Khoirul Faiq Muzakka

Global Nuclear PDF Analysis with Neutrino DIS and LHC Data

- 2022 -

WESTFÄLISCHE WILHELMS-UNIVERSITÄT MÜNSTER

INSTITUT FÜR THEORETISCHE PHYSIK

Global Nuclear PDF Analysis with Neutrino DIS and LHC Data

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften im Fachbereich Physik
der Mathematisch-Naturwissenschaftlichen Fakultät
der Westfälischen Wilhelms-Universität Münster

vorgelegt von
Khoirul Faiq Muzakka
aus Rembang

- 2022 -

Dekan :	Prof. Dr. Michael Rohlfing
Erster Gutachter :	Prof. Dr. Michael Klasen
Zweiter Gutachter :	PD Dr. Karol Kovařík
Tag der Disputation :	.....
Tag der Promotion :	.....

Zusammenfassung

Nukleare Parton-Verteilungsfunktionen (nPDFs) sind unverzichtbar, um Vorhersagen für Collider-Experimente mit Atomkernen in den Anfangszuständen unter Verwendung des kollinearen Fakto-risierungs-Ansatzes der QCD zu treffen. nPDFs werden mit einem datengetriebenen Ansatz über eine globale Analyse mit verschiedenen nuklearen Daten bestimmt. Da diese Daten typischerweise nur für bestimmte Kombinationen von PDFs einzelner Parton-Flavours empfindlich sind, werden mehr Daten aus verschiedenen harten Prozessen benötigt, um genauere nPDFs aus einer verbesserten Flavor-Trennung zu erhalten. Die tiefinelastische Neutrino-Nukleus-Streuung (DIS) ist einer jener Prozesse, die komplementäre Informationen über Valenz- und Down-Typ-Quark-PDFs liefern. Der exklusiverer Prozess, die Charm-Dimuon-Produktion, bietet starke Einschränkungen für die Strange Quark-PDFs. Das Einbeziehen dieser Neutrino Daten ist jedoch aufgrund von Spannungen mit einigen geladenen Lepton-DIS-Daten nicht einfach. In dieser Arbeit untersuchen wir die Kompatibilität von Neutrino-DIS-Daten von CCFR, NuTeV, Chorus und CDHSW im Rahmen von nCTEQ PDF Fits. Wir haben mehrere Kompatibilitätskriterien eingeführt, um die Spannungen zu bewerten und kinematische Regionen zu identifizieren, die sie erzeugen. Wir schlagen mehrere Lösungen vor, um die Spannungen abzubauen und einen konsistenten globalen Fit zu erzielen, und vergleichen die resultierenden Theorievorhersagen mit den Charm-Dimuon-Daten von NOMAD und CDHS-Daten.

Im zweiten Teil dieser Arbeit untersuchen wir Targetmassenkorrekturen (TMCs) in Lepton-Nukleus-DIS, die in der hohen x- und niedrigen Q^2-Region signifikant sind. Während eine Reihe von Masterformeln für TMCs verfügbar sind, geht die Herleitung von einem einzelnen Nukleon als Target aus. In dieser Arbeit wird die Gültigkeit der TMC-Masterformel für Lepton-Nukleus-DIS untersucht. Die Auswirkungen von TMCs auf die DIS-Strukturfunktionen werden ebenfalls untersucht. Wir schlagen eine Reihe von Parametrisierungen vor, um die Verhältnisse zu parametrisieren, damit die TMCs schnell berechnet werden können. Die Fähigkeit, TMC-Strukturfunktionen schnell auszuwerten, ist äußerst wichtig für nPDF Fits, bei denen DIS-Theorievorhersagen viele Male während der Suche der Parameter berechnet werden müssen.

Im letzten Teil dieser Arbeit wird die Möglichkeit untersucht, CMS-Dijet-Daten von Proton-Blei-Kollisionen bei $\sqrt{5}$ TeV einzubeziehen. Wir beginnen mit der Analyse der pp-Spektren und zeigen, dass die pp-Daten von allen modernen Protonen-PDFs nicht gut beschrieben werden können. Wir zeigen auch, dass die pPb-Spektren von den Blei-PDFs aus den jüngsten EPPS21- und HIXNEU-CJ2 Fits nicht reproduziert werden können. Der Fit HIXNEU-CJ2 repräsentiert eine globale Analyse mit verbesserter Methodik und fast allen Datensätzen, die in den nCTEQ15HIX- und BaseDimuChorus-Analysen verwendet wurden. Allerdings werden die Dijet-pPb/pp-Verhältnisse mit Rapidität $\eta_{dijet} \leq 2$ bereits gut durch den Fit HIXNEU-CJ2 beschrieben. Wir erweitern dann die HIXNEU-CJ2-Analyse, indem wir die Dijet-pPb/pp-Daten einbeziehen, was zu einer wesentlichen Reduzierung der Gluon-PDF-Unsicherheiten führt.

Abstract

Nuclear parton distribution functions (nPDFs) are indispensable in making predictions for collider experiments with nuclei in the initial states using the QCD collinear factorization framework. nPDFs are determined using a data-driven approach via a global analysis with various nuclear data. As these data typically are sensitive to only certain combinations of PDFs of individual parton flavors, more data from different hard processes are needed to obtain more accurate nPDFs from an improved flavor separation. Neutrino-nucleus deep inelastic scattering (DIS) is one of those processes that provide complementary information on valence and down-type quark PDFs. Its more exclusive process, the charm-dimuon production, places strong constraints on the strange quark PDFs. However, including these neutrino data is not straightforward due to tensions with some charged lepton DIS data. In this thesis, we investigate the compatibility of neutrino DIS data from CCFR, NuTeV, Chorus, and CDHSW in the nCTEQ PDF fitting framework. We introduce several compatibility criteria to assess the level of tensions and identify kinematical regions that generate them. We propose several solutions to relieve the tensions and compare the resulting theory predictions with the charm-dimuon data from NOMAD and nuclear ratio data from CDHS.

In the second part of this work, we study target mass corrections (TMCs) in lepton-nucleus DIS. TMCs are significant in the high x and low Q^2 regions and therefore important if data with low hadronic invariant mass are included. While a set of master formulas for TMCs is available, the derivation did not emphasize the use of a nucleus as the DIS target. In this work, the validity of the TMC master formula for lepton-nucleus DIS is studied. The impact of TMCs on the DIS structure functions are also investigated. We propose a set of parameterizations for the ratios, so that the TMCs can be quickly calculated. The ability to quickly evaluate TMC structure functions is very important for nPDF fitting, where DIS theory predictions need to be calculated many times during the fitting loop.

In the last part of this thesis, the viability of including CMS dijet data from proton-lead collisions at $\sqrt{5}$ TeV is investigated. We start by analyzing the pp spectra, and show that the pp data can not be well-described by all modern proton PDFs. We also show that the pPb spectra can not be reproduced by the lead PDFs from the recent EPPS21 and the HIXNEU-CJ2 fits. The HIXNEU-CJ2 represents a global analysis with almost all the data sets used in the nCTEQ15HIX and BaseDimuChorus analyses, with an improved methodology. We show that the dijet pPb/pp ratio data with $\eta_{dijet} \leq 2$ can, however, already be well-described by the HIXNEU-CJ2. We then extend the HIXNEU-CJ2 fit by including the dijet pPb/pp resulting in substantial reductions of the gluon PDF uncertainties.

List of Publications

The work presented in this thesis has contributed to a number of publications listed below. Additionally, results from this thesis have been presented in talks at various conferenecs. The main authors are underlined.

- **Impact of LHC vector boson production in heavy ion collisions on strange PDFs.**
  A. Kusina, T. Ježo, D.B. Clark, P. Duwentaster, E. Godat, T.J. Hobbs, J. Kent, M. Klasen, K. Kovařík, F. Lyonnet, K.F. Muzakka, F.I. Olness, I. Schienbein, J.Y. Yu
  Eur.Phys.J.C 80 (2020) 10, 968

- **Extending nuclear PDF analyses into the high-x , low-Q2 region.**
  E.P. Segarra, T. Ježo, A. Accardi, P. Duwentaster, O. Hen, T.J. Hobbs, C. Keppel, M. Klasen, K. Kovařík, A. Kusina, J.G. Morfiin, K.F. Muzakka, F.I. Olness, I. Schienbein, J.Y. Yu
  Phys.Rev.D 103 (2021) 11, 114015

- **Impact of inclusive hadron production data on nuclear gluon PDFs.**
  P. Duwentaster, L.A. Husova, T. Ježo, M. Klasen, K. Kovařík, A. Kusina, K.F. Muzakka, F.I. Olness, I. Schienbein, J.Y. Yu
  Phys.Rev.D 104 (2021) 9, 094005

- **Impact of heavy quark and quarkonium data on nuclear gluon PDFs.**
  P. Duwentaster, T. Ježo, M. Klasen, K. Kovařík, A. Kusina, K.F. Muzakka, F.I. Olness, R. Ruiz, I. Schienbein, J.Y. Yu
  Accepted Phys.Rev.D

- **Compatibility of neutrino DIS data and its impact on nuclear parton dis- tribution functions.**
  K.F. Muzakka, P. Duwentaster, T.J. Hobbs, T. Ježo, M. Klasen, K. Kovařík, A. Kusina, J.G. Morfin, F.I. Olness, R. Ruiz, I. Schienbein, J.Y. Yu
  Submitted to Phys.Rev.D

- **Constraining the nuclear gluon PDF with inclusive hadron production data.**
  P. Duwentaster, L.A. Husova, T. Ježo, M. Klasen, K. Kovařík, A. Kusina, K.F. Muzakka, F.I. Olness, I. Schienbein, J.Y. Yu
  Published in SciPost Physics Proceedings DIS2021.

- **Impact of W and Z Production Data and Compatibility of Neutrino DIS Data in Nuclear Parton Distribution Functions.**
  K.F. Muzakka, P. Duwentöster, T.J. Hobbs, T. Ježo, M. Klasen, K. Kovařík, A. Kusina, J.G. Morfiin, F.I. Olness, R. Ruiz, I. Schienbein, J.Y. Yu
  Published in SciPost Physics Proceedings DIS2021

Contents

Chapter 1

Introduction

Humans have been pondering and gazing away to the farthest edge of the cosmos using sophisticated telescopic apparatus and deep down to the most elementary constituents of matters through particle colliders. The more we see physics occurring in the domains beyond what can be seen by the naked eye, the more mysterious things become. Newton's laws and Maxwell's theory on electromagnetism largely do not contradict human intuition. It was not until the advent of Quantum Mechanics, initially developed by Heisenberg, Schrodinger and many others in the 1920s, which describe physics at ultra-microscopic scale, and special relativity theory, mainly developed by Einstein, which describes physics at high energy, that mind-bending laws of nature started to become widely known. The development of relativistic quantum mechanic, also initiated in the 1920s, as a unified framework to merge quantum mechanics and special relativity, was a natural continuation of both theories that applies at short distance and high energy. To explain multi-particle productions and annihilations, field theory concepts were introduced, leading to the development of quantum electrodynamics, which was later unified with weak interaction theory by Glashow[1] and later perfected by Weinberg and Salam[2, 3] into a modern form of electroweak theory by incorporating the Higgs mechanism. This theory is now part of the standard model of particle physics (the SM for short), which is often regarded as one of the pinnacles of human knowledge in understanding the laws of nature at ultra-microscopic scales. The SM describes electromagnetic, weak, and strong interactions in a unified way using quantum field theory as the main framework, imbued with some symmetry principles, such as Lorentz invariance and local gauge symmetry of $SU(3)_c \otimes SU(2)_L \otimes U(1)_Y$.

Being the focus of this thesis, QCD is based on an $SU(3)$ gauge group, with the gluons as the gauge bosons that mediate the interaction. The strong interaction is only active for colored particles, such as quarks. Together with the gluons, quarks constitute hadrons. QCD is very interesting for a variety of reasons. It beautifully explains hadron spectrosocopy from early bubble chamber experiments in the 1950s[4, 5], the theory predicts asymptotic freedom at high energy and therefore explains confinement nature of hadrons, and it implies factorization of long- and short-distance physics for some processes[6], such as deep inelastic scattering (DIS) and Drell-Yan lepton pair production (DY). The factorization of long and short distance physics, implies the existence of a unique set of the so-called parton distribution functions (PDFs), which describe the structure of hadrons.

In lepton-hadron and hadron-hadron colliders, PDFs are very important as inputs to make

theory predictions for the data (typically cross sections) that the experimentalists measure. they can not be predicted from first principle (QCD) using perturbation theory due to a large QCD coupling constant at low energy. However, PDFs are universal, namely the same PDFs are entered in the factorization-based theory predictions for various scattering processes. Therefore, once PDFs are determined from one process, they can be used to calculate theory prediction for other processes. This opens up the possibility of using data-driven methods to determine PDFs. In fact, modern PDFs[7–10] are determined in this way, by using various scattering data from many experiments conducted at the Large Hadron Collider (LHC) at CERN in Geneva, HERA particle accelerator at DESY in Hamburg, Tevatron collider at Fermilab in Batavia, Illinois. Many of these experiments collide protons and therefore, PDFs of the proton were historically the first to be determined.

In 1983, results from muon-iron and muon-deuterium DIS experiments conducted by the European Muon Collaboration (EMC)[11] suggested that the cross sections of DIS of a nucleus A is different from that of a free nucleon. This is unexpected, as the binding energy of protons and neutrons inside the nucleus is very small compared to the exchange energy of the DIS process. These findings suggest that non-trivial nuclear effects contribute to the DIS process, although there is still no consensus about definitive explanations for this EMC effect. This experiment later inspired a series of follow up experiments by different groups, such as the one from New Muon Collaboration (NMC)[12, 13], SLAC (Stanford Linear Accelerator)[14–16], BCDMS[17, 18]and recently the CLAS collaboration in Jefferson Lab (JLab)[19, 20]. These experiments essentially confirm evidence of nuclear modifications for a wide range of kinematical regions. The results from all these experiments can also be regarded as evidence of nuclear modification of free proton PDFs, which leads to the concept of nuclear parton distribution functions.

Nuclear Parton Distribution Functions (nPDFs) are PDFs of nuclei, and therefore, generalizations of the proton PDFs. Similar to the free proton PDFs, they are determined by a global analysis using various nuclear data. nPDFs are currently less precise than the proton ones, mainly due to the lack of (precise) data. All modern proton PDFs[7, 21] are from global analysis using next-to-next- to-leading order (NNLO) of perturbative QCD, while in the nuclear case, the development of NNLO nPDF analyses only started several years ago[22, 23], while widely used nPDFs in the market[24–26] are still at next-to-leading order (NLO) accuracy. This is not a big issue, though, as most nuclear data are old and not precise enough to require going to the next perturbative orders of pQCD to describe them.

In this work, we are interested in better determining nPDFs by including new data, improved theory predictions for data in extreme kinematical regions, and an improved fitting methodology. The new data that we are trying to include is the neutrino-nucleus DIS data from NuTeV, CCFR, CDHSW and Chorus experiments. Compared to the charged lepton DIS data previously used in the past nCTEQ analyses, these new data are six times more abundant. However, we will see that some of these data sets have irreconcilable tensions with the charged lepton DIS data, making the analysis not as straightforward as one initially expected.

The inclusion of DIS data with high Bjorken x and low virtuality Q^2 demands a proper treatment of the so-called target mass corrections (TMCs). A set of TMC master formula based on operator product expansion (OPE) for lepton-nucleon DIS has been known for quite some time, but the validity of the same master formula for lepton-nucleus case has not been properly addressed. In this thesis, we will show that the master formula is still valid, thanks to the rescaling formalism employed in the nPDF framework. We will also show that the size of TMCs seems to be universal for all nuclei, allowing us to parameterize the corrections. The parameterizations are useful in an nPDF fitting, where such slow TMC calculation need to be done many times.

In the last part of this thesis, we will analyze the CMS data from dijet production process in proton-lead collisions at $\sqrt{s} = 5$ TeV. This data is very important in nPDF determination, as it can provide a strong constraint to the nuclear gluon PDF at low and high x regions. We will see that the pp data can not be well-described by all modern proton PDFs available in the literature. This makes the inclusion of pPb or pPb/pp data problematic.

This thesis is organized as follows. In Chapter II, we give an overview of the standard pQCD and nPDF framework used in this analysis. This includes a summary of basic principles of QCD, parton model and factorization, nuclear modifications of structure functions, and finally, nPDFs and rescaling formalism.

In chapter III, we discussed in detail the statistical aspects of PDF fitting. First, derivations of the maximum likelihood-based loss function will be given. Then, three different methods for error estimation will discussed and compared. In particular, we focus on how model misspecifications (which happen, for example, when there are tensions between data sets, or when the PDF parameterization is insufficiently flexible) impact the estimated PDF uncertainties. We then give a detailed discussion on Bayesian reweighting techniques as a method to assess the impact of new data on the fitted PDFs without doing an actual fit. Finally, we discuss several compatibility criteria which is useful to assess and quantify the tensions between data sets.

In chapter IV, we report the detailed results of the global analysis with neutrino DIS data. We start by discussing the neutrino data and the extracted nuclear corrections. We then discuss the baseline fit (nCTEQ15WZSIHdeut) that represents the charged lepton data. A neutrino data alone fit is then discussed, and the resulting nPDFs and predictions are contrasted to the ones from nCTEQ15WZSIHdeut. We then discuss several combined fits and their compatibility assessment. Finally, we compare the predictions from the combined fits to the neutrino-induced dimuon production process with the data from the NOMAD experiment and nuclear ratio data from CDHS.

In chapter V, discussions on the target mass corrections (TMCs) for lepton-nucleus will be given. First, we discuss the sketch of TMC master formula derivation using OPE formalism and then compare the TMC structure functions with that from ACOT. The universality of $F_i^{A,TMC}/F_i^{leading}$ is then shown and discussed. Then, we discuss how we parameterize these ratios in terms of ${}_2F_1$ hypergeometric function and compare the fitted parameterizations to the exact results.

In Chapter VI, we discuss the viability of including the CMS dijet production data. We will start by discussing the pp spectra and how well all the modern proton PDFs can describe the data. We then move to discuss the pPb spectra and the ratio pPb/pp data, which naturally bring to the discussion on a combined fit with the ratio pPb/pp data.

Finally, chapter VII presents the summary and outlook of this work.

Chapter 2

Foundation

2.1 Quantum Chromodynamics (QCD)

This section aims at providing an overview of the theory of strong interaction, otherwise known as QCD. The overview is centered around specific aspects of QCD that are directly related to the main part of this thesis, which is nuclear parton distribution functions.

2.1.1 The Lagrangian

QCD is a quantum field theory (QFT) featuring $SU(3)$ local gauge symmetry which provides an explanation for phenomena related to the strong interaction. Being part of the SM, QCD is different from the electroweak theory in the sense that it predicts asymptotic freedom and confinement property of parton (quarks and gluons) inside hadrons. The word "chromo" in QCD literally means "color", which refers to the charge associated with the strong interaction. The color charge was historically introduced to reconcile the quark model of hadrons with the Pauli exclusion principle[27, 28].

The pure QCD Lagragian takes the following form :

$$\mathcal{L}_{0,QCD} = -\frac{1}{4} G^{a\mu\nu} G^a_{\mu\nu} + \overline{\psi}^f \left(i\not{D} - m^f \right) \psi^f \,. \tag{2.1}$$

This Lagrangian describes the self-interactions between gluon fields A^a_μ, and the interaction between the fermionic field ψ^f with the gluon field A^a_μ. The interaction term between the two resides in the covariance derivative D_μ, defined as :

$$D_\mu = \partial_\mu - i g_s A^a_\mu T^a, \tag{2.2}$$

where g_s is the QCD coupling and T^a, $a = 1, ..., 8$ are the generators of $SU(3)$ in the fundamental representation. The gluonic self interactions come from the kinetic term, which is expressible in terms of the gluon field strength tensor:

$$G^a_{\mu\nu} = \partial_\mu A^a_\mu - \partial_\nu A^a_\mu + g_s f^{abc} A^b_\mu A^c_\nu, \tag{2.3}$$

where f^{abc} is the totally antisymmetric $SU(3)$ structure constant.

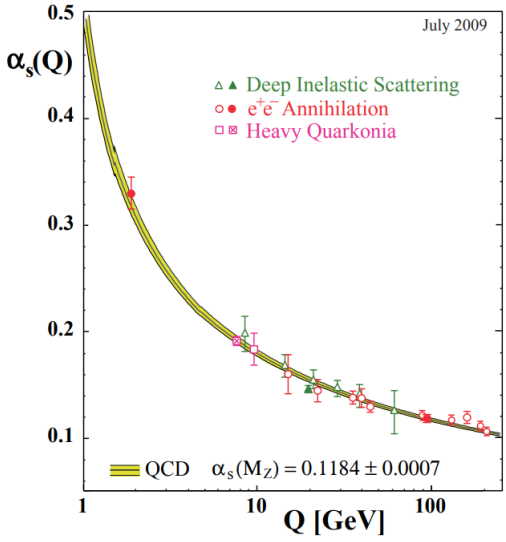


FIGURE 2.1: The running of the strong coupling constant. Figure is taken from [32]

In quantum field theory, to deal with multi-particle productions and annihilations, the classical fields ψ^f and A^a_μ are promoted to quantum fields, which behave as operators in the Fock space. The physical observables, such as cross sections and decay rates, can be computed once the matrix elements are known. The matrix element can be calculated using the LSZ reduction formalism[29]. The calculation requires to know the relevant correlation functions for the process under consideration. Up until this stage, everything is exact. To compute correlation functions, perturbation theory is usually employed, which, through path integral formalism or Wick contractions, leads to a set of Feynman rules.

Perturbation theory assumes that the interaction parts of the Lagrangian can be regarded as small perturbations on top of the free theory. The Feynman diagrams and rules can then be built by first quantizing the free theory. While the quantizing of a free fermionic (Dirac) field is rather straightforward, the quantization of a non-abelian gauge field is not. Due to the gauge freedom, only the physical gauge boson configurations can be taken into account. In the path integral formalism, one can achieve this by including a gauge-fixing term and the so-called ghost term in the Lagrangian. The full derivation can be found in any standard text book of QFT, such as [30, 31]. The full QCD Lagrangian is then given by

$$\mathcal{L}_{QCD} = -\frac{1}{4}G^{a\mu\nu}G^a_{\mu\nu} + \overline{\psi}^f\left(i\not{D} - m^f\right)\psi^f - \frac{1}{2\xi}(\partial_\mu A^{a\mu})^2 + \partial_\mu\bar{c}^a\left(\partial^\mu\delta_{ad} - g_s f^{abd}A^\mu_b\right)c_d. \quad (2.4)$$

Here, c^a and $\bar{c}^a$ are the fermionic scalar ghost and anti-ghost fields. Using this Lagrangian, one can derive Feynman rules for propagators and interaction vertices.

2.1.2 UV Divergences and Asymptotic Freedom

As in quantum electrodynamics (QED), going beyond the tree level in perturbation theory when computing Green functions generally leads to infinities or divergences. A divergence related to virtual particle's momenta going to infinities is called ultraviolet (UV) divergence. The standard way to remove these infinities is through a renormalization program, which basically absorbs the divergences into the so-called bare fields and couplings. A theory is said to be renormalizable if all the UV divergences can be absorbed into a finite set of the so-called 'bare' fields, parameters (such as masses), and couplings. At the end of the program, any observable computed in this framework will be finite and depends only on the renormalized couplings and parameters, which are now functions of the chosen renormalization scale μ. The dependency of the couplings and parameters on μ are governed by renormalization group equations (RGEs). In QCD, the strong coupling g_s, is running according to:

$$\mu \frac{dg_s}{d\mu} = \beta(g_s), \tag{2.5}$$

where $\beta(g_s)$ can be computed order-by-order in perturbation theory. At one loop using the $\overline{MS}$ scheme, it is given by

$$\beta(g_s) = -g_s \frac{g_s^2}{16\pi} \left[\frac{11}{3} C_A - \frac{2}{3} N_f \right], \tag{2.6}$$

where C_A is the Casimir constant for the $SU(3)$ group and N_f is the number of quark flavors. For $N_f \leq 33C_A/6$, which is the case for the SM, we see that β_s is always negative, thus the QCD coupling decreases in strength as the scale increases. By solving the RGE (2.5), one obtains

$$\alpha_s(\mu) = \frac{12\pi}{(11C_A - 2N_f) \ln\left(\mu^2/\Lambda_{QCD}^2\right)}, \tag{2.7}$$

where $\alpha_s = g_s^2/4\pi$ and $\Lambda_{QCD} \sim 200$ MeV is the scale at which g_s diverges, otherwise known as the Landau pole. Eqs. (2.7) implies that QCD becomes perturbative and asymptotically free at high energy. Conversely, at $\mu \sim \Lambda_{QCD}$, the theory ceases to be perturbative, hence the perturbation theory can not be used. One has to resort to a computationally intensive Lattice QCD technique if one wants to obtain meaningful QCD predictions. In Fig. 2.1, we show the value of α_s as a function of the renormalization scale Q. We can clearly see that α_s is a monotonically decreasing function of Q.

2.2 Deep Inelastic Scattering, Parton Model and Factorization

Deep inelastic scattering (DIS) is a process where energetic leptons collide with the hadronic target, typically proton or nucleus, such that the target is destroyed and the detectors measure the products of the collisions. Schematically, the DIS process for $lp^+ \to l'X$ is shown in Fig. 2.2. The term "deep" indicates that the target is completely destroyed. Let $W^2 = p_X^2$, where

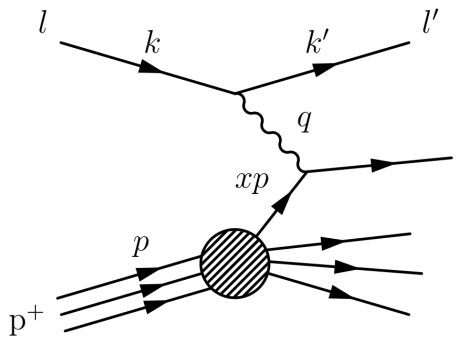


FIGURE 2.2: Feynman diagram for $l + p^+ \to l' + X$ DIS.

p_X is the total momentum of the debris X. Then a scattering process $lp^+ \to l'X$ is categorized as a DIS process only if $W^2 \gg M^2$, where M is the hadron mass. If $W^2 = M^2$, we say that the process is elastic. The term shallow (or semi) inelastic scattering (SIS) is often introduced to refer to a scattering process where the target is excited to a resonance state, which then decayed into debris. This typically happens for W values not far above the proton mass ($W \lesssim 2$ GeV).

Let's discuss the typical kinematical variables used to describe DIS process. For massless leptons, it is easy to show that $q^2 = (k - k')^2 = -EE'(1 - \cos(\theta) \leq 0$, where θ is the angle between the outgoing and incoming lepton. Then we define $Q^2 = -q^2 \geq 0$. Note that $Q^2 = 0$ only if $\theta = 0$. Then we define

$$\nu \equiv \frac{p \cdot q}{M} = (E - E')_{\text{lab}}, \qquad x \equiv \frac{Q^2}{2p \cdot q} = \frac{Q^2}{2M\nu}, \qquad y = \frac{p \cdot q}{p \cdot k} = \frac{E - E'}{E}\Big|_{\text{lab}}. \tag{2.8}$$

It is easy to show that the DIS condition $W^2 > M^2$ implies $\nu \geq 0$, $x \in [0, 1]$ and $y \in [0, 1]$. To show this, we can start by writing W^2 in terms of x and Q^2:

$$W^2 = p_X^2 = (p + q)^2 = M^2 + Q^2 \left(\frac{1}{x} - 1\right). \tag{2.9}$$

The condition $W^2 \geq M^2$ implies :

$$\frac{1}{x} - 1 \geq 0\,. \tag{2.10}$$

From this inequality, one can see that x can not be negative, otherwise, this inequality can never be true. Thus $x \geq 0$. As x is always positive, this inequality (2.10) implies $x \leq 1$. Thus $x \in [0, 1]$ has been shown. As $Q^2 \geq 0$, from the definition of x, the inequality $x \geq 0$ implies that ν can not be negative, or $\nu \geq 0$. This further implies $y \geq 0$. As the energy of the outgoing lepton is always positive, $E' \geq 0$, this gives $y \leq 1$, demonstrating that $y \in [0, 1]$. In the literature, x is often called the Bjorken x, while Q^2 is often referred to as the virtuality.

