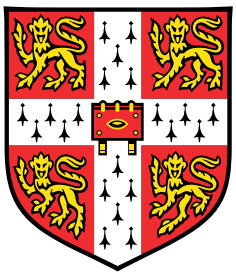
Precision QCD and effective field theories with machine learning



Shayan Iranipour

Department of Applied Mathematics and Theoretical Physics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Girton College

April 2022

Declaration

The work presented in this text is based on published research. This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared here and specified in the text.

The material presented in chapter 3 is based on [1, 2] and was done as part of my contribution to the NNPDF collaboration. Contributions to the public code base has formed a significant portion of the work of my PhD and so detail is given where appropriate.

Chapter 4 is published in [3] and was done in collaboration with F. Faura, E. R. Nocera, J. Rojo, and M. Ubiali

The first half of the work presented in chapter 5 is published in [4] and was done in collaboration with S. Carrazza, C. Degrande, J. Rojo, and M. Ubiali. The latter half was instead published in [5] and was done in collaboration with A. Greljo, Z. Kassabov, M. Madigan, J. Moore, J. Rojo, M. Ubiali, and C. Voisey.

The work of chapter 6 is based on my own work and published in [6] with M. Ubiali.

This thesis has not been submitted in whole or in part for consideration for any other degree or qualification at the University of Cambridge, or any other university.

Shayan Iranipour

April 2022

Abstract

Precision QCD and effective field theories with machine learning
Shayan Iranipour

The Standard Model (SM) serves as one the best descriptions of fundamental physics we have and the quest for its falsification has led to it being tested to an unprecedented degree. Despite its flawless performance, there are many theoretical and phenomenological indications that the SM cannot be a complete description of nature; though, so far, no direct evidence for new physics at the TeV scale has been gathered at colliders. Far from being discouraging, the precision level reached by current experiments gives us the unique opportunity to investigate the effects of new particles whose masses are far above the TeV scale, but still produce observable effects at the scales within the direct kinematical reach of the Large Hadron Collider (LHC). Unlike for direct searches, which are limited by the energy reach of the collider, indirect searches are limited only by the theoretical and experimental control over the processes under inspection.

A robust understanding of Quantum Chromodynamics (QCD) is crucial in order to achieve precision theoretical predictions in the era of initial state hadron colliders such as the LHC. An important ingredient therein are the parton distribution functions (PDFs) which parameterize the proton structure in terms of its elementary quark and gluon constituents. These quantities are non-perturbative and obtained from data using a global QCD analysis. In tandem, Effective Field Theories (EFTs), provide a convenient framework to capture the indirect effects of possible BSM resonances in low energy observables. Constraints on the EFT then translate to constraints on the nature of BSM physics.

This manuscript serves to marry these two endeavours. We present machine learning-based approaches to PDF determination and specifically highlight how deep learning algorithms form ideal candidates to parameterize the PDFs in an unbiased fashion. We present the NNPDF4.0 PDF set which serves as the latest and most precise determination of proton structure delivered by such a methodology. We show how a

precise determination of the PDFs has important consequences on LHC phenomenology by presenting a precision determination of the strange content of the proton and a number of key phenomenological applications.

We then discuss the interplay between EFT dynamics and the PDFs; analysing the extent to which the fit of PDFs may absorb possible BSM signals and assess the implications a consistent treatment of PDFs in EFT fits has on phenomenological studies. For this, we use legacy deep inelastic scattering data from HERA and later some more modern measurements from high-mass Drell-Yan observables at the Large Hadron Collider (LHC) to investigate the back-reaction of EFT dynamics on the PDFs.

The considerations presented in the above study then act as an impetus to develop a methodology that is capable of simultaneously determining proton structure alongside BSM dynamics in a consistent framework. We present a novel methodology, `SIMUnet`, which delivers a robust and accurate determination of PDFs and general theory parameters, of which BSM dynamics are a subset. We show how this state-of-the-art methodology can, for the first time, extract and disentangle the PDFs from BSM dynamics from a global dataset paving the way for a truly global and simultaneous interpretation of indirect searches in the context of precision physics.

Acknowledgements

I would like to start by thanking my PhD supervisor, Maria Ubiali, for all the support and guidance she has given me during my graduate studies. Her expertise and wealth of knowledge in physics has proven invaluable over the years and her encouragement to pursue various ideas have led to a vast number of the results presented in this text. Thank you for giving me the opportunity to pursue research into particle physics; a field which I have been incredibly passionate about since the start of my physics career. Having worked under her tutelage is a privilege which I shall cherish.

I have had the pleasure of working with the NNPDF and PBSP collaboration during my time as a graduate student. From both I have learned an incredible amount and have met many talented and wonderful collaborators. To both groups I am thankful, but special mention must go to Zahari Kassabov, who has taught me so much, in particular regarding topics in machine learning and coding. These are skills which I am grateful for and I am very fortunate to have been taught them by a friend so well versed in the field.

My time at DAMTP and Cambridge has been made particularly special thanks to the staff and students that make up its members. This text would be significantly lengthened if were I to mention them all, but I wish to particularly thank Daniel Zhang, Sam Crew, Ben Day, Theo Björkmo, Bogdan Ganchev, James Delaney, Robin Croft, Hasan Mahmood, Maeve Madigan, João Melo, Ward Haddadin, James Moore, Manuel Morales, Cameron Voisey, Matthew Wales, and Manda Stagg.

A particular thank you must go to Ashkan for his company during some particularly long study sessions and his incredible sense of humour that made them considerably more bearable. Special thanks also to Kimiya for her unconditional support and encouragement; making the good times better and the tough times easier.

Above all else I wish to thank my parents, Nahid and Hossain. You truly knew the value of education and for that I will always be indebted. Thank you for believing in me; this thesis is for you.

تقدیم به پدر و مادر عزیزم که علم را به من هدیه دادند

Table of contents

Chapter 1

Introduction

The Standard Model (SM) [7–13] serves as the crowning achievement of modern scientific endeavour. Its ability to explain such a vast range of phenomena, such as the prediction of the Higgs boson [14–16] and precision measurements of the electron anomalous magnetic moment [17–19], has been unsurpassed by any other theory and has therefore planted itself as one of the best successes in our fundamental understanding of nature. Despite its copious and broad range of triumphs, we know as a matter of fact that it cannot be the whole story: there are simply too many known unknowns. For example, and perhaps most glaringly, one of the four fundamental forces of nature, gravity, eludes description by the Standard Model [20]; not to mention the neutrino oscillations [21, 22] or the lack of a dark matter sector [23].

Research is therefore not only rife in trying to extend and thus complete this already successful theory, but also to seek its falsification. Despite the battery of precision tests from the likes of the Large Electron-Positron collider (LEP), Tevatron, and the Large Hadron Collider (LHC), the SM remains resilient in defeat; with experiment yielding no direct evidence of a resonance on top of the SM background. This is not to say, however, that signs of new physics beyond the Standard Model (BSM) are non-existent. For example, strong evidence for lepton flavour universality violation exists in rare B-meson decays [24–26] or in measurements of the muon anomalous magnetic moment, $(g-2)_\mu$, [27–31]. It is not unreasonable, therefore, to suppose some BSM resonance lies slightly beyond the direct kinematic reach of our experimental apparatus, the effects of whom we will be indirectly sensitive to in the tails of some differential distributions.

In the strive to make the most of such indirect searches, precision, both experimental and theoretical, is above all else one of the key ingredients in the drive for BSM searches. A strong grasp of all sources of uncertainty, whether from detector design or missing

higher orders in the perturbative expansion, is tantamount if a genuine deviation from the SM can be attributed to new physics rather than simply statistical fluctuations or an imperfect formulation of SM predictions. Thanks to the enhanced luminosity and kinematic reach of the LHC, incredibly precise measurements of exotic processes can be made with high statistics. This can only be made better by the upcoming LHC run III and the subsequent High-Luminosity (HL-LHC) upgrade. Theoretical precision is also making similarly impressive progress with state-of-the-art observables, computed within the context of Quantum Chromodynamics (QCD), being calculated to unprecedented accuracy, in some cases accounting for several orders in perturbation theory [32–34]. However, perseverance in the perturbative expansion alone is not enough to constitute a precision theoretical prediction in the era of hadron colliders. Non-perturbative objects, known as parton distribution functions (PDFs) [1, 35–37], are also an essential component of the phenomenology program. Such quantities encode the inner structure of the proton and other hadronic initial states, but, unlike the hard cross sections, are not computable from first principles within perturbation theory, being instead extracted from experimental measurements. As such, accompanied with the PDF is another source of theoretical uncertainty associated with the finite resolution in the experimental measurements that they are fit to. These so-called PDF uncertainties are in fact one of the dominant sources of theoretical uncertainty in key theoretical predictions that enable a precise characterization of the Higgs boson or other SM background processes. It is clear then that if indirect searches are to be fruitful, these PDF uncertainties must be tamed as much as possible.

A convenient way to parameterize the effects of new physics in the infrared (IR) is through the use of Effective Field Theories (EFTs). Schematically, these are QFTs which we acknowledge are valid only up to a certain energy scale, beyond which the resolution of our theory becomes too coarse and we become sensitive to ultraviolet (UV) effects, for example the dynamics of the heavy modes. The IR theory is constructed from the UV using Wilsonian renormalization, whereby the heavy degrees of freedom are integrated out from the path integral, the effect of which is to introduce novel interactions between the light particles. Experimental measurements can then be used to place constraints on these new interactions [4, 5, 38–46] which in turn restrict the possible space of UV theories [47].

There is, however, a source of inconsistency with this approach. The PDFs are fitted to data assuming the validity of the SM at all scales, indeed the hard cross section being computed assuming the SM matter content and interactions. However, if

we are to believe that embedded within the data are signals of BSM physics, then it is entirely likely that the PDFs absorb these effects in order to obtain an adequate fit quality of the high-energy data that enter the fits. The astute physicist is then forced to exclude these BSM sensitive datasets from the PDF fit when performing an EFT study. However, doing so sacrifices constraining power on these reduced PDFs and thereby increases the PDF uncertainty which, as mentioned above, was paramount to reduce.

This apparent dichotomy is the concern of this manuscript. In what follows, chapter 2 provides an overview of the theoretical background needed for some of the ensuing discussion. We outline the theory of strong interactions: QCD, discussing its lagrangian and asymptotically free nature, before introducing PDFs by considering deep inelastic scattering experiments. Included in this chapter is also an overview of effective field theories and importantly the Standard Model Effective Field Theory (SMEFT) which is the EFT of choice for much of the later analysis. In chapter 3, we present the neural network approach to PDF fitting. An overview of deep learning is given, before we discuss how this important class of machine learning algorithm can be deployed to understand the structure of the proton, highlighting important methodological improvements in the latest NNPDF4.0 [1, 2] release and its phenomenological implications. The importance of precision PDFs is highlighted in chapter 4. Here, a precise determination of the strange quark content of the proton is provided, using some of the latest strange-sensitive measurements available [3]. We again show how an enhanced precision in the strange PDF has important consequence for phenomenology at the LHC. Chapter 5 then addresses the issue of the interplay between BSM dynamics and the PDFs. We use deep inelastic scattering data [4], and later high-mass Drell-Yan measurements from the LHC [5], to consider to what degree this interplay has implications on bounds one obtains on Wilson coefficients in the IR and to what extent the neural network determination of proton structure is susceptible to fitting away these BSM signals. We see that the effect is mild, though if the High-Luminosity LHC data is to be used, then one is at great peril from possible BSM contamination. In turn, this motivates chapter 6, which addresses the issue of a simultaneous determination of PDFs and EFT parameters [6]. For the first time, we present PDFs that have been fitted, simultaneously, alongside EFT coefficients, with no compromises on dataset selection. This new generation of fitting methodology, dubbed `SIMUnet`, provides a robust and accurate method to fit an arbitrary number of external theory parameters alongside PDFs, free from the worry that one may be

fitting away BSM signals or compromising constraining power due to a reduced dataset. Concluding remarks and future directions of the work considered in this text are then given in chapter 7.

Chapter 2

Theoretical overview

We begin the discussion of this manuscript by outlining a review of the various theoretical concepts which will prove relevant for subsequent discussions. We begin by providing an overview of Quantum Chromodynamics (QCD) which is an example of a non-abelian gauge theory describing the strong interactions. Hidden within a seemingly simple lagrangian is a wealth of emergent phenomena, but we shall concern ourselves in particular with an important class of non-perturbative objects known as parton distribution functions (PDFs). These quantities describe the momentum distribution of the constituents of the proton and we show how they arise naturally within the parton model. Moreover, we show that if one considers QCD corrections to the parton model, then the PDFs obtain an anomalous dimension using renormalization group methods and hence evolve with the energy scale. We then go on to discuss the fact that PDFs are process independent quantities and thus characterize the long distance (low energy) phenomena, with the hard cross section computable within the framework of perturbation theory. As such the PDFs are said to factorize the non-perturbative physics: a result that has been proven across a broad range of processes, known as the factorization theorems.

The second important topic to discuss is that of effective field theories (EFTs). These quantum field theories are valid up to some energy scale Λ beyond which we acknowledge the theory to no longer be valid. As such, requirements such as renormalizability may be abandoned with reasonable predictions being made by generally non-renormalizable theories. We shall provide a toy example which shows how such a theory can arise as the low energy IR limit of a UV complete theory by integrating out a heavy mass mode from the partition function. We show how the two theories give equivalent predictions in low energy regimes, so long as heavy modes resides only in the internal legs of

scattering graphs in the UV theory. Finally, we introduce an important example of an EFT known as the Standard Model Effective Field Theory (SMEFT). We specify its lagrangian and show how we may use it to obtain convenient handles on heavy resonances beyond the direct kinematic reach of modern colliders, but whose effects we may be indirectly sensitive to.

2.1 Review of Quantum Chromodynamics

We start by providing an overview of QCD. We shall present the classical level lagrangian which is a locally $SU(3)$ gauge invariant field theory with fermionic matter content. The gauge group is famously non-abelian which presents various nuances upon quantization. We shall discuss these and specifically how ghost particles are required in order to remove non-physical degrees of freedom from the theory.

At 1-loop and beyond, ultraviolet divergences arise due to momentum integrals and we show how the renormalization process then necessitates a running of the QCD coupling. This leads to a phenomenon known as asymptotic freedom which forms the fundamental basis for much of the discussion that is to come.

2.1.1 The QCD lagrangian

The theory of strong interactions, QCD, is the $SU(3)$ gauge theory where the matter content is minimally coupled through the use of the gauge covariant derivative. The matter content are spin-1/2 particles, which we shall refer to as quarks [11–13] and transform in the fundamental representation under the action of the gauge group. These particles come in N_f different *flavours* each with mass m_f. The matter content are thus N_f copies of Dirac spinor-valued $SU(3)$ triplets:

$$\psi^f = \begin{pmatrix} \psi^f_{\text{red}} \\ \psi^f_{\text{green}} \\ \psi^f_{\text{blue}} \end{pmatrix}. \tag{2.1}$$

The subscript red, green, and blue, denote the colour charge under $SU(3)$. These spinors transform as vectors under the action of the $\text{Spin}(1,3)$ group [1]. Despite the fundamental degrees of freedom of QCD being quarks and gluons, they are never observed as free particles, but rather as colourless (colour singlet) bound states known

[1] The universal covering of the Lorentz group [48].

as hadrons. Conceivable ways of forming hadrons is by having quark-antiquark bound states (such that the colours cancel, for example, red and anti-red) which form *mesons* or an odd number of valence quarks which form *baryons*, common examples of these are protons and neutrons (collectively nucleons) which form colour singlets by red, green, and blue forming a colourless state (akin to modular arithmetic).

The fact that the fundamentally realised units of nature at low energies are gapped hadronic bound states is known as *confinement* and stands as an open problem in fundamental physics at the time of writing. Though one can show analytically [49] confinement to be emergent in 4d $\mathcal{N} = 2$ supersymmetric [50] Yang-Mills models it is not clear how it can be done for QCD. Additionally, the low energy dynamics of hadron physics can be modelled directly using the framework of chiral perturbation theory ($\mathcal{X}$PT) [51] using light scalar pions as fundamental degrees of freedom. Though a vast number of phenomena can be explained using this framework a direct construction of $\mathcal{X}$PT from QCD using renormalization based techniques remains an open problem.

At the classical level, the dynamics of QCD is governed by the lagrangian:

$$\mathcal{L} = \sum_{f=1}^{N_f} \bar{\psi}_i^f (i\not{D} - m_f)_{ij} \psi_j^f - \frac{1}{4} F_{\mu\nu}^a F^{\mu\nu a} \tag{2.2}$$

where $f = 1, \ldots, N_f$ is the flavour index, i, j are colour indices and run over the fundamental representation space, and a (summation implied) enumerates Lie algebra indices. We employ the slashed notation: $\not{D} = \gamma^\mu D_\mu$ and Greek indices μ, ν enumerate spacetime coordinates. The γ matrices satisfy the Clifford algebra:

$$\{\gamma^\mu, \gamma^\nu\} \equiv \gamma^\mu \gamma^\nu + \gamma^\nu \gamma^\mu = 2\eta^{\mu\nu} \tag{2.3}$$

and provides a construction for the spin group. The metric $\eta^{\mu\nu} = \mathrm{diag}(1, -1, -1, -1)$ is the metric on Minkowski spacetime. The gauge covariant derivative, D_μ is defined in order to preserve local $SU(3)$ gauge invariance and couples the quarks with the gluons. Suppressing colour indices the gauge covariant derivative is given by:

$$D_\mu = \partial_\mu + igA_\mu^a(x)t^a \tag{2.4}$$

with an implied summation on a. Here g is the coupling to the gauge fields and A_μ^a is a Lorentz vector valued function of spacetime and is interpreted as the gluon field. The t^a are the Lie algebra generators in the fundamental representation of the $\mathfrak{su}(3)$

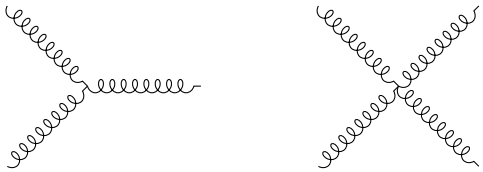


Figure 2.1: The 3-point (left) and 4-point (right) gluon Feynman diagrams in QCD.

Lie algebra which are given by

$$t^a = \lambda^a/2 \tag{2.5}$$

where λ^a are the 8, 3×3 Gell-Mann matrices. Since $\dim(\mathfrak{su}(3)) = 8$ we have $a = 1, \ldots, 8$ and correspondingly 8 independent gluons. In order to preserve local gauge invariance we require that the gauge fields transform under the adjoint representation. The Lie algebra generators satisfy:

$$[t^a, t^b] \equiv t^a t^b - t^b t^a = i f^{abc} t^c \tag{2.6}$$

$$\mathrm{Tr}(t^a t^b) = \frac{1}{2}\delta^{ab} \tag{2.7}$$

where f^{abc} are the structure constants of the Lie algebra. The gauge kinetic term of equation 2.2 comes from standard massless Yang-Mills theory [52]. For QCD the Killing form reduces to give the kinetic term as a simple sum in colour space. The field strength tensor is given by

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - g f^{abc} A^b_\mu A^c_\nu. \tag{2.8}$$

The fact that $SU(3)$ is non-Abelian implies the structure constants are non-vanishing. Thus the final term in equation 2.8 gives rises to gluon-gluon self-interaction Feynman diagrams shown in figure 2.1. Unlike the photon of Quantum Electrodynamics (QED), the gluon is thus charged under its own gauge symmetry and carries colour charge in a way that charge is preserved at every interaction vertex.

The quantization of a gauge theory, however, introduces subtleties. These generally stem from the fact that the gauge symmetry is in fact more of a redundancy in our lagrangian: in much the same way that the 4-potential of classical electromagnetism is not the measurable field, but rather the electromagnetic fields that arise from it. The gauge configurations must then be thought of as elements of an equivalence class, identifying all gauge fields related to each other by a gauge transformation. This is equivalent to imposing a *gauge fixing condition* $F(A) = 0$ at the level of the path

Figure 2.2: The new Feynman rule introduced by ghosts (left) shown by the dotted lines. However, ghosts can only be used as internal legs, since they do not correspond to physical states, and so can contribute to the gluon self-energy at 1-loop (right).

integral (for some gauge fixing function F) with the use of a delta function $\delta(F(a))$ [2]. This follows the Faddeev-Popov prescription [54] where the Jacobian arising from the δ-function gives rise to a *ghost lagrangian* [55]. A common choice for F is the R_ξ class of gauges, which amounts to the addition of:

$$-\frac{1}{2\xi}(\partial^\mu A^a_\mu)^2 \tag{2.9}$$

to the lagrangian. To fully pick a gauge we must also specify ξ. Common choices are the Landau gauge ($\xi \to 0$), Feynman-'t Hooft gauge ($\xi = 1$), or the unitary gauge ($\xi \to \infty$), though the result of any physically observable calculation should of course not depend on ξ. Accounting for the ghost lagrangian, the QCD lagrangian now reads:

$$\begin{aligned}\mathcal{L} = \sum_{f=1}^{N_f} \bar{\psi}^f_i (i\not{D} - m_f)_{ij} \psi^f_j - \frac{1}{4} F^a_{\mu\nu} F^{\mu\nu a} \\ - \frac{1}{2\xi}(\partial^\mu A^a_\mu)^2 + \partial_\mu \bar{c}^a \, \partial^\mu c^a + g f^{abc} (\partial^\mu \bar{c}^a) A^b_\mu c^c,\end{aligned} \tag{2.10}$$

where c^a are anti-commuting, Grassmann valued, Lorentz scalar (spin-0) fields known as *ghosts*. They thus violate the spin-statistics theorem, but are required as internal lines in order to remove unphysical gluon degrees of freedom: the longitudinal and timelike components. The Feynman rules they introduce are shown in the left panel of figure 2.2, but their unphysical nature restricts them to internal legs and so they contribute to the gluon self energy at 1-loop, shown in the right of the same figure. The abelian nature of $U(1)$ (the gauge group of QED) means the structure constants vanish and thus ghosts are not required for the quantization of electrodynamics.

[2] The Gauge slice corresponding to the solution to $F(A) = 0$ in general can intersect the Gauge orbits more than once and thus admits multiple solutions (representatives) for a given orbit. This problem is known as the Gribov ambiguity [53].

2.1.2 Renormalization and asymptotic freedom

When computing a scattering graph containing loops, the resulting amplitude are often divergent quantities. These infinities, called ultraviolet divergences, can be traced back to the undetermined loop momenta which must be integrated over. To handle these divergences, a *regulator* is required. Several choices of regulators exist [56, 57], but a popular one which we shall consider is *dimensional regularization* which analytically continues the spacetime dimension to $d = 4 - 2\epsilon$ [58, 59]. In doing so, the precise nature and form of the divergences can be isolated (for example often appearing as $1/\epsilon$ poles in the contribution to the scattering graph).

One then treats the original lagrangian (such as equation 2.2) as a *bare lagrangian* where the bare parameters are treated as formally infinite. To this bare lagrangian one then adds counter terms, whose coupling are set to cancel the divergences leaving behind a purely finite part. In this way, one then relates the parameters of the theory to physically observable quantities such as cross sections or decay rates. This process of relating to physical observables is known as *renormalization* and is a scheme dependent process (to which exact physical observable the parameters have been related to). For example, the on-shell renormalization scheme relates the mass parameter m_f of the lagrangian in equation 2.2 to the physical mass defined by the poles in the propagator. Alternatives are devised by theorists to help simplify calculations such as the minimal subtraction scheme which removes the divergence and nothing else; equivalently, the counter term has no finite part. The more popular modified minimal subtraction scheme ($\overline{\text{MS}}$) removes the divergence as well as the Euler-Mascheroni constant, γ, and $\log 4\pi$.

By construction then, the resulting observables are finite quantities. However, remnant from the renormalization process are artifacts of the regulator used. For example, in dimensional regularization, a mass parameter, μ, is introduced in order to ensure the coupling constant g remains dimensionless when going to $d = 4 - 2\epsilon$ dimensions. The dependence on the renormalization scale, μ, is of course arbitrary, since it did not exist when we defined our theory, and no physical quantity can therefore depend on it. However, owing to the fact that we work to a finite order in perturbation theory, dependence on this mass scale remains. We remove the dependence on μ by requiring that any physical quantity, R, cannot depend on this, arbitrary, parameter:

$$\mu^2 \frac{dR}{d\mu^2} = 0. \tag{2.11}$$

In much the same way in QCD the strong coupling defined as

$$\alpha_s = \frac{g^2}{4\pi} \tag{2.12}$$

gains a scale dependence on the energy scale, μ, due to the physicality constraint of equations similar to equation 2.11. This then gives a *running coupling* via the Renormalization Group Equation:

$$\mu^2 \frac{\partial \alpha_s(\mu^2)}{\partial \mu^2} = \beta\left(\alpha_s(\mu^2)\right). \tag{2.13}$$

The β-function [3] admits a power expansion in α_s:

$$\beta(\alpha_s(\mu^2)) = -\alpha_s^2(\mu^2)\left(\beta_0 + \beta_1\alpha_s(\mu^2) + \cdots\right) \tag{2.15}$$

which has been computed up to 5 loops [60–64]. At leading order:

$$\beta_0 = \frac{33 - 2N_f}{12\pi} \tag{2.16}$$

which means that the β function is negative at 1-loop if $N_f < 17$. For QCD, the number of flavours is $N_f = 6$ and so the QCD coupling gets smaller with increasing energy scale. To see this, we solve explicitly the resulting Renormalization Group equation to relate α_s at energy scale μ with α_s at μ_0:

$$\alpha_s(\mu^2) = \frac{\alpha_s(\mu_0^2)}{1 + \beta_0\alpha_s(\mu_0^2)\ln\frac{\mu^2}{\mu_0^2}}. \tag{2.17}$$

Thus we see that with $\beta_0 > 0$ the strong coupling asymptotically flows to a free theory. This property is known as *asymptotic freedom* [65, 66] and so QCD is said to be asymptotically free. Though μ is arbitrary and by construction we can set it to be whatever value we like, for use in calculations it is reasonable to set it around the energy scale of the process in question, Q. The reason for this is that one in general has terms of the form $\ln \mu/Q$ which systematically appear as factors in front of α_s

[3] In the case of renormalizing an operator in the lagrangian, such as the mass term $m\bar{\psi}\psi$, one would have an anomalous dimension instead of a β-function:

$$\mu^2 \frac{\partial m}{\partial \mu^2} = -\gamma_m(\mu^2)m \tag{2.14}$$

so-named because it alters the naive classical scaling dimension.

order-by-order in perturbation theory. These so-called large-logarithms can be made arbitrarily large and can thus spoil the validity of truncating the perturbative series. In setting $\mu \sim Q$, one may resum these large logs thereby preserving the perturbative approximation.

Note further that there is an intrinsic scale at which the strong coupling diverges, the so-called *Landau pole*:

$$\ln \Lambda_{\mathrm{QCD}} = \ln \mu_0^2 - \frac{1}{\beta_0 \alpha_s(\mu_0^2)}. \tag{2.18}$$

The scale $\Lambda_{\mathrm{QCD}} \sim 200 - 500$ MeV separates the energy scale between perturbative and non-perturbative QCD. This characteristic energy scale has appeared after quantization without any dependence on the quark masses; indeed in the massless theory this energy scale would still emerge, a phenomenon known as *dimensional transmutation*. The asymptotic freedom of QCD plays a key role for phenomenology at particle accelerators: with energy scales far surpassing Λ_{QCD} one is able to make perturbative calculations at kinematic regions of phase space that may otherwise not have been possible.

2.2 Collinear factorization theorem

Asymptotic freedom implies that perturbative QCD calculations are only valid at high energies well beyond the Landau pole Λ_{QCD}. Otherwise a truncation of the perturbative expansion is not valid since the subsequent terms will not necessarily be subleading. The coupling constant is therefore not a valid expansion parameter in such regimes and for low energy objects, such as for example the internal structure of hadronic bound states, a perturbative understanding proves difficult; though work has been done for an expansion in the number of flavours [67, 68].

The factorization theorems, however, provide an incredibly important and powerful result for physics involving long distance quantities, such as scattering amplitudes involving initial state hadrons. They state that a general observable decomposes into a long distance part and a short distance part. The short distance part is computable using perturbative QCD (pQCD) while the long distance component is parameterized into what are known as parton distribution functions (PDFs). These PDFs are universal and process independent and as such can be obtained from data using a global QCD analysis.

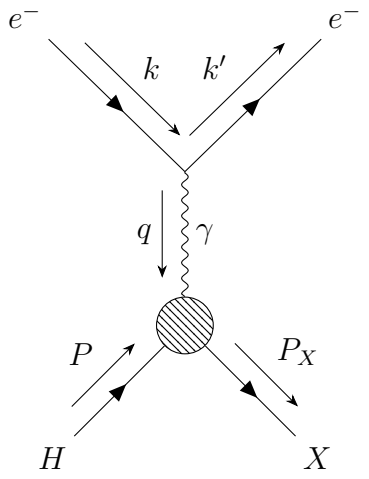


Figure 2.3: The Born level diagram for deep inelastic scattering of an electron and hadron mediated by virtual photon exchange.

In this section we consider how PDFs arise naturally in the parton model when considering lepton-hadron scattering experiments. We shall also show how their running with scale arises at next-to-leading order in QCD: a result that forms a major triumph for the theory.

2.2.1 Deep inelastic scattering

To motivate the factorization theorems, we begin by considering one of the simplest processes in QCD: deep inelastic scattering (DIS). For this observable, protons are collided with leptons (for simplicity suppose electrons) where we assume a kinematic regime at sufficiently high energies such that the proton fragments into some final state products X (which differentiates DIS from elastic compton scattering $e^- p^+ \to e^- p^+$). The Born level diagram is shown in figure 2.3 for photon mediated DIS; the only diagram that contributes at this order being the t-channel exchange. We label the initial (final) state electron to have 4-momentum k (k'), while the initial state proton (final state hadrons) has 4-momentum P (P_X). The amplitude for this process can be computed in the Feynman gauge to be:

$$i\mathcal{M} = (-ie)\bar{u}(k')\gamma^\mu u(k)\left(-\frac{i\eta_{\mu\nu}}{q^2}\right)\langle X|j_h^\nu(0)|H\rangle \tag{2.19}$$

where e is the electron electric charge, u ($\bar{u}$) are the positive (negative) frequency Dirac spinor solutions and j_h^ν is the hadronic electromagnetic current. We do however

quickly reach an impasse when trying to compute this amplitude. The matrix element $\langle X|j_h^\nu(0)|H\rangle$ contains hadronic states $|H\rangle$ and $|X\rangle$ as initial and final states which have been known to be composite bound states much before deep inelastic experiments. In the language of the second quantization, these are to be understood as the creation operators of many matter and gauge fields acting on the vacuum to create a non-perturbative bound state. Our ignorance lies in our inability to write down an expression for these Hilbert space elements. Perturbation theory calculations of S-matrix elements rely on the initial and final states of the scattering experiments at times $\pm\infty$ be pure states in the Fock space, such that one may then use Wick's theorem to eventually annihilate the vacuum. Clearly we cannot do this unless we have a firm understanding on the precise nature of the hadronic state.

We can, however, still make good progress by computing the squared matrix element. Using the fact that our experiment is unpolarized (insensitive to spin alignments) and so our matrix element lives inside a spin sum average as well as a phase space integral:

$$|\mathcal{M}|^2 = \frac{e^2}{4q^4} \sum_{\text{spins}} \sum_{X,P_X} \Big[\bar{u}(k')\gamma_\mu u(k)\, \bar{u}(k)\gamma_\nu u(k')\Big] \cdot \Big[\, \langle X|j_h^\mu(0)|H\rangle \, \langle H|j_h^{\nu\dagger}(0)|X\rangle \,\Big] (2\pi)^4 \delta^{(4)}(P_X - P - q). \quad (2.20)$$

We now define the leptonic tensor which may be calculated straight forwardly:

$$\begin{aligned} L_{\mu\nu} &= \frac{1}{2} \sum_{\text{spins}} \bar{u}(k')\gamma_\mu u(k)\, \bar{u}(k)\gamma_\nu u(k') \\ &= \frac{1}{2} \mathrm{Tr}\left(\not{k}'\gamma_\mu \not{k}\gamma_\nu\right) \\ &= 2\left(k_\mu k'_\nu + k_\nu k'_\mu - \eta_{\mu\nu} k\cdot k'\right). \end{aligned} \quad (2.21)$$

The hadronic tensor, however, encapsulates the hadronic interaction:

$$W^{\mu\nu} = \frac{1}{2} \sum_{X,P_X} \langle X|j_h^\mu(0)|H\rangle \, \langle H|j_h^{\nu\dagger}(0)|X\rangle \, (2\pi)^4 \delta^{(4)}(P_X - P - q) \quad (2.22)$$

such that the matrix element for the entire process is given by contracting the two tensors $L_{\mu\nu}W^{\mu\nu}$. At this point, it is worth defining some Lorentz invariant kinematical

quantities that are ubiquitous in DIS discussions:

$$Q^2 = -q^2 = -(k - k')^2 \tag{2.23}$$

$$x = \frac{Q^2}{2P \cdot q} \tag{2.24}$$

$$y = (q \cdot P)/(k \cdot P). \tag{2.25}$$

The quantity Q^2 is understood as the exchanged momentum from the lepton to the proton, Bjorken-x has the interpretation of being the fraction of proton momentum carried by a struck constituent (this interpretation is not obvious and follows after manipulating the 4-momentum conserving δ-function of equation 2.29 with the parton model described below) and y is the inelasticity. We can use these quantities to parameterize the Lorentz structure of the hadronic tensor. We note that the Ward identity requires $q_\mu W^{\mu\nu} = q_\nu W^{\mu\nu} = 0$ and that the only 4-vectors which we can use are P and q since the others are integrated over in the phase space integral. Furthermore, the hadronic tensor is symmetric in its Lorentz indices and so we may write:

$$W^{\mu\nu} = \left(-\eta^{\mu\nu} + \frac{q^\mu q^\nu}{q^2}\right) F_1(x, Q^2) + \frac{1}{P \cdot q}\left(P^\mu - \frac{P \cdot q}{q^2} q^\mu\right)\left(P^\nu - \frac{P \cdot q}{q^2} q^\nu\right) F_2(x, Q^2) \tag{2.26}$$

where F_1 and F_2 are known as the *structure functions*. In the case of the full neutral current contribution where there is a Z-boson mediated diagram, there will also be a parity-violating structure function F_3. It is useful to define the longitudinal structure function:

$$F_L(x, Q^2) = F_2(x, Q^2) - 2xF_1(x, Q^2) \tag{2.27}$$

which has the interpretation of parameterizing the proton's ability to interact with a virtual photon whose polarization is longitudinal with respect to the beam axis.

2.2.2 The parton model

To proceed further we use the *parton model* [69, 70] first introduced by Feynman to be able to elucidate the surprising results of early DIS experiments. The principal assumption here is that the proton is a bound state of essentially free *partons*. Accompanying each such parton is a *parton distribution function* (PDF) $f_i(\xi)$, which has the classical interpretation of being the number density for partons of species i carrying a fraction ξ of the total hadron momentum. With the success of the quark model in explaining

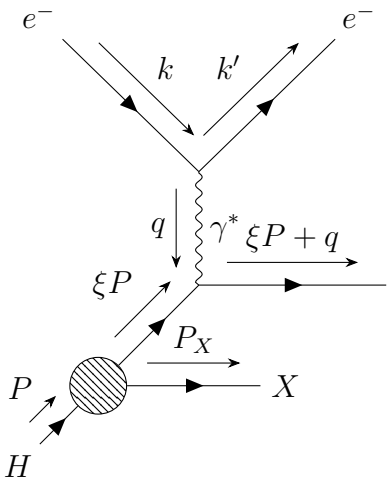


Figure 2.4: Born level diagram for deep inelastic scattering in the parton model. The gauge boson interacts with one constituent parton in proton.

hadron properties, such as the eightfold way [71] and the asymptotic freedom of QCD, the partons quickly became identified with the quarks and gluons of QCD, as well as more formally any other SM particle [72–75].

The justification for the parton model is that for DIS experiments $Q \gg \Lambda_{\mathrm{QCD}}$ while the typical time scale of momentum transfer between partons within the proton is $\Lambda_{\mathrm{QCD}}^{-1}$. This time scale is much slower than the time scale probed by the gauge boson and so, in effect, the proton appears frozen at the instant of the interaction, the gauge boson striking, therefore, only one constituent.

Nowadays the PDFs can be defined in a quantum field theoretic way using the operator product expansion [76–78] which we present for completeness. The quark PDF may be expressed as the matrix element of the quark number operator on proton states [4]:

$$f_i(\xi) = \int \frac{dy^-}{4\pi} e^{-i\xi P^+ y^-} \langle P|\bar{\psi}(0, y^-, \mathbf{0}_T) W[y, 0] \gamma^+ \psi_i(0)|P\rangle \tag{2.28}$$

where the superscript $+$ and $-$ refer to light-cone coordinates, ψ is the (renormalized) quark fields, and $W[y, 0]$ is a Wilson line (path ordered exponential of the gluon field) along the light-like straight line from the point 0 to $(0, y^-, \mathbf{0}_T)$. Such a definition is useful for formal proofs of the factorization theorems, of which the parton model is a leading order approximation, however, this formal definition will not be of further interest or use for our discussion.

[4] The gluon PDF can also be expressed by a similar expression.

The born level diagram for DIS under the parton model is shown in figure 2.4, whereby one constituent parton has interacted with the gauge boson. With the help of the parton model we can complete the calculation of the full DIS cross section, σ. We do so by computing a *partonic cross section*, $\hat{\sigma}(e^- p_i \to e^- X)$ [5] , for a parton p_i having 4-momentum ξP^μ. The electron-hadron cross section is then given by averaging over parton momenta and flavours:

$$\sigma(e^- P \to e^- X) = \sum_i \int_0^1 d\xi \, f_i(\xi) \hat{\sigma}(e^- p_i \to e^- X). \tag{2.29}$$

A formal proof of equation 2.29 is possible [79], with higher order corrections, known as *higher twists*, being found to be suppressed by powers of Q. Though proofs exist for DIS and other processes such as Drell-Yan, a process-independent proof of the so-called *collinear factorization theorem* does not exist, though is often simply assumed with resounding success.

The proton PDFs must satisfy the following *sum rules*. The first is the *valence sum rule* and states that up-valence distribution must integrate to 2, while the down-valence distribution must integrate to unity and the valence strange distribution to zero in order to satisfy the proton quantum numbers:

$$\int_0^1 d\xi \, (f_u(\xi) - f_{\bar{u}}(\xi)) = 2, \tag{2.30}$$

$$\int_0^1 d\xi \, (f_d(\xi) - f_{\bar{d}}(\xi)) = 1, \tag{2.31}$$

$$\int_0^1 d\xi \, (f_s(\xi) - f_{\bar{s}}(\xi)) = 0, \tag{2.32}$$

as well as the fact that the constituent momenta must integrate to give the parent hadron momenta giving the so-called *momentum sum rule*:

$$\int_0^1 d\xi \, \sum_i \xi f_i(\xi) = 1, \tag{2.33}$$

where we highlight that the summation is over quarks, anti-quarks, as well as gluons. As a final remark before continuing the main discussion, we note that the approximate $SU(2)$ isospin symmetry relates the neutron PDFs to the proton PDFs: $f_u^n = f_d$ and $f_d^n = f_u$.

[5] By convention quantities with a hat are partonic.

Inspired by the parton model, we can attempt to compute the DIS process for the partonic case, which we then relate to the full hadron-lepton scattering through equation 2.29. Indeed, asymptotic freedom tells us that at the high energies of DIS, the coupling will be small and thus perturbation theory holds true for the hard matrix element calculation. Doing so, and after a lengthy computation, one arrives at the equations for the DIS structure functions:

$$F_1(x, Q^2) = \frac{1}{2} \sum_i e_i^2 f_i(x) \tag{2.34}$$

$$F_2(x, Q^2) = \sum_i x e_i^2 f_i(x). \tag{2.35}$$

We see that at leading order:

$$F_L = F_2 - 2xF_1 = 0 \tag{2.36}$$

which is the *Callan-Gross equation* [80] and confirms that the partons are spin-1/2 particles. The equivalent relation, had they been Lorentz scalars instead, would be $F_2(x, Q^2) = 0$. Importantly, note that the structure functions of equations 2.34 and 2.35, though able to depend on Q, in fact do not. This property is known as *Bjorken scaling*, the breaking of which has been both observed by experiments and predicted by QCD.

Finally, we conclude by introducing the Mellin convolution, which for arbitrary functions, f and g, is given by:

$$(f \otimes g)(x) = \int_x^1 \frac{dy}{y} f(y) g\left(\frac{x}{y}\right). \tag{2.37}$$

The Mellin convolution is used to relate partonic quantities (such as the hadronic tensor, $\hat{W}^{\mu\nu}$, or structure functions, $\hat{F}$) with the full hadronic quantity. For example, the partonic structure functions C^i (often referred to as coefficient functions) can be computed within pQCD and are related to the full F_2 structure function by convolving with the PDFs:

$$F_2(x, Q^2) = \sum_i f_i \otimes C^i + \mathcal{O}\left(\frac{\Lambda_{\mathrm{QCD}}^2}{Q^2}\right), \tag{2.38}$$

where the higher-twists are suppressed by inverse powers of Q^2 and we assume to be probing a kinematical regime where $Q \gg \Lambda_{\mathrm{QCD}}$.

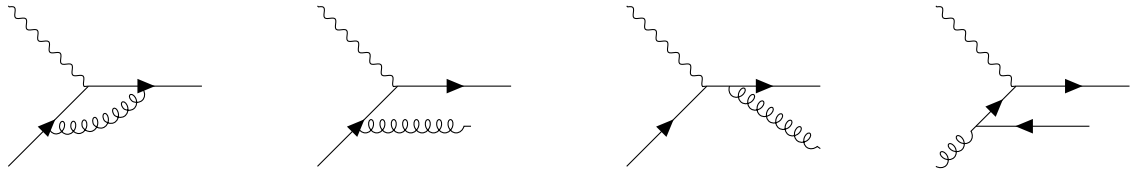


Figure 2.5: Next-to-leading order in α_s contributions to DIS. From left to right we have: virtual emission, real emission as initial state radiation, final state radiation, and gluon-boson fusion.

2.2.3 QCD corrections to the parton model

With the remarkable ability of the parton model to explain the early DIS measurements, one is naturally led to ask whether this behaviour is true for higher values of Q and to what extent does the picture change once next-to-leading order (NLO) QCD corrections are applied. At NLO in QCD 4 additional diagrams must be accounted for as shown in figure 2.5. These correspond to virtual emission, whereby the initial state quark emits a gluon that is absorbed by the final state quark; initial (final) state radiation by real gluon emission from the initial (final) state quark; and gluon-boson fusion whereby the splitting process $g \to q\bar{q}$ results in one of the quarks interacting with the mediating boson. Note that since we are considering fully inclusive DIS whereby we integrate over the phase space of all possible final products, final state radiation is a valid diagram to consider. We mention here that, at the squared amplitude level, the virtual emission graph must be paired with the Born graph, while the others can pair with each other to give a correction of $\mathcal{O}(\alpha_s^2)$ to the cross section.

Each diagram possesses divergences of different flavours. The virtual emission results in UV divergences owing to divergent momenta contained within the loop. Moreover, IR divergences occur due to soft and collinear emission of quarks and gluons; that is, gluons that have zero transverse momentum relative to the parent particle, and soft divergences owing to the massless nature of gluons. Isolating these divergences requires the use of a regulator such as dimensional regularization.

The UV divergence can be handled using renormalization techniques, for example, using the $\overline{\text{MS}}$ subtraction scheme. The IR divergences, however, require more care. The Kinoshita-Lee-Nauenberg (KLN) theorem [81, 82] (as well as the related Bloch-Nordsieck theorem [83] though violated by QCD [84]) state that so long as the phase space integral over all degenerate final (and initial) states is performed, then the infrared divergences cancel, so long as one is computing an *infrared safe* quantity, after summing all the Feynman diagrams at a given perturbative order. An n-particle

quantity is infrared safe if the observable is equal to the analogous $(n-1)$-particle quantity when any pair of particles become collinear (have parallel momenta). For deep inelastic scattering, however, we do not sum over all degenerate initial states, instead isolating a particular proton state of definite momentum. As such the cancellation of the collinear singularities is not protected by the KLN theorem and in fact persists after all the diagrams of figure 2.5 are added together.

This remnant collinear singularity is an incredibly important result. In fact, once the dust has settled (and for now omitting the gluon-boson fusion diagram):

$$F_1(x,Q^2) = \frac{1}{2}\sum_i e_i^2 \int_x^1 \frac{dy}{y} f_i(y) \left[\delta\left(1-\frac{x}{y}\right) - \frac{\alpha_s}{2\pi}\left(P_{qq}\left(\frac{x}{y}\right)\left(\frac{1}{\epsilon} + \log 4\pi - \gamma + \log\frac{\mu^2}{Q^2}\right) + R_{qq}\left(\frac{x}{y}\right)\right)\right] \quad (2.39)$$

where P_{qq} is known as the quark-quark splitting function (given below), γ is the Euler-Mascheroni constant, R_{qq} is the process dependent quark-quark remainder function, and μ is the arbitrary mass parameter first introduced in dimensional regularization to maintain that the coupling is dimensionless. The $1/\epsilon$ pole is the aforementioned collinear singularity, isolated using dimensional regularization and its removal is done by analogy to the renormalization group. We note the left hand side of equation 2.39 is a measurable physical quantity which should thus be finite. The right hand side, however, has a divergent pole upon taking $\epsilon \to 0$. The PDF, however, is not a measurable quantity, fundamentally defined by the matrix element of equation 2.28 which in the framework of QFT is no more measurable than the couplings are: only cross section and decay rates are measurable entities. We thus treat the PDFs in equation 2.39 as *bare PDFs*, with the $\overline{\text{MS}}$ *renormalized PDFs* (referred to henceforth as simply PDFs) defined by a collinear subtraction:

$$f_i^{\overline{\text{MS}}}(x,\mu_F) = \int_x^1 \frac{dy}{y} f_i(y) \left[\delta\left(1-\frac{x}{y}\right) - \frac{\alpha_s}{2\pi} P_{qq}\left(\frac{x}{y}\right)\left(\frac{1}{\epsilon} - \log\frac{\mu^2}{\mu_F^2} - \gamma + \log 4\pi\right)\right] \quad (2.40)$$

where μ_F is an arbitrary mass (or energy scale) parameter called the *factorization scale* and plays the same role as the renormalization scale for the strong coupling running. Its value too is arbitrary and can be set to anything independently of the renormalization scale, though again it is beneficial to set it equal to the typical energy scale of the process in question: $\mu_F \approx Q$. The resulting structure function thus reads

(dropping the $\overline{\text{MS}}$ superscript):

$$F_1(x,Q^2) = \frac{1}{2}\sum_i e_i^2 \int_x^1 \frac{dy}{y} f_i(x,\mu_F)\left[\delta\left(1-\frac{x}{y}\right) + \frac{\alpha_s}{2\pi}P_{qq}\left(\frac{x}{y}\right)\log\frac{Q^2}{\mu_F^2}\right] \tag{2.41}$$

and is a perfectly finite quantity. The μ_F dependence drops out of this expression due to the PDF evolution discussed below and the logarithm in Q is responsible for the Bjorken scaling violation.

By differentiating equation 2.40, the μ_F dependence of the PDFs is then given by the Dokshitzer-Lipatov-Gribov-Altarelli-Parisi (DGLAP) integro-differential equation [85–87]:

$$\mu_F\frac{d}{d\mu_F}f_i(x,\mu_F) = \frac{\alpha_s}{\pi}\int_x^1 \frac{dy}{y} f_i(y,\mu_F)P_{qq}\left(\frac{x}{y}\right) + \mathcal{O}(\alpha_s^2) \tag{2.42}$$

which, after reintroducing the gluon-boson fusion diagram, causes quark-gluon mixing upon DGLAP evolution by the matrix equation:

$$\mu_F\frac{d}{d\mu_F}\begin{pmatrix} f_i(x,\mu_F) \\ f_g(x,\mu_F)\end{pmatrix} = \frac{\alpha_s}{\pi}\sum_j\int_x^1\frac{dy}{y}\begin{pmatrix} P_{q_iq_j} & P_{q_ig} \\ P_{gq_j} & P_{gg}\end{pmatrix}\begin{pmatrix} f_j(y,\mu_F) \\ f_g(y,\mu_F)\end{pmatrix} + \mathcal{O}(\alpha_s^2). \tag{2.43}$$

The splitting functions (or Altarelli-Parisi kernels) P, are known up to NNLO [88, 89]. Under the assumption of $SU(N_f)$ isospin symmetry and charge conjugation invariance they do not depend on quark flavour and are the same for both quarks and anti-quarks [90]. At LO they read [91]:

$$P_{qq}(z) = \frac{4}{3}\left[\frac{1+z^2}{(1-z)_+} + \frac{3}{2}\delta(1-z)\right], \tag{2.44}$$

$$P_{qg}(z) = \frac{4}{3}\left[\frac{1+(1-z)^2}{z}\right], \tag{2.45}$$

$$P_{gq}(z) = \frac{1}{2}\left[z^2+(1-z)^2\right], \tag{2.46}$$

$$P_{gg}(z) = 6\left[\frac{1-z}{z} + \frac{z}{(1-z)_+} + z(1-z) + \left(\frac{11}{12} - \frac{N_f}{18}\right)\delta(1-z)\right] \tag{2.47}$$

where the plus distribution is defined by its behaviour within an integral. For two functions f and g:

$$\int_0^1 dz[g(z)]_+ f(z) = \int_0^1 dz\, g(z)\left(f(z)-f(1)\right). \tag{2.48}$$

In order to solve the DGLAP evolution equations one works in the maximally diagonal basis known as the *evolution basis*. First we define:

$$f_i^{\pm} = f_i \pm \bar{f}_i \tag{2.49}$$

which defines the valence distributions:

$$V_i = f_i^- \qquad i = 1, \ldots, N_f \tag{2.50}$$

and the triplet distributions:

$$T_3 = u^+ - d^+ \tag{2.51}$$
$$T_8 = u^+ + d^+ - 2s^+ \tag{2.52}$$
$$T_{15} = u^+ + d^+ + s^+ - 3c^+ \tag{2.53}$$
$$T_{24} = u^+ + d^+ + s^+ + c^+ - 4b^+ \tag{2.54}$$
$$T_{35} = u^+ + d^+ + s^+ + c^+ + b^+ - 5t^+. \tag{2.55}$$

Together the valence and the triplet distributions form what is known as the *non-singlet sector* and each evolve independently of the other according to equation 2.42. The singlet distribution:

$$\Sigma = \sum_{i=1}^{N_f} f_i^+, \tag{2.56}$$

however, evolves according to the coupled DGLAP evolution of equation 2.43 and couples with the gluon.

Thus far we have been discussing the DGLAP evolution equations in the so-called x-space. It is, however, not obvious how to go about solving these rather complicated systems of equations. We do note that throughout this discussion, the Mellin convolution of equation 2.37 has been used throughout, though not been made explicit: the non-singlet evolution of equation 2.42 is nothing but the convolution of the PDF with the splitting functions. However, by performing the *Mellin transformation*:

$$f_i(N, \mu_F) = \int_0^1 dx\, x^{N-1} f_i(x, \mu_F), \tag{2.57}$$

whereby the difference between Mellin-space and x-space objects is manifest only in terms of their arguments, the convolution can be made into a simple product. For

example, in the case of non-singlet evolution at leading order in QCD:

$$\mu_F \frac{d}{d\mu_F} f_{NS}(N, \mu_F) = \frac{\alpha_s(\mu_F)}{\pi} \gamma_{qq}^{NS}(N) f_{NS}(N, \mu_F), \tag{2.58}$$

where the Mellin transform of the non-singlet splitting function (often referred to as the anomalous dimension) is given by the quantity γ_{qq}^{NS} by convention. We choose $\mu_R = \mu_F = Q$ in order to resum large logarithms. The Mellin-space evolution is more easily solved analytically with the inverse transformation being much more amenable to numerical evaluation than solving the x-space evolution directly.

Finally, we mention that the Callan-Gross equation of equation 2.36 is violated at NLO in QCD [90] with the non-zero nature being a measured phenomenon at HERA [92].

2.2.4 Universality of parton distributions

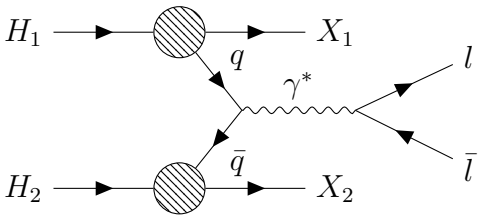


Figure 2.6: The tree level diagram for the Drell-Yan process involving two initial state hadrons. Under factorization, two partons are ejected from the hadrons which annihilate to a virtual photon which decays further into a dilepton pair. The hadron fragments, X_1 and X_2 are integrated over in the phase space integral and are treated as beam remnants.

So far we have motivated the parton model by considering lepton-hadronic scattering in the context of DIS. In this section we extend the discussion to the scenario where two hadrons are present in the initial state for example at the Tevatron or the LHC in fixed-target or collider hadron-hadron experiments. A ubiquitous process at such experiments is Drell-Yan (DY) shown schematically at Born level in figure 2.6. In this process, under factorization, two partons, ejected from the hadrons, interact by annihilating to a virtual photon which then decays into an on-shell dilepton pair. The obvious extension of equation 2.29 is thus to simply include two parton distribution

functions in the convolution, one for each initial state hadron:

$$\sigma\left(H_1H_2 \to l\bar{l} + X_1X_2\right) = \sum_{ij} \int_0^1 \int_0^1 d\xi_1 d\xi_2 \, f_i^{H_1}(\xi_1, \mu_F^2) f_j^{H_2}(\xi_2, \mu_F^2) \, \hat{\sigma}(p_i p_j \to l\bar{l}) + \mathcal{O}\left(\frac{\Lambda_{\rm QCD}^2}{M^2}\right). \quad (2.59)$$

which again can be more rigorously proven using field theoretic arguments with the collinear factorization theorem of DY processes [93–95]. To this end it is useful to define the *parton luminosity*:

$$\mathcal{L}_{ij} = f_i \otimes f_j \quad (2.60)$$

such that the cross sections are given by convolving the luminosity with the process dependent coefficient functions, C_{ij}, for partonic channels i and j:

$$\sigma = \sum_{ij} \mathcal{L}_{ij} \otimes C_{ij}. \quad (2.61)$$

We emphasize here that the PDFs in equation 2.59 are precisely the same as those of equation 2.40 which we used when considering deep inelastic scattering (so long as the initial state hadrons are the same e.g protons). In this way PDFs are *universal* objects, parameterizing the long distance physics of the proton structure, while the process dependent hard cross section is computed in pQCD for any observable we wish to construct. While the scale dependence of PDFs are determined by the DGLAP evolution equations, the Bjorken-x dependence cannot be computed perturbatively and so must be fitted to experimental data using a global QCD analysis. This shall be discussed further in chapter 3, but for now we remark that the power of PDF universality is used to employ data from a host of various processes to constrain the parton distributions.

2.3 Effective field theories

We now turn our attention to discuss the second important theoretical framework relevant for this text: the notion of effective field theories [51, 96, 97]. A remarkable fact of nature is that its phenomena disassociate from one another at varying length (energy) scales. Indeed, it is quite peculiar that few is the scholarly correspondence between the particle physicist and, for example, the biologist; the latter of whom has

little concern for the precise nature of fundamental particles and their interactions, despite the fact that their field of study is emergent from precisely these fundamental building blocks.

Such a fact can be understood through the lense of effective field theory (EFT). An effective field theory is a quantum field theory which we maintain is valid only to a certain energy scale, beyond which its reliability falls apart. We shall see how we can construct such a theory in the IR by integrating out heavy modes from an UV theory. The IR and UV theory both agree on low energy phenomena, far below the heavy mass scale, despite the IR theory having no heavy mode explicit in its lagrangian. Such a construction is thus *top down* whereby the UV completion is known (or assumed) and the IR is constructed from it.

We then see how this naturally leads us to consider all possible UV completions that flow to the Standard Model under Wilsonian renormalization group flow. This allows us to construct the Standard Model Effective Field Theory, a *bottom up* approach where by the UV completion is not known, but the effects of which manifest in the IR through the presence of higher dimensional operators.

2.3.1 The Wilsonian effective action

To motivate the concept of effective field theory, we start with a pedagogical example. Consider a scalar field, ϕ with mass m and a vastly heavier scalar field, Φ which has mass $M \gg m$. We assume this is the field content of the UV theory $\mathcal{L}_{\text{UV}}$ which is able to resolve phenomena at all scales. However, if our experimental apparatus has a resolution comparable to m, then the precise nature of the interactions between ϕ and Φ is irrelevant, since Φ will in general be very off-shell and thus very short lived.

If one considers the generating functional, $Z[J_\phi, J_\Phi]$, for the UV theory

$$Z_{\text{UV}}[J_\phi, J_\Phi] = \int D\phi\, D\Phi\, e^{\left(iS + i\int d^4x \left(J_\phi \phi + J_\Phi \Phi\right)\right)} \tag{2.62}$$

then we can construct the dynamics of the IR theory, by *integrating out* the heavy field Φ. Since we are not interested in the dynamics of the heavy field, but rather processes involving only correlation functions of the light scalar, we set the associated current, J_Φ, to 0 without loss of generality. By formally performing the integration over the

heavy scalar first:

$$Z[J_\phi, J_\Phi = 0] = \int D\phi\, e^{\int d^4x J_\phi \phi} \left(\int D\Phi\, e^{iS} \right) \tag{2.63}$$

$$\equiv \int D\phi\, e^{\int d^4x J_\phi \phi} e^{iS_{\rm IR}} \tag{2.64}$$

we can define the Wilsonian effective action, $S_{\rm IR}$, which governs the dynamics of the IR. By performing the integration in this way, we capture entirely all diagrams containing the heavy field as internal legs of Feynman diagrams. The effective action admits a perturbative expansion in powers of $\hbar$

$$S_{\rm IR} = S_{\rm IR}^{(0)} + \hbar S_{\rm IR}^{(1)} + \mathcal{O}(\hbar^2), \tag{2.65}$$

with higher order quantum corrections arising from heavy scalar loops.

The quantity $S_{\rm IR}$ is then the Wilsonian effective action, a coarse grained, less detailed, description of the UV theory, but more appropriate for use in the computation of physical observables at the light scale, m. The tree-level contribution, $S_{\rm IR}^{(0)}$, describes all tree-level amplitudes involving external light scalars, while the NLO term $S_{\rm IR}^{(1)}$, describes all amplitudes containing 1-loop corrections and so on.

To compute the effective action, we expand the integral of equation 2.64 using the saddle point approximation. By Wick rotating, the exponential is well peaked at the classical field configuration allowing for a perturbative expansion to be made. Letting Φ_c denote the classical equation of motion satisfying the functional derivative equation:

$$\left.\frac{\delta S}{\delta \Phi}\right|_{\Phi_c} = 0 \tag{2.66}$$

the saddle point approximation decomposes the heavy field to be $\Phi = \Phi_c + \eta$ where η are the quantum fluctuations that we force to vanish at the boundaries. The integral of equation 2.64 expands to read:

$$e^{iS_{\rm IR}} = e^{iS[\Phi_c]} \int D\eta\, e^{\frac{i}{2}\eta \left.\frac{\delta^2 S}{\delta \eta^2}\right|_{\Phi_c} \eta} \tag{2.67}$$

which can be evaluated using standard Gaussian integral results to give:

$$e^{iS_{\rm IR}} = C \cdot e^{iS[\Phi_c]} \det \left(-\frac{\delta^2 S}{\delta \eta^2} \right)\Bigg|_{\Phi_c} + \dots \tag{2.68}$$

where the ellipses refer to higher loop corrections and C is a constant which does not affect the physical dynamics. Simplifying this expression gives us the IR effective action at 1-loop:

$$S_{\text{IR}} = S[\Phi_c] + \frac{i}{2}\text{Tr}\log\left(-\frac{\delta^2 S}{\delta\eta^2}\right) \tag{2.69}$$

where the functional trace is taken over the space of field configurations as well as any internal indices, such as spin or colour, that the quantum field being integrated over possesses.

We see then that the task of computing the tree-level effective action is as simple as substituting the classical equation of motion into the UV action. The one-loop matching and above requires the non-trivial task of computing the functional trace which we shall not burden ourselves with but merely acknowledge that it can be done with some computational effort [98].

2.3.2 Tree level matching

As a pedagogical example, consider the UV theory governed by the lagrangian:

$$\mathcal{L}_{\text{UV}} = \frac{1}{2}\partial^\mu\phi\partial_\mu\phi - \frac{1}{2}m^2\Phi^2 + \frac{1}{2}\partial^\mu\Phi\partial_\mu\Phi - \frac{1}{2}M^2\phi^2 - \frac{\lambda_0}{4!}\phi^4 - \frac{\lambda_1}{2}M\phi^2\Phi \tag{2.70}$$

where as before the scalar fields ϕ and Φ have masses m and M with the latter being considered heavy ($M \gg m$). We construct the IR theory related to this UV theory by integrating out the heavy degree of freedom at tree-level. What will happen, as should be relatively self-evident, is that the IR lagrangian will contain within it novel light scalar interactions that were not present in the full UV theory. This can be thought of as point like interactions whereby all heavy scalar propagators are infinitely short lived.

The classical equation of motion for the heavy scalar field Φ satisfies:

$$\left(\Box + M^2\right)\Phi_c = -\frac{\lambda_1}{2}M\phi^2 \tag{2.71}$$

where we follow the notation that the d'Alembert operator, $\Box$, denotes the contraction of the derivative with itself: $\Box = \partial^\mu\partial_\mu$. The subscript, Φ_c, denotes that this is the solution of the classical equations of motion and is to be understood as a function of the light scalar field.

We can formally invert the Klein-Gordon operator to obtain the solution for the classical field:

$$\Phi_c(\phi) = -\frac{\lambda_1}{2} M \left(\Box + M^2\right)^{-1} \phi^2. \tag{2.72}$$

Following the result from the saddle-point approximation, the tree-level effective action is given by simply substituting the classical solution into the UV Lagrangian:

$$\begin{aligned}\mathcal{L}_{\rm IR}^{(0)} = \mathcal{L}_{\rm UV}(\phi, \Phi_c(\phi)) &= \frac{1}{2}\partial^\mu\phi\partial_\mu\phi - \frac{1}{2}m^2\phi^2 - \frac{\lambda_0}{4!}\phi^4 - \frac{\lambda_1}{2}M\phi^2\Phi_c(\phi) \\ &= \frac{1}{2}\partial^\mu\phi\partial_\mu\phi - \frac{1}{2}m^2\phi^2 - \frac{\lambda_0}{4!}\phi^4 - \frac{\lambda_1^2}{4}M^2\phi^2\left(\Box + M^2\right)^{-1}\phi^2.\end{aligned} \tag{2.73}$$

The step that follows is the one that one would naively expect; namely expanding non-local $(\Box + M^2)^{-1}$ operator according to its Taylor expansion. We can justify this with the argument that the $\Box$ operator will lead to terms proportional to the squared 4-momenta of the light scalar field, which we assume to be far smaller than the heavy mass, M. A more rigorous derivation follows the covariant derivative expansion (CDE) [98]. Thus to leading order in the inverse mass expansion, the IR theory reads [6]

$$\mathcal{L}_{\rm IR}^{(0)} = \frac{1}{2}\partial^\mu\phi\partial_\mu\phi - \frac{1}{2}m^2\phi^2 - \frac{1}{4!}\left(\lambda_0 - 3\lambda_1^2\right)\phi^4 - \frac{\lambda_1^2}{6M^2}\phi^3\Box\phi + \mathcal{O}(M^{-4}). \tag{2.74}$$

We see that in the low energy limit, not only has the ϕ^4 coupling been corrected by the vertex coupling the light and heavy degrees of freedom, but a new interaction altogether, pertaining to the dimension 6 operator $\phi^3\Box\phi$, has been generated as a result of integrating out the heavy scalar. This lagrangian can by simplified further using the equations of motion for the ϕ field [7]. Indeed, shifting the lagrangian by a term proportional to the classical equations of motion has been shown to not change the S-matrix, even at the loop level [99]. The dynamics of the lagrangian above is

[6] We have used integration by parts to obtain $\phi^2\Box\phi^2 = \frac{4}{3}\phi^3\Box\phi$.

[7] Which gives the identity

$$\frac{1}{M^2}\phi^3\Box\phi = -\frac{m^2}{M^2}\phi^4 - \left(\frac{\lambda_0 - 3\lambda_1^2}{6M^2}\right)\phi^6 + \mathcal{O}(M^{-4}).$$

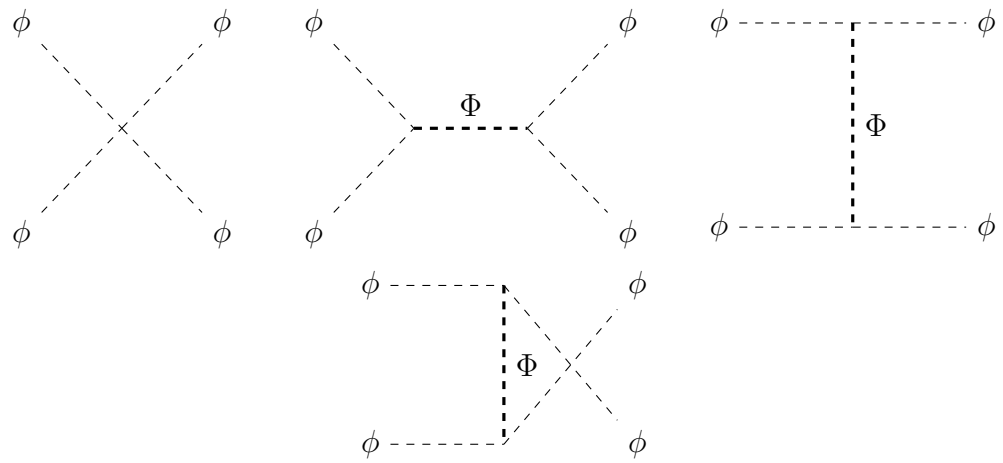


Figure 2.7: The diagrams contributing to 2-to-2 scattering at tree-level for $\phi\phi \to \phi\phi$. In order: the 4-point interaction is shown along with the s, t and u channel exchange of the heavy scalar field. For the IR theory, only the 4-point diagram contributes.

therefore completely identical to the lagrangian:

$$\begin{aligned}\mathcal{L}_{\text{IR}}^{(0)} = &\frac{1}{2}\partial^\mu\phi\partial_\mu\phi - \frac{1}{2}m^2\phi^2 \\ &- \frac{1}{4!}\left(\lambda_0 - 3\lambda_1^2 - 4\lambda_1^2\frac{m^2}{M^2}\right)\phi^4 \\ &- \frac{1}{6!}\left(\frac{20\lambda_1^2(3\lambda_1^2-\lambda_0)}{M^2}\right)\phi^6 + \mathcal{O}(M^{-4}).\end{aligned} \tag{2.75}$$

Computing observables

We are now in a position to showcase the ability of effective field theory. Consider the 2-to-2 scattering $\phi\phi \to \phi\phi$ of the light scalar field. For the full UV theory, the diagrams that contribute at tree-level are the 4-point contact interaction vertex and the s, t and u channel exchange of the heavy scalar field. The diagrams are shown in figure 2.7. Using the corresponding Mandelstam variables, the UV amplitude reads:

$$\mathcal{A}_{\text{UV}} = -\lambda_0 - \lambda_1^2 M^2\left[\frac{1}{s-M^2} + \frac{1}{t-M^2} + \frac{1}{u-M^2}\right] \tag{2.76}$$

$$= -\lambda_0 + 3\lambda_1^2 + 4\lambda_1^2\frac{m^2}{M^2} + \mathcal{O}(M^{-4}), \tag{2.77}$$

where we use the well known property that by 4-momentum conservation the Mandelstam variables sum to give four times the particle invariant mass.

