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 XXXVII

Faithful estimation of uncertainties in modern PDF extractions.

Settore Scientifico Disciplinare FIS/02

Supervisore: Prof. Stefano Forte

Coordinatore: Prof. Aniello Mennella

Tesi di Dottorato di:

Andrea Barontini

Anno Accademico 2024-2025

Commission of the final examination:

External Referee:
Prof. Giuliano Panico
Prof. Katarzyna Wichmann

External Member:
Prof. Fabio Maltoni
Prof. Katarzyna Wichmann

Internal Member:
Prof. Marco Zaro

Final examination:

Date 15/11/2024

Università degli Studi di Milano, Dipartimento di Fisica, Milano, Italy

MIUR subjects:

FIS/02 - Fisica Teorica, Modelli e Metodi Matematici

Abstract

The primary focus of this Ph.D. thesis is the precise determination and validation of the uncertainties associated with parton distribution functions (PDFs). We introduce and implement the theory covariance method within the NNPDF4.0 framework, enabling the incorporation of theoretical uncertainties into the PDF determination process. Additionally, we revisit and expand upon the closure tests framework, which serves as a tool for validating the PDF extraction methodology in a controlled environment. This framework is applied to a dataset that has been deliberately constructed to be inconsistent, allowing for a rigorous assessment of the methodology's robustness. Furthermore, we utilize this framework to validate the extraction of the strong coupling constant using the correlated replica method. Finally, we present a novel theoretical pipeline, which introduces several technical advancements and underpins all the results discussed in this thesis.

Contents

Introduction

With the Large Hadron Collider (LHC) currently on its Phase III, the experimental precision will significantly improve due to an expected luminosity of approximately $300\ \text{fb}^{-1}$. This LHC run is also projected to enhance its discovery potential, operating at a center-of-mass energy around 14 TeV. The new kinematic regions explored during this phase are likely to be pertinent to some of the most critical unresolved issues of the Standard Model. It is essential to maximize the potential of the forthcoming data to address these *beyond-the-standard-model* (BSM) questions. To achieve this is necessary for the theoretical prediction to match the experimental precision, which is now recognized to be at the percent level.

On the theoretical front, one of the primary sources of uncertainty arises from the depiction of the internal structure of the colliding hadrons, most commonly protons. This depiction is encapsulated in the *Parton Distribution Functions* (PDFs). These functions are essential for interpreting any hadron collision event, as they describe the constituent particles of the proton, known as *partons*. Since PDFs are connected to the low-energy dynamics occurring inside the proton, they cannot be determined within the framework of perturbative Quantum ChromoDynamics (QCD), complicating their accurate determination.

While alternative methods exist, the most prevalent approach to extracting PDFs involves leveraging their *universality*, which means that they remain the same across different collision processes. This universality allows for the extraction of a set of PDFs from a limited dataset, which, in theory, can predict any other proton collision event[1]. In particular, thanks to the *collinear factorization theorem*, a longitudinal cross-section σ for a certain process can be written as the convolution

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

where f is the PDF and $\hat{\sigma}$, called *partonic cross-section*, describes the high-energy dynamics happening between the partons in the collision. Given that the partonic cross-sections are related to high-energy dynamics, they can be calculated within the framework of perturbative QCD. Consequently, by utilizing a dataset that measures the value of σ for a set

[1] In practice, this is not straightforward, as various complications can undermine PDF universality. For instance, the dataset used for determination might not be comprehensive enough, potentially failing to constrain certain combinations and/or kinematic regions of the PDFs. These and other complications are discussed in more detail in chapter 2.

of processes and the corresponding partonic cross-sections $\hat{\sigma}$ computed through perturbation theory, one can solve the inverse problem to extract the PDFs. The quality of the PDF determination and the associated uncertainties will depend on three main factors: the experimental data, the theoretical predictions (of the partonic cross-sections), and the fitting methodology. While the experimental aspect will not be further discussed, the advancements in theoretical predictions and in the methodology form the central theme of this thesis.

Regarding the theoretical predictions, this thesis addresses a critical issue: how to estimate theoretical uncertainties, primarily arising from neglected higher-order terms in the perturbative series, and incorporate them into the PDF fits (chapter 2). This problem, previously overlooked, has now become essential for obtaining reliable PDF uncertainties and central values, given that theoretical errors are comparable to experimental errors.

On the methodology side, one of the most significant challenges is assessing the reliability of PDF uncertainties. This task is particularly complex for NNPDF, which employs a Neural Network approach that currently lacks a comprehensive theoretical foundation. Therefore, it becomes crucial to rigorously test and validate the resulting PDFs under controlled conditions. This validation can be achieved through a closure test, which evaluates the fit results using appropriate statistical metrics in an artificial environment where the correct answer is known. Due to its versatility, the closure test framework allows for testing the robustness of a fit in various scenarios, including cases where there may be inconsistencies in the experimental data or challenges in precise parameter estimation. These aspects are explored in detail in chapter 3.

All the results presented in this thesis necessitated technical improvements in the production of theoretical predictions, culminating in the development of an entirely new theory pipeline, detailed in chapter 4.

Outline of the thesis

Chapter 1: QCD and Parton Distribution Functions.

> We review the fundamental properties of *Quantum Chromodynamics* (QCD), the fundamental theory describing the strong interactions (sections 1.1 and 1.2). We focus on the parton model (section 1.3) and, in particular, on the *Parton Distribution Functions* (PDFs), which are the common thread that connects all the topics covered in this thesis. We also provide some details on the different ways in which heavy quarks effects can be taken into account (section 1.4). This introduction is mainly based on [1–3].

Chapter 2: Inclusion of theory errors in PDF fitting.

> We present the first NNLO PDF extraction with inclusion of theory errors due to missing higher orders (MHO), based on the methodology first introduced in [4]. We also adapt this methodology to produce an approximate N3LO PDF set where different sources of theory errors, due to missing or incomplete theoretical calculations, are taken into account. This chapter is based on [5–7].

Chapter 3: Validation of the methodology: Closure Tests.

> We provide a complete description of the *Closure Test* tool, a validation framework already adopted in [8] (section 3.1). We then discuss the statistical estimators that

can be used in the context of a *multiclosure test* to assess the quality of a fitting methodology and we propose some improved variants (section 3.1.1). Making use of the improved estimators, we test the NNPDF4.0 methodology on data that are inconsistent by design, i.e. data whose nominal uncertainties are smaller than their real uncertainties (section 3.2). We also make use of the closure tests framework to validate our estimation of the strong coupling α_s (section 3.3). Most of what is discussed in this chapter is based on [8–10] [11]

Chapter 4: Technical Improvements: The Pineline.

We present the *Pineline*, a new set of tools, adopted by NNPDF as the theory predictions production pipeline, whose goal is to standardize and make more efficient the process of producing high-energy theory predictions (sections 4.1 and 4.2). We also present a specific example in which we show that adopting the Pineline is both simple and advantageous from a performance point of view. This chapter is mainly based on [12].

A substantial part of the research presented in this thesis was conducted in collaboration with colleagues from the NNPDF collaboration. Throughout the presentation of results, emphasis has been placed on areas where the author believes their contributions have been particularly noteworthy. Unless explicitly stated otherwise in the captions, all figures presented in this thesis have been generated by the author or have previously appeared in publications co-authored by the author.

Also, note the colors of the citations in the text: the cyan is used for citation to paper co-authored by the author, while green is used for all the others.

The following publications co-authored by the author are not discussed in this thesis:

- *NNPDF4.0 aN3LO PDFs with QED corrections* [7].
- *Photons in the proton: implications for the LHC* [13].

CHAPTER 1

QCD and Parton Distribution Functions

In this chapter, Quantum ChromoDynamics (QCD), the model that is currently used to describe the strong interactions happening inside the hadrons, is described in some details. QCD was first introduced in the 1960s and, since then, its predictive power was confirmed by many experiments, making it the main tool for the computation of theoretical predictions at the hadron colliders.

It is the theory of quarks, gluons and their interactions and it is a gauge theory, like Quantum Electrodynamics (QED). Both theories share several similarities; for example, just as electrons carry electric charge, quarks carry the QCD charge, referred to as color. However, unlike the single type of electric charge, color comes in three types: red, green, and blue. Additionally, while photons are electrically neutral, gluons are not color-neutral. Instead, gluons can be thought of as carrying a combination of color and anti-color charges, resulting in eight distinct combinations. These and other differences stem from the fact that QCD is a non-abelian gauge theory, unlike QED. This fundamental distinction leads to many unique features in QCD that are absent in QED, as will be illustrated in the following sections.

Another significant distinction between QCD and QED lies in their respective coupling behaviors. The strong coupling constant, α_s, approaches zero at high energy scales, a phenomenon known as *asymptotic freedom*. In contrast, the electromagnetic coupling constant, α_{EM}, increases with rising energy scales. At the energy levels of the LHC, α_s varies from approximately 0.08 at a scale of 5 TeV, an energy range conducive to the application of perturbation theory, to about 1 at 0.5 GeV. The high value of α_s at lower energy scales facilitates the aggregation of quarks into color-neutral states, known as hadrons, a phenomenon referred to as *confinement*. However, this high value also makes perturbation theory ineffective for making predictions at low energies. To address this issue, modern high-energy scattering predictions are computed using the improved parton model, an enhancement of Feynman's original parton model that incorporates QCD corrections.

In the following section (1.1), the QCD lagrangian and its group structure are described. In section 1.2, the fundamentals aspects of perturbative QCD are recalled, as well as UV and IR divergences treatment. In section 1.3, the improved parton model is described is some details, focusing on the introduction of the Parton Distribution Functions (PDFs) and on their evolution equations. Finally, in section 1.4, some details about the treatment of the heavy quarks effects are provided. Most of the discussion of this chapter is based on [1–3].

1.1 Lagrangian and group structure

The fields entering the QCD Lagrangian are the quark fields, ψ_a, which are spinors (since quarks are fermions) that carry the color index a ranging from 1 to 3, and the gluon fields, $\mathcal{A}^C_\mu$, which are vector fields that carry the color index C ranging from 1 to 8. The theory is constructed to be gauge invariant under local $SU(3)$ symmetry group, i.e. invariant under the field transformations

$$\begin{aligned} \psi_a &\to e^{i\theta^C(x)t^C_{ab}}\psi_b \\ \mathcal{A}^C_\mu t^C &\to e^{i\theta^D(x)t^D}\left(\mathcal{A}^C_\mu t^C - \frac{1}{g_s}\partial_\mu\theta^C(x)t^C\right)e^{-i\theta^E(x)t^E}\,, \end{aligned} \tag{1.1}$$

where $\theta^C(x)$ are eight arbitrary real functions of the space-time position x, t^C are the eight $SU(3)$ group generators, the index μ is a Lorentz index and, as in the rest of this thesis, the repeated indices are summed over, following Einstein notation.

In eq. (1.1) the flavour index has been kept implicit, as it will be done in the rest of this chapter. There are six quark flavours, which can be categorized into three families based on their physical masses and electric charges. The first family includes the up (u) and down (d) quarks. These are the lightest, with masses of approximately 2 MeV and 5 MeV, respectively, and electric charges of $e_u = 2/3$ and $e_d = -1/3$. The second family consists of the charm (c) and strange (s) quarks, with masses around 1 GeV and 100 MeV, respectively. The third family comprises the top (t) and bottom (b) quarks, with significantly larger masses of approximately 170 GeV and 5 GeV. Despite having the same electric charge structure, the quark masses increase substantially across these families. The underlying physical reason for the existence of multiple essentially equivalent quark families, as well as the possibility of additional families, remains unknown.

While quark fields transform according to the fundamental representation of $SU(3)$, gluon fields transform according to the *adjoint* representation. However, the parameters $\theta^C(x)$ are functions of the space-time coordinate x. Ensuring the invariance of the theory under a local group transformation is the conventional method for constructing gauge theories such as QCD and QED.

The $SU(3)$ group generators t^C are hermitian matrices which have to follow the so-called Lie algebra of the group

$$[t^A, t^B] = if^{ABC}t^C\,, \tag{1.2}$$

where f^{ABC} is a completely antisymmetric tensor whose entries are called *structure constants* of $SU(3)$. It is important to note that the non-abelian nature of the $SU(3)$ group is indicated by the fact that f^{ABC} is non-zero. Adopting the Gell-man convention

$$\mathrm{Tr}(t^A t^B) = T_R\delta_{AB} \quad T_R = \frac{1}{2}\,, \tag{1.3}$$

it is possible to write the generators of $SU(3)$ esplicitly as

$$t^1 = \frac{1}{2}\begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} t^2 = \frac{1}{2}\begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} t^3 = \frac{1}{2}\begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{pmatrix} t^4 = \frac{1}{2}\begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$
$$t^5 = \frac{1}{2}\begin{pmatrix} 0 & 0 & -i \\ 0 & 0 & 0 \\ i & 0 & 0 \end{pmatrix} t^6 = \frac{1}{2}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} t^7 = \frac{1}{2}\begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix} t^8 = \frac{1}{2\sqrt{3}}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -2 \end{pmatrix}.$$

From this explicit representation of the generators, we can observe that there are two diagonal matrices, t^3 and t^8. Since they are diagonal, they commute with each other, indicating that the rank of $SU(3)$ is 2 (generally, the rank of $SU(N)$ is $N-1$). This is significant because the rank corresponds to the number of *Casimir* operators of the group, which are associated with important properties of particles.

The generators in the adjoint representation are instead 8×8 matrices defined as

$$(T^A)^{BC} = -if^{ABC}\,. \tag{1.4}$$

In QCD computations there are some color related quantities which usually appear in the cross-sections. In particular, when quarks are involved one recurring combination is

$$\sum_A t^A_{ab} t^A_{bc} = C_F \delta_{ac}\,, \quad C_F = \frac{N^2-1}{2N} = \frac{4}{3}\,, \tag{1.5}$$

while, when gluons are involved, we often have

$$\mathrm{Tr}(T^A T^B) = C_A \delta^{AB}\,, C_A = N = 3\,. \tag{1.6}$$

Having discussed the group structure of QCD in some detail, we will now focus the remainder of this section on the Lagrangian formulation of the theory. The QCD Lagrangian can be written as

$$\mathcal{L}_{\text{QCD}} = \mathcal{L}_q + \mathcal{L}_{\text{G}} + \mathcal{L}_{\text{quantum}}\,, \tag{1.7}$$

where $\mathcal{L}_q$ is the quark part, $\mathcal{L}_{\text{G}}$ the purely gluonic part and $\mathcal{L}_{\text{quantum}}$ is needed for the quantization of the theory. The quark part is

$$\mathcal{L}_q = \overline{\psi}_a (i \not{D}_{ab} - m\delta_{ab})\psi_b\,, \tag{1.8}$$

where

$$\not{D}_{ab} = \gamma^\mu \partial_\mu \delta_{ab} + i g_s \gamma^\mu t^C_{ab} \mathcal{A}^C_\mu\,, \tag{1.9}$$

is the *covariant derivative* used, in place of the standard derivative, to promote the global $SU(3)$ symmetry to the local version of eq. (1.1). The covariant derivative includes an additional component compared to the standard derivative, which governs the interactions between quarks and gluons. As evident from eq. (1.9), these interactions do not preserve the diagonal nature of color states, implying that when a gluon interacts with a quark, the quark typically changes its color.

The second part contains the dynamics and the interactions of the gluons with them-

selves. Defining the *field-strength tensor* as

$$F^A_{\mu\nu} = \partial_\mu \mathcal{A}^A_\nu - \partial_\nu \mathcal{A}^A_\mu - g_s f^{ABC} \mathcal{A}^B_\mu \mathcal{A}^C_\nu \,, \tag{1.10}$$

the gluonic lagrangian can be written as

$$\mathcal{L}_{\rm G} = -\frac{1}{4} \mathrm{Tr}(F_{\mu\nu} F^{\mu\nu}) \,. \tag{1.11}$$

It is crucial to observe that the term $g_s f^{ABC} \mathcal{A}^B_\mu \mathcal{A}^C_\nu$ represents one of the primary distinctions from Quantum Electrodynamics (QED). This term arises due to the non-abelian nature of QCD and plays a pivotal role, as will be discussed in the subsequent section, in the formation of three- and four-gluon vertices, phenomena absent for the photons in QED.

The final component of the Lagrangian is not gauge-invariant and must be included for proper quantization of the theory. Due to the gauge symmetry of the theory, there exist redundant degrees of freedom that render canonical quantization inadequate for QCD. To address this issue, the *Faddeev-Popov* method involves eliminating these redundant degrees of freedom by introducing a gauge-fixing term into the Lagrangian. It reads

$$\mathcal{L}_{g.f.} = \frac{1}{2\xi} \sum_A |\partial_\mu \mathcal{A}^\mu_A|^2 \,, \tag{1.12}$$

from which one gets, for example, the Feynman gauge setting $\xi = 1$ or the Landau gauge for $\xi = 0$. Additionally, to complete the quantization part of the lagrangian, we must add the *ghost* term

$$\mathcal{L}_{\rm ghost} = \overline{\eta}_A \partial_\mu D^\mu_{AB} \eta_B \,, \tag{1.13}$$

which introduces another kind of fields, called *ghost* fields, that are complex scalar fields but that obey Fermi statistics. This property makes these degrees of freedom unphysical and hence not directly associated with observable particles. However, for the computation of measurable quantities, these degrees of freedom generally need to be taken into account.

1.2 Perturbative QCD

Perturbation theory involves expressing a given observable as a series expansion in a small parameter. In Quantum Chromodynamics (QCD) the strong coupling constant α_s serves this role. Therefore, an observable quantity computed using perturbation theory up to α_s^n order in QCD can be expressed as:

$$F = f^{(0)} + f^{(1)} \alpha_s + f^{(2)} \alpha_s^2 + \cdots + f^{(n)} \alpha_s^n + \mathcal{O}(\alpha_s^{n+1}) \,. \tag{1.14}$$

Notice the round parenthesis notation denoting the perturbative order that will be adopted in the rest of the thesis.

For perturbation theory to provide reliable predictions, it is essential not only that the expansion parameter itself is small, but also that the coefficients $f^{(i)}$ do not include terms that grow excessively with i. In other words, the critical condition for the reliability of perturbation theory is that $f^{(i)} \alpha_s^i$ constitutes a parametrically decreasing function in i. This requirement imposes constraints on both the coupling constant and the coefficients

themselves.

Regarding the coupling constant, as mentioned earlier in this chapter, in Quantum Chromodynamics (QCD) it decreases with increasing energy scale of the process. This behavior is a consequence of renormalizing ultraviolet (UV) divergences (further details on running coupling and renormalization can be found in section 1.2.1). The value of the QCD coupling constant varies by up to an order of magnitude within commonly observed kinematic ranges, making perturbation theory not always reliable across all energy scales. Thus, in QCD, there exist distinct kinematic regions: the perturbative region, typically above 1 GeV, and the non-perturbative region. This dichotomy implies that standard techniques cannot reliably predict the low-energy internal dynamics of objects such as hadrons, nor processes involving them. However, the Feynman parton model provides a means to circumvent these limitations and make predictions for processes initiated by hadrons, such as those observed at the LHC. The parton model will be discussed in section 1.3.

Regarding the coefficients $f^{(i)}$, they are primarily computed using Feynman diagram techniques. Starting from the QCD Lagrangian discussed in the preceding section, one can derive the QCD Feynman rules (fig. 1.1), which govern the construction of Feynman diagrams associated with these coefficients $f^{(i)}$. It is noteworthy that the second and third rules in fig. 1.1 correspond precisely to the last term of eq. (1.10), highlighting them as significant differences between QCD and QED, as mentioned earlier. Once the coefficients $f^{(i)}$ are computed, it may occur that calculations involving multiple energy scales introduce logarithms of the ratios of these scales. In certain kinematic regions, these logarithmic terms can become significant and, depending on their exponents, can cause the coefficients $f^{(i)}$ to become too large for perturbation theory to be applicable. In such cases, it becomes necessary to resum these logarithmic contributions to all orders.

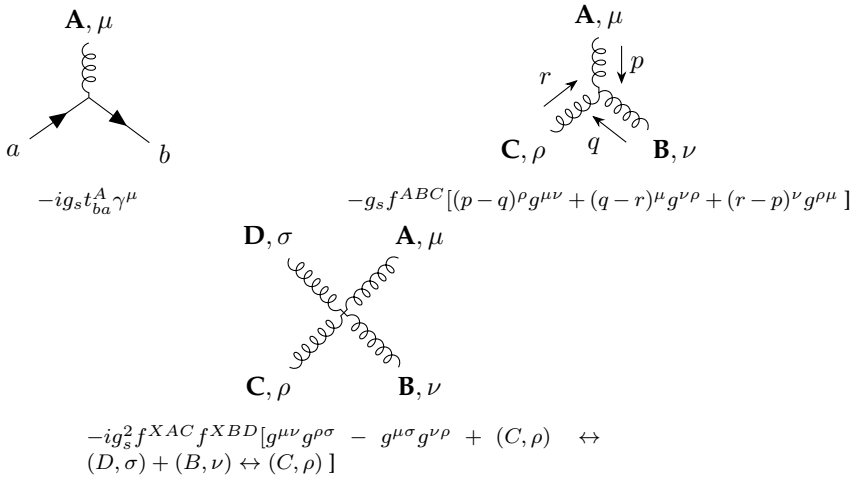


Figure 1.1: Interactions vertices of the Feynman rules of QCD

A class of such logarithms is composed by the *collinear logarithms*. These logarithmic terms arise from the phase-space integration of, for instance, the $g \to q\bar{q}$ splitting process in the limit where the final quarks become collinear. If the final quarks are treated as heavy, meaning their masses are non-zero, the logarithms take the form $\log^k\left(Q^2/m_q^2\right)$, where m_q is the mass of the quark, Q is the hard energy scale of the process, and the

power k is less than or equal to the perturbative order i of the coefficient $f^{(i)}$ being computed. If the quarks are treated as light, these collinear logarithms become infrared (IR) collinear divergences that must be regularized, for example, using dimensional regularization. In section 1.2.2 we discuss collinear divergences and, more in general, how IR divergences are treated, while, in section 1.4, we discuss how the resummation of collinear logarithms is performed, focusing on the Deep Inelastic Scattering (DIS) case.

1.2.1 UV divergences: running coupling

When computing a quantity in perturbation theory beyond leading order, phase-space integrals can become divergent in both the ultraviolet (UV) region, i.e. high energy, and the infrared (IR) region, i.e. low energy. The methods for addressing these divergences vary depending on their nature. In this and the subsequent section, we will examine the main aspects of their regularization. For a detailed description, the reader may refer to [14, 15].

The standard procedure to regularize UV divergences is renormalization. It involves redefining both the fields and the constant terms, such as the coupling, in the Lagrangian in a manner that allows them to absorb the infinities. In the QCD case this means to rescale quark and gluon fields, quark masses and the coupling as follows

$$\psi^b = \sqrt{Z}\psi\,, \quad \mathcal{A}^b = \sqrt{Z_3}\mathcal{A}\,, \quad m^b = \frac{Z_m}{Z_2}m\,, \quad g_s^b = \frac{Z_1}{Z_2\sqrt{Z_3}}g_s\,, \tag{1.15}$$

where we use the apex b to denote a *bare*, i.e. not renormalized yet, quantity.

In dimensional regularization, which is the standard method used to regularize IR divergences as well, the Lagrangian is constructed to have a dimension of $d = 4 - 2\epsilon$. Consequently, the energy dimensions of fields and constants are altered as

$$[\psi] = \frac{d-1}{2} = \frac{3}{2} - \epsilon \tag{1.16}$$

$$[\mathcal{A}] = \frac{d-2}{2} = 1 - \epsilon \tag{1.17}$$

$$[g_s] = d - 2[\psi] - [\mathcal{A}] = \frac{4-d}{2} = \epsilon\,. \tag{1.18}$$

In order to keep working with a dimensionless coupling, one may define

$$\alpha_s^b = \frac{(g_s^b)^2}{4\pi} = \frac{Z_1^2}{Z_2^2 Z_3}\alpha_s\tilde{\mu}^{2\epsilon}\,, \quad \tilde{\mu}^{2\epsilon} \equiv \frac{\mu^2 e^\gamma}{4\pi}\,, \tag{1.19}$$

where γ is the Euler's gamma and μ, called *renormalization scale*, is a fictitious scale in which the energy scale is retained. The second part of the last equation defines a particular scheme of renormalization called *modified minimal subtraction scheme* ($\overline{\text{MS}}$), which is the scheme adopted in this chapter.

After this redefinition, Z terms can be fixed, order by order in perturbation theory in such a way the observables are UV finite. However, by means of eq. (1.19), the bare coupling acquires an energy dependence which we must impose to vanish, as

$$\mu^2 \frac{d}{d\mu^2}\log\alpha_s^b = 0\,. \tag{1.20}$$

In turn, this equation leads to the *renormalization group equation* (RGE) of the renormalized coupling, that reads

$$\mu^2 \frac{d}{d\mu^2}\alpha_s(\mu^2) = \beta(\alpha_s(\mu^2)) \tag{1.21}$$

and that fixes, order by order in perturbation theory, the dependence of the renormalized coupling α_s on the renormalization scale μ. The β function appearing on the right hand side of the latest equation can be expressed as

$$\beta(\alpha_s(\mu^2)) = -\epsilon\alpha_s - (\beta_0\alpha_s^2(\mu^2) + \beta_1\alpha_s^3(\mu^2) + \mathcal{O}(\alpha_s^4))\,, \tag{1.22}$$

with the coefficients β_i computed in perturbation theory (currently they are known up to five loops [16]). In particular,

$$\beta_0 = \frac{11C_A - 4n_f T_F}{12\pi} = \frac{33 - 2n_f}{12\pi}\,, \tag{1.23}$$

that is positive for $n_f < 17$ (thus is positive in our case, given that, as far as we currently know, $n_f = 6$). The solution of eq. (1.21) at leading order, with $d = 4$, is

$$\alpha_s(\mu^2) = \frac{\alpha_s(\mu_0^2)}{1 + \alpha_s(\mu_0^2)\beta_0 \log \mu^2/\mu_0^2}\,, \tag{1.24}$$

where μ_0^2 is an arbitrary scale. It is then clear that, with $\beta_0 > 0$, the value of the running coupling decreases logarithmically to 0 as the energy scale of the process increases. As already mentioned, this property is called asymptotic freedom, and makes QCD an asymptotically free theory. Note that in eq. (1.24) we still need an initial condition in order to compute the value fo the running coupling at all energy scales μ^2. This initial condition is usually obtained from experiments, which quote the value of the running coupling at the mass of the Z boson, $\alpha_s(M_Z^2)$ (fig. 1.2).

In eq. (1.24), the fixed coupling still depends on the arbitrary scale μ_0. In certain cases, it may be desirable to eliminate this dependence. This is commonly achieved by replacing it with a dimensionful parameter Λ_{QCD}, which roughly corresponds to the energy scale at which the theory becomes strongly coupled. It allows to write eq. (1.24) as

$$\alpha_s(\mu^2) = \frac{1}{\beta_0 \log \frac{\mu^2}{\Lambda_{\text{QCD}^2}}}\,, \tag{1.25}$$

thanks to its definition

$$\log \frac{\mu^2}{\Lambda_{\text{QCD}}^2} = -\int_{\alpha_s(\mu^2)}^{\infty} \frac{dx}{\beta(x)}\,. \tag{1.26}$$

Its value is approximately 200 MeV, although its precise definition depends on the chosen renormalization scheme. Together with the renormalization group equations (RGE) describing the running of the coupling, Λ_{QCD} enables us to replace the dependence on the dimensionless parameter g_s, which, as we have just seen, is not a constant.

1.2.2 IR divergences

The other type of divergence encountered when computing a QCD observable is the IR divergence. It arises from the low-energy region of phase-space integrals and can be cat-

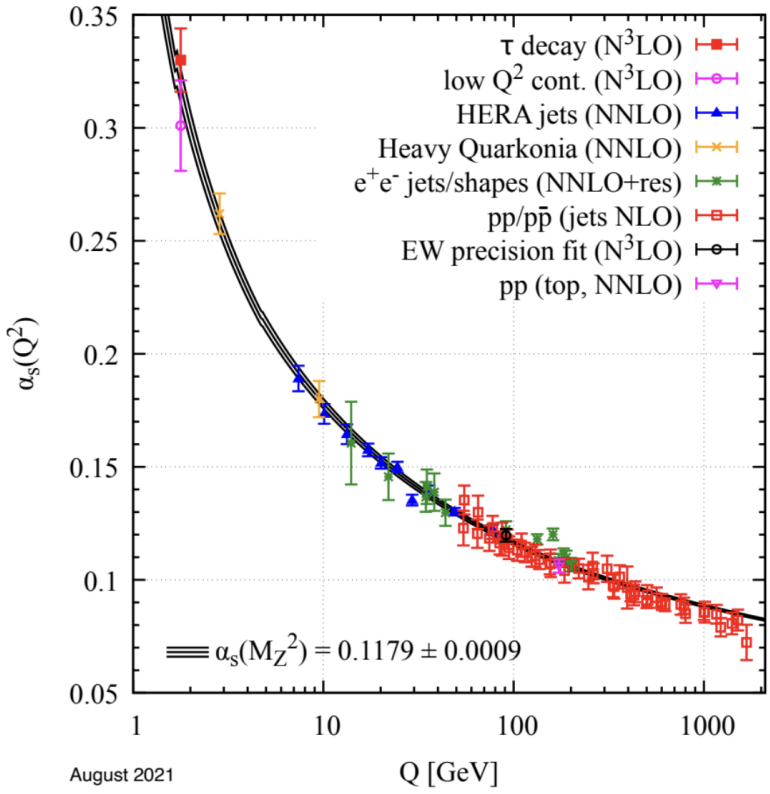


Figure 1.2: Running coupling measurements at different scales and different perturbative order. This figure is taken from [17].

egorized into two types: soft divergences, associated with the low energy of the particles, and collinear divergences, related to the collinearity between emitted and emitting particles. Both types can originate from either initial or final state particles, and the methods for their regularization differ significantly in these two cases, as a consequence of these two theorems [18]:

- *Bloch-Nordsieck* theorem: IR singularities cancel between real and virtual diagrams when summing up all resolution-indistinguishable final states at a certain perturbative order.
- *Kinoshita-Lee-Nauenberg* (KLN) theorem: mass singularities ($m \to 0$) of external particles (i.e. both initial and final) are cancelled if all mass-degenerate states are summed up.

The first theorem ensures the cancellation of both collinear and soft final-state divergences when all virtual and real diagrams at a given perturbative order are combined. However, it is essential to clarify the concept of resolution-indistinguishable final states. The key point is that real diagrams, which must be added to the virtual ones, include additional real emissions of QCD particles. Consequently, they do not share the same final state and, in principle, should not be considered together. Nonetheless, the crucial observation is that, in both the soft and collinear limits, the real-emission process

becomes experimentally (and theoretically) indistinguishable from the no-emission process, thereby justifying their combined consideration.

The second theorem indicates that, since summing over the *initial* degenerate states is not typically performed, the cancellation of initial-state divergences is not assured. However, it can be demonstrated that soft divergences do cancel in the initial state. Consequently, only the collinear initial-state divergences persist. A different method of regularization is required for these divergences.

The regularization of collinear divergences is based on the fact that it can be demonstrated that only the non-singular part of the cross-section is process-dependent, while the singular part is entirely universal. Specifically, it has been found that a quark emitting a gluon introduces a collinear divergence proportional, at $\mathcal{O}(\alpha_s)$, to the so-called Altarelli-Parisi quark-quark splitting function

$$P_{qq} = \frac{\alpha_s}{2\pi} C_F \left(\frac{1+z^2}{1-z} \right)_+ + \mathcal{O}(\alpha_s^2)\,, \tag{1.27}$$

where the *plus-distribution* is defined as

$$\int_0^1 dz f(z)[g(z)]_+ \equiv \int_0^1 [f(z) - f(1)]g(z)\,. \tag{1.28}$$

The other Altarelli-Parisi [19] splitting functions arise from the computation of the other types of splittings: a gluon remaining a gluon (P_{gg}), a gluon becoming a quark (P_{qg}), and a quark becoming a gluon (P_{gq}). It is important to emphasize that, while the splitting functions are process-independent, they are not scheme-independent. In the $\overline{\text{MS}}$ scheme they are

$$\begin{aligned}
P_{gg} &= \frac{\alpha_s}{4\pi}\left(4C_A \left[\frac{z}{(1-z)_+} + \frac{1-z}{z} + z(1-z) \right] + \frac{11C_A - 4T_F n_f}{6}\delta(1-z) \right) + \mathcal{O}(\alpha_s^2)\,, \\
P_{qg} &= \frac{\alpha_s}{2\pi} n_f [z^2 + (1-z)^2] + \mathcal{O}(\alpha_s^2)\,, \\
P_{gq} &= \frac{\alpha_s}{2\pi} C_F \frac{1+(1-z)^2}{z} + \mathcal{O}(\alpha_s^2)\,.
\end{aligned} \tag{1.29}$$

The universality of the splitting functions is the fundamental concept underlying the collinear factorization method, which is used to regularize collinear divergences. However, to understand its operation, it is first necessary to introduce the Feynman parton model. Consequently, the detailed presentation of this topic is deferred to the next section (1.3).

1.3 Parton Model

From section 1.2, it should be evident that computing a cross-section of a hadron-initiated process solely from first principles is impossible within standard perturbation theory. The reason is that, even if the center-of-mass energy is sufficiently high to fall within the perturbative region of QCD, the hadrons themselves are intrinsically low-energy objects, and thus their internal structure is governed by non-perturbative dynamics.

The initial solution to this problem was the parton model, developed by *Richard Feynman* in the late 1960s. Its fundamental concept is to factorize the cross-section of

a hadron-initiated process into two components: the cross-section of the high-energy process occurring between the partons (i.e. quarks and gluons), known as the *partonic cross-section*, and a process-independent part representing the probability distribution of extracting a particular parton from the hadron. This approach allows the first part to be computed using perturbation theory. However, a method to compute the second part, called the *parton distribution function* (PDF), is still required.

The key point, which is the central topic of this thesis and that will be discussed in detail in chapter 2, is that PDFs can be fitted from experimental data, and once determined from a particular process, they can be used for other processes due to their process-independent nature.

To clarify the discussion on the parton model, it is advantageous to specialize it on the case of Deep Inelastic Scattering (DIS), which is introduced in the following section. The discussion on the parton model applied to the DIS case is then resumed in section 1.3.2.

1.3.1 Deep Inelastic Scattering

Deep Inelastic Scattering (DIS) involves the collision of a lepton with a hadronic target, resulting in the destruction of the target. This contrasts with elastic or slightly inelastic scattering, where the target remains intact. DIS provides a precise method for testing Quantum Chromodynamics (QCD), as it allows the hadron (typically a proton) to be probed with a structureless particle, usually an electron. Historically, DIS experiments have been crucial for advancing the understanding of perturbative QCD. Even today, DIS measurements remain significant for the determination of parton distribution functions (PDFs). Examples of such measurements, that are currently used in modern PDF determinations, are those permormed at SLAC [20], BCDMS [21] and HERA (H1 [22] and ZEUS [23]). Given its importance, this section is dedicated to defining the kinematics and the observables associated with the DIS process.

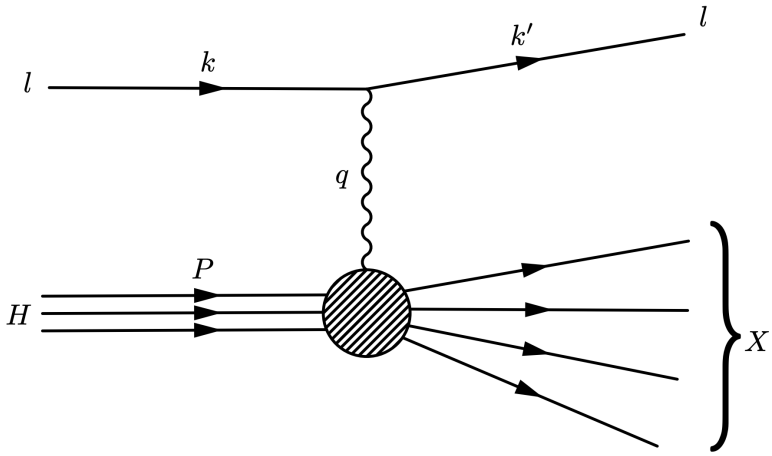


Figure 1.3: Schematic representation of the Deep Inleastic Scattering process of the charged lepton l with the hadron target H[3].

In DIS, a charged lepton with initial momentum k and final momentum k' scatters

off a hadron target of momentum P producing a final state X(fig. 1.3):

$$l(k) + H(P) \to l(k') + X\,. \tag{1.30}$$

For the rest of this section we will only consider the case of an electron scattering off a proton, through a virtual photon. This means we are only considering the *electromagnetic* (EM) contribution. If we considered also the contribution given by a Z boson mediator, we would fully describe the *neutral current* (NC) sector, as opposed to the *charged current* (CC) contribution that is mediated by $W^{\pm}$ bosons. This approximation is valid as long as the energy scale is well below the Z mass M_Z.

The centre-of-mass energy is

$$s = (P + k)^2\,, \tag{1.31}$$

and the invariant mass of the final state X is

$$W^2 = (P + q)^2\,. \tag{1.32}$$

We can then define the standard DIS kinematic variables

$$\begin{aligned} Q^2 &= -q^2\,, \\ x &= \frac{Q^2}{2P\cdot q}\,, \\ y &= \frac{P\cdot q}{P\cdot k} = \frac{Q^2}{xs}\,. \end{aligned} \tag{1.33}$$

The variable x, known as the *Bjorken scaling variable*, ranges between 0 and 1. At $x = 1$, it corresponds to elastic scattering. The deep inelastic scaling region is then characterized by $Q^2 \gg \Lambda^2_{\text{QCD}}$ for fixed and sufficiently small x.

It can be shown that the Feynman amplitude of DIS can be decomposed into a leptonic and a hadronic part, as

$$\frac{1}{4}\sum_{\text{spin}} |\mathcal{M}|^2 = \frac{e^4}{Q^4} L^{\mu\nu} h^X_{\mu\nu}\,, \tag{1.34}$$

with the *leptonic tensor* $L_{\mu\nu}$ that reads

$$L^{\mu\nu} = k^{\mu}k'^{\nu} + k'^{\mu}k^{\nu} - g^{\mu\nu}k\cdot k'\,. \tag{1.35}$$

The hadronic part can be also expressed in terms of an *hadronic tensor*, as

$$W_{\mu\nu} = \sum_X \int d\Phi h^X_{\mu\nu}\,, \tag{1.36}$$

where $d\Phi$ is the phase-space factor. By requiring Lorentz symmetry and gauge invari-

ance, we derive a general formulation of the hadronic tensor, expressed as

$$W_{\mu\nu}(P,q) = -\left(g_{\mu\nu} + \frac{q_\mu q_\nu}{q^2}\right)F_1(x,Q^2) + \frac{1}{P\cdot q}\left(P_\mu - q_\mu\frac{P\cdot q}{q^2}\right)\left(P_\nu - q_\nu\frac{P\cdot q}{q^2}\right)F_2(x,Q^2)\,. \tag{1.37}$$

The functions F_1 and F_2 are called *structure functions* and they are the main observables in the context of DIS [1].

We can also express the differential cross-section of the DIS process in terms of the structure functions, as

$$\frac{d\sigma}{dxdQ^2} = \frac{2\pi\alpha^2}{Q^4}\left[(1+(1-y)^2)F_T(x,Q^2) + \frac{2(1-y)}{x}F_L(x,Q^2)\right], \tag{1.38}$$

with $\alpha = e^2/(4\pi)$ and

$$\begin{aligned} F_L &= F_2 - 2xF_1\,, \\ F_T &= 2F_1\,. \end{aligned} \tag{1.39}$$

An example of the F_2 structure function as measured by the SLAC, BCDMS, H1 and

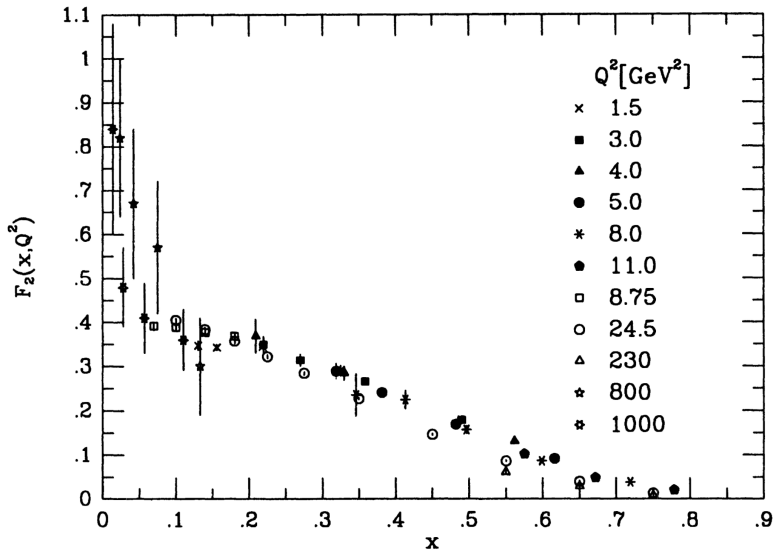


Figure 1.4: The F_2 structure function as measured by the SLAC, BCDMS, H1 and ZEUS collaborations for different values of the scale Q^2 [1].

ZEUS collaborations for different values of the scale Q^2 is shown in fig. 1.4.

[1] If one allows for parity violating effects in the charged current sector, a third structure function F_3 is needed to fully parametrize the hadronic tensor.

1.3.2 Parton Model in DIS case

For a process with a single hadron in the initial state, like *deep-inelastic scattering* (DIS), the parton model takes the form

$$F(Q^2) = \sum_q \int_0^1 f_q(x) C_q(x) dx + \mathcal{O}\left(\frac{\Lambda^2_{QCD}}{Q^2}\right), \tag{1.40}$$

where C_q is the partonic cross-section, called *coefficients functions* in DIS case, for parton q, f_q is the PDF for parton q, x is the fraction of inital hadron momentum carried by the parton and Q is the energy scale of the process.

The parton model formula of eq. (1.40), as explicitly written, is valid up to corrections of the order $\Lambda^2_{\text{QCD}}/Q^2$, thus applying only to energy scales within the perturbative region of QCD. Formally, this has been rigorously proved only for deep inelastic scattering (DIS); nonetheless, the parton model is currently utilized for all QCD processes. The graphical interpretation of eq. (1.40) is illustrated in fig. 1.5b (applied to the DIS case), while fig. 1.5a depicts the version for two initial hadrons.

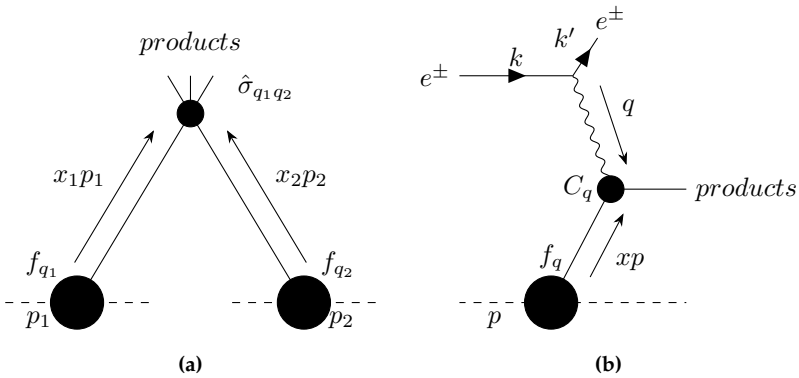


Figure 1.5: Graphical version of the LO parton model formulas applied in both the two initial hadrons (a) and the single-initial hadron (b) cases. The bubbles $\hat{\sigma}$ or C_q represent the partonic cross-sections, the other bubbles are the initial hadrons and the f_{q_i} are the PDFs.

However, this model, as for eq. (1.40), is valid only at leading order (LO), i.e. at the first order in perturbation theory, since it does not incorporate radiative QCD corrections. The inclusion of these corrections leads to what is known as the *improved parton model*, which enables the computation of observables beyond LO but alters the interpretation of parton distribution functions (PDFs) as probability distributions. The form of the improved parton model for deep inelastic scattering (DIS) is

$$F(Q^2, x) = \sum_q \int_x^1 \frac{dy}{y} f_q\left(\frac{x}{y}, \mu_F^2\right) C_q(y, Q^2/\mu_F^2, \alpha_s(\mu_R^2)) + \mathcal{O}\left(\frac{\Lambda_{\text{QCD}^2}}{\mu_F^2}\right). \tag{1.41}$$

As discussed in section 1.2.2, beyond LO the coefficients C_q contains unregularized IR collinear divergences. However, the PDFs, devoid of their distribution function interpretation, can also contain such divergences. Thus, to achieve a finite result from eq. (1.41), the PDFs are presumed to incorporate infrared (IR) divergences in a manner that offsets

those present in the coefficient functions. This forms the foundational concept of the collinear factorization method, which is elaborated upon in detail in section 1.3.3. As a consequence of the collinear factorization, PDFs acquire a dependence on a *factorization scale* μ_F, which will be also discussed in section 1.3.3.

For completeness, the parton model form in the two initial hadrons case is

$$\sigma_X(s, M_X) = \sum_{q_1 q_2} \int_0^1 dx_1 dx_2 f_{q_1}(x_1, \mu_F^2) f_{q_2}(x_2, \mu_F^2) \hat{\sigma}_{q_1 q_2 \to X}\left(x_1, x_2, \alpha_s(\mu_R^2), \frac{Q^2}{\mu_F^2}\right), \tag{1.42}$$

where $\hat{\sigma}_{q_1 q_2 \to X}$ is the partonic cross section for the partons q_1 and q_2 to produce a certain final state X.

1.3.3 Collinear Factorization and DGLAP evolution

Collinear factorization is the method used to regularize initial-state collinear infrared (IR) divergences in QCD, based on the factorization theorem [24–26]. Denoting the coefficients functions, which still include IR singular terms, as $\overline{C}_i$, the theorem in dimensional regularization can be expressed as

$$\overline{C}_i(x, \alpha_s, \epsilon) = \int_x^1 \frac{dz}{z} C_j\left(\frac{x}{z}, \alpha_s, \epsilon\right) \Gamma_{ij}(z, \alpha_s, \epsilon)\,, \tag{1.43}$$

where C_j are the IR regularized coefficients functions, i.e. they do not have *poles* in ϵ, and Γ_{ij} are called *collinear counter-terms* and are the objects containing the divergent terms.

Defining the *Mellin transform* of $f(z)$ as

$$f(N) \equiv \int_0^1 dz z^{N-1} f(z)\,, \tag{1.44}$$

the factorization theorem can be equivalently written in Mellin space as

$$\overline{C}_i(N, \alpha_s, \epsilon) = C_j(N, \alpha_s, \epsilon) \Gamma_{ij}(N, \alpha_s, \epsilon)\,. \tag{1.45}$$

Note the property of the Mellin space that converts a convolution, like the one in eq. (1.43), in a product. We will make use of Mellin space for most of this section.

These expressions imply that it is feasible to separate out the universal collinear singularities, which are not process-dependent (see section 1.2.2), contained within Γ_{ij}, from the raw coefficient functions such that the residual parts, C_i, are finite with respect to infrared (IR) regularization. Since observable are computed through the convolution of the coefficients functions with the PDFs, we can then absorb the collinear counter-terms Γ_{ij} in the PDF definition as

$$\begin{aligned} F(N, Q^2) &= \overline{C}(N, \alpha_s(\mu^2), \epsilon) \overline{f}_i(N, \epsilon) \\ &= C_j(N, \alpha_s(\mu^2), \epsilon) \Gamma_{ij}(N, \alpha_s(\mu^2), \epsilon) \overline{f}_i(N, \epsilon) \\ &= C_j(N, \alpha_s(\mu^2), \epsilon) f_j(N, (\mu^2)) + \mathcal{O}(\epsilon)\,, \end{aligned} \tag{1.46}$$

where we denoted with $\overline{f}_i$ the bare PDFs. Of course, the last line makes sense only if the divergences in Γ_{ij} effectively cancel out with those in the bare parton distribution functions (PDFs), as assumed. Although we will not show this explicitly, it can be proven that

this condition holds true, thereby allowing for the regularization of initial-state infrared (IR) divergences in this manner.

From the last line of eq. (1.46), it is clear that PDFs acquire an energy dependence through the dependence on μ^2 of the strong coupling α_s. Since the all-order PDFs are assumed to be scale independent, we can get the renormalization group equation for the PDFs

$$\mu_F^2 \frac{d}{d\mu_F^2} f_i(N, \mu_F^2) = -\gamma_{ij}(N, \alpha_s(\mu_F^2)) f_j(N, \mu_F^2)\,, \tag{1.47}$$

where the functions γ_{ij} are called *anomalous dimensions* and we have identified the scale μ^2 with the factorization scale μ_F^2. Note the convention of the minus sign in front of the anomalous dimensions on the right hand side of the equation. In x-space it reads

$$\mu_F^2 \frac{d}{d\mu_F^2} f_i(x, \mu_F^2) = \int_x^1 \frac{dz}{z} P_{ij}\left(\frac{x}{z}, \alpha_s(\mu_F^2)\right) f_j(z, \mu_F^2)\,, \tag{1.48}$$

that is expressed in terms of the Altarelli-Parisi splitting functions. Note that the anomalous dimensions are the Mellin transforms of the corresponding Altarelli-Parisi splitting functions.

These equations are known as the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) equations [19, 27] and they describe the evolution of parton distribution functions (PDFs) with respect to the factorization scale. The DGLAP equations enable us to define PDFs at an initial scale Q_0 and evolve them up to a higher scale Q relevant to a hard process. This capability allows PDFs parametrized at an initial scale to be constrained by processes occurring at different energy scales.

Due to the $SU(n_f)$ flavour symmetry present in QCD in the limit where quark masses are neglected, it is feasible to establish a basis of flavor states that remain invariant under evolution with the matrix P_{ij}. One approach to constructing such a basis involves partitioning the system of equations into two subsystems referred to as the singlet and non-singlet sectors. Given a system consisting of six quarks u, d, s, c, t, b, their anti-quarks, as well as the gluon, we define

$$f_i^{\pm} \equiv f_i \pm \overline{f}_i\,, \tag{1.49}$$

where we denoted with $\overline{f}_i$ the PDF of the anti-quark i. The f^- PDFs are called *valence* PDFs and they are usually denoted as

$$V_i \equiv f_i^-\,, \tag{1.50}$$

or with the subscript V, e.g. the valence up quark is $u_V = u - \bar{u}$. Some combination of the f^+ PDFs instead define the *triplet* states

$$\begin{aligned}
T_3 &= u^+ - d^+ \\
T_8 &= u^+ + d^+ - 2s^+ \\
T_{15} &= u^+ + d^+ + s^+ - 3c^+ \\
T_{24} &= u^+ + d^+ + s^+ + c^+ - 4b^+ \\
T_{35} &= u^+ + d^+ + s^+ + c^+ + b^+ - 5t^+\,,
\end{aligned} \tag{1.51}$$

where u, d, s, c, b, t are the PDFs of the corresponding quark flavour. The valence and triplet states comprise the so-called *non-singlet* sector. It can be demonstrated straight-

forwardly that the non-singlet sector decouple from the others, meaning it evolves according to the equation

$$\mu_F^2 \frac{d}{d\mu_F^2} f^{\mathrm{NS}}(x,\mu_F^2) = \frac{\alpha_s(\mu_F^2)}{2\pi} \int_x^1 \frac{dz}{z} P(z,\alpha_s) f^{\mathrm{NS}}\left(\frac{x}{z},\mu_F^2\right), \tag{1.52}$$

where valence states evolve with P_- and triplet states evolve with P_+. The definition of P_- and P_+ stem from the following definitions of *singlet* (S) and *non-singlet* (NS) components of the splitting functions,

$$\begin{aligned} P_{q_i q_k} &= \delta_{ik} P_{qq}^V + P_{qq}^S \\ P_{q_i \bar{q}_k} &= \delta_{ik} P_{q\bar{q}}^V + P_{q\bar{q}}^S \\ P_{\pm} &= P_{qq}^V \pm P_{q\bar{q}}^V . \end{aligned} \tag{1.53}$$

The *singlet* sector is instead composed by the gluon g and the singlet distribution

$$\Sigma = \sum_{i=1}^{n_f} f_i^+ , \tag{1.54}$$

which evolve according to the coupled equation

$$\mu_F^2 \frac{d}{d\mu_F^2} \begin{pmatrix} \Sigma(x,\mu_F^2) \\ g(x,\mu_F^2) \end{pmatrix} = \frac{\alpha_s(\mu_F^2)}{2\pi} \int_x^1 \frac{dz}{z} \begin{pmatrix} P_{qq} & P_{qg} \\ P_{gq} & P_{gg} \end{pmatrix} \begin{pmatrix} \Sigma(z,\mu_F^2) \\ g(z,\mu_F^2) \end{pmatrix} . \tag{1.55}$$

We can also rewrite eqs. (1.52) and (1.55) in Mellin space as

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

In practice, the DGLAP equations are solved using iterative numerical procedures. Various software packages have been developed for this purpose. Some, like HOPPET [28], QCDNUM [29], and APFEL [30], solve the evolution equations directly in momentum space. Others, such as PEGASUS [31] and EKO [32], employ the Mellin space approach.

It is important to note that the DGLAP equations are formulated in terms of splitting functions, which are computed using perturbation theory. Therefore, while eq. (1.56) hold true to all orders theoretically, in practice they are applied up to a certain finite order of perturbation theory. We will then talk about evolution performed at LO, NLO, NNLO and so on, according to the perturbative order accuracy of the splitting functions employed in the evolution. Examples of PDFs evolved from $Q_0 = 1.65$ GeV to $Q = 3.2$ GeV and $Q = 100$ GeV are shown respectively in fig. 1.6a and fig. 1.6b.

Before proceeding further, it is also beneficial to examine some general properties of the PDFs. These properties arise from fundamental observations about their nature and serve as important guidelines in the context of PDF determination.

Since the PDFs must reproduce the quantum numbers that characterize the proton,

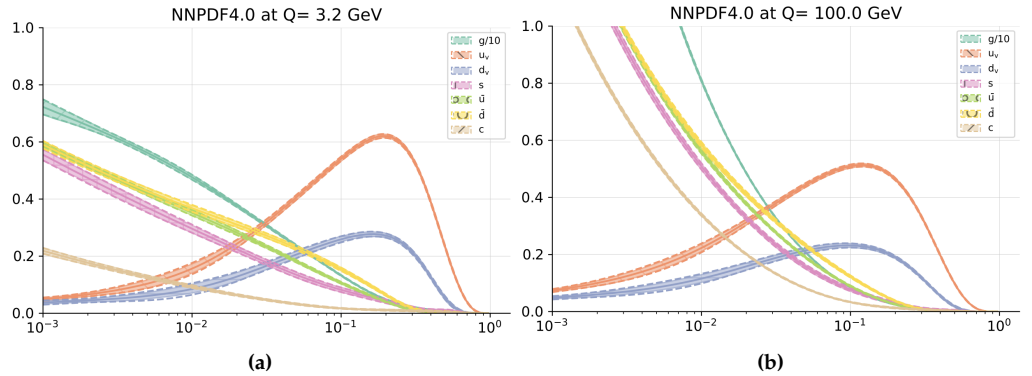


Figure 1.6: The NNPDF4.0 PDFs [8] evolved from the initial scale $Q_0 = 1.65$ GeV to $Q = 3.2$ GeV (left) and $Q = 100$ GeV (right).

it follows that

$$\int_0^1 dx(d(x,Q^2) - \bar{d}(x,Q^2)) = \int_0^1 d_V(x,Q^2) = 1\,, \tag{1.57}$$

for the valence down-quark, and

$$\int_0^1 dx(u(x,Q^2) - \bar{u}(x,Q^2)) = \int_0^1 u_V(x,Q^2) = 2\,, \tag{1.58}$$

for the valence up-quark. These relations, referred to as the *valence sum rules*, also require the PDFs to be integrable across the entire range of x.

The longitudinal momenta of all the constituent partons within a hadron must collectively equal the total longitudinal momentum of the hadron itself. This requirement is formalized by the momentum sum rule, which is expressed as

$$\sum_{i=q,\bar{q},g} \int_0^1 dx x f_i(x,Q^2) = 1\,. \tag{1.59}$$

1.4 Heavy Quarks

Quarks are conventionally categorized into light quarks, which have a mass significantly below $\Lambda_{\rm QCD}$, and heavy quarks, whose mass exceeds $\Lambda_{\rm QCD}$. According to this classification, the up, down, and strange quarks are considered light quarks. For these quarks, the massless approximation yields accurate results. For the remaining quarks, the massless approximation is no longer appropriate, in particular for processes in which the typical hard scale Q is of the same order of magnitude of the quark mass. In fact, in the latter case, mass power correction, i.e. terms like m_q^2/Q^2, contribute significantly to the final prediction. On the other hand, in the region $Q \gg m_q$ the collinear logarithms, introduced in section 1.2, become relevant and they can spoil the perturbative accuracy of the series if not resummed at all orders.

It is then clear that the appropriate way, or *scheme*, to treat the heavy quark contributions strongly depends on the kinematic region of interest. For this reason, *variable flavour number schemes* (VFNS) are usually adopted to obtain accurate predictions for

datasets with a large range in the hard scale Q. In particular, for each heavy quark with mass m_h, we have three relevant kinematic regions:

- $Q \ll m_h$: The mass of the heavy quark is significantly larger than the hard scale of the process. In this scenario, the heavy quark can be decoupled [33] and treated as a purely final state particle, which means that it does not contribute to DGLAP evolution and to the running of the coupling. The scheme that is accurate in this region is known as the *fixed flavor number* (FFN) scheme and the number of considered flavours does not include the heavy quark.

- $Q \sim m_h$: Since the mass power corrections are relevant in this region, the partonic calculation contains the exact dependence on m_h but the heavy quark is still considered a *non-active* flavour and does not contribute to the evolutions. Moreover, m_h acts as an IR regulator, and thus the collinear singularities, given by the gluon splitting into a h-pair, produce $\log Q^2/m_h^2$ terms. In this regime, these terms are considered small and are included in the fixed order expansion.

- $Q \gg m_h$: The h quark is considered an active flavor, so its renormalization scheme is switched from decoupling to $\overline{\text{MS}}$ and it also contributes to DGLAP evolution and running coupling. The collinear logarithms are not small in this regime, thereby spoiling the accuracy of the fixed order expansion. Consequently, these logarithms are resummed to all orders in an effective heavy quark PDF. In this case, the calculation of the partonic cross sections is carried out in the $m_h \to 0$ limit, called *zero mass* (ZM) scheme, because the mass power corrections can be safely neglected.

To ensure a smooth transition between the three regions for all the heavy quarks, the so-called *general mass variable flavour number* (GM-VFN) schemes are usually adopted. While different formulations of such schemes exist, which will be briefly described in the next sections, their general approach is to interpolate between the Fixed Flavor Number (FFN) and Zero Mass (ZM) schemes, ensuring that double-counting of terms is avoided. In the following sections we will describe FFN (section 1.4.1), ZM (section 1.4.2) and GM-VFN (section 1.4.3) schemes, focusing on the Deep Inelastic Scattering case for simplicity's sake[2].

1.4.1 Fixed Flavour Number (FFN) scheme

Let us first consider the region where the mass of the heavy quark is approximately equal to the hard scale (the threshold region) or larger than the hard scale of the process $Q \lesssim m_h$. In this region, the heavy quark is considered a purely final state particle, i.e. it does not contribute to DGLAP and running coupling. We can then write the structure function in the FFN scheme as[3]

$$F^{[n_l]}(Q^2, m_h^2) = \sum_i^{n_l} C_i^{[n_l]}\left(\frac{m_h^2}{Q^2}, \frac{Q^2}{\mu^2}\right) \otimes f_i^{[n_l]}(\mu^2)\,, \tag{1.60}$$

[2]The decision to focus on Deep Inelastic Scattering (DIS) is also motivated by the fact that, in practice, DIS datasets are the primary instance where the use of a General-Mass Variable Flavor Number (GM-VFN) scheme is truly necessary. In most non-DIS datasets, the typical hard energy scale is sufficiently high that the Zero Mass (ZM) scheme is usually sufficiently accurate.