The cross section for the DIS process can be written in terms of the leptonic $L^{\mu\nu}$ and $W^{\mu\nu}$ as

$$d\sigma = \frac{1}{F}\frac{e^4}{q^4} L^{\mu\nu} W_{\mu\nu} 4\pi \, \frac{d^3k'}{(2\pi)^3 2E'}\,, \tag{2.11}$$

where $F = 4\sqrt{(p.k) - M^2 m_l^2} = 2S$ is the Moller flux factor. The leptonic and dimensionless hadronic tensors are defined as:

$$L_{\mu\nu}(k,k')\Big|_{\text{QED}} = \frac{1}{2}\sum_{sl,sl'} \overline{u}(l')\gamma_\mu u(l)\overline{u}(l)\gamma_\nu u(l') = 2\Big\{k_\mu k'_\nu + k_\nu k'_\mu - (k\cdot k')g_{\mu\nu}\Big\}\,, \tag{2.12}$$

$$4\pi W_{\mu\nu}(p,q) = \sum_X d\Pi_X (2\pi)^4\,\delta^{(4)}(p+q-p_X)\,\Big\langle \langle N(p)|J^\dagger_\nu(0)|X\rangle\langle X|J_\mu(0)|N(p)\rangle\Big\rangle_{\text{spin}}\,. \tag{2.13}$$

Here, $d\Pi_X = \sum_{i\in X}\frac{d^3p_i}{(2\pi)^3 2p_i^0}$ denotes the invariant phase space measure for the hadronic final state X and $|N(p)\rangle$ denotes the nucleon state. From the Lorentz structure of $W_{\mu\nu}$, one can express $W_{\mu\nu}$ in terms of $W_i(x,Q^2)$ structure function as :

$$W_{\mu\nu}(p,q) = -g_{\mu\nu}W_1 + \frac{p_\mu p_\nu}{M^2}W_2 - i\epsilon_{\mu\nu\alpha\beta}\frac{p^\alpha q^\beta}{M^2}W_3 + \frac{q_\mu q_\nu}{M^2}W_4 + \frac{(p_\mu q_\nu \pm p_\nu q_\mu)}{M^2}W_{5,6}\,. \tag{2.14}$$

Here, the structure functions $W_i(x,Q^2)$ are real. These W_is are related to the measurable F_i structure function as

$$\Big\{F_1,\ F_2,\ F_3,\ F_4,\ F_{5\ (6)}\Big\} = \Big\{W_1,\ \frac{Q^2}{2xM^2}W_2,\ \frac{Q^2}{xM^2}W_3,\ \frac{Q^2}{2M^2}W_4,\ \frac{Q^2}{2xM^2}W_{5\ (6)}\Big\}\,. \tag{2.15}$$

When contracting $W_{\mu\nu}$ with the $L_{\mu\nu}$, we see that not all the structure functions W_is contribute to the total cross section. For example, as $L_{\mu\nu}$ is symmetric, then the contributions from antisymmetric terms such as the one from W_3 and W_6 are zero. As $q^\mu q^\nu L_{\mu\nu} \propto m_l^2 = 0$, the contribution from W_4 also vanishes. Similarly also for W_5, as $p^\mu q^\nu L_{\mu\nu} \propto m_l^2 = 0$. After evaluating the contraction, we obtain

$$\frac{d\sigma}{d\Omega dE'} = \frac{\alpha_e^2}{4ME^2\sin^4\frac{\theta}{2}}\left[2W_2(x,Q)\cos^2\frac{\theta}{2} + W_1(x,Q)\sin^2\frac{\theta}{2}\right]\,. \tag{2.16}$$

We have managed to derive a cross section formula for a DIS process in terms of W_i structure functions. The calculation for W_i can be performed using parton model approach[33]. Alternatively, a more rigorous operator product expansion (OPE)[34] technique can also be employed.

Parton model was first introduced by Feynman[35], Bjorken and Paschos[33] in 1969 as an explanation for the Bjorken scaling behavior of the DIS structure function. In the parton model, the incoming lepton is assumed to scatter off point-like constituents (called partons, which we identify as quarks and gluon) of the hadron, whose probability distributions are given by the so-called parton distribution functions (PDFs). At high energy, these partons are effectively massless and assumed to carry a fraction x of the hadron momentum p. Then the cross section of DIS process can be expressed as a convolution of hard cross section with parton distribution functions (PDFs) :

$$\sigma(e^- p^+ \to e^- X) = \sum_i \int d\xi f_i(\xi)\hat{\sigma}(e^- q_i \to e^- X)\,. \tag{2.17}$$

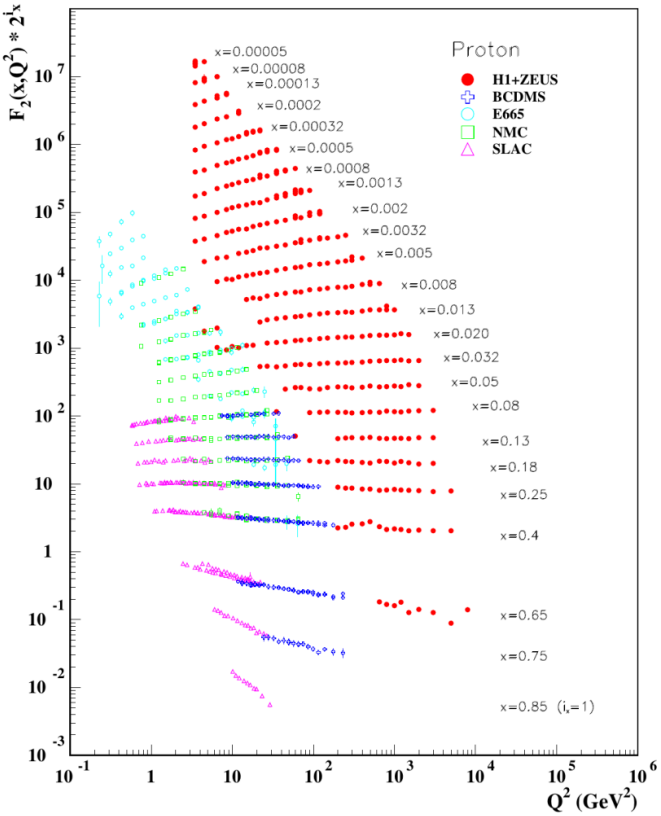


FIGURE 2.3: Proton structure functions measured by several experiments. The figure is taken from [30].

The hard cross section for $e^- q_i \to e^- q_i$ can be computed easily using Feynman diagram. It is given by

$$\frac{d\hat{\sigma}}{d\Omega dE'} = \frac{\alpha_e^2 Q_i^2}{4E^2 \sin^4 \frac{\theta}{2}} \left[\cos^2 \frac{\theta}{2} + \frac{Q^2}{2m_q^2} \sin^2 \frac{\theta}{2} \right] \delta \left(E - E' - \frac{Q^2}{2m_q} \right) . \tag{2.18}$$

Here, Q_i, m_q are the fractional charge and mass of the participating quark. Inserting this hard cross section to the factorization formula (2.17) and comparing to (2.16), we obtain

$$F_1(x, Q^2) = \frac{1}{2} \sum_i Q_i^2 f_i(x) , \tag{2.19}$$

$$F_2(x, Q^2) = 2xF_1 = \sum_i Q_i^2 x\, f_i(x) . \tag{2.20}$$

Thus, the parton model predicts that both structure functions are independent of Q^2. This is the scaling behavior mentioned before. The relation $F_2 = 2xF_1$ is often referred to as the Callan-Gross relation.

In reality, Bjorken scaling is violated in the small and high x regions. In Fig. 2.3, we show proton structure function F_2 for wide range of x and Q^2 region. We can clearly see that F_2 is not independent of Q^2. To better describe the data in these regions, it is necessary to include contributions from QCD radiations. The improvement leads to a QCD-improved parton model,

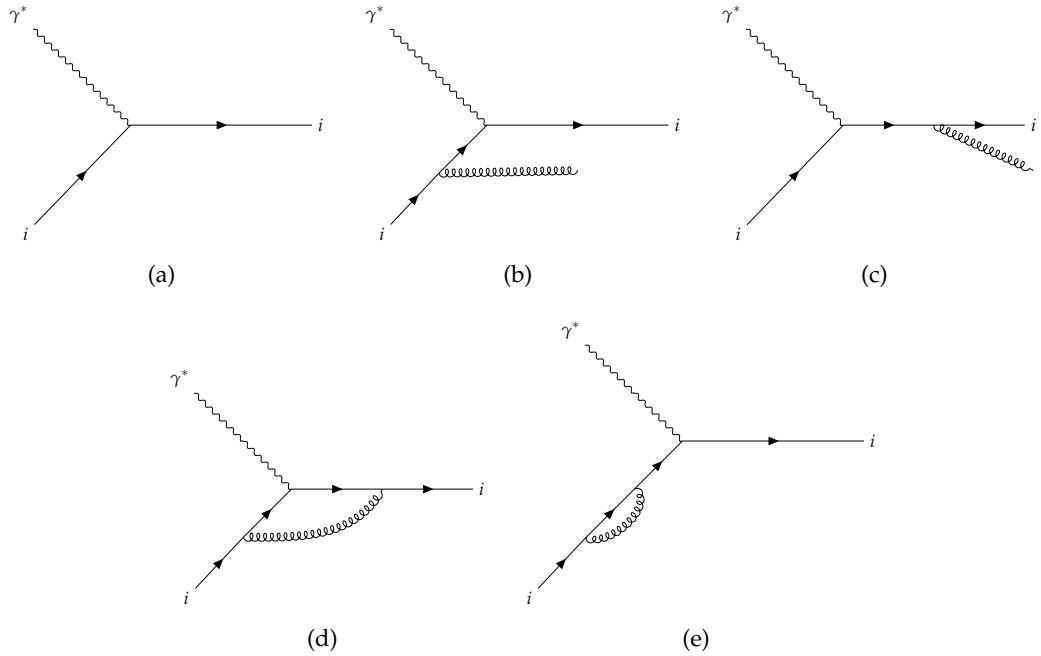


FIGURE 2.4: Feynman diagrams for the hard processes in DIS. (a) Born (tree level) diagram for γ^* + quark of type i scattering. (b) Diagram for a gluon radiation from the initial quark i (initial state radiation). (c) Diagram for a gluon radiation from the final state quark i (final state radiation). (d) Diagram for one loop QCD correction to the $\gamma q_i q_i$ vertex. (e) Same as (a), but now QCD self energy correction to the initial state quark i is included.

as discussed in the following.

At next-to-leading order (NLO), or $\mathcal{O}(\alpha_s)$, QCD corrections enter in: 1) initial state radiation (shown in Fig. 2.4(b)), 2) virtual correction (shown in Fig. 2.4(d)) and 2.4(e)), and 3) final state radiation(shown in Fig. 2.4(c)). As the partons are massless, the contributions from these processes to the total partonic cross section contain soft and collinear divergences. Fortunately, most of the divergences cancel, and the remaining collinear divergence, which is Q^2-dependent, is absorbed into PDFs. The PDFs then contain some Q^2 dependence, whose behavior is governed by the DGLAP evolution equations[36–38]. As the PDFs are Q^2 dependent, the calculated structure functions, therefore, have some Q^2 dependency as well.

The derivation of QCD corrections from these diagrams can be found in various textbooks on QCD, such as [30, 31]. However, here, we would like to stress the importance of choosing a gauge for the gluon field. If one uses Feynman gauge, such as in [30], all the diagrams : initial, final and virtual radiations contribute to the hard process. As the remaining infrared (IR) divergences from these diagrams will eventually be absorbed into PDFs, the interpretation of PDF being a probability distribution for finding *initial* state parton becomes less obvious, as it receive contributions from the *final* state radiation. In this respect, Feynman gauge is not the most convenient choice of gauge, as it leads to an obscure interpretation of PDFs. Using the axial gauge, however, as shown in [39], leads to a clearer interpretation of PDFs as initial state distributions, as all contributions from the final state radiations vanish.

In the following, we outline the derivation of DGLAP equation. Readers can refer to [39] for pedagogical details.

- *Phase Space Factorization*: Using Sudakov decomposition for collinear branching of initial state parton, it can be shown that the $n+1$ final state particle phase space measure factorizes as

$$d\Phi^{(n+1)}(k_1,k) = d\Phi^{(n)}(zk_1,k)\,\frac{d|\vec{k}_T|^2}{|\vec{k}_T|^{2\epsilon}}\frac{1}{16\pi^2}\frac{(4\pi)^\epsilon}{\Gamma(1-\epsilon)}\frac{dz}{1-z}\,, \tag{2.21}$$

  where k_1 is the momentum of branching parton, $k = (1-z)k_1$ is the momentum of the splitted particle, and $\vec{k}_T$ is its transverse momentum.

- *Cross section factorization* : For initial state quark-quark splitting (namely: $q \to q+g$), the cross section for producing $n+1$ final states factorizes as

$$d\hat{\sigma}^{(n+1)}(k_1,...) \to d\hat{\sigma}^{(n)}(zk_1)\frac{\alpha_s}{2\pi}\hat{P}_{qq}(z)dz\frac{d|\vec{k}_T|^2}{|\vec{k}_T|^2}\,. \tag{2.22}$$

  Here, $\hat{P}_{qq}(z)$ is the unregulated Alterelli-Parisi splitting function for the quark-quark splitting :

$$\hat{P}_{qq}(z) = C_F\frac{1+z^2}{1-z^2}\,, \tag{2.23}$$

  where $C_F = 4/3$ is the color factor. We can see that the $n+1$ cross section become divergent when $|\vec{k}_T| \to 0$ (collinear radiation) and/or $z \to 1$ (soft radiation). Note that besides $q \to qg$ splitting (initial state quark, split into a quark that participates in the hard process and a gluon radiation), there are also $q \to gq$ (now the gluon participates in the hard process), $g \to qq$ and $g \to gg$ splittings which contribute to the hard cross sections.

- One can work out the $|\vec{k}_T|$ integral directly, with the lower and upper bound given by m^2 and Q^2. The lower bound m^2 is the quark mass which serve as an IR regulator. The $|\vec{k}_T|$ integral gives $\ln(Q^2/m^2)$. Upon adding the contribution from the real emission from $q \to qg$ splitting, we see that the cross section calculated using parton model maintains the same form (2.17), but now the PDFs are modified as

$$f_q(x) \to f_q(x,Q^2) \equiv \left[1+\frac{\alpha_s}{2\pi}\ln(Q^2/m^2)\hat{P}_{qq}\right]\otimes f_q\,. \tag{2.24}$$

  Here, the convolution symbol $\otimes$ means

$$\hat{P}_{qq}\otimes f_q \equiv \int_x^1\frac{dz}{z}\hat{P}_{qq}(z)f_q\left(\frac{x}{z}\right) = f_q\otimes\hat{P}_{qq}\,, \tag{2.25}$$

$$1\otimes f_q \equiv \int_x^1\frac{dz}{z}\delta(1-z)f_q\left(\frac{x}{z}\right) = f_q(x)\,. \tag{2.26}$$

- The redefinition in (2.24) however is not infrared safe: one needs to know the detail of the IR regulator m^2 to make predictions. The solution of this problem is to resum a whole tower of diagrams, where multiple gluons are emitted. Resumming all such diagrams is equivalent to redefining PDFs as :

$$f_q(x) \to f_q(x, Q^2) \equiv \exp\left[\frac{\alpha_s}{2\pi} \ln(Q^2/m^2) P_{qq}\right] \otimes f_q \,. \tag{2.27}$$

  Thus, by taking the derivative with respect to Q^2, the PDFs now satisfy:

$$Q^2 \frac{\partial}{\partial Q^2} f_q(x, Q^2) = \frac{\alpha_s}{2\pi} \hat{P}_{qq} \otimes f_q(x, Q^2) \,. \tag{2.28}$$

  However, we still have some problems : 1) the splitting function is divergent for $z \to 1$. 2) The strong coupling α_s is set to be a (fixed) constant and not yet renormalized.

- The first problem can be cured by considering gluon radiations that get reabsorbed by the radiating initial state quark (see Fig. 2.4(e)). This $\mathcal{O}(\alpha_s)$ self-energy diagram modifies the wave function renormalization of the quark, which soften the IR divergence in the splitting function. After taking into account the self-energy contribution, the splitting function is now regulated as

$$\hat{P}(z) \to P(z) \equiv C_F \left(\frac{1+z^2}{1-z}\right)_+ \,. \tag{2.29}$$

  The plus distribution f_+ is defined such that for a test function $h(x)$:

$$\int dz f(z)_+ h(z) = \int f(z)(h(z) - h(1)) \,. \tag{2.30}$$

- Finally, taking into account the virtual correction in the γqq vertex (see Fig. 2.4(d)) amounts to replace α_s with the renormalized one. By including contributions from all other splittings, the PDFs can be shown to satisfy :

$$Q^2 \frac{\partial f_i}{\partial Q^2} = \frac{\alpha_s(Q^2)}{2\pi} \sum_j \hat{P}_{ij} \otimes f_j(x, Q^2) \,. \tag{2.31}$$

  This set of equations is often called as the *Dokshitzer-Gribov-Lipatov-Alterelli-Parisi* (DGLAP) evolution equations [36–38].

We have seen that when including NLO QCD corrections in our calculation, we can maintain the main statement of parton model by promoting the PDFs to be Q^2 dependence. In turns out this is also true if one goes beyond NLO. Specifically, by writing the hard cross section as

$$d\hat{\sigma} = \sum_i \left(\frac{\alpha_s}{2\pi}\right)^i d\hat{\sigma}^{(i)} \,, \tag{2.32}$$

one can compute the hadron-level cross section using the factorization formula

$$d\sigma(Q^2) = \sum_i \int d\xi d\hat{\sigma}_i)(Q^2) f_i(\xi, Q^2) \equiv \sum_i d\hat{\sigma}_i(Q^2) \otimes f_i(Q^2)\,, \tag{2.33}$$

with the PDFs $f_i(x, Q^2)$ satisfy the DGLAP equations (2.31). We note here that we can also define PDFs in other way than (2.27), giving different *factorization scheme*. Given a set of PDFs $f_i(x, Q^2)$, one can define another set $f'_i(x, Q^2)$ as

$$F'_i(x, Q^2) = \sum_j C_{ij} \otimes f_j(Q^2)\,, \tag{2.34}$$

$$C_{ij}(z) = 1 + \frac{\alpha_2}{2\pi} C_{ij}^{(1)} + \left(\frac{\alpha_2}{2\pi}\right) C_{ij}^{(2)} + \ldots\,. \tag{2.35}$$

The hadron-level cross section should not depends on the choice of scheme, thus :

$$\sigma = \sum_i \sigma_i \otimes f_i = \sum_i \sigma'_i \otimes f'_i\,, \tag{2.36}$$

with $\sigma'_i = \sum_i \sigma_j \otimes C^{-1}_{ji}$. To retain the same form of evolution equation, the splitting functions must also be modified as[39]

$$P'_{ij} = \left[C \otimes P \otimes C^{-1} - 2\pi\beta(\alpha_s)\frac{dC}{d\alpha_s} C^{-1}\right]_{ij}\,. \tag{2.37}$$

Thus, beyond leading order, the hard coefficient is always intertwined with the splitting functions. We note here that while the fixed order prediction for the total cross section is scheme-dependent, the all-order predictions are the same for different schemes.

Besides having freedom to choose factorization scheme, one is also be free to choose the factorization scale while keeping the hadron-level prediction the same:

$$\sigma(Q^2) = \sum_i d\hat{\sigma}_i(Q^2) \otimes f_i(Q^2) = \sum_i d\hat{\sigma}_i(Q^2, Q_F^2) \otimes f_i(Q_F^2)\,. \tag{2.38}$$

The factorization scale Q_f, is typically chosen $Q_F = cQ$, with c is between 0.5 and 2. As PDFs at different scales are related via the DGLAP equation, they satisfy

$$f_i(x, Q^2) = \sum_j D_{ij}(Q^2/Q_F^2) \otimes f_j(Q_F^2) \tag{2.39}$$

for some D_{ij}. Therefore, the hard cross section at different factorization scales are related by

$$\hat{\sigma}_i(x, Q^2, Q_F^2) = \sum_j \hat{\sigma}_j \otimes D_{ji}(Q^2/Q_F^2)\,. \tag{2.40}$$

The all-order predictions should be independent of the choice of Q_F. However, in practice, $\hat{\sigma}_i$ is calculated at a fixed order. Therefore some differences are expected when comparing

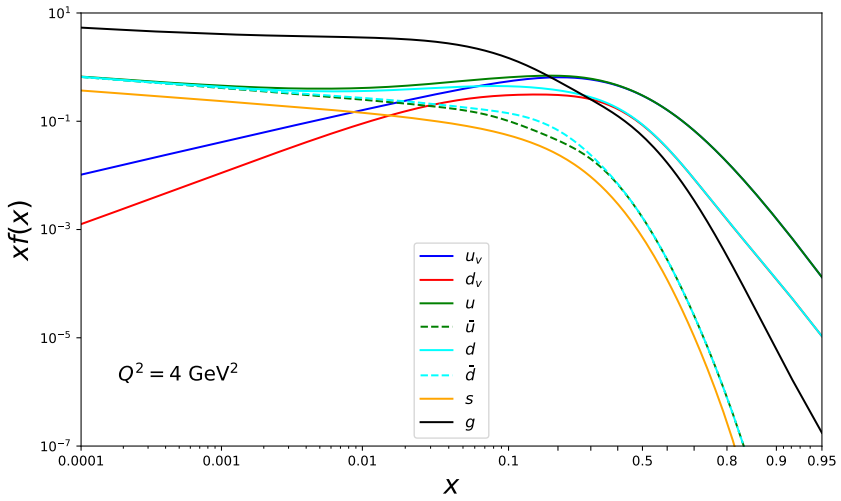


FIGURE 2.5: CJ15 NLO[10] PDFs at $Q = 2$ GeV.

predictions from different factorization scales.

We have shown that for a DIS process, the total cross section can be written as convolutions of the cross sections from hard process with scale-dependent parton distribution functions (PDFs). It turns out that factorization of short- and long- distance physics is not unique in DIS. There are other processes where the factorization also applies, such as Drell-Yan (DY) $A + B \to l + l' + X$ and single inclusive hadron production (SIH) $A + B \to C + X$ cross sections. For SIH process :

$$\sigma(A + B \to C + X) = \sum_{i,j,k} \int_0^1 dx_A dx_B dz_C \, f_i^A(x_A, \mu_F) f_j(x_B, \mu_F) \times \hat{\sigma}_{ij \to cX}(x_A p_A, x_B p_B, \frac{p_C}{z}, \mu_F) \, D_{C/c}(z_C, \mu_F) \, . \tag{2.41}$$

Here, $D_{C/c}$ is the fragmentation function for a parton c fragments into a hadron C. Note that the same PDFs enter into the factorization formula for DIS, DY, and SIH. Thus, PDFs are universal. Once they are determined at some scale Q_0 from one process, it can be used to predict other processes.

2.3 Parton Distribution Functions

As discussed in the previous section, factorization theorem predicts the existence of universal (process-independent) parton distribution functions (PDFs). To understand how the PDFs look like, in Fig. 2.5, we show proton PDFs obtained from the CJ15 analysis[10]. From this figure, we can learn several important points:

- At low x ($x \lesssim 0.1$), gluon PDF is much higher than all quark PDFs. This explain why Bjorken scaling is maximally violated in this region.

- At low x ($x \lesssim 0.1$), all the sea quark ($u, d, s, \bar{u}, \bar{d}$) PDFs are higher than the valence quark ($u_v = u - \bar{u}$ and $d_v = d - \bar{d}$) PDFs. This is understandable, as these quarks are generated from gluon radiation, which are abundant in this region.
- At higher x ($x \gtrsim 0.1$), valence quarks dominate, with $u_v > d_v$. Furthermore, $u \gtrsim u_v$ and $d \gtrsim d_v$ as $\bar{u}$ and $\bar{d}$ are orders of magnitudes smaller than u and d,
- At the highest x, u and u_v are orders of magnitudes higher than the rest: $u \gtrsim u_v \gg d \gtrsim d_v \gg g \gg f_{else}$.
- In all x regions, $d > \bar{d}$, $u > \bar{u}$ and $s < f_{else}$.

These points are useful when assessing PDF sensitivities of an observable. For example, the charged lepton DIS structure function at leading order can be written as

$$F_2^{NC}(x) = \frac{4}{9}\left(u + \bar{u} + c + \bar{c}\right) + \frac{1}{9}\left(d + \bar{d} + s + \bar{s}\right) \tag{2.42}$$

This means that this observable is, for instance, 4 times more sensitive to $\bar{u}$ than d, however, we see that d is orders of magnitude bigger than $\bar{u}$ at large x. Therefore the contribution of $\bar{u}$ to the process at high x is meaningless compared to that of d, despite having lesser sensitivity.

PDF represent a probability distribution of finding certain parton with momentum fraction within ξ and $\xi + d\xi$, therefore, the momentum fractions carried by the partons must sum up to unity. Namely, the total momenta carried by the partons are equal to the momentum of the nucleon. Thus:

$$\sum_i \int d\xi\, \xi f_i(\xi) = 1\,. \tag{2.43}$$

In the modern view, partons here refer to quarks and gluon. Given that proton consist of two up and one down quarks in total, we also have the following valence sum rules

$$\int d\xi\, (u - \bar{u})(\xi) d\xi = 2\,, \tag{2.44}$$

$$\int d\xi\, (d - \bar{d})(\xi) d\xi = 1\,. \tag{2.45}$$

2.4 Global QCD Analysis

Due to the non-perturbative nature of QCD at low energy, PDFs can only be reliably determined by fitting the functions to various high energy scattering data. There are some efforts (see, for example, [40] and references therein) to calculate PDFs directly from first principle using Lattice QCD method, however, so far, the uncertainties of the resulting PDFs are still substantial. Therefore, PDFs from the data-driven approach are still widely used for comparing factorization-based theory prediction to the experimental data.

There are several research groups that study PDFs as the main goal. For proton PDFs, the main players are : NNPDF[9], CT[7], MSHT[8], and CJ[10]. While these groups use overlapping

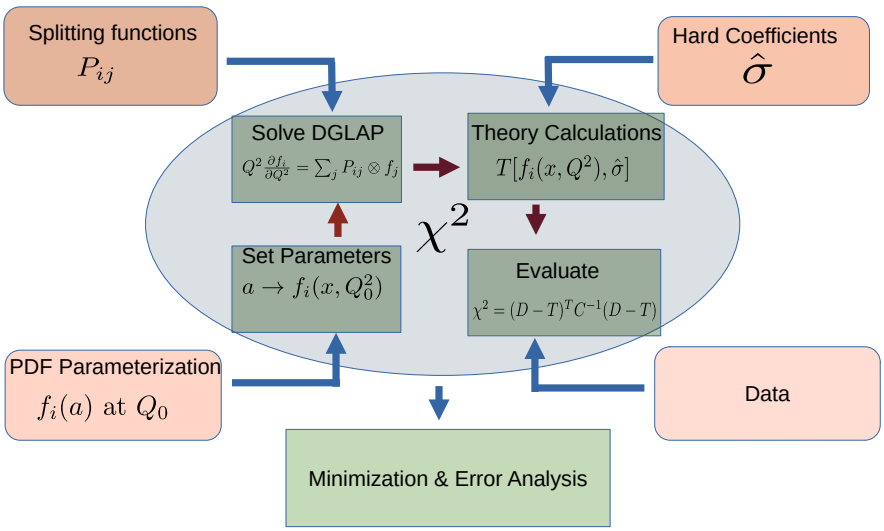


FIGURE 2.6: The generic flowchart of a global QCD analysis.

data sets to constrain the PDFs, they differ in how they parameterize the PDFs at the input scale, fitting program/code, different included data sets and kinematical cuts, and theory corrections, such as treatment of higher twist, deuteron correction, and nuclear corrections. They also use different methods to estimate the uncertainties of the fitted PDFs. To illustrate the difference, let's take for example NNPDF4.0[9] and CT18[7] analyses. NNPDF4.0, used neural networks (NNs) to parameterize the PDFs at the input scale. In contrast, CT18 PDFs, used Bernstein polynomials. To estimate the uncertainty of the PDFs, NNPDF4.0 analysis used a monte-carlo replica method, while CT18 used the Hessian method. Given the differences in the fitting methodology, it is no surprise that the PDFs from these groups are not identical. However, the fact that the central values are within the respective error bars means that a PDF set determined by one group is reproducible by the other groups, albeit using a different methodology.

The determination of PDFs from the data is an inverse problem. The term "Global QCD analysis" is often used to indicate that PDFs are fitted from various data sets from different processes. The analysis is started with data selection. This includes choosing which data sets to include and the kinematical cuts to impose. If there are tensions between data sets, one should also look further into the data sets and investigate possible explanations. Once the data sets have been selected, one can proceed to run the fitting procedure, summarized in Fig. 2.6. The central part of the fitting is loss function χ^2 minimization. To calculate the χ^2 function, theory predictions need to be computed, which typically depend on the PDFs at the scale Q of the process. The values of the PDFs at the scale Q can be computed by solving the DGLAP evolution equations, given the parameterized PDFs at the input scale Q_0 as the initial

condition. The whole χ^2 function is then effectively a function of the PDF parameters. Once the minimization is finished, the uncertainty of the fitted PDFs can then be determined using Hessian method or monte-carlo method as discussed in Chapter 3.

2.5 Nuclear Corrections

A naive view to describe the dynamics of nucleons inside a nucleus is that the nucleons are non-interacting with each other and thus essentially free. This means the total DIS structure function is an arithmetic sum of the structure functions from the individual nucleons. However, this naive view is challenged by the measurements of the structure function ratio F_2^A / F_2^D, where F_2^A are the average (per nucleon) structure function for a nucleus with the mass number A and F_2^D is the average deuteron structure function. In Fig. 2.7, we display the measurement of $R = F_2^{Fe} / F_2^D$ from several experiments. The figure shows deviation from the unity (as expected from the naive argument) in three characteristic regions :

- Shadowing region, $R \leq 1$, $x \lesssim 0.1$.
- Anti-shadowing region, $R \geq 1$, $0.1 \lesssim x \lesssim 0.3$.
- EMC region, $R \leq 1$, $0.3 \lesssim x \lesssim 0.7$.
- Fermi motion, $R \geq 1$, $x \gtrsim 0.7$.

These regions represent different physics contributing to the nuclear correction. Similar shapes can also be observed for other nuclei, albeit the strength (the deviation from unity) slowly varies with A. In [41], it was shown that by fitting $R[F_2^A] = a + b\ln(A)$ for each x bins of various nuclear ratio data, it can be inferred that : q) the size of shadowing at low x increase with $\ln(A)$. 2) For $0.07 < x < 0.3$, which corresponds to the anti-shadowing region, the size of anti-shadowing is independent or slightly increasing with A. 3) For $0.2 < x < 0.8$, the ratio $R[F_2^A]$ decreases with $\ln(A)$ (or the shadowing strength in this region increases with $\ln(A)$).

