UNIVERSITÀ DEGLI STUDI DI MILANO

Scuola di Dottorato in Fisica, Astrofisica e Fisica Applicata

Dipartimento di Fisica

Corso di Dottorato in Fisica, Astrofisica e Fisica Applicata

Ciclo XXXV

Statistical learning for Standard Model phenomenology

Settore Scientifico Disciplinare FIS/02

Supervisore: Prof. Stefano Forte

Coordinatore: Prof. Matteo Paris

Tesi di Dottorato di:

Roy Stegeman

Anno Accademico 2021-2022

Committee of the final examination:

External Referees:
Dr. Frédéric Dreyer
Dr. Luca Rottoli

External Members:
Prof. Luigi Del Debbio
Prof. Andrea Wulzer

Internal Member:
Prof. Raoul Röntsch

Final examination:

December 19, 2022
Università degli Studi di Milano, Dipartimento di Fisica, Milano, Italy

MIUR subjects:

FIS/02 - Fisica Teorica, Modelli e Metodi Matematici

Funding information:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 740006.

Abstract

The focus of this thesis is the accurate determination of parton distribution functions (PDFs), with a particular emphasis on modern machine learning tools used within the NNPDF approach. We first present NNPDF4.0, currently the most recent and most precise set of PDFs based on a global dataset. We then provide suggestions for improvements to the machine learning tools used for the NNPDF4.0 determination, both in terms of parametrization and model selection. We discuss different sources of PDF uncertainty. First, we elucidate the nontrivial aspects of averaging over the space of PDF determinations by explicitly calculating the data-driven correlation between different sets of PDFs. Then, we lay out certain fundamental properties of the sampling as performed within NNPDF methodology through explicit examples, and discuss how one may gain insight into the results of a neural network fit despite it being a black box model. Finally, we show how the flexibility of the NNPDF methodology allows for it to be applied to problems other than PDF determination, in particular we present a determination of neutrino inelastic structure functions.

Acknowledgements

During my PhD I have had the fortune to meet many brilliant people who in one way or another played an important role in the development of this thesis. First and foremost of whom is my supervisor, Stefano Forte, who provided the opportunity and support to work on interesting problems in a highly collaborative environment. I am thankful for his guidance throughout the entire three years and for encouraging independence while still always being available to discuss any problem or idea. I would also like to thank Stefano Carrazza for his support to work on problems even if the path is not always clear from the beginning.

I would like to thank the members of the N3PDF group for the friendly work environment. Juan Cruz-Martinez, Jesus Urtasun Elizari, Tanjona Rabemananjara and Christoper Schwan, for welcoming as a member of the group. Alessandro Candido, Felix Hekhorn and Kirill Kudashkin, who arrived in Milan at the same time as me, I learned a lot from each of you. In particular I would like to thank Alessandro Candido, who started the PhD alongside me for figuring out many things together. Niccolò Laurenti and Andrea Barontini who joined the last year, thank you for making the office pleasant environment and the many morning coffees.

Special thanks go to Petra Dell'Arme without whose help I'm really not sure how I would have ever managed to navigate all the bureaucracy involved in obtaining a PhD. From guiding me through my application to the PhD school before I even arrived in Milan, to helping with the final little things as my PhD is about to end.

I am deeply indebted to my collaborators of the NNPDF collaboration. In particular to Maria Ubiali and Juan Rojo for inviting me to visit and collaborate with their local groups, and Richard Ball for granting me the opportunity to continue as a postdoc in Edinburgh.

Besides my NNPDF and N3PDF colleagues, I am grateful to the member of the Milan physics department and in particular the members of the phenomenology group with who I had the pleasure to interact during the past years; Alessandro Vicini, Giancarlo Ferrera and Raoul Röntsch.

The last year of my PhD would not have been the same without the many short breaks to have a coffee or lunch with Davide Morgante, Davide Maria Tagliabue, and Jacopo D'Alberto. Thank you also Davide for your patience with teaching me Italian.

Finally, I would like to thank my parents and sister Michelle for always supporting me in my endeavors.

Contents

Introduction

The successful operation of the Large Hadron Collider (LHC) at CERN enables us to test the fundamental laws of nature at a high precision across a large kinematic range. In July 2012 this led to the detection of the final missing piece in the Standard Model, the Higgs boson, by both the ATLAS [1] and CMS [2] experiments. Despite the great successes of the Standard Model, there are strong theoretical arguments pointing towards beyond the Standard Model (BSM) physics. Currently one of the main focuses of the particle physics community is the determination of properties of the Standard Model to a high enough precision such that deviations from experimental measurements become evident. In particular, the precision of theoretical predictions has to keep up with that of the corresponding experimental measurements.

Experiments in particle physics at the LHC aim to probe the fundamental building blocks of nature through high energy proton collisions. As such, the interpretation of their results requires a precise understanding of the constituent particles – the so-called partons – of the proton. Collisions probing these partons happen at high energy scales where the dynamics of the partons can be described within the theoretical framework of perturbative quantum chromodynamics (QCD). However, a description of the low energy, long range interactions needed for a complete understanding of the initial state of the proton cannot be obtained from perturbative QCD. This makes its accurate determination challenging.

The theoretical predictions corresponding to the experimental data rely on collinear factorization arguments, allowing for the separation of the short distance, perturbative, contributions from the large distance, non-perturbative, contributions. In the framework of collinear factorization a longitudinal cross-section σ can be written as

$$\sigma = \hat{\sigma} \otimes f,$$

where $\hat{\sigma}$ is a partonic cross-section describing the short distance dynamics which is convoluted with a PDF f encoding the partonic structure of the proton. While PDFs cannot be calculated from first principles in the framework of perturbative QCD, they may be extracted from experimental data, and since the PDFs are a universal quantity, they can then be used for the calculation of predictions for other experimental observables. This is why PDFs are a fundamental ingredient to test the faithfulness of the Standard Model.

Because of its important role in theoretical predictions of the Standard Model, a precise determination of the PDFs is the subject of ongoing research with various groups regularly releasing sets of PDFs. One such group is the NNPDF collaboration

which uses a methodology that differs from the standard approach in a number of important ways, leading to a PDF determination with a reduced parametrization bias.

It is now generally accepted that the frontiers of high-energy collider physics require percent-level accuracy from both experiment and theory. To achieve this it is critical for the PDF uncertainties to decrease, while ensuring all factors that impact their determination are well understood and the final PDF uncertainty is accurate. This thesis focuses on the determination of a new set of PDFs, NNPDF4.0, and in particular the methodology used in its determination. It aims to provide a better understanding of the impact of various sources of PDF uncertainties in the NNPDF determinations, and proposes several improvements to the methodology along the way.

Outline of the thesis

Chapter 1: *Parton distribution functions*

We provide an introduction on parton distribution functions mainly based on Refs. [3–5]. It reviews the theoretical formalism of QCD and in particular how parton distribution functions emerge in the calculation of deep inelastic structure functions and Drell-Yan cross-sections. We also discuss theoretical aspects relevant for the determination of PDFs, in particular the evolution of PDFs with respect to their energy scale and further theoretical constraints.

Chapter 2: *NNPDF4.0: towards PDFs with percent-level accuracy*

We present NNPDF4.0 [6], the latest set of PDFs released by the NNPDF collaboration. We focus on the methodological framework underpinning the determination, and in particular the use of a gradient descent based optimization algorithm and an automated selection of the model hyperparameters resulting in a significantly improved efficiency of the fitting algorithm. We then discuss the experimental dataset and some of the important features of the NNPDF4.0 PDF set. Finally, we present the open-source NNPDF code [7] and list the main packages with their corresponding functionalities included in the code.

Chapter 3: *Advanced machine learning tools*

We highlight certain aspects of the NNPDF4.0 methodology and propose how they may be improved upon for a potential future release of NNPDF PDFs. First, we propose an alternative data-based scaling of the momentum fraction x first presented in Ref. [8]. This scaling facilitates the removal of a preprocessing prefactor present in the NNPDF parametrization of the PDFs, thereby significantly simplifying the methodology without a loss of efficiency. Then we propose an extension to the hyperoptimization framework used to determine the model hyperparameters in NNPDF4.0. The proposed methodology relies on the automated construction of representative subsets of data to improve test of the methodology's generalizability, as well as a statistical measure for the detection of overfitting.

Chapter 4: *Methodological uncertainties in PDFs*

We discuss different sources of PDF uncertainties with a particular focus on methodological uncertainties. In the first part of the chapter we study correlation between different sets of PDFs and examine the extent to which the correlation between them is due to the underlying data. We then discuss how this knowledge

can be used to assess the efficiency of methodologies used for PDF determination. We also show that the use of data-driven correlations for the combination of different PDF sets can lead to inconsistent results. In the second part of this chapter we clarify certain fundamental aspects of the statistical framework underpinning the sampling as performed within the NNPDF methodology.

Chapter 5: *The neural network approach for neutrino structure functions*
We demonstrate how the NNPDF methodology can be applied to problems closely related to, but fundamentally different from, PDF determination. In particular, we present a determination of neutrino inelastic structure functions for a wide range of scattering energies.

Much of the work in this thesis has been done in collaboration with colleagues from the NNPDF collaboration. Where results are presented, I tried to emphasize the parts where I believe that I have made a significant contribution. Unless otherwise stated in the caption I have generated the figures for this thesis or they have appeared in previous publications co-authored by me.

Chapter 1

QCD and parton distribution functions

This chapter introduces parton distribution functions, a fundamental component of QCD, factorizing the non-perturbative long range dynamics corresponding to the hadronic states. When convoluted with the perturbative, partonic cross-sections they allow for the calculation of predictions of LHC processes.

We start by providing a brief overview of some of the fundamental properties of QCD, the gauge theory of the strong interaction. We then explore how PDFs arise in the framework of perturbative QCD, and in particular we discuss collinear factorization theorems in QCD using the case of the deep inelastic scattering (DIS) of a lepton off a hadronic target as a basic example. We will finally review some properties of PDFs exploited in their determination from experimental data.

1.1 Basics of quantum chromodynamics

The observation of a symmetry corresponding to the special unitary group of degree 3, $SU(3)$, in the spectrum of mesons and baryons lead to the idea of quarks whose interactions are described by quantum chromodynamics, a gauge quantum field theory based on the non-Abelian gauge group $SU(3)$ [9–11]. The classical Lagrangian of QCD is fully determined from the requirement to satisfy the $SU(3)$ symmetries and renormalizibility, it reads[1]

$$\mathcal{L} = \sum_{i=1}^{n_f} \bar{\psi}_i^a \left(i\gamma^\mu D_\mu - m_i\right)_{ab} \psi_i^b - \frac{1}{4} F^{A\mu\nu} F^A_{\mu\nu}. \tag{1.1}$$

This term describes the interactions of the massless spin-1 gluons, as well as the spin-$\frac{1}{2}$ quark fields ψ_i^a of mass m_i, where the label i runs over all n_f flavors. The index a is the color index for the fundamental triplet representation ψ_i^a, while A is the color index in the adjoint representation corresponding to the eight color degrees of freedom of the

[1] The Lagrangian of Eq. (1.1) can be extended with another gauge invariant term proportional to $\epsilon_{\mu\nu\gamma\rho} F^{A\mu\nu} F^A_{\gamma\rho}$. This can be written as a total derivative leaving the Euler-Lagrange equations unchanged. In reality there are some additional subtleties to this argument [12], but for the discussion in this chapter the term will be ignored.

gluon field $\mathcal{A}^A_\mu$. The index μ is the Lorentz index running over the four dimensions of spacetime. The gamma matrices γ^μ satisfy the anti-commutation relation

$$\{\gamma^\mu, \gamma^\nu\} = 2g^{\mu\nu}, \tag{1.2}$$

with $g^{\mu\nu}$ the Minkowski metric. The covariant derivative is defined as

$$D_\mu = \partial_\mu + ig_s(\mathcal{A}^A_\mu t^A). \tag{1.3}$$

where g_s is the gauge coupling representing the strength of the interaction between colored states, it is a free parameter of the theory. t^A are the generators of the group in the fundamental representation satisfying the commutation relations

$$\left[t^A, t^B\right] = if^{ABC}t^C, \tag{1.4}$$

where f^{ABC} are the structure constants of the $SU(3)$ group. A representation for the generators t^A is given by

$$t^A = \frac{1}{2}\lambda^A, \tag{1.5}$$

with λ^A corresponding to the eight 3×3 Hermitian traceless Gell-Mann matrices, with the normalization of the generators conventionally chosen as

$$\mathrm{Tr}\left(t^A t^B\right) = T_R\delta^{AB}, \quad T_R = \frac{1}{2}. \tag{1.6}$$

Finally, $F^A_{\mu\nu}$ is the field strength tensor and can be defined in terms of the gluon fields and the structure constant as

$$F^A_{\mu\nu} = \partial_\mu\mathcal{A}^A_\nu - \partial_\nu\mathcal{A}^A_\mu - g_s f^{ABC}\mathcal{A}^B_\mu\mathcal{A}^C_\nu. \tag{1.7}$$

It is worth pointing out that the final term in Eq. (1.7) corresponds to the gluon self-interaction. An equivalent term is not present for the virtual photon fields in quantum electrodynamics (QED) as it is a feature of non-Abelian gauge theories. In QCD this leads to the important property of asymptotic freedom to be discussed below.

From the Lagrangian of Eq. (1.1), it is now possible to calculate observables such as cross-sections or decay rates in terms of expansions in the strong coupling constant α_s, which is defined in terms of the QCD gauge coupling from Eq. (1.7) as $\alpha_s = g_s^2/4\pi$. However, radiative quantum corrections introduce divergences beyond leading order in the calculation of a physical observable. These divergences are treated by a renormalization procedure to remove ultraviolet (UV) divergences, which requires that the coupling must be redefined to absorb the dependence on the renormalization scale. This dependence of the running coupling $\alpha_s(\mu^2)$ on the scale μ^2 at which the subtraction of the ultraviolet poles is performed is given by the renormalization group equation:

$$\frac{d\alpha_s(\mu^2)}{d\log\mu^2} = \beta\left(\alpha_s\left(\mu^2\right)\right), \tag{1.8}$$

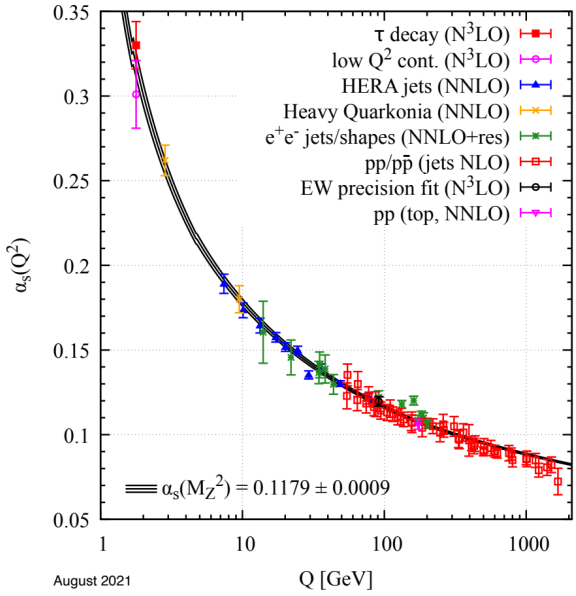

Figure 1.1: An overview of measurements of α_s as a function of the energy scale Q. The perturbative order of the QCD calculation used in the extraction is indicated in the parentheses. The figure is taken from Ref. [16]

where the β function can be calculated as a series in $\alpha_s(\mu^2)$

$$\beta\left(\alpha_s\left(\mu^2\right)\right) = -\alpha_s^2\left(\mu^2\right)\left(\beta_0 + \beta_1\alpha_s\left(\mu^2\right) + \mathcal{O}\left(\alpha_s^2\right)\right), \tag{1.9}$$

with

$$\beta_0 = \frac{33 - 2n_f}{12\pi}, \quad \beta_1 = \frac{153 - 19n_f}{2\pi\left(33 - 2n_f\right)}. \tag{1.10}$$

While, for illustrative purposes, the β function is only explicitly shown up to NLO, it is known up to five loops [13].

The solution to the renormalization group equation Eq. (1.8) at leading log in the expansion of the inverse powers of $\log\left(\mu^2\right)$ can be written as

$$\alpha_s\left(\mu^2\right) = \frac{\alpha_s\left(\mu_0^2\right)}{1 + \beta_0\alpha_s\left(\mu_0^2\right)\log\frac{\mu^2}{\mu_0^2}}, \tag{1.11}$$

where μ_0^2 is an arbitrary initial scale. This highlights an important property of QCD. Namely, since β_0 is positive for $n_f < 17$, the value of the running coupling $\alpha_s(\mu^2)$ decreases logarithmically to 0 as the energy scale of the process increases. This property is called asymptotic freedom, and makes QCD an asymptotically free theory [14, 15].

Perturbative QCD thus tells us how the coupling constant depends on the scale, but it does not provide us with its value. This has to be obtained experimentally. Commonly the value of the coupling constant is quoted at the mass of the Z boson,

thus $\alpha_s(M_Z^2)$. Using Eq. (1.11), the value of the coupling constant at any other scale can then be obtained. Fig. 1.1 shows an overview of different measurements of the strong coupling $\alpha_s(\mu^2)$ for a range of scales.

In Eq. (1.11) the fixed coupling still depends on the arbitrary scale μ_0. In some cases we may wish to remove this dependence, which is commonly done by replacing it with a dimensionful parameter Λ roughly corresponding to the point at which the theory becomes strongly coupled. It is defined as

$$\log \frac{\mu^2}{\Lambda^2} = -\int_{\alpha_s(\mu^2)}^{\infty} \frac{dx}{\beta(x)}, \tag{1.12}$$

and allows us to write Eq. (1.11) as

$$\alpha_s(\mu^2) = \frac{1}{\beta_0 \log \frac{\mu^2}{\Lambda^2}} \tag{1.13}$$

Its value is around 200 MeV, though its precise definition depends on the choice of renormalization scheme. Along with the RGE equations describing the running of the coupling, Λ allows us to replace the dependence on the dimensionless parameter g_s, which – as we have just seen – is not a constant.

Asymptotic freedom is the fundamental property of QCD that allows us to perform calculations perturbatively in the limit $\mu \gg \Lambda$, since there we have that $\alpha_s(\mu^2) \ll 1$ and the dynamics of the quarks can be approximated by that of free particles. However, the same scaling relation tells us that for low energies the theory is strongly coupled and thus perturbation theory cannot be applied reliably in this regime.

1.2 Collinear factorization and the parton model

A description of the low energy bound state of the hadrons is required to make predictions for collisions involving hadrons. However, as we have just seen, at low energies the running coupling becomes large and thus perturbative QCD is not accurate in this regime. Here we will discuss the collinear factorization theorems that allow for the factorization of the short distance effects that can be computed perturbatively, and the long distance effects that have to be extracted from data. We will consider as an example the process of deep inelastic scattering. In particular we will see how PDFs arise in this framework.

1.2.1 Kinematics of deep inelastic scattering

Deep inelastic scattering is the process of colliding a lepton with a hadronic target, destroying the target in the process (compared to elastic or slightly inelastic scattering, where the target is not destroyed). This is a very clean way of testing QCD since it allows us to probe the hadron (commonly a proton) with a structureless probe (commonly an electron). Historically, DIS experiments have played a vital role in obtaining a deeper understanding of perturbative QCD, and currently these measurements still play an important role in the determination of PDFs. Important

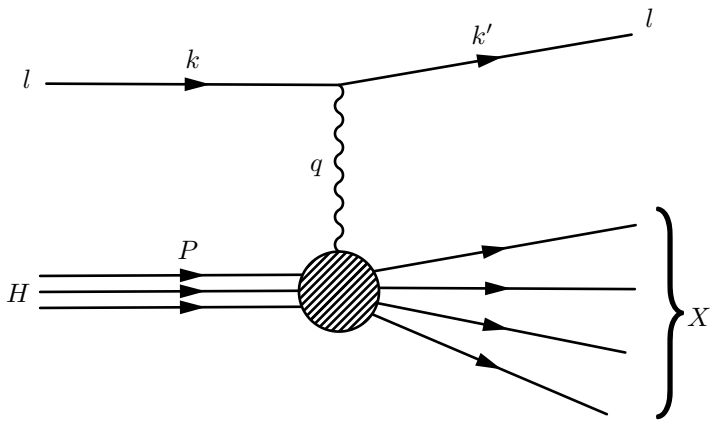

Figure 1.2: Schematic representation of the deep inelastic scattering of a charged lepton l off a hadronic target H.

of such measurements are those performed at SLAC [17], BCDMS [18], and HERA (H1 [19] and ZEUS [20]).

Fig. 1.2 shows a schematic representation of a DIS process where a charged lepton l with momentum k probes a hadron H with momentum P thereby breaking it apart into a hadronic final state X:

$$l(k) + H(P) \to l(k') + X. \tag{1.14}$$

Let us here consider the case of a proton probed by an electron, which involves neutral current (NC) scattering through a virtual photon. We will consider only the contribution associated to the photon exchange, which is a valid assumption for energy scales well below M_Z. To describe this process we need to parametrize the interaction of the photon with the proton. As such, a sensible choice for the parametrization of the cross-section would be to describe it in terms of the momentum of the proton P and the momentum of the photon $q = k - k'$. The center of mass energy is denoted by

$$s = (P + k)^2, \tag{1.15}$$

while the invariant mass of the final state X is given by

$$W^2 = (P + q)^2. \tag{1.16}$$

Then, we define the standard DIS kinematic variables:

$$Q^2 = -q^2, \tag{1.17}$$

$$x = \frac{Q^2}{2P \cdot q}, \tag{1.18}$$

$$y = (P \cdot q)/(P \cdot k) = \frac{Q^2}{xs}. \tag{1.19}$$

In the kinematic domain where the scales Q^2 and W^2 are both much greater than the mass of the proton, the mass of the electron and the quarks can be neglected. Here y is the relative energy loss, and can alternatively be written as $y = 1 - E'/E$, with E energy of the incoming electron and E' the energy of the outgoing electron. The variable x, known as the Bjorken x scaling variable, can take values between 0 and 1, where $x = 1$ corresponds to elastic scattering. The deep inelastic scaling region then corresponds to $Q^2 \gg \Lambda^2$ for fixed x and sufficiently small x. As will be discussed in more detail below, the Bjorken x is of fundamental importance in the understanding of DIS processes in QCD.

The idea of this parametrization is that by measuring the kinematics of the outgoing electron, the structure of the proton can be obtained in terms of the characteristics of the probe such as x, Q^2 and y.

The leading order matrix element corresponding to the DIS process of Fig. 1.2 can be written in Feynman gauge as

$$\mathcal{M} = ie^2 \bar{u}\left(k'\right) \gamma^\mu u(k) \left(i \frac{g_{\mu\nu}}{Q^2}\right) \langle X \left| J^\nu \right| H \rangle , \tag{1.20}$$

where spin labels have been omitted. Here $|H\rangle$ represents the state of the incoming hadron, $|X\rangle$ represents the hadronic final state, and J_μ is the electromagnetic current. It is worth noting that the hadronic states cannot be computed in perturbation theory as a result of the large value of the coupling constant discussed in Sect. 1.1 above.

If we want to describe the cross-section of a DIS process, a natural starting point is thus to separate the both the phase-space factor and the Feynman amplitude into a leptonic and a hadronic part as

$$d\Phi = \frac{d^3k'}{(2\pi)^3 2E'} d\Phi_X = \frac{ME}{8\pi^2} y dy dx d\Phi_X, \tag{1.21}$$

$$\frac{1}{4}\sum_{\text{spin}} |\mathcal{M}|^2 = \frac{e^4}{Q^4} L^{\mu\nu} h_{X\mu\nu}, \tag{1.22}$$

with a leptonic tensor $L_{\mu\nu}$, and the hadronic part of the amplitude denoted by $h_{X\mu\nu}$. The leptonic tensor can be calculated explicitly in QED and reads

$$\begin{aligned} L^{\mu\nu} &= \frac{1}{4}\sum_{\text{spin}} \bar{u}\left(k'\right) \gamma^\mu u(k) \bar{u}(k) \gamma^\nu u\left(k'\right), \\ &= \frac{1}{4} \operatorname{tr}\left[\not{k} \gamma^\mu \not{k}' \gamma^\nu\right], \\ &= k^\mu k'^\nu + k'^\mu k^\nu - g^{\mu\nu} k \cdot k'. \end{aligned} \tag{1.23}$$

The hadronic part of phase space and the amplitude can be combined into a hadronic tensor

$$W_{\mu\nu} = \sum_X \int d\Phi h_{X\mu\nu}. \tag{1.24}$$

Then, by requiring Lorentz symmetry and gauge invariance we find that a general formulation of the hadronic tensor can be written as

$$\begin{aligned} W^{\mu\nu}(P,q) = &- \left(g^{\mu\nu} + \frac{q^\mu q^\nu}{q^2}\right) F_1\left(x, Q^2\right) \\ &+ \left(P^\mu - q^\mu \frac{P\cdot q}{q^2}\right)\left(P^\nu - q^\nu \frac{P\cdot q}{q^2}\right) \frac{1}{P\cdot q} F_2\left(x, Q^2\right), \end{aligned} \tag{1.25}$$

where the functions F_1 and F_2 are called structure functions. More general structures including a third structure function F_3 can be found if we allow for parity violating interaction mediated by a W or Z boson as will be discussed in Chapter 5.

The cross-sections corresponding to the DIS process can be calculated using

$$\sigma = \sum_X \frac{1}{4ME} \int d\Phi \frac{1}{4} \sum_{\text{spin}} |\mathcal{M}|^2, \tag{1.26}$$

where combining the expressions collected above gives

$$\frac{d\sigma}{dx dQ^2} = \frac{2\pi\alpha^2}{Q^4}\left[\left[1 + (1-y)^2\right] F_T\left(x, Q^2\right) + \frac{2(1-y)}{x} F_L\left(x, Q^2\right)\right], \tag{1.27}$$

with $\alpha = e^2/(4\pi)$ the fine structure constant quantifying the strength of the electromagnetic interaction, and the transverse structure function F_T and the longitudinal structure functions F_L defined as

$$F_L = F_2 - 2xF_1, \tag{1.28}$$

$$F_T = 2F_1. \tag{1.29}$$

Note that the distinction between F_1 and F_2 (or F_L and F_T) can be made based on the y dependence of the prefactor.

Thus far the only assumptions that we have made about the $W_{\mu\nu}$ tensor is that it satisfies both gauge and Lorentz invariance. Let us now also assume that the proton is formed as a bound state of constituent particles. If we then consider the DIS process in a reference frame where the hadron moves very fast and the energy of the process is large, it will be Lorentz contracted in the direction of the collision and the lifetime of particles inside the hadron increases. What this means in practice for a DIS process, is that upon probing the hadron with an external lepton, the interaction can be though of as the interaction between a lepton and a single, pointlike, particle while the other particles inside the hadron do not interfere. Interactions happening in the final state, however, for similar reasons occur on large timescale therefore not interfering with the hard scattering process either.

This is the basic intuition leading to Feynman's parton model [21], which assumes that the proton is formed as a bound state of constituent spin-$\frac{1}{2}$ objects called partons. It suggests that short distance physics describing the electron-parton interactions can

be separated from long distance physics describing the hadron. Using the parton model, the DIS cross-section can then be written as

$$\frac{d^2\sigma}{dxdQ^2} = \int_0^1 \frac{d\xi}{\xi} \sum_i f_i(\xi) \frac{d^2\hat{\sigma}}{d\hat{x}dQ^2} \left(\frac{x}{\xi}, Q^2\right), \tag{1.30}$$

where $f_i(\xi)$ represents the probability of finding a parton of flavor i inside the hadron carrying a momentum fraction ξ of the total momentum of the proton (and thus the parton carries momentum ξP), and $d^2\hat{\sigma}/(d\hat{x}dQ^2)$ is the cross-section for the scattering of the electron with a parton. Such a factorized expression is accurate up to corrections that are suppressed by powers of Λ^2/Q^2 corresponding to so-called higher twist terms.

We now have the tools needed to discuss scale dependence in the parton model. To this end, let consider the leading order cross-section of the scattering of a lepton off a parton, $e^- q \to e^- q$. Using the DIS variables we can express the cross-section differential in Q^2 and x as

$$\frac{d^2\hat{\sigma}}{dQ^2dx} = \frac{4\pi\alpha^2}{Q^4} \frac{1}{2} \left[1 + (1-y)^2\right] \delta(x-\xi), \tag{1.31}$$

implying that the Bjorken variable x is equal to the momentum fraction ξ at leading order. From Eq. (1.31) we can read the expressions for the partonic structure functions

$$\hat{F}_2 = 2x\hat{F}_1 = xe^2\delta(x-\xi). \tag{1.32}$$

Finally, if we compare the cross-section as described in terms of structure function in Eq. (1.27) to the cross-section as described in the parton model of Eq. (1.30) with Eq. (1.31), we find

$$F_2(x) = 2xF_1 = x \sum_{i=q,\bar{q}} \int_0^1 d\xi f_i(\xi) e_q^2 \delta(x-\xi) = x \sum_{i=q,\bar{q}} e_q^2 f_i(x). \tag{1.33}$$

Eq. (1.33) shows how DIS experiments can probe the structure of the proton in terms of its quark and gluon constituents. It further explicitly shows that the structure functions in the limit of large Q^2 only depend on x and not on Q^2. This is known as Bjorken scaling [22] and establishes that DIS must be described in terms of the scattering process of the parton with a photon. It also shows that $F_L = F_2 - 2xF_1 = 0$ holds at leading order. This equation is known as the Callan-Gross relation [23], and is a consequence of the fact that for a longitudinally polarized photon scattering off a spin-$\frac{1}{2}$ particle the cross-section vanishes [24].

It is also worth highlighting that here the quark q and antiquark $\bar{q}$ are indistinguishable, hence to independently determine the corresponding PDFs, charged current (CC) processes are required.

1.2.2 Deep inelastic scattering in QCD

Early DIS experiments showed good agreement with the parton model, thereby providing strong support for QCD as a theory for the strong interaction, and Feynman's partons were soon associated with the quarks in the model based on the

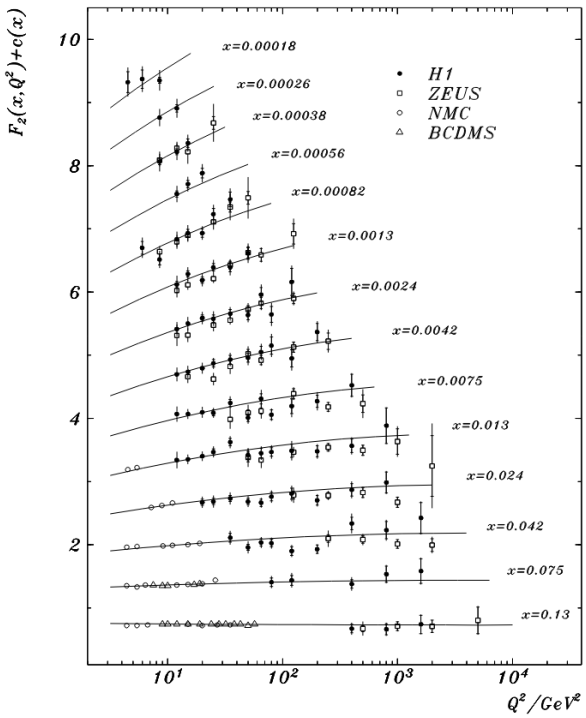

Figure 1.3: Measurement of the $F_2(x, Q^2)$ structure function from various different experiments. To the F_2 values for each x a scale term $c(x) = 0.6(i - 0.4)$ has been added, where i is the bin number in x starting at $i = 1$ for $x = 0.13$. The figure is taken from Ref. [25]

$SU(3)$ gauge symmetry by Gell-Mann in 1964 [10]. However, in our discussion so far we have neglected higher order QCD corrections. Such corrections would correspond to logarithms of the scale of the process Q^2, and thus introduce a dependence on Q^2 ignored by the large Q^2 assumption leading to the observation of Bjorken scaling. Indeed, measurements of the F_2 structure function such as those by the H1 Collaboration [25] shown in Fig. 1.3 reveal a violation of the Bjorken scaling of the structure function. We will now extend the parton model to include the first order of QCD corrections.

We have seen how the idea of the parton model without QCD corrections allowed us to obtain the result of Eq. (1.33). The inclusion of higher order QCD corrections is then obtained through a generalization of this result to all orders in QCD motivated by the factorization theorem [4]. We can then write any structure function F in a

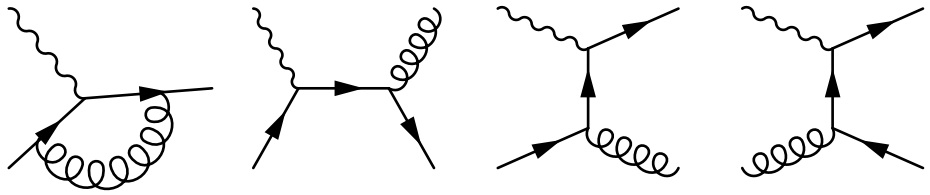

Figure 1.4: Feynman diagrams describing QCD corrections to the $qq\gamma^*$ vertex. The first three Feynman diagrams correspond to α_s corrections to the LO process, while the final (right-most) diagram corresponds to photon-gluon fusion.

factorized form where the partonic structure functions, also called Wilson coefficients, $C_i(x, Q^2)$ are weighted by the PDFs as

$$\begin{aligned} F\left(x, Q^2\right) &= \sum_{i=q,\bar{q},g} \int_x^1 \frac{d\xi}{\xi} C_i\left(\frac{x}{\xi}, Q^2\right) f_i(\xi), \\ &= \sum_{i=q,\bar{q},g} C_i\left(\frac{x}{\xi}, Q^2\right) \otimes f_i(\xi), \end{aligned} \tag{1.34}$$

where $\otimes$ denotes the Mellin convolution product defined as

$$f(x) \otimes g(x) \equiv \int_x^1 \frac{dy}{y} f\left(\frac{x}{y}\right) g(y). \tag{1.35}$$

The Wilson coefficients encode information about the high energy process and can be calculated as a perturbative series in α_s.

Let us now explicitly consider QCD correction by again studying the case of the $qq\gamma^*$ vertex. The leading order diagram is the same as in the parton model, while at NLO we find both virtual corrections of the self-energy diagram and real corrections corresponding to the gluon emission from the incoming or outgoing fermion line. We also find a diagram corresponding to photon-gluon fusion. The corresponding Feynman diagrams are shown in Fig. 1.4 and contain both infrared (IR) and UV divergences.

The infrared divergences cancel between the real and virtual corrections, since the Standard Model is perturbatively infrared finite [26, 27]. Nevertheless, a regulator is introduced to accommodate intermediate steps of the calculation which is later removed. The infrared divergences originate from the treatment of partons as massless particles, as such one may regulate these divergences by introducing a small mass for the partons. In practice, dimensional regularization is generally preferred (in particular for calculations beyond NLO), but for illustrative purposes we will introduce quark masses m_q and the gluon mass m_g.

Accounting for the contribution to the LO vertex corresponding to the Feynman diagrams of Fig. 1.4 gives

$$\hat{F}_2^q = e_q^2 x \left[\delta(1-x) + \frac{\alpha_S}{4\pi}\left[P_{qq}(x) \log \frac{Q^2}{m_g^2} + C_2^q(x)\right]\right], \tag{1.36}$$

where the $\delta(1-x)$ term corresponds to the LO vertex while the $\mathcal{O}(\alpha_s)$ term corresponds to the first three diagrams in Fig. 1.4. The final diagram of Fig. 1.4 then gives

$$\hat{F}_2^g = \sum_q e_q^2 x \frac{\alpha_S}{4\pi}\left[P_{qg}(x)\log\frac{Q^2}{m_q^2} + C_2^g(x)\right]. \tag{1.37}$$

The P_{ij} in Eq. (1.36) and Eq. (1.37) represent the Altarelli-Parisi splitting functions describing the probability that a parton j splits into a parton i and another parton carrying a momentum fraction x of the incoming parton j. The splitting functions are universal and can be calculated perturbatively in QCD. Currently they are known up to NNLO [28,29], with parts of the splitting functions known at N3LO [30–32].

The presence of large logarithms of Q^2/m^2 in Eq. (1.36) and Eq. (1.37) indicate that not all UV and soft divergences have canceled. Namely, a collinear divergence remains in the case where an emitted gluon is collinear to the incoming quark. These large logarithms now contain all the residual long-range physics left after resumming the real and virtual corrections. While these infrared divergences appear at the parton level, a physical observable should not be sensitive to infrared divergences. The physical observable can be obtained using Eq. (1.34) to give

$$\begin{aligned} F_2^q\left(x,Q^2\right) =& x\sum_{i=q,\bar{q}} e_q^2 \\ &\times\left[f_{i,0}(x) + \frac{\alpha_S}{2\pi}\int_x^1\frac{d\xi}{\xi}f_{i,0}(\xi)\left[P_{qq}\left(\frac{x}{\xi}\right)\log\frac{Q^2}{m_g^2} + C_2^q\left(\frac{x}{\xi}\right)\right]\right], \end{aligned} \tag{1.38}$$

where $f_{q,0}$ are the bare PDFs. We may also define renormalized PDFs

$$f_q\left(x,\mu_F\right) \equiv f_{q,0}(x) + \frac{\alpha_S}{2\pi}\int_x^1\frac{d\xi}{\xi}f_{q,0}(\xi)P_{qq}\left(\frac{x}{\xi}\right)\log\frac{\mu_F^2}{m_q^2} + z_{qq}, \tag{1.39}$$

where we have introduced a factorization scale μ_F defining the threshold between long and short distance physics, and the dependence on the IR cutoff has been absorbed into the definition of $f_{q,0}\left(x,\mu_F\right)$. This is possible because the divergences are universal and thus so are the renormalized PDFs. The finite term z_{qq} depends on the factorization choice and is a calculable quantity.

Finally, this allows us to write the structure function as a factorized expression:

$$\begin{aligned} F_2^q\left(x,Q^2\right) =& x\sum_{i=q,\bar{q}} e_q^2\int_x^1\frac{d\xi}{\xi}f_i\left(\xi,\mu_F^2\right) \\ &\times\left[\delta\left(1-\frac{x}{\xi}\right) + \frac{\alpha_S\left(\mu_R\right)}{2\pi}\left[P_{qq}\left(\frac{x}{\xi}\right)\log\frac{Q^2}{\mu_F^2} + C_2^q\left(\frac{x}{\xi}\right) - z_{qq}\right]\right]. \end{aligned} \tag{1.40}$$

This is an important expression for understanding DIS in QCD, and some point are worth stressing here. In particular, all long distance effects are encoded in the PDFs $f_i\left(\xi,\mu_F^2\right)$ which depend on a factorization scale μ_F in such a way as to exactly cancel the dependence at all orders in perturbation theory. Since the final result is given as an expansion in α_s, it does depend on the choice of the renormalization scale

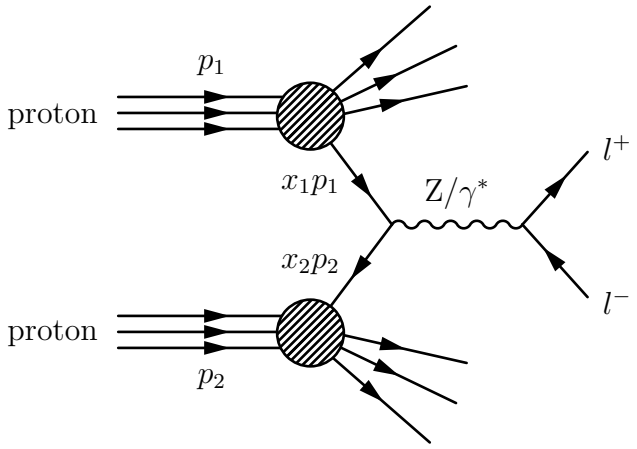

Figure 1.5: A diagrammatic representation of a neutral current Drell-Yan process in a proton-proton collision.

μ_R. The short distance effects are encoded in the factorization scale dependent wilson coefficients and can be calculated perturbatively.

1.2.3 Hadron-hadron collisions

So far we have only considered the factorization for DIS processes where we have only a single hadron in the initial state, however the formalism may be extended [33–35] to processes with two hadrons in the initial state. This allows us to study processes at proton-proton colliders such as the LHC. Arguably the most relevant process in proton-proton collisions is the neutral current Drell-Yan (DY) process shown in Fig. 1.5, in which a quark from one proton and an antiquark from the other proton annihilate, creating a Z boson or virtual photon which finally decays into a lepton pair l^+l^-.

The collinear factorization theorem for a process involving two incoming hadrons reads[2]

$$\begin{aligned}\sigma_X(s, M_X) = &\sum_{a,b} \int_0^1 dx_1 dx_2 f_a\left(x_1, \mu_F^2\right) f_b\left(x_2, \mu_F^2\right) \\ &\times \hat{\sigma}_{ab\to X}\left(x_1, x_2, \alpha_S\left(\mu_R^2\right), \frac{Q^2}{\mu_F^2}, \frac{Q^2}{\mu_R^2}\right),\end{aligned} \tag{1.41}$$

Here $\hat{\sigma}_{ab\to X}$ is the partonic cross-section for the production of a hadronic final state X, and it encodes the short-distance behavior for incoming quarks or gluons of flavors a and b that can be calculated as an expansion in α_s.

Eq. (1.40) and Eq. (1.41) tell us how to connect calculations of hard scattering cross-sections in perturbative QCD with two partons in the initial state to observables resulting from collisions with hadrons. Since PDFs encode the initial state of hadrons and are by definition non-perturbative, they cannot be calculated using perturbation theory. Instead they have to be extracted through the analysis of experimental collider

[2]It should be noted that the factorization theorem has not been formally proven for all processes considered at the LHC, but even in those cases factorization is generally treated in the same way.

data. This is only possible because PDFs are universal objects, and thus the PDFs appearing in Eq. (1.40) are the same as those appearing in Eq. (1.41).

1.3 Scale dependence of PDFs

The observables calculated using the factorized expressions correspond to measurable quantities, and must therefore be independent of the factorization scale μ_F introduced in a renormalization procedure as a way of treating initial state divergences. This observation leads to the renormalization group equation for the structure functions

$$\mu_F^2 \frac{dF_2\left(x, Q^2\right)}{d\mu_F^2} = 0, \tag{1.42}$$

while the renormalization group equation for the quark distributions reads

$$\mu_F^2 \frac{d}{d\mu_F^2} f_q\left(x, \mu_F^2\right) = \frac{\alpha_s\left(\mu_F^2\right)}{2\pi} \int_x^1 \frac{d\xi}{\xi} P_{qq}\left(\frac{x}{\xi}, \alpha_s\left(\mu_F^2\right)\right) f_q\left(\xi, \mu_F^2\right), \tag{1.43}$$

which are again expressed in terms of the Altarelli-Parisi splitting functions.

These expressions are known as the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) equations [36–38] and describe how the PDFs evolve with respect to the factorization scale. The DGLAP equations allow us to define PDFs at a given scale Q_0 and evolve them up to the scale Q of a hard process. This is why processes at different energy scales can all be used to constrain the PDF that is parametrized at an initial scale Q_0

Due to the flavor symmetries present in QCD in the limit where quark masses are neglected, it is possible to define a basis of flavor states that are preserved under evolution with the matrix P_{ij}. One such basis can be constructed by dividing the system of equations into two subsystems known as the singlet and non-singlet sectors. Given a system with thirteen partons with $n_f = 6$ consisting of six quarks $f_i = \{u, d, s, c, t, b\}$, their anti-quarks, as well as the gluon, we define

$$f_i^{\pm} \equiv f_i \pm \bar{f}_i. \tag{1.44}$$

Here we make the distinction between valence f^- commonly denoted by[3]

$$V_i \equiv f_i^-, \tag{1.45}$$

[3] Another common notational convention that will appear in this thesis to denote a valence quark is by writing a subscript V, e.g. the valence up quark is written as $u_V = u - \bar{u}$.

and the triplet states

$$\begin{aligned} T_3 &\equiv u^+ - d^+, \\ T_8 &\equiv u^+ + d^+ - 2s^+, \\ T_{15} &\equiv u^+ + d^+ + s^+ - 3c^+, \\ T_{24} &\equiv u^+ + d^+ + s^+ + c^+ - 4b^+, \\ T_{35} &\equiv u^+ + d^+ + s^+ + c^+ + b^+ - 5t^+. \end{aligned} \tag{1.46}$$

The valence and triplet states comprise the non-singlet sector and evolve according to

$$\mu_F^2 \frac{d}{d\mu_F^2} f^{\mathrm{NS}}\left(x, \mu_F^2\right) = \frac{\alpha_s\left(\mu_F^2\right)}{2\pi} \int_x^1 \frac{d\xi}{\xi} P\left(\xi, \alpha_S\right) f^{\mathrm{NS}}\left(\frac{x}{\xi}, \mu_F^2\right), \tag{1.47}$$

where valence states evolve with P_- and the triplet states with P_+. At leading order the splitting functions read

$$P_-^{(0)}(x) = P_+^{(0)}(x) = \frac{C_F}{2\pi}\left(\frac{1+x^2}{1-x}\right)_+. \tag{1.48}$$

For the singlet sector we define the singlet distribution

$$\Sigma \equiv \sum_{i=1}^{n_f} f_i^+, \tag{1.49}$$

which couples to the gluon PDF, and the evolution equations of the corresponding system read

$$\mu_F^2 \frac{d}{d\mu_F^2} \begin{pmatrix} \Sigma\left(x, \mu_F^2\right) \\ g\left(x, \mu_F^2\right) \end{pmatrix} = \frac{\alpha_s\left(\mu_F^2\right)}{2\pi} \int_x^1 \frac{d\xi}{\xi} \begin{pmatrix} P_{qq} & P_{qg} \\ P_{gq} & P_{gg} \end{pmatrix} \begin{pmatrix} \Sigma\left(\xi, \mu_F^2\right) \\ g\left(\xi, \mu_F^2\right) \end{pmatrix}. \tag{1.50}$$

The convolution of Eq. (1.34) can be written in a more convenient way that will allow us to find an analytic solution for the DGLAP equation by performing a Mellin transform defined as

$$f(N) \equiv \int_0^1 dx x^{N-1} f(x). \tag{1.51}$$

Namely, by performing a Mellin transform of a convolution it can be written as a simple product:

$$\begin{aligned} \int_0^1 dx x^{N-1} \left[\int_x^1 \frac{dy}{y} f(y) g\left(\frac{x}{y}\right)\right] &= \int_0^1 dx x^{N-1} \int_0^1 dy \int_0^1 dz \delta(x - zy) f(y) g(z) \\ &= \int_0^1 dy \int_0^1 dz (zy)^{N-1} f(y) g(z) \\ &= f(N) g(N). \end{aligned} \tag{1.52}$$

Using Eq. (1.34), and noting that from the parton model we have $F_2 \propto \sum_i f_i \otimes \hat{F}_2$, we find that the Mellin transform of the renormalization group equation for the structure functions reads

$$\frac{df_q\left(N,\mu_F\right)}{d\log\mu_F}\hat{F}_2\left(N,\frac{\mu_F}{Q}\right)+f_q\left(N,\mu_F^2\right)\frac{d\hat{F}_2\left(N,\frac{\mu_F}{Q}\right)}{d\log\mu_F}=0, \tag{1.53}$$

where we can separate the PDF and structure function coefficient terms as

$$\frac{d\log\hat{F}_2\left(N,\frac{Q}{\mu_F}\right)}{d\log(Q/\mu_F)}=\frac{d\log f_q\left(N,\mu_f\right)}{d\log\mu_F}=-\gamma_{qq}(N). \tag{1.54}$$

Here $\gamma_{ij}(N)$ denote the anomalous dimensions, which are the Mellin transforms of the corresponding splitting functions P_{ij}.

The solution of the evolution equation in Mellin space Eq. (1.54) can then be written as

$$f_q(N,Q)=f_q\left(N,Q_0\right)e^{-\gamma_{qq}(N)\log\left(\frac{\mu_F}{Q_0}\right)}. \tag{1.55}$$

These evolution equations can be used to evolve the PDFs from an initial scale Q_0 to a general scale Q.

The full DGLAP equations in Mellin space then become

$$\begin{aligned}\frac{d}{d\mu_F^2}f_i^{\rm NS}\left(N,Q^2\right)&=\frac{\alpha_s(\mu_F^2)}{2\pi}\gamma_{qq}^{\rm NS}\left(N,\alpha_s(\mu_F^2)\right)f_i^{\rm NS}\left(N,Q^2\right)\\ \frac{d}{d\mu^2}\begin{pmatrix}\Sigma\left(N,Q^2\right)\\ g\left(N,Q^2\right)\end{pmatrix}&=\frac{\alpha_s(\mu_F^2)}{2\pi}\begin{pmatrix}\gamma_{qq}&2n_f\gamma_{qg}\\ \gamma_{gq}&\gamma_{gg}\end{pmatrix}\begin{pmatrix}\Sigma\left(N,Q^2\right)\\ g\left(N,Q^2\right)\end{pmatrix}\end{aligned} \tag{1.56}$$

In practice the DGLAP equations are solved using iterative numerical procedures. For this purpose several codes have been developed that either solve the evolution equations directly in momentum space such as HOPPET [39], QCDNUM [40] and APFEL [41], or in Mellin space such as PEGASUS [29] or EKO [42]. The Mellin space approach has also been used by the internal NNPDF evolution code FastKernel discussed in Refs. [43–45].

An example of the result of PDF evolution is shown in Fig. 1.6 where we show PDFs evolved from the initial scale $Q_0^2 = 1.65\,\text{GeV}$ to $Q = 3.2\,\text{GeV}$ (left) and $Q = 100\,\text{GeV}$ (right). PDF evolution is fundamental to the extraction of PDFs from data, namely it allows us to parametrize the PDFs at an initial scale and evolve to the experimental scale in order make predictions that can be compared to measurements.

1.4 Treatment of heavy quarks

This section is based on the discussion in Ref. [5].

Quarks are conventionally divided into light quarks with a mass well below Λ, and heavy quarks with a mass greater than Λ. Following this definition, the up, down and strange quark are considered light quarks, and for these quarks the approximation leads to accurate results. For the other quarks the argument becomes more subtle, in particular the approximation is no longer accurate when treating processes with a hard

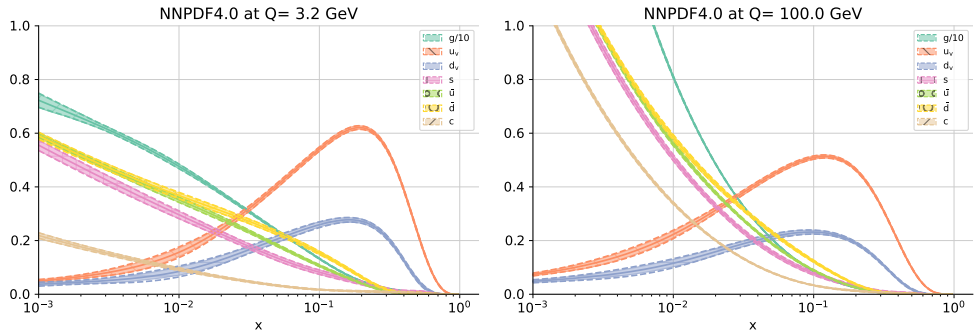

Figure 1.6: The NNPDF4.0 PDFs [6] evolved from the initial scale $Q_0^2 = 1.65\,\text{GeV}$ to $Q = 3.2\,\text{GeV}$ (left) and $Q = 100\,\text{GeV}$ (right).

scale Q that is close to a quarks mass. In those cases a mass independent scheme can no longer be applied reliably and contributions coming from mass dependent terms should be accounted for, in particular when the evolution crosses a scale equal to a quark mass. Because of this, variable flavor number (VFN) schemes are used to obtain an accurate treatment of datasets with a large range in the hard scale.

When considering VFN schemes generally a distinction is made between three kinematic regions:

- $Q \ll m_h$: The mass of the heavy quark is much larger than the hard scale of the process. In this case the heavy quarks can be decoupled [46–48] and treated as a purely final state particle. The scheme accurate in this region is the fixed flavor number (FFN) scheme.

- $Q \sim m_h$: The mass of the heavy quark is of the same order as the hard scale of the process, and should be treated as a large parameter. Heavy quark contributions contribute to the Wilson coefficients or in renormalizations such as in the case where the heavy quark is decoupled.

- $Q \gg m_h$: The mass of the heavy quark is much smaller than the energy scale of the process. In this region the heavy quark is neglected in the Wilson coefficients and instead a heavy quark PDF is introduced. The scheme accurate in this region is the so called zero mass variable flavor number (ZM-VFN) scheme.

To ensure both the decoupling at low scales as provided by the FFN scheme and resummation of logs of Q^2/m_h^2 at high scales as provided by the ZM-VFN scheme, so-called general mass variable flavor number (GM-VFN) schemes have been constructed to interpolate between the FFN and ZM-VFN schemes.

Fixed flavor number scheme

Let us first consider the region where the mass of the heavy quark is roughly equal to the hard scale (the threshold region) or larger than the hard scale of the process, $Q \lesssim m_h$. In this region only the light quarks are treated as partons while the heavy quarks are treated as a purely final state particle. Under these assumptions we only

need to consider the light quarks and the gluon in the theory. Setting the factorization and normalization scales equal to a scale μ allows us to write the calculation of the structure function of Eq. (1.34) as

$$F\left(n_l, Q^2, m_h^2\right) = \sum_{i=1}^{n_l} C_i\left(n_l, \frac{Q^2}{m_h^2}, \frac{\mu^2}{m_h^2}, \frac{Q^2}{\mu^2}\right) \otimes f_i\left(n_l, \mu^2\right), \tag{1.57}$$

where the index i runs over the n_l light quarks, and x dependence has been omitted. We can then separate the structure function into a part corresponding to contributions that only involve light quarks F^L and a part corresponding to the contributions that involve heavy quarks F^H as

$$F\left(n_l, Q, m_h\right) = F^L\left(n_l, Q\right) + F^H\left(n_l, Q, m_h\right). \tag{1.58}$$

F^H first contributes at $\mathcal{O}(\alpha_s)$ through the production of a quark anti-quark pair from the splitting of a gluon.

Zero mass variable flavor number scheme

Although the FFN scheme is accurate in the region where $Q \lesssim m_h$, this scheme does not resum logs of Q^2/m_h^2 that become large in the region much larger than the heavy quark mass. This can be resolved by using the ZM-VFN scheme in which the heavy quark is treated as a parton at scales above the heavy quark mass, allowing for the resummation of the logs of Q^2/m_h^2 through DGLAP evolution. This scheme differs from the FFN scheme only through the additional parton, and thus the equation for the structure function analogue to Eq. (1.57) can be written as

$$F\left(n_l + 1, x, Q^2\right) = \sum_{i=1}^{n_l+1} C_i\left(n_l + 1, \frac{Q^2}{\mu^2}\right) \otimes f_i\left(n_l + 1, \mu^2\right). \tag{1.59}$$

In this scheme, the heavy quark PDFs are set to zero at scales below the quark mass, and evolved using DGLAP on the same footing as the light partons above the quark mass. This resolves the problem at large scales of the unresummed logs of Q^2/m_h^2. However, since it assumes the heavy quarks to be massless, the mass contributions to the coefficient functions C_i are not accounted for. As a result, the accuracy of the ZM-VFN scheme decreases in regions where m_h/Q becomes large.

