the vast majority of QCD events exhibit values of this variable closer to 0, while events at larger values are almost exclusively due to background. To ensure a high sample purity of QCD events and further limit the impact

of noise, a cut is imposed on this value, rejecting events above a threshold value of 0.3. This value is chosen in order to remain consistent with previous QCD analyses [4, 5].

A final quality criterion concerns the effect of noise on the jet reconstruction. The presence of detector noise can cause spurious energy clusters in the calorimeters with a sufficiently high energy to be misidentified as jets or to make up a significant portion of the energy of genuine jets.

Since these jets form a background to QCD jet production, they must be accounted for in the final measurement. For this purpose, a set of *jet identification criteria* are recommended within CMS. These are based on the multiplicity of the particle flow jet constituents and the energy carried by different constituent types, as a fraction of the total jet energy. Jets which do not pass these criteria are marked as invalid and removed from the event.

Due to the different subdetector configurations, different criteria are applied depending on the pseudorapidity of the jet. For instance, the contribution of charged constituents is only considered within tracker coverage ($|\eta| < 2.4$), extending to $|\eta| < 2.7$ for muons. A summary of the criteria can be found in table 5.3. These have been optimized to obtain a high noise jet rejection rate, while keeping genuine jets unaffected. It is found that the associated selection efficiency remains above 99% in all phase space regions considered in this analysis.

Table 5.3 – Summary of the jet identification criteria. Different criteria are used depending on the pseudorapidity of the jet.

	Pseudorapidity regions			
	$\lvert\eta\rvert < 2.4$	$2.4 \leq \lvert\eta\rvert < 2.7$	$2.7 \leq \lvert\eta\rvert < 3.0$	$3.0 \leq \lvert\eta\rvert$
Energy fractions:				
Neutral electromagnetic	< 0.9	< 0.9	> 0.01	< 0.9
Neutral hadrons	< 0.9	< 0.9	< 0.98	–
Muons	< 0.8	< 0.8	–	–
Charged hadrons	> 0	–	–	–
Charged electromagnetic	< 0.9	–	–	–
Number of constituents:				
Total	> 1	> 1	–	–
Neutral	–	–	> 2	> 10
Charged	> 0	–	–	–

5.3 Phase space and binning

The differential dijet cross section is measured as a function of three observable quantities: y^*, y_b, and a third kinematic variable, chosen to be either the average transverse momentum of the two leading jets $\langle p_\mathrm{T} \rangle_{1,2}$ or the invariant mass of the dijet system m_jj.

The quantity y^* ("y-star") is defined as half the absolute value of the difference between the rapidity of the leading jet $y_{\mathrm{jet}\,1}$ and the subleading jet $y_{\mathrm{jet}\,2}$. Analogously, y_b ("y-boost") is defined as half the absolute value of the rapidity sum:

$$y^* = \frac{1}{2}\left|y_{\mathrm{jet}\,1} - y_{\mathrm{jet}\,2}\right|, \qquad y_\mathrm{b} = \frac{1}{2}\left|y_{\mathrm{jet}\,1} + y_{\mathrm{jet}\,2}\right| \tag{5.3}$$

The quantities $\langle p_\mathrm{T} \rangle_{1,2}$ and m_jj are defined via the transverse momenta of the two leading jets $p_\mathrm{T}^{\mathrm{jet}\,1}$ and $p_\mathrm{T}^{\mathrm{jet}\,2}$, and the four-vectors $p_\mu^{\mathrm{jet}\,1}$ and $p_\mu^{\mathrm{jet}\,2}$, respectively:

$$\langle p_\mathrm{T} \rangle_{1,2} = \frac{1}{2}\left(p_\mathrm{T}^{\mathrm{jet}\,1} + p_\mathrm{T}^{\mathrm{jet}\,2}\right) \tag{5.4}$$

$$m_\mathrm{jj} = \sqrt{\left(p_\mu^{\mathrm{jet}\,1} + p_\mu^{\mathrm{jet}\,2}\right)\left(p^{\mu,\mathrm{jet}\,1} + p^{\mu,\mathrm{jet}\,2}\right)} \tag{5.5}$$

5.3.1 Binning

The cross section measurement is performed in bins of y^*, y_b, and either $\langle p_\mathrm{T} \rangle_{1,2}$ or m_jj. Thus, an adequate binning scheme that is suited to the present analysis must be established. The choice of variables exhibits a number of particularities that are addressed in the following.

First, for a simultaneous binning in y^* and y_b, it must be considered that the overall phase space results from a linear combination of the underlying rapidity phase spaces of the two leading jets. This leads to a characteristic "triangular" shape, which can only partially be covered by conventional rectangular bins.

In addition, an adequate range and bin size must be chosen for $\langle p_\mathrm{T} \rangle_{1,2}$ and m_jj, which is largely governed by the accessible phase space and the detector resolution. An additional particularity of the m_jj measurement is that the accessible phase space shifts toward higher values in the high-y^* region as a consequence of the invariant mass definition.

Binning in (y^*, y_b)

From the definitions in (5.3), it follows that y^* and y_b can take values between 0 and an upper bound of y_max, which is the maximum absolute rapidity of any of the two leading jets, as required by the selection criteria.

However, these two quantities are not independent, since they are formed from linear combinations of $y_{\text{jet}\,1}$ and $y_{\text{jet}\,2}$, both of which are limited by the selection criteria. As a consequence, not every combination of (y^*, y_b) values within those bounds is possible, and it can be shown that the following restriction holds: $y^* + y_\text{b} < y_\text{max}$. This results in a characteristic triangular shape of the combined phase space in the (y^*, y_b) plane.

In this analysis, an identical equidistant binning is chosen for y^* and y_b, starting from 0 and with a bin size of 0.5 in each variable. The upper bound of the binning in each y^* and y_b region is determined in such a way as to ensure that all resulting bins lie fully inside the available (y^*, y_b) phase space, as determined by y_max.

Going up to a value of $y_\text{max} = 3.0$, this results in a total of 15 (y^*, y_b) bins. In the region with $0 \leq y_\text{b} < 0.5$, there are five bins starting with $(0.0 \leq y^* < 0.5)$ and going up to $(2.0 \leq y^* < 2.5)$. For each of the subsequent y_b regions, there is one less y^* bin than in the previous one, ending with a single bin $(0.0 \leq y^* < 0.5)$ for $2.0 \leq y_\text{b} < 2.5$. The resulting configuration is shown in figure 5.4.

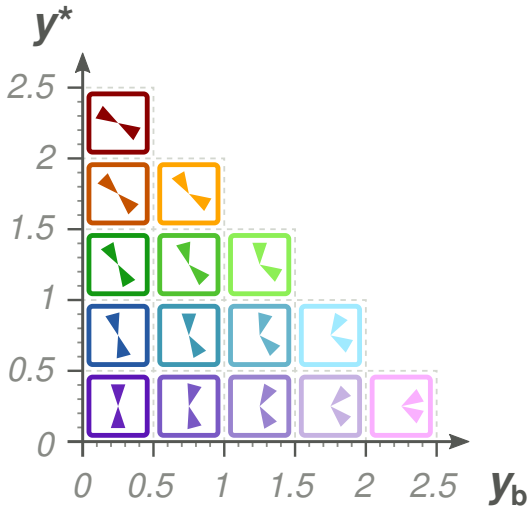

Figure 5.4 – Representation of the binning in (y^*, y_b), showing the characteristic triangular shape of the phase space. The cone pairs represented schematically in each bin are suggestive of the typical orientation of the leading dijet pair, with the horizontal direction corresponding to the beam pipe.

Binning in $\langle p_\text{T} \rangle_{1,2}$

Before deciding on a concrete binning scheme in $\langle p_\text{T} \rangle_{1,2}$, a number of criteria must first be established as to what constitutes a suitable binning for a precision measurement of the dijet cross section. In this analysis, three main criteria are considered, as outlined in the following.

event yield First, the number of events per bin is required to be of the order of at least 100 events. At the higher end of the spectrum, where the number of events naturally tapers off due to the exponential decrease of the differential cross section, this criterion establishes an upper bound on $\langle p_T \rangle_{1,2}$. At lower values, events are rejected by the trigger threshold requirements and the offline selection criteria, thus leading to an effective lower bound on $\langle p_T \rangle_{1,2}$.

trigger efficiency As discussed in section 4.3, the trigger efficiency can cause the loss of events in the "turn-on" region near the trigger threshold. To minimize the impact that a trigger inefficiency would have on the cross section measurement, only phase space regions where at least one trigger path has been determined to be fully efficient are considered. This requirement affects the lower part of the spectrum, leading to an exclusion of the "turn-on" region of the lowest-threshold trigger.

resolution Another important factor is the finite $\langle p_T \rangle_{1,2}$ resolution. This causes a spread of the reconstructed $\langle p_T \rangle_{1,2}$ with respect to the true value, leading to a systematic migration of events between bins. While this effect is accounted for using an unfolding procedure (see section 5.5), the stability of this procedure depends on the size of the migration effect. To ensure that the proportion of events that undergo bin-to-bin migration is acceptably low, the chosen bin width must be larger than or equal to the $\langle p_T \rangle_{1,2}$ resolution.

A further thing to consider is that both the trigger efficiency and the $\langle p_T \rangle_{1,2}$ resolution can exhibit differences depending on the (y^*, y_b) region, potentially resulting in a different lower bound or optimal bin width. Despite this, it is preferable to maintain a single consistent $\langle p_T \rangle_{1,2}$ binning, derived in such a way as to satisfy the criteria outlined above for all (y^*, y_b) regions simultaneously. This is achieved as follows.

In a first stage, a preliminary binning model is determined using simulated samples based on the $\langle p_T \rangle_{1,2}$ resolution, without distinguishing between the different (y^*, y_b) regions. Starting from a value of 100 GeV (which corresponds to the lower bound of the leading jet p_T, as required by the selection criteria), events are binned in very fine $\langle p_T \rangle_{1,2}$ increments of 1 GeV, resulting in a collection of *narrow bins*. The $\langle p_T \rangle_{1,2}$ resolution is then determined by taking the root-mean-square (RMS) of the absolute response distribution in each bin.

Next, consecutive narrow bins are merged one after another to form a *wide bin*, recalculating the resolution each time. The merging continues until the value of the resolution in the resulting wide bin is found to be less than or equal to half the bin width. Once this condition is satisfied, the resulting wide bin is added to the model, and process is repeated with the remaining narrow bins. This is done until the number of events in the remaining bins drops below the required minimum of 100.

The binning obtained via the above procedure is then validated separately in each (y^*, y_b) region. The absolute $\langle p_T \rangle_{1,2}$ resolution is verified to ensure that it remains comparable to the bin width. Furthermore, any bins at the upper end of the spectrum with event

counts below the required minimum of 100 are discarded. Finally, the trigger efficiency is evaluated in data as described in section 5.2.1, discarding all $\langle p_T \rangle_{1,2}$ bins at the lower end of the spectrum where no trigger path with an efficiency above 99.5% is found. The resulting $\langle p_T \rangle_{1,2}$ binning is given in table 5.4.

Binning in m_{jj}

For establishing the binning used for the measurement as a function of m_{jj}, an analogous procedure as for $\langle p_T \rangle_{1,2}$ is employed. However, an important difference to note here is that the typical range of m_{jj} at high values of y^* is shifted towards higher values with respect to the low-y^* region. This is because the invariant mass of the dijet system increases as a function of the dijet rapidity separation. The same does not hold for y_b, since m_{jj} is invariant under Lorentz boosts of the dijet system. The m_{jj} binning is given in table 5.5.

5.3.2 Active trigger ranges

As discussed in section 4.3, dijet triggers exhibit a "turn-on" behavior, causing a sharp decrease in the number of events accepted by a trigger path at $\langle p_T \rangle_{1,2}$ values below the trigger threshold. Furthermore, the reduction in effective luminosity caused by trigger prescaling leads to a progressive drop in event numbers as the value of $\langle p_T \rangle_{1,2}$ increases. As a result, each individual trigger path effectively only covers a part of the total $\langle p_T \rangle_{1,2}$ spectrum. The is also true for m_{jj}, which is highly correlated to $\langle p_T \rangle_{1,2}$.

In order to ensure that full trigger efficiency and an adequate event yield are maintained throughout the entire spectrum, the $\langle p_T \rangle_{1,2}$ (or m_{jj}) phase space is divided into several contiguous, nonoverlapping regions, each of which is assigned a single trigger path, referred to as the *active* trigger. This is done independently in each (y^*, y_b) region, as follows.

First, the efficiency of every trigger path is determined in each measurement bin using the emulation method described in section 5.2.1. Starting with the trigger path with the highest $\langle p_T \rangle_{1,2}$ threshold as the active trigger, the bins are then traversed in order from highest to lowest, evaluating the efficiency of the current active trigger for each bin.

As long as the efficiency of the current trigger path remains above a threshold value of 99.5%, it is assigned to the current bin. Otherwise, the trigger is switched to the next-lowest path, for which the efficiency is again above threshold. This continues until the lowest bin is reached. To avoid a bias of the efficiency measurement due to low event counts, efficiency measurements derived from a reference of fewer than 20 events are not considered during this process.

Once the active trigger regions have been established, the full $\langle p_T \rangle_{1,2}$ (or m_{jj}) spectrum is obtained by normalizing the number of events accepted by the active trigger path in

Table 5.4 – Binning scheme for $\langle p_T \rangle_{1,2}$. A general binning model is first established by evaluating the accessible phase space. The lower bound of the model results from trigger efficiency considerations, the upper bound is chosen based on the observed event counts, while the bin width is optimized based on the resolution. The actual $\langle p_T \rangle_{1,2}$ binning used in each (y^*, y_b) bin represents a contiguous subset of bins from the general model, with the bounds adjusted to the specific conditions in that bin.

Model for $\langle p_T \rangle_{1,2}$ bin edges (GeV)
122, 147, 175, 207, 243, 284, 329, 380, 437, 499, 569, 646, 732, 827, 931, 1046, 1171, 1307, 1458, 1621, 1806, 2003, 2217, 2453, 2702

y^* range	**y_b range**	**$\langle p_T \rangle_{1,2}$ bin edges (GeV)**	**Number of bins**
[0.0, 0.5)	[0.0, 0.5)	122, …, 2702	24
[0.0, 0.5)	[0.5, 1.0)	122, …, 2217	22
[0.0, 0.5)	[1.0, 1.5)	122, …, 1621	19
[0.0, 0.5)	[1.5, 2.0)	122, …, 1046	15
[0.0, 0.5)	[2.0, 2.5)	122, …, 732	12
[0.5, 1.0)	[0.0, 0.5)	122, …, 2453	23
[0.5, 1.0)	[0.5, 1.0)	122, …, 1806	20
[0.5, 1.0)	[1.0, 1.5)	122, …, 1307	17
[0.5, 1.0)	[1.5, 2.0)	122, …, 931	14
[1.0, 1.5)	[0.0, 0.5)	122, …, 1806	20
[1.0, 1.5)	[0.5, 1.0)	122, …, 1458	18
[1.0, 1.5)	[1.0, 1.5)	122, …, 1046	15
[1.5, 2.0)	[0.0, 0.5)	122, …, 1307	17
[1.5, 2.0)	[0.5, 1.0)	122, …, 1046	15
[2.0, 2.5)	[0.0, 0.5)	122, …, 827	13
(all)	*(all)*	–	264

Table 5.5 – Binning scheme for m_{jj}. A general binning model is first established by evaluating the accessible phase space. The lower bound of the model results from trigger efficiency considerations, the upper bound is chosen based on the observed event counts, while the bin width is optimized based on the resolution. The actual m_{jj} binning used in each (y^*, y_b) bin represents a contiguous subset of bins from the general model, with the bounds adjusted to the specific conditions in that bin.

Model for m_{jj} bin edges (GeV)
249, 306, 372, 449, 539, 641, 756, 887, 1029, 1187, 1361,1556, 1769, 2008, 2273, 2572, 2915, 3306, 3754, 4244, 4805, 5374

y^* range	**y_b range**	**m_{jj} bin edges (GeV)**	**Number of bins**
[0.0, 0.5)	[0.0, 0.5)	249, …, 5374	21
[0.0, 0.5)	[0.5, 1.0)	249, …, 4244	19
[0.0, 0.5)	[1.0, 1.5)	249, …, 2915	16
[0.0, 0.5)	[1.5, 2.0)	249, …, 2008	13
[0.0, 0.5)	[2.0, 2.5)	249, …, 1187	9
[0.5, 1.0)	[0.0, 0.5)	372, …, 5374	19
[0.5, 1.0)	[0.5, 1.0)	372, …, 4244	17
[0.5, 1.0)	[1.0, 1.5)	372, …, 2915	14
[0.5, 1.0)	[1.5, 2.0)	372, …, 2008	11
[1.0, 1.5)	[0.0, 0.5)	539, …, 5374	17
[1.0, 1.5)	[0.5, 1.0)	539, …, 4244	15
[1.0, 1.5)	[1.0, 1.5)	539, …, 2915	12
[1.5, 2.0)	[0.0, 0.5)	756, …, 5374	15
[1.5, 2.0)	[0.5, 1.0)	756, …, 4244	13
[2.0, 2.5)	[0.0, 0.5)	1187, …, 5374	12
(all)	*(all)*	–	223

each bin to the corresponding effective luminosity. Events that lie outside the region where the trigger path(s) that accepted them are active are rejected. This is achieved by applying a trigger weight to each event as follows:

$$w_{\text{trigger}}(x) = \begin{cases} 1/L_{\text{eff.}}\left(T_{\text{a}}(x)\right) & \text{event accepted by trigger path } T_{\text{a}}(x) \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

Here, the shorthand notation x is used to mean the values of the main event variables (y^*, y_{b}, and $\langle p_{\text{T}} \rangle_{1,2}$ or m_{jj}), $T_{\text{a}}(x)$ refers to the active trigger assigned to the bin that contains x, and $L_{\text{eff.}}\left(T_{\text{a}}(x)\right)$ denotes the effective integrated luminosity of the active trigger (see table 5.2).

5.3.3 Prefiring efficiency

Section 4.3.1 discusses the inefficiency induced by L1 trigger prefiring, which causes events to be falsely rejected already at the trigger level. The impact of this effect cannot be estimated using standard techniques such as the trigger emulation method considered for active trigger assignment outlined in previous section, and must be handled using a dedicated method.

Since the L1 trigger prefiring is caused by mistimed ECAL trigger primitives in the forward region of the detector with $2 < |\eta| < 3$, it is closely associated to jets reconstructed in this region. By counting the L1 jet trigger rates in this region using a special subset of *unprefireable* events, an estimate of the per-jet prefiring probability can be obtained (cf. figure 4.6). These are obtained from a comparatively small event sample and are therefore subject to a statistical uncertainty, which is provided alongside them. In addition, an inherent systematic uncertainty equal to 20% of the prefiring rate is assigned.

The total efficiency for an event is then calculated as the probability of a jet *not* to cause prefiring, which is in turn given by the product of the nonprefiring probabilities attributed to each jet in the event:

$$\begin{aligned} \varepsilon_{\text{prefiring}} &= 1 - P(\text{prefiring}) \\ &= \prod_{j \in \text{jets}} \left(1 - \varepsilon^{\text{jet}}_{\text{prefiring}}(\eta_j, p_{\text{T},j})\right) \end{aligned} \tag{5.7}$$

The resulting efficiency estimates derived for the present cross section measurement are shown in figure 5.5. As can be seen, the largest impact occurs in the (y^*, y_{b}) regions at the edge of the analysis phase space, where it can drop to values of 75%. These regions are particularly susceptible to prefiring inefficiency since they are more likely to contain a jet in the affected pseudorapidity region ($2 < |\eta| < 3$). Conversely, all other rapidity regions are significantly less affected, exhibiting efficiencies of 95% and above.

Also shown in figure 5.5 is the uncertainty of the prefiring efficiency. It is obtained by systematically shifting the per-jet prefiring rates in equation (5.7) upward and downward by a factor corresponding to their total uncertainty, which is estimated as the sum in quadrature of the statistical and systematic uncertainty (cf. section 4.3.1).

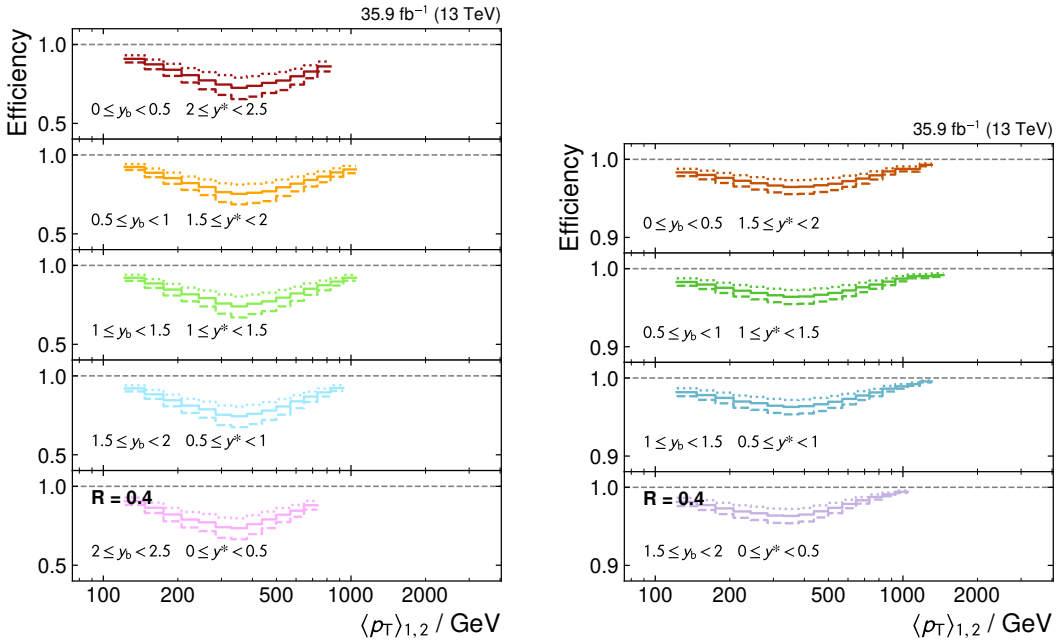

Figure 5.5 – Efficiency due to trigger prefiring. Shown on the left are the five (y^*, y_b) regions at the outer edge of the rapidity phase space, where large contributions from forward jets cause the efficiency to decrease to levels close to 75%. In contrast, in the immediately neighboring four rapidity regions shown on the right, the efficiency remains above 95% at all times, while in all other regions (not shown), the prefiring efficiency exceeds 99%. The dashed and dotted lines indicate the total uncertainty of the efficiency estimation.

5.3.4 Cross section definition

For the cross section measurement, events are binned in the three-dimensional phase space spanned by the observables (y^*, y_b, X), where X refers to either $\langle p_T \rangle_{1,2}$ or m_{jj}. The triple-differential cross section is then given by:

$$\frac{d^3\sigma}{dy^* dy_b dX} = \frac{1}{\varepsilon L_{\text{eff.}}} \frac{N_{\text{events}}}{\Delta y^* \Delta y_b \Delta X} \tag{5.8}$$

In the above, N_{events} denotes the number of events reconstructed in each bin, and Δy^*,

Δy_b, ΔX refer to the bin dimensions. The quantity $L_{\text{eff.}}$ represents the integrated effective luminosity, which is accounted for in recorded data by the trigger-specific event weight defined in (5.6), and ε stands for the total event selection efficiency. The latter comprises the trigger efficiency, which remains above 99.5% in all bins by construction (cf. section 5.3.2), and the total prefiring efficiency discussed in the previous section.

5.4 Comparison of recorded data to simulation

Predictions of observable quantities obtained from MC simulations are a useful tool for modeling systematic effects on the measurement. In particular, the unfolding approach described in a later section crucially relies on the accurate modeling of detector-related effects in simulated samples. It is therefore necessary to ensure that the distributions of the variables of interest and other closely related observables predicted by the simulations are compatible to those observed in data.

Due to the steeply falling cross section, events in MC simulations are generated using different weights in different phase space regions, so that the entire phase space is covered more efficiently. Similarly, recorded events are accepted by trigger paths with different prescaling factors depending on the phase space region. To recover the differential cross section across the entire spectrum, simulated events are assigned a weight corresponding to the inverse of the equivalent luminosity of the simulated sample (cf. equation (5.1)), and events in recorded data are weighted with the inverse of the effective luminosity of the trigger path by which they were accepted (cf. equation (5.6)).

Furthermore, in order to compare the recorded and simulated distributions on an equal footing, the inefficiency incurred due to L1 trigger prefiring must be corrected in data, since it is not accounted for in the simulated samples. This is done by applying an additional event weight corresponding to the inverse of the prefiring efficiency (5.7). In addition, the nominal distribution of pileup interactions assumed in the simulation must be adjusted to the actual data-taking conditions. This is done using a reweighting procedure, as described in the following.

5.4.1 Pileup reweighting

In order to perform a reweighting of the pileup profile, the amount of pileup per event for both data and simulation is first quantified in terms of observable quantities.

A suitable quantity for this purpose is the expected average number μ of additional proton collisions per bunch crossing. Given an estimated minimum bias cross section σ_{pp} of 69.2 mb [71] and the nominal bunch crossing frequency f of 40 MHz, an estimate of μ can be obtained from the average instantaneous luminosity $\langle L_{\text{inst}} \rangle$:

$$\mu = \sigma_{\text{pp}} \langle L_{\text{inst}} \rangle / f \tag{5.9}$$

In recorded data, the value of $\langle L_{\text{inst}} \rangle$ is determined from dedicated luminosity measurements for each luminosity block and varies with time according to data taking conditions. Events recorded during the luminosity block are assigned a corresponding value of μ.

For Monte Carlo samples, μ is determined for each event by sampling from a target distribution that roughly corresponds to the expected pileup profile in data. The actual number of pileup interactions overlaid on top of the hard scattering event is drawn from a Poisson distribution with a parameter μ (cf. 5.1).

To reweight the pileup profile in simulation, event weights are derived based on the distribution of μ in data and simulation, binned equidistantly between 0 and 80 expected pileup interactions. A simple reweighting scheme would then consist of taking the ratio of these distributions in each bin of μ, normalizing the distributions to unity beforehand in order to preserve the total sum of weights.

It is observed, however, that the overall distribution of μ is different depending on the trigger path. To take this difference into account when reweighting, a different set of weights is determined based on the pileup distribution observed for each trigger path. For a simulated event accepted by a trigger path T_k, the pileup weight is given by:

$$w_{\text{PU}}(\mu_i, T_k) = \frac{N_{\text{data}}(\mu_i, T_k)/N_{\text{data}}(T_k)}{N_{\text{MC}}(\mu_i, T_k)/N_{\text{MC}}(T_k)} \tag{5.10}$$

In the above, $N_S(T_k)$ refers the to the total number of events in a sample S that were accepted by a trigger path T_k, and $N_S(\mu_i, T_k)$ denotes the subset of those events in the i-th pileup bin μ_i. In the rare case that an event is accepted by multiple trigger paths, the weight is derived based on the path with the highest $\langle p_{\text{T}} \rangle_{1,2}$ threshold.

A further pileup-related observable is the *number of primary vertices* n_{PV} reconstructed in the event. However, this quantity is sensitive to the particularities of the track reconstruction, and therefore not directly suitable for pileup reweighting. Nevertheless, it is a useful quantity for performing a cross-check of the reweighting procedure.

The result can be seen in figure 5.6, indicating a greatly improved agreement between data and simulation after reweighting.

5.4.2 Reconstructed spectra

After reweighting the simulated samples based on the pileup distribution in data and applying the prefiring corrections to the data, the distributions of the main observable quantities are compared. In addition to the main variables y^*, y_{b}, $\langle p_{\text{T}} \rangle_{1,2}$ and m_{jj}, the distributions of the dijet separation in the azimuthal angle $\Delta\phi_{1,2}$ and the Euclidean dijet separation in (η, ϕ) space $\Delta R_{1,2}$ are also checked.

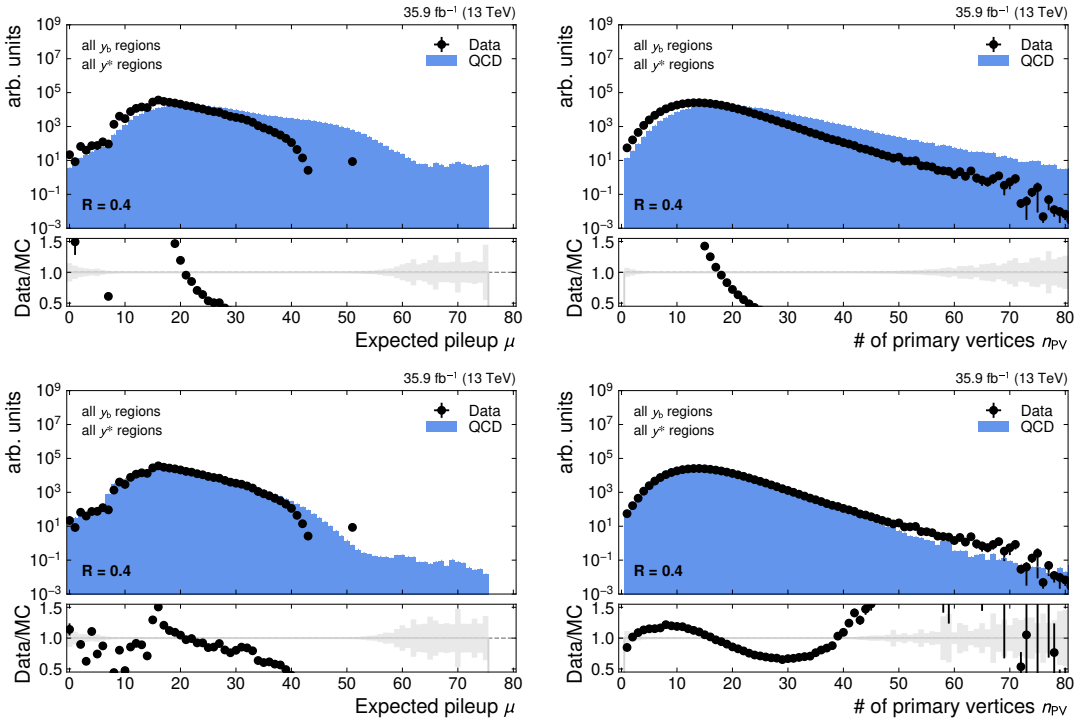

Figure 5.6 – Distribution of pileup-related quantities before (*top*) and after (*bottom*) pileup reweighting. Shown are the expected number of pileup interactions μ and the number of primary vertices in the event n_{PV}. The reweighting results in an improved agreement for both observables.

As can be seen in figure 5.7, the shapes of the distributions in data and simulated samples are largely compatible. Some deviations remain, which is expected on account of the simulations being based only on leading-order QCD calculations.

5.5 Unfolding

One of the main sources of bias incurred by measurements at particle colliders is the finite degree of accuracy and precision to which the properties of high-level analysis objects such as particles and jets can be reconstructed.

This is characterized in terms of the detector *response*, which is defined as the difference (or ratio) between the value of an observable, as measured with the help of the detector, and its reference or "true" value. Since the reconstruction process introduces a degree of randomness, the response is modeled as a random variable characterized by a particular mean value (also sometimes referred to as the *response*) and a measure of the spread of the distribution, which is called the *resolution*.