We now compute the same process using our EFT lagrangian equation 2.75. The only diagram that contributes is the first diagram of figure 2.7. Our lagrangian of equation 2.75 promises to reproduce all tree-level diagrams involving the light scalar as external legs, thus we are limited to tree-level amplitudes in our calculation. Calculations involving loops requires the computation of the 1-loop matching of the UV with the IR. At tree-level in the IR the calculation is far simpler:

$$\mathcal{A}_{\text{IR}} = -\left(\lambda_0 - 3\lambda_1^2 - 4\lambda_1^2\frac{m^2}{M^2}\right) + \mathcal{O}(M^{-4}) \tag{2.78}$$

precisely the same solution as that obtained with Taylor expanding the prediction from the UV. In this way we can significantly simplify all tree-level n-point amplitudes, so long as we restrict the region of phase space to be far below the heavy mass scale M.

2.3.3 Non-renormalizable quantum field theories

When obtaining the infrared lagrangian of equation 2.75, the process of integrating out the heavy degrees of freedom generated new interactions such as a new ϕ^6 interaction for the light scalar field. Such behaviour is typical in the context of effective field theories and so we write the general IR lagrangian as

$$\mathcal{L}_{\text{IR}} = \mathcal{L}_{d\leq 4} + \frac{\mathcal{L}_5}{\Lambda} + \frac{\mathcal{L}_6}{\Lambda^2} + \cdots \tag{2.79}$$

where Λ is the characteristic energy scale of new physics which was taken as the heavy field mass in the above example. The subscripts denote the mass dimensions of the operators present for each lagrangian, for example, $\mathcal{L}_5$ contains operators of mass dimension exactly 5. By dividing by increasing powers of Λ we ensure the lagrangians have the canonical mass dimension of 4 and the operator couplings can remain dimensionless. If we consider constructing an amplitude $\mathcal{A}$ with a single insertion of an operator of dimension d, then, by dimensional grounds this part of the scattering graph must contribute to the overall amplitude as

$$\mathcal{A} \sim \left(\frac{p}{\Lambda}\right)^{d-4} \tag{2.80}$$

where p is obtained from the various kinematic variables of the process such as the external momenta. This simple power counting argument leads to a rather profound result. We see that as we flow deeper into the IR and p gets smaller: operators with

dimension less than 4 begin to contribute more while those with dimension greater than 4 have less and less importance. As such, operators with dimension < 4 are called *relevant* while those with dimension > 4 are called *irrelevant*. Operators with $d = 4$ are neither and are called *marginal* with the behaviour often being determined by considering quantum effects to the scaling behaviour (the so-called *anomalous dimension*). This important scaling behaviour illustrates the fact that lower dimensional operators, such as the mass, are important for macroscopic long-distance physics while the finer details are resolved with the addition of higher and higher dimensional operators.

However, the introduction of such irrelevant operators poses a risk to the predictability of our QFT. Generally speaking, the UV divergences arising from divergent loop momenta are regulated by the addition of counter terms. We treat the *bare couplings* of the original lagrangian as formally infinite and add counter terms to the lagrangian which cancel these infinities under some renormalization scheme, such as $\overline{\text{MS}}$ dimensional regularization [8]. For lagrangians possessing operators of dimension ≤ 4 the renormalization of the bare couplings can be done using a finite number of counter terms: requiring operators already present within the theory. However, for irrelevant operators, an infinite number of counter terms is required to regulate the UV divergences. To see this consider inserting a collection of operators with dimensions d_i, for example when renormalizing some interaction vertex, to generalize the result of equation 2.80 to:

$$\mathcal{A} \sim \left(\frac{p}{\Lambda}\right)^{\sum_i (d_i - 4)} . \tag{2.81}$$

Using as example a scattering graph involving two insertions of a dimension 5 operator, we see then that the amplitude scales as $(p/\Lambda)^2$ and thus the counter term required will be of mass dimension 6. This procedure continues infinitely, with the addition of each counter term requiring the addition of an additional counter term, but of higher dimension. Including each such counter term introduces a new unknown parameter into our lagrangian that must be determined by experimental measurement. As such, our EFT has infinitely many parameters, thus requiring infinitely many experimental measurements before any prediction can be made: a theory like this is certainly no scientific theory.

[8] A mass-independent regulator is in fact an important property for any regulator in order to preserve the power counting behaviour. Dimensional regularization thus forms the ideal candidate as opposed to a momentum cut-off regulator which has cut-off scale comparable to the new physics scale. The independence on regularization scheme between these two procedures can, however, be recovered using resummation techniques [100].

This apparent problem is resolved by accepting a finite accuracy in our predictions. By truncating the expansion at some finite order in the power counting parameter, p/Λ, we can obtain finite results from our non-renormalizable theory; even at the loop level. Indeed, note that our calculation in equation 2.78 is truncated at quadratic level in the power counting parameter. Moreover, we ubiquitously quote cross sections computed to some finite order in the strong coupling, α_s, or to a finite number of loops thereby truncating in powers of the reduced Plank's constant, $\hbar$. Thus truncating in the power counting perturbative is a natural thing to do and crucially allows for non-renormalizable EFTs to retain their predictive power.

2.3.4 The Standard Model as an EFT

Despite the vast repertoire of successful phenomena explained by the Standard Model, large gaps of unexplained details remain: neutrino masses [101], dark matter [102], and gravity, to name but a few. In this light, one is forced to set aside their hubris and resign to the fact that the Standard Model is an approximation (albeit an incredibly good one) of some more, as yet unknown, fundamental theory.

We can thus treat the Standard Model as the IR limit of some UV completion in precisely the same way that the lagrangian of equation 2.75 was obtained from the UV complete theory of equation 2.70. In doing so, we should thus allow for higher dimensional operators than those presently found in the Standard Model. We saw this happen in the example above, whereby integrating out the heavy scalar field led to operators involving the light scalar field of mass dimension 6.

The lagrangian obtained in this way leads to the Standard Model Effective Field Theory (SMEFT) [103]:

$$\mathcal{L}_{\text{SMEFT}} = \mathcal{L}_{\text{SM}} + \sum_{d=5}^{\infty} \sum_{i=1}^{N_d} \frac{c_i^{(d)}}{\Lambda^{d-4}} \mathcal{O}_i^{(d)} \tag{2.82}$$

where for each mass dimension, d, we allow for the existence of all operators, $\mathcal{O}^{(d)}$, constructed from the Standard Model matter content [9] and abiding by the $SU(3)_c \times SU(2)_L \times U(1)_Y$ gauge symmetry, we generically assume there are N_d non-redundant operators which contribute non-trivially to the S-matrix. The coefficients $c_i^{(d)}$ are

[9] Models in which the Higgs is no longer an $SU(2)$ doublet are not captured by the SMEFT, instead being the realm of the Higgs Effective Field Theory (HEFT) [104]. Such EFTs can have composite Higgs UV completions [105, 106] as a possible mechanism for electroweak spontaneous symmetry breaking.

dimensionless parameters known as Wilson coefficients and parameterize the couplings of the operators and Λ is the typical energy scale of new, beyond the Standard Model (BSM), physics assumed to be well above the electroweak scale.

The first contribution to the SMEFT is at dimension 5 and was shown [107] that at this dimension the only operator that contributes (up to hermitian conjugate) is

$$\mathcal{O}_5^{(5)} = \epsilon_{ij}\epsilon_{kl} H^i H^k (l_p^j)^T C\, l_q^l \tag{2.83}$$

where i, j, k, l are $SU(2)$ indices while p and q are flavour indices. The term ϵ_{ij} is the totally anti-symmetric tensor with $\epsilon_{12} = 1$, H and l are the Higgs and lepton $SU(2)$ doublets respectively and C is the charge conjugation matrix which in the Dirac representation reads $C = i\gamma^2\gamma^0$. This operator violates lepton number conservation; though such accidental symmetries of the SM are not required to be preserved by the SMEFT, they are nevertheless heavily constrained, so in the remainder of this text we shall consider dimension 6 operators in the SMEFT expansion. The complete minimal basis for dim-6 operators (known as the *Warsaw basis*) has been tabulated, whereby no operator from the basis can be removed using field redefinitions and integration by parts [108]. Under flavour universality, there are 59 such non-redundant operators, extending to 2499 after relaxing this flavour universality condition. Recently the complete set of non-redundant dimension 8 operators have also been tabulated [109, 110].

In the same way that the Wilson coefficients of the example EFT from equation 2.75 contains useful information regarding the UV completion: bounds on the Wilson coefficients of the SMEFT can give useful insights into possible UV completions of the Standard Model. Thanks to its model independent parameterization of BSM effects, the SMEFT will be our choice of EFT in order to parameterize new physics resonances using high-energy observables.

Chapter 3

The neural network determination of proton structure

In the context of the factorization theorems, the short distance, high energy, component of a scattering process can be computed with the tools of perturbative QCD. The component of the matrix element pertaining to the hadronic states, however, is non-perturbative and is captured by the PDFs which cannot be computed perturbatively. While the Q dependence of the PDF is described by the DGLAP evolution equations [85–87], the Bjorken-x dependence must be ascertained by comparing to experimental measurements in a global QCD analysis. Such problems of function fitting to data are the domain of machine learning [111], with a particularly important and powerful class of algorithm being the deep learning approach to function fitting [112].

The NNPDF approach to the determination of proton structure employs artificial neural networks to parameterize the PDFs. Such an approach enjoys the lack of bias introduced by choosing a particular functional form as well as the flexibility to mimic any sufficiently well behaved function. However, these virtues come at the price of possibly overlearning the training data and fitting the measurement noise rather than the underlying law.

The focus of the present section is to provide a brief overview of the salient concepts of machine learning and specifically the artificial neural network approach to function modelling. After having defined the feed-forward, fully connected, neural network, we explain how gradient descent based, supervised learning algorithms can be used to train the model by minimizing a loss function computed to a set of training data, while being conscious of the perils of overfitting.

The NNPDF4.0 approach is then outlined with detailed discussion on novel implementations of the Monte Carlo approach to ensemble learning and covariance matrix generation. We discuss how these two ingredients can be used to train a neural network to parameterize parton distribution functions, which abide by stringent positivity constraints and the various sum rules. The problem of PDF fitting is an inverse problem, whereby we attempt to discern the form of an underlying function, related to the observed measurement by a non-trivial forward map. We explain how the forward map can be achieved using the so-called Fast-Kernel method. Much of the discussion of the aforementioned points will be pertinent to later chapters and so detail is provided where necessary.

We shall also present the new parser implemented for NNPDF4.0 to implement data cuts in a declarative and human-readable way as well as the phenomenological implications the precision NNPDF4.0 PDF sets have at the LHC.

3.1 Artificial neural networks as unbiased parametrizations

The field of machine learning aims to deduce the functional dependence of some parameter, Y (the dependent variable), on some collection of independent variables (or covariates) X by means of statistical inference on measurements of Y. Often such approaches rely on vast datasets, on the order of several thousand measurements, in order to achieve statistically sound results. In general, we assume $Y \in \mathbb{R}^n$ to be some real-valued random n-vector denoting the output variable which is assumed to be a function of the input features $X \in \mathbb{R}^p$ where p is the number of covariates.

We assume the two are linked by a continuous function $f : \mathbb{R}^p \to \mathbb{R}^n$, through:

$$Y = f(X) + \epsilon \tag{3.1}$$

where ϵ is a random n-vector with zero mean and finite variance which encapsulates the inherent statistical fluctuations in the measurement. The principal issue is that the function f is not a known function, but rather must be inferred through its effect on the output variable upon changes to the input. Such is the domain of machine learning, and numerous techniques exist to approximate f, all with varying levels of prior knowledge on the form of f [111].

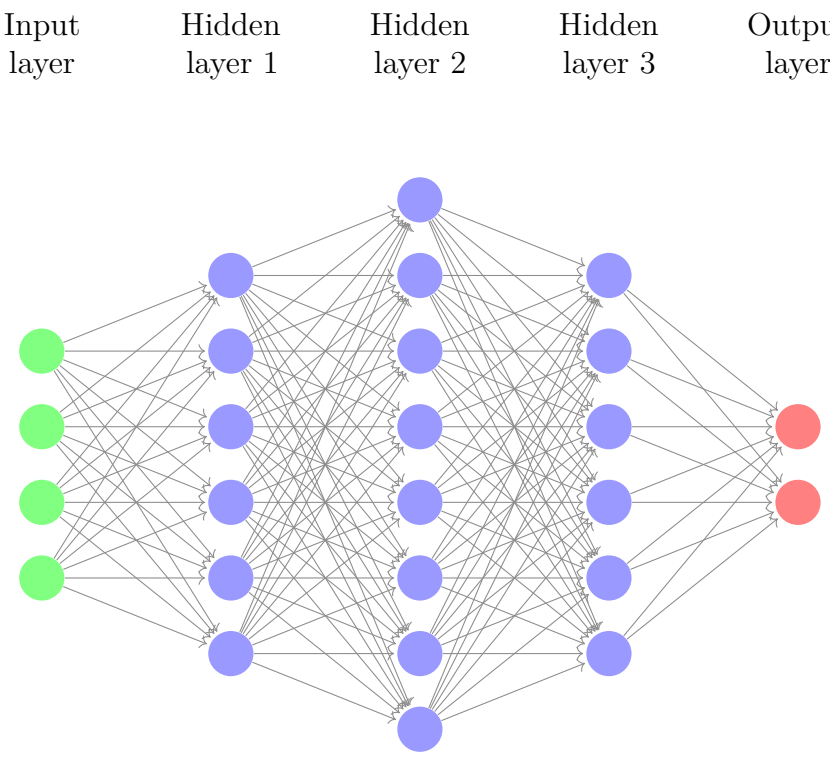


Figure 3.1: An example of a feedforward, fully connected, deep neural network with three hidden layers. The input layers are shown in green, with the hidden layers in blue. The output layer is shown in red. Each constituent node in the graph denotes a neuron while each directed edge has a weight associated to it. The architecture of the network refers to the number of neurons in each layer and so this example is dubbed a 4-6-8-6-2 neural network.

However, in sufficiently complex scenarios, where knowledge of the underlying function is essentially non-existent, one may appeal to the use of artificial neural networks [112, 113] in order to parameterize f. The field of non-perturbative QCD forms one such arena and provides a natural use-case for artificial neural networks since there is an extremely limited level of prior knowledge on the PDFs. Artificial neural networks form a class of powerful machine learning algorithms inspired by biological neural networks such as the brain. Such algorithms exist in many guises each adapted to a particular field such as natural language processing [114] (using recurrent neural networks) or image classification [115] (using convolutional neural networks), but we restrict our interest to the ubiquitous feedforward, fully connected, deep neural network architecture. In what follows, unless specified, we shall refer to this particular class of deep learning algorithm as simply *neural networks*. We can mathematically define a neural network using a sequence of definitions

Table 3.1: Popular choices of activation function for artificial neural networks. The input vector is $x \in \mathbb{R}^p$ and it is to be understood that the output is obtained by applying the activation function elementwise.

Activation Function	$\sigma(x)$
ReLU (rectified linear unit)	$\max(0, x)$
Leaky ReLU	$\max(0, x) + \alpha \min(0, x)$
sigmoid	$\frac{1}{1+e^{-x}}$
tanh	$\tanh x$
linear	x

Definition 3.1 (Activation function) *An activation function is a function $\sigma : \mathbb{R}^p \to \mathbb{R}^p$ which applies a (generally) non-linear transformation, $g : \mathbb{R} \to \mathbb{R}$, to the input vector x elementwise:*

$$\sigma(x) = \Big(g(x_1), \quad g(x_2), \quad \cdots, \quad g(x_p)\Big)^T. \tag{3.2}$$

Popular choices of activation function are tabulated in table 3.1.

Definition 3.2 (Artificial neuron) *A neuron with label i is a function $a_i : \mathbb{R}^p \to \mathbb{R}$ defined by weights $\{w_{ij} \in \mathbb{R} : j = 1, \ldots, p\}$ and a bias $b_i \in \mathbb{R}$ along with an activation function, σ. The activation of a neuron is given by:*

$$a_i(x) = \sigma\left(\sum_{j=1}^{p} w_{ij} x_j + b_i\right) \tag{3.3}$$

Definition 3.3 (Dense layer) *A dense layer of size m is a set of m neurons; for the k'th dense layer we have the collection of neurons $\{a_i^{(k)} : i = 1, \ldots, m\}$ where it is to be understood that each neuron in the set has the same domain. The activation of layer k, $A^{(k)} : \mathbb{R}^p \to \mathbb{R}^m$, is given by:*

$$A^{(k)}(x) = \Big(a_1^{(k)}(x), \quad a_2^{(k)}(x), \quad \cdots, \quad a_m^{(k)}(x)\Big)^T \tag{3.4}$$

The activation of a dense layer can be understood more easily as a linear algebra operation and indeed, owing to the efficiency of matrix multiplication, this is how modern deep learning libraries implement dense layers.

Definition 3.4 (Kernel) *The kernel of a dense layer of size m is a matrix of weights $w \in \mathbb{R}^{m \times p}$, where the i'th row is formed by the weights of the i'th neuron in the layer.*

Similarly the bias extends to a vector of biases, $b \in \mathbb{R}^m$, formed from the individual biases of the neurons.

The activation of the layer now reads:

$$A^{(k)}(x) = \sigma(w \cdot x + b) \tag{3.5}$$

and in this way we can see the action of a dense layer is a linear transformation followed by some non-linear mapping determined by the activation function.

Definition 3.5 (Neural network) *A neural network of depth d is a set of d dense layers that forms a map $f : \mathbb{R}^p \to \mathbb{R}^n$ defined by the action $f : x \mapsto f(x)$ where*

$$f(x) = A^{(d)} \circ A^{(d-1)} \circ \cdots \circ A^{(2)} \circ A^{(1)}(x) \tag{3.6}$$

where $\circ$ denotes functional composition. The codomain of any layer, $A^{(i)}$, must be identical to the next layer's ($A^{(i+1)}$) domain (that is if $A^{(i)} : \mathbb{R}^a \to \mathbb{R}^b$ then it must be the case that $A^{(i+1)} : \mathbb{R}^b \to \mathbb{R}^c$).

Such neural networks can be represented using directed acyclic graphs as depicted in figure 3.1. Each neuron is depicted as a node in the graph and each edge is a connection between two neurons. For each such edge there is a scalar weight value corresponding to an entry in the kernel of that particular layer. Note that each neuron is connected to every neuron in the preceding layer, and as such the layer is known to be *dense* and since the neural network is constructed of such dense layers, it is an example of a fully connected network. Moreover, the directed acyclic nature of the graph gives the network its *feedforward* attribute. While recurrent neural networks, which use later layers to feed back into earlier layers, can be used, they are more adept to time series or natural language processing tasks. Passing an input vector through the network in order to obtain an output value is known as a *forward pass* of the network. In the literature, the first layer, $A^{(1)}$, is often referred to as the *input layer* and is mathematically equivalent to the identity map. The last layer, $A^{(d)}$, is the *output layer* and any intermediary layers are *hidden layers*. If the number of hidden layers is greater than one then the network is a *deep network* otherwise it is *shallow*.

The *architecture* of a neural network is the size of each layer that constitutes the network design. In practice, there is no clear prescription for choosing the architecture, nor the activation function for that matter. These parameters, as well as others which determine the design of the model (such as the optimizer, or the initialization of the

weights and biases) are examples of *hyperparameters* and are discussed in section 3.2.2, but in practice there are two guiding principles. The first concerns the number of parameters in the model. In general the weights and biases of the model are to be tuned by some optimization algorithm to best fit the available data, the precise nature of this process is discussed in section 3.1.1. The general mantra in machine learning is that the number of data points must far exceed the number of model parameters in order to avoid overfitting the data and learning the statistical fluctuations as opposed to any underlying pattern. Note that a neural network with depth d and layers of size $\{N_i : i = 1, \ldots, d\}$ will have

$$\sum_{i=2}^{d} (N_{i-1} \cdot N_i + N_i) \tag{3.7}$$

independent parameters to be fitted. This number can in general grow very quickly and so the desire to achieve parsimony is particularly strong when using deep learning techniques in that the model should be as complex as needed and no more. Generically speaking, models that have too few parameters are prone to not having enough degrees of freedom to learn the data, and as such perform poorly in mimicking f (have a high bias), but this behaviour would be true of any training dataset and so they are said to have low variance. Conversely, models with too many parameters are prone to overlearning the data and thus have a low bias, but then they have learned that particular training data too well and so perform poorly on new, unseen, data and thus have a high variance. In this view, the model architecture should be selected to optimize for this bias-variance tradeoff [111]. This key machine learning concept is illustrated by figure 3.2 whereby 10 data points have been generated according to the rule $y = 2x + \epsilon$ where ϵ is homoscedastic standard normal noise[1]. To this data two models have been fitted, the first is a simple linear fit and the second is a degree 9 polynomial. The loss function is the usual ordinary least squares loss. We note that the degree 9 polynomial can fit the data exactly, since there is an equal number of data points and fit parameters, thus in principle it performs better than the linear fit when considering purely the goodness of fit metric given by the residual sum of squares (RSS). However, in this special scenario, where the underlying function, f is known to be linear, the best fit is then surely the linear fit, despite the poorer RSS. Indeed, this is made concrete by considering the generalization error of both parameterizations. When faced with a new data point, the high complexity polynomial fit will in general give a higher error than the linear fit and thus has high variance.

[1]Noise is referred to as homoscedastic if all random variables have the same, finite, variance.

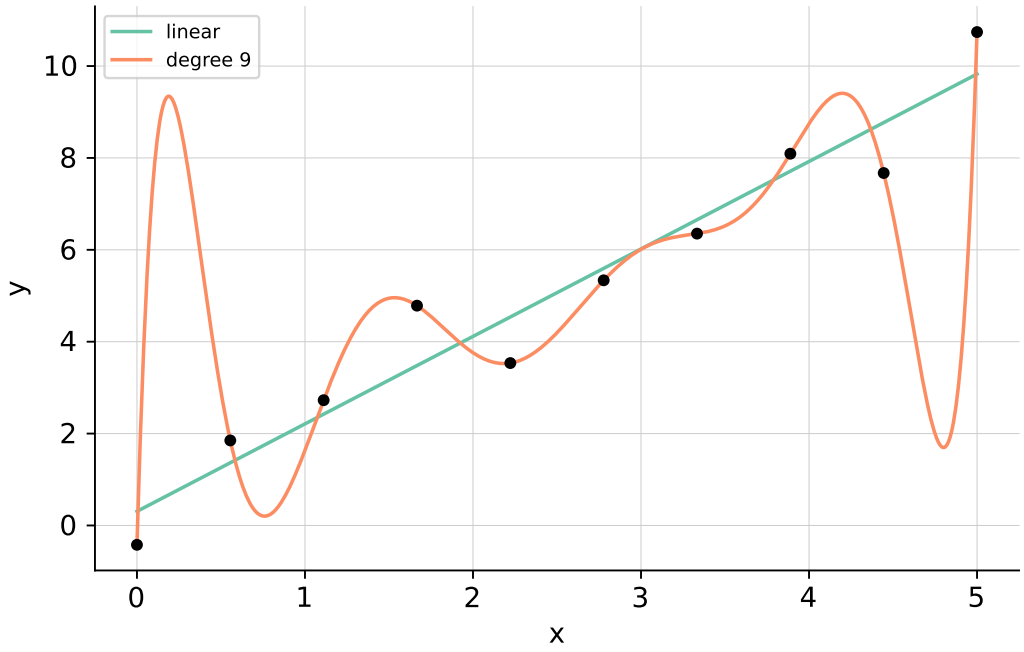


Figure 3.2: Ten evenly separated data points generated according to the rule $y = 2x + \epsilon$ where ϵ are independent samples from a standard normal distribution, shown by the black dots. To this data we perform a linear regression fit (green) and a polynomial of degree 9 fit (orange).

The use of activation function is particularly key in order to give the neural network approach its flexibility. Their non-linear nature is what allows the network to capture a wide spectrum of functional forms. Activation functions should generally be easily computed using mathematical primitives (such as max, exponentiation etc.) and their derivative should also be easily obtained for reasons which will become apparent in section 3.1.1. Note that, for example, the vastly popular rectified linear unit (ReLU) activation function has the Heaviside step function as its derivative which additionally helps avoid the so-called vanishing gradient problem [116, 117].

The neural network is a powerful parameterization of the underlying function f thanks to the following theorem [118]:

Theorem 3.1 *(Universal Approximation Theorem for Width-Bounded ReLU Networks) For any Lebesgue-integrable function*[2] $f : \mathbb{R}^n \to \mathbb{R}$ *and arbitrary* $\epsilon > 0$*, there exists a full-connected neural network with ReLU activation functions and with width*

[2]Those satisfying $\int_{\mathbb{R}^n} |f(x)|\, dx < \infty$.

$w \leq n + 4$, such that the output of this network, $n(x)$, satisfies

$$\int_{\mathbb{R}^n} |f(x) - n(x)|\, dx < \epsilon. \tag{3.8}$$

The width of a neural network is simply the size of the largest layer. There exists an analogous theorem for a neural network of one hidden layer, but with arbitrary width [119]. This important theorem highlights the expressive power of neural networks. A sufficiently deep (or wide) network can be used as an unbiased estimator for any sufficiently well-behaved function.

3.1.1 Supervised learning and training a neural network

The neural network model can now be used to map an input feature to an output vector, however, for now the model will possess little predictive power owing to the fact that the weights and biases have not been tuned properly. The weights and biases form the parameter set, $\{\theta\}$, of the model and together are referred to as *trainable* or *learnable* parameters. They will be adjusted during optimization to best fit the available training dataset and the precise nature of this process is the principle concern of the present section.

The parameters generally are initialized randomly, however, it is well known [120] that this initialization procedure is a rather important step to allow the neural network to converge both quickly and also towards the true global minimum. Various techniques [121] exist that automate the initialization procedure and are readily available in standard machine learning libraries and we shall not concern ourselves with the particular nature of this process.

The training of a neural network model is an example of a supervised learning task. We assume we have a training dataset which consists of tuples of input data and the correct output data which we attempt to learn. The input features are then said to be labelled and we wish to train the model on the training set such that it can generalize and perform well on unseen data.

The first ingredient we require is a quantitative measure of the model's performance, called a figure of merit. Mathematically, we define a *loss function*, $J(\theta)$, which will quantify how well the model is performing in a given configuration of its parameters. By convention the loss function is something we wish to minimize, so smaller losses are better than larger ones. Various choices exist for the loss function, but the one we shall use henceforth is the χ^2 figure of merit, which is interpreted as the (negative)

log-likelihood of a multivariate normal distribution, ignoring constant terms. In this way, the trained neural network will be interpreted as the maximum likelihood estimator for the underlying function, f. We restrict our attention to the case of a single scalar output, $y^{(i)}$ and a single input variable, $x^{(i)}$.

Letting

$$\mathbf{y} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(n)} \end{pmatrix} \tag{3.9}$$

be the vector of n observed values and

$$\mathbf{t}(\theta) = \begin{pmatrix} f(x^{(1)};\theta) \\ f(x^{(2)};\theta) \\ \vdots \\ f(x^{(n)};\theta) \end{pmatrix} \tag{3.10}$$

be the corresponding vector of neural network predictions when the network is configured to have parameters θ, then the loss function is defined as

$$\chi^2(\theta) = (\mathbf{d} - \mathbf{t}(\theta))^T C^{-1} (\mathbf{d} - \mathbf{t}(\theta)) \tag{3.11}$$

where C is the symmetric, positive semi-definite covariance matrix which encapsulates the uncertainty in the observed values and also the correlation between them. The optimal choice of weights and biases we seek, $\hat{\theta}$, is then given by:

$$\hat{\theta} = \arg\min_{\theta} \chi^2(\theta). \tag{3.12}$$

The minimum of the loss function is obtained using gradient descent based optimization techniques [122]. The most basic form is simply stepping in a direction antiparallel to the gradient of the loss function

$$\theta \leftarrow \theta - \eta \cdot \nabla_\theta J(\theta) \tag{3.13}$$

where η is the *learning rate* (another example of a hyperparameter) which controls the step size. For reference, the value of the learning rate is typically in the unit interval and is the same for all parameters in the model. This approach is dubbed

batch gradient descent and requires that the entire dataset is passed through the network in its entirety before computing the gradient of the loss. This has the practical drawback that the entire training dataset must be loaded into computer memory before evaluating the model. Such a task may be infeasible for sufficiently large datasets since the memory requirement may exceed what is available. An alternative is to pass the dataset through the network one datapoint at a time and compute the gradient after each forward pass. This has the advantage of requiring much less available memory, but the drawback is that the descent on the loss surface occurs in a sporadic way, since data points which strongly affect the model occur in a random order. As such this approach is dubbed stochastic gradient descent. The best-of-both-worlds approach is to pass the dataset through the model in batches and compute the gradient after each batch. This is called mini-batch gradient descent and the size of the batch is called the batch size, again, another hyperparameter to be considered. As a terminology note, the number of times the full dataset is passed through the model before convergence is achieved is known as the number of *epochs*.

However, the above vanilla gradient descent approaches suffer from a number of drawbacks. For one, the learning rate, η, is set at the beginning of training by the user and left fixed throughout the learning process. As such a small learning rate will cause the model's convergence to be too slow and possibly will lead to the global minimum not being found before the final epoch is reached. On the contrary, if the learning rate is too large then the gradient descent algorithm will overshoot the true minimum and oscillate around the it, or worse: diverge entirely. Moreover, in the large dimensional parameter space of neural networks, the existence of false minima become more scarce. This reason for this can be seen by considering the fact that a false minimum requires each direction to be increasing about a stationary point, which becomes less likely as the number of independent directions increases. As such, saddle points become far more frequent which becomes the principle problem of convex optimization [123]. The solution to the above two problems are given by adaptive learning rate optimizers (such as Adam [124]) and (Nesterov [125]) momentum respectively. Such techniques are used extensively in the deep learning community and in particular are the optimization techniques employed in the NNPDF4.0 methodology. By maintaining a record of the step size at each iteration during gradient descent, a momentum term is added to the update rule to encourage the model to follow directions which lead to the global minimum. Similarly, a record of past loss function gradients allows the optimizer to

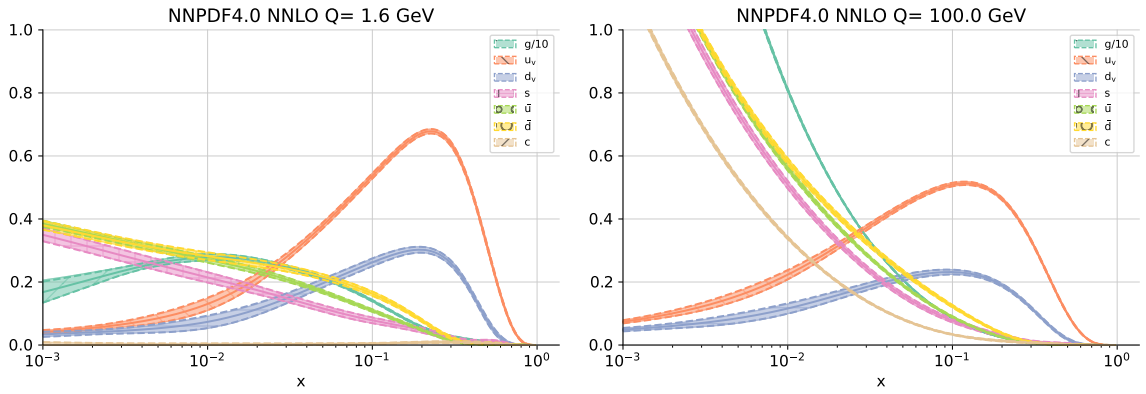


Figure 3.3: The NNPDF4.0 NNLO PDF sets at the fitting scale of 1.6 GeV (left) and at 100 GeV (right).

update important parameters slowly while those parameters which have less impact on the loss are updated more quickly (hence an adaptive learning rate).

Note that for all gradient descent algorithms, the gradient of the loss function is required. This is done using automatic differentiation [126], where by the exact gradient is computed using *backpropagation* by considering the execution graph generated by the forward pass of the network and then applying the chain rule in reverse, starting first from the loss and working backwards to the inputs themselves. This proves to be a highly efficient method, filling in the rows of the Jacobian matrix using only one backwards breadth first traversal of the execution graph.

3.2 The NNPDF4.0 methodology

The NNPDF4.0 PDF set serves as the latest major release from the NNPDF collaboration. The methodology toolchain has been majorly overhauled, using the latest in cutting edge machine learning libraries such as `TensorFlow` and the accompanying wrapper library `Keras` [127]. As such the NNPDF4.0 approach boasts a significantly improved performance and accuracy, where a fully global fit can be achieved in ~ 5 hours as compared to the previous release's ~ 24 hour fit time.

In this section we present the methodology employed to parameterize the parton distribution functions of the proton using neural networks. We highlight salient methodological improvements over the previous releases both in terms of fitting methodology, but also in the new datasets introduced. Presented in figure 3.3 are the PDF sets at the fitting scale of 1.6 GeV and evolved, using the DGLAP evolution, to a higher Q

value of 100 GeV. These constitute the most precisely determined PDF sets NNPDF has provided to date and it is the objective of the present section to outline how we arrive at these PDFs.

3.2.1 Error propagation and Monte Carlo PDFs

The covariance matrix

The data used in a PDF fit are obtained from experimental measurements and as such, associated with them, are uncertainties and correlations between them. This correlation information is encapsulated by the covariance matrix and the information must be propagated down to the PDF level. The sources of uncertainty can be categorized as follows [128]

- Statistical uncertainties are associated with the inherent randomness of the measurement being made. This source of uncertainty originates from the fact that an experimental measurement is the result of a finite sample of some unknown population.
- Systematic uncertainties instead are associated to a particular aspect of the experimental setup, such as the nature of the measuring apparatus, and are in general correlated between different measurements (the beam luminosity being an archetypal example). We shall break these uncertainties down further into two categories. Additive uncertainties, σ^{add}, which do not depend on the measured value, and multiplicative uncertainties, σ^{mul}, which are proportional to the measurement.

The uncertainty breakdown, provided from experimental collaborations, can be used to generate a covariance matrix encapsulating uncertainty and correlation information. As a motivating example, suppose we have a series of observed measurements D_i. We assume that the true underlying value is x_i, but then acknowledge that each measurement is affected, not only by a statistical error of σ_i^{stat}, but also a systematic error of σ_i^{sys} the source of which is common to all measurements. Letting the expectation value of a random variable, $X : \Omega \to \mathbb{R}$, with an event space Ω and distribution p_ω be denoted by:

$$\mathbb{E}(X) = \sum_{\omega \in \Omega} p_\omega X(\omega) \tag{3.14}$$

then it is true that:

$$D_i = x_i + X_i + S_i \tag{3.15}$$

$$\text{with} \qquad \mathbb{E}(X_i) = \mathbb{E}(S_i) = 0 \quad \mathbb{E}(S_i^2) = (\sigma_i^{\text{sys}})^2 \quad \mathbb{E}(X_i^2) = (\sigma_i^{\text{stat}})^2. \tag{3.16}$$

Then, noting that a constant shift of a random variable does not affect its covariance:

$$\begin{aligned} \text{cov}\,(D_i, D_j) &= \mathbb{E}(D_i\, D_j) - \mathbb{E}(D_i)\,\mathbb{E}(D_j) \\ &= \mathbb{E}\Big((X_i + S_i)(X_j + S_j)\Big) - \mathbb{E}(X_i + S_i)\mathbb{E}(X_j + S_j) \\ &= \mathbb{E}(X_i\, X_j) + \sigma_i^{\text{sys}}\sigma_j^{\text{sys}} \\ &= \begin{cases} (\sigma_i^{\text{stat}})^2 + (\sigma_i^{\text{sys}})^2 & i = j \\ \sigma_i^{\text{sys}}\sigma_j^{\text{sys}} & i \neq j \end{cases} \end{aligned} \tag{3.17}$$

where we have used the fact that S and X are independent and so $\mathbb{E}(SX) = \mathbb{E}(S)\mathbb{E}(X) = 0$. We see that for a given data point the total variance is given by summing in quadrature the sources of uncertainty (the diagonal entries of the covariance matrix), while the covariance between two distinct points (the off-diagonal entries) is the product of the correlated systematic uncertainties. Thus the full covariance between experimental measurements p and q generalizes to:

$$C_{pq} = \delta_{pq}\sigma_p^{\text{uncorr}}\sigma_q^{\text{uncorr}} + \sum_i \sigma_{ip}^{\text{add}}\sigma_{iq}^{\text{add}} + \left(\sum_j \sigma_{jp}^{\text{mult}}\sigma_{jq}^{\text{mult}}\right) \cdot D_p D_q \tag{3.18}$$

where we allow for systematic uncertainties to be possibly uncorrelated and so σ_p^{uncorr} is the total uncorrelated uncertainty obtained by adding the individual sources in quadrature. A description of this procedure is provided in appendix A.

An example of a correlation matrix (covariance matrix normalized by the total uncertainty) is shown in figure 3.4, whereby a 3,092×3,092 correlation matrix is generated using the procedure outlined above. We categorize the data points according to their process type and observe the strong intra-process correlation. Indeed, the correlation matrix is in block-diagonal form, owing to the fact that, for example, a data point as measured by the HERA experiments (CC and NC DIS) has no correlation with those measured by the LHC (DY, top and jets).

Although the covariance matrix, as given by equation 3.18, allows us to define the objective loss function of equation 3.11; it suffers from the problem that the minimum

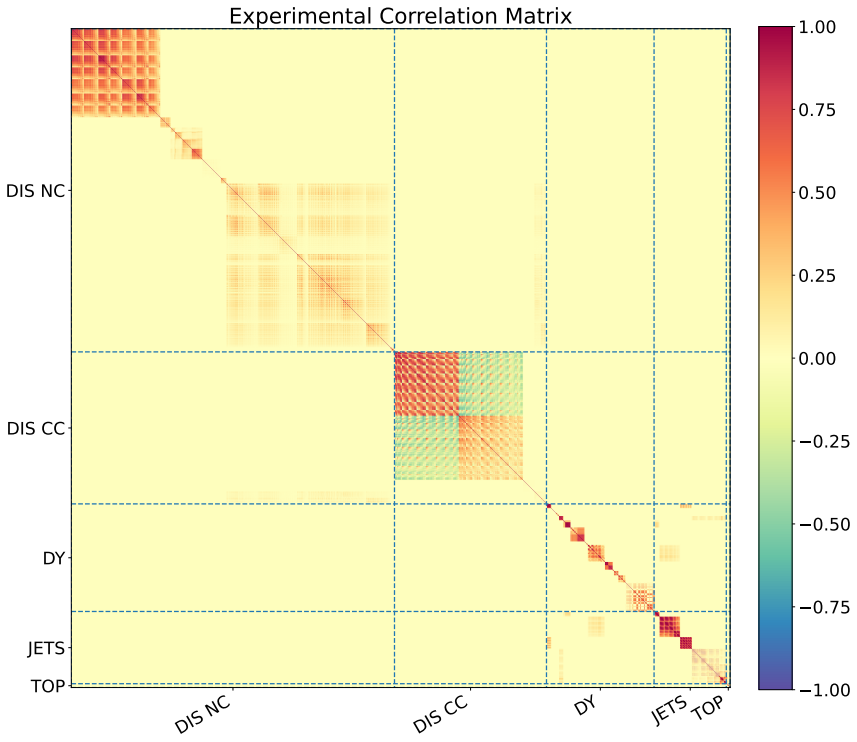


Figure 3.4: A 3,092×3,092 correlation matrix encoding the correlation and uncertainty information of the various data points available in the NNPDF framework.

obtained from such an approach is not an unbiased estimator [129]. This bias, known as the d'Agostini bias, can often be appreciable and its presence can be circumvented using the so-called t_0 approach [130], by replacing the experimental central values of the right most term in equation 3.18, with the theoretical predictions obtained by convolving the partonic cross section with some previous PDF determination, T_p:

$$C_{pq} = \delta_{pq}\sigma_p^{\text{uncorr}}\sigma_q^{\text{uncorr}} + \sum_i \sigma_{ip}^{\text{add}}\sigma_{iq}^{\text{add}} + \left(\sum_j \sigma_{jp}^{\text{mult}}\sigma_{jq}^{\text{mult}}\right) \cdot T_p T_q \tag{3.19}$$

which is known as the t_0 covariance matrix, and is the covariance matrix used for computing the figure of merit during training of the neural network model. In practice, only one iteration of the PDF set used for the t_0 matrix calculation is required.

Monte Carlo pseudodata generation

There are many approaches to incorporate training data uncertainty in a machine learning model [131], however, the method employed in the NNPDF4.0 methodology pertains to that of ensemble learning. Using this approach the data central value and covariance matrix defines a probability distribution in the space of experimental data. From this distribution, Monte Carlo (MC) samples of the data are made (often referred to as pseudodata) and to each pseudodata sample, a neural network of the form presented in section 3.2.2 is trained.

The pseudodata generation mechanism is described in [132] and proceeds by producing fluctuations about the experimental data central value, $D_p^{(0)}$, by an amount proportional to the experimental uncertainty. The k'th MC pseudodata replica for data point p then reads

$$D_p^{(k)} = \left(D_p^{(0)} + \sum_i X_{ip}^{(k)} \sigma_{ip}^{\text{add}}\right) \prod_j \left(1 + Y_{jp}^{(k)} \sigma_{jp}^{\text{mult}}\right) \tag{3.20}$$

where $X_{ip}^{(k)}$ and $Y_{jp}^{(k)}$ are a collection of independent and identically distribution $\mathcal{N}(0,1)$ random variables which allows for fluctuations about the experimental data central value to be generated. As discussed above the uncertainty breakdown can be composed of correlated systematics which in general introduce correlations between different measurements p and p'. In this particular case, the random variables associated with the systematic are set equal: $X_{ip}^{(k)} = X_{ip'}^{(k)}$ (or $Y_{ip}^{(k)} = Y_{ip'}^{(k)}$ if they are multiplicative uncertainties).

As with the covariance matrix generation, the pseudodata generation implementation has been entirely overhauled, eliminating altogether the need for the legacy `C++` codebase. The pseudodata generation has proven to be a major performance bottleneck in previous releases, but thanks to the new implementation a performance boost of 2 orders of magnitude is achieved. Further performance can be gained using just-in-time compilation through the use of `Numba` [133], however, such performance gains were not deemed necessary for the current release.

A low level implementation of the procedure is implemented in appendix A with the actual implementation being heavily vectorized in order to achieve performance.

A model, as described in section 3.2.2, is then trained for each such MC replica, generating an ensemble of N_{rep} neural networks. The observable, $T^{(k)}$, is then obtained by convolving the partonic cross section with the k'th replica. The central value

(expected value) and uncertainty (variance) is given by the weak law of large numbers as

$$\mathbb{E}_k\left(T^{(k)}\right) = \frac{1}{N_{\text{rep}}}\sum_{k=1}^{N_{\text{rep}}} T^{(k)} \tag{3.21}$$

$$\text{Var}_k\left(T^{(k)}\right) = \frac{1}{N_{\text{rep}}-1}\sum_{k=1}^{N_{\text{rep}}}\left(T^{(k)} - \mathbb{E}_k\left(T^{(k)}\right)\right)^2. \tag{3.22}$$

In general ~ 100 MC replicas are sufficient for a faithful reproduction of the experimental uncertainties at percent level, with $\sim$ 1,000 replicas achieving the correlations at the same precision [134].

3.2.2 Model design

Functional form

Contrary to previous releases of NNPDF [130, 135–137] the NNPDF4.0 release of PDFs parameterizes all parton flavours using a single neural network. The basis used to fit the partons is the evolution eigenbasis:

$$\begin{aligned}
\Sigma &= u + \bar{u} + d + \bar{d} + s + \bar{s} + 2c \\
T_3 &= \left(u + \bar{u}\right) - \left(d + \bar{d}\right) \\
T_8 &= \left(u + \bar{u} + d + \bar{d}\right) - 2\left(s + \bar{s}\right) \\
V &= \left(u - \bar{u}\right) + \left(d - \bar{d}\right) + \left(s - \bar{s}\right) \\
V_3 &= \left(u - \bar{u}\right) - \left(d - \bar{d}\right) \\
V_8 &= \left(u - \bar{u} + d - \bar{d}\right) - 2\left(s - \bar{s}\right) \\
T_{15} &= \left(u + \bar{u} + d + \bar{d} + s + \bar{s}\right) - 3\left(c + \bar{c}\right) \\
g &
\end{aligned} \tag{3.23}$$

and it is assumed that at the fitting scale $c = \bar{c}$. Here, u, d, s and c refer to the up, down strange and charm quark distributions with the bar denoting the anti-quark distribution. The g PDF refers to the gluon distribution. We shall refer equations 3.23 as the evolution basis in this text.

It is important to note that the parameterization captures the Bjorken-x dependence of the PDFs, since the Q dependence is fully determined thanks to the DGLAP evolution of equation 2.42 and equation 2.43. As such we fit the x dependence at

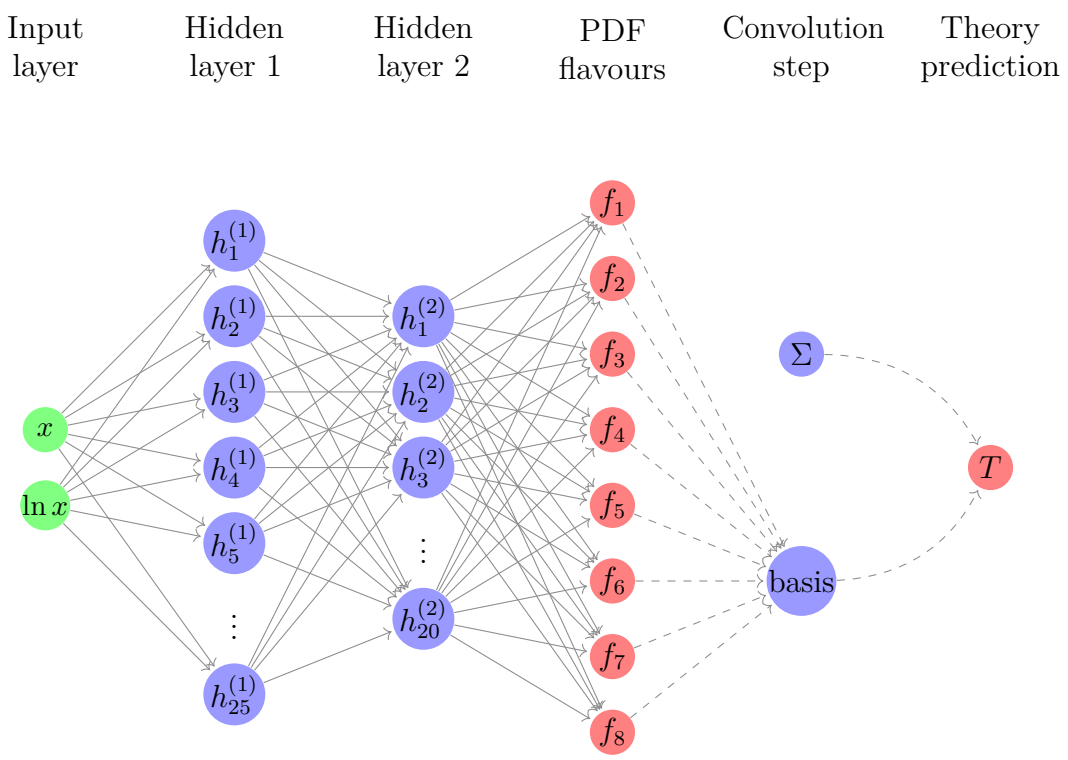


Figure 3.5: The neural network architecture used to fit the parton distributions. The input layer is Bjorken-x and its logarithm. The output layer is the PDFs in the evolution basis. All solid edges are learned parameters, the dashed edges are non-trainable and depict the convolution of the PDFs with the partonic cross (FK table in section 3.2.3). The architecture used is a 2-25-20-8 design.

fixed scale $Q_0 = 1.65$ GeV before evolving to the relevant process scale using the DGLAP equations. This value of initial scale ensures that the PDFs are fitted above the charm mass threshold, in the case where the intrinsic charm content of the proton is independently determined.

The PDFs themselves are parameterized by neural networks complemented by a preprocessing term that characterizes the small and large-x extrapolation behaviour and aids in the parameterization convergence:

$$f_i(x, Q_0^2) = A_i\, x^{-\alpha_i}(1-x)^{\beta_i}\mathrm{NN}_i(x), \qquad i \in \{1, \ldots, 8\} \tag{3.24}$$

where i enumerates the parton flavour, NN_i is the i'th output of the neural network of figure 3.5 and A_i is a normalizing constant that allows the parameterization to abide by the momentum and valence sum rules. Note that the valence sum rules of equations

2.30-2.33 in the evolution basis read

$$\int_0^1 dx\, V(x,Q^2) = \int_0^1 dx\, V_8(x,Q^2) = 3 \qquad \int_0^1 dx\, V_3(x,Q^2) = 1 \tag{3.25}$$

$$\int_0^1 dx x \left(\Sigma(x,Q^2) + g(x,Q^2)\right) = 1 \tag{3.26}$$

and so setting:

$$\begin{aligned} A_g &= \frac{1 - \int_0^1 dx x\widetilde{\Sigma}(x)}{\int_0^1 dx\, x\widetilde{g}(x)}, \qquad & A_V &= \frac{3}{\int_0^1 dx\, \widetilde{V}(x)}, \\ A_{V_3} &= \frac{1}{\int_0^1 dx\, \widetilde{V}_3(x)}, \qquad & A_{V_8} &= \frac{1}{\int_0^1 dx\, \widetilde{V}_8(x)} \end{aligned} \tag{3.27}$$

allows for equation 3.24 to satisfy the momentum and valence sum rules. The notation here is such that $\widetilde{f}$ refers to the unnormalized PDFs, that is, equation 3.24, but without the normalization constant.

The exponents of the preprocessing term, α_i and β_i, are determined iteratively according to the prescription outlined in [136]. The favoured exponents of a previous PDF fit are used to define a uniform distribution to sample new exponents from and this process is repeated until convergence is achieved. This, typically, requires only one iteration of the preprocessing exponents.

Early stopping and optimization

With deep learning models often having large numbers of free parameters, the risk of the model being overly complex and thus learning the statistical noise is high. Such a regime is known as *overfitting* and can occur when the number of parameters is comparable to the number of training samples. Numerous ways exist to circumvent this problem, such as dropout mechanisms which stochastically drop links between neurons thus preventing certain connections from developing too strongly [138], or regularization techniques such as $L2$-regularization (weight decay) [139] which add a penalty term to the loss function, forbidding the model parameters from growing too large. The regularization approach employed in NNPDF4.0 is the so-called *early stopping* approach which halts the training process once the onset of overlearing occurs.

The approach followed by this method begins by splitting the global dataset into two disjoint subsets (such that their intersection is empty and the union is the entire dataset) referred to henceforth as training and validation sets. The precise nature of this split is left to the user, but in general, a $3:1$ ratio is recommended. The training

dataset is then used to train the model using the gradient descent based algorithms outlined in section 3.1.1. After each iteration of gradient descent the model χ^2 is evaluated on the validation set, which is in principle out-of-sample and thus can be treated as unseen data. The general trend is that the training and validation χ^2 both begin to initially decrease as the model learns the underlying laws common to both subsets. However, while the training χ^2 will generally continue to decrease *ad infinitum* since the model can be tuned to asymptotically fit the data noise, the validation metric does not enjoy such improvements. Rather, after a particular number of iterations the goodness of fit on out-of-sample data begins to cease to improve and instead starts to deteriorate. The point at which this transition occurs is effectively the parsimonious point of the model at which the underlying law is learned as best as possible and any further training will only serve to overfit the model. At this point training is halted, despite not being at the global minimum of the loss-surface: hence early stopping.

This process is depicted schematically in figure 3.6 whereby the training loss decreases monotonically while the validation loss initially follows the trend, but eventually begins to deteriorate. We choose to stop training the model at the optimal stopping point which is depicted by the vertical line that divides the under and overfitting regimes. In practice, however, the minimum of the validation loss is not so clear cut and instead we end training once the validation χ^2 has not improved for a certain number of iterations known as the *patience* parameter.

As discussed in section 3.1, a variety of choices for the neural network model are determined by the user. For example, the precise choice of neural network architecture, activation functions, initializer, optimizer and parameters therein, and loss function, amongst others. Collectively, all such parameters are referred to as *hyperparameters*. In general, there is no algorithmic way to determine the best hyperparameter choice and so it is the task of the user to use their intuition and various guiding principles to choose hyperparameters that are best suited to the machine learning problem at hand.

A naive, brute force, method would be simply to perform a scan of all reasonable hyperparameters and converge on the set of hyperparameters that achieved the lowest loss. Such a task is, however, in general infeasible owing to the fact that the complexity grows exponentially in the number of hyperparameters. However, it is possible to employ Bayesian approaches [140] whereby the information gained from one random trial educates the parameter selection for the following trial. Indeed, this methodology, dubbed `Hyperopt` [140], is employed for the NNPDF4.0 hyperparameter selection. The result of this process is tabulated in table 3.2. The most salient difference between the

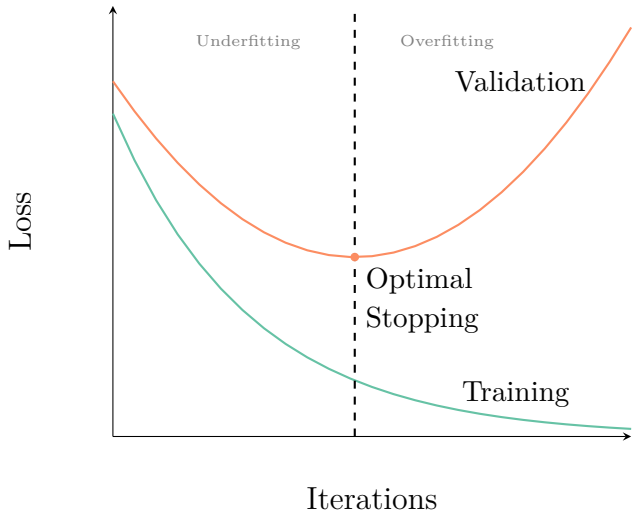


Figure 3.6: Schematic depiction of the early stopping mechanism. For each iteration in the learning process, the χ^2 of the training (blue) and validation (orange) set is monitored. Training is halted once the minimum of the validation χ^2 is obtained. Configurations to the left of this point are underfitted and overfitted to the right.

hyperparameter choices of NNPDF4.0 and previous iterations is the use of gradient descent optimizers as opposed to the legacy genetic algorithm approach [141]. The improved optimization algorithm leads to an improved fit quality and the substantial performance gains advertised. The fact that hyperparameter optimization in this way is possible is credit to the fact that an individual fit using `TensorFlow` is much quicker and the fact that K-fold cross validation [142] is used as a hyperoptimization figure of merit. Using the legacy `C++` implementations of NNPDF3.1 would have meant that such a scan would have been simply too computationally intensive and thus infeasible.

Table 3.2: The hyperparameter selection used for NNPDF4.0, obtained using the `Hyperopt` methodology.

Parameter	Choice
Architecture	2-25-20-8
Activation	$\tanh - \tanh -$linear
Initializer	glorot-normal [143]
Optimizer	Nadam [144]
Learning Rate	2.6×10^{-3}
Max. epochs	$17,000$
Patience	10% max. epochs

3.2.3 From neural networks to theoretical predictions

As mentioned before, the task of PDF fitting is an example of an inverse problem, where we are given the observations (cross section measurements) and we attempt to reconstruct the causal factor that generated them (the PDF). Recalling the discussion of section 2.2, the forward problem is given by convolving the PDF with a hard partonic cross section computed in pQCD. However, with the PDF model being given by the output of a neural network, it is not obvious how this convolution should happen, indeed, the convolution involves a non-trivial integral which must be performed every time a prediction is needed. Moreover, ignoring the problem of feasibility and assuming such a procedure exists, it need also be an extremely fast operation. The convolution happens for each data point a vast number of times during learning. For every iteration of the neural network parameters the convolution must be performed, and so an approximate figure for the maximum time taken for a convolution must be on the order of milliseconds or better. With the choice to perform the optimization using gradient descent methods in NNPDF4.0, the procedure must also be composed of mathematical primitives, such that they are differentiable, using backpropagation, to obtain the gradients required for the optimization problems of section 3.1.1. To further complicate matters, the factorization and renormalization scales must also be evolved to the kinematic scale in a similarly efficient manner. As described by equations 2.42 and 2.43, the former is governed by the DGLAP evolution equations and the latter by the relevant beta functions.

The Fast Kernel (FK) approach [134, 145, 146] provides precisely the prescription required. The non-trivial convolution is reduced to a tensor product with FK tables that have been precomputed and stored. The tensor product is clearly a differentiable operation and an extremely efficient one too since tensor manipulations are heavily optimized with readily available libraries. The overview of the FK table approach, presented here, will prove pertinent for much of the discussions in chapter 6.

As a pedagogical example, consider a lepton-hadron interaction, as in DIS, measured at kinematical point (x_I, Q_I^2). We can use the evolution kernel operator (EKO), $\Gamma_{ij}(x_I, Q_I^2, Q_0^2)$, to evolve the PDFs from the fitting scale, Q_0^2, to the scale of the observable, Q_I^2:

$$f_i(x_I, Q_I^2) = \int_{x_I}^1 \frac{dy}{y} \Gamma_{ij}\left(\frac{x_I}{y}, Q_I^2, Q_0^2\right) f_j(y, Q_0^2) \tag{3.28}$$

where summation on repeated indices is implied throughout this discussion. Then, convolving $f_j(x, Q^2)$ with the coefficient function appropriate for this process, C_j where

j indexes the partonic channel, we may write the theoretical cross section for this specific point as:

$$\sigma(x_I, Q_I^2) = \int_{x_I}^1 \frac{dy}{y} C_j\left(\frac{x_I}{y}, Q_I^2\right) f_j(y, Q_I^2) \tag{3.29}$$

$$= \int_{x_I}^1 \frac{dy}{y} K_j\left(\frac{x_I}{y}, Q_I^2, Q_0^2\right) f_j(y, Q_0^2) \tag{3.30}$$

where we have defined a modified evolution operator, K_j, that also incorporates the convolution with the coefficient function:

$$K_j(x_I, Q_I^2, Q_0^2) = \int_{x_I}^1 \frac{dy}{y} C_k\left(\frac{x_I}{y}, Q_I^2\right) \Gamma_{kj}(y, Q_I^2, Q_0^2). \tag{3.31}$$

We can factorize the PDFs at fixed scale, $f_j(y, Q_0^2)$, from equation 3.30 by introducing a basis of interpolation functions, $\{\mathcal{I}^\alpha\}$, and a grid of x values which is a monotonically strictly increasing sequence, $\{x_\alpha\}$, for $\alpha \in \{1, \ldots, N_x\}$. In the jargon, the x_α are known as *knots* and represent points at which the function is known precisely. The interpolation functions satisfy the useful property that for each knot there exists one and only one interpolation function that is non-zero at that point while all others are vanishing. We refer the reader to [134] for the precise nature of the x-grid and choice of the interpolation basis.

Linear combinations of the interpolation basis can be used to approximate an arbitrary function and we can thus approximate the PDFs at fixed scale by:

$$f_j(y, Q_0^2) \sim \sum_{\alpha=1}^{N_x} \mathcal{I}^{(\alpha)}(y) f_j(x_\alpha, Q_0^2) \tag{3.32}$$

with the accuracy of the interpolation being governed by coarseness of the x-grid. The term, $f_j(x_\alpha, Q_0^2)$, denotes the output of the neural network parameterization of equation 3.24 evaluated at x_α.

By interpolating the PDF parameterization on a grid of x values in this way we may rewrite equation 3.30 as:

$$\sigma(x_I, Q_I^2) \sim \sum_{\alpha=1}^{N_x} \left[\int_{x_I}^1 \frac{dy}{y} K_j\left(\frac{x_I}{y}, Q_I^2, Q_0^2\right) \mathcal{I}^{(a)}(y)\right] f_j(x_\alpha, Q_0^2)$$

$$\equiv \sum_{\alpha=1}^{N_x} \Sigma_{\alpha j} f_j(x_\alpha, Q_0^2) \tag{3.33}$$

where the summation on α is left explicit. The quantity Σ is the FK-table which reduces the convolution integral to a tensor product. Note that the FK-table is free from the dependence on the PDF parameterization: this key point allows the FK tables to be pre-computed once and stored on disk thereafter. As such, the forward map from PDF to observable is an extremely quick operation, on the order of a few milliseconds, once the FK tables have been computed, making the FK table approach ideal for PDF fitting. The computation of the FK-table is done using the public PDF evolution code `APFEL` [147].

The extension to hadronic observables is a more involved example of the prescription outlined above, requiring a rank-4 tensor contracted with the PDF luminosity.

3.2.4 Positivity and integrability of PDFs

The interpretation that parton distributions are understood as probability distributions (and thus non-negative) breaks down beyond LO in QCD. The reason for this is that at NLO the collinear subtraction is scheme dependent and so whether a PDF is positive or not depends on the scheme chosen. That being said, however, it can be shown [148] that $\overline{\text{MS}}$ PDFs are positive even at NLO. In addition, any physical observable obtained by convolving the parton distributions with a partonic cross must be non-negative at all orders in perturbation theory by the simple argument that a cross-section is related to a probability distribution.

The positivity of NLO PDFs themselves are new to NNPDF4.0 and are implemented by modifying the target loss function with the addition of a Lagrange multiplier, Λ_k, penalty term for each independently parameterized parton flavour. The inclusion of this term penalizes network configurations that violate the positivity. Noting that this positivity applies to the flavour basis, $\widetilde{f}_k$ (related to the fitting basis of equation 3.23 by a linear system of equations), and so we append to the χ^2 loss:

$$\chi^2 \to \chi^2 + \sum_{k=1}^{8} \Lambda_k \sum_{i=1}^{n_i} \text{Elu}_\alpha \left(-\widetilde{f}(x_i, Q^2) \right) \tag{3.34}$$

where

$$\text{Elu}_\alpha(t) = \begin{cases} t & t > 0 \\ \alpha \left(e^t - 1 \right) & t < 0 \end{cases} \tag{3.35}$$

with α set to 10^{-7}. The Lagrange multipliers increase exponentially during fitting, reaching a maximal value by the final epoch. The initial values, on the other hand, are determined by hyperoptimization.

The positivity of physical observables is imposed in a similar manner, with a Lagrange multiplier being introduced to impose positivity on pseudo-observables constraining linear combinations of PDFs and their luminosities.

Also new to NNPDF4.0 is the requirement that the PDFs are integrable. By equations 3.25 and 3.26 we see that for $q \in \{V, V_3, V_8, xg, x\Sigma\}$ (and also T_3 and T_8 by the Gottfried sum rules [149]) the behavior in the small-x region must satisfy

$$\lim_{x\to 0^+} xq(x, Q_0) = 0 \tag{3.36}$$

in order for the valence and momentum sum rules to be satisfied. In a similar fashion to the positivity, we enforce the integrability conditions by adding a Lagrange multiplier to the loss function proportional to

$$\Big(xq(x_i, Q_0)\Big)^2 \tag{3.37}$$

for a set of sample points, x_i, concentrated in the small-x region.

Finally, as a final *a posteriori* check, once fitting has successfully terminated, we check that all final PDFs do not violate positivity and integrability conditions too strongly. Indeed, due to the vastly improved methodology used in NNPDF4.0, far fewer replicas are discarded by the post-fit selection (as compared to NNPDF3.1) meaning the user need run only a small fraction more replicas than the number desired.

3.3 Declarative data cut selections

One of the core mantras embedded in the principles of the NNPDF software design is that all inputs to the vast code base must be not only reproducible, but also human readable. That is to say a given `runcard` must declare, in a readable way, all inputs (leaves) for the code infrastructure (execution graph) such that an execution of the `runcard` will always produce an identical output regardless of when or where it is run. As such, the NNPDF codebase is developed based off of the `reportengine` [150] framework which processes `YAML` inputs and produces the code output(s) as specified by the user.

However, a major point which was lacking this code philosophy was in the data cut selection. When data is implemented into the fitting framework, it is often the case that a small subset of points are in kinematical regions of phase space that are not appropriate to be included in a PDF fit. For example, a given DIS measurement may have low Q^2 (relative to Λ_{QCD}) and thus higher twist corrections to the collinear factorization are not suppressed and become non-negligible. Similarly some data points have large electroweak corrections related to them for which we do not have an adequate theoretical description [3]. For such particular cases, it is important that these data points are not included in a fit in order to avoid introducing tensions which only exist because of our lack of theoretical understanding.

In previous releases, with approximately 3,000 data points, the cut selection was implemented by simply considering a large `if-else` statement, thus making it very difficult to trace down the reason why a given data point was, or was not, cut from the fit. Such an implementation thus lacks human readability, often getting in the way of physics (since one must consider why a data point, which may aid in constraining a PDF has not been considered), but is also prone to errors; in that a data point has been cut when it should not have been or vice versa[4].