General mass variable flavor number scheme

Thus far we have introduced the FFN scheme suffering from unresummed logs of Q^2/m_h^2 spoiling the accuracy of the scheme outside the region $Q \lesssim m_h$, and the ZM-VFN scheme suffering from missing correction in powers of m_h/Q that spoil the accuracy of the scheme outside the region $Q \gg m_h$. Let us now discuss the GM-VFN schemes that interpolate between the FFN scheme and the ZM-VFN scheme to provide a single scheme which reduces the impact of missing corrections when heavy quark masses are involved.

A requirement of a GM-VFN scheme is that the FFN scheme and the ZM-VFN scheme match at very large scales, $Q \gg m_h$, where the heavy quark mass dependence

of Eq. (1.58) in the FFN scheme can be neglected. As such, PDFs in the two schemes are related through a perturbative transformation in the matching point at threshold $\mu = m_h$. Specifically, the n_l flavors up to the matching point are related to the $n_l + 1$ flavors above the matching point through a $n_l \times (n_l + 1)$ transformation matrix A

$$f_i\left(n_l+1,\mu^2\right) = \sum_{j=1}^{n_l} A_{ij}\left(n_l, \frac{\mu^2}{m_h^2}\right) \otimes f_j\left(n_l,\mu^2\right), \tag{1.60}$$

where A_{ij} are known up to NNLO [49,50]. To ensure continuity across the matching point any VFN scheme needs to satisfy the condition

$$\begin{aligned} F^{\mathrm{GM}}\left(m_h^2\right) &= \sum_{i=1}^{n_l} C_i^{\mathrm{GM}}\left(n_l, m_h^2\right) \otimes f_i\left(n_l\right) \\ &= \sum_{i=1}^{n_l+1} C_i^{\mathrm{GM}}\left(n_l+1, m_h^2\right) \otimes f_i\left(n_l+1\right) \\ &= \sum_{i=1}^{n_l+1}\sum_{j=1}^{n_l} C_i^{\mathrm{GM}}\left(n_l+1, m_h^2\right) \otimes A_{ij}\left(n_l, m_h^2\right) \otimes f_j\left(n_l\right), \end{aligned} \tag{1.61}$$

where to obtain the last line we used Eq. (1.60).

From this matching condition follows a minimal description of a GM-VFN scheme [51]:

$$C_j^{\mathrm{GM}}\left(n_l, m_h^2\right) = \sum_{i}^{n_l+1} C_i^{\mathrm{GM}}\left(n_l+1, m_h^2\right) \otimes A_{ij}\left(n_l, m_h^2\right). \tag{1.62}$$

Here it should be noted that the definition of the GM-VFN is not unique. Specifically, the matrix A_{ij} transforms a $n_l + 1$ dimensional vector into a n_l dimensional vector there is a single degree of freedom that allows for terms proportional to powers of m_h/Q to be included in either of the Wilson coefficients in Eq. (1.62).

This freedom allows one to make a scheme choice, which has led to the introduction of a number of GM-VFN schemes. Some of these include:

- The ACOT scheme [52] provided the first GM-VFNS. It ensures Eq. (1.62) is satisfied by including the mass dependence in the Wilson coefficients. It has since been superseded by the simplified-ACOT, or S-ACOT, scheme [53, 54] which uses the freedom in the definition of the transition matrix A_{ij} to allow for a simpler calculation of observables. This is build on the realization that heavy quarks Wilson coefficients can be computed in the massless limit since massive contributions to the Wilson coefficients do vanish in the limit $Q \ll m_h$ and therefore do not spoil the interpolation.

- The TR scheme [55] instead uses the freedom of the definition of the massive Wilson coefficient to constrain the threshold point by ensuring that derivatives of structure functions are continuous. The TR scheme has later been extended to NNLO in the so-called TR' scheme [56].

- The FONLL scheme [57, 58], which is the scheme used within the NNPDF determinations, is based on the idea of summing the observables calculated in the n_l flavor scheme and $(n_l + 1)$ flavor scheme, and subtracts the double counting terms.

1.5 Constraints on PDFs

Having now discussed how PDFs at a scale Q can be determined from a PDF at an initial scale Q_0 through the DGLAP evolution equations, and by exploiting one of the various heavy mass schemes for the treatment of heavy quark distributions, let us now turn to the determination of the x dependence of the PDFs at an initial scale $f(x, Q_0)$.

From the kinematics of the factorized expressions for processes with a single proton in the initial state in Eq. (1.40) or with two protons in the initial state described in Eq. (1.41), it is clear that experimental measurements of the DIS structure function or cross-sections can provide constraints on the PDFs. Nevertheless, beyond constraining the PDFs with data, there are some general statements that can be made about PDFs that will aid us in their determination.

The analysis of DIS experimental data made it possible to obtain the first insights of the structure of the proton. In particular, since the PDFs must yield the quantum numbers that characterize the proton, that it consists of one valence down-quark

$$\int_0^1 dx \left(d\left(x, Q^2\right) - \bar{d}\left(x, Q^2\right)\right) = \int_0^1 dx d_v(x) = 1, \tag{1.63}$$

and two valence up-quarks

$$\int_0^1 dx \left(u\left(x, Q^2\right) - \bar{u}\left(x, Q^2\right)\right) = \int_0^1 dx u_v(x) = 2, \tag{1.64}$$

carrying the proton charge and baryon number, and a so-called sea of light quark pairs $q\bar{q}$. These relations are known as the valence sum rules (VSR).

By definition, the sum of the longitudinal momenta of the constituent partons of a hadron must be equal to the total longitudinal momentum of the hadron. This leads to the following relation which is known as the momentum sum rule (MSR):

$$\sum_{i=q,\bar{q},g} \int_0^1 dx x f_i(x, Q^2) = 1. \tag{1.65}$$

A further requirement on the PDFs, suggested by the momentum sum rules, is that they should vanish as $x \to 1$:

$$f_i(x = 1, Q) = 0, \tag{1.66}$$

since no intrinsic partons can exist with $x > 1$. At the same time, the valence sum rules require the corresponding distributions to be integrable on the entire range in x.

These three constraints can each aid in the determination of the PDFs and lead to the universally applied parametrization choice of the valence-like, singlet and gluon PDFs:

$$f_i\left(x, Q_0\right) = A_i x^{-\alpha_i} (1 - x)^{\beta_i} \mathcal{P}_i(y(x)). \tag{1.67}$$

Here the α and β exponents control the functional form outside the data region, while N is an overall normalization. α and β thus need to be chosen such that the three constraints can be satisfied, while the normalization is enforced via A. Finally, $\mathcal{P}_i(y(x))$ is a parametrization choice that mainly determines the PDFs in the data region. This part of the parametrization is an important subject of current research and much of this thesis is dedicated to its determination.

Chapter 2

NNPDF4.0: towards PDFs with percent-level accuracy

In the previous chapter we have seen how, using collinear factorization theorems, short distance physics corresponding to parton level events where the cross-section can be calculated using perturbative QCD can be separated from the long distance physics encoded in universal structures called parton distribution functions. We have furthermore seen how a PDF at any scale Q^2 can be obtained from a PDF parametrized at a scale Q_0^2 through DGLAP evolution. In this chapter we will discuss how, in the NNPDF4.0 PDF determination [6], the x dependence of the PDFs has been extracted from experimental measurements.

NNPDF4.0 is, at the time of writing of this thesis, the latest set of PDFs released by the NNPDF collaboration and supersedes the previously released NNPDF3.1 [59]. With respect to NNPDF3.1 it includes a wealth of new data from 44 different (mostly LHC) datasets. The methodology has seen some significant improvements, including a novel fitting algorithm based on stochastic gradient descent [60]. Further improvements include a systematic implementation of positivity constraints [61] and integrability of sum rules

Here we provide a summary of the NNPDF4.0 determination, with an emphasis on the fitting methodology. We start by discussing the NNPDF4.0 methodology in Sect. 2.1, in particular we discuss the propagation of data uncertainties, the parametrization of PDFs using neural networks, their training, and the determination of the model hyperparameters. In Sect. 2.2 we present the experimental data on which the NNPDF4.0 determination is based, we emphasize the datasets that have not been included in earlier releases and discuss the stability of PDFs upon the removal or inclusion of individual datasets. Finally, in Sect. 2.3 we list some of the main characteristics of NNPDF4.0.

2.1 Methodology

The determination of PDFs from discrete data is an example of a pattern recognition problem where the aim of a PDF fitter is to provide an accurate representation of an unknown underlying function. This while only the very limited information of

their functional form discussed in Sect. 1.5 is known. Furthermore, the problem of PDF determination has certain characteristic features that should be taken into consideration when developing a fitting framework. First, in most standard pattern recognition problems the output of the model is directly compared to data, instead for PDFs one cannot associate a pair consisting of an input and an output of the model with a single data point. Rather, as can be observed from Eq. (1.40) and Eq. (1.41), each observable depends in a non-linear way on the multiple output PDF functions in the full range of x. The second characteristic is that in order for PDFs to be useful in calculating predictions of observables, it is necessary to provide a description of the full PDF correlations. PDF uncertainties need to reflect the various sources of uncertainty affecting the experimental data, and in practice PDF uncertainties are often the dominant source of uncertainty when calculating predictions of observables in scattering processes [62]. It is further interesting to note that, unlike most pattern recognition problems, PDFs correspond to probability distributions of observables. This is because the observables correspond to stochastic events as a result of the quantum mechanical nature of the interactions.

In this section we will review the general strategy that NNPDF employs for PDF fitting and the propagation of the data uncertainties to PDF uncertainties. Many of the main principles that will be discussed here are not unique to the NNPDF4.0 release and have instead been used within the NNPDF framework for many years. Nevertheless, the release of NNPDF4.0 introduces a number of major methodological improvements with respect to NNPDF3.1. In this section we will mainly focus on these improvements.

2.1.1 Monte Carlo method for error propagation

Various PDF fitting groups aiming to extract the proton PDFs from data employ different techniques to achieve this goal. Of the most commonly used modern PDF sets, MSHT20 [63], CT18 [64], and ABMP16 [65] have been determined using the so-called Hessian method whereby the PDF are parametrized using a polynomial functional form and the PDF uncertainties are represented by symmetric eigenvectors. The NNPDF collaboration on the other hand parametrizes the PDFs using a neural network. This replaces the functional form used within the Hessian method, thereby removing a potential source of bias in the determination of the unknown PDFs from data. Then, to make a faithful estimation of the data uncertainties NNPDF uses the concept of artificial Monte Carlo pseudodata replicas for error propagation.

The result of a PDF determination using the NNPDF framework is a set of $N_{\rm rep}$ Monte Carlo PDF replicas $f^{(r)}$ with $r = 1, \ldots, N_{\rm rep}$ that provide an importance sampling of the probability distribution of the PDFs. Each replica is equally probable, so replicas are statistically uncorrelated, and estimators of functions of the PDFs are given by simple averages over the replicas:

$$\langle X\left[f\right]\rangle = \frac{1}{N_{\rm rep}} \sum_{r=1}^{N} X\left[f^{(r)}\right]. \tag{2.1}$$

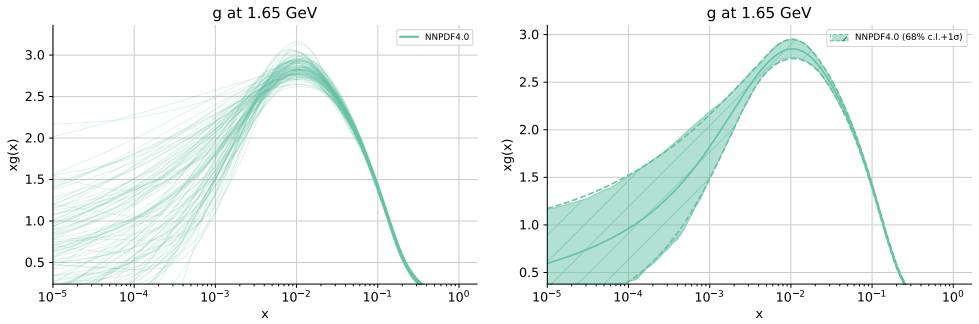

Figure 2.1: A distribution of 100 PDF replicas (left) and the corresponding 1σ interval and 68% confidence level as computed using Eq. (2.1) and Eq. (2.2) with X the identity operator (right). Both are shown for the gluon distribution at 1.65 GeV.

The contribution of the PDFs to the variance of such an estimator is then

$$\mathrm{Var}\left[X\left[f\right]\right] = \frac{1}{N_{\mathrm{rep}}}\sum_{r=1}^{N_{\mathrm{rep}}}\left(X\left[f^{(r)}\right] - \langle X\left[f\right]\rangle\right)^2. \tag{2.2}$$

In this way the uncertainty bands corresponding to any confidence level can be computed from the posterior Monte Carlo distribution, where it can be checked that indeed the 68% confidence interval and the 1σ uncertainty band are in agreement. This is shown explicitly for the gluon PDF in Fig. 2.1 which were computed using Eq. (2.1) and Eq. (2.2) with X the identity operator.

To understand why the PDF replicas are equally probable, it is necessary to understand that they are obtained by performing N_{rep} fits to a corresponding set of N_{rep} independent and identically distributed pseudodata replicas that provide a faithful description of the statistical properties of the experimental dataset. Specifically, the pseudodata replicas are generated in a standard way [44] by shifting the original dataset with a multigaussian distribution given by the covariance matrices corresponding to this dataset.

Producing pseudodata replicas thus provides a way of propagating the experimental uncertainties to the PDFs. As will be discussed in detail in Sect. 4.1, the uncertainty of the experimental data is not the only source of uncertainty present in the PDFs. Other sources of uncertainty include theoretical uncertainties related to the missing contributions of higher orders in the perturbative calculations, tensions between datasets, and uncertainties as the result of an imperfect optimization strategy.

2.1.2 PDF parametrization

The core problem of PDF determination is the extraction of a continuous function from a discrete set of data. This is in itself an ill-defined problem, though by constructing a prior the problem becomes tractable. The PDFs as a function of x only have to be parametrized at a single parametrization scale Q_0 using Eq. (1.67), where the PDFs at any other scale Q can be obtained by solving the DGLAP evolution equations discussed

in Sect. 1.3. One thus needs to choose a parametrization with a level of complexity that allows for a faithful description of the underlying data. A parametrization that is not sufficiently complex will introduce a bias in the resulting PDFs. One place where this problem is still apparent today is in the parametrization of PDFs using a fixed functional from. Namely, within the Hessian approach mentioned before, uncertainties on the fit parameters are determined by performing a least square fit to the data where the PDFs are parametrized using different functional forms constructed from a polynomials in x and $\sqrt{x}$, followed by a standard error propagation [66, 67]. The uncertainties obtained in this naive way generally underestimate the uncertainties of the corresponding predictions, therefore an inflation of the chi-squared distribution (to be discussed in more detail below) corresponding to 1σ is introduced a posteriori using a "tolerance" factor. An important reason for the underestimated uncertainties observed before applying tolerance is that using a fixed functional form provides a PDF parametrization that is too restrictive and thereby introduces a bias.

To address this shortcoming of PDF determination, the idea of employing a neural network parametrization for to the problem of PDF fitting was suggested back in 2002 in Ref. [68]. In this work a neural network was first applied to the determination of the DIS structure function F_2. The relevant underlying principle leading to the idea of applying neural networks to solve the problem of a biased functional form is that, in the limit of an infinite number of parameters, neural networks can reproduce any differentiable function as per the universal approximation theorem [69].

After this initial study, the NNPDF collaboration continued to further develop the idea in Ref. [70], and expand it to the problem of fitting the non-singlet quark distributions in Ref. [43]. Over time more PDF determinations based on the NNPDF methodology were released, for each subsequent release gradually improving the methodology and expanding the dataset. The main intermediate releases were presented in Refs. [44, 45, 59, 71–73], before releasing NNPDF4.0 [6] in 2021.

The parametrization used in the NNPDF4.0 determination can be written as

$$xf_i\left(x, Q_0\right) = A_i x^{(1-\alpha_i)}(1-x)^{\beta_i}\mathrm{NN}_i(x), \tag{2.3}$$

which is a specific variant of the parametrization in Eq. (1.67) in Sect. 1.5 where $\mathrm{NN}_i(x)$ represents a single neural network with a different output parametrizing each flavor i. The neural network is supplemented with a prefactor A_i a polynomial prefactor $x^{(1-\alpha_i)}(1-x)^{\beta_o}$ to improve convergence and ensure the constraints as discussed in Sect. 1.5 are satisfied.

Let us briefly review the neural network model used in the NNPDF4.0 determination. For an extensive review on the subject of neural networks and machine learning, the reader is referred to references such as Ref. [74]. A neural network provides a non-linear mapping from an input space (in this case x) to an output space (in this case the space of PDFs). It does this by utilizing a directed graph structure consisting of multiple layers where the nodes of the consecutive layers are fully connected. A schematic representation of the graph – though without explicitly denoting the direction of the edges – used for the NNPDF4.0 determination is shown in Fig. 2.2. In this figure the blue circles correspond to the nodes of the graph, of which each has an associated function called an activation function. Here the input to each activation function corresponds to the set of all outputs of the previous layer as

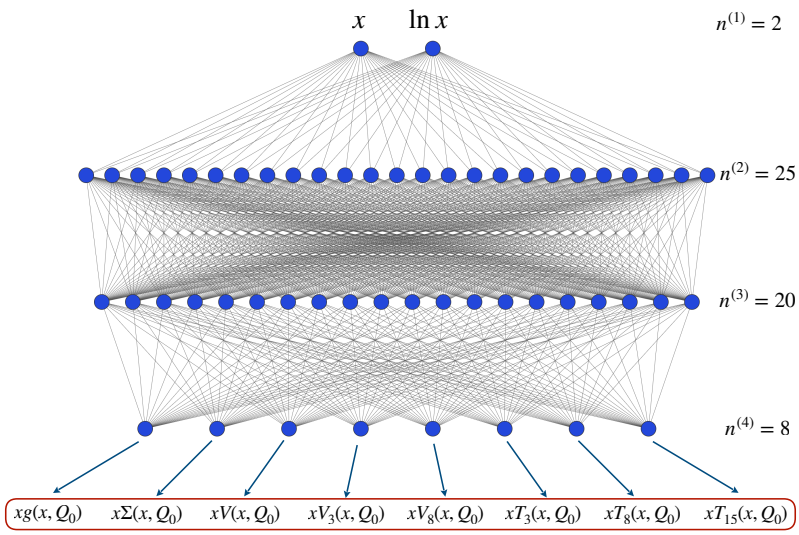

Figure 2.2: neural network parametrization of the PDFs used in NNPDF4.0

represented by the edges. As such, if we know the activation functions of each node, we can evaluate the neural network explicitly and obtain a the function encoded by the neural network.

To obtain this function for the NNPDF4.0 neural network shown in Fig. 2.2, we note that the nodes of the input layer are set as x and $\log x$. This is because PDFs are believed to scale logarithmically at small x and linearly in the large x region [68]. The output of the i-th node in the l-th is then given by

$$\xi_i^{(l)} = g\left(\sum_j w_{ij}^{(l)} \xi_j^{(l-1)} + b_i^l\right). \tag{2.4}$$

where $g(x)$ is the activation function, and the weights $w_{ij}^{(l)}$ and biases b_i^l are the free parameters of the neural network. Note that, as mentioned, the output of the node in the l-th layer is obtained by taking a weighed sum of the outputs of the nodes in the $(l-1)$-th layer. Different choices can be made for the activation function, though it needs to be nonlinear and monotonic. A neural network constructed with only linear activation functions would reduce to a simple linear regression model. A common choice for the activation function is the sigmoid function $g(x) = \frac{1}{1+e^{-x}}$. This function has an two asymptotes: $g(x) = 1$ as $x \to \infty$ and $g(x) = 0$ as $x \to -\infty$, as such it can be thought of as a differentiable function that approximates a step function. The idea of the activation function as a step function provides an intuitive illustration of the connection with neurons in a biological brain, which, depending on the inputs to a neuron either send a signal or not.

Finally, it should be noted that the parameters defining the model – commonly referred to as the models' hyperparameters – such as the number of layer and the

number of nodes per layer, are determined through a semi-automated procedure to be discussed in Sect. 2.1.4.

To prevent the polynomial prefactor of Eq. (2.3) from restricting the functional form and therefore lead to underestimated uncertainties, the α and β exponents are randomly sampled from a range that is determined in a self-consistent manner [71,75]. Specifically, upon making a change to the methodology or dataset, an initial fit is performed for which the effective exponents are calculated for each distribution using

$$\alpha_{\text{eff},i}(x) = \frac{\log f_i(x)}{\log 1/x}, \quad \beta_{\text{eff},i}(x) = \frac{\log f_i(x)}{\log(1-x)}. \tag{2.5}$$

Then for a subsequent fit, the sampling distribution for the α and β exponents is taken to be uniform on the interval determined by taking twice the 68% confidence interval of the corresponding effective exponent. This process is iterated until the sampling domain stabilizes.

The output nodes are parametrized in the evolution-basis defined to simplify evolution, per the discussion in Sect. 1.3, as

$$\begin{aligned}
\Sigma &= u + \bar{u} + d + \bar{d} + s + \bar{s} + 2c, \\
V &= (u - \bar{u}) + (d - \bar{d}) + (s - \bar{s}), \\
V_3 &= (u - \bar{u}) - (d - \bar{d}), \\
V_8 &= (u - \bar{u} + d - \bar{d}) - 2(s - \bar{s}), \\
T_3 &= (u + \bar{u}) - (d + \bar{d}), \\
T_8 &= (u + \bar{u} + d + \bar{d}) - 2(s + \bar{s}), \\
T_{15} &= (u + \bar{u} + d + \bar{d} + s + \bar{s}) - 3(c + \bar{c}), \\
c^+ &= c + \bar{c}, \\
g &= g.
\end{aligned} \tag{2.6}$$

Alternatively one may consider performing a PDF fit in the flavor basis, in which the PDFs are parametrized as $f_i = u, \bar{u}, d, \bar{d}, s, \bar{s}, c, g$. It has however been tested explicitly that the resulting PDFs remain largely unchanged upon changes to the choice of parametrization basis in section 8.4 of Ref. [6]. To perform this check two sets of PDFs were generated, one corresponding to a fit using each basis, it was then observed that the resulting PDFs agree within 1σ. One may then wonder whether, to obtain a conservative estimate of the PDF uncertainty, one should combine the PDFs determined from a fit in both bases. This can be checked by explicitly performing a combination of the two PDF fits using the PDF4LHC15 prescription [76]. The combination method used by the PDF4LHC working group is described in section 4.2 of Ref. [76], and consists of adding the replicas of multiple PDF sets into a single PDF set whereby each replica is given equal weight. This is believed to be the most reliable method of combining PDF sets [77], and will be discussed in more detail in Sect. 4.1. The result of this combination is shown in Fig. 2.3 for the antidown and gluon PDFs, where it is clear that the uncertainties remain unchanged upon performing this conservative combination of the PDFs. The behavior of these two PDFs is representative for the full basis of PDFs.

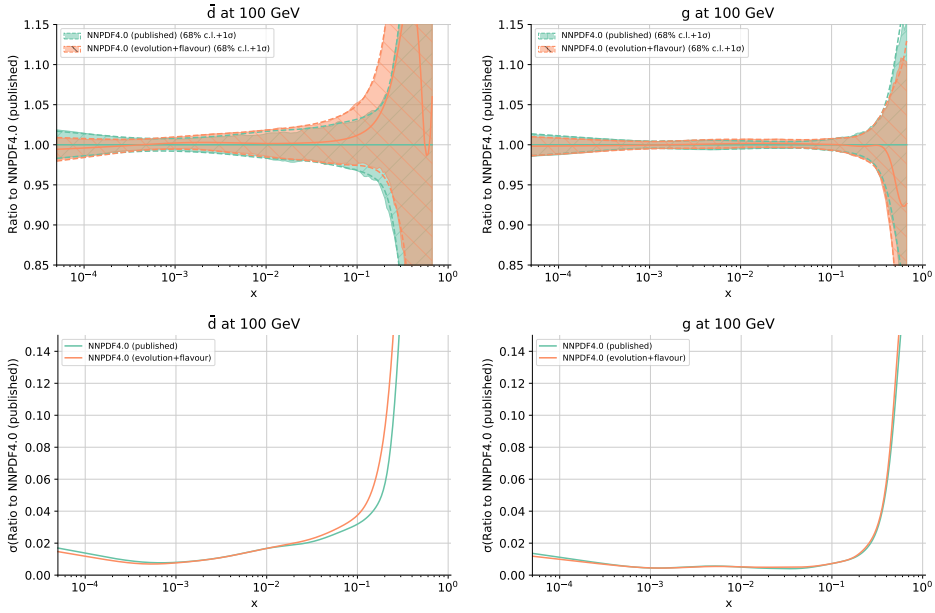

Figure 2.3: The antidown and gluon PDFs from the published NNPDF4.0 baseline set fit in the evolution basis compared to a PDF4LHC-like combination of the baseline fit and a fit in flavor basis.

2.1.3 Fitting framework

After selecting the experimental datasets and defining a model, the free parameters, $w_{ij}^{(l)}$ and $b_i^{(l)}$, of the neural network discussed in Sect. 2.1.2, need to be optimized to obtain a faithful description of PDFs.

A diagrammatic representation of the fitting framework used to train the neural network is shown in Fig. 2.4. The framework takes three external inputs. First, FK tables encoding the partonic cross-sections and evolution equations in a pre-computed format, possibly extended with QCD or electroweak K-factors. Second, a configuration of the hyperparameters determined through a hyperoptimization routine to be discussed in Sect. 2.1.4 below. Finally, the experimental data along with covariance matrices as stored in a common format. This is used to optimize a figure of merit in a computational loop shown in more detail in Fig. 2.6. After the fit has completed the `APFEL` [41] package is used to determine the PDFs at different Q^2 scales. Then, a post fit selection is applied to filter replicas of insufficient quality, before finally storing the replicas that pass the post fit filter in the LHAPDF6 format [78].

Evaluating cross-sections and the modular code structure

Fig. 2.5 shows a schematic representation of the part of the NNPDF fitting code that evaluates the cross-sections and by extension the loss functions. In Fig. 2.4 this is the part enclosed in the blue box.

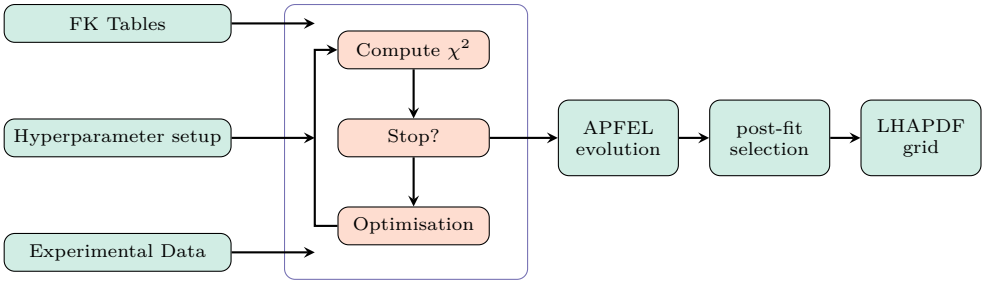

Figure 2.4: Diagrammatic representation of the NNPDF fitting framework. The blue box contains the minimization of the χ^2 figure of merit, whose computation is illustrated in Fig. 2.5.

It takes as input a matrix in x denoted by $\{x_n^{(k)}\}$, where n denotes the experimental dataset, and k labels the nodes in the corresponding x-grid. This matrix is passed to the model presented in Eq. (2.3) consisting of the neural network along with preprocessing and a normalization prefactor. The outputs of the neural network correspond to PDFs $f_i(x_n^{(k)})$ of flavor i at an input scale Q_0. The output can be presented in different linearly dependent bases, though for convenience the evolution basis of Eq. (2.6) is commonly used. The outputs of the neural network are then convoluted with FK tables encoding the theory calculations and the evolution from the parametrization scale to the scale of the hard process. This convolution provides the corresponding observable $\mathcal{O}_n$. For hadronic observables the corresponding calculation is

$$\mathcal{O}_n = \mathrm{FK}^n_{ijkl} f_i(x_n^{(k)}, Q_0) f_j(x_n^{(l)}, Q_0), \tag{2.7}$$

while for DIS observables it reduces to

$$\mathcal{O}_n = \mathrm{FK}^n_{ik} f_i(x_n^{(k)}, Q_0). \tag{2.8}$$

Finally, these predicted observables can be compared to the corresponding experimental values. The distance between the two is expressed using the chi-squared distribution of Eq. (2.9).

In the final step shown in Fig. 2.5, the observables are separated into a training and validation set, resulting in corresponding training loss χ^2_{tr} and validation loss χ^2_{vl}. This is part of the cross-validation technique used to regularize the fitting procedure of which a detailed discussion follows below.

It should be noted that the fitting code – as presented in Ref. [60] – is designed to have a modular structure. This means that each block in Fig. 2.5, as well as the backend used to initialize and train the neural network, can be adjusted independently of the others. An example of this where the default `Tensorflow` backend is replaced by the `evolutionary_keras` package will be discussed below. Another example can be found in Ref. [79]. Here the neural network parametrization is replaced with a simulated quantum circuit implemented using the `Qibo` package [80], and optimization is performed using the `L-BFGS-B` algorithm [81] as implemented in the `scipy` package [82].

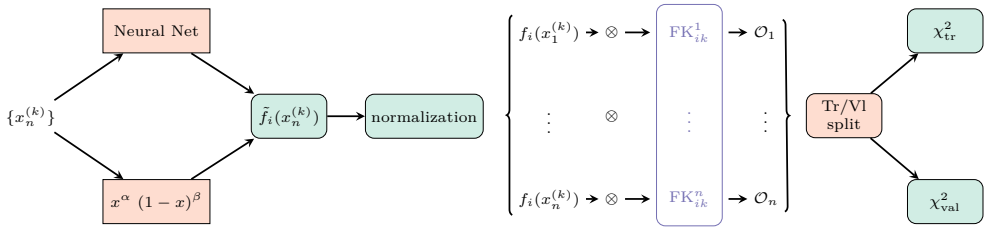

Figure 2.5: Diagrammatic representation of the calculation of the χ^2 in the NNPDF fitting framework as a function of the values of $\{x_n^{(k)}\}$ for the different datasets. Each block indicates an independent component.

The target loss function

Since it is assumed that the experimental uncertainties are Gaussian, a natural choice for the target function is the chi-squared statistic defined as

$$\chi^2 = \sum_{i,j=1}^{N_{\mathrm{data}}} (D_i - P_i)\, \mathrm{cov}_{ij}^{-1}\, (D_j - P_j)\,, \tag{2.9}$$

where D_i are the experimental values of datapoint i, P_i the corresponding prediction of the NNPDF model, and cov_{ij} denotes the covariance between the datapoints with label i and j. The experimental covariance matrix reads

$$\begin{aligned}(\mathrm{cov}_{\mathrm{exp}})_{ij} =& \delta_{ij}\sigma_i^{(\mathrm{uncorr})}\ \sigma_j^{(\mathrm{uncorr})} \\ &+ \left(\sum_{m=1}^{N_{\mathrm{mult}}} \sigma_{i,m}^{(\mathrm{norm})}\sigma_{j,m}^{(\mathrm{norm})} + \sum_{l=1}^{N_{\mathrm{corr}}} \sigma_{i,l}^{(\mathrm{corr})}\sigma_{j,l}^{(\mathrm{corr})}\right) D_i D_j\,,\end{aligned} \tag{2.10}$$

where $\sigma_i^{(\mathrm{uncorr})}$ are the uncorrelated uncertainties obtained by adding the uncorrelated systematic uncertainties and statistical uncertainties in quadrature, m runs over the N_{norm} multiplicative normalization uncertainties, $\sigma_{i,m}^{(\mathrm{norm})}$, and l runs over the N_{corr} other correlated systematic uncertainties, $\sigma_{i,l}^{(\mathrm{corr})}$.

The agreement of a fit to the data is expressed in terms of the experimental χ^2, which is defined as Eq. (2.9) with the covariance matrix Eq. (2.10). This agrees with the usual measure adopted by the community to assess the quality of a fit. However, to avoid the so-called D'Agostini bias [83] that would ensue in the presence of multiplicative uncertainties (such as the luminosity uncertainty) if the covariance matrix as published by experimental collaboration were used for minimization, the t_0 prescription [84] is applied when performing a fit.

In this prescription a so-called t_0 χ^2 is minimized which is defined by a corresponding t_0 covariance matrix that reads

$$\begin{aligned}(\mathrm{cov}_{t_0})_{ij} =& \delta_{ij}\sigma_i^{(\mathrm{uncorr})}\,\sigma_j^{(\mathrm{uncorr})} + \sum_{m=1}^{N_{\mathrm{norm}}} \sigma_{i,m}^{(\mathrm{norm})}\sigma_{j,m}^{(\mathrm{norm})} P_i^{(0)} P_j^{(0)} \\ &+ \sum_{l=1}^{N_{\mathrm{corr}}} \sigma_{i,l}^{(\mathrm{corr})}\sigma_{j,l}^{(\mathrm{corr})} D_i D_j,\end{aligned} \tag{2.11}$$

where $P_i^{(0)}$ corresponds to the central value of a theoretical prediction computed before the fit using as input PDFs the best fit of the previous iteration.

This procedure thus introduces the need for an iterative determination of the t_0 covariance matrix, where the input PDF – sometimes called the t_0 PDF – is iterated until the covariance matrix stabilizes[1]. In practice it has been observed that usually two or three iterations suffice to obtain stability.

Early stopping and post-fit criteria

As discussed before, the commonly used Hessian method of PDF determination applies regularization by relying on a functional form that aims to accommodate the complexity of experimental data without being too flexible. A neural network, on the other hand, removes the need to enforce a specific functional form, and instead the flexibility of a neural network allows for the effective functional form to be determined by the optimization algorithm during the fit. A (sufficiently large) neural network, however, is able to optimize on the experimental data to such an extent that also noise present in the data is learned by the methodology, as opposed to limiting the extraction of information from the data to only genuine features of the data. This is a phenomenon called overfitting, and a regularization procedure is required to prevent overfitting from taking place.

In the NNPDF framework this regularization procedure mainly relies on an early stopping algorithm based on cross-validation as represented by the flowchart shown in Fig. 2.6. The purpose of the stopping algorithm is twofold: determining the best instance of the neural network parameters encountered during training, and deciding when to stop looking for a better instance and instead stop the training.

To identify the best instance of the neural network – this being the instance that generalizes the best to unseen data – a cross-validation method is applied. With cross-validation the full global NNPDF4.0 dataset is divided into a validation dataset and a training dataset, where per experimental dataset a random fraction of 75% of the datapoints is placed in the training set, and 25% is placed in the validation set. Fig. 2.7 illustrates how this split into a training and a validation set is used to identify the optimal instance of the neural network. Namely, during fitting the training set is used to define a training error function χ^2_{tr} which is the target of the optimizer, and thus in principle can be reduced indefinitely as it vanishes asymptotically. This corresponds to the blue curve shown. The validation set, on the other hand, is not seen by the optimizer but nevertheless the corresponding error function χ^2_{val} to this

[1] These iterations can be performed simultaneous with the iteration of the sampling range of the α and β exponents discussed in Sect. 2.1.2.

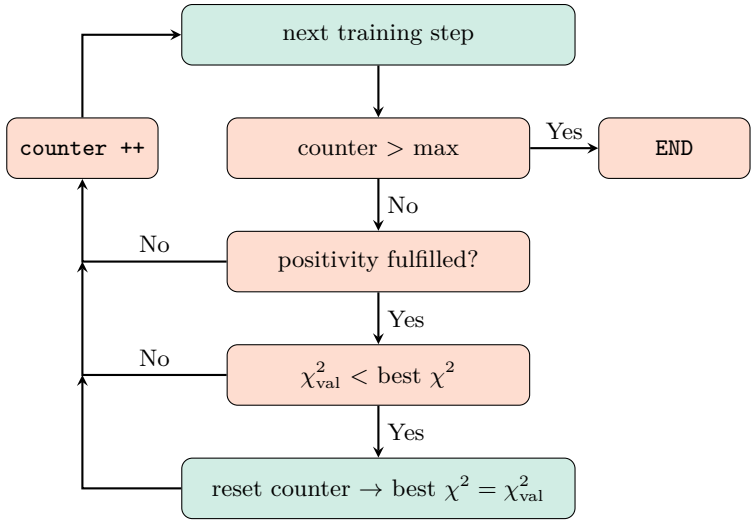

Figure 2.6: Flowchart describing the early stopping algorithm used in NNPDF4.0 to determine the optimal stopping point of the fit based on the look-back cross-validation method.

subset of the data is evaluated at each training epoch. Its value is represented by the orange line. As can be seen, after reaching a minimum value for the χ^2_{val} just before 6000 epochs, it increases again. This can be understood as a result of overlearning, where the optimizer is fitting even the noise present int he training data but no longer generalizes well to unseen data. The final result of the fitting procedure corresponds to the instance at which χ^2_{val} has the smallest value. In Fig. 2.7 the epoch corresponding to the best instance of the neural network is highlighted by the vertical dashed line.

To recognize when a neural network has completed its training, a counter is started when the validation loss χ^2_{val} drops below a certain threshold value. From this point the counter keeps track of the number of epochs that have passed and the training ends if the validation loss has not improved for a given number of epochs. This number is a hyperparameter. If this happens, the training is ended and the model is reset to the instance with the best validation loss. If at no point during training this threshold value for the validation loss is reached, the fit is not considered to be in sufficient agreement with the data and is therefore discarded. Furthermore, for an instance to be considered acceptable, it is checked wether certain positivity criteria [6] are satisfied to ensure that the up, down and strange quark and antiquark PDFs, and the gluon PDF are positive. These constraints follow from Ref. [61] in which it was shown that for PDFs for the individual quark flavors and the gluon as defined in the $\overline{\text{MS}}$ factorization scheme are non-negative. Finally, there is a hard threshold for the number of epochs for which the model is allowed to be trained. If the model was still improving by the time it reaches this threshold, the training will be ended nonetheless.

Once the training of the full set of replicas has completed, certain post-fit criteria are checked and those replicas that do not satisfy all of the post-fit criteria are discarded. As a result, any replica with an arc-length, or χ^2 value as calculated

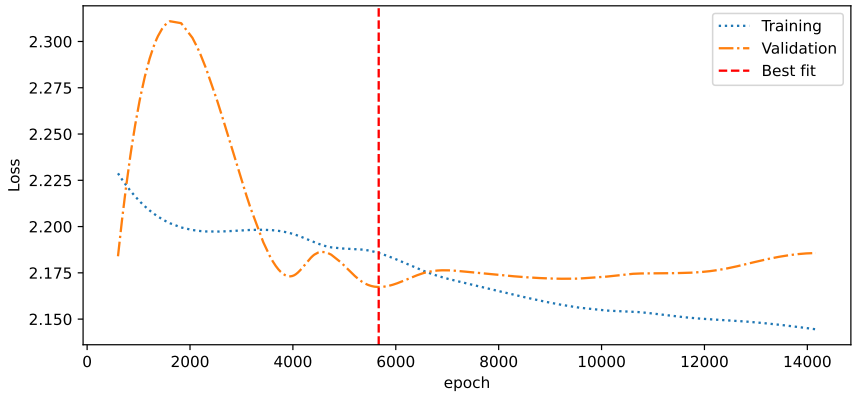

Figure 2.7: Idealized profile of the training (dotted, blue) and validation (dashed, orange) loss in a typical PDF fitting. For ilustration purposes the profiles have been smoothed out and loss penalties applied to the training loss have been removed. The optimization algorithm continues improving the training loss, however, by monitoring the validation loss it is possible to stop the training at the optimal point before we enter in the overfitting domain.

to the experimental data, that is more than 4σ away from the central value of their distribution are discarded. The post-fit check also ensures that integrability of the solutions is satisfied by checking that the inequality

$$\sum_k \left| x_{\rm int}^{(k)} f_i\left(x_{\rm int}^{(k)}, Q^2\right)\right| < \frac{1}{2}, \tag{2.12}$$

is fulfilled for $f_i = V, V_3, V_8, T_3, T_8$, $x_{\rm int}^{(k)} \in \left\{10^{-9}, 10^{-8}, 10^{-7}\right\}$ and the PDFs are evaluated at $Q^2 = 5\,\text{GeV}^2$. The cumulative effect of all post-fit criteria described here is that roughly 1% of the replicas are discarded.

Impact of gradient descent based minimization

Since the methodological update between NNPDF3.1 and NNPDF4.0 includes a complete re-writing of the fitting framework, many changes occurred simultaneously making it challenging to assess the impact that each individual feature that was changed has on the resulting PDFs. One of the main differences though, is the choice of optimization algorithm. Where in NNPDF3.1 a nodel genetic algorithm (NGA) was used to train the neural network, in NNPDF4.0 this has been replaced by stochastic gradient descent (SGD) based algorithms. In particular, as backend to initialize the model and perform the optimization for NNPDF4.0, the `TensorFlow` [85] package provides the necessary tools. This also means that any of the optimization algorithms implemented in `TensorFlow` – such as `RMSprop` [86], `Adagrad` [87], and `Adam` [88] – can be used in the NNPDF framework. As will be discussed in Sect. 2.1.4, for the NNPDF4.0 release the Adam optimizer has been used.

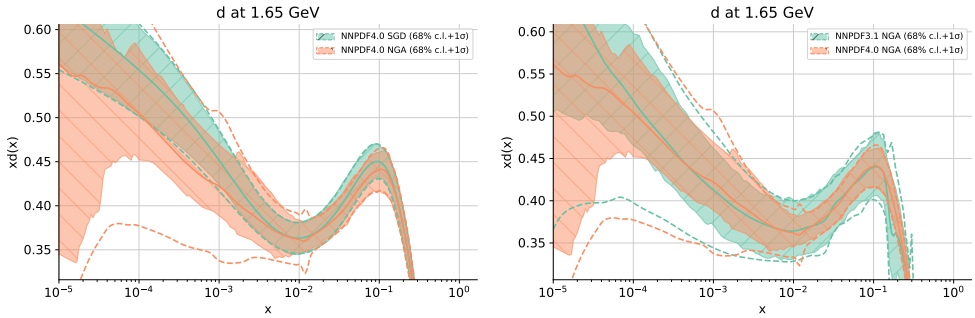

Figure 2.8: Left: comparison of two PDFs generated using the NNPDF4.0 framework, and fitted using the nodal genetic algorithm (orange) and `Adadelta` (green).
Right: comparison of a fits performed with the nodal genetic algorithm in the NNPDF3.1 framework (orange) and the NNPDF4.0 framework (green).
All fits are performed to the DIS-only dataset as defined in Tab. 1 of Ref. [59].

To assess the impact of the change of optimization algorithm a library named `evolutionary_keras` [89, 90] has been developed. It extends the `Model` class of the `Keras` [91] interface to the `TensorFlow` library with genetic based algorithms and allows for the easy implementation of further custom genetic algorithms. It does this while retaining compatibility with `Keras` and can be used in the same way any of the standard gradient descent based optimizers would.

Fig. 2.8 contains plots showing the impact of the gradient descent based algorithm used in NNPDF4.0 compared to the nodal genetic algorithm used in NNPDF3.1. In the left plot of Fig. 2.8 it can be observed that the gradient descent algorithm `Adadelta` [92] results in much smoother and more Gaussian PDFs than the NGA. This difference in smoothness is in fact the main qualitative difference observed between the new NNPDF4.0 framework and the old NNPDF3.1 fitting framework. In Fig. 2.8 it can clearly be seen that the increased smoothness obtained with the NNPDF4.0 framework can – at least for the main part – be attributed to the change in optimization algorithm.

Another genetic optimization algorithm that has been used with the NNPDF framework is the covariance matrix adaptation evolution strategy (CMA-ES) [93, 94]. The CMA-ES has been applied to the determination of structure functions NNFF1.0 [95], and also tested within the NNPDF3.1 framework [96]. In Ref. [95] it was observed that the CMA-ES obtained improved agreement with data, improved consistency, and reduced complexity when compared with the NGA.

`evolutionary_keras` also provides support for the CMA-ES allowing us to check whether fits with the CMA-ES will show similar features compared to its NGA counterpart when it is applied within the NNPDF4.0 framework. Fig. 2.9 shows a comparison between a fit performed using the CMA-ES and a fit performed using the NGA, both within the NNPDF4.0 framework. In this plot a number of differences between the NGA and CMA-ES algorithms can be observed. In particular the PDFs produced using the CMA-ES is are smoother and have larger uncertainties in the large-x region than of those found with the NGA. Furthermore, the number of outliers

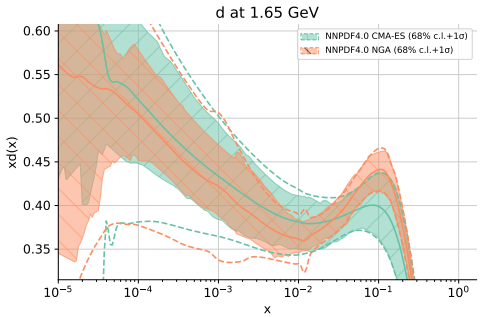

Figure 2.9: Comparison of a fit performed with the nodal genetic algorithm (orange) and the covariance matrix adaptation evolution strategy (green), both using the NNPDF4.0 fitting framework

is reduced, indicating greater consistency, and the agreement with data as measured in terms of the χ^2 has improved.

While the results obtained with the CMA-ES share many features with those obtained with SGD based algorithms, computational costs are much larger and thus it should not be considered an alternative to the SGD based algorithms. This is particularly relevant because the decreased computational cost of the SGD based algorithms allow for the quick and convenient investigation of various setups. An example of what this reduced computational costs makes possible, is the testing of many different combinations of hyperparameters as will be discussed in Sect. 2.1.4.

2.1.4 Hyperoptimization

An important aspect of the NNPDF4.0 methodology involves the determination of the model's hyperparameters. Where in previous NNPDF releases these were determined through a manual, and labour intensive, process of trial and error, for NNPDF4.0 this has largely been replaced by an algorithmic hyperoptimization procedure. In short, the automatic hyperoptimization routine makes use of the improved efficiency achieved with the `TensorFlow` framework. This enables us to test $\mathcal{O}(10^3)$ different hyperparameter setups by performing fits with them and ranking the setups through a k-folds cross-validation algorithm, to be discussed below.

The scan over hyperparameter setups is implemented using the `hyperopt` [97] package which employs a Bayesian algorithm [98] to determine the best combination of hyperparameters.

The output of such a scan is a ranking of the tested setups based on how well they generalize to unseen data as quantified using a figure of merit, Eq. (2.13), to be discussed below. An example of the output of such a scan of around 1500 setups is shown in Fig. 2.10, where each dot corresponds to a different setup. It shows the loss – which is the figure of merit Eq. (2.13) – for a subset of the model's hyperparameters: the number of hidden layers model, the distribution for the initialization of the neural network parameters, the learning rate of the optimizer, and the optimization algorithm. A lower loss corresponds to better model performance. It should be noted that setups with a loss value above 2.5 do exist but correspond to such poor

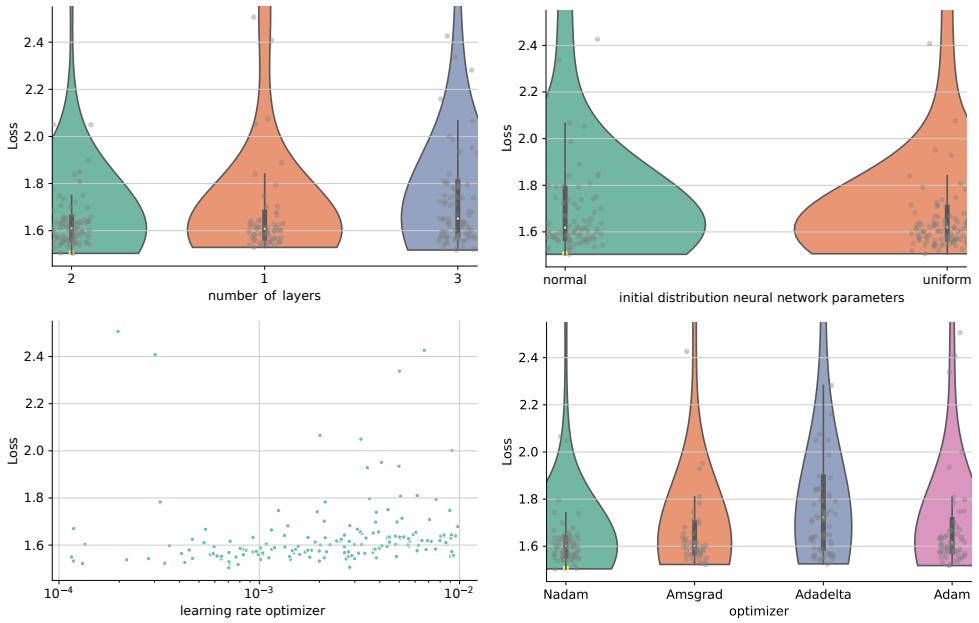

Figure 2.10: Graphical representation of the hyperoptimization loss function L corresponding to a subset of the hyperparameters in a scan based on 1500 configurations. Note that here we only show the subset of parametrizations with $L < 2.5$.

performance that they are not included in these figures. While these figures are not used to make the ranking, as this is done purely based on the loss values, it can provide us with some intuition about how different parameters impact the fit. For example, if we look at the plot comparing the optimization algorithms in the right bottom of the figure, it can be observed that the `Nadam` [99] (which extends the `Adam` optimizer [88] by adding a Nesterov-accelerated adaptive moment [100]) is not only able to obtain a lower loss than any of the other optimizers, but the width of the violin plot also indicates it is able to achieve low losses more consistently than any of the other optimizers. For similar reasons it can be observed that the choice of distribution used for the initialization of the neural networks parameters – as shown in the top right plot – does not seem to have a significant impact on the model performance. It should however be stated that this figure gives a rather naive view of the impact of different hyperparameters. In particular, where this figure shows each hyperparameter in isolation, the performance of the model depends on the combination of hyperparameters and correlations between them, therefore one should be careful when drawing conclusions from the plots in Fig. 2.10.

Hyperparameter correlation

An important motivation for using the `hyperopt` package as opposed to manually tuning the hyperparameters is that best value for a given hyperparameter depends on

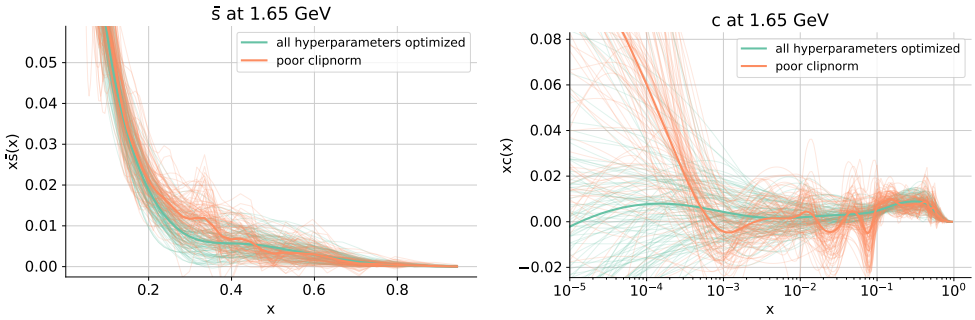

Figure 2.11: Comparison between the results for the antistrange (left) and charm (right) PDFs in two fits, one with all hyperparameters optimized and another where the `clipnorm` value is not optimized.

the settings of the other hyperparameters and cannot be determined independently. To shows this explicitly we will consider the tuning of the `clipnorm` hyperparameter as an example. The value of the `clipnorm` parameter indicates the maximum allowed value for the L2-norm of a tensor corresponding to the gradient calculated during optimization. If the L2-norm of the gradient tensor is less than, or equal to the `clipnorm` value, nothing changes. If, on the other hand, the L2-norm of the gradient tensor is greater than the `clipnorm` value, the tensor is normalized such that the L2-norm is equal to the `clipnorm` value.

Clipping of the gradients is a regularization technique preventing large updates to the neural network parameters as this can cause a numerical overflow leading to instabilities in the training of a neural network. However, one should be careful when setting the value of the `clipnorm` parameter, since a too large value can lead to an insufficiently regularized fit and by extension to overfitting while a value that is too small can prevent convergence. Fig. 2.11 shows a comparison between the antistrange PDF in the large-x region fitted with two different hyperparameter setups (left), and a corresponding comparison for the small-x charm PDF (right). During the determination of one PDF (green), all hyperparameters shown in Table 2.2 are optimized, while for the other PDF (orange), the value of the `clipnorm` parameter is fixed to a large value before optimizing the other hyperparameters. While in both cases the training and validation losses are similar, the resulting PDFs are different and the setup with the fixed `clipnorm` value clearly leads to an overfitted result. This example illustrates the importance of considering all possible hyperparameters when defining the model.

Figure of merit and stability

A sensible choice for the figure of merit is vital for a reliable hyperoptimization routine. Specifically, the figure of merit should quantify the quality of the fit. An obvious choice might be to the validation loss as the figure of merit during hyperoptimization, that is $L = \chi^2_{\text{val}}$. However, because of the stopping algorithm shown in Fig. 2.6, it is already the target of the fitting algorithm itself. Using the validation loss as both a target and a measure of quality is risky, since a target can be obtained in ways that

do not necessarily mean that the outcome is of a high quality, this is also known as "Goodhart's law" [101]. In practice, what this means for us, is that if we were to use validation loss as a measure of quality, the algorithm will select for setups that result in overfitting.

Instead, the validation of a model is performed using k-folds cross-validation (see e.g. chapter 7 of Ref. [102]) schematically represented in Fig. 2.12. The main idea of k-folds cross-validation is to divide the dataset not only into training and validation subsets, but to instead also define a separate test set. The fit will still be performed in the usual manner using the training and validation sets along with the stopping algorithm, but the figure of merit relevant for hyperparameter selection will be defined based on the agreement of the fit to the test set. Since the test set has not been used during the training of the neural network, this provides a way of testing how well the methodology generalizes to unseen data.