While a shift in the mean response value is typically compensated by the overall calibration, the effect of the resolution on a measurement remains and must be taken into account separately. The main effect here consists in the reconstruction of events at a different point in phase space compared to the "true" location obtained in the ideal case of infinite resolution.

In binned measurements, this manifests as a migration of events from the bin that corresponds to the "true" observable values to a different bin based on the actually measured observable values. *Unfolding* is a technique used to reverse these migrations and enable the reconstruction of the "true" distributions.

5.5.1 Simulated response

A prerequisite for unfolding is the characterization of the detector response, as it applies to a particular measurement. Since the reference quantities required for this are not available in data, the response estimation relies instead on simulated samples, where the final states coming from a Monte Carlo event generator are propagated through a full detector simulation. This enables the estimation of the detector response and the resolution of detector-level quantities by using generator-level information as a reference.

In the present analysis, the effect of the detector response on all three of the analysis variables y^*, y_b and $\langle p_T \rangle_{1,2}$ (or m_{jj}) must be determined, with the corresponding reference values given by their generator-level equivalents. The latter quantities are calculated from jets clustered from all stable final-state generator particles, with the exception of neutrinos.

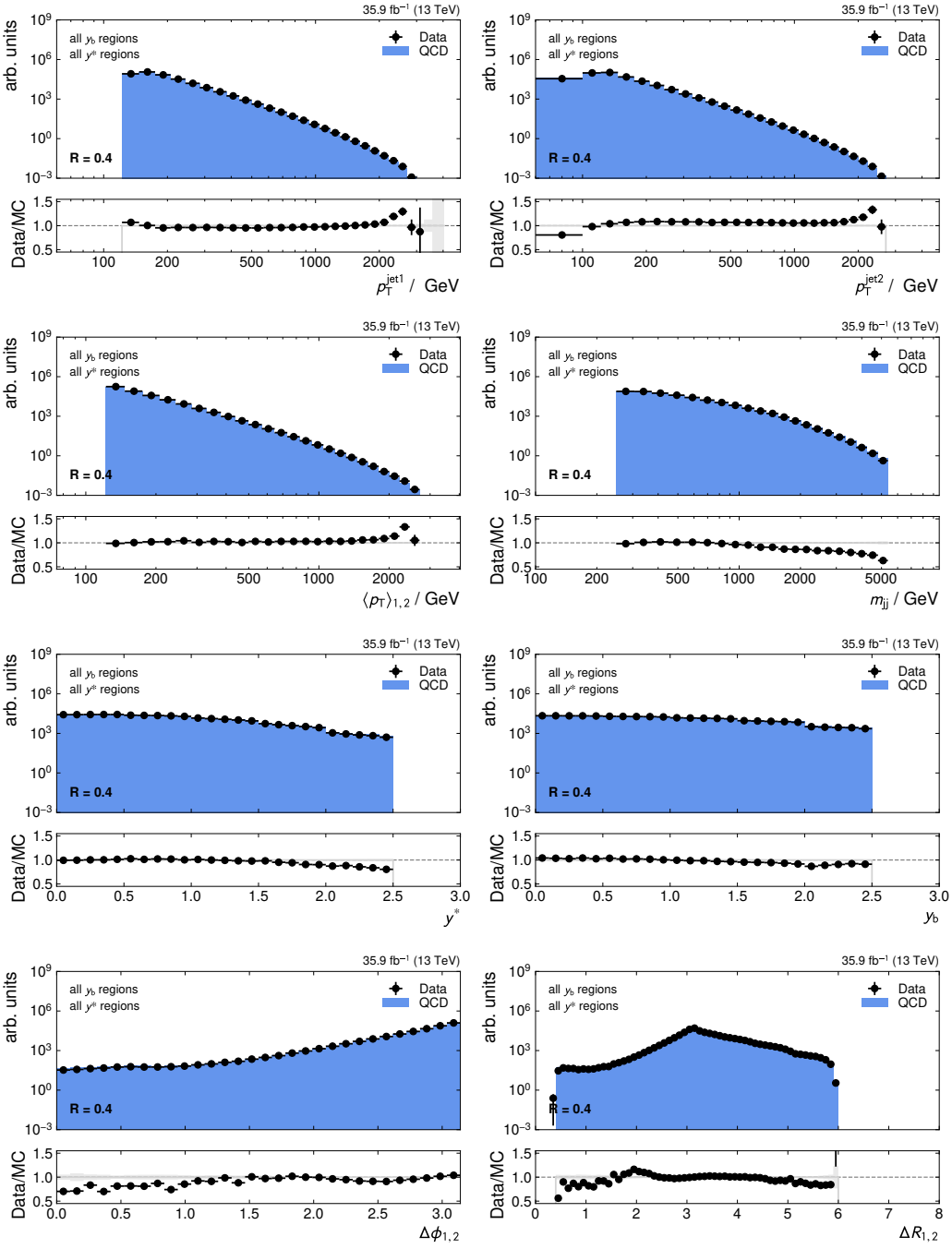

Figure 5.7 – Comparison of the variable distributions in data and MC simulation, shown for the transverse momenta of the two leading jets (*top row*), the average dijet transverse momentum and the dijet invariant mass (*second row*), the dijet rapidity separation y^* and total dijet boost y_b (*third row*), as well as the dijet angle separation in azimuthal plane ($\Delta\phi$) and the Euclidean dijet separation in the η–ϕ plane (*bottom row*). Similar plots for all rapidity regions can be found in section A.5 of the appendix.

It is important to note that the p_T ordering of jets may change between the generator and detector levels, potentially causing the analysis variables to be calculated based on unrelated objects. One way to restore the correspondence would be to perform a matching between the detector-level and the generator-level jets based on a criterion such as their separation in (η, ϕ) space. While such a matching would be useful in determining the resolution of individual jets, it would not result in an accurate description of the total detector response in the analysis variables, which are defined on an event-by-event basis strictly based on the two p_T-leading jets. Accordingly, no such matching is done when determining the detector response for unfolding.

Phase space unraveling

On account of the triple-differential phase space, which is spanned by three observables, bin migration effects would need to be modeled starting from three-dimensional distributions. However, this is both highly impractical and would make the effect considerably more difficult to visualize. Instead, a more accessible representation is achieved by rearranging the bins of into a one-dimensional bin sequence. This is called *phase space unraveling*.

The unraveling scheme used here results from first sorting the bins based on their characteristic y^*, y_b and $\langle p_T \rangle_{1,2}$ (or m_{jj}) values, in that order. The bins are then labeled based on the *global bin index* corresponding to their position in the sorted sequence. The unraveling scheme is illustrated in figure 5.8.

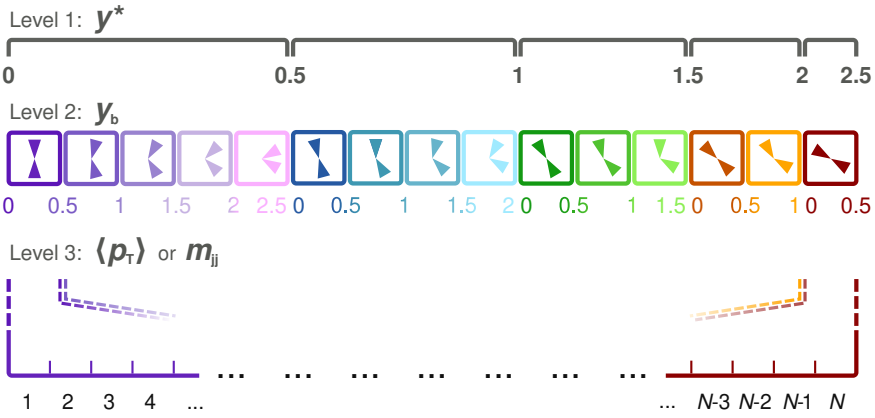

Figure 5.8 – Phase space unraveling scheme for unfolding. The analysis bins are sorted based on their characteristic y^*, y_b and $\langle p_T \rangle_{1,2}$ (or m_{jj}) values and arranged in a linear sequence. A bin is uniquely identified by its position in the resulting sorted sequence, which is referred to as its *global bin index*.

Marginal distributions

Using the unraveling scheme described above, it is possible to represent the binned distribution of events as a one-dimensional vector of dimension N, indicating the total number of bins. This is referred to as a *marginal distribution* and can be determined by binning events as a function of either detector-level or generator-level quantities.

In the following, the detector-level marginal distribution is denoted $\mathbf{r}$, with r_i representing the effective number of events in the bin with the global index i. Likewise, $\mathbf{g}$ stands for the generator-level marginal distribution, with g_j being the effective number of events in bin j.

It is important to note that, in addition to the binning, any other selection criteria applied to the event sample that involve jets must also be evaluated based on the corresponding observable level. Thus, the event selection used for obtaining the generator-level distribution must be applied on the generator-level jets, while for the detector-level distribution, it must be applied on the detector-level jets.

Migration matrix

The migration of events between bins in the unraveled phase space is described by a two-dimensional distribution. Formally, this corresponds to an $N \times N$ matrix $\mathbf{M}$ called the *migration matrix*, with the entry M_{ij} representing the effective number of events that have been generated in bin j and reconstructed in bin i on the detector level.

In order for an event to contribute to the migration matrix, it is necessary for it to pass the selection criteria on both the generator and the detector levels.

If the migration matrix is normalized so that the sum of all the entries in each column is equal to one, a new matrix $\mathbf{A}$ called the *response matrix* is obtained, where A_{ij} is an estimate of the probability for an event generated in bin j to be reconstructed in bin i (not yet taking into account events that migrate outside of the measurement phase space). This matrix is shown in figure 5.9.

The response matrix is an essential part of the unfolding process, since it expresses a linear map between the generator-level and the detector-level distributions. The essence of unfolding consists of finding an inverse of this map, which can then be applied to a given detector-level distribution to recover the underlying true distribution, as determined from generator-level quantities. However, in order to correctly account for any events migrating in and out of the measurement phase space, additional information is needed.

Fakes and acceptance losses

Events that contribute to the reconstructed detector-level spectrum but cannot be attributed to one of the generator-level bins or otherwise fail the selection criteria on the

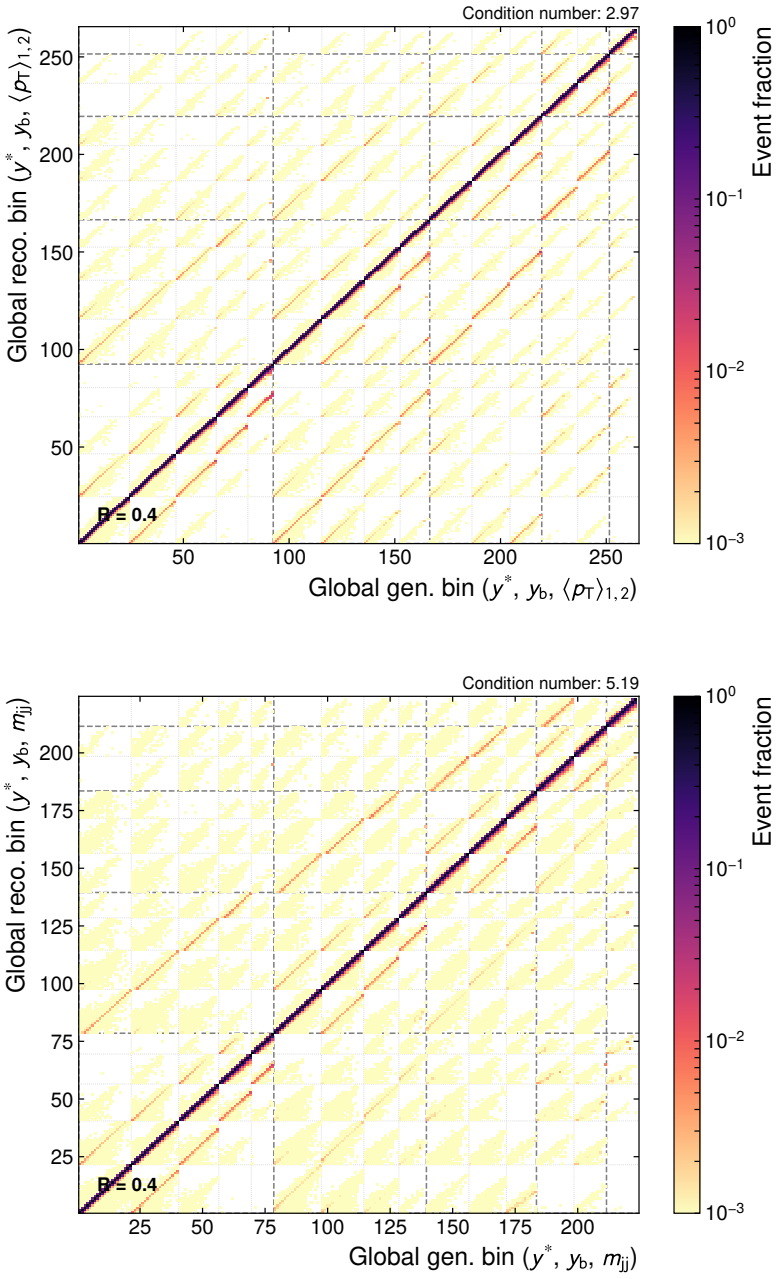

Figure 5.9 – Migration probability matrices. The matrix entries indicate the probability for an event to migrate from the respective generator-level bin on the x-axis to the detector-level bin indicated on the y-axis. The most significant portion of migrations is observed between immediately neighboring bins in $\langle p_T \rangle_{1,2}$ (m_{jj}) and, to a lesser extent, in y^* and y_b.

generator level are considered to be "fakes". These can be caused by either spurious reconstructions or by events migrating into the detector-level phase space from a region outside the measurement phase space on the generator level.

Conversely, events that originate from a generator-level bin for which no corresponding detector-level can be found are considered "acceptance losses". This can result from migration or events to regions outside the measurement phase space or by an event failing to pass the selection criteria based on the detector-level requirements.

Both "fakes" and "acceptance losses" need to be taken into account when performing the unfolding. The effective number of "fake" events in every detector-level bin, as well as the effective number of "lost" events in every generator-level bin can be estimated by comparing the marginal distributions with the distribution of "true" (**t**) and "accepted" (**a**) events, respectively. These are obtained by forming the row-wise and column-wise sums of the migration matrix:

$$a_j = \sum_i M_{ij}, \quad t_i = \sum_j M_{ij} \tag{5.11}$$

Figure 5.10 shows the fraction of "fake" events per detector-level bin, as well as the fraction of "accepted" events per generator-level bin. These fractions remain close to 0 and 1, respectively, with the notable exception of bins located at the overall phase space boundaries. This is expected, since these regions are susceptible to events migrating into and out of the measurement phase space, which is not taken into account by the migration matrix, which only covers event migrations that remain within the overall phase space.

5.5.2 Unfolding procedure

The unfolding procedure is applied on an input detector-level distribution **y** measured in data. Information about the detector response is passed to the unfolding procedure in the form of the migration matrix **M** and the corresponding marginal distributions on the generator **g** and detector **r** levels, as determined in simulation. The result of the unfolding procedure is the unfolded distribution, denoted **x**.

The input distribution **y** is obtained by first correcting the detector-level distribution measured in data for "fakes". The correction is given by the proportion of "true" events in each detector-level bin in simulation:

$$y_i = r_i^{\text{data}} \cdot \frac{t_i^{\text{MC}}}{r_i^{\text{MC}}} \tag{5.12}$$

The overall efficiency due to "acceptance losses" is taken into account by normalizing the columns of the response matrix **A** to the overall acceptance probability (as opposed

to unity). This is achieved by dividing each entry of the migration matrix by the total marginal generator-level distribution **g**, where the acceptance losses are still present, instead of the column-wise sum, where they are not:

$$A_{ij} = M_{ij}/g_j \tag{5.13}$$

The unfolding itself is performed using the software package TUnfold [72], which provides a way to obtain the unfolded distribution from a specified detector-level input distribution, while also propagating the statistical uncertainty of the input distribution to the output.

In the TUnfold approach, unfolding is treated as a minimization problem. The function to minimize is constructed analogously to the established χ^2 statistic used in least-squares optimization by comparing the input detector-level distribution **y** to the result of applying the linearized detector response **A** to the unfolded distribution **x**:

$$\chi^2 = (\mathbf{y} - \mathbf{A}\mathbf{x})^{\mathsf{T}} (\mathbf{V}_y)^{-1} (\mathbf{y} - \mathbf{A}\mathbf{x}) \tag{5.14}$$

In the above expression, $\mathbf{V}_y$ is the covariance matrix of the input distribution, which in this case is diagonal and represents the statistical uncertainty. The χ^2 is then minimized with respect to the entries x_j, yielding the final unfolded distribution **x**. The covariance matrix of **x** is obtained by linear error propagation of $\mathbf{V}_y$.

In this analysis, the number of bins on the detector level is identical to the number of generator-level bins, so that the response matrix **A** is a square matrix. In this case, the best estimate for the unfolded cross section distribution resulting from the minimization approach used by TUnfold, as well as its covariance matrix can be written directly as a function of the inverse matrix $\mathbf{A}^{-1}$:

$$\hat{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{y} \tag{5.15}$$

$$\mathbf{V}_x = \mathbf{A}^{-1}\mathbf{V}_y(\mathbf{A}^{-1})^{\mathsf{T}} \tag{5.16}$$

Procedure stability

As with all inverse problems, the stability of the unfolding procedure is highly dependent on the input parameters. Most importantly, the stability is limited by the magnitude of the spread in the analysis variables caused by the detector resolution. If this quantity is too large, the effect of event migrations can no longer be reliably reversed.

To quantify this effect and provide a rough estimate the expected impact on the unfolding stability, it is useful to study two statistics derived from the response matrix called the *purity* and *stability*. The purity P_i is defined as the fraction of events in a detector-level bin i that originates in the same bin on the generator level. Analogously, the stability S_j is the fraction of events in a generator-level bin that is reconstructed in the

same bin on the detector level. Using the definitions of "true" and "accepted" events in equation (5.11), the purity and stability can be expressed as:

$$P_i = M_{ii}/a_i, \quad S_j = M_{jj}/t_j, \tag{5.17}$$

The unfolding method displays the highest stability when both of these quantities are high, typically remaining above the 50% threshold. As can be seen in figure 5.10, this condition is satisfied in a majority of analysis bins. A notable exception to this occurs in the high-y^* regions for the m_{jj} measurement, where lower values are reached due to a worsening of the resolution. Nevertheless, the number of affected bins remains small, so that these do not have a significant impact the unfolding stability.

A further useful quantity in evaluating the stability of unfolding, is the *condition number* of the response matrix, defined as the ratio between its largest and smallest eigenvalues. As a rule of thumb, condition numbers less than 10 indicate that the unfolding is sufficiently well-conditioned. In this analysis, this condition is satisfied, with the largest condition number observed being approximately 7.

A further effect that can influence the unfolding stability are the statistical fluctuations present in the simulated sample used to estimate the response matrix and the corresponding marginal distributions. These are propagated to the unfolded cross sections, and thus represent a new source of uncertainty, in addition to the statistical uncertainty of the input data itself.

5.5.3 Validation

The consistency of the unfolding procedure is evaluated by unfolding the detector-level distribution measured in simulation and comparing the unfolding result with the generator-level distribution. These are expected to be identical, since the unfolding parameters are derived from the same simulated sample. This is explicitly confirmed, indicating that the unfolding procedure is consistent.

In addition to the above consistency check, the result of applying the unfolding procedure on data is evaluated. This is done by comparing the size of the *unfolding correction* in data and simulation, which is defined as the ratio between the unfolded and the detector-level cross section. The correction size is illustrated in figure 5.11 for the $\langle p_{\text{T}} \rangle_{1,2}$ measurement in four representative (y^*, y_{b}) bins. In simulation, it is observed to be generally less than unity and is subject to small fluctuations.

The general trend is consistent with the exponential decrease of the cross folded with a largely symmetric response. This causes more events to migrate into a bin from lower $\langle p_{\text{T}} \rangle_{1,2}$ (m_{jj}) values, causing a net increase in the number of events in the folded distribution. Unfolding reverses these migrations, thus yielding a lower cross section estimate than on the detector level.

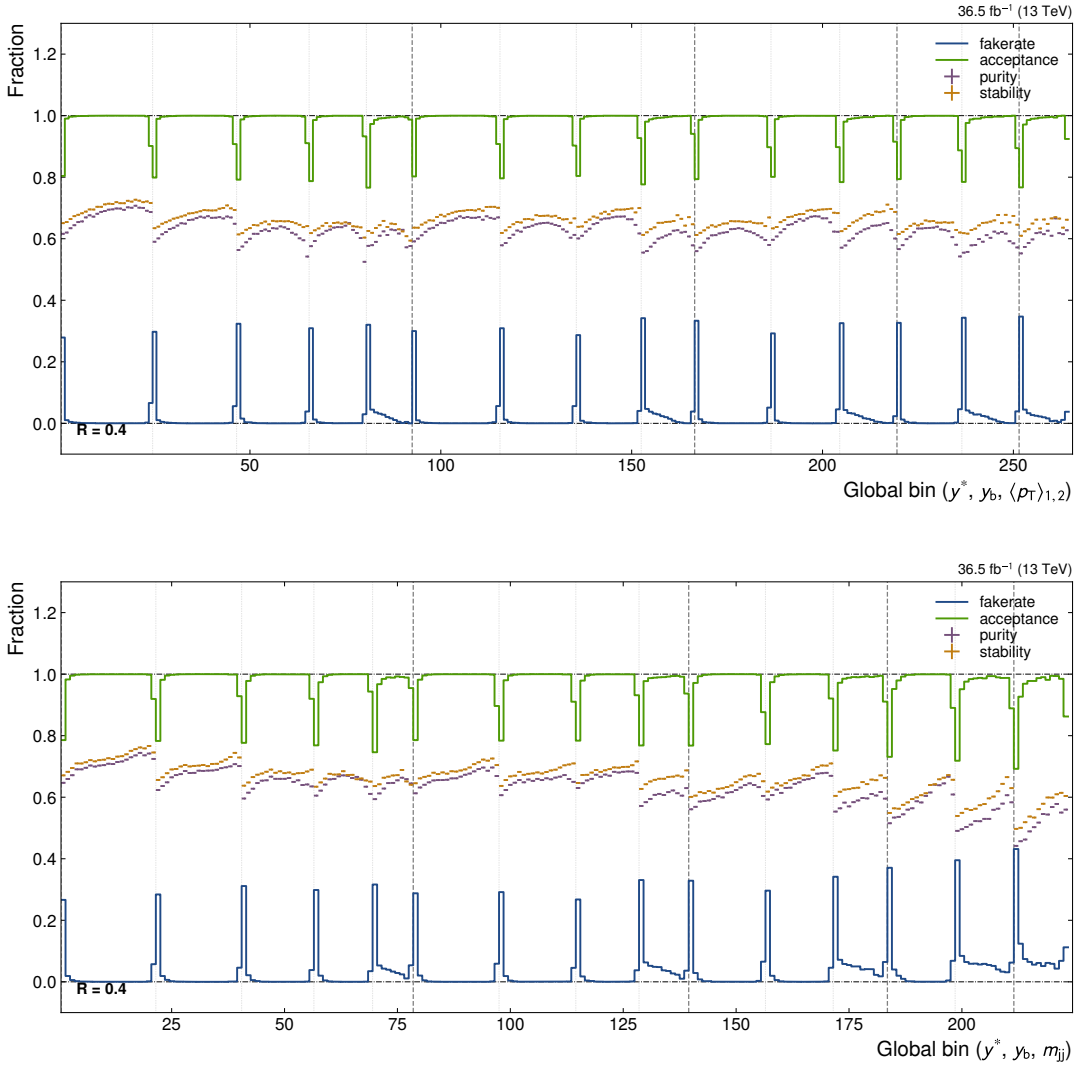

Figure 5.10 – Characteristics extracted from the response matrix. The x axis shows the bin index in the unraveled phase space. The purity and proportion of ”fake” reconstructions are shown as function of the detector-level bin index, while the stability and proportion of accepted events are shown as a function of the generator-level bin index. The thick and thin vertical dashed lines indicate the transition points between the different y^* and y_b regions, respectively. The fraction of ”fake” and ”lost” events increases in bins close to the overall phase space boundary.

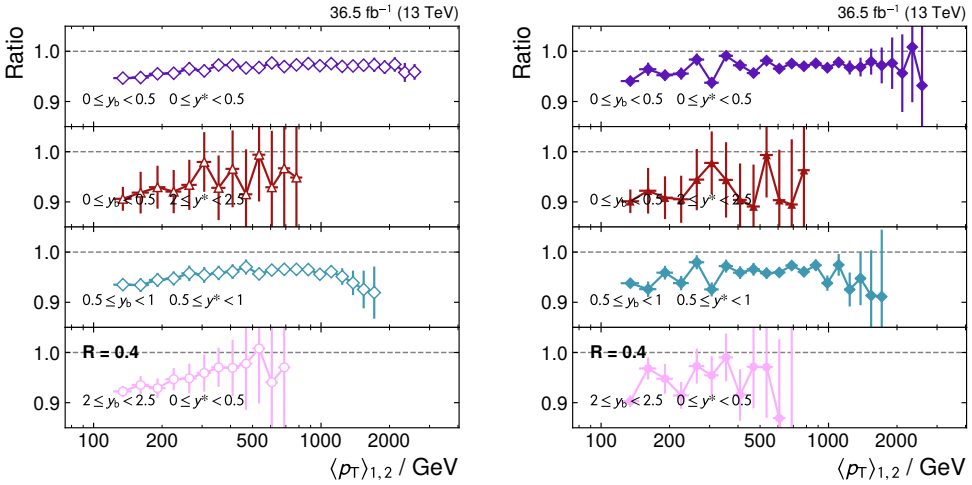

Figure 5.11 – Size of the unfolding correction in four representative (y^*, y_b) regions in QCD simulation (*left*) and data (*right*), with no smoothing applied. Fluctuations are observed in data which are not present to the same degree in simulation. This suggests the presence of a bias in the data reconstruction.

The fluctuations observed in the unfolding corrections in simulation are largely due to the limited sample size and appear compatible with the statistical uncertainty. In data, however, it is immediately apparent that the level of fluctuations is increased compared to the simulation. In particular, the fluctuations seen here appear too large to be covered by the statistical uncertainty.

Since the unfolding parameters are derived from the simulated samples and thus are identical in both cases, the origin of these enlarged fluctuations cannot be attributed to the unfolding procedure. Rather, they suggest the presence of an additional systematic effect that is specific to the recorded data, leading to a significant difference between the input detector-level distribution observed in data and that in simulation. This discrepancy propagates through the unfolding procedure, resulting in the observed fluctuations.

5.5.4 Smoothing

To gain a better understanding of the enlarged fluctuations observed in data, as described in the previous section, the input detector-level distributions are analyzed in terms of a smooth parametrization.

The dependence of the dijet cross section on $\langle p_T \rangle_{1,2}$ and m_{jj} is predicted by QCD theory to be smooth (i.e. exhibiting no localized fluctuations or discontinuities). This also holds for the detector-level spectra, provided that the detector response in these variables can

be accurately modeled by a continuous and monotonous (not necessarily linear) function.

While this is generally a reasonable assumption to make, given the complexity of the experimental setup, the online and offline reconstruction processes, and the analysis techniques employed, the measured distributions may exhibit deviations from the theorized smooth dependence. This can either be due to the statistical fluctuations caused by the limited number of events in a particular phase space region, or due to a systematic effect causing a local perturbation of the response. In any case, the deviation from a smooth spectrum is an experimental effect and must therefore be accounted for in the final uncertainty model.

Before the level of smoothness exhibited by a distribution can be assessed, it first needs to be approximated by a smooth function. Here, the distribution to approximate consists of a series of discrete differential cross section measurements y_i, each corresponding to a $\langle p_\mathrm{T} \rangle_{1,2}$ (or m_jj) value x_i and with an associated uncertainty σ_i. The goal is to obtain a smooth function $f(x)$, whose values at the different x_i values are a good approximation of the cross section estimates y_i.

The approach taken here is to use a linear combination of *Chebyshev polynomials*, which form a set of orthogonal functions over the interval $[-1, 1]$, and are given by the following recursive formula:

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x), \quad \text{with } T_0(x) = 1,\ T_1(x) = x \tag{5.18}$$

An arbitrary function defined over $[-1, 1]$ can then be approximated by a linear combination of the first few Chebyshev polynomials up to a maximum order of p. Denoting the vector of the corresponding coefficients as $\theta = (\theta_0, \ldots, \theta_p)$, this can be expressed as:

$$S_p(x, \theta) = \sum_{k=0}^{p} \theta_k \, T_k(x) \tag{5.19}$$

However, a polynomial approach is not suitable for modeling the cross section dependence on $\langle p_\mathrm{T} \rangle_{1,2}$ (or m_jj) directly, since this is observed to be exponential. This can be remedied by performing an approximation of the logarithm of the cross section instead. In addition, due to the large range in x and the uneven spacing that results from the binning scheme, a better approximation is obtained by taking the logarithm $(\ln x)$ of the independent variable. Finally, since Chebyshev polynomials are defined over $[-1, 1]$, the variable $(\ln x)$ must be mapped onto this range, which is readily accomplished via a linear transformation.

With the modifications outlined above, the full expression for approximating the cross section as a function of x using the Chebyshev polynomials up to an order of p is given

by:

$$f_p(x,\theta) = \exp\left(S_p(\tilde{x},\theta)\right), \quad \text{with } \tilde{x} = \frac{2\,\ln x/x_{\text{min}}}{\ln x_{\text{max}}/x_{\text{min}}} - 1 \tag{5.20}$$

In the above, x_{min} and x_{max} indicate the total range in $\langle p_{\text{T}}\rangle_{1,2}$ (or m_{jj}) covered by the cross section measurement.

The final step consists of finding adequate values for the Chebyshev coefficients θ. This is achieved via a least-squares fit of the approximation function in equation (5.20), with the χ^2 function given by:

$$\chi^2 = \sum_i \left(\frac{y_i - f_p(x_i,\theta)}{\sigma_i}\right)^2 \tag{5.21}$$

In (5.21), the expression in parentheses is called the *pull* of a data point. After the parametrization is obtained by minimizing the value of the χ^2 function with respect to the parameters θ, the pulls indicate the level of agreement between the individual data points and the parametrization, taking their uncertainty into account.

Assuming a Gaussian uncertainty model, pulls follow a standard normal distribution, having an expectation value of 0 and a standard deviation of 1. Pulls with high absolute values beyond a few standard deviations therefore indicate a localized tension between the data point and the fitted parametrization.

The value of the pulls obtained can be seen in figure 5.12. They are observed to be significantly larger in data, which indicates that the data distribution deviates significantly more from the smooth parametrization than the simulated distribution.

The overall level of agreement between the data and the parametrization can be assessed from the χ^2 value obtained after minimization. As the sum of squares of standard normally-distributed variables, this quantity follows a χ^2 distribution with a number of degrees of freedom n_{dof} given by the number of data points, minus the number of fit parameters:

$$n_{\text{dof}} = n_{\text{points}} - n_{\text{parameters}} \tag{5.22}$$

The ratio χ^2/n_{dof} has an expectation value of 1 and deviations to higher values therefore indicate an overall incompatibility between the measurement and the smooth parametrization.