With the large influx of new data points, spanning a broad spectrum of processes and experiments, being introduced into NNPDF4.0 and future releases, this method of implementing cuts was no longer sustainable and was rather impeding the stream of incoming datasets. As such, in NNPDF4.0, a new cut selection procedure has been implemented which boasts readability as the main improvement, but also performance enhancements as a by-product.

In doing so, we have implemented a new `YAML` parser with the ability of understanding domain specific constructs [152]. A syntax is developed to allow for a readable and easily implemented declaration of cut policies which can be added, removed or edited with minimal effort. We begin by defining a list of *filter rules* which are atomic declarations of a specific cuts policy. These filters declare to which dataset or process type the rule applies. For example, we may wish that a rule applies to all DIS data points, or perhaps to only a specific experiment from the HERA collider. Accompanying this information is the cut policy itself. This is a string written in valid `Python` syntax which should return a boolean value. If the value of the expression is `False` then this

[3] Though good progress is being made in the full inclusion of electroweak corrections using the `PineAPPL` library [151].

[4] A quick check on the issue tracking of the code repository reveals several such bugs did indeed exist and have been since caught thanks to the new style cuts implementation.

rule discards the point in question. If, however, the rule evaluates to `True` then we move on to the next rule (importantly a return value of `True` does not imply that the point is kept). One may further define optional arguments to the filter rule. These include specifying if the cut should be applied only if the corresponding theory is at a particular perturbative order. The user can also specify additional `local_variables` which are used to define variables specific only to a given rule. These generally help simplify the syntax for the rule itself. The user has the option to use typical mathematical primitives in this namespace, such as exponentials or logarithms; with more niche functions being available so long as it is available in the `NumPy` library. The final, optional argument is a `reason` field which is simply a string which explains why the cut policy is there in the first place.

As a pedagogical example one such filter rule is given below in the corresponding `YAML` syntax and indeed is a filter rule from the NNPDF4.0 release.

```
- dataset: CMSDY2D11
  reason: Remove data points for which electroweak corrections are large.
  PTO: NNLO+
  local_variables:
    M: sqrt(M2)
    max_rapidity: 2.2
    max_M: 200.0
  rule: M <= max_M and etay <= max_rapidity
```

It is clear that the rule applies to the CMS Drell-Yan double differential measurement from 2011 [153]. A quick glance at the corresponding article [153] reveals the distribution is binned in dimuon absolute rapidity (referred to by the variable `etay`) and the dimuon invariant mass (`M2` in the code which is simply the squared invariant mass). These two variables, alongside the centre of mass energy (`sqrts` in the code), define the three kinematic variables for this measurement and form the basis of our `rule` definition. We see that any point with invariant mass above 200 GeV or absolute rapidity above 2.2 is discarded. For convenience, we have defined the variable `M` to be invariant mass and also the maximum rapidity and invariance mass have their own variables too, which ensures that it is clear that these values are upper bounds. The `PTO` variable declares that the cut should only be used if a NNLO or better (denoted by the `+`) theory is employed. The rationale here being that $\alpha_{\text{EW}} \sim \alpha_s^2$ and so electroweak corrections

are comparable to the NNLO QCD corrections. Finally, the `reason` field provides a brief reason for the cut and we see that is is to avoid regions of phase space with large electroweak corrections.

A more non-trivial example is given by the fixed target Drell-Yan measurement from experiment E605 at Fermilab [154].

```
- dataset: DYE605
  reason: |
    Remove data points for which the fixed-order
    perturbative expansion is not reliable since
    resummation effects are large. A justification
    of these cuts can be found in arXiv:1507.01006.
  process_type: DY
  local_variables:
    tau: M2 / sqrts**2
    ymax: -0.5 * log(tau)
    maxTau: 0.080
    maxY: 0.663
  rule: tau <= maxTau and fabs(y/ymax) <= maxY
```

This particular filter rule pertains to threshold resummation and removes datapoints deemed to be too close to the production threshold which would spoil the validity of the resummation. The `process_type` field focuses on fixed target Drell-Yan measurements and applies for all perturbative orders. The `local_variables` field here plays an important role. The standard choice of kinematic variables for this dataset are the squared invariant mass of the dimuon pair, the (square root) beam centre of mass energy, and absolute rapidity. The rule itself, however, concerns the kinematic variable τ, defined through the relation

$$\tau = \frac{m^2}{s} \tag{3.38}$$

which is of course different for each measured value. Moreover, the allowed maximum rapidity is defined through τ through the upper-bound $y_{\max} = -\frac{1}{2}\log\tau$. Writing the `rule` entry, explicitly using these relations, will spoil the readability of the rule. As such, we choose to define them symbolically in the `local_variables` field and then use these newly defined variables in the `rule`. The rule itself is then trivial, cutting out data points that exceed a maximal τ and rapidity value. The `fabs` and `log` operations

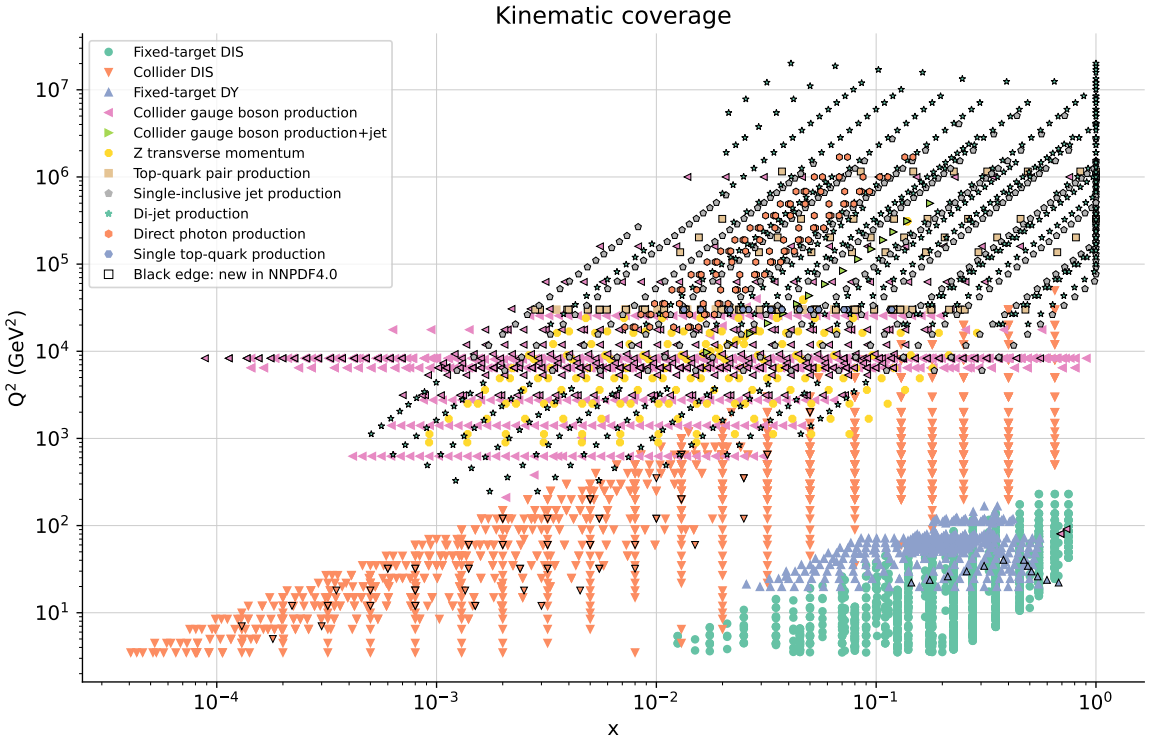


Figure 3.7: The kinematic coverage of the datasets used in NNPDF4.0, grouped according to their process, mapped in the (x, Q^2)-plane. Points marked with a black edge are new to NNPDF4.0.

are floating point absolute value and the base-e logarithm respectively, which the parser understands and applies correctly. The `reason` entry for this particular rule is paramount, stating that the threshold resummation will be violated by data points not abiding by this cut policy and conveniently links a reference for more details [155].

The development of this custom `YAML` parser also provides the user with a convenient framework to investigate the effect adding or removing particular data points has on the resulting fit. Indeed, for the various chapters pertaining to PDF fitting presented throughout this work, this method of applying cuts is used extensively and seamlessly.

3.4 New datasets in NNPDF4.0

Accompanying the various methodological improvements, discussed extensively throughout this chapter, is a host of new datasets spanning a broad range of new processes. The addition of an extra 333 (131) data points at NNLO (NLO) as compared to the previous NNPDF3.1 release [137] aids in a markedly improved precision in the latest

PDF sets. The additional processes introduced are: gauge boson with jets, single top production, inclusive isolated (prompt) photon production, and dijet production. The data points are depicted in the (x, Q^2) plane in figure 3.7 with the markers grouped according to their process. We depict the new points with a black border and highlight the extended kinematic reach of, for example, jet production. The introduction of new data points are a key ingredient to parton distribution determination. By exploiting different processes as well as the universality of PDFs, one can constrain different regions of Bjorken-x and further consolidate regions already well determined. However, despite the advantages of using data spanning more distant regions of kinematical phase space, one is at peril of being sensitive to possible BSM resonances residing beyond the highest energy bins. We shall address this issue in chapter 5, but for now we assume this not to be the case and thereby assume the correct description of LHC phenomenology is the Standard Model when fitting PDFs.

3.5 High precision parton distribution functions

We now move on to present the results of these numerous methodological improvements and the impact of the dataset extension. In figure 3.8 we compare the NNPDF4.0 PDF set with the NNPDF3.1 release [137]. Both PDFs are composed of 1000 MC replicas for enhanced statistics. We compare two representative choices of parton flavours: shown are both the up and gluon distributions at $Q = 100$ GeV. We see a strong agreement in both flavours: the up quark central value remains virtually identical across the entire Bjorken-x domain. The gluon PDF shows a good agreement too, with the gluon suppression at medium/large-x being due to the new single-inclusive jet, dijet, and top-pair production datasets. The fact that it is precisely these datasets that modify the gluon is determined by performing fits with various training datasets and determining which variant has the greatest impact on the gluon PDF [1]. An important point of note is the reduction in PDF uncertainties, also shown in figure 3.8, this time in the lower panel. Indeed, in various kinematical regions we see a reduction in PDF uncertainties by up to 50% relative to the NNPDF3.1 PDF sets, with the absolute PDF uncertainty reaching a low of $\sim 1\%$. This crucial figure acts as a milestone for precision in PDF determinations and sets a precedence in achieving a gold-standard of 1% PDF uncertainties across a broad kinematic range as well as for all the various PDF flavours. This reduces the PDF contribution to theoretical uncertainties and acts

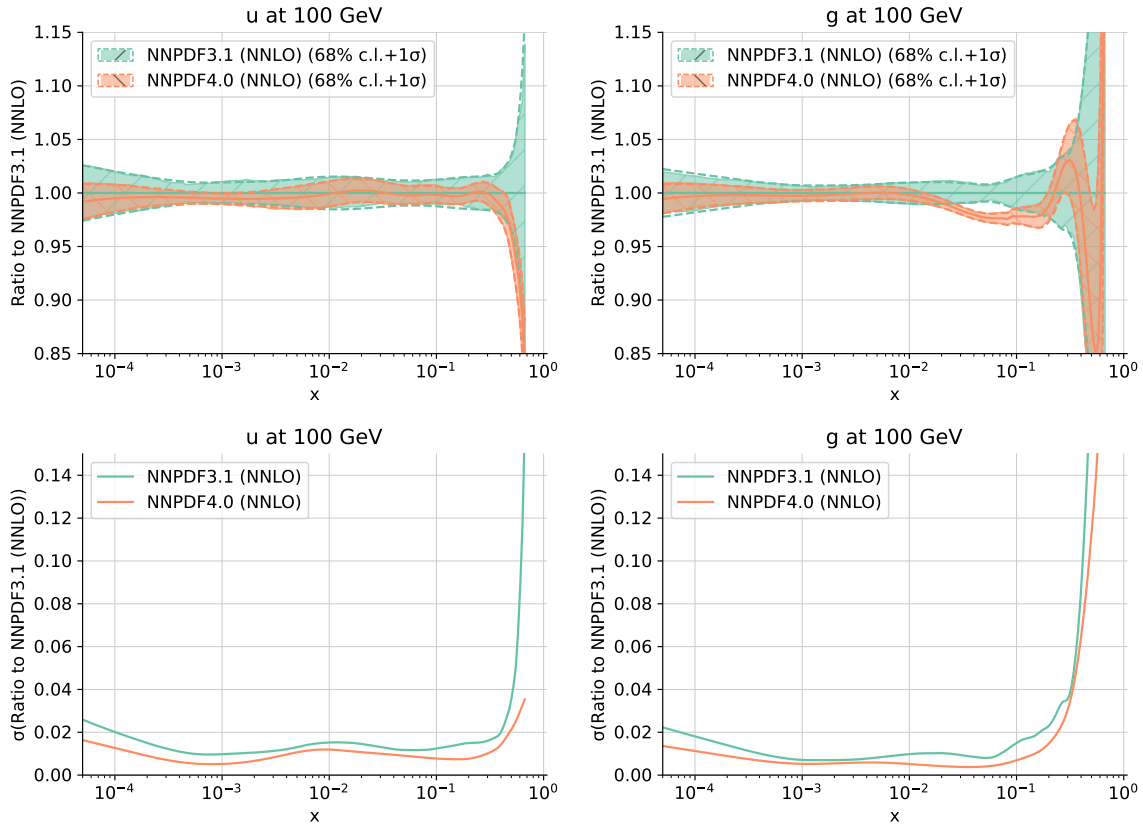


Figure 3.8: Comparison of NNPDF4.0 PDFs (orange) normalized to those obtained in the NNPDF3.1 release (green) at $Q = 100$ GeV. Shown are the up (left) and gluon (right) distributions (top) and the relative 1σ uncertainties (lower panel). Solid and dashed bands correspond to 68% confidence level and one-sigma uncertainties, respectively.

as an impetus for a strong understanding in the missing higher order uncertainties [156, 157].

The importance of a well determined PDF set is presented at a phenomenological level in figure 3.9 where we perform a data-theory comparison for the HERA combined dataset [158] as well as the CMS measurement of double differential Drell-Yan cross sections at the LHC [159]. We use both NNPDF4.0 and NNPDF3.1 PDF sets to make the comparison. We bring the reader's attention to the improved central value agreement of NNPDF4.0 as compared to NNPDF3.1. Moreover, and equally as important, is the reduction in PDF uncertainties highlighted by the shrinkage in error bars. It is worth mentioning that the pQCD part of the theoretical calculation is identical in both cases. That is: the partonic cross section is computed at NNLO in QCD and the PDFs themselves are fitted assuming NNLO DGLAP evolution. The increased precision is thus totally attributed to the improved methodology and enlarged input dataset for

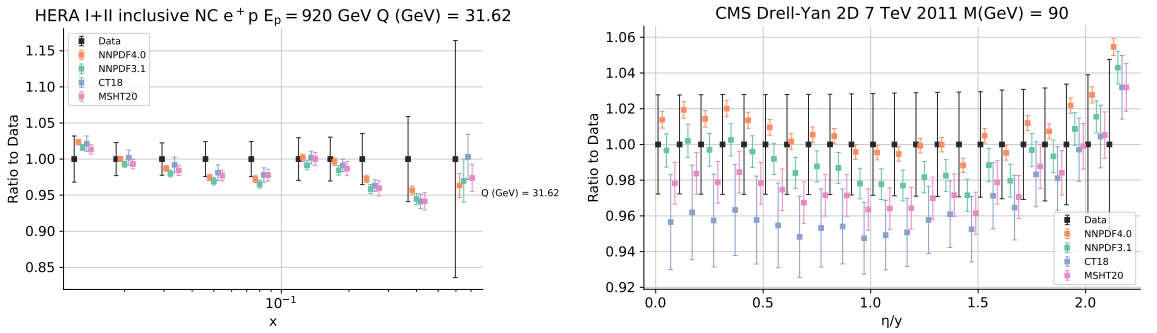


Figure 3.9: Data-theory comparisons using NNPDF4.0 (orange), NNPDF3.1 (green), CT18 [35], (blue), and MSHT20 (pink) [160]. Shown are the H1 and ZEUS combined dataset measurement of neutral current positron-proton scattering with proton beam energy of 920 GeV and $Q = 31.62$ GeV [158] (left) and the double differential CMS Drell-Yan measurement at 7 TeV binned in dimuon rapidity at an invariant mass of 90 GeV [159] (right).

the NNPDF4.0 PDF fit. Phenomenological comparisons between the NNPDF4.0 PDF set against the CT18 [35] and MSHT20 [160] PDF sets are also made in figure 3.9. We see that even when compared to non-neural network based approaches, the PDF uncertainties are vastly reduced with the NNPDF4.0 PDFs.

Finally, the picture for the global dataset is captured by table 3.3 where we tabulate the χ^2 fit quality for the various processes used in this work. As is usual, we use the experimental covariance matrix (equation 3.18) to make the χ^2 calculation. The fit quality is generally compatible, being close to unity for virtually all the processes. At NNLO in QCD the largest χ^2 is 1.36 for inclusive gauge boson production; though this dataset has by the far the greatest experimental precision. At the opposite end of the spectrum, single top production has the lowest χ^2 of 0.36, but correspondingly has the largest uncertainties. We note here that an arbitrarily low χ^2 is not in general desired since this can be a sign of overfitting. We expect a χ^2 per degree of freedom close to unity since then the standard fluctuations around the central value are comparable to the uncertainties; while if they are too low then we are, in a sense, fitting too well, and perhaps fitting the noise rather than an underlying law. However, the number of data points for this dataset is rather low, and combined with the large experimental uncertainties, we can safely conclude that this dataset has been adequately captured by the NNPDF4.0 PDFs. The increase in DIS χ^2 from NNPDF3.1 to NNPDF4.0 is attributed to enhanced sensitivity to small-x resummation effects [161] being omitted in the fit. Indeed, with reference to figure 3.7, we note the addition of small-x DIS measurements not previously present in the NNPDF fits. These logarithmic enhancement effects arise order-by-order in perturbation theory and

lead to large logarithms in the small-x region, the resummation of which requires the BFKL equation [162–165]. We note the remarkable result that the overall NNPDF4.0 fit quality to the global dataset is comparable to the NNPDF3.1 goodness-of-fit, despite the substantial increase in the number of data points in the entire data set (333) at NNLO.

Table 3.3: Summary of χ^2 per degree of freedom for the various processes used in NNPDF4.0. The number of data points are shown in parentheses. We tabulate values using NNPDF4.0 (central column) and NNPDF3.1 (right column). Prompt photon and single top production were not included in NNPDF3.1 and so we do not tabulate their χ^2 values. The χ^2 is computed using the experimental covariance matrix of equation 3.18.

Dataset	NNPDF4.0	NNPDF3.1
DIS NC (fixed-target)	1.26 (973)	1.12 (973)
DIS CC (fixed-target)	0.86 (908)	1.08 (908)
DIS NC (collider)	1.19 (1127)	1.15 (1130)
DIS CC (collider)	1.28 (81)	1.18 (81)
Drell-Yan (fixed-target)	1.00 (195)	1.25 (189)
Tevatron W, Z inclusive production	1.09 (65)	1.29 (74)
LHC W, Z production (inclusive)	1.37 (483)	1.37 (314)
LHC W, Z production (Zp_T and W+jets)	0.98 (150)	1.00 (120)
LHC top-quark pair production	1.21 (66)	1.08 (19)
LHC jet production	1.26 (500)	0.94 (470)
LHC isolated γ production	0.77 (53)	—
LHC single t production	0.36 (17)	—
Total	1.16 (4618)	1.15 (4285)

3.6 The open source NNPDF code

Up until the NNPDF4.0 PDF release, the related code base for all prior NNPDF releases has been an internally held, developed, tested, and executed infrastructure. In fact, this statement is true of all global PDF sets, whereby the fitting code of

other PDF collaborations has never been made publicly available to the high energy physics community. However, for the first time the NNPDF code has been made public and open source [2] in conjunction with the PDF sets themselves. In addition to the fitting code itself, this release includes the original and filtered experimental data, the fast NLO interpolation grids relevant for the computation of hadronic observables, and whenever available the bin-by-bin NNLO QCD and NLO electroweak K-factors for all processes entering the fit. Furthermore, the code comes accompanied by a battery of plotting, statistical, and diagnosis tools providing the user with an extensive characterization of the PDF fit output. The entirety of the code base is accompanied by a comprehensive and user-friendly documentation resource and a host of pedagogical example use cases. Doing so sets a precedence for transparency, reproducibility, and scrutiny of the code base in keeping with the ethos of the scientific method.

The availability of the NNPDF open-source code, along with its detailed online documentation, will enable users to perform new PDF analyses based on the NNPDF methodology and modifications thereof. Some examples of potential applications include assessing the impact of new measurements in the global fit; producing variants based on reduced datasets, carrying out PDF determinations with different theory settings, for example as required for studies of the strong coupling or heavy quark mass sensitivity, or with different electroweak parameters; and quantifying the role of theoretical uncertainties from missing higher orders to nuclear effects [156, 157]. One could also deploy the NNPDF code as a toolbox to pin down the possible effects of beyond the Standard Model physics at the LHC, such as Effective Field Theory corrections in high-p_T tails of chapter 5 or modified DGLAP evolution from new BSM light degrees of freedom [166]. Furthermore, while the current version of the NNPDF code focuses on unpolarized parton distributions, its modular and flexible infrastructure makes it amenable to the determination of closely related non-perturbative collinear QCD quantities such as polarized PDFs, nuclear PDFs [167, 168], fragmentation functions, or even the parton distributions of mesons like pions and kaons [169].

Chapter 4

Constraining the strange content of the proton

As the energy at which the proton is probed is increased, one is able to resolve the proton with greater and greater resolution. With sufficiently high energies, the sea distributions, formed by quark-antiquark creation from gluon splitting, begin to contribute at a level on-par with the valence distributions. This is manifest from figure 3.3, where evolution to $Q = 100$ GeV, well above the charm production threshold, shows the sea distributions competing with the valence up and down quarks. Moreover, in a quantum field theory with gauge group $SU(3)_C$ and N_f flavours of identical mass spinors transforming in the fundamental representation, the isospin symmetry $SU(N_f)_f$ relating to unitary transformations of the flavours is an exact symmetry. However, QCD does not enjoy this exact symmetry, owing to the different quark masses, but despite this, an observational fact is that the up and down quark masses are very similar, and so the $SU(2)$ is an approximate isospin symmetry in QCD. With larger values of Q this approximation becomes better and better, being promoted to $SU(3)$ well above the strange mass. So at these high-Q regimes, the second generation quarks are equally as important as the first generation quarks and are thus of great phenomenological interest at the LHC. For example, the determination of Standard Model parameters such as the W-boson mass [170], the Weinberg angle [171], or general electroweak parameter determinations using LHC measurements will benefit greatly from an improved knowledge of the strange or charm quark distribution. However, with the relative lack of processes entering a global fit of PDFs that are sensitive to the strange quark distribution, the strange PDF is in general much less constrained than the up and down distributions.

The determination of the proton strangeness will be addressed in this section. We will employ the NNPDF3.1 methodology to determine the strange quark distribution using various cutting edge theoretical calculations supplemented by the addition of LHC strange sensitive measurements. We shall consider the strangeness ratio

$$R_s(x,Q^2) = \frac{s(x,Q^2)+\bar{s}(x,Q^2)}{\bar{u}(x,Q^2)+\bar{d}(x,Q^2)} \tag{4.1}$$

which measures the sea-strangeness relative to the other light sea-quark combinations. The strangeness ratio of equation 4.1 has gained a lot of interest recently, with tensions arising from ATLAS measurements of W and Z boson rapidity measurements favouring a ratio of $R_s(0.023, 1.6^2\ \text{GeV}^2) \sim 1$ [172, 173] strongly incompatible with the $R_s(0.023, 1.6^2\ \text{GeV}^2) \lesssim 0.5$ obtained from the CT18, CT18A (CT18A a variant of CT18 which includes the ATLAS W and Z rapidity distributions) [35], MMHT14 [36], and ABMP16 [37] PDF analyses.

4.1 Data sensitive to the strange distribution

The main bulk of the data used in this study stems from the precision strong coupling determination of [174]. In particular, this contains measurements of dimuon (anti)neutrino-nucleus DIS cross sections from NuTeV [175, 176]. The NuTeV experiment measured charged current (CC) (anti)neutrino-iron collisions which leads to charm production events through the strange quark (or the Cabibbo suppressed down quark [1]) interactions. This leads to charm quark production which semimuonically decays to a dimuon final state whereby the muons have opposite charge. The neutrino beam energy varied in the range 20-400 GeV. Additionally, inclusive gauge boson production in proton-(anti)proton collisions from various Tevatron and LHC experiments [173, 178–181] are also present in this data selection.

Accompanying these strange-sensitive datasets are additional measurements which are introduced to better constrain the strange distribution. The NOMAD [182] experiment measures neutrino-iron deep inelastic scattering events giving rise to dimuon final states. The process is shown schematically in figure 4.1. The semi-leptonic decay of the charm quark results in dimuon production where the final state muon is oppositely charged to that coming from the charm quark decay. The dimuon presents a clear signal by which to differentiate this process from other charged current background

[1] The CKM matrix elements to compare are $|V_{cs}| = 0.997 \pm 0.017$ with $|V_{cd}| = 0.218 \pm 0.004$ [177].

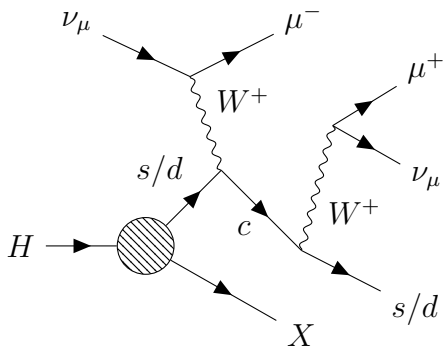


Figure 4.1: The dominant mechanism for neutrino induced charged current deep inelastic scattering charm dimuon production measured by the NOMAD experiment [182]. The processes stemming from an emitted down quark is Cabibbo suppressed.

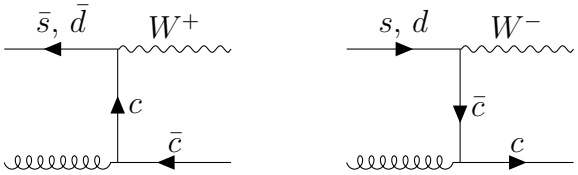


Figure 4.2: The dominant mechanism for the associated $W + c$ production. The d and $\bar{d}$ contributions are Cabibbo suppressed.

interactions. Crucially, while the charm dimuon production can proceed via an emitted down quark, this process is Cabibbo suppressed by CKM matrix elements and so the NOMAD mechanism provides a precise probe for the strangeness of the proton. The observable delivered by NOMAD is the charm dimuon cross section ratio normalized by the total charged current cross section:

$$R(E_\nu) = \frac{\sigma_{\mu\mu}(E_\nu)}{\sigma_{CC}(E_\nu)}, \tag{4.2}$$

where we bin the measurements in E_ν, the neutrino beam energy. The NOMAD experiment introduces a total of 19 additional data points. The observable of equation 4.2 is ideal for constraining strangeness as the ratio leads to cancellation of experimental uncertainties as well as the fact that the denominator is largely insensitive to the strange distribution since the inclusive cross section does not suffer from a Cabibbo suppressed first generation quark contribution.

In the case of proton-proton collisions the ATLAS 7 TeV inclusive gauge boson production [173] is adjusted to also include the off-peak[2] and forward rapidity bins

[2]Dilepton invariant mass significantly above and below the Z peak.

(which were originally omitted from [174]) making for a total of 91 data points from this experiment. Measurements of $W + c$-jet from ATLAS at 7 TeV [183] and CMS at 7 [184] and 13 TeV [185] are also included. The dominant mechanism for this process is shown in figure 4.2 which again uses the Cabibbo suppression of the first generation quark channel to probe the strange content most dominantly. These datasets together contribute 37 extra measurements. Finally, 32 extra data points are introduced corresponding to the ATLAS W+jets measurement at 8 TeV differential in the W-boson transverse momentum, p_T [186]. The total number of data points in this study is thus 4096; optimized for a precision determination of the proton strangeness.

4.2 Theoretical considerations

4.2.1 The NOMAD observables

To compute the theoretical predictions corresponding to the numerator and denominator of the ratio observable, equation 4.2, we require the DIS double differential cross section:

$$\frac{d^2\sigma_i}{dxdQ^2} = \frac{G_F^2 M_W^2}{4\pi}\frac{1}{(Q^2+M_W^2)^2} \cdot \left[\left(Y_+ - \frac{2m_p^2x^2y^2}{Q^2}\right)F_2^i(x,Q^2) - y^2F_L^i(x,Q^2) + Y_-xF_3^i(x,Q^2)\right]K^i \quad (4.3)$$

where $Y_\pm = 1 \pm (1-y)^2$ related to the inelasticity $y = Q^2/(2m_pE_\nu x)$. The quantity G_F is the Fermi constant and M_W is the W-boson mass. The index i enumerates whether we are considering the numerator ($i = \mu\mu$) or the denominator ($i =$CC) of equation 4.2. Correspondingly, the overall factor K^i is unity for the inclusive cross section ($i =$CC) while for the charm dimuon measurement it is the charm semi-leptonic branching ratio and has the slightly more involved functional form:

$$K^{(\mu\mu)} = \int_0^1 dz \sum_h f_h D_c^h(z)\mathcal{B}(h \to \mu X) \quad (4.4)$$

where f_h is the production fraction of a charmed hadron, h, and $\mathcal{B}(h \to \mu X)$ is the inclusive branching ratio for the muon decays. The quantity $D_c^h(z)$ is the charm fragmentation function [187], which is required when specific final states are identified (exclusively or semi-inclusively) and can be thought of as the probability that an unpolarized charm quark fragments into an unpolarized charmed hadron, h, where the

hadron carries a fraction z of the parton momentum. As with PDFs, this probabilistic interpretation breaks down beyond LO in QCD [188]. Akin to PDFs, fragmentation functions are non-perturbative objects in QCD and require to be fitted to data assuming some parameterization. The NOMAD experiment assume the Collins-Spiller parameterization [189], as well as a universal fragmentation function, $D_c(z)$, for all charmed hadrons:

$$D_c(z) = \left[\frac{1-z}{z} - \epsilon_c \frac{2-z}{1-z}\right](1+z)^2 \left[1 - \frac{1}{z} - \frac{\epsilon_c}{1-z}\right]^{-2} \tag{4.5}$$

where ϵ_c is a free parameter and determined, using NOMAD data as well as data from E531 at Fermilab [190], to be:

$$\epsilon_c = 0.165 \pm 0.025. \tag{4.6}$$

The production fraction and branching ratio are then combined in an effective semi-leptonic branching ratio for muon production:

$$\mathcal{B}_\mu = \sum_h f_h \mathcal{B}(h \to \mu X) \tag{4.7}$$

with the functional form of $\mathcal{B}_\mu(E_\nu) = a(1 + b/E_\nu)^{-1}$ being assumed. The fit parameters are again determined by NOMAD [182] to be $a = 0.097 \pm 0.003$ and $b = 6.7 \pm 1.8$. These uncertainties are included as systematic uncertainties for inclusion in the experimental covariance matrix to be used in the PDF fits as discussed in section 3.2.1.

Both the charm ($i = \mu\mu$) and total ($i =$CC) structure functions $F^i_{2,L,3}(x, Q^2)$ of equation 4.3 are computed using `APFEL` [147] and have been benchmarked against an independent computation based on [191]. The agreement is on the permille level, reaching a few percent in the lowest E_ν bins. Finally, the observable of equation 4.2 is related to the differential cross sections of equation 4.3 by integrating over the fiducial phase space:

$$\sigma^i(E_\nu) = \int_{Q^2_{\min}}^{Q^2_{\max}} dQ^2 \int_{x_{\min}}^{x_{\max}} dx \frac{d^2\sigma^i}{dx dQ^2}(x, Q^2, E_\nu) \tag{4.8}$$

which is handled numerically.

4.2.2 NNLO massive corrections in neutrino DIS

We incorporate charm-quark massive corrections in the theoretical description of the neutrino DIS structure functions [191, 192] (computed at NNLO in QCD) of NOMAD [182] and NuTeV [176]. These massive corrections are not ready to be implemented directly in a PDF fit owing to their computational intensity [191] and so are delivered in the K-factor approximation, defined by convolving an NNLO PDF set, f_{NNLO}, with the relevant partonic cross section. Schematically:

$$K = \frac{f_{\text{NNLO}} \otimes \sigma(m_c \neq 0)}{f_{\text{NNLO}} \otimes \sigma(m_c = 0)} \tag{4.9}$$

and so the differential cross section of equation 4.3 including massive corrections reads:

$$\left.\frac{d^2\sigma^i}{dxdQ^2}\right|_{m_c} = K \cdot \left.\frac{d^2\sigma^i}{dxdQ^2}\right|_{m_c=0}. \tag{4.10}$$

4.2.3 Nuclear corrections in neutrino DIS

The neutrino DIS measurements of NuTeV and NOMAD used in this study use an iron (Fe) target. It is a known fact that the PDFs of free nucleons are not identical to those when the nucleons are bound in a nucleus [193, 168]. This effect is, however, ignored in our study since they are expected to be sub-dominant relative to other sources of uncertainty. For the case of NuTeV, the effect is shown explicitly to be moderate [194], while for NOMAD the effect is expected to approximately cancel in the ratio observable. The validity of the latter statement was checked by computing equation 4.2 using the NLO Fe nuclear PDF set `nNNPDF2.0` [167] and compared with the equivalent NLO free proton PDF set. The difference was on the permille level, reaching $\sim 3\%$ in the lowest E_ν bin: vastly smaller than the data and PDF uncertainties.

4.2.4 NNLO corrections for collider gauge boson production

The theoretical predictions for inclusive W and Z-boson production as well as for W-boson production in association with charm quarks or light jets are evaluated at NLO using `MCFM+APPLgrid` [195, 196] and are supplemented by NNLO QCD K-factors. These K-factors are computed using `FEWZ` [197] for the inclusive gauge boson production, and N_{jetty} [198, 199] for the W production with light jets. However, for the $W + c$-jet production data, NNLO QCD corrections have only recently been made available [200]

and are not ready to be used in a PDF fit. As such, NNLO QCD corrections are omitted and the missing higher order uncertainty is implemented using the so-called 9-point (renormalization and factorization) scale variations of the NLO calculation [156, 157], which allows a component to be added to the t_0 covariance matrix to account for the theoretical uncertainties that affect this observable in the fit.

4.2.5 Positivity of cross sections

Using the approach outlined in section 3.2.4, positivity of the charm structure function, F_2^c, is imposed alongside positivity constrains on light quark distributions. This ensures the intrinsic charm PDF does not become unphysically negative.

4.3 PDF fit strategy

We now assess the impact of the above experimental data and theoretical considerations on the PDF fits. We will perform PDF fits at NNLO, where available, and all the PDFs are obtained using the NNPDF3.1 methodology [136]. The reason for this is that, at the time this study was performed, the NNPDF4.0 approach and PDF sets were not yet available. An interesting future study would be to perform this study presented in this section using the enhanced accuracy of the NNPDF4.0 approach.

The first PDF fit is our baseline, referred to henceforth as `str_base`. This fit corresponds to [174], with the exception that NNLO charm-mass K-factors for NuTeV data are included, positivity on F_2^c is enforced and the 2010 and 2011 ATLAS inclusive gauge boson production of [173, 180] are omitted. The reason for this last point is that these datasets introduce tensions in the PDF fit and so by isolating them in this way, their impact can be more critically assessed.

The inclusive gauge boson production data is then reinstated and all the new LHC data described in section 4.1 are now included to yield the PDF fit we shall call `str_prior`.

Finally, the `str_prior` PDF set is supplemented with the NOMAD data using the Bayesian reweighting and unweighting procedure [201, 202] with the prior being the `str_prior` set. The reason the NOMAD data must be added in this way is that the two-dimensional fiducial integral of equation 4.8 is too computationally costly to do during a PDF fit. The reweighting procedure, however, allows us to compute the integral only once, so long as the number of new data points is relatively small

Table 4.1: Value of the χ^2 per data point for the various strangeness-sensitive datasets considered in this work. We display values for all 3 PDF sets described in section 4.3. The totals for each sub-categories are also shown, which account for correlations across datasets. Values in square brackets are for datasets not included in the corresponding fit.

Process	Dataset	n_{dat}	$\chi^2_{\text{str\_base}}$	$\chi^2_{\text{str\_prior}}$	χ^2_{str}
νDIS ($\mu\mu$)	NuTeV [176]	76/76/76	0.70	0.71	0.53
	NOMAD [182]	—/—/19	[9.0]	[8.8]	0.55
Total		76/76/95	0.70	0.71	0.53
W, Z (incl.)	ATLAS [180, 173]	—/91/91	[2.61]	1.52	1.44
$W + c$-jet	CMS [184, 185]	—/15/15	[1.10]	0.98	0.96
	ALTAS [183]	—/22/22	[0.53]	0.48	0.42
Total		—/37/37	[0.76]	0.68	0.60
W+jets	ATLAS [186]	—/32/32	[1.58]	1.18	1.18
Total		3917/4077/4096	1.17	1.17	1.17

relative to the prior, which is indeed the case here. The baseline fit, `str_base`, is composed of $N_{\text{rep}} = 100$ Monte Carlo replicas while the prior, `str_prior`, is composed of $N_{\text{rep}} = 500$ replicas since the reweighting method requires a large prior in order for the posterior to have an appreciable ensemble size. After reweighting, we construct the PDF set, composed of $N_{\text{rep}} = 100$ MC replicas, containing all the strange sensitive data of section 4.1, which we shall refer to as `str`.

In table 4.1 we summarize the values of the χ^2 per data point obtained using the 3 PDF sets described above. We group the χ^2 values according to the process and consider only the strange sensitive measurements. We see that initially the inclusive gauge boson production from ATLAS and the NOMAD measurements are poorly described, having a χ^2 per degree of freedom of 2.61 and 9.0 respectively. Including the ATLAS data in the `str_prior` fit reduced the χ^2 to a much better value of 1.52, while the reweighting procedure further improves this to 1.44, thanks to the better constrained strange PDF arising from now including the NOMAD data. Similarly for the NOMAD data, the χ^2 is improved to 8.8 at first when considering the LHC measurements, but the reduction is much more significant when NOMAD is introduced, dropping all the way to 0.55 in the full `str` PDF. The χ^2 of all the other datasets, however, continue to be consistently well described and we therefore concluded that

the global dataset is overall consistent and satisfactorily described by the final `str` PDF.

4.4 Data theory comparisons

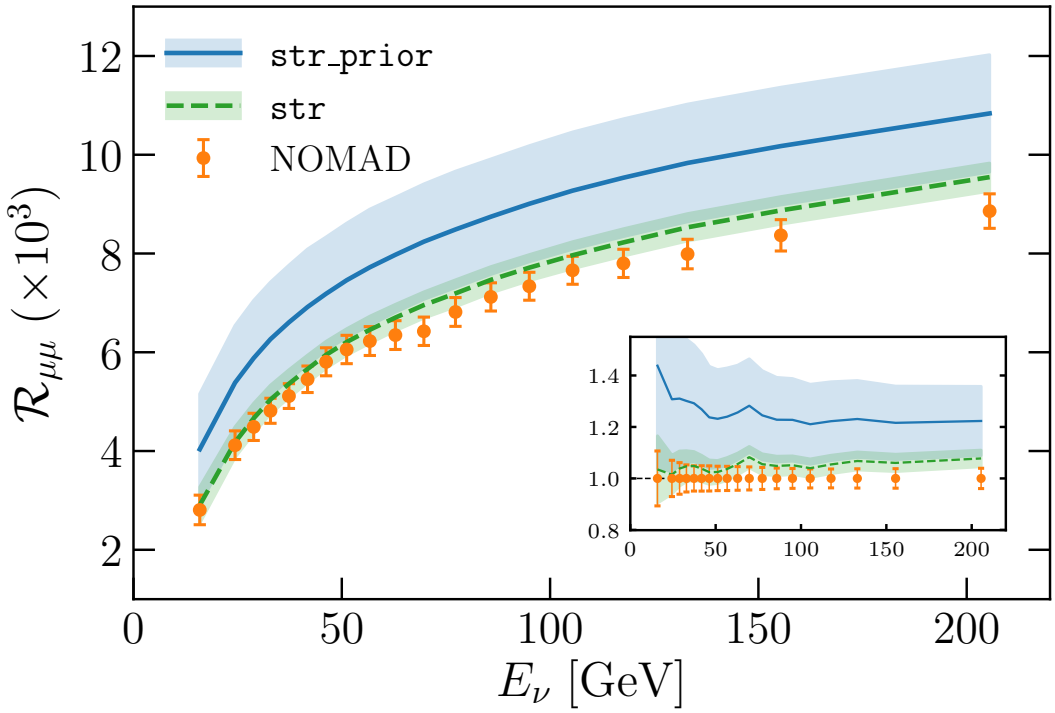


Figure 4.3: Comparison between the theoretical predictions against the data for the NOMAD experiment [182] as a function of the neutrino beam energy E_ν. The inset displays the ratio to the central value of each data point. The error bands depict the 1σ PDF uncertainties. The PDFs used are `str_prior` (blue) and `str` (green) where only the latter has the NOMAD data included.

We now move on to compare the strangeness-sensitive datasets included in our analysis with the corresponding theoretical predictions. We aim to assess the impact of introducing the various datasets at the observable level, before discussing the PDF level impact in section 4.5.

We begin by considering the impact of neutrino-DIS observables. In figure 4.3 we display the comparison for the NOMAD measurements against the theoretical predictions as a function of the neutrino beam energy, E_ν. We see that when predictions are made using the `str_prior` PDF set, which omits the NOMAD data, an overshoot

of the data points of about 20% is present. We see all the data points consistently lie beyond the PDF uncertainty error bands and hence lead to the large χ^2 value of 8.8 reported in table 4.1. After the NOMAD data is included in the fit using the reweighting procedure, the data is now well described (evidenced by the $\chi^2/n_{\text{dat}} = 0.55$). We see the interesting result that once the data is considered by the fit, not only is the central values in better agreement, but the PDF uncertainty is shrunk consistently by approximately a factor of 4 at the observable level.

This has a knock-on effect of better describing the other neutrino-DIS dataset considered in this work: the NuTeV dimuon production [176]. In figure 4.4 we plot the theory prediction normalized to the data for both the neutrino and anti-neutrino beams used in the NuTeV experiment. Shown are values for both the `str_base` baseline PDF and the full `str` PDF set, including NOMAD. We see that not only is the central value improved, but there is a further reduction in uncertainties across all measurements. This is also reflected by table 4.1, whereby the introduction of the NOMAD data causes the NuTeV χ^2 to drop from 0.70 to 0.53, despite the fact that this dataset was present in both PDF determinations.

We now turn our attention to the LHC experiments. In particular we consider the ATLAS [183] and CMS [184, 185] measurements of $W + c$-jet production and present the comparisons for ATLAS inclusive gauge boson production in appendix B (which are in keeping with the present discussion). Figure 4.5 shows the theory predictions, made using `str_base` and `str`, normalized to the data central value. The measurements are binned in lepton invariant mass which are produced by the decay of the W-boson. A consistent pattern of reduced uncertainties is observed, with a general trend in the central value moving closer to the central data point. Indeed, a few data points are poorly reproduced, but we highlight to the reader the rather large experimental uncertainties in these measurements. The findings presented in this section demonstrate the importance of a well constrained PDF set. The interplay between datasets is strong, often leading to improved descriptions, at the observable level, in distant regions of phase space. Indeed, such precision is necessarily required in order to achieve strong (5σ) discrepancies needed to drive the search for new physics forward.

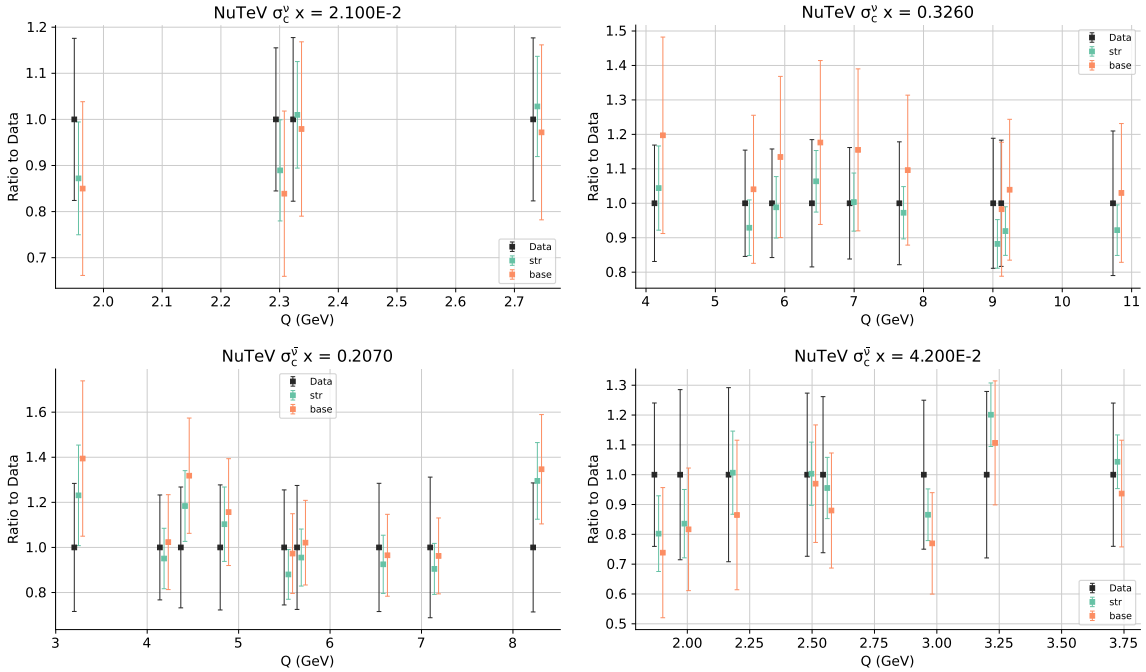


Figure 4.4: Comparison between theoretical predictions and experimental data for the neutrino (upper) and antineutrino (lower) charm dimuon cross sections measured by the NuTeV experiment [176]. The low-x region is shown (left) as well as for high-x (right). The theoretical predictions are normalized to the data central value (black). Predictions are made using `str_base` (orange) and `str` (green).

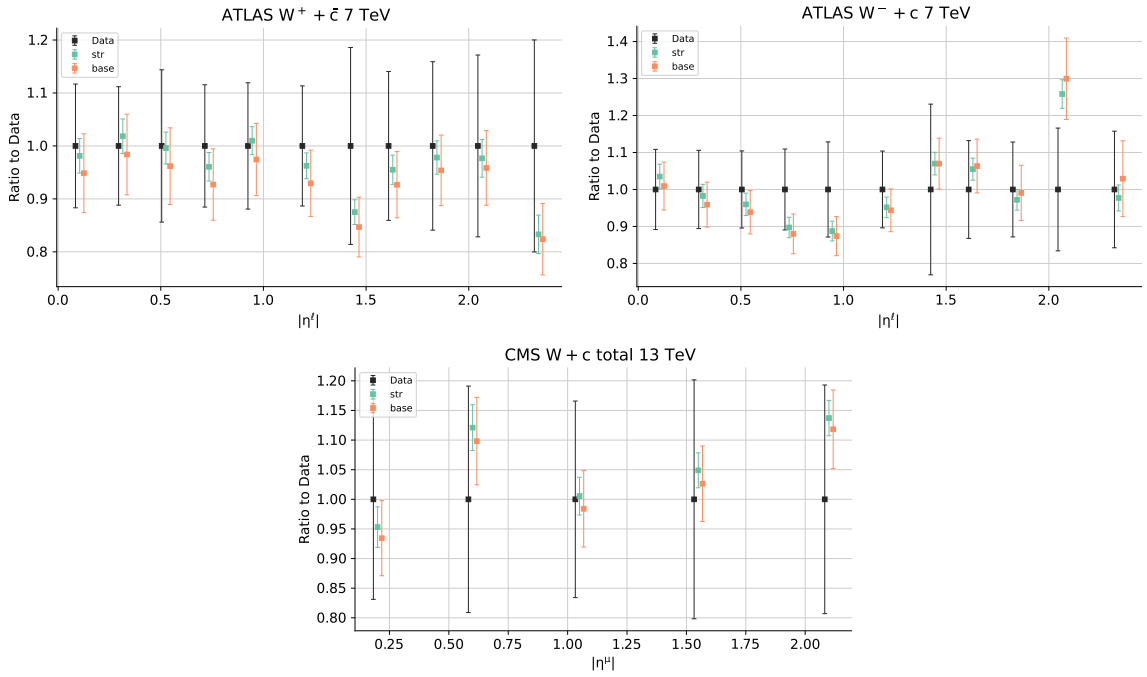


Figure 4.5: Data-theory comparison for ATLAS (top) and CMS (bottom) $W + c$-jet production measured in lepton rapidity.

4.5 PDF sets with precision strange distributions

We now consider the impact of the enhanced strange-sensitivty on the PDFs themselves. In figure 4.6 we plot the strange and anti-strange distributions as a function of Bjorken-x at $Q = 10$ GeV. We present the three different PDFs considered in this work as well as the relative uncertainties defined by $\delta s^i/s^i$, where δs^i is the uncertainty of PDF set i and s^i is the central value of the reference PDF set . The inclusion of the LHC datasets in `str_prior` does not greatly alter the central value of the PDF, lying well within the 1σ uncertainty band of the baseline PDF. The effect of this dataset, however, is to narrow the PDF uncertainty across the entire x range giving a more precise determination of the strange distribution. Introducing the NOMAD data causes the strange distributions to be suppressed in the region $x \gtrsim 0.1$ corresponding to the kinematic region of the NOMAD measurements. The PDF uncertainty is reduced by up to a third in the same region as a result.

In figure 4.7 we show a similar plot, except this time for the up and gluon PDFs. We see a modification of the gluon PDF in the small-x region corresponding to processes such as that shown in figure 4.2. Indeed, the deflection is within confidence levels, though not negligible; the uncertainty, however, remains identical in all 3 fits.

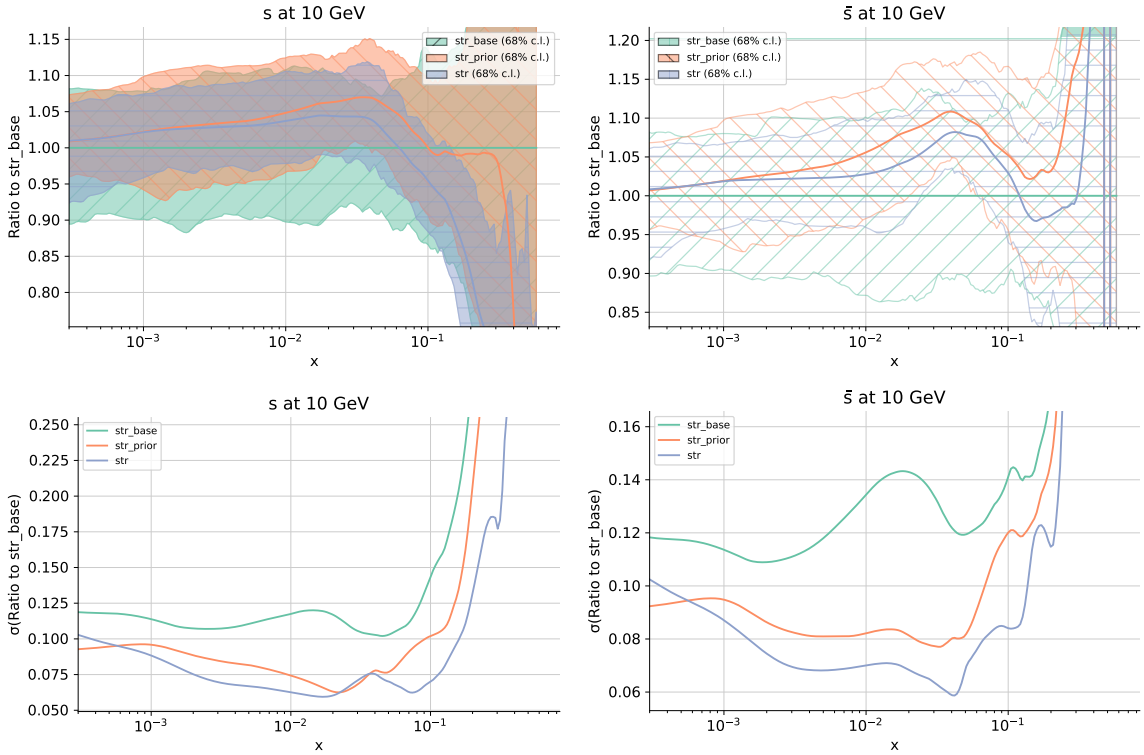


Figure 4.6: The total strange (left) and anti-strange (right) PDFs (top panel), for the 3 fits considered in this work: `str_base` (green), `str_prior` (orange), and `str` (blue) at $Q = 10$ GeV. We also plot the relative PDF uncertainties (lower panel) defined by $\delta s^i / s^i$. We normalize these plots to `str_base`.

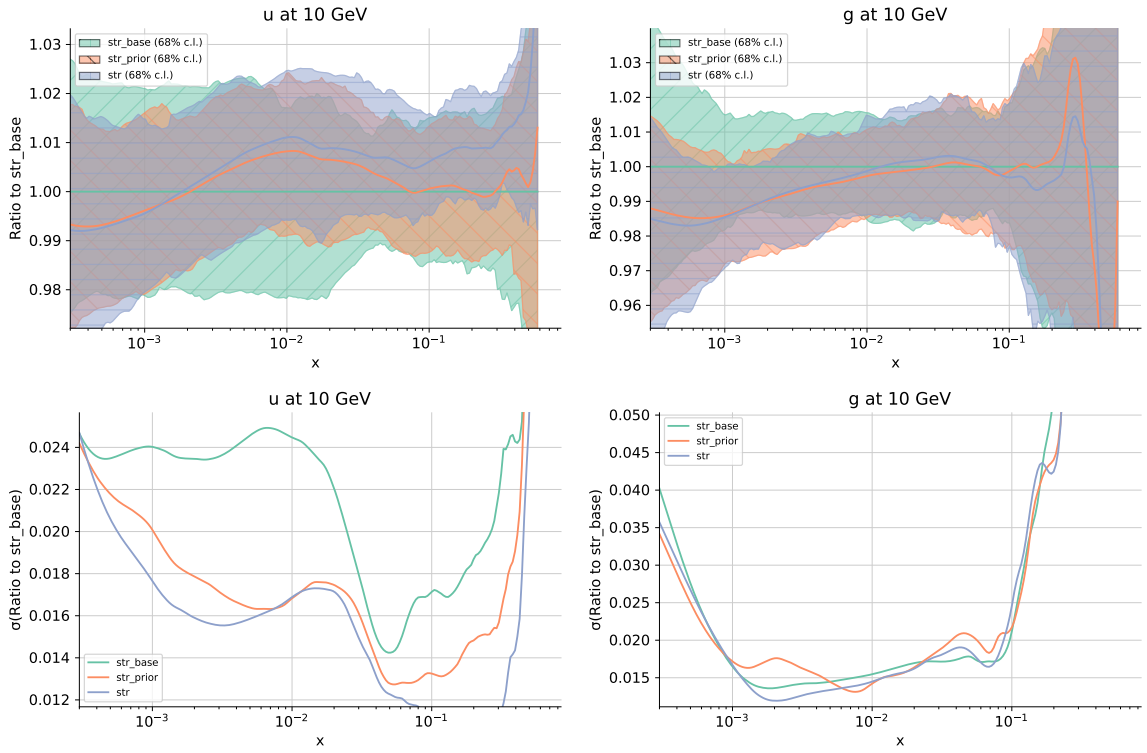


Figure 4.7: Same as figure 4.6, but for up (left) and gluon (right) distributions.

The up (and other light) quarks, however, are altered in central value very little, with the LHC datasets constraining instead the uncertainty. Indeed, a considerable reduction in uncertainty levels is seen in the $x \lesssim 0.1$ region corresponding to the LHC measurements.

4.6 Strangeness ratio

We now return to the earlier mention of the strangeness ratio defined in equation 4.1. This is a measure of the fraction of sea-quarks that are strange or antistrange. A closely related object is $K_s(Q^2)$ which measures the momentum fraction carried by the strange quarks (excluding contributions from the gluons, up, and down quarks):

$$R_s(x, Q^2) = \frac{s(x, Q^2) + \bar{s}(x, Q^2)}{\bar{u}(x, Q^2) + \bar{d}(x, Q^2)} \tag{4.11}$$

$$K_s(Q^2) = \frac{\int_0^1 dx\, x\, [s(x, Q^2) + \bar{s}(x, Q^2)]}{\int_0^1 dx\, x\, \left[\bar{u}(x, Q^2) + \bar{d}(x, Q^2)\right]}. \tag{4.12}$$

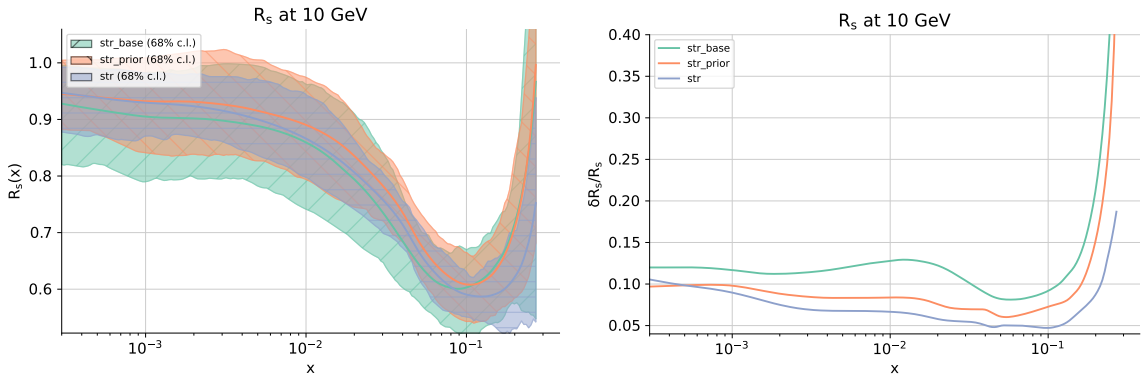


Figure 4.8: The strangeness ratio, defined in equation 4.11 as a function of Bjorken-x at $Q = 10$ GeV (left). The relative uncertainties are also plotted (right). We use `str_base` (green), `str_prior` (orange), and `str` (blue) for the computation.

We plot R_s as well as its relative uncertainties [3] at $Q = 10$ GeV in figure 4.8 for the various PDFs considered in this work. The effect of the LHC data is to cause R_s to rise slightly from the baselines central value, though the effect is very mild. The NOMAD data has a similarly slight impact, but for $x \gtrsim 0.1$ the value of R_s is suppressed, in keeping with the strange-PDF suppression of figure 4.6. Similarly, the uncertainties are reduced (right of figure 4.8) by approximately 4% in $x \lesssim 0.1$ with the inclusion of the LHC data and further suppressed by the NOMAD data in the $x \gtrsim 0.1$ region corresponding to the NOMAD phase space. At sufficiently high-x the PDFs are essentially unconstrained due to a lack of data points and so the PDF uncertainties diverge: this is typical behaviour in the so-called *extrapolation region*.

We compare our strangeness determination with the ATLAS strangeness determination [172, 173]. A value of $R_s = 1.13 \pm 0.11$ at $x = 0.023$ and $Q = 1.6$ GeV [4] was reported using a combined analysis of HERA deep inelastic scattering data as well as the inclusive gauge boson production of the same study. The PDF fit was done using the `xFitter` framework [203]. This result suggests a much more prominent strange sea, while results using the CT18, CT18A [35], MMHT14 [36], and ABMP16 [37] PDF sets suggest a far more suppressed strange sea. We compare our values of strangeness ratio against those of the above PDF sets in figure 4.9 as well as K_s of equation 4.12 for representative value of kinematic variables. We see that our results for R_s and K_s are compatible with the above PDF sets, but not with the ATLAS ratio delivered in [173]. The reason for this can likely be attributed to their using a reduced dataset

[3] Defined by $\delta R_s / R_s$.

[4] This particularly peculiar choice of kinematic point corresponds to the region of enhanced sensitivity of the study's inclusive gauge boson production dataset.

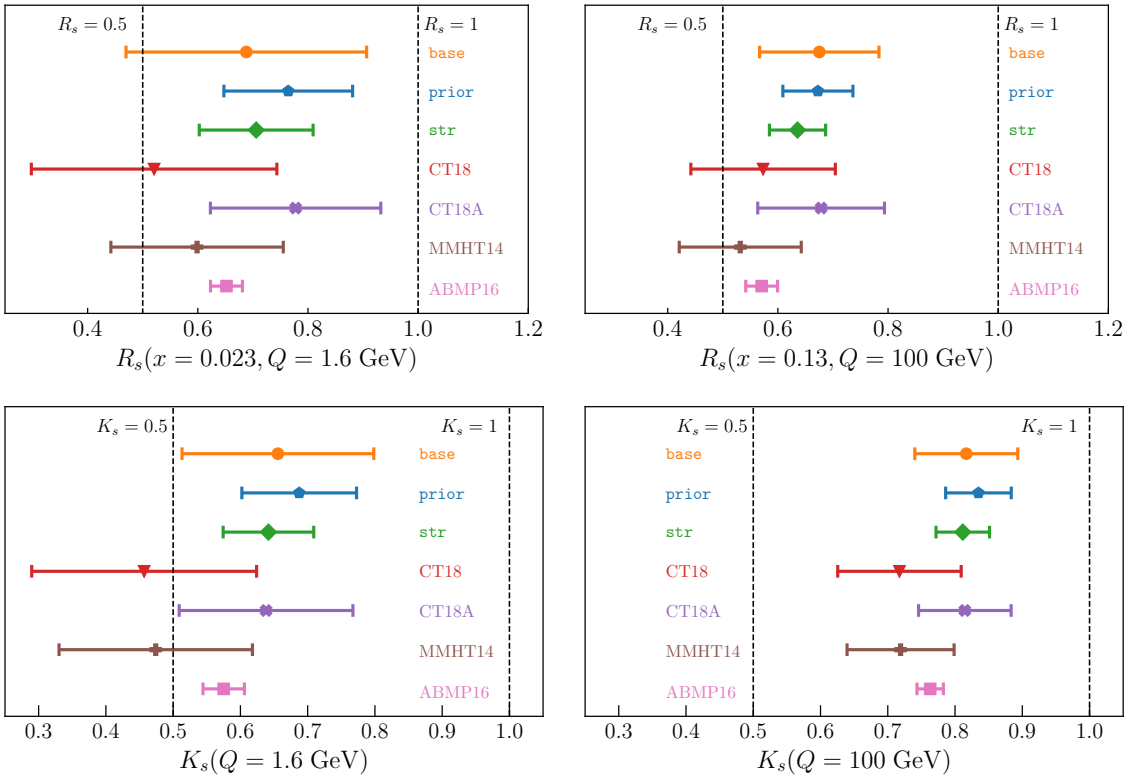


Figure 4.9: The strangeness ratio (top) and momentum fraction (bottom) as defined in equations 4.11 and 4.12. Comparisons are made using the `str_base`, `str_prior`, and `str` PDF sets of this work (respectively: orange, blue, and green) as well as PDF sets using the CT18(A) [35], MMHT14 [36], and ABMP16 [37] methodologies (respectively: red, purple, brown, and pink). Shown also are the 1σ PDF uncertainties. The values of R_s are at $x = 0.023$ and $Q = 1.6$ GeV (left) and $x = 0.13$ and $Q = 100$ GeV (right). Likewise K_s is shown for $Q = 1.6$ GeV (left) and $Q = 100$ GeV (right).

or the more restricted functional form PDF fitting methodology [203]. For the `str` PDF set we observe a reduction in both the central value and uncertainties of R_s for larger values of Bjorken-x corresponding to the enhanced sensitivity of NOMAD data at high-x. Similar arguments apply to K_s with the large-Q behaviour explained by the extended high-Q reach at the LHC. Our results thus suggest a strangeness fraction in between the highly suppressed and strange dominated regimes.

The precision strange determination, referred to as `str` in this text, serves as a deliverable of this study [3] ready for public consumption in the `LHAPDF` format [204]. We have illustrated how the precision determination of sea quark distributions is paramount to LHC phenomenology and how the measurement and use of unfolded, strange sensitive, observables is vital to obtain a good understanding of how strange

the proton is. Similarly, the charm distribution is a topic of similar interest whose precision determination has been studied in a similar fashion as this work [205] with an improved determination, using the NNPDF4.0 methodology, to be made available in a future publication.

Chapter 5

Disentangling new physics effects from PDFs

The PDFs constructed using the NNPDF methodology outlined in chapter 3 implicitly assume the validity of the Standard Model (SM) at all kinematic regions of phase space spanned by the input data. This assumption manifests through the fact that the partonic cross sections and DGLAP evolution kernels of equation 3.31 are computed at some finite order in perturbation theory in the framework of the Standard Model and that even the high-energy observables are not modified by any new physics imprints. The PDFs are in this sense SM PDFs. While new physics corrections are necessarily suppressed at the LHC and thus assuming SM theoretical predictions is fine for phenomenological studies, such as that presented in chapter 4; this assumption is a possible source of inconsistency when a beyond the Standard Model (BSM) analysis is performed. Indeed, the precise nature of such a work mandates that the Standard Model must be extended, for example, by extending the gauge group or matter content of the theory. In turn, this requires that the partonic cross sections or DGLAP evolution to be corrected accordingly, possibly yielding significantly different PDFs.

While a direct search for BSM physics does not suffer from this problem (looking instead for a Breit-Wigner resonance in event count atop a continuously decreasing SM background [15, 16]) indirect searches must be most cautious of this inconsistency. In indirect searches, one attempts to exploit subtle deviations from the SM signal which can be induced by possible heavy degrees of freedom beyond the direct collider kinematical reach. Indirect approaches to constraining BSM dynamics often make use of the effective field theory framework to study the higher-dimensional operators

which are generated by integrating out heavy degrees of freedom (see section 2.3 and references therein) which thereby restricts the class of UV completions which may exist. The EFT then provides a powerful tool to identify and parameterize the new physics in a model-independent, bottom-up, manner. It is clear then that not only is a firm grasp on experimental uncertainties tantamount to the success of a BSM study, but so too is the precision of theoretical ingredients such as the perturbative order of the partonic cross section or PDFs and their uncertainties. As such, simply omitting the BSM sensitive datasets used in an indirect search from a PDF fit is unsatisfactory since one loses constraining power on the PDFs: a quantity whose precision is of utmost importance. On the other hand, inclusion of the BSM datasets in both is perhaps even more unsatisfactory since the logical inconsistency may lead to incorrect conclusions.

