UNIVERSITÀ
DI TORINO

Università degli Studi di Torino

Corso di Laurea Magistrale in Fisica

Impact of Semi-Inclusive Deep-Inelastic Scattering Data on the Determination of Parton Distribution Functions

Tesi di Laurea Magistrale

Relatore
Dr. Emanuele R. Nocera

Co-relatore
Dr. Andrea Signori

Controrelatore
Prof.ssa Mariaelena Boglione

Candidato
Lorenzo Canzian
Matricola 862980

Contents

Acknowledgements **5**

1 Introduction **7**

2 QCD and perturbation theory **9**
2.1 Deep inelastic scattering . . . 9
2.2 The parton model and QCD . . . 14
2.3 Factorization . . . 16
2.4 Scaling violation and the DGLAP equation . . . 17
2.4.1 Solution of the leading order DGLAP equations . . . 20

3 PDF determination, SIDIS, and FFs **23**
3.1 PDF determination . . . 23
3.2 Semi Inclusive Deep Inelastic Scattering . . . 25
3.3 Fragmentation Functions . . . 27
3.4 SIDIS facilities and experiments . . . 28

4 Methodology **31**
4.1 Iterative Procedure . . . 31
4.2 Fit of fragmentation functions . . . 32
4.3 LHAPDF Grid . . . 34
4.4 Theoretical Predictions . . . 34
4.5 Reweighting . . . 35
4.5.1 Weights . . . 35
4.5.2 Measuring information Loss and Consistency . . . 36
4.6 Unweighting . . . 37
4.6.1 Unweighting Algorithm . . . 38
4.7 Distances between PDFs: definition and meaning . . . 39

5 Results **41**
5.1 χ^2 distributions at NLO and NNLO . . . 41
5.2 Weights distributions at NLO and NNLO . . . 42
5.3 Comparison plot on the impact of Pions and Kaons data . . . 43
5.4 Impact of SIDIS data on Parton Distribution Functions . . . 45
5.5 PDFs distances, Rs plot and Data-Prediction comparison . . . 47

6 Conclusion **59**

A Proof of the weight formula **61**

B Additional results **65**

C Bibliography **71**

Acknowledgments

Before diving into the details of this project, I would like to express my gratitude to all the people who have supported me throughout my studies. I have met many amazing people who have made my experience richer and more enjoyable.

I will never have enough words to thank my thesis advisors, Dr. Emanuele Roberto Nocera and Dr. Andrea Signori. With kindness and a friendly attitude, they provided me with thoughtful advice and continuous support while facing the challenges of this thesis. They always encouraged me to ask as many questions as necessary, answering each one in a clear and precise manner. I always looked forward to our weekly meetings, as they were both enjoyable and highly instructive. I am also deeply grateful for the time they dedicated to reviewing this thesis and for their truly helpful feedback.

A special thanks goes to my girlfriend, who supported me in every possible way. Without her, this journey would have been ten times harder. With kindness and love, she offered me a safe place where I could relax and catch my breath. I can't count how many times she listened to my problems, always doing everything she could to help me.

Vorrei ringraziare la mia famiglia, in particolare i miei genitori, senza i quali nulla di tutto questo sarebbe stato possibile. Li ringrazio per la pazienza e il supporto che mi hanno dimostrato in questi anni, soprattutto nei momenti più difficili. Grazie al loro costante impegno, sono riuscito a seguire gli studi libero da preoccupazioni. Ringrazio anche mio fratello Davide e mia sorella Martina, che con il loro affetto e la loro fiducia mi hanno sempre spronato ad essere una figura d'esempio. Infine, vorrei ringraziare i miei nonni, che sono sempre stati pronti a confortarmi e proteggermi come solo loro sanno fare. Spero, con questo traguardo, di averli resi orgogliosi.

Lorenzo Canzian

Chapter 1

Introduction

Nucleons (protons and neutrons) are bound states that constitute all atomic nuclei and, consequently, most of the visible matter in the Universe. A deeper understanding of hadronic structure in terms of their elementary constituents, known as partons, is a key ingredient to interpret cross section measurements for a wide array of processes at hadron colliders. Such understanding is rooted in the theoretical framework of the Standard Model, which defines elementary particles and their interactions through two fundamental components: the spontaneously broken $SU(2) \times U(1)$ electroweak theory and the unbroken $SU(3)$ color gauge theory, known as Quantum Chromodynamics (QCD).

Since energy increases with the separation of color charges, one of the defining features of QCD, partons are confined to exist only in color-neutral combinations, i.e., hadrons. Hadrons, typically protons and neutrons, are probed through scattering experiments using beams of leptons or protons/antiprotons in large-momentum-transfer processes. As elementary interactions occur at length scales much smaller than the confinement scale, the measurable cross section of such processes can be determined by convoluting the partonic cross section with Parton Distribution Functions (PDFs). The former encodes the scattering of quasi-free partons in terms of process-dependent kernels computed perturbatively in QCD, while the latter describe the momentum distribution of partons involved in the elementary scattering process through universal functions. Parton distribution functions are universal objects and thus serve as essential tools for interpreting experimental data from a wide range of hard-scattering processes. This has led to extensive studies aimed at determining PDFs and their associated uncertainties as accurately as possible, typically through global fits to diverse experimental datasets.

The objective of this thesis is to determine the impact of a specific process, Semi-Inclusive Deep Inelastic Scattering (SIDIS), on a set of precomputed PDFs at next-to-leading order (NLO) and next-to-next-to-leading order (NNLO) in perturbative QCD. In this process, a proton is probed by a lepton beam, and both the scattered lepton and one produced hadron are observed in the final state. The measurable cross section for SIDIS requires, in addition to partonic cross sections and PDFs, an additional contribution from Fragmentation Functions (FFs). These functions describe the hadronization process, in which a quasi-free parton transforms into a hadron in the final state.

Including SIDIS data in PDF determination is particularly interesting for two main reasons. First, there may be an interplay between PDFs and FFs within the SIDIS process. Second, a new experiment, the Electron-Ion Collider (EIC), is currently under development and will primarily study SIDIS over a broad kinematic range. Gaining a deeper understanding of the impact of SIDIS processes is therefore valuable for the preparation and implementation of this new experiment.

In general, the fitting procedures used to determine PDFs and FFs are complex and computationally demanding. Moreover, incorporating SIDIS data presents an additional challenge, as it requires simultaneous fitting of PDFs and FFs due to their correlation. A limited number of previous studies have attempted to include SIDIS in PDF determination. One approach employs a sequential fit, where FFs are first determined through fitting, followed by a second fit for PDFs while keeping FFs fixed [1, 2]. An alternative approach is to fit only the FFs and subsequently apply a technique called reweighting to the PDFs, which allows for evaluating the impact of new data on a Monte Carlo replica set [3].

In this thesis, we follow the latter approach and implement an iterative procedure based on reweighting. The idea is to compute a set of weights using only the χ^2 values from the new data. Each weight quantifies the importance of a specific replica of the PDF set in describing these new data. This results in a new weighted PDF set that incorporates the information from the included datasets, thereby revealing the impact of SIDIS

data on the PDFs. Standard PDF sets require that all replicas in a set share the same weight, equal to 1. For this reason, we employ an additional technique called unweighting, which generates a new PDF set that retains all the information from the reweighted set while ensuring uniform weights.

This study is particularly timely because previous works have explored the impact of SIDIS only at NLO and, in some cases, with a more limited dataset for determining the initial PDF set. Recently, partonic cross section calculations have become available at NNLO, and the two primary SIDIS experiments, COMPASS and HERMES, have concluded their data collection. We are now able to investigate the impact of SIDIS at NNLO with the largest available dataset for both FFs and PDFs.

This thesis is organized as follows.

In Chapter 2, I will discuss fundamental properties of QCD and their application in perturbation theory. The structure of this chapter is as follows: in Section 2.1, I introduce a basic example of a hard-scattering process, Deep Inelastic Scattering (DIS), which facilitates the definition of PDFs. In Section 2.2, I describe the Parton Model, leading to the introduction of factorization theorems in Section 2.3. In Section 2.4, I present the evolution equations for PDFs, commonly known as the DGLAP equations.

Chapter 3 details the fitting procedures used to determine PDFs and FFs, along with an overview of the techniques applied in this project. Section 3.2 introduces the SIDIS process, while Section 3.3 describes fragmentation functions. Finally, Section 3.4 briefly reviews relevant SIDIS experiments.

Chapter 4 presents the iterative procedure developed for this study, explaining each step in detail. In Section 4.2, I describe the FF fitting process. Sections 4.3 and 4.4 outline the construction of LHAPDF grids and the computation of theoretical predictions using the MontBlanc code from the MAP Collaboration. Sections 4.5 and 4.6 introduce the two techniques I developed to evaluate the impact of SIDIS data on the initial PDF set.

Chapter 5 presents the results obtained in this study. Sections 5.1, 5.2, and 5.3 contain χ^2 distributions, weight distributions, and comparisons of quark distributions at NLO and NNLO for both pion and kaon data. In Section 5.4, I integrate these findings by sequentially incorporating pion and kaon data in the reweighting procedure. Section 5.5 discusses PDF distances, the R_s distribution, which quantifies the fraction of strange quarks in the sea, and comparisons of theoretical predictions before and after applying the reweighting procedure.

The thesis is completed by two appendices. In Appendix A I present the proof of Eq. (4.12), which is at the core of the reweighting procedure. In Appendix B I present some additional results not shown in Sect. 5.

Chapter 2

QCD and perturbation theory

The theory describing the strong interaction among hadrons, Quantum Chromodynamics (QCD), is a non-Abelian gauge theory within the framework of quantum field theory. QCD consists of two fundamental types of fields: quark fields and gluon fields. The particles associated with the quark fields serve as the basic constituents of hadrons, while the gluon fields mediate the strong interaction, binding quarks together. A fundamental property of QCD is confinement, which dictates that quarks and gluons cannot be observed in isolation but instead exist only within hadrons. Another crucial feature of QCD is asymptotic freedom, which implies that the effective coupling strength of QCD decreases at short distances or, equivalently, at high energies. This behavior allows for the use of perturbative methods to analyze short-distance processes in QCD (pQCD). However, many phenomena, including the formation of bound states such as hadrons, cannot be studied using perturbative techniques due to the dominance of strong interactions at long distances.

Despite these limitations, perturbative QCD remains a powerful tool in high-energy hadronic physics. The key principle enabling its applicability is factorization, which allows the separation of short-distance perturbative interactions from long-distance non-perturbative effects. In this approach, hadrons are described as collections of quasi-free quarks and gluons, known as partons. The probability distributions of these partons, called parton distribution functions (PDFs), are extracted from experimental data. The evolution of these distributions with energy scales is then computed using the perturbative DGLAP (Dokshitzer-Gribov-Lipatov-Altarelli-Parisi) equations, which serve as the QCD analog of the renormalization group equation. Through factorization, many physical cross sections can be predicted by convoluting PDFs with short-distance partonic cross sections, which are calculated using Feynman diagrams involving free partons. The universality of these non-perturbative parton distributions is a crucial feature that grants pQCD its predictive power.

The structure of this Chapter is as follows. In Sect. 2.1, I will introduce a fundamental process known as Deep Inelastic Scattering (DIS), which provides the foundation for a more technical discussion of parton distribution functions. In Sect. 2.3, I will explore the concept of factorization and its application to the DIS cross section. Finally, in Sect. 2.4, I will present the DGLAP equations and explain how they account for scaling violations in QCD.

2.1 Deep inelastic scattering

One of the key experimental discoveries that led to the establishment of QCD as the theory of strong interactions was Deep Inelastic Scattering (DIS). This process involves the high-energy scattering of a charged lepton, l, off a target hadron, H. Let k^μ and k'^μ be the four-momenta of the incoming and outgoing lepton, respectively, p^μ be the four-momentum of the target hadron, and $q^\mu = k^\mu - k'^\mu$ be the four-momentum transferred to the hadron:

$$l(k) + H(p) \longrightarrow l^{'}(k^{'}) + X. \tag{2.1}$$

Here, X denotes the sum over all possible hadronic final states, meaning that we consider an inclusive cross section, which is differential in the final-state lepton momentum k'^μ.

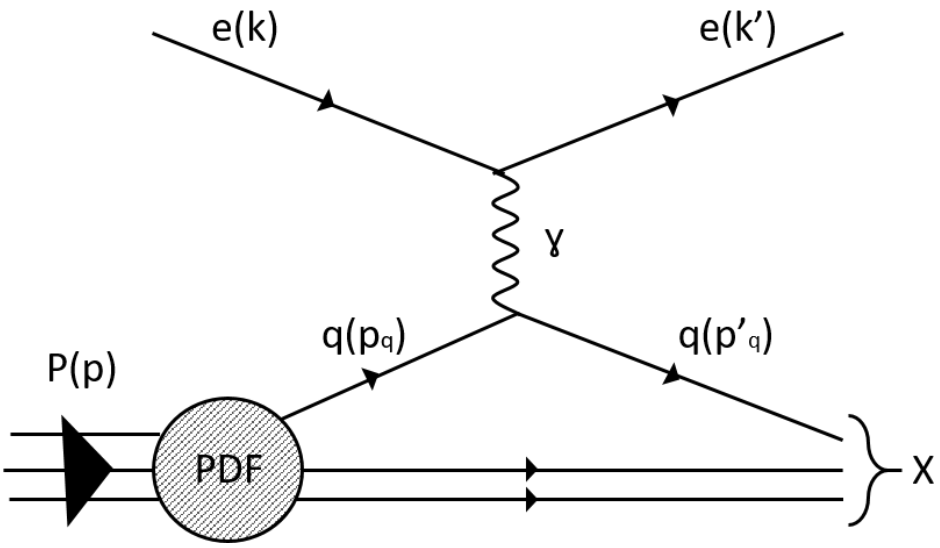

Figure 2.1: Feynman diagram representing an electron with momentum k scattering off a proton with momentum p. The interaction is mediated by a virtual boson with momentum q.

The standard kinematic variables used to describe DIS are:

$$\begin{aligned} Q^2 &= -q^2, \\ M^2 &= p^2, \\ \nu &= p \cdot q = M(E^{'} - E), \\ x &= \frac{Q^2}{2\nu} = \frac{Q^2}{2M(E - E')}, \\ y &= \frac{q \cdot p}{k \cdot p} = 1 - \frac{E^{'}}{E}, \end{aligned} \tag{2.2}$$

where Q^2 is the squared four-momentum transfer, M is the mass of the hadron, and ν represents the energy transferred to the hadron in its rest frame. The variable x is the Bjorken scaling variable, which is constrained kinematically to the range $Q^2/(s + Q^2) \leq x \leq 1$, where s is the center-of-mass energy squared. Later, we will see that x provides an estimate of the fraction of the initial hadron's momentum carried by the struck parton. The variable y lies between 0 and 1 and, in the target rest frame, represents the fractional energy loss of the lepton: $(E - E^{'})/E$, where E and $E^{'}$ are the initial and final lepton energies, respectively.

In this discussion, we consider the scattering of an electron or muon beam off a proton, where the interaction is mediated by the exchange of a virtual photon, as depicted in Fig. 2.1. To describe the DIS cross section, we introduce the structure functions $F_i(x, Q^2)$, $i = 1, 2$, which encode information about the target hadron as probed by the virtual photon. For charged lepton scattering, $lH \longrightarrow lX$, the differential cross section is given by:

$$\frac{d^2\sigma^{em}}{dxdy} = \frac{8\pi\alpha^2 ME}{Q^4}\left[\left(\frac{1 + (1 - y)^2}{2}\right) 2xF_1^{em} + (1 - y)(F_2^{em} - 2xF_1^{em}) - (M/2E)xyF_2^{em}\right], \tag{2.3}$$

where α is the electromagnetic coupling. For neutrino scattering, $\nu H \longrightarrow lX$, the cross section takes the form:

$$\frac{d^2\sigma^{\nu}}{dxdy} = \frac{G_F^2 ME}{\pi}\left[\left(1 - y - \frac{M}{2E}xy\right) F_2^{\nu} + y^2 xF_1^{\nu} y\left(1 - \frac{1}{2}y\right) xF_3^{\nu}\right], \tag{2.4}$$

where G_F is the Fermi constant.

In the Bjorken limit, defined as $Q^2, \nu \to \infty$ with x fixed, the structure functions obey an approximate scaling law, meaning they depend only on x:

$$F_i(x, Q^2) \to F_i(x). \tag{2.5}$$

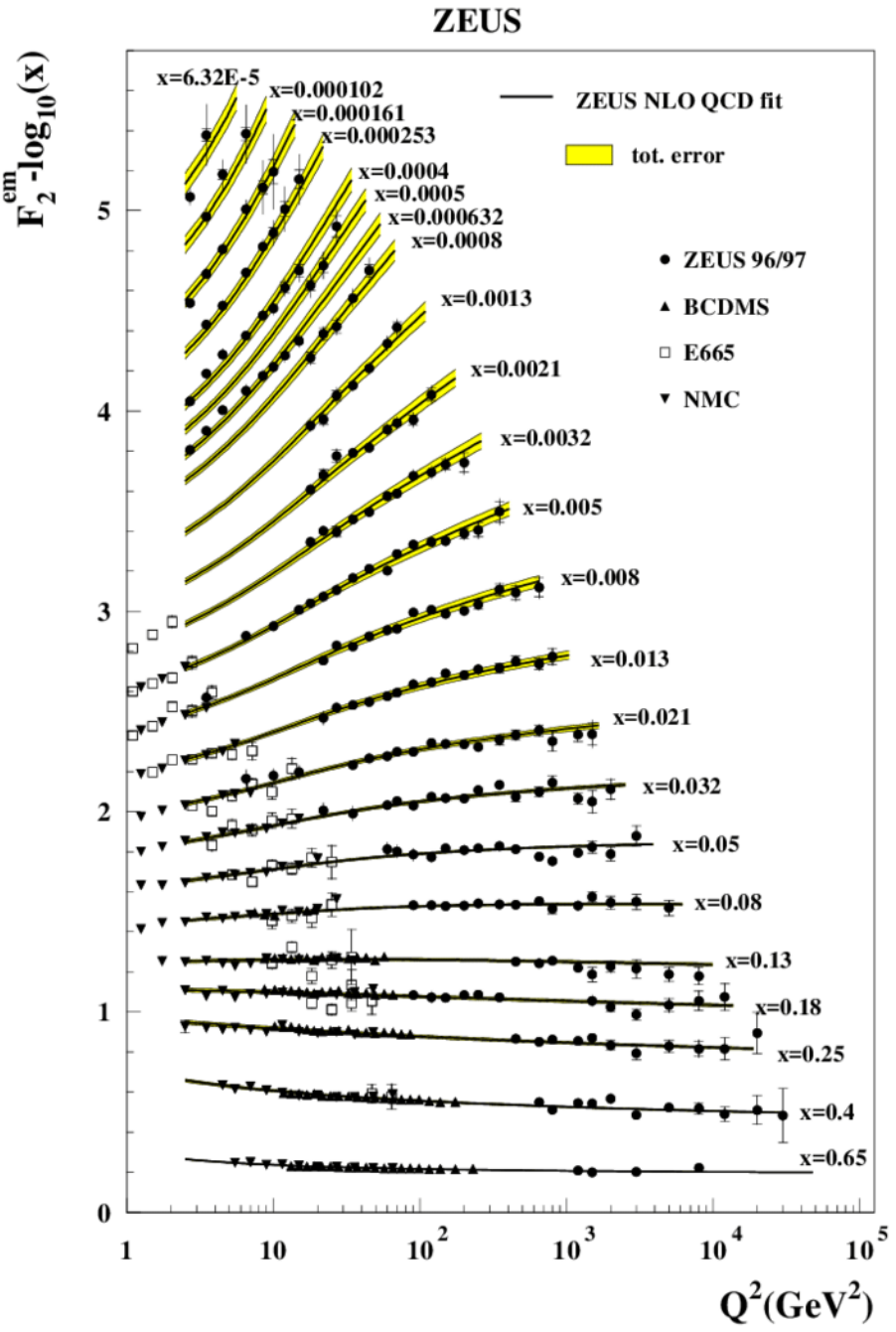

Figure 2.2: F_2 measured from experimental data. In the naive parton model, F_2 does not depend on Q^2 at fixed x, in accordance with Bjorken scaling. However, this scaling is broken logarithmically due to QCD effects [4].

Experimental data confirm this approximate scaling behavior. Fig. 2.2 shows measurements of the structure function F_2 at different energy scales Q^2. As observed, F_2 remains nearly constant at fixed x, indicating its approximate independence from Q^2, as expected from Bjorken scaling. However, QCD predicts a slow logarithmic violation of this scaling due to gluon radiation, leading to the DGLAP evolution equations.

This Bjorken scaling behavior suggests that the virtual photon scatters off point-like constituents inside the hadron. If the hadron had an intrinsic finite size, the structure functions would depend on Q/Q_0, with $1/Q_0$ being a characteristic length scale. This intuition led to the so-called parton model. The parton model interpretation of DIS is more intuitive in the infinite momentum frame, where the proton is highly boosted,

$p^\mu = (P, 0, 0, P)$ with $P \gg M$. In this frame, the virtual photon scatters off a quasi-free quark, which carries a fraction ξ of the proton's momentum. By comparing the theoretical cross section with experimental data, we obtain expressions for the structure functions and infer the presence of parton distributions within the hadron. Neglecting the proton mass M, we can rewrite Eq. (2.3) as

$$\frac{d^2\sigma}{dxdQ^2} = \frac{4\pi\alpha^2}{Q^4}\left[[1+(1-y)^2]F_1 + \frac{(1-y)}{x}(F_2 - 2xF_1)\right]. \tag{2.6}$$

Now we would like to find an expression for F_1 and F_2. In order to do that we have to compute the matrix element squared for the process,

$$e^-(k) + q(p_q) \longrightarrow e^-(k') + q(p'_q) \tag{2.7}$$

hence a process were the electron scatters off a single parton. The result is

$$\overline{\sum}|\mathcal{M}|^2 = 2e_q^2 e^4 \frac{\hat{s}^2 + \hat{u}^2}{\hat{t}^2} \tag{2.8}$$

where e_q is the quark charge, e the electron charge, the sum denotes the average (sum) over initial (final) colors and spins and $u = (p_q - k')^2$, $t = (k - k')^2$, $s = (k + p_q)^2$ are the Mandelstam variables. Using Eq. (2.2) we can substitute for the deep inelastic variables: $u = s(y-1)$, $t = -Q^2$, $s = \xi Q^2/xy$. Using the standard result for the cross section for massless $2 \longrightarrow 2$ scattering,

$$\frac{d\sigma}{dt} = \frac{1}{16\pi\hat{s}^2}\overline{\sum}|\mathcal{M}|^2 \tag{2.9}$$

and substituting for the kinematic variables in the matrix element squared, we obtain:

$$\frac{d\sigma}{dQ^2} = \frac{2\pi\alpha^2 e_q^2}{Q^4}[1+(1-y)^2]. \tag{2.10}$$

The mass shell constraint for the outgoing quark,

$$p'^2_q = (p_q + p)^2 = q^2 + 2p_q \cdot q = -2p \cdot q(x - \xi) = 0, \tag{2.11}$$

implies $x = \xi$, i.e x represent the momentum of the quark which is a fraction of the hadron's momentum. By writing $\int_0^1 dx\delta(x-\xi) = 1$, we obtain the double differential cross section for the quark scattering process:

$$\frac{d^2\hat{\sigma}}{dxdQ^2} = \frac{4\pi\alpha^2}{Q^4}[1+(1-y)^2]\frac{1}{2}e_q^2\delta(x-\xi). \tag{2.12}$$

By comparing Eqs. (2.6) and (2.12), we see that the structure functions in this simple model are:

$$\hat{F_2} = xe_q^2\delta(x-\xi) = 2x\hat{F_1}. \tag{2.13}$$

This result suggests that $F_2(x)$ "probes" a quark constituent with momentum fraction $\xi = x$. From experimental results, the measured structure function is a distribution in x rather than a delta function, which indicates that the quark constituents carry a range of momentum fractions.

The above ideas are incorporated in what is known as the "naive parton model":

- $q(\xi)d\xi$ represents the probability that a quark q carries momentum fraction between ξ and $\xi + d\xi$, where $0 \leq \xi \leq 1$;
- the virtual photon scatters incoherently off the quark constituents.

In order to obtain the proton structure function we have to weight the quark structure functions with the probability distribution $q(\xi)$,

$$F_2(x) = 2xF_1(x) = \sum_{q,\bar{q}}\int_0^1 d\xi q(\xi)xe_q^2\delta(x-\xi) = \sum_{q,\bar{q}} e_q^2 xq(x). \tag{2.14}$$

summed over quarks and anti-quarks. One of the most commonly encountered parton model deep inelastic structure function is

$$2xF_1 = F_2 \tag{2.15}$$

This relation is called Callan-Gross relation and is a direct consequence of the spin $\frac{1}{2}$ property of the quarks. To understand this, we note that the two terms in the square brackets on the RHS of Eq. (2.6) correspond to the absorption of transversely (F_1) and longitudinally ($F_2 - 2xF_1$) polarized virtual photons. In fact the combination

$$F_L(x,Q^2) = \left(1 + \frac{4M^2x^2}{Q^2}\right)F_2(x,Q^2) - 2xF_1(x,Q^2) \xrightarrow{Q^2\to\infty} F_2 - 2xF_1, \tag{2.16}$$

called the longitudinal structure function, is sometimes used instead of F_1 or F_2. The Callan-Gross relation follows from the fact that a quark of spin $\frac{1}{2}$ cannot absorb a longitudinally polarized vector boson. In contrast, spin 0 quarks cannot absorb transversely polarized vector bosons and so would have $F_1 = 0$ i.e $F_L = F_2$ in the Bjorken limit. Measurements show that $F_L \ll F_2$ confirming the spin $\frac{1}{2}$ property of quarks.

At very high energy there are other contributions to the $ep \longrightarrow eX$ scattering, like an exchange of a Z or W boson. The generalization of Eq. (2.6) which incorporates the complete neutral current (γ and Z boson) exchange for $e^-p \to e^-X$ scattering is

$$\frac{d^2\sigma_{NC}}{dxdQ^2} = \frac{4\pi\alpha^2}{xQ^4}[xy^2F_1^{NC} + (1-y)F_2^{NC} + y(1-\frac{1}{2}y)F_3^{NC}(x,Q^2)] \tag{2.17}$$

where

$$\begin{aligned} F_2^{NC}(x) = 2xF_1^{NC}(x) &= \sum_q x[q(x) + \bar{q}(x)]C_q(Q^2) \\ xF_3^{NC}(x) &= \sum_q x[q(x) - \bar{q}(x)]D_q(Q^2) \end{aligned} \tag{2.18}$$

with

$$\begin{aligned} C_q(Q^2) &= e_q^2 - 2e_qV_eV_qP_Z + (V_e^2 + A_e^2)(V_q^2 + A_q^2)P_Z^2 \\ D_q(Q^2) &= -2e_qA_eA_qP_Z + 4V_eA_eV_qA_qP_Z^2 \\ P_Z &= \frac{Q^2}{Q^2 + M_Z^2} \end{aligned} \tag{2.19}$$

where V_i and A_i represent the vector and axial coupling to the boson. The charged-current (W-exchange) contribution also becomes significant at high Q^2. The corresponding parton-distribution decomposition of the cross section is

$$\frac{d^2\sigma_{CC}}{dxdQ^2} = \frac{(1-\lambda_e)\pi\alpha^2}{8sin^4\theta_W(Q^2 + M_W^2)^2} \times \sum_{i,j}[\mid V_{u_id_j} \mid^2 u_i(x) + (1-y)^2 \mid V_{u_jd_i} \mid^2 d_i(x)] \tag{2.20}$$

where λ_e is the helicity of the electron, u_i and d_i refer to up- and down- type quarks respectively, and the $V_{u_id_j}$ are the elements of the CKM matrix, which describes the couplings of the quarks to the charged weak current.

If we consider the complete relations for the cross section we can find the following picture. The proton consists of three valence quarks (uud) which carry its electric charge and baryon quantum numbers, and an infinite sea of light $q\bar{q}$ pairs. When probed at momentum scale Q, the sea contains all quark flavors with mass $m_q \ll Q$. Thus, at a scale of $\mathcal{O}(1$ GeV), and assuming the sea to be symmetric in the quark flavors, we would have

$$\begin{aligned} u(x) &= u_V(x) + S(x) \\ d(x) &= d_V(x) + S(x) \\ S(x) &= \bar{u}(x) = \bar{d}(x) = s(x) = \bar{s}(x) \end{aligned} \tag{2.21}$$

hence the quark distribution are a combination of the valence quark and sea quark distributions. These respect the sum rules:

$$\int_0^1 dx u_V(x) = 2, \qquad \int_0^1 dx d_V(x) = 1. \tag{2.22}$$

Regarding the quark sea, it is not SU(3) flavor symmetric. There are some asymmetries between u and d quarks while the strange quark distribution is typically a factor of 2 smaller than light quark distributions.

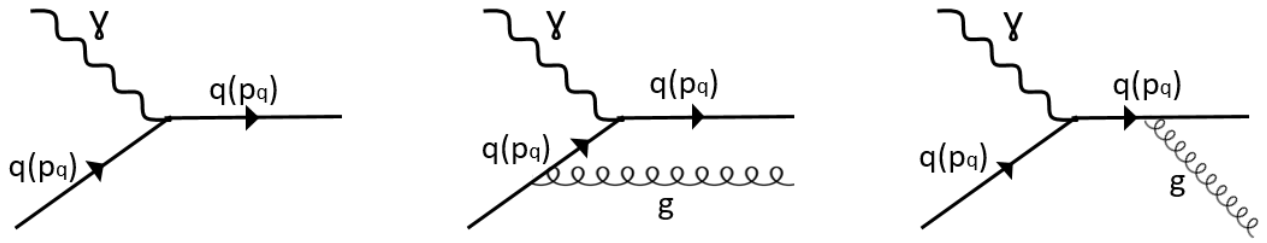

Figure 2.3: Feynman diagrams for Deep inelastic scattering off a quark.

2.2 The parton model and QCD

In the naive parton model, the structure functions exhibit exact scaling, meaning that $F(x, Q^2) \to F(x)$ in the asymptotic limit $Q^2 \to \infty$ with x fixed. However, in QCD, this scaling is broken by logarithmic dependencies on Q^2 due to gluon radiation and higher-order corrections. To understand the origin of these scaling violations, we begin by decomposing the quark four-momentum k^μ in terms of the hadron momentum p^μ, the light-like vector n^μ, and a transverse component k_T^μ:

$$k^\mu = \xi p^\mu + \frac{k^2 + k_T^2}{2\xi} n^\mu + k_T^\mu, \tag{2.23}$$

where the vectors satisfy the constraints:

$$p^2 = n^2 = n \cdot k_T = p \cdot k_T = 0. \tag{2.24}$$

A key observation is that the parton's transverse momentum k_T is not constrained to be small. A quark can radiate a gluon, acquiring a large transverse momentum k_T with a probability proportional to $\alpha_s \frac{dk_T^2}{k_T^2}$ at high k_T. Here, $\alpha_s = \frac{g^2}{4\pi}$ is the strong coupling constant. The integral over k_T^2 extends up to the kinematic limit $k_T^2 \sim Q^2$, leading to logarithmic contributions of the form $\alpha_s \log Q^2$, which break naive scaling. These logarithmic scaling violations are a distinctive feature of renormalizable gauge theories with point-like fermion-vector boson interactions. To explicitly demonstrate this violation in QCD, we will compute the structure function of a quark that can emit a gluon, i.e., the $\mathcal{O}(\alpha_s)$ correction to the parton model result $\hat{F}_2 = eq^2 x\delta(x-\xi)$.

To establish the normalization we calculate the scattering of a virtual photon off a free quark with momentum p, represented by Fig. 2.3(a):

$$\gamma^*(q) + q(p) \longrightarrow q(l). \tag{2.25}$$

The invariant matrix element for this process is

$$\mathcal{M}_\alpha = -ie_q \bar{u}(l)\gamma^\alpha u(p), \tag{2.26}$$

where γ^α are Dirac matrices and u, $\bar{u}$ are the fermionic fields. Now we compute the squared matrix element (summed and averaged over spins and colors) and we project out the F_2 contribution

$$n^\alpha n^\beta \overline{\sum} |\mathcal{M}_{\alpha\beta}|^2 = 4e_q^2. \tag{2.27}$$

The one dimensional phase space is

$$d\Phi_1 = 2\pi\delta((p+q)^2). \tag{2.28}$$

Inserting the flux factor of $\frac{1}{4\pi}$ we obtain

$$\hat{F}_2(x) = e_q^2 \delta(1-x), \tag{2.29}$$

which is the naive parton model result with $\xi = 1$. The $\hat{}$ indicates that we are referring to a quark, rather then proton, target.

