Parton Distribution Functions at the High Precision LHC



Shaun Bailey

Wadham College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity Term 2021

Abstract

In this thesis, we discuss current progress made in the determination of the Parton Distribution Functions (PDFs) in the context of the high–precision LHC era. The study of these quantities not only furthers our understanding of the composite nature of the proton but is also a vital ingredient for many Standard Model and Beyond the Standard Model phenomenological studies performed at the LHC. To this end, we examine four recent studies in this area, covering a wide range of topics.

We first discuss the challenges currently being faced due to the highly correlated and precise data that are available at the LHC and how this can be overcome in the context of top–quark pair production. We then analyse new theoretical predictions for the dimuon process and place this in context of the recent ATLAS 7 TeV Drell–Yan data. While these datasets can be broadly fit together, they have different pulls on the strange PDF. We investigate the impact of these new calculations on this possible tension. Next, we look to the future of PDFs at the High–Luminosity phase of the LHC and the proposed Large Hadron–electron Collider. We generate projections of what the PDF uncertainty may look like at such detectors, demonstrating that they will place significant constraints on the PDFs. Finally we analyse WW production in the context of the newly formulated structure function approach to photon–initiated processes. This method has the potential to significantly reduce the uncertainty when compared to the standard photon PDF approach. We focus on the key concept of gauge dependency, which is highly relevant for this new formalism.

Statement of Authenticity

This thesis contains published work [1–3] that has been produced collaboratively as well as a project that has yet to be published. In the vast majority of cases, the results and plots that will appear in the thesis are directly produced by myself. This is not the case for the results in sections 5.3.5, 5.4 and 5.5. For section 5.3.5, I present a natural extension of my results generated by a collaborator [1] that is necessary for the context of my work and is thus briefly mentioned. For sections 5.4 and 5.5, I played a significant role in generating initial tests, data handling, and engaged in discussions to guide the study. Thus, these results will be included within the thesis. Finally Fig. 5.8 contains my results, although I did not specifically generate the plot.

Acknowledgements

First and foremost I would like to thank my supervisor Lucian Harland–Lang for his guidance and support throughout my degree. In particular, for the numerous fruitful discussions and his encouragements to develop my own ideas, which has enabled my research skills to grow over the past four years. I would also like to thank Claire Gwenlan for her continued support during this time. Over the past eight years I have had the pleasure to call Wadham College my home, and I am particularly grateful for the tutors there, whose excellent teaching has enabled me to be where I am now. On a similar note, I would also like to thank the Physics department and those within who have aided me during my time at Oxford. I am lucky to have the love and care of my family and Daniel, who have provided the much-needed support throughout my time as a DPhil student. Finally, I would like to thank the STFC for their funding through the studentship ST/R505006/1.

Contents

1 Introduction **1**

2 PDFs: Theory and Practice **4**

2.1 Theoretical Background 4

2.1.1 Quantum Chromodynamics 4

2.1.2 Parton Distribution Functions 7

2.2 Experimental Data 15

2.2.1 DIS data 16

2.2.2 LHC Precision Data 17

2.3 Methods for PDFs 17

2.3.1 Data and Theory Considerations 19

2.3.2 Error Determination 20

2.4 PDFs in practise 22

2.4.1 MSHT20 23

2.4.2 Other PDF fitting groups 24

2.4.3 Current Challenges and Advancements in PDF fitting 27

2.4.4 Future of PDFs 30

3 Differential $t\bar{t}$ production **32**

3.1 Introduction . . . 32
3.2 Data Sets and Theoretical Calculation . . . 36
3.3 Lepton + Jets Channel at ATLAS . . . 37
3.3.1 Fitting within the MMHT framework . . . 37
3.3.2 Understanding the poor fit quality: the role of correlated systematic errors . . . 39
3.3.3 Theoretical precision . . . 49
3.4 Dilepton Channel at ATLAS . . . 51
3.5 Lepton + jets Channel at CMS . . . 53
3.6 Combining Datasets . . . 55
3.7 Implications for the MSHT20 Fit . . . 58
3.8 Summary and Outlook . . . 58

4 Updates to the Strange PDF from Dimuon Data 62
4.1 Introduction . . . 62
4.2 Theoretical Background . . . 63
4.3 Impact of full NNLO calculation on MMHT . . . 73
4.4 Summary . . . 79

5 PDFs at the HL–LHC and LHeC 81
5.1 Introduction . . . 81
5.2 Fitting Methodology . . . 83
5.3 HL–LHC . . . 87
5.3.1 Data Selection . . . 87
5.3.2 Theory calculations and pseudo–data generation . . . 91
5.3.3 Fits to individual datasets . . . 95
5.3.4 Ultimate PDFs at the HL–LHC . . . 100

5.3.5 Implications for HL–LHC phenomenology 102
5.4 LHeC 104
5.4.1 Pseudo–datasets 104
5.4.2 Impact of the LHeC pseudo–dataset 107
5.5 HL–LHC and LHeC Combined Pseudo–dataset 109
5.6 Summary 111

6 WW Production via Photon Interactions 113
6.1 Introduction 113
6.2 The Structure Function Approach 115
6.3 The Unitary Gauge 117
6.4 The Axial Gauge 122
6.5 WW Production at the LHC 127
6.6 Summary 130

7 Conclusion 131

Chapter 1

Introduction

Since the confirmation of the point–like nature of the proton's constituent particles from electron–proton collisions at SLAC [4], the study of these fundamental particles, commonly referred to as partons, has been a long–standing and extensive field of research. With Quantum Chromodynamics (QCD) forming our best understanding of these particles and their interactions, these partons were identified as the quarks and gluons of the Standard Model and numerous distinct and important avenues of research have opened up to help understand these particles and the underlying physics. The behaviour of the partons as they interact at high–energy remains an important field of study and underpins our knowledge of the composite nature of the proton. Such behaviour is parametrised by the Parton Distribution Functions (PDFs), which are not currently calculable from first principles due to the low scales involved within the proton. While there is progress with lattice determinations of such quantities [5], a data–oriented approach is typically used to compute these functions, with one of the first global NLO PDF fits involving a substantial range of data performed in [6]. Since then, numerous additional experiments have collected data to constrain the PDFs, with the bulk of this coming from the HERA experiments [7–11]. Following the start–up of the Large Hadron Collider (LHC) in 2009, an enormous wealth of data has been collected and analysed, providing constraints on

the PDFs that now begin to rival the HERA data [12]. The importance of this constraining power cannot be understated, with these data providing information on regions of phase space and PDF combinations that could not be analysed by previous Deep Inelastic Scattering (DIS) experiments. Thus, fitting the PDFs to the LHC data provides an impressive handle on these complex objects.

As the precision of data from colliders increased over time, it was clear that the uncertainty on the PDFs and its propagation to observables became important to quantify and understand. This is now key in modern data, with this being one of the dominant theoretical uncertainties for some Standard Model (SM) parameters, Beyond the Standard Model (BSM) processes and background determinations, particularly those that involve kinematical regions of the PDFs that currently lack constraints. Thus, the reduction of PDF uncertainties, and the analysis of processes that constrain new regions of phase space has become a significant area of research in recent years. With the current uncertainties decreasing as we progress through the LHC precision era, new and interesting challenges in how to deal with such precise data face the PDF community, with constant progress being made.

In this thesis we continue this long–standing avenue of research in the precision era at the LHC. We begin by briefly describing the theoretical foundation of PDFs and methodology employed in their determination in chapter 2. Here we also describe the current state of PDF fits, the challenges currently faced, and the expected future of PDFs. In chapter 3, we analyse the impact of differential $t\bar{t}$ data from the LHC [13–15] in the context of the MMHT14 PDFs [16]. Here we describe complications that can arise from highly correlated data and a potential method to remedy this. In chapter 4 we describe updates to the theoretical calculation [17] of neutrino–ion interactions at CCFR/NuTeV [18, 19] and its impact on the MMHT14 PDFs. Both of these studies have been implemented in the new MSHT20 PDF set [12].

Then, in chapter 5, we present projected PDF uncertainties at the upcoming High Luminosity phase of the LHC (HL–LHC) and the proposed Large Hadron–electron collider (LHeC). We show that both detectors can provide substantial constraints on the PDFs and are complementary to each other. In chapter 6, we describe new methods that can be used to generate precise predictions of photon–initiated processes in proton–proton collisions [20]. We perform these calculations in the context of WW pair production, paying particular attention to gauge dependent effects. Finally, in chapter 7, we conclude.

Chapter 2

Parton Distribution Functions: Theory and Practice

In this chapter we discuss the theoretical framework upon which the Parton Distribution Functions are defined. The discussion given here is based upon standard textbook results found in, for example, [21, 22]. We also describe modern methodology used to determine the PDFs, their current status, and their projected future.

2.1 Theoretical Background

2.1.1 Quantum Chromodynamics

We begin our discussion of PDFs by examining the current best theory of the quarks and gluons: Quantum Chromodynamics (QCD). The QCD Lagrangian is given by:

$$\mathcal{L} = -\frac{1}{4}F^a_{\mu\nu}F^{a\mu\nu} + \sum_i \bar{\psi}_i(i\not{D} - m)\psi_i + \mathcal{L}_{\text{gf}} + \mathcal{L}_{\text{ghost}}, \tag{2.1}$$

where ψ_i are the quark fields, and $F^a_{\mu\nu}$ are the field strength tensors given by:

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu + g f^{abc} A^b_\mu A^c_\nu, \tag{2.2}$$

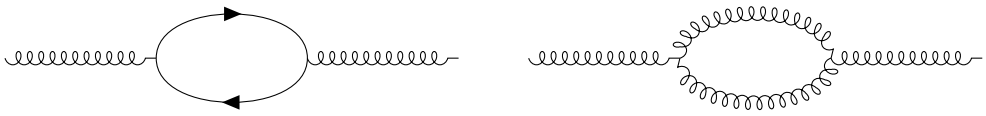


Figure 2.1: NLO corrections to the gluon propagator due to a fermion loop (left) and a gluon loop (right).

where A^a_μ are the gluon fields, g is the coupling constant and f^{abc} are the structure constants. D_μ are the covariant derivatives, defined as:

$$D_\mu = \partial_\mu - igA_\mu. \tag{2.3}$$

$\mathcal{L}_{\text{gf}}$ is the gauge fixing Lagrangian which can be chosen to simplify calculations, while physical observables are independent of this choice. In chapter 6 we will perform a calculation for which the choice of an equivalent term in the electroweak Lagrangian plays an important role. $\mathcal{L}_{\text{ghost}}$ is the ghost Lagrangian describing the interactions of the unphysical ghost particles. This term is important for higher order calculations which are not performed here, thus we will not discuss this term further.

The QCD Lagrangian is invariant under $SU(3)$ gauge transformations and the non–abelian nature of this group results in more complicated interactions when compared to a theory satisfying an abelian gauge symmetry, such as QED. One of the most important differences here is that the gluons themselves carry colour charge, resulting in a self–interaction term. This has many interesting effects on the theory, such as their impact on the running of the coupling, which will now discuss.

The NLO Feynman diagrams for the gluon propagator are shown in Fig. 2.1. It can be shown that these diagrams, along with vertex and higher order corrections, can be absorbed into the coupling constant, resulting in an 'effective' coupling that

varies with the scale of the system. For NLO this is given by:

$$\alpha(Q^2) = \frac{\alpha(\mu^2)}{1 + b_0 \ln \frac{Q^2}{\mu^2}}, \tag{2.4}$$

where Q^2 is the scale at which we wish to measure the coupling at, μ^2 is some reference scale and b_0 is a constant that depends on the detailed calculation of the NLO diagrams. For QED we have $b_0 < 0$, resulting purely from charged fermionic loop corrections, similar to Fig. 2.1 (left), and gives a coupling that becomes weaker at lower scales. This is often interpreted as the virtual charged pairs providing a shielding effect between the two real charges at larger distances (lower scales). For QCD we have both the fermion and gluon loop corrections and, due to the bosonic nature of the gluon, these result in contributions of differing signs. Further, we must also include NLO corrections to the gluon–fermion vertex as these contribute to the effective coupling in QCD, while the equivalent diagrams in QED do not. Overall, it can be shown that:

$$b_0 = \frac{33 - 2n_f}{12\pi}, \tag{2.5}$$

where n_f is the number of active quarks, typically taken to be 3–6 depending on the energy scale. It should be noted that, for all of these values of n_f, we have $b_0 > 0$, resulting in a coupling that decreases at higher energies, in complete contrast to the QED case. This results in two distinct regions of QCD that display completely different physical behaviour. At higher scales, the coupling is sufficiently weak that the quarks and gluons can be considered as free, and their interactions accurately described by the corresponding Feynman diagrams. In particular, the matrix element

can be written as an expansion in the strong coupling constant:

$$\mathcal{M} = \sum_{k=0}^{\infty} g^k(\mu^2)\mathcal{M}^{(k)}, \tag{2.6}$$

where $g = \sqrt{4\pi\alpha_S}$ is the coupling constant and $\mathcal{M}^{(k)}$ is the matrix element calculated using Feynman diagrams of order k. At low scales the coupling is stronger, so much so that this perturbative expansion breaks down and can no longer be used. Since the proton's constituent partons interact with each other at these scales, a different formalism is required to describe them. As both of these regimes are present in high energy collisions of hadrons, careful disentanglement of the different physical effects must be performed to enable an accurate calculation of observables.

2.1.2 Parton Distribution Functions

Due to the presence of low energy physics within the protons as they collide, an accurate determination of the impact of the non–perturbative effects present is required. Since we lack a perturbative expansion in α_S, a data–driven approach is usually taken. We therefore must establish a theoretical framework in which we can extract and parametrise the low energy behaviour. With this in mind, we analyse the Deep Inelastic Scattering (DIS) process, where a lepton interacts with a hadron at high energy. A typical diagram for such a process is shown in Fig. 2.2 (left), where we have obscured the interaction between the proton and the boson as this is not currently calculable. Instead, we parametrise this cross section as:

$$\sigma \sim L_{\mu\nu}W^{\mu\nu}, \tag{2.7}$$

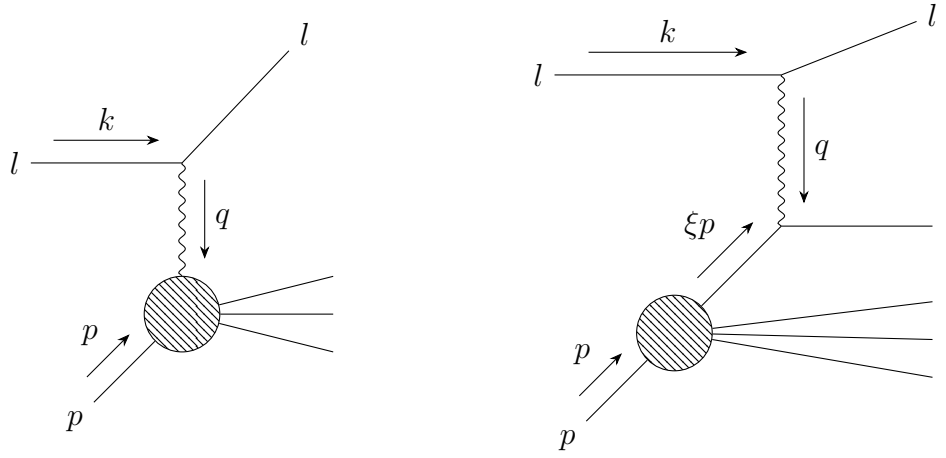


Figure 2.2: Electron–proton collisions. (left): standard Feynman diagram in terms of structure functions. (right): diagram split into low– and high–energy regimes using the factorisation theorem.

where $L^{\mu\nu}$ represents the (calculable) interaction between the lepton and the virtual boson, while $W^{\mu\nu}$ represents the same for the proton. By analysing the Lorentz structure of $W^{\mu\nu}$ and taking into account charge conservation, we can write this term as:

$$\begin{aligned} W^{\mu\nu} = &- \left(g^{\mu\nu} + \frac{q^\mu q^\nu}{Q^2} \right) F_1(x, Q^2) + \frac{(2xp^\mu - q^\mu)(2xp^\nu - q^\nu)}{Q^2} F_2(x, Q^2) \\ &- i\epsilon^{\mu\nu\alpha\beta} p_\alpha q_\beta \frac{x}{Q^2} F_3(x, Q^2), \end{aligned} \tag{2.8}$$

where p^μ and q^μ are the momenta of the incoming proton and virtual boson respectively, $Q^2 = -q^2$, and $x = Q^2/(2p.q)$ is known as the Bjorken-x variable. $F_i(x, Q^2)$ are known as the structure functions and are measurable in experiments. Note that F_3 is only present for parity violating interactions. Further, it is common to define $F_L = F_2 - 2xF_1$ to extract the longitudinal component of the interaction. Other useful kinematical variables that will be used later are the invariant mass of the proton–photon system, $W = (p + q)^2$, and the inelasticity, $y = p.q/p.k$, where k is the momentum of the incoming lepton. The original measurements of these structure

functions showed a perhaps surprising fact: they are approximately independent of Q^2 in the limit $Q^2, p.q \to \infty$ [23, 24]. This is known as Bjorken scaling and suggests that there cannot be any additional scales in the interaction; therefore, the photon must be interacting with point–like massless particles. This evidence formed the backbone of the Quark–Parton Model (QPM), where the photon interacts directly with a constituent quark of the proton. Such an interaction only makes sense in the infinite momentum frame, where the proton is moving at such a speed that its mass can be ignored. Here, the interactions within the proton are slowed down by a factor of $1/\sqrt{Q^2}$ due to time dilation and can be ignored. Further, due to the high proton momentum in this frame, all momenta transverse to this can be neglected and we can write the momenta of the quark as $p_q^\mu = \xi p^\mu$. The low scale contribution is parametrised by a Parton Distribution Function (PDF), $q_i(\xi)$, representing the probability that a particular quark carries a fraction of the proton's momentum, ξ, and participates in the hard scattering. Such an interaction is shown in Fig.2.2 (right), and a calculation of this results in:

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}x\mathrm{d}Q^2} = \frac{2\pi\alpha^2}{xQ^4}[1 + (1-y)^2] \sum_i e_i^2 x q_i(x). \tag{2.9}$$

Here we have summed over all possible quarks, each with charge e_i. This can be directly compared to Eq. (2.8) to give:

$$F_2(x) = \sum_i e_i^2 x q_i(x), \qquad F_L(x) = 0, \tag{2.10}$$

where the latter result is due to the fermionic nature of the quarks. Experimental data [7] demonstrates that F_L is, in general, non-zero and F_2 does have some dependence on Q^2. It is therefore clear that Bjorken scaling breaks down (typically

referred to as 'scaling violation') at low Q^2 and thus the Quark–Parton Model cannot capture the entire picture. It is here that higher order QCD corrections have significant impact on the interaction and cannot be ignored. To calculate the contribution from such corrections, one requires a new framework to separate the high– and low–energy scales. This can be done using the factorization theorem [25], which for DIS can be written as:

$$F_i(x, Q^2) = \sum_j \int_x^1 \frac{\mathrm{d}\xi}{\xi} f_j \left(\frac{x}{\xi}\right) C_{ij}(\xi, Q^2) + \mathcal{O}\left(\frac{\Lambda^2_{\mathrm{QCD}}}{Q^2}\right), \tag{2.11}$$

where C_{ij} is the contribution from the high energy hard–scattering cross section between the parton and the lepton on the structure function F_i with the j^{th} parton in the initial state, and $f_j(x, Q^2)$ are the PDFs (which can now include the gluon and other particles). Such a formula can be derived through the operator product expansion [22]. The corrections to this are referred to as 'higher twist' and represents the contributions from diagrams involving additional partons, which are suppressed by Λ_{QCD}, typically taken to be of the order of 100 MeV. The LO contribution is:

$$C^{(0)}_{2q_j}(\xi, Q^2) = e_j^2 x \delta(1 - \xi), \tag{2.12}$$

where the number in brackets refers to the order. This gives the same result as the QPM.

With this new framework, we are now able to calculate NLO correction to this DIS process. These include the emission of a real gluon as shown in Fig.2.3 (left and centre) and it can be shown that these diagrams give a contribution to F_2:

$$C^{(1)}_{2q_j}(\xi, Q^2) = \frac{e_j^2 x \alpha_S}{2\pi} \left(P_{qq}(\xi) \ln \frac{Q^2}{\kappa^2} + K_q(\xi) \right). \tag{2.13}$$

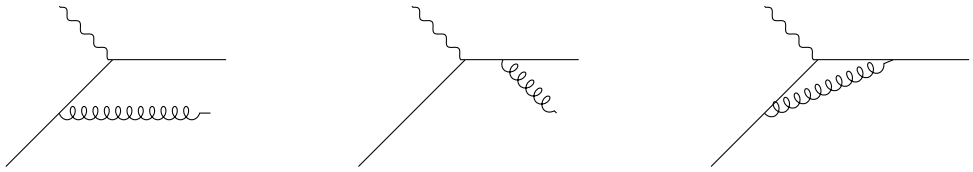


Figure 2.3: Real (left and centre) and a representative virtual (right) corrections to the fermion–gluon interaction.

Here we can already see that the coefficient, and hence the structure functions, depend on the scale, Q^2, and thus break Bjorken scaling. The function $P_{qq}(\xi)$ is the $q \to qg$ splitting function defined as:

$$P_{qq}(\xi) = C_F \frac{1+\xi^2}{1-\xi}, \tag{2.14}$$

where $C_F = 4/3$ is the Casimir for the fundamental representation of $SU(3)$ and $K_q(\xi)$ is a (sub–leading) non–logarithmic term. This NLO contribution contains two divergences: first, $P_{qq}(\xi)$ diverges when the gluon becomes very soft, i.e. in the $\xi \to 1$ limit. Secondly, the derivation of the above result involves the integral over the gluon's transverse momentum relative to the initial parton: $\int \mathrm{d}p_t^2/p_t^2$. This integral diverges in the collinear limit and has been regulated by a low momentum cut–off κ; however, including such a regulator in the final result is rather unsatisfactory.

The soft gluon divergence can be corrected by simply including the virtual corrections to the quark propagator and the photon vertex (see Fig. 2.3 (right) for a representative diagram). Such contributions alter the form of the splitting function:

$$P_{qq}(\xi) = C_F \left(\frac{1+\xi^2}{(1-\xi)_+} + \frac{3}{2}\delta(1-\xi) \right), \tag{2.15}$$

where we have used the 'plus prescription', defined for any sufficiently smooth function $f(\xi)$ to be:

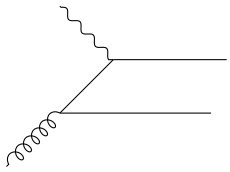


Figure 2.4: Gluon initiated DIS at NLO.

$$\int_0^1 \mathrm{d}\xi \frac{f(\xi)}{(1-\xi)_+} = \int_0^1 \mathrm{d}\xi \frac{f(\xi)-f(1)}{1-\xi}. \tag{2.16}$$

This form of the splitting function is not divergent in the soft gluon limit. Before tackling the collinear divergence, we first look at the gluon–initiated process, as shown in Fig. 2.4. This process has coefficient function:

$$C_{2g}^{(1)} = x\left(\sum_i e_i^2\right)\frac{\alpha_S}{2\pi}\left(P_{qg}(\xi)\ln\frac{Q^2}{\kappa^2} + K_g(\xi)\right), \tag{2.17}$$

where the $g \to qq$ splitting function is given by:

$$P_{qg}(\xi) = \frac{1}{2}\left[\xi + (1-\xi)^2\right]. \tag{2.18}$$

We can see that, while this process does not contain a soft divergence, it still requires the regulator κ to remove the collinear divergence. It is best to deal with both the gluon and quark contributions simultaneously to remove this singularity. To do this, one must realise that the PDF, as defined in Eq. (2.11), is not directly measurable. We can thus proceed by defining this quantity as the bare PDF, $f_j^0(x)$ and define a renormalized PDF. We begin by looking at the LO and NLO contributions to F_2 in terms of the bare PDF:

$$\begin{aligned} F_2(x,Q^2) = x\sum_j e_i^2 \Bigg\{ & q_j^0(x) + \frac{\alpha_S}{2\pi}\int_x^1 \frac{\mathrm{d}\xi}{\xi} q_j^0(\xi)\left[P_{qq}\left(\frac{x}{\xi}\right)\ln\frac{Q^2}{\kappa^2} + K_q\left(\frac{x}{\xi}\right)\right] \\ & + \frac{\alpha_S}{2\pi}\int_x^1 \frac{\mathrm{d}\xi}{\xi} g^0(\xi)\left[P_{qg}\left(\frac{x}{\xi}\right)\ln\frac{Q^2}{\kappa^2} + K_g\left(\frac{x}{\xi}\right)\right]\Bigg\}. \end{aligned} \tag{2.19}$$

By carefully defining the renormalized PDF, the collinear divergences can be absorbed into a factorisation scale μ_F:

$$\begin{aligned} q_j(x,\mu_F^2) = q_j^0(x) &+ \frac{\alpha_S}{2\pi}\int_x^1 \frac{\mathrm{d}\xi}{\xi} q_j^0(\xi)\left[P_{qq}\left(\frac{x}{\xi}\right)\ln\frac{\mu_F^2}{\kappa^2} + K_q\left(\frac{x}{\xi}\right)\right] \\ &+ \frac{\alpha_S}{2\pi}\int_x^1 \frac{\mathrm{d}\xi}{\xi} g^0(\xi)\left[P_{qg}\left(\frac{x}{\xi}\right)\ln\frac{\mu_F^2}{\kappa^2} + K_g\left(\frac{x}{\xi}\right)\right], \end{aligned} \tag{2.20}$$

and similarly for the gluon. Using this, we can define a renormalised coefficient function C_{ij}:

$$\begin{aligned} C_{2q}^{(0,1)}(\xi,\mu_F^2) &= xe_q^2\left[\delta(1-\xi) + \frac{\alpha_S}{2\pi}P_{qq}(\xi)\ln\frac{Q^2}{\mu_F^2}\right], \\ C_{2g}^{(0,1)}(\xi,\mu_F^2) &= x\left(\sum_i e_i^2\right)\frac{\alpha_S}{2\pi}P_{qg}(\xi)\ln\frac{Q^2}{\mu_F^2}, \end{aligned} \tag{2.21}$$

such that substituting the renormalized quantities into Eq. (2.11) will result in the same result as using the bare equivalents. We note that this particular definition of the renormalized PDF has removed the finite functions, $K_{q,g}(\xi)$, completely from the result. This can be done at all orders in what is known as the DIS factorisation scheme, such that the result $F_2(x,Q^2) = x\sum_j e_j^2 q_j(x,Q^2)$ is exact to all orders in perturbation theory. Thus, since the structure functions are measurable experimentally, it is clear that the renormalized PDFs are also measurable. However, this is not the only scheme one can use, with the $\overline{\mathrm{MS}}$ scheme also being common. Here, the coefficient functions at LO and NLO are given by:

$$C_{2q_j,\overline{\mathrm{MS}}}^{(0,1)}(\xi,\mu_F^2) = xe_j^2\left[\delta(1-\xi) + \frac{\alpha_S}{2\pi}P_{qq}(\xi)\ln\frac{Q^2}{\mu_F^2} + K_q^{\overline{\mathrm{MS}}}(\xi)\right]. \tag{2.22}$$

The dependence of the PDFs on the scale μ_F can be extracted by differentiating

Eq. (2.20):

$$\frac{\partial q_i(x,\mu_F^2)}{\partial \ln \mu_F} = \frac{\alpha_S}{2\pi}\int_x^1 \frac{\mathrm{d}\xi}{\xi}\left[q_i(\xi,\mu_F^2)P_{qq}\left(\frac{x}{\xi}\right) + g(\xi,\mu_F^2)P_{qg}\left(\frac{x}{\xi}\right)\right]. \tag{2.23}$$

This, along with a similar formula for the gluon, form the DGLAP evolution equations [26], and are analogous to the running of the coupling in Eq. (2.4). While the dependence on Q^2 can be explicitly determined through DGLAP, the x–dependence cannot currently be calculated from first principles.

The above analysis can be performed with hadron–hadron collisions, where the cross section is given by:

$$\sigma = \sum_{ab}\int \frac{\mathrm{d}z_1}{z_1}\int \frac{\mathrm{d}z_2}{z_2} f_a(z_1,\mu_F^2) f_b(z_2,\mu_F^2)\hat{\sigma}_{ab}(z_1,z_2,\mu_F^2,\mu_R^2) + \mathcal{O}\left(\frac{\Lambda_{\mathrm{QCD}}^2}{Q^2}\right), \tag{2.24}$$

where $\hat{\sigma}_{ab}$ are the hard scattering cross sections, calculable in perturbation theory. It should be noted that this formula is strictly only proven for Drell–Yan processes, where it can be shown that the interactions between the two protons outside of the hard scattering are sub–leading [21]. For processes involving coloured particles in the final state, interactions between these and the initial protons has not yet been proven to be sub–leading, but it is generally assumed to be, as will be done here.

It can be difficult to disentangle the effects of the two PDFs and it is typical to define a luminosity that combines the two:

$$\mathcal{L}_{ij}(\hat{s},\mu_F^2) = \frac{1}{s}\int_{\hat{s}/s}^1 \frac{\mathrm{d}\xi}{\xi} f_i\left(\frac{\hat{s}}{s\xi},\mu_F^2\right) f_j(\xi,\mu_F^2), \tag{2.25}$$

where $\sqrt{\hat{s}}$ and $\sqrt{s}$ are the centre of mass energies for the parton–parton and proton–proton systems respectively. With this definition, we can calculate the cross section

to be:

$$\sigma = \sum_{ij} \int_{M^2}^{s} \mathrm{d}\hat{s} \mathcal{L}_{ij}(\hat{s}, \mu_F^2) \hat{\sigma}_{ij}(\hat{s}, M^2, \mu_R^2, \mu_F^2) + \mathcal{O}\left(\frac{\Lambda_{\mathrm{QCD}}^2}{Q^2}\right). \tag{2.26}$$

The luminosity enables one to quickly identify the impact certain PDF combinations have on the overall cross section.

2.2 Experimental Data

Given the lack of a perturbative expansion in α_S, the determination of PDFs is a largely data–driven endeavour; thus the choice and treatment of the experimental data forms a significant part of PDF fitting. We wish to pick data that have a strong dependence on the PDFs as well as a wide range of processes that explore the different parton distributions over a diverse range of x. However, care must be taken when selecting data, for example, we do not want to select data at particularly low Q^2, as these may be susceptible to higher twist corrections as seen in Eq. (2.11). A cut, usually around $Q^2 \gtrsim 2$ GeV^2 and $W^2 = (p+q)^2 \gtrsim 15$ GeV^2, is typically placed on the data to avoid this. Further, we would like to avoid any datasets that have potential new physics hidden within them. If we were to fit such a dataset, the new physics would be absorbed into the parametrised PDF and any future analysis would be unlikely to detect any signals associated with these theories [27–29]. Finally, care must be taken to analyse any possible tensions that exist between datasets that are included within any particular fit (see [12, 30, 31] for examples). This is particularly prevalent if data from multiple experiments are used, where no perfect statistical description can be formulated.

2.2.1 DIS data

We first describe Deep Inelastic Scattering (DIS) data, for which the theory was detailed in section 2.1.2. Data from DIS have been available since before QCD was fully formulated at experiments such as SLAC [4], and still to this day forms a significant backbone of many PDF fits [17], of which the main contribution comes from the combined HERA I + II [7–11] dataset. These data are presented in terms of structure functions and thus one can easily determine the PDF combination most constrained by examining the LO result such as in Eq. (2.10). Therefore, it is useful to use a variety of different structure functions to better constrain the PDFs. For example, one could analyse the neutrino–induced charged current interactions to gain information on the flavour separation as seen in the LO contribution:

$$F_2^{\bar{\nu}p} - F_2^{\nu p} = 2x\left[(u(x) - \bar{u}(x)) - (d(x) - \bar{d}(x)) - (s(x) - \bar{s}(x))\right]. \tag{2.27}$$

This is complementary to F_2 in the neutral current, which gives information on the charge weighted sum as seen in Eq. (2.10). Further, higher order corrections can also be analysed and their impact on F_L, as well as violations of Bjorken scaling, provide invaluable information on the gluon PDF. This is particularly useful in the low x region, where the variation with respect to Q^2 is much larger. In addition, the final state particles can also be identified to enable determinations of new PDF combinations through exclusive processes. For example, neutrino induced DIS at CCFR/NuTeV [18, 19] can induce two muons in the final state, one of which is the result of a charm decay. Due to the large CKM matrix element between the strange and charm quarks, this provides invaluable constraints on the strange PDF. This PDF is not well constrained by the HERA data, as the available combinations have

little sensitivity to this PDF. Recent theoretical advancements to this process are discussed in chapter 4.

2.2.2 LHC Precision Data

Over the past decade, the LHC has provided a wealth of precise data that can constrain the PDFs. The nature of the proton–proton collisions allows for a wider range of processes to be measured compared to what is available in lepton–proton collisions and enables kinematical regions to be constrained that data from DIS have little handle on. This includes the high x gluon, which can be determined through, for example, jet or top quark pair production, and the sea quarks at high x, measured through Drell–Yan processes at high invariant mass. However, the addition of a second proton into the collision can result in significantly more QCD effects in the problem, sometimes producing less clean signals compared to that in DIS experiments. This can result in larger uncertainties from, for example, Monte Carlo simulations that attempt to handle these effects. We discuss an example of this in chapter 3 in the context of top quark pair production.