This apparent dichotomy is the concern of this chapter. We will assess the question of how the PDF and EFT interplay may affect conclusions on bounds for BSM physics and answer the question of whether the new physics effects can be *reabsorbed* into a flexible parameterization of the PDFs. In what follows, section 5.1 describes how one can explore the space of EFT parameters in a way that the back-reaction of non-zero Wilson coefficients on the PDFs can be studied. Section 5.2, shows how bounds on the Wilson coefficients are then obtained, accounting for the PDF uncertainties as well as the finite size uncertainties owing to the fact that a finite ensemble of PDF MC replicas are used. In section 5.3, we then deploy this methodology to study the interplay of SMEFT operators and the PDFs using deep inelastic scattering data. Section 5.4 then extends this to include high-mass Drell-Yan processes from the LHC.

5.1 PDF exploration of EFT space

In this section we will outline the procedure for how one can explore the EFT parameter space and simultaneously account for the back-reaction effect on the PDFs. For our studies we will employ the Standard Model Effective Field Theory (SMEFT) lagrangian of equation 2.82 due to the numerous advantages this approach enjoys (outlined in section 2.3.4), while the prescription below can equally well be applied to other EFTs such as, for example, the HEFT. The general approach corresponds to picking a subset of $N_{\rm op}$ non-renormalizable operators with corresponding Wilson coefficients

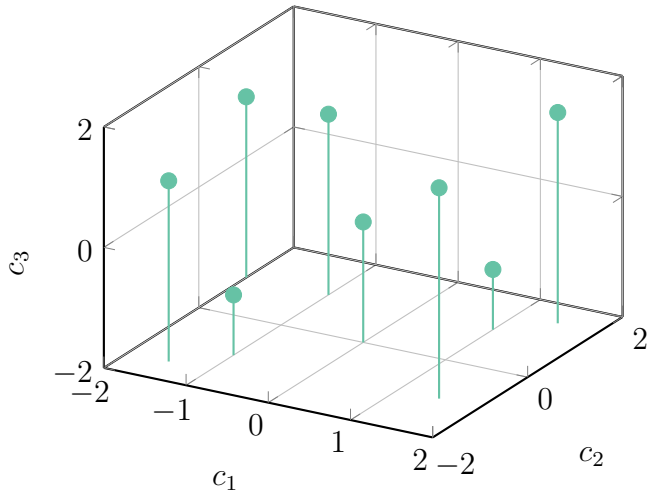


Figure 5.1: Schematic depiction of an example choice of benchmark points corresponding to a BSM scenario with 3 Wilson coefficients. Each mark depicts a particular choice of benchmark point in arbitrary units. The point $(0, 0, 0)$ corresponds to the SM. The vertical lines are to aid the reader.

$\{c_1, \ldots, c_{N_{\text{op}}}\}$. We shall collect these Wilson coefficients in vector form as

$$\mathbf{c} = \begin{pmatrix} c_1 \\ \vdots \\ c_{N_{\text{op}}} \end{pmatrix} \tag{5.1}$$

which we imagine as living in a N_{op} dimensional vector space. An exploration of this vector space can be made by performing a scan of the Wilson coefficients. We do so by choosing a set of *benchmark points*:

$$\mathcal{P} = \{\mathbf{c}^1, \ldots, \mathbf{c}^{N_{\text{op}}}\} \tag{5.2}$$

chosen to be sufficiently well spread so as to achieve a good resolution of the Wilson coefficient space, but not too far from the SM point (the origin) in order to avoid spoiling the validity of the perturbative expansion. An illustration in the 3-dimensional case is shown in figure 5.1.

We initially construct a PDF set by assuming as theory input the SM ($\mathbf{c} = \mathbf{0}$). Correspondingly the theory prediction entering the figure of merit of equation 3.11 is the purely SM prediction. This is the typical case whereby one assumes the PDF is not sensitive to the BSM dynamics. We shall refer to the outcome of a PDF fit determined in this way as SM PDFs. Thus the theory prediction, $\mathbf{t}$, varies upon moving in Wilson

coefficient space only through the partonic cross section and so we have an explicit dependence on $\mathbf{c}$: $\mathbf{t} = \mathbf{t}(\mathbf{c})$. We generalize this notion further by supposing that the PDF also varies as we explore the EFT parameter space. This corresponds to the fully consistent scenario, whereby the training of the PDFs is done in the presence of non-vanishing EFT operators in equation 3.11, with the couplings determined by the particular choice of benchmark point, $\mathbf{c}$. The PDF then acquires implicit dependence on the Wilson coefficient, which we write as $f(\mathbf{c})$. As such, the theory prediction vector acquires dependence on the Wilson coefficients not only through the partonic cross section, but also implicitly through the PDF dependence on $\mathbf{c}$: $\mathbf{t} = \mathbf{t}(\mathbf{c}, f(\mathbf{c}))$. The PDF sets constructed in this way will be referred to as SMEFT PDFs henceforth: with one SMEFT PDF set for each benchmark point.

We highlight here that at the time of performing the work in the present chapter, the NNPDF4.0 methodology and PDF sets were in their infancy. As such, the work done here is performed using the NNPDF3.1 approach. The major difference between the two approaches is the use of genetic algorithms [141] to optimize the figure of merit rather than the gradient descent based approaches, discussed in chapter 3.

5.2 The Hessian approach to EFT bounds

Having performed the scan of EFT space by trialing various benchmark points, we now turn our attention to how this process can give us bounds on the Wilson coefficients.

We pick a set of benchmark points $\mathcal{P} = \{\mathbf{c}^i : i = 1, \ldots, N_{\text{BP}}\}$ and for each element of this set, we compute a χ^2 value to the data:

$$\chi^2_i = \Big(\mathbf{d} - \mathbf{t}(\mathbf{c}^i, f(\mathbf{c}^i))\Big)^T C^{-1} \Big(\mathbf{d} - \mathbf{t}(\mathbf{c}^i, f(\mathbf{c}^i))\Big) \qquad i = 1, \ldots, N_{\text{BP}} \tag{5.3}$$

where $\mathbf{d}$ is the vector of experimental measurements and $\mathbf{t}(\mathbf{c}^i, f(\mathbf{c}^i))$ the corresponding vector of theoretical predictions, computed as the mean across replicas. Note that in the case of SM PDFs, the PDF used to compute $\mathbf{t}$ does not depend on $\mathbf{c}^i$. The covariance matrix, C, is constructed under the t_0 prescription outlined in section 3.2.1.

For reasons that will become apparent shortly, we wish to treat the χ^2 as a continuous function of the Wilson coefficients. For now it is only known at each benchmark point. We do so by fitting a functional form to the list of known, but discrete, χ^2 values given by computing equation 5.3 for each benchmark point. To motivate the functional form

we use, note that the χ^2 can be written near its minimum as:

$$\chi^2(\mathbf{c}) = \chi^2_0 + \frac{1}{2}\sum_{p,q=1}^{N_{\text{op}}} (\mathbf{c}-\mathbf{c}_0)_p(\mathbf{c}-\mathbf{c}_0)_q \left.\frac{\partial^2\chi^2}{\partial c_p \partial c_q}\right|_{\mathbf{c}=\mathbf{c}_0} + \mathcal{O}(|\mathbf{c}-\mathbf{c}_0|^3) \tag{5.4}$$

where $\mathbf{c}_0$ is the location of the minimum and χ^2_0 is the value of the χ^2 at this minimum. Note that in the case of SM PDFs and truncating at $\mathcal{O}(1/\Lambda^2)$, this functional approximation is exact globally. The reason for this is that the theory predictions are at most linear in the Wilson coefficients and so the χ^2 is at most quadratic. Thus, the quadratic form above exactly captures the χ^2 dependence on $\mathbf{c}$. However, in the case of SMEFT PDFs, it is an approximation, justified by the fact that we are expanding the χ^2 around the minimum and restricting to within a small neighborhood about this point. The non-linear dependence of the PDFs on $\mathbf{c}$ being the reason why the above argument does not hold in the SMEFT PDF case.

In principle, higher order terms in equation 5.4 are subleading in the neighborhood of the minimum. However, it is often advantageous to include quartic terms in the case where one considers the effect of $\mathcal{O}(1/\Lambda^4)$ terms in the cross section. Again, in the case of SM PDFs, the functional form thus becomes exact, while for SMEFT PDFs it remains an approximation. This will be required in the discussion of section 5.4.6 and the following arguments must then be replaced by a numerical optimization approach.

Writing the Hessian, H_{pq}, as the matrix of mixed derivatives (with the factor of 1/2) we may write this as:

$$\chi^2(\mathbf{c}) = \chi^2_0 + (\mathbf{c}-\mathbf{c}_0)^T H (\mathbf{c}-\mathbf{c}_0) + \cdots . \tag{5.5}$$

We then treat χ^2_0, $\mathbf{c}_0$, and H as parameters to be fitted to the known values of the function at the benchmark points as computed by equation 5.3. We package these fit parameters in vector form as

$$\beta = (\chi^2_0,\ \mathbf{c}_0,\ H) \tag{5.6}$$

such that we treat the functional form of equation 5.5 to additionally be a function of β, $\chi^2(\mathbf{c};\beta)$. Note that the Hessian is symmetric and so we have

$$\dim\beta = 1 + N_{\text{op}} + \frac{1}{2}N_{\text{op}}(N_{\text{op}}+1) = 1 + \frac{1}{2}N_{\text{op}}(N_{\text{op}}+3). \tag{5.7}$$

We can obtain β by minimizing the usual ordinary least squares function:

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N_{\mathrm{BP}}} \left| \chi_i^2 - \chi^2(\mathbf{c}^i; \beta) \right|^2 . \tag{5.8}$$

The minimization problem of equation 5.8 in fact has a closed form solution. To see this note that we may write equation 5.5 as:

$$\chi^2(\mathbf{c}) = \begin{pmatrix} 1 & -2c_p & c_p c_q \end{pmatrix} \begin{pmatrix} \chi_0^2 + \mathbf{c}_0^T H \mathbf{c}_0 \\ H_{pq}(\mathbf{c}_0)_q \\ H_{pq} \end{pmatrix} \tag{5.9}$$

where summation on repeated indices after the tensor multiplication is implied. The convention here is that $p, q = 1, \ldots, N_{\mathrm{op}}$ indices run over components of $\mathbf{c}$ while $i = 1, \ldots, N_{\mathrm{BP}}$ runs over the benchmark points. The vector of χ^2 values evaluated at all the benchmark points thus reads:

$$\begin{pmatrix} \chi^2(\mathbf{c}^1) \\ \vdots \\ \chi^2(\mathbf{c}^{N_{\mathrm{BP}}}) \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & -2c_p^1 & c_p^1 c_q^1 \\ \vdots & \vdots & \vdots \\ 1 & -2c_p^{N_{\mathrm{BP}}} & c_p^{N_{\mathrm{BP}}} c_q^{N_{\mathrm{BP}}} \end{pmatrix}}_{X} \underbrace{\begin{pmatrix} \chi_0^2 + \mathbf{c}_0^T H \mathbf{c}_0 \\ H_{pq}(\mathbf{c}_0)_q \\ H_{pq} \end{pmatrix}}_{\delta} \tag{5.10}$$

which defines the (known) *design matrix* X and allows us to write the functional form of equation 5.5 to be linear in the newly defined covariates δ, which are themselves a non-linear, though invertible, function of the original parameters of interest β.

The solution to equation 5.8 is then the well known solution from linear regression:

$$\hat{\delta} = (X^T X)^{-1} X^T \cdot \begin{pmatrix} \chi^2(\mathbf{c}^1) \\ \vdots \\ \chi^2(\mathbf{c}^{N_{\mathrm{BP}}}) \end{pmatrix} \tag{5.11}$$

which, by the Gauss-Markov theorem, will give the best [1] linear unbiased estimator for δ. Having computed $\hat{\delta}$, we can invert the relation to find $\hat{\beta}$ by using the values of H_{ij} to solve for $\mathbf{c}_0$ and then solving for χ_0^2.

[1] In the sense that it has the least sampling variance.

5.2.1 Confidence intervals

Having fitted the functional form of equation 5.5 to the grid of benchmark points, the maximum likelihood estimator (assuming the data is distributed according to a multivariate normal distribution) for the Wilson coefficients, $\hat{\mathbf{c}}$, is then given by:

$$\hat{\mathbf{c}} = \arg\min \chi^2(\mathbf{c}). \tag{5.12}$$

However, in general we will be more interested in a $\alpha\%$ ($0 \leq \alpha \leq 100$) confidence interval rather than a best fit value. This will be a random interval that will, on average, contain the true Wilson coefficients $\alpha\%$ of the time samples are made.

This can be done using the prescription outlined in [206]. The region in parameter space described by the subspace of Wilson coefficients:

$$\mathcal{R}_\alpha = \{\mathbf{c} : \chi^2(\mathbf{c}) - \chi^2_0 \leq \Delta_{\alpha,N_{\text{op}}}\} \tag{5.13}$$

contains all the points in Wilson coefficient space corresponding to a confidence level of α. The constant $\Delta_{\alpha,N_{\text{op}}}$ represents the increment in χ^2 corresponding to a (α significance) deviation from the maximum likelihood estimator. It can be obtained from any standard χ^2 distribution table and corresponds to the inverse of the cumulative function for a χ^2 distribution having N_{op} degrees of freedom. Using equation 5.5, this has a very simple geometric interpretation:

$$\chi^2(\mathbf{c}) - \chi^2_0 = (\mathbf{c} - \mathbf{c}_0)^T H (\mathbf{c} - \mathbf{c}_0) = \Delta_{\alpha,N_{\text{op}}} \tag{5.14}$$

and remembering that H is the Hessian matrix evaluated at a minimum, then equation 5.14 is a quadratic form tracing out a N_{op} dimensional ellipsoid which is the boundary of the confidence region $\mathcal{R}_\alpha$. In 1 dimension this surface is an interval contained within the real number line, while in 2 dimensions this surface corresponds to an ellipse. The marginalized bounds for each of the orthogonal directions are then given by projecting the ellipse onto each axis as shown in figure 5.2 [207]. Equivalently the ellipse is placed within a minimally enclosing hypercube (square in two dimensions) that is aligned with the principal directions. The dimensions of this hypercube then yield the confidence interval for the corresponding directions. This problem can be formulated using constrained optimization. We wish to extremize c_i subject to the constraint that

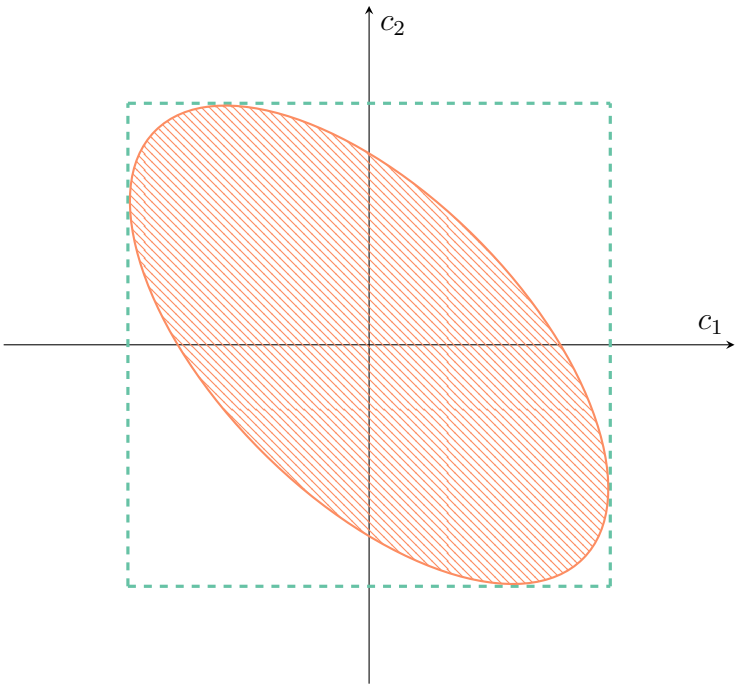


Figure 5.2: An example of an ellipse in a 2-dimensional Wilson coefficient space which forms the boundary of a confidence region (orange). The green lines form the minimally enclosing hypercube, the dimensions of which give the bounds for the marginalized directions and are given by the solutions to equations 5.19 and 5.20.

we remain on the boundary of $\mathcal{R}_\alpha$. The target function is then:

$$\mathcal{L}_i = c_i - \lambda_i\Big((\mathbf{c}-\mathbf{c}_0)^T H(\mathbf{c}-\mathbf{c}_0) - \Delta_{\alpha,N_{\text{op}}}\Big) \tag{5.15}$$

where λ_i is a Lagrange multiplier and we have one for each direction (operator). Differentiating this target function then gives:

$$\frac{\partial \mathcal{L}_i}{\partial c_j} = \delta_{ij} - 2\lambda_i H_{jk}(\mathbf{c}-\mathbf{c}_0)_k = 0 \tag{5.16}$$

which gives the set of equations

$$1 - 2\lambda_i\, H_{ik}(\mathbf{c}-\mathbf{c}_0)_k = 0 \qquad \text{No sum on } i \tag{5.17}$$

$$-2\lambda_i\, H_{jk}(\mathbf{c}-\mathbf{c}_0)_k = 0 \qquad \forall j \neq i. \tag{5.18}$$

The Lagrange multiplier is then strictly non-vanishing for any direction, i, that we consider, since otherwise the first of the two equations would give a contradiction.

These conditions then reduce further to give:

$$H_{jk}(\mathbf{c}-\mathbf{c}_0)_k = 0 \qquad \forall j \neq i \tag{5.19}$$

$$(\mathbf{c}-\mathbf{c}_0)^T H(\mathbf{c}-\mathbf{c}_0) = \Delta_{\alpha,N_{\rm op}} \tag{5.20}$$

which gives two solutions for c_i corresponding to the upper and lower bounds. The implementation of equations 5.19-5.20 has been done using the symbolic mathematics library `SymPy` [208].

5.2.2 Including PDF uncertainty

For the case of the fixed SM PDF analysis, it is important to consider the effect of the PDF uncertainty on the bounds obtained. The prescription for this is very akin to the above discussion, except for the fact that bounds are obtained for each of the $N_{\rm rep}$ replicas rather than the central replica. One thus obtains a set of bounds

$$\left[c_{\rm min}^{(k)}, c_{\rm max}^{(k)}\right] \qquad k = 1, \ldots, N_{\rm rep} \tag{5.21}$$

which gives the bounds with 1σ-PDF uncertainty after taking the 68% envelope or 2σ-PDF uncertainty with the 95% envelope. The case for SMEFT modified PDFs, however, incorporates the PDF uncertainty by construction since it was constructed from a global set of PDFs.

5.2.3 Methodological uncertainty and the bootstrap method

As discussed above, while exact for the SM PDFs, the functional form of equation 5.5 is an approximation for the SMEFT PDF case inspired by truncating the Taylor expansion around the EFT minimum. The reason for this can be traced backed to the fact that the χ^2 is not only dependent on the Wilson coefficients through the partonic cross section, but implicitly through the fact that the PDFs too are non-linear functions of $\mathbf{c}$. As such, for the SMEFT modified PDF analysis, one must consider the impact of the finite number of replicas used in the ensemble. Indeed, computing $\chi^2(\mathbf{c}_i)$ for a given ensemble will vary depending on which member replica is used in the calculation. The uncertainty from this finite sampling of an unknown population can be computed using the bootstrap method [174, 209].

We provide a general overview of the bootstrap prescription before applying it specifically to our particular use case concerning the statistical fluctuations of the χ^2 in varying the ensemble member.

Consider independent and identically distributed random variables $X_1, \ldots, X_n$ each taking values in $\mathcal{X}$ with common cumulative distribution function F. We can construct an estimator for some unknown parameter θ as a function, T, of our data

$$\hat{\theta} = T(X_1, \ldots, X_n). \tag{5.22}$$

In general, we will be interested in the accuracy of our point estimate, $\hat{\theta}$, as well as the value itself. Noting that $\hat{\theta}$ is a function of random variables so is itself a random variable, then one may enquire about its variance, $\mathrm{Var}(\hat{\theta})$. If F is known analytically, then we can also compute $\mathrm{Var}(\hat{\theta})$ analytically [2]:

$$\mathrm{Var}(\hat{\theta}) = \int_{\mathcal{X}^n} \left(\hat{\theta} - \mathbb{E}(\hat{\theta})\right)^2 dF \tag{5.23}$$

$$\mathbb{E}(\hat{\theta}) = \int_{\mathcal{X}^n} \hat{\theta}\, dF. \tag{5.24}$$

If instead, the distribution is not known, or the above integrals are intractable, but the distribution is easily sampled from (through Markov Chain Monte Carlo [210, 211] or accept-reject sampling), then a Monte Carlo estimate of the variance is an adequate substitute. In this regime, we construct N samples of the data

$$\{X_1^i, \ldots, X_n^i\} \qquad i = 1, \ldots, N \tag{5.25}$$

which in turn generates N samples of the estimator $\hat{\theta}_i = T(X_1^i, \ldots, X_n^i)$. Then the variance of $\hat{\theta}$ can be estimated using the sample variance:

$$\mathrm{Var}(\hat{\theta}) = \frac{1}{N-1} \sum_{i=1}^{N} \left(\hat{\theta}_i - \frac{1}{N} \sum_{j=1}^{N} \hat{\theta}_j\right)^2 \tag{5.26}$$

where we choose the unbiased (Bessel corrected) estimator for the variance.

However, a ubiquitous scenario occurs when F is sufficiently complex that it is neither known nor easily sampled from: as such any attempt at an analytic calculation or Monte Carlo estimate will prove futile. Instead, we consider the case where we have made a sample of the data once to obtain $\mathcal{B} = \{X_1, \ldots, X_n\}$, but are unable to make

[2] We use the notation that the measure is to be understood as $dF = \prod_{i=1}^{n} F'(x_i) d^n x$

further samples. To proceed we come up with a notion of the cumulative distribution function obtainable purely from $\mathcal{B}$.

Definition 5.1 (Empirical cumulative distribution function) *Let $X_1, \ldots, X_n$ be independent and identically distributed random variables. The empirical cumulative distribution function, $\widetilde{F}(t)$, is defined by*

$$\widetilde{F}(t) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq t) \tag{5.27}$$

where I is the indicator function:

$$I(X_i \leq t) = \begin{cases} 1 & \text{if } X_i \leq t \\ 0 & \text{otherwise} \end{cases}. \tag{5.28}$$

The empirical cumulative distribution function is attractive because it is unbiased:

$$\mathbb{E}\left(\widetilde{F}(t)\right) = \frac{1}{n} \sum_{i=1}^{n} P(X_i < t) = P(X_1 < t) = F(t) \tag{5.29}$$

and by the weak law of large numbers: asymptotically tends to the true cumulative distribution function in the limit of large n. The bootstrapping prescription simply suggests using the empirical distribution instead of the true distribution to sample the data points of equation 5.25. The corresponding sampling distribution is then uniform on the set of measured values, $\mathcal{B}$. The procedure is then straightforward. From the set of measured values, obtain a *bootstrap sample*:

$$\mathcal{B}_i^* = \{X_{1i}^*, \ldots, X_{ni}^*\} \tag{5.30}$$

by sampling uniformly with replacement from $\mathcal{B}$. As such, each $\mathcal{B}_i^*$ differs from the original dataset $\mathcal{B}$ in that it includes duplicates and missing values. We construct $i = 1, \ldots, N_{\text{res}}$ such bootstrap resamples, which gives N_{res} copies of the estimator $\hat{\theta}$:

$$\hat{\theta}_i^* = T(X_{1i}^*, \ldots, X_{ni}^*) \tag{5.31}$$

with the bootstrap variance being estimated by:

$$\text{Var}(\hat{\theta}) = \frac{1}{N_{\text{res}} - 1} \sum_{i=1}^{N_{\text{res}}} \left(\hat{\theta}_i^* - \frac{1}{N_{\text{res}}} \sum_{i=1}^{N_{\text{res}}} \hat{\theta}_i^*\right)^2. \tag{5.32}$$

We now return to the original problem of estimating the statistical fluctuations of $\chi^2(\mathbf{c}_i)$ at each benchmark point, $\mathbf{c}_i$, due to using different PDF ensemble members in the theory calculation. For all the data points, we construct a corresponding vector of theory predictions $\mathbf{t}_k$ using the k'th replica of the ensemble, where $\dim \mathbf{t}_k = N_{\text{data}}$. This generates a set of theory predictions of size N_{rep}:

$$\mathcal{B} = \{\mathbf{t}_1, \ldots, \mathbf{t}_{N_{\text{rep}}}\} \tag{5.33}$$

from which we resample with replacement N_{res} times to obtain the bootstrap samples:

$$\mathcal{B}_i^* = \{\mathbf{t}_{1i}, \ldots, \mathbf{t}_{N_{\text{rep}}i}\} \qquad i = 1, \ldots, N_{\text{res}}. \tag{5.34}$$

For each bootstrap, compute the mean theory prediction:

$$\mathbf{t}_i^* = \frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} \mathbf{t}_{ki} \tag{5.35}$$

and use it to compute the χ^2 to the data to obtain a scalar value for each bootstrap χ_i^{2*}. The variance of the χ^2 is then the sample variance across bootstraps:

$$\text{Var}(\chi^{2*}) = \frac{1}{N_{\text{res}} - 1} \sum_{i=1}^{N_{\text{res}}} \chi_i^{2*}. \tag{5.36}$$

A value for N_{res} of 10^4 was determined to be adequate in order to obtain stable results that are independent on the seeding of the random number generator that performs the bootstrap sampling.

5.3 Constraining the SMEFT with lepton-proton scattering

As an initial proof-of-concept study, we choose to constrain a subset of 4 Warsaw basis operators using DIS data. The precise choice of operators and dataset is discussed in section 5.3.1. The DIS only study provides a convenient initial framework to assess the interplay of PDFs with BSM effects due to a number of reasons. The first is that, thanks to the collider DIS data from HERA [158], a sufficiently broad kinematic coverage of Q is achieved. As such we suspect the highest energy bins to be sensitive to the heavy mass effects and thus provide constraints on the Wilson coefficients; while

the large number of DIS data points (3092 in total) provides strong constraints on the PDFs. Secondly, the modifications to the DIS structure functions arising from these operators can be easily calculated in a perturbative framework. This calculation is presented in section 5.3.2 and subsequently implemented in the `APFEL` library [147]: responsible for generating DIS predictions for the corresponding observable. The calculation is done to leading order in QCD while a separate analysis shows that QCD corrections to the new SMEFT scattering graphs are negligible. The SM partonic cross sections are computed at NNLO, however. The DIS data form approximately 75% of the total global dataset in a PDF fit and is thus very much the backbone of any PDF determination. The assessment of its interplay with BSM dynamics is therefore a crucial first starting point. For this particular study, we question to what extent the neural network parameterization of PDFs is prone to fitting away any possible BSM effects present within DIS data. We then analyse the effect that a consistent treatment of the PDFs has on imposing bounds on the Wilson coefficients. We do so by obtaining bounds assuming a fixed SM PDF, and then compare these bounds when we allow for the PDFs to vary appropriately as we explore the Wilson coefficient space using the approach outline in section 5.2.

5.3.1 Dataset selection and BSM scenario

The operators we choose to study are a subset of the Warsaw basis [108] restricted to the form

$$\mathcal{O}_f = \frac{c_f}{\Lambda^2}(\bar{l}_R\gamma_\mu l_R)\ (\bar{q}_R^f\gamma^\mu q_R^f) \tag{5.37}$$

where l are the charged lepton Dirac spinor fields taken in this study to be the electron or muon field and q^f are the quark fields corresponding to flavour f which are taken to be the independently parameterized up, down, strange and charm flavours. We assume a universal coupling to the lepton families as suggested by LEP precision data [212], but non-universally to the quarks and as such there are 4 independent Wilson coefficients, c_f. The new physics scale is denoted by Λ taken in this study to be 1 TeV well above the highest Q value of 173 GeV reached by the neutral current HERA experiments. Note that the spinor fields have been projected onto their right-handed components, $l_R = \frac{1}{2}(1+\gamma^5)l$, thus transforming trivially under $SU(2)_L$. The operators of the form presented in equation 5.37 form a reasonable case study for analyzing the interplay of BSM dynamics with the parton distribution functions. As is shown in section 5.3.2, the modifications to the neutral current structure functions scale linearly

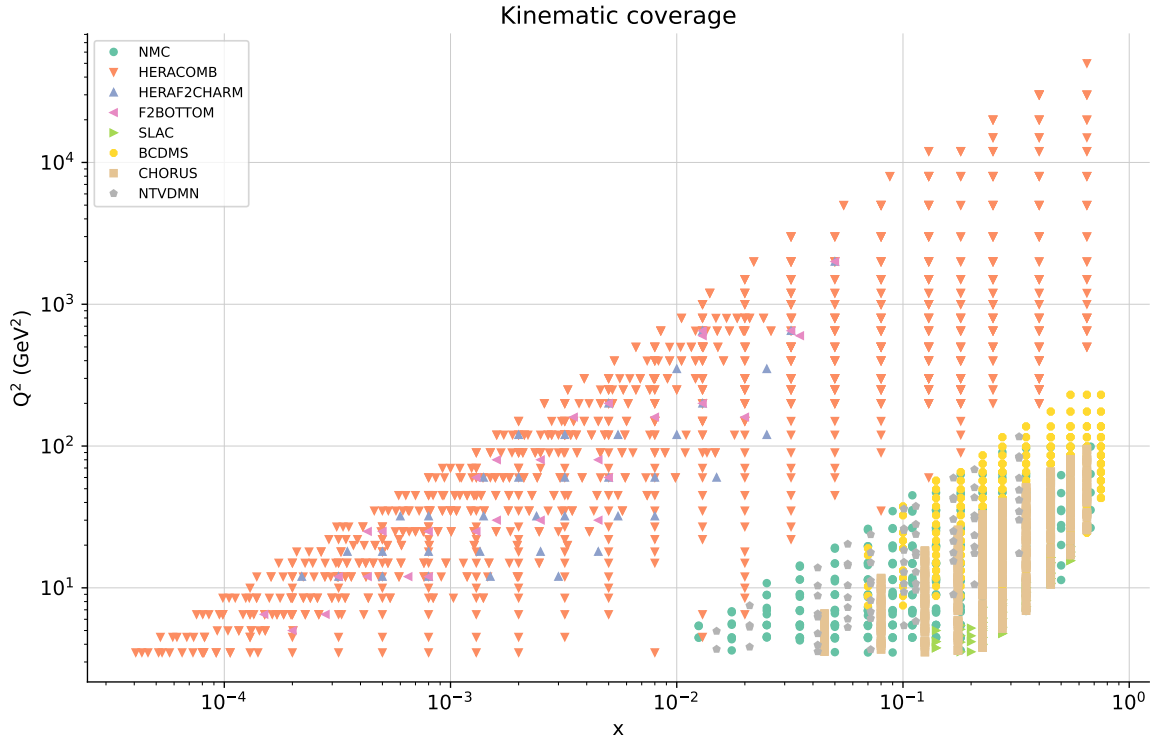


Figure 5.3: Kinematic coverage of the DIS datasets used in this study. The markers are grouped according to their experiment.

in Q^2 (at leading order in the Wilson coefficients) leading to an energy growing effect at the observable level that is faster than the SM: the sensitivity to which should be present within the highest energy DIS measurements available.

As discussed earlier, for this initial study we shall use data from inclusive fixed-target deep inelastic scattering data, from NMC [213, 214], SLAC [215], BCDMS [216], CHORUS [217], NuTeV [175] as well as the HERA combined collider data from H1 and ZEUS [158]. We also include the HERA measurements for heavy flavour (charm and bottom) production [218] which proceeds via gluon-boson fusion and thus constrains the gluon. The kinematic coverage in the (x, Q^2) plane for all these DIS datasets is presented in figure 5.3, with the markers grouped according to their experiment.

5.3.2 SMEFT-modified DIS observables

When computing the full amplitude for neutral-current DIS scattering we have contributions from photon and Z mediated exchange as well as a new contribution from the dimension 6 operators. To maintain generality, we consider fields having chirality λ for

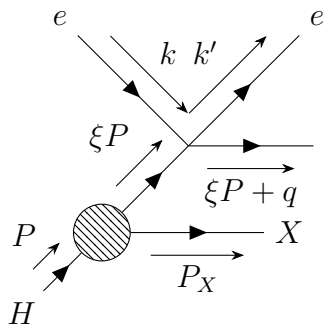


Figure 5.4: Diagram corresponding to deep inelastic scattering of $eH \to eX$ by the added BSM 4-fermion contact interaction in the parton model. The variable $q = k - k'$ is defined to be the exchanged momentum.

the leptons and λ' for the quarks:

$$\mathcal{O}_f = \frac{c_f}{\Lambda^2} \bar{l} \gamma_\mu (1 + \lambda \gamma^5) l \; \bar{q}^f \gamma^\mu (1 + \lambda' \gamma^5) q^f . \tag{5.38}$$

The squared amplitude then introduces 3 additional terms not present in the Standard Model calculation:

$$\begin{aligned} |\mathcal{A}|^2 = &|\mathcal{A}_\gamma|^2 + |\mathcal{A}_Z|^2 + 2\mathrm{Re}(\mathcal{A}_\gamma \mathcal{A}_Z^*) + \\ &|\mathcal{A}_{\mathrm{BSM}}|^2 + 2\mathrm{Re}(\mathcal{A}_{\mathrm{BSM}} \mathcal{A}_\gamma^*) + 2\mathrm{Re}(\mathcal{A}_{\mathrm{BSM}} \mathcal{A}_{\mathrm{Z}}^*). \end{aligned} \tag{5.39}$$

Suppressing the spinor indices, the Feynman diagrams evaluate to read:

$$i\mathcal{A}_\gamma = (-ie)\bar{u}(k')\gamma^\mu u(k) \left(\frac{-ig_{\mu\nu}}{q^2} \right) \langle H(P)|j^\nu_{\mathrm{EM}}(0)|X(P_X)\rangle \tag{5.40}$$

$$i\mathcal{A}_Z = \left(-\frac{ig}{\cos\theta_W} \right) \bar{u}(k')\gamma^\mu (V_Z - A_Z\gamma^5) u(k) \left(\frac{-ig_{\mu\nu}}{q^2 - M_Z^2} \right) \langle H(P)|j^\nu_{\mathrm{Z}}(0)|X(P_X)\rangle \tag{5.41}$$

$$i\mathcal{A}_{\mathrm{BSM}} = (-i)\bar{u}(k')\frac{1}{2}\gamma_\mu (1 + \lambda\gamma^5) u(k) \langle H(P)|j^\mu_{\mathrm{BSM}}(0)|X(P_X)\rangle . \tag{5.42}$$

with the vector and axial couplings with the Z boson being:

$$V_Z = \frac{1}{2} I_3 - e \sin^2\theta_W \qquad A_Z = \frac{1}{2} I_3 \tag{5.43}$$

where e is the particles electric charge in units of the positron charge, θ_W is the weak mixing angle and I_3 is the third component of the weak isospin. We mention here the fact that the BSM vertex is a 4-fermion contact interaction. As such, the Feynman rule for the hadronic matrix element must omit the factor of $-i$ since it has already been

included in the leptonic part of the matrix element. Similarly, for notational clarity, we have left the c_f/Λ^2 factor with the hadronic current insertion. The calculation of all three additional contributions follows the same procedure as one another and so we focus here on the SMEFT interference term with the photon. This term corresponds to a cross term and so we must recall to take the real part and multiply by 2. This term alone then reads:

$$\mathcal{A}_{\rm BSM}\mathcal{A}^*_\gamma = -\frac{ie}{Q^2}\bar{u}(k')\frac{1}{2}\gamma_\mu(1+\lambda\gamma^5)u(k)\,\bar{u}(k)\gamma_\nu u(k') \\ \langle H(P)|j^\mu_{\rm BSM}(0)|X(P_X)\rangle\,\langle X(P_X)|j^{\mu\dagger}_{\rm EM}(0)|H(P)\rangle\,. \quad (5.44)$$

Remembering that in order to compute a cross section, this term lives inside a spin sum average, we can compute the leptonic tensor as before [3]:

$$L_{\mu\nu} = -\frac{ie}{2Q^2}{\rm Tr}\Big(\not{k}'\frac{1}{2}\gamma_\mu(1+\lambda\gamma^5)\not{k}\gamma_\nu\Big) \quad (5.45)$$

$$= -\frac{ie}{Q^2}\Big(k'_\mu k_\nu - (k'\cdot k)g_{\mu\nu} + k_\mu k'_\nu + i\lambda k'^\beta k^\alpha\epsilon_{\mu\nu\alpha\beta}\Big). \quad (5.46)$$

We now turn our attention to the hadronic tensor. As before, this term will be computed within the parton model:

$$W^{\mu\nu} = \frac{1}{8xP\cdot q}\sum_q f_q(x)\sum_{\rm spins}\langle q(xP)|j^\mu_{\rm BSM}(0)|q(xP+q)\rangle\,\langle q(xP+q)|j^{\nu\dagger}_{\rm EM}(0)|q(xP)\rangle \quad (5.47)$$

with the matrix elements now being computed using pQCD:

$$W^{\mu\nu} = \frac{ie}{8xP\cdot q}\sum_q f_q(x)\frac{c_f Q_f}{\Lambda^2}\sum_{\rm spins}\bar{u}(xP+q)\frac{1}{2}\gamma^\mu(1+\lambda'\gamma^5)u(xP)\bar{u}(xP)\gamma^\nu u(xP+q). \quad (5.48)$$

It is important to highlight here that the matrix elements have been computed assuming the ejected parton is a quark. In the case of an anti-quark, the form is much the same except for the fact that the momenta must be swapped around and the positive frequency modes, u, must be replaced with the negative frequency modes, v. We shall

[3] With the additional help of ${\rm Tr}(\gamma_\mu\gamma_\nu\gamma_\rho\gamma_\sigma\gamma^5) = -4i\epsilon_{\mu\nu\rho\sigma}$.

comment further on this momentarily. The spin-sum average simplifies to read:

$$\begin{aligned}
&\sum_{\text{spins}} \bar{u}(xP+q)\frac{1}{2}\gamma^{\mu}(1+\lambda'\gamma^5)u(xP)\bar{u}(xP)\gamma^{\nu}u(xP+q) \\
&\quad = \frac{1}{2}\mathrm{Tr}\Big((x\not{P}+\not{q})\gamma^{\mu}(1+\lambda'\gamma^5)x\not{P}\gamma^{\nu}\Big) \\
&\quad = 2(x^2P_{\alpha}P_{\beta}+xq_{\alpha}P_{\beta})\Big(g^{\alpha\mu}g^{\beta\nu}-g^{\alpha\beta}g^{\mu\nu}+g^{\alpha\nu}g^{\mu\beta}-i\lambda'\epsilon^{\alpha\mu\beta\nu}\Big).
\end{aligned} \tag{5.49}$$

At this point a pause for reflection will prove fruitful. Recall that in equation 5.48 we computed the matrix elements assuming the ejected parton is a quark. For an anti-quark the fact that the negative frequency modes are needed is immaterial since the massless assumption still implies

$$\sum_{\text{spins}} v(p)\bar{v}(p) = \not{p} \tag{5.50}$$

however the reversal of momenta causes the α and β indices to swap. The net effect is to introduce an additional minus sign for the term proportional to $\epsilon^{\alpha\mu\beta\nu}$ (upon relabelling summation indices) while the remainder of the expression is symmetric under swapping $\alpha \leftrightarrow \beta$. This results in the structure function F_3 being proportional to $(f_q - f_{\bar{q}})$ while F_1 and F_2 will be proportional to $(f_q + f_{\bar{q}})$. This gives the overall hadronic tensor as [4]

$$W^{\mu\nu} = \frac{ie}{2xP\cdot q}\sum_q f_q(x)\frac{c_f Q_f}{\Lambda^2}\Big(2xP^{\mu}P^{\nu}-xP^2g^{\mu\nu}+P^{\mu}q^{\nu}-(P\cdot q)g^{\mu\nu}P^{\nu}q^{\mu}-i\lambda' P_{\alpha}q_{\beta}\epsilon^{\alpha\mu\beta\nu}\Big). \tag{5.51}$$

Decomposing the DIS cross section into its constituent structure functions, this time accounting for the parity violating structure function, F_3, reads:

$$\frac{d^2\sigma}{dxdy} = \frac{4\pi\alpha^2}{xyQ^2}\left(xy^2F_1(x,Q^2)+\left(1-y-\frac{x^2y^2M^2}{Q^2}\right)F_2(x,Q^2)+\left(y-\frac{y^2}{2}\right)F_3(x,Q^2)\right) \tag{5.52}$$

with $\alpha = e/4\pi$ the fine structure constant. We remind the reader this expression is obtained by contracting the leptonic tensor with the hadronic tensor:

$$\frac{d^2\sigma}{dxdy} = \frac{y}{8\pi}L_{\mu\nu}W^{\mu\nu}. \tag{5.53}$$

[4]The additional factor of two, originating from the fact that this will be a cross term in the overall squared amplitude, has been included here.

Performing the exercise in contracting 4-vectors and using the Ward identity $q^\mu L_{\mu\nu} = q^\nu L_{\mu\nu} = 0$, we obtain the photon-BSM induced modifications to the DIS structure functions read:

$$F_2 = F_2^{\rm SM} + Q^2 \sum_q \frac{e^2 Q_f c_f}{2\Lambda^2}(f_q(x) + f_{\bar{q}}(x)) \tag{5.54}$$

$$F_3 = F_3^{\rm SM} + Q^2 \sum_q \lambda\lambda' \frac{e^2 Q_f c_f}{2\Lambda^2}(f_q(x) - f_{\bar{q}}(x)) \tag{5.55}$$

$$F_1 = \frac{1}{2x} F_2. \tag{5.56}$$

The contribution from the other terms in equation 5.39 follow by a directly analogous computation the details of which we spare the reader of. The full correction, with the appropriate chirality choice of $\lambda = \lambda' = 1$ corresponding to equation 5.37, reads:

$$\begin{aligned} F_2(x,Q^2) =& F_2^{\rm SM}(x,Q^2) \\ &+ \frac{x}{12e^4}\Bigg[(3c_d^2\frac{Q^4}{\Lambda^4} - 2c_d e^2\frac{Q^2}{\Lambda^2}(1+4K_Z\sin^4\theta_W))(d(x,Q^2)+\bar{d}(x,Q^2)) \\ &\quad +(3c_u^2\frac{Q^4}{\Lambda^4} + 4c_u e^2\frac{Q^2}{\Lambda^2}(1+4K_Z\sin^4\theta_W))(u(x,Q^2)+\bar{u}(x,Q^2)) \\ &\quad +(3c_s^2\frac{Q^4}{\Lambda^4} - 2c_s e^2\frac{Q^2}{\Lambda^2}(1+4K_Z\sin^4\theta_W))(s(x,Q^2)+\bar{s}(x,Q^2)) \\ &\quad +(3c_c^2\frac{Q^4}{\Lambda^4} + 4c_c e^2\frac{Q^2}{\Lambda^2}(1+4K_Z\sin^4\theta_W))(c(x,Q^2)+\bar{c}(x,Q^2))\Bigg] \end{aligned} \tag{5.57}$$

$$\begin{aligned} F_3(x,Q^2) =& F_3^{\rm SM}(x,Q^2) \\ &+ \frac{1}{12e^4}\Bigg[(3c_d^2\frac{Q^4}{\Lambda^4} - 2c_d e^2\frac{Q^2}{\Lambda^2}(1+4K_Z\sin^4\theta_W))(d(x,Q^2)-\bar{d}(x,Q^2)) \\ &\quad +(3c_u^2\frac{Q^4}{\Lambda^4} + 4c_u e^2\frac{Q^2}{\Lambda^2}(1+4K_Z\sin^4\theta_W))(u(x,Q^2)-\bar{u}(x,Q^2)) \\ &\quad +(3c_s^2\frac{Q^4}{\Lambda^4} - 2c_s e^2\frac{Q^2}{\Lambda^2}(1+4K_Z\sin^4\theta_W))(s(x,Q^2)-\bar{s}(x,Q^2)) \\ &\quad +(3c_c^2\frac{Q^4}{\Lambda^4} + 4c_c e^2\frac{Q^2}{\Lambda^2}(1+4K_Z\sin^4\theta_W))(c(x,Q^2)-\bar{c}(x,Q^2))\Bigg] \end{aligned} \tag{5.58}$$

$$F_1(x,Q^2) = \frac{1}{2x}F_2(x,Q^2) \tag{5.59}$$

where θ_W is the weak mixing angle and for convenience we define

$$K_Z = \frac{Q^2}{4\cos^2\theta_W \sin^2\theta_W (Q^2 + M_Z^2)}. \tag{5.60}$$

with M_Z the Z-boson pole mass. We note that the Callan-Gross relation is preserved implying the longitudinal structure function retains its SM form. Recalling the discussion from section 2.3 we see that the term corresponding to $|\mathcal{A}_{\rm BSM}|^2$ is suppressed by $O(1/\Lambda^4)$ and is thus sub-leading relative to the $\mathcal{O}(1/\Lambda^2)$ terms. As such this study omits the quartic corrections, keeping only the EFT interference with the SM diagrams. Indeed, it was explicitly verified that the quadratic effects yielded virtually identical results to the case where they were omitted in the individual operator scenario.

Since the operators of equation 5.37 do not affect the QCD splitting functions, the QCD collinear structure is preserved. As such the DGLAP evolution equations retain their Standard Model form. Similarly, there is no running of the Wilson coefficients with scale since no loop calculations are done meaning no UV divergent momenta. We show in appendix C that even if QCD loop corrections were made, the particular Wilson coefficients for the operators of the present discussion have vanishing beta function at NLO. The structure function modifications of equation 5.57-5.59 require a trivial modification of the `APFEL` library [147] which generates the DIS theory predictions for the NNPDF codebase.

5.3.3 BSM absorption by PDFs

In this section we present the results of PDFs obtained by performing an NNPDF3.1 DIS only fit at various benchmark point choices. Note that since we perform a manual scan of the Wilson coefficient space, we are restricted to PDFs at representative choices of **c**. We reserve discussion of a first truly simultaneous PDF extraction for chapter 6. For lepton-proton scattering, the phenomenological quantity of interest are the PDFs themselves, while for hadron-hadron processes, the luminosity is of principal concern and we reserve presentation of these quantities for section 5.4

A representative point of $(c_u, c_d, c_s, c_c) = (-1.3, 1.3, 0.0, 0.0)$ is used to present the modifications induced on the PDFs in figure 5.5. A reminder that these Wilson coefficients are in units of $1/\Lambda^2$ where $\Lambda = 1$ TeV. These PDFs are obtained by performing a fit in the presence of the operators of equation 5.37 with the couplings set according to this benchmark point. With reference to table 5.3, this point corresponds to the PDF set with the largest deterioration in fit quality (as measured according

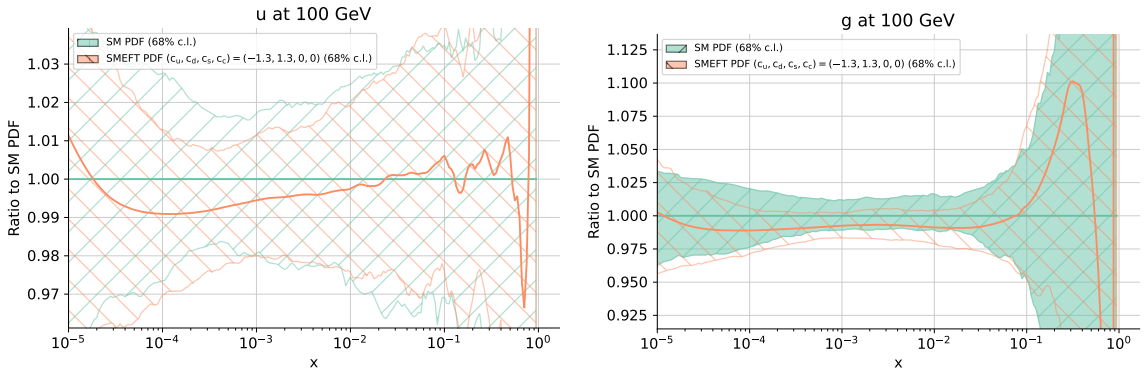


Figure 5.5: The SMEFT modified PDFs fitted in a BSM scenario corresponding to $(c_u, c_d, c_s, c_c) = (-1.3, 1.3, 0.0, 0.0)$ (orange) normalized to the corresponding SM PDFs (green) at $Q = 100$ GeV. The bands illustrate 68% confidence levels.

to the χ^2 increase to the data relative to the SM). We present representative parton flavours corresponding to the valence up quark and gluon distributions in figure 5.5. We see that the modification in the PDF central value is moderate, with the largest deflection observed in the gluon distribution approaching the 68% confidence level band of the baseline SM PDF. The largest deflection occurs in the Bjorken-x region corresponding to the area containing the highest concentration of data: the so-called *data region*. Correspondingly, this is where the PDF is best constrained. Despite the deflection in the gluon PDF, the effect remains moderate implying a mild absorption of the BSM effects by the PDFs. Indeed, as we show throughout this chapter and further on in this text, for currently available data, this will prove to be a common theme. It takes the inclusion of projected data for the High Luminosity upgrade of the LHC for a considerable effect to be observed.

Indeed the largest deflection is seen in the gluon, where the deflection approaches the 1σ uncertainty bands of the SM PDF fit. The reason for this is attributed to the fact that the gluon is most constrained by the Bjorken scaling violations of high Q data relative to the low Q measurements. This behaviour is strongly amplified by the SMEFT operators which have an energy growing effect linear in Q^2 at the $\mathcal{O}(1/\Lambda^2)$ level. On the other hand, the quark distributions are predominantly determined by the moderate Q datasets and thus are less affected by the high-Q measurements that are sensitive to the EFT operators. Indeed, making reference to figure 5.3, the large-x quark distributions are most strongly constrained by the BCDMS, SLAC, and CHORUS measurements which preside in a region of phase space corresponding to small values of Q^2 (< 300 GeV^2) where EFT effects are suppressed. Interestingly, the effect of the

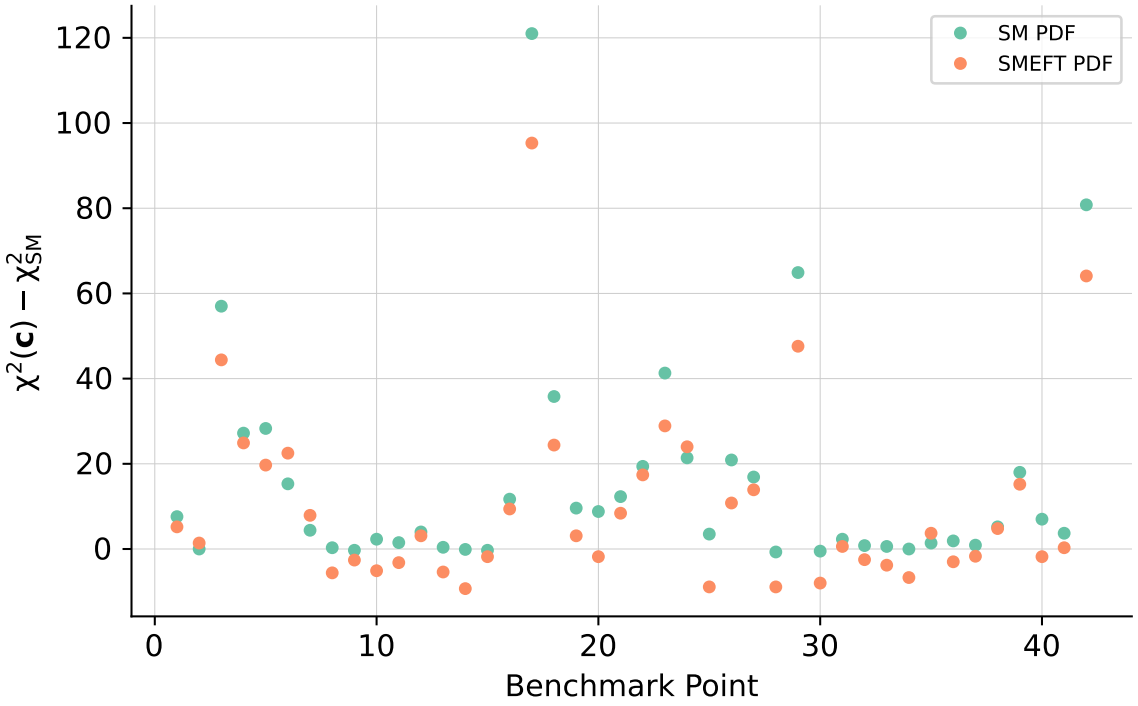


Figure 5.6: Distribution of χ^2 differences to the DIS data for each choice of benchmark point relative to the Standard Model χ^2: shown for both SM PDFs (green) and SMEFT PDFs (orange). The χ^2 is computed using the experimental covariance matrix of equation 3.18. The values of $\mathbf{c}$ for each benchmark point are tabulated in table 5.3.

EFT operators on the uncertainty bands is to leave them unchanged. This is to be expected since the number of PDF constraining measurements is the same, regardless of which benchmark point one uses. We will show, however, that in the scenario were the Wilson coefficients are treated on the same footing as the PDFs in chapter 6, then this property no longer holds. The discussion above is applicable to the vast majority of the benchmark points (barring small statistical fluctuations) used in this study and tabulated in table 5.3, but with varying levels of manifestation. Indeed, the PDFs of figure 5.5 are the most affected by the EFT operators, and hence are ideal for illustration purposes.

In figure 5.6, we plot the χ^2 difference to the data, relative to the SM χ^2, for each choice of benchmark point. A clear trend of a slight drop in the χ^2 in going from the SM PDFs to SMEFT PDFs is observed; implying the absorption of SMEFT effects into the PDF remains particularly mild at the energy scales probed by the DIS experiments.

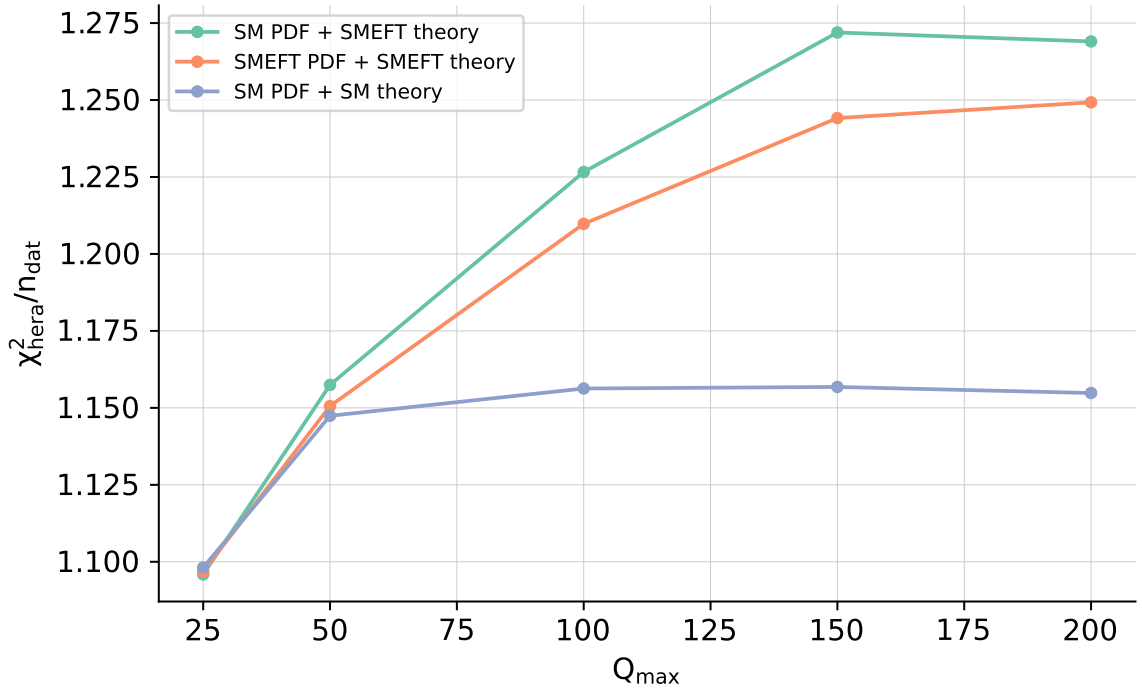


Figure 5.7: The χ^2 per data point to the HERA combined data as a function of the maximum Q cut, Q_{max}. The experimental covariance matrix of equation 3.18 is used in the computation. Shown are values for SM PDFs convolved with SMEFT partonic cross sections (green), SMEFT PDFs convolved with SMEFT partonic cross sections (orange), and SM PDFs convolved with SM partonic cross sections (blue). The benchmark point assumed here is $(c_u, c_d, c_c, c_s) = (-1.3, 1.3, 0.0, 0.0)$.

The scale dependence of the structure functions can be exploited to assess the degree of this reabsorption. Recall that in purely SM calculations, the Bjorken scaling is broken by terms logarithmic in Q^2. However, the analysis of section 5.3.2 reveals that the SMEFT operators introduce a scale dependence linear in Q^2 and hence energy growing effects faster than the purely SM case. In figure 5.7 we plot the HERA χ^2 per degree of freedom against the maximum Q that enters the evaluation of the χ^2. This is done for the SM PDF + SM theory case, but also for a representative choice of Wilson coefficient $c_u = -c_d = -1.3$ and $c_s = c_c = 0$ both for the case of fixed SM PDFs and SMEFT PDFs. While for $Q_{\text{max}} = 25$ GeV the χ^2 in all three cases is virtually identical (since the vast majority of SMEFT sensitive data has been filtered out by the Q_{max} cut) a rapid degradation in fit quality is observed as more and more EFT sensitive measurements are included, while the SM curve (in blue) quickly reaches saturation. Importantly, we note that the drop in the SM PDF + SMEFT theory (green) to the SMEFT PDF + SMEFT theory curve (orange) is much smaller than the distance to

the SM PDF + SM theory curve (blue). This important observation again highlights a slight reabsorption into the PDFs: in the case where a more marked absorption occurs, the improvement in fit quality would be closer to the SM PDF + SM theory curve. Again, this benchmark point is chosen for illustration purposes with virtually all other benchmark points behaving in the same way, as is evident from figure 5.6.

5.3.4 Bounds on Wilson coefficients using DIS data

We now turn our attention to discuss the second concern of this study. Namely, the effect one would obtain on the Wilson coefficient bounds by considering a fully consistent treatment whereby the PDFs are allowed to vary in EFT parameter space (by producing SMEFT PDFs for each benchmark point); versus a more rudimentary analysis of keeping the PDFs fixed as SM PDFs. Moreover, the effect of fitting multiple EFT operators in the presence of one another is known to have a considerable impact on the bounds as compared to when they are fitted individually owing to the correlations between operators [43, 219]. This effect too will be assessed in light of varying PDFs alongside the Wilson coefficients.

We begin by presenting 90% CL bounds assuming SM PDFs in table 5.1. Included are bounds for individual Wilson coefficients as well as marginalized bounds for the full global analysis incorporating all 4 SMEFT operators, where for the latter we include the PDF uncertainty bounds as explained in section 5.2.2. We find that the most stringent bounds are obtained for c_u followed by c_d and the widest bounds for c_c and c_s. This can be understood through the observation that the corresponding quark distributions are constrained in the same order by DIS data. Indeed, this is attributed to the Wilson coefficients being multiplied by the corresponding PDF flavour in equations 5.57 and 5.58. Moreover, the bounds broaden significantly upon considering the full 4-dimensional Wilson coefficient space as opposed to the 1-dimensional individual case. Indeed, this property is ubiquitous and highlights the importance of considering as global a setting as possible: the more operators considered the better. Note however, that in the setting of a DIS only study, bounds on, for example, top quark or Higgs sector operators would not be possible and can safely be discarded; however, it does highlight the pertinent issue that all relevant operators to a set of measurements should be considered. The generic problem associated with the addition of further operators, however, is the existence of flat directions in the Wilson coefficient space. This makes it very difficult for a Hessian approach to be viable and we will address this issue in chapter 6 by presenting a more sophisticated methodological improvement.

Table 5.1: The 90% CL bounds obtained on the Wilson coefficients using SM PDFs and the approach outlined in section 5.2. The first column of bounds corresponds to the scenario where Wilson coefficients are fitted individually and so the χ^2 profiles are parabolas. The second set of columns are the bounds in the fully simultaneous case. The third set includes PDF uncertainty for the simultaneous case using the method outlined in section 5.2.2.

Wilson Coefficient	Individual	Marginalised	
		no PDF unc	PDF unc
c_u	$[-0.1, +0.4]$	$[-2.4, +1.4]$	$[-3.6, +2.7]$
c_d	$[-1.6, +0.4]$	$[-13, +3.9]$	$[-19, +11]$
c_s	$[-2.8, +4.2]$	$[-18, +29]$	$[-36, +47]$
c_c	$[-2.6, +1.2]$	$[-13, +7.0]$	$[-21, +15]$

In table 5.2 we present bounds, but this time allow for variations of the PDFs as we explore the Wilson coefficient space. Indeed, much of the results stay the same. We see the same hierarchy of bounds and a significant broadening going from the individual scenario to the marginalized 4-dimensional case. We see the bounds are

Table 5.2: Same as table 5.1, but using SMEFT PDFs instead. Bounds with PDF uncertainty are included by construction and so the "no PDF uncertainty" column is omitted.

Wilson Coefficient	Individual	Marginalised
c_u	$[0.0, +0.5]$	$[-0.4, +2.4]$
c_d	$[-1.1, +0.8]$	$[-4.4, +4.5]$
c_s	$[-4.5, +3.6]$	$[-61, +39]$
c_c	$[-2.4, +0.7]$	$[-29, +2.7]$

roughly comparable with the fixed SM PDF case. The slight variation in the bounds is attributed to the discussion in the previous section implying that the SMEFT effects are only partially absorbed by the PDFs. The numerical differences, however, could potentially lead to misleading conclusions in the context of a BSM study and are suggestive of the need for a consistent framework to disentangle the interplay of BSM dynamics and the PDFs.

5.3.5 Fit quality

In table 5.3 we present the total χ^2 values for the various benchmark points used in this study. The computation of the χ^2 is made using the experimental covariance matrix of equation 3.18. We show the χ^2 to the global DIS dataset as well as to the

HERA combined dataset only; for both SM PDFs and SMEFT PDFs. In principle, this information is sufficient to reproduce the Hessian matrix of section 5.2 and to reproduce the bounds of tables 5.1 and 5.2. The total number of data points amount to 3092 for the global DIS dataset, while the number of data points for the HERA inclusive structure function data (both neutral current and charged current, but excluding heavy quark structure functions) is 1145.

The vast majority of benchmark points enjoy an improved fit quality, with a handful of data points having the opposite behaviour. However, these violations are particularly small and most likely a result of the number of replicas used in the PDF ensemble ($N_{\text{rep}} = 300$).

5.4 Parton distributions and the SMEFT from high-mass Drell-Yan tails

The lepton-proton scattering analysis presented above serves as a convenient framework to provide a proof-of-concept study for the interplay between PDFs and BSM dynamics. We find that for the kinematic coverage of the DIS data and the high energy reach of the HERA experiments, a mild reabsorption of the BSM dynamics occurs by the neural network parameterization of PDFs. However, with reference to figure 3.7, the span of the DIS data is restricted to a low to medium energy region of kinematical phase space. With the advent of the latest analyses from the LHC, a kinematic coverage is achieved surpassing the vast majority of DIS measurements. In addition, for the DIS study, the number of EFT sensitive data points is few in comparison to those which are not affected by the EFT operators. This results in the PDFs being highly restricted, since the effects of the EFT are suppressed relative to fitting the PDFs to the low energy points. We expect this narrative to change if we equalled the number of EFT sensitive data points with those that reside at lower Q. One is naturally led to consider the impact of the above analysis applied to more exotic measurements from proton-proton collider data, especially those stemming from high-mass Drell-Yan measurements from the LHC. In this section we extend the discussion of the DIS only study to further include hadronic measurements from some of the highest energy processes we have available. Moreover, we select SMEFT operators with a similarly fast energy growing effect as those of equation 5.37, thus allowing for a strong control

Table 5.3: The χ^2 figure of merit calculated using the experimental covariance matrix. Shown are values for the full DIS dataset as well as to only the HERA combined dataset. We use both SM PDFs as well as SMEFT PDFs and for convenience show the difference for each BP to the corresponding SM χ^2. We also specify the values of benchmark points used in this study.

