University of Genova

school of mathematical, physical and natural sciences

Physics Department

Master Degree in Physics

Master Thesis

Search for intrinsic charm in the proton at the ATLAS experiment

Candidate:
Lucrezia Boccardo

Supervisors:
Prof. Federico Sforza
Dr. Francesco Armando Di Bello

Co-Supervisor:
Prof. Simone Marzani

Academic Year 2022/2023

CERN-THESIS-2023-367
21/09/2023

Science is no more than an investigation of a miracle we can never explain, and art is an interpretation of that miracle.

— Ray Bradbury, *The Martian Chronicles*

Contents

Chapter 1

Introduction

The dynamics of strong interactions and the internal content of the proton have been a matter of interest for particle physics ever since the discovery of quarks. This thesis focuses on the analysis of processes involving heavy quarks in the final states and, in particular, on a phenomenon called **intrinsic charm**, which is a model that predicts the existence of charm quarks in the initial state of the proton. The study of charm quarks relies on heavy flavour b- and c-jets, which can be produced in hadronic collisions. The work was done in the context of the Z+heavy flavour group, one of the groups performing precise measurements in the ATLAS experiment of the cross section of the production of the Z boson.

This allowed me to familiarize with all the high energy physics analysis techniques, from event selection to background estimate and Monte Carlo event generators. In particular, my focus was on developing predictions for Z+b-jets in a particular calculation scheme with massive quarks (called 4 Flavour Number Scheme), on intrinsic charm sensitivity optimization and intrinsic charm predictions. A second part of my thesis involved Parton Density Function parameterization, with a focus on an innovative Machine Learning technique called symbolic regression. Before delving into my analysis work, in Chapter 2 I will give a theoretical background to the physics that will be discussed. Chapter 3 will be dedicated to the experimental apparatus of LHC and ATLAS, and Chapter 4 will be focused on object reconstruction and identification procedures. Chapters 5 and 6 will contain the actual analysis work, from the measurements of Z boson cross sections, to the study of variable sensitivity to intrinsic charm. The last part, Chapter 7, deals with Machine Learning applied to Parton Density Functions parameterization via symbolic regression, which extracts simple functional forms given a dataset, to simplify and quicken the PDF parameterization process.

Chapter 2

An introduction to theory and phenomenology

The Standard Model (SM) is one of the most brilliant triumphs of particle physics, if not of the whole of physics. In the span of half a century, the Model was developed and later confirmed by several experimental evidences which culminated with the announcement of the observation of the Higgs boson by the ATLAS and CMS collaborations in 2012, the last building block of one of the most accurate and at the same time simple theories in modern physics.
Despite all its successes, to this day the Standard Model is still incomplete: there are many physical phenomena beyond our current understanding. Their research is comprehensively called *Beyond Standard Model* (BSM) physics, and it includes several branches of particle physics, from neutrinos to dark matter, from supersymmetry to Grand Unification theories, which all have in common an attempt to expand the Standard Model. These lines of research deal with events that are very hard to detect for a number of reasons: they are very rare because they have a low cross section and they are hard to detect with the instruments currently at our disposal. For this reason, there is a discipline parallel to BSM which works on precise measurements, with the goal of perfecting our knowledge of Standard Model compliant events in order to be able to identify with as much precision as possible those that are not.

2.1 Particles and forces in the Standard Model

The Standard Model is the quantum field theory that embodies all of our current understanding of particle physics. According to it, all interactions in nature happen between particles through three types of force: electromagnetic (EM), weak, and strong (QCD). The fourth type of known interaction, gravity, is for now not included in the Standard Model. The forces are carried by spin-1 particles called vector bosons. Each force is associated to a Quantum Field Theory (QFT) and to a symmetry. As seen in Figure 2.1, there are four types of boson (although, there are 12 vector bosons, 8 for QCD and 4 for electromagnetic and weak interactions).

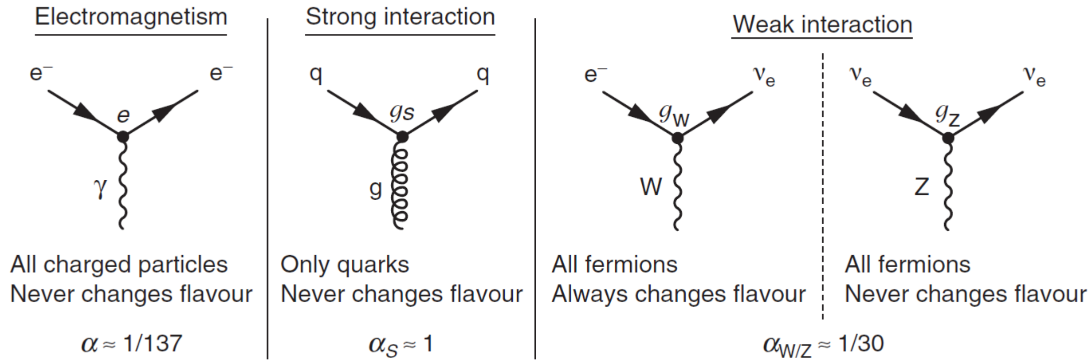


Figure 2.1: Interactions as they appear in Feynman diagrams [1]: the coupling constant at the vertex, whose approximate value is given in natural units below each diagram, is directly related to the relative strength of the interaction.

The rest of the particles in the Standard Model are spin $\frac{1}{2}$ fermions: leptons, who have integer charge, and quarks, who have fractional charge and, unlike leptons, can only be detected as bound states called hadrons (e.g. the proton and the neutron). There are 12 fermions in the SM. In QFT, particles are described by fields. In particular, spin $\frac{1}{2}$ particles are described by spinors ψ, whose free dynamics is described by the Dirac equation:

$$(i\gamma^\mu \partial_\mu - m)\,\psi = 0 \tag{2.1}$$

Bosons		
Boson	Mass [GeV]	Charge [e]
Photon	0	0
Gluon	0	0
W boson	80	$\pm$ 1
Z boson	91	0
Quarks		
Quark	Mass [GeV]	Charge [e]
up	0.005	$+\frac{1}{3}$
down	0.003	$-\frac{1}{3}$
charm	1.3	$+\frac{1}{3}$
strange	0.1	$-\frac{1}{3}$
top	4.5	$+\frac{1}{3}$
bottom	172	$-\frac{1}{3}$
Leptons		
Lepton	Mass [GeV]	Charge [e]
e	0.0005	-1
ν_e	0	0
μ	0.105	-1
ν_μ	0	0
τ	1.7	-1
ν_τ	0	0

Table 2.1: Particles in the SM

It is customary to divide leptons and quarks in three generations, based on mass and flavour. Most of common matter is made up of particles from the first generations. There is very strong experimental evidence that confirms that there are no more than three generations.

Even though the Standard Model is the only theory that, so far, has matched almost perfectly with experimental data, there is still something missing from the theory. First of all, the fourth fundamental interaction, gravity, is yet to be added to the Standard Model, and its vector boson still has not been observed. Another pressing issue is the neutrino mass: it has been shown experimentally, by studying solar neutrinos, that they undergo flavour oscillations. This is possible only if they have a mass: however, according to the Standard Model they do not. Current advances in neutrino physics have set an upper limit to the neutrino mass, but so far only the relative mass difference between mass eigenstates is known.

2.2 Symmetries

The Standard model relies on gauge symmetries to describe the interaction between particles mathematically. The vector bosons described earlier are, in the language of Quantum Field Theory, gauge fields that need to be introduced to maintain local gauge invariance, which is in turn needed to keep the Lagrangian symmetric and to ensure renormalizability, which is of the utmost importance to remove divergences in high-order calculations and to obtain finite results. The Standard Model Lagrangian is invariant under $\mathcal{U}(1) \times \mathcal{SU}(2) \times \mathcal{SU}(3)$ transformations, which correspond to the electromagnetic, weak and strong interactions.

2.2.1 Electromagnetism

The concept of gauge invariance is already familiar in electromagnetism without the addition of quantum mechanics: the vector potentials $\phi = A^0$ and $\mathbf{A}$, associated to the electric field and the magnetic field respectively, were invariant under the gauge transformation

$$A_\mu \to A'_\mu = A_\mu - \partial_\mu \chi \tag{2.2}$$

The concept of gauge invariance translated into quantum mechanics becomes a request of invariance under the local phase transformation $\mathcal{U}(1)$. This means that the Lagrangian must be invariant for

$$\psi(x) \to \psi'(x) = \hat{\mathcal{U}}(x)\psi(x) = e^{iq\chi(x)}\psi(x) \tag{2.3}$$

However, the Dirac equation (Eq. 2.1) is not invariant under such transformations, because it does not involve interactions (it is a free theory). In order to become invariant, a new degree of freedom must be added:

$$\left(i\gamma^\mu \partial_\mu + iqA_\mu\right)\psi - m\psi = 0 \tag{2.4}$$

This new addition is the massless gauge boson field of the photon.

2.2.2 Weak interaction

Weak interactions were first studied by Enrico Fermi to justify the β-decay $n \to e^- + p + \bar{\nu}_e$. He proposed an effective theory, which admitted parity violation and had a mass scale of $\Lambda = 300$ GeV. His theory, however brilliant and formally correct, presented a few issues, namely it was not renormalizable and was inaccurate at energies greater than Λ, since some cross sections diverged above it. Fermi's theory can be transformed into a gauge theory, solving both problems: gauge theories are always renormalizable, and the massive gauge boson will be the cut-off necessary to avoid divergences. It was already known, at the time, that weak interactions only involved left-handed particles and right-handed antiparticles. The associated Noether current is:

$$J_\mu^+ = \bar{\nu}_\ell \gamma_\mu \frac{1}{2}(1-\gamma_5)\,\ell = \bar{\ell}_L \gamma_\mu \sigma^+ \ell_L \tag{2.5}$$

where $\sigma^+ = \frac{1}{2}(\sigma_1 + i\sigma_2)$ is a linear combination of Pauli matrices. The gauge symmetry has been found, and it is $\mathcal{SU}(2)$, whose generators are precisely Pauli matrices. Left handed leptons will participate in the interaction and be placed in $\mathcal{SU}(2)$ doublets, like $\begin{pmatrix} \nu_{e_L} \\ e_L \end{pmatrix}$, whereas right handed leptons will be placed in singlets. $\mathcal{SU}(2)$ has three generators, which means there are two more currents: J_μ^- and J_μ^3, which is neutral. $J_\mu^\pm$ can be easily identified as the currents associated to the vector bosons $W^\pm$. However, J_μ^3 and its boson W_μ^3 are not associated to the Z boson, because the current does not admit coupling to right handed particles, while the Z bosons couples to them, albeit with different intensity. For the same reason, this neutral current is not the electromagnetic current.

2.2.3 Electroweak unification

In the 1970s, Glashow, Salam and Weinberg proposed a solution to the issue of the neutral weak current. In order to add a second neutral vector boson to the theory, it was sufficient to include a $\mathcal{U}(1)$ invariance, that is requesting the invariance under a transformation analogous to the one described in Eq. 2.3. The conserved charge $\frac{Y}{2}$ is called hypercharge and the boson field is B_μ. The Lagrangian for the neutral current interaction is

$$\mathcal{L}_n = \bar{\Psi}\gamma_\mu \left[g T_3 W_3^\mu + g' \frac{Y}{2} B^\mu \right] \Psi \tag{2.6}$$

Ψ is a vector which includes both right- and left-handed particles. T_3 is the quantum number of weak isospin, which is 0 for singlets and $\pm\frac{1}{2}$ for doublets. The hypercharge quantum number can now be set in order to obtain the electromagnetic current in Eq. 2.6. First, the two neutral bosons are rotated to obtain A_μ and Z_μ.

$$B^\mu = A^\mu \cos\theta_W - Z^\mu \sin\theta_W \tag{2.7}$$

$$Z^\mu = A^\mu \sin\theta_W + Z^\mu \cos\theta_W \tag{2.8}$$

Equation 2.6 now becomes:

$$\mathcal{L}_n = \bar{\Psi}\left(g\sin\theta_W T_3 + \frac{Y}{2} g' \cos\theta_W \right) \Psi A^\mu + \bar{\Psi}\left(g\cos\theta_W T_3 - \frac{Y}{2} g' \sin\theta_W \right) \Psi Z^\mu \tag{2.9}$$

The two terms are, respectively, the electromagnetic current and the Z neutral current. Conventionally, $Y(\ell_L)$ = -1, which means that

$$g\sin\theta_W = g'\cos\theta_W = e \tag{2.10}$$

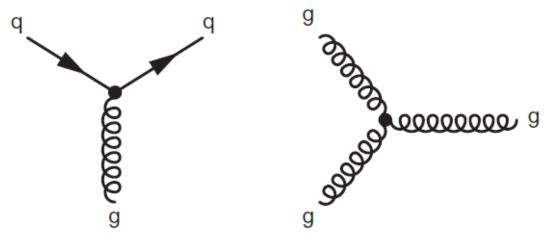


Figure 2.2: The three vertices involving quarks and gluons predicted by $\mathcal{SU}(3)$ invariance [1].

Hadrons can be added to the Langrangian in the same way, but the doublet is usually written as $\begin{pmatrix} u_L \\ d'_L \end{pmatrix}$, where

$$\begin{pmatrix} d'_L \\ s'_L \\ b'_L \end{pmatrix} = V \begin{pmatrix} d_L \\ s_L \\ b_L \end{pmatrix} \tag{2.11}$$

in which V is the CKM matrix of flavour mixing.

2.2.4 Strong interaction

The strong interaction is the most intense of the four fundamental forces. It is 137 times stronger than electromagnetism, and almost 10^6 times stronger than the weak interaction. It is described by Quantum Chromodynamics (QCD), which is a non-abelian gauge theory associated to the $\mathcal{SU}(3)$ symmetry. Due to its being non-abelian, the theory predicts self-interaction between the mediators, which is why Feynman diagrams that involve 3 or 4 gluons are allowed, as seen in Figure 2.2. $\mathcal{SU}(N)$ symmetries have $N^2 - 1$ generators, which means that QCD has 8 massless vector bosons. Furthermore, their representations are 3×3 matrices called Gell-Mann matrices. This means that there are three degrees of symmetry, which are identified by the colors red, green and blue. Gluons only couple to particles that have non-zero color charges, that is quarks. Due to confinement, QCD bounds them together in zero-color bound states called hadrons: two quarks form a meson, three quarks form a baryon. At very high energies or at very short distances, the intensity of QCD is greatly reduced, and quarks and gluons are free. This concept is called asymptotic freedom, and it allows experimentalists to study the behaviour of these fundamental particles when the energy is high enough. The strong interaction does not distinguish between quark flavours. However, since quarks have vastly different masses, the flavour symmetry in the QCD Lagrangian is approximate.

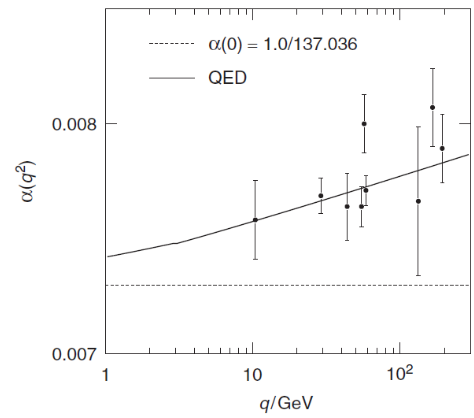


Figure 2.3: Running of the QED coupling constant [1]

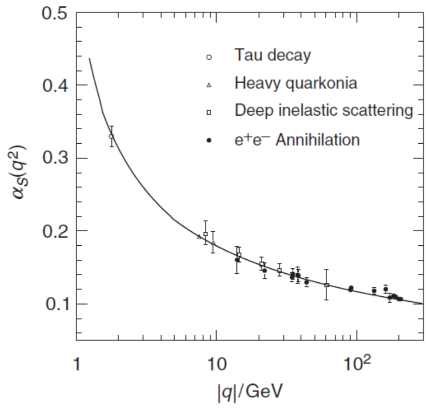


Figure 2.4: Running of the QCD coupling constant [1]

2.2.5 Running couplings

The relative intensity of the fundamental forces is determined by their coupling constants. In the language of Feynman diagrams, they appear in the vertices, determining the strength of the interactions. As seen when discussing QCD, the strength of the interaction is neither fixed nor constant, it depends on the energy scale. For example, the intensity of the strong interaction decreases with energy (the so-called asymptotic freedom), while the electromagnetic's increases. The evolution of the coupling constant is governed by the β function [2]:

$$\beta(\alpha) = \mu \frac{d\alpha}{d\mu} \tag{2.12}$$

where α is the generic coupling constant and μ is the energy scale. In the case of the coupling of the electromagnetic force,

$$\alpha(\mu) = \frac{2\pi}{\beta_0} \frac{1}{\ln \frac{\mu}{\Lambda_{QED}}} \tag{2.13}$$

where β_0 is the first-order expansion of the β function and $\Lambda_{QED} = 10^{286}$ is the Landau pole, the energy scale at which the interaction is too strong for a perturbative approach. From Figures 2.3 and 2.4, the behaviour previously described is evident: the QCD coupling constant decreases with energy, the electromagnetic one increases.

2.3 Parton Density Functions

2.3.1 The internal structure of the proton

As seen in Section 2.1, the proton is not one of the fundamental particles of the Standard Model. It has a composite structure, whose nature was first investigated in electron-proton scattering. Whenever these two particles interact electromagnetically, a virtual photon is exchanged with a wavelength λ that is inversely proportional to the center of mass energy. If the energy is sufficiently high, the photon's wavelength is smaller than the classical radius of the proton r_p and the interaction consists of deep inelastic scattering between the electron and one of the fundamental components of the proton, as seen in Figure 2.5.

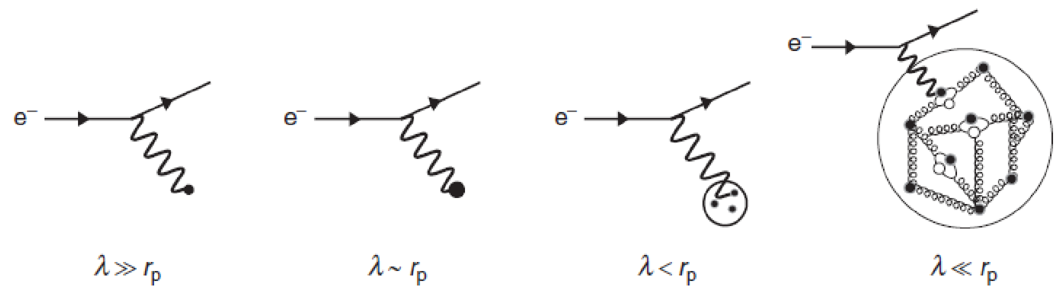


Figure 2.5: Electron-proton interaction at different energy scales [1]

In general, the elastic $e^-p \rightarrow e^-p$ cross section is:

$$\frac{d\sigma}{d\Omega} = \frac{\alpha^2}{4E_1^2\sin^4(\theta/2)} \frac{E_3}{E_1} \left(\cos^2\frac{\theta}{2} + \frac{Q^2}{2m_P^2}\sin^2\frac{\theta}{2} \right) \tag{2.14}$$

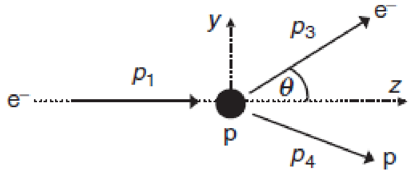


Figure 2.6: As seen in Eq. 2.14

where $Q^2 = -q^2$. In Eq. 2.14, the term in $sin^2(\theta/2)$ comes exclusively from spin-spin interactions. However, it assumes that the proton is a point-like particle, whereas it actually has a radius, as stated earlier, and electric and magnetic charge distributions. The previous equation can therefore be corrected in the Rosenbluth formula (Eq. 2.15) by adding $G_E(Q^2)$ and $G_M(Q^2)$, two form factors depending on the virtual photon four-momentum and connected to the Fourier transforms of the distributions.

$$\frac{d\sigma}{d\Omega} = \frac{4\pi\alpha^2}{Q^4} \left[\frac{G_E^2 + \frac{Q^2}{m_P}G_M^2}{1 + \frac{Q^2}{m_P}} \left(1 - y - \frac{m_P^2 y^2}{Q^2} \right) + \frac{1}{2}y^2 G_M^2 \right] \tag{2.15}$$

Due to these distributions, the elastic cross section at high energies decreases rapidly:

$$\left(\frac{d\sigma}{d\Omega}\right)_{elastic} \propto \frac{1}{Q^6} \tag{2.16}$$

This means that the higher the energy, the higher the chances of the proton breaking up and there being an inelastic scattering interaction instead of an elastic one. In the inelastic interaction, the cross section depends on another Lorentz-invariant (LI) kinematic variable, which is $x = 1$ in elastic interactions:

$$x = \frac{Q^2}{2p_2 \cdot q} \tag{2.17}$$

so that

$$\frac{d^2\sigma}{dxdQ^2} = \frac{4\pi\alpha^2}{Q^4}\left[(1-y)\frac{F_2(x,Q^2)}{x} + y^2 F_1(x,Q^2)\right] \tag{2.18}$$

where it is assumed that $\mathrm{Q}^2 \gg m_p^2 y^2$. $F_1(x,Q^2)$ and $F_2(x,Q^2)$ are related to G_E and G_M and are called *structure functions*. A very important experimental observation is **Bjorken scaling**, which shows that structure functions do not depend on Q^2. This was one of the first hints to an internal point-like structure of the proton. The suspicion was furthermore confirmed by the **Callan-Gross relation**, that proved that

$$F_2(x) = 2xF_1(x) \tag{2.19}$$

which means that an inelastic proton-electron interaction is *actually* an elastic scattering of electrons with half-spin internal components of the proton: the quarks. In particular, it is a $eq \to eq$ interaction, whose Feynman diagram is in Figure 2.7.

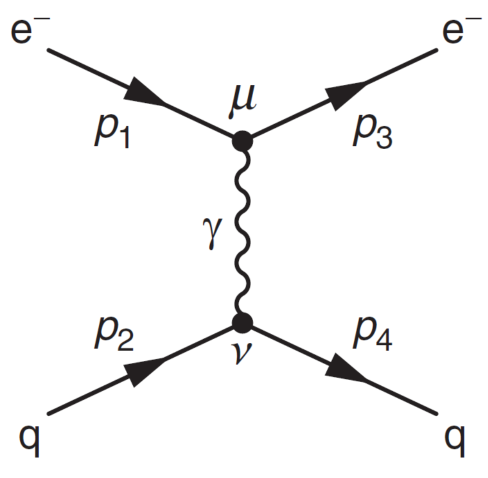


Figure 2.7: Leading order quark-electron scattering Feynman diagram

The matrix element for this process is:

$$\mathcal{M}_{fi} = \frac{Q_q e^2}{q^2} \left[\bar{u}(p_3)\gamma^{\mu} u(p_1)\right] g_{\mu\nu} \left[\bar{u}(p_4)\gamma^{\nu} u(p_2)\right] \tag{2.20}$$

and the differential cross section for deep inelastic scattering as a function of θ^* of Figure 2.8 comes out as

$$\frac{d\sigma}{d\Omega^*} = \frac{Q_q^2 e^4}{8\pi^2 s} \frac{\left[1 + \frac{1}{4}(1 + cos\theta^*)^2\right]}{(1 - cos\theta^*)^2} \tag{2.21}$$

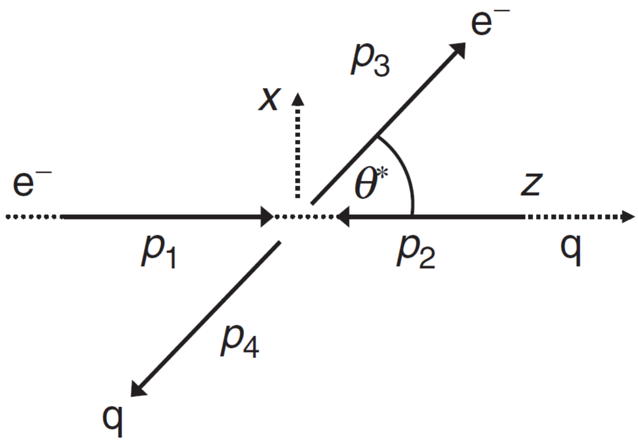


Figure 2.8: Quark-electron scattering in the center of mass frame [1]

2.3.2 The parton model

According to the parton model, the proton is made up of fundamental constituents called partons, and the total deep inelastic scattering cross section is the sum of all their individual contributions. The proton is energetic enough to be considered massless, and every transverse component of the parton momentum can be neglected. Each parton carries a fraction ξ of the total momentum of the proton, so that

$$p_q = \xi p_2 = (\xi p_2, 0, 0, \xi p_2) \tag{2.22}$$

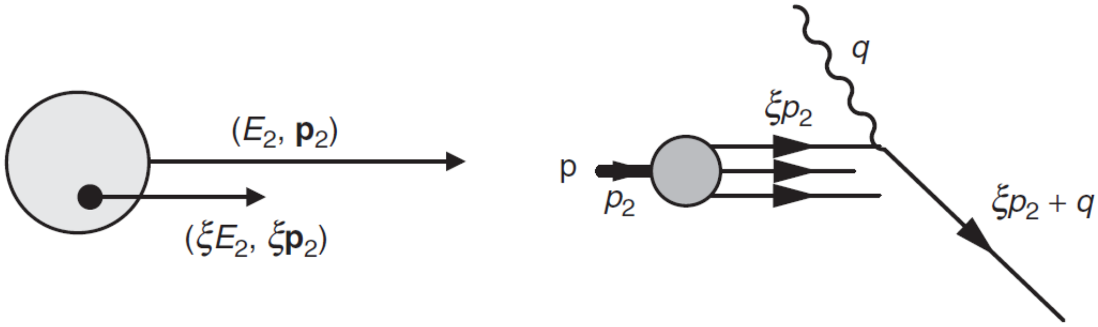


Figure 2.9: Parton model interaction

It can be shown that the fraction of momentum is the same as Bjorken x, because as seen in Figure 2.9

$$(\xi p_p + q)^2 = m_q^2 \tag{2.23}$$

and given that both the parton and the proton are massless

$$\xi = \frac{-q^2}{2p_2 \cdot q} = x \tag{2.24}$$

which is exactly the definition of the Bjorken x.
Hence, the center of mass energy of the quark-electron interaction is

$$s_q = (p_1 + xp_2)^2 = 2xp_1 \cdot p_2 = xs \tag{2.25}$$

and the differential cross section for elastic scattering is

$$\frac{d\sigma}{dq^2} = \frac{2\pi\alpha^2 Q_q^2}{q^4}\left[1 + \left(1 + \frac{q^2}{s_q^2}\right)\right] \tag{2.26}$$

Defining analagous variables as those of Sec. 2.3.1,

$$Q^2 = -q^2 \tag{2.27}$$

$$y = \frac{p_q \cdot q}{p_q \cdot p_1} = \frac{xp_2 \cdot q}{xp_2 \cdot p_1} \tag{2.28}$$

and substituting them in Eq. 2.26, it becomes:

$$\frac{d\sigma}{dQ^2} = \frac{4\pi\alpha^2 Q_q^2}{Q^4}\left[(1-y) + \frac{y^2}{2}\right] \tag{2.29}$$

This is the same structure as the deep inelastic scattering Eq. 2.15. In order to understand the physical meaning behind this analogy, a bit of Quantum Chromodynamics is needed.

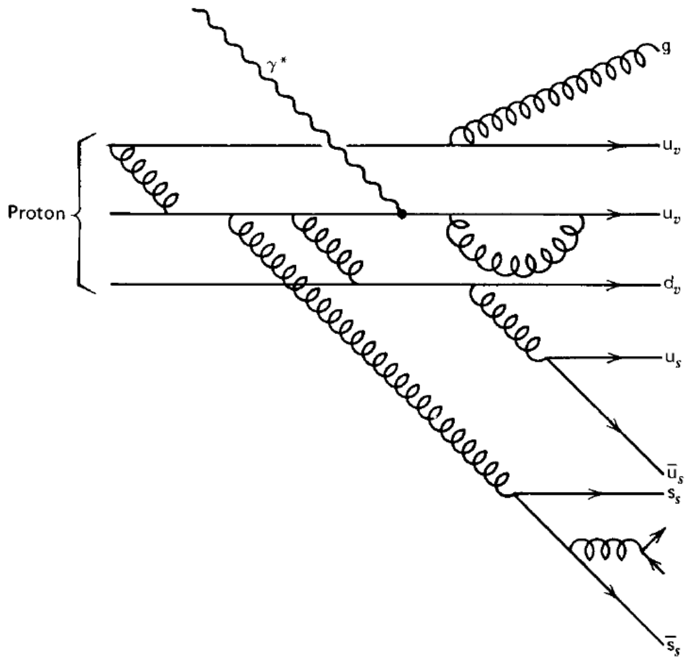


Figure 2.10: QCD interactions inside the proton [3]

2.3.3 Parton Density Functions

The interactions of quarks inside the proton are governed by their momenta distributions, called Parton Density Functions, or PDFs. Each quark will carry a fraction of the momentum of the proton and, since at high energy the scattering happens between electron and quark, these distributions, which will be called $q_i(x)$ with i identifying the flavour, will be related to the structure functions of Eq. 2.18. At these energy scales, the functions can not be inferred because QCD can not be treated perturbatively, due to the high value of its coupling constant $\alpha_s = \mathcal{O}(1)$: they are usually determined experimentally. In order to understand the relationship between PDFs and structure functions, the cross section for electron-quark scattering needs to be generalized. PDFs are defined so that $q_i(x)dx$ are quarks carrying a momentum fraction between x and $x + dx$. Hence, the double differential cross section for all flavours will be:

$$\frac{d^2\sigma}{dxdQ^2} = \frac{4\pi\alpha^2}{Q^4}\left[(1-y) + \frac{y^2}{2}\right]\sum_i Q_i^2 q_i(x) \tag{2.30}$$

It is straightforward then to compare Eq. 2.30 and Eq. 2.18 to find out that PDFs are related to structure functions by

$$F_2(x, Q^2) = 2xF_1(x, Q^2) = x\sum_i Q_i^2 q_i(x) \tag{2.31}$$

Assuming that third-generation quarks are negligible inside the proton due to their size:

$$F_2(x) = \left(\frac{2}{3}\right)^2 [u(x) + \bar{u}(x) + c(x) + \bar{c}(x)] + \left(\frac{1}{3}\right)^2 \left[d(x) + \bar{d}(x) + s(x) + \bar{s}(x)\right] \tag{2.32}$$

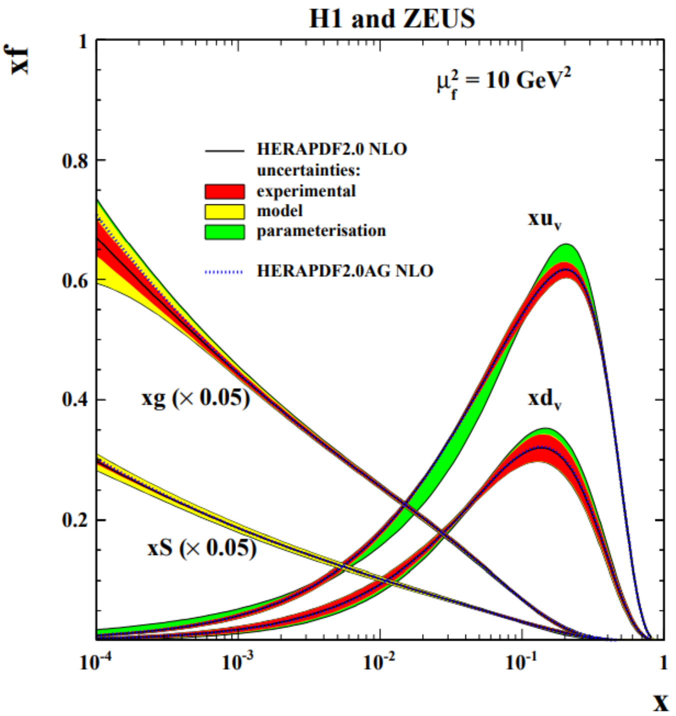


Figure 2.11: Valence and sea PDFs, HERA collaboration [4]

Since quantum numbers associated with QCD still need to be conserved, the structure of the proton can be described as **three valence quarks** $u_v u_v d_v$ and many quark-antiquark pairs that constitute the **sea quarks** $q_S \bar{q}_S$, which are radiated from valence quarks, as in Fig. 2.10. For this reason, gluons and sea quarks have PDFs that grow logarithmically as $x \to 0$, as seen in Fig. 2.11, while valence quarks have peaks.

2.3.4 DGLAP equations

QCD is a scale-dependant interaction, and PDFs make no exception. Their dependance on the exchanged momentum Q^2 is regulated by the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) equations [5–7], which are different for quarks q and gluons g:

$$\frac{dq(x,Q^2)}{d\ln Q^2} = \frac{\alpha_s(Q^2)}{2\pi}\int_x^1 \frac{dz}{z} P_{qq}\left(\frac{x}{z}, \alpha_s(Q^2)\right) q(z,Q^2) + \frac{\alpha_s(Q^2)}{2\pi}\int_x^1 \frac{dz}{z} P_{qg}\left(\frac{x}{z}, \alpha_s(Q^2)\right) g(z,Q^2) \tag{2.33}$$

$$\frac{dg(x,Q^2)}{d\ln Q^2} = \frac{\alpha_s(Q^2)}{2\pi}\int_x^1 \frac{dz}{z}\,P_{gq}\left(\frac{x}{z},\alpha_s(Q^2)\right)\,q(z,Q^2) + \frac{\alpha_s(Q^2)}{2\pi}\int_x^1 \frac{dz}{z}\,P_{gg}\left(\frac{x}{z},\alpha_s(Q^2)\right)\,g(z,Q^2) \tag{2.34}$$

They take into consideration both the emission and absorption of softer partons through the **splitting functions** P_{ij} [2]

2.3.5 Proton-proton collisions

So far, only electron-proton collisions have been discussed. In the LHC, however, the collisions are between two hadrons, which means that there are parton density functions on both sides of the interaction. There are, then, three unknown variables: Q^2, x_1 and x_2. As said earlier, the final states of a hadron collisions are jets, which is why the observables have to be jet-related. It is customary to use the two jet-angles and one of their transverse momentum

$$p_T = \sqrt{p_x^2 + p_y^2} \tag{2.35}$$

However, since the center of mass of the protons is not the same as the center of mass of the partons, it is better to use a Lorentz invariant for the jet angles as well: the rapidity y. It is defined as

$$y = \frac{1}{2}\ln\frac{E+p_z}{E-p_z} \tag{2.36}$$

Whenever the energies of the jets are high enough, the jet mass can be neglected and the pseudorapidity η can be used

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right) \tag{2.37}$$

As seen in Section 2.5.1, hard scattering, which is the actual interaction between protons, can be treated perturbatively: for an $ij \to X$ interaction (partons i and j with final state X) there exists a Feynman diagram, like those in Figure 2.12, and a cross section can be computed as a function of the Mandelstam variables of the parton-parton system:

$$\hat{\sigma}_{ij\to X}(\hat{s},\hat{t},\hat{u}) = \frac{1}{\hat{s}}\int |\mathcal{M}_{ij\to X}|^2\,(2\pi)^4\delta^{(4)}(p_i + p_j - \sum_f p_f)\prod_f \frac{d^3p_f}{(2\pi)^3 2E_f}. \tag{2.38}$$

However, once this cross section is inserted in the global $pp \to X$ cross section, Parton Density Functions have to be taken in consideration, because they determine the momentum distribution of the i and j partons:

$$\sigma = \sum_{i,j}\int dx_1 dx_2\, f_i(x_1,Q^2)\, f_j(x_2,Q^2)\, \hat{\sigma}_{ij\to X}(\hat{s},\hat{t},\hat{u}) \tag{2.39}$$

Thanks to the **factorization theorem**, the perturbative and non-perturbative calculations can be dealt with separately.