[3]Note that here we are setting the factorization scale to be equal to the renormalization scale, $\mu_R = \mu_F = \mu$.

where the index i runs over the n_l light quarks and the x dependence is omitted. Note the square bracket notation denoting the number of flavours in the scheme in which each quantity has been computed. In particular, $f_i^{[n_l]}$ denotes PDFs evolved with n_l active flavours in DGLAP. This means that $f_i^{[n_l]} \equiv 0$ for $i > n_l$ at all scales. In the same way, $C_i^{[n_l]}$ means that the running of the coupling is performed with n_l active flavours. We will denote it with $\alpha_s^{[n_l]}$.

1.4.2 Zero Mass (ZM) 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 $Q \gg m_h$. 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. On the other hand, in the $Q \gg m_h$ region, the power mass corrections become irrelevant and we can safely carry out the calculation in the $m_h \to 0$ limit. This scheme differs from the FFN scheme only through the additional parton, and thus the equation for the structure function analogue to eq. (1.60) can be written as

$$F^{[n_l+1]}(Q^2) = \sum_i^{n_l+1} C_i^{[n_l+1]}\left(\frac{Q^2}{\mu^2}\right) \otimes f_i^{[n_l+1]}(\mu^2)\,. \tag{1.61}$$

It is important to note that in this case we get the contribution of the heavy quark PDF $f_h^{[n_l+1]}$. This PDF is generated by DGLAP evolution performed with n_l+1 active flavours and thus it is non-zero only for $\mu^2 > m_h^2$[4]. This condition holds as long as the intrinsic component, which is non-zero also in the $\mu^2 < m_h^2$ region and not generated by DGLAP, of the heavy quark PDF is neglected. The generalized equations that apply to the intrinsic heavy quark case can be found, for example, in [34].

1.4.3 General Mass Variable Flavour Number (GM-VFN) scheme

Previously, we have examined the Fixed Flavor Number (FFN) scheme, which is compromised by unresummed logarithms of Q^2/m_h^2, diminishing its accuracy beyond the region $Q \lesssim m_h$. Conversely, the Zero Mass Variable Flavor Number (ZM-VFN) scheme lacks corrections proportional to m_h/Q, affecting its precision outside the domain $Q \gg m_h$.

We then turn our attention to the General Mass Variable Flavour Number (GM-VFN) schemes. These schemes are designed to interpolate between the FFN and ZM-VFN approaches, thereby offering a unified framework that mitigates the impact of missing corrections when heavy quark masses are present. This interpolation ensures improved accuracy over a wider range of energy scales.

We first need to note that PDFs evolved with a different number of active flavours

[4]To be more precise, the scale where the heavy quark PDF is generated is called the *threshold* scale, denoted by μ_h, and it is another unphysical scale of the same kind of the renormalization and factorization scale. Conventionally, it is often chosen to equal the heavy quark mass, m_h, yet this selection is not obligatory and the final results depend perturbatively on this choice.

are related at the *matching* (or *threshold*) scale by

$$f_i^{[n_l+1]}(\mu_h^2) = \sum_{j=1}^{n_l} A_{ij}^{[n_l+1]\leftarrow[n_l]}(\mu_h^2/m_h^2) \otimes f_j^{[n_l]}(\mu_h^2)\,, \tag{1.62}$$

where A_{ij} are known as *matching conditions* and are known up to NNLO [35, 36]. Since the structure function in the n_l scheme,

$$F^{[n_l]}(Q^2) = \sum_{i=1}^{n_l} C_i^{[n_l]}(Q^2/\mu^2) \otimes f_i^{[n_l]}(\mu^2)\,, \tag{1.63}$$

and in the $n_l + 1$ scheme,

$$F^{[n_l+1]}(Q^2) = \sum_{i=1}^{n_l+1} C_i^{[n_l+1]}(Q^2/\mu^2) \otimes f_i^{[n_l+1]}(\mu^2)\,, \tag{1.64}$$

must match at the matching scale μ_h, i.e.

$$F^{[n_l]}(\mu_h^2) = F^{[n_l+1]}(\mu_h^2)\,, \tag{1.65}$$

we can get, using eq. (1.62),

$$C_j^{[n_l]}(\mu_h^2) = \sum_{i=1}^{n_l+1} C_i^{[n_l+1]}(\mu_h^2) \otimes A_{ij}^{[n_l+1]\leftarrow[n_l]}(\mu_h^2/m_h^2)\,. \tag{1.66}$$

It is important to note that the last equation contains a degree of arbitrariness. Specifically, the transformation matrix A_{ij} converts an $n_l + 1$-dimensional vector into an n_l-dimensional vector. This introduces a degree of freedom, linked to the terms proportional to powers of m_h/Q, that one can exploit to simplify the construction of the GM-VFN scheme. This degree of freedom allows for a scheme choice, which has led to the introduction of several GM-VFN schemes, such as:

- The ACOT scheme [37] and the S-ACOT scheme [38, 39].
- The TR scheme [40] and the TR' scheme [41].
- The FONLL scheme [42].
- The BPT scheme [43].

The FONLL scheme is particularly significant within the context of this thesis, as it is adopted by the NNPDF collaboration. It serves as the framework employed for generating theoretical predictions that underpin most of the results presented in this thesis. However, in the following we will adopt the BPT scheme choices and notation. This scheme shares many similarities with FONLL but simplifies the expression of the final result. In particular, the FONLL and BPT schemes are exactly equivalent at all orders, even if the construction itself is obtained with different steps, and they start to differ only in the organization of the perturbative expansion.

The fundamental concept behind constructing the GM-VFN scheme is to combine observables computed in the n_l and $n_l + 1$ scheme, while carefully subtracting the terms

that would otherwise be counted twice. So we write the structure function as

$$F^{\rm GM} = F^{[n_l+1]} + F^{\rm nons}\,, \tag{1.67}$$

where $F^{[n_l+1]}$ is the resummed result of eq. (1.61) while $F^{\rm nons}$ contains all and only the mass power corrections and thus vanishes in the limit $m_h \to 0$. This last condition ensures that correctly $F_{\rm GM} \to F^{[n_l+1]}$ in the high-energy limit. We then need to require that

$$F^{\rm nons} = F^{[n_l]} - F^{\rm sing}\,, \tag{1.68}$$

where $F^{[n_l]}$ is the FFNS result of eq. (1.60) and $F^{\rm sing}$ must contain all the double counting terms. In particular, this implies that $F^{\rm sing}$ is the fixed-order expansion of the resummed result $F^{[n_l+1]}$. We can then obtain its expression just evaluating $F^{[n_l+1]}$ in $\mu_h = Q$ as

$$F^{\rm sing} = F^{[n_l+1]}|_{\mu_h=Q} = \sum_{i,j=1}^{n_l} [C_j^{[n_l+1]} A_{ji}^{[n_l+1]\leftarrow[n_l]}(m_h,Q) + C_h^{[n_l+1]} A_{hi}^{[n_l+1]\leftarrow[n_l]}(m_h,Q)] f_i^{[n_l]}(Q)\,. \tag{1.69}$$

where we used the $f^{[n_l+1]}$ expression of eq. (1.62). We can then plug eq. (1.69) in the definition of $F^{\rm nons}$ to get

$$\begin{aligned} F^{\rm nons} &= F^{[n_l]} - F^{\rm sing} = \\ &= \sum_{i,j=1}^{n_l} [D_i^{[n_l]} - C_j^{[n_l+1]} A_{ji}^{[n_l+1]\leftarrow[n_l]}(m_h,Q) - C_h^{[n_l+1]} A_{hi}^{[n_l+1]\leftarrow[n_l]}(m_h,Q)] f_i^{[n_l]}(Q)\,. \end{aligned} \tag{1.70}$$

Note that the coefficients functions $D_i^{[n_l]}$ are obtained from the $C_i^{[n_l]}$, re-expanding their perturbative series in terms of the n_l+1 running coupling $\alpha_s^{[n_l+1]}$[5]. This is needed to ensure that all the expressions in the last equation are expanded in terms of the same coupling $\alpha_s^{[n_l+1]}$.

While in principle we have everything now to compute $F^{\rm GM}$, we would like to express $F^{[n_l+1]}$ and $F^{\rm nons}$ in terms of the same PDFs set. In fact, the former is expressed in terms of the n_l+1 PDFs, while the latter (eq. (1.70)) is expressed in terms of the n_l PDFs. We then rewrite

$$F^{\rm nons} = \delta C_i^{\rm nons}(Q,m_h) f_i^{[5]}(Q)\,, \tag{1.71}$$

and we fix the coefficients $C_i^{\rm nons}$ comparing the last equation to eq. (1.70). We then get

$$\delta C_i^{\rm nons} = \sum_{j=1}^{n_l} \Big[D_j^{[n_l]} - \sum_{k=1}^{n_l+1} [C_k^{n_l+1} A_{kj}^{[n_l+1]\leftarrow[n_l]}] \Big] A_{ji}^{[n_l]\leftarrow[n_l+1]}\,. \tag{1.72}$$

Here we note explicitly the ambiguity caused by the inverse $A_{ji}^{[n_l]\leftarrow[n_l+1]}$ of the rectangular matrix $A_{kj}^{[n_l+1]\leftarrow[n_l]}$. In the BPT scheme, we exploit these two degrees of freedom to impose

$$\delta C_h^{\rm nons} = \delta C_{\bar h}^{\rm nons} = 0\,, \tag{1.73}$$

[5]This is done using the expression $\alpha_s^{[n_l+1]}(\mu^2) = \alpha_s^{[n_l]}(\mu^2) + \frac{(\alpha_s^{[n_l]})^2}{6\pi} \log \frac{\mu^2}{\mu_h^2} + \mathcal{O}(\alpha_s^2)$.

which simplifies the practical implementation of the GM-VFN scheme. The final result can be written as

$$F = \sum_{i=1}^{n_l} \tilde{C}_i(Q, m_h) f_i^{n_l+1} + C_h^{[n_l+1]}(Q) f_h^{[n_l+1]}(Q)\,, \tag{1.74}$$

where C_h does not contain mass power correction as consequence of the choice made in eq. (1.73) and

$$\tilde{C}_i(Q, m_h) = C_i^{[n_l+1]}(Q) + \delta C_i^{\text{nons}}(Q, m_h)\,. \tag{1.75}$$

CHAPTER 2

Inclusion of theory errors in PDF fitting

The uncertainty associated with parton distribution functions (PDFs) represents a significant bottleneck in achieving precision physics at the Large Hadron Collider (LHC). Recent advancements in methodology, particularly the application of machine learning techniques and the accumulation of experimental data, have culminated in the development of NNPDF4.0 [8]. This version of PDFs claims a nominal precision at the percent level. It is imperative to evaluate whether this purported precision is indeed reliable and whether it corresponds to a comparable level of accuracy.

Significant efforts have been dedicated to evaluating the impact of the methodologies employed in the determination of PDFs on their associated uncertainties. This includes, in particular, the manner in which the information contained in the data is propagated to the PDF uncertainties ([9, 44]). Nevertheless, the uncertainties provided in all standard PDF sets, such as NNPDF4.0 [8], CT18 [45], MSHT20 [46], or ABMP16 [47], do not account for theoretical uncertainties. The sole exceptions are the parametric uncertainty related to the value of the strong coupling constant, α_s, which has been routinely included since the early days of LHC physics [48], and nuclear uncertainties affecting data such as deep-inelastic scattering on nuclear targets (e.g., neutrino DIS data), which are incorporated in the NNPDF4.0 PDF determination [49, 50].

Theoretical uncertainties can, in principle, originate from a variety of sources, both parametric (such as the values of heavy quark masses) and non-parametric (such as the aforementioned nuclear corrections). Theoretical uncertainties associated with missing higher orders in Quantum Chromodynamics (QCD) computations—hereafter referred to as MHOUs—are particularly pertinent, as they influence any prediction. The current standard perturbative accuracy of QCD computations is next-to-next-to-leading order (NNLO), with next-to-next-to-next-to-leading order (N3LO) corrections only known in a limited number of instances [51]. At NNLO, MHOUs are typically on the order of a few percent or greater. For LHC precision observables utilized in PDF determination, such as gauge boson or top-pair production, this uncertainty is comparable to experimental systematic uncertainties and often exceeds experimental statistical uncertainties.

Given that the uncertainties in both the experimental measurements and theoretical predictions symmetrically contribute to the figure of merit used for PDF determination, it is unjustifiable to include the former without accounting for the latter if they are of comparable magnitudes. In the subsequent sections, the NNPDF4.0 approach to incorporating MHOUs into the PDF determination is presented in detail.

This chapter is organized as follows. In section 2.1 we briefly review the NNPDF4.0 methodology, emphasizing the aspects that are most pertinent to the main topic of this chapter (for further details the reader can refer to [3, 8, 52]). In section 2.2, we review the

theory covariance matrix framework, as initially presented in [4, 53]. We also provide some details about the estimation of the missing higher orders (2.2.1) and about the application of the method to the N3LO determination (2.2.3). In sections 2.3 and 2.4, we validate our estimation on the known NLO and present the results at the PDF and observable levels [5, 6].

2.1 NNPDF4.0 methodology

The determination of PDFs from discrete data exemplifies a pattern recognition problem, wherein the objective of a PDF fitter is to deliver an accurate representation of an unknown underlying function. However, the problem of PDF determination exhibits specific characteristic features that must be considered when developing a fitting framework. Firstly, unlike in most standard pattern recognition problems where the model output is directly compared to data, in the case of PDF determination, one cannot associate a single data point with a pair consisting of an input and an output of the model. Instead, as indicated by eqs. (1.41) and (1.42), each observable depends non-linearly on multiple output PDF functions across the entire range of x. Secondly, for PDFs to be effective in predicting observables, it is essential to provide a description of the full correlations among PDFs.

In the following, we will describe some of the most relevant aspects of the NNPDF4.0 determination, focusing on the propagation of data uncertainties to PDF uncertainties, without pretense of completeness.

2.1.1 Error propagation: Monte Carlo method

The most commonly adopted method to propagate data uncertainties to PDFs is the *Hessian* method [54, 55], which represents PDF uncertainties by symmetric eigenvectors. On the other hand, NNPDF utilizes a *Monte Carlo* (MC) replica approach.

The MC replica method involves generating a set of fit outcomes to approximate the posterior probability distribution of the PDF model based on a given set of experimental input data. This technique relies on a known data generating distribution, typically a multivariate normal distribution, which is used to generate N_{reps} pseudo-data samples. Each sample is subsequently fitted to the forward model employed to describe the data. For a more detailed mathematical treatment, refer to Ref. [56], where the authors present an analytical expression for the posterior distribution of the model derived from this method.

If we assume that the observational noise of the experimental data can be modeled as a vector drawn from a multivariate normal distribution with a specified covariance matrix C, which is measured in the experiment, the central experimental values y_0 are given by

$$y_0 = f + \eta\,, \tag{2.1}$$

where $f \in \mathbb{R}^{N_{\text{data}}}$ is the vector of true, thus *unknown*, observable values, which we will also refer to as Level-0 (L0) data[1], and $\eta \sim \mathcal{N}(0, C)$ represents the observational noise drawn from a Gaussian distribution centered at zero with covariance matrix C. We will also refer to the y_0 vector as Level-1 (L1) data. Then, the pseudo-data replicas are gener-

[1] This notation is typical of *closure tests* which will be described in chapter 3.

ated by augmenting y_0 with some noise $\epsilon^{(k)} \sim \mathcal{N}(0, C)$, as

$$\mu^{(k)} = y_0 + \epsilon^{(k)} = f + \eta + \epsilon^{(k)} \,, \tag{2.2}$$

where the index k runs over the number of replicas N_{reps}. Each realization of the noise $\epsilon^{(k)}$ is independently drawn from the same multivariate Gaussian distribution as the observational noise. The vector $\mu^{(k)}$ is known as Level-2 (L2) data. We can also write the last equation explicitly as

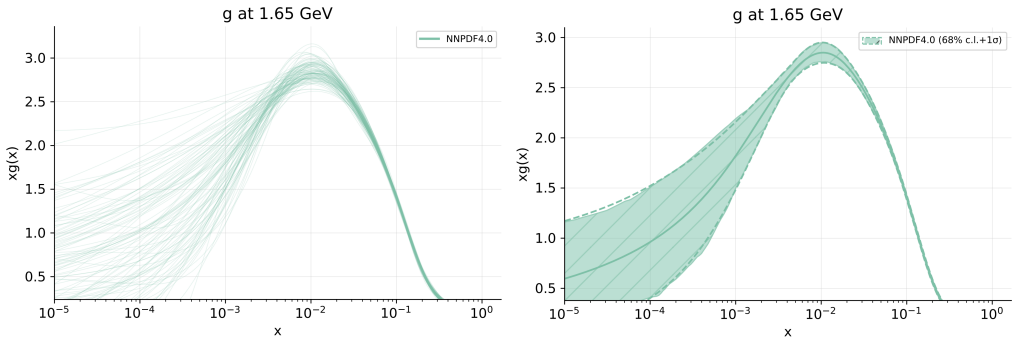
$$\mu^{(k)} = y_0 + \sum_{j=1}^{N_{\text{data}}} (\sqrt{C})_{i,j} r_j^{(k)} \,, \tag{2.3}$$

where $r_j^{(k)}$ are random numbers generated from a standard normal distribution and N_{data} is the total number of data used in the fit.

Therefore, the outcome of a PDF determination using the NNPDF framework consists of a set of N_{reps} Monte Carlo PDF replicas $f^{(k)}$ with $k = 1, \ldots, N_{\text{reps}}$, which represent an importance sampling of the probability distribution of the PDFs. Each PDF replica $f^{(k)}$ is obtained from its corresponding pseudo-data replica $\mu^{(k)}$. Estimators for functions of the PDFs, as well as their variances, are computed by simple averages over the replicas:

$$\begin{aligned} \langle X[f] \rangle &= \frac{1}{N_{\text{reps}}} \sum_{k=1}^{N_{\text{reps}}} X[f^{(k)}] \,, \\ \text{Var}[X[f]] &= \frac{1}{N_{\text{reps}}} \sum_{k=1}^{N_{\text{reps}}} (X[f^{(k)}] - \langle X[f] \rangle)^2 \,. \end{aligned} \tag{2.4}$$

Thus, uncertainty bands corresponding to any confidence level can be computed from the posterior Monte Carlo distribution. For instance, it can be verified that the 68% confidence interval aligns with the 1σ uncertainty band. This alignment is explicitly demonstrated for the gluon PDF at $Q = Q_0 = 1.65$ GeV in fig. 2.1. In particular, fig. 2.1a shows the distribution of $N_{\text{reps}} = 100$ gluon PDF replicas and fig. 2.1b shows the resulting 1σ interval and 68% confidence level.



(a) **(b)**

Figure 2.1: The distribution of $N_{\text{reps}} = 100$ gluon PDF replicas (left) and the resulting 1σ interval and 68% confidence level (right) [3].

Thus far, we have outlined how the MC replica method propagates experimental uncertainties to the PDFs. In section 2.2, we will demonstrate its generalization to accommodate additional sources of uncertainties, with particular focus on those arising from theoretical errors. Before proceeding, we will briefly introduce the PDF parametrization (section 2.1.2) and the fitting procedure (section 2.1.3).

2.1.2 PDF parametrization

The fundamental challenge in PDF determination is extracting a continuous function from a discrete dataset. This inherently poses an ill-defined problem, but it can be made tractable by constructing a suitable prior. It is the goal of this section to describe such *prior* in some details.

First, remind that the PDFs as functions of x need to be parameterized at a single scale Q_0, where the PDFs at any other scale Q can be derived by solving the DGLAP evolution equations discussed in section 1.3.3. It is crucial to select a parametrization that is complex enough to accurately describe the underlying data. A parametrization that is too simplistic may introduce biases in the resulting PDFs. For this reason, the NNPDF collaboration parametrizes the PDFs using a Neural Network, replacing the more common fixed functional form adopted, for instance, by CT18 [45], MSHT20 [46] and ABMP16 [47]. In the latter approach, the PDFs are parametrized using various functional forms constructed from polynomials in x and $\sqrt{x}$, followed by Hessian propagation methods. However, uncertainties obtained in this straightforward manner often underestimate the uncertainties of the corresponding predictions. To address this, a posteriori adjustment is made by inflating the chi-squared distribution corresponding to 1σ using a *tolerance* factor. The Neural Network (NN) approach adopted by NNPDF then mitigates this potential source of bias as it is known that, in the limit of infinite number of parameters, neural networks can reproduce any differentiable functions (*universal approximation theorem* [57]).

The NNPDF4.0 parametrization can be expressed as

$$xf_i(x, Q_0) = A_i x^{(1-\alpha_i)}(1-x)^{\beta_i}\mathrm{NN}_i(x)\,, \tag{2.5}$$

where the prefactors A_i and $x^{(1-\alpha_i)}(1-x)^{\beta_i}$ ensure that the known PDF constraints (section 1.3.3) are satisfied.

The Neural Network model NN_i, provides a non-linear mapping from an input space (in this case x) to an output space (in this case the space of PDFs). This is achieved through a directed graph structure consisting of multiple layers, where nodes in consecutive layers are fully connected. A schematic representation of this graph used in the NNPDF4.0 determination is depicted in fig. 2.2. Note that the we have two inputs, x and $\ln x$, because PDFs are expected to scale logarithmically at small x and linearly at large x [58].

In the figure, the blue circles represent the nodes of the graph, each associated with an *activation function*. Here, the input to each activation function corresponds to the set of all outputs of the previous layer, represented by the edges. Therefore, if we know the activation functions of each node, we can explicitly evaluate the neural network and obtain the function encoded by it. In particular, the output of the $i-$th node of the $l-$th layer is

$$\xi_i^{(l)} = g\left(\sum_j w_{ij}^{(l)}\xi_j^{(l-1)} + b_i^{(l)}\right), \tag{2.6}$$

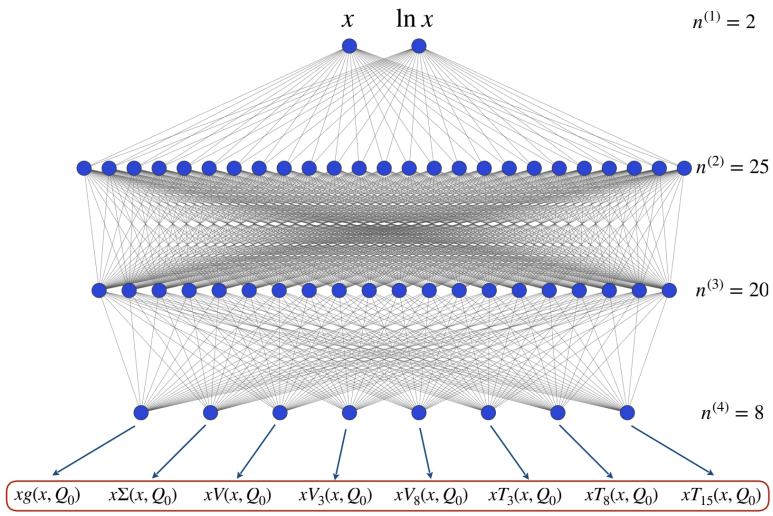


Figure 2.2: The graphical representation of the neural network parametrizaion adopted in NNPDF4.0 [8].

where $g()$ is the activation function of the node and the *weights* $w_{ij}^{(l)}$ and the *biases* $b_i^{(l)}$ are the free parameters of the NN. These are the parameters that are optimized during the fitting procedure, as described in section 2.1.3. The other parameters of the Neural Network, such as the number of layers, the number of units, the activation functions and so on, are called *hyperparameters* and, while they are kept fixed during each fit, their choice must be optimized. In [59] an automated algorithm that is able to optimize the hyperparameters, finding the best possible NN architecture among a finite space of possible architectures, has been proposed.

Various choices exist for the activation function, but it must be nonlinear and monotonic. A neural network constructed solely with linear activation functions would reduce to a basic linear regression model. A common choice for the activation function is the *sigmoid* function given by $g(x) = \frac{1}{1+e^{-x}}$. This function exhibits two asymptotes: $g(x) = 1$ as $x \to \infty$ and $g(x) = 0$ as $x \to -\infty$, making it a differentiable function that approximates a *step function*. The concept of the activation function as a step function provides an intuitive analogy to neurons in a biological brain, where neurons either fire a signal or do not based on their inputs.

To mitigate the potential restriction imposed by the polynomial prefactor in eq. (2.5) and to avoid underestimating uncertainties, the exponents α and β are randomly sampled from a range determined in a self-consistent manner [60, 61]. Specifically, when changes are made to the methodology or dataset, an initial fit is conducted to calculate the effective exponents for each distribution using

$$\alpha_i(x) = \frac{\log f_i(x)}{\log 1/x}, \quad \beta_i(x) = \frac{\log f_i(x)}{\log 1 - x}. \tag{2.7}$$

Then, for each subsequent fits, the α and β exponents are sampled from an uniform

distribution on a interval corresponding to twice the confidence interval of the respective effective exponent.

Finally, it is noteworthy from fig. 2.2 that the output nodes are parameterized in the evolution basis $\{g, \Sigma, V, V_3, V_8, T_3, T_8, T_{15}\}$ (see section 1.3.3). This choice is motivated by the observation that, as explained in section 1.3.3, the DGLAP evolution solutions are more straightforward in the evolution basis compared to the flavour basis. In [8], it has been demonstrated that the PDFs remain stable, i.e. agree within the 1σ level, when changing the parametrization basis.

2.1.3 Fitting procedure

The NNPDF fitting procedure is depicted in fig. 2.3. The optimization of the free parameters of the NN is carried out by the so-called *stopping algorithm* (described in more detail later on), that is contained in the blue box in the figure. The inputs of the stopping algorithm are the theoretical predictions in *FK table* format (see chapter 4), the experimental data along with their covariance matrices and the hyperparameters of the Neural Networks optimized as described above. The output of the stopping algorithm are the PDFs parametrized a the fitting scale $Q_0 = 1.65$ GeV. They are then evolved to a predefined Q grid by a software implementing DGLAP equations[2], and selected by some *post-fit* criteria. Finally they are exported in LHAPDF6 standard format [62].

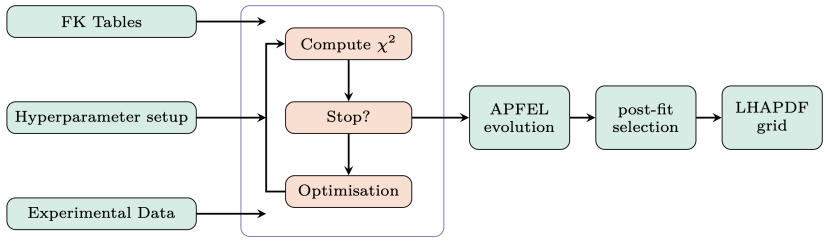


Figure 2.3: Diagrammatic representation of the NNPDF fitting procedure. The orange boxes describe the *stopping* algorithm with which the χ^2 is minimized. The inputs of the stopping algorithm (on the left) are the experimental data, the theoretical predictions (in *FK Table* format, see chapter 4) and the hyperparameters of the Neural Network. The outputs are the PDFs at the fitting scale Q_0 which are first evolved to a predefined grid of Q values and then filtered by some *post-fit* selection criteria. The final grids are then provided in the standard LHAPDF format [3].

Stopping algorithm. Given the assumption that the experimental uncertainties are Gaussian, a natural choice for the target function is the chi-squared statistic, defined as

$$\chi^2 = \sum_{i,j=1}^{N_{\text{data}}} (D_i - P_i) C_{ij}^{-1} (D_j - P_j) \,, \tag{2.8}$$

where D is the vector of experimental datapoints, P is the corresponding vector of theoretical prediction at a certain fitting step and, as introduced above, C is the experimental

[2]It used to be APFEL[30], but now we use EKO[32] which is part of the new theory pipeline described in chapter 4.

covariance matrix[3]. The latter can be expressed in terms of experimental uncertainties as

$$C_{ij} = \delta_{ij}\sigma_i^{(\text{uncorr})}\sigma_j^{(\text{uncorr})} + \left(\sum_{m=1}^{N_{\text{mult}}}\sigma_{i,m}^{(\text{norm})}\sigma_{j,m}^{(\text{norm})} + \sum_{l=1}^{N_{\text{corr}}}\sigma_{i,l}^{(\text{corr})}\sigma_{j,l}^{(\text{corr})}\right)D_i D_j \,, \tag{2.9}$$

where $\sigma_i^{(\text{uncorr})}$ are the uncorrelated systematic uncertainties, $\sigma_{i,m}^{(\text{norm})}$ are the multiplicative normalization uncertainties and $\sigma_{i,l}^{(\text{corr})}$ are the remaining correlated uncertainties.

A sufficiently large neural network is capable of optimizing the experimental data to such an extent that it also learns the noise present in the data, rather than restricting the extraction of information to only the genuine features of the data. This phenomenon is known as overfitting, and the stopping algorithm (fig. 2.4), together with the *cross-validation*, are specifically designed to prevent it.

Adopting a cross-validation method means that the full global NNPDF4.0 dataset is divided into a *training* dataset and a *validation* dataset. For each experimental dataset, a random fraction of 75% of the data points is placed in the training set, while the remaining 25% is placed in the validation set. Fig. 2.5 illustrates the use of the split into training and validation sets to identify the optimal instance of the neural network. During the fitting process, the training set is utilized to define a training error function χ^2_{tr}, which serves as the target for the optimizer. This function can, in principle, be reduced indefinitely, asymptotically approaching zero. This behavior is depicted by the blue curve in the figure.

Conversely, the validation set is not directly used by the optimizer, but the corresponding error function χ^2_{val} for this subset of data is evaluated at each training epoch, represented by the orange line. As shown in the figure, χ^2_{val} reaches its minimum value just before 6000 epochs, after which it begins to increase. This increase indicates overfitting, where the optimizer starts fitting the noise present in the training data and loses its ability to generalize well to unseen data.

The optimal result of the fitting procedure corresponds to the epoch at which χ^2_{val} is minimized. In fig. 2.5, the epoch representing the best instance of the neural network is highlighted by a vertical dashed line.

This criteria is implemented in the stopping algorithm, depicted in fig. 2.4. To determine when a neural network has completed its training, a counter is initiated once the validation loss χ^2_{val} falls below a specified threshold. From this point, the counter tracks the number of epochs that have passed, and the training terminates if the validation loss does not improve for a predetermined number of epochs. This number of epochs is a hyperparameter. Should this occur, the training is halted and the model is reset to the instance with the lowest validation loss. If at no stage during the training process does the validation loss reach this threshold value, the fit is considered insufficiently consistent with the data and is consequently discarded.

Additionally, for an instance to be deemed acceptable, it must satisfy certain positivity criteria, ensuring that the up, down, and strange quark and antiquark PDFs, as well

[3]In practice, in order to avoid the so-called D'Agostini bias [63], due to the presence of multiplicative uncertainties, during the optimization the t_0 covariance matrix is used instead [64]. This implies using the theoretical prediction P, obtained with a certain t_0 PDF set, in place of the datapoints D for the normalization part in eq. (2.9).

as the gluon PDF, are positive. These constraints are derived from [65], which demonstrated that the PDFs for individual quark flavors and the gluon, as defined in the $\overline{\text{MS}}$ factorization scheme, are non-negative.

Lastly, there is an upper limit on the number of epochs for which the model is allowed to be trained. If the model is still improving when this limit is reached, the training will nevertheless be terminated.

Once the training of the full set of replicas is complete, specific post-fit criteria are evaluated. Replicas that do not satisfy all of these criteria are discarded. As a result, any replica with an arc-length or a χ^2 value, calculated relative to the experimental data, that exceeds 4σ from the central value of its distribution is discarded. The post-fit check also verifies the integrability of the solutions by ensuring that the inequality

$$\sum_k |x_{\text{int}}^{(k)} f_i(x_{\text{int}}^k, Q^2)| < \frac{1}{2} \tag{2.10}$$

is satisfied for $f_i = V, V_3, V_8, T_3, T_8$ evaluated at $Q^2 = 5\,\text{GeV}^2$ and $x_{\text{int}}^{(k)} \in 10^{-9}, 10^{-8}, 10^{-7}$. Roughly 1% of the replicas are discarded by the post-fit criteria.

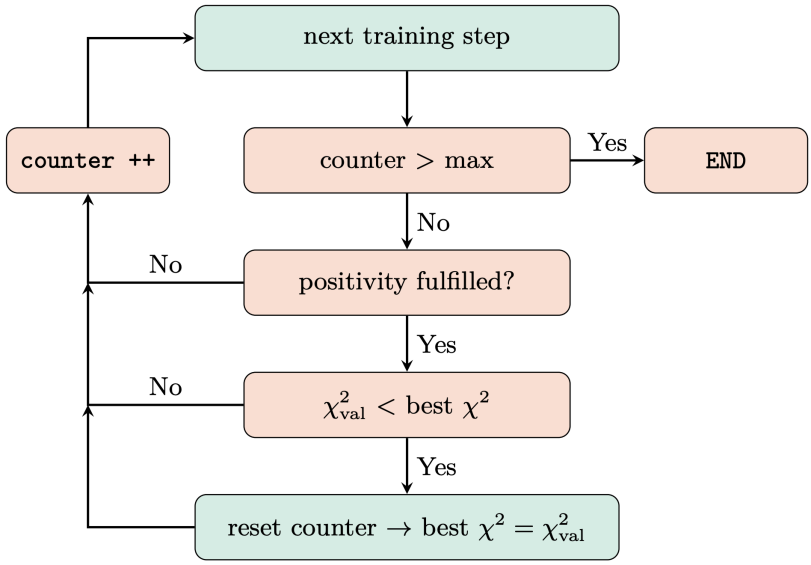


Figure 2.4: Diagram showing the stopping algorithm used to choose the optimal minimization step (or *epoch*) to stop the fit, based on the *look-back* algorithm [3].

2.2 The theoretical covariance matrix framework

The NNPDF methodology can be extended, under certain assumptions, to incorporate theoretical errors expressed by a theoretical covariance matrix. In this section, we first outline the theoretical foundation of this approach, which can accommodate any source of theoretical uncertainty. Subsequently, in section 2.2.1, we delve into the problem of estimating Missing Higher Order Uncertainties (MHOUs), addressed through the scale variation method. Detailed insights into the construction of the theory covariance matrix

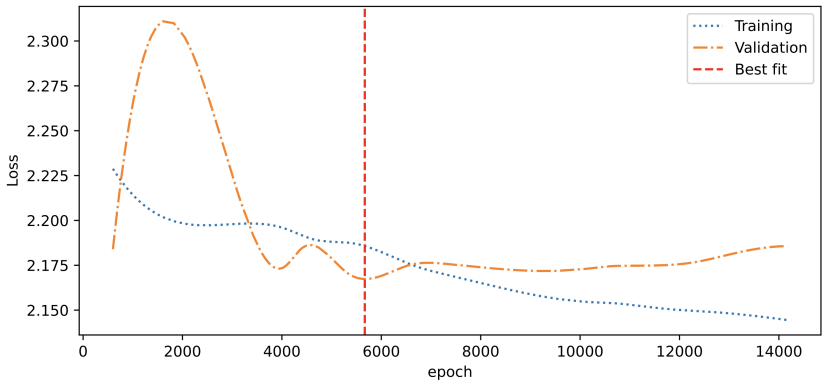


Figure 2.5: Typical profile of the training and validation χ^2 as a function of the fitting step. The optimal stopping point, in which the validation chi^2 reaches its minimum, is highlighted by a red dashed line [3].

are provided in section 2.2.2. Finally, the application of this framework to the broader challenge of extracting PDFs at N3LO is briefly discussed in section 2.2.3.

We begin by noting that each experimental data point D_i is associated with a "true" value $\mathcal{T}_i$, which represents the value given by Nature. Due to imperfections in experimental measurements, $\mathcal{T}_i$ cannot be determined exactly, but Bayesian probability can be used to estimate the likelihood of a hypothesis for $\mathcal{T}_i$. Assuming Gaussian distribution of experimental results around this hypothetical true value, the conditional probability for the true values $\mathcal{T}$ given the measured cross-sections D is

$$P(\mathcal{T}|D) = P(D|\mathcal{T}) \propto \exp\left(-\frac{1}{2}(\mathcal{T}_i - D_i)C_{ij}^{-1}(\mathcal{T}_j - D_j)\right). \tag{2.11}$$

Although the true values $\mathcal{T}_i$ are unknown, theoretical predictions P_i can be computed for each data point D_i. These predictions are derived from a theoretical framework that is typically incomplete, such as being based on a fixed-order truncation of a perturbative expansion or omitting higher-twist effects, nuclear effects, or other difficult-to-calculate factors. Additionally, these theory predictions P_i rely on PDFs evolved to a suitable scale, also using an incomplete theory. While the theory predictions may correspond to various observables and processes, they all hinge on the same underlying (universal) PDFs.

Now, following a similar approach used in estimating experimental systematics, we assume that the true values $\mathcal{T}_i$ are centered on the theory predictions P_i, and are distributed Gaussianly around these predictions. Ideally, these distributions would coincide if the theory were exact and the PDFs were known with certainty. The conditional probability for the true values $\mathcal{T}$ given theoretical predictions P is

$$P(\mathcal{T}|P) = P(P|\mathcal{T}) \propto \exp\left(-\frac{1}{2}(\mathcal{T}_i - P_i)S_{ij}^{-1}(\mathcal{T}_j - P_j)\right), \tag{2.12}$$

where the *theory covariance matrix* S_{ij} has to be estimated (section 2.2.1).

PDFs are determined by maximizing the marginalized probability of the theory given the data $P(P|D)$, where the true values $\mathcal{T}$ remain unknown. Using Bayes' theorem we

have

$$P(\mathcal{T}|DP)P(D|P) = P(D|\mathcal{T}P)P(\mathcal{T}|P)\,, \tag{2.13}$$

where

$$P(D|\mathcal{T}P) = P(D|\mathcal{T})\,, \tag{2.14}$$

given that experimental data do not depend on the theoretical predictions P but only on the true values $\mathcal{T}$. Thus, we can obtain

$$P(D|P) = \int D^N\mathcal{T}P(\mathcal{T}|D)P(\mathcal{T}|P)\,, \tag{2.15}$$

where the $N-$dimensional integral is over all possible values of $\mathcal{T}$ and which stems from

$$\int D^N\mathcal{T}P(\mathcal{T}|PD) = 1\,. \tag{2.16}$$

Having marginalized over the true values $\mathcal{T}$, now the probability of the experimental data is conditional to the theoretical predictions P.

We can rewrite this probability in terms of the difference between the true values and the theoretical predictions, i.e.

$$\Delta_i \equiv \mathcal{T} - P_i\,, \tag{2.17}$$

as

$$P(D|P) \propto \int D^N\Delta \exp\left(-\frac{1}{2}(D_i - P_i - \Delta_i)C_{ij}^{-1}(D_j - P_j - \Delta_j) - \frac{1}{2}\Delta_i S_{ij}^{-1}\Delta_j\right), \tag{2.18}$$

which is obtained exploiting the Gaussianity assumption. We can now perform this integral explicitly [4], exploiting the fact that both C and S are symmetric matrices, to get

$$P(P|D) = P(D|P) \propto \exp\left(-\frac{1}{2}(D-P)^T(C^{-1} - C^{-1}(C^{-1}+S^{-1})C^{-1})(D-P)\right). \tag{2.19}$$

Now, noting that

$$(C^{-1}+S^{-1}) = (C^{-1}(C+S)S^{-1})^{-1} = S(C+S)^{-1}C\,, \tag{2.20}$$

we can write

$$C^{-1} - C^{-1}(C^{-1}+S^{-1})^{-1}C^{-1} = C^{-1} - C^{-1}S(C+S)^{-1} = (C+S)^{-1}\,, \tag{2.21}$$

which allows us to write the final result

$$P(P|D) \propto \exp\left(-\frac{1}{2}(D_i - P_i)(C+S)^{-1}_{ij}(D_j - P_j)\right). \tag{2.22}$$

Comparison of eq. (2.22) with eq. (2.11) indicates that when replacing the true $\mathcal{T}_i$ by the theoretical predictions P_i in the expression for the χ^2 of the data, the theoretical covariance matrix S_{ij} should simply be added to the experimental covariance matrix C_{ij} [66]. This implies that, at least within this Gaussian approximation, when determining PDFs, theoretical uncertainties can be treated akin to additional experimental systematics: they

represent additional uncertainties considered when seeking to derive the truth from the data based on a specific theoretical prediction. The experimental and theoretical uncertainties are added in quadrature because they are assumed to be uncorrelated.

The question remains of how to estimate the theory covariance matrix S_{ij}. We need a method to estimate the shifts Δ_i, often referred to as *nuisance parameters* in the context of systematic error determination, that accounts for theoretical correlations among different kinematic points within the same dataset, across different datasets measuring the same physical process, and between datasets corresponding to different processes (involving different initial state hadrons). It should be noted that theory correlations persist even among different processes due to universal parton distributions; processes involving only leptons in the initial state are the only ones with truly independent theoretical uncertainties, though they are irrelevant for PDF determination.

The most commonly used method for estimating the theoretical corrections due to Missing Higher-Order Uncertainties (MHOUs), which naturally incorporates all theoretical correlations, is *scale variation*. This method is discussed in section 2.2.1 and then applied in section 2.2.2 to formulate specific procedures for constructing the theory covariance matrix S_{ij}. Other approaches discussed in the literature involve estimating MHOUs based on the behavior of known perturbative orders [67–70]; however, these approaches do not currently offer a sufficiently well-established formalism of broad applicability and it is currently not clear how they can provide theoretical correlations between observables. It should be noted that the formalism presented in this section is independent of the specific method used to estimate the correlated theoretical shifts Δ_i.

2.2.1 MHOUs from scale variations

Theoretical predictions at hadron colliders rely on two quantities computed perturbatively: the partonic cross sections or coefficient functions, eq. (1.41), and the anomalous dimensions, eq. (1.47), which determine the scale dependence, eq. (1.48), of the PDF. Both quantities can be expressed as a series in the strong coupling $\alpha_s(Q^2)$, which itself is given perturbatively through the solution to eq. (1.24) in terms of the value of the strong coupling at a reference scale, typically $\alpha_s(M_Z)$. The Missing Higher-Order Uncertainty (MHOU) on the predictions arises from the truncation of these perturbative expansions at a given order.

In principle, if a variable-flavour-number scheme is used (section 1.4), a further MHOU is introduced by the truncation of the perturbative expansion of the matching conditions that relate PDFs in schemes with different numbers of active flavors (see eq. (1.62)). These uncertainties, especially those related to the matching at the charm threshold, are very significant if one is interested in PDFs below the charm threshold, such as when determining the intrinsic charm PDF [71, 72]. However, for precision LHC phenomenology, physics predictions are produced in an $n_f = 5$ scheme, and PDFs are also determined by comparing to data predictions, the vast majority of which are computed in the $n_f = 5$ scheme. Therefore, the matching uncertainties only affect the small amount of data below the bottom threshold μ_b (no data below the charm threshold are used), and then only through the MHOU at the bottom threshold, which is very small. Consequently, the MHOU related to the matching conditions are subdominant, and we will neglect them here.

We thus focus on MHOUs on the hard cross-sections and anomalous dimensions. The estimation of these MHOUs from scale variations is obtained by producing various expressions for a perturbative result to a given accuracy, which differ by the subleading

terms generated when varying the scale at which the strong coupling is evaluated. Starting with the coefficient function, we construct a scale varied N^{k}LO coefficient function

$$\overline{C}(\alpha_s(\mu^2),\rho_r) = \alpha_s^m(\mu^2)\sum_{j=0}^{k}(\alpha_s(\mu^2))^j\overline{C}_j(\rho_r) \tag{2.23}$$

requiring that

$$\overline{C}(\alpha_s(\rho_r Q^2),\rho_r) = C(\alpha_s(Q^2))[1+\mathcal{O}(\alpha_s)]\,, \tag{2.24}$$

which determines the scale varied coefficients $\overline{C}_j(\rho_r)$ in terms of the starting C_j. Note that ρ_r is defined as the square of the ratio between the renormalization scale and the typical scale of the process, i.e. $\rho_r = \mu_r^2/Q^2$. We provide explicit expressions up to N^{3}LO in appendix A.

At any given order, C and $\overline{C}$ differ by *subleading* terms: their difference is taken as an estimate of the missing higher orders, and it is used for the construction of a theory covariance matrix, as described in section 2.2.2. We refer to this method of estimating MHOUs on partonic cross-sections as *renormalization scale variation*, as it has to do with the scale dependence of α_s in the coefficient functions expansion.

Through the same procedure, we may obtain an estimate of the MHOU on the anomalous dimension. Namely, we construct a scale-varied N^{k}LO anomalous dimension

$$\overline{\gamma}(\alpha_s(\mu^2),\rho_f) = \alpha_s(\mu^2)\sum_{j=0}^{k}(\alpha_s(\mu^2))^j\overline{\gamma}_j(\rho_f)\,, \tag{2.25}$$

requiring

$$\overline{\gamma}(\alpha_s(\rho_f Q^2),\rho_f) = \gamma(\alpha_s(Q^2))[1+\mathcal{O}(\alpha_s)]\,, \tag{2.26}$$

which fixes $\overline{\gamma}(\rho_f)$ in terms of γ_j. Note that this is just a specific instance of eq. (2.23), obtained choosing $m = 1$. The definition of ρ_f is analogous to the definition of ρ_r, i.e. $\rho_f = \mu_f^2/Q^2$. The subleading difference between γ and $\overline{\gamma}$ can be taken as an estimate of the MHOU on anomalous dimensions. This uncertainty then translates into a MHOU on the PDF $f(Q^2)$ when expressed through eq. (1.47) in terms of the PDFs at the parametrization scale. We refer to this estimate of the MHOU on the scale dependence of the PDF as *factorization scale variation*.

By substituting the scale-varied anomalous dimension $\overline{\gamma}(\alpha(Q^2),\rho_f)$ into the expression of the PDF in eq. (1.47), it can be shown [4] that factorization scale variation can equivalently be performed directly at the level of the PDF. To do so, we start by espressing the solution of the DGLAP evolution equations from Q_0^2 to Q^2 in terms of an *evolution kernel operator* (EKO)[32], as

$$f(Q^2) = E(Q^2\leftarrow Q_0^2) = \mathcal{P}\exp\left(-\int_{Q_0^2}^{Q^2}\frac{d\mu^2}{\mu^2}\gamma(\alpha_s(\mu^2))\right)f(Q_0^2)\,, \tag{2.27}$$

where $\mathcal{P}$ denotes the *path ordering*. Then, we define a scale-varied PDF $\overline{f}(Q^2,\rho_f)$, as a PDF whose scale dependence is governed by a scale varied EKO $\overline{E}(Q^2\leftarrow Q_0^2,\rho_f)$, as

$$\overline{f}(Q^2,\rho_f) = \overline{E}(Q^2\leftarrow Q_0^2,\rho_f)f(Q_0^2)\,. \tag{2.28}$$

The scale varied N^{k}LL[4] EKO, as usual, differs by subleading terms from the original EKO:

$$\overline{E}(Q^2 \leftarrow Q_0^2, \rho_f) = E(Q^2 \leftarrow Q_0^2)[1 + \mathcal{O}(\alpha_s)]\,. \tag{2.29}$$

We can construct the scale varied EKO as

$$\overline{E}(Q^2 \leftarrow Q_0^2, \rho_f) = K(\alpha_s(\rho_f Q^2), \rho_f) E(\rho_f Q^2 \leftarrow Q_0^2)\,, \tag{2.30}$$

where the additional N^{k}LL evolution kernel $K(\alpha_s(\rho_f Q^2), \rho_f)$ is given by the expansion

$$K(\alpha_s(\rho_f Q^2), \rho_f) = \sum_{j=0}^{k} (\alpha_s(\rho_f Q^2))^j K_j(\rho_f)\,. \tag{2.31}$$

By substituting this expansion into eq. (2.29), all coefficients $K_j(\rho_f)$ are determined in terms of γ_j. Their expressions are provided up to N^{3}LO in appendix A. Collectively, eqs. (2.29) and (2.30) imply that the scale-varied evolution kernel evolves from Q_0^2 to $\rho_f Q^2$, and then from $\rho_f Q^2$ back to Q^2, with the latter evolution expanded to fixed N^{k}LO.

The two methods for performing factorization scale variation, whether on anomalous dimensions (as in eq. (2.26)) or directly on PDFs (as in eq. (2.29)), are equivalent. When executed at N^{k}LO, both methods generate the same subleading N^{k+1}LO terms, although higher-order terms may differ. Specifically, this equivalence ensures that the theoretical predictions are consistent within the limits of the calculated perturbative order. These two approaches to factorization scale variation, involving variations in the scale of the anomalous dimension and the scale of the PDF, were distinguished as *Scheme A* and *Scheme B* in [53].

A third way, referred to as *Scheme C* in [53], consists of using the scale-varied PDF, eq. (2.28), namely

$$\overline{f}(Q^2, \rho_f) = K(\alpha_s(\rho_f Q^2), \rho_f) E(\rho_f Q^2 \leftarrow Q_0^2) f(Q_0^2)\,, \tag{2.32}$$

but including $K(\alpha_s(\rho_f Q^2), \rho_f)$ in the coefficient functions instead of the PDF, exploiting the factorization theorem. For completeness, also the explicit expressions needed to adopt this scheme are given in appendix A.

In standard practice, factorization scale variation is typically implemented using Scheme C, as it avoids the need to alter the PDFs, which are commonly obtained from an external source. However, in the context of PDF determination, factorization scale variation via Scheme B (eq. (2.28)) is preferred for its simplicity. This approach involves modifying only the EKO used in computing PDF evolution. Hence, we will adopt Scheme B for factorization scale variation.

2.2.2 Prescriptions for the theory covariance matrix

The missing higher-order uncertainty (MHOU) arising from the perturbative truncation of the partonic cross-sections and the scale dependence of the parton distribution functions (PDFs) are respectively estimated through renormalization scale variation, as described in eq. (2.24), and factorization scale variation according to scheme B of [4],

[4]The N^{k}LL solutions to the DGLAP evolution equations resum the N^{k}LL collinear logarithms (see section 1.4). In practice, it means that the anomalous dimensions are computed up to N^{k}LO.

detailed in eq. (2.30). These uncertainties are incorporated into the fit through the construction of a MHOU covariance matrix. We review here the details of the construction of such theory covariance matrix, together with a description of the possible prescriptions that can be adopted in its definition[4, 5, 53].

The estimation of the shifts Δ_i (as defined in section 2.2) given by the scale variation procedure is

$$\Delta_i(\rho_f, \rho_r) \equiv P_i(\rho_f, \rho_r) - P_i(0,0)\,, \tag{2.33}$$

where $P_i(\rho_f, \rho_r)$ is the prediction for the i-th datapoint obtained by varying the renormalization and factorization scale by a factor ρ_r, ρ_f respectively.

Next, we need to choose a correlation pattern for scale variation, as follows:

- Factorization scale variation is correlated for all datapoints because the scale dependence of PDFs is universal.
- Renormalization scale variation is correlated for all datapoints belonging to the same category. This category could either be the same observable, such as fully inclusive DIS cross-sections, or different observables pertaining to the same process, for example, the transverse momentum and rapidity distributions of the Z boson.

Note that this approach necessitates a categorization of processes. For instance, charged-current (CC) and neutral-current (NC) deep inelastic scattering are treated as distinct processes. The specific categorization adopted for the results presented in sections 2.3 and 2.4, is detailed in section 2.3.1.

These choices correspond to the assumption that factorization and renormalization scale variations fully capture the MHOU on anomalous dimensions and partonic cross-sections, respectively, and that missing higher-order terms are of a similar nature and thus of a similar magnitude across all processes within a given process category. Alternative assumptions are possible. For instance, one could decorrelate the renormalization scale variation from contributions to the same process originating from different partonic sub-channels, or introduce an additional variation of the process scale on top of the renormalization and factorization scale variations discussed above (see Section 4.3 of [4] for a more detailed discussion).

The MHOU covariance matrix is then defined as

$$S_{ij} = n_m \sum_{V_m} \Delta_i(\rho_f, \rho_{r_i}) \Delta_j(\rho_f, \rho_{r_j})\,, \tag{2.34}$$

where the sum runs over the space V_m of the m scale variations that are included. The factorization scale is always varied in a correlated manner, while the renormalization scales, corresponding to ρ_{r_i} and ρ_{r_j}, are varied in a correlated manner ($\rho_{r_i} = \rho_{r_j}$) if datapoints i and j belong to the same category. However, they are varied independently if i and j belong to different categories.

The normalization factor n_m is nontrivial to compute because it must account for the mismatch between the dimension of the space of scale variations when two datapoints are in the same category (hence, there is only one correlated set of renormalization scale variations) and when they are not (thus, there are two independent sets of variations). These normalization factors were computed for various choices of the space V_m and for various values of m in [4]. A detailed description of the possible prescriptions, as well as the computation of such normalization factors is provided in appendix B.

As in [4], and as is commonly done, we consider scale variations by a factor 2, so that

$$\rho_f, \rho_r = \{1/4, 1, 4\}\,. \tag{2.35}$$

In [4], various choices for the space of allowed variations were examined. Among these were the 9-point prescription, where ρ_r and ρ_f are allowed to take all values $\{1/4, 1, 4\}$, resulting in $m = 8$ (eight variations around the central value). Another commonly used prescription is the 7-point prescription, with $m = 6$, derived from the 9-point prescription by discarding the two outermost variations, specifically where $\rho_r = 4, \rho_f = 1/4$ and $\rho_r = 1/4$ and $\rho_f = 4$. We will demonstrate in section 2.3 that, upon validating the MHOU covariance matrix [5], the 7-point and 9-point prescriptions exhibit similar behavior, consistent with the findings of [4]. Other prescriptions, which involve a more limited set of independent scale variations, were shown in [4] to be less effective, and therefore, will not be considered further. The explicit expressions for the MHOU covariance matrix using the 7-point and 9-point prescriptions are provided in Eqs. (4.18-4.19) and Eq. (4.15) of [4], respectively.

The assumptions concerning the correlation patterns of renormalization and factorization scale variations, the categorization of processes, the range of scale variations, and the specific selection of variation points inherently involve some degree of arbitrariness (part of which is discussed in appendix B). This arises because the MHOU represents an estimate of the probability distribution for an unknown quantity with a unique true value, making it intrinsically Bayesian. The validation of such estimates relies on comparing their performance against cases where the true value is known, as we will demonstrate in section 2.3.

2.2.3 Application to N3LO

Calculations of hard-scattering cross-sections at the fourth perturbative order in the strong coupling (N3LO), have been available for a considerable period for massless deep-inelastic scattering (DIS) [73–76]. More recently, these calculations have been extended to encompass a rapidly expanding range of processes at hadron colliders. These processes include inclusive Higgs production through gluon-fusion [77, 78] and bottom-fusion [79], as well as in association with vector bosons [80] and through vector-boson fusion [81]. Additionally, N3LO calculations have been conducted for Higgs pair production [82], inclusive Drell-Yan production [83, 84], differential Higgs production [85–89], and differential Drell-Yan distributions [90, 91]. For a comprehensive overview, see [92].

In order to exploit this theoretical accuracy, one must combine these partonic cross-sections with PDFs at the same perturbative order. The primary challenge hindering progress in this endeavor is the absence of complete expressions for the N3LO splitting functions that dictate the scale dependence of the PDFs. Currently, only partial information is accessible [93–103], which includes a collection of integer N-Mellin moments, terms proportional to n_k^f with $k \geq 1$, and insights into the behavior at large and small values of x. Despite these partial findings, it is possible to approximate the N3LO splitting functions by combining available data [102, 104], mirroring successful strategies previously employed at NNLO [105]. Although we will not discuss the construction of this approximation in the context of this thesis, it is fully described in [6].

Therefore, achieving a global PDF determination at N3LO involves working with incomplete information: we have approximate knowledge of splitting functions and com-

plete knowledge of partonic cross-sections only for certain processes. A preliminary effort in this direction was undertaken in [104], where the unknown theoretical aspects of N3LO calculations were modeled using a set of nuisance parameters. These parameters were then simultaneously determined with the PDFs through a fit to experimental data.

Here, we employ a distinct approach. Specifically, we utilize the theory covariance matrix framework described in section 2.2 to address the gaps in perturbative information. In particular, we build, in the same way we built the MHOU covariance matrix in section 2.2.2, an *incomplete higher-order uncertainties* (IHOUs) covariance matrix that incorporates the uncertainties due to incomplete knowledge of N3LO theory, specifically for the splitting functions approximation and for the massive DIS coefficient functions. Armed with these theory covariance matrices, we are poised to conduct a determination of PDFs at *approximate N3LO* (referred to as aN3LO hereafter). In this approach, the theory covariance matrix plays a crucial role by accommodating both the incomplete understanding of N3LO splitting functions and the massive coefficient functions (IHOUs), as well as the absence of N3LO corrections in partonic cross-sections for hadronic processes (MHOUs).

Certainly, the purpose of this thesis does not delve into the detailed construction of the IHOUs covariance matrix, as comprehensively outlined in [6]. Instead, this section has underscored how the framework of theory covariance matrices can be extended to encompass challenges beyond its initial domain. Moreover, it serves as an introduction to the findings that will be presented in section 2.4, which include results that have been derived using aN3LO PDFs.

2.3 Validation on known perturbative order

In this section, we compute and validate the MHOU covariance matrix following the methodology outlined in the previous section. Initially, we describe the construction of the matrix based on an appropriate categorization of the dataset. Subsequently, we validate the matrix at the Next-to-Leading Order (NLO) level, where the next-order corrections are known, allowing for the exact determination of the true MHOUs.

2.3.1 Dataset and categorization of processes

To determine the covariance matrix, we must first select an appropriate dataset and process categorization. The dataset utilized for the determination of the NNPDF4.0MHOU PDFs is identical to that used for the determination of the NNLO NNPDF4.0 PDFs, as detailed in Ref. [8]. This same dataset is employed for both the NLO and NNLO NNPDF4.0MHOU PDFs discussed herein. In Ref. [8], a slightly different dataset was used for the NLO PDF determination, specifically excluding data points for which NNLO corrections are substantial and including some data at NLO for which NNLO corrections were not available at the time. Here, we aim to use the same dataset at both NLO and NNLO to analyze the impact of including MHOUs on perturbative convergence, without any changes in the dataset acting as a confounding factor.

As explained in the previous section, process categories correspond to classes of processes for which the missing higher-order terms are likely to originate similarly. Consequently, the correlation between the MHOU on any pair of predictions for processes within the same category can be approximated as if they were two data points from the

same physical process. We thus classify processes into nine categories: neutral-current deep-inelastic scattering (DIS NC); charged-current deep-inelastic scattering (DIS CC); and the following seven hadronic production processes: top-pair; Z, i.e., neutral-current Drell-Yan (DY NC); $W^{\pm}$, i.e., charged-current Drell-Yan (DY CC); single top; single-inclusive jets; prompt photon; and dijet. For more details about the process categorization see tables 2.2 to 2.5.

With these classifications, the covariance matrices at NLO and NNLO can be computed using eq. (2.34). The results at NLO and NNLO, computed using the 7-point prescription, are shown in fig. 2.6. As anticipated, the absolute value of the matrix elements is smaller at NNLO compared to NLO, with a reduction of nearly an order of magnitude. However, the pattern of correlations remains relatively stable across different perturbative orders. It is important to note that all data points, including those from different experiments, are correlated through MHOUs on perturbative evolution. This represents a significant difference compared to a typical experimental covariance matrix.