The k-folds cross-validation algorithm does not use only a single test set, but instead divides the full dataset into k subsets of data. These subsets are also called folds. Here it is important that each subset is representative of the full dataset both in terms of kinematic range and scattering processes. Because of this, subsets have to be carefully selected, where the $k = 4$ folds used in the NNPDF4.0 determination are shown in Table 2.1. For the NNPDF4.0 release this task has been performed manually, though in Sect. 3.2.1 we propose a method to automate the construction of the folds. Then, $k - 1$ of these folds are divided into training and validation datasets that are used to do a fit, while leaving out a k-th fold that will be used as a test set. This is repeated k times, resulting in k fits where for each fit a different fold is used as the test set. Here a fit is a single replica fit to central values of the experimental data, as opposed to the usual fits performed to artificial pseudodata. This is to save on computational costs since the aim of hyperoptimization is to test a large (order 10^3) number of hyperparameters, so testing each hyperparameter setup by performing fits to many pseudodata replicas is not feasible.

As a proxy for the quality of the fit, the target function of the hyperoptimization algorithm is defined as

$$L = \frac{1}{k} \sum_{i=1}^{k} \chi_i^2, \tag{2.13}$$

where χ_i^2 is the χ^2 evaluated to the datasets in the i-th fold using the PDF obtained with the i-th fit, where for the determination of the i-th PDF, the i-th fold was left out. The optimal hyperparameter setup is the setup for which L is minimized.

Alternative definitions of the figure of merit may also be used, though if two figures of merit are equally well motivated the result should be the same in both cases. For example, instead of defining the loss as the average value of the χ_i^2, one could consider defining it as the worst χ_i^2 obtained with a given hyperparameter setup:

$$L = \max\left(\chi_1^2, \chi_2^2, \chi_3^2, \ldots, \chi_k^2\right). \tag{2.14}$$

This has been checked explicitly, Fig. 2.13 shows a comparison between a setup found by optimizing for the "average" as defined in Eq. (2.13) and a setup found by optimizing for the lowest "max" loss as defined in Eq. (2.14). The specific hyperparameter setups for both cases are shown in Table 2.2. It is clear from Fig. 2.13 that the results in

Fold 1		
CHORUS σ^{ν}_{CC}	HERA I+II $\sigma^{p}_{\rm NC}$ e^+ (920 GeV)	BCDMS F_2^p
LHCb $Z \to ee$ 7 TeV	ATLAS W, Z 7 TeV ($\mathcal{L} = 35$ pb^{-1})	CMS Z p_T 8 TeV
E605 σ^p	CMS DY 2D 7 TeV	CMS 3D dijets 8 TeV
ATLAS single t 7 TeV ($1/\sigma d\sigma/dy_{\bar{t}}$)	ATLAS single t R_t 7 TeV	CMS $t\bar{t}$ ℓ+jets 8 TeV ($1/\sigma d\sigma/dy_{t\bar{t}}$)
CMS single t R_t 8 TeV		

Fold 2		
HERA I+II $\sigma^{p}_{\rm CC}$ e^-	HERA I+II $\sigma^{p}_{\rm NC}$ e^+ (460 GeV)	HERA I+II $\sigma^{b}_{\rm NC}$
NMC $\sigma^{{\rm NC},p}$	NuTeV $\sigma^{\bar{\nu}}_{CC}$	LHCb $Z \to ee$ 8 TeV
CMS W electron asymmetry 7 TeV	ATLAS Z p_T 8 TeV ($p_T, m_{\ell\ell}$)	D0 W muon asymmetry
E866 σ^p (NuSea)	ATLAS isolated γ prod. 13 TeV	ATLAS dijets 7 TeV, R=0.6
ATLAS single t 8 TeV ($1/\sigma d\sigma/dy_{\bar{t}}$)	CMS $\sigma^{\rm tot}_{tt}$ 7,8 TeV	CMS single t $\sigma_t + \sigma_{\bar{t}}$ 7 TeV

Fold 3		
HERA I+II $\sigma^{p}_{\rm CC}$ e^+	HERA I+II $\sigma^{p}_{\rm NC}$ e^+ (575 GeV)	NMC F_2^d/F_2^p
NuTeV σ^{ν}_{CC}	LHCb $W, Z \to \mu$ 7 TeV	LHCb $Z \to ee$ 13 TeV
ATLAS $\sigma^{\rm tot}_{tt}$ 7,8 TeV	ATLAS W^++jet 8 TeV	ATLAS high-mass DY 7 TeV
CMS W muon asymmetry 7 TeV	E866 $\sigma^d/2\sigma^p$ (NuSea)	CDF Z differential
ATLAS W, Z 7 TeV ($\mathcal{L} = 4.6$ fb^{-1}) central	ATLAS single t 8 TeV ($1/\sigma d\sigma/dy_t$)	CMS $\sigma^{\rm tot}_{tt}$ 5 TeV
CMS $t\bar{t}$ 2D 2ℓ 8 TeV ($1/\sigma d\sigma/dy_t dm_{t\bar{t}}$)		

Fold 4		
CHORUS $\sigma^{\bar{\nu}}_{CC}$	HERA I+II $\sigma^{p}_{\rm NC}$ e^+ (820 GeV)	LHCb $W, Z \to \mu$ 8 TeV
ATLAS single t R_t 13 TeV	LHCb $Z \to \mu\mu$ 13 TeV	ATLAS W^-+jet 8 TeV
ATLAS low-mass DY 7 TeV	ATLAS Z p_T 8 TeV (p_T, y_Z)	CMS W rapidity 8 TeV
D0 Z differential	CMS dijets 7 TeV	ATLAS single t 8 TeV ($1/\sigma d\sigma/dy_t$)
ATLAS W, Z 7 TeV ($\mathcal{L} = 4.6$ fb^{-1}) forward	CMS single t R_t 13 TeV	

Table 2.1: The four folds in which the NNPDF4.0 dataset is divided for the k-folds hyperoptimisation procedure represented in Fig. 2.12.

both cases are equivalent, even though the hyperparameters are completely different. This shows the stability of the hyperoptimization routine.

It should be noted however, that the value of L for each individual hyperparameter setup tested in this way is susceptible to non-negligible random fluctuations and hence it is ill-advised to select the model with lowest L without a second thought. While the hyperoptimization routine is useful in producing a ranking of good hyperparameter configurations, the configuration ranked first is not necessarily the best. To confidently identify the best setup among those with hyperparameter configurations that resulted in a low loss L, the hyperoptimization routine is followed by a production of PDFs with the default sample size of 100 replicas. As a final step of the hyperparameter selection, these PDFs consisting of 100 replicas are closely studied to identify the preferred

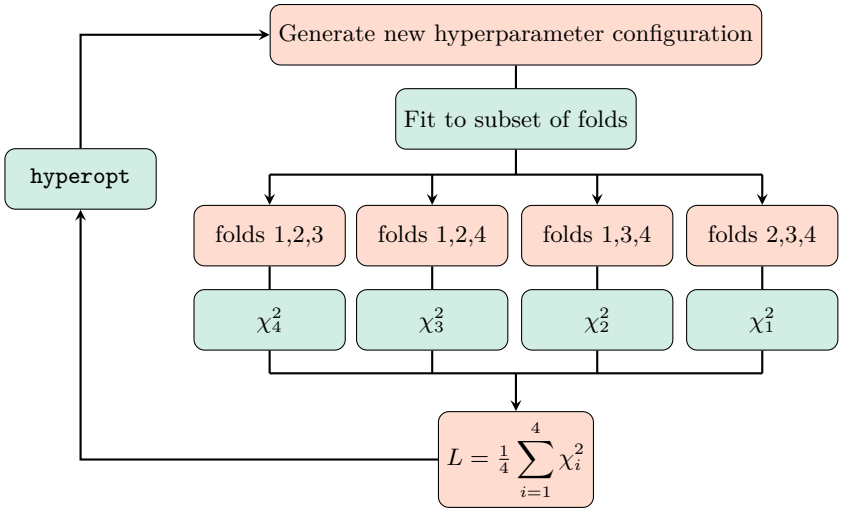

Figure 2.12: Diagrammatic representation of the k-fold algorithm used for the hyperparameter optimization. Here the number of folds equals four, i.e. $k = 4$.

setup and discard those in which more subtle features of overfitting or underfitting are recognized. Indeed, the procedure is still not fully automated and even though the selection of hyperparameters has been greatly improved, human experience is still needed in the final step.

An improvement to the hyperoptimization routine is proposed in Sect. 3.2, which includes a quantitative measure for the detection of overfitting which will be introduced in Sect. 3.2.2.

Baseline hyperparameters for NNPDF4.0

A k-folding hyperoptimization, as described above, has been performed to determine the best values of the hyperparameters that have been used for the NNPDF4.0 determination. These are listed in Table 2.2. The hyperparameters include the network architecture, the type of activation function, the Glorot-type [103] initializer, the optimizer, the values of the learning rate and of `clipnorm`, the maximum number of iterations and the stopping patience, and the initial values of the Lagrange multipliers for the PDF positivity and integrability constraints (see Sect. 3.1 of Ref. [6] for a discussion on the implementation of integrability and positivity constraints using Lagrange multipliers). The ranges of the hyperparameters that are sampled by the hyperoptimization algorithm are chosen empirically: we start out conservatively with very wide ranges, and once we are confident that the optimal value of a given hyperparameter falls within a sub-domain of this (conservative) range, we adjust the sampled domain accordingly to limit the runtime and computational resources of the hyperparameter scan.

In Table 2.2 we show both the optimal hyperparameters for our default methodology, based on the hyperoptimization loss defined in Eq. (2.13), as well as the hyperparameter values obtained with the different choice of loss function Eq. (2.14).

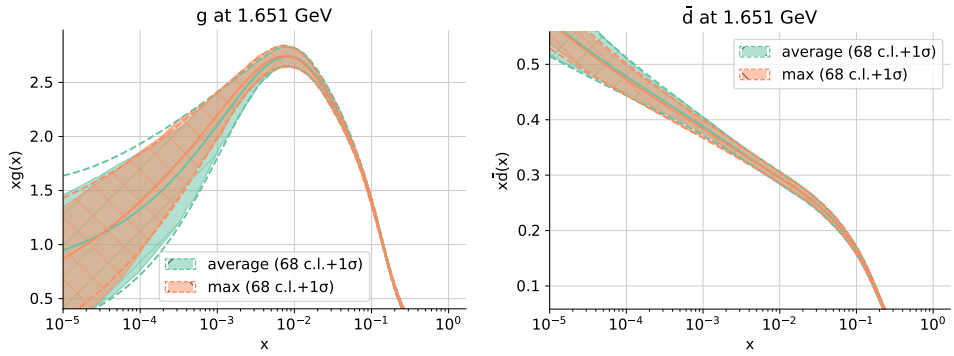

Figure 2.13: Comparison between the gluon (left) and antidown (right) PDFs at $Q = 1.65$ GeV found by using methodologies in which hyperparameters are selected based on the "average" loss function Eq. (2.13) (green) or the "max" loss function Eq. (2.14) (orange).

As mentioned, both different choices of loss function (see Fig. 2.13) lead to equivalent results, but the corresponding hyperparameter values can be quite different. For instance, the optimal architecture for fits based on the alternative loss function Eq. (2.14) has more than twice the number of neurons in the hidden layers compared to the baseline settings.

We now specifically discuss the hyperoptimization and its results for our default choice. Concerning the network architecture, until NNPDF3.1, each PDF was parametrized with an individual neural network. While the number of independently parametrized PDFs was gradually increased, this remained unchanged since NNPDF1.0 [44]. Now the hyperoptimization scan is run with a single network which outputs the value of all PDFs. So while in all NNPDF fits up to and including NNPDF3.1 $\mathrm{NN}_i(x)$ in Eq. (2.3) denotes the i-th neural network, in NNPDF4.0 it indicates the activation state of the i-th neuron in the last layer of the neural net. The architecture selected by the hyperoptimization is 2-25-20-8 with hyperbolic activation functions except for the final linear layer, and it is shown in Fig. 2.2.

The NNPDF4.0 architecture has 763 free parameters, to be compared to a total of 296 parameters for the NNPDF3.1 neural nets. We emphasize however that a larger network does not necessarily imply better performance, and that for a given dataset there exists a lower bound to the number of required free network parameters but probably not an upper one. Given comparable performance, smaller networks are preferred in order to reduce the computational costs.

Hyperoptimization stability

The main goal of the hyperoptimization procedure is to identify the best optimization settings for the current problem of determining the PDFs. This raises the question of deciding in which cases a new hyperoptimization would be required. Our current understanding encompasses changes to the experimental data, the theoretical description, and methodological choices (such as the choice of PDF basis).

Parameter	NNPDF4.0	L as in Eq. (2.14)
Architecture	2-25-20-8	2-70-50-8
Activation function	hyperbolic tangent	hyperbolic tangent
Initializer	`glorot_normal`	`glorot_uniform`
Optimizer	`Nadam`	`Adadelta`
Clipnorm	6.0×10^{-6}	5.2×10^{-5}
Learning rate	2.6×10^{-3}	2.5×10^{-3}
Maximum # epochs	17×10^{3}	45×10^{3}
Stopping patience	10% of max epochs	12% of max epochs
Initial positivity $\Lambda^{(\mathrm{pos})}$	185	166
Initial integrability $\Lambda^{(\mathrm{int})}$	10	10

Table 2.2: The baseline hyperparameter configuration (left) selected using the k-folds hyperoptimization procedure with hyperoptimization loss Eq. (2.13) and used to perform the NNPDF4.0 fits in the evolution basis. We also show an configuration selected using the alternative hyperoptimization loss Eq. (2.14) (right).

We have checked that the procedure is quite stable upon reasonably small changes of the dataset. In particular, the datasets included in Table 2.1 do not correspond exactly to the datasets included in the final dataset as listed in App. A, since the final appraisal of the data to be included was performed after the methodology was set. Furthermore, when removing datasets the given methodology remains viable, though in principle there might be a computationally more efficient one giving the same results for the small datasets. Of course in principle the only way of being absolutely certain whether a new hyperoptimization is needed or not is to actually perform it.

On the other hand, a substantial change in methodology or dataset generally needs a new hyperoptimization. An example of this is the flavor basis plot included in the combination using the PDF4LHC15 prescription shown in Fig. 2.3. Likewise, the addition of a large number of new datasets affecting kinematic regions or PDF combinations for which currently there is little or no information might have an impact on the fit sufficient to warrant a new run of the hyperoptimization procedure.

Note that the need for a re-hyperoptimization upon large changes to the dataset does not imply that the uncertainties obtained with the methodology are not robust. The hyperparameters are selected to accurately fit a given dataset; for example if a specific subset of data were to be removed such that the dataset spans a smaller kinematic range, presumably a less 'aggressive' methodology is required to accurately describe the data (though likely at the cost of an increase in uncertainty). Likewise, the inverse would be true for a dataset spanning a much larger kinematic range; in such a scenario, if the methodology were not updated, likely not all features of the data would be fitted to the same precision as could be obtained with a more aggressive methodology. This then leads to a less precise – though generally not less accurate – determination of the PDFs than could be obtained upon a re-hyperoptimization of the hyperparameters. These are hypothetical scenarios to sketch a picture of how the hyperparameters can depend on changes to the dataset, for this reason the definition of aggressiveness of the methodology is intentionally left arbitrary.

2.2 Experimental data

Having discussed the methodology employed for the NNPDF4.0 determination, we continue by providing a brief overview of the global dataset that is the basis for the NNPDF4.0 release, as well as the theoretical calculations corresponding to the datasets.

The kinematic coverage in the (x, Q^2) plane of the NNPDF4.0 dataset entering the default NNLO fit is shown in Fig. 2.14. For hadronic data the corresponding kinematic values have been determined using LO kinematics. Whenever an observable is integrated over rapidity, the center of the integration range is used to compute the values of x. The datapoints corresponding to datasets that are new in NNPDF4.0 are indicated with a black edge.

In essence, the baseline NNPDF4.0 dataset is largely a superset of the baseline NNPDF3.1 dataset, extending the NNPDF3.1 global dataset with 44 new datasets. There are a few exception to this rule, namely, the single-inclusive jet data in NNPDF3.1 have been replaced with corresponding dijet datasets, and various minor changes have been made to certain datasets already present in NNPDF3.1, or their theoretical treatment. This includes datasets being replaced by more recent measurements of the same cross-sections, and in some cases more differential distributions of the same process have been included due to correlations becoming available. In terms of theoretical treatment some minor changes have been made as well, for example, the fixing of a bug in `APFEL` affecting the computation of the NLO charged current structure functions, and updating the branching ratio of charmed hadrons into muons with the value from PDG 2020 [104]. For a detailed discussion of changes made regarding datasets also included in the NNPDF3.1 determination, the reader is referred to Chapter 2 of Rev. [7].

We will briefly go over the datasets that have not been included in earlier determinations by NNPDF. For the first time, this includes data from the LHC Run II at a collision energy of 13 TeV. While some of these new measurements correspond to processes already present in the baseline NNPDF3.1 dataset, they also include data from direct photon production, single top production, dijet production, and gauge boson production with jets which have not been included in any previous NNPDF determination.

In particular, the new datasets with respect to NNPDF3.1 are the following:

DIS. The ration $\mathcal{R}_{\mu\nu}$ of dimuon to inclusive neutrino-nucleus CC DIS cross-sections from NOMAD [105] as a function of the neutrino beam energy. The H1 [106] and ZEUS [107] measurement of charm and bottom production cross-sections in DIS have been replaced with the combined measurement of Ref. [108].

DIS jet. DIS single-inclusive jet and dijet production data from ZEUS [109–111] in the high-Q region and H1-HeraII [112, 113] in the high- and low-Q regions.

Fixed-target DY. The recent SeaQuest [114] measurement of the production of a Z boson decaying into muon pairs.

Incl. W and Z. The ATLAS measurements of the W and Z differential cross-section at $\sqrt{s} = 7$ TeV in the central and forward rapidity regions [115], of double and triple differential DY lepton pair production at $\sqrt{s} = 8$ TeV [116, 117], of

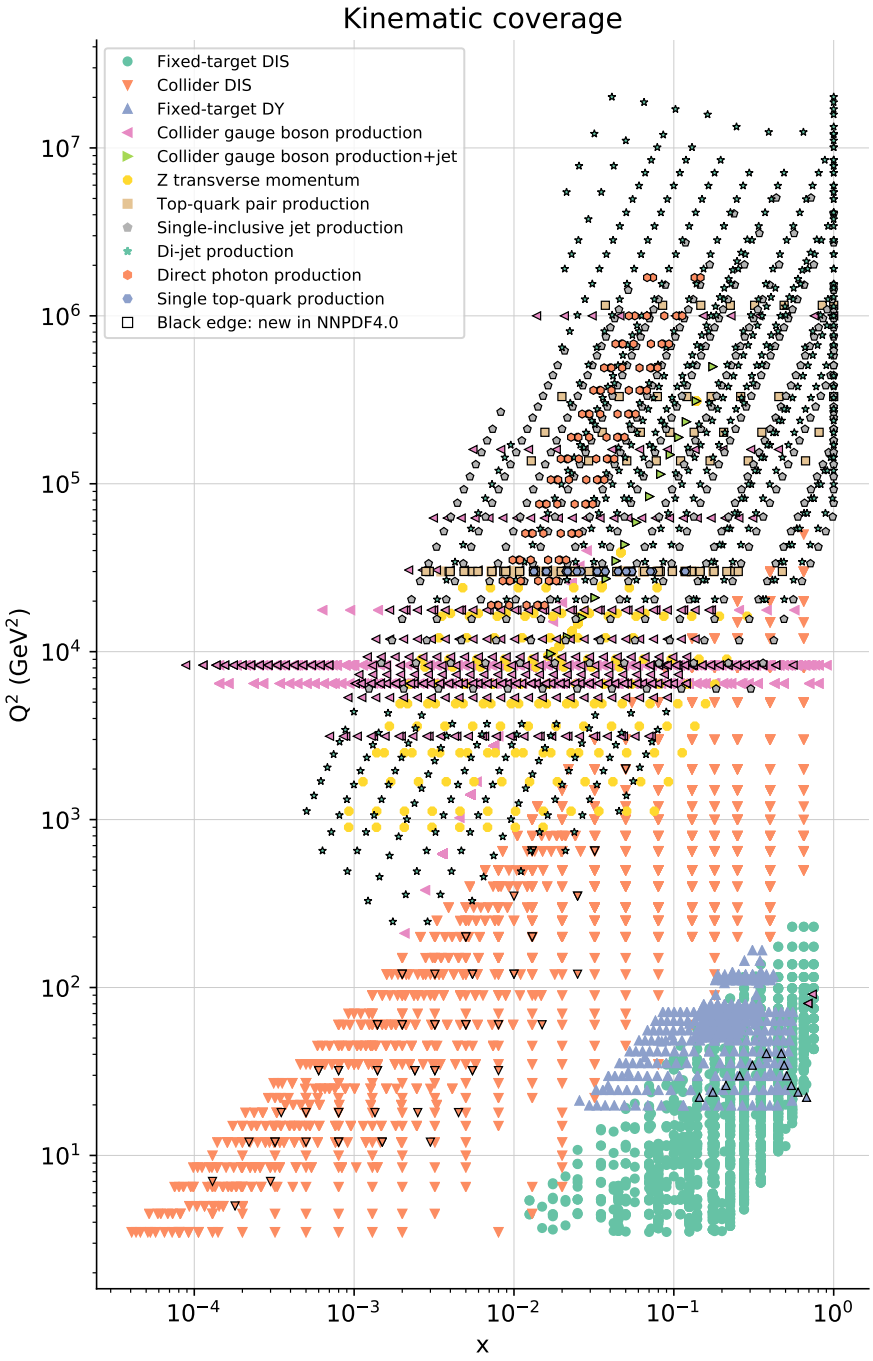

Figure 2.14: The kinematic coverage of the NNPDF4.0 dataset in the (x, Q^2) plane. Points with a black edge around it are new in NNPDF4.0 and were not included in any previous NNPDF release.

W production and decay at $\sqrt{s} = 8$ TeV [118], and of W and Z decay into leptons at $\sqrt{s} = 13$ TeV [119]. The LHCb measreument of the Z cross-section at $\sqrt{s} = 13$ TeV [120].

W+jet. The measurement of W boson production with additional jets from ATLAS [121] at $\sqrt{s} = 8$ TeV. The measreuments of W production with a charm jet from ATLAS [122] at $\sqrt{s} = 7$ TeV and CMS [123] at $\sqrt{s} = 13$ TeV.

Top pair. The ATLAS [124] differential and CMS [125] double differential normalized cross-sections measured at $\sqrt{s} = 8$ TeV. The ATLAS total cross-section [126] and the CMS absolute differential distributions in the lepton+jets channel [127] and in the dilepton channel [128].

Dijet. Single-inclusive jet production from ATLAS [129] and CMS [130] at $\sqrt{s} = 8$ TeV. Dijet production from ATLAS [131] and CMS [132] measurements at $\sqrt{s} = 7$ TeV and the CMS measurement [133] at $\sqrt{s} = 8$ TeV.

Direct photon. Isolated photon production measurements from ATLAS at $\sqrt{s} = 8\,\text{TeV}$ [134] and $\sqrt{s} = 13\,\text{TeV}$ [135].

Single top. Single top production from ATLAS [136–138] and CMS [139–141] measurements at $\sqrt{s} = 7$, 8 and 13 TeV.

Of these processes, jet production in dis, top pair production, dijet production, direct photon production and t-channel single top production have not been included in previous NNPDF releases. For an exhaustive list of all included datasets in the NNPDF4.0 determination we refer the reader to App. A. For a description of how the theoretical predictions corresponding to the measurements are obtained, we refer the reader to Sect. 2 of Ref. [6].

2.2.1 The impact of datasets with tension

The baseline dataset described above is constructed to be maximally consistent, this is discussed in detail in section 4 of Ref. [6]. However, here we will briefly assess the sensitivity of the resulting PDF upon the exclusion of those datasets where some indication of inconsistency was found but that were eventually included in the baseline dataset.

To identify possible dataset inconsistencies, three quantities are used. First is the total χ^2 per datapoint, second is the distance in terms of standard deviations that the χ^2 per datapoint evaluated for a given dataset differs from its expected value:

$$n_\sigma \equiv \frac{\chi^2 - 1}{\sigma\left[\chi^2\right]} = \frac{\chi^2 - 1}{\sqrt{2/N_{\text{dat}}}}, \tag{2.15}$$

and third is the inverse of the smallest eigenvalue of the experimental correlation matrix. This provides a proxy for the stability of the experimental covariance matrix. For each of these quantities a threshold value is decided, and for those datasets where the threshold is crossed, additional checks are performed. In particular this involves giving additional weight to the identified datasets one-by-one, and seeing how this impacts the fits.

Since the decision of whether to include a given dataset or not is not a simple one, but rather is based on a combination of many different factors, one may wonder how much these decisions impact the PDFs. One way to observe that these decisions do not have a significant impact is shown in Fig. 2.15. Here seven new PDF determinations have been performed, each one by removing one of the datasets with particularly large values for the estimators described above, these are ATLAS 7 TeV dijets [131], NMC σ_p [142], BCDMS F_2^p [18], HERA I+II charm [108], CMS 7 TeV dijets [132], HERA inclusive [143], and E866 σ_p [144]. These seven PDFs are then combined using the PDF4LHC15 prescription, which means that the combined PDF is an unweighted combination of the seven different PDF sets.

Fig. 2.15 compares the up, antiup, gluon and down quark PDFs and their uncertainties of the NNPDF4.0 baseline determination to this combination of the seven PDF sets where different datasets have been omitted. From this it can be concluded that the results are stable and changes are comparable with statistical fluctuations.

2.3 Important features of NNPDF4.0

Here we will discuss some of the main features of the NNPDF4.0 determination. So far in this chapter we discussed two main sources of changes to the NNPDF4.0 determination with respect to the earlier NNPDF3.1 determination: the dataset and the methodology. Here we will study the impact of the new methodology, as well as the impact of the new dataset, independently. We will then study the implications of the NNPDF4.0 PDF set for hadron collider phenomenology. Finally, we will discuss the implications of the independent determination of the charm PDF.

2.3.1 Impact of the new data

As discussed above in Sect. 2.2, the NNPDF4.0 dataset is not a pure extension of the NNPDF3.1 dataset. Instead, also some changes have been made regarding the treatment of data already included in the NNPDF3.1 determination. These updates mostly consist of updated measurements of the same observable, and updates to the theory calculations. The global dataset that incorporates these updates to the NNPDF3.1 dataset has been dubbed the NNPDF3.1-like dataset, which is the dataset that we will compare the full NNPDF4.0 dataset to here.

To assess the impact of the new data – and also the impact of the new methodology in Sect. 2.3.2 – we study a quantity relevant for LHC physics, namely, the parton luminosities as a function of the invariant mass of the final state m_X at $\sqrt{s} = 14\,\text{TeV}$. While various definitions of the parton luminosity as a function of the invariant mass can be used, here we will define it as

$$\mathcal{L}_{ij}\left(M_X, \sqrt{s}\right) \equiv \sum_{ij}^{\text{channels}} \frac{1}{s} \int_\tau^1 \frac{dx}{x} f_i\left(x, M_X\right) f_j\left(\tau/x, M_X\right), \tag{2.16}$$

where i and j are the parton flavor indices, and $\tau = M_x^2/s$.

The impact of the new data is assessed by comparing the luminosity of the baseline NNPDF4.0 PDF set to a PDF set determined using the same NNPDF4.0 methodology,

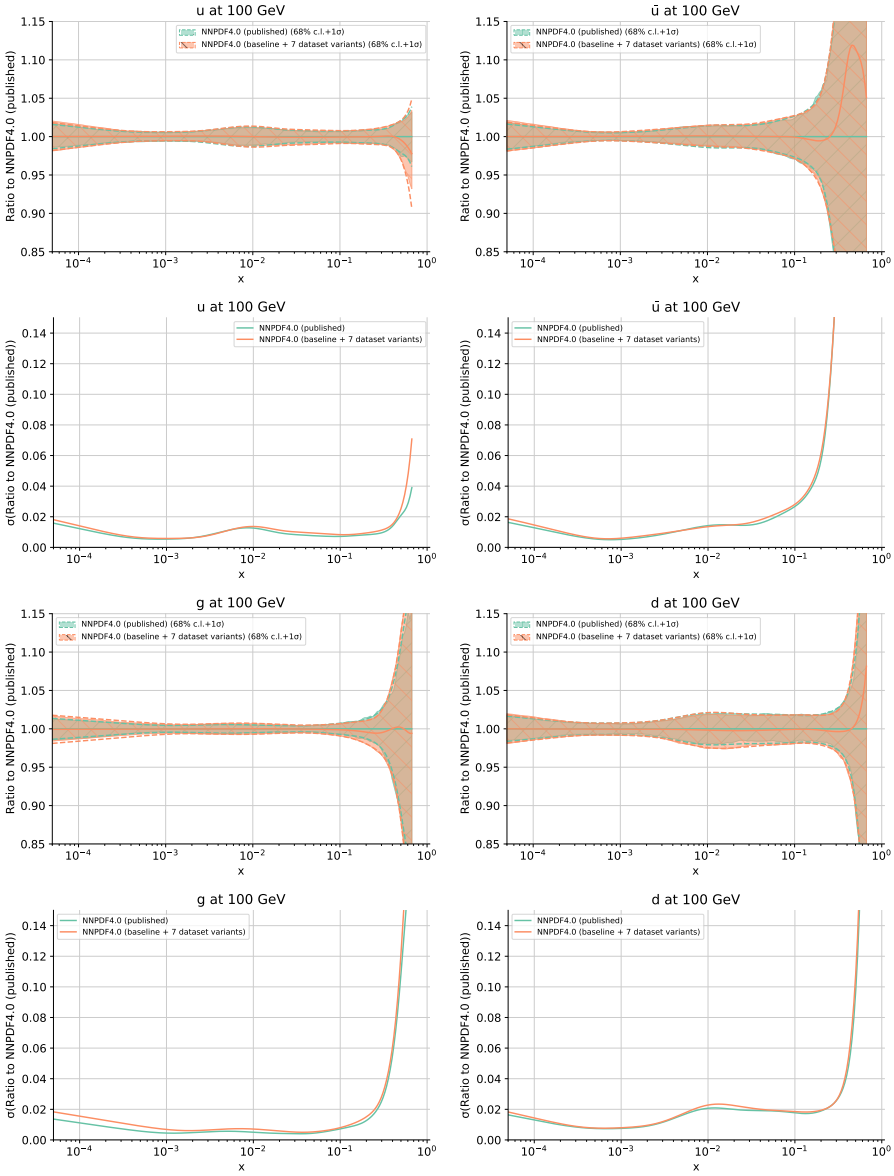

Figure 2.15: The NNPDF4.0 baseline PDFs and their relative uncertainties compared to a combination of PDFs where in their determination datasets with poor statistical estimators have been omitted one-by-one. The combination is performed using the PDF4LHC15 prescription.

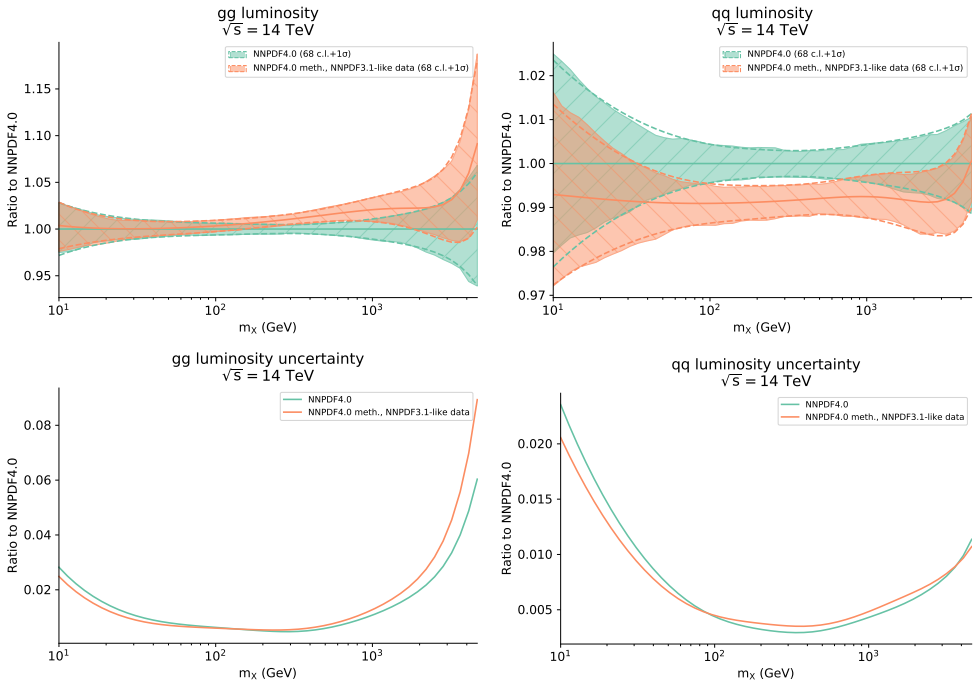

Figure 2.16: A comparison of the gg and qq luminosities Eq. (2.16) as a function of the invariant mass (top) and their relative 1σ uncertainties (bottom), between the baseline NNPDF4.0 PDF set (green) and a PDF set determined using the same NNPDF4.0 methodology but fitted to the NNPDF3.1-like dataset (orange).

but instead fitted to the NNPDF3.1-like dataset. In Fig. 2.16 we compare the gluon-gluon and quark-quark channel luminosities of both PDF sets along with their 1σ uncertainties. Even though the uncertainties remain largely unchanged there is a clear shift of the central value that, for some values of the invariant mass, at the 2σ level. From this it can be concluded that while the extended NNPDF4.0 dataset does not improve the precision of the PDF fit, the NNPDF4.0 dataset does result in an improved accuracy as a result of the larger amount of information included in the NNPDF4.0 dataset.

2.3.2 Impact of the new methodology

Similar to the assessment of the impact of the new data above, here we will again study the luminosities as a function of the invariant mass for two different PDF sets.

In Fig. 2.17 we compare the luminosities of the baseline NNPDF4.0 PDF set to a PDF set determined using the NNPDF3.1 methodology and the NNPDF4.0 dataset. From these plots it is clear that while the two methodologies are in perfect agreement, the PDF set obtained with the NNPDF4.0 methodology achieves a much higher precision that that obtained using the NNPDF3.1 methodology.

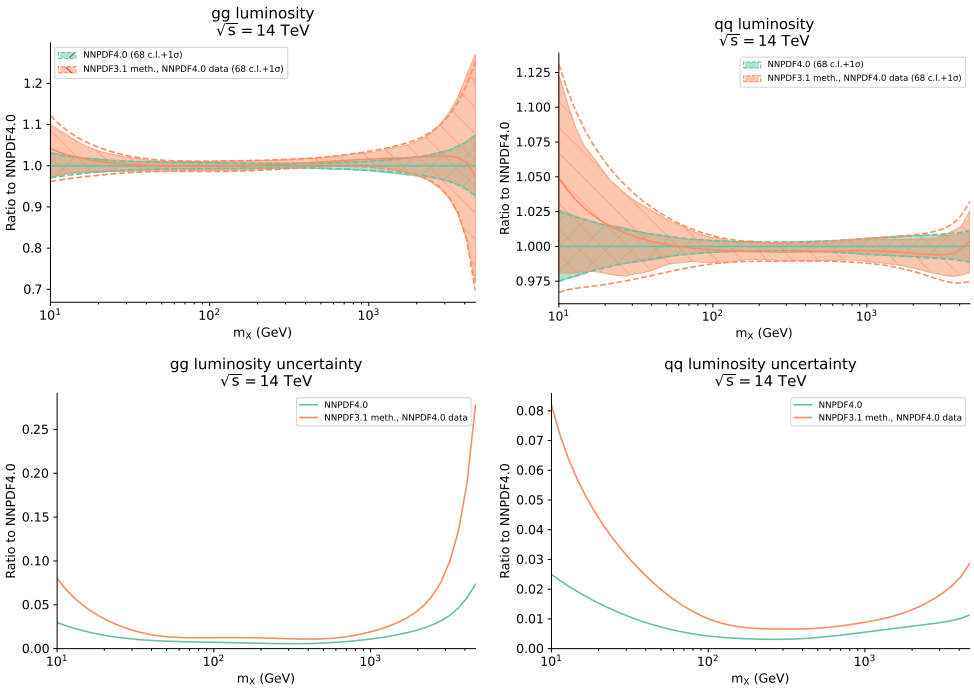

Figure 2.17: Same as Fig. 2.16, but here the baseline NNPDF4.0 PDF (green) set is compared to a PDF set determined using the NNPDF3.1 methodolgy and the NNPDF4.0 dataset (orange).

Combining this observation with the previous observation about the impact of the new data, it can be concluded that while the change in central value of any predictions is due to the change in dataset, the change in uncertainty is entirely due to the change in methodology.

2.3.3 Implications for phenomenology

To demonstrate precision that NNPDF4.0 provides, we study here a quantity that is relevant for LHC phenomenology: the PDF uncertainty on the luminosity differential in rapidity y at an energy scale of $\sqrt{s} = 14\,\text{TeV}$. This can be written as

$$\widetilde{\mathcal{L}}_{ij}\left(M_X, y, \sqrt{s}\right) = \sum_{ij}^{\text{channels}} \frac{1}{s} f_i\left(\frac{M_X e^{y}}{\sqrt{s}}, M_X\right) f_j\left(\frac{M_X e^{-y}}{\sqrt{s}}, M_X\right), \tag{2.17}$$

from which Eq. (2.16) can be obtained by integrating over rapidity:

$$\mathcal{L}_{ij}\left(M_X, \sqrt{s}\right) = \sum_{ij}^{\text{channels}} \int_{-\log\sqrt{s}/M_X}^{\log\sqrt{s}/M_X} dy \widetilde{\mathcal{L}}_{ij}\left(M_X, y, \sqrt{s}\right). \tag{2.18}$$

In Fig. 2.18 the relative PDF uncertainty on the luminosity differential in rapidity is presented as a function of both the invariant mass m_X and the rapidity y of the final state. Here the impact of the decrease in uncertainty as previously observed in Sect. 2.3.2 is clearly visible. In particular it can be seen that where NNPDF3.1 reaches uncertainties of around 1% in a limited range of phase space, for NNPDF4.0 a precision of around 1% is obtained for a much larger kinematic domain and for several parton channels.

2.3.4 The charm PDF

In the baseline NNPDF4.0 determination the charm PDF is treated on the same footing as the light quark PDFs and fitted independently. Such a treatment has various advantages [145], among with is the fact that it allows for a non-perturbative intrinsic charm component.

Fig. 2.19 compares the independently determined charm PDF at the parametrization scale of $Q_0 = 1.65\,\text{GeV}$ to the perturbatively calculated charm PDF at the same scale. The two PDFs are very different, and in particular the uncertainties of the perturbatively generated charm PDF do not seem faithful. The independently fitted charm PDF is less sensitive to the choice of the charm mass m_c and the uncertainties are significantly larger. Nevertheless, the independently determined charm PDF shows a clear valence-like bump around $x \gtrsim 0.1$ approaching 3σ significance.

Discovery of an intrinsic charm contribution to the proton however cannot be claimed based on this PDF determination since it is given in a four-flavor-number scheme with in which up, down, strange and quark are sensitive to radiative corrections and mix with each other and the gluon. To accurately determine the intrinsic component of the charm PDF, it needs to be determined in the three-flavor-number-scheme in which only the three light quarks are sensitive to radiative corrections. Note though, that the valence-like peak is found at large values of x where charm is radiatively generated only at a low rate and thus one would not expect the scheme change to have a major impact on the valence-like bump. This exercise was performed explicitly in Ref. [146], where indeed the valence-like peak remained largely unchanged upon this transformation of flavor number scheme and thereby provides evidence for intrinsic charm.

2.4 Open-source code

Along with the release of the NNPDF4.0 PDF sets, also the NNPDF code [7] has been made publicly available under the GNU General Public License v3.0, allowing for the freedom to run, study, share and modify the software . It can be found on the Github page of the NNPDF collaboration

https://github.com/NNPDF/,

along with user-friendly, and continuously updated documentation

https://docs.nnpdf.science/.

This release contains not only the NNPDF fitting framework, but also the codes needed to transform experimental data into a common format, to produce FK-tables,

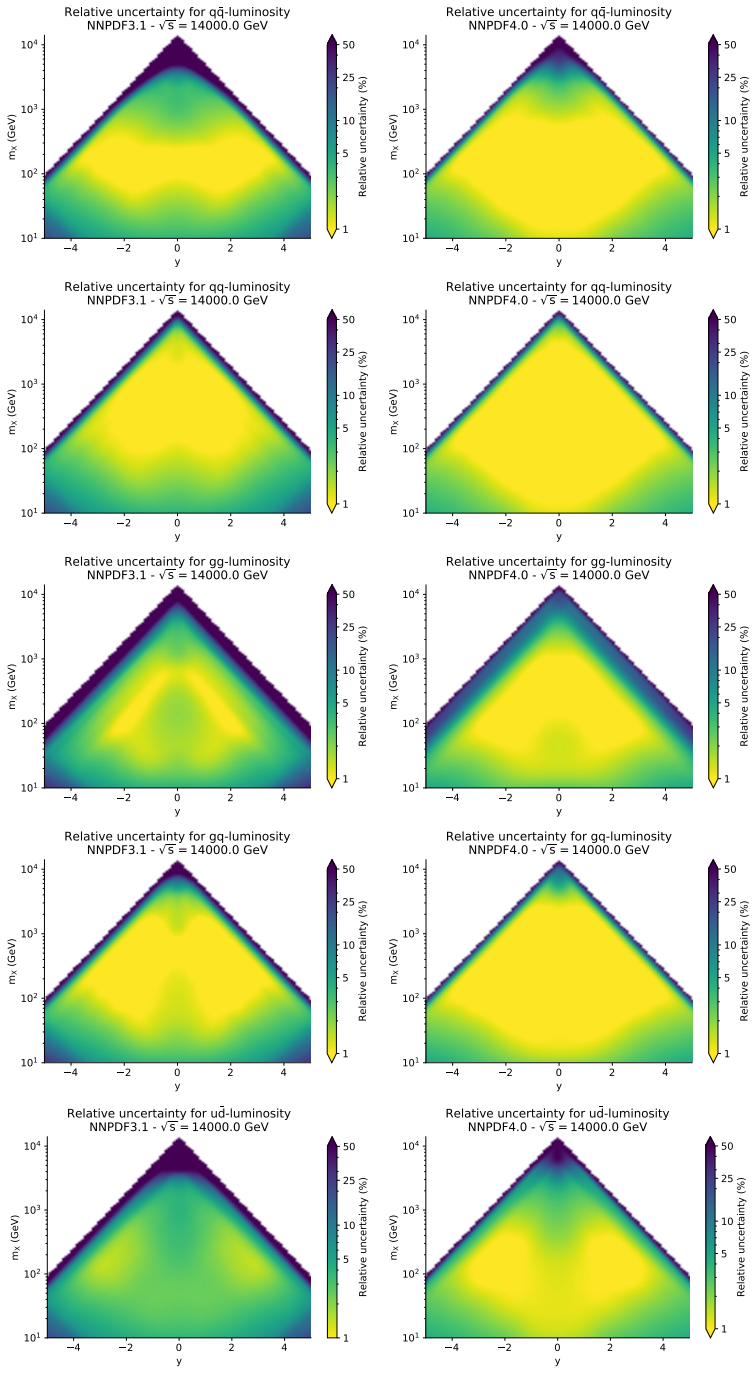

Figure 2.18: The relative PDF uncerainty on the luminosities Eq. (2.17) for NNPDF3.1 (left) and NNPDF4.0 (right) plotted as a function of the invariant mass m_X and rapidity y of the final state.

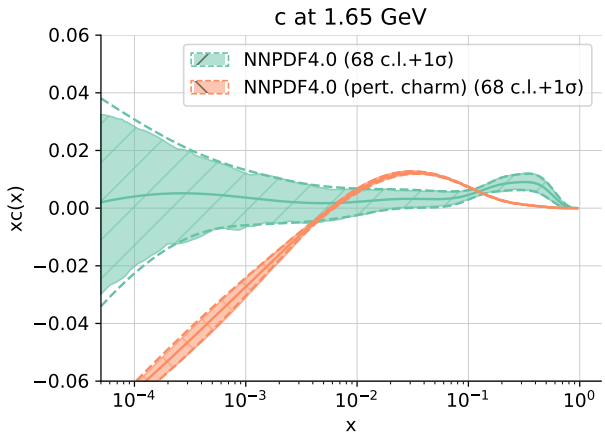

Figure 2.19: A comparison between the baseline NNPDF4.0 charm PDF (green) which has been parametrized independently, and the charm PDF determined by perturbative matching (orange). The charm mass is $m_c = 1.51\,\text{GeV}$ in both cases.

and to perform the data analysis and visualization. In addition to the codes, the public release includes the original and filtered experimental data, the fast NLO interpolation grids for the computation of hadronic observables, and whenever available the bin-by-bin NNLO QCD and NLO electroweak K-factors.

For previous releases of NNPDF only the PDF sets produced with the framework were made publicly available as through the LHAPDF framework [78] as LHAPDF interpolation grids, while the code itself remained private. As a result, the code itself could not be scrutinized and results could not be reproduced by external parties. It also meant that the only method of obtaining variations of existing PDFs – for example a PDF set determined using a reduced dataset – was by requesting them from the NNPDF authors. In practice this was a limitation to benchmarking studies such as those performed by the PDF4LHC working group [147]. In such a benchmarking study the differences between PDF determinations from the different fitting collaborations are attempted to be understood, but often differences are the result of a complex combination of various factors. Studies such as these are therefore aided by the code being open access.

The NNPDF code consists of the following main packages:

- The `buildmaster` code for handling experimental data is a `C++` code that can be used to take experimental as provided by experimental collaboration – for example through the HEPData interface [148] – and generate files containing information about the data and the treatment of uncertainties in a format that can conveniently be used within the rest of the NNPDF framework.

- The `APFELcomb` code for the generation of FK-tables. It does this by taking matrix elements, such as those obtained from `APPLgrid`, `FastNLO` for hadronic observables or `APFEL` for DIS observables, and combining them with DGLAP evolution kernels from `APFEL`.

- The `validphys` framework for the analysis and visualization of data related to PDF determinations. The `validphys` framework is build on top of the `reportengine` framework [149], which is a data science framework supporting declarative inputs in YAML format and checks constraints at initialization time while building a computational graph. Most of the plots in this thesis have been produced entirely with `validphys`, and in many other cases the `validphys` API has been used to obtain data about the PDFs.

- The `n3fit` fitting framework that takes as input the experimental data and FK-tables, and produces the PDF replicas as has been discussed in detail in Sect. 2.1.3.

The public availability of the NNPDF code opens up a number of possibilities for users to perform their own analysis of PDFs using the NNPDF framework or extensions thereof. Examples of possibly interesting applications for users that the NNPDF code allows are the assessment of the impact of a specific dataset or group of datasets by producing fits based on a reduced dataset, or by implementing a dataset that has not yet been included in the NNPDF framework. The open-source code also allows users to perform fits with different settings of the theory calculations. This enables for example the study of α_s dependence by performing fits to theories with different values of α_s [150], the estimation of missing higher order uncertainties (MHOU) by varying the factorization and renormalization scales [151,152], study the sensitivity to heavy quark masses by varying those. Finally, since the open-source code is public, users can extend its functionality. For example, one can extend the framework to allow for the simultaneous determination of PDFs and Wilson coefficients in the Standard Model Effective Field Theory framework [153], or even apply it to the determination of other non-perturbative QCD quantities such as nuclear PDFs [154], fragmentation functions [155], or polarized PDFs [156].

It should be noted that some of the functionality described above is already available in the `xFitter` framework [157,158]. However the NNPDF framework provides some complementary functionalities, specifically by offering a PDF parametrization based on state-of-the-art machine learning tools, a more extensive experimental dataset, and a great number of tools for the statistical analysis and visualization of data.

Chapter 3

Advanced machine learning tools

In the previous chapter we introduced the NNPDF4.0 PDF determination. We observed how an increase in data quality, in particular as a result of the large number of new processes, resulted in more accurate PDFs with no significant impact on their precision. We then observed how the improved methodology instead resulted in a significant reduction of the PDF uncertainties. Thus, while both NNPDF3.1 and NNPDF4.0 provide an accurate determination of the PDFs, NNPDF4.0 is an improvement over NNPDF3.1. Similarly, in this chapter we discuss two main directions of improvements to the methodology that may be utilized for future releases.

In Sect. 3.1 we present a method that allows us to replace the $(x, \log x)$ splitting in the first layer of the neural network with a data-based scaling, and we will see how this further enables us to remove the preprocessing prefactor, and by extension the corresponding iterative procedure for their determination, present in the parametrization Eq. (2.3). In Sect. 3.2 we propose a way to further automate the hyperoptimization routine thereby reducing the need for human intervention. In particular, we propose a method to optimize the selection of folds for the k-folds cross-validation, and a measure to quantify the degree of overfitting that occurred during the fitting of the PDFs. In this way, we address two main directions for improvement of the hyperoptimization routine discussed in Sect. 2.1.4.

3.1 Improved PDF parametrization

All of the most used PDF sets are parametrized at some input scale Q_0 by a function of the form in Eq. (1.67)

$$x f_i\left(x, Q_0\right) = A_i x^{(1-\alpha_i)}(1-x)^{\beta_i}\mathcal{P}_i(x), \tag{3.1}$$

where the indices i correspond to the type of parton, and $\mathcal{P}_i$ is a functional form that is different between PDF fitting groups. This is a generalization of Eq. (2.3) for the parametrization employed by the NNPDF collaboration in which $\mathcal{P}_i$ is represented by a single neural network. As also mentioned before, for other modern PDF sets such as MSHT20 [63], CT18 [64], and ABMP16 [65], $\mathcal{P}_i$ represents a polynomial in functions of x, such as $\sqrt{x}$.

The PDFs are kinematically constrained at $x = 1$ per Eq. (1.66) which is enforced through the $(1-x)^{\beta_i}$ component in Eq. (1.67) by all PDF fitting collaborations . The motivation for this term stems from the constituent counting rules [159]. In the fitting methodologies this component not only ensures that the condition of Eq. (1.66) is satisfied, but it partially controls the large-x extrapolation region where data is unavailable. The small-x behavior instead is controlled by the prefactor $x^{(1-\alpha_i)}$. The introduction of this factor was inspired by Regge theory [160]. While enforcing this behavior implies a methodological bias [58], studies on the extrapolation behavior of PDF determinations confirm both Regge theory and counting rules for the valence distributions [161]. Regardless, the effect of the exponents α_i and β_i as a source of bias [44] is mitigated in the NNPDF methodology by independently and randomly sampling them from a uniform distribution per replica, and freezing their values during the fit. This is to be contrasted with the approach of other collaborations where the α_i and β_i are treated as parameters that are to be optimized during the fit. The boundaries defining the distributions from which the exponents are sampled in the NNPDF approach are determined through the iterative procedure described in Sect. 2.1.2.

Despite its generalized use in PDF determination, the fixed functional form is a part of the methodology that leaves room for improvement. Namely, if we are able to remove the preprocessing entirely, this provides two main benefits. First, and perhaps most obviously, if we are able to remove this preprocessing from the NNPDF methodology this also reduces the need for the iterative procedure to determine the ranges of the exponents. Second, the required sampling of the exponents corresponds to an additional source of data-independent replica-by-replica fluctuations. This may affect the hyperoptimization procedure of Sect. 2.1.4, namely a relatively poor methodology may perform well on the hyperoptimization metric if the randomly sampled preprocessing exponents result in a better agreement of the fit to the data than the expected performance of that methodology. Inversely, a relatively good methodology may perform poorer during hyperoptimization as a result of the same fluctuations in the preprocessing exponents. Finally, it is worth mentioning that the PDFs are only based on data in the domain $10^{-5} \lesssim x \lesssim 0.75$ while PDF grids are delivered in the domain $10^{-9} \leq x \leq 1$ and as a result the preprocessing impacts the extrapolation behavior of the PDFs. While there are theoretical arguments [162] that suggest the power-like behavior of PDFs in the limits of $x \to 1$ and $x \to 0$ as described by the preprocessing function, it is not clear that this is also true for the finite region in which the PDF is provided through the LHAPDF6 interface. Though, even if we assume that this is indeed the case, it is still not clear at which scale Q^2 the exponential scaling should hold given that it is not preserved under the evolution equations discussed in Sect. 1.3.

Another aspect of the PDF parametrization that is the result of an explicit human choice is the $(x, \log x)$ split in the first layer of the neural network as shown in Fig. 2.2. The choice for this splitting of the input results from the observation that typically PDFs show logarithmic behavior at small-x ($x \lesssim 0.01$) and linear behavior at large-x ($x \gtrsim 0.01$) [68], and together with the prefactor it ensures convergence of the optimization algorithm in the small-x region.

In what follows we will show, in Sect. 3.1.1, how the scaling of the x-grids that are given as input to the neural network can be automated to replace the $(x, \log x)$ split

in a way that allows for a more flexible parametrization. Then, in Sect. 3.1.2, we show how this scaling allows to remove the preprocessing from the parametrization entirely.

3.1.1 A data-based scaling of the input x-grids

Often, in machine learning problems, the input data can be unbalanced or span several orders of magnitude. Such is the case of PDF fitting, where the input is concentrated at small-x.

This can be a problem because, as we will explicitly show below, having input features of different magnitudes introduces an artificial impact on the importance of each feature within the network. This problem is exacerbated for gradient descent based algorithms where the issue propagates to the learning rate of the weights of the network. Thus, even if the algorithm is still able to find the global minimum, the rate of convergence is not equal for all features. In the case we are interested in (the NNPDF methodology with an early stopping algorithm) this can lead to locally overfitted or underfitted results in different regions of the kinematic domain. Ideally the fitting methodology should result in a uniform rate of convergence across all input scales.

In short, the problem is that while the data spans multiple orders of magnitude, the fitting methodology requires the inputs to be of the same length scale. Below we discuss the impact of the input scaling on the PDFs, and provide a methodology that takes an arbitrary input grid and scales it such that the optimizer always has a good resolution across the entire input grid.

At this point one may note that the input x-grids of the neural network are the grids defined in the FK-tables as shown in Eq. (2.7) and Eq. (2.8), which may differ from the x-values of the corresponding experimental datasets. While this is true, from the perspective of the fitting methodology, the grid choice is arbitrary and thus the problem remains.

In NNPDF fits, the input variable is mapped to $(x, \log x)$ in the first layer of the neural network which facilitates the methodology in learning features of the PDF that scale either linearly or logarithmically in x. As mentioned before, these scales are carefully chosen in accordance with the typical scaling of PDFs which is logarithmic in the small-x region while it is linear in the large-x region. In the structure function fit presented in Ref. [68] it was further noted that the choice of input scales could affect the rate of convergence but not the final result. However, the $(x, \log x)$ split can have an effect on the shape of the PDFs when determined using the modern framework. This is seen in Fig. 3.1, where we compare the gluon PDF of the NNPDF4.0 fit to a PDF generated using the same data, theory and methodological settings, but with the $(x, \log x)$ input scaling replaced with only an (x) input. While the NNPDF4.0 methodology was sensitive to the small-x region (where the logarithmic behavior is expected) when we remove $(\log x)$ from the input we can observe a hint of saturation in said region. Despite the fact that the (x) and $(\log x)$ variables contain the same information the split has a noticeable effect on the fit. We will now present an alternative data-based scaling and show that this scaling finds agreement with the results found using the $(x, \log x)$ scaling, thereby providing evidence that the $(x, \log x)$ does not introduce an inefficiency in the methodology.