Next, we have to consider more complicated processes represented by Fig. 2.3(b,c) in which the quark emits a gluon:

$$\gamma^*(q) + q(p) \longrightarrow g(r) + q(l), \tag{2.30}$$

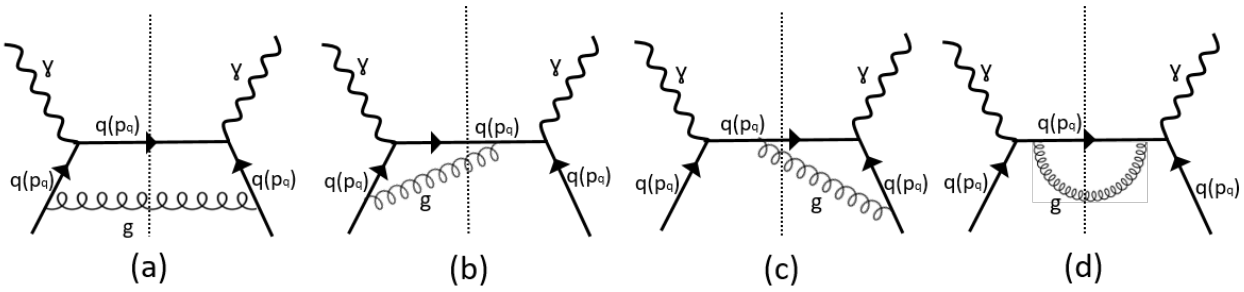

Figure 2.4: Gluon emission diagrams contributing to deep inelastic scattering.

where q, p, r, l denotes respectively the momentum of the photon, initial quark, gluon and quark in the final state. The square Feynman diagrams involving real radiation are shown in Fig. 2.4. We start from the diagram of Fig. 2.4(a). The Lorentz invariant phase space in this case is

$$d\Phi_2 = \int \frac{d^4r}{(2\pi)^3}\frac{d^4l}{(2\pi)^3}\delta^+(r^2)\delta^+(l^2)(2\pi)^4\delta^4((k+q)^2). \tag{2.31}$$

If k^μ is the momentum of the struck parton line, we obtain

$$d\Phi_2 = \frac{1}{4\pi^2}\int d^4k\delta^+((p-k)^2)\delta^+((k+q)^2). \tag{2.32}$$

Rewriting k^μ in terms of p^μ, n^μ, k_T^μ we can find

$$\begin{aligned} k^\mu &= \xi p^\mu + \frac{k_T^2 - |k^2|}{2\xi}n^\mu + k_T^\mu, \\ d^4k^\mu &= \frac{d\xi}{2\xi}dk^2d^2k_T \\ (p-k)^2 &= (1-\xi)\frac{|k^2|}{\xi} - \frac{k_T^2}{\xi}, \\ (k+q)^2 &= 2\xi\nu - Q^2 - |k^2| - 2q_T\cdot k_T, \end{aligned} \tag{2.33}$$

so the phase space becomes

$$d\Phi_2 = \frac{1}{16\nu\pi^2}\int d\xi dk^2 dk_t^2 d\theta\delta(k_T^2 - (1-\xi)|k^2|)\times\delta\left(\xi - x - \frac{|k^2| + 2q_T\cdot k_T}{2\nu}\right), \tag{2.34}$$

with $0 < \theta < \pi$. The corresponding matrix element is

$$\mathcal{M}^\alpha = -ige_q\bar{u}(l)\gamma^\alpha\frac{1}{\not k}\not\epsilon t^A u(p) \tag{2.35}$$

where t^A is the SU(3) color matrix, g is the strong coupling and $C_F = \frac{4}{3}$ in QCD. Squaring and averaging over color and spins gives

$$\overline{\sum}|\mathcal{M}|^2_{\alpha\beta} = \frac{1}{2}e_q^2g^2\sum_{pol}C_FTr[\gamma^\alpha\not k\not\epsilon\not p\not\epsilon^*]\frac{1}{k^4}, \tag{2.36}$$

To perform the sum over the polarization of the real gluon we use the projector

$$\sum_{pol}\epsilon_\mu(r)\epsilon^*_\nu(r) = -g_{\mu\nu} + \frac{n_\mu r_\nu + n_\nu r_\mu}{n\cdot r}. \tag{2.37}$$

Thus in addition to the Lorentz condition $\epsilon\cdot r = 0$, the gluon satisfies the (light-cone) gauge condition $\epsilon\cdot n = 0$. This ensures that only two physical polarization propagate. We can again project out the F_2 contribution by using the vector n. Using the kinematic relations implicit in the phase space terms we find

$$\frac{1}{4\pi}n^\alpha n^\beta\overline{\sum}|\mathcal{M}|^2_{\alpha\beta} = \frac{8e_q^2\alpha_s}{|k^2|}\xi P(\xi), \tag{2.38}$$

where the $P(\xi)$ function is known as splitting function:

$$P(\xi) = C_F \frac{1+\xi^2}{1-\xi}. \tag{2.39}$$

Its form is specific to the qqg vertex of QCD. Putting everything together and performing the k_T and θ integration gives

$$\hat{F}_2 = e_q^2 \frac{\alpha_s}{2\pi^2} \int_0^{2\nu} \frac{d\left|k^2\right|}{\left|k^2\right|} \int_{\xi_-}^{\xi_+} d\xi \frac{\xi P(\xi)}{\sqrt{(\xi_+ - \xi)(\xi - \xi_-)}} \tag{2.40}$$

where

$$\begin{aligned} \xi_\pm(z,x) &= x + z - 2xz \pm \sqrt{4x(1-x)z(1-z)} \\ z &= \frac{\left|k^2\right|}{2\nu} \quad 0 < z < 1. \end{aligned} \tag{2.41}$$

As already discussed k_T is not restricted to small values. Moreover the integral is logarithmically divergent at small $\left|k^2\right|$. This is the only diagram which gives a logarithmic divergence, the other diagrams gives finite corrections to the structure function.

Introducing a small cut-off κ and considering all contributions from the diagrams of Fig. 2.4 we find

$$\hat{F}_2(x,Q^2) = e_q^2 x \left[\delta(1-x) + \frac{\alpha_s}{2\pi} \left(P(x) \ln \frac{Q^2}{\kappa^2} + C(x) \right) \right] \tag{2.42}$$

where $C(x)$ is a calculable function. Thus, we showed that beyond the leading order, the structure function is Q^2 dependent, with Bjorken scaling broken by logarithms of Q^2. The corresponding quark distribution function to this order in perturbation theory is

$$q(x,Q^2) = \delta(1-x) + \frac{\alpha_s}{2\pi} \left(P(x) \ln \frac{Q^2}{\kappa^2} + C(x) \right). \tag{2.43}$$

Similar expressions can be obtained for the anti-quark and gluon distributions, at NLO and beyond.

2.3 Factorization

Let us now discuss a more convenient approach to handling the singularity at $|k^2| = 0$. In the previous section, we introduced a small cutoff κ^2, but alternative renormalization methods exist. Notably, the singularity arises when the gluon is emitted parallel to the quark ($k_T = 0$). The limit $k_T^2 \to 0$ corresponds to the long-range component of the strong interaction, which is not calculable using perturbation theory.

To obtain a proton structure function, we must convolute the quark structure function $\hat{F}_2$ from Eq. (2.42) with a bare quark distribution q_0 inside the proton and sum over quark flavors, similar to the procedure in the naive parton model. This yields:

$$F_2(x,Q^2) = x \sum_{q,\bar{q}} e_q^2 \left[q_0(x) + \frac{\alpha_S}{2\pi} \int_x^1 \frac{d\xi}{\xi} q_0(\xi) \left\{ P(\frac{x}{\xi}) \ln \frac{Q^2}{\kappa^2} + C(\frac{x}{\xi}) \right\} + \dots \right] \tag{2.44}$$

where the dots represent higher-order perturbative contributions. Here, $q_0(x)$ is an unobservable bare distribution.

The collinear singularities are absorbed into this bare distribution at a factorization scale μ, which plays a role analogous to the renormalization scale. We thus define a renormalized distribution $q(x,\mu^2)$ by:

$$q(x,\mu^2) = q_0(x) + \frac{\alpha_s}{2\pi} \int_x^1 \frac{d\xi}{\xi} q_0(\xi) \left\{ P(\frac{x}{\xi}) \ln \frac{Q^2}{\kappa^2} + C(\frac{x}{\xi}) \right\} + \dots \tag{2.45}$$

which leads to:

$$F_2(x,Q^2) = x \sum_{q,\bar{q}} e_q^2 \int_x^1 \frac{d\xi}{\xi} q(\xi,\mu^2) \times \left\{ \delta(1 - \frac{x}{\xi} + \frac{\alpha_s}{2\pi} P(\frac{x}{\xi}) \ln \frac{Q^2}{\mu^2} + \dots \right\}. \tag{2.46}$$

The distribution $q(x, \mu^2)$ cannot be derived from first principles in perturbation theory, as it includes long-distance strong interaction effects. Instead, it is determined from structure function data at a given scale, since $F_2(x, Q^2) = x \sum e_q^2 q(x, Q^2)$.

The ability to factorize contributions in this way is a fundamental property of the theory. However, while factorization prescribes how to treat the logarithmic singularities, there remains an arbitrariness in handling the finite parts. The extent to which finite contributions are factored out defines what is known as the factorization scheme. A notable scheme is the DIS scheme, in which all finite contributions are absorbed into the quark distribution.

To obtain a complete description of deep inelastic structure functions in terms of parton distributions, we must include an additional contribution: the $O(\alpha_s)$ correction from the process $\gamma^* q \to qg$. The calculation of this contribution follows the same approach as the $\gamma^* q \to qg$ process discussed earlier. The resulting structure function is:

$$\hat{F}2^g(x, Q^2) = x \sum q, \bar{q} e_q^2 \frac{\alpha_S}{2\pi} \left(P_{qg}(x) \ln \frac{Q^2}{\kappa^2} + C_g(x) \right). \tag{2.47}$$

Once again, we observe a logarithmic singularity associated with vanishing quark virtuality. The splitting coefficient in this case is:

$$P_{qg}(x) = T_R[x^2 + (1-x)^2], \tag{2.48}$$

where $T_R = \frac{1}{2}$ is the color factor, and the sum is taken over the n_f massless quark and antiquark flavors that contribute.

To obtain the physical structure function, this contribution must be convoluted with a bare gluon distribution g_0, then added to the expression obtained in Eq. (2.45) as follows:

$$q(x, \mu^2) = q_0(x) + \frac{\alpha_S}{2\pi} \int_x^1 \frac{d\xi}{\xi} q_0(\xi) \left\{ P_{qq}(\frac{x}{\xi}) \ln \frac{Q^2}{\kappa^2} + C_q(\frac{x}{\xi}) \right\} \tag{2.49}$$

$$+ \frac{\alpha_S}{2\pi} \int_x^1 \frac{d\xi}{\xi} g_0(\xi) \left\{ P_{qg}(x) \ln \frac{\mu^2}{\kappa^2} + C_g(\frac{x}{\xi}) \right\} + \ldots. \tag{2.50}$$

In the DIS scheme, we finally obtain:

$$F_2(x, Q^2) = x \sum_{q,\bar{q}} e_q^2 q(x, Q^2), \tag{2.51}$$

where all finite contributions are absorbed into the parton distribution functions.

2.4 Scaling violation and the DGLAP equation

It is important to remark that the calculation of the parton distributions is beyond the scope of perturbation theory. What can be done instead is calculate how these distributions evolve for variation of the factorization scale μ. The right-hand-side of Eq.(2.45) must be μ independent so taking the log derivative of both sides give us a first order differential equation for the μ dependence of $q(x, \mu^2)$ and hence for the Q^2 dependence of the deep inelastic structure function. If we define $t = \mu^2$ and take the $\ln t$ partial derivative of Eq.(2.45) we obtain

$$t \frac{\partial}{\partial t} q(x, t) = \frac{\alpha_S(t)}{2\pi} \int_x^1 \frac{d\xi}{\xi} P\left(\frac{x}{\xi}\right) q(\xi, t). \tag{2.52}$$

This equation is known as the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) equation. More generally, the DGLAP equation is a $(2n_f + 1)$ dimensional matrix equation in the space of quarks, antiquarks and gluon,

$$t \frac{\partial}{\partial t} \left(q_i(x, t) g(x, t) \right) = \frac{\alpha_S(t)}{2\pi} \int_x^1 \frac{d\xi}{\xi} \times \begin{pmatrix} P_{q_i q_j}(\frac{x}{\xi}, \alpha_S(t)) & P_{q_i g}(\frac{x}{\xi}, \alpha_S(t)) \\ P_{g q_j}(\frac{x}{\xi}, \alpha_S(t)) & P_{gg}(\frac{x}{\xi}, \alpha_S(t)) \end{pmatrix} \begin{pmatrix} q_j(\xi, t) \\ g(\xi, t) \end{pmatrix}, \tag{2.53}$$

where each splitting function is calculable as a power series in α_s,

$$P_{q_iq_j}(z,\alpha_s) = \delta_{ij}P^{(0)}_{qq}(z) + \frac{\alpha_s}{2\pi}P^{(1)}_{q_iq_j}(z) + \dots \tag{2.54}$$

$$P_{qg}(z,\alpha_s) = P^{(0)}_{qg}(z) + \frac{\alpha_s}{2\pi}P^{(1)}_{qg}(z) + \dots \tag{2.55}$$

$$P_{gq}(z,\alpha_s) = P^{(0)}_{gq}(z) + \frac{\alpha_s}{2\pi}P^{(1)}_{gq}(z) + \dots \tag{2.56}$$

$$P_{gg}(z,\alpha_s) = P^{(0)}_{gg}(z) + \frac{\alpha_s}{2\pi}P^{(1)}_{gg}(z) + \dots \tag{2.57}$$

Note that because of charge conjugation invariance and $SU(n_f)$ flavor symmetry we have

$$P_{q_iq_j} = P_{\bar{q}_i\bar{q}_j} \tag{2.58}$$

$$P_{q_i\bar{q}_j} = P_{\bar{q}_iq_j} \tag{2.59}$$

$$P_{q_ig} = P_{\bar{q}_ig} = P_{qg} \tag{2.60}$$

$$P_{gq_i} = P_{g\bar{q}_i} = P_{gq} \tag{2.61}$$

i.e. the splitting functions P_{qg} and P_{gq} are independent of the quark flavor and the same for quarks and antiquarks.

The leading order DGLAP splitting functions $P^{(0)}_{ab}(x)$ can be interpreted as the probability of finding a parton type a in a parton type b with a fraction x of the longitudinal momentum of the parent parton and a transverse momentum squared much less than μ^2. This interpretation implies that the splitting functions are positive definite for $x < 1$, and thus satisfy the sum rules

$$\int_0^1 dx\; P^{(0)}_{qq}(x) = 0 \tag{2.62}$$

$$\int_0^1 dx\; x\; [P^{(0)}_{qq}(x) + P^{(0)}_{gq}(x)] = 0 \tag{2.63}$$

$$\int_0^1 dx\; x\; [2n_f P^{(0)}_{qg}(x) + P^{(0)}_{gg}(x)] = 0 \tag{2.64}$$

which correspond to quark number and momentum conservation in the splitting of quarks and gluons respectively.

For the leading order splitting function we have

$$P^{(0)}_{qq}(x) = C_F\left[\frac{1+x^2}{(1-x)_+} + \frac{3}{2}\delta(1-x)\right] \tag{2.65}$$

$$P^{(0)}_{qg}(x) = T_R[x^2 + (1-x)^2] \qquad T_R = \frac{1}{2}, \tag{2.66}$$

$$P^{(0)}_{gq}(x) = C_F\left[\frac{1+(1-x)^2}{x}\right] \tag{2.67}$$

$$P^{(0)}_{gg}(x) = 2C_A\left[\frac{x}{(1-x)_+} + \frac{1-x}{x} + x(1-x)\right] + \delta(1-x)\frac{(11C_A - 4n_fT_R)}{6} \tag{2.68}$$

The problem with the NLO correction is that the flavor structure of the $P_{q_iq_j}$ function is non-trivial. Using $SU(n_f)$ flavor symmetry, we can write this in terms of flavor singlet (S) and non-singlet (V) quantities:

$$P_{q_iq_k} = \delta_{ik}P^V_{qq} + P^S_{qq} \tag{2.69}$$

$$P_{q_i\bar{q}_k} = \delta_{ik}P^V_{q\bar{q}} + P^S_{q\bar{q}} \tag{2.70}$$

$$P^{\pm} = P^V_{qq} \pm P^V_{q\bar{q}} \tag{2.71}$$

At next-to-leading order the functions P^S_{qq} and $P^S_{q\bar{q}}$ are non-zero, but we have the additional relation

$$P^S_{qq} = P^S_{q\bar{q}} \tag{2.72}$$

which simplifies the treatment of the non singlet contributions.

An efficient method for calculating the evolution for individual quark distributions is to introduce a sequence of non-singlet combinations:

$$V_i = q_i^- \tag{2.73}$$
$$T_3 = u^+ - d^+ \tag{2.74}$$
$$T_8 = u^+ + d^+ - 2s^+ \tag{2.75}$$
$$T_{15} = u^+ + d^+ + s^+ - 3c^+ \tag{2.76}$$
$$T_{24} = u^+ + d^+ + s^+ + c^+ - 4b^+ \tag{2.77}$$
$$T_{35} = u^+ + d^+ + s^+ + c^+ + b^+ - 5t^+ \tag{2.78}$$

where

$$q_i^\pm = q_i \pm \bar{q}_i \tag{2.79}$$

and u, d, s, c, b, t are the distributions of the various flavors of quarks. The one remaining combination of quark distributions is the singlet distribution,

$$\Sigma(x,t) = \sum_i q_i^+(x,t) \equiv \sum_i [q_i(x,t) + \bar{q}_i(x,t)] \tag{2.80}$$

whose evolution is coupled with that of the gluon distribution:

$$t\frac{\partial}{\partial t}\begin{pmatrix}\Sigma(x,t)\\ g(x,t)\end{pmatrix} = \frac{\alpha_s(t)}{2\pi}\int_x^1 \frac{d\xi}{\xi} \times \begin{pmatrix}P_{qq}(\frac{x}{\xi},\alpha_s(t)) & 2n_f P_{qg}(\frac{x}{\xi},\alpha_s(t))\\ P_{gq}(\frac{x}{\xi},\alpha_s(t)) & P_{gg}(\frac{x}{\xi},\alpha_s(t))\end{pmatrix}\begin{pmatrix}\Sigma(\xi,t)\\ g(\xi,t)\end{pmatrix} \tag{2.81}$$

where now

$$P_{qq} = P^+ + n_f(P_{qq}^S + P_{q\bar{q}}^S). \tag{2.82}$$

Knowing the 6 V_i combinations, the 5 T_j combinations and Σ one can solve for each of the 12 individual quark and anti-quark distributions.

An alternative formulation of the evolution equations is in terms of the moments (Mellin transforms) of the parton distribution:

$$f(j,t) = \int_0^1 dx\; x^{j-1} f(x,t), \qquad f = q_i, g. \tag{2.83}$$

In terms of these moments, the t dependence of a non-singlet quark distribution function is given by

$$t\frac{\partial}{\partial t} q_{NS}(j,t) = \frac{\alpha_s}{2\pi}\gamma_{qq}(j,\alpha_s(t)) q_{NS}(j,t), \tag{2.84}$$

where the anomalous dimension γ_{qq} is defined as

$$\gamma_{qq}(j,\alpha_s) = \int_0^1 dx\; x^{j-1} P_{qq}(x,\alpha_s). \tag{2.85}$$

This method allows us to simplify the evolution equations reducing the convolution integral to a simple product. Similar equations hold for the evolution of the singlet quark and gluon distributions:

$$t\frac{\partial}{\partial t}\begin{pmatrix}\Sigma(j,t)\\ g(j,t)\end{pmatrix} = \frac{\alpha_s}{2\pi}\begin{pmatrix}\gamma_{qq}(j,\alpha_s(t)) & 2n_f\gamma_{qg}(j,\alpha_s(t))\\ \gamma_{gq}(j,\alpha_s(t)) & \gamma_{gg}(j,\alpha_s(t))\end{pmatrix}\begin{pmatrix}\Sigma(j,t)\\ g(j,t)\end{pmatrix} \tag{2.86}$$

where $\Sigma(j,t)$ and $g(j,t)$ are the moments of the singlet quark and gluon distributions respectively. The complete

set of leading order anomalous dimensions is

$$\gamma_{qq}^{(0)}(j) = C_F\left[-\frac{1}{2} + \frac{1}{j(j+1) - 2\sum_{k=2}^{j}\frac{1}{k}}\right] \tag{2.87}$$

$$\gamma_{qg}^{(0)}(j) = T_R\left[\frac{(2+j+j^2)}{j(j+1)(j+2)}\right] \tag{2.88}$$

$$\gamma_{gq}^{(0)}(j) = C_F\left[\frac{(2+j+j^2)}{j(j^2-1)}\right] \tag{2.89}$$

$$\gamma_{gg}^{(0)}(j) = 2C_A\left[-\frac{1}{12} + \frac{1}{j(j-1)} + \frac{1}{(j+1)(j+2)} - \sum_{k=2}^{j}\frac{1}{k}\right] - \frac{2}{3}n_f T_R. \tag{2.90}$$

Using the results

$$\int_0^1 dx\, x^{j-1}\frac{1}{x} = \frac{1}{j-1} \tag{2.91}$$

$$\int_0^1 dx\, x^{j-1}\frac{1}{(1-x)_+} = -\int_0^1 dx\, \frac{x^{j-1}-1}{(x-1)} \sim -\ln j, \quad \text{for } j \to \infty, \tag{2.92}$$

we see that all the anomalous dimensions have poles at $j = 1$, and that the diagonal anomalous dimensions grow as $\ln(j)$ at large j.

2.4.1 Solution of the leading order DGLAP equations

The solution of the DGLAP equation is simplest for (flavor) non-singlet combinations of quark distributions, e.g. $V = q_i - q_j$. In such combinations the mixing with the flavor singlet gluon distribution dropout and we have simply

$$t\frac{\partial}{\partial t}V(x,t) = \frac{\alpha_s(t)}{2\pi}[P_{qq}(\xi) \otimes V(z,t)] \tag{2.93}$$

where we have introduced $\otimes$ as a shorthand notation for the convolution integral. We can again write this equation in terms of moments:

$$t\frac{\partial}{\partial t}V(j,t) = \frac{\alpha_s(t)}{2\pi}\gamma_{qq}^{(0)}(j)V(j,t) \tag{2.94}$$

with $\gamma_{qq}^{(0)}(j)$ given in Eq. (2.90). Inserting the lowest order form for the running couplings $\alpha_s = ccccc$, we obtain the solution for the moments of non-singlet distribution,

$$V(j,t) = V(j,t_0)\left(\frac{\alpha_s(t_0)}{\alpha_s(t)}\right)^{d_{qq}(j)}, \quad d_{qq}(j) = \frac{\gamma_{qq}^{(0)}}{2\pi b} \tag{2.95}$$

where $b = \frac{(33-2n_f)}{12\pi}$. Finally, the distribution in x space can be obtained using the inverse Mellin transform integral,

$$V(x,t) = \frac{1}{2\pi i}\int_C dj\, x^{-j}V(j,t) \tag{2.96}$$

where the integration contour in the complex j plane is parallel to the imaginary axis and to the right of all singularities of the integrand. Except for some very special cases, the inverse Mellin transform has to be performed by numerical integration. It can be shown that $d_{qq}(1) = 0$ and that $d_{qq}(j) < 0$ for $j \geq 2$. This implies that as μ^2 increases the non-singlet distribution function decreases at large x and increases at small x. Physically, this can be understood as an increase in the phase space for the gluon emission by the quarks as μ^2 increases, with a corresponding reduction in quark momentum. We now turn to the flavor singlet combination of quark distributions. At leading order we have

$$t\frac{\partial\Sigma}{\partial t} = \frac{\alpha_s(t)}{2\pi}[P_{qq}^{(0)} \otimes \Sigma + 2n_f P_{qg}^{(0)} \otimes g] + O(\alpha_s^2(t)) \tag{2.97}$$

$$t\frac{\partial g}{\partial t} = \frac{\alpha_s(t)}{2\pi}[P_{gq}^{(0)} \otimes \Sigma + P_{gg}^{(0)} \otimes g] + O(\alpha_s^2(t)) \tag{2.98}$$

These equations are most easily solved by direct numerical integration in x space starting with input distributions obtained from data. From the solution of the evolution equations for the moments of non-singlet and singlet combinations of quark distributions, the evolution of the moments of any individual flavor of quark distribution can be determined. The x distributions themselves are then obtained by an inverse Mellin transformation,

$$f_a(x,\mu^2) = \frac{1}{2\pi i}\int_C djx^{-j} f_a(j,\mu^2), \quad a = q_i, g \tag{2.99}$$

Now it's better to stop for a moment and summarize what we have seen so far. Starting from Deep inelastic scattering we were able to introduce some central concept for this thesis project like the factorization property of QCD and most importantly the main object of this study, the Parton density functions. In the next chapter we will see why the determination of these PDFs is important and how they are usually found.

Chapter 3

PDF determination, SIDIS, and FFs

In Sect. 2.4, we established that perturbative QCD (pQCD) predicts the Q^2 evolution of PDFs rather than determining their shape directly. A precise and reliable determination of PDFs and their uncertainties is a crucial component for future advancements in hadronic physics. This is achieved through global fits of datasets obtained from hadronic collision experiments, where the quality of the final result is directly dependent on the accuracy and breadth of the data used in the fits. Modern PDF determinations include a number of experimental measurements for a wide array of hard-scattering processes, which however do not typically cover SIDIS. The theoretical description of SIDIS in terms of QCD factorization is indeed complicated by the fact that an additional non-perturbative object, the FF of a parton hadronizing in the final-state hadron, is required.

The structure of this Chapter is as follows. In Sect. 3.1, I explain how PDFs are determined from a global analysis of a wide array of experimental measurements. In Sect. 3.2, I introduce SIDIS and I discuss its relevance to possibly constrain PDFs. In Sect. 3.3, I dive into the concept of FF. Finally, in Sect. 3.4, I review the available SIDIS experimental data that will enter my analysis in the Chapter 5.

3.1 PDF determination

In this section, we outline the typical procedure for determining PDFs and fragmentation functions, illustrated in Fig. 3.1. The general approach begins with constructing a parametrization of the parton distribution functions at a reference scale Q_0. This parametrization can, for example, be implemented using a Neural Network. Using the DGLAP evolution equations, we can then determine the PDFs at any energy scale perturbatively.

Given such a parametrization, theoretical predictions can be computed and compared against datasets from various hadronic experiments. Since PDFs are universal functions that depend only on the parton type they describe, data from multiple processes can be included in the fit. Based on these comparisons, an optimization algorithm iteratively updates the initial parameters to improve the agreement between predictions and data. The process continues until the algorithm reaches convergence, yielding a final set of parameters that best describe the PDFs.

To estimate PDF uncertainties, it is essential to propagate errors from the data space to the parameter space. A widely used approach for this purpose is the Monte Carlo method [5]. The underlying assumption is that the data follow a multivariate Gaussian distribution:

$$\mathcal{G}(\mathbf{x}^{(k)}) \propto \exp\left[(\mathbf{x}^{(k)} - \boldsymbol{\mu})^T \cdot C^{-1} \cdot (\mathbf{x}^{(k)} - \boldsymbol{\mu})\right], \tag{3.1}$$

where C is the covariance matrix, $\mathbf{x}^{(k)} = (x_1^{(k)}, x_2^{(k)}, \ldots, x_{N_{dat}}^{(k)})$ are equally probable replicas ($k = 1, \ldots, N_{rep}$) of a set of N_{dat} measured quantities, and $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_{N_{dat}})$ represents the expectation values corresponding to the measured data.

The elements of the covariance matrix are defined as:

$$C_{ij} = \delta_{ij}\sigma_{i,unc}^2 + \sum \beta \sigma_{i,corr}^{(\beta)} \sigma_{j,corr}^{(\beta)}, \tag{3.2}$$

where $\sigma_{i,unc}$ represents the total uncorrelated uncertainty of the i-th data point, and $\sigma_{i,corr}^{(\beta)}$ denotes the correlated uncertainty from source β.

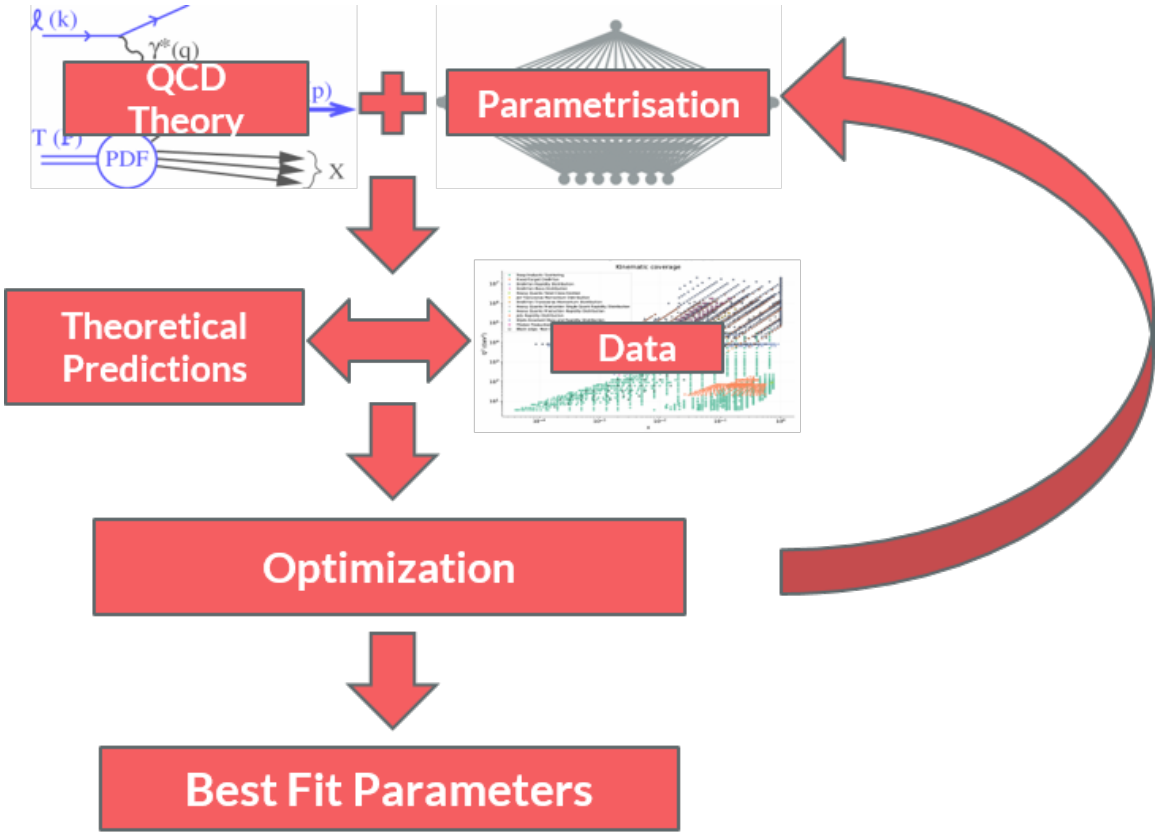

Figure 3.1: General procedure for global fits of PDFs and fragmentation functions.

To generate N_{rep} data replicas, we sample from this distribution using the Cholesky decomposition $\mathbf{L}$ of the covariance matrix, where $\mathbf{C} = \mathbf{L} \cdot \mathbf{L}^T$. The data replicas are then computed as:

$$\boldsymbol{x}^{(k)} = \boldsymbol{\mu} + \boldsymbol{L} \cdot \boldsymbol{r}^{(k)}, \tag{3.3}$$

where $\boldsymbol{r}^{(k)}$ is an N_{dat}-dimensional normal random vector.

For a sufficiently large number of replicas, the sampled data satisfy the following statistical properties:

$$\frac{1}{N_{rep}} \sum_{k}^{N_{rep}} x_i^{(k)} \simeq \mu_i, \qquad \frac{1}{N_{rep}} \sum_{k}^{N_{rep}} x_i^{(k)} x_j^{(k)} \simeq \mu_i \mu_j + C_{ij}. \tag{3.4}$$

Repeating the fitting procedure for each replica results in N_{rep} sets of PDF parameters. From these fits, PDF replicas are constructed, each represented on a discrete grid. These grids contain PDF values at various energy scales and x values, obtained through DGLAP evolution. Specific values of a PDF replica are extracted via interpolation.

One way to define a parametrization is through Neural Networks, which function as flexible function generators. In this approach, the optimization algorithm minimizes an objective function known as the cost function. During training, the neural network explores the parameter space, guided by the optimization algorithm, and updates its parameters iteratively to refine predictions.