For shadowing, the underlying mechanism is coherent multiple scatterings of the gauge boson probe with different nucleons inside the nucleus, leading to a destructive interference and hence a reduced cross section (see for example a review in [42] and references therein). In the coherent limit, the hadronic fluctuation of the off-shell γ^* that interacts with the nucleus has roughly a lifetime $\tau \sim 1/(2M_N x)$, where M_N is the nucleon mass, x is the Bjorken x. Thus, the lifetime increase with decreasing x. To interact with the nucleus at a whole, the lifetime must exceed the radius of the nucleus $\tau > R_A \sim A^{1/3}$, which gives $x \lesssim 0.1A^{-1/3}$. This simple argument demonstrates why the shadowing effect occurs at low x. From various experimental data taken on different nuclei, it can be inferred that : 1) Shadowing increases as x decreases. At the smallest value of x, it is consistent with either saturation or mild decrease (plateau). 2) Shadowing increases as A.

The nuclear effect in the anti-shadowing region is less well understood, although similar constructive interferences from multiple scattering of γ^* are suspected as the underlying

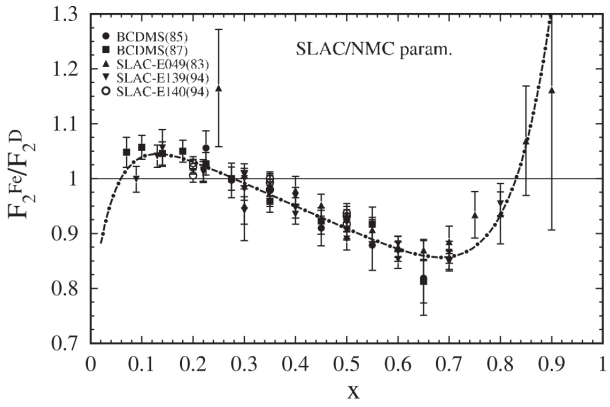


FIGURE 2.7: Nuclear ratio of F_2 structure functions as measured by several experiments. The figure is taken from [43].

mechanism[44–46]. It is worth noting that while in the DIS, anti-shadowing can be clearly seen at $0.1 \lesssim x \lesssim 0.3$, it is not the case in the Drell-Yan (DY) lepton pair production process. The nuclear ratio data for the DY process is consistent with unity, namely, no anti-shadowing (this can be seen in Fig. 14 of Refs. [25]). As the valence quarks dominate in the DIS process, the anti-shadowing in the DIS then translates to the shadowing of the valence quark PDFs. Similarly, the absence of anti-shadowing in the DY process in this kinematical region translates to no shadowing in the sea quark PDFs.

In the EMC region, again, we observed a suppression of the ratio $R[F_2]$. Even though this region was the first to be experimentally measured to provide the evidence of nuclear effects, there is still no widely accepted model. In fact, there are a plethora of models that try to explain this effect, see a review in Refs. [47]. Among them are models based on : nuclear binding [48, 49], pion excess[50, 51], multi-quark clusters [52, 53], dynamical rescaling[54, 55], and short-range correlations (SRCs)[56, 57].

The models mentioned in the above passage tries to explain nuclear corrections in a specific kinematical region. There are also models which try to explain these for all x-region. Notably, among these are Kulagin-Petti model[58] and Aligarch-Valencia model[59]. Arguably, a more natural approach is to treat these nuclear corrections as effects originated from non-perturbative QCD, just like in the proton case, and hence can be absorbed into PDFs of the nucleus. Thus, one simply assumes the factorization theorem and hypothesizes that all the nuclear effects can be absorbed into non-perturbative nuclear parton distribution functions (nPDFs), just like in the proton case. This approach leads to nPDF framework adopted in this work.

A nucleus with $A = 2$ is a special case due to its simpler nature. A microscopic model adopted in the CJ15 analysis[10], for example, assume a nuclear smearing model, where the deuteron PDFs are assumed to be a convolution of the PDFs $\tilde{f}$ of the off-shell nucleon with a smearing function $f_{N/d}$, which describes momentum distribution of nucleons inside deuteron. The smearing function can be obtained in the weak-binding approximation in terms of the deuteron wave function, which is obtained by fitting nucleon-nucleon scattering data as done

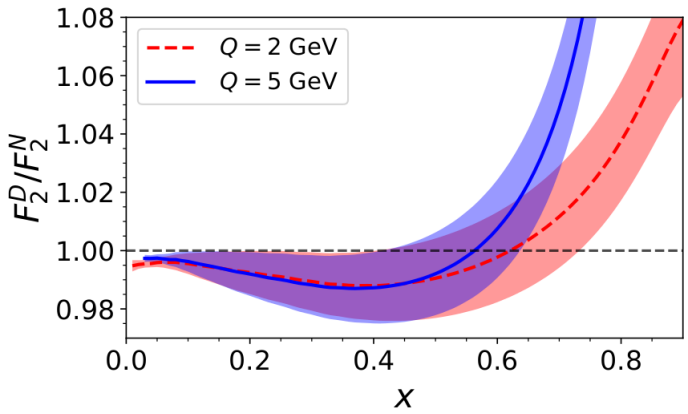


FIGURE 2.8: The ratio of deuteron structure function to isoscalar structure function from the CJ15 analysis[10]. The figure is taken from [60].

in AV18[61], CD-Bonn[62], WJC-1 and WJC-2[63]. For the PDFs $\tilde{f}$ of the off-shell nucleon, one can further decompose it as an on-shell contribution (which is essentially the same as the free proton PDFs) and an off-shell one. The off-shell correction is then parameterized and fitted together with the parameters of the proton PDFs to the data. In Fig. 2.8, we show the extracted the nuclear correction $R[F_2^D]$, where the denominator is the isoscalar stucture function $F_2^N = F_2^p + F_2^n$. We can see that it differs to the isoscalar one by $\sim 1\%$ in the EMC region.

2.6 Nuclear Parton Densities

As mentioned before, nuclear parton distribution functions (nPDFs) are essentially a generalization of proton PDFs to accomodate nuclear effects in a nucleus. They carry the same probabilistic interpretation as in the proton case and their scale evolution is governed by the same DGLAP equation. Let $\tilde{f}_i^A(x_A, \mu_F)$ be a nPDF of nucleus A for the parton flavor i at the scale μ_F and $0 \leq x_A \leq 1$ is the momentum fraction with respect to the nucleus momentum. Then the DGLAP equation reads

$$\frac{d\tilde{f}_i^A}{d\ln\mu_F^2} = \frac{\alpha_s(\mu_F^2)}{2\pi}\sum_j \int_{x_A}^1 \frac{dy_A}{y_A} P_{ij}\left(\frac{x_A}{y_A}\right)\tilde{f}_j^A(x_A, \mu_F^2)\,. \tag{2.46}$$

Being probability distributions, nPDFs satisfy the following number and momentum sum rules

$$\int_0^1 \tilde{u}_v^A(x_A, \mu_F)dx_A = 2Z + N\,, \tag{2.47a}$$

$$\int_0^1 \tilde{d}_v^A(x_A, \mu_F)dx_A = Z + 2N\,, \tag{2.47b}$$

$$\int_0^1 x_A \sum_i \tilde{f}_i^A(x_A, \mu_F) = 1\,. \tag{2.47c}$$

Here, Z is the atomic number and $N = A - Z$ is the number of neutron in the nucleus. The number sum rules (2.47a) and (2.47b) basically represent the external constraints, that a nucleus with a mass number A contains Z protons and N neutrons, where each proton carries two valence up quarks and one valence down quark and each neutron carries one valence up quark and two valence down quarks. To calculate a theory prediction for an observable in nucleus-nucleus or lepton nucleus collider, the same factorization theorem can be used. For a DIS cross section in lepton-nucleus scattering $l + A \rightarrow l' + X$, the total cross section is given by the convolution

$$\sigma(lA \rightarrow l' + X) = \sum_i \int dx_A \tilde{f}_i^A(x_A, Q^2) \hat{\sigma}_{iA \rightarrow l'X}(x_A p_A, Q^2)\,, \tag{2.48}$$

where $\hat{\sigma}_{iA \rightarrow l'X}$ is the hard scattering cross section. Similarly, for nucleus A and B collision with an identified hadron C in the final state, the cross section is given by

$$\begin{aligned}\sigma(A + B \rightarrow C + X) = \sum_{i,j,k} \int_0^1 dx_A dx_B dz_C\, \tilde{f}_i^A(x_A, \mu_F) \tilde{f}_j^A(x_B, \mu_F) \\ \times \hat{\sigma}_{ij \rightarrow cX}(x_A p_A, x_B p_B, \frac{p_C}{z}, \mu_F)\, D_{C/c}(z_C, \mu_F)\,,\end{aligned} \tag{2.49}$$

where $\hat{\sigma}_{ij \rightarrow cX}$ is the hard scattering $i + j \rightarrow c + X$ cross section and $D_{C/c}$ is the non-perturbative fragmentation function for an outgoing parton c fragments into a hadron C.

2.6.1 Rescaling

While the nPDFs formalism discussed in the previous section are just a straightforward generalization of proton PDFs, they are not the most useful as the support region for the nPDFs are A-dependent, hence it is not easy to compare the shape of nPDFS from different nuclei. For this reason, a rescaling formalism that will be discussed in this section, is useful. Another reason for using rescaling formalism is that all the DIS nuclear data are represented in terms of the rescaled Bjorken x_N, defined as

$$x_N = A x_A = \frac{Q^2}{2 M_N \nu} \leq A\,, \tag{2.50}$$

$$\nu = \frac{p_N . q}{M_N}, \qquad p_N = \frac{p_A}{A}, \qquad M_N = \frac{M_A}{A}\,. \tag{2.51}$$

Here, M_A and p_A are the mass and the momentum of the nucleus A. Therefore, it is indeed natural to work in the QCD factorization framework that uses the rescaled PDFs.

Given the unrescaled nPDF $\tilde{f}_i^A(x_A, Q^2)$, the rescaled version, $f_i^A(x, Q^2)$, with $x = A x_A$, is defined such that

$$f_i^A(x, Q^2) dx := \tilde{f}_i^A(x_A, Q^2) dx_A\,. \tag{2.52}$$

This definition ensures that the probability is conserved. It should be stressed once more that $x \in [0, A]$, while $x_A \in [0, 1]$. While x_A represent parton momentum fraction with respect to the nucleus momentum p_A, x represent parton momentum fraction with respect to the average

nucleon momentum $p_N = p_A/A$. If all the nucleons do not move to each other, then the momentum for each nucleon is exactly the same as p_N. However, due to Fermi motion, nucleons are moving within the nucleus, therefore, some nucleons can have momentum larger than p_N. A parton with $x > 1$ means its momentum is larger than p_N. As the parton itself is inside a nucleon, the nucleon that carries this parton must have momentum much bigger than p_N. This extreme uneven distribution of the nucleon momentum is very rare, therefore, we expect that $f_i^A(x > 1)$ is very small. In fact, in all recent nPDF analyses, it is always assumed that $f_i^A(x > 1) = 0$.

Given the sum rules (2.47), it is easy to derive the corresponding sum rules for the rescaled nPDFs :

$$\int_0^A dx\, u_v^A(x, Q^2) = 2Z + N\,, \tag{2.53a}$$

$$\int_0^A dx\, d_v^A(x, Q^2) = Z + 2N\,, \tag{2.53b}$$

$$\int_0^A dx\, x \sum_i f_i^A(x, Q^2) = A\,. \tag{2.53c}$$

The rescaled nPDFs also satisfy the analogous DGLAP evolution equations:

$$\frac{df_i^A(x, Q^2)}{d\ln Q^2} = \frac{\alpha_s(Q^2)}{2\pi} \int_x^A \frac{dy}{y}\, P_{ij}\left(\frac{x}{y}\right)\, f_j^A(y, Q^2)\,. \tag{2.54}$$

As said, most nPDF analyses assume $f_i^A(x > 1) = 0$, In this case, we have

$$\frac{df_i^A(x, Q^2)}{d\ln Q^2} = \begin{cases} \frac{\alpha_s(Q^2)}{2\pi} \int_x^1 \frac{dy}{y}\, P_{ij}\left(\frac{x}{y}\right)\, f_j^A(y, Q^2) & : 0 < x \le 1 \\ 0 & : 1 < x \le A\,. \end{cases} \tag{2.55}$$

This implies that, if $f_i(x > 1, Q_0) = 0$ for some initial scale Q_0, then $f_i(x > 1, Q) = 0$ for all. Thus, DGLAP equations do not change the vanishing values of PDFs at $x > 1$.

It is interesting to see how the factorization formula should be modified under the rescaling. Given that $dx_A \tilde{f}_i^A(x_A, Q^2) = dx f_i^A(x, Q^2)$ and $x_A p_A = x p_N$, we thus have the same factorization form :

$$\sigma(lA \to l' + X) = \sum_i \int dx_A f_i^A(x, Q^2) \hat{\sigma}_{iA\to l'X}(x p_N, Q^2)\,, \tag{2.56}$$

$$\begin{aligned} \sigma(A + B \to C + X) = \sum_{i,j,k} \int dx dy dz\, f_i^A(x, \mu_F) f_j^B(y, \mu_F) \\ \times \hat{\sigma}_{ij\to cX}(x p_N, y p_N, \frac{p_C}{z}, \mu_F)\, D_{C/c}(z, \mu_F)\,. \end{aligned} \tag{2.57}$$

Note here that the upper bound for x and y integrals is A, while the upper bound for z is still 1.

From (2.53), we can see that the sum of the total parton momentum computed using the rescaled PDFs f_i^A is equal to A. To have proper comparison with free nucleon PDFs, the sum

must be unity. Therefore, in practice, a second rescaling is performed :

$$f_i^A(x,Q^2) \to \frac{f_i^A(x,Q^2)}{A}\,. \tag{2.58}$$

Note that the argument of the PDFs are the same, it is just the values of the PDFs are rescaled by a factor of $1/A$. In terms of the second-rescaled PDFs, the sum rules (2.53) becomes

$$\int_0^A dx\, u_v^A(x,Q^2) = \frac{2Z+N}{A}\,, \tag{2.59a}$$

$$\int_0^A dx\, d_v^A(x,Q^2) = \frac{Z+2N}{A}\,, \tag{2.59b}$$

$$\int_0^A dx\, x \sum_i f_i^A(x,Q^2) = 1\,. \tag{2.59c}$$

while DGLAP evolution (2.55) stays the same. The factorization formula for DIS and SIH becomes :

$$\frac{1}{A}\sigma(lA \to l' + X) = \sum_i \int dx_A f_i^A(x,Q^2)\hat{\sigma}_{iA\to l'X}(xp_N,Q^2)\,, \tag{2.60}$$

$$\frac{1}{AB}\sigma(A+B\to C+X) = \sum_{i,j,k}\int dxdydz\, f_i^A(x,\mu_F) f_j^B(y,\mu_F) \times \hat{\sigma}_{ij\to cX}(xp_N, yp_N, \frac{p_C}{z}, \mu_F)\, D_{C/c}(z,\mu_F)\,. \tag{2.61}$$

Thus, using the second-rescaled PDFs, we obtain formulas for calculating the per-nucleon cross section. As it turns out, most nuclear data are also rescaled with A, giving a per-nucleon cross section instead of the total one. Therefore, the theory calculated using the second-rescaled PDFs can be directly compared to the data, without any need to rescale back. For brevity, from now on, we will call the second-rescaled nPDFs as simply nPDFs.

In all nPDF analyses, the nPDFs are decomposed in terms of effective bound proton $f_i^{p/A}$ and neutrino $f_i^{n/A}$ PDFs as as:

$$f_i^{(A,Z)}(x,Q) = \frac{Z}{A} f_i^{p/A}(x,Q) + \frac{A-Z}{A} f_i^{n/A}(x,Q) \quad . \tag{2.62}$$

We note that in the nPDF framework, at the theoretical level, it is actually unnecessary to decompose the full nPDFs in terms of the bound nucleon PDFs. The benefit of using this decomposition is that it is easy to implement isospin symmetry:

$$\{u^{p/A}, \bar{u}^{p/A}\} \leftrightarrow \{d^{n/A}, \bar{d}^{n/A}\}\,, \tag{2.63}$$

such that these bound nucleon PDFs really resemble the free nucleon ones in terms of valence number sum rules:

$$\int_0^A u_v^{p/A}(x,Q^2)dx = 2, \qquad \int_0^A u_v^{n/A}(x,Q^2)dx = 1\,, \tag{2.64}$$

$$\int_0^A d_v^{p/A}(x,Q^2)dx = 1, \qquad \int_0^A d_v^{n/A}(x,Q^2)dx = 2\,. \tag{2.65}$$

The main consequence of the decomposition (2.62) and isospin symmetry (2.63) is that for an isoscalar nucleus with $N = Z = A/2$, we have

$$u^A(x,Q^2) = d^A(x,Q^2), \qquad \bar{u}^A(x,Q^2) = \bar{d}^A(x,Q^2), \qquad u_v^A(x,Q^2) = d_v^A(x,Q^2)\,. \tag{2.66}$$

This implies that data taken on isoscalar nucleus can not separate u and d.

2.6.2 nCTEQ Fitting Framework

nPDFs are determined by a global analysis with world data, therefore, they are dependent on the fitting methodology used in the analysis. It is therefore not so surprising that nPDFs from different nPDF fitting groups, such as nCTEQ[25], EPPS[24], and nNNPDF[26] are slightly different. Their differences can eventually be traced back to the differences in : 1) choice of data sets, 2) choice of proton PDF baseline, 3) parameterizations of the nPDFs at the input scale Q_0, and 4) the method of error estimation.

In this section, we present fitting framework employed in nCTEQ15 analysis[25] and the recent nCTEQ15HQ analysis[64]. The full nPDFs are decomposed in terms of the effective bound nucleon PDFs, as discussed in the previous section. At the input scale $Q_0 = 1.3$, the bound proton PDFs are parametrized as [25]:

$$xf_i^{p/A}(x,Q_0) = c_0 x^{c_1}(1-x)^{c_2}e^{c_3 x}\,(1+e^{c_4}x)^{c_5}\,, \tag{2.67}$$

$$\frac{\bar{d}(x,Q_0)}{\bar{u}(d,Q_0)} = c_0 x^{c_1}(1-x)^{c_2} + (1+c_x)(1-x)^{c_4}, \tag{2.68}$$

where the flavor index i runs over $i = u_v, d_v, g, \bar{u}+\bar{d}, s+\bar{s}$. Here u_v and d_v are the up- and down-quark valence distributions, and $g, \bar{u}, \bar{d}, s, \bar{s}$ are the gluon, anti-up, anti-down, strange, and anti-strange quark distributions, respectively. The A, Z-dependence in encoded in c_k via

$$c_k(A,Z) = p_k + a_k(1-A^{b_k}). \tag{2.69}$$

This parameterization ensures that for $A = 1$, we obtain $c_k(A=1,Z) = p_k$, where p_k is the proton PDF parameters obtained from the old CTEQ6 fit[65]. In other word, the proton PDF baseline is the CTEQ6 PDFs.

The nPDF fitting then can be done by a flowchart shown in Fig. 2.6. First, we must perform data selection. In nCTEQ15 analysis, the charged lepton DIS, Drell-Yan lepton pair productions,

and Pion production data were used. The fit can be done by minimizing the χ^2 function :

$$\chi^2(a) = (D - T(a))^T C^{-1} (D - T(a)) \,, \tag{2.70}$$

where D, T, and C represent the data, theory prediction, and the data covariance matrix. The theory predictions are calculated using (2.60), which requires us to know the PDFs at the scale Q of the data. The PDFs at the scale Q can be obtained by using DGLAP evolution from the input scale Q_0, at which the PDFs are parameterized. The theory predictions then are functions of PDF parameters a. In nCTEQ analyses, the minimization is done using Minuit Migrad[66]. After the optimal PDF parameters a are found, then the errors of the fitted PDFs are determined using Hessian method.

Chapter 3

Statistics Aspects of nPDF Fitting

The determination of parton distribution functions (PDFs) is usually done through a global QCD analysis with a set of data taken from various high-energy physics experiments. The global analysis is performed by iteratively comparing perturbative QCD (pQCD) predictions with the experimental data until an optimal set of PDFs is obtained. PDF fitting is an infinite dimensional inverse problem, which is not necessarily well-defined as one tries to infer an infinite number of parameters from a finite number of data. In practice, however, the PDFs are usually parameterized in terms of a finite number of parameters. In this case, the cost function is no longer functional, but merely a function.

PDF fitting estimates the parameters from the data, which can be considered as a random variable. Therefore, several statistical details needs to be properly taken care off. This includes: choosing the right loss/cost function that takes into account correlated systematic uncertainties, how the uncertainties of the data are propagated into the fitted parameters, how the impact of specific data on the fitted parameters, and how the tensions between data sets can be quantified and assessed.

In this chapter, we will give an overview of the statistics aspect of nPDF global analysis by considering it as a finite-dimensional inverse problem. As this topic is usually not covered in the standard textbooks on pQCD and generally only found in the publications of the PDF fitting experts, we give quite a detailed discussion on this. We first explain the derivation of χ^2 function used in the nCTEQ analysis based on the normality assumption of the data. As the data itself is a random variable, we also discuss several methods to propagate the statistical properties of the data to the fitted nPDFs. Here, we emphasize more on Hessian method as it is widely used in the literature. we will then discuss replica and Bayesian method. At the end of this chapter, we will discuss how to quantify tensions between data sets.

3.1 Loss or χ^2 Function

Let a set of pairs $D = \{(x_1, y_1), (x_2, y_2), ..., (x_{N_D}, y_{N_D})\}$ be the data. The datum is denoted as $D_i = (x_i, y_i)$, where x_i is the input (for example, Bjorken x, rapidity y, p_T, etc), and y_i denotes the measured observable (cross section, structure function, etc). The goal is to derive a cost function whose minimum serves as an ideal estimator (unbiased, consistent, and efficient) for

the theory parameter a_μ, $\mu = 1, 2, 3, ...N_p$. For this purpose, we can use the maximum likelihood method. The maximum likelihood estimation is a standard estimator whose (asymptotic) statistical properties are widely known. Let $L(D|a)$ the probability density of the whole data D given the theory parameter a. This probability density is often called the likelihood. Then we can define the cost function as

$$\chi^2(D, a) = -2 \ln p(D|a) \tag{3.1}$$

Therefore, maximizing the likelihood is then equivalent to minimizing the cost function. The likelihood can be derived by modeling the data generation process. The standard assumption on the data generation is that the data is normally distributed with the variance given by the square of the data uncertainty. Based on how the uncertainty is related to the data, one can distinguish an additive and multiplicative uncertainty. An additive uncertainty means that the ratio of the data uncertainty to the data is not constant for all data points. On the other hand, a multiplicative uncertainty is always tied to the data linearly, such that the ratio of uncertainty to the data is always constant. The multiplicative uncertainty, which is related to the conversion of event count to cross section data, is often called normalization uncertainty. For the rest of this thesis, we will use the terms normalization and multiplicative uncertainty interchangeably. The difference between the nature of the additive and multiplicative uncertainties results in different likelihood functions, as discussed below.

3.1.1 Additive Errors

Given a data set D used in a global fit that contains N_D data points. For each data point, there is a statistical uncertainty, N_{unc} uncorrelated systematic errors, and N_{corr} correlated systematic errors. We assume that all the systematic uncertainties are additive. We model the distribution of the data point D_i in a way such that D_i is related to the true value $\langle D_i \rangle$ as

$$D_i = \langle D_i \rangle + \sigma_i r_i + \sum_{\alpha}^{N_{corr}} \bar{\sigma}_{i\alpha} \bar{r}_\alpha + \sum_{\alpha}^{N_{unc}} \tilde{\sigma}_{i\beta} \tilde{r}_{i\beta} \,. \tag{3.2}$$

Here, r_i is distributed with standard normal distribution $r_i \sim \mathcal{N}(0, 1)$. At this stage, $\bar{r}_\alpha$ and $\tilde{r}_{i\beta}$ are nuisance parameters whose values are unknown. The likelihood function can be obtained by replacing $\langle D_i \rangle$ with the theory prediction $T_i(a)$, where a is the theory parameters, and taking the expectation value of the delta function

$$L(a) \propto \prod_{i,\alpha,\beta} \int dr_i dr_{i\alpha} dr_\beta \, e^{-\frac{{r_i}^2}{2}} e^{-\frac{r_{i\alpha}^2}{2}} e^{-\frac{{r_\beta}^2}{2}} \times \delta \left(D_i - T_i(a) - \sigma_i r_i - \sum_{\alpha}^{N_{corr}} \bar{\sigma}_{i\alpha} \bar{r}_\alpha - \sum_{\alpha}^{N_{unc}} \tilde{\sigma}_{i\beta} \tilde{r}_{i\beta} \right) , \tag{3.3}$$

where we have include a standard normal prior in order to eliminate the nuisance parameters. Performing the integration over r_i and $\tilde{r}_{i\beta}$, we obtain

$$L(a) \propto \int d\bar{r}_\alpha \exp\left[-\frac{1}{2}\sum_i \frac{\left(D_i - T_i - \sum_\alpha \bar{\sigma}_{i\alpha}\bar{r}_\alpha\right)^2}{\sigma_i^{unc2}} - \frac{1}{2}\sum_\alpha \bar{r}_\alpha^2\right] \equiv \exp\left[-\frac{1}{2}\chi^2(a)\right], \tag{3.4}$$

where

$$\sigma_i^{unc2} \equiv {\sigma_i}^2 + \sum_\beta \sigma_{i\beta}^2 . \tag{3.5}$$

The integrated likelihood, given the Gaussian prior $\Pi(r'_\alpha) \propto \exp(-{r'_\alpha}^2/2)$, is then given by

$$L(a) \propto \prod_\alpha \int \exp\left[-\frac{1}{2}\left(\sum_i \frac{\left(D_i - T_i - \sum_\alpha \bar{\sigma}_{i\alpha} r'_\alpha\right)^2}{\sigma_i^2} + \sum_\beta {r'}_\beta^2\right)\right] dr'_\alpha \propto \exp\left[-\frac{1}{2}\chi^2(a)\right] . \tag{3.6}$$

The Gaussian integral can be evaluated analytically and thus, $\chi^2(a)$ is given by

$$\chi^2(a) = \sum_{i,j}(D_i - T_i(a))(D_j - T_j(a)) \times \left[\frac{\delta_{ij}}{\sigma_i^2} - \frac{1}{\sigma_i^2\sigma_j^2}\sum_{\alpha,\beta}\bar{\sigma}_{i\alpha}({A'}^{-1})_{\alpha\beta}\bar{\sigma}_{j\beta}\right] \tag{3.7}$$

$$\equiv (D-T)^T C^{-1}(D-T) , \tag{3.8}$$

where $A'_{\alpha\beta} = \delta_{\alpha\beta} + \sum_i \frac{\bar{\sigma}_{i\alpha}\bar{\sigma}_{j\alpha}}{\sigma_i^2}$ and C is the covariance matrix, defined by

$$C_{ij} = \sigma_i^2\delta_{ij} + \sum_\alpha \bar{\sigma}_{i\alpha}\bar{\sigma}_{j\alpha} . \tag{3.9}$$

Thus, Maximum likelihood principle reduce to the familiar weighted least square method.

3.1.2 Normalization Uncertainty

Normalization uncertainty is a scale uncertainty that affects both the central data and its uncertainties. It arises when converting an event number to a physical cross section. As such, the error is usually written in percentage and constant for all data points.

Here, we review several prescriptions available in the literature :

- The D-method. Here, one use the following χ^2 function:

$$\chi^2_D(a,r) = (rD - T(a))^T C^{-1}(rD - T(a)) + \frac{(1-r)^2}{\sigma^2_{norm}}. \tag{3.10}$$

As χ^2_D is quadratic in r, one can minimize r analytically to obtain

$$\tilde{\chi}^2_D \equiv \min_r \chi^2_D(a,r) = (D-T)^t C_D^{-1}(D-T) , \tag{3.11}$$

where :

$$C_{D,ij} = C_{ij} + \sigma^2_{norm} D_i D_j \,, \tag{3.12}$$

$$C_{ij} = \sigma_i^2 \delta_{ij} + \sum_\alpha \bar{\sigma}_{i\alpha} \bar{\sigma}_{j\alpha} \,, \tag{3.13}$$

and σ_{norm} is the normalization uncertainty. The symbol ^{t} in 3.11 refers to transpose operation. Thus, (3.10) and (3.11) are therefore *equivalent*.

To prove (3.11), one can start by writing the optimal value for r, given the data D and theory T. For the derivations presented below, it is useful to define :

$$A = 1 + \sigma^2_{norm} T^T C^{-1} T \,, \tag{3.14a}$$

$$B = 1 + \sigma^2_{norm} D^T C^{-1} T \,, \tag{3.14b}$$

$$E = 1 + \sigma^2_{norm} D^T C^{-1} D \,. \tag{3.14c}$$

We can therefore write

$$\chi^2_D(a, r) = \frac{1}{\sigma^2_{norm}} \left[E \left(r - \frac{B}{E} \right)^2 + \left(A - \frac{B^2}{E} \right) \right] .$$

The fitted normalization r_D is then given by :

$$r_D \equiv \arg\min_r \chi^2_D = \frac{B}{E} \,. \tag{3.15}$$

Therefore,

$$\tilde{\chi}^2_D = \min_r \chi^2(a, r) = \chi^2(a, r_D) = \frac{1}{\sigma^2_{norm}} \left(A - \frac{B^2}{E} \right) = (D - T)^t C_D^{-1} (D - T), \tag{3.16}$$

where we have used the Sherman-Morrison formula [67] to write the inverse of C_D as :

$$C_D^{-1} = C^{-1} - \frac{\sigma^2_{norm} C^{-1} D D^T C^{-1}}{1 + \sigma^2_{norm} D^T C^{-1} D} \,. \tag{3.17}$$

The main drawback of this method is that it can lead to the d'Agostini bias [68] which causes the fitted theory to be much lower than expected. Furthermore, the bias becomes worse as the number of data points increases [68, 69]. This can be understood as follows. When one used (3.10), or equivalently (3.11), the minimizer algorithm will always prefer to have $r \leq 1$. The scale $r \leq 1$ makes the data smaller without rescaling the uncertainties. Therefore, relative to the data, the errors become larger, and hence smaller χ^2. As a result, the theory T is also shifted downward, leading to a biased fit. Note that as we have more data points, the penalty term $(1 - r)^2 / \sigma^2_{norm}$, which makes the fit to prefer $r = 1$, is becoming less relevant, and hence the bias is more apparent.