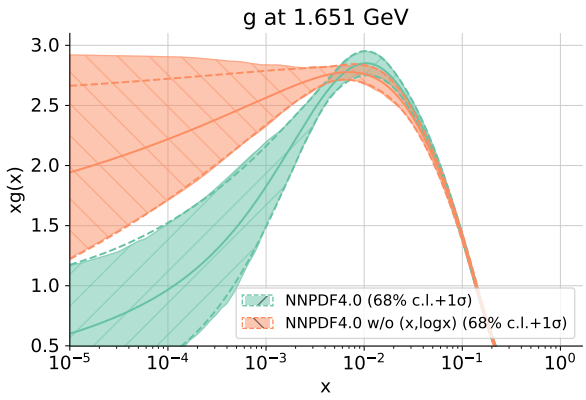

Figure 3.1: Comparison between the gluon PDF generated with the standard NNPDF4.0 methodology (green) and our modification in which we have removed the splitting layer of x to $(x, \log x)$ (orange). While we observe good compatibility between both PDFs in the large-x region, as we enter in the small-x region our modified PDF saturates. This is evident also in the χ^2 of the modified fit which was blocked at $\chi^2 = 1.20$ while NNPDF4.0 is able to get it to $\chi^2 = 1.16$.

In order for the optimization algorithm to be able to easily learn features across many orders of magnitude we can perform a feature scaling of the training input x such that the distances between all points are of the same order of magnitude. In particular, we can consider mapping the combined training input x-grid from the FK-tables of all datasets as discussed in Sect. 2.1 to an empirical cumulative distribution function (eCDF) of itself. The eCDF is defined as a step function that starts at 0 and increases by $1/N_x$ at each point of the input x-grids, with N_x the total number of nodes in the x-grids. If the x-grids of n FK-tables share a common point in x, the step-size corresponding to this point is instead n/N_x. This results in a function whose value at any x corresponds to the fraction of points in the x-grids that are less than or equal to x. In other words, while the x values present in the FK table are not uniformly distributed on the domain $0 \leq x \leq 1$, applying the eCDF makes it that they are. A density plot of the distribution of input points without scaling, logarithmically scaled, and after applying the eCDF is shown in Fig. 3.2. This figure also clearly shows that both inputs to the neural network as used in NNPDF4.0 [6] have a high density of points on the same scale.

Applying the eCDF results in a distribution on the domain $0 \leq x \leq 1$. However, for the results presented in this paper the eCDF transformation is followed by a linear scaling, resulting in a total transformation of the input $\tilde{x} = 2 \cdot \mathrm{eCDF(x)} - 1$, meaning that the input values to the neural network are in the range $-1 \leq \tilde{x} \leq 1$. This is done to ensure that the input is symmetric around 0 which results in improved convergence for many of the commonly used activation functions in neural networks.

Since using the eCDF means that we apply a discrete scaling only for values present in the input x-grids, we need to also add both an interpolation and an extrapolation function to extract PDF values at values of the momentum fraction that do not coincide with the input x-grids. Here it is important to note that the PDFs are made publicly

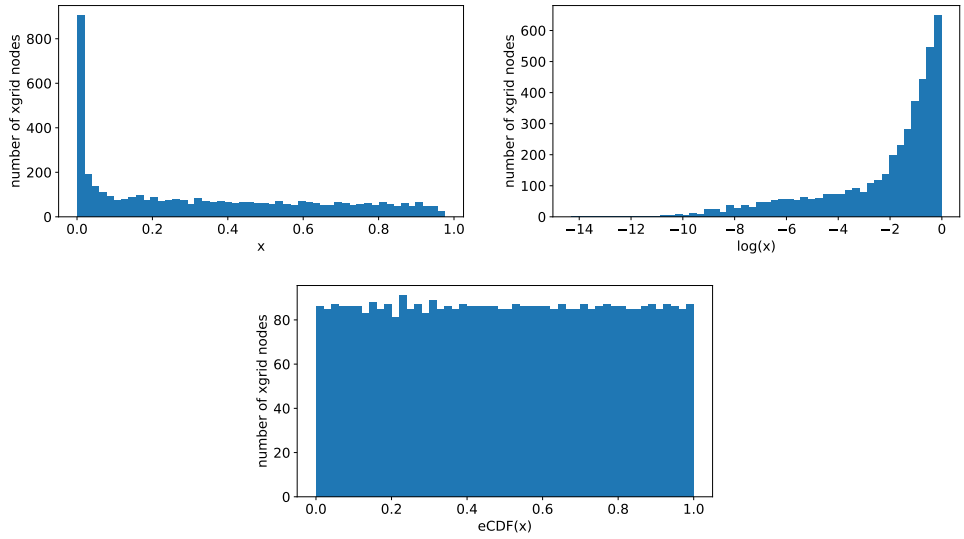

Figure 3.2: Histograms showing the distribution of the unscaled x points in the FK-table x-grids (top-left), as well as the distribution of the input points after scaling with $\log x$ (top-right) and eCDF (bottom).

available through the LHAPDF interface, and that they are correspondingly stored in the LHAPDF grid format [78]. Because LHAPDF grids are provided on the domain $10^{-9} \leq x \leq 1$, the problem of extrapolation can be turned into an interpolation problem by including the points $x = 10^{-9}$ and $x = 1$ in the input x-grid before determining the eCDF, and defining a methodology for interpolation.

The simplest option for an interpolation function is a "nearest neighbor" mapping, whereby we map any input on the continuous domain $0 \leq x \leq 1$ to the nearest node in the x-grids of the FK-table. We can nevertheless improve this simple mapping by using instead a continuous function. A requirement of any such interpolation function is that it needs to be monotonically increasing. However, if we determine the interpolation between each two points of the FK-table x-grids the optimization algorithm will be agnostic to the existence of this interpolation function as it is never probed. Ideally, in particular for the evaluation of validation data of which the corresponding FK-tables were not included when defining the eCDF scaling, we want the optimizer to probe the interpolation functions such that it is able to learn its properties and as a result provide a more accurate prediction in the interpolation region as well. As such, the interpolation functions are not defined between each neighboring pair of values in the input x-grid, but rather we select N_{int} evenly distributed points (after the eCDF transformation) between which to define interpolation functions. Here N_{int} is a new hyperparameter, though not necessarily one that needs to be free during hyperoptimization of the methodology. To obtain a monotonic interpolation function, we propose determining the interpolation functions using cubic Hermite splines [163].

By scaling the input in this way, we remove any restrictions on the PDF resulting from the input features while simultaneously simplifying the model architecture by

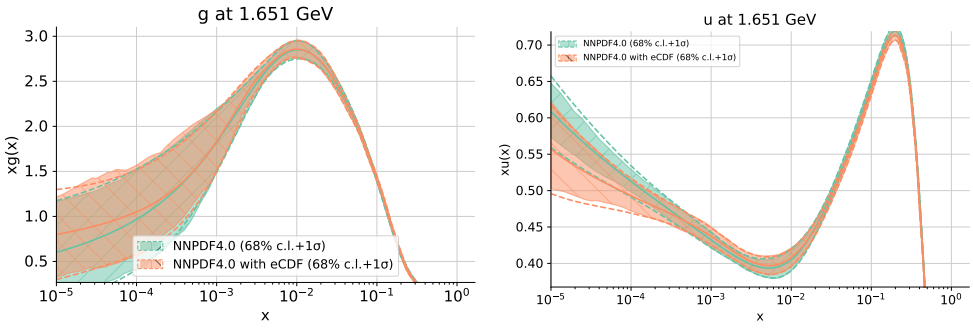

Figure 3.3: Comparison between the gluon and up PDFs determined using the NNPDF4.0 methodology (green) and a PDF determined using input scaling based on the eCDF (orange) with all other parameters the same.

getting rid of the mixing of two different orders of magnitude in the first layer. In Fig. 3.3 we compare the gluon PDF generated using the NNPDF4.0 methodology, to a PDF generated using the same data and theory settings, but with the $(x, \log x)$ input scaling replaced with the eCDF input scaling as described above. This comparison of the gluon PDF is representative for all flavors, and shows that the PDFs produced with this new scaling are in agreement with those found using the $(x, \log x)$ input. If the PDFs had not been in agreement that would have suggested that the PDFs have a component that scales neither linearly nor logarithmically, and was therefore missed when enforcing the $(x, \log x)$ scaling.

3.1.2 Removing the prefactor

In the previous section we discussed a new way of treating the input for the PDF fitting by rescaling the input in a systematic way that depends only on the fitted data itself. This is a purely data-driven approach and thus free of sources of bias due to the choice of functional form. As explained, the data-based scaling of the input grid in x will also allow us to remove the prefactor entirely.

In what follows we will discuss the consequence of removing the prefactor. Specifically, by "removing the prefactor", we understand a treatment which is equivalent to setting $\alpha_i = 1$ and $\beta_i = 0$ in Eq. (2.3), while enforcing the condition of Eq. (1.66). As a result the PDF model is simply written as

$$x f_i(x, Q_0) = A_i \left[\mathrm{NN}_i(x) - \mathrm{NN}_i(1)\right]. \tag{3.2}$$

A similar model, without the model-agnostic input scaling, has previously been applied to the study of fragmentation functions [95]. We will focus on the effects of the change in the small-x and large-x extrapolation regions where the lack of data makes the fit particularly prone to methodological biases.

Earlier we mentioned that the motivation to include the prefactor in NNPDF is to improve convergence during optimization and that its effect as a source of bias in the extrapolation region was mitigated by randomly sampling the exponents α_i and β_i from a uniform distribution and keeping their values fixed as opposed to

allowing the optimizer to determine their values. However, removing the preprocessing entirely not only has the advantage of avoiding bias, but also removes the replica-by-replica fluctuations introduced by the different values of the exponents used for each replica. These fluctuations are an inefficiency of the fitting methodology, which may in particular affect the hyperoptimization routine discussed in Sect. 2.1.4. The reason for this is that the performance of each hyperparameter configuration is tested by performing $k = 4$ fits, hence the sample size (and thus potential impact of random fluctuations) is non-negligible.

The improvements in efficiency achieved with the NNPDF4.0 with respect to the NNPDF3.1 methodology discussed in Sect. 2.1 allow us to remove the prefactor without a significant change in computational costs. Therefore any possible benefit of the prefactor in terms of convergence no longer outweighs its disadvantages. As an example of where fluctuations between replicas as a result of the randomized exponents of the prefactor can limit the development of the methodology, one can consider the hyperoptimization procedure previously discussed in Sect. 2.1.4. Namely, in the current scenario an otherwise good hyperparameter setup with poor exponents in the prefactor can return a worse figure of merit during hyperparameter optimization than a relatively poorer hyperparameter setup with very suitable exponents. As a result many more hyperparameter combinations need to be tested to overcome the statistical noise. Removing the replica-by-replica random sampling of the exponents removes this effect from hyperoptimization.

The uncertainties of the fit in the extrapolation region are closely related to the ranges the prefactor exponents are sampled from. Removing them from the parametrization also removes the random sampling. Therefore, we will next validate the obtained small-x and large-x uncertainties.

For brevity and clarity, we will from now on refer to the proposed methodology without the prefactor and with the eCDF input scaling as the "feature scaling" methodology.

3.1.3 Validation of the updated parametrization

After any significant change to the fitting methodology, it is important to re-evaluate the choice of the hyperparameters of the model. The model parameters obtained through the hyperoptimization procedure applied to the feature scaling methodology are given in Tab. 3.1. Note that the selected activation function does not saturate asymptotically for large or small values of x, thus preventing saturation outside the data region. The choice of activation function was however not fixed during the selection of hyperparameters, this activation function has been selected by the hyperoptimization algorithm among a selection of both saturating and non-saturating activation functions.

Having identified the best settings for the hyperparameters, we can analyze the effect that changing the parametrization has on the PDFs and the predictions made with them. The χ^2 values obtained with the updated methodology are shown in Fig. 3.4 where they are compared to those of NNPDF4.0. From this it is clear that the feature scaling methodology is able to find agreement to the data that is as good as NNPDF4.0.

Architecture	1-59-49-48-42-8
Activation function	$\|x\|\tanh(x)$
Initializer	`glorot_normal`
Optimizer	`Nadam`
Clipnorm	1.5×10^{-5}
Learning rate	4.3×10^{-3}
Maximum # epochs	19×10^{3}
Stopping patience	24% of max epochs
Initial positivity $\Lambda^{(pos)}$	34
Initial integrability $\Lambda^{(int)}$	10
N_{int}	40

Table 3.1: The hyperparameter configuration used to perform the feature scaling fits. The configuration has been selected using the hyperoptimization routine of Sect. 2.1.4.

In what follows we will study the implications of the methodology in more detail, in many cases by comparing it to a PDF based on the same experimental dataset and theory setting, but produced using the NNPDF4.0 methodology. Specifically, we will perform various tests to validate the PDFs both in the extrapolation regions, as well as in the data region. These tests comprise the validation of the NNPDF4.0 methodology, and we will show that the performance of feature scaling is very similar to that of NNPDF4.0.

Validation of the small-x extrapolation region

To begin with, we need the PDFs to accurately describe the kinematic domain from which the methodology has not seen data during training. If we are able to determine the χ^2 for this unseen data, that would provide some insight into the generalization of our methodology in the extrapolation region.

By definition, testing the accuracy in a region where there is no data to test against is impossible. Given that waiting for a future collider to become operational could take decades, the next best thing we can do is to perform a fit to a "historic" dataset representing the knowledge available at an earlier point in time. To this end we utilize the "future test" technique introduced in Ref. [164], and used to validate the extrapolation region of the NNPDF4.0 PDFs. For consistency we keep the same datasets as presented in the original future test paper (pre-HERA and pre-LHC). In short, the test goes as follows: if the prediction from our methodology is able to accommodate (within uncertainties) currently available data that was not included in the fit, then the test is successful and we consider the generated uncertainties to be faithful.

Since the aim of doing a future test is to determine the ability of a methodology for PDF determination to provide a generalized fit, we need to take into account not only the uncertainty of the experimental data but also the uncertainty of the PDF itself. This is done by redefining the covariance matrix in the chi-squared-distribution Eq. (2.9) as

$$(\mathrm{cov}_{\mathrm{tot}})_{ij} = (\mathrm{cov}_{\mathrm{exp}})_{ij} + (\mathrm{cov}_{\mathrm{pdf}})_{ij}, \tag{3.3}$$

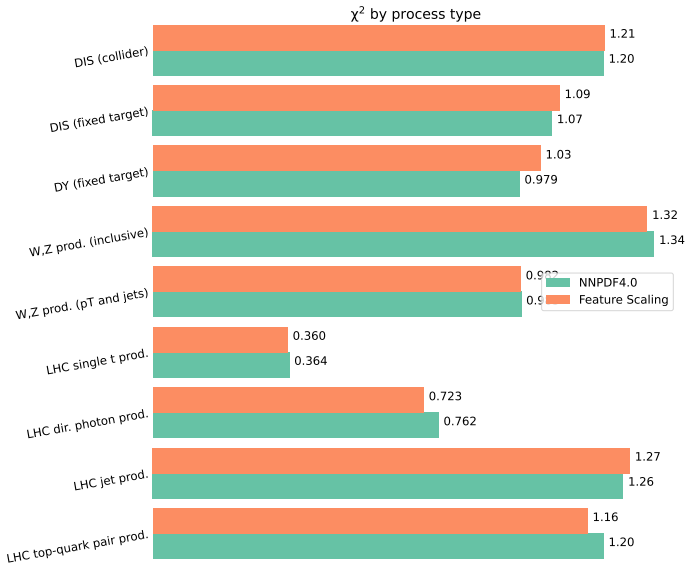

Figure 3.4: A comparison of the χ^2 per process type between NNPDF4.0 (green) and feature scaling (orange), the total χ^2 of feature scaling is 1.17 while that of NNPDF is 1.16.

where cov_{exp} corresponds to the covariance matrix defined in Eq. (2.10) without the t_0 prescription applied, while cov_{pdf} corresponds to the covariance matrix of the observables calculated from PDF predictions:

$$(\text{cov}_{\text{pdf}})_{ij} = \frac{1}{N_{\text{rep}}} \sum_{r=1}^{N_{\text{rep}}} P_i^r P_j^r - \frac{1}{N_{\text{rep}}} \sum_{r=1}^{N_{\text{rep}}} P_i^r \frac{1}{N_{\text{rep}}} \sum_{k=1}^{N_{\text{rep}}} P_j^k, \tag{3.4}$$

where P_i^r is the prediction of the i-th datapoint using the r-th PDF replica.

As can be seen in Fig. 3.5, where we compare the gluon and upquark PDFs of the NNPDF4.0 fit, to a PDF generated using the feature scaling methodology, the plots show good agreement between the two PDFs. While only the two partons are shown, this is representative of all flavors. The prediction of the feature scaling methodology in the extrapolation region is validated by performing a future test of the feature scaling methodology. The results of this future test results shown in Tab. 3.2. Each column corresponds to a fit perform using all previous datasets (for instance, the pre-LHC fit includes all the data in pre-HERA as well). Instead, each row corresponds to the partial dataset used to compute the χ^2. We make a distinction between χ^2 inside parentheses with the experimental covariance matrix, and the χ^2 without parentheses corresponding to a covariance matrix as defined in Eq. (3.3). Before seeing these results one may wonder whether, because all datasets are sensitive to the same large-x region, the datasets are consistent and thus the test is trivial. The answer to this becomes clear by looking at the χ^2 values inside the parentheses which indicate that

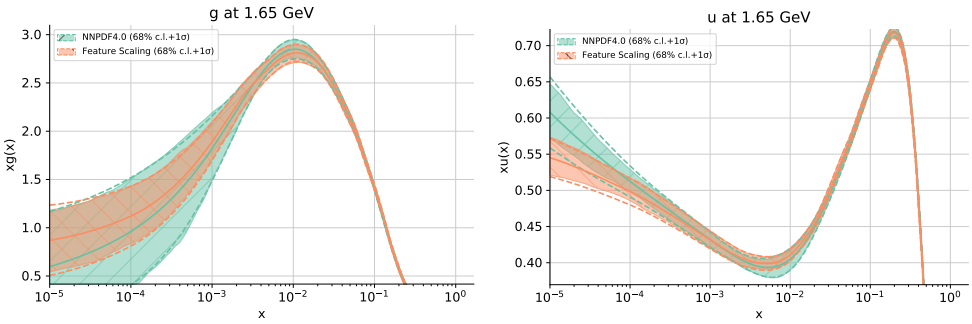

Figure 3.5: Comparison of the gluon and upquark PDFs between a fit performed with the NNPDF4.0 methodology (green), and one with the feature scaling methodology (orange).

when the PDF uncertainties are not considered the fit quality is very poor for unseen data.

We can analyze the result starting on the third row corresponding to the NNPDF4.0 dataset. For the fit that included the entire dataset (third column) it makes virtually no difference whether or not the PDF uncertainties are taken into account. This is quite different for the pre-HERA fit (first column): even though the central PDF is off ($\chi^2 = 7.23$), once its uncertainties are considered, the quality of the fit is comparable to that of NNPDF4.0 with with a χ^2 of 1.29 compared to 1.21. In the second row instead the pre-LHC dataset is considered. Both the NNPDF4.0 and the pre-LHC fit, where the dataset is included, produce a trivially good χ^2 for their fitted data. When we compute the prediction using the pre-HERA fit instead the number is much worse. Once again, upon considering the PDF uncertainties, the number is of order one, though still significantly larger than the corresponding values in the fits with pre-LHC or NNPDF4.0 data. This suggests that qualitatively good agreement is obtained but stability upon changes to the dataset can still be improved.

It should be noted that in all cases the methodology used has been hyperoptimized for the full NNPDF4.0 dataset. While one may argue that the fits to the historic datasets require re-hyperoptimization and to redo the iteration required for the t_0 procedure discussed in Sect. 2.1.3, the main purpose of the exercise performed here is to compare the future test results obtained with the feature scaling methodology to those obtained with the NNPDF4.0 methodology in Ref. [6].

If we compare these results as presented in Tab. 3.2 for the feature scaling methodology, to the results for the NNPDF4.0 methodology shown in Tab. 3.3, we observe much the same properties. Indeed, even in cases in which the out-of-sample χ^2 differs greatly between both methodologies, the results are compatible once the PDF uncertainties are considered. This confirms that, when PDF uncertainties are considered, the agreement to out-of-sample data is of a similar level as that of fitted data where the PDF uncertainty is not considered.

We must note however a deterioration of the results in Tab. 3.2 with respect to those of Tab. 3.3 which points to a greater dependence on the considered dataset with the feature scaling methodology. In part this may be a consequence of the fact that

Dataset	$N_{\rm dat}$	pre-HERA fit	pre-LHC fit	NNPDF4.0 fit
pre-HERA	2076	0.87 (0.92)	0.91 (1.03)	0.98 (1.08)
pre-LHC	1273	**1.35 (5.61)**	1.17 (1.27)	1.18 (1.20)
NNPDF4.0	1269	**1.29 (7.23)**	**1.22 (4.72)**	1.21 (1.29)

Table 3.2: χ^2 values per datapoint as obtained during a future test of the feature scaling methodology. The columns correspond to fits based on a given dataset, while the rows correspond to the datasets for which the χ^2 values are shown. While for the fit the dataset are inclusive (i.e., the NNPDF4.0 fit includes also the pre-LHC and pre-HERA datasets) the χ^2 is computed in an exclusive manner (i.e., the χ^2 as calculated for the NNPDF4.0 dataset only uses "post-LHC" data). The values in bold represent the performance on datasets that were not part of the training. The values inside parentheses correspond to a χ^2 defined with σ as defined in Eq. (3.3), while those without parenthesis are defined with only the experimental covariance matrix.

the datasets used in the NNPDF4.0 determination have been carefully selected by analyzing their impact on PDF fits using the NNPDF4.0 methodology (see Sect. 4 of Ref. [6]), while those same datasets have here been used to validate the feature scaling methodology instead of performing again the appraisal of the datasets for the feature scaling methodology. Nevertheless, in Sect. 2.2.1 we analyzed the impact of seven datasets which all had particularly large values for measures used to determine whether or not to include the corresponding dataset in the PDF fit, and the results obtained there suggest that no significant impact is to be expected upon the removal of these datasets from the fit. However, a more likely explanation for this difference is related to the fact that the preprocessing ranges of the NNPDF4.0 methodology are determined using the global NNPDF4.0 dataset. Some information on the full NNPDF4.0 dataset may thus be encoded in the preprocessing exponents and therefore be present in the fits performed using the NNPDF4.0 methodology, even if the data used during training was the pre-LHC or pre-HERA subset of the global dataset. Thus, while the preprocessing ranges of the fits with the NNPDF4.0 methodology have been determined using the NNDPF4.0 dataset, one of the main improvements provided by the feature scaling methodology is that it is able to accommodate directly different datasets without the need to determine the range of preprocessing exponents. This may also explain why the out-of-sample χ^2 of feature scaling is actually better than that achieved by NNPDF4.0. One may explicitly check this hypothesis by determining the ranges of the preprocessing exponents one would obtain through the iterative procedure described in Sect. 2.1.2 if only the pre-HERA or pre-LHC datasets were available, and repeating the future test using the resulting methodologies. This however is left for future work. Here it suffices to note that, when considering the PDF uncertainties, the χ^2 of the PDF predictions and the "future datasets" excluded from the corresponding fit is close to one.

Finally, having removed the preprocessing, one may consider further constraining the small-x region using different methods. One such possibility is proposed in Ref. [165] where a Gaussian Process is used to sample pseudodata in the extrapolation region by explicitly learning the correlation of the DIS data in the small-x region.

Dataset	N_{dat}	pre-HERA fit	pre-LHC fit	NNPDF4.0 fit
pre-HERA	2076	0.87 (0.91)	0.94 (1.01)	1.01 (1.06)
pre-LHC	1273	**1.22 (26.1)**	1.18 (1.21)	1.17 (1.20)
NNPDF4.0	1269	**1.28 (22.6)**	**1.28 (2.15)**	1.23 (1.29)

Table 3.3: Same as Tab. 3.2 for the NNPDF4.0 methodology.

Evaluation of large-x extrapolation

Upon removing the prefactor, we not only affect the small-x extrapolation region of the PDFs, but also the large-x extrapolation region. It is difficult to apply the idea of the future test to also validate the faithfulness of the predictions in the large-x region due to the limitations of the datasets that do not contain any large-x datapoints (irrespective of how we define large-x precisely). For example, removing all datasets which contain a point in $x \gtrsim 0.3$ leaves a set of datasets which do not provide sufficient constraints on the PDF to perform the future test. Nevertheless, here we will assess the large-x extrapolation behavior of the PDF produced with feature scaling.

To do so, let us visually inspect the PDFs themselves in this region, and see how the PDFs based on the NNPDF4.0 methodology compare to those that have been produced with feature scaling. A comparison of the gluon and strange PDF in the domain $0.6 < x < 1$ is shown in Fig. 3.6. Note here that there is no data available for $x > 0.75$, meaning that what is shown is mostly extrapolation region, and these representative examples show a good agreement between the NNPDF4.0 PDF and the feature scaling counterpart. We further want to point out that due to the lack of data in this region different parametrization choices can lead to significantly different results. In particular this can be seen by comparing the NNPDF4.0 PDFs to those produced by MSHT or CT, where the observed difference may be related to the more flexible PDF parametrization used by NNPDF4.0 [166].

As a more rigorous check of the large-x extrapolation region one could create pseudodata based on predictions corresponding to PDFs that have a different (exponential) behavior in the extrapolation region, e.g. a change of the β_i exponent outside the data region. One can then perform a future test to this pseudodata, to quantify how well the PDFs generalize in the extrapolation region. The development of such a test, however, is left for future work.

Validation of the data region

Where previously we performed a future test to validate the faithfulness of the PDFs in the extrapolation region where the PDFs are not constrained by data. Here, instead, we will validate the faithfulness of the PDFs in the data region by performing a closure test as first introduced in Ref. [75] and extended in the NNPDF4.0 paper and Ref. [167]. Below we repeat the closure test as performed in Sect. 6.1 of the NNPDF4.0 paper, but this time for the feature scaling methodology. Unless stated otherwise, the same settings are used.

When fitting experimental data we are subject to complexities in the data such as inconsistencies between datasets or limitations of the theoretical calculations. These complexities make it more difficult to assess the performance of a fitting methodology

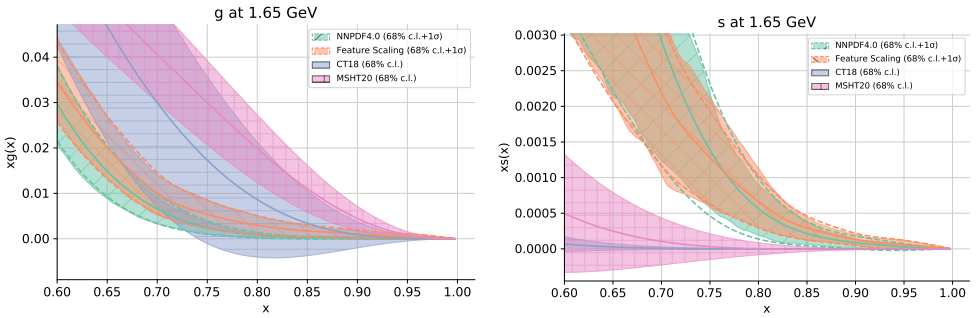

Figure 3.6: Comparison of the large-x extrapolation regions of the gluon (top) and the strange (bottom) PDFs between NNPDF4.0 (green), feature scaling (orange), CT18 [64] (blue), and MSHT20 [63] (pink).

by analyzing the result of a fit to experimental data. This realization is what led to the idea of a closure test, where, instead of fitting to experimental data, a fit to pseudodata is performed. This pseudodata is generated by taking a fitted PDF as input, and from that calculating the observables corresponding to those in the experimental datasets, thereby creating a dataset with an associated, and known, underlying PDF. This allows us to test whether our methodology is able to faithfully reproduce the underlying PDF. To test whether our methodology was successful, a number of statistical estimators are considered that we will discuss next. For a detailed motivation of these estimators we refer the reader to section 6 of Ref. [6]. As underlying truth we use one non-central replica from a feature scaling fit.

A first statistical estimator to consider is the Δ_{χ^2}

$$\Delta_{\chi^2} = \chi^2[f^{(cv)}] - \chi^2[f^{(ul)}], \tag{3.5}$$

where $\chi^2[f^{(cv)}]$ is the loss evaluated for the expectation value of the fitted model predictions, while $\chi^2[f^{(ul)}]$ is the loss evaluated for the predictions of the PDF used as underlying law. The latter loss does not vanish, because the pseudodata includes a Gaussian random noise on top of the central value predictions made using the underlying law. As such, Δ_{χ^2} can be understood as an indicator for overfitting or underfitting: if $\Delta_{\chi^2} > 0$, that indicates underfitting, while $\Delta_{\chi^2} < 0$ indicates overfitting. For the feature scaling methodology, the average Δ_{χ^2} as evaluated over observables corresponding to the full NNPDF4.0 dataset is $\Delta_{\chi^2} = -0.002$ (compared to $\Delta_{\chi^2} = -0.009$ for NNPDF4.0), which is at the per mille level indicating a negligible amount of overfitting. The Δ_{χ^2} estimator has some shortcomings as will be discussed in Sect. 3.2.2. It is included here to provide a validation using the all the metrics that have been used for the NNPDF4.0 determination.

Let us estimate the faithfulness of the PDF uncertainty at the level of observables. For this we use the bias over variance ratio as defined in Eq. (6.15) of Ref. [6]. Here bias can be understood as a measure of the fluctuations of the observable values with respect to the central value prediction of the fitted PDF, while variance can be understood as the fluctuations of the fitted PDF with respect to its central value prediction. Thus if the methodology has faithfully reproduced the uncertainties in the underlying data

(bias), this uncertainty should be equal to the uncertainty in the predictions of the PDFs (variance), and hence the bias to variance ratio $R_{\mathrm{b}v}$ is expected to be one. To test this, the value of $R_{\mathrm{b}v}$ is determined for out-of-sample data. Specifically, we fit the PDFs to the NNPDF3.1-like dataset as defined in Ref. [6], and then evaluate the value of $R_{\mathrm{b}v}$ for the data that is part of the NNPDF4.0 dataset but has not already been included in the NNPDF3.1-like dataset. This allows us to test how well the predication made using a PDF fitted with a given methodology generalizes to unseen data. The value of the bias to variance ratio found for the new, feature scaling, methodology is $R_{\mathrm{b}v} = 1.03 \pm 0.04$ (compared to $R_{\mathrm{b}v} = 1.03 \pm 0.05$ for NNPDF4.0), where again the uncertainty corresponds to a 1σ bootstrap error, meaning the agreement to the expected value of $R_{\mathrm{b}v} = 1$ is at the 1σ level.

To estimate the faithfulness of the PDF uncertainty at the level of the PDF we calculate a quantile estimator in PDF space $\xi_{1\sigma}^{(\mathrm{p}df)}$. This quantity corresponds to the number of fits for which the 1σ uncertainty band covers the PDF used as underlying law. This is determined for fits performed to pseudodata covering the full NNPDF4.0 dataset. The result is $\xi_{1\sigma}^{(\mathrm{p}df)} = 0.70 \pm 0.02$ (compared to $\xi_{1\sigma}^{(\mathrm{p}df)} = 0.71 \pm 0.02$ for NNPDF4.0), where the uncertainty is a 1σ uncertainty determined through bootstrapping [168, 169]. Thus the observed $\xi_{1\sigma}^{(\mathrm{p}df)}$ value is in agreement with the expected value of 0.68 within 1σ.

An analogous estimator can be calculated for the theory predictions in data space as opposed to PDF space, providing a generalization to quantile statics of the bias of variance ratio $R_{\mathrm{b}v}$. Similar to the bias over variance ratio, also for this estimator the values are calculated on out-of-sample data, where the PDFs have been determined using NNPDF3.1-like data. The expected value of this quantile estimator depends on the bias over variance ratio is $\mathrm{erf}(R_{\mathrm{b}v}/\sqrt{2}) = 0.67 \pm 0.02$ (compared to $\mathrm{erf}(R_{\mathrm{b}v}/\sqrt{2}) = 0.67 \pm 0.03$ for NNPDF4.0), which is in agreement with the calculated value of $\xi_{1\sigma}^{(exp)} = 0.69 \pm 0.02$ (and $\xi_{1\sigma}^{(exp)} = 0.68 \pm 0.02$ for NNPDF4.0).

The validation tests carried out show that the feature scaling methodology produces faithful results. This way the feature scaling methodology achieves two important objectives. First, it validates the NNPDF4.0 determination by removing two possible sources of bias or inefficiencies without a significant change of the results. Second, it simplifies the PDF parametrization by automatizing steps that until now required human intervention and it removes a source of statistical fluctuations that interferes with the hyperoptimization routing. As such it provides a vital step towards the improvement of the hyperparameter selection protocol which we further develop in the next pages.

3.2 Improved hyperparameter selection

The k-folds hyperoptimization as described in Sect. 2.1.4 aims to obtain the best methodology, this being the one that provides the most accurate generalization of the data. While the automated hyperoptimization provides a useful tool to aid in the selection of the model hyperparameters, improvements can be made along two main trajectories to be discussed below, both of which improve the efficiency of the methodology by reducing the need for human interaction. Similar to the feature scaling proposed in the previous section, here we will propose directions for improvement

that allow to automate a human interaction, and show that the result confirms the faithfulness of NNPDF4.0.

The first improvement we will propose is based on the observation that for the problem of PDF determination not all datasets are equal: different datasets may constrain different kinematic ranges and correspond to different processes. As such, the choice of the folds can have a non-negligible impact on the hyperoptimization procedure. For instance, if we use two folds, one with only high-x data and another with only low-x data, the k-folding would be useless, since the extrapolation values will be, for all intents and purposes, random. It is then important to curate the folds in a way that ensures they are representative of the whole range of the problem. In NNPDF4.0 (see Table 3.2 of Ref [6]) the fold selection was a completely manual process aided by a very extensive appraisal of the individual datasets considered for the NNPDF4.0 release. This careful curation of the folds would, in principle, have to be repeated each time a change is made to the target dataset. In practice though, this does not have to be redone upon small changes to the dataset since there is a certain redundancy on the kinematic coverage of the data. Nevertheless, even the decision not to redo the selection is one that needs to be taken with care.

The second improvement we will propose relies on the observation that it is possible that the hyperparameter setup corresponding to the lowest hyperoptimization loss Eq. (2.13) results in overlearning. This is a consequence of the simplified nature of the fits that are performed during optimization as well as random samplings that affect the performance of the individual fits. One such random sampling that affected the model choice during hyperoptimization as performed for the NNPDF4.0 determination is the sampling of the preprocessing exponents for which we proposed a solution in Sect. 3.1. Other fluctuations include those due to the randomized initialization of the neural network and optimization algorithm. After deciding on a hyperparameter configuration it is necessary to validate the chosen methodology by performing future tests and closure tests. However, in particular closure testing requires a lot of computational resources and is therefore not feasible to perform for more than a select few configurations. Furthermore, while the closure test provides a method of testing the faithfulness of the PDF uncertainty, it lacks an adequate measure for the detection of inefficiencies due to overfitting. As a result, the fit corresponding to the antistrange and gluon PDFs presented in Fig. 2.11 passed a closure test, though visual inspection suggests that the clear wiggles present in the PDF replicas are unlikely to correspond to features of the underlying PDF of nature.

One may, and arguably should, wonder whether the features shown in Fig. 2.11 do truly correspond to overfitting. In Sect. 3.2.2 below, we will therefore define a measure for the degree of overfitting, and using this measure we will show that those features indeed correspond to overfitting.

3.2.1 Automated fold selection for hyperoptimization

The assessment of data is a very time-consuming task and, for the purposes of hyperparameter configuration, the only information we are interested in is which regions of the (flavour, x) space of the PDFs depend on a given dataset. Since we have an extensive corpus of data, small inefficiencies in the dataset selection lead to similarly small inefficiencies in the hyperparameter optimization. In practice this means that

suboptimal configurations are selected by the algorithm that need to be discarded further down the line. However, as we introduce more data these inefficiencies will only increase and on top of this, when new data is introduced it is not straightforward to define new folds.

The following algorithm aims to optimize the selection of the datasets for each fold in an automated way, offering a recipe that can be applied for every dataset change without losing previous information, and which is expected to produce an equally good result:

1. Create a level-0 dataset (see App. D) based on the predictions from a selected PDF replica.
2. Perform a fit to the dataset generated at 1. This fit will to serve as a reference.
3. Perform many fits, $\mathcal{O}(10^3)$, with a random selection of the datasets.
4. Train a neural network to learn the distance with respect to the reference fit based on a mask corresponding to a choice of fold.

In order to assess which datasets have an effect in which part of (flavor, x) space we start by removing any possible sources of inconsistencies. To this end, we create level-0 data[1]: fake data generated from a given underlying law (for instance, a replica from the NNPDF4.0 set) using theoretical predictions. The data created in this way is perfectly consistent since it eliminates all unknowns (known and otherwise) impacting experimental measurements and the corresponding predictions.

We then create a full fit targeting said fake data. Since the data still spans a finite range and it allows for a certain level of interpolation, doing a full fit with random initial states generates some spread of acceptable fits.

For the automated fold selection we train a neural network to learn how well the PDFs are constrained in (flavor, x) space based on a choice of folds as encoded in a mask. To this end we need to generate training data to train our model. This is done by performing many fits, $\mathcal{O}(10^3)$, to random selections of the datasets (note that for NNPDF4.0 the number of datasets is $\mathcal{O}(80)$ and thus the total number of possible combinations is completely impossible to test explicitly). Full fits are not needed, as the reference fit provides some spread. Intuitively, when we (randomly) hit a good fold we should find a PDF that it is very close to the reference PDF as shown in the left plot of Fig. 3.7 and instead, when we hit a very bad fold, we should find that, in some regions, the distance between the PDF and the reference is very bad as shown in the right plot of Fig. 3.7.

The random selection of datasets then acts as a boolean mask that, applied to the input collection of datasets, moves the resulting PDF away from the reference PDF. We can then select a measure of the distance between the reference and the result, and train a neural network to learn the optimal mask. Examples of such measures may be the Kullback-Leibler divergence [170], Hellinger distance [171], or the usual NNPDF distance defined in App. C. The loss function of this procedure is then defined as the integration of this distance over x, summed for all flavors. To obtain the results of Fig. 3.7 the Hellinger distance has been used.

[1] The concept of level-0 data was originally introduced in the context of closure tests as discussed in App D

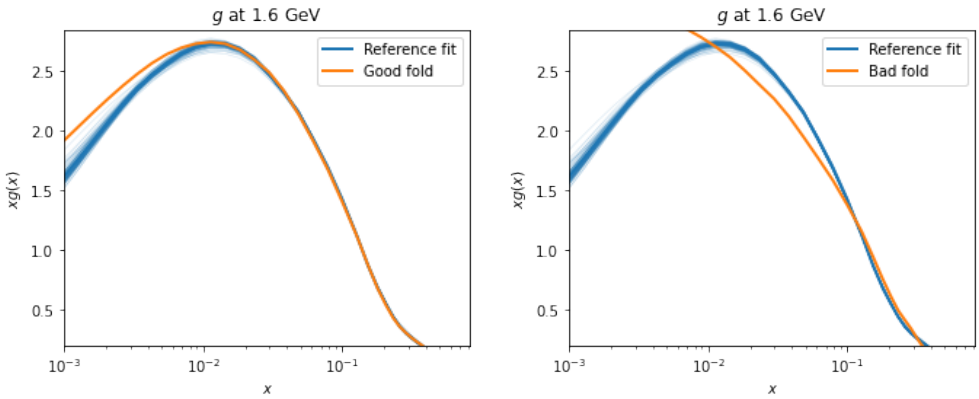

Figure 3.7: Resulting gluon for a good (left) and bad (right) choice of fold. The reference fit with its spread is represented in blue. The fold-fits (orange) are performed so that they can only access half of the datasets. In the left plot this subset of data is such that the gluon is well constrained in the entire x range. Instead, in the right plot the fold chosen doesn't contain enough information at small x for the gluon and the resulting PDF looks nothing like the reference.

This method allows us to automatically select folds with a cost equivalent to $\mathcal{O}(10)$ of the common PDF determinations. An added benefit of this method is that the library of fits can be reused for future dataset appraisals since for any extra dataset, the fits already performed correspond to masks where any new dataset is folded away.

3.2.2 Detecting overfitting

The automated hyperparameter selection based on k-folds cross-validation is useful to preselect potentially good hyperparameter setups, and to provide some constraint on the optimal range for certain parameters. Nevertheless, simply picking the setup that corresponds to the smallest hyperoptimization figure of merit without further inspection of the resulting PDF, can lead to significantly overfitted or underfitted results.

Since an overfitted solution correspond to a model that is, roughly speaking, more complex than the underlying law, the PDF arc-length and the spread thereof are commonly used as a diagnostic tool to analyze features related to overfitting. The arc-length of a PDF roughly corresponds to the complexity of the PDF replicas: if both the mean value and the spread in arc-length are relatively small, this can indicate a methodology that is not sufficiently flexible and therefore leads to underlearning, whereas if those quantities are relatively large, they may indicate a methodology that is too flexible and is learning noise in the data. However, the problem with the arc-length, and related measures of model complexity, is that it is unknown what the arc-length is expected to be from theoretical arguments. The PDF arc-length can be used to compare different fitting methodologies, but it does not provide an absolute measure of overfitting. While useful insights can be gained from studying how certain

changes to the methodology or dataset impact the arc-length, when it comes to the detection of overfitting or underfitting its benefits are very limited.

In practice the way in which the final hyperparameter selection was done in NNPDF4.0, is that after the hyperoptimization routine provided a ranking of $\mathcal{O}(10^3)$ setups, the ones with the lowest hyperoptimization loss as defined in Eq. (2.13) were selected and their distribution studied in detailed by creating Monte Carlo representations of the PDFs consisting of the standard number of 100 replicas fitted to the pseudodata replicas (compared to the fits to the experimental central values as performed during hyperoptimization). At this point human interaction is required to decide where the threshold of overfitting is by balancing a lower χ^2 loss against the increased complexity that this in general implies. While the determination of this threshold is a potential source of bias, we will show that no significant overfitting is present in the NNPDF4.0 determination.

In what follows, we will first review the criterion that has been applied in previous releases of NNPDF to avoid overfitting, after which we propose a statistical estimator for the degree of overfitting thus removing the need for a human intuition to detect overfitting. Although applied here to the NNPDF methodology, in principle it can be generalized to a variety of machine learning problems that rely on performing an ensemble of fits.

Detecting overfitting in a closure test

Once the hyperoptimization procedure has found a good candidate methodology we need to test (among other things) how well it generalizes to unseen data, i.e., we need to ensure that the methodology is not prone to overfitting.

The first step in the quality control of a methodology is the closure test which ensures that the uncertainties are faithfully reproduced. In fact, the closure test checks both overfitting and underfitting. This is based on the observation that in level-1 and level-2 closure tests (see App. D), the χ^2 calculated using the mean of the fitted PDFs should agree with the χ^2 computed using the input PDF. Thus in particular $\chi^2\left[\langle\mathcal{T}[f_{\text{fit}}]\rangle, \mathcal{D}_1\right] \approx \chi^2\left[\mathcal{T}\left[f_{\text{in}}\right], \mathcal{D}_1\right]$, where $\mathcal{T}\left[f_{\text{fit}}\right]$ is a theory prediction from a single fitted replica, $\mathcal{T}\left[f_{\text{in}}\right]$ is the theoretical prediction corresponding to the input PDF, $\mathcal{D}_1$ indicates that level-1 pseudodata has been used, and the brackets $\langle\cdot\rangle$ denote an averaging over replicas.

Whether a certain methodology leads to underfitting or overfitting is then tested using the Δ_{χ^2} metric defined as

$$\Delta_{\chi^2} = \frac{\chi^2\left[\langle\mathcal{T}[f_{\text{fit}}]\rangle, \mathcal{D}_1\right] - \chi^2\left[\mathcal{T}\left[f_{\text{in}}\right], \mathcal{D}_1\right]}{\chi^2\left[\mathcal{T}\left[f_{\text{in}}\right], \mathcal{D}_1\right]}, \tag{3.6}$$

which has first been introduced in Ref. [75]. This metric provides a measure for the difference between the χ^2 of the central prediction of the closure test fits, and the χ^2 of the input PDF set, both defined with respect to the same closure test dataset. This estimator thus provides a measure for how well the methodology is able to produce the theoretical predictions of the input PDF.

In particular, if the optimization algorithm is sufficiently efficient, it may decrease $\chi^2\left[\langle\mathcal{T}[f_{\text{fit}}]\rangle, \mathcal{D}_1\right]$ to a point where $\Delta_{\chi^2} = 0$ suggesting that the underlying law has been perfectly reproduced. However, in practice we may also find Δ_{χ^2} to be larger

than 0 indicating that the optimal χ^2 has not been reached, thus corresponding to an underlearning methodology. Or alternatively, we may find Δ_{χ^2} to be well below 0, this may indicate that the methodology has learned noise in the data, and is thus overfitted.

Nevertheless, a replica distribution can be perfectly sampled from the posterior distribution, while still resulting in a negative Δ_{χ^2}. This means that a good methodology can still result in a negative value, which was indeed the case for the NNPDF4.0 determination. In such a case the negative value could also correspond to a combination of smaller correlation with the level-1 data and a smaller bias. If, in such a scenario, the PDF uncertainties are reproduced correctly as indicated by the other estimators that constitute the closure test, we say the methodology has passed the closure test regardless of whether the Δ_{χ^2} estimator has a negative value or not.

The Δ_{χ^2} metric may nevertheless be used as a diagnostic tool when a methodology fails to produce the correct results for the other statistical estimators of the closure test. However, failing to produce the correct results for other estimators would already indicate a failed closure test, and as such the closure test is not able to discard methodologies purely based on overfitting.

In order to solve these problems, the future test targeting only overfitting has been introduced in NNPDF4.0. In the future test, several subsets of the global dataset are constructed. The future test passes if the uncertainties generated by the all subsets are such that the quality of the fit is compatible with the fit to the global dataset when accounting for PDF uncertainties.

While in principle the combination of future and closure tests should be able to discard bad methodologies, in practice they suffer of important drawbacks. The first is that a closure test is expensive to perform: to obtain values for the statistical estimators at the required level of accuracy it is needed to repeat the same fit around 25 times. The measure we will propose in the next section will only require a single fit. While the future test is computationally cheaper, it still requires several fits to different dataset variations. In addition it is, as previously mentioned, a relative measure in that compares the fit quality of the fit to a subset of the data to the fit performed to the full global datasets. Because of this, the future test can only be utilized if we know that the uncertainties of the full dataset fit are correctly propagated and it will always follow a closure test.

In what follows, we propose a metric which instead requires one single fit and provides an absolute number. This metric tackles both the problem of the computing cost of our current tests and allows for it to be used in an automated procedure.

A new overfitting metric

The basic idea of the overfit metric we propose here is that, since the validation data is used to test the generalization of the parametrization during the fit, no information about the validation data should be present in the final fitted PDFs. In practice some information about the validation pseudodata used during the fitting of the PDF replica may be present in the final result. This happens because the validation and training datasets are not fully uncorrelated, as such, a sufficiently efficient setup of hyperparameters may succeed at fitting even the validation pseudodata. This

phenomenon renders the early stopping algorithm an insufficient tool to prevent overfitting entirely.

So how do we test if the methodology has learned features from the validation pseudodata? Let us consider a fit of a given PDF replica $f^{(r)}$ to an underlying data replica $\mathcal{D}^{(r)}$, where r labels the replica number. If a PDF replica $f^{(r)}$ does not contain information on the specific data replica $\mathcal{D}^{(r)}$, then

$$\chi^2_{\text{val,r}}\left[\mathcal{T}\left[f^{(r)}\right],\mathcal{D}^{(r)}\right] = \frac{1}{N}\sum_{r'=1}^{N}\chi^2_{\text{val,r}}\left[\mathcal{T}\left[f^{(r)}\right],\mathcal{D}^{(r')}\right] \quad \text{if} \quad N\to\infty, \tag{3.7}$$

where the validation mask in the definition of $\chi^2_{\text{val,r}}$ equal to the mask used during the fitting of PDF relica $f^{(r)}$. Here the left hand side corresponds to the usual definition of the validation loss corresponding to the training of a PDF replica $f^{(r)}$. The right hand side corresponds to the expected validation loss of the PDF $f^{(r)}$ to independently sampled pseudodata replicas $\{D^{(r')} : r' = 1,\ldots,N\}$ not seen by the optimizer during training.

Using this insight, one may define as a measure of overfitting the difference between the left hand side and the right hand side of Eq. (3.7):

$$\mathcal{R}_O = \chi^2_{\text{val,r}}\left[\mathcal{T}\left[f^{(r)}\right],\mathcal{D}^{(r)}\right] - \frac{1}{N}\sum_{r'=1}^{N}\chi^2_{\text{val,r}}\left[\mathcal{T}\left[f^{(r)}\right],\mathcal{D}^{(r')}\right]. \tag{3.8}$$

A negative value of $\mathcal{R}_O$ is then a characteristic of an overfitting methodology.

Let us now, as an example, consider the methodology with a poor value of the `clipnorm` for which the resulting antistrange and gluon PDFs are shown in Fig. 2.11. Previously we stated a suspicion of overfitting, and here using the $\mathcal{R}_O$ metric we will confirm this suspicion. This is a fit that intuitively appears overfitted from the PDF plot in Fig. 2.11, though it passes all the quantitative validation checks used in NNPDF4.0. This methodology could therefore only be rejected based on close visual inspection of the resulting PDFs. Thus, at the very least, the lack of an objective measure for overfitting leads to a sizeable amount of wasted computing resources and human effort.

If we calculate the $\mathcal{R}_O$ value of Eq. (3.8) for PDFs consisting of 100 replicas, we find $\mathcal{R}_O = -0.023 \pm 0.012$, where the uncertainty is the 1σ uncertainty as determined through the bootstrapping method [168,169]. This indicates that the $\mathcal{R}_O$ value found for this fit is 1.9σ away from the $\mathcal{R}_O = 0$ point corresponding to no overfitting. We can then compare this to the $\mathcal{R}_O = -0.001 \pm 0.013$ found for the baseline NNPDF4.0 methodology.

This suggests that the intuitive impression that the fit without a carefully tuned gradient clipping is overfitted is correct and we have a metric that allows to measure to what extent the resulting PDF is overfitted. It further provides a confirmation that upon calculating $\mathcal{R}_O$ for the NNPDF4.0 baseline methodology, we find the expected result of $\mathcal{R}_O = 0$ with 0.1σ agreement.

The overfitting metric during model selection

So far we have seen how the $\mathcal{R}_O$ metric can be used to detect overfitting, making it a useful additional tool for the validation of a fitting methodology. We will now see how, due to the low computational costs required, it can be used to largely reduce the arbitrariness in model selection present in the current hyperoptimization procedure. In particular, it addresses the manual identification of the single best hyperparameter configuration among the ranking that the hyperoptimization routine generates. In practice this manual identification mainly involves balancing between a low χ^2 value or a low model complexity. While, all other things being equal, a lower model complexity is preferred per Occam's razor, but in general PDFs that have better agreement with the data are more complex (which may indeed indicate overfitting). In the hyperoptimization procedure used in NNPDF4.0 it was still up to the PDF fitter to strike a balance between the two and effectively pinpoint the threshold between underfitting and overfitting.

With the $\mathcal{R}_O$ overfitting metric of Eq. (3.8) this choice can be made algorithmically instead, even though we only have a measure of to identify overfitting while for all non-overfitted and underfitted the expected value of $\mathcal{R}_O$ vanishes. One may think that also a measure for underfitting is required, to make sure that a given potential candidate is not underlearned, but this is not necessary. Namely, an underfitted result will by definition have poorer agreement to the data than a well-fitted or overfitted result and as such will not be a preferred setup based on the corresponding χ^2. Thus, after running the hyperoptimization scan we can perform the following steps:

1. Perform regular N_{rep} replica fits for the best N_{meth} methodologies as ranked by the hyperoptimization routine.
2. Determine the $\mathcal{R}_O$ for these N_{meth} methodologies.
3. Discard all replicas with $\mathcal{R}_O < 0$ with more than $N_\sigma\sigma$ confidence.
4. Of the remaining configurations, select the one with the lowest χ^2_{val} averaged over replicas.

With the $\mathcal{R}_O$ we are now able to apply a completely algorithmic approach to the selection of the hyperparameter configuration. Nevertheless, it is clear that even this algorithm requires to make a choice about certain parameters, specifically we need to choose values for N_{rep}, N_{meth} and N_σ.

Of these, N_{meth} is rather inconsequential. It has to be sufficiently large to ensure that a spectrum of fits from underfitted to overfitted are included in the N_{meth} setups, but at some point (which in practice is found to be of order $\mathcal{O}(10)$ setups) this requirement is sufficiently satisfied. The reason that this is largely inconsequential can be understood from the fact that many configurations with very different values for the hyperparameters allow for equally complex solutions. This has been shown explicitly in section 3.3.4 of Ref. [6], where two very different hyperparameter configurations are shown to produce PDFs with good agreement.

Both N_{rep} and N_σ are more consequential. A larger vale of N_{rep} will result in a decrease of the bootstrap standard error as $1/\sqrt{N_{\text{rep}}}$, while N_σ defines our confidence threshold for when we consider a deviation $\mathcal{R}_O$ from 0 to be a signal rather than the result of statistical noise. Reasonable values for these that we recommend are $N_{\text{rep}} =$

100 as this is the standard sample size used to provide PDFs as part of any NNPDF release and is considered to produce a result with sufficient statistical accuracy for most purposes. At the same time N_σ can be taken to be conservatively small, we recommend a value around $N_\sigma = 0.5$. This will filter a significant number of configurations as a result of statistical noise instead of being truly overfitting, though, because many different configurations will produce the same results this is not a problem: if N_{meth} is chosen sufficiently large the likelihood of removing all "best" setups due to statistical noise is very small.

Reasonable choices for the parameters N_{rep}, N_{meth} and N_σ need to be made, though for the reason discussed above the final result will not be very sensitive to minor changes to these values, as long as the parameter values are chosen conservatively. What "reasonable" here may change depending on the complexity of the optimization problem but for the problem of fitting PDFs to the global NNPD4.0 dataset, the mentioned values are recommended.