However, this process involves balancing two competing effects. On the one hand, an overly complex neural network may overfit the dataset, memorizing the data rather than capturing its underlying structure. This results in a low training error but poor generalization, as the model becomes highly sensitive to noise. On the other hand, an overly simple network fails to adequately fit the data, leading to poor generalization as well. To address this, an appropriate stopping criterion is incorporated into the optimization algorithm. One commonly used technique is early stopping, where training is halted if the generalization error increases for a predetermined number of steps. In this framework, the cost function typically used is the χ^2, defined as [5]:

$$\chi^2 \equiv \left(\boldsymbol{T}(\boldsymbol{\theta}^{(k)}) - \boldsymbol{x}^{(k)} \right)^T \cdot C^{-1} \cdot \left(\boldsymbol{T}(\boldsymbol{\theta}^{(k)}) - \boldsymbol{x}^{(k)} \right), \tag{3.5}$$

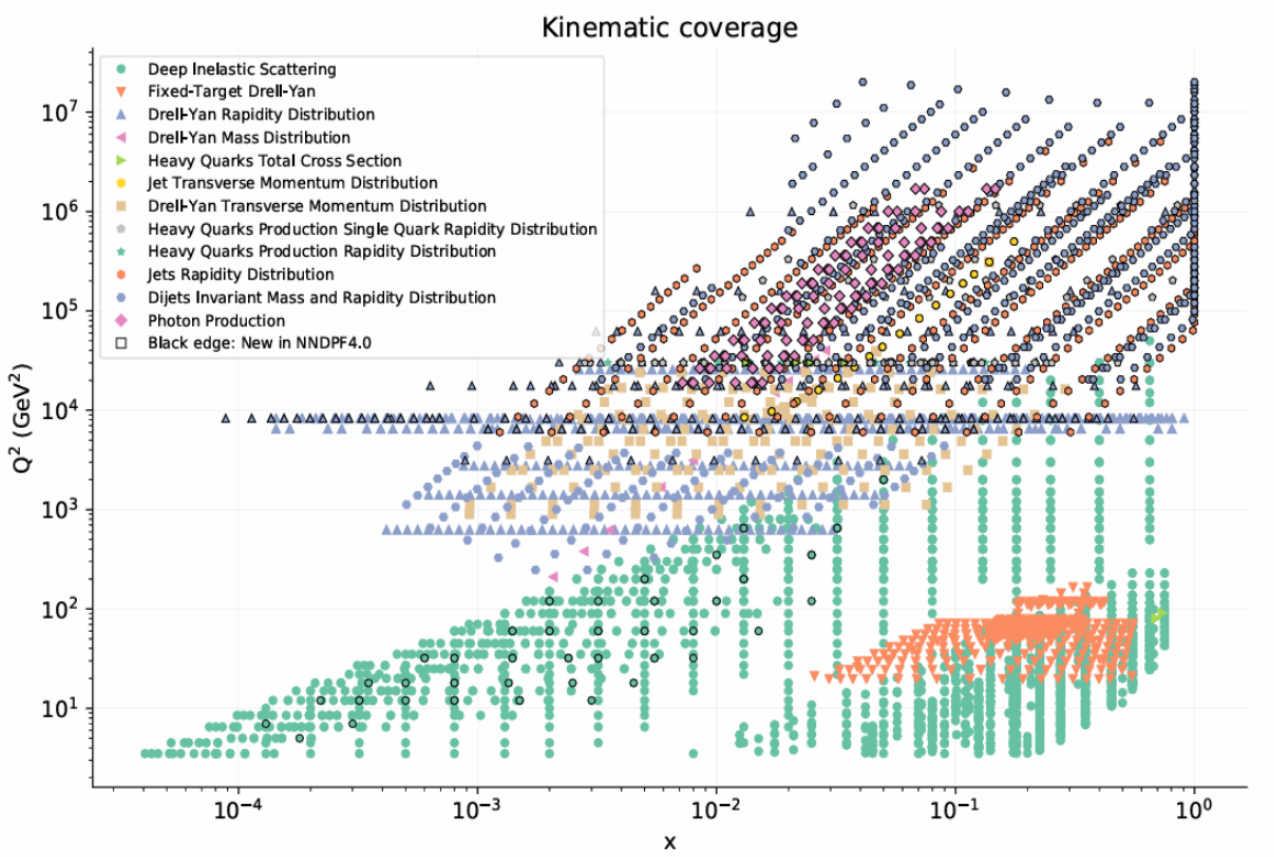

Figure 3.2: Datasets included for PDF determination. The main contributions come from DIS, Drell-Yan, and SIA experiments [6].

where $\boldsymbol{T}(\boldsymbol{\theta}^{(k)})$ represents the theoretical predictions based on the neural network parametrization.

The goal of the optimization algorithm is to find the set of parameters that minimizes χ^2, ensuring optimal agreement between theory and experiment. Fig. 3.2 illustrates the kinematic coverage of datasets typically included in global PDF fits. Most data originate from Deep Inelastic Scattering (DIS), Semi-Inclusive Annihilation (SIA), and Drell-Yan processes. Additional processes can provide valuable constraints on PDFs, one of which is the focus of this thesis: Semi-Inclusive Deep Inelastic Scattering (SIDIS). In the next section, we will explore the SIDIS process, examining its relevance in PDF determination and discussing why it has historically been excluded from standard fits.

3.2 Semi Inclusive Deep Inelastic Scattering

In Fig. 3.2, we observed the most commonly used processes for determining Parton Distribution Functions (PDFs). However, one significant process is missing: Semi-Inclusive Deep Inelastic Scattering (SIDIS).

Let us begin by considering the following reaction:

$$l(k) + N(P) \rightarrow l(k^{'}) + h(P_h) + X, \tag{3.6}$$

where l denotes the beam lepton, N represents the nucleon target, h is the produced hadron, and the four-momenta of these particles are given in parentheses. The masses of the nucleon and the hadron h are denoted by M and M_h, respectively.

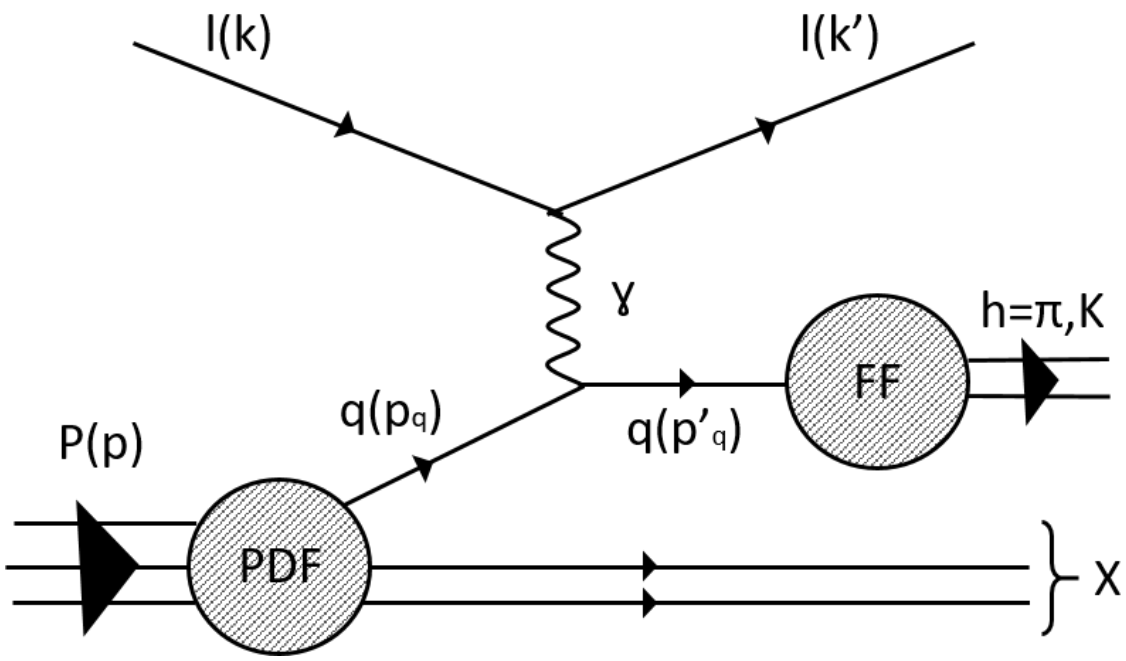

Figure 3.3: Feynman diagram for Semi-Inclusive Deep Inelastic Scattering. Unlike in DIS, both the scattered lepton and a final-state hadron are detected [7].

As discussed in Sect. 2.1, we can describe this process using the following kinematic variables:

$$\begin{aligned} Q^2 &= -q^2, \\ x &= \frac{Q^2}{2\nu} = \frac{Q^2}{2M(E-E')}, \\ z &= \frac{p \cdot p_\pi}{p \cdot q}, \\ y &= \frac{q \cdot p}{k \cdot p} = 1 - \frac{E^{'}}{E}, \end{aligned} \tag{3.7}$$

where Q^2 is the invariant mass of the virtual vector boson, x represents the fraction of the nucleon's momentum carried by the incoming parton, z is the fraction of the outgoing parton's momentum carried by the observed hadron h, and y is the inelasticity parameter.

The SIDIS cross-section can be expressed in terms of structure functions, similar to the case of DIS. A graphical representation of SIDIS is shown in Fig. 3.3.

In the following discussion, we focus on a specific SIDIS process: the inclusive production of a charged pion, $\pi^{\pm}$, in lepton-nucleon scattering:

$$l(k) + N(p) \rightarrow l(k^{'}) + \pi^{\pm}(p_{\pi^{\pm}}) + X. \tag{3.8}$$

Under the assumption $Q \ll M_Z$ (which holds for all SIDIS data considered in this project), only the exchange of a virtual photon is relevant. The triple-differential SIDIS cross-section is given by:

$$\frac{d^3\sigma}{dxdQdz} = \frac{4\pi\alpha^2}{xQ^3}[(1+(1-y)^2)F_2(x,z,Q^2) - y^2 F_L(x,z,Q^2)], \tag{3.9}$$

where α is the fine-structure constant, and F_2 and F_L are dimensionless structure functions.

Within the framework of collinear factorization, structure functions can be expressed as:

$$F_i(x,z,Q) = x\sum_{q\bar{q}} e_q^2 \Big\{ [C_{i,qq}(x,z,Q) \otimes f_q(x,Q) + C_{i,qg}(x,z,Q) \otimes f_g(x,Q)] \otimes D_q^{\pi^{\pm}}(z,Q) \tag{3.10}$$

$$+ [C_{i,gq}(x,z,Q) \otimes f_q(x,Q)] \otimes D_g^{\pi^{\pm}}(z,Q) \Big\}, \tag{3.11}$$

where the convolution symbol $\otimes$ is defined as:

$$C(x,z)\otimes f(x)\otimes D(z)=\int_x^1\frac{dx^{'}}{x^{'}}\int_z^1\frac{dz^{'}}{z^{'}}C(x^{'},z^{'})f(\frac{x}{x^{'}})D(\frac{z}{z^{'}}). \tag{3.12}$$

The sum in Eq. (3.10) runs over all active quark and antiquark flavors at the scale Q. Here, e_q represents the electric charge of the quark flavor q, C_i are perturbatively calculable coefficient functions, and $f_{q(g)}$ denote the collinear quark (gluon) PDFs.

The description of the SIDIS cross-section necessitates the introduction of an additional non-perturbative quantity, the Fragmentation Functions (FFs) $D^{\pi^{\pm}}_{q(g)}$. These functions describe the hadronization process, where a parton fragments into a hadron. We will examine FFs in greater detail later, but for now, we define them as functions that characterize the probability of a parton producing a particular hadron in the final state.

Returning to the cross-section expression, the coefficient functions C in Eq. (3.10) admit a perturbative expansion:

$$C(x,z,Q)=\sum_{n=0}\left(\frac{\alpha_s(Q)}{4\pi}\right)^n C^{(n)}(x,z), \tag{3.13}$$

where the reference value of the strong coupling constant is taken as $\alpha_s(M_Z)=0.118$.

A crucial property of the perturbative coefficients C is that, for $n=0,1$, the functions $C^{(n)}(x,z)$ can be expressed as bilinear combinations of single-variable functions:

$$C^{(n)}(x,z)=\sum_t c_t O^{(1)}t(x)O^{(2)}t(z), \tag{3.14}$$

where c_t are numerical coefficients. This property allows us to separate the double convolution integral in Eq. (3.12) into a linear combination of single integrals:

$$C^{(n)}(x,z)\otimes f(x)\otimes D(z)=\sum_t c_t\left[O_t^{(1)}(x)\otimes f(x)\right]\left[O_t^{(2)}(z)\otimes D(z)\right]. \tag{3.15}$$

This observation significantly accelerates the numerical computation of SIDIS cross-sections.

To assess the impact of SIDIS data on PDFs, we must also determine the non-perturbative Fragmentation Functions (FFs). Therefore, in the next section, we will study these functions in greater detail, following a similar approach to our discussion of PDFs.

3.3 Fragmentation Functions

We have seen in Sect. 2.3 that factorization theorems allow the separation of the perturbatively calculable part of the cross-section from the non-perturbative contributions.

When specific particles are identified in the final state, parton fragmentation functions (FFs) frequently appear as non-perturbative ingredients in QCD factorization formulas. In this section, we introduce some fundamental concepts related to FFs. For more detailed discussions, see Ref.[8].

Fragmentation functions describe how color-carrying quarks and gluons transform into color-neutral particles, such as hadrons or photons. The most well-studied FF is $D_1^{h/i}(z)$, which characterizes the fragmentation of an unpolarized parton of type i into an unpolarized hadron of type h. Here, z represents the fraction of the parton's momentum carried by the hadron along the parton's direction of motion. Consequently, $D_1^{h/i}(z)$ is often referred to as the collinear fragmentation function.

Based on this definition, we can interpret $D_1^{h/i}(z)dz$ as the number of hadrons h produced from a parton i within the momentum fraction range $[z,z+dz]$. However, as with parton densities, this intuitive interpretation holds only at low perturbative orders. Beyond leading order, while the factorization formula remains valid, the direct probability interpretation no longer applies.

For the purposes of this thesis, we focus on the unpolarized integrated fragmentation function $D_1^{h/i}(z)$.

As in the case of parton distributions, the fundamental definitions of fragmentation functions contain ultraviolet (UV) divergences. The factorization procedure employs renormalized FFs, which are related to the bare fragmentation functions through a renormalization formula of the form:

$$D^{h/i}(z;\mu)=\lim_{\epsilon\to 0}\sum_{j'}\int_{z^-}^{1^+}\frac{d\rho}{\rho}D^{h/i}_{(0)}(z/\rho)L_{j'j}(\rho;g(\mu),\epsilon), \tag{3.16}$$

where $\epsilon = 2 - n/2$ with n being the space-time dimension.

The bare distribution can be expressed as:

$$D^{h/i}(z) = \frac{1}{12\pi} \sum_X \int dy^- e^{-i\frac{P^+}{z}y^-} \mathrm{Tr}\big[\gamma^+ \langle 0 \mid \psi(y)\mathcal{P} \mid h(P)X\rangle\langle h(P)X \mid \mathcal{P}^{'}\overline{\psi}(0) \mid 0\rangle\big], \tag{3.17}$$

where we have used light-cone coordinates, defined as $y^{\pm} = (y^0 \pm y^z)/\sqrt{2}$, and $\mathcal{P}$ and $\mathcal{P}^{'}$ denote appropriate gauge links.

We now explore the evolution equations governing fragmentation functions and their role in QCD phenomenology.

In general, for a given process where a hadron is detected in the final state, there exist 11 different fragmentation functions (FFs), one for each quark, antiquark, and the gluon. For example, in the case of the SIDIS process described in Eq. (3.8), we would have two sets of FFs: one for π^+ and one for π^-. Fortunately, the number of independent FFs can be reduced by exploiting symmetries such as charge conjugation and isospin symmetry.

Considering the fragmentation of up quarks, down quarks, and gluons into pions, charge conjugation symmetry leads to the following exact relations:

$$D_1^{\pi^+/u} = D_1^{\pi^-/\overline{u}}, \quad D_1^{\pi^+/\overline{u}} = D_1^{\pi^-/u}, \quad D_1^{\pi^+/d} = D_1^{\pi^-/\overline{d}}, \quad D_1^{\pi^+/\overline{d}} = D_1^{\pi^-/d}, \quad D_1^{\pi^+/g} = D_1^{\pi^-/g}. \tag{3.18}$$

Additionally, isospin symmetry of the strong interaction provides the following relations:

$$D_1^{\pi^+/u} = D_1^{\pi^-/d}, \qquad D_1^{\pi^+/d} = D_1^{\pi^-/u}, \tag{3.19}$$

which are only broken by small electromagnetic effects.

For integrated FFs, there is no debate regarding their universality, i.e., their independence from the specific process in which they are measured. It is generally assumed that $D_1^{h/i}(z)$ remains the same across different processes such as e^+e^- annihilation, SIDIS, and hadronic collisions.

Due to QCD dynamics, FFs depend on an additional parameter: the renormalization scale μ.

The evolution equations for unpolarized integrated FFs take the general form:

$$\frac{d}{d\ln\mu^2} D_1^{h/i}(z,\mu^2) = \frac{\alpha_s(\mu^2)}{2\pi} \sum_j \int_z^1 \frac{du}{u} P_{ij}(u,\alpha_s(\mu^2)) D_1^{h/j}\left(\frac{z}{u},\mu^2\right), \tag{3.20}$$

which is structurally similar to the evolution equations for PDFs. The main difference is that, in the case of FFs, the time-like splitting functions P_{ji} appear instead of the space-like splitting functions P_{ij} found in the case of PDFs.

Typically, the system of evolution equations is decomposed into flavor non-singlet and flavor singlet sectors. The splitting functions P_{ji} have a perturbative expansion of the form:

$$P_{ij}(u,\alpha_s(\mu)) = P_{ij}^{(0)}(u) + \frac{\alpha_s(\mu^2)}{2\pi} P_{ij}^{(1)}(u) + \left(\frac{\alpha_s(\mu^2)}{2\pi}\right)^2 P_{ij}^{(2)}(u) + \ldots \tag{3.21}$$

where the leading-order (LO) splitting functions $P_{ji}^{(0)}$ coincide with the well-known LO space-like DGLAP splitting functions.

3.4 SIDIS facilities and experiments

For the inclusion of the SIDIS process in the determination of PDFs, we consider data from two main experiments: the COMPASS experiment at CERN and the HERMES experiment at DESY.

COMPASS utilizes a muon beam with energy $E_\mu = 160$ GeV and a ^{6}LiD target, while HERMES employs electron and positron beams with an energy of $E_e = 27.6$ GeV using hydrogen or deuterium targets.

The quantity measured by both HERMES and COMPASS is not the absolute cross-section but rather an integrated multiplicity, defined as:

$$\frac{dM}{dz} = \left[\int_{Q_{min}}^{Q_{max}} dQ \int_{x_{min}}^{x_{max}} dx \int_{z_{min}}^{z_{max}} dz \frac{d^3\sigma}{dxdQdz}\right] \Bigg/ \left[\delta z \int_{Q_{min}}^{Q_{max}} dQ \int_{x_{min}}^{x_{max}} dx \frac{d^2\sigma}{dxdQ}\right], \tag{3.22}$$

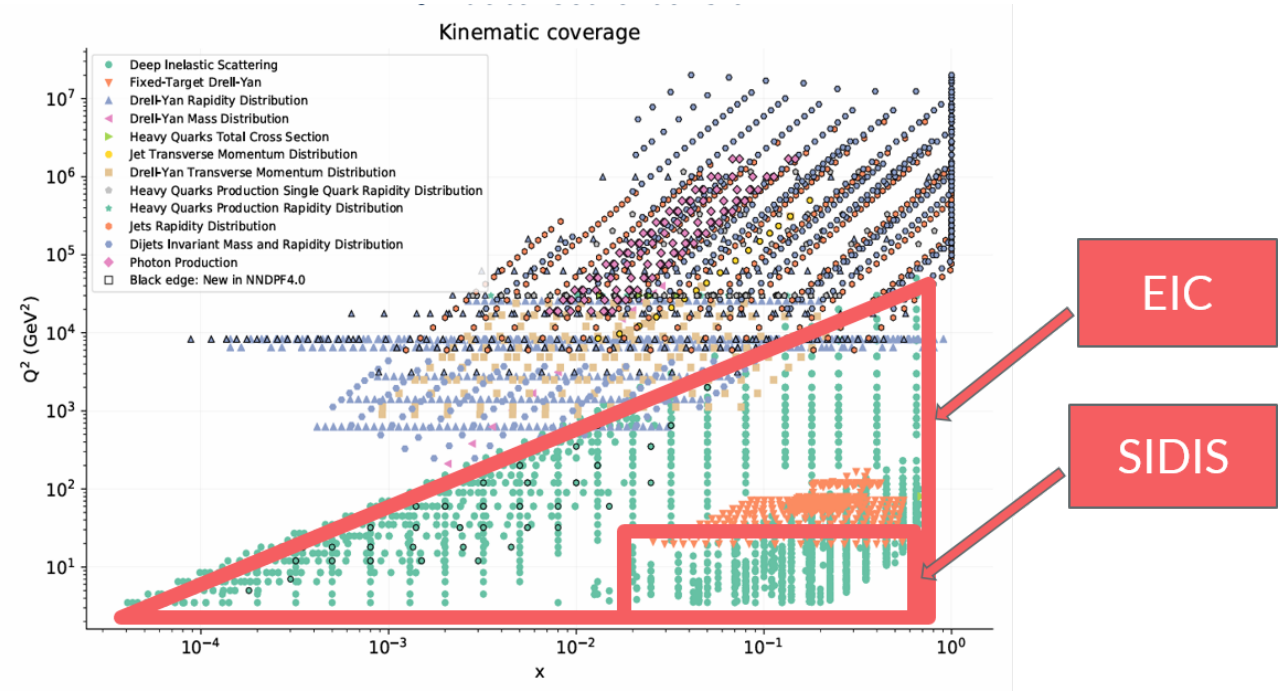

Figure 3.4: Kinematic regions covered by current SIDIS experiments and the future EIC.

where the integration limits define the specific kinematic bin, and $\delta z = z_{max} - z_{min}$. The denominator corresponds to the DIS cross-section inclusive with respect to the final state, which is independent of the FFs.

While the multiplicities measured by HERMES are binned in the variables x, Q^2, z, exactly matching Eq. (3.22), those from COMPASS are binned in x, y, z. In the latter case, theoretical predictions require adjusting the integration limits in Q and x:

$$Q_{min} = \sqrt{x_{min} y_{min} s}, \qquad Q_{max} = \sqrt{x_{max} y_{max} s}, \tag{3.23}$$

and

$$x_{min} \to \max\left[x_{min}, \frac{Q^2}{y_{max} s}\right], \qquad x_{max} \to \min\left[x_{max}, \frac{Q^2}{y_{min} s}\right], \tag{3.24}$$

where y_{min} and y_{max} define the bin boundaries in y.

Furthermore, both HERMES and COMPASS measure cross-sections within a specific fiducial region given by:

$$W = \sqrt{\frac{(1-x)Q^2}{x}} \geq W_{low}, \qquad y_{low} \leq y \leq y_{up}, \tag{3.25}$$

where the values of W_{low}, y_{low}, and y_{up} are listed in Table 1. These constraints reduce the phase space for some bins located at the boundaries of the fiducial region. Consequently, the integration limits in Eq. (3.22) are modified as follows:

$$x_{min} \to \overline{x}_{min} = \max\left[x_{min}, \frac{Q^2}{s y_{up}}\right], \qquad x_{max} \to \overline{x}_{max} = \min\left[x_{max}, \frac{Q^2}{s y_{low}}, \frac{Q^2}{Q^2 + W^2_{low}}\right], \tag{3.26}$$

where x_{min} and x_{max} are interpreted as in Eq. (3.24) for COMPASS.

In our determination of FFs, all integrals in Eq. (3.22) are computed explicitly during the fit. The effect of performing these integrations, compared to evaluating cross-sections at the central bin points, is relatively small for COMPASS but significant for HERMES. However, in both cases, proper integration improves the description of the data.

Both HERMES and COMPASS measure multiplicities separately for π^+, π^-, K^+, and K^-. However, charge conjugation symmetry relates these distributions, allowing one to be obtained from the other by exchanging quark and antiquark distributions while keeping the gluon unchanged:

$$D^{h^-}_{q(\overline{q})}(x, Q) = D^{h^+}_{\overline{q}(q)}(x, Q), \qquad D^{h^-}_{g}(x, Q) = D^{h^+}_{g}(x, Q). \tag{3.27}$$

Now that we have introduced the necessary background on SIDIS and fragmentation functions, we can address the questions raised at the end of the previous section: Why is SIDIS relevant for studying PDFs, and why has it not been included in PDF fits until now?

First, SIDIS explores a kinematic region that is currently not well covered by other processes. Additionally, the interplay between PDFs and FFs in SIDIS provides valuable insights that could enhance our understanding of PDFs. Furthermore, a new experiment under construction, the Electron-Ion Collider (EIC), will primarily investigate SIDIS in an even larger kinematic region than HERMES and COMPASS. Assessing the impact of existing SIDIS datasets on PDFs now can provide useful information for optimizing the future experimental program. The kinematic regions covered by current SIDIS experiments and the planned EIC are shown in Fig. 3.4.

Regarding the second question, SIDIS cross-sections can be expressed using factorization theorems as:

$$\sigma^{lN\to lhX} = \hat{\sigma} \otimes PDF \otimes FF, \tag{3.28}$$

where $\hat{\sigma}$ is the partonic cross-section computed from Feynman diagrams, f represents the PDFs, and D represents the FFs. Each of these contributions must be evaluated at the same perturbative order. The NNLO calculations of partonic cross-sections for SIDIS have only been made available in recent years [9, 10, 11].

Moreover, both COMPASS and HERMES have completed their experimental programs [12, 13]. This makes it an opportune time for the inclusion of SIDIS data at NNLO in global PDF fits.

Chapter 4

Methodology

In Chap. 3, we discussed how factorization theorems in QCD allow us to compute a class of observables, such as cross-sections, by separating perturbative and non-perturbative contributions. The process of interest in this study is SIDIS, where the non-perturbative contributions consist of Parton Distribution Functions (PDFs) and Fragmentation Functions (FFs). In the cross-section formula of Eq. (3.9), these two contributions are intertwined, meaning that, ideally, they should be fitted simultaneously. However, this is a highly complex task, and no definitive solution currently exists.

A common approach is to introduce an iterative procedure in which each contribution is determined separately. One strategy is a sequential fit, where FFs are determined first and then used to fit PDFs while keeping the FFs fixed at their central values [1, 2]. Another approach is to fit the FFs and assess their impact on the PDFs using a technique called reweighting, thereby avoiding a second fit [3].

A recent study [14] applies a fragmentation function fit in combination with reweighting to evaluate the impact of SIDIS on PDFs. Although similar to the procedure developed in this project, that approach does not incorporate multiple iterations, assuming that all relevant information is successfully integrated into the PDFs in a single iteration.

The structure of this chapter is as follows. In Sect. 4.1, I will discuss the strengths of this project and introduce the iterative procedure. From Sect. 4.2 to Sect. 4.6, I will provide a step-by-step description of the method, with a particular emphasis on the application of reweighting (Sect. 4.5). Finally, in Sect. 4.7, I will explore the concept of statistical distances and their significance in this analysis.

4.1 Iterative Procedure

The iterative procedure designed to evaluate the impact of SIDIS data on PDFs consists of the following steps.

I start with a complete set of PDFs provided by the NNPDF collaboration [6] and perform a fit of the fragmentation functions (FFs) while keeping the PDFs fixed at their central values.

Next, I compute theoretical predictions for SIDIS data using each PDF replica while keeping the FFs fixed. These predictions are then used in the reweighting procedure, which allows us to assess the impact of the new datasets on the PDFs. Finally, I apply another technique called unweighting to generate a new PDF set that incorporates the information obtained from reweighting while maintaining the standard format for PDF sets.

This approach offers two significant advantages compared to previous works. First, it enables the use of perturbative corrections at NNLO, which were unavailable at the time for many earlier studies. Second, the dataset used for determining the PDF set in sequential fits is more limited compared to the comprehensive dataset employed by the NNPDF collaboration. Additionally, using only a single iteration in the fitting process may lead to less precise and potentially inconsistent results. Therefore, this project is conducted at the highest available perturbative order and employs the most extensive dataset for both PDF and FF determination, while also incorporating multiple iterations for improved accuracy.

Regarding the perturbative contributions to SIDIS, we account for both the partonic cross-sections, which are computed at NNLO, and the scale dependence of PDFs and FFs as determined by evolution equations.

The steps described above are part of an iterative procedure, illustrated in Fig.4.1, which is structured as follows:

1. Fit of fragmentation functions at NLO and NNLO
2. Construction of LHAPDF grids

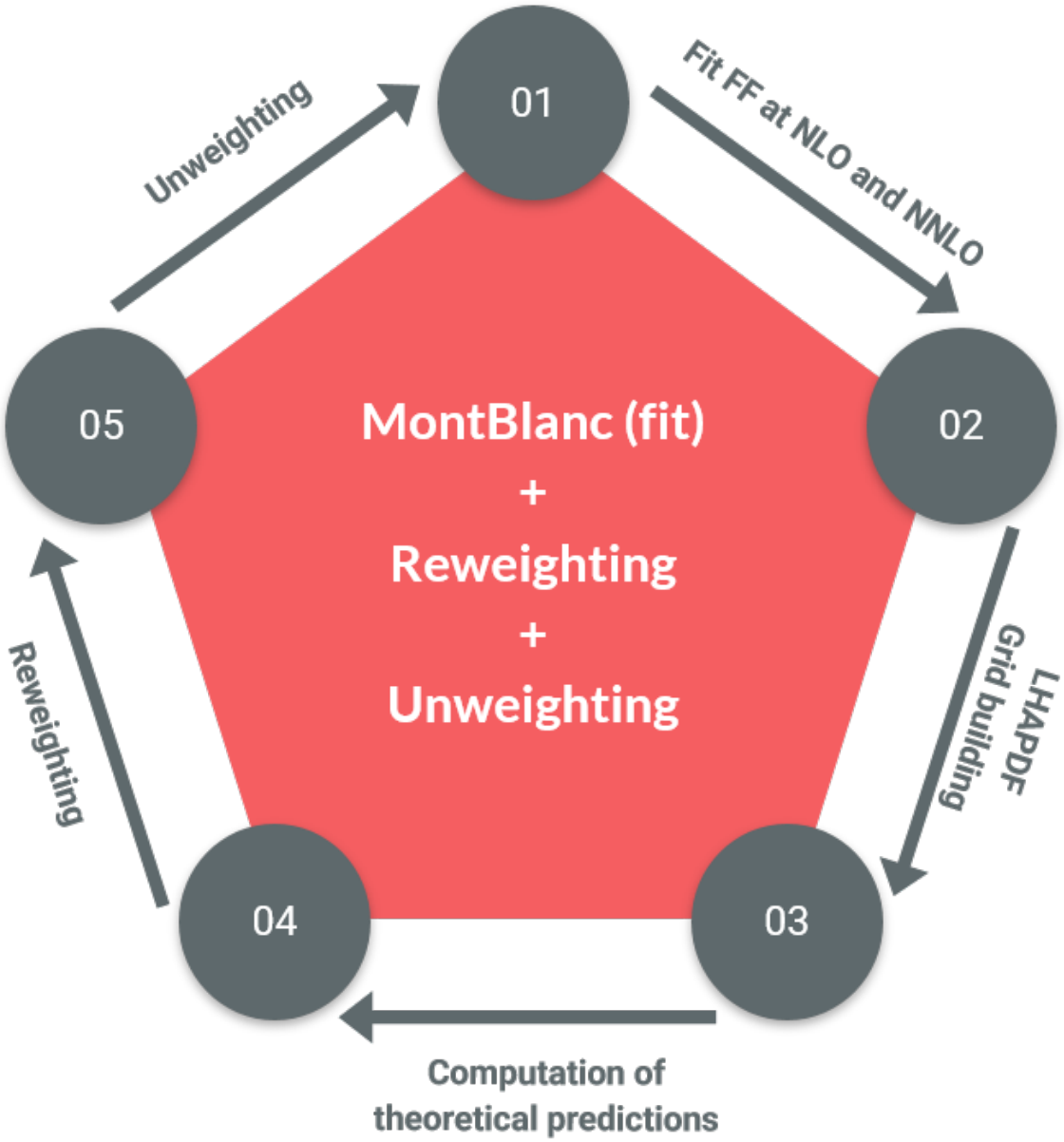

Figure 4.1: Iterative procedure developed in this thesis for studying the impact of SIDIS data on PDFs.