To illustrate the bias in a real PDF fit, we performed fits with neutrino DIS data from

NuTeV[70] and Chorus[71]. These data have the same normalization uncertainty of 2.1%, but NuTeV data is much more numerous than Chorus (2136 vs. 824). After the fits, we obtained $\chi^2/N = 0.86$ and $\chi^2/N = 0.95$ for the NuTeV and Chorus fit respectively. We plot the weighted average of data/theory in the left panel of Fig. 3.1. The figure shows that the theory is severely below the data. We can also see that the bias in the NuTeV fit is more severe than in the Chorus fit. This is because NuTeV has much more data points than Chorus.

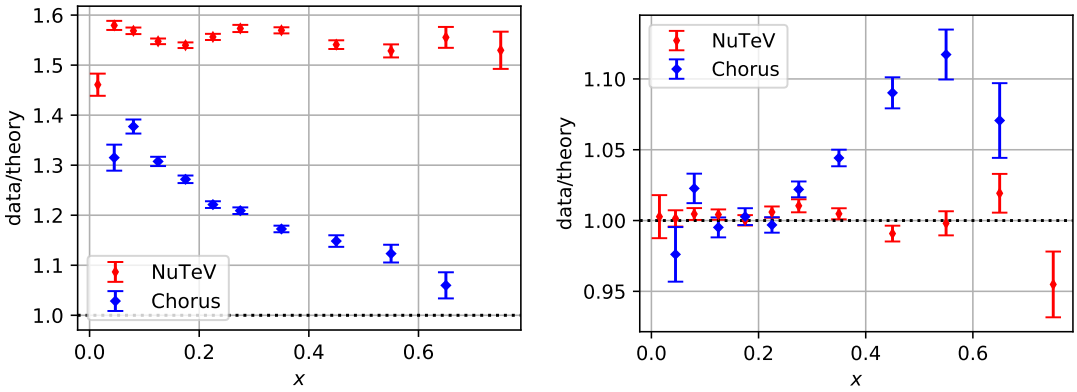


FIGURE 3.1: The weighted average of the data/theory from fits with NuTeV and Chorus data where the normalization uncertainties are treating using (3.10) (top panel) and using the method adopted in this work (3.19) (bottom panel).

- d' Agostini method[68]. Now, the errors of the data are also rescaled by a factor r, and thus, effectively rescaling the theory :

$$\chi^2_{1/r}(a,r) = \sum_{i,j}\left(D_i - \frac{T_i}{r}\right) C^{-1}_{ij}\left(D_j - \frac{T_j}{r}\right) + \left(\frac{1-r}{\sigma_{norm}}\right)^2 . \tag{3.18}$$

This method requires to fit the normalization fluctuation, r, directly to the data. The main drawback of this approach is that the number of normalization parameters can become large, and in case there are many data sets in the global fit, even comparable to the number of PDF parameters. This causes the fit to be prone to numerical problems, such as saddle point or local minimum trap. The larger number of parameters also means the computing cost will increase. Furthermore, the uncertainty of the nuisance parameter r must also be taken into account when estimating the uncertainty of observables. Further discussion on this is given in Section 3.2.1.

- T-method[72, 73]. Instead of rescaling the theory by a factor of $1/r$, a factor r is used instead :

$$\chi^2_r(a,r) = \sum_{i,j}(D_i - rT_i)C^{-1}_{ij}(D_j - rT_j) + \frac{(1-r)^2}{\sigma^2_{norm}} . \tag{3.19}$$

The main advantage of using (3.19) is that the normalization r can now be minimized analytically, hence there is no need to open more free parameters in the fit. We can rewrite

(3.19) as

$$\chi^2_r(a,r) = \frac{1}{\sigma^2_{norm}}\left[A\left(r - \frac{B}{A}\right)^2 + \left(E - \frac{B^2}{A}\right)\right] \tag{3.20}$$

It is clear now that the fitted normalization is given by :

$$r_T = \arg\min_r \chi^2_r(a,r) = \frac{B}{A} \tag{3.21}$$

Inserting $r = r_T$ into (3.19), we obtain

$$\begin{aligned}\tilde{\chi}^2_T(a) &\equiv \min_r \chi^2_r(a,r) = \frac{1}{\sigma^2_{norm}}\left(E - \frac{B^2}{A}\right) \\ &= (D-T)^T C_T^{-1}(D-T)\end{aligned} \tag{3.22}$$

where

$$C_{T,ij}(a) = C_{ij} + \sigma^2_{norm} T_i(a) T_j(a) \tag{3.23}$$

and we have used the following formula for the inverse of C_T :

$$C_T^{-1} = C^{-1} - \frac{\sigma^2_{norm} C^{-1} T T^T C^{-1}}{1 + \sigma^2_{norm} T^T C^{-1} T} \tag{3.24}$$

Thus, the fitting normalization uncertainty using (3.19) is equivalent to using an effective covariance matrix C_T. The advantage of using this approach is that the nuisance parameters are completely eliminated, and the Hessian error method, to be discussed in the next section, automatically takes into account the uncertainty of the nuisance parameters in the estimation of PDF uncertainties. As the difference between T-method and d' Agostini method essentially comes from the penalty term, they are equivalent if the optimal normalization parameter r_T is not far from unity, which is usually the case.

It is trivial to generalize this method to a case where there are more than one data set that share the same normalization. In such case, (3.21) still holds, but A, B and C are modified as

$$A = 1 + \sum_s \sigma^2_{norm} T^{s\,T} C_s^{-1} T^s \tag{3.25}$$

$$B = 1 + \sum_s \sigma^2_{norm} D^{s\,T} C_s^{-1} T^s \tag{3.26}$$

$$E = 1 + \sum_s \sigma^2_{norm} D^{s\,T} C_s^{-1} D^s \tag{3.27}$$

where s denotes the data set s and the sum is done over all data sets that share the same normalization.

In order to contrast the fit results obtained with D-method, we performed analogical fits using (3.19)and show the weighted average of the data/theory in the right panel of Fig. 3.1. We can see that for both NuTeV and Chorus fits, the ratio becomes much closer

to unity. The relatively high data/theory values for the Chorus fit at $x > 0.4$ is related to large systematic uncertainties (hence large systematic theory shifts). As far as the χ^2 is concerned, we obtain much larger χ^2_r/N (compared to the one form the D-method) : $\chi^2_r/N = 1.36$ and $\chi^2_r/N = 1.07$ for the NuTeV and Chorus fits respectively.

- T_0-method[73]. An alternative method to include normalization uncertainties in a global fit is to use t_0-method as explained in detail in [73]. This method sets the covariance matrix as:

$$C_{t_0,ij} = C_{ij} + \sigma^2_{norm} T_{0i} T_{0j} \tag{3.28}$$

where T_{0i} is the theory prediction from the previous iteration of the fit and C_{ij} is the original covariance matrix without normalization uncertainties. This method eliminates the nuisance parameters from the χ^2 function, and hence their uncertainties are automatically included. As T_0 is frozen during the fit, this method has the advantage of having a simpler χ^2 function (as the covariance matrix is not a function of theory parameters a), hence the minimizer should be easier to find the global minimum. However, several fit iterations need to be done in order to use this method, hence it is more computationally expensive.

3.2 Error Estimations

One of the main issue in PDF global analysis is how the uncertainties of the fitted PDFs are estimated. The usual way is to use the Hessian method, for which the χ^2 function is assumed to be quadratic with respect to the PDF parameters. It is then possible to derive a formula that can be used to estimate the uncertainty of the fitted PDFs. However, this method depends on the choice of the so-called $\Delta\chi^2$ tolerance, whose value depends on: 1) tensions between data sets, 2) fitting methodology mistakes, such those that are related to the inflexibility of the PDF parameterization, assumption about the distribution of the data uncertainties, treatment of the correlated systematic uncertainties, and theory uncertainties, and 3) issues with the experimental data, such as missing correlation, inaccuracy of the data and/or its uncertainties. To estimate the tolerance, there are several prescriptions in the literature. The global tolerance, where a single value of $\Delta\chi^2$ is used for all the Hessian eigenvector directions, is used in [24, 25, 74]. Dynamical tolerance criterion was adopted in CT PDFs[7, 8], which assign different $\Delta\chi^2$ values for each eigenvector direction. The resulting $\Delta\chi^2$ generally ranges from ~ 10 to ~ 100.

The PDF uncertainties obtained using the Hessian method with an enlarged tolerance can be regarded as uncertainties in the hypothesis testing sense. This is because the tolerance is typically determined using a hypothesis testing approach, by requiring all the data sets used in the global analysis to be explainable by the error PDFs. This type of uncertainty is used in some PDF analyses because some of the high energy physics data have misestimated uncertainties and correlations, therefore, the standard frequentist uncertainties based on χ^2 tolerance of $T^2 =$

1, can lead to PDF uncertainties that can not even explain the data (in the hypothesis testing sense).

Omitting the requirement that PDF uncertainties must be able to explain all the data used in the analysis, we have the standard frequentist uncertainties. In this approach, the PDF uncertainties are obtained by propagating the uncertainties of the data. We will show later that, the presence of tensions between data sets and model misspecifications do not affect the validity of using $T^2 = 1$ in the Hessian approach.

The hessian method is not the only method used to estimate the PDF uncertainties. NNPDF group[9, 75] use a replica-based method to sample the distribution of the fitted PDFs. The appeal of this approach is that it does not assume χ^2 to be quadratic, but it is computationally much more expensive than the Hessian method. Both the Hessian and replica methods are based on the so-called frequentist view of probability distribution. An alternative picture is a bayesian view, which in the PDF fitting context, estimates the PDFs and their uncertainties by sampling the posterior distribution of the PDF parameters from the data. We will show later that, In the case of a sufficiently linear model and Gaussian errors, the Hessian, replica, and bayesian methods should be equivalent. Therefore, choosing either of them is a matter of choice with consideration on computational cost in mind.

3.2.1 Error Estimation : Hessian Method

Hessian error method is the most common way to propagate the uncertainties of the data to the uncertainty of observables. The main appeal of this method is its simplicity. In the statistics literature (see for example : [76–78]), two related methods are often encountered to estimate the uncertainties of the fitted parameters : the likelihood-based method and the normal theory method. The former is based on Wilk's likelihood ratio statistic $l = L(a)/L(\hat{a})$ (here $L(a)$ is the likelihood for a given theory parameter a), whose $-2\ln L(a)$ is distributed asymptotically as a χ^2-distribution with N_p degrees of freedom (here N_p is the number of theory parameters). This method has been used by physicists from a long time and in Minuit program[66], this is called the MINOS error. The latter is based on asymptotic properties of the maximum likelihood estimation. The hessian method, which appears in PDF-related literature, is the same as the normal theory method. In the Minuit framework, it is called "parabolic error". The likelihood-based interval is generally preferred when the likelihood function is multimodal and the confidence interval contains multiply connected regions. Using normal theory method in such a case can severely underestimate the resulting parameter errors.

Linear Error Propagation

Let a_μ denotes the PDF parameters at the given initial scale Q_0. Here, $\mu = 1, 2, 3, ..., N_p$, where N_p is the number of parameters. Let $\chi^2(a, D)$ be the cost function to be minimized to get the

best fit for a, then the fitted PDF parameters, $\hat{a}_\mu$, can be written as

$$\hat{a}(D) = \arg\min_a \chi^2(a, D) \tag{3.29}$$

Since $\hat{a}$ is a function of D which is a random variable, then it is also a random variable. The uncertainties of any quantity X can be obtained by propagating the uncertainties of $\hat{a}$. We note here that the distribution of the fitted parameters, $p(\hat{a})$, does not care whether the data are consistent or not. It only depends on the functional form of $\chi^2(a, D)$ and how D is distributed.

Formally, let $p(\hat{a}_1, ..., \hat{a}_p)$ the probability distribution of the fitted parameters $\hat{a}$, then the uncertainty of PDF $f_i(x, Q)$ can be obtained by the standard deviation

$$\Delta f_i(x, Q) := \sqrt{\mathrm{Var}(f_i(x, Q))} \tag{3.30}$$

For any observable X which is a function of the PDF parameters, we can propagate the uncertainties of the fitted parameters $\hat{a}$ using the standard method. Let $\langle \hat{a} \rangle$ be the mean of $\hat{a}$, then around $\hat{a} = \langle \hat{a} \rangle$, one expand X as

$$X(\hat{a}) \approx X(\langle \hat{a} \rangle) + \left. \frac{\partial X}{\partial \hat{a}_\mu} \right|_{\hat{a}=\langle \hat{a} \rangle} (\hat{a} - \langle \hat{a} \rangle) \tag{3.31}$$

This means

$$\langle X(\hat{a}) \rangle = X(\langle \hat{a} \rangle) \tag{3.32}$$

$$\mathrm{Var}(X(\hat{a})) = \sum_{\mu,\nu} \frac{\partial X}{\partial \hat{a}_\mu} \mathcal{C}_{\mu\nu} \frac{\partial X}{\partial \hat{a}_\nu} \tag{3.33}$$

where $\mathcal{C}_{\mu\nu} = \mathrm{Cov}\left(\hat{a}_\mu, \hat{a}_\nu\right)$ is the covariance matrix. Here, the derivatives are understood to be evaluated at $\hat{a} = \langle \hat{a} \rangle$.

Under suitable assumptions, the distribution of the fitted parameters is Gaussian. This normality is also automatic if the estimator (3.29) is the maximum likelihood and the number of data is large. Denoting

$$\mathcal{H}_{\mu\nu} = \mathcal{C}^{-1}_{\mu\nu} \tag{3.34}$$

the joint probability distribution is then given by

$$p(\hat{a}_1, ..., \hat{a}_{N_p}) = \frac{1}{\sqrt{(2\pi)^{N_p} \det \mathcal{H}}} \exp\left(-\frac{1}{2} \left(\hat{a} - \langle \hat{a} \rangle\right)^T \mathcal{H} \left(\hat{a} - \langle \hat{a} \rangle\right) \right) \tag{3.35}$$

It is useful to work in a basis such that the fitted parameters are independent (uncorrelated). This requires us to assume that the estimator is invariance, namely, if $\hat{a}$ is the estimator for a, then the estimator for any function of a, $f(a)$, is just $f(\hat{a})$. Fortunately, the maximum likelihood

estimator, which we will assume in this paper, satisfies this invariance property. Now, define:

$$z = \mathcal{H}_{diag}^{1/2} V^T (a - \langle a \rangle), \tag{3.36}$$

with V is $N_p \times N_p$ matrix that diagonalize $\mathcal{H}$, namely $\mathcal{H} = V \mathcal{H}_{diag} V^T$. Let $\hat{z}$ be the estimator for z, then from the invariance properties and (3.35), $\hat{z}$ is distributed as

$$p(\hat{z}) = \prod_{\mu} p(\hat{z}_\mu) \tag{3.37}$$

with $p(\hat{z}_\mu) \sim \mathcal{N}(0,1)$. Thus, $\mathrm{Cov}(\hat{z}_\mu, \hat{z}_\nu) = \delta_{\mu\nu}$. Therefore,

$$\mathrm{Var}(X(\hat{z})) = \sum_{\mu} \left(\frac{\partial X}{\partial \hat{z}_\mu} \right)^2 \tag{3.38}$$

We can estimate the derivative $\partial X / \partial \hat{z}$ by using finite difference. Let $S_\mu^{\pm} = \{a_1, a_2, ..., a_{N_p}\}$ be a set of parameter a that correspond to z whose μ-th component is ± 1, then

$$\frac{\partial X}{\partial \hat{z}} \approx \frac{X(S_\mu^+) - X(S_\mu^-)}{2} \tag{3.39}$$

Inserting this to (3.38), we find

$$\mathrm{Var}(X) = \frac{1}{4} \sum_{\mu} \left(X(S_\mu^+) - X(S_\mu^-) \right)^2 \tag{3.40}$$

Having derived almost all the necessary formula to propagate the uncertainty of the fitted parameters to an observable, the remaining puzzle is to estimate the covariance of the fitted parameter.

Given the covariance matrix $\mathcal{C}_{\mu\nu}$, one can obtain p% the confidence region of the fitted parameters. To find the region, it is easier to work with z instead of a as z is uncorrelated and distributed according to the standard normal distribution. Let's define[1] $\chi_z^2 = \sum_\mu z_\mu^2$. As z_μ is a standard normal variable, χ_z^2 is distributed according to χ^2 distribution with N_p degrees of freedom. Let $\chi^2_{N_p, p\%}$ be the p% percentile of the χ^2 distribution with N_p degrees of freedom, then the confidence region for the fitted parameter with probability content p% is given by :

$$R_{p\%} = \left\{ z \;\middle|\; \sum_\mu z_\mu^2 \leq \chi^2_{N_p, p\%} \right\} \tag{3.41}$$

As an example, for $N_p = 20$ and $p = 90\%$, the confidence region is then given by a hypersphere with radius $\sqrt{\chi^2_{20,90\%}} = 5.3$.

[1] In the statistics literature, it is called Wald statistic.

Hessian Matrix as the Inverse of The Covariance Matrix

To obtain the covariance of the fitted parameters, one can utilize asymptotic property of MLE. Let $\hat{a}(D)$ be the MLE estimate (fitted parameters) from a fit with the data D and a^0 be the true parameters. Then $(\hat{a}(D) - a^0)$ is asymptotically normal with mean 0 and covariance matrix $\mathcal{C}_{\mu\nu}$ given by :

$$\mathcal{C}^{-1}_{\mu\nu} = -E\left[\frac{\partial^2 \ln L(a^0)}{\partial a_{0,\mu}\,\partial a_{0,\nu}}\right] = E\left[\frac{1}{2}\frac{\partial \chi^2(a^0)}{\partial a_{0,\mu}\,\partial a_{0,\nu}}\right] \tag{3.42}$$

Here, the expectation value E is evaluated with respect to the probability of the data. As the true parameters and the data probability are unknown, one can approximate $\mathcal{C}_{\mu\nu}$ by using $\hat{a}$ as

$$\mathcal{C}^{-1}_{\mu\nu} \approx \frac{1}{2}\frac{\partial \chi^2(\hat{a})}{\partial a_{0,\mu}\,\partial a_{0,\nu}} = H_{\mu\nu} \tag{3.43}$$

The matrix $H_{\mu\nu}$ is the Hessian matrix, obtained by taking a second derivative to the χ^2 function at $a = \hat{a}$.

Given that the inverse of the covariance matrix is just the Hessian, then there is a direct relation between the χ^2 and confidence region of the fitted parameters. As now $\mathcal{H} = H$, then z defined in (3.36) is just the parameter a in the eigenvector basis. Writing

$$\chi^2(a) = \chi^2(\hat{a}) + (a-\hat{a})^t H (a-\hat{a}) = \chi^2(\hat{a}) + \sum_\mu z_\mu^2 \tag{3.44}$$

This means, as individually $z_\mu \sim \mathcal{N}(0,1)$, then 1σ deviation of the parameter z_μ while fixing the others to zero corresponds to :

$$\Delta\chi^2 \equiv \chi^2(a(z=1)) - \chi^2(\hat{a}) = 1 \tag{3.45}$$

In the case where tensions between data sets are present, the so-called global tolerance method is sometimes used to modify the covariance matrix. In this case, the covariance matrix is enlarged by a factor T^2:

$$\mathcal{C}_{\mu\nu} = T^2 H^{-1}_{\mu\nu} \tag{3.46}$$

where $T^2 \geq 1$ is the global tolerance. The tolerance T^2 is usually determined by using hypothesis testing, such that all data sets can be explained at some p% (usually $p = 90$) confidence level by all points inside the one σ confidence region of the fitted parameters specified by $\mathcal{C}_{\mu\nu}$. This method is used in nPDF analyes by nCTEQ[25, 60, 64, 72, 79] and EPPS group[24, 80]. In this case, one can relate $\chi^2(a)$ to z_μ defined in (3.36) as:

$$\chi^2(a) = \chi^2(\hat{a}) + (a-\hat{a})^t H (a-\hat{a}) = \chi^2(\hat{a}) + T^2 \sum_\mu z_\mu^2 \tag{3.47}$$

Thus, $z_\mu = 1$ (1σ deviation) while fixing the others to zero corresponds to $\Delta\chi^2 = T^2$.

Nuisance Parameters in the Hessian Method

In section 3.1, we discussed how the nuisance parameters were introduced to accommodate correlated systematic and normalization uncertainties. In this section, we will discuss how one can take into account the uncertainty of the nuisance parameters in the estimation of the errors of observables.

If the nuisance parameters are eliminated from the χ^2 as in (3.8) and (3.22), then the Hessian formalism automatically includes the contributions from the nuisance parameters. If, on the other hand, the nuisance parameters are not eliminated, such as in (3.18), then one needs to take into account the contribution from the nuisance parameters manually, by using, for example, profile likelihood method. For a $\chi^2(a_\mu, r_i)$ function, where a_μ, $\mu = 1, ..., N$ denotes the parameters of interest (PDF parameters) and r_i, $i = 1, ..., M$ are the nuisance parameters, one defines the "profile" χ^2 function as

$$\chi^2_p(a) := \min_r \chi^2(a, r) \tag{3.48}$$

which is a function of theory (PDF) parameters only. As the nuisance parameters are eliminated in χ^2_p, one can use the Hessian method using χ^2_p, and it will automatically include the contribution from r. However, the computation of χ^2_p is expensive as there is no closed-form solution for χ^2_p. Hence this method is impractical.

An alternative method is to treat the nuisance parameters in the same footing as the theory parameters. Then their errors can be determined using an effective Hessian matrix, which is given by $N \times N$-submatrix of the inverse of the full $(N + M) \times (N + M)$ Hessian matrix [78].

To prove this statement, let $H^p_{\mu\nu}$ be the second derivative of $\chi^2_p(a)$ with respect to the theory parameters a_μ and a_ν, where $\mu, \nu = 1, ..., N$. Let $H_{\mu i}$, $H_{\mu\nu}$, and H_{ij} be the second deriative of χ^2 with respect to a_μ and r_i, a_μ and r_i and r_j. Here, $i = 1, ..., M$. By implicit differentiation, the Hessian $H^p_{\mu\nu}$ can be written as

$$H^p_{\mu\nu} = H_{\mu\nu} + H_{\mu i}\frac{\partial \hat{r}}{\partial a_\nu} \tag{3.49}$$

where $\hat{r}(a) = \arg\min_r \chi^2(a, r)$. The derivative $\partial\hat{r}/\partial a_\nu$ evaluated at any a is hard to be calculated as the explicit function $\hat{r}(a)$ is unknown. However, for $a = \hat{a} = \arg\min_a \chi^2_p(a)$, we can express the derivative as

$$\frac{\partial \hat{r}_i(\hat{a})}{\partial a_\nu} = -{H_r}^{-1}{}_{ij} H_{j\nu} \tag{3.50}$$

where H_r is an $M \times M$ matrix whose components are the same as H_{ij}. Note that all the Hessian matrices on the RHS are evaluated at the minimum $\hat{a}$. Inserting this to (3.49), we obtain

$$H^p_{\mu\nu} = H_{\mu\nu} - H_{\mu i}{H_r}^{-1}{}_{ij} H_{i\nu} \tag{3.51}$$

For any block matrix :

$$P = \begin{pmatrix} A & B \\ C & D \end{pmatrix} \tag{3.52}$$

with A, B, C, D are $N \times N$, $N \times M$, $M \times N$, and $M \times M$ matrices, the first (upper left) $N \times N$ component of P^{-1} is given by $(A - BD^{-1}C)^{-1}$. Therefore, one immediately see that

$$H^{p-1} = H^{-1}\Big|_{N\times N}, \tag{3.53}$$

and thus the statement was proven.

3.2.2 Justification of Hessian Method under Linear Approximation

In the previous section, we show how the asymptotic properties of MLE lead to the Hessian matrix as the inverse of the covariance matrix. In this section, we will show how a sufficiently linear theory/model naturally leads to a normally distributed fitted parameters with covariance matrix given by the inverse of the Hessian matrix. We will also discuss cases in which ´the predictions from Hessian formalism are inaccurate. In this section, we assume that the cost function takes the following form

$$\chi^2(a) = (D - T)^T C^{-1} (D - T) \tag{3.54}$$

For a general case with normalization uncertainties, the cost function (3.19) can be shown to take the form of (3.54) as the normalization fluctuation r can be regarded as a theory parameter which is fitted to the data.

Let a^0 be some reference theory parameters. In our context, this can be the true theory parameter (if the model is correct), or it can be the average of the fitted parameters obtained by repeating the data acquisitions. Then, assuming that the theory prediction is linear in the vicinity of a^0

$$T_i(a) \approx T_i(a^0) + T'_{i\mu}(a^0)(a - a^0)_\mu \tag{3.55}$$

where $T'_{i\mu} = \partial T_i / \partial a_\mu$. Inserting this approximation to (3.54), one can show that the cost function is minimized at

$$\hat{a}_\mu \equiv \arg\min_a \chi^2(a, D) = a^0 + \sum_\nu H^{-1}_{\mu\nu} d^0_\nu \tag{3.56}$$

where

$$H_{\mu\nu} = \sum_{i,j} T'_{i\mu}(a^0) C^{-1}_{i,j} T'_{j\nu}(a^0) = \frac{1}{2} \frac{\partial^2 \chi^2(a^0)}{\partial a_\mu \partial a_\nu} \tag{3.57}$$

$$d_\mu = \sum_{i,j} (D - T(a^0))_i C^{-1}_{ij} T'_{j\mu}(a^0) \tag{3.58}$$

Inserting (3.56) into (3.54), one find the χ^2 at the minimum as

$$\chi^2(\hat{a}) = (D - T(a^0))^T C^{-1}(D - T(a^0)) - d^T H^{-1} d \tag{3.59}$$

To obtain the distribution of the fitted parameters $\hat{a}$ and the $\chi^2(\hat{a})$, we assume that the data D_i is normally distributed with mean $\langle D_i \rangle$ and covariance matrix $\langle (D_i - \langle D_i \rangle)(D_j - \langle D_j \rangle) \rangle = C_{ij}^{true}$. Given the data distribution, it is easy to show that the fitted parameters are also normally distributed with

$$\langle \hat{a}_\mu \rangle = a^0 + \sum_\nu H^{-1}_{\mu\nu} \langle d_\nu \rangle \tag{3.60}$$

$$\hat{a}_\mu - \langle \hat{a}_\mu \rangle = \sum_\nu H^{-1}_{\mu\nu} \, \Delta d_\nu \tag{3.61}$$

$$\langle (\hat{a}_\mu - \langle \hat{a}_\mu \rangle)(\hat{a}_\nu - \langle \hat{a}_\nu \rangle) \rangle = \sum_{\lambda\gamma} H^{-1}_{\mu\lambda} \langle \Delta d_\lambda \Delta d_\gamma \rangle H^{-1}_{\nu\gamma} \tag{3.62}$$

where

$$\Delta d_\mu \equiv d^0_\mu - \langle d^0_\mu \rangle = \sum_{i,j} (D_i - \langle D_i \rangle) C^{-1}_{ij} T'_{j\mu}(a^0) \tag{3.63}$$

$$\langle \Delta d_\mu \Delta d_\nu \rangle = \sum_{i,j,k,l} T'_{i\mu} C^{-1}_{ij} C^{true}_{jk} C^{-1}_{kl} T'_{l\nu} \tag{3.64}$$

Having derived the distribution of the fitted parameters, we can now proceed further to determine the distribution of the minimum of the cost function. From (3.59), as D and d^0 are normally distributed, then it follows that $\chi^2(\hat{a})$ is distributed according to χ^2 distribution. Here, we are interested in the mean and the variance of the distribution. First, we rewrite (3.59) as

$$\chi^2(\hat{a}) = (D - T(a^0)) K^{-1} (D - T(a^0)) \tag{3.65}$$

where

$$K^{-1}_{ij} = C^{-1}_{ij} - \sum_{k,l,\mu,\nu} C^{-1}_{ik} T'_{k\mu} H^{-1}_{\mu\nu} T'_{l\nu} C^{-1}_{lj} \tag{3.66}$$

Writing $D - T(a^0) = D - \langle D \rangle + \langle D \rangle - T(a^0)$ and using the fact that $\langle (D - \langle D \rangle) \rangle = 0$, we find

$$\langle \chi^2(\hat{a}) \rangle = \mathrm{Tr}\left(K^{-1} C^{true} \right) + \left(\langle D \rangle - T(a^0) \right) K^{-1} \left(\langle D \rangle - T(a^0) \right) \tag{3.67}$$

Similarly, we can also calculate the variance of $\chi^2(\hat{a})$. They are given by :

$$\mathrm{Var}\left(\chi^2(\hat{a}) \right) = 2\mathrm{Tr}\left(K^{-1} C^{true} K^{-1} C^{true} \right) + 4 (\langle D \rangle - T(a^0))^T K^{-1} C^{true} K^{-1} (\langle D \rangle - T(a^0)) \tag{3.68}$$

It is possible to slightly simplify the formula (3.68) and (3.67) by choosing

$$a^0 = \langle \hat{a} \rangle \tag{3.69}$$

Eqs. (3.60) then implies

$$H^{-1}\langle d\rangle = 0 \tag{3.70}$$

This leads to

$$K^{-1}\left(\langle D\rangle - T(a^0)\right) = C^{-1}\left(\langle D\rangle - T(a^0)\right) \tag{3.71}$$

We can then rewrite

$$\langle\chi^2(\hat{a})\rangle = \mathrm{Tr}\left(K^{-1}C^{true}\right) + \left(\langle D\rangle - T(a^0)\right)C^{-1}\left(\langle D\rangle - T(a^0)\right) \tag{3.72}$$

$$\mathrm{Var}\left(\chi^2(\hat{a})\right) = 2\mathrm{Tr}\left(K^{-1}C^{true}K^{-1}C^{true}\right) + 4(\langle D\rangle - T(a^0))^T C^{-1}C^{true}C^{-1}(\langle D\rangle - T(a^0)) \tag{3.73}$$

The advantage of choosing $a^0 = \langle\hat{a}\rangle$ is now apparent. As C and C^{true} are positive definite, the second term in the RHS of (3.72) and (3.73) are always positive, and zero if $T(a^0) = \langle D\rangle$. Thus, model misspecification *always* leads to an increase of $\langle\chi^2(\hat{a})\rangle$ and $\mathrm{Var}\left(\chi^2(\hat{a})\right)$. In contrast, if the uncertainties of the data are misestimated, then $C \neq C^{true}$, the first term in RHS of (3.72) and (3.73) can be smaller or larger from the ideal values.