Figure 5.13 shows the value of χ^2/n_{dof} as a function of the maximum order of the Chebyshev polynomials used in the smooth fit. For small p, this value is large, since the polynomial complexity is not sufficient to describe the cross section dependence. Conversely, for larger p, the value of χ^2/n_{dof} decreases as the agreement between the measurement and the parametrization improves, finally stabilizing for values greater than 6. This value is therefore chosen as the nominal value of p for the smooth parametrization.

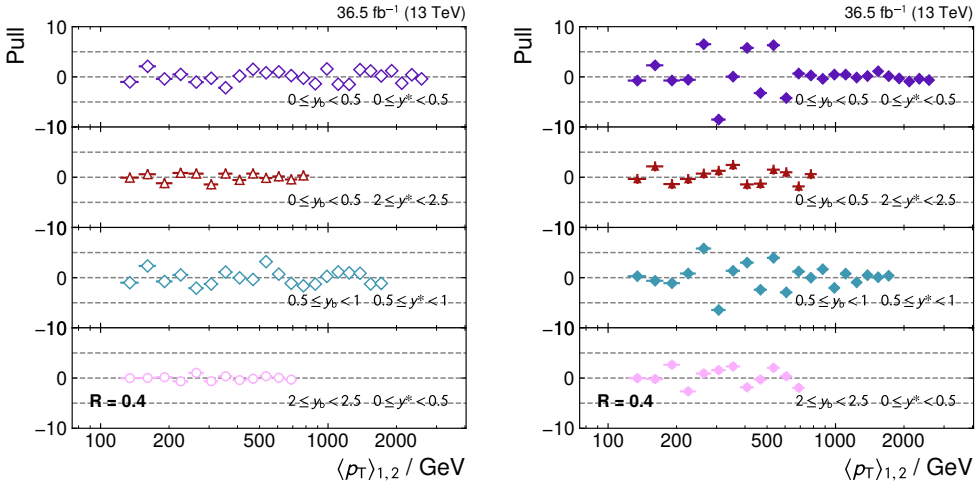

Figure 5.12 – Pulls of the smooth parametrization in four representative (y^*, y_b) regions in QCD simulation (*left*) and data (*right*), calculated with respect to the statistical uncertainty. The magnitude of the pulls observed in data are significantly larger than in simulation, reaching deviations of up to 9σ.

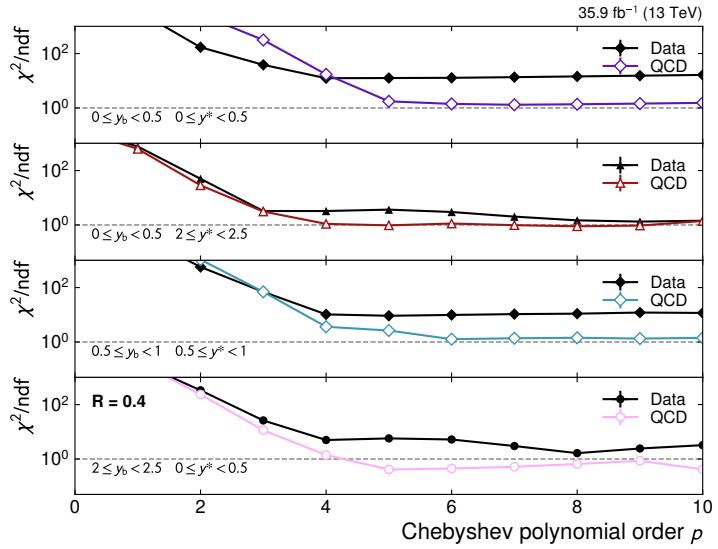

Figure 5.13 – Goodness of fit of the smooth cross section parametrization. Shown are the values of χ^2/n_{dof} obtained for a parametrization using Chebyshev polynomials of at most order p, in four representative (y^*, y_b) bins. The fit quality initially improves as more orders are added, until finally stabilizing beyond $p = 6$. In simulation (open colored markers), the parametrization stabilizes at $\chi^2/n_{\mathrm{dof}} \approx 1$, indicating that the smooth parametrization agrees with the estimates. In data (black markers), significantly higher values are observed.

While the value of χ^2/n_{dof} beyond $p = 6$ is close to the expectation value of 1 in simulation, this is not the case in data, where the values of χ^2 are significantly higher, only reaching down to values of ≈ 20 in some (y^*, y_b) bins.

This is an indication of an additional systematic effect in data that causes a deviation from the smooth parametrization. To account for this effect, an additional systematic uncertainty is introduced, taken as a bin-by-bin uncorrelated uncertainty relative to the measurement value. With this additional uncertainty a, the χ^2 definition becomes:

$$\chi^2 = \sum_i \left(\frac{y_i - f(x_i, \theta)}{\sigma_i + a\, y_i} \right)^2 \tag{5.23}$$

The final value of the additional uncertainty a is determined separately in each (y^*, y_b) bin by repeating the fit used to obtain the smooth parametrization with multiple values of a. The smallest value of a for which a value of $\chi^2/n_{\text{dof}} < 1$ is reached is taken as the final value of the additional uncertainty. The values obtained for a can be seen in table 5.6.

The inclusion of the additional uncertainty included in the χ^2 definition results in a smooth parametrization of the measurement $f(x, \theta)$ that remains statistically compatible with it, while also accounting for the enlarged data fluctuations by construction. The smoothed cross section values and associated uncertainties are obtained by evaluating the smooth parametrization:

$$y_i^{\text{smooth}} = f(x_i, \theta) \tag{5.24}$$

$$\sigma_i^{\text{smooth}} = \sigma_i + a\, y_i^{\text{smooth}} \tag{5.25}$$

Effect on unfolding

To assess the effect that the uncovered discontinuities in the detector-level distribution in data have on the unfolding result, the unfolding is performed using the smooth parametrization and its associated enlarged uncertainties as an input. The resulting unfolded cross sections are then compared to the unfolded cross sections obtained from simulation.

Figure 5.14 shows the resulting data/simulation ratios with and without input smoothing. If the smoothing is not applied, the ratios exhibit discontinuities that are not covered by the uncertainty. These are not present when input smoothing is used, and any remaining fluctuations are compatible within the uncertainties.

Owing to the factors outlined above, it is therefore concluded that performing a smooth parametrization of the unfolding inputs in data results in a smaller systematic bias on the unfolded cross section estimates. Accordingly, the smooth parametrization is used as an input for obtaining the final unfolded cross sections. The potential bias incurred

Table 5.6 – Additional uncorrelated uncertainty introduced in order to reflect the enlarged fluctuations observed in data compared to the simulation. The values are determined by performing a smooth parametrization of the cross sections as a function of $\langle p_T \rangle_{1,2}$ and m_{jj} in each (y^*, y_b) bin, and choosing a value of a so that the χ^2/n_{dof} of the parametrization is ≈ 1. Different sets of values are obtained for the two jet radii.

y^* **range**	y_b **range**	**Additional uncorrelated uncertainty** a **(%)**			
		$\langle p_T \rangle_{1,2}$		m_{jj}	
		$R = 0.4$	$R = 0.8$	$R = 0.4$	$R = 0.8$
[0.0, 0.5)	[0.0, 0.5)	0.5	1.3	0.6	1.7
[0.0, 0.5)	[0.5, 1.0)	0.8	1.8	0.6	2.0
[0.0, 0.5)	[1.0, 1.5)	0.9	1.4	0.7	2.0
[0.0, 0.5)	[1.5, 2.0)	0.9	1.2	0.8	2.3
[0.0, 0.5)	[2.0, 2.5)	1.8	2.1	0.1	1.0
[0.5, 1.0)	[0.0, 0.5)	0.8	1.4	1.0	1.8
[0.5, 1.0)	[0.5, 1.0)	0.6	1.4	0.5	1.9
[0.5, 1.0)	[1.0, 1.5)	1.0	1.4	1.2	1.8
[0.5, 1.0)	[1.5, 2.0)	1.3	2.5	3.2	1.5
[1.0, 1.5)	[0.0, 0.5)	0.9	1.5	0.7	2.0
[1.0, 1.5)	[0.5, 1.0)	0.6	1.1	0.3	2.0
[1.0, 1.5)	[1.0, 1.5)	1.2	1.9	2.3	0.7
[1.5, 2.0)	[0.0, 0.5)	0.6	0.8	0.8	1.3
[1.5, 2.0)	[0.5, 1.0)	0.2	2.1	0.0	1.3
[2.0, 2.5)	[0.0, 0.5)	2.1	2.4	0.0	2.3

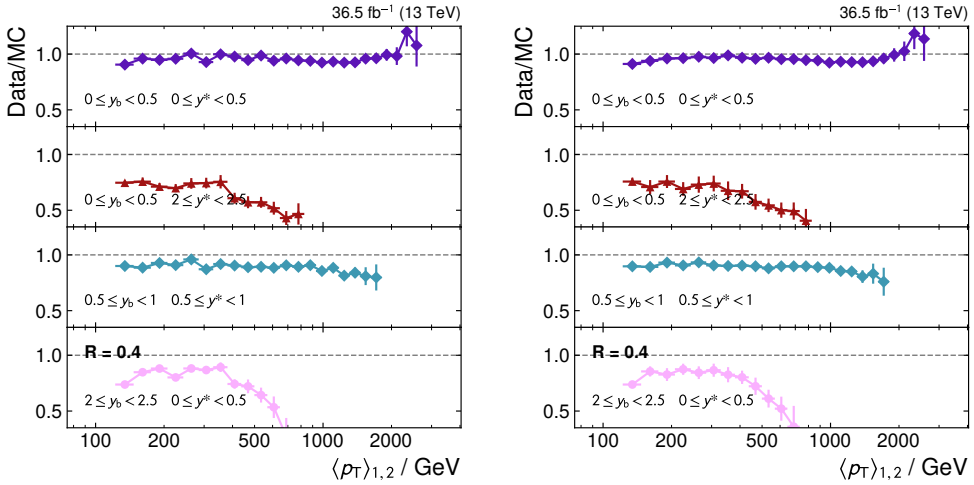

Figure 5.14 – Ratio of the unfolded cross sections in data and simulation in four selected (y^*, y_b) bins, shown without (*left*) and with (*right*) applying smoothing to the unfolding input in data. If left unsmoothed, the observed fluctuations in data which are not observed in the simulation cause the unfolding to perform poorly, yielding jumps in the data/simulation ratio which are not covered by the uncertainty. With smoothing, these jumps are attenuated, and the associated additional uncertainty covers the remaining fluctuations.

by the smoothing itself is accounted for in the final uncertainty model by the additional, binwise uncorrelated uncertainty.

5.6 Uncertainties

The cross section measurement is subject to experimental uncertainties due to a number of statistical and systematic effects. The main uncertainty contributions are studied in detail and are described in the following.

Statistical and additional uncorrelated uncertainty

Due to the finite sample size, the cross section estimates are subject to a statistical uncertainty typically modeled by a Poisson distribution. In this analysis, event numbers are sufficiently high for this to be approximated by a Gaussian distribution instead. The uncertainty size is computed from the standard deviation of the Gaussian approximation, which is given by the square root of the sum of the squares of all event weights in a bin

While the statistical uncertainties of the detector-level distribution are uncorrelated between different bins, such a correlation is induced by the unfolding procedure. An esti-

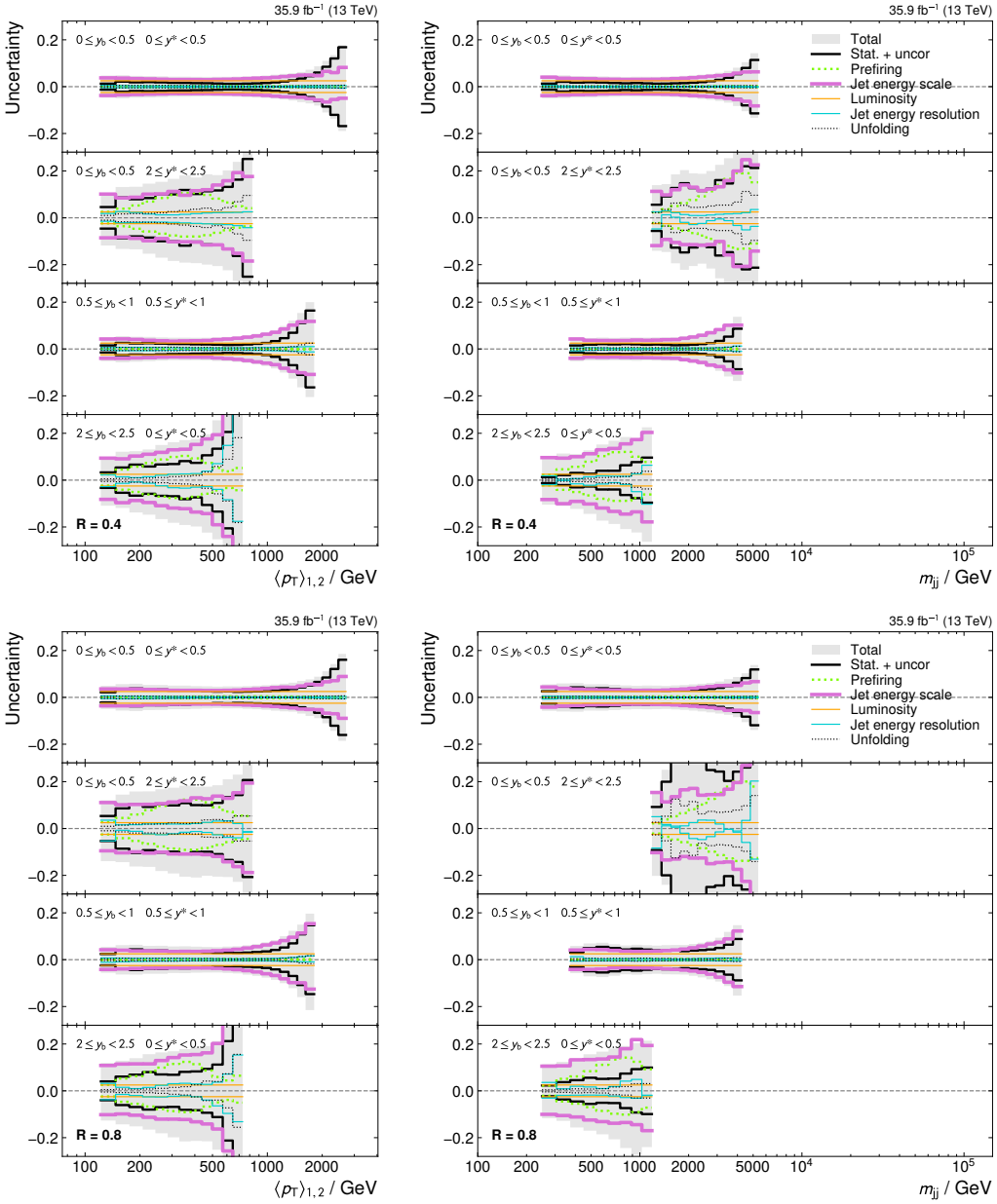

Figure 5.15 – Relative uncertainty on the cross section, showing the contributions from different uncertainty sources in four representative (y^*, y_b) regions. Overall, the dominant contribution is given by the jet energy scale uncertainty. A complete set of plots showing the uncertainties in other rapidity regions can be found in the appendix (section A.2).

mate of the correlated statistical uncertainty on the unfolded cross sections is obtained by propagating the covariance matrix through the unfolding procedure, as described in equation (5.16). The resulting correlations are strongest between neighboring $\langle p_{\mathrm{T}} \rangle_{1,2}$ (or m_{jj}) bins, which exhibit negative correlation coefficients of the order of -50%, while bins separated by one intermediate bin exhibit a weaker, positive correlation. The correlation matrices can be seen in figures A.9 and A.10 in the appendix (section A.3).

Furthermore, as discussed in section 5.5.4, fluctuations in the detector-level distributions are observed in data to a higher degree than in the simulation. In order to avoid an adverse effect on the unfolded cross sections, the unfolding input in data is parametrized by a smooth function. An additional binwise uncorrelated uncertainty is introduced to account for the potential bias incurred by the smoothing. The uncertainty values are given in table 5.6. Since these are propagated through the unfolding procedure, they exhibit the same correlation structure as the statistical uncertainty on the unfolded cross section.

Jet energy scale (JES)

Since corrections are applied to the jet energy in multiple stages (see section 4.2), the transverse momentum of reconstructed jets is subject to multiple sources of systematic bias. The uncertainty on the jet energy associated with this bias is on the level of a few percent. Propagated to the cross section, the effect is amplified by the steeply falling spectrum, and thus represents the largest source of systematic bias in this analysis.

An estimate of the jet energy scale uncertainty is obtained by repeating the analysis with systematic upward and downward shifts applied to the transverse momentum of each reconstructed jet, according to the total jet energy scale uncertainty. The latter is provided as a function of the individual jet kinematics and other event properties along with the centrally-derived jet energy corrections.

The detector-level cross sections resulting from the shifted analyses are propagated through the unfolding procedure, resulting in an upward and downward shifted estimate of the unfolded cross section. To prevent a bias on the uncertainty estimate resulting from any residual fluctuations, the unfolded cross sections obtained from the shifted inputs are parametrized by a smooth function, as described in section 5.5.4. Comparing the shifted cross section estimated to the nominal ones results in an asymmetric confidence interval, which is taken as the final uncertainty.

Jet energy resolution (JER)

The detector response used as an input for the unfolding procedure is based on the jet energy resolution assumed in simulation. While this is explicitly adjusted to correspond to the resolution seen in data (see section 4.2.4), the scale factors used remain subject to an uncertainty, which is centrally provided alongside them.

The impact of the jet energy resolution modeling on the unfolded cross section is estimated by repeating the unfolding procedure with alternative responses derived using systematic variations of the jet energy resolution. This is achieved by varying the simulation-to-data scale factor used in the jet energy resolution (JER) smearing within its uncertainty. Both an upward and downward variation are provided.

As with the jet energy scale, the final uncertainty due to JER is obtained from an asymmetric confidence interval around the nominal cross section, as given by the shifted estimates.

Prefiring uncertainty

The correction of the L1 trigger prefiring inefficiency in recorded data introduces an additional uncertainty (see section 5.3.3). This is estimated by repeating the analysis with a different set of prefiring weights, which correspond to an upward and downward variation of the prefiring correction, as determined by the total uncertainty of the prefiring efficiency.

This uncertainty is most significant in the (y^*, y_b) regions at the edge of the phase space, where it can reach values of 10–20%. This is due to the high likelihood for an event in these regions to contain a jet in the pseudorapidity region affected by prefiring ($2 < |\eta| < 3$). Outside these regions, however, the uncertainty is negligible.

Luminosity uncertainty

The effective integrated luminosities that enter into the cross section calculation are obtained from measurements of the instantaneous luminosity performed live during data taking. The systematic uncertainty of the luminosity inherent to these measurements is estimated at 2.5% and propagated to the cross section.

Unfolding

As mentioned in section 5.5, an additional source of uncertainty introduced by the unfolding process is due to the finite size of the simulated sample used to derive the detector response. The response matrix obtained is therefore subject to a statistical uncertainty which can cause fluctuations to appear in the unfolded cross sections even when none are present in the input detector-level distribution. This effect is accounted for by introducing an additional unfolding uncertainty.

This is achieved by first dividing the simulated sample randomly into N statistically independent subsamples. The unfolding procedure is then applied repeatedly to the

input distribution, each time using a different response matrix, derived from each individual subsample. This results in N different estimates for the unfolded cross section, which are treated as a statistical ensemble.

The standard error of the mean, as calculated in each measurement bin from these N estimates, is a measure of the statistical uncertainty on the unfolded cross section induced by the unfolding procedure. It is obtained by dividing the standard deviation $\sigma(x)$ of the cross section estimates $\{x\}$ by the square root of the number of estimates:

$$\delta(\langle x \rangle) = \sigma(x)/\sqrt{N} \tag{5.26}$$

It is worth noting that the standard deviation used here is particularly sensitive to outliers. To increase the robustness of the uncertainty estimator, a truncation is first performed by removing the outlying 10% of estimates in each bin. An unbiased uncertainty estimate is then obtained from the standard deviation of the truncated sample, correcting for the truncation by an analytic factor derived from the variance of a truncated normal distribution.

In this approach, N is a free parameter, the choice of which represents a tradeoff between two kinds of statistical precision. On the one hand, if the chosen value for N is too large, the number of events in each subsample will no longer be sufficient for obtaining a reliable estimate of the response matrix. In particular, this increases the likelihood of distributions containing empty bins, in which case the response matrix becomes singular and the unfolding is no longer possible. On the other hand, small values of N do not provide a sufficient number of independent cross section estimates, decreasing the reliability of the method. For this analysis, a value of $N = 100$ is found to be an adequate choice.

The unfolding uncertainties obtained are of the order of 1%, remaining mostly below this value in well-populated areas of the phase space. A cross-check is performed for an alternative choice of $N = 50$, which yields comparable estimates in the vast majority of bins. The only exception to this are bins at the upper edge of the $\langle p_{\mathrm{T}} \rangle_{1,2}$ (or m_{jj}) phase space, where the lower number of reconstructed events induces a larger spread of the estimates of the unfolded cross section.

5.7 Comparison to fixed-order theory

Fixed-order theory predictions of the triple-differential cross section are obtained using the software package NNLOJET [73]. These calculations are used to derive interpolation tables with fastNLO [74, 75], which allow cross section predictions to be obtained for arbitrary PDFs.

The calculations are performed at NLO and NNLO accuracy in perturbative QCD for jets reconstructed with both of the radius parameters used for the measurement: $R = 0.4$

and $R = 0.8$. The dijet invariant mass m_{jj} is chosen as the characteristic scale of the hard process. This is in line with theoretical investigations of the perturbative convergence which have identified this variable to be a suitable choice for predictions of dijet production [76].

The unfolded triple-differential cross sections in all 15 (y^*, y_b) regions are presented in figure 5.16, along with the corresponding theory predictions. Overall, the shape of the spectrum corresponds to the expected exponential dependence in all of the examined phase space, with some deviations observed at high values of y_b where the highest sensitivity to the PDFs is expected.

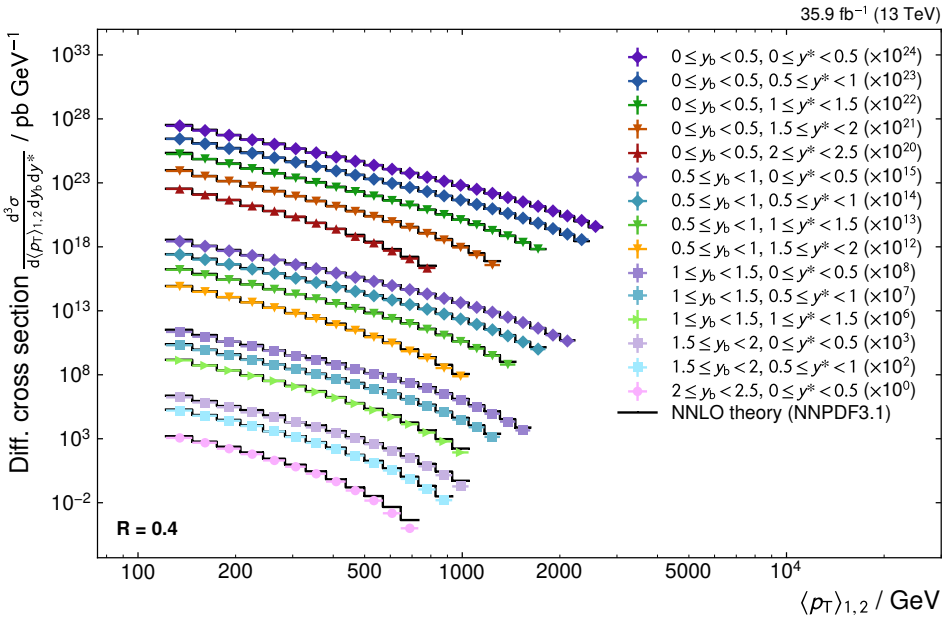

Figure 5.16 – Overview of the unfolded cross sections, compared to fixed-order theory at NNLO accuracy. The corresponding cross section values in each (y^*, y_b) region are scaled by a constant factor to minimize overlap for illustration purposes. The overall spectrum is well described by the theory, with some deviations seen at high values of y_b. Similar plots for the other measurements can be found in section A.3 of the appendix.

The agreement is investigated in more detail in each (y^*, y_b) region by examining the ratio of the measured cross sections to theory predictions obtained for different PDFs. A selection of four recent PDF sets from global determinations at NNLO accuracy are considered: `NNPDF3.1` [18], `CT14` [19], `MMHT2014` [20], and `ABMP16` [21]. Figures 5.17 and 5.18 show the predictions obtained for these PDFs in a representative selection of rapidity regions and the corresponding measurements as a ratio to the `NNPDF3.1` prediction for jets with $R = 0.4$ and $R = 0.8$, respectively. A full set of such plots can be found in section A.4 of the appendix.

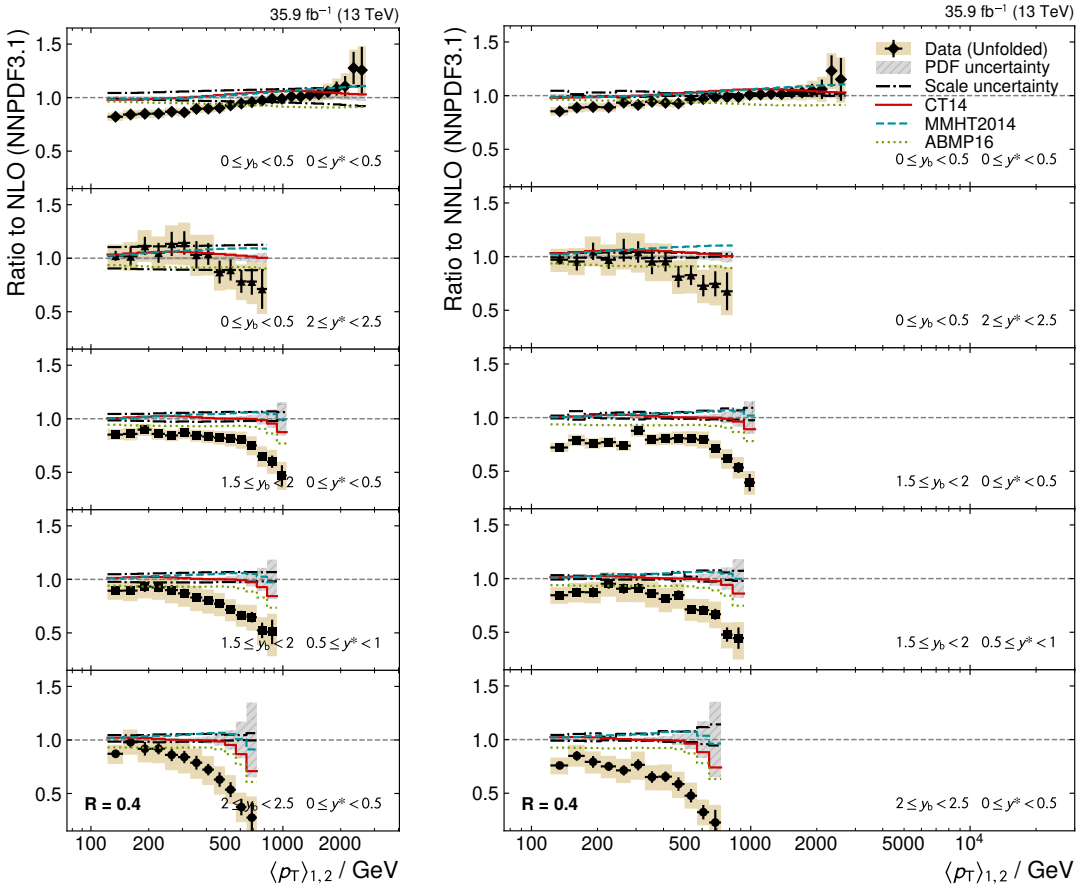

Figure 5.17 – Ratio of the unfolded cross sections to NLO (*left*) and NNLO (*right*) theory predictions, shown for jets with $R = 0.4$ in the central (y^*, y_b) region (*top*), the high-y^* region, and three bins at highest y_b (*bottom 3*). The reference predictions and PDF uncertainties are obtained with the NNPDF3.1 PDF set, with other PDFs also shown for comparison. The error bars and the yellow shaded area represent the statistical and total uncertainty of the data, respectively. The fluctuations observed for NNLO are due to the limited statistical precision of the calculation.

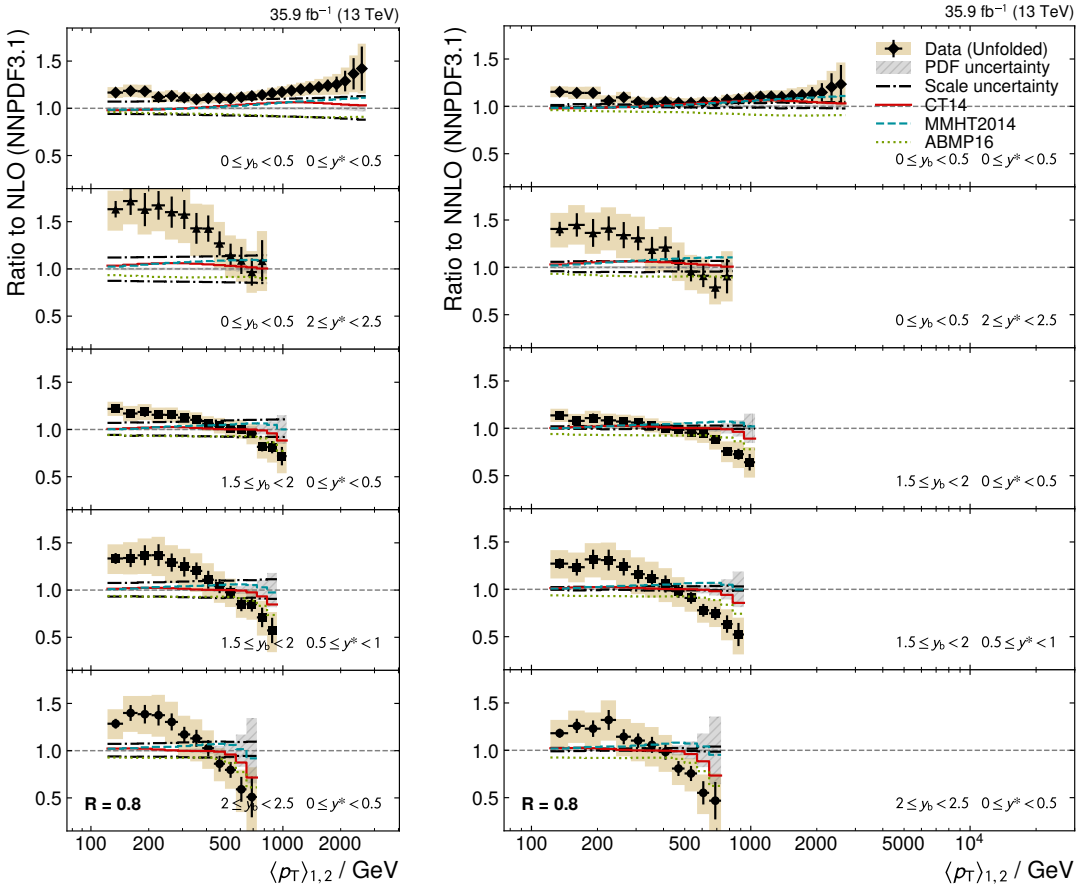

Figure 5.18 – Ratio of the unfolded cross sections to NLO (*left*) and NNLO (*right*) theory predictions, shown for jets with $R = 0.8$ in the central (y^*, y_b) region (*top*), the high-y^* region, and three bins at highest y_b (*bottom 3*). The reference predictions and PDF uncertainties are obtained with the NNPDF3.1 PDF set, with other PDFs also shown for comparison. The error bars and the yellow shaded area represent the statistical and total uncertainty of the data, respectively. The fluctuations observed for NNLO are due to the limited statistical precision of the calculation.