3. Evaluation of theoretical predictions
4. Application of reweighting
5. Application of unweighting

In the following sections, each of these steps will be examined in detail. A significant part of the project utilizes the MontBlanc code [15], developed by the MAP Collaboration, for fitting, grid construction, and the computation of theoretical predictions. These codes were adapted to meet the specific requirements of this study. For the implementation of the reweighting and unweighting procedures, I developed two standalone codes that utilize the results obtained from the previous steps.

4.2 Fit of fragmentation functions

The first step of the procedure consists of a fit of fragmentation functions (FFs). In this section, we present some technical details of the MontBlanc code, as discussed in [5].

The data included in the fits come from Single-Inclusive Annihilation (SIA) and SIDIS measurements conducted by various experiments. The SIA data were collected by CERN (ALEPH, DELPHI, OPAL), DESY (TASSO, BELLE, TOPAZ), and SLAC (BABAR, TPC, SLD), while SIDIS data were obtained from CERN (COMPASS) and DESY (HERMES).

In Sect.3.4, we discussed how charge conjugation symmetry allows us to reduce the number of independent FFs. Specifically, we express the π^- (K^-) FFs in terms of the π^+ (K^+) FFs, effectively extracting only the latter.

The fitting algorithm relies on a Monte Carlo approach to generate a set of N_{rep} FF replicas, which are parameterized using a neural network. The input variable for the network is the momentum fraction z carried by the final hadron. To determine the best set of independent FF combinations in the parametrization basis, three options were considered:

1. **11 independent flavors:** This is the most general case, where all FF flavors and the gluon FF are disentangled. However, this parametrization is overly redundant, as the available dataset does not sufficiently constrain all 11 combinations.

2. **7 independent flavors:** The sea distributions are assumed to be partially symmetric, such that $D_q^{\pi^+} = D_{\bar{q}}^{\pi^+}$ for $q = s, c, b$, and $D_d^{\pi^+} = D_{\bar{u}}^{\pi^+}$. This reduces the number of independent distributions to 7 without significantly deteriorating the quality of the fits. This is chosen as the baseline parametrization.

3. **6 independent flavors:** Imposing approximate $SU(2)$ isospin symmetry would further constrain the FFs by setting $D_d^{\pi^+} = D_{\bar{u}}^{\pi^+}$, reducing the number of independent FFs to 6. However, this assumption leads to a deterioration of the fit quality.

The final set of 7 independent FF combinations parametrized in the fit is:

$$D_u^{\pi^+}, D_{\bar{d}}^{\pi^+}, D_d^{\pi^+} = D_{\bar{u}}^{\pi^+}, D_s^{\pi^+} = D_{\bar{s}}^{\pi^+}, D_c^{\pi^+} = D_{\bar{c}}^{\pi^+}, D_b^{\pi^+} = D_{\bar{b}}^{\pi^+}, D_g^{\pi^+}. \tag{4.1}$$

The same approach is used for Kaons, where the 7 independent FF combinations are:

$$D_u^{K^+}, D_{\bar{d}}^{K^+}, D_d^{K^+} = D_{\bar{u}}^{K^+}, D_s^{K^+} = D_{\bar{s}}^{K^+}, D_c^{K^+} = D_{\bar{c}}^{K^+}, D_b^{K^+} = D_{\bar{b}}^{K^+}, D_g^{K^+}. \tag{4.2}$$

The parametrization is introduced at the initial scale $\mu_0 = 5$ GeV and consists of a single-layer feed-forward neural network $\mathcal{N}_i(z;\theta)$, where θ denotes the set of parameters. This network has:

- One input node corresponding to the momentum fraction z,
- 20 intermediate nodes with a sigmoid activation function,
- 7 output nodes, with a linear activation function, corresponding to the flavor combinations in Eq.(4.1) or Eq.(4.2).

This architecture $[1, 20, 7]$ results in a total of 187 free parameters. The kinematic constraint $D_i^{\pi^+}(z = 1) = 0$ is imposed by subtracting the neural network output at $z = 1$. Additionally, the FFs are constrained to be positive-definite by squaring the outputs:

$$zD_i^{\pi^+}(z, \mu_0 = 5 \text{ GeV}) = \Big(\mathcal{N}i(z;\theta) - \mathcal{N}i(1;\theta)\Big)^2. \tag{4.3}$$

This choice prevents FFs from becoming unphysically negative.

The fit is performed by maximizing the log-likelihood $\mathcal{L}(\theta|x^k)$, which corresponds to minimizing the χ^2 function:

$$\chi^2 \equiv \left(\boldsymbol{T}(\boldsymbol{\theta}^{(k)}) - \boldsymbol{x}^{(k)}\right)^T \cdot C^{-1} \cdot \left(\boldsymbol{T}(\boldsymbol{\theta}^{(k)}) - \boldsymbol{x}^{(k)}\right), \tag{4.4}$$

where $\boldsymbol{T}(\boldsymbol{\theta}^{(k)})$ represents the theoretical predictions obtained from the neural network parametrization.

To avoid overfitting, a cross-validation procedure is implemented. Data sets with more than 10 points are randomly split into training and validation subsets, each containing half of the points, with only the training set used in the fit. Data sets with 10 or fewer points are fully included in the training set. The χ^2 of the

validation set is monitored, and the fit is terminated when the validation χ^2 reaches its minimum. Replicas with a total χ^2 per point exceeding 3 are discarded.

The MontBlanc determination of FFs employs the zero-mass variable-flavor-number scheme (ZM-VFNS), treating all active partons as massless while introducing partial heavy-quark mass dependence at flavor thresholds. The chosen mass thresholds are $m_c = 1.51$ GeV and $m_b = 4.92$ GeV. In this approach, inactive-flavor FFs, such as charm and bottom FFs below their respective thresholds, are not set to zero but remain constant below threshold. This allows heavy-quark FFs to contribute to cross-section computations even below threshold, although their impact in SIDIS is suppressed by PDFs and appears only at NLO.

4.3 LHAPDF Grid

The fit parameters obtained in the previous step were computed at an initial scale μ. Consequently, the resulting set of fragmentation functions is valid only at this specific scale. However, just like PDFs, FFs obey an evolution equation that allows us to compute their values at any energy scale Q perturbatively.

This concept forms the basis of grid-based parameterizations [16], where a grid is generated for each FF replica in a given set. These grids collectively form what we refer to as a fragmentation function (or parton distribution function) set.

A typical FF set consists of multiple files:

- **Info file:** Contains essential metadata about the set, including the number of replicas, mass schemes, perturbative approximations, kinematic variable ranges, and other relevant details.
- **Central replica:** Represents the average over all individual replica grids.
- **Replica grids:** Each replica in the set has a corresponding file containing the grid values for that particular replica.

Each grid consists of multiple blocks, where:

- Each block corresponds to a specific range of energy scales.
- The grid contains eleven columns, one for each quark, antiquark, and gluon.
- Inside each block, the FF values for each parton are stored at different kinematic points.

Finally, specific values of the FFs are extracted from the grid via interpolation, ensuring accurate evaluation at arbitrary energy scales.

4.4 Theoretical Predictions

As discussed in Chapt. (3.22) as the ratio between integrated SIDIS and DIS cross-sections.

To compare theoretical predictions with experimental data, we need to compute these multiplicities. This requires three main ingredients:

- Fragmentation functions, which were determined in the previous steps,
- Partonic cross-sections, implemented in the MontBlanc framework,
- PDFs, which serve as the baseline for assessing the impact of SIDIS data.

The PDF sets used in this project are provided by the NNPDF collaboration. Specifically, we employ the `NNPDF31_pch_0118_1000` sets, which contain 1000 replicas at both NLO and NNLO.

To accurately assess the impact of SIDIS data on PDFs, we must account for the variability of the PDFs in our procedure. To achieve this, theoretical predictions are computed while keeping the FFs fixed at their central values and iterating over the PDF replicas. The final outcome is a set of predictions for each replica and each SIDIS dataset included in the fit. These predictions serve as a crucial input for the reweighting procedure.

4.5 Reweighting

The determination of parton distribution functions (PDFs) and their uncertainties through global fits to datasets obtained from deep inelastic scattering and hadronic collision experiments is a key component in the analysis of current and future experiments. There is a strong correlation between the quality of the data and the reliability of the fits; therefore, whenever new datasets become available, PDF fits must be updated. However, this process is time-consuming and computationally demanding. Furthermore, to ensure consistency, all fits should ideally be performed using the same software framework.

Fortunately, there exists an alternative approach to incorporate the effects of new data into PDFs without requiring a full refit. This method, known as reweighting [17], only requires knowledge of the χ^2 values of the new dataset for each PDF replica in the ensemble. With this information, one can assess the impact of the new data on the PDFs, their consistency with previous datasets, their effect on the shape and precision of individual PDFs, and ultimately their influence on physical observables such as cross-sections and predictions for new physics scenarios.

Reweighting is based on statistical inference. In the NNPDF approach, an ensemble of N PDF replicas, $\mathcal{E} = f_k, k = 1, ..., N$, is generated through a Monte Carlo procedure. Each of these replicas is fitted to a data replica generated according to the experimental uncertainties and their correlations, as provided by experimental collaborations.

Each PDF is parameterized using a highly redundant neural network to minimize parametrization bias, which could otherwise compromise the procedure. As in the case of FFs, cross-validation is used to determine the stopping criterion for the fit of each replica, preventing overfitting. The final PDF ensemble provides an accurate representation of the probability distribution of PDFs, conditional on the input data and the chosen assumptions.

Given a PDF ensemble, any quantity or experimental observable $\mathcal{O}[f]$ that depends on the PDFs can be computed for each replica and then averaged. The integral over the space of functions is well approximated by an average over the ensemble $\mathcal{E}$, so that the mean value of $\mathcal{O}[f]$ is given by:

$$\langle \mathcal{O} \rangle = \int \mathcal{O}[f]\mathcal{P}(f)Df = \frac{1}{N}\sum_{k=1}^{N} \mathcal{O}[f_k]. \tag{4.5}$$

Each replica f_k carries equal weight because they were generated using importance sampling: the replicas were fitted to a data replica drawn from the probability distribution of the experimental data, using an unbiased fitting procedure.

The effect of a new independent dataset can be incorporated without performing a new fit by computing a set of weights w_k for the existing PDF replicas, which quantify the likelihood that each replica f_k agrees with the new data. The reweighted ensemble then represents the probability distribution of PDFs conditioned on both the old and new data. The weights are computed by evaluating the χ^2 of the new data for each replica. The mean value of the observable $\mathcal{O}[f]$ after incorporating the new data is then given by the weighted average:

$$\langle \mathcal{O} \rangle_{new} = \int \mathcal{O}[f]\mathcal{P}_{new}(f)Df = \frac{1}{N}\sum k = 1^{N} w_k \mathcal{O}[f_k]. \tag{4.6}$$

Reweighting allows us to avoid a computationally expensive refit, but it comes with a trade-off: the effective number of replicas is reduced, either because the new data impose strong constraints or because they are inconsistent with the existing data. If the new data are both precise and consistent, the reduction in the effective number of replicas may be so significant that a full refit becomes necessary.

Reweighting is also conceptually important. As more data are included through reweighting, the resulting PDFs become increasingly independent of the initial prior PDF set. Furthermore, PDFs obtained in this way inherently satisfy the principles of statistical inference — such as the proper propagation of uncertainties — ensuring that they evolve according to standard statistical rules upon the inclusion of new data.

4.5.1 Weights

Consider the situation where a set of experimental data has been used to construct a probability distribution for PDFs, $\mathcal{E} = f_k, k = 1, \ldots, N$. As shown in the previous section, any observable can be evaluated by averaging over this PDF ensemble.

The idea behind reweighting is to use statistical inference to compute a set of weights, w_k, for the PDF ensemble such that the new weighted set incorporates information from additional datasets. From this perspec-

tive, the updated probability distribution $\mathcal{P}_{new}(f)$ can be interpreted as an improved version of the original probability distribution $\mathcal{P}_{old}(f)$.

To be specific, consider a set of n new data points that were not included in the determination of the initial probability density distribution:

$$y = y_1, y_2, \ldots, y_n. \tag{4.7}$$

Each instance of this dataset corresponds to a point y in an n-dimensional real space. The experimental uncertainties are summarized by the $n \times n$ experimental covariance matrix C_{ij}, which reduces to a diagonal matrix in cases where correlated systematic uncertainties are unavailable.

We assume that these new data points are statistically independent of any of the data included in the original fit. Using statistical inference, the initial probability density $\mathcal{P}_{old}(f)$ can be updated to incorporate the new data, yielding an improved probability density $\mathcal{P}_{new}(f)$. To achieve this, we need to determine the relative probabilities of the new data for different choices of PDFs. Since the new data are assumed to follow a Gaussian distribution, these probabilities are proportional to the probability density of the χ^2 function conditional on f:

$$\mathcal{P}(\chi|f) \propto (\chi^2(y,f))^{\frac{1}{2}(n-1)} e^{-\frac{1}{2}\chi^2(y,f)} \tag{4.8}$$

where, if $y_i[f]$ is the predicted value for the data point y_i given the PDF f,

$$\chi^2(y,f) = \sum_{i,j=1}^{n} (y_i - y_i[f])\sigma_{ij}^{-1}(y_j - y_j[f]). \tag{4.9}$$

By the law of multiplication of probabilities, and given the statistical independence of the old and new data,

$$\mathcal{P}_{new}(f) = \mathcal{N}\chi\mathcal{P}(\chi|f)\mathcal{P}_{old}(f), \tag{4.10}$$

where $\mathcal{N}\chi$ is a normalization factor independent of f.

Multiplying both sides by an observable $\mathcal{O}[f]$ and integrating over the PDFs, we obtain:

$$\langle\mathcal{O}\rangle_{new} = \int \mathcal{O}[f]\mathcal{P}_{new}(f)Df \quad = \mathcal{N}\chi \int \mathcal{O}[f]\mathcal{P}(\chi|f)\mathcal{P}_{old}(f)Df \; = \frac{1}{N}\sum k = 1^N \mathcal{N}\chi\mathcal{P}(\chi|fk)\mathcal{O}[f_k], \tag{4.11}$$

where in the last step we used Eq. (4.5).

Thus, we can sample the probability density $\mathcal{P}_{new}(f)$ using the N replicas fk, but reweighted. Instead of Eq. (4.9). The normalization factor $\mathcal{N}\chi^{'}$ is determined by requiring that the new probability density $\mathcal{P}_{new}(f)$ is properly normalized. Setting the expectation value $\langle 1\rangle_{new} = 1$, the final expression for the weights is:

$$w_k = \frac{(\chi k^2)^{\frac{1}{2}(n-1)} e^{-\frac{1}{2}\chi k^2}}{\frac{1}{N}\sum_{k=1}^{N}(\chi k^2)^{\frac{1}{2}(n-1)} e^{-\frac{1}{2}\chi k^2}}. \tag{4.12}$$

The weights w_k, when divided by N, represent the probabilities of the replicas f_k given the χ^2 values for the new data.

4.5.2 Measuring information Loss and Consistency

The original ensemble of replicas $\mathcal{E} = f_k, k = 1, \ldots, N$ is constructed through importance sampling of the probability density $\mathcal{P}_{old}(f)$. Consequently, each replica has equal weight, ensuring that the ensemble is maximally efficient — meaning that for a given number of replicas N, this is the best possible representation of the underlying density $\mathcal{P}_{old}(f)$. The only way to improve this representation is by increasing N.

After reweighting, however, this is no longer the case, as the weights w_k assign different levels of importance to the replicas. As a result, the reweighted ensemble is less efficient: for a given N, the accuracy of the representation of the updated distribution $\mathcal{P}_{new}(f)$ is reduced compared to what would be achieved by performing a full refit.

The loss of efficiency can be quantified using Shannon entropy to compute the effective number of replicas remaining after reweighting:

$$N_{eff} \equiv \exp\left(\frac{1}{N}\sum_{k=1}^{N} w_k \ln\left(\frac{N}{w_k}\right)\right). \tag{4.13}$$

Clearly, $0 < N_{eff} < N$: the reweighted ensemble has the same accuracy as a refit performed with N_{eff} replicas. If N_{eff} becomes too small, the reweighting procedure may no longer be reliable. This may occur for two reasons:

- The new data provide a significant amount of additional information on the PDFs, requiring a full refit with a larger number of replicas.

- The new data are inconsistent with the previous dataset, indicating possible issues with the experimental uncertainties or systematic biases.

These two cases can be distinguished by examining the χ^2 profile of the new data. If, in the reweighted fit, very few replicas exhibit a χ^2 per data point of order unity, it is likely that the uncertainties in the new dataset have been underestimated. This profile can be evaluated using:

$$\mathcal{P}(\chi^2) = \frac{1}{N}\sum_k w_k, \tag{4.14}$$

where the sum is taken over all replicas k for which $\chi^2_k \in [\chi^2, \chi^2 + d\chi^2]$.

Alternatively, inconsistent data can be interpreted as data whose uncertainties have been underestimated. In this case, we can introduce a scaling factor α for the uncertainties and use inverse probability to compute the probability density for α:

$$\mathcal{P}(\alpha) = \frac{1}{\alpha}\sum_{k=1}^{N} w_k(\alpha), \tag{4.15}$$

where $w_k(\alpha)$ are the reweighting factors computed by replacing χ^2 with χ^2/α^2. Averaging these weights in the reweighted fit provides an estimate for the probability density of α.

If the probability density $\mathcal{P}(\alpha)$ peaks close to one, the new data are consistent with previous measurements. However, if it peaks significantly above one, it is likely that the errors in the new data have been underestimated, suggesting a need for further investigation.

4.6 Unweighting

The standards for PDF sets require that all replicas have equal weights. Therefore, to share and use a reweighted set, an additional step is necessary: unweighting.

We begin with a set of N_{rep} reweighted replicas, where each replica, indexed by $k = 1, \ldots, N_{rep}$, carries a weight w_k determined by comparing each replica of the original unweighted distribution to the new experimental data. The goal of unweighting is to obtain a new set of $N^{'}_{rep}$ replicas, all with equal weight, while preserving the probability distribution of the original weighted set.

This procedure, known as unweighting [18], is achieved by selecting replicas from the weighted set such that replicas with relatively high weights are chosen multiple times, while those with very small weights are removed from the final unweighted set.

The method is illustrated in Fig. 4.2. We begin by subdividing a unit-length interval into N_{rep} segments, ensuring that the length of each segment is proportional to the weight of the corresponding replica, and ordering them randomly. To extract a set of $N^{'}_{rep}$ replicas that accurately represents this distribution, we draw another unit-length interval below the first, dividing it into $N^{'}_{rep}$ segments of equal length, $1/N^{'}_{rep}$. Replicas from the original weighted set are then selected based on the number of lower-segment right edges that fall within the corresponding upper segment.

This selection process ensures that all $N^{'}_{rep}$ replicas are chosen according to the probabilities defined by the N_{rep} replicas in the original set.

If $N^{'}_{rep}$ is sufficiently large, at least one lower segment will fall within each upper segment, and the original probability distribution will be faithfully reproduced. However, in practice, this would require $N^{'}_{rep}$ to be as large as the ratio between the highest and lowest weight, which can be prohibitively large.

Choosing a very large $N^{'}_{rep}$ is unnecessary because the effective information contained in the weighted set is quantified by its Shannon entropy, which determines the effective number of unweighted replicas N_{eff}. By construction, $N_{eff} \leq N_{rep}$.

For larger weights, multiple unweighted segments are contained within a single weighted segment, while for smaller weights, there are often none. Since the ordering of replicas is random, the choice among equally small-weighted replicas is also random.

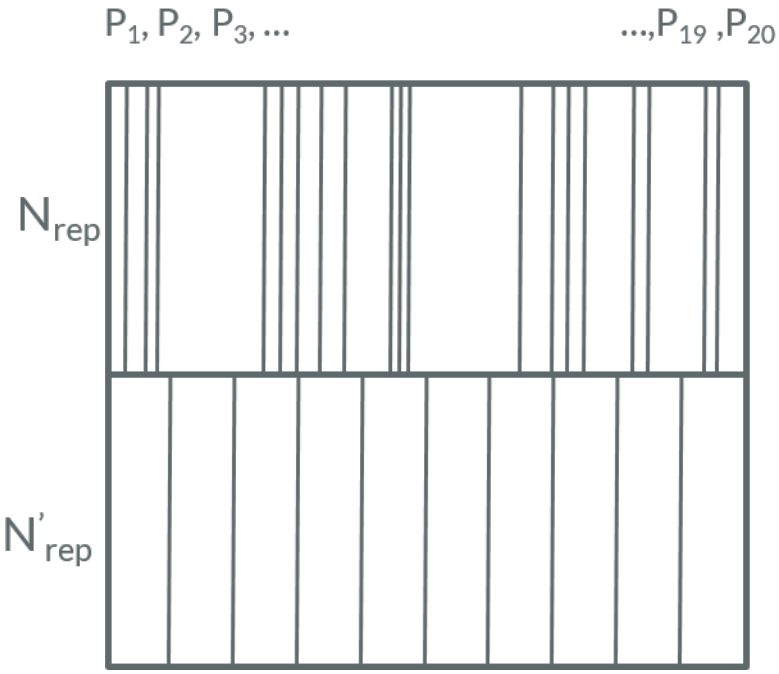

Figure 4.2: Graphical representation of the unweighting process. In the lower section, $N_{rep}^{'}$ (vertical lines) are selected, each corresponding to the nearest replica in the upper section, which initially contains N_{rep} replicas.

4.6.1 Unweighting Algorithm

We now present the quantitative formulation of the unweighting algorithm. Given an initial set of N_{rep} replicas with weights w_k, we normalize the weights as:

$$\sum_{k=1}^{N_{rep}} w_k = N_{rep}. \tag{4.16}$$

The probability of each replica is then given by:

$$p_k = \frac{w_k}{N_{rep}}. \tag{4.17}$$

We define cumulative probabilities:

$$P_k \equiv P_{k-1} + p_k = \sum_{j=0}^{k} p_j, \tag{4.18}$$

where we set $P_0 = 0$. By construction, $0 \leq P_k \leq 1$ and $P_{k-1} \leq P_k$. These cumulants define the positions of the right edges of the upper segments in Fig. 4.2.

The unweighted set is then constructed as follows. We start with N_{rep} weights w_k and determine $N_{rep}^{'}$ new weights:

$$w_k^{'} = \sum_{j=1}^{N_{rep}^{'}} \theta\left(\frac{j}{N_{rep}^{'}} - P_{k-1}\right) \theta\left(P_k - \frac{j}{N_{rep}^{'}}\right). \tag{4.19}$$

The weights $w_k^{'}$ are either zero or positive integers, satisfying the normalization condition:

$$N_{rep}^{'} \equiv \sum_{k=1}^{N_{rep}} w_k^{'}. \tag{4.20}$$

The unweighted set is then built by selecting $w_k^{'}$ copies of the k-th replica for all $k = 1, \ldots, N_{rep}$. The probability of the k-th replica in the new unweighted set is given by:

$$p_k^{'} = \frac{w_k^{'}}{N_{rep}^{'}}. \tag{4.21}$$

As a consequence, in the limit of a large sample size, the unweighted set reproduces the probability distribution of the weighted set. Although the probability distributions of the reweighted and unweighted sets are identical in the limit of Eq. (4.13). For practical applications, it is advisable to choose $N_{rep}^{'} \leq N_{eff}$. While there is no fundamental issue with taking $N_{rep}^{'} > N_{eff}$, this would result in a highly redundant replica set without adding new information.

4.7 Distances between PDFs: definition and meaning

Plotting the unweighted (posterior) and initial (prior) PDF sets to visualize the differences between them — i.e., the impact of new data — is a challenging task. The primary issue is that in certain regions, the changes in the unweighted set relative to the prior set are minimal, making it difficult to quantify these differences accurately. One possible approach is to normalize the results by dividing by the prior set, thereby enhancing the differences. However, since PDFs tend to zero at high x, this can introduce divergences in the plots.

To address this issue, we introduce a method based on defining a distance metric. Given two sets of $N_{rep}^{(1)}$ and $N_{rep}^{(2)}$ replicas, we seek to determine whether they correspond to different instances of the same underlying probability distribution or whether they originate from distinct distributions. Since $N_{rep}^{(i)}$ is finite, this question can only be answered statistically. To this end, we define the square distance between two estimators based on these samples as the squared difference between the estimators divided by the corresponding variance. By construction, the expectation value of this distance is one.

Given a set of $N_{rep}^{(k)}$ replicas $q_i^{(k)}$ for some quantity q, the estimator for its expected (true) value is given by the mean:

$$\langle q^{(k)} \rangle = \frac{1}{N_{rep}^{(i)}} \sum_{i=1}^{N_{rep}^{(i)}} q_i^{(k)}. \tag{4.22}$$

The squared distance between the two estimates of the expected value from sets $q_i^{(1)}$ and $q_i^{(2)}$ is given by:

$$d^2(\langle q^{(1)} \rangle, \langle q^{(2)} \rangle) = \frac{(\langle q^{(1)} \rangle - \langle q^{(2)} \rangle)^2}{\sigma_{(1)}^2[\langle q^{(1)} \rangle] + \sigma_{(2)}^2[\langle q^{(2)} \rangle]}, \tag{4.23}$$

where the variance of the mean is given by:

$$\sigma_{(i)}^2[\langle q^{(i)} \rangle] = \frac{1}{N_{rep}^{(i)}} \sigma_{(i)}^2[q^{(i)}], \tag{4.24}$$

and the variance $\sigma_{(i)}^2[q^{(i)}]$ is computed as:

$$\sigma_{(i)}^2[q^{(i)}] = \frac{1}{N_{rep}^{(i)} - 1} \sum_{k=1}^{N_{rep}^{(i)}} (q_k^{(i)} - \langle q^{(i)} \rangle)^2. \tag{4.25}$$

A similar definition applies to quantify the difference in uncertainties. Given a set of $N_{rep}^{(k)}$ replicas $q_i^{(k)}$, the estimator for the squared uncertainty of q is given by the variance of the replica sample. The distance between the two estimates of the squared uncertainty from sets $q_i^{(1)}$ and $q_i^{(2)}$ is defined as:

$$d^2(\sigma_{(1)}^2, \sigma_{(2)}^2) = \frac{(\bar{\sigma}(1)^2 - \bar{\sigma}(2)^2)^2}{\sigma_{(1)}^2[\bar{\sigma}(1)^2] + \sigma(2)^2[\bar{\sigma}(2)^2]}, \tag{4.26}$$

where, for brevity, we define $\bar{\sigma}(i)^2 \equiv \sigma_{(i)}^2[q^{(i)}]$.

The variances $\sigma^2_{(i)}[\bar{\sigma}(i)^2]$ of the squared uncertainties can be estimated from the replica sample by computing variances from multiple subsets and then computing the variance of these resulting variances. However, for finite $Nrep$, this method may lead to statistical inaccuracies. Instead, we use the expression:

$$\sigma^2_{(i)}[\bar{\sigma}(i)^2] = \frac{1}{Nrep^{(i)}} \left[m_4[q^{(i)}] - \frac{N^{(i)}_{rep} - 3}{N^{(i)}_{rep} - 1} (\bar{\sigma}(i)^2)^2 \right], \tag{4.27}$$

where the fourth moment $m4$ of the probability distribution is estimated as:

$$m_4[q^{(i)}] = \frac{1}{N^{(i)}_{rep}} \sum_{k=1}^{N^{(i)}_{rep}} (q^{(i)}_k - \langle q^{(i)} \rangle)^4. \tag{4.28}$$

In practice, for small replica samples, the distances defined in Eqs. (63) and (66) exhibit significant statistical fluctuations. These distances measure whether the given samples originate from the same underlying probability distribution. Specifically, Eq. (63) tests whether the two distributions have the same mean, while Eq. (66) assesses whether they have the same standard deviation.

By construction, the probability distribution of the distance follows a χ^2 distribution with one degree of freedom, meaning it has an expectation value of $\langle d \rangle = 1$ and satisfies $d \lesssim 2.3$ at the 90% confidence level.

It is important to note that asking whether two PDF determinations originate from the same underlying distribution is a more stringent criterion than simply checking whether they are consistent within uncertainties. Consider, for example, two PDF sets where one is based on a dataset that is a subset of the other. Even if all data are consistent, the two determinations will not originate from the same underlying distribution because the PDF set based on the larger dataset will exhibit reduced uncertainties. However, they will still be consistent within uncertainties.

The precision of estimating moments of the underlying distribution improves as the number of replicas increases, with the accuracy of the expectation value scaling as $1/\sqrt{N_{rep}}$. Thus, if the underlying probability distributions are different, the distance metric will grow as $\sqrt{N_{rep}}$ in the large-N_{rep} limit. In this limit, the distance between central values is given by the rescaled distance:

$$\delta(\sigma^2(1), \sigma^2(2)) \equiv \frac{1}{\sqrt{N_{rep}}} d(\sigma^2(1), \sigma^2(2)). \tag{4.29}$$

For all distances computed in this study with $N_{rep} = 100$, one standard deviation corresponds to $d = \sqrt{50} \approx 7$.

Chapter 5

Results

In this Chapter I will present the main results of this Thesis, namely the impact of SIDIS data on the determination of proton PDFs. I obtain these results by applying the iterative procedure described in Sect.4.1. Specifically, I proceed as follows. First, the procedure is tested separately on pion and kaon data at NLO and NNLO. This will show the differences between the two hadronic species and the two perturbative orders, and it will serve as a test on the correct implementation of the procedure. Second, these results will be combined, only at NNLO, applying the procedure on pion data for two iterations and then on kaon data for another two iterations. The ensuing posterior PDF set is the final result from which, by comparison with the prior PDF set, I will gauge the impact of SIDIS data on PDFs.

This Chapter is organized as follows. In Sects. 5.1-5.2 I present the χ^2 and weight distributions, respectively. In Sect. 5.3 I will introduce the comparison plots between the PDF set used at the beginning of an iteration, which we will call prior set, and the unweighted set obtained at the end of an iteration which we will call the posterior set. In the end, the same structure will be repeated for the final results where we will try to summarize what we have found.

5.1 χ^2 distributions at NLO and NNLO

In this section, I present the results obtained for Pions and Kaons in the first part of the study. The primary objectives are twofold: first, to verify whether the iterative procedure converges with the number of iterations, and second, to compare the impact of SIDIS data on the PDFs at NLO and NNLO.

Starting with the χ^2 distributions, we recall that χ^2 is computed using Eq. (4.9), and that, in applying the Monte Carlo method, we assume that the data follow a Gaussian distribution. Therefore, we expect that, iteration after iteration, the empirical distribution will shift towards the expected distribution, which corresponds to a χ^2 distribution with a number of degrees of freedom (df) equal to the number of points in the dataset. Additionally, we anticipate that with more iterations, the empirical distribution will become more symmetric. This is because the Monte Carlo method maps data uncertainties onto the parameter space, allowing us to construct PDF and FF replicas. If the new datasets are perfectly incorporated into the PDFs, our predictions should describe these data well. Consequently, the χ^2 values computed for each PDF replica should follow the same distribution as the data, namely, a Gaussian distribution.

In Fig. 5.4. As with Pions, the blue histogram represents the empirical distribution, while the orange curve and dotted line correspond to the expected distribution and its mean. At NLO, we observe that the empirical distribution is closer to the expected one at both iterations compared to the case of Pions. A similar pattern is seen at NNLO, though it is less pronounced and only evident at the second iteration. For Kaon data, the choice of $N_{rep}^{'} = 100$ appears to be less restrictive. Indeed, after one iteration, the distribution obtained from experimental data shifts towards the expected curve and also becomes more symmetric. From these observations, we anticipate that at NLO, the impact of SIDIS data on the PDFs will be smaller for Kaons than for Pions. Furthermore, we expect that at NNLO, Pion data will have a lower impact, whereas Kaon data should exhibit a significant impact in the first iteration and a smaller impact in the second iteration, confirming that the dataset is being effectively incorporated into the PDFs.