Having derived the distribution of the $\hat{a}$ and $\chi^2(\hat{a})$ for a general case without assuming the correctness of the model and the data uncertainties, we can now discuss specific cases where either of the theory or the uncertainties are correct. We say that theory prediction $T_i(a)$ is *correctly specified* if there exist a set of parameters, $\bar{a}$, such that $\langle D_i\rangle = T_i(\bar{a})$ for all data points i. Using this definition, one can say that in the case of tensions the theory is misspecified, as there is no $\bar{a}$ that satisfy $\langle D_i\rangle = T_i(\tilde{a})$ for all data points i. We say that the data uncertainties are *correctly estimated* if $C_{ij} = C_{ij}^{true}$. We discuss further the distribution of the fitted parameters and the cost function at the minimum for each case in the following.

Correct Theory and Uncertainties

If the theory is correct, then it is easy to show that the estimator is unbiased, which means $\langle\hat{a}\rangle = \bar{a}$. As explained above, as consequence of choosing $a^0 = \langle a\rangle$, we have $H^{-1}\langle d\rangle = 0$. If the theory is correct, then

$$\langle d\rangle = T'^T C^{-1}\left(T(\tilde{a}) - T(\langle a\rangle)\right) = T'^T C^{-1}T'((\tilde{a} - \langle a\rangle) = 0$$

Therefore, $H^{-1}\langle d\rangle = 0$ implies $\bar{a} = \langle a\rangle$. This proves our claim. We note here that we did not make any assumption about the correctness of the data uncertainties. Irrespective of the correctness of the data uncertainties, the estimator is always unbiased.

If, additionally, the data uncertainties are correct, then by definition $C^{true} = C$. This further implies

$$\mathrm{Tr}\left(K^{-1}C^{true}\right) = \mathrm{Tr}\left((K^{-1}C^{true})^2\right) = N_D - N_p \tag{3.74}$$

Using the choice $a^0 = \tilde{a}$, we have $\langle d_\mu^0 \rangle = 0$. Furthermore, $H = H^0$. Thus, the fitted parameters are normally distributed with

$$\langle \hat{a}_\mu \rangle = \bar{a}_\mu \tag{3.75}$$

$$\mathrm{Cov}(\hat{a}_\mu, \hat{a}_\nu) = H^{-1}_{\mu\nu} \tag{3.76}$$

Thus, the covariance matrix of the fitted parameters is given precisely by the inverse Hessian matrix. Therefore, it amounts to using a χ^2 tolerance of $T^2 = 1$ at 1σ.

The cost function at the minimum is distributed according to χ^2-distribution with $N_D - N_p$ degrees of freedom, such that :

$$\langle \chi^2(\hat{a}) \rangle = N_D - N_p \tag{3.77a}$$

$$\mathrm{Var}(\chi^2(\hat{a})) = 2(N_D - N_p) \tag{3.77b}$$

Correct Theory and Incorrect Uncertanties

Besides the central value, experimental data usually also specify the statistical as well as systematical uncertainties. However, it is often the case that the individual correlated systematic uncertainty is not provided. Instead, the systematic errors are just the total correlated and uncorrelated ones added in quadrature. Furthermore, the uncertainties could be overestimated or underestimated. Therefore, the reported uncertainties can be different than the actual data uncertainties that specify the data distribution. To highlight the departure from an ideal case, we write the true data covariance matrix as

$$C^{true} = C + \Delta C \tag{3.78}$$

As C^{true} is unknown, so is ΔC. Let $\bar{a}$ be the correct (true) PDF parameters, satisfying $\langle D_i \rangle = T_i(\bar{a})$ for all i (the model is correctly specified). Then as before, we can set $a^0 = \bar{a}$. The mean and the covariance of the fitted parameters are

$$\langle \hat{a}_\mu \rangle = \bar{a}_\mu \tag{3.79}$$

$$\mathrm{Cov}(\hat{a}_\mu, \hat{a}_\nu) = \sum_\alpha H^{-1}_{\mu\alpha}(\delta_{\alpha\nu} + \Delta_{\alpha\nu}) \tag{3.80}$$

with

$$\Delta_{\alpha\nu} = \sum_{\sigma,\lambda} H^{-1}_{\alpha\sigma} \left[\sum_{i,j,k,l} T'_{i\sigma} C^{-1}_{ij} \Delta C_{jk} C^{-1}_{kl} T'_{l\lambda} \right] H^{-1}_{\lambda\nu} \tag{3.81}$$

Thus, the estimator is still unbiased, but the covariance matrix is now modified. This means, if one uses the plain Hessian method, one could easily underestimate or overestimate the uncertainty of fitted parameters. The expectation value of the cost function at the minimum is given

by

$$\langle \chi^2(\hat{a}) \rangle = N_D - N_p - \mathrm{Tr}\left(C^{-1} T' {H^0}^{-1} T'^T C^{-1} \Delta C \right) \tag{3.82}$$

If ΔC is positive definite, then we can see that $\langle \chi^2(\hat{a}) \rangle \leq (N_D - N_p)$. However, in general ΔC does not have a specific definitiness, so $\langle \chi^2(\hat{a}) \rangle$ can be larger than $N_D - N_p$ as well.

Incorrect Theory and Correct Uncertainties

When the uncertainties are correct, then $C^{true} = C$. Let

$$R = \langle D \rangle - T\left(\langle \hat{a} \rangle \right) \tag{3.83}$$

Then

$$\langle \chi^2(\hat{a}) \rangle = N_D - N_p R^T C^{-1} R \tag{3.84}$$

$$\mathrm{Var}\left(\chi^2(\hat{a}) \right) = 2(N_D - N_p) + 4RC^{-1}R \tag{3.85}$$

We can see that the mean and variance are always larger than those that correspond to the ideal case. We remark here, while the mean is larger, this does not mean the χ^2 at the minimum is larger than $N_D - N_p$. In fact, in an overfitting case, $\chi^2(\hat{a})$ could be smaller than $N_D - N_p$ and conversely for the underfitting case. It is just when the whole experiment and parameter fitting are repeated many times, the mean is always larger than the ideal case $N_D - N_p$.

Turning to the distribution of the fitted parameters, we can see that the covariance matrix is the same as in the ideal case :

$$\mathrm{Cov}(\hat{a}_\mu, \hat{a}_\nu) = H^{-1}_{\mu\nu} \tag{3.86}$$

Thus, the Hessian method with tolerance $T^2 = 1$ still applies in this case. Note that this also includes the case of tensions between data sets.

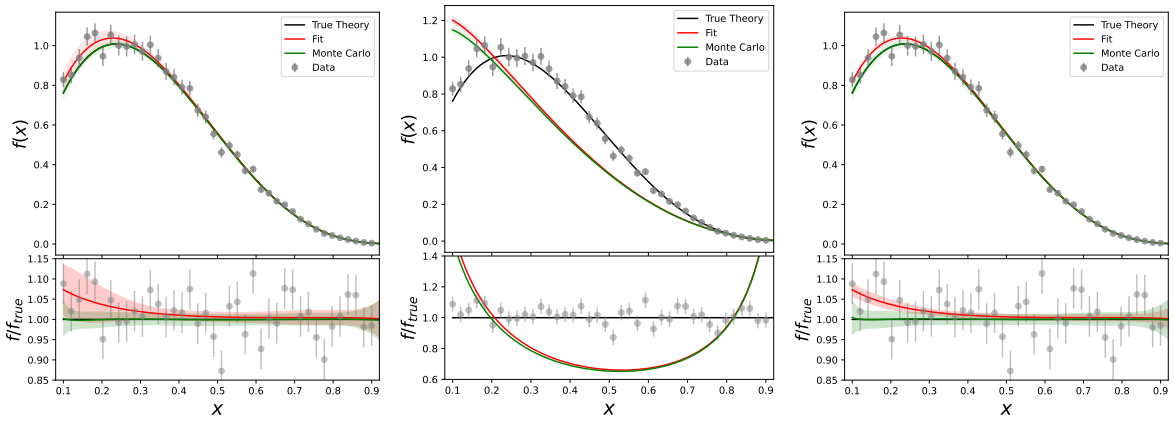
Numerical Experiment

All derivations presented above assume that the theory predictions are sufficiently linear near the reference parameter a^0. In practice, such linearity condition applies quite generally. To see how well the Hessian error method in non-linear curve fittings, we perform fits with the data generated from the following function :

$$f_{true}(x) = N\, x^{a_1}(1-x)^{a_2}(1 + a_3\sqrt{x} + a_4 x + a_5 x^{3/2}) \tag{3.87}$$

The data is generated according to :

$$f_{data}(x_i) = f_{true}(x_i) + \sigma_i r_i \tag{3.88}$$



(a) (b) (c)

FIGURE 3.2: Data-theory comparison for fits with artifical data. The black lines shows the true theory functions use to generate the data. The green lines shows the results from Monte Carlo fit (see text for more detail). The red lines and bands show the results of theory fits and their Hessian errors. (a) shows data-theory with $f_{model} = f_{true}$. (b) Shows data-theory with f_{model} given by Eqs. (3.89). (c) shows data-theory with $f_{model} = f_{true}$ and the uncertainty of the data is artifically shrinked by a factor of 2.

where $\sigma_i = 0.05 f_{true}(x_i)$ is the statistical uncertainty for the i-th data point and $r_i \sim \mathcal{N}(0,1)$ is a standard normal random number. Note that f_{true} defined here is similar to the CJ15 parameterization for u_v PDF, therefore the fit that we are doing here mimicks a real PDF fit.

After we generate 40 data points, we fit a model function $f_{model}(x)$, which is identical to f_{true} to the data (let's call this *fit-a*). Thus, the model is correctly specified. We obtain the final $\chi^2 = 40.28$, which correspond to $\chi^2/dof = 1.14$, where $dof = N_D - N_p = 35$. In Fig. 3.2(a), we show the data-theory comparison and also the Hessian errors (with $T^2 = 1$) of the fitted model. In Fig. 3.2(a), we also show results from 1000 Monte Carlo (MC) fits. For each Monte Carlo fit, we generate the data using different sample of r_i and redo the fitting process. The green line in Fig. 3.2(a) shows the average of the fitted model and the green band show the standard deviation. Thus, essentially, Monte Carlo fits is just sampling the distribution of the fitted parameters and thus the resulting model errors are representative of the actual errors propagated from the data uncertainties. The mean and variance of χ^2 in the MC fits are $\langle \chi^2 \rangle = 34.82$ and $Var(\chi^2) = 71.12$ respectively. These values are very close to the expected values from (3.77) : $\langle \chi^2 \rangle = 35$ and $Var(\chi^2) = 70$.

To see the impact of model misspecification on the uncertainty of the fitted parameters, perform a second fit (let's call it *fit-b*), where now :

$$f_{model}(x) = N\, x^{a_1}(1-x)^{a_2} \tag{3.89}$$

Thus, the model is now severely misspecified. The final $\chi^2/N = 34.26$, much larger than the ideal case. In Fig. 3.2(b), data-theory from this fit is shown. In Fig. 3.2(b), we also show the

results from MC fits. One can see that the fitted theory is very similar to MC one. Furthermore, the Hessian uncertainty band (with $T^2 = 1$) is also similar to the errors from MC fits.

In Fig. 3.2(c), we show data-theory from fit with correctly specified model $f_{model} = f_{true}$, but the uncertainty is shrunk by a factor of 2. As expected, the Hessian error is now shrunk by the same factor compared to the MC uncertainties.

To summarize, we have shown that for sufficiently linear model and correctly estimated data uncertainties, the Hessian formalism with $\Delta\chi^2 = 1$ is still valid, even if the model is severely misspecified[2]. If the data uncertainties is misestimated, then the Hessian errors with $\Delta\chi^2 = 1$ are not valid and can lead to overestimated/underestimated uncertainties. ´

3.2.3 Error Estimation : Replica Method

The hessian method relies on the linearity of the model with respect to the theory parameters. When the theory is not linear, then the Hessian estimates can be inaccurate. The replica method is a way to estimate the uncertainties of the fitted parameters without assuming the theory to be linear. The idea of the replica method is similar to the Monte Carlo method. Namely, one samples the data generation distribution and repeats the parameter estimation many times. As the data generating distribution is unknown, it can be approximated as the probability distribution from the observed data. Thus, if the data is Gaussian, then the mean of the data generating distribution is just the observed data, with the covariance matrix given by the one of the observed data. One should remember that the reported data is not the true value of the measured observable, but rather a sample of the data generating distribution with an unknown true value, which is hopefully close to the measured data. Thus, generating replica from the data will cause some bias. However, as the data should be randomly distributed around the true value, the bias is hopefully small and thus, fitting a replica is equivalent to sampling the distribution of the fitted parameters.

As said, one generates a replica $\tilde{D}$ of the original data D according to

$$\tilde{D}_i = (1 + \sigma_{norm} r_{norm}) \left(D_i + \sigma_i r_i + \sum_\alpha \bar{\sigma}_{i\alpha} \bar{r}_\alpha \right) \tag{3.90}$$

where σ_{norm}, σ_i and $\bar{\sigma}_{i\alpha}$ are the normalization, statistical and correlated systematical uncertainties for the i-th data point and from α-th source of correlated systematics. D_i represent the original i-th data point. All fluctuations r_{norm}, r_i and $\bar{r}_\alpha$ are sampled from a standard normal distribution. Let's forget normalization uncertainty for the moment, then (3.90) implies that the replica is Gaussianly distributed with mean D_i and the same covariance matrix as the original data :

$$C_{ij} = \sigma_i^2 \delta_{ij} + \sum_{alpha} \bar{\sigma}_{i\alpha} \bar{\sigma}_{j\alpha} \tag{3.91}$$

[2]Note that the presence of tensions between data sets also means that the model is incorrect/misspecified.

Thus,

$$\langle \tilde{D}_i \rangle_{rep} = D_i \tag{3.92}$$

$$\langle (\tilde{D} - T)_i (\tilde{D}^k - T)_k \rangle_{rep} = C_{ij} \tag{3.93}$$

To fit nPDFs, we minimize the standard χ^2 function for each replica :

$$\chi^2_{\tilde{D}} = (\tilde{D} - T)^t C^{-1} (\tilde{D} - T) \tag{3.94}$$

If, for each replica, a set $\hat{\tilde{a}}$ of optimized theory parameters is obtained, then the sample mean of such $\hat{\tilde{a}}$ for all replicas then provides the best estimate for a. Furthermore, the sample covariance of $\tilde{a}$ approximates the true covariance of the fitted parameters. Therefore it can be used to estimate the PDF uncertainties. In the linear approximation, it is possible to analytically derive the statistical properties of the replicas. Let a^0 be some reference parameter as discussed in the previous section. We can write the linearized theory $T_i(a)$ for a that is close to the reference parameter a^0 as :

$$T_i(a) = T_i(a^0) + \sum_\mu T'_{i\mu} (a - a^0)_\mu \tag{3.95}$$

where $T'_{i\mu} = \partial T_i(a) / \partial a_\mu |_{a^0}$. At $\hat{\tilde{a}}$, the χ^2 satisfies $\partial \chi^2_{\tilde{D}}(a) / \partial a_\mu$ for all μ. This implies

$$\sum_\nu H_{\mu\nu} (\hat{\tilde{a}}_\mu - a^0_\mu) = \tilde{d}_\mu \tag{3.96}$$

where

$$H_{\mu\nu} = \sum_{i,j} T'_{i\mu} C^{-1}_{ij} T'_{j\nu} \approx \frac{1}{2} \frac{\partial^2 \chi^2_D(a)}{\partial a_\mu \partial a_\nu} \bigg|_{a^0} \tag{3.97}$$

$$\tilde{d}_\mu = \sum_{i,j} T'_{i\mu} C^{-1}_{ij} (\tilde{D} - T(a^0))_j \tag{3.98}$$

Note that H here approximate the Hessian matrix and will become exact if $T(a)$ is exactly linear in a. The fitted parameters $\hat{a}$ is then given by

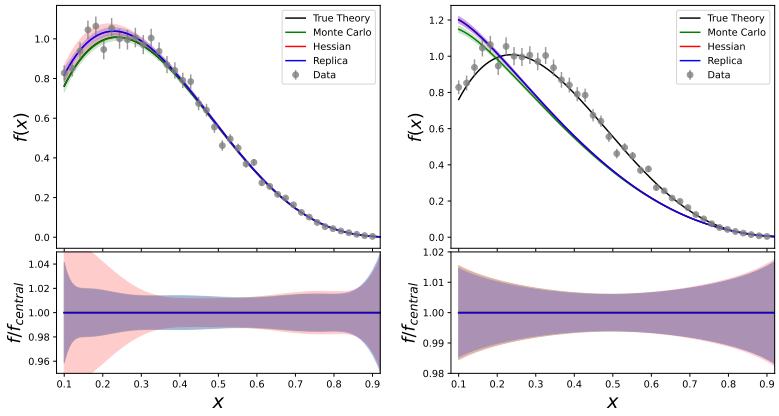
$$\hat{\tilde{a}}_\mu (\tilde{D}) = a^0_\mu + \sum_\nu H^{-1}_{\mu\nu} \tilde{d}_\nu \tag{3.99}$$

As $\tilde{D}$ is Gaussian, then $\hat{\tilde{a}}$ is also so. The mean and covariance of the fitted parameters $\hat{\tilde{a}}$ are therefore given by

$$\langle \hat{\tilde{a}} \rangle_{rep} = a^0_\mu + \sum_\nu H^{-1}_{\mu\nu} d_\mu = \hat{a}_\mu \tag{3.100}$$

$$\left\langle \left(\hat{\tilde{a}} - \langle \hat{\tilde{a}} \rangle \right)_\mu \left(\hat{\tilde{a}} - \langle \hat{\tilde{a}} \rangle \right)_\nu \right\rangle_{rep} = H^{-1}_{\mu\nu} \tag{3.101}$$

where d_μ is the same as (3.98). The above equations shows that under replica variation, the



(a) (b)

FIGURE 3.3: Data-theory comparison for fits with artifical data. The black lines shows the true theory functions use to generate the data, the green, red,blue lines and bands show the results from Monte Carlo, Hessian, and replica methods. (a) shows data-theory with $f_{model} = f_{true}$. (b) Shows data-theory with f_{model} given by Eqs. (3.89).

mean of the fitted parameters is the same as the ones from a fit with the original data, with covariance matrix given by the inverse of the Hessian matrix. This provides an equivalence between both Monte Carlo and the Hessian approach.

It is interesting how the minimum $\chi^2_{min} \equiv \chi^2(\hat{\tilde{a}})$ is distributed in this linear approximation. From (3.94), (3.95), (3.99), and (3.97), we then have

$$\chi^2(\hat{\tilde{a}})_{\tilde{D}} = (\tilde{D} - T(a^0))C^{-1}(\tilde{D} - T(a^0)) - \tilde{d}_\mu H^{-1}_{\mu\nu}\tilde{d}_\nu \tag{3.102}$$

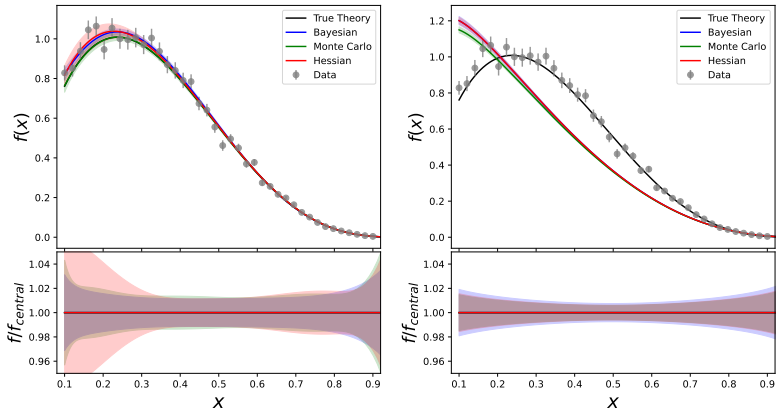
In then straightforward to prove that

$$\langle \chi^2_{\tilde{D}}(\hat{\tilde{a}})\rangle_{rep} = \chi^2_D(\hat{a}) + N_D - N_p \tag{3.103}$$

where $\chi^2_D(\hat{a}^0)$ is the minimum of the χ^2 of the original data. This is rather surprising, as we naively expect that $\langle \chi^2_{\tilde{D}}(\hat{\tilde{a}})\rangle_{rep} = N_D - N_p$. (3.103) basically says that the replica average of the $\chi^2_{\tilde{D}}$ is just $N_D - N_p$ plus the bias term χ^2_D as we generate the data replica from the original data D, instead of the true theory function. For a good fit, we expect $\chi^2_D(\hat{a}) \sim N_D - N_p$, thus $\langle \chi^2_{\tilde{D}}(\hat{\tilde{a}})\rangle_{rep}/(N_D - N_p) \sim 2$. Thus, the expected χ^2/dof for each replica fit is around 2.

In practice, when running many replica fits, some of the fits converge to local minima or even do not converge at all. As a result, the replica mean of the χ^2 at minimum is larger than the expected value (3.103). Thus, (3.103) can be used to check if the replica method works as expected.

To see how the errors from replica method compared to the ones from Hessian (with $T^2 = 1$) and from Monte Carlo sampling, we repeat *fit-a* and *fit-b* in section 3.2.2, but now we also evaluate the uncertainty of the fitted parameters using replica method with 1000 replicas. For



(a) (b)

FIGURE 3.4: Data-theory comparison for fits with an artifical data. The black lines shows the true theory functions use to generate the data, the green, red,blue lines and bands show the results from Monte Carlo, Hessian, and Bayesian methods. (a) shows data-theory with $f_{model} = f_{true}$. (b) Shows data-theory with f_{model} given by Eqs. (3.89).

fit-a and *fit-b*, we obtain $\langle \chi^2_D \rangle_{rep}$ of 75.64 and 1336.30 respectively. Given the χ^2_D values for *fit-a* and *fit-b* are 40.28 and 1301.85, and the $N_D - N_p$ values are 35 and 38 respectively, the values of $\langle \chi^2_D \rangle_{rep}$ are in excellent agreement with the expected ones from (3.103) : 75.28 for *fit-a* and 1339.85 for *fit-b*. In the top panels of Fig. 3.3, we show the comparison between the results from different methods of error estimation. We can see that the central curves from the replica method (blue lines) perfectly coincide with the ones from Hessian (red lines), as expected from (3.100). In the bottom panel, we show the ratio of the errors to the central values. We can see that all these three methods lead to similar uncertainty bands, even though we have a severe model misspecification in *fit-b*. The Hessian uncertainty band in the low x region in *fit-a* is significantly larger than the one from Monte Carlo and Replica method. The likely explanation for this is that the linear approximation breaks down for the model and thus, the Hessian method gives an inaccurate results.

3.2.4 Error Estimation : Bayesian Approach

Estimating the uncertainty of the fitted PDFs involves an implicit assumption that the resulting uncertainty of the PDFs should match the spread of fits from repeated measurements. An alternative to this frequentist method is the Bayesian approach. In this method, the distribution of the PDFs, given the data D, is obtainable via Bayes theorem, $p(f|D) \propto p(D|f)p(f)$, where $p(D|f)$ is the likelihood and $p(f)$ is the prior distribution. Usually, in the absence of prior information, a uniform distribution is taken to be the prior, then the posterior PDF distribution is just the likelihood. As shown in section 3.1, by construction, the χ^2 function is proportional to log-likelihood. Therefore, with a uniform distribution as the prior, the posterior distribution

is given by

$$p(f|D) \propto \exp\left(-\frac{\chi^2(f,D)}{2}\right) \tag{3.104}$$

PDF determination using this approach can be done by sampling the posterior distribution. From a large number of samples, one can then obtain the mean as a point estimate for the PDF parameters. Alternatively, one can construct a credibility interval, within which some percentage of the samples fall in. As another remark, one can relate this method to the maximum likelihood approach if the χ^2 function is symmetric around the minimum. In this case, the best (point) estimate for the PDF parameters is given by the global maximum of the likelihood function, or equivalently, the minimum of $\chi^2(a, D)$.

In the context of PDF fitting, the most considerable appeal of this approach lies in its capability to reliably estimate the inferred parameters without assuming linearized theory prediction as in the Hessian approach or requiring multiple fits of data replicas, which are prone to get stuck in local minima, as in the MC replica method. The downside of this method is the same as in the MC replica method. Namely it requires a rather high computational cost.

If the theory is linear and symmetric enough near the best estimate of the parameters, one can actually show that this method is equivalent to the Hessian one. Let $\langle a \rangle$ be the best estimate for PDF parameters. Then one expand the χ^2 function near $\langle a \rangle$ as

$$\chi^2(a, D) = \chi^2(\langle a \rangle) + \frac{1}{2}\sum_{\mu\nu} \frac{\partial^2 \chi^2}{\partial a_\mu \partial a_\nu}(a - \langle a \rangle)_\mu (a - \langle a \rangle)_\nu \tag{3.105}$$

$$= \chi^2(\langle a \rangle) + z^T z, \tag{3.106}$$

Thus, from (3.104), one can immediately see that the posterior distribution is Gaussian, centered around $\langle a \rangle$, with covariance matrix given by the inverse of the Hessian. This shows that Hessian, Bayesian and MC replica methods are equivalent for sufficiently linear theory functions.

In Fig. 3.4, we compare the uncertainty estimates for the model fitted to the artificial data as discussed in the previous sections. To obtain the parameter errors from Bayesian method, we sample 1000 samples of (3.104) using Markov Chain Monte Carlo with Metropolis-Hasting algorithm. To obtain 1000 samples, we collected 10000 accepted samples from the algorithm, and discarded the first 9000 ones to account the burn-in period. From Fig. 3.4, we can see that central values from the Bayesian method are identical to those from the Hessian method. Looking at the error estimates, one can see that the errors from all these three methods are similar, even when the model is misspecified as shown in Fig.3.4(b).

3.2.5 Hypothesis-Testing vs Frequenstist Uncertainty

Previously, it has been shown that the Hessian method is equivalent to both replica and Bayesian approach, provided that the χ^2 function is sufficiently linear around the average of the fitted parameters. Note that while the linearity condition seems restricting, in practice, this condition

is quite general and applies in many cases. In section 3.2.1, it has also been shown that the tensions between data sets and model misspecification do not influence the validity of the Hessian method (with $\Delta\chi^2 = 1$). In this case, enlarging the uncertainty by using global or dynamic tolerance is not needed.

One should note that the Hessian and Monte Carlo methods that we discussed previously give estimate the uncertainty of the PDF parameter in frequentist way. Namely, the uncertainty is exactly the standard deviation when the experiments are repeated many times, and the resulting data sets are used to fit the PDFs using exactly the same methodology. Thus, the uncertainty estimation does not care whether the data sets agree with each other or if the theory is correct. It is just a propagation of uncertainty from a set D of random variables to another $\hat{a}(D)$. For future reference, we will call this type of uncertainty as frequentist uncertainty[3]. The main concern of using this type of uncertainty is that the resulting error bands for the theory can be quite far away from the true theory that generates the data in the case of a severe model misspecification. If the tensions between data sets are large, the error bands may not even reproduce all the data. Thus, the estimation of the uncertainty is as good (in the sense of covering the true theory and explaining the data) as the data and the model used in the fit.

In order to enforce the error bands to explain all the data (in the hypothesis testing sense), the frequentist uncertainty is artificially enlarged by choosing $T^2 \geq 1$. the global and dynamic tolerance methods fall in this category. The resulting uncertainty is therefore called hypothesis-testing uncertainty. As the tolerance $T^2 \geq 1$, this uncertainty is always larger than the frequentist one. This uncertainty does not reflect the probability of the fitted parameters, so its propagation (to observable errors) does not have probabilistic interpretation. It also does not reduce to the frequentist uncertainty even if tensions are absent and the theory is perfect. The dramatic difference between frequentist and hypothesis testing uncertainty can be illustrated as follows: Let D_1 represent the data from an experiment E, with N data points. Suppose that the experiment is later repeated with the same set-up, and a new data D_2 is obtained. If one performs a model fitting to D_1 (let's call it fit1) and to both D_1 and D_2 (fit2), then frequentist uncertainty of the fit2 is shrunk by a factor of $1/\sqrt{2}$. In contrast, the hypothesis-uncertainty will not shrink at all. This illustrate that type of uncertainty does not respond well with the addition of new data.