Due to the long processing times required for the NNLO calculations, they were only available with a limited statistical precision at the time of writing, leading to increased fluctuations in the ratios. Despite this, the available precision is sufficient to recognize the general trend of the predictions and assess the data/theory agreement within uncertainties.

In addition to the experimental uncertainties, the uncertainty on the theory predictions due to the PDFs is evaluated. It is shown to increase as a function of $\langle p_\mathrm{T} \rangle_{1,2}$ and is generally largest in the high-y_b region. Furthermore, the uncertainty due to missing higher orders in perturbation theory ("scale uncertainty") is estimated from independent variations of the renormalization and factorization scales by factors of 2 and 0.5, taking the envelope of the resulting predictions as the uncertainty. In general, it is reduced at NNLO accuracy compared to NLO.

In most regions, the data are seen to be in agreement with the calculations, within the experimental and theory uncertainties. However, a number of deviations are observed, with the data exhibiting different trends for the $R = 0.4$ and $R = 0.8$ measurements, which are addressed in the following.

For $R = 0.4$, the data generally indicate a preference for smaller cross sections. At low values of y_b and y^*, this trend is strongest in the lower part of the $\langle p_\mathrm{T} \rangle_{1,2}$ spectrum. In this region, contributions from nonperturbative effects are known to be significant. Since they are not accounted for in the parton-level calculations presented here, this may explain part of the observed tension. Further corrections from electroweak radiative effects also contribute, becoming sizeable beyond 1 TeV, particularly at low y^* [5].

In the high-y^* region, the data are largely compatible with the predictions, with only a small deviation to lower values beyond 500 GeV. In contrast, the data in the high-y_b regions indicate a clear preference for consistently lower cross sections across the entire $\langle p_\mathrm{T} \rangle_{1,2}$ range, exhibiting deviations from the theory predictions that are noticeably larger than the confidence interval defined by the PDF, scale and experimental uncertainties. Barring fluctuations due to the limited statistical precision of the NNLO calculation, a greater tension is observed for the NNLO theory compared to NLO.

For $R = 0.8$, an overall difference in shape is observed between the data and the theory, with the former generally showing a preference for larger cross sections in the lower part of the spectrum. This is in contrast to the $R = 0.4$ measurement and is consistent with nonperturbative corrections, which depend strongly on the jet radius. At high $\langle p_\mathrm{T} \rangle_{1,2}$, the same trend is seen as for the $R = 0.4$ measurement in all rapidity regions, consistent with the expectation of similar electroweak corrections, for which the jet radius dependence is minimal. Some tension is observed at low $\langle p_\mathrm{T} \rangle_{1,2}$ in all rapidity regions. At high y_b, this is also the case in the upper part of the spectrum where the PDF uncertainties become sizable, although the effect is not as pronounced as for $R = 0.4$.

While a full evaluation would require an improved statistical precision of the NNLO calculation and the derivation of corrections for nonperturbative and electroweak radiative effects, the present analysis already provides a strong indication of the sensitivity of

the measurement to PDFs in the boosted region, where the parton densities are probed simultaneously at different fractions of the proton momentum. The inclusion of the present measurement in PDF determinations is therefore expected to further constrain the parton densities in these regions.

6 Conclusion and Outlook

In this thesis, a triple-differential measurement of the dijet cross section has been presented, using proton-proton collision data recorded by the CMS experiment at a center-of-mass energy of 13 TeV in 2016, corresponding to 35.9 fb^{-1}.

The triple-differential cross section was measured in parallel using anti-k_T jets with radius parameters of $R = 0.4$ and $R = 0.8$ as a function of two sets of kinematic variables: the dijet rapidity separation y^*, the total boost of the dijet system y_b, and either the average dijet transverse momentum $\langle p_\text{T} \rangle_{1,2}$ or the dijet invariant mass m_jj as the third variable. Such measurements are performed for the first time at 13 TeV, with the present analysis also being the first to include a systematic study of the triple-differential jet cross section for both $\langle p_\text{T} \rangle_{1,2}$ and m_jj, as well as for two jet radii simultaneously.

The dijet spectrum is reconstructed by combining data from multiple triggers across a wide kinematic range, starting at a value of 122 GeV in $\langle p_\text{T} \rangle_{1,2}$ and 249 GeV in m_jj. The measurement is performed in 15 bins of the joint (y^*, y_b) phase space, going up to 2.5 in each variable. Detector effects are corrected through a three-dimensional unfolding procedure, which corrects for bin-to-bin migrations in all three measurement variables simultaneously. Through an appropriate choice of binning with respect to the resolution of the measurement variables, a well-conditioned response matrix is achieved, thus eliminating the need for further regularization to suppress response-induced fluctuations.

A detailed study of the experimental sources of uncertainty on the cross section is performed. The jet energy scale is identified as the largest single uncertainty source, ranging from 5% to more than 25% towards the upper end of the spectrum at high rapidities. A further notable uncertainty is due to *trigger prefiring*, which is a source of trigger inefficiency linked to the presence of forward jets. In most of the investigated phase space, the uncertainty is of the order of 1% or below, rising to 10% only in the outer rapidity regions where there is a significant contribution from forward jets.

The goal of these measurements is to provide inputs for PDF determinations. The choice of y^* and y_b as rapidity variables serves to better disentangle the PDF dependence of the cross section from the dependence on the fixed-order matrix elements. In addition, the large amount of data makes a triple-differential measurement possible with a high statistical precision.

To assess the potential impact of the current measurement on PDF determinations, the unfolded cross sections are compared to fixed-order calculations at NNLO accuracy in perturbative QCD. In contrast to previous impact studies, such as the one performed in reference [6], which makes use of interpolation tables accurate to NLO and binwise k-factors to estimate NNLO predictions, the theory treatment here consists in the computation of full interpolation tables at NNLO accuracy. While these offer a higher degree of flexibility for this purpose, their calculation requires significantly longer processing times to accumulate the statistical precision needed for PDF determinations at NNLO.

Nevertheless, with the statistical precision currently available it is possible to give a first look into NNLO phenomenology. Taking the current calculations as a baseline for the comparison, it is observed that they are largely in agreement with the data in most phase space regions. However, potentially significant deviations from the fixed-order theory are observed. Particularly at high y_b, where the PDF-induced uncertainty is largest, the deviation from theory exceeds the experimental and theory uncertainty, underlining the benefit of including the presented measurements in PDF determinations.

In light of these findings, several avenues of study can be identified for further improving the results. A crucial aspect here consists in enhancing the statistical precision of the NNLO fixed-order calculation. In addition, the inclusion of further data from the 2017 and 2018 data-taking periods would increase the integrated luminosity by more than a factor of three, leading to an improvement of the statistical precision of the measurement in the less populated areas of the phase space at high p_T.

Furthermore, in order to improve the description of the data, correction factors accounting for nonperturbative effects, as well as electroweak contributions, must be derived and applied to the fixed-order theory.

With the above improvements in place, it will be possible to evaluate the full benefit of the current measurement for the determination of PDFs and the strong coupling constant α_s by performing a fit of the NNLO theory predictions to the data.

A Appendix

A.1 Monte Carlo samples

The following tables list the Monte Carlo samples used throughout this analysis. The Pythia 8 QCD signal samples are listed in A.1, indicating the exact sample names and the number of events, as they appear in the CMS Data Aggregation System (DAS). The cross sections are taken from the CMS Cross Section Database (XSDB).

Table A.1 – Signal MC samples

DAS name		
/QCD_Pt_⟨p_T *slice*⟩_TuneCUETP8M1_13TeV_pythia8/ RunIISummer16MiniAODv3-PUMoriond17_94X_mcRun2_asymptotic_v3-v*/MINIAODSIM		
p_T slice	**Cross section** / pb	**# of events**
30to50	138900000	9980050
50to80	19100000	9954370
80to120	2735000	6986740
120to170	467500	6708572
170to300	117400	6958708
300to470	7753	4150588
470to600	642.1	3959986
600to800	185.9	3896412
800to1000	32.05	3992112
1000to1400	9.365	2999069
1400to1800	0.8398	396409
1800to2400	0.1124	397660
2400to3200	0.006752	399226
3200toInf	0.0001626	391735

The background samples are obtained with MadGraph and Pythia 8 and comprise simulations of W boson production (denoted "W($\to \ell\nu$) + jets"), production of a Z boson decaying to neutrinos ("Z($\to \nu\bar{\nu}$) + jets"), and top quark pair production ("$t\bar{t}$ + jets"). They are listed in table A.2.

Table A.2 – Background MC samples

Sample
$Z(\to \nu\bar{\nu})$ + jets
`/ZJetsToNuNu_Zpt-*_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/`
`RunIISummer16MiniAODv3-PUMoriond17_94X_mcRun2_asymptotic_v3-v2/MINIAODSIM`
$W(\to \ell\nu)$ + jets
`/WJetsToLNu_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/`
`RunIISummer16MiniAODv3-PUMoriond17_94X_mcRun2_asymptotic_v3-v2/MINIAODSIM`
$t\bar{t}$ + jets
`/TTJets_TuneCUETP8M1_13TeV-madgraphMLM-pythia8/`
`RunIISummer16MiniAODv3-PUMoriond17_94X_mcRun2_asymptotic_v3-v1/MINIAODSIM`

A.2 Uncertainties on the unfolded cross section

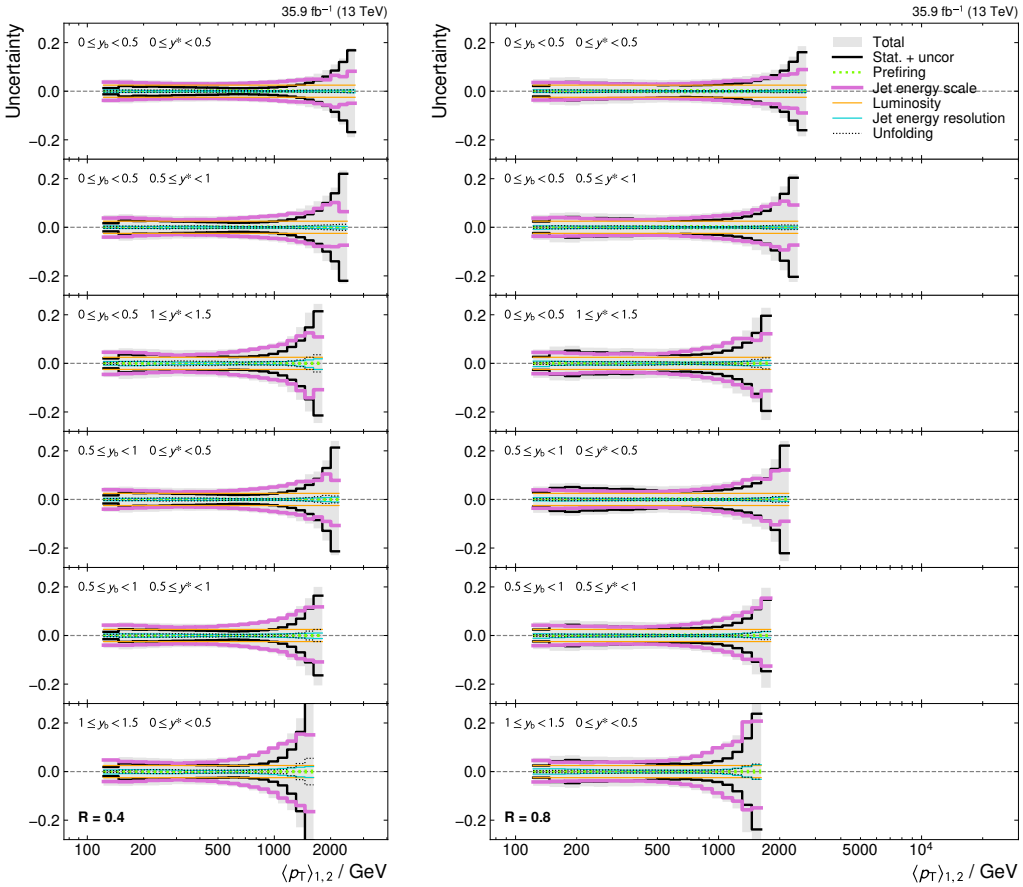

Figure A.1 – Relative uncertainty on the cross section measurement as a function of (y^*, y_b, $\langle p_T \rangle_{1,2}$), displaying the contributions from different sources. Shown are 6 out of 15 rapidity bins at central rapidities. The remaining 9 rapidity regions are shown in figure A.2.

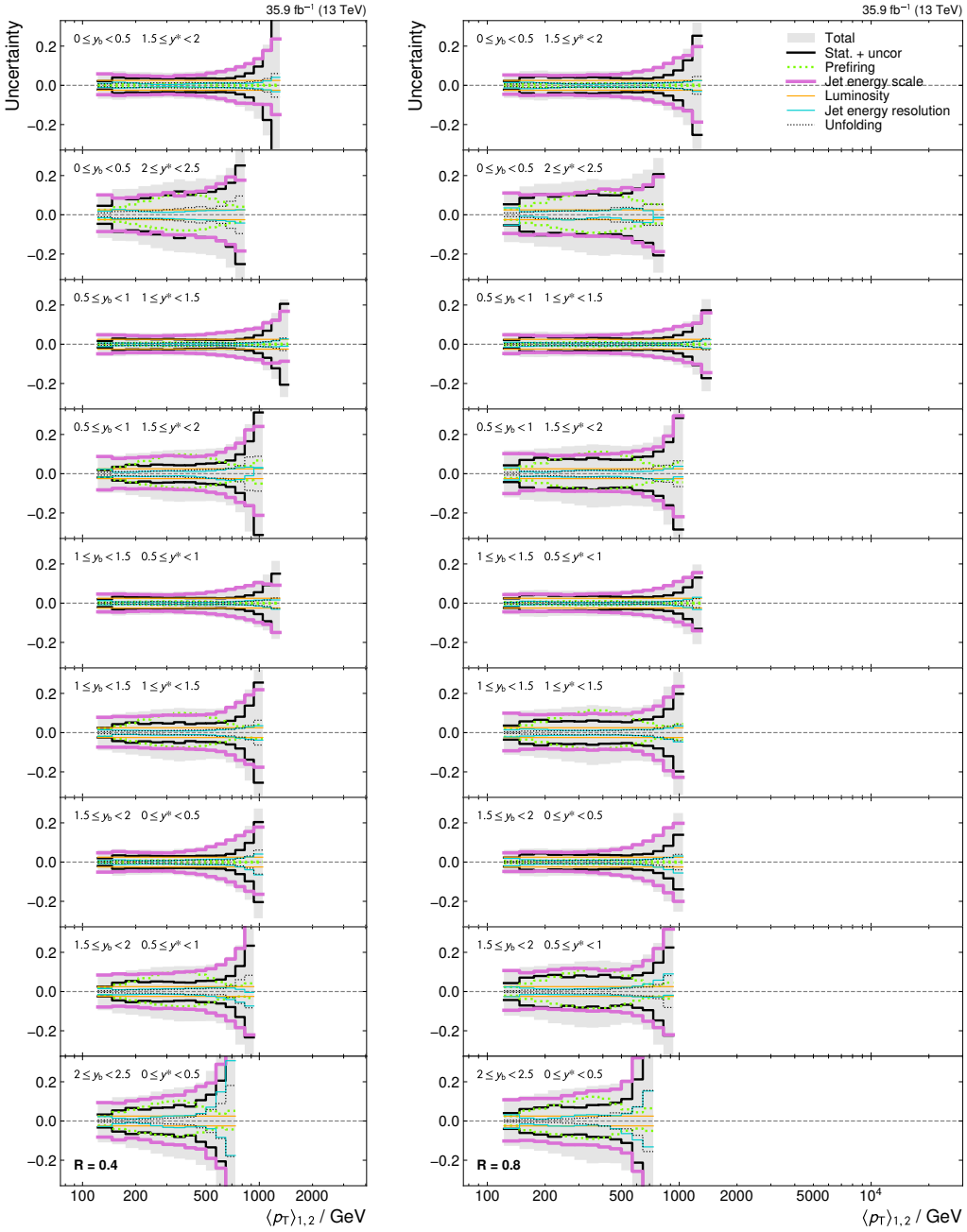

Figure A.2 – (*continued from figure A.1*) Relative uncertainty on the cross section measurement as a function of (y^*, y_b, $\langle p_T \rangle_{1,2}$), displaying the contributions from different sources.

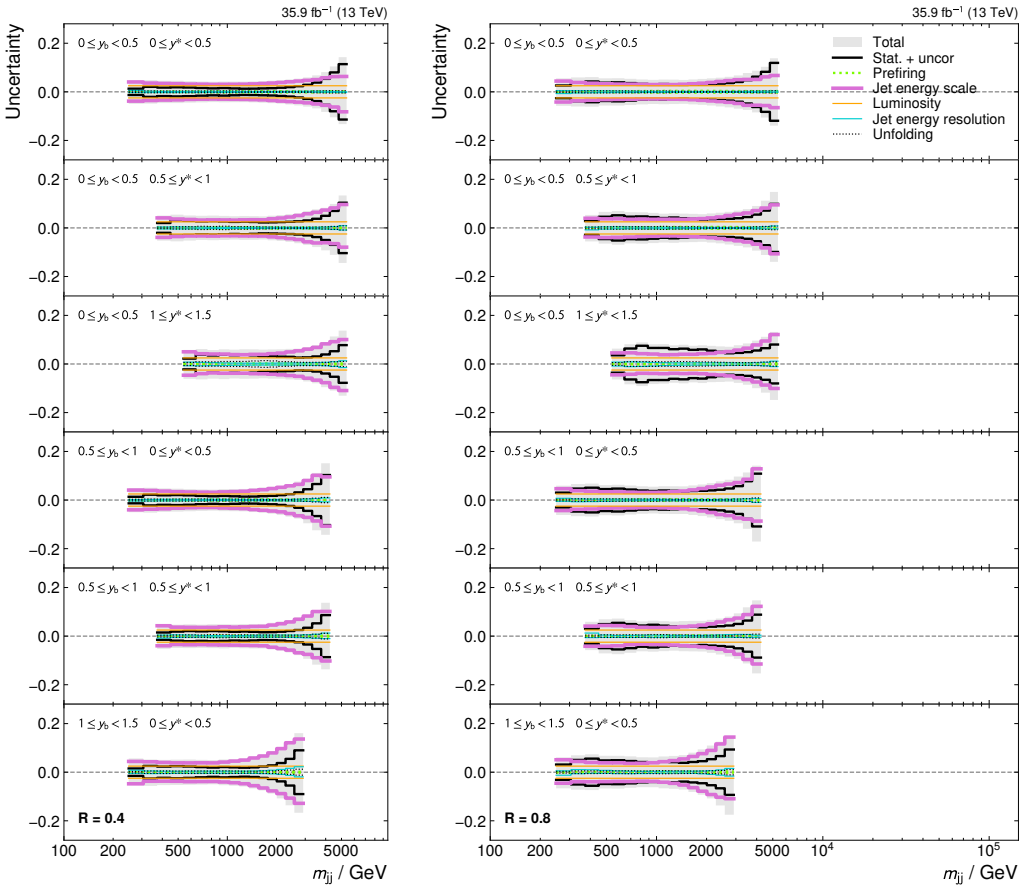

Figure A.3 – Relative uncertainty on the cross section measurement as a function of (y^*, y_b, m_{jj}), displaying the contributions from different sources. Shown are 6 out of 15 rapidity bins at central rapidities. The remaining 9 rapidity regions are shown in figure A.4.

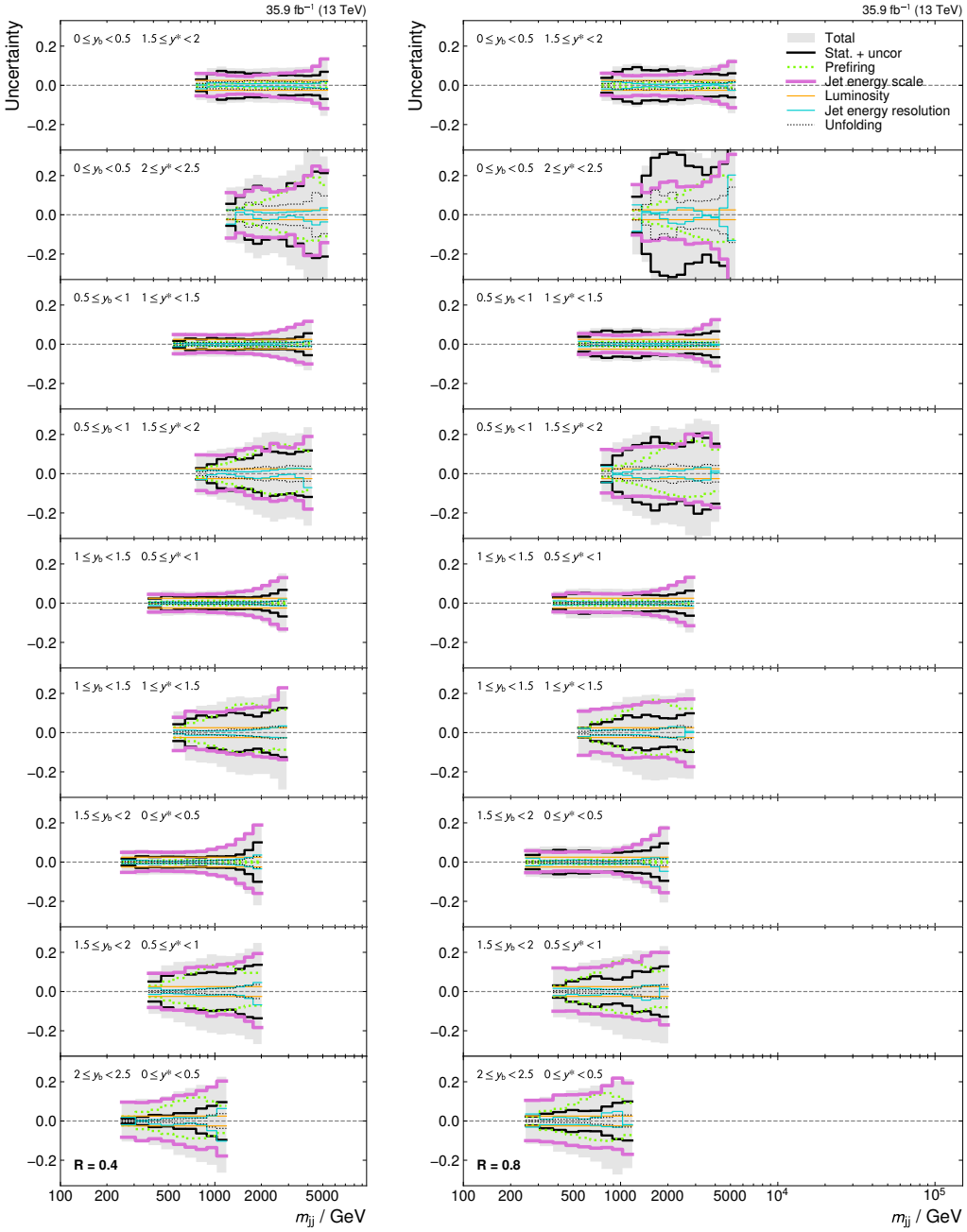

Figure A.4 – (*continued from figure A.3*) Relative uncertainty on the cross section measurement as a function of (y^*, y_b, m_{jj}), displaying the contributions from different sources.

A.3 Supplementary material

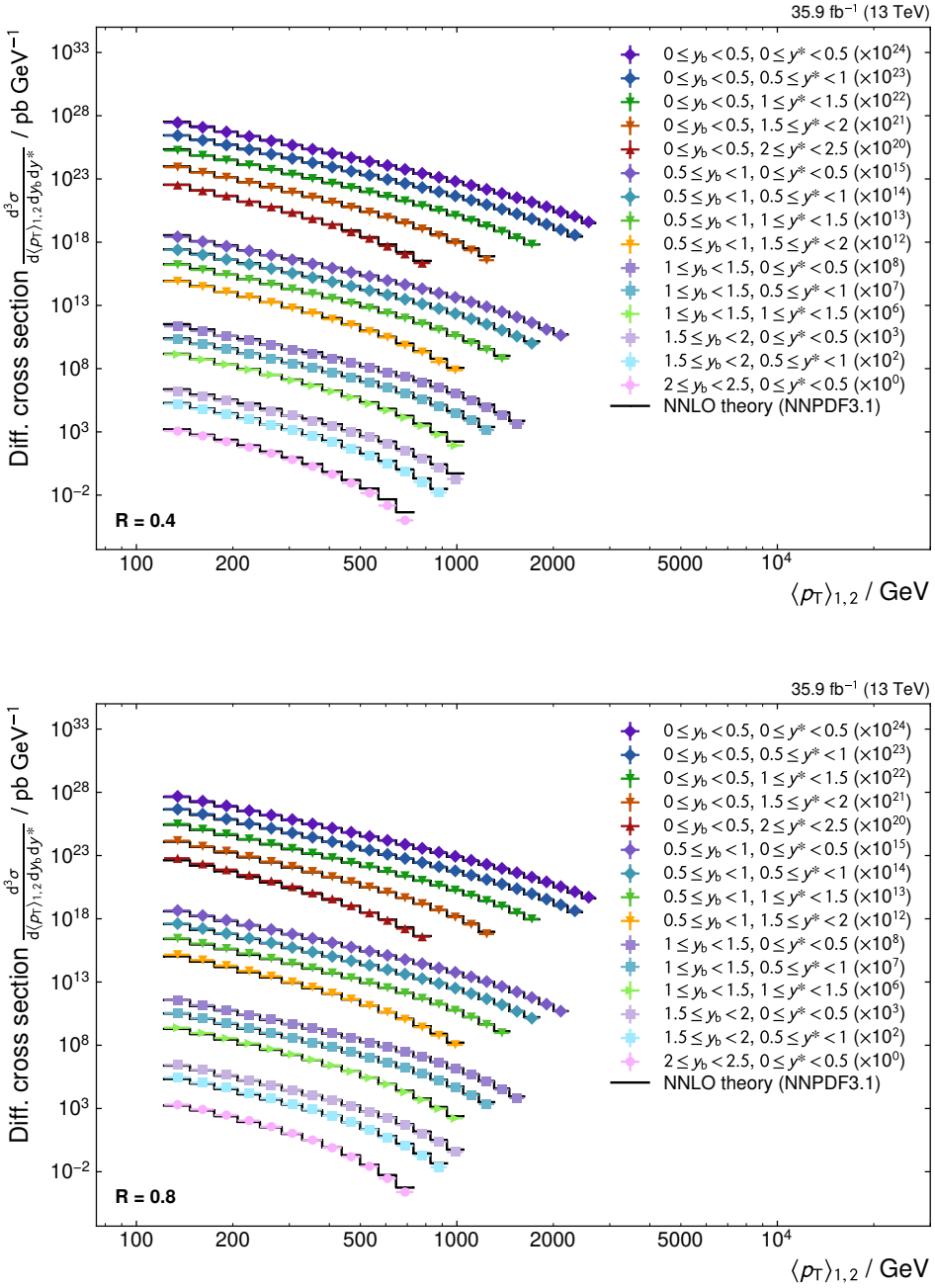

Figure A.5 – Overview of the unfolded cross sections as a function of $\langle p_T\rangle_{1,2}$ for jets with $R = 0.4$ (*top*) and $R = 0.8$ (*bottom*), compared to fixed-order theory at NNLO accuracy.

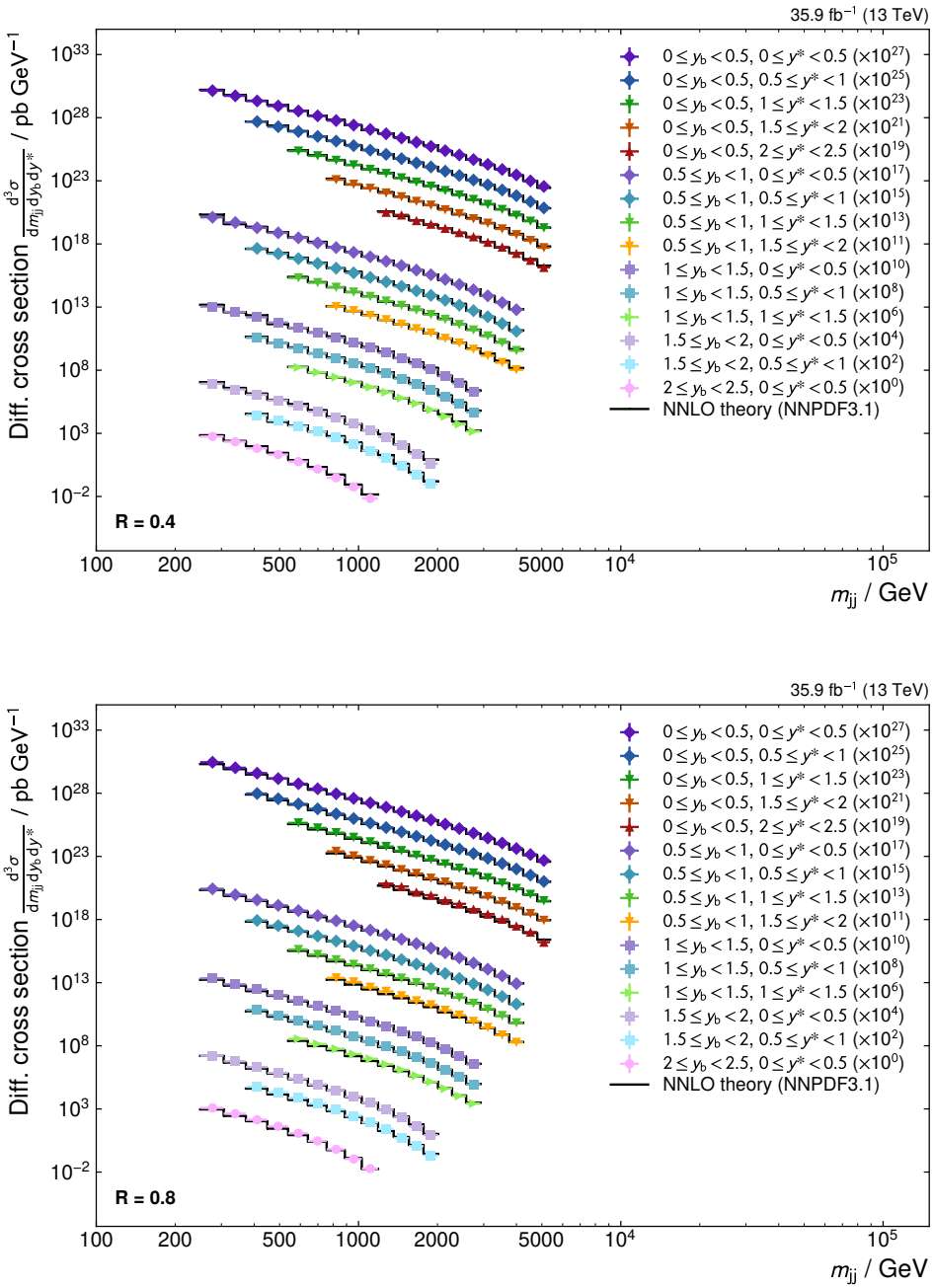

Figure A.6 – Overview of the unfolded cross sections as a function of m_{jj} for jets with $R = 0.4$ (*top*) and $R = 0.8$ (*bottom*), compared to fixed-order theory at NNLO accuracy.