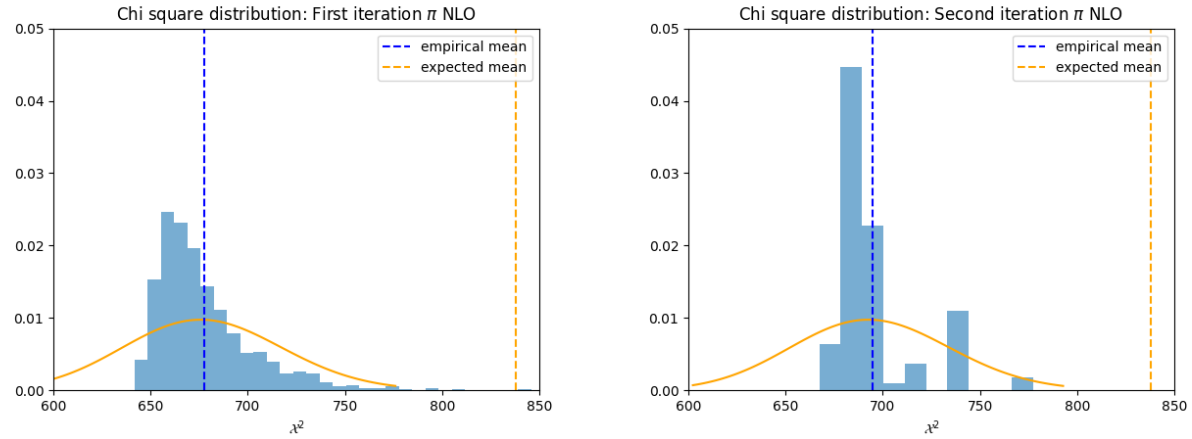

Figure 5.1: Chi square distributions for Pions first(left) and second(right) iteration at NLO. The orange curve represent the expected distribution which have been centered on the empirical mean to be able to compare the shape of the two distributions. The orange dotted line represent the mean of the orange distribution.

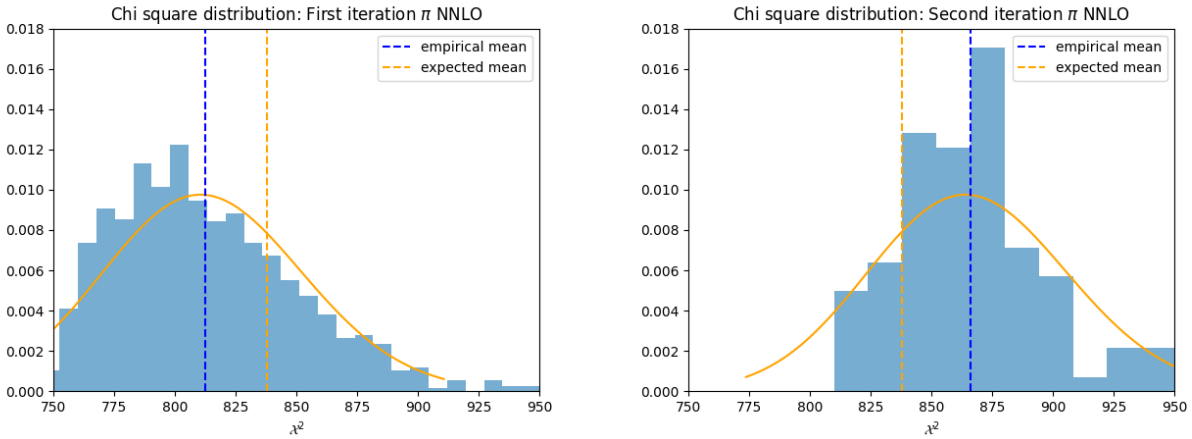

Figure 5.2: Chi square distributions for Pions first(left) and second(right) iteration at NNLO. The orange curve represent the expected distribution which have been centered on the empirical mean to be able to compare the shape of the two distributions. The orange dotted line represent the mean of the orange distribution.

5.2 Weights distributions at NLO and NNLO

The knowledge of the χ^2 values allows us to apply reweighting to the prior PDF set. In this section, we present the results for weight distributions at NLO and NNLO for Pions and Kaons separately. As discussed in Sect. 4.5, reweighting is used to incorporate new experimental information into a given PDF set. Initially, each PDF replica carries the same weight, $w_k = 1$, meaning that all replicas contribute equally to the evaluation of observables such as cross-sections. After applying reweighting, a new set of weights is assigned based on the new data being included. At this stage, unweighting is used to construct a final set in which all replicas have equal weights, while still preserving the information contained in the reweighted set.

If all experimental information were fully incorporated in the first iteration of the procedure, a subsequent application of reweighting using the same data should have no further effect on the weights. In such a scenario, the weight distribution would form a Dirac delta function centered at $w_k = 1$. However, in practice, this is not the case, as some information is inevitably lost during the unweighting step. Therefore, as the iterative procedure progresses, we expect the width of the weight distribution to decrease while the mean shifts toward $w_k = 1$.

Starting with Pions, Fig. 5.5 shows a comparison between the weight distributions after the first iteration (blue) and after the second iteration (orange). At NLO, even after two iterations, the empirical distribution

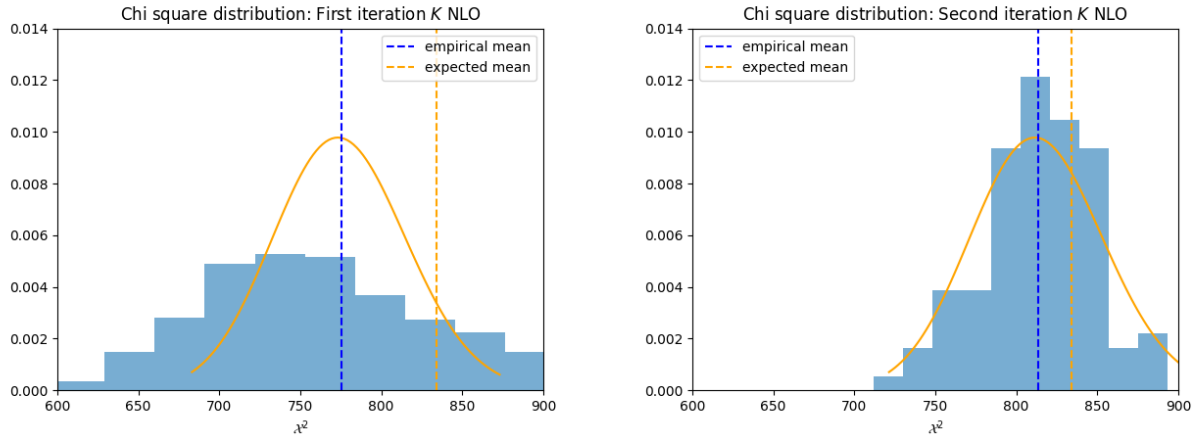

Figure 5.3: Chi square distributions for Kaons first(left) and second(right) iteration at NLO. The orange curve represent the expected distribution which have been centered on the empirical mean to be able to compare the shape of the two distributions. The orange dotted line represent the mean of the orange distribution.

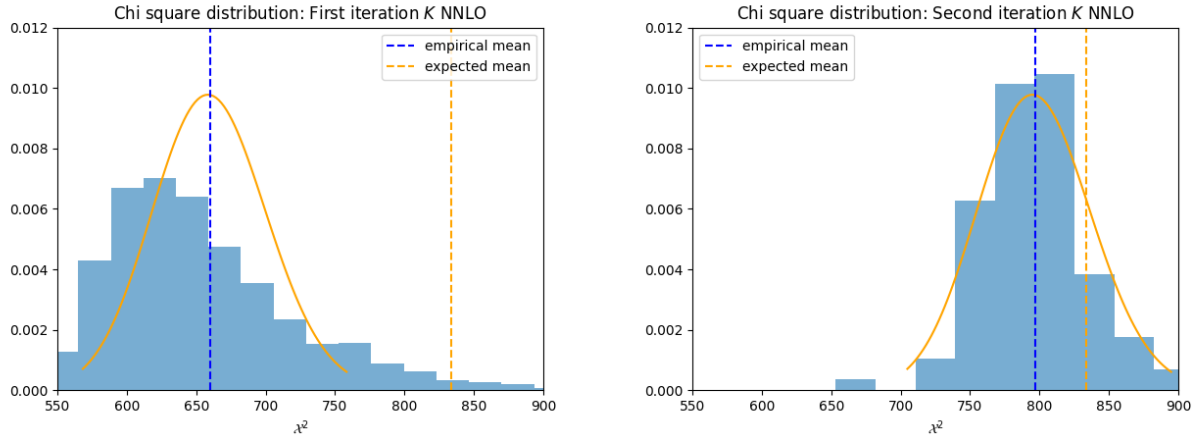

Figure 5.4: Chi square distributions for Kaons first(left) and second(right) iteration at NNLO. The orange curve represent the expected distribution which have been centered on the empirical mean to be able to compare the shape of the two distributions. The orange dotted line represent the mean of the orange distribution.

remains far from the expected delta function, but we observe a decrease in the width of the distribution, indicating some degree of convergence. At NNLO, the improvement in distribution width is less noticeable. However, considering that the unweighted set at the end of the second iteration consists of only 100 replicas, we observe that approximately 75

For Kaons, the weight distributions are shown in Fig. 5.6. Compared to Pions, the distributions appear more concentrated around $w_k = 1$ at both perturbative orders, aside from a few outliers. These results provide concrete evidence that the iterative procedure is converging, although the effect is less pronounced at NLO, particularly for Pions.

5.3 Comparison plot on the impact of Pions and Kaons data

In this section, as a final validation of the procedure, we present a comparison between the PDF set used at the beginning of each iteration (Prior) and the set obtained at the end of each iteration (Posterior). Specifically, we examine the parton distributions for the valence quarks of Pions and Kaons. In these plots, central values are represented by solid curves, while uncertainties are depicted as shaded bands. The prior set is shown in blue, and the posterior set is shown in red. To emphasize the differences between the PDFs, both central values and uncertainties are normalized to the prior set.

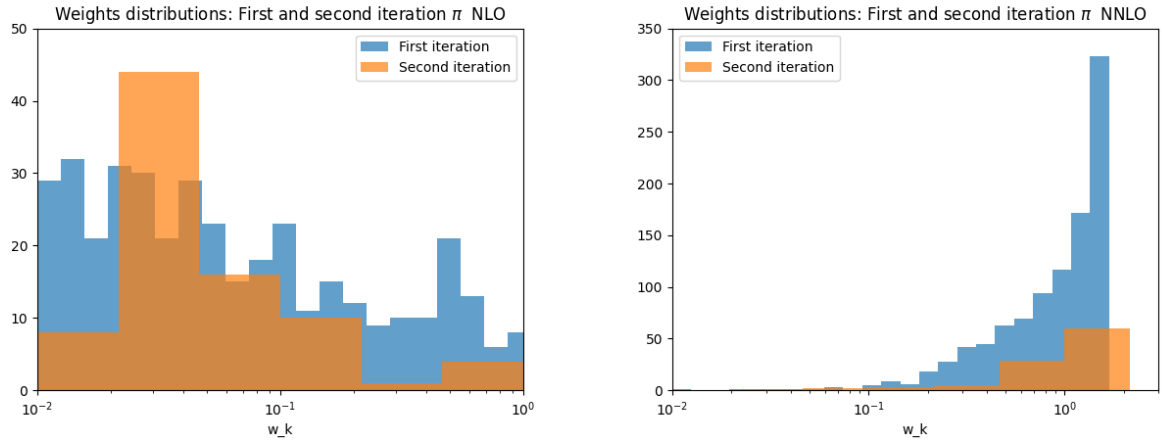

Figure 5.5: Weights distribution for Pions at NLO(left) and NNLO(right). The orange histogram shows the second iteration while the blue one the first.

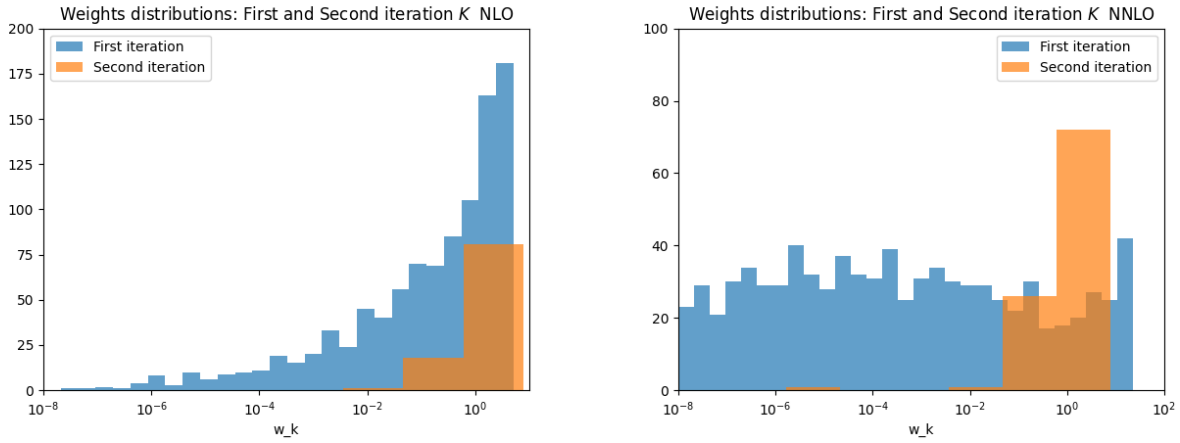

Figure 5.6: Weights distribution for Kaons at NLO(left) and NNLO(right). The orange histogram shows the second iteration while the blue one the first.

At NLO, Pion data appear to have a non-negligible impact on the PDFs. The number of effective replicas at each iteration, i.e., the number of replicas that retain most of the information in the set, is approximately 4% in the first iteration and increases to 30% in the second iteration.

However, when comparing these results with those at NNLO, we observe a different picture. The number of effective replicas is already high at the first iteration (92%) and increases slightly to 94% in the second iteration. Consistently, the comparison plots in Figs. 5.7, 5.8, 5.9, 5.10, 5.11, 5.12, 5.13, and 5.14 show that the impact at NNLO is relatively small. This suggests that part of the experimental uncertainty is influencing the impact, amplifying the effects at NLO.

A similar analysis can be performed for Kaons, with the corresponding plots shown in Figs. 5.15, 5.16, 5.17, 5.18, 5.19, 5.20, 5.21, and 5.22.

As expected from the χ^2 and weight distributions, at NLO the impact is less pronounced compared to the case of Pions, while at NNLO there is a noticeable improvement between the first and second iteration. These behaviors align with the number of effective replicas shown in Table 5.1. At NLO, we find $N_{eff} = 48\%N_{rep}$ for the first iteration and $N_{eff} = 90\%N_{rep}$ for the second iteration. At NNLO, the corresponding values are $N_{eff} = 10\%N_{rep}$ and $N_{eff} = 85\%N_{rep}$, respectively.

We now summarize the key findings obtained so far. We have demonstrated that the iterative procedure converges, although at NLO, the lack of perturbative accuracy affects the impact of the data. Furthermore, to improve the interpretation of the χ^2 and weight distribution plots, it would be beneficial to work with a larger PDF set, keeping the set size between iterations fixed at N_{eff} while maintaining $N_{rep} = 100$ only for the final

	Pions		Kaons	
	NLO	NNLO	NLO	NNLO
First iteration	4%	92%	48%	10%
Second iteration	30%	94%	90%	85%

Table 5.1: Values for the number of effective replica N_{eff} after the reweighting respect to the number of replica in the initial PDF set.

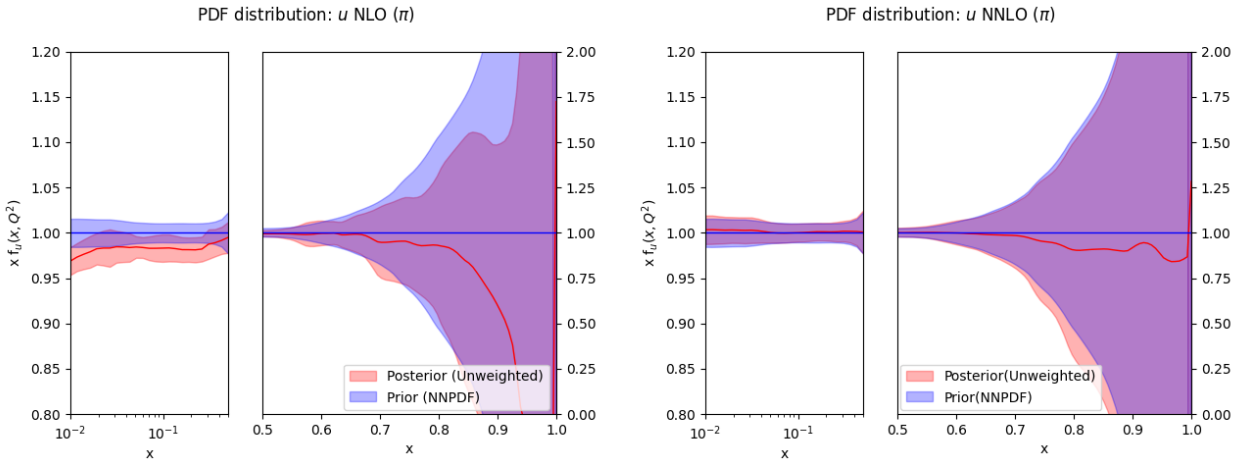

Figure 5.7: Quark u distribution, first iteration, at NLO(left) and NNLO(right) obtained from Pions data. In blue we have the NNPDF set while in red we have the unweighted set built at the end of the iteration. Central values and uncertainties are normalized over NNPDF set.

unweighted set. With this consideration in mind, a comprehensive study of the impact of SIDIS data on PDFs has been conducted. Since the results obtained at NLO indicate that a much larger PDF set would be required for an effective application of the procedure, we opted to repeat the analysis only at NNLO. Finally, while the comparison plots provide insight into the impact of SIDIS data on PDFs, a more quantitative approach is preferable for the final results. As introduced at the end of Chapter 4, we employ the concept of distance between PDFs' central values and uncertainties to quantify these effects.

5.4 Impact of SIDIS data on Parton Distribution Functions

We aim to assess the combined impact of Pion and Kaon data on the PDFs. The most straightforward approach is to utilize the iterative procedure, initially incorporating one dataset and, after a few iterations, introducing the other. As discussed in Sect. 4.5, applying reweighting can significantly reduce the number of replicas. If the number of effective replicas becomes too low, the statistical sample may be insufficient, necessitating a full refit. For this reason, I opted to drop the NLO analysis and proceed exclusively at the second perturbative order. Additionally, we observed that the impact of Pion data on the PDFs is relatively small compared to Kaon data, leading to a higher number of effective replicas, close to N_{rep}.

Consequently, we begin by incorporating Pion data, ensuring that the number of replicas remains sufficiently large before subsequently including Kaon data. Furthermore, given that two iterations were sufficient for both datasets individually, I chose to apply two iterations with Pion data, followed by two iterations with Kaon data. Examining the χ^2 distributions in Figs. 5.24 corresponds to the third iteration of the procedure, it also represents the first instance in which Kaon data were included. As a result, the agreement between distributions slightly deteriorates. The central values for the empirical χ^2 are reported in Table 5.2.

The weight distributions are shown in Fig. 5.25. In this case as well, the improved statistics enhance the

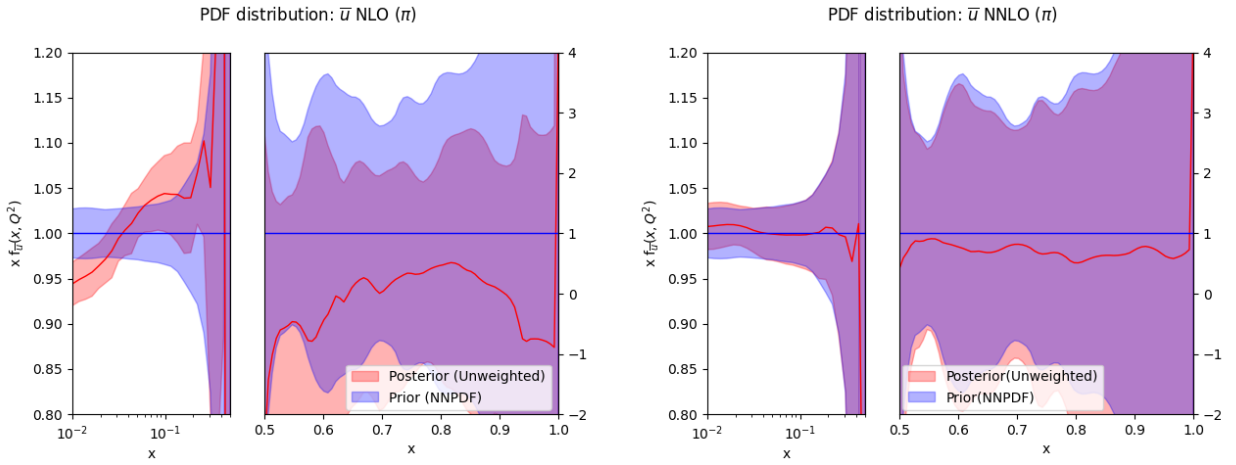

Figure 5.8: Quark $\bar{u}$ distribution, first iteration, at NLO(left) and NNLO(right) obtained from Pions data. In blue we have the NNPDF set while in red we have the unweighted set built at the end of the iteration. Central values and uncertainties are normalized over NNPDF set.

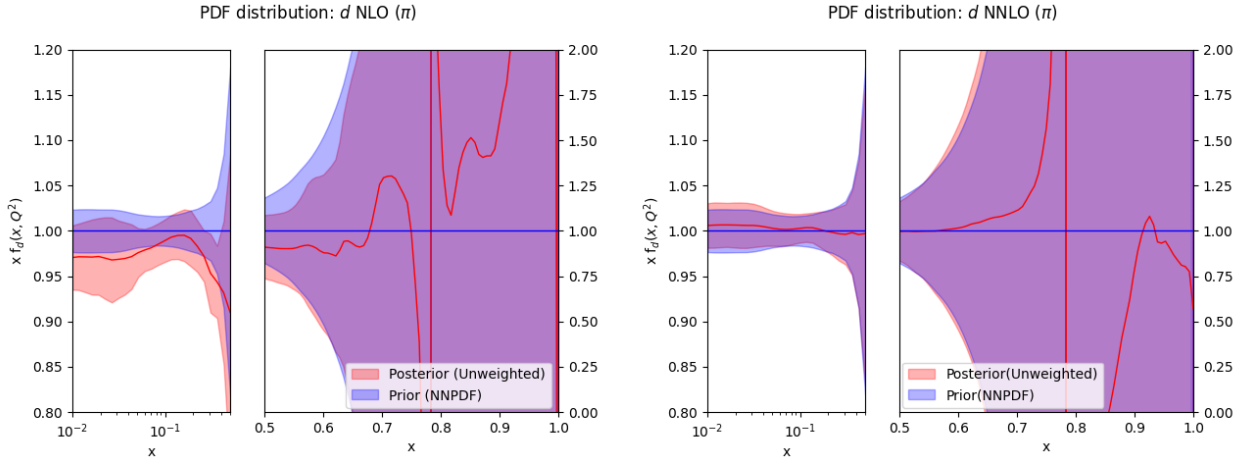

Figure 5.9: Quark d distribution, first iteration, at NLO(left) and NNLO(right) obtained from Pions data. In blue we have the NNPDF set while in red we have the unweighted set built at the end of the iteration. Central values and uncertainties are normalized over NNPDF set.

Pions		Kaons	
First iteration	Second iterazion	Third iteration	Fourth iteration
0.97	1.01	0.79	0.95

Table 5.2: Values for the empirical reduced χ^2 at each iteration.

readability of the plot. During the first two iterations for Pions, we observe a decrease in the width of the distribution, with more weights clustering around unity. For Kaons, this improvement is even more pronounced, where after two iterations, the distribution transitions from a broad shape with very small weights to a much more concentrated form around the expected result.

Finally, we present the comparison plots between the initial PDF set used for the analysis and the unweighted set obtained after the last iteration. As in the separate cases, these plots are normalized to the Prior (NNPDF) set, which is represented in blue, while the Posterior (Unweighted) set is shown in red. In this section, we

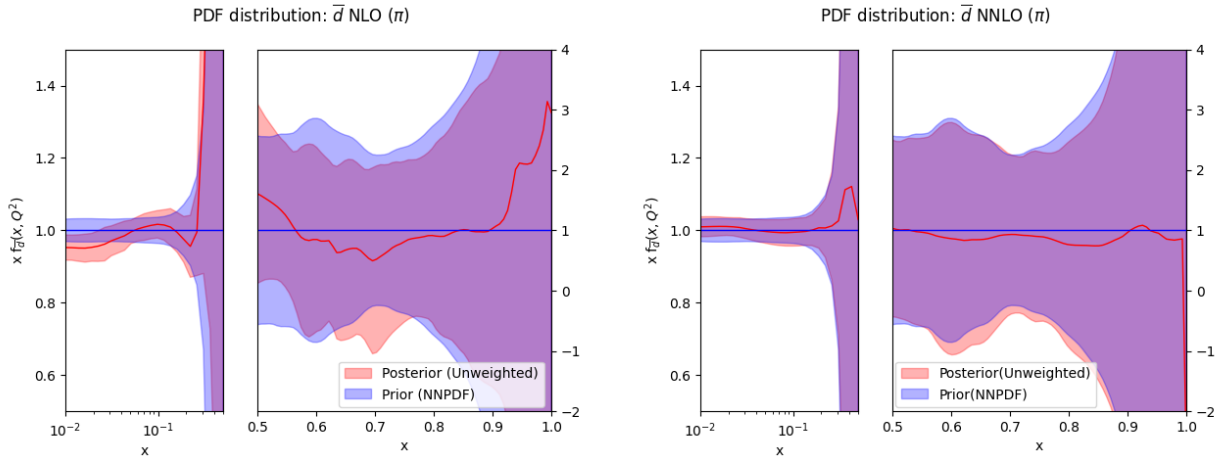

Figure 5.10: Quark $\bar{d}$ distribution, first iteration, at NLO(left) and NNLO(right) obtained from Pions data. In blue we have the NNPDF set while in red we have the unweighted set built at the end of the iteration. Central values and uncertainties are normalized over NNPDF set.

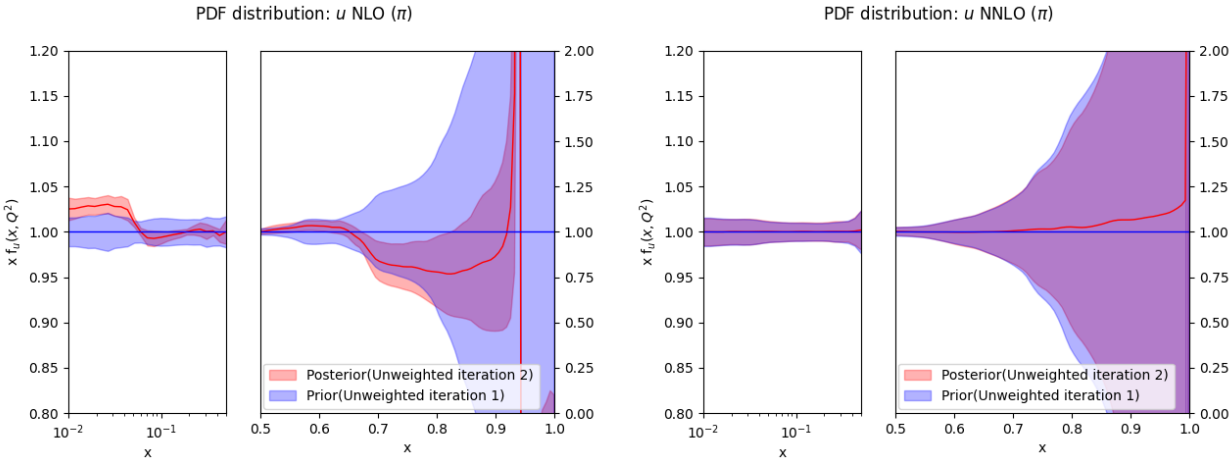

Figure 5.11: Quark u distribution, second iteration, at NLO(left) and NNLO(right) obtained from Pions data. In blue we have the unweighted set obtained at the first iteration(prior) while in red we have the unweighted set built at the end of the second. Central values and uncertainties are normalized over the prior set.

display only the valence quark distributions for Pions and Kaons. Additional plots can be found in Appendix B. Fig. 5.26 shows the u and $\bar{u}$ distributions. For the u distribution at small x, we observe a shift of nearly 1σ in the posterior central value relative to the prior, while at higher x, the impact is minimal. For $\bar{u}$, this pattern is not observed, and the impact is negligible.

A similar behavior is seen for the d and $\bar{d}$ distributions in Fig. 5.27, where a noticeable impact is present only at small x for the d quark. Finally, for the s and $\bar{s}$ quark distributions, we observe some effects at low x, particularly for the s distribution, where the central value shifts by nearly one σ. From these plots, we conclude that in most cases, the impact of SIDIS data on the PDFs is negligible, with more significant effects appearing at low x for the u, s, and d quarks.

5.5 PDFs distances, Rs plot and Data-Prediction comparison

As a last test of consistency, I have computed the PDFs distances introduced in Chapt. 4 with the help of the validphys code developed by NNPDF collab. This measure will give a more quantitative information about the impact of the data on the PDFs and represent how much the Unweighted set obtained at the end of the last iteration is statistically different from the Prior set. Since the posterior set have $N_{rep} = 62$ a distance value of 8 means that the two distributions differ by one σ. I will show separately the results for the u, $\bar{u}$, d, $\bar{d}$, s, $\bar{s}$

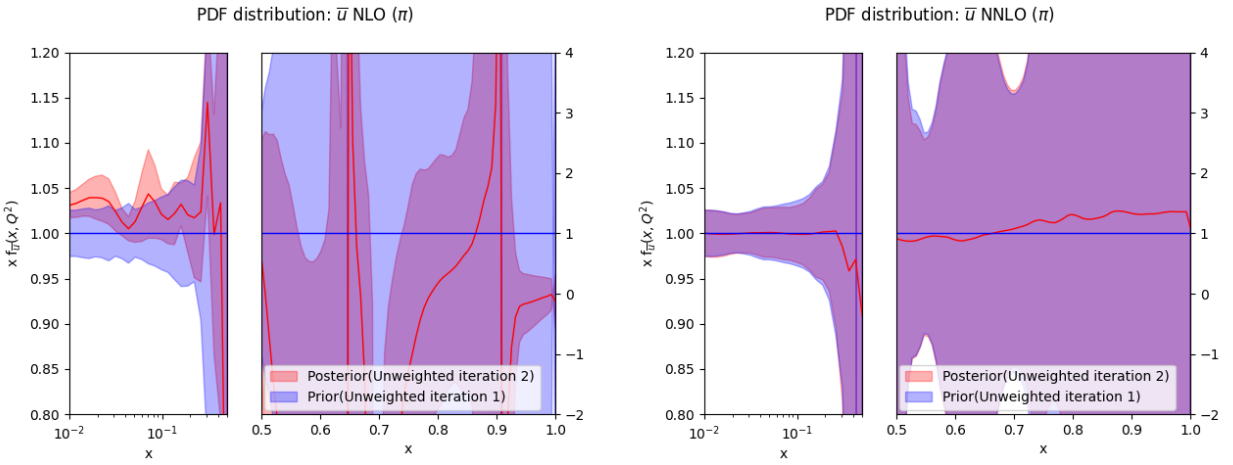

Figure 5.12: Quark $\bar{u}$ distribution,second iteration, at NLO(left) and NNLO(right) obtained from Pions data. In blue we have the unweighted set obtained at the first iteration(prior) while in red we have the unweighted set built at the end of the second(posterior). Central values and uncertainties are normalized over the prior set.

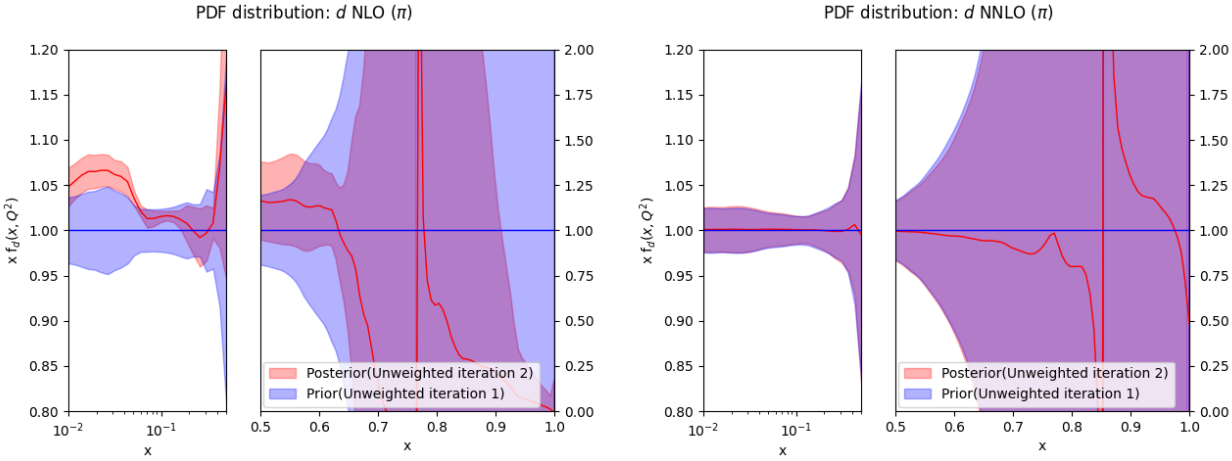

Figure 5.13: Quark d distribution, second iteration, at NLO(left) and NNLO(right) obtained from Pions data. In blue we have the unweighted set obtained at the first iteration(prior) while in red we have the unweighted set built at the end of the second(posterior). Central values and uncertainties are normalized over the prior set.

distributions.