Generally speaking, when we speak about the uncertainty of the fitted parameters, the frequentist uncertainty should be used, no matter if the model is misspecified (which includes the presence of tensions between data sets) or not. This means, if one uses the Hessian method, the tolerance should be set to $T^2 = 1$ for 1σ deviation. However, in PDF fittings, it is desirable that the PDF uncertainties can explain the data. Neglecting this requirement can make some data to have large χ^2, which implies (assuming the data and the uncertainties are correct), for example, breakdown of the factorization theorem, which is a conclusion that we do not want to take easily.

[3] In PDF fitting literature, this type of uncertainty is often referred to as the uncertainty in the parameter estimation sense.

Regardless, it is always useful to have hypothesis-testing uncertainty when comparing a theory with a data. This uncertainty band in this case represents the regions where the theory curves can still explain the data used in the fit. Comparing data S and theory in this case is equivalent to assess the compatibility of the data S with all the data used in the fit.

3.3 Assessing the Impact of New Data: Reweighting Technique

Reweighting, in the context of PDF fitting, is a technique to obtain a new set of PDFs by including new data that was not included in the original PDFs. This technique is very useful when one does not have the access to the fitting code used in the original PDFs, or the theory predictions for the new data simply take too long to compute, making the new data impractical to be included in the original fit. By comparing the new set of PDFs with the original ones, one can study the impact of the new data in PDF extraction.

Reweighting is originally formulated in the Bayesian framework[81]. The core concept is Bayes theorem : $p(f|D) = \mathcal{N}(D)\, p(D|f)p(f)$, where $p(f|D)$ represents the posterior distribution of the PDFs, generically denoted as f, after seeing the new data D. $p(D|f)$ denotes the likelihood and $p(f)$ is the prior distribution, which is usually given by the PDF distribution of the original fit (without the new data D). $\mathcal{N}(D)$ here is a normalization factor, such that $p(f|D)$ is properly normalized. The normalization is then given by

$$\mathcal{N}(D) = \left(\int \mathcal{D}f\, p(f|D)p(f) \right)^{-1} . \tag{3.107}$$

Here, $\mathcal{D}f'$ is the functional integration measure. Now, by construction, the loss function χ^2 is $\chi^2(f, D) = -2 \log p(D|f)$, hence

$$p(D|f) = \exp\left[-\frac{1}{2}\chi^2(f, D) \right] . \tag{3.108}$$

Given the posterior distribution $p(f|D)$, one can obtain the expectation value and variance of an observable $Q(f)$ as

$$\langle Q \rangle = \mathcal{N}(D) \int \mathcal{D}f\, Q(f)p(D|f)p(f)\,, \tag{3.109a}$$

$$\mathrm{Var}(Q) = \mathcal{N}(D) \int \mathcal{D}f\, \left(Q(f) - \langle Q \rangle\right)^2 p(D|f)p(f)\,. \tag{3.109b}$$

The functional integrations in (3.107) and (3.109) can be computed using a Monte Carlo approach. As the prior distribution $p(f)$ is known, one can generate sufficiently large number of samples from $p(f)$, and the integrations are reduce to a simple summation. Thus :

$$\int \mathcal{D}f\, Q(f)\, p(f|D)p(f) = \frac{1}{N_{rep}} \sum_f Q(f)\, \exp\left(-\frac{1}{2}\chi^2(f, D) \right) , \tag{3.110}$$

where N_{rep} is the number of PDF samples. The expectation value and variance of the observable Q can then be approximated as

$$\langle Q \rangle = \sum_f w(f) Q(f) \,, \tag{3.111a}$$

$$\mathrm{Var}(Q) = \sum_f w(f) \left(Q(f) - \langle Q \rangle \right)^2 \,, \tag{3.111b}$$

where the weight $w(f)$ is given by

$$w(f) = \frac{\exp\left(-\chi^2(f, D)/2\right)}{\sum_{f'} \exp\left(-\chi^2(f', D)/2\right)}. \tag{3.112}$$

Note that the weight is normalized to one ;

$$\sum_f w(f) = 1 \,. \tag{3.113}$$

Given the original PDFs $f(x)$, one can the update the PDFs after including the data D using (3.111), with $Q(f) = f$. The uncertainty of the new PDFs can also be determined in the same way.

The reweighting procedure described above requires knowing how to sample the prior distribution. If the distribution of the original PDFs are represented as a collection of PDF replicas (samples) as in NNPDF case, then one can immediately use the replicas to do the reweighting. If the Hessian error PDFs are known, then one needs to devise a way to sample the distribution from the provided error PDFs. The general strategy for sampling PDFs from Hessian error PDFs is as follows. Recall from section 3.2.1 that, written in terms of $z(a)$, the distribution of the fitted parameters is given by

$$p(z) \propto \exp\left(-\frac{1}{2T_{std}^2} z^T z \right) \,. \tag{3.114}$$

Here, as in section 3.2.1, we use a to denote a point in the PDF parameter space, while $z(a)$ is the same point, written in the rescaled eigenvector basis. T_{std} denotes the tolerance that corresponds to 68% percentile. In most modern PDFs, T_{std} is significantly larger than one, to account for data or theory inconsistencies. Given $z(a)$, the value of PDF with flavor i at $a(z)$ can be obtained by a linear expansion :

$$f_i(z) = f(z = 0) + \sum_\mu \frac{\partial f_i}{\partial z_\mu} z_\mu \,. \tag{3.115}$$

The first derivative can be obtained by using finite difference method :

$$\frac{\partial f_i}{\partial z_\mu} = \frac{f(z_\mu^+) - f(z_\mu^-))}{2T_{lhapdf}} \,. \tag{3.116}$$

Here, T_{lhapdf} denotes the value of z_μ used in the LHAPDF distribution. Typically, one use 90% percentile, which correspond to $T = 1.645$. In the case of 68% percentile, then $T = 1.0$. Note that $T_{std} \neq T_{lhapdf}$, as in most case, the given error PDFs in the LHAPDF package is the ones that correspond to the 90% percentile. The procedure to generate a PDF replica is then very simple. First, one sample the D-dimensional z from the normal distribution (3.114). The associated PDF replica is then given by (3.115).

While the Bayesian approach is clean and simple, one still needs to interpret the updated PDFs from PDF fitting point of view. To simplify things a bit, let us assume that the PDFs are parameterized in terms of a finite number of theory parameters a_μ, $\mu = 1, ..., D$ and hence the distribution of the PDFs are completely specified if one knows the distribution of a_μ. It is therefore more convenient to directly work with a_μ and its representation in the (rescaled) eigenvector basis z_μ rather than the infinite-dimensional functional space f. Let's assume that the theory predictions near the original PDFs are sufficiently linear. Let $\chi^2_0(z)$ be χ^2 function that correspond to the data used in the original PDF analysis. Near the minimum ($z = 0$), one has

$$\chi^2_0(z) = \chi^2_0(z = 0) + z^T z \,. \tag{3.117}$$

The prior distribution for the fitted parameters z_μ, $\mu = 1, ..., D$ can then be expressed in term of $\chi^2_0(z)$ as

$$p(z) \propto \exp\left(-\frac{1}{2T^2_{std}} z^T z\right) \propto \exp\left(-\frac{\chi^2_0(z)}{2T^2_{std}}\right) . \tag{3.118}$$

Thus, the posterior distribution is given by

$$p(z|D) \propto \exp\left[-\frac{1}{2T^2_{std}}\left(T^2_{std}\chi^2_D(z) + \chi^2_0(z)\right)\right] , \tag{3.119}$$

where $\chi^2_D = \chi^2(z, D)$ is the χ^2 function for the data D. Defining an effective χ^2_{eff} as

$$\begin{aligned} \chi^2_{eff}(z) &= T^2_{std}\chi^2_D(z) + \chi^2_0(z) \\ &\approx \chi^2_{eff}(z = 0) + \sum_\mu \frac{\partial \chi^2_{eff}}{\partial z_\mu} z_\mu + \sum_{\mu,\nu} \frac{1}{2} z_\mu \frac{\partial^2 \chi^2_{eff}}{\partial z_\mu \partial z_\nu} z_\nu \\ &= \chi^2_{eff}(z_{new}) + (z - z_{new})^T H_{new} (z - z_{new}) \,. \end{aligned} \tag{3.120}$$

Note that all derivatives appearing in the above expressions are evaluated at $z = 0$. Here, z_{new} and H_{new} are given by

$$z_{new} = -\frac{1}{2} H^{-1}_{new,\mu\nu} \frac{\partial \chi^2_{eff}}{\partial z_\nu} = \arg\min_z \chi^2_{eff}(z) \,, \tag{3.121a}$$

$$H_{new,\mu\nu} = \frac{1}{2} \frac{\partial^2 \chi^2_{eff}}{\partial z_\mu \partial z_\nu} \,. \tag{3.121b}$$

If the original PDFs has hessian error PDFs, in the linear approximation, one can calculate H_{new} and $\partial\chi^2_{eff}/\partial z_\mu$ in terms of the error PDFs directly :

$$H_{new,\mu\nu} = \delta_{\mu\nu} + \sum_{i,j} \frac{T_i(z)}{\partial z_\mu} C^{-1}_{ij} \frac{T_i(z)}{\partial z_\mu} = \delta_{\mu\nu} + \sum_{i,j} \frac{T(z^+_\mu) - T(z^-_\mu))}{2T_{lhapdf}} C^{-1}_{ij} \frac{T(z^+_\nu) - T(z^-_\nu))}{2T_{lhapdf}}, \quad (3.122a)$$

$$\frac{\partial\chi^2_{eff}}{\partial z_\mu} = \frac{\partial\chi^2_D}{\partial z_\mu} = \frac{\chi^2_D(z^+_\mu) - \chi^2_D(z^-_\mu))}{2T_{lhapdf}}, \quad (3.122b)$$

where, $T_i(z)$ is the theory prediction for the i-th data point of D, C_{ij} is the covariance matrix of the data D, and $z^\pm_\mu$ is a point in the parameter space that correspond to the μ-th error PDFs. Given z_{new} and H_{new}, The posterior distribution then can be rewritten as

$$p(z|D) \propto \exp\left(-\frac{1}{2T^2_{std}}(z - z_{new})^T H_{new}(z - z_{new})\right). \quad (3.123)$$

This shows, that in the linear approximation (as also assumed in the Hessian error method) :

1. The expectation value of the PDFs after seeing the data D is the same as fitting D together with the other data sets used in the original PDF analysis with loss function given by the $\chi^2_{eff}(z)$ defined in (3.120).

2. The covariance matrix of the PDFs after including the data D using reweighting method is given by $T^2_{std}H^{-1}_{new}$.

3. Reweighting method with the weights given by (3.112) is then completely equivalent with refitting PDFs with the original and the new data, with the loss function given by χ^2_{eff}.

4. The increase of the χ^2_0 before and after reweighting can be estimated as

$$\Delta\chi^2_0 = z^T_{new} z_{new}\,. \quad (3.124)$$

   if $\chi^2_0(z = 0) + \Delta\chi^2_0$ is outside the 95% χ^2-percentile of the χ^2_0 distribution, this indicates that the data D has too strong tensions with the data sets used in the original PDF analysis.

5. To have $\chi^2_{eff} = \chi^2_D + \chi^2_0$, one can modify the likelihood to $p(D|f) = \exp\left(-\chi^2(f, D)/(2T_{std})\right)$ leading to a modified weight :

$$w(f) = \frac{\exp\left(-\chi^2(f, D)/(2T^2_{std})\right)}{\sum_{f'} \exp\left(-\chi^2(f', D)/(2T^2_{std})\right)}. \quad (3.125)$$

   Thus, if one wants the reweighting technique to be equivalent to refitting the PDFs with the new data, one should use this weight.

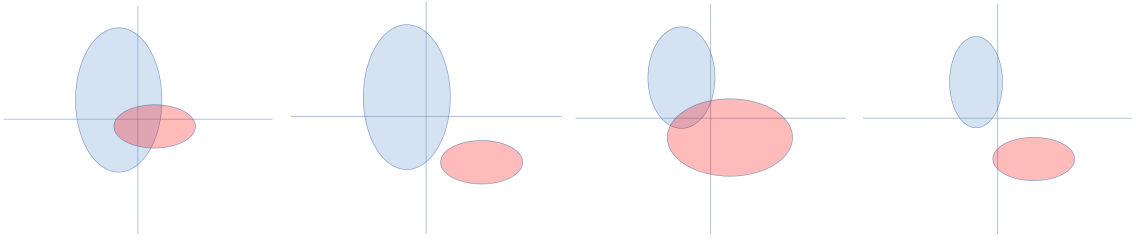


FIGURE 3.5: Illustrations of confidence regions of S-fit (blue) and $\bar{S}$-fit (red). The origin of each figure here represent the minimum of the combined S+$\bar{S}$ fit. (a) shows both S and $\bar{S}$ are mutually compatible. (b) shows that minimum from the combine fit are outside of $\bar{S}$-confidence region. (c) shows that minimum from the combine fit are outside of S-confidence region. (d) shows that both S and $\bar{S}$ are incompatible.

6. If the original PDFs admit Hessian errors PDFs, it is possible to perform a reweighting method without sampling PDF distribution. The so-called Hessian reweighting technique[82] is based on (3.121), with the z_{new} and H_{new} are computed directly from the error PDFs, as shown in (3.122). The PDFs are then updated as

$$f_i(z) = f_i(z=0) + \sum_{\mu} \frac{f(z_\mu^+) - f(z_\mu^-)}{2T_{lhapdf}} z_{new,\mu} \,. \tag{3.126}$$

Here, the dependency of PDFs to parton momentum fraction x and factorization scale has been suppressed.

3.4 Tensions Between Data Sets

In a global analysis with many data sets, sometimes some data sets do not perfectly agree with each other, in the sense that they prefer different minima. In fact, even for perfectly mutually compatible data sets, some slight difference in the preferred minima is often observed. Therefore, a question arises: how much different can the minima be in order for two data sets S and $\bar{S}$ to be said incompatible?

We can use the hypothesis testing method to answer this question, which leads to a set of conditions that defines compatibility. Stronger statements can be made by using the confidence region of the fitted parameters in th frequentist view, however, the resulting conditions will be too strict to be applied in a PDF fitting. Given a data set S, we define the p% (hypothesis-testing) confidence region $R_S^{p\%}$ of S as

$$R_S^{p\%} = \left\{ a \middle| \chi_S^2(a) \leq \chi^2_{N_S,p\%} \right\} , \tag{3.127}$$

where $\chi_S^2(a)$ is the χ^2 value of S for given theory parameters a and $\chi^2_{N_S,p\%}$ is the p% percentile of χ^2 distribution with N_S degrees of freedom. Here, N_S is the number of data points in S.

In Fig. 3.5, we show the illustrations of (hypothesis testing) confidence regions of fits with S (blue oval) and $\bar{S}$ (red oval). In all panels in Fig. 3.5, the origins represent the minimum of the combined fit with S and $\bar{S}$. When we say two data sets are mutually compatible, we mean that the minimum of the combined fit with $S + \bar{S}$ is inside the confidence regions of S and $\bar{S}$. This implies that the confidence regions of S and $\bar{S}$ must overlap. Note that, as shown in Fig. 3.5(c), the converse is not true : overlapping confidence region does not imply compatibility.

For future reference, we also define other kinds of compatibility:

- *Mutual compatibility* : the minimium of the combined fit with S and $\bar{S}$ is inside the confidence regions of both S and $\bar{S}$ (see Fig. 3.5(a)). This means, the combined fit can explain both the data at p% confidence level (CL).
- *S-compatibility* : the minimium of the combined fit with S and $\bar{S}$ is inside the confidence regions of S (see Fig. 3.5(b)). Thus, this type of compatibility ensures that the combined fit can still explain S at p% CL.
- *$\bar{S}$-compatibility* : the minimium of the combined fit with S and $\bar{S}$ is inside the confidence regions of $\bar{S}$ (see Fig. 3.5(c)). This type of compatibility ensures that the combined fit can explain $\bar{S}$ at p% CL.

If the combined fit with S and $\bar{S}$ are outside of the confidence regions of both S and $\bar{S}$, then S and $\bar{S}$ are said to be *mutually incompatible*. Note that, if S- and $\bar{S}$-compatibilities are satisfied, the two data are automatically mutually compatible.

In PDF global analysis, where many data sets are included, S and $\bar{S}$ represent collections of data sets. In this case, it is reasonable to check if all data sets are well described by the the combined fit. For this, one can use χ^2/N metric, where N is the size of the data. If χ^2/N correspond to χ^2 value that exceeds the 90% percentile, then we can say that the combined fit can not describe the data. Using χ^2/N as a metric to assess the description of the data is not the most convenient, however. This is because the value of χ^2/N that corresponds to 90% percentile of χ^2 distribution strongly depends on N. For example, for $N = (5, 10, 20, 100, 1000)$, one has $\chi^2_{N,90\%}/N = (1.85, 1.60, 1.42, 1.18, 1.06)$ respectively. A better way is to use the S_E variable[83] :

$$S_E(\chi^2(N), N) = \sqrt{2\chi^2(N)} - \sqrt{2N-1} \sim \mathcal{N}(0,1) \tag{3.128}$$

which is distributed according to the standard normal distribution for $N \gtrsim 10$. We can therefore define :

- *S_E-compatibility* : the distribution of S_E variable for *all* data sets in S and $\bar{S}$ in the combined fit is approximately standard normal.

When testing two collections S and $\bar{S}$ of data sets in nPDF fit, ideally, one should use the three compatibility criteria : S-, $\bar{S}$, and S_E-compatibilities. However, it is often the case that $\bar{S}$-compatibility can not be evaluated as the fit with $\bar{S}$ alone is not available, or not reliable enough due to limited flavor separations. In this case, one can only use the S-compatibility criterion, which assesses the impact of the pulls of the new data $\bar{S}$ to S.

Chapter 4

Global Analysis with Netrino Data

This chapter is based on the work presented in [72].

A reliable determination of nPDFs based on a global analysis of all available data requires that all the flavors are sufficiently separated. With the additional A, Z degrees of freedom, more data are required for an nPDF global analysis to have the same level of precision as the proton PDF analyses. In reality, the amount of nuclear data is less abundant and less precise, making the nPDFs have large uncertainties. As an example, the nCTEQ15WZSIH analysis[84] used 940 data points, much less than 4600 pts in NNPDF4.0 analysis[21]. This shows that nPDF analysis is in a dire need of including more data, preferably taken from different nuclei.

A type of nuclear data, which has been known for quite a while in the nPDF fitting community is the neutrino DIS taken on iron and lead, measured by NuTeV, CCFR, CDHSW and Chorus collaborations. In fact, some of these data, for example, the neutrino data from Chorus and the neutrino-induced dimuon data from NuTeV, are still used in modern proton PDF analyses, such as CT18[7] and NNPDF4.0[21]. In nuclear PDF case, however, only Chorus was ever used (as in the EPPS21 analysis[24]). This is because the other data have been shown to have irreconcilable tensions with some of the charged lepton DIS data. It is the primary purpose of this study to investigate this issue further.

4.1 Review of Past nPDF Analyses with Neutrino Data

In this section, we give a review of studies on the compatibility between the neutrino and charged lepton data. In Ref. [85], it was shown by conducting a global analysis of neutrino DIS data from NuTeV and dimuon data from NuTeV and CCFR, that the extracted iron PDFs using the nCTEQ framework leads to a nuclear ratio of the charged-current structure function F_2 that is flatter and significantly different from the Kulagin-Petti model [58] and the SLAC/NMC parametrization [86], which are usually used to correct charged-lepton DIS data in proton PDF analyses. In particular, the lack of shadowing of the charged-current structure function ratio in the low-x ($x \leq 0.1$) region is quite atypical. Although Ref. [85] did not study the compatibility with other data, the behavior of the extracted nuclear ratio clearly shows some signs of tension. In a follow-up study by Kovarík *et al.* [87], by performing a global analysis that included charged-lepton and Drell-Yan (DY) data as well as neutrino DIS from NuTeV [70]

and Chorus [71], it was even concluded that these neutrino DIS data are incompatible with the charged-lepton data.

A contradictory conclusion was obtained in a global fit by De Florian *et al* [88]. This study analyzed charged-lepton DIS, DY, pion production, and F_2 and F_3 neutrino data from CDHSW, NuTeV, and Chorus. All correlated uncertainties were added in quadrature, hence ignoring the point-by-point correlations of the data. The study shows that, qualitatively, these neutrino data sets are reproduced within their respective uncertainties by the combined fit, although the χ^2/datum = 1.41 for the F_2 NuTeV data is well above unity. A study on the issue of neutrino DIS was later done by the EPPS group [89]. This time, the more abundant differential cross section data was used. The study suggested that data normalization might be the reason for the apparent incompatibility between the neutrino and charged-lepton DIS. By normalizing the cross section data with the integrated cross section in each energy bin and using Hessian reweighting analysis, it was shown that the neutrino DIS data, in particular the one from NuTeV, could be included in a global analysis with charged-lepton data without causing significant tension. It is worth noting that the NuTeV data used in Ref. [89] was without the data correlations, which was shown to possess a very good χ^2, even in the analysis of Ref. [87].

Another intriguing study was done by Kalantarians *et al* [90]. There, F_2^{Fe}/F_2^D data from BCDMS and NMC were brought into F_2^{Fe} data by multiplying it with F_2^D from the NMC parameterization[86]. This neutral current F_2^{Fe} data is then compared with charge current F_2^{Fe} data from NuTeV, CCFR, and CDHSW, after correcting them using the famous 18/5-rule. Although agreement at the valence region ($x > 0.3$) can be shown, around 15% discrepancies at $x < 0.15$ are visible. However, we point out here that such discrepancy could still be explained by the factorization framework at NLO of pQCD when heavy quark effects are properly treated. As shown in Fig. 4.1, using the nCTEQ15WZ nPDFs[79] with S-ACOT scheme [91, 92] for the heavy quark treatments, the NLO predictions for $F_2^{l^\pm A}$ and $F_2^{CC} \equiv (F_2^{\nu A} + F_2^{\bar{\nu} A})/2$ differ at around 15% at low x.

It is important to stress that the notion of compatibility in general is always dependent on the specific nPDF fitting framework employed, the compatibility criteria used to quantify tensions, the type of neutrino data (differential cross section or structure function data), and how the uncertainties of the data are treated during the fitting procedure. Studies from the nCTEQ, EPPS, and De Florian *et al* groups used different nPDF frameworks. In particular, the proton PDF baseline and the A-dependence parametrizations differ. They also used different compatibility criteria that are not necessarily equivalent. As for the type of data, while De Florian *et al* groups used F_2 and F_3 structure function data, the use of cross section data is arguably preferred as extracting the latter always involves more assumptions .[1] Because of these differences, it is therefore not completely unexpected that the resulting conclusions about the neutrino data compatibility are different.

[1] In particular, some inputs for the cross section ratio of longitudinally to transversely polarized W bosons $R_L(x, Q^2)$ and $\Delta x F_3 \equiv x F_3^{\nu} - x F_3^{\bar{\nu}}$ are needed to extract the structure function. In the case of NuTeV data, $R_L(x, Q^2)$ was obtained from a fit to $e - p$ and $e - d$ world data at that time [93], and therefore ignore any nuclear effects. $\Delta x F_3 \equiv x F_3^{\nu} - x F_3^{\bar{\nu}}$ was calculated using the QCD parton model with the MRSTW PDFs [94].

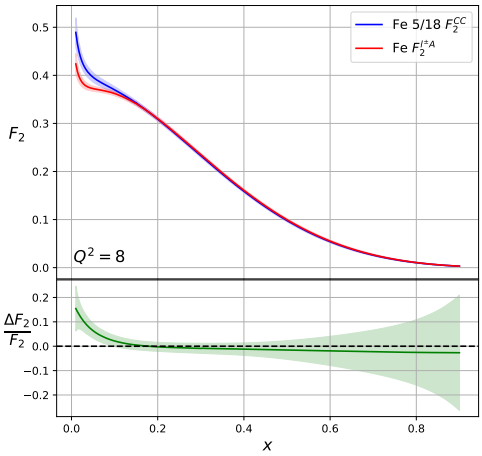


FIGURE 4.1: Top panel: The comparison between neutral current $F_2^{l^\pm A}$ and the charge current $F_2^{CC} \equiv 1/2\,(F_2^{\nu A} + F_2^{\bar{\nu} A})$ structure function predictions for iron (A=56, Z=28) using the nCTEQ15WZ nPDFs. Bottom panel: The ratio $\Delta F_2/F_2 \equiv (5F_2^{CC}/18 - F_2^{l^\pm A})/F_2^{l^\pm A}$.

4.2 Neutrino DIS Sensitivity

With the addition of neutrino data, it is necessary to understand how these data help constrain PDF components in a global fit. The cross section of (anti-)neutrino scattering on a nucleus with a mass number A can be computed using the formula

$$\frac{1}{E}\frac{d^2\sigma^{\nu A(\bar{\nu}A)}}{dx\,dy} = \frac{G_F^2 M_W^4}{\left(Q^2 + M_W^2\right)^2}\frac{M}{\pi}\left[xy^2 F_1^{\nu A(\bar{\nu}A)}\left(1 - y - \frac{xyM}{2E}\right)F_2^{\nu A(\bar{\nu}A)} \pm xy\left(1 - \frac{y}{2}\right)F_3^{\nu A(\bar{\nu}A)}\right]. \tag{4.1}$$

Here, E is the energy of incident (anti-)neutrino in the lab frame, G_F is the Fermi constant, M is the proton mass, and M_W is the W boson's mass. The $F_i^{\nu A(\bar{\nu}A)}$ are nuclear structure functions of (anti-)neutrino DIS on a target A. For $\nu A(\bar{\nu}A)$, the sign preceeding F_3 is $+(-)$. The three relativistically invariant variables Q, x, y are related by $Q^2 = 2MExy$.

The Collinear Factorization Theorem stipulates that the cross section can be written as convolutions of parton-level cross section with the PDFs

$$d\sigma = \sum_i d\hat{\sigma} \otimes f_k. \tag{4.2}$$

The Wilson coefficient C_{ik} can be computed order-by-order in perturbative QCD. In this section, for simplicity, we will work at leading order (LO) and assume the first two generations of quarks are massless and we negelect the contributions of bottom and top quarks.

Suppressing the dependence on x and Q, the structure functions can be expressed as

$$F_1^{\nu A} = d + s + \bar{u} + \bar{c}, \tag{4.3}$$
$$F_2^{\nu A} = 2x\,(d + s + \bar{u} + \bar{c})\,, \tag{4.4}$$
$$F_3^{\nu A} = 2\,(d + s - \bar{u} - \bar{c})\,, \tag{4.5}$$
$$F_{1,2}^{\bar{\nu} A} = F_{1,2}^{\nu A}[q \leftrightarrow \bar{q}], \quad F_3^{\bar{\nu} A} = -F_3^{\nu A}[q \leftrightarrow \bar{q}], \tag{4.6}$$

where $u, d, \ldots$ are the full (not bound) up, down, $\ldots$ quark PDFs and we have used assumed that the Callan-Gross relation holds. The analogous formula for charged lepton DIS is given by

$$F_2^{l^\pm A} = x\frac{1}{9}\left[4(u + \bar{u}) + (d + \bar{d}) + 4(c + \bar{c}) + (s + \bar{s})\right]. \tag{4.7}$$

We can now write the LO expression for neutrino cross section formula. Using Eqs. (4.3)-(4.6), and assuming $xyM/2E \ll 1$, we obtain

$$\frac{1}{E}\frac{d^2\sigma^{\nu A}}{dx\,dy} \propto 2x\left[(d + s) + (1 - y)^2\,(\bar{u} + \bar{c})\right], \tag{4.8}$$
$$\frac{1}{E}\frac{d^2\sigma^{\bar{\nu} A}}{dx\,dy} \propto 2x\left[(\bar{d} + \bar{s}) + (1 - y)^2\,(u + c)\right]. \tag{4.9}$$

Here, the proportionality factor is equal to the one before in Eq. (4.1). Due to the suppression factor $(1 - y)^2$, the (anti-)neutrino cross section is more sensitive to d and s ($\bar{d}$ and $\bar{s}$) than $\bar{u}$ and $\bar{c}$ (u and c). For an isoscalar nucleus, due to the quantum number sum rule and isospin symmetry, we have further constraints $d = u$ and $\bar{d} = \bar{u}$. Thus, the (anti-)neutrino data can be used to constrain $(\bar{u})u$ PDFs.

From Eqs. (4.8) and (4.9), it is easy to understand that for a nucleus with more neutrons than protons, such as lead, the neutrino scattering differential cross section will be larger compared to an isoscalar nucleus at the same E, x, y. Similarly, the anti-neutrino counterpart will be smaller. This is because the neutron contains a larger d content and a smaller $\bar{d}$ content compared to proton.