Finally, it should be noted that while the $\mathcal{R}_O$ overfitting metric and the corresponding hyperoptimization algorithm have been designed with the NNPDF methodology in mind, its application is not limited to the NNPDF framework. In fact, the $\mathcal{R}_O$ value is particularly suitable for the detection of overfitting in the context of ensemble learning methods.

Chapter 4

Methodological uncertainties in PDFs

In chapter 2 we discussed the NNPDF4.0 PDF determination where the data uncertainties are propagate to the PDFs by producing Monte Carlo replicas of the experimental data and performing fits to these data replicas to produce Monte Carlo PDF replicas. As the determination of PDFs approaches the precision domain, the need for a careful understanding and validation of PDF uncertainties becomes increasingly important. In the previous chapter we therefore stressed the importance of testing the faithfulness of these uncertainties, as is currently mainly done using closure tests in the data region, and future tests in the extrapolation region. We also proposed an extension of the validation procedure by introducing a measure for overfitting that we suggest to be used in addition to the existing validation checks.

In this chapter we will discuss PDF uncertainties, specifically those related to the chosen parametrization. In Sect. 4.1 we study the correlation between different sets of PDFs, and we will see that a significant component of the correlations is not due to the underlying data because the data do not determine the PDFs uniquely. Specifically, we show that data-driven correlations can be used to assess the efficiency of methodologies and consider the feasibility of using data-driven correlations for the combination of different PDFs into a joint set. In Sect. 4.2 we briefly elucidate some aspects of the sampling method that is fundamental to the NNPDF methodology. In particular, we will discuss the minimization target used for the NNPDF4.0 determination and what consequences this has for the posterior distribution. We also introduce the kinetic energy of the PDF as a measure of complexity in order to provide an understanding of the likelihood given to certain candidate PDF functions that is usually obscured by the non interpretability of a neural network model.

4.1 Correlations between PDFs

Just as full information on data uncertainties requires knowledge of the experimental covariance matrix – and not just the uncertainties on individual datapoints – full information on PDF uncertainties is encoded in the point-by-point covariance matrix between all pairs of PDFs at any pair of points. Correlations between PDFs play a fundamental role in PDF determination. The study of correlations between PDFs and observables is a commonly used tool to understand the impact of PDF uncertainties on

theoretical predictions of observables, as well as the impact of individual observables on the PDF determination. This relevance was first emphasized in Ref. [172] and subsequently applied to Higgs phenomenology in section 3.2 of Ref. [173], and the construction of minimal PDF sets using the methodology discussed in Ref. [174].

The covariance matrix between PDFs reads

$$\mathrm{Cov}\left[f_a^p, f_a^q\right](x,x') = E\left[f_a^p\left(x,Q_0^2\right) f_a^q\left(x',Q_0^2\right)\right] - E\left[f_a^p\left(x,Q_0^2\right)\right] E\left[f_a^q\left(x',Q_0^2\right)\right], \quad (4.1)$$

while the related correlation matrix is

$$\rho\left[f_a^p, f_a^q\right](x,x') = \frac{\mathrm{Cov}\left[f_a^p, f_a^q\right](x,x')}{\sqrt{\mathrm{Var}\left[f_a^p\right](x)\,\mathrm{Var}\left[f_a^q\right](x')}}, \quad (4.2)$$

where $f_a^p(x,Q_0^2)$ is the p-th PDF in the set Φ_a with a momentum fraction x and at a scale Q_0^2, and variance

$$\mathrm{Var}\left[f_a^p\right](x) = \mathrm{Cov}\left[f_a^p, f_a^p\right](x,x) = E\left[f_a^p\left(x,Q_0^2\right)^2\right] - E\left[f_a^p\left(x,Q_0^2\right)\right]^2, \quad (4.3)$$

with E the average over the probability distribution of PDFs. In what follows, we will use indices a, b to label the PDF sets (NNPDF3.1, NNPDF4.0, MSHT20, CT18, ...) and the indices p, q to label the PDF flavors (up, down, anti-up, gluon, ...). In this chapter we will limit ourselves to the study of correlations where $x = x'$ and for brevity of notation the x and Q_0^2 dependence will be suppressed.

The PDF covariance matrix $\mathrm{Cov}\left[f_a^p, f_a^q\right](x,x')$ is the second moment of the joint distribution of PDFs. It can be computed in a standard way [76], given a representation of this distribution as a multigaussian in parameter space for a given PDF parameterization, or as a Monte Carlo sample of PDF replicas.

The correlation between different PDF flavors of the same PDF set as given in Eq. (4.2) has been widely computed and used. However, one may also define the correlation between different PDF sets [175,176]. Each PDF set is a determination of the universal true PDFs. As such, one may consider different PDF sets as independent determinations of the same physical quantity. Generally, a pair of independent determinations of the same quantity are characterized by an uncertainty and the correlation between them. In the presence of uncertainties, each determination may be thought of as a random variable. So for a distinct pair of determinations one can define their covariance and correlation, which may then be combined using the standard methodology that is used for the combination of correlated measurements [83,177]. Indeed, the correlation between two determinations expresses the amount of new information that each determination introduces. Namely, if two determinations are fully correlated they are repetitions of the same determination and hence a combination will not provide any additional information or allow for an improvement upon the uncertainty of the individual determination, while if two determinations are fully uncorrelated this indicates that the determinations are completely independent and the two determinations together contain more information than the individual determinations, thus allowing for an increased precision upon performing a combination.

In this picture we can view any PDF set Φ_a consisting of a PDF $f_a^p(x,Q_0^2)$ and its estimated error as an instance in the probability distribution of PDF

determinations [178], in the same way a measurement is an instance in the probability distribution of measurement outcomes. Thus we can generalize the covariance matrix of Eq. (4.1) by extending it to the characterization of different PDF sets Φ_a and Φ_b:

$$\text{Cov}\left[f_a^p, f_b^q\right] \equiv \text{Cov}\left[f_a^p, f_b^q\right](x,x) = E\left[f_a^p f_b^q\right] - E\left[f_a^p\right] E\left[f_b^q\right], \tag{4.4}$$

where in this case the corresponding correlation matrix reads

$$\rho\left[f_a^p, f_b^q\right] = \frac{\text{Cov}\left[f_a^p, f_b^q\right]}{\sqrt{\text{Var}\left[f_a^p\right]\text{Var}\left[f_b^q\right]}}. \tag{4.5}$$

Here again we assume that $x = x'$, and dependence of the PDFs on x and Q_0^2 has been suppressed in the notation.

Eq. (4.4) and Eq. (4.5) provide the covariance and correlation between two PDF flavors, p and q, from two different PDF sets, Φ_a and Φ_b. We will therefore refer to the covariance matrix of Eq. (4.4) as cross-covariance and to the corresponding correlation matrix of Eq. (4.5) as cross-correlation. The expected value in Eq. (4.4) is now defined as the average over the full probability distribution of PDF sets, hat is, over distinct determinations of which Φ_a and Φ_b are two generic instances.

If we set $p = q$, the cross-covariance Eq. (4.4) reduces to the correlation between two different determinations of the p-th PDF in the sets Φ_a and Φ_b, henceforth S-covariance:

$$\text{Cov}\left[f_a^p, f_b^p\right] = E\left[f_a^p f_b^p\right] - E\left[f_a^p\right] E\left[f_b^p\right], \tag{4.6}$$

while the same restriction on the cross-correlation Eq. (4.5) provides the S-correlation

$$\rho\left[f_a^p, f_b^p\right] = \frac{\text{Cov}\left[f_a^p, f_b^p\right]}{\sqrt{\text{Var}\left[f_a^p\right]\text{Var}\left[f_b^p\right]}}. \tag{4.7}$$

In this section it will be important to clearly distinguish the S-correlation Eq. (4.7) defined as a correlation in the space of different PDF sets, from the previously mentioned correlation defined in the space of PDFs within a given set Eq. (4.2). For clarity and brevity of notation, we will henceforth refer to the latter as F-correlation and the corresponding covariance Eq. (4.1) as F-covariance. Both the S-covariance and F-covariance are special cases of the cross-covariance Eq. (4.4).

F-covariance is a well understood and commonly used concept. If we consider a PDF set Φ_a of Monte Carlo PDF replicas denoted by $\{f_a^{p,(r)}(x, Q_0^2)\}$, where r labels the replica number, the F-covariance can be obtained by simply averaging over the replicas. The S-covariance rather less trivial since it requires the construction of PDF replicas that span the space of possible independent determinations of a given PDF – including, the results that might have been found by different groups using different methodologies. The S-covariance was explicitly determined for the first time in Ref. [77], and in this section we will discuss the non-trivial aspects of averaging over the space of PDF determinations.

The reason why averaging over the space of PDF determinations is subtle, is because the outcome of a PDF determination does only depend on the underlying data. The problem of PDF determination is that of determining a probability distribution in a space of functions [178] from a discrete set of data. Hence, the result is not

unique. If the PDFs are fitted to a functional parametrization with a relatively small number of free parameters when compared to the number of datapoints, such as the parametrization employed for the CT18 PDF determination [64], a unique best fit exists. However, in this determination, a fit is repeated with different parametrization choices, thereby leading to a distribution of best fits for a fixed dataset. In the NNPDF framework the PDFs are parametrized using a neural network that provides a very general functional form. As such, there is an ensemble of best fits of equal quality to fixed underlying data. This ensemble of best fits will be discussed in more detail in Sect. 4.2 of this chapter.

The distribution of the best fits for a fixed underlying dataset adds a contribution to the PDF uncertainty, and more generally to the corresponding covariance matrix, that is not driven by the covariance matrix of the underlying dataset. When one calculates the S-correlation as defined in Eq. (4.7), there is thus a contribution to it which comes from integrating over a space of PDF determinations from fixed underlying data. Since the correlation by construction has a value on the interval between -1 and +1, and there is a non-zero contribution to the correlation corresponding due to non data-driven components, this means that the data-driven component is strictly smaller than one. In what follows we will refer to the non data-driven component of the correlation as the "functional correlation" for notational brevity (without committing ourselves to its precise origin). It should be noted that this non data-driven component also contains uncertainties that are unrelated to the parametrization, for example choices in theory settings such as the value of the strong coupling constant, or missing higher order corrections. These can however in principle be accounted for by simply varying the relevant parameters (though in practice one immediately sees how this is problematic for non-parametric parameters), and we will not discuss them here: we will always consider theory assumptions to be fixed.

In this section, we will investigate the relative sizes of the data-driven and functional components. This will be done by computing explicitly the data-driven S-correlation for PDFs determined from the same underlying data.

Knowledge of the relative magnitude of the data-driven component of the correlation provides information relevant for applications. For example, as mentioned before, the F-correlation Eq. (4.2) is a commonly used tool for the assessment of the impact of a dataset on the determination of the PDFs. In this context it is then surely important to know how much of the PDF is determined from the underlying data. Furthermore, correlations between different determinations of the same quantity can generally be used to perform a combination of those different determinations. Combined PDF sets, such as the PDF4LHC21 set [147], can be viewed in a similar way, namely, as the combination of different determinations of the true PDF. At present, these combinations are performed by simply assuming that all PDF sets in the combination are equally likely. However, one might think that they should instead be combined as correlated measurements, and that a determination of the data-driven S-correlation might be useful to this goal [176]. It is then interesting to investigate possible ways to implement such a procedure, and their consequences.

In Sect. 4.1.2 we will present results for the data-driven component of the S-correlation, both between different PDF sets determined from the same underlying data and with the same methodology, and between pairs of PDF sets determined using the same data, but different methodologies. We will use the results to shed light

on the origin of the PDF S-correlation, and we will explain how S-correlations can be used as a diagnostic tool when comparing different methodologies. In Sect. 4.1.2 we will discuss the implications of our result for the construction of combined PDF sets.

4.1.1 PDF cross-correlations

In what follows we will explicitly calculate the cross-covariance given in Eq. (4.4), and from the knowledge of the covariance matrix determine the cross-correlation matrix of Eq. (4.5). To achieve this goal we will use PDFs consisting of Monte Carlo replicas produced using the NNPDF methodology, specifically the methodologies used for the NNPDF3.1 and NNPDF4.0 determinations. Even though here we limit the study to PDF sets in the Monte Carlo representation, this is not restrictive since PDFs in a Hessian representation can be converted to PDFs in a Monte Carlo representation using the methodology developed in Ref. [179]. A PDF set Φ_a is represented by a set of N PDF replicas $\{f_a^{p(r)} : r = 1, \ldots, N\}$ of the p-th PDF flavor, that provides an importance sampling of the probability distribution of the PDF set. We will assume that the number of PDF replicas N is fixed and sufficiently large to provide an accurate determination of the correlations. To this end, we will provide the uncertainty as a result of finite size effects along with the central value whenever we plot the correlations. Finally we assume that all the PDFs considered provide a faithful representation of the underlying data distribution such as they are tested in a closure test [6, 167].

Correlated replicas

Data replicas, and by extension PDF replicas, are independent and identically distributed random samples and hence the expected value of a statistical estimator X that is a function of the PDFs can be obtained through a simple averaging over replicas as given by Eq. (2.1) in Sect. 2.1.1:

$$\langle X\left[f_a^p\right]\rangle = \frac{1}{N_{\text{rep}}}\sum_{r=1}^{N} X\left[f_a^{p(r)}\right]. \tag{4.8}$$

In particular, the mean and F-covariance matrix Eq. (4.1) of a PDF set are thus given by

$$E\left[f_a^p\right] = \langle f_a\rangle\,, \quad \text{Cov}\left[f_a^p, f_a^q\right] = \langle f_a^p f_a^q\rangle - \langle f_a^p\rangle\langle f_a^q\rangle\,. \tag{4.9}$$

Henceforth, we will use the angle brackets to denote the average over replicas, while the symbol E represents a generic average.

PDF replicas are constructed by generating a Monte Carlo representation of the underlying data (i.e. an ensemble of data replicas) which can then be fitted using any chosen methodology to obtain a corresponding Monte Carlo representation in PDF space (i.e. an ensemble of PDF replicas). The details of this procedure have previously been discussed in Sect. 2.1.1. It should be noted that this approach is not unique to the neural network parametrization employed by NNPDF, instead, also polynomial PDF parametrization such as those employed by the MSTW analysis [180] can be used in conjunction with the Monte Carlo approach, which has been shown

to produce PDF uncertainties that are in good agreement with the corresponding uncertainties resulting from the Hessian approach [179].

When computing the F-covariance between PDFs of different flavors p and q in the same PDF set Φ_a, both flavors have been fitted to the same data replicas, so in particular

$$\langle f_a^p f_a^q \rangle = \frac{1}{N} \sum_{r=1}^{N} f_a^{p(r)} f_a^{q(r)}. \tag{4.10}$$

Here it is important to understand that a PDF replica $f_a^{p(r)}$ has been fitted to a data replica with the same index r. However, the PDF replica $f_a^{p(r)}$ is generally not a unique solution for a data replica with index r, instead there exists a probability distribution of PDFs that correspond to a given underlying data replica. Within the NNDPF methodology this can be explained as a result of the complexity of the neural network being larger than model complexity preferred by the data, and thus the best-fit neural network trained to fixed underlying data is not unique. In practice, because of the stochastic nature of the initialization of the NNPDF model, as well as the optimization algorithm, this means that upon repeatedly fitting the same data replica, multiple equally good solutions will be found. This will be explicitly shown in Sect. 4.2. In the methodology employed by other PDF fitting groups a similar scenario unfolds, however in these cases it stems from fitting each data replica with an ensemble of different functional forms, each resulting in a different best fit solution [63,64,181]. It is clear that this spread in the PDFs fitted to the same fixed set of underlying data is unrelated to the covariance matrix of the data. This is why, to calculate the covariance, in Eq. (4.10) $f_a^{p(r)}$ and $f_a^{p(r)}$ need to correspond to the same PDF replica, that is two replicas that have been fitted to the same underlying data replica.

Now let's turn our attention to the computation of the S-covariance and S-correlation between two PDF sets Φ_a and Φ_b, where the S-covariance can be written as

$$\mathrm{Cov}\left[f_a^p, f_b^p\right] = \langle f_a^p f_b^p \rangle - \langle f_a^p \rangle \langle f_b^p \rangle, \tag{4.11}$$

while the S-correlation can be written as

$$\rho\left[f_a^p, f_b^p\right] = \frac{\mathrm{Cov}\left[f_a^p, f_b^p\right]}{\sqrt{\mathrm{Var}\left[f_a^p\right] \mathrm{Var}\left[f_b^p\right]}}. \tag{4.12}$$

Naively, one may think that the S-correlation can be calculated between any two PDF sets, $\{f_a^{p(r)}\}$ and $\{f_b^{p(r)}\}$, however it should be noted that if the PDF replicas are independently sampled then the S-covariance, and thus the S-correlation Eq. (4.7), will vanish. In fact, the same applies to the F-covariance, and F-correlation Eq. (4.2). Namely if we were to consider a PDF set Φ_a consisting of $2N$ independent PDF replicas, we find

$$\frac{1}{N} \sum_{r=1}^{N} f_a^{p(r)} f_b^{p(N+r)} = \langle f_a^p \rangle \langle f_b^p \rangle. \tag{4.13}$$

and thus the S-covariance Eq. (4.11) vanishes within finite size effects resulting from the Monte Carlo representation.

To determine the correlation between different PDF sets, we thus need to calculate the S-covariance between two PDF sets Φ_a and Φ_b where the constituent replicas of both sets are correlated. To clarify what this means, let us assume, for the sake of argument, that the data does uniquely correspond to a best fit PDF. In this case, each replica $f_a^{p(r)}$ is in fact uniquely determined from the underlying data replica with the same index r. If under this assumption we consider two PDF sets Φ_a and Φ_b that have been fitted to the same underlying data replicas, but using different methodologies, the unique best fits are nevertheless not exactly the same, i.e. $f_a^{p(r)} \neq f_b^{p(r)}$. The S-correlation between those sets however, can now be calculated using Eq. (4.11) and Eq. (4.12). The S-correlation will generally be non-zero since the PDF replicas $f_a^{p(r)}$ and $f_b^{p(r)}$ are correlated through the underlying data replica r. It is now obvious that if $a = b$, the S-correlation is equal to unity. Namely, this is because the unique solution $f_a^{p(r)}$ is used in the computation of averages.

However, we know that in practice this is not the case since the data replicas do not uniquely determine the corresponding PDF replicas. In fact, the replicas correspond to a distribution of best fits $\{f_a^{p(r,r')}\}$. In the same way r runs over data replicas, r' runs over "functional" or methodological replicas. For each data replica with index r, the index r' labels all other aspects that determine the answer. For example, if we consider a methodology in which the functional uncertainty is estimated by varying over an ensemble of different functional forms, such as in Refs. [63, 64, 181], the index r' would label the different functional forms. If $a = b$, then for fixed r and fixed r' the same answer is obtained and the correlation is one: each PDF set has unit correlation to itself. The full S-correlation is thus obtained, not only by varying over data replicas r, but also requires varying over "methodology replicas" r' in a correlated manner. However, if only r is varied in a correlated manner between Φ_a and Φ_b, while variations in r' are uncorrelated, this again leads to a vanishing of the corresponding component following Eq. (4.13).

Thus, if only r is correlated, but r' is uncorrelated, the computation of the S-correlation will only account for the data-driven component of the S-correlation while the functional component vanishes. If we then set $a = b$ in Eq. (4.11), the S-covariance matrix reduces to variance and the full S-correlation should be equal to unity within statistical fluctuations due to finite size effects. After all, we are calculating the correlation of a PDF set to itself. However, as discussed, if between two determinations of Φ_a, only the data replicas r are varied in a correlated way while the methodological replicas r' are completely independent, the functional component of the S-correlation is not accounted for in the calculation and may come out to be less than unity. Thus, the amount by which the data-driven component of the S-correlation deviates from unity is a measure of the correlation as a result of not integrating over functional replicas in a correlated manner. It tells us how significant the functional component of the S-correlation is, with respect to the data-driven component. This procedure thus tells us, crudely speaking, to what extent the PDFs are determined from the underlying data.

One can then wonder if the functional component of the S-correlation can be calculated directly. To do this, it would be needed to produce two sets of PDF replicas $\{f_a^{p(r,r')}\}$ such that varying the index r' fully spans the possible best fits obtained from a fixed underlying data replica with index r. It would be needed to

produce "functional replicas" as well as data replicas, if you will. This may – at least in principle – be possible for certain parametric model parameters contained in r' such as the value of the strong coupling constant α_s, and the preprocessing exponents discussed in Sect. 2.1.3 for which we will calculate the impact explicitly below. However, this appears to be nontrivial in general for non-parametric aspects of the methodology that determine in which of the many equivalent best fits a particular minimization will end up.

In what follows we will focus on the computation of the data-driven component of the S-correlation using Eq. (4.11). This means that we will compute the S-correlation between PDF sets Φ_a and Φ_b, consisting of the PDF replicas $\{f_a^{p(r)}\}$ and $\{f_b^{p(r)}\}$ respectively, where we assume that in both cases the index r corresponds to the same underlying data replicas. We will also "compare a PDF set to itself". By this we mean that we calculate the cross-correlation or S-correlation between two sets of PDF replicas that have been fitted using the same methodology $\{f_a^{p(r,r')}\}$, and $\{f_a^{p(r,r'')}\}$. In this case again, r also labels the underlying data replica that is the same for both determinations, while the indices r' and r'' span the space of best fits to the fixed data replica.

One may write

$$\langle f_a^p f_a^p \rangle = \frac{1}{N} \sum_{r=1}^{N} f_a^{p(r,r')} f_a^{p(r,r'')}, \tag{4.14}$$

though since we can not vary r' and r'' in a correlated manner between two fits, in practice $\{f_a^{p(r,r')}\}$, and $\{f_a^{p(r,r'')}\}$ are two different fits from Φ_a to the same set of underlying data replica. To better understand what this implies, it can be contrasted with the F-covariance as calculated using Eq. (4.10), where the PDFs are taken from a single fit and therefore corresponds to having $r' = r''$ in Eq. (4.14). As stated above, in what follows we will refer to the situation of Eq. (4.14) as comparing a PDF to itself and the resulting quantity will be referred to as the data-driven component of the cross-correlation or S-correlation.

The data induced S-correlation and cross-correlation

We compute the data-induced component of the S-correlation, both between a pair of PDF sets determined from the same underlying data using two different methodologies, and between a PDF set and itself, as defined above. Specifically, we consider the NNPDF3.1 methodology as presented in Ref. [59], and the NNPDF4.0 methodology. The differences between both methodologies have been discussed in Sect. 2.1, though here the details of these methodologies are not relevant, and it suffices to know that they are both faithful, and compatible with each other as shown in the luminosity plots in Sect. 2.3.2 (and will again be verified explicitly for the PDFs here).

In order to compute the F-correlations and S-correlations to be discussed below, we have generated several PDF sets of approximately 1000 PDF replicas each, using the open-source NNPDF code. These PDF replicas have been generated by performing fits to a dataset which consists of the deep-inelastic scattering (DIS) data used for the NNPDF3.1 PDF determination, as listed in Table 1 of Ref. [59]. The kinematic coverage of this data in the (x, Q^2) plane is shown in Fig. 4.1. The reasons for choosing a DIS-only dataset are to limit the use of computational resources, as well as to deal

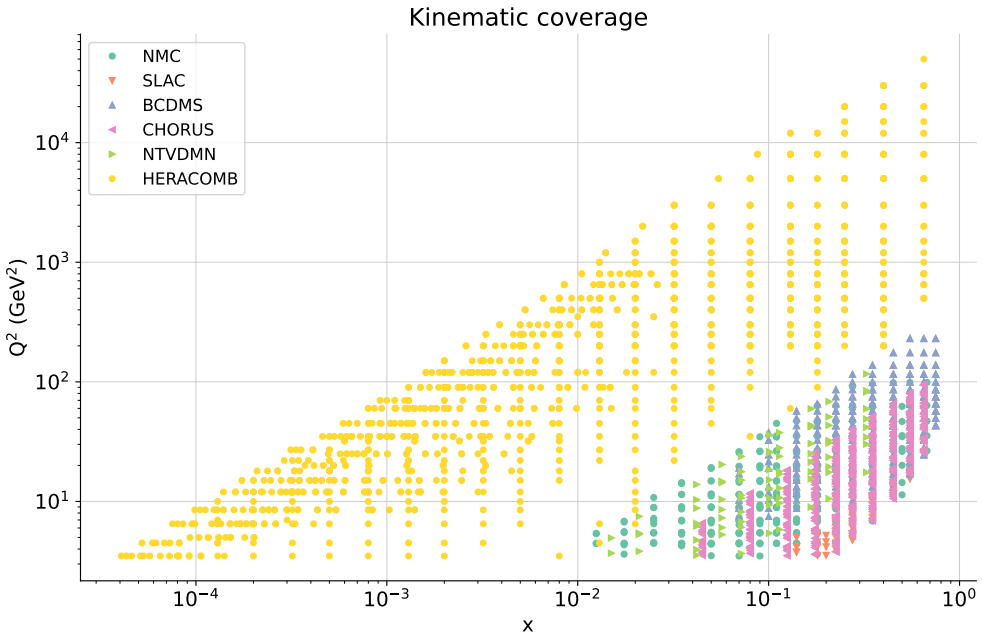

Figure 4.1: The gluon PDF as determined with the NNPDF3.1 and NNPDF4.0 methodologies, on a linear (left) and logarithmic (right) scale in x.

with a dataset involving a single, well-understood process, thereby avoiding possible complications related to tensions between data, slow perturbative convergence, and other issues that could obscure our conclusions. Uncertainties on the S-correlations due to the finite size of the replica sample are estimated using a bootstrapping procedure. Further details on the computation of the S-correlation are given in Appendix B.

To begin with, we have constructed four PDF sets, all determined from the same underlying data replicas: two using the NNPDF3.1 methodology, and two using the NNPDF4.0 methodology. Whereas a detailed comparison of PDFs produced using the NNPDF4.0 and NNPDF3.1 methodologies is given in Ref. [6] and for the luminosities in Sect. 2.3.2, in Fig. 4.2 we show a representative comparison of the gluon PDF as determined using these two methodologies. The general features of the comparison discussed in Sect. 2.3.2 are apparent from this example: namely, first, that results found with either methodology are compatible within uncertainties and central values are generally quite similar, and second, that the NNPDF4.0 methodology leads to rather smaller uncertainties, so generally whereas the NNPDF4.0 central value is within the NNPDF3.1 uncertainty band, the NNPDF3.1 central value is not within the rather smaller NNPDF4.0 uncertainty band.

As in all NNPDF determinations, it can be checked explicitly that independently determined PDF replicas all provide a consistent representation of the same underlying probability distribution. Namely, we can check that the standard deviation of the mean of $N_{\rm rep}$ replicas is equal to $\sigma/\sqrt{N_{\rm rep}}$. In order to perform this check, we have generated yet another set of PDF replicas with the NNPDF4.0 methodology, now based

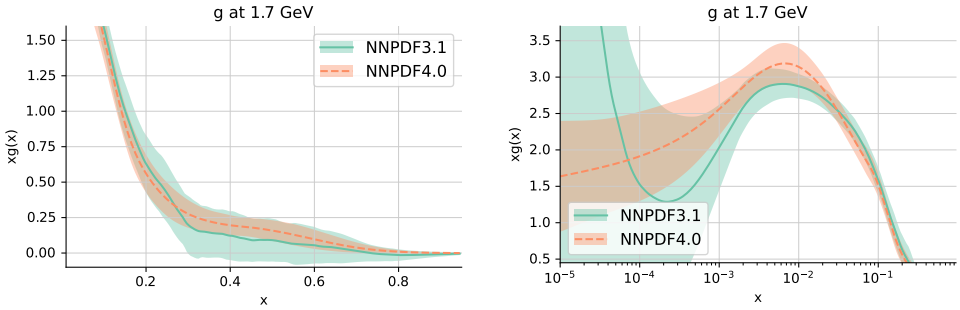

Figure 4.2: The gluon PDF as determined with the NNPDF3.1 and NNPDF4.0 methodologies, on a linear (left) and logarithmic (right) scale in x

on a new set of data replicas. In Fig. 4.3 we show the distance between the central values and uncertainties of the two different sets of PDF replicas determined using the NNPDF4.0 methodology trained on different sets of data replicas. The distance (defined in Appendix C) is the mean square difference of central values in units of the standard deviation of the mean. It is apparent that indeed its value is of order one, as it ought to be. This shows that as the number of replicas increases, both central values and uncertainties of PDFs converge to the mean and standard deviation of the underlying probability distribution.

Having verified that samples of PDF replicas behave as expected, we now proceed to the computation the data-induced component of the S-correlation Eq. (4.7). Results are shown in Fig. 4.4, where we show the S-correlation between two sets of replicas determined with the NNPDF3.1 methodology (orange), between two sets of replicas determined with the NNPDF4.0 methodology (green), and between a set of replicas determined with the NNPDF3.1 methodology and a set of replicas determined with the NNPDF4.0 methodology (blue). The error bands show the 2σ uncertainty due to the finite size of the replica set, estimated using bootstrapping (see Appendix B).

It is apparent from the plots that the data-induced PDF S-correlation drops very quickly to zero outside the data region, as it ought to. For light quarks, the correlations drop to zero for $x \lesssim 10^{-4}$ and $x \gtrsim 0.4$, while the data region is rather smaller for the gluon and heavy quarks. This is because in a DIS-only PDF determination the gluon PDF is only determined indirectly by scaling violations, while the PDFs for heavier quarks are determined by charged current data, i.e. mostly by fixed-target neutrino DIS data. However, interestingly, even in the middle of the data region the S-correlation of pairs of PDF sets determined using the NNPDF3.1 methodology is typically around 40% and never exceeds 60%. The S-correlation of PDF sets determined using the NNPDF4.0 methodology, in turn, is typically around 60% and never exceeds 80%. The S-correlation between PDF sets determined using the NNPDF3.1 and NNPDF4.0 methodologies, finally, is very similar to the S-correlation between the pair of NNPDF3.1 PDF sets.

As discussed before, the deviation from unity of the data-driven S-correlation when comparing of a PDF set to itself is a measure of the size of the functional component of the S-correlation. Both for the NNPDF3.1 and NNPDF4.0 methodologies, this

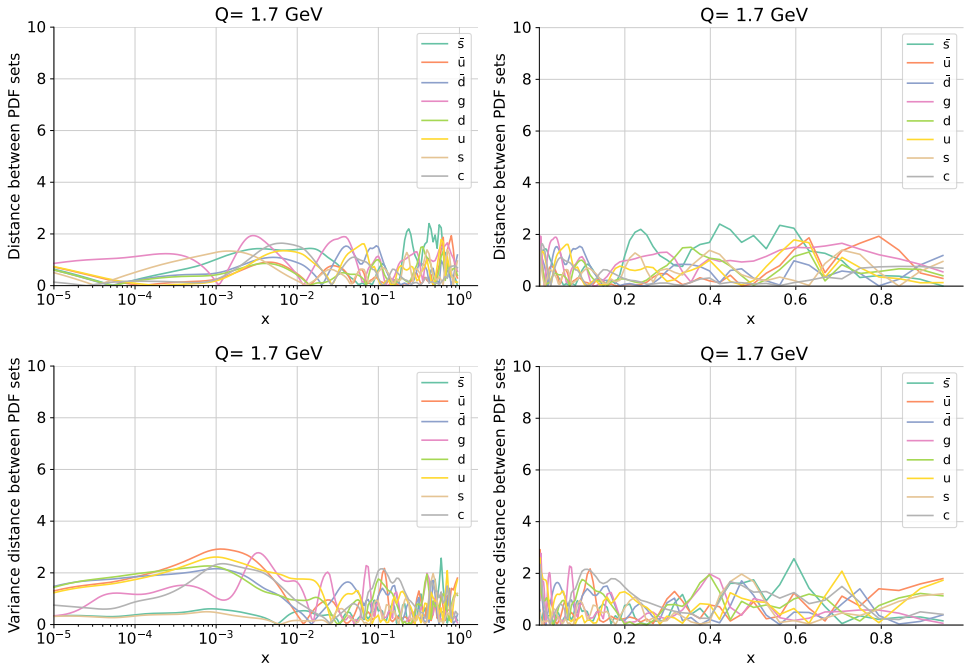

Figure 4.3: Distances between central values Eq. (C.5) (top) and uncertainties Eq. (C.6) (bottom) of two PDF sets determined using the NNPDF4.0 methodology. Results are shown both on a logarithmic (left) and linear (right) scale in x. The distances are discussed in more detail in Appendix C.

deviation is substantial. Note that, as shown in Fig. 4.3, any two independent sets of replicas for the same set have the same mean and uncertainty within finite-size fluctuations, and that these fluctuations scale as expected and in particular go to zero in the limit of a large number replicas. Note also that the deviation of the S-correlation from 100% is much larger than the uncertainty due to the finite size of the replica sample, so the correlation loss cannot just be due to an insufficient number of replicas.

The fact that the S-correlation of NNPDF3.1 PDFs is smaller than that of NNPDF4.0 PDFs is consistent with the fact that the NNPDF4.0 methodology leads to smaller uncertainties than the NNPDF3.1 methodology, even though both can be shown to be faithful using closure tests. Indeed, the only way PDF sets determined from the same underlying data can have different uncertainties is if one of the two has a smaller functional (i.e. non data-driven) component of the uncertainty. But then we would expect that the methodology characterized by a smaller functional uncertainty also has a smaller functional S-correlation: i.e. that it is determined to a greater extent by the underlying data. This is indeed what happens here: NNPDF4.0 has a smaller uncertainty for a fixed dataset, and accordingly a larger S-correlation.

It is interesting to observe that the data-induced component of the S-correlation between NNPDF3.1 and NNPDF4.0 PDFs is almost always similar to that of the

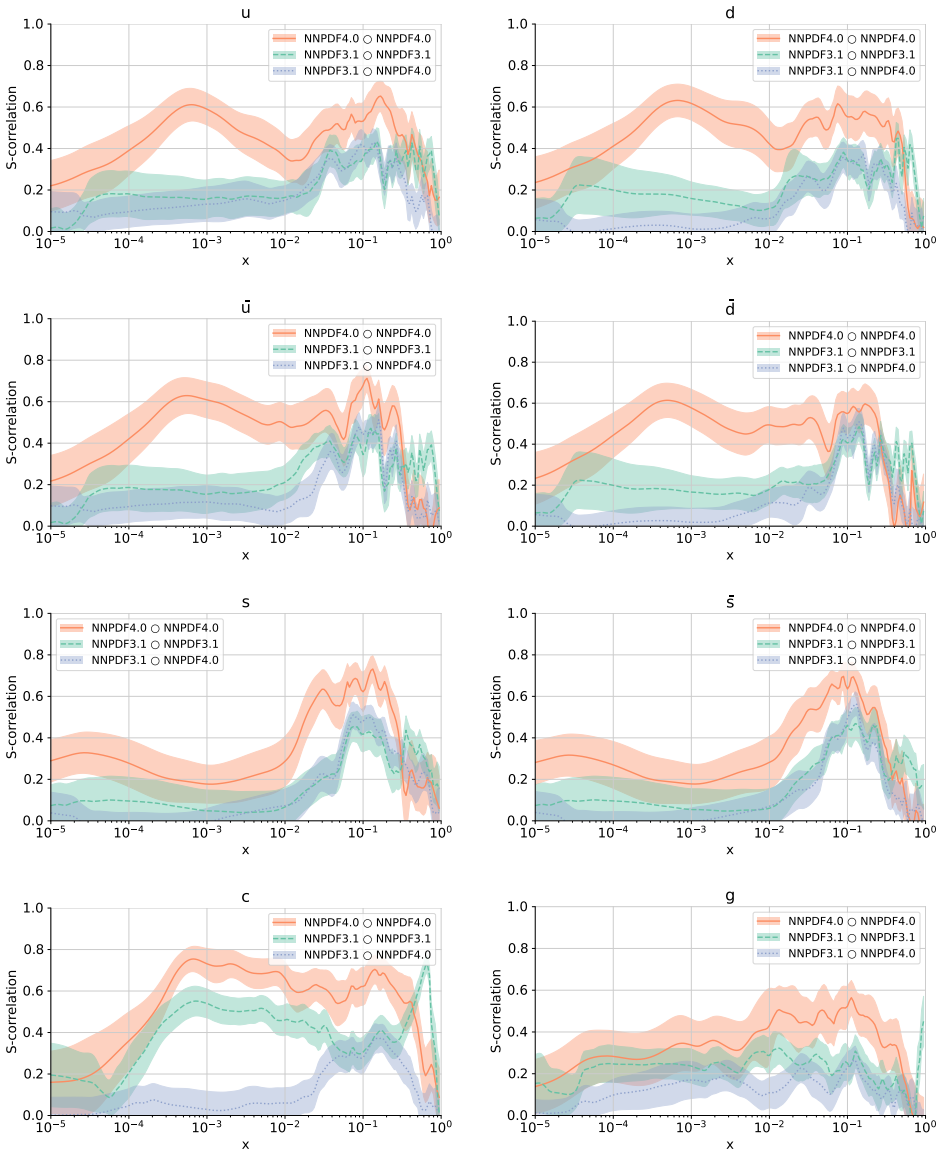

Figure 4.4: The data-driven component of the S-correlation Eq. (4.11). Results are shown for all PDF flavors and the gluon. We consider PDF sets determined with the NNPDF3.1 methodology or the NNPDF4.0 methodology, and the three curves shown correspond to comparing pairs of sets with either methodology to themselves (NNPDF3.1: green, NNPDF4.0: orange) or with each other (blue). The shaded band for each curve is the 2σ uncertainty due to the finite size of the replica sample, estimated by bootstrapping (see Appendix B). Here and below ◯ denotes the operation of comparing quantities computed from two sets of replicas that are correlated by being fitted to the same underlying data.

PDF set that has the smallest correlation to itself, namely NNPDF3.1 (a possible exception being the charm PDF, which is a special case because in a DIS-only fit it is almost undetermined). This suggests a "weakest-link" explanation: the data are more weakly correlated to NNPDF3.1 than to NNPDF4.0, and so inevitably the data-driven correlation between NNPDF3.1 and NNPDF4.0 is dominated by this weaker correlation. This is apparent, for instance, in the small x region, where the data-driven S-correlation of NNPDF3.1 to itself is significantly weaker than that of NNPDF4.0.

All this suggests that the data-driven component of the S-correlation between PDF sets for a given methodology can be used as a criterion for the assessment of the efficiency of the methodology itself, with the interpretation that a methodology leading to higher cross-correlation is more efficient. Namely, PDFs determined using a methodology characterized by higher S-correlation have a smaller functional component of the S-correlation, i.e. they are to a greater extent determined by the underlying data. So for instance the weaker S-correlation of NNPDF3.1 at small x suggests that in this region the NNPDF3.1 uncertainties could be reduced without loss of accuracy, as is indeed the case [6].

In order to further investigate the functional component of the S-correlation, we have produced sets of PDF replicas in which some methodological choices are correlated or decorrelated. First, we have produced a set of replicas in which preprocessing is also correlated. To understand this, recall that neural networks used to parametrize PDFs include a preprocessing function Eq. (2.3), whose parameters are randomly varied between replicas. We have thus produced a new pair of NNPDF4.0 replicas in which not only the data, but also the preprocessing exponents are correlated: so in Eq. (4.11), replicas $f_a^p(x,Q_0^2)^{(r)}$ and $f_b^p(x,Q_0^2)^{(r)}$ are not only fitted to the same underlying data, but also have the same value of the preprocessing exponents.

Results are shown in green in Fig. 4.5, compared to the previous results of Fig. 4.4, shown in orange. It is clear that in the data region, where the data-induced S-correlation is largest, the extra correlation due to preprocessing is negligible. As PDFs extrapolate further away from the data region, the contribution due to preprocessing is increasingly large: for instance for the gluon at large $x \gtrsim 0.4$ the data-induced correlation rapidly drops to zero as $x \to 1$, but the correlation due preprocessing makes up for the decrease and in fact it somewhat exceeds it as the kinematic boundary at $x = 1$ is approached.

Furthermore, we have produced a PDF set, based on the NNPDF4.0 methodology, but with a different architecture of the neural net, i.e., different number of layers and layer sizes. This is thus effectively a variation of the NNPDF4.0 methodology. Results, also shown in Fig. 4.5 (in blue), demonstrate that this specific aspect of the methodology has little impact: the S-correlation is essentially the same as in the case where the architecture of the neural networks is the same in the two sets being compared. This shows that these two methodologies lead to very similar results, which in turn suggests that correlating the neural network architecture would have a less significant impact than that of correlating preprocessing.

These two examples illustrate how, at least in principle, all components of the S-correlation could be determined, namely, by correlating all methodological aspects that determine the final result. As already discussed, whereas this is easily done for parametric choices (like the values of the preprocessing exponents), it is rather more difficult for non-parametric aspects, such as, for instance, the choice of minimization

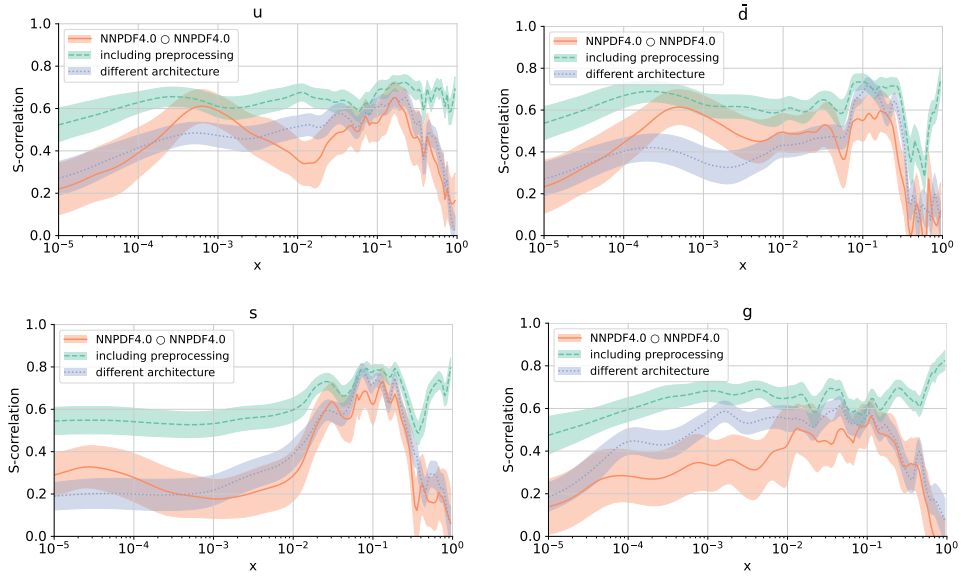

Figure 4.5: The data-driven component of the S-correlation Eq. (4.11) between PDFs determined with the NNPDF4.0 methodology (same as Fig. 4.4) compared to the case in which also the preprocessing-induced component is included in the correlation (green), and the case in which the neural network architecture is changed with all other aspects of the methodology kept fixed (blue). Results are shown for the up, anti-down, strange and gluon PDFs.

settings. These aspects are of course closely tied to the non-uniqueness of the best fit for given data, which leads to functional uncertainties.

Finally, we compare, for a fixed methodology, the data-driven component of the S-correlation for a pair of PDF flavors, to the cross-correlation between them: in Fig. 4.6 results are shown for the up and the down PDFs, both for the NNPDF3.1 methodology and the NNPDF4.0 methodology. Clearly, unlike the diagonal S-correlation in the flavor basis Eq. (4.7), the correlation between two different PDF flavors need not be positive. This is indeed seen in the figure, namely, the up and down PDFs turn out to be anti-correlated at large x. Furthermore, one would generally expect any cross-correlation between two different PDF flavors to be weaker than the S-correlation — indeed, if all sources of S-correlation were included, the S-correlation would be 100%. This is again borne out by the explicit computation, that shows that the cross-correlation is generally smaller in modulus than the S-correlation. Note that for the less efficient NNPDF3.1 methodology, for which all S-correlations are smaller, as already discussed, the cross-correlation is accordingly smaller in modulus.

In the same figure we also compare, again for a fixed methodology, the data-induced component of the cross-correlation between the up and down PDFs to the standard F-correlation Eq. (4.2). Note that if the cross-correlation was entirely data-driven, for a fixed methodology these two quantities would coincide. But in actual fact they differ because in the computation of the F-correlation, Eq. (4.10) is used (the replicas are

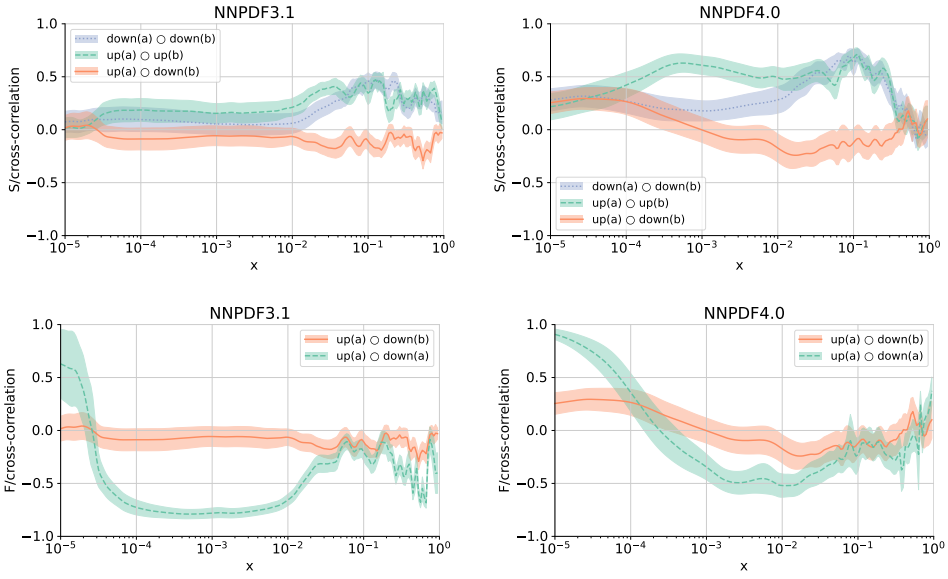

Figure 4.6: Top: comparison of the data-driven component of the S-correlation Eq. (4.11) of the up and down PDFs and the up-down cross-correlation Eq. (4.5). Bottom: comparison between the data-driven component of the up-down cross-correlation Eq. (4.5) and the standard F-correlation Eq. (4.2) for the up and down quark PDFs. Results are shown for the NNPDF3.1 (left) and the NNPDF4.0 methodology (right). In all plots (a) and (b) denote two distinct sets of correlated replicas (i.e. fitted to the same underlying data), so (a) ◯ (b) denotes the case in which the replicas are correlated but distinct while (a) ◯ (a) denotes the case in which the replicas are identical.

fully correlated), while in the computation of the cross-correlation, Eq. (4.14) is used (the replicas are only correlated through data). Whenever the S-correlation is sizable, the data-induced component of the cross-correlation, and the F-correlation are very close to each other. This means that the functional component of the PDF uncertainty is essentially uncorrelated between different PDFs, i.e. that the standard F-correlation is due to the correlation between the underlying data. This is as one expects, and justifies using PDF correlations to estimate the impact of data uncertainties on PDFs uncertainties and conversely.

However, when the data-induced component of the S-correlation is small, then the S-correlation can differ significantly from the F-correlation. This is clearly seen when comparing the up-down correlation in the region $10^{-4} \lesssim x \lesssim 10^{-2}$ computed with the NNPDF3.1 methodology to that computed with the NNPDF4.0 methodology. With the NNPDF4.0 methodology, the data-driven S-correlation in this region is large, and the standard up-down correlation is quite close to the cross-correlation. With the NNPDF3.1 methodology, on the other hand, the data-driven S-correlation is almost vanishing. This means that, with NNPDF3.1 methodology, PDFs in this region are completely dominated by functional uncertainty, which then screen out the

PDF correlation when computing the cross-correlation. In other words, the functional component of the S-correlation (between a PDF and itself) is generally non-negligible, while the functional component of the cross-correlation (between two different PDFs or two different x values) is generally quite small.

4.1.2 Combined PDF sets

As mentioned in the introduction of Sect. 4.1, combined PDF sets have been produced [76, 182], with the goal of providing a common, conservative PDF determination. The underlying idea is that so-called global PDF sets, namely PDFs determined using the widest possible amount of experimental information available at a given time, differ due to theoretical and methodological assumptions. If these assumptions all satisfy reasonable criteria of reliability, (as will be specified below) these different PDF determinations are considered to be equally likely, and thus a conservative choice is to combine them into a single determination. One may then reasonably ask whether this combination might be constructed in such a way as to explicitly take account of the cross-correlation between PDF sets.

The PDF4LHC15 combination

The PDF4LHC15 prescription [76] for combining different PDF sets Φ_a, $\Phi_b, \ldots$ may be described as follows. All sets being combined are turned into a Monte Carlo representation, i.e. each of them is represented by a set of N PDF replicas $\{f_i^p(r) : r = 1, \ldots N\}$, where the index i runs over the sets that are being combined. The combined set is then defined simply as the union of the replicas in the individual sets: specifically to combine two sets Φ_a and Φ_b, we select randomly $N/2$ replicas from each set, and then define the replicas for the combined set $\{F^p\} = \{F^{p(r)} : r = 1, \ldots N\}$ as

$$F^{p(r)} = \begin{cases} f_a^{p(r)} & \text{for } r = 1, \ldots, \frac{1}{2}N, \\ f_b^{p(r)} & \text{for } r = \frac{1}{2}N + 1, \ldots, N. \end{cases} \tag{4.15}$$

The combination assumes that, in the absence of an objective criterion for deciding on the relative probability of the different sets, they are equally probable, and thus that we should take the same number of replicas from each set.

The combined set $\{F^p\}$ is then treated in the same way as the individual sets for the calculation of estimators, with the averages taken over all the replicas in the set according to Eq. (2.1). Thus in particular the mean of any PDF in the combined set is immediately seen to be given by

$$E[F^p] = \tfrac{1}{2}(\langle f_a^p \rangle + \langle f_b^p \rangle), \tag{4.16}$$

while the F-covariance between two PDFs is

$$\begin{aligned} \mathrm{Cov}[F^p, F^q] =& \langle F^p F^q \rangle - \langle F^p \rangle \langle F^q \rangle \\ =& \tfrac{1}{2}(\langle f_a^p f_a^q \rangle + \langle f_b^p f_b^q \rangle) - \tfrac{1}{4}(\langle f_a^p \rangle + \langle f_b^p \rangle)(\langle f_a^q \rangle + \langle f_b^q \rangle) \\ =& \tfrac{1}{2}(\mathrm{Cov}[f_a^p, f_a^q] + \mathrm{Cov}[f_b^p, f_b^q]) \\ & + \tfrac{1}{4}(\langle f_a^p \rangle - \langle f_b^p \rangle)(\langle f_a^q \rangle - \langle f_b^q \rangle). \end{aligned} \tag{4.17}$$

Thus in the combined set, the central PDF is the mean of the central PDFs in each set (up to the usual statistical uncertainties of order $1/\sqrt{N}$), whereas the uncertainties are always greater than the mean of the uncertainties, the extra term being due to the spread of the central predictions. This is as it should be: when the sets used in the combination disagree, the uncertainty is increased.

The expressions Eq. (4.16) and Eq. (4.17) can be generalized straightforwardly to the combination of n PDF sets, by taking N/n replicas at random from each (always keeping $N \gg n$ of course): then

$$F^{p(r)} = \begin{cases} f_{a_1}^{p(r)} & \text{for } r = 1, \ldots, \frac{N}{n}; \\ f_{a_2}^{p(r)} & \text{for } r = \frac{N}{n} + 1, \ldots, 2\frac{N}{n}, \\ \vdots & \vdots \\ f_{a_n}^{p(r)} & \text{for } r = \frac{n-1}{n} N + 1, \ldots, N, \end{cases} \tag{4.18}$$

and

$$\begin{aligned} E[F^p] &= \tfrac{1}{n} \sum_a \langle f_a^p \rangle, \\ \mathrm{Cov}[F^p, F^q] &= \tfrac{1}{n} \sum_a \mathrm{Cov}[f_a^p, f_a^q] + \tfrac{1}{n^2} \sum_{a \neq b} (\langle f_a^p \rangle - \langle f_b^p \rangle)(\langle f_a^q \rangle - \langle f_b^q \rangle). \end{aligned} \tag{4.19}$$

For large n the extra term in the covariance of the combination increases the result according to the covariance of the distribution of central values, since in the pairwise sum there are $n(n-1)$ terms.

The PDF4LHC15 prescription is based on the assumption that the PDF sets that are being combined, viewed as measurements of the true underlying PDF, are all equally likely, which means that they have approximately the same uncertainty and are approximately 100% correlated, i.e. they are not independent. Indeed, independent (uncorrelated or partly correlated) measurements of the same quantity a priori bring in new information on the true value, so the uncertainty on their combination is always smaller or equal to the uncertainty of any of the measurements that are being combined, (see e.g. Ref. [177]), as will be discussed further below.

In fact, an important property of the PDF4LHC15 combination prescription is that if the constituent PDF sets Φ_a are perfectly correlated, meaning that taking averages over replicas of the different PDF sets all give the same result, the combination also gives this result. Note that perfectly correlated sets will still be distinct, in the sense that the replicas will not be the same: it is only the averages over the full ensemble of replicas that are the same, in the limit when the number of replicas N becomes very large. An example are the various replica sets considered in Sect. 4.1.1, all based on the NNPDF4.0 methodology: both the two sets compared in Fig. 4.3, with replicas based on different underlying data replicas, and those compared in Fig. 4.4, based on correlated underlying data replicas. For these pairs of sets, $\langle f_a^p \rangle = \langle f_b^p \rangle$ for all a, b, and $\mathrm{Cov}[f_a^p, f_a^q] = \mathrm{Cov}[f_b^p, f_b^q]$. This is of course true irrespective of the number of sets n used in the combination: it is just the same as when combining several batches of PDF replicas from a given PDF set (such as NNPDF4.0) into a single larger replica set.

To ensure that these assumptions are reasonable, that is, the PDF sets used in the PDF4LHC15 combination are highly correlated, and as such can reasonably be combined by giving equal weight to each set, a number of criteria was adopted [76]:

- Each set is based on a global dataset, and in practice these global datasets are very similar, both in size and content.
- Theoretical calculations are performed to the same order in α_s, using a VFNS, and benchmarked against one another.
- External parameters such as α_s and quark masses are given common values where possible.
- Each PDF determination includes procedural and functional uncertainties in the adopted methodology.

Furthermore, an extensive benchmarking was performed in order to make sure that indeed uncertainties from the various sets were approximately equal, and that these criteria were sufficient to ensure that the PDF sets used in the combination could be be meaningfully assigned equal probability in the combination.