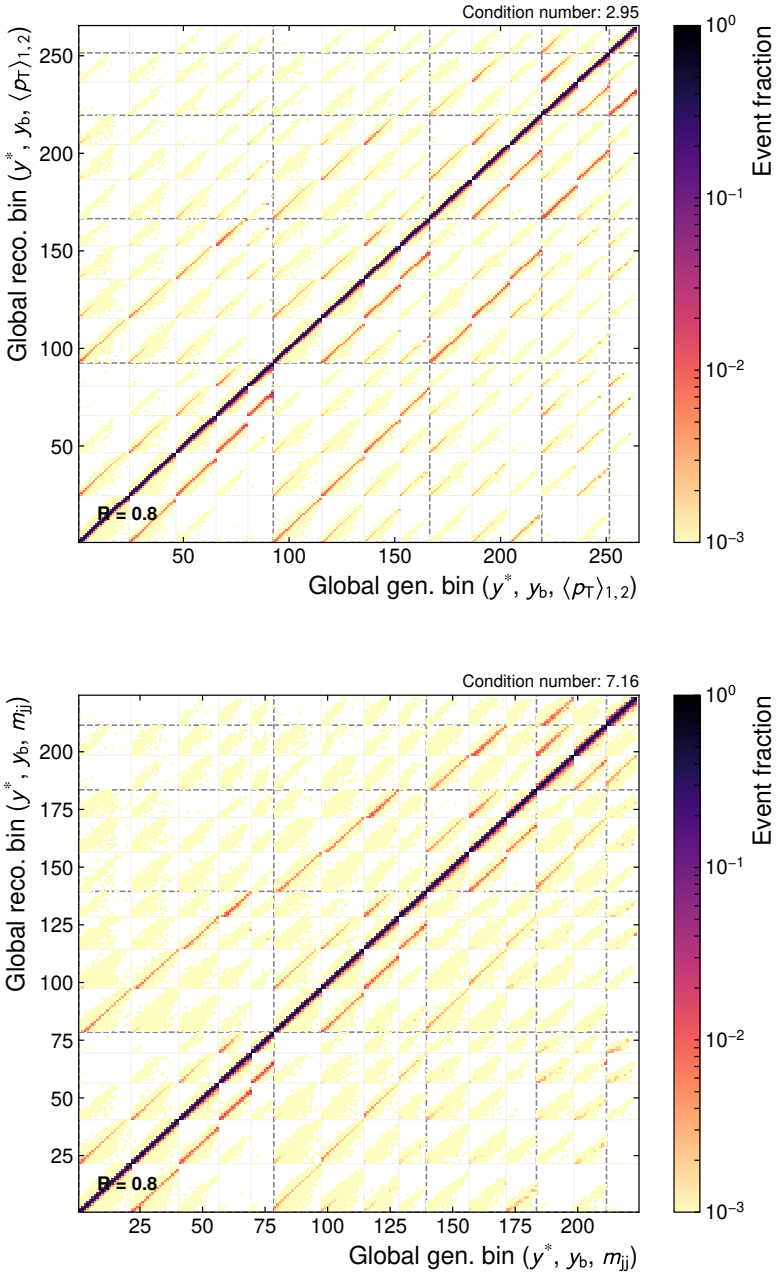

Figure A.7 – Migration probability matrices for jets with $R = 0.8$. The matrix entries indicate the probability for an event to migrate from the respective generator-level bin on the x-axis to the detector-level bin indicated on the y-axis. The most significant portion of migrations is observed between immediately neighboring bins in $\langle p_T \rangle_{1,2}$ (m_{jj}) and, to a lesser extent, in y^* and y_b.

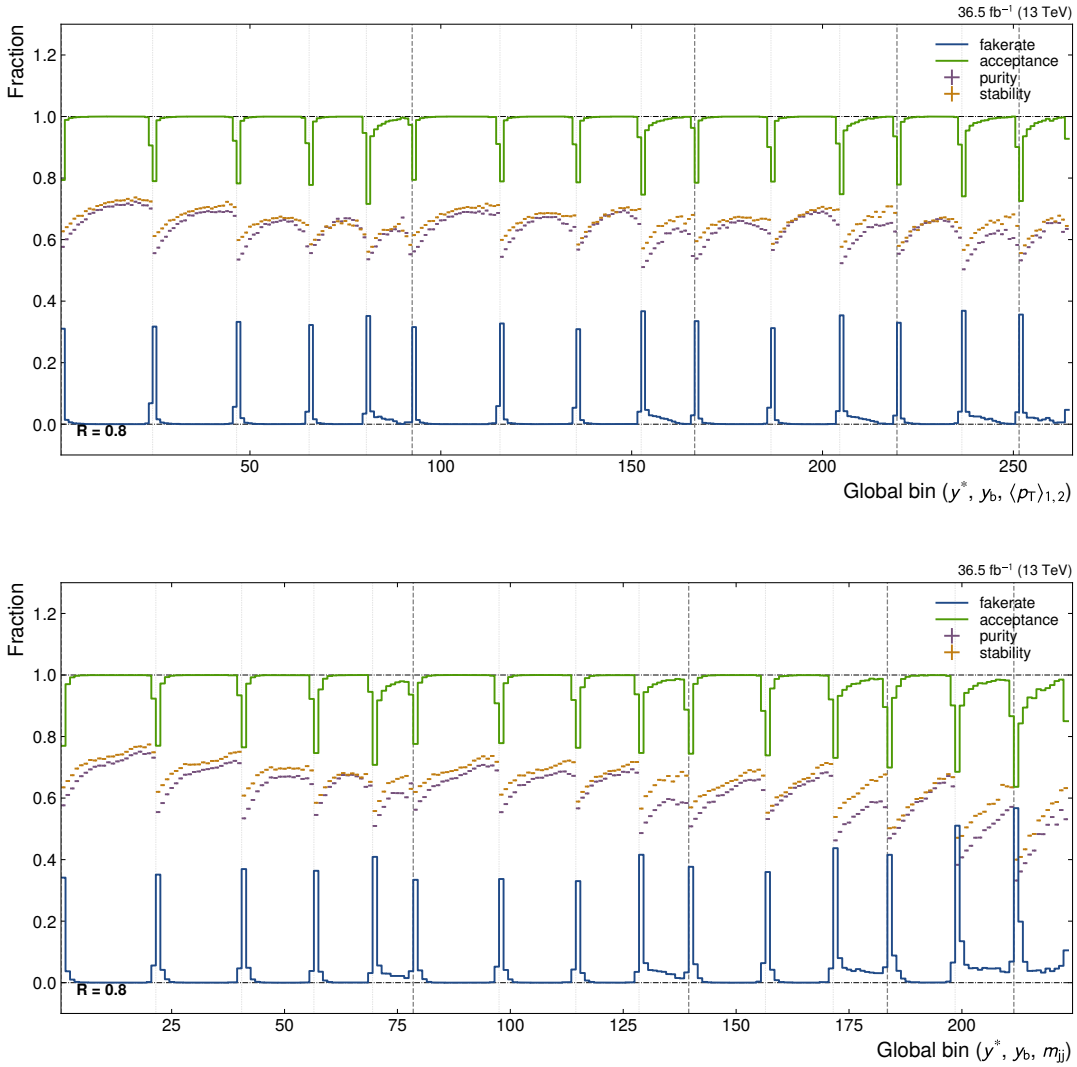

Figure A.8 – Characteristics extracted from the response matrix for jets with $R = 0.8$. The x-axis shows the bin index in the unraveled phase space. The purity and proportion of "fake" reconstructions are shown as function of the detector-level bin index, while the stability and proportion of accepted events are shown as a function of the generator-level bin index. The thick and thin vertical dashed lines indicate the transition points between the different y^* and y_b regions, respectively. The fraction of "fake" and "lost" events increases in bins close to the overall phase space boundary.

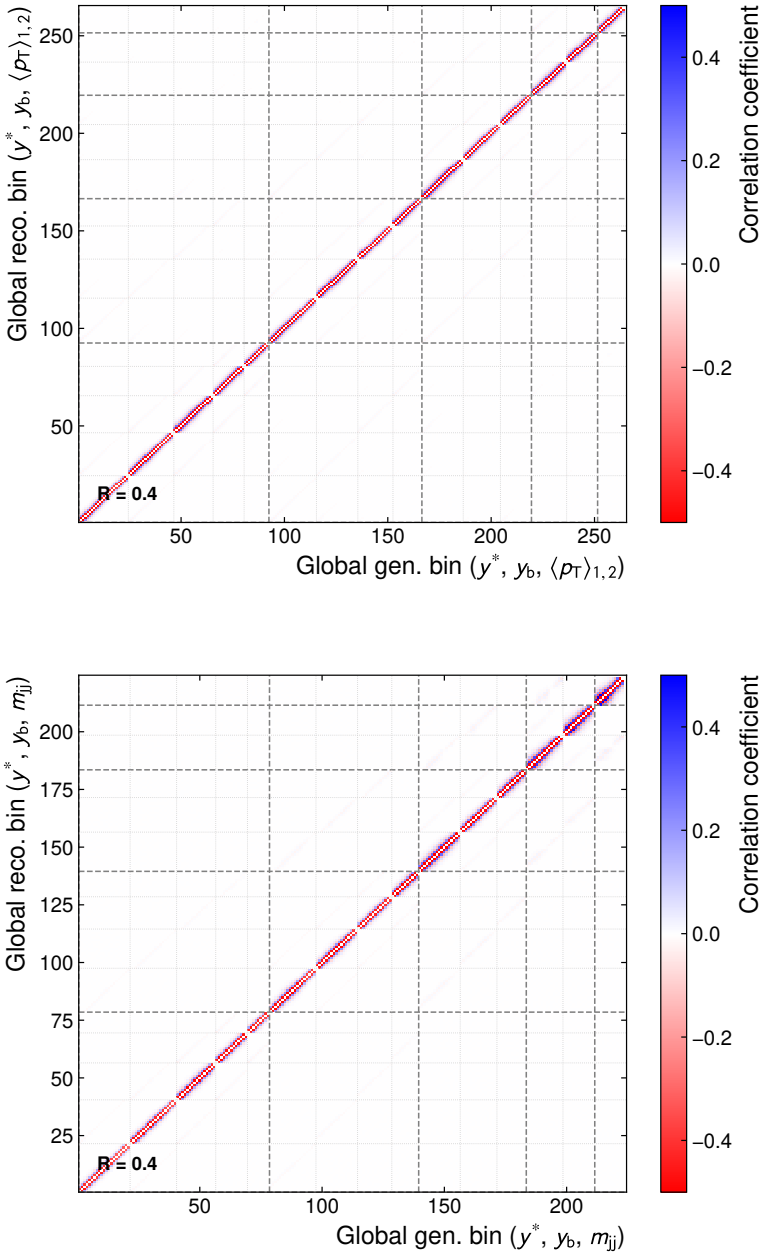

Figure A.9 – Correlation matrix of the statistical uncertainty on the unfolded cross sections for jets with $R = 0.4$. The strongest correlations appear between immediately neighboring bins, where negative coefficients of the order of -50% are observed, while bins separated by one intermediate bin exhibit a weaker, positive correlation. Off-diagonal correlations between different (y^*, y_b) regions are negligible. The correlation coefficients on the diagonal are equal to unity by definition and are not explicitly displayed.

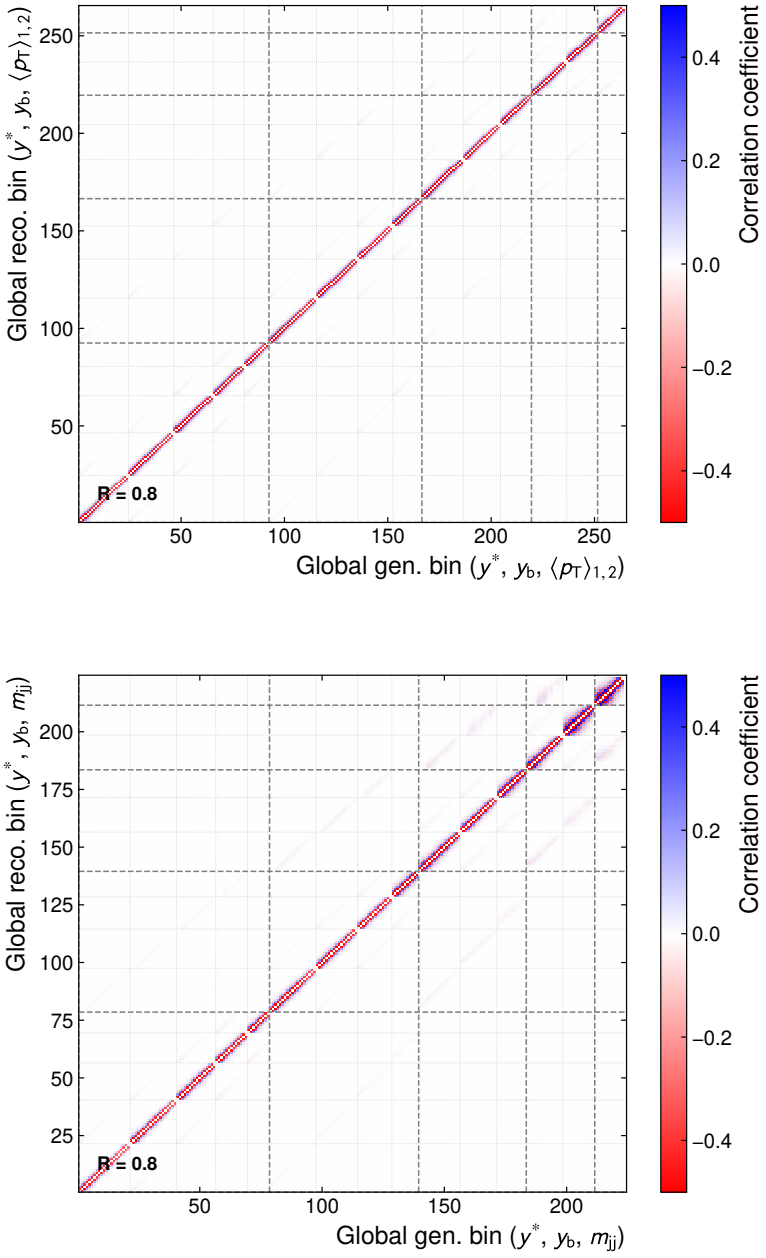

Figure A.10 – Correlation matrix of the statistical uncertainty on the unfolded cross sections for jets with $R = 0.8$. The strongest correlations appear between immediately neighboring bins, where negative coefficients of the order of -50% are observed, while bins separated by one intermediate bin exhibit a weaker, positive correlation. Off-diagonal correlations between different (y^*, y_b) regions are negligible. The correlation coefficients on the diagonal are equal to unity by definition and are not explicitly displayed.

A.4 Cross section comparison to fixed-order theory

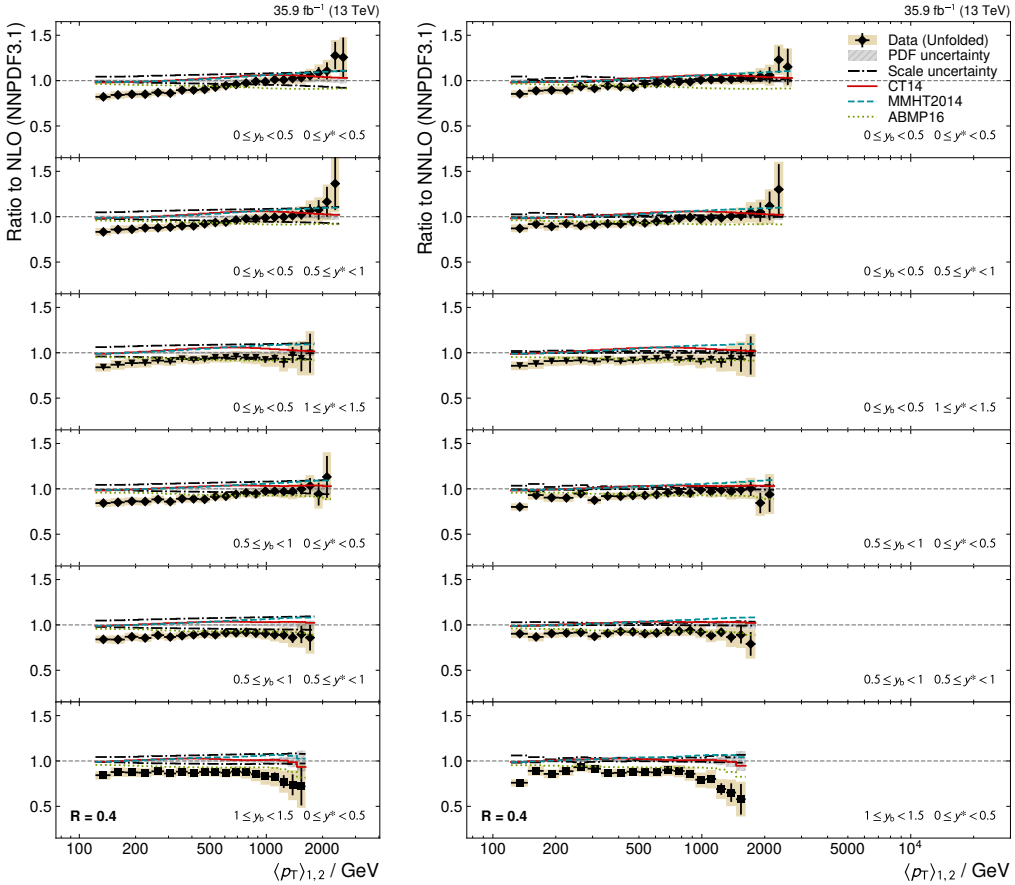

Figure A.11 – Ratio of unfolded cross sections measured in data to NLO (*left*) and NNLO (*right*) theory calculations as a function of (y^*, y_b, $\langle p_T \rangle_{1,2}$) for jets with $R = 0.4$. Shown are 6 out of 15 rapidity bins at central rapidities. The remaining 9 rapidity regions are shown in figure A.12.

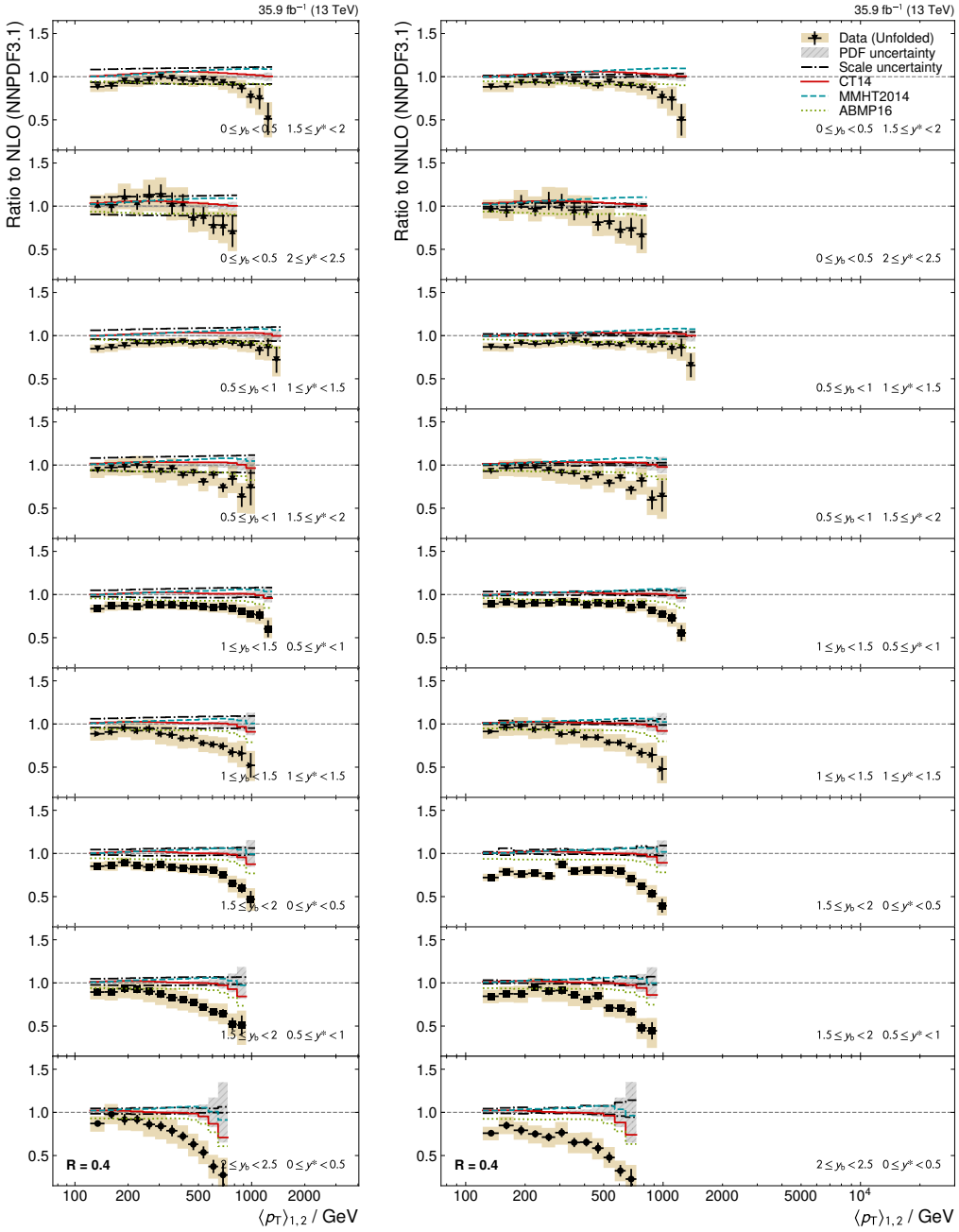

Figure A.12 – (*continued from figure A.11*) Ratio of unfolded cross sections measured in data to NLO (*left*) and NNLO (*right*) theory calculations as a function of (y^*, y_b, $\langle p_T \rangle_{1,2}$) for jets with $R = 0.4$.

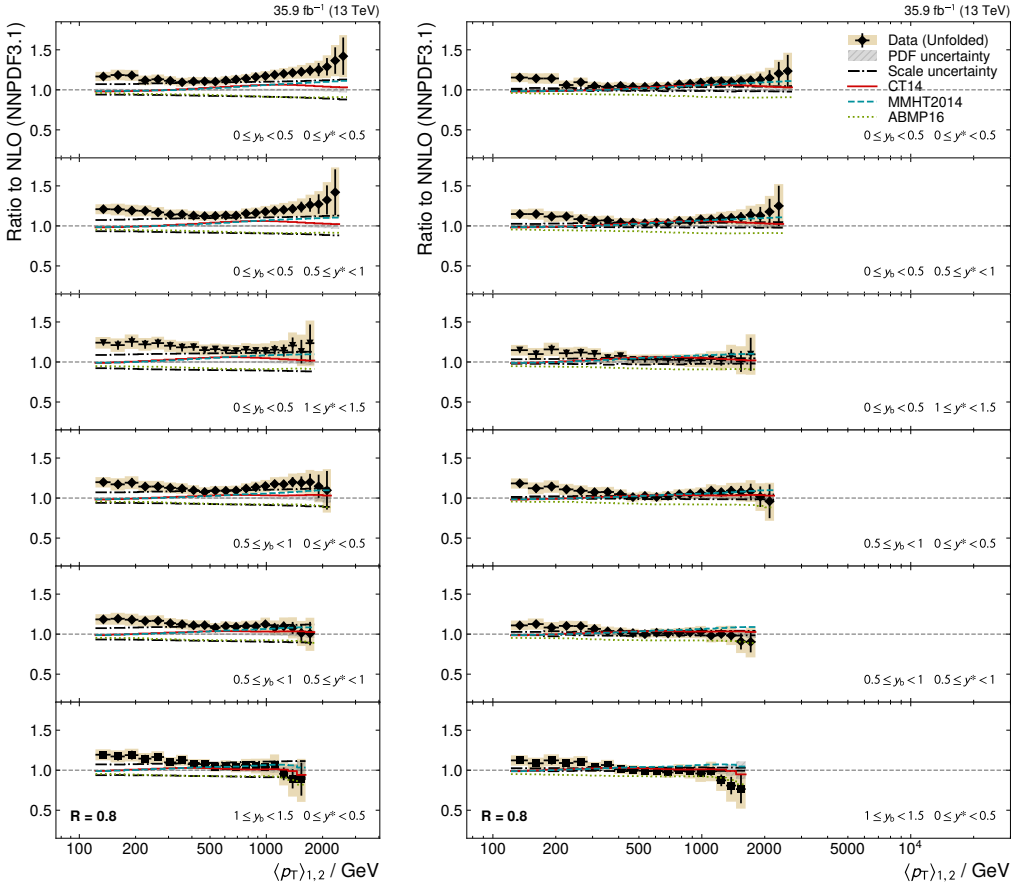

Figure A.13 – Ratio of unfolded cross sections measured in data to NLO (*left*) and NNLO (*right*) theory calculations as a function of (y^*, y_b, $\langle p_T \rangle_{1,2}$) for jets with $R = 0.8$. Shown are 6 out of 15 rapidity bins at central rapidities. The remaining 9 rapidity regions are shown in figure A.14.

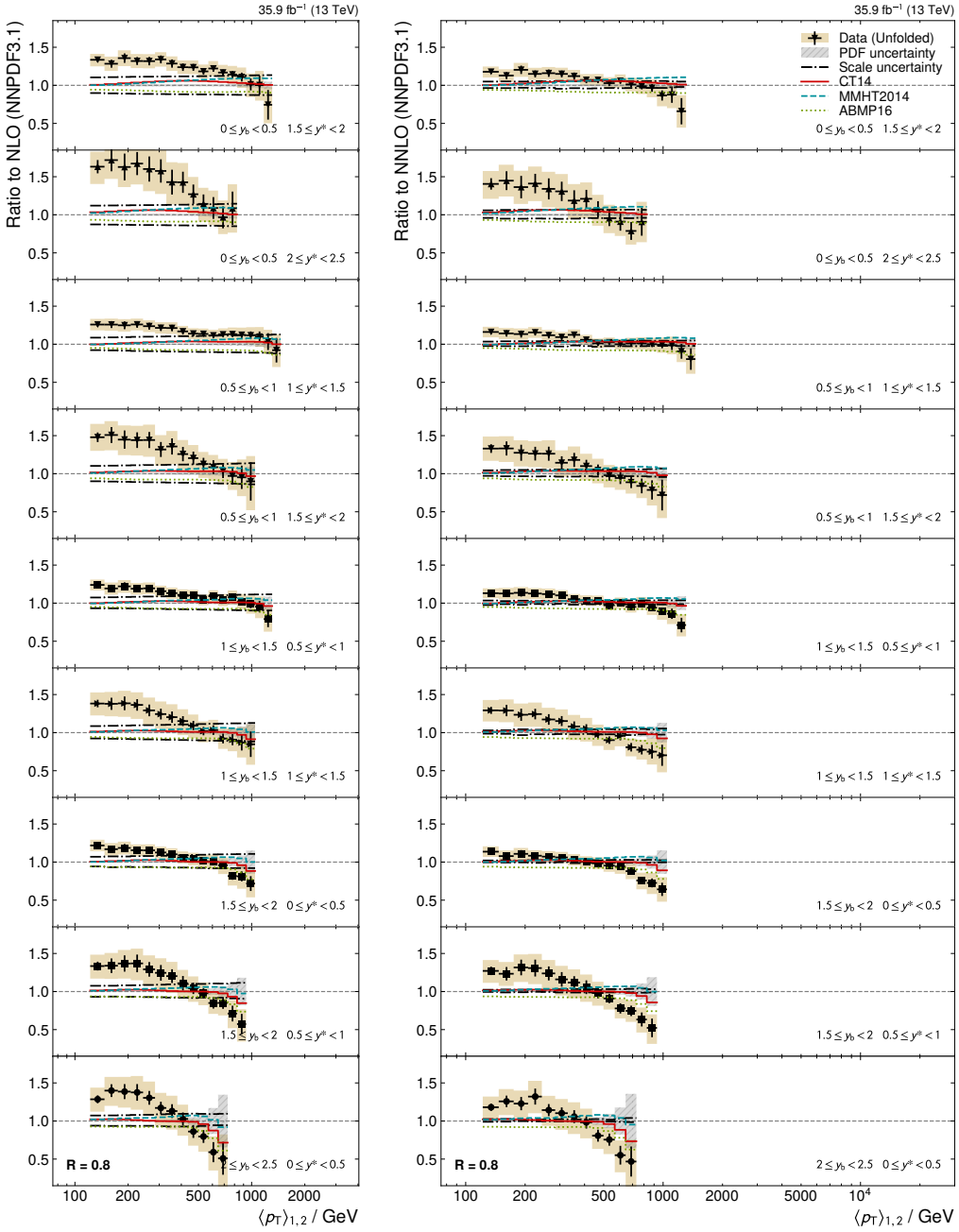

Figure A.14 – (*continued from figure A.13*) Ratio of unfolded cross sections measured in data to NLO (*left*) and NNLO (*right*) theory calculations as a function of (y^*, y_b, $\langle p_T \rangle_{1,2}$) for jets with $R = 0.8$.