BP	c_u	c_d	c_s	c_c	SM PDF			SMEFT PDF		
					χ^2_{tot}	χ^2_{HERA}	$\Delta\chi^2$	χ^2_{tot}	χ^2_{HERA}	$\Delta\chi^2$
SM	0	0	0	0	3445.8	1311.8	-	3445.8	1311.8	-
BP1	-0.28	0.1	0.1	-0.28	3453.4	1319.4	7.6	3451.0	1314.6	5.2
BP2	-0.04	-0.19	-0.19	-0.04	3445.8	1311.7	0.0	3447.2	1312.4	1.4
BP3	-1.0	0.7	-0.7	1.0	3502.8	1368.9	57.0	3490.2	1354.9	44.4
BP4	-0.7	0.5	0.0	3.0	3473.0	1338.7	27.2	3470.7	1331.1	24.9
BP5	1.0	0.0	0.0	0.0	3474.1	1339.9	28.3	3465.5	1341.7	19.7
BP6	-0.5	0.0	0.0	0.0	3461.1	1327.1	15.3	3468.3	1324.1	22.5
BP7	0.5	0.0	0.0	0.0	3450.2	1316.1	4.4	3453.7	1316.9	7.9
BP8	0.3	0.0	0.0	0.0	3446.1	1312.0	0.3	3440.2	1313.7	-5.6
BP9	0.0	-1.0	0.0	0.0	3445.5	1311.4	-0.3	3443.2	1312.5	-2.6
BP10	0.0	0.5	0.0	0.0	3448.1	1314.1	2.3	3440.7	1315.4	-5.1
BP11	0.0	-1.5	0.0	0.0	3447.3	1313.2	1.5	3442.5	1318.9	-3.2
BP12	0.0	-1.9	0.0	0.0	3449.8	1315.6	4.0	3448.9	1317.4	3.1
BP13	0.0	0.0	-0.7	0.0	3446.2	1312.1	0.4	3440.4	1312.5	-5.4
BP14	0.0	0.0	0.0	-0.2	3445.7	1311.6	-0.1	3436.5	1314.2	-9.3
BP15	0.0	0.0	0.0	-1.0	3445.5	1311.6	-0.3	3444.0	1311.3	-1.8
BP16	0.9	0.9	0.0	0.0	3457.5	1323.4	11.7	3455.2	1325.5	9.4
BP17	-1.3	1.3	0.0	0.0	3566.8	1433.0	121.0	3541.1	1405.7	95.3
BP18	0.0	0.0	5.0	-5.0	3481.6	1347.8	35.8	3470.1	1337.6	24.4
BP19	0.0	0.0	-2.0	2.0	3455.4	1321.2	9.6	3448.9	1323.4	3.1
BP20	0.3	0.0	10.0	0.0	3454.6	1320.6	8.8	3451.4	1321.6	-1.8
BP21	0.3	0.0	-5.0	0.0	3458.1	1324.0	12.3	3454.2	1327.0	8.4
BP22	-0.3	0.0	0.0	5.0	3465.2	1330.6	19.4	3463.2	1326.7	17.4
BP23	0.0	1.2	10.0	0.0	3487.1	1353.2	41.3	3474.7	1343.1	28.9
BP24	0.0	-1.8	-5.0	0.0	3467.2	1332.9	21.4	3469.8	1336.3	24.0
BP25	0.3	1.2	-5.0	-5.0	3449.3	1315.4	3.5	3436.9	1311.3	-8.9
BP26	0.3	-1.8	10.0	5.0	3466.7	1332.0	20.9	3456.6	1327.1	10.8
BP27	0.3	-1.8	10.0	-5.0	3462.7	1328.8	16.9	3459.7	1324.5	13.9
BP28	0.3	-1.8	-5.0	-5.0	3445.1	1311.0	-0.7	3437.1	1310.9	-8.7
BP29	-0.3	-1.8	-5.0	5.0	3510.7	1375.9	64.9	3493.4	1365.7	47.6
BP30	0.0	-0.2	0.0	0.0	3445.3	1311.2	-0.5	3437.8	1308.3	-8.0
BP31	0.0	0.0	-2.6	0.0	3448.1	1314.0	2.3	3446.4	1313.2	0.6
BP32	0.0	0.0	2.6	0.0	3446.6	1312.5	0.8	3443.3	1307.8	-2.5
BP33	0.0	0.0	-1.0	0.0	3446.4	1312.3	0.6	3442.0	1313.9	-3.8
BP34	0.0	0.0	1.0	0.0	3445.8	1311.7	0.0	3439.1	1312.0	-6.7
BP35	0.0	0.0	0.0	0.8	3447.2	1313.1	1.4	3449.5	1314.3	3.7
BP36	0.0	0.0	0.0	1.0	3447.7	1313.6	1.9	3442.8	1315.7	-3.0
BP37	0.0	0.0	0.0	-2.0	3446.7	1312.8	0.9	3444.1	1312.2	-1.7
BP38	0.0	0.0	0.0	2.0	3451.0	1316.8	5.2	3450.6	1317.0	4.8
BP39	0.0	0.0	10.0	0.0	3464.5	1330.6	18.7	3461.0	1327.4	15.2
BP40	0.0	0.0	-5.0	0.0	3452.8	1318.7	7.0	3444.0	1317.4	-1.8
BP41	0.0	0.0	0.0	-3.0	3449.5	1315.6	3.7	3446.1	1314.9	0.3
BP42	0.0	0.0	20.0	0.0	3526.6	1392.7	80.8	3509.9	1368.9	64.1

on the Wilson coefficient bounds as well as the PDFs. We remind the reader that both analyses performed in this chapter are done with the NNPDF3.1 machinery.

5.4.1 BSM sensitive Drell-Yan data

The dataset used in this study is identical to the DIS data selection of section 5.3.1 supplemented by a number of hadron collider processes. We extend the lepton-proton measurements with Drell-Yan measurements making for 4022 post cut data points, spanning a broad range of processes including neutral and charged current DIS data to high mass Drell-Yan measurements from ATLAS and CMS at the LHC. The full dataset selection is an extension of the strangeness study presented in chapter 4 and NNPDF3.1 [137].

The Drell-Yan measurements can be categorized into low-mass, on shell, and high-mass measurements: referring to the produced dilepton invariant mass, $m_{\ell\ell}$. In table 5.4 we summarize the Drell-Yan measurements corresponding to the first two categories. These include measurements from the Tevatron at Fermilab as well as modern LHC measurements at CERN. These amount to 609 total measurements, and we tabulate the center-of-mass-energy as well as the measured observable.

Accompanying these low-invariant mass measurements are the neutral current Drell-Yan measurements from ATLAS at 7, and 8 TeV [233, 234] and CMS at 7, 8, and 13 TeV [159, 235, 236]. These are high-mass Drell-Yan measurements from the LHC that are used to constrain, not only the PDF, but are also sufficiently sensitive to the BSM scenario considered in this work to also constrain the EFT operators.

We tabulate these experiments in table 5.5, highlighting the total integrated luminosity, final states, whether the distribution is 1 or 2 dimensional (that is: differential in the dilepton invariant mass or dilepton invariant mass and rapidity), the number of data points, and the bin-edges for the highest energy binnings. Some of these datasets are available in terms of both Born and dressed leptons, such as ATLAS at 7 TeV, for which we use the Born data to avoid the need for electroweak corrections corresponding to final-state QED radiation of the leptons. However, the CMS data at 13 TeV is only available for dressed lepton measurements and so we do include the corresponding final state radiation corrections. The CMS data at 13 TeV is the only high-mass set that is provided in terms of individual final state products (either dielectron or dimuon) or the combined channel. The benefit of this is that it allows one to assess the effect of new physics scenarios that couple differently across the charged-lepton flavours (such as that outlined in section 5.4.3), while flavour universal models (such as in section 5.4.2)

Table 5.4: The low-mass and on-shell Drell-Yan datasets used in the present study. For each dataset we indicate the experiment, the centre-of-mass energy $\sqrt{s}$, the publication reference, the physical observable, and the number of data points.

Exp.	$\sqrt{s}$/TeV	Ref.	Observable	$n_{\rm dat}$
E886	0.8	[220]	$d\sigma^d_{\rm DY}/d\sigma^p_{\rm DY}$	15
E886	0.8	[221, 222]	$d\sigma^p_{\rm DY}/(dy\, dm_{\ell\ell})$	89
E605	0.04	[154]	$\sigma^p_{\rm DY}/(dx_F\, dm_{\ell\ell})$	85
CDF	1.96	[178]	$d\sigma_Z/dy_Z$	29
D0	1.96	[179]	$d\sigma_Z/dy_Z$	28
D0	1.96	[223]	$d\sigma_{W\to\mu\nu}/d\eta_\mu$ asy.	9
ATLAS	7	[180]	$d\sigma_W/d\eta_l, d\sigma_Z/dy_z$	30
ATLAS	7	[224]	$d\sigma_{Z\to e^+e^-}/dm_{e^+e^-}$	6
ATLAS	7	[173]	$d\sigma_W/d\eta_l, d\sigma_Z/dy_z$	61
ATLAS	7	[225]	$d\sigma_{W+c}/dy_c$	22
ATLAS	8	[226]	$d\sigma_Z/dp_T$	82
ATLAS	8	[186]	$d\sigma_{W+j}/dp_T$	32
CMS	7	[227]	$d\sigma_{W\to l\nu}/d\eta_\ell$ asy.	22
CMS	7	[184]	$d\sigma_{W+c}/dy_c$	5
CMS	7	[184]	$d\sigma_{W^++c}/d\sigma_{W^-+c}$	5
CMS	8	[228]	$d\sigma_Z/dp_T$	28
CMS	8	[181]	$d\sigma_{W\to\mu\nu}/d\eta_\mu$	22
CMS	13	[185]	$d\sigma_{W+c}/dy_c$	5
LHCb	7	[229]	$d\sigma_{Z\to\mu^+\mu^-}/dy_{\mu^+\mu^-}$	9
LHCb	7	[230]	$d\sigma_{W,Z}/d\eta$	29
LHCb	8	[231]	$d\sigma_{Z\to e^+e^-}/dy_{e^+e^-}$	17
LHCb	8	[232]	$d\sigma_{W,Z}/d\eta$	30
Total				**659**

will enjoy the reduced systematic uncertainties of the combined channel. In all, an additional 270 data points are introduced by the high-mass data, or 313 if we consider the individual CMS 13 TeV channels. The kinematic coverage of all the data points used in this study are shown in figure 5.8. The points are shown in (x, Q^2) space with the data points that are modified by the EFT operators highlighted with a border, such points thus also constrain the Wilson coefficients as well as the PDFs.

Table 5.5: Same as table 5.4 for the neutral-current high-mass Drell-Yan datasets considered in this work. We also indicate the final-state, whether the distribution is 1D (which are differential in the invariant mass, $m_{\ell\ell}$, of the final-state leptons) or 2D (which are differential in both the invariant mass of the leptons, $m_{\ell\ell}$, and in their rapidity, $y_{\ell\ell}$), and the values of the $m_{\ell\ell}$ bin edges for the most energetic bin.

Exp.	$\sqrt{s}$/TeV	Ref.	$\mathcal{L}$ (fb^{-1})	Channel	1D/2D	$n_{\rm dat}$	$m_{\ell\ell}^{\rm max}$/TeV
ATLAS	7	[233]	4.9	e^-e^+	1D	13	[1.0, 1.5]
ATLAS	8	[234]	20.3	$\ell^-\ell^+$	2D	46	[0.5, 1.5]
CMS	7	[159]	9.3	$\mu^-\mu^+$	2D	127	[0.2, 1.5]
CMS	8	[235]	19.7	$\ell^-\ell^+$	1D	41	[1.5, 2.0]
CMS	13	[236]	5.1	e^-e^+, $\mu^-\mu^+$ $\ell^-\ell^+$	1D	43, 43 43	[1.5, 3.0]
Total						**270 (313)**	

5.4.2 SMEFT scenario I: the oblique corrections

An important class of BSM models are the *universal* new physics models [237–239]. These BSM scenarios can include new heavy vector bosons which mix with the SM gauge bosons [240–243] or new charged degrees of freedom [244]. Such models manifest in the low energy by modifications to the electroweak gauge boson vacuum polarization tensor, $\Pi_{VV'}(q^2)$, [245] and thus modify the boson self-energy. The effects of the new physics can be captured in the IR using the *oblique parameters*. One may expand the gauge boson self-energies in powers of q^2 (the exchanged 4-momentum) and truncating at order q^4, while imposing normalization and symmetry constraints, implies the need for only 4 parameters: $\hat{S}$, $\hat{T}$, $\hat{W}$, and $\hat{Y}$ [246–248] [5]. The parameters $\hat{S}$ and $\hat{T}$ grow as $\mathcal{O}(q^0)$ and $\mathcal{O}(q^2)$ respectively and are heavily constrained by precision LEP measurements [237]. However, $\hat{W}$ and $\hat{Y}$ both grow as $\mathcal{O}(q^4)$ and so form an ideal operator selection for use in conjunction with the LHC high-mass Drell-Yan data. We shall leverage this energy growing effect to examine the possible modifications to the PDFs; an effect that up until now has been ignored in the literature, though could possibly affect a number of various BSM interpretations [249]. Though in principle the same sentiment applies to the $\hat{S}$ and $\hat{T}$ parameters, we expect from the onset the

[5] We place hats on the oblique parameters so as to not confuse them with the gauge boson W or hypercharge Y.

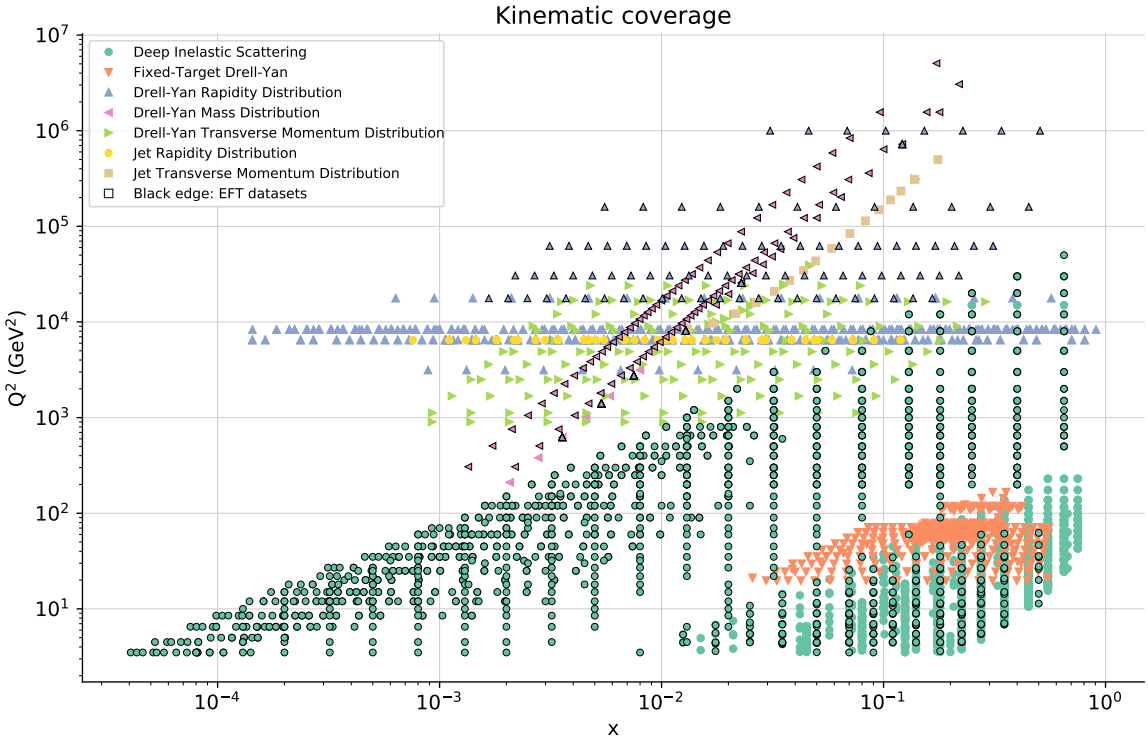


Figure 5.8: The kinematic coverage of the data points used in this study grouped according to their process. Points marked with a black edge are also used to constrain the EFT operators.

effect to be mild. Thus, since our primary objective is to assess the impact the BSM and PDF interplay has on BSM bounds, we shall choose to omit these parameters. In principle a fully global analysis should account for these parameters, perhaps using the methodology presented in chapter 6, we here restrict the effect to those operators most sensitive to the interplay.

The inclusion of the $\hat{W}$ and $\hat{Y}$ parameters necessitates the addition of the following dimension-6 operators to the Standard Model lagrangian

$$\mathcal{L} = \mathcal{L}_{\text{SM}} - \frac{\hat{W}}{4m_W^2}\left(D_\rho W^a_{\mu\nu}\right)^2 - \frac{\hat{Y}}{4m_W^2}\left(\partial_\rho B_{\mu\nu}\right)^2 \tag{5.61}$$

where D_ρ is the usual covariant derivative, m_W is the W-boson mass, and $W^a_{\mu\nu}$ ($B_{\mu\nu}$) is the field strength tensor associated with the unbroken $SU(2)_L$ ($U(1)_Y$) gauge symmetry. The index a enumerates the $\mathfrak{su}(2)$ Lie algebra generators.

These operators can be decomposed into the Warsaw basis via the equations of motion as [5, 238]:

$$\mathcal{L} = \mathcal{L}_{\rm SM} - \frac{g^2\hat{W}}{4m_W^2}\mathcal{O}_{lq}^{(3)} - \frac{g_Y^2\hat{Y}}{m_W^2}\Bigg(Y_lY_d\mathcal{O}_{ld} + Y_lY_u\mathcal{O}_{lu} + Y_lY_q\mathcal{O}_{lq}^{(1)} + Y_eY_d\mathcal{O}_{ed} + Y_eY_u\mathcal{O}_{eu} + Y_eY_q\mathcal{O}_{qe}\Bigg) \quad (5.62)$$

Where g and g_Y are the corresponding electroweak gauge couplings and the hypercharges are:

$$\begin{pmatrix} Y_q \\ Y_l \\ Y_u \\ Y_d \\ Y_e \end{pmatrix} = \begin{pmatrix} \frac{1}{6} \\ -\frac{1}{2} \\ \frac{2}{3} \\ -\frac{1}{3} \\ -1 \end{pmatrix}. \quad (5.63)$$

The various operators are defined as

$$\begin{gathered} \mathcal{O}_{ld} = (\bar{l}\gamma_\mu l)(\bar{d}\gamma^\mu d), \qquad \mathcal{O}_{lu} = (\bar{l}\gamma_\mu l)(\bar{u}\gamma^\mu u), \qquad \mathcal{O}_{lq}^{(1)} = (\bar{l}\gamma_\mu l)(\bar{q}\gamma^\mu q), \\ \mathcal{O}_{ed} = (\bar{e}\gamma_\mu e)(\bar{d}\gamma^\mu d), \qquad \mathcal{O}_{eu} = (\bar{e}\gamma_\mu e)(\bar{u}\gamma^\mu u), \qquad \mathcal{O}_{qe} = (\bar{q}\gamma_\mu q)(\bar{e}\gamma^\mu e), \\ \mathcal{O}_{lq}^{(3)} = (\bar{l}\sigma^a\gamma_\mu l)(\bar{q}\sigma^a\gamma^\mu q). \end{gathered} \quad (5.64)$$

Here l and q represent the lepton and quark $SU(2)_L$ left-handed doublets respectively, while u, d and e are respectively the up, down, and electron right-handed singlets. The Pauli matrices, σ^a, act on $SU(2)$ space, while the gamma matrices as ever act on the Lorentz spin structure. There is an implicit summation on the left-handed generations; that is we have

$$(\bar{l}\gamma_\mu l) \equiv \bar{l}^1\gamma_\mu l^1 + \bar{l}^2\gamma_\mu l^2 + \bar{l}^3\gamma_\mu l^3. \quad (5.65)$$

The modifications induced by the $\hat{W}$ and $\hat{Y}$ operators have been implemented and cross checked with the `SMEFTsim` package [250]. As in the case of the DIS only study, we restrict the analysis for Scenario I to $\mathcal{O}(1/\Lambda^2)$. These operators have gained much attention in the context of high-energy LHC data and the following bounds have been reported for individual $\hat{W}$ and $\hat{Y}$ operators (that is: assuming the presence of one

operator at a time) at 95% confidence level [249]:

$$\begin{aligned}
&\hat{W} \in [-3, 15] \times 10^{-4} \qquad \text{(ATLAS 8 TeV, 20.3 fb}^{-1}\text{ [234])} \\
&\hat{W} \in [-5, 22] \times 10^{-4} \qquad \text{(CMS 8 TeV, 19.7 fb}^{-1}\text{ [235])} \\
&\hat{Y} \in [-4, 24] \times 10^{-4} \qquad \text{(ATLAS 8 TeV, 20.3 fb}^{-1}\text{ [234])} \\
&\hat{Y} \in [-7, 41] \times 10^{-4} \qquad \text{(CMS 8 TeV, 19.7 fb}^{-1}\text{ [235])}
\end{aligned} \tag{5.66}$$

which have been obtained assuming SM PDFs. In this study we will benchmark these results and then compare them to the consistent treatment of using SMEFT modified PDFs. We note here that a known flat direction exists if one attempts to simultaneously extract $\hat{W}$ and $\hat{Y}$ operators together using only neutral current Drell-Yan data [249] which we will demonstrate further in chapter 6. As such, we will perform the simultaneous fit with the HL-LHC data which includes charged current processes in section 5.5.

5.4.3 SMEFT scenario II: left-handed muon-philic lepton-quark interactions

The second benchmark scenario we choose to consider follows that of [45] and is sensitive to flavour physics. We consider the gauge invariant four-fermion operators built from the quark and lepton $SU(2)_L$ doublets. From equation 5.4.2, these are $\mathcal{O}_{lq}^{(1)}$ and $\mathcal{O}_{lq}^{(3)}$. Performing the enumeration over $SU(2)_L$ indices, we restrict to operators of the form:

$$\mathcal{L} = \mathcal{L}_{\text{SM}} + \frac{\mathbf{C}_{ij}^{U\mu}}{v^2}(\bar{u}_L^i \gamma_\mu u_L^j)(\bar{\mu}_L \gamma^\mu \mu_L) + \frac{\mathbf{C}_{ij}^{D\mu}}{v^2}(\bar{d}_L^i \gamma_\mu d_L^j)(\bar{\mu}_L \gamma^\mu \mu_L) \tag{5.67}$$

where $v \simeq 246$ GeV is the Higgs vacuum expectation value and $\mathbf{C}_{ij}^{U\mu}$ and $\mathbf{C}_{ij}^{D\mu}$ are matrices of Wilson coefficients. The indices $i, j = 1, 2, 3$ run over quark flavours; however, we choose to consider those operators that couple solely to the second lepton family. Here, μ_L is the left-handed muon Dirac spinor field, while u_L^i (d_L^i) is the spinor field corresponding to left-handed up-type (down-type) quarks of flavour i. The operators of equation 5.67 form an attractive choice of BSM scenario due to the lepton flavour universality violating LHCb anomalies reported in rare B-meson decays [24–26]. The operator structure of equation 5.67 are reminiscent of the flavour-changing CKM structure of charged current weak decays and thus affect processes such as $b \to s\mu^+\mu^-$ within $pp \to \mu^+\mu^-$ charged current Drell-Yan scattering at the LHC. Modifications

to this channel would be responsible for the discrepancies between the SM prediction for $R(K^{(*)})$ and the observed experimental values. Models successfully explaining the B-anomalies [251] suggest the dominant channel in the EFT is $b\bar{b} \to \mu^+\mu^-$. While the direct production channel stemming from an operator such as $(\bar{b}_L\gamma_\mu s_L)(\bar{\mu}_L\gamma^\mu\mu_L)$ does exist; it is Cabibbo suppressed by V_{ts} after rotating from the weak eigenbasis to the mass eigenbasis in the Yukawa sector. As such the only Wilson coefficient of interest is $\mathbf{C}_{33}^{D\mu}$ with the corresponding lagrangian for BSM scenario II being

$$\mathcal{L} = \mathcal{L}_{\text{SM}} + \frac{\mathbf{C}_{33}^{D\mu}}{v^2}(\bar{b}_L\gamma_\mu b_L)(\bar{\mu}_L\gamma^\mu\mu_L) \tag{5.68}$$

where b_L the left-handed bottom quark field. Note that the electron channel is still governed by the Standard Model. This property provides a useful handle to constrain both the PDFs and the Wilson coefficients. Using dielectron production data, we can constrain heavily the PDF while the dimuon measurements constrain both. Moreover, this particular Wilson coefficient poses a complication which we have up until now been neglecting. Previously, we deemed the SMEFT-SMEFT interference terms in the squared amplitude to be subleading relative to the SM-SMEFT cross term diagrams. We argued this occurs due to a suppression of $\mathcal{O}(1/\Lambda^4)$ in the former and $\mathcal{O}(1/\Lambda^2)$ in the latter. However, it is known that at linear level in the EFT, this particular Wilson coefficient is virtually unconstrained [45] and so the $\mathcal{O}(1/\Lambda^4)$ cannot be neglected as they are required to provide the necessary constraining power. As such, instead of performing a 1-dimensional quadratic fit as in equation 5.5, we will use the full quartic polynomial. Note that this operator is of the form considered in equation 5.37 and so affects the heavy quark structure function measurements from HERA. We will, however, neglect this modification owing to the fact that the bottom quark distribution is very small in a broad kinematic range and that DIS measurements probe low energies. As a benchmark, the ATLAS search data of [252] was used in [45] to obtain

$$\mathbf{C}_{33}^{D\mu} \in [-2.6, 2.1] \times 10^{-2} \qquad \text{(ATLAS 13 TeV, 36.1 fb}^{-1}\text{ [252])}. \tag{5.69}$$

5.4.4 Theory modifications

In the DIS study, we showed how the operators of equation 5.37 modify the DIS structure functions by the calculation outlined in section 5.3.2. By the exact same token, the operators introduced in the various BSM scenarios above will also modify

the corresponding theory predictions for the various measurements in this study too, which crucially include proton-proton scattering processes.

The corrections from these SMEFT operators are introduced using the K-factor approximation. Consider a general SMEFT lagrangian expansion at dimension-6 as:

$$\mathcal{L} = \mathcal{L}_{\rm SM} + \sum_{n=1}^{N_{\rm op}} \frac{c_n}{v^2} \mathcal{O}_n \tag{5.70}$$

where we allow for $N_{\rm op}$ dimension-6 operators, $\mathcal{O}_n$, enumerated by the index n. Here v is some new physics scale introduced to make the Wilson coefficients, c_n, dimensionless. The linear effect of these operators at the cross section level can be expressed as:

$$R^{(n)}_{\rm SMEFT} \equiv \left(\mathcal{L}^{\rm NNLO}_{ij} \otimes d\widehat{\sigma}^{(n)}_{ij,\rm SMEFT}\right) \Big/ \left(\mathcal{L}^{\rm NNLO}_{ij} \otimes d\widehat{\sigma}_{ij,\rm SM}\right) , \quad n = 1 \ldots, N_{\rm op} , \tag{5.71}$$

with $\mathcal{L}^{\rm NNLO}_{ij}$ being the usual partonic luminosity evaluated at NNLO QCD relating partons of flavour i and j:

$$\mathcal{L}_{ij}(x,m) = \int_x^1 \frac{dy}{y}\, f_i(y,m) f_j\left(\frac{x}{y}, m\right) , \tag{5.72}$$

$d\widehat{\sigma}_{ij,\rm SM}$ the bin-by-bin partonic SM cross section, and $d\widehat{\sigma}^{(n)}_{ij,\rm SMEFT}$ the corresponding partonic cross section associated to the interference between $\mathcal{O}_n$ and the SM amplitude $\mathcal{A}_{\rm SM}$ when setting $c_n = 1$. The notation $\otimes$ is as ever the Mellin convolution. Likewise, the ratio encapsulating the quadratic effects is defined as

$$R^{(n,m)}_{\rm SMEFT} \equiv \left(\mathcal{L}^{\rm NNLO}_{ij} \otimes d\widehat{\sigma}^{(n,m)}_{ij,\rm SMEFT}\right) \Big/ \left(\mathcal{L}^{\rm NNLO}_{ij} \otimes d\widehat{\sigma}_{ij,\rm SM}\right) , \quad n,m = 1 \ldots, N_{\rm op} , \tag{5.73}$$

with the bin-by-bin partonic cross section $d\widehat{\sigma}^{(n,m)}_{ij,\rm SMEFT}$ now being evaluated from the squared amplitude $\mathcal{A}_n\mathcal{A}_m$ associated to the operators $\mathcal{O}_n$ and $\mathcal{O}_m$ when $c_n = c_m = 1$. The SMEFT partonic cross sections are computed in this study at leading order in QCD. The overall K-factor which maps a SM prediction to one which includes SMEFT modifications is then:

$$K_{\rm EFT} = 1 + \sum_{n=1}^{N_{\rm op}} c_n R^{(n)}_{\rm SMEFT} + \sum_{n,m} c_n c_m R^{(n,m)}_{\rm SMEFT} \tag{5.74}$$

with the mapping being related by a simple multiplicative factor:

$$d\sigma_{\rm SMEFT} = d\sigma_{\rm SM} \times K_{\rm EFT} \tag{5.75}$$

where the $d\sigma_{\rm SM}$ is the state-of-the-art SM prediction including NNLO QCD and NLO EW corrections. Clearly, the K-factor is a PDF dependent object, since it depends on the PDF luminosity. In [5] we show that the dependence on varying the input PDF is at the permil level for Scenario I while for Scenario II the effect is slightly more pronounced, but still only at the few percent level. As such this effect will be ignored here henceforth, while in chapter 6 we introduce an approach which eliminates the need for K-factors altogether, instead building on the FK-table approach of section 3.2.3.

5.4.5 Constraints on oblique parameters from high-mass Drell-Yan measurements: Scenario I

We deploy the methodology outlined in sections 5.1 and 5.2 to perform a 1-dimensional fit of $\hat{W}$ and $\hat{Y}$ where only one parameter is allowed to be non-zero at a time. The reason for this is the flat direction that exists for neutral-curent Drell-Yan measurements which we shall lift in section 5.5 with the use of charged-current Drell-Yan from the High-Luminosity LHC projections. We perform an exploration of Wilson coefficient space by again constructing a set of benchmark points for each operator. For each benchmark point we obtain PDF fits of ensemble size equal to $N_{\rm rep} = 100$ MC replicas. In the case of $\hat{Y}$ we use 21 sampling values equally spaced in the closed interval $\hat{Y} = [-20, 20] \times 10^{-4}$ while for $\hat{W}$ we found it convenient to use 15 equally spaced points in the closed interval $\hat{W} = [-14, 14] \times 10^{-4}$ supplemented by two additional points at $\hat{W} = -18 \times 10^{-4}$ and -22×10^{-4}.

In figure 5.9 we display the parabolic fits corresponding to equation 5.5 for the case of SMEFT PDFs. The error bars are obtained using the bootstrap methodology outlined in section 5.2.3 and correspond to the methodology uncertainties. The horizontal line corresponds to $\Delta\chi^2 = 4$. The points corresponding to the intersection of the horizontal line with the parabolas corresponds to a 2σ or 95% CL. We see that for both scenarios the SM value is contained in the 95% confidence levels. We highlight here an important statement about the computation of χ^2_i in equation 5.8. The datasets entering the computation are only those which are affected by the EFT corrections: namely the DIS measurements that have a reach in Q above 120 GeV and the high-mass data

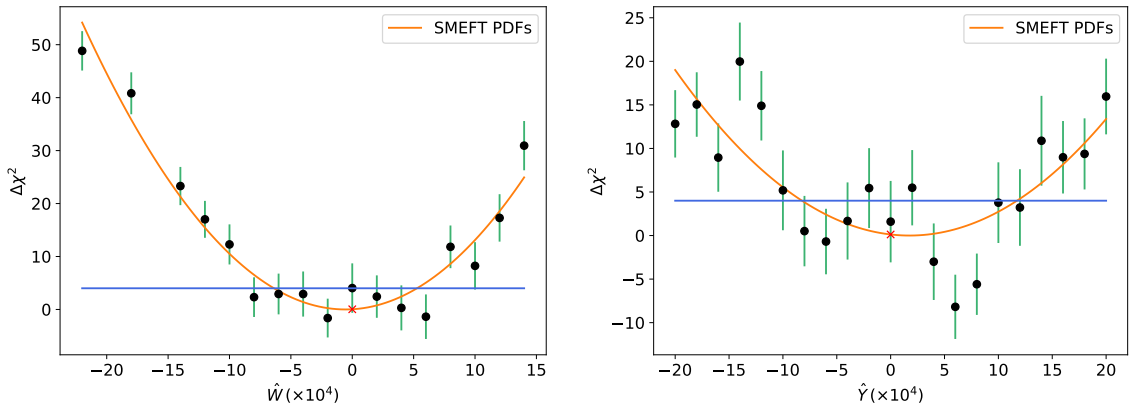


Figure 5.9: The $\Delta\chi^2$ parabolic fits, corresponding to the Taylor expansion of equation 5.5, for both the $\hat{W}$ (left) and $\hat{Y}$ (right) scenarios in the case of SMEFT PDFs. The error bars are calculated using the bootstrap method of section 5.2.3. The only data entering the χ^2_i computation (black dots) are those affected by the SMEFT corrections. The horizontal line in blue depicts $\Delta\chi^2 = 4$ which corresponds to a 2σ ($\sim 95\%$ CL) interval. The red cross depicts the SM point.

of table 5.5. We shall refer to the χ^2 obtained in this way as the partial-χ^2 rather than the global-χ^2 that would in principle use all the data of figure 5.8. The need for this approximation is down to the fact that the statistical fluctuations corresponding to the global-χ^2 are too large to obtain stable estimates on the Wilson coefficient bounds. The only way to tame these fluctuations is to significantly increase the number of MC replicas or to increase the density of benchmark points in the region of EFT parameter space being explored. This approximation is, however, well justified since the dominant contribution to the global-χ^2, as a function of the benchmark points, originates from the SMEFT corrections to the partonic cross section and from the impact of the SMEFT modifications on the PDFs. The datasets most sensitive to these effects are precisely those that induced them in the first place, that is: the $Q > 120$ GeV DIS data and the high-mass Drell-Yan measurements. As such, restricting to the partial-χ^2 captures the dominant effects, whilst simultaneously minimizing the level of statistical fluctuations. We will show how we can eliminate this approximation altogether using the `SIMUnet` approach of chapter 6 by considering the impact of the global dataset on the PDFs and Wilson coefficients from the ground-up determination of both.

In figure 5.10 we compare these same parabolas with those one would obtain assuming fixed SM PDFs. For both operators a broadening is observed reminiscent of those seen in the DIS only study of section 5.3. While the best-fit value of $\hat{W}$

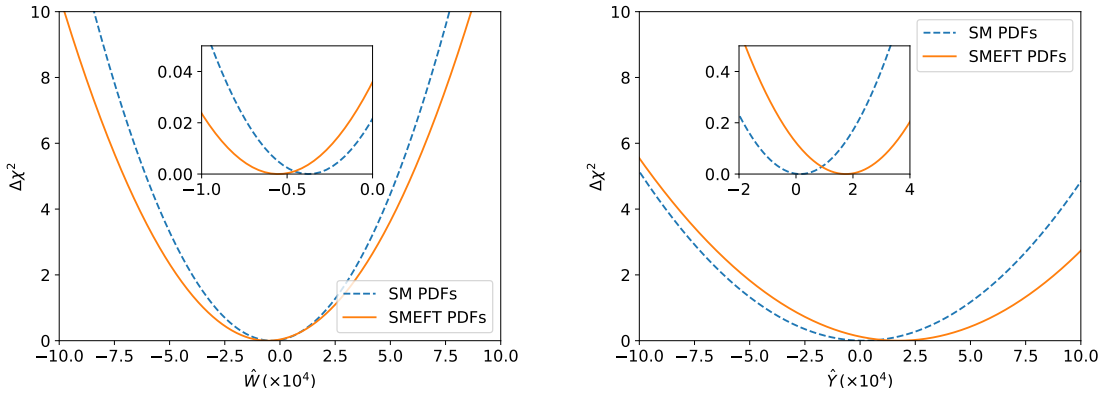


Figure 5.10: Comparison between the $\Delta\chi^2$ parabolic fits obtained using SMEFT PDFs (orange) to those one would obtain assuming fixed SM PDFs (blue dashed) for individual $\hat{W}$ (left) and $\hat{Y}$ (right) scenarios. The insets zoom on the region close to the minima.

remains approximately similar in both the SM and SMEFT PDF case, we see that the simultaneous determination results in a shift in the best-fit of $\hat{Y}$ by approximately $+2 \times 10^{-4}$. Though the effect of the broadening in confidence levels occurs in a region excluded by LEP [237] the effect would most certainly affect electroweak precision tests of the SMEFT at the LHC [249].

We define two useful quantities when considering the bounds in this part of the study. The *best-fit shift* corresponds to the difference in best fit values of $\hat{W}$ when using SMEFT and SM PDFs:

$$\text{best-fit shift} = \left(\hat{W}^{(0)}\Big|_{\text{SMEFT PDF}} - \hat{W}^{(0)}\Big|_{\text{SM PDF}}\right) \tag{5.76}$$

where $\hat{W}^{(0)}$ is the best-fit value of $\hat{W}$ obtained using either SMEFT or SM PDFs. The fractional increase in confidence level width relative to the SM PDF bounds defines the *broadening*:

$$\text{broadening} = \left(\Delta\hat{W}^{(0)}\Big|_{\text{SMEFT PDF}} - \Delta\hat{W}^{(0)}\Big|_{\text{SM PDF}}\right) \Big/ \Delta\hat{W}^{(0)}\Big|_{\text{SM PDF}} \tag{5.77}$$

where $\Delta\hat{W}^{(0)}$ is the width of the confidence intervals obtained using SM or SMEFT PDFs. Analogous expressions are defined for the $\hat{Y}$ oblique parameter. These expressions, along with the SM and SMEFT PDF bounds at 68% and 95% are tabulated in table 5.6

For the SM PDF case, we present bounds with and without PDF uncertainties, where the contribution from PDF uncertainties can be included using the procedure

Table 5.6: The 68% CL and 95% CL bounds on the $\hat{W}$ and $\hat{Y}$ parameters obtained from the corresponding parabolic fits to the $\Delta\chi^2$ values calculated from either the SM or the SMEFT PDFs. For the SM PDF results, we indicate the bounds obtained with (lower) and without (upper entry) PDF uncertainties accounted for using the approach outlined in section 5.2.2. The SMEFT PDF bounds already include PDF uncertainties by construction: being constructed from a global set of PDFs. The fourth and fifth column indicate the absolute shift in best-fit values, equation 5.76, and the percentage broadening of the EFT parameter uncertainties, equation 5.77, when the SMEFT PDFs are consistently used instead of the SM PDFs.

	SM PDFs	SMEFT PDFs	best-fit shift	broadening
$\hat{W} \times 10^4$ (68% CL)	$[-3.0, 2.2]$	$[-3.5, 2.4]$	-0.2	$+13\%$
	$[-4.3, 3.8]$		-0.3	-27%
$\hat{W} \times 10^4$ (95% CL)	$[-5.5, 4.7]$	$[-6.4, 5.3]$	-0.2	$+15\%$
	$[-6.8, 6.3]$		-0.3	-11%
$\hat{Y} \times 10^4$ (68% CL)	$[-4.4, 4.7]$	$[-3.4, 6.9]$	$+1.6$	$+13\%$
	$[-6.7, 7.5]$		$+1.4$	-27%
$\hat{Y} \times 10^4$ (95% CL)	$[-8.8, 9.2]$	$[-8.3, 11.8]$	$+1.6$	$+12\%$
	$[-11.1, 12.0]$		$+1.3$	-13%

of section 5.2.2; the bounds for SMEFT PDFs incorporate the PDF uncertainties by construction. The interesting result here is that, as with the DIS study, a broadening of the results is observed when PDF uncertainties are neglected, however, when accounted for the bounds in fact are improved upon. One would expect that with a better control of the PDF uncertainties, the same broadening behaviour would be seen in both cases. Indeed an interesting study would be to perform this analysis employing the NNPDF4.0 approach. In this case, the reduced PDF uncertainties could in principle be sufficient to achieve this regime. We reserve such an analysis for a future study, but remark here that the change in bounds remains mild, but not negligible; indicative of a slight absorption of the new physics effects by the neural network parameterization. Such a light effect can be understood by the scarcity of the high-mass Drell-Yan datasets available. Indeed, the highest data comes from LHC Run II and only one dataset at the full kinematic reach of 13 TeV. Indeed comparing table 5.4 with table 5.5, we see that the number of EFT sensitive datasets is eclipsed by the low-mass measurements

as well as the 3092 DIS measurements. This achieves a strong control over the PDF determination thus not allowing them to be heavily modified by the new physics effects.

We now turn our attention to the modification of the PDFs themselves. Note that since we are considering hadron-hadron processes, involving the ejection of a parton from each, we are phenomenologically interested in the PDF luminosities (convolution of the PDFs) and not the PDFs themselves. We plot the luminosities for the gluon-quark, quark-antiquark, and quark-quark channels in figure 5.11 at $\sqrt{s} = 14$ TeV as a function of the produced final state invariant mass, M_X. Shown are both the $\hat{W}$ (upper) and $\hat{Y}$ (lower) luminosities for representative benchmark points corresponding to the boundaries of the 95% confidence intervals. For clarity of the figure we display the SM luminosity along with its 68% PDF uncertainty, while for the SMEFT luminosities we show only the central values. We have verified that the PDF uncertainty is unchanged in the SMEFT PDFs as in the DIS only study. In this case, the luminosity deviations from the SM remain small, with the largest deflection seen in the $q\bar{q}$ channel. This can be explained by the fact that in neutral-current DY, the dominant production mechanism is mediated primarily by the $u\bar{u}$ and $d\bar{d}$ combination and so the modifications to this process induced by the Scenario I operators of equation 5.62 mostly affect this particular channel. Certainly in all cases the deviation is contained within the PDF uncertainties, illustrative again of the fact that the interplay between the EFT effects and PDFs remains moderate in the tails of high-mass Drell-Yan distributions considered in this work.

An important point here is how one would disentangle these EFT-induced shifts in PDF luminosity from other possible sources of deviations, such as internal inconsistencies and tensions between datasets or missing higher orders in the perturbative expansion of the partonic cross section. We can resolve this issue in a similar fashion as to what was done for figure 5.7 in the DIS only study. We use the energy-growing effects associated with the higher-dimensional EFT operators which in turn translate into an enhanced sensitivity to the $\hat{W}$ and $\hat{Y}$ parameters for large values of dilepton invariant mass, $m_{\ell\ell}$. Thus, a useful ratio to consider is:

$$R_{\chi^2}\left(m_{\ell\ell}^{(\max),\hat{W},\hat{Y}}\right) = \frac{\chi^2\left(m_{\ell\ell}^{(\max)},\hat{W},\hat{Y}\right)}{\chi^2\left(m_{\ell\ell}^{(\max)} = 120\text{ GeV},\hat{W},\hat{Y}\right)} \tag{5.78}$$

where $m_{\ell\ell}$ is the upper bound on the value of dilepton invariant mass bins that enter the χ^2 calculation. As in figure 5.7, the χ^2 calculation is done for both SM and SMEFT PDFs at various benchmark point. The denominator is the χ^2 in the kinematic region

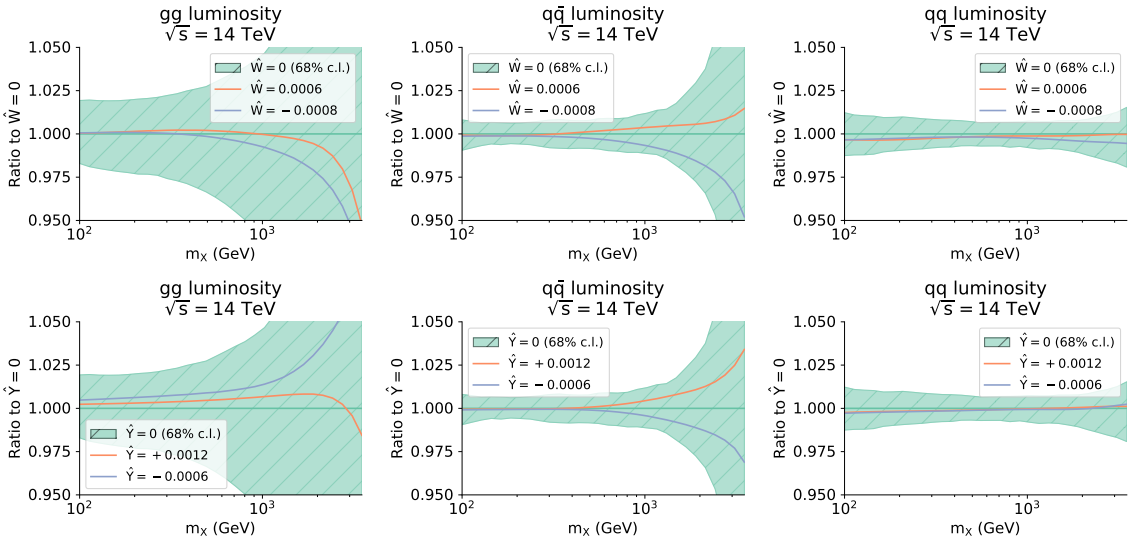


Figure 5.11: Comparison between the SM PDF luminosities with their SMEFT counterparts, displayed as ratios to the central value of the SM luminosities, for representative values of the $\hat{W}$ (upper) and $\hat{Y}$ (lower panel) parameters. The values of $\hat{W}$ and $\hat{Y}$ are chosen to be close to the upper and lower limits of the 95% CL intervals reported in table 5.6. From left to right we show the gluon-gluon, quark-antiquark, and quark-quark channels.

where sensitivity to the EFT operators is negligible. We show R_{χ^2} for representative choices of Wilson coefficients in figure 5.12. The R_{χ^2} estimator is approximately unity for small choices of $m_{\ell\ell}^{(\max)}$ consistent with the fact that all EFT sensitive binnings have been cut out. For increasing values of $m_{\ell\ell}^{(\max)}$, the ratio includes contributions from $m_{\ell\ell}$ bins more sensitive to the EFT effects. Thus, for the case where R_{χ^2} is computed at a non-zero choice of $\hat{W}$ and $\hat{Y}$, but using SM PDFs (orange), we expect a degradation in fit quality, since there is a mismatch between the partonic cross section used in the χ^2 computation with that used during the fits of the PDFs. Correcting for this mismatch and using instead SMEFT PDFs (green) causes R_{χ^2} to improve in the high invariant mass bins. However, we note the improvement here is mild, dropping only slightly below the mismatched case and still far from the SM PDF and $\hat{W} = \hat{Y} = 0$ case (blue). This is precisely the same behaviour as seen in figure 5.7 indicative again of a mild interplay between the EFT effects and PDFs.

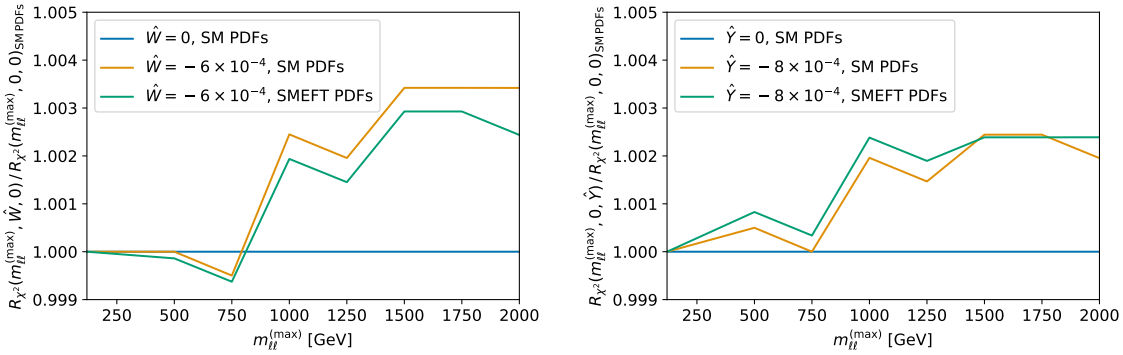


Figure 5.12: The R_{χ^2} estimator as defined in equation 5.78 normalized to its SM value (that is: $\hat{W} = \hat{Y} = 0$ and using SM PDFs) as a function of $m_{\ell\ell}^{(\max)}$. Shown are representative choice of $\hat{W}$ (left) and $\hat{Y}$ (right). We display the results obtained both with SM PDFs (orange and blue) and SMEFT PDFs (green).

5.4.6 Constraints on muonphilic operators with high-mass Drell-Yan data: Scenario II

We now consider the BSM scenario II of section 5.4.3 which adds a muonphilic BSM operator to the SM lagrangian, coupling preferentially to the dimuon DY process and leaving the dielectron production to be described by SM dynamics. This property alone implies that to constrain the $\mathbf{C}_{33}^{Du}$ Wilson coefficient, we require high-mass Drell-Yan measurements with an exclusive measurement of dimuon final states. From table 5.5 we see the only measurements satisfying this condition are CMS at 7 and 13 TeV amounting to a total of 170 data points. As such, we expect the PDF-EFT interplay to be even milder than in the oblique parameter scenario and as such we perform here a fixed PDF determination of $\mathbf{C}_{33}^{Du}$, reserving the joint determination for when HL-LHC projections are included in section 5.5.

In figure 5.13 we display the results of three quartic fits to the $\chi^2\left(\mathbf{C}_{33}^{Du}\right)$ profile in benchmark scenario II, akin to the functional form of equation 5.5, but extended to include the quartic terms. As in the above discussion, the calculation of χ_i^2 for each benchmark point is the partial-χ^2 and includes only the contributions from the two available DY measurements with a dimuon final state. We present fits based on cross sections that account only for the linear, only for the quadratic, and for both the linear and quadratic terms in the EFT expansion. In all cases, these cross sections are computed using the baseline SM PDF set. The inset displays the outcome of the linear EFT fit with an enlarged x-axis range. Figure 5.13 nicely illustrates the fact that the muonphilic operator is effectively completely unconstrained at the $\mathcal{O}(1/\Lambda^2)$

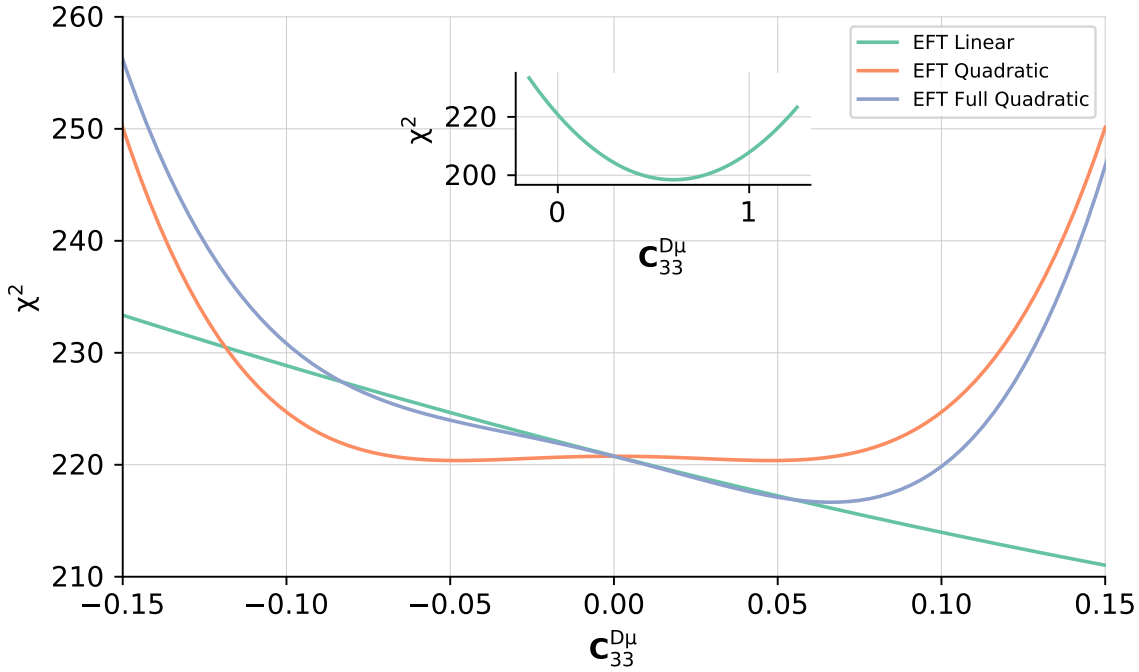


Figure 5.13: The results of quartic polynomial fits to $\chi^2\left(\mathbf{C}_{33}^{D\mu}\right)$, in scenario II. The χ^2 computed at each benchmark point includes only the contributions from the two DY measurements in the dimuon final state. We display results for fits based on cross sections that account only for the linear, only for the quadratic, and for both linear and quadratic terms in the EFT expansion, in all cases using the baseline SM PDF set. The inset displays the fit to the linear EFT values with an enlarged x-axis range.

level: requiring a much extended x-axis in the inset for the parabolic fit to be visibly concave. Only once the $\mathcal{O}(1/\Lambda^4)$ terms, arising from the SMEFT-SMEFT diagram in the squared amplitude, are included can this operator be constrained at any appreciable level. This behaviour is known [45] and can be traced back to the fact that the SM interference with this operator is suppressed. The 95% confidence levels obtained when using the full linear and quadratic cross section reads:

$$\mathbf{C}_{33}^{Dl} \in [-1.2, 10.7] \times 10^{-2} \tag{5.79}$$

which can be directly compared to equation 5.69 which is obtained in [45] and uses the ATLAS search data [252]. The source of discrepancy between the two bounds can be attributed to the fact that the dilepton search data benefits from an extended coverage of $m_{\ell\ell}$ as compared to the unfolded DY data used in our study.

5.5 PDF and EFT interplay at the High-Luminosity LHC

The discussions of sections 5.4.5 and 5.4.6 show that even with the inclusion of LHC measurements of high-mass Drell-Yan events, the interplay between PDFs and EFTs remains moderate; although not entirely negligible. However, with the High-Luminosity upgrade of the LHC set to accumulate more data, it is entirely possible that in the era of future colliders the effects will become more enhanced. In view of this, we repeat the analyses of sections 5.4.5 and 5.4.6, but this time include HL-LHC projections in the fit strategy. We follow the procedure of [253] (see also [254, 255]) to generate HL-LHC pseudo-data for NC and CC high-mass Drell-Yan processes at a center-of-mass energy of $\sqrt{s} = 14$ TeV making for a total integrated luminosity of $\mathcal{L} = 6$ ab^{-1}; with ATLAS and CMS each contributing $\mathcal{L} = 3$ ab^{-1}. The theoretical predictions for these forecasted observables are computed at NNLO in QCD and NLO electroweak (EW) corrections are additionally included. We include both dielectron and dimuon production cross sections through neutral and charged current Drell-Yan, the latter being the crucial process which will allow for the flat direction to be broken in the combined $(\hat{W}, \hat{Y})$ space. The NC distribution is binned in dilepton invariant mass, $m_{ll} \in [500, 4000]$ GeV while the CC distribution is binned in dilepton transverse mass, $m_T \in [500, 3500]$ GeV. For the precise details of the construction of these projections we refer the reader to [5].

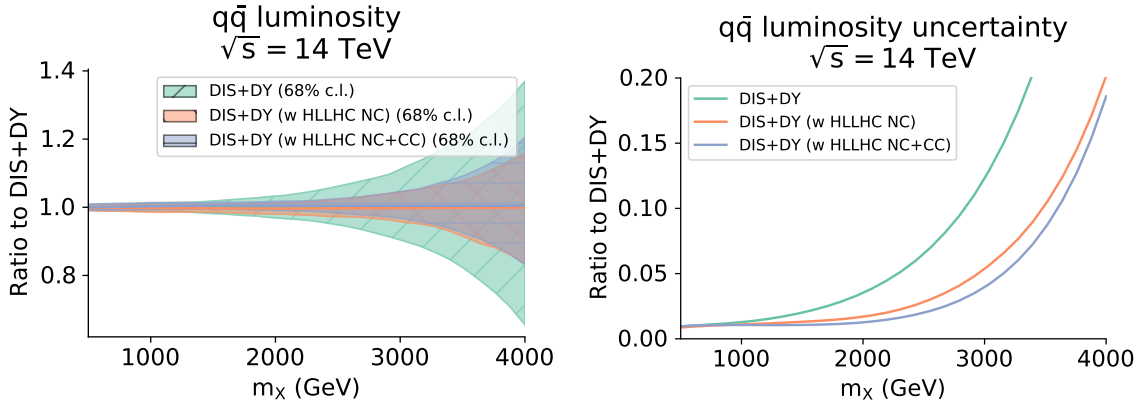


Figure 5.14: Impact of the HL-LHC pseudo-data on the quark-antiquark luminosity for the SM PDF fits as a function of the final state invariant mass, m_X. Shown are the luminosities (left) for the DIS+DY fit (green) and the corresponding fits including the HL-LHC pseudo-data, with only NC (orange) or also with CC (blue) cross sections, presented as a ratio to the central value of the former. Also shown is the relative PDF uncertainty in $\mathcal{L}_{q\bar{q}}$ (right) (with the central value of the DIS+DY baseline as reference) for the same fits.

In figure 5.14 we compare the quark-antiquark luminosities, at $\sqrt{s} = 14$ TeV, obtained with a fit to the DIS and DY data to those which include the NC HL-LHC projections and with the NC+CC HL-LHC projections. We see a marked reduction in the luminosity uncertainty (highlighted by the right panel) while the central value remains much the same. This is as expected since the DY process is dominated by the $u\bar{u}$ and $d\bar{d}$ channel. On the other hand, the gluon-quark and quark-quark luminosities experience little improvement after introducing the HL-LHC compared to the DIS+DY only fits. With this motivation, we thus expect an enhancement in PDF-EFT interplay as compared to the current LHC reach. Moreover, we now are able to present the results of a combined $(\hat{W}, \hat{Y})$ fit since we incorporate CC DY data. Furthermore, a simultaneous extraction of the muonphilic Wilson coefficient $\mathbf{C}_{33}^{D\mu}$ as well as the PDFs is presented, since the HL-LHC data is generated for both the dielectron and dimuon channel which can help constrain BSM scenario II.

5.5.1 The oblique parameters and the HL-LHC

The benchmark points for the simultaneous fit of $(\hat{W}, \hat{Y})$ are now a list of tuples, for which we construct a total of 35 benchmark points, scanning the 2-dimensional parameter space. Of these, 25 points are equally spaced in either $\hat{W} \in [-1.6, 1.6] \times 10^{-5}$ and $\hat{Y} \in [-8.0, 8.0] \times 10^{-5}$ and the remaining 10 points are equally spaced along the diagonal directions. We have verified that the addition of 12 extra benchmark points

Table 5.7: Same as table 5.6 for the 68% CL and 95% CL marginalised bounds on the $\hat{W}$ and $\hat{Y}$ parameters obtained from the two-dimensional ($\hat{W}$,$\hat{Y}$) fits that include the HL-LHC pseudo-data for NC and CC Drell-Yan distributions. As in table 5.6, for the SM PDFs we indicate the bounds obtained with (lower) and without (upper entry) PDF uncertainties accounted for.

	SM PDFs	SMEFT PDFs	best-fit shift	broadening
$\hat{W} \times 10^5$ (68% CL)	$[-0.7, 0.5]$	$[-4.5, 6.9]$	1.3	850%
	$[-1.0, 0.9]$		1.3	500%
$\hat{W} \times 10^5$ (95% CL)	$[-1.0, 0.8]$	$[-8.1, 10.6]$	1.4	940%
	$[-1.4, 1.2]$		1.4	620%
$\hat{Y} \times 10^5$ (68% CL)	$[-1.8, 3.2]$	$[-6.4, 8.0]$	0.1	190%
	$[-3.7, 4.7]$		0.3	70%
$\hat{Y} \times 10^5$ (95% CL)	$[-3.4, 4.7]$	$[-11.1, 12.6]$	0.1	190%
	$[-5.3, 6.3]$		0.3	110%

(8 further away from the origin and 4 along the principal axes) does not affect the confidence level contours obtained, thus verifying the stability of our procedure.

In table 5.7 we tabulate the 68% and 95% CL marginalise bounds on the Wilson coefficients obtained in a simultaneous fit of $(\hat{W}, \hat{Y})$ both for the case of SM PDFs and SMEFT PDFs. Bounds are shown at 68% and 95% as well as the best-fit shift and broadening defined respectively in equation 5.76 and equation 5.77. For the SM PDF bounds we give intervals with and without PDF uncertainties. In the case of HL-LHC, the interplay is very significant and cannot, at all, be neglected. Indeed, in this scenario, one would obtain completely inconsistent bounds if the back-reaction of the Wilson coefficients on the PDFs is neglected. The SM PDFs (which ignore this back-reaction) give bounds significantly tighter that the SMEFT PDFs, with the broadening for $\hat{W}$ ($\hat{Y}$) being 620% (110%) at 95% CL when PDF uncertainties are accounted for. Clearly, the situation would be worse if PDF uncertainties are neglected.

In table 5.8 we present the same results as table 5.7, but having replaced the SM PDFs with *conservative* SM PDFs. A conservative PDF is the one a fastidious particle physicist would use: conscious of the perils of an inconsistent treatment between the EFT and PDFs, they would choose to omit the high-mass Drell-Yan and HL-LHC data from the PDF fit when performing a BSM analysis. Thus, the conservative PDF set is

Table 5.8: Same as table 5.7 for the 68% and 95% CL marginalised bounds on the $\hat{W}$ and $\hat{Y}$ parameters obtained from the two-dimensional ($\hat{W}$,$\hat{Y}$) fits that include the HL-LHC pseudo-data for NC and CC Drell-Yan distributions. The input PDF set for the analysis done using fixed SM PDFs (corresponding to the results displayed in the column "SM cons. PDFs") is a conservative PDF set that does not include any of the high-mass distributions or the HL-LHC projections. The limits obtained from the simultaneous fit of PDFs and Wilson coefficients (corresponding to the results displayed on the column "SMEFT PDFs") are the same as those in table 5.7.

	SM cons. PDFs	SMEFT PDFs	best-fit shift	broadening
$\hat{W} \times 10^5$ (68% CL)	$[-1.0, 0.0]$	$[-4.5, 6.9]$	1.7	1000%
	$[-4.0, 2.8]$		1.8	70%
$\hat{W} \times 10^5$ (95% CL)	$[-1.4, 0.4]$	$[-8.1, 10.6]$	1.8	940%
	$[-4.3, 3.1]$		1.9	150%
$\hat{Y} \times 10^5$ (68% CL)	$[2.1, 7.0]$	$[-6.4, 8.0]$	-3.7	190%
	$[-3.4, 11.2]$		-3.6	-1%
$\hat{Y} \times 10^5$ (95% CL)	$[0.5, 8.5]$	$[-11.1, 12.6]$	-3.7	200%
	$[-5.0, 13.7]$		-3.6	30%

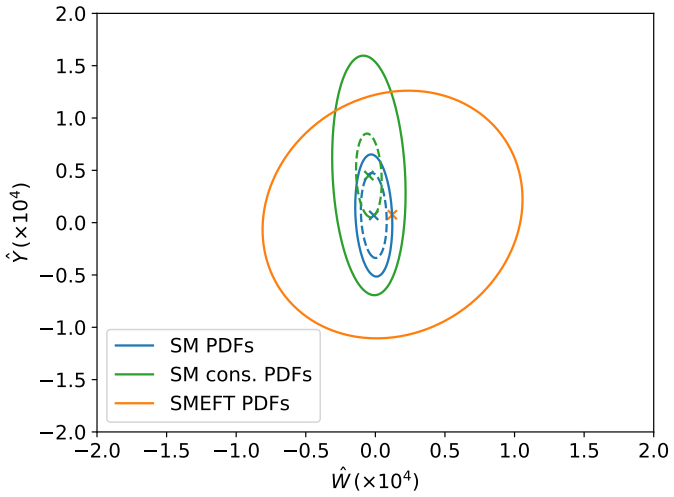


Figure 5.15: The 95% confidence level contours in the $(\hat{W}, \hat{Y})$ plane obtained using SM PDFs (blue), conservative (which omit the high-mass DY and HL-LHC data) PDFs (green), and the SMEFT PDFs (orange). The dashed ellipses account for the PDF uncertainty. The crosses depict the best-fit value.

fitted to the DIS data, low-mass DY and on-shell DY data of table 5.4. We observe that using conservative PDFs and accounting for PDF uncertainties, the EFT bounds increase when compared to the full SM PDF bounds, albeit with a much smaller broadening value. As a result, the size of the bounds obtained by keeping fixed SM PDFs is closer to the size obtained from the simultaneous fits, although still slightly underestimated. At the same time, the shift in the best-fit becomes more marked. We owe this behaviour to the fact that the fewer data points used in the fit causes the PDF to be less well constrained. Thus the bounds with PDF uncertainty are significantly wider than when they are omitted: highlighting the importance of incorporating and taming the PDF uncertainties. The results are displayed graphically in figure 5.15, where 95% confidence level contours in the $(\hat{W}, \hat{Y})$ plane are obtained using SM PDFs, conservative PDFs, and the consistent SMEFT PDF treatment. We also show the effect of including PDF uncertainties, which has the greatest effect on the conservative PDFs.

The increased role that the interplay between PDFs and EFT coefficients will play at the HL-LHC can also be illustrated by comparing the expected behaviour of the quark-antiquark luminosity, shown in figure 5.16, for the SMEFT PDFs corresponding to representative values of $(\hat{W}, \hat{Y})$ relative to the SM PDFs. Indeed, the central value of the quark-antiquark luminosity for SMEFT PDFs lies well outside the 1σ error band

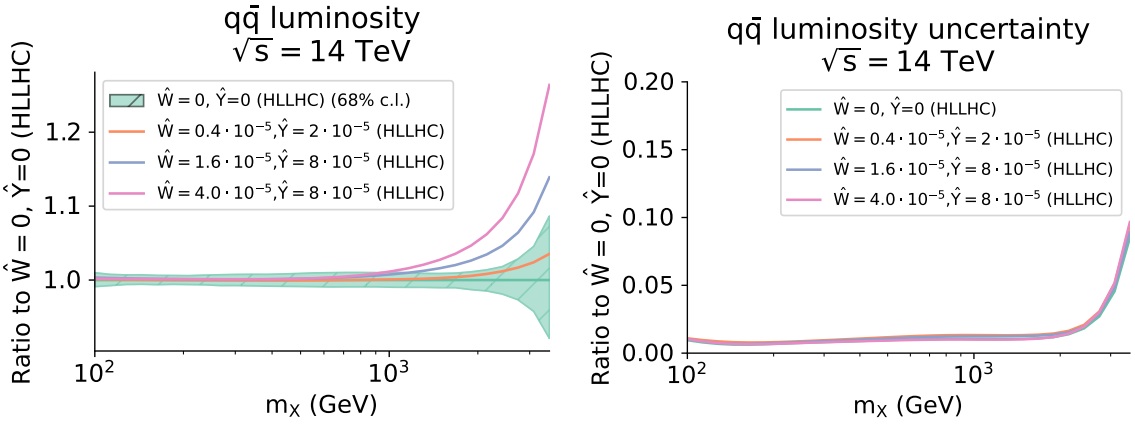


Figure 5.16: Same as figure 5.14, now comparing the quark-antiquark SM PDF luminosities in the fits including HL-LHC pseudo-data with those obtained in the SMEFT PDF fits for representative values of $(\hat{W}, \hat{Y})$.

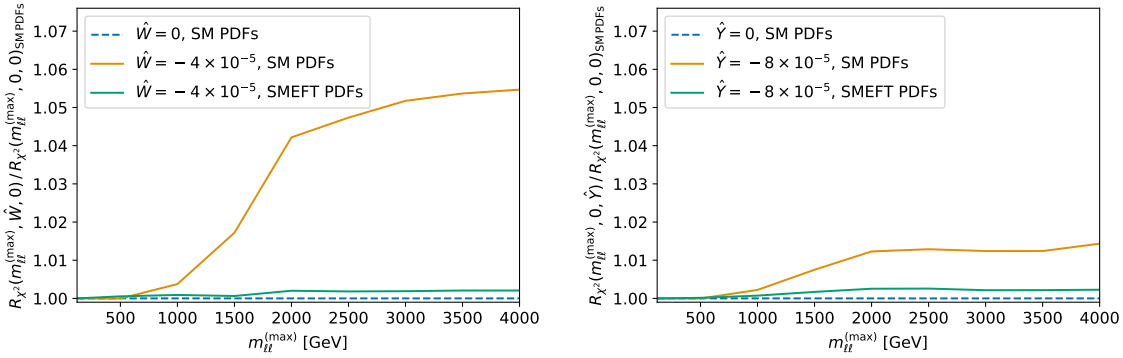


Figure 5.17: Same as figure 5.12, but now for the fits including the HL-LHC pseudo-data. In the charged current case $m_{\ell\ell}^{(\max)}$ refers to the transverse mass, m_T.

of the SM PDFs, while the PDF uncertainties themselves are unchanged. This change in central value of the large-x PDFs reabsorbs the EFT effects in the partonic cross section induced by the SMEFT operators and leads to better χ^2 values as compared to those obtained with the SM PDFs.