In Fig. 5.29 distances plot are shown for u and $\bar{u}$ quarks. The distance is evaluated between central values of the two distributions and is plotted in logarithmic scale to better show the effects at low x. In accord to what we have seen with the comparison plots of the u distribution, the distance reach its maximun at $x = 10^{-1}$ where a value of 10 means that the posterior distribution is different of 1 sigma. At the same time for the $\bar{u}$ the impact appear to be milder than the one seen in the last section. Like the u quark also the d and $\bar{d}$ quarks distributions show a similar behavior to the comparison plot. In this case we reach a peak value of 8 and 7 for the d and $\bar{d}$ quarks respectevely. Finally, the s and $\bar{s}$ distribution show less impact than what I expected for the s distribution where it reach a maximum value of 5. Regarding $\bar{s}$ the distance plot reflect the expected andamento where the peak at high x is probably given by computation instabilities. In the end, these plots allow me to confirm some of the conclusion I reached in the previous section giving more informations on the impact on the distributions.

Finally, one observable of interest is R_s which is defined as

$$R_s = \frac{f_s + f_{\bar{s}}}{f_{\bar{u}} + f_{\bar{d}}} \tag{5.1}$$

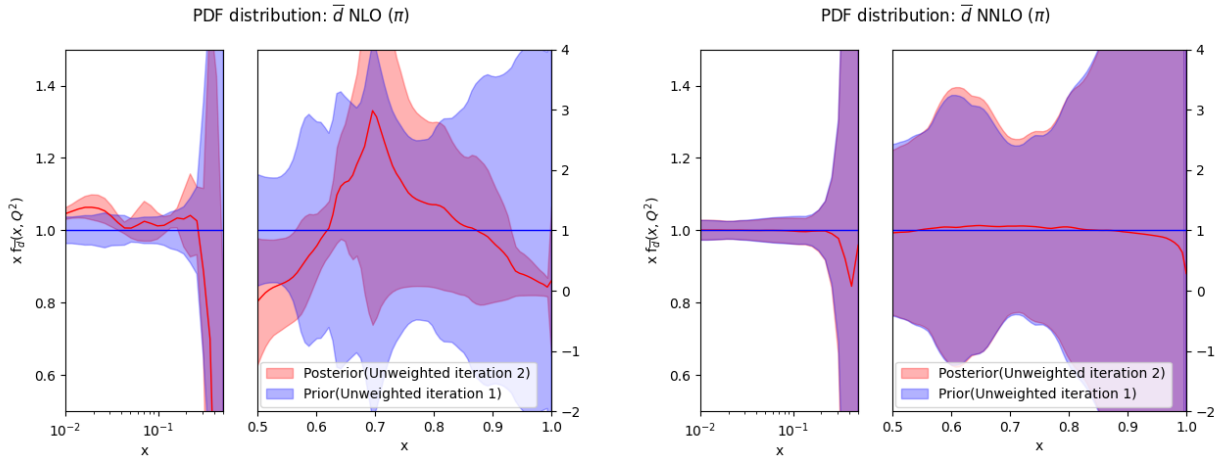

Figure 5.14: Quark $\bar{d}$ distribution, second iteration, at NLO(left) and NNLO(right) obtained from Pions data. In blue we have the unweighted set obtained at the first iteration while in red we have the unweighted set built at the end of the second(posterior). Central values and uncertainties are normalized over the prior set.

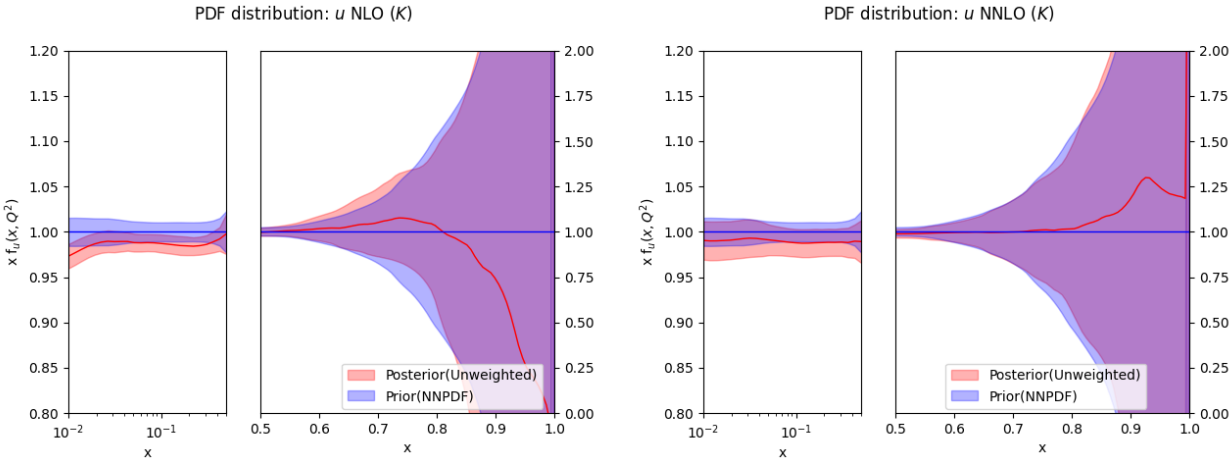

Figure 5.15: Quark u distribution, first iteration, at NLO(left) and NNLO(right) obtained from Kaons data. In blue we have the NNPDF set while in red we have the unweighted set built at the end of the iteration. Central values and uncertainties are normalized over NNPDF set.

which represent the number of s and $\bar{s}$ quarks in the sea respect to the number of $\bar{u}$ and $\bar{d}$ quarks. The plot is shown in fig.5.32. At low x R_s show a similar behavior with a little increase in the uncertainty in the posterior case. At higher x numerical instabilities are present due to the fact that PDFs distributions tends to small values. Despite that we can see that central values remains similar while we have an high decrease in the uncertainties respect to the prior results. This allow me to assume that even though there are some mild impacts on the u, d, s distributions this does not effect the knowledge we already have on the fraction of the strange in the sea. However, the reduction in the uncertainties, at higher x is significant, leading to a more precise measure of this observable.

I have studied the impact of SIDIS data on a PDF set through the reweighting method analyzing the results with comparison plot between the PDFs distribution and computing PDFs distances. There is another analysis which may give more information about this topic, a comparison between the predictions obtained with the prior and posterior sets respect to the experimental Kaons and Pions production data. One plot at a specific binning have been chosen for the different configurations of HERMES and COMPASS datasets. For both, there is a separation between positive and negative Kaons/Pions while for HERMES we have two different targets, Proton and Deuteron leading to 12 plot at each bin. First, I have done two new fit of fragmentation functions for Kaons and Pions separately with the Posterior set fixed at the central value. Then I have computed predictions for each replica of the PDFs sets in order to find the central values and the uncertainties for both Prior and

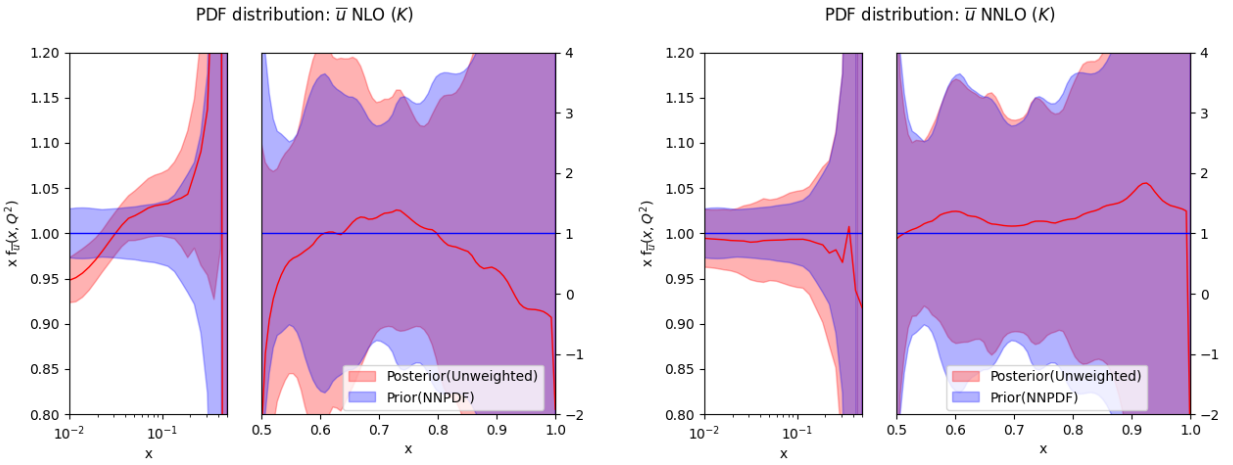

Figure 5.16: Quark $\bar{u}$ distribution, first iteration, at NLO(left) and NNLO(right) obtained from Kaons data. In blue we have the NNPDF set while in red we have the unweighted set built at the end of the iteration. Central values and uncertainties are normalized over NNPDF set.

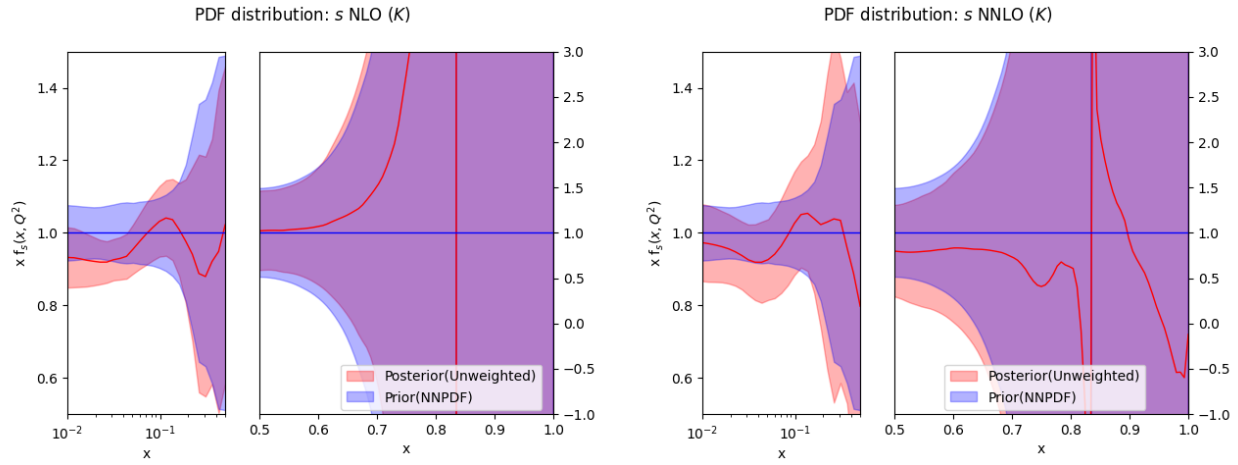

Figure 5.17: Quark s distribution, first iteration, at NLO(left) and NNLO(right) obtained from Kaons data. In blue we have the NNPDF set while in red we have the unweighted set built at the end of the iteration. Central values and uncertainties are normalized over NNPDF set.

Posterior. These plots are shown in Figs. 5.33 5.34 5.35for Kaons and in Figs. 5.36 5.37 5.38 for Pions.

In black, I show the experimental data with their uncertainties. Then, I show in blue predictions and uncertainties for the prior set while in red the predictions obtained with the posterior set. Generally, as I expected, the predictions obtained from the two sets are consistent with each other. For some bins the posterior looks closer to the data while for others is the opposite. However, these shifts are rather small reinforcing the fact that the impact on the prior is moderate.

Even if with some subtlety, every test presented in this section points to the same direction. I can safely assume that, from these results, the impact of SIDIS data on the PDF is rather moderate. This enhance the knowledge we already have on these important objects while it would be interesting to see if the data that will be produced by the EIC could give us more information on the matter.

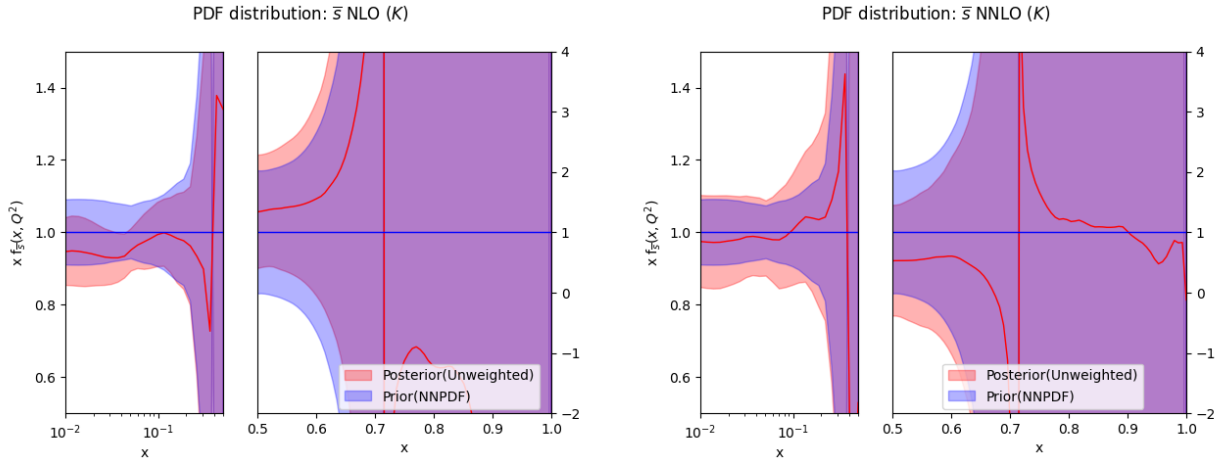

Figure 5.18: Quark $\bar{s}$ distribution, first iteration, at NLO(left) and NNLO(right) obtained from Kaons data. In blue we have the NNPDF set while in red we have the unweighted set built at the end of the iteration. Central values and uncertainties are normalized over NNPDF set.

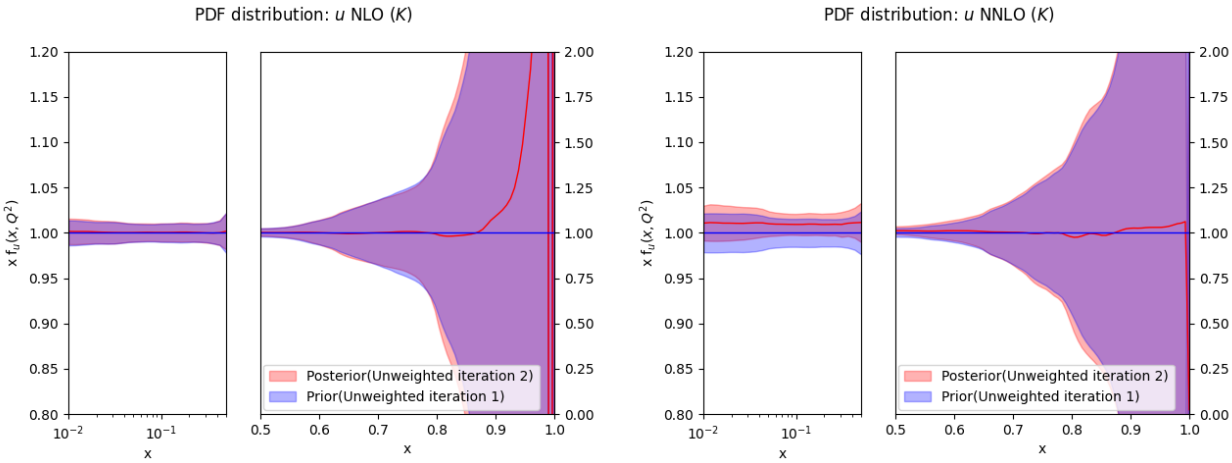

Figure 5.19: Quark u distribution, second iteration, at NLO(left) and NNLO(right) obtained from Kaons data. In blue we have the unweighted set obtained at the first iteration(prior) while in red we have the unweighted set built at the end of the second. Central values and uncertainties are normalized over the prior set.

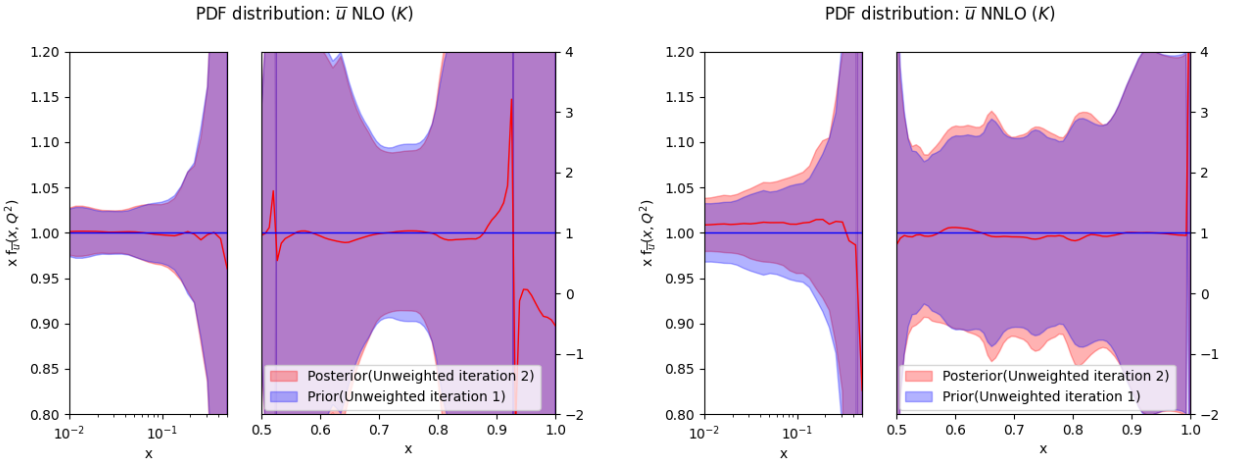

Figure 5.20: Quark $\bar{u}$ distribution,second iteration, at NLO(left) and NNLO(right) obtained from Kaons data. In blue we have the unweighted set obtained at the first iteration(prior) while in red we have the unweighted set built at the end of the second(posterior). Central values and uncertainties are normalized over the prior set.

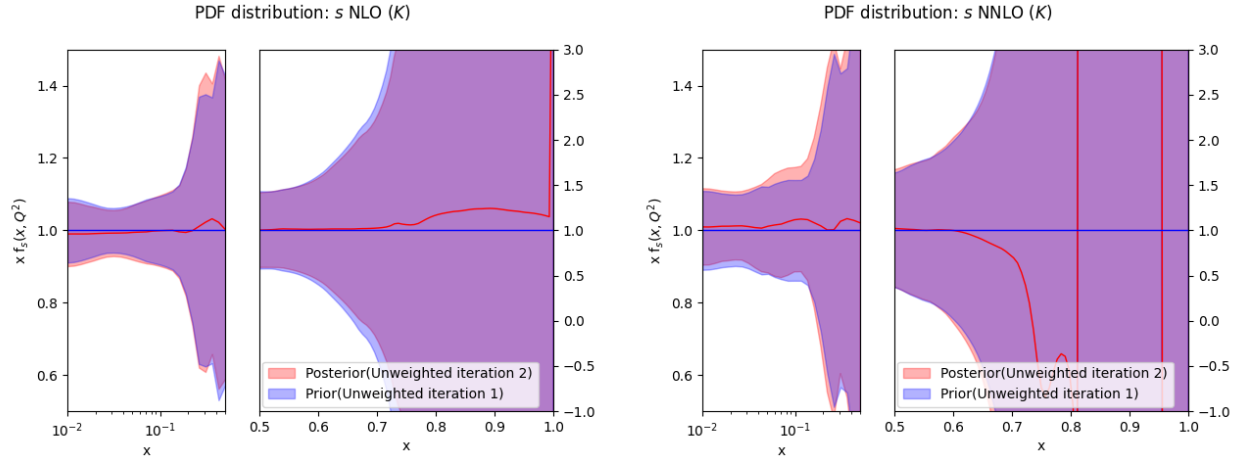

Figure 5.21: Quark s distribution, second iteration, at NLO(left) and NNLO(right) obtained from Kaons data. In blue we have the unweighted set obtained at the first iteration(prior) while in red we have the unweighted set built at the end of the second(posterior). Central values and uncertainties are normalized over the prior set.

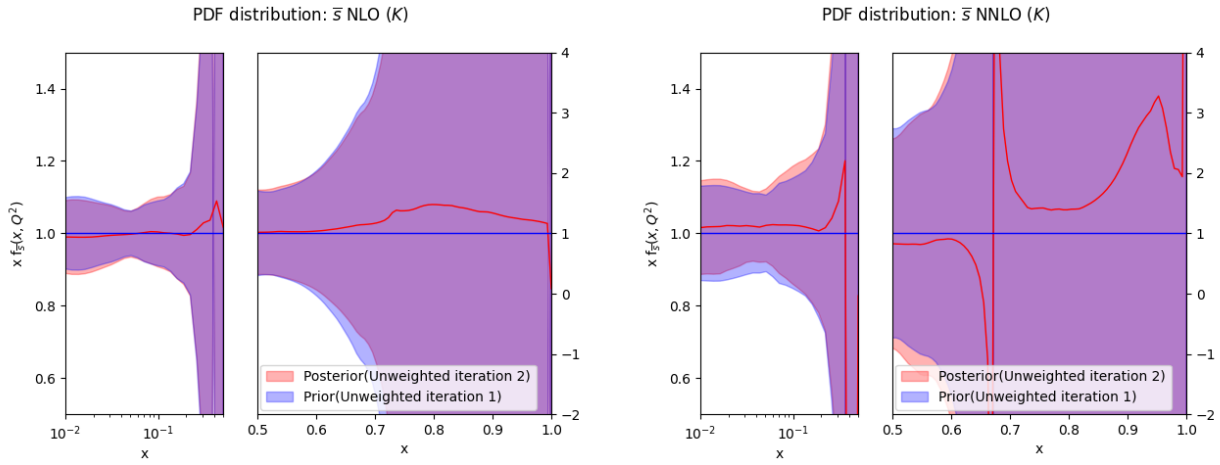

Figure 5.22: Quark $\bar{s}$ distribution, second iteration, at NLO(left) and NNLO(right) obtained from Kaons data. In blue we have the unweighted set obtained at the first iteration while in red we have the unweighted set built at the end of the second(posterior). Central values and uncertainties are normalized over the prior set.

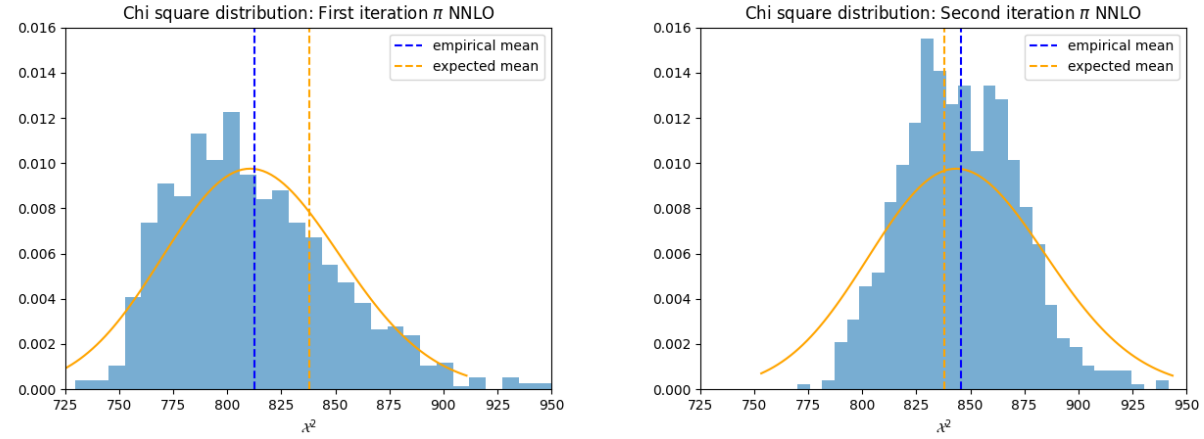

Figure 5.23: Chi square distributions for Pions first(left) and second(right) iteration at NNLO. The orange curve represent the expected distribution which have been centered on the empirical mean to be able to compare the shape of the two distributions. The orange dotted line represent the mean of the orange distribution.

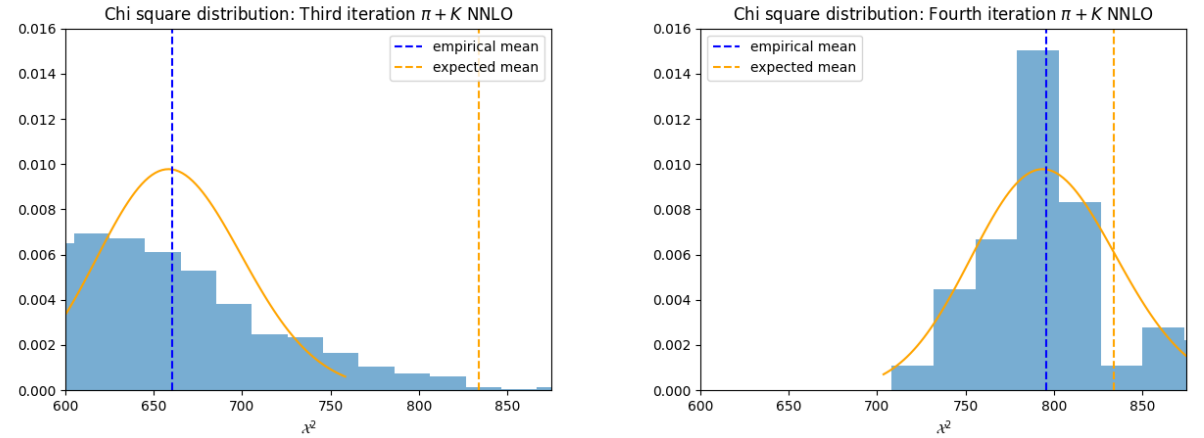

Figure 5.24: Chi square distributions for Kaons first(left) and second(right) iteration at NNLO. The orange curve represent the expected distribution which have been centered on the empirical mean to be able to compare the shape of the two distributions. The orange dotted line represent the mean of the orange distribution.

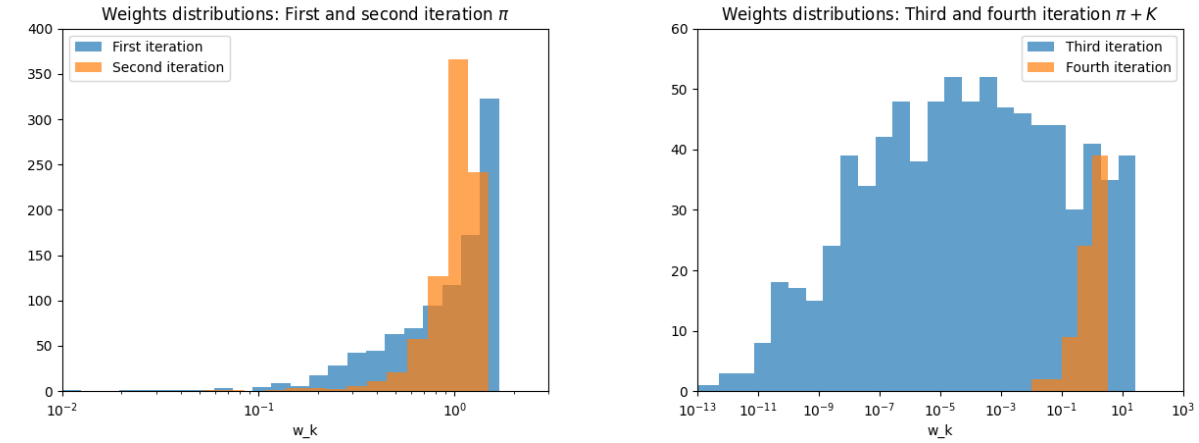

Figure 5.25: Weights distribution for Pions at NLO(left) and NNLO(right). The orange histogram shows the second iteration while the blue one the first.

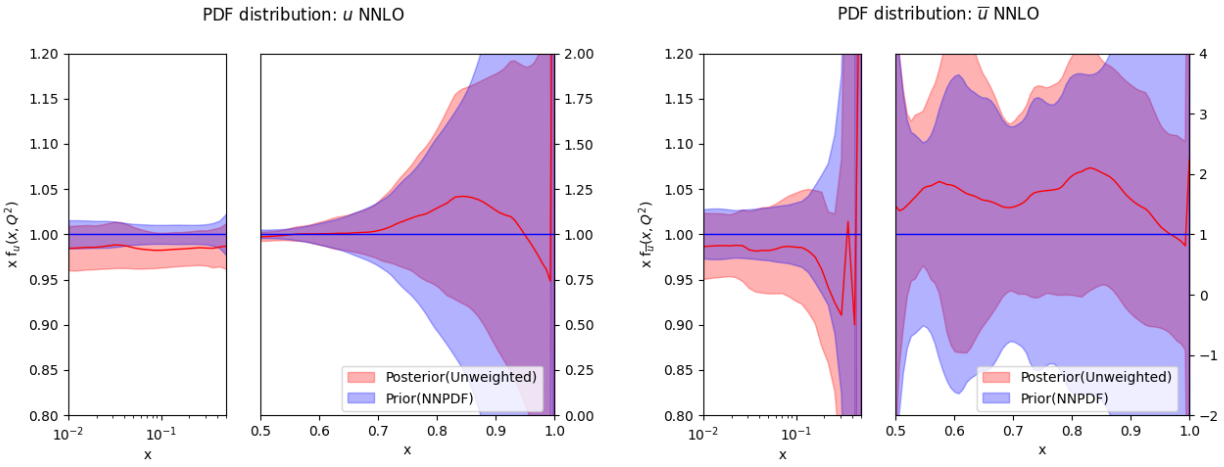

Figure 5.26: Quark u(left) and $\bar{u}$(right) distributions. In blue, the initial PDF set(prior) while in red, the unweighted set built at the end of the fourth iteration(posterior). Central values and uncertainties are normalized over the prior set.

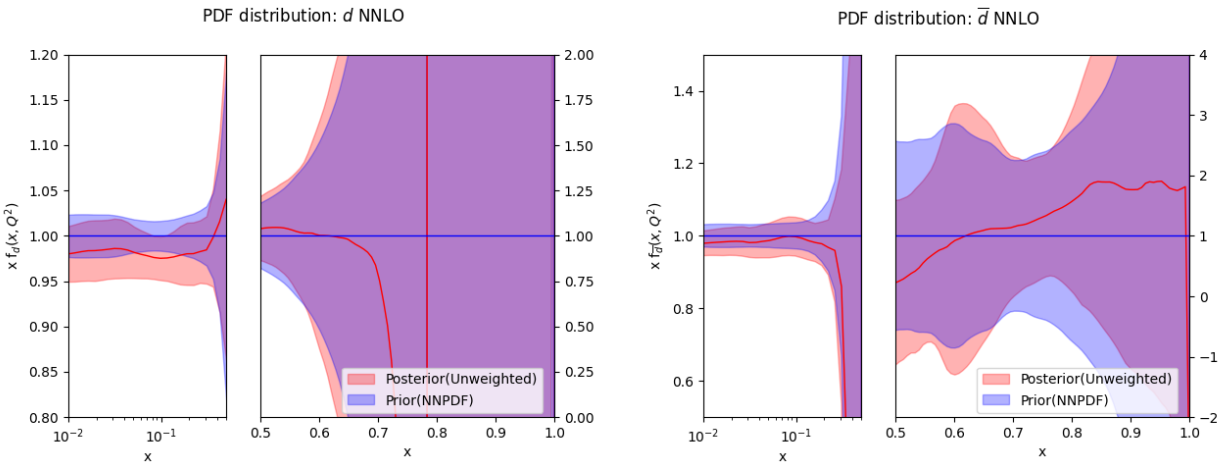

Figure 5.27: Quark d(left) and $\bar{d}$(right) distributions. In blue, the initial PDF set(prior) while in red, the unweighted set built at the end of the fourth iteration(posterior). Central values and uncertainties are normalized over the prior set.