2.3 Methods for PDFs

We now move on to the more practical aspects of PDF fitting. To begin one selects a particular scale, Q_0^2, and parametrises the PDFs, $f_i(x, Q_0^2)$, with respect to the variable x. The exact choice of parametrisation is up to those performing the fit, generally chosen to accommodate the types of data that will be included. It is common to include a factor of $x^\alpha(1-x)^\beta$, where α and β are to be fit. These represent the $x \to 0, 1$ limits described by Regge theory [22] and constituent counting

rules [32]. Overall, a PDF parametrization will typically take the form of:

$$f_i(x, Q_0^2) = A_i x^{\alpha_i}(1-x)^{\beta_i}\left(1 + \sum_k^{n_{\text{par}}} a_{i,k} C_k(x)\right), \tag{2.28}$$

where i is the flavour index, A_i, α_i, β_i and $a_{i,k}$ are parameters to be fit, and C_k are the first n_{par} functions of some expansion to be chosen. Note that writing all PDFs in this way will result in some redundant parameters due to the QCD sum rules, which need to be taken into account. Once the PDF has been parametrised, it can be evolved with the DGLAP equations to obtain the PDFs at any scale. The next ingredient to be added is the hard scattering cross section, represented by C_{ij} in Eq. (2.11) and $\hat{\sigma}_{ab}$ in Eq. (2.24). As this theoretical prediction will be required numerous times during the course of the fit, they are often tabulated for fast access, with `APPLGRID` [33] and `fastNLO` [34, 35] being common software to do this. With the hard regime calculated and the soft scale parametrised, we can now perform the convolutions in Eqs. (2.11) and (2.24). This is then compared to an experimental result while varying the parameters in the PDF, such that a minimisation with respect to some statistical measure can be performed. For this, it is typical to take the χ^2 measure:

$$\chi^2 = \sum_i^{N_{\text{dat}}} \frac{1}{s_i^2}\left(D_i - T_i + \sum_\alpha^{N_{\text{sys}}} \beta_{i\alpha}\lambda_\alpha\right)^2 + \sum_\alpha^{N_{\text{sys}}} \lambda_\alpha^2, \tag{2.29}$$

where D_i are the data points, T_i are the theoretical predictions, which include the parametrised PDFs, s_i are the uncorrelated errors, $\beta_{i\alpha}$ are the correlated uncertainties for the error labelled α, and N_{dat} and N_{sys} are the number of data points and correlated errors respectively. Finally, λ_α are the nuisance parameters that we will minimise with respect to, along with the PDF parameters. Once the fitted PDF is

obtained, these are then often tabulated and presented in the `LHAPDF` [36] format, such that they can be used in the prediction of new observables.

2.3.1 Data and Theory Considerations

When performing a fit, numerous details need to be carefully considered and choices must be made to define how the fit is performed. It is these choices that lead to a numerous PDF fitting groups that all produce slightly different answers. We now discuss some of the more important general differences as well as introduce some of these fitting groups.

One of the more important choices is which datasets should be included into the fit. This splits most PDF fits into two groups: the so–called 'global' fits, where a wide range of data from many experiments are included, and those fits which focus on one or two experiments. The most notable global fitting collaborations are MSHT [12], NNPDF [37] and CTEQ [30], while ATLAS-epWZtop18 [38] and HERAPDF2.0 [39] represent PDFs that are focused on a smaller dataset. The choice of data also strongly impacts the parametrisation, with global PDFs opting for a more general parametrisation to incorporate many datasets, while the restricted PDFs will typically choose parameters that will be largely impacted by the chosen datasets, such as HERAPDF2.0 selecting parameters that strongly depend on the DIS structure functions. The datasets chosen to be included into a global fit can also vary between collaborations, although typically similar datasets are chosen due to the shared desire to include data that constrain the PDFs.

PDF fits can also vary by certain theoretical choices, the most notable of which is the value of the strong coupling $\alpha_S(Q^2)$. Most PDF groups choose the value of $\alpha_S(M_Z) = 0.118$, agreeing with that given by the Particle Data Group (PDG) [40];

however, one notable exception is ABM [41], who find a value of $\alpha_S(M_Z) = 0.1147$ fits the data best at NNLO. It should be noted that simultaneous fits to α_S and the PDF parameters can be performed, with results typically agreeing with the PDG value [42, 43].

Another theoretical consideration is the treatment of heavy quarks and how to include their mass into a factorizable framework. This is typically done using either an evolved quark PDF, where the heavy quark is generated by DGLAP, or an intrinsic quark [44], where a new parametrised heavy quark PDF is introduced. The evolved quark methodology is explained in more detail in chapter 4, where it is applied to dimuon production.

Finally, there are numerous theoretical considerations that are specific to particular datasets, such as nuclear corrections and branching ratios that need to be taken into account.

2.3.2 Error Determination

One important aspect of PDF fits is propagating the experimental error originating from the data used in the fit into the PDFs themselves. These errors need to be presented in an acceptable format such that they can be propagated again into new observables. The two most common ways of doing this are the Hessian and Monte Carlo approaches. We begin by describing the former, which is used by PDFs that utilize a more traditional parametrisation, such as MSHT, CTEQ and ATLAS–epWZtop18. Here, a Hessian matrix is calculated, defined as:

$$H_{ij} = \frac{\partial^2 \chi^2}{\partial p_i \partial p_j} \tag{2.30}$$

where p_i are the parameters used in the fit. H_{ij} can then be diagonalized to obtain its eigenvectors, p'_i such that:

$$\chi^2 = \chi^2_{\text{min}} + \sum_{ij}^{n_{\text{eig}}} p'_i H_{ij} p'_j, \tag{2.31}$$

where n_{eig} is the number of eigenvectors. The parameters are then varied away from the minimum along p'_i, until the figure of merit has increased such that $\Delta\chi^2 = T^2$, where T is referred to as the tolerance, and we define an error PDF at this point in parameter space. This procedure is then repeated for each eigenvector of H_{ij} until a set of error PDFs is obtained. With these, one can estimate the error when predicting of a new observable, $\mathcal{O}$, using the equation [45]:

$$\sigma_{\mathcal{O}} = \frac{1}{2}\left(\sum_{i}^{n_{\text{eig}}} [\mathcal{O}(S_i^+) - \mathcal{O}(S_i^-)]^2\right)^{1/2}, \tag{2.32}$$

where S_i^+ are the error PDFs along the positive eigenvector direction while S_i^- are those along the negative direction. While one might expect to choose a tolerance of unity as is done in, for example, ATLAS–epWZtop18, in the case of global fits there is sufficient tension between datasets that this is no longer acceptable. If such a value were taken, then fits to different subsets of the data can be inconsistent with each other. Thus a value of $T > 1$ is often chosen to enable a more accurate description of the uncertainties within the fit while also somewhat allowing for parametrization effects and non–Gaussian uncertainties to be accounted for. For CTEQ and MSHT, T is chosen dynamically such that an optimal value can be determined during the fit and allows for the value to vary between eigenvector directions. These methods typically predict a value of $T \sim 3$. An example of one of these procedures will be described in section 2.4.1.

The other method to represent the uncertainty on a PDF is to use a Monte Carlo representation by utilizing replicas. This is most notably used by NNPDF, who use a neural network to parametrise the PDF and thus does not lend well to the Hessian framework. Instead, the experimental data are varied randomly according to their presented errors to generate a data replica. Machine Learning technology is then implemented to fit the PDF to these to create a PDF replica. This is then repeated numerous times, typically of the order $N_{\text{rep}} \sim 100$, to ensure a statistically significant result is obtained. Predictions are then made by convoluting each replica with the hard cross section and taking an average, with the errors defined by the variance [45]:

$$\Delta\mathcal{O} = \sqrt{\frac{1}{N_{\text{rep}} - 1} \sum_{i=1}^{N_{\text{rep}}} \left(\mathcal{O}(f_i) - \langle\mathcal{O}(f)\rangle\right)^2}, \tag{2.33}$$

where f_i is the i^{th} PDF replica and $\langle\mathcal{O}(f)\rangle$ is the average over all PDF replicas.

2.4 PDFs in practise

In this section we describe current state–of–the–art PDF fits. We begin by discussing some of the commonly used PDFs: MSHT, NNPDF, CTEQ and ATLAS–epWZtop18. We pay particular focus on the PDFs generated by the MSHT collaboration as the studies in chapters 3 and 4 are based upon these. Further, the statistical combination of the first three PDFs, known as PDF4LHC15 will be discussed, which will be used as a baseline in chapter 5. We then move on to discuss some of the main topics currently under investigation, including current issues facing the determination of PDFs and ongoing developments.

2.4.1 MSHT20

The MSHT20 PDF set was released towards the end of 2020 as the successor to the MMHT14 PDF set [46] used in numerous studies. The analyses in chapters 3 and 4 build upon MMHT14 and their results formed crucial parts of this update.

The MSHT20 parametrization is an expansion in terms of Chebyshev polynomials in $y = 1-2\sqrt{x}$. This choice results in well–behaved parameters that are typically around or smaller than unity and is easily extended, with additional terms enabling higher frequency details to be determined. The errors of the MSHT PDF set are generated using the Hessian methodology, with a total of 32 eigenvectors. We note that this is smaller than the total number of free parameters, 52, due to redundancy found between some of resulting eigenvector directions. MSHT20 determines the tolerance for each eigenvector direction dynamically. This is done by varying the parameters along each eigenvector direction, keeping track of the χ^2 of each individual dataset. Once the χ^2 of any particular dataset leaves the 68% confidence level region, the variation is stopped and the change in the total χ^2 compared to the best fit value is chosen to be the tolerance. This procedure ensures that all datasets used in the fitting procedure can be accurately described within the overall PDF errors.

The MSHT20 PDF set includes many new datasets into the fit, primarily LHC data, including ATLAS Drell–Yan [47], CMS $W+c$ [48], and single differential $t\bar{t}$ data from ATLAS and CMS at 8 TeV [13–15], the latter of which is discussed in detail in chapter 3. This update also extended the parametrisation of the PDFs, going to the sixth order of Chebyshev polynomials in general, up from four in MMHT14. The exceptions to this are the gluon, for which there is a slightly different parametrisation, although the same number of additional terms have been added, and the difference between strange and anti–strange distribution, $s - \bar{s}$, which has a very simple form

due to the poor constraints that still exist on this quantity. Another significant change is to parametrise the ratio of the anti–up and anti–down quarks, $\bar{d}/\bar{u}$, rather than the difference, $\bar{d} - \bar{u}$. The requirement that $\bar{d} - \bar{u} \to 0$ as $x \to 0$ that was present in MMHT14 is relaxed such that now it is only required that $\bar{d}/\bar{u}$ tends to a constant in this limit. However, fits to data reproduce the original constraint to good precision. The change from a difference to a ratio also allows for an improved central fit and a better representation of the error in the low x region.

The overall impact the updated data and formalism has on the PDFs is varied and a full discussion is not given here. In general, some kinematic regions of the PDF experience decrease in uncertainty, resulting from the increase in the amount of data used. However, there are some regions for which the uncertainty increases when moving from MMHT14 to MSHT20 due to the increased parametrisation. This typically happens in the low x region and enables a better representation of these errors and will allow for a more accurate determination of the uncertainty when dealing with observables dependent on these regions of phase space.

In Fig. 2.5 we plot the PDF distributions from the MSHT20 fit. Note that these include the valence distributions $u_v = u - \bar{u}$ and $d_v = d - \bar{d}$.

2.4.2 Other PDF fitting groups

We now briefly discuss some of the other collaborations that generate commonly used PDF sets. We begin with NNPDF, who are the only major PDF group to use a neural network when fitting. The most recent iteration of this PDF set is NNPDF4.0 released in 2021 [49, 50] as the successor to the NNPDF3.1 set [37]. This update includes more data from the LHC, similar to the MSHT fit, although it should be noted that some of the data new to MSHT20 were already included

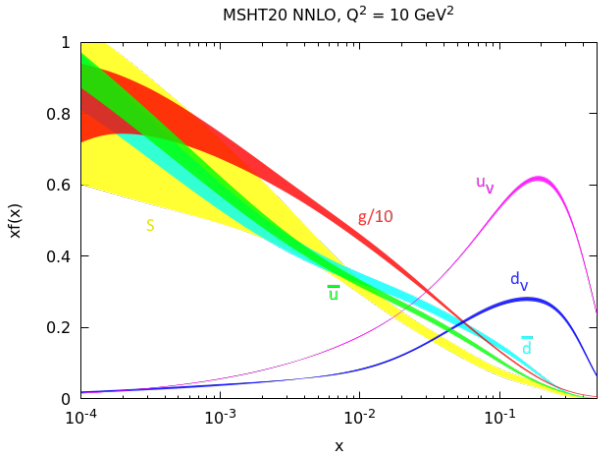


Figure 2.5: The PDF distributions as fit by MSHT20 at $Q^2 = 10$ GeV.

in the NNPDF3.1 release. NNPDF4.0 also brings about significant changes to the methodology, changing the genetic algorithm used for the minimisation to one using gradient decent and utilizing hyper–optimisation [51], enabling parameters related to the structure of the neural network to be varied and fit during run time. This allows for NNPDF fits to be generated faster and optimise the use of the neural network.

We now move onto the CTEQ collaboration with their most up to data PDF set labelled CT18. This superseded the previous CT14 [52] set and includes many new datasets from the LHC and other experiments, similar to those included in the MSHT20 fit. A notable choice made by the collaboration was to generate several distinct PDF sets for different choices of data and methodology. In the base fit, CT18, they exclude the ATLAS 7 TeV WZ data due to tensions with previous datasets and instead include these data in a fit labelled CT18A. In addition to this, they also generate a PDF set, CT18X, where the factorization scale is changed at low x in DIS to mimic resummation effects in this region. Both the ATLAS 7 TeV WZ data and the altered scale are combined into a final fit, CT18Z. The PDFs provided by CTEQ are also in the Hessian formalism and implement dynamical tolerance.

However, the methodology is different to that of MSHT20, using a Tier–2 penalty, which rapidly grows when any experiment is outside of its 90% confidence region [52].

Next, we briefly discuss the ATLAS–epWZtop PDF. This particular set is distinct from those discussed above as it uses a more limited dataset. In particular, it only includes data from the HERA experiment as well as the ATLAS 7 TeV WZ data [53] and the ATLAS 8 TeV differential top quark pair distribution [38]. Due to the reliance on the DIS data at HERA, this PDF set also has a more restrictive parametrization compared to the CTEQ, NNPDF and MSHT fits, with a focus on the particular PDF combinations that are sensitive to these data. In addition, this PDF set also determines the errors in Hessian formalism but chooses a fixed tolerance of $T = 1$ due to the smaller input dataset, resulting results in reduced uncertainties than if a larger tolerance were taken.

Finally we discuss the PDF4LHC15 [54]. This set is not generated by a single collaboration but is instead a statistical combination of the MMHT14, NNPDF3.0 [55] and CT14 PDFs generated in 2015. This set was designed to represent the current best knowledge of PDFs at the time, combining numerous sources of data and distinct methodologies into a single PDF set and was recommended for use in the LHC run II. To generate this, a benchmarking procedure was undertaken to iron out any major differences in the PDFs used in its generation while still allowing for their individuality. Since then, the MSHT, NNPDF and CTEQ collaborations have diverged somewhat in certain regions of phase space and new benchmarking study is under way to analyse the source of these differences and generate a new updated PDF4LHC set [56]. This will include new data from the LHC that are now part of the component sets as well as theoretical advancements that have been made in recent years.

2.4.3 Current Challenges and Advancements in PDF fitting

We now describe some of the current topics generating significant discussion within the PDF community. The relation between these topics and the studies presented in this thesis are also noted.

As the precision of data increases with time, fitting such data becomes more difficult. Sometimes, this difficultly will present itself as tension between two datasets, where they cannot be fit consistently within a single framework. In chapter 4 we describe one such example that has recently drawn attention from the community: the possible tension between the relatively new ATLAS 7 TeV WZ data [47] and the older dimuon data from CCFR/NuTeV [18, 19]. Potential tensions, such as this one, that may exist between datasets require extensive studies to discover, and potentially remedy. More recently, tensions within a single highly correlated and systematic dominated datasets are being discovered, for example in the ATLAS 7 TeV jet data [57] and the ATLAS 8 TeV differential $t\bar{t}$ dataset measured through the lepton + jet channel [13]. Here the correlations between bins within the dataset strongly constrain how the data behave in such a way that a reliable fit cannot be obtained. In chapter 3 we discuss in more detail how such complications can arise and how one could go about carefully decorrelating the data such that a reasonable fit can be obtained. These methods are not completely watertight as any manipulation of the data can result in the introduction of bias; however, detailed analysis of the source of the correlated errors and a deep understanding of the data can lead to a reasonable decorrelation and a better fit.

The increased precision of the data input into the PDFs also results in the need for more precise theory predictions. To this end, not only are NNLO QCD corrections being included into the hard scattering cross section, but so are NLO electroweak

corrections, which can be of a similar order of magnitude. Thus photon [58, 59] and lepton [60] PDFs can also be included to account for these electroweak corrections in the soft scales present in the proton. We focus on the photon PDF, which is more common, with most implementations stemming from the LuxQED photon formalism [61]. Here, the equivalent photon approximation (EPA) [62] is used to justify setting the photon on–shell at leading order, as this is where a majority of the photon contribution will lie. This allows for the factorization theorem to apply, separating the soft contributions into a photon PDF, which can be directly determined by structure function data. Recently, a different methodology has been developed where the structure functions are used to directly predict new observable, bypassing the need to define a photon PDF in some cases [63, 64]. This can lead to significant reduction of errors and does not require the photon to be on–shell. This new method will be discussed in the context of WW pair production in chapter 6.

With PDF errors reducing with the addition of new precise data from the LHC, those in the PDF fitting community are beginning to quantify how the theoretical uncertainties within the hard scattering cross section can be propagated into the PDF error and what impact this has [65]. This can relate to a number of different theoretical uncertainties; however, the main focus has been on the Missing Higher Order (MHO) uncertainties obtained when one varies the renomalization scale, μ_R, or the factorisation scale, μ_F. Such variation is also performed when using the PDF to predict a new observable and potential correlations between the MHO uncertainties generated when fitting, and those produced when making new predictions may need be taken into account to avoid potential double counting. This is particularly difficult for μ_R, as the MHO uncertainties will be correlated between process with similar higher order corrections, but uncorrelated otherwise. As little to no direct information on the datasets used in the fit is present in the final PDF, assessing these

potential correlations will be challenging when new predictions will be made. Current discussions regarding these effects and if the impact is significant are ongoing [66,67].

The treatment of the heavy quark has also generated discussions in recent years, with NNPDF now treating the charm as having its own parametrised PDF [44]. Traditionally, the heavy quarks are described by an evolved PDF that is turned on when the scale is around the mass of the quark and is then evolved via the DGLAP equations. As the factorisation theorem only applies for massless partons, the heavy quark PDF is only truly defined for scales much higher than the quark mass. The effect of adding in the evolved PDF at scales just above the quark mass is determined by a scheme that is chosen by the PDF fitting group. A more detailed discussion of the evolved PDF is given in 4.

Work is also ongoing to extend the PDF formalism beyond the standard unpolarized proton. The most notable of these are the extensions to nuclear PDFs, where an additional variable for the number of nucleons in the nucleus is included. Nuclear PDFs have already been generated by nNNPDF [68] and nCTEQ [69]; however, due to the limited data available on nuclear collisions, the errors are significantly larger than their proton counterpart. In addition, work has also been done on polarized PDFs, where the incoming proton or lepton has a definite spin value [70], and transverse momentum PDFs where the parton is allowed a small momentum perpendicular to the proton [71].

Finally, work is under way to determine the PDFs directly via lattice gauge theory [5]. Until recently it was thought that such quantities could not be determined in this way due to their kinematic nature; however, recent work has revealed that this is possible via what is known as a pseudo–PDF. While these currently lie far away from the fitted PDFs, it is hoped that these will become more reliable, and be used to constrain the PDFs further.

2.4.4 Future of PDFs

We now discuss the predicted future of PDFs, which can broadly be split into two categories: data improvements and theoretical advancements.

We first describe the data improvements, which will result from the new colliders that will be running in the future as well as upgrades to existing ones. One of these is the High–Luminosity phase of the LHC (HL–LHC) [72, 73], which, as the name suggests, will upgrade the LHC and provide a significant boost to the instantaneous luminosity. This will allow for a considerable improvement to the statistics collected by the detectors, enabling more precise data, and subsequently PDFs, to be obtained. One proposed detector is the Large Hadron–electron Collider (LHeC) [74] which intends to use the proton beam from the LHC to collide with electrons. This experiment will, in general, suffer less from complicated QCD effects and thus potentially provide cleaner results than the LHC. The predicted impact of this detector and the HL–LHC on the PDF uncertainty is described in chapter 5. Additionally, a completely new electron–ion collider will be constructed, the Electron Ion Collider (EIC) [75]. Here, high energy and polarized electrons will collide with protons and ion beams at a variable centre of mass energy of 20–100 GeV with a large luminosity. This collider will also place some constraints on the PDFs, particularly at high x, and will potentially have a significant impact on the currently poorly determined nuclear PDFs due to the variety of ion beams that would be available.

With more precise data, comes the necessity to provide improved theoretical calculations. The most obvious improvement will come from the higher order determinations of the splitting functions and the hard–scattering cross section. For the latter, currently most predictions are computed at NLO with NNLO and electroweak k–factors included. This will likely move on to full NNLO determinations being pro-

vided, as is already done with some processes, with the `APPLfast` project [76] intending to make this more widespread. Another theoretical consideration is the changing the DGLAP equations to the more general BFKL dynamics [77, 78], which include resummation effects at low x. Evidence suggesting the need for these equations has already been found in the context of the NNPDF formalism [79].

Chapter 3

Differential Top Quark Pair Production at the LHC: Challenges for PDF Fits

In this chapter we present the results of PDF fits to differential top quark pair production at the LHC within the MMHT framework. We in particular consider ATLAS data in the lepton + jet and dilepton channels and CMS data in the lepton + jet channel, both at 8 TeV. We focus on the ATLAS lepton + jet data, for which correlations of the statistical and systematic errors are provided across the four relevant distributions for PDF determination, namely p_T, M_{tt}, y_t and y_{tt}. The results of this study have been published in [3], and have helped guide the inclusion of this dataset into the MSHT20 fit [12].

3.1 Introduction

One area of particular phenomenological relevance in the fitting of PDFs is the determination of the gluon PDF at high x, which can, for example, play an important role in predictions for new heavy BSM states via gluon–initiated production. However, until more recently, this has suffered from a relative lack of direct constraints, resulting in rather large PDF uncertainties. However, with the advent of high precision

LHC data at higher masses and transverse momenta, this situation has in principle changed rather dramatically. As discussed in [80–82], the inclusion of inclusive jet, Z boson transverse momentum, and differential top quark pair production in PDF fits places important constraints on the high x gluon that were lacking in earlier PDF fits. In all cases, these benefit from cutting edge high precision theory calculations at NNLO in QCD [83–86].

On the other hand, the inclusion of such processes in PDF fits is not without issues. In the case of jet production, it is well known [57, 80, 87] that the ATLAS jet production data at 7 and 8 TeV cannot be well fit across all rapidity bins. For the Z boson transverse momentum distribution, issues with fitting the ATLAS 7 TeV (normalized) data [88], and the CMS 8 TeV data [89] in the highest rapidity bin have been reported [82], while more generally in this study an additional source of uncorrelated error, assumed to be due to residual theoretical uncertainties and possible underestimated experimental errors, has to be introduced to achieve a good fit.

For differential top quark pair production, the first study of its inclusion in PDF fits was presented in [90] with approximate NNLO theory and later in [81], where the full NNLO calculation was used in the context of a global fit with rather encouraging results. Here, 8 TeV data from ATLAS [13], and CMS [14], both in the lepton + jet channel, were considered. These data were presented differentially in the top quark pair invariant mass, $m_{t\bar{t}}$, and rapidity, $y_{t\bar{t}}$, and the individual top quark/anti–quark transverse momentum, p_T, and rapidity, y_t. However, as the corresponding statistical correlations across these different distributions were not then available, only one of these distributions from each dataset could be included at once. In all of the above cases, the data were corrected back to the top quark parton level.

This situation has changed quite recently, with the ATLAS collaboration pro-

viding the corresponding statistical correlations for their 8 TeV data [38], while the systematic errors were provided in the original data. This therefore allows a simultaneous fit to all four distributions to be performed for the first time, in principle providing a greater discriminating power in the determination of the high x gluon. However, as discussed in [38], when even two distributions are fit simultaneously within the context of the ATLAS PDF fit, only a very poor fit quality can be achieved, while a similar observation has been made by the MMHT [91, 92] and CT [93] collaborations. This therefore casts doubt on the reliability of including such data in PDF fits. It should in particular be emphasised that this issue is not bypassed by fitting to one individual distribution alone, which would artificially mask the issues in fitting the complete dataset, while also risking introducing a bias into the fit, through the particular choice of distribution that is made. A different conclusion is found by NNPDF [94], although here they concentrate on the normalized distributions while for this study we focus on the absolute case due to the lack of available systematic and statistical correlations between the total cross section and the normalized distributions.

Further study of the above effects is clearly essential, and indeed in [38], the source of this poor fit quality was traced to a small subset of dominant two–point systematic uncertainties associated with the choice of Monte Carlo generator or input parameters used in the data unfolding to the parton–level results which are included in the PDF fit. A rather significant improvement in the fit quality, with a relative stability in the resultant gluon, was then observed by introducing some decorrelation in a given systematic error between distributions, reminiscent of the studies [80, 87] for the case of jet production. On the other hand, this study was only performed within the ATLAS PDF framework, fitting to a limited dataset and with a rather restrictive parametrisation, while fits to only two distributions were made rather

than the complete set of four.

In this chapter, we therefore present the first consistent fit to the full ATLAS lepton + jets dataset, within the global MMHT framework. We study in detail the variation in the fit quality that comes from fitting individual datasets in comparison to the combination, the effect that introducing some degree of decorrelation in the two–point experimental systematics has, and the impact of this on the resultant gluon PDF. We consider the impact of including the NNLO QCD and NLO EW corrections to the top quark production matrix elements on the gluon, finding that these are rather smaller than those due to the prescription for treating the experimental systematic correlations. As we will discuss, this is equally true if an individual distribution alone is fit to, which in effect implicitly assumes a complete decorrelation of all systematics between distributions.

We in addition consider the ATLAS data in the dilepton channel [15], finding that here a good fit to the individual distributions can be achieved. While a fit to the total dataset cannot be performed, as the corresponding statistical correlations are not available, this suggests a more reliable fit might be achieved here. Finally, we consider the CMS lepton + jet data, for which distributions must also be fit individually due to a lack of available statistical correlations between them. We particularly focus on the rapidity distributions, y_t and y_{tt}, and find that a fair description can be achieved in the latter case, while the fit quality is rather poor in the former. However, we note that the systematic correlations are all presented as positive, a situation that is unlikely for normalized data, raising questions on its validity.

3.2 Data Sets and Theoretical Calculation

We consider ATLAS 8 TeV data on differential top quark production in lepton + jet [13] and dilepton [15] channels. The lepton + jet data are presented as single differential distributions with respect to a number of parton–level kinematic variables. The four which are useful for this analysis are: the average transverse momentum of the $t(\bar{t})$–quark (p_T), the invariant mass of the $t\bar{t}$ pair (M_{tt}), the average rapidity of the $t(\bar{t})$–quark (y_t) and the rapidity of the $t\bar{t}$ pair (y_{tt}). Correlations between individual systematic errors are available both within and between distributions, as well as a full statistical correlation matrix. For the dilepton dataset, this is again presented as single differential distributions with respect to two relevant kinematic variables for PDF fitting, namely the rapidity and invariant mass of the $t\bar{t}$ pair, however the systematic and statistical correlations are not available between these distributions. In this study we will consider individual fits to the two cases.

We in addition consider the CMS 8 TeV in the lepton + jets channel [14], which are given in terms of the same kinematic variables as in the ATLAS case. These data are presented as normalized distributions, which we simply multiply by the total cross section measured in the same channel [95], taking the total statistical, systematic and luminosity errors associated with this as fully correlated between all bins. However, we do not take into account the correlations of systematic uncertainties between the total cross section and the normalized bins as these are not available. The statistical correlation matrix within each distribution is provided, but not across distributions. Therefore, as in the ATLAS dilepton case, only a single distribution at a time can be fit.

For the theoretical input, predictions at NNLO in QCD for the ATLAS lepton + jets and CMS data are provided in [96] as fastNLO grids [97], while NLO elec-

troweak K–factors are evaluated using the predictions of [98], as provided using the PDF4LHC15LUXqed [61] set. As the K–factors should be relatively insensitive to the PDF set used, this should provide a good estimate for the current case. For the ATLAS dilepton data, NLO QCD Applgrids [33], NNLO QCD K–factors and NLO electroweak K–factors were supplied by the ATLAS collaboration [99]. As NNLO grid files for the CMS dilepton data are not currently available, we do not consider these in what follows.

3.3 Lepton + Jets Channel at ATLAS

We first consider the ATLAS data at 8 TeV, in the lepton + jet channel, using the recently provided full statistical correlation matrix [38]. As discussed in the introduction, significant issues in achieving a good fit to these data have been found by ATLAS, with similar observation made by CT and MMHT. We analyse this in detail below.

3.3.1 Fitting within the MMHT framework

As a baseline, we use the MMHT14 PDF set [46], but including the final HERA I + II combination data [100] along with some new total $t\bar{t}$ cross section data, and finally including some minor improvements to the fitting code. When we add in the ATLAS data to this while also removing the corresponding total cross section measurement to avoid double counting, though the effect of this is rather small.

In the fit we use the figure of merit:

$$\chi^2 = \sum_{i,j=1}^{N_{\rm data}} \left(D_i - T_i - \sum_{\alpha=1}^{N_{\rm sys}} \beta_{i,\alpha}\lambda_\alpha \right) (\mathrm{cov}_{\rm stat})^{-1}_{ij} \left(D_j - T_j - \sum_{\gamma=1}^{N_{\rm sys}} \beta_{j,\gamma}\lambda_\gamma \right) + \sum_{\alpha=1}^{N_{\rm sys}} \lambda_\alpha^2, \tag{3.1}$$

where D_i and T_i are the data and theory predictions, $\beta_{i,\alpha}$ are the systematic errors, cov_{stat} is the statistical covariance matrix and λ_α are the nuisance parameters, due to the experimental correlated systematic errors. Here we have symmetrized the systematic errors and have shifted the central values of the data points accordingly to account for this. While in principle one can minimise the above χ^2 analytically, by profiling with respect to these nuisance parameters, we have kept them explicit here to aid the discussion which follows.

The χ^2 per data point after fitting, for the combined fit to all four distributions, as well as various individual fits to be discussed further below, is shown in Table 3.1. We can see immediately that the fit quality in the combined case is extremely poor, with a χ^2 of ~ 7 per point. If we consider instead a subset of two distributions, as in the ATLAS study [38], the fit quality is better but still very poor. To be concrete, we show the result of a fit to the p_T and M_{tt} distributions, though for other choices the results are rather similar. We also show results excluding the statistical correlations to assess their impact and find that these can play an important role, but that even when excluding them artificially, the fit quality is still poor. Finally, we find that even fitting to the distributions individually, while a good description of the p_T and M_{tt} cases is possible, the fit to the rapidity variables is still rather poor, with a $\chi^2 \sim 3$ per point. This is consistent with the results of the ATLAS study [38], as well as CT [93], but interestingly not with the original NNPDF study of [81], where a good description of all individual distributions was found. Here, the impact of statistical correlations within the distributions, which were not available for the NNPDF study, is seen to be minimal, and so cannot explain this difference. A subsequent NNPDF analysis [94] included these correlations and an updated systematic covariance matrix, finding a poor fit quality when combining all four absolute distributions simultaneously.

We now consider the impact on the gluon itself, shown in Fig. 3.1, for fits to

Distribution	Statistics Correlated	Statistics Uncorrelated
p_T	0.53	0.50
y_t	3.12	3.16
y_{tt}	3.51	3.51
M_{tt}	0.70	0.60
$p_T + M_{tt}$	5.73	2.47
Combined	7.00	3.28

Table 3.1: χ^2/N_{data} values for fits to different distributions within the ATLAS 8 TeV lepton + jet data, as well as for the combined fit to all four distributions. The left (right) columns correspond to the case that the statistical correlations are included (excluded).

different distributions, both individual and in combination. Broadly speaking, we can see that, as expected, the data has a noticeable impact on the high x gluon, both in terms of the central value and uncertainty. However, on closer inspection, we can see that the impact of the individual fits on the gluon is somewhat different, with the rapidity variables tending to decrease the gluon at high x, while the p_T and M_{tt} tend to pull in the opposite direction. While the resulting gluon PDFs in all cases agree within their error bands, indicating that there is no strong tension between them, a difference in the overall pulls is clear. This difference in trend is also observed in the ATLAS study [38].

3.3.2 Understanding the poor fit quality: the role of correlated systematic errors

In the previous section we observed a very poor fit quality to the ATLAS 8 TeV lepton + jet data when two or more distributions were fit simultaneously. Here, we wish to try and understand this effect further. In particular, as discussed in the ATLAS analysis [38], the experimental uncertainties for this dataset are completely dominated by the experimental systematic errors. For the rapidity distribution, the

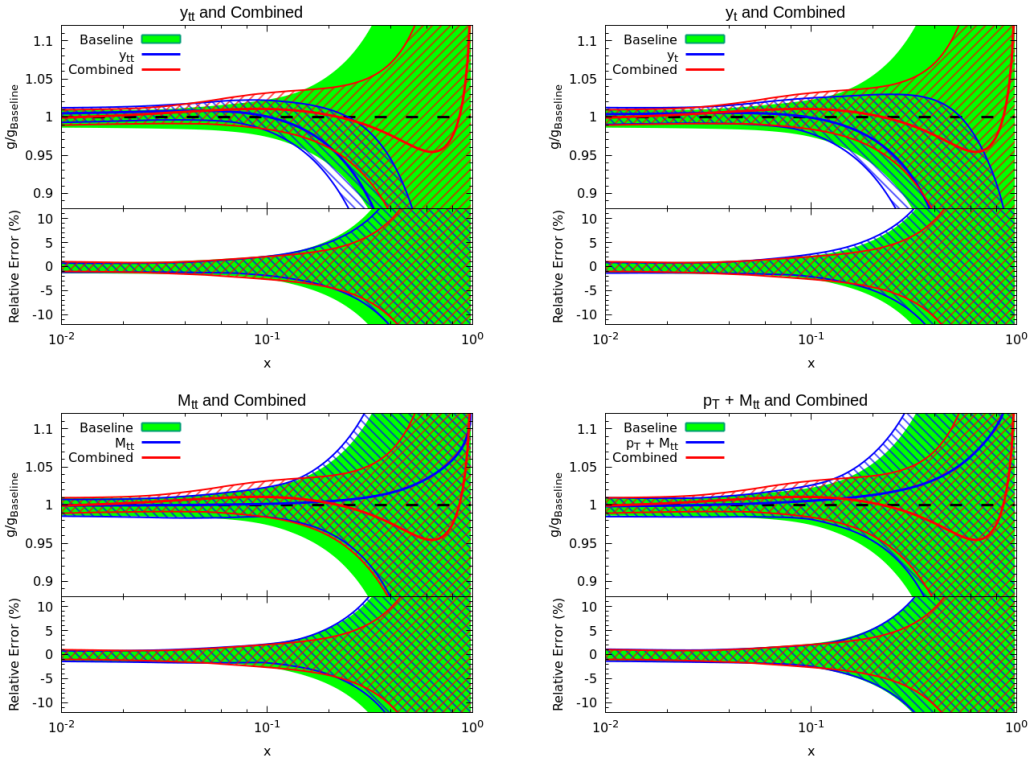


Figure 3.1: Impact of the ATLAS 8 TeV lepton + jet data on the gluon PDF. Results for fits to individual distributions as well as the combined $p_T + M_{tt}$ case are shown, while the result from a combined fit to all four distributions is shown for comparison in all cases.

three largest sources of systematic error are $\sim 3-6$ % (hard–scattering model), $\sim 6-9$ % (ISR/FSR) and ~ 3 % (parton shower model), depending on the distribution and rapidity bin, while the statistical uncertainty is only $\sim 0.6-1.3$ %. We will return to these systematic errors shortly, but for now simply note that the fit quality will naturally be driven by the precise treatment of the correlations in these errors. In particular, as we will see, it is the correlation across different distributions that drives the observed deterioration in fit quality.