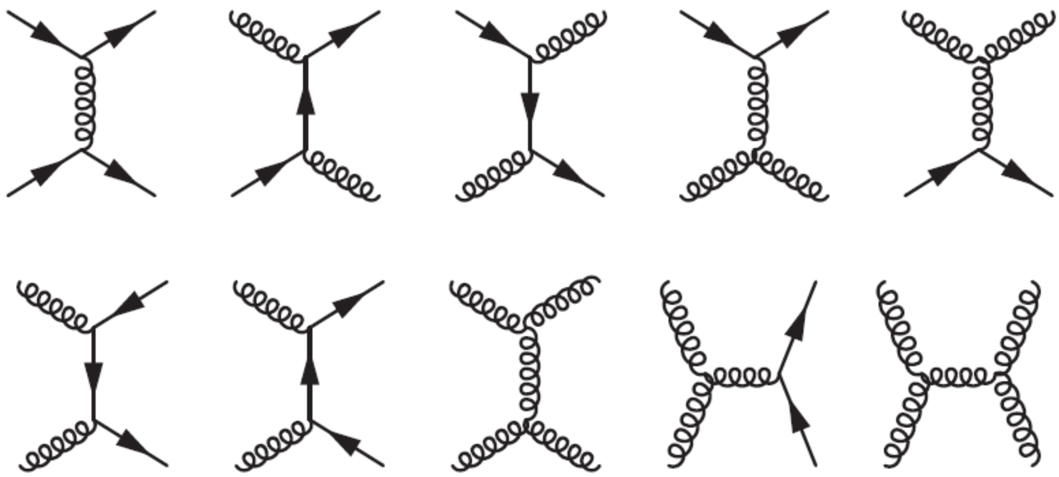


Figure 2.12: Examples of Feynman diagrams in $ij \to 2$ jets [1]

2.3.6 Determination of PDFs

Since PDFs are *de facto* probability distributions, they can not be determined theoretically. Current approaches at producing PDFs usually start from different sets of data (deep inelastic scattering, Drell-Yann, jets etc.) and, after reconstructing the events and applying QCD calculations, they use various fitting methods to extract the PDF. There are many types of PDF, and each collaboration is different from the other both in the kind of data they use to produce the sets, and in the fitting method. The previously mentioned HERA collaboration [4] uses xFitter [8], a QCD fit-framework that can evolve PDFs with the DGLAP equations, parameterise sets and test different theoretical predictions. Other collaborations use different methods: the NNPDF collaboration [9] uses neural networks like the one in Figure 2.13, to avoid the bias of a predetermined fitting function.

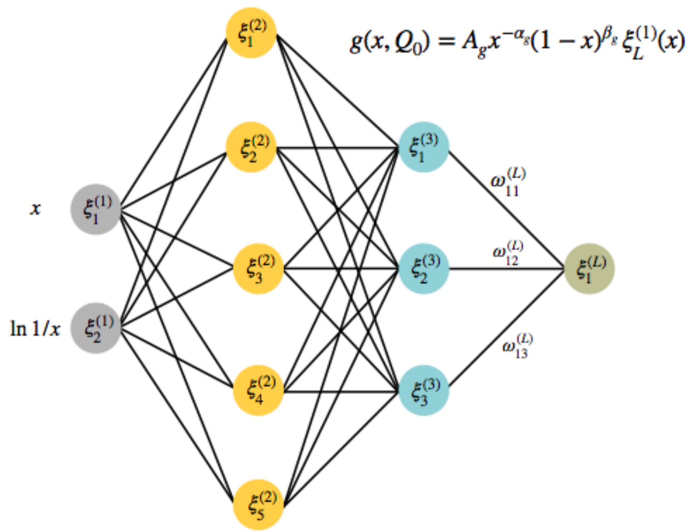


Figure 2.13: Neural network for PDF fitting [10]. In the top right corner, the equation shows the PDF functional form, where ξ_L are the NN node weights.

2.4 Intrinsic charm

Intrinsic charm as a theory was first proposed by Brodsky, Hoyer, Peterson and Sakai as the BHPS model [11] in response to a series of experimental observations in fixed-target proton experiments that did not match predictions. In particular, they cite a higher charm cross section (above the 10-15 μb predicted by perturbative QCD) and an excess of production of D^+ and Λ_c^+ in the forward region. The latter was unexpected in particular for the D meson, because, since it does not share any valence quarks with the proton (it has a $\bar{u}c$ structure), its production is expected to be reduced in the so-called proton fragmentation region of the proton. This region is also broadly identified with the central region, which explained the different configurations for valence and sea quarks, as seen in the previous Section: valence quarks transport a higher momentum fraction and are more easily described by perturbative QCD, and their interactions involve a lower momentum transfer, hence the products are produced at a smaller angle with respect to the z-axis.

2.4.1 Models for intrinsic charm

Several theoretical models for intrinsic charm have been developed. In the BHPS model, perhaps the most notorious, *intrinsic* charm is defined in a way that is subtly different from a full-on pentaquark approach. It is, rather, a matter of introducing a $|uudc\bar{c}\rangle$ component in the Fock space decomposition which is non-negligible with respect to the $|uud\rangle$ one. Its origin is non-perturbative, as opposed to sea quarks which come from radiative QCD processes. Therefore, while in the penta-quark state there are five valence quarks, in the intrinsic charm model the charm quark has a *valence-like* behaviour which allows him to exist inside the proton over longer time scales than sea quarks. The model is defined around three parameters: the **charm content**, which is the fraction of momentum carried by intrinsic charm quarks, the charm PDF, and the scale energy at which intrinsic charm is relevant. Studies for the presence of intrinsic charm have been performed both on fixed target experiments, such as muon-electron DIS, and collider ones, and they all agree on putting an upper limit on IC content around 2%.

2.4.2 Experimental and phenomenological state of the art

The intrinsic charm model has been around since the 1980s, and many collaborations have attempted to measure it.
At LHCb, a study on the production of the Z boson and charm jets in the forward region used $\mathcal{R}_c^j = \frac{\sigma(Zc)}{\sigma(Zj)}$ to show an excess of charm in the region consistent with the model, although a charm percentage could not be given [12]. Also, the CMS collaboration has performed several measurments of $Z + c$ inclusive and differential cross sections [13]. In the summer of 2022, the NNPDF collaboration published a paper claiming evidence of intrinsic charm in the proton [14]. Their approach, instead on focusing on one dataset and studying, for instance, the Forward/Central Ratio, exploits the techniques used in producing Parton Density Functions to extract a non-radiative PDF for the charm. Through NNLO calculations, and by comparing matching conditions in 3 and 4 flavour schemes (i.e. PDFs in which either the b quark or both the b and the c are considered massive and integrated away), they managed to extract a valence-like shape for the charm PDF (see Figure 2.14), which matches the required conditions: it is compatible with zero at low x and has a peak roughly in the middle of the $[0, 1]$ interval.

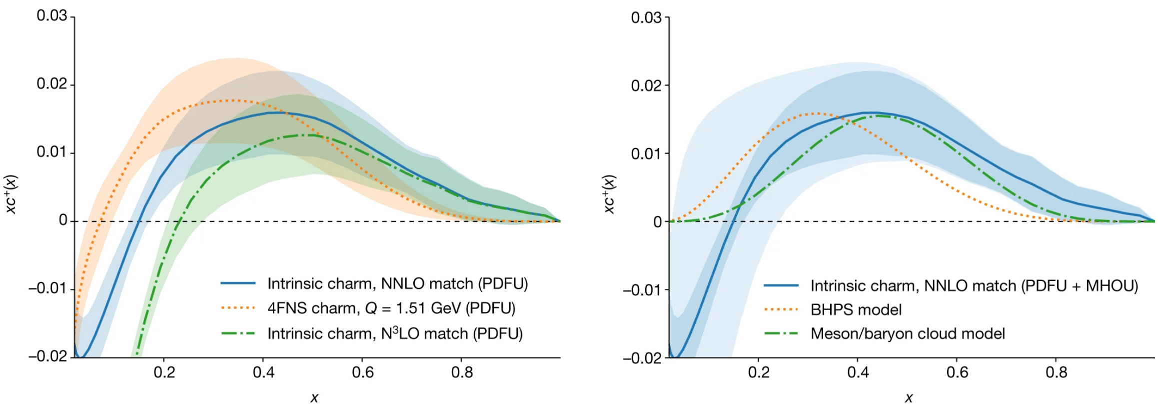


Figure 2.14: Comparison of intrinsic charm PDF with PDFs that also contain radiative charm components at different matching orders (left) and comparison with different intrinsic charm models, the BHPS and the meson-baryon cloud model (right). The charm PDF on the right is consistent with an IC model, in that it is compatible with 0 at lower x and has a peak roughly around 0.4-0.6 [14].

NNPDF was not the first collaboration to include intrinsic charm in their PDF: in this thesis, the CTEQ14 PDF will also be mentioned. The main difference between the two is that in the CTEQ14 PDF the intrinsic charm component was manually added as predicted by the BHPS, the model presented by Brodsky in [11], whereas in NNPDF the intrinsic charm component was fit in the PDF.

2.5 Introduction to jets and hadron physics

2.5.1 Hadronisation and jets

A QCD process can be divided in different phases starting from the parton interaction:

1. Hard process: this is the actual parton interaction, as seen in the Feynman diagram. It is perturbative and the matrix element can be calculated.

2. Parton shower: additional processes of quark or gluon emissions.

3. Hadronisation: non-perturbative phenomenon in which the free quarks combine to form hadrons and, subsequently, jets.

4. Underlying events, multi-particle events, pile-up: extra interactions that do not come from the primary parton-parton interaction.

Since hadronisation is non-perturbative, the events have to be reconstructed phenomenologically, via jet algorithms that will be described in Chapter 4. Through heavy meson identification, the flavour of the parton that originated the jet can be inferred.

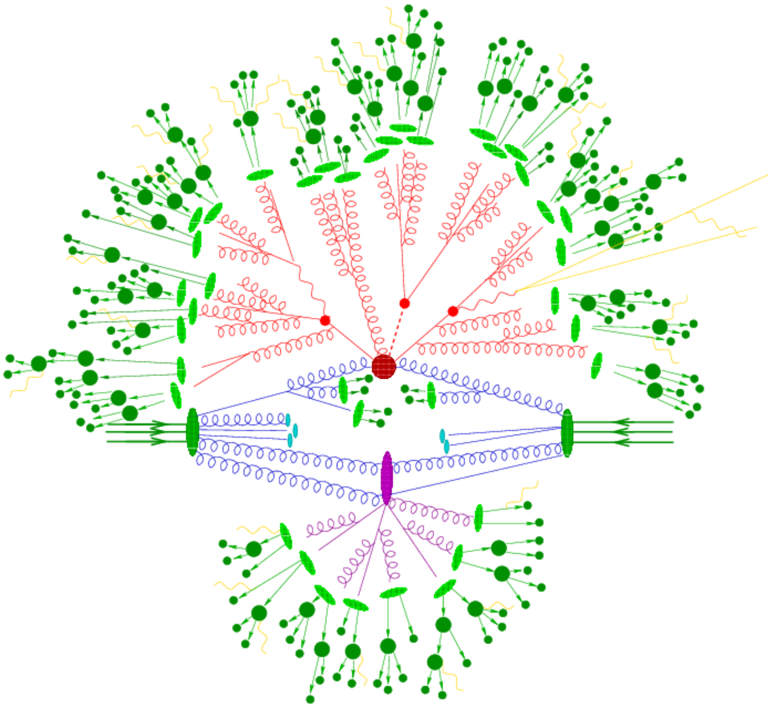


Figure 2.15: Sketch of a hadron-hadron interaction [15]

2.5.2 Heavy mesons

The heavy mesons used in flavour identification are B and D mesons. They are bound $q\bar{q}'$ states, in particular B mesons contain one b-quark and D mesons one c-quark. Due to the different masses of the two quarks, these mesons have different masses and lifetimes, which is why they can be easily distinguished and used to identify the flavour of the jet. An overview of the two families can be seen in Tables 2.2 and 2.3. Both D and B mesons undergo oscillations in which they change into their own antiparticle.

Meson	Quark content	Charge [e]	Mass [GeV]	Lifetime [s]
D^+	$c\bar{d}$	+1	1.87	$1.04 \cdot 10^{-12}$
D_0	$c\bar{u}$	0	1.86	$4.10 \cdot 10^{-13}$
D_s^0	$c\bar{s}$	0	1.97	$5.00 \cdot 10^{-13}$

Table 2.2: D mesons

Meson	Quark content	Charge [e]	Mass [GeV]	Lifetime [s]
B^+	$u\bar{b}$	+1	5.28	$1.64 \cdot 10^{-12}$
B_0	$d\bar{b}$	0	5.28	$1.52 \cdot 10^{-12}$
B_s^0	$s\bar{b}$	0	5.37	$1.52 \cdot 10^{-12}$
B_c^+	$c\bar{b}$	+1	6.27	$0.51 \cdot 10^{-12}$

Table 2.3: B mesons

2.5.3 Monte Carlo Generators and QCD calculations

Perturbative quantum field theory is the best way to solve QCD interactions. The basis of the calculations are Matrix Elements (MEs), which are the quantum mechanics amplitudes for any given process and to which differential cross sections are proportional. They are actually a perturbative series, which can be known within a certain order: the first term is the leading order (LO), and the following are next-to-leading order (NLO), next-to-next-to-leading order (NNLO) and so on. At tree level, ME calculations are usually fairly easy. The reality is quite different: in an actual QCD interaction, real-emission corrections, also called *legs*, and virtual corrections, or *loops*, must be introduced. At all orders, the matrix element differential cross section for a certain variable $\mathcal{O}$ for a final state F is:

$$\left.\frac{d\sigma_F}{d\mathcal{O}}\right|_{ME} = \sum_{k=0}^{\infty} \int d\Phi_{F+k} \, |\sum_{\ell=0}^{\infty} \mathcal{M}_{F+k}^{(\ell)}|^2 \delta(\mathcal{O} - \mathcal{O}(\Phi_{F+k})) \tag{2.40}$$

The sum over k takes into account real-emission corrections (which, in Feynman diagrams, appear as extra legs in the initial and final states, hence the name), while the sum over ℓ

represents loops. If $k = 0$ and $\ell = 0$, it stands for the leading order calculation. If $\ell = 0$ but $k = n$, it is the LO for the final state $F + n$ jets. If $k + \ell \leq n$, it is the N^{n}LO calculation. Multileg Monte Carlo generators like Madgraph or Sherpa perform these ME calculations, which, however, have no physical meaning unless paired with a parton shower generator, like Pythia or Alpgen, whose output are truth-level events, which means they contain information on the particles as they come out of the interaction. The main issue ME generators are faced with is double counting: whenever different order calculations are performed, the same physical process might be counted several times, leading to an overestimation of its contribution. This can be taken care of with matching techniques, which ensure that the combination of different-order calculations and parton showers happens smoothly and without double counting. There are several matching schemes, like CKKW or FxFx, that mostly rely on three techniques, slicing, subtraction and unitarity, as seen in [16].

2.5.4 Flavour schemes

Figure 2.16: Z+b-quarks in 4 and 5 flavour scheme, respectively: the first diagram gives rise to one b-jet, the second to two.

All quarks are not born equal, and their vastly different masses have a profund impact on proton physics. Quarks are usually distinguished between heavy and light. Heavy quarks are defined as those whose mass is such that the strong coupling constant $\alpha_s(m^2)$ at their mass scale is in the perturbative region. This is the case for bottom and top quarks.

Calculations involving b-quarks can be performed in two separate ways, depending on whether the quark mass is considered negligible or not. The first is the massive, or 4-flavour (4FNS) option. Since b-quarks are more massive than the proton, they can only be produced in pairs through gluon splitting at high Q^2. There are no b-partons involved and it is, in a sense, an effective theory in which the b-quarks do not have a PDF of their own (they are "integrated away") and do not contribute to the calculation of the running coupling constant.
The second approach is massless, and is called the 5-flavour (5FNS) scheme. This approach is valid whenever the energy scale is higher than the b mass. Higher order perturbative corrections may introduce logarithmic terms of the type $\log\frac{Q^2}{m_b^2}$ which lead to ultraviolet divergences. Assuming that the b-quark mass is negligible at this energy scale, logarithmic

terms can be resummed with DGLAP equations into a distribution function for initial states (the PDF of the b-quark) and in perturbative functions for final states.

Chapter 3

The Large Hadron Collider and the ATLAS experiment

3.1 The Large Hadron Collider

The Large Hadron Collider (LHC) is a two-ring superconducting hadron accelerator and collider. It is located at the border between France and Switzerland, near Geneva, on the site of CERN, the European Center for Nuclear Research. Its construction started in the early 2000s in the circular tunnel that once hosted the LEP (Large Electron Positron) collider. Inside the collider, which has a 26.7 km circumference, two types of particles are accelerated: protons and, for a shorter period of time during the year, heavy ions. The four major experiments operating on the ring are ATLAS, CMS, ALICE, and LHCb. Collisions happen at their locations, also called Interaction Points (IP).
LHC is characterised by periods during which it is fully operational, the so-called Runs, and periods during which it undergoes reparations and upgrades, the Long Shutdowns. Run 1 started in 2011, with an integrated luminosity of 30 fb^{-1} and center of mass (COM) energy of up to 8 TeV, and ended in 2013 with the Long Shutdown 1. Run 2 began in 2015 with much higher center of mass energy (13 TeV) and luminosity (190 fb^{-1}). After the second Long Shutdown, which started in 2019 and during which all major experiments underwent several upgrades, in 2022 LHC began colliding again, with COM energy 13.6 TeV energy and luminosity 450 fb^{-1}. As of summer 2023, LHC is in Run 3 phase, which is set to last until 2025 and will be followed by an important update to LHC machinery that will be discussed in Section 3.1.3

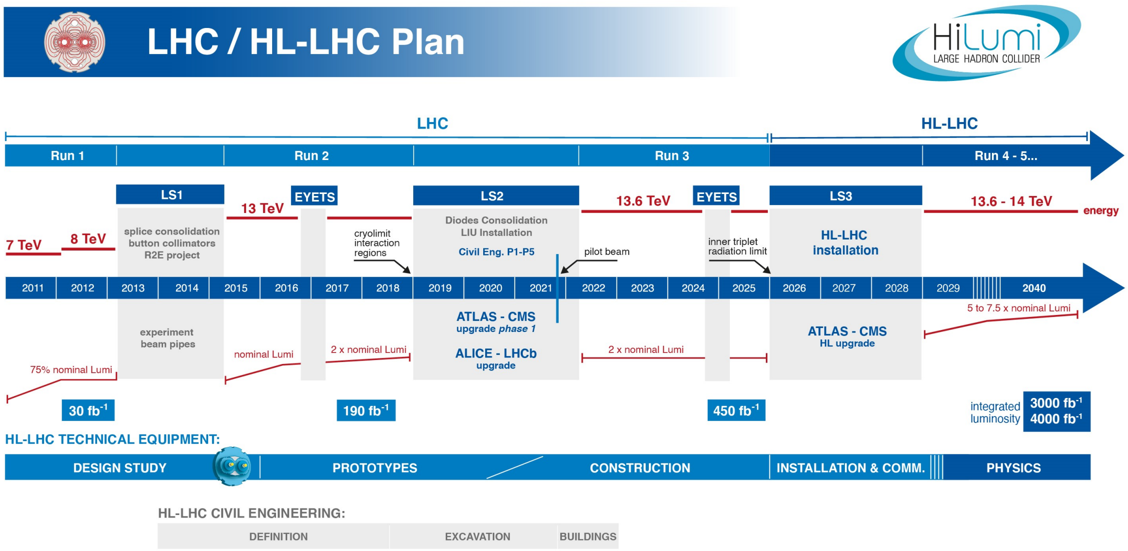


Figure 3.1: Timeline of LHC data taking, with center of mass energy on top of each running year: Run1 in 2011-2012, Run2 (which is the dataset analyzed in this thesis) in 2015-2018, the on-going Run3 and the planned High-Lumi LHC.

3.1.1 Structure and performance

The main accelerator ring, the *actual* LHC, is too big to be the only accelerating structure. Instead, protons (i.e. hydrogen atoms stripped of electrons) undergo a series of successive accelerations, that begin in LINAC 2, a linear accelerator that injects them in the Proton Synchrotron Booster (PSB). They are later accelerated up to 1.4 GeV and enter the Super Proton Synchrotron, from where they are injected into the main ring. Protons travel in bunches of $\mathcal{O}(10^{11})$ and have an interaction rate of 140 MHz.

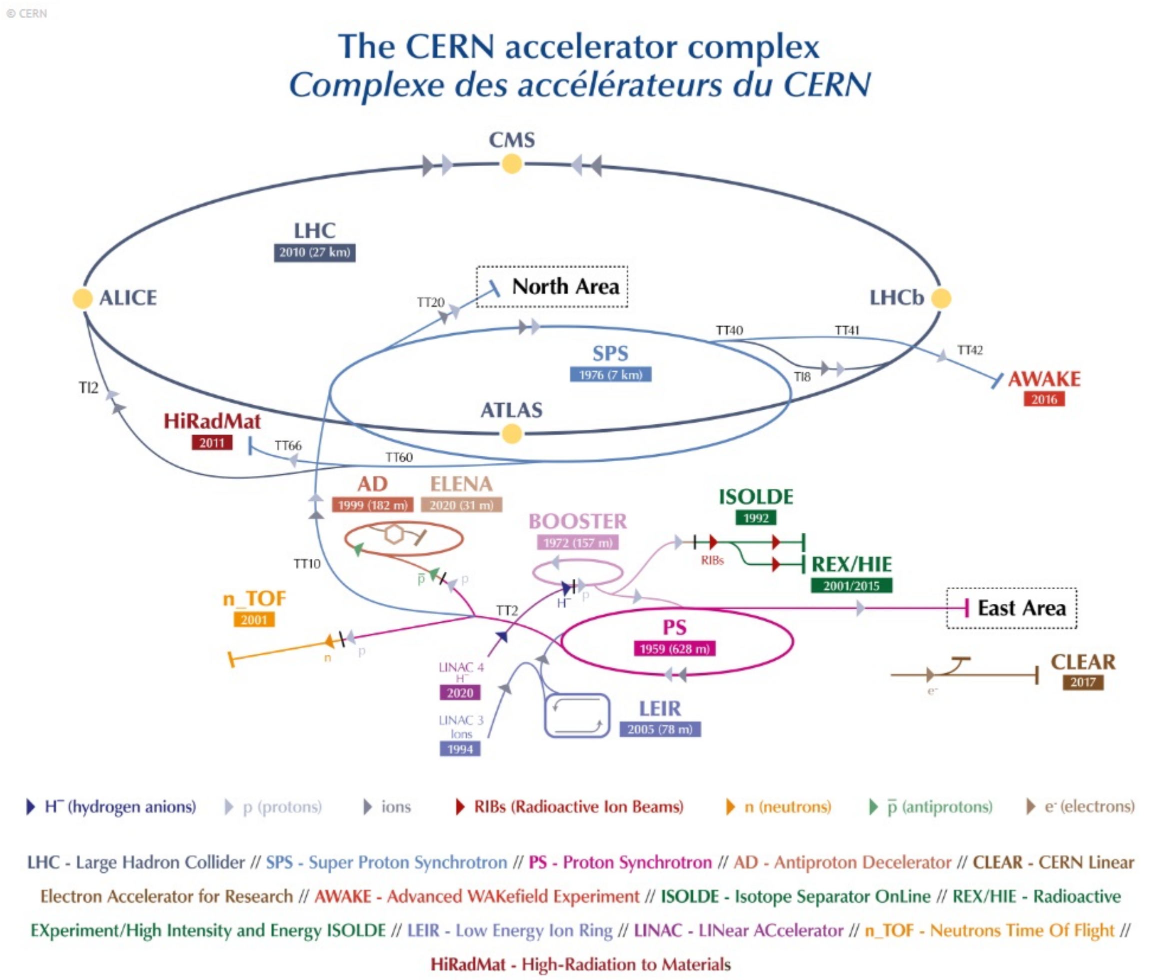


Figure 3.2: CERN Accelerating Complex

The acceleration of the particles is performed along the rings by radiofrequency (RF) cavities and superconducting magnets.
Radiofrequency cavities are metallic chambers that contain an electromagnetic field oscillating at 400 MHz. The field transfers energy to the protons, nudging them forward and accelerating them. The periodic nature of RF cavities is also responsible for separating the beam in bunches. There are 8 cavities for each beam, and they are all installed in the same region of the ring.
Superconducting magnets [17] are present in LHC as dipole or quadrupoles. Some of them can generate fields of up to 8.3 T. These types of magnet need to operate at temperatures as low as 1.9 K, which is why LHC is the coldest place on Earth. Inside the magnets, there are superconducting cables made of NbTi, inside which runs a current of tens of thousands of Ampères. Dipole magnets guide the beam along the collider, keeping it stable and

aligned with the desired trajectory. They are installed on $\frac{2}{3}$ of the circumference: there are 1232 dipole magnets, each 15 meters long and weighing 35 tons. The rest of the ring has quadrupole magnets, whose role is tightening the beam, making the bunches more compact before collision. After impact, more magnets are responsible for decreasing the energy of the proton and for dumping the used beams.

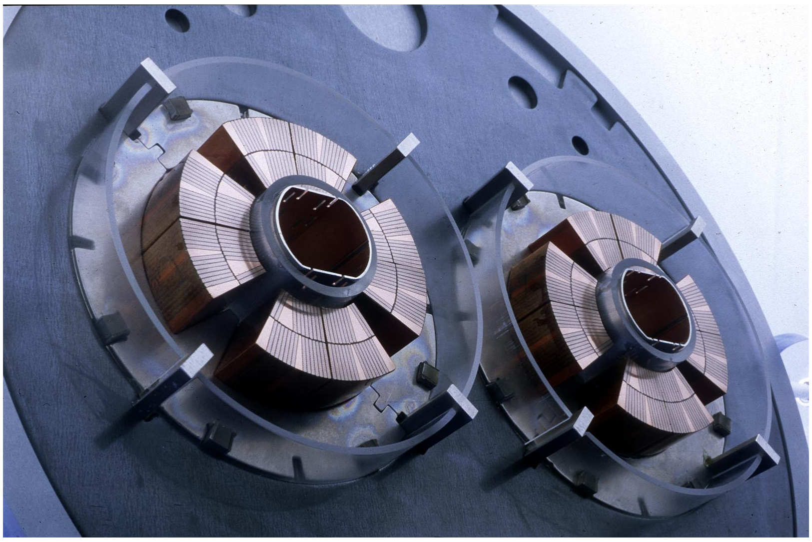


Figure 3.3: Cross section of a quadrupole superconductor used in LHC

As of Run2 [18], LHC has reached **center of mass energy** 13 TeV, which means that each beam has an energy of 6.5 TeV, since COM energy is defined as

$$s = \left(\sum_{i=1}^{2} E_i\right) - \left(\sum_{i=1}^{2} p_i\right) \tag{3.1}$$

The goal of a particle physics analysis is comparing experimental data to theoretical predictions, which are given as cross sections. Since the formula for the number of signal events of a certain process with final state x, taking into account the acceptance A of the detector and the efficiency ε of final state reconstruction, is

$$N_{x,obs} - N_{x,bkg} = A_{x,sig}\ \varepsilon_{x,sig}\ \sigma_{i \to x} \int \mathcal{L}(t)\ dt \tag{3.2}$$

the instantaneous **luminosity** of the colliding beams has to be accurately known in order to perform the analysis. It is a function of the number of particles colliding n, the collision rate f and, assuming the beams collide head-on and have a Gaussian profile, their root mean square along the x- and y-axis σ

$$\mathcal{L} = f \frac{n_1 n_2}{4\pi\sigma_x\sigma_y} \tag{3.3}$$

The maximum instantaneous luminosity achieved by LHC in Run 2 has been 2.01 $\cdot 10^{-34}$ cm^{-2} s^{-1}.

3.1.2 Experiments

LHC is the largest and most powerful accelerator ever built, so it is no surprise that it hosts numerous and varied detectors. ATLAS, the largest of the nine and the experiment in which this thesis was conducted, will be described in Section 3.2. The remaining eight are:

- **CMS** (Compact Muon Solenoid) is the second largest. It is a general purpose detector that deals with several branches of research, from Standard Model to dark matter and supersymmetry;
- **LHCb** specializes in studying the asymmetry between matter and anti-matter by focusing on the b-quark;
- **ALICE** (A Large Ion Collider Experiment) uses heavy ion collisions to study a primordial state of matter, the quark-gluon plasma, recreating extreme temperature conditions similar to those of instants after the Big Bang;
- **LHCf** simulates cosmic rays by using particles escaped from collisions;
- **Totem** (Total, Elastic And Diffractive cross section Measurements) studies proton physics by analysing them after the collisions;
- **MoEDAL** (Monopole and Exotics Detector At the LHC) is specialized in Beyond Standard Model Physics. It searches for instance for magnetic monopoles and gravitons;
- **SND@LHC** (Scattering and Neutrino Detector) is one of the two neutrino-based experiments at LHC. It studies collider muon neutrinos.
- **FASER** (Forward Search Experiment) studies weakly interacting particles like neutrinos, and searches for dark matter and supersymmetric particles.

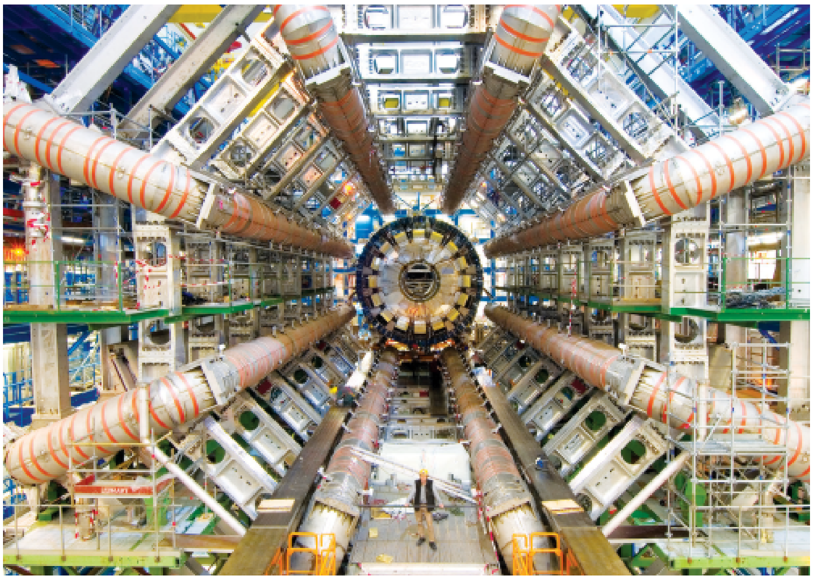


Figure 3.4: The ATLAS experiment

3.1.3 High-Luminosity Upgrade

The Large Hadron Collider machinery will undergo a period of major upgrade during the second half of the 2020s, the High-Lumi upgrade. The goal is increasing the luminosity to probe rarer events. It will increase more than five times, reaching over 4000 fb^{-1} of integrated luminosity in 10 years. This will allow detectors to collect more data than ever, probing rarer events and furthering their studies in SM and BSM physics.

3.2 The ATLAS experiment

ATLAS (A Toroidal LHC Apparatus) is the largest experiment operating in LHC. It is a general-purpose detector, which means that it is equipped to detect almost all kinds of particles. The research conducted there is varied: from precise Standard Model to search for new, beyond Standard Model Physics.
The detector has the shape of a cylinder 46 m long and with a diameter of 25 m. It weighs 7000 tonnes and it lays 100 m deep in a cavern near the town of Meyrin, Switzerland. As of 2023, more than 6000 scientists from 47 countries participate in the collaboration.
ATLAS has a concentric structure: layers of detectors wrap around the interaction point, each with its own function of registering the momentum, the energy, or the position of the particles. The construction of the detector finished in 2008, but it only became fully operational in 2009. During Run 2, it recorded the record instantaneous luminosity of $1.2 \cdot 10^{34}$ fb^{-1} [19] and accumulated 140 fb^{-1} worth of data, as seen in Figure 3.5.

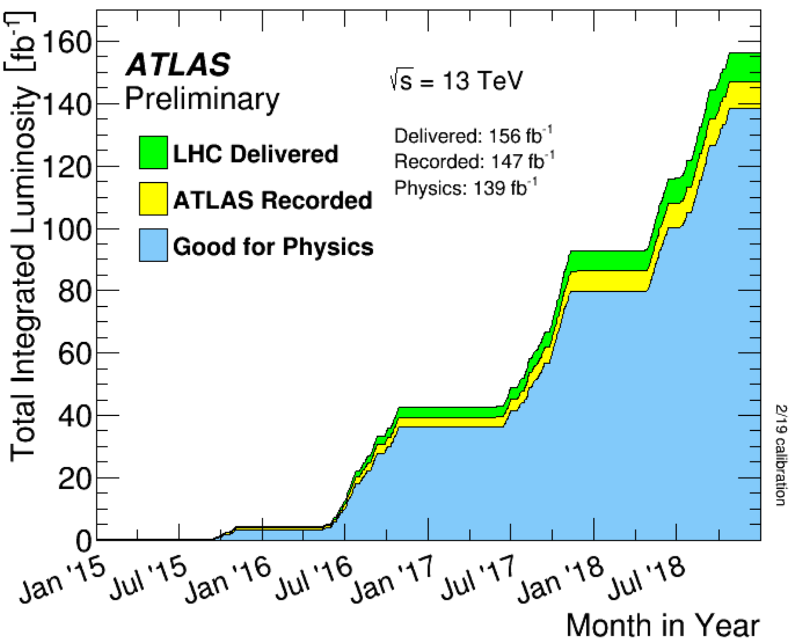


Figure 3.5: Integrated luminosity during Run 2: the plot shows histograms of the luminosity delivered by LHC (green), collected by the ATLAS experiment (yellow) and good for physics (blue) from 2015 to 2018 [19].

LHC has almost full coverage in the solid angle around the Interaction Point, that is it has full forward-backwards symmetry. Coordinates are defined as seen in Figure 3.6: it is a right-hand system with origin in the IP. The z-axis is then parallel to the beam.

Pseudorapidity is defined as

$$\eta = -\ln\left[\tan\left(\frac{\theta}{2}\right)\right] \tag{3.4}$$

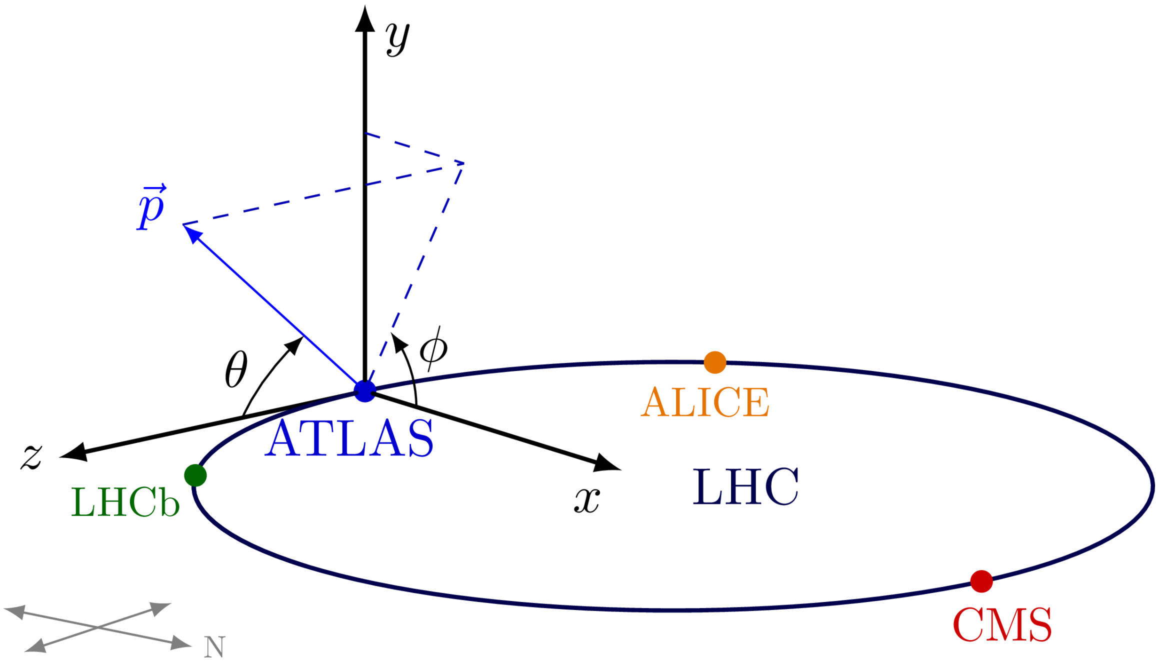


Figure 3.6: Definition of coordinates in the ATLAS experiment

The main components of the detector are described in the following Sections, including an overview of the Data Acquisition System.

3.2.1 Inner Detector

The inner detector is the first component of ATLAS that records charged particles after the collision. It is very compact: 6.2 m in length and 1.2 in height. Its goal is tracking momentum, charge, and directions of the particles thanks to a magnetic field of 2 T generated by the central solenoid. It is composed of three different detectors that have similar functions but different construction features. It covers a range of $|\eta| < 2.5$.