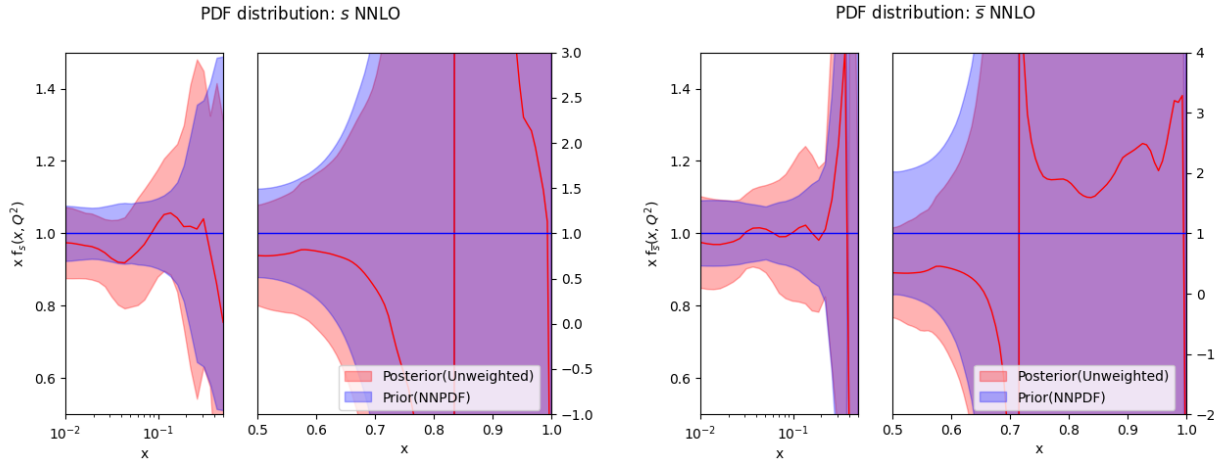

Figure 5.28: Quark s(left) and $\bar{s}$(right) distributions. In blue, the initial PDF set(prior) while in red, the unweighted set built at the end of the fourth iteration(posterior). Central values and uncertainties are normalized over the prior set.

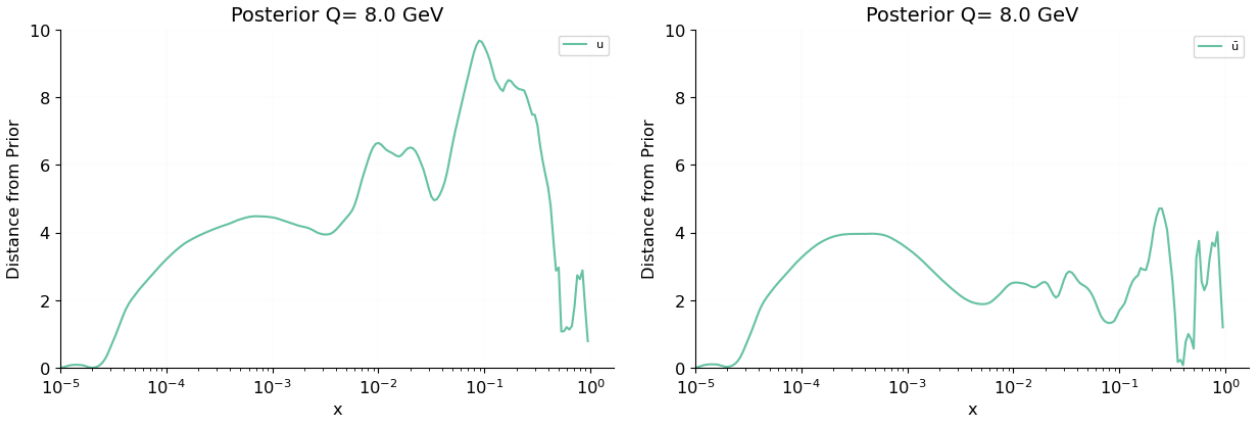

Figure 5.29: Distances between prior and posterior for Quark u(left) and $\bar{u}$(right).

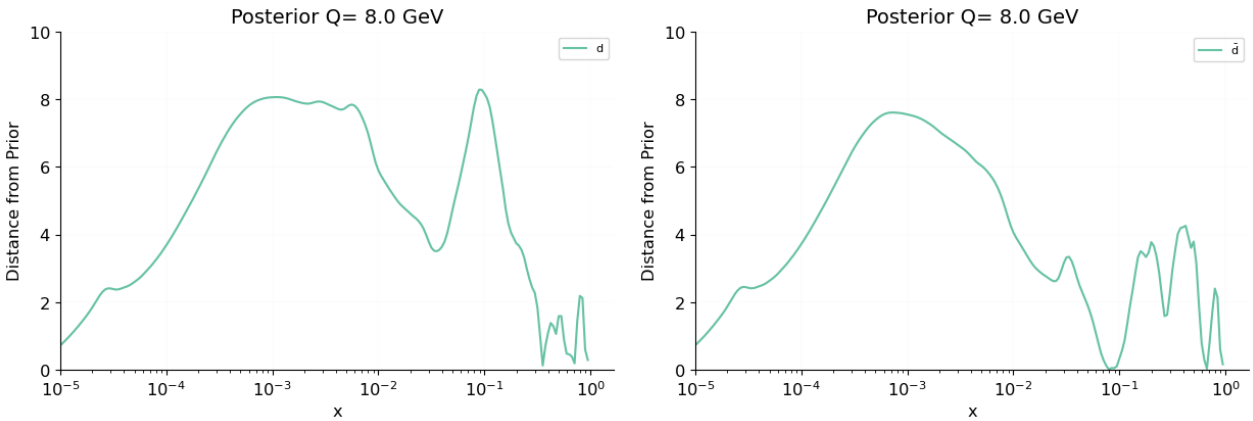

Figure 5.30: Distances between prior and posterior for Quark d(left) and $\bar{d}$(right).

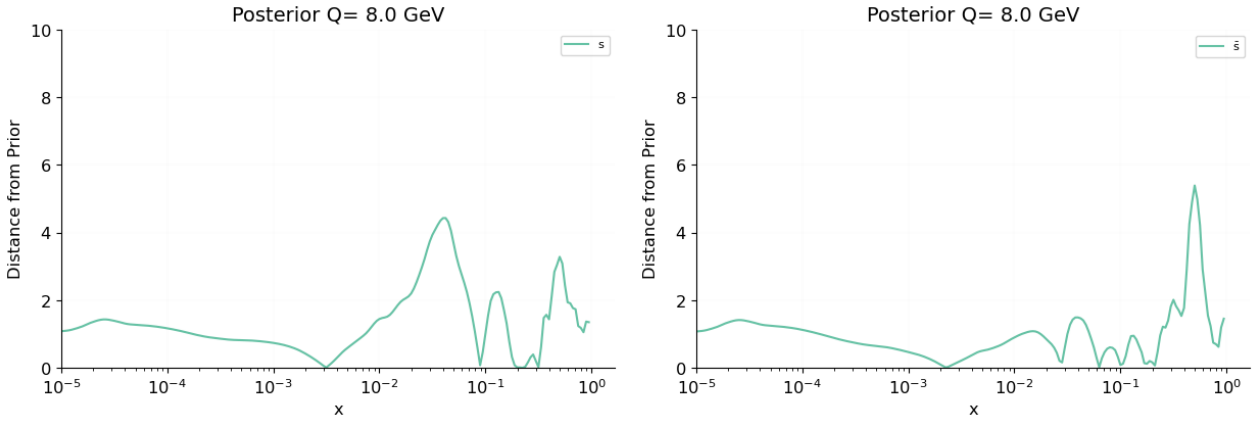

Figure 5.31: Distances between prior and posterior for Quark s(left) and $\bar{s}$(right).

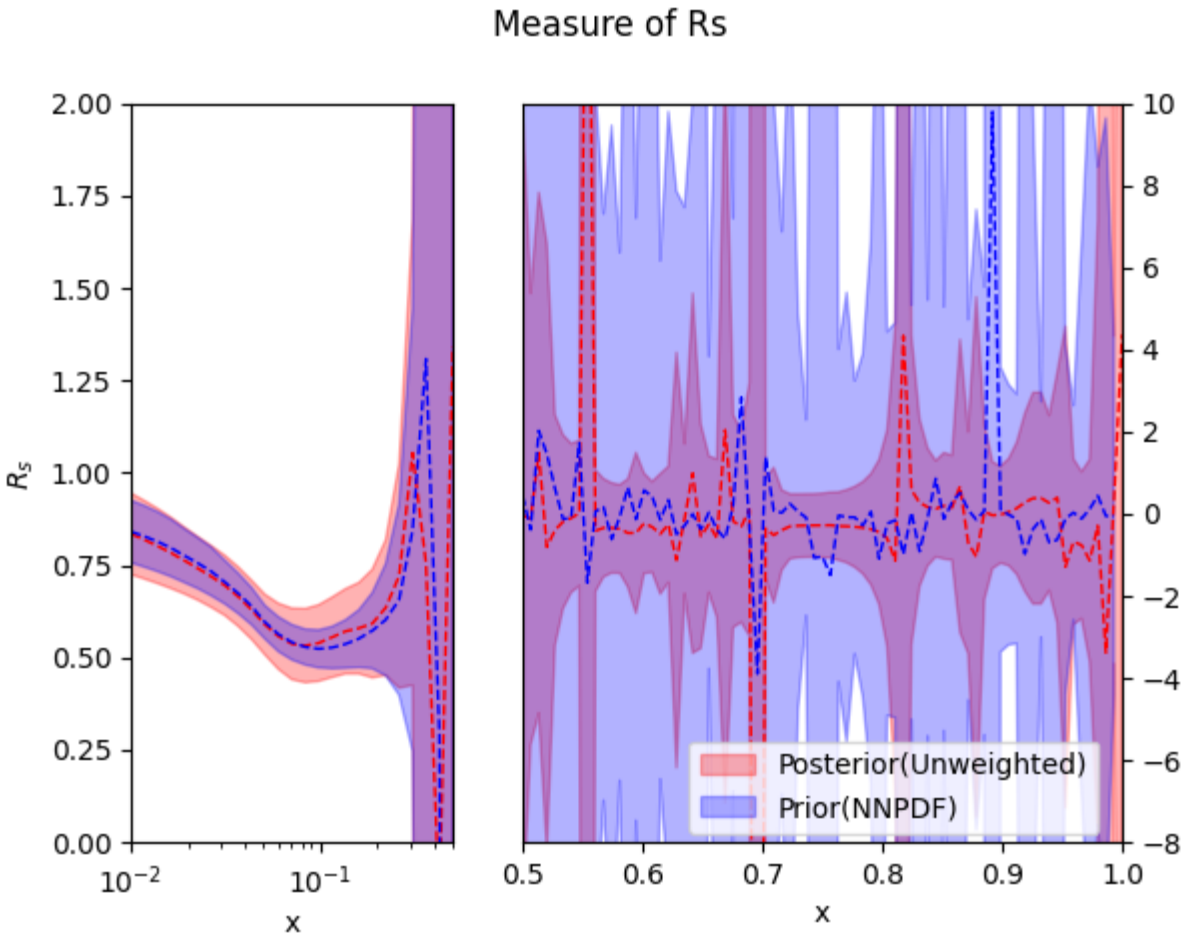

Figure 5.32: Plot of R_s which represent the number of s and $\bar{s}$ quarks in the sea respect to the number of u and $\bar{u}$ quarks. R_s have been computed for the prior set (blue) and posterior set (red).

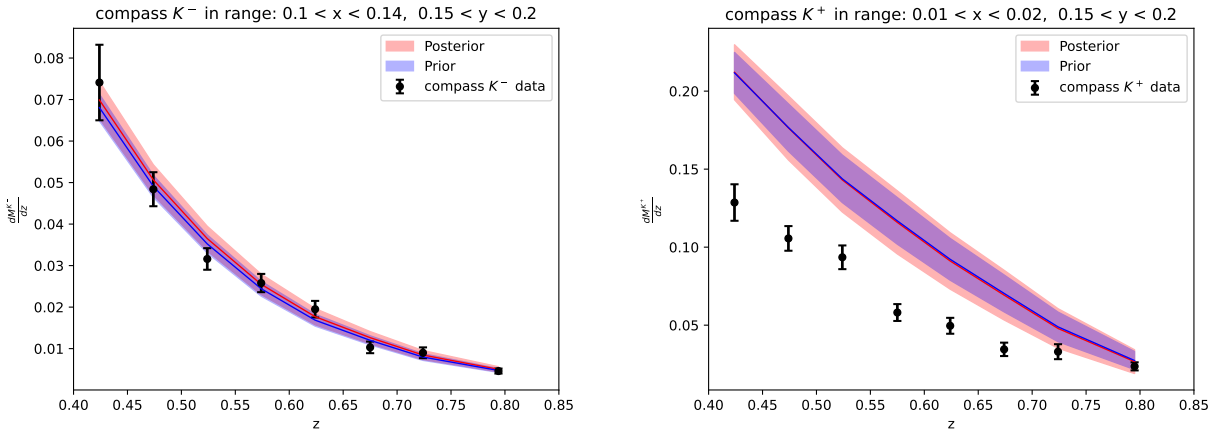

Figure 5.33: Comparison between the predictions computed with Prior and Posterior set and the experimental data from COMPASS experiment.

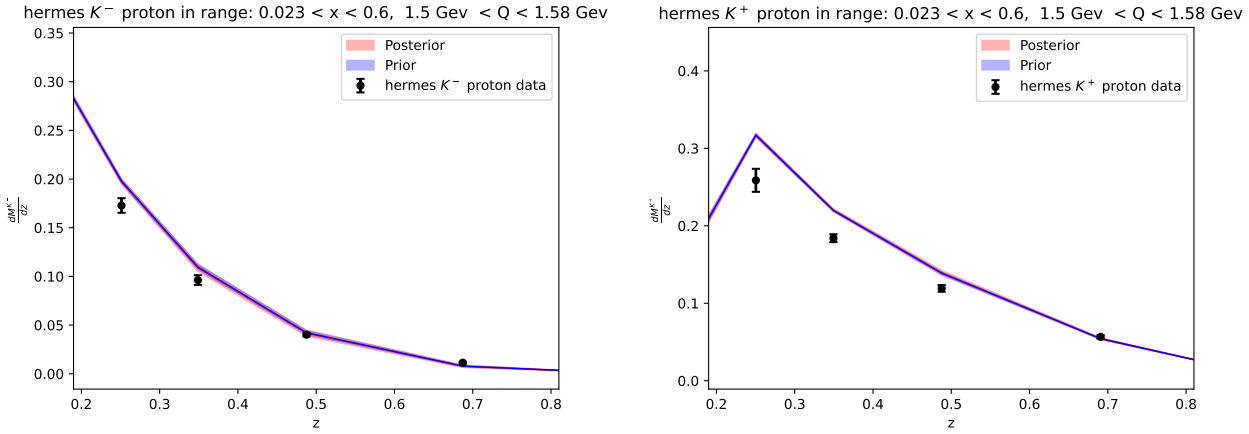

Figure 5.34: Comparison between the predictions computed with Prior and Posterior set and the experimental data from HERMES experiment with a Proton target.

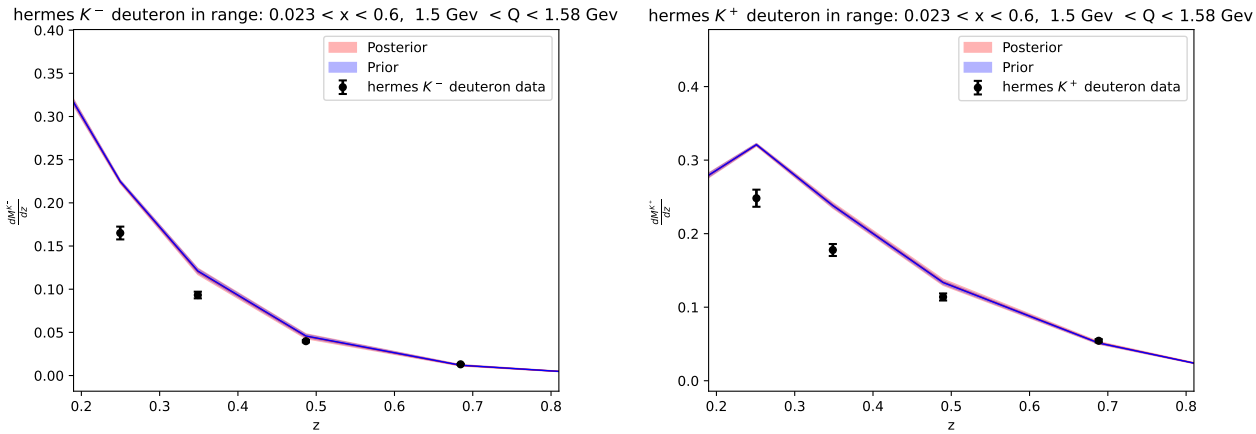

Figure 5.35: Comparison between the predictions computed with Prior and Posterior set and the experimental data from HERMES experiment with a Deuteron target.

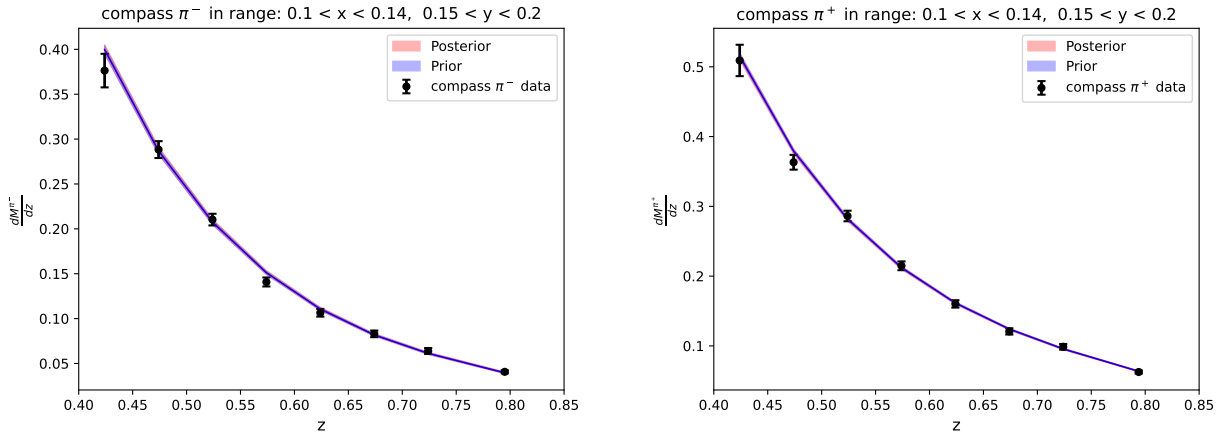

Figure 5.36: Comparison between the predictions computed with Prior and Posterior set and the experimental data from COMPASS experiment.

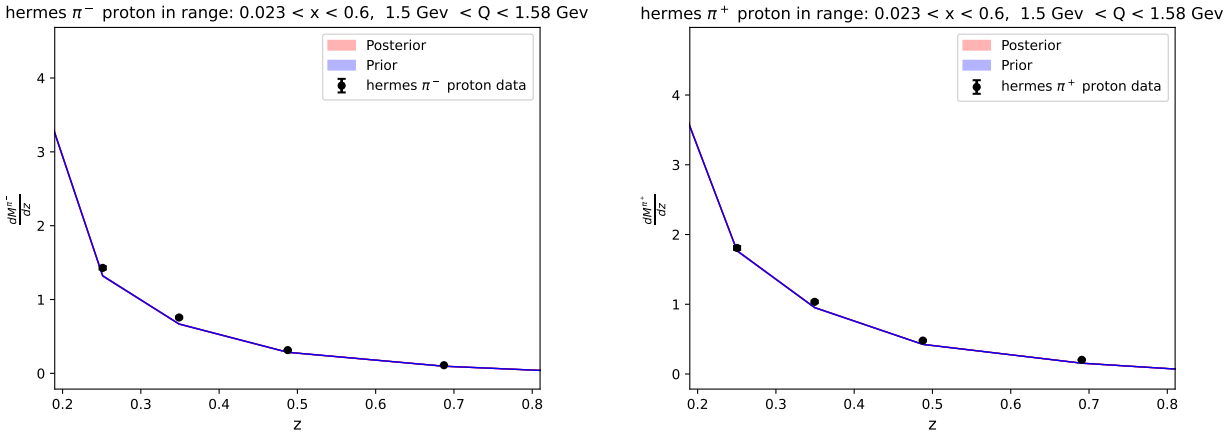

Figure 5.37: Comparison between the predictions computed with Prior and Posterior set and the experimental data from HERMES experiment with a Proton target.

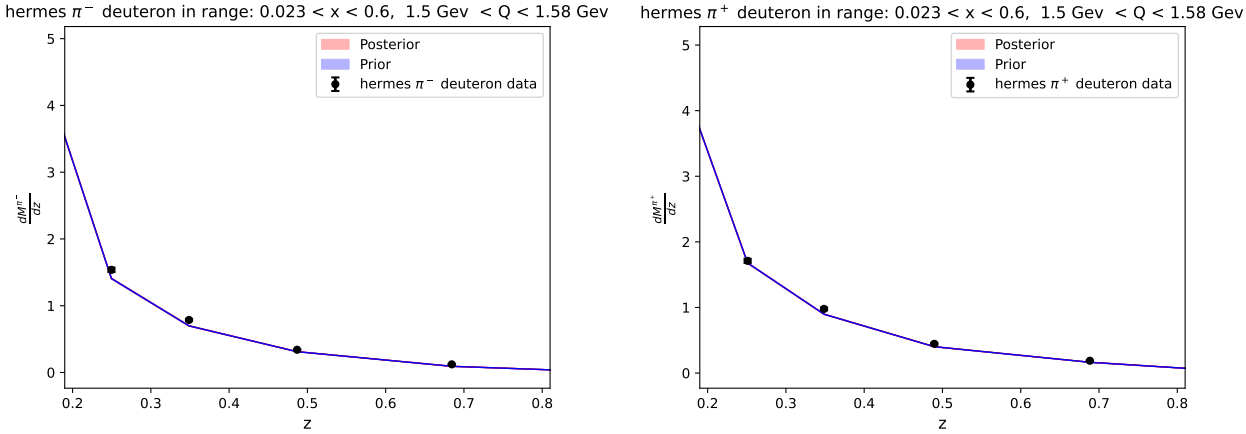

Figure 5.38: Comparison between the predictions computed with Prior and Posterior set and the experimental data from HERMES experiment with a Deuteron target.

Chapter 6

Conclusion

In this thesis, I performed an in-depth analysis of the impact of Semi-Inclusive Deep Inelastic Scattering (SIDIS) data on Parton Distribution Functions (PDFs) at Next-to-Leading Order (NLO) and Next-to-Next-to-Leading Order (NNLO).

This was accomplished through an iterative procedure involving an initial fit of fragmentation functions with fixed PDFs, followed by the application of a reweighting method. This approach allows for the inclusion of new datasets without requiring a complete refit, relying only on the knowledge of the χ^2 function. However, a limitation of this method is that the weights assigned to the PDF replicas are not necessarily equal to 1. To address this, an additional procedure known as unweighting was applied, enabling the construction of a PDF set in the LHAPDF format while retaining all information obtained through reweighting.

The analysis was conducted using datasets from two different experiments: COMPASS and HERMES. Both experiments measure hadron multiplicities for positive and negative Kaon and Pion production.

An initial study at NLO revealed a significant impact on the central values and uncertainties of the PDFs. However, the same level of impact was not observed at NNLO for both Pions and Kaons. This discrepancy arises from the fact that the increased perturbative accuracy at NNLO affects the influence of new data, reducing its relative impact. Consequently, I decided to continue the analysis exclusively at NNLO.

The iterative procedure was first applied separately to Pions and Kaons over two iterations each. To assess the convergence of the method, a study of the χ^2 and weight distributions was performed. In both cases, it was observed that, with each iteration, the χ^2 distribution approached the expected theoretical behavior, while the weights concentrated more around 1. This indicates that the procedure effectively integrates new information into the PDF set. This conclusion was further supported by comparison plots between the initial (prior) and final (posterior) PDF sets.

Generally, I found that the impact of SIDIS data on Pions is smaller than on Kaons, with the latter exhibiting more pronounced effects, particularly at low x. The final step of the analysis involved applying the procedure to both Pions and Kaons together. To achieve this, I first applied two iterations to the Pion data, which had a lesser impact on the PDFs, followed by two iterations incorporating the Kaon data. This yielded the final PDF set, which will be used for further studies.

To quantify the impact of SIDIS data more precisely, I computed the statistical distance between the central values of the prior and posterior sets. A distance of $d \sim 8$ corresponds to a 1σ difference between the two distributions. The results are consistent with the comparison plots, showing a more substantial impact at low x for the u, d, and s quarks, which in some cases reach a maximum distance of $d \sim 9$.

Given the observed impact, albeit mild, I also investigated the fraction of strange quarks in the sea, denoted as R_s. This quantity was computed and compared for both the prior and posterior PDF sets. The results indicate no significant shift in the central value of R_s, although a notable reduction in uncertainties was observed.

Finally, I compared the prior and posterior PDFs with the experimental SIDIS data. As expected, the predictions from both sets are consistent with each other. Some minor shifts are present in specific kinematic regions, but no clear systematic trend emerges.

In conclusion, the information provided by the current SIDIS data has only a mild impact on our knowledge of the Parton Distribution Functions. From the perspective of this thesis, it would be valuable to explore the potential impact of the future Electron-Ion Collider (EIC) experiment on our understanding of PDFs. Such a study could provide useful insights into the minimum performance requirements the experiment should meet to significantly improve our knowledge of the nucleon structure.

Appendix A

Proof of the weight formula

In this Appendix, I provide a proof of Eq. (4.12). We begin by considering the probability $P(f)$ for the PDF f. This is the probability $P(f \mid K)$, where K denotes all the data used in the determination and their associated errors, the values of parameters such as α_s and heavy quark masses used in the computation of the data expected from the PDF, and finally also the theoretical framework used. If we then wish to extend the dataset by including new data y, the new probability $P_{new}(f)$ is then $P(f \mid yK)$: besides K we now also assume the new data y.

The new probability is then determined from the old probability using the sampling distribution $P(y \mid fK)$ and multiplicative rule for probabilities (Bayes theorem):

$$P(AB \mid C) = P(A \mid BC)P(B \mid C) = P(B \mid AC)P(A \mid C) \tag{A.1}$$

Naively applying this result in the present case we have

$$P(f \mid yK)P(y \mid K) = P(y \mid fK)P(f \mid K) \tag{A.2}$$

whence (replacing $P(f \mid K)$ with $\mathcal{P}(f \mid K)Df, P(f \mid yK) with \mathcal{P}(f \mid yK)Df$

$$\mathcal{P}(f \mid yK) = \frac{P(y \mid fK)\mathcal{P}(f \mid K)}{P(y \mid K)} \tag{A.3}$$

Note that $P(y \mid K)$ does not depend on the PDF f, and can thus be determined simply by insisting that $\mathcal{P}(f \mid yK)$ is properly normalized: we then find

$$P(y \mid K) = \int P(y \mid fK)\mathcal{P}(f \mid K)Df \tag{A.4}$$

so

$$\mathcal{P}(f \mid yK) = P(y \mid fK)\mathcal{P}(f \mid K) / \int P(y \mid fK)\mathcal{P}(f \mid K)Df \tag{A.5}$$

where everything on the right hand side is now known. This argument would work without problems if the data y could only take discrete values. The difficulty in the present case is that our data are continuous, so rather than the probability $P(y \mid fK)$ we have to work with a multi-dimensional probability density $\mathcal{P}(y \mid fK)d^n y$, in a limit in which the volume element $d^n y$ goes to zero. Of course in this limit the probabilities $P(y \mid fK)$ and $P(y \mid K)$ also go to zero, and we find a ratio of two zeros in EQ... The conditional probability $P(f \mid yK)$ is then only well defined if we specify carefully the way in which the limit is to be taken: probabilities conditional on sets of measure zero are ambiguous. Failure to specify the limiting process can result in contradictions.

Consider then the probability density for the data y. Assuming that the new experiments are not correlated with any of the experiments used in the determination of the initial probability density, the probability density of y is then given by Eq. (A.4):

$$\mathcal{P}(y \mid K) = \int \mathcal{P}(y \mid fK)\mathcal{P}(f \mid K)Df = \frac{1}{N}\sum_{k=1}^{N}\mathcal{P}(y \mid f_k K), \tag{A.6}$$

where in the second step we used Eq. (A.4). The density $\mathcal{P}(y \mid fK)$ gives the probability that new data lie in an infinitesimal volume $d^n y$ centered at y in the space of possible data given a particular choice of PDF f: it is

often called the sampling distribution or the likelihood function. Assuming that the uncertainties in the data are purely Gaussian,

$$\mathcal{P}(y \mid fK)d^n y = (2\pi)^{-n/2}(det\sigma_{ij})^{-1/2}e^{-\frac{1}{2}\chi^2(y,f)}d^n y, \tag{A.7}$$

where $\chi^2(y,f)$ is calculated using Eq. (4.9). The volume element $d^n y$ is independent of f: without a specific prediction, all data are assumed equally likely. Since to compute $\mathcal{P}(\chi \mid fK)$ it is sufficient to compute $\chi^2(y,f)$, we can consider the probability density for the χ^2 to the new dataset:

$$\mathcal{P}(\chi \mid fK)d\chi = 2^{1-n/2}(\Gamma(n/2))^{-1}(\chi(y,f))^{n-1}e^{-\frac{1}{2}\chi^2(y,f)}d\chi \tag{A.8}$$

where $\chi(y,f) \equiv (\chi^2(y,f))^{1/2}$. This distribution may be readily derived from Eq. (A.7) by diagonalising the covariance matrix and rescaling the data to a set Y_i of independent Gaussian variables each with unit variance. Then $d^n y = (det\sigma_{ij})^{1/2}d^n Y$, and $\chi^2 = \sum_{i=1}^n Y_i^2$. Choosing n-dimensional spherical co-ordinates in the space of data(with χ as the radial co-ordinate, and thus $y = y[f]$ as the origin), we may write $d^n Y = A_n \chi^{n-1} d\chi d^{n-1}\Omega$, where $d^{n-1}\Omega$ is the measure on the sphere and $A_n = 2\pi^{n/2}(\Gamma(n/2))^{-1}$ is the area of the unit sphere in n-dimensions. The probability Eq. (A.7) may thus be written

$$\begin{aligned}\mathcal{P}(y \mid fK)d^n y &= (2\pi)^{-n/2}e^{-\frac{1}{2}\chi^2(y,f)}d^n Y \\ &= 2^{1-n/2}(\Gamma(n/2))^{-1}(\chi(y,f))^{n-1}e^{-\frac{1}{2}\chi^2(y,f)}d\chi d^{n-1}\Omega\end{aligned} \tag{A.9}$$

which can be written as

$$\mathcal{P}(y \mid fK)d^n y = \mathcal{P}(\chi \mid fL)d\chi d^{n-1}\Omega \tag{A.10}$$

Again the probability density $\mathcal{P}(\chi \mid K)$ for the χ of the new dataset is obtained by averaging over replicas:

$$\mathcal{P}(\chi \mid K) = \int \mathcal{P}(\chi \mid fK)\mathcal{P}(f \mid K)Df = \frac{1}{N}\sum_{k=1}^{N}\mathcal{P}(\chi \mid f_k K) \tag{A.11}$$

so combining Eq. (A.6), Eq. (A.10), Eq. (A.11)

$$\mathcal{P}(y \mid K)d^n y = \frac{1}{N}\sum_{k=1}^{N}\mathcal{P}(\chi \mid f_k K)d\chi d^{n-1}\Omega = \mathcal{P}(\chi \mid K)d\chi d^{n-1}\Omega \tag{A.12}$$

since both the volume factor $d^{n-1}\Omega$ and the interval $d\chi$ are independent of the choice of replica, and may thus be taken out of the sum: this follows directly from the assumption that the measure $d^n y$ in Eq. .. is independent of f. The advantage of using $\mathcal{P}(\chi \mid fK)$ when evaluating Eq. (A.3) is that $\mathcal{P}(\chi \mid fK)$ is only a one dimensional density, so taking the limit in which the volume element goes to zero is straightforward and unambiguous. We may write Eq. .(A.3) as

$$\mathcal{P}(f \mid \chi K)Df\mathcal{P}(\chi \mid K)d\chi = \mathcal{P}(\chi \mid fK)d\chi\mathcal{P}(f \mid K)Df. \tag{A.13}$$

The marginalization Eq. (A.11) follows directly on integration over f, since if $\mathcal{P}(f \mid \chi)$ is correctly normalized, $\mathcal{P}(f \mid \chi K)Df = 1$. Now, canceling the $d\chi$ from either side of Eq. (A.13)

$$\mathcal{P}(f \mid \chi K)Df = \frac{\mathcal{P}(\chi \mid fK)}{\mathcal{P}(\chi \mid K)}\mathcal{P}(f \mid K)Df. \tag{A.14}$$

Multiplying on both sides by some observable $\mathcal{O}[f]$ and integrating over the PDFs,

$$\langle\mathcal{O}\rangle_{new} = \int \mathcal{O}[f]\mathcal{P}(f \mid \chi K)Df \tag{A.15}$$

$$= \int \mathcal{O}[f]\frac{\mathcal{P}(\chi \mid fK)}{\mathcal{P}(\chi \mid K)}\mathcal{P}(f \mid K)Df \tag{A.16}$$

$$= \frac{1}{N}\sum_{k=1}^{N}\frac{\mathcal{P}(\chi \mid f_k K)}{\mathcal{P}(\chi \mid K)}\mathcal{O}[f_k] \tag{A.17}$$

where in the last line we used Eq. (4.5). This corresponds to the reweighting with weights

$$w_k = \frac{\mathcal{P}(\chi \mid f_k K)}{\mathcal{P}(\chi \mid K)} \tag{A.18}$$

Combining Eq. (A.18) with Eq. (A.8) and Eq. (A.11) we obtain Eq. (4.12) Note that a further application of Beys' theorem to Eq. (A.18) gives the alternative form

$$w_k = \frac{P(f_k \mid \chi K)}{P(f_k \mid K)} = NP(f_k \mid \chi K), \tag{A.19}$$

since because the replicas are uniformly distributed, $P(f_k \mid K) = 1/N$. Thus w_k/N is the probability of replica f_k given the χ to the new data.