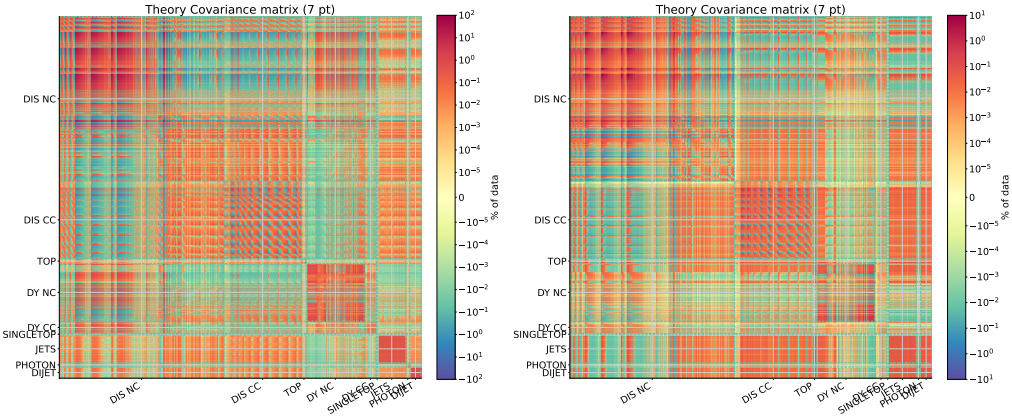


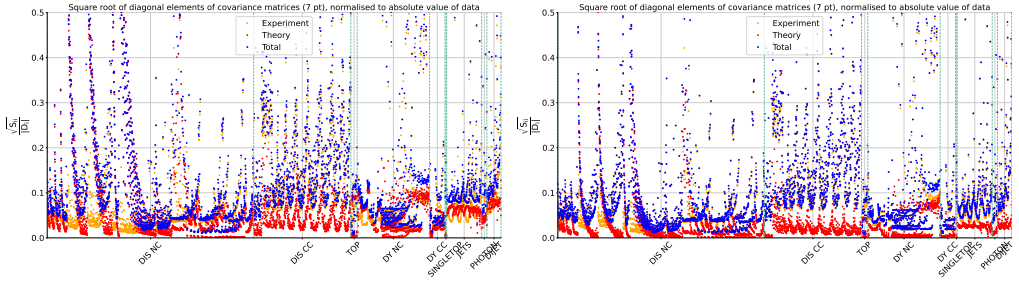
Figure 2.6: The theory covariance matrices computed at NLO (left) and at NNLO (right) [5].

The relative uncertainties on individual points (i.e. the square root of the diagonal covariance matrix elements) before and after the inclusion of the MHOU, as well as the MHOU itself, are compared in fig. 2.7 at both NLO and NNLO. It is evident that, at the NLO level, the MHOU uncertainty is comparable to the other components of the uncertainty, whereas, at NNLO, it is clearly subdominant. Consequently, we might expect the effect of MHOUs at NNLO to manifest primarily through correlations, thus impacting the PDF central values more significantly than the PDF uncertainties.

2.3.2 Validation procedure

The MHOU covariance matrix at NLO can be validated by comparing it to the known difference between NLO and NNLO predictions. This comparison can be performed using various estimators, originally proposed in [4]. We present here the results of this validation, both for our default 7-point prescription and for the 9-point prescription discussed in section 2.2.2 and in appendix B.

We define a normalized shift vector, whose i-th component δ_i is the normalized shift of the i-th datapoint due to the change in theory prediction from NLO to NNLO for a



(a) **(b)**

Figure 2.7: The diagonal elements of the theory covariance matrix at NLO (left) and at NNLO (right) compared with the experimental uncertainty and the total uncertainty, i.e. obtained as the sum in quadrature of the two [5].

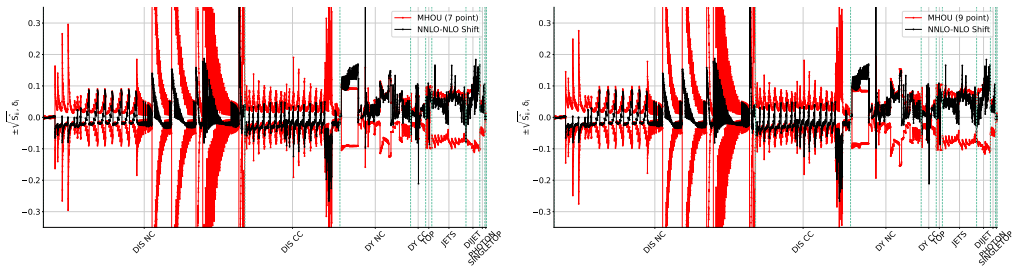
fixed PDF, namely

$$\delta_i = \frac{P_i^{\text{NNLO}} - P_i^{\text{NLO}}}{P_i^{\text{NLO}}}\,, \tag{2.36}$$

where P_i^{NNLO} and P_i^{NLO} are respectively the NNLO and NLO theory predictions both computed using the NLO PDF set. The simplest validation consists of comparing the shift δ_i to the uncertainty on individual points (also normalized), i.e. to the square root of the diagonal entries of the normalized NLO MHOU covariance matrix

$$\hat{S}_{ij}^{\text{NLO}} = \frac{S_{ij}^{\text{NLO}}}{P_i^{\text{NLO}} P_j^{\text{NLO}}}\,. \tag{2.37}$$

Results are presented in fig. 2.8 for both the 7-point (2.8a) and the 9-point (2.8b) prescriptions, where we compare δ_i to $\pm\sqrt{\hat{S}_{ii}^{\text{NLO}}}$. It is evident that for deep-inelastic scattering, both 7-point and 9-point scale variations at NLO provide a very conservative uncertainty estimate that significantly overestimates the NLO-NNLO shift. Conversely, for hadronic processes, the shift and scale variation estimates are generally comparable in size. However, for Drell-Yan (DY) processes, scale variations occasionally underestimate the shift.



(a) **(b)**

Figure 2.8: Comparison of the shifts of eq. (2.36) to $\pm\sqrt{\hat{S}_{ii}^{\text{NLO}}}$ (eq. (2.37)) as obtained with the 7-points prescription (left) and with the 9-points prescription (right) [5].

However, this validation method is rather crude as it does not test correlations. Correlations can be examined by comparing the eigenvalues of the covariance matrix to the projection of the shift along its eigenvectors. It is crucial to note that the shift δ_i is a vector in the N_{data}-dimensional space of data, whereas the independent eigenvectors of the covariance matrix span a smaller subspace S with dimension $N_{\text{sub}} \ll N_{\text{data}}$. In our case, $N_{\text{data}} = 4616$ while $N_{\text{sub}} = 22$ for the 7-point scale variation, and $N_{\text{sub}} = 48$ for the 9-point scale variation (see formulas in Appendix A of [4] with $p = 9$ process classes). Therefore, an additional nontrivial condition is that the shift vector δ_i should predominantly lie within the subspace S.

We can perform both tests quantitatively as follows. First, we compute the eigenvectors $\mathbf{e}_i^\alpha$ and eigenvalues $\lambda^\alpha = (s^\alpha)^2$ of the MHOU covariance matrix. Subsequently, we determine the projections δ^α of the shift vector δ_i onto these eigenvectors $\mathbf{e}_i^\alpha$, namely

$$\delta^\alpha = \sum_{i=1}^{N_{\text{data}}} \delta_i \mathbf{e}_i^\alpha \,, \quad \alpha = 1, \ldots, N_{\text{sub}} \,. \tag{2.38}$$

Finally, we determine the component of the shift vector in the N_{sub} dimensional subspace S:

$$\delta_i^S = \sum_{\alpha=1}^{N_{\text{sub}}} \delta^\alpha \mathbf{e}_i^\alpha \,. \tag{2.39}$$

The orthogonal component

$$\delta_i^{\text{miss}} = \delta_i - \delta_i^S \,, \tag{2.40}$$

is the part of the shift vector that is missed by the MHOU covariance matrix.

We can now evaluate whether correlated uncertainties are accurately represented by comparing the magnitudes of s^α and δ^α. Under the assumption of Gaussian distribution of MHO terms, approximately 68% of δ^α should be less than or equal to s^α. Additionally, we can assess how much of the shift vector lies outside the subspace S by determining the magnitude $|\delta^{\text{miss}}|$ of the missed vector, and examining the angle between the full shift vector and its component contained within the S subspace, namely

$$\theta = \arccos \frac{|\delta^S|}{|\delta|} \,. \tag{2.41}$$

Clearly, if the shift vector δ were entirely accounted for by the MHOU covariance matrix, then $|\delta^{\text{miss}}| = 0$, $|\delta^S| = |\delta|$, and $\theta = 0$.

The shift projections $|\delta^\alpha|$ and covariance matrix eigenvalues $|s^\alpha|$ are compared for both the 7-point and the 9-point prescriptions in fig. 2.9, where we also illustrate the length of the missed component $|\delta^{\text{miss}}|$. There is generally good agreement between shift projections and predicted MHOUs for the largest eigenvalues using both prescriptions. For smaller eigenvalues, there is also good agreement, although the 9-point prescription tends to slightly underestimate the size of individual components of the shift.

The size of the missed component of the shift vector can be observed in fig. 2.9 to be comparable to its largest eigenvector component for both prescriptions, indicating it is relatively small compared to the full shift. This observation holds true especially for the first ten components, which are of similar magnitude. This relationship is further supported by examining the angle given by eq. (2.41) between the shift vector and its projection in the subspace S, as detailed in table 2.1 for both individual datasets and the

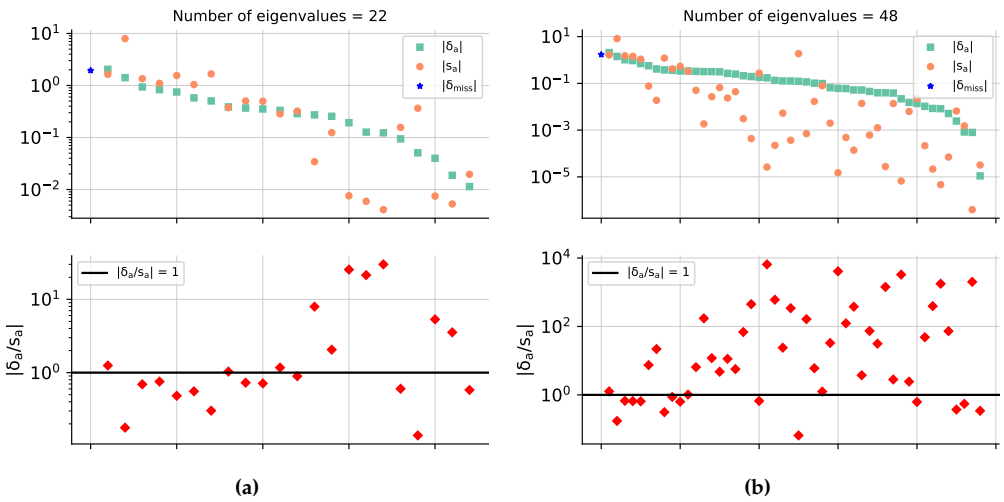


Figure 2.9: Comparison of the shift $|\delta^\alpha|$ with the MHOU covariance matrix eigenvalues $|s^\alpha|$ for the 7-point prescription (left) and the 9-point prescription (right). We also show the lenght of the missed component $|\delta^{\text{miss}}|$ [5].

entire dataset.

Remarkably, both the 7-point and 9-point prescriptions perform well despite the small size of the S subspace, showing minimal differences between them despite the 9-point prescription having a subspace size more than twice that of the 7-point prescription. Across almost all datasets, the direction of the shift and its projection in the S subspace are closely aligned, with some instances being nearly identical, except for cases like NC DIS and to a lesser extent DY, particularly CC.

Prescription	N_{sub}	DIS NC	DIS CC	TOP	DY NC	DY CC	SINGLETOP	JETS	PHOTON	DIJET	TOTAL
		θ [°]									
7-point	22	39	18	24	23	38	14	15	12	12	32
9-point	48	37	15	20	23	34	12	13	7	12	28

Table 2.1: The angle θ, eq. (2.41), for the 7-point and the 9-point prescriptions for individual process categories and for the total dataset. The dimension N_{sub} of the S subspace is also shown.

In summary, we conclude that the NLO MHOU covariance matrix effectively captures the uncertainty arising from missing NNLO corrections. The scale variation slightly underestimates the uncertainty in DY predictions, and both the 7-point and 9-point prescriptions perform comparably well. Specifically, the 9-point prescription excels in accurately delineating the subspace containing the shift, while the 7-point prescription better estimates the magnitude of uncertainty within that subspace. Consequently, we adopt the 7-point prescription as our default choice going forward.

2.4 PDFs with theoretical errors

We now focus on the primary outcomes of this study, namely the NNPDF4.0 NLO and NNLO PDF sets incorporating MHOUs. These sets are derived by re-executing the NNPDF4.0 PDF determinations, but with the inclusion of a MHOU covariance matrix determined using a 7-point prescription as described in section 2.2.2. The dataset used is identical to that employed for the determination of the NNPDF4.0 NNLO PDFs [8]. As discussed in section 2.3.1, we use the exact same dataset for both NLO and NNLO evaluations, contrasting with [8] where a somewhat different dataset was utilized at NLO. Consequently, we compare four distinct PDF sets in this study: NLO and NNLO, each with and without MHOUs, all derived from the same underlying dataset.

It is important to note that the NLO PDFs without MHOUs presented in this study are not suitable for phenomenological applications. They include data points for which NNLO corrections are substantial, leading to their exclusion from the NNPDF4.0NLO dataset as detailed in Ref. [8]. However, for the purpose of this study, we choose to compare PDFs generated using identical code and datasets, differing only in perturbative order and the inclusion of MHOUs. This approach allows us to assess the impact of MHOUs without any additional confounding effects, however minimal they may be.

Given that we are also interested in illustrating the perturbative convergence from NLO to N3LO and evaluating the impact of including MHOU in the PDF determination on phenomenology, we have produced a NNPDF4.0 NLO PDF with MHOU which utilizes the exact same dataset as described in [8]. Such results are presented in section 2.4.3.

Furthermore, the NNPDF4.0 NNLO PDFs presented here without MHOUs are equivalent but not identical to the published NNPDF4.0 PDFs [8]. They differ due to the correction of minor bugs in data implementation and the use of a new theory pipeline [12] (see chapter 4) for predictions computation, which includes a revised treatment of heavy quark mass effects involving subleading terms. The negligible impact of these changes was evaluated in Appendix A of [13]. Therefore, for practical applications, the NNPDF4.0 NNLO MHOU PDFs presented in this study can be considered as counterparts to the published NNPDF4.0 NNLO PDFs (without MHOU).

2.4.1 Fit quality

The tables 2.2 to 2.5 present the number of data points and the χ^2 per data point in the NLO and NNLO NNPDF4.0 PDF determinations before and after the inclusion of MHOUs. When MHOUs are not included, the covariance matrix is defined as in eq. (2.9), which consists of the sum of the experimental covariance matrix C and a theory covariance matrix only accounting for missing nuclear corrections $S^{(\text{nucl})}$, as determined in Refs. [49, 50]. The impact of $S^{(\text{nucl})}$ is discussed in Section 8.6 of [8]. When MHOUs are included, the covariance matrix also includes the contribution from eq. (2.34) discussed in section 2.2.2, denoted as $S^{(7\text{pt})}$.

It is important to note that the MHOU contribution is either excluded or included in the definition of the χ^2 used by the NNPDF algorithm, which includes pseudodata generation, training, and validation loss functions. Similarly, it affects the covariance matrix used to compute the values reported in tables 2.2 to 2.5. Additionally, the experimental covariance matrix used to compute these values differs from that used in the NNPDF algorithm. The NNPDF algorithm employs the t_0 method [106] for treating multiplicative uncertainties to mitigate the d'Agostini bias, while the published experimental covariance matrix is used as-is.

In table 2.2, datasets are aggregated according to the process categorization outlined in section 2.3.1. Individual datasets are detailed in table 2.3 (NC and CC DIS), table 2.4 (NC and CC DY), and table 2.5 (top pairs, single-inclusive jets, dijets, isolated photons, and single top).

Dataset	$N_{\rm dat}$	NLO		NNLO	
		$C+S^{(\rm nucl)}$	$C+S^{(\rm nucl)}+S^{(\rm 7pt)}$	$C+S^{(\rm nucl)}$	$C+S^{(\rm nucl)}+S^{(\rm 7pt)}$
DIS NC	2100	1.30	1.22	1.23	1.20
DIS CC	989	0.92	0.87	0.90	0.90
DY NC	736	2.01	1.71	1.20	1.15
DY CC	157	1.48	1.42	1.48	1.37
Top pairs	64	2.08	1.24	1.21	1.43
Single-inclusive jets	356	0.84	0.82	0.96	0.81
Dijets	144	1.52	1.84	2.04	1.71
Prompt photons	53	0.59	0.49	0.75	0.67
Single top	17	0.36	0.35	0.36	0.38
Total	4616	1.34	1.23	1.17	1.13

Table 2.2: The number of data points and the χ^2 per data point for the NLO and NNLO NNPDF4.0 PDF sets without and with MHOUs. Datasets are grouped according to the process categorization of section 2.3.1 [5].

Dataset	$N_{\rm dat}$	NLO		NNLO	
		$C+S^{(\rm nucl)}$	$C+S^{(\rm nucl)}+S^{(\rm 7pt)}$	$C+S^{(\rm nucl)}$	$C+S^{(\rm nucl)}+S^{(\rm 7pt)}$
NMC F_2^d/F_2^p	121	0.87	0.87	0.87	0.88
NMC $\sigma^{\rm NC,p}$	204	1.96	1.29	1.62	1.33
SLAC F_2^p	33	1.72	0.84	0.97	0.68
SLAC F_2^d	34	1.08	0.75	0.63	0.54
BCDMS F_2^p	333	1.60	1.26	1.41	1.29
BCDMS F_2^d	248	1.06	1.01	1.01	0.99
HERA I+II $\sigma_{\rm NC}$ e^-p	159	1.39	1.39	1.39	1.39
HERA I+II $\sigma_{\rm NC}$ e^+p ($E_p=460$ GeV)	204	1.11	1.04	1.08	1.04
HERA I+II $\sigma_{\rm NC}$ e^+p ($E_p=575$ GeV)	254	0.89	0.87	0.92	0.88
HERA I+II $\sigma_{\rm NC}$ e^+p ($E_p=820$ GeV)	70	1.08	0.96	1.12	0.95
HERA I+II $\sigma_{\rm NC}$ e^+p ($E_p=920$ GeV)	377	1.19	1.17	1.30	1.25
HERA I+II $\sigma^c_{\rm NC}$	37	1.83	1.66	2.03	1.75
HERA I+II $\sigma^b_{\rm NC}$	26	1.46	1.03	1.45	1.11
CHORUS σ^{ν}_{CC}	416	0.96	0.95	0.97	0.97
CHORUS $\sigma^{\bar\nu}_{CC}$	416	0.90	0.87	0.88	0.87
NuTeV σ^{ν}_{CC} (dimuon)	39	0.22	0.22	0.31	0.33
NuTeV $\sigma^{\bar\nu}_{CC}$ (dimuon)	37	0.58	0.39	0.56	0.64
HERA I+II $\sigma_{\rm CC}$ e^-p	42	1.39	1.18	1.25	1.29
HERA I+II $\sigma_{\rm CC}$ e^+p	39	1.33	1.25	1.22	1.25

Table 2.3: Same as table 2.2, for the DIS NC (top) and DIS CC (bottom) datasets [5].

Dataset	$N_{\rm dat}$	NLO		NNLO	
		$C+S^{(\rm nucl)}$	$C+S^{(\rm nucl)}+S^{(\rm 7pt)}$	$C+S^{(\rm nucl)}$	$C+S^{(\rm nucl)}+S^{(\rm 7pt)}$
E866 $\sigma^d/2\sigma^p$ (NuSea)	15	0.66	0.52	0.53	0.51
E866 σ^p (NuSea)	89	1.35	0.85	1.63	1.00
E605 $\sigma^d/2\sigma^p$ (NuSea)	85	0.44	0.42	0.46	0.45
E906 $\sigma^d/2\sigma^p$ (SeaQuest)	6	1.23	3.20	0.90	0.90
CDF Z differential	28	1.36	1.26	1.23	1.18
D0 Z differential	28	0.75	0.73	0.64	0.64
ATLAS low-mass DY 7 TeV	6	13.3	8.97	0.87	0.78
ATLAS high-mass DY 7 TeV	5	1.60	1.64	1.60	1.67
ATLAS Z 7 TeV ($\mathcal{L}=35$ pb^{-1})	8	0.77	0.49	0.60	0.57
ATLAS Z 7 TeV ($\mathcal{L}=4.6$ fb^{-1}) CC	24	5.00	3.29	1.73	1.68
ATLAS W,Z 7 TeV ($\mathcal{L}=4.6$ fb^{-1}) CF	15	1.82	1.21	1.07	1.02
ATLAS low-mass DY 2D 8 TeV	60	1.73	1.04	1.21	1.08
ATLAS high-mass DY 2D 8 TeV	48	1.48	1.34	1.12	1.08
ATLAS $\sigma_Z^{\rm tot}$ 13 TeV	1	0.48	0.43	0.30	0.60
ATLAS Z p_T 8 TeV ($p_T, m_{\ell\ell}$)	44	1.05	0.93	0.90	0.91
ATLAS Z p_T 8 TeV (p_T, y_Z)	48	0.74	0.69	0.88	0.70
CMS DY 2D 7 TeV	110	3.66	1.10	1.35	1.32
CMS Z p_T 8 TeV	28	1.66	1.58	1.40	1.41
LHCb $Z\to ee$ 7 TeV	9	1.51	1.36	1.64	1.53
LHCb $Z\to \mu$ 7 TeV	15	1.01	0.85	0.78	0.73
LHCb $Z\to ee$ 8 TeV	17	1.67	1.21	1.25	1.26
LHCb $Z\to \mu$ 8 TeV	16	1.40	1.05	1.46	1.59
LHCb $Z\to ee$ 13 TeV	16	1.34	1.61	0.96	1.80
LHCb $Z\to \mu\mu$ 13 TeV	15	1.88	1.13	1.75	0.99
D0 W muon asymmetry	9	2.48	1.92	1.99	1.95
ATLAS W 7 TeV ($\mathcal{L}=35$ pb^{-1})	22	1.23	1.15	1.13	1.12
ATLAS W 7 TeV ($\mathcal{L}=4.6$ fb^{-1})	22	2.74	2.26	2.15	2.16
ATLAS $\sigma_W^{\rm tot}$ 13 TeV	2	0.10	0.40	1.21	1.60
ATLAS W^++jet 8 TeV	15	1.68	1.15	0.79	0.79
ATLAS W^-+jet 8 TeV	15	1.82	1.31	1.49	1.45
CMS W electron asymmetry 7 TeV	11	0.85	0.96	0.83	0.85
CMS W muon asymmetry 7 TeV	11	2.05	1.75	1.74	1.73
CMS W rapidity 8 TeV	22	0.92	0.71	1.39	1.03
LHCb $W\to \mu$ 7 TeV	14	1.76	1.44	2.76	1.99
LHCb $W\to \mu$ 8 TeV	14	0.76	0.51	0.96	0.92

Table 2.4: Same as table 2.2, for the DY NC (top) and DY CC (bottom) datasets [5].

Dataset	$N_{\rm dat}$	NLO		NNLO	
		$C+S^{(\rm nucl)}$	$C+S^{(\rm nucl)}+S^{(7\rm pt)}$	$C+S^{(\rm nucl)}$	$C+S^{(\rm nucl)}+S^{(7\rm pt)}$
ATLAS $\sigma_{tt}^{\rm tot}$ 7 TeV	1	11.7	3.66	4.66	2.40
ATLAS $\sigma_{tt}^{\rm tot}$ 8 TeV	1	2.28	0.87	0.03	0.03
ATLAS $\sigma_{tt}^{\rm tot}$ 13 TeV ($\mathcal{L}$=139 fb^{-1})	1	4.58	1.18	0.56	0.41
ATLAS $t\bar{t}$ ℓ+jets 8 TeV $(1/\sigma d\sigma/dy_t)$	4	3.39	1.89	3.01	3.70
ATLAS $t\bar{t}$ ℓ+jets 8 TeV $(1/\sigma d\sigma/dy_{t\bar{t}})$	4	7.19	3.85	3.65	5.80
ATLAS $t\bar{t}$ 2ℓ 8 TeV $(1/\sigma d\sigma/dy_{t\bar{t}})$	4	1.80	1.76	1.57	1.86
CMS $\sigma_{tt}^{\rm tot}$ 5 TeV	1	0.72	0.95	0.01	0.01
CMS $\sigma_{tt}^{\rm tot}$ 7 TeV	1	6.37	1.82	1.10	0.50
CMS $\sigma_{tt}^{\rm tot}$ 8 TeV	1	4.39	1.21	0.31	0.17
CMS $\sigma_{tt}^{\rm tot}$ 13 TeV	1	1.06	0.36	0.04	0.01
CMS $t\bar{t}$ ℓ+jets 8 TeV $(1/\sigma d\sigma/dy_{t\bar{t}})$	9	1.67	1.61	1.20	1.59
CMS $t\bar{t}$ 2D 2ℓ 8 TeV $(1/\sigma d\sigma/dy_t dm_{t\bar{t}})$	15	2.03	1.84	1.32	1.25
CMS $t\bar{t}$ 2ℓ 13 TeV $(d\sigma/dy_t)$	10	0.77	0.71	0.51	0.59
CMS $t\bar{t}$ ℓ+jet 13 TeV $(d\sigma/dy_t)$	11	0.54	0.26	0.56	0.66
ATLAS incl. jets 8 TeV, $R=0.6$	171	0.70	0.73	0.71	0.64
CMS incl. jets 8 TeV	185	0.97	0.81	1.19	0.95
ATLAS dijets 7 TeV, $R=0.6$	90	1.48	1.82	2.16	1.69
CMS dijets 7 TeV	54	1.59	2.07	1.84	1.74
ATLAS isolated γ prod. 13 TeV	53	0.59	0.50	0.75	0.67
ATLAS single t R_t 7 TeV	1	0.38	0.30	0.48	0.57
ATLAS single t R_t 13 TeV	1	0.05	0.03	0.06	0.07
ATLAS single t 7 TeV $(1/\sigma d\sigma/dy_t)$	3	0.84	0.82	0.97	0.94
ATLAS single t 7 TeV $(1/\sigma d\sigma/dy_{\bar{t}})$	3	0.06	0.06	0.06	0.06
ATLAS single t 8 TeV $(1/\sigma d\sigma/dy_t)$	3	0.35	0.33	0.24	0.26
ATLAS single t 8 TeV $(1/\sigma d\sigma/dy_{\bar{t}})$	3	0.19	0.21	0.19	0.19
CMS single t $\sigma_t+\sigma_{\bar{t}}$ 7 TeV	1	0.91	0.96	0.76	0.84
CMS single t R_t 8 TeV	1	0.13	0.09	0.17	0.20
CMS single t R_t 13 TeV	1	0.32	0.28	0.35	0.38

Table 2.5: Same as table 2.2, for (from top to bottom) top pair, single-inclusive jet, isolated photon and single top production [5].

The tables 2.2 to 2.5 demonstrate that upon inclusion of the MHOU covariance matrix, the total χ^2 decreases for both the NLO and NNLO fits, with a more pronounced

decrease observed at NLO. However, even after including the MHOU, the NLO χ^2 remains somewhat higher than the NNLO χ^2. Examination of tables 2.3 to 2.5 reveals that this discrepancy is primarily attributed to a few datasets, specifically the ATLAS low-mass Drell-Yan dataset. Further investigation confirms that this is due to a small number of highly precise data points (excluded by the NLO cuts in [8]), for which NNLO corrections are significantly underestimated by scale variation.

Nevertheless, for the majority of data points and process categories, the MHOU covariance matrix effectively addresses the discrepancy between data and theoretical predictions at NLO arising from missing NNLO terms. This finding is consistent with the validation results discussed in section 2.3.

2.4.2 PDFs and PDF uncertainties

Individual PDFs at NLO and NNLO, with and without MHOUs, are compared in fig. 2.10 at $Q = 100$ GeV. We present the gluon, singlet, valence (V, $V3$, $V8$), and triplet ($T3$, $T8$, $T15$) distributions (defined in section 1.2.2), each shown as a ratio to the NNLO PDFs with MHOUs. Corresponding one sigma uncertainties are depicted in fig. 2.11.

The inclusion of MHOUs results in generally moderate changes in central values at NNLO. However, at NLO, significant changes are observed for the gluon and singlet distributions, while changes are more moderate for the other PDF combinations.

The PDF uncertainty at NNLO generally shows a slight reduction or remains unchanged upon inclusion of MHOUs. This somewhat counterintuitive observation, where the inclusion of an additional source of uncertainty leads to a reduction in PDF uncertainty in χ^2, has been previously noted in Refs. [49, 50]. It demonstrates the improved compatibility of the data due to the MHOU.

At NLO, a similar effect is observed in the nonsinglet sector, where the PDF uncertainty is reduced by the inclusion of MHOU. However, in the singlet sector, the PDF uncertainty increases upon inclusion of MHOU. This is consistent with the findings in section 2.3, where it was observed that at NLO, the MHOU from scale variation does not fully account for the substantial shift observed from NLO to NNLO for some datasets.

A quantitative assessment of PDF uncertainties on physics predictions can be obtained through the ϕ estimator, introduced in [61] (see Eq. (4.6) there) and also discussed in [4] (see sec. 6 there). The ϕ estimator calculates the ratio of the average correlated PDF uncertainty to the data uncertainty, and thus provides an estimate of the consistency of the data. In particular, a value $\phi < 1$ means that on average the uncertainties in the predictions are smaller than those of the original data, indicating that consistent data are being collectively described successfully by the underlying theory.

The value of ϕ, before and after inclusion of the MHOUs, is presented at NLO and NNLO in table 2.6. It is evident that ϕ decreases upon inclusion of MHOUs for most data categories, with a more pronounced decrease observed at NNLO and a smaller decrease at NLO. This observation further confirms that at NNLO, the inclusion of MHOU improves data compatibility, although this is not uniformly observed at NLO.

Dataset	NLO		NNLO	
	$C + S^{(\mathrm{nucl})}$	$C + S^{(\mathrm{nucl})} + S^{(7\mathrm{pt})}$	$C + S^{(\mathrm{nucl})}$	$C + S^{(\mathrm{nucl})} + S^{(7\mathrm{pt})}$
DIS NC	0.14	0.13	0.15	0.13
DIS CC	0.12	0.12	0.12	0.12
DY NC	0.19	0.17	0.18	0.17
DY CC	0.37	0.30	0.35	0.32
Top pairs	0.19	0.16	0.17	0.17
Single-inclusive jets	0.13	0.12	0.13	0.13
Dijets	0.10	0.09	0.11	0.10
Prompt photon	0.06	0.06	0.06	0.06
Single top	0.04	0.04	0.04	0.04
Total	0.16	0.15	0.17	0.15

Table 2.6: The ϕ estimator for PDFs at NLO and NNLO with and without MHOUs for the process categories of section 2.3.1[5].

A priori, PDF sets with and without MHOUs should not necessarily be compatible

within uncertainties, as the latter do not account for an additional source of uncertainty. However, they do agree in the nonsinglet sector, where MHOU are generally comparable in magnitude to the PDF uncertainty before their inclusion. In contrast, agreement is not typically observed in the singlet sector, where NNLO corrections can be quite substantial. The inclusion of MHOUs typically shifts the NLO PDFs towards NNLO, thereby improving perturbative convergence, except for the gluon distribution. Even for the singlet sector, where the NLO PDFs move towards NNLO upon inclusion of MHOUs, the NNLO results often lie well outside the NLO uncertainty band, especially at small x. This discrepancy indicates that in the singlet sector, there exist large NNLO corrections to the NLO result that are underestimated by MHOUs determined through scale variation. At small x, this discrepancy can be attributed to unresummed small-x logarithms [107], whose increase with perturbative order is not adequately captured by scale variation. A more detailed discussion about perturbative convergence is provided in section 2.4.3, where the aN^{3}LO fit is also shown.

In summary, the inclusion of MHOUs estimated through scale variation at NNLO enhances data compatibility, resulting in a reduction of uncertainties and a moderate shift in central values. At NLO, it partially addresses the effect of MHOUs on fit quality, with a moderate impact on both PDF uncertainties and central values in the nonsinglet sector. However, in the singlet sector, it leads to a more significant impact on central values along with an increase in uncertainties, while still not fully accounting for the largest missing NNLO corrections.

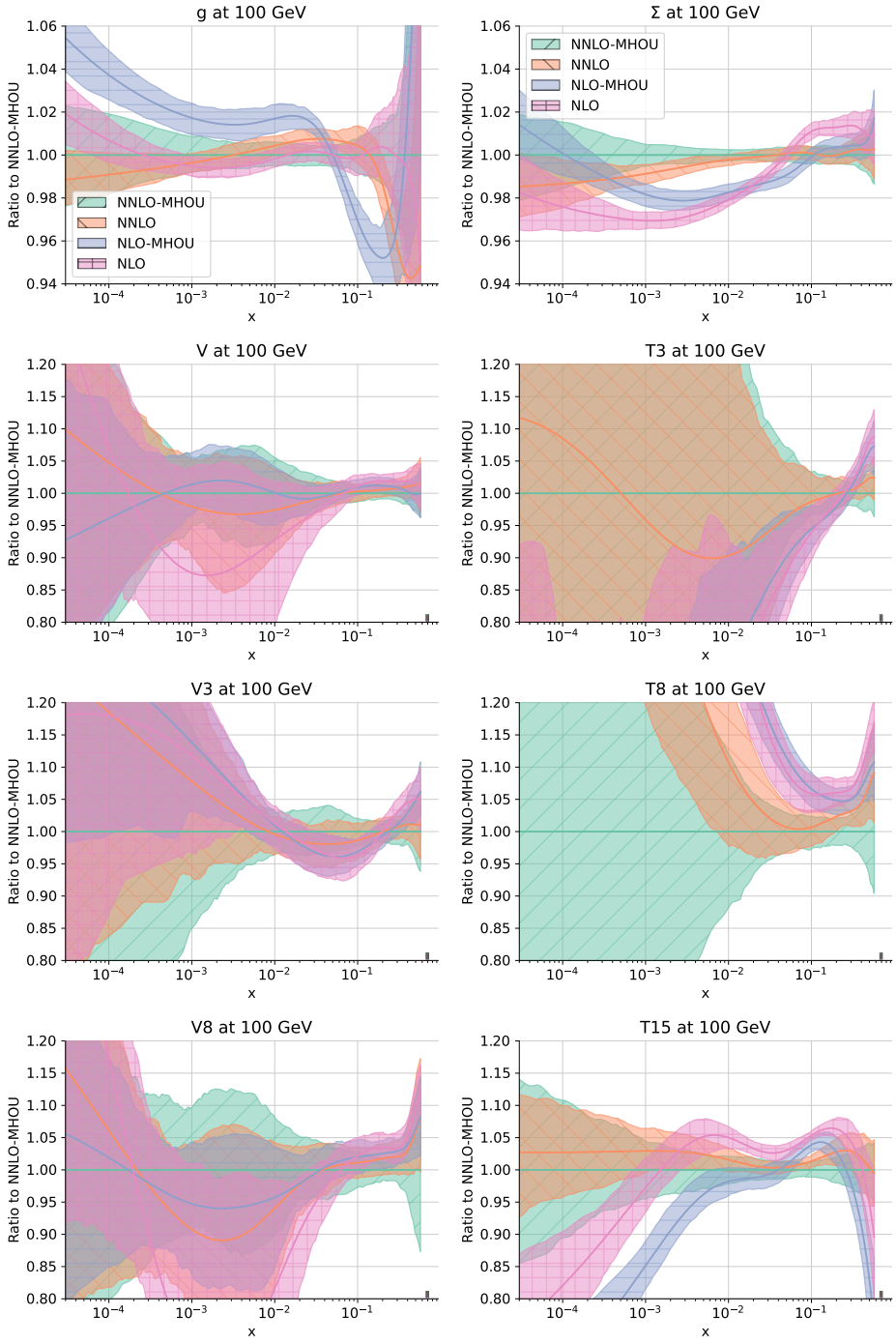


Figure 2.10: The NLO and NNLO PDFs with and without MHOUs at $Q = 100$ GeV. The gluon, singlet, valence (V, V_3, V_8), and triplet (T_3, T_8, T_{15}) PDFs are shown. All curves are normalized to the NNLO with MHOUs. The bands correspond to one sigma uncertainty [5].

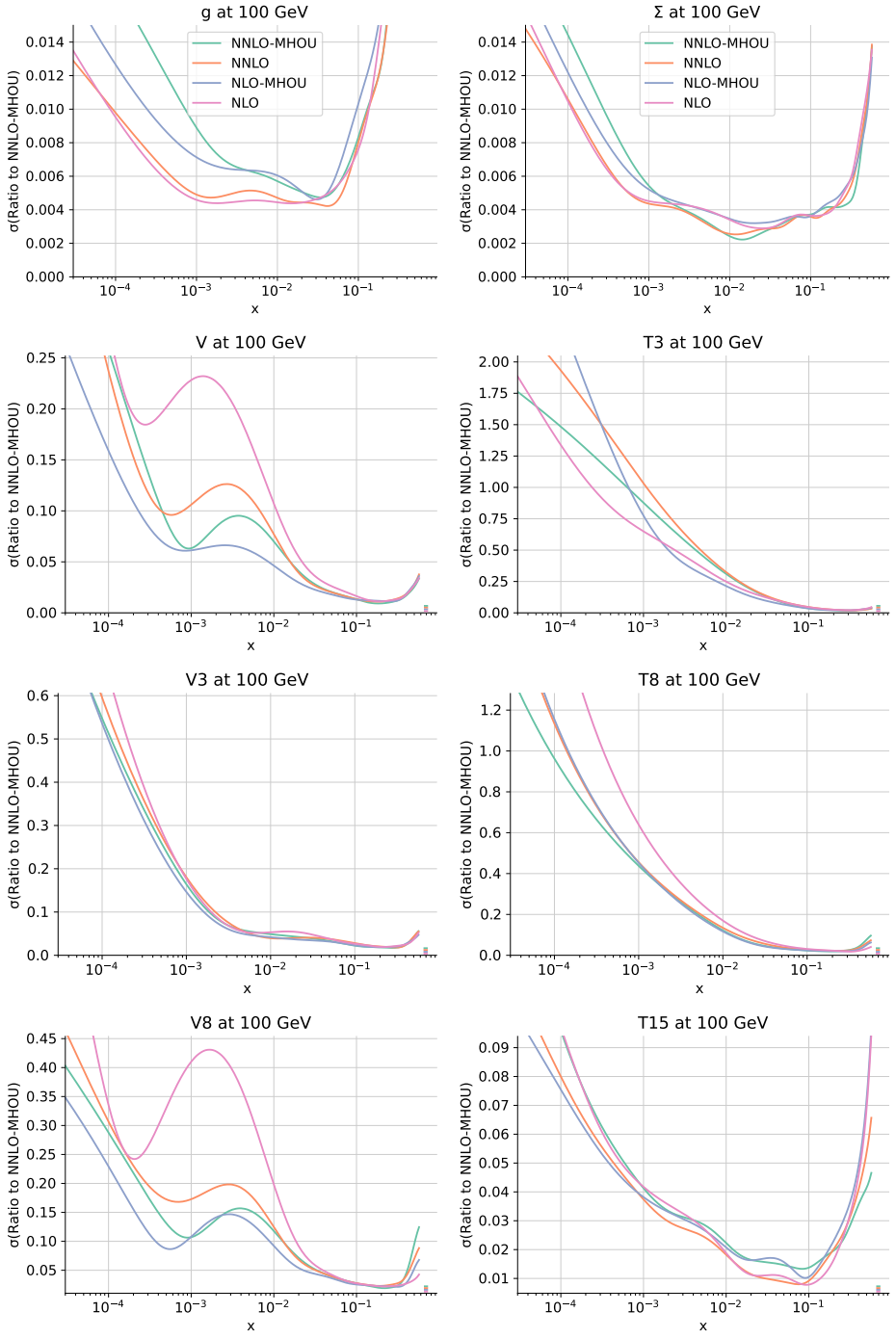


Figure 2.11: Relative one sigma uncertainties for the PDFs shown in fig. 2.10. All uncertainties are normalized to the corresponding central NNLO PDFs with MHOUs [5].

2.4.3 Perturbative convergence and phenomenology

In this section, we examine the perturbative convergence of PDFs and observables across NLO, NNLO, and aN^{3}LO (see section 2.2.3) fits, both with and without the inclusion of MHOUs. It should be noted that, while the NNLO fit discussed here is identical to that presented in the previous section, the NLO fit utilizes a different dataset. Specifically, this dataset excludes datapoints with significant NNLO corrections, as described in [8].

The value of the total χ^2 per data point is shown as a function of the perturbative order in fig. 2.12. It is observed that, in the absence of MHOUs, fit quality improves as the perturbative order increases. Conversely, when MHOUs are included, the fit quality becomes almost independent of the perturbative order within uncertainties (noting that, with $N_{\rm data} = 4616$, $\sigma_{\chi^2} = 0.03$). This indicates that the MHOU covariance matrix estimated through scale variation accurately reproduces the true MHOUs. Furthermore, at aN^{3}LO, the fit quality remains consistent within uncertainties, regardless of whether MHOUs are included. This observation strongly suggests that, given the current experimental uncertainties, the present methodology, and the existing dataset, the perturbative expansion has converged. Thus, the inclusion of higher-order QCD corrections beyond N^{3}LO is unlikely to further enhance the fit quality to the current data.

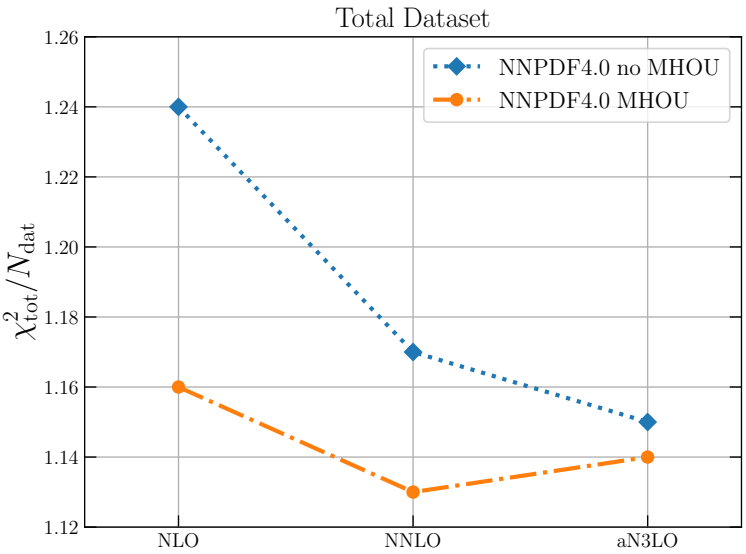


Figure 2.12: The values of the total χ^2 per data point in the NNPDF4.0 NLO, NNLO, and aN^{3}LO fits without and with MHOUs [6].

We compare the NLO, NNLO, and aN^{3}LO NNPDF4.0 PDFs, obtained without and with the inclusion of MHOUs, in figs. 2.13 and 2.14 and figs. 2.15 and 2.16, respectively. Specifically, we present the up, antiup, down, antidown, strange, antistrange, charm, and gluon PDFs at $Q = 100$ GeV, normalized to the aN^{3}LO result, as a function of x in both logarithmic and linear scales. Error bands correspond to one sigma PDF uncertainties, which either include (MHOU sets) or exclude (no MHOU sets) MHOUs on all theory predictions used in the fit.

The excellent perturbative convergence observed in the fit quality is also evident at

the level of PDFs. In particular, the NNLO PDFs are either very close to or indistinguishable from their aN^{3}LO counterparts. The inclusion of MHOUs further enhances the consistency between NNLO and aN^{3}LO PDFs, with the two lying almost on top of each other. This indicates that the NNLO PDFs are made more accurate by the inclusion of MHOUs, and that the aN^{3}LO PDFs have converged, as discussed above. Exceptions to this stability are observed for the charm and gluon PDFs, where aN^{3}LO corrections have a significant impact. For the charm PDF, these corrections lead to an enhancement of the central value by approximately 4% for $x \sim 0.05$, and for the gluon PDF, to a suppression of about $2-3\%$ for $x \sim 0.005$. In both cases, the inclusion of MHOUs increases PDF uncertainties by about $1-2\%$, making the NNLO and aN^{3}LO charm PDFs compatible within uncertainties, and the NNLO and aN^{3}LO gluon PDFs with MHOUs almost compatible.

Figure 2.17 presents a comparison similar to that of figs. 2.13 to 2.16 for the gluon-gluon, gluon-quark, quark-quark, and quark-antiquark parton luminosities. These are shown integrated in rapidity as a function of the invariant mass of the final state m_X for a center-of-mass energy $\sqrt{s} = 14$ TeV. Their definition follows Eqs. (1)-(4) of [108].

As already observed for PDFs, perturbative convergence is excellent and improves upon the inclusion of MHOUs. The NNLO and aN^{3}LO results are compatible within uncertainties for the gluon-quark, quark-quark, and quark-antiquark luminosities. Some differences are observed for the gluon-gluon luminosity, consistent with the differences seen in the gluon PDF. Specifically, the aN^{3}LO corrections lead to a suppression of the gluon-gluon luminosity by $2-3\%$ for $m_X \sim 100$ GeV. This effect is somewhat compensated by an increase in uncertainty of about 1% upon inclusion of MHOUs. Indeed, the NNLO and aN^{3}LO gluon-gluon luminosities for $m_X \sim 100$ GeV differ by about 2.5σ without MHOU, but become almost compatible within uncertainties when MHOUs are included.

Overall, these results indicate that aN^{3}LO corrections are generally small, except for the gluon PDF, and that at aN^{3}LO the perturbative expansion has nearly converged, with NNLO and aN^{3}LO PDFs very close to each other, especially upon inclusion of MHOUs. They also show that MHOUs generally improve the accuracy of PDFs, though at aN^{3}LO they have a very small impact.

LHC phenomenology at aN^{3}LO accuracy. We present a preliminary assessment of the implications of aN^{3}LO PDFs for LHC phenomenology by examining processes for which N^{3}LO results are publicly available, namely the Drell-Yan and Higgs total inclusive cross-sections.

At each perturbative order, the uncertainty on the cross-section is determined by combining the PDF uncertainty with the MHOU on the hard matrix element, obtained by performing 7-point renormalization and factorization scale variation and taking the envelope of the results. This procedure, commonly used for estimating the total uncertainty in hadron collider processes, is followed here for ease of comparison with existing results. In a more refined approach, MHOUs on the hard cross-section could be included through a theory covariance matrix for the hard cross-section itself, similar to the MHOUs and IHOUs on the PDF. This would allow tracking the correlation between different sources of uncertainty [109–111]. However, to disentangle the contribution of the MHOU in the processes used for PDF determination from that in the matrix element, we present results with the PDF uncertainty evaluated using both MHOU and no-MHOU PDF sets.

We present results for inclusive charged-current and neutral-current gauge boson production cross-sections, followed by their decays into dilepton final states. Cross-sections are evaluated using the `n3loxs` code [80] for various ranges in the final-state dilepton invariant mass, $Q = m_{\ell\ell}$ for neutral-current, and $Q = m_{\ell\nu}$ for charged-current scattering. Figure 2.18 displays the inclusive neutral-current Drell-Yan cross-section $pp \to \gamma^*/Z \to \ell^+\ell^-$, and figs. 2.19 and 2.20 show the charged-current cross-sections $pp \to W^\pm \to \ell^\pm\nu_\ell$. We consider one low-mass bin (30 GeV $\leq Q \leq$ 60 GeV), the mass peak bin (60 GeV $\leq Q \leq$ 120 GeV), and two high-mass bins (120 GeV $\leq Q \leq$ 300 GeV and 2 TeV $\leq Q \leq$ 3 TeV), which are relevant for high-mass new physics searches [112].

In all cases, we compare the NLO, NNLO, and aN^{3}LO predictions, with the same perturbative order in matrix element and PDFs, with and without MHOUs. In general, we observe good perturbative convergence, with predictions at successive orders agreeing within uncertainties, and generally improved convergence upon including MHOUs in the PDF. The difference between PDFs with and without MHOUs, while moderate, remains non-negligible even at N^{3}LO, where it starts being comparable to the overall uncertainty. Thus, it must be included in precision calculations.

We now consider Higgs production through gluon fusion, associated production with vector bosons, and vector-boson fusion (VBF). Predictions are obtained using the `ggHiggs` code [113] for gluon fusion, `n3loxs` for associated production, and `proVBFH` code [114] for VBF. Results are depicted in Figure 2.21 for gluon fusion and VBF, and in Figure 2.22 for associated production with W^+ and Z bosons.

In these cases, we observe generally good perturbative convergence, even for gluon fusion, which is known for its notoriously slow converging expansion. The impact of MHOUs on the PDFs is generally minor compared to the PDF uncertainty at all perturbative orders and is almost negligible for gluon fusion. For associated production, the inclusion of MHOUs marginally improves perturbative convergence.

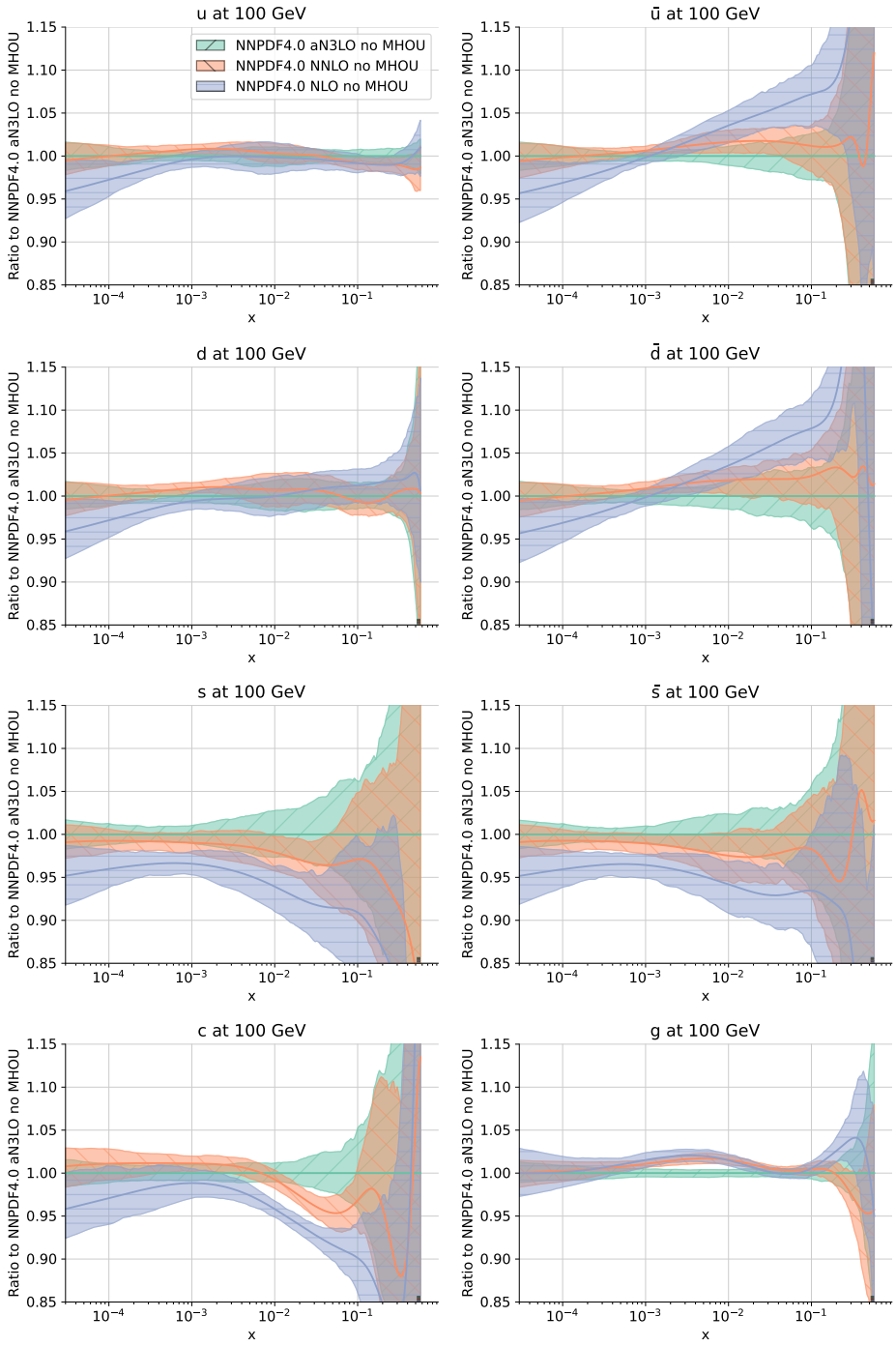


Figure 2.13: The NLO, NNLO and aN^{3}LO NNPDF4.0 PDFs at $Q = 100$ GeV. We display the up, antiup, down, antidown, strange, antistrange, charm and gluon PDFs normalized to the aN^{3}LO result. Error bands correspond to one sigma PDF uncertainties, not including MHOUs on the theory predictions used in the fit [6].

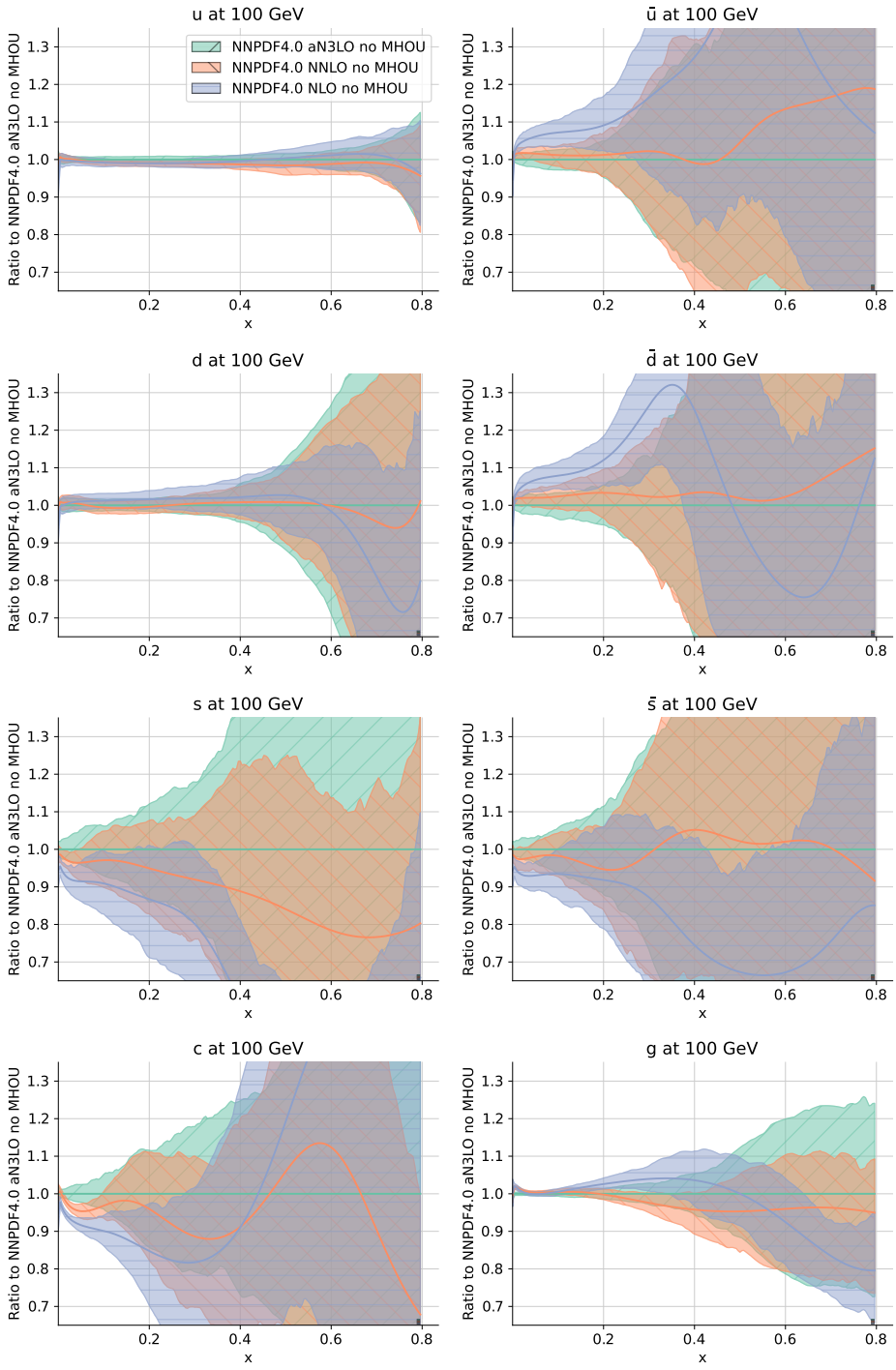


Figure 2.14: Same as fig. 2.13 in linear scale [6].

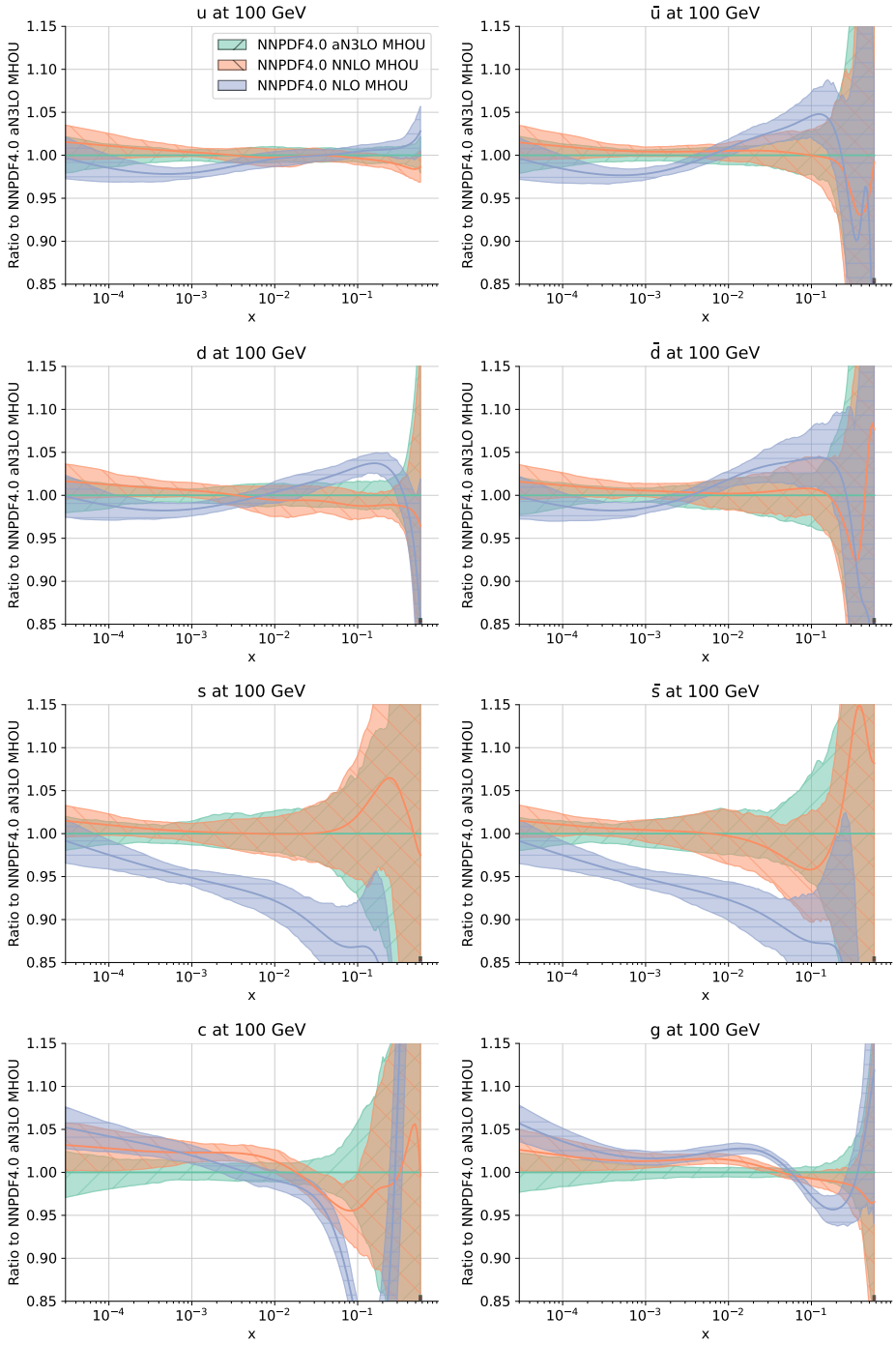


Figure 2.15: Same as fig. 2.13 for NNPDF4.0MHOU PDF sets. Error bands correspond to one sigma PDF uncertainties also including MHOUs on the theory predictions used in the fit [6].