Correlated PDF combination

It is clear that even though the PDF sets included in the PDF4LHC15 combination are highly correlated, the correlation is manifestly not complete. Even assuming that the benchmarking and parameter settings can achieve complete agreement, there will still be some decorrelation through the choice of global dataset, and in the different methodologies used by the different groups. It has therefore been suggested [176] that a more precise and accurate result might be obtained if different PDFs are combined as independent, partly correlated measurements of the underlying true PDF. The logic is that, even in the presence of a common underlying dataset, each PDF determination, based on a different methodology, might be extracting different information from the data, just like different detectors could provide partly independent though correlated information on the same physical phenomenon. A correlated combination might then be advantageous because it would lead to a more precise and accurate prediction.

Unbiased correlated measurements of the same underlying observable can be combined in a standard way (see e.g. Sect. 7.6 of Ref. [177]). Specifically, viewing the expectation values $E[f_a^p]$ $a = 1, 2, \ldots$ of PDFs as measurements of an underlying true value, their correlated combination is a weighted average, that we can in turn view as the expectation value of the probability distribution for a combined determination $\tilde{F}^p$:

$$E[\tilde{F}^p] = \sum_a w_a^p E[f_a^p], \tag{4.20}$$

with weights w_a^p given by

$$w_a^p = \frac{\sum_b \mathrm{Cov}^{-1}[f_a^p, f_b^p]}{\sum_{c,d} \mathrm{Cov}^{-1}[f_c^p, f_d^p]}, \tag{4.21}$$

where $\text{Cov}^{-1}[f_a^p, f_b^p]$ is the matrix inverse of the S-covariance. The square uncertainty on the combination Eq. (4.20) is the variance of the probability distribution of $\tilde{F}^p$, given by

$$\text{Var}[\tilde{F}^p] = \sum_{a,b} w_a^p w_b^p \text{Cov}[f_a^p, f_b^p]. \tag{4.22}$$

Of course, all this relies on the assumption that the measurements are unbiased, and that the correlation between them, namely, the S-correlation Eq. (4.7), can be reliably computed.

Even so, the combination Eq. (4.20) is subject to several caveats. Specifically, the weights w_a^p Eq. (4.21) depend not only on p but also on x because the cross-correlation does (recall Eq. (4.4)), and consequently, the combined PDF Eq. (4.20) does not automatically satisfy sum rules. Furthermore, for the same reason, the result of the combination will generally depend on the scale at which it is performed, because with x-dependent weights even if the PDF sets f_a^p, for each a satisfy QCD evolution equations, the combination does not. Finally, in order to compute physical observables using the combined PDFs, knowledge of the diagonal uncertainty Eq. (4.22) is not sufficient: rather, the full cross-covariance matrix for all p, q and all x, y would be required. This could be done in principle by sampling the PDFs, computing their F-correlation, and then turning the result into a Hessian representation by using techniques similar to those of Ref. [183]. A way out of all these problems might be to perform the weighted combination Eq. (4.20) not at the level of PDFs, but rather at the level of physical observables. However, this has the further disadvantage that cross-covariances and weights would have to be re-computed for each new observable.

Be all that as it may, our goal here is not to investigate the most efficient way to implement the weighted combination, but rather, to explore the implications of performing a correlated weighted PDF combination according to Eqs. (4.20-4.22). As discussed above, in practice the PDF sets in the PDF4LHC15 combination have approximately equal uncertainties. When this is the case, the weights Eq. (4.21) are all approximately equal, and constant (independent of x), and then all the aforementioned problems can be ignored. Indeed, when we combine two PDF sets Φ_a and Φ_b such that $\text{Var}[f_a^p] = \text{Var}[f_b^p]$ the S-covariance is

$$\text{Cov}[f_a^p, f_b^p] = \big(\delta_{ab} + (1-\delta_{ab})\rho[f_a^p, f_b^p]\big)\text{Var}[f_a^p], \tag{4.23}$$

from which it follows, using Eq. (4.21), that $w_a = w_b = \frac{1}{2}$.

This equal weight situation can be very simply implemented in a Monte Carlo approach, in a completely equivalent way. Indeed, assuming that Monte Carlo replicas are available for two PDFs, Φ_a and Φ_b, the correlated combination is found by combining the two sets of replicas into a single replica set given by

$$\tilde{F}^{p(r)} = \tfrac{1}{2}(f_a^{p(r)} + f_b^{p(r)}) \tag{4.24}$$

for $r = 1, \ldots, N$. Then $E[\tilde{F}^p] = \langle F^{p(r)} \rangle = \frac{1}{2}(\langle f_a^{p(r)} \rangle + \langle f_b^{p(r)} \rangle)$, is in agreement with Eq. (4.20) when $w_a = w_b = \frac{1}{2}$. The F-covariance Eq. (4.17) evaluated over the replica set Eq. (4.24) is now given by[1]

$$\begin{aligned}\mathrm{Cov}[\tilde{F}^p, \tilde{F}^q] &= \langle \tilde{F}^p \tilde{F}^q \rangle - \langle \tilde{F}^p \rangle \langle \tilde{F}^q \rangle \\ &= \tfrac{1}{4}\left(\langle f_a^p f_a^q \rangle + \langle f_b^p f_b^q \rangle + \langle f_a^p f_b^q \rangle + \langle f_b^p f_a^q \rangle\right) \\ &\quad - \tfrac{1}{4}(\langle f_a^p \rangle + \langle f_b^p \rangle)(\langle f_a^q \rangle + \langle f_b^q \rangle) \\ &= \tfrac{1}{4}(\mathrm{Cov}[f_a^p, f_a^q] + \mathrm{Cov}[f_b^p, f_b^q]) + \tfrac{1}{4}(\mathrm{Cov}[f_a^p, f_b^q] + \mathrm{Cov}[f_b^p, f_a^q]).\end{aligned} \tag{4.25}$$

Note that when expressed in terms of the replicas from the original sets Φ_a and Φ_b the F-covariance between two PDFs in the combined set now depends on the cross-covariance between the corresponding PDFs of the original sets $\mathrm{Cov}[f_a^p, f_b^q]$ Eq. (4.4). Considering the diagonal case $p = q$ in Eq. (4.25), the variance of the PDFs of the combined set now depends on the S-correlation, and, using $\mathrm{Var}[f_a^p] = \mathrm{Var}[f_b^p]$, Eq. (4.25) with $p = q$ reduces to

$$\mathrm{Var}[\tilde{F}^p] = \tfrac{1}{2}\left(1 + \rho[f_a^p, f_b^p]\right)\mathrm{Var}[f_a^p], \tag{4.26}$$

which is the same as Eq. (4.23) when $a = b$.

Using the correlated Monte Carlo approach, the properties of the correlated combination are especially transparent. Specifically, it is clear that the uncertainty computed using Eq. (4.25) is always smaller than that found using the PDF4LHC15 combination. To see this, note that the correlation $|\rho[f_a, f_b]| \leq 1$ or equivalently $|\mathrm{Cov}[f_a, f_b]| \leq \sqrt{\mathrm{Var}[f_a]}\sqrt{\mathrm{Var}[f_b]}$. It follows that the square uncertainty on $F^p(x, Q_0^2)$ satisfies the inequality

$$\begin{aligned}\mathrm{Var}[\tilde{F}^p] &\leq \tfrac{1}{4}(\mathrm{Var}[f_a^p] + \mathrm{Var}[f_b^p]) + \tfrac{1}{2}\sqrt{\mathrm{Var}[f_a^p]}\sqrt{\mathrm{Var}[f_b^p]} \\ &\leq \tfrac{1}{2}(\mathrm{Var}[f_a^p] + \mathrm{Var}[f_b^p]) \\ &\leq \mathrm{Var}[F^p],\end{aligned} \tag{4.27}$$

where in going from the first to the second inequality we have trivially made use of the fact that $\frac{1}{2}(x + y)^2 \leq x^2 + y^2$, and the third inequality follows from the observation that the second term in Eq. (4.17) is non-negative.

The second inequality Eq. (4.27) has the obvious implication that, as already mentioned before, and seen explicitly from Eq. (4.26), whenever the correlation $\rho[f_a^p, f_b^p] < 1$ the uncertainty of the correlated combination $\tilde{F}^p$ is smaller than either of the uncertainties on f_a^p or f_b^p, that have been assumed to be approximately equal when forming the correlated combination according to Eq. (4.24), $\mathrm{Var}[f_a^p] \approx \mathrm{Var}[f_b^p]$.

The third inequality Eq. (4.27) has the perhaps less obvious implication that even if the two sets are fully correlated, so $\rho[f_a^p, f_b^p] = 1$, the uncertainty on the PDF4LHC15 combination F^p, can be larger that the uncertainty of either f_a^p or f_b^p, even though the uncertainty on the correlated combination $\tilde{F}^p$ is the same as that on both f_a^p and f_b^p. This happens whenever the central values of the two sets are not the same $\langle f_a^p \rangle \neq \langle f_b^p \rangle$.

[1] We note that Eq. (26) of the publication [77] that corresponds to Eq. (4.25) in this thesis contains a mistake in the second line, namely the parentheses are not placed correctly. The final result of line three is nevertheless correct in Ref. [77]

This is a situation that the combination formula Eq. (4.20) cannot accommodate. Indeed, the uncertainty of this correlated combination can never exceed that of the two determinations that are being combined. This follows from the assumption that the two determinations are unbiased estimators of the same underlying true value. Upon these assumptions, unit correlation means that the covariance matrix has a vanishing eigenvalue, so the two determinations have the same central value and uncertainty.

However, it is clearly possible to have two random variables that have unit correlation but do not have the same central value (or uncertainty). In particular the correlation of any two sets of random variables f_1^r and f_2^r is invariant under the linear transformations $f_1^r \to c_1 f_1^r + k_1$, $f_2^r \to c_2 f_2^r + k_2$, for any constants c_1, c_2, k_1, k_2, which change their mean values and variances. In the Bayesian combination one simply takes the point of view that the two measurements are equally likely determinations of the underlying true quantity, so a priori they might be fully correlated, and yet their mean values and variances might differ. In such a situation, the variance of the PDF4LHC15 combination always comes out larger than those of the determinations that are being combined. So, both the second and the third inequalities Eq. (4.27) become equalities only if PDF sets Φ_a and Φ_b are identical, i.e. they have the same central value, uncertainty, and unit correlation.

These results are easily generalized to the case of n PDF sets:

$$\begin{aligned} E[\tilde{F}^p] &= \frac{1}{n}\sum_a \langle f_a^p \rangle, \\ \mathrm{Cov}[\tilde{F}^p, \tilde{F}^q] &= \frac{1}{n^2}\sum_a \mathrm{Cov}[f_a^p, f_a^q] + \frac{1}{n^2}\sum_{a\neq b} \mathrm{Cov}[f_a^p, f_b^q]. \end{aligned} \tag{4.28}$$

In this case

$$\begin{aligned} \mathrm{Var}[\tilde{F}^p] &\leq \frac{1}{n^2}\sum_a \mathrm{Var}[f_a^p] + \frac{1}{n^2}\sum_{a\neq b} \sqrt{\mathrm{Var}[f_a^p]}\sqrt{\mathrm{Var}[f_b^p]} \\ &\leq \frac{1}{n}\sum_a \mathrm{Var}[f_a^p] \\ &\leq \mathrm{Var}[F^p], \end{aligned} \tag{4.29}$$

and again equality can only be achieved when there is complete equivalence between all the PDF sets in the combination, i.e. when $f_a^{p(r)} = f_b^{p(r)}$ for all r, p and for all pairs a, b. Otherwise, the correlated combination inevitably reduces uncertainties.

Even disregarding the issue related to PDF sets that have different central values despite being very highly correlated, the main problem with the reduction in uncertainty in the correlated combination is that it is reliable only if the cross-correlation has been correctly estimated. In particular, if the cross-correlation is underestimated, then the uncertainty on the combination is underestimated: in the most extreme case, in which two PDFs are fully correlated, but the cross-correlation is incorrectly determined to be very small, the uncertainty on the combination is underestimated by a factor $\sqrt{n}$ (assuming again the uncertainties of the n starting PDFs are approximately the same).

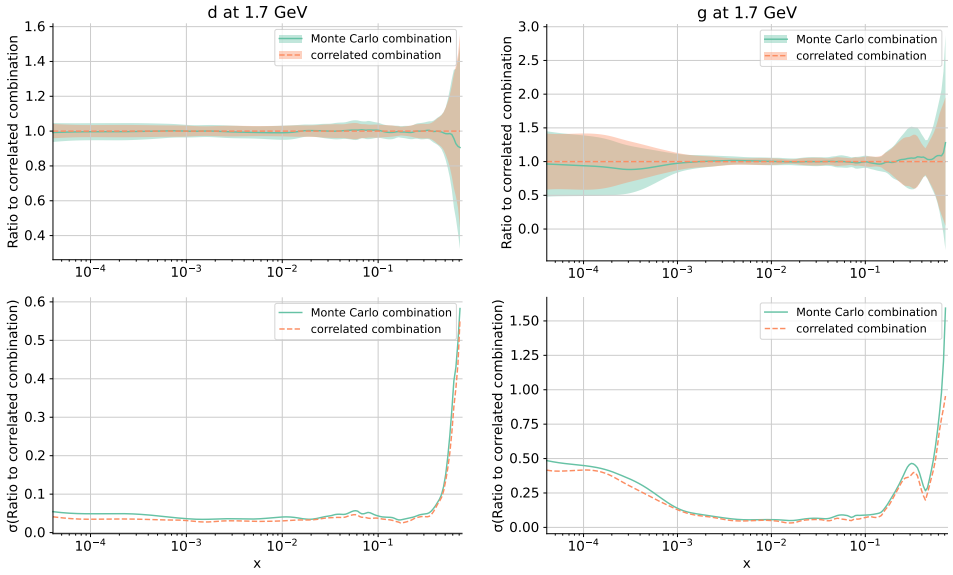

Figure 4.7: Check that the Monte Carlo combination Eq. (4.24) and the correlated combination Eq. (4.20) of PDF sets yield the same answer. Ten sets sets of 43 NNPDF4.0 PDF replicas are combined : 1) as the correlated weighted average Eq. (4.20) of the ten result from the ten sets determined using the data-driven component Eq. (4.11) of the S-correlation (correlated combination); or 2) as the average set of 43 replicas obtained using Eq. (4.28) from the ten replicas determined from each data replica (Monte Carlo combination). Results are shown for the down (left) and gluon (right) PDF. We show the PDFs normalized to the correlated weighted PDF (top), and the relative 1σ uncertainty (bottom).

In practice, the problem resides in the construction of the correlated replicas to be used combination Eq. (4.24): this ought to be done in such a way that the averages over replicas in Eq. (4.25) lead to a faithful determination of the F-covariance and the S-covariance. As we discussed before, if the sets of replicas $\{f_a^{p(r)}\}$, $\{f_b^{p(r)}\}$ that are being combined in Eq. (4.24) are randomly selected from the two sets, then the correlation vanishes regardless of its true value, see Eq. (4.13). If one selects replicas $f_a^{p(r)}$, $f_a^{p(r)}$ that are fitted to the same underlying data replica, then the S-correlation does not vanish, but it is generally underestimated because it only includes its data-driven component, as explicitly shown in Sect. 4.1.1.

The problem is especially severe when combining n different sets, because in this case underestimating the correlation between each pair of sets will lead to an increasingly large underestimation of the uncertainty on the combination as the number of sets increases. This is because in this case one is effectively assuming that the differences between the different determinations are due to each of them being a partly independent measurement, and as such doing more and more determinations reduces the uncertainty indefinitely.

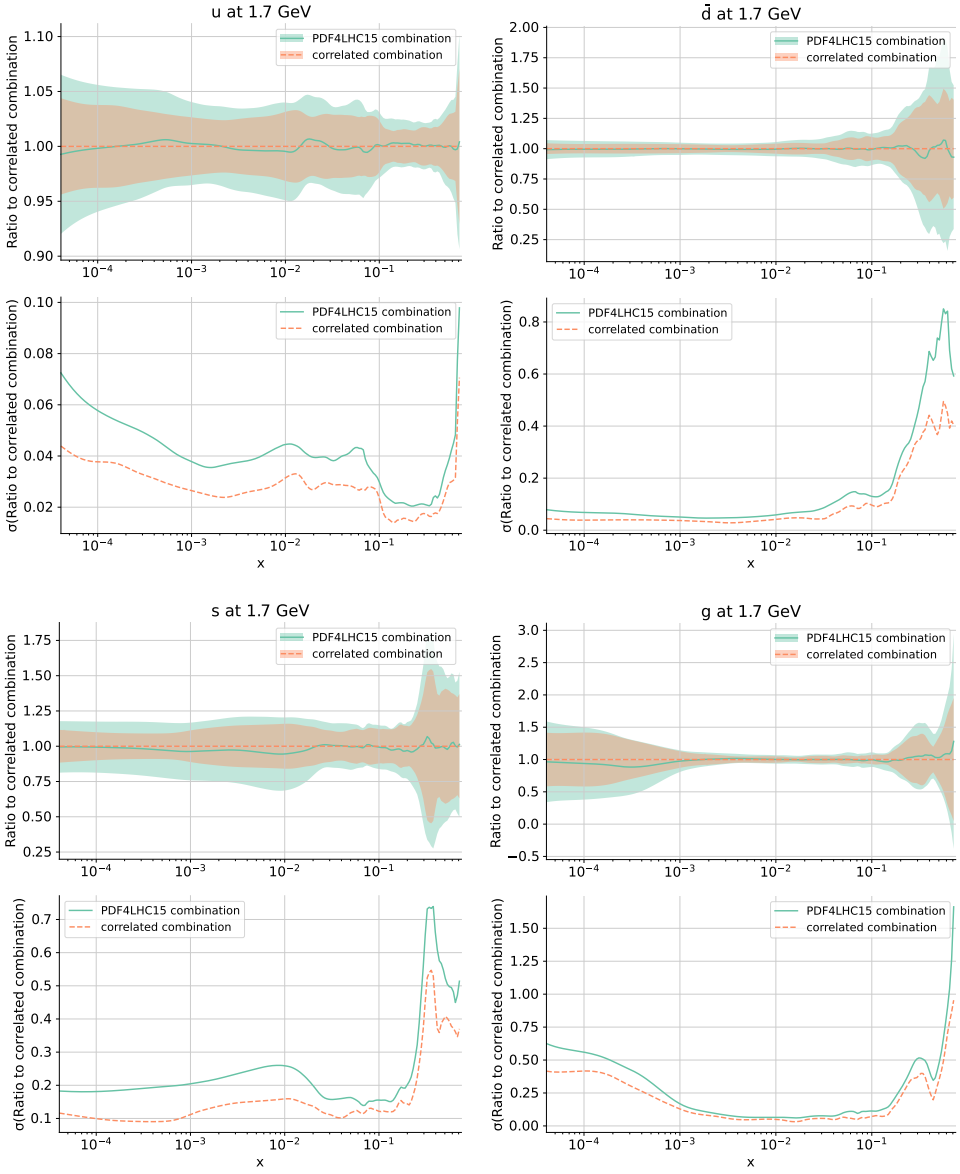

Figure 4.8: Comparison of the uncertainty on the correlated PDF combination and the PDF4LHC15 combination for 10 sets of 43 PDF NNPDF4.0 replicas. The correlated combination is obtained as the correlated weighted average of the ten results Eq. (4.20) (same as shown in Fig. 4.7); the PDF4LHC15 combination is found by simply combining all replicas in a single 430-replica set.

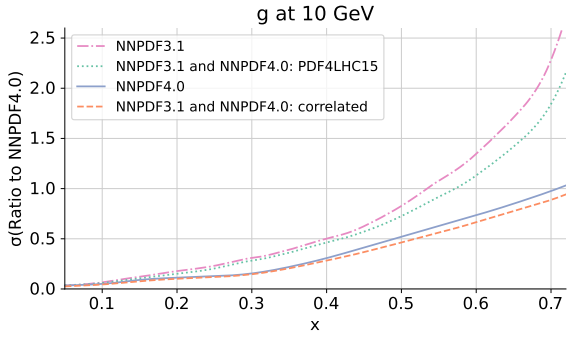

Figure 4.9: Comparison of the relative uncertainty on the large x gluon PDF determined using (from top to bottom) NNPDF3.1 methodology (purple, dot-dashed), NNPDF3.1 and NNPDF4.0 uncorrelated (PDF4LHC15) combination (green, dotted), NNPDF4.0 methodology (blue, solid), correlated combination (orange, dashes).

In order to expose the problem, we have considered an implementation of the combination Eq. (4.20). We have constructed ten sets of $N_{\rm rep}$ PDF replicas, all determined from the same $N_{\rm rep}$ underlying data replicas. In practice we take $N_{\rm rep} = 43$ because this is the largest number we got after applying the procedure discussed in Appendix B. We have then computed the ten by ten S-correlation matrix Eq. (4.11) for each PDF and each x value, and we have combined the ten sets using Eq. (4.20). We have explicitly checked that this is equivalent to instead using Eq. (4.28) to combine the ten sets in a single set with 43 replicas. This is demonstrated in Fig. 4.7 where two representative PDFs determined using either method are compared and seen to agree. This shows that the correlated Monte Carlo combination Eq. (4.28) is equivalent to the combination using the correlation matrix Eq. (4.20).

We have then compared this correlated combination to the PDF4LHC15 combination. The latter of course simply consists of putting together all replicas in a single 430 replica PDF set. Results are shown in Fig. 4.8. It is evident that while the PDF4LHC15 combination gives by construction the correct answer (since in this case the PDF is simply being combined to itself), the correlated combination leads to a rather smaller uncertainty. Clearly this is absurd. The reduction in uncertainty is the consequence of the fact that the S-correlation computed using Eq. (4.11) only includes the data-induced component. This underestimates the true correlation, because as we have seen in Sect. 4.1.1 (see in particular Fig. 4.4) it leads to a S-correlation which is rather lower than one, while in actual fact all these PDFs are fully correlated. The uncertainty reduction is amplified by having combined ten different sets.

We thus see that combining PDFs determined from the same underlying data as if they were correlated measurements leads to an incorrect answer because it neglects the fact that a sizable component of the PDF correlation is not data-driven. Indeed, if one did determine PDF uncertainties in this way, one would reach the paradoxical conclusion that PDF uncertainties can be made smaller at will by simply repeating many times the PDF determination with the same underlying data replicas.

Because of the difficulty in accurately estimating the non-data-driven component of the self-correlation, which is generally significant, this will be the generic scenario. As an especially striking example of this situation, in Fig. 4.9 we compare the relative uncertainty on the gluon PDF that we find if the gluon is determined using the NNPDF4.0 methodology, the NNPDF3.1 methodology, or the uncorrelated (PDF4LHC15) or correlated combination. As already discussed (see Fig. 4.2), the uncertainty found using the NNPDF3.1 methodology is rather larger than that found using the NNPDF4.0 methodology. Because, as also discussed, central values are very close, the uncertainty of the PDF4LHC15 combination, Eq. (4.17) is essentially the average of the uncertainties with the two methodologies. However, the uncertainty on the correlated combination is actually smaller than either of the uncertainties with the two methodologies that are being combined. One would thus reach the paradoxical conclusion that combining PDFs obtained with the more precise NNPDF4.0 methodology with the previous less precise NNPDF3.1 would actually lead to a reduction in uncertainty.

We thus conclude that a correlated combination inevitably leads to uncertainty underestimation and it cannot be considered as an alternative to the PDF4LHC15 combination, even though the latter might lead to uncertainties that are a little conservative.

4.2 Replica sampling for faithful PDF uncertainties

Having shown in the previous section that a large contribution to the PDF uncertainties does not correspond to the data uncertainties but rather has a methodological component, it is important to ensure that these uncertainties are not underestimated. In the NNPDF framework this is tested using closure tests [167] for uncertainties in the data region, and future tests [164] for uncertainties in the extrapolation region. These tests form only a small fraction of the NNPDF framework that has led to the NNPDF4.0 determination, the full framework in its current state has been constructed over a period of two decades and presented in dozens of papers. This spread of information over time and papers may have contributed to certain parts of the methodology being misunderstood [184]. Therefore let us use this section to elucidate certain features of the posterior sample of PDFs as constructed using the NNPDF4.0 framework.

A faithful determination of the PDFs is achieved by computing the probability measure in the functional space $\mathcal{F}$ of possible functions $f \in \mathcal{F}$ describing the PDFs at the initial scale. In this respect, PDF fitting is an example of inverse problem: the is aim to find a posterior probability of f given the data. As such, PDFs are defined, and can be extracted from data only within a well-defined theoretical framework (e.g. using perturbative QCD in a given scheme and at a given order of perturbation theory). These choices, together with any of the theoretical assumptions such as integrability, positivity and sum rules, determine the prior for the probability measure.

In the NNPDF approach, since the very beginning [68] the probability measure is represented by a Monte Carlo ensemble of functions $\{f_i : k = 1, ..., N_{\rm rep}\}$, which is obtained following the procedure discussed in Sect. 2.1.1. This ensemble of fitted functions yields a representation of the probability density in $\mathcal{F}$, which can be used to compute the probability distribution of any quantity that depends on the PDFs

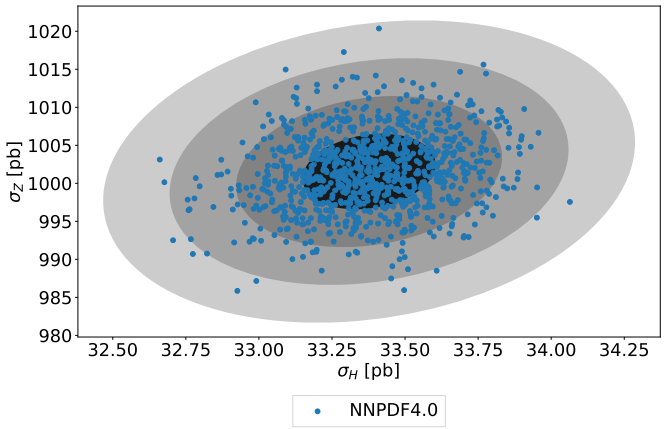

Figure 4.10: The NNPDF4.0 PDF replicas (1000 replica sample) in the σ_Z-σ_H plane (referred to as ZH in the main text). The one, two, three and four σ contours are shown.

using Eq. (2.1). It is crucial to realize that the Monte Carlo ensemble is not a random exploration, but rather an importance sampling of $\mathcal{F}$. As a result, the number of replicas needed to obtain a faithful representation in the PDF space does not require an exponentially growing number of replicas, with an exponent proportional to the number of parameters used to parametrize the function f, rather a Monte Carlo sample of $N_{\rm rep} \sim 1000$ is enough to reproduce the correlations of the experimental data to 1% accuracy, thus to determine the probability in $\mathcal{F}$ with the same accuracy, given a prior. The very same principle is at the basis of the PDF4LHC15 [76] and PDF4LHC21 [147] combinations.

We illustrate the outcome of such a random sampling by showing probability contours for the final replica distribution from the NNPDF4.0 PDF determination. Of course, a PDF set is a set of functions, so confidence levels for it should be shown in a space of functions [178], which is difficult to visualize. We can instead consider a projection of the PDF on a finite dimensional space. In order to make contact with the discussion in Ref. [184], we choose a two-dimensional space of LHC cross sections, that of the Z and Higgs total production cross section (ZH plane, henceforth). This is a useful choice in that the Higgs cross section is gluon driven and the Z cross section is quark-driven: so points in the ZH plane can be interpreted in terms of the size of the quark and gluon luminosities. Cross sections are computed as in Ref. [6]. In particular, partonic cross sections accurate to NLO in the strong coupling are convoluted with PDFs accurate to NNLO. A center-of-mass energy of 14 TeV is assumed, and cross sections are integrated in the fiducial phase space specified in Sect. 9.2 of Ref. [6]. Contours in this plane provide a test of the fact that the PDFs are correctly sampled, given that the cross sections depend on several different combinations of PDFs, evolved to the appropriate scale and convoluted over x with hard cross sections.

In Fig. 4.10 we show the results for 1000 NNPDF4.0 PDF replicas in the ZH plane, along with contours representing the standard deviations up to 4σ. The replicas are

distributed as expected given that the underlying distribution of the experimental uncertainties is Gaussian and it is expected (and has explicitly shown [185]) to lead to a Gaussian distribution of physical predictions. Indeed, more replicas are concentrated at the center and less in the tails, with no local distortions or clusters. For example, the two dimensional three σ contour corresponds to a 98.9% confidence level, so one would expect 11 NNPDF4.0 replicas outside the three σ contour, to be compared to 14 found in Fig. 4.10 in good agreement with the expected value. This result displays no evidence for sampling bias in the NNPDF4.0 replica sample and instead confirms that the replica sample is representative of the probability distribution in the ZH plane.

4.2.1 A lower χ^2 does not equal a more likely PDF

PDF fitting collaborations often use the experimental χ^2 as the optimization figure the best possible fit to the central value of the data. However, this statement needs to be better defined. Specifically, in the case of the NNPDF collaboration, the value of the χ^2 used during the minimization is the t_0 χ^2 introduced in Ref. [84], which is defined using the corresponding covariance matrix cov_{t_0} of Eq. (2.11) which serves to avoid the so-called D'Agostini bias.

In addition, as described in the discussion of the NNPDF4.0 methodology in Sect. 2.1.3, even though the χ^2 of the central value PDF (i.e., the average of the replicas) to the experimental data is often quoted, this value is never known to the fitting methodology (not the raw experimental χ^2 nor the t_0 χ^2). Instead, the fit uses for optimization the χ^2 to its own replica data and only that corresponding to the training set. This means the fitting methodology will take (per replica) the path that minimizes the χ^2 of the training set, however, it is not allowed to reach the absolute minimum of this quantity. In order to avoid overfitting, a small amount of data is hidden from the fit and used to validation its generalization power. The fit is then stopped when the validation χ^2 stops improving, regardless of the value of the training χ^2.

However, this still does not complete all the components of the target function minimized within the NNPDF approach. Namely, in order to include theoretical constraints in the optimization metric we use lagrange multipliers that are added to the training χ^2 and which are thus considered by the optimization algorithm.

It is commonly, though erroneously, understood that expected value of the PDFs as obtained by summing over replicas corresponds to the absolute minimum of the χ^2 while all deviations from the corresponding central PDF should result in an increase of the χ^2. As explained above, this is not enforced in any way. Instead, the randomly sampled PDF replicas may result in a solution closer to the underlying law from which they are sampled than the central PDF, though this is a low probability event.

In the case of the NNPDF4.0 determination, a baseline PDF set consisting of 1000 replicas has been released, and in this case each of the individual replicas corresponds to a larger t_0 χ^2 to the central data than the t_0 χ^2 of the central PDF. This is shown explicitly in Fig. 4.11 where the t_0 χ^2 of each of the replicas is computed and plotted in a frequency histograms with the χ^2 of the central value PDFs instead marked as vertical lines. However, if we continue generating replicas, we can generate replicas which correspond to a lower t_0 χ^2 than the central value of the set of replicas, in

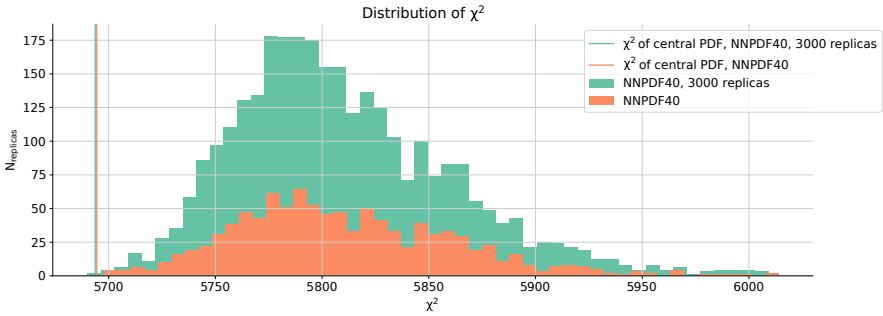

Figure 4.11: Histogram of the t_0 χ^2 for sets of 1000 (orange) and 3000 (green) NNPDF replicas. The corresponding vertical lines denote the t_0 χ^2 corresponding to the central PDF, i.e. the PDF obtained as an average over replicas. Note that the χ^2 is the total χ^2 for all 4618 datapoints.

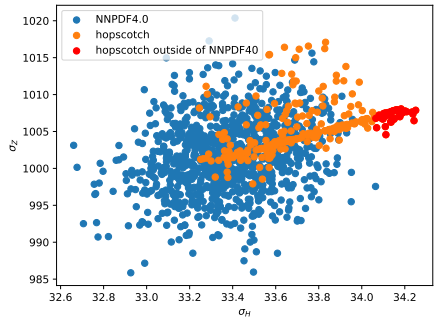

Figure 4.12: Scatter plot in the $\sigma_Z - \sigma_H$ plane including 1000 replicas from NNPDF4.0 (blue), 300 hopscotch functions with a t_0 χ^2 below the NNPDF4.0 t_0 χ^2 and the subset of 40 hopscotch functions (red) which fall outside of the space covered by the NNPDF4.0 replicas

Fig. 4.11 we show what happens if one continues fitting replicas until a set of 3000 replicas is obtained, in this case there are replicas which have a smaller χ^2.

An example where this misunderstanding becomes clear is in Ref. [184], in which candidate PDF functions, henceforth called hopscotch PDFs, have been generated by sampling along the eigenvector direction of the Hessian representation of the NNPDF4.0 1000 replica set. Of these hopscotch PDFs some have a marginally lower t_0 χ^2 to the central value of the datapoints than the NNPDF4.0 central PDF while, as can be seen in Fig. 4.12, simultaneously falling outside the range covered by the NNPDF4.0 predictions in the space of predictions of the inclusive Higgs cross-section. We will now explicitly show that in the Monte Carlo approach no such correlation exists between the training and validation χ^2 of a replica, and its distance from the central prediction of the Higgs cross-section.

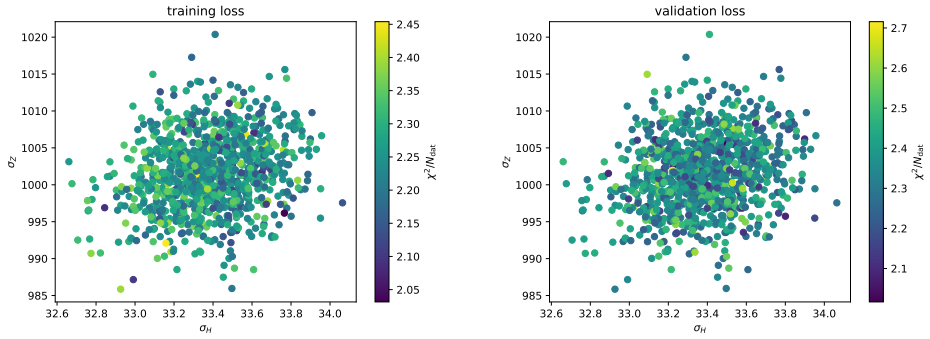

Figure 4.13: Scatter plot in the $\sigma_Z - \sigma_H$ plane for all NNPDF4.0 replicas showing the t_0 χ^2 to, the training (left) and validation (right) data used by each individual replica (i.e., including the Monte Carlo shift and training/validation masks).

The χ^2 values per replica for both the training and validation data for the NNPDF4.0 replicas are shown in Fig. 4.13. It is clear that there is no visible correlation between the position in the ZH plane, specifically the position along the σ_H axis, and the value of the χ^2. In fact, the fit quality of each PDF replica to its data replica is similar, and essentially independent of the position in the ZH plane. This means that outlier replicas are fitted equally well as replicas close to the center of the distribution. Outlier replicas simply correspond to unlikely data fluctuations. The NNPDF4.0 methodology has no difficulties in fitting PDFs that correspond to large (or small) values of the Higgs (or Z) cross-section.

To ensure that there is not some inherent inflexibility in the NNPDF methodology preventing it from generating replicas similar to the HS PDFs, we check explicitly that we can fit the HS PDFs if we assume them to be the underlying truth. To this purpose, we have performed a fit to level 0 closure test data (see App. D). Because the data are fitted at zero uncertainty, the fit can obtain vanishing χ^2. We have picked as an underlying true PDF the HS PDF that gives the largest Higgs cross-section. We have then fitted 100 PDF replicas to it, with standard methodology (including training-validation split). We find $\langle\chi\rangle_{\rm tr} = 0.03 \pm 0.01$, $\langle\chi\rangle_{\rm val} = 0.04 \pm 0.02$, so indeed we reach a near-prefect fit. A scatter plot of results in the ZH plane is shown in Fig. 4.14. It is clear that even though all the fits are equally good and fit the data perfectly (with zero uncertainty) there is still a distribution of results, due to the fact that of course the data do not determine the PDF completely. Each replica can be thought of as a different equally good interpolation of the given data, distributed in the ZH plane. Several of these results have values of σ_H that are in fact larger than the input underlying truth. So there is surely no "hard wall", and we must conclude that the NNPDF4.0 methodology has no difficulty in producing replicas with large Higgs cross section and/or of fitting the HS PDF exactly.

The hopscotch functions do however raise an interesting question: do we understand why they are given a low probability in the NNPDF4.0 determination? Here we touch upon the reason that neural networks are often called "black box" models. Namely, while a neural network is a universal approximator, studying the internal structure of

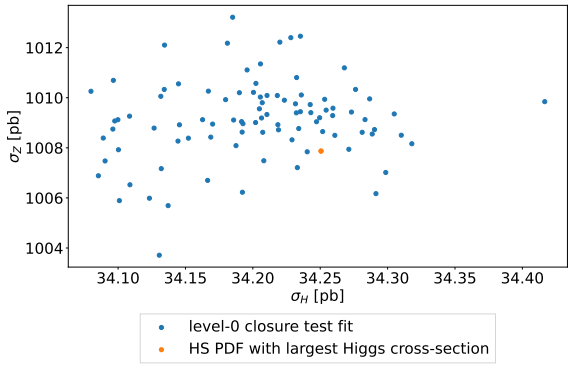

Figure 4.14: Scatter plot in the $\sigma_h - \sigma_Z$ plane of PDF replicas determined in level-0 closure test with the HS PDF with largest Higgs cross-section taken as underlying truth, shown as an orange point.

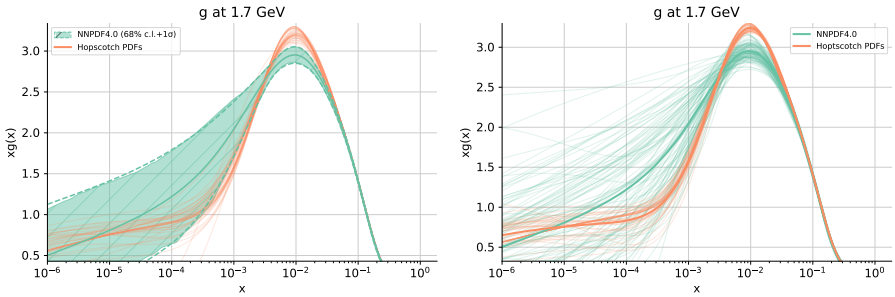

Figure 4.15: The gluon for HS PDFs with low χ values (orange) compared to the NNPDF4.0 gluon (green) 68%c.l. (left) and 1000 replica sample (right).

a neural network does not provide an insights into its the function it approximates. In principle, it is thus at the moment impossible to answer such questions as the one we are asking ourselves here. Nevertheless, we may aim to elucidate certain specific aspects of the black box.

4.2.2 Kinetic energy of the PDF

Having shown that the NNPDF4.0 methodology has no particular difficulty in fitting the HS PDF, we now address the question of why these PDFs are unlikely in the NNPDF methodology.

As a preliminary observation, we note that, given the extremely flexible NNPDF parametrization (with close to 800 free parameters), it is very easy to obtain fits with a value of χ that is much lower than that of the reference NNPDF4.0 determination. In particular since it should be kep in mind that the NNPDF4.0 methodology is carefully tuned with the aim of avoiding overfitting without while still optimally extracting information from the data. As a result, if the methodology were to be made slightly

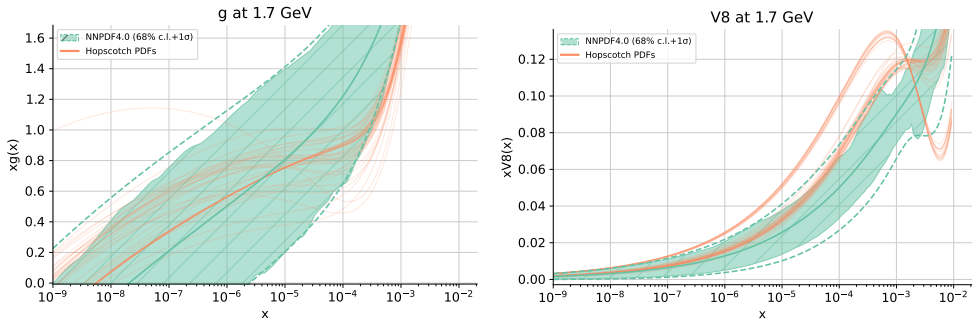

Figure 4.16: Left: same as the left plot of Fig. 4.15, now showing a detail of the small-x region. Right: the same as the left plot, but now for the valence octet PDF.

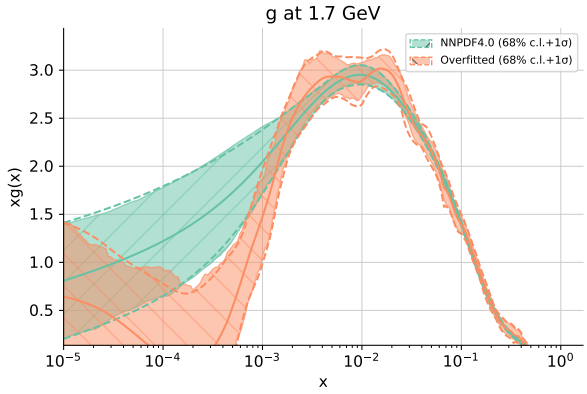

Figure 4.17: The gluon obtained in a overfitted PDF determination, in which the final χ value is by about 0.8 smaller than that of the default NNPDF4.0.

more aggressive – such as was done to generate the HS PDFs – this is is likely to result in overfitting.

As an example of such an overfitted PDF, in Fig. 4.17 we show the gluon distribution obtained in a fit in which we have artificially modified the minimization procedure in order to obtain a very low value of χ. Indeed, in this overfitted result, the final value of the χ of the central PDF to the experimental data is by about $\delta = 0.08$ smaller than that of the default NNPDF4.0, i.e. $\delta \times N_{\rm dat} \approx 300$, a difference that is about one order of magnitude bigger than that of the HS PDFs. The unphysical behavior of the PDFs thus obtained is manifest and representative of overfitting.

In light of this observation, in Fig. 4.15 we compare the NNPDF4.0 PDF gluon PDF to the HS PDFs: for NNPDF we show both the central value and 1σ uncertainty (left plot) and the corresponding replica set (right), while for HS we can only show the set of individual PDFs since their ensemble has no statistical meaning. It is apparent that the HS PDFs are characterized by a kink in the region $10^{-5} \lesssim x \lesssim 10^{-3}$. This kink is absent both in the central NNPDF4.0 gluon, and also in individual replicas

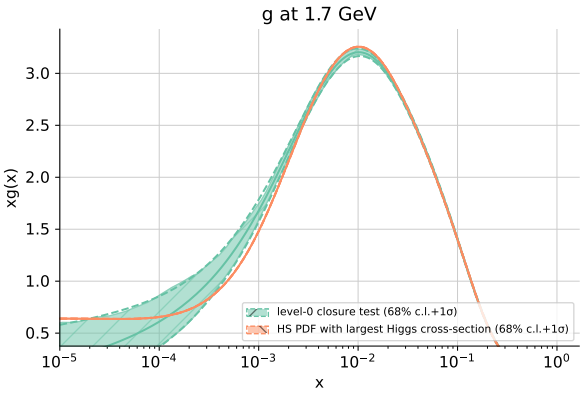

Figure 4.18: The gluon PDF obtained from the level-0 PDF replicas of Fig. 4.14 (green), compared to the underlying assumed truth, namely the HS PDF with largest Higgs cross-section taken as underlying truth (orange).

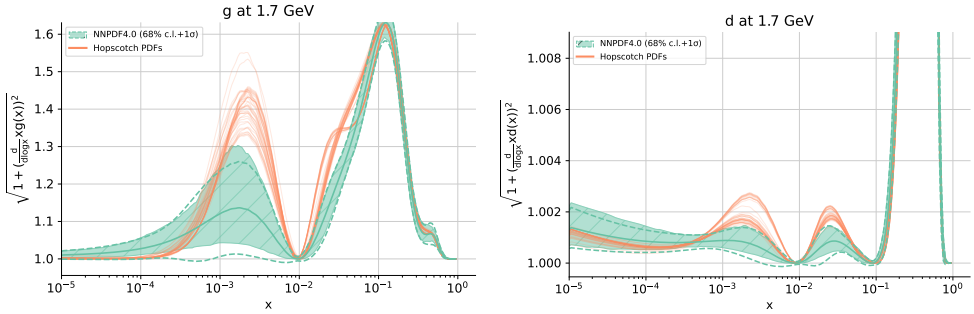

Figure 4.19: The kinetic energy Eq. (4.30) for the gluon (left) and down PDFs (right), for NNPDF4.0 (green) and for the HS PDFs (orange).

In Fig. 4.16 we show a detail of the small x region, in which the kink is clearly visible. For comparison, in Fig. 4.16 we also show a similar comparison for the octet valence combination ($V_8 = u^- + d^- - 2s^-$, with $q_i^- = q - \bar{q}$) in which even more pronounced kinks are seen in the HS PDFs.

It is important to observe that there are essentially no data constraining the gluon in the region of $x \lesssim 10^{-4}$, so the kink displayed by the HS PDFs is likely not data-driven. We can actually prove this explicitly by looking at the PDFs obtained in the fit to the level 0 closure test data of Fig. 4.14. Recall that this PDF produces a perfect fit to the data, i.e. it has vanishing χ^2. In Fig. 4.18 the gluon PDF from this set is shown and compared to the underlying assumed truth HS PDF. Even though the assumed truth has the kink that characterizes the HS PDFs, a perfect fit to data as predicted by a HS PDF has no kink. This indicates that the HS kink is not data driven, but rather an overfitted feature.

The overfitting metric proposed in Sect. 3.2.2 has the limitation that it can determine whether a given methodology is overfitted, but it cannot be applied to individual PDFs.

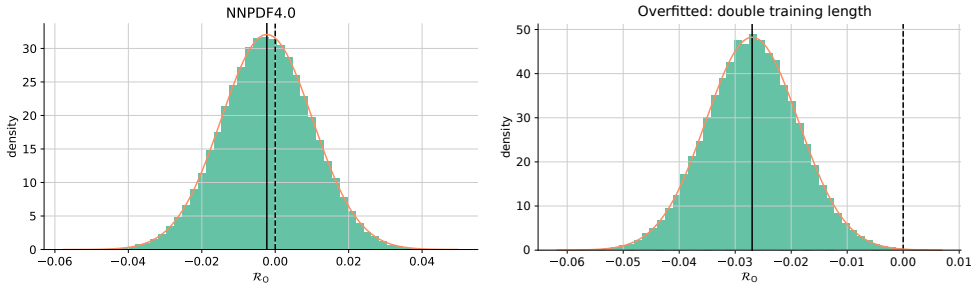

Figure 4.20: The overfitting metric Eq. (3.8) for the default NNPDF4.0 fit and for the artificially overfitted variant discussed in text (right).

As an alternative, we can construct a quantitative overfitting estimator by defining the PDF kinetic energy

$$\mathrm{KE} = \sqrt{1 + \left(\frac{d}{d\log x} x f(x, Q^2)\right)^2}\,. \tag{4.30}$$

This quantity, integrated between any two values of x gives the arclength of the curve that $xf(x)$ traverses, viewed as a function of $\log x$. The kinetic energy is thus a local measure of "wiggliness": given a pair of curves with fixed extremes, the one with greater kinetic energy joins the two extremes with a longer curve. It coincides of course with the Lagrangian of a relativistic free particle (with xf interpreted as space and $\log x$ as time), hence its name, with its integral being equal to the action.

In Fig. 4.19 we compare the kinetic energy of the HS PDFs to that of NNPDF4.0: we show the gluon and also, for reference, the down quark. It is clear that the HS PDFs are characterized by higher kinetic energy, specifically for the gluon, but in fact for all HS PDFs. Furthermore, the kinetic energy itself displays greater fluctuations for the HS PDFs. We conclude that the HS PDFs are characterized by having a feature which is not data driven and that corresponds to the given curve being further away from a least-action geodesic, which is disfavored by the NNPDF methodology.

The fact that the HS PDFs display signs of overfitting should not come as a surprise, given that they have been constructed by varying NNPDF4.0 replicas in such a way to further reduce the χ^2 to the central data, while the NNPDF4.0 replicas have been constructed in such a way as to marginally avoid overfitting. This suggests that PDF replicas with features similar to the HS PDFs could be obtained by forcing overfitting in the NNPDF4.0 methodology.

In order to check this explicitly, we have introduced overfitting in the NNPDF4.0 methodology by changing by hand the fit settings (which are set by the hyperoptimization procedure). In particular, we doubled the training length, which in turn implies increasing some parameters (such as the stopping patience) that are determined as functions of the training length. This leads to a decrease of χ^2 by about 0.01, i.e. similar to the greatest reduction observed in the HS PDFs. We can explicitly check that these PDFs are overfitted using the overfitting metric of Sect. 3.2.2. This metric vanishes for a proper fit, and it is negative for an overfit. Results are shown

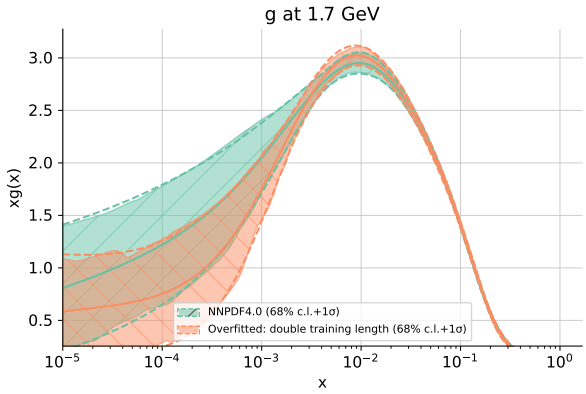

Figure 4.21: The gluon PDF obtained from the artificially overfitted variant discussed in text, compared to the default NNPDF4.0 gluon.

in Fig. 4.20, where we compare this metric for the default NNPDF4.0 PDFs and for this overfitted variant. We see that while for the default $R_O = -0.001 \pm 0.013$, for the overfitted variant $R_O = -0.027 \pm 0.001$, which indicates an overfit at the three σ level.

A comparison of the gluon PDF in this overfitted variant to the default NNPDF4.0 gluon in Fig. 4.21 shows that it starts developing features that are similar to that of the HS PDFs, specifically a kink in the small x region and a somewhat higher peak. This provides further evidence that the HS PDFs are overfitted.

Summarizing, we have shown that the HS gluon is characterized by a feature that is not data-driven and that corresponds to being further away from a least-action curve, and that similar features can be obtained in NNPDF4.0 replicas by forcing overfitting. We conclude that NNPDF4.0 replicas that look like the HS PDFs are disfavored by the NNPDF methodology because they correspond to overfitting solutions. Nevertheless, PDF replicas leading to results similar to the HS PDFs in the ZH plane (i.e. leading to similar values of the Higgs and Z cross section) can be obtained as proper fits to unlikely data fluctuations, given a large enough replica sample.

In this section we have thus seen a specific example of how a careful analysis using well-understood metrics can lead to a partial understanding of what is going on inside the black box that is a neural network. To obtain full insight into how prior assumptions lead to the posterior distribution of a fitting procedure one would need to use an interpretable model [186]. Where interpretability can colloquially be described as "the collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g., classification or regression)" [187]. For neural networks this is still far away, and the topic of ongoing research, nevertheless other machine learning tools such as gaussian process regression may be considered as alternative to address this shortcoming of neural networks in PDF fits.

Chapter 5

The neural network approach for neutrino structure functions

The NNPDF approach used for the determination of unpolarized PDFs was first applied to DIS neutral current structure functions [68, 70], and its flexibility allows for it to be applied to the determination of various other physical quantities. To date, the neural network approach has been applied to the determination of helicity PDFs [188], nuclear PDFs [154], fragmentation functions [189], and most recently even a simultaneous determination of the unpolarized PDFs with Standard Model Wilson coefficients within the framework of Standard Model Effective Field Theory [153].

In this chapter, to emphasize the wide range of applicability of the NNPDF approach, we will present a determination of neutrino inelastic structure functions based on matching a neural network parametrization of structure function data with NNLO perturbative QCD calculations based on recent analyses of proton and nuclear parton distributions. In Sect. 5.1 we motivate the need for a modern determination of the neutrino structure functions at low-Q and present our strategy, in Sect. 5.2 we discuss the theoretical formalism of neutrino inelastic structure functions, and in Sect. 5.3 we present the fitting strategy.

5.1 Modelling neutrino structure functions at low-Q

The scattering of neutrinos and anti-neutrinos with proton and nuclear targets [190] can be expressed in terms of the neutrino-nucleus structure functions, $F_i^{\nu N}(x, Q^2)$, which are analogous to the electron-proton structure functions discussed in Chapter 1. The DIS component dominates the inclusive neutrino-nucleus scattering cross-section for neutrino energies with $E_\nu \gtrsim 1$ TeV, and subleading processes such as (in)elastic scattering off the photon field of nucleons, coherent scattering off the photon field of nuclei, and scattering on atomic electrons are also well-understood and implemented in public codes [191].

On the other hand, the description of neutrino structure functions for momentum transfers, Q, of a few GeV becomes complicated by the poor convergence of the perturbative expansion in α_s, higher-twist effects, heavy quark and target mass corrections, and nuclear effects for nuclear targets. Eventually, for momentum

transfers below the proton mass ($Q \lesssim 1$ GeV), the perturbative framework breaks down and structure functions cannot be expressed anymore in terms a factorized convolution of PDFs with hard-scattering partonic coefficient functions. For these reasons, neutrino-nucleus structure functions at low and intermediate values of Q cannot be evaluated within the perturbative QCD framework and hence alternative strategies must be adopted.

The phase space region where neutrino structure functions are probed for $Q \lesssim$ few GeV represents a sizable contribution to the total inclusive cross-section for neutrino energies E_ν up to several TeV. For instance, structure functions for momentum transfers between $Q = 1.64$ GeV and $Q = 2.2$ GeV amount to around 10% of the inclusive cross-section for $E_\nu \simeq 100$ GeV [192]. In this low-Q region, one can distinguish between quasi-elastic scattering, for final-state invariant masses defined in Eq. (1.16), of $W < 1.07$ GeV, the resonance region, with 1.1 GeV $\leq W \leq 1.8$ GeV, and the inelastic region for $W \gtrsim 1.8$ GeV, with the latter being the main subject of the work discussed in this chapter. Given that perturbative QCD calculations cannot be applied, a widely used strategy to model low-Q region inelastic neutrino structure functions is the phenomenological Bodek-Yang (BY) model [193–198]. The BY approach is based on evaluating structure functions in terms of effective leading-order PDFs based on GRV98 determination [199], suitably extended with scaling variables for mass effects and with K-factor rescaling to improve the agreement with experimental data. The BY model is for instance the default option for inelastic structure function in the `GENIE` neutrino Monte Carlo event generator [200, 201].