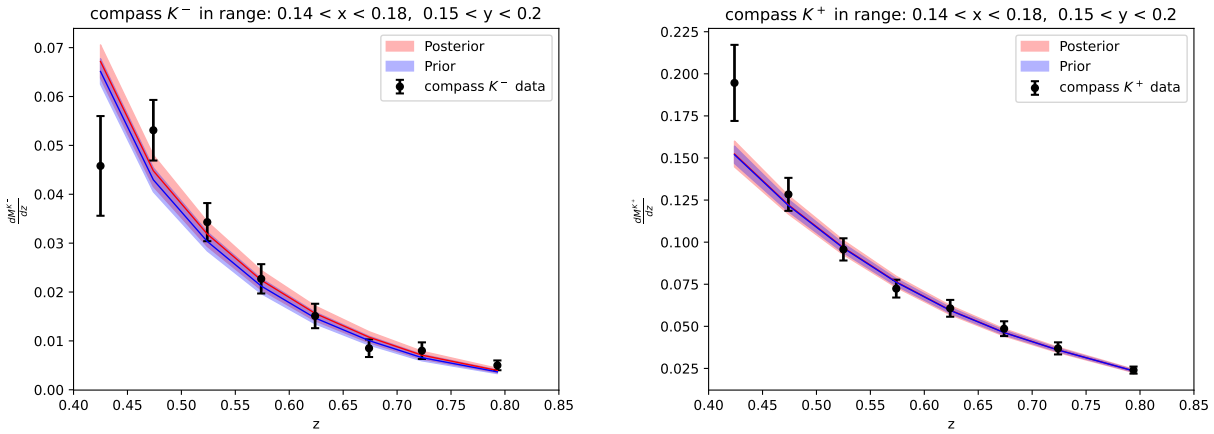

Figure B.1: Comparison between the predictions computed with Prior and Posterior set and the experimental data from COMPASS experiment.

Appendix B

Additional results

In this appendix I present some of the plots that I have not shown in the main part of the thesis. These results are obtained with the final unweighted set, i.e. the PDFs where both Kaons and Pions data are taken in consideration. The list of figures is as follow:

- In Figs. B.1, B.2, B.3, B.4, B.5, I show the comparison between the theoretical predictions obtained with the Prior (NNPDF3.1) and Posterior (Unweighted) sets with Kaon production data.
- In Figs. B.6, B.7, B.8,B.9, B.10, I show the comparison between the theoretical predictions obtained with the Prior (NNPDF3.1) and Posterior (Unweighted) sets with Pion production data.
- In Figs. B.11, B.12, B.13, I show the impact of SIDIS data on the other quarks and gluon. In blue I show the Prior set while in red the Posterior. Central values and uncertainties are both normalized over the Prior.
- In Fig. B.14, I show all the distances for the $u, d, s, \overline{u}, \overline{d}, \overline{s}$ quarks at a linear scale.
- In Fig. B.15, I show all the distances for the $c, b, \overline{c}, \overline{b}$ quarks and the gluon in both linear and logarithmic scale.

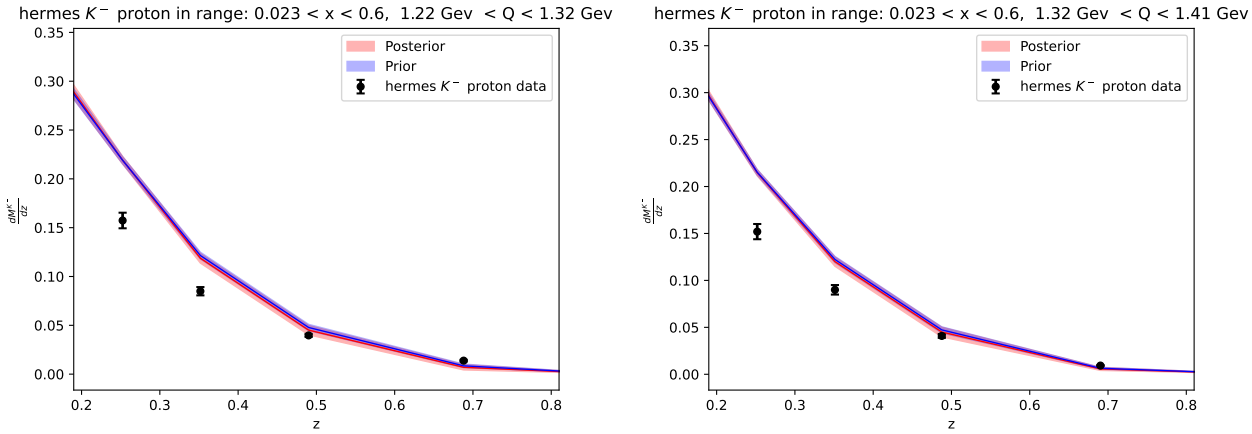

Figure B.2: Comparison between the predictions computed with Prior and Posterior set and the experimental data from HERMES experiment with a Proton target.

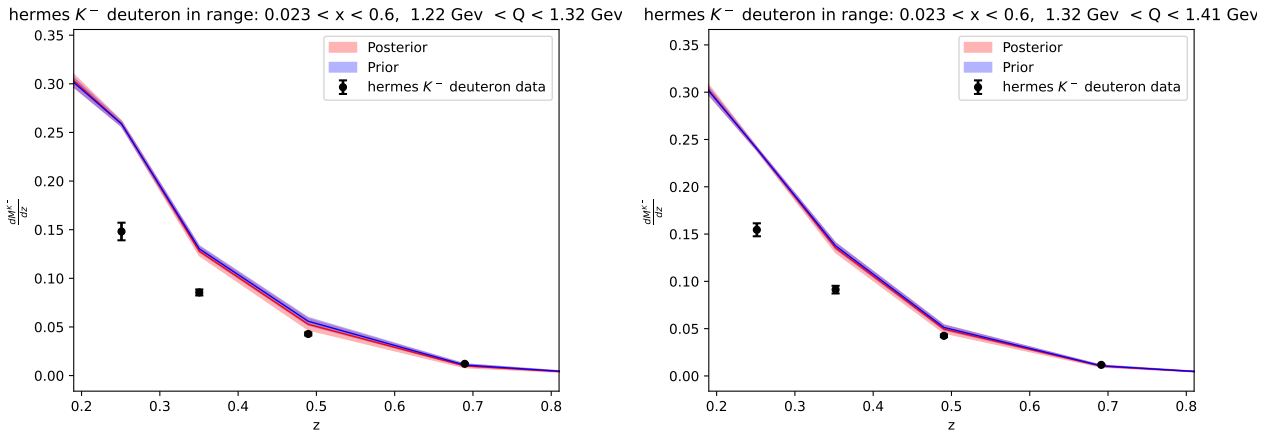

Figure B.3: Comparison between the predictions computed with Prior and Posterior set and the experimental data from HERMES experiment with a Deuteron target.

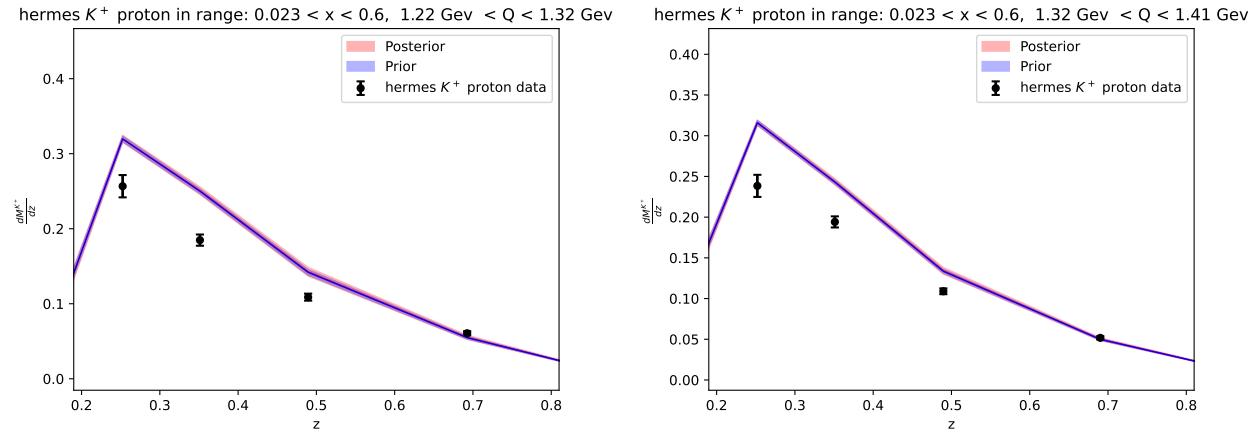

Figure B.4: Comparison between the predictions computed with Prior and Posterior set and the experimental data from HERMES experiment with a Proton target.

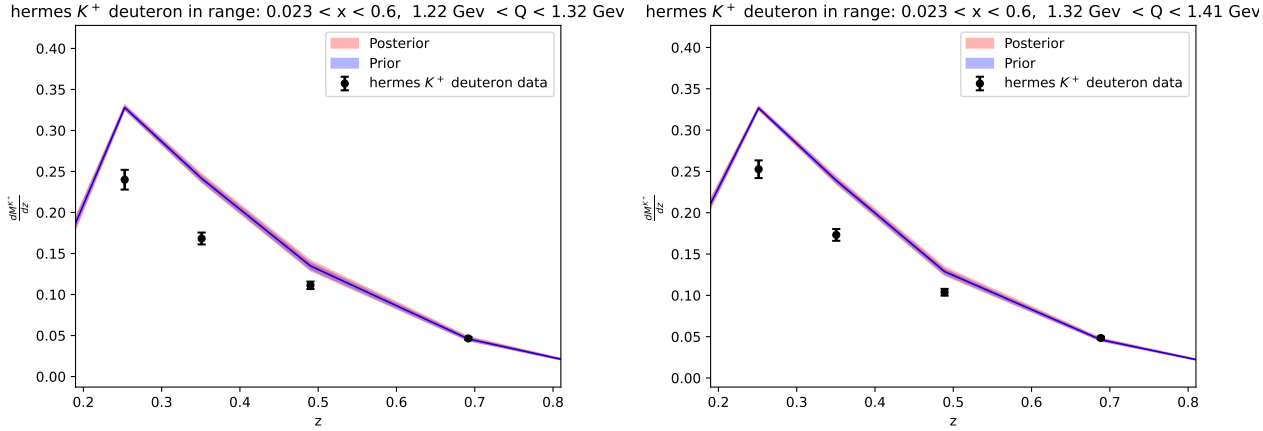

Figure B.5: Comparison between the predictions computed with Prior and Posterior set and the experimental data from HERMES experiment with a Deuteron target.

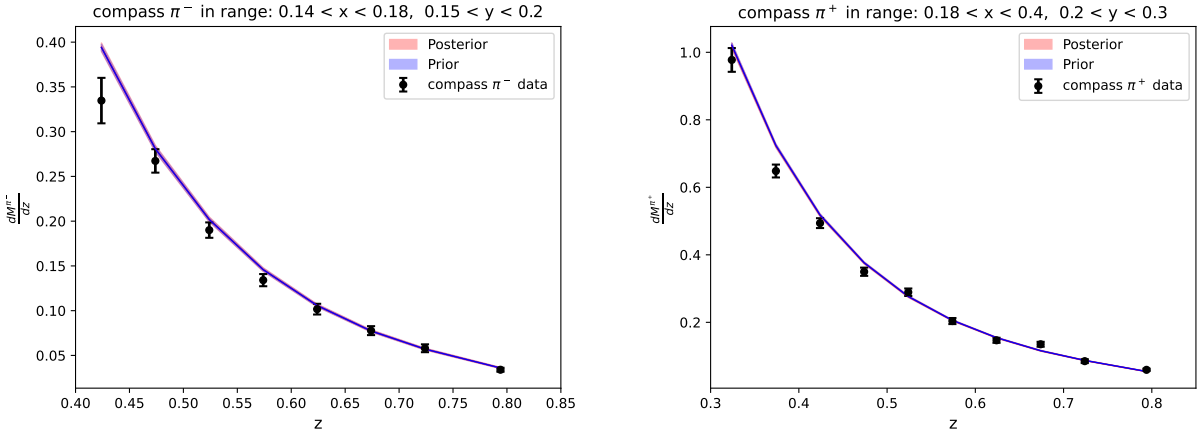

Figure B.6: Comparison between the predictions computed with Prior and Posterior set and the experimental data from COMPASS experiment.

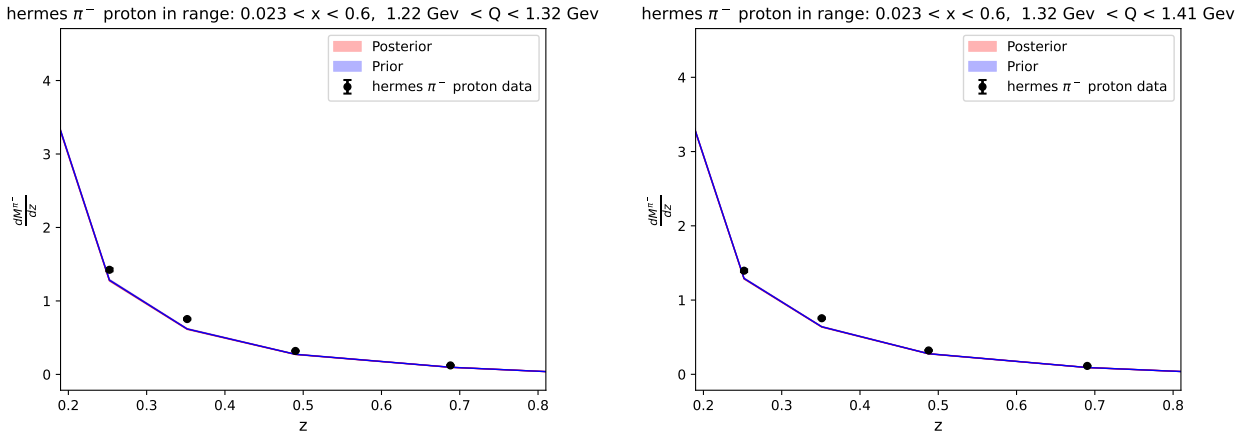

Figure B.7: Comparison between the predictions computed with Prior and Posterior set and the experimental data from HERMES experiment with a Proton target.

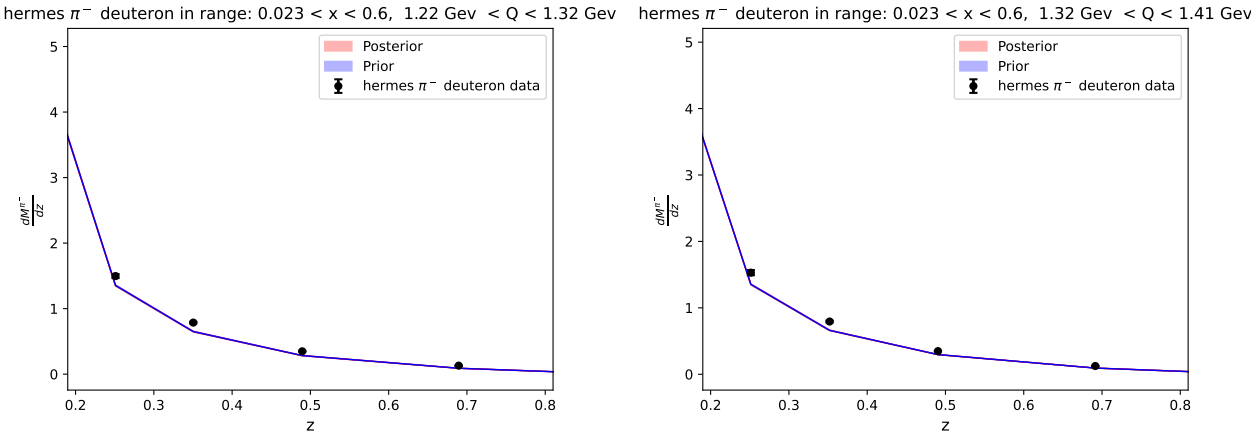

Figure B.8: Comparison between the predictions computed with Prior and Posterior set and the experimental data from HERMES experiment with a Deuteron target.

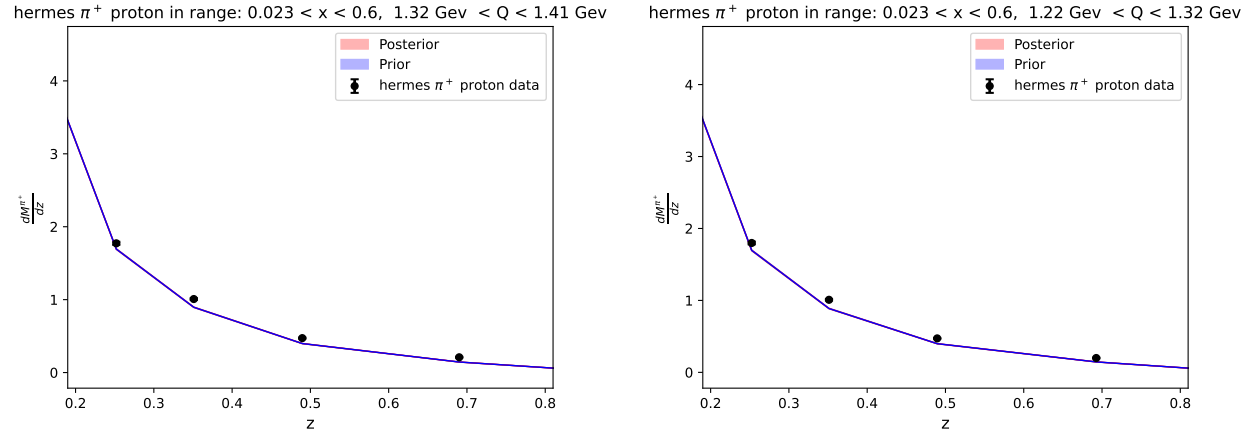

Figure B.9: Comparison between the predictions computed with Prior and Posterior set and the experimental data from HERMES experiment with a Proton target.

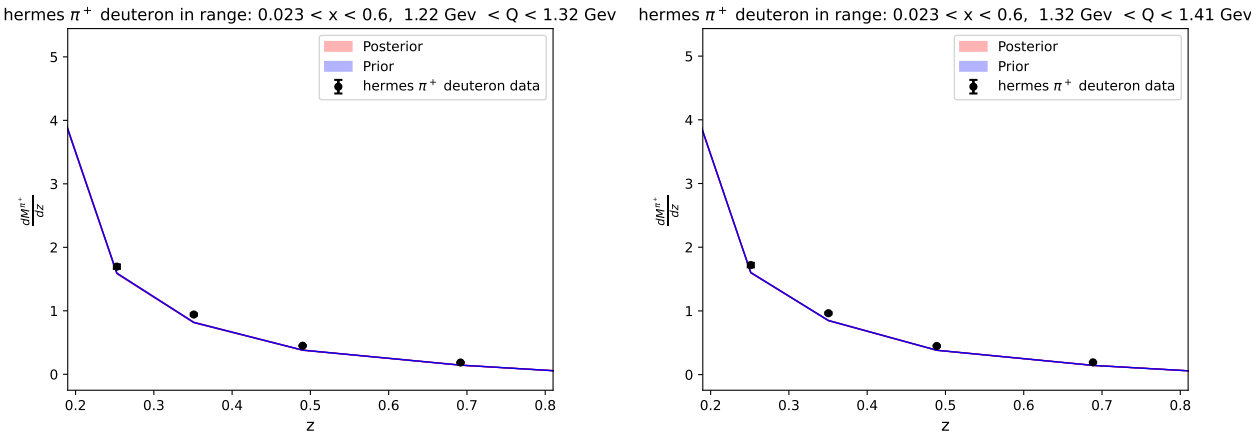

Figure B.10: Comparison between the predictions computed with Prior and Posterior set and the experimental data from HERMES experiment with a Deuteron target.

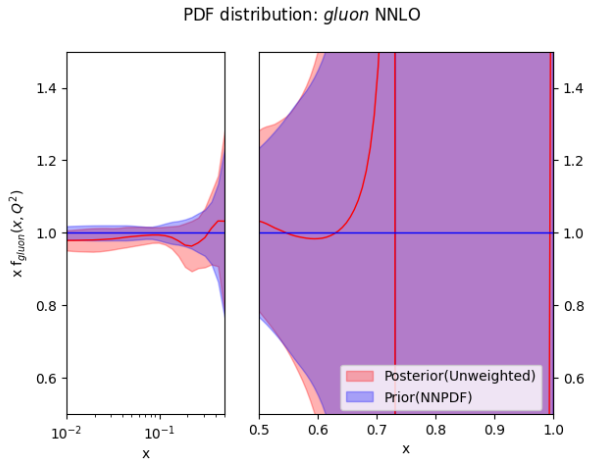

Figure B.11: Gluon distribution at final iteration with the Kaons + Pions data. In blue we have the unweighted set obtained at the first iteration(prior) while in red we have the unweighted set built at the end of the second. Central values and uncertainties are normalized over the prior set.

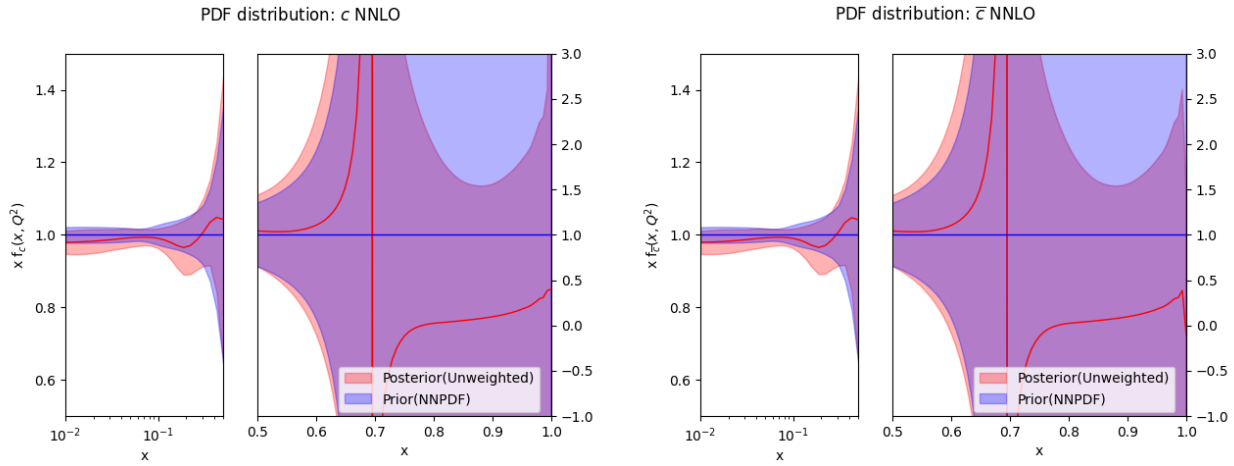

Figure B.12: Quark c and $\bar{c}$ distributions at final iteration with the Kaons + Pions data. In blue we have the unweighted set obtained at the first iteration(prior) while in red we have the unweighted set built at the end of the second(posterior). Central values and uncertainties are normalized over the prior set.

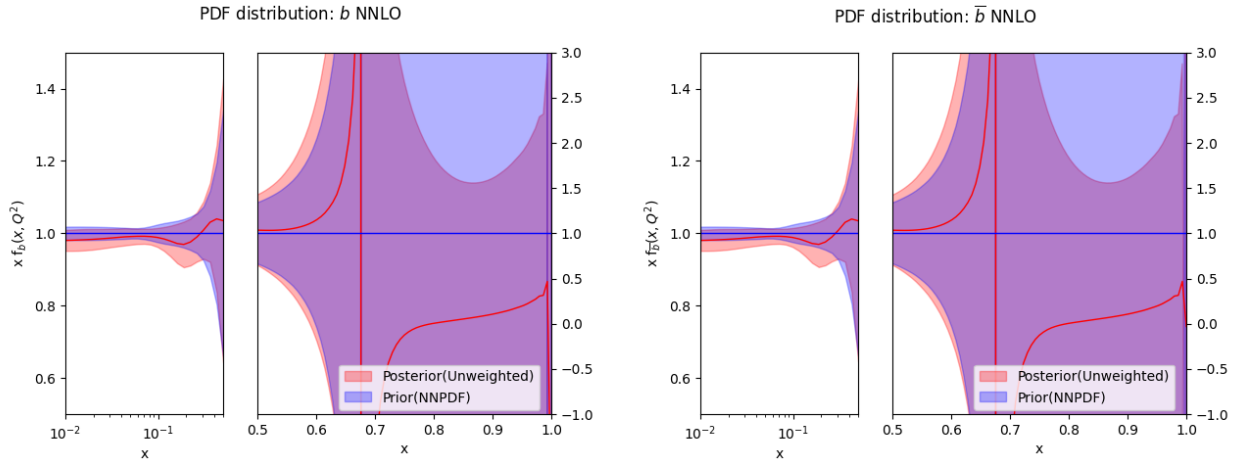

Figure B.13: Quark b and $\bar{b}$ distributions at final iteration with the Kaons + Pions data. In blue we have the unweighted set obtained at the first iteration(prior) while in red we have the unweighted set built at the end of the second(posterior). Central values and uncertainties are normalized over the prior set.

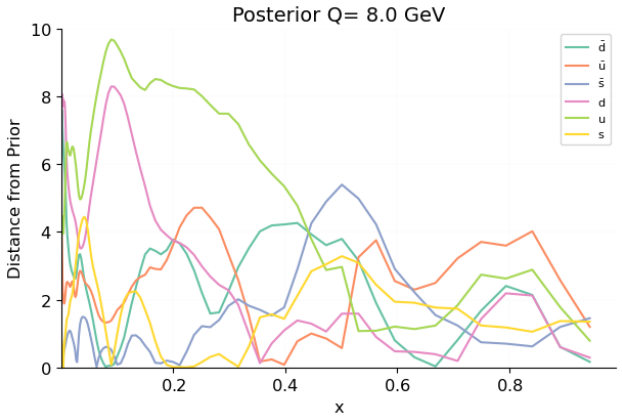

Figure B.14: Distances between prior and posterior for u,d,s, $\overline{u}$, $\overline{s}$, $\overline{s}$ quarks in linear scale.

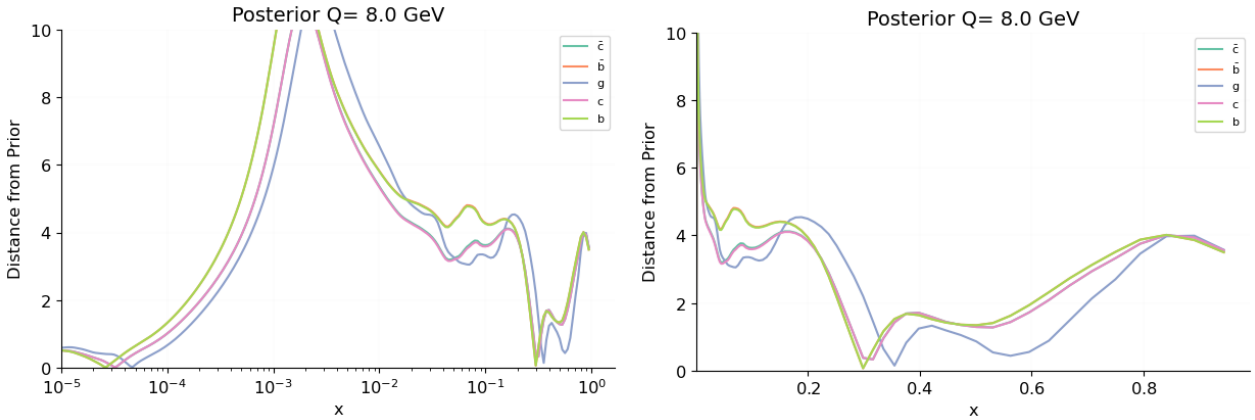

Figure B.15: Distances between prior and posterior for c,b, $\overline{c}$, $\overline{b}$ quarks and gluon at linear(left) and logarithmic(right) scale.

Appendix C

Bibliography

[1] JAM collaboration, N. Sato, C. Andres, J. J. Ethier and W. Melnitchouk, *Strange quark suppression from a simultaneous Monte Carlo analysis of parton distributions and fragmentation functions*, *Phys. Rev. D* **101** (2020) 074020, [1905.03788].

[2] Jefferson Lab Angular Momentum (JAM) collaboration, E. Moffat, W. Melnitchouk, T. C. Rogers and N. Sato, *Simultaneous Monte Carlo analysis of parton densities and fragmentation functions*, *Phys. Rev. D* **104** (2021) 016015, [2101.04664].

[3] I. Borsa, R. Sassot and M. Stratmann, *Probing the Sea Quark Content of the Proton with One-Particle-Inclusive Processes*, *Phys. Rev. D* **96** (2017) 094020, [1708.01630].

[4] ZEUS collaboration, A. Kappes, *Structure function results from ZEUS*, *Nucl. Phys. B Proc. Suppl.* **117** (2003) 247, [hep-ex/0210032].

[5] MAP (Multi-dimensional Analyses of Partonic distributions) collaboration, R. A. Khalek, V. Bertone and E. R. Nocera, *Determination of unpolarized pion fragmentation functions using semi-inclusive deep-inelastic-scattering data*, *Phys. Rev. D* **104** (2021) 034007, [2105.08725].

[6] *Towards NNPDF4.0: The Structure of the Proton to One-Percent Accuracy*, 2021.

[7] COMPASS collaboration, A. Moretti, *Azimuthal asymmetries and transverse momentum distributions of charged hadrons in muon-proton deep inelastic scattering*, *Int. J. Mod. Phys. A* **37** (2022) 2240005.

[8] A. Metz and A. Vossen, *Parton Fragmentation Functions*, *Prog. Part. Nucl. Phys.* **91** (2016) 136–202, [1607.02521].

[9] M. Abele, D. de Florian and W. Vogelsang, *Approximate nnlo qcd corrections to semi-inclusive dis*, *Phys. Rev. D* **104** (Nov, 2021) 094046.

[10] S. Goyal, S.-O. Moch, V. Pathak, N. Rana and V. Ravindran, *Next-to-next-to-leading order qcd corrections to semi-inclusive deep-inelastic scattering*, *Phys. Rev. Lett.* **132** (Jun, 2024) 251902.

[11] L. Bonino, T. Gehrmann and G. Stagnitto, *Semi-inclusive deep-inelastic scattering at next-to-next-to-leading order in qcd*, *Phys. Rev. Lett.* **132** (Jun, 2024) 251901.

[12] HERMES collaboration, A. Airapetian et al., *Multiplicities of charged pions and kaons from semi-inclusive deep-inelastic scattering by the proton and the deuteron*, *Phys. Rev. D* **87** (2013) 074029, [1212.5407].

[13] COMPASS collaboration, C. Adolph et al., *Multiplicities of charged kaons from deep-inelastic muon scattering off an isoscalar target*, *Phys. Lett. B* **767** (2017) 133–141, [1608.06760].

[14] J. Gao, X. Shen, H. Xing, Y. Zhao and B. Zhou, *Fragmentation functions of charged hadrons at next-to-next-to-leading order and constraints on proton PDFs*, 2502.17837.

[15] MAP Collaboration, *Montblanc*, 2022.

[16] A. Buckley, J. Ferrando, S. Lloyd, K. Nordström, B. Page, M. Rüfenacht et al., *LHAPDF6: parton density access in the LHC precision era*, *Eur. Phys. J. C* **75** (2015) 132, [1412.7420].

[17] NNPDF collaboration, R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, A. Guffanti et al., *Reweighting NNPDFs: the W lepton asymmetry*, *Nucl. Phys. B* **849** (2011) 112–143, [1012.0836].

[18] R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, A. Guffanti et al., *Reweighting and Unweighting of Parton Distributions and the LHC W lepton asymmetry data*, *Nucl. Phys. B* **855** (2012) 608–638, [1108.1758].