The importance of considering the PDF-EFT interplay at the HL-LHC is nicely illustrated by figure 5.17, where we again plot the R_{χ^2} estimator defined in equation 5.78, now for the HL-LHC case. The same monotonic growth of R_{χ^2} is seen as in figure 5.12, with the reasoning being identical to that discussed in the surrounding text. However, the most striking difference is that the SMEFT PDF curve (green) is virtually identical to the SM PDF with vanishing Wilson coefficients curve (blue). This is indeed the smoking gun we've been waiting to see: it shows that the EFT effects

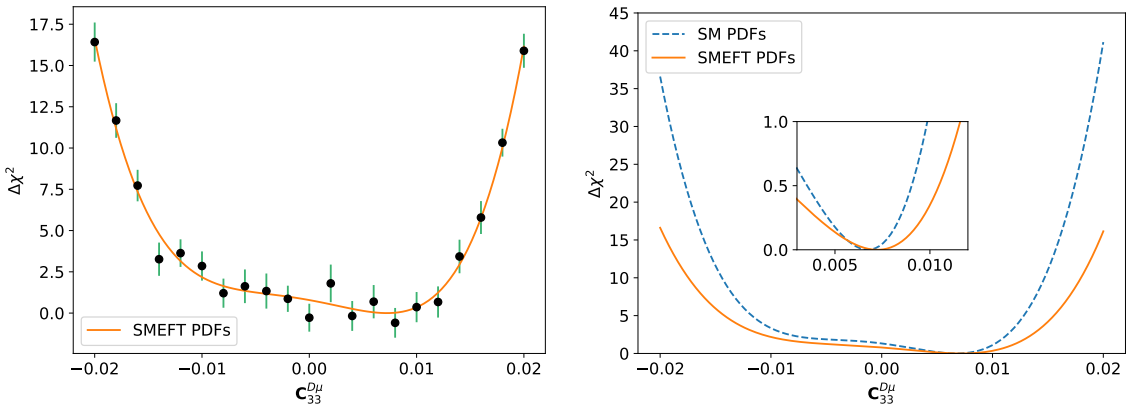


Figure 5.18: The values of $\Delta\chi^2$ obtained for the SMEFT PDFs as a function of $\mathbf{C}_{33}^{D\mu}$ using the HL-LHC data as well as the corresponding quartic polynomial fit (left). We also compare the polynomial fit for SMEFT PDFs to that obtained with the SM PDF counterpart (right). The inset zooms on the neighborhood of the global minimum.

are almost entirely absorbed into the PDFs during fitting. The findings here act as an impetus to devise a methodology which can systematically disentangle PDF fitting from possible BSM effects.

5.5.2 Lepton flavour universality violating operators at the HL-LHC

With the addition of separated dielectron and dimuon channels from the HL-LHC projection pseudo-data, we now have sufficient constraining power to attempt a simultaneous determination of PDFs and $\mathbf{C}_{33}^{D\mu}$ of scenario II. We recall that the CMS high-mass measurements are also provided in separate channels and so this dataset is also used in the determination.

The benchmark points used for this study are 21 points evenly distributed in the closed interval $\mathbf{C}_{33}^{D\mu} \in [-0.02, 0.02]$. Following on from the results obtained in section 5.4.6, we include the $\mathcal{O}(1/\Lambda^4)$ terms in the partonic cross section and perform quartic polynomial fits to the corresponding χ_i^2: that is expanding equation 5.5 to quartic order.

The result of this process is shown in figure 5.18, for both SM and SMEFT PDFs. Again the error bars are computed using the bootstrap method outlined in section 5.2.3. The corresponding 68% and 95% confidence levels are shown in table 5.9. We have the interesting result that in the case of scenario II, the interplay of PDFs and EFT dynamics remains moderate, even at the HL-LHC. Despite the consistent treatment of

Table 5.9: Same as table 5.7, now for the $\mathbf{C}_{33}^{Dμ}$ parameter from EFT benchmark scenario II.

	SM PDFs	SMEFT PDFs	best-fit shift	broadening
$\mathbf{C}_{33}^{Dμ} \times 10^2$ (68% CL)	$[-0.1, 1.1]$	$[-0.3, 1.2]$	0.06	25%
$\mathbf{C}_{33}^{Dμ} \times 10^2$ (95% CL)	$[-1.0, 1.2]$	$[-1.2, 1.4]$	0.06	18%

PDFs and Wilson coefficients of scenario I (see figure 5.15 and surrounding text), the bounds obtained for scenario II loosen by around only 30%. The origin of this behaviour is explained by the fact the dielectron channel does not receive EFT corrections, hence all the information provided by the e^-e^+ channel goes in to exclusively constraining the PDFs. The muon channel distributions then determine the allowed range for $\mathbf{C}_{33}^{Dμ}$, which is now restricted by the well-constrained large-x quark and antiquark distributions thanks to the electron data. This finding emphasizes how the availability of separated leptonic final states is of utmost importance to test BSM models that account for lepton flavour universality violation.

In this chapter we have presented a first analysis of the interplay between PDFs and possible BSM dynamics. Indeed, we may be indirectly sensitive to such BSM effects in the highest energy unfolded data that are included in both PDF and EFT fits. We assessed the questions of to what extent can a PDF parameterization absorb possible BSM effects and also the effect of a consistent analysis of PDFs and EFTs would have on bounds obtained on each. We did so by considering first a DIS only study, before including proton-proton measurements at the LHC in the form of high mass Drell-Yan processes. The effect was deemed ostensibly mild with current observations; however, the use of HL-LHC projected data suggested that the effect can not only become more pronounced at the HL-LHC, but the results one would obtain by neglecting this EFT induced back-reaction would be completely misleading.

We acknowledge that the benchmark point based methodology, used in this work, though able to disentangle PDFs from BSM effects, cannot be deemed a *truly* simultaneous determination. Indeed, it is not possible to plot the PDF set that one would obtain from this approach, restricted instead to only displaying representative choices of benchmark point. Moreover, a deliverable PDF for phenomenological use is also not provided by this approach. Finally, we remark that, while enforced upon us due to stability reasons, the use of the partial-χ^2 in section 5.4.5 is not entirely satisfactory if a global analysis of all available datasets is to be performed. With these points in

mind, we consider the `SIMUnet` methodology next, which serves as a new generation of simultaneous determination methodology.

Chapter 6

A new generation of simultaneous global fits

DESPITE the ability to disentangle the interplay between PDF and BSM dynamics using the approach outlined in chapter 5, there are various extensions which the methodology outlined in section 5.1 cannot provide. For example, we can obtain bounds on the Wilson coefficients by scanning benchmark points in the EFT coefficient space, but it is not clear what the resulting PDF from this process should be. Indeed, for the various PDF and luminosity plots presented, one is restricted to show only representative choices of benchmark points. As such this discretized scan of EFT space cannot promise to provide a single simultaneously determined PDF set that is free from the effects of possible contamination that arises from BSM resonances. Moreover, a truly global fit of the LHC data, in the context of indirect searches for new physics, would not only account for possible BSM operators, but also treat all the free parameters of the theory, such as the strong coupling, the heavy quark masses, the gauge boson masses, etc. on an equal footing and fit the entire set of independent parameters simultaneously. Indeed, it has been shown that unless PDFs and $\alpha_s(M_Z)$ are fitted simultaneously (when using hadronic processes), then one necessarily achieves inconsistent results: failing to obtain the most optimal values for both [256]. The case of α_s and PDFs is special, since the two are highly correlated thanks to the DGLAP evolution, but in principle the same argument applies to all the other aforementioned parameters.

In this chapter we present a novel methodology which can fit, truly simultaneously, the PDFs alongside any external parameter that affects the theory predictions, whether within the SM or beyond it. The methodology, dubbed `SIMUnet` [6], is based on the

NNPDF4.0 methodology of chapter 3 and introduces a new sequential layer to capture the data dependence on any external parameters with the PDF sector of the neural network architecture free to capture the data dependence on the PDFs. We showcase the methodology by performing a truly simultaneous fit of the Standard Model Effective Field Theory operators considered in section 5.4. The results of chapter 5 thus forms a benchmark for the present chapter. We demonstrate how this ground-up fit of PDFs and external parameters can be made without making the partial-χ^2 approximation of section 5.4. We demonstrate the robustness of `SIMUnet` by performing a closure test [219, 257] whereby we contaminate our data with artificially chosen choices of Wilson coefficient and show how the methodology can not only retrieve these Wilson coefficients, but can also replicate a known underlying PDF law. The `SIMUnet` methodology thus acts as a significant step forward towards a global fit of the Standard Model and forms the methodology of choice for all future studies akin to those of chapter 5.

6.1 Fast interface to theory predictions

In this section we outline a general formalism that allows for the fast interface of theory predictions and to isolate their dependence on the physical parameters that one may want to fit simultaneously with the PDFs. We then focus on simplifying this approach through the use of multiplicative K-factors before discussing in the next section how we extend the NNPDF4.0 framework to fit PDFs along with other physical parameters.

In order to fit a general set of external parameters alongside the PDFs, we first need to isolate their dependence during a fit. Recall from the discussion of section 3.2.3, all theory parameters are encapsulated in the FK tables. Using the notation employed in [145], we can write the theoretical prediction for any hadronic cross section as

$$T_I^{\rm hh} = \sum_{i,j=1}^{N_{\rm pdf}} \sum_{\alpha,\beta=1}^{N_x} \Sigma^I_{\alpha\beta ij}\, N^0_{\alpha i}\, N^0_{\beta j} \equiv N^0 \cdot \Sigma_I \cdot N^0, \tag{6.1}$$

where I indicates a specific hadronic observable included in a PDF fit; α, β are the indices of the interpolation grids in x-space for the first and second parton respectively; i, j are the indices of the PDFs of the initial-state partons that contribute to the observable I and N^0 are the neural network parametrization of the independent PDFs at the initial scale Q_0 (equation 3.24). The computation of the hadronic observables is reduced to a bilinear product over an interpolation grid in the $x_{1,2}$ space and the basis of the input PDFs for a given process. The quantity Σ is the FK-table, which

incorporates both the evolution of the PDFs from the initial scale to the scale of the measured observable and the partonic cross sections associated to each of the partonic channels that enter the computation of the hadronic cross section. For processes involving only one hadron and one lepton, like DIS, the expression is even simpler and reads:

$$T_I^{\rm hl} = \sum_{i=1}^{N_{\rm pdf}} \sum_{\alpha=1}^{N_x} \Sigma_{\alpha i}^I \, N_{\alpha i}^0 \equiv \Sigma_I \cdot N^0. \tag{6.2}$$

We can collectively refer to the theory prediction for a generic observable - whether it involves one or two hadrons in the initial states - as

$$T_I = \Sigma_I \cdot L^0, \tag{6.3}$$

where L^0 indicates either the parametrisation of one independent PDF at the initial scale or the product of two of them.

If we now consider how the theoretical predictions T_I for each observable I explicitly depend upon a single parameter c, that could be for example the strong coupling constant, α_s, or the top mass, m_t, the entire dependence upon this parameter is contained in the FK-tables Σ, since L^0 only captures the initial scale parametrization of PDFs, which is fitted from the data. For clarity of notation we focus here on the case where only one single parameter is fitted alongside PDFs, but the generalization to more parameters is a straightforward extension of the following argument, simply requiring the use of multi-variate Taylor expansions. Schematically we can write the FK-tables as

$$\Sigma_I(c) = \left[\hat{\sigma}(c) \otimes \Gamma(c)\right]_I \tag{6.4}$$

where $\hat{\sigma}$ is the partonic cross section and Γ are the evolution kernels that evolve the PDFs from the initial scale to the scale of T_I. The shorthand $\otimes$ denotes the usual convolution given by equation 2.37. For a standard PDF-only determination, the parameter c is typically fixed to a certain value during the computation of the FK table. In general both $\hat{\sigma}$ and Γ depend on the given parameter (the strong coupling being one such example), whilst in some other cases only the partonic cross sections, $\hat{\sigma}$, depends on the parameter under consideration (the Wilson coefficients of the SMEFT expansion being one such example).

If we now want to fit the parameter c alongside the PDFs, we need a fast interface to the dependence of each $\Sigma_I(c)$ upon the parameter c[1]. One possible way to achieve such a fast interface is to assume that both the evolution kernel and the partonic cross section are suitably analytic such that they can be accurately described by their Taylor expansion about some point c^*. The closer the point c^* is to the actual parameter, the greater the validity of truncating the Taylor series. Dropping from now on the observable index I, we can write

$$\begin{aligned}\Sigma(c) &= \sum_{p,q} \frac{(c-c^*)^{p+q}}{p!q!} \frac{\partial^p \hat{\sigma}(c^*)}{\partial c^p} \otimes \frac{\partial^q \Gamma(c^*)}{\partial c^q} \\ &= \sum_k (c-c^*)^k \sum_{\substack{p,q: \\ p+q=k}} \frac{1}{p!q!} \frac{\partial^p \hat{\sigma}(c^*)}{\partial c^p} \otimes \frac{\partial^q \Gamma(c^*)}{\partial c^q} \\ &= \sum_k (c-c^*)^k \Sigma_k(c^*), \end{aligned} \tag{6.5}$$

whereby for each power of c we have an order-by-order FK table, $\Sigma_k(c^*)$, that can be pre-computed before the fit, with the dependence on the parameter c being isolated from the FK table. In this way the task of querying the FK table for various values of c has been reduced to computing individual FK tables for each order and taking a weighted sum of these tensors which is a computationally trivial and fast operation. Importantly, such an operation can be implemented using the purely `TensorFlow` functionality already present in the NNPDF4.0 methodology, thus allowing for the gradients to be computed using automatic differentiation techniques [258].

6.1.1 Observable dependence on the strong coupling

To make the above discussion more concrete, we explicitly consider the case whereby we wish to isolate the FK table dependence on the strong coupling constant at the Z-pole, such that, following our above notation, we have $c = \alpha_s(m_Z)$ and $c^* = \alpha_s^{\rm PDG}(m_Z) = 0.1179(10)$ is the PDG value for the strong coupling evaluated at the Z-boson mass [177].

[1]In section 6.3 we will see that the reason for this is that gradient descent will assess the optimal value of c at every step during learning by repeatedly evaluating theory predictions with different c values.

The partonic cross section admits a power expansion in α_s using perturbation theory, which allows us to write the exact expression

$$\hat{\sigma}(c) = \sum_p (c - c^*)^p \, \hat{\sigma}_p(c^*) \tag{6.6}$$

where we have simply centred the order by order sum around c^*. For illustrative purposes, we restrict ourselves to the case where the process depends solely on the non-singlet quark distribution. The evolution kernel in Mellin space to leading order in the anomalous dimension (the Mellin transform of the splitting function) γ_0, is then given by [132]:

$$\Gamma(N, \alpha_s(Q^2), \alpha_S(Q_0^2)) = \left(\frac{\alpha_s(Q^2)}{\alpha_s(Q_0^2)}\right)^{-\gamma^0(N)/\beta_0} . \tag{6.7}$$

If we consider the leading-log evolution of α_s via renormalisation group equation from M_Z^2 to Q^2 or Q_0^2 and set $c = \alpha_s(M_Z^2)$, we get

$$\alpha_s(Q^2) = \frac{c}{1 + \beta_0 c \ln \frac{Q^2}{M_Z^2}} \qquad \alpha_s(Q_0^2) = \frac{c}{1 + \beta_0 c \ln \frac{Q_0^2}{M_Z^2}} \tag{6.8}$$

where $\beta_0 = 11 - \frac{2}{3} N_f$. Expanding the evolution kernel Γ according to equation 6.5, taking $c^* = \alpha_s^{\mathrm{PDG}}(M_Z) = 0.1179$, it is easy to see that we get

$$\Gamma(c) = \Gamma_0 + (c - c^*)\Gamma_1 + \cdots , \tag{6.9}$$

where

$$\Gamma_0 = \frac{1 + c^* \beta_0 \ln \frac{Q_0^2}{M_Z^2}}{1 + c^* \beta_0 \ln \frac{Q^2}{M_Z^2}} \tag{6.10}$$

$$\Gamma_1 = -\frac{\gamma_0(N)}{\beta_0} \Gamma_0^{-\frac{\gamma_0(N)}{\beta_0} - 1} \left(\frac{\beta_0 \ln \frac{Q_0^2}{M_Z^2}}{1 + c^* \beta_0 \ln \frac{Q^2}{M_Z^2}} - \frac{\left(1 + c^* \beta_0 \ln \frac{Q_0^2}{M_Z^2}\right) \beta_0 \ln \frac{Q^2}{M_Z^2}}{\left(1 + c^* \beta_0 \ln \frac{Q^2}{M_Z^2}\right)^2} \right). \tag{6.11}$$

Finally, once equation 6.9 is combined with equation 6.6 we obtain the order by order expansion for the FK tables:

$$K = \hat{\sigma}_0 \otimes \Gamma_0 + (c - c^*)\Big(\hat{\sigma}_1 \otimes \Gamma_0 + \hat{\sigma}_0 \otimes \Gamma_1\Big) + \cdots \tag{6.12}$$

$$\equiv K_0 + (c - c^*) K_1 + \cdots \tag{6.13}$$

which can, in principle, be computed and stored permanently in the usual way. In this way we can capture the FK table dependence on the strong coupling parameter in the neighborhood of some prior choice c^*.

So far, we have limited ourselves to the leading order evolution of the PDFs and the strong coupling constant, however a similar expression can be straight forwardly obtained at NLO and NNLO. The only caveat is that, from NLO onwards terms of the form $\ln\left(1+\beta_0 c \ln\frac{Q^2}{M_Z^2}\right)$ arise which spoil the validity of the Taylor expansion. The term $\ln\frac{Q^2}{M_Z^2}$ can in principle be made arbitrarily large thus exiting the unit disc which is the region of analyticity of $\ln(1+x)$. This problem can be circumvented, however, by noting:

$$\ln\left(1+\beta_0 c \ln\frac{Q^2}{M_Z^2}\right) = \ln\left(1+\beta_0 c^* \ln\frac{Q^2}{M_Z^2}\right) + \ln\left(1+\frac{\beta_0(c-c^*)\ln\frac{Q^2}{M_Z^2}}{1+\beta_0 c^*\ln\frac{Q^2}{M_Z^2}}\right) \tag{6.14}$$

where the rightmost term can now be Taylor expanded since $c - c^*$ can be made arbitrarily small so as to suppress the large logarithm. Indeed, note that in the worst case:

$$\lim_{Q^2\to\infty}\left((c-c^*)\frac{\beta_0\ln\frac{Q^2}{M_Z^2}}{1+\beta_0 c^*\frac{Q^2}{M_Z^2}}\right) = \frac{c-c^*}{c^*} \leq 1 \tag{6.15}$$

$$\implies c \leq 2c^* \tag{6.16}$$

which can be implemented within the optimizer to restrict it from venturing to values greater than $2c^*$.

Finally if one was to include electroweak corrections to the DGLAP evolution equation, as is for example done in `APFEL` [147], then electroweak parameters will in general also be present in the combined QCD and QED evolution operator. For such a scenario, the prescription outlined in this section must then be followed. However, the corrections to the pure QCD splitting functions introduced by electroweak considerations have no dependence on the CKM matrix elements, the weak mixing angle, θ_W, or gauge boson masses, amongst others. Such quantities manifest solely in the partonic cross section and in general yield a prescription much simpler than that outlined here.

6.1.2 Interpolation of Fast Kernel tables

Despite the appeal of the method of Taylor expanding the FK tables using equation 6.5, the practicality of this method may be rather restricted for the specific case of fitting the strong coupling at the Z-pole. The reason for this is that the computational implementation of equations such as 6.10 and 6.11 will be rather cumbersome in the (`Fortran` based) `APFEL` library [147]. Realistically, the picture will only get worse once non-singlet evolution is considered and NNLO expansions of the evolution kernel are used.

A viable alternative to the Taylor expansion approach for the determination of α_s and the PDFs is to compute various FK tables using preset values of the strong coupling and then perform an element-wise interpolation between these tensors. In this way one can avoid having to implement the Taylor series expansion, while still accurately replicating the linear behaviour of the FK table. More concretely, one selects a set of values for the strong coupling, $\mathcal{A} = \{\alpha_s^{(1)}, \ldots, \alpha_s^{(p)}\}$ (perhaps linearly spaced in the interval $[0.115, 0.122]$) and for each one an FK table is computed using `APFEL`: $\mathcal{S} = \{\Sigma^{(1)}, \ldots, \Sigma^{(p)}\}$. Then the FK table for a general choice of α_s is given by interpolating the FK table elementwise. For example, using a piecewise linear interpolant, the FK table as a function of α_s would be given by:

$$\Sigma(\alpha_s) = \Sigma(\alpha_s^{\downarrow}) + (\alpha_s - \alpha_s^{\downarrow}) \cdot \frac{\Sigma(\alpha_s^{\uparrow}) - \Sigma(\alpha_s^{\downarrow})}{\alpha_s^{\uparrow} - \alpha_s^{\downarrow}} \tag{6.17}$$

where $\alpha_s^{\uparrow}$ ($\alpha_s^{\downarrow}$) is the nearest neighbor from above (below) of α_s for elements in the set $\mathcal{A}$; that is:

$$\alpha_s^{\uparrow} = \underset{x \in \mathcal{A}_{\geq \alpha_s}}{\arg\min}(x - \alpha_s) \qquad \alpha_s^{\downarrow} = \underset{x \in \mathcal{A}_{\leq \alpha_s}}{\arg\max}(\alpha_s - x). \tag{6.18}$$

This method is appealing because such a piecewise linear interpolation is in an extremely quick computation with much less implementational overhead than implementing equations 6.10 and 6.11. However, the accuracy of mimicking the precise behaviour of Σ around the neighborhood of the knots $\alpha_s^{(i)}$ may be restricted if one uses too coarse a selection of $\mathcal{A}$. This problem can be overcome by choosing a more dense set of knots since in principle any continuous function can be replicated using an infinite number of piecewise linear functions. We shall reserve this endeavour for a future, though ongoing, study.

6.2 Observable dependence on the Wilson coefficients

We now focus on the parameters that we shall use to showcase the potential of our approach. Specifically, we are interested in fitting N parameters $\{c_n\}$ with $n = 1, \ldots, N$, each of which is a parameter associated to a Wilson coefficient of a given operator in the SMEFT expansion of equation 2.82.

As we have seen in chapter 5, to include the effects of the corrections coming from the dim-6 operators included in the SMEFT expansion in the theoretical prediction T_I for any observable included in the fit, one has to augment the SM partonic cross sections with the effects of the relevant operators with the linear and quadratic modifications of the SM cross section that the operators induce. Note that, given that the SMEFT operators that we consider here do not modify the PDF DGLAP evolution, the corrections will only appear in the partonic cross sections $\hat{\sigma}$ and, unlike in the case of the Taylor expansion around the PDG value of α_s discussed in section 6.1.1, here the sum is exact (and not approximated) when the Taylor expansion is truncated at order $k = 2$ (corresponding to linear and quadratic corrections). Recall from the discussion of equation 5.74, the EFT corrections can be adequately implemented using the K-factor approximation, by defining:

$$K(\{c_n\}) = 1 + \sum_{n=1}^{N} c_n R^{(n)}_{\text{SMEFT}} + \sum_{n,m=1}^{N} c_n c_m R^{(n,m)}_{\text{SMEFT}} \,, \tag{6.19}$$

where $R^{(n)}_{\text{SMEFT}}$ ($R^{(n,m)}_{\text{SMEFT}}$) corresponds to the SM-EFT (EFT-EFT) matrix element arising due to operator, n (n with m), normalized to the pure SM matrix element. This allows us to express a general cross section accounting for the dim-6 operators of equation 2.82 as

$$T = T^{\text{SM}} \times K(\{c_n\}) \tag{6.20}$$

where T is the SMEFT-modified theoretical prediction, T^{SM} is the state-of-the-art SM theoretical prediction including NNLO QCD and NLO EW corrections and $K(\{c_n\})$ are the SMEFT K-factors defined in equation 6.19. In this approach, the SMEFT predictions inherit factorisable higher-order radiative correction [45, 259]. The coefficients associated with the linear (quadratic) corrections $R^{(n)}_{\text{SMEFT}}$ ($R^{(n,m)}_{\text{SMEFT}}$) in equation 6.19 can be precomputed before the fit using a reference PDF set and then kept fixed. The impact of the coefficients $\{c_n\}$ can thus be included in the FK-tables, Σ, by a

simple multiplicative factor. This is convenient because we are then able to factorize the K-factors and thereby isolate the $\{c_n\}$ dependence. Schematically this reads

$$\Sigma(\{c_n\}) = [\hat{\sigma} \otimes \Gamma] \times K(\{c_n\}) = \Sigma^{\rm SM} \times K(\{c_n\}), \tag{6.21}$$

where the first factor does not depend on the Wilson coefficients and it is given by the same FK-tables that one computes for the standard NNPDF4.0 fits [1] which implicitly assume the SM to be valid at all scales at which observables are measured. The possibility of factoring the whole dependence upon the SMEFT parameters in a multiplicative factor simplifies the procedure highlighted in equation 6.5 and it thus simplifies the way in which the dependence upon the parameters $\{c_n\}$ is fitted within `SIMUnet`, as will be outlined below.

6.3 Methodology

In this section we move on to discuss the details of the new `SIMUnet` methodology. We show how, by extending the NNPDF4.0 methodology discussed in chapter 3, we can exploit the fast interface of the theory dependence on the external parameters presented in sections 6.1 and 6.2 to simultaneously fit the PDFs and external parameters. We show how the Monte Carlo method can be used to build a representation of not only the PDFs, but also of $\{c_n\}$. We also highlight various points of the NNPDF4.0 methodology that remain pertinent to our study, in particular the hyperparameter selection as well as the cross-validation techniques employed to avoid overfitting.

6.3.1 Neural network design

The NNPDF4.0 fit [1] that our methodology is built upon, shares the features of the previous NNPDF releases [137, 145], specifically the use of a Monte Carlo representation of PDF uncertainties and correlations, and the use of neural networks as basic interpolating functions. As such all the details of the fitting methodology, such as the choice of neural network architecture and the minimization algorithm, are now selected through an automated hyperoptimization procedure [260]. As such, our methodology employs state-of-the-art deep-learning techniques through publicly available and highly optimised Machine Learning libraries such as `TensorFlow` [127] and `Keras` [140]. As a result, it boasts both performance and improved fit quality using the cutting edge in optimiser technology. Moreover, through the use of `TensorFlow`

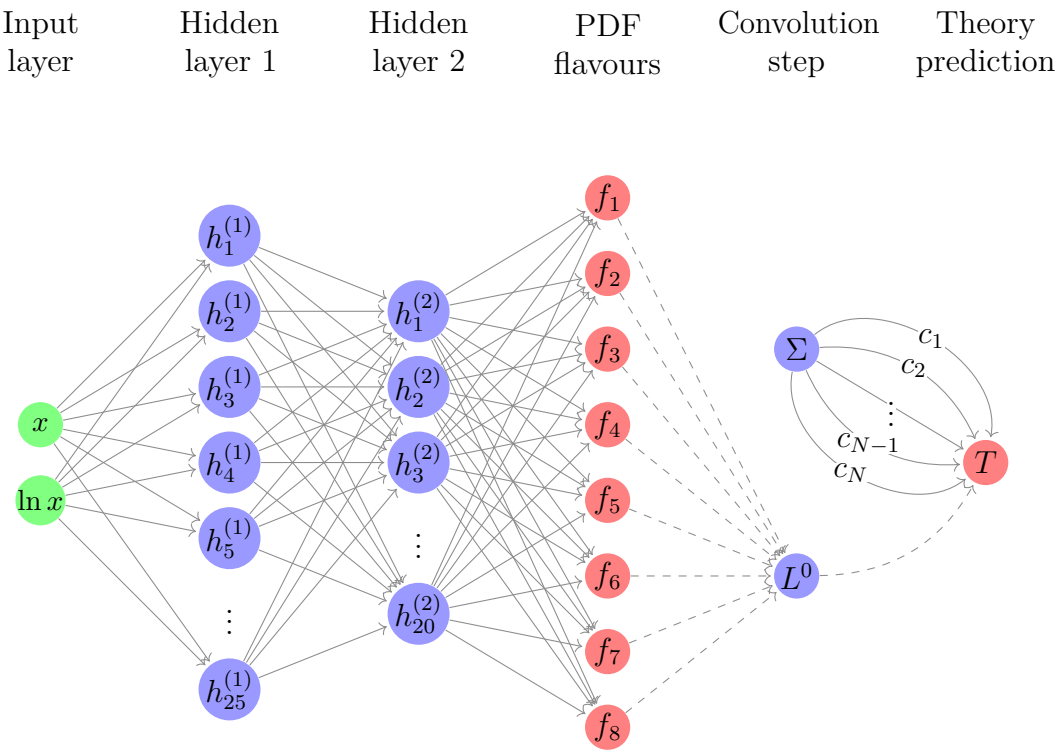


Figure 6.1: Schematic depiction of the `SIMUnet` methodology. The input nodes (shown in green) are Bjorken-x and its logarithm. The forward pass through the deep hidden layers (blue) are performed as in equation 3.6 to yield the output PDFs at the initial scale (red). The initial scale PDFs are then combined in the initial scale luminosity L^0, defined in equation 6.3. The initial scale luminosity is then convolved with the pre-computed FK-tables Σ (shown in blue) to obtain the theoretical prediction T (shown in red), which enters the figure of merit (equation 6.22) and is minimized during the fit. The Σ dependence on the parameters $\{c_n\}$ is fed into theoretical prediction T via the trainable edges of the combination layer. All trainable edges are shown by solid edges and are thus learned parameters determined through gradient descent, while dashed edges are non-trainable.

graph based execution, the code enjoys the readability `Python` is famed for, whilst still maintaining the performance of more traditional, statically typed, compiled languages. Additionally, we still retain parallel based execution capabilities, allowing for the ability to fit many replicas in a scalable way, whether on a local machine (using central or graphical processing units) or on a cluster.

The key feature of the `SIMUnet` methodology is the use of a custom *combination layer*, which captures the dependence of the theoretical predictions upon the external parameters $\{c_n\}$, with $n = 1, \ldots, N$, that we fit alongside the PDFs. The edges of the combination layer are fitted simultaneously with the weights and biases associated with the parametrization of the PDFs at the initial scale Q_0.

The approach is represented schematically in figure 6.1, whereby the theory prediction, T, for each experimental observable included in the fit depends on a dynamical choice of $\{c_n\}$. The values of $\{c_n\}$ are associated with the weights of the trainable edges which determine the FK table, Σ, as in equation 6.5. Such dependence enters the theoretical prediction T via the bilinear product between $\Sigma(\{c_n\})$ and the initial scale PDFs, which in equation 6.3 we refer to as L^0, where L^0 indicates either the parametrization of one independent PDF at the initial scale or the product of two of them.

Letting θ denote the set of trainable neural network parameters (the weights and biases) that parameterize the PDFs and $\{c_n\}$ the parameters that we fit alongside the PDFs, `SIMUnet` fits the joint $\hat{\theta} = \theta \cup \{c_n\}$ parameter set, by letting gradient descent determine their optimum value in order to minimize the figure of merit used in the fit, which is defined as

$$\chi^2(\hat{\theta}) = \frac{1}{N_{\text{dat}}}(\mathbf{D} - \mathbf{T}(\hat{\theta}))^T C^{-1}(\mathbf{D} - \mathbf{T}(\hat{\theta})), \tag{6.22}$$

with $\mathbf{D}$ being the vector of experimental central values, $\mathbf{T}$ the vector of theoretical predictions and C the covariance matrix encapsulating the experimental uncertainties and the correlations therein. As in NNPDF4.0, the covariance matrix is constructed using the t_0 prescription in order to avoid the d'Agostini bias [129]; we refer the reader to section 3.2.1 for further details. If one wanted to include also correlated sources of theoretical uncertainties, such as those associated with missing higher order uncertainties in the theory predictions, we could include them using the method outlined in [156, 157, 261]. We leave this endeavour to a future analysis, once the theory covariance matrix for missing higher orders will be available at NNLO.

6.3.2 Parameter fitting using linearisation

In the case of dim-6 operators, discussed in section 6.2, including only the interference of the SMEFT corrections with the SM diagrams is trivial, as we only add a linear dependence upon the Wilson coefficients. Indeed, the identity of equation 6.20 allows us to write the theoretical predictions by `SIMUnet` at a particular configuration $\hat{\theta}$ as

$$T(\hat{\theta}) = \Sigma(\{c_n\}) \cdot\ L^0(\theta) = T^{\text{SM}}(\theta) \cdot \left(1 + \sum_{n=1}^{N} c_n R^{(n)}_{\text{SMEFT}}\right), \tag{6.23}$$

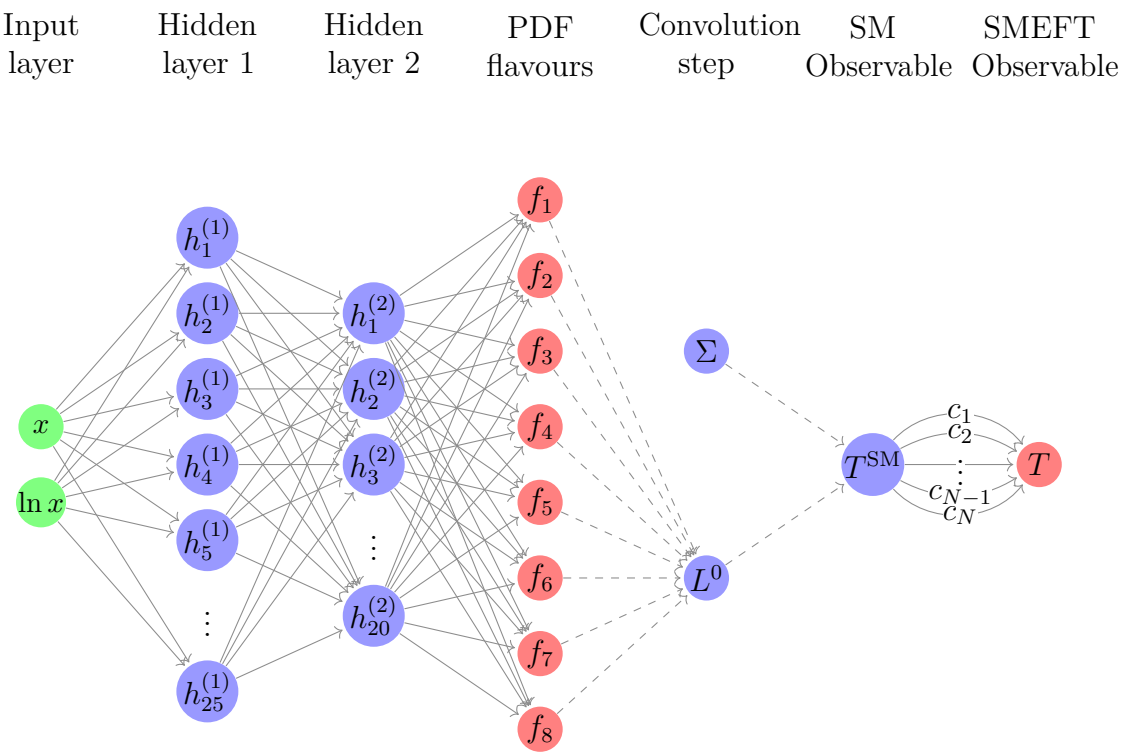


Figure 6.2: Schematic representation of the architecture used by `SIMUnet` in the case of the fit of SMEFT coefficients, in which the dependence of the theoretical prediction T upon the parameters $\{c_n\}$ can be factored into a multiplicative K-factor, as in equation 6.21. The scheme is the same as the one of figure 6.1, however the initial scale PDFs L^0 are first convolved with the relevant SM FK Tables to obtain the Standard Model theory prediction ($T^{\rm SM}$). The SM predictions are then incremented by the addition of the linear SMEFT corrections via a final linear *combination layer*. All solid edges are trainable and thus modified during gradient descent. The precise nature of the manipulation performed by the final layer is outlined in equation 6.23.

where $T^{\rm SM}(\theta) = \Sigma^{\rm SM} \cdot L^0(\theta)$ is the SM theoretical prediction for each observable and corresponds to $\{c_n = 0\}$. Given that in this case the dependence upon the parameters $\{c_n\}$ is factored out of the FK-tables into a multiplicative factor, we can visualize the dependence upon the parameter in the simplified schematic representation given in figure 6.2. Thanks to linearisation, the bracketed term is implemented using a `Keras` custom layer, which takes the usual SM observable predicted by the network at some interim configuration and maps it to a SMEFT modified observable with the strength of the new physics interaction being determined by the weights of the combination layer, which in this case is simply an extra sequential layer that maps $T^{\rm SM}$ into the theoretical prediction T which enters the figure of merit defined in equation 6.22.

The SMEFT-modified theory prediction, T, is then split into disjoint training and validation splits. For each split we compute a training and validation χ^2. Gradient descent attempts to minimize the training χ^2 by descending the loss surface in $\hat{\theta}$-space while the validation χ^2 is monitored to assess the network's out of sample performance. The validation χ^2 decreases initially (since the network is learning the shared underlying laws that are common to both sets), but upon the onset of overfitting, the out of sample performance begins to deteriorate as the network fits the noise of the training data. This particular point in the training process corresponds to the validation χ^2 ceasing to decrease and instead beginning to increase. Upon reaching this regime, training is halted and various checks, such as positivity and integrability of the resulting PDF are assessed. It is at this point that the best fit values of the Wilson coefficients $\{c_n\}$ are obtained, since gradient descent has modified the combination layer's weights such that they best fit the input data. Using this cross-validation procedure [132, 135, 262] we ensure the fitting of statistical fluctuations are avoided as much as possible and it is the shared underlying physical laws that are being fitted instead. At the same time, in order to ensure that the $\{c_n\}$ are not overfitted, we set the datasets that will be modified by these parameters to have a good representation both within the training set and validation set. Thus in this study we split such datasets to have their training and validation fractions equal: $f_{\text{tr}} = f_{\text{val}} = 0.5$, while keeping all other training and validations fractions as in the fits of section 5.4.

As with all deep learning studies, the user has freedom to choose the various hyperparameters of their model at their own discretion. Such parameters include the architecture of the network, the particular choice of initializer that sets the initial values of network parameters $\{\theta\}$, or the choice of minimizer (and the specific settings therein) that tunes these parameters to their optimal values such that the performance metric is minimized. Techniques exist to automate this process such as the `hyperopt` library [263] employing Bayesian methods to perform this optimization. Indeed this is employed by NNPDF4.0 and we choose to use the same settings that were found to be optimal there [1, 2] (see table 3.2). In this study, the hyperoptimization procedure has not been performed to reassess the optimality of the various neural network settings. Indeed, this is justified, since we expect, *a priori* that the PDF modifications due the presence of the aforementioned SMEFT operators will be moderate, and thus the hyperparameters selected assuming a SM-only scenario will remain adequate. However, this assumption does not hold anymore in the event that a more marked PDF modification is expected or obtained *a posteriori*. This point is also worth considering

if one is to introduce a large number (relative to NNPDF4.0) of new measurements or any measurements which introduce tensions with other datasets. Should this be the case, hyperoptimization should be performed again on the layers preceding the combination layer.

The connections to the combination layer (one for each of the c_n) are initialized to zero, since we assume *a priori* the Wilson coefficients will be small; although we observe that if they are initialized according to a normal distribution, then virtually identical results are obtained. These connections are then modified during back propagation. It is worth mentioning that for small values of Wilson coefficients (such as those in this study), it is highly desirable to scale the units such that the Wilson coefficients are $\mathcal{O}(1)$ and thus comparable to the learning rate. This will assist the optimiser during gradient descent to converge upon the minimum in a timely fashion, since the step size will be more suited to the characteristic scale of the Wilson coefficients. In practice, the appropriate scaling is usually determined *a posteriori*, where one can analyse the typical values for the Wilson coefficients and refit with the normalization set accordingly.

6.3.3 Incorporating non-linear effects

The effect of including the SMEFT self-interference diagrams can often introduce a marked effect on the bounds obtained for a given SMEFT scenario [5, 264]. Moreover, with the impetus to produce high precision theoretical predictions for the LHC era, the inclusion of higher order corrections in the SMEFT scattering graphs are becoming particularly pertinent [265–268]. Such considerations introduce a non-linear dependence on the Wilson coefficients in the space of observables: quadratic in the former and quadratic in the highest order in the QCD expansion in the latter. This point serves as a major advantage of our methodology which can accommodate these effects during the fit by the simple addition of non-trainable edges.

When computing the amplitude of a Feynman diagram related to some process one has the schematic form $\mathcal{A} = \mathcal{A}^{\mathrm{SM}} + \sum \mathcal{A}_i$ where $\mathcal{A}_i$ is the amplitude corresponding to the operator $\mathcal{O}^{(i)}$ computed to some order in perturbation theory. We assume here that it is LO in the Wilson coefficients, but need not be in general and the extension to higher orders in the Wilson coefficient expansion is discussed at the end of the present section. When computing the observable, the matrix element, $|\mathcal{A}|^2$, introduces terms of the form $\mathcal{A}_i\mathcal{A}_j$. Since these amplitudes are computed to LO in perturbation theory,

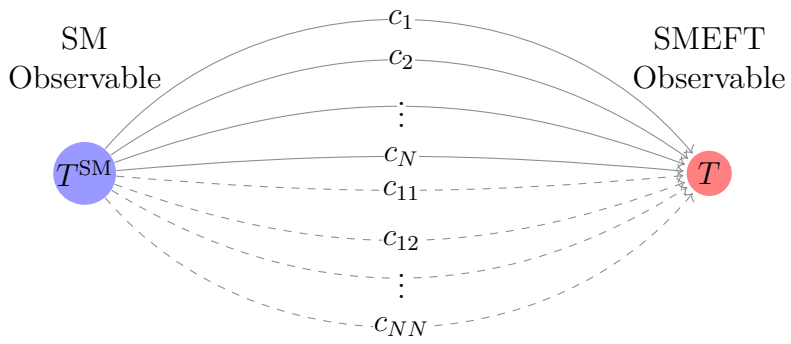


Figure 6.3: Schematic representation of how `SIMUnet` allows for the effects of the SMEFT self-interaction diagrams (dim-6)^2 to be included. We show here how the SM observable can be transformed to a SMEFT observable which includes the $\mathcal{O}(1/v^4)$ terms. The preceding PDF layers are omitted for clarity. The linear contributions are included in the usual way, with the strength of the SMEFT couplings being determined by the trainable edges, c_n: shown by the solid lines. The SMEFT-SMEFT interference contributions are instead non-trainable edges with their value being fixed by the strength of the corresponding pair of trainable edges. These are shown by the dashed lines. There will in general be N trainable edges, and $N(N+1)/2$ non-trainable edges.

we can rewrite equation 6.23 as

$$T(\hat{\theta}) = T^{\mathrm{SM}}(\theta) \cdot \left(1 + \sum_{n=1}^{N} c_n R^{(n)}_{\mathrm{SMEFT}} + \sum_{1 \leq n \leq m \leq N} c_n c_m R^{(n,m)}_{\mathrm{SMEFT}} \, , \right) \tag{6.24}$$

with $R^{(n)}_{\mathrm{SMEFT}}$ ($R^{(n,m)}_{\mathrm{SMEFT}}$) being defined as in equation 5.71 (5.73). Including this contribution in the simultaneous fit is a straightforward task and simply requires that the manipulation performed by the combination layer correspond to that of equation 6.24 instead of equation 6.23.

In practice, however, it is more convenient to rewrite the terms in the right most sum of equation 6.24 by defining $c_{nm} = c_n c_m$:

$$T(\hat{\theta}) = T^{\mathrm{SM}}(\theta) \cdot \left(1 + \sum_{n=1}^{N} c_n R^{(n)}_{\mathrm{SMEFT}} + \sum_{1 \leq n \leq m \leq N} c_{nm} R^{(n,m)}_{\mathrm{SMEFT}} \, , \right) \tag{6.25}$$

we see that both summations are of the same form and so the manipulation required to incorporate the quadratic effects of the squared SMEFT amplitude is reduced to introducing $N(N+1)/2$ additional edges to the combination layer, one for each c_{nm}. These additional connections, however, are not trainable, that is to say their

value is not determined by gradient descent during learning; since c_{nm} is completely determined by the product of trainable edges n and m. This is shown schematically in figure 6.3, whereby the trainable edges determine the non-trainable edges that perform the manipulation corresponding to the right most term in the brackets of equation 6.24. Loops involving SMEFT operator insertions can thus also be fitted in this way alongside the higher dimensional operators in the SMEFT expansion (such as adding dimension-8 operators [109, 110]) that contribute at the same order in the power counting parameter.

Moreover, to incorporate quantum corrections arising from Next-to-Leading Order (NLO) terms in the QCD perturbative expansion of the SMEFT corrections is a similarly straightforward task, simply requiring the computation of the corresponding K-factor and including an additional edge in the combination layer whose value is pinned to be the value of the corresponding trainable edge raised to some power. If one wanted to include also the scale dependence of the Wilson coefficients [269–271], one would have to fit the scale dependent Wilson coefficients at some fixed scale, Q_0, and use the relevant β-functions to evolve the operator to the relevant scale using some pre-computed evolution kernel, which would be factored into the pre-computed FK-tables.

Finally, it is well-known that one can obtain prior knowledge on the Wilson coefficients by assuming a UV completion of the EFT exists that is local, unitary and causal. Such matching conditions can often impose positivity bounds [272] on functions, f_i, of the Wilson coefficients by leveraging these standard conditions on the UV theory. To incorporate such a prior within our framework is analogous to the way positivity on physical cross sections is achieved within the NNPDF methodology [134, 136]. The minimizer is informed of the prior by modifying the training loss function to penalize negative values of Wilson coefficients, with one such choice of penalty term being

$$\text{loss} = \text{loss} + \lambda \sum_i \Theta\Big(- f_i(\{c\})\Big), \tag{6.26}$$

with Θ the usual Heaviside step function which takes unit value for positive arguments and vanishes otherwise. The parameter λ is a non-negative scalar which can be treated as a hyperparameter and encapsulates the degree of belief in the prior: larger values impose the positivity more strictly while smaller values allow for slight violations of the positivity constraints. In a similar way to NNPDF it is also possible to filter the replicas

a posteriori by discarding those that are deemed to violate positivity constraints too strongly.

6.3.4 Fixed PDF analysis

A desirable feature of any simultaneous fitting methodology is to be able to benchmark the simultaneous extraction with the analogous fixed PDF analysis. Indeed, this is precisely what is done in chapter 5 and the broadening of the Wilson coefficients bounds there being the key message of the chapter.

This too is achievable with our methodology and proceeds as follows. We begin by first freezing the combination layer parameters by making them non-trainable. Mathematically, the combination layer thus performs an identity transformation, effectively removing itself from the process. At this stage we are in effect performing a regular NNPDF4.0 PDF fit. Upon successfully achieving the stopping criterion for the PDF only portion of the fit, we then freeze the PDF sector of the neural network (again by making the weights and biases non-trainable) and reinstate the combination layer, allowing for gradient descent to optimize the external parameters, c_i, only. In this way we eliminate the cross-talk between the PDFs and the external parameters and so the external parameters fitted in this way are equivalent to an analogous study keeping the PDFs fixed to the baseline.

6.4 A first simultaneous determination of PDFs and Wilson Coefficients

In this section we present the results we obtain by applying the `SIMUnet` methodology to fit the Wilson coefficients alongside the PDFs in the two benchmark scenarios considered in section 5.4. We describe the results we obtain in the two scenarios, both in terms of the resulting PDFs and bounds on the Wilson coefficients. We stress here that these PDFs are the first to have been truly fitted alongside the Wilson coefficients. We compare our findings to those obtained when PDFs are kept fixed to the SM baseline versus those fitted simultaneously alongside the Wilson coefficients. Finally, we summarize our results, by exploring the correlations between PDFs and the Wilson coefficients that we consider in this study and quantifying the agreement with respect to the previous findings of section 5.4.

We reiterate here that the dataset consider in this work is precisely that of section 5.4.1 and in particular uses high mass Drell-Yan measurements to achieve strong handles on the EFT operators. We shall also include the High-Luminosity projections from section 5.5 when considering the simultaneous fit of (W, Y) and PDFs. The two EFT benchmark scenarios considered are the oblique parameters of section 5.4.2 and the muonphilic operator of section 5.4.3. The ability to accommodate a broader span of Wilson coefficients was not directly assessed in this study, though a detailed analysis of the top sector using top quark sensitive data from the LHC is an on going analysis and reserved as a future publication. In this context the ability for `SIMUnet` to scale with the number of external parameters will be presented. It is, however, well expected that our methodology can accommodate a vastly greater number of Wilson coefficients than just those presented in the present section. This is chiefly due to the fact that the number of experimental measurements available in the global analysis far outweighs the number of external parameters. Even if one is to account for the degrees of freedom in the PDF sector of the neural network architecture, the resulting optimization problem is still well-posed.

6.4.1 Results for Benchmark Scenario I

We first present the results obtained within the first Benchmark Scenario outlined in section 5.4.2, in which the linear effects are dominant as compared to the quadratic effect and, as a result, SMEFT-modified theoretical predictions are defined as in equation 6.23. We start by employing `SIMUnet` to individually constrain the W and Y operators with the ATLAS and CMS neutral-current (NC) high mass DY data from Run I and Run II (table 5.5). In the next subsection we will be able to constrain them simultaneously thanks to the inclusion of charged-current (CC) HL-LHC projections.

In figure 6.4 we show the distribution of the optimal values of W and Y determined for each of the 1000 replicas of the Monte Carlo representation of the experimental data that we obtain during gradient descent in the simultaneous fit of either W or Y and the PDFs. The best-fit values are normally distributed in keeping with the Monte-Carlo pseudodata replica generation that the neural network replicas are fit to. We compare the distribution (in green) with the one that we obtain by keeping the PDFs fixed to the SM baseline (in orange), the latter obtained by using the methodology outlined in section 6.3.4. Both distributions are centred around zero, illustrating that the high-mass DY datasets are compatible with the SM-only hypothesis, but admit non-zero values, with Y being less constrained by the data than W. The distribution

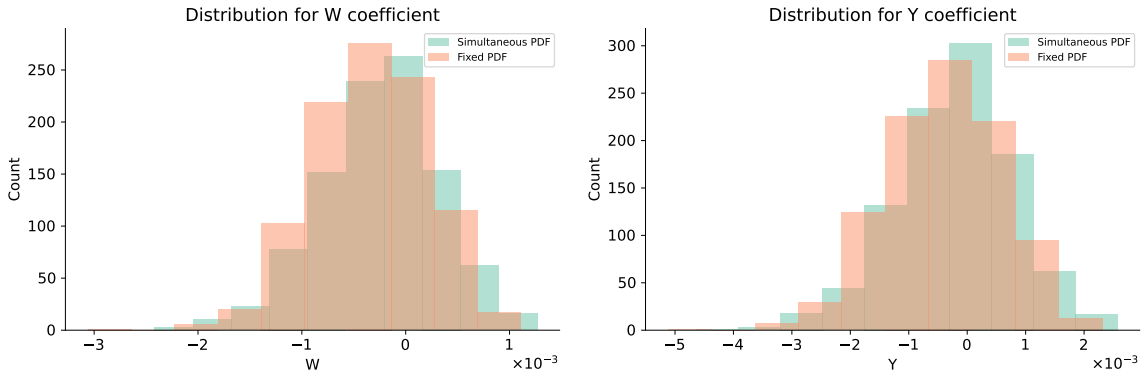


Figure 6.4: Distribution of best fit W (left panel) and Y values (right panel) obtained across 1000 replicas by fitting each of the Wilson coefficients alongside PDFs with the `SIMUnet` methodology (green) compared to the distribution that one would obtain by keeping the PDFs fixed to the SM baseline (orange).

Table 6.1: The 68% and 95% CL bounds on the W and Y parameters obtained either for a fit in which PDFs are kept fixed (SM PDFs) or in a fit in which PDFs are fitted simultaneously with either W or Y (SMEFT PDFs). The fourth and fifth column indicate the absolute shift in best-fit values, equation 5.76, and the percentage broadening of the SMEFT bounds, equation 5.77, when the PDFs are allowed to change alongside the Wilson coefficients.

	SM PDFs	SMEFT PDFs	best-fit shift	broadening
$W \times 10^4$ (68% CL)	$[-9.0, 2.1]$	$[-8.2, 3.2]$	$+1.0$	$+2.6\%$
$W \times 10^4$ (95% CL)	$[-14.5, 7.6]$	$[-13.9, 8.9]$	$+1.0$	$+3.2\%$
$Y \times 10^4$ (68% CL)	$[-13.8, 5.7]$	$[-12.2, 7.9]$	$+1.9$	$+3.1\%$
$Y \times 10^4$ (95% CL)	$[-23.5, 15.5]$	$[-22.2, 18.0]$	$+1.9$	$+3.1\%$

of best fits obtained in a simultaneous fit looks similar to the distribution of best fits obtained by keeping the PDFs fixed to the SM baseline.

For a more quantitative comparison, in table 6.1 we compare the bounds for the individual W and Y obtained in the simultaneous fits to those obtained by keeping the PDFs fixed to the SM baseline. We indicate the shift and the broadening of the bounds that are obtained once the Wilson coefficients are fitted alongside PDFs as in equations 5.76 and 5.77 respectively. As in section 5.4, the effect of the back-reaction on the PDFs induced by the Wilson coefficients on the interpretation of high-mass DY constraints is entirely moderate, with the bounds of the simultaneous fit being only slightly looser (by a factor around 3%) than those obtained by keeping the PDFs fixed.

In figure 6.5 we show the quark-antiquark channel luminosity plots defined in equation 5.72 with the error bands showing 68% confidence levels, normalised to the

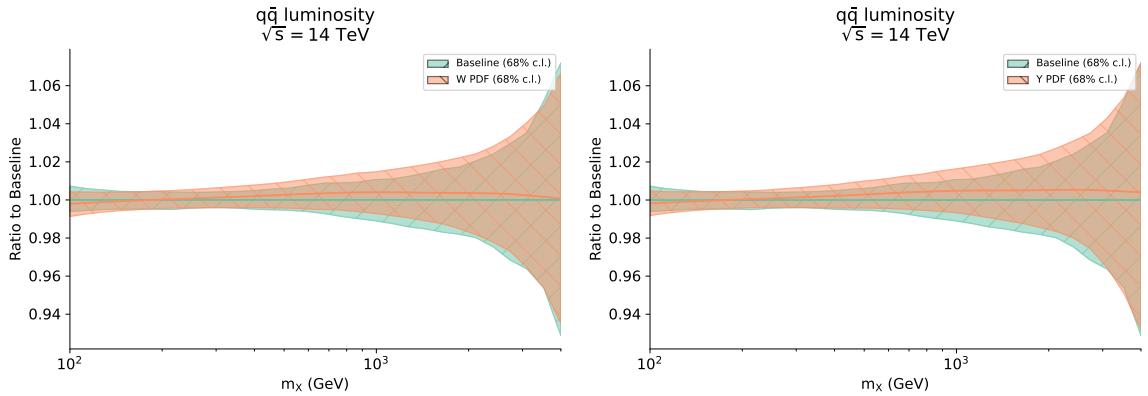


Figure 6.5: The one dimensional luminosity in the $q\bar{q}$ channel for a PDF fitted in the presence of a non-zero W operator (left panel) or of a non-zero Y operator (right panel) shown in orange, normalized to the SM baseline PDF shown in green.

baseline SM PDFs. We notice that, while in the previous analysis of section 5.4, we could only produce sets of PDFs obtained at fixed values of W and Y (corresponding to the Benchmark Points that were taken under consideration), here we do really produce a set of PDFs obtained out of a simultaneous fit alongside the Wilson coefficients. We see the luminosity modification due to the simultaneous fit remains moderate, with a slightly larger deviation found at higher values of the produced lepton invariant mass, where the PDFs exhibit slightly larger uncertainties. This is indeed a result consistent with the indications given in section 5.4. Indeed, the dominant luminosity channel for NC DY is $u\bar{u}$ and $d\bar{d}$ with the valence quark distribution being strongly constrained by DIS and the SMEFT modification being small relative to the experimental uncertainties of the highest invariant mass bins probed by current experimental data. Again, this finding shows that the interplay between EFT effects and PDFs remains moderate when one performs a truly simultaneous determination of both.

It is interesting to observe that, if we try to fit W and Y at the same time using only the current NC DY data, we identify both the flat direction and the strong anti-correlation between W and Y, which are known to exist in this case [249]. Results are shown in figure 6.6. Both the flat direction and the anti-correlation can be retrieved within the framework of the `SIMUnet` methodology, without the user having to be aware of the existence or particular nature of any flat directions which may exist. Indeed, the optimizer cannot preferentially differentiate one point from another within this landscape valley and so the points are distributed tightly along the flat direction. The true minimum is thus a soft attractor in this case. The preference for the upper left quadrant in the W-Y plane is consistent with the findings of [249] with the failure

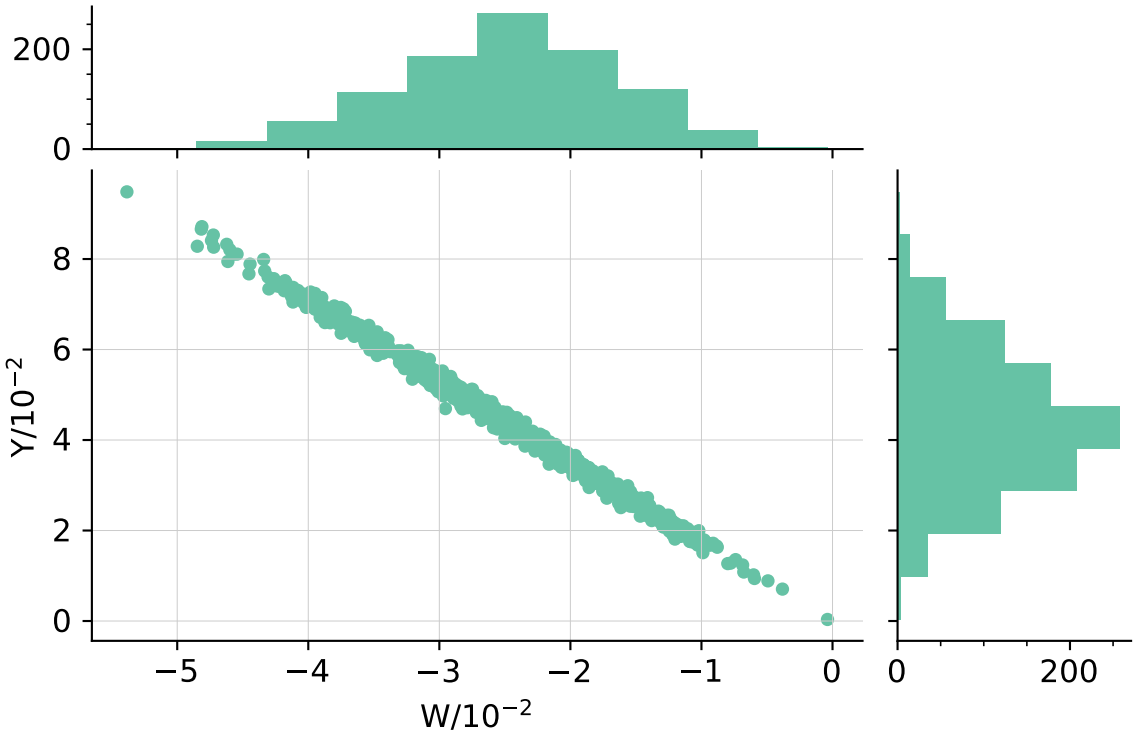


Figure 6.6: Scatter plot for the best fit values of (W, Y) per replica using ATLAS and CMS high mass Drell-Yan data, which are exclusively NC observables. The upper and rightmost panels are histograms in their respective directions. A clear flat direction has been detected along with a strong anti-correlation.

to capture the origin possibly due to the lack of inclusion of $\mathcal{O}(1/\Lambda^4)$ terms in the partonic cross section which is known to ease such tensions when only $\mathcal{O}(1/\Lambda^2)$ terms are present, as it was pointed out in reference [264], where the effect of dim-8 operator is carefully assessed. While easily spotted in a two operator analysis by simply plotting the best fit values of Wilson coefficients, it is significantly more challenging to notice a flat direction in a many operator scenario. A plot like figure 6.6 can no longer be made in higher-dimensions and in such a case it is wise to perform a principle component analysis (PCA) to assess directions of high variance in best fit values [273, 274].

6.4.2 Inclusion of the HL-LHC projections

The flat direction illustrated in figure 6.6 can be eliminated with the inclusion of Charged Current (CC) DY data as we explicitly demonstrated in section 5.5. No unfolded measurements of the high-mass transverse mass m_T distribution have been yet released at 13 TeV, thus we will base our analysis on the High Luminosity LHC

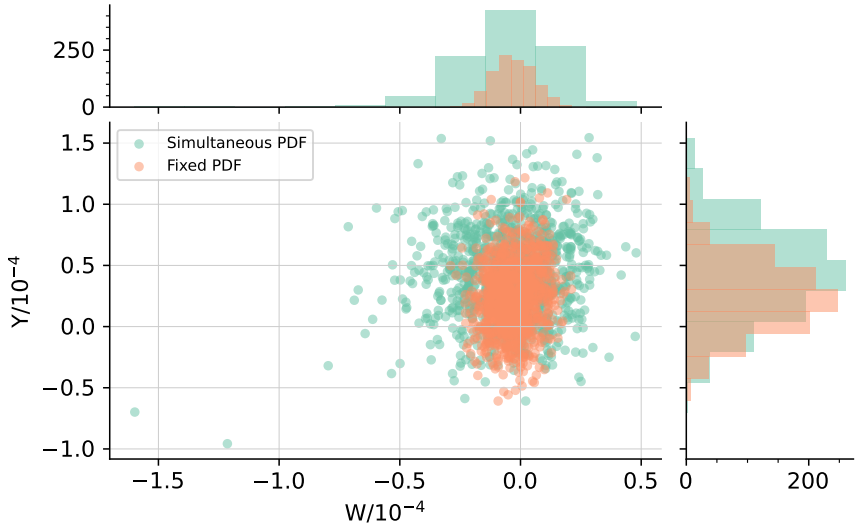


Figure 6.7: Scatter plot for best fit tuples of (W, Y) for each replica obtained in the simultaneous fit (green) compared to those obtained when PDFs are kept fixed to the SM baseline (orange). The upper and rightmost panels are histograms in their respective directions.

(HL-LHC) high-mass Drell-Yan projections that we used in chapter 5; inspired by the HL-LHC projections studied in [253]. We perform a simultaneous fit of the PDFs and the two (W, Y) parameters by adding two trainable edges in the combination layer displayed in figure 6.2 and appending the aforementioned HL-LHC projected data to the already present DIS and Drell-Yan data. The best fit values of (W, Y) obtained for each of the 1000 replicas of the Monte Carlo ensemble are plotted in figure 6.7. We see that not only is the flat direction of figure 6.6 broken, but the HL-LHC projections heavily favour vanishing Wilson coefficients. Indeed, the origin is now covered in both the W and Y axes and much more heavily constrained, enjoying roughly two orders of magnitude tighter constraints in both directions. We notice a slightly favorable pull towards the upper-left quadrant in the W-Y plane, that the ATLAS and CMS datasets seem to prefer. Comparing the best-fit distribution that we get in a simultaneous fit (displayed in green) to the one we get in a fit in which PDFs are kept fixed to the SM baseline (displayed in orange) we can see that the bounds are visibly tighter once PDFs are kept fixed to the SM baseline and are not allowed to consistently vary alongside the Wilson coefficients. This is precisely the same behaviour as figure 5.15.

In table 6.2 we compare the bounds for the individual W and Y fits obtained in the simultaneous fits to those obtained by keeping the PDFs fixed to the SM baseline. We indicate the shift and the broadening of the bounds according to the definitions given

Table 6.2: Same as table 6.1 for the 68% C.L. and 95% C.L. marginalized bounds on the W and Y parameters obtained from the two-dimensional (W, Y) fits that include the HL-LHC pseudo-data for NC and CC Drell-Yan distributions.

	SM PDFs	SMEFT PDFs	best-fit shift	broadening
$W \times 10^5$ (68% CL)	$[-1.1, 0.5]$	$[-2.4, 1.5]$	-0.2	$+144\%$
$W \times 10^5$ (95% CL)	$[-2.0, 1.4]$	$[-4.3, 3.4]$	-0.2	$+126\%$
$Y \times 10^5$ (68% CL)	$[-0.4, 5.2]$	$[0.6, 8.0]$	$+1.9$	$+32\%$
$Y \times 10^5$ (95% CL)	$[-3.2, 8.1]$	$[-3.1, 11.7]$	$+1.9$	$+31\%$

in equations 5.76 and 5.77. Consistently to what was found in our previous study of section 5.5 we observe that including high-mass data at the LHC both in a fit of PDFs and in a fit of SMEFT coefficients and neglecting the interplay between them could result in a significant underestimate of the uncertainties associated to the SMEFT parameters. Indeed, the marginalized bounds on the $W(Y)$ parameter increase by about 150% (30%) once a simultaneous fit of the PDFs and the (W, Y) parameters is performed. The broadening is smaller than observed in section 5.5, but it is still significant. A detailed comparison is given in section 6.4.4.

As far as the quark-antiquark luminosity is concerned, we can see in figure 6.8 that once the PDFs are fitted simultaneously with the (W, Y) parameters two things happen. First of all the central values of the luminosity shift upwards for large values of the invariant mass ($M_X \gtrsim 1$ TeV) towards the edge of the 68% C.L. error band. Second the error band significantly increases. The shift in the central value is compatible to what we observed before, namely that the luminosity plots, once PDFs are fitted at some representative values of the W and Y parameters, do change significantly, well outside the 1σ error band of the SM PDFs, while the PDF uncertainties themselves are unchanged (figure 5.16). However, in this case the PDF uncertainty does in fact increase since here we are actually performing a simultaneous fit and, as a result, the PDF error band increases proportionally to the width of the range of W and Y that the data allow. This is a very interesting result and shows that PDF error bands at large-x inherently have an extra source of theory uncertainty related with possible BSM effects that the data do not exclude.