While reasonably successful in describing neutrino and electron inelastic scattering data, the phenomenological BY model introduces a potential source of theoretical uncertainty in the calculations of inclusive neutrino-nucleus scattering whose magnitude is difficult to assess. First of all, the BY model is based on an obsolete set of proton PDFs that ignores all experimental constraints provided in the last 25 years, assumes an approximate treatment of quark and target mass corrections, ignores QCD effects beyond those present at the Born level, does not account for recent progress in our understanding of nuclear structure, and finally lacks a systematic assessment of the associated model uncertainties. A consequence of this restrictive character is that the BY model structure functions cannot be smoothly matched to state-of-the-art calculations based on modern PDFs and higher-order QCD calculations. Improving the description of these low-Q neutrino structure functions would hence strengthen the theoretical interpretation of ongoing and future reactor, accelerator, and atmospheric neutrino experiments involving energies E_ν between a few GeV and a few TeV. For instance, low-Q neutrino structure functions will be most relevant both for the Deep Underground Neutrino Experiment (DUNE) [202], involving E_ν up to several GeV, and the LHC-based experiments Faserν [203], SND@LHC [204], and the Forward Physics Facility [205, 206], where the E_ν peaks at a few hundreds of GeV.

We apply a strategy based on the NNPDF approach to the fitting of PDFs to describe inelastic neutrino structure functions valid for all values of the momentum transfer Q. It combines a neural network based parametrization of neutrino structure functions at low and moderate Q^2 values matched to NNLO perturbative QCD calculations at large Q^2. For the former, we perform a fit to all available measurements on low- and intermediate-Q neutrino structure functions, such that the dependence of $F_i^{\nu N}(x, Q, A)$, with $i = 2, 3, L$, on x, Q, and A is entirely determined from the data,

and where uncertainties are estimated and propagated with the Monte Carlo replica method discussed in Sect. 2.1. For the latter, NNLO QCD predictions for neutrino structure functions are computed with `YADISM` [207] using as input NNPDF4.0 and the nNNPDF3.0 [154] global determinations of the nuclear structure. In this matching procedure, theoretical predictions are weighted by their total uncertainty consisting on missing higher order (MHO) and PDF errors.

The outcome of this approach is a robust prediction for the neutrino structure functions $F_i(x, Q, A)$ valid in the whole range of Q with a faithful estimate of the uncertainties. These predictions can then be used to evaluate the inclusive neutrino cross-sections without the need to rely on ad-hoc models, hence providing a valuable input for the interpretation of measurements from a wide variety of neutrino experiments.

5.2 Theoretical formalism

We will now present an overview of the theoretical formalism underpinning neutrino-nucleus inelastic scattering cross-sections in terms of structure functions and of the calculation of the latter in perturbative QCD. The discussion here draws many similarities to that of Sect. 1.2, with an important difference that here we will consider the parity violating interaction giving rise to a third structure function. To improve readability of this chapter we will, where needed, repeat variables already defined in Sect. 1.2. We will not include a discussion of the calculations of neutrino structure functions based on YADISM, and limit the discussion to a description of the general framework needed to motivate restrictions implemented in the parametrization to be discussed in Sect. 5.3.3.

The double-differential cross-section for neutrino-nucleus scattering can be decomposed [208, 209] in terms of three independent structure functions $F_i^{\nu A}(x, Q^2)$ with $i = 1, 2, 3$. Focusing on the charged-current scattering case, mediated by the exchange of a W^+ weak boson, the differential cross-section reads

$$\begin{aligned}\frac{d^2\sigma^{\nu A}(x,Q^2,y)}{dxdy} =& \frac{G_F^2 s}{2\pi\left(1+Q^2/m_W^2\right)^2}\Bigg[(1-y)F_2^{\nu A}(x,Q^2) \\ &+ y^2 x F_1^{\nu A}(x,Q^2) + y\left(1-\frac{y}{2}\right)xF_3^{\nu A}(x,Q^2)\Bigg],\end{aligned} \tag{5.1}$$

where $s = 2m_N E_\nu$ is the neutrino-nucleon center of mass energy squared, m_N is the nucleon mass, E_ν is the incoming neutrino energy, and the inelasticity y is defined as $y = Q^2/(xs)$. An analogous expression holds for antineutrino scattering, mediated by the exchange of a W^- weak boson, with the only difference being a sign change in front of the parity-violating structure function xF_3,

$$\begin{aligned}\frac{d^2\sigma^{\bar{\nu} A}(x,Q^2,y)}{dxdy} =& \frac{G_F^2 s}{2\pi\left(1+Q^2/m_W^2\right)^2}\Bigg[(1-y)F_2^{\bar{\nu} A}(x,Q^2) \\ &+ y^2 x F_1^{\bar{\nu} A}(x,Q^2) - y\left(1-\frac{y}{2}\right)xF_3^{\bar{\nu} A}(x,Q^2)\Bigg].\end{aligned} \tag{5.2}$$

While the differential cross-section depends on three kinematic variables, (x, Q^2, y), the structure functions themselves depend only on x and Q^2. Furthermore, both the cross-section and the structure functions depend on the atomic mass number A of the target nucleus via nuclear modifications of the free-nucleon structure functions.

Alternatively, Eq. (5.1) can be expressed in terms of the longitudinal structure function $F_L^{\nu A}(x, Q^2)$ defined by $F_L = F_2 - 2xF_1$ given in Eq. (1.29), leading to

$$\begin{aligned}\frac{d^2\sigma^{\nu A}(x,Q^2,y)}{dxdy} =& \frac{G_F^2 s}{4\pi\left(1+Q^2/m_W^2\right)^2}\Big[Y_+ F_2^{\nu A}(x,Q^2) \\ &- y^2 F_L^{\nu A}(x,Q^2) + Y_- x F_3^{\nu A}(x,Q^2)\Big],\end{aligned} \tag{5.3}$$

where $Y_\pm = 1 \pm (1-y)^2$ and with a counterpart expression for anti-neutrino scattering,

$$\begin{aligned}\frac{d^2\sigma^{\bar{\nu} A}(x,Q^2,y)}{dxdy} =& \frac{G_F^2 s}{4\pi\left(1+Q^2/m_W^2\right)^2}\Big[Y_+ F_2^{\bar{\nu} A}(x,Q^2) \\ &- y^2 F_L^{\bar{\nu} A}(x,Q^2) - Y_- x F_3^{\bar{\nu} A}(x,Q^2)\Big],\end{aligned} \tag{5.4}$$

Expressing the differential cross-section as in Eq. (5.3) is advantageous because remember from Eq. (1.33) that in the parton model (and in perturbative QCD at leading order) the longitudinal structure function vanishes, $F_L^{\nu A}(x, Q^2) = 0$. The combination of neutrino and antineutrino measurements makes possible disentangling the different structure functions, for example the cross-section difference

$$\begin{aligned}\frac{d^2\sigma^{\nu A}(x,Q^2,y)}{dxdy} - \frac{d^2\sigma^{\bar{\nu} A}(x,Q^2,y)}{dxdy} =& \frac{G_F^2 s Y_-}{4\pi\left(1+Q^2/m_W^2\right)^2} \\ &\times \left[x F_3^{\nu A}(x,Q^2) + x F_3^{\bar{\nu} A}(x,Q^2)\right],\end{aligned} \tag{5.5}$$

is proportional to the parity-violating structure function xF_3 averaged over neutrinos and antineutrinos. Furthermore, the kinematic considerations leading to Eq. (1.66) imply that these structure functions vanish in the elastic limit $x \to 1$.

Depending on the values of the momentum transfer squared Q^2 and of the hadronic final-state invariant mass W given by

$$W^2 = m_N^2 + Q^2\frac{(1-x)}{x}, \tag{5.6}$$

different processes contribute to the neutrino structure functions as depicted in Fig. 5.1. At the lowest values of Q^2 and W^2, quasi-elastic (QE) scattering dominates, e.g. $\nu + n \to \ell^- + p$, where the target nucleon changes without breaking up. As Q^2 is increased excited resonances (RES) can be produced, such as $\nu + n \to \ell^- + \Delta^+$, provided the final-state mass W is above the resonance mass. The Δ^+ baryon is a higher mass spin-excitation of the proton, it can decay as $\Delta^+ \to p + \pi^0$. Once Q^2 becomes large enough that target nucleon breaks up, $\nu + n \to \ell^- + X$, one enters the regime known as inelastic scattering which is the main focus of this work. In the

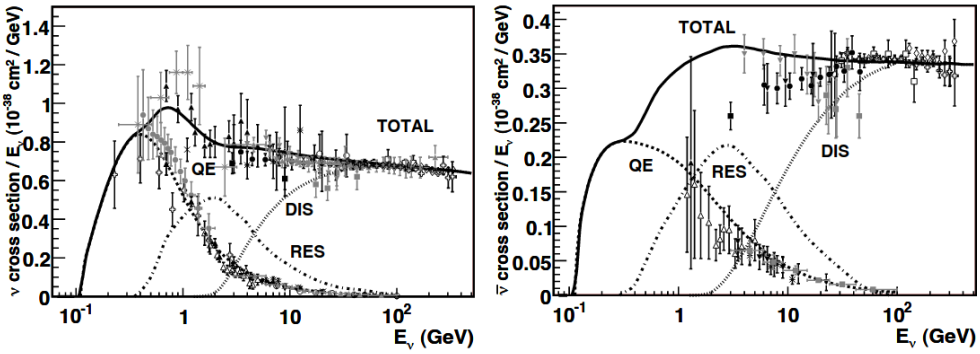

Figure 5.1: Total neutrino (left) and antineutrino (right) charged-current cross-section measurements [210] and corresponding predictions [211] as a function of the neutrino energy E_ν (solid). The various contributing processes are also shown independently and include quasi-elastic scattering (dashed), resonance production (dot-dash), and deep inelastic scattering (dotted). The figure is taken from Ref. [210].

following we assume that $W \gtrsim 2$ GeV to ensure that the inelastic scattering regime dominates.

For $Q^2 \gtrsim$ few GeV^2 we have access to the deep inelastic regime, where structure functions can be evaluated in perturbative QCD in terms of a factorised convolution of process-dependent partonic scattering cross-sections and of process-independent parton distribution functions as discussed in Sect. 1.2.2. The relevant factorization equation is Eq. (1.34), though here we also consider dependence on the atomic mass number A.

For massless quarks, charged-current neutrino DIS coefficient functions have been evaluated up to N3LO in [212, 213]. For massive quarks, the calculation of strange-to-charm transitions with charm mass effects has been performed at NNLO in [214]. Mass effects can be incorporated in the massless calculation by means of a general-mass variable-flavour-number scheme as described in Sect. 1.4. As we discuss below, perturbative corrections to neutrino DIS structure functions are moderate unless Q^2 approaches the boundary of the non-perturbative region, $Q^2 \simeq 1$ GeV^2.

A key feature of neutrino and anti-neutrino deep inelastic scattering is that each of the structure functions depends on a different combination of quark and antiquark PDFs, bringing in a unique sensitivity to quark flavour separation in nucleons and nuclei. At leading order and considering a proton target, four active quark flavours, neglecting charm mass effects, and assuming a diagonal CKM matrix, we can express the $F_2^{\nu p}$ and $xF_3^{\nu p}$ structure functions as follows

$$\begin{aligned} F_2^{\nu p} &= \quad 2x\left(\bar{u} + d + s + \bar{c}\right), \\ F_2^{\bar{\nu} p} &= \quad 2x\left(u + \bar{d} + \bar{s} + c\right), \\ xF_3^{\nu p} &= 2x\left(-\bar{u} + d + s - \bar{c}\right), \\ xF_3^{\bar{\nu} p} &= \quad 2x\left(u - \bar{d} - \bar{s} + c\right), \end{aligned} \tag{5.7}$$

where the dependence on x and Q have been suppressed. The corresponding expressions for a neutron target or isoscalar target can be obtained from isospin symmetry, for instance the LO structure functions for neutrino-neutron scattering are

$$\begin{aligned} F_2^{\nu n} &= 2x\left(\bar{d}+u+s+\bar{c}\right), \\ F_2^{\bar{\nu} n} &= 2x\left(d+\bar{u}+\bar{s}+c\right), \\ xF_3^{\nu n} &= 2x\left(-\bar{d}+u+s-\bar{c}\right), \\ xF_3^{\bar{\nu} n} &= 2x\left(d-\bar{u}-\bar{s}+c\right). \end{aligned} \tag{5.8}$$

From Eq. (5.7) and Eq. (5.8) we see how different combinations of neutrino structure functions provide access to different PDF combinations. For instance, the cross-section difference Eq. (5.5) for an isoscalar target without nuclear effects, so that $xF_3^{\nu A} = (xF_3^{\nu p} + xF_3^{\nu n})/2$, is given in this approximation by

$$\begin{aligned} \frac{d^2\sigma^{\nu A}(x,Q^2,y)}{dxdy} - \frac{d^2\sigma^{\bar{\nu} A}(x,Q^2,y)}{dxdy} = & \frac{G_F^2 s Y_-}{2\pi\left(1+Q^2/m_W^2\right)^2} \\ & \times \left[xu_V + xd_V + xs_V + xc_V\right], \end{aligned} \tag{5.9}$$

in terms of the valence combinations Eq. (1.45). The dependence of the valence PDFs on x and Q has been suppressed.

As opposed to the case of unpolarized PDF determinations, where one must impose the momentum and valence sum rules, structure functions in neutrino scattering do not need to satisfy all-orders sum rules. The exception is the parity-violating structure function xF_3 which enters the Gross-Llewellyn Smith (GLS) sum rule [215], calculable in perturbative QCD, and given for an isoscalar ($N = (p+n)/2$) target by

$$\int_0^1 \frac{dx}{x} xF_3^{\nu N}(x,Q^2) = 3\left(1+\sum_{k=1}^{3}\left(\frac{\alpha_s(Q^2)}{\pi}\right)^k c_k(n_f)\right), \tag{5.10}$$

where the coefficients c_k are known. The leading-order contribution to Eq. (5.10) follows from the partonic decomposition of the isoscalar $xF_3^{\nu N}$ in terms of the valence quark PDFs, as also indicated in Eq. (5.9). In this respect, the GLS sum rule is closely related to the valence sum rules of Eq. (1.63) and Eq. (1.64) imposed in PDF fits. While the Gross-Llewellyn Smith sum rule Eq. (5.10) will be satisfied by the QCD calculation, here we will not impose it explicitly in the data-driven parametrization but rather verify that it is satisfied within uncertainties a posteriori in the region of applicability of pQCD. We note that experimentally one cannot access the $x \to 0$ region and hence the evaluation of Eq. (5.10) depends on the modelling of the small-x extrapolation region.

5.3 Fitting methodology

Let us now describe the methodology used to determine the (anti)neutrino-nucleus inelastic structure functions and their associated uncertainties across the whole range of Q values relevant for the interpretation of present and future experiments. Our approach is based on the combination of the direct constraints provided by

experimental data at low- and intermediate-Q with those from the perturbative QCD calculations in the intermediate and high-Q regions. First, we present the general strategy indicating how the different components of the calculation are assembled. We then review the experimental data on neutrino structure functions and cross-sections used here to constrain the machine learning parametrization. Subsequently we introduce the main features of the ML algorithm, such as the choice of hyperparameters, the training procedure, and the matching to the perturbative QCD calculations.

5.3.1 General strategy

A schematic representation of the strategy adopted to describe the inelastic neutrino structure functions in the complete range of Q is presented in Fig. 5.2. It is based on classifying neutrino structure functions in three disjoint regions of Q and then evaluating them separately as follows:

Region I. At low momentum transfers $Q \lesssim Q_{\rm I}$, with $Q_{\rm I}$ of the order of a few GeV, the perturbative calculation of neutrino inelastic structure functions is either not valid (factorization breaks down) or is affected by large uncertainties related to e.g. highest twist, missing higher perturbative orders, and potential large-x resummation effects. In this region it is hence advantageous to directly parametrize the structure functions from existing experimental data on neutrino-nucleus scattering. Following the NNPDF methodology, this parametrization is achieved through a combination of neural networks with the Monte Carlo replica method for the uncertainty estimate. The experimental data that enters this parametrization is detailed in Sect. 5.3.2 and consists on all available measurements of inelastic neutrino-nucleus scattering.

Region II. The region of intermediate momentum transfers, defined by $Q_{\rm I} \lesssim Q \lesssim Q_{\rm II}$ where $Q_{\rm II}$ is of the order of a few tens of GeV, is well described by perturbative QCD calculations at NNLO which are provided by `YADISM` with the NNPDF4.0 and nNNPDF3.0 determinations of the proton and nuclear PDFs as inputs respectively. To ensure that the data-driven parametrization from region I is matched smoothly to this perturbative QCD calculation, level 2 pseudodata is generated using `YADISM` which is fitted in the same way as the experimental data in region I. For a description of level 2 pseudodata see App. D. Given that NNPDF4.0 and nNNPDF3.0 already include constraints from neutrino measurements, in particular from CHORUS and NuTeV, in region II no neutrino data needs to be added to the structure function parametrization.

Region III. In the region of high momentum transfers, $Q \gtrsim Q_{\rm II}$, the machine learning model predictions are replaced by those obtained from the `YADISM` perturbative QCD calculation, which already entered the fit as pseudo-data in region II. Hence in region III central prediction and uncertainties coincide with the `YADISM` outcome, with a mild smoothing procedure applied to eliminate residual discontinuities between this and region II. These `YADISM` predictions extend up to $Q = 10^6$ GeV, the highest values of Q relevant for phenomenology in particular for ultra-high-energy neutrino cross-sections.

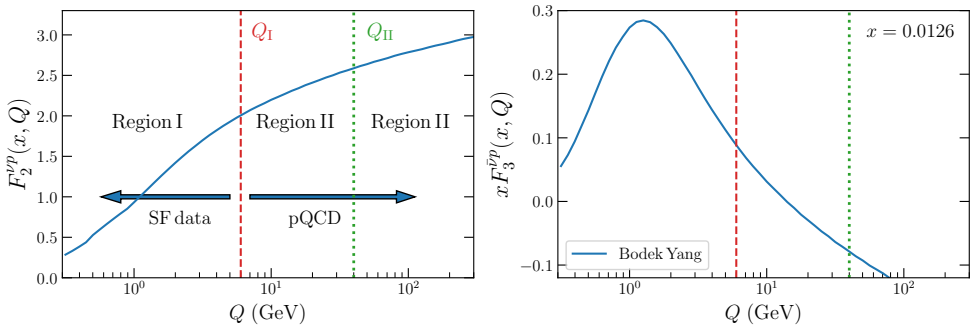

Figure 5.2: Schematic representation of the strategy adopted to parametrize the neutrino structure function as a function of Q, here represented by the Bodek-Yang calculation as implemented in `GENIE` for $x = 0.0126$.

In the schematic of Fig. 5.2 the neutrino structure functions are represented by the Bodek-Yang calculation as implemented in `GENIE`. Furthermore, the specific values of the hyperparameters $Q_{\rm I} = 6\,\text{GeV}$ and $Q_{\rm II} = 40\,\text{GeV}$ shown here, are representative of those used in the fit, and results for the structure function predictions are stable with respect to moderate variations of the chosen values. Nevertheless, the boundaries of the three regions should always satisfy the requirements stated above.

5.3.2 Experimental data

In order to parametrize the inelastic structure functions at low- and intermediate-Q values corresponding to region I in Fig. 5.2, we consider all available data on neutrino structure functions and double differential cross-sections available in the literature. We restrict ourselves to those measurements where the incoming neutrino energy E_ν is sufficiently large to ensure that the contribution from the inelastic region dominates. For this reason we do not consider here neutrino scattering measurements such as those provided the by ArgoNeuT [216], MicroBooNE [217], T2K [218], or MINERνa [219] experiments, since there the neutrino energy range is small enough such that inelastic scattering represents a subleading contribution to the scattering rates.

In Table 5.1 we display the datasets included in this work to parametrize the inelastic neutrino structure functions in region I. For each dataset we indicate the publication reference, the range of x covered, the smallest value of Q available, the observables measured, and the scattering target. A cut of $W \geq 2$ GeV is applied to restrict structure functions to the inelastic region, while data with $Q \geq Q_{\rm I}$ is excluded from the fit. Fig. 5.3 then displays the kinematic coverage in the (x, Q^2) plane of the measurements listed in Table 5.1.

5.3.3 Neural network parametrization

As explained in Sect. 5.3.1, in the low- and intermediate-Q region we construct a parametrization of the neutrino structure functions with its associated uncertainties by training a neural network model to available experimental data.

Datasets	Ref	$[x_{\min},\ x_{\max}]$	$Q_{\min}$ (GeV)	Observables	Target
BEBCWA59	[220]	$[0.028, 0.649]$	0.40	F_2, xF_3	Ne
CCFR	[221]	$[0.015, 0.650]$	1.12	F_2, xF_3	Fe
CHARM	[222]	$[0.015, 0.800]$	0.42	$F_2, xF_3, \bar{Q}$	$CaCO_3$
CHORUS	[223]	$[0.020, 0.650]$	0.57	$F_2, xF_3, \sigma_L/\sigma_T, d^2\sigma/(dxdy)$	Pb
CDHS	[224]			$F_2, d^2\sigma/(dxdy)$	D, Fe
CDHSW	[225]	$[0.015, 0.650]$	0.44	$F_1, F_2, xF_3, d^2\sigma/(dxdy)$	Fe
NUTEV	[226]	$[0.015, 0.750]$	0.44	$F_2, xF_3, d^2\sigma/(dxdy)$	Fe

Table 5.1: The datasets included in this work to parametrize the inelastic neutrino structure functions in region I defined in Fig. 5.2.

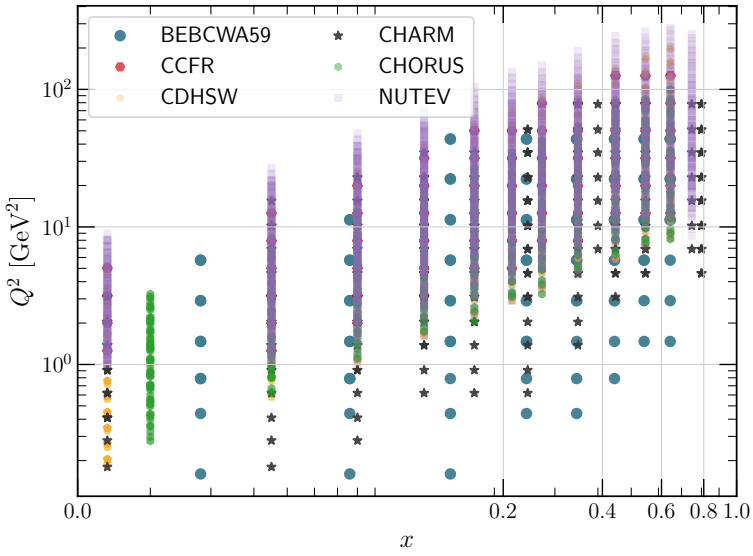

Figure 5.3: The kinematic coverage in the (x, Q^2) plane of the inelastic neutrino scattering cross-section data listed in listed in Table 5.1. A kinematic cut of $W \geq 2$ GeV is applied to restrict structure functions to the inelastic scattering region, while data with $Q \geq Q_I$ is excluded from the fit and replaced by perturbative QCD calculations.

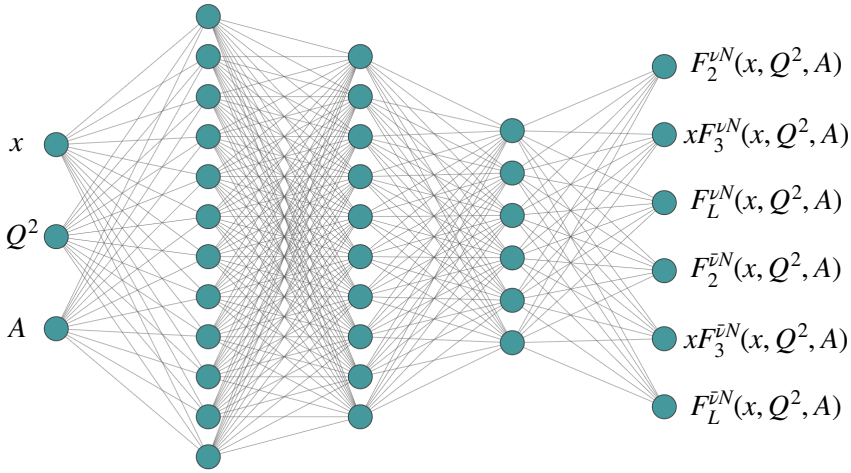

Figure 5.4: The neural network architecture used to parametrize the (anti-)neutrino-nucleus structure functions.

Given that the double-differential neutrino-nucleus cross-section can be expressed in terms of three structure functions, we need to parametrize six independent quantities from experimental data, namely F_2^ν, xF_3^ν, F_L^ν, $F_2^{\bar{\nu}}$, $xF_3^{\bar{\nu}}$, $F_L^{\bar{\nu}}$. For all we determine their dependence on x, Q^2 and A. Hence, we require a mapping from the inputs (x, Q^2, A) to the outputs F_i^ν and $F_i^{\bar{\nu}}$ with $i = 2, 3, L$. This mapping is provided by an artificial neural network with 3 input and 6 output neurons, and whose free parameters, the weights and biases, are adjusted to reproduce experimental data at low- and medium-Q (region I) and to the perturbative QCD predictions at medium- and high-Q (region II). Fig. 5.4 displays a representative example of the neural network architectures used in this work to parametrize neutrino structure functions. The architecture is 3-12-10-6-6, which means 3 hidden layers with 12, 10, and 6 neurons each. We note however that here, as for the NNPDF4.0 determination, the architecture is not fixed by hand, but rather automatically optimized together with other hyperparameters, such as the gradient descent algorithm and learning rates by means of the hyperoptimization procedure discussed in Sect. 2.1.4.

Our parametrization of structure functions implements the physical condition of vanishing in the elastic limit $x \to 1$. This is implemented in a way similar to that used in the feature scaling parametrization discussed earlier in this thesis, Eq. (3.2), by relating the output of the neural network to the structure functions as

$$F_i^\nu(x, Q^2, A) = \mathrm{NN}_i(x, Q^2, A) - \mathrm{NN}_i(x = 1, Q^2, A), \tag{5.11}$$

for all $i = 2, 3, L$, with NN_i the activation state of one of the output layer neurons of Fig. 5.4. The same is done for the anti-neutrino structure functions. This way, one ensures that by construction $F_i^\nu(x = 1, Q^2, A) = 0$ as required for any value of Q^2 and A. Furthermore, the inputs of the network are preprocessed by means of the feature scaling method discussed in Sect. 3.1 such that these inputs are read by the network in a way that maximizes its sensitivity. We note structure functions are not required

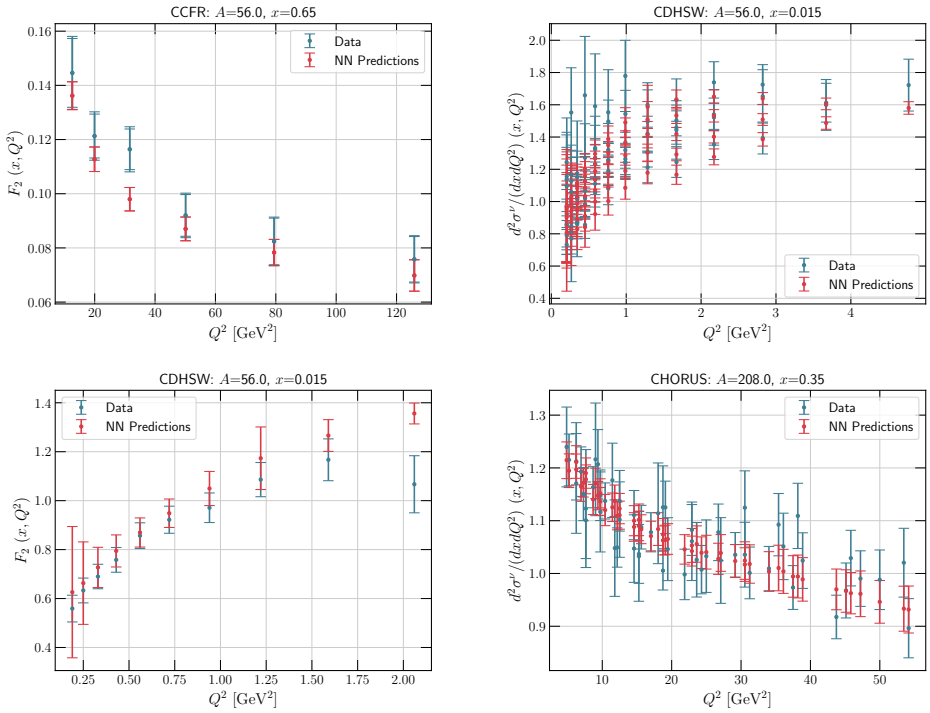

Figure 5.5: Representative comparisons between experimental data and the NN based parametrization of inelastic neutrino structure functions as a function of Q^2 For each panel the title indicates the dataset (which in turn specifies the corresponding physical observable) and the values of x and A selected for the comparison. The uncertainties in the prediction are obtained from the standard deviation over $N_{\rm rep} = 100$ Monte Carlo replicas.

to be positive-definite but it is found that negative configurations are excluded by the data.

The rest of the methodology is analogous to that used in the NNPDF4.0 determination as laid out in Sect. 2.1. Namely, the figure of merit that determines the goodness-of-fit of a given theory prediction when compared to the experimental data is the standard t_0 χ^2, error propagation is performed using the Monte Carlo replica method, and regularization is implemented through cross-validation.

Fig. 5.5 presents representative comparisons between experimental data and our parametrization of inelastic neutrino structure functions as a function of Q^2 For each panel the title indicates the dataset (which in turn specifies the corresponding physical observable) and the values of x and A selected for the comparison. The uncertainties in the predictions are obtained from the standard deviation over $N_{\rm rep} = 100$ Monte Carlo replicas.

The parameterization of the inelastic neutrino-structure functions presented here provide reliable predictions in the complete range of Q^2 values relevant for neutrino phenomenology, from very low to very high-Q^2 values. These results will be

implemented in the GENIE neutrino Monte Carlo event generator and as interpolation grids in the LHAPDF6 format. They will make possible the evaluation of the inclusive neutrino scattering cross-sections for a range of energy values without the need to rely on model assumptions.

Chapter 6

Summary

In this thesis we have presented various studies with as a central theme the NNPDF machine learning framework for the determination of parton distribution functions.

In Chapter 2 we discussed the most recent global determination of parton distribution functions by the NNPDF collaboration: NNPDF4.0. We first discussed the general strategy employed for this determination with a focus on the improved implementation as available through the open-source code. We then briefly reviewed the datasets included in this determination, while explicitly highlighting those datasets and processes that have for the first time been used in a PDF determination. We demonstrated that the results are stable upon the exclusion of those datasets which show the strongest signals of inconsistencies by performing a combination using the PDF4LHC15 prescription of various PDF determination each with a different dataset removed and show that the results remain the same up to statistical fluctuations. Afterwards, we discussed some of the main features of NNPDF4.0 relevant for phenomenology. Specifically, we independently assessed the impact on the luminosity of the changes to the methodology and the changes to the global dataset with respect to NNPDF3.1. We demonstrated that the qualitative improvements in the data result in an improved accuracy without affecting the uncertainty, while the updated methodology results in a more precise determination mainly as a consequence of the application of a gradient descent based algorithm. We closed the discussion of the main features of NNPDF4.0 by emphasizing how independent determination of the charm provides a strong indication for an intrinsic charm component in the proton. Finally, we have provided a brief overview of the functionality and possibilities of the open-source NNPDF code.

In Chapter 3 we proposed two main improvements to the methodology used for the NNPDF4.0 determination. First, we discussed the theoretical assumptions underpinning the parametrization of the PDFs in the NNPDF framework, which are enforced through a preprocessing factor and a scaling of the input x-grids. We demonstrated how a data-based scaling of x can avoid the problem of the inability of gradient descent based algorithms to efficiently learn features across multiple orders of magnitude without making assumptions on the scaling of the PDFs in x. We further demonstrated how this data based scaling allows us to remove the preprocessing exponents and by extension the replica-by-replica noise corresponding to the random sampling of the exponents as well as the iterative determination of their distribution of

the exponents. We then discussed how an important part of the hyperparamtrization procedure still requires human intervention for both the selection of the folds in k folds cross-validation, as well as the detection of overfitting in a PDF fit. For the former we have proposed a method for the automatic selection of the folds by training a neural network to learn which regions of the PDF (x, flavor) space are constrained by which datasets. For the latter we have presented a quantitative metric for the detection of overfitting by testing explicitly if the fit is fully agnostic to the validation dataset used during training, and thus ensuring the regularization as implemented through a training-validation split fully prevents overlearning.

In Chapter 4 we studied the methodological contribution to the total PDF uncertainties. First, we studied the correlation between different sets of PDFs, specifically, by viewing different PDF sets as distinct determinations, generally correlated, of the same underlying physical quantity. We examined the extent to which the correlation between them is due to the underlying data and show that correlations have a sizable component that is not due to the underlying data, because the data do not determine the PDFs uniquely. We then showed that the data-driven correlations can be used to assess the efficiency of methodologies used for PDF determination. Finally, we have also shown that the use of data-driven correlations for the combination of different PDFs into a joint set can lead to inconsistent results. In the second section of this chapter we addressed commonly misunderstood features related to sampling method of the NNPDF approach. In particular, we introduced the concept of the kinetic energy of the PDF, which has allowed us to explain certain aspects of the results provided by the trained neural network that are otherwise considered to be hidden in the black box model that is the neural network.

In Chapter 5 we have demonstrated the wide applicability of the NNPDF strategy to the determination of physical quantities other than PDFs. Specifically, we have presented a strategy for the determination of neutrino inelastic structure functions valid for the full range of the momentum transfer across a domain where different processes dominate the cross-section in different regions.

Appendix A

The NNPDF4.0 global dataset

Dataset	Ref.	N_{dat}	x	Q [GeV]
NMC F_2^d/F_2^p	[227]	260	[0.012, 0.680]	[2.1, 10.]
NMC $\sigma^{\mathrm{NC},p}$	[142]	292	[0.012, 0.500]	[1.8, 7.9]
SLAC F_2^p	[17]	211	[0.140, 0.550]	[1.9, 4.4]
SLAC F_2^d	[17]	211	[0.140, 0.550]	[1.9, 4.4]
BCDMS F_2^p	[18]	351	[0.070, 0.750]	[2.7, 15.]
BCDMS F_2^d	[18]	254	[0.070, 0.750]	[2.7, 15.]
CHORUS σ_{CC}^{ν}	[223]	607	[0.045, 0.650]	[1.9, 9.8]
CHORUS $\sigma_{CC}^{\bar{\nu}}$	[223]	607	[0.045, 0.650]	[1.9, 9.8]
NuTeV σ_{CC}^{ν} (dimuon)	[228, 229]	45	[0.020, 0.330]	[2.0, 11.]
NuTeV $\sigma_{CC}^{\bar{\nu}}$ (dimuon)	[228, 229]	45	[0.020, 0.210]	[1.9, 8.3]
[NOMAD $\mathcal{R}_{\mu\mu}(E_\nu)$] (*)	[105]	15	[0.030, 0.640]	[1.0, 28.]
[EMC F_2^c]	[230]	21	[0.014, 0.440]	[2.1, 8.8]
HERA I+II $\sigma^p_{\mathrm{NC},CC}$	[143]	1306	[$4{\cdot}10^{-5}$, 0.65]	[1.87, 223]
HERA I+II σ^c_{NC} (*)	[108]	52	[$7{\cdot}10^{-5}$, 0.05]	[2.2, 45]
HERA I+II σ^b_{NC} (*)	[108]	27	[$2{\cdot}10^{-4}$, 0.50]	[2.2, 45]

Table A.1: The DIS datasets analyzed in the NNPDF4.0 PDF determination. For each of them we indicate a description of the dataset, the corresponding reference, the number of data points in the fits before kinematic cuts (see Sect.4 of Ref. [6]), and the kinematic coverage in the relevant variables after cuts. Datasets not previously considered in NNPDF3.1 are indicated with an asterisk. Datasets not included in the baseline determination are indicated in square brackets. The Q coverage indicated for NOMAD is to be interpreted as an integration range.

Dataset	Ref.	N_{dat}	Q^2 [GeV^2]	p_T [GeV]
[ZEUS 820 (HQ) (1j)] **(*)**	[109]	30	[125,10000]	[8,100]
[ZEUS 920 (HQ) (1j)] **(*)**	[110]	30	[125,10000]	[8,100]
[H1 (LQ) (1j)] **(*)**	[112]	48	[5.5,80]	[4.5,50]
[H1 (HQ) (1j)] **(*)**	[113]	24	[150,15000]	[5,50]
[ZEUS 920 (HQ) (2j)] **(*)**	[111]	22	[125,20000]	[8,60]
[H1 (LQ) (2j)] **(*)**	[112]	48	[5.5,80]	[5,50]
[H1 (HQ) (2j)] **(*)**	[113]	24	[150,15000]	[7,50]

Table A.2: Same as Table A.1 for DIS jet data.

Dataset	Ref.	N_{dat}	$y_{\ell\ell}$	$m_{\ell\ell}$ [GeV]
E866 $\sigma^d/2\sigma^p$ (NuSea)	[231]	15	[0.07, 1.53]	[4.60, 12.9]
E866 σ^p (NuSea)	[144]	184	[0.00, 1.36]	[4.50, 8.50]
E605 σ^p	[232]	119	[-0.20, 0.40]	[7.10, 10.9]
E906 $\sigma^d/2\sigma^p$ (SeaQuest) **(*)**	[114]	6	[0.11, 0.77]	[4.71, 6.36]

Table A.3: Same as Table A.1 for fixed-target DY data.

Dataset	Ref.	N_{dat}	Kin$_1$	Kin$_2$ [GeV]
CDF Z differential	[233]	29	$0.0 \leq y_{\ell\ell} \leq 2.9$	$66 \leq m_{\ell\ell} \leq 116$
D0 Z differential	[234]	28	$0.0 \leq y_{\ell\ell} \leq 2.8$	$66 \leq m_{\ell\ell} \leq 116$
[D0 W electron asymmetry]	[235]	13	$0.0 \leq y_e \leq 2.9$	$Q = m_W$
D0 W muon asymmetry	[236]	10	$0.0 \leq y_\mu \leq 1.9$	$Q = m_W$
ATLAS low-mass DY 7 TeV	[237]	6	$\lvert\eta_\ell\rvert \leq 2.1$	$14 \leq m_{\ell\ell} \leq 56$
ATLAS high-mass DY 7 TeV	[238]	13	$\lvert\eta_\ell\rvert \leq 2.1$	$116 \leq m_{\ell\ell} \leq 1500$
ATLAS W,Z 7 TeV ($\mathcal{L} = 35$ pb^{-1})	[239]	30	$\lvert\eta_\ell, y_Z\rvert \leq 3.2$	$Q = m_W, m_Z$
ATLAS W,Z 7 TeV ($\mathcal{L} = 4.6$ fb^{-1}) **(*)**	[115]	61	$\lvert\eta_\ell, y_Z\rvert \leq 2.5, 3.6$	$Q = m_W, m_Z$
CMS W electron asymmetry 7 TeV	[240]	11	$\lvert\eta_e\rvert \leq 2.4$	$Q = m_W$
CMS W muon asymmetry 7 TeV	[241]	11	$\lvert\eta_\mu\rvert \leq 2.4$	$Q = m_W$
CMS DY 2D 7 TeV	[242]	132	$\lvert\eta_{\ell\ell}\rvert \leq 2.2$	$20.0 \leq m_{\ell\ell} \leq 200$
LHCb $Z \to ee$ 7 TeV	[243]	9	$2.0 \leq \eta_\ell \leq 4.5$	$Q = m_Z$
LHCb $W,Z \to \mu$ 7 TeV	[244]	33	$2.0 \leq \eta_\ell \leq 4.5$	$Q = m_W$
[ATLAS W 8 TeV] **(*)**	[118]	22	$\lvert\eta_\ell\rvert < 2.4$	$Q = m_W$
ATLAS low-mass DY 2D 8 TeV **(*)**	[117]	84	$\lvert y_{\ell\ell}\rvert < 2.4$	$46 \leq m_{\ell\ell} \leq 200$
ATLAS high-mass DY 2D 8 TeV **(*)**	[116]	48	$\lvert y_{\ell\ell}\rvert < 2.4$	$116 \leq m_{\ell\ell} \leq 1500$
CMS W rapidity 8 TeV	[245]	22	$\lvert\eta_\ell\rvert \leq 2.3$	$Q = m_W$
LHCb $Z \to ee$ 8 TeV	[246]	17	$2.00 < \lvert\eta_e\rvert < 4.25$	$Q = m_Z$
LHCb $W,Z \to \mu$ 8 TeV	[247]	34	$2.00 < \lvert\eta_\mu\rvert < 4.25$	$Q = m_Z$
[LHCb $W \to e$ 8 TeV] **(*)**	[248]	8	$2.00 < \lvert\eta_e\rvert < 4.25$	$Q = m_W$
ATLAS $\sigma^{tot}_{W,Z}$ 13 TeV **(*)**	[119]	3	—	$Q = m_W, m_Z$
LHCb $Z \to ee$ 13 TeV **(*)**	[120]	17	$2.00 < \lvert y_Z\rvert < 4.25$	$Q = m_Z$
LHCb $Z \to \mu\mu$ 13 TeV **(*)**	[120]	18	$2.00 < \lvert y_Z\rvert < 4.50$	$Q = m_Z$

Table A.4: Same as Table A.1 for collider inclusive gauge boson production data.

Dataset	Ref.	N_{dat}	Kin_1	Kin_2 [GeV]
ATLAS $W^{\pm}+c$ 7 TeV **(*)**	[122]	22	$\lvert\eta_\ell\rvert < 2.5$	$Q = m_W$
CMS $W^{\pm}+c$ 7 TeV	[249]	10	$\lvert\eta_\ell\rvert < 2.1$	$Q = m_W$
CMS $W^{\pm}+c$ 13 TeV **(*)**	[123]	5	$\lvert\eta_\ell\rvert < 2.4$	$Q = m_W$
ATLAS $W^{\pm}$+jet 8 TeV **(*)**	[121]	32	$0 \leq p_T^W \leq 800$ GeV	$Q = m_W$
ATLAS Z p_T 8 TeV $(p_T, m_{\ell\ell})$	[250]	64	$12 \leq m_{\ell\ell} \leq 150$ GeV	$30 \leq p_T^Z \leq 900$
ATLAS Z p_T 8 TeV (p_T, y_Z)	[250]	120	$\lvert y_Z\rvert < 2.4$	$30 \leq p_T^Z \leq 150$
CMS Z p_T 8 TeV	[245]	50	$\lvert y_Z\rvert < 1.6$	$30 \leq p_T^Z \leq 170$
CMS σ_{tt}^{tot} 5 TeV **(*)**	[251]	1	—	$Q = m_t$
ATLAS σ_{tt}^{tot} 7, 8 TeV	[252]	2	—	$Q = m_t$
CMS σ_{tt}^{tot} 7, 8 TeV	[253]	2	—	$Q = m_t$
ATLAS σ_{tt}^{tot} 13 TeV ($\mathcal{L}$=139 fb^{-1}) **(*)**	[126]	1	—	$Q = m_t$
CMS σ_{tt}^{tot} 13 TeV	[254]	1	—	$Q = m_t$
[ATLAS $t\bar{t}$ ℓ+jets 8 TeV $(1/\sigma d\sigma/dp_T^t)$]	[255]	8	$0 \leq p_T^t \leq 500$ GeV	$Q = m_t$
ATLAS $t\bar{t}$ ℓ+jets 8 TeV $(1/\sigma d\sigma/dy_t)$	[255]	5	$\lvert y_t\rvert < 2.5$	$Q = m_t$
ATLAS $t\bar{t}$ ℓ+jets 8 TeV $(1/\sigma d\sigma/dy_{t\bar{t}})$	[255]	5	$\lvert y_{t\bar{t}}\rvert < 2.5$	$Q = m_t$
[ATLAS $t\bar{t}$ ℓ+jets 8 TeV $(1/\sigma d\sigma/dm_{t\bar{t}})$]	[255]	7	$345 \leq m_{t\bar{t}} \leq 1600$ GeV	$Q = m_t$
ATLAS $t\bar{t}$ 2ℓ 8 TeV $(1/\sigma d\sigma/dy_{t\bar{t}})$ **(*)**	[124]	5	$\lvert y_{t\bar{t}}\rvert < 2.8$	$Q = m_t$
CMS $t\bar{t}$ ℓ+jets 8 TeV $(1/\sigma d\sigma/dy_{t\bar{t}})$	[256]	10	$-2.5 < y_{t\bar{t}} < 2.5$	$Q = m_t$
CMS $t\bar{t}$ 2D 2ℓ 8 TeV $(1/\sigma d\sigma/dy_t dm_{t\bar{t}})$ **(*)**	[125]	16	$\lvert y_t\rvert < 2.5$	$340 \leq m_t \leq 1500$
CMS $t\bar{t}$ ℓ+jet 13 TeV $(d\sigma/dy_t)$ **(*)**	[127]	10	$\lvert y_t\rvert < 2.5$	$Q = m_t$
CMS $t\bar{t}$ 2ℓ 13 TeV $(d\sigma/dy_t)$ **(*)**	[128]	11	$\lvert y_t\rvert < 2.5$	$Q = m_t$
[ATLAS incl. jets 7 TeV, R=0.6]	[257]	90	$\lvert y^{jet}\rvert < 3.0$	$100 \leq p_T^{jet} \leq 1992$
[CMS incl. jets 7 TeV]	[258]	133	$\lvert y^{jet}\rvert < 2.5$	$100 \leq p_T^{jet} \leq 2000$
ATLAS incl. jets 8 TeV, R=0.6 **(*)**	[129]	171	$\lvert y^{jet}\rvert < 3.0$	$70 \leq p_T^{jet} \leq 2500$
CMS incl. jets 8 TeV **(*)**	[130]	185	$\lvert y^{jet}\rvert < 3.0$	$74 \leq p_T^{jet} \leq 2500$
ATLAS dijets 7 TeV, R=0.6 **(*)**	[131]	90	$0.0 \leq y^* \leq 3.0$	$260 \leq m_{jj} \leq 4270$
CMS dijets 7 TeV **(*)**	[132]	54	$\lvert y_{max}\rvert < 2.5$	$200 \leq m_{jj} \leq 5000$
[CMS 3D dijets 8 TeV] **(*)**	[133]	122	$0.0 < y_b, y^* < 3.0$	$133 \leq p_{T,avg} \leq 1780$
[ATLAS isolated γ prod. 8 TeV] **(*)**	[134]	49	$\lvert\eta_\gamma\rvert < 2.37$	$E_T^\gamma < 1500$
ATLAS isolated γ prod. 13 TeV **(*)**	[135]	53	$\lvert\eta_\gamma\rvert < 2.37$	$E_T^\gamma < 1500$
ATLAS single t R_t 7 TeV **(*)**	[136]	1	—	$Q = m_t$
CMS single t $\sigma_t + \sigma_{\bar{t}}$ 7 TeV **(*)**	[139]	1	—	$Q = m_t$
ATLAS single t R_t 8 TeV **(*)**	[137]	1	—	$Q = m_t$
CMS single t R_t 8 TeV **(*)**	[140]	1	—	$Q = m_t$
ATLAS single t R_t 13 TeV **(*)**	[138]	1	—	$Q = m_t$
CMS single t R_t 13 TeV **(*)**	[141]	1	—	$Q = m_t$
ATLAS single t 7 TeV $(1/\sigma d\sigma/dy_t)$ **(*)**	[136]	4	$\lvert y_t\rvert < 3.0$	$Q = m_t$
ATLAS single t 7 TeV $(1/\sigma d\sigma/dy_{\bar{t}})$ **(*)**	[136]	4	$\lvert y_{\bar{t}}\rvert < 3.0$	$Q = m_t$
ATLAS single t 8 TeV $(1/\sigma d\sigma/dy_t)$ **(*)**	[137]	4	$\lvert y_t\rvert < 2.2$	$Q = m_t$
ATLAS single t 8 TeV $(1/\sigma d\sigma/dy_{\bar{t}})$ **(*)**	[137]	4	$\lvert y_{\bar{t}}\rvert < 2.2$	$Q = m_t$

Table A.5: Same as Table A.1 for other LHC processes.

Appendix B

Computing cross-correlations in the NNPDF framework

We provide here some details on the computation of the cross-covariance Eq. (4.4) and S-covariance Eq. (4.11)using NNPDF methodology.

In the NNPDF methodology the data replicas are generated based on a Monte Carlo method with random initialization. Furthermore, input data are split into a training subset used by the optimization algorithm and a validation subset used to validate the optimization [75]. This split is performed randomly for each PDF replica. In order to compute the data-induced component of the cross-correlation therefore we have made sure that the two PDF sets that are being compared are fitted to the same data replicas, with the same training-validation split.

Furthermore, not all fits end up in the final PDF set, but only those that pass post-fit criteria specified in Ref. [75]. Because these criteria are applied a posteriori, it might happen that, for a given underlying data replica, the criteria are only passed by one of the two PDF replicas that are being compared. For the computation of the cross-correlation, we only include in the final set PDF replicas for which both sets have passed the criteria.

The S-covariance is then computed using Eq. (4.11), or its obvious generalization in the case of the cross-covariance Eq. (4.4).

In order to estimate the uncertainty on final results due to the finite size of the replica sample we have used a bootstrapping method [168,169]. Specifically, we apply a Monte Carlo algorithm to perform a resampling "with replacement" of the PDF replicas. This is of course done synchronously for both PDF sets between which the correlation is calculated. We then calculate the PDF correlation using these resampled PDF sets. This routine is repeated many times to obtain a precise estimate of the standard error of the PDF correlation. The magnitude of the uncertainty decreases with the inverse square root of the number of PDF replicas used to determine the PDF correlation. We have performed this procedure with 200 resampled sets. The value was chosen comparing the 2σ standard deviation and the 95% confidence interval, and checking that for any flavor and any value of x they agree.

Finally, we have by default computed the cross-correlation at the scale $Q_0 = 1.7$ GeV, and we have checked that by repeating the computation with different choices of Q_0 up to 100 GeV results are unchanged.

Appendix C

Distance estimators

Distance plots such as those shown in Fig. 4.3 provide a measure of the distance between to PDFs provided in a Monte Carlo representation. This distance estimator has been introduced in Ref. [75] and is commonly exploited in many of the NNPDF analyses reference in this thesis, where it can be used as a test of the statistical equivalence between the PDF sets.

Given a Monte Carlo sample of $N_{\rm rep}$ replicas representing the probability distribution of a given PDF set, $\{f^{(k)}\}$, the expectation value of the distribution can be determined using Eq. (2.1). It is given by

$$\langle f \rangle = \frac{1}{N_{\rm rep}} \sum_{r=1}^{N_{\rm rep}} f^{(r)}, \tag{C.1}$$

where the index r runs over all the replicas in the sample, the brackets denotes an average over replicas, and the x and Q^2 dependence of the PDFs is suppressed. The variance of the sample can similarly be obtained using Eq. (2.2), and reads

$$\mathrm{Var}\left[f\right] = \frac{1}{N_{\rm rep}-1} \sum_{r=1}^{N_{\rm rep}} \left(f^{(r)} - \langle f \rangle\right)^2. \tag{C.2}$$

The variance of the mean is then

$$\mathrm{Var}\left[\langle f \rangle\right] = \frac{1}{N_{\rm rep}} \mathrm{Var}\left[f\right], \tag{C.3}$$

while the variance of the variance itself can be written as

$$\mathrm{Var}\left[\mathrm{Var}\left[f\right]\right] = \frac{1}{N_{\rm rep}} \left[\frac{1}{N_{\rm rep}} \sum_{r=1}^{N_{\rm rep}} \left(f^{(r)} - \langle f \rangle\right)^4 - \frac{N_{\rm rep}-3}{N_{\rm rep}-1} \left(\mathrm{Var}\left[f\right]\right)^2 \right]. \tag{C.4}$$

The distance between two PDF sets denoted by $\{f^{(r)}\}$ and $\{g^{(r)}\}$, can be defined as the square root of the square difference of the PDF central values in units of the uncertainty of the mean, that is

$$d\left(\langle f\rangle, \langle g\rangle\right) = \sqrt{\frac{\left(\langle f\rangle - \langle g\rangle\right)^2}{\mathrm{Var}[\langle f\rangle] + \mathrm{Var}[\langle g\rangle]}}. \tag{C.5}$$

Analogously, the distance for the variances of the two samples can be defined as

$$d\left(\mathrm{Var}[f], \mathrm{Var}[g]\right) = \sqrt{\frac{\left(\mathrm{Var}[f] - \mathrm{Var}[g]\right)^2}{\mathrm{Var}\left[\mathrm{Var}[f]\right] + \mathrm{Var}\left[\mathrm{Var}[g]\right]}}. \tag{C.6}$$

According to these definitions a distance of around $d \sim 1/N_{\mathrm{rep}}$ indicates a 1σ disagreement in units of the corresponding denominator. As such, these distances can be used as a test to check whether the distributions from which the PDFs are sampled have the same mean and variance.

Appendix D

Closure testing

The basic idea of closure testing [259] is to perform a PDF determination based on artificial data that have been generated with perfect statistical properties from a known underlying input PDF, $f_{\rm in}$, given a theoretical model. The statistical properties of such a set of pseudodata is then completely known and controlled, it furthermore does not contain any internal inconsistencies, and is perfectly consistent with the theoretical model used in their generation. This then allows to test the fitting methodology by performing a fit to this pseudodata and check if the underlying PDF is reproduced with the correct uncertainties. Closure tests were first used to validate certain aspects of a PDF fitting methodology in Ref. [179], and later adapted by NNPDF to validate the NNPDF3.0 release [75].

The result of the closure tests fit is an ensemble of PDFs which are to be compared to the input PDF, $f_{\rm in}$. The fits are performed in the usual way through the minimization of a χ^2 loss function, Eq. (2.9) with the t_0 covariance matrix defined as Eq. (2.11). In the context of a closure test the true solution is known, namely, it is given by the input PDF $f_{\rm in}$, and the ensemble of fitted PDF replicas should correctly reproduce this known input PDF within the statistical uncertainties determined by the fit.