Using Eqs. (4.4) and (4.6), it is possible to derive the so-called "18/5 rule" mentioned in the introduction section. For a nucleus with an equal number of protons and neutrons (isoscalar), then we have $u = d$ and $\bar{u} = \bar{d}$ from isospin symmetry and the quantum number sum rule. It is then straightforward to derive the relation

$$\left(\frac{F_2^{l^\pm A}}{F_2^{CC,A}}\right)_{ISO} = \frac{5}{18}\left[1 + \frac{3}{5}\frac{c + \bar{c} - s - \bar{s}}{q_S}\right], \tag{4.10}$$

where $F_2^{CC,A} = (F_2^{\nu A} + F_2^{\bar{\nu} A})/2$, and $q_S = \sum_i (q_i + \bar{q}_i)$ is the quark singlet PDF. Without the second term in the parenthesis, Eq. (4.10) is often called the "5/18 rule" and is used to compare F_2 data from charged-lepton and neutrino scattering. Deviations from the 5/18 rule are expected, for example, when radiative corrections and/or heavy quark mass effects are large [95], when

isosopin symmetry does not hold to some degree [96], and when the ratio $|c+\bar{c}-s-\bar{s}|/q_S \gg 0$. The latter could happen, for example, by having a large $s+\bar{s}$ density or having a large strange-antistrange asymmetry [97].

The more exclusive process from neutrino scattering is the charm dimuon production, given by

$$\nu_\mu(\overline{\nu_\mu}) \; + \; A \; \rightarrow \; \mu^\mp \; + \; h^\mp \; + \; X. \tag{4.11}$$

Here h is some charmed hadron ($D^0, D^+, D_s^+, \Lambda_c^+$, etc) and X is the remnant of the target. For (anti)neutrino probes, an (anti)charm quark is produced at LO by the interaction of a down-type quark a time-like exchange of a $W^+(W^-)$ gauge boson. The charm quark then fragments into a charmed hadron $h^\mp$, carrying a fraction z of the charm quark momentum. The hadron then decays weakly into a neutrino and a (secondary) muon that has an opposite charge as the (primary) muon from the initial $\nu - A$ interaction. The cross section of charm-tagged dimuon production is then

$$\frac{d\sigma_{\mu\mu}}{dx\,dy\,dz} = \frac{d\sigma_c}{d\xi dy} \sum_h f_h \, D_c^h(z) \, \mathrm{Br}(h \rightarrow \mu X) \, . \tag{4.12}$$

Here, x, y are the usual DIS scaling variable, $\xi = x(1 + m_c^2/Q^2)(1 - x^2M^2/Q^2)$ is the momentum fraction of the struck quark (m_c is the charm quark mass, M is the nucleon mass, Q^2 is the virtuality), $\sigma_c(x, y)$ is the charm production cross section, f_h is the production fraction of a charmed hadron h, D_c^h is the charm fragmentation function into a charmed hadron h, f_h, and $\mathrm{Br}(h \rightarrow \mu X)$ is the branching ratio of hadron h into a muon. At LO, the $\nu A \rightarrow \mu c + X$ production cross section is given by

$$\begin{aligned} \frac{d\sigma_c}{d\xi dy} = & \frac{G_F^2 M_W^4}{\left(Q^2 + M_W^2\right)^2} \frac{M}{\pi} \left(1 - \frac{m_c^2}{2ME_\nu \xi}\right) \\ & \times \left\{ |V_{cs}|^2 s(\xi, Q^2) + |V_{cd}|^2 \left[u(\xi, Q^2) + d(\xi, Q^2) \right] \right\} . \end{aligned} \tag{4.13}$$

which shows us the sensitivity to the strange PDF due to the CKM matrix.

4.3 Neutrino Data

The neutrino DIS data that we analyze here are the data from the NuTeV, CCFR, Chorus, and CDHSW experiments. We refer to these data sets simply as the neutrino data. To help better constrain the strange quark PDF, we also analyze the neutrino-induced charm dimuon production data from NuTeV and CCFR. Although this data is technically obtained from a neutrino DIS experiment, to distinguish it from the inclusive cross section one, we will simply refer to this as the dimuon data. Electromagnetic radiative corrections were applied to NuTeV, CCFR and Chorus data. The point-by-point correlated systematic uncertainties are taken into account wherever available. Specifically for the NuTeV data, these correlated uncertainties are obtained from the NuTeVPack package [70]. Table 4.1 shows the summary of the newly added data sets used in our analysis. The total number of data points is 7123 points after kinematical cuts.

Data set	Nucleus	$E_{\nu/\bar{\nu}}$(GeV)	#pts	Corr.sys.	Ref.
CDHSW ν	Fe	23 - 188	465	No	[98]
CDHSW $\bar{\nu}$			464		
CCFR ν	Fe	35 - 340	1109	No	[95]
CCFR $\bar{\nu}$			1098		
NuTeV ν	Fe	35 - 340	1170	Yes	[70]
NuTeV $\bar{\nu}$			966		
Chorus ν	Pb	25 - 170	412	Yes	[71]
Chorus $\bar{\nu}$			412		
CCFR dimuon ν	Fe	110 - 333	40	No	[99]
CCFR dimuon $\bar{\nu}$		87 - 266	38		
NuTeV dimuon ν	Fe	90 - 245	38	No	[99]
NuTeV dimuon $\bar{\nu}$		79 - 222	34		

TABLE 4.1: New neutrino data sets used in this analysis.

Experiment	#pts	Relative Error(%)
CDHSW ν	59	8.36
CDHSW $\bar{\nu}$	59	10.75
CCFR ν	54	6.01
CCFR $\bar{\nu}$	54	16.90
NuTeV ν	55	5.88
NuTeV $\bar{\nu}$	54	10.29
Chorus ν	65	7.70
Chorus $\bar{\nu}$	65	18.32

TABLE 4.2: Relative experimental uncertainties (in percent) of various data sets at $E_\nu \sim 85$ GeV where all the data sets overlap.

In Table 4.2, we show the average absolute and relative total uncertainties (statistical and systematical errors are added in quadrature) for (anti)neutrino data that we use in our fits for an incoming neutrino energy of $E_\nu \sim 85$ GeV where all these data overlap. We can see that the NuTeV neutrino data has the smallest uncertainties followed by the CCFR, Chorus, and CDHSW. For the anti-neutrino counterpart, the order is slightly different : NuTeV anti-neutrino data is the most precise, followed by the CDHSW, CCFR and then Chorus data.

We note here that other sources of neutrino data are available in the literature, but they are not used in our fits. The latest results come from MINERνA neutrino scatterings on polystyrene, graphite, iron, and lead. The collaboration published the ratio of neutrino scattering single differential cross section $d\sigma/dx$ [100] as function of x and neutrino energy E_ν. However, the average virtuality $\langle Q^2 \rangle$ is below our $Q^2 = 4\,\text{GeV}^2$ threshold for the kinematic cuts, which makes the data unusable in our global analysis. The NOMAD experiment is one of the first fined-grained experiments that allows for high-resolution measurements of exclusive states [101]. However, NOMAD has yet to officially publish the data. Preliminary data from NOMAD have been available since 2005 [102], but their presentation takes the form of a plot and not a table. There

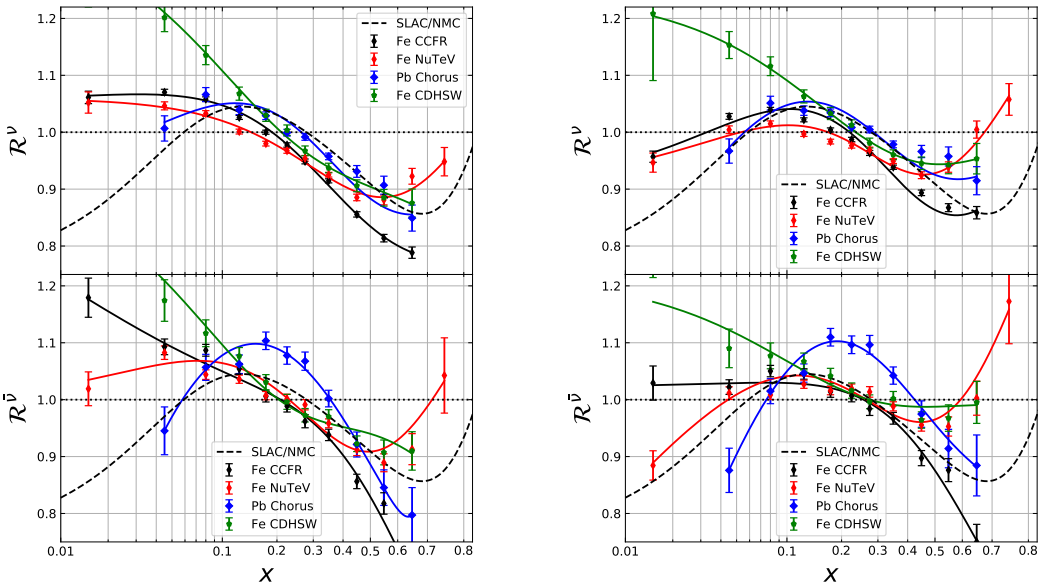


FIGURE 4.2: The weighted average of the cross-section ratios for $Q^2 > 4$ GeV^2 and $W^2 > 12.25$ GeV^2 from CDHSW, CCFR, NuTeV, and Chorus data. The denominator (σ_{free}) is computed using nCTEQ15 proton baseline (left) and CT18 (no-ν-A) NLO proton PDFs without neutrino data of Ref. [105] (right).

are also charm-tagged dimuon production data from Chorus [103] and Nomad [104], which should help constrain the strange quark PDF. In the present analysis, we do not include them in our global fit. We do, however, compare theory predictions from our fits to these data in section 4.8.

4.4 Nuclear Corrections from Neutrino Data

The nuclear PDF framework assumes that there exists a unique set of PDFs that universally can be used to calculate theory predictions using the factorization theorem. Given that charged lepton and neutrino DIS are different processes, it is expected the shape of nuclear corrections are different. The corrections will be the same if the bound nucleon PDFs receive the same nuclear corrections. This is roughly true for $0.01 \leq x \leq 0.6$, which is the region where most nuclear data are in. Thus, we expect that the shape of nuclear correction for neutrino DIS is rather similar to the one in charged lepton DIS.

To study nuclear corrections of neutrino DIS cross sections directly from the data, we construct a ratio :

$$R^{\sigma} = \frac{\sigma(x,y,E)}{\sigma_{free}(x,y,E)}, \quad \text{and} \quad \Delta R^{\sigma} = \frac{\Delta\sigma(x,y,E)}{\sigma_{free}(x,y,E)}, \tag{4.14}$$

Here, σ represents the cross section data for a given x, y, E. $\Delta\sigma(x,y,E)$ is the total sum of statistical and systematical uncertainties added in quadrature, except for the uncertainty in normalization. If the data prefers bound-nucleon PDFs that are identical to the free ones, then the ratio R^{σ} is expected to be around unity. Thus, deviations of R away from unity measure

the nuclear correction for the observable σ. To get an x-dependence for the ratio, we construct a weighted average, such that for a given x the weighted-average ratio and its uncertainty are:

$$\mathcal{R}_x = \sum_i w_i R_i^\sigma, \tag{4.15}$$

$$\Delta\mathcal{R}_x = \left(\sum_i w_i^2 (\Delta R_i^\sigma)^2\right)^{1/2}, \tag{4.16}$$

$$w_i = \left(\sum_j \frac{1}{(\Delta R_j^\sigma)^2}\right)^{-1} \frac{1}{(\Delta R_i^\sigma)^2}, \tag{4.17}$$

where the sums over data point indices i, j are done for all points with the same x. This averaging procedure is similar to the one used in Ref. [89], although there are differences in the weight w_i and $\Delta\mathcal{R}$. Our averaging method is based on maximum likelihood estimation of a single quantity R, given multiple data $\{R_1 \pm \Delta R_1, R_2 \pm \Delta R_2, ...R_N \pm \Delta R_N\}$. The uncertainty $\Delta\mathcal{R}_x$ can be understood as the "spread" of ratio data if the experiment is repeated many times. This averaging procedure could be understood in the limit that the nuclear ratio R is independent of the kinematic variables that were averaged out, in this case, y and E, or equivalently y and Q^2. Thus we assume that the dependence on y and Q^2 is largely canceled out in the ratio. We have checked that such an assumption is reasonably valid for a wide range of Q^2 and y. Deviation from this assumption can be observed at $x = 0.015$ and $x = 0.75$ where R can be seen spread around unity quite widely. Therefore, any inference based on this averaging procedure at $x = 0.015$ and $x = 0.75$ should be done with caution. To emphasize the shape of the nuclear correction, we also plot the interpolation (solid lines), obtained from fits to the following ratio parametrization [70]

$$\mathcal{R}(x) = a_1 + a_2 x + a_3 e^{a_4 x} + a_5 x^{a_6}. \tag{4.18}$$

In the left panel of Fig. 4.2, we show the shape of the cross section ratio where σ_{free} is computed using our proton PDF baseline[65]. We can observe the rough x-shape for the nuclear correction for (anti)neutrino scattering process. For the CCFR and NuTeV data, although they generally agree at low x, at $x > 0.4$, the NuTeV data is consistently above the CCFR data sets. This aligns with the observation in Ref. [70]. Overall, for the iron neutrino data (CDHSW, CCFR and NuTeV), there is no obvious shadowing at low x ($x \leq 0.1$). This is even more so for the CDHSW data. However, one should remember that the bin center correction was not applied for the CDHSW data, which affects largely low- and high-x data points [70]. Chorus anti-neutrino data shows rather typical nuclear correction expected from charged-lepton DIS, represented by the SLAC/NMC curve[86]. In short, while the shape of nuclear corrections at $x \gtrsim 0.1$ in neutrino-iron DIS is similar to those in charged lepton DIS, a striking difference can be observed at low x ($x \lesssim 0.1$).

It is worth noting that the nuclear corrections presented in the left panel of Fig. 4.2 always depend on the free nucleon PDFs used to compute the denominator in the ratio R^σ. Therefore, one may have different results when other proton PDFs are used. On the right panel of Fig. 4.2,

we show the weighted average of cross section ratio, where the denominator σ_{free} is computed using CT18NLO (no-ν-A)[105] proton PDFs instead. One can see that, for neutrino-iron data, the nuclear correction curves are closer to the SLAC/NMC one at low x, while deviating further at high x. This should serve as a warning to draw conclusions about the existence of shadowing in neutrino data from observables, which are not purely data driven and depend on some assumptions such as the proton PDFs.

4.5 The Base Fit : nCTEQ15WZSIHdeut

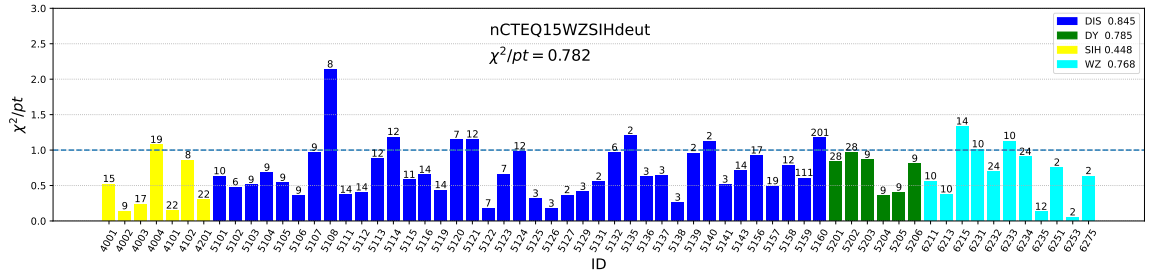


FIGURE 4.3: Values of χ^2/pt for the nCTEQ15WZSIHdeut fit for individual experiments. The IDs of the experiments can be found in Tabs. I-IV of Ref. [25], Tab. II of Ref. [79] and Tab. I of Ref. [84].

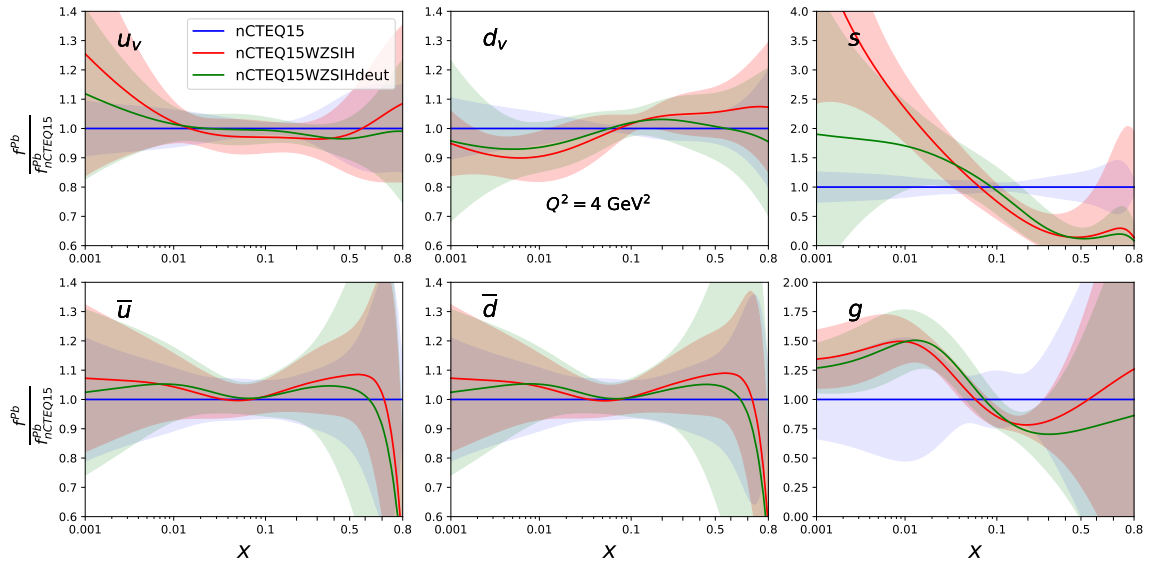


FIGURE 4.4: The ratio of nuclear parton distribution functions of the nCTEQ15WZSIH and nCTEQ15WZSIHdeut analyses with respect to the nCTEQ15 analysis for lead at the scale $Q^2 = 4$ GeV^2.

To study the compatibility of the neutrino data with the other data used in the previous analysis [84], we need to set up a reference fit to which the combined fit with neutrino data will be compared. The reference fit is called nCTEQ15WZSIHdeut, and it is based on the nCTEQ15WZSIH analysis, which used the following data sets :

	ATLAS Run I			CMS Run I			CMS Run II		ALICE		LHCb		DIS	DY	SIH	W,Z LHC	**Total**
	W^-	W^+	Z	W^-	W^+	Z	W^-	W^+	W^-	W^+	Z						
nCTEQ15	(1.38)	(0.71)	(2.88)	(6.13)	(6.38)	(0.05)	(9.65)	(13.20)	(2.30)	(1.46)	(0.70)		0.91	0.73	(0.25)	(6.20)	**1.66**
nCTEQ15WZSIH	0.64	0.26	1.76	1.31	1.16	0.11	0.74	1.14	0.76	0.04	0.56		0.91	0.78	0.41	0.91	**0.83**
nCTEQ15WZSIHdeut	0.56	0.37	1.33	1.01	1.13	0.13	0.70	0.90	0.75	0.05	0.63		0.85	0.79	0.45	0.77	**0.78**

TABLE 4.3: Comparison of the χ^2/pt for the nCTEQ15, nCTEQWZSIH and nCTEQ15WZSIHdeut analyses for selected data sets. Numbers appearing inside brackets show the χ^2/pt values for data sets that are not used in the corresponding fits.

- Charged lepton DIS data from [12–18, 106–115]
- Drell-Yan lepton pair productions data from [116, 117]
- Single Inclusive Hadron data from [118–124]
- Drell-Yan W and Z production data from LHC [125–132]

In the nCTEQ15WZSIHdeut analysis, we impose the following kinematical cuts. For DIS data, we apply our standard kinematic cuts : $Q^2 > 4$ GeV^2 and $W^2 = M_p^2 + Q^2(1-x)/x > 12.25$ GeV^2, where M_p is the nucleon mass. As in [84], we use the same strict $p_T \geq 3$ GeV cut for all single inclusive hadron data (compared to $p_T \geq 1.7$ GeV in the nCTEQ15 and EPPS16 analyses). The main difference with the analysis in [84] is that now deuteron nuclear effects are taken into account when calculating deuteron structure function F_2^D. There are three approaches that one can use to apply deuteron nuclear effects to the theory predictions :

- 'ISO-CJ' method :

$$F_2^D \equiv F_2^{ISO,CTEQ} \frac{F_2^{D,CJ15}}{F_2^{ISO,CJ15}} \tag{4.19}$$

- 'P-CJ' method :

$$F_2^D \equiv F_2^{p,CTEQ} \frac{F_2^{D,CJ15}}{F_2^{P,CJ15}} \tag{4.20}$$

- 'D-CJ' method :

$$F_2^D = F_2^{D,CJ15} \tag{4.21}$$

Here, $F_2^{ISO,CTEQ}$ is an isoscalar structure function using our base proton PDFs[65]. These three methods are equivalent if the base proton PDFs used to compute F_2^{ISO} and F_2^P are the CJ15 PDFs. In this work, we use P-CJ method as it gives the best χ^2 with the conservative cuts used in this analysis. Note that similar approach was also adopted in the previous nCTEQ analysis [60].

In the nCTEQ15WZSIHdeut fit, we enlarged the set of free parameters from 19 to 27 :

$a_1^{u_v}$, $a_2^{u_v}$, $a_4^{u_v}$, $a_5^{u_v}$, $b_1^{u_v}$, $b_2^{u_v}$, $a_1^{d_v}$, $a_2^{d_v}$, $a_4^{d_v}$, $a_5^{d_v}$, $b_1^{d_v}$, $b_2^{d_v}$, $a_1^{\bar{u}+\bar{d}}$, $a_2^{\bar{u}+\bar{d}}$, $a_5^{\bar{u}+\bar{d}}$, a_1^g, a_4^g, a_5^g, b_0^g, b_1^g, b_4^g, b_5^g, $a_0^{s+\bar{s}}$, $a_1^{s+\bar{s}}$, $a_2^{s+\bar{s}}$, $b_0^{s+\bar{s}}$, $b_2^{s+\bar{s}}$.

Beside these PDF parameters, there are 10 additional normalisation parameters which are also determined in the fit using the approach discussed in section 3.1.2. Specifically, 7 normalisation

Dimuon		NuTeV ν		NuTeV $\bar{\nu}$		CCFR ν		CCFR $\bar{\nu}$		Chorus ν		Chorus $\bar{\nu}$		CDHSW ν		CDHSW $\bar{\nu}$		Total	
χ^2/pt	#pts	χ^2/pt	#pts	χ^2/pt	#pts	χ^2/pt	#pts	χ^2/pt	#pts	χ^2/pt	#pts	χ^2/pt	#pts	χ^2/pt	#pts	χ^2/pt	#pts	χ^2/pt	#pts
1.06	150	1.51	1170	1.25	966	1.00	824	1.00	826	1.21	412	1.09	412	0.68	465	0.72	464	1.12	5689

TABLE 4.4: χ^2/pt value for each data set from the DimuNeu fit.

parameters are used to describe the single inclusive hadron experimental data and 3 normalisations are used for the description of the W- and Z-boson production measurements from the LHC. Note that, in the original nCTEQ15WZSIH analysis[84], only 7 normalization parameters are fitted, while normalizations for the LHC data are taken from nCTEQ15WZ analysis[79]. After fitting 940 data points from the same experiments that were also used in the nCTEQ15WZSIH analysis [84], we obtain a $\chi^2 = 735$ corresponding to $\chi^2/\text{pt} = 0.782$.

In Tab. 4.3, we compare the quality of the new nCTEQ15WZSIHdeut fit with the previous nCTEQ15WZSIH and the nCTEQ15 analyses. The values of χ^2/pt for each experiment are displayed in Fig. 4.3. The resulting PDFs are then compared for all relevant flavours at the scale $Q^2 = 4$ GeV^2 in Fig. 4.4. All PDF uncertainties in the figure are constructed using the same $\Delta\chi^2 = 45$ tolerance. There are several differences that can be observed between the original nCTEQ15WZSIH and the nCTEQ15WZSIHdeut analyses. In all parton flavors, we observe larger uncertainties compared to the nCTEQ15WZSIH analysis. This is connected to the enlarged number of free parameters, which can more realistically describe the true uncertainty. The differences in the central values for the up- and down-quark parton distributions are the expected consequences of removing the isoscalar corrections and of the different treatment of the deuterium in DIS data together with a slightly larger number of free parameters. The differences seen in the gluon distribution can be attributed to different free parameters used to describe the gluon PDF as well as secondary effects on the gluon from altered scaling violations coming from the modified deuteron data. In the case of the strange quark, the only constraint comes from the W and Z boson data from the LHC as well as the sum rules linking all the PDFs together. Given the lack of data constraining the strange quark, we conclude that what is displayed in Fig. 4.4 is just the parametrization bias where even our parametrization with a large number of free parameters cannot reproduce the true uncertainty in the determination of the strange quark PDF, which should be regarded as much wider than the plotted bands in Fig. 4.4. It is here where the neutrino DIS data could play a major role in a global PDF analysis, providing additional sensitivity to the strange quark PDF.

4.6 Neutrino DIS Data Fit

nPDF global analysis assumes QCD factorization theorem, which posits the existence of unique and universal PDFs for all processes. Therefore, one way to check, before doing a combined a global analysis, if a set of data S is compatible with another set $\bar{S}$ is to compare the extracted nPDFs from S and $\bar{S}$.

In this section, we will compare the nPDFs from a fit with neutrino and dimuon data alone, which we call the DimuNeu fit, to those from reference fit. Note that extracting a reliable set of

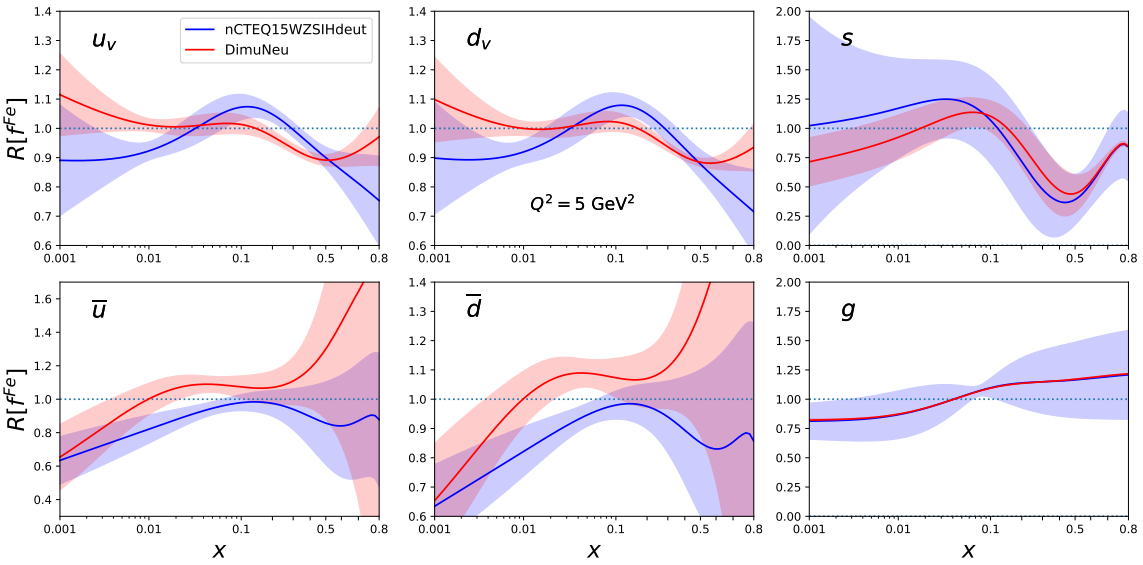


FIGURE 4.5: The ratio of nuclear parton distribution functions for the full nuclei - iron ($A = 56, Z = 26$) - to the nPDF of full nuclei made up of free protons and neutrons both at the scale $Q^2 = 5\,\text{GeV}^2$.

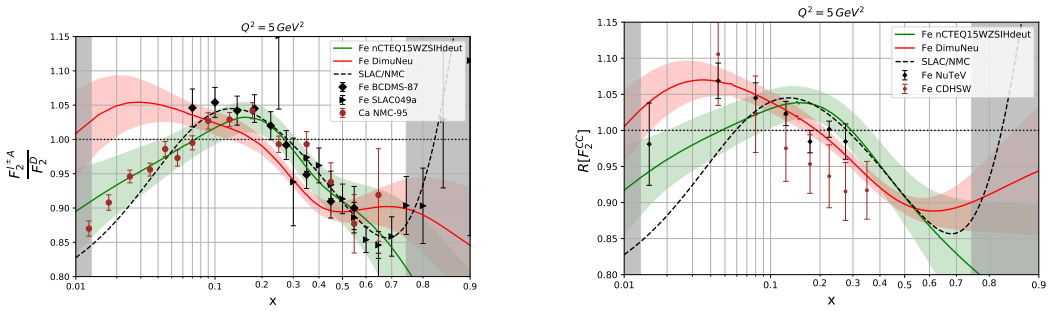


FIGURE 4.6: The structure function ratio predictions from DimuNeu and nCTEQ15WZSIHdeut fits. The grey bands on the left and on the right highlight the regions without any data points passing the kinematic cuts.