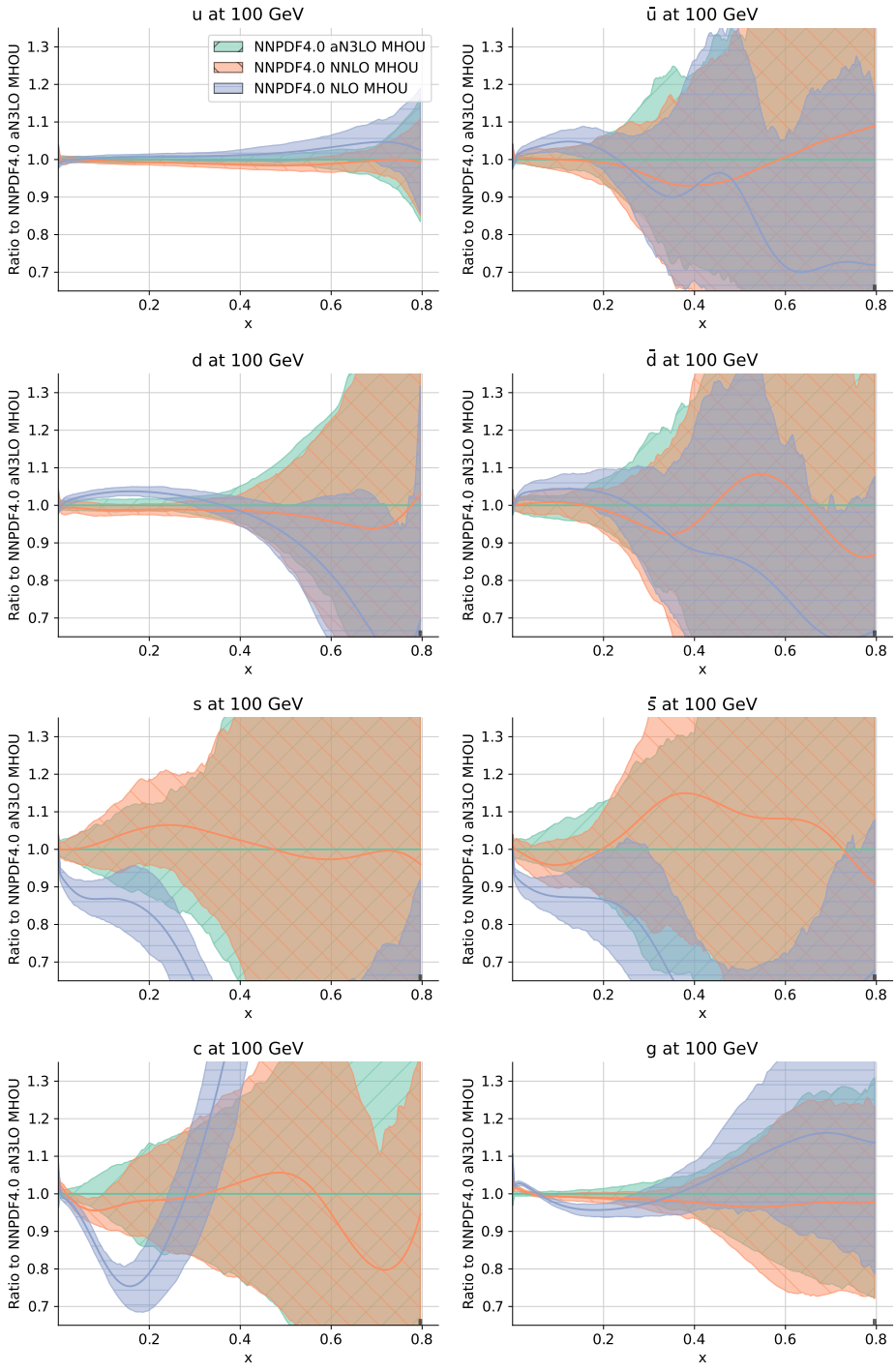


Figure 2.16: Same as fig. 2.15 in linear scale [6].

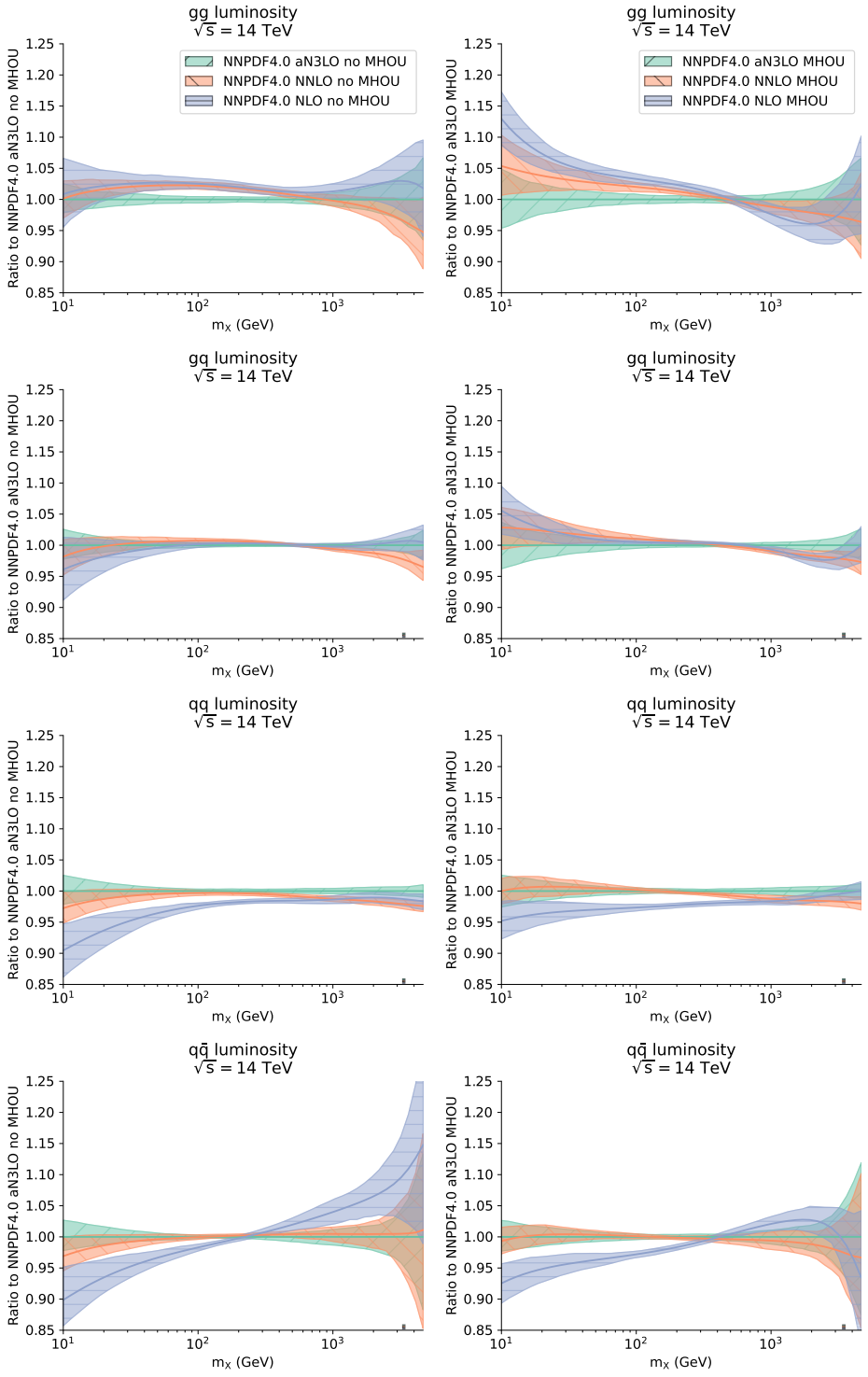


Figure 2.17: The gluon-gluon, gluon-quark, quark-quark, and quark-antiquark parton luminosities as a function of m_X at $\sqrt{s} = 14$ TeV, computed with NLO, NNLO and aN^{3}LO NNPDF4.0 PDFs without MHOUs (left) and with MHOUs (right), all shown as a ratio to the respective aN^{3}LO results [6].

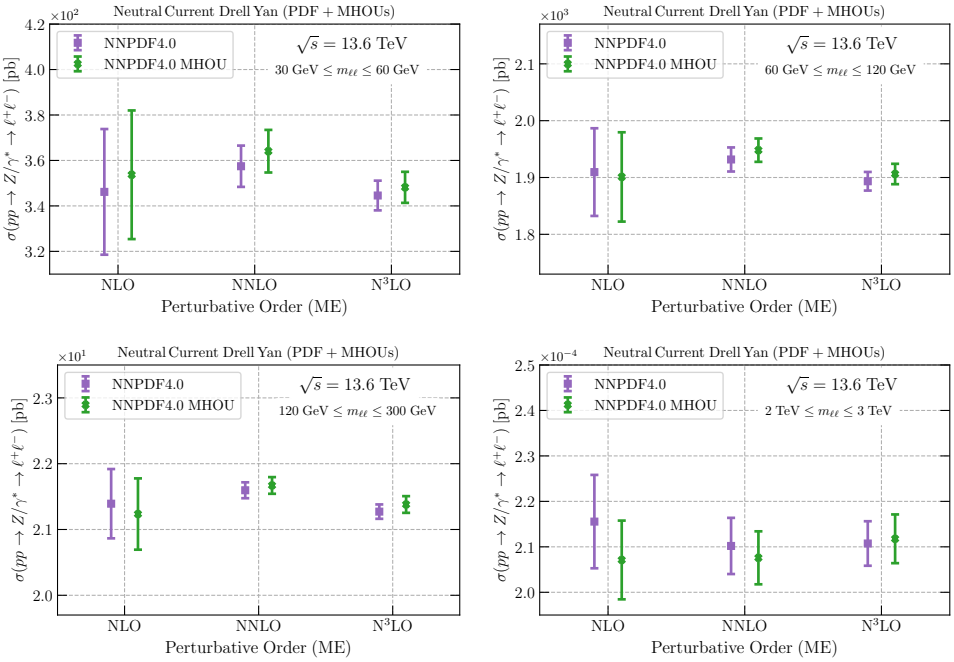


Figure 2.18: The inclusive neutral-current Drell-Yan production cross-section, $pp \to \gamma^*/Z \to \ell^+\ell^-$, for different ranges of the dilepton invariant mass $Q = m_{\ell\ell}$, from low to high invariant masses. Results are shown comparing NLO, NNLO and aN^{3}LO with matched perturbative order in the matrix element and PDF, with PDFs without and with MHOUs [6].

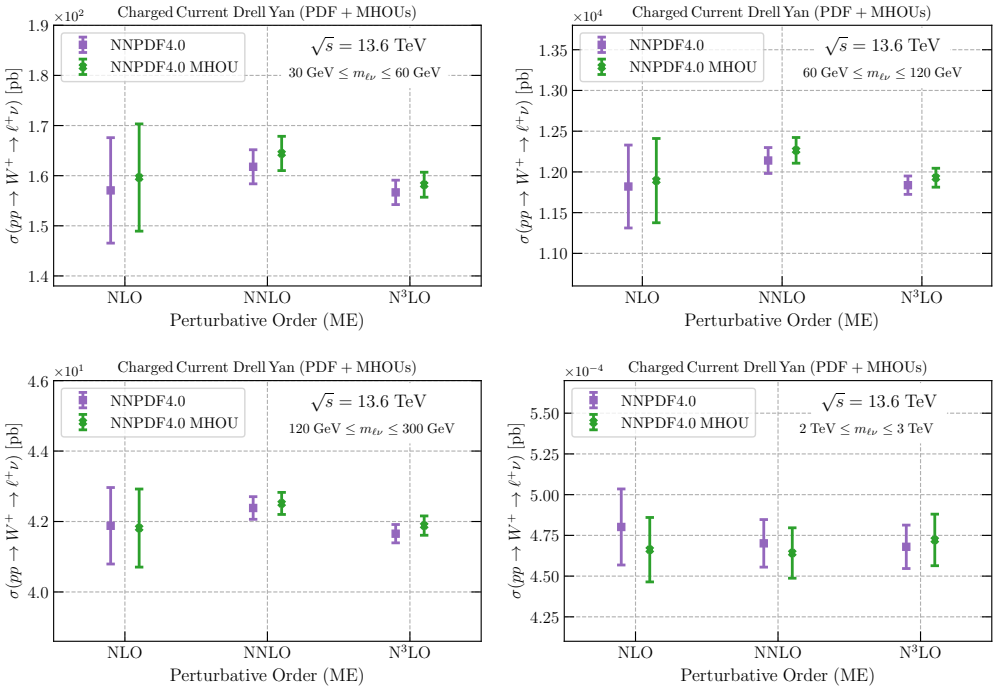


Figure 2.19: Same as fig. 2.18 for the inclusive charged-current Drell-Yan production cross-section, $pp \to W^+ \to \ell^+\nu_\ell$ [6].

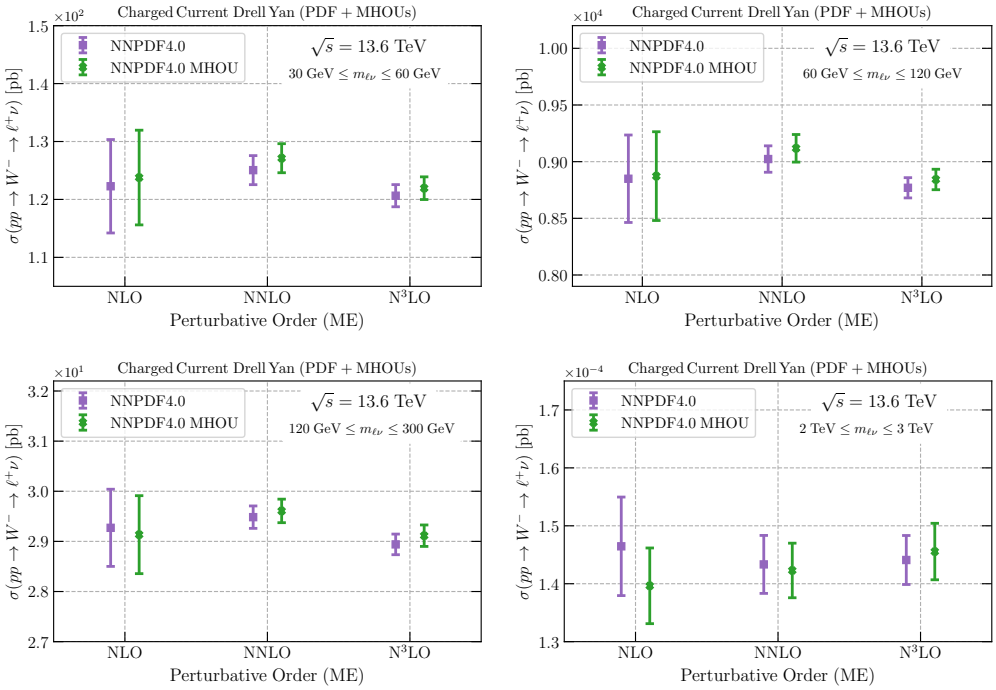


Figure 2.20: Same as fig. 2.18 for the inclusive charged-current Drell-Yan production cross-section, $pp \to W^- \to \ell^- \bar{\nu}_\ell$ [6].

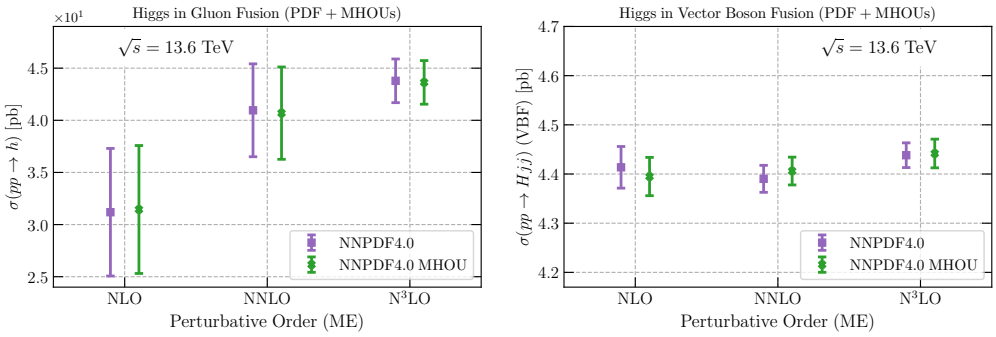


Figure 2.21: Same as fig. 2.18 for Higgs production in gluon-fusion and via vector-boson fusion.

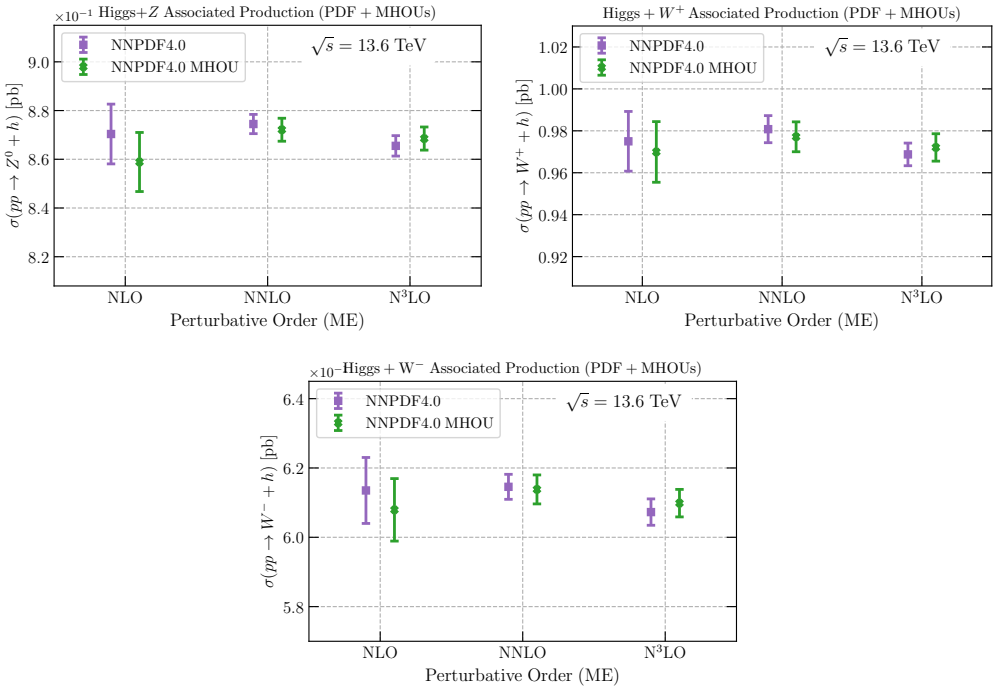


Figure 2.22: Same as fig. 2.18 for Higgs production in association with W^+ and Z gauge bosons: from top to bottom, Zh, W^+h, and W^-h.

CHAPTER 3

Validation of the methodology: Closure Tests

Parton Distribution Functions (PDFs) are a fundamental component of theoretical predictions at hadron colliders (see chapter 1); for recent reviews, see [115, 116].

The determination of PDFs from a finite set of experimental data is a classic example of an inverse problem, inferring a model from noisy and sometimes incompatible observations. Inverse problems are inherently challenging, and Machine Learning (ML) has become an increasingly prominent tool for addressing them; comprehensive reviews can be found in [117, 118]. Within the context of PDF fitting, the NNPDF collaboration has been leveraging ML techniques for over a decade (see section 2.1), in particular adopting a deep neural network to parametrize PDFs combined with a bootstrap procedure to propagate data fluctuations into the fitted PDFs.

The closure testing methodology, introduced in [119] and following suggestions from [120] and studies in [121], assesses the robustness and effectiveness of global PDF fits. A detailed theoretical discussion of the statistical basis for this methodology is presented in [9]. Closure tests involve fitting artificial data generated from a *known* set of input PDFs. Since the underlying law is known, this allows for a direct comparison of fit results to the true values, thus evaluating whether the fitting methodology can accurately reproduce the central values of the underlying law and correctly propagate experimental uncertainties.

So far [8], closure tests have been conducted on a set of artificially generated data that is inherently consistent, as it is produced using the *known* underlying law with the experimental uncertainties and correlations provided by experimentalists. In this chapter, we discuss the impact of inconsistencies of experimental origin within the training dataset. Closure tests with inconsistent training data aim to simulate scenarios where specific systematic uncertainties may have been either underestimated or entirely overlooked by experimentalists, leading to tension between different experimental observations. By employing closure tests in a context where inconsistency is deliberately introduced into the data, the conditions of the test are made more representative of real-world situations.

As a by-product of our investigation, we develop a more robust estimate of the reliability of uncertainties in a closure test. We achieve this by refining the evaluation of a key indicator, the bias-to-variance ratio, compared to the definition previously provided in [9].

We also extend the application of the closure test methodology to another significant objective: the precise determination of the strong coupling constant, α_s. Specifically, we demonstrate how closure tests can be utilized to validate various methodologies employed to extract the strong coupling from experimental data. This validation is crucial, considering the unprecedented accuracy reached in such determination.

This chapter is organized as follows. In section 3.1 we introduce the notation and the definitions that are relevant in the context of a closure test. We also review the statistical estimators first introduced in [9] and we propose some improvements. Then we discuss the problem of assessing the impact of inconsistent data on a NNPDF fit, by means of a closure test, in section 3.2 [11]. Finally, in section 3.3, we review some of the methodologies that can be employed to extract α_s from experimental data and we describe how the closure tests framework can be utilized to validate such methodologies.

3.1 Notation and Definitions

In this section, we introduce the concept of a closure test, as well as the required definitions, and we introduce the notation. Some of these definitions have already been outlined in section 2.1, to which we will refer. In section 3.1.1, we revisit some of the statistical estimators utilized in [8] and explain how they can be refined to yield more reliable results.

We start by expressing the χ^2 of eq. (2.8) in a different notation, which will be useful in the following. For the k-th replica, we have

$$\chi^{2(k)} = \frac{1}{N_{\text{data}}} \left(\mathcal{G}(u_k) - \mu^{(k)}\right)^T C^{-1} \left(\mathcal{G}(u_k) - \mu^{(k)}\right) , \tag{3.1}$$

where $\mu^{(k)}$ are the Level-2 data for replica k (see eq. (2.3)), C is the experimental covariance matrix (see eq. (2.9)) and $\mathcal{G}(u_k)$ represents the forward map from the PDF model to the observable space. The u_k are in fact the model parameters defining the k-th replica PDF. In chapter 2 we denoted $\mathcal{G}(u_k)$ with P, but here we need to switch to a more detailed notation. Note that the map itself is often referred to as *FK-table*. We will discuss in detail the *FK-table* interface in chapter 4.

The fundamental concept of closure tests is based on the assumption that the *true* model of nature, denoted as w, is known. This true model is used to compute the true values of each measurable observable $i \in (1, \ldots, N_{\text{data}})$, by means of the forward map $\mathcal{G}$, as

$$L_{0,i} = \mathcal{G}(w)_i \, . \tag{3.2}$$

As mentioned in section 2.1, we refer to $L_{0,i}$ as Level-0 data.

In an L_0 closure test executed with the NNPDF methodology, no stochastic noise is introduced to the L_0 data. Consequently, the N_{reps} fits are performed on an identical set of data, but with varying seeds for the initialization of the random numbers employed in the minimization process. Therefore, the fit quality can be arbitrarily high, implying that the χ^2 for each replica should tend towards zero. This renders the L_0 test a significant evaluation of the minimization algorithm's efficiency.

At Level 1 (L_1) the experimental central values are artificially generated according to eq. (2.1), with

$$y_0 = \mathcal{G}(w) + \eta \, , \tag{3.3}$$

where the observational noise η is pseudo-randomly generated from the assumed distribution. Specifically, each L_1 data is given by

$$L_{1,i} = L_{0,i} + \sum_j (\sqrt{C})_{i,j} r_j \, , \tag{3.4}$$

where i indicates the datapoint and $\sqrt{C}$ represents the Cholesky decomposition of the experimental covariance matrix including the statistic, additive and multiplicative systematic experimental uncertainties and model uncertainties for each dataset, while r_i are random numbers generated from a standard normal distribution[1].

At L_1, the same underlying data is used for each replica fit. However, similar to the L_0 level, a different random seed is employed for the initialization of the random numbers used in the minimization process. Since the L_1 data are, on average, fluctuated by one standard deviation away from the L_0 values, it is expected that in L_1 closure tests, the χ^2 of the best fit will be approximately 1.

Finally, at Level 2 (L_2), starting from the shifted pseudo-data in eq. (3.4), we generate N_{reps} pseudo-data $L_2^{(k)}$ given by

$$L_{2,i}^{(k)} = L_{1,i} + \sum_j (\sqrt{C})_{i,j} r_j^{(k)}, \tag{3.5}$$

where $r_i^{(k)}$, as in eq. (3.4), are random numbers sampled from a standard normal distribution. In a L_2 fit, we expect the final error function to be close to 2, since the data are generated by adding an additional layer of fluctuations compared to L_1. In the following we will only discuss L_2 closure tests.

3.1.1 Statistical estimators

In [8, 9], various statistical estimators were introduced to evaluate the faithful propagation of experimental uncertainties into the PDF space within the context of a closure test [2]. Such estimators are defined within the framework of a multi-closure test, where they are computed across multiple closure test fits, with each fit performed on a different instance of L_1 data. In this work, we use the index l to indicate one of the N_{fits} that we perform across instances of the L_1 data, and the index k to indicate the N_{reps} pseudo-data replicas fitted for each of the N_{fits}. In total, we have ensembles of $N_{\text{reps}} \times N_{\text{fits}}$ replicas.

A key estimator in our analysis is the *bias* computed on each of the individual l fits, which measures the distance between the central value of the model replica predictions and the vector of the true observable values, $f \equiv L_{0,i}$, in units of the covariance matrix. It reads

$$B^{(l)}(C) = \left(\mathbb{E}_\epsilon \mathcal{G}(u_{*,k}) - f\right)^T C^{-1} \left(\mathbb{E}_\epsilon \mathcal{G}(u_{*,k}) - f\right) , \tag{3.6}$$

where $\mathbb{E}_\epsilon$ denotes the expectation value over replicas.

Another estimator is the *variance*, which characterizes the fluctuations of model predictions around their mean value in units of the covariance matrix. It is defined as

$$V^{(l)}(C) = \mathbb{E}_\epsilon \left[\left(\mathbb{E}_\epsilon \mathcal{G}(u_{*,k}) - \mathcal{G}(u_{*,k})\right)^T C^{-1} \left(\mathbb{E}_\epsilon \mathcal{G}(u_{*,k}) - \mathcal{G}(u_{*,k})\right) \right] . \tag{3.7}$$

Note that the variance can be computed also for fits to real experimental data, not only in the context of a closure test. In contrast, the bias can only be evaluated within the

[1] It is important to note that in a closure test, the L_0 data are used instead of experimental data for the generation of the multiplicative uncertainties contribution to the covariance matrix. Specifically, this modification involves replacing the experimental covariance matrix with the t_0 covariance matrix [122].

[2] In this section, we will refrain from discussing statistical estimators that are not pertinent to the results presented in this thesis. For an exhaustive review, readers are encouraged to consult [8, 9].

framework of a closure test, as the true underlying law f is typically unknown.

To assess the fidelity of the PDF uncertainties, one approach involves computing the expectation $\mathbb{E}_\eta$ of both bias and variance across the N_{fits} fits, each obtained with a randomly selected value of y_0, i.e. for different instances of L_1 data. The evaluation then proceeds by taking the square root of the *bias-to-variance ratio* (since both are squared quantities), defined as

$$R_{bv} = \sqrt{\frac{\mathbb{E}_\eta B^{(l)}(C)}{\mathbb{E}_\eta V^{(l)}(C)}}\,. \tag{3.8}$$

If the uncertainties associated with the PDF replicas are accurate, the bias-to-variance ratio should ideally equal one. This implies that the average discrepancy between the central prediction from the replicas and the true value matches the mean-square difference between replica predictions and their central values [8, 9].

The quantity R_{bv} serves as an indicator of how much the uncertainty might have been over- or under-estimated. Specifically, if R_{bv} deviates from unity, it suggests that the uncertainty for a given fit is, on average, over- or under-estimated by a factor of $1/R_{bv}$.

Improved estimators

In the course of our investigation, we have identified that the previously defined bias-to-variance ratio may lead to biased outcomes. To illustrate this issue, let us consider a scenario involving two experimentally uncorrelated datasets. For such cases, the ratio is expressed as

$$R_{bv} = \sqrt{\frac{\mathbb{E}_\eta B^{(l)}(C_{d_1}) + \mathbb{E}_\eta B^{(l)}(C_{d_2})}{\mathbb{E}_\eta V^{(l)}(C_{d_1}) + \mathbb{E}_\eta V^{(l)}(C_{d_2})}}\,, \tag{3.9}$$

where $B^{(l)}(C_{d_i})$ and $V^{(l)}(C_{d_i})$ denote the bias and variance computed using the portion of the total covariance matrix associated with dataset i for fit l. This formulation reveals that R_{bv} tends to favor datasets with larger absolute values in both bias and variance.

Specifically, in situations where $B^{(l)}(C_{d_1}) \gg B^{(l)}(C_{d_2})$ and $V^{(l)}(C_{d_1}) \gg V^{(l)}(C_{d_2})$, the ratio approximately simplifies to

$$R_{bv} \approx \sqrt{\frac{\mathbb{E}_\eta B^{(l)}(C_{d_1})}{\mathbb{E}_\eta V^{(l)}(C_{d_1})}}\,. \tag{3.10}$$

It is noteworthy that under the previously employed definition, both bias and variance are independent of the number of data points in each dataset. Consequently, this can result in the unpleasant scenario where a single data point carries significantly more weight than a larger dataset.

The crux of the issue lies in the covariance matrix used in the fit, which incorporates correlations among observables induced by both experimental and model factors. This matrix is not the appropriate metric for assessing the proximity between central replicas and the underlying law. Instead, the relevant covariance matrix should reflect the PDF-induced correlations among observables included in the fit.

To clarify this concept, consider a population of random multivariate observables

computed from each replica in a specific fit l, based on a given instance of the L_1 dataset:

$$\mathcal{P}_1^{(l)} : \left\{ \mathcal{G}(u_{*,1}^{(l)}), \dots, \mathcal{G}(u_{*,N_{\text{reps}}}^{(l)}) \right\} . \tag{3.11}$$

Assuming Gaussianity, we can estimate the covariance matrix of this population $\mathcal{P}_1^{(l)}$ as

$$\left(C_{\text{obs}}^{\text{rep},(l)}\right)_{ij} = \frac{N_{\text{reps}}}{N_{\text{reps}}-1} \left(\mathbb{E}_\epsilon \left[\mathcal{G}(u_{*,k}^{(l)})_i \mathcal{G}(u_{*,k}^{(l)})_j \right] - \mathbb{E}_\epsilon \mathcal{G}(u_{*,k}^{(l)})_i \mathbb{E}_\epsilon \mathcal{G}(u_{*,k}^{(l)})_j \right) , \tag{3.12}$$

where the expectation value is computed over the replicas k for a given fit l. It is worth noticing that in the large N_{reps} limit only the mean of $\mathcal{P}_1^{(l)}$ depends on the fit index l, while the variance should be approximatively independent of it.

To ensure stable results, we estimate $C_{\text{obs}}^{\text{rep}}$ as the average of the covariance matrices estimated from each fit l:

$$C_{\text{obs}}^{\text{rep}} = \frac{1}{N_{\text{fits}}} \sum_{l=1}^{N_{\text{fits}}} C_{\text{obs}}^{\text{rep},(l)} . \tag{3.13}$$

Using the covariance matrix defined in eq. (3.12), we redefine the variance estimator as

$$V^{(l)}(C_{\text{obs}}^{\text{rep}}) = \mathbb{E}_\epsilon \left[\left(\mathbb{E}_\epsilon \mathcal{G}(u_{*,k}^{(l)}) - \mathcal{G}(u_{*,k}^{(l)}) \right)^T (C_{\text{obs}}^{\text{rep}})^{-1} \left(\mathbb{E}_\epsilon \mathcal{G}(u_{*,k}^{(l)}) - \mathcal{G}(u_{*,k}^{(l)}) \right) \right] . \tag{3.14}$$

This quantity is a random variable dependent on the shift η and follows a χ^2 distribution with N_{data} degrees of freedom. Therefore, taking the expectation value over the fits yields

$$\mathbb{E}_\eta V^{(l)}(C_{\text{obs}}^{\text{rep}}) = N_{\text{data}} , \tag{3.15}$$

which has been explicitly verified in section 3.2.

Now, to test the faithfulness of the uncertainties in a PDF closure test fit, we aim to verify that the central value of a fit l is "close enough" to the underlying law f. Here, "close enough" is defined in terms of the variance of the fit. Specifically, we consider a fit faithful if its central value lies within 1σ of the underlying law.

To this end, we define the bias as

$$B^{(l)}(C_{\text{obs}}^{\text{rep}}) = \left(\mathbb{E}_\epsilon \mathcal{G}(u_{*,k}^{(l)}) - f \right)^T (C_{\text{obs}}^{\text{rep}})^{-1} \left(\mathbb{E}_\epsilon \mathcal{G}(u_{*,k}^{(l)}) - f \right) . \tag{3.16}$$

The new bias-variance ratio, in terms of these estimators, is given by

$$R_{bv} = \mathbb{E}_\eta \sqrt{\frac{B^{(l)}(C_{\text{obs}}^{\text{rep}})}{V^{(l)}(C_{\text{obs}}^{\text{rep}})}} = \mathbb{E}_\eta \sqrt{\frac{B^{(l)}(C_{\text{obs}}^{\text{rep}})}{N_{\text{data}}}} \tag{3.17}$$

and can be interpreted as an average distance between the central values of the fits and the underlying law. In appendix C, we discuss in further detail the impact of the new definition in eq. (3.17) within a real-case scenario. Note however that the application of eq. (3.17) requires some care, as we will describe in the following section.

Another equivalent estimator, as considered in [9, 52], is a quantile estimator that

measures the fraction of fits for which the input PDF lies within the 1σ interval of the central PDF, averaged over PDF flavors and values of x. In our current analysis, we compute an analogous estimator in the space of experimental data rather than PDF space, given by

$$\xi_{1\sigma} = \frac{1}{N_{\rm data}} \sum_{i=1}^{N_{\rm data}} \frac{1}{N_{\rm fits}} \sum_{l=1}^{N_{\rm fits}} I_{[-\sigma_i', \sigma_i']} \left(\delta_{im} W_{mn} (\mathbb{E}_\epsilon \mathcal{G}(u_{*,k}^{(l)}) - f)_n \right), \tag{3.18}$$

where W is the $N_{\rm data} \times N_{\rm data}$ dimensional matrix that diagonalizes the matrix $C_{\rm obs}^{\rm rep}$, such that

$$\Lambda = W^T C_{\rm obs}^{\rm rep} W, \text{ with } \Lambda_{ii} = (\sigma_i')^2, \ \Lambda_{ij} = 0 \text{ for } i \neq j, \tag{3.19}$$

and σ_i denotes the PDF uncertainty associated with the experimental observation i. The function $I_A(x)$ represents the indicator function of the interval A, which equals one if its argument lies within the interval A, and zero otherwise. Note that in eq. (3.18), the sum over the repeated indices m and n is implicit.

For a successful closure test, one should find that $\xi_{1\sigma} \sim 0.68$ if the PDF uncertainties are correctly estimated. It is important to note that this relies on the assumption that both the PDF replicas and the expectation values of the observables across fits are Gaussianly distributed. This assumption holds by construction for the closure test data, and it is likely valid for the observables computed with the fitted closure test PDFs for those observables that are sensitive to PDF combinations and kinematic regions well-constrained by the fitting data.

The uncertainties of $\xi_{1\sigma}$ and R_{bv} are determined as the standard deviation over a bootstrap sample performed on both fits and replicas, as detailed in appendix D.

A useful graphical representation of the quantile estimator defined in eq. (3.18) is achieved by binning the difference between the mean value (over replicas) of the theory predictions and the corresponding true observable values, normalized by the PDF uncertainties. This difference is defined as

$$\delta_i^{(l)} = \frac{W_{ij} (\mathbb{E}_\epsilon \mathcal{G}(u_{*,k}^{(l)}) - f)_j}{\sigma_i'}, \tag{3.20}$$

and we will present several such plots in section 3.2.

PCA and single data point analysis definition

The definition of the bias in eq. (3.16) presupposes the invertibility of the covariance matrix $C_{\rm obs}^{\rm rep}$. However, a practical challenge may arise when computing $B^{(l)}(C_{\rm obs}^{\rm rep})$ on a given dataset. Depending on the dataset's size, which can range from a handful of points to the entire set of data points included in the fit, strong PDF-induced correlations may render the covariance matrix estimated from the samples ill-defined and non-invertible. To address this issue, we have identified two alternative approaches, each providing complementary information:

- Restrict the computation of R_{bv} to single data points.
- Regularize the covariance matrix using a *Principal Component Analysis* (PCA) approach and utilize the regularized matrix to compute R_{bv} for arbitrary groups of

datasets.

When R_{bv} is computed on a dataset or on a group of datasets, we can assess whether PDF-induced correlations on the observables are faithfully estimated. However, in this case we are forced to rely on PCA, thus we lose the ability to interpret the impact of specific experimental points on R_{bv}.

Computing R_{bv} for each individual data point ignores correlations but retains specific information and simplifies the computation. Maintaining both sets of information allows us to investigate the bias-variance ratio problem from multiple perspectives.

While the latter approach is straightforward, the PCA method involves a certain degree of arbitrariness and some complications, especially when R_{bv} is computed for a heterogeneous group of datasets. Thus, we provide practical details on performing PCA.

The main idea of PCA is to project the matrix onto a lower-dimensional space defined by a basis that maximizes variance among theoretical predictions. The following steps outline the regularization of the covariance matrix $C_{\text{obs}}^{\text{rep}}$ as defined in eq. (3.12).

1. Choose the number of components to retain based on the explained variance ratio (EVR), i.e. retain N_{pc} components such that

$$\frac{\sum_{i=1}^{N_{\text{pc}}} \lambda_i}{\sum_{i=1}^{N_{\text{data}}} \lambda_i} \tag{3.21}$$

   meets a specified threshold, e.g., 0.99, where λ_i are the eigenvalues of the covariance matrix.

2. Diagonalize the covariance matrix: $C_{\text{obs}}^{\text{rep}} = W\Lambda W^T$ and construct the matrix of reduced components (eigenvectors) $\tilde{W} \in \mathbb{R}^{N_{\text{data}} \times N_{\text{eig}}}$ by retaining only the N_{pc} components with the largest eigenvalues.

3. Construct the regularized $N_{\text{pc}} \times N_{\text{pc}}$ covariance matrix: $\tilde{C}_{\text{obs}}^{\text{rep}} = \tilde{W}^T C_{\text{obs}}^{\text{rep}} \tilde{W}$.

4. Compute the variance (and bias with corresponding regularization) in the reduced space:

$$V^{(l)} = \mathbb{E}_\epsilon \left(\mathbb{E}_\epsilon \mathcal{G}(u_*^{(k)}) - \mathcal{G}(u_*^{(k)}) \right)^T \tilde{W} \tilde{\Lambda}^{-1} \tilde{W}^T \left(\mathbb{E}_\epsilon \mathcal{G}(u_*^{(k)}) - \mathcal{G}(u_*^{(k)}) \right), \tag{3.22}$$

   where $\tilde{\Lambda}$ is the $N_{\text{pc}} \times N_{\text{pc}}$ matrix of eigenvalues.

When this methodology is applied to a heterogeneous group of datasets, where the entries of the covariance matrix $C_{\text{obs}}^{\text{rep}}$ are expressed in different units, we apply steps 1 and 2 to the *correlation matrix* instead,

$$(\rho_{\text{obs}}^{\text{rep}})_{ij} = \frac{(C_{\text{obs}}^{\text{rep}})_{ij}}{(C_{\text{obs}}^{\text{rep}})_{ii} (C_{\text{obs}}^{\text{rep}})_{jj}}, \tag{3.23}$$

which, being normalized, avoids the units of measure problem. Once the reduced components $\tilde{W}$ for the correlation matrix are obtained, the reduced covariance matrix can be constructed as in step 3, and bias and variance can be computed as in step 4. In section 3.2, this solution will be employed whenever R_{bv} is computed for more than one dataset.

It is crucial to note that the first step in this methodology introduces an element of arbitrariness. Selecting an EVR that is too low might result in the loss of valuable information from the samples, whereas choosing an EVR too close to one could lead to an unstable covariance matrix.

To balance the number of components and the EVR, we evaluate the L_2 condition number of the covariance matrix in Eq. (3.12), which provides insight into its stability as the number of components varies. The L_2 condition number of the covariance matrix is computed as the ratio of the largest to the smallest eigenvalues:

$$\kappa(C_{\text{obs}}^{\text{rep}}) = \frac{|\lambda_{\max}(C_{\text{obs}}^{\text{rep}})|}{|\lambda_{\min}(C_{\text{obs}}^{\text{rep}})|}. \tag{3.24}$$

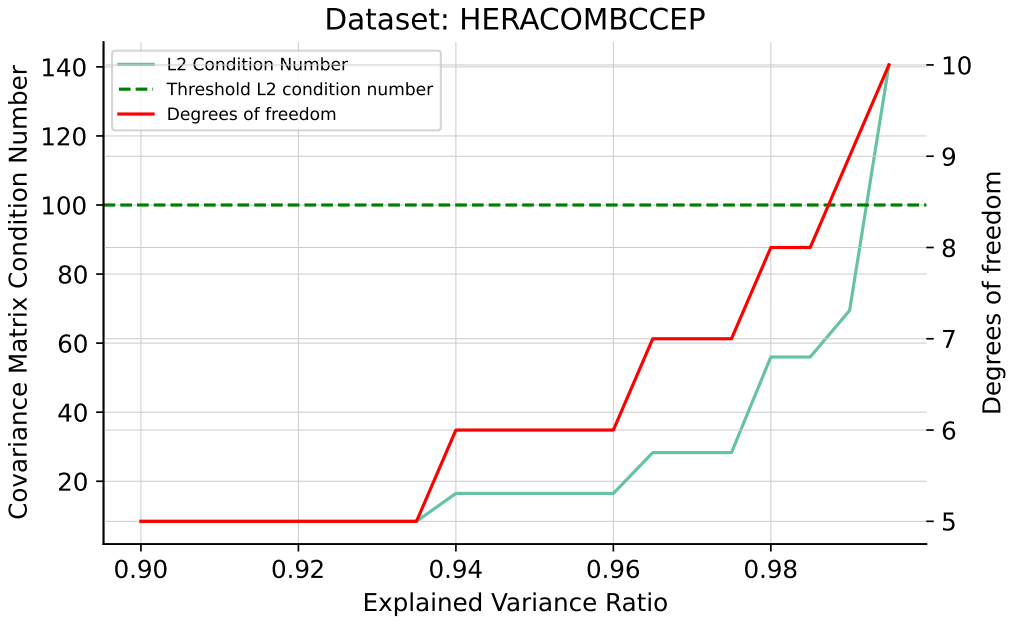


Figure 3.1: L_2 condition number for the HERA I+II inclusive CC dataset as a function of the explained variance ratio.

Fig. 3.1 illustrates the condition number as a function of the EVR for a charged current DIS dataset. The plot demonstrates that the explained variance ratio, and consequently the number of components to retain, can be chosen based on a threshold value for the covariance matrix condition number. Specifically, the threshold value is indicated by the green dashed line in fig. 3.1, which intersects the condition number curve (the green solid line) at about 0.99. In this case a threshold value of 100 has been chosen.

3.2 Tests on inconsistent Data

The fundamental premise of closure testing involves utilizing a specified PDF set that serves as a proxy for the *true* proton structure, alongside a theoretical model calculated at a particular perturbative accuracy, to compute the partonic cross sections and generate

a set of artificial data points. These data points are ideal in the sense that they exhibit known statistical properties, lack internal inconsistencies, and are fully compatible with the theoretical model used in their creation. Consequently, in a standard closure test, the objective is to accurately reproduce the underlying PDF within the appropriate margins of uncertainty.

In this section, we introduce the next stage in the evolution of closure tests: *inconsistent closure tests*. We begin in section 3.2.1 by clearly defining what we mean by *inconsistent* and detailing the practical methods for introducing experimental inconsistencies into the artificial data set. In section 3.2.2, we outline the practical settings for the implementation. In section 3.2.3 we analyze the results of this study in some cases in terms of the statistical estimators introduced in section 3.1.1.

3.2.1 Methodology

As discussed at the beginning of this chapter, inconsistencies of an experimental nature arise when some experimental uncertainties are either underestimated or entirely overlooked. Consequently, the nominal standard deviation is smaller than the *true* one, and due to the correlated multi-dimensional nature of measurements, correlations might be miscalculated, leading to tension between different experimental observations.

Although experimentalists strive to precisely estimate the uncertainties associated with measurements, accurately determining systematic uncertainties and their correlations is a complex task, especially in the presence of highly correlated data with small statistical uncertainties. Therefore, exploring the outcomes of closure tests in the presence of experimental *inconsistencies* makes the closure test setup more representative of real-world scenarios.

To formalize the definition of inconsistency within the context of a NNPDF-like closure test, we first explicitly express the covariance matrix as

$$(C)_{ij} = \delta_{ij}\sigma_i^{(\text{uncorr})}\sigma_j^{(\text{uncorr})} + \sum_{m=1}^{N_{\text{mult}}} \sigma_{i,m}^{(\text{mult})}\sigma_{j,m}^{(\text{mult})} D_i D_j + \sum_{k=1}^{N_{\text{add}}} \sigma_{i,k}^{(\text{add})}\sigma_{j,k}^{(\text{add})} + (C_{\text{th}})_{ij} \ ,$$

where $\sigma_i^{(\text{uncorr})}$ denotes the uncorrelated systematics, $\sigma_i^{(\text{mult})}$ represents the correlated multiplicative systematics, and $\sigma_i^{(\text{add})}$ denotes the correlated additive systematics. In principle, $(C_{\text{th}})_{ij}$ incorporates all contributions of theoretical nature, not just the model uncertainties of the nuclear datasets (discussed in section 2.1). However, in the context of a closure test, the theoretical model is by definition correct, meaning that there are no Missing Higher-Order Uncertainties (MHOU) associated with the employed theoretical predictions. A possible exception to this would be generating L_1 data with a NNLO theory and then performing the fit with a NLO theory. In this case, we could utilize the closure test framework to assess the impact of adopting the methodology described in chapter 2. This would be an interesting study, but it is outside the scope of this work.

To simulate the scenario in which the experimental portion of the total covariance matrix has been inaccurately estimated, we introduce C_1 and C_2, representing the covariance matrices used to generate Level-1 and Level-2 data respectively, i.e.

- Level-1 data: $L_1 = L_0 + \eta$, where $\eta \sim \mathcal{N}(0, C_1)$.
- Level-2 data: $L_2^{(k)} = L_1 + \epsilon^{(k)}$, where $\epsilon^{(k)} \sim \mathcal{N}(0, C_2)$, for $k = 1, \ldots, N_{\text{reps}}$.

In a consistent closure test, the covariance matrices for generating L_1 and L_2 data are identical:

$$C_1 = C_2 = C\,. \tag{3.25}$$

In a closure test where we introduce experimental inconsistencies, we have:

$$C_1 = C \quad \text{and} \quad C_2 = C(\lambda)\,, \tag{3.26}$$

with

$$\begin{aligned}[C(\lambda)]_{ij} = \delta_{ij}\sigma_i^{(\text{uncorr})}\sigma_j^{(\text{uncorr})} + \sum_{m=1}^{N_{\text{mult}}} \lambda_m^{(\text{mult})}\sigma_{i,m}^{(\text{mult})}\lambda_m^{(\text{mult})}\sigma_{j,m}^{(\text{mult})} D_i D_j \\ + \sum_{k=1}^{N_{\text{add}}} \lambda_k^{(\text{add})}\sigma_{i,k}^{(\text{add})}\lambda_k^{(\text{add})}\sigma_{j,k}^{(\text{add})} + (C_{\text{th}})_{ij}\,,\end{aligned} \tag{3.27}$$

where the rescaling factors $\lambda_m^{(\text{mult})}$ and $\lambda_k^{(\text{add})} \in [0,1]$ can vary for each systematic uncertainty.

These modifications clearly simulate a miscalculation in the published experimental covariance matrix: we generate the L_2 data replicas *as if* C had underestimated correlations, adjusted by the λ_m factor, relative to the true covariance matrix C_1, potentially causing shifts in the central values from the underlying truth.

To visualize the effect of rescaling multiplicative systematics by a factor $\lambda < 1$, fig. 3.2 presents a simple $2D$ case to illustrate the impact of tuning the λ parameter. In this illustrative example, X and Y represent two observables, with underlying true values $X = 0$ and $Y = 0$. The covariance matrix is given by

$$C(\lambda) = \begin{pmatrix} 1.1 & \lambda \\ \lambda & 1.1 \end{pmatrix}. \tag{3.28}$$

In this example, $\lambda = \text{Var}(X, Y)$ and it represents the tuned correlated uncertainties. By scanning λ from 1, corresponding to the usual closure test with consistent data, to $\lambda = 0$, representing the maximum level of inconsistency in our setup, we observe that as λ decreases, the information on correlations between observables is lost. Consequently, the ellipses change shape until they become circles.

This simple example illustrates how changes in R_{bv}, as the inconsistency injected into the artificial data becomes more severe, indicate the extent to which a given experimental inconsistency can compromise the accuracy of the PDF uncertainties. Essentially, we can directly examine this statistic to determine whether the fitting methodology *absorbs* or *flags* the experimental inconsistency that we artificially introduce.

3.2.2 Details on the setup

In all results presented in section 3.2.3, the L_1 data are generated using one of the replicas of the `NNPDF4.0` set as the underlying law f, consistent with the approach described in Sect. 5 of [9]. This replica is randomly drawn from a previous NNPDF fit to experimental data and typically exhibits more structure than the final central PDF, making it a more comprehensive choice than any single central fit. We refer to this as the *underlying law*, and the corresponding predictions are considered the true observable values. Note that,

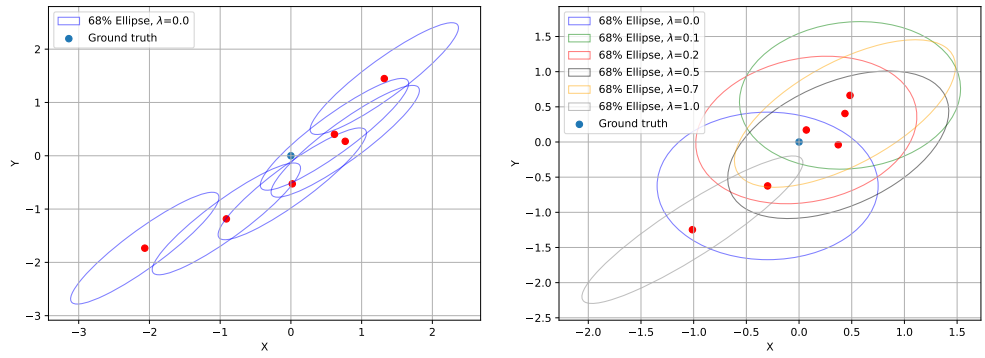


Figure 3.2: Left panel: the red dots are the L_1 generated instances, while the confidence ellipses show the distribution of the L_2 data, conditional on the respective L_1 data instances. The L_1 instances are generated according to a multi Gaussian centred in the underlying true value (the origin in this plot) with the same confidence ellipses as the L_2 ones. Right panel: The L_1 data are generated according to a multi Gaussian with same ellipse as the consistent case, but λ is varied to show how the generation of L_2 replicas changes as λ decreases.

while theoretically any function could serve as the underlying law, using a realistic input is practical and justified.

We conduct $N_{\text{fits}} = 25$ closure tests, each utilizing a different randomly sampled type of L_1 noise, consisting of $N_{\text{reps}} = 100$ replicas each. This choice is motivated by previous studies [8].

In each subsequent subsection, we specify the dataset(s) to which an inconsistency is introduced, the value(s) of λ associated with each correlated systematic uncertainty, and the strength of the inconsistency, which ranges from maximal ($\lambda = 0$) to minimal ($\lambda = 1$). Across the analyses presented in the following sections, we set $\lambda_k^{(\text{add})} = 1$ for all additive systematic uncertainties k, while for multiplicative systematic uncertainties m in selected datasets, we set $\lambda_m^{(\text{mult})} = \lambda \leq 1$. It is noteworthy that multiplicative systematics constitute the majority of the uncertainties in the HERA and LHC data considered here. Many of these uncertainties are correlated not only across different kinematic bins within the same measurement but also across different datasets within the same experiment. For instance, various ATLAS datasets are correlated through the luminosity uncertainty.

When presenting results for the ratio of bias to variance obtained using PCA regularization of the covariance matrix in the PDF space, we adopt an Explained Variance Ratio (EVR) of 0.99 across all datasets. This choice, as explained in section 3.1.1, aims to retain maximum information from the samples while regularizing the covariance matrix.

3.2.3 Results

In this section, we present some of the results obtained in several scenarios. We begin by illustrating the results obtained for DIS data only. Subsequently, we move to results obtained on global datasets. In particular, we discuss the cases in which the inconsistency is injected in Drell-Yan and in inclusive jets data.

Deep Inelastic Scattering

We begin by performing a multi-closure fit on DIS data, encompassing all DIS datasets included in the `NNPDF4.0` analysis [8] for lepton-nucleon and neutrino-nucleus scattering processes. This includes fixed-target neutral current (NC) structure function data from NMC [123, 124], SLAC [125], and BCDMS [126], fixed-target inclusive and dimuon charged current (CC) cross-section data from CHORUS [127] and NuTeV [128, 129], as well as collider NC and CC cross-section data from the HERA legacy combination [130] and combined measurements from H1 and ZEUS for reduced electron-proton NC DIS cross-sections involving open charm and bottom quarks [131].

As is customary in all NNPDF fits, we exclude the region where higher twist corrections might affect the reliability of the perturbative expansion ($Q^2 < 3.5\,\text{GeV}^2$ and $W^2 < 12.5\,\text{GeV}^2$). Additionally, for fits involving the charm PDF, a stricter Q^2 cut is applied to the HERA I+II σ^c_{NC} dataset at NNLO ($Q^2 < 8\,\text{GeV}^2$) to minimize potential impacts from missing NNLO terms related to initial-state charm (see Sect. 2.2 in [132]).

The DIS dataset is partitioned into an *in-sample* subset, which is included in the fit, and an *out-of-sample* subset, which is excluded from the fit. The partition shown in fig. 3.3

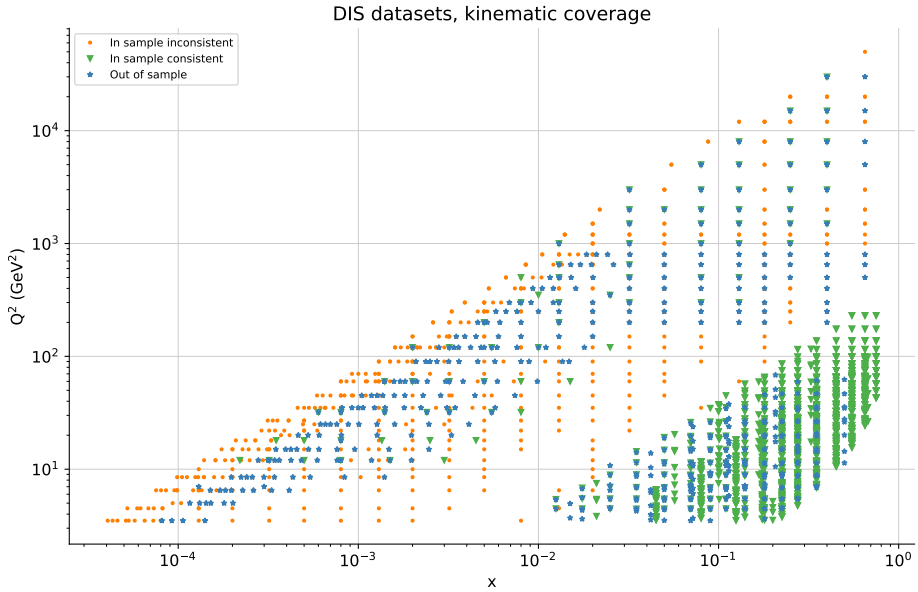


Figure 3.3: Kinematic coverage in (x, Q^2) of the data included in the closure tests on DIS-only fits. The orange dot marker indicate data that are included in the training set with an inconsistency built in according to the procedure described in section 3.2.1. The green inverted triangle indicates the data that are included in the training set that are consistent. Finally, the blue stars indicate the out-of-sample DIS data that are not included in the fit and that we use as test set.

is selected such that the kinematic coverage of the two samples is similar, facilitating the investigation of whether the fitted model generalizes effectively across comparable data samples.

Table table 3.1 provides a list of all DIS datasets included in this analysis. The table details the number of data points passing kinematic cuts and specifies whether each dataset is included in the fit (*in-sample* sets) or excluded from the fit (*out-of-sample* sets).

The last column indicates datasets into which experimental inconsistencies are injected according to the methodology described in section 3.2.1 and the specific configurations outlined in section 3.2.2. Various levels of inconsistency, parameterized by $\lambda \in [0, 1]$, are

Datasets	N_{data}	in/out sample	Inconsistency
NMC F_2^d/F_2^p [123]	121	in	
SLAC F_2^p [125]	33	in	
SLAC F_2^d [125]	34	in	
BCDMS F_2^p [126]	333	in	
BCDMS F_2^d [126]	248	in	
CHORUS σ_{CC}^{ν} [127]	416	in	
CHORUS $\sigma_{CC}^{\bar{\nu}}$ [127]	416	in	
NuTeV σ_{CC}^{ν} (dimuon) [128, 129]	39	in	
HERA I+II $\sigma_{\text{NC}}^{e^-p}$ $E_p = 920$ GeV [130]	159	in	✓
HERA I+II $\sigma_{\text{NC}}^{e^+p}$ $E_p = 575$ GeV [130]	254	in	✓
HERA I+II $\sigma_{\text{NC}}^{e^+p}$ $E_p = 820$ GeV [130]	70	in	✓
HERA I+II $\sigma_{\text{NC}}^{e^+p}$ $E_p = 920$ GeV [130]	377	in	✓
HERA I+II $\sigma_{\text{CC}}^{e^+p}$ [130]	39	in	
HERA I+II $\sigma_{\text{NC}}^{\text{charm}}$ [131]	37	in	
NMC $\sigma^{\text{NC},p}$ [124]	204	out	
NuTeV $\sigma_{CC}^{\bar{\nu}}$ (dimuon) [128, 129]	37	out	
HERA I+II $\sigma_{\text{NC}}^{e^+p}$ $E_p = 460$ GeV [130]	204	out	
HERA I+II $\sigma_{\text{CC}}^{e^-p}$ [130]	42	out	
HERA I+II $\sigma_{\text{NC}}^{\text{bottom}}$ [131]	26	out	

Table 3.1: List of the DIS dataset included in our analysis. For each dataset we indicate the number of datapoints included in the fit (after the standard kinematic cuts have been applied), whether the dataset belongs to the "in-sample" or "out-of-sample" set, and whether an experimental inconsistency is introduced.

introduced into the inclusive DIS HERA NC data, specifically in NC e^-p collisions at proton energy $E_p = 920$ GeV and in NC e^+p collisions at proton energies $E_p = 575$ GeV, $E_p = 820$ GeV, and $E_p = 920$ GeV.