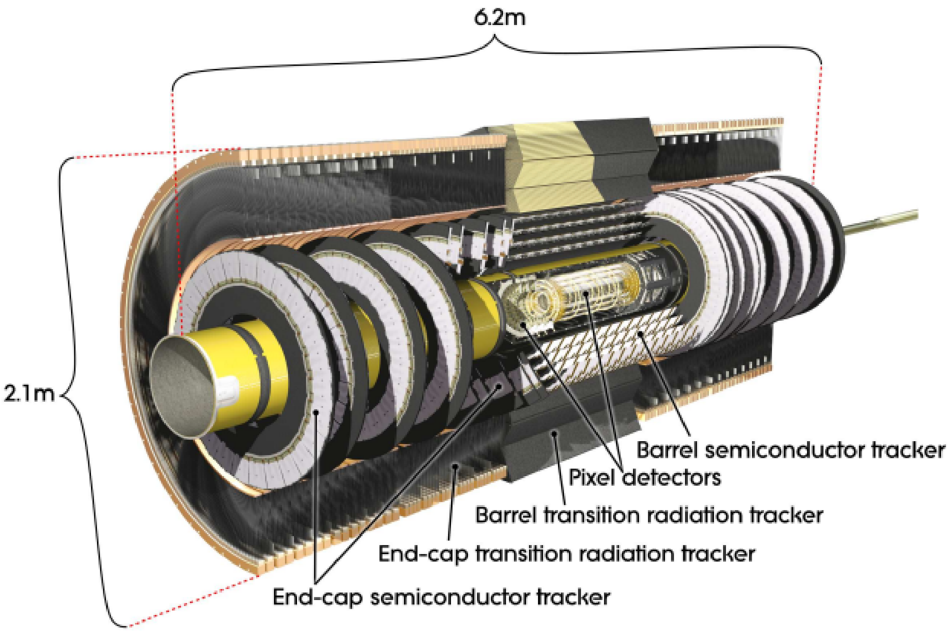


Figure 3.7: Cut-away of the inner detector [20]

The Pixel Detector is the closest to the interaction point, and it is composed of roughly 92 million pixels of size as small as 50×250 μm for the inner part, and 50×400 μm for the outer part. The pixels are mounted on 1736 modules with over 45 000 readout channels per module. The high granularity gives rise to extreme tracking precision. They are organised in 4 cylindrical layers and 3 disks per endcap.
The Semiconductor Tracker has the same role as the Pixel Detector, but it exploits silicon microstrips instead of pixels for detecting purposes; they are organised in 4 barrel layers and 9 disks per endcap. It contains over 6 million microstrips and has a sensitivity of 25 μm.
The Transition Radiation Tracker complements the other two detectors providing particle identification. It is composed of planes of straw tubes with a tungsten wire and exploits the ionization of gaseous Xenon to identify the radiating particle.

3.2.2 Calorimeters

Calorimeters measure the energy of the passing particle. There are two types of calorimeter in ATLAS: one that exploits liquid argon, the other scintillating plastic tiles. Both calorimeters use the sampling technique: they alternate the active material with layers of passive materials like lead, tungsten and steel. In the LAr calorimeter, the particles in the metal generate an electromagnetic shower that ionizes the liquid gas, generating a current that is dependent on their energy.

In the Tile Calorimeter, specialized in hadronic showers, the particles interact with the scintillating material producing photons which are later converted to an electrical current. It surrounds the LAr calorimeter detecting the particles that escape it. It is the heaviest component of the ATLAS detector, weighing more than 2900 tonnes. Most of the particles produced in the interaction are absorbed inside the calorimeter, except muons and neutrinos.

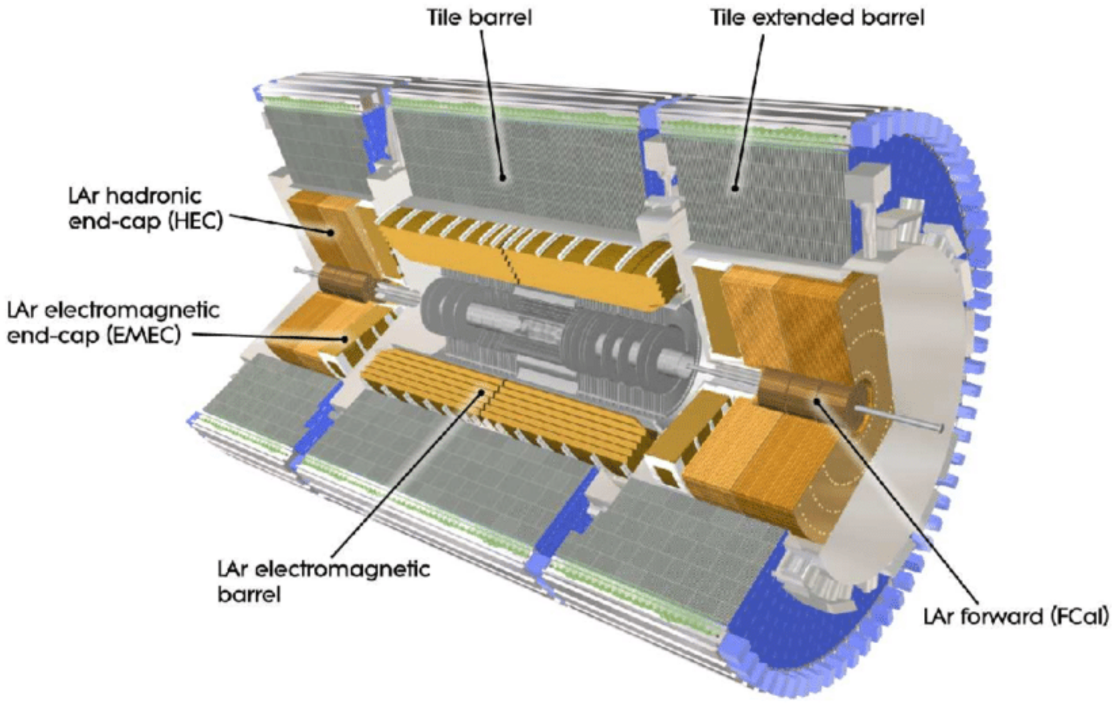


Figure 3.8: Cut-away of the calorimeter [20]

3.2.3 Muon system

Since muons do not generate showers, they leave the calorimeters mostly unscathed. For this reason, a muon spectrometer is installed to detect their tracks through magnetic deflection. The field is generated by two different sets of magnets: the barrel toroid for $|\eta| < 1.4$, two smaller endcap magnets for $1.6 < |\eta| < 2.7$ and a combination of the two in the middle region. The muon chambers wrap around the beam axis in the barrel region, and are orthogonal to it in the endcap regions. In most of the rapidity range, the chambers contain drift tubes, but at high $|\eta|$ they contain Cathode Strip Chambers.

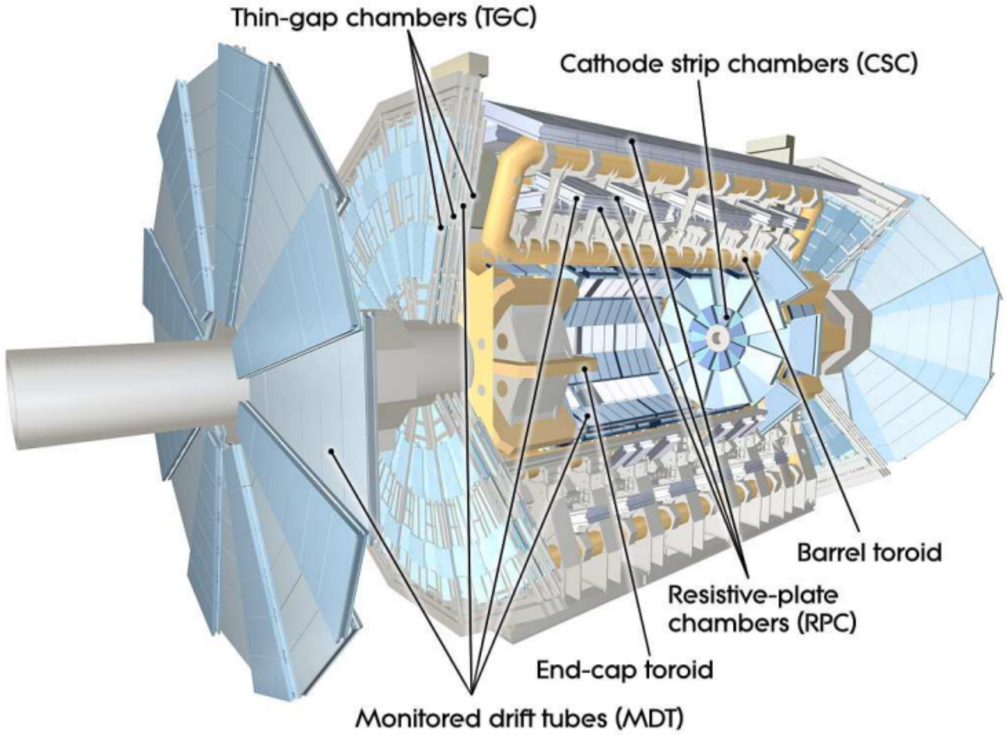


Figure 3.9: Cut-away of the muon system [20]

3.2.4 Forward Detectors

There are three smaller detectors that lie in ATLAS forward region:

- LUCID (Luminosity measurement using Cerenkov Integrated Detector), an online luminosity monitor;
- ALFA (Absolute Luminosity for ATLAS), another luminosity monitor that lies 1 mm from the beam and uses scintillating fibers;
- ZDC (Zero-Degree Calorimeter), employed in heavy ion collisions.

3.2.5 Trigger and data acquisition

The volume of data generated by interactions in ATLAS is immense: each proton-proton collision produces roughly 1.6 MB of data which, taking into account 40 MHz of beam crossing frequency means requiring 1 Pb/s of data acquisition. This is obviously impossible, both for storage reasons and for the incredibly high data flux for each channel. For this

reason, there are triggers inside the experiment whose purpose is deciding if the data that is being collected is worth saving. ATLAS employs a first level trigger, L1, which is hardware based, and a High Level Trigger which is software based.
The L1 trigger analyzes the information coming from coincidences of calorimeters and muon detectors to determine whether and event could be of interest. This manages to reduce the volume of data of a factor of 400. As seen in Figure 3.10, there are dedicated triggers for various types of particle. In Section 5, electron and muon triggers will be employed in the analysis. They are always labeled `EL` and `MU`.

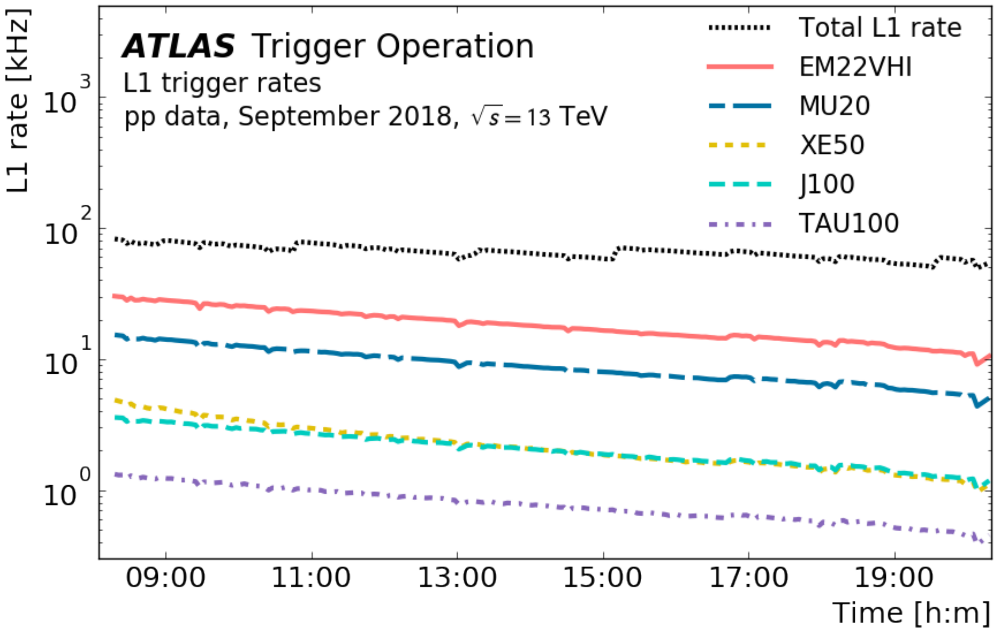


Figure 3.10: L1 Trigger [21]

The HLT further reduces the data of a factor 100 by using all the information from the detectors and a farm of more than 30 000 CPUs that has complete access to the data recorded by calorimeters, muon chambers, and internal trackers.

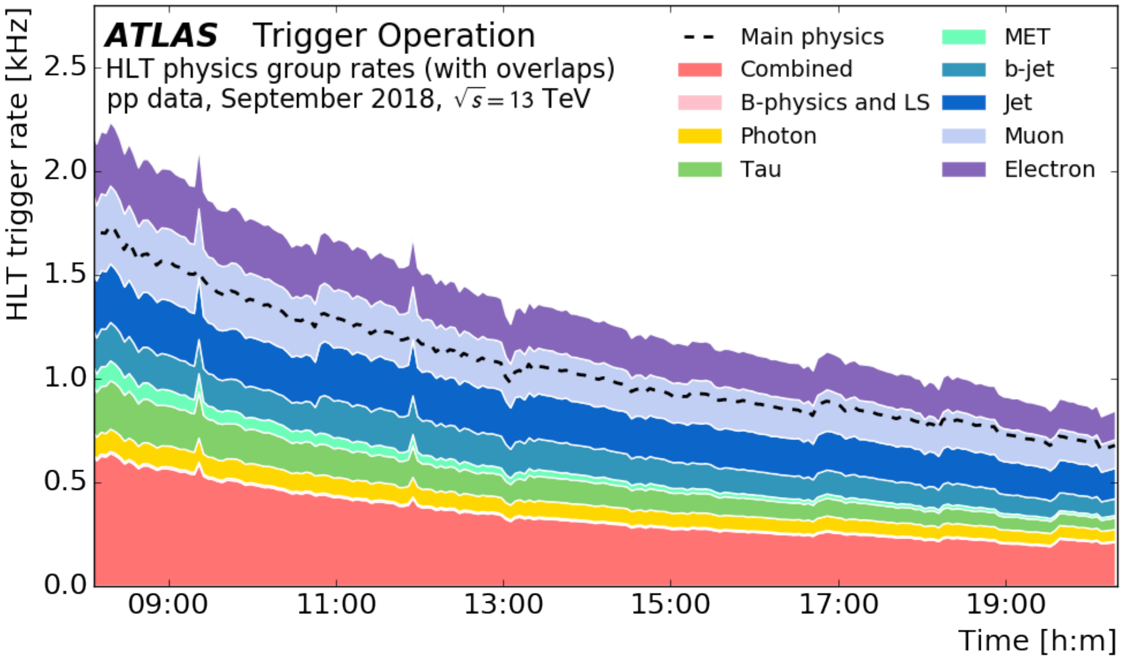


Figure 3.11: High level trigger [21]

At the end of the triggering system, of the billions of collisions only a few thousand remain: these events are channeled to storage by the Data Acquisition System, to be analyzed later. The offline analysis of this data is performed by the ATLAS Computing Service, and all memebers of the experiment have access to it thanks to the Worldwide LHC Computing Grid. The ATLAS collaboration uses a C++/Python based framework called Athena 21.0 which encompasses all the tools necessary for the offline analysis. Truth-level simulated events can go through a GEANT4 based detector simulation and data reconstruction, after which they are called reco-level data and can be compared to experimental data.

Chapter 4

Object reconstruction and identification

As described in Section 2.5.1, a hadronic collider is a very complex environment, in which millions of events happen at once. It is therefore of paramount importance to be able to correctly identify and recognize the events of interest. The final states at LHC are produced from parton interactions, or hard scattering, and, after undergoing the processes of parton shower and hadronization, manifest themselves to experimentalists as particles or, more often, collections of particles called jets. The real challenge is then recognizing the so-called **physical objects** from the signals in the various parts of the detector and being able to trace them back to the phenomena that originated them. In the following Sections, the techniques used by the ATLAS collaboration to identify jets and their flavour, leptons, and tracks and vertices in general will be discussed.

4.1 Tracks and vertices

All charged particles have their trajectory bent when immersed in a magnetic field. The ATLAS experiment exploits a solenoid magnet whose field is directed along the Z-axis combined with the ionization in the Inner Detector (ID) to determine the track of charged particles. The trajectory of a charged particle in a magnetic field is a helix, which is described by 5 parameters:

- d_0: the transverse impact parameter, which is the minimum distance from the z-axis;
- z_0: the longitudinal impact parameter;
- ϕ_0: azimuthal angle at minimum distance
- θ: polar angle
- q/p: curvature (it has the same sign as the charge of the particle);

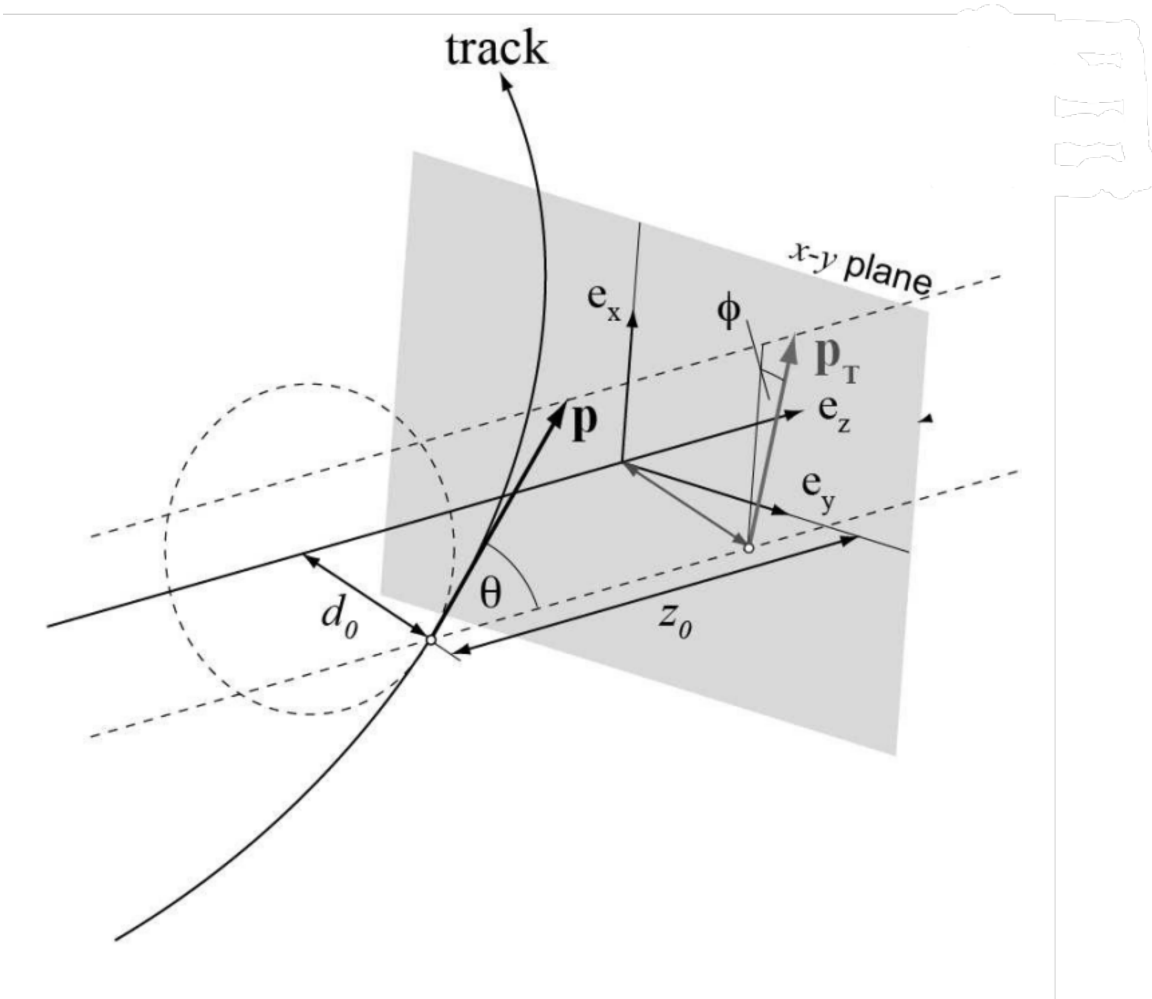


Figure 4.1: An overview of the helix parameters used in charged particles track reconstruction. The z-axis runs along the proton beam axis.

When passing in the ID, which is composed of silicon detectors, the particles deposit energy. The **hits**, which are the signals in the detector, are grouped in **clusters**, and **track seeding** identifies the vertex of origin of the object.
Once the vertex has been identified (the vertex with the most tracks pointing to it is called **primary vertex**) and there is a series of points that could belong to the track, the **track fitting** fits the trajectory parameters, taking into account the deflection from the magnetic field, Coulomb scattering and energy loss.

4.2 Leptons

Leptons are fundamental to every analysis because they are easy to unequivocally identify. They represent ≈ 10% of the weak bosons decays and, since they participate in a weak interaction, it is a very clear signature in an otherwise overwelming background of hadrons and jets produced in proton-proton collisions. Lepton identification relies on algorithms which use a combination of signals from various parts of the detector, because every lepton behaves differently and has a unique signal.

4.2.1 Electrons

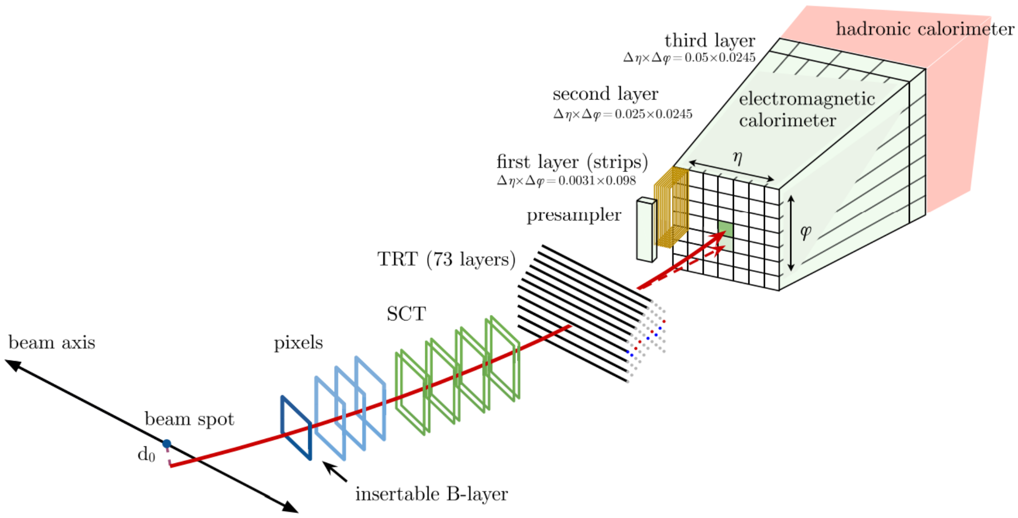


Figure 4.2: Path of an electron inside the ATLAS detector (red line). The dashed red line is a photon produced inside the tracker.

High-energy electrons are charged particles, which means they leave a track in the Inner Detector (ID) which can be reconstructed as seen in the previous Section, and in the calorimeter where they deposit energy via bremssttrahlung. Their reconstruction is articulated in three phases. **Seed-cluster** reconstruction is based on the clusters of energy deposited in the ECal. **Track reconstruction** exploits the techniques previously described to identify the tracks in the inner detector. The information from these two phases is combined in the **electron candidate** reconstruction, in which clusters and tracks are matched to identify the electron candidate.
Electron identification is carried out with a likelihood approach, by using a product of the distributions of the variables for signal (S) and background(B):

$$L_{S(B)}(\boldsymbol{x}) = \prod_{i=1}^{n} P_{s(B),i}(x_i) \tag{4.1}$$

and the optimal discriminant is defined as

$$d_L = \frac{L_S}{L_S + L_B} \tag{4.2}$$

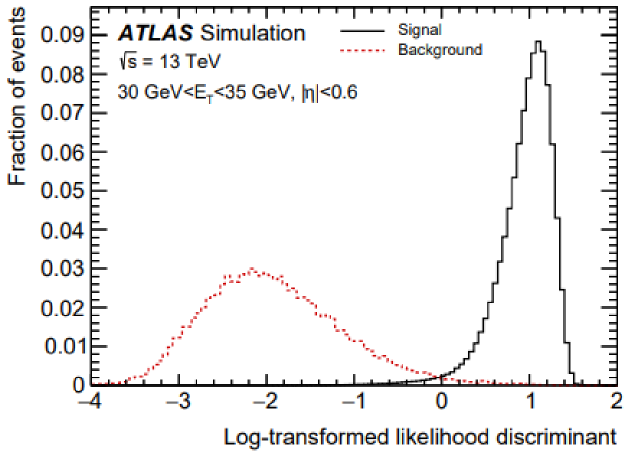


Figure 4.3: Transformed likelihood-based discriminant for reconstructed electron candidates. Signal (black histogram) is prompt electrons in a $Z \to ee$ simulation sample, background (red dashed line) is a generic two-to-two process simulation sample [22].

The signal are elctrons coming from the original interacton (i.e. not produced in secondary vertices), meanwhile background is composed of "false" electrons (which are other objects mimicking an electron signal), and electrons coming from interactions within the detector or hadronic decays.
Different analysis groups may have different requirements when it comes to background rejection, which is why the discriminant has different operating points: **Very-Loose**, **Loose** whose efficiency for electrons at $E = 30$ GeV is 98% and requires at least 2 hits in the pixel detector and seven in the whole ID, and **Medium** with 88% and **Tight** at 80%, which also both require hits in the innermost layer.

4.2.2 Muons

Muon identification relies mainly on the muon spectrometer (MS), the ID track and the energy they deposit in the calorimeters as minimum ionizing particles, or MIP. There are four main algorithms that use combinations of this information to identify muons.
In the **Combined** approach, the muon is reconstructed via a global fit that uses hits from both the tracker and the spectrometer. In the **Segment-tagged**, the ID track must be matched to at least a track segment in the muon chambers. In the **Calorimeter-tagged**, the ID track must be associated with hits in the calorimeter compatible with a MIP. Finally,

in the **Extrapolated**, the track is reconstructed only from the MS, as long as it's compatible with the vertex of origin.
As it is the case for electrons, muons are identified via a likelihood based approach, which has another four operating points: **Loose**, **Medium**, **Tight**, **High** p_T.

4.2.3 Isolation requirements

Leptons produced in weak interactions are independent from hadron events, which is why they are expected to have very little activity around them. Isolation variables, based on the amount of energy deposited in the cone around the lepton, are therefore an important feature for the identification of prompt leptons, because they increase the purity of the selection of those produced in the hard interaction against those coming from quark decays. The radius of the cone is defined as $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$.
Isolation can be calorimeter based, by looking at the sum of the transverse energies of a cluster inside a cone of $R = 0.2$, or track-based, in which the variable is the sum of transverse momenta in a cone of R = min (0.2, 10 GeV/E_T). As is the case for LH identification, isolation also has several operating points with different variables and constraints. Looser constraints can be helpful especially with background studies, meanwhile, tighter constraints produce purer samples, at the expense of efficiency.

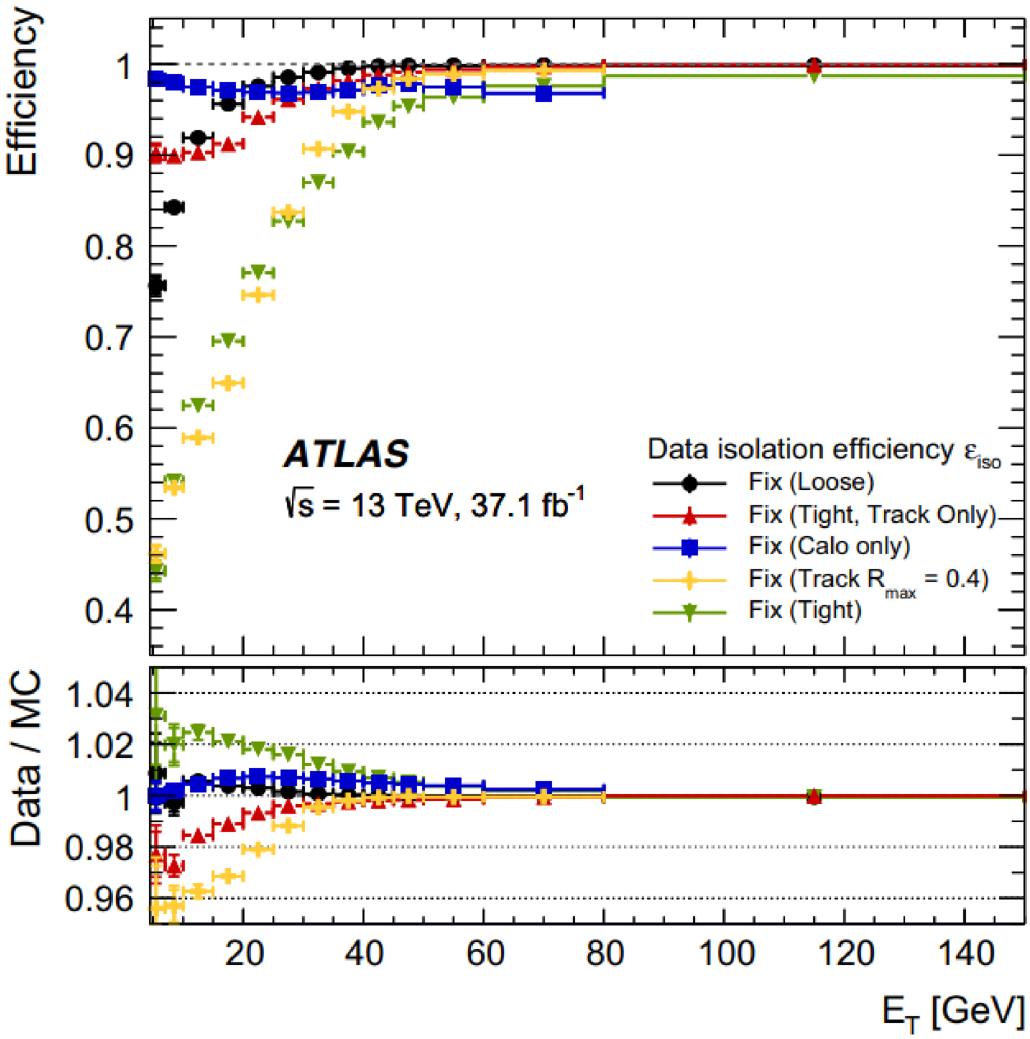


Figure 4.4: Data isolation efficiency for different operating points. Data from 2015-2016 LHC Run2 at $\sqrt{s} = 13$ TeV. [22]

4.2.4 Tau leptons

Tau leptons are a lot trickier to identify, because their lifetime is very short. They can decay both leptonically and hadronically, thanks to their large mass. In the first case, the electrons and muons from the decay might be confused for prompt leptons, but there is evidence of missing energy because the decay involves two neutrinos. In hadronic decays, mostly pions are produced: the cone is similar to a jet, but more collimated. Tau leptons will not be considered further in this analysis.

4.3 Jets

As seen in Section 2.5.1, the products of a QCD interaction are jets of particles. The goal is being able to trace back the jets to the parton-parton interaction that produced them. As seen in Figure 4.5, a single event in the ATLAS experiment is very complex, which is why one of the biggest challenges is the reconstuction and identification of jets, which relies heavily on algorithms.

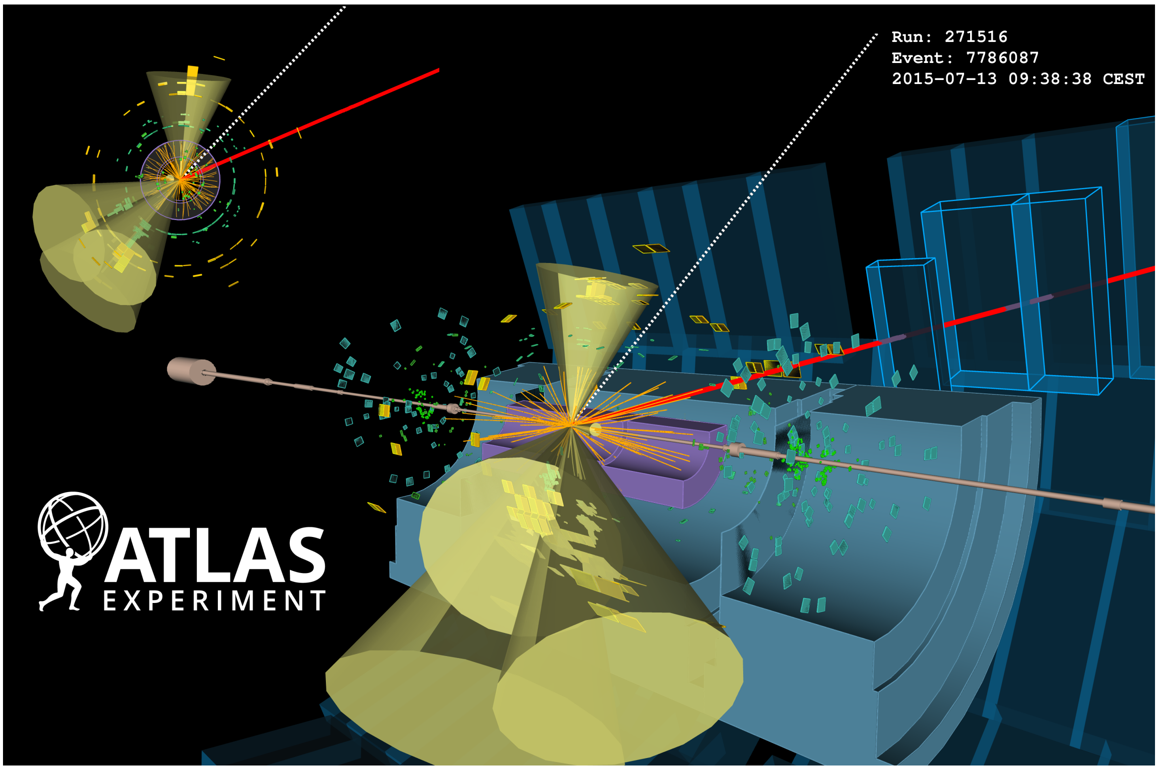


Figure 4.5: Display of a candidate boosted top quark pair production event from proton-proton collisions recorded by ATLAS with LHC stable beams at a collision energy of 13 TeV. [23]

4.3.1 Jet Reconstruction

The so-called **jet finding** algorithms have the task of grouping together detected particles that have a common origin, in order to understand the physical properties of their source. Every algorithm uses different procedures and variables and must be calibrated accordingly. However, there are phenomenological restraints to these algorithms that must be respected in order to have results that are comparable to theoretical calculations: **infrared** and **collinear** safety (IRC). In the first case, soft radiation must not affect the outcome of the reconstruction, in the second, reconstruction must not be different if there are more collinear particles with smaller impulse with respect to fewer particles with higher impulse.

Mathematically speaking, an observable $\mathcal{O}$ is IRC safe if it is invariant in the limit of collinear splitting or soft emission.

$$\mathcal{O}(X; p_1, ..., p_n, p_n + 1 \to 0) = \mathcal{O}(X; p_1, ..., p_n) \tag{4.3}$$

Algorithms can be top-down, which means they draw stable cones by identifying regions of particles with the correct variables. This is the case for **cone algorithms** [24] like the Iterative Seed Fixed Cone Jet Algorithm,or SISCone. However fast, these types of algorithm are not IRC safe, because they often rely on variables such as transverse impulse. Bottom-up algorithms are a better choice when it comes to IRC safety. The most important ones are **Recursive Recombination Jet Algorithms** [25], which are used in the ATLAS experiment. They iteratively measure the distance between pairs of particles (i,j) and compare them to a characteristic distance for a single particle:

$$d_{ij} = \min(d_i, d_j) \frac{\Delta R^2_{i,j}}{R^2} \tag{4.4}$$

$$d_i = p_T^2 n \tag{4.5}$$

If the minimum is d_{ij}, the particles are grouped together to form a single object in the list; if it is d_i the particle becomes a jet and is removed from the list. R is a parameter of the algorithm and is usually $R \approx 0.4 - 0.7$ (in ATLAS, $R = 0.4$). If $n = 1$ the algorithm is called k_T **algorithm**, if $n = 0$ it is the **Cambridge Aachen algorithm**, if $n = -1$ it is the **anti-k_T algorithm**. Different algorithms may produce different jet configurations, as seen in Figure 4.8

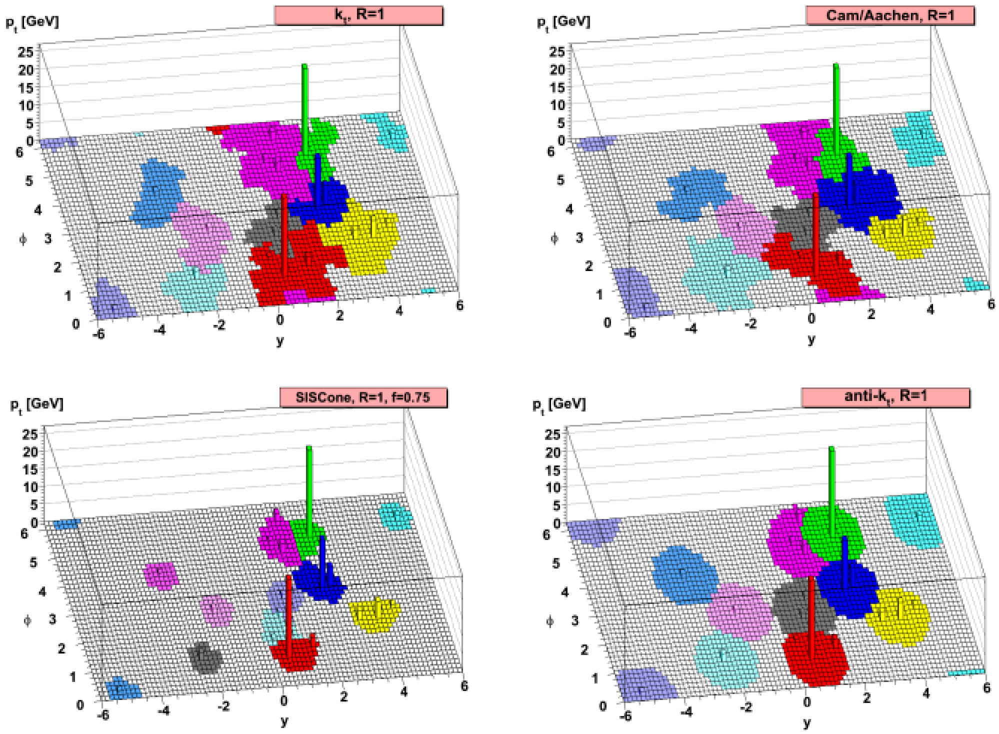


Figure 4.6: Different jet reconstruction outputs by 4 algorithms. [26]

All algorithms use as input **ParticleFlow** objects, which are lists of particles obtained by dedicated algorithms that combine information from the tracker and the calorimeter.