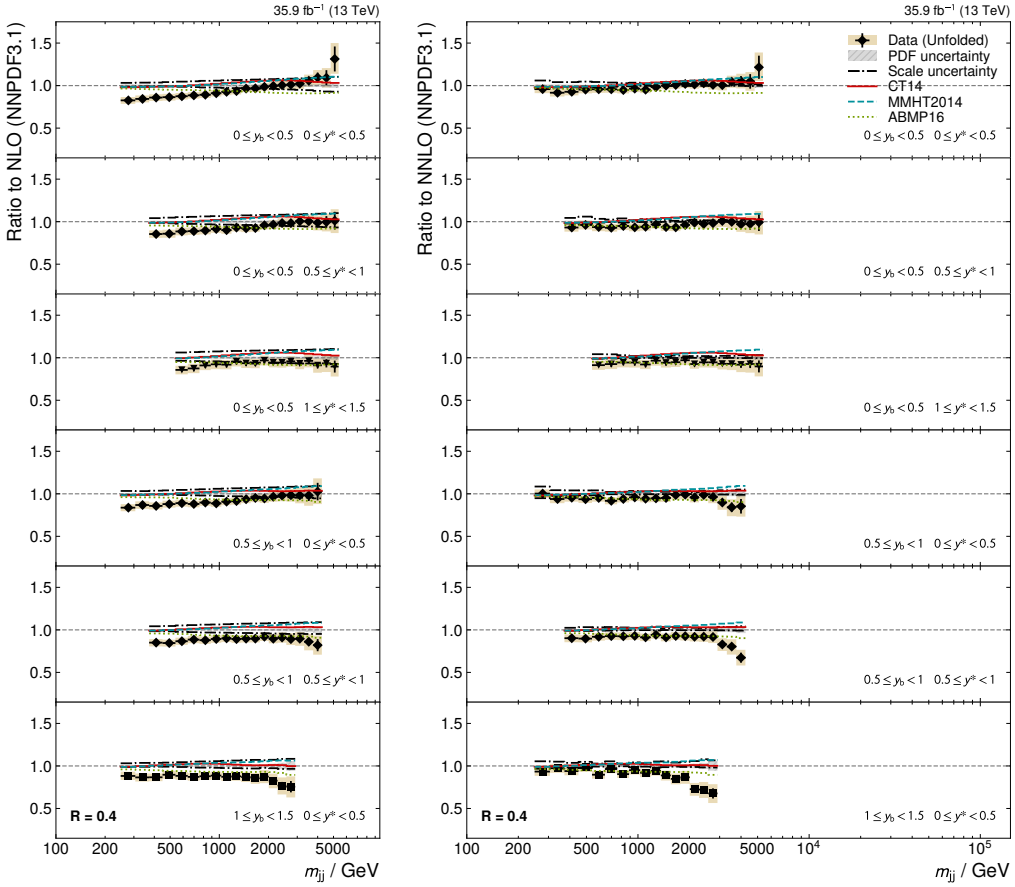

Figure A.15 – Ratio of unfolded cross sections measured in data to NLO (*left*) and NNLO (*right*) theory calculations as a function of (y^*, y_b, m_{jj}) for jets with $R = 0.4$. Shown are 6 out of 15 rapidity bins at central rapidities. The remaining 9 rapidity regions are shown in figure A.16.

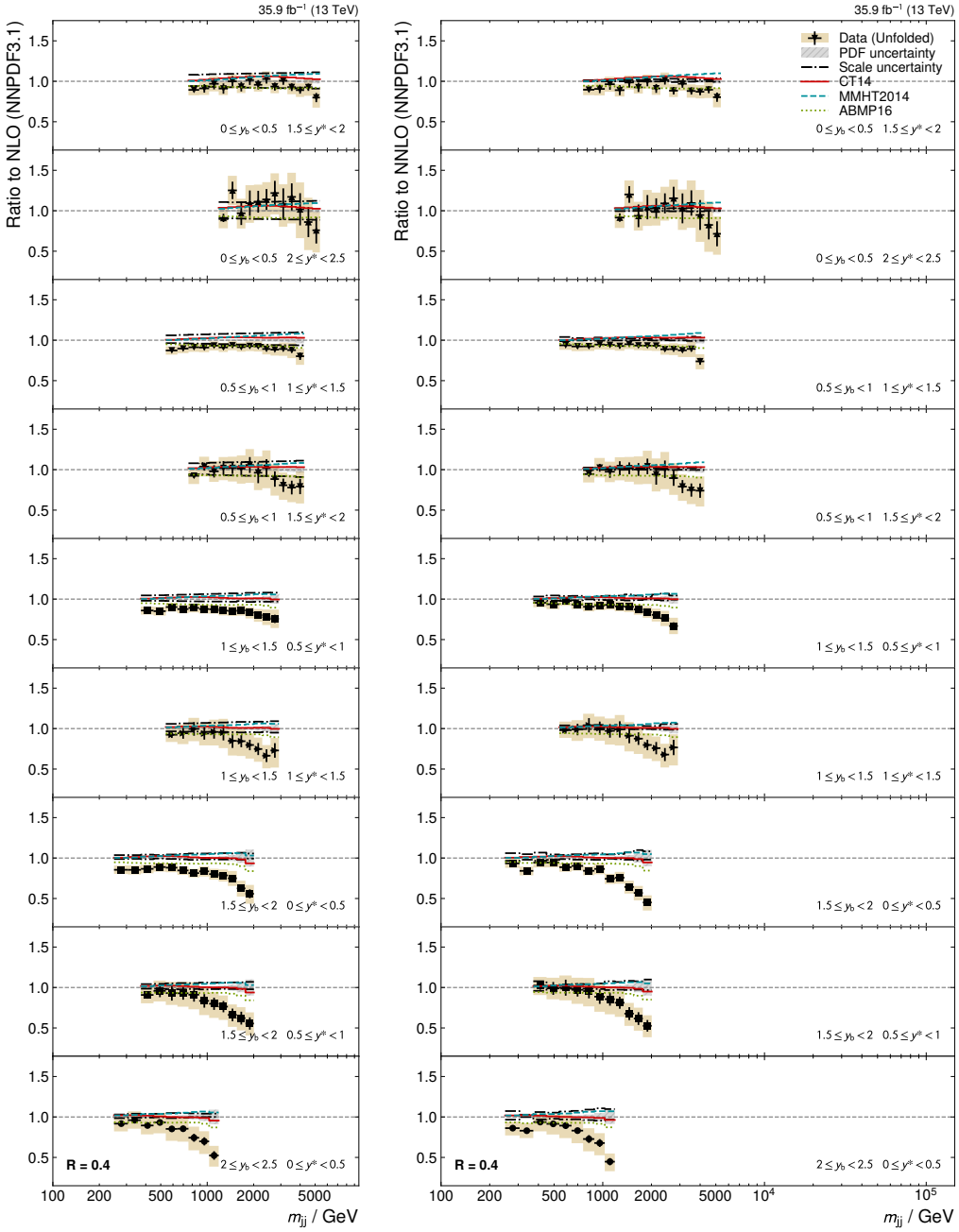

Figure A.16 – (*continued from figure A.15*) Ratio of unfolded cross sections measured in data to NLO (*left*) and NNLO (*right*) theory calculations as a function of (y^*, y_b, m_{jj}) for jets with $R = 0.4$.

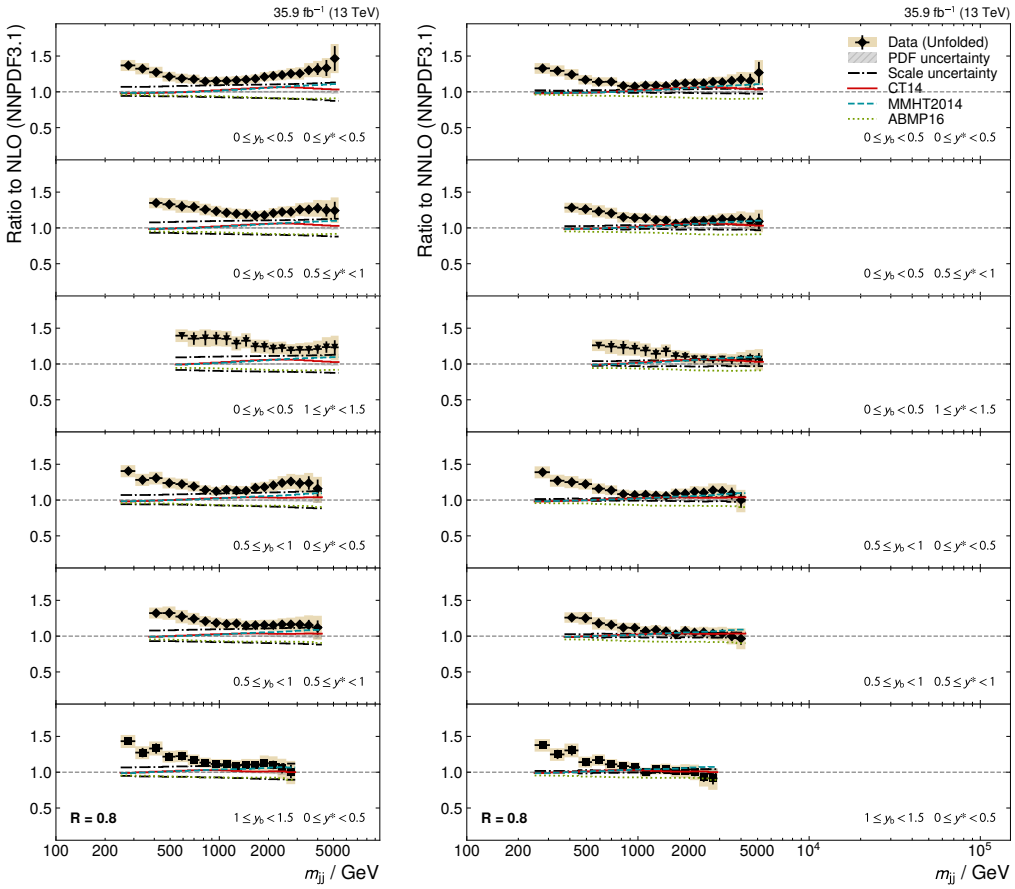

Figure A.17 – Ratio of unfolded cross sections measured in data to NLO (*left*) and NNLO (*right*) theory calculations as a function of (y^*, y_b, m_{jj}) for jets with $R = 0.8$. Shown are 6 out of 15 rapidity bins at central rapidities. The remaining 9 rapidity regions are shown in figure A.18.

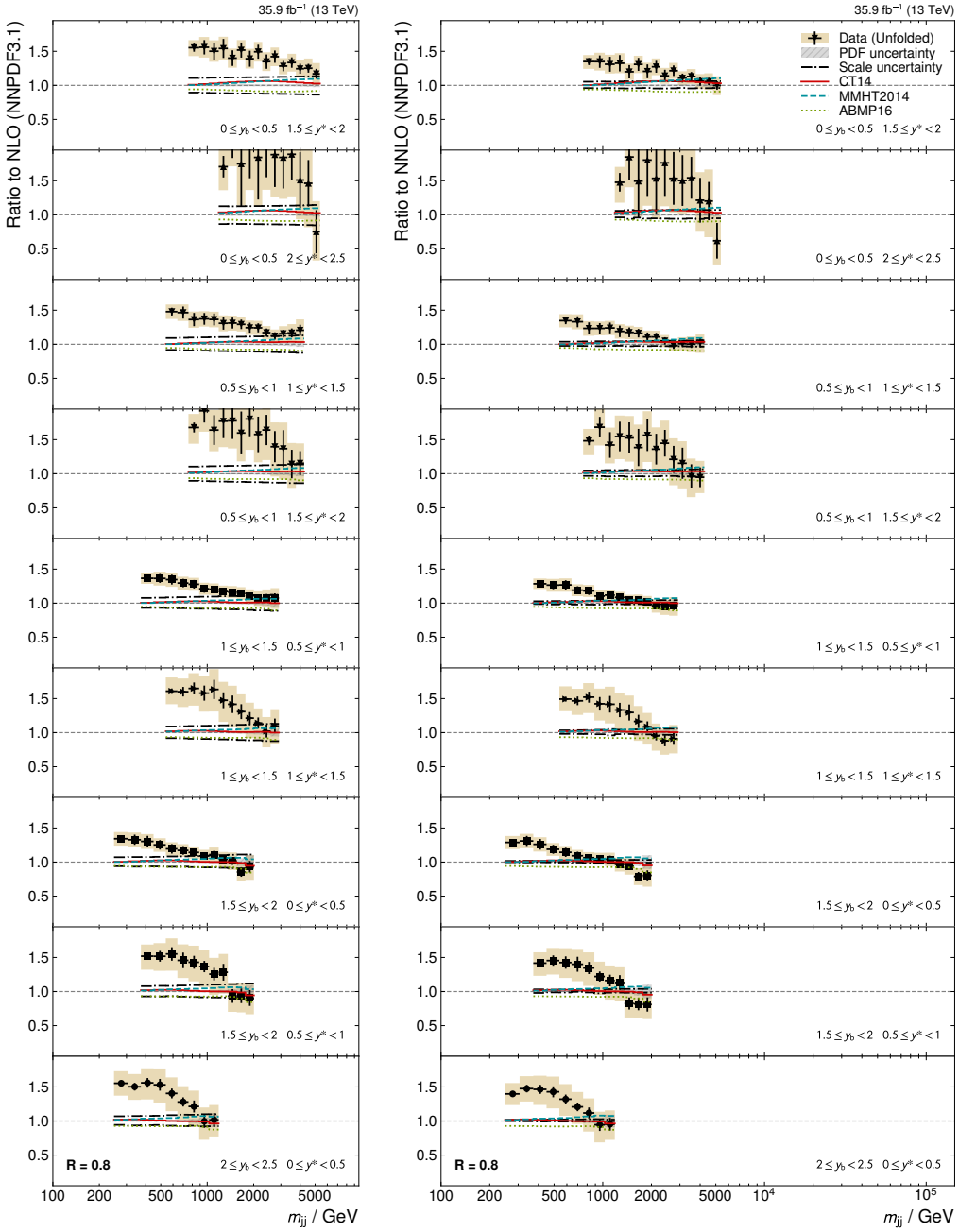

Figure A.18 – (*continued from figure A.17*) Ratio of unfolded cross sections measured in data to NLO (*left*) and NNLO (*right*) theory calculations as a function of (y^*, y_b, m_{jj}) for jets with $R = 0.8$.

A.5 Variable distributions

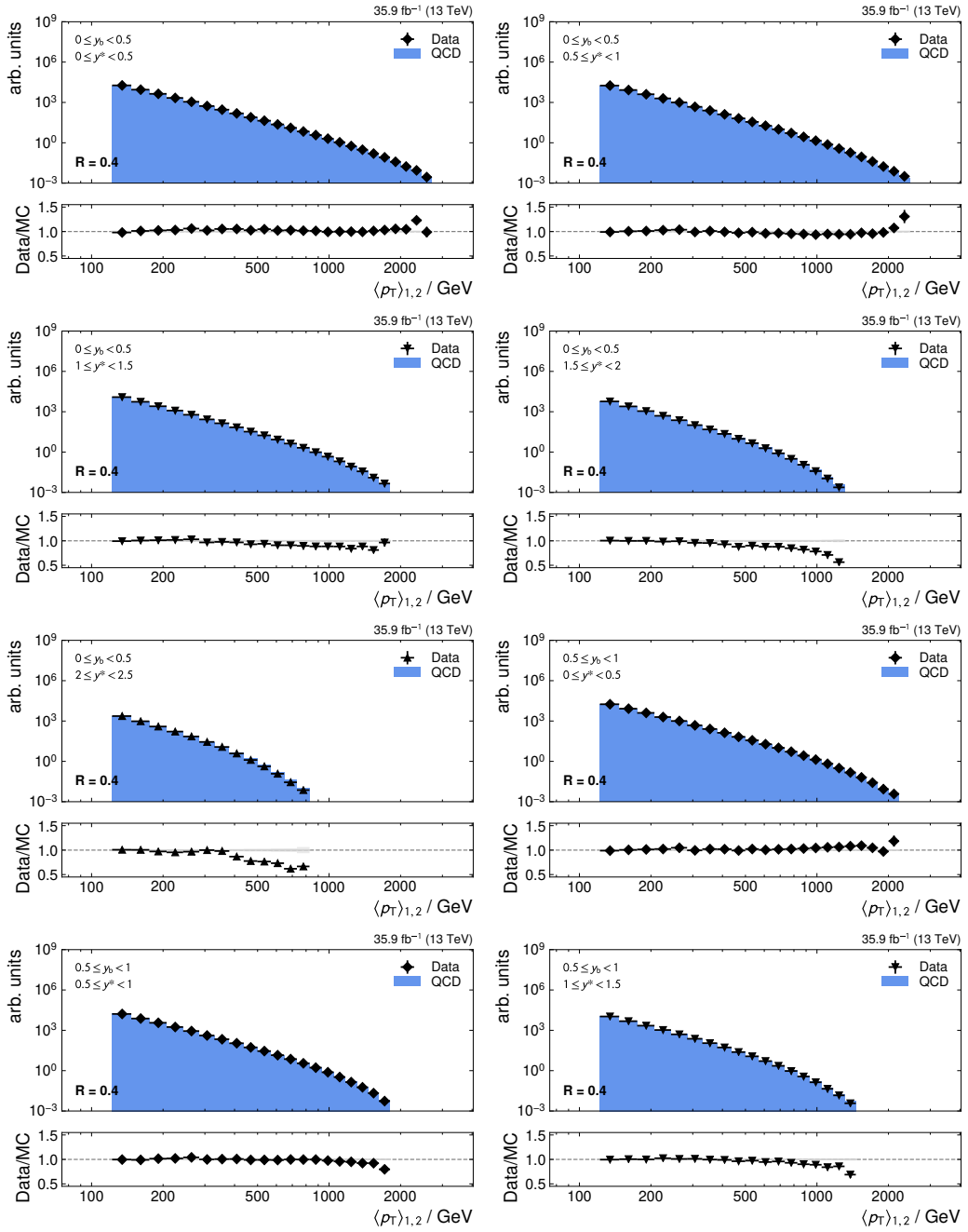

Figure A.19 – Comparison of the distribution of the average transverse momentum of the two leading jets $\langle p_T \rangle_{1,2}$ in data and MC simulation, shown for jets with $R = 0.4$ in 8 out of 15 rapidity regions. The remaining ones are shown in figure A.20.

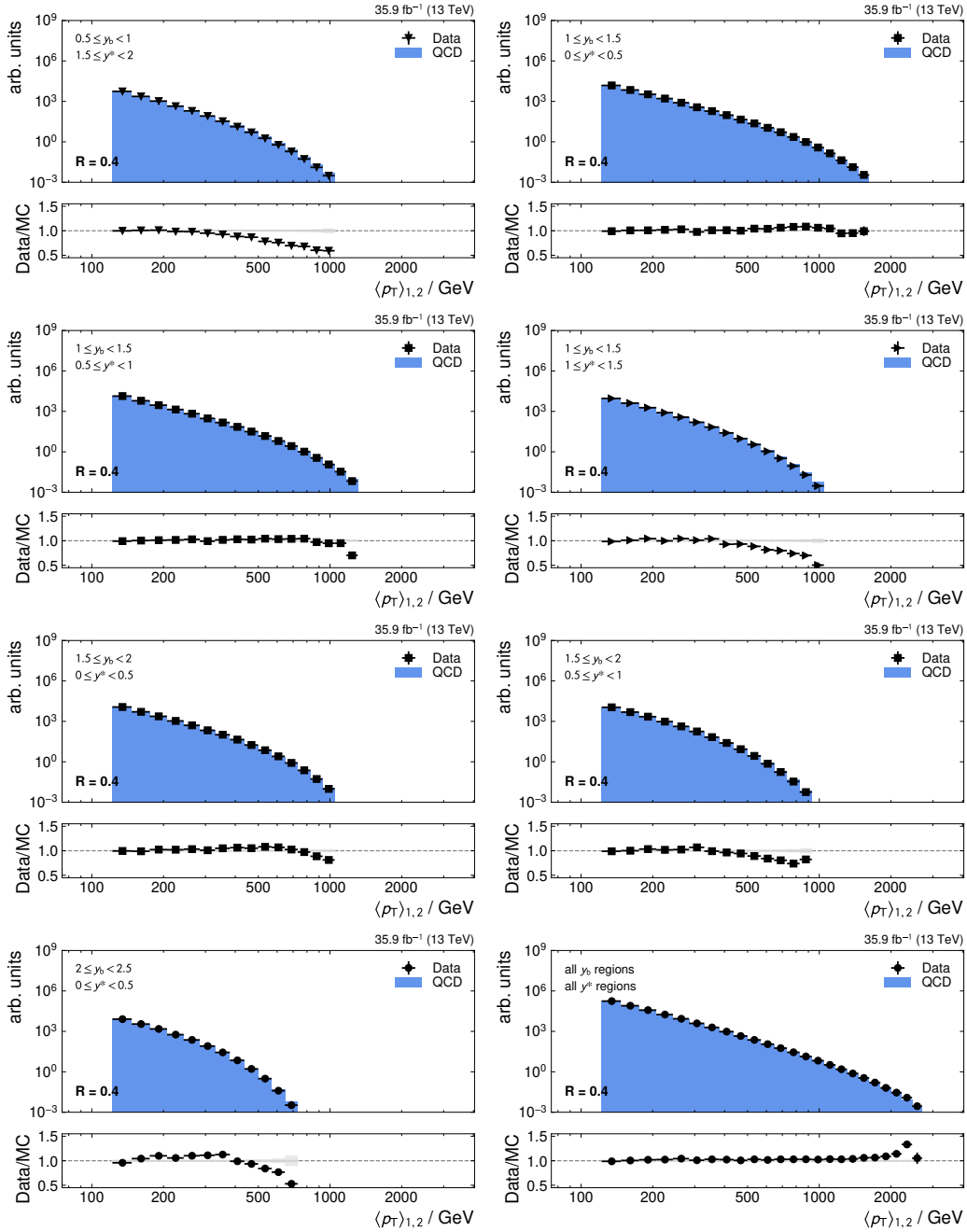

Figure A.20 – (*continued from figure A.19*) Comparison of the distribution of the average transverse momentum of the two leading jets $\langle p_T \rangle_{1,2}$ in data and MC simulation, shown for jets with $R = 0.4$ in 7 out of 15 rapidity regions, as well as for the combination of all rapidity regions (*bottom right*).

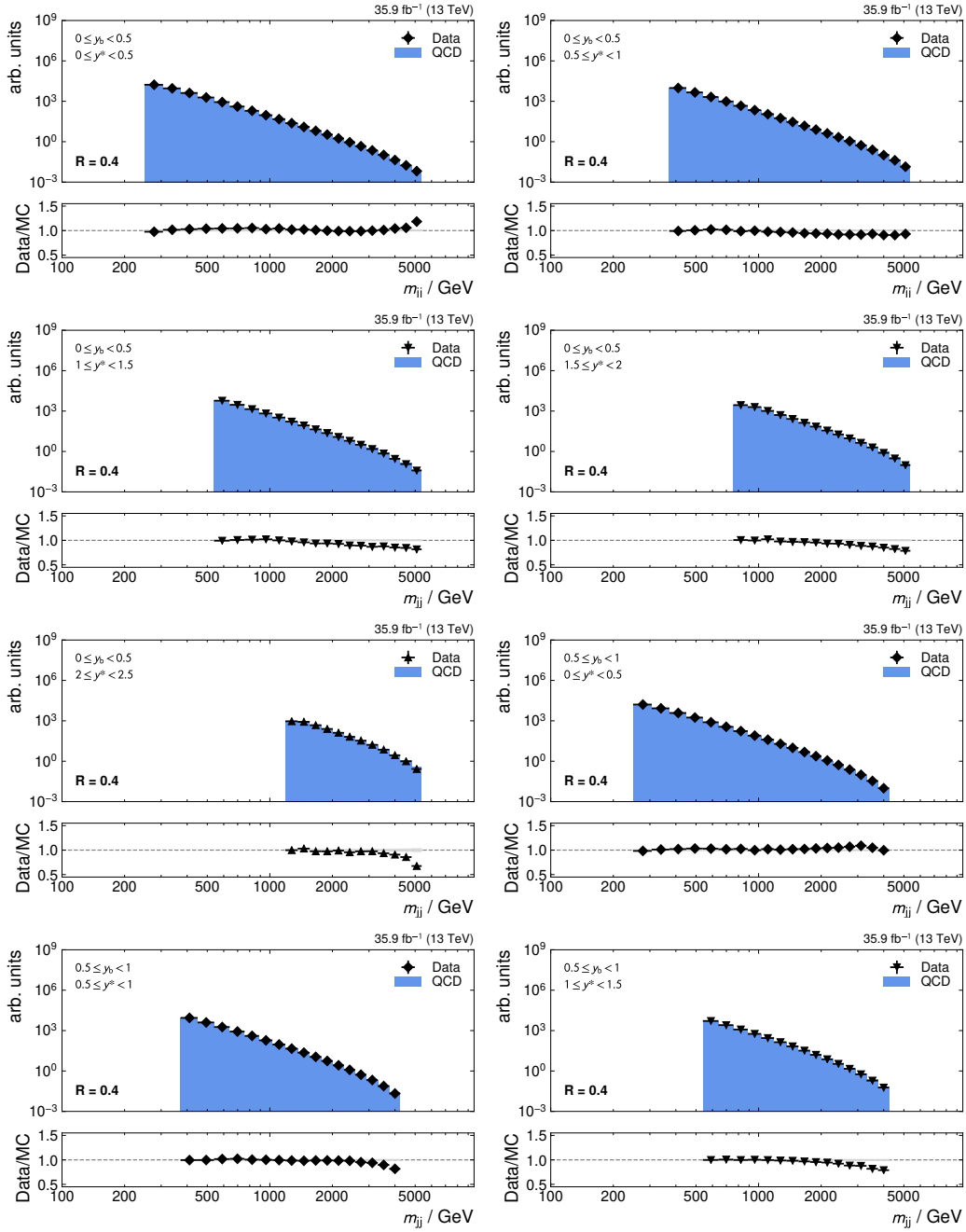

Figure A.21 – Comparison of the distribution of the invariant mass of the two leading jets m_{jj} in data and MC simulation, shown for jets with $R = 0.4$ in 8 out of 15 rapidity regions. The remaining ones are shown in figure A.22.

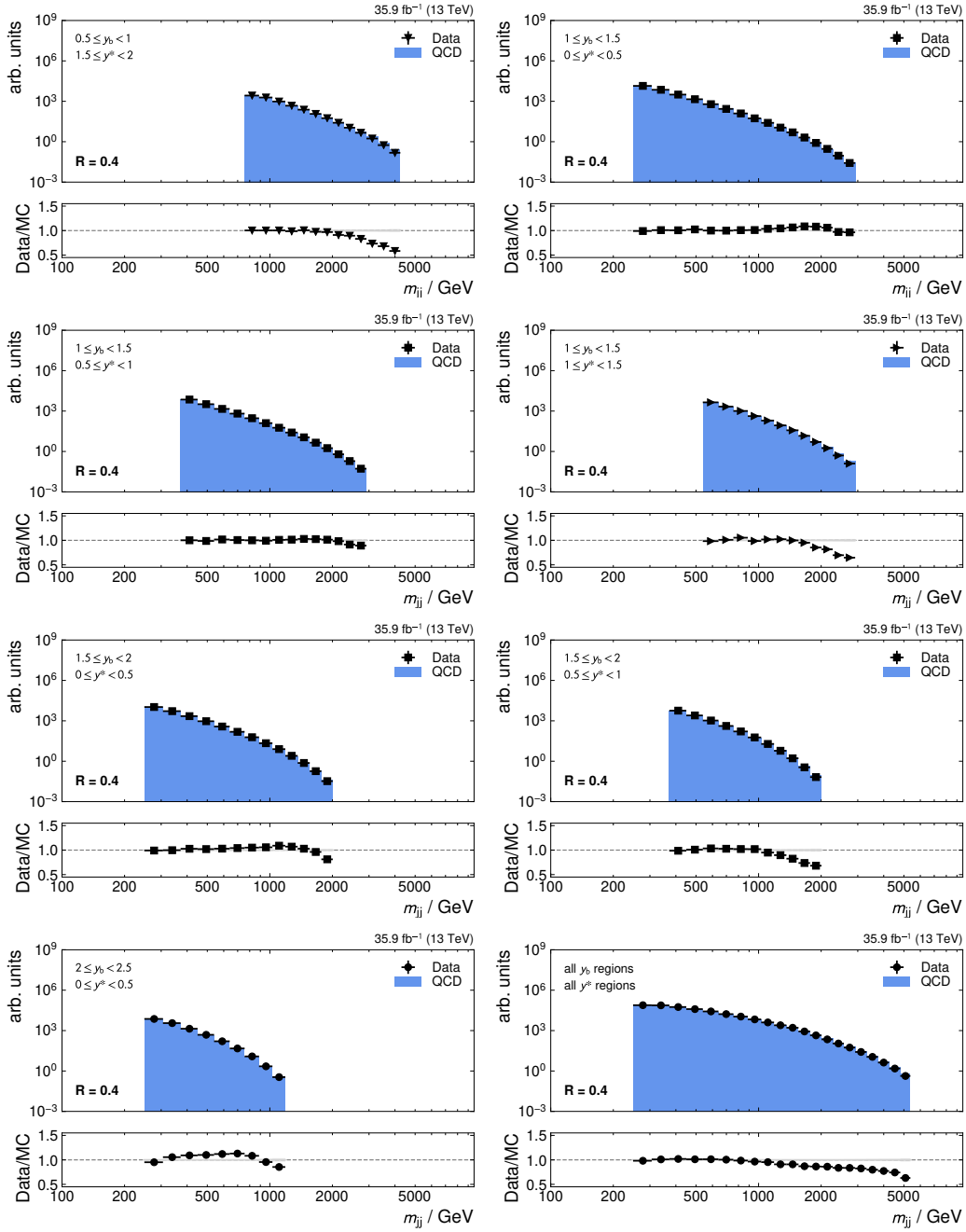

Figure A.22 – (*continued from figure A.21*) Comparison of the distribution of the invariant mass of the two leading jets m_{jj} in data and MC simulation, shown for jets with $R = 0.4$ in 7 out of 15 rapidity regions, as well as for the combination of all rapidity regions (*bottom right*).

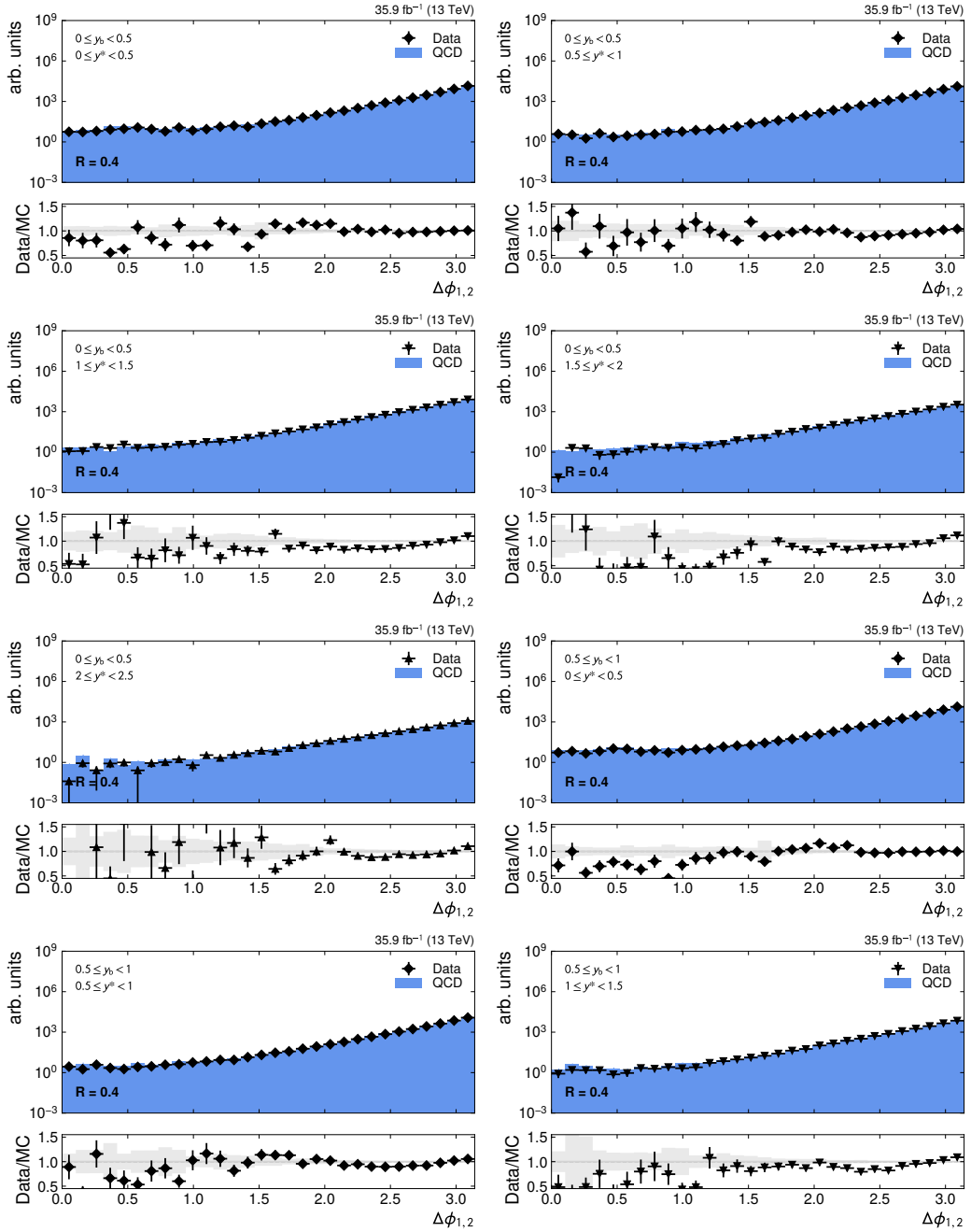

Figure A.23 – Comparison of the distribution of the dijet angle separation in azimuthal plane η–ϕ plane $\Delta\phi_{1,2}$ in data and MC simulation, shown for jets with $R = 0.4$ in 8 out of 15 rapidity regions. The remaining ones are shown in figure A.24.