The scan in λ covers:

- $\lambda = 1.0$: No inconsistency injected (baseline case).
- $\lambda = 0.7$: Mild inconsistency corresponding to a 30% underestimate of systematic multiplicative experimental uncertainties.
- $\lambda = 0.4$: Strong inconsistency corresponding to a 60% underestimate of systematic multiplicative experimental uncertainties.
- $\lambda = 0.0$: Maximal inconsistency where multiplicative uncertainties are completely disregarded by the experimentalists.

In total, $N_{\text{inc}} = 860$ out of the $N_{\text{tr}} = 2576$ datapoints included in the fit are affected by these experimental inconsistencies. The inconsistencies pertain to all systematic multiplicative uncertainties, which constitute the majority of the systematics in the DIS HERA data and are correlated among these four datasets where inconsistencies are injected.

In fig. 3.4, we present the ratio bias-variance computed according to eq. (3.17) for all DIS data included in our analysis, encompassing both the *in-sample* and *out-of-sample* datasets.

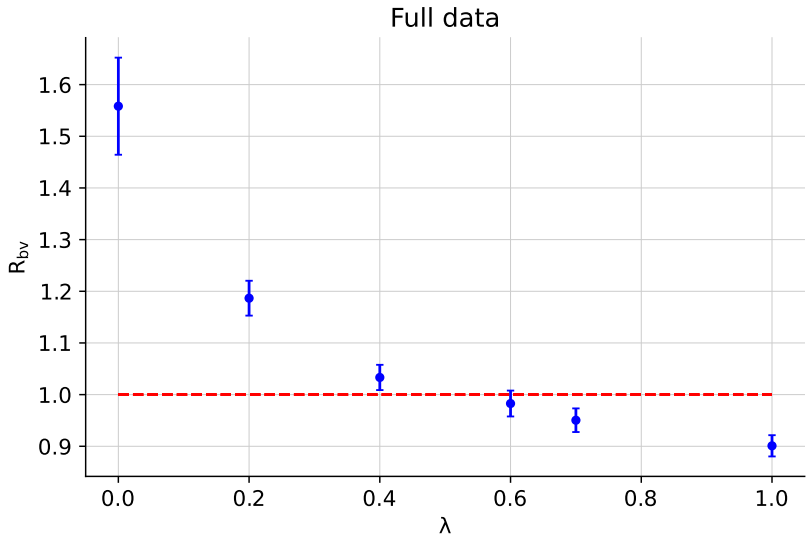


Figure 3.4: Ratio bias-variance, eq. (3.17), and its bootstrap uncertainty (see appendix D) as a function of λ computed on the entire DIS dataset.

The figure illustrates how the ratio bias-variance increases as we incrementally introduce experimental inconsistencies into the datasets marked by ticks in table 3.1, ranging from the fully consistent case ($\lambda = 1.0$) to the maximally inconsistent case ($\lambda = 0.0$).

Notably, as observed in [8], the PDF uncertainty in DIS-only fits tends to be slightly overestimated, as indicated by $R_{bv} \lesssim 1$. Interestingly, in intermediate scenarios ($\lambda = 0.7, 0.6, 0.4$), the Neural Network effectively assimilates the injected inconsistency, resulting in closure test outcomes that demonstrate PDF uncertainties are faithfully represented.

This trend is also evident in the $\xi_{1\sigma}$ quantile estimator, which we present in table 3.2 for various degrees of inconsistency. The table shows that even when λ is reduced to

λ	$\xi_{1\sigma}$
1.0	0.73 ± 0.01
0.7	0.71 ± 0.02
0.6	0.69 ± 0.01
0.4	0.68 ± 0.01
0.2	0.62 ± 0.01
0.0	0.52 ± 0.02

Table 3.2: Values of the $\xi_{1\sigma}$ quantile estimator in the observable space, eq. (3.18), and their bootstrap uncertainties.

0.4, the observables computed with the fitted PDFs include those calculated with the

underlying law 68% of the time. This indicates that the model effectively accommodates moderate inconsistencies without overestimating PDF uncertainties.

The normalized distribution of relative differences $\delta_i^{(l)}$ is depicted in fig. 3.5 for two scenarios: $\lambda = 1.0$ (left panel) and $\lambda = 0.0$ (right panel). Note that although the PCA algorithm starts from the same number of DIS data points in both cases, the number of degrees of freedom—specifically, the number of principal components retained when regularizing the covariance matrix—varies. This variation occurs because changes in the experimental covariance matrix lead to adjustments in the PDF-induced covariance matrix and hence its eigenvalues.

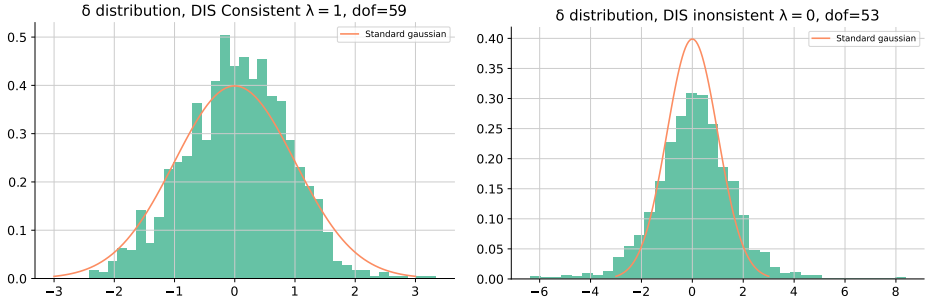


Figure 3.5: Normalized distribution of relative differences $\delta_i^{(l)}$ in the observable space, eq. (3.20), for $\lambda = 1.0$ (left panel) and $\lambda = 0.0$ (right panel). A univariate zero-mean Gaussian distribution is shown for reference in both cases.

From the plots in fig. 3.5, it is observed that the spread of the distribution increases moderately as the degree of inconsistency (measured by λ) increases, although it remains close to a normal distribution, as shown by the reference Gaussian plot.

After analyzing the estimators on the full DIS dataset included in the analysis, we now examine their values on each individual dataset.

Table 3.3 presents the ratio bias-variance for each dataset, categorized into *in-sample* and *out-of-sample* sets. Table 3.3 shows that across different DIS datasets, both *in-sample* and *out-of-sample*, the ratio bias-variance R_{bv} remains close to 1, indicating that the fitted PDF uncertainties are consistent with the underlying law without significant overestimation or underestimation. However, as the level of inconsistency increases, R_{bv} shows a slight increase for datasets directly affected by the inconsistency and for some *out-of-sample* datasets that probe similar kinematics, such as HERA I+II $\sigma_{\rm NC}^{e^+p}$ with $E_p = 460$ GeV. This effect remains mild, and R_{bv} values remain compatible with 1 within the uncertainties, even when systematic uncertainties are underestimated by 60%.

To visually illustrate the impact of experimental inconsistency on the ratio bias-variance (R_{bv}) across different datasets, we present fig. 3.6, which shows selected results from table 3.3. Each plot includes the bootstrap uncertainty alongside R_{bv} computed for various values of λ, representing different levels of injected inconsistency.

Overall observations from fig. 3.6 reveal that under a consistent closure test scenario ($\lambda = 1.0$), uncertainties tend to be slightly overestimated ($R_{bv} \lesssim 1$), consistent with previous findings [8]. Both *in-sample* and *out-of-sample* datasets exhibit similar behavior, suggesting effective generalization of the PDF model across different datasets with similar kinematic properties.

Datasets	N_{data}	R_{bv}			
		$\lambda = 1.0$	$\lambda = 0.7$	$\lambda = 0.4$	$\lambda = 0.0$
NMC F_2^d/F_2^p [123]	121	0.8	0.7	0.8	1.0
SLAC F_2^p [125]	33	0.7	0.7	0.8	1.1
SLAC F_2^d [125]	34	0.8	0.8	0.8	0.9
BCDMS F_2^p [126]	333	0.8	0.8	0.8	**1.1**
BCDMS F_2^d [126]	248	0.9	0.9	0.9	**1.1**
CHORUS σ^{ν}_{CC} [127]	416	0.8	0.9	0.8	0.9
CHORUS $\sigma^{\bar{\nu}}_{CC}$ [127]	416	0.9	1.0	1.0	**1.2**
NuTeV σ^{ν}_{CC} (dimuon) [128, 129]	39	0.8	0.9	0.9	**1.2**
HERA I+II $\sigma^{e^+p}_{\text{CC}}$ [130]	39	0.8	0.9	1.0	**1.2**
HERA I+II $\sigma^{\text{charm}}_{\text{NC}}$ [131]	37	1.0	**1.1**	**1.1**	**1.2**
(*) HERA I+II $\sigma^{e^-p}_{\text{NC}}$ $E_p = 920$ GeV [130]	159	0.9	1.0	**1.2**	**2.2**
(*) HERA I+II $\sigma^{e^+p}_{\text{NC}}$ $E_p = 575$ GeV [130]	254	0.8	0.9	**1.3**	**2.4**
(*) HERA I+II $\sigma^{e^+p}_{\text{NC}}$ $E_p = 820$ GeV [130]	70	0.8	0.9	**1.2**	**2.3**
(*) HERA I+II $\sigma^{e^+p}_{\text{NC}}$ $E_p = 920$ GeV [130]	377	0.8	0.9	**1.1**	**2.1**
NMC $\sigma^{\text{NC},p}$ [124]	204	0.9	0.9	**1.1**	**1.6**
NuTeV $\sigma^{\bar{\nu}}_{CC}$ (dimuon) [128, 129]	37	0.8	0.9	0.9	**1.1**
HERA I+II $\sigma^{e^+p}_{\text{NC}}$ $E_p = 460$ GeV [130]	204	0.9	1.0	**1.2**	**2.4**
HERA I+II $\sigma^{e^-p}_{\text{CC}}$ [130]	42	0.9	1.0	**1.1**	**1.4**
HERA I+II $\sigma^{\text{bottom}}_{\text{NC}}$ [131]	26	0.9	1.0	**1.2**	**1.9**

Table 3.3: Ratio bias variance, as defined in eq. (3.17), for all DIS dataset included in the DIS analysis (both the *in-sample* at the top of the table, with the ones unaffected by the inconsistency in the top section and the ones affected by the inconsistency in the middle section, and *out-of-sample* ones at the bottom of the table) as we increase the experimental inconsistency injected in the datasets in the middle section (marked by an asterisk) from $\lambda = 1.0$ (fully consistent datasets) to $\lambda = 0.0$ (maximally inconsistent datasets) and intermediate steps in between. We highlight in bold all instances in which $R_{bv} > 1$.

In contrast, when the level of inconsistency is increased ($\lambda < 1.0$), particularly noticeable in datasets directly affected by systematic underestimation (e.g., HERA I+II $\sigma^{e^+p}_{\text{NC}}$ with $E_p = 920$ GeV in the top left plot of fig. 3.6), R_{bv} increases above 1. This indicates a significant underestimation of uncertainties in the fit, highlighting the impact of experimental inconsistency. Furthermore, *out-of-sample* datasets that overlap kinematically with the inconsistent datasets, such as HERA I+II $\sigma^{e^+p}_{\text{NC}}$ with $E_p = 460$ GeV (bottom left plot in fig. 3.6), also exhibit increased R_{bv} values, albeit to a lesser extent compared to directly affected datasets.

Interestingly, even *in-sample* datasets like HERA I+II $\sigma^{e^+p}_{\text{CC}}$ (top right plot in fig. 3.6), which are not directly affected by inconsistency, can still show higher R_{bv} values under extreme inconsistency scenarios ($\lambda \to 0$), due to their correlation with the affected datasets.

These observations underscore the sensitivity of the PDF uncertainties to experimental inconsistencies, demonstrating the importance of robustness tests like the ones conducted here to evaluate the impact of such inconsistencies on global PDF fits.

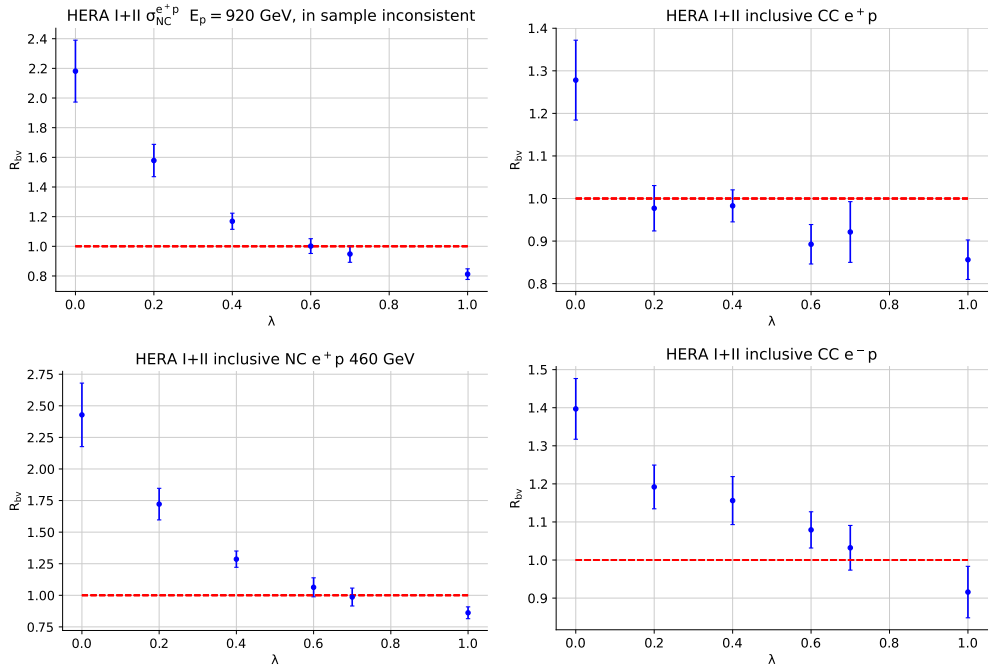


Figure 3.6: Ratio bias-variance R_{bv} computed as a function of λ, eq. (3.17), with bootstrap uncertainties, for different DIS datasets: (Top left) HERA I+II $\sigma_{\mathrm{NC}}^{e^+p}$ with $E_p = 920$ GeV, directly affected by systematic underestimation. (Top right) HERA I+II $\sigma_{\mathrm{CC}}^{e^+p}$, an *in-sample* dataset not directly affected by inconsistency. (Bottom left) HERA I+II $\sigma_{\mathrm{NC}}^{e^+p}$ with $E_p = 460$ GeV, an *out-of-sample* dataset overlapping kinematically with inconsistent datasets. (Bottom right) HERA I+II $\sigma_{\mathrm{CC}}^{e^-p}$, an *out-of-sample* dataset.

To analyze the impact of experimental inconsistency on individual data points within the HERA I+II dataset measuring $\sigma_{\mathrm{NC}}^{e^+p}$ with $E_p = 920$ GeV (directly affected) and $\sigma_{\mathrm{NC}}^{e^+p}$ with $E_p = 460$ GeV (indirectly affected), we present fig. 3.7. This figure displays R_{bv} computed for each data point under the maximally inconsistent scenario ($\lambda = 0.0$).

Observations of fig. 3.7 indicate that the largest contributions to R_{bv} are from data points with smaller statistical uncertainties, primarily constraining the x region between 10^{-3} and 10^{-2}. These data points exhibit R_{bv} values that tend to deviate more significantly from 1 under the influence of experimental inconsistency. Conversely, data points at larger x and low Q^2, where uncertainties are larger, show R_{bv} values that remain closer to 1 even in the presence of inconsistency.

Further insight into the correlation patterns between the inconsistent datasets and PDFs is provided in appendix E. Specifically, it is shown that the gluon distribution in the small-intermediate x region ($x \sim 10^{-3} - 10^{-2}$) and the up quark distribution in the medium x region ($x \sim 10^{-2} - 10^{-1}$) are most affected by the inconsistency. This is consistent with the findings in fig. 3.8, where the gluon distribution shows a more pronounced shift compared to the underlying law under maximally inconsistent conditions, while the up quark distribution is less affected due to constraints from other datasets in the medium x region.

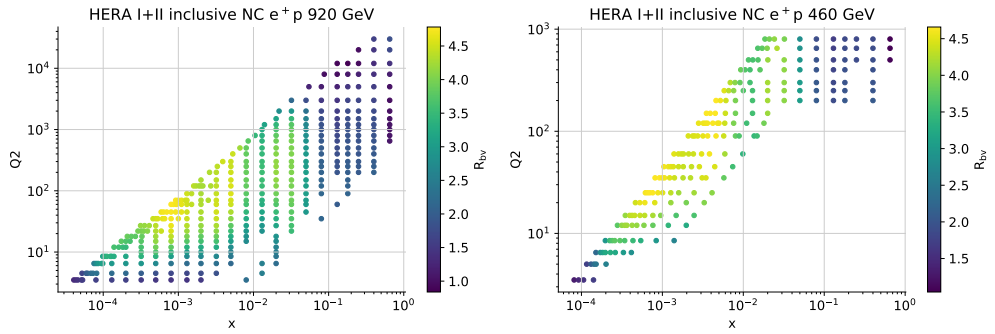


Figure 3.7: R_{bv} per data point in the maximally inconsistent scenario ($\lambda = 0.0$) computed on the datasets: (Left panel) HERA I+II $\sigma_{\mathrm{NC}}^{e^+p}$ with $E_p = 920$ GeV, directly affected by experimental inconsistency. (Right panel) HERA I+II $\sigma_{\mathrm{NC}}^{e^+p}$ with $E_p = 460$ GeV, an *out-of-sample* dataset indirectly affected.

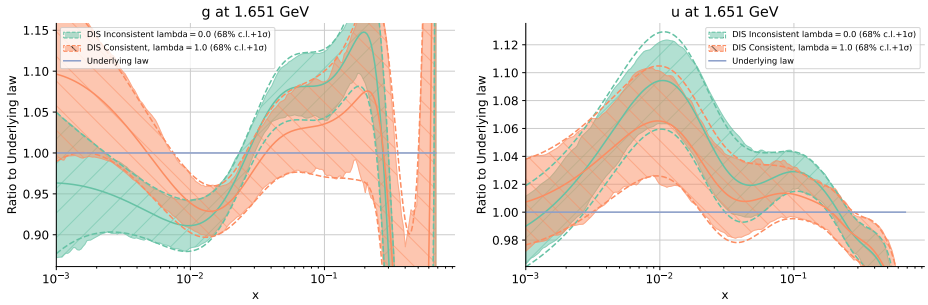


Figure 3.8: Comparison between underlying law (blue line) and the result of a L_2 consistent DIS-only closure test $\lambda = 1$ (orange band) and maximally inconsistent DIS-only closure test $\lambda = 0$ (green band). The gluon distribution (left panel) and the up quark distribution (right panel) are plotted at the initial scale $Q = 1.65$ GeV and normalised to the central values of the underlying law.

Overall, these results highlight the localized impact of experimental inconsistency on specific PDFs within the global fit, particularly affecting regions where data constraints are strongest and inconsistencies are pronounced.

Drell-Yan

In this section, we extend our analysis to include the entire NNPDF4.0 dataset [8], incorporating both DIS data and hadronic observables. Also in this case, we partition the dataset into *in-sample* and *out-of-sample* subsets, as depicted in fig. 3.9. This partition is carefully designed to ensure that both subsets have similar kinematic coverage, facilitating an investigation into the generalization performance of the fitted model across comparable data samples. This approach allows us to assess whether the fitted PDF model can effectively generalize to new, similar data that were not part of the training set. Such a study is crucial for evaluating the robustness of the PDF fit and its predictive power beyond the specific dataset used for training.

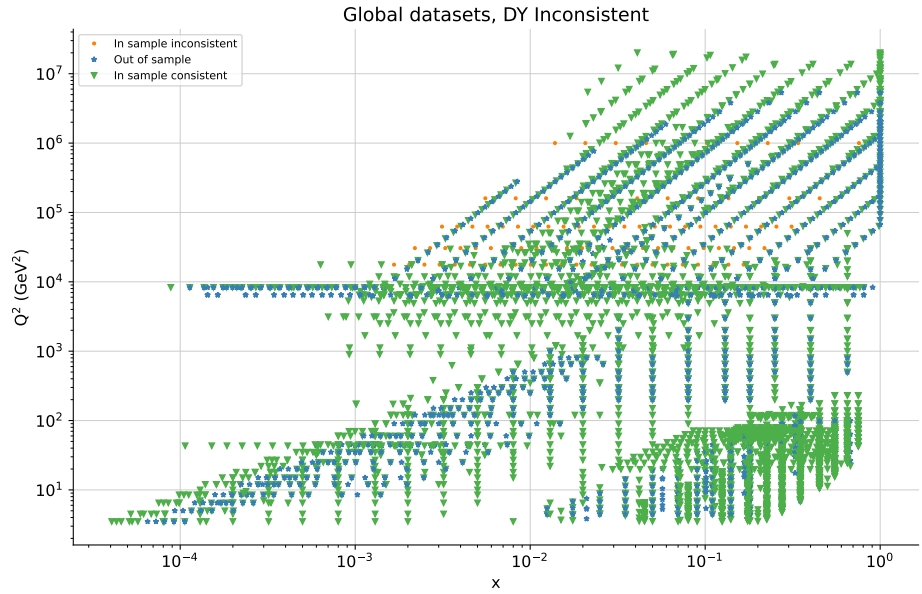


Figure 3.9: Kinematic coverage of the *in-sample* and *out-of-sample* subsets used in the closure tests on the PDF global fit with an experimental inconsistency injected into the ATLAS 8 TeV high-mass Drell-Yan dataset [133]

The rationale behind the dataset split draws inspiration primarily from the established practice of *k-folds* separation [134], ensuring a balanced partitioning that considers the full kinematic coverage of the dataset. To tailor the split for our study, slight adjustments were made to the conventional *k-fold* approach. These modifications were implemented specifically to enhance the similarity between *out-of-sample* datapoints and the inconsistent data subset in terms of their kinematic characteristics. All details regarding the split are listed in tables 3.4 and 3.5.

The first case study involves the inconsistent Drell-Yan scenario, where an inconsistency is introduced into the double differential high-mass Drell-Yan cross section measured by ATLAS at $\sqrt{s} = 8$ TeV [133]. Similar to the DIS case, various degrees of inconsistency are parameterized by the factor $\lambda \in [0, 1]$. The dataset affected by inconsistency comprises only $N_{\text{inc}} = 48$ datapoints, whereas the total number of training datapoints is $N_{\text{tr}} = 3772$.

It is crucial to note a distinction in this scenario: certain multiplicative uncertainties, inherently underestimated, impact *all* observables measured by the ATLAS detector, particularly the luminosity uncertainty. Consequently, more datapoints beyond just the ATLAS double differential Drell-Yan measurement are affected, resulting in a total of $N_{\text{tot, inc}} = 607$ inconsistent datapoints.

Similar to the DIS case, we investigate extreme and intermediate levels of inconsistency: $\lambda = 0.0$ corresponds to complete disregard of correlated multiplicative uncertainties at the experimental level for the ATLAS 8 TeV high-mass Drell-Yan dataset, along with intermediate cases $\lambda = 0.4, 0.8$.

The trend of R_{bv} as a function of λ, obtained through the PCA procedure applied to the entire NNPDF 4.0 dataset, is depicted in fig. 3.10.

Datasets	N_{data}	in/out sample	Inconsistency
DY E866 σ^p_{DY}	29	in	
DY E605 σ^p_{DY}	85	in	
DY E906 $\sigma^d_{\text{DY}}/\sigma^p_{\text{DY}}$ (SeaQuest)	6	in	
D0 Z rapidity	28	in	
D0 $W \to \mu\nu$ asymmetry ($\mathcal{L} = 7.3$ fb^{-1})	9	in	
ATLAS W, Z 7 TeV ($\mathcal{L} = 35$ pb^{-1})	30	in	
ATLAS low mass DY 7 TeV	6	in	
ATLAS W, Z 7 TeV ($\mathcal{L} = 4.6$ pb^{-1}) CF	15	in	
ATLAS low-mass DY 2D 8 TeV	60	in	
ATLAS $\sigma_{W,Z}$ 13 TeV	3	in	
ATLAS W^-+jet 8 TeV	15	in	
ATLAS Z p_T 8 TeV (p_T, m_{ll})	44	in	
ATLAS Z p_T 8 TeV (p_T, y_Z)	48	in	
CMS W electron asymmetry 7 TeV	11	in	
CMS DY 2D 7 TeV	110	in	
CMS W rapidity 8 TeV	22	in	
LHCb $Z \to ee$ 7 TeV	9	in	
LHCb $Z \to ee$ 8 TeV ($\mathcal{L} = 2$ fb^{-1})	17	in	
LHCb $W, Z \to \mu$ 8 TeV	30	in	
ATLAS high-mass DY 2D 8 TeV	48	in	$\checkmark^{\text{DY}}$
ATLAS W, Z 7 TeV ($\mathcal{L} = 4.6$ pb^{-1}) CC	46	out$^{\text{JETS}}$ in$^{\text{DY}}$	
DY E866 $\sigma^d_{\text{DY}}/2\sigma^p_{\text{DY}}$ (NuSea) [131]	15	out$^{\text{JETS}}$ in$^{\text{DY}}$	
CDF Z rapidity	28	out$^{\text{JETS}}$ in$^{\text{DY}}$	
CMS Z p_T 8 TeV	28	in$^{\text{JETS}}$ out$^{\text{DY}}$	
LHCb $Z \to ee$ 13 TeV	16	out	

Table 3.4: Same as Table 3.1 for the analyses presented in Sect. 3.2.3. We specify the settings that differ in the two analyses with a superscript. The DIS datasets are omitted and details are given in the text.

Datasets	N_{data}	in/out sample	Inconsistency
ATLAS dijets 7 TeV, R=0.6	90	in	
ATLAS direct photon production 13 TeV	53	in	
ATLAS single t R_t 7 TeV	1	in	
ATLAS single t R_t 13 TeV	1	in	
ATLAS single t 7 TeV $(1/\sigma d\sigma/dy_t)$	3	in	
ATLAS single t 7 TeV $(1/\sigma d\sigma/dy_{\bar{t}})$	3	in	
ATLAS single t 8 TeV $(1/\sigma d\sigma/dy_{\bar{t}})$	3	in	
ATLAS σ_{tt}^{tot} 13 TeV ($\mathcal{L} = 139$ fb^{-1})	1	in	
ATLAS $t\bar{t}$ $l+$jets 8 TeV $(1/\sigma d\sigma/dy_t)$	4	in	
ATLAS $t\bar{t}$ $l+$jets 8 TeV $(1/\sigma d\sigma/dy_{t\bar{t}})$	4	in	
ATLAS $t\bar{t}$ $2l$ 8 TeV $(1/\sigma d\sigma/dy_{t\bar{t}})$	4	in	
CMS dijets 7 TeV	54	in	
CMS σ_{tt}^{tot} 7, 8, 13 TeV	3	in	
CMS $t\bar{t}$ $l+$jets 8 TeV $(1/\sigma d\sigma/dy_{t\bar{t}})$	9	in	
CMS $t\bar{t}$ $2l$ 13 TeV $(d\sigma/dy_t)$	10	in	
CMS $t\bar{t}$ $l+$ jets 13 TeV $(d\sigma/dy_t)$	11	in	
CMS single t 7 TeV $(\sigma_t + \sigma_{\bar{t}})$	1	in	
CMS single t 8 TeV R_t	1	in	
CMS single t 13 TeV R_t	1	in	
ATLAS single-inclusive jets 8 TeV, R=0.6	171	in	$\checkmark^{\text{JETS}}$
LHCb $W, Z \to \mu$ 7 TeV	29	out	
LHCb $Z \to \mu\mu$ 13 TeV	15	out	
ATLAS W^++jet 8 TeV	15	out	
CMS W muon asymmetry 7 TeV	11	out	
ATLAS σ_{tt}^{tot} 7 TeV	1	out	
ATLAS σ_{tt}^{tot} 8 TeV	1	out	
ATLAS single t 8 TeV $(1/\sigma d\sigma/dy_t)$	3	out	
CMS σ_{tt}^{tot} 5 TeV	1	out	
CMS $t\bar{t}$ 2D $2l$ 8 TeV $(1/\sigma d\sigma/dy_t dm_{t\bar{t}})$	15	out	
CMS single-inclusive jets 8 TeV	185	out*	

Table 3.5: Same as Table 3.4 for the top and jets datasets. The * category refers to the exercise described in Sect. **??** in which the CMS single-inclusive jets at $\sqrt{s} = 8$ TeV are kept in the *in-sample* set.

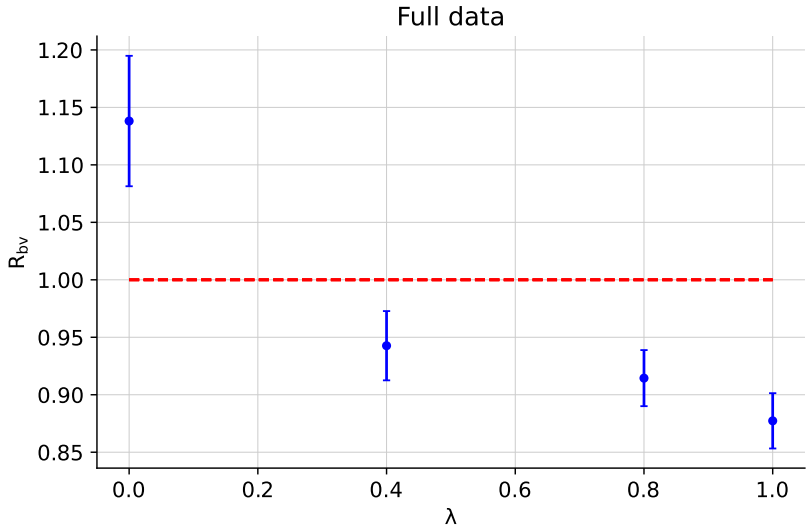


Figure 3.10: R_{bv} as a function of λ for the inconsistent multi-closure test in the Drell-Yan sector. The estimator is computed over the entire NNPDF 4.0 dataset.

It is observed that the impact of inconsistency in the ATLAS 8 TeV high-mass Drell-Yan dataset is minimal. Similar to the DIS-only fit scenario, standard consistent closure tests show a slight overestimation of PDF uncertainties, with $R_{bv} \lesssim 1$. Even in the presence of mild to moderate inconsistencies, the ratio remains below one, indicating an overall underestimation of PDF uncertainties only in the extreme $\lambda = 0$ case.

The mild impact of inconsistency in the Drell-Yan sector is further corroborated by examining the values of the quantile estimator in the observable space $\xi_{1\sigma}$ computed across the entire dataset, including both the *in-sample* and *out-of-sample* subsets. As shown in table 3.6, we observe that $\xi_{1\sigma}$ decreases as the level of inconsistency, parameterized by λ, increases, which aligns with our expectations. However, it is noteworthy that while we expect $\xi_{1\sigma} \sim 0.68$ for accurately estimated uncertainties, the computed values remain above this threshold until $\lambda = 0.4$, dipping slightly below it only in the maximally inconsistent case.

λ	$\xi_{1\sigma}$
1.00	0.75 ± 0.02
0.80	0.73 ± 0.02
0.40	0.71 ± 0.02
0.00	0.66 ± 0.03

Table 3.6: Quantile estimator $\xi_{1\sigma}$ computed in the case of inconsistency in the Drell-Yan sector.

Additionally, in fig. 3.11, we depict the normalized distribution of relative differences $\delta_i^{(l)}$ for the extreme cases $\lambda = 1.0$ and $\lambda = 0.0$. Similar to the DIS case, the histogram represents the global dataset without distinguishing between *in-sample* and *out-of-sample* subsets. It is evident that the histogram broadens slightly for $\lambda = 0.0$, indicating a marginally increased variance in the relative differences. However, the overall shape

of the histogram remains consistent with a normal distribution, as indicated by the reference curve plotted for comparison.

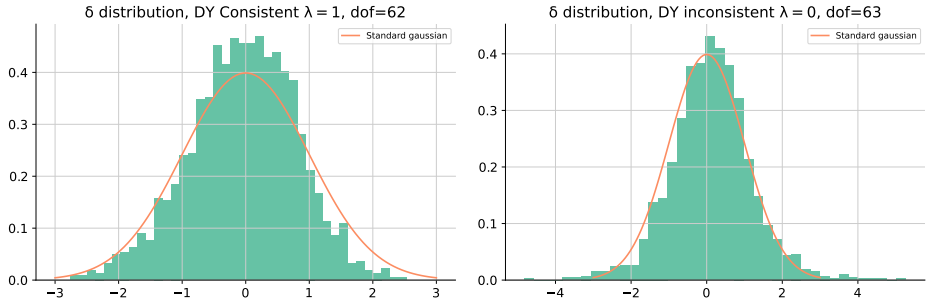


Figure 3.11: Normalized distribution of relative differences $\delta_i^{(l)}$ in the case of inconsistency in the Drell-Yan sector. Left: $\lambda = 1.0$. Right: $\lambda = 0.0$.

After examining the estimators on the global datasets, we proceed to analyze their values on each individual dataset. Table 3.7 presents the R_{bv} values for the most significant datasets, specifically the *out-of-sample* datasets and the *in-sample* inconsistent one.

Datasets	N_{data}	R_{bv}			
		$\lambda = 1.0$	$\lambda = 0.8$	$\lambda = 0.4$	$\lambda = 0.0$
(*) ATLAS DY 2D 8 TeV low mass [135]	48	0.8	0.8	**1.1**	**3.2**
HERA I+II $\sigma_{\text{CC}}^{e^+p}$ [130]	39	0.9	1.0	1.0	**1.1**
HERA I+II $\sigma_{\text{NC}}^{e^\pm p}$ $E_p = 575$ GeV [130]	254	0.7	0.7	0.8	0.8
NMC F_2^d/F_2^p [123]	121	0.9	0.9	0.9	0.8
NuTeV σ_{CC}^{ν} (dimuon) [128, 129]	39	1.0	**1.1**	**1.1**	**1.1**
LHCb $W, Z \to \mu$ 7 TeV [136]	29	0.9	0.9	1.0	**1.4**
LHCb $Z \to \mu\mu$ 13 TeV [137]	15	0.9	0.9	1.0	**1.1**
ATLAS W^++jet 8 TeV [138]	15	0.7	0.7	1.0	**1.5**
CMS W muon asymmetry 7 TeV [139]	11	0.7	0.7	0.7	0.8
ATLAS σ_{tt}^{tot} 8 TeV [140]	1	0.92	0.8	0.9	0.9
ATLAS high mass DY 7 TeV [141]	5	1.0	1.0	**1.5**	**3.4**
ATLAS single t 8 TeV $(1/\sigma d\sigma/dy_t)$ [142]	3	0.9	0.9	0.8	0.9
CMS σ_{tt}^{tot} 5 TeV [143]	1	0.8	0.8	0.9	0.7
CMS $t\bar{t}$ 2D $2l$ 8 TeV $(1/\sigma d\sigma/dy_t dm_{t\bar{t}})$ [144]	15	0.7	0.7	0.8	0.8
CMS single-inclusive jets 8 TeV [145]	185	0.7	0.7	0.7	0.9

Table 3.7: Ratio bias-variance R_{bv} as a function of λ, for all the *out-of-sample* datasets and for the *in-sample* inconsistent one, marked with an asterisk. The *in-sample* consistent datasets are not shown for presentation reasons but are discussed in the text.

The results are also illustrated in fig. 3.12 for the two datasets that exhibit the most significant effects. As depicted in the plot, uncertainties are underestimated only in the most extreme case where all correlated systematics are completely ignored ($\lambda = 0$). Even for $\lambda = 0.4$, the ratio bias-variance remains consistent with 1 within the 1σ uncertainty.

Aside from the inconsistent 8 TeV dataset itself, which exhibits $R_{bv} \gtrsim 3$ for $\lambda = 0.0$, the most affected out-of-sample dataset is the high-mass Drell-Yan dataset measured by

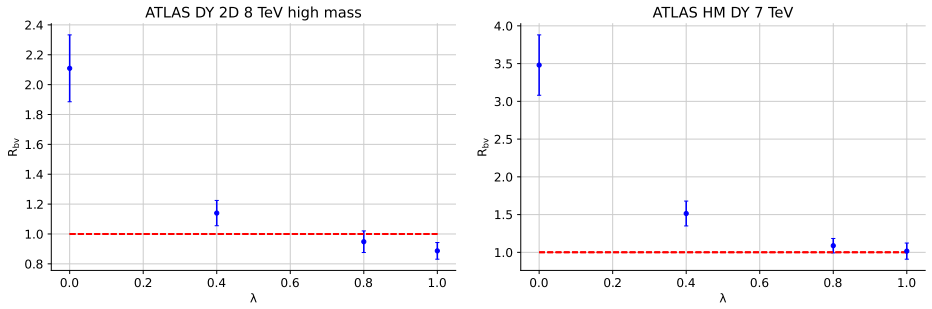


Figure 3.12: Ratio bias-variance R_{bv} computed with PCA with explained variance ratio $= 0.99$, for the double differential high-mass ATLAS DY distribution at $\sqrt{s} = 8$ TeV, which is the dataset directly made inconsistent during training (left panel), and for the ATLAS DY high-mass $\sqrt{s} = 7$ TeV, which is the most affected dataset in the out-of-sample subset (right panel).

ATLAS at $\sqrt{s} = 7$ TeV, for which R_{bv} also exceeds 3 for $\lambda = 0.0$. Other datasets remain largely unaffected. This observation is understandable, as the ATLAS high-mass Drell-Yan measurement at $\sqrt{s} = 7$ TeV is the only out-of-sample dataset that explores the same large-x kinematic region for quark and antiquarks, which is not solely dominated by statistical uncertainties. Similarly, the on-shell LHCb distributions at large rapidities are mildly affected, as they overlap with the large-x kinematic region but not entirely, and they also feature larger statistical uncertainties.

Looking at R_{bv} computed on each single data point in fig. 3.13 for both the ATLAS 8 TeV *in-sample* inconsistent high-mass Drell-Yan dataset and for the ATLAS 7 TEV *out-of-sample* dataset we observe that, while all points of the ATLAS 7 TeV dataset contribute to the large value of R_{bv} featured by the dataset and these points probe the region in $x \approx (10^{-2}, 10^{-1})$, the points in ATLAS 8 TeV that contribute the most to R_{bv} are the two lowest bins in the invariant mass, and that the kinematic region in x that is maximally contributing is still the one around $x \approx (10^{-2}, 10^{-1})$.

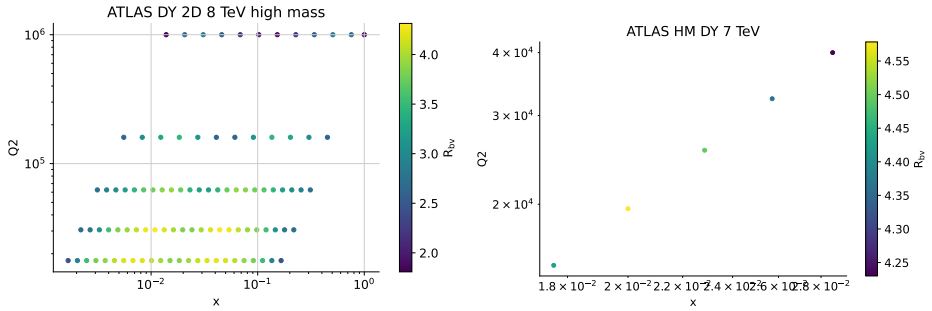


Figure 3.13: R_{bv} per data point in the maximally inconsistent scenario ($\lambda = 0.0$) computed on the double differential high mass ATLAS DY distribution at $\sqrt{s} = 8$ TeV, which is the dataset directly made inconsistent during the training (left panel) and on the ATLAS DY high-mass distributiom at $\sqrt{s} = 7$ TeV, which is *out of sample* (right panel).

Considering the correlations between the ATLAS high-mass data in the invariant mass bins that most contribute to a large R_{bv} value and the individual PDFs, which we

display in fig. E.2, we see that the correlation is maximal with the gluon around $x \sim 10^{-3}$ and with light quark and antiquark around $x \sim 10^{-2}$. In fig. 3.14 we plot these PDF combinations and compare the agreement with the underlying law and the uncertainties of the PDFs obtained from a consistent closure test on the global dataset ($\lambda = 1$) and those obtained from a maximally inconsistent closure test ($\lambda = 0$) in the Drell-Yan sector. We see that there is almost no change in the PDFs agreement with the underlying law. The only visible effect is a mild reduction in the uncertainty bands, which is to be expected in an inconsistent closure test, and also confirmed at the observables level given the increase in the ratio bias variance.

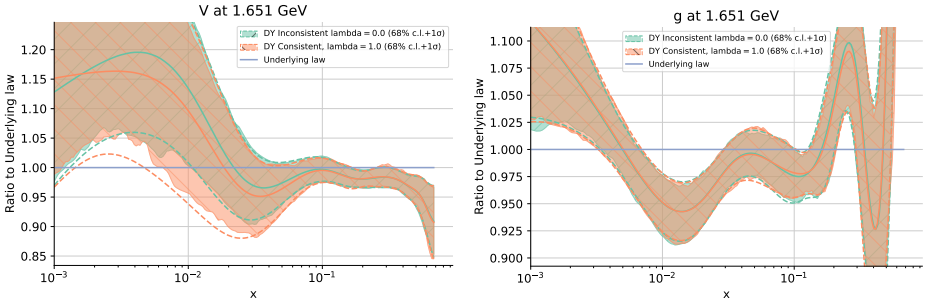


Figure 3.14: Comparison between underlying law (blue line) and the result of a L_2 consistent global closure test $\lambda = 1$ (orange band) and closure test in which a maximal inconsistency is injected in the Drell-Yan sector, $\lambda = 0$ (green band). The gluon distribution (left panel) and the total valence distribution (right panel) are plotted at the initial scale $Q = 1.65$ GeV and normalised to the central values of the underlying law.

The most plausible explanation is that the high-mass Drell-Yan data are not the dominant constraint of the PDFs in the intermediate-to-large x region and that the raw number of inconsistent points is too small to have a visible effect on the PDF themselves. This confirms what has been observed throughout this section, namely that an inconsistency of experimental origin in the high-mass Drell-Yan data would not distort the results of a fit, nor would undermine in any dramatic way the faithfulness of PDF uncertainties.

Inclusive Jets

In this section, we present the results of closure tests where an experimental inconsistency has been introduced into one of the inclusive jet measurements included in the NNPDF4.0 global analysis. The inconsistency specifically affects the measurement of the inclusive jet cross-sections in pp collisions at $\sqrt{s} = 8$ TeV performed by ATLAS [146]. Similar to the Drell-Yan case, this inconsistency acts on all correlated systematic uncertainties, thereby impacting all other ATLAS datasets.

The split of the dataset into in-sample and out-of-sample subsets is displayed in fig. 3.15. This split is designed to ensure similar kinematic coverage between the two subsets, facilitating the study of how well the fitted PDF model generalizes to similar data samples. As in the Drell-Yan case, we consider the whole NNPDF 4.0 dataset. There are minor differences with respect to the Drell-Yan case, listed in detail in Tables table 3.4 and table 3.5.

Note that the single-inclusive jets measurement at $\sqrt{s} = 8$ TeV conducted by CMS [147] is included in the global NNPDF4.0 dataset. Therefore, we have two datasets, the

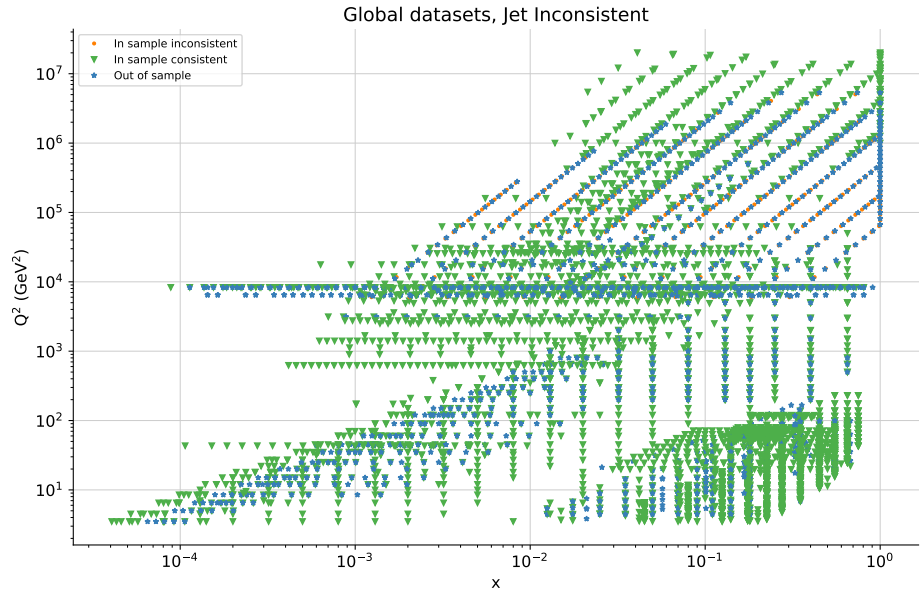


Figure 3.15: Same as fig. 3.3 for the kinematic coverage of the samples included in the closure tests on global PDF fit with inconsistent inclusive jets data.

ATLAS and CMS inclusive jet measurements, which observe the same physical observable but were obtained independently by two different experimental collaborations. This setup allows us to investigate the behavior of neural networks when trained with a single inconsistent instance of data compared to encountering two sets of almost equivalent data, where one set is consistent and the other is inconsistent.

We will begin by discussing the results of the closure test in the scenario where the CMS dataset is considered out of sample, meaning the inconsistent ATLAS inclusive jet dataset predominantly constrains the large-x gluon. Conversely, we will observe that maintaining the CMS consistent dataset in sample notably mitigates the impact of the inconsistency.

Starting with the scenario where the CMS inclusive jets 8 TeV data are *out-of-sample*, the neural network (NN) model is trained using a total of $N_{\text{tr}} = 3793$ data points. The ATLAS inconsistent dataset consists of $N_{\text{inc}} = 171$ data points, but considering all ATLAS inconsistent data brings the total to $N_{\text{tot, inc}} = 607$ data points affected by inconsistency.

The trend of R_{bv} as a function of λ, obtained when applying the PCA procedure to the entire NNPDF4.0 dataset, is illustrated in fig. 3.16. It is evident from the figure that the impact of inconsistency within the ATLAS 8 TeV inclusive jet dataset is most pronounced at $\lambda = 0$. In this case, the R_{bv} value deviates significantly, approaching nearly 6σ from 1.0. However, a discernible trend in the R_{bv} value for increasing λ values suggests that at the default setting ($\lambda = 1.0$) and even in scenarios with mild inconsistencies, PDF uncertainties tend to be slightly overestimated. This trend is supported by the $\xi_{1\sigma}$ quantile estimator shown in table 3.8 for various degrees of inconsistency, where the effect of inconsistency is primarily noticeable in the $\lambda = 0$ bin, but marginal for other values.

In fig. 3.17, we present the normalized distribution of relative differences $\delta_i^{(l)}$ for the

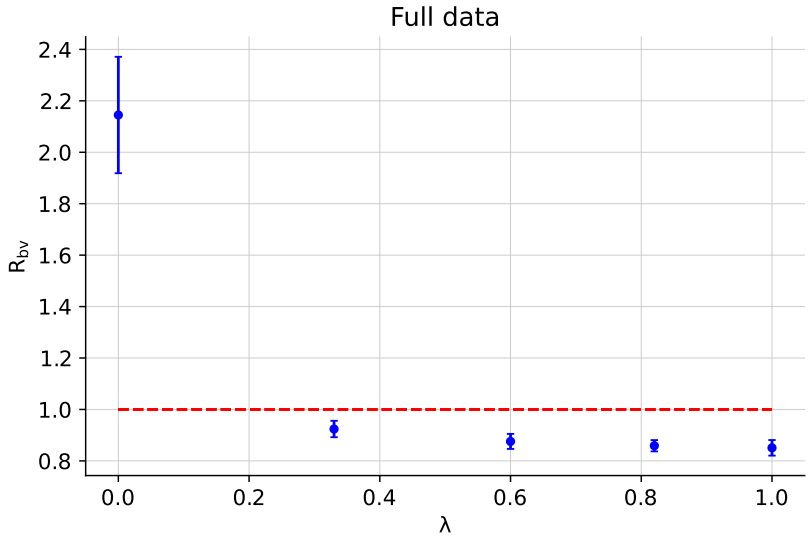


Figure 3.16: R_{bv} as a function of λ for the JETS inconsistent multi-closure test. The estimator is computed over the whole NNPDF 4.0 dataset.

λ	$\xi_{1-\sigma}$
1.00	0.75 ± 0.01
0.82	0.75 ± 0.01
0.60	0.74 ± 0.01
0.33	0.73 ± 0.02
0.00	0.49 ± 0.03

Table 3.8: Values of the quantile estimator $\xi_{1\sigma}$ (eq. (3.18)) with bootstrap uncertainty, for different λ values.

extreme cases of $\lambda = 0$ and $\lambda = 1$, similar to the Drell-Yan case discussed before. The $\lambda = 1$ case aligns with the Drell-Yan scenario, as expected given their consistency and nearly identical setups except for minor variations in training data (see table 3.4). As observed in the Drell-Yan case, significant deformation of the histogram shape is not evident even in the most inconsistent scenario $\lambda = 0$.

We now focus on presenting our estimators within more local scenarios, narrowing our analysis to individual datasets. The trend of R_{bv} as a function of λ is shown in table 3.9 for a subset of datasets. We focus our attention on the inconsistent dataset and the *out-of-sample* datasets, as the inconsistency in the ATLAS inclusive jet data at 8 TeV does not significantly impact the *in-sample* datasets.

As anticipated, the CMS inclusive jets cross section measured at $\sqrt{s} = 8$ TeV is notably influenced by the inconsistency introduced in the corresponding ATLAS dataset. However, this effect is only observable in the most extreme case of $\lambda = 0$, which aligns with observations in the DIS and DY cases. Specifically, the Neural Network demonstrates capability in handling moderate inconsistencies of experimental origin.

Furthermore, it is noteworthy that datasets measuring the cross section values of $t\bar{t}$ production also show significant susceptibility to inconsistency. This is evidenced by the

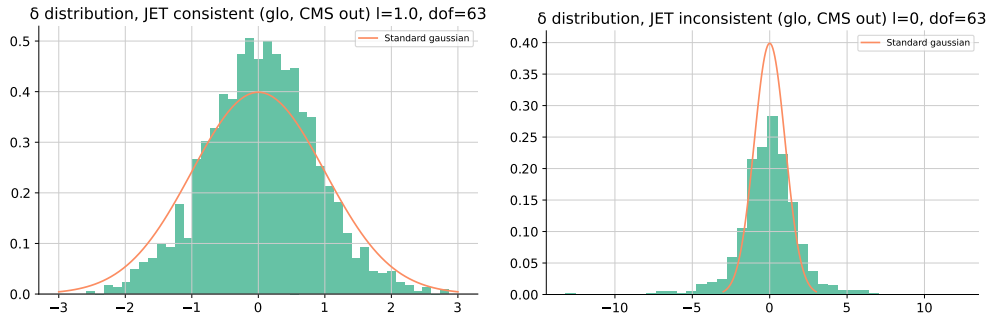


Figure 3.17: δ plots for all data for $\lambda = 1.0$ (right) and $\lambda = 0$ (left) cases.

bias variance ratio deviating markedly from unity in the $\lambda = 0$ case.

In fig. 3.18, we illustrate the trend of R_{bv} for the two most affected *out-of-sample* datasets discussed above: the CMS double differential $t\bar{t}$ distribution measured in the lepton channel at $\sqrt{s} = 8$ TeV, and the CMS inclusive jets cross section measured at $\sqrt{s} = 8$ TeV. We observe that the bias-variance ratio increases above 1 as λ decreases below 0.3; otherwise, the effect of the inconsistency is almost negligible.

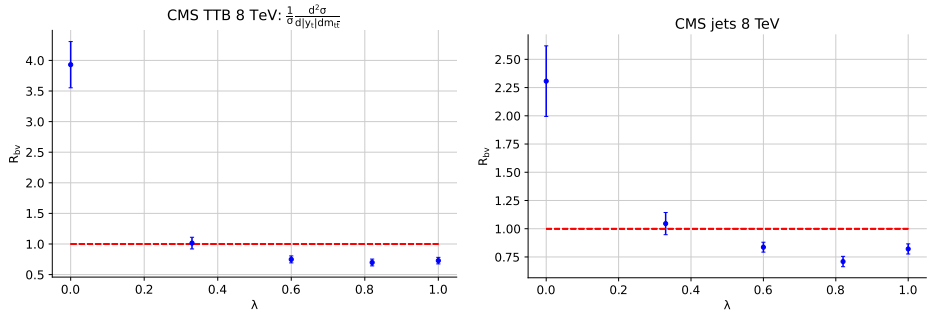


Figure 3.18: Bias-variance ratio R_{bv} computed with PCA with an explained variance ratio of 0.99, for the CMS double differential $t\bar{t}$ distribution measured in the lepton channel at $\sqrt{s} = 8$ TeV (left panel) and for the CMS inclusive jets cross section measured at $\sqrt{s} = 8$ TeV (right panel).

In fig. 3.19, we present the single data point analysis obtained with $\lambda = 0$, for the ATLAS inconsistent dataset and the double differential $t\bar{t}$ distribution measured in the lepton channel at $\sqrt{s} = 8$ TeV by CMS.

In both cases, the region in x that contributes most strongly to R_{bv} is between $x \approx 0.03$ and $x \approx 0.4$. For the CMS inclusive jets, the bins at larger p_T contribute the most.

Now we move to the scenario where the CMS jets dataset is *in-sample*, briefly presenting the main differences compared to the *out-of-sample* case.

In fig. 3.20, we compare the single data point values of R_{bv} for the CMS jets dataset in both cases. We select $\lambda = 0.33$ to illustrate the impact of inconsistency within a moderately severe scenario. Notably, on average, the R_{bv} values are larger in the *out-of-sample* case (indicated by the prominent yellow area in the center of the left plot), as expected. Interestingly, in the *in-sample* case, several data points exhibit the highest R_{bv} values, predominantly located at the edges of the kinematic plot.

Datasets	N_{data}	R_{bv}			
		$\lambda = 1.0$	$\lambda = 0.6$	$\lambda = 0.33$	$\lambda = 0.00$
(*) ATLAS jets 8 TeV, $R = 0.6$	171	0.8	0.8	1.0	**2.3**
HERA I+II $\sigma^{e^+p}_{\text{CC}}$ [130]	39	1.0	0.8	1.0	**1.2**
HERA I+II $\sigma^{e^\pm p}_{\text{NC}}$ $E_p = 575$ GeV [130]	254	0.8	0.8	0.7	**1.6**
NMC F_2^d/F_2^p [123]	121	0.7	0.7	0.8	0.9
NuTeV σ^{ν}_{CC} (dimuon) [128, 129]	39	0.9	1.0	1.0	**1.1**
LHCb $W, Z \to \mu$ 7 TeV [136]	29	0.8	0.9	0.9	**1.1**
LHCb $Z \to \mu\mu$ 13 TeV [137]	15	0.8	0.8	0.8	**1.4**
ATLAS W, Z 7 TeV ($\mathcal{L} = 4.6$ pb^{-1}) CC [148]	46	0.7	0.6	0.6	0.9
ATLAS W^++jet 8 TeV [138]	15	0.8	0.8	1.0	**3.2**
ATLAS high mass DY 7 TeV [141]	5	1.0	0.9	1.0	**1.2**
CMS W muon asymmetry 7 TeV [139]	11	0.7	0.7	0.7	0.7
DY E866 $\sigma^d_{\text{DY}}/2\sigma^p_{\text{DY}}$ (NuSea) [131]	15	0.8	0.8	0.8	0.9
CDF Z rapidity [149]	15	0.7	0.8	0.7	0.9
ATLAS σ^{tot}_{tt} 7 TeV [148]	1	0.7	0.8	1.0	**5.9**
ATLAS σ^{tot}_{tt} 8 TeV [138]	1	0.7	0.7	1.0	**5.3**
ATLAS single t 8 TeV ($1/\sigma d\sigma/dy_t$) [142]	3	1.0	1.0	1.0	**1.5**
CMS σ^{tot}_{tt} 5 TeV [143]	1	0.7	0.9	**1.2**	**5.4**
CMS $t\bar{t}$ 2D 2l 8 TeV ($1/\sigma d\sigma/dy_t dm_{t\bar{t}}$) [144]	15	0.7	0.7	1.0	**3.8**
CMS single-inclusive jets 8 TeV [145]	185	0.8	0.8	1.0	**2.2**

Table 3.9: Ratio bias variance, as defined in eq. (3.17) as a function of λ, for all the *out-of-sample* datasets and for the *in-sample* inconsistent one, marked with an asterisk. The estimator evaluated on the *in-sample* consistent datasets is not shown for presentation reasons but is discussed in the text.

Next, we examine the PDFs themselves, comparing the cases where the CMS jets dataset is *in-sample* and *out-of-sample*. Given that the inconsistency was injected into a dataset that measures jets, which strongly constrain the gluon (see appendix E), this comparison is crucial. In fig. 3.21, we show the ratio to the underlying law of the gluon PDF as obtained in the consistent case and in the $\lambda = 0$ case, comparing the *in-sample* and *out-of-sample* scenarios. Notably, as expected, in the *out-of-sample* case, the effect of the inconsistency is significant. In the fully inconsistent case, the gluon PDF is compatible neither with the underlying law nor with the consistent case, particularly in the $x = 10^{-2} - 10^{-1}$ region, with lesser impact in the small- and large-x regions. However, this effect almost completely fades when the CMS jets dataset is included in the fitted datasets. In fact, the gluons in the two most extreme cases are fully compatible in the *in-sample* scenario.

These results strongly indicate that the Neural Network employed by NNPDF is generally capable of mitigating the effect of data inconsistency and can almost completely eliminate such an effect when provided with a consistent dataset constraining the same PDF features.

3.3 Validation of strong coupling determination

The determination of the strong coupling constant, $\alpha_s(M_Z)$, represents a significant source of uncertainty in the computation of various processes at the Large Hadron Collider (LHC). This uncertainty is frequently combined with that of parton distribution

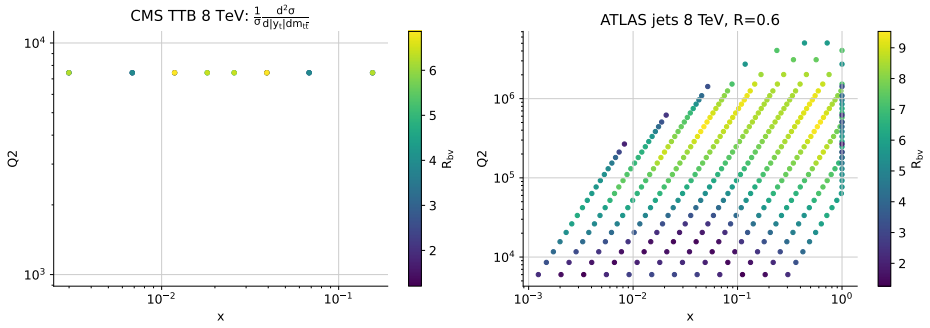


Figure 3.19: Bias-variance ratio R_{bv} per data point for $\lambda = 0.0$, computed for the CMS $t\bar{t}$ double differential cross section at 8 TeV (left panel) and the ATLAS inconsistent dataset (right panel).