4.3.2 Flavour tagging

Once the jet has been reconstructed, it is necessary to find out which parton originated it. Flavour tagging allows experimentalists to categorize jets in three flavours: b-jets, from b-quarks, c-jets from c-quarks and light jets from any other parton. Jets coming from b- and c-quarks contain heavy B and D mesons, which are used to identify them. As described in Section 2.5.2, these mesons have longer lifetimes and higher masses, which means that their decays happen after a longer time, in a displaced location that identifies a secondary vertex which can be as far from the primary vertex as some cm for B mesons. Their decays can be identified because they have higher p_T with respect to the jet-axis. Also, at the end of the chain of decays that they trigger there is often a non-isolated lepton.

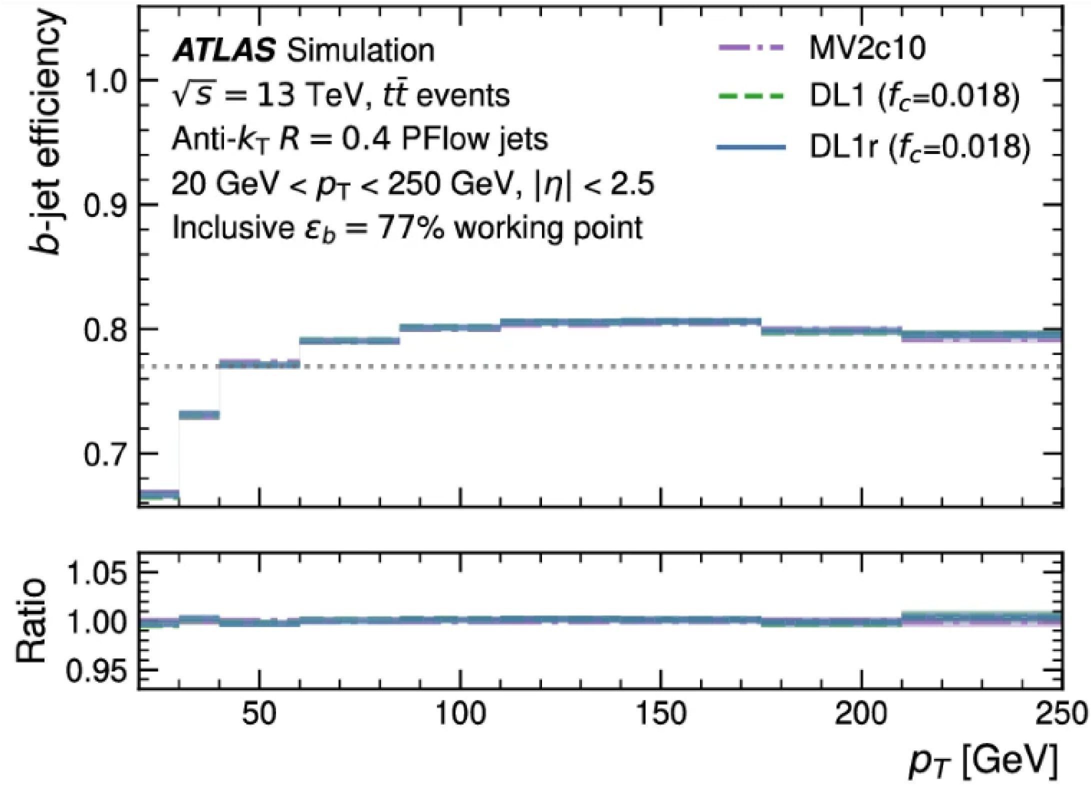


Figure 4.7: Run 2 b-tag efficiency by three different algorithms as a function of p_T [27].

Flavour tagging algorithms take as input kinematic variables from the particles in the jet, like secondary vertices locations and impact parameters, and combine them to properly identify the flavour of the jet. Some algorithms, like the Soft Muon Tagging, also look at low-momentum, non-isolated leptons for the identification of hadronic decays.

The same algorithm is always used on both reco-level data and experimental data, which is why some calibration has to be performed in order for the efficiency of the algorithm to match in the two cases. This is achieved through scale factors (SF): the efficiencies for the simulated data are corrected bin-by-bin:

$$SF = \frac{\varepsilon_{data}}{\varepsilon_{MC}} \tag{4.6}$$

Scale factors are computed separately by calculating tagging efficiencies and mistagging rates for MC $t\bar{t}$ samples and a $t\bar{t}$ rich and pure data sample. The efficiencies affected concern lepton triggers, lepton selection, jet vertexing and flavour tagging.
The performance of algorithms is usually evaluated using single-cut operating points (OPs). The discriminant distributions used by the flavour tagging algorithm are divided in *pseudo-continuous* bins identified by increasing efficiency percentage. Different OPs offer trade-

offs between efficiency and mistagging rates: higher efficiency implies higher chances of misidentification, while lowering efficiency means improving purity. This technique is called pseudo-continuous b-tagging (PCBT).

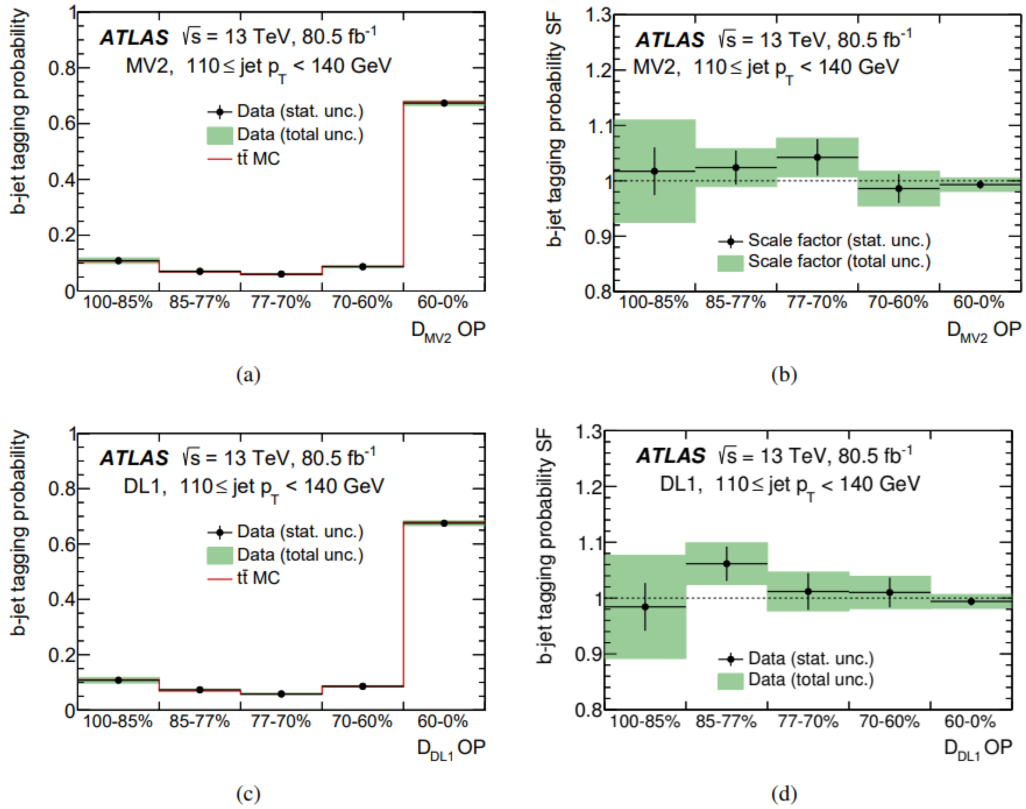


Figure 4.8: b-jet tagging probability (a,c) and simulation-to-data SFs (b,d) for two algorithms, MV2 (a,b) and DL1 (c,d) with pseudo-continuous binning. Vertical error bars are statistical uncertainties, while green bands also include systematic uncertainties.

4.4 Missing transverse energy

Whenever particles are produced in an interaction, but not leave a signal in the detector, there is an imbalance in the energy and momentum in the transverse plane, where they should otherwise add to 0, because negligible momentum is present in the transverse direction of the collision. This is the case whenever a neutrino is produced (in leptonic decays, for example for the W boson). It could also be the manifestation of some BSM phenomenon, which is why missing energy reconstruction is very important. Since the full kinematics of the parton-parton interaction is not known, only the transverse component can be inferred, by looking at the conservation of momentum on the plane that is transverse to the beam axis. An estimate of the missing transverse energy is performed by listing all the reconstructed object, separating between hard and soft interactions.

$$\mathbf{E}_T^{miss} = -\sum_{\text{selected electrons}} \mathbf{p}_T^e - \sum_{\text{selected photons}} \mathbf{p}_T^\gamma - \sum_{\text{accepted } \tau} \mathbf{p}_T^\tau - \sum_{\text{accepted jets}} \mathbf{p}_T^{jet} - \sum_{\text{unused soft tracks}} \mathbf{p}_T^{tracks} \tag{4.7}$$

There is an important contribution to the soft term from pile-up interactions, which are secondary collisions from protons in the same bunch as the ones from the main collision event.

Chapter 5

The Z boson and heavy flavours

In this thesis, the focus was on the production of a Z boson in association with heavy flavour jets (Z+HF), i.e. b- or c-tagged jets. These types of final states are of interest especially because they represent a test for QCD. The analysis has been ongoing for a couple of years by a team within the ATLAS collaboration on the full Run2 dataset, with the goal of extracting inclusive and differential cross sections. Z+HF production can be used to probe the proton structure, to study pertubative QCD predictions and compare Monte Carlo event generators at different order of perturbative accuracy and evaluated with different calculation schemes. Z+b-jets also constitutes a background for other analyses, especially those involving the Higgs boson, whose primary decay is in two b-quarks. Furthermore, the most important feature for this thesis are final states involving c-jets, because they will be fundamental to analyze the charm component of the PDF. In particular, the final state which includes only one c-jet, as seen in Figure 5.1, is a direct way of probing the charm PDF. Results involving intrinsic charm are extensively described in the next chapter. The fiducial measurement of the Z+HF differential cross section proceeds according to the following steps, which will be described in later sections of this Chapter:

- identification of events in the phase space region of interest through cuts in p_T and flavour selection;
- estimate and subtraction of the main backgrounds using a partially data-driven technique that allows for different fractions of Z+c-jet, Z+b-jet and Z+l-jet to be extracted from data using MC templates;
- extraction of cross section measurements with statistical unfolding, removing efficiency effects and detector resolution.

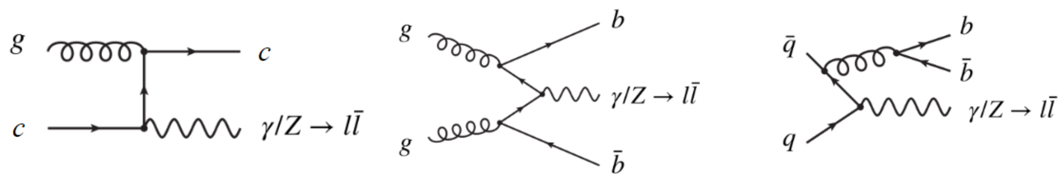


Figure 5.1: Examples of leading order (LO) and next-to-leading order (NLO) Feynman diagrams for Z production.

In an analysis, events must be accurately selected. In Section 5.4, the criteria used for the Z+heavy flavour analysis will be reported and discussed, and in Section 5.3 the simulation process will be described and the data sampled listed. Backgrounds and systematics are discussed in Sections 5.5 and 5.8 respectively, and results are presented in Section 5.9.

5.1 State of the art

Several measurements of both inclusive and differential cross sections of Z+heavy jets production have been made before this one by various collaborations: at $\sqrt{s} = 1.96$ TeV there are the CDF [28] and D0 [29] collaborations; at $\sqrt{s} = 7$ Tev, $\sqrt{s} = 8$ TeV and $\sqrt{s} = 13$ TeV (with a partial data-set from Run2) the ATLAS and CMS collaborations [30–33].

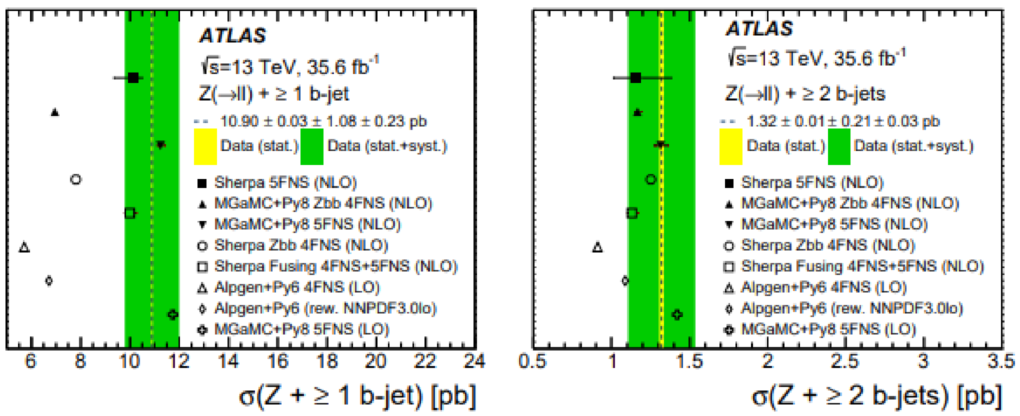


Figure 5.2: cross section results: comparison between data and several MC predictions from a Run 2 partial ATLAS dataset for $Z + b$ and $Z + bb$ at $\sqrt{s} = 13$ TeV [32]

5.2 Measurements performed and fiducial phase space

The Z+HF measurement is performed in a phase space region defined through p_T and rapidity cuts on final state leptons and jets, but also through flavour selection (according to hadrons within the jets). Table 5.1 describes the criteria used in event selection: fiducial phase space is defined first through kinematic criteria on leptons produced in the decay, which must be of opposite sign (OS), and jets, and later through flavour constraints, thus identifying 3 fiducial regions: `Zge1B_truth` for events with $\geq 1b$-jet, `Zge2B_truth` for events with $\geq 2b$-jets, and `Zge1C_truth` for events with ≥ 1 c-jet but no b-jets. A b-jet ($c-$jet) is tagged if there is a B (D) hadron within $\Delta R = 0.3$ of the center of the jet, where $\Delta R = \sqrt{(\Delta\phi)^2 + (\Delta y)^2}$. The following differential distributions are reconstructed in the identified phase space:

1. Jet multiplicity, jet properties, Z and Z-jet distributions, which are used in QCD predictions, PDF models sensitivity studies and MC event generators modeling of $Z + b$-jets:
   - Z+b-jet production cross section as a function of the inclusive jet multiplicities for 1 and 2 b-jets: $\mathrm{N}_{b-\mathrm{jets}} \geq 1{,}2$
   - differential Z + b-jet production cross section for events with $\mathrm{N}_{b-\mathrm{jets}} \geq 1$ as a function of
     (i) p_T of the leading b-jet;
     (ii) p_T of the Z boson;
     (iii) x_F variable, which is defined as $x_F = 2p_T|sinh(\eta)|/\sqrt{s}$;
   - differential Z+b-jet cross section as a function of the difference in radius ΔR between the Z boson and the leading b-jet
2. Jet multiplicity, jet properties, Z and Z-jet distributions, which are used in QCD predictions, PDF models sensitivity studies and MC event generators modeling of $Z + c$-jets:
   - Z+c-jet production cross section as a function of the inclusive jet multiplicities for 1 c-jet: $\mathrm{N}_{c-\mathrm{jets}} \geq 1$
   - differential Z + c-jet production cross section for events with $\mathrm{N}_{c-\mathrm{jets}} \geq 1$ as a function of
     (i) p_T of the leading c-jet;
     (ii) p_T of the Z boson;
     (iii) x_F variable;
   - the ratio of the differential $Z + c$-jets production cross section as a function of the p_T of the Z boson, for events with $\mathrm{N}_{c-\mathrm{jets}} \geq 1$, measured in two ranges of Z rapidity, defined as "central" if $|y| < 1.2$ and "forward" if $|y| > 1.2$. It is expected to be sensitive to intrinsic charm.

3. Di-b-jets distributions used in QCD predictions or in topologies sensitive to Higgs and BSM physics:

   - the differential Z+b-jet production cross section as a function of the invariant mass m_{bb} of the 2 leading b-jets for events with $N_{b-\text{jets}} \geq 2$;
   - the differential Z+b-jet production cross section as a function of the azimuthal angle difference between the 2 leading b-jets $\Delta\Phi_{bb}$ for events with $N_{b-\text{jets}} \geq 2$.

Object selection	Acceptance cuts
$Z \to \mu\mu$	2 OS μ with $p_T > 27$ GeV, $\eta < 2.5$, $m_{\mu\mu} = 91$ GeV ± 15 GeV
$Z \to ee$	2 OS e with $p_T > 27$ GeV, $\eta < 2.5$, $m_{ee} = 91$ GeV ± 15 GeV
b-jet	$p_T > 20$ GeV and $\lvert y \rvert < 2.5$, ΔR(b-jet, ℓ)> 0.4
c-jet	$p_T > 20$ GeV and $\lvert y \rvert < 2.5$, ΔR(c-jet, ℓ)> 0.4
Event selection	**Acceptance cuts**
$Z \geq 1b$-jet (`Zge1B_truth`)	Z and 1 b-jet
$Z \geq 2b$-jet (`Zge2B_truth`)	Z and 2 b-jet
$Z \geq 1c$-jet (`Zge1C_truth`)	Z and 1 c-jet

Table 5.1: Fiducial regions for measurements at particle level.

5.3 Monte Carlo samples and data

Experimental data used in this thesis was recorded by the ATLAS detector during Run2 and consists of a total integrated luminosity of 140 fb^{-1} with an uncertainty of 0.8% of $p-p$ collisions at $\sqrt{s} = 13$ TeV.
The simulated samples used in this thesis were all produced using either the `MadGraph5_AMC@NLO` or `SHERPA 2.2.11` programs. MadGraph is a parton-level generator that produces matrix elements, which are the probability amplitudes associated with all the possible Feynman diagrams of the parton in the hard interactions, before the hadronisation process. The NLO calculations were performed in 5FS and 4FS. The NNLO calculations were performed in 5FS, and the different jet multiplicities were merged via the FxFx [34] prescription. The PDFs used in the generations were `NNPDF 3.1 NLO`, `NNPDF 3.1 NLO 4FS`, `NNPDF 3.1 NLO 3FS`. Three series of samples were generated and later merged, `MC_16_a`, `MC_16_d` and `MC_16_e`. The three samples correspond to three phases of data taking: 2015-2016, 2017, and 2018 respectively. It is important to differentiate the years of data taking, because to each correspond different detector conditions, luminosities (as seen in Figure 5.3) and pile-up collision profiles, which are the secondary proton-proton collisions that happen in the same bunch as the primary interaction.

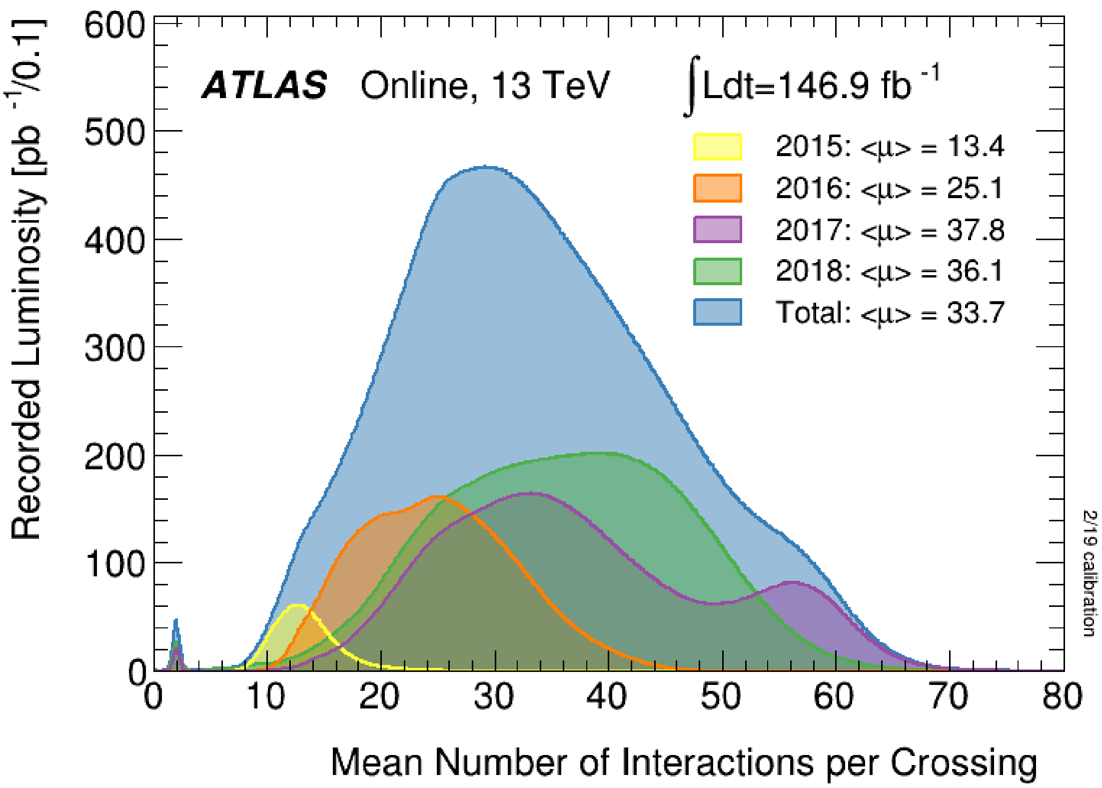


Figure 5.3: Plot of the different luminosities and mean interactions per bunch crossing in the different years of data-taking (Run2).

Reweights with other PDFs will be discussed in the next chapter. The output of a Madgraph simulation is a Les Houches file `.lhe` which contains the output of the parton-level interaction. In order for the results to have physical meaning the hadronisation and parton shower processes must be simulated. The Pythia8 [35] interface takes care of these two steps. The output is an `HEPMC` file (High Energy Physics Monte Carlo). To analyze the results, the RIVET toolkit [36] has been used: it is a truth-level analysis, which means that is the data not affected by detector effects, or reconstruction or experimental uncertainties unless explicitely addedd.

5.3.1 Theory uncertainties

A contribution to systematic uncertainties, which will be discussed later, comes from the Monte Carlo generators. There are two kinds of theoretical uncertainties: QCD scale uncertainties and PDF uncertainties. The former come from the perturbative nature of QCD and they are based on the 7-point variations of two parameters: the renormalization scale, μ_R, which is a free parameter that intervenes at the energy scale at which the strong coupling constant α_S is fixed, and the factorization scale, μ_F, which is used to determine the scale at which QCD goes from perturbative to non-perturbative. PDF uncertainties correspond to the Root Mean Square of the replica sample of the PDF which was chosen in the generator and α_S uncertainties. However, PDF uncertainties were shown to have negligible impact on the analysis, and were therefore not considered in the final results.

5.4 Selection criteria, object definition, signal

The Z boson is identified through its decay in a lepton-antilepton pair, because it is the easiest channel to reconstruct due to the absence of jets. Most events in a proton-proton collision only have jets in the final state, because they have a much higher cross section in QCD interactions. Selecting events with a leptonic signature means then reducing by many orders of magnitude the rate of events that are not of interest for the analysis. Both data and simulated predictions have to respect a strict set of rules in order to be admissible for the analysis. Furthermore, the data has to come from events that have not had issues in neither the liquid argon, nor the tile calorimeter systems, no errors in the inner detector, and are not incomplete. They have to respect a combination of High Level Triggers (Tab. 5.2) that depend on the flavour of the Z-generated lepton.

Muon channel	
2015	HLT_mu20_iloose_L1MU15 HLT_mu20_iloose_L1MU15
2016-2018	HLT_mu26_ivarmedium HLT_mu50
Electron channel	
2015	HLT_e24_lhmedium_L1EM20VH HLT_e60_lhmedium HLT_e120_lhloose
2016-2018	HLT_e26_lhtight_nod0_ivarloose HLT_e60_lhmedium_nod0 HLT_e140_lhloose_nod0 HLT_e300_etcut

Table 5.2: High Level Triggers on lepton identification: the name of the triggers identifies whether it is for electrons `e` or muons `mu`, then the next number is the p_T threshold. If the name has an `L1` label, it indicates the threshold at L1 trigger for the leptons. The labels that follow `lh` are the likelihood based levels of identification at trigger level, whereas `i` indicates the isolation requirements.

Collision vertices are identified through two ID tracks with $p_T > 500$ MeV. The vertex with the highest total squared transverse momentum is the primary vertex (PV).

5.4.1 Object reconstruction

As seen in Chapter 4, object reconstruction relies on algorithms and discriminants to identify leptons. Electrons must have ID tracks matched to clusters in the EM calorimeter, and come from the PV. In order to identify the vertex of origin, conditions on the longitudinal and transverse impact parameter are set. There are also constraints on the electromagnetic shower shape, the energy released in it, and the tracks within a cone of $\Delta R = 0.2$ of the presumed electron.

Muons do not shower in the EM calorimeter, so their identification relies on the muon spectrometer. It is performed either by matching tracks in the inner detector (ID) to tracks in the spectrometer, or by using standalone tracks in the latter. They must be isolated, and have similar requirements on impact parameter and tracks in the nearby cone. Both leptons have to have a $p_T > 27$ GeV and have also requirements on pseudorapidity.

Muon channel	
ID	Medium
Isolation	`FCTight`
Vertex association	$\lvert d_0 sign\rvert < 3$ and $z_0 \sin\theta < 0.5$ mm
p_T	27 GeV
η	$\lvert\eta\rvert < 2.5$
Electron channel	
ID	Tight
Isolation	`Tight_VarRad`
Vertex association	$\lvert d_0 sign\rvert < 5$ and $z_0 \sin\theta < 0.5$ mm
p_T	27 GeV
η	$\lvert\eta\rvert < 1.37$ or $1.52 < \lvert\eta\rvert < 2.47$

Table 5.3: Lepton Selection Criteria, as described in Section 4.2: for ID and isolation the operating points of the discriminant are given. Vertex association relies on the conditions on track parameters,as seen in Figure 4.1. The gap in electron rapidity $|\eta|$ is due to the crack region between endcap and barrel calorimeters.

Jets	
Algorithm	anti-k_t, R = 0.4, `AntiKt4EMPFlowJets`
p_T	20 GeV
η	$\lvert\eta\rvert < 2.5$
JVT	`Tight`
***b*-tagging**	Cut on DL1r discriminant with 85% efficiency
Rejection rate for $c-$jets	3 (corresponding to a selection efficiency of $\approx$ 30%)
Rejection rate for light jets	40

Table 5.4: Jet Selection Criteria: JVT is the Jet Vertex Tagger, which matches the jet to the primary vertex.

Jets are harder to reconstruct: they come from PFlow objects and are reconstructed using the anti-k_T algorithm. The requirements are $p_T > 20$ GeV, rapidity $|y| < 2.5$, ΔR > 0.4. There are stricter requirements for jets coming from pileup vertices. The b-tagging efficiency is roughly 85% and b-jets are required to have $p_T > 20$ GeV and pseudorapidity $|\eta| < 2.5$.

5.4.2 Flavour classification and scale factors

Flavour tagging uses cone-based criteria to assign b,- c- or light jets labels: a b-jet is classified as one if there is at least one b-hadron with $p_T > 5$ GeV within a cone of $\Delta\text{R} = 0.3$. The check is analogous for c-hadrons. If neither check is positive, the jet is labelled as *light*. Simulated jet events are used to define classifications: events with only one b-jet are called $Z + b$, then there are $Z + bb$ who have two. Events with 1 or 2 c-jets and no b-jets are $Z + c$ and $Z + cc$, and then there are $Z + l$, the remaining ones. These categories are then used to define signal and background regions. Since the efficiencies of reconstruction and classification may differ between simulations and actual data, the former are corrected using scale factors. The efficiencies affected concern lepton triggers, lepton selection, jet vertexing and flavour tagging.

5.4.3 Signal region

Events with a Z-boson are selected through the di-lepton channel. They must have exactly two oppositely charged leptons of the same flavour that respect the aforementioned requirements of Table 5.3. Furthermore, their invariant mass must be in the range 76 GeV $< m_{\ell\ell} <$ 106 GeV. There are additional selection cuts on missing transverse energy if the Z has a lower transverse impulse, because of the significant $t\bar{t}$ background: it has to be $E_T^{miss} < 60$ GeV if the transverse impulse is $p_T < 150$ GeV. A summary of the parameters can be found in table 5.5. Signal regions are defined based on the presence of b-jets, as summarized in Table 5.6. In the first signal region, `ZGeq1B`, both $Z + c-$jets and $Z + b-$jets are measured, taking first the former and then the latter as background. In the second region, the signal is composed of just $Z + 2b-$jets.

Z selection	
Leptons	exactly 2 of the same flavour and opposite charge
Di-lepton mass	76GeV $< m_{\ell\ell} <$ 106GeV
Missing transverse energy	$E_T^{miss} < 60$ GeV if $p_T < 150$ GeV

Table 5.5: Parameters for Z boson selection through its leptonic decay

Signal regions	
`ZGeq1B`	at least 1 good b-jet
`ZGeq2B`	at least 2 good b-jets

Table 5.6: Definition of signal regions based on jet flavour.

5.5 Backgrounds

The primary source of background in this analysis are jets of unwanted flavours, which correspond to light jets in all cases, and jets of c (b)-flavour whenever studies on b(c)-jets are made. These backgrounds are taken care of with the flavour fit method, described in Section 5.5.2. The rest of the background contribution comes from weak interactions (single and diboson) and top quarks. In general, background evaluation is performed either by using Monte Carlo samples or through a data-driven approach. In the data-driven approach, which was used to estimate $t\bar{t}$ and multijet backgrounds, instead of relying solely on theoretical predictions, a control region is chosen in the data (it has different parameters than the signal region, so that the background of interest dominates) and it is enriched with the background process until it makes up for roughly 90% of the contribution. A transfer factor is then used to scale the control region to the signal region. This approach reduces the theoretical uncertainties that come from generators. Another data-driven approach, that uses template fits, is performed to estimate multijet background contributions. This contribution comes from lepton fakes that contaminate the signal selection, and is very difficult to simulate. In this analysis, multijet background contributions were found to be negligible.

5.5.1 Monte Carlo background estimates

The MC estimates of background and signal samples are in Tables 5.7 and 5.8 (percentage) and 5.9 and 5.10 (total yield). In the tables, the signal $Z + ee$ or $Z + \mu\mu$ is not yet divided by jet flavour.

Sample	$Z \geq 1b$−jet MC16a	MC16d	MC16e	$Z \geq 2b$−jet MC16a	MC16d	MC16e
$Z \rightarrow \mu^+\mu^-$	97.6	97.84	97.81	84.36	85.97	85.84
Top	1.7	1.51	1.53	13.54	12.04	12.17
$Z \rightarrow \tau\tau$	0.02	0.01	0.01	0.0	0.01	0.01
VV $\rightarrow \ell\ell/\ell\nu/\nu\nu + q\bar{q}$	0.62	0.58	0.59	1.71	1.62	1.62
VH	0.06	0.06	0.06	0.38	0.36	0.36

Table 5.7: MC estimates for signal and background (percentage), $Z \rightarrow \mu^+\mu^-$. The first line represents the signal, while the three column represent the three Monte Carlo samples from the different periods of data taking, as seen in Section 5.3

Sample	$Z \geq 1b$−jet MC16a	MC16d	MC16e	$Z \geq 2b$−jet MC16a	MC16d	MC16e
$Z \rightarrow e^+e^-$	97.36	97.58	97.54	83.49	85.04	84.71
Top	1.89	1.71	1.74	14.34	12.88	13.19
$Z \rightarrow \tau\tau$	0.01	0.0	0.01	0.0	0.01	0.01
VV $\rightarrow \ell\ell/\ell\nu/\nu\nu + q\bar{q}$	0.67	0.64	0.64	1.75	1.67	1.69
VH	0.07	0.07	0.07	0.42	0.39	0.4

Table 5.8: MC estimates for signal and background (percentage), $Z \rightarrow e^+e^-$.The first line represents the signal, while the three column represent the three Monte Carlo samples from the different periods of data taking, as seen in Section 5.3

Sample	$Z \geq 1b-$jet MC16a	MC16d	MC16e
$Z \to \mu^+\mu^-$	664821 ± 1467	8732556 ± 1678	1133170 ± 1991
single-top	986.4 ± 11.3	1163.4 ± 12.3	1534.5 ± 14.1
ttbar	11323.0 ± 39.4	12977.2 ± 42.2	17247.6 ± 48.7
$Z \to \tau\tau$	67.8 ±12.2	66.6 ±14.3	66.6 ±14.3
VV $\to \ell\ell/\ell\nu/\nu\nu + q\bar{q}$	4612.8 ± 14.4	5548.7 ± 14.8	7346.4 ± 20.9
ZH	192.9 ± 0.4	227.7 ± 1.3	302.7 ± 0.5

Sample	$Z \geq 2b-$jet MC16a	MC16d	MC16e
$Z \to \mu^+\mu^-$	41725 ± 298	54643 ± 338	70484 ± 390
single-top	221.5 ± 5.3	249.1 ± 5.7	327.9 ± 6.5
ttbar	6237.0 ± 29.2	7154.3 ± 31.3	9492.8 ± 36.1
$Z \to \tau\tau$	0.6 ±2.0	1.3 ± 2.3	1.3 ± 2.3
VV $\to \ell\ell/\ell\nu/\nu\nu + q\bar{q}$	817.3 ± 5.4	1004.7 ± 6.9	1318.9 ± 8.3
ZH	99.2 ± 0.3	118.1 ± 1.0	156.4 ± 0.3

Table 5.9: MC estimates for signal and background (total yield), $Z \to \mu^+\mu^-$. The first line represents the signal, while the three column represent the three Monte Carlo samples from the different periods of data taking, as seen in Section 5.3

Sample	Z $\geq 1b-$jet MC16a	MC16d	MC16e
$Z \rightarrow e^{+}e-$	383517 ± 1072	485377 ± 1226	634467 ± 1417
single-top	645.7 ± 9.0	727.3 ± 9.8	960.2 ± 11.3
ttbar	7254.9 ± 31.2	8193.7 ± 33.7	10921 ± 39.1
$Z \rightarrow \tau\tau$	48.9 ±10.6	46.6 ±12.4	46.6 ±12.4
VV $\rightarrow \ell\ell/\ell\nu/\nu\nu + q\bar{q}$	2877.2 ± 11.4	3416.5 ± 11.4	4491.0 ± 16.2
ZH	128.4 ± 0.3	149.4 ± 1.1	199.0 ± 0.4

Sample	Z $\geq 2b-$jet MC16a	MC16d	MC16e
$Z \rightarrow e^{+}e-$	24759 ± 215	31746 ± 248	42218 ± 228
single-top	129.9 ± 4.0	167.4 ± 4.7	206.5 ± 5.3
ttbar	4011.1 ± 23.2	4544.5 ± 25.1	6049.9 ± 291
$Z \rightarrow \tau\tau$	6.2 ±2.2	4.7 ± 2.4	4.7 ± 2.4
VV $\rightarrow \ell\ell/\ell\nu/\nu\nu + q\bar{q}$	508.7 ± 4.2	615.4 ± 5.4	810.8 ± 6.6
ZH	66.4 ± 0.2	78.3 ± 0.8	104.5 ± 0.3

Table 5.10: MC estimates for signal and background (total yield), $Z \rightarrow e^{+}e^{-}$. The first line represents the signal, while the three column represent the three Monte Carlo samples from the different periods of data taking, as seen in Section 5.3

5.5.2 Flavour fits

Once both event selection and background rejection have been performed, flavour fits are necessary to to extract the background distributions in the signal region and improve the agreement between data and simulation, because the flavour fractions provided by the Monte Carlo are not always reliable. This is achieved via a maximum-likelihood bin-by-bin fit, for $Z+c$, $Z+b$ and $Z+l$ in the `Geq1BiJ` region 5.4, and $Z+bb$, $Z+b$, $Z+c$ and $Z+l$ in the `Geq2BiJ` region 5.5. The fits are performed with the jet PCBT output described in Section 4.3.2 on intervals of the variables of interest, for example in the [20,24] interval of p_T in Figure 5.5. In the same plot, the odd shapes come from the combined PCBT output of the two b-jets.