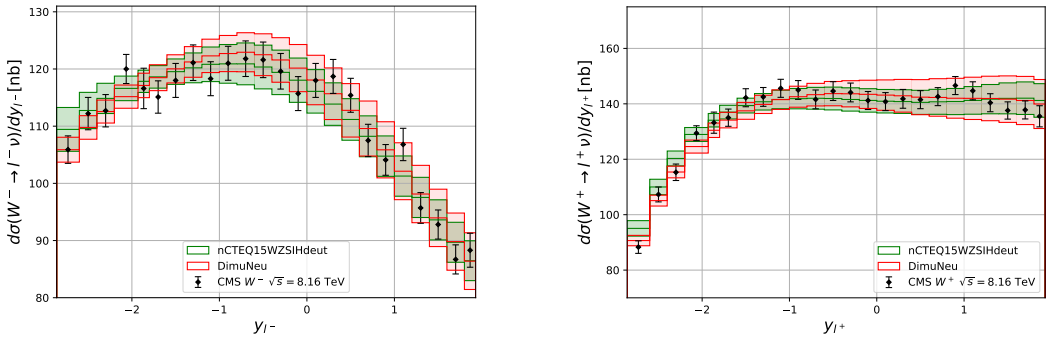


FIGURE 4.7: Comparison between CMS $W^{\pm}$ boson production cross section data with the theory predictions from our fits. The green (red) bands show the theory uncertainties from nCTEQ15WZSIHdeut (DimuNeu) error PDFs. All theory predictions have been shifted by their respective fitted normalization shift.

nPDFs from the neutrino data alone is not straightforward due to the lack of sufficient flavor separations. During the fitting procedure, we want to have a reasonable fit, indicated by a good goodness-of-fit. On the other hand, we also want to make sure that the data are not overfitted by varying the least constrained flavor PDFs. To avoid this, we are forced to use certain assumptions, which could potentially introduce additional theory biases. Therefore, we have to make sure that the additional assumptions do not affect the fit too much. This can be checked, for example, by requiring that upon using the assumptions, the final χ^2 value differ from the one without the assumptions only by less than a few percents. In the DimuNeu fit, we set the gluon PDF parameters to be the same as in the nCTEQ15WZSIHdeut fit. Furthermore, we set $\bar{d}/\bar{u}$ ratio to be the same as in the free proton case, as we assume that the nuclear correction to $\bar{u}$ and $\bar{d}$ are similar and cancel in the ratio [85].

In Table 4.4 we show the total χ^2 per degree of freedom as well as the χ^2 per degree of freedom per data set. Based on the total χ^2 in Tab. 4.4, we see that the DimuNeu result can decently describe all the neutrino data. We see however that not all data are described equally well. For example, the NuTeV data still relatively large χ^2/N. As was stated in previous analyses and verified also in the course of this analysis, the NuTeV neutrino data cannot be adequately described in this nPDF framework even if the data are fitted alone.

In Fig. 4.5, we compare the nPDFs extracted from the neutrino data to nCTEQ15WZSIHdeut nPDFs as discussed in section 4.5. We observe that the results from the DimuNeu and nCTEQ15-WZSIHdeut analyses are distinctly different for the valence quark PDFs as well as for the non-valence quark PDFs. The strange quark nPDF also differs between the two analyses but they are still within the uncertainties. The gluon PDF parameters were fixed and so the gluon PDF is the same in both analyses.

To qualitatively examine the tension between the neutrino and charged-lepton data, we compare the predictions from the nCTEQ15WZSIHdeut and DimuNeu fits to charged lepton nuclear ratio and F_2 data from NuTeV[70] and CDHSW[98]. On the right panel of Fig. 4.6, we plot for $Q^2 = 5$ the theory predictions for charged current (CC) structure function ratio $R(F_2)$, defined by:

$$R[F_2^{CC}] = \frac{F_2^{CC}[f_i^A]}{F_2^{CC}[f_i^{A,\text{free}}]}, \tag{4.22}$$

where $F_2^{CC} = (F_2^{\nu A} + F_2^{\bar{\nu} A})/2$ and $f_i^{A,\text{free}}$ is the free proton PDF. Looking at $R[F_2^{CC}]$ predictions from DimuNeu fit, one can see very similar overall pattern as in R^{ν}-curves in Fig. 4.2. On the left panel in Fig. 4.6, the usual charged-lepton F_2^{Fe}/F_2^D predictions for the same $Q^2 = 5$ GeV^2 are shown. The error bands from the DimuNeu fit are obtained with the same χ^2 tolerance as in nCTEQ15WZSIHdeut, namely $\Delta\chi^2 = 45$. For both charged and neutral current processes, some discrepancies between DimuNeu and nCTEQ15WZSIHdeut predictions can be observed basically at all x. In particular, DimuNeu prediction generally is significantly higher at low x and at $x > 0.1$ than that of nCTEQ15WZSIHdeut. At glance, it seems that the tension can be relaxed by enlarging the normalization factor of the red curve by $\lesssim 3\%$. However, doing so will increase the tensions with the green curve at low x which is driven mainly by NMC calcium

data[107] as shown on the left panel of Fig. 4.6.

In Fig. 4.7, we show the predictions of DimuNeu and nCTEQ15WZSIHdeut for $W^{\pm}$ production at the LHC. The data from CMS run II is also shown. We can see that both fits can describe the data well. This is actually expected as both fits have the same gluon PDF, which in turn gives a dominant contribution to the vector boson production cross section.

4.7 Combined Analysis

Analysis name	χ^2_S/N	χ^2_S/pt	$\chi^2_{\bar{S}}/N$	$\chi^2_{\bar{S}}/pt$	$\Delta\chi^2_S$	$\Delta\chi^2_{\bar{S}}$	$p_S/p_{\bar{S}}$
nCTEQ15WZSIHdeut	735/940	0.78	-	-	0	-	0.500 / -
DimuNeu	-	-	6383/5689	1.12	-	0	- / 0.500
DimuChorus	-	-	1059/974	1.09	-	0	- / 0.500
BaseDimuNeu	866/940	0.92	6666/5689	1.17	131	283	0.99987/0.990
BaseDimuNeuU	861/940	0.92	5569/5689	0.98	126	-	0.99978 / -
BaseDimuNeuX	781/940	0.83	5032/4644	1.08	46	-	0.908 / -
BaseDimuChorus	740/940	0.79	1117/974	1.15	5	58	0.559 / 0.885

TABLE 4.5: Statistical information such as the total χ^2 and the number of data points for all analyses discussed here are presented. Moreover, the χ^2-percentiles with respect to the default data sets of the reference fit nCTEQ15WZSIHdeut (denoted S) and to the neutrino only fits (denoted $\bar{S}$) are also given if applicable.

In this section, we discuss the inclusion of the neutrino data in a global analysis with data sets used in the nCTEQ15WZSIHdeut fit. We will discuss several fits that reflect the way we lessen the tensions. In all these fits, different compatibility definitions as discussed in Section 3.4 will be assessed. First, we will discuss the BaseDimuNeu fit, which is just a combined fit that includes data sets used in nCTEQ15WZSIHdeut and DimuNeu analyses. We will see that the BaseDimuNeu analysis does not satisfy all S-, $\bar{S}$-, and S_E- compatibility criteria. We then discuss other combined fits that contain fewer data points than in the BaseDimuNeu fit, whose focus is to lessen the tensions based on different approaches :

1. Large tensions can often be caused by very precise experimental data, and a compromise can be reached if it is believed that the estimate of the experimental errors is underestimated. In such a case, the errors might be artificially enlarged. This approach leads to BaseDimuNeuU fit.

2. If the tensions can be attributed to a specific kinematic region, they can be removed by imposing a kinematic cut on the neutrino data. This approach leads to BaseDimuNeuX.

3. The last option is to identify experiments that are still consistent with the bulk of the original data and include only those in our analysis. This leads to BaseDimuChorus fit.

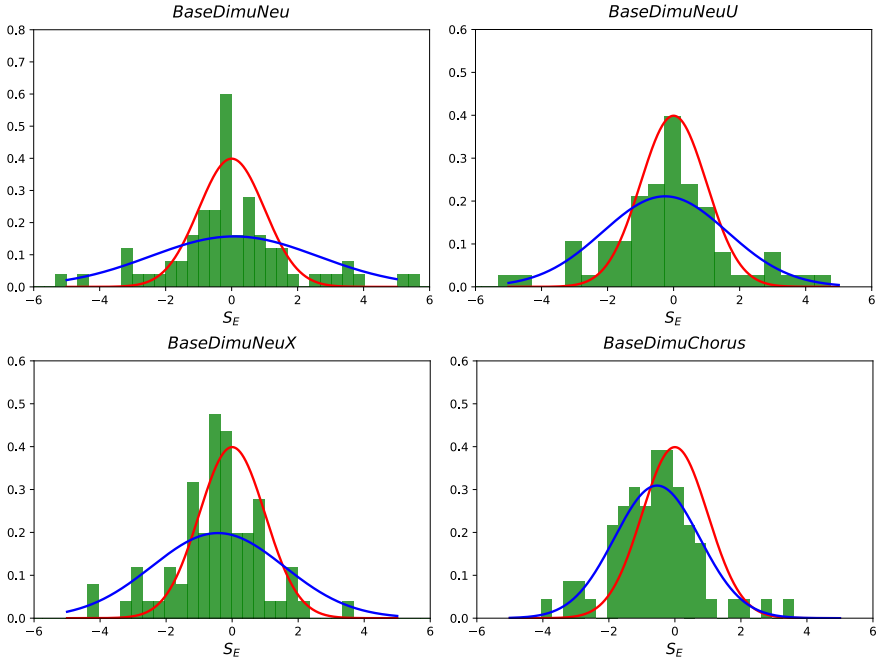


FIGURE 4.8: Distribution of the variable S_E for all experiments in the BaseDimuNeu, BaseDimuNeuU, BaseDimuNeuX and BaseDimuChorus fits. All panels show the fitted Gaussian distribution to the actual S_E distribution (blue) compared to the ideal Gaussian S_E distribution with $\mu = 0$ and $\sigma = 1$ (red).

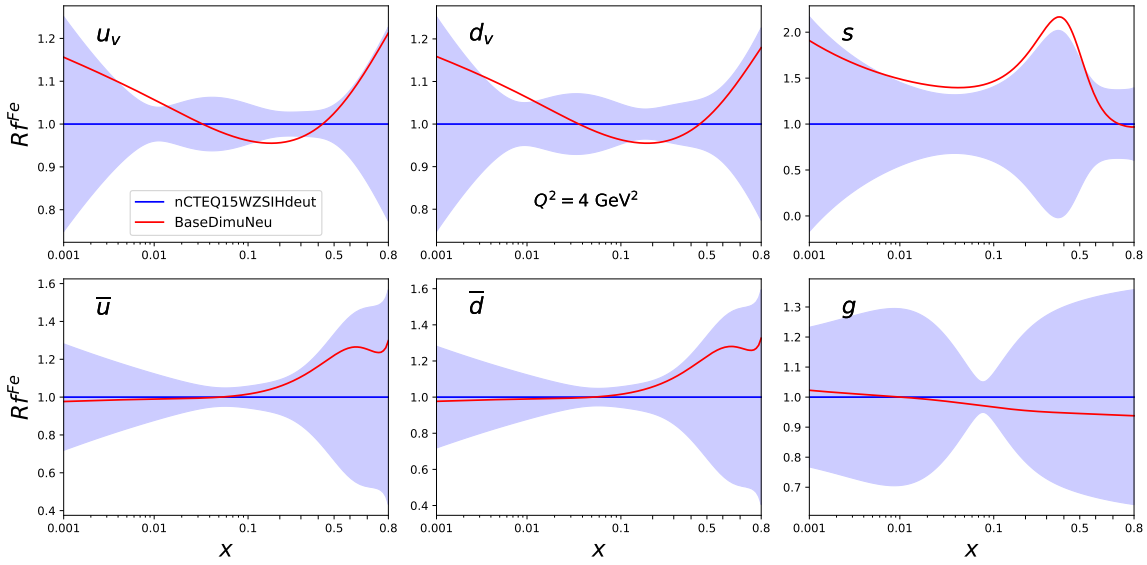


FIGURE 4.9: Ratio of the full iron PDFs to the corresponding PDFs from nCTEQ15WZSIHdeut fit at $Q^2 = 4$ GeV^2. All uncertainty bands are obtained using the Hessian method with $\Delta\chi^2 = 45$.

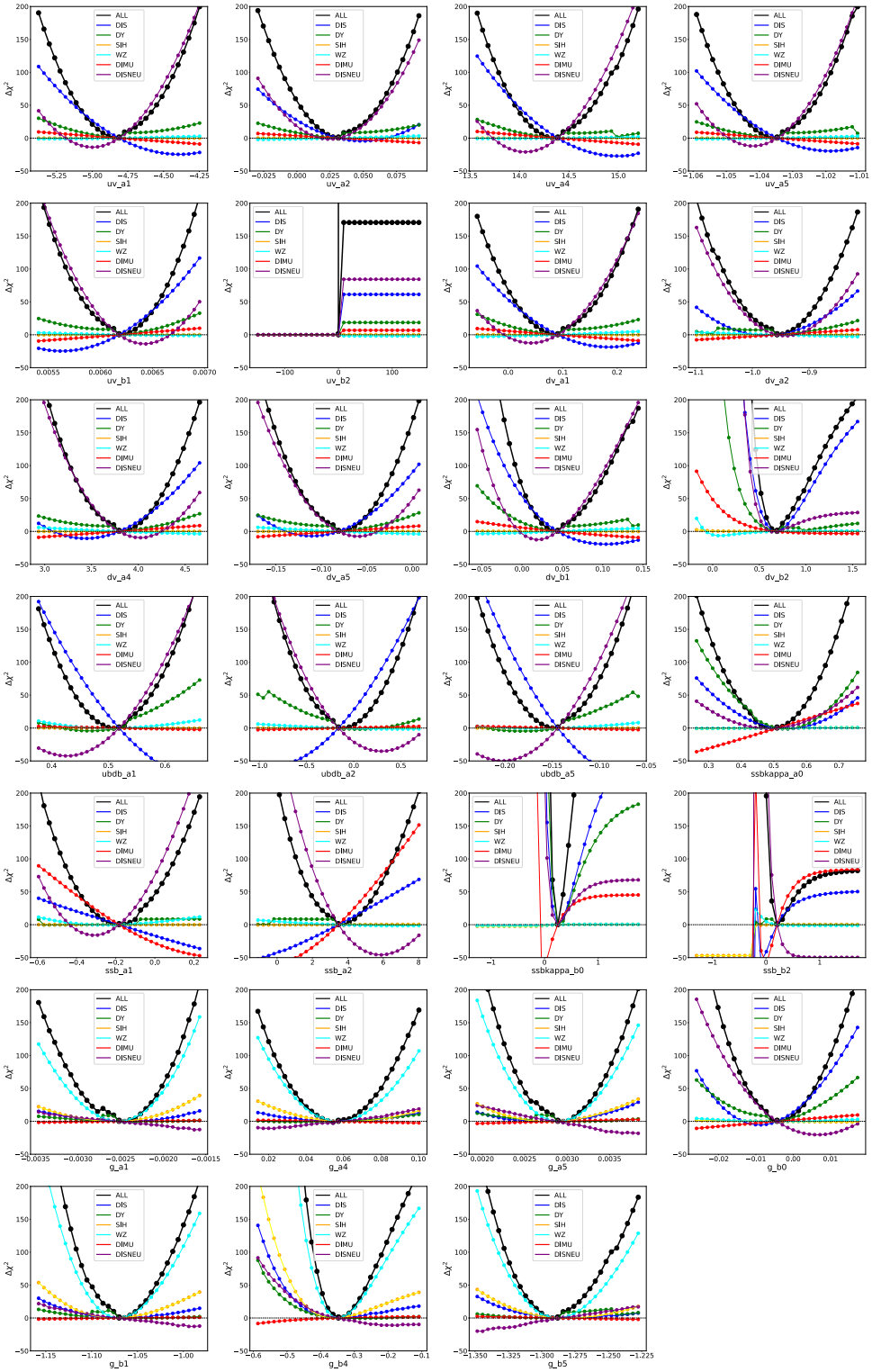


FIGURE 4.10: Scans of the χ^2 function along the PDF parameter directions varying always one free parameter at a time while other parameters were left fixed at the global minimum of the BaseDimuNeu analysis. The breakdown into χ^2 for classes of experimental data is also shown.

4.7.1 BaseDimuNeu

In this section, we discuss the combined fit of these two. The fit is referred to as BaseDimuNeu, which contains all the data from the reference nCTEQ15WZSIHdeut analysis and all inclusive (anti-)neutrino DIS data from the CDHSW, Chorus, CCFR and NuTeV experiments as well as semi-inclusive di-muon data from CCFR and NuTeV. The total number of the included data points is 6629 pts. We open the same 27 free parameters to determine nuclear PDFs. We obtain $\chi^2 = 7532$ or $\chi^2/\text{pt} = 1.14$.

Before going further, let's discuss the compatibility between S=nCTEQ15WZSIHdeut and $\bar{S}$ =DimuNeu. We can use several definitions of compatibility as discussed in Section 3.4, namely S-, $\bar{S}$-, and S_E-compatibility. To assess these compatibility criteria, in table 4.5, we show statistical information for the BaseDimuNeu fit and all other fits we consider in this work. To assess the $S(\bar{S})$-compatibility, we can take a look at the percentile $p_{S(\bar{S})}$ of the $\chi^2_{S(\bar{S})}$ in the combined fit. For BaseDimuNeu case, we can see that $p_S = 0.99987$ and $p_{\bar{S}} = 0.99$, much higher than $p_{tresh} = 0.95$ threshold value. This suggests that both S and $\bar{S}$ are not S- and $\bar{S}$-compatible. To assess S_E-compatibility, in Fig. 4.8, we show the distribution of S_E variable from each data for the BaseDimuNeu and other fits that will be discussed later. For the BaseDimuNeu fit, we can see that the distribution is much flatter than the ideal standard normal one, suggesting many data sets are pulled outside their respective confidence regions. Thus, we see that the BaseDimuNeu fit fails in S_E-compatibility test.

We can identify the origin of the inconsistencies by examining Tab. 4.6 and Tab. 4.7, which shows the χ^2/pt and S_E values for selected experiments. We can see that the description of the NuTeV, Chorus, and the di-muon data in the BaseDimuNeu analysis is much worse than in the DimuNeu fit. Moreover, if one examines the shifts in the description of the experiments in the reference the nCTEQ15WZSIHdeut analysis, we can discover large shifts in χ^2/pt or alternatively in the S_E variable, especially in precise DIS experiments taken on iron, calcium and carbon.

To see the level of pulls from DimuNeu data, let's take a look at the comparison between the fitted PDFs from BaseDimuNeu and nCTEQ15WZSIHdeut for iron. To better see the difference, we rescale the PDFs with the central values of PDFs from nCTEQ15WZSIHdeut and show the ratio in Fig. 4.9. We can clearly see that the up- and down-quark valence PDF distributions as well as the strange-quark nuclear PDF from the global analysis including all the neutrino data lie outside or at the edge of the error band of the reference nCTEQ15WZSIHdeut analysis.

To better see how the tensions realize at the level of nPDF parameters, we show the χ^2 scan by varying one parameter at a time in Fig. 4.10. In Fig. 4.10 we see that for many quark parameters the result of the BaseDimuNeu analysis is a compromise between the neutral current DIS data already present in the nCTEQ15WZSIHdeut analysis (labeled DIS in Fig. 4.10) and the newly added inclusive neutrino DIS data (labeled DISNEU). The DIS and DISNEU subsets show clear sensitivity to the quark valence and the sea $\bar{u} + \bar{d}$ parameters, but the minimum preferred by those data are widely separated. The situation is different in the case of strange quark, where neutrino data seems to provide much better constraints compared to the charged

Data set	#pts	χ^2/pt (S_E) DimuNeu	χ^2/pt (S_E) BaseDimuNeu
CDHSW ν	465	0.68 (-5.29)	0.59 (-7.01)
CDHSW $\bar{\nu}$	464	0.73 (-4.47)	0.69 (-5.22)
CCFR ν	824	0.99 (-0.09)	1.03 (0.56)
CCFR $\bar{\nu}$	826	1.00 (0.07)	1.02 (0.45)
NuTeV ν	1170	1.51 (11.12)	1.61 (13.05)
NuTeV $\bar{\nu}$	966	1.25 (5.16)	1.27 (5.50)
Chorus ν	412	1.21 (2.85)	1.25 (3.40)
Chorus $\bar{\nu}$	412	1.09 (1.26)	1.25 (3.35)
CCFR dimuon ν	40	1.70 (2.79)	2.52 (5.32)
CCFR dimuon $\bar{\nu}$	38	0.79 (-0.89)	0.64 (-1.68)
NuTeV dimuon ν	38	0.98 (-0.06)	2.11 (4.01)
NuTeV dimuon $\bar{\nu}$	34	0.73 (-1.16)	1.16 (0.70)

TABLE 4.6: Statistical information on the description of the neutrino data sets used in different analyses.

Experiment	Target	ID	#pts	χ^2/pt (S_E) Reference	χ^2/pt (S_E) BaseDimuNeu
NMC-95	C/D	5113	12	0.88 (-0.20)	1.70 (1.59)
NMC-95,re	C/D	5114	12	1.18 (0.53)	2.16 (2.40)
NMC-95	Ca/D	5121	12	1.15 (0.46)	2.98 (3.66)
BCDMS	Fe/D	5101	10	0.63 (-0.81)	2.00 (1.97)
BCDMS	Fe/D	5102	6	0.48 (-0.93)	1.62 (1.09)

TABLE 4.7: Statistical information on the description of the selected neutral current DIS data sets used in the reference nCTEQ15WZSIHdeut and BaseDimuNeu analyses.

lepton DIS data. It is interesting here to see that the dimuon data (labelled DISDIMU) and the charged lepton DIS shows the same preference for the strange quark PDF parameters, contradicting that of neutrino data. This tension can also be seen in Tab. 4.6 where the listed χ^2/pt of the di-muon data signifies that they are described much worse than in the neutrino only DimuNeu analysis.

To summarize, we have seen that combining data sets in Base=nCTEQ15WZSIHdeut and DimuNeu analyses leads to S-, $\bar{S}$-, and S_E- incompatibilities. The tensions seems to originates from the difference in both valence and sea quark PDFs. Interestingly, both charged lepton DIS and dimuon data prefers similar strange quark PDFs, while at the same time contradicting the preference from the neutrino data.

4.7.2 BaseDimuNeuU

It has been suggested in [87, 133] that ignoring the correlation of the NuTeV data can relax the tensions with the charged lepton data. In this section, we will investigate further whether

the impact of ignoring correlation is significant enough for the combined fit to pass all the compatibility criteria.

We repeat the BaseDimuNeu analysis, but now the correlation from NuTeV data is ignored. We call the new fit BaseDimuNeuU. We report the statistical information of BaseDimuNeuU fit in Tab. 4.5. Examining the table, we can see the neutrino data are described much better (χ^2/pt=0.98). However, the tension with the neutral current data is unchanged. In the combined fit BaseDimuNeuU, the increase of χ^2 of nCTEQ15WZSIHdeut data is similar to that of BaseDimuNeu, showing that both the data in DimuNeu and nCTEQ15WZSIHdeut are S-incompatible. Some details of the tensions are visible in the S_E-distribution shown in Fig. 4.8, where the standard deviation of the distribution is much larger than unity (σ = 1.89), showing that both the data do not pass the S_E-compatibility criterion. Large S_E contributions can be traced back to the neutrino di-muon data from both CCFR (S_E=4.77) and NuTeV (S_E=3.19), which as we have seen before prefer a different strange quark PDF compared to the inclusive neutrino data. The tensions with the neutral current DIS data have also not improved but got worse compared to the BaseDimuNeu analysis. The largest S_E contributions still come from the Ca/D and C/D data from the NMC collaboration (S_E=3.91 and S_E=2.45 respectively). These data support shadowing at low x, and therefore contradict the preference from the neutrino data. Therefore, we conclude that the use of correlated systematic errors for the NuTeV data has no effect on the compatibility of the neutrino data with the rest of the scattering data and neglecting the correlations does not reduce the tensions, even though the neutrino data seem to be described well overall.

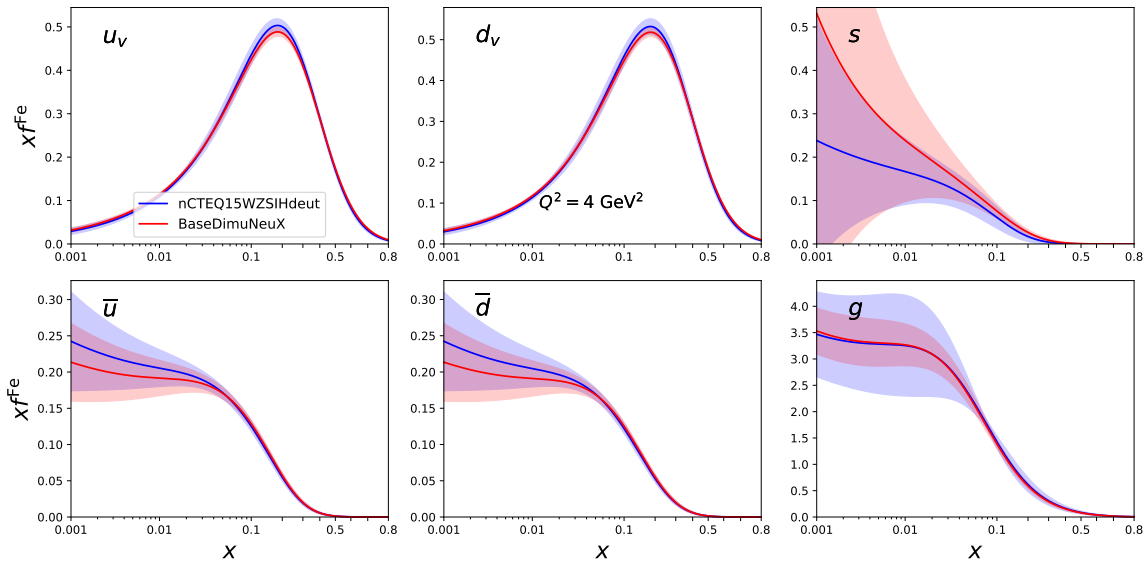


FIGURE 4.11: The full iron PDFs at $Q^2 = 4$ GeV^2. All uncertainty bands are computed using the Hessian method with $\Delta\chi^2 = 45$.

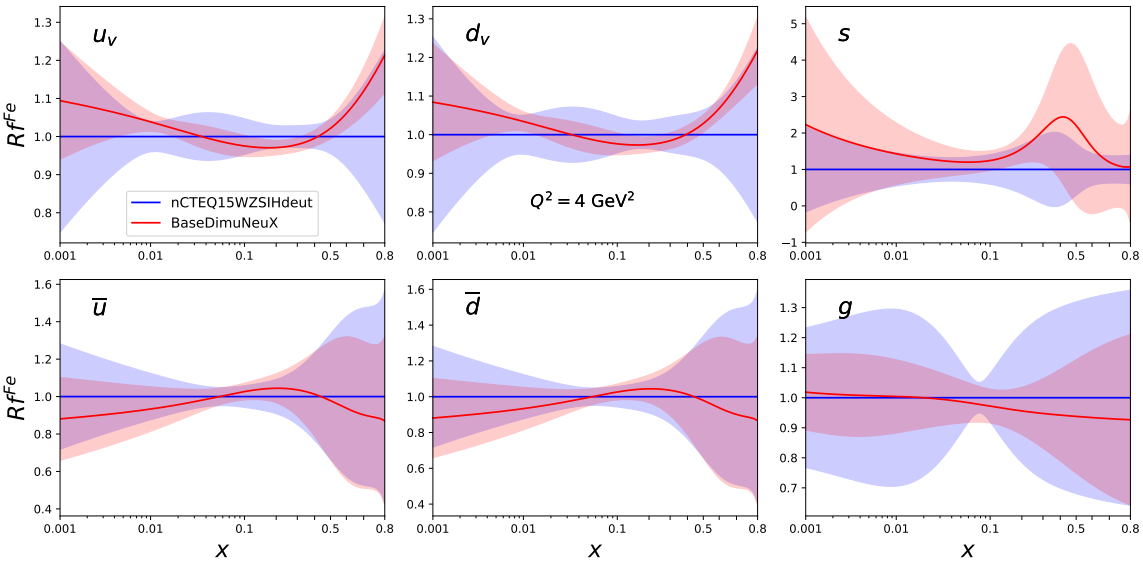


FIGURE 4.12: The fitted iron PDF ratio to nCTEQ15WZSIHdeut. All uncertainty bands are obtained using the Hessian method with $\Delta\chi^2 = 45$.

4.7.3 BaseDimuNeuX

Given that the nuclear ratio extracted directly from the neutrino iron data, as shown in Fig. 4.2, displays no-shadowing at low x, we expect that the tensions between the charged lepton DIS data and the neutrino data are maximal at low x. Therefore, it is natural to ask if excluding the neutrino data from this region in a combined fit can relax the tensions. Before going further, we note that using arbitrary cuts to remove the data which cause the largest tensions in each experiment is not in line with the philosophy of a global analysis, because it introduces a bias or preference to one data over another. The cut is justified only if there is some physical motivation behind it. In this section, we will assume that the large tensions in the low-x region may be due to e.g. a different mechanism for nuclear shadowing in charged current DIS [134] which is not properly included in our theoretical framework, hence causing the tensions.

We proceed further by applying the low x cut (excluding all data with $x \leq 0.1$) to *all* neutrino data, including the dimuon one. We then include these data in a global analysis with nCTEQ15WZSIHdeut. We will call the new fit the BaseDimuNeuX fit. This fit use the same fitting methodology as the BaseDimuNeu fit. The kinematic cut removes 1045 data points from the low-x region of neutrino scattering data. The result of this analysis has $\chi^2/\text{pt} = 1.04$. Further details and the breakdown of the χ^2 for the usual data subsets are listed in Tab. 4.5.

Examining the statistical information of BaseDimuNeuX in the Tab. 4.5, we can see a dramatic reduction in tension between both data, as now we see $\Delta\chi^2 = 46$, which corresponds to 91% percentile of χ^2 of nCTEQ15WZSIHdeut. However, if we examine the S_E distribution as shown in Fig. 4.8, we can see that the distribution is not that much different from that of BaseDimuNeu and BaseDimuNeuU. However, most experiments are fitted well, and the distribution is distorted by a few outliers. The tensions are experienced by the NuTeV neutrino cross-section