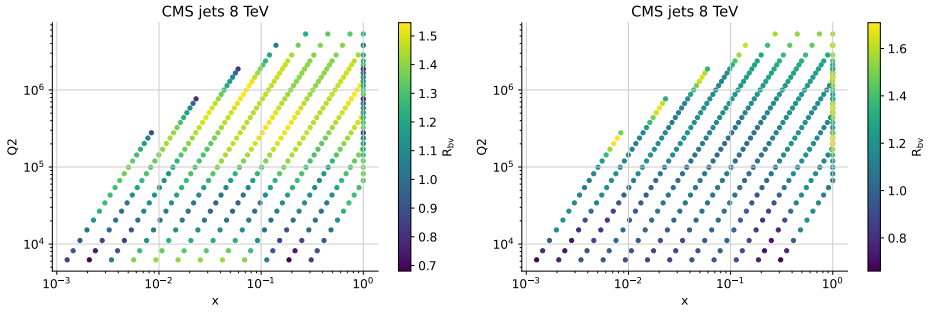


Figure 3.20: Bias-variance ratio R_{bv} per data point at $\lambda = 0.33$ computed for the CMS single jet dataset at 8 TeV. On the left, we show the result for the global fit with the same dataset *out-of-sample*, and on the right for the fit which includes it in the *in-sample* subset.

functions (PDFs), with which it exhibits a strong correlation. Among the methods for determining $\alpha_s(M_Z)$, those that do not require knowledge of PDFs, such as the global electroweak fit [150], are considered some of the most reliable. These methods are advantageous as they are not subject to the potential biases that could influence PDF determinations and, consequently, the derived value of $\alpha_s(M_Z)$.

Conversely, determining $\alpha_s(M_Z)$ in conjunction with PDFs has the benefit of being informed by a vast array of experimental measurements across multiple processes. This approach is beneficial because the uncertainties associated with specific measurements, whether of theoretical or experimental origin, are generally uncorrelated and thus tend to average out in the final determination of $\alpha_s(M_Z)$. Furthermore, a simultaneous global fit of $\alpha_s(M_Z)$ and PDFs is likely to yield a more precise, and potentially more accurate, determination than those based on pre-existing PDF sets. This enhanced precision is attributed to the comprehensive utilization of the global dataset, which appropriately accounts for the correlation between $\alpha_s(M_Z)$ and the underlying PDFs.

A determination of this nature has been conducted in [151], where the novel *correlated replica method* was introduced for the first time. In this work, we aim to validate this methodology, employed in NNPDF4.0, utilizing the closure test framework.

We begin by providing a brief review of the methodology in section 3.3.1. Then we present results of the validation in section 3.3.2.

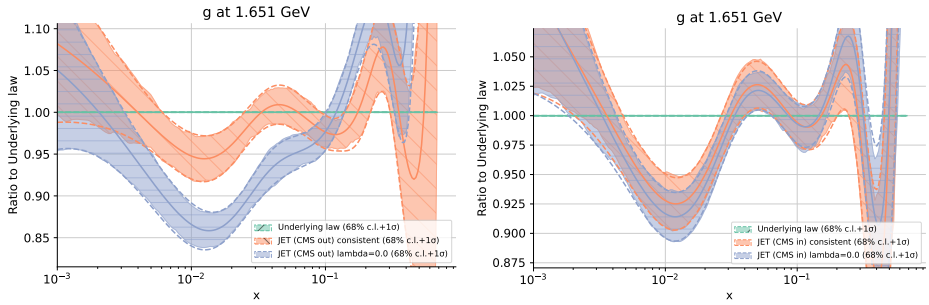


Figure 3.21: Ratio to the underlying law of the gluon PDF at 1.651 GeV, as obtained in the consistent case and in the $\lambda = 0$ case for the *out-of-sample* scenario (left) and the *in-sample* scenario (right). All central values and replicas come from fits with the same instance of L_1 data, *out-of-sample* (left) and *in-sample* (right).

3.3.1 The correlated replica method

In standard NNPDF determinations (see chapter 2), $\alpha_s(M_Z)$ is treated as a fixed parameter, alongside other theoretical parameters such as quark masses, CKM matrix elements, the fine structure constant, and similar quantities. However, it is well established (see, for example, [152] for an early reference) that PDFs are strongly correlated with the value of $\alpha_s(M_Z)$. Consequently, determining the combined PDF+α_s uncertainty for a process that depends on both requires knowledge of the PDFs as $\alpha_s(M_Z)$ varies. In light of this, NNPDF sets are routinely released for different fixed values of $\alpha_s(M_Z)$. The procedure involves generating data replicas $\mu^{(k)}$ and determining PDF replicas based on the best-fit parameters $\theta^{(k)}$, which is repeated multiple times for various values of $\alpha_s(M_Z)$.

Ideally, we seek a method for determining $\alpha_s(M_Z)$ in which the uncertainty associated with $\alpha_s(M_Z)$ is evaluated on the same basis as the uncertainty in the PDFs, thereby yielding the full probability distribution for $\alpha_s(M_Z)$, marginalized with respect to the PDF parameters. The objective is to treat $\alpha_s(M_Z)$ on equal footing with the vector of parameters θ that define the PDFs, such that the figure of merit is minimized simultaneously with respect to both $\alpha_s(M_Z)$ and θ. This approach is challenging in practice due to the dependence of theoretical predictions on $\alpha_s(M_Z)$, which are, for reasons of computational efficiency, provided as pre-computed FK tables that are determined prior to the fit using the pineline framework (see chapter 4).

This challenge can be addressed through the *correlated replica method*, as we now describe. The method is based on the concept of a *correlated replica*, or *c-replica* for short. A c-replica is a correlated set of PDF replicas, all obtained by determining the best-fit parameters $\theta^{(k)}$, but with different fixed values of $\alpha_s(M_Z)$. Given the data replica $\mu^{(k)}$, the minimization described in section 2.1 is performed multiple times across a range of fixed values of $\alpha_s(M_Z)$. Consequently, a c-replica corresponds to as many standard NNPDF replicas as the number of α_s values for which minimization has been executed, all derived from the same underlying data replica $\mu^{(k)}$.

To determine the best-fit value $\alpha_s^{(k)}(M_Z)$ for the k-th c-replica, we minimize the figure of merit χ^2, as a function of $\alpha_s(M_Z)$, computed with $\theta^{(k)}(\alpha_s)$ while varying $\alpha_s(M_Z)$ for a fixed k. Specifically, we first define the figure of merit for the k-th c-replica as

$$\chi^{2(k)}(\alpha_s) = \chi^2\left(\alpha_s, \theta^{(k)}(\alpha_s), \mu^{(k)}\right),$$

which can be interpreted as a function of $\alpha_s(M_Z)$. It is important to note that the dependence of the theoretical prediction P, and thus of the figure of merit, on $\alpha_s(M_Z)$ is both explicit and implicit through the best-fit parameters $\theta^{(k)}(\alpha_s)$. The best-fit value of $\alpha_s(M_Z)$ for the k-th c-replica is then determined as

$$\alpha_s^{(k)} = \text{argmin}\left[\chi^{2(k)}(\alpha_s)\right].$$

Therefore, in practice, $\alpha_s^k(M_Z)$ can be determined by fitting a parabola to the discrete set of values of $\chi^{2(k)}(\alpha_s)$ for each replica, and finding the minimum of the parabola.

For each data replica $\mu^{(k)}$, this procedure yields a best-fit value $(\alpha_s^{(k)}, \theta^{(k)})$ for both $\alpha_s(M_Z)$ and the PDF parameters. In other words, from each c-replica, a single best-fit value $\alpha_s^{(k)}$—an α_s *replica*—is extracted, treated equivalently to all other fit parameters. The ensemble of $\alpha_s^{(k)}(M_Z)$ values obtained from all c-replicas then provides a representation of the probability density of $\alpha_s(M_Z)$, from which standard statistical analysis can be performed. This implies that not only can the best-fit value of $\alpha_s(M_Z)$ and its uncertainty be computed as the mean and standard deviation (or as a 68% confidence interval) using the $\alpha_s(M_Z)$ replicas, but also the correlation between $\alpha_s(M_Z)$ and the PDFs, or any other PDF-dependent quantity.

In summary, the correlated replica method is analogous to the standard NNPDF methodology in that it begins by generating a set of replicas of the original data. However, it further utilizes these to construct a set of correlated α_s-dependent PDF replicas, the c-replicas, which correspond to parameters $\theta^{(k)}(\alpha_s)$ as k runs over the replica sample and $\alpha_s(M_Z)$ takes on a number of discrete values. From each c-replica, a best-fit $\alpha_s^{(k)}$ can then be determined, resulting in an $\alpha_s(M_Z)$ replica.

The correlated replica method thus leverages the fact that, within the NNPDF framework, it is sufficient to determine the best-fit set of parameters for each replica, with all other relevant information encapsulated within the replica sample. However, the trade-off for this approach is that the statistical demands of fitting $\alpha_s(M_Z)$ are inherently quite intensive, as it requires fitting a distinct parabola for each c-replica.

3.3.2 Validation of the methodology

In this section, we outline how the closure tests framework can be utilized as a validation tool for the correlated replica method.

The approach involves extending the closure tests framework, described in section 3.1, by selecting a *true* value for $\alpha_s(M_Z)$, alongside the usual *true* PDF w. Regarding this choice, as long as the chosen value lies within a reasonable range suitable for the application of perturbation theory, the specific value of this true α_s is not critical. However, to maintain consistency with the world average reported by the PDG [153], we select

$$\alpha_s^{\text{true}} \equiv \bar{\alpha}_s = 0.118\,. \tag{3.29}$$

With this choice, we can then construct the Level-0 data, as defined in eq. (3.2), as

$$L_{0,i} = \mathcal{G}(w, \bar{\alpha}_s)_i\,, \tag{3.30}$$

where it is specified that the forward map, i.e. the FK table, is computed for $\alpha_s(M_Z) = \bar{\alpha}_s$. Remind that the index i runs over the datapoints in the dataset employed in the determination.

Subsequently, the Level-1 (eq. (2.1)) and Level-2 (eq. (2.3)) data can be constructed in the usual manner. The application of the correlated replica method is then straightforward: one simply employs an instance of the Level-2 data in place of each data replica μ^k and proceeds as described in the previous section. Specifically, each Level-2 instance is fitted using various fixed values of $\alpha_s(M_Z)$, thereby generating the complete set of c-replicas. It is important to note that while the value of $\alpha_s(M_Z)$ used in each fit varies, the value employed to generate the data remains fixed at $\bar{\alpha}_s$, as specified in eq. (3.30).

In order to apply the statistical analysis described in section 3.1, a multi-closure test must be conducted. This involves generating N_{fits} instances of Level-1 data and repeating the aforementioned procedure for each instance. Upon completion, we obtain a population of N_{fits} $\alpha_s(M_Z)$ distributions, each comprising the N_{reps} $\alpha_s(M_Z)$ values derived from the c-replicas. We can then analyze the statistical properties of these populations in relation to the true underlying $\alpha_s(M_Z)$ value, $\bar{\alpha}_s$.

Having outlined the procedure used to validate the NNPDF4.0 determination of $\alpha_s(M_Z)$ through the correlated replica method, we are now prepared to provide details regarding the specific implementation of this validation (section 3.3.3) and to present the results (section 3.3.4).

3.3.3 Details of the implementation

In this section we provide some details about the specific implementation of the multi-closure test described in the previous section.

Multi-Closure test settings. The number of fits, N_{fits}, and the number of replicas in each fit, N_{reps}, has been chosen to be the same employed in the tests on inconsistent data of section 3.2. Therefore we have

$$\begin{aligned} N_{\text{fits}} &= 25 \\ N_{\text{reps}} &= 100\,. \end{aligned} \tag{3.31}$$

The perturbative order of the predictions is NLO, although in a closure test this is irrelevant, and the dataset employed in the validation is the full NNPDF4.0 dataset.

Range of $\alpha_s(M_Z)$ values. It is crucial to carefully select the list of $\alpha_s(M_Z)$ values used to fit each Level-2 data instance. Since we need to fit a parabola to the χ^2 values corresponding to each of these $\alpha_s(M_Z)$ values, it is essential to have a sufficiently broad range. Additionally, the range should be approximately symmetric around the true value $\bar{\alpha}_s$ to ensure adequate coverage of both smaller and larger values. For this purpose, we employed the following discrete set of values:

$$\alpha_s(M_Z)_i \in \{0.106, [0.114, 0.125], 0.130\}\,, \tag{3.32}$$

where the range $[0.114, 0.125]$ denotes the inclusive range from 0.114 to 0.125 with increments of 0.001.

Batches of fits. In [151], it was suggested to increase the number of fits by performing N_{batch} fits for each Level-1 data instance. By retaining only the minimum value of $\chi^{2(k)}(\alpha_s)$ across the batches for each $\alpha_s(M_Z)$ value, this method aims to mitigate the influence of potential outliers. For performance reasons, N_{batch} was set to a relatively

small value, $N_{\text{batch}} = 3$, which was nonetheless found to be sufficient. In this study, we sought to validate this choice and to investigate the effects of alternative approaches for utilizing the additional batch fits, such as averaging instead of selecting the minimum. These validations are further discussed in section 3.3.4.

Covariance matrices. A crucial aspect of this validation's settings is the selection of the covariance matrices used in the χ^2 definition for each $\alpha_s(M_Z)$ value. Recall that, to avoid the D'Agostini bias (see section 2.1), we employ the t0 prescription for the multiplicative component of the covariance matrices. This involves computing the multiplicative contribution using theoretical predictions derived from a selected t0 PDF set, rather than the PDF being fitted, and utilizing the same forward map as the fit itself. However, in the context of the fits discussed in this section, we discovered that this approach could introduce a bias of the same nature of the D'Agostini bias. Specifically, if the theoretical predictions used in the t0 prescription are calculated with the same $\alpha_s(M_Z)$ value as that used in the fit, the final $\alpha_s(M_Z)$ value obtained via the correlated replica method tends to be significantly overestimated[3].

This issue can be addressed by fixing the $\alpha_s(M_Z)$ value used to compute the theoretical predictions in the t0 prescription. For this validation, a natural choice for this fixed value is the true value $\bar{\alpha}_s$, which we adopted. However, when applying the correlated replica method to real data, the true value of α_s is, of course, unknown. In such cases, a feasible approach is to start with a reasonable initial value and proceed iteratively until the result stabilizes. Testing this iterative procedure through closure tests, we found that after just one iteration, the result was already sufficiently stable.

3.3.4 Results of the validation

The results of the determination of $\alpha_s(M_Z)$ for a single instance of Level-1 data, utilizing the settings described in the previous section, are presented in fig. 3.22. Specifically, we compare the determinations obtained using the correlated replica method with various approaches to exploit the additional batches:

- **CRM min** (3.22a): Retains only the minimum value of χ^2 for each $\alpha_s(M_Z)$. This is the original method proposed in [151].
- **CRM mean** (3.22b): Computes the average of the χ^2s values for each $\alpha_s(M_Z)$.
- **CRM min (cubic)** (3.22c): Similar to the CRM min method but fits a cubic polynomial instead of a parabola, to evaluate the impact of potential non-quadratic terms in the χ^2 profile.
- **CRM mean (full sample)** (3.22d): Combines all the replicas from the batches into a single sample.
- **CRM single batch** (3.22e): Ignores the additional fits. Included to assess the impact of using batches.

[3]The value of $\alpha_s(M_Z)$ obtained without fixing the theory employed in the t_0 prescription is in fact $\alpha_s(M_Z) = 0.11994\,(33)$, which is evidently incompatible with the true value $\bar{\alpha}_s = 0.118$. The specifics of this investigation will be detailed in an upcoming publication.

Additionally, in fig. 3.22f, we present the results obtained using a simpler methodology, here referred to as the *experimental* method. In this approach, each point in fig. 3.22f represents the χ^2 value obtained from the central replica of the fit for the corresponding $\alpha_s(M_Z)$ value. The $\alpha_s(M_Z)$ value is then extracted by fitting a parabola to these points. As argued in [154], this methodology may lead to underestimated uncertainties, a problem that the correlated replica method is designed to address. We include the experimental method in our validation to evaluate whether this issue arises.

The $\alpha_s(M_Z)$ values obtained in fig. 3.22 are all consistent with each other, as they fall within their respective 68% confidence intervals. These intervals either narrowly include or just fail to include the true value $\bar{\alpha}_s = 0.118$, which is entirely expected for a 68% confidence interval, as will be further discussed in the multi-closure test results.

All the distributions produced by the correlated replica method are approximately Gaussian, with the CRM mean distribution exhibiting the highest degree of symmetry. In contrast, the CRM single batch distribution is, as expected, the most asymmetric.

Before examining the results of the multi-closure tests, we first describe three additional variants of the correlated replica method that are included in our validation. These variants are distinguished by the fact that the fit performed on the χ^2 values for each replica is conducted in the $(\log(\alpha_s), \chi^2)$ plane rather than the (α_s, χ^2) plane. This modification is motivated by the natural logarithmic scaling of the strong coupling in QCD (see section 1.2.1). The three variants are as follows:

- **CRM LOG min**: Similar to CRM min, but the fit is performed on $\log(\alpha_s)$.
- **CRM LOG min (cubic)**: Similar to CRM LOG min, but a cubic polynomial is used for the fit.
- **CRM LOG min (quartic)**: Similar to CRM LOG min, but a quartic polynomial is used for the fit.

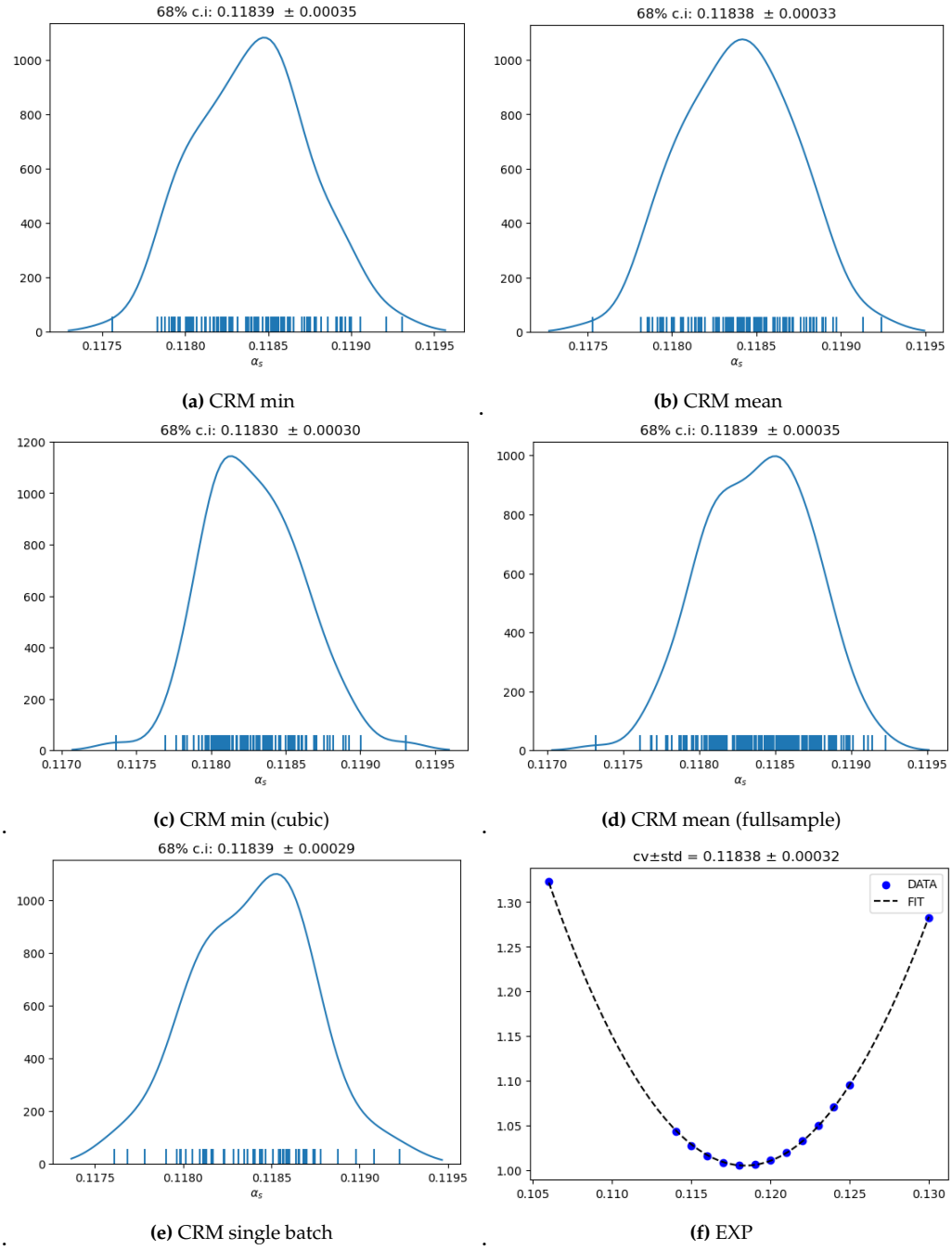
The results of the multi-closure test analysis are shown in fig. 3.23 for both the LOG case (fig. 3.23b) and the non-LOG case (fig. 3.23a). The extracted $\alpha_s(M_Z)$ values for each variant are also presented in the first column of table 3.10.

All the distributions shown in fig. 3.23a are compatible with each other and with the true value $\bar{\alpha}_s = 0.118$, although they are all shifted towards higher values. This shift appears to be a statistical fluctuation due to the relatively small number of Level-1 instances tested.

In fig. 3.23b, it can be observed that the distribution obtained with the CRM LOG min variant is not compatible with the other distributions shown in fig. 3.23a. This issue seems to stem from the presence of non-quadratic contributions in the LOG fit, as the CRM LOG (cubic) and CRM LOG (quartic) distributions are, by contrast, compatible with the others. This implies that, in a real-case scenario, it is necessary to include at least cubic terms in the fit when fitting $\log(\alpha_s)$.

In summary, the differences between the distributions obtained with each methodology are relatively minor and do not allow for a precise comparison. Therefore, in order to assess the performance of each methodology, we apply a variant of the R_{bv} estimator defined in section 3.1.1 to this analysis. Specifically, in this case, it is defined as:

$$R_{bv} = \frac{1}{N_{\text{fits}}} \sum_{i=1}^{N_{\text{fits}}} \frac{|\alpha_s^{(i)} - \bar{\alpha}_s|}{\sigma^{(i)}}\,, \tag{3.33}$$



(a) CRM min

(b) CRM mean

(c) CRM min (cubic)

(d) CRM mean (fullsample)

(e) CRM single batch

(f) EXP

Figure 3.22: Results of the determination of $\alpha_s(M_Z)$ for a single instance of Level-1 data using both the correlated replica method and the experimental method. The correlated replica method is applied in various forms, differing in the manner in which batches are utilized, as described in section 3.3.4.

where $\alpha_s^{(i)}$ is the value of $\alpha_s(M_Z)$ extracted for the i-th instance of Level-1 data, and $\sigma^{(i)}$ is its uncertainty. Remind that $R_{bv} = 1$ indicates that the uncertainties have been estimated correctly.

Additionally, to assess the uncertainty associated with the R_{bv} computed as in eq. (3.33), we apply the same bootstrap procedure used in section 3.2, as described in appendix D.

The values and uncertainties of the R_{bv} estimator for each methodology are presented in the second column of table 3.10 and illustrated in a bar plot in fig. 3.24. Although the differences among the methodologies are relatively small, it is observed that the CRM mean method performs notably better than both the EXP and CRM single batch variants, as anticipated. Interestingly, despite having the central value closest to the true value, the CRM min (cubic) method appears to perform poorly. The same observation applies to the CRM mean (fullsample) method. The performance of all the LOG variants is quite worse than the CRM min and CRM mean performances.

This validation demonstrates that the correlated replica method is effective in accurately extracting the true value of $\alpha_s(M_Z)$ and in providing well-estimated uncertainties. Among the various tested variants, the CRM mean method has been found to deliver the best performance. Therefore, it is advisable to employ this method for the extraction of $\alpha_s(M_Z)$ from real data.

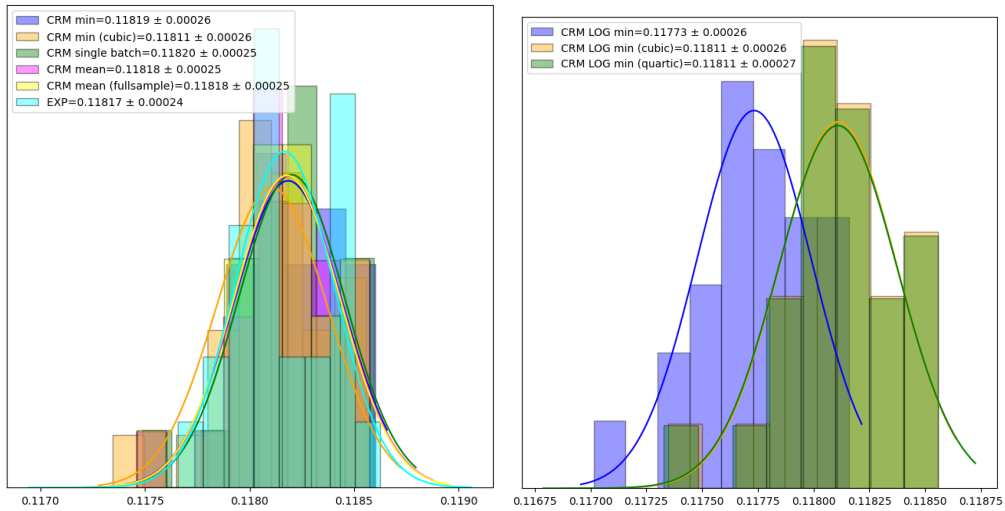


Figure 3.23: Results of the multi-closure tests for all the methodologies described in section 3.3.4 are presented. The outcomes for the LOG case are displayed separately on the right.

Method	$\alpha_s(M_Z)$	R_{bv}
CRM min	0.11819 ± 0.00026	0.95 ± 0.11
CRM min (cubic)	0.11811 ± 0.00026	0.84 ± 0.12
CRM single batch	0.11820 ± 0.00025	0.95 ± 0.10
CRM mean	0.11818 ± 0.00025	0.97 ± 0.10
CRM mean (fullsample)	0.11818 ± 0.00025	0.90 ± 0.11
EXP	0.11817 ± 0.00024	0.92 ± 0.11
CRM LOG min	0.11773 ± 0.00026	1.15 ± 0.18
CRM LOG min (cubic)	0.11811 ± 0.00026	0.86 ± 0.12
CRM LOG min (quartic)	0.11811 ± 0.00027	0.85 ± 0.11

Table 3.10: Results of the multi-closure tests validation described in section 3.3.4. We show the final value of $\alpha_s(M_Z)$ and the R_{bv} estimator (3.33) obtained for each methodology.

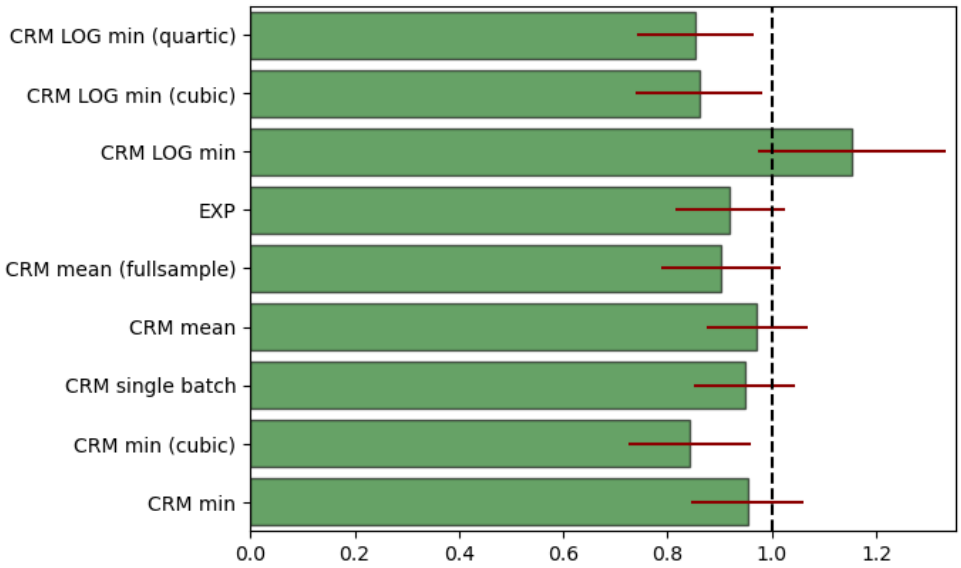


Figure 3.24: Visualization of the R_{bv} estimator and its uncertainty obtained in the multi-closure test for each methodology tested. The black dashed line highlights the $R_{bv} = 1$ value, which indicates a methodology that correctly estimates the uncertainties.

CHAPTER 4

Technical Improvements: The Pineline

Modern particle physics phenomenology increasingly depends on complex theoretical calculations which precision must align with highly accurate measurements, particularly those from experiments conducted at the Large Hadron Collider (LHC) [115]. Enhancing the accuracy of these predictions is associated with computing higher orders in the strong and/or electroweak couplings for partonic cross sections, typically executed by numerical programs, which we shall refer to as *generators* throughout this chapter. Given that these computations are demanding in terms of runtime, memory, and storage, these generators are often optimized for and limited to calculating a small set of observables. Moreover, they frequently employ different conventions and strategies. Thus, the ability to generate, store, and exchange predictions in suitable formats for a wide range of processes, allowing their use in various analyses, is highly advantageous.

We propose a framework, named `pineline`, designed to generate theoretical predictions by (1) developing a translation layer from a common input format to each of the different generators and (2) implementing a common output format for all of them. This concept, which we term *industrialization*, addresses the limitation that while specific generators suffice for the calculation of individual processes, no single generator can calculate all processes, including those beyond the LHC, such as deep-inelastic scattering processes. By interfacing with multiple generators and thus connecting them in an assembly line or *pipeline*, we can efficiently run the generator best suited for a particular process. Additionally, having a common input format facilitates parameter variations, such as those required for parameter scans.

The motivation for this project is the fitting of parton distribution functions (PDFs) [8, 155–157] (see chapters 2 and 3), although the output generated by `pineline` could be utilized in any fit or analysis requiring theoretical predictions. A notable aspect of a PDF fit in this context is the necessity of a vast number of predictions, complicating the tracking of the theoretical parameters used. While this issue is manageable for a limited number of predictions, it becomes critical for a comprehensive PDF fit to ensure that different processes utilize compatible parameter sets. Centralized parameter tracking thus facilitates the rerunning of predictions if parameter adjustments are needed. It is crucial to emphasize that PDFs are a fundamental component in any observable involving hadrons in the initial state and, therefore, must be meticulously controlled in all applications.

This chapter is based on [12, 158, 159] and it is organized as follows: in section 4.1 we outline the abstract concepts that guided the design of the `pineline` framework. In section 4.2 we provide a technical overview of the actual implementation, briefly describing the individual softwares as well. In section 4.3 we show an explicit example of application of the `pineline`.

4.1 Industrialization of high-energy theory predictions

Our goal is to align theory predictions in high-energy physics with the FAIR principles [160] (findability, accessibility, interoperability, and reusability) to promote sustainable and reproducible research.

4.1.1 Input and output formats

Our framework is designed to generate and store theory predictions in a unified format from a common set of inputs. By standardizing the input across different generators, we can enforce consistency in theoretical settings. Furthermore, by storing the predictions in a single format, we ensure they can be utilized and analyzed regardless of their original computation method.

To illustrate the diversity of generators, consider the NNPDF4.0 [8] study, which employed predictions from more than ten different programs: `APFEL` [161], `DYNNLO` [162, 163], `FEWZ` [164–166], `Madgraph5_aMC@NLO` [167, 168], `MCFM` [169–172], `Njetti` [173, 174], `NNLOjet` [175], `NLOjet++` [176], `Top++` [177], `Vrap` [178], and `SHERPA` [179]. Each of these programs requires a distinct set of inputs and parameters, and even similar inputs are provided in different formats. To address this issue, we propose a layout featuring a global *theory runcard*, which, through an appropriate generator-dependent translation layer, is fed into the target program.

The output of these programs is a hadronic observable, which has already been folded with non-perturbative objects, such as the PDF. By standardizing the output of all generators to an interpolation grid, we can reanalyze the same prediction in different scenarios without the need for costly recomputation. The evaluation of results for different sets of PDFs becomes almost instantaneous, facilitating parameter fits for objects that depend on these quantities.

In the context of PDF fitting, two common scenarios are:

- The inclusion of new data points into the fit, from existing or new experiments [180–182].
- Investigating the impact of theory settings, such as the reference value of the strong coupling $\alpha_s(M_Z)$ (see section 3.3) [154].

Both scenarios necessitate the (re-)computation of theory predictions for a large number of data points. To exemplify the scale of this task, consider NNPDF4.0, which fits more than 4500 data points across almost 100 different datasets (see section 2.1). Meeting the increasing demands from the theoretical side requires more automation to avoid time-consuming and error-prone manual processes.

The objects we work with in practice are interpolation grids [183–185], which store theory predictions independently of PDFs and the strong coupling. Interfaces for these grids to some generators are available [186–188]. Since they are independent of PDFs, they are ideally suited for PDF fits, where they have been widely adopted, but their use is not limited to this area. Note that by re-fitting the PDFs, any observable depending on them will change. However, the partonic cross sections do not depend on the PDFs. By storing them as interpolation grids, one can update all predictions without recomputing the most computationally intensive part of the observables.

In summary, our objective is to provide a reliable and user-friendly workflow that integrates the necessary intermediate steps and can scale to accommodate any amount of data.

4.1.2 Reproducibility

A crucial aspect of integrating various generators into a pipeline is ensuring the reproducibility of results. It is imperative that every prediction can be traced back to its inputs, enabling any result to be independently verified by a third party and allowing the impact of changes from a base set of parameters to be assessed. To achieve this, each interpolation grid and all intermediate objects must contain all the necessary (meta)data required for their recalculation and verification of compatibility.

Specifically, this metadata includes: the programs used, their version numbers and random seeds, the values of relevant Standard Model parameters, renormalization scheme choices, phase space cuts, and Monte Carlo uncertainties. It is noteworthy that many interpolation grids publicly available on hepdata [189] and ploughshare [190] do not include this information, though it can sometimes be inferred from associated publications. However, this data is often unavailable, complicating and prolonging the process of making comparisons.

In our framework, this metadata is explicitly embedded in the grids and all other outputs, ensuring it can be reliably and easily extracted. This practice not only facilitates reproducibility but also enhances transparency and efficiency in high-energy physics research.

4.1.3 Open-source Software

All software utilized within this framework is open source, which facilitates its distribution, use, and maintenance. In addition to the code, the data are also available online in formats that can be analyzed using open-source tools. Specifically, we store all metadata in the widely used YAML[1] format, while interpolation grids are stored as PineAPPL grids, which can be interfaced with many programming languages.

Furthermore, this work can be seen as a continuation of the effort initiated with the publication of the NNPDF fitting code [52], providing the community with all necessary tools to reproduce and perform theoretical variations of NNPDF fits.

4.2 The Pineline flowchart

In the following, we describe the technical implementation of the ideas highlighted above into the `pineline`. To do so, it is most straightforward to follow the *deliverables*, i.e. the objects that the `pineline` produces. These are illustrated in fig. 4.1 and are the oval objects, namely: (1) PineAPPL grids, (2) EKOs, and (3) fast-kernel (FK) tables.

PineAPPL grids, like APPLgrids and fastNLO tables, store theoretical predictions independently of their PDFs and the strong coupling. EKOs and FK tables are tailored towards PDF fits and translate interpolation grids to use a single factorization scale.

An extended discussion of the technical details of the various programs is beyond the scope of this thesis. Instead, we refer the interested reader to the relevant documentation and development repositories of each tool.

4.2.1 Mathematical overview

Let us consider the calculation of a single observable σ, which, for the sake of readability, we assume to contain only a single convolution, as in the case of a DIS structure func-

[1] https://yaml.org

tion. The extension to multiple convolutions is straightforward (see chapter 1). Eq. 4.1 shows the defining property of interpolation grids, namely how convolutions with PDFs $f_a(x, \mu_F^2)$ are performed:

$$\sigma = \sum_{i,j,k} \sum_a f_a(x_i, \mu_{Fj}^2) \alpha_s^{n+k}(\mu_{Rj}^2) \sigma_a^{(k)}(x_i, \mu_{Fj}^2, \mu_{Rj}^2) \,. \tag{4.1}$$

The grid itself is the set of values $\left\{\sigma_a^{(k)}(x_i, \mu_{Fj}^2, \mu_{Rj}^2)\right\}$ for all partons a and perturbative orders k. Note that the PDFs are interpolated and therefore evaluated at specific momentum fractions $\{x_i\}$ and (squared) factorization scales $\{\mu_{Fj}^2\}$, just as the partonic cross sections σ_a. For simplicity, we assume the renormalization scale equals the factorization scale $\mu = \mu_R^2 = \mu_F^2$, but the choice of scale is completely free.

The interpolation transforms the convolution integral into a sum, resulting in the grid being a PDF-independent quantity. In particular, the PDF is expanded over an interpolation basis, with the expansion coefficients being the values of the PDF on specific nodes. This means the specific interpolation basis is only used in the construction of the grid but is not relevant for the construction of the PDF table (and thus not of concern for any PDF user).

To represent interpolation grids, we use the PineAPPL library[188]. The source code can be inspected from its repository:

https://github.com/NNPDF/pineappl

and the associated documentation can be consulted at:

https://nnpdf.github.io/pineappl/ .

For the specific case of PDF fits, interpolation grids are not the most efficient representation, given that the factorization dependence of the PDFs is known perturbatively and consequently not fitted. We can therefore rewrite eq. (4.1) to refer only to a single factorization scale μ_0, which in PDF fits is known as the initial scale or the fitting scale:

$$\sigma = \sum_i \sum_a f_a(x_i; \mu_0^2) \, \mathrm{FK}_a(x_i; \mu_0^2) \,. \tag{4.2}$$

The object $\{\mathrm{FK}_a(x_i; \mu_0^2)\}$ is known as a fast-kernel (FK) table[191] and is a special case of an interpolation grid that:

- Uses a single factorization scale, and
- Contains the resummed evolution, thus combining various perturbative orders and consuming the dependence on the strong coupling.

An FK table can be computed using EKOs (see section 2.2),

$$\mathrm{FK}_a(x_i; \mu_0^2) = \sum_{b,j,k,l} \alpha_s^{n+k}(\mu_j^2) \, \mathrm{EKO}_{a,i}^{b,l,j} \, \sigma_b^{(k)}(x_l, \mu_j^2) \,, \tag{4.3}$$

where $\mathrm{EKO}_{a,i}^{b,l,j}$ are the (linear) operators resulting from the evolution equations. FK tables are ideally suited for PDF fits because the time- and memory-consuming evolutions are done only once and not during the fit.

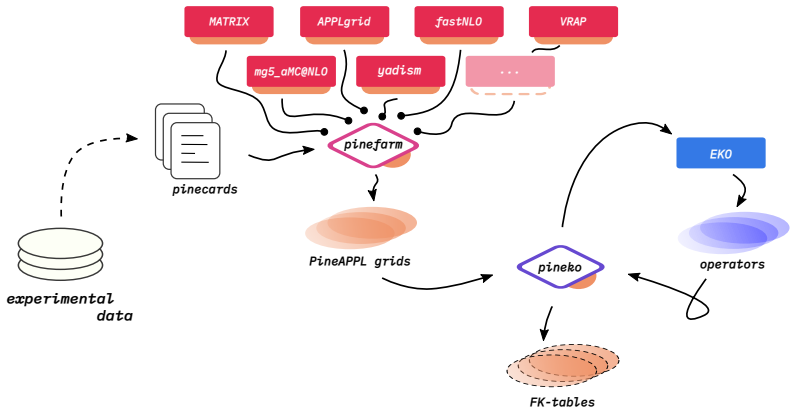


Figure 4.1: Flow diagram showing the overall pipeline architecture and deliverables in the case of parameter fits. Arrows indicate the flow of information (together with the execution order), and the orange insets on other elements indicate an interface to PineAPPL. The programs `pinefarm` and `pineko` act as interfaces between other programs and the deliverable objects, represented by ovals. These objects can be PineAPPL grids (orange) or Evolution Kernel Operators (blue) [158].

What we gain are theoretical predictions $\{\sigma\}$, represented as FK tables, which allow us to perform convolutions with a set of one-dimensional PDFs $f_a(x;\mu_0^2)$ very efficiently. However, the price we pay is that we need a set of tools that calculate all the required objects:

1. A numerical calculation must generate interpolation grids for each observable σ that we want to incorporate in a fit.
2. Next, we need to calculate the EKOs, for the corresponding choices in each observable calculated previously and the choices made in the fit.
3. Finally, we need to evolve the interpolation grids using the EKOs to generate FK tables.

In the subsequent sections, we briefly review the various programs dedicated to each step.

Note that the assumption of a single scale is chosen here only to simplify the notation, but this is not present in the actual implementation. In fact, having chosen a modularized composition of the `pineline` allows for a simplified implementation of scale variations: scale variation, as described in detail in chapter 2, can be divided into renormalization scale variation, related to the ultraviolet structure of the partonic matrix elements and which can thus only act on the level of grids, and factorization scale variation related to the collinear factorization theorem, which can either affect the split between PDFs and grids or directly EKOs. We can use such scale variations to estimate the uncertainty associated with the limited perturbative knowledge of perturbative QCD [4][5].

4.2.2 Generating grids: `pinefarm`

PineAPPL itself is agnostic to physics applications, necessitating the integration with a parton-level generator to effectively create and populate grids. This involves inter-

facing with PineAPPL, where relevant phase-space information such as $x, \mu_F, a, \ldots$ is provided. PineAPPL efficiently stores this data in a compact data structure representing $\left\{\sigma_a^{(k)}(x_i, \mu_j^2)\right\}$ (see eq. (4.1)). Practical implementation is facilitated through interfaces available in C, C++, Fortran, Python, and Rust.

Currently, PineAPPL interfaces with several generators:

- Madgraph5_aMC@NLO[167, 168] for LHC processes including NLO EW and QCD–EW corrections,
- `yadism`[192, 193] for NC and CC DIS processes,
- a modified version[2] of Vrap[178] for fixed-target Drell–Yan processes, and
- an interface to MATRIX[194] under development.

Moreover, `PineAPPL` can convert existing APPLgrids and fastNLO tables into its format using a command-line interface (CLI). Refer to Appendix A of [12] for an illustrative example.

The program `pinefarm`, presented here for the first time, abstracts away differences among various generators. For the listed generators, it manages diverse input file formats specifying the desired physical observables. Additionally, it incorporates substitutions from a theory parameters database and directly executes generators to produce predictions and aggregate necessary interpolation grids. The extensibility to more generators is facilitated by the open-source nature of `PineAPPL` and `pinefarm`.

The source code can be accessed from its repository:

`https://github.com/NNPDF/pinefarm`

and comprehensive documentation is available at:

`https://pinefarm.readthedocs.io`

4.2.3 Generating evolution kernel operators: eko

While grids $\left\{\sigma_a^{(k)}(x_i, \mu_j^2)\right\}$ are convoluted with PDFs evaluated at higher scales μ_j^2, FK tables $\left\{\mathrm{FK}_a(x_i; \mu_0^2)\right\}$ are convoluted with PDFs evaluated at the fitting scale μ_0^2, reducing the dimensionality to two dimensions for DIS observables (parton flavor index and momentum fraction), and four for hadronic observables. This reduction leverages the DGLAP equation[195–197] which dictates the scale dependence of PDFs (see section 1.3.3).

`EKO`[198, 199] has been developed specifically to solve these equations in terms of EKOs:

$$f_b(x_l, \mu_j^2) = \sum_i \sum_a \mathrm{EKO}_{a,i}^{b,l,j} \, f_a(x_i; \mu_0^2) \tag{4.4}$$

In contrast to other programs[161, 200–202], `EKO` focuses on computing these operators directly, enabling their integration within the described pipeline to generate FK tables. The PDF-independence of the operator allows for reuse across different PDF sets, enhancing the efficiency of theoretical computations.

The source code is accessible from its repository:

[2] `https://github.com/NNPDF/hawaiian_vrap`

https://github.com/NNPDF/eko

and comprehensive documentation can be found at:

https://eko.readthedocs.io

4.2.4 Generating FK tables: `pineko`

The `pineko` program, introduced here for the first time, integrates interpolation grids and EKOs to produce FK tables as described by eq. (4.3). Specifically, `pineko` retrieves essential data from a grid and a theory runcard (containing all pertinent theory parameters), and either selects or computes the required EKO as outlined in section 4.2.3. Once the EKO is calculated, `pineko` loads the grid and applies the EKO to evolve it, ultimately generating the final FK table.

Since eq. (4.2) is a specific case of eq. (4.1), PineAPPL can represent FK tables in the same format as interpolation grids. This uniform representation is crucial as it allows any theory prediction, whether derived from a Monte Carlo generator, converted from other interpolation grids, or directly computed FK tables, to be treated consistently as a PineAPPL grid at any stage of the pipeline. Consequently, the same set of tools can be utilized for analysis and manipulation across all these formats.

The division between EKO computation and grid convolution offers computational advantages. To illustrate, consider two scenarios:

- Studies involving variations in $\alpha_s(M_Z)$ [154], where only EKOs need recalculating, without affecting the grids (notably, eq. (4.1) factors out the strong coupling).
- Studies focusing on variations in M_W, where only grids require recalculating, leaving EKOs unchanged.

For transparency and accessibility, the source code for `pineko` can be accessed from its repository:

https://github.com/NNPDF/pineko

Detailed documentation is also available at:

https://pineko.readthedocs.io

Utilities

`Pineko` serves as the central user interface for the entire `pineline`, providing not only its core functionality of computing FK tables but also offering several useful utilities. One such utility, developed within the context of the study presented in chapter 2, involves integrating renormalization scale variation contributions into interpolation grids initially generated without them. The terms required for implementing scale variations at a specific perturbative order depend solely on preceding perturbative orders and known constants (see appendix A). Consequently, these terms can be calculated post-production of the interpolation grid and seamlessly integrated into it, enabling the utilization of scale-varied predictions. It is important to note that while the incorporation of renormalization scale variation contributions in `pineko` is fully implemented up to N3LO, the same functionality for factorization scale variations (see eq. (A.20)) is currently in development.

When performing scale variations at NNLO using the contributions obtained as described, relying on k-factors for NNLO predictions becomes inadequate. To address this issue, one approach is to extract the NNLO contributions directly from the k-factor and include them in the interpolation grid as if they were originally computed there. We elaborate on this issue here.

The NNLO k-factor is defined by

$$\mathrm{K}_C^{\mathrm{NNLO}} = \frac{\sum_i^N \alpha_s^2(\mu_{R,i})(m+2,0)_i + \alpha_s(\mu_{R,i})(m+1,0)_i + (m,0)_i}{\sum_i^M \alpha_s(\mu_{R,i})(m+1,0)_i + (m,0)_i}\,, \tag{4.5}$$

where (n,m) denotes the contribution of QCD order n proportional to ρ_r^m (see section 2.2). The summation over renormalization scales $\mu_{R,i}$ is necessary because each bin typically comprises contributions computed at different renormalization scales. Note that the number of such scales may vary even across different perturbative orders, as illustrated in eq. (4.5).

We use the k-factor approximation precisely because we lack the $(m+2,0)_i$ terms required for (renormalization) scale-varied predictions, for instance,

$$\begin{aligned}\mathrm{P}_{\mathrm{REN}}^{\mathrm{NNLO}} = &\sum_i^N \alpha_s^{m+2}(\rho_r\mu_{R,i})(m+2,1)_i + \alpha_s^{m+1}(\rho_r\mu_{R,i})(m+1,0)_i \\ &+ \alpha_s^{m+1}(\rho_r\mu_{R,i})(m+1,1)_i + \alpha_s^m(m,0)_i\,.\end{aligned} \tag{4.6}$$

Consequently, we are forced to use the k-factor. However, two issues arise:

- The NLO part in eq. (4.6) is computed using α_s evaluated at the shifted scale $\rho_r\mu_{R,i}$, whereas the k-factor definition (eq. (4.5)) assumes the central scale.

- Using the k-factor we are forced to multiply all the contributions, including $(m+2,1)_i$ and $(m+1,1)_i$ which is incorrect.

Two potential solutions present themselves. The first involves extracting the orders $(m+2,0)_i$ from the k-factor and directly incorporating them into the grid. The second solution involves estimating the varied k-factor $\mathrm{K}_V^{\mathrm{NNLO}}$, which maintains the same definition as the central k-factor but evaluates α_s at the varied scale. Adopting the latter solution would also necessitate rescaling the scale variation orders by $1/\mathrm{K}_V^{\mathrm{NNLO}}$. We refrain from further discussing the latter approach.

To address the absence of the term $(m+2,0)_i$ in eq. (4.5), our goal is to extract it and include it in the grid. Specifically, we start with the relationship:

$$\sum_i^N \alpha_s^2(\mu_{R,i})(m+2,0)_i = (\mathrm{K}_C^{\mathrm{NNLO}} - 1)\left(\sum_i^N \alpha_s(\mu_{R,i})(m+1,0)_i + (m,0)_i\right). \tag{4.7}$$

Initially, isolating individual terms $(m+2,0)_i$ might seem infeasible, due to the sum over the N renormalization scales. However, leveraging the differential application of the k-factor, for a specific index $i=j$, we derive:

$$\alpha_s^2(\mu_{R,j})(m+2,0)_j = (\mathrm{K}_C^{\mathrm{NNLO}} - 1)\left(\alpha_s(\mu_{R,j})(m+1,0)_j + (m,0)_j\right), \tag{4.8}$$

enabling the expression of $(m+2,0)_j$ as:

$$(m+2,0)_j = \frac{(\mathrm{K}_C^{\mathrm{NNLO}} - 1)\,(\alpha_s(\mu_{R,j})(m+1,0)_j + (m,0)_j)}{\alpha_s^2(\mu_{R,j})}. \tag{4.9}$$

The latter equation allows the extraction of individual contributions $(m+2,0)_i$ for each bin i, which can be then incorporated in the interpolation grid. This is the approach that is available within `pineko`.

4.2.5 DIS predictions: `yadism`

The provider that is responsible for the production of DIS grids is `yadism` [193, 198]. `Yadism` includes most of the currently available results in literature, specifically allowing for the computation of polarized [203] and unpolarized structure functions up to next-to-next-to-next-to-leading order (N3LO) [6] in QCD. Thanks to its modular design, the library can be easily extended as new computational results become available. The coefficients, whenever possible, have been benchmarked against `APFEL++` [202] and `QCDNUM` [29].

`Yadism` provides consistent implementations of both renormalization and factorization scale variations [5] up to any desired order. The currently implemented coefficients support renormalization scale variations up to N3LO and factorization scale variations up to NNLO.

`Yadism`, in conjunction with `EKO`, facilitates the construction of general-mass variable flavor number schemes (GM-VFNS) using coexisting PDFs with different numbers of active flavors. This approach avoids [204] the perturbative expansion of the evolution kernel, as typically done in the construction of the FONLL scheme [42]. We discuss this implementation in the following.

`Yadism` adopts a uniform treatment for all heavy quarks, ensuring that features available for charm quarks are also applicable to bottom and top quarks. This strategy enables computations involving an intrinsic bottom quark.`Yadism` offers calculations for both the fixed-flavor number scheme (FFNS) and zero-mass variable-flavor number scheme (ZM-VFNS), as well as the asymptotic limit where $Q^2 \gg m^2$ of the FFNS (FFN0), which is essential for constructing the FONLL scheme.

The source code for `yadism` can be accessed from its repository:

`https://github.com/NNPDF/yadism`

Detailed documentation is also available at:

`https://yadism.readthedocs.io`

We do not discuss the details of the implementation of `yadism`, which can be found in [193, 198].

FONLL implementation

As introduced in section 1.4, several approaches have been suggested [205–210] to include heavy quark mass effect into theoretical predictions. Here, we focus on the FONLL approach (also discussed in section 1.4), initially proposed for heavy flavor hadroproduction [211], and subsequently extended to DIS [34, 42]. The fundamental idea behind

the FONLL scheme is to combine fixed-order calculations that retain all heavy quark mass effects with collinear resummed calculations.

We present a new construction of the FONLL scheme, that is perturbatively equivalent to the original formulation up to higher-order corrections. Importantly, our approach directly addresses several shortcomings of the previous prescriptions.

A complication with the original prescription [42] arises from how the final coefficient functions are constructed: the method reformulates all expressions using a single PDF (a common practice in existing heavy quark mass schemes), which can be challenging to implement in practical applications. This limitation becomes more pronounced at high perturbative orders (e.g. N^3LO) or in hadronic collisions such as those at the LHC, where multiple PDFs are involved. Our new approach diverges by not requiring the assumption of using a single PDF. Instead, we leverage the capabilities of the EKO package [199] for solving the DGLAP evolution equations, enabling the computation of coexisting flavor number PDFs for a given factorization scale. This innovation allows for a fresh implementation of the FONLL scheme.

By utilizing coexisting flavor number PDFs, our approach achieves a distinct separation between evolution and partonic matrix elements. This separation facilitates precise control over the accuracy of both fixed-order and collinear resummed calculations, thereby enabling straightforward application across scenarios involving any number of parton distributions.

Furthermore, the earlier FONLL framework solely addressed single-mass scenarios without providing clear guidance on handling multi-mass situations. In practice, this poses a relevant issue given that the masses of charm and bottom quarks are of comparable magnitude. Our new approach specifically tackles this challenge and demonstrates how it can seamlessly handle multi-mass scenarios in a natural manner.

A comprehensive discussion of the specifics of this alternative FONLL implementation exceeds the purview of the present thesis. For a detailed exposition, the reader is referred to [212].

4.3 An example of application: K-factors vs. exact predictions

As an application of the tools described earlier, we have integrated Vrap[178] into `pinefarm` and interfaced it with `PineAPPL` to generate FK tables for fixed-target Drell–Yan observables (FTDY) up to next-to-next-to-leading order (NNLO) precision in the strong coupling constant. A step-by-step guide for implementing these results using the latest version of the `pineline` framework is documented at:

`https://nnpdf.github.io/pineline/examples/vrap`

with the final step specific to the NNPDF framework.

In this study, we apply the framework presented in this paper along with the procedures outlined in the tutorial above to perform fits similar to NNPDF4.0 [8], but with variations in the treatment of predictions for the FTDY datasets: E605 [213], E866 [214, 215], and SeaQuest [216].

Specifically, we explore these predictions in three scenarios:

1. Inclusion of FTDY datasets only at NLO QCD,

2. Inclusion of NNLO predictions approximated as K-factors (as in NNPDF4.0), and

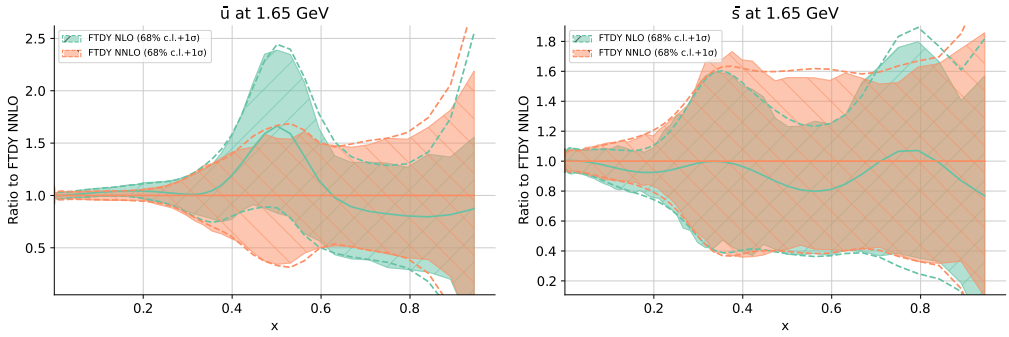


Figure 4.2: Comparison of PDF fits with and without NNLO contributions for FTDY datasets. In both cases, all other datasets are included at NNLO, differing only in the exact treatment of NNLO contributions for FTDY.

3. Inclusion of exact NNLO predictions using interpolation grids.

It's noteworthy that the majority of hadron-hadron collider data (especially all Drell–Yan Z and W production at the LHC) in all PDF fits are limited to NNLO K-factors. However, K-factors are susceptible to accidental cancellations between different partonic channels[217], suggesting that using interpolation grids for a truly NNLO-accurate PDF fit is preferable despite their computational challenges and potential lack of availability.

Figure 4.2 illustrates the impact of fitting FTDY datasets at NLO QCD (green), normalized against fits incorporating exact NNLO QCD predictions (orange). In fig. 4.3, we further investigate the influence of NNLO contributions on predictions: exact NNLO (orange) versus NLO results multiplied by bin-dependent K-factors (green).

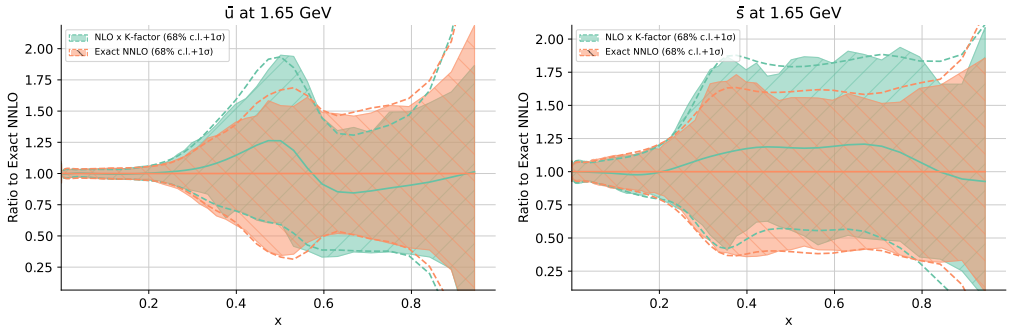


Figure 4.3: Comparison of PDF fits incorporating FTDY datasets up to NNLO: exact NNLO predictions in FK tables (orange) versus NLO results multiplied by K-factors (green). The orange fit corresponds to that in Figure 4.2.

In the case of FTDY datasets, as seen in fig. 4.2, the effect of NNLO corrections is localized within a specific region of the PDF space. Figure 4.3 further demonstrates that while fitting with K-factors shifts results towards the expected direction (as shown in fig. 4.2), K-factors do not fully capture the subtleties of NNLO contributions. A similar observation applies to the $\bar{s}$ PDF. Nonetheless, these discrepancies fall within acceptable uncertainties, and the impact of using K-factor approximations in this context appears

negligible. Quantitative differences between PDFs derived from exact NNLO calculations and those from K-factors are detailed in fig. 4.4, where differences remain insignificant, consistently well below half a standard deviation.