5.6 Detector level distributions

In this section, detector level distributions for the main variables of interest are presented. First, the inclusive jet distributions are shown, before the b-jets selection: the events in the

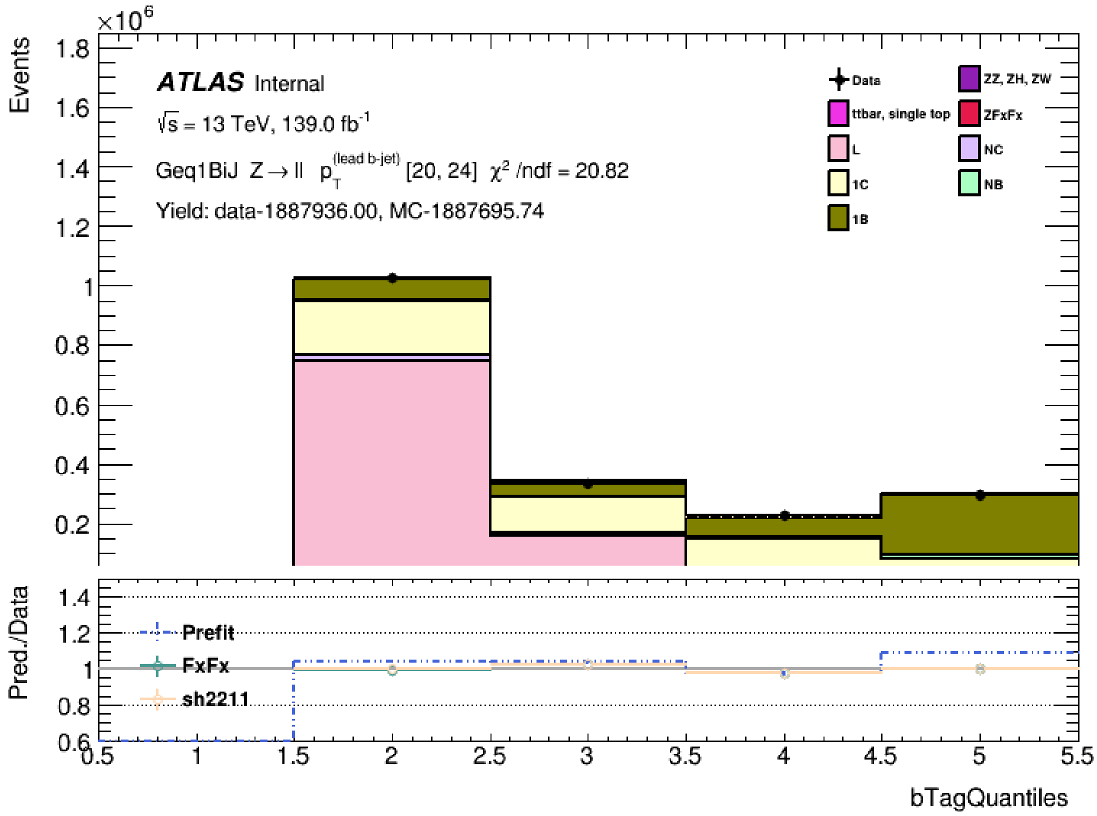


Figure 5.4: The pre- and post-fit PCBT distributions with the [20, 24] bin of p_T^{LBJ} of the combined lepton channels in the `Geq1BiJ` region.

plots have a Z boson candidate and at least one jet with $p_T > 20$ GeV and $|y| < 2.5$. Then, distributions in the signal region are shown before and after the flavour fits, rescaling the Z+jets background. In the bottom part of the plots, the ratio of predictions over data is shown. Predictions include all MC signal, plus background. In all plots, the shaded part corresponds to systematic uncertainties, described in Section 5.8.

5.6.1 Distributions in the Z+jets inclusive region

The variables shown are the the transverse impulses p_T of the leading jet (Fig.5.6) and of the Z boson (Fig. 5.7), and the dijet invariant mass (Fig. 5.8).

5.6.2 Pre- and post-flavour fit distributions in the signal regions

The plots shown are always pre-(on the left) and post-flavour fit (on the right). MC predictions slightly overestimate the data in the `ZGeq1BiJ` region, while agreement in the `ZGeq2BiJ` is better.

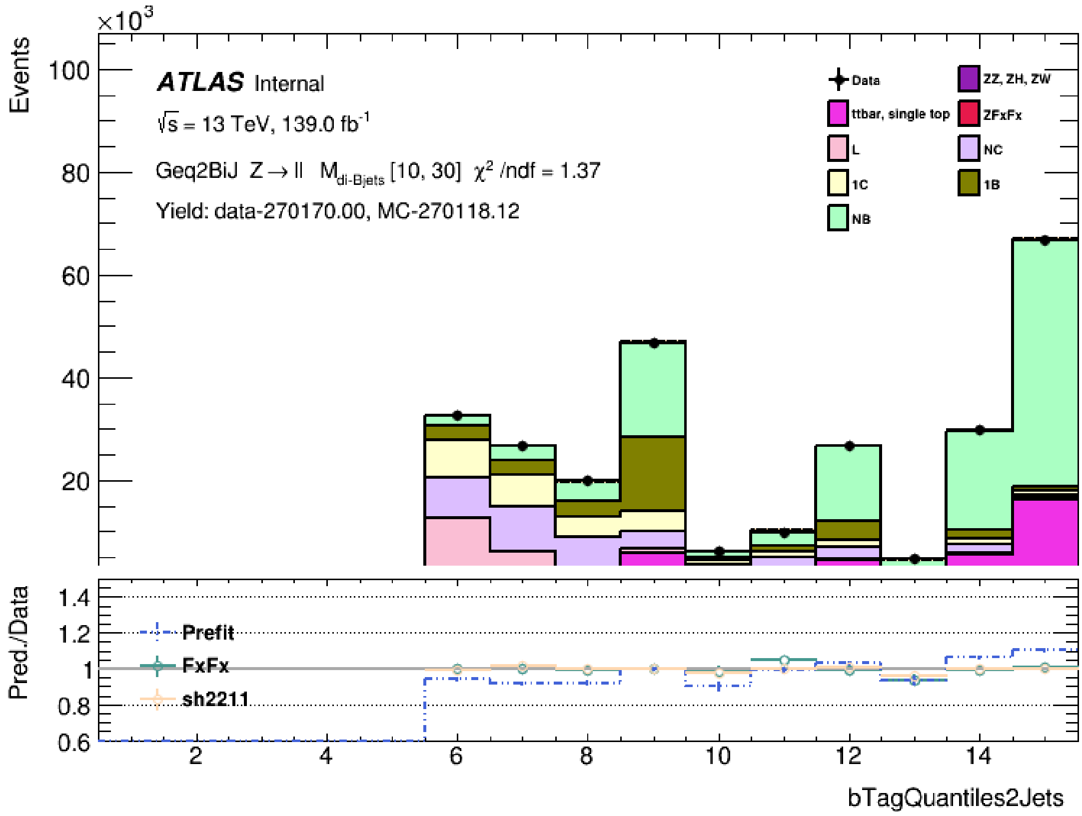


Figure 5.5: The pre- and post-fit PCBT distributions with the [10, 30] bin of $m_{\ell\ell}$ of the combined lepton channels in the `Geq2BiJ` region.

5.6.2.1 `Geq1BiJ`

The variables of choice are the transverse impulse p_T of the leading b-jet (Fig. 5.9) and of the Z boson (Fig. 5.10), and the leading b-jet x_F (Fig. 5.11).

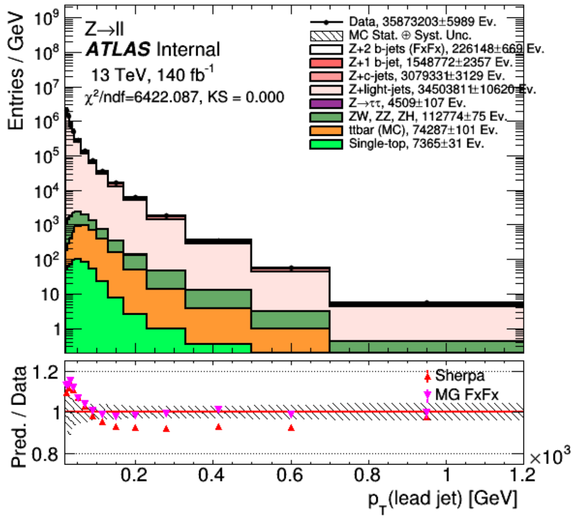


Figure 5.6: Detector level plots of leading jet p_T distributions in the inclusive region.

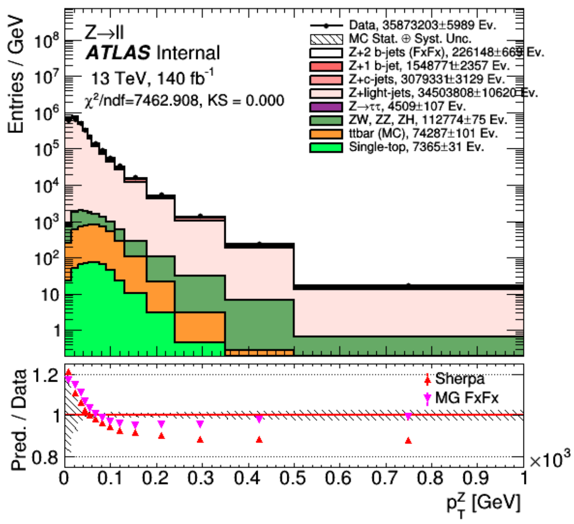


Figure 5.7: Detector level plots of Zp_T distributions in the inclusive region.

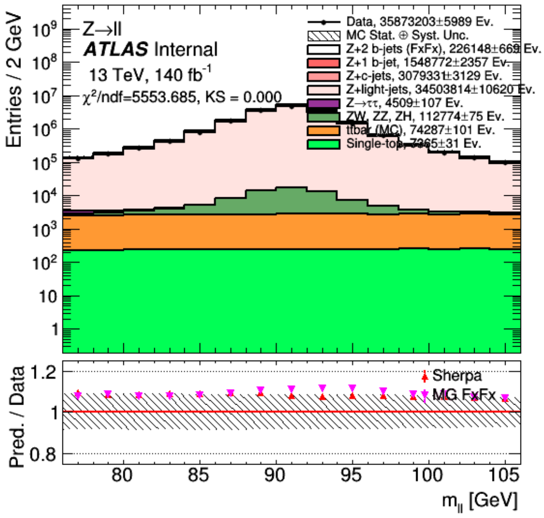


Figure 5.8: Detector level plots of di-lepton invariant mass distributions in the inclusive region.

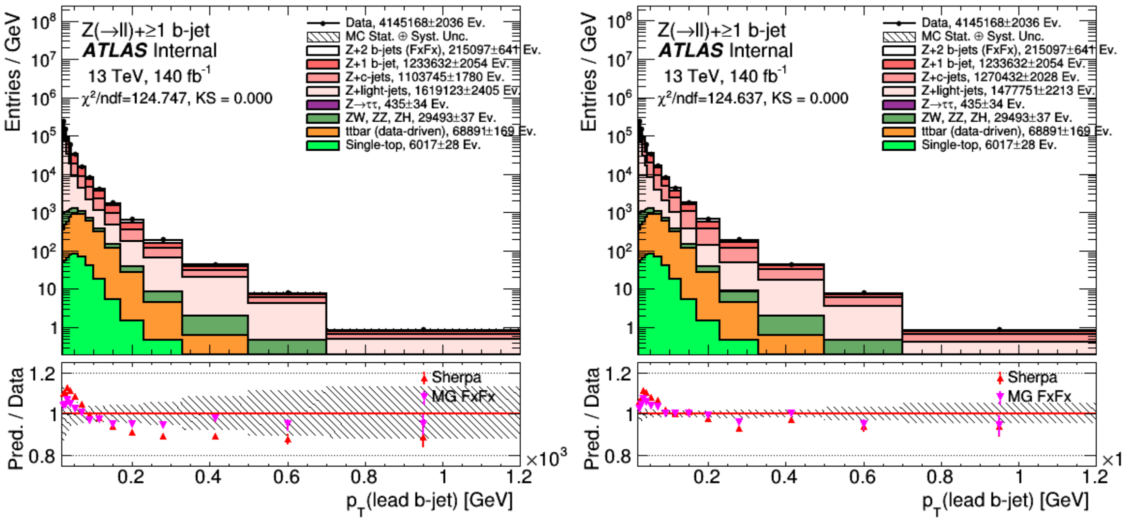


Figure 5.9: Detector level plot of leading b-jet p_T distributions before and after the flavour fit.

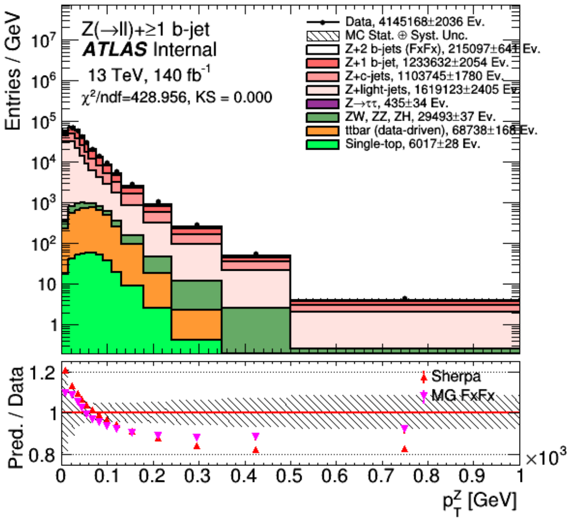


Figure 5.10: Detector level plot of Z p_T distributions before and after the flavour fit.

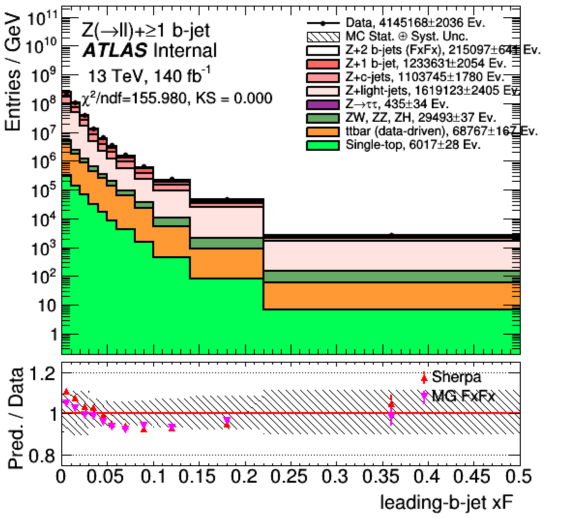


Figure 5.11: Detector level plot of leading b-jet x_F distributions before and after the flavour fit.

5.6.2.2 `Geq2BiJ`

Figure 5.12 shows the di-lepton invariant mass distributions before and after the flavour fit.

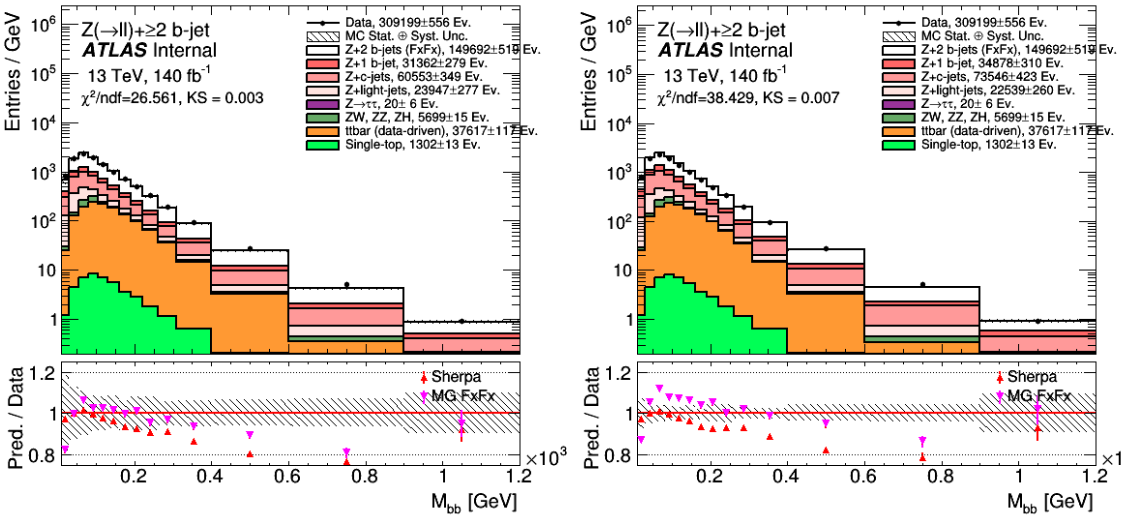


Figure 5.12: Detector level plot of di-lepton invariant mass before and after the flavour fit.

5.7 Unfolding

As seen in the previous Sections, Monte-Carlo generated events can be dealt with at **truth-level**, before the interaction with the detector, or at **reco-level**, when the final states have been reconstructed and detector effects taken into account. When it comes to measured events, however, the data only presents itself at reco-level: the truth-level distributions must be extracted in order to remove the distortion that comes from detector interaction, especially if a data comparison is to be made with different experiments or new theoretical predictions. The technique employed to reverse engineer reco-level data is called **unfolding** [37]. It is a linear inverse problem based on the integral equation

$$f_{reco}(x) = \int_{\Omega} R(x|y) f_{truth}(y) dy \tag{5.1}$$

where x and y are respectively the reco and truth variables with their respective distributions. $R(x|y)$ is the response matrix, which can be defined as

$$R_{ij} = P(\text{observed in bin } i | \text{truth value in bin } j) \tag{5.2}$$

The unfolding technique employed in this analysis is iterative Bayesian [38], and is implemented in the RooUnfold package [39].

5.8 Systematic uncertainties

Systematic uncertainties incorporate detector-level uncertainties and performance issues and have to be applied to both unfolded data and Monte Carlo generated signal, where

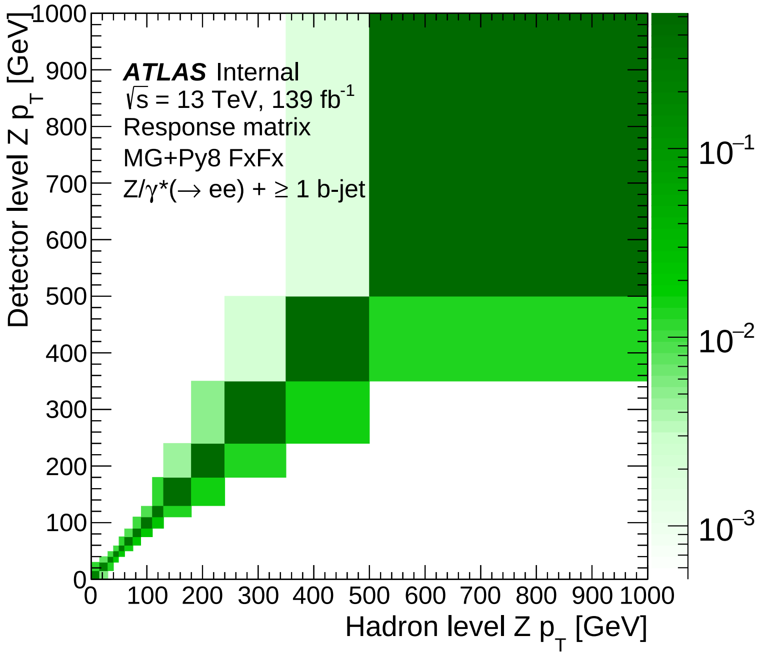


Figure 5.13: Response matrix for the p_T of the Z in the `ZGeq1BiJ` fiducial region. It relates true-level to reco-level distributions.

they are applied to reconstruction-level plots and summed in quadrature. They include, among others:

- `LUMI`: uncertainty on the luminosity;
- `EG`: uncertainties on electron reconstruction and identification. These include energy scale and resolution, and reconstruction, identification and trigger efficiencies;
- `MUON`, uncertainties on muon reconstruction and identification. These include smearing of the muon track, scale corrections and reconstruction, identification and trigger efficiencies;
- `MET`: uncertainties on scale and resolutions of the soft tracks used in missing transverse enrgy calculations;
- `JES`: uncertainties on Jet Energy Scale, which is the calibration performed to match the energy of the jet to that of the particles that compose it;
- `JER`: uncertainties on Jet Energy Resolution, which quantifies the fractional uncertainty on the reconstructed energy of the jet;
- `PRW`: uncertainties on pileup vertices;
- `FTAG`: flavour tagging uncertainties, which are mainly composed of variations in the calibrated efficiencies, i.e. corrections to the scale factors.

As seen in Tables 5.11, 5.12 and 5.13, flavour tagging and jet reconstruction are among the main sources of uncertainty in measured cross sections.

$Z \to \ell^+\ell^- \geq 1b$-jet	
Syst. source	Rel. uncertainty (%)
BTagging	± 3.58
DIBOSON	± 0.07
EGAMMA	± 0.24
FxFx	± 0.48
Jet	± 2.41
Luminosity	± 0.80
MET	± 0.42
MUON	± 0.11
PRW	± 0.60
SINGLETOP	± 0.03
TTBAR	± 0.12
UNFOLDING	± 4.69
VjBkg	± 0.60
ZH	± 0.00
ZTAUTAU	± 0.00
Statistics	± 0.16

Table 5.11: Breakdown of relative systematic uncertainties on measured cross section for events coming from Z decay in both electron and muon channel + $\geq 1b-$jet.

$Z \to \ell^+\ell^- \geq 2b$-jet	
Syst. source	Rel. uncertainty (%)
BTagging	± 5.71
DIBOSON	± 0.18
EGAMMA	± 0.25
FxFx	± 0.33
Jet	± 4.34
Luminosity	± 0.87
MET	± 0.51
MUON	± 0.11
PRW	± 0.57
SINGLETOP	± 0.06
TTBAR	± 0.28
UNFOLDING	± 5.53
VjBkg	± 1.49
ZH	± 0.02
ZTAUTAU	± 0.00
Statistics	± 0.44

Table 5.12: Breakdown of relative systematic uncertainties on measured cross section for events coming from Z decay in both electron and muon channel + $\geq 2b-$jet.

$Z \to \ell^+\ell^- \geq 1c$-jet	
Syst. source	Rel. uncertainty (%)
BTagging	±10.27
DIBOSON	0.11
EGAMMA	±0.38
FxFx	–
Jet	±6.45
Luminosity	±0.74
MET	±0.34
MUON	±0.11
PRW	±0.17
SINGLETOP	±0.03
TTBAR	±0.01
UNFOLDING	±5.62
VjBkg	±1.58
ZH	±0.00
ZTAUTAU	±0.00
Statistics	±0.32

Table 5.13: Breakdown of relative systematic uncertainties on measured cross section for events coming from Z decay in both electron and muon channel + $\geq 1c-$jet.

5.9 Results

5.9.1 Inclusive cross sections

The inclusive cross sections for regions `ZGeq2B`, `ZGeq1B`, `ZGeq1C` are presented in this section. The tables feature the cross section as extracted via Unfolding compared to the cross section predicted by the `MadGraph` simulations. Data results are presented in the form σ[pb] ± stat. ± syst.. MC results are presented as σ[pb]± MC stat.

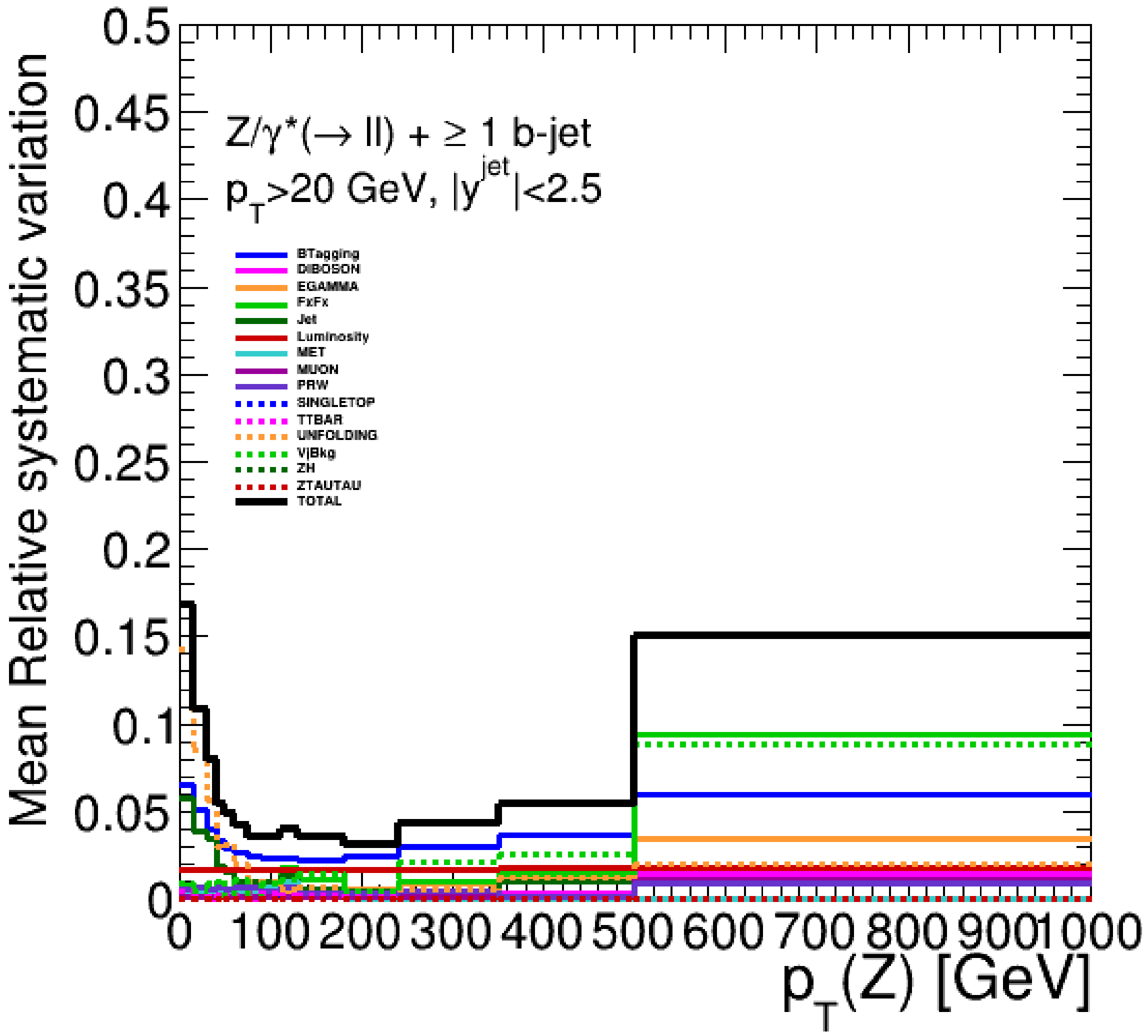


Figure 5.14: Uncertainty breakdown on the unfolded results of the Z boson p_T for an event with at least $1b-$ jet, in the electron-muon channel.

	$Z \geq 1b-$jet	
Signal	Data [pb]	MC [pb]
$Z \rightarrow \mu^+\mu^-$	$10.438 \pm 0.020 \pm 0.741$	11.736 ± 0.017
$Z \rightarrow e^+e^-$	$10.611 \pm 0.025 \pm 0.614$	11.662 ± 0.017
$Z \rightarrow \ell^+\ell^-$	$10.490 \pm 0.017 \pm 0.684$	11.699 ± 0.012

Table 5.14: Inclusive cross sections in the $Z+ \geq 1b-$jet signal region

	$Z \geq 2b-$jet	
Signal	Data [pb]	MC [pb]
$Z \to \mu^+\mu^-$	$1.379 \pm 0.008 \pm 0.131$	1.568 ± 0.005
$Z \to e^+e^-$	$1.423 \pm 0.010 \pm 0.130$	1.556 ± 0.005
$Z \to \ell^+\ell^-$	$1.394 \pm 0.006 \pm 0.129$	1.562 ± 0.003

Table 5.15: Inclusive cross sections in the $Z+ \geq 2b-$jet signal region

	$Z \geq 1c-$jet	
Signal	Data [pb]	MC [pb]
$Z \to \mu^+\mu^-$	$20.706 \pm 0.083 \pm 2.887$	18.370 ± 0.021
$Z \to e^+e^-$	$21.091 \pm 0.106 \pm 2.774$	18.390 ± 0.021
$Z \to \ell^+\ell^-$	$20.890 \pm 0.068 \pm 2.818$	18.380 ± 0.015

Table 5.16: Inclusive cross sections in the $Z+ \geq 1c-$jet signal region

5.9.2 Differential cross sections

Differential cross sections plots are shown for various kinematic variables of interest. For signal regions with one b-jet, plots for the transverse impulse p_T of the leading b-jet and of the Z boson, and the leading b-jet x_F are presented for electrons and muons separately (Figures 5.15, 5.17 and 5.20 respectively) and after combination (Figures 5.16, 5.19 and 5.18 respectively). The Z p_T and leading c-jet variables are presented for events with one c-jet in Figures 5.21 and 5.22. For the two b-jets signal region, the di-lepton invariant mass $m_{\ell\ell}$ is shown in Figure 5.23.
In all plots, the black line represents unfolded data and the red and blue lines represent Madgraph and Sherpa predictions, respectively.

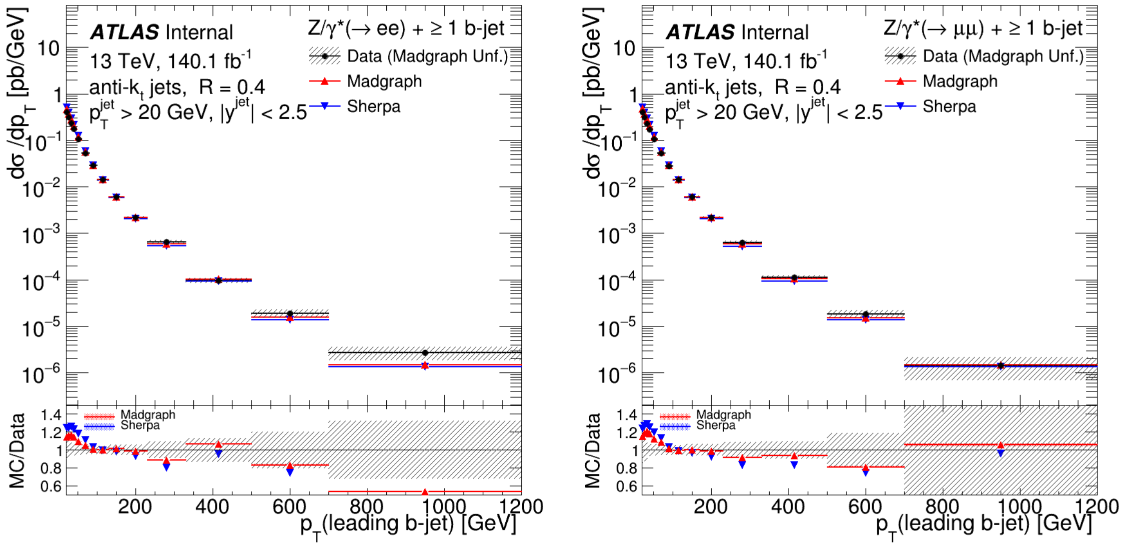


Figure 5.15: Differential cross section for leading b-jet p_T: comparison of unfolded data and predictions from MG5+PY8 FxFx and SHERPA v2.2.11 for events with at least one b-jet, electron (left) and muon (right) channels.

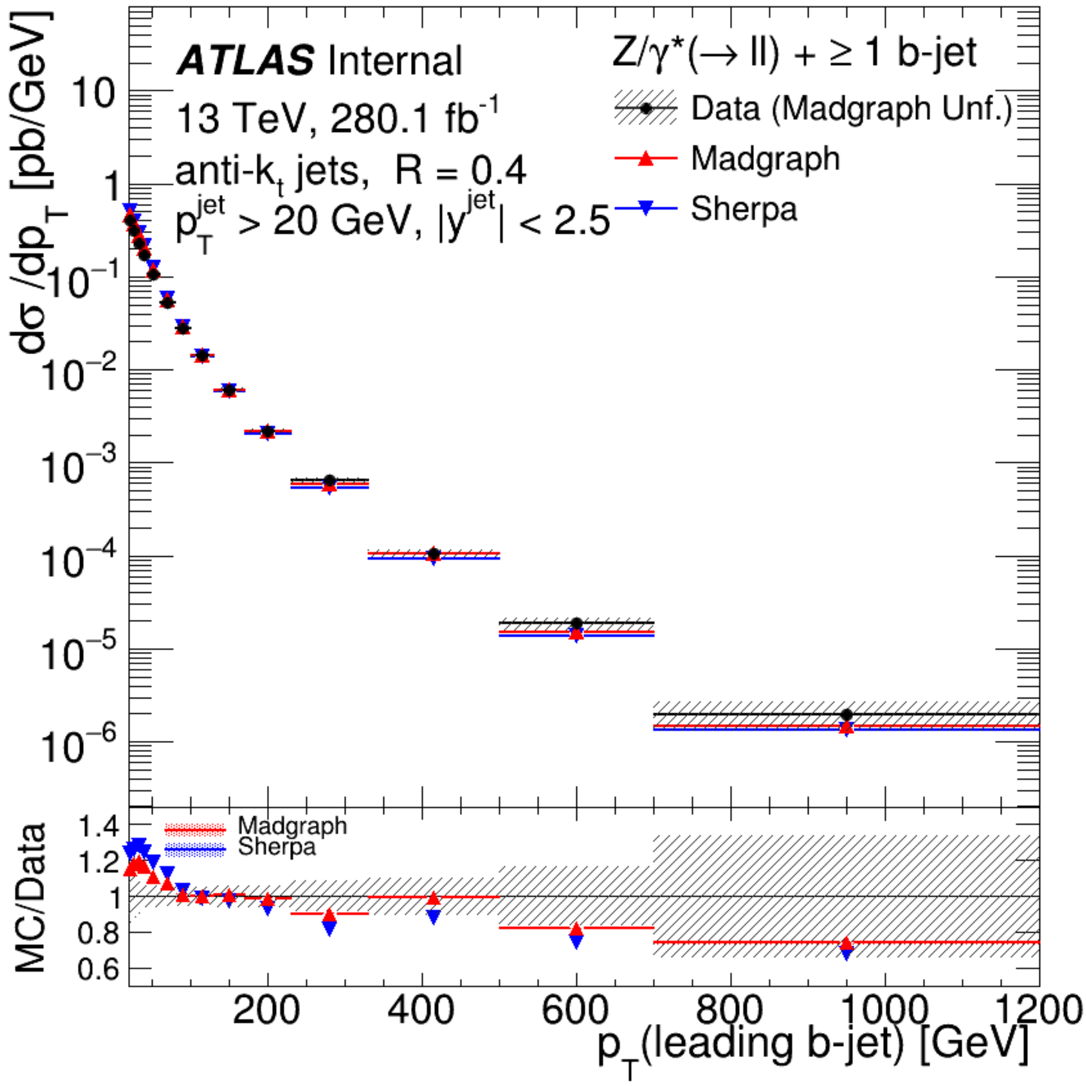
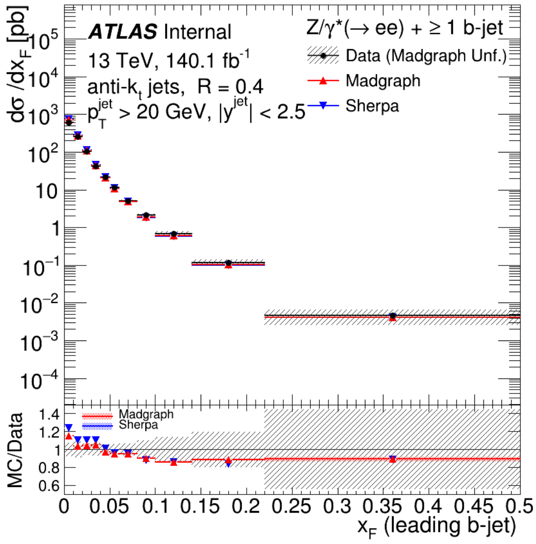


Figure 5.16: Differential cross section for leading b-jet p_T: comparison of unfolded data and predictions from MG5+PY8 FxFx and SHERPA v2.2.11 for events with at least one b-jet, electron+muon channel.