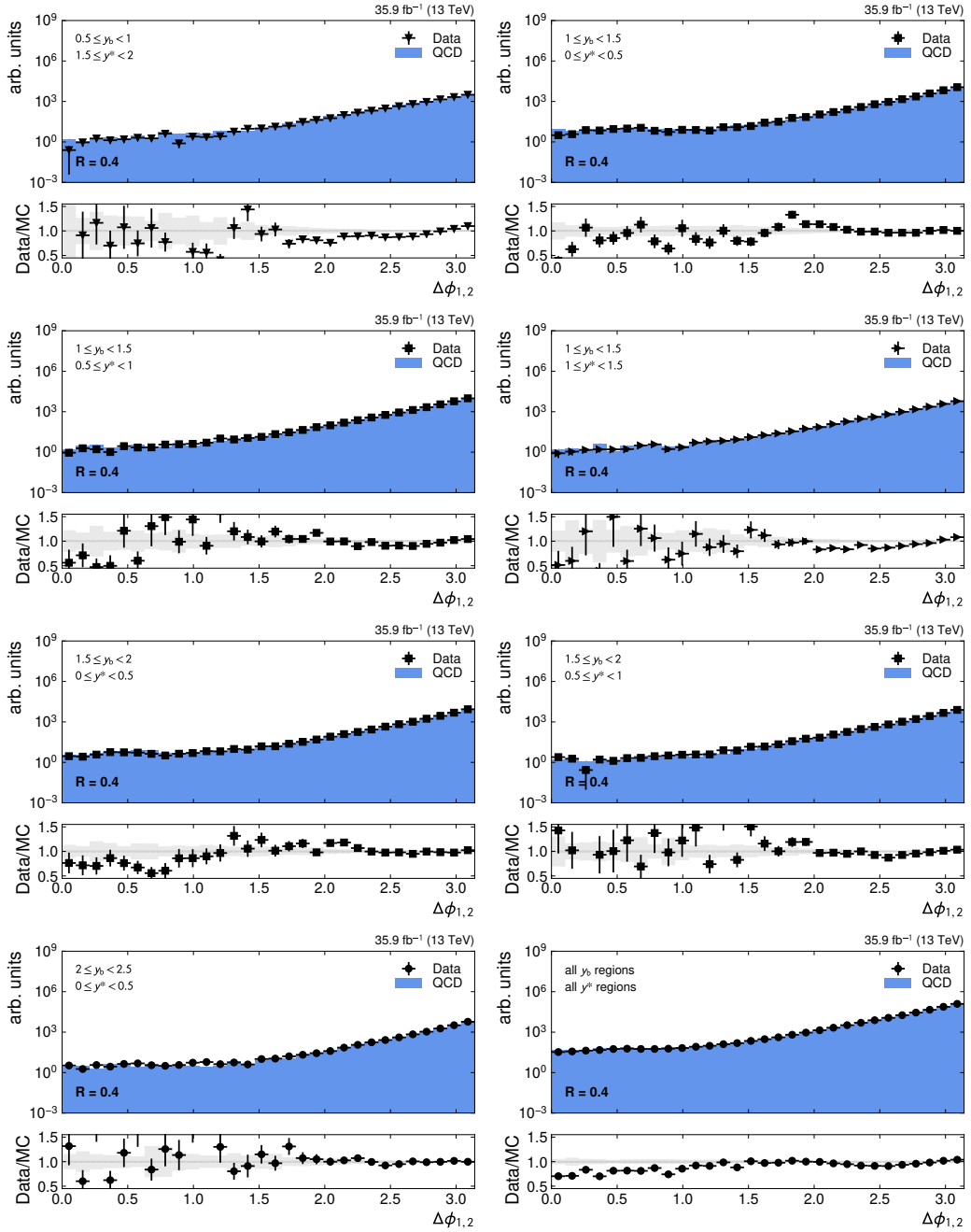

Figure A.24 – (*continued from figure A.23*) Comparison of the distribution of the dijet angle separation in azimuthal plane η–ϕ plane $\Delta\phi_{1,2}$ in data and MC simulation, shown for jets with $R = 0.4$ in 7 out of 15 rapidity regions, as well as for the combination of all rapidity regions (*bottom right*).

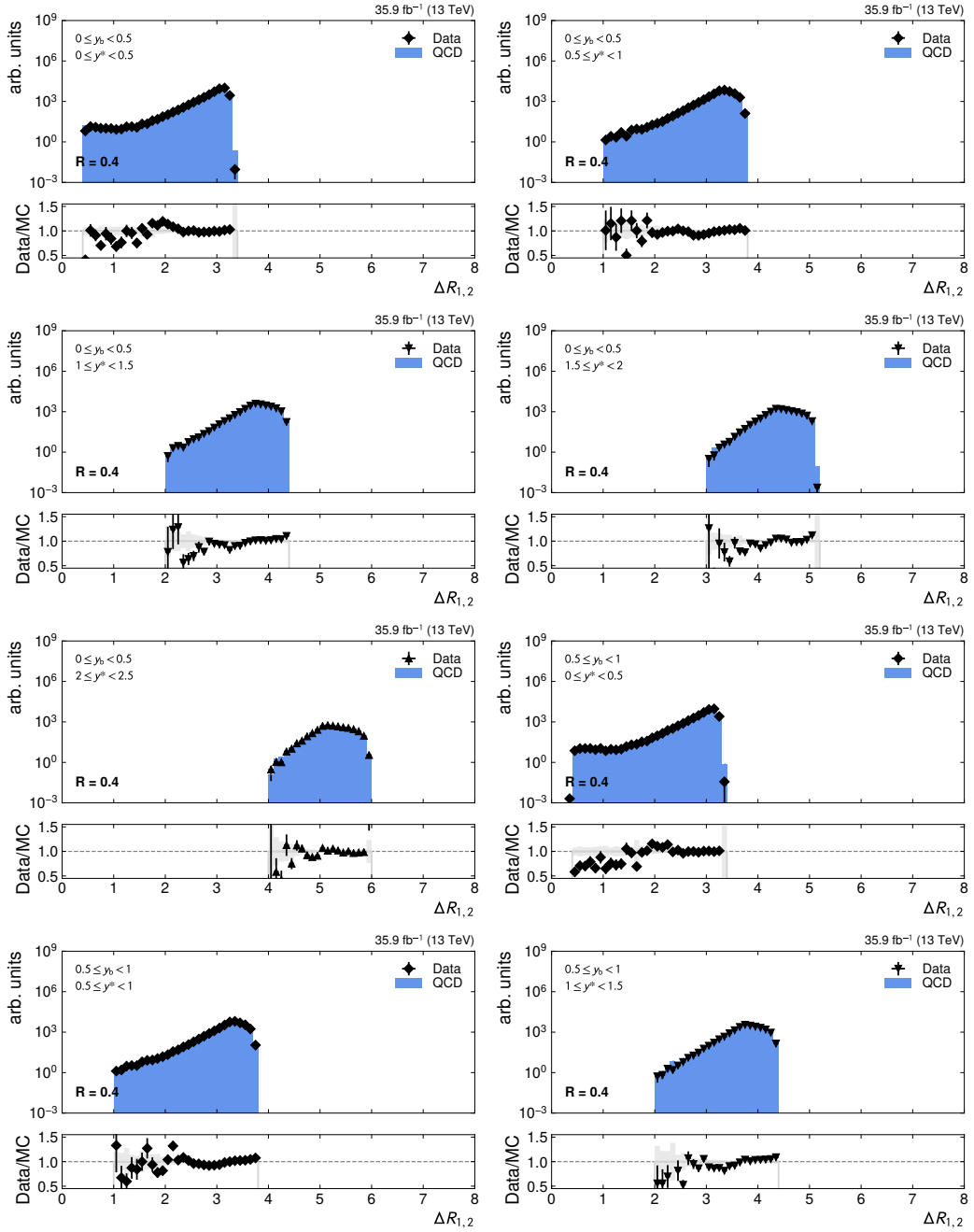

Figure A.25 – Comparison of the distribution of the rapidity of the subleading jet $\Delta R_{1,2}$ in data and MC simulation, shown for jets with $R = 0.4$ in 8 out of 15 rapidity regions. The remaining ones are shown in figure A.26.

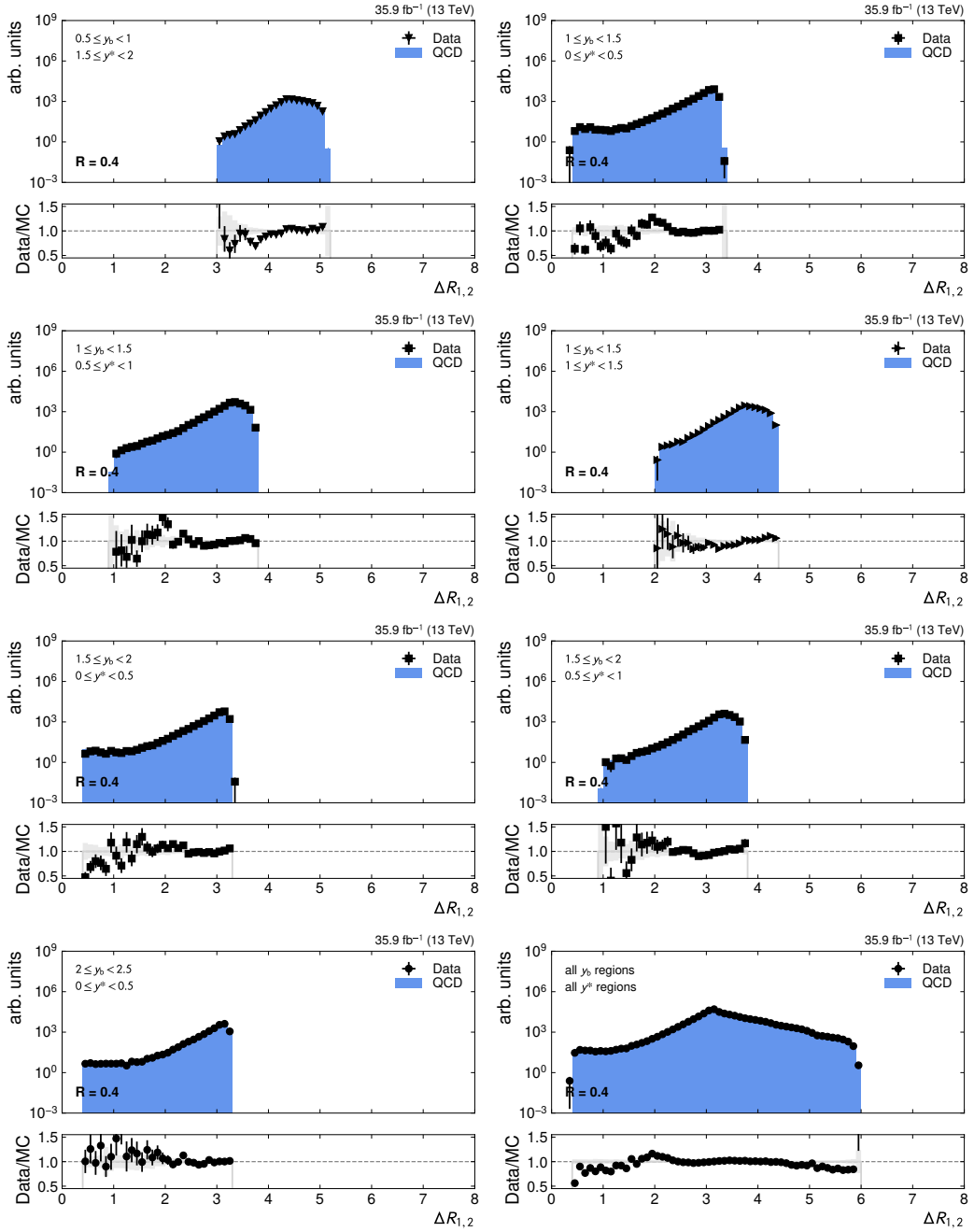

Figure A.26 – (*continued from figure A.25*) Comparison of the distribution of the rapidity of the subleading jet $\Delta R_{1,2}$ in data and MC simulation, shown for jets with $R = 0.4$ in 7 out of 15 rapidity regions, as well as for the combination of all rapidity regions (*bottom right*).

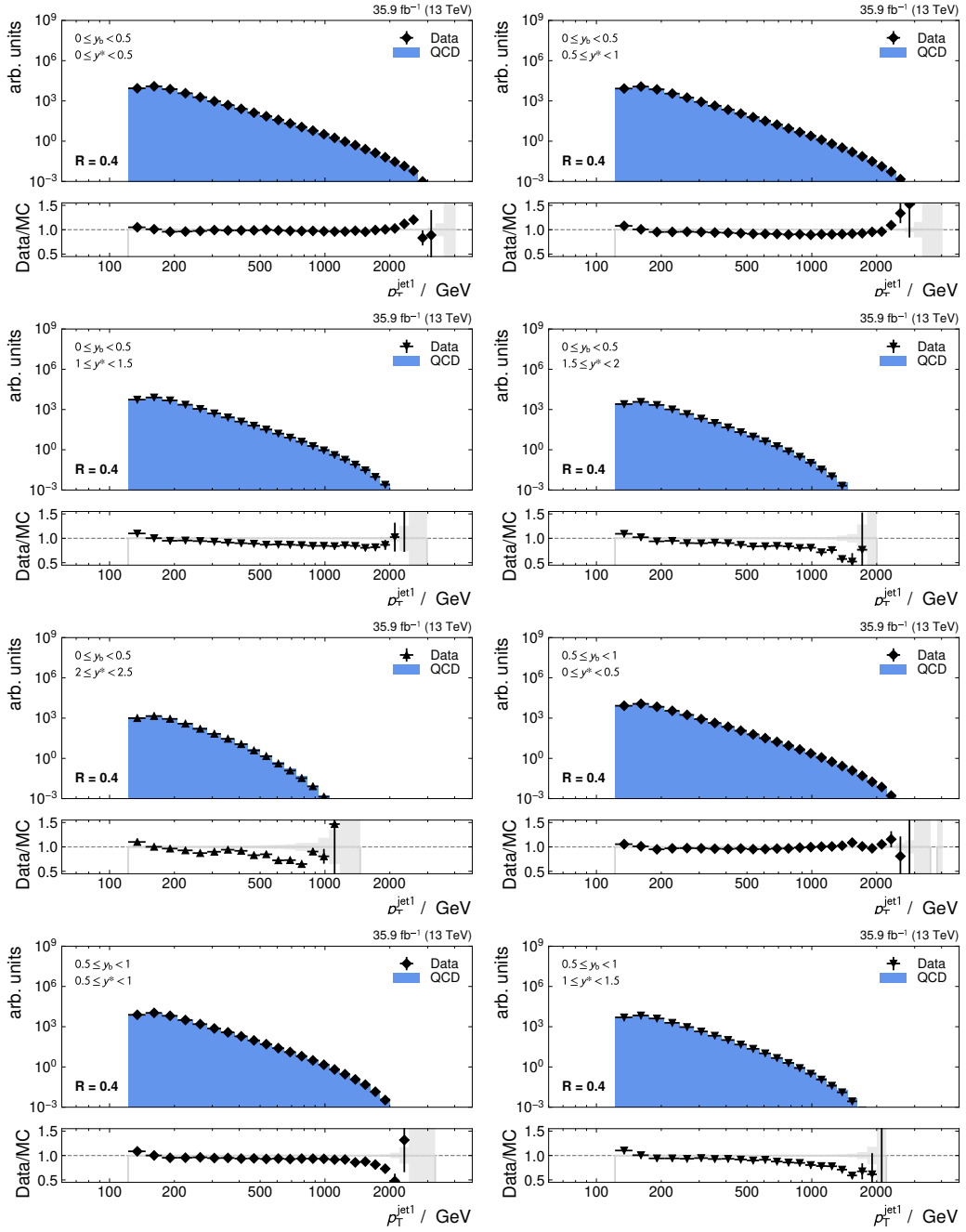

Figure A.27 – Comparison of the distribution of the transverse momentum of the leading jet $p_{\mathrm{T}}^{\mathrm{jet}\,1}$ in data and MC simulation, shown for jets with $R = 0.4$ in 8 out of 15 rapidity regions. The remaining ones are shown in figure A.28.

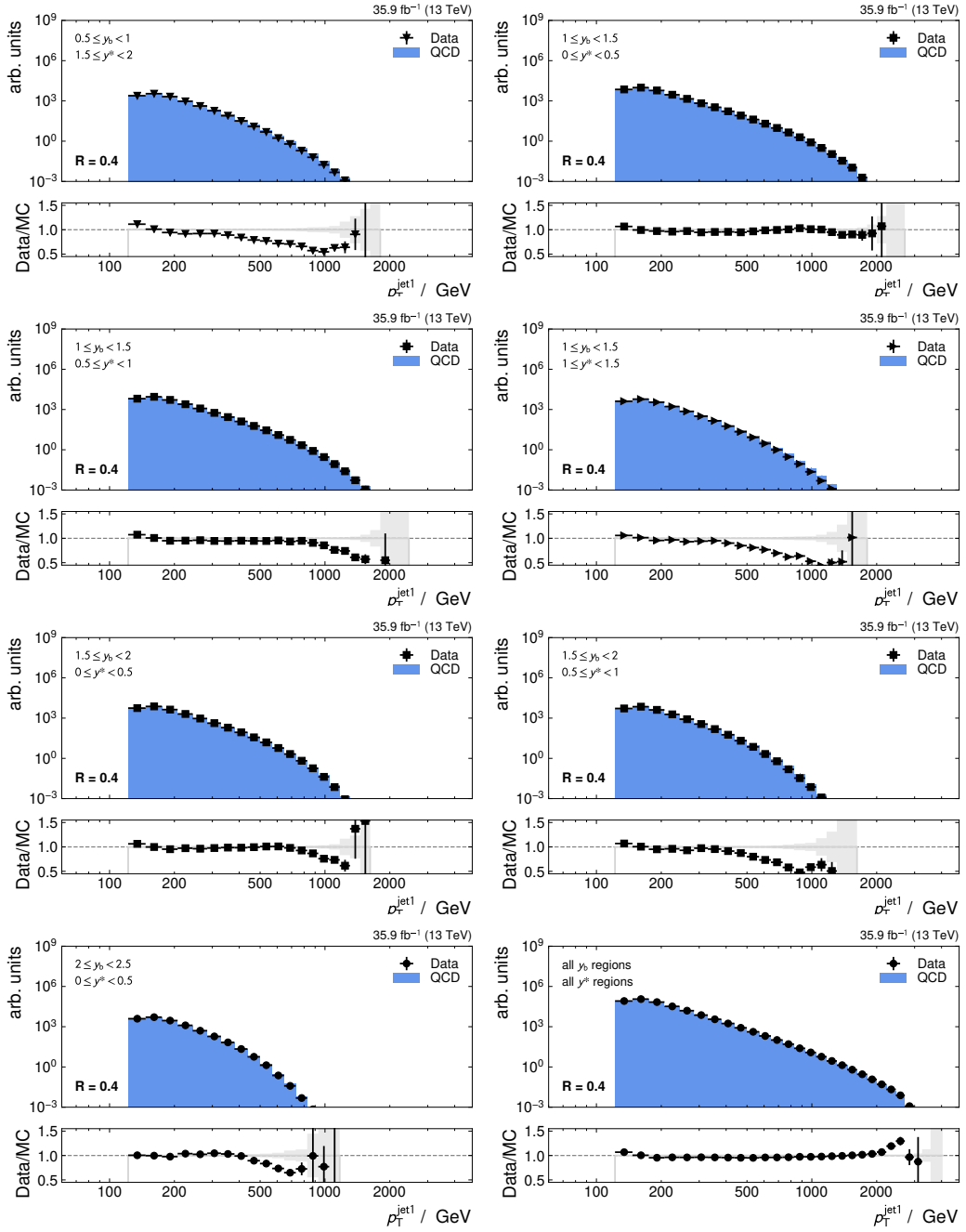

Figure A.28 – (*continued from figure A.27*) Comparison of the distribution of the transverse momentum of the leading jet $p_{\mathrm{T}}^{\mathrm{jet}\,1}$ in data and MC simulation, shown for jets with $R = 0.4$ in 7 out of 15 rapidity regions, as well as for the combination of all rapidity regions (*bottom right*).

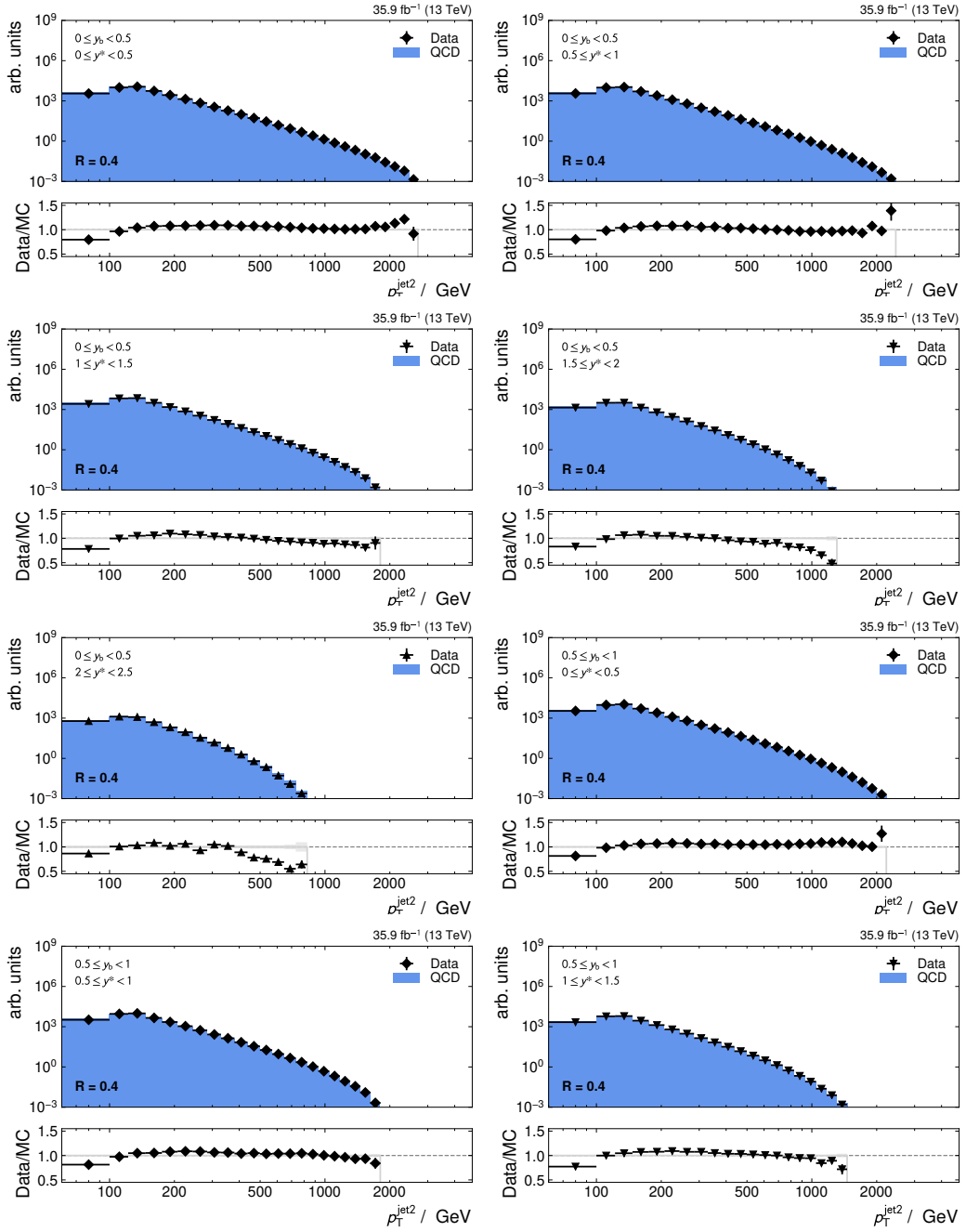

Figure A.29 – Comparison of the distribution of the transverse momentum of the subleading jet $p_{\mathrm{T}}^{\mathrm{jet}\,2}$ in data and MC simulation, shown for jets with $R = 0.4$ in 8 out of 15 rapidity regions. The remaining ones are shown in figure A.30.

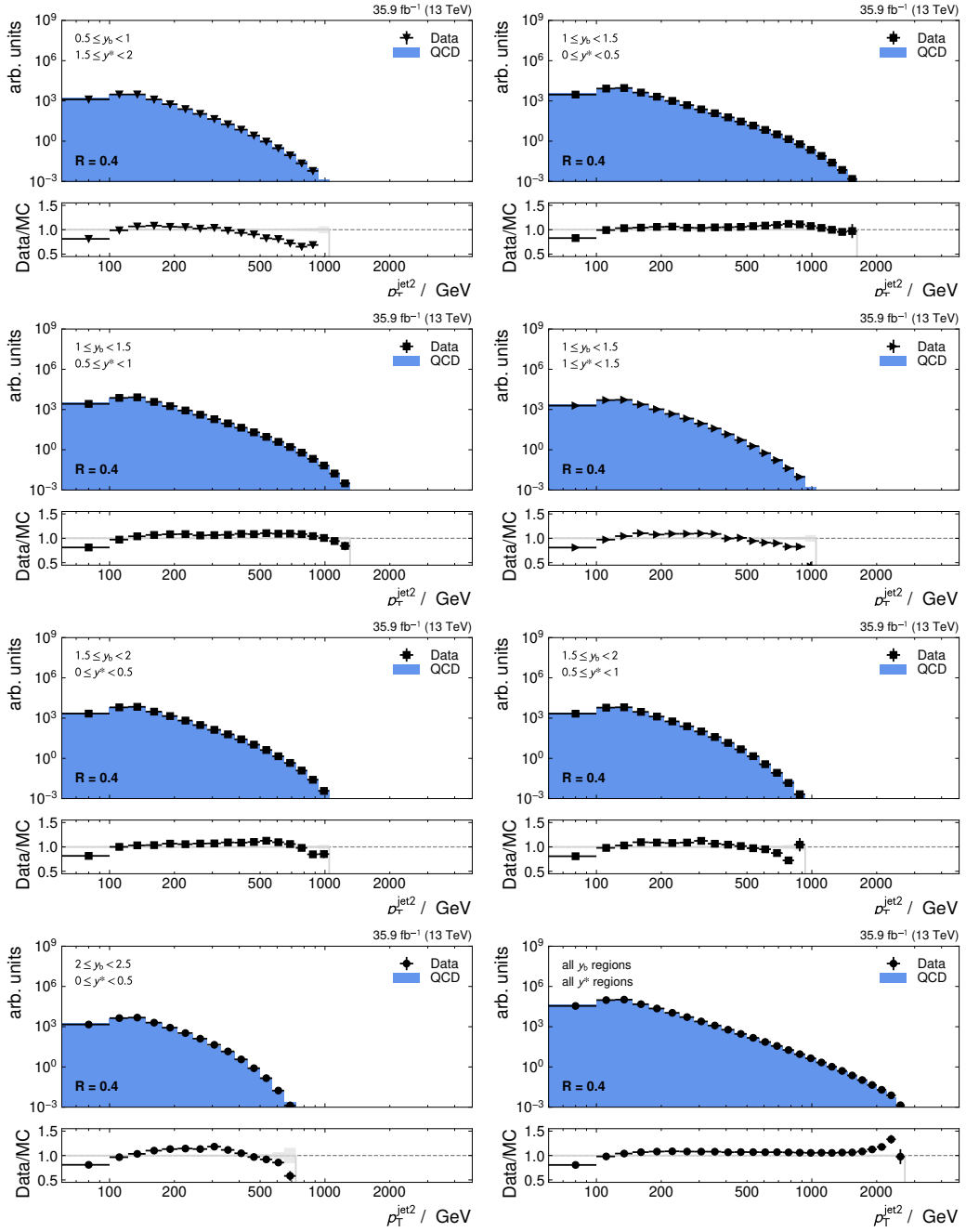

Figure A.30 – (*continued from figure A.29*) Comparison of the distribution of the transverse momentum of the subleading jet $p_{\mathrm{T}}^{\mathrm{jet}\,2}$ in data and MC simulation, shown for jets with $R = 0.4$ in 7 out of 15 rapidity regions, as well as for the combination of all rapidity regions (*bottom right*).