To investigate the above effect further, we first follow the study of [80] and evaluate the variation in the preferred values of the corresponding systematic shifts in Eq. (3.1) when fitting to the different distributions individually. Any significant tension between these values will then indicate that when fitting the distributions

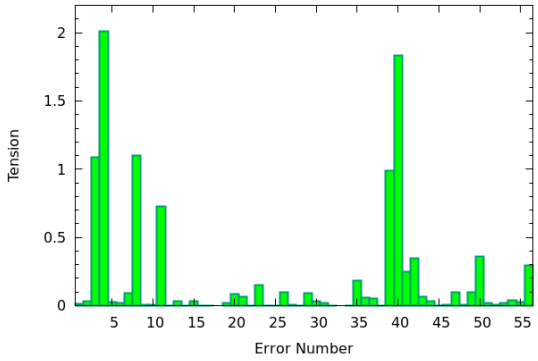


Figure 3.2: The tension between shifts for the ATLAS 8 TeV lepton + jet data, between the individual distributions. This is calculated as $(1/6)\sum_{j=1}^{4}\sum_{l>j}^{4}(\lambda_{i,j}-\lambda_{i,l})^2$ where $\lambda_{i,j}$ is the i^{th} nuisance parameter calculated when fitting the j^{th} distribution.

in combination, some deterioration in the fit quality will occur, and moreover that a less restrictive degree of correlation between the distribution for the corresponding shift may be preferred by the data/theory comparison. However, it should be stressed that this can clearly only be taken as a guide, as one must of course also consider whether it is reasonable to consider a different correlation scenario for the corresponding shift, or whether this is already precisely determined by the experimental analysis, with no further decorrelation being possible. The result is shown in Fig. 3.2, and we find that the jet energy scale (4) and ISR/FSR (40) errors provide the largest source of tension. As the correlation in the former may reasonably be assumed to be well determined experimentally, we will not consider this further. On the other hand, in the latter case this is less clear, as we shall now discuss (see also [38]).

In fact, further investigation reveals that the errors relating to the hard–scattering model and the parton shower are very similar to the ISR/FSR error for most bins, as well as being the largest systematic errors, as mentioned above. This suggests that there is some redundancy in our analysis, with the large tension in one of these errors

masking the tension in the others. Thus, it is reasonable to consider all three of these errors as a potential source of tension. Now, the important point here is that in all three cases these are 2–point errors evaluated using two choices of Monte Carlo (MC) generator or generator inputs. In particular, a given choice of MC generator and input parameters must be taken to unfold the observed data back to the parton–level distributions which we use for our PDF fits. To evaluate the systematic uncertainty due to this choice, a second generator/set of input parameters is used to unfold MC signal events generated with the default MC. The difference between this result and parton–level signal is then assigned to be the systematic uncertainty in each bin. More specifically, in the case of the hard scattering the default signal is generated with `MC@NLO` + `Herwig` and the unfolding is performed with `POWHEG`+`Herwig`, while for the parton shower the difference is between `POWHEG`+`Herwig` and `POWHEG`+`Pythia`, and for ISR/FSR a variation in the input parameters of `POWHEG`+`Pythia` is taken.

Now, crucially in all cases the correlation in the systematic error is also taken from this procedure, with it being assumed that any correction factor evaluated by this two–point procedure should be applied in a fully correlated way across all bins. To some extent this is a reasonable assumption: if a different choice of MC generator/input parameters for the unfolding leads to a larger parton–level result at larger $m_{t\bar{t}}$ and p_t, say, then there may well be physics or kinematic reasons for these effects to be correlated. However, there is clearly some uncertainty in this degree of correlation, and indeed if one performed the same study as above with a third choice of MC/input parameters, there is no reason to assume that in all cases the corresponding correction would always lie within the correlation two–point band. More specifically, when determining the systematic errors, it is assumed that the result of a potential third MC or choice of input parameters would be perfectly describable by taking some particular value of the single λ_α shift assigned to this error

source. Put another way, the effect of this is constrained to be a linear combination of the two baselines used to evaluate the systematic error, with the combination being the same across all bins. This may not necessarily be the case, and indeed, given the issues with fit quality observed above, which as we will see are to a large extent driven by these correlated systematics, it is reasonable to investigate the effect of loosening the assumed correlation scenarios for these.

A further related issue is the assumption implied in Eq. (3.1) that the corresponding systematic uncertainties are standard symmetric Gaussian sources of error. As the discussion above makes clear, these two–point systematic uncertainties, as well as not providing a well–defined degree of correlation, are in addition expected to be inherently non–Gaussian. For example, there is no particular reason to associate a 1–σ uncertainty, as opposed to a smaller or larger allowed variation, with the uncertainty bands calculated in the above way. These error sources can in addition be rather asymmetric. While the parton shower and hard scattering errors are provided by ATLAS after symmetrisation, the ISR/FSR error does have some degree of asymmetry, particularly in the high invariant mass bins (though less so in the rapidity distributions, where already we see a poor individual fit quality). In the latter case we choose to simply symmetrize as described above, but this is only one possible choice. More generally, it would be desirable to investigate the impact of different assumptions about the above effects, and of including a non–Gaussian probability distribution for these error sources. From a practical point of view this makes a direct application of Eq. (3.1) much less tractable, as a simple analytic minimisation with respect to the corresponding nuisance parameters is not possible. A full analysis of these issues is beyond the scope of the current study but is clearly important in the future.

Here, we therefore concentrate solely on the question of correlation, and in par-

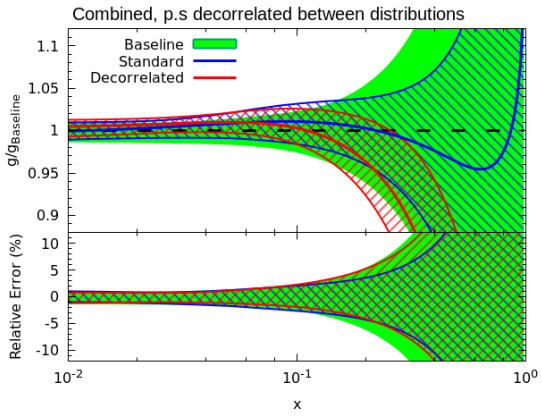


Figure 3.3: Impact of the ATLAS 8 TeV lepton + jet data on the gluon PDF. The result of the default to the combined dataset is shown, as well as decorrelating the parton shower systematic uncertainty between the four distributions.

Distribution	p.s. correlated	p.s. decorrelated
Combined	7.00	1.80
$p_T + M_{tt}$	5.73	0.66

Table 3.2: χ^2/N_{data} values for fits to the ATLAS 8 TeV lepton + jet data, including decorrelation of the parton shower systematic uncertainty described in the text.

ticular investigate the effect of loosening the correlation in the systematic error associated with parton shower, guided by the discussion above, and following the choice made in the ATLAS analysis [38]. We in particular consider the same combined fit to the four distributions as before, including all statistical correlations, but now completely decorrelating the parton shower systematic error between the four distributions. That is, we split this source of uncertainty into four sources, with each only contributing to a given individual distribution, while keeping all correlations within the distributions for now untouched. The result is shown in Table 3.2, and we can see that the effect is dramatic, leading to a factor of ~ 4 decrease in the χ^2. This confirms the discussion above, showing that indeed the assumed degree of correlation in these two–point systematics has a significant impact on the fit quality. In Fig. 3.3 we show the impact this has on the corresponding gluon. We can see that there is

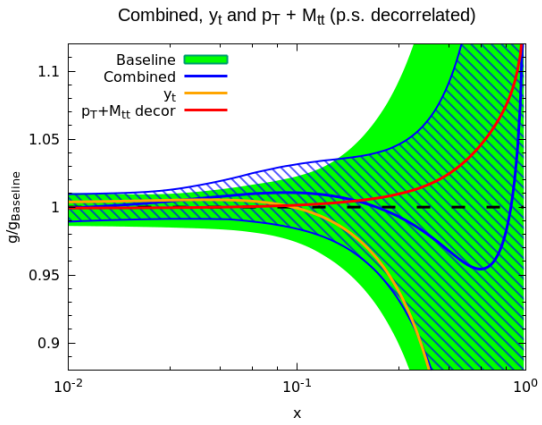


Figure 3.4: As in Fig. 3.1, for fits to different distributions taken in the literature: y_t and $p_T + M_{tt}$ (with decorrelation of the parton shower systematic error), as well as the combined fit.

some non–negligible difference in the result, such that the gluon is not completely stable under different treatments of the systematic correlations. Moreover, we can see that now the pull on the gluon is in fact close to the pull from the individual fits to the rapidity variables, see Fig. 3.1, suggesting these distributions are having a more dominant effect after the decorrelation.

We summarise the situation, directly comparing to choices that have been made in the literature, in Fig. 3.4. Here we show the impact of fits to difference distributions, namely the individual rapidity y_t, the combined $p_T + M_{tt}$ (including parton shower decorrelation), and the fit to all four distributions, on the same plot. A rather large spread in results is observed, emphasising the need for further investigation.

Given that the rapidity distributions appear to be driving the gluon impact after decorrelating across distributions, it is worth investigating the fit these distributions individually further. In particular, given that the individual fit quality to these is rather poor, see Table 3.1, we should investigate the impact of the above correlations within the rapidity distributions on the extracted gluon. According to the arguments above, it is equally possible that the degree of correlation taken within the rapidity

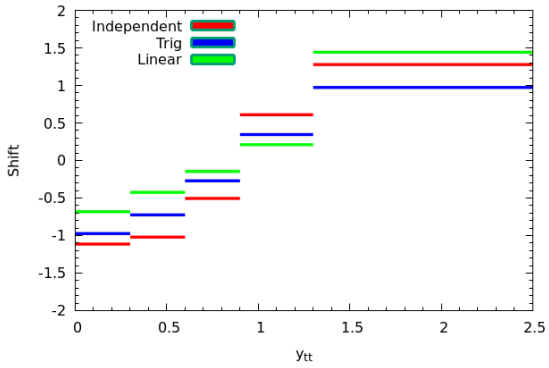


Figure 3.5: Values for individual shift parameters (λ_i in Eq. (3.1)) for the parton shower error when fitting the y_{tt} distribution of the ATLAS 8 TeV lepton + jet data, while decorrelating the error between each data point.

distribution, as well as across distributions, for the two–point MC uncertainties may be overly restrictive. To investigate this, we will consider the $y_{t\bar{t}}$ distribution, though similar results are found for y_t. We first allow the parton–shower uncertainty to be completely uncorrelated across the bins of the rapidity distribution, with one corresponding nuisance parameter per bin that are each allowed to shift independently. The result is shown in Fig. 3.5, and we can see that rather a smooth shape in the shifts is preferred. While in all cases the magnitude of the shift is ~ 1, indicating that the shifts are not being pushed particularly outside the 1σ bands corresponding to the magnitude of the systematic errors in each bin, the shape is clearly quite different from that required by the default correlation prescription.

Motivated by this, we allow for a degree of decorrelation within the rapidity distribution, following the approach of [87], where a similar decorrelation was introduced for various sources of experimental systematic and theoretical uncertainty within the rapidity and $p_\perp$ bins of the ATLAS 8 TeV inclusive jet data. In particular, we split the parton shower error into two components for each bin, with the first error following some function of the distribution variable, and the other chosen such

that they add in quadrature to the original error. We choose two simple approaches, a linear function with the errors given by:

$$\beta_i^{(1)} = \left(\frac{y_{tt,i} - y_{tt,\min}}{y_{tt,\max} - y_{tt,\min}}\right)\beta_i^{\text{tot}}, \qquad \beta_i^{(2)} = \left[1 - \left(\frac{y_{tt,i} - y_{tt,\min}}{y_{tt,\max} - y_{tt,\min}}\right)^2\right]^{\frac{1}{2}}\beta_i^{\text{tot}} \tag{3.2}$$

and a trigonometric function given by:

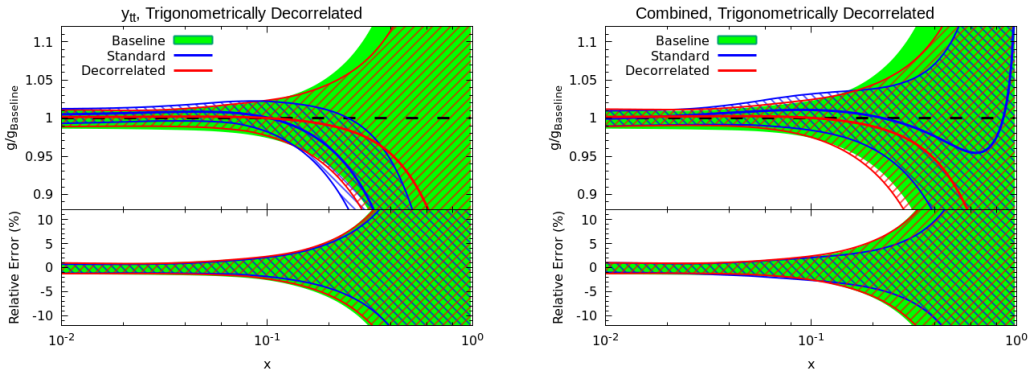
$$\beta_i^{(1)} = \cos\left[\pi\left(\frac{y_{tt,i} - y_{tt,\min}}{y_{tt,\max} - y_{tt,\min}}\right)\right]\beta_i^{\text{tot}}, \quad \beta_i^{(2)} = \sin\left[\pi\left(\frac{y_{tt,i} - y_{tt,\min}}{y_{tt,\max} - y_{tt,\min}}\right)\right]\beta_i^{\text{tot}}. \tag{3.3}$$

We note that the trigonometric decorrelation has the advantage of naturally adding in quadrature to the magnitude of the original error, while allowing some additional freedom in the relative sign of the shift across the distribution. Although it may seem preferable to instead use a factor of $\pi/2$ in the arguments for sine and cosine to preserve symmetry, the above choice provides a better description of the desired shifts and a higher quality fit.

The results of the fit to the y_{tt} distribution are shown in Table 3.3. We can see that the effect of the trigonometric decorrelation is significant, reducing the χ^2 per point to ~ 1, over a factor of 3 lower than in the default case. For the linear case the effect is rather less pronounced, though some reduction is seen, suggesting that the trigonometric decorrelation allows for the data to be fit more consistently within our framework. This is demonstrated in Fig. 3.5, where it can be seen that this decorrelation more closely represents the shifts seen in the completely decorrelated case. In Fig. 3.6a we show the impact on the gluon PDF for the trigonometric case. We can see that some shift in the central value of the gluon occurs, albeit within PDF uncertainties. More significantly, the resultant gluon uncertainty is rather larger, and indeed very similar to the uncertainty on the baseline set. It would therefore appear

Distribution	p.s. correlated	p.s. decorrelated
$y_{t\bar{t}}$ (linear)	3.51	2.62
$y_{t\bar{t}}$ (trig.)	3.51	1.02
Combined ($y_{t\bar{t}}$ (trig.))	7.00	1.62

Table 3.3: χ^2/N_{data} values for fits to the ATLAS 8 TeV lepton + jet rapidity distributions, including decorrelation of the parton shower systematic uncertainty described in the text.



(a) (b)

Figure 3.6: Fits to the y_{tt} distribution (left) and all distributions together (right) with different decorrelation procedures along y_{tt}, compared to fits with no decorrelation.

that by allowing for a weaker degree of correlation, much of the constraining power of this particular distribution that appeared to be present in the default case (albeit with a rather poor fit quality, raising questions about the reliability of this) has been washed out. In essence, by allowing a greater freedom in the correlation of the systematic error associated with parton shower, to a large extent it appears that the data can be fit simply by shifts of the corresponding nuisance parameters, resulting in the end in a rather small constraining power on the gluon itself. We note that for this effect to occur, it is not necessary for the original fit quality to be bad and can happen even if the original fit quality is good. Nonetheless, if the correlations taken for such dominant sources of systematic error are too constraining, this may still result in an artificially large constraint on the corresponding gluon PDF.

Finally, we should consider the impact of this within a fit to the combined distributions, that is by decorrelating the parton shower error between each set and along the y_{tt} distribution using the trigonometric decorrelation. The result, shown in Table 3.3, is a χ^2 of 1.62 per point, only marginally better than correlating solely between sets. The impact on the gluon is shown in Fig. 3.6a (b), where we see a relatively small shift in the central value, but again an increase in the corresponding PDF uncertainty to the level of the baseline set. Thus, again in this case the constraining power of the data appears to be somewhat washed out. We note that the fit quality of 1.62 per point is still relatively poor, being driven by the fact that the y_t distribution is still not well described, see Table 3.1. We have investigated the effect of introducing the same decorrelation as above to the y_t distribution, and indeed find an improvement in the fit, while the gluon is quite similar to Fig. 3.6a. Simply omitting this rapidity distribution gives a similar result to this.

3.3.3 Theoretical precision

In the previous section, we have observed that the degree of correlation taken for the dominant two–point systematic errors (taking the parton shower as a test case) can have a rather large effect on the extracted gluon from the fit to the ATLAS lepton + jet data. This is clearly a cause of concern, in particular in light of the high precision being aimed for in the PDF fit. In particular, as discussed in the introduction of this chapter, we can perform a fit using cutting edge NNLO in QCD theoretical predictions for the parton–level top quark pair production process [86], and including NLO EW corrections on top of this. The theoretical precision in these predictions is therefore high, but while one would be tempted to claim a similar degree of precision in the corresponding PDF extraction, the above results cast some

doubt on this, at least for the ATLAS lepton + jet data.

To investigate this point further, we evaluate the impact of including NNLO QCD and NLO EW corrections to the top quark pair production matrix element on the extracted gluon PDF. To be specific, we will evaluate the top quark pair production cross section at NLO, NNLO and NNLO×EW order, while keeping all other fit settings the same, i.e. we in particular use NNLO PDFs throughout, in order to isolate the effect of these corrections. The impact on the fit quality is shown in Table 3.4, and one can see that the NNLO QCD corrections lead to a better description of the data in comparison to the NLO, as we would hope for, while interestingly the EW corrections leads to some deterioration in the fit quality. On the other hand, the impact of these effects on the extracted gluon PDF when fitting to the m_{tt} and y_{tt} distributions, shown in Fig. 3.7, is seen to be rather small and for the other distributions, the difference is smaller still.

Further, in Fig. 3.8, we compare the impact on the gluon of the theoretical precision in the cross section calculation to the treatment of the parton shower error in the data, for the case that all four distributions are fit simultaneously. We in particular show the result of a fit including NNLO×EW corrections, with and without decorrelation, as well as the result of a fit including NLO corrections alone, with no decorrelation. We can see that the difference between the NLO and NNLO×EW cases is indeed rather smaller than the difference between the NNLO×EW case with decorrelation to the case without. Thus, if we are to extract the gluon PDF from such data with an accuracy to match the high precision provided by the NNLO×EW calculation of the underlying cross section, it will be crucial to have a clearer understanding of these dominant two–point systematic uncertainties and the uncertainty on their correlation. However, in the meantime we can decorrelate these systematic errors if there is a clear reason to do so, in order for a reasonable fit to be obtained.

Distribution	NLO	NNLO	NNLO+EW
p_T	0.65	0.36	0.53
y_t	2.99	2.98	3.12
y_{tt}	4.06	3.30	3.51
M_{tt}	1.33	0.57	0.70
All	7.88	6.61	7.00

Table 3.4: χ^2/N_{data} values for fits to different distributions within the ATLAS 8 TeV lepton + jet data, using NLO, NNLO and NNLO+EW theory for the top quark pair production cross section.

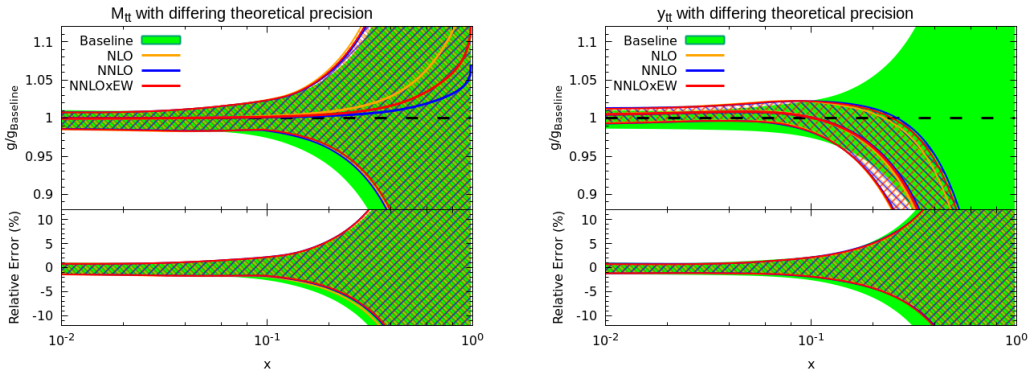


Figure 3.7: Impact of the ATLAS 8 TeV lepton + jet data on the gluon PDF, for different levels of precision in the theoretical prediction for the top quark pair production cross section, when fitting the M_{tt} (left) and y_{tt} (right) distributions.

3.4 Dilepton Channel at ATLAS

In the preceding section, we have considered a fit the ATLAS 8 TeV data collected in the lepton + jet channel, however data in the dilepton channel [15] at 7 and 8 TeV are also available, of which we fit to the higher precision latter dataset. Unfortunately, the statistical and systematic correlations are not provided across distributions, and hence a complete study cannot be performed. We therefore instead consider fit to the y_{tt} and M_{tt} distributions individually. The results are shown in Table 3.5, with the fit quality found to be very good, consistent with [38], and indeed for the M_{tt} case in particular, the fit quality is anomalously low, suggesting that the experimental

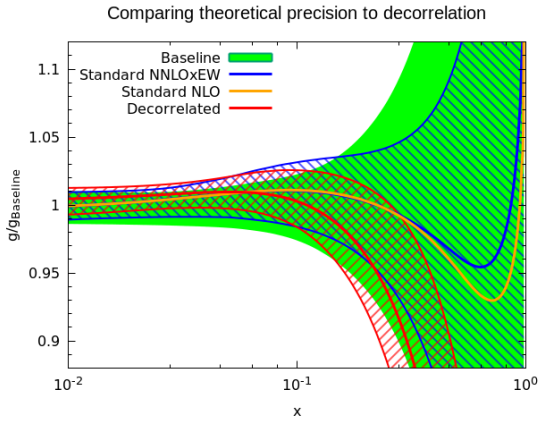


Figure 3.8: Extracted gluon from a fit to the ATLAS 8 TeV lepton + jet data (all four distributions), from a fit including NNLO×EW in the cross section with ('decorrelated') and without ('standard') decorrelation of the parton shower systematic error, and from a fit with pure NLO in the cross section calculation, without decorrelation.

errors may be overestimated. The impact on the gluon PDF is shown in Fig. 3.9, and are seen to be broadly consistent within errors, though the rapidity prefers a somewhat lower gluon. In both cases, some reduction in the PDF uncertainty at higher x is seen, comparable in size with the same distributions in the lepton + jet channels. Indeed, it is interesting to note that the y_{tt} distribution has a similar pull to the individual and combined decorrelated fits for the lepton + jet channel.

We can therefore see that the issues present in the lepton + jet case are not apparent in the dilepton channel. Indeed, this may be natural in light of the fact that the impact of the type of two–point MC uncertainties present in the lepton + jet case should be somewhat smaller in this cleaner channel. On the other hand, given that the most significant issues in the lepton + jet case came about when one considered a combined fit to all available distributions, it is difficult to make a firm statement without performing a combined fit to the y_{tt} and M_{tt} distributions. However, the fact that the fit quality is rather low in the case of the rapidity, and anomalously low in the case of the M_{tt} distribution may warrant some further investigation.

Distribution	χ^2/N_{data}
M_{tt}	0.06
y_{tt}	0.66

Table 3.5: χ^2/N_{data} values for fits to the ATLAS 8 TeV dilepton data.

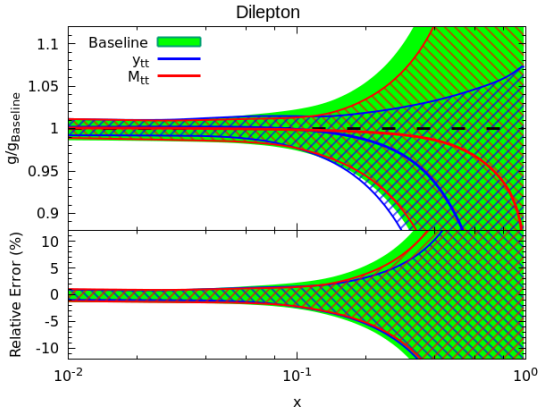


Figure 3.9: Impact of the ATLAS 8 TeV dilepton data on the gluon PDF, for the y_{tt} and M_{tt} distributions.

3.5 Lepton + jets Channel at CMS

The final dataset to be considered in this analysis is the lepton + jets channel measured by CMS [14]. While statistical correlations are provided within each distribution, they are not available across the different distributions, and hence fits can only be performed on these individually. Further, we note that CMS provide a break–down for each systematic error in each bin into individual sources which, as in [81,93], we treat as a set of correlated errors across the individual distribution, one for each source. However, with our findings from the ATLAS case in mind, we compare to the case that these systematic errors are instead treated as uncorrelated. We note that the values for the errors given in the breakdown are all positive, strongly calling into question their interpretation as a correlated source of errors across the bins of normalized distributions. The data are only presented as normalized, and hence

Distribution	Correlated	Uncorrelated
$p_\perp$	3.14	1.49
y_t	2.71	1.25
y_{tt}	1.70	1.39
M_{tt}	5.81	2.97

Table 3.6: $\chi^2/N_{\rm data}$ values for fits to the CMS 8 TeV lepton + jets data with the systematic errors taken as correlated and uncorrelated, as described in the text.

we multiply by the corresponding total cross section measurement [101] to translate to the absolute case here. We treat the total systematic, statistical and luminosity uncertainty as additional sources of correlated systematic in all cases. Correlations between the total cross section and the normalized bins are ignored as these are not currently available. Finally, as in [102], we remove the final bin in each distribution so that the covariance matrices for these normalised data are non–singular.

The results of these fits are shown in Table 3.6. The fit quality is fair in the case of the y_{tt} distribution, while for the others it is noticeably worse, particularly for the M_{tt}. Interestingly, if we assume the systematic errors as uncorrelated we see a dramatic improvement in the fits to p_T, y_t and M_{tt}, with some improvement in y_{tt}. We leave a more detailed analysis of these effects for further study, in particular given the possible questions about the precise degree of correlation one should assume. The impact on the gluon PDF is shown in Fig. 3.10. We can see that all distributions have an impact at high x, with the y_{tt} distribution having a larger constraining power. We also note that, in all cases, the pull is in the same direction as the ATLAS rapidity distributions, see Figs. 3.1 and 3.9.

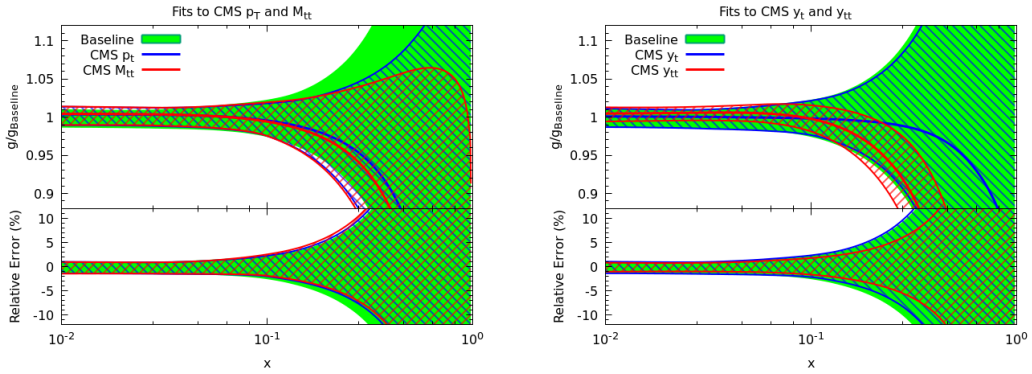


Figure 3.10: Impact of the CMS 8 TeV lepton+jets data on the gluon PDF.

3.6 Combining Datasets

Finally we move on to combining the datasets already discussed into a final fit. We take the combination chosen by ATLAS in [38], namely the M_{tt} and p_T distributions with the parton shower error decorrelated from the lepton + jets channel, and the y_{tt} distribution from the dilepton channel (henceforth referred to as ATLAS default). On top of this, we add in the CMS distributions measured through the lepton + jets channel individually. The results for these are presented in Table 3.7.

The 'ATLAS default' combination was chosen so as to be in line with the choice made by ATLAS in [38], and as such a direct comparison of the corresponding PDF impact can be made as shown in Fig. 3.11. Here we see that at high x, the $t\bar{t}$ data causes the two baseline PDF sets (which themselves fit to rather different datasets) to converge. While they still disagree with each other in terms of error bands at rather high x, this is certainly encouraging to see.

The 'ATLAS default' combination results in a deterioration in the ATLAS dilepton y_{tt} distribution, however this is reduced upon adding in the CMS data in most cases, suggesting a good description of the ATLAS data can be achieved in conjunction with the CMS data. Interestingly this is not the case for the CMS y_t distribution.

Distribution	$p_T + M_{tt}$ ATLAS lepton+jets (0.66)	y_{tt} ATLAS dilepton (0.66)	CMS lepton+jets	Total
ATLAS default	0.74	1.26	N/A	0.87
ATLAS default + CMS $p_\perp$	0.83	0.93	3.13 (3.14)	1.45
ATLAS default + CMS y_t	0.74	1.27	2.71 (2.17)	1.44
ATLAS default + CMS y_{tt}	0.90	0.63	1.79 (1.70)	1.13
ATLAS default + CMS M_{tt}	0.86	0.85	5.87 (5.81)	2.02

Table 3.7: χ^2/N_{data} breakdown for combined fits to ATLAS and CMS $t\bar{t}$ data. The fourth column corresponds to the relevant distribution from CMS that is used in the fit. The numbers in brackets indicate the χ^2/N_{data} when this distribution is fit individually.

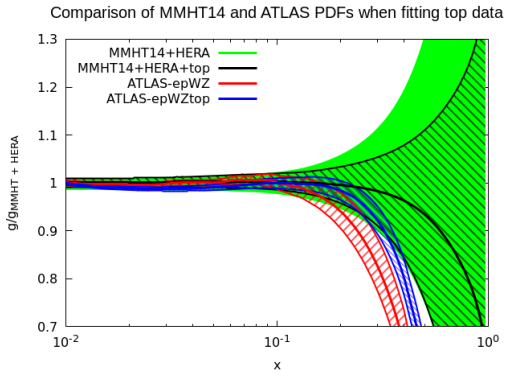


Figure 3.11: Impact of the ATLAS default combination on the MMHT14 and ATLAS-epWZ [103] PDF sets.

The impact of these final fits on the gluon is shown in Fig. 3.12, where we can see clearly that the ATLAS default combination provides good constraints on the high x gluon as expected. The CMS data further improves on the constraints imposed, with the y_{tt} distribution having a stronger impact as before, while the others have very minimal constraining power. Considering the dynamically determined tolerances (see section 2.4.1), we find that three eigenvector directions are constrained by the ATLAS dilepton and one by the ATLAS lepton + jets data in most of the fits

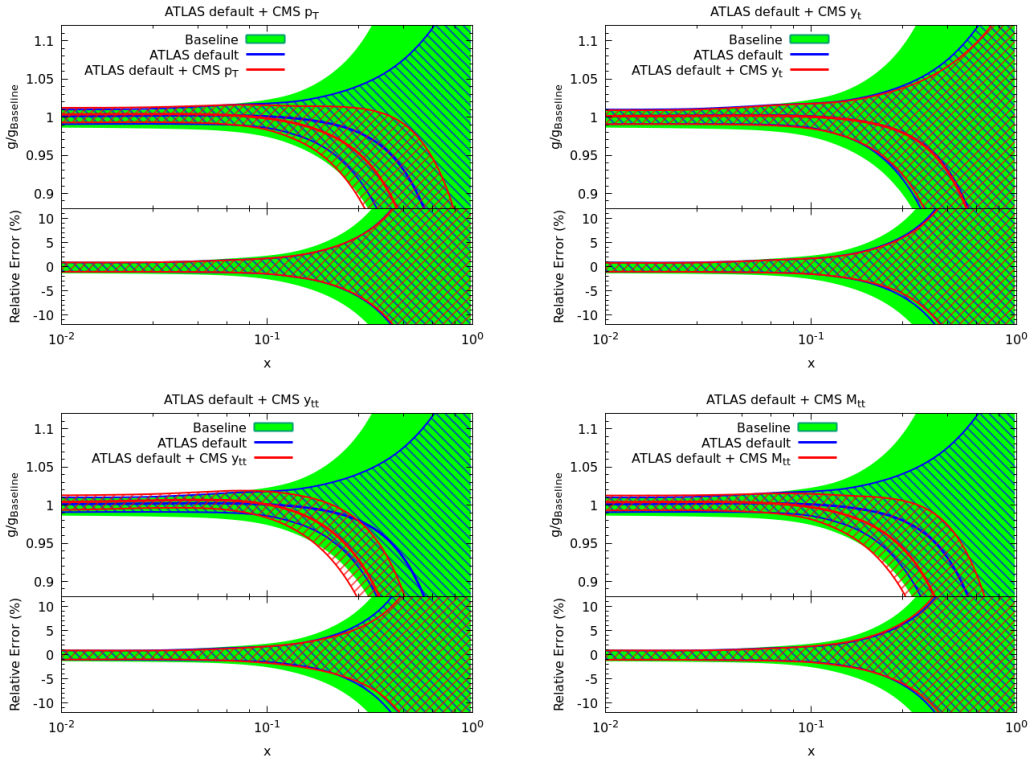


Figure 3.12: Impact of combining the default ATLAS combination of p_T and M_{tt} distributions with the parton shower error decorrelated from the lepton+jets channel and the y_{tt} distribution from the dilepton with the CMS rapidity distributions on the gluon PDF.

to the ATLAS default combination, i.e. both with and without the CMS data. The only exception to this is the fit to the CMS y_{tt} distribution, where this distribution and the ATLAS dilepton data confine three directions each, further demonstrating the constraining power of this distribution from CMS. Finally, we note that in all fits presented here, both individual and combined, no significant deterioration is observed in the fit quality to other datasets present in the fit, indicating that the $t\bar{t}$ datasets have no obvious tension with these.