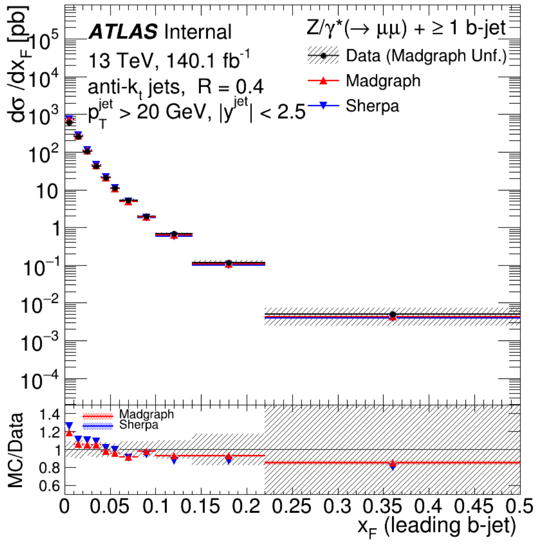


Figure 5.17: Differential cross section for leading b-jet x_F: comparison of unfolded data and predictions from MG5+PY8 FxFx and SHERPA v2.2.11 for events with at least one b-jet, electron (left) and muon (right) channels.

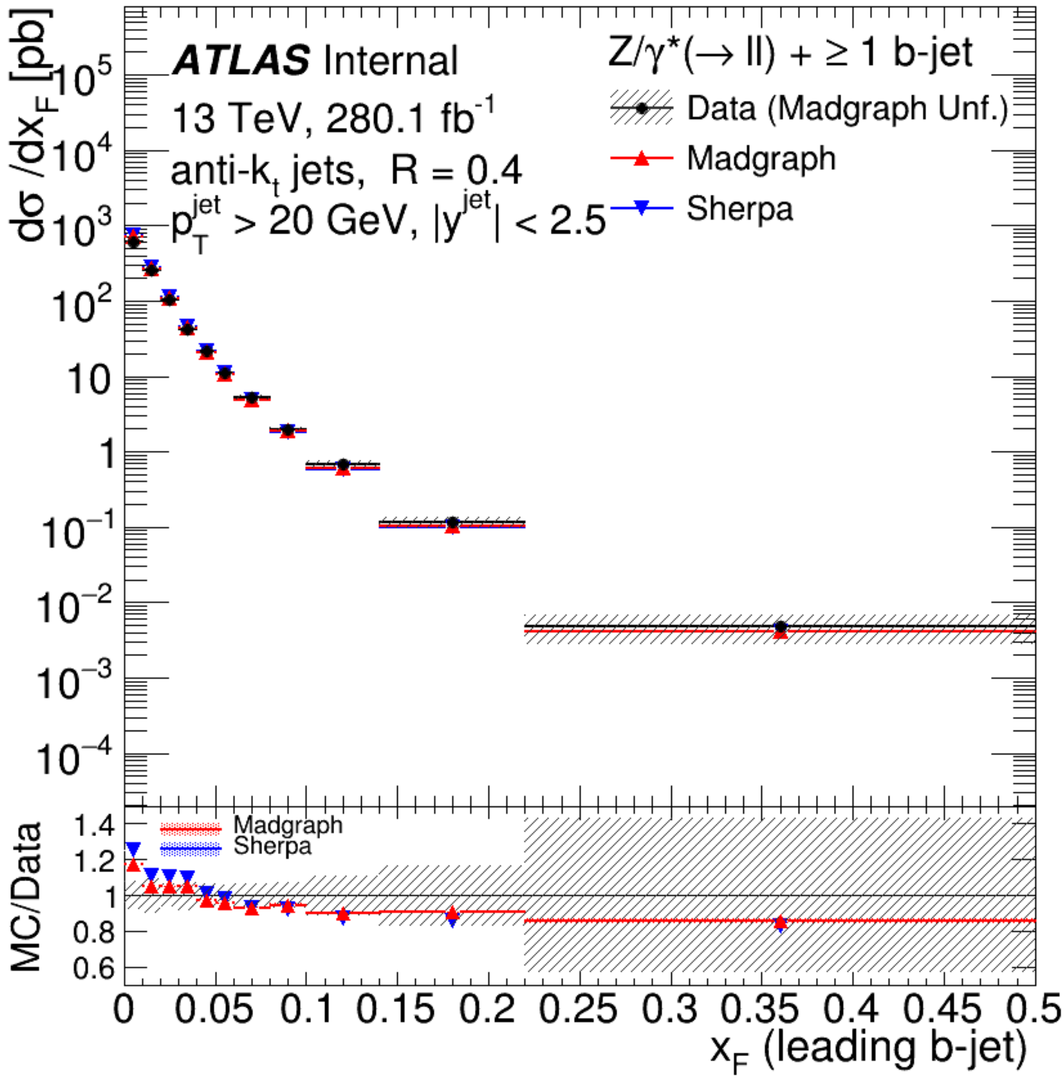


Figure 5.18: Differential cross section for leading b-jet x_F: comparison of unfolded data and predictions from MG5+PY8 FxFx and SHERPA v2.2.11 for events with at least one b-jet, electron+muon channel.

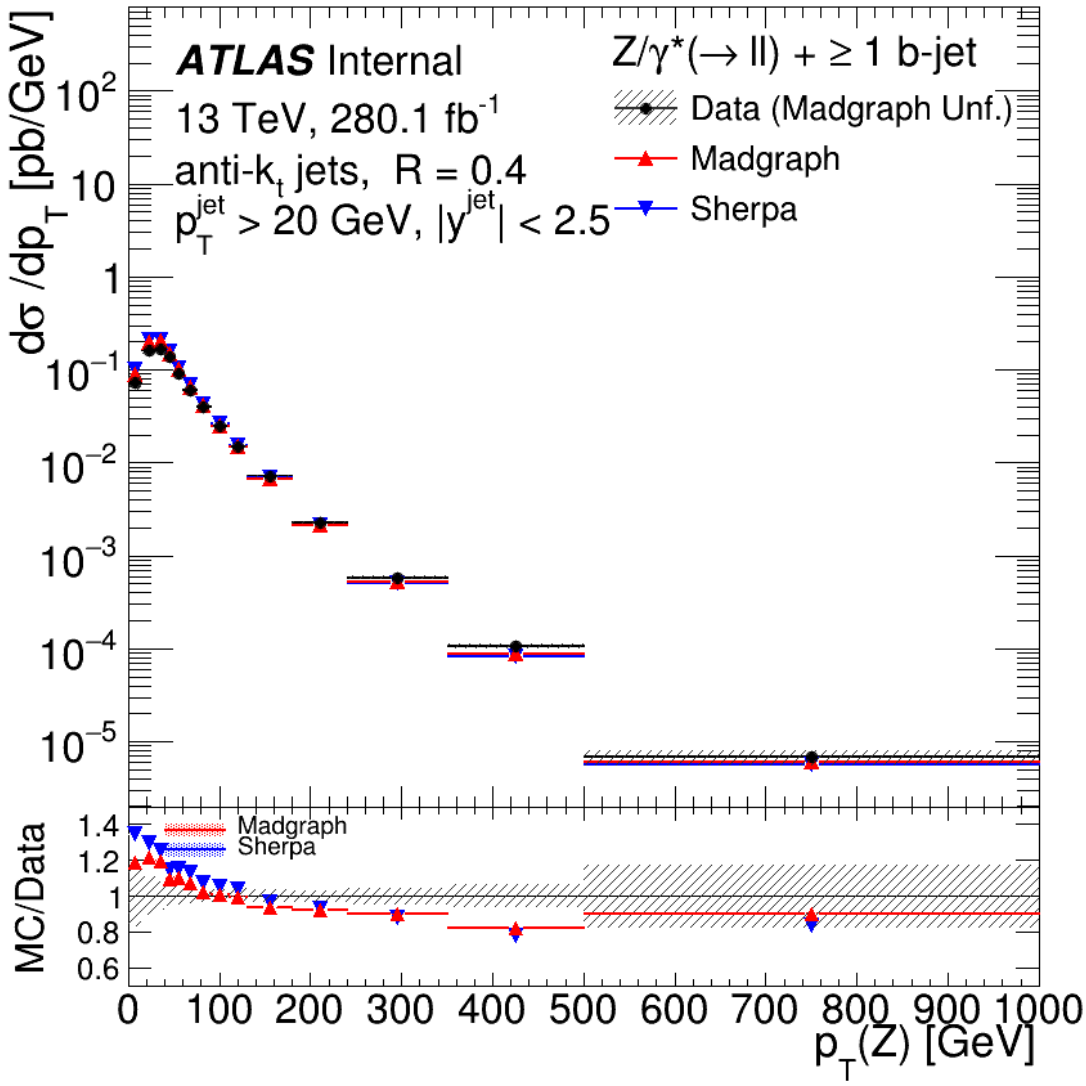


Figure 5.19: Differential cross section for Z boson p_T: comparison of unfolded data and predictions from MG5+PY8 FxFx and SHERPA v2.2.11 for events with at least one b-jet, electron+muon channel.

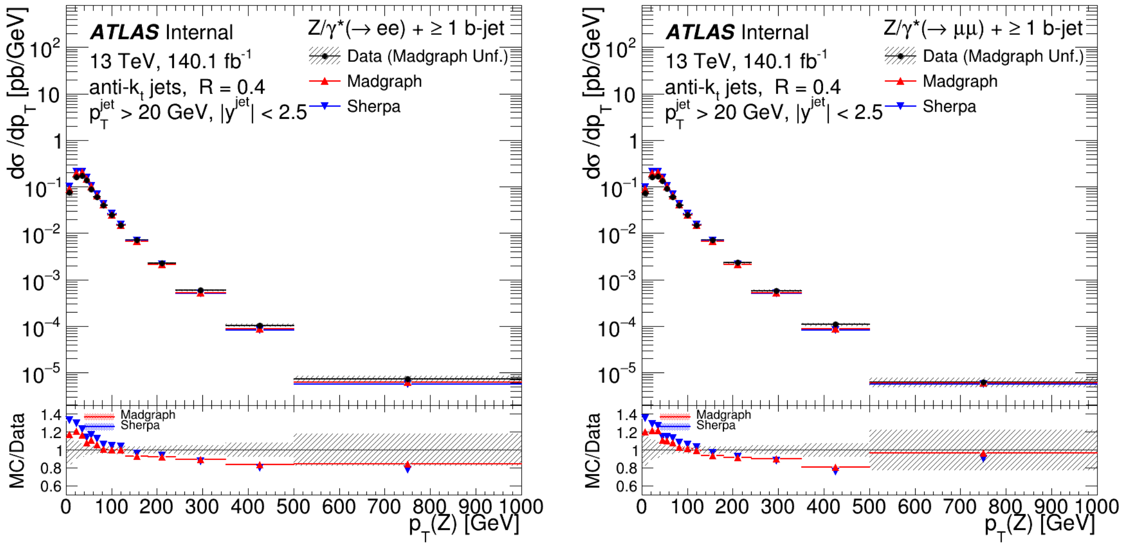


Figure 5.20: Differential cross section for Z boson p_T: comparison of unfolded data and predictions from MG5+PY8 FxFx and SHERPA v2.2.11 for events with at least one b-jet, electron (left) and muon (right) channels.

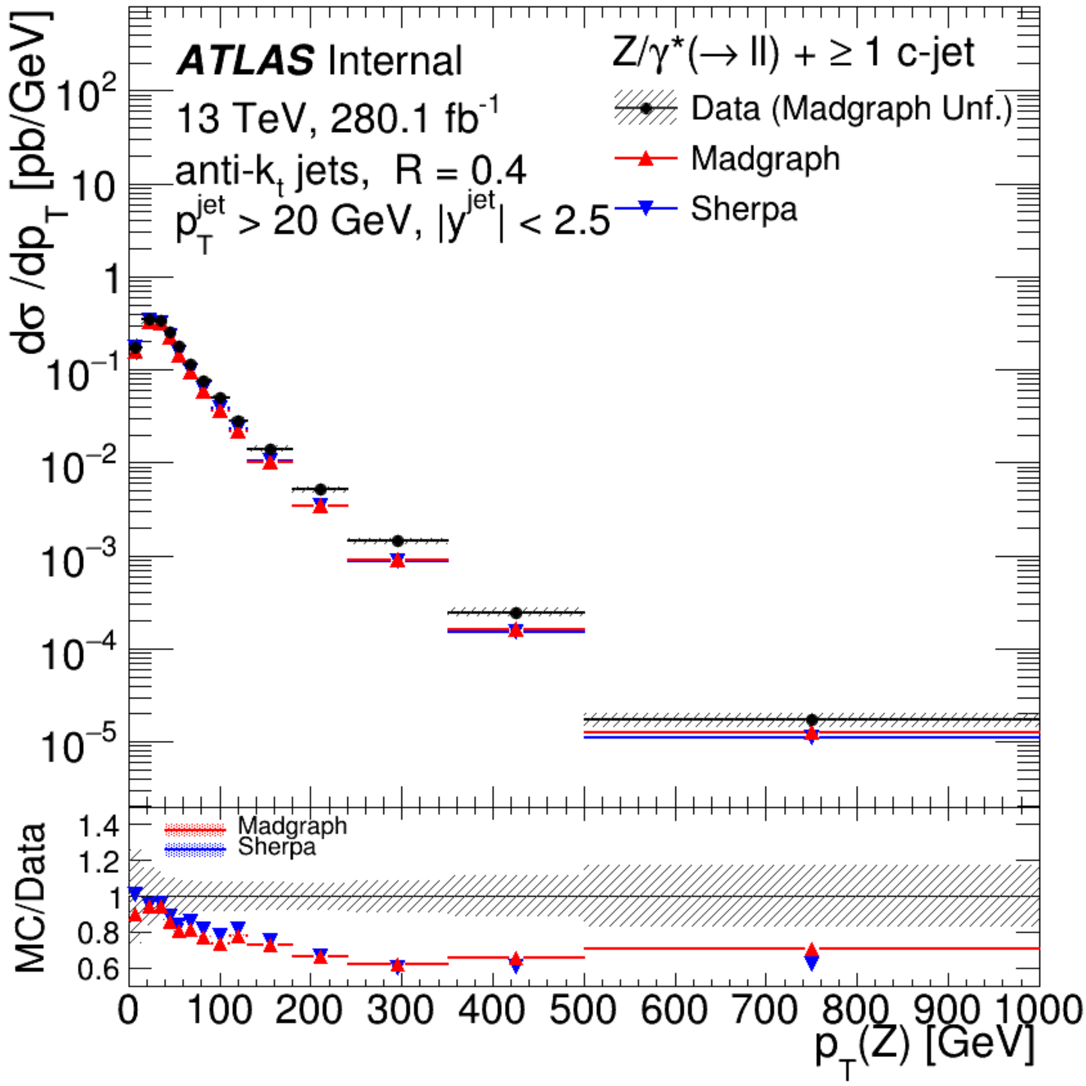


Figure 5.21: Differential cross section for Z boson p_T: comparison of unfolded data and predictions from MG5+PY8 FxFx and SHERPA v2.2.11 for events with at least one c-jet, electron+muon channel.

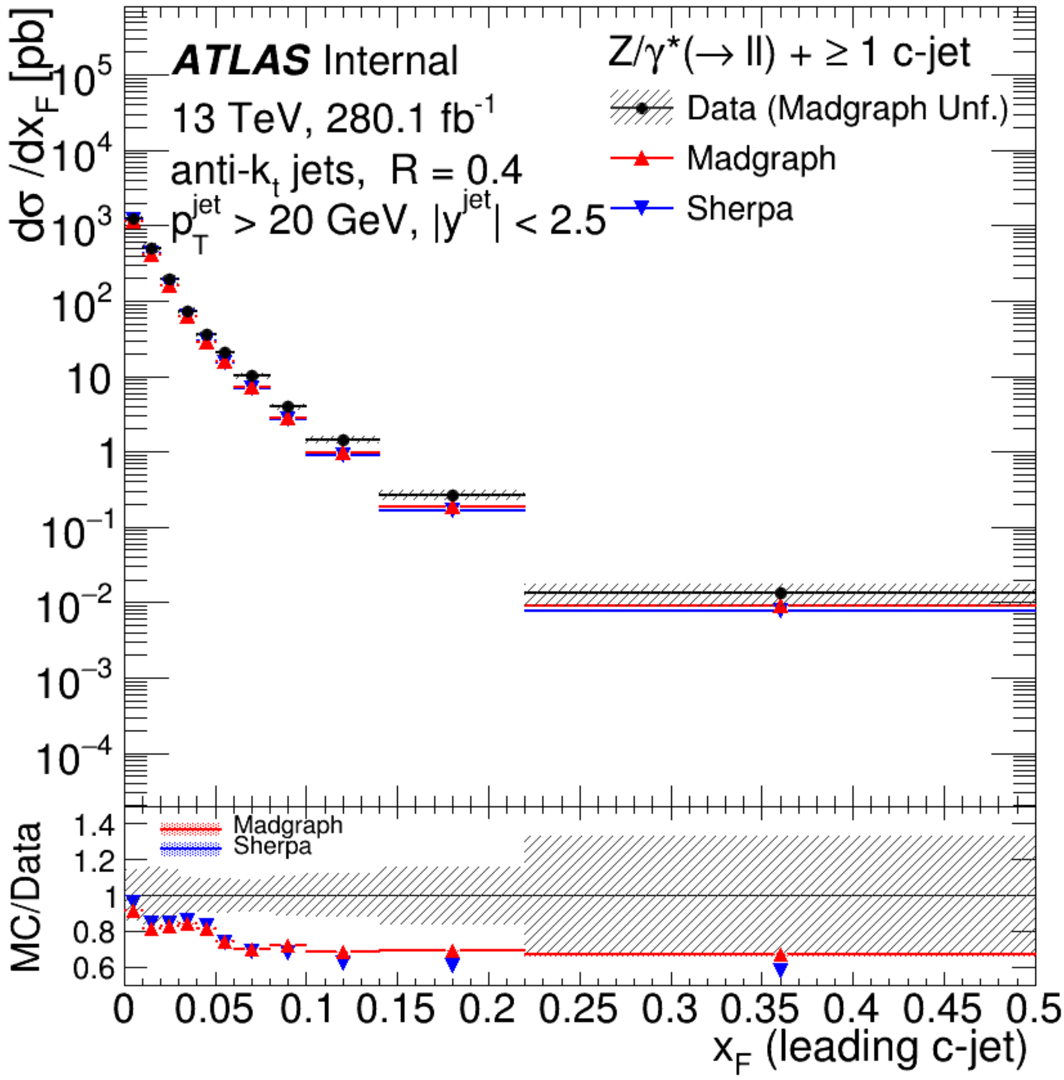


Figure 5.22: Differential cross section for leading c-jet x_F: comparison of unfolded data and predictions from MG5+PY8 FxFx and SHERPA v2.2.11 for events with at least one c-jet, electron+muon channel.

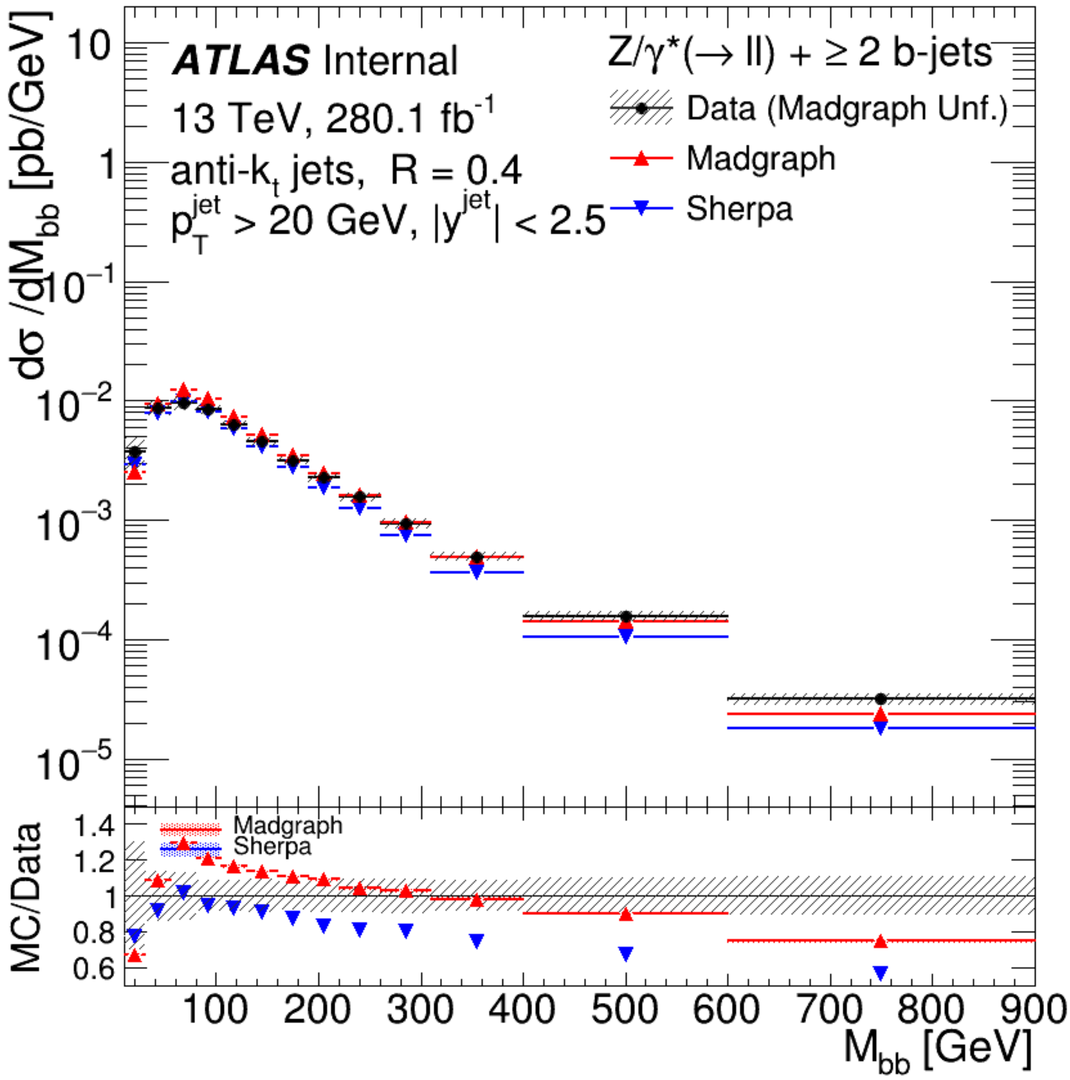


Figure 5.23: Differential cross section for lepton invariant mass: comparison of unfolded data and predictions from MG5+PY8 FxFx and SHERPA v2.2.11 for events with at least two b-jets, electron+muon channel.

5.10 5FNS and 4FNS comparison

As explained in detail in Section 2.5.4, a major contribution to the final states of an hadronic collision comes from PDFs. It is then a relevant exercise to compare results in samples generated with different PDFs. In this section, a comparison of differential cross sections calculated in both 5FNS and 4FNS to unfolded data is presented. As expected, the cross sections in 4FNS are consistentely smaller, because the sole contribution from b quarks comes from gluon splitting, and there is no further contribution from perturbative b-quarks. After the plots, integrated cross sections for three different variables, the transverse impulse p_T of both leading b-jet and Z and the di-jet invariant mass are presented. The invariant mass is the variable that is least influenced by the difference in the two schemes.

	σ [pb]		
Variable [GeV]	Unfolded data	Madgraph 4FNS	Madgraph 5FNS
p_T(leading b-jet)	10.3693 $\pm$ 0.0127	3.3255 $\pm$ 0.0455	11.7046 $\pm$ 0.0123
M_{bb}	1.3916 $\pm$ 0.0024	1.4083 $\pm$ 0.0513	1.5621 $\pm$ 0.0034

Table 5.17: cross sections for data and different flavour schemes simulations.

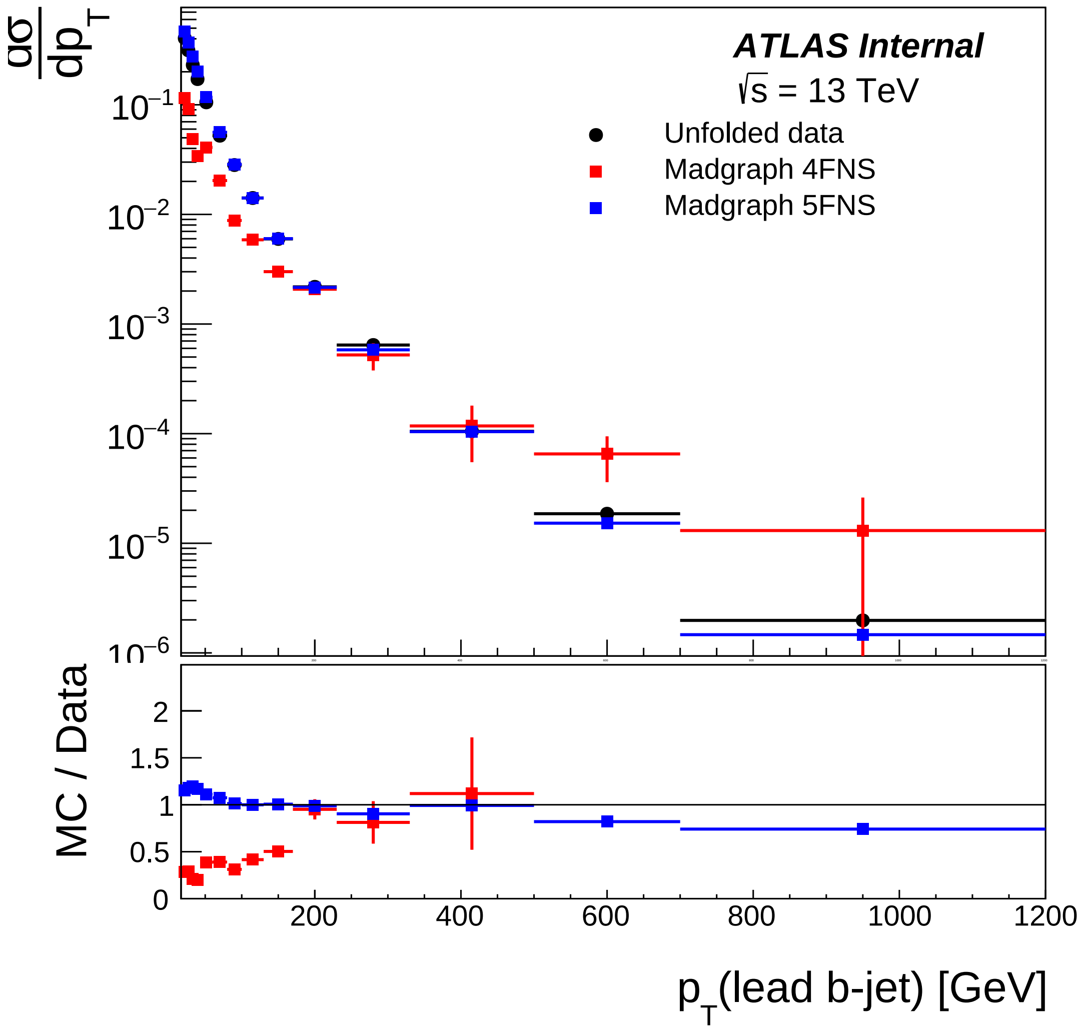


Figure 5.24: Differential cross section for leading b-jet p_T: comparison of unfolded data and predictions from Madgraph at 4FNS (with statistical error bands) and 5FNS for events with at least one b-jet, electron+muon channel.

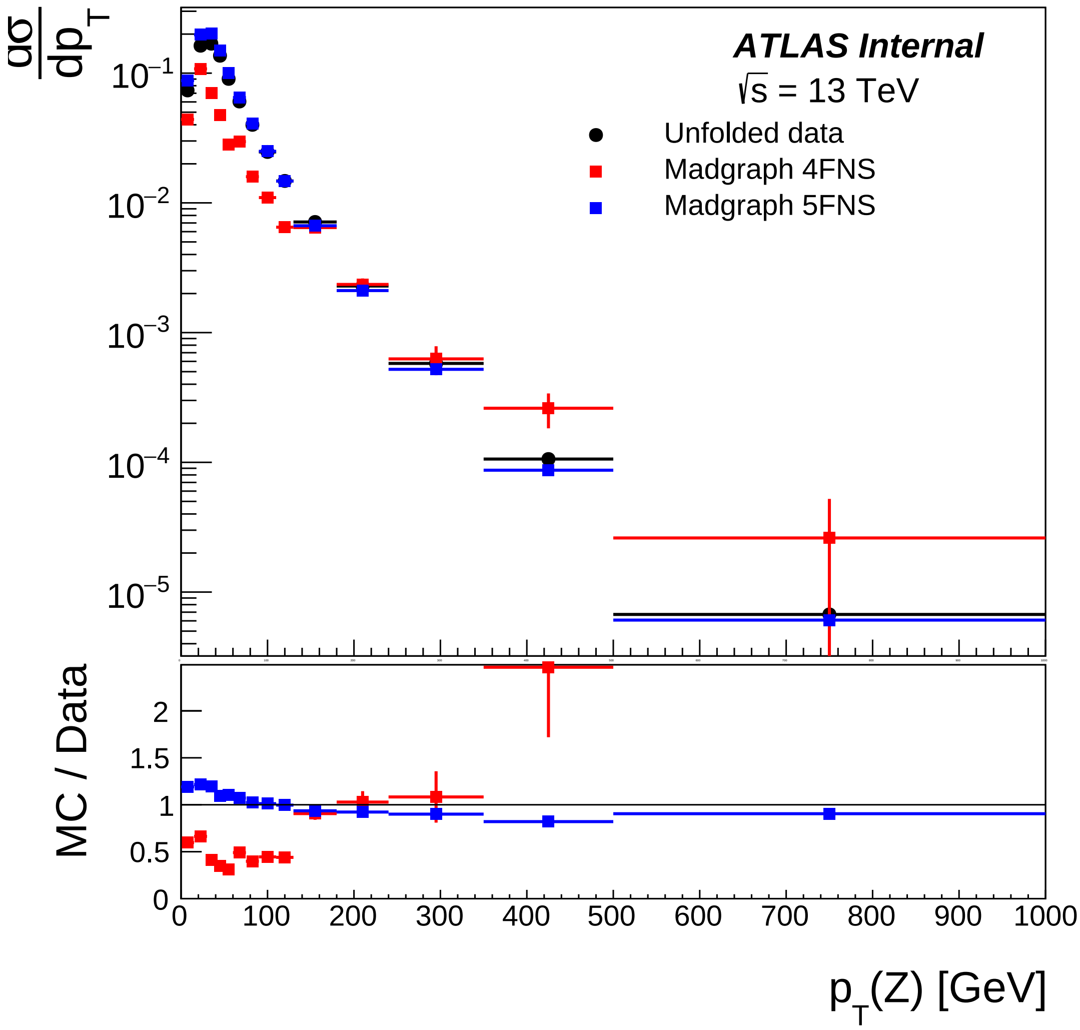


Figure 5.25: Differential cross section for Z p_T: comparison of unfolded data and predictions from Madgraph at 4FNS (with statistical error bands) and 5FNS for events with at least one b-jet, electron+muon channel.

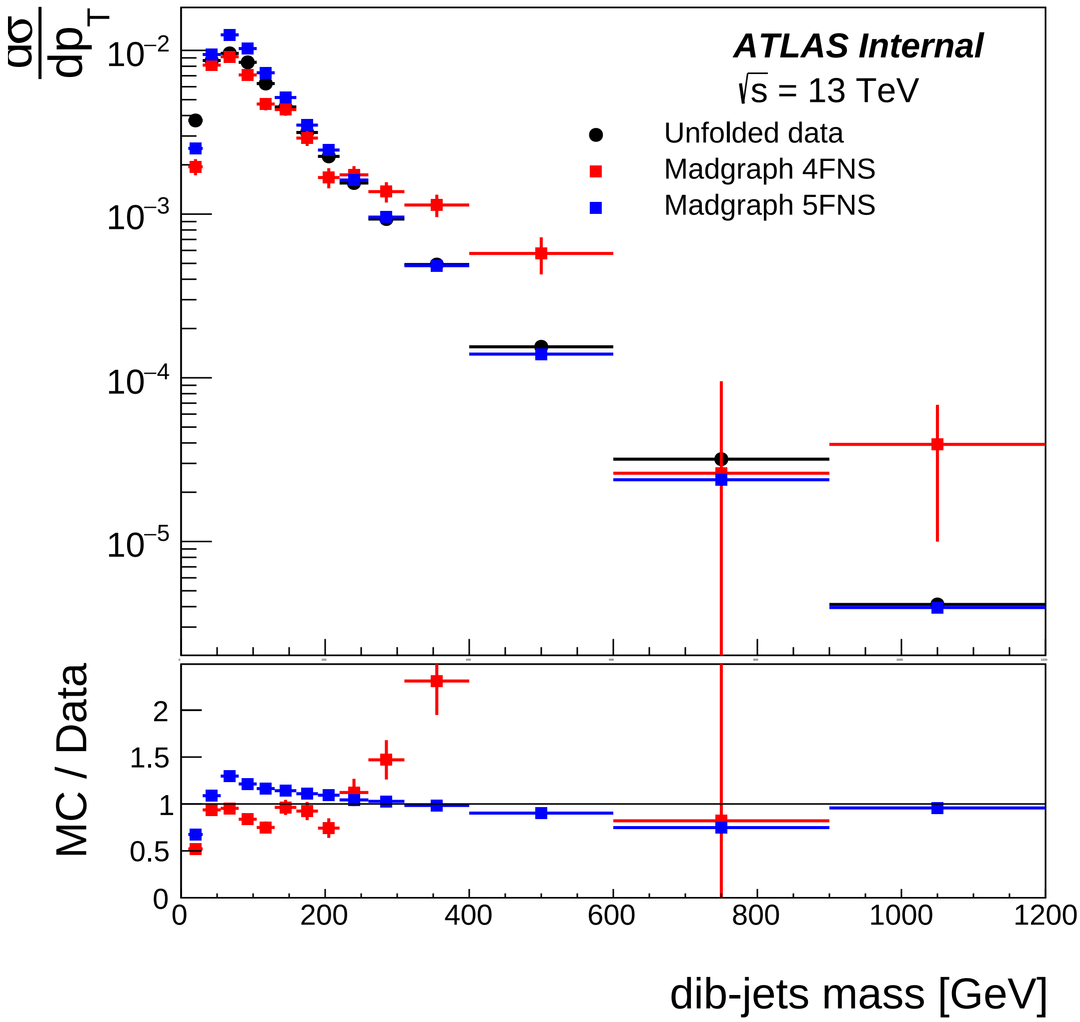


Figure 5.26: Differential cross section for di-jet invariant mass: comparison of unfolded data and predictions from Madgraph at 4FNS (with statistical error bands) and 5FNS for events with at least two b-jets, electron+muon channel.

Chapter 6

Intrinsic charm sensitivity

As stated in previous chapters, the presence of intrinsic charm in the proton can be investigated by studying the final states which contain one c-jet. In particular, the differential cross section is of the utmost importance: the first intrinsic charm theories, like BHPS [11], originated from discrepancies in the cross sections for the production of D mesons and unexpected values in the comparison between forward and central regions.
This gives rise to two questions: which is the optimal kinematic variable to use for the differential cross section, and which is the optimal way of defining the forward and central regions. The regions are defined from rapidity, which is a kinematic variable that identifies the relative angle between the particle and the beam, as seen in Eq. 2.36. The b-jet signal from the data used in this analysis cover a rapidity range of $|y| \leq 2.5$. Setting a cut at a fixed $\bar{y}$ means defining the central region as wherever $y < \bar{y}$, and the forward region wherever $y > \bar{y}$. In the next sections, the distribution of kinematic variables as the percentage of intrinsic charm changes will be analyzed through the reweighting procedure, and a scan of rapidities will be performed, to determine the optimal variable at the optimal cut to calculate the ratio of cross sections in the forward and central regions. The ratio is chosen for this analysis because correlated uncertainties between the two regions cancel out, improving the sensitivity.

6.1 PDF Reweighting for intrinsic charm sensitivity studies

PDF reweighting is a method that allows to better understand the impact that different PDF choices have on the output of the simulation. The reweighting process is, in a way, an *a posteriori* PDF choice to correct a sample generated with an old PDF to a new one. Weights are calculated as follows:

$$w(x_1, x_2, Q^2) = \frac{f_i^{new}(x_1, Q^2) f_j^{new}(x_2, Q^2)}{f_i^{old}(x_1, Q^2) f_j^{old}(x_2, Q^2)} \tag{6.1}$$

The reweighting process was performed once with a PDF that contained intrinsic charm, and once with one that did not. The PDFs used in reweighting can be found in Table 6.1

Reweight PDF	Notes
`NNPDF_3.1_NNLO`	contains IC component in unknown percentage
`NNPDF_3.1_NNLO_pch`	only contains perturbative charm
`CTEQ14_NNLO_IC/0000`	does not contain intrinisc charm
`CTEQ14_NNLO_IC/0001`	contains 0.6% intrinsic charm according to BHPS models
`CTEQ14_NNLO_IC/0002`	contains 2.1% intrinsic charm according to BHPS models

Table 6.1: PDFs chosen for the reweight process, as can be found in the LHPDF repository from the NNPDF [40] and CTEQ14 collaborations [41].