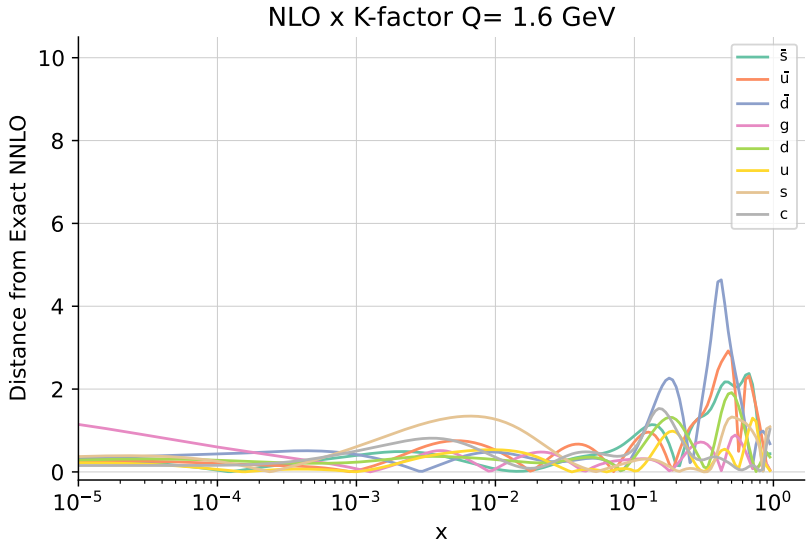


Figure 4.4: Distance plots between PDFs derived from exact NNLO calculations and those from K-factors, as computed according to Eq. (48) of [119]. A distance of 10 units corresponds to a difference of one standard deviation between the two sets of PDFs.

This example demonstrates the versatility of the framework presented in this paper. Through a single run of Vrap, we extracted predictions at NLO, NLO multiplied by K-factors, and exact NNLO (QCD) predictions, each transformed into FK tables using the same NNLO EKOs. Consequently, three different FK tables were produced for three distinct fits. For further details and to reproduce these results, refer to the `pineline` website (`https://nnpdf.github.io/pineline`), where a tutorial is available.

Summary

In this thesis, we have conducted a series of studies concerning the NNPDF methodology for the extraction of parton distribution functions (PDFs), with a particular emphasis on the estimation and validation of their associated uncertainties.

In chapter 2, we introduced the theory covariance method, which facilitates the incorporation of theoretical uncertainties arising from missing higher orders (MHO) in the determination of PDFs. Initially, we validated the approach of using scale variations to estimate these MHO uncertainties. Subsequently, the theory covariance method was applied to the NNPDF4.0 framework at NLO, NNLO, and aN3LO, where we analyzed the resulting improvements in both fit quality and perturbative convergence. Additionally, we evaluated the impact of including theoretical uncertainties in the computation of certain theoretical predictions pertinent to phenomenology.

In chapter 3, we revisited and enhanced the closure test framework, a tool that enables the validation of fitting methodologies within a controlled environment. Notably, we refined the bias-to-variance ratio estimator from its original formulation and proposed more robust alternatives. This improved closure test framework, along with the new estimators, was then employed to evaluate a scenario deliberately designed to be inconsistent, thereby testing the response of the NNPDF4.0 methodology to varying degrees of data inconsistency. We presented results for different cases, where inconsistency was injected into DIS data, DY data, and inclusive jet data. Our findings indicate that the NNPDF4.0 methodology is generally capable of providing reliable PDF uncertainties and central values, provided that the inconsistency is not excessively severe or that the affected kinematic region is not otherwise poorly constrained.

We also utilized the closure test framework to validate the extraction of the strong coupling constant, $\alpha_s(M_Z)$, via the correlated replica method. This validation process involved applying the previously discussed closure test estimators to the distribution of $\alpha_s(M_Z)$ values obtained through the correlated replica method. We validated the choices made in the original formulation of this method and assessed the impact of adopting slightly different approaches. The results of our analysis will serve as guideline for future studies on real data.

In chapter 4, we introduced a novel theoretical predictions pipeline, termed the *pineline*, designed to *industrialize*, i.e. , standardize and automate, the production of theoretical predictions necessary for QCD studies. We began by outlining the guiding principles that underpin the design of the *pineline*. Following this, we provided a brief overview of each software component within the *pineline*, with particular attention to the features

that were not described in the original publication. Finally, we demonstrated a specific example where employing the *pineline* offers advantages both in terms of performance and in the accuracy of the final predictions.

Appendices

APPENDIX A

Explicit scale-varied expressions

We collect here explicit expressions for the perturbative expansion coefficients up to N^{3}LO that are needed in order to perform scale variation according to the prescriptions discussed in Section 2.2.1.

Running of α_s. The perturbative solution of eq. (1.24) is

$$\begin{aligned}\alpha_s(\lambda\mu^2) = \alpha_s(\mu^2) &- \left(\alpha_s(\mu^2)\right)^2 \beta_0 \log\lambda + \left(\alpha_s(\mu^2)\right)^3 \left((\beta_0)^2 \log^2\lambda - \beta_1 \log\lambda\right) \\ &- \left(\alpha_s(\mu^2)\right)^4 \left((\beta_0)^3 \log^3\lambda - \frac{5}{2}\beta_0\beta_1 \log^2\lambda + \beta_2 \log\lambda\right) \\ &+ O\left(\left(\alpha_s(\mu^2)\right)^5\right). \end{aligned} \tag{A.1}$$

PDF evolution. The perturbative solution of eq. (1.48) is

$$\begin{aligned}E(\lambda\mu^2 \leftarrow \mu^2) = 1 &- \alpha_s(\mu^2)\gamma_0 \log\lambda + \left(\alpha_s(\mu^2)\right)^2 \left[\frac{1}{2}\gamma_0\left(\beta_0 + \gamma_0\right)\log^2\lambda - \gamma_1 \log\lambda\right] \\ &- \left(\alpha_s(\mu^2)\right)^3 \left[\frac{1}{6}\gamma_0\left(2\left(\beta_0\right)^2 + 3\beta_0\gamma_0 + \left(\gamma_0\right)^2\right)\log^3\lambda\right. \\ &\left. - \frac{1}{6}\left(\beta_1\gamma_0 + 2\beta_0\gamma_1 + \gamma_0\gamma_1 + \gamma_1\gamma_0\right)\log^2\lambda + \gamma_2 \log\lambda\right] \\ &+ \mathcal{O}\left(\left(\alpha_s(\mu^2)\right)^4\right). \end{aligned} \tag{A.2}$$

Scale variation of cross-sections and anomalous dimensions. The expression of the scale-varied coefficients $\overline{C}_j(\rho)$ eq. (2.23) in terms of the expansion coefficients C_j is

$$\overline{C}_0(\rho) = C_0\,, \tag{A.3}$$

$$\overline{C}_1(\rho) = C_1 + mC_0\beta_0 \log\rho\,, \tag{A.4}$$

$$\overline{C}_2(\rho) = C_2 + \frac{m(m+1)}{2} C_0 \left(\beta_0\right)^2 \log^2\rho + \left((m+1)C_1\beta_0 + mC_0\beta_1\right)\log\rho\,, \tag{A.5}$$

$$\begin{aligned}\overline{C}_3(\rho) = C_3 &+ \frac{m(m+1)(m+2)}{6} C_0 \left(\beta_0\right)^3 \log^3\rho \\ &+ \left(\frac{(m+1)(m+2)}{2} C_1 \left(\beta_0\right)^2 + \frac{m(2m+3)}{2} C_0\beta_0\beta_1\right)\log^2\rho \\ &+ \left((m+2)C_2\beta_0 + (m+1)C_1\beta_1 + mC_0\beta_2\right)\log\rho\,.\end{aligned} \tag{A.6}$$

The expression of the scale-varied coefficients $\overline{\gamma}_j(\rho)$ eq. (2.26) in terms of the expansion coefficients γ_j of course is the same, with $m = 1$ and $C \to \gamma$.

Scale variation of PDFs. The expression of the coefficients $K_j(\rho)$ in terms of the expansion coefficients γ_j can be obtained by setting $\lambda = 1/\rho$ in eq. (A.2). They are given by

$$K_0(\rho) = 1\,, \tag{A.7}$$

$$K_1(\rho) = \gamma_0 \log\rho\,, \tag{A.8}$$

$$K_2(\rho) = \frac{1}{2}\gamma_0\left(\beta_0 + \gamma_0\right)\log^2\rho + \gamma_1\log\rho\,, \tag{A.9}$$

$$\begin{aligned}K_3(\rho) = &\frac{1}{6}\gamma_0\left(2\left(\beta_0\right)^2 + 3\beta_0\gamma_0 + \left(\gamma_0\right)^2\right)\log^3\rho \\ &+ \frac{1}{2}\left(\beta_1\gamma_0 + 2\beta_0\gamma_1 + \gamma_0\gamma_1 + \gamma_1\gamma_0\right)\log^2\rho + \gamma_2\log\rho\,.\end{aligned} \tag{A.10}$$

Factorization scale variation in coefficient functions. Substituting eq. (2.32) in eq. (1.41) and switching to Mellin space, the factorized expression for the physical observable after factorization scale variation is

$$\begin{aligned}F(Q^2) &= C(Q^2)\overline{f}(Q^2,\rho_f) \\ &= C(Q^2)K(\alpha_s(\rho_f Q^2),\rho_f)E(\rho_f Q^2 \leftarrow \mu_0^2)f(\mu_0^2) &\text{(A.11)}\\ &= \overline{\overline{C}}(Q^2,\rho_f)f(\rho_f Q^2)\left[1 + O(\alpha_s)\right], &\text{(A.12)}\end{aligned}$$

where we defined

$$\overline{\overline{C}}(Q^2,\rho_f) = C(Q^2)K'(\alpha_s(Q^2),\rho_f) = \alpha_s^m(Q^2)\sum_{j=0}^{k}\left(\alpha_s(Q^2)\right)^j \overline{\overline{C}}_j(\rho_f), \tag{A.13}$$

and K' is in turn found by re-expressing $K(\rho_f Q^2,\rho_f)$ as a series in $\alpha_s(Q^2)$, namely letting

$$K'(\alpha_s(Q^2),\rho_f) = \sum_{j=0}^{k}\left(\alpha_s(Q^2)\right)^j K'_j(\rho_f) \tag{A.14}$$

with the requirement

$$K(\alpha_s(\rho_f Q^2), \rho_f) = K'(\alpha_s(Q^2), \rho_f)\,[1 + O(\alpha_s)]\,. \tag{A.15}$$

We get

$$K'_0(\rho) = 1\,, \tag{A.16}$$
$$K'_1(\rho) = \gamma_0 \log\rho, \tag{A.17}$$
$$K'_2(\rho) = \frac{1}{2}\gamma_0\,(-\beta_0 + \gamma_0)\log^2\rho + \gamma_1 \log\rho\,, \tag{A.18}$$
$$\begin{aligned} K'_3(\rho) = {} & \frac{1}{6}\gamma_0\left(2\,(\beta_0)^2 - 3\beta_0\gamma_0 + (\gamma_0)^2\right)\log^3\rho \\ & + \frac{1}{2}\,(-\beta_1\gamma_0 - 2\beta_0\gamma_1 + \gamma_0\gamma_1 + \gamma_1\gamma_0)\log^2\rho + \gamma_2\log\rho\,, \end{aligned} \tag{A.19}$$

which, substituted in eq. (A.13), leads to

$$\begin{aligned} \overline{\overline{C}}_0(\rho) &= C_0\,, \\ \overline{\overline{C}}_1(\rho) &= C_1 + \gamma_0 C_0 \log\rho\,, \\ \overline{\overline{C}}_2(\rho) &= C_2 + \frac{1}{2}\gamma_0(-\beta_0 + \gamma_0)C_0\log^2\rho + (\gamma_0 C_1 + \gamma_1 C_0)\log\rho\,, \\ \overline{\overline{C}}_3(\rho) &= C_3 + \frac{1}{6}\gamma_0\left(2\,(\beta_0)^2 - 3\beta_0\gamma_0 + (\gamma_0)^2\right)C_0\log^3\rho \\ &\quad + \frac{1}{2}\,(-\beta_1\gamma_0 - 2\beta_0\gamma_1 + \gamma_0\gamma_1 + \gamma_1\gamma_0)\,C_0\log^2\rho \\ &\quad + \frac{1}{2}\gamma_0\,(-\beta_0 + \gamma_0)\,C_1\log^2\rho \\ &\quad + (\gamma_0 C_2 + \gamma_1 C_1 + \gamma_2 C_0)\log\rho\,. \end{aligned} \tag{A.20}$$

APPENDIX B

MHOU covariance matrix prescriptions

There are two conditions that we want to satisfy in constructing the theory covariance matrix, in order to support the interpretation as the covariance matrix of our theory prior distribution:

1. We want the theory covariance to be **generated by some shift vectors** $\Delta_i(\vec{\rho})$[1]; the vectors should be proportional to the difference of predictions obtained by a theory variation $P_i(\vec{\rho})$ and the default theory in which $\vec{\rho} = \vec{\rho}_0$:

$$\Delta_i(\vec{\rho}) = c_i(\vec{\rho})\,(P_i(\vec{\rho}) - P_i(\vec{\rho}_0)) \tag{B.1}$$

$$S_{ij} = \sum_{\vec{\rho}\in V_{ij}} \Delta_i(\vec{\rho})\Delta_j(\vec{\rho}) \tag{B.2}$$

2. We want it to be **positive semi-definite**, as required for any covariance matrix

$$v_i S_{ij} v_j > 0 \qquad \forall v \in \mathbb{R}^{n_{\text{data}}} \tag{B.3}$$

Derivation

Once all the elements in eqs. (B.1) and (B.2) are spelled out, we have a clear recipe on how to compute the covariance matrix S_{ij}.

For this reason, we are going to exploit all the properties that are required or desirable (advantageous), in order to limit the available degrees of freedom: anything left, it has to be regarded as being part of the *prescription*.

The current degrees of freedom are:

1. the choice of the $p+1$ dimensional space V_{ij} of all the accounted variations (p renormalization scales, 1 factorization scale)
2. the choice of normalization coefficients $c_i(\vec{\rho}) \in \mathbb{R}$ [2]
3. the choice of the default value $\vec{\rho}_0$

The last element is trivial: it's going to be part of the prescription, but in the following we will always write $\vec{\rho}_0 = \vec{0}$ for definiteness (it's simple to replace this in the final result with $\vec{\rho}_0$ in any case).

[1] We denote $\vec{\rho} = (\rho_f, \rho_r)$.

[2] Not all values of $\mathbb{R}$ make sense, but there is quite a wide range of interesting variations: $\mathbb{N}$ for repeated points, or $\mathbb{Q}^+$ for normalizations (possibly coming from repeated points), or 0 for masking. At this level, we are just not excluding anything that has no special reason to be excluded.

Extra scales We know that the predictions for each data point only depend on two scales: the common factorization scale, and the related renormalization scale, but not the others. For this reason, it makes no sense to pick the normalization for point i dependent on the other scales, since it would introduce a dependency of the shifts on those scales that was not present in the unnormalized shifts. Thus:

$$c_i(\vec{\rho}) \equiv c_i(\rho_f, \rho_{r_i}) \tag{B.4}$$

Per-pair space Next, we claim that the space V_{ij} can not actually depend on the element ij of the covariance matrix been constructed. Indeed this stems directly for the necessity to prove eq. (B.3) that is done in the following way:

$$\sum_{i,j} v_i S_{ij} v_j = \sum_{i,j} \sum_{\vec{\rho} \in V_{ij}} v_i \Delta_i(\vec{\rho}) \Delta_j(\vec{\rho}) v_j = \tag{B.5}$$

$$= \sum_{\vec{\rho} \in V} \sum_{i,j} v_i \Delta_i(\vec{\rho}) \Delta_j(\vec{\rho}) v_j = \tag{B.6}$$

$$= \sum_{\vec{\rho} \in V} \left(\sum_i v_i \Delta_i(\vec{\rho}) \right)^2 > 0 \tag{B.7}$$

If the space V were actually dependent on ij, it would have not been possible to swap the two sums in the second step.

Space choice On the other hand, it is desirable to define the prescription only on the space of relevant scales for the given point ij. This means the factorization scale ρ_f and

off-diagonal two renormalization scales ρ_{r_i} and ρ_{r_j}, or

diagonal even a single one, if the two points are related to the same process, i.e.

$$\rho_{r_i} = \rho_{r_j}$$

We would like our expressions not to depend on the number of scales present, and only account for the scale relevant for the pair ij being considered. The easiest choice is to pick the space V to be fully factorized in the various dimensions of $\vec{\rho}$. This means that it can be written as

$$V = \prod_{i=1}^{p+1} v_i, \tag{B.8}$$

with v_i the one-dimensional space representing the variation of the single scale labeled with i.

Alternative This is not the only choice available, it is just the simplest. There is only one more option that guarantees the independence of the projection on the pair ij, i.e. factorize the space for each possible value of ρ_f. This option will be explored in appendix B.0.2.

In the case of a fully factorized space, the complex choice of the space is reduced on $p+1$ choices for one dimensional spaces. But if there is no reason to distinguish processes at this level, it is reasonable to pick the same space for each renormalization scale.

In practice, the basic one dimensional space will be always the same[3]:

$$v = \{1/4, 1, 4\} \equiv \{-, 0, +\} \tag{B.9}$$

and the overall space will be just the product:

$$V = v^{p+1} \tag{B.10}$$

Normalization At this point, all the arbitrariness left for the prescription is encoded in the normalization coefficients. With our simple choice of the space there is no reason to choose complex coefficients, thus we will define the following prescriptions:

$$c_i(\vec{\rho}) = \begin{cases} 1/\sqrt{N_m} & \rho \in V_m^i \\ 0 & \text{else} \end{cases} \tag{B.11}$$

The spaces V_m^i now defines our point prescription, together with the overall normalization N_m, since the $c_i(\vec{\rho})$ are acting as *masks* on the points $\vec{\rho}$ not belonging to the space. For the former we'll choose:

$$V_m^i = v_m^i \times \{-, 0, +\}^{p-1} \tag{B.12}$$

where the two dimensional spaces v_m^i are always the same space v_m, but for the scales (ρ_f, ρ_{r_i}), while the other scales are free to assume any possible value.

For the normalizations instead, there is no strict nor reasonable way to fix it completely, but it is possible to fix the scaling in the case of a space v_m and v with an hypothetically large number of point: since we don't want the normalization of the theory covariance matrix to depend on the number of points being in the prescription, we'll choose

$$N_m \propto |v_m| \cdot |v| = m \cdot 3^{p-1} \tag{B.13}$$

B.0.1 Examples of prescriptions

Since the presence of many processes have been reconciled at a theoretical (even though abstract) level, here we will concentrate on fully spelled out examples, in the simplest case of only two data points (1 and 2) belonging to two distinct processes.

Again, the following is in no way a proof, which has been spelled out in details in the previous section, for which considering more than two processes is extremely relevant.

We will show the actual results of the obtained prescriptions for the on-diagonal, S_{11}, and off-diagonal, S_{12} cases.

Notice that, with respect to [4], here we have not yet introduced the factor s, but it would still be allowed by eq. (B.13). In order to make the comparison with [4] easier, in this section we'll define the actual normalization including this factor, so:

$$N_m = \frac{m \cdot 3^{p-1}}{s_m} \tag{B.14}$$

[3]The one spelled out is only an option, any other space would work equally well.

For convenience, the unnormalized shifts will be called δ, i.e.:

$$\delta_i(\vec{\rho}) \equiv \Delta_i(\vec{\rho}) \cdot \sqrt{N_m} \tag{B.15}$$

In general, the expressions for the *diagonal* and *off-diagonal* cases with only two process, $p = 2$, are the following:

diagonal effectively two-dimensional, since both the shifts depend only on two scales

$$S_{11} = \sum_{\vec{\rho} \in V} \Delta_1(\vec{\rho}) \Delta_1(\vec{\rho}) = \tag{B.16}$$

$$= \frac{s_m}{3 \cdot m} \sum_{\vec{\rho} \in V_m^1} \delta_1(\vec{\rho})^2 = \tag{B.17}$$

$$= \frac{s_m}{m} \sum_{(\rho_f, \rho_{r_1}) \in v_m^1} \delta_1(\rho_f, \rho_{r_1}, 0)^2 \tag{B.18}$$

where in the last step a single value has been chosen for ρ_{r_2}, since δ_1 does not depend on this scale.

off-diagonal effectively three-dimensional, that only for this specific problem coincide with the whole space (for a greater number of processes, would be itself a projection)

$$S_{12} = \sum_{\vec{\rho} \in V} \Delta_1(\vec{\rho}) \Delta_2(\vec{\rho}) = \tag{B.19}$$

$$= \frac{s_m}{3 \cdot m} \sum_{\vec{\rho} \in V_m^1 \cap V_m^2} \delta_1(\vec{\rho}) \delta_2(\vec{\rho}) \tag{B.20}$$

$$= \frac{s_m}{3 \cdot m} \sum_{\vec{\rho} \in V_m^1 \cap V_m^2} \delta_{12}(\vec{\rho}) \tag{B.21}$$

where in the last step we defined $\delta_{12}(\vec{\rho}) \equiv \delta_1(\vec{\rho}) \delta_2(\vec{\rho})$.

9 points

The easiest prescription is the *so-called* 9 points prescription, because it corresponds to consider the whole two dimensional space as V_9^i, thus the two elements to be fixed are:

$$v_9 = \{-, 0, +\}^2 \tag{B.22}$$

$$N_9 = \frac{8 \cdot 3}{2} = 12 \tag{B.23}$$

with $s_9 = 2$ (naïvely because two scales are involved).

In the following, the expressions for the diagonal and off-diagonal cases are formatted in order to stress the connection with the various pictures in this section. Concerning the *diagonal* expressions they are formatted on three lines, with three terms each, such that each term correspond to one point in the two-dimensional diagram. Since *off-diagonal* would correspond to a three-dimensional picture, this picture is ideally sliced in two-dimensional planes, and each plane is displayed in the equation as a block of terms

in square brackets, and slightly indented with respect to previous blocks. In order to preserve the shape, and to stress the effect of zero values in the $c_i(\vec{\rho})$, missing terms are explicitly marked with zeros.

diagonal for this prescription, we effectively have only 8 shifts, since out of the 9 theory predictions, one shift vanishes, just because it is used as the reference

$$\begin{aligned} S_{11} = \frac{1}{4}\Big[&\delta_1(-,-,0)^2 + \delta_1(-,0,0)^2 + \delta_1(-,+,0)^2 \; + \\ &\delta_1(0,-,0)^2 \; + \quad 0 \quad + \delta_1(0,+,0)^2 \; + \\ &\delta_1(+,-,0)^2 + \delta_1(+,0,0)^2 + \delta_1(+,+,0)^2\Big] \end{aligned} \tag{B.24}$$

off-diagonal the two Δ_i combine in three dimensions: each one contains 3 zero elements (relative to the two dimensional central value), but the two are overlapping over the central point $(\rho_f, \rho_{r_1}, \rho_{r_2}) = (0,0,0)$, leading to only 5 zero elements out of $3^3 = 27$ total elements, see appendix B.0.1; thus the 22 non-vanishing elements are the following:

$$\begin{aligned} S_{12} = \frac{1}{12}\Big\{&\Big[\delta_{12}(-,-,-) + \delta_{12}(-,-,0) + \delta_{12}(-,-,+) \; + \\ &\delta_{12}(-,0,-) \; + \delta_{12}(-,0,0) \; + \delta_{12}(-,0,+) \; + \\ &\delta_{12}(-,+,-) + \delta_{12}(-,+,0) + \delta_{12}(-,+,+)\Big] + \\ &\Big[\delta_{12}(0,-,-) + \quad 0 \quad + \delta_{12}(0,-,+) \; + \\ &\quad 0 \qquad + \quad 0 \quad + \quad 0 \qquad + \\ &\delta_{12}(0,+,-) + \quad 0 \quad + \delta_{12}(0,+,+)\Big] + \\ &\Big[\delta_{12}(+,-,-) + \delta_{12}(+,-,0) + \delta_{12}(+,-,+) \quad + \\ &\delta_{12}(+,0,-) \; + \delta_{12}(+,0,0) \; + \delta_{12}(+,0,+) \quad + \\ &\delta_{12}(+,+,-) + \delta_{12}(+,+,0) + \delta_{12}(+,+,+)\Big]\Big\} \end{aligned} \tag{B.25}$$

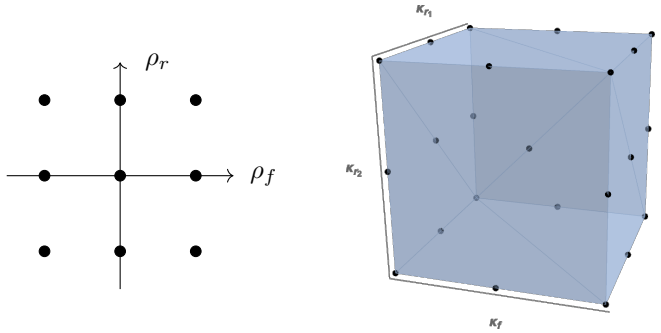


Figure B.1: Visualization of the 9 points prescription for the diagonal (2 dimensional) and off-diagonal (3 dimensional) elements.

5 points

Another interesting prescription is the 5 points one, since it is a rather minimal prescription involving both renormalization and factorization scale.

$$v_5 = \{(-,0),(0,-),(+,0),(0,+)\} \tag{B.26}$$

$$N_5 = \frac{4 \cdot 3}{2} = 6 \tag{B.27}$$

with $s_5 = 2$ (same reason of eq. (B.22)).

diagonal for this prescription, we effectively have only 4 shifts, since only 5 theory predictions are taken into account[4], and, as for the 9 points, one is used as reference

$$\begin{aligned} S_{11} = \frac{1}{2}\Big[\quad 0 \quad & + \delta_1(-,0,0)^2 + \quad 0 \quad + \\ \delta_1(0,-,0)^2 + \quad 0 \quad & + \delta_1(0,+,0)^2 + \\ 0 \quad & + \delta_1(+,0,0)^2 + \quad 0 \quad \Big] \end{aligned} \tag{B.28}$$

off-diagonal in this case the two two-dimensional normalizations combine into one three-dimensional pattern, where non-zero elements are arranged in the shape of a double square pyramid: only central value is allowed for $\rho_f \neq 0$, while the four

[4] with the shape of a Greek cross, as the $+$ symbol

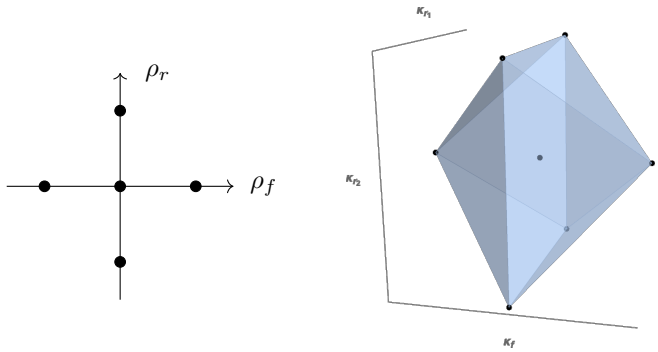


Figure B.2: Visualization of the 5 points prescription for the diagonal (2 dimensional) and off-diagonal (3 dimensional) elements.

corners are left for $\rho_f = 0$ (same as the 9 points in this case), see appendix B.0.1

$$
\begin{aligned}
S_{12} = \frac{1}{6}\Bigg\{ & \begin{bmatrix} 0 & + & 0 & + & 0 & + \\ 0 & + & \delta_{12}(-,0,0) & + & 0 & + \\ 0 & + & 0 & + & 0 & \end{bmatrix} + \\
& \begin{bmatrix} \delta_{12}(0,-,-) & + & 0 & + & \delta_{12}(0,-,+) & + \\ 0 & + & 0 & + & 0 & + \\ \delta_{12}(0,+,-) & + & 0 & + & \delta_{12}(0,+,+) & \end{bmatrix} + \\
& \begin{bmatrix} 0 & + & 0 & + & 0 & + \\ 0 & + & \delta_{12}(+,0,0) & + & 0 & + \\ 0 & + & 0 & + & 0 & \end{bmatrix} \Bigg\}
\end{aligned}
\tag{B.29}
$$

$\bar{5}$ points

Just another option with renormalization and factorization scale, with same two dimensional volume, but a different geometry.

$$v_5 = \{(-,-),(-,+),(+,-),(+,+)\} \tag{B.30}$$

$$N_5 = \frac{4 \cdot 3}{2} = 6 \tag{B.31}$$

with $s_{\bar{5}} = 2$ (same reason of eq. (B.22)).

diagonal for this prescription, we effectively have only 4 shifts, since only 5 theory predictions are taken into account[5], and, as for the 9 points, one is used as reference

$$S_{11} = \frac{1}{2}\Big[\delta_1(-,-,0)^2 + \quad 0 \quad + \delta_1(-,+,0)^2 \; + \\ 0 \qquad + \quad 0 \quad + \quad 0 \qquad + \\ \delta_1(+,-,0)^2 + \quad 0 \quad + \delta_1(+,+,0)^2\Big] \tag{B.32}$$

off-diagonal also in this case the two two-dimensional normalizations $c_i(\vec{\rho})$ have the combined effect of setting to zero a lot of elements in the three dimensional space, this leaving the shape of an empty cube: the four corners are now left for $\rho_f \neq 0$, and no point is left for $\rho_f = 0$

$$S_{12} = \frac{1}{6}\Big\{\Big[\delta_{12}(-,-,-) + \quad 0 \quad + \delta_{12}(-,-,+) \; + \\ 0 \qquad + \quad 0 \quad + \quad 0 \qquad + \\ \delta_{12}(-,+,-) + \quad 0 \quad + \delta_{12}(-,+,+)\Big] + \\ \Big[\; 0 \; + \; 0 \; + \; 0 \; + \\ 0 \; + \; 0 \; + \; 0 \; + \\ 0 \; + \; 0 \; + \; 0 \;\Big] + \tag{B.33}$$

$$\Big[\delta_{12}(+,-,-) + \quad 0 \quad + \delta_{12}(+,-,+) \quad + \\ 0 \qquad + \quad 0 \quad + \quad 0 \qquad + \\ \delta_{12}(+,+,-) + \quad 0 \quad + \delta_{12}(+,+,+)\Big]\Big\}$$

B.0.2 Alternative space: ρ_f slices

Previously, we made a set choices for the degrees of arbitrariness exposed at the beginning. All of them were yield by a strict requirement (needed to obtain a property, like $S_{ij} \geq 0$) or by a reasonable request (e.g. not adding further dependencies with normalizations, which led to eq. (B.4)). Only in one single case we made an assumption based on an unneeded simplicity: the choice of the space as fully factorized.

This choice is sensible for the renormalization scales: why should the space look different seen from the perspective of different processes? Why different processes should be correlated by the space? On the other hand, it is completely arbitrary for the factoriza-

[5] with the shape of St. Andrew's cross, as the $\times$ symbol

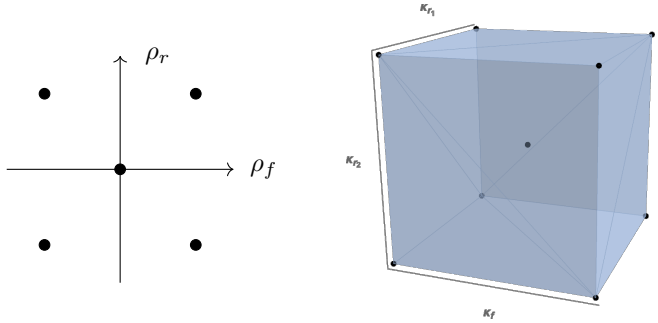


Figure B.3: Visualization of the $\bar{5}$ points prescription for the diagonal (2 dimensional) and off-diagonal (3 dimensional) elements.

tion scale. Since factorization scale ρ_f is treated separately from renormalization scales ρ_{r_i}, no surprise if even the space symmetry somehow is broken on ρ_f[6].

Thus, we can have a different factorized space for each different value of ρ_f:

$$V = \bigsqcup_{\rho_f \in v_f} V(\rho_f) \tag{B.34}$$

$$V(\rho_f) \equiv v(\rho_f)^p \tag{B.35}$$

where v_f is the space of possible values of ρ_f (usually it will be just v of eq. (B.9)), and $v(\rho_f)$ is instead the space of renormalization scales related to that single value of the factorization scale.

In this case also the definition of the normalizations $c_i(\vec{\rho})$ should change with respect to those defined in eq. (B.11) in order to account for this, since the different spaces contain different numbers of points. We decide to normalize the elements such that once the full space is projected over each of the two dimensional spaces (ρ_f, ρ_{r_i}), the coefficients of the various shifts are equal to one, thus:

$$c_i(\vec{\rho})^2 \propto \frac{1}{\sum_{\rho'_f} v(\rho'_f)} \frac{|v(\rho_f)|}{|V(\rho_f)|} = \frac{1}{m \cdot |v(\rho_f)|^{p-1}} \tag{B.36}$$

since the scales projected are all renormalization scales but a single one, that is the relevant one for the given i, and together with ρ_f make the two dimensional space, whose volume is $\sum_{\rho'_f} v(\rho'_f) = m$.

B.0.3 Examples of prescriptions

In this case as well, for better comparison with [4], we introduce the factor of s in the normalization of eq. (B.36), thus

$$c_i(\vec{\rho})^2 = \frac{s_m}{m \cdot |v(\rho_f)|^{p-1}} \tag{B.37}$$

[6]For the ρ_{r_i}, choosing them factorized and uniform as argued, a permutation invariance is present, and makes sense. No reason to extend it to ρ_f.

Furthermore, same as in appendix B.0.1 (on purpose, to stress comparison) we consider the case of only two data points (1 and 2) belonging to two distinct processes. With this limited case it is harder to appreciate the difference of the constructions, since it actually lies in the way the different three dimensional shapes for pair of processes are reconciled in the full $p+1$-dimensional space. However, this difference has already been stressed in the abstract construction of the two classes of prescriptions, thus the purpose of this examples is different: to showcase the different expressions obtained fully explicitly. For this aim the choice of considering just two points is fully satisfactory.

For this second set of examples there is no need to rewrite the full set of terms: they are the exact same of appendix B.0.1, the only difference will be in the coefficients, that now might depend on the value of ρ_f because of the space structure (and they will always depend on it).

Thus, the expressions for the *diagonal* and *off-diagonal* cases with only two process, $p = 2$, in this second class of prescriptions are the following:

diagonal effectively two-dimensional, since both the shifts depend only on two scales

$$S_{11} = \sum_{\vec{\rho}\in V} \Delta_1(\vec{\rho})\Delta_1(\vec{\rho}) = \tag{B.38}$$

$$= \sum_{\rho_f\in v_f} \frac{s_m}{|v(\rho_f)|\cdot m} \sum_{\vec{\rho}_R\in V(\rho_f)} \delta_1(\vec{\rho})^2 = \tag{B.39}$$

$$= \frac{s_m}{m} \sum_{\rho_f\in v_f} \sum_{\rho_{r_1}\in v(\rho_f)} \delta_1(\rho_f,\rho_{r_1},0)^2. \tag{B.40}$$

where in the last step a single value has been chosen for ρ_{r_2}, since δ_1 does not depend on this scale (this trivial sum cancels with the factor of $|v(\rho_f)|$ in the denominator).

Notice that the last sum $\sum_{\rho_f\in v_f}\sum_{\rho_{r_1}} = \sum_{(\rho_f,\rho_{r_1})\in v^1_m}$, thus the finally formula for the diagonal case is the same of eq. (B.18). While this is not a proof of the general case, it is simple to show (in essentially the same way of above) that this is the formula obtained for any number of processes p.

off-diagonal effectively three-dimensional, that only for this specific problem coincide with the whole space

$$S_{12} = \sum_{\vec{\rho}\in V} \Delta_1(\vec{\rho})\Delta_2(\vec{\rho}) = \tag{B.41}$$

$$= \sum_{\rho_f\in v_f} \frac{s_m}{|v(\rho_f)|\cdot m} \sum_{\vec{\rho}_r\in V(\rho_f)} \delta_1(\vec{\rho})\delta_2(\vec{\rho}) \tag{B.42}$$

$$= \frac{s_m}{m} \sum_{\rho_f\in v_f} \frac{1}{|v(\rho_f)|} \sum_{\vec{\rho}_r\in V(\rho_f)} \delta_{12}(\rho_f,\rho_{r_1},\rho_{r_2}). \tag{B.43}$$

Since the space of this second class is engineered to give the same terms of the first one (both diagonal and off-diagonal), and the normalizations are chosen such to obtain uniform coefficients for the diagonal case (and then they are the exact same of the first class, as noted above), the only difference will be in the **coefficients of the off-diagonal** case, and they can only depend on the factorization scale ρ_f. For this reason, we will not

repeat the full construction of the previous section, but just adopt a concise notation to make the different coefficients explicit in the off-diagonal expressions:

$$S_{12} = \frac{s_m}{m \cdot k_m} \left(c_m(-)\delta_{12}(-, \cdots) + c_m(0)\delta_{12}(0, \cdots) + c_m(+)\delta_{12}(+, \cdots) \right) \tag{B.44}$$

where:

- k_m is the least common multiple of the $|v(\rho_f)|$, in order to leave integer coefficients in the sum
- $c_m(\rho_f)$ is the leftover the $1/|v(\rho_f)|$ once $1/k_m$ has been factored out
- $\delta_{12}(\rho_f, \cdots)$ is a placeholder for all the terms with that value of ρ_f, as they have been spelled out in the corresponding prescription in appendix B.0.1

9 points

The specification of this prescription is almost the same of the corresponding one for the first class:

$$v_9(-) = v_9(+) = \{-, 0, +\} \tag{B.45}$$
$$v_9(0) = \{-, +\} \tag{B.46}$$
$$\tag{B.47}$$

Therefore, the resulting *off-diagonal* expression is:

$$S_{12} = \frac{2}{8 \cdot 6} \left(2\, \delta_{12}(+, \cdots) + 2\, \delta_{12}(-, \cdots) + 3\, \delta_{12}(0, \cdots) \right) \tag{B.48}$$
$$= \frac{1}{24} \left(2\, \delta_{12}(+, \cdots) + 2\, \delta_{12}(-, \cdots) + 3\, \delta_{12}(0, \cdots) \right) \tag{B.49}$$

5 points

For this prescription, the difference is a bit more relevant, mainly in terms of the overall factor, since no one of the $v_5(\rho_f)$ spaces has the maximal allowed cardinality, i.e. 3[7]

$$v_5(-) = v_9(+) = \{0\} \tag{B.50}$$
$$v_5(0) = \{-, +\} \tag{B.51}$$
$$\tag{B.52}$$

Therefore, the resulting *off-diagonal* expression is:

$$S_{12} = \frac{2}{4 \cdot 2} \left(2\, \delta_{12}(+, \cdots) + 2\, \delta_{12}(-, \cdots) + \delta_{12}(0, \cdots) \right) \tag{B.53}$$
$$= \frac{1}{4} \left(2\, \delta_{12}(+, \cdots) + 2\, \delta_{12}(-, \cdots) + \delta_{12}(0, \cdots) \right) \tag{B.54}$$

[7] Of course even 3 is completely arbitrary, as explained in eq. (B.9), and the related note, but both classes of prescriptions are perfectly *adaptive* w.r.t. this value, i.e. their definitions work perfectly fine in the general case.

$\bar{5}$ points

It is worth to analyze separately also this prescription: the former two are enough to exemplify the regular cases, but this one is slightly *degenerate*. Indeed, one of the spaces is actually empty.

$$v_5(-) = v_9(+) = \{-, +\} \tag{B.55}$$
$$v_5(0) = \{\} \tag{B.56}$$
$$\tag{B.57}$$

We need to generalize a bit the definition given above: k_m is chosen to be the least common multiple of all **non-zero** coefficients. Finally, the *off-diagonal* expression for this prescription is:

$$S_{12} = \frac{2}{4 \cdot 2} \left(\delta_{12}(+, \cdots) + \delta_{12}(-, \cdots)\right) \tag{B.58}$$
$$= \frac{1}{4} \left(\delta_{12}(+, \cdots) + \delta_{12}(-, \cdots)\right) \tag{B.59}$$

Note that the expressions obtained in this section, with the assumption of having a space sliced in ρ_f, are the same proposed in [4].

APPENDIX C

Impact of the improved estimators

In this appendix, we compare the results obtained using the new definition of the bias-to-variance ratio given in eq. (3.17) with the one previously used in [9]. The primary focus is to highlight the necessity of introducing a different definition for the consistency estimator in a multiclosure fit.

Consider a typical global multiclosure fit with $N_{\text{fits}} = 25$ and $N_{\text{rep}} = 100$. The bias-to-variance ratio is chosen as the summary statistic to evaluate the consistency of the NNPDF fitting procedure. To perform this evaluation, a specific testing set defined by certain datasets is selected, and the bias-to-variance ratio of the fit's output is computed on the observables included in these datasets. We aim to compare the behavior of the PCA-based R_{bv} with the previously adopted definition, demonstrating why the previous estimator lacked interpretability and faithfulness.

Define the testing data to consist solely of two experimentally uncorrelated datasets. Given the old definition, it is straightforward to see that the global bias-to-variance ratio would be given by:

$$R_{bv} = \sqrt{\frac{\mathbb{E}_\eta[B^{(l)}_{\text{ds1}}] + \mathbb{E}_\eta[B^{(l)}_{\text{ds2}}]}{\mathbb{E}_\eta[V^{(l)}_{\text{ds1}}] + \mathbb{E}_\eta[V^{(l)}_{\text{ds2}}]}}. \tag{C.1}$$

The problem with this definition is that we desire a global test statistic to reflect on *how many data points* the fitting procedure can be defined as consistent. This requirement boils down to the quantities B_{ds1} and B_{ds2} being distributed according to a χ^2 distribution with $N_{\text{dof}} = N_{\text{data}}$. This differs from the PCA procedure, where only a certain number of degrees of freedom survive after the regularization of the (PDF-induced) covariance matrix.

It can be shown that the previously adopted definition fails to meet this requirement. Consider, for instance, two datasets: one from Drell-Yan and another from HERA NC. The first consists of 15 data points, while the second comprises 254 data points. Given this proportionality between the sizes of the datasets, the HERA dataset should weigh much more than the Drell-Yan dataset. However, examining the distribution of the bias and variance previously computed in a multiclosure fit, we observe that this is not the case (fig. C.1).

As can be clearly seen from the figures, the distributions for the two different datasets do not follow a correctly normalized χ^2 distribution. In fact, when computing the global R_{bv} by merging the two datasets, the Drell-Yan dataset will have a much higher weight than the DIS dataset, despite having fewer data points. Specifically, following the previous notation, the means for the bias and variance of the two datasets are listed in

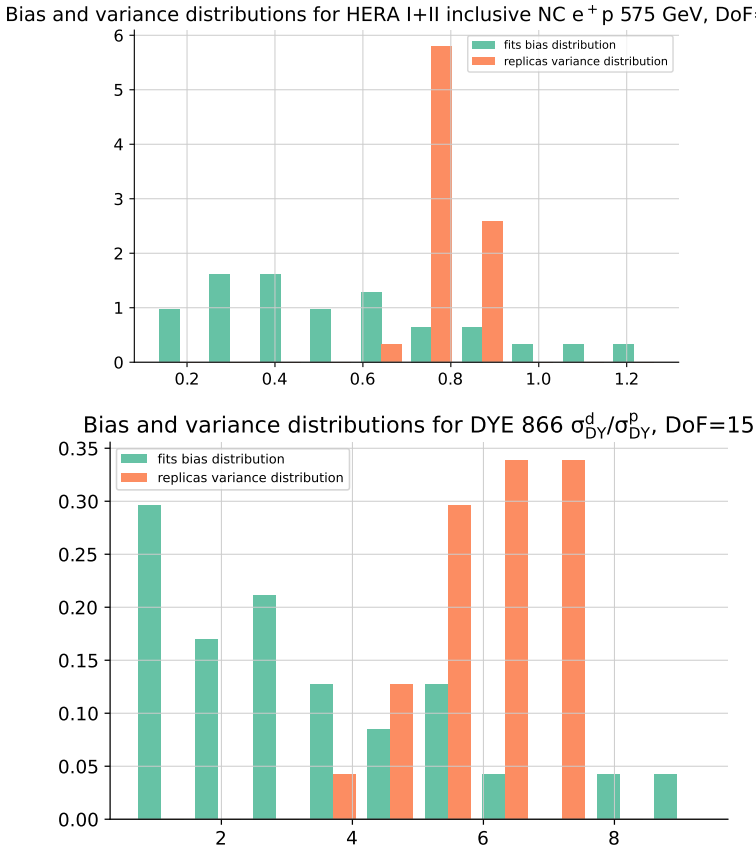


Figure C.1: Distribution of mean variance and bias across fits for the inclusive DIS neutral current HERA dataset (top) and for the differential Drell-Yan cross section (bottom).

table C.1. Note that the value of the bias-to-variance ratio is

$$R_{bv} \approx 0.76\,, \tag{C.2}$$

which is confirmed to be mostly driven by the DY dataset.

Dataset	N_{data}	$\mathbb{E}_\eta[B]$	$\mathbb{E}_\eta[V]$	R_{bv}
HERA I+II inclusive NC e^+p 575 GeV	254	0.6	0.8	0.87
DYE 886 $\sigma^d_{DY}/\sigma^p_{DY}$	15	3.3	6.0	0.74

Table C.1: Central values for bias and variance computed following the "old" definition.

When computing the global R_{bv} with the new definition, we cannot make the same point: in particular, we do not know beforehand which degrees of freedom will be re-

moved by the PCA. Therefore, it is incorrect to state that the global R_{bv} is given by:

$$R_{bv} = \sqrt{\frac{\mathbb{E}_\eta[B^{(l)}_{\text{ds1, PCA}}] + \mathbb{E}_\eta[B^{(l)}_{\text{ds2, PCA}}]}{\mathbb{E}_\eta[V^{(l)}_{\text{ds1, PCA}}] + \mathbb{E}_\eta[V^{(l)}_{\text{ds2, PCA}}]}}. \tag{C.3}$$

Nevertheless, by plotting the global histogram for the two merged datasets (fig. C.2), we can clearly see that the bias and variance quantities more closely follow a correctly normalized χ^2 distribution. This indicates that each individual data point has the correct weight, and merging datasets in a final analysis with PCA provides the correct proportionality concerning the total number of degrees of freedom.

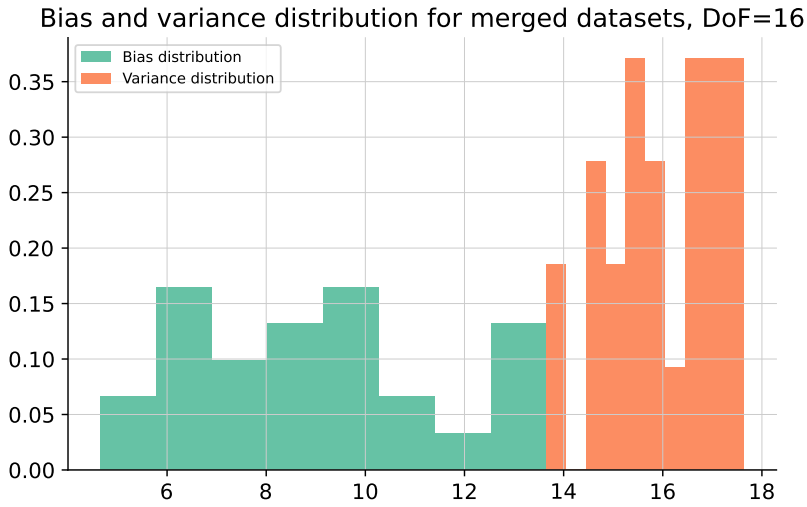


Figure C.2: Distribution of bias and variance across fits for the two merged datasets, computed with PCA.

APPENDIX D

Bootstrap algorithm definition

To quantify the uncertainty associated with the multi-closure test estimators utilized in this study, such as the bias-variance ratio R_{bv}, we implemented a bootstrapping procedure on the closure test fits. This approach enables us to achieve more reliable uncertainty quantification than merely computing the variance of the estimator on the 25 closure test fits. Given n closure test fits

$$F_1, \ldots, F_n \overset{\text{i.i.d.}}{\sim} \hat{P}_n, \tag{D.1}$$

the algorithm involves the following steps:

1. **Bootstrap Sample Generation**: Randomly select, with replacement, n closure tests from the n available closure tests (typically 25). Within each selected closure test, randomly select, again with replacement, m replicas (we use $m = 60$, as shown in Table D.1, which demonstrates the stability of the estimator as a function of m) from the total of 100 replicas. This process creates a bootstrap sample comprising n closure tests, each containing m replicas.

$$F_1^*, \ldots, F_n^* \overset{\text{i.i.d.}}{\sim} \hat{P}_n \tag{D.2}$$

2. **Bootstrapped Estimator Calculation**: Compute the value of the estimator, e.g. R_{bv},

$$R_{bv}^* = R_{bv}(F_1^*, \ldots, F_n^*), \tag{D.3}$$

using the n closure tests within the bootstrap sample.

3. **Repetition**: Repeat steps 1 and 2 for a total of B (we choose $B = 100$) iterations, generating B instances of R_{bv}:

$$R_{bv}^{*,1}, \ldots, R_{bv}^{*,B}. \tag{D.4}$$

4. **Inference**: Compute the mean and the variance of the B calculated R_{bv} values to

estimate the estimator's bootstrap expectation value and uncertainty,

$$\mathbb{E}^*[R^*_{bv}] \approx \frac{1}{B}\sum_{i=1}^{B} R^{*,i}_{bv}, \tag{D.5}$$

$$\mathrm{Var}^*(R^*_{bv}) \approx \frac{1}{B-1}\sum_{i=1}^{B}\left(R^{*,i}_{bv} - \frac{1}{B}\sum_{j=1}^{B} R^{*,j}_{bv}\right)^2. \tag{D.6}$$

m	R_{bv}	ΔR_{bv}
100	0.89	2.3e-2
80	0.89	1.9e-2
60	0.90	2.1e-2

Table D.1: Bootstrapped values of the bias-variance ratio computed on the full DIS dataset. The estimator is computed for different numbers of replicas of each fit. The table shows that the bootstrap result dependence on m is mild and that the result obtained with $m = 60$ is consistent with the others.

The bootstrapping procedure outlined here is used to produce all the uncertainties quoted in the results shown in section 3.2.

APPENDIX E

Correlation between PDFs and observables

To assess the correlation between the *inconsistent* HERA I+II dataset measuring $\sigma_{\rm NC}^{e^+p}$ with $E_p = 920$ GeV and the various PDF flavours, in fig. E.1 we plot the correlation defined in [218]. The correlation function is defined as:

$$\rho(j,x,\mathcal{O}) \equiv \frac{N_{\rm rep}}{N_{\rm rep}-1}\left(\frac{\langle f_j(x,Q)\,\mathcal{O}\rangle_{\rm reps} - \langle f_j(x,Q)\rangle_{\rm reps}\langle\mathcal{O}\rangle_{\rm reps}}{\Delta_{\rm PDF}f(x,Q)\,\Delta_{\rm PDF}\mathcal{O}}\right), \tag{E.1}$$

where the PDFs are evaluated at a given scale $Q = Q_0$ and the observable $\mathcal{O}$ is computed with the set of PDFs f, j is the PDF flavour, $N_{\rm rep}$ is the number of replicas in the baseline PDF set and $\Delta_{\rm PDF}$ are the PDF uncertainties. In the figures we show two Q^2 bins that feature the largest R_{bv} in fig. 3.7, namely $Q^2 = 60$ GeV^2 (left panel) and $Q^2 = 75$ GeV^2 (right panel). We observe that the dataset is mostly correlated with the gluon in the kinematical region in which uncertainties are overestimated as an experimental inconsistency is introduced.

We can look at this same quantity for the other two inconsistent closure test setups, namely the Drell-Yan and JET ones. For the Drell-Yan we show in fig. E.2 the correlation between observables and flavours for two datasets: the one in which which the inconsistency is directly injected during training and the most affected one, which are respectively the ATLAS high-mass Drell-Yan measurements at 8 and 7 TeV. These are the same datasets shown in fig. 3.13 for the single data point analysis.

Finally also for the case of the inconsistent JET multiclosure test we follow the same logic and show the PDF-observable correlation for the datasets shown in fig. 3.19. These are respectively the ATLAS single jet and the CMS $t\bar{t}$ double differential cross section at 8 TeV measurements. As one can see the gluon is the most correlated PDF, and we show the highlighted region in fig. E.3

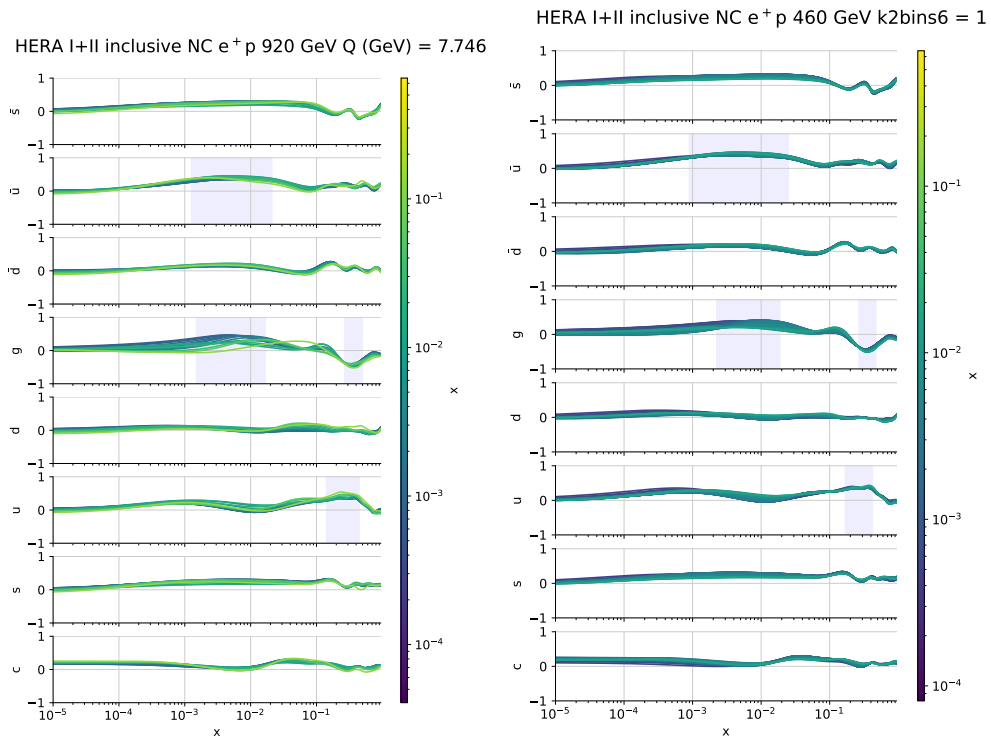


Figure E.1: Correlation coefficient ρ defined in eq. (E.1) between the flavour PDFs of the baseline set at $Q_0 = 1.651$ GeV^2 and the HERA I+II dataset measuring: on the left $\sigma_{\rm NC}^{e^+p}$ with $E_p = 920$ for $Q^2 = 60$ GeV^2, and on the right for $E_p = 460$ and $Q^2 \in (10, 36)$ GeV^2. The highlighted region corresponds to $\rho > 0.7\rho_{\rm max}$.

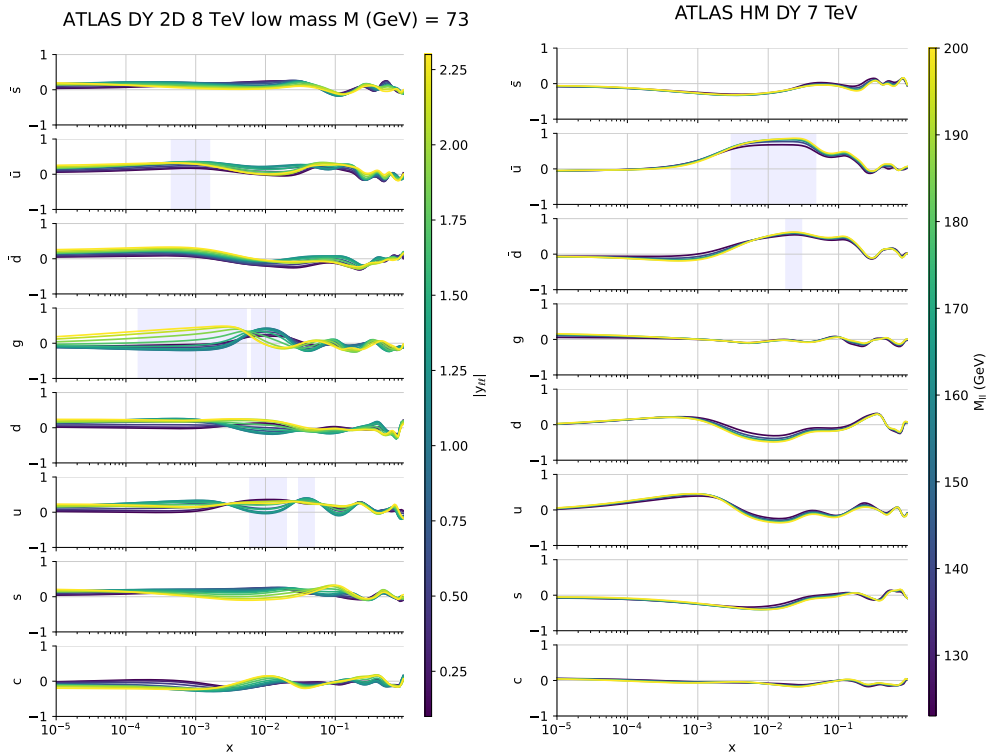


Figure E.2: Correlation coefficient ρ defined between the flavour PDFs of the baseline set at $Q_0 = 1.651$ GeV^2 and the ATLAS high-mass Drell-Yan measurements at 8 (left) and 7 TeV (right). The highlighted region corresponds to $\rho > 0.7\rho_{\max}$.

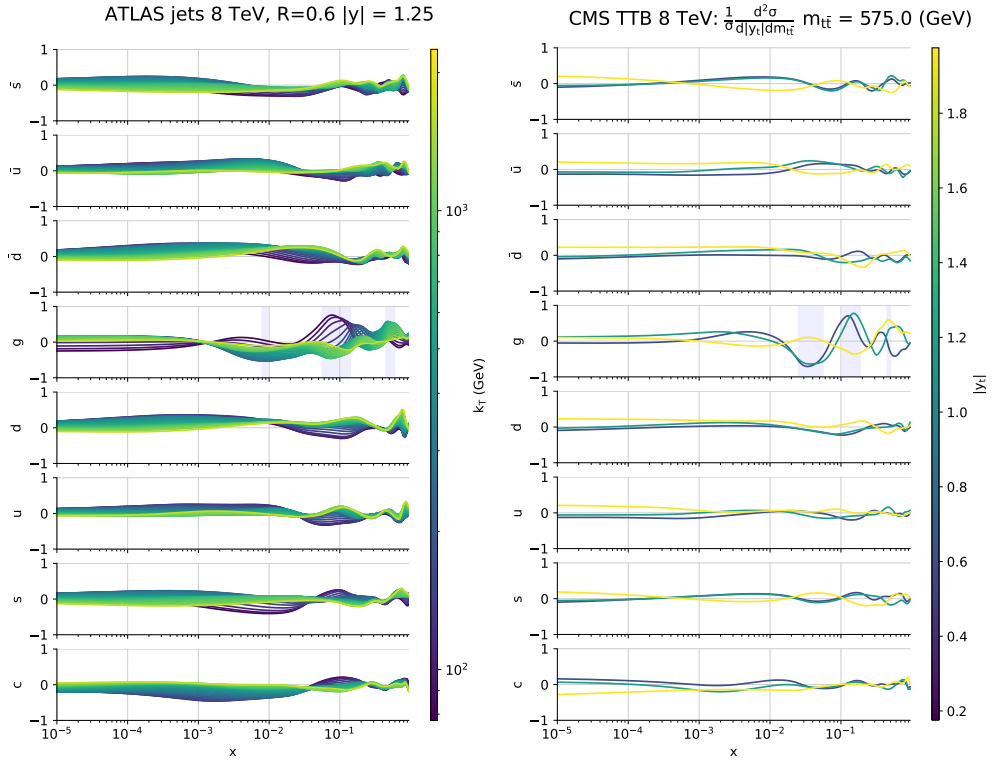


Figure E.3: Correlation coefficient ρ defined between the flavour PDFs of the baseline set at $Q_0 = 1.651$ GeV^2 and the CMS $t\bar{t}$ (left) and the ATLAS single jet (right). The highlighted region corresponds to $\rho > 0.7\rho_{\max}$.