3.7 Implications for the MSHT20 Fit

The results of this study have been used to guide the implementation of the ATLAS 8 TeV lepton + jets within the MSHT fit. Here all four distributions were used in the fit and the trigonometric decorrelation model was applied to all of them instead of just y_{tt}. This resulted in a decrease in $\chi^2/N_{\rm dat}$ from 6.84 to 1.04, enabling this dataset to be fit consistently. This was however shown to have large impacts on the gluon, similar what we have seen in the MMHT14 case (see Fig. 3.6b).

The trigonometric decorrelation model was also used to improve the fit to the ATLAS 7 TeV inclusive jet distributions [57] which also show similar complications [104]. Here, the multi–jet balance asymmetry, the jet flavour response and the multi–jet fragmentation uncertainties were decorrelated as these all correspond to genuine MC two–point uncertainties. It is shown that such decorrelation has minimal impact on the fit quality (going from $\chi^2/N_{\rm dat} \sim 1.8$ to $\chi^2/N_{\rm dat} \sim 1.6$) when compared to a full decorrelation ($\chi^2/N_{\rm dat} \sim 0.9$). The overall impact of these data is to pull the gluon down at high x, with the decorrelation model weakening this by a noticeable amount, albeit within the uncertainty bands. This can be compared to the effects of the decorrelation model on the ATLAS 8 TeV $t\bar{t}$ data measured though the lepton + jets channel, where a similar effect on the gluon was observed in Fig. 3.6.

3.8 Summary and Outlook

In this study we have investigated, in detail, the effect of including LHC differential top quark production data within the global MMHT fit framework. Data for this have been provided by ATLAS and CMS in both the lepton + jet and dilepton decay channels, corrected back to the top quark parton level. These are then pre-

sented differentially in terms of various kinematic variables, namely the top quark pair invariant mass, $m_{t\bar{t}}$, and rapidity, $y_{t\bar{t}}$, and the individual top quark/antiquark transverse momentum, p_T, and rapidity, y_t. Such data are in particular a sensitive probe of the gluon PDF at high x.

The main focus of this study has been on the ATLAS 8 TeV data presented in the lepton + jet channel, which is currently the only dataset for which the corresponding statistical and systematic error correlations have been provided across all distributions. The reason we concentrate on this is that it provides in principle the most constraining overall dataset, allowing all distributions to be fit at the same time without introducing the potential bias of choosing a particular distribution to fit. Moreover, this allows us to test whether full consistent dataset can in fact be described by the NNLO parton–level theory entering the PDF fit in a way that can only be partially done by considering individual distributions.

We find, consistent with the ATLAS internal analysis, that only a very poor fit quality can be achieved to the full ATLAS dataset. This raises questions about the reliability of including such data with default correlations in PDF fits, and in particular, in simply choosing one distribution to fit, which would simply mask the underlying issues rather than resolving them. With this in mind we investigated the causes of this poor quality in detail, finding that it is dominated by the correlation prescription provided for the dominant sources of experimental error, namely two–point MC errors associated with the data unfolding back to parton–level; specifically due to the MC generator/input parameters used for the underlying hard scattering process, for the resulting parton shower and for ISR and FSR effects. We then discuss how the correlations provided for these errors across and within the different distributions may well be overly restrictive and find that, with some fairly mild loosening of these correlations, a significantly better fit quality can be achieved. Un-

fortunately, this has a non–negligible impact on the central value and uncertainties of the extracted high x gluon, resulting in particular in a rather smaller reduction in the corresponding PDF uncertainties after this decorrelation is introduced. Moreover, we explicitly compare this effect on the gluon with that of either excluding the NNLO QCD or NLO EW corrections to the top quark cross section and find that it is significantly larger. Therefore, caution is required in claiming such data with default uncertainties as a high precision probe of the gluon PDF, with the impact on the PDFs noticeably reduced upon decorrelation.

We have in addition considered both the ATLAS 7 and 8 TeV data in the dilepton channel and the CMS 8 TeV data in the lepton + jet channel. In the former case we find a very good fit quality, with a reasonable impact on the gluon PDF from the top quark rapidity that is in fact rather similar to the pull from the corresponding distribution in the lepton + jet channel. This might indicate that the somewhat cleaner dilepton channel, for which the impact from the dominant systematic errors in the lepton + jet case should be smaller, may be a more promising channel to consider. However, it should be emphasised that, as here the full experimental correlations across distributions have not been provided, there is a potential danger that similar underlying issues might be masked by the fits to individual distributions that we must necessarily perform. For the CMS data, we have considered the 8 TeV lepton + jets channel, using the same distributions as in the ATLAS case. We find a reasonable description of the y_{tt} distribution can be achieved, while the fit quality is poor for the others. These results are however highly sensitive to the treatment of the systematic error correlations. This is again found to be sensitive to the precise degree of correlation one assumes in the underlying systematic errors, though we have not studied this in detail here. Again, as the statistical correlations between the distributions are not available, we cannot perform a full analysis, including all

distributions simultaneously.

In summary, we have found that the prescription for treating the correlated systematic errors in the case of the most comprehensive ATLAS lepton + jet data on differential top quark pair production plays a significant role in the overall constraining power of this dataset, which is in fact greater than the impact of either NNLO QCD or NLO EW corrections on the extracted gluon PDF. While we consider the above data explicitly, it should be emphasised that this effect is not limited to the ATLAS lepton + jet case but may simply be masked in the other considered ATLAS and CMS datasets, where full correlations across distributions are not provided. We have introduced some reasonable procedures for decorrelating the dominant two–point MC uncertainties to judge the (rather large) impact on the fit, but this is clearly just a first step, rather than a firm prescription. Rather, a more complete understanding from both the experimental and theoretical point of view of the precision we can ascribe to the determined correlations of these uncertainties will be essential in the future. Moreover, the current common approach of fitting to an individual distribution risks introducing bias in the fit, and arguably will mask rather than resolve the above issues. Certainly, if the assumptions about the understanding of these correlations are too strong, we have found evidence that this may lead to an unreliable PDF determination, making a more conservative approach desirable. Alternatively, it might be that considering the relatively cleaner dilepton channel and/or alternative kinematic variables, perhaps say at the level of the decayed leptons rather than the corrected parton–level top quarks, might provide a cleaner probe of the gluon PDF.

Chapter 4

Updates to the Strange PDF from Dimuon Data

In this chapter we discuss updates to the theoretical prediction of the dimuon process as measured at CCFR and NuTeV. These data are particularly interesting for their constraints on the strange PDF. We perform fits to assess the impact of these updates, with a particular focus on potential tensions with the ATLAS Drell–Yan data, which places competing constraints on the strange quark. Some results shown in this chapter were presented in [92] and the updated theoretical predictions were used in the MSHT20 fit [12].

4.1 Introduction

The extensive and wide–ranging data from electron–proton collisions at experiments such as HERA provide excellent constraints on the PDFs and form a significant portion of most PDF fits. However, as discussed in section 2.3.1, such experiments only constrain certain combinations of the PDFs, with some sensitivity beyond this when including NLO effects. One example of a PDF that is poorly determined from these data is the strange quark distribution. While information can be obtained through, for example, charged current structure functions, the constrains placed are rather poor as it is difficult to extract this PDF from the combinations relevant for DIS

(see Eqs. (2.10) and (2.27)). To this end, detectors have been developed to probe different DIS interactions, of which we turn our attention to the CCFR and NuTeV detectors [18, 19], where (anit–)neutrinos interact with Fe nuclei. Here we focus on the dimuon production, where two distinct muons are detected in the final state. This particular process has the potential to put tight constraints on the strange quark and has been included within PDF fits for some time [105]. Since then, ATLAS W/Z production data at 7 TeV [47] were presented, placing further constraints on the strange quark. However, the original analysis of this dataset resulted in a much larger strangeness [103] than had been predicted by global fits at the time, suggesting some tension between this and the dimuon data. Until recently, only an approximate NNLO calculation was available for the dimuon process; however, now full heavy mass dependent NNLO corrections to the Fixed Flavour Number Scheme (FFNS) calculation are available [17] and have already been applied in the context of NNPDF [106] and CTEQ [30] PDF fits. In this study we analyse the impact of these corrections to the dimuon process in the context of a MMHT14 fit, paying particular attention on the tensions between this dataset and the ATLAS W/Z production data. We also implement updates to the General Mass–Variable Flavour Number Scheme (GM–VFNS) contributions calculated using the TR$'$ scheme.

4.2 Theoretical Background

We begin with a description of the dimuon process and how the cross section is calculated, taking into account the effects of heavy quarks. As with typical DIS

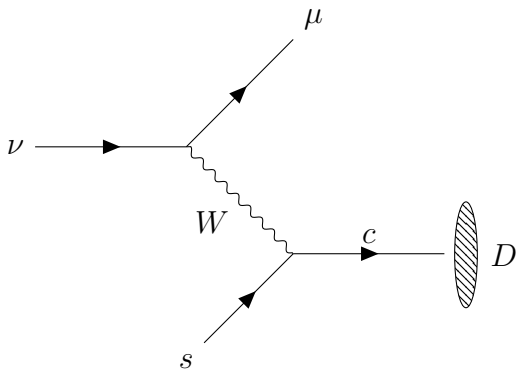


Figure 4.1: Dimuon production at CCFR/NuTeV. Here the D meson decays into a μ.

processes, this cross section is written in terms of the structure functions:

$$\frac{\mathrm{d}^2\sigma}{\mathrm{d}x\mathrm{d}Q^2} = \frac{G_F^2}{4\pi}\frac{M_W^2}{(Q^2+M_W^2)^2}\left[\left(Y_+ - \frac{2m_p^2x^2y^2}{Q^2}\right)F_2(x,y) - y^2F_L(x,y) \pm Y_- xF_3(x,y)\right], \tag{4.1}$$

where y is the inelasticity, m_p is the mass of the proton, M_W is the mass of the W boson, G_F is the Fermi constant and $Y_\pm = 1 \pm (1-y)^2$. Here the plus sign refers to neutrino scattering, while the minus sign is for the anti–neutrino case. We present the leading order strange–initiated diagram for dimuon production in Fig. 4.1. Here, the outgoing charm quark hadronises, forming a D meson, which then subsequently decays into a muon. Due to the larger CKM matrix element for the strange–initiated diagram compared to that of the down quark, these get an enhancement. This results in an overall ∼50% of neutrino collisions involving a strange quark and ∼90% of anti–neutrino collisions induced by an anti–strange [107]. The former is significantly lower due to the large down valence competing with the effect of the CKM matrix elements. This, along with the already constrained d quark, enables the dimuon process to place strong constraints on the strange quark.

However, one must also consider charm–initiated diagrams which contribute even

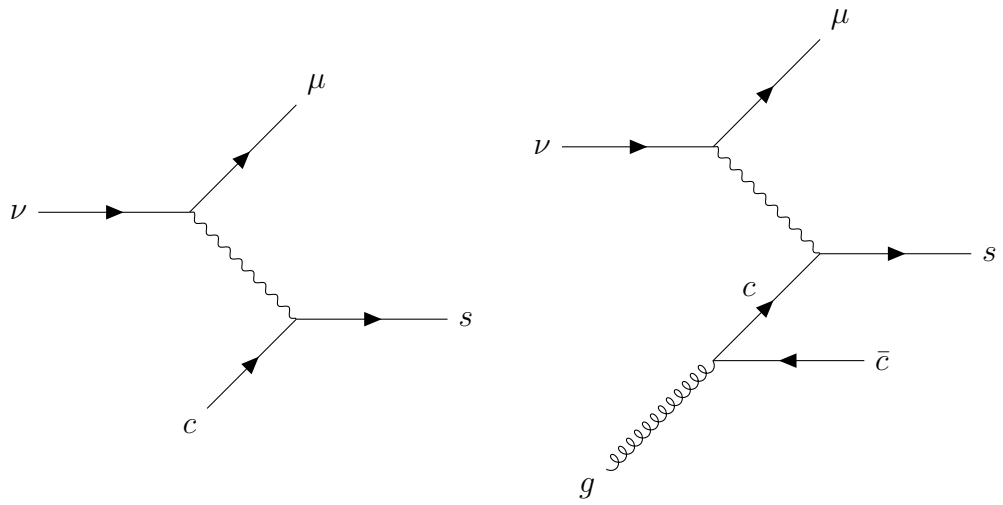


Figure 4.2: Left: the charm–initiated LO diagram for dimuon production. Here there is an implicit anti–charm within the decay products of the Fe ion. Right: gluon initiated dimuon production at NLO. Here the additional charm is explicit.

at leading order as shown in Fig. 4.2 (left). It should be noted that, even though there is no explicit charm quark in the final state, we know it must be there implicitly due to the flavour conserving nature of the strong force. This diagram and the equivalent higher order contributions were not removed from the experimental acceptance corrections [16], and, given the nature of the fixed–target process we are considering, one would expect that the charm quarks emitted collinear to the incoming neutrino would still be measured. We therefore include such diagrams in this calculation. How exactly a heavy quark PDF should be treated is an open question, as the factorization theorem assumes that the partons are massless. At high scales, this is approximately true and thus a charm PDF should provide an accurate description of the process in this region. We could therefore describe the high energy limit by introducing a heavy quark PDF and taking its mass to be zero. This is referred to as the Zero Mass–Variable Flavour Number Scheme (ZM–VFNS). However, such a scheme is unreliable at low scales of $Q^2 \sim m_c^2$, as mass dependent terms are ignored. Contributions from $c \to s$ diagrams can also be included without the need of a charm

PDF by including the gluon splitting $g \to c\bar{c}$ contributions at one order higher as shown in Fig. 4.2 (right). Here the charm in the final state is now explicit and the full dependence on the charm mass is included. Adding in the charm quark in this way is referred to as the Fixed Flavour Number Scheme (FFNS) as we only include the light quark and gluon PDFs. However, this calculation is logarithmically enhanced with terms proportional to $\ln(Q^2/m_c^2)$. This results in the FFNS giving inaccurate results at high scales, whereas the ZM–VFNS has the same contributions resummed to all orders via the DGLAP equations, enabling a much better description in this kinematic region [108, 109]. Thus we overall have two schemes that work well in two distinct regions of phase space: the FFNS, which includes the full dependence on the heavy quark mass and thus behaves well when $Q^2 \sim m_c^2$, and the ZM–VFNS, which resums logarithmic terms, enabling a good description for $Q^2 \gg m_c^2$. The advantages of both calculations can be combined into one scheme, known as the General Mass Variable Flavour Number Scheme (GF–VFNS).

We now derive the GF–VFNS for the inclusion of a general heavy quark of mass m_H in charged current interactions and describe different ways to treat the $Q^2 \sim m_H^2$ region. To begin, we relate the PDFs defined in a FFNS containing n_f flavours, to those defined within a GM–VFNS that contain $n_f + 1$ flavours. This can be done perturbatively as follows:

$$f_i^{\mathrm{VF}}(x, \mu_F^2) = \sum_j \int_x^1 \frac{\mathrm{d}\xi}{\xi} A_{ij}\left(\xi, \frac{\mu_F^2}{m_H^2}\right) f_j^{\mathrm{FF}}\left(\frac{x}{\xi}, \mu_F^2\right), \tag{4.2}$$

where f_i^{FF} and f_i^{VF} are the PDFs of parton i in a FFNS and GM–VFNS respectively, evaluated at a factorization scale μ_F and parton momentum fraction ξ. Here, the coefficients $A_{jk}(\xi, \mu_F^2/m_H^2)$ are perturbative matrix elements describing the relationship between the two sets of PDFs and will contain the splitting functions with the

logarithmic factor $\ln(\mu_F^2/m_H^2)$. It is typical to turn on and start evolving the heavy quark PDF from threshold, taken to be m_H^2 in this case due to the single heavy quark in the final state. Below this scale, the heavy quark PDFs are set to zero, i.e. $f_H^{\rm VF}(\xi,\mu_F^2) = 0$ for $\mu_F^2 < m_H^2$. It is worth noting that, starting from NNLO contributions, there are terms within the coefficient functions, $A_{ij}(\mu_F^2/m_H^2)$, that are independent of $\ln(\mu_F^2/m_H^2)$ [110], generating a discontinuity in the heavy quark PDF at threshold. However, in any sensible scheme, this will not result in a discontinuity in any observable quantity such as the structure functions or cross sections.

With the FFNS PDF set we can calculate the structure functions:

$$F_i(x,Q^2) = \int_\chi^1 \frac{{\rm d}\xi}{\xi} C_{ik}^{\rm FF}\left(\frac{\chi}{\xi}, \frac{Q^2}{m_H^2}, \frac{Q^2}{\mu_R^2}\right) f_k^{\rm FF}(\xi,\mu_F^2), \tag{4.3}$$

where $C_{ik}^{\rm FF}$ are the hard scattering coefficients for the structure function F_i with an incoming parton k calculated within the FFNS and μ_R is the renormalization scale. Here we have introduced a new variable:

$$\chi = x\left(1 + \frac{m_H^2}{Q^2}\right). \tag{4.4}$$

This is done to ensure that the invariant mass of the products, W, is at least as large as the single heavy quark in the final state, $W \geq m_H$. The same calculation using the GM–VFNS gives:

$$\begin{aligned} F_i(x,Q^2) &= \int_\zeta^1 \frac{{\rm d}\xi}{\xi} C_{ij}^{\rm VF}\left(\frac{\zeta}{\xi}, \frac{Q^2}{m_H^2}, \frac{Q^2}{\mu_R^2}\right) f_j^{\rm VF}(\xi,\mu_F^2) \\ &= \int_\zeta^1 \frac{{\rm d}\xi}{\xi} C_{ij}^{\rm VF}\left(\frac{\zeta}{\xi}, \frac{Q^2}{m_H^2}, \frac{Q^2}{\mu_R^2}\right) \int_\xi^1 \frac{{\rm d}\xi'}{\xi'} A_{jk}\left(\frac{\xi'}{\xi}, \frac{\mu_F^2}{m_H^2}\right) f_k^{\rm FF}(\xi',\mu_F^2), \end{aligned} \tag{4.5}$$

where $C_{ij}^{\rm VF}$ are the hard scattering coefficients calculated in the GM-VFNS. To derive

this, we have used the relationship between the FFNS PDFs and GM–VFNS PDFs given in Eq. (4.2). Here we have introduced a different variable, ζ and, unlike χ earlier, this variable is more ambiguous, parametrising the kinematical impact of including a heavy quark in the initial state, despite this being emitted collinearly from the proton. It is unsurprising that such an undetermined quantity appears here, as the definition of a heavy quark PDF is unclear near threshold. We are therefore able to choose a suitable value for this within a particular scheme so long as it obeys the high energy limit, $Q^2 \gg m_H^2$, where GM–VFNS must approach the result of a massless particle: $\zeta \to x$. Given that we would like to have a smooth transition around threshold, we ensure that the GM–VFNS only has contributions for $W \geq m_H$. We thus set $\zeta = \chi$ for the rest of this discussion as this quantity ensures this constraint. Later we will describe two schemes that further define the behaviour of the GF–VFNS coefficients around threshold, both of which make this choice of ζ.

Using Eqs. (4.3) and (4.5), a relationship between the FFNS and GM–VFNS coefficients can be found:

$$C_{ik}^{\mathrm{FF}}\left(x, \frac{Q^2}{m_H^2}, \frac{Q^2}{\mu_R^2}\right) = \int_x^1 \frac{\mathrm{d}\xi}{\xi} C_{ij}^{\mathrm{VF}}\left(\xi, \frac{Q^2}{m_H^2}, \frac{Q^2}{\mu_R^2}\right) A_{ij}\left(\frac{x}{\xi}, \frac{\mu_F^2}{m_H^2}\right). \tag{4.6}$$

When considering the above equation, it is important to note that the value of $\alpha_S(\mu^2)$ is dependent on the number of flavours in the system, and thus will differ between C_{ik}^{FF} and C_{ik}^{VF}, impacting the above relationship at NNLO and beyond. By taking $\mu_R^2 = \mu_F^2 = Q^2$ and performing this calculation for the gluon–initiated processes at

NLO we get:

$$C_{ig}^{\mathrm{VF},(1)}\left(x,\frac{Q^2}{m_H^2}\right) = C_{ig}^{\mathrm{FF},(1)}\left(x,\frac{Q^2}{m_H^2}\right) - \int_x^1 \frac{\mathrm{d}\xi}{\xi} C_{iH}^{\mathrm{VF},(0)}\left(\xi,\frac{Q^2}{m_H^2}\right) P_{qg}^{(1)}\left(\frac{x}{\xi}\right) \ln\left(\frac{Q^2}{m_H^2}\right), \tag{4.7}$$

where the bracketed number represents the order to which the calculation was performed (see Eq. (2.6)). The negative term on the right–hand side of the above equation is an example of a 'subtraction' term and is present to take into account contributions that exist in both the perturbative expansion of the FFNS and the charm evolution in the GM–VFNS. At NLO, this double counting is due to the $g \to c\bar{c}$ splitting, which is present in both the FFNS diagram and the GM–VFNS DGLAP equations. We note that at NNLO, there is a subtraction term containing the matrix element, A_{Hg}^2, which is non–zero at threshold. This precisely cancels the discontinuity present in the heavy quark PDF, resulting in observables with a smooth distribution around $Q^2 \sim m_H^2$. While the above equation relates the coefficients to each other, it does not uniquely define the GM–VFNS coefficients, $C_{ij}^{\mathrm{VF},(n)}$. As discussed before, the only uniquely defined region is $Q^2/m_H^2 \to \infty$, where the GM–VFNS coefficients must tend to the limit where the heavy quark can be treated as massless, which is defined by the ZM–VFNS with coefficients $C_{ij}^{\mathrm{ZM},(n)}$. A perhaps obvious choice would be to define the GM–VFNS heavy quark coefficients, $C_{iH}^{\mathrm{VF},(n)}$, to be exactly what one would obtain from calculating the heavy quark scattering cross section. These will clearly tend to the massless limit when $Q^2/m_H^2 \to \infty$. Contributions from the gluon and light–quark initiated diagrams in this scheme could then be determined using Eq. (4.6). Such a choice was the basis of the original ACOT(χ) scheme [111]. However, simpler choices exist, for example, one could set the GM–VFNS heavy quark coefficients to be identical to those in the ZM–VFNS.

This would, by definition, satisfy the high energy limit and provide a more straightforward calculation compared to including the full mass dependence. This choice was adopted to form a simpler description in an updated ACOT(χ) scheme [112], were they set:

$$C_{iH}^{\mathrm{VF}}\left(\xi, \frac{Q^2}{m_H^2}\right) = C_{iH}^{\mathrm{ZM}}\left(\xi\right). \tag{4.8}$$

The treatment in the $Q^2 \sim m_H^2$ region has already been taken care of by setting $\zeta = \chi$ earlier. With this, the contribution to the structure functions from the heavy quarks in this scheme is given by:

$$F_i^H(x, Q^2) = \int_\chi^1 \frac{\mathrm{d}\xi}{\xi} C_{iH}^{\mathrm{ZM}}\left(\frac{\chi}{\xi}\right) h(\xi, Q^2), \tag{4.9}$$

where $h(\xi, Q^2)$ is the heavy quark PDF. The presence of χ in this equation ensures that the structure functions tend smoothly to zero when approaching threshold from above.

A different scheme that is often used to include heavy quark PDFs is the TR′ scheme [109]. This offers similar results compared to the simplified ACOT(χ) procedure; however, it is often justified differently. Here it is noted that the charm in the dissociated proton's final state places kinematic constraints on x, in particular:

$$x < x_{\mathrm{max}} = \frac{Q^2}{Q^2 + m_H^2}. \tag{4.10}$$

Further, terms of order m_H^2/Q^2 can be swapped between the GM–VFNS coefficients in Eq. (4.7) while still keeping the high Q^2 behaviour. Thus one can modify the coefficients to ensure the above kinematical constraint is satisfied while still complying

with the high energy limit. We can thus write:

$$C_{2H}^{\mathrm{VF}}\left(\xi, \frac{Q^2}{m_H^2}\right) = C_{2H}^{\mathrm{ZM}}\left(\frac{\xi}{x_{\max}}\right). \tag{4.11}$$

With this, the leading order contribution to F_2 from the heavy quark is given by:

$$\begin{aligned} F_2^{H,(0)}(x,Q^2) &= \int_x^{x_{\max}} \frac{\mathrm{d}\xi}{\xi} C_{2H}^{\mathrm{ZM},(0)}\left(\frac{\xi}{x_{\max}}\right) h\left(\frac{x}{\xi}, Q^2\right) \\ &= \frac{x}{x_{\max}} h\left(\frac{x}{x_{\max}}, Q^2\right). \end{aligned} \tag{4.12}$$

We note here the different limits on the integral when compared to the ACOT(χ) scheme (see e.g. Eq. (4.9)), however, these integrals are related by a simple change of variables. Further, we have additionally made the replacement of $x \to x/x_{\max}$ in the prefactor by again swapping terms of order m_H^2/Q^2 in Eq. (4.7).

If a similar formula was also chosen for $F_1^{H,(0)}$, then one would obtain non–zero contributions to the longitudinal structure function, $F_L = F_2 - 2xF_1$. However in the case of the TR$'$ scheme, one instead fixes the leading order contribution to F_L from the heavy quarks to be zero while altering the higher order contributions:

$$C_{LH}^{\mathrm{VF},(n)}\left(\xi, \frac{Q^2}{m_H^2}\right) = 2\left(\frac{1}{1+m_H^2/Q^2} - \frac{1}{2}\right) C_{LH}^{\mathrm{ZM},(n)}\left(\frac{\xi}{x_{\max}}\right), \qquad n \geq 1. \tag{4.13}$$

This choice ensures that the high Q^2 behaviour is correct and F_L tends to zero at threshold. This also suppresses F_L at scales just above this, avoiding a potential kink in the structure function in this region [113].

However, this definition leads to the expansion of F_L starting at different orders of α_S in different limits. Below threshold we only have the FFNS contribution, which begins at $\mathcal{O}(\alpha_S^0)$ due to the effects of the heavy quark in the final state. In particular,

the contribution from the strange PDF at leading order is [105]:

$$F_L^{s,(0)} = \frac{m_c^2}{Q^2} xs\left(\frac{x}{x_{\max}}, Q^2\right). \tag{4.14}$$

However, at high energies this tends to zero, leaving us with only contributions from the GM–VFNS which, as defined above, has its expansion start at $\mathcal{O}(\alpha_S)$. Hence, we have two distinct regions that begin at different orders and a choice of ordering must be made. Below threshold we simply include the FFNS calculated at the desired order. Above threshold, we include contributions from the FFNS and GM–VFNS at the same order and add in terms calculated at one order higher but modulated by a factor of $(1 - m_H^2/Q^2)$. Thus at threshold we make the transition:

$$\begin{aligned} F_L^{\mathrm{FF},(n)}(x,Q^2) \to & F_L^{\mathrm{FF},(n)}(x,Q^2) + F_L^{\mathrm{VF},(n)}(x,Q^2) \\ & + \left(1 - \frac{m_H^2}{Q^2}\right)\left(F_L^{\mathrm{FF},(n+1)}(x,Q^2) + F_L^{\mathrm{VF},(n+1)}(x,Q^2)\right). \end{aligned} \tag{4.15}$$

Therefore, around threshold, $Q^2 \sim m_H^2$, F_L is calculated up to order α_S^n, while at high energies we calculate this to order α_S^{n+1}, as the expansion starts at one order higher and smoothly transition between these with a factor of m_H^2/Q^2. For this study, we will calculate F_L up to NNLO around threshold; however, we will not include N^{3}LO corrections at high energies as would be suggested by the above equation. These are only expected to make a significant impact in the small x and Q^2 region, which is currently not well measured in charged current interactions [105]. We thus omit these terms for simplicity.

4.3 Impact of full NNLO calculation on MMHT

We now implement the full NNLO calculation into the MMHT14 framework. The dimuon process has already been included within the MMHT14 PDF for some time now, with the current implementation using only approximate NNLO [114]. As was done in chapter 3, as a baseline we use the MMHT14 PDF set [46], including the final HERA I + II combination data [100] along with some new total $t\bar{t}$ cross section data, and finally including some minor improvements to the fitting code. We shall refer to this combination as 'baseline' throughout this section. Nuclear corrections are applied using results of de Florian et al., shown in Fig. 14 of [115]. These consist of multiplicative factors, $R_f(x, Q^2, A)$, relating the proton PDFs to the nuclear PDFs with A nucleons. Further, we assume that the heavy quark corrections are the same as that for the strange quark. We allow for the nuclear corrections to vary in normalization and shape by multiplying by a flavour independent correction factor [116]:

$$\text{NUCMOD}(x) = \begin{cases} (1 + 0.03r_1)[1 + 0.015r_2 \ln^2(x_m/x)] & x < x_m \\ (1 + 0.03r_1)[1 + 0.015r_3 \ln^2(x/x_m)] & x \geq x_m, \end{cases} \tag{4.16}$$

where r_i are free parameters allowed to vary without penalty and $\ln(x_m) = -2.5$ is chosen to be roughly in the centre of the x values spanned by the neutrino data included within the MMHT14 fit. This factor typically allows for a variation of a few percent. Acceptance corrections [117] accounting for the 5 GeV cut on the energy of the muon resulting from the charm decay are also applied. It should be noted that both the nuclear and acceptance corrections are calculated up to NLO and are thus currently less precise than the now updated FFNS calculation. The VFNS determi-

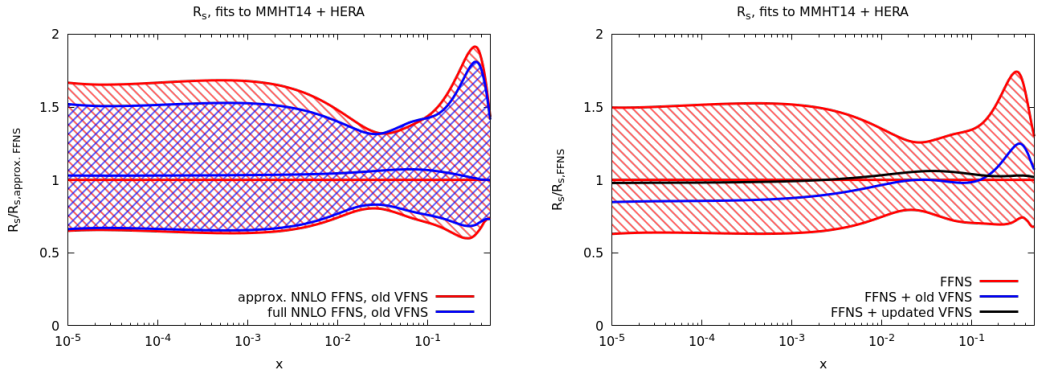


Figure 4.3: Left: Impact of the full NNLO calculation on the fitted value of R_s in the MMHT14 framework with the old VFNS determination. Right: Determination of R_s using various calculations for the VFNS in the MMHT14 framework with the FFNS contribution calculated to full NNLO.

nation is not included with the full NNLO grids, and thus we continue to use the approximate NNLO contributions that were already present in the previous analyses. However, when implementing the updated FFNS, an inconsistency was discovered in the calculation of the VFNS contribution. Here, the subtraction terms were calculated using the TR′ scheme, as was intended; however, the charm–initiated diagrams were inadvertently calculated using the ACOT(χ) scheme. Thus Eq. (4.7) is not satisfied, resulting in a discontinuity in the structure functions due to the NNLO Q^2 independent terms for the charm quark that is not cancelled by the subtraction terms. This error has now been corrected and we will discuss the impact of this as well as the introduction of the full NNLO calculation of the FFNS coefficients.