The PDFs chosen for the reweighting are two: NNPDF 3.1 and CTEQ 14. The reason for this choice is the fact that both come in two versions: one that contains intrinsic charm and one that does not. In NNPDF, one PDF contains intrinsic charm, while the other only contains perturbative charm. CTEQ14 PDFs come in two variations to the central value `CTEQ14_NNLO_IC/0000` that has no intrinsic charm:, `CTEQ14_NNLO_IC/0001` contains a fraction of intrinsic charm of 0.6% and `CTEQ14_NNLO_IC/0002` contains a fraction of intrinsic charm of 2.1%, in both cases fit in according to the BHPS model.
Reweighting was performed on samples from $Z \rightarrow \mu^+\mu^-$ in the signal area `ZGeq1C`, which contains at least one c-jet and no b-jets. The studies focused on the p_T of both jets and Z boson, and x_F of the leading jet, because according to several studies, for instance [42] and [43], these variables are the ones that are most likely to be impacted by the presence of intrinsic charm, which should manifest itself in the forward region.

As seen in Figures 6.1, 6.2 and 6.3, at high p_T and x_F, which correspond to higher values of $|y|$ (i.e. in the Forward region), up to 30% increases in data are shown for PDFs containing higher percentages of intrinsic charm.

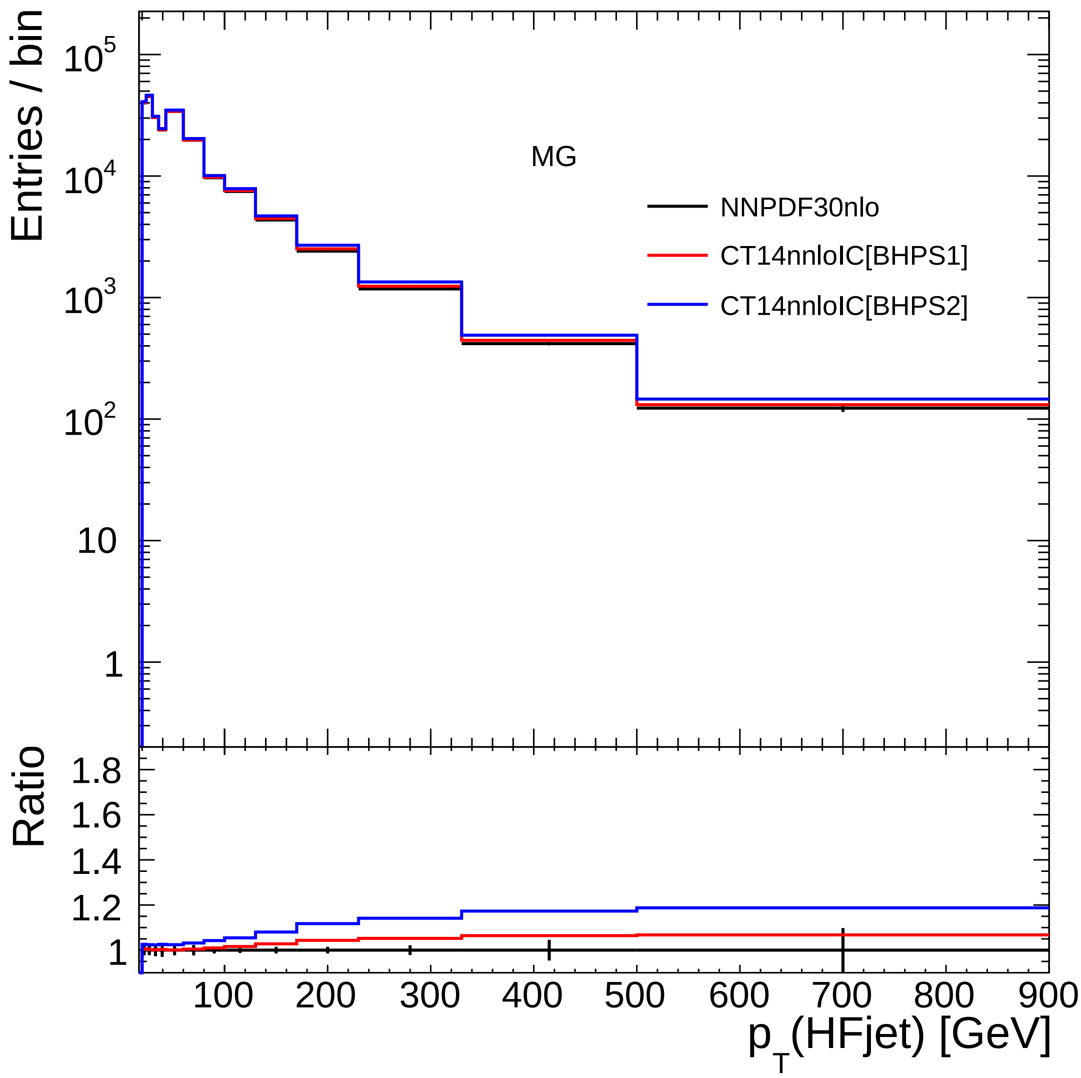


Figure 6.1: Reweighting for leading jet p_T: `NNPDF30nnlo` is reweighted to `CT14nnloIC` with both BHPS1 and BHPS2.

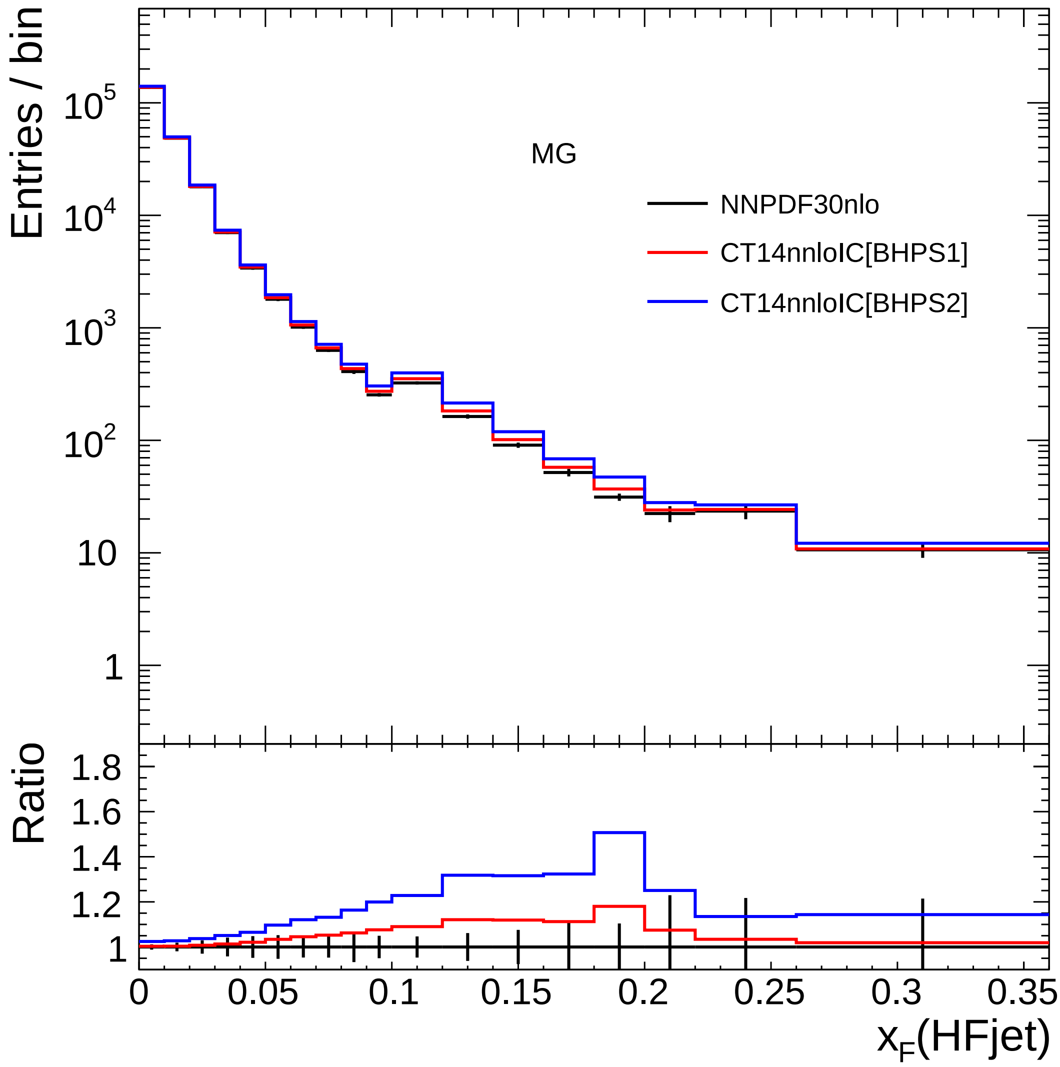


Figure 6.2: Reweighting for leading jet x_F: `NNPDF30nnlo` is reweighted to `CT14nnloIC` with both BHPS1 and BHPS2.

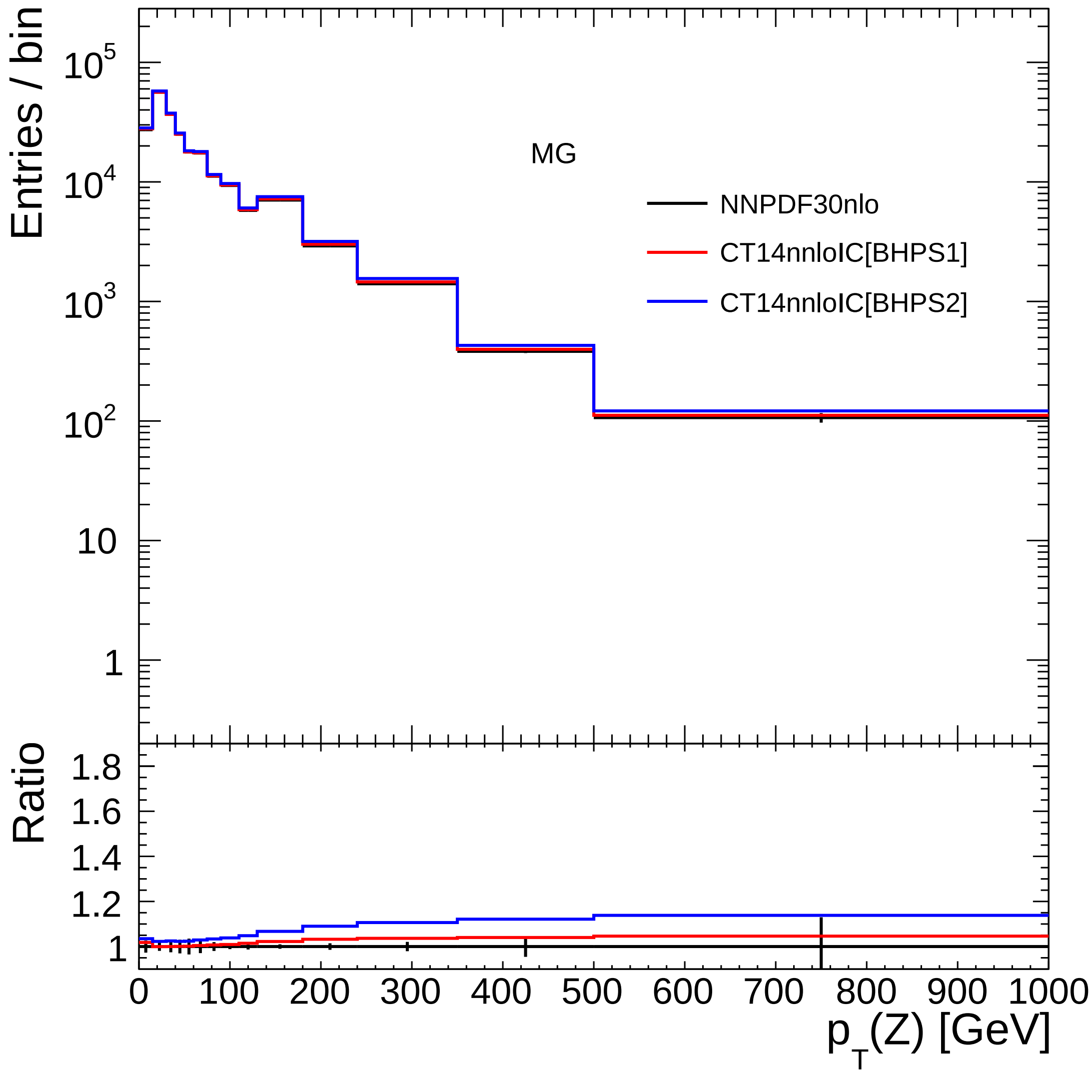


Figure 6.3: Reweighting for Z p_T: `NNPDF30nnlo` is reweighted to `CT14nnloIC` with both BHPS1 and BHPS2.

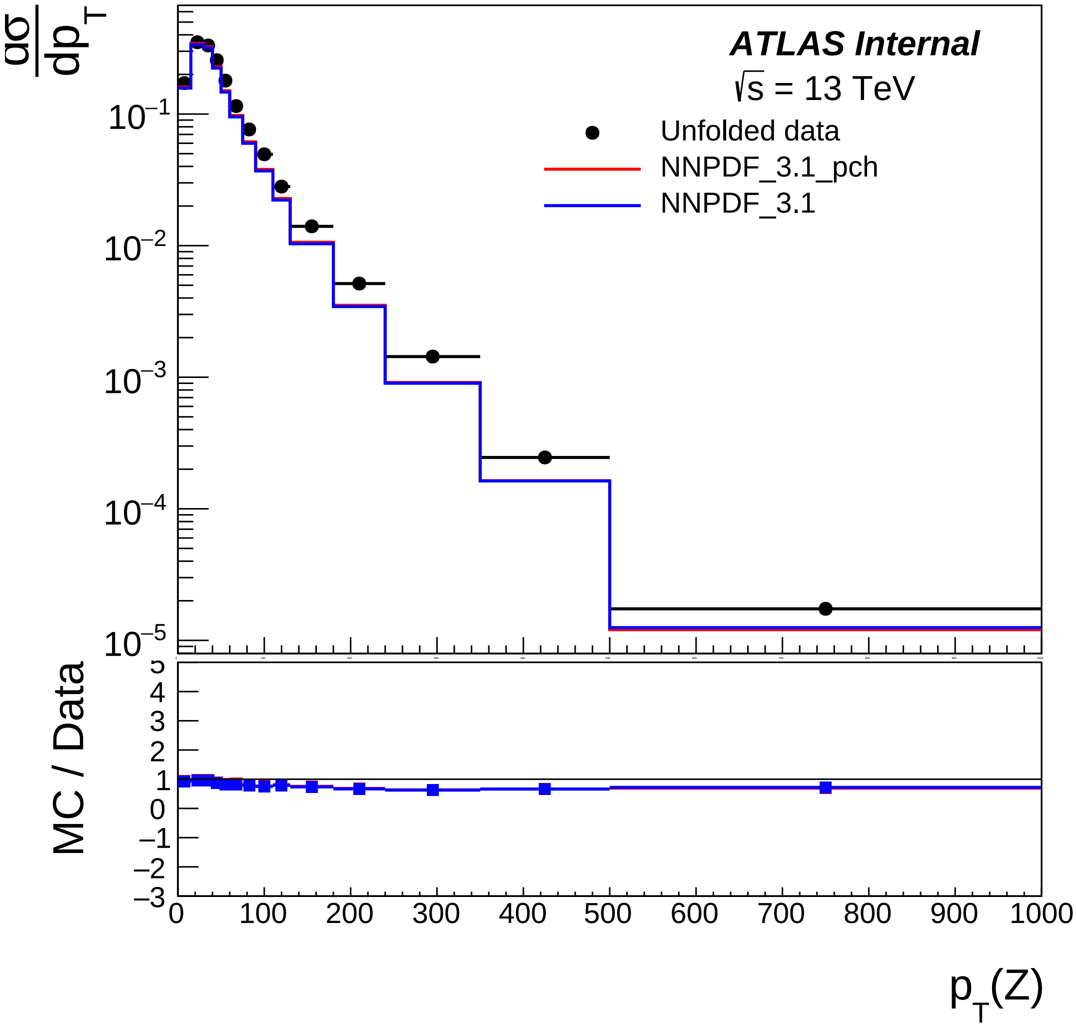


Figure 6.4: Differential cross section for leading b-jet p_T: comparison of unfolded data and predictions from MG5+PY8 FxFx reweighted with two PDFs containing different fractions of IC. Events with at least one b-jet, electron+muon channel.

6.2 Variables and uncertainties

Once the events have been reweighted, a range of y is decided to determine the forward and central region. The scan was performed from $y = 1.2$ to $y = 2$, with a step $\Delta y = 0.2$. Examining the forward/central ratio allows for many uncertainties to cancel. The kinematic variables taken into consideration were either the p_T of the Z boson or of the leading b-jet

(LBJ), or their x_F. The rapidity cut was performed either on the Z or on the LBJ. A detailed overview of the variables of interest is given in Table 6.2

Variable	Cut on y_Z	Cut on y_{LBJ}
p_{T_Z}	✓	✓
$p_{T_{LBJ}}$	✓	✓
x_{F_Z}	X	X
$x_{F_{LBJ}}$	✓	X

Table 6.2: Variables subject to optimization

For each variable, the ratio of Forward over Central region was calculated. Both statistical (Poisson) and systematic uncertainties have been taken into consideration. An exhaustive list of uncertainties can be found in Section 5.8. Theoretical uncertainties (scale and PDF variations) have been excluded because of their negligible effect. The total error is, bin by bin, the square sum of the contribution of each uncertainty. Calculating the ratio helps reducing the impact of systematic uncertainties.

6.3 Significance

In order to quantify the sensitivity to the presence of intrinsic charm, significance was calculated. For each variable, the ratios produced from samples generated with (1%) or without (0%) intrinsic charm were compared in this formula:

$$\chi^2 = \sum_{i=1}^{N} \frac{(v^i_{0\%} - v^i_{1\%})^2}{\sigma^2_{0\%}} \tag{6.2}$$

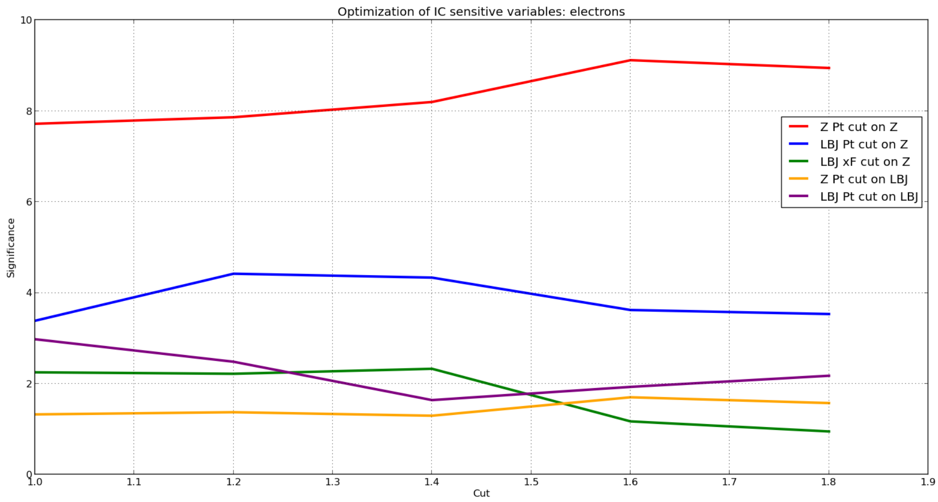
where σ is the sum in quadrature of all the errors of the ratio histograms for each bin:

$$\sigma = \varepsilon^2_{stat} + \sum_{j=1}^{n} \varepsilon^2_{syst,j} \tag{6.3}$$

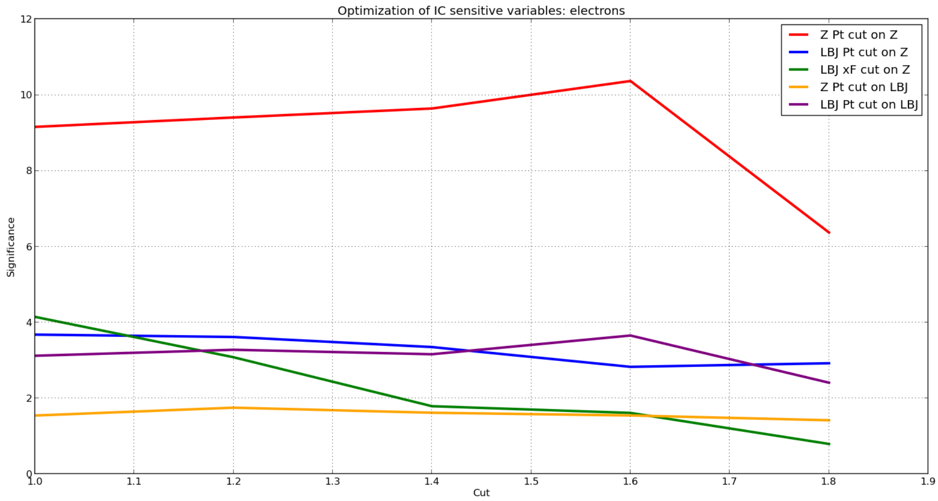
As seen in Figure 6.6 , the optimal variable is the p_T of the Z boson, with a cut on y_Z. The significance does not vary much across cuts, so the choice fell on $y = 1.2$. A more conservative choice allows to have higher statistics in the forward region. Figures 6.5a and 6.6a show significances for electrons and muons final states reweighted with the CTEQ14 PDFs, whereas Figures 6.5b and 6.6b with NNPDF 3.1.

6.4 Results

The Central/Forward differential cross sections in the p_T of the Z boson ratios are presented in Figures 6.7a, 6.7b and 6.8. The systematics breakdown for combined leptonic signals can be seen in Figure 6.9. Figure 6.10 shows the Forward/Central cross section ratio for Z p_T.

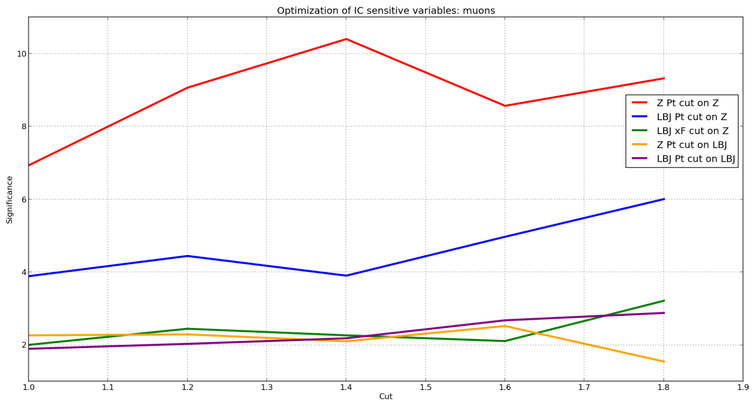


(a) CTEQ 14

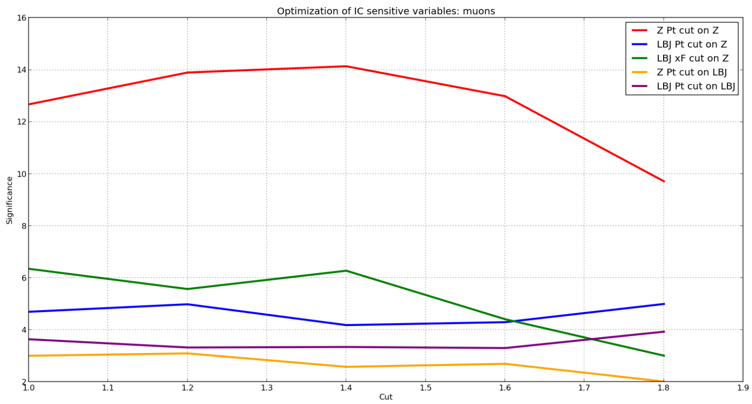


(b) NNPDF 3.1

Figure 6.5: Rapidity cut scans for kinematic variables for two different sets of PDFs, $Z \to e^+e^-$.

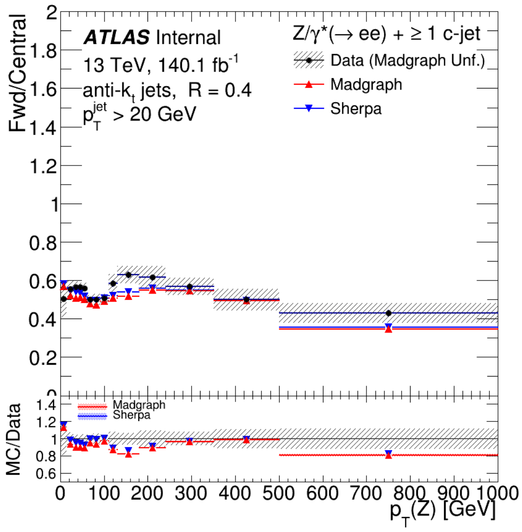


(a) CTEQ 14

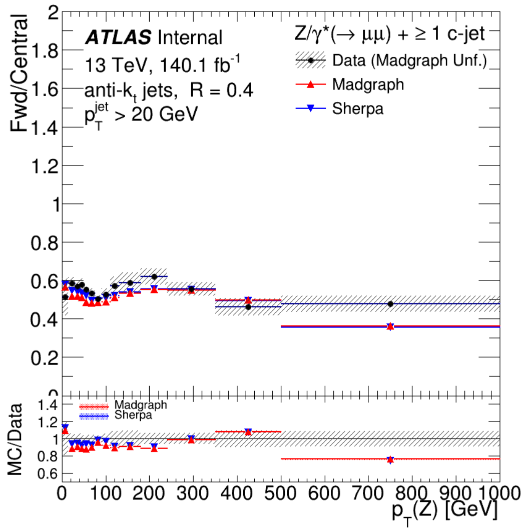


(b) NNPDF 3.1

Figure 6.6: Rapidity cut scans for kinematic variables for two different sets of PDFs, $Z \to \mu^+\mu^-$.



(a) Signal is $Z \to e^+e^-$, `ZGeq1C`.



(b) Signal is $Z \to \mu^+\mu^-$, `ZGeq1C`.

Figure 6.7: Differential cross section Forward/Central ratio for Z p_T in the two leptonic channels

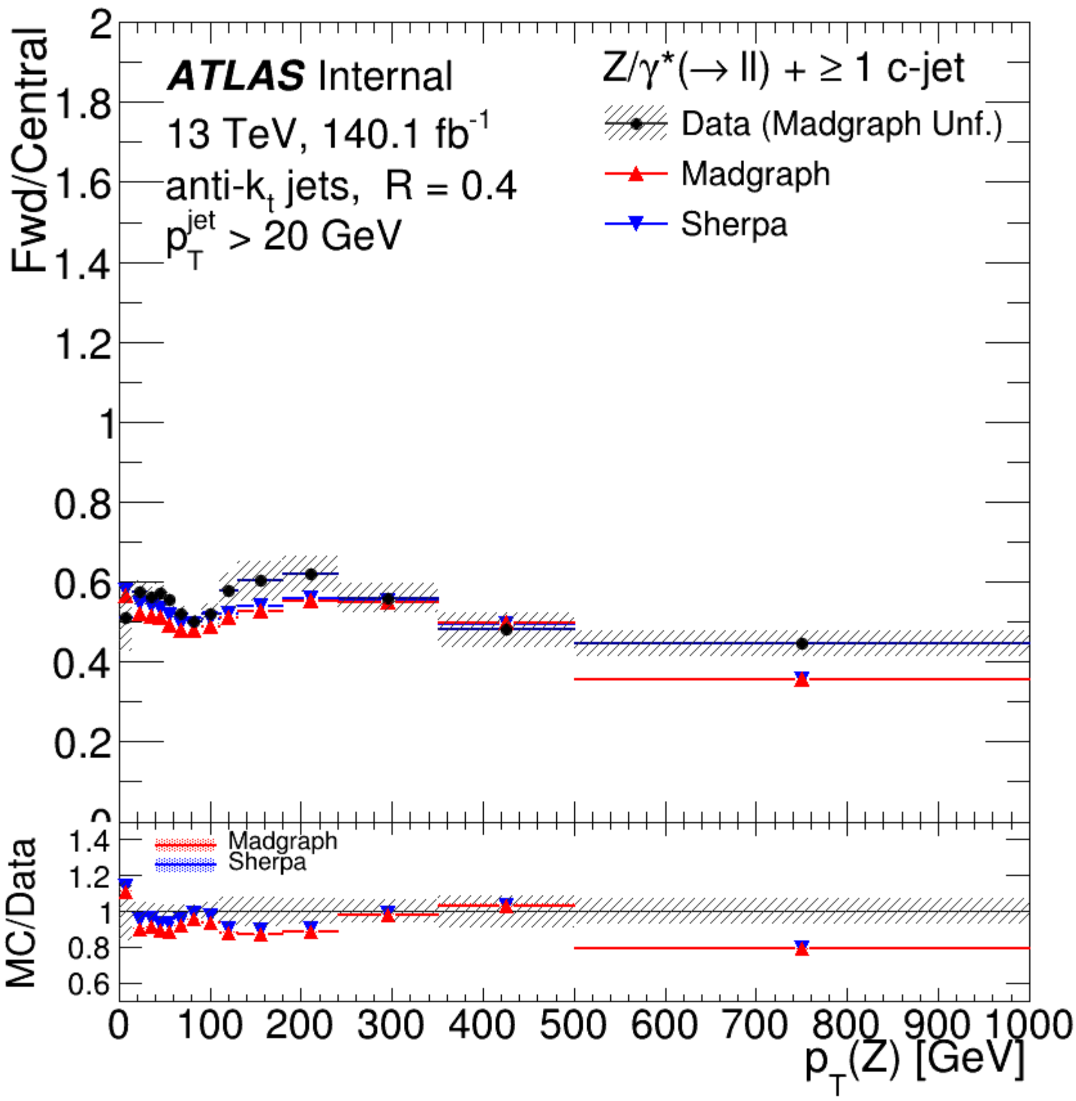


Figure 6.8: Differential cross section Forward/Central ratio for Z p_T. Signal is $Z \to \ell^+\ell^-$, `ZGeq1C`.

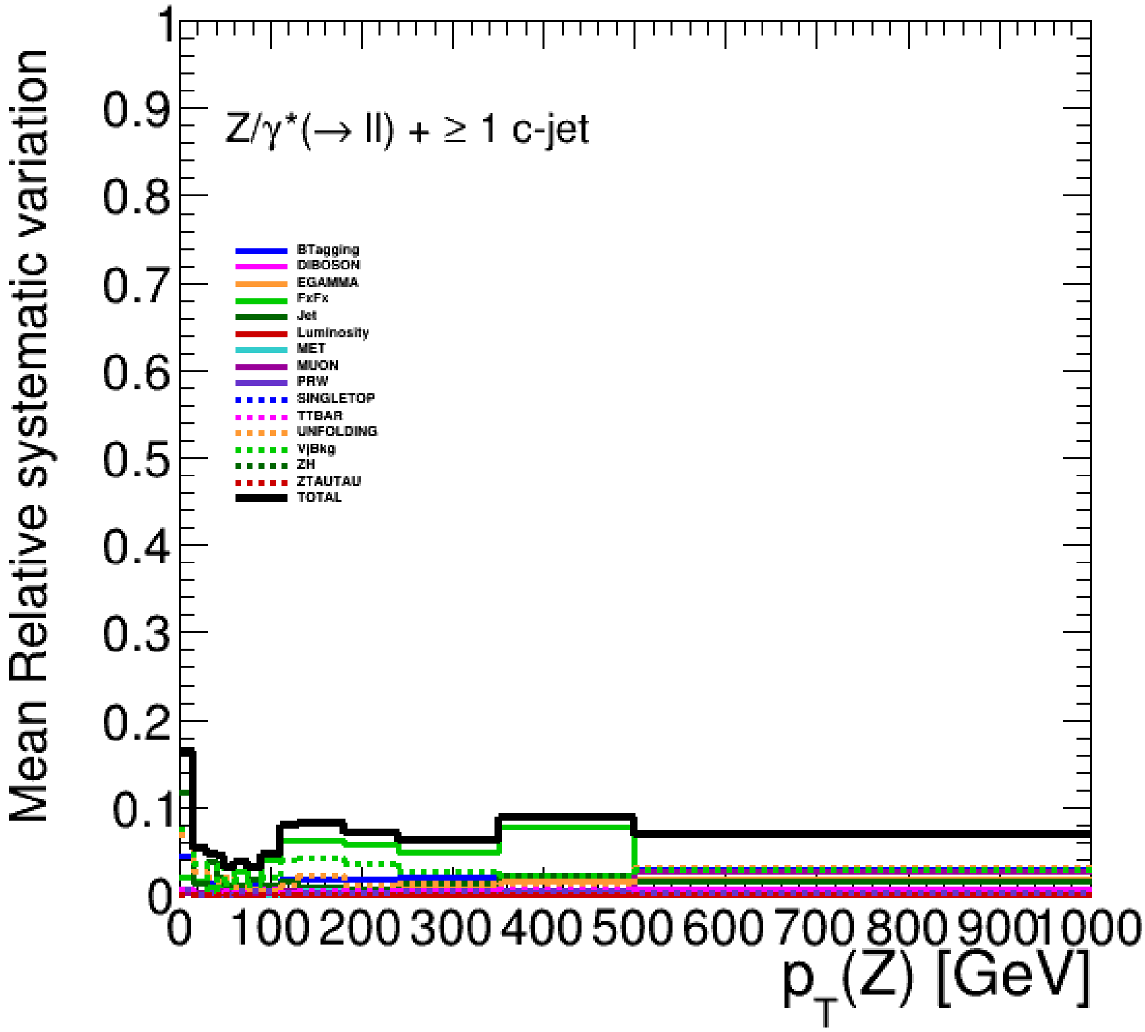


Figure 6.9: Systematics breakdown for Z p_T. Signal is $Z \to \ell^+\ell^-$, `ZGeq1C`.

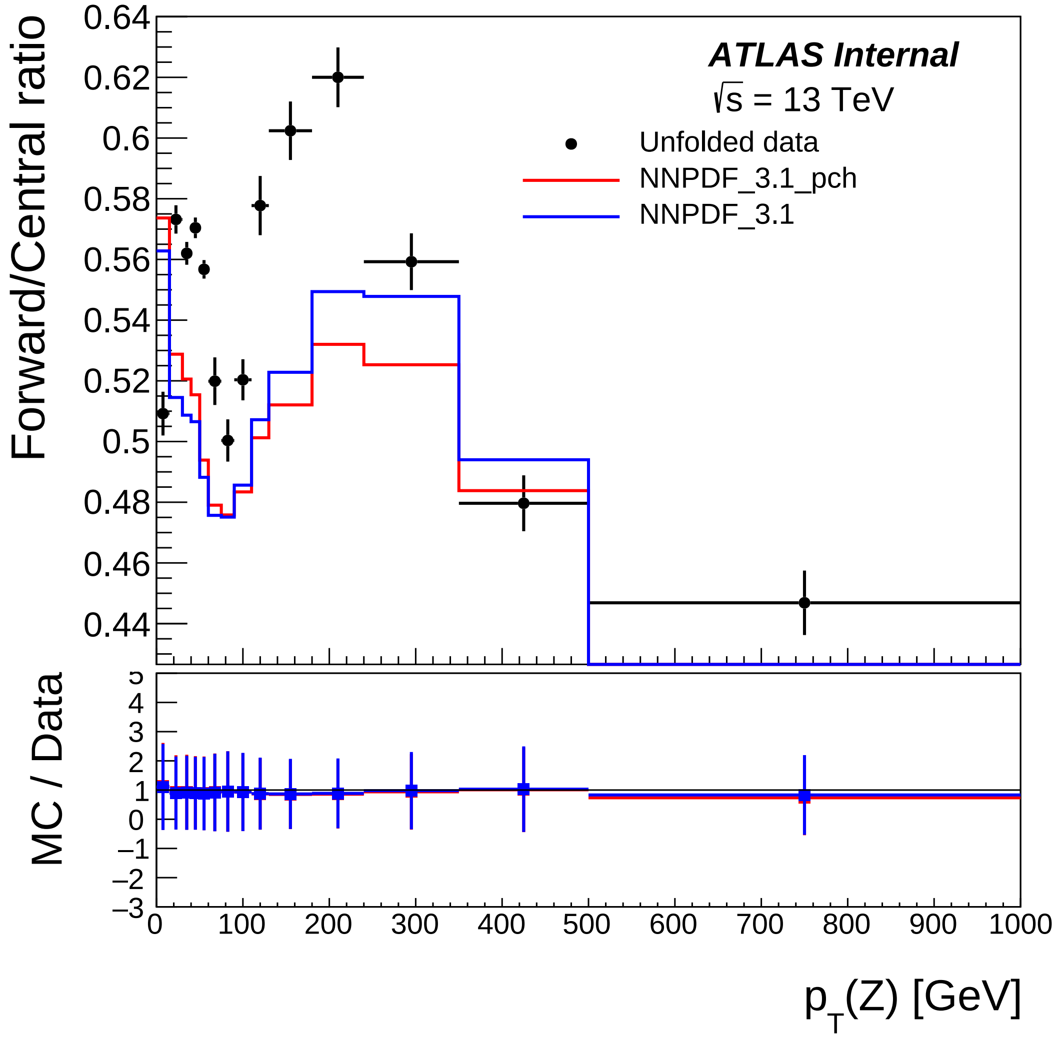


Figure 6.10: Forward/Central ratio for unfolded data and simulated samples reweighted with different fractions of intrinsic charm: in red, with `NNPDF_3.1_pch` which only contains perturbative charm, in blue `NNPDF_3.1` with intrinsic charm. Signal is $Z \to \ell^+\ell^-$, `ZGeq1C`. Bars represent statistical errors.