We distinguish three "levels" of closure tests, corresponding to different levels of stochastic noise added to the pseudodata generated from the input PDF. Here, the stochastic noise is generated by sampling the experimental covariance matrix of the relevant datasets. The three levels are defined as follows:

level 0:

- The central value of the pseudodata is given by predictions of the input PDF
- Each replica is fit to the same data directly corresponding to the predictions of the input PDF.

level 1:

- The central values of the pseudodata are shifted by some noise generated in agreement with the experimental covariance matrix. A level 1 dataset is comparable to an experimental dataset in that it does not sit exactly on top of the underlying law, but rather contain a layer of noise in accordance with the experimental uncertainty.

- Each replica is determined using the same shifted data. There is however a difference in the split of training and validation data used for stopping. The observed PDF covariance matrix in a PDF ensemble produced with these settings is due to this split in addition to any methodological uncertainty.

level 2:

- A further level of Monte Carlo noise is added on top of the level 1 shift.
- The level 2 shift changes replica-by-replica during sampling. Thus a level 2 closure test fit is comparable to a real fit within the NNPDF framework (with known underlying law and statical properties).

Bibliography

[1] **ATLAS** Collaboration, G. Aad *et al.*, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC," *Phys. Lett. B* **716** (2012) 1–29, arXiv:1207.7214 [hep-ex].

[2] **CMS** Collaboration, S. Chatrchyan *et al.*, "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC," *Phys. Lett. B* **716** (2012) 30–61, arXiv:1207.7235 [hep-ex].

[3] M. E. Peskin and D. V. Schroeder, *An Introduction to quantum field theory*. Addison-Wesley, Reading, USA, 1995.

[4] R. K. Ellis, W. J. Stirling, and B. R. Webber, *QCD and collider physics*. Cambridge University Press, 1996.

[5] N. Hartland, *Proton structure at the LHC*. PhD thesis, University of Edinburgh, 11, 2014.

[6] **NNPDF** Collaboration, R. D. Ball *et al.*, "The path to proton structure at 1% accuracy," *Eur. Phys. J. C* **82** no. 5, (2022) 428, arXiv:2109.02653 [hep-ph].

[7] **NNPDF** Collaboration, R. D. Ball *et al.*, "An open-source machine learning framework for global analyses of parton distributions," *Eur. Phys. J. C* **81** no. 10, (2021) 958, arXiv:2109.02671 [hep-ph].

[8] S. Carrazza, J. M. Cruz-Martinez, and R. Stegeman, "A data-based parametrization of parton distribution functions," *Eur. Phys. J. C* **82** no. 2, (11, 2022) 163, arXiv:2111.02954 [hep-ph].

[9] M. Gell-Mann, "Symmetries of baryons and mesons," *Phys. Rev.* **125** (1962) 1067–1084.

[10] M. Gell-Mann, "A schematic model of baryons and mesons," *Phys. Lett.* **8** (1964) 214–215.

[11] G. Zweig, "An su_3 model for strong interaction symmetry and its breaking; version 2,". http://cds.cern.ch/record/570209. Version 1 is CERN preprint 8182/TH.401, Jan. 17, 1964.

[12] G. Altarelli, P. Nason, and G. Ridolfi, "A Study of ultraviolet renormalon ambiguities in the determination of alpha-s from tau decay," *Z. Phys. C* **68** (1995) 257–268, arXiv:hep-ph/9501240.

[13] 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* **02** (2017) 090, arXiv:1701.01404 [hep-ph].

[14] D. J. Gross and F. Wilczek, "Ultraviolet behavior of nonabelian gauge theories," *Phys. Rev. Lett.* **30** (1973) 1343–1346.

[15] H. D. Politzer, "Reliable perturbative results for strong interactions?" *Phys. Rev. Lett.* **30** (1973) 1346–1349.

[16] **Particle Data Group** Collaboration, R. L. Workman, "Review of Particle Physics," *PTEP* **2022** (2022) 083C01.

[17] 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* **282** (1992) 475–482.

[18] **BCDMS** Collaboration, 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.* **B223** (1989) 485.

[19] **H1** Collaboration, F. D. Aaron *et al.*, "Measurement of the Proton Structure Function F_L at Low x," *Phys. Lett.* **B665** (2008) 139–146, arXiv:0805.2809 [hep-ex].

[20] **ZEUS** Collaboration, A. Cooper Sarkar, "Measurement of high-Q^2 neutral current deep inelastic e+p scattering cross sections with a longitudinally polarised positron beam at HERA," arXiv:1208.6138 [hep-ex].

[21] R. P. Feynman, "The behavior of hadron collisions at extreme energies," *Conf. Proc. C* **690905** (1969) 237–258.

[22] J. D. Bjorken, "Asymptotic Sum Rules at Infinite Momentum," *Phys. Rev.* **179** (1969) 1547–1553.

[23] C. G. Callan, Jr. and D. J. Gross, "High-energy electroproduction and the constitution of the electric current," *Phys. Rev. Lett.* **22** (1969) 156–159.

[24] **CTEQ** Collaboration, R. Brock *et al.*, "Handbook of perturbative QCD: Version 1.0" *Rev. Mod. Phys.* **67** (1995) 157–248.

[25] **H1** Collaboration, T. Ahmed *et al.*, "A Measurement of the proton structure function f2 (x, Q**2)," *Nucl. Phys. B* **439** (1995) 471–502, arXiv:hep-ex/9503001.

[26] T. Kinoshita, "Mass singularities of Feynman amplitudes," *J. Math. Phys.* **3** (1962) 650–677.

[27] T. D. Lee and M. Nauenberg, "Degenerate Systems and Mass Singularities," *Phys. Rev.* **133** (1964) B1549–B1562.

[28] E. B. Zijlstra and W. L. van Neerven, "Order alpha-s**2 QCD corrections to the deep inelastic proton structure functions F2 and F(L)," *Nucl. Phys. B* **383** (1992) 525–574.

[29] A. Vogt, S. Moch, and J. A. M. Vermaseren, "The Three-loop splitting functions in QCD: The Singlet case," *Nucl. Phys. B* **691** (2004) 129–181.

[30] J. Davies, A. Vogt, B. Ruijl, T. Ueda, and J. A. M. Vermaseren, "Large-n_f contributions to the four-loop splitting functions in QCD," *Nucl. Phys. B* **915** (2017) 335–362.

[31] S. Moch, B. Ruijl, T. Ueda, J. A. M. Vermaseren, and A. Vogt, "Four-Loop Non-Singlet Splitting Functions in the Planar Limit and Beyond," *JHEP* **10** (2017) 041.

[32] S. Moch, B. Ruijl, T. Ueda, J. A. M. Vermaseren, and A. Vogt, "On quartic colour factors in splitting functions and the gluon cusp anomalous dimension," *Phys. Lett. B* **782** (2018) 627–632.

[33] 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* **152** (1979) 285–329.

[34] G. T. Bodwin, "Factorization of the Drell-Yan Cross-Section in Perturbation Theory," *Phys. Rev. D* **31** (1985) 2616. [Erratum: Phys.Rev.D 34, 3932 (1986)].

[35] J. C. Collins, D. E. Soper, and G. F. Sterman, "Factorization for Short Distance Hadron - Hadron Scattering," *Nucl. Phys. B* **261** (1985) 104–142.

[36] V. N. Gribov and L. N. Lipatov, "Deep inelastic electron scattering in perturbation theory," *Phys. Lett. B* **37** (1971) 78–80.

[37] Y. L. Dokshitzer, "Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics.," *Sov. Phys. JETP* **46** (1977) 641–653.

[38] G. Altarelli and G. Parisi, "Asymptotic Freedom in Parton Language," *Nucl. Phys. B* **126** (1977) 298–318.

[39] G. P. Salam and J. Rojo, "A Higher Order Perturbative Parton Evolution Toolkit (HOPPET)," *Comput. Phys. Commun.* **180** (2009) 120–156, arXiv:0804.3755 [hep-ph].

[40] M. Botje, "QCDNUM: Fast QCD Evolution and Convolution," *Comput. Phys. Commun.* **182** (2011) 490–532, arXiv:1005.1481 [hep-ph].

[41] V. Bertone, S. Carrazza, and J. Rojo, "APFEL: A PDF Evolution Library with QED corrections," *Comput. Phys. Commun.* **185** (2014) 1647–1668, arXiv:1310.1394 [hep-ph].

[42] A. Candido, F. Hekhorn, and G. Magni, "EKO: Evolution Kernel Operators," arXiv:2202.02338 [hep-ph].

[43] **NNPDF** Collaboration, L. Del Debbio, S. Forte, J. I. Latorre, A. Piccione, and J. Rojo, "Neural network determination of parton distributions: The Nonsinglet case," *JHEP* **03** (2007) 039, arXiv:hep-ph/0701127.

[44] **NNPDF** Collaboration, 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* **809** (2009) 1–63, arXiv:0808.1231 [hep-ph]. [Erratum: Nucl.Phys.B 816, 293 (2009)].

[45] **The NNPDF** Collaboration, 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* **838** (2010) 136–206, arXiv:1002.4407 [hep-ph].

[46] T. Appelquist and J. Carazzone, "Infrared Singularities and Massive Fields," *Phys. Rev. D* **11** (1975) 2856.

[47] T. Appelquist and J. Carazzone, "Physical Processes and the Infrared Problem in Gauge Theories," *Nucl. Phys. B* **120** (1977) 77–95.

[48] K. Symanzik, "Infrared singularities and small distance behavior analysis," *Commun. Math. Phys.* **34** (1973) 7–36.

[49] M. Buza, Y. Matiounine, J. Smith, R. Migneron, and W. L. van Neerven, "Heavy quark coefficient functions at asymptotic values Q**2 >> m**2," *Nucl. Phys. B* **472** (1996) 611–658, arXiv:hep-ph/9601302.

[50] M. Buza, Y. Matiounine, J. Smith, and W. L. van Neerven, "Charm electroproduction viewed in the variable flavor number scheme versus fixed order perturbation theory," *Eur. Phys. J. C* **1** (1998) 301–320.

[51] R. S. Thorne and W. K. Tung, "PQCD Formulations with Heavy Quark Masses and Global Analysis," in *HERA and the LHC: 4th Workshop on the Implications of HERA for LHC Physics*, pp. 332–351. 9, 2008. arXiv:0809.0714 [hep-ph].

[52] J. C. Collins, F. Wilczek, and A. Zee, "Low-Energy Manifestations of Heavy Particles: Application to the Neutral Current," *Phys. Rev. D* **18** (1978) 242.

[53] J. C. Collins, "Hard scattering factorization with heavy quarks: A General treatment," *Phys. Rev. D* **58** (1998) 094002.

[54] M. Kramer, 1, F. I. Olness, and D. E. Soper, "Treatment of heavy quarks in deeply inelastic scattering," *Phys. Rev.* **D62** (2000) 096007, arXiv:hep-ph/0003035.

[55] R. S. Thorne and R. G. Roberts, "A Practical procedure for evolving heavy flavor structure functions," *Phys. Lett. B* **421** (1998) 303–311.

[56] R. S. Thorne, "A Variable-flavor number scheme for NNLO," *Phys. Rev. D* **73** (2006) 054019, arXiv:hep-ph/0601245 [hep-ph].

[57] M. Cacciari, M. Greco, and P. Nason, "The P(T) spectrum in heavy flavor hadroproduction," *JHEP* **05** (1998) 007, arXiv:hep-ph/9803400.

[58] S. Forte and S. Carrazza, "Parton distribution functions," arXiv:2008.12305 [hep-ph].

[59] **NNPDF** Collaboration, R. D. Ball *et al.*, "Parton distributions from high-precision collider data," *Eur. Phys. J. C* **77** no. 10, (2017) 663, arXiv:1706.00428 [hep-ph].

[60] S. Carrazza and J. Cruz-Martinez, "Towards a new generation of parton densities with deep learning models," *Eur. Phys. J. C* **79** no. 8, (2019) 676, arXiv:1907.05075 [hep-ph].

[61] A. Candido, S. Forte, and F. Hekhorn, "Can $\overline{\text{MS}}$ parton distributions be negative?," *JHEP* **11** (2020) 129, arXiv:2006.07377 [hep-ph].

[62] **HL-LHC, HE-LHC Working Group** Collaboration, P. Azzi *et al.*, "Standard Model Physics at the HL-LHC and HE-LHC," *CERN Yellow Rep. Monogr.* **7** (2019) 1–220, arXiv:1902.04070 [hep-ph].

[63] 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* **81** no. 4, (2021) 341, arXiv:2012.04684 [hep-ph].

[64] T.-J. Hou *et al.*, "New cteq global analysis of quantum chromodynamics with high-precision data from the lhc," *Phys. Rev. D* **103** no. 1, (2021) 014013, arXiv:1912.10053 [hep-ph].

[65] S. Alekhin, J. Blümlein, S. Moch, and R. Placakyte, "Parton distribution functions, α_s, and heavy-quark masses for lhc run ii," *Phys. Rev. D* **96** no. 1, (2017) 014011, arXiv:1701.05838 [hep-ph].

[66] D. Stump, J. Pumplin, R. Brock, D. Casey, J. Huston, J. Kalk, H. L. Lai, and W. K. Tung, "Uncertainties of predictions from parton distribution functions. 1. The Lagrange multiplier method," *Phys. Rev. D* **65** (2001) 014012, arXiv:hep-ph/0101051.

[67] J. Pumplin *et al.*, "Uncertainties of predictions from parton distribution functions. 2. The Hessian method," *Phys. Rev.* **D65** (2001) 014013, arXiv:hep-ph/0101032.

[68] S. Forte, L. Garrido, J. I. Latorre, and A. Piccione, "Neural network parametrization of deep inelastic structure functions," *JHEP* **05** (2002) 062, arXiv:hep-ph/0204232.

[69] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks* **2** no. 5, (1989) 359–366.

[70] **NNPDF** Collaboration, L. Del Debbio, S. Forte, J. I. Latorre, A. Piccione, and J. Rojo, "Unbiased determination of the proton structure function f(2)**p with faithful uncertainty estimation," *JHEP* **03** (2005) 080, arXiv:hep-ph/0501067.

[71] **NNPDF** Collaboration, R. D. Ball, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, A. Piccione, J. Rojo, and M. Ubiali, "Precision determination of electroweak parameters and the strange content of the proton from neutrino deep-inelastic scattering," *Nucl. Phys. B* **823** (2009) 195–233, arXiv:0906.1958 [hep-ph].

[72] **The NNPDF** Collaboration, R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, J. Rojo, and M. Ubiali, "Impact of heavy quark masses on parton distributions and lhc phenomenology," *Nucl. Phys. B* **849** (2011) 296–363, arXiv:1101.1300 [hep-ph].

[73] R. D. Ball *et al.*, "Parton distributions with lhc data," *Nucl. Phys. B* **867** (2013) 244–289, arXiv:1207.1303 [hep-ph].

[74] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.

[75] **NNPDF** Collaboration, R. D. Ball *et al.*, "Parton distributions for the LHC Run II," *JHEP* **04** (2015) 040, arXiv:1410.8849 [hep-ph].

[76] J. Butterworth *et al.*, "Pdf4lhc recommendations for lhc run ii," *J. Phys. G* **43** (2016) 023001, arXiv:1510.03865 [hep-ph].

[77] R. D. Ball, S. Forte, and R. Stegeman, "Correlation and combination of sets of parton distributions," *Eur. Phys. J. C* **81** no. 11, (2021) 1046, arXiv:2110.08274 [hep-ph].

[78] A. Buckley, J. Ferrando, S. Lloyd, K. Nordström, B. Page, M. Rüfenacht, M. Schönherr, and G. Watt, "Lhapdf6: parton density access in the lhc precision era," *Eur. Phys. J. C* **75** (2015) 132, arXiv:1412.7420 [hep-ph].

[79] A. Pérez-Salinas, J. Cruz-Martinez, A. A. Alhajri, and S. Carrazza, "Determining the proton content with a quantum computer," *Phys. Rev. D* **103** no. 3, (2021) 034027, arXiv:2011.13934 [hep-ph].

[80] S. Efthymiou, S. Ramos-Calderer, C. Bravo-Prieto, A. Pérez-Salinas, D. García-Martín, A. Garcia-Saez, J. I. Latorre, and S. Carrazza, "Qibo: a framework for quantum simulation with hardware acceleration," *Quantum Sci. Technol.* **7** no. 1, (2022) 015018, arXiv:2009.01845 [quant-ph].

[81] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on scientific computing* **16** no. 5, (1995) 1190–1208.

[82] P. Virtanen, *et al.*, "Scipy 1.0: fundamental algorithms for scientific computing in python," *Nature methods* **17** no. 3, (2020) 261–272.

[83] G. D'Agostini, *Bayesian Reasoning in Data Analysis: A Critical Introduction*. World Scientific, Singapore, 2003. https://cds.cern.ch/record/642515.

[84] **NNPDF** Collaboration, 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* **05** (2010) 075, arXiv:0912.2276 [hep-ph].

[85] M. Abadi, *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. https://www.tensorflow.org/. Software available from tensorflow.org.

[86] T. Tieleman, G. Hinton, *et al.*, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA: Neural networks for machine learning* **4** no. 2, (2012) 26–31.

[87] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research* **12** no. 61, (2011) 2121–2159. http://jmlr.org/papers/v12/duchi11a.html.

[88] 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. http://arxiv.org/abs/1412.6980.

[89] R. Stegeman, J. Cruz-Martinez, and S. Carrazza, "evolutionary_keras: a genetic algorithm library," Jan., 2021. https://doi.org/10.5281/zenodo.4415396.

[90] J. M. Cruz-Martinez, S. Carrazza, and R. Stegeman, "Studying the parton content of the proton with deep learning models," *PoS* **AISIS2019** (2020) 008, arXiv:2002.06587 [physics.comp-ph].

[91] F. Chollet *et al.*, "Keras," 2015. https://github.com/fchollet/keras.

[92] M. D. Zeiler, "Adadelta: an adaptive learning rate method," *arXiv preprint arXiv:1212.5701* (2012) .

[93] N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evolutionary computation* **9** no. 2, (2001) 159–195.

[94] N. Hansen, "The CMA evolution strategy: A tutorial," 1604.00772. http://arxiv.org/abs/1604.00772.

[95] **NNPDF** Collaboration, V. Bertone, S. Carrazza, N. P. Hartland, E. R. Nocera, and J. Rojo, "A determination of the fragmentation functions of pions, kaons, and protons with faithful uncertainties," *Eur. Phys. J. C* **77** no. 8, (2017) 516, arXiv:1706.07049 [hep-ph].

[96] S. Carrazza and N. P. Hartland, "Minimisation strategies for the determination of parton density functions," *J. Phys. Conf. Ser.* **1085** no. 5, (2018) 052007, arXiv:1711.09991 [hep-ph].

[97] J. Bergstra, D. Yamins, and D. Cox, "Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures," in *Proceedings of the 30th International Conference on Machine Learning*, S. Dasgupta and D. McAllester, eds., vol. 28 of *Proceedings of Machine Learning Research*, pp. 115–123. PMLR, Atlanta, Georgia, USA, 17–19 jun, 2013. https://proceedings.mlr.press/v28/bergstra13.html.

[98] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, "Algorithms for hyper-parameter optimization," in *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, eds., vol. 24. Curran Associates, Inc., 2011. https://proceedings.neurips.cc/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

[99] T. Dozat, "Incorporating nesterov momentum into adam," in *Proceedings of the 4th International Conference on Learning Representations*. 2016.

[100] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, "On the importance of initialization and momentum in deep learning," in *International conference on machine learning*, pp. 1139–1147, PMLR. 2013.

[101] D. M. Hawkins, "The problem of overfitting," *J. Chem. Inf. Comput. Sci.* **44** (2004) 1.

[102] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, vol. 2. Springer, 2009.

[103] Y. Bengio and X. Glorot, "Understanding the difficulty of training deep feed forward neural networks," *International Conference on Artificial Intelligence and Statistics* (01, 2010) 249–256.

[104] **Particle Data Group** Collaboration, P. A. Zyla *et al.*, "Review of Particle Physics," *PTEP* **2020** no. 8, (2020) 083C01.

[105] **NOMAD** Collaboration, O. Samoylov *et al.*, "A precision measurement of charm dimuon production in neutrino interactions from the nomad experiment," *Nucl.Phys.* **B876** (2013) 339, arXiv:1308.4750 [hep-ex].

[106] **H1** Collaboration, F. D. Aaron *et al.*, "Measurement of the charm and beauty structure functions using the h1 vertex detector at hera," *Eur. Phys. J.* **C65** (2010) 89–109, arXiv:0907.2643 [hep-ex].

[107] **ZEUS** Collaboration, H. Abramowicz *et al.*, "Measurement of beauty and charm production in deep inelastic scattering at hera and measurement of the beauty-quark mass," *JHEP* **09** (2014) 127, arXiv:1405.6915 [hep-ex].

[108] **H1, ZEUS** Collaboration, 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* **78** no. 6, (2018) 473, arXiv:1804.01019 [hep-ex].

[109] **ZEUS** Collaboration, S. Chekanov *et al.*, "Inclusive jet cross-sections in the breit frame in neutral current deep inelastic scattering at hera and determination of alpha(s)," *Phys. Lett. B* **547** (2002) 164–180, arXiv:hep-ex/0208037.

[110] **ZEUS** Collaboration, S. Chekanov *et al.*, "Inclusive-jet and dijet cross-sections in deep inelastic scattering at hera," *Nucl. Phys. B* **765** (2007) 1–30, arXiv:hep-ex/0608048.

[111] **ZEUS** Collaboration, H. Abramowicz *et al.*, "Inclusive dijet cross sections in neutral current deep inelastic scattering at hera," *Eur. Phys. J. C* **70** (2010) 965–982, arXiv:1010.6167 [hep-ex].

[112] **H1** Collaboration, V. Andreev *et al.*, "Measurement of jet production cross sections in deep-inelastic ep scattering at hera," *Eur. Phys. J. C* **77** no. 4, (2017) 215, arXiv:1611.03421 [hep-ex].

[113] **H1** Collaboration, V. Andreev *et al.*, "Measurement of multijet production in *ep* collisions at high q^2 and determination of the strong coupling α_s," *Eur. Phys. J. C* **75** no. 2, (2015) 65, arXiv:1406.4709 [hep-ex].

[114] **SeaQuest** Collaboration, J. Dove *et al.*, "The asymmetry of antimatter in the proton," *Nature* **590** no. 7847, (2021) 561–565, arXiv:2103.04024 [hep-ph].

[115] **ATLAS** Collaboration, 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.* **C77** no. 6, (2017) 367, arXiv:1612.03016 [hep-ex].

[116] **ATLAS** Collaboration, 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* **08** (2016) 009, arXiv:1606.01736 [hep-ex].

[117] **ATLAS** Collaboration, M. Aaboud *et al.*, "Measurement of the Drell-Yan triple-differential cross section in pp collisions at $\sqrt{s} = 8$ TeV," *JHEP* **12** (2017) 059, arXiv:1710.05167 [hep-ex].

[118] **ATLAS** Collaboration, G. Aad *et al.*, "Measurement of the cross-section and charge asymmetry of W bosons produced in proton–proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector," *Eur. Phys. J. C* **79** no. 9, (2019) 760, arXiv:1904.05631 [hep-ex].

[119] **ATLAS** Collaboration, G. Aad *et al.*, "Measurement of $W^{\pm}$ and Z-boson production cross sections in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector," *Phys. Lett.* **B759** (2016) 601–621, arXiv:1603.09222 [hep-ex].

[120] **LHCb** Collaboration, R. Aaij *et al.*, "Measurement of the forward Z boson production cross-section in pp collisions at $\sqrt{s} = 13$ TeV," *JHEP* **09** (2016) 136, arXiv:1607.06495 [hep-ex].

[121] **ATLAS** Collaboration, 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* **05** (2018) 077, arXiv:1711.03296 [hep-ex]. [Erratum: JHEP 10, 048 (2020)].

[122] **ATLAS** Collaboration, 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* **1405** (2014) 068, arXiv:1402.6263 [hep-ex].

[123] **CMS** Collaboration, 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* **79** no. 3, (2019) 269, arXiv:1811.10021 [hep-ex].

[124] **ATLAS** Collaboration, M. Aaboud *et al.*, "Measurement of top quark pair differential cross-sections in the dilepton channel in pp collisions at $\sqrt{s} = 7$ and 8 TeV with ATLAS," *Phys. Rev. D* **94** no. 9, (2016) 092003, arXiv:1607.07281 [hep-ex]. [Addendum: Phys.Rev.D 101, 119901 (2020)].

[125] **CMS** Collaboration, A. M. Sirunyan *et al.*, "Measurement of double-differential cross sections for top quark pair production in pp collisions at $\sqrt{s} = 8$ TeV and impact on parton distribution functions," *Eur. Phys. J.* **C77** no. 7, (2017) 459, arXiv:1703.01630 [hep-ex].

[126] **ATLAS** Collaboration, G. Aad *et al.*, "Measurement of the $t\bar{t}$ production cross-section in the lepton+jets channel at $\sqrt{s} = 13$ TeV with the ATLAS experiment," *Phys. Lett. B* **810** (2020) 135797, arXiv:2006.13076 [hep-ex].

[127] **CMS** Collaboration, A. M. Sirunyan *et al.*, "Measurement of differential cross sections for the production of top quark pairs and of additional jets in lepton+jets events from pp collisions at $\sqrt{s} = 13$ TeV," *Phys. Rev.* **D97** no. 11, (2018) 112003, arXiv:1803.08856 [hep-ex].

[128] **CMS** Collaboration, A. M. Sirunyan *et al.*, "Measurements of $t\bar{t}$ differential cross sections in proton-proton collisions at $\sqrt{s} = 13$ TeV using events containing two leptons," *JHEP* **02** (2019) 149, arXiv:1811.06625 [hep-ex].

[129] **ATLAS** Collaboration, M. Aaboud *et al.*, "Measurement of the inclusive jet cross-sections in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector," *JHEP* **09** (2017) 020, arXiv:1706.03192 [hep-ex].

[130] **CMS** Collaboration, V. Khachatryan *et al.*, "Measurement and QCD analysis of double-differential inclusive jet cross sections in pp collisions at $\sqrt{s} = 8$ TeV and cross section ratios to 2.76 and 7 TeV," *JHEP* **03** (2017) 156, arXiv:1609.05331 [hep-ex].

[131] **ATLAS Collaboration** Collaboration, G. Aad *et al.*, "Measurement of dijet cross sections in *pp* collisions at 7 tev centre-of-mass energy using the atlas detector," *JHEP* **1405** (2014) 059, arXiv:1312.3524 [hep-ex].

[132] **CMS** Collaboration, S. Chatrchyan *et al.*, "Measurements of differential jet cross sections in proton-proton collisions at $\sqrt{s} = 7$ TeV with the CMS detector," *Phys.Rev.* **D87** no. 11, (2013) 112002, arXiv:1212.6660 [hep-ex]. [Erratum: Phys. Rev.D87,no.11,119902(2013)].

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

[134] **ATLAS** Collaboration, G. Aad *et al.*, "Measurement of the inclusive isolated prompt photon cross section in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector," *JHEP* **08** (2016) 005, arXiv:1605.03495 [hep-ex].

[135] **ATLAS** Collaboration, M. Aaboud *et al.*, "Measurement of the cross section for inclusive isolated-photon production in *pp* collisions at $\sqrt{s} = 13$ tev using the atlas detector," *Phys. Lett. B* **770** (2017) 473–493, arXiv:1701.06882 [hep-ex].

[136] **ATLAS** Collaboration, G. Aad *et al.*, "Comprehensive measurements of t-channel single top-quark production cross sections at $\sqrt{s} = 7$ TeV with the ATLAS detector," *Phys. Rev. D* **90** no. 11, (2014) 112006, arXiv:1406.7844 [hep-ex].

[137] **ATLAS** Collaboration, M. Aaboud *et al.*, "Fiducial, total and differential cross-section measurements of t-channel single top-quark production in pp collisions at 8 tev using data collected by the atlas detector," *Eur. Phys. J. C* **77** no. 8, (2017) 531, arXiv:1702.02859 [hep-ex].

[138] **ATLAS** Collaboration, M. Aaboud *et al.*, "Measurement of the inclusive cross-sections of single top-quark and top-antiquark t-channel production in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector," *JHEP* **04** (2017) 086, arXiv:1609.03920 [hep-ex].

[139] **CMS** Collaboration, S. Chatrchyan *et al.*, "Measurement of the Single-Top-Quark t-Channel Cross Section in pp Collisions at $\sqrt{s} = 7$ TeV," *JHEP* **12** (2012) 035, arXiv:1209.4533 [hep-ex].

[140] **CMS** Collaboration, V. Khachatryan *et al.*, "Measurement of the t-channel single-top-quark production cross section and of the $| V_{tb} |$ CKM matrix element in pp collisions at $\sqrt{s}$= 8 TeV," *JHEP* **06** (2014) 090, arXiv:1403.7366 [hep-ex].

[141] **CMS** Collaboration, A. M. Sirunyan *et al.*, "Cross section measurement of t-channel single top quark production in pp collisions at $\sqrt{s} = 13$ tev," *Phys. Lett. B* **772** (2017) 752–776, arXiv:1610.00678 [hep-ex].

[142] **New Muon** Collaboration, M. Arneodo *et al.*, "Measurement of the proton and deuteron structure functions, f_2^p and f_2^d, and of the ratio σ_l/σ_t," *Nucl. Phys.* **B483** (1997) 3–43, arXiv:hep-ph/9610231 [hep-ph].

[143] **ZEUS, H1** Collaboration, 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.* **C75** no. 12, (2015) 580, arXiv:1506.06042 [hep-ex].

[144] **NuSea** Collaboration, J. C. Webb *et al.*, "Absolute drell-yan dimuon cross sections in 800-gev/c p p and p d collisions," arXiv:hep-ex/0302019.

[145] **NNPDF** Collaboration, 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* **76** no. 11, (2016) 647, arXiv:1605.06515 [hep-ph].

[146] **NNPDF** Collaboration, R. D. Ball, A. Candido, J. Cruz-Martinez, S. Forte, T. Giani, F. Hekhorn, K. Kudashkin, G. Magni, and J. Rojo, "Evidence for intrinsic charm quarks in the proton," *Nature* **608** no. 7923, (2022) 483–487, arXiv:2208.08372 [hep-ph].

[147] **PDF4LHC Working Group** Collaboration, R. D. Ball *et al.*, "The pdf4lhc21 combination of global pdf fits for the lhc run iii," *J. Phys. G* **49** no. 8, (2022) 080501, arXiv:2203.05506 [hep-ph].

[148] E. Maguire, L. Heinrich, and G. Watt, "Hepdata: a repository for high energy physics data," *J. Phys. Conf. Ser.* **898** no. 10, (2017) 102006, arXiv:1704.05473 [hep-ex].

[149] Z. Kassabov, "Reportengine: A framework for declarative data analysis." https://doi.org/10.5281/zenodo.2571601, Feb., 2019.

[150] **NNPDF** Collaboration, 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* **78** no. 5, (2018) 408, arXiv:1802.03398 [hep-ph].

[151] **NNPDF** Collaboration, R. Abdul Khalek *et al.*, "A first determination of parton distributions with theoretical uncertainties," *Eur. Phys. J.* **C** (2019) 79:838, arXiv:1905.04311 [hep-ph].

[152] **NNPDF** Collaboration, R. Abdul Khalek *et al.*, "Parton distributions with theory uncertainties: General formalism and first phenomenological studies," *Eur. Phys. J. C* **79** no. 11, (2019) 931, arXiv:1906.10698 [hep-ph].

[153] S. Iranipour and M. Ubiali, "A new generation of simultaneous fits to lhc data using deep learning," *JHEP* **05** (2022) 032, arXiv:2201.07240 [hep-ph].

[154] R. Abdul 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," *Eur. Phys. J. C* **82** no. 6, (2022) 507, arXiv:2201.12363 [hep-ph].

[155] R. A. Khalek, V. Bertone, A. Khoudli, and E. R. Nocera, "Pion and kaon fragmentation functions at next-to-next-to-leading order," arXiv:2204.10331 [hep-ph].

[156] **NNPDF** Collaboration, E. R. Nocera, R. D. Ball, S. Forte, G. Ridolfi, and J. Rojo, "A first unbiased global determination of polarized pdfs and their uncertainties," *Nucl. Phys. B* **887** (2014) 276–308, arXiv:1406.5539 [hep-ph].

[157] S. Alekhin *et al.*, "Herafitter," *Eur. Phys. J. C* **75** no. 7, (2015) 304, arXiv:1410.4412 [hep-ph].

[158] **xFitter** Collaboration, H. Abdolmaleki *et al.*, "xfitter: An open source qcd analysis framework. a resource and reference document for the snowmass study," 6, 2022. https://arxiv.org/abs/2206.12465.

[159] S. J. Brodsky and G. R. Farrar, "Scaling Laws at Large Transverse Momentum," *Phys. Rev. Lett.* **31** (1973) 1153–1156.

[160] H. D. I. Abarbanel, M. L. Goldberger, and S. B. Treiman, "Asymptotic properties of electroproduction structure functions," *Phys. Rev. Lett.* **22** (1969) 500–502.

[161] R. D. Ball, E. R. Nocera, and J. Rojo, "The asymptotic behaviour of parton distributions at small and large x," *Eur. Phys. J.* **C76** no. 7, (2016) 383, arXiv:1604.00024 [hep-ph].

[162] R. G. Roberts, *The Structure of the proton: Deep inelastic scattering*. Cambridge University Press, 1993.

[163] F. N. Fritsch and R. E. Carlson, "Monotone piecewise cubic interpolation," *SIAM Journal on Numerical Analysis* **17** no. 2, (1980) 238–246.

[164] J. Cruz-Martinez, S. Forte, and E. R. Nocera, "Future tests of parton distributions," *Acta Phys. Polon. B* **52** (2021) 243, arXiv:2103.08606 [hep-ph].

[165] R. Stegeman, S. Carrazza, and J. Cruz-Martinez, "Small x extrapolation for parton distributions," *PoS* **EPS-HEP2021** (2022) 371.

[166] R. D. Ball, A. Candido, S. Forte, F. Hekhorn, E. R. Nocera, J. Rojo, and C. Schwan, "Parton Distributions and New Physics Searches: the Drell-Yan Forward-Backward Asymmetry as a Case Study," arXiv:2209.08115 [hep-ph].

[167] L. Del Debbio, T. Giani, and M. Wilson, "Bayesian approach to inverse problems: an application to nnpdf closure testing," *Eur. Phys. J. C* **82** no. 4, (2022) 330, arXiv:2111.05787 [hep-ph].

[168] B. Efron, "Bootstrap methods: Another look at the jackknife," *Annals Statist.* **7** no. 1, (1979) 1–26.

[169] B. Efron and R. Tibshirani, "An introduction to the bootstrap," *Statist. Sci.* **57** no. 1, (1986) 54–75.

[170] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics* **22** no. 1, (1951) 79–86.

[171] E. Hellinger, "Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen.," *Journal für die reine und angewandte Mathematik* **1909** no. 136, (1909) 210–271.

[172] P. M. Nadolsky, H.-L. Lai, Q.-H. Cao, J. Huston, J. Pumplin, D. Stump, W.-K. Tung, and C. P. Yuan, "Implications of cteq global analysis for collider observables," *Phys. Rev. D* **78** (2008) 013004, arXiv:0802.0007 [hep-ph].

[173] S. Dittmaier *et al.*, "Handbook of LHC Higgs Cross Sections: 2. Differential Distributions," arXiv:1201.3084 [hep-ph].

[174] S. Carrazza, S. Forte, Z. Kassabov, and J. Rojo, "Specialized minimal pdfs for optimized lhc calculations," *Eur. Phys. J. C* **76** no. 4, (2016) 205, arXiv:1602.00005 [hep-ph].

[175] **HERAFitter developers' Team** Collaboration, P. Belov *et al.*, "Parton distribution functions at lo, nlo and nnlo with correlated uncertainties between orders," *Eur. Phys. J. C* **74** no. 10, (2014) 3039, arXiv:1404.4234 [hep-ph].

[176] **LHC precision EW working group**, "Proposal for PDF benchmarking exercise using LHC precision EW data and pseudodata, https://indico.cern.ch/event/775325/contributions/3241729/attachments/1769767/2875062/PDFnote_benchmarking_031218.pdf." Unpublished note, 2018.

[177] G. Cowan, *Statistical data analysis*. Oxford University Press, 2002.

[178] W. T. Giele, S. A. Keller, and D. A. Kosower, "Parton distribution function uncertainties," arXiv:hep-ph/0104052.

[179] G. Watt and R. S. Thorne, "Study of monte carlo approach to experimental uncertainty propagation with mstw 2008 pdfs," *JHEP* **08** (2012) 052, arXiv:1205.4024 [hep-ph].

[180] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt, "Parton distributions for the lhc," *Eur. Phys. J. C* **63** (2009) 189–285, arXiv:0901.0002 [hep-ph].

[181] **H1, ZEUS** Collaboration, F. D. Aaron *et al.*, "Combined measurement and qcd analysis of the inclusive e+- p scattering cross sections at hera," *JHEP* **01** (2010) 109, arXiv:0911.0884 [hep-ex].

[182] M. Botje *et al.*, "The pdf4lhc working group interim recommendations," arXiv:1101.0538 [hep-ph].

[183] S. Carrazza, S. Forte, Z. Kassabov, J. I. Latorre, and J. Rojo, "An unbiased hessian representation for monte carlo pdfs," *Eur. Phys. J. C* **75** no. 8, (2015) 369, arXiv:1505.06736 [hep-ph].

[184] A. Courtoy, J. Huston, P. Nadolsky, K. Xie, M. Yan, and C. P. Yuan, "Parton distributions need representative sampling," arXiv:2205.10444 [hep-ph].

[185] S. Carrazza, S. Forte, Z. Kassabov, and J. Rojo, "Chapter II.2 of Les Houches 2015: Physics at TeV Colliders Standard Model Working Group Report," in *9th Les Houches Workshop on Physics at TeV Colliders*. 5, 2016. arXiv:1605.04692 [hep-ph].

[186] P. J. Phillips, C. A. Hahn, P. C. Fontana, D. A. Broniatowski, and M. A. Przybocki, "Four principles of explainable artificial intelligence," *Gaithersburg, Maryland* (2020) .

[187] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital signal processing* **73** (2018) 1–15.

[188] **The NNPDF** Collaboration, R. D. Ball *et al.*, "Unbiased determination of polarized parton distributions and their uncertainties," *Nucl.Phys.* **B874** (2013) 36–84, arXiv:1303.7236 [hep-ph].

[189] **NNPDF** Collaboration, V. Bertone, N. P. Hartland, E. R. Nocera, J. Rojo, and L. Rottoli, "Charged hadron fragmentation functions from collider data," *Eur. Phys. J. C* **78** no. 8, (2018) 651, arXiv:1807.03310 [hep-ph].

[190] **NuSTEC** Collaboration, L. Alvarez-Ruso *et al.*, "NuSTEC White Paper: Status and challenges of neutrino–nucleus scattering," *Prog. Part. Nucl. Phys.* **100** (2018) 1–68, arXiv:1706.03621 [hep-ph].

[191] A. Garcia, R. Gauld, A. Heijboer, and J. Rojo, "Complete predictions for high-energy neutrino propagation in matter," *JCAP* **09** (2020) 025, arXiv:2004.04756 [hep-ph].

[192] V. Bertone, R. Gauld, and J. Rojo, "Neutrino Telescopes as QCD Microscopes," *JHEP* **01** (2019) 217, arXiv:1808.02034 [hep-ph].

[193] U.-K. Yang and A. Bodek, "Parton distributions, d/u, and higher twist effects at high x," *Phys. Rev. Lett.* **82** (1999) 2467–2470, arXiv:hep-ph/9809480.

[194] A. Bodek and U. K. Yang, "Modeling deep inelastic cross-sections in the few GeV region," *Nucl. Phys. B Proc. Suppl.* **112** (2002) 70–76, arXiv:hep-ex/0203009.

[195] A. Bodek and U. K. Yang, "Modeling neutrino and electron scattering inelastic cross-sections in the few GeV region with effective LO PDFs TV Leading Order," in *2nd International Workshop on Neutrino-Nucleus Interactions in the Few GeV Region*. 8, 2003. arXiv:hep-ex/0308007.

[196] A. Bodek, I. Park, and U.-k. Yang, "Improved low Q**2 model for neutrino and electron nucleon cross sections in few GeV region," *Nucl. Phys. B Proc. Suppl.* **139** (2005) 113–118, arXiv:hep-ph/0411202.

[197] A. Bodek and U.-k. Yang, "Axial and Vector Structure Functions for Electron- and Neutrino- Nucleon Scattering Cross Sections at all Q^2 using Effective Leading order Parton Distribution Functions," arXiv:1011.6592 [hep-ph].

[198] A. Bodek, U. K. Yang, and Y. Xu, "Inelastic Axial and Vector Structure Functions for Lepton-Nucleon Scattering 2021 Update," arXiv:2108.09240 [hep-ph].

[199] M. Glück, E. Reya, and A. Vogt, "Dynamical parton distributions revisited," *Eur. Phys. J. C* **5** (1998) 461–470, arXiv:hep-ph/9806404.

[200] C. Andreopoulos *et al.*, "The GENIE Neutrino Monte Carlo Generator," *Nucl. Instrum. Meth. A* **614** (2010) 87–104, arXiv:0905.2517 [hep-ph].

[201] C. Andreopoulos, C. Barry, S. Dytman, H. Gallagher, T. Golan, R. Hatcher, G. Perdue, and J. Yarba, "The GENIE Neutrino Monte Carlo Generator: Physics and User Manual," arXiv:1510.05494 [hep-ph].

[202] **DUNE** Collaboration, B. Abi *et al.*, "Deep Underground Neutrino Experiment (DUNE), Far Detector Technical Design Report, Volume II: DUNE Physics," arXiv:2002.03005 [hep-ex].

[203] J. L. Feng, I. Galon, F. Kling, and S. Trojanowski, "ForwArd Search ExpeRiment at the LHC," *Phys. Rev. D* **97** no. 3, (2018) 035001, arXiv:1708.09389 [hep-ph].

[204] **SHiP** Collaboration, C. Ahdida *et al.*, "SND@LHC," arXiv:2002.08722 [physics.ins-det].

[205] L. A. Anchordoqui *et al.*, "The Forward Physics Facility: Sites, experiments, and physics potential," *Phys. Rept.* **968** (2022) 1–50, arXiv:2109.10905 [hep-ph].

[206] J. L. Feng *et al.*, "The Forward Physics Facility at the High-Luminosity LHC," arXiv:2203.05090 [hep-ex].

[207] A. Candido, F. Hekhorn, and G. Magni, "N3pdf/yadism: Fonll-b," Feb., 2022. https://doi.org/10.5281/zenodo.6285149.

[208] M. L. Mangano *et al.*, "Physics at the front-end of a neutrino factory: A quantitative appraisal," arXiv:hep-ph/0105155.

[209] J. M. Conrad, M. H. Shaevitz, and T. Bolton, "Precision measurements with high-energy neutrino beams," *Rev. Mod. Phys.* **70** (1998) 1341–1392, arXiv:hep-ex/9707015.

[210] J. A. Formaggio and G. P. Zeller, "From eV to EeV: Neutrino Cross Sections Across Energy Scales," *Rev. Mod. Phys.* **84** (2012) 1307–1341, arXiv:1305.7513 [hep-ex].

[211] D. Casper, "The Nuance neutrino physics simulation, and the future," *Nucl. Phys. B Proc. Suppl.* **112** (2002) 161–170, arXiv:hep-ph/0208030.

[212] S. Moch, M. Rogal, and A. Vogt, "Differences between charged-current coefficient functions," *Nucl. Phys.* **B790** (2008) 317–335, arXiv:0708.3731 [hep-ph].

[213] S. Moch, J. A. M. Vermaseren, and A. Vogt, "Third-order QCD corrections to the charged-current structure function F(3)," *Nucl. Phys. B* **813** (2009) 220–258, arXiv:0812.4168 [hep-ph].

[214] J. Gao, "Massive charged-current coefficient functions in deep-inelastic scattering at NNLO and impact on strange-quark distributions," *JHEP* **02** (2018) 026, arXiv:1710.04258 [hep-ph].

[215] D. J. Gross and C. H. Llewellyn Smith, "High-energy neutrino - nucleon scattering, current algebra and partons," *Nucl. Phys. B* **14** (1969) 337–347.

[216] **ArgoNeuT** Collaboration, C. Anderson *et al.*, "First Measurements of Inclusive Muon Neutrino Charged Current Differential Cross Sections on Argon," *Phys. Rev. Lett.* **108** (2012) 161802, arXiv:1111.0103 [hep-ex].

[217] **MicroBooNE** Collaboration, P. Abratenko *et al.*, "First Measurement of Inclusive Muon Neutrino Charged Current Differential Cross Sections on Argon at E_ν ∼0.8 GeV with the MicroBooNE Detector," *Phys. Rev. Lett.* **123** no. 13, (2019) 131801, arXiv:1905.09694 [hep-ex].

[218] **T2K** Collaboration, K. Abe *et al.*, "First measurement of the ν_μ charged-current cross section on a water target without pions in the final state," *Phys. Rev. D* **97** no. 1, (2018) 012001, arXiv:1708.06771 [hep-ex].

[219] **MINERvA** Collaboration, J. Mousseau *et al.*, "Measurement of Partonic Nuclear Effects in Deep-Inelastic Neutrino Scattering using MINERvA," *Phys. Rev. D* **93** no. 7, (2016) 071101, arXiv:1601.06313 [hep-ex].

[220] **BEBC WA59** Collaboration, K. Varvell *et al.*, "Measurement of the Structure Functions F2 and Xf3 and Comparison With QCD Predictions Including Kinematical and Dynamical Higher Twist Effects," *Z. Phys. C* **36** (1987) 1.

[221] E. Oltman *et al.*, "Nucleon structure functions from high energy neutrino interactions," *Z. Phys. C* **53** (1992) 51–71.

[222] **CHARM** Collaboration, F. Bergsma *et al.*, "Experimental Study of the Nucleon Structure Functions and of the Gluon Distribution from Charged Current Neutrino and anti-neutrinos Interactions," *Phys. Lett. B* **123** (1983) 269.

[223] **CHORUS** Collaboration, G. Onengut *et al.*, "Measurement of nucleon structure functions in neutrino scattering," *Phys. Lett.* **B632** (2006) 65–75.

[224] H. Abramowicz *et al.*, "Measurement of ν and $\bar{\nu}$ structure functions in hydrogen and iron," *Z. Phys. C* **25** (1984) 29–43.

[225] J. P. Berge *et al.*, "A Measurement of Differential Cross-Sections and Nucleon Structure Functions in Charged Current Neutrino Interactions on Iron," *Z. Phys. C* **49** (1991) 187–224.

[226] **NuTeV** Collaboration, M. Tzanov *et al.*, "Precise measurement of neutrino and anti-neutrino differential cross sections," *Phys. Rev. D* **74** (2006) 012008, arXiv:hep-ex/0509010.

[227] **New Muon** Collaboration, M. Arneodo *et al.*, "Accurate measurement of f_2^d/f_2^p and $r_d - r_p$," *Nucl. Phys.* **B487** (1997) 3–26, arXiv:hep-ex/9611022.

[228] **NuTeV** Collaboration, M. Goncharov *et al.*, "Precise measurement of dimuon production cross-sections in ν_μFe and $\bar{\nu}_\mu$Fe deep inelastic scattering at the Tevatron," *Phys. Rev.* **D64** (2001) 112006, arXiv:hep-ex/0102049 [hep-ex].

[229] D. A. Mason, "Measurement of the strange - antistrange asymmetry at nlo in qcd from nutev dimuon data,". FERMILAB-THESIS-2006-01.

[230] **European Muon** Collaboration, J. J. Aubert *et al.*, "Production of charmed particles in 250-gev μ^+ - iron interactions," *Nucl. Phys.* **B213** (1983) 31–64.

[231] **FNAL E866/NuSea** Collaboration, R. S. Towell *et al.*, "Improved measurement of the anti-d/anti-u asymmetry in the nucleon sea," *Phys. Rev.* **D64** (2001) 052002, arXiv:hep-ex/0103030 [hep-ex].

[232] G. Moreno *et al.*, "Dimuon production in proton - copper collisions at $\sqrt{s}$ = 38.8-GeV," *Phys. Rev.* **D43** (1991) 2815–2836.

[233] **CDF** Collaboration, 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.* **B692** (2010) 232–239, arXiv:0908.3914 [hep-ex].

[234] **D0** Collaboration, 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 $\sqrt{s}$=1.96-TeV," *Phys. Rev.* **D76** (2007) 012003, arXiv:hep-ex/0702025.

[235] **D0** Collaboration, 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.* **D88** (2013) 091102, arXiv:1309.2591 [hep-ex].

[236] **D0** Collaboration, V. M. Abazov *et al.*, "Measurement of the electron charge asymmetry in $\boldsymbol{p\bar{p} \to W + X \to e\nu + X}$ decays in $\boldsymbol{p\bar{p}}$ collisions at $\boldsymbol{\sqrt{s} = 1.96}$ TeV," *Phys. Rev.* **D91** no. 3, (2015) 032007, arXiv:1412.2862 [hep-ex]. [Erratum: Phys. Rev.D91,no.7,079901(2015)].

[237] **ATLAS** Collaboration, G. Aad *et al.*, "Measurement of the low-mass Drell-Yan differential cross section at $\sqrt{s} = 7$ TeV using the ATLAS detector," *JHEP* **06** (2014) 112, arXiv:1404.1212 [hep-ex].

[238] **ATLAS** Collaboration, 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.* **B725** (2013) 223, arXiv:1305.4192 [hep-ex].

[239] **ATLAS** Collaboration, 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.* **D85** (2012) 072004, arXiv:1109.5141 [hep-ex].

[240] **CMS** Collaboration, 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.* **109** (2012) 111806, arXiv:1206.2598 [hep-ex].

[241] **CMS** Collaboration, S. Chatrchyan *et al.*, "Measurement of the muon charge asymmetry in inclusive pp to WX production at $\sqrt{s} = 7$ TeV and an improved determination of light parton distribution functions," *Phys.Rev.* **D90** no. 3, (2014) 032004, arXiv:1312.6283 [hep-ex].

[242] **CMS** Collaboration, 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* **1312** (2013) 030, arXiv:1310.7291 [hep-ex].

[243] **LHCb** Collaboration, R. Aaij *et al.*, "Measurement of the cross-section for $Z \to e^+e^-$ production in *pp* collisions at $\sqrt{s} = 7$ TeV," *JHEP* **1302** (2013) 106, arXiv:1212.4620 [hep-ex].

[244] **LHCb** Collaboration, R. Aaij *et al.*, "Measurement of the forward Z boson production cross-section in *pp* collisions at $\sqrt{s} = 7$ TeV," *JHEP* **08** (2015) 039, arXiv:1505.07024 [hep-ex].

[245] **CMS** Collaboration, 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.* **B749** (2015) 187–209, arXiv:1504.03511 [hep-ex].

[246] **LHCb** Collaboration, R. Aaij *et al.*, "Measurement of forward Z $\to e^+e^-$ production at $\sqrt{s} = 8$ TeV," *JHEP* **05** (2015) 109, arXiv:1503.00963 [hep-ex].

[247] **LHCb** Collaboration, R. Aaij *et al.*, "Measurement of forward W and Z boson production in *pp* collisions at $\sqrt{s} = 8$ TeV," *JHEP* **01** (2016) 155, arXiv:1511.08039 [hep-ex].

[248] **LHCb** Collaboration, R. Aaij *et al.*, "Measurement of forward $W \to e\nu$ production in *pp* collisions at $\sqrt{s} = 8$ TeV," *JHEP* **10** (2016) 030, arXiv:1608.01484 [hep-ex].

[249] **CMS** Collaboration, S. Chatrchyan *et al.*, "Measurement of associated W + charm production in pp collisions at $\sqrt{s} = 7$ TeV," *JHEP* **02** (2014) 013, arXiv:1310.1138 [hep-ex].

[250] **ATLAS** Collaboration, 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.* **C76** no. 5, (2016) 291, arXiv:1512.02192 [hep-ex].

[251] **CMS** Collaboration, A. M. Sirunyan *et al.*, "Measurement of the inclusive $\mathrm{t\bar{t}}$ cross section in pp collisions at $\sqrt{s} = 5.02$ TeV using final states with at least one charged lepton," *JHEP* **03** (2018) 115, arXiv:1711.03143 [hep-ex].

[252] **ATLAS** Collaboration, G. Aad *et al.*, "Measurement of the $t\bar{t}$ production cross-section using $e\mu$ events with b-tagged jets in pp collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS detector," *Eur. Phys. J.* **C74** no. 10, (2014) 3109, arXiv:1406.5375 [hep-ex]. [Addendum: Eur. Phys. J.C76,no.11,642(2016)].

[253] S. Spannagel, "Top quark mass measurements with the cms experiment at the lhc," *PoS* **DIS2016** (2016) 150, arXiv:1607.04972 [hep-ex].

[254] **CMS** Collaboration, V. Khachatryan *et al.*, "Measurement of the top quark pair production cross section in proton-proton collisions at $\sqrt{(}s) = 13$ tev," *Phys. Rev. Lett.* **116** no. 5, (2016) 052002, arXiv:1510.05302 [hep-ex].

[255] **ATLAS** Collaboration, G. Aad *et al.*, "Measurements of top-quark pair differential cross-sections in the lepton+jets channel in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector," *Eur. Phys. J.* **C76** no. 10, (2016) 538, arXiv:1511.04716 [hep-ex].

[256] **CMS** Collaboration, V. Khachatryan *et al.*, "Measurement of the differential cross section for top quark pair production in pp collisions at $\sqrt{s} = 8\,\mathrm{TeV}$," *Eur. Phys. J.* **C75** no. 11, (2015) 542, arXiv:1505.04480 [hep-ex].

[257] **ATLAS** Collaboration, G. Aad *et al.*, "Measurement of the inclusive jet cross-section in proton-proton collisions at $\sqrt{s} = 7$ TeV using 4.5 fb^{-1} of data with the ATLAS detector," *JHEP* **02** (2015) 153, arXiv:1410.8857 [hep-ex]. [Erratum: JHEP09,141(2015)].

[258] **CMS** Collaboration, S. Chatrchyan *et al.*, "Measurement of the Ratio of Inclusive Jet Cross Sections using the Anti-k_T Algorithm with Radius Parameters R=0.5 and 0.7 in pp Collisions at $\sqrt{s} = 7$ TeV," *Phys. Rev. D* **90** no. 7, (2014) 072006, arXiv:1406.0324 [hep-ex].

[259] L. Demortier, *Proceedings, PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN,Geneva, Switzerland 17-20 January 2011*, ch. Open Issues in the Wake of Banff 2011. 2011.