Bibliography

[1] R. K. Ellis, W. J. Stirling, and B. R. Webber. *QCD and Collider Physics*. Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology. Cambridge University Press, 1996.

[2] Gavin P. Salam. *Elements of QCD for hadron colliders*. 2011. arXiv: 1011.5131.

[3] Roy Stegeman. "Statistical Learning for Standard Model Phenomenology". PhD thesis. Milan U., 2022.

[4] Rabah Abdul Khalek et al. "Parton Distributions with Theory Uncertainties: General Formalism and First Phenomenological Studies". In: *Eur. Phys. J. C* 79.11 (2019), p. 931. DOI: 10.1140/epjc/s10052-019-7401-4. arXiv: 1906.10698 [hep-ph].

[5] Richard D. Ball et al. "Determination of the theory uncertainties from missing higher orders on NNLO parton distributions with percent accuracy". In: (Jan. 2024). arXiv: 2401.10319 [hep-ph].

[6] Richard D. Ball et al. "The Path to N^{3}LO Parton Distributions". In: (Feb. 2024). arXiv: 2402.18635 [hep-ph].

[7] Andrea Barontini, Niccolo Laurenti, and Juan Rojo. "NNPDF4.0 aN^{3}LO PDFs with QED corrections". In: *31st International Workshop on Deep-Inelastic Scattering and Related Subjects*. June 2024. arXiv: 2406.01779 [hep-ph].

[8] Richard D. Ball et al. "The path to proton structure at 1% accuracy". In: *Eur. Phys. J. C* 82.5 (2022), p. 428. DOI: 10.1140/epjc/s10052-022-10328-7. arXiv: 2109.02653 [hep-ph].

[9] Luigi Del Debbio, Tommaso Giani, and Michael Wilson. "Bayesian approach to inverse problems: an application to NNPDF closure testing". In: *Eur. Phys. J. C* 82.4 (2022), p. 330. DOI: 10.1140/epjc/s10052-022-10297-x. arXiv: 2111.05787 [hep-ph].

[10] Richard D. Ball et al. "Precision NNLO determination of $\alpha_s(M_Z)$ using an unbiased global parton set". In: *Phys.Lett.* B707 (2012), pp. 66–71. DOI: 10.1016/j.physletb.2011.11.053. arXiv: 1110.2483 [hep-ph].

[11] Andrea Barontini et al. *Evaluating the Faithfulness of PDF uncertainties in the presence of Inconsistent Data, in preparation*. 2024.

[12] Andrea Barontini et al. "Pineline: Industrialization of high-energy theory predictions". In: *Comput. Phys. Commun.* 297 (2024), p. 109061. DOI: 10.1016/j.cpc.2023.109061. arXiv: 2302.12124 [hep-ph].

[13] Richard D. Ball et al. "Photons in the proton: implications for the LHC". In: (Jan. 2024). arXiv: 2401.08749 [hep-ph].

[14] Steven Weinberg. *The Quantum Theory of Fields*. Vol. 2. Cambridge University Press, 1996. DOI: 10.1017/CBO9781139644174.

[15] M. Srednicki. *Quantum field theory*. Cambridge University Press, Jan. 2007. ISBN: 978-0-521-86449-7, 978-0-511-26720-8.

[16] F. Herzog et al. "The five-loop beta function of Yang-Mills theory with fermions". In: *JHEP* 02 (2017), p. 090. DOI: 10.1007/JHEP02(2017)090. arXiv: 1701.01404 [hep-ph].

[17] Particle Data Group et al. "Review of Particle Physics". In: *Progress of Theoretical and Experimental Physics* 2022.8 (Aug. 2022), p. 083C01. ISSN: 2050-3911. DOI: 10.1093/ptep/ptac097. eprint: https://academic.oup.com/ptep/article-pdf/2022/8/083C01/49175539/ptac097.pdf. URL: https://doi.org/10.1093/ptep/ptac097.

[18] Guido Altarelli. *Collider Physics within the Standard Model: a Primer*. 2013. arXiv: 1303.2842v2 [hep-ph].

[19] Guido Altarelli and G. Parisi. "ASYMPTOTIC FREEDOM IN PARTON LANGUAGE". In: *Nucl. Phys.* B126 (1977), p. 298.

[20] L.W. Whitlow et al. "Precise measurements of the proton and deuteron structure functions from a global analysis of the SLAC deep inelastic electron scattering cross sections". In: *Physics Letters B* 282.3 (1992), pp. 475–482. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(92)90672-Q. URL: https://www.sciencedirect.com/science/article/pii/037026939290672Q.

[21] A.C. Benvenuti et al. "A high statistics measurement of the proton structure functions F2(x, Q2) and R from deep inelastic muon scattering at high Q2". In: *Physics Letters B* 223.3 (1989), pp. 485–489. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(89)91637-7. URL: https://www.sciencedirect.com/science/article/pii/0370269389916377.

[22] F.D. Aaron et al. "Measurement of the proton structure function F_L at low x". In: *Physics Letters B* 665.4 (July 2008), pp. 139–146. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2008.05.070. URL: http://dx.doi.org/10.1016/j.physletb.2008.05.070.

[23] H. Abramowicz et al. "Measurement of high-Q^2 neutral current deep inelastic $e + p$ scattering cross sections with a longitudinally polarized positron beam at HERA". In: *Physical Review D* 87.5 (Mar. 2013). ISSN: 1550-2368. DOI: 10.1103/physrevd.87.052014. URL: http://dx.doi.org/10.1103/PhysRevD.87.052014.

[24] G. Curci, W. Furmanski, and R. Petronzio. "Evolution of parton densities beyond leading order". In: *Nuclear Physics B* 175 (1980).

[25] R. K. Ellis et al. "Perturbation theory and the parton model in QCD". In: *Nuclear Physics B* 152 (1978).

[26] R. K. Ellis et al. "Factorization and the parton model in QCD". In: *Physics letters* 78B (1978).

[27] V. N. Gribov and L. N. Lipatov. "Deep inelastic ep scattering in perturbation theory". In: *Sov. J. Nucl. Phys.* 15 (1972), pp. 438–450.

[28] G.P. Salam and J. Rojo. "A Higher Order Perturbative Parton Evolution Toolkit (HOPPET)". In: *Computer Physics Communications* 180.1 (Jan. 2009), pp. 120–156. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2008.08.010. URL: http://dx.doi.org/10.1016/j.cpc.2008.08.010.

[29] M. Botje. "QCDNUM: Fast QCD evolution and convolution". In: *Computer Physics Communications* 182.2 (Feb. 2011), pp. 490–532. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2010.10.020. URL: http://dx.doi.org/10.1016/j.cpc.2010.10.020.

[30] Valerio Bertone, Stefano Carrazza, and Juan Rojo. "APFEL: A PDF evolution library with QED corrections". In: *Computer Physics Communications* 185.6 (June 2014), pp. 1647–1668. ISSN: 0010-4655. DOI: 10.1016/j.cpc.2014.03.007. URL: http://dx.doi.org/10.1016/j.cpc.2014.03.007.

[31] A. Vogt, S. Moch, and J.A.M. Vermaseren. "The three-loop splitting functions in QCD: the singlet case". In: *Nuclear Physics B* 691.1 (2004), pp. 129–181. ISSN: 0550-3213. DOI: https://doi.org/10.1016/j.nuclphysb.2004.04.024. URL: https://www.sciencedirect.com/science/article/pii/S0550321304003074.

[32] Alessandro Candido, Felix Hekhorn, and Giacomo Magni. "EKO: evolution kernel operators". In: *The European Physical Journal C* 82.10 (Oct. 2022). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-022-10878-w. URL: http://dx.doi.org/10.1140/epjc/s10052-022-10878-w.

[33] T. Appelquist and J. Carazzone. "Infrared singularities and massive fields". In: *Physical review D* 11 (May 1975), pp. 2856–2861.

[34] Richard D. Ball, Marco Bonvini, and Luca Rottoli. "Charm in Deep-Inelastic Scattering". In: *JHEP* 11 (2015), p. 122. DOI: 10.1007/JHEP11(2015)122. arXiv: 1510.02491 [hep-ph].

[35] M. Buza et al. "Heavy quark coefficient functions at asymptotic values $Q^2 \gg m^2$". In: *Nuclear Physics B* 472.3 (July 1996), pp. 611–658. ISSN: 0550-3213. DOI: 10.1016/0550-3213(96)00228-3. URL: http://dx.doi.org/10.1016/0550-3213(96)00228-3.

[36] M. Buza et al. "Charm electroproduction viewed in the variable-flavour number scheme versus fixed-order perturbation theory". In: *Eur. Phys. J.* C1 (1998), pp. 301–320. eprint: hep-ph/9612398.

[37] J. Collins, F. Wilczek, and A. Zee. "Low-energy manifestations of heavy particles: Application to the neutral current". In: *Phys. Rev. D* 18 (1 July 1978), pp. 242–247. DOI: 10.1103/PhysRevD.18.242. URL: https://link.aps.org/doi/10.1103/PhysRevD.18.242.

[38] J. C. Collins. "Hard-scattering factorization with heavy quarks: A general treatment". In: *Phys. Rev. D* 58 (9 Sept. 1998), p. 094002. DOI: 10.1103/PhysRevD.58.094002. URL: https://link.aps.org/doi/10.1103/PhysRevD.58.094002.

[39] Michael Krämer, Fredrick I. Olness, and Davison E. Soper. "Treatment of heavy quarks in deeply inelastic scattering". In: *Phys. Rev. D* 62 (9 Oct. 2000), p. 096007. DOI: 10.1103/PhysRevD.62.096007. URL: https://link.aps.org/doi/10.1103/PhysRevD.62.096007.

[40] R.S. Thorne and R.G. Roberts. "A practical procedure for evolving heavy flavour structure functions". In: *Physics Letters B* 421.1 (1998), pp. 303–311. ISSN: 0370-2693. DOI: https://doi.org/10.1016/S0370-2693(97)01580-3. URL: https://www.sciencedirect.com/science/article/pii/S0370269397015803.

[41] R. S. Thorne. "Variable-flavor number scheme for next-to-next-to-leading order". In: *Physical Review D* 73.5 (Mar. 2006). ISSN: 1550-2368. DOI: 10.1103/physrevd.73.054019. URL: http://dx.doi.org/10.1103/PhysRevD.73.054019.

[42] Stefano Forte et al. "Heavy quarks in deep-inelastic scattering". In: *Nucl. Phys. B* 834 (2010), pp. 116–162. DOI: 10.1016/j.nuclphysb.2010.03.014. arXiv: 1001.2312 [hep-ph].

[43] Marco Bonvini, Andrew Papanastasiou, and Frank Tackmann. "Matched predictions for the b b ¯ H $b\bar{b}H$ cross section at the 13 TeV LHC". In: *Journal of High Energy Physics* 2016 (Oct. 2016). DOI: 10.1007/JHEP10(2016)053.

[44] The NNPDF Collaboration et al. *Response to "Parton distributions need representative sampling"*. 2022. arXiv: 2211.12961.

[45] Tie-Jiun Hou et al. "New CTEQ global analysis of quantum chromodynamics with high-precision data from the LHC". In: *Physical Review D* 103.1 (Jan. 2021). ISSN: 2470-0029. DOI: 10.1103/physrevd.103.014013. URL: http://dx.doi.org/10.1103/PhysRevD.103.014013.

[46] S. Bailey et al. "Parton distributions from LHC, HERA, Tevatron and fixed target data: MSHT20 PDFs". In: *The European Physical Journal C* 81.4 (Apr. 2021). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-021-09057-0. URL: http://dx.doi.org/10.1140/epjc/s10052-021-09057-0.

[47] S. Alekhin et al. "Parton distribution functions, α_s, and heavy-quark masses for LHC Run II". In: *Physical Review D* 96.1 (July 2017). ISSN: 2470-0029. DOI: 10.1103/physrevd.96.014011. URL: http://dx.doi.org/10.1103/PhysRevD.96.014011.

[48] Federico Demartin et al. "Impact of parton distribution function and α_s uncertainties on Higgs boson production in gluon fusion at hadron colliders". In: *Physical Review D* 82.1 (July 2010). ISSN: 1550-2368. DOI: 10.1103/physrevd.82.014002. URL: http://dx.doi.org/10.1103/PhysRevD.82.014002.

[49] Richard D. Ball, Emanuele R. Nocera, and Rosalyn L. Pearson. "Nuclear uncertainties in the determination of proton PDFs: NNPDF Collaboration". In: *The European Physical Journal C* 79.3 (Mar. 2019). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-019-6793-5. URL: http://dx.doi.org/10.1140/epjc/s10052-019-6793-5.

[50] Richard D. Ball, Emanuele R. Nocera, and Rosalyn L. Pearson. "Deuteron uncertainties in the determination of proton PDFs". In: *The European Physical Journal C* 81.1 (Jan. 2021). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-020-08826-7. URL: http://dx.doi.org/10.1140/epjc/s10052-020-08826-7.

[51] Julien Baglio et al. "Inclusive production cross sections at N3LO". In: *Journal of High Energy Physics* 2022.12 (Dec. 2022). ISSN: 1029-8479. DOI: 10.1007/jhep12(2022)066. URL: http://dx.doi.org/10.1007/JHEP12(2022)066.

[52] Richard D. Ball et al. "An open-source machine learning framework for global analyses of parton distributions". In: *Eur. Phys. J. C* 81.10 (2021), p. 958. DOI: 10.1140/epjc/s10052-021-09747-9. arXiv: 2109.02671 [hep-ph].

[53] The NNPDF Collaboration et al. *A First Determination of Parton Distributions with Theoretical Uncertainties*. 2019. arXiv: 1905.04311.

[54] D. Stump et al. "Uncertainties of predictions from parton distribution functions. I. The Lagrange multiplier method". In: *Physical Review D* 65.1 (Dec. 2001). ISSN: 1089-4918. DOI: 10.1103/physrevd.65.014012. URL: http://dx.doi.org/10.1103/PhysRevD.65.014012.

[55] J. Pumplin et al. "Uncertainties of predictions from parton distribution functions. II. The Hessian method". In: *Physical Review D* 65.1 (Dec. 2001). ISSN: 1089-4918. DOI: 10.1103/physrevd.65.014013. URL: http://dx.doi.org/10.1103/PhysRevD.65.014013.

[56] Mark N. Costantini et al. "A critical study of the Monte Carlo replica method". In: (Apr. 2024). arXiv: 2404.10056 [hep-ph].

[57] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5 (1989), pp. 359–366. ISSN: 0893-6080. DOI: https://doi.org/10.1016/0893-6080(89)90020-8. URL: https://www.sciencedirect.com/science/article/pii/0893608089900208.

[58] Stefano Forte et al. "Neural network parametrization of deep-inelastic structure functions". In: *Journal of High Energy Physics* 2002.05 (May 2002), pp. 062–062. ISSN: 1029-8479. DOI: 10.1088/1126-6708/2002/05/062. URL: http://dx.doi.org/10.1088/1126-6708/2002/05/062.

[59] Stefano Carrazza and Juan Cruz-Martinez. "Towards a new generation of parton densities with deep learning models". In: *Eur. Phys. J. C* 79.8 (2019), p. 676. DOI: 10.1140/epjc/s10052-019-7197-2. arXiv: 1907.05075 [hep-ph].

[60] Richard D. Ball et al. "Precision determination of electroweak parameters and the strange content of the proton from neutrino deep-inelastic scattering". In: *Nuclear Physics B* 823.1–2 (Dec. 2009), pp. 195–233. ISSN: 0550-3213. DOI: 10.1016/j.nuclphysb.2009.08.003. URL: http://dx.doi.org/10.1016/j.nuclphysb.2009.08.003.

[61] Richard D. Ball et al. "Parton distributions for the LHC run II". In: *Journal of High Energy Physics* 2015.4 (Apr. 2015). ISSN: 1029-8479. DOI: 10.1007/jhep04(2015)040. URL: http://dx.doi.org/10.1007/JHEP04(2015)040.

[62] Andy Buckley et al. "LHAPDF6: parton density access in the LHC precision era". In: *The European Physical Journal C* 75.3 (Mar. 2015). ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-015-3318-8. URL: http://dx.doi.org/10.1140/epjc/s10052-015-3318-8.

[63] Giulio D'Agostini. *Bayesian Reasoning in Data Analysis: A Critical Introduction*. Singapore: World Scientific, 2003. DOI: 10.1142/5262. URL: https://cds.cern.ch/record/642515.

[64] Richard D. Ball et al. "Fitting parton distribution data with multiplicative normalization uncertainties". In: *Journal of High Energy Physics* 2010.5 (May 2010). ISSN: 1029-8479. DOI: 10.1007/jhep05(2010)075. URL: http://dx.doi.org/10.1007/JHEP05(2010)075.

[65] Alessandro Candido, Stefano Forte, and Felix Hekhorn. "Can

$$\overline{\text{MS}}$$

parton distributions be negative?" In: *Journal of High Energy Physics* 2020.11 (Nov. 2020). ISSN: 1029-8479. DOI: 10.1007/jhep11(2020)129. URL: http://dx.doi.org/10.1007/JHEP11(2020)129.

[66] Richard D. Ball and A. Deshpande. *The Proton Spin, Semi-Inclusive processes, and a future Electron Ion Collider*. 2018. arXiv: 1801.04842 [hep-ph]. URL: https://arxiv.org/abs/1801.04842.

[67] Matteo Cacciari and Nicolas Houdeau. "Meaningful characterisation of perturbative theoretical uncertainties". In: *Journal of High Energy Physics* 2011.9 (Sept. 2011). ISSN: 1029-8479. DOI: 10.1007/jhep09(2011)039. URL: http://dx.doi.org/10.1007/JHEP09(2011)039.

[68] André David and Giampiero Passarino. "How well can we guess theoretical uncertainties?" In: *Physics Letters B* 726.1–3 (Oct. 2013), pp. 266–272. ISSN: 0370-2693. DOI: 10.1016/j.physletb.2013.08.025. URL: http://dx.doi.org/10.1016/j.physletb.2013.08.025.

[69] Emanuele Bagnaschi et al. "An extensive survey of the estimation of uncertainties from missing higher orders in perturbative calculations". In: *Journal of High Energy Physics* 2015.2 (Feb. 2015). ISSN: 1029-8479. DOI: 10.1007/jhep02(2015)133. URL: http://dx.doi.org/10.1007/JHEP02(2015)133.

[70] Marco Bonvini. "Probabilistic definition of the perturbative theoretical uncertainty from missing higher orders". In: *Eur. Phys. J. C* 80.10 (2020), p. 989. DOI: 10.1140/epjc/s10052-020-08545-z. arXiv: 2006.16293 [hep-ph].

[71] Richard D. Ball et al. "Evidence for intrinsic charm quarks in the proton". In: *Nature* 608.7923 (Aug. 2022), pp. 483–487. ISSN: 1476-4687. DOI: 10.1038/s41586-022-04998-2. URL: http://dx.doi.org/10.1038/s41586-022-04998-2.

[72] Richard D. Ball et al. *The intrinsic charm quark valence distribution of the proton*. 2024. arXiv: 2311.00743 [hep-ph]. URL: https://arxiv.org/abs/2311.00743.

[73] J. A. M. Vermaseren, A. Vogt, and S. Moch. "The third-order QCD corrections to deep-inelastic scattering by photon exchange". In: *Nucl. Phys.* B724 (2005), p. 3. DOI: 10.1016/j.nuclphysb.2005.06.020. arXiv: hep-ph/0504242.

[74] S. Moch, J. A. M. Vermaseren, and A. Vogt. "The longitudinal structure function at the third order". In: *Phys. Lett.* B606 (2005), p. 123. DOI: 10.1016/j.physletb.2004.11.063. arXiv: hep-ph/0411112.

[75] S. Moch, M. Rogal, and A. Vogt. "Differences between charged-current coefficient functions". In: *Nucl. Phys.* B790 (2008), pp. 317–335. DOI: 10.1016/j.nuclphysb.2007.09.022. arXiv: 0708.3731 [hep-ph].

[76] S. Moch, J. A. M. Vermaseren, and A. Vogt. "Third-order QCD corrections to the charged-current structure function F(3)". In: *Nucl. Phys. B* 813 (2009), pp. 220–258. DOI: 10.1016/j.nuclphysb.2009.01.001. arXiv: 0812.4168 [hep-ph].

[77] Charalampos Anastasiou et al. "Higgs Boson Gluon-Fusion Production in QCD at Three Loops". In: *Phys. Rev. Lett.* 114.21 (2015), p. 212001. DOI: 10.1103/PhysRevLett.114.212001. arXiv: 1503.06056 [hep-ph].

[78] Bernhard Mistlberger. "Higgs boson production at hadron colliders at N^3LO in QCD". In: *JHEP* 05 (2018), p. 028. DOI: 10.1007/JHEP05(2018)028. arXiv: 1802.00833 [hep-ph].

[79] Claude Duhr, Falko Dulat, and Bernhard Mistlberger. "Higgs Boson Production in Bottom-Quark Fusion to Third Order in the Strong Coupling". In: *Phys. Rev. Lett.* 125.5 (2020), p. 051804. DOI: 10.1103/PhysRevLett.125.051804. arXiv: 1904.09990 [hep-ph].

[80] Julien Baglio et al. "Inclusive production cross sections at N^3LO". In: *JHEP* 12 (2022), p. 066. DOI: 10.1007/JHEP12(2022)066. arXiv: 2209.06138 [hep-ph].

[81] Frédéric A. Dreyer and Alexander Karlberg. "Vector-Boson Fusion Higgs Production at Three Loops in QCD". In: *Phys. Rev. Lett.* 117.7 (2016), p. 072001. DOI: 10.1103/PhysRevLett.117.072001. arXiv: 1606.00840 [hep-ph].

[82] Long-Bin Chen et al. "Higgs boson pair production via gluon fusion at N^3LO in QCD". In: *Phys. Lett. B* 803 (2020), p. 135292. DOI: 10.1016/j.physletb.2020.135292. arXiv: 1909.06808 [hep-ph].

[83] Claude Duhr, Falko Dulat, and Bernhard Mistlberger. "Charged current Drell-Yan production at N^3LO". In: *JHEP* 11 (2020), p. 143. DOI: 10.1007/JHEP11(2020)143. arXiv: 2007.13313 [hep-ph].

[84] Claude Duhr and Bernhard Mistlberger. "Lepton-pair production at hadron colliders at N^3LO in QCD". In: *JHEP* 03 (2022), p. 116. DOI: 10.1007/JHEP03(2022)116. arXiv: 2111.10379 [hep-ph].

[85] Falko Dulat, Bernhard Mistlberger, and Andrea Pelloni. "Differential Higgs production at N^3LO beyond threshold". In: *JHEP* 01 (2018), p. 145. DOI: 10.1007/JHEP01(2018)145. arXiv: 1710.03016 [hep-ph].

[86] Falko Dulat, Bernhard Mistlberger, and Andrea Pelloni. "Precision predictions at N^3LO for the Higgs boson rapidity distribution at the LHC". In: *Phys. Rev. D* 99.3 (2019), p. 034004. DOI: 10.1103/PhysRevD.99.034004. arXiv: 1810.09462 [hep-ph].

[87] X. Chen et al. "Fully Differential Higgs Boson Production to Third Order in QCD". In: *Phys. Rev. Lett.* 127.7 (2021), p. 072002. DOI: 10.1103/PhysRevLett.127.072002. arXiv: 2102.07607 [hep-ph].

[88] Georgios Billis et al. "Higgs pT Spectrum and Total Cross Section with Fiducial Cuts at Third Resummed and Fixed Order in QCD". In: *Phys. Rev. Lett.* 127.7 (2021), p. 072001. DOI: 10.1103/PhysRevLett.127.072001. arXiv: 2102.08039 [hep-ph].

[89] Stefano Camarda, Leandro Cieri, and Giancarlo Ferrera. "Drell–Yan lepton-pair production: qT resummation at N3LL accuracy and fiducial cross sections at N3LO". In: *Phys. Rev. D* 104.11 (2021), p. L111503. DOI: 10.1103/PhysRevD.104.L111503. arXiv: 2103.04974 [hep-ph].

[90] Xuan Chen et al. "Dilepton Rapidity Distribution in Drell-Yan Production to Third Order in QCD". In: *Phys. Rev. Lett.* 128.5 (2022), p. 052001. DOI: 10.1103/PhysRevLett.128.052001. arXiv: 2107.09085 [hep-ph].

[91] Xuan Chen et al. "Transverse mass distribution and charge asymmetry in W boson production to third order in QCD". In: *Phys. Lett. B* 840 (2023), p. 137876. DOI: 10.1016/j.physletb.2023.137876. arXiv: 2205.11426 [hep-ph].

[92] Fabrizio Caola et al. "The Path forward to N^{3}LO". In: *Snowmass 2021*. Mar. 2022. arXiv: 2203.06730 [hep-ph].

[93] J. Davies et al. "Large-nf contributions to the four-loop splitting functions in QCD". In: *Nucl. Phys. B* 915 (2017), pp. 335–362. DOI: 10.1016/j.nuclphysb.2016.12.012. arXiv: 1610.07477 [hep-ph].

[94] S. Moch et al. "Four-Loop Non-Singlet Splitting Functions in the Planar Limit and Beyond". In: *JHEP* 10 (2017), p. 041. DOI: 10.1007/JHEP10(2017)041. arXiv: 1707.08315 [hep-ph].

[95] J. Davies et al. "Resummation of small-x double logarithms in QCD: inclusive deep-inelastic scattering". In: *JHEP* 08 (2022), p. 135. DOI: 10.1007/JHEP08(2022)135. arXiv: 2202.10362 [hep-ph].

[96] Johannes M. Henn, Gregory P. Korchemsky, and Bernhard Mistlberger. "The full four-loop cusp anomalous dimension in $\mathcal{N} = 4$ super Yang-Mills and QCD". In: *JHEP* 04 (2020), p. 018. DOI: 10.1007/JHEP04(2020)018. arXiv: 1911.10174 [hep-th].

[97] Claude Duhr, Bernhard Mistlberger, and Gherardo Vita. "Soft integrals and soft anomalous dimensions at N^{3}LO and beyond". In: *JHEP* 09 (2022), p. 155. DOI: 10.1007/JHEP09(2022)155. arXiv: 2205.04493 [hep-ph].

[98] S. Moch et al. "Low moments of the four-loop splitting functions in QCD". In: *Phys. Lett. B* 825 (2022), p. 136853. DOI: 10.1016/j.physletb.2021.136853. arXiv: 2111.15561 [hep-ph].

[99] G. Soar et al. "On Higgs-exchange DIS, physical evolution kernels and fourth-order splitting functions at large x". In: *Nucl. Phys. B* 832 (2010), pp. 152–227. DOI: 10.1016/j.nuclphysb.2010.02.003. arXiv: 0912.0369 [hep-ph].

[100] G. Falcioni et al. "Four-loop splitting functions in QCD – The quark-quark case". In: (Feb. 2023). arXiv: 2302.07593 [hep-ph].

[101] G. Falcioni et al. "Four-loop splitting functions in QCD – The gluon-to-quark case". In: *Phys. Lett. B* 846 (2023), p. 138215. DOI: 10.1016/j.physletb.2023.138215. arXiv: 2307.04158 [hep-ph].

[102] S. Moch et al. "Additional moments and x-space approximations of four-loop splitting functions in QCD". In: *Phys. Lett. B* 849 (2024), p. 138468. DOI: 10.1016/j.physletb.2024.138468. arXiv: 2310.05744 [hep-ph].

[103] G. Falcioni et al. "The double fermionic contribution to the four-loop quark-to-gluon splitting function". In: *Phys. Lett. B* 848 (2024), p. 138351. DOI: 10.1016/j.physletb.2023.138351. arXiv: 2310.01245 [hep-ph].

[104] J. McGowan et al. "Approximate N^{3}LO parton distribution functions with theoretical uncertainties: MSHT20aN^{3}LO PDFs". In: *Eur. Phys. J. C* 83.3 (2023). [Erratum: Eur.Phys.J.C 83, 302 (2023)], p. 185. DOI: 10.1140/epjc/s10052-023-11236-0. arXiv: 2207.04739 [hep-ph].

[105] W. L. van Neerven and A. Vogt. "Improved approximations for the three loop splitting functions in QCD". In: *Phys. Lett. B* 490 (2000), pp. 111–118. DOI: 10.1016/S0370-2693(00)00953-9. arXiv: hep-ph/0007362.

[106] G. D'Agostini. "On the use of the covariance matrix to fit correlated data". In: *Nucl.Instrum.Meth.* A346 (1994), pp. 306–311. DOI: 10.1016/0168-9002(94)90719-6.

[107] Richard D. Ball et al. "Parton distributions with small-x resummation: evidence for BFKL dynamics in HERA data". In: *Eur. Phys. J.* C78.4 (2018), p. 321. DOI: 10.1140/epjc/s10052-018-5774-4. arXiv: 1710.05935 [hep-ph].

[108] M. L. Mangano et al. "Physics at a 100 TeV pp collider: Standard Model processes". In: (2016). arXiv: 1607.01831 [hep-ph].

[109] L. A. Harland-Lang and R. S. Thorne. "On the Consistent Use of Scale Variations in PDF Fits and Predictions". In: *Eur. Phys. J.* C79.3 (2019), p. 225. DOI: 10.1140/epjc/s10052-019-6731-6. arXiv: 1811.08434 [hep-ph].

[110] Richard D. Ball and Rosalyn L. Pearson. "Correlation of theoretical uncertainties in PDF fits and theoretical uncertainties in predictions". In: *Eur. Phys. J. C* 81.9 (2021), p. 830. DOI: 10.1140/epjc/s10052-021-09602-x. arXiv: 2105.05114 [hep-ph].

[111] Zahari Kassabov, Maria Ubiali, and Cameron Voisey. "Parton distributions with scale uncertainties: a Monte Carlo sampling approach". In: *JHEP* 03 (2023), p. 148. DOI: 10.1007/JHEP03(2023)148. arXiv: 2207.07616 [hep-ph].

[112] Richard D. Ball et al. "Parton distributions and new physics searches: the Drell–Yan forward–backward asymmetry as a case study". In: *Eur. Phys. J. C* 82.12 (2022), p. 1160. DOI: 10.1140/epjc/s10052-022-11133-y. arXiv: 2209.08115 [hep-ph].

[113] Marco Bonvini et al. "Updated Higgs cross section at approximate N^3LO". In: *J. Phys.* G41 (2014), p. 095002. DOI: 10.1088/0954-3899/41/9/095002. arXiv: 1404.3204 [hep-ph].

[114] Frédéric A. Dreyer and Alexander Karlberg. "Vector-Boson Fusion Higgs Pair Production at N^3LO". In: *Phys. Rev. D* 98.11 (2018), p. 114016. DOI: 10.1103/PhysRevD.98.114016. arXiv: 1811.07906 [hep-ph].

[115] S. Amoroso et al. "Snowmass 2021 Whitepaper: Proton Structure at the Precision Frontier". In: *Acta Phys. Polon. B* 53.12 (2022), 12–A1. DOI: 10.5506/APhysPolB.53.12-A1. arXiv: 2203.13923 [hep-ph].

[116] Maria Ubiali. "Parton Distribution Functions and Their Impact on Precision of the Current Theory Calculations". In: Apr. 2024. arXiv: 2404.08508 [hep-ph].

[117] Dan Guest, Kyle Cranmer, and Daniel Whiteson. "Deep Learning and its Application to LHC Physics". In: *Ann. Rev. Nucl. Part. Sci.* 68 (2018), pp. 161–181. DOI: 10.1146/annurev-nucl-101917-021019. arXiv: 1806.11484 [hep-ex].

[118] Kim Albertsson et al. "Machine Learning in High Energy Physics Community White Paper". In: *J. Phys. Conf. Ser.* 1085.2 (2018), p. 022008. DOI: 10.1088/1742-6596/1085/2/022008. arXiv: 1807.02876 [physics.comp-ph].

[119] Richard D. Ball et al. "Parton distributions for the LHC Run II". In: *JHEP* 04 (2015), p. 040. DOI: 10.1007/JHEP04(2015)040. arXiv: 1410.8849 [hep-ph].

[120] 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". In: ed. by Harrison B. Prosper and Louis Lyons. 2011. Chap. Open Issues in the Wake of Banff 2011. DOI: 10.5170/CERN-2011-006.

[121] G. Watt and R. S. Thorne. "Study of Monte Carlo approach to experimental uncertainty propagation with MSTW 2008 PDFs". In: *JHEP* 1208 (2012), p. 052. DOI: 10.1007/JHEP08(2012)052. arXiv: 1205.4024 [hep-ph].

[122] Richard D. Ball et al. "Fitting Parton Distribution Data with Multiplicative Normalization Uncertainties". In: *JHEP* 05 (2010), p. 075. DOI: 10.1007/JHEP05(2010)075. arXiv: 0912.2276 [hep-ph].

[123] M. Arneodo et al. "Accurate measurement of F_2^d/F_2^p and $R_d - R_p$". In: *Nucl. Phys.* B487 (1997), pp. 3–26. DOI: 10.1016/S0550-3213(96)00673-6. arXiv: hep-ex/9611022.

[124] 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". In: *Nucl. Phys.* B483 (1997), pp. 3–43. DOI: 10.1016/S0550-3213(96)00538-X. arXiv: hep-ph/9610231.

[125] L. W. Whitlow et al. "Precise measurements of the proton and deuteron structure functions from a global analysis of the SLAC deep inelastic electron scattering cross-sections". In: *Phys. Lett.* B282 (1992), pp. 475–482. DOI: 10.1016/0370-2693(92)90672-Q.

[126] 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". In: *Phys. Lett.* B223 (1989), p. 485. DOI: 10.1016/0370-2693(89)91637-7.

[127] G. Onengut et al. "Measurement of nucleon structure functions in neutrino scattering". In: *Phys. Lett.* B632 (2006), pp. 65–75. DOI: 10.1016/j.physletb.2005.10.062.

[128] 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". In: *Phys. Rev.* D64 (2001), p. 112006. DOI: 10.1103/PhysRevD.64.112006. arXiv: hep-ex/0102049.

[129] David Alexander Mason. "Measurement of the strange - antistrange asymmetry at NLO in QCD from NuTeV dimuon data". In: (). FERMILAB-THESIS-2006-01. DOI: 10.1103/PhysRevLett.99.192001.

[130] H. Abramowicz et al. "Combination of measurements of inclusive deep inelastic $e^{\pm}p$ scattering cross sections and QCD analysis of HERA data". In: *Eur. Phys. J.* C75.12 (2015), p. 580. DOI: 10.1140/epjc/s10052-015-3710-4. arXiv: 1506.06042 [hep-ex].

[131] H. Abramowicz et al. "Combination and QCD analysis of charm and beauty production cross-section measurements in deep inelastic ep scattering at HERA". In: *Eur. Phys. J. C* 78.6 (2018), p. 473. DOI: 10.1140/epjc/s10052-018-5848-3. arXiv: 1804.01019 [hep-ex].

[132] Richard D. Ball et al. "Parton distributions from high-precision collider data". In: *Eur. Phys. J.* C77.10 (2017), p. 663. DOI: 10.1140/epjc/s10052-017-5199-5. arXiv: 1706.00428 [hep-ph].

[133] Georges 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". In: *JHEP* 08 (2016), p. 009. DOI: 10.1007/JHEP08(2016)009. arXiv: 1606.01736 [hep-ex].

[134] Stefano Forte and Stefano Carrazza. "Parton distribution functions". In: (Aug. 2020). arXiv: 2008.12305 [hep-ph].

[135] Georges Aad et al. "A precise measurement of the Z-boson double-differential transverse momentum and rapidity distributions in the full phase space of the decay leptons with the ATLAS experiment at $\sqrt{s} = 8$ TeV". In: *Eur. Phys. J. C* 84.3 (2024), p. 315. DOI: 10.1140/epjc/s10052-024-12438-w. arXiv: 2309.09318 [hep-ex].

[136] Roel Aaij et al. "Measurement of the forward Z boson production cross-section in pp collisions at $\sqrt{s} = 7$ TeV". In: *JHEP* 08 (2015), p. 039. DOI: 10.1007/JHEP08(2015)039. arXiv: 1505.07024 [hep-ex].

[137] Roel Aaij et al. "Measurement of the forward Z boson production cross-section in pp collisions at $\sqrt{s} = 13$ TeV". In: *JHEP* 09 (2016), p. 136. DOI: 10.1007/JHEP09(2016)136. arXiv: 1607.06495 [hep-ex].

[138] Morad 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". In: *JHEP* 05 (2018). [Erratum: JHEP 10, 048 (2020)], p. 077. DOI: 10.1007/JHEP05(2018)077. arXiv: 1711.03296 [hep-ex].

[139] Serguei Chatrchyan et al. "Measurement of the Muon Charge Asymmetry in Inclusive $pp \to W + X$ Production at $\sqrt{s} = 7$ TeV and an Improved Determination of Light Parton Distribution Functions". In: *Phys. Rev. D* 90.3 (2014), p. 032004. DOI: 10.1103/PhysRevD.90.032004. arXiv: 1312.6283 [hep-ex].

[140] Georges 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". In: *Eur. Phys. J. C* 74.10 (2014). [Addendum: Eur.Phys.J.C 76, 642 (2016)], p. 3109. DOI: 10.1140/epjc/s10052-016-4501-2. arXiv: 1406.5375 [hep-ex].

[141] Georges 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". In: *Phys. Lett. B* 725 (2013), pp. 223–242. DOI: 10.1016/j.physletb.2013.07.049. arXiv: 1305.4192 [hep-ex].

[142] Morad 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". In: *Eur. Phys. J. C* 77.8 (2017), p. 531. DOI: 10.1140/epjc/s10052-017-5061-9. arXiv: 1702.02859 [hep-ex].

[143] 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". In: *JHEP* 03 (2018), p. 115. DOI: 10.1007/JHEP03(2018)115. arXiv: 1711.03143 [hep-ex].

[144] Albert 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". In: *Eur. Phys. J. C* 77.7 (2017), p. 459. DOI: 10.1140/epjc/s10052-017-4984-5. arXiv: 1703.01630 [hep-ex].

[145] Vardan 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". In: *JHEP* 03 (2017), p. 156. DOI: 10.1007/JHEP03(2017)156. arXiv: 1609.05331 [hep-ex].

[146] Morad Aaboud et al. "Measurement of the inclusive jet cross-sections in proton-proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector". In: *JHEP* 09 (2017), p. 020. DOI: 10.1007/JHEP09(2017)020. arXiv: 1706.03192 [hep-ex].

[147] Vardan 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". In: *JHEP* 03 (2017), p. 156. DOI: 10.1007/JHEP03(2017)156. arXiv: 1609.05331 [hep-ex].

[148] Morad Aaboud et al. "Precision measurement and interpretation of inclusive W^+, W^- and Z/γ^* production cross sections with the ATLAS detector". In: *Eur. Phys. J. C* 77.6 (2017), p. 367. DOI: 10.1140/epjc/s10052-017-4911-9. arXiv: 1612.03016 [hep-ex].

[149] Timo Antero 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". In: *Phys. Lett. B* 692 (2010), pp. 232–239. DOI: 10.1016/j.physletb.2010.06.043. arXiv: 0908.3914 [hep-ex].

[150] J. de Blas et al. "Electroweak precision observables and Higgs-boson signal strengths in the Standard Model and beyond: present and future". In: *Journal of High Energy Physics* 2016.12 (Dec. 2016). ISSN: 1029-8479. DOI: 10.1007/jhep12(2016)135. URL: http://dx.doi.org/10.1007/JHEP12(2016)135.

[151] Richard D. Ball et al. "Precision determination of the strong coupling constant within a global PDF analysis". In: *Eur. Phys. J. C* 78.5 (2018), p. 408. DOI: 10.1140/epjc/s10052-018-5897-7. arXiv: 1802.03398 [hep-ph].

[152] A.D. Martin, W.J. Stirling, and R.G. Roberts. "The α_s dependence of parton distributions". In: *Physics Letters B* 356.1 (Aug. 1995), pp. 89–94. ISSN: 0370-2693. DOI: 10.1016/0370-2693(95)00808-x. URL: http://dx.doi.org/10.1016/0370-2693(95)00808-X.

[153] Particle Data Group et al. "Review of Particle Physics". In: *Progress of Theoretical and Experimental Physics* 2022.8 (Aug. 2022), p. 083C01. ISSN: 2050-3911. DOI: 10.1093/ptep/ptac097. eprint: https://academic.oup.com/ptep/article-pdf/2022/8/083C01/49175539/ptac097.pdf. URL: https://doi.org/10.1093/ptep/ptac097.

[154] Stefano Forte and Zahari Kassabov. "Why α_s cannot be determined from hadronic processes without simultaneously determining the parton distributions". In: *Eur. Phys. J. C* 80.3 (2020), p. 182. DOI: 10.1140/epjc/s10052-020-7748-6. arXiv: 2001.04986 [hep-ph].

[155] S. Bailey et al. "Parton distributions from LHC, HERA, Tevatron and fixed target data: MSHT20 PDFs". In: *Eur. Phys. J. C* 81.4 (2021), p. 341. DOI: 10.1140/epjc/s10052-021-09057-0. arXiv: 2012.04684 [hep-ph].

[156] Tie-Jiun Hou et al. "New CTEQ global analysis of quantum chromodynamics with high-precision data from the LHC". In: *Phys. Rev. D* 103.1 (2021), p. 014013. DOI: 10.1103/PhysRevD.103.014013. arXiv: 1912.10053 [hep-ph].

[157] S. Alekhin et al. "Parton distribution functions, α_s, and heavy-quark masses for LHC Run II". In: *Phys. Rev.* D96.1 (2017), p. 014011. DOI: 10.1103/PhysRevD.96.014011. arXiv: 1701.05838 [hep-ph].

[158] Andrea Barontini et al. "Theory pipeline for PDF fitting". In: *PoS* ICHEP2022 (2022), p. 784. DOI: 10.22323/1.414.0784. arXiv: 2211.10447 [hep-ph].

[159] Andrea Barontini et al. "Theory prediction in PDF fitting". In: *21th International Workshop on Advanced Computing and Analysis Techniques in Physics Research: AI meets Reality*. Mar. 2023. arXiv: 2303.07119 [hep-ph].

[160] Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1 (Mar. 2016), p. 160018. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. URL: https://doi.org/10.1038/sdata.2016.18.

[161] Valerio Bertone, Stefano Carrazza, and Juan Rojo. "APFEL: A PDF Evolution Library with QED corrections". In: *Comput.Phys.Commun.* 185 (2014), p. 1647. DOI: 10.1016/j.cpc.2014.03.007. arXiv: 1310.1394 [hep-ph].

[162] Stefano Catani and Massimiliano Grazzini. "An NNLO subtraction formalism in hadron collisions and its application to Higgs boson production at the LHC". In: *Phys. Rev. Lett.* 98 (2007), p. 222002. DOI: 10.1103/PhysRevLett.98.222002. arXiv: hep-ph/0703012.

[163] Stefano Catani et al. "Vector boson production at hadron colliders: a fully exclusive QCD calculation at NNLO". In: *Phys. Rev. Lett.* 103 (2009), p. 082001. DOI: 10.1103/PhysRevLett.103.082001. arXiv: 0903.2120 [hep-ph].

[164] Ryan Gavin et al. "FEWZ 2.0: A code for hadronic Z production at next-to-next-to-leading order". In: *Comput. Phys. Commun.* 182 (2011), pp. 2388–2403. DOI: 10.1016/j.cpc.2011.06.008. arXiv: 1011.3540 [hep-ph].

[165] Ryan Gavin et al. "W Physics at the LHC with FEWZ 2.1". In: *Comput.Phys.Commun.* 184 (2013), pp. 208–214. DOI: 10.1016/j.cpc.2012.09.005. arXiv: 1201.5896 [hep-ph].

[166] Ye Li and Frank Petriello. "Combining QCD and electroweak corrections to dilepton production in FEWZ". In: *Phys.Rev.* D86 (2012), p. 094034. DOI: 10.1103/PhysRevD.86.094034. arXiv: 1208.5967 [hep-ph].

[167] J. Alwall et al. "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations". In: *JHEP* 1407 (2014), p. 079. DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph].

[168] R. Frederix et al. "The automation of next-to-leading order electroweak calculations". In: *JHEP* 07 (2018), p. 185. DOI: 10.1007/JHEP07(2018)185. arXiv: 1804.10017 [hep-ph].

[169] John M. Campbell and R. Keith Ellis. "An Update on vector boson pair production at hadron colliders". In: *Phys. Rev. D* 60 (1999), p. 113006. DOI: 10.1103/PhysRevD.60.113006. arXiv: hep-ph/9905386.

[170] John M. Campbell, R. Keith Ellis, and Ciaran Williams. "Vector boson pair production at the LHC". In: *JHEP* 07 (2011), p. 018. DOI: 10.1007/JHEP07(2011)018. arXiv: 1105.0020 [hep-ph].

[171] John M. Campbell, R. Keith Ellis, and Walter T. Giele. "A Multi-Threaded Version of MCFM". In: *Eur. Phys. J. C* 75.6 (2015), p. 246. DOI: 10.1140/epjc/s10052-015-3461-2. arXiv: 1503.06182 [physics.comp-ph].

[172] John Campbell and Tobias Neumann. "Precision Phenomenology with MCFM". In: *JHEP* 12 (2019), p. 034. DOI: 10.1007/JHEP12(2019)034. arXiv: 1909.09117 [hep-ph].

[173] Radja Boughezal et al. "W-boson production in association with a jet at next-to-next-to-leading order in perturbative QCD". In: *Phys. Rev. Lett.* 115.6 (2015), p. 062002. DOI: 10.1103/PhysRevLett.115.062002. arXiv: 1504.02131 [hep-ph].

[174] A. Gehrmann-De Ridder et al. "Precise QCD predictions for the production of a Z boson in association with a hadronic jet". In: *Phys. Rev. Lett.* 117.2 (2016), p. 022001. DOI: 10.1103/PhysRevLett.117.022001. arXiv: 1507.02850 [hep-ph].

[175] D. Britzger et al. "Calculations for deep inelastic scattering using fast interpolation grid techniques at NNLO in QCD and the extraction of α_s from HERA data". In: *Eur. Phys. J. C* 79.10 (2019). [Erratum: Eur.Phys.J.C 81, 957 (2021)], p. 845. DOI: 10.1140/epjc/s10052-021-09688-3. arXiv: 1906.05303 [hep-ph].

[176] Zoltan Nagy. "Three jet cross-sections in hadron hadron collisions at next-to-leading order". In: *Phys.Rev.Lett.* 88 (2002), p. 122003. DOI: 10.1103/PhysRevLett.88.122003. arXiv: hep-ph/0110315 [hep-ph].

[177] Michal Czakon and Alexander Mitov. "Top++: A Program for the Calculation of the Top-Pair Cross-Section at Hadron Colliders". In: *Comput. Phys. Commun.* 185 (2014), p. 2930. DOI: 10.1016/j.cpc.2014.06.021. arXiv: 1112.5675 [hep-ph].

[178] Charalampos Anastasiou et al. "High precision QCD at hadron colliders: Electroweak gauge boson rapidity distributions at NNLO". In: *Phys. Rev.* D69 (2004), p. 094008. DOI: 10.1103/PhysRevD.69.094008. arXiv: hep-ph/0312266.

[179] T. Gleisberg et al. "Event generation with SHERPA 1.1". In: *JHEP* 02 (2009), p. 007. DOI: 10.1088/1126-6708/2009/02/007. arXiv: 0811.4622 [hep-ph].

[180] Jun Gao, Lucian Harland-Lang, and Juan Rojo. "The Structure of the Proton in the LHC Precision Era". In: *Phys. Rept.* 742 (2018), pp. 1–121. DOI: 10.1016/j.physrep.2018.03.002. arXiv: 1709.04922 [hep-ph].

[181] A. Accardi et al. "Electron Ion Collider: The Next QCD Frontier: Understanding the glue that binds us all". In: *Eur. Phys. J. A* 52.9 (2016). Ed. by A. Deshpande, Z. E. Meziani, and J. W. Qiu, p. 268. DOI: 10.1140/epja/i2016-16268-9. arXiv: 1212.1701 [nucl-ex].

[182] Daniele P. Anderle et al. "Electron-ion collider in China". In: *Front. Phys. (Beijing)* 16.6 (2021), p. 64701. DOI: 10.1007/s11467-021-1062-0. arXiv: 2102.09222 [nucl-ex].

[183] Tancredi Carli et al. "A posteriori inclusion of parton density functions in NLO QCD final-state calculations at hadron colliders: The APPLGRID Project". In: *Eur.Phys.J.* C66 (2010), p. 503. DOI: 10.1140/epjc/s10052-010-1255-0. arXiv: 0911.2985 [hep-ph].

[184] Daniel Britzger et al. "New features in version 2 of the fastNLO project". In: *20th International Workshop on Deep-Inelastic Scattering and Related Subjects*. 2012, pp. 217–221. DOI: 10.3204/DESY-PROC-2012-02/165. arXiv: 1208.3641 [hep-ph].

[185] Christopher Schwan et al. *NNPDF/pineappl: v0.5.9*. Version v0.5.9. Jan. 2023. DOI: 10.5281/zenodo.7499507. URL: https://doi.org/10.5281/zenodo.7499507.

[186] Luigi Del Debbio, Nathan P. Hartland, and Steffen Schumann. "MCgrid: projecting cross section calculations on grids". In: *Comput.Phys.Commun.* 185 (2014), pp. 2115–2126. DOI: 10.1016/j.cpc.2014.03.023. arXiv: 1312.4460 [hep-ph].

[187] Valerio Bertone et al. "aMCfast: automation of fast NLO computations for PDF fits". In: *JHEP* 08 (2014), p. 166. DOI: 10.1007/JHEP08(2014)166. arXiv: 1406.7693 [hep-ph].

[188] S. Carrazza et al. "PineAPPL: combining EW and QCD corrections for fast evaluation of LHC processes". In: *JHEP* 12 (2020), p. 108. DOI: 10.1007/JHEP12(2020)108. arXiv: 2008.12789 [hep-ph].

[189] Eamonn Maguire, Lukas Heinrich, and Graeme Watt. "HEPData: a repository for high energy physics data". In: *J. Phys. Conf. Ser.* 898.10 (2017). Ed. by Richard Mount and Craig Tull, p. 102006. DOI: 10.1088/1742-6596/898/10/102006. arXiv: 1704.05473 [hep-ex].

[190] *Ploughshare*. URL: http://ploughshare.web.cern.ch/.

[191] Richard D. Ball et al. "Parton distributions for the LHC Run II". In: *JHEP* 04 (2015), p. 040. DOI: 10.1007/JHEP04(2015)040. arXiv: 1410.8849 [hep-ph].

[192] Alessandro Candido, Felix Hekhorn, and Giacomo Magni. *N3PDF/yadism: FONLL-B*. Version v0.11.0. Feb. 2022. DOI: 10.5281/zenodo.6285149. URL: https://doi.org/10.5281/zenodo.6285149.

[193] Alessandro Candido et al. "Yadism: Yet Another Deep-Inelastic Scattering Module". In: (Jan. 2024). arXiv: 2401.15187 [hep-ph].

[194] Massimiliano Grazzini, Stefan Kallweit, and Marius Wiesemann. "Fully differential NNLO computations with MATRIX". In: *Eur. Phys. J.* C78.7 (2018), p. 537. DOI: 10.1140/epjc/s10052-018-5771-7. arXiv: 1711.06631 [hep-ph].

[195] Guido Altarelli and G. Parisi. "Asymptotic Freedom in Parton Language". In: *Nucl. Phys.* B126 (1977), pp. 298–318. DOI: 10.1016/0550-3213(77)90384-4.

[196] V. N. Gribov and L. N. Lipatov. "Deep inelastic e p scattering in perturbation theory". In: *Sov. J. Nucl. Phys.* 15 (1972). [Yad. Fiz.15,781(1972)], pp. 438–450.

[197] Yuri L. Dokshitzer. "Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics." In: *Sov. Phys. JETP* 46 (1977). [Zh. Eksp. Teor. Fiz.73,1216(1977)], pp. 641–653.

[198] Alessandro Candido, Felix Hekhorn, and Giacomo Magni. *N3PDF/eko: Paper*. Version v0.8.5. Mar. 2022. DOI: 10.5281/zenodo.6340153. URL: https://doi.org/10.5281/zenodo.6340153.

[199] Alessandro Candido, Felix Hekhorn, and Giacomo Magni. "EKO: evolution kernel operators". In: *Eur. Phys. J. C* 82.10 (2022), p. 976. DOI: 10.1140/epjc/s10052-022-10878-w. arXiv: 2202.02338 [hep-ph].

[200] A. Vogt. "Efficient evolution of unpolarized and polarized parton distributions with QCD-PEGASUS". In: *Comput. Phys. Commun.* 170 (2005), pp. 65–92. DOI: 10.1016/j.cpc.2005.03.103. arXiv: hep-ph/0408244.

[201] M. Botje. "QCDNUM: Fast QCD Evolution and Convolution". In: *Comput.Phys.Commun.* 182 (2011), pp. 490–532. DOI: 10.1016/j.cpc.2010.10.020. arXiv: 1005.1481 [hep-ph].

[202] Valerio Bertone. "APFEL++: A new PDF evolution library in C++". In: *PoS* DIS2017 (2018). Ed. by Uta Klein, p. 201. DOI: 10.22323/1.297.0201. arXiv: 1708.00911 [hep-ph].

[203] Felix Hekhorn et al. *Heavy Quarks in Polarised Deep-Inelastic Scattering at the Electron-Ion Collider*. 2024. arXiv: 2401.10127 [hep-ph]. URL: https://arxiv.org/abs/2401.10127.

[204] Andrea Barontini et al. *An FONLL prescription with coexisting flavor number PDFs*. in preparation. 2024.

[205] M. A. G. Aivazis et al. "Leptoproduction of heavy quarks. 2. A Unified QCD formulation of charged and neutral current processes from fixed target to collider energies". In: *Phys. Rev. D* 50 (1994), pp. 3102–3118. DOI: 10.1103/PhysRevD.50.3102. arXiv: hep-ph/9312319.

[206] R. S. Thorne and R. G. Roberts. "A Practical procedure for evolving heavy flavor structure functions". In: *Phys. Lett. B* 421 (1998), pp. 303–311. DOI: 10.1016/S0370-2693(97)01580-3. arXiv: hep-ph/9711223.

[207] Michael Krämer, Fredrick I. Olness, and Davison E. Soper. "Treatment of heavy quarks in deeply inelastic scattering". In: *Phys. Rev. D* 62 (2000), p. 096007. DOI: 10.1103/PhysRevD.62.096007. arXiv: hep-ph/0003035.

[208] Wu-Ki Tung, Stefan Kretzer, and Carl Schmidt. "Open heavy flavor production in QCD: Conceptual framework and implementation issues". In: *J. Phys. G* 28 (2002). Ed. by Guenter Grindhammer et al., pp. 983–996. DOI: 10.1088/0954-3899/28/5/321. arXiv: hep-ph/0110247.

[209] Pavel M. Nadolsky and Wu-Ki Tung. "Improved Formulation of Global QCD Analysis with Zero-mass Matrix Elements". In: *Phys. Rev. D* 79 (2009), p. 113014. DOI: 10.1103/PhysRevD.79.113014. arXiv: 0903.2667 [hep-ph].

[210] Marco Guzzi et al. "General-Mass Treatment for Deep Inelastic Scattering at Two-Loop Accuracy". In: *Phys. Rev. D* 86 (2012), p. 053005. DOI: 10.1103/PhysRevD.86.053005. arXiv: 1108.5112 [hep-ph].

[211] Matteo Cacciari, Mario Greco, and Paolo Nason. "The p_T spectrum in heavy-flavour hadroproduction." In: *JHEP* 05 (1998), p. 007. DOI: 10.1088/1126-6708/1998/05/007. arXiv: hep-ph/9803400.

[212] Andrea Barontini et al. *An FONLL prescription with coexisting flavor number PDFs*. 2024. arXiv: 2408.07383 [hep-ph]. URL: https://arxiv.org/abs/2408.07383.

[213] G. Moreno et al. "Dimuon production in proton - copper collisions at $\sqrt{s}$ = 38.8-GeV". In: *Phys. Rev.* D43 (1991), pp. 2815–2836. DOI: 10.1103/PhysRevD.43.2815.

[214] J. C. Webb et al. "Absolute Drell-Yan dimuon cross sections in 800-GeV/c p p and p d collisions". In: (2003). arXiv: hep-ex/0302019.

[215] R. S. Towell et al. "Improved measurement of the anti-d/anti-u asymmetry in the nucleon sea". In: *Phys. Rev.* D64 (2001), p. 052002. DOI: 10.1103/PhysRevD.64.052002. arXiv: hep-ex/0103030.

[216] J. Dove et al. "The asymmetry of antimatter in the proton". In: *Nature* 590.7847 (2021), pp. 561–565. DOI: 10.1038/s41586-021-03282-z. arXiv: 2103.04024 [hep-ph].

[217] Claude Duhr, Falko Dulat, and Bernhard Mistlberger. "Drell-Yan Cross Section to Third Order in the Strong Coupling Constant". In: *Phys. Rev. Lett.* 125.17 (2020), p. 172001. DOI: 10.1103/PhysRevLett.125.172001. arXiv: 2001.07717 [hep-ph].

[218] Stefano Carrazza et al. "Specialized minimal PDFs for optimized LHC calculations". In: *Eur. Phys. J.* C76.4 (2016), p. 205. DOI: 10.1140/epjc/s10052-016-4042-8. arXiv: 1602.00005 [hep-ph].

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my advisor, Stefano Forte, for his unwavering support, guidance, and encouragement throughout my Ph.D. journey. His exceptional ability to foster a stimulating and collaborative environment has been truly invaluable.

I am sincerely thankful to all the members of the NNPDF collaboration and of the phenomenology group. Their profound expertise, insightful feedback, and continuous mentorship have played a crucial role in shaping this work and in my development as a researcher.

I am especially grateful to the younger members of NNPDF and N3PDF for making me feel like a part of the team from the very beginning. In particular, I would like to thank Alessandro Candido, Felix Hekhorn, Juan Cruz Martinez, and Roy Stegeman, who warmly welcomed me to the University of Milan and helped make the experience far less lonely.

My heartfelt thanks go to Michele Caresana, Jacopo D'Alberto, Davide Morgante, Niccolò Laurenti, and Davide Maria Tagliabue, with whom I had the pleasure of being both a colleague and a friend (and a housemate as well). Without your companionship, the rainy days in Milan might have brought my Ph.D. journey to a premature end (yes Davide, Milan is ugly for 90% of the year). I would also like to extend my thanks to all the friends I made in the Physics Department, who have been an essential part of this journey.

On a personal note, I owe a deep debt of gratitude to my family for their unconditional love, patience, and encouragement throughout my academic endeavors. To my parents, Paolo and Alessandra, and my brother Federico. Your unwavering belief in me has been a constant source of strength.

I also wish to express my appreciation to my friends, who, despite being far away throughout the course of my Ph.D., were always close at heart. You know who you are: though distant, you were never absent in spirit. Thank you for your understanding, support, and for being there when I needed you the most.

Finally, a special thank you to Francesca, my infinite source of motivation and so much more. Your patience and constant encouragement to be the best version of myself have been invaluable. This journey, now drawing to a close, is just the first step in our future together.