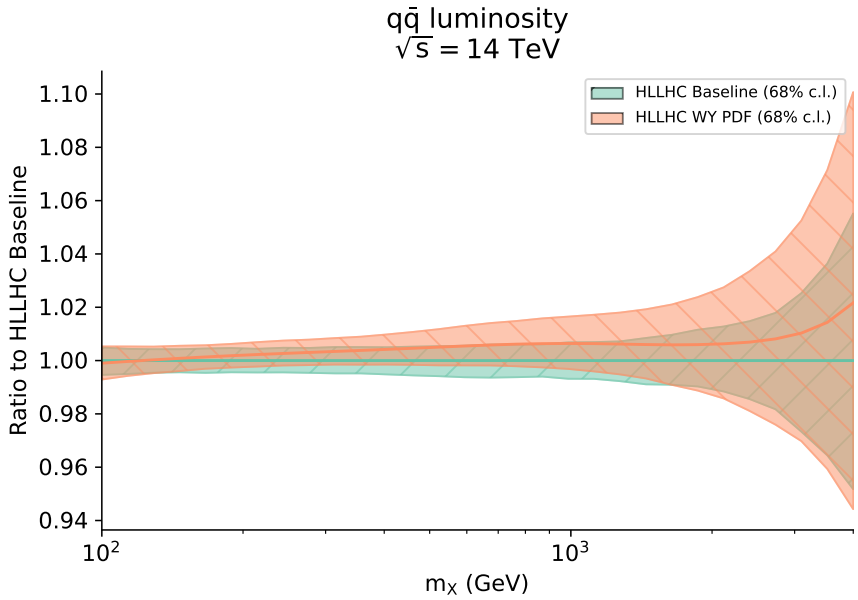


Figure 6.8: The $q\bar{q}$ luminosity channel of the PDF fitted in the presence of W and Y parameters (orange) fitted to the ATLAS and CMS high mass Drell-Yan data as well as the NC and CC DY HL-LHC projections, normalized to the appropriate baseline SM PDF (green).

6.4.3 Results for Benchmark Scenario II

In this section we employ the left-handed muon-philic operator $\mathbf{C}_{33}^{D\mu}$ to showcase our methodology's ability to constrain Wilson coefficients whilst accounting for the effect of quadratic dim-6 effects using the approach discussed in section 6.3.3. The fact that this second Benchmark SMEFT scenario is effectively unconstrained [5] at the linear level serves to act as the ideal setting to assess the ability to fit EFT operators whilst simultaneously accounting for their quadratic contributions, as in equation 6.24. Furthermore, since the $\mathbf{C}_{33}^{D\mu}$ operator affects only muon final state observables, while electron final states are described by the SM, the combination layer uses only those datasets that have a muon in the final state to constrain $\mathbf{C}_{33}^{D\mu}$. In particular, the CMS high-mass measurements at 13 TeV, which up until now have been for the combined decay channel, is again separated into the electron and muon channels. As a result of splitting this particular dataset into separate channels, we have accordingly generated a new baseline PDF used in the comparisons.

The best-fit values of $\mathbf{C}_{33}^{D\mu}$ across the 1000 replicas in the fit are shown on the right-hand panel of figure 6.9. We compare the distribution obtained out of a simultaneous fit (green) with the one obtained when PDFs are kept fixed to the SM baseline (orange).

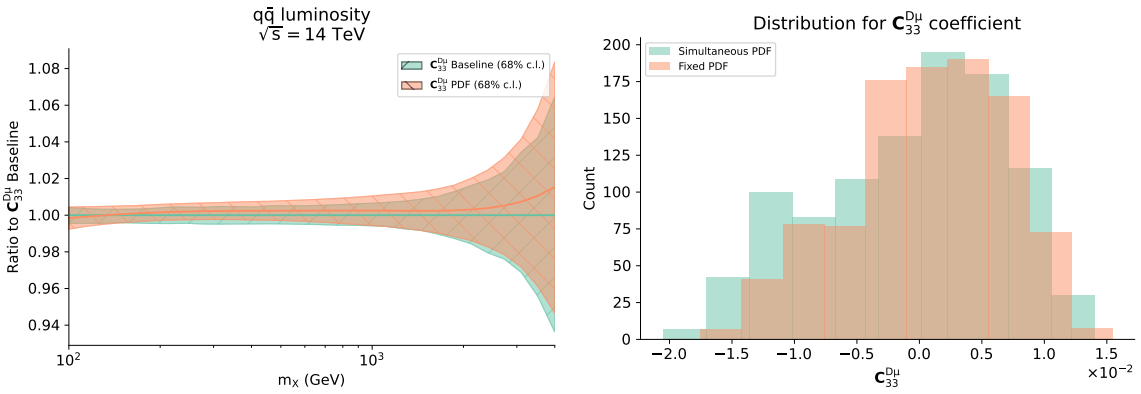


Figure 6.9: Left: $q\bar{q}$ luminosity channel of the PDF fitted in the presence of the $\mathbf{C}_{33}^{D\mu}$ parameter normalized to the baseline SM PDF. Right: histogram plot for best fit values for $\mathbf{C}_{33}^{D\mu}$ for each replica. The best-fit distribution over 1000 replicas obtained out of a simultaneous fit (green) is compared to the one obtained by keeping PDFs fixed to the SM baseline (orange).

Table 6.3: Same as Table 6.1 for the 68% CL and 95% CL bounds on the $\mathbf{C}_{33}^{D\mu}$ Wilson Coefficient, including both linear and quadratic terms in the SMEFT expansion.

	SM PDFs	SMEFT PDFs	best-fit shift	broadening
$\mathbf{C}_{33}^{D\mu} \times 10^3$ (68% CL)	$[-5.6, 6.9]$	$[-8.0, 6.7]$	-0.9	$+18\%$
$\mathbf{C}_{33}^{D\mu} \times 10^3$ (95% CL)	$[-11.9, 13.1]$	$[-15.3, 14.0]$	-0.9	$+17\%$

We see that the distribution of best fit values is centred at the origin, although it is skewed towards $\mathbf{C}_{33}^{D\mu} \approx 5 \times 10^{-3}$. The shape of the distribution is what we expect once quadratic terms are allowed in the fit of the Wilson coefficients by keeping PDFs fixed [44], and it is interesting to see this feature not only holds but it is actually enhanced once the Wilson coefficient is fitted alongside PDFs (green histogram).

For a quantitative comparison of the bounds obtained in the simultaneous fit to those obtained in a Wilson coefficient-only fit, in table 6.3 we show the bounds that we obtain in the two cases. The interplay between PDFs and SMEFT coefficients is quite moderate in this particular scenario. In contrast with the marked effects in Benchmark Scenario I; in Benchmark Scenario II the obtained bounds on this Wilson coefficient would loosen by around 20%. The origin of this rather different behaviour can be traced back to the fact that in this scenario the electron channel data do not receive EFT corrections, and hence all the information that they provide makes it possible to exclusively constrain the PDFs. The muon channel distributions then determine the allowed range for the $\mathbf{C}_{33}^{D\mu}$, restricted by the well-constrained large-x quarks

and antiquark PDFs from the electron data. This claim is backed by the luminosity comparison, displayed on the left-hand panel of figure 6.9: the shift and the increase in the PDF uncertainty of the $q\bar{q}$ luminosity are visible, but less enhanced than in the first Benchmark Scenario. Again, our findings here are corroborated by the discussion of section 5.5.2 where we found that even at the HL-LHC the broadening of $\mathbf{C}_{33}^{D\mu}$ remains moderate owing to the relative lack of measurements that differentiate between lepton flavours.

6.4.4 Results overview

To conclude this section, we present an overview of our results. We have seen in section 6.4.1 that the current high-mass Drell-Yan data from LHC Run I and Run II do not allow to simultaneously fit W and Y in Benchmark Scenario I and loosely constrain $\mathbf{C}_{33}^{D\mu}$ in Benchmark Scenario II. This is mostly due to the lack of unfolded CC Drell-Yan data, which would remove the flat direction that our algorithm is able to detect (see figure 6.6). Moreover, our analysis confirms what was outlined in section 5.4 namely that the interplay between the individual fits of W and Y and the fit of the large-x quark distributions is mild at the level of current DY data. However, once the high-statistics data projections from the HL-LHC are included, the flat direction disappears and one is able to obtain strong constraints both on the (W, Y) plane and on the individual $\mathbf{C}_{33}^{D\mu}$ coefficient. From the point of view of showcasing our methodology, the two scenarios are interesting, as in the Benchmark Scenario I, once the HL-LHC projections are included, we can simultaneously fit both W and Y alongside the PDFs, including only the linear SMEFT corrections while in the Benchmark Scenario II we can fit individually $\mathbf{C}_{33}^{D\mu}$ alongside the PDFs, including both the linear and the quadratic SMEFT corrections. In the first scenario, we observe that there is a strong interplay between SMEFT and PDF fits, as the bounds for the SMEFT coefficients significantly broaden once PDFs are allowed to vary alongside W and Y (see figure 6.7) and the PDFs themselves display a sizeable shift (see figure 6.8). The interplay is more moderate in the second scenario, given that BSM effects only affect the data with muons in the final states, while the data with electrons in the final states constrain the large-x quark distributions.

In this section, we focus on the results obtained including the HL-LHC projections. We first explore the correlation patterns between the PDFs and the SMEFT coefficients in both Benchmark Scenarios. These correlation coefficients can be evaluated as in [4]. For example in the case of the gluon and the W Wilson coefficient, the correlation

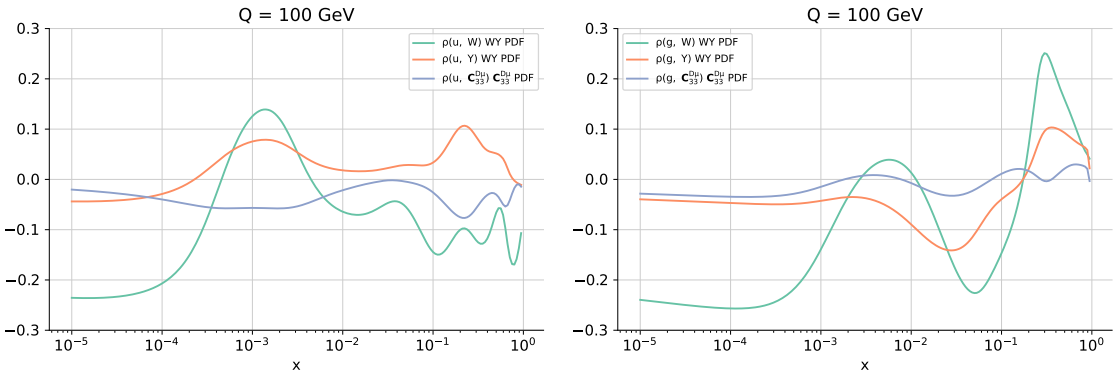


Figure 6.10: Correlation coefficients, defined in equation 6.27, between the W (green), Y (orange), $\mathbf{C}_{33}^{D\mu}$ (blue) and the up quark PDF (left panel) and gluon PDF (right panel) computed at $Q = 100$ GeV as a function of the momentum fraction x. The PDFs used to compute the correlation coefficients are those obtained in the simultaneous fit of (W, Y) and the PDFs described in section 6.4.2 (in the case of the green and orange curves) and those obtained in the simultaneous fit of $\mathbf{C}_{33}^{D\mu}$ and the PDFs described in section 6.4.3 (in the case of the blue curve).

coefficient is defined as follows

$$\rho\left(W, g(x,Q)\right) = \frac{\left\langle W^{(\text{best}-\text{fit})} g(x,Q)\right\rangle - \left\langle W^{(\text{best}-\text{fit})}\right\rangle \left\langle g(x,Q)\right\rangle}{\sqrt{\left\langle W^{(\text{best}-\text{fit})2}\right\rangle - \left\langle W^{(\text{best}-\text{fit})}\right\rangle^2}\sqrt{\left\langle g(x,Q)^2\right\rangle - \left\langle g(x,Q)\right\rangle^2}}\,, \tag{6.27}$$

where $W^{(\text{best}-\text{fit})}$ is the best-fit of the W coefficient for each replica in the simultaneous fit, in which PDFs are allowed to vary alongside W and Y. In equation 6.27, averages are computed over the $N_{\text{rep}} = 1000$ replicas. We show in appendix D, that even restricting to $N_{\text{rep}} = 100$ replicas, as is usually done in PDF fitting, then virtually identical results are obtained. This correlation coefficient provides a measure of how the variations in the PDFs translate into modifications of the best fit value of the Wilson coefficients. In figure 6.10 we show these correlation coefficients between the up quark distribution and the gluon PDFs at $Q = 100$ GeV. Each of the curves corresponds to one of Wilson coefficients considered in the HL-LHC analysis, the W and Y being fitted simultaneously alongside the PDFs (hence the "WY PDF" label) and the $\mathbf{C}_{33}^{D\mu}$ being fitted individually alongside the PDFs (hence the "$\mathbf{C}_{33}^{D\mu}$ PDF" label) . Although correlations are moderate, it is interesting to observe that the correlation/anticorrelation between W and the PDFs is much stronger than the correlation with the other Wilson coefficients. This explains why the broadening of the bounds in the W direction are more marked than those in the Y directions.

Throughout this study we found that the results obtained with the `SIMUnet` methodology are in line with those presented in section 5.4. In table 6.4 we make this comparison more quantitative, by focussing on the results in which the effects of the interplay between the SMEFT coefficients and the PDFs are more visible, namely the fits obtained by using the NC and CC projections from the HL-LHC. The results are comparable, although the effect of fitting the Wilson coefficients along with the PDFs is more moderate. This is not surprising, as there are two crucial differences in the two analyses. On the one hand the PDF set that we use here in the fixed SM PDF case is different compared to the one we used in the previous analysis, being based on the same dataset as before but on the NNPDF4.0 methodology rather than the NNPDF3.1 methodology. Secondly, because the previous methodology was based on the use of Benchmark Points in the Wilson coefficients parameter space, we determined the bounds on the parameters by using the partial χ^2 including only the data affected by the SMEFT corrections, rather than the global χ^2. Recall this approximation was forced because the statistical fluctuations of the global χ^2 were found to be significantly larger than those of the partial χ^2 and could only be tamed by running a very large batch of replicas for each Benchmark Point and by increasing the density of Benchmark Points in the region that is explored. This approximation is no longer necessary within `SIMUnet`, because we no longer rely on the interpolation over Benchmark Points, rather we perform a truly simultaneous fit of the PDFs and the Wilson coefficients based on the global χ^2. Additionally, the minimizer used in this study employs a momentum driven stochastic gradient descent based algorithm (Nesterov-accelerated adaptive moment estimation [124, 125] available from the `Keras` library), while in section 5.4 we employed a more traditional genetic algorithm approach using legacy in-house implementations. As such, one in general expects to achieve an improved fit quality with our methodology. Finally, the bounds quoted in our study are defined using Monte Carlo based statistical estimators, whereas in the latter approach, the geometry of the χ^2 profile is used to define bounds on the Wilson coefficients. Notwithstanding, we observe that the results and the trends are consistent with each other, although the use of the partial χ^2 over-emphasizes the broadening of the bounds.

Results in the HL-LHC scenario are displayed in figure 6.11. Shown are the results obtained including the current high-mass Drell-Yan data and the projected HL-LHC NC and CC Drell-Yan pseudo-data. In Scenario I, W and Y are fitted simultaneously by keeping only the linear term in the SMEFT expansion, while in Scenario II the $\mathbf{C}_{33}^{Dμ}$ coefficients is fitted individually including the $\mathcal{O}(1/\Lambda^4)$ quadratic terms in the SMEFT

Table 6.4: 95% CL bounds on the simultaneous fit of the W and Y Wilson coefficients in Benchmark Scenario I and of the individual fit of the $\mathbf{C}_{33}^{D\mu}$ Wilson Coefficient in Benchmark Scenario II, based on a fit including the HL-LHC projections, compared to those obtained in the previous analysis presented in section 5.4. The fourth and fifth column indicate the absolute shift in best-fit values, equation 5.76 and the percentage broadening of the SMEFT bounds, equation 5.77, when the PDFs are allowed to change alongside the Wilson coefficients.

	SM PDFs	SMEFT PDFs	best-fit shift	broadening
$W \times 10^5$ (this work)	$[-2.0, 1.4]$	$[-4.3, 3.4]$	-0.2	$+126\%$
$W \times 10^5$ section 5.4	$[-1.4, 1.2]$	$[-8.1, 10.6]$	-1.4	$+620\%$
$Y \times 10^5$(this work)	$[-3.2, 8.1]$	$[-3.1, 11.7]$	$+1.9$	$+31\%$
$Y \times 10^5$ section 5.4	$[-5.3, 6.3]$	$[-11.1, 12.6]$	$+0.3$	$+110\%$
$\mathbf{C}_{33}^{D\mu} \times 10^3$ (this work)	$[-11.9, 13.1]$	$[-15.3, 14.0]$	-0.9	$+17\%$
$\mathbf{C}_{33}^{D\mu} \times 10^3$ section 5.4	$[-10.4, 12.3]$	$[-12.5, 14.6]$	-0.6	$+18\%$

expansion. The results from both studies are compatible, although the bounds obtained from this study are slightly more conservative with differences being explained by the differences between the methodologies employed that are outlined in this section.

6.5 Fit quality

We now move on to discuss the fit quality of not only the final PDF sets, but also the contribution arising from the best-fit Wilson coefficients. For each MC replica there is the pair $(\mathbf{f}_i, \mathbf{c}_i)$ which are the best fit PDFs and Wilson coefficients respectively. Together they can be used to generate the corresponding theory predictions for each dataset, noting that any datasets which were not modified by the SMEFT operators will effectively have $\mathbf{c} = \mathbf{0}$. We thus have an ensemble of N_{rep} vectors of theory predictions, $\mathbf{T}$, with the central theory prediction being given by the average across replicas:

$$\langle \mathbf{T} \rangle = \frac{1}{N_{\text{rep}}} \sum_{i=1}^{N_{\text{rep}}} \mathbf{T}(\mathbf{f}_i, \mathbf{c}_i). \tag{6.28}$$

The central χ^2 per data point is then computed in the usual way

$$\chi^2 = \frac{1}{N_{\text{dat}}} (\mathbf{d} - \langle \mathbf{T} \rangle)^T C^{-1} (\mathbf{d} - \langle \mathbf{T} \rangle) \tag{6.29}$$

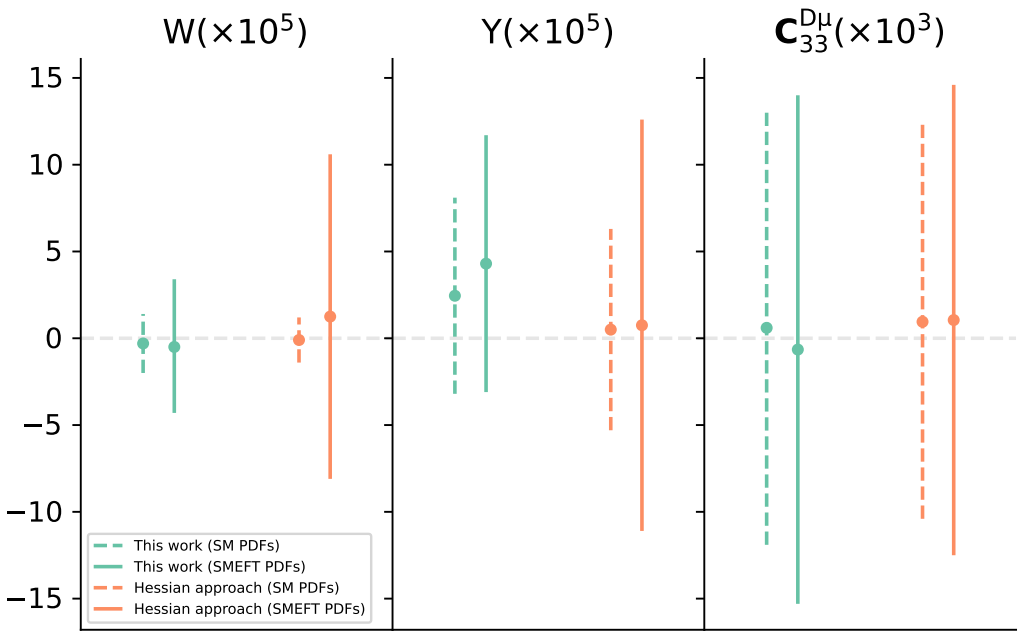


Figure 6.11: Comparison of 95% CL bounds between this study (green) and those determined using the Hessian approach of section 5.4 (orange). The W and Y bounds are determined simultaneously in the HL-LHC scenario, $\mathbf{C}_{33}^{D\mu}$ is fitted individually including SMEFT quadratic terms in the HL-LHC scenario. The bounds obtained by keeping the PDFs fixed to the SM baseline (dashed lines) are compared to those obtained in a simultaneous fit of PDFs and Wilson coefficients (solid lines).

with $\mathbf{d}$ being the vector of experimental central values and C the covariance matrix encapsulating the experimental uncertainties and the correlations therein. These values are tabulated in table 6.5 for each of the various SMEFT scenarios considered in this work. We also tabulate the χ^2 for various groupings of these datasets, such as DIS only, including or excluding the high-mass DY measurements etc. For these particular entries various correlated systematics may exist between datasets, such as the uncertainty in beam luminosity, which introduces off-diagonal entries in the covariance matrix; as such the grouped χ^2 is not necessarily equal to the weighted average of the individual χ^2 values constituting the grouping. This effect is particularly marked for the HL-LHC entries.

In all scenarios considered, the added degrees of freedom result in the χ^2 per data point to drop when a simultaneous determination is performed when contrasted to the purely Standard Model fits. The SM χ^2 in this sense serves as an upper bound, since the optimizer is free to determine $\mathbf{c} = 0$ and so the goodness of fit can be no worse than the SM fit. In particular we see the SMEFT sensitive high-mass DY

Table 6.5: Values of the χ^2 per data point across all datasets used in this study. We tabulate values for the baseline PDF set as well as those obtained in the various SMEFT scenarios. Shown also is the χ^2 for the HL-LHC scenario. The rows indicating total χ^2 values are computed accounting for any relevant correlated systematic errors. Values in italics indicate the dataset was not used in the corresponding fit.

Dataset	n_{data}	χ^2/n_{data}				
		SM Baseline	W	Y	HL-LHC SM Baseline	HL-LHC (W,Y)
SLAC	67	0.866	0.855	0.854	0.837	0.850
BCDMS	581	1.285	1.265	1.265	1.294	1.266
NMC	325	1.320	1.318	1.318	1.333	1.320
CHORUS	832	1.208	1.209	1.210	1.206	1.208
NuTeV	76	0.444	0.486	0.487	0.472	0.497
HERA inclusive	1145	1.188	1.183	1.183	1.190	1.184
HERA charm	37	1.435	1.384	1.382	1.463	1.391
HERA bottom	29	1.113	1.110	1.110	1.118	1.110
Total DIS	**3092**	**1.202**	**1.197**	**1.197**	**1.206**	**1.198**
E886 $\sigma^d_{\text{DY}}/\sigma^p_{DY}$	15	0.669	0.609	0.600	0.974	0.679
E886 σ^p_{DY}	89	1.572	1.593	1.604	1.566	1.631
E605 σ^p_{DY}	85	1.197	1.200	1.205	1.197	1.213
CDF $d\sigma_Z/dy_Z$	29	1.613	1.546	1.548	1.623	1.563
D0 $d\sigma_Z/dy_Z$	28	0.612	0.610	0.610	0.613	0.613
D0 $W \to \mu\nu$ asy.	9	1.845	1.510	1.503	2.052	1.587
ATLAS W, Z 2010	30	1.021	1.014	1.014	1.017	1.017
ATLAS low-mass $Z \to ee$	6	0.923	0.921	0.921	0.923	0.921
ATLAS W, Z 2011 CC	46	2.095	2.005	2.006	2.091	2.010
ATLAS W, Z 2011 CF	15	1.062	1.072	1.073	1.066	1.071
ATLAS $W + c$ rapidity	22	0.453	0.463	0.463	0.447	0.457
ATLAS Z_{p_T}	92	0.959	0.938	0.936	0.939	0.928
ATLAS W_{p_T} jets	32	1.685	1.672	1.670	1.676	1.665
CMS W e asy.	11	0.785	0.790	0.789	0.815	0.804
CMS W μ asy.	11	1.767	1.732	1.733	1.765	1.732
CMS σ_{W+c} 7 TeV	5	0.513	0.518	0.516	0.502	0.504
CMS $\sigma_{W^++c}/\sigma_{W^-+c}$ 7 TeV	5	1.822	1.791	1.796	1.884	1.848
CMS Z_{p_T}	28	1.303	1.312	1.311	1.287	1.306
CMS $W \to \mu\nu$ rapidity	22	1.472	1.337	1.340	1.422	1.310
CMS $W + c$ rapidity 13 TeV	5	0.719	0.722	0.721	0.712	0.711
LHCb $Z \to \mu\mu$	9	1.503	1.545	1.550	1.506	1.549
LHCb $W, Z \to \mu$ 7 TeV	29	2.043	1.973	1.977	2.066	2.005
LHCb $W, Z \to ee$	17	1.249	1.236	1.236	1.220	1.231
LHCb $W, Z \to \mu$ 8 TeV	30	1.621	1.497	1.502	1.615	1.543
Total DY (excl. HM)	**670**	**1.302**	**1.274**	**1.276**	**1.307**	**1.286**
ATLAS DY high-mass 7 TeV	13	1.680	1.575	1.609	1.654	1.626
ATLAS DY high-mass 8 TeV	46	1.174	1.177	1.171	1.175	1.171
CMS DY high-mass 7 TeV	117	1.694	1.671	1.676	1.677	1.669
CMS DY high-mass 8 TeV	41	0.923	0.944	0.941	0.893	0.914
CMS DY high-mass 13 TeV	43	2.003	2.064	2.037	2.000	2.005
Total DY (HM only)	**260**	**1.531**	**1.529**	**1.527**	**1.517**	**1.515**
Total (excl. HL-LHC)	**4022**	**1.245**	**1.236**	**1.237**	**1.248**	**1.238**
HL-LHC CC e	16	*1.119*	*119.9*	*0.922*	0.588	0.544
HL-LHC CC μ	16	*1.414*	*112.8*	*1.162*	0.894	0.803
HL-LHC NC e	12	*1.164*	*17.65*	*7.495*	1.104	0.961
HL-LHC NC μ	12	*1.041*	*13.32*	*5.048*	0.964	1.071
Total HL-LHC only	**56**	***1.298***	***72.88***	***5.546***	**0.894**	**0.836**
Total	**4078**	***1.246***	***2.220***	***1.296***	**1.243**	**1.232**

datasets experience a large overall improvement in fit quality, largely driven by the CMS measurements at 7 TeV owing to the large weight carried by this dataset: forming just under half the entire high-mass DY data points considered in this study. Moreover, the DIS only grouping experiences a marked improvement in fit quality with the χ^2 per data point dropping by 0.014 across 3092 data points in the case of the simultaneous (W, Y) determination. The reader is again reminded that the HERA combined dataset, which forms the majority of the DIS data points, is included in the set of datasets that are modified by the W and Y operators. It is interesting to observe that the improvement in fit quality is propagated down to those datasets, such as the low-mass measurements, that are not used to explicitly constrain the Wilson coefficients. Such datasets, however, are correlated through shared sources of systematic errors, such as the luminosity uncertainty or detector effects, thus improving the fit to one such dataset necessarily affects others. Also tabulated are the fit quality values for the HL-LHC projections, even for those fits which do not incorporate these projections in their training data. These entries illustrate the pull the HL-LHC projections have on the Wilson coefficients. We see that not including these data points in the fit renders the fit virtually useless in the context of the projected data. This is indeed reflected in the values of the K-factors used for these projections, reaching $K \simeq 5$ for the highest transverse mass bins.

The χ^2 values for the muonphilic operator, $\mathbf{C}_{33}^{D\mu}$, are tabulated in table 6.6 where the quadratic effects of the EFT operator are included both in the fit and in the χ^2 calculation. The story is much the same here, albeit with the improvement across the high mass Drell-Yan measurements being rather mild; the reason being again due to the lack of measurements differentiating between the dielectron and dimuon channel. Here, all datasets are improved upon, except for the 8 TeV measurement from CMS. We see that despite the total high-mass Drell-Yan measurements being slightly less well explained by the simultaneous fit, the overall dataset is indeed improved upon thanks again to the strong weight carried by the HL-LHC projections.

6.6 Methodology validation and closure testing

In this section, we assess the robustness of our approach through the closure testing framework [219, 257]. Here we do not address the robustness of the PDF part of the fit, which in a sense comes from the results presented in the NNPDF4.0 release [1] and

Table 6.6: Values of the χ^2 per data point across all high mass Drell-Yan measurements for the muonphilic operator. We tabulate values for the baseline PDF set as well as those obtained in the EFT scenario II. Shown also is the χ^2 for the HL-LHC projections. Quadratic terms in the EFT parameter are used in the χ^2 calculation. The rows indicating total χ^2 values are computed accounting for any relevant correlated systematic errors. Datasets marked with an asterisk indicate that they were not used to constrain $\mathbf{C}_{33}^{D\mu}$.

Dataset	n_{data}	SM Baseline	$\mathbf{C}_{33}^{D\mu}$
ATLAS DY high-mass 7 TeV	13	1.631	1.625
ATLAS DY high-mass 8 TeV	46	1.179	1.168
CMS DY high-mass 7 TeV	117	1.666	1.664
CMS DY high-mass 8 TeV	41	0.858	0.893
CMS DY high-mass 13 TeV ee*	43	2.579	2.574
CMS DY high-mass 13 TeV $\mu\mu$	43	0.836	0.837
Total DY (HM only)	**303**	1.501	1.503
HL-LHC CC e*	16	0.437	0.424
HL-LHC CC μ*	16	0.841	0.752
HL-LHC NC e*	12	1.108	1.130
HL-LHC NC μ	12	0.943	0.969
Total HL-LHC only	**56**	0.830	0.826

the following dedicated study [257], but rather focus on the robustness of the fit of the Wilson coefficients and of the PDFs in a simultaneous fit.

Crucially, the central values of the SMEFT-sensitive datasets are modified by artificially contaminating them with pre-chosen, extreme, values of Wilson coefficients. This will emulate a situation where a dataset will favour a non-vanishing EFT operator. In the first part of this section we show that our methodology is sufficiently flexible to recover these chosen values. In the second part of this section we consider the case where the input datasets are replaced by theory predictions from a known underlying PDF set, rather than experimental central values. We show that even in the case of fixing both the underlying PDF and Wilson coefficients to some *a priori* values, the methodology is sufficiently robust that it can simultaneously reproduce both.

6.6.1 Closure test results on the Wilson Coefficients

In this section we will fix the value of the Wilson coefficients *a priori* and study how effectively we can retrieve these values using our methodology. To perform this kind of

closure test, we artificially choose extreme values of the Wilson coefficients considered in Benchmark Scenario I, where the interplay between PDF and SMEFT fits is more marked, in two separate analyses:

1. The individual fit of the W Wilson coefficient including the current high-mass DY data from LHC Run I and Run II along with all other datasets listed in section 5.4.1. Specifically we set $W = 1 \times 10^{-3}$, which is outside the 95% C. L. bounds that are displayed in figure 6.4

2. The combined fit of the (W, Y) parameters including the HL-LHC projections. Specifically we set $(W, Y) = (1, -1) \times 10^{-4}$, which is also far outside the 95% C. L. contours displayed in figure 6.7.

The way we input these non-zero values of the Wilson coefficients in the underlying law is by multiplying the Monte Carlo pseudodata central values by the SMEFT K-factors obtained by setting the Wilson coefficient(s) to the aforementioned values. The fitting methodology proceeds as before. Importantly, the entire tool chain has no knowledge of what the value of the Wilson coefficient it is looking for are set to.

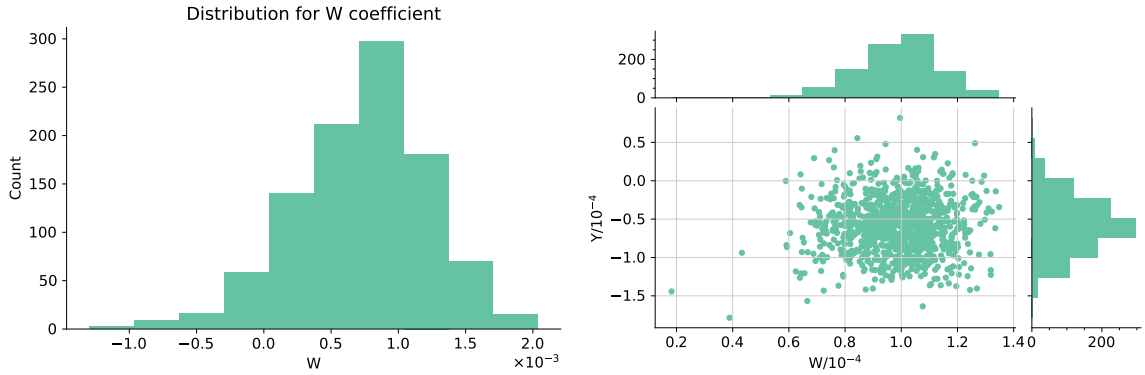


Figure 6.12: Result of the closure testing framework for our methodology. Left: histogram for the distribution of the W parameter when the input data has been modified by setting $W = 1 \times 10^{-3}$. Right: distribution of (W, Y) when fitting to data that has been modified by setting $(W, Y) = (1, -1) \times 10^{-4}$. The upper and right panels show the histograms for the distribution of the best fit values in their respective directions.

The result of the closure test are shown in figure 6.12. On the left-hand panel we see that the distribution of best fit values of W across 1000 replicas, that were previously centred at the origin, now moves considerably towards the *a priori* value of $W = 1 \times 10^{-3}$. Similarly, in the second analysis displayed on the right-hand panel, we see that the presence of HL-LHC data continues to eliminate the flat direction, with

the distribution of best fit Wilson coefficients resembling that of figure 6.7, but with the best-fits values of the (W, Y) parameters consistently pushed towards the pre-selected *a priori* values of $(W, Y) = (1, -1) \times 10^{-4}$. We see that, despite an extremal choice of the injected values of the Wilson coefficients, the methodology is sufficiently flexible and robust to recover them.

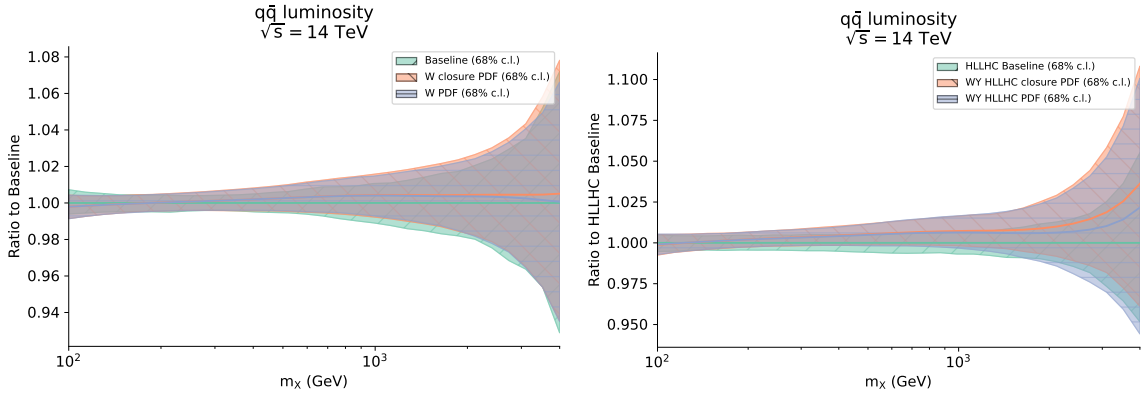


Figure 6.13: The $q\bar{q}$ luminosity obtained from the closure tests when the experimental central values are modified to encode a specific SMEFT benchmarking scenario ("closure PDF" in orange), compared to the results of the simultaneous fits ("PDF" in blue) normalized to the corresponding baseline (in green). Shown in the left panel is the $W = 1 \times 10^{-3}$ scenario while the $(W, Y) = (1, -1) \times 10^{-4}$ scenario is displayed in the right panel.

The PDFs generated in the context of the closure tests are displayed in figure 6.13. On the left panel, we plot the quark-antiquark luminosity obtained when the experimental central values are modified by inputting $W = 1 \times 10^{-3}$ (called "W closure PDF") and compare it to the ones that we obtain in the simultaneously fit of the PDFs and W presented in section 6.4.1 (called "W PDF"), both normalized to the SM baseline. We observe that the PDFs generated with our closure test for $W = 1 \times 10^{-3}$ are similar to those that we obtain in the simultaneous fit, despite the fact that the training dataset has been heavily modified by the extreme *a priori* choice of W. This is to be expected, since one can view the combination layer as capturing the data's dependence on the Wilson coefficients, whilst the complementary PDF sector of the network architecture captures the data's dependence on the underlying PDF, which of course remains unchanged. The combination layer, in effect, subtracts off the EFT dependence, leaving behind the pure SM contribution for the PDF sector to parameterize. This exact same sentiment is echoed in the case of the HL-LHC based closure test where the data is contaminated with a choice of $(W, Y) = (1, -1) \times 10^{-4}$ displayed on the right-hand panel of the figure. Again, the "WY HLLHC closure PDF"

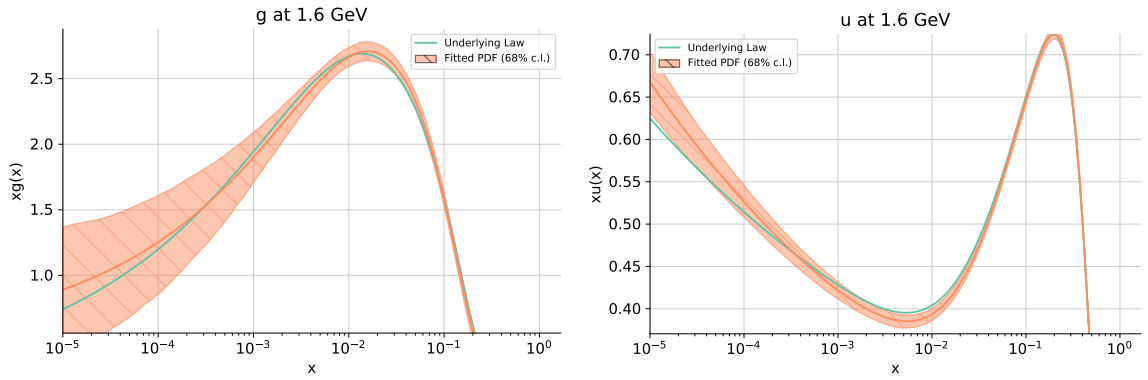


Figure 6.14: The gluon (left) and up quark (right) PDFs obtained from the closure test framework in which both the underlying PDF set and Wilson coefficients are known. Shown in green is the PDF replica used as the underlying law which generates the fake data used to train our model. The resulting PDFs are shown in orange along with their 68% confidence level bands. The fake data generated by the underlying law is subsequently modified so as to encode the $(W, Y) = (1, -1) \times 10^{-4}$ condition.

obtained when performing a fit to the contaminated data is virtually indistinguishable from the simultaneous fits presented in figure 6.8, labelled "WY HLLHC PDF".

This result is indeed a remarkable feature of our methodology and serves to prove its robustness. By performing a simultaneous fit of the PDFs and Wilson coefficients using the `SIMUnet` methodology, one is guaranteed that the resulting PDFs are free from any possible BSM contamination that may be present in the data: so long as the contamination is entirely captured by the choice of EFT operators.

6.6.2 Closure test results on the simultaneous fit

The natural extension of the closure test described in the previous subsection is to assess the degree to which `SIMUnet` is able to replicate, not only fixed Wilson coefficients, but also a known underlying PDF. For this scenario we employ the NNPDF level 2 [136] closure test strategy. In the context of a simultaneous fitting methodology this amounts to generating Standard Model predictions using a known PDF set (referred to henceforth as the underlying law) and mapping these to SMEFT observables by multiplying the SM theory predictions with SMEFT K-factors scaled by a previously determined choice of Wilson coefficients. These SMEFT observables, generated by the underlying law, replace the usual MC pseudodata replicas and are used to train the neural network in the usual way.

Through this approach we are able to assess the degree to which the parameterization is able to capture not only an underlying choice of Wilson coefficients, but ensures it is sufficiently flexible to adequately replicate a known PDF. For such a closure test, we use the HL-LHC baseline used in this study as the underlying law with the SMEFT scenario being again the simultaneous (W, Y) determination, with the input data being adjusted to have $(W, Y) = (1, -1) \times 10^{-4}$ encoded within it. The PDFs generated in this way are shown in figure 6.14 for representative choices of parton flavour: we display both the resulting PDFs as well as the underlying law. We see that, despite modifying the training data with an extreme choice of SMEFT benchmark, the methodology is sufficiently robust so as to be able to recover the true PDF set with good precision. Indeed, the distribution of best fit (W, Y) values plotted in figure 6.15 shows that not only did the methodology retrieve the underlying law, but also managed to recover the chosen SMEFT scenario: being able to correctly determine the $(W, Y) = (1, -1) \times 10^{-4}$ condition. Such behaviour is reminiscent of the above closure test whereby the combination layer is able to parameterise the BSM dynamics while the preceding layers of the model are left to parameterise the PDFs. This conclusion thus illustrates how our methodology is able to correctly disentangle the interplay between PDFs and BSM dynamics.

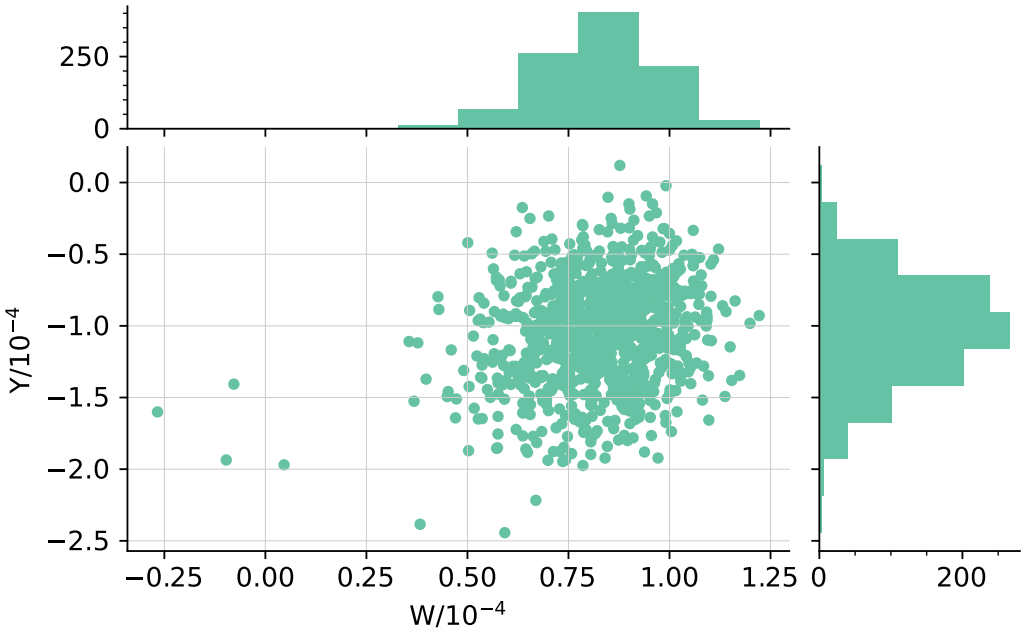


Figure 6.15: Distribution of best fit (W, Y) parameters for each Monte Carlo replica. The data used for the fit was generated by a pre-selected PDF before the SM predictions were transformed to a SMEFT observable with $(W, Y) = (1, -1) \times 10^{-4}$. The upper panel is the histogram of best fit values in the W axis while the right panel is the histogram for the Y axis.

Chapter 7

Concluding remarks and outlook

In this thesis we have presented a discussion of the precision determination of proton subnuclear structure in terms of its elementary constituents; the quarks and gluons. Knowledge of non-perturbative hadron dynamics is a vital consideration for precision QCD predictions in the era of hadron based collider experiments. To this end we have demonstrated how neural networks, an important subclass of machine learning based approaches, form the ideal parameterization for the parton distribution functions of the proton. An outline of deep learning was presented, before discussing how it can be applied to precision PDF determinations in a global QCD analysis. This led us to present the NNPDF4.0 PDF set, which forms the latest and one of the most precisely determined PDF sets to date. Alongside this, the entire codebase used for the PDF fit is made publicly available and open source: the first global PDF fitting platform to do so. This allows for the high energy physics community to extend and contribute to the framework as well as to use it for their own phenomenological studies.

We showed how the precision determination of the strange quark content of the proton is not only of utmost importance for the determination of various Standard Model quantities, but also a topic of great interest in current LHC phenomenology. To this end a dedicated study of the proton strange content was performed using a host of strange sensitive measurements from neutrino deep inelastic scattering experiments to gauge boson production based processes at the LHC.

We then turned our attention to consider the interplay of PDFs with possible BSM manifestations at the LHC. The issue of the validity of assuming the Standard Model at all scales was brought into question by considering how the SMEFT induced back-reaction on PDFs can alter the interpretation of Wilson coefficient bounds. We started to study this question by considering the effect of lepton-quark dimension-

6 operators of the SMEFT in the context of deep inelastic scattering experiments. This formed the ideal arena to study this interplay due to the relative theoretical simplicity of hadron-lepton scattering as well as the precision of the HERA combined datasets. Moreover, we exploited the broad kinematic reach of DIS measurements to achieve sensitivity not only on the PDFs, but also on the EFT operators thanks to the high-Q reach of HERA. The interplay was found to be mild, with possible BSM reabsorption effects by the PDF being slight. Inspired by the results of this study one was then naturally led to extend this approach to incorporate high-mass Drell-Yan measurements from the LHC, which covers a much larger region of Q. We selected two beyond the Standard Model scenarios which we motivated by selecting operators deemed to be sufficiently sensitive to these Drell-Yan observables. The methodology naturally extended to accommodate these processes and operators, with the interplay being particularly strong with the EFT oblique parameters at the High-Luminosity LHC. We saw that in this particular case, if one was to neglect the EFT back-reaction, then significantly misleading results would be obtained.

This then motivated the need for a new generation of fitting methodology that can systematically disentangle EFT effects from the PDF. To this end, we introduced the `SIMUnet` methodology which, for the first time, allowed for a truly simultaneous determination of PDFs and external parameters. This approach extends the NNPDF4.0 neural network architecture by the inclusion of an additional sequential layer, the architecture of which is atypical in the deep learning community. We showcased its ability by performing a PDF and BSM simultaneous determination using high-mass Drell-Yan data and the oblique parameters of the above study. We showed how the methodology can find the presence of a flat direction without the need for prior user knowledge or input. This flat direction was then broken by introducing charged-current Drell-Yan data from the High-Luminosity LHC projected data. We showed how the need for linearity in Wilson coefficients, at the matrix element level, is not at all imposed and how the quadratic and beyond terms can be readily accounted for by the simple inclusion of non-trainable edges. The robustness of the `SIMUnet` methodology was then assessed, whereby we showed that our approach is sufficiently flexible to not only recover a prior choice of Wilson coefficient, but also an underlying choice of PDF set. This so-called closure test contaminates the data with a defined choice of extreme Wilson coefficient before providing it to the `SIMUnet` model. We showed how the contamination yielded the same PDFs as the non-closure test PDFs, despite the large difference in numerical value of training data. This then led us to conclude that

the architecture is such that the combination layer captures the data dependence on the Wilson coefficients, leaving the PDF layers free to replicate the data dependence on the true, underlying, PDF.

The next steps in this program are to apply the `SIMUnet` approach to the top sector of the SMEFT operators and employ top quark measurements from the LHC. In the literature, this sector has gained a lot of interest thanks to its suspected strong sensitivity to possible BSM dynamics. Moreover, we discussed how `SIMUnet` can be employed for a precision determination of PDFs and the strong coupling, α_s. Again, this is a topic of great interest thanks to the strong correlation between these two objects. In the past, these approaches were based on the Hessian approach presented in chapter 5, but we discussed in chapter 6 how an interpolation in FK-table space based approach can be used to obtain a better determination. This would, again, be the first ever truly simultaneous determination of PDFs and α_s and could in principle use the entire global dataset entering a PDF fit.

Work on both these extensions is under way, the results of which are left to future publications. In the longer term, a strong case for the `SIMUnet` methodology, presented here, has been presented to be the approach of choice to have a first global, precision, determination of Standard Model quantities, fitting electroweak parameters, the strong coupling, PDFs themselves, and possibly the entirety of the Warsaw basis of Wilson coefficients. Such an endeavour is crucial for indirect new physics searches, the feasibility of which has only now been made possible thanks to the work presented in this text.

References

[1] R. D. Ball *et al.*, "The Path to Proton Structure at One-Percent Accuracy," 9 2021.

[2] R. D. Ball *et al.*, "An open-source machine learning framework for global analyses of parton distributions," *Eur. Phys. J. C*, vol. 81, no. 10, p. 958, 2021.

[3] F. Faura, S. Iranipour, E. R. Nocera, J. Rojo, and M. Ubiali, "The Strangest Proton?," *Eur. Phys. J. C*, vol. 80, no. 12, p. 1168, 2020.

[4] S. Carrazza, C. Degrande, S. Iranipour, J. Rojo, and M. Ubiali, "Can New Physics hide inside the proton?," *Phys. Rev. Lett.*, vol. 123, no. 13, p. 132001, 2019.

[5] A. Greljo, S. Iranipour, Z. Kassabov, M. Madigan, J. Moore, J. Rojo, M. Ubiali, and C. Voisey, "Parton distributions in the SMEFT from high-energy Drell-Yan tails," *JHEP*, vol. 07, p. 122, 2021.

[6] S. Iranipour and M. Ubiali, "A new generation of simultaneous fits to LHC data using deep learning," 1 2022.

[7] A. Salam, "Weak and Electromagnetic Interactions," *Conf. Proc. C*, vol. 680519, pp. 367–377, 1968.

[8] S. L. Glashow, "Partial Symmetries of Weak Interactions," *Nucl. Phys.*, vol. 22, pp. 579–588, 1961.

[9] S. Weinberg, "A Model of Leptons," *Phys. Rev. Lett.*, vol. 19, pp. 1264–1266, 1967.

[10] M. Kobayashi and T. Maskawa, "CP Violation in the Renormalizable Theory of Weak Interaction," *Prog. Theor. Phys.*, vol. 49, pp. 652–657, 1973.

[11] M. Gell-Mann, "A Schematic Model of Baryons and Mesons," *Phys. Lett.*, vol. 8, pp. 214–215, 1964.

[12] G. Zweig, *An SU(3) model for strong interaction symmetry and its breaking. Version 2*, pp. 22–101. 2 1964.

[13] G. Zweig, "Concrete Quarks," *Subnucl. Ser.*, vol. 52, pp. 81–116, 2017.

[14] P. W. Higgs, "Broken Symmetries and the Masses of Gauge Bosons," *Phys. Rev. Lett.*, vol. 13, pp. 508–509, 1964.

[15] G. Aad *et al.*, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC," *Phys. Lett. B*, vol. 716, pp. 1–29, 2012.

[16] S. Chatrchyan *et al.*, "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC," *Phys. Lett. B*, vol. 716, pp. 30–61, 2012.

[17] J. S. Schwinger, "On Quantum electrodynamics and the magnetic moment of the electron," *Phys. Rev.*, vol. 73, pp. 416–417, 1948.

[18] T. Aoyama, M. Hayakawa, T. Kinoshita, and M. Nio, "Tenth-Order QED Contribution to the Electron g-2 and an Improved Value of the Fine Structure Constant," *Phys. Rev. Lett.*, vol. 109, p. 111807, 2012.

[19] T. Aoyama, M. Hayakawa, T. Kinoshita, and M. Nio, "Tenth-Order Electron Anomalous Magnetic Moment — Contribution of Diagrams without Closed Lepton Loops," *Phys. Rev. D*, vol. 91, no. 3, p. 033006, 2015. [Erratum: Phys.Rev.D 96, 019901 (2017)].

[20] J. F. Donoghue, "General relativity as an effective field theory: The leading quantum corrections," *Phys. Rev. D*, vol. 50, pp. 3874–3888, 1994.

[21] Y. Fukuda *et al.*, "Evidence for oscillation of atmospheric neutrinos," *Phys. Rev. Lett.*, vol. 81, pp. 1562–1567, 1998.

[22] Q. R. Ahmad *et al.*, "Direct evidence for neutrino flavor transformation from neutral current interactions in the Sudbury Neutrino Observatory," *Phys. Rev. Lett.*, vol. 89, p. 011301, 2002.

[23] F. Zwicky, "On the Masses of Nebulae and of Clusters of Nebulae," *Astrophys. J.*, vol. 86, pp. 217–246, 1937.

[24] R. Aaij *et al.*, "Test of lepton universality using $B^+ \to K^+\ell^+\ell^-$ decays," *Phys. Rev. Lett.*, vol. 113, p. 151601, 2014.

[25] R. Aaij *et al.*, "Test of lepton universality with $B^0 \to K^{*0}\ell^+\ell^-$ decays," *JHEP*, vol. 08, p. 055, 2017.

[26] R. Aaij *et al.*, "Search for lepton-universality violation in $B^+ \to K^+\ell^+\ell^-$ decays," *Phys. Rev. Lett.*, vol. 122, no. 19, p. 191801, 2019.

[27] H. N. Brown *et al.*, "Precise measurement of the positive muon anomalous magnetic moment," *Phys. Rev. Lett.*, vol. 86, pp. 2227–2231, 2001.

[28] G. W. Bennett *et al.*, "Measurement of the positive muon anomalous magnetic moment to 0.7 ppm," *Phys. Rev. Lett.*, vol. 89, p. 101804, 2002. [Erratum: Phys.Rev.Lett. 89, 129903 (2002)].

[29] G. W. Bennett *et al.*, "Measurement of the negative muon anomalous magnetic moment to 0.7 ppm," *Phys. Rev. Lett.*, vol. 92, p. 161802, 2004.

[30] B. Abi *et al.*, "Measurement of the Positive Muon Anomalous Magnetic Moment to 0.46 ppm," *Phys. Rev. Lett.*, vol. 126, no. 14, p. 141801, 2021.

[31] S. Borsanyi *et al.*, "Leading hadronic contribution to the muon magnetic moment from lattice QCD," *Nature*, vol. 593, no. 7857, pp. 51–55, 2021.

[32] C. Anastasiou, C. Duhr, F. Dulat, F. Herzog, and B. Mistlberger, "Higgs Boson Gluon-Fusion Production in QCD at Three Loops," *Phys. Rev. Lett.*, vol. 114, p. 212001, 2015.

[33] M. Czakon, P. Fiedler, and A. Mitov, "Total Top-Quark Pair-Production Cross Section at Hadron Colliders Through $O(\alpha_S^4)$," *Phys. Rev. Lett.*, vol. 110, p. 252004, 2013.

[34] B. Mistlberger, "Higgs boson production at hadron colliders at N^3LO in QCD," *JHEP*, vol. 05, p. 028, 2018.

[35] T.-J. Hou *et al.*, "New CTEQ global analysis of quantum chromodynamics with high-precision data from the LHC," *Phys. Rev. D*, vol. 103, no. 1, p. 014013, 2021.

[36] L. A. Harland-Lang, A. D. Martin, P. Motylinski, and R. S. Thorne, "Parton distributions in the LHC era: MMHT 2014 PDFs," *Eur. Phys. J. C*, vol. 75, no. 5, p. 204, 2015.

[37] S. Alekhin, J. Blümlein, S. Moch, and R. Placakyte, "Parton distribution functions, α_s, and heavy-quark masses for LHC Run II," *Phys. Rev. D*, vol. 96, no. 1, p. 014011, 2017.

[38] I. Brivio, S. Bruggisser, F. Maltoni, R. Moutafis, T. Plehn, E. Vryonidou, S. Westhoff, and C. Zhang, "O new physics, where art thou? A global search in the top sector," *JHEP*, vol. 02, p. 131, 2020.

[39] J. Ellis, C. W. Murphy, V. Sanz, and T. You, "Updated Global SMEFT Fit to Higgs, Diboson and Electroweak Data," *JHEP*, vol. 06, p. 146, 2018.

[40] A. Buckley, C. Englert, J. Ferrando, D. J. Miller, L. Moore, M. Russell, and C. D. White, "Constraining top quark effective theory in the LHC Run II era," *JHEP*, vol. 04, p. 015, 2016.

[41] S. Brown, A. Buckley, C. Englert, J. Ferrando, P. Galler, D. J. Miller, L. Moore, M. Russell, C. White, and N. Warrack, "TopFitter: Fitting top-quark Wilson Coefficients to Run II data," *PoS*, vol. ICHEP2018, p. 293, 2019.

[42] R. Gomez-Ambrosio, "Studies of Dimension-Six EFT effects in Vector Boson Scattering," *Eur. Phys. J. C*, vol. 79, no. 5, p. 389, 2019.

[43] J. Ellis, M. Madigan, K. Mimasu, V. Sanz, and T. You, "Top, Higgs, Diboson and Electroweak Fit to the Standard Model Effective Field Theory," *JHEP*, vol. 04, p. 279, 2021.

[44] J. J. Ethier, G. Magni, F. Maltoni, L. Mantani, E. R. Nocera, J. Rojo, E. Slade, E. Vryonidou, and C. Zhang, "Combined SMEFT interpretation of Higgs, diboson, and top quark data from the LHC," *JHEP*, vol. 11, p. 089, 2021.

[45] A. Greljo and D. Marzocca, "High-p_T dilepton tails and flavor physics," *Eur. Phys. J. C*, vol. 77, no. 8, p. 548, 2017.

[46] J. Ellis, V. Sanz, and T. You, "The Effective Standard Model after LHC Run I," *JHEP*, vol. 03, p. 157, 2015.

[47] J. de Blas, J. C. Criado, M. Perez-Victoria, and J. Santiago, "Effective description of general extensions of the Standard Model: the complete tree-level dictionary," *JHEP*, vol. 03, p. 109, 2018.

[48] V. Bargmann, "On Unitary ray representations of continuous groups," *Annals Math.*, vol. 59, pp. 1–46, 1954.

[49] N. Seiberg and E. Witten, "Electric - magnetic duality, monopole condensation, and confinement in N=2 supersymmetric Yang-Mills theory," *Nucl. Phys. B*, vol. 426, pp. 19–52, 1994. [Erratum: Nucl.Phys.B 430, 485–486 (1994)].

[50] S. Weinberg, *The quantum theory of fields. Vol. 3: Supersymmetry.* Cambridge University Press, 6 2013.

[51] S. Weinberg, "Phenomenological Lagrangians," *Physica A*, vol. 96, no. 1-2, pp. 327–340, 1979.

[52] C.-N. Yang and R. L. Mills, "Conservation of Isotopic Spin and Isotopic Gauge Invariance," *Phys. Rev.*, vol. 96, pp. 191–195, 1954.

[53] V. N. Gribov, "Quantization of Nonabelian Gauge Theories," *Nucl. Phys. B*, vol. 139, p. 1, 1978.

[54] L. D. Faddeev and V. N. Popov, "Feynman Diagrams for the Yang-Mills Field," *Phys. Lett. B*, vol. 25, pp. 29–30, 1967.

[55] S. Weinberg, *The Quantum Theory of Fields*, vol. 2. Cambridge University Press, 1996.

[56] W. Pauli and F. Villars, "On the Invariant regularization in relativistic quantum theory," *Rev. Mod. Phys.*, vol. 21, pp. 434–444, 1949.

[57] S. Weinberg, *The Quantum Theory of Fields*, vol. 1. Cambridge University Press, 1995.

[58] K. G. Wilson, "Quantum field - theory models in less than 4 dimensions," *Phys. Rev. D*, vol. 7, pp. 2911–2926, May 1973.

[59] G. 't Hooft and M. J. G. Veltman, "Regularization and Renormalization of Gauge Fields," *Nucl. Phys. B*, vol. 44, pp. 189–213, 1972.

[60] P. A. Baikov, K. G. Chetyrkin, and J. H. Kühn, "Five-Loop Running of the QCD coupling constant," *Phys. Rev. Lett.*, vol. 118, no. 8, p. 082002, 2017.

[61] T. Luthe, A. Maier, P. Marquard, and Y. Schröder, "Towards the five-loop Beta function for a general gauge group," *JHEP*, vol. 07, p. 127, 2016.

[62] T. Luthe, A. Maier, P. Marquard, and Y. Schroder, "The five-loop Beta function for a general gauge group and anomalous dimensions beyond Feynman gauge," *JHEP*, vol. 10, p. 166, 2017.

[63] F. Herzog, B. Ruijl, T. Ueda, J. A. M. Vermaseren, and A. Vogt, "The five-loop beta function of Yang-Mills theory with fermions," *JHEP*, vol. 02, p. 090, 2017.

[64] K. G. Chetyrkin, G. Falcioni, F. Herzog, and J. A. M. Vermaseren, "Five-loop renormalisation of QCD in covariant gauges," *JHEP*, vol. 10, p. 179, 2017. [Addendum: JHEP 12, 006 (2017)].

[65] D. J. Gross and F. Wilczek, "Ultraviolet Behavior of Nonabelian Gauge Theories," *Phys. Rev. Lett.*, vol. 30, pp. 1343–1346, 1973.

[66] H. D. Politzer, "Reliable Perturbative Results for Strong Interactions?," *Phys. Rev. Lett.*, vol. 30, pp. 1346–1349, 1973.

[67] G. 't Hooft, "A Planar Diagram Theory for Strong Interactions," *Nucl. Phys. B*, vol. 72, p. 461, 1974.

[68] E. Witten, "Baryons in the 1/n Expansion," *Nucl. Phys. B*, vol. 160, pp. 57–115, 1979.

[69] R. P. Feynman, "Very high-energy collisions of hadrons," *Phys. Rev. Lett.*, vol. 23, pp. 1415–1417, 1969.

[70] J. D. Bjorken and E. A. Paschos, "Inelastic Electron Proton and gamma Proton Scattering, and the Structure of the Nucleon," *Phys. Rev.*, vol. 185, pp. 1975–1982, 1969.

[71] M. Gell-Mann, "The eightfold way: A theory of strong interaction symmetry," in *The Eightfold Way*, pp. 11–57, CRC Press, 2018.

[72] A. Manohar, P. Nason, G. P. Salam, and G. Zanderighi, "How bright is the proton? A precise determination of the photon parton distribution function," *Phys. Rev. Lett.*, vol. 117, no. 24, p. 242002, 2016.

[73] A. V. Manohar, P. Nason, G. P. Salam, and G. Zanderighi, "The Photon Content of the Proton," *JHEP*, vol. 12, p. 046, 2017.

[74] B. Fornal, A. V. Manohar, and W. J. Waalewijn, "Electroweak Gauge Boson Parton Distribution Functions," *JHEP*, vol. 05, p. 106, 2018.

[75] L. Buonocore, P. Nason, F. Tramontano, and G. Zanderighi, "Leptons in the proton," *JHEP*, vol. 08, no. 08, p. 019, 2020.

[76] M. D. Schwartz, *Quantum field theory and the standard model*. Cambridge University Press, 2014.

[77] J. C. Collins and D. E. Soper, "Parton Distribution and Decay Functions," *Nucl. Phys. B*, vol. 194, pp. 445–492, 1982.

[78] J. C. Collins, "What exactly is a parton density?," *Acta Phys. Polon. B*, vol. 34, p. 3103, 2003.

[79] J. C. Collins, D. E. Soper, and G. F. Sterman, "Factorization of Hard Processes in QCD," *Adv. Ser. Direct. High Energy Phys.*, vol. 5, pp. 1–91, 1989.

[80] C. G. Callan, Jr. and D. J. Gross, "High-energy electroproduction and the constitution of the electric current," *Phys. Rev. Lett.*, vol. 22, pp. 156–159, 1969.

[81] T. Kinoshita, "Mass singularities of Feynman amplitudes," *J. Math. Phys.*, vol. 3, pp. 650–677, 1962.

[82] T. D. Lee and M. Nauenberg, "Degenerate Systems and Mass Singularities," *Phys. Rev.*, vol. 133, pp. B1549–B1562, 1964.

[83] F. Bloch and A. Nordsieck, "Note on the Radiation Field of the electron," *Phys. Rev.*, vol. 52, pp. 54–59, 1937.

[84] R. Doria, J. Frenkel, and J. C. Taylor, "Counter Example to Nonabelian Bloch-Nordsieck Theorem," *Nucl. Phys. B*, vol. 168, pp. 93–110, 1980.

[85] Y. L. Dokshitzer, "Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics.," *Sov. Phys. JETP*, vol. 46, pp. 641–653, 1977.

[86] G. Altarelli and G. Parisi, "Asymptotic Freedom in Parton Language," *Nucl. Phys. B*, vol. 126, pp. 298–318, 1977.

[87] V. N. Gribov and L. N. Lipatov, "Deep inelastic e p scattering in perturbation theory," *Sov. J. Nucl. Phys.*, vol. 15, pp. 438–450, 1972.

[88] E. B. Zijlstra and W. L. van Neerven, "Order α_s^2 QCD corrections to the deep inelastic proton structure functions F2 and F(L)," *Nucl. Phys. B*, vol. 383, pp. 525–574, 1992.

[89] A. Vogt, S. Moch, and J. A. M. Vermaseren, "The Three-loop splitting functions in QCD: The Singlet case," *Nucl. Phys. B*, vol. 691, pp. 129–181, 2004.

[90] R. K. Ellis, W. J. Stirling, and B. R. Webber, *QCD and collider physics*, vol. 8. Cambridge University Press, 2 2011.

[91] M. E. Peskin and D. V. Schroeder, *An Introduction to quantum field theory*. Reading, USA: Addison-Wesley, 1995.

[92] S. Chekanov *et al.*, "Measurement of the Longitudinal Proton Structure Function at HERA," *Phys. Lett. B*, vol. 682, pp. 8–22, 2009.

[93] J. C. Collins, D. E. Soper, and G. F. Sterman, "Factorization for Short Distance Hadron - Hadron Scattering," *Nucl. Phys. B*, vol. 261, pp. 104–142, 1985.