We begin by assessing the impact of the full FFNS calculation on the PDF fits in the context of the old VFNS calculation. We plot the quantity $R_s = (s+\bar{s})/(\bar{u}+\bar{d})$ in Fig. 4.3 (left) at a scale $Q^2 = 4$ GeV, where we expect the result to be impacted by the choice of scheme due to the proximity to m_c^2. Here there is minimal impact on the overall fitted value of R_S, with a mild increase in the region of $x \sim 10^{-2} - 10^{-1}$. There is however some reduction of the overall uncertainty in the upper error bound over a

significant region of x values, particularly for $x < 10^{-2}$. We note that NNPDF also found a slight enhancement of the strangeness in the same region when including the full NNLO FFNS calculation for the NuTeV data into their 'str_base' fit [106], which is comparable to the fits shown in Fig. 4.3 (left). We next move onto the updated VFNS calculation in Fig. 4.3 (right), for which we fix the FFNS contribution to be that from the full NNLO determination in all cases. Here we see that, when compared to the old VFNS calculation, the updated VFNS has minimal impact on the fitted R_s value. Overall, when going from the old calculation to the new one, R_s goes up over a large range of x by about 20%. The high x region is an exception to this, with the opposite occurring. The χ^2 figure of merit for these fits and those that follow are shown in Table 4.1. We note that for the NuTeV data, there is a high degree of correlation between neighbouring bins which can be accounted for by introducing an effective number of points ($\sim$40) that is much less than the actual number of data points (84). This method is only approximate, with modern datasets instead accounting for these effects in the form of a statistical correlation matrix. Further, MSHT20, NNPDF3.1 and CT18 all present their results using the total number of data points, and therefore we do the same here so that an easy comparison can be made. However, we note that, in all fits performed within the MMHT14 framework, we still use the effective number of points to determine the 68% confidence region and its impact on the dynamical tolerance (see section 2.4.1). From the χ^2 values, we see that the updated theory has opposite effects on the two detectors, with CCFR preferring the updated calculation, while NuTeV prefers the old determination. This overall results in a few points of χ^2 being added to the combination of both when updating the FFNS calculation to include the full NNLO determination, while updating the VFNS results in little overall impact.

We next move on to adding in the ATLAS WZ data. Fits assessing the impact

Dataset	CCFR	NuTeV	ATLAS WZ	Total
Baseline (approx FFNS + old VFNS)	0.90	0.51		0.71
Baseline (Full FFNS + old VFNS)	0.84	0.62		0.73
Baseline (Full FFNS + + updated VFNS)	0.80	0.66		0.73
Baseline (approx FFNS + old VFNS) + ATLAS WZ	0.83	0.67	1.78	1.02
Baseline (Full FFNS + old VFNS) + ATLAS WZ	0.80	0.92	1.77	1.10
Baseline (Full FFNS + updated VFNS) + ATLAS WZ	0.76	0.85	1.75	1.05
MSHT20	0.79	0.70	1.91	1.05
$N_{\rm dat}$	86	84	61	170/231

Table 4.1: Overall fit quality, $\chi^2/N_{\rm dat}$, for the CCFR/NuTeV and ATLAS WZ datasets for fits including various datasets and different theoretical calculations of the dimuon process.

of this dataset on the baseline PDF set are shown in Fig. 4.4 (top left). Here we see an overall increase in the value of R_s over the entire kinematic range. This is most noticeable at higher values of x, where the best fit values of the two PDFs diverge. However, it should be noted that the error bands of these two fits broadly overlap, thus this is only suggestive of a potential tension between this dataset and those within the baseline PDF. We now update the FFNS calculation to include the full NNLO contributions and perform the same comparison, displaying the fitted PDFs in Fig. 4.4 (top right). Here we appear to see a reduction in the apparent

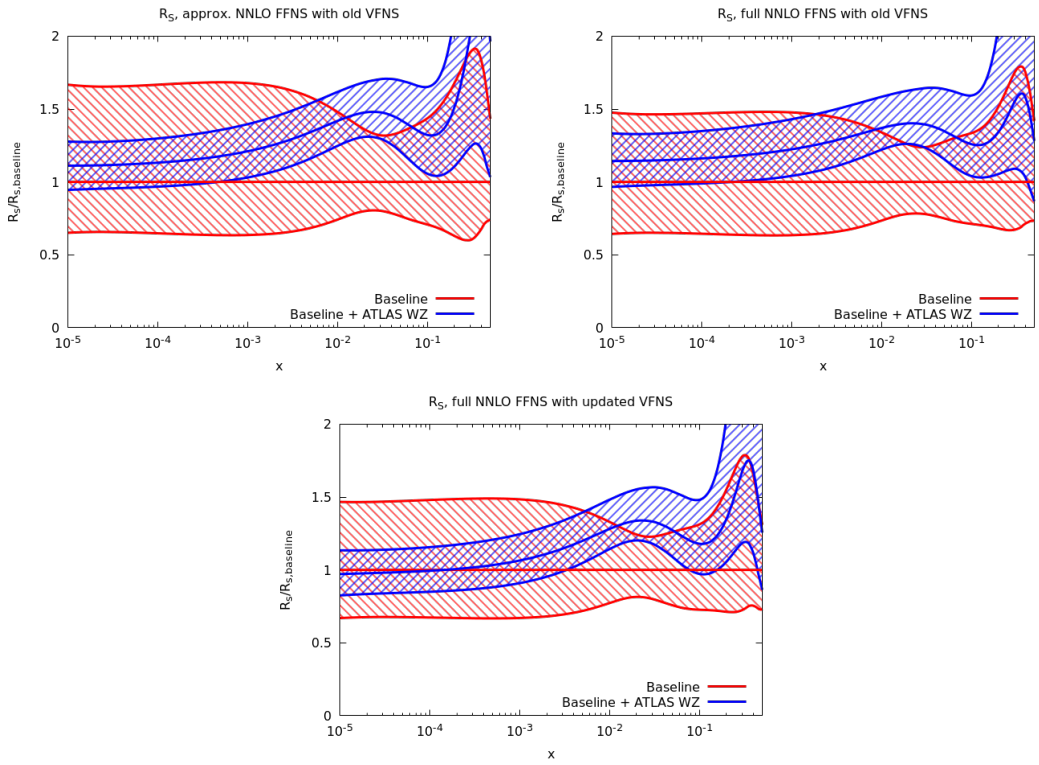


Figure 4.4: Fits including and excluding the ATLAS 7 TeV Drell–Yan data with various determinations of the theoretical prediction for dimuon production. Top left: approximate NNLO FFNS and old VFNS. Top right: Full NNLO FFNS, old VFNS. Bottom: Full NNLO FFNS, updated VFNS.

tension between the ATLAS and dimuon datasets at high x, with the inclusion of the ATLAS data not resulting in a central fit that is outside of the error bands of the baseline PDF set. However, examination of the χ^2 values in Table 4.1 show that the overall fit quality has deteriorated upon updating the FFNS calculation while including the ATLAS data. This is due to a potential tension between the NuTeV and ATLAS data being compounded with the deterioration of the NuTeV fit quality when updating the theory calculation. Further to this, we do not see improvement in the tension for intermediate x values around $10^{-2} - 10^{-1}$. Next, we make the same comparison, but including the updated VFNS calculation on top of the full NNLO FFNS determination as shown in Fig. 4.4 (bottom). Here we see that

there is less agreement between fits including the ATLAS data to those excluding it when compared to Fig. 4.4 (top right) at high x, although this is still within errors. However, we do see a mild improvement in the low and intermediate x regions, with the ATLAS data agreeing with the data included within the MMHT14+HERA fit at low x values. In contrast to the updated FFNS contribution, the updates to the VFNS result in a better fit quality for all three datasets considered here as shown in Table 4.1. This is particularly interesting for the NuTeV dataset, for which these updates deteriorated the fit quality when the ATLAS data was not included. Overall, the updates to both the FFNS and VFNS result in a slight degradation in fit quality when including the ATLAS data, with a mild shift towards agreeing on a value of R_s.

For comparison, we also include the χ^2 per point for each dataset as fit by the MSHT20 PDF set in Table 4.1. This set uses the theoretical updates described above and includes additional datasets not considered here, in particular the CMS 7 TeV $W + c$ data [118] and the ATLAS 8 TeV W/Z data [119, 120]. Overall this fit sees a reduction in the dimuon χ^2 when compared to the MMHT14 fit including the updated theoretical prediction and the ATLAS 7 TeV W/Z data, while the opposite is seen for the ATLAS data. This could in part be driven by the CMS $W + c$ data, which pulls the strange PDF in the same direction as the dimuon data, although only mildly. This happens despite the ATLAS 8 TeV W/Z data pulling the strange distribution up, similar to the effect of the same measurement at 7 TeV.

One important quantity to keep track of in the fits is the value of the branching ratio, $\mathrm{B}(c \to \mu)$. In the MMHT framework, this is given a value of 0.092, as determined by direct measurements [121], and has a 10% error by which it can vary. This is in slight disagreement with the value of 0.099 which is preferred by NuTeV [122]. In Table 4.2 we present the value of the branching ratio that is favoured by each

Dataset	B($c \to \mu$)
Baseline (approx FFNS + old VFNS)	0.090
Baseline (Full FFNS + old VFNS)	0.098
Baseline (Full FFNS + updated VFNS)	0.101
Baseline (approx FFNS + old VFNS) + ATLAS WZ	0.073
Baseline (Full FFNS + old VFNS) + ATLAS WZ	0.082
Baseline (Full FFNS + updated VFNS) + ATLAS WZ	0.086
MSHT20	0.089

Table 4.2: Best fit values for the $c \to \mu$ branching ratio for various fits with different theoretical determinations of the dimuon process and different datasets included.

fit. Here we see that inclusion of the updated FFNS and VFNS calculations results in increasing the preferred value of the branching ratio, more in line with what is determined by the NuTeV experiment. However, it can be seen that the ATLAS experiment disagrees with this value and reduces the determination of the branching ratio by 15%–20%. This overall results in a branching ratio that is much closer to the direct experimental value compared to a fit without the ATLAS dataset (but still including the updated theory calculation for the dimuon process) and is well within the 10% uncertainty. We also present the value found in the MSHT20 fit, where the addition of the CMS 7 TeV $W + c$ and ATLAS 8 TeV W/Z data pull the optimal branching ratio even closer to the value determined by direct measurement.

4.4 Summary

In this chapter we implemented recently calculated NNLO FFNS corrections for the dimuon production at CCFR/NuTeV with full charm mass dependence into the MMHT14 framework. Along with this, we corrected the VFNS calculation so that

this contribution is now determined more consistently. We found that the updated FFNS has minimal impact on the determination of R_s, while the updated VFNS has a more, although still mild impact, with an overall increase at low to mid x values and a decrease at high x. However, this results in an overall slight decrease in fit quality, mostly due to the NuTeV data. We also found a slightly better agreement for R_s between the dimuon data and the ATLAS Drell–Yan data at 7 TeV, with the FFNS corrections decreasing the apparent tension at high x while the corrected VFNS does the same at low to mid x. This improvement is only mild and again a deterioration in fit quality is still found, mostly due to the NuTeV dataset.

Chapter 5

Towards Ultimate Parton Distribution Functions at the HL–LHC and LHeC

In this chapter we quantify the impact of the high luminosity phase of the LHC (HL–LHC) on the PDFs though projected data using the Hessian profiling technique. We then describe the extension of this methodology to the proposed LHeC and combine both results to obtain a general picture of what one might expect the status of PDF fits to be in the future. The results shown in this section have been published in [1,2].

5.1 Introduction

The data already collected from the LHC has provided a wealth of information on a variety of a different particle physics topics and has enabled unique constraints to be imposed on the PDFs that would not have been possible with other experiments [16, 37, 52, 123] . Now, with the continued analysis of the Run II dataset from the LHC, many in the particle physics community are looking ahead to the future of this detector. In around 2023, a major upgrade of the LHC accelerator and detector systems will enable the beginning of its High–Luminosity (HL) operation phase. This will result in a ten–fold increase in its instantaneous luminosity and generate a

dataset with significantly more statistics than was previously possible, with the HL–LHC expected to deliver around $\mathcal{L} = 3$ ab^{-1} to ATLAS and CMS and $\mathcal{L} = 0.3$ ab^{-1} to LHCb. This unprecedented dataset will enable high precision data to be obtained and open new exciting physics opportunities, which have been summarised in the CERN Yellow Report [124]. These immense data samples will allow for a significant extension of the kinematic coverage of PDF–sensitive measurements in addition to a marked improvement in their statistical and systematic uncertainties. With this in mind, we attempt to quantify the impact of the future HL–LHC measurements on the uncertainty of the proton PDFs. The result of this can then be readily fed into other related projections for processes which will benefit from the reduction of PDF errors at the HL–LHC. Further, it has been proposed to collide electrons and protons within the LHC in a facility known as the Large Hadron Electron Collider (LHeC), which would run concurrently with the HL–LHC. This will provide a significantly larger and higher–energy dataset of lepton–proton collisions in comparison to the existing HERA structure function measurements [125]. Since the HERA measurements still to this day form a significant portion of the constraints on the PDFs, it is clear that the LHeC could enable impressive bounds to be imposed and enable our understanding and precision of the PDFs to improve significantly. In addition, lepton–proton collisions are in general cleaner than their proton–proton counterparts that will occur at the HL–LHC due to the reduction of QCD effects. Thus, despite the lower luminosity and centre of mass energy, it is expected that the LHeC will have a similar impact on the PDF uncertainty when compared to the HL–LHC. In this study we analyse the projected impact of the HL–LHC and LHeC on the PDF uncertainties using the Hessian Profiling method to perform an approximate fit. It should be noted that this type of study has already been performed on the LHeC projected data [126–134], however, this study represents the first time such analysis

has been undertaken on a proton–proton collider. In addition to this, we perform the analysis of both detectors in the same framework, enabling the datasets to be combined and the overall impact of both experiments to be estimated.

5.2 Fitting Methodology

Before explaining in detail how the projected dataset and the theoretical predictions are generated, we first describe the fitting methodology used, as this is the same for both experiments. We will be using the Hessian profiling formalism [135, 136] which can be used to estimate the constraints of a dataset on a baseline PDF set defined within the Hessian formalism as described in section 2.3.2. Here, one can take the following figure of merit:

$$\chi^2(\beta_{\rm th}) = \sum_{i,j=1}^{N_{\rm dat}} (\sigma_i^{\rm exp} - \sigma_i^{\rm th} + \sum_k^{N_{\rm eig}} \Gamma_{ik}^{\rm th}\,\beta_{k,\rm th})(\rm cov)_{ij}^{-1}(\sigma_j^{\rm exp} - \sigma_j^{\rm th} + \sum_m^{N_{\rm eig}} \Gamma_{jm}^{\rm th}\,\beta_{m,\rm th}) + T^2 \sum_k^{N_{\rm eig}} \beta_{k,\rm th}^2 \ , \tag{5.1}$$

where $\sigma_i^{\rm exp}$ are the data used, $\sigma_i^{\rm th}$ are the theoretical predictions made using the baseline PDF, $\rm cov_{ij}$ is the experimental covariance matrix, $\Gamma_{ik}^{\rm th}$ represent the PDF uncertainty on the measurement along the $k^{\rm th}$ eigenvector, $\beta_{k,\rm th}$ are a set of nuisance parameters, and T is the tolerance. $N_{\rm dat}$ and $N_{\rm eig}$ are the number of data points used in the fit and eigenvectors in the baseline PDF respectively. We have defined the experimental covariance matrix as follows:

$$(\rm cov)_{ij} = \delta_{ij}(\delta_{{\rm unc},i}^{\rm exp}\sigma_i^{\rm th})^2 + \Gamma_{i,\rm corr}^{\rm exp}\Gamma_{j,\rm corr}^{\rm exp} \ , \tag{5.2}$$

where $\delta^{\rm exp}_{{\rm unc},i}$ are the relative uncorrelated experimental uncertainties and $\Gamma^{\rm exp}_{i,{\rm corr}}$ are the correlated experimental uncertainties. We note that this figure of merit is resilient to the impact of D'Agostini bias [137, 138]. The first term of Eq. (5.1) encodes the χ^2 change from the newly added dataset. Here we allow the PDFs to vary by their uncertainty such that the theoretical prediction for the observed quantity can shift and allow a range of values to be acceptable. These shifts will clearly have an impact on the data that have already been included in the baseline PDF, which is taken into account by the second term, commonly referred to as a penalty term. The fit result can then be determined by the minimisation of Eq. (5.1) with respect to the nuisance parameters, which results in:

$$\beta^{\rm min}_{k,{\rm th}} = -\sum_l H^{-1}_{kl} a_l \ , \tag{5.3}$$

where we have defined

$$H_{kl} = \sum_{i,j} \sigma^{\rm th}_i \Gamma^{\rm th}_{ik} ({\rm cov})^{-1}_{ij} \sigma^{\rm th}_j \Gamma^{\rm th}_{jl} + T^2 \delta_{kl} \ , \tag{5.4}$$

$$a_k = \sum_{i,j} \sigma^{\rm th}_i \Gamma^{\rm th}_{ik} ({\rm cov})^{-1}_{ij} \left(\sigma^{\rm th}_j - \sigma^{\rm exp}_j\right) \ . \tag{5.5}$$

To obtain the uncertainties on the profiled PDFs, one can expand the χ^2 around this minimum:

$$\chi^2_{\rm profiled} = \chi^2_{\rm profiled}|_{\beta_{k,{\rm th}}=\beta^{\rm min}_{k,{\rm th}}} + \sum_{kl} \delta\beta_{k,{\rm th}} H_{kl} \delta\beta_{l,{\rm th}} \ , \tag{5.6}$$

where $\delta\beta_{k,{\rm th}} = \beta_{k,{\rm th}} - \beta^{\rm min}_{k,{\rm th}}$. Thus H_{kl} can be diagonalised to obtain the new eigenvector directions, $\hat{v}^{(k)}$, with which the updated error PDFs can be defined. We can

then move along these directions until we reach a point where $\Delta\chi^2 = T^2$:

$$\delta\beta_{l,\mathrm{th}}^{(k)} = T v_l^{(k)} \sqrt{\frac{1}{\hat{\epsilon}_k}} \,, \tag{5.7}$$

where $\hat{\epsilon}_k$ is the k^{th} eigenvalue of the Hessian matrix, H_{ij}. For this analysis, we select PDF4LHC15 [54] as a baseline, in particular we choose the Hessian set with 100 eigenvectors. This is the most recent combination of the global PDFs, representing the overall understanding of the PDFs at the time and has been recommend for use at the LHC. Since this PDF set is a combination of PDFs defined in the Hessian and replica methodologies, there is no precisely defined tolerance. We therefore take a uniform tolerance of $T = 3$, which is a rough average of the dynamically determined tolerance found by MMHT14 and CT14.

There is often discussion as to the choice of tolerance, with a choice of $T > 1$ typically taken such that data from a wide range of experiments can be described consistently. However, one could argue that, if the new data to be profiled is sufficiently clean, then the tolerance should be taken as unity, with anything more than this resulting in the projected impact of the new data potentially being hampered. This has been argued regarding the LHeC pseudo-dataset and so we now discuss the impact of the choice of tolerance on the profiled PDF. Despite the apparent linear dependence of the error on the tolerance in Eq. (5.7), the true relationship will depend on the interplay between the newly added datasets and those already included in the baseline PDF. This is due to the second term in the updated Hessian matrix, Eq. (5.4), also depending on this. When the data is unconstraining, this term dominates and the eigenvalues are given by $\epsilon_k \approx T$, resulting in uncertainties that are independent of the tolerance and equal to the baseline errors, as consistency would dictate. However, when the data are constraining, the errors of the profiled PDF

are linearly dependent on the tolerance. In reality, the relationship will lie somewhere in–between these extremes, and be different for each eigenvector, depending on which regions of the PDFs the data constrain. On the other hand, the tolerance is implicitly encoded into the baseline PDF and thus is in principle fixed. Changes to the tolerance would represent a reweighing of the data already present in this PDF by altering the second term of Eq. (5.1). Therefore, such changes can modify the apparent impact that the new dataset has, with a lower tolerance resulting in an artificial reduction of the profiled uncertainty. One could try to get around this by defining a new tolerance T' that applies to the profiled PDF and is different from that of the baseline PDF. However, including completely unconstrained data with this method will result in the uncertainty being scaled by a factor of T'/T and is thus inconsistent. One potential method to resolve this would be to apply a form of dynamic tolerance as is done in MSHT20 and CT18. This would allow for the baseline PDF to keep its defined tolerance while that of the profiled PDF will depend on the input dataset in a consistent way. However, doing this would require detailed knowledge of the datasets included in the baseline PDF, significantly reducing the simplicity of the Hessian Profiling method, which is its main benefit over a full fit. Further, we note that taking a constant tolerance of $T > 1$ can also help to account for possible tensions that could be present in the true HL–LHC and LHeC datasets but are lacking in the projections. Thus we simply take $T = 3$ throughout the calculation, but keep in mind the impact that this choice has on the results of the LHeC fits, where the profiled uncertainty could potentially be reduced if a lower or dynamical tolerance were taken.

5.3 HL–LHC

We first concentrate on the High–Luminosity phase of the LHC. Here, since at the time of this study there were no official data projections, we generate our own pseudo–data and fit to it. For this, we choose processes we believe to best demonstrate the constraining power the increase in luminosity will provide. The original analysis of this pseudo–data and additional details can be found in [1].

5.3.1 Data Selection

We begin by selecting processes that are sensitive to the PDFs and are expected to have significantly improved constraining power due to the larger luminosity that will be available at the HL–LHC. This could either be due to the process being rare, and thus are dominated by the statistical uncertainty, or those which will benefit from information in regions of phase space that previous experiments could not provide enough statistics. We choose a wide range of processes that place constraints on different regions of the PDFs such that an overall picture can be obtained. We do however focus on the mid to high x regions as this is where the HL–LHC is expected to perform better. This is not strictly necessary as processes such as inclusive D meson production can play a role at low x [139], however this particular process is unlikely to benefit from the HL–LHC due to the already high statistics. Nonetheless, there may be other rare processes that could be important for this region of phase space. For each process, a reference dataset was chosen such that, as will be described in more detail later, a reasonable assessment of the uncertainties can be obtained. Where relevant, we also describe the binning used, which is often extended beyond the reference dataset to account for the greater kinematical reach of the HL–LHC.

The particular processes chosen for this analysis are:

- High–mass Drell–Yan, specifically the dilepton invariant mass differential distributions $d\sigma(pp \to ll)/dm_{ll}$ for $m_{ll} \gtrsim 110$ GeV for a central rapidity acceptance, $|\eta_l| \leq 2.4$. This process is particularly useful for quark flavour separation, specifically to constrain the poorly known large x sea quarks.

  Here the ATLAS 8 TeV measurement of differential Drell–Yan cross sections [140] is taken as reference, with additional bins in the high m_{ll} region included to benefit from the enhanced kinematic coverage.

- The differential distributions for on–peak W and Z boson production in the forward region, $2.0 \leq \eta_l\,(\eta_{ll}) \leq 4.5$, covered by the LHCb experiment. These measurements constrain quark flavour separation, including the strange and charm content of the proton, in the large and small x region [141] complementary to the data from the central region.

  The reference analysis is the LHCb measurement of the rapidity distributions of W and Z bosons in the muon final state at 8 TeV [142]. As compared to the reference measurement, a finer binning by a factor of 2 to 5 has been adopted as allowed by the increased event rates. Events are selected if $p_T^l \geq 20$ GeV, the lepton rapidities fall in the LHCb acceptance, and, in the case of Z production, there is the additional requirement that 60 GeV $\leq m_{ll} \leq$ 120 GeV.

- Differential distributions in top–quark pair production, providing direct information on the large x gluon [143]. Specifically, we consider the top–quark transverse momentum p_T^t and rapidity y_t, as well as for the top–quark pair rapidity $y_{t\bar{t}}$ and invariant mass $m_{t\bar{t}}$.

  The reference measurements here are the ATLAS 8 TeV differential distri-

butions in the lepton+jets final state [144]. We assume that the statistical correlations between different distributions will be available, as is the case for the 8 TeV data [38], and therefore include the four distributions simultaneously in the fit. To account for the increased statistics of the HL–LHC, the number of bins in the rapidity distributions is doubled, while the p_T^t and $m_{t\bar{t}}$ distributions are extended to higher values in the TeV region.

- The transverse momentum distribution of the Z bosons in the dilepton final state, $20\,\text{GeV} < p_T^{ll} < 3.5\,\text{TeV}$ region for central rapidities $|\eta_Z| \leq 2.4$ and different bins of the dilepton invariant mass m_{ll}. This process is relevant to constrain the gluon and the anti–quarks at intermediate values of x [82].

  As the reference analysis, we take the ATLAS measurements of the transverse momentum of lepton pairs at 8 TeV [145]. The pseudo–data is generated for six different bins of the dilepton invariant mass m_{ll}, with boundaries 12, 20, 30, 40, 66, 116, and 150 GeV respectively. In each of the invariant mass m_{ll} bins, additional bins are added to the p_T^{ll} distribution to exploit the improved coverage of the large transverse momentum region.

- The production of W bosons in association with charm quarks. This process provides a sensitive handle on the strangeness content of the proton [146, 147], which is the least well known of the light quark PDFs. The pseudo–data for this process has been generated as function of the lepton rapidity η_l from the W boson decay, and is inclusive over the kinematics of the charm quark provided it satisfies the selection cuts.

  For this process, pseudo–data has been generated both for the central rapidity region relevant for ATLAS and CMS as we as for the forward region covered by LHCb. In the central rapidity region, $|\eta^l| \leq 2.4$, the reference measurement

is the CMS analysis at 13 TeV [148], where events are selected provided that $p_T^c \geq 5$ GeV and $p_T^l \geq 26$ GeV with l indicating the result of the $W \to l\nu$ decay. At forward rapidities, $2 \leq \eta^l \leq 5$, we use a dedicated selection strategy with $2.2 \leq \eta^c \leq 4.2$, $p_T^\mu \geq 20$ GeV, $p_T^c \geq 20$ GeV, and $p_T^{\mu+c} \geq 20$ GeV [149]. We take the acceptance to be 30% due to c–jet tagging and an overall normalisation error of 5%.

- Prompt isolated photon production represents a complementary probe of the gluon PDF at intermediate values of x [150]. Here the pseudo–data has been generated as differential distributions in the photon transverse momentum p_T^γ for different bins in the photon rapidity η^γ.

  The reference measurements here is the ATLAS 13 TeV analysis of [151], where additional bins have been added to the p_T^γ distribution in each rapidity bins to benefit from the improved coverage of the large p_T^γ region.

- Finally, we consider the inclusive production of hadronic jets in different bins of their rapidity up to $|y_{\rm jet}| \leq 3$ as a function of their $p_T^{\rm jet}$. This process provides direct information on the gluon and the valence quarks at large x. Here the jets are reconstructed using the anti–k_T algorithm with $R = 0.4$ as radius parameter.

  The reference measurement here is the 13 TeV ATLAS analysis of inclusive jets and dijet cross sections based on a luminosity $\mathcal{L} = 3.2$ fb^{-1} from the 2015 data–taking period. As compared to this baseline, the coverage of the high–p_T region has been extended, making possible to access transverse momenta up to $p_T^{\rm jet} \simeq 2 - 3$ TeV.

The LO Feynman diagrams for these processes are shown in Fig. 5.1. We emphasise that this list is not exhaustive and other processes could also be considered. For

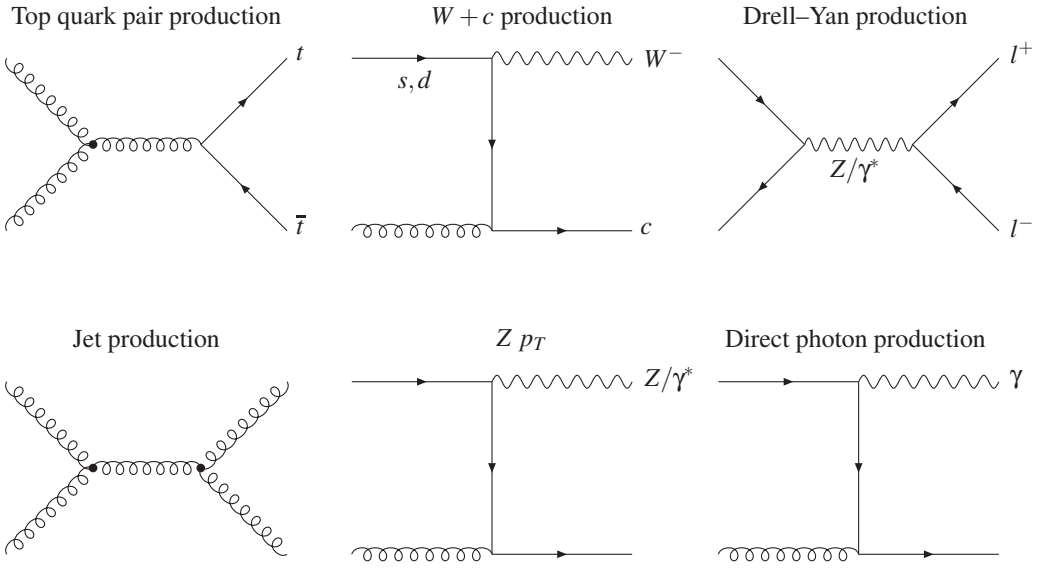


Figure 5.1: Representative Feynman diagrams at the Born level of the six types of collider processes for which HL–LHC pseudo–data has been generated in this analysis.

example, one could use dijet production [152] to provide information on the gluon, or further constrain the quark flavour separation by including single top production [153]. Further, one may expect that developments in experimental and theoretical methods may result in new novel processes being added to the PDF fitting toolkit and enable further constraints on the PDF uncertainty.

5.3.2 Theory calculations and pseudo–data generation

Theoretical predictions for the above processes have been calculated to NLO using `MCFM` [154] interfaced to `APPLgrid` [33] to generate the relevant fast interpolation grids. The exception to this is inclusive jet production, for which the NLO calculation is obtained using `NLOJET++` [155]. Since this study relies on pseudo–data, higher–order QCD effects or electroweak corrections are not needed. This is because this analysis is a case of a PDF closure test [156], and we are only interested in the reduction of uncertainties, with the central value essentially unchanged.

To generate the pseudo–data, we take the theoretical prediction and vary according to some error that we estimate. This gives the central value to be:

$$\sigma_i^{\rm exp} = \sigma_i^{\rm th}(1 + r_i \cdot \delta_{{\rm tot},i}^{\rm exp} + \lambda \cdot \delta_{\mathcal{L}}^{\rm exp} + s \cdot \delta_{\mathcal{N}}^{\rm exp}) \,, \tag{5.8}$$

where r_i, λ, and s are univariate Gaussian random numbers, $\sigma_i^{\rm th}$ is the theoretical prediction for this bin, $\delta_{{\rm tot},i}^{\rm exp}$ is the total (relative) experimental uncertainty corresponding to this specific bin (excluding the luminosity and normalization uncertainties), and $\delta_{\mathcal{L}}^{\rm exp}$ is the luminosity uncertainty, which is fully correlated among all the pseudo–data bins of the same experiment (but uncorrelated among different experiments). We take this luminosity uncertainty to be $\delta_{\mathcal{L}}^{\rm exp} = 1.5\%$ for the three LHC experiments. $\delta_{\mathcal{N}}^{\rm exp}$ are possible additional normalization uncertainties as in the case of forward W boson production in association with charm quarks.

With the exception of forward W + charm, which will be detailed later, we estimate the experimental uncertainty by looking to previous datasets on the same process. We then split these uncertainties into two types: statistical and systematic. We evaluate the statistical uncertainty as:

$$\delta_{{\rm stat},i}^{\rm exp} = (f_{\rm acc} \times N_{{\rm ev},i})^{-1/2} \,, \tag{5.9}$$

where $N_{{\rm ev},i}$ is the predicted number of events in bin i and $f_{\rm acc}$ is the acceptance correction that accounts for detector effects and/or branching ratios reducing the measured number of events. The acceptance correction is extracted from reference dataset by comparing the statistical error on this measurement to the expected number of events, in a similar fashion to the above equation.

For systematic errors, we treat all of them as uncorrelated except for the luminos-

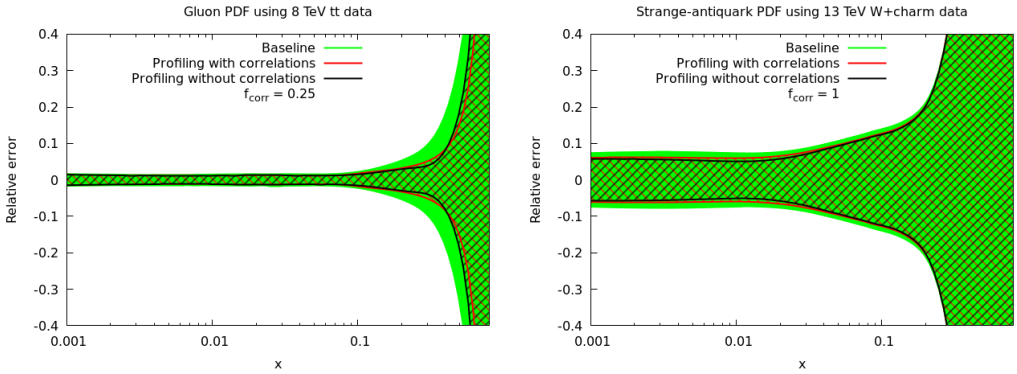


Figure 5.2: Comparison between the baseline PDF4LHC15 set and the profiled sets with LHC data, either with or without the correlations between the experimental systematic uncertainties accounted for. In the latter case, the $f_{\rm corr}$ factor is chosen to reproduce the results of the profiling when the correlations are included, see text. We show here the results of profiling with the top differential distributions at $\sqrt{s} = 8$ TeV with $f_{\rm corr} = 0.25$ (left) and the W+charm rapidity distribution at $\sqrt{s} = 13$ TeV with $f_{\rm corr} = 1$ (right plot). A tolerance factor of $T = 1$ has been used for this specific comparison.

ity error which is treated separately as in Eq. (5.8). This choice is motivated by the fact that it is challenging to even estimate the amount by which the systematics will reduce, let alone their correlations which is a delicate topic even in already measured datasets. We thus add the systematic errors from the reference data in quadrature. However, in doing so, we may artificially reduce the constraining power of the data and overestimate the uncertainty. We thus multiply the systematic error by a factor of $f_{\rm corr}$ to account for the impact that this decorrelation. To estimate this value, we take two of the reference datasets and and profile them both twice: first with the full correlations, and then with the systematics added in quadrature but reduced by the factor $f_{\rm corr}$. We then vary this factor and compare the resulting profiled PDFs by eye until they are roughly the same. For this comparison, we take a tolerance of $T = 1$ to exaggerate the impact of the dataset.

We performed this analysis with the differential top quark pair production at $\sqrt{s} = 8$ TeV, for which a value of $f_{\rm corr} \sim 0.25$ was found, and the central $W+$ charm

rapidity distribution at $\sqrt{s} = 13$ TeV, for which a value of $f_{\text{corr}} = 1$ was preferred. The results of these profiling exercises can be seen in Fig. 5.2. It is therefore clear that the precise value for this correction is heavily dependent on the process and reference dataset chosen and thus may bring into question this approach. However, we have explicitly selected datasets that will benefit from an improvement in the statistical uncertainty and not systematics. Therefore, as we will show later, the specific value of this parameter should not have a large impact on the overall results of this study. We therefore simply take $f_{\text{corr}} = 0.5$, which represents a rough average of the tested datasets but weighted more towards the top quark differential data as this is expected to have a larger impact on the PDFs and thus is more important to describe accurately

Due to the much larger dataset that will be measured at the HL–LHC, it is expected that there will be better calibration of the systematic errors. Further, the detector improvements made during the course of the shut–down will further reduce these uncertainties. To take into account these effects, we include an additional factor of $f_{\text{red}} \leq 1$ which also multiplies the systematic uncertainties. This value is varied between 1 (0.5) and 0.4 (0.2) to represent conservative and optimistic scenarios for reference data taken at 8 TeV (13 TeV). These datasets are treated differently due to the fact that the 13 TeV reference data are based on a much smaller dataset and is therefore expected to have larger systematic errors than the 8 TeV case and are likely to be reduced as more data is taken.

Overall, the relative experimental uncertainty is given by:

$$\delta^{\text{exp}}_{\text{tot},i} = \sqrt{\left(\delta^{\text{exp}}_{\text{stat},i}\right)^2 + \left(f_{\text{corr}} f_{\text{red}} \delta^{\text{exp,ref}}_{\text{sys},i}\right)^2} \tag{5.10}$$

where $\delta^{\text{exp,ref}}_{\text{sys},i}$ is the total systematic error for bin i from the reference dataset. Due

to the similar effects of $f_{\rm corr}$ and $f_{\rm red}$, we will often write these as $F = f_{\rm red} f_{\rm corr}$.

Since we have chosen some datasets that will benefit from the extended kinematic range, we have chosen to extend the binning in variables such as transverse momentum and invariant mass. We thus choose the binning to roughly match those of the reference data and proceed logarithmically, with each bin containing at least 30 predicted events to ensure that the errors remain approximately Gaussian. For the rapidity distributions, finer binning has been used to account for the increased statistics in most datasets. The variables $f_{\rm acc}$ and $\delta^{\rm exp,ref}_{{\rm sys},i}$ are taken from the closest bin and are extrapolated beyond the reference dataset's kinematic range if necessary.

The exception to all of the above is the forward W+charm pseudo–dataset to be measured at the LHCb. No measurement for this process currently exists at the time of this study, and thus we follow recommendations from experimentalists [149] and take $f_{acc} = 0.3$ due to c–jet tagging and $\delta^{\rm exp}_{\mathcal{N}} = 0.05$ as an overall normalisation error.

A summary of the overall pseudo–dataset used for this study is presented in Table 5.1. Here, we also indicate the reference datasets that have been used to generate the uncertainties. A total number of $N_{\rm data} = 768$ pseudo–data points are used in this study.

5.3.3 Fits to individual datasets

We now present results for fits to individual pseudo–datasets to assess their potential constraining power and demonstrate how the HL–LHC can pin down the PDFs. Here, we concentrate on the top quark pair production and the W+charm processes, with details of the other pseudo–datasets found in the original study [1]. For this section, we will be taking the most optimistic scenario, with $f_{\rm red} = 0.4$ (0.2) for

Process	Kinematics	$N_{\rm dat}$	$f_{\rm corr}$	$f_{\rm red}$	Baseline
Z p_T	$20\,\text{GeV} \leq p_T^{ll} \leq 3.5$ TeV $12\,\text{GeV} \leq m_{ll} \leq 150$ GeV $\lvert y_{ll}\rvert \leq 2.4$	338	0.5	$(0.4, 1)$	[145] (8 TeV)
High–Mass Drell–Yan	$p_T^{l1(2)} \geq 40(30)\,\text{GeV}$ $\lvert\eta^l\rvert \leq 2.5$, $m_{ll} \geq 116\,\text{GeV}$	32	0.5	$(0.4, 1)$	[140] (8 TeV)
Top Quark Pair	$m_{t\bar{t}} \simeq 5$ TeV, $\lvert y_t\rvert \leq 2.5$	52	0.5	$(0.4, 1)$	[144] (8 TeV)
W+charm (central)	$p_T^{\mu} \geq 26\,\text{GeV}$, $p_T^{c} \geq 5\,\text{GeV}$ $\lvert\eta^{\mu}\rvert \leq 2.4$	12	0.5	$(0.2, 0.5)$	[148] (13 TeV)
W+charm (forward)	$p_T^{\mu} \geq 20\,\text{GeV}$, $p_T^{c} \geq 20\,\text{GeV}$ $p_T^{\mu+c} \geq 20\,\text{GeV}$ $2 \leq \eta^{\mu} \leq 5$, $2.2 \leq \eta^{c} \leq 4.2$	12	0.5	$(0.4, 1)$	LHCb projection
Direct photon	$E_T^{\gamma} \lesssim 3$ TeV, $\lvert\eta_{\gamma}\rvert \leq 2.5$	118	0.5	$(0.2, 0.5)$	[151] (13 TeV)
Forward W, Z	$p_T^{l} \geq 20\,\text{GeV}$, $2.0 \leq \eta^{l} \leq 4.5$ $2.0 \leq y_{ll} \leq 4.5$ $60\,\text{GeV} \leq m_{ll} \leq 120\,\text{GeV}$	90	0.5	$(0.4, 1)$	[142] (8 TeV)
Inclusive jets	$\lvert y\rvert \leq 3$, $R = 0.4$	58	0.5	$(0.2, 0.5)$	[157] (13 TeV)
Total		712			

Table 5.1: Summary of the features of the HL–LHC pseudo-data generated for the present study. For each process we indicate the kinematic coverage, the number of pseudo–data points used across all detectors $N_{\rm dat}$, the values of the correction factors $f_{\rm corr}$ and $f_{\rm red}$; and finally the reference from the 8 TeV or 13 TeV measurement used as baseline to define the binning and the systematic uncertainties of the HL–LHC pseudo–data, as discussed in the text.

processes with reference datasets measured at 8 TeV (13 TeV).

We start with the differential top quark pair production. This analysis is based on the ATLAS measurement at 8 TeV though the lepton + jets channel, where differential distributions of the transverse momenta and rapidity of a top quark (p_T, y_t) as well as the invariant mass and rapidity of the $t\bar{t}$ pair (m_{tt}, y_{tt}) are provided with full correlations. The reference data records bins in m_{tt} up to 1600 GeV which we extend to $\sim$ 6000 GeV due to the higher statistics that will be available at the HL–LHC allowing for these to be measured. We also generate finer binning for both rapidity distributions, going from 5 bins each to 12, with bins spread evenly across the same range.

It is expected that this process will constrain the high x gluon which can be demonstrated by plotting the correlation coefficients which, for this particular case, are defined as:

$$\rho = \frac{1}{\sigma_{\mathcal{O}}\sigma_f}\sum_{i}^{N_{\text{eig}}}(\mathcal{O}(S_i^+) - \mathcal{O}(S_i^-))(f(S_i^+) - f(S_i^-)), \tag{5.11}$$

where $\mathcal{O}$ is the observable to be measured, f is the PDF, and $S_i^{\pm}$ are the error PDFs along eigenvector direction i. Here $\sigma_{\mathcal{O}}$ and σ_f are defined as in Eq. (2.32). This coefficient is plotted in Fig. 5.3, where it can be clearly seen that the m_{tt} pseudo–data are highly correlated with the gluon PDF at high x and indicates that this distribution can place strong constraints in this region.

In Fig. 5.4 (left) we show the impact of profiling the baseline PDF with the m_{tt} distribution on the PDF uncertainties. At low m_{tt}, the pseudo–data has little impact on the PDF due to the fact that the reference dataset is already systematic dominated and thus is not expected to significantly improve at the HL–LHC. However, as we move past the region measured by the reference study, we see that the impact of this

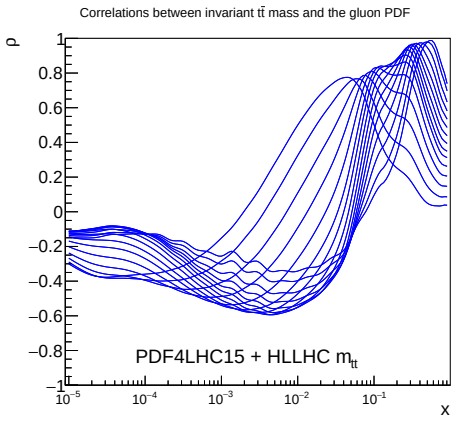


Figure 5.3: The correlation coefficient between the gluon PDF and the various bins of $m_{t\bar{t}}$, the invariant mass distribution of top quark pairs in $t\bar{t}$ production.

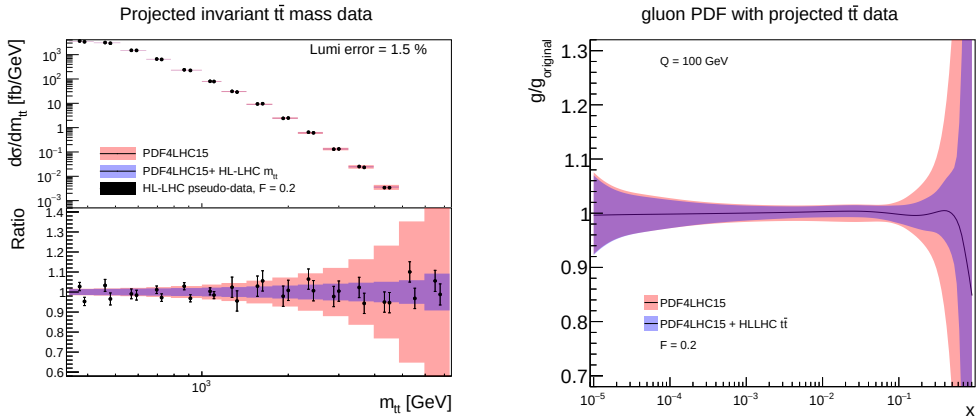


Figure 5.4: Left: Fits to the $m_{t\bar{t}}$ distribution in top quark pair production pseudo–data. Right: Impact on the gluon PDF after including the HL–LHC $t\bar{t}$ pseudo–data (all distributions) in the fit.

pseudo–dataset significantly improves. This demonstrates the HL–LHC's ability to constrain this extended kinematic region and reduce the PDF uncertainty using datasets that are already systematic dominated.

In Fig. 5.4 (right), we plot the overall impact this pseudo–dataset has on the gluon when combining all four distributions. Here we see, as expected, a significant improvement can be seen at high x, demonstrating the good constraining power of this type of measurement.

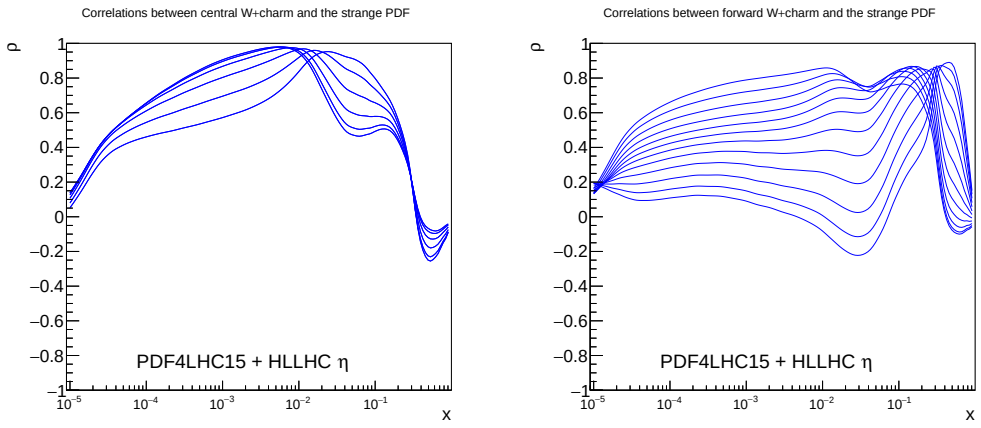


Figure 5.5: The correlation coefficient between the strange PDF and the lepton rapidity distributions in W+charm production pseudo–data in the central rapidity region (left) and in the forward region (right).

We now move on to the W+charm process, for which we separately consider the central rapidity region, which can be measured by ATLAS and CMS, and the forward region, to be measured by LHCb. For the former, we use the CMS analysis at 13 TeV [148] as reference, while for the latter we use a dedicated selection strategy described above. This process places bounds on the relatively unknown strange quark which is only constrained by a handful of processes and, as discussed in chapter 4, some of these are potentially in tension with each other. Thus having precise data from the HL–LHC will aid in analysing this possible tension. To demonstrate the impact this process has on the strange PDF, we again plot the correlation coefficient relating these quantities in Fig. 5.5. Here we see that there is strong correlations across a large range of x values in the central rapidity region, thus allowing this process to place constraints on the strange quark over a broad range of kinematics. Further, we see that the forward W+charm process can allow for the determination of the high x strange quark, which is particularly poorly constrained.

The constraints that this pseudo–data has on the PDF uncertainties when predicting the same process are shown in Fig. 5.6. Here we see that central region

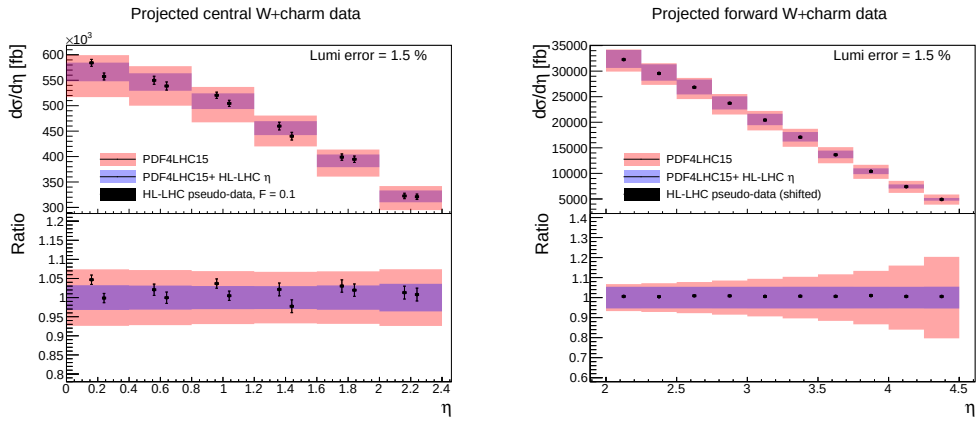


Figure 5.6: Fits to the W+charm production pseudo–data in the central (left) and forward (right) rapidity regions. In the right plot, only the statistical errors are shown, while the data have been shifted by the dominant correlated source of uncertainty, namely the 5% normalization uncertainty.

experiences a significant improvement over the whole range of rapidities due to the increased statistics while the forward region constrains the very high rapidity bins that have not yet been measured. Further, we note that the error bars shown for the forward case comprise entirely of the statistical uncertainty and are significantly smaller than the PDF error. This suggests that the normalization uncertainty dominates here, with the HL–LHC thus providing a systematic dominated dataset with the statistical errors being negligible.

The impact this has on the strange quark is shown in Fig.5.7. Here we see, as expected, the central region has significant impact over a wide range of x values, while the forward region has much better constraining power at high x.

5.3.4 Ultimate PDFs at the HL–LHC

Finally, we apply the entire HL–LHC pseudo–dataset to the PDF4LHC15 PDF set. Here, we use two different values of f_{red} as described earlier. We choose scenario A (conservative), where $f_{\text{red}} = 1$ (0.5) and scenario C (optimistic), where $f_{\text{red}} = 0.4$

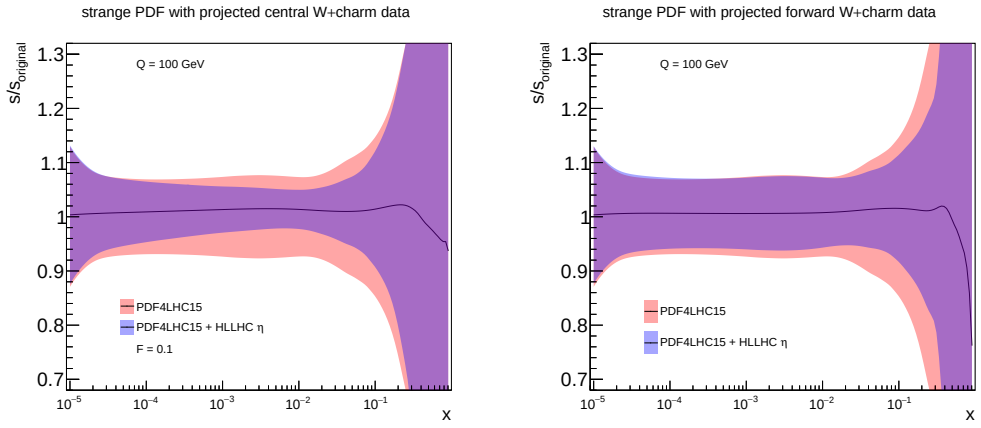


Figure 5.7: Impact on the strangeness quark PDF when including the HL–LHC pseudo–data on W+charm production in the central (left) and in the forward region (right).

(0.2) for reference dataset at 8 (13) TeV. We present the results of this profiling in Fig. 5.8, were we show the impact on different PDFs. Here we observe substantial reduction of the PDF uncertainties in all cases over a wide range of x. This is particularly striking for the strange quark, which is reduced by almost a factor of 4, from 15% to just a few percent. This is mostly due to the W+charm measurements, particularly the forward region which has a unique handle on this PDF and demonstrates the importance of this process when fitting PDFs. We also see a good reduction in the gluon PDF for a wide range of x. This is due to the numerous datasets that constrain different regions of x for this PDF, including the differential top quark pair production and direct photon.

An important feature of these plots is the similarity between the two scenarios. This is by design; we have chosen dataset that will benefit from the improved statistics that will be available at the HL–LHC. Thus, the precise treatment of the systematics has a much smaller impact on the constraining power when compared to the improved statistics. This also demonstrates our analysis is robust in regards to how we treat experimental systematic errors.

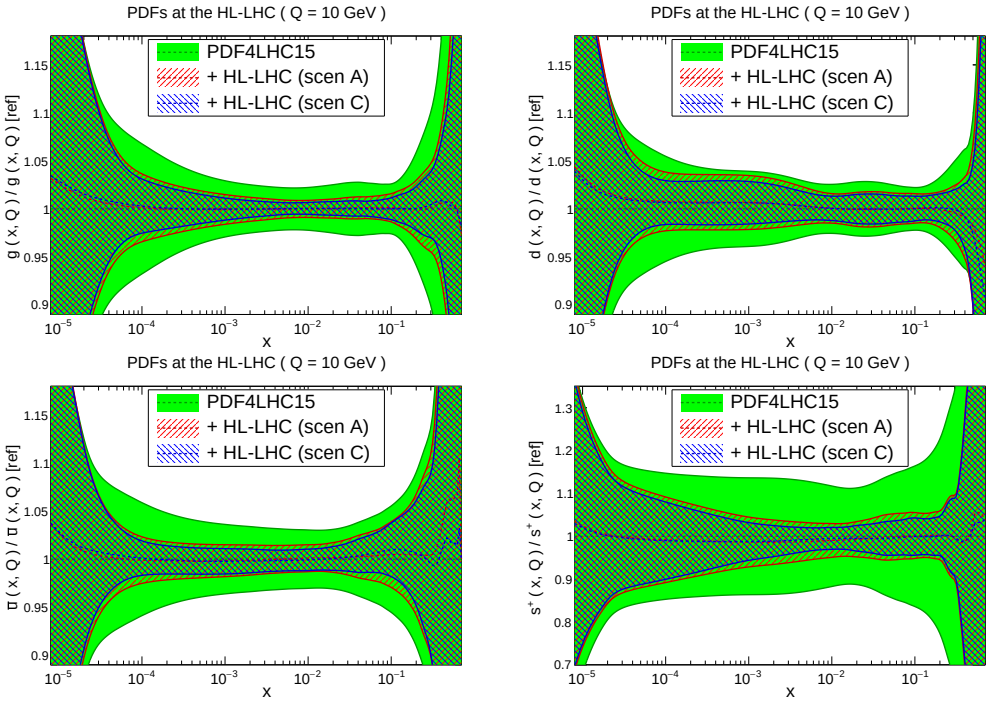


Figure 5.8: Comparison of the PDF4LHC15 set with the profiled sets with HL–LHC pseudo–data in scenarios A and C, defined in the text. We show the gluon, down quark, up anti–quark, and total strangeness at $Q = 10$ GeV, normalized to the central value of the PDF4LHC15 baseline.

5.3.5 Implications for HL–LHC phenomenology

The PDFs profiled with the HL–LHC pseudo–data have been published in the `LHAPDF6` [36] format, allowing for easy use to make phenomenological predictions of Standard Model and Beyond the Standard Model quantities at the HL–LHC. Some of these analyses were performed in the original study, for which we present some examples.

First, we discuss the PDF uncertainties on the inclusive Higgs production via gluon fusion, which we plot as a function of the minimum jet transverse momentum in Fig. 5.9 (left). These calculations have been performed in `MCFM8.2` [154] using the LO matrix elements. Here we see that the constraints the HL–LHC places on the

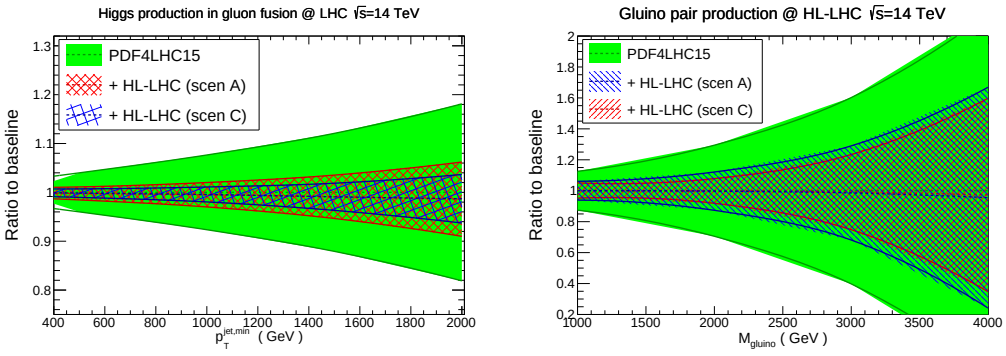


Figure 5.9: Projected PDF uncertainties for Higgs production in gluon fusion (left) and gluino pair production (right) at the HL–LHC. Scenarios A and C are defined in section 5.3.4

PDFs result in a significant reduction of PDF errors, down to the $\sim 2\%$ level over the kinematical region relevant for the HL–LHC. This particular process is sensitive to corrections from heavy particle loops in the high Higgs transverse momentum region and is thus susceptible to certain BSM models such as partial Higgs compositeness.

Next, we describe an example of a BSM calculation that can be performed with these PDFs: gluino pair production. This calculation has been performed with `Pythia8.325` [158] to LO and is presented in Fig. 5.9 (right). Here we see that there is a significant reduction in the PDF uncertainty of a factor of $2-3$ over a large range of potential gluino masses.

It is interesting to note that for both phenomenological studies, the PDF uncertainties are only mildly affected by the choice of scenario. This further confirms the observation that the exact treatment of the systematics does not have a large impact on the overall results.

5.4 LHeC

We now move onto the proposed Large Hadron–electron Collider. Previous studies have been performed to estimate the constraints that this detector has on the PDFs [126–128, 134]. Here, we present our own study in the context of Hessian Profiling using the same pseudo–dataset. The original study can be found in [2].

5.4.1 Pseudo–datasets

For the profiling of the LHeC dataset, we use the most recently available LHeC projection for both neutral–current (NC) and charged–current (CC) scattering. This pseudo–dataset analyses the collisions between electrons/positrons of energy 60 GeV with protons of energy 1 and 7 TeV. While the $E_p = 7$ TeV pseudo–datasets provide the dominant constraining power, the lower energy pseudo–data can extend this constraining power to higher x values. Further, by taking a range of energies, one can place strong constraints on quantities highly dependent on the variation of these scales, such as F_L, which is important for pinning down the gluon PDF. The pseudo–dataset also includes information on the heavy quark structure functions. Of these, the charm production through charge current in $e^- p$ scattering provides an interesting constraint on the anti–strange PDF via the $\bar{s} + W^+ \to \bar{c}$ vertex. We apply a cut of $Q \geq 2$ GeV to ensure that the impact of higher twist corrections can be neglected. We also neglect the polarization of the lepton beam as it is known that the impact of this on the PDF determination is small, particularly in the context of a closure test. The integrated luminosity is taken to be $\mathcal{L} = 1$ ab^{-1} for electron–proton collisions with $E_p = 7$ TeV, which is roughly three orders of magnitude more than that at HERA. For all other collisions we take $\mathcal{L} = 0.1$ ab^{-1}. We note that

most of the pseudo–data points are dominated by the systematic uncertainty, and thus, the overall results do not depend strongly on the choice of luminosity, with somewhat lower values of this resulting in similar conclusions. A detailed summary of the complete pseudo–dataset is given in Table 5.2. We note that additional processes at the LHeC could be considered, for example, jet production, which can place good constraints on the high x gluon, that would otherwise be difficult to measure in lepton–hadron collisions.

The theoretical predictions for the LHeC structure functions have been evaluated at NLO using `APFEL` [160] with the `FONLL--B` General Mass–Variable Flavour Number Scheme [161]. As is the case for the HL–LHC predictions, we ignore NNLO and electroweak corrections as these are not expected to have significant impact on a closure test. However, it is worth noting that, due to this pseudo–dataset often constraining the PDFs to the per–mille level, other contributions to the error such as theoretical uncertainties, for which only recently have studies of these effects been done [66, 162], could have a substantial impact on the PDF errors. Variation of the strong coupling constant and heavy quark masses is also expected to have an impact at this level of precision [163].

In contrast to the HL–LHC pseudo–dataset, where we determined the uncertainty using reference datasets, we instead use a dedicated LHeC projection that contains a detailed correlation model. We thus modify the method to generate the pseudo–data point for bin i:

$$\sigma_i^{\text{exp}} = \sigma_i^{\text{th}} \left(1 + \delta_{\text{unc},i}^{\text{exp}} \cdot r_i + \sum_k \Gamma_{ik}^{\text{exp}} s_k \right) , \tag{5.12}$$

where $\delta_{\text{unc},i}^{\text{exp}}$ and Γ_{ik}^{exp} are the uncorrelated and correlated errors respectively. s_k are univariate Gaussian random numbers that allow for a correlated variation. This should be compared to Eq. (5.8), where the other variables are also defined.

Observable	E_p	Kinematics	$N_{\rm dat}$	$\mathcal{L}_{\rm int}$ [ab^{-1}]
$\tilde{\sigma}^{\rm NC}$ (e^-p)	7 TeV	$5\times10^{-6}\leq x\leq 0.8$, $5\leq Q^2\leq 10^6$ GeV^2	150	1.0
$\tilde{\sigma}^{\rm CC}$ (e^-p)	7 TeV	$8.5\times10^{-5}\leq x\leq 0.8$, $10^2\leq Q^2\leq 10^6$ GeV^2	114	1.0
$\tilde{\sigma}^{\rm NC}$ (e^+p)	7 TeV	$5\times10^{-6}\leq x\leq 0.8$, $5\leq Q^2\leq 5\times10^5$ GeV^2	148	0.1
$\tilde{\sigma}^{\rm CC}$ (e^+p)	7 TeV	$8.5\times10^{-5}\leq x\leq 0.7$, $10^2\leq Q^2\leq 5\times10^5$ GeV^2	109	0.1
$\tilde{\sigma}^{\rm NC}$ (e^-p)	1 TeV	$5\times10^{-5}\leq x\leq 0.8$, $2.2\leq Q^2\leq 10^5$ GeV^2	128	0.1
$\tilde{\sigma}^{\rm CC}$ (e^-p)	1 TeV	$5\times10^{-4}\leq x\leq 0.8$, $10^2\leq Q^2\leq 10^5$ GeV^2	94	0.1
$F_2^{c,\rm NC}$ (e^-p)	7 TeV	$7\times10^{-6}\leq x\leq 0.3$, $4\leq Q^2\leq 2\times10^5$ GeV^2	111	1.0
$F_2^{b,\rm NC}$ (e^-p)	7 TeV	$3\times10^{-5}\leq x\leq 0.3$, $32\leq Q^2\leq 2\times10^5$ GeV^2	77	1.0
$F_2^{c,\rm CC}$ (e^-p)	7 TeV	$10^{-4}\leq x\leq 0.25$, $10^2\leq Q^2\leq 10^5$ GeV^2	14	1.0
Total			945	

Table 5.2: Overview of the main features of the LHeC pseudo–data [159] included in our PDF projections. For each process, we indicate the kinematic coverage, the integrated luminosity, the proton energy, and the number of pseudo–data points, $N_{\rm dat}$, after the $Q\geq 2$ GeV kinematic cut. Note that in all cases the incoming lepton energy is fixed to be $E_l = 60$ GeV. We ignore the effect of the incoming lepton beam polarization.

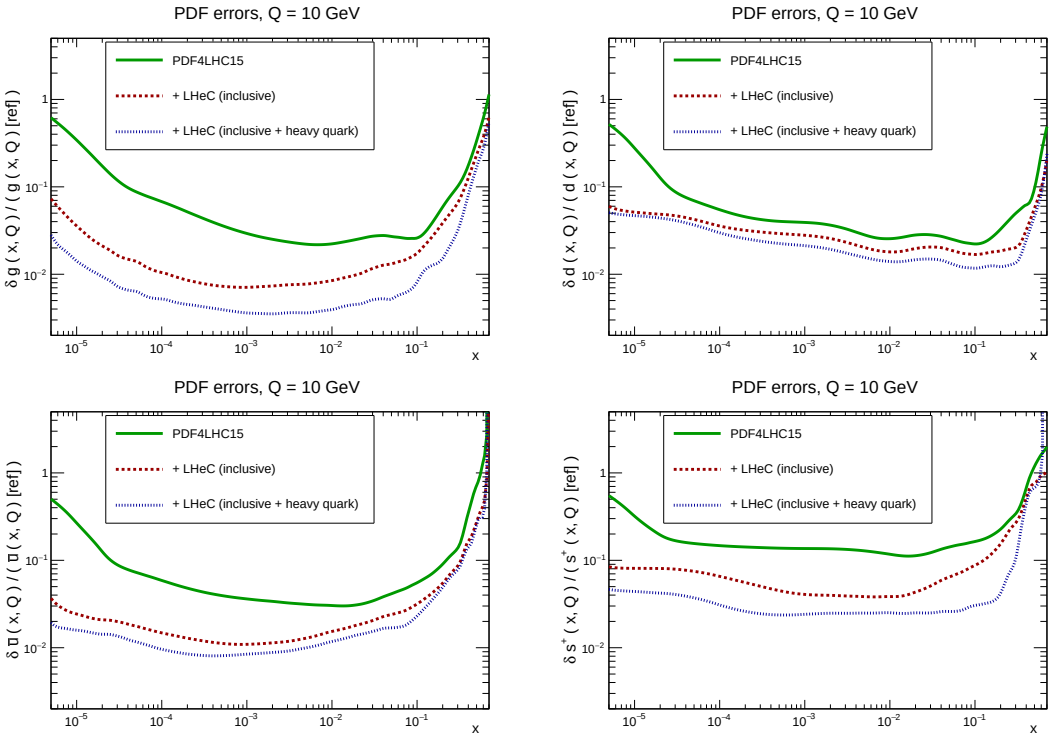


Figure 5.10: Impact of LHeC on the 1–σ relative PDF uncertainties of the gluon, down quark, anti–up quark and strangeness distributions, with respect to the PDF4LHC15 baseline set. In this comparison, shown at $Q = 10$ GeV, we indicate the results of both profiling with the inclusive LHeC measurements alone and also with the semi–inclusive heavy quark structure functions.

5.4.2 Impact of the LHeC pseudo–dataset

We now present the impact of profiling the LHeC pseudo–data using the PDF4LHC15 PDF set as a baseline. This is shown in Fig. 5.10, where we demonstrate the significant constraining power of both the inclusive dataset alone and in combination with the semi–inclusive heavy quark pseudo–data. We see that, while the inclusive data are able to place good constraints on the gluon in the low to medium x regions through the direct measurements of their gradient with respect to Q^2, this is further improved at by the heavy quark data which also extends this constraining power to high x. For the down quark (and similarly for the up quark, which is not shown),

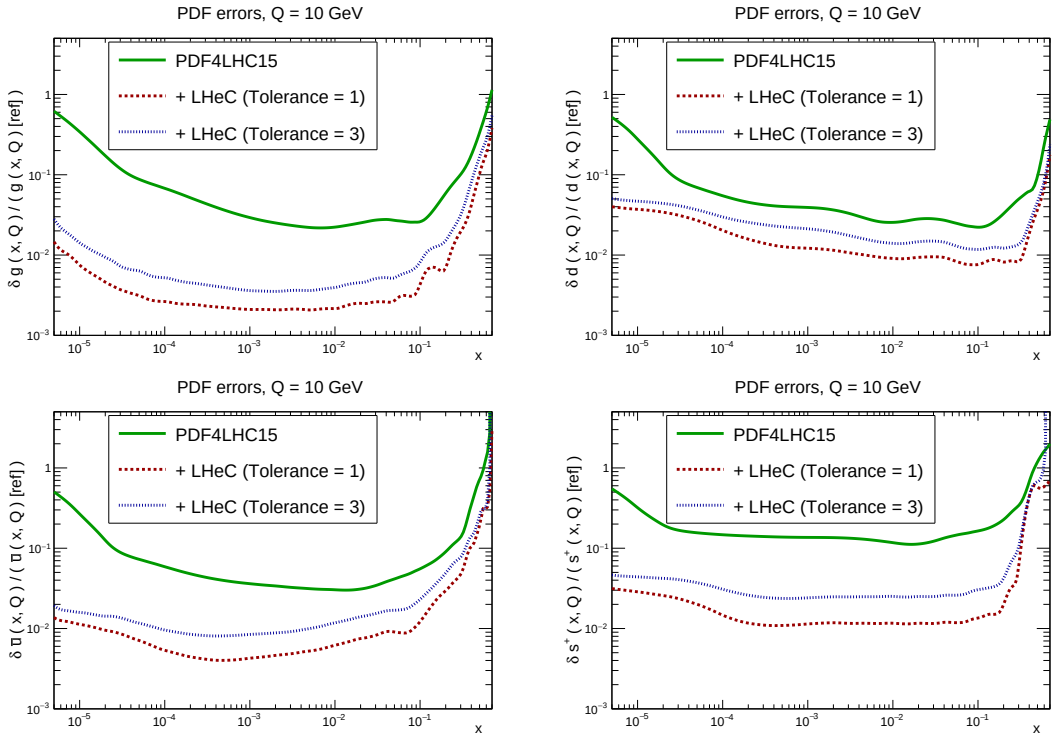


Figure 5.11: Same as Fig. 5.10 for the complete LHeC dataset, now comparing the baseline results obtained with a tolerance $T = 3$ with those based on $T = 1$.

the pseudo–data enables some improvement in the uncertainty over a wide range of x. This is particularly significant at high x, where the large reductions on the valence quark uncertainty dominates. We note that, for the first–generation quarks, the heavy quark data is less important, although not negligible. Finally, the inclusive data places impressive constraints on the strangeness, which is further improved by the addition of the heavy quark data which can also push these constraints to the high x region.

In light of the discussion given in section 5.2 and of pre–existing LHeC PDF projection [126–128, 134], we examine the impact of choosing $T = 1$ rather than $T = 3$ as is default. This is shown in Fig. 5.11, where we see that this change results in up to a factor of 2 reduction in the PDF uncertainty. We also see that this effect

is reduced when the impact of the LHeC pseudo–data is smaller, as was expected. If a dynamical tolerance was taken, one might expect that the true result would lie somewhere in–between the $T = 1$ and $T = 3$ cases, although the precise location would intimately depend on the data already included within the PDF4LHC15 PDF set, information which is not available. However, it is clear that the $T = 1$ case misrepresents the data within the baseline PDF and thus is difficult to justify. It is also worth noting that, given that this is based on pseudo–data, there are many additional factors, such as tension and additional processes, that could alter the results here by a similar magnitude compared to the tolerance. We thus take the $T = 3$ case as default, since this best describes the PDF4LHC15 baseline and can account for the tension that could be present at the HL–LHC and LHeC, with the $T = 1$ case presented as a demonstration. However, we note that with the true LHeC dataset, a dynamic tolerance would be used when fitting with a global Hessian PDF, which may give a slightly different picture.

5.5 HL–LHC and LHeC Combined Pseudo–dataset

Finally, we compare the impact the HL–LHC and the LHeC pseudo–datasets have on the PDF uncertainties and combine both projections to form a complete picture of the impact both detectors will have in the determination of the PDFs. We show the results of this profiling in Fig. 5.12, where we see that the LHeC has the most dominate constraints in the low x region for all PDFs. This is not unexpected, the LHeC projections provides excellent coverage of this kinematical region, while the HL–LHC pseudo–dataset focused more on the mid to high x regions. For intermediate values of x, we see that both detectors have comparable constraints on the PDFs. However, at high x values we see that the HL–LHC provides somewhat

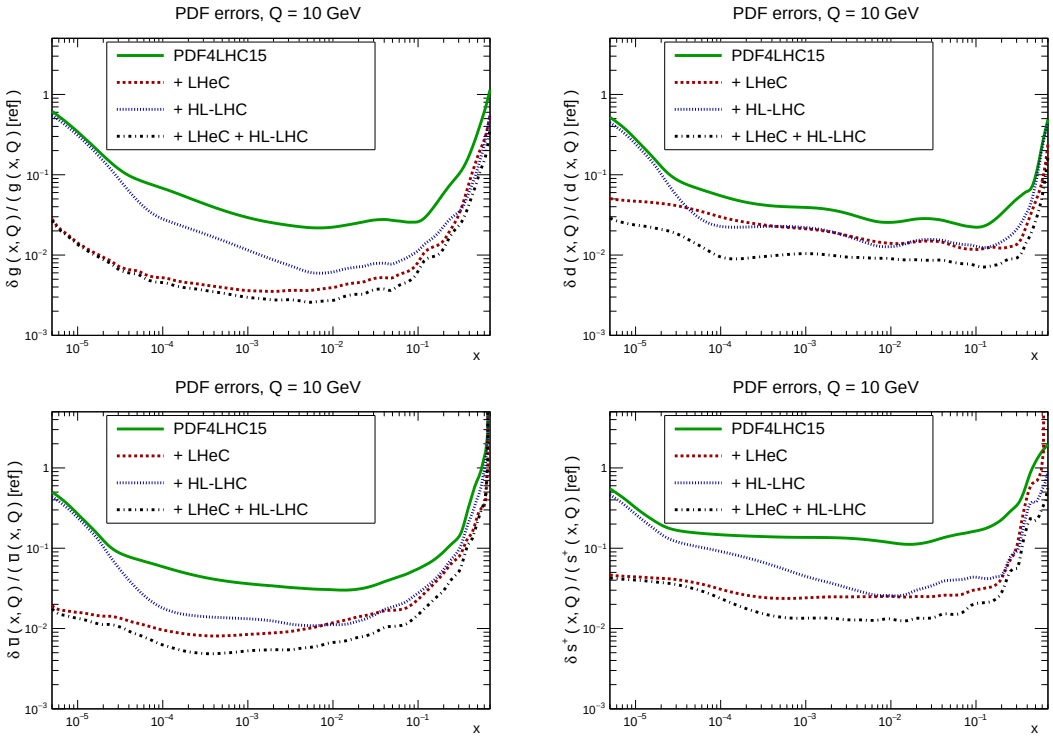


Figure 5.12: Fits to the combined HL–LHC and LHeC pseudo–datasets.

stronger constraints on the gluon and strange PDFs while the LHeC pseudo–data has more impact on the first generation quarks. Overall we see that the HL–LHC and LHeC detectors are complementary in their PDF constraining power, placing bounds on different regions of the PDFs. When combined, these enable a significant reduction in the PDF uncertainty over a very broad range of x.

We note that this analysis is rather dependent on the processes chosen for the projections. For the HL–LHC pseudo–data, we chose processes that will impact the mid to high x PDFs and thus it is unsurprising that the LHeC dominates in the low x region, where the projections contain a significant number of pseudo–data points. However, additional processes could be considered for HL–LHC that will impact the low x region, such a D meson production. One could also do the same for the LHeC pseudo–data, such as include jet production which would impact the high x gluon.

5.6 Summary

In this chapter we have quantified the expected impact the HL–LHC and LHeC detectors will have on the PDF uncertainties. We have additionally generated a combined result, demonstrating the complimentary nature of both detectors and presenting an overall picture of what one could expect the PDFs to look like once the full dataset of these detectors has been collected. We show that these experiments will place significant constraints on the PDF uncertainty, reducing it to the per–mille level over an extremely broad range of x. Such constraints will enable new kinematic regions of the PDFs to be analysed and enable the study novel QCD phenomena such a BFKL dynamics and saturation effects at low x. Furthermore, this dramatic error reduction will propagate to observables within the SM and BSM models, enabling extremely precise determination of such quantities and potentially lead to the discovery of physics beyond the SM.

It should be noted that this analysis has a number of caveats that may alter these results compared to how the true datasets will impact the PDFs. Firstly, we have ignored potential tensions that may exist within a particular dataset or between two different datasets. Examples of this were shown in chapters 3 and 4. These effects will in general increase the overall uncertainties compared to this study. The chosen tolerance of $T = 3$ may help to describe such possible tension, however, the precise details of their effects will not be known until the data are available. In contrast, we could have considered additional processes for the generation of the pseudo–datasets, allowing for more data to be used and further constraints to be placed on the PDFs. This could also include novel processes that are not currently included within the PDF fitting toolbox but could be discovered before the data from these experiments become available. Additional effects from the uncertainty of

the strong coupling constant and the mass of heavy quarks can also have impacts on the PDF uncertainties at this level of precision and will need to be taken into account, increasing the overall PDF error.

The PDFs generated in this study have been made publicly available in the `LHAPDF` format [36] for use in further projections at these detectors. Details of these can be found in the original studies [1, 2].

Chapter 6

WW Production via Photon Interactions: The Structure Function Approach

In this chapter we discuss the production of WW pairs at the LHC through the interaction of two photons. We analyse this in terms of the new structure function approach, assessing the impact of off–shell photons and how these can be included in a consistent manner through a careful choice of gauge.

6.1 Introduction

Since the discovery of the Higgs boson at the LHC in 2012 [164], a wealth of data have been collected to analyse its physical properties [165,166]. With electroweak symmetry breaking providing an intimate relationship between the Higgs and electroweak sectors, analysis of the photon, W and Z bosons and their interactions has become a high–interest topic. The Standard Model precisely predicts these couplings, the masses of the heavy bosons and their relationship. Thus experimental determination of these quantities is important to examine if any of them vary from their predicted value, with a number of Beyond the Standard Model theories anticipating these to be anomalous [167]. Particular interest is on the triple and quartic couplings between

multiple bosons, with these being difficult to measure through hadron collisions due to the dominant QCD interactions. However, with EW corrections now being included in the calculation of many processes, such determinations are becoming viable. For this to be done, accurate calculations of the photon–initiated diagrams are required, as these include the potentially anomalous couplings even at leading order. These contributions have typically been calculated through the photon PDF, with the LUXqed formalism [61, 168] being the standard. Here, the photon PDF is determined directly through the structure functions and has been implemented within global fits such as MMHT2015qed [169] and NNPDF31luxQED [58]. While the photon PDF can be determined very precisely to around a few percent, the overall error on photon–induced processes is high at leading order due to scale variation.

One such process is the production of WW pairs through photon interactions, which is sensitive to the γWW and $\gamma\gamma WW$ couplings. Evidence for this process was found in the ATLAS 8 TeV data [170] and the CMS combined 7 TeV and 8 TeV datasets [171], and has been recently confirmed and measured within the ATLAS 13 TeV dataset [172]. In addition to the standard inclusive events, this process can occur as semi–exclusive and exclusive events, where one or both of the protons remain intact after the interaction. Such events can be measured using dedicated proton tagging detectors or through the observation of rapidity gaps [173]. While the photon PDF may apply well for inclusive processes, for these exclusive events a more direct approach may be more precise, calculating the relevant observables straight from the structure function data without the need to define a photon PDF in some cases [20]. This has already been applied in the context of lepton production [64] and has been shown to generate smaller errors than the equivalent photon PDF calculation for some regions of phase space.

In this chapter, we calculate the cross section for the production of WW pairs

at the LHC using this structure function approach and carefully analyse the impact that gauge dependency has on the result, particularly relating to the longitudinal modes of the W boson. We show that the axial gauge performs better than the more common unitary gauge and compare this to the photon PDF approach.

6.2 The Structure Function Approach

The structure function approach has been a central element in the calculation of Higgs production via vector boson fusion at the LHC [174–176]. Recently, this formalism has been generalised for photon–initiated processes [20], and we now apply this to the production of W boson pairs. Here the cross section is calculated according to:

$$\sigma_{pp} = \frac{1}{2s}\int \frac{\mathrm{d}^3p_1\mathrm{d}^3p_2\mathrm{d}\Gamma}{E_1E_2}\alpha(Q_1^1)\alpha(Q_2^2)\frac{\rho_1^{\mu\mu'}\rho_2^{\nu\nu'}\mathcal{M}^*_{\mu'\nu'}\mathcal{M}_{\mu\nu}}{Q_1^2Q_2^2}\delta^{(4)}(q_1+q_2-k), \tag{6.1}$$

where q_i are the momenta of the photons with $Q_i^2 = -q_1^2$, p_i are the momenta of the outgoing hadronic systems, and $k = \sum_i k_i$ is the total momentum of the particles generated in the $\gamma\gamma$ sub–process. Further, $\sqrt{s}$ is the centre of mass energy, $\mathrm{d}\Gamma$ is the phase space measure for the $\gamma\gamma$ sub–process and $\mathcal{M}_{\mu\nu}$ is its amplitude. Finally, $\rho_i^{\mu\mu'}$ are the density matrices for a virtual photon given by:

$$\begin{aligned}\rho_i^{\mu\nu} = 2\int \frac{\mathrm{d}M_i^2}{Q_i^2}\Bigg[&-\left(g^{\mu\nu}+\frac{q_i^\mu q_i^\mu}{Q_i^2}\right)F_1(x_i,Q_i^2)\\ &+\frac{(2xp_i^\mu - q_i^\mu)(2xp_i^\nu - q_i^\nu)}{2Q_i^2}F_2(x,Q_i^2)\Bigg],\end{aligned} \tag{6.2}$$

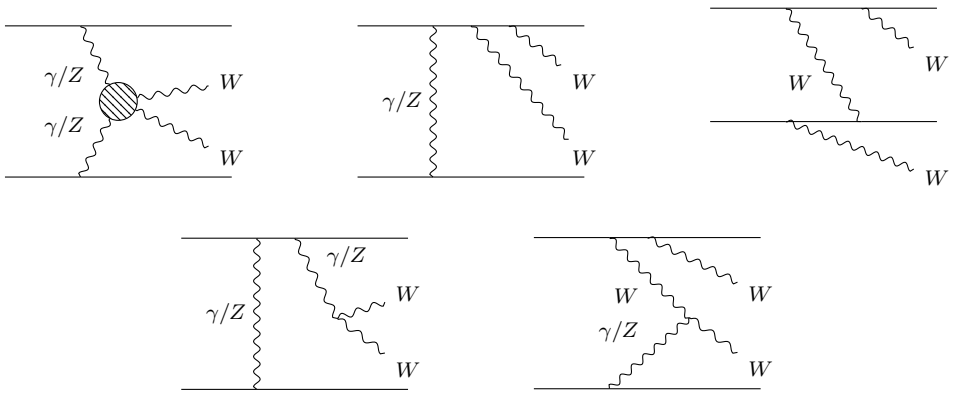


Figure 6.1: Representative Feynman diagrams for the $qq \to qW^+W^-q$ process. Only the top right digram represents structure function type processes, while all others are non–factorizable. The fermion mass is taken to be zero, otherwise additional diagrams would be present. Further, additional diagrams that can contribute if the initial fermions are anti–particles of each other are not considered here.

where M_i are the masses of the outgoing hadronic systems. Comparisons can be made to the hadron tensor $W^{\mu\nu}$ in Eq. (2.8), specifically:

$$\rho_i^{\mu\nu} = 2 \int \frac{\mathrm{d}M_i^2}{Q_i^2} W_i^{\mu\nu}. \tag{6.3}$$

All uncertainties that are included in the structure function approach are also present in the photon PDF calculation, with the most obvious example being the experimental uncertainties on the structure functions themselves. These typically result in an error of the order of a few percent for the structure function approach [20]. A further uncertainty introduced in both methods is the missing non–factorizable diagrams, shown in Fig. 6.1, that cannot be written in the above form. These diagrams are expected to be negligible in the limit $Q_i^2 \ll M_{WW}^2$, where the Equivalent Photon Approximation (EPA) applies. We also ignore, for now, factorizable Z–initiated diagrams which can be assumed to have minimal impact for $Q_i^2 \ll M_Z^2$, which is only a slightly stronger constraint than that for the non–factorizable processes for most of the relevant phase space. Therefore, we expect the Z–initiated diagrams to become

important at similar values of Q^2 compared to the non–factorizable contributions. At leading order, these limits are already imposed in the photon PDF formalism, with the EPA used in its derivation resulting in an on–shell photon at leading order. However, in the case of the structure function approach, such limits are not assumed, and care must be taken to ensure that these are valid. Further, the photon is not taken to be on–shell, even at leading order, and the exclusion of non–factorizable and Z–initiated diagrams means that the observable calculated is not truly physical and does not need to be gauge invariant. Thus the gauge that we choose will impact the overall result and care must be taken to avoid gauges where the structure function diagrams have a substantial contribution to the cross section outside of the above limits. We will show that, if this is the case, then the neglected diagrams can have a significant impact on the overall cross section.

6.3 The Unitary Gauge

The first gauge that we implement is the unitary gauge, which is commonly used to calculate LO processes due to its simplicity. Here we choose the gauge fixing term:

$$\begin{aligned}\mathcal{L}_{\text{gf}}^{\text{unitary}} = \lim_{\xi\to\infty}\Bigg[&-\frac{1}{2\xi}(\partial_\mu A^\mu)^2 - \frac{1}{2\xi}(\partial_\mu Z^\mu - \xi M_Z \varphi_Z)^2 \\ &-\frac{1}{\xi}(\partial^\mu W_\mu^+ - i\xi M_W \varphi_W^+)(\partial^\mu W_\mu^- + i\xi M_W \varphi_W^-)\Bigg],\end{aligned} \tag{6.4}$$

where $\varphi_W^\pm$ and φ_Z are the Goldstone bosons resulting from electroweak symmetry breaking. It should be noted that the mixing between these and the heavy bosons cancels the same contributions in the Higgs kinetic term. Further, by taking the $\xi \to \infty$ limit, we remove these Goldstone bosons from the problem entirely.

It is commonly known that processes involving longitudinally polarized massive

bosons can contain significant cancellations between diagrams. This can easily be seen by examining the form of the longitudinal polarization for a boson of mass m and momentum of $k = (E_k, \vec{k})$ in the unitary gauge:

$$\varepsilon_L^\mu(k) = \frac{1}{m}(|\vec{k}|, E_k\hat{k}). \tag{6.5}$$

In the high–energy limit we have $\varepsilon_L^\mu(k) \sim E_k/m$. Thus individual diagrams can grow unbounded as the energy of the heavy boson increases; however, the combination of all diagrams does not suffer from this behaviour due to gauge invariance. We note that this does not impact the photon PDF determination, with the sub–process $\gamma\gamma \to WW$ involving two real photons and thus this matrix element describes a physical process and must be gauge invariant. However, this will impact the structure function approach as the photons here are off–shell. To assess the impact of these polarizations on this methodology, we calculate the amplitude for the process $qq \to qWWq$ both using all diagrams (see Fig. 6.1), and with only those that contain the $\gamma\gamma \to WW$ sub–process. We take both quarks to be right–handed for simplicity as there are fewer non–factorizable diagrams due to the lack of coupling between the W boson and right–handed fermions. We further take the quarks to be massless such that any interactions with the Higgs can be ignored. For normalization, we simply multiply by the leading order contribution to the inelastic F_2 structure function (i.e. the charged weighted sum of the quark PDFs), calculated using the MMHT2015qed_nnlo PDF set [169]. Note that this normalization is rather arbitrary, as we are only interested in the relative difference of the calculations. We take the centre of mass energy to be 13 GeV, relevant for the LHC, and apply cuts rapidity of the WW pair, y_{WW}, such that $|y_{WW}| < 2.5$. We also apply the same cut to the rapidity of the lepton in the W boson's decay product as well as ensure that its transverse momentum is greater

than 10 GeV. We additionally apply cuts to the photon virtuality, $Q_i^2 > 1$ GeV^2, such that we avoid the collinear divergence in the $q \to q\gamma$ splitting and focus on the regions where we expect non–factorizable diagrams to contribute. To quantify the impact of the missing diagrams, we define a new variable:

$$\kappa_Z = \frac{Q_1^2 Q_2^2}{M_Z^4}, \tag{6.6}$$

which parametrizes the impact of the missing Z–initiated diagrams. We could also include a similar variable to quantify the impact of non–factorizable contributions, however, as discussed in the previous section, we expect that these will become relevant at a similar scale to the Z–initiated diagrams contribution. Thus, either variable will suffice, and we choose κ_Z as we expect the Z–initiated processes to become significant at slightly lower values of Q^2 compared to the non–factorizable diagrams. The dependence of the cross section on the invariant mass of the WW pair, M_{WW}, is shown in Fig. 6.2 (left), where we see that the structure function approach predicts a significantly larger cross section compared to the full calculation by several orders of magnitude. In particular, we see that the cross section is still significant even for $M_{WW} \sim 1$ TeV, whereas the complete calculation predicts a negligible cross section in this region. In Fig. 6.2 (right), we see that the structure function calculation and the full determination agree well in the $\kappa_Z \ll 1$ limit as expected, however, as κ_Z increases, we see a significant difference between these two approaches, with the structure function approach giving a drastically larger cross section within this region, directly corresponding to the poorly behaved invariant mass distribution. It is within this region we expect the Z–initiated and non–factorizable contributions to have a non–negligible impact, as is clearly the case here, with these interfering destructively with the photon–initiated diagrams in the full calculation, resulting in

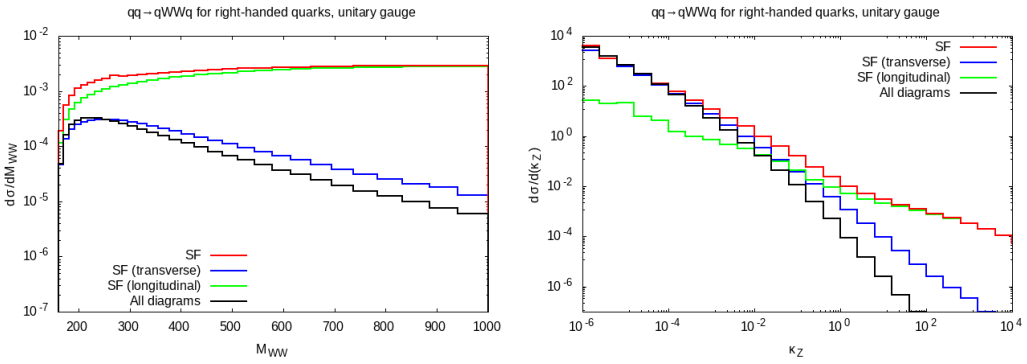


Figure 6.2: Cross section for the $qq \to qWWq$ process with right–handed fermions as calculated by the structure function approach using the unitary gauge and when including all diagrams. For the structure function calculation, we also show contributions from transverse and longitudinal W bosons.

significantly reduced cross section compared to when these are ignored. To further demonstrate the origin of this cancellation, we present contributions to the structure function approach resulting from purely transverse and longitudinal W bosons in the same plot. As expected from Eq. (6.5), we see that the longitudinal polarized W bosons are responsible for nearly the entire cross section in the $\kappa_Z \gg 1$ region that is causing the bad behaviour. We further see that the transverse structure function contributions distinctly lack these properties, providing a cross section that is slightly larger, but of the same order, as the full calculation summed over all polarizations. However, the transverse contributions should be smaller than this total sum; therefore, there still exists some cancellations between photon–initiated diagrams and non–factorizable processes, with this effect being milder for these polarizations compared to the longitudinal case.

To remedy this, we attempt to include further diagrams into the calculation, the simplest of which are the Z–initiated structure function type diagrams, excluding those that involve the Higgs boson. We note that, for interactions involving only right–handed fermions, the situation is even simpler, with the overall contribution

from the couplings in Z–initiated diagrams being identical to that of the photon case up to a sign, leaving the Z/γ propagators being the only difference between these. Further, the only additional diagrams that are present with right–handed fermions are the structure function diagrams involving the Higgs and those involving the $\gamma \to WW$ splitting (see bottom left diagram of Fig. 6.1). These contain additional factors of $1/M_{WW}^2$, resulting in significantly different kinematics compared to the other diagrams. It is therefore expected that there is unlikely to be significant cancellation between the two sets of diagrams. We thus define a new scheme, unitary$'$, which includes the Z–initiated diagrams (excluding those involving the Higgs), with a matrix element that can be used in Eq. (6.1) given by:

$$\mathcal{M}^{\mu\nu}_{\text{unitary}'} = \left(1 - \frac{Q_1^2}{Q_1^2 + M_Z^2}\right)\left(1 - \frac{Q_2^2}{Q_2^2 + M_Z^2}\right)\mathcal{M}^{\mu\nu}_{\text{unitary}}. \tag{6.7}$$

Here we note that this contains explicit factors of M_Z^2/Q^2 in the $Q^2 \gg M_Z^2$ limit, resulting in this region being suppressed. The cross section calculated in this scheme is presented in Fig. 6.3, where we again compare to the calculation involving all diagrams. Here we see that the difference between the structure function determination and the full cancellation is minimal, with the invariant mass distribution being near identical. The only significant difference is in the $\kappa_Z > 1$ region, where we expect the Higgs and non–factorizable diagrams to be relevant. We also see that the longitudinal W bosons provide a substantially smaller contribution to the cross section when compared to the same calculated in the standard unitary gauge without the Z–initiated diagrams (see Fig. 6.2). We do however see that this polarization does provide the largest contribution to the cross section in the $\kappa_Z > 1$ region, due to the same enhancement seen before. However, this is now suppressed by the pre–factor in Eq. (6.7), ensuring that this does not have a significant contribution to the overall

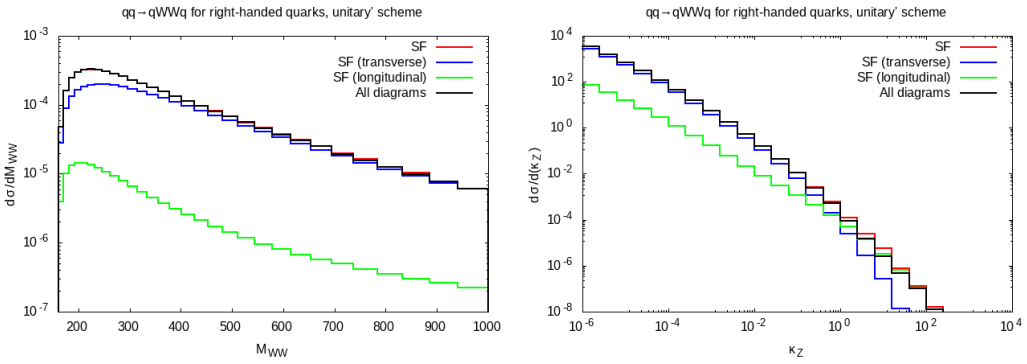


Figure 6.3: Same as Fig. 6.2 except using the unitary′ scheme for the structure function calculation.

cross section.

Overall we see that the inclusion of Z–initiated diagrams allow for a sensible description of the $qq \to qWWq$ cross section when using the structure function approach with the unitary′ scheme, with this agreeing well with the full calculation. However, this has been done purely in the case of right–handed quarks, with the couplings in the Z–initiated processes being different from the equivalent photon diagrams and additional non–factorizable diagrams present in the left–handed case. Therefore the unitary′ scheme is not expected to accurately describe this process for quarks of this helicity.

6.4 The Axial Gauge

We now consider the axial gauge [177, 178], defined by the gauge fixing term:

$$\mathcal{L}_{\text{gf}}^{\text{axial}} = \lim_{\lambda\to\infty}\left[-\frac{1}{2}\lambda n^{\mu}A_{\mu}^{a}A_{\nu}^{a}n^{\nu} - \frac{1}{2}\lambda(n.B)^{2}\right], \tag{6.8}$$

where A_{μ}^{a} and B_{μ} are the gauge fields for the $SU(2)$ and $U(1)$ symmetries respectively, that will later mix to form the photon and W/Z bosons through electroweak

symmetry breaking. In this theory, the Goldstone bosons need to be taken into account, and particular care is needed to deal with mixing terms as these are not cancelled by other contributions as they are in the unitary gauge. Here we diagonalise the Lagrangian as in [178], although one can write these mixing terms as explicit Feynman rules as was done in [179]. In the above equation, n is any constant 4–vector so long as $p.n \neq 0$ where p is the momentum of any W, Z or photon, real or virtual, in the problem. The question of gauge dependency now turns into asking what is the best choice of n that minimises the cancellations between the diagrams. It is clear that a space–like value is ill–advised, as we can easily run into a situation where we can have $p.n = 0$ for the real W bosons in the final state for some kinematical regions. For other choices of n, it is best to examine terms where the n dependence can explicitly appear. First of all is the W boson propagator:

$$\Delta_W^{\mu\nu}(k) = \frac{-i}{k^2 - m_W^2}\left(g^{\mu\nu} - \frac{n^\mu k^\nu + n^\nu k^\mu}{n.k} + k^\mu k^\nu \frac{n^2}{(n.k)^2}\right), \tag{6.9}$$

where m_W is the mass of the W boson. This propagator appears in both t–channel and u–channel diagrams. It is clear that this is divergent as $n.k \to 0$ and will dominate the amplitude in regions where this dot product is almost zero. This is not ideal since these terms cancel if the full calculation was performed, however, since that is not done in the structure function approach, we desire for these terms to be small to reduce the impact of these large cancellations. To examine this term further, we consider the impact of the kinematics on this, writing:

$$n.k = n.p_p - n.p'_p - n.p_W, \tag{6.10}$$

where p_p, p'_p and p_W are the momenta of an incoming proton, the outgoing hadronic system of the same proton, and one of the outgoing W bosons respectively. By taking $n = (1, \vec{n})$ with $|\vec{n}|^2 \leq 1$, all three terms in the above expression are positive. It can also be shown that, for any choice of n, $n.k$ can be negative for certain kinematical configurations. We therefore must choose a value of n for which $n.k < 0$ for all possible values of k, otherwise our amplitude will be undefined in certain regions of phase space. We should also attempt to maximise the value of $|n.k|$, such that the n dependent terms are minimised. To do this, we attempt to minimise the value of $n.p_p$ as this is the only term in Eq. (6.10) for which we have complete control over. This is given by:

$$n.p_p = E_p(1 - |\vec{n}|\beta_p \cos\theta_p), \tag{6.11}$$

where E_p is the energy of the proton, β_p is the speed of the proton and θ_p is the angle between the proton's 3–momentum and $\vec{n}$. It is clear that the minimum of this occurs for $|\vec{n}| = 1$ and $\theta = 0$, resulting in $n = (1, 0, 0, 1)$ being the optimal value when the proton is moving along the z–direction. We note that the same argument holds when taking the other proton in the above equation, so $n = (1, 0, 0, -1)$ is equivalent to this choice. It can be explicitly shown that, with this choice of n, we have:

$$|n.k| \gtrsim \frac{1}{\sqrt{s}}\left(4M^2 + 4m_W^2 - 2m_p^2\right), \tag{6.12}$$

where M is the invariant mass of the outgoing hadronic system of the proton travelling in the positive z–direction and m_p is the mass of the proton.

Other n dependent terms can appear in the longitudinal polarization:

$$\epsilon_L^\mu(p_W) = \sqrt{\frac{-p_W^2}{(p_W.n)^2 - n^2 p_W^2}}\left(n^\mu - \frac{n^2}{p_W.n}p_W^\mu\right) \overset{n^2=0}{=} \frac{im_W}{n.p_W}n^\mu, \tag{6.13}$$

where we have explicitly taken a light–like value of n given the above discussion. From this, one might be concerned that the value of $n.p_W$ may become small, however, using the value of $n = (1, 0, 0, 1)$, we see that this quantity is bounded below by:

$$n.p_W \gtrsim \frac{2m_W^2}{\sqrt{s}}. \tag{6.14}$$

It should be noted that in both Eqs. (6.12) and (6.14), the bound has been produced by requiring the W boson to move along the z–direction at relativistic speeds. This means that the boson is moving directly down the beam line, with the chance to measure its decay products suppressed by the headlight effect. Thus, the bounds in these equations are perhaps too strict for real experiments, with much higher values typically found. The same value of n has also been found to generate the most stable results in the context of $WW \to WW$ scattering in hadron–hadron collisions [180].

Using the axial gauge with $n = (1, 0, 0, 1)$, we again calculate the cross section for the $qq \to qWWq$ process with right–handed quarks in the structure function approach and compare to the full determination in Fig. 6.4. Here we see that the structure function approach using the axial gauge predicts an overall larger cross section compared to including all diagrams. This is likely due to the exclusion of Z–initiated diagrams, with the $Z\gamma \to WW$ sub–process interfering destructively with the purely photon–initiated diagrams. This can further be seen by the enhanced cross section in the $\kappa_Z > 1$ region, where these diagrams are expected to be relevant. Similar to the unitary$'$ scheme, the longitudinally polarized W bosons provide a substantially smaller contribution to the overall cross section compared to their transversely polarized counterparts. Additionally, while we see that these do have a larger contribution to the cross section at high κ_Z, this is still sub–dominant to the transverse polarizations at $\kappa_Z \sim 100$. This is in contrast to the unitary$'$ scheme,

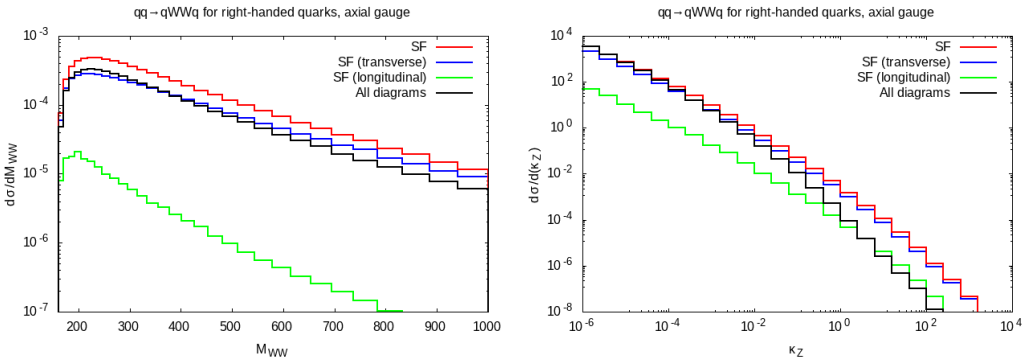


Figure 6.4: Same as Fig. 6.2 except using the axial for the structure function calculation.

where the transverse polarizations has a negligible impact in the same region (see Fig. 6.3 (right)). Finally, we see that the structure function approach using the axial gauge is a significantly better approximation to the full calculation compared to using the standard unitary gauge. This is driven by the longitudinal polarization, which for the unitary case grows with energy, as seen in Eq. (6.5), while in the axial case this, is bounded for the value of n we have chosen, as seen in Eq. (6.14). A similar result has been reported in the context of $WW \to WW$ processes [180], with a similar difference between the structure function calculation using the axial gauge and the full determination observed in this process.

Given the results of the axial gauge and the unitary$'$ scheme presented here, one might be tempted to suggest that the latter provides more accurate results. However, we have only considered right–handed quarks, with the left–handed case likely to differ significantly due to the different couplings between the quarks and the W/Z bosons. Therefore, it is unlikely that the unitary$'$ scheme will provide an acceptable description of this process, while a judgement for the axial gauge can only be made once a full calculation is available.

6.5 WW Production at the LHC

We now move onto the full calculation of the production of WW pairs through photons at a hadron–hadron collider. We compare the results of the structure function approach using the axial gauge to the calculation using the photon PDF. We also present the structure function calculation using the unitary$'$ scheme, although, as this is not a true gauge, this is simply to demonstrate the impact of gauge/scheme dependency on the final result. The structure function cross section is evaluated using (6.1), applying the relevant elastic or inelastic structure functions determined by the appropriate MMHT2015qed PDF set. The matrix elements are evaluated in either the axial gauge described in section 6.4 or in the unitary$'$ scheme in Eq. (6.7). For the photon PDF determination, we will be using the same PDF set as the structure function approach, which employs the LUXqed formalism to determine the photon PDF. With this, the cross section can be calculated in the standard form:

$$\sigma_{pp} = \int \mathrm{d}x_1 \mathrm{d}x_2 f_1^\gamma(x, \mu_F^2) f_2^\gamma(x, \mu_F^2) \hat{\sigma}_{\gamma\gamma}(x_1, x_2, \mu_R^2), \tag{6.15}$$

where similarities can be seen with Eq. (2.24). The PDFs are evolved using `APFEL` [160] and we take $\mu_F = \mu_R = M_{WW}$ as the nominal scale, varying μ_F and μ_R by a factor of two simultaneously in both directions to estimate the uncertainties due to scale variation. We note that, although the cross section $\hat{\sigma}_{\gamma\gamma}$ is only evaluated to leading order, it still depends on the scale through the running coupling (see Eq. (2.4)). We will not be including other uncertainties, such as PDF errors, when calculating the cross section either through the photon PDF or the structure function approach as these are expected to be much smaller [20].

We begin with the exclusive case where both protons remain intact, for which we

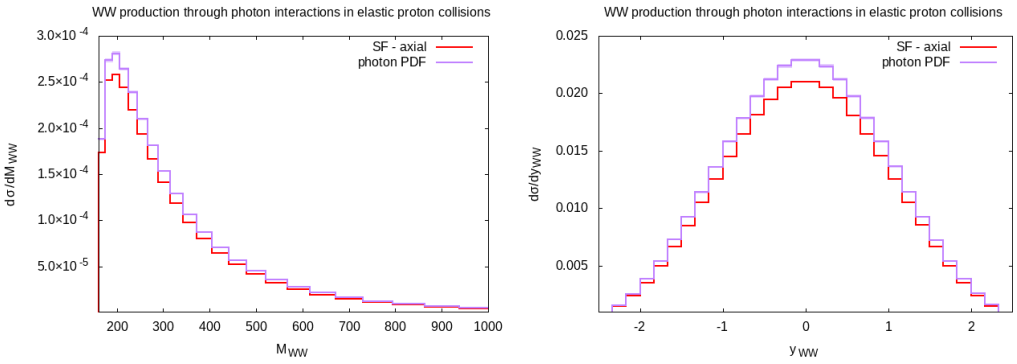


Figure 6.5: Cross section of elastic $pp \to pWWp$ via photon–initiated processes calculated using the photon PDF and the structure function approach.

use the MMHT2015qed_nnlo_elastic PDF set and apply the same set of cuts that were used in the calculation of the $qq \to qWWq$ process in the previous section, excluding those on the photons' invariant mass. The results of these calculations are shown in Fig. 6.5. We only show the structure function calculation with a single gauge as the elastic structure functions are highly peaked at $Q^2 = 0$ and thus the gauge dependency and the inclusion of non–factorizable contributions are negligible. We see that the photon PDF calculation predicts a slightly higher cross section compared to the structure function approach by around $\sim 10\%$. This is larger than the small scale uncertainties on the PDF calculation, suggesting a disagreement between the two methodologies. The structure function approach contains the full Q^2 dependence of the process while the photon PDF does not. Therefore, one could argue that this method is likely to be more accurate, even if such dependence is suppressed by the elastic structure functions. Regardless, the difference between these calculations is much smaller than the current uncertainty on the exclusive WW production as measured by ATLAS at 8 TeV [170], where they estimated an uncertainty of 39%, driven by the low statistics available.

We next move onto the double–dissociative case where both protons interact in-

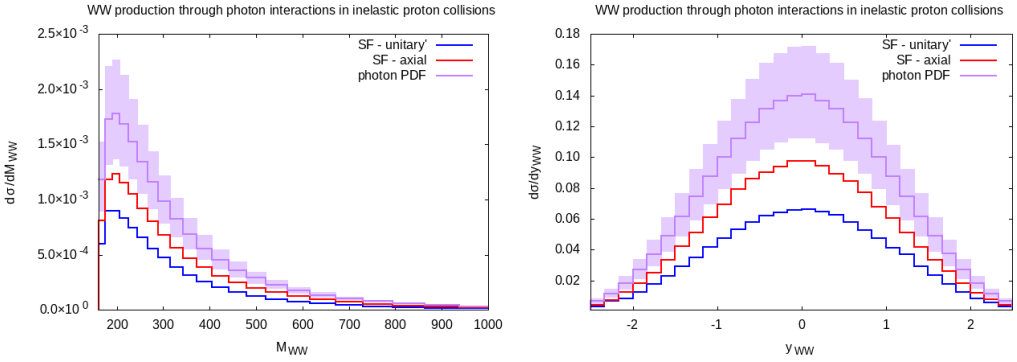


Figure 6.6: Same as Fig. 6.5 but with both protons interacting inelastically.

elastically (we omit the single–dissociative case for simplicity). Here, the structure functions are much broader and thus $Q^2 \neq 0$ effects are more prominent. This can be seen in Fig. 6.6, where the gauge dependency is clear, with the unitary$'$ scheme predicting a smaller cross section compared to the axial calculation. We further see that the LUXqed formalism, now calculated using the MMH2015qed_nnlo_inelastic PDF, predicts a larger cross section compared to the axial determination by a factor of $\sim$40% around the M_{WW} peak. It is clear that the dependence on methodology is rather large, with a wide range of predictions made. Further, each exist outside of the uncertainty bands of the other predictions, suggesting disagreement. A similar observation has already been seen in the case of dilepton production [64], with the structure function approach again predicting a smaller cross section, outside of the factorization scale variation uncertainty. Given the significant difference, further investigation is required to obtain a reliable prediction for WW production via photons in inelastic proton collisions.

6.6 Summary

In this chapter we have calculated the cross section for the production of WW pairs at the LHC through photon–initiated processes. This is of particular interest due to the dependence on triple and quartic boson interaction which are of great importance in some Beyond the Standard Model theories. We have performed this calculation using both the photon PDF and the new structure function approach, taking different gauges for the latter. We find that, for both exclusive and double–dissociation processes, these two calculations differ outside of scale variation uncertainties, with the structure function approach predicting a smaller cross section overall. Further, for the inelastic case, we also see that there is a large gauge dependency in the structure function approach due to missing non–factorizable and Z–initiated diagrams, with the unitary gauge providing an unreasonable prediction, while the axial gauge performs substantially better. Further investigation is required to produce a reliable cross section for this process that minimises the impact of these missing diagrams. In particular, a future study will assess the impact of non–factorizable contributions for left–handed quarks. The gauge dependency in exclusive production is negligible, as the strongly peaked elastic structure functions force the photons to be approximately on–shell. Additionally, here the difference between the structure function approach and the photon PDF calculation are much smaller than the current experimental uncertainty, suggesting that either calculation is reliable at the current level of precision. However, one would expect that such measurements will become more precise in the near future, with the difference in theoretical calculations becoming significant. Thus understanding this difference and predicting a reliable observable will be important.

Chapter 7

Conclusion

The study of the PDFs continues to be of vital importance to the wider fields of QCD and particle physics as we move through the precision era at the LHC. Not only do they reveal the fundamental structure of the proton as they collide at high–energies, but their determination is an essential ingredient in the measurement of Standard Model quantities and in constraining Beyond the Standard Model theories. In this thesis we have presented several studies detailing the current progress from data and theory perspectives, as well as project results that should become available in the near future. These advancements allow for tighter constraints on the PDFs to be obtained, which will subsequently propagate to the measurement of observables at the LHC and beyond and enable further understanding of the proton structure.

In chapter 3 we detailed the analysis of the differential top quark pair production measured at the LHC in the context of an MMHT14 fit. We showed that these data are difficult to fit as presented, with the fully correlated systematic uncertainties resulting in an overall poor fit. These difficulties are only representative of a wider problem of highly correlated data at the LHC, with other datasets showing similar complications. We demonstrated that decorrelating some of the two–point Monte Carlo uncertainties enables this data to be fit consistently. Such methodology has also been applied to other datasets within the MSHT20 fit with good results. How-

ever, this approach should be used with caution, with each decorrelation requiring justification on a case–by–case basis.

We then moved from tension found within a dataset, to ones found between experiments in chapter 4. In particular, we focused on a possible tension found between the CCFR/NuTeV dimuon data with the more recent ATLAS WZ data. Both processes are important for the determination of the poorly constrained strange PDF, which is subsequently a substantial ingredient for the study of certain electroweak parameters. We examined this apparent tension in the context of a PDF fit, with a recently calculated FFNS theory prediction and an updated MMHT14 VFNS determination. We showed that, while there is mild deterioration of the fit quality, there is an overall slightly better agreement of R_s at high x. These potential tensions will need to be monitored closely moving forward, particularly as more data is collected from processes sensitive to the strange quark such as W production in association of a charm quark.

In chapter 5 we looked to the future of PDF determinations at the HL–LHC and LHeC. With the impressive luminosity that will be provided at such detectors, they will supersede the current data available from the LHC and HERA experiments. Here, we profiled projected data using the Hessian Profiling methodology to place constraints on the PDF uncertainties and showed that both detectors are able to provide impressive bounds on the PDFs and are complementary of each other. We saw that, in general, the HL–LHC performs better within the high x region while LHeC determines the low x physics to better precision. Such conclusions are dependent on the specifics of how the pseudo–data were generated, in particular, additional processes could be chosen to enable both detectors to place bounds on regions for which the current projections do not. These projections have been subsequently used in other calculations of SM and BSM quantities that will be analysed at these

detectors.

Finally, in chapter 6 we calculated the cross section of WW pair production at the LHC through photon–initiated processes. This is of particular interest in certain Beyond the Standard Model theories, with these predicting the couplings involved in this process to be anomalous. We performed this calculation using the newly developed structure function approach and showed that careful consideration of the gauge used when calculating the inelastic process is needed for a reasonable result to be obtained. Comparison between this method and the standard photon PDF approach for exclusive processes showed that, while these disagree outside of error bands, the difference is much smaller than the uncertainty from the current best experimental measurement.

While these studies have broad implications for the determination of PDFs at the LHC and beyond, they merely represent a snapshot of the current research into PDFs and the wider field of QCD. With the ever more precise data that is becoming available and the continuous advancements in theoretical calculations, our understanding of these topics is ever improving. As we progress further into these fields, we look towards the potential discoveries that they may uncover.

Bibliography

[1] R. Abdul Khalek, S. Bailey, J. Gao, L. Harland-Lang, and J. Rojo, Eur. Phys. J. **C78**, 962 (2018), 1810.03639.

[2] R. Abdul Khalek, S. Bailey, J. Gao, L. Harland-Lang, and J. Rojo, SciPost Phys. **7**, 051 (2019), 1906.10127.

[3] S. Bailey and L. Harland-Lang, Eur. Phys. J. C **80**, 60 (2020), 1909.10541.

[4] G. Miller *et al.*, Phys. Rev. D **5**, 528 (1972).

[5] M. Constantinou, Eur. Phys. J. A **57**, 77 (2021), 2010.02445.

[6] A. D. Martin, R. G. Roberts, and W. J. Stirling, Phys. Rev. D **37**, 1161 (1988).

[7] H1, ZEUS, F. D. Aaron *et al.*, JHEP **01**, 109 (2010), 0911.0884.

[8] H. Abramowicz *et al.*, The European Physical Journal C **70**, 945963 (2010).

[9] H1, F. D. Aaron *et al.*, Eur. Phys. J. C **71**, 1579 (2011), 1012.4355.

[10] H1, F. D. Aaron *et al.*, JHEP **09**, 061 (2012), 1206.7007.

[11] ZEUS, H. Abramowicz *et al.*, Phys. Rev. D **87**, 052014 (2013), 1208.6138.

[12] S. Bailey, T. Cridge, L. A. Harland-Lang, A. D. Martin, and R. S. Thorne, Eur. Phys. J. C **81**, 341 (2021), 2012.04684.

[13] ATLAS, G. Aad *et al.*, Eur. Phys. J. **C76**, 538 (2016), 1511.04716.

[14] CMS, V. Khachatryan *et al.*, Eur. Phys. J. **C75**, 542 (2015), 1505.04480.

[15] ATLAS, M. Aaboud *et al.*, Phys. Rev. **D94**, 092003 (2016), 1607.07281.

[16] L. A. Harland-Lang, A. D. Martin, P. Motylinski, and R. S. Thorne, Eur. Phys. J. **C75**, 204 (2015), 1412.3989.

[17] J. Gao, L. Harland-Lang, and J. Rojo, Physics Reports **742**, 1121 (2018).

[18] NuTeV, M. Tzanov *et al.*, Phys. Rev. **D74**, 012008 (2006), hep-ex/0509010.

[19] NuTeV, M. Goncharov *et al.*, Phys. Rev. **D64**, 112006 (2001), hep-ex/0102049.

[20] L. A. Harland-Lang, JHEP **03**, 128 (2020), 1910.10178.

[21] R. K. Ellis, W. J. Stirling, and B. R. Webber, *QCD and collider physics*Cambridge monographs on particle physics, nuclear physics, and cosmology ; 8 (Cambridge University Press, Cambridge, 1996).

[22] R. Devenish and A. Cooper-Sarkar, *Deep inelastic scattering* (Oxford University Press, Oxford, 2004).

[23] T. A. et al., Nuclear Physics B **439**, 471 (1995).

[24] A. B. et al., Physics Letters B **223**, 485 (1989).

[25] J. C. Collins and D. E. Soper, Ann. Rev. Nucl. Part. Sci. **37**, 383 (1987).

[26] G. Altarelli and G. Parisi, Nuclear Physics B **126**, 298 (1977).

[27] ZEUS, H. Abramowicz *et al.*, Phys. Rev. D **99**, 092006 (2019), 1902.03048.

[28] S. Carrazza, C. Degrande, S. Iranipour, J. Rojo, and M. Ubiali, (2019), 1905.05215.

[29] A. Greljo *et al.*, JHEP **07**, 122 (2021), 2104.02723.

[30] T.-J. Hou *et al.*, Phys. Rev. D **103**, 014013 (2021), 1912.10053.

[31] PDF4LHC21 combination group, T. Cridge, PDF4LHC21: Update on the benchmarking of the CT, MSHT and NNPDF global PDF fits, in *28th International Workshop on Deep Inelastic Scattering and Related Subjects*, 2021, 2108.09099.

[32] S. J. Brodsky and G. R. Farrar, Phys. Rev. Lett. **31**, 1153 (1973).

[33] T. Carli *et al.*, Eur. Phys. J. **C66**, 503 (2010), 0911.2985.

[34] fastNLO, M. Wobisch, D. Britzger, T. Kluge, K. Rabbertz, and F. Stober, (2011), 1109.1310.

[35] fastNLO, D. Britzger, K. Rabbertz, F. Stober, and M. Wobisch, New features in version 2 of the fastNLO project, in *20th International Workshop on Deep-Inelastic Scattering and Related Subjects*, 2012, 1208.3641.

[36] A. Buckley *et al.*, Eur.Phys.J. **C75**, 132 (2015), 1412.7420.

[37] NNPDF, R. D. Ball *et al.*, Eur. Phys. J. C **77**, 663 (2017), 1706.00428.

[38] ATLAS Collaboration, CERN Report No. ATL-PHYS-PUB-2018-017, 2018 (unpublished).

[39] H1, ZEUS, H. Abramowicz *et al.*, Eur. Phys. J. C **75**, 580 (2015), 1506.06042.

[40] Particle Data Group, P. A. Zyla *et al.*, PTEP **2020**, 083C01 (2020).

[41] S. Alekhin, J. Blümlein, S. Moch, and R. Placakyte, Phys. Rev. D **96**, 014011 (2017), 1701.05838.

[42] NNPDF, R. D. Ball *et al.*, Eur. Phys. J. C **78**, 408 (2018), 1802.03398.

[43] T. Cridge, L. A. Harland-Lang, A. D. Martin, and R. S. Thorne, (2021), 2106.10289.

[44] NNPDF, R. D. Ball *et al.*, Eur. Phys. J. C **76**, 647 (2016), 1605.06515.

[45] J. Gao, L. Harland-Lang, and J. Rojo, Phys. Rept. **742**, 1 (2018), 1709.04922.

[46] L. A. Harland-Lang, A. D. Martin, P. Motylinski, and R. S. Thorne, Eur. Phys. J. **C75**, 204 (2015), 1412.3989.

[47] ATLAS, G. Aad *et al.*, Phys. Rev. D **85**, 072004 (2012), 1109.5141.

[48] CMS, S. Chatrchyan *et al.*, JHEP **02**, 013 (2014), 1310.1138.

[49] R. D. Ball *et al.*, (2021), 2109.02653.

[50] R. D. Ball *et al.*, (2021), 2109.02671.

[51] S. Carrazza and J. Cruz-Martinez, Eur. Phys. J. C **79**, 676 (2019), 1907.05075.

[52] S. Dulat *et al.*, Phys. Rev. **D93**, 033006 (2016), 1506.07443.

[53] ATLAS, G. Aad *et al.*, Phys. Rev. Lett. **109**, 012001 (2012), 1203.4051.

[54] J. Butterworth *et al.*, J. Phys. **G43**, 023001 (2016), 1510.03865.

[55] NNPDF, R. D. Ball *et al.*, JHEP **04**, 040 (2015), 1410.8849.

[56] R. Rojo, Update on the pdf4lhc20 benchmarking exercise, 2020.

[57] ATLAS, G. Aad *et al.*, JHEP **02**, 153 (2015), 1410.8857, [Erratum: JHEP09,141(2015)].

[58] NNPDF, V. Bertone, S. Carrazza, N. P. Hartland, and J. Rojo, SciPost Phys. **5**, 008 (2018), 1712.07053.

[59] L. A. Harland-Lang, A. D. Martin, R. Nathvani, and R. S. Thorne, Eur. Phys. J. C **79**, 811 (2019), 1907.02750.

[60] L. Buonocore, P. Nason, F. Tramontano, and G. Zanderighi, JHEP **08**, 019 (2020), 2005.06477.

[61] A. Manohar, P. Nason, G. P. Salam, and G. Zanderighi, Phys. Rev. Lett. **117**, 242002 (2016), 1607.04266.

[62] V. M. Budnev, I. F. Ginzburg, G. V. Meledin, and V. G. Serbo, Phys. Rept. **15**, 181 (1975).

[63] L. A. Harland-Lang, JHEP **03**, 128 (2020), 1910.10178.

[64] L. A. Harland-Lang, (2021), 2101.04127.

[65] NNPDF, R. Abdul Khalek *et al.*, Eur. Phys. J. C **79**, 931 (2019), 1906.10698.

[66] L. A. Harland-Lang and R. S. Thorne, Eur. Phys. J. **C79**, 225 (2019), 1811.08434.

[67] R. D. Ball and R. L. Pearson, (2021), 2105.05114.

[68] R. Abdul Khalek, J. J. Ethier, J. Rojo, and G. van Weelden, JHEP **09**, 183 (2020), 2006.14629.

[69] I. Schienbein *et al.*, Phys. Rev. D **80**, 094004 (2009), 0907.2357.

[70] Asymmetry Analysis, Y. Goto *et al.*, Phys. Rev. D **62**, 034017 (2000), hep-ph/0001046.

[71] C. Lorcé, B. Pasquini, and P. Schweitzer, JHEP **01**, 103 (2015), 1411.2550.

[72] P. Azzi *et al.*, CERN Yellow Rep. Monogr. **7**, 1 (2019), 1902.04070.

[73] M. Cepeda *et al.*, CERN Yellow Rep. Monogr. **7**, 221 (2019), 1902.00134.

[74] LHeC, FCC-he Study Group, P. Agostini *et al.*, (2020), 2007.14491.

[75] R. Abdul Khalek *et al.*, (2021), 2103.05419.

[76] D. Britzger *et al.*, Eur. Phys. J. C **79**, 845 (2019), 1906.05303.

[77] V. S. Fadin, E. Kuraev, and L. Lipatov, Phys.Lett. **B60**, 50 (1975).

[78] I. I. Balitsky and L. N. Lipatov, Sov. J. Nucl. Phys. **28**, 822 (1978).

[79] R. D. Ball *et al.*, Eur. Phys. J. **C78**, 321 (2018), 1710.05935.

[80] L. A. Harland-Lang, A. D. Martin, and R. S. Thorne, Eur. Phys. J. **C78**, 248 (2018), 1711.05757.

[81] M. Czakon, N. P. Hartland, A. Mitov, E. R. Nocera, and J. Rojo, JHEP **04**, 044 (2017), 1611.08609.

[82] R. Boughezal, A. Guffanti, F. Petriello, and M. Ubiali, JHEP **07**, 130 (2017), 1705.00343.

[83] J. Currie, E. W. N. Glover, and J. Pires, Phys. Rev. Lett. **118**, 072002 (2017), 1611.01460.

[84] A. Gehrmann-De Ridder, T. Gehrmann, E. W. N. Glover, A. Huss, and T. A. Morgan, JHEP **11**, 094 (2016), 1610.01843, [Erratum: JHEP10,126(2018)].

[85] R. Boughezal *et al.*, Phys. Rev. Lett. **116**, 152001 (2016), 1512.01291.

[86] M. Czakon, D. Heymes, and A. Mitov, Phys. Rev. Lett. **116**, 082003 (2016), 1511.00549.

[87] ATLAS, M. Aaboud *et al.*, JHEP **09**, 020 (2017), 1706.03192.

[88] ATLAS, G. Aad *et al.*, JHEP **09**, 145 (2014), 1406.3660.

[89] CMS, V. Khachatryan *et al.*, Phys. Lett. **B749**, 187 (2015), 1504.03511.

[90] M. Guzzi, K. Lipka, and S.-O. Moch, JHEP **01**, 082 (2015), 1406.0386.

[91] L. A. Harland-Lang, Talk at ATLAS SM Workshop, London, September 2018 `https://indico.cern.ch/event/716301/contributions/3113778/attachments/1710873/2758180/atlas_PDF.pdf`, 2019.

[92] R. S. Thorne *et al.*, Updates of PDFs in the MMHT framework, in *27th International Workshop on Deep Inelastic Scattering and Related Subjects (DIS 2019) Torino, Italy, April 8-12, 2019*, 2019, 1907.08147.

[93] T.-J. Hou, Z. Yu, S. Dulat, C. Schmidt, and C. P. Yuan, Phys. Rev. D **100**, 114024 (2019), 1907.12177.

[94] S. Amoroso *et al.*, Les Houches 2019: Physics at TeV Colliders: Standard Model Working Group Report, in *11th Les Houches Workshop on Physics at TeV Colliders: PhysTeV Les Houches*, 2020, 2003.01700.

[95] CMS, V. Khachatryan *et al.*, Eur. Phys. J. **C77**, 15 (2017), 1602.09024.

[96] M. Czakon, D. Heymes, and A. Mitov, (2017), 1704.08551.

[97] T. Kluge, K. Rabbertz, and M. Wobisch, FastNLO: Fast pQCD calculations for PDF fits, in *14th International Workshop on Deep Inelastic Scattering*, 2006, hep-ph/0609285.

[98] M. Czakon *et al.*, JHEP **10**, 186 (2017), 1705.04105.

[99] Amanda Cooper-Sakar, private comunication.

[100] L. A. Harland-Lang, A. D. Martin, P. Motylinski, and R. S. Thorne, Eur. Phys. J. **C76**, 186 (2016), 1601.03413.

[101] CMS, V. Khachatryan *et al.*, JHEP **08**, 029 (2016), 1603.02303.

[102] CMS, A. M. Sirunyan *et al.*, Eur. Phys. J. **C77**, 459 (2017), 1703.01630.

[103] ATLAS, M. Aaboud *et al.*, Eur. Phys. J. **C77**, 367 (2017), 1612.03016.

[104] L. A. Harland-Lang, A. D. Martin, and R. S. Thorne, Eur. Phys. J. **C78**, 248 (2018), 1711.05757.

[105] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt, Eur. Phys. J. **C63**, 189 (2009), 0901.0002.

[106] F. Faura, S. Iranipour, E. R. Nocera, J. Rojo, and M. Ubiali, Eur. Phys. J. C **80**, 1168 (2020), 2009.00014.

[107] CCFR, A. O. Bazarko *et al.*, Z. Phys. **C65**, 189 (1995), hep-ex/9406007.

[108] M. A. G. Aivazis, J. C. Collins, F. I. Olness, and W.-K. Tung, Phys.Rev. **D50**, 3102 (1994), hep-ph/9312319.

[109] R. S. Thorne, Phys.Rev. **D73**, 054019 (2006), hep-ph/0601245.

[110] M. Buza, Y. Matiounine, J. Smith, and W. L. van Neerven, Eur. Phys. J. **C1**, 301 (1998), hep-ph/9612398.

[111] W.-K. Tung, S. Kretzer, and C. Schmidt, J. Phys. G **28**, 983 (2002), hep-ph/0110247.

[112] S. Kretzer, H. L. Lai, F. I. Olness, and W. K. Tung, Phys. Rev. D **69**, 114005 (2004), hep-ph/0307022.

[113] R. S. Thorne and R. G. Roberts, Phys. Rev. **D57**, 6871 (1998), hep-ph/9709442.

[114] W. L. van Neerven and A. Vogt, Nucl. Phys. **B568**, 263 (2000), hep-ph/9907472.

[115] D. de Florian, R. Sassot, P. Zurita, and M. Stratmann, Phys. Rev. D **85**, 074028 (2012).

[116] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt, The European Physical Journal C **63**, 189285 (2009).

[117] D. e. a. Mason, Phys. Rev. Lett. **99**, 192001 (2007).

[118] CMS, S. Chatrchyan *et al.*, JHEP **02**, 013 (2014), 1310.1138.

[119] ATLAS, G. Aad *et al.*, Eur. Phys. J. C **79**, 760 (2019), 1904.05631.

[120] ATLAS, M. Aaboud *et al.*, JHEP **12**, 059 (2017), 1710.05167.

[121] T. Bolton, (1997), hep-ex/9708014.

[122] D. Mason *et al.*, Phys. Rev. Lett. **99**, 192001 (2007).

[123] S. Alekhin, J. Blümlein, S. Moch, and R. Placakyte, Phys. Rev. **D96**, 014011 (2017), 1701.05838.

[124] A. Dainese *et al.*, editors, *Report on the Physics at the HL-LHC,and Perspectives for the HE-LHC*, CERN Yellow Reports: Monographs Vol. 7/2019 (CERN, Geneva, Switzerland, 2019).

[125] ZEUS, H1, H. Abramowicz *et al.*, Eur. Phys. J. **C75**, 580 (2015), 1506.06042.

[126] LHeC Study Group, J. Abelleira Fernandez *et al.*, J.Phys. **G39**, 075001 (2012), 1206.2913.

[127] LHeC Study Group, J. L. Abelleira Fernandez *et al.*, (2012), 1211.5102.

[128] LHeC study Group, H. Paukkunen, PoS **DIS2017**, 109 (2018), 1709.08342.

[129] NNPDF, R. D. Ball *et al.*, Phys. Lett. **B728**, 524 (2014), 1310.0461.

[130] C. Marquet, M. R. Moldes, and P. Zurita, Phys. Lett. **B772**, 607 (2017), 1702.00839.

[131] E. C. Aschenauer, S. Fazio, M. A. C. Lamont, H. Paukkunen, and P. Zurita, Phys. Rev. **D96**, 114005 (2017), 1708.05654.

[132] E. C. Aschenauer, R. Sassot, and M. Stratmann, Phys. Rev. **D92**, 094030 (2015), 1509.06489.

[133] D. Boer *et al.*, (2011), 1108.1713.

[134] LHeC study Group, A. M. Cooper-Sarkar, PoS **DIS2016**, 274 (2016), 1605.08579.

[135] H. Paukkunen and P. Zurita, JHEP **12**, 100 (2014), 1402.6623.

[136] C. Schmidt, J. Pumplin, C. P. Yuan, and P. Yuan, (2018), 1806.07950.

[137] The NNPDF, R. D. Ball *et al.*, JHEP **05**, 075 (2010), 0912.2276.

[138] R. D. Ball *et al.*, JHEP **1304**, 125 (2013), 1211.5142.

[139] R. Gauld and J. Rojo, Phys. Rev. Lett. **118**, 072001 (2017), 1610.09373.

[140] ATLAS, G. Aad *et al.*, JHEP **08**, 009 (2016), 1606.01736.

[141] NNPDF, J. Rojo, PoS **DIS2017**, 198 (2018), 1705.04468.

[142] LHCb, R. Aaij *et al.*, JHEP **01**, 155 (2016), 1511.08039.

[143] M. Czakon, N. P. Hartland, A. Mitov, E. R. Nocera, and J. Rojo, JHEP **04**, 044 (2017), 1611.08609.

[144] ATLAS, G. Aad *et al.*, Eur. Phys. J. **C76**, 538 (2016), 1511.04716.

[145] ATLAS, G. Aad *et al.*, Eur. Phys. J. **C76**, 291 (2016), 1512.02192.

[146] W. Stirling and E. Vryonidou, Phys.Rev.Lett. **109**, 082002 (2012), 1203.6781.

[147] CMS, S. Chatrchyan *et al.*, Phys.Rev. **D90**, 032004 (2014), 1312.6283.

[148] CMS Collaboration, CERN Report No. CMS-PAS-SMP-17-014, 2018 (unpublished).

[149] Will Barter, Stephen Farry, private communication.

[150] J. M. Campbell, J. Rojo, E. Slade, and C. Williams, Eur. Phys. J. **C78**, 470 (2018), 1802.03021.

[151] ATLAS, M. Aaboud *et al.*, Phys. Lett. **B770**, 473 (2017), 1701.06882.

[152] J. Currie *et al.*, Phys. Rev. Lett. **119**, 152001 (2017), 1705.10271.

[153] E. L. Berger, J. Gao, and H. X. Zhu, JHEP **11**, 158 (2017), 1708.09405.

[154] R. Boughezal *et al.*, Eur. Phys. J. **C77**, 7 (2017), 1605.08011.

[155] Z. Nagy, Phys. Rev. **D68**, 094002 (2003), hep-ph/0307268.

[156] NNPDF, R. D. Ball *et al.*, JHEP **04**, 040 (2015), 1410.8849.

[157] ATLAS, M. Aaboud *et al.*, JHEP **05**, 195 (2018), 1711.02692.

[158] T. Sjostrand, S. Mrenna, and P. Z. Skands, Comput. Phys. Commun. **178**, 852 (2008), 0710.3820.

[159] `http://hep.ph.liv.ac.uk/~mklein/lhecdata/` and `http://hep.ph.liv.ac.uk/~mklein/heavyqdata/`.

[160] V. Bertone, S. Carrazza, and J. Rojo, Comput.Phys.Commun. **185**, 1647 (2014), 1310.1394.

[161] S. Forte, E. Laenen, P. Nason, and J. Rojo, Nucl. Phys. **B834**, 116 (2010), 1001.2312.

[162] NNPDF, R. Abdul Khalek *et al.*, Eur. Phys. J. **C**, 79:838 (2019), 1905.04311.

[163] C. Gwenlan, Talk at the DIS19, Turin, Italy, `https://indico.cern.ch/event/749003/contributions/3330763/attachments/1828248/2992940/DIS19-LHeC_Gwenlan.pdf`, 2019.

[164] G. Aad *et al.*, Physics Letters B **716**, 1 (2012).

[165] LHC Higgs Cross Section Working Group, J. R. Andersen *et al.*, (2013), 1307.1347.

[166] LHC Higgs Cross Section Working Group, D. de Florian *et al.*, (2016), 1610.07922.

[167] D. R. Green, P. Meade, and M.-A. Pleier, Rev. Mod. Phys. **89**, 035008 (2017).

[168] A. V. Manohar, P. Nason, G. P. Salam, and G. Zanderighi, JHEP **12**, 046 (2017), 1708.01256.

[169] L. A. Harland-Lang, A. D. Martin, R. Nathvani, and R. S. Thorne, Eur. Phys. J. C **79**, 811 (2019), 1907.02750.

[170] ATLAS, M. Aaboud *et al.*, Phys. Rev. D **94**, 032011 (2016), 1607.03745.

[171] CMS, V. Khachatryan *et al.*, JHEP **08**, 119 (2016), 1604.04464.

[172] ATLAS, G. Aad *et al.*, Phys. Lett. B **816**, 136190 (2021), 2010.04019.

[173] L. A. Harland-Lang, V. A. Khoze, and M. G. Ryskin, (2021), 2104.13392.

[174] T. Figy, C. Oleari, and D. Zeppenfeld, Phys.Rev. **D68**, 073005 (2003), hep-ph/0306109.

[175] P. Bolzoni, F. Maltoni, S.-O. Moch, and M. Zaro, Phys.Rev.Lett. **105**, 011801 (2010), 1003.4451.

[176] F. A. Dreyer and A. Karlberg, Phys. Rev. Lett. **117**, 072001 (2016), 1606.00840.

[177] E. Accomando, Eur. Phys. J. C **54**, 239 (2008), hep-ph/0604273.

[178] C. Dams and R. Kleiss, Eur. Phys. J. C **34**, 419 (2004), hep-ph/0401136.

[179] J. Chen, T. Han, and B. Tweedie, JHEP **11**, 093 (2017), 1611.00788.

[180] E. Accomando, A. Ballestrero, A. Belhouari, and E. Maina, Phys. Rev. D **74**, 073010 (2006), hep-ph/0608019.