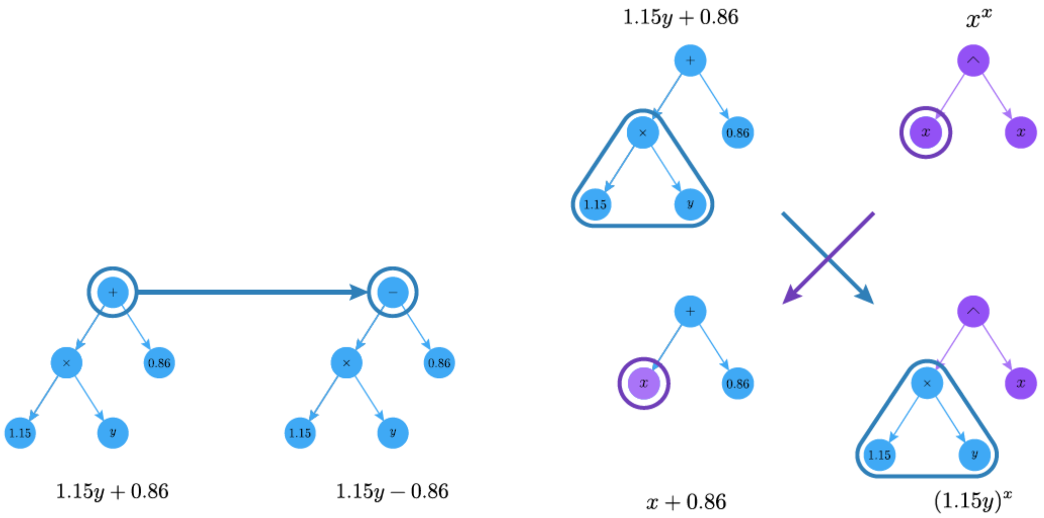
Chapter 7

Symbolic regression

Parton Density Functions, as seen in Chapter 2.3, give a description of the momenta distribution of the quarks inside the proton. Whenever strong interactions are involved in a simulation, a choice of PDF must be made. The largest pre-computed PDF repository is LHAPDF (Les Houches Accord PDF), which hosts PDFs of all kinds from different collaborations. The PDFs are impossible to be determined theoretically, they have to be extracted from data and each collaboration employs a different technique: the NNPDF [40] collaboration uses neural networks trained on toy-samples, HERAPDF uses global fits on deep inelastic scattering, meanwhile CTEQ [41] uses data from DIS, Tevatron and more.
These PDFs are presented in the repository as a collection of points at various energy scales. They are the results of fits on experimental data and the evolution from the DGLAP equations Parameterizing these functions is without a doubt an interesting exercise, that has been approached in different ways in the last years. In this chapter, a novel approach, which involves a machine learning technique called symbolic regression, will be discussed.

7.1 Introduction to machine learning

Machine learning is a form of artificial intelligence that uses algorithms that can learn from data without being explicitly programmed. It plays a fundamental role in particle physics because it can be applied to a multitude of cases, whenever it is necessary to interpret complex systems. It is most commonly used in tracking and vertex finding, flavour tagging and event classification. Every ML problem is unique in its own way, but most of them have some features in common. For instance, data is always split between training and validation sets: the former is used by the model to learn, the latter to validate the outcome of the training. The discrepancy between the predictions made by the model and the target values is quantified by the loss function $\mathcal{L}$. The goal is minimizing this function, whose values are an expression of the distance between the values predicted by the model and the actual values. There are many types of loss functions: in regression problems, the most common one is the Mean Squared Error function, whereas in classification problems log-likelihood and binary cross entropy functions are most widely used.



(a) Mutation (b) Crossover

Figure 7.1: Examples of crossovers and mutations [45].

7.2 Symbolic regression

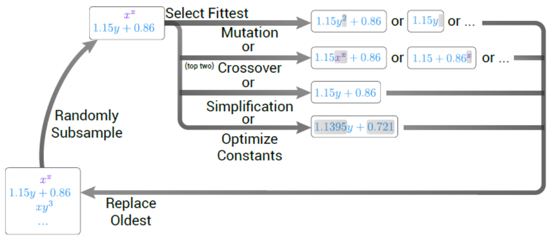
Symbolic regression is a machine learning technique whose goal is extracting functional forms from datasets. It is, more precisely, a supervised learing task which spans the space of mathematical expressions and combines them until they approximate in a satisfying way the data it is given. In this thesis, the program that was used to implement it is PySR [44], which is a multi-population evolutionary algorithm. Generally, evolutionary algorithms are based on tournament selection and use mutations and crossover to generate new individuals, as seen in Figure 7.1a and 7.1b.

Given a population, the algorithm randomly selects a subset of individuals that undergo a tournament to determine the fittest. The winner randomly receives a mutation and replaces the weakest member of the population (in the specific case of PySR, the evolution is "age-regularized", which means that the first substitutions are for older memebers rather than least fit). The tournament is parallelized between populations, and asynchronous migration is implemented between groups.
PySR introduces three modfications to the classic evolutionary algorithm:

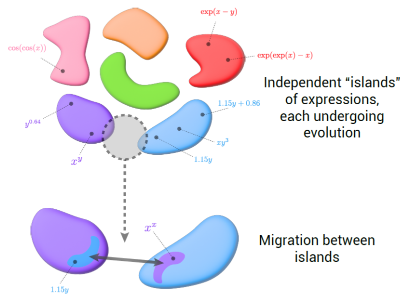
1. **annealing**: to speed up the selection process, there is a non-zero probability of rejection of a mutation based on the difference of fitness pre- and post-mutation;

2. **evolve-simplify-optimize loop**: during the evolution process, equations are occasionally simplified and optimize to obtain simpler equations with numerical constants;

3. **adaptive parsimony**: it punishes complexity and excessive early specialization



(a) Inner loop



(b) Outer loop

Figure 7.2: Inner and outer PySR loops, in a single population and between populations respectively [45].

adding a term to the loss $\ell_{pred}(E)$

$$\ell(E) = \ell_{pred}(E) \cdot \exp(\text{frecency})[\text{C(E)}] \tag{7.1}$$

where *frecency* is a combination of frequency and recency at complexity C(E).

7.3 HERA PDF

In order to fine tune the program and get the hang of its parameters, a first attempt at parameterizing a known PDF was made. The choice fell on the HERAPDF 2.0 NLO, because a parameterization with xFitter [4, 46] was already available to be compared with the regression results. The parameters for symbolic regression are in Table 7.1

Parameter	Value
`npop`	15
`population_size`	33
`niterations`	90
`ncyclesperiteration`	5000
`parsimony`	10^{-7}
`unary_operators`	$1-x$, x, $log(x)$
`binary_operators`	$+$, $-$, $\times$, $\wedge$, $\div$

Table 7.1: Main parameters in PySR program.

The HERA collaboration presented the following parameterization for the valence down PDF:

$$xd_v^{(H)} = 3.15x^{0.806}(x-1)^{4.08} \tag{7.2}$$

Meanwhile, the output of PySR was:

$$xd_v^{(SR)} = 3.37x^{0.837}(x-1)^{4.0} \tag{7.3}$$

with loss $\mathcal{L} = 1.6 \cdot 10^{-6}$ As seen in Figure 7.3, this function approximates the PDF in a satisfying way, even better (Figure 7.4) than xFitter's alternative.

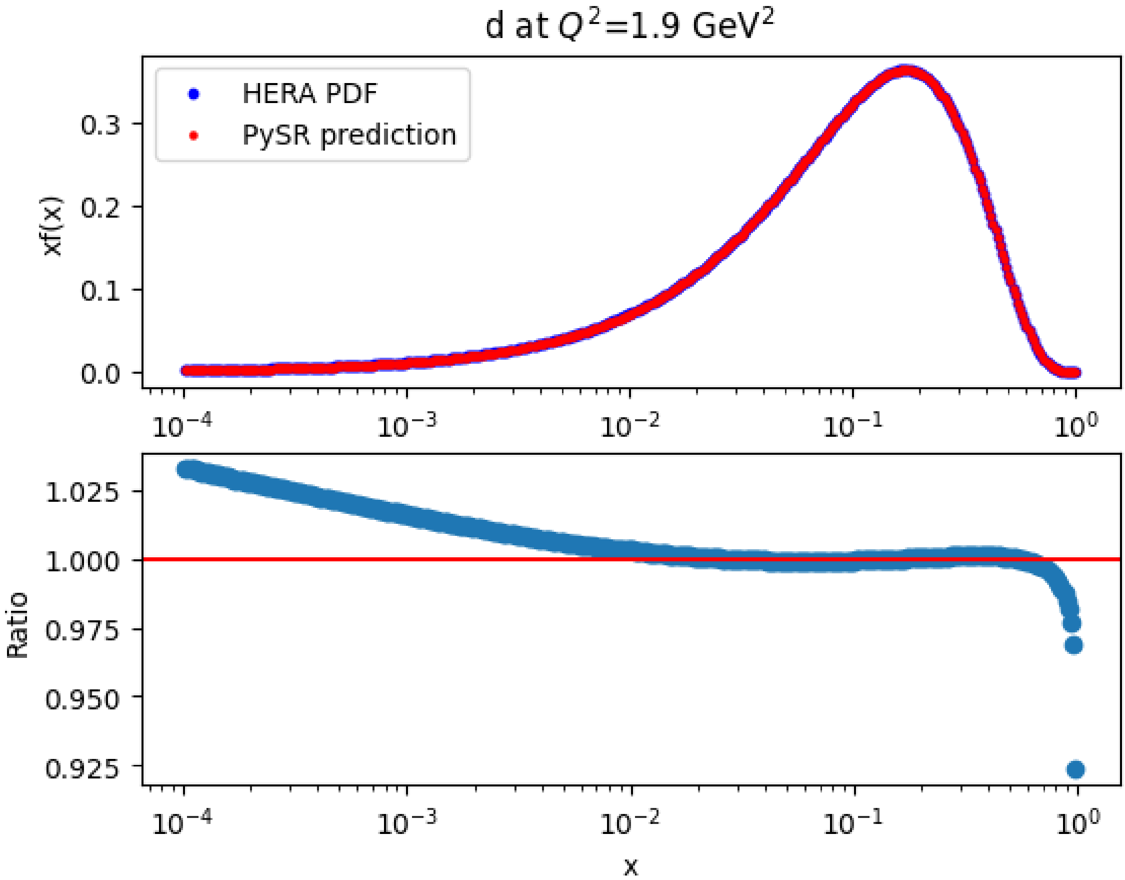


Figure 7.3: Symbolic regression (red) on HERA PDF(blue): the chosen PDF is the down PDF at $Q^2 = 1.9\text{GeV}^2$.

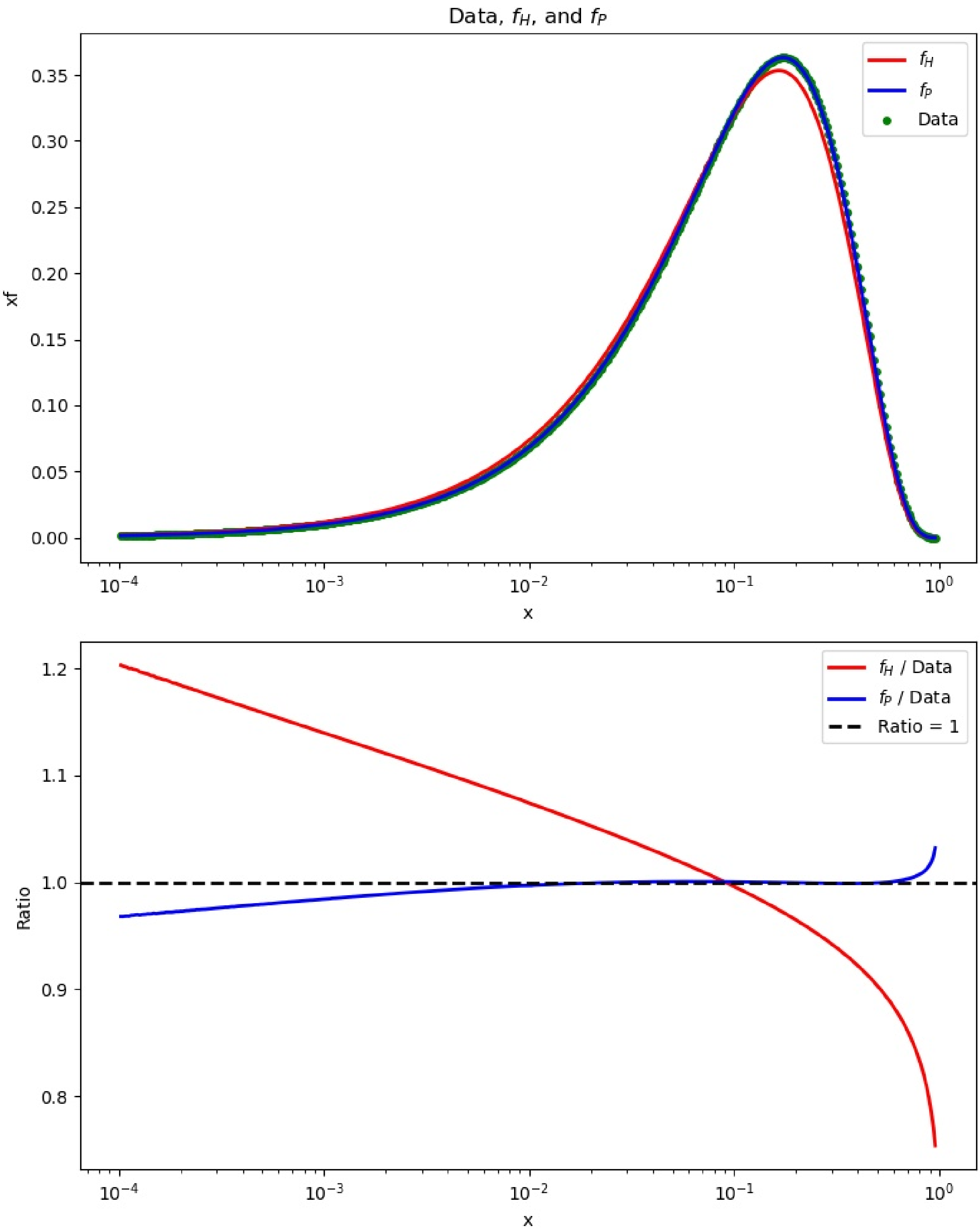


Figure 7.4: Comparison of parameterizations: f_H is the HERA collaboration PDF (red), while f_P is the PySR output (blue)

7.4 Intrinsic charm in NNPDF 3.1 and CTEQ14

Both **NNPDF 3.1 NNLO** and **CT14 NNLO** are PDF sets which include intrinsic charm, although in a different way. NNPDF 3.1 has the intrinsic charm component fit in: that is, the neural network that fits the function to the data takes into consideration the intrinsic charm model. On the contrary, CT14 has variations which have the intrinsic charm component manually added in. This is evident in Figure 7.5, where the "manually" added intrinisc charm component appears smoother than the fit one. Both components were obtained by subtracting the charm PDF that only contains perturbative charm from the total charm PDF. Nonetheless, they both have valence-like shapes: a peak in the middle of the x-interval, and a compatibility with 0 at low x. In the specific case of NNPDF, the advantage of using symbolic regression on the output of a neural network is two-fold: generally speaking, neural networks are black boxes, and extracting their functional form could have interesting phenomenological applications. On the other hand, NNPDF's neural network has 9 free parameters, whereas most PDF parameterization have fewer, so determining the differences in the two approaches could be of interest when it comes to optimizing the degrees of freedom.

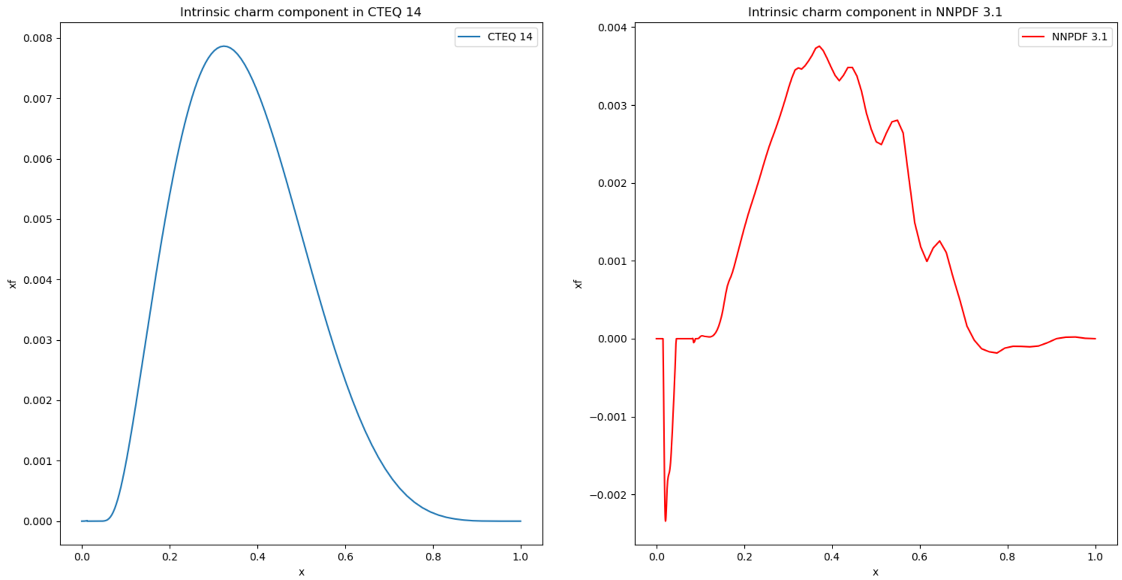


Figure 7.5: Intrinsic charm components for NNPDF 3.1 and CTEQ 14.

In the following tables are possible candidates to model the intrinsic charm component in the two PDFs.

Equation	Complexity	Loss	Score
$y = 0$	1	$1.04 \cdot 10^{-6}$	0.0
$y = 0.00253x$	2	$6.87 \cdot 10^{-7}$	0.416
$y = 0.0101x^{1.09} \cdot (1 - x)$	4	$2.79 \cdot 10^{-7}$	0.451
$y = 0.0103x^{1.09} \cdot (1 - 1.09x)$	5	$2.45 \cdot 10^{-7}$	0.129
$y = 0.0114x^{1.09} \cdot (1 - 1.09x)$	6	$2.39 \cdot 10^{-7}$	0.0278
$y = 0.0473x^2 (1 - x)^2$	8	$1.75 \cdot 10^{-7}$	0.154
$y = x^2 \cdot (0.0858 - 0.0858x) (0.775 - x)$	9	$7.91 \cdot 10^{-8}$	0.795
$y = 0.105x \cdot (0.766 - x) (1 - x) (x - 0.0497)$	10	$6.67 \cdot 10^{-8}$	0.170
$y = 0.105x \cdot (0.771 - x) (1 - 1.04x) (x - 0.0497)$	11	$5.69 \cdot 10^{-8}$	0.160
$y = x \cdot (0.121 - 0.127x) (0.752 - x) (x - 0.0655)$	12	$5.30 \cdot 10^{-8}$	0.0702
$y = 0.400x^{2.84} \cdot (0.727 - x) (1 - x)^2$	13	$3.42 \cdot 10^{-8}$	0.438
$y = x^{3.28} \cdot (0.738 - x) (1 - x)^3$	14	$1.77 \cdot 10^{-8}$	0.657
$y = x^{1.90} \cdot (0.727 - x) (1 - x)^3 (x - 0.0968)$	15	$1.13 \cdot 10^{-8}$	0.449
$y = x^{1.90} \cdot (0.738 - x) (1 - 1.08x) (1 - x)^2 (x - 0.0948)$	16	$9.75 \cdot 10^{-9}$	0.150
$y = x^2 (1 - x)^4 \cdot (1.16x \cdot (1 - x) - 0.103)$	17	$8.89 \cdot 10^{-9}$	0.0917
$y = x^2 \cdot (1 - 1.08x) (1 - x)^3 \cdot (1.28x \cdot (1 - x) - 0.119)$	18	$8.51 \cdot 10^{-9}$	0.0441
$y = x \cdot (1 - 1.08x) (1 - x)^3 (x - 0.00206) (1.28x \cdot (1 - x) - 0.119)$	19	$8.40 \cdot 10^{-9}$	0.0132
$y = x \cdot (0.834 - x) (1 - x)^3 (x - 0.102) (1.54x \cdot (1 - x) - 0.0393)$	20	$7.53 \cdot 10^{-9}$	0.109

Table 7.2: NNPDF 3.1: `hall of fame` of PySR.

Equation	Complexity	Loss	Score
$y = -0.000693$	1	$7.76 \cdot 10^{-6}$	0.0
$y = 0.00142$	2	$4.97 \cdot 10^{-6}$	0.447
$y = 0.00576x$	3	$4.04 \cdot 10^{-6}$	0.205
$y = 0.00440x + 0.000597$	4	$3.85 \cdot 10^{-6}$	0.0488
$y = 0.783x \cdot (0.0283 - 0.0283x)$	5	$1.42 \cdot 10^{-6}$	0.998
$y = 0.0291x \cdot (1 - 1.09x)$	6	$1.28 \cdot 10^{-6}$	0.105
$y = 0.619x \cdot (0.0593 - 0.0593x)(1 - x)$	7	$9.89 \cdot 10^{-7}$	0.257
$y = x^{1.33} \cdot (0.0677 - 0.0677x)(1 - x)$	8	$8.23 \cdot 10^{-7}$	0.184
$y = 0.0792x^{1.60} \cdot (1 - 1.25x)(1 - x)$	9	$6.08 \cdot 10^{-7}$	0.302
$y = 1.94x^{1.58} \cdot (0.0525 - 0.0691x)(1 - x)$	10	$3.52 \cdot 10^{-7}$	0.547
$y = 2.42x^{0.568} \cdot (0.0476 - 0.0652x)(1 - x)(x - 0.0273)$	11	$2.95 \cdot 10^{-7}$	0.176
$y = 2.42x^{0.557} \cdot (0.0504 - 0.0691x)(1 - 1.06x)(x - 0.0273)$	12	$2.36 \cdot 10^{-7}$	0.223
$y = 0.0813x \cdot (1 - 1.35x)(1 - x)^2 \cdot \left(3.34x^{0.545} - 0.696\right)$	13	$7.64 \cdot 10^{-8}$	1.13
$y = 0.0813 \cdot (1 - 1.35x)(1 - x)^2 \cdot \left(3.34x^{0.545} - 0.696\right)(x - 0.00117)$	14	$7.00 \cdot 10^{-8}$	0.0875
$y = (0.762 - x)(0.630x - 0.493)(x - 0.0480)\left(x^2 - x\right)$	15	$4.55 \cdot 10^{-8}$	0.430
$y = (0.762 - x)(0.641x - 0.502)(x - 0.0480)\left(1.03x^2 - x\right)$	16	$3.96 \cdot 10^{-8}$	0.139
$y = (0.762 - x)(0.641x - 0.502)(x - 0.0476)\left(1.03x^2 - x + 0.00166\right)$	17	$3.88 \cdot 10^{-8}$	0.0215
$y = (0.762 - x)(0.653x - 0.512)(x - 0.0476)\left(1.03x^2 - x + 0.00280\right)$	18	$3.82 \cdot 10^{-8}$	0.0153
$y = (0.653x - 0.512)((0.762 - x)(x - 0.0476) - 0.000283)\left(1.03x^2 - x + 0.00280\right)$	19	$3.80 \cdot 10^{-8}$	0.00510
$y = (0.767 - 1.02x)(0.647x - 0.508)(x - 0.0473)\left(1.04x^2 - 1.01x + 0.00448\right)$	20	$3.76 \cdot 10^{-8}$	0.0120

Table 7.3: CTEQ 14: `hall of fame` of PySR.

Bibliography

[1] Mark Thomson. *Modern particle physics*. Cambridge University Press, 2013.

[2] Matthew D Schwartz. *Quantum field theory and the standard model*. Cambridge university press, 2014.

[3] Francis Halzen and Alan D Martin. *Quark & Leptons: An introductory course in modern particle physics*. John Wiley & Sons, 2008.

[4] ZEUS Collaborations et al. Combination of measurements of inclusive deep inelastic scattering cross sections and qcd analysis of hera data. *arXiv preprint arXiv:1506.06042*, 2015.

[5] Guido Altarelli and Giorgio Parisi. Asymptotic freedom in parton language. *Nuclear Physics B*, 126(2):298–318, 1977.

[6] Yuri L Dokshitzer. Calculation of the structure functions for deep inelastic scattering and e+ e-annihilation by perturbation theory in quantum chromodynamics. *Zh. Eksp. Teor. Fiz*, 73:1216, 1977.

[7] Vladimir Naumovich Gribov and Lev N Lipatov. Deep inelastic ep-scattering in a perturbation theory. Technical report, Inst. of Nuclear Physics, Leningrad, 1972.

[8] The xFitter Developers' Team, :, H. Abdolmaleki, S. Amoroso, V. Bertone, M. Botje, D. Britzger, S. Camarda, A. Cooper-Sarkar, J. Fiaschi, F. Giuli, A. Glazov, C. Gwenlan, F. Hautmann, H. Jung, A. Kusina, A. Luszczak, T. Mäkelä, I. Novikov, F. Olness, R. Sadykov, P. Starovoitov, M. Sutton, and O. Zenaiev. xfitter: An open source qcd analysis framework. a resource and reference document for the snowmass study, 2022. arXiv:2206.12465.

[9] Stefano Forte, Lluí s Garrido, José I Latorre, and Andrea Piccione. Neural network parametrization of deep-inelastic structure functions. *Journal of High Energy Physics*, 2002(05):062–062, may 2002. URL: https://doi.org/10.1088%2F1126-6708%2F2002%2F05%2F062, doi:10.1088/1126-6708/2002/05/062.

[10] https://nnpdf.mi.infn.it/research/neural-networks/.

[11] Stanley J Brodsky, P Hoyer, C Peterson, and N Sakai. The intrinsic charm of the proton. *Physics Letters B*, 93(4):451–455, 1980.

[12] Roel Aaij, Ahmed Sameh Wagih Abdelmotteleb, C Abellan Beteta, FJ Abudinen Gallego, Thomas Ackernley, Bernardo Adeva, Marco Adinolfi, Hossein Afsharnia, Christina Agapopoulou, Christine Angela Aidala, et al. Study of z bosons produced in association with charm in the forward region. *Physical review letters*, 128(8):082001, 2022.

[13] the CMS collaboration. Measurement of differential cross sections for z bosons produced in association with charm jets in pp collisions at $$ \sqrt{s} $$ = 13 TeV. *Journal of High Energy Physics*, 2021(4), apr 2021. URL: https://doi.org/10.1007%2Fjhep04%282021%29109, doi:10.1007/jhep04(2021)109.

[14] Evidence for intrinsic charm quarks in the proton. *Nature*, 608(7923):483–487, 2022.

[15] Stefan Höche. Introduction to parton-shower event generators, 2015. arXiv:1411.4085.

[16] Peter Skands. Qcd for collider physics, 2012. arXiv:1104.2863.

[17] Lucio Rossi. The lhc main dipoles and quadrupoles toward series production. *IEEE transactions on applied superconductivity*, 13(2):1221–1228, 2003. doi:https://doi.org/10.1109/TASC.2003.812639.

[18] Jorg Wenninger. Sissa: Lhc status and performance. *PoS*, page 001, 2019. doi:https://doi.org/10.22323/1.339.0001.

[19] Nicoletta Garelli. Performance of the atlas detector in run-2. In *EPJ Web of Conferences*, volume 164, page 01021. EDP Sciences, 2017. doi:https://doi.org/10.1051/epjconf/201716401021.

[20] Georges Aad, Xabier Sebastian Anduaga, S Antonelli, M Bendel, B Breiler, F Castrovillari, JV Civera, T Del Prete, Maria Teresa Dova, S Duffin, et al. The atlas experiment at the cern large hadron collider. 2008. doi:https://doi.org/10.1088/1748-0221/3/08/S08003.

[21] https://twiki.cern.ch/twiki/bin/view/AtlasPublic/TriggerOperationPublicResults.

[22] The ATLAS collaboration. Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton–proton collision data at $$\sqrt{s} = 13$$ $$\text {TeV}$$. *The European Physical Journal C*, 79(8), aug 2019. URL: https://doi.org/10.1140%2Fepjc%2Fs10052-019-7140-6, doi:10.1140/epjc/s10052-019-7140-6.

[23] https://twiki.cern.ch/twiki/bin/view/AtlasPublic/EventDisplayRun2Physics.

[24] Gavin P Salam. Towards jetography. *The European Physical Journal C*, 67:637–686, 2010.

[25] Stefano Catani, Yu L Dokshitzer, Michael H Seymour, and Bryan R Webber. Longitudinally-invariant kt-clustering algorithms for hadron-hadron collisions. *Nuclear Physics B*, 406(1-2):187–224, 1993.

[26] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. The anti-kt jet clustering algorithm. *Journal of High Energy Physics*, 2008(04):063, 2008.

[27] ATLAS Collaboration et al. Atlas flavour-tagging algorithms for the lhc run 2pp collision dataset. *arXiv preprint arXiv:2211.16345*, 2022.

[28] CDF Collaboration. Measurement of cross sections for b jet production in events with a z boson in p p^{-} collisions at s= 1.96 tev. *Physical Review D*, 79(5), mar 2009. URL: https://doi.org/10.1103%2Fphysrevd.79.052008, doi:10.1103/physrevd.79.052008.

[29] D0 Collaboration. Measurement of the ratio of differential cross sections σ (p p$^{-}\rightarrow$ z+ b jet)/σ (p p$^{-}\rightarrow$ z+ jet) in p p^{-} collisions at s= 1.96 tev. *Physical Review D*, 87(9), may 2013. URL: https://doi.org/10.1103%2Fphysrevd.87.092010, doi:10.1103/physrevd.87.092010.

[30] ATLAS Collaboration. Measurement of differential production cross-sections for a z boson in association with b-jets in 7 TeV proton-proton collisions with the ATLAS detector. *Journal of High Energy Physics*, 2014(10), oct 2014. URL: https://doi.org/10.1007%2Fjhep10%282014%29141, doi:10.1007/jhep10(2014)141.

[31] CMS Collaboration. Measurement of the production cross sections for a z boson and one or more b jets in pp collisions at s \$\$ \sqrt{s} \$\$ = 7 TeV. *Journal of High Energy Physics*, 2014(6), jun 2014. URL: https://doi.org/10.1007%2Fjhep06%282014%29120, doi:10.1007/jhep06(2014)120.

[32] ATLAS collaboration. Measurements of the production cross-section for a z boson in association with b-jets in proton-proton collisions at \$\$ \sqrt{s} \$\$ = 13 TeV with the ATLAS detector. *Journal of High Energy Physics*, 2020(7), jul 2020. URL: https://doi.org/10.1007%2Fjhep07%282020%29044, doi:10.1007/jhep07(2020)044.

[33] ATLAS collaboration. Cross-section measurements for the production of a z boson in association with high-transverse-momentum jets in pp collisions at \$\$ \sqrt{s} \$\$ = 13 TeV with the ATLAS detector. *Journal of High Energy Physics*, 2023(6), jun 2023. URL: https://doi.org/10.1007%2Fjhep06%282023%29080, doi:10.1007/jhep06(2023)080.

[34] Rikkert Frederix and Stefano Frixione. Merging meets matching in MC@NLO. *Journal of High Energy Physics*, 2012(12), dec 2012. URL: https://doi.org/10.1007%2Fjhep12%282012%29061, doi:10.1007/jhep12(2012)061.

[35] Christian Bierlich, Smita Chakraborty, Nishita Desai, Leif Gellersen, Ilkka Helenius, Philip Ilten, Leif Lönnblad, Stephen Mrenna, Stefan Prestel, Christian Tobias Preuss, et al. A comprehensive guide to the physics and usage of pythia 8.3. *SciPost Physics Codebases*, page 008, 2022.

[36] Andy Buckley, Jonathan Butterworth, David Grellscheid, Hendrik Hoeth, Leif Lönnblad, James Monk, Holger Schulz, and Frank Siegert. Rivet user manual. *Computer Physics Communications*, 184(12):2803–2819, 2013.

[37] Glen Cowan. A survey of unfolding methods for particle physics. In *Conf. Proc. C*, volume 203181, pages 248–257, 2002.

[38] Giulio D'Agostini. A multidimensional unfolding method based on bayes' theorem. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 362(2-3):487–498, 1995.

[39] https://hepunx.rl.ac.uk/~adye/software/unfold/RooUnfold.html.

[40] Richard D Ball, Valerio Bertone, Stefano Carrazza, Luigi Del Debbio, Stefano Forte, Patrick Groth-Merrild, Alberto Guffanti, Nathan P Hartland, Zahari Kassabov, José I Latorre, et al. Parton distributions from high-precision collider data: Nnpdf collaboration. *The European Physical Journal C*, 77:1–75, 2017.

[41] Sayipjamal Dulat, Tie-Jiun Hou, Jun Gao, Joey Huston, Jon Pumplin, Carl Schmidt, Daniel Stump, and C-P Yuan. Intrinsic charm parton distribution functions from cteq-tea global analysis. *Physical Review D*, 89(7):073004, 2014.

[42] V.A. Bednyakov, M.A. Demichev, G.I. Lykasov, T. Stavreva, and M. Stockton. Searching for intrinsic charm in the proton at the LHC. *Physics Letters B*, 728:602–606, jan 2014. URL: https://doi.org/10.1016%2Fj.physletb.2013.12.031, doi:10.1016/j.physletb.2013.12.031.

[43] P.-H. Beauchemin, V. A. Bednyakov, G. I. Lykasov, and Yu. Yu. Stepanenko. Search for intrinsic charm in vector boson production accompanied by heavy-flavor jets. *Physical Review D*, 92(3), aug 2015. URL: https://doi.org/10.1103%2Fphysrevd.92.034014, doi:10.1103/physrevd.92.034014.

[44] Miles Cranmer, Alvaro Sanchez Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. Discovering symbolic models from deep learning with inductive biases. *Advances in Neural Information Processing Systems*, 33:17429–17442, 2020.

[45] Miles Cranmer. Interpretable machine learning for science with pysr and symbolicregression.jl, 2023. arXiv:2305.01582.

[46] Marco Bonvini and Francesco Giuli. A new simple pdf parametrization: improved description of the hera data. *The European Physical Journal Plus*, 134(10):531, 2019.

Appendix A Madgraph Runcard

Example of a Madgraph runcard: after specifying the model, in this case the Standard Model with loop diagrams, the contents of the proton are specified. The chosen process is then generated, specifying that QCD has to be taken into consideration. The output is then saved to a directory, from where the events can later be simulated, specifying the number of events, the PDF and the systematics.

```
import model loop_sm

define p = g u c d s u~ c~ d~ s~

define j = g u c d s u~ c~ d~ s~

generate p p > l+ l- b b~ [QCD]

output zb
```

Appendix B PySR model

An example of PySR model, specifying the operators that the algorithm must combine, their complexity and their nesting. An early stop condition is also implemented, commanding the model to stop if it reaches convergence with a simple enough function.

```
model = PySRRegressor(

    loss="loss(prediction,target) =(prediction - target)^2",

    niterations=50,
    binary_operators=["plus", "sub", "mult", "pow"],
    unary_operators=["neg", "xminus(x)=x-1","linear(x)=x","quad(x)=x^2","
    cube(x)=x^3","log"],

    extra_sympy_mappings={"xminus": lambda x: 1-x, "linear": lambda x: x, "
    quad": lambda x: x**2, "cube": lambda x: x**3}
    constraints={ "pow" : (7,1)},

    complexity_of_operators={
        "pow": 0
        "linear": 1,
        "xminus": 1
    },
    nested_constraints={
        "pow":{"pow":0,"linear":0,"xminus":0},
        "linear":{"xminus":0, "^":0},
        "xminus":{"linear":0, "^":0}
    },
    maxsize=7,
    early_stop_condition=(
        "stop_if(loss, complexity) = loss < 1e-5 && complexity < 8"
        # Stop early if we find a good and simple equation
    ),
    maxdepth=15
    complexity_of_constants=0
```