[94] R. K. Ellis, H. Georgi, M. Machacek, H. D. Politzer, and G. G. Ross, "Perturbation Theory and the Parton Model in QCD," *Nucl. Phys. B*, vol. 152, pp. 285–329, 1979.

[95] G. T. Bodwin, "Factorization of the Drell-Yan Cross-Section in Perturbation Theory," *Phys. Rev. D*, vol. 31, p. 2616, 1985. [Erratum: Phys.Rev.D 34, 3932 (1986)].

[96] H. Georgi, "An Effective Field Theory for Heavy Quarks at Low-energies," *Phys. Lett. B*, vol. 240, pp. 447–450, 1990.

[97] H. Georgi, "Effective field theory," *Ann. Rev. Nucl. Part. Sci.*, vol. 43, pp. 209–252, 1993.

[98] B. Henning, X. Lu, and H. Murayama, "How to use the Standard Model effective field theory," *JHEP*, vol. 01, p. 023, 2016.

[99] C. Arzt, "Reduced effective Lagrangians," *Phys. Lett. B*, vol. 342, pp. 189–195, 1995.

[100] A. V. Manohar, "Effective field theories," *Lect. Notes Phys.*, vol. 479, pp. 311–362, 1997.

[101] R. N. Mohapatra and A. Y. Smirnov, "Neutrino mass and new physics," *Annu. Rev. Nucl. Part. Sci.*, vol. 56, pp. 569–628, 2006.

[102] G. Bertone and D. Hooper, "History of dark matter," *Reviews of Modern Physics*, vol. 90, no. 4, p. 045002, 2018.

[103] I. Brivio and M. Trott, "The Standard Model as an Effective Field Theory," *Phys. Rept.*, vol. 793, pp. 1–98, 2019.

[104] T. Cohen, N. Craig, X. Lu, and D. Sutherland, "Is SMEFT Enough?," *JHEP*, vol. 03, p. 237, 2021.

[105] D. B. Kaplan and H. Georgi, "SU(2) x U(1) Breaking by Vacuum Misalignment," *Phys. Lett. B*, vol. 136, pp. 183–186, 1984.

[106] D. B. Kaplan, H. Georgi, and S. Dimopoulos, "Composite Higgs Scalars," *Phys. Lett. B*, vol. 136, pp. 187–190, 1984.

[107] S. Weinberg, "Baryon and Lepton Nonconserving Processes," *Phys. Rev. Lett.*, vol. 43, pp. 1566–1570, 1979.

[108] B. Grzadkowski, M. Iskrzynski, M. Misiak, and J. Rosiek, "Dimension-Six Terms in the Standard Model Lagrangian," *JHEP*, vol. 10, p. 085, 2010.

[109] C. W. Murphy, "Dimension-8 operators in the Standard Model Eective Field Theory," *JHEP*, vol. 10, p. 174, 2020.

[110] H.-L. Li, Z. Ren, J. Shu, M.-L. Xiao, J.-H. Yu, and Y.-H. Zheng, "Complete set of dimension-eight operators in the standard model effective field theory," *Phys. Rev. D*, vol. 104, no. 1, p. 015026, 2021.

[111] J. H. Friedman, *The elements of statistical learning: Data mining, inference, and prediction*. springer open, 2017.

[112] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[113] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[114] S. Wang and J. Jiang, "Learning natural language inference with LSTM," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (San Diego, California), pp. 1442–1451, Association for Computational Linguistics, June 2016.

[115] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[116] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[117] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[118] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, "The expressive power of neural networks: A view from the width," *Advances in neural information processing systems*, vol. 30, 2017.

[119] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of control, signals and systems*, vol. 2, no. 4, pp. 303–314, 1989.

[120] D. Mishkin and J. Matas, "All you need is a good init," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2016.

[121] J. Y. Yam and T. W. Chow, "A weight initialization method for improving training speed in feedforward neural network," *Neurocomputing*, vol. 30, no. 1-4, pp. 219–232, 2000.

[122] S. Ruder, "An overview of gradient descent optimization algorithms," *arXiv preprint arXiv:1609.04747*, 2016.

[123] Y. N. Dauphin, R. Pascanu, C. Gulcehre, K. Cho, S. Ganguli, and Y. Bengio, "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization," *Advances in neural information processing systems*, vol. 27, 2014.

[124] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (Y. Bengio and Y. LeCun, eds.), 2015.

[125] Y. Nesterov, "A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$," *Proceedings of the USSR Academy of Sciences*, vol. 269, pp. 543–547, 1983.

[126] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *NIPS 2017 Workshop on Autodiff*, 2017.

[127] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283, 2016.

[128] P. K. Sinervo, "Definition and treatment of systematic uncertainties in high energy physics and astrophysics," *Statistical Problems in Particle Physics, Astrophysics, and Cosmology*, pp. 122–129, 2003.

[129] G. D'Agostini, "On the use of the covariance matrix to fit correlated data," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 346, no. 1-2, pp. 306–311, 1994.

[130] R. D. Ball, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, and M. Ubiali, "Fitting Parton Distribution Data with Multiplicative Normalization Uncertainties," *JHEP*, vol. 05, p. 075, 2010.

[131] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, 2021.

[132] R. D. Ball, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, A. Piccione, J. Rojo, and M. Ubiali, "A Determination of parton distributions with faithful uncertainty estimation," *Nucl. Phys. B*, vol. 809, pp. 1–63, 2009. [Erratum: Nucl.Phys.B 816, 293 (2009)].

[133] S. K. Lam, A. Pitrou, and S. Seibert, "Numba: A llvm-based python jit compiler," in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pp. 1–6, 2015.

[134] R. D. Ball, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, and M. Ubiali, "A first unbiased global NLO determination of parton distributions and their uncertainties," *Nucl. Phys. B*, vol. 838, pp. 136–206, 2010.

[135] L. Del Debbio, S. Forte, J. I. Latorre, A. Piccione, and J. Rojo, "Neural network determination of parton distributions: The Nonsinglet case," *JHEP*, vol. 03, p. 039, 2007.

[136] R. D. Ball *et al.*, "Parton distributions for the LHC Run II," *JHEP*, vol. 04, p. 040, 2015.

[137] R. D. Ball *et al.*, "Parton distributions from high-precision collider data," *Eur. Phys. J. C*, vol. 77, no. 10, p. 663, 2017.

[138] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[139] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Advances in neural information processing systems*, pp. 950–957, 1992.

[140] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, and D. D. Cox, "Hyperopt: a Python library for model selection and hyperparameter optimization," *Computational Science and Discovery*, vol. 8, p. 014008, Jan. 2015.

[141] J. McCall, "Genetic algorithms for modelling and optimisation," *Journal of computational and Applied Mathematics*, vol. 184, no. 1, pp. 205–222, 2005.

[142] S. Raschka, "Model evaluation, model selection, and algorithm selection in machine learning," *arXiv preprint arXiv:1811.12808*, 2018.

[143] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics* (Y. W. Teh and M. Titterington, eds.), vol. 9 of *Proceedings of Machine Learning Research*, (Chia Laguna Resort, Sardinia, Italy), pp. 249–256, PMLR, 13–15 May 2010.

[144] T. Dozat, "Incorporating Nesterov Momentum into Adam," in *Proceedings of the 4th International Conference on Learning Representations*, pp. 1–4.

[145] R. D. Ball *et al.*, "Parton distributions with LHC data," *Nucl. Phys. B*, vol. 867, pp. 244–289, 2013.

[146] V. Bertone, S. Carrazza, and N. P. Hartland, "APFELgrid: a high performance tool for parton density determinations," *Comput. Phys. Commun.*, vol. 212, pp. 205–209, 2017.

[147] V. Bertone, S. Carrazza, and J. Rojo, "APFEL: A PDF Evolution Library with QED corrections," *Comput. Phys. Commun.*, vol. 185, pp. 1647–1668, 2014.

[148] A. Candido, S. Forte, and F. Hekhorn, "Can $\overline{MS}$ parton distributions be negative?," *Journal of High Energy Physics*, vol. 2020, Nov 2020.

[149] S. Forte, "The Gottfried sum rule and the light flavor content of the nucleon," *Phys. Rev. D*, vol. 47, pp. 1842–1853, 1993.

[150] Z. Kassabov, "Reportengine: A framework for declarative data analysis," Feb. 2019.

[151] S. Carrazza, E. R. Nocera, C. Schwan, and M. Zaro, "PineAPPL: combining EW and QCD corrections for fast evaluation of LHC processes," *JHEP*, vol. 12, p. 108, 2020.

[152] M. Mernik, J. Heering, and A. M. Sloane, "When and how to develop domain-specific languages," *ACM Comput. Surv.*, vol. 37, p. 316–344, dec 2005.

[153] S. Chatrchyan, V. Khachatryan, A. M. Sirunyan, A. Tumasyan, W. Adam, T. Bergauer, M. Dragicevic, J. Erö, C. Fabjan, and et al., "Measurement of the differential and double-differential drell-yan cross sections in proton-proton collisions at $\sqrt{s} = 7$ tev," *Journal of High Energy Physics*, vol. 2013, Dec 2013.

[154] G. Moreno, C. N. Brown, W. E. Cooper, D. Finley, Y. B. Hsiung, A. M. Jonckheere, H. Jostlein, D. M. Kaplan, L. M. Lederman, Y. Hemmi, K. Imai, K. Miyake, T. Nakamura, N. Sasao, N. Tamura, T. Yoshida, A. Maki, Y. Sakai, R. Gray, K. B. Luk, J. P. Rutherfoord, P. B. Straub, R. W. Williams, K. K. Young, M. R. Adams, H. Glass, D. Jaffe, R. L. McCarthy, J. A. Crittenden, and S. R. Smith, "Dimuon production in proton-copper collisions at $\sqrt{s} = 38.8$ gev," *Phys. Rev. D*, vol. 43, pp. 2815–2835, May 1991.

[155] M. Bonvini, S. Marzani, J. Rojo, L. Rottoli, M. Ubiali, R. D. Ball, V. Bertone, S. Carrazza, and N. P. Hartland, "Parton distributions with threshold resummation," *Journal of High Energy Physics*, vol. 2015, Sep 2015.

[156] R. Abdul Khalek *et al.*, "Parton Distributions with Theory Uncertainties: General Formalism and First Phenomenological Studies," *Eur. Phys. J. C*, vol. 79, no. 11, p. 931, 2019.

[157] R. Abdul Khalek *et al.*, "A first determination of parton distributions with theoretical uncertainties," *Eur. Phys. J.*, vol. C, p. 79:838, 2019.

[158] H. Abramowicz *et al.*, "Combination of measurements of inclusive deep inelastic $e^{\pm}p$ scattering cross sections and QCD analysis of HERA data," *Eur. Phys. J. C*, vol. 75, no. 12, p. 580, 2015.

[159] S. Chatrchyan *et al.*, "Measurement of the Differential and Double-Differential Drell-Yan Cross Sections in Proton-Proton Collisions at $\sqrt{s} = 7$ TeV," *JHEP*, vol. 12, p. 030, 2013.

[160] S. Bailey, T. Cridge, L. A. Harland-Lang, A. D. Martin, and R. S. Thorne, "Parton distributions from LHC, HERA, Tevatron and fixed target data: MSHT20 PDFs," *Eur. Phys. J. C*, vol. 81, no. 4, p. 341, 2021.

[161] R. D. Ball, V. Bertone, M. Bonvini, S. Marzani, J. Rojo, and L. Rottoli, "Parton distributions with small-x resummation: evidence for BFKL dynamics in HERA data," *Eur. Phys. J. C*, vol. 78, no. 4, p. 321, 2018.

[162] L. N. Lipatov, "Reggeization of the Vector Meson and the Vacuum Singularity in Nonabelian Gauge Theories," *Sov. J. Nucl. Phys.*, vol. 23, pp. 338–345, 1976.

[163] E. A. Kuraev, L. N. Lipatov, and V. S. Fadin, "Multi - Reggeon Processes in the Yang-Mills Theory," *Sov. Phys. JETP*, vol. 44, pp. 443–450, 1976.

[164] E. A. Kuraev, L. N. Lipatov, and V. S. Fadin, "The Pomeranchuk Singularity in Nonabelian Gauge Theories," *Sov. Phys. JETP*, vol. 45, pp. 199–204, 1977.

[165] I. I. Balitsky and L. N. Lipatov, "The Pomeranchuk Singularity in Quantum Chromodynamics," *Sov. J. Nucl. Phys.*, vol. 28, pp. 822–829, 1978.

[166] E. L. Berger, M. Guzzi, H.-L. Lai, P. M. Nadolsky, and F. I. Olness, "Constraints on color-octet fermions from a global parton distribution analysis," *Phys. Rev. D*, vol. 82, p. 114023, 2010.

[167] R. Abdul Khalek, J. J. Ethier, J. Rojo, and G. van Weelden, "nNNPDF2.0: quark flavor separation in nuclei from LHC data," *JHEP*, vol. 09, p. 183, 2020.

[168] R. A. Khalek, R. Gauld, T. Giani, E. R. Nocera, T. R. Rabemananjara, and J. Rojo, "nNNPDF3.0: Evidence for a modified partonic structure in heavy nuclei," 1 2022.

[169] B. Adams *et al.*, "Letter of Intent: A New QCD facility at the M2 beam line of the CERN SPS (COMPASS++/AMBER)," 8 2018.

[170] M. Aaboud *et al.*, "Measurement of the W-boson mass in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector," *Eur. Phys. J. C*, vol. 78, no. 2, p. 110, 2018. [Erratum: Eur.Phys.J.C 78, 898 (2018)].

[171] A. M. Sirunyan *et al.*, "Measurement of the weak mixing angle using the forward-backward asymmetry of Drell-Yan events in pp collisions at 8 TeV," *Eur. Phys. J. C*, vol. 78, no. 9, p. 701, 2018.

[172] G. Aad *et al.*, "Determination of the strange quark density of the proton from ATLAS measurements of the $W \to \ell\nu$ and $Z \to \ell\ell$ cross sections," *Phys. Rev. Lett.*, vol. 109, p. 012001, 2012.

[173] M. Aaboud *et al.*, "Precision measurement and interpretation of inclusive W^+ , W^- and Z/γ^* production cross sections with the ATLAS detector," *Eur. Phys. J. C*, vol. 77, no. 6, p. 367, 2017.

[174] R. D. Ball, S. Carrazza, L. Del Debbio, S. Forte, Z. Kassabov, J. Rojo, E. Slade, and M. Ubiali, "Precision determination of the strong coupling constant within a global PDF analysis," *Eur. Phys. J. C*, vol. 78, no. 5, p. 408, 2018.

[175] M. Goncharov *et al.*, "Precise Measurement of Dimuon Production Cross-Sections in ν_μ Fe and $\bar{\nu}_\mu$ Fe Deep Inelastic Scattering at the Tevatron.," *Phys. Rev. D*, vol. 64, p. 112006, 2001.

[176] D. Mason *et al.*, "Measurement of the Nucleon Strange-Antistrange Asymmetry at Next-to-Leading Order in QCD from NuTeV Dimuon Data," *Phys. Rev. Lett.*, vol. 99, p. 192001, 2007.

[177] P. Zyla *et al.*, "Review of Particle Physics," *PTEP*, vol. 2020, no. 8, p. 083C01, 2020.

[178] T. A. Aaltonen *et al.*, "Measurement of $d\sigma/dy$ of Drell-Yan e^+e^- pairs in the Z Mass Region from $p\bar{p}$ Collisions at $\sqrt{s} = 1.96$ TeV," *Phys. Lett. B*, vol. 692, pp. 232–239, 2010.

[179] V. M. Abazov *et al.*, "Measurement of the Shape of the Boson Rapidity Distribution for $p\bar{p} \to Z/\gamma^* \to e^+e^- + X$ Events Produced at $\sqrt{s}$ of 1.96 TeV," *Phys. Rev. D*, vol. 76, p. 012003, 2007.

[180] G. Aad *et al.*, "Measurement of the inclusive $W^\pm$ and Z/γ cross sections in the electron and muon decay channels in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector," *Phys. Rev. D*, vol. 85, p. 072004, 2012.

[181] V. Khachatryan *et al.*, "Measurement of the differential cross section and charge asymmetry for inclusive pp $\to \mathrm{W}^\pm + X$ production at $\sqrt{s} = 8$ TeV," *Eur. Phys. J. C*, vol. 76, no. 8, p. 469, 2016.

[182] O. Samoylov *et al.*, "A Precision Measurement of Charm Dimuon Production in Neutrino Interactions from the NOMAD Experiment," *Nucl. Phys. B*, vol. 876, pp. 339–375, 2013.

[183] G. Aad, T. Abajyan, B. Abbott, J. Abdallah, S. Abdel Khalek, O. Abdinov, R. Aben, B. Abi, M. Abolins, and et al., "Measurement of the production of a W boson in association with a charm quark in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector," *Journal of High Energy Physics*, vol. 2014, May 2014.

[184] S. Chatrchyan *et al.*, "Measurement of Associated W + Charm Production in pp Collisions at $\sqrt{s} = 7$ TeV," *JHEP*, vol. 02, p. 013, 2014.

[185] A. M. Sirunyan *et al.*, "Measurement of associated production of a W boson and a charm quark in proton-proton collisions at $\sqrt{s} = 13$ TeV," *Eur. Phys. J. C*, vol. 79, no. 3, p. 269, 2019.

[186] M. Aaboud *et al.*, "Measurement of differential cross sections and W^+/W^- cross-section ratios for W boson production in association with jets at $\sqrt{s} = 8$ TeV with the ATLAS detector," *JHEP*, vol. 05, p. 077, 2018. [Erratum: JHEP 10, 048 (2020)].

[187] A. Metz and A. Vossen, "Parton Fragmentation Functions," *Prog. Part. Nucl. Phys.*, vol. 91, pp. 136–202, 2016.

[188] J. Collins, *Foundations of perturbative QCD*, vol. 32. Cambridge University Press, 11 2013.

[189] P. D. B. Collins and T. P. Spiller, "The Fragmentation of Heavy Quarks," *J. Phys. G*, vol. 11, p. 1289, 1985.

[190] N. Ushida *et al.*, "Production Characteristics of Charmed Particles in Neutrino Interactions," *Phys. Lett. B*, vol. 206, pp. 380–384, 1988.

[191] J. Gao, "Massive charged-current coefficient functions in deep-inelastic scattering at NNLO and impact on strange-quark distributions," *JHEP*, vol. 02, p. 026, 2018.

[192] E. L. Berger, J. Gao, C. S. Li, Z. L. Liu, and H. X. Zhu, "Charm-Quark Production in Deep-Inelastic Neutrino Scattering at Next-to-Next-to-Leading Order in QCD," *Phys. Rev. Lett.*, vol. 116, no. 21, p. 212002, 2016.

[193] J. J. Ethier and E. R. Nocera, "Parton Distributions in Nucleons and Nuclei," *Ann. Rev. Nucl. Part. Sci.*, vol. 70, pp. 43–76, 2020.

[194] R. D. Ball, E. R. Nocera, and R. L. Pearson, "Nuclear Uncertainties in the Determination of Proton PDFs," *Eur. Phys. J. C*, vol. 79, no. 3, p. 282, 2019.

[195] R. Boughezal, J. M. Campbell, R. K. Ellis, C. Focke, W. Giele, X. Liu, F. Petriello, and C. Williams, "Color singlet production at NNLO in MCFM," *Eur. Phys. J. C*, vol. 77, no. 1, p. 7, 2017.

[196] T. Carli, D. Clements, A. Cooper-Sarkar, C. Gwenlan, G. P. Salam, F. Siegert, P. Starovoitov, and M. Sutton, "A posteriori inclusion of parton density functions in NLO QCD final-state calculations at hadron colliders: The APPLGRID Project," *Eur. Phys. J. C*, vol. 66, pp. 503–524, 2010.

[197] R. Gavin, Y. Li, F. Petriello, and S. Quackenbush, "FEWZ 2.0: A code for hadronic Z production at next-to-next-to-leading order," *Comput. Phys. Commun.*, vol. 182, pp. 2388–2403, 2011.

[198] R. Boughezal, C. Focke, X. Liu, and F. Petriello, "W-boson production in association with a jet at next-to-next-to-leading order in perturbative QCD," *Phys. Rev. Lett.*, vol. 115, no. 6, p. 062002, 2015.

[199] A. Gehrmann-De Ridder, T. Gehrmann, E. W. N. Glover, A. Huss, and T. A. Morgan, "Precise QCD predictions for the production of a Z boson in association with a hadronic jet," *Phys. Rev. Lett.*, vol. 117, no. 2, p. 022001, 2016.

[200] M. Czakon, A. Mitov, M. Pellen, and R. Poncelet, "NNLO QCD predictions for W+c-jet production at the LHC," *JHEP*, vol. 06, p. 100, 2021.

[201] R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, and M. Ubiali, "Reweighting NNPDFs: the W lepton asymmetry," *Nucl. Phys. B*, vol. 849, pp. 112–143, 2011. [Erratum: Nucl.Phys.B 854, 926–927 (2012), Erratum: Nucl.Phys.B 855, 927–928 (2012)].

[202] R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, A. Guffanti, N. P. Hartland, J. I. Latorre, J. Rojo, and M. Ubiali, "Reweighting and Unweighting of Parton Distributions and the LHC W lepton asymmetry data," *Nucl. Phys. B*, vol. 855, pp. 608–638, 2012.

[203] S. Alekhin *et al.*, "HERAFitter," *Eur. Phys. J. C*, vol. 75, no. 7, p. 304, 2015.

[204] D. Bourilkov, R. C. Group, and M. R. Whalley, "LHAPDF: PDF use from the Tevatron to the LHC," in *TeV4LHC Workshop - 4th meeting*, 5 2006.

[205] R. D. Ball, V. Bertone, M. Bonvini, S. Carrazza, S. Forte, A. Guffanti, N. P. Hartland, J. Rojo, and L. Rottoli, "A Determination of the Charm Content of the Proton," *Eur. Phys. J. C*, vol. 76, no. 11, p. 647, 2016.

[206] Y. Avni, "Energy spectra of x-ray clusters of galaxies," *The Astrophysical Journal*, vol. 210, pp. 642–646, 1976.

[207] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*. Cambridge University Press, second ed., 1992.

[208] A. Meurer, C. P. Smith, M. Paprocki, O. Čertík, S. B. Kirpichev, M. Rocklin, A. Kumar, S. Ivanov, J. K. Moore, S. Singh, T. Rathnayake, S. Vig, B. E. Granger, R. P. Muller, F. Bonazzi, H. Gupta, S. Vats, F. Johansson, F. Pedregosa, M. J. Curry, A. R. Terrel, v. Roučka, A. Saboo, I. Fernando, S. Kulal, R. Cimrman, and A. Scopatz, "Sympy: symbolic computing in python," *PeerJ Computer Science*, vol. 3, p. e103, Jan. 2017.

[209] B. Efron, "Bootstrap methods: another look at the jackknife," in *Breakthroughs in statistics*, pp. 569–593, Springer, 1992.

[210] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The journal of chemical physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[211] W. K. Hastings, *Monte Carlo sampling methods using Markov chains and their applications*. Oxford University Press, 1970.

[212] Z. Han and W. Skiba, "Effective theory analysis of precision electroweak data," *Phys. Rev. D*, vol. 71, p. 075009, 2005.

[213] M. Arneodo *et al.*, "Accurate measurement of F2(d) / F2(p) and R^d - R^p," *Nucl. Phys. B*, vol. 487, pp. 3–26, 1997.

[214] M. Arneodo *et al.*, "Measurement of the proton and deuteron structure functions, F2(p) and F2(d), and of the ratio sigma-L / sigma-T," *Nucl. Phys. B*, vol. 483, pp. 3–43, 1997.

[215] L. W. Whitlow, E. M. Riordan, S. Dasu, S. Rock, and A. Bodek, "Precise measurements of the proton and deuteron structure functions from a global analysis of the SLAC deep inelastic electron scattering cross-sections," *Phys. Lett. B*, vol. 282, pp. 475–482, 1992.

[216] A. C. Benvenuti *et al.*, "A High Statistics Measurement of the Proton Structure Functions $F_2(x, Q^2)$ and R from Deep Inelastic Muon Scattering at High Q^2," *Phys. Lett. B*, vol. 223, pp. 485–489, 1989.

[217] G. Onengut *et al.*, "Measurement of nucleon structure functions in neutrino scattering," *Phys. Lett. B*, vol. 632, pp. 65–75, 2006.

[218] H. Abramowicz *et al.*, "Combination and QCD analysis of charm and beauty production cross-section measurements in deep inelastic *ep* scattering at HERA," *Eur. Phys. J. C*, vol. 78, no. 6, p. 473, 2018.

[219] N. P. Hartland, F. Maltoni, E. R. Nocera, J. Rojo, E. Slade, E. Vryonidou, and C. Zhang, "A Monte Carlo global analysis of the Standard Model Effective Field Theory: the top quark sector," *JHEP*, vol. 04, p. 100, 2019.

[220] R. S. Towell *et al.*, "Improved measurement of the anti-d / anti-u asymmetry in the nucleon sea," *Phys. Rev. D*, vol. 64, p. 052002, 2001.

[221] J. C. Webb *et al.*, "Absolute Drell-Yan dimuon cross-sections in 800 GeV / c pp and pd collisions," *AIP Conference Proceedings*, vol. 698, no. 1, pp. 1–12, 2004.

[222] J. C. Webb, *Measurement of continuum dimuon production in 800-GeV/C proton nucleon collisions*. PhD thesis, New Mexico State U., 2003.

[223] V. M. Abazov *et al.*, "Measurement of the Muon Charge Asymmetry in $p\bar{p} \to$ W+X $\to \mu\nu$ + X Events at $\sqrt{s}$=1.96 TeV," *Phys. Rev. D*, vol. 88, p. 091102, 2013.

[224] G. Aad *et al.*, "Measurement of the low-mass Drell-Yan differential cross section at $\sqrt{s} = 7$ TeV using the ATLAS detector," *JHEP*, vol. 06, p. 112, 2014.

[225] G. Aad *et al.*, "Measurement of the production of a W boson in association with a charm quark in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector," *JHEP*, vol. 05, p. 068, 2014.

[226] G. Aad *et al.*, "Measurement of the transverse momentum and ϕ^*_η distributions of Drell–Yan lepton pairs in proton–proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector," *Eur. Phys. J. C*, vol. 76, no. 5, p. 291, 2016.

[227] S. Chatrchyan *et al.*, "Measurement of the Electron Charge Asymmetry in Inclusive W Production in pp Collisions at $\sqrt{s} = 7$ TeV," *Phys. Rev. Lett.*, vol. 109, p. 111806, 2012.

[228] V. Khachatryan *et al.*, "Measurement of the Z boson differential cross section in transverse momentum and rapidity in proton–proton collisions at 8 TeV," *Phys. Lett. B*, vol. 749, pp. 187–209, 2015.

[229] R. Aaij *et al.*, "Inclusive W and Z production in the forward region at $\sqrt{s} = 7$ TeV," *JHEP*, vol. 06, p. 058, 2012.

[230] R. Aaij *et al.*, "Measurement of the forward Z boson production cross-section in pp collisions at $\sqrt{s} = 7$ TeV," *JHEP*, vol. 08, p. 039, 2015.

[231] R. Aaij *et al.*, "Measurement of the cross-section for $Z \to e^+e^-$ production in pp collisions at $\sqrt{s} = 7$ TeV," *JHEP*, vol. 02, p. 106, 2013.

[232] R. Aaij *et al.*, "Measurement of forward W and Z boson production in pp collisions at $\sqrt{s} = 8$ TeV," *JHEP*, vol. 01, p. 155, 2016.

[233] G. Aad *et al.*, "Measurement of the high-mass Drell–Yan differential cross-section in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector," *Phys. Lett. B*, vol. 725, pp. 223–242, 2013.

[234] G. Aad *et al.*, "Measurement of the double-differential high-mass Drell-Yan cross section in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector," *JHEP*, vol. 08, p. 009, 2016.

[235] V. Khachatryan *et al.*, "Measurements of differential and double-differential Drell-Yan cross sections in proton-proton collisions at 8 TeV," *Eur. Phys. J. C*, vol. 75, no. 4, p. 147, 2015.

[236] A. M. Sirunyan *et al.*, "Measurement of the differential Drell-Yan cross section in proton-proton collisions at $\sqrt{s} = 13$ TeV," *JHEP*, vol. 12, p. 059, 2019.

[237] R. Barbieri, A. Pomarol, R. Rattazzi, and A. Strumia, "Electroweak symmetry breaking after LEP-1 and LEP-2," *Nucl. Phys. B*, vol. 703, pp. 127–146, 2004.

[238] J. D. Wells and Z. Zhang, "Effective theories of universal theories," *JHEP*, vol. 01, p. 123, 2016.

[239] J. D. Wells and Z. Zhang, "Renormalization group evolution of the universal theories EFT," *JHEP*, vol. 06, p. 122, 2016.

[240] E. Salvioni, G. Villadoro, and F. Zwirner, "Minimal Z-prime models: Present bounds and early LHC reach," *JHEP*, vol. 11, p. 068, 2009.

[241] F. del Aguila, J. de Blas, and M. Perez-Victoria, "Electroweak Limits on General New Vector Bosons," *JHEP*, vol. 09, p. 033, 2010.

[242] J. de Blas, J. M. Lizana, and M. Perez-Victoria, "Combining searches of Z' and W' bosons," *JHEP*, vol. 01, p. 166, 2013.

[243] D. Pappadopulo, A. Thamm, R. Torre, and A. Wulzer, "Heavy Vector Triplets: Bridging Theory and Data," *JHEP*, vol. 09, p. 060, 2014.

[244] M. Cirelli, N. Fornengo, and A. Strumia, "Minimal dark matter," *Nucl. Phys. B*, vol. 753, pp. 178–194, 2006.

[245] A. Djouadi and P. Gambino, "Electroweak gauge bosons selfenergies: Complete QCD corrections," *Phys. Rev. D*, vol. 49, pp. 3499–3511, 1994. [Erratum: Phys.Rev.D 53, 4111 (1996)].

[246] M. E. Peskin and T. Takeuchi, "Estimation of oblique electroweak corrections," *Phys. Rev. D*, vol. 46, pp. 381–409, 1992.

[247] G. Altarelli, R. Barbieri, and S. Jadach, "Toward a model independent analysis of electroweak data," *Nucl. Phys. B*, vol. 369, pp. 3–32, 1992. [Erratum: Nucl.Phys.B 376, 444 (1992)].

[248] G. Altarelli and R. Barbieri, "Vacuum polarization effects of new physics on electroweak processes," *Phys. Lett. B*, vol. 253, pp. 161–167, 1991.

[249] M. Farina, G. Panico, D. Pappadopulo, J. T. Ruderman, R. Torre, and A. Wulzer, "Energy helps accuracy: electroweak precision tests at hadron colliders," *Phys. Lett. B*, vol. 772, pp. 210–215, 2017.

[250] I. Brivio, Y. Jiang, and M. Trott, "The SMEFTsim package, theory and tools," *JHEP*, vol. 12, p. 070, 2017.

[251] A. Greljo, P. Stangl, and A. E. Thomsen, "A model of muon anomalies," *Phys. Lett. B*, vol. 820, p. 136554, 2021.

[252] M. Aaboud *et al.*, "Search for new high-mass phenomena in the dilepton final state using 36 fb^{-1} of proton-proton collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector," *JHEP*, vol. 10, p. 182, 2017.

[253] R. Abdul Khalek, S. Bailey, J. Gao, L. Harland-Lang, and J. Rojo, "Towards Ultimate Parton Distributions at the High-Luminosity LHC," *Eur. Phys. J. C*, vol. 78, no. 11, p. 962, 2018.

[254] P. Azzi *et al.*, "Report from Working Group 1: Standard Model Physics at the HL-LHC and HE-LHC," *CERN Yellow Rep. Monogr.*, vol. 7, pp. 1–220, 2019.

[255] M. Cepeda *et al.*, "Report from Working Group 2: Higgs Physics at the HL-LHC and HE-LHC," *CERN Yellow Rep. Monogr.*, vol. 7, pp. 221–584, 2019.

[256] S. Forte and Z. Kassabov, "Why α_s cannot be determined from hadronic processes without simultaneously determining the parton distributions," *Eur. Phys. J. C*, vol. 80, no. 3, p. 182, 2020.

[257] L. Del Debbio, T. Giani, and M. Wilson, "Bayesian Approach to Inverse Problems: an Application to NNPDF Closure Testing," 11 2021.

[258] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind, "Automatic differentiation in machine learning: a survey," *Journal of Marchine Learning Research*, vol. 18, pp. 1–43, 2018.

[259] R. Torre, L. Ricci, and A. Wulzer, "On the W&Y interpretation of high-energy Drell-Yan measurements," *JHEP*, vol. 02, p. 144, 2021.

[260] S. Carrazza and J. Cruz-Martinez, "Towards a new generation of parton densities with deep learning models," *Eur. Phys. J. C*, vol. 79, no. 8, p. 676, 2019.

[261] R. D. Ball, E. R. Nocera, and R. L. Pearson, "Deuteron Uncertainties in the Determination of Proton PDFs," *Eur. Phys. J. C*, vol. 81, no. 1, p. 37, 2021.

[262] C. M. Bishop *et al.*, *Neural networks for pattern recognition*. Oxford university press, 1995.

[263] J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *International conference on machine learning*, pp. 115–123, PMLR, 2013.

[264] R. Boughezal, E. Mereghetti, and F. Petriello, "Dilepton production in the SMEFT at O(1/Λ4)," *Phys. Rev. D*, vol. 104, no. 9, p. 095022, 2021.

[265] C. Grojean, E. E. Jenkins, A. V. Manohar, and M. Trott, "Renormalization Group Scaling of Higgs Operators and $h \to \gamma\gamma$ Decay," *JHEP*, vol. 04, p. 016, 2013.

[266] J. Baglio, S. Dawson, and S. Homiller, "QCD corrections in Standard Model EFT fits to WZ and WW production," *Phys. Rev. D*, vol. 100, no. 11, p. 113010, 2019.

[267] C. Degrande, G. Durieux, F. Maltoni, K. Mimasu, E. Vryonidou, and C. Zhang, "Automated one-loop computations in the standard model effective field theory," *Phys. Rev. D*, vol. 103, no. 9, p. 096024, 2021.

[268] S. Dawson and P. P. Giardino, "New physics through Drell-Yan standard model EFT measurements at NLO," *Phys. Rev. D*, vol. 104, no. 7, p. 073004, 2021.

[269] M. González-Alonso, J. Martin Camalich, and K. Mimouni, "Renormalization-group evolution of new physics contributions to (semi)leptonic meson decays," *Phys. Lett. B*, vol. 772, pp. 777–785, 2017.

[270] E. E. Jenkins, A. V. Manohar, and M. Trott, "Renormalization Group Evolution of the Standard Model Dimension Six Operators I: Formalism and lambda Dependence," *JHEP*, vol. 10, p. 087, 2013.

[271] E. E. Jenkins, A. V. Manohar, and M. Trott, "Renormalization Group Evolution of the Standard Model Dimension Six Operators II: Yukawa Dependence," *JHEP*, vol. 01, p. 035, 2014.

[272] J. Davighi, S. Melville, and T. You, "Natural selection rules: new positivity bounds for massive spinning particles," *JHEP*, vol. 02, p. 167, 2022.

[273] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, vol. 2, no. 11, pp. 559–572, 1901.

[274] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, p. 20150202, 2016.

[275] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, *et al.*, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[276] R. Alonso, E. E. Jenkins, A. V. Manohar, and M. Trott, "Renormalization Group Evolution of the Standard Model Dimension Six Operators III: Gauge Coupling Dependence and Phenomenology," *JHEP*, vol. 04, p. 159, 2014.

[277] A. Grozin, "Lectures on QED and QCD," in *3rd Dubna International Advanced School of Theoretical Physics*, 8 2005.

Appendix A

Low level implementations for NNPDF4.0

In this appendix we provide an overview of the low-level algorithmic implementations for the covariance matrix construction and Monte Carlo pseudodata generation used in NNPDF4.0. As mentioned in the main text, the actual implementations enjoy vectorized operations implemented natively by the `NumPy` and `Pandas` libraries, thus maintaining a great deal of the performance obtained in the legacy `C++` codebase while retaining readability.

A.1 Covariance matrix construction

The experimental covariance matrix can be constructed from the systematic uncertainty breakdown of a set of experimental measurements. A mathematical definition is given by equation 3.18 with a low level code implementation given by algorithm A.1.1. In the NNPDF4.0 methodology the implementation is in a purely `Python` framework enjoying not only readability, but also substantial performance gains thanks to using optimized libraries such as `NumPy` [275] and `Pandas`, reducing the covariance matrix computation down to order a few seconds, as opposed to minutes.

Note that the covariance matrix constructed using the t_0 prescription of equation 3.19 only requires a trivial modification of this algorithm.

Algorithm A.1.1 Algorithm to obtain the covariance between two datapoints given their uncertainty breakdown.

```
1: procedure COVARIANCE(datapoint1, datapoint2)
2:     if datapoint1 is datapoint2 then
3:         same ← True                                          ▷ Diagonal entry
4:     else
5:         same ← False                                         ▷ Off-diagonal entry
6:     end if
7:     covariance ← 0
8:     for (unc1, unc2) in (datapoint1.uncertainties, datapoint2.uncertainties) do
9:         if same then                  ▷ Add uncorrrolated uncertainties to diagonal
10:            covariance ← covariance + unc1.uncorr * unc2.uncorr
11:        end if
12:        if unc1 is additive then ▷ Both uncertainties have same type so only check first
13:            covariance ← covariance + unc1 * unc2
14:        else                                          ▷ Then it was multiplicative
15:            covariance ← covariance +
                            unc1 * datapoint1.central *
                            unc2 * datapoint2.central
16:        end if
17:    end for
18: end procedure
```

A.2 Monte Carlo pseudodata generation

The Monte Carlo data replicas have historically been a considerable bottleneck in the NNPDF fitting methodology. By replacing this algorithm, significant performance gains have been made. The low level code implementation is presented in algorithm A.2.1. A major bottleneck of the current implementation is the `while` condition which produces MC replicas until a positive replica is found. This ensures that all observables, that are not asymmetry measurements, are positive definite, but the stochastic approach employed to achieve this condition carries a large performance footprint. Alternative solutions include simply relaxing this condition or to sample from a truncated gaussian, both of which are left as future work.

Algorithm A.2.1 Algorithm for the generation of pseudodata. We assume access to a random number generator (`RNG`) capable of generating univariate standard normal random variables.

```
1: procedure GeneratePseudodata(datapoints)
2:     replicas ← [ ]                                          ▷ Empty list
3:     for datapoint in datapoints do
4:         positive ← False
5:         correlates ← { }                                    ▷ Empty hashtable
6:         while not positive do
7:             replica ← datapoint.central
8:             for uncertainty in datapoint.uncertainties.additive do
9:                 if uncertainty is in correlates then        ▷ Check if this uncertainty
                                                                 source has occurred be-
                                                                 fore
10:                    random ← correlates[uncertainty]
11:                else
12:                    random ← RNG( )
13:                    correlates[unceratinty] ← random
14:                end if
15:                replica ← replica + random * uncertainty
16:            end for                                         ▷ End additive loop
17:            mult ← 1          ▷ Order matters. Apply multplicative after additive
18:            for uncertainty in datapoint.uncertainties.multiplicative do
19:                if uncertainty is in correlates then
20:                    random ← correlates[uncertainty]
21:                else
22:                    random ← RNG( )
23:                    correlates[unceratinty] ← random
24:                end if
25:                mult ← mult + random * uncertainty
26:            end for                                         ▷ End multiplicative loop
27:            replica ← replica * mult
28:            if replica > 0 or datapoint is asymmetry then
29:                positive ← True
30:                replicas append replica
31:            end if
32:        end while                                           ▷ End positivity check loop
33:    end for                                                 ▷ End data points loop
34: end procedure
```

Appendix B

Data theory comparisons for inclusive W, Z-boson production

We present here data-theory comparisons for ATLAS inclusive gauge boson production [180, 173] which supplement the discussion of section 4.4. In figure B.3 we plot the data-theory comparison for inclusive W, Z gauge boson production. Comparisons are made for each of the gauge bosons and binned in (di)lepton rapidity from W (Z) decay. The central and forward rapidity distributions are shown in figures B.2 and B.1 respectively. For the Z-boson measurements, the data is provided in low-mass, on-shell, and high-mass dilepton invariant mass binnings (respectively Z_{low}, Z_{peak}, and Z_{high} in the figures).

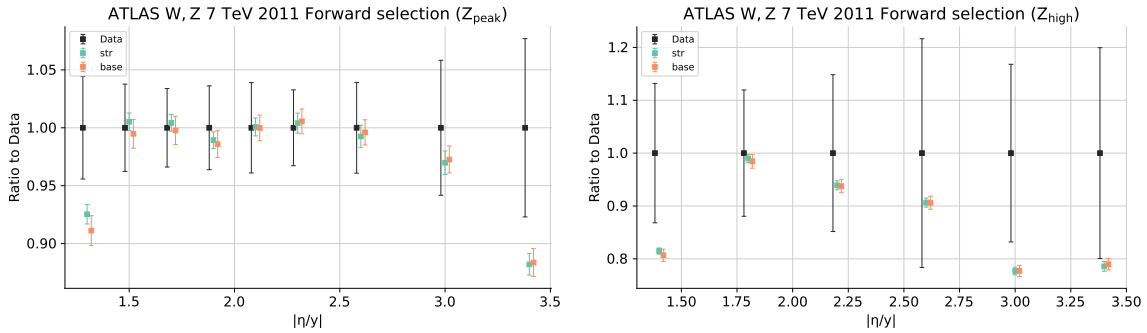


Figure B.1: Data-theory comparison for precision ATLAS inclusive gauge boson at 7 TeV [173]. We show here the forward rapidity selection. Values are normalized to the data central value. Only Z-boson measurements at the Z-peak and above are delivered for this dataset. Predictions are made using `str_base` (orange) and `str` (green).

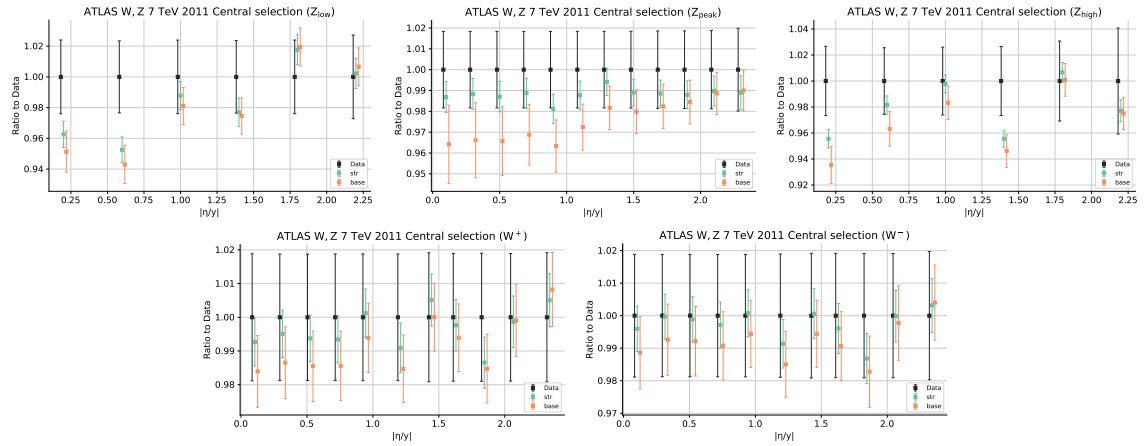


Figure B.2: Same as figure B.1 but for the central rapidity region. The top row shows the Z boson production for low-mass, on-shell, and high mass dilepton invariant mass bins (read left to right). The bottom row shows W^+ and W^- production respectively. The binnings are in (di)lepton rapidity from W (Z) decay.

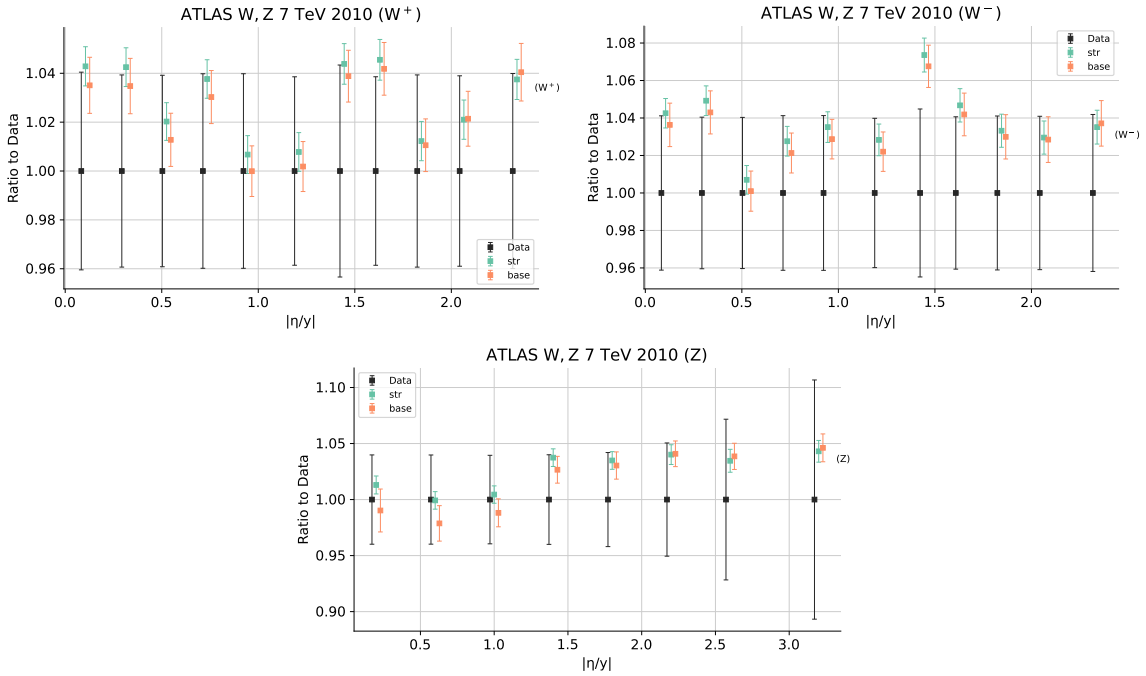


Figure B.3: Data-theory comparison for ATLAS inclusive gauge boson production at 7 TeV [180]. Values are reported for each gauge boson and measured in (di)lepton rapidity from W (Z) decay (from left to right, top to bottom, we have W^+, W^- and Z channels.) Values are normalized to the data central value. Predictions are made using `str_base` (orange) and `str` (green).

Appendix C

The QCD non-renormalization of Wilson coefficients at NLO

We here outline that the NLO pure QCD correction of the Wilson coefficients of equation 5.37 exactly cancels the one loop gluon self-energy. Note that the anomalous dimensions of dimension-6 operators are not in general vanishing if one considers Yukawa couplings [270, 271, 276], but here we concern ourselves only with QCD running.

Consider the following massless Yang-Mills theory with the addition of a 4-fermion contact interaction:

$$\mathcal{L} = \bar{q} i \not{D} q - \frac{1}{4} F^a_{\mu\nu} F^{\mu\nu a} + \frac{c}{4\Lambda^2} \bar{q}\gamma^\mu(1+\lambda'\gamma^5)q \; \bar{l}\gamma_\mu(1+\lambda\gamma^5)l \tag{C.1}$$

where the λ, λ' denote some chirality configuration and a runs over Lie algebra generators. If we naively use this theory to one loop order, then we will find that our loop integrals give divergences which must be regulated. To do so, we add to the bare lagrangian of equation C.1 counter-terms, $\delta\psi$ and $c\delta c$; the values of which will be determined in a way to match the residue of the poles corresponding to these UV divergences. The lagrangian with the appropriate counter-terms thus reads:

$$\mathcal{L} = (1+\delta\psi)\bar{q} i \not{D} q - \frac{1}{4} F^a_{\mu\nu} F^{\mu\nu a} + \frac{c + c\delta c}{4\Lambda^2} \bar{q}\gamma^\mu(1+\lambda'\gamma^5)q \; \bar{l}\gamma_\mu(1+\lambda\gamma^5)l. \tag{C.2}$$

By convention it is usual to rescale the fields such that the coefficient of the quark kinetic term is unity. We perform the field redefinition $q_R = \sqrt{1+\delta\psi}q$ and so in terms

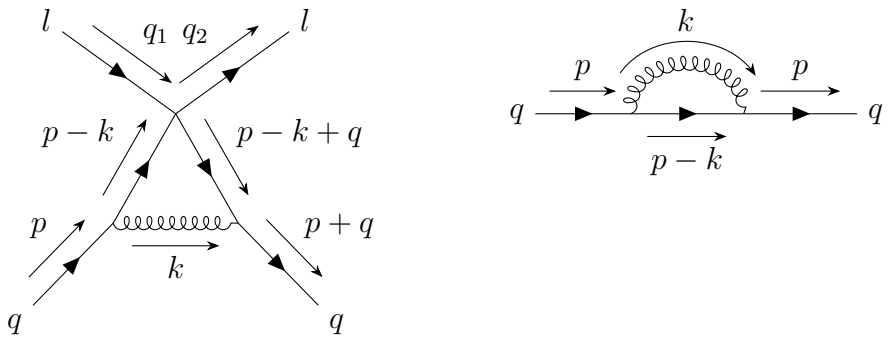


Figure C.1: The diagrams contributing to the running of the Wilson coefficient corresponding to the 4-fermion contact operator connecting leptons and quarks. Virtual emission (left) is shown to cancel exactly with the wavefunction renormalization of the quark self-energy (right).

of the renormalized fields and couplings we have:

$$\mathcal{L} = \bar{q_R} \not{D}\, q_R - \frac{1}{4} F^a_{\mu\nu} F^{\mu\nu a} + \frac{1+\delta c}{1+\delta\psi} \frac{c}{4\Lambda^2}\, \bar{q_R} \gamma^\mu (1+\lambda'\gamma^5) q_R\, \bar{l} \gamma_\mu (1+\lambda\gamma^5) l. \tag{C.3}$$

We shall demonstrate perturbatively that $\delta c = \delta\psi$ such that the Wilson coefficient receives no corrections at this order (charge conservation arguments suggest this holds at all orders, but this argument is omitted). The pure vertex correction diagram is shown in figure C.1 (left) as well as the quark propagator self-energy (right) at NLO in QCD. We work in $d = 4 - 2\epsilon$ dimensions using dimensional regularization [59] in order to regularize the divergence. The matrix element associated with the amputated diagram reads [1]:

$$i\mathcal{M}^\mu = \int \frac{d^d k}{(2\pi)^d} (-ig_s)^2 \mu^{2\epsilon} C_F \gamma_\nu \frac{i}{\not{p} + \not{q} - \not{k}} \left(-i \frac{c}{4\Lambda^2} \gamma^\mu (1+\lambda'\gamma^5) \right) \frac{i}{\not{p} - \not{q}} \gamma^\nu \left(\frac{-i}{k^2} \right) \tag{C.4}$$

where C_F is the colour factor arising from traces over products of Gell-Mann matrices and μ is a physically arbitrary mass scale introduced to ensure the gauge field coupling, g_s, is dimensionless. The notation follows the convention

$$\frac{1}{\not{p}} = \frac{\not{p}}{p^2}. \tag{C.5}$$

[1] We do not include the usual $u, \bar{u}$ spinor fields as we are concerned in the correction to the BSM vertex: which may be an operator insertion at an arbitrary position in a scattering graph.

The matrix element then simplifies to

$$i\mathcal{M}^\mu = -g_s^2\mu^{2\epsilon}C_F\int\frac{d^dk}{(2\pi)^d}\frac{\gamma_\nu(\not{p}+\not{q}-\not{k})\left(\frac{c}{4\Lambda^2}\gamma^\mu(1+\lambda'\gamma^5)\right)(\not{p}-\not{k})\gamma^\nu}{(p+q-k)^2(p-k)^2k^2}. \tag{C.6}$$

For notational clarity we will define $\mathcal{O}^\mu$ to capture the Lorentz structure of our BSM operator:

$$\mathcal{O}^\mu = \frac{c}{4\Lambda^2}\gamma^\mu(1+\lambda'\gamma^5). \tag{C.7}$$

The use of Feynman parameters allows us to separate the product in the denominator into a sum:

$$i\mathcal{M}^\mu = -2g_s^2\mu^{2\epsilon}C_F\int_0^1du_1\int_0^1du_2\int_0^1du_3\int\frac{d^dk}{(2\pi)^d} \frac{\gamma_\nu(\not{p}+\not{q}-\not{k})\mathcal{O}^\mu(\not{p}-\not{k})\gamma^\nu}{\Big((p+q-k)^2u_1+(p-k)^2u_2+k^2u_3\Big)^3}\delta(1-u_1-u_2-u_3). \tag{C.8}$$

Using this trick allows us to write the denominator as:

$$u_1(p^2+q^2+k^2+2p\cdot q-2p\cdot k-2q\cdot k)+u_2(p^2+k^2-2p\cdot k)+u_3k^2 \tag{C.9}$$

which, after performing the u_3 integral, gives:

$$= k^2+u_1(p^2+q^2+2p\cdot q-2p\cdot k-2q\cdot k)+u_2(p^2-2p\cdot k) \tag{C.10}$$

and after completing the square, we obtain:

$$= (k-u_1p-u_1q-u_2p)^2-(u_1p+u_1q+u_2p)^2+(u_2p^2+u_1p^2+u_1q^2+2u_1p\cdot q). \tag{C.11}$$

Defining

$$\Delta \equiv (u_1p+u_1q+u_2p)^2-(u_2p^2+u_1p^2+u_1q^2+2u_1p\cdot q). \tag{C.12}$$

the diagram we are computing reads:

$$i\mathcal{M}^\mu = -2g_s^2\mu^{2\epsilon}C_F\int_0^1du_1\int_0^1du_2\int\frac{d^dk}{(2\pi)^d}\frac{\gamma_\nu(\not{p}+\not{q}-\not{k})\mathcal{O}^\mu(\not{p}-\not{k})\gamma^\nu}{\Big((k-u_1p-u_1q-u_2p)^2-\Delta\Big)^3}. \tag{C.13}$$

Focusing for now only on the integral and using translation invariance of the phase space measure, the integral reads

$$\int \cdots = \int_0^1 du_1 \int_0^1 du_2 \int \frac{d^d k}{(2\pi)^d} \frac{\gamma_\nu(\not{p} + \not{q} - \not{k} - u_1\not{P} - u_1\not{q} - u_2\not{p})\mathcal{O}^\mu(\not{p} - \not{k} - u_1\not{p} - u_1\not{q} - u_2\not{p})\gamma^\nu}{(k^2 - \Delta)^3} \tag{C.14}$$

Noting that the terms linear in k will result in an odd integrand and will thus evaluate to zero we get:

$$\int \cdots = \int_0^1 du_1 \int_0^1 du_2 \int \frac{d^d k}{(2\pi)^d} \left[\frac{\gamma_\nu \not{k} \mathcal{O}^\nu \not{k} \gamma^\mu}{(k^2 - \Delta)^3} + \frac{\gamma_\nu(\not{p} + \not{q} - u_1\not{p} - u_1\not{q} - u_2\not{p})\mathcal{O}^\mu(\not{p} - u_1\not{p} - u_1\not{q} - u_2\not{p})\gamma^\nu}{(k^2 - \Delta)^3} \right] \tag{C.15}$$

A simple power counting argument reveals the latter term in the integrand to be finite. We thus focus on the divergent part of the integral arising due to the former term. Examining the numerator of this term reveals the following simplification:

$$\begin{aligned} \gamma_\nu \not{k} \mathcal{O}^\mu \not{k} \gamma^\nu &= \frac{c}{4\Lambda^2} \gamma_\nu \not{k} \gamma^\mu (1 + \lambda' \gamma^5) \not{k} \gamma^\nu \\ &= \left((2 - 2\epsilon) k^2 \gamma^\mu - 2k^\mu (2 - 2\epsilon) k^\nu \gamma_\nu \right) \frac{c}{4\Lambda^2} (1 + \lambda' \gamma^5), \end{aligned} \tag{C.16}$$

where we have used $\not{k} \cdot \not{k} = k^2$ and $\gamma^\mu \gamma_\mu = 4 - 2\epsilon$ which can be obtained directly from the Clifford algebra of equation 2.3.

The first term in the parenthesis will give a term that is proportional to the Lorentz structure, $\mathcal{O}^\mu$, of our BSM operator. The right term will also do so after some further manipulation. Noting that this term lives inside the phase space integral, we isolate its contribution as:

$$I^{\alpha\beta} \equiv \int \frac{d^d k}{(2\pi)^d} \frac{k^\alpha k^\beta}{(k^2 - \Delta)^3} \tag{C.17}$$

Noting $I^{\alpha\beta} = I^{\beta\alpha}$ and that I is a Lorentz tensor, it must be the case that $I^{\alpha\beta} = A\eta^{\alpha\beta}$ for some constant A. Contracting both sides of the equation with η_{α_β} we find:

$$(4 - 2\epsilon)A = \int \frac{d^d k}{(2\pi)^d} \frac{k^2}{(k^2 - \Delta)^3} \tag{C.18}$$

We can use the master equation [100]:

$$\int \frac{d^dk}{(2\pi)^d}\frac{(k^2)^a}{(k^2-\Delta)^b} = \frac{(-1)^{b-a}i\Gamma(b-a-\frac{1}{2}d)\Gamma(a+\frac{1}{2}d)}{(4\pi)^{\frac{d}{2}}\Gamma(b)\Gamma\left(\frac{d}{2}\right)}\left(\frac{1}{\Delta}\right)^{b-a-\frac{d}{2}} \tag{C.19}$$

to solve for A giving:

$$A = \frac{i\Gamma(\epsilon)\Gamma(3-\epsilon)}{(4\pi)^{2-\epsilon}\Gamma(3)\Gamma(2-\epsilon)(4-2\epsilon)}\left(\frac{1}{\Delta}\right)^{\epsilon} \tag{C.20}$$

where the Γ function is defined as:

$$\Gamma(x) = \int_0^\infty z^{x-1}e^{-z}dz \tag{C.21}$$

and famously evaluates to $(n-1)!$ for integer arguments. Returning to our integral and using $\Gamma(3) = 2$, we thus find:

$$\int \frac{d^dk}{(2\pi)^d}\frac{\gamma_\mu \not{k}\mathcal{O}^\mu \not{k}\gamma^\mu}{(k^2-\Delta)^3} = \frac{i(2\epsilon-2)\Gamma(\epsilon)\Gamma(3-\epsilon)}{2(4\pi)^{2-\epsilon}\Gamma(2-\epsilon)(4-2\epsilon)}\left(\frac{1}{\Delta}\right)^{\epsilon}\mathcal{O}^\mu, \tag{C.22}$$

but we must remember that this was the integrand of a $du_1\,du_2$ integral thus giving:

$$i\mathcal{M}^\mu = -2g_s^2\mu^{2\epsilon}C_F\frac{i(2\epsilon-2)\Gamma(\epsilon)\Gamma(3-\epsilon)}{2(4\pi)^{2-\epsilon}\Gamma(2-\epsilon)(4-2\epsilon)}\mathcal{O}^\mu\int_0^1 du_1\int_0^1 du_2\left(\frac{1}{\Delta}\right)^{\epsilon} + \text{finite terms} \tag{C.23}$$

but the integrand

$$\left(\frac{1}{\Delta}\right)^{\epsilon} = e^{-\epsilon\log\Delta} = 1 + O(\epsilon) \tag{C.24}$$

will not contribute to the divergent part of this diagram. Taking the limit of $\epsilon \to 0$ and using the definition $\alpha_s \equiv \frac{g_s^2}{4\pi}$ we have:

$$i\mathcal{M}^\mu = i\frac{\alpha_s C_F}{4\pi\epsilon}\mathcal{O}^\mu + \text{finite terms.} \tag{C.25}$$

We see that the counter-term to add in the minimal subtraction (MS) scheme is thus:

$$\delta_C = \frac{\alpha_s C_F}{4\pi\epsilon} \tag{C.26}$$

precisely the same counter-term arising from the wavefunction renormalization of the quark propagator [277].

Appendix D

`SIMUnet` stability on replica number

When performing purely PDF fits to experimental data, it is often the case one will have to compute ~ 100 Monte Carlo PDF replicas in order to achieve a percent level, faithful, uncertainty estimation [134]. As the study presented in chapter 6 acts as a proof-of-concept for our `SIMUnet` methodology, we perform high statistic fits composed of 1000 MC replicas. However, with each replica requiring approximately 3 hours of compute time, one necessarily requires access to a cluster of nodes in order to asynchronously compute the roughly 3000 hours of total wall clock time that is needed for each fit. It is possible, however, to reduce this time by an order of magnitude by instead computing ~ 100 MC replicas per fit and in this appendix we show that doing so poses little risk in underestimating the statistics when compared to a high replica fit.

From each of our high replica fits we randomly sample, without replacement, a set of 100 replicas thereby effectively emulating the scenario whereby the user would have performed a low statistics fit. In figure D.1 we plot the $q\bar{q}$ luminosity for the various SMEFT scenarios considered in this work. We plot the low replica luminosity normalized to the analogous high replica set. We see that the luminosities remain virtually identical, with no discernible difference at the ensemble level. Such behaviour is inherited directly from NNPDF4.0 where a typical fit will typically only possess 100 MC replicas after the post-fit selection has filtered poorly performing replicas.

In figure D.2 we present the distributions of best-fit W and Y values using both the high replica and reduced sets. We see that the distribution of best fit Wilson coefficients are accurately reproduced by the low statistics set: implying that, had we stopped the fitting process with 100 MC replicas, the additional 900 replicas would have changed the ensemble statistics such as the mean, standard deviation, or bounds

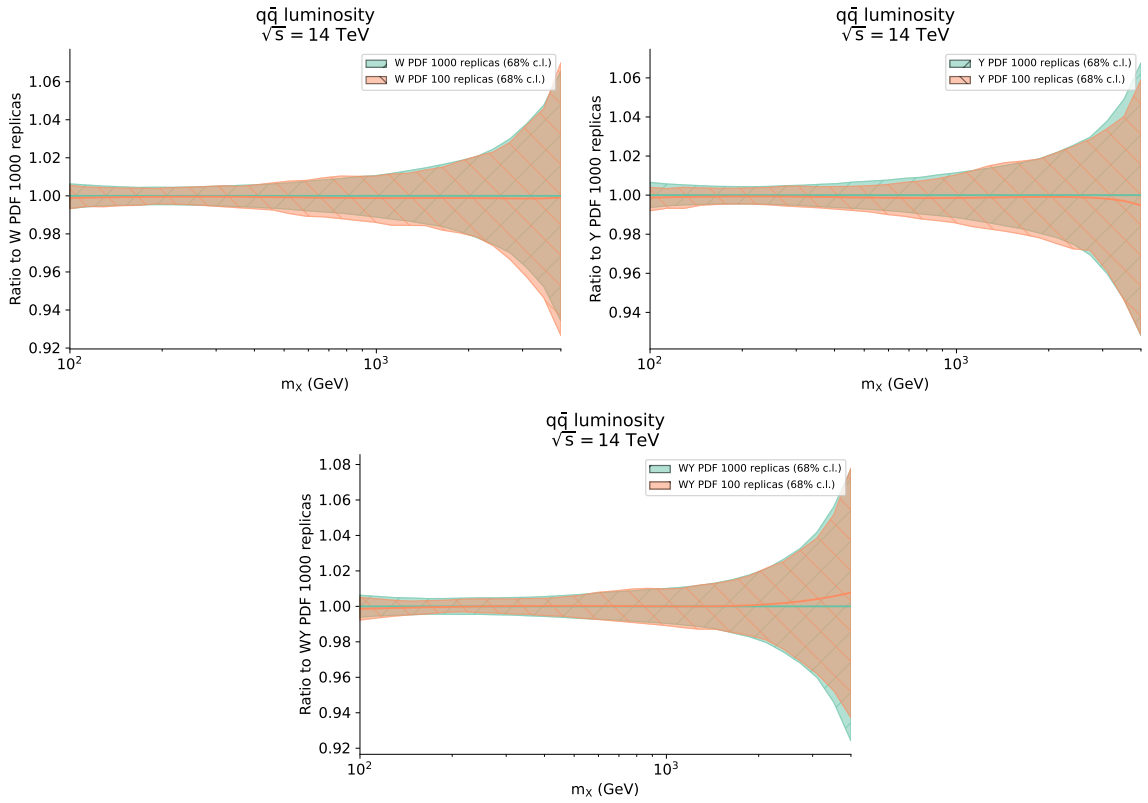


Figure D.1: The $q\bar{q}$ luminosity of the 100 replica (low statistics) fits normalized to the 1000 replica (high statistics) fits for various SMEFT scenarios considered in this work. Shown in the top left (top right) is the individual W (Y) scenario, while the lower panel is the combined (W, Y) scenario with HL-LHC projections being used in the fit.

very little. This is a typical pattern in the Monte Carlo approach to PDF fitting, whereby one quickly reaches saturation after ~ 100 MC replicas and further replicas only serve to accurately reproduce the experimental correlations.

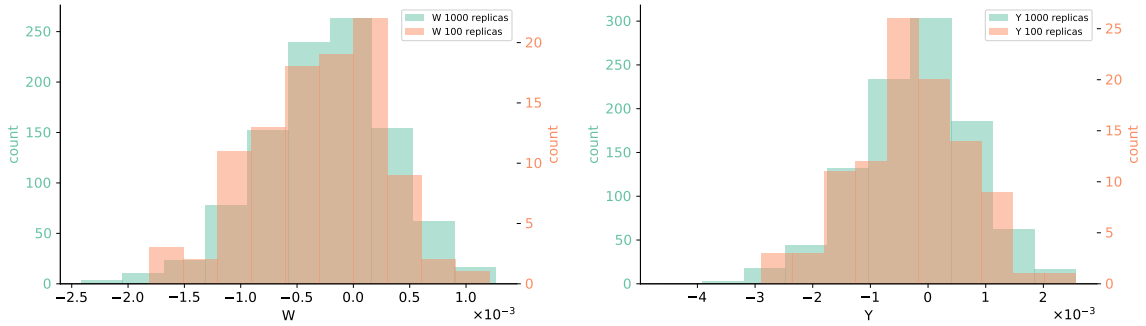


Figure D.2: Distributions of the best fit W (left) and Y (right) parameters using both high ($N_{\text{rep}} = 1000$) and low ($N_{\text{rep}} = 100$) statistic fits. We bring the readers attention to the different y axes for the histogram overlays.