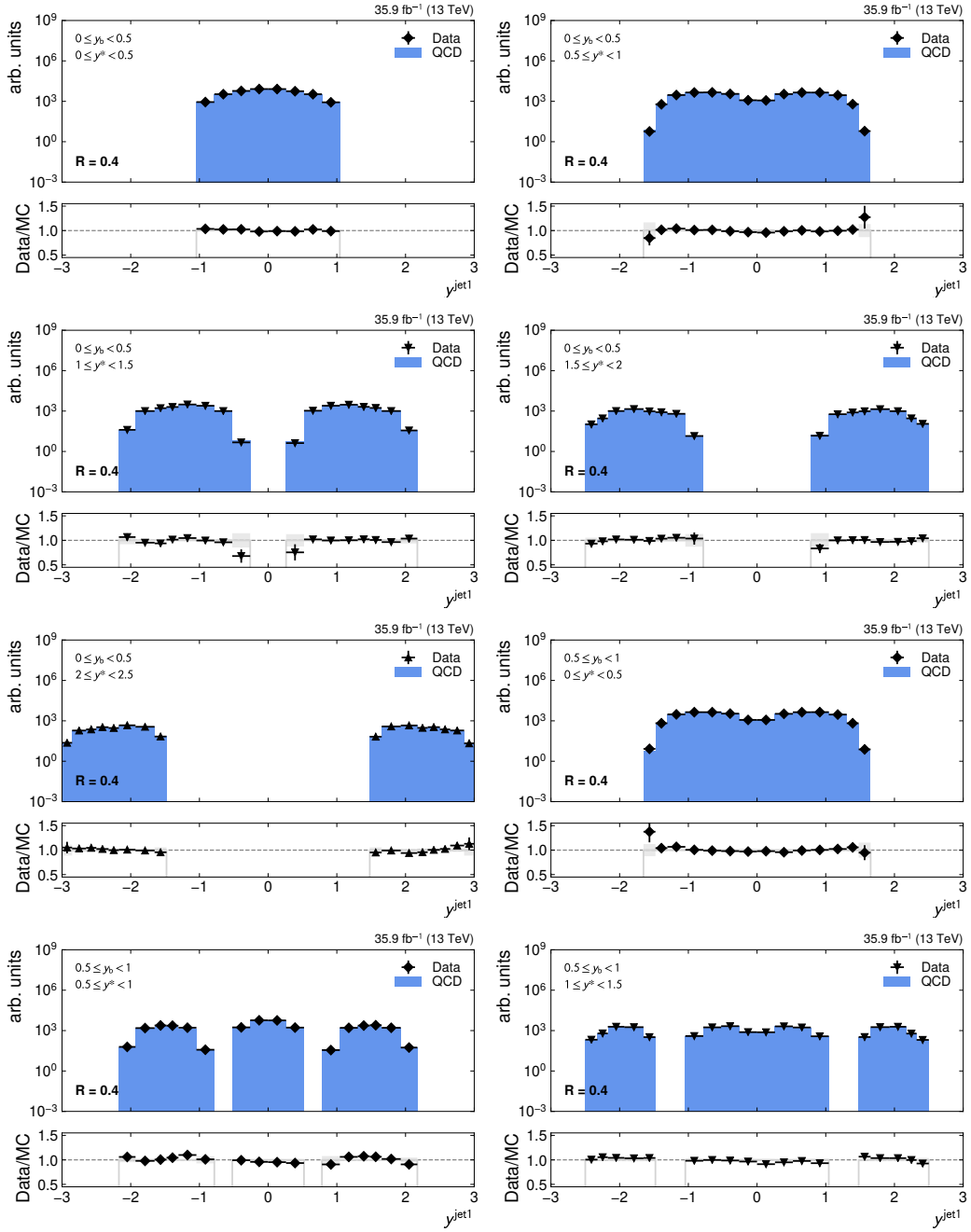

Figure A.31 – Comparison of the distribution of the rapidity of the leading jet $y_{\text{jet}\,1}$ in data and MC simulation, shown for jets with $R = 0.4$ in 8 out of 15 rapidity regions. The remaining ones are shown in figure A.32.

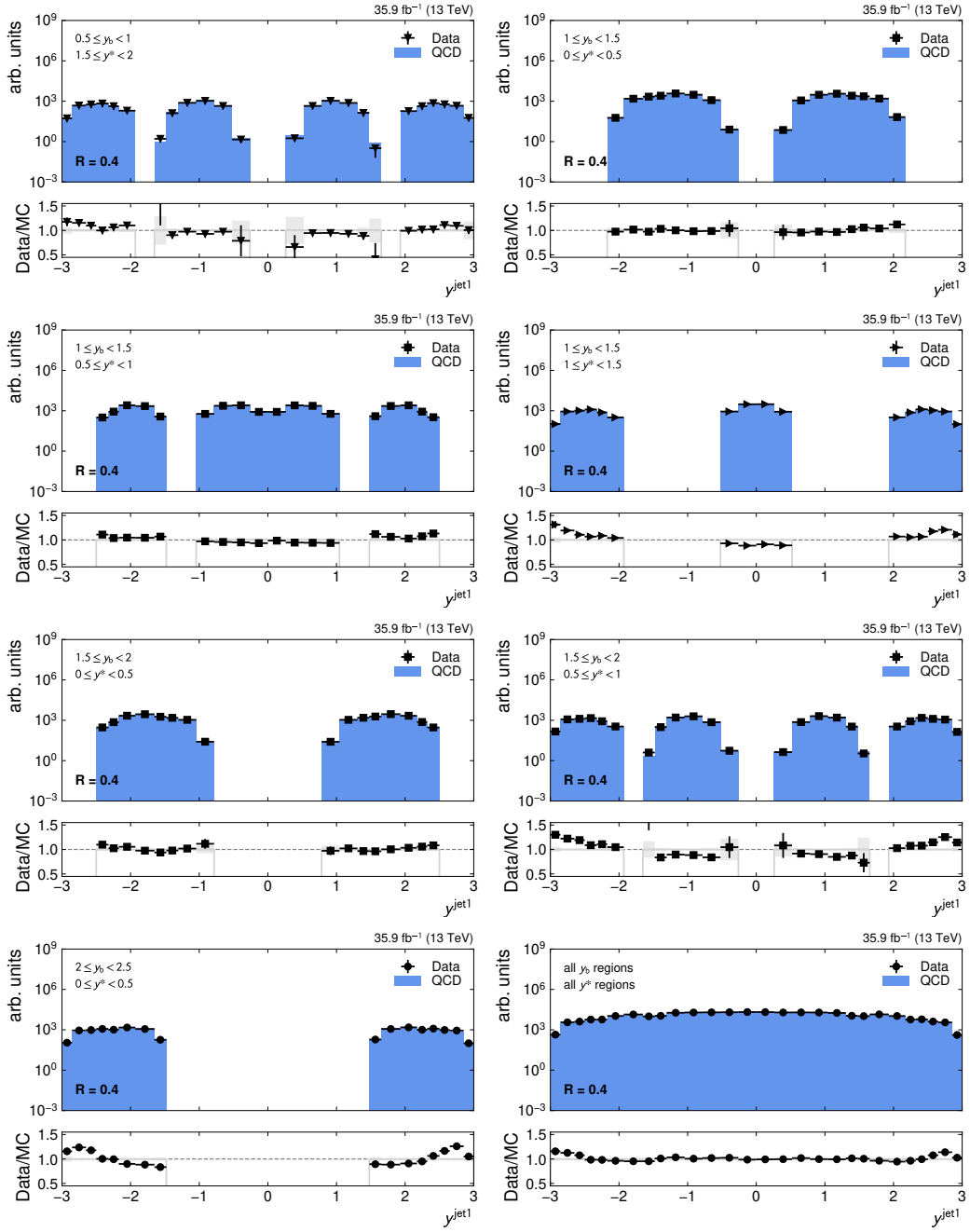

Figure A.32 – (*continued from figure A.31*) Comparison of the distribution of the rapidity of the leading jet $y_{\text{jet }1}$ in data and MC simulation, shown for jets with $R = 0.4$ in 7 out of 15 rapidity regions, as well as for the combination of all rapidity regions (*bottom right*).

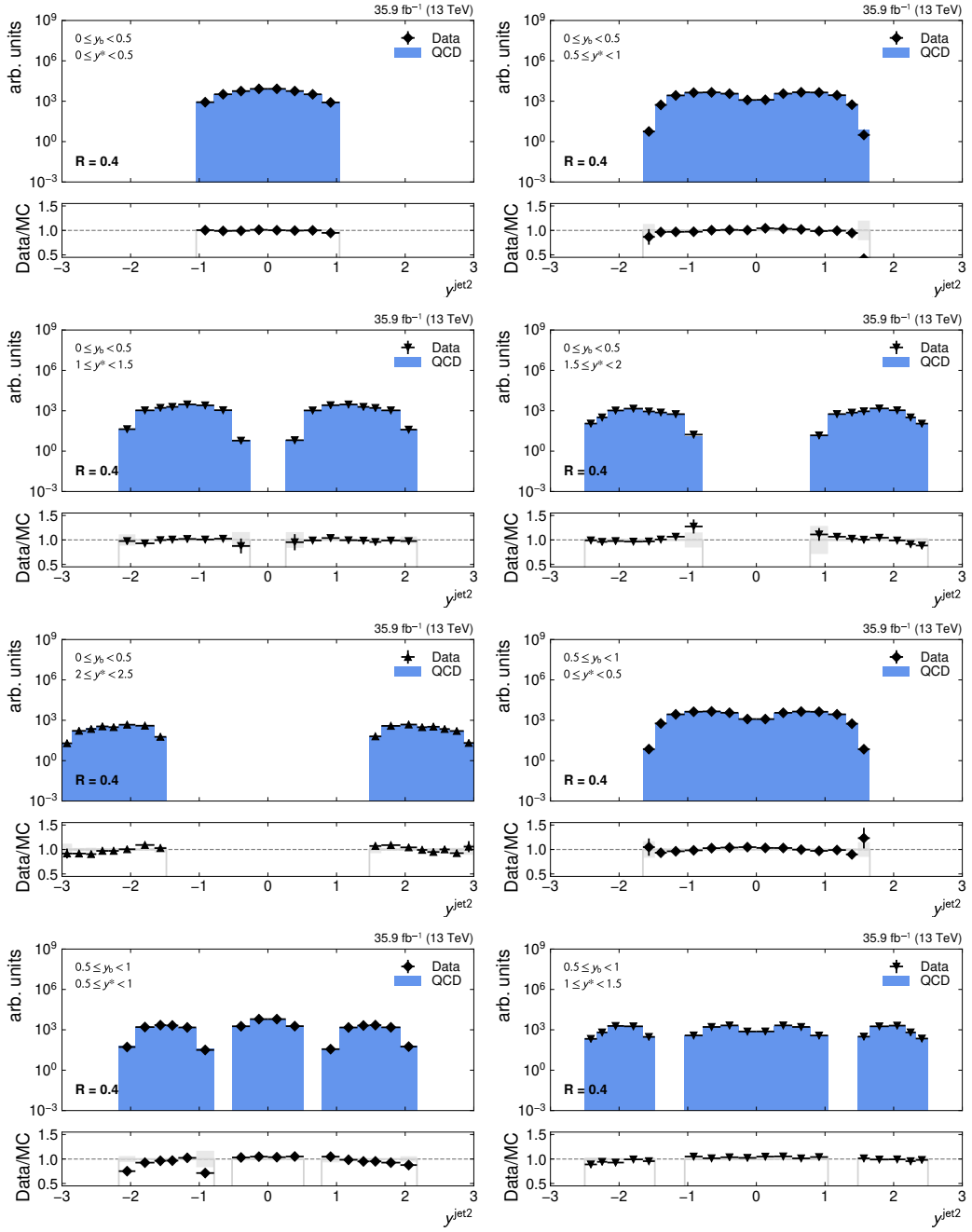

Figure A.33 – Comparison of the distribution of the rapidity of the subleading jet $y_{\text{jet 2}}$ in data and MC simulation, shown for jets with $R = 0.4$ in 8 out of 15 rapidity regions. The remaining ones are shown in figure A.34.

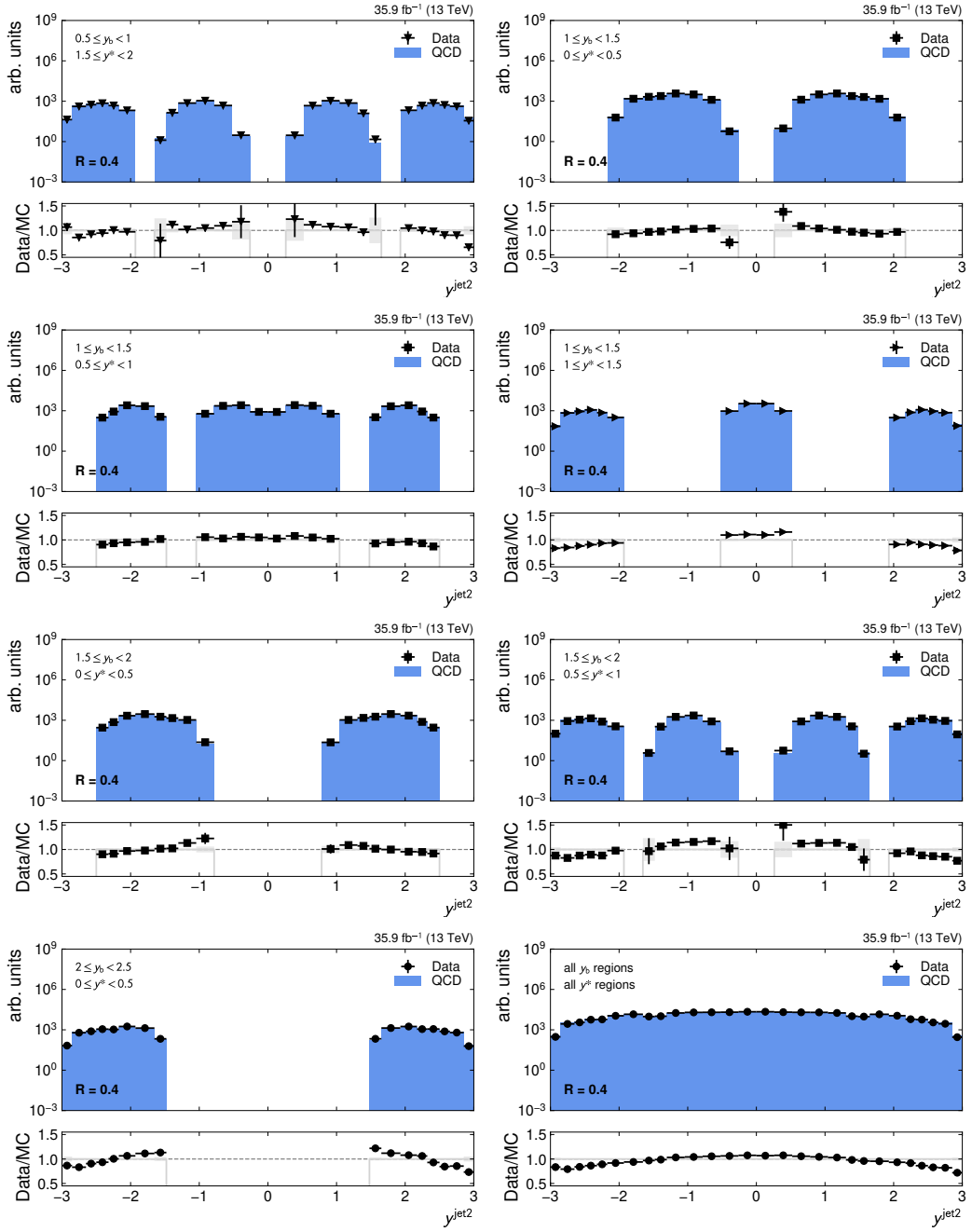

Figure A.34 – (*continued from figure A.33*) Comparison of the distribution of the rapidity of the subleading jet $y_{\text{jet 2}}$ in data and MC simulation, shown for jets with $R = 0.4$ in 7 out of 15 rapidity regions, as well as for the combination of all rapidity regions (*bottom right*).

List of Tables

List of Figures

References

[1] J. Currie, E. Glover, and J. Pires, "Next-to-Next-to Leading Order QCD Predictions for Single Jet Inclusive Production at the LHC," *Phys. Rev. Lett.* **118** (2017), no. 7, 072002, arXiv:1611.01460. doi:10.1103/PhysRevLett.118.072002.

[2] J. Currie, A. Gehrmann-De Ridder, T. Gehrmann, E. Glover, A. Huss, and J. Pires, "Precise predictions for dijet production at the LHC," *Phys. Rev. Lett.* **119** (2017), no. 15, 152001, arXiv:1705.10271. doi:10.1103/PhysRevLett.119.152001.

[3] A. Gehrmann-De Ridder, T. Gehrmann, E. Glover, A. Huss, and J. Pires, "Triple Differential Dijet Cross Section at the LHC," *Phys. Rev. Lett.* **123** (2019), no. 10, 102001, arXiv:1905.09047. doi:10.1103/PhysRevLett.123.102001.

[4] G. Sieber, "Measurement of triple-differential dijet cross sections with the CMS detector at 8 TeV and PDF constraints". PhD thesis, Karlsruhe Institute of Technology (KIT), 2016. doi:10.5445/IR/1000055723.

[5] **CMS** Collaboration, A. M. Sirunyan et al., "Measurement of the triple-differential dijet cross section in proton-proton collisions at $\sqrt{s} = 8$ TeV and constraints on parton distribution functions," *Eur. Phys. J. C* **77** (2017), no. 11, 746, arXiv:1705.02628. doi:10.1140/epjc/s10052-017-5286-7.

[6] R. Abdul Khalek et al., "Phenomenology of NNLO jet production at the LHC and its impact on parton distributions," arXiv:2005.11327.

[7] M. E. Peskin and D. V. Schroeder, "An introduction to Quantum Field Theory". CRC Press, 1995.

[8] R. Ellis, W. Stirling, and B. Webber, "QCD and Collider Physics". Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology. Cambridge University Press, 2003.

[9] F. Mandl and G. Shaw, "Quantum Field Theory". Wiley, 2010.

[10] **Particle Data Group** Collaboration, P. Zyla et al., "Review of Particle Physics," *PTEP* **2020** (2020), no. 8, 083C01. doi:10.1093/ptep/ptaa104.

[11] F. Englert and R. Brout, "Broken Symmetry and the Mass of Gauge Vector Mesons," *Phys. Rev. Lett.* **13** (1964) 321–323. doi:10.1103/PhysRevLett.13.321.

[12] P. W. Higgs, "Broken Symmetries and the Masses of Gauge Bosons," *Phys. Rev. Lett.* **13** (1964) 508–509. doi:10.1103/PhysRevLett.13.508.

[13] **ATLAS** Collaboration, G. Aad 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. B* **716** (2012) 1–29, arXiv:1207.7214. doi:10.1016/j.physletb.2012.08.020.

[14] **CMS** Collaboration, S. Chatrchyan et al., "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC," *Phys. Lett. B* **716** (2012) 30–61, arXiv:1207.7235. doi:10.1016/j.physletb.2012.08.021.

[15] G. 't Hooft and M. J. G. Veltman, "Regularization and Renormalization of Gauge Fields," *Nucl. Phys. B* **44** (1972) 189. doi:10.1016/0550-3213(72)90279-9.

[16] G. 't Hooft, "Dimensional regularization and the renormalization group," *Nucl. Phys. B* **61** (1973) 455–468. doi:10.1016/0550-3213(73)90376-3.

[17] S. Weinberg, "New approach to the renormalization group," *Phys. Rev. D* **8** (1973) 3497–3509. doi:10.1103/PhysRevD.8.3497.

[18] **NNPDF** Collaboration, R. D. Ball et al., "Parton distributions from high-precision collider data," *Eur. Phys. J. C* **77** (2017), no. 10, 663, arXiv:1706.00428. doi:10.1140/epjc/s10052-017-5199-5.

[19] S. Dulat, T. J. Hou, J. Gao, M. Guzzi, J. Huston, et al., "The CT14 Global Analysis of Quantum Chromodynamics," arXiv:1506.07443.

[20] L. Harland-Lang, A. Martin, P. Motylinski, and R. Thorne, "Parton distributions in the LHC era: MMHT 2014 PDFs," *Eur. Phys. J. C* **75** (2015), no. 5, 204, arXiv:1412.3989. doi:10.1140/epjc/s10052-015-3397-6.

[21] S. Alekhin, J. Blümlein, S. Moch, and R. Placakyte, "Parton distribution functions, α_s, and heavy-quark masses for LHC Run II," *Phys. Rev. D* **96** (2017), no. 1, 014011, arXiv:1701.05838. doi:10.1103/PhysRevD.96.014011.

[22] V. Gribov and L. Lipatov, "Deep inelastic ep scattering in perturbation theory," *Sov. J. Nucl. Phys.* **15** (1972) 438.

[23] Y. L. Dokshitzer, "Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics.," *Sov. Phys. JETP* **46** (1977) 641.

[24] G. Altarelli and G. Parisi, "Asymptotic Freedom in Parton Language," *Nucl. Phys. B* **126** (1977) 298. doi:10.1016/0550-3213(77)90384-4.

[25] S. Höche, "Introduction to parton-shower event generators," in *Theoretical Advanced Study Institute in Elementary Particle Physics: Journeys Through the Precision Frontier: Amplitudes for Colliders*, pp. 235–295. 2015. arXiv:1411.4085.

[26] A. Buckley et al., "General-purpose event generators for LHC physics," *Phys. Rept.* **504** (2011) 145–233, arXiv:1101.2599. doi:10.1016/j.physrep.2011.03.005.

[27] B. Andersson, G. Gustafson, G. Ingelman, and T. Sjöstrand, "Parton Fragmentation and String Dynamics," *Phys. Rept.* **97** (1983) 31. doi:10.1016/0370-1573(83)90080-7.

[28] B. Andersson, G. Gustafson, and B. Söderberg, "A General Model for Jet Fragmentation," *Z. Phys. C* **20** (1983) 317. doi:10.1007/BF01407824.

[29] T. Sjostrand, S. Mrenna, and P. Z. Skands, "A Brief Introduction to PYTHIA 8.1," *Comput. Phys. Commun.* **178** (2008) 852–867, arXiv:0710.3820. doi:10.1016/j.cpc.2008.01.036.

[30] B. Webber, "A QCD Model for Jet Fragmentation Including Soft Gluon Interference," *Nucl. Phys. B* **238** (1984) 492. doi:10.1016/0550-3213(84)90333-X.

[31] J.-C. Winter, F. Krauss, and G. Soff, "A modified cluster hadronization model," *Eur. Phys. J. C* **36** (2004) 381, arXiv:hep-ph/0311085. doi:10.1140/epjc/s2004-01960-8.

[32] D. Amati and G. Veneziano, "Preconfinement as a Property of Perturbative QCD," *Phys. Lett. B* **83** (1979) 87–92. doi:10.1016/0370-2693(79)90896-7.

[33] M. Bahr et al., "Herwig++ Physics and Manual," *Eur. Phys. J. C* **58** (2008) 639–707, arXiv:0803.0883. doi:10.1140/epjc/s10052-008-0798-9.

[34] **Sherpa** Collaboration, E. Bothmann et al., "Event Generation with Sherpa 2.2," *SciPost Phys.* **7** (2019), no. 3, 034, arXiv:1905.09127. doi:10.21468/SciPostPhys.7.3.034.

[35] C. Buttar et al., "The Underlying Event," in *HERA and the LHC: A Workshop on the Implications of HERA for LHC Physics: CERN - DESY Workshop 2004/2005 (Midterm Meeting, CERN, 11-13 October 2004; Final Meeting, DESY, 17-21 January 2005)*, pp. 192–217. CERN, Geneva, 2005.

[36] **CMS** Collaboration, V. Khachatryan et al., "Event generator tunes obtained from underlying event and multiparton scattering measurements," *Eur. Phys. J. C* **76** (2016), no. 3, 155, arXiv:1512.00815. doi:10.1140/epjc/s10052-016-3988-x.

[37] G. F. Sterman and S. Weinberg, "Jets from Quantum Chromodynamics," *Phys. Rev. Lett.* **39** (1977) 1436. doi:10.1103/PhysRevLett.39.1436.

[38] G. P. Salam, "Towards Jetography," *Eur. Phys. J. C* **67** (2010) 637–686, arXiv:0906.1833. doi:10.1140/epjc/s10052-010-1314-6.

[39] S. Catani, Y. L. Dokshitzer, M. Olsson, G. Turnock, and B. R. Webber, "New clustering algorithm for multi - jet cross-sections in e+ e- annihilation," *Phys. Lett. B* **269** (1991) 432. doi:10.1016/0370-2693(91)90196-W.

[40] S. Catani, Y. L. Dokshitzer, M. Seymour, and B. Webber, "Longitudinally invariant K_t clustering algorithms for hadron hadron collisions," *Nucl. Phys. B* **406** (1993) 187–224. doi:10.1016/0550-3213(93)90166-M.

[41] R. K. Ellis, "Quantum Chromodynamics and Deep Inelastic Scattering," *Adv. Ser. Direct. High Energy Phys.* **26** (2016) 61–78. doi:10.1142/9789814733519_0003.

[42] Y. L. Dokshitzer, G. Leder, S. Moretti, and B. Webber, "Better jet clustering algorithms," *JHEP* **08** (1997) 001, arXiv:hep-ph/9707323. doi:10.1088/1126-6708/1997/08/001.

[43] M. Wobisch and T. Wengler, "Hadronization corrections to jet cross-sections in deep inelastic scattering," in *Workshop on Monte Carlo Generators for HERA Physics (Plenary Starting Meeting)*, pp. 270–279. 4, 1998. arXiv:hep-ph/9907280.

[44] S. Marzani, G. Soyez, and M. Spannowsky, "Looking inside jets: an introduction to jet substructure and boosted-object phenomenology", volume 958. Springer, 2019. doi:10.1007/978-3-030-15709-8.

[45] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-k_t jet clustering algorithm," *JHEP* **04** (2008) 063, arXiv:0802.1189. doi:10.1088/1126-6708/2008/04/063.

[46] **CMS** Collaboration, S. Chatrchyan et al., "The CMS Experiment at the CERN LHC," *JINST* **3** (2008) S08004. doi:10.1088/1748-0221/3/08/S08004.

[47] E. Mobs, "The CERN accelerator complex - August 2018. Complexe des accélérateurs du CERN - Août 2018." OPEN-PHO-ACCEL-2016-013-1, CERN Document Server, https://cds.cern.ch/record/2636343, Aug, 2018.

[48] **ATLAS** Collaboration, G. Aad et al., "The ATLAS Experiment at the CERN Large Hadron Collider," *JINST* **3** (2008) S08003. doi:10.1088/1748-0221/3/08/S08003.

[49] **ALICE** Collaboration, K. Aamodt et al., "The ALICE experiment at the CERN LHC," *JINST* **3** (2008) S08002. doi:10.1088/1748-0221/3/08/S08002.

[50] **LHCb** Collaboration, J. Alves, A.Augusto et al., "The LHCb Detector at the LHC," *JINST* **3** (2008) S08005. doi:10.1088/1748-0221/3/08/S08005.

[51] **CMS** Collaboration, D. Barney, "CMS Slice." CMS-OUTREACH-2018-017, CERN Document Server, http://cds.cern.ch/record/2628641, Feb, 2015.

[52] G. Herrera, C. Hojat, and M. Sheaff, "Experimental Techniques in High Energy Physics", pp. 31–37. World Scientific, 1994.

[53] J. Varela, "CMS L1 Trigger Control System," Technical Report CMS-NOTE-2002-033, CERN, Geneva, Sep, 2002. http://cds.cern.ch/record/687458.

[54] G. Bauer et al., "Operational Experience with the CMS Data Acquisition System," *J. Phys. Conf. Ser.* **396** (2012) 012007. doi:10.1088/1742-6596/396/1/012007.

[55] C. Jones et al., "The new CMS data model and framework," in *Proocedings of the Conference on Computing in High Energy Physics, Mumbai India*, volume 213. 2006.

[56] **CMS** Collaboration, A. Sirunyan et al., "Particle-flow reconstruction and global event description with the CMS detector," *JINST* **12** (2017), no. 10, P10003, arXiv:1706.04965. doi:10.1088/1748-0221/12/10/P10003.

[57] **CMS** Collaboration, "Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET," Technical Report CMS-PAS-PFT-09-001, CERN, Geneva, Apr, 2009. https://cds.cern.ch/record/1194487.

[58] W. Adam, B. Mangano, T. Speer, and T. Todorov, "Track reconstruction in the CMS tracker," Technical Report CMS-NOTE-2006-041, CERN, Geneva, Dec, 2005. http://cds.cern.ch/record/934067.

[59] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet User Manual," *Eur. Phys. J. C* **72** (2012) 1896, arXiv:1111.6097. doi:10.1140/epjc/s10052-012-1896-2.

[60] **CMS** Collaboration, S. Chatrchyan et al., "Description and performance of track and primary-vertex reconstruction with the CMS tracker," *JINST* **9** (2014), no. 10, P10009, arXiv:1405.6569. doi:10.1088/1748-0221/9/10/P10009.

[61] **CMS** Collaboration, "Pileup Removal Algorithms," Technical Report CMS-PAS-JME-14-001, CERN, Geneva, 2014.

[62] M. Cacciari, G. P. Salam, and G. Soyez, "The Catchment Area of Jets," *JHEP* **04** (2008) 005, arXiv:0802.1188. doi:10.1088/1126-6708/2008/04/005.

[63] M. Cacciari and G. P. Salam, "Pileup subtraction using jet areas," *Phys. Lett. B* **659** (2008) 119–126, arXiv:0707.1378. doi:10.1016/j.physletb.2007.09.077.

[64] **CMS** Collaboration, V. Khachatryan et al., "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV," *JINST* **12** (2017), no. 02, P02014, arXiv:1607.03663. doi:10.1088/1748-0221/12/02/P02014.

[65] K. Rabbertz, "Jet Physics at the LHC: The Strong Force beyond the TeV Scale", volume 268 of *Springer Tracts in Modern Physics*. Springer, Berlin, 2017.

[66] **CMS** Collaboration, "Jet energy scale and resolution performance with 13 TeV data collected by CMS in 2016-2018." Detector Performance note CMS-DP-2020-019, https://cds.cern.ch/record/2715872, Apr, 2020.

[67] S. Agostinelli et al., "Geant4 – a simulation toolkit," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **506** (2003), no. 3, 250 – 303. doi:10.1016/S0168-9002(03)01368-8.

[68] **CMS** Collaboration, "Summary performance Plots for 2016 data." https://twiki.cern.ch/twiki/bin/view/CMSPublic/HLTplotsSummary2016, 2016. Public TWiki article, topic revision r2 (2017-08-23).

[69] V. Kuznetsov, D. Evans, and S. Metson, "The CMS data aggregation system," *Procedia Computer Science* **1** (2010), no. 1, 1535 – 1543. ICCS 2010. doi:https://doi.org/10.1016/j.procs.2010.04.172.

[70] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations," *Journal of High Energy Physics* **2014** (Jul, 2014). doi:10.1007/jhep07(2014)079.

[71] **CMS** Collaboration, A. M. Sirunyan et al., "Pileup mitigation at CMS in 13 TeV data," arXiv:2003.00503.

[72] S. Schmitt, "TUnfold, an algorithm for correcting migration effects in high energy physics," *Journal of Instrumentation* **7** (oct, 2012) T10003–T10003. doi:10.1088/1748-0221/7/10/t10003.

[73] T. Gehrmann et al., "Jet cross sections and transverse momentum distributions with NNLOJET," *PoS* **RADCOR2017** (2018) 074, arXiv:1801.06415. doi:10.22323/1.290.0074.

[74] T. Kluge, K. Rabbertz, and M. Wobisch, "fastNLO: Fast pQCD calculations for PDF fits," in *14th International Workshop on Deep-Inelastic Scattering (DIS 2006)*, p. 483. Tsukuba, Japan, 20-24 Apr 2006, 2006. arXiv:hep-ph/0609285.

[75] **fastNLO** Collaboration, D. Britzger, K. Rabbertz, F. Stober, and M. Wobisch, "New features in version 2 of the fastNLO project," in *20th International Workshop on Deep-Inelastic Scattering and Related Subjects*, pp. 217–221. 2012. arXiv:1208.3641.

[76] J. Currie, A. Gehrmann-De Ridder, T. Gehrmann, N. Glover, A. Huss, and J. Pires, "Jet cross sections at the LHC with NNLOJET," *PoS* **LL2018** (2018) 001, arXiv:1807.06057. doi:10.22323/1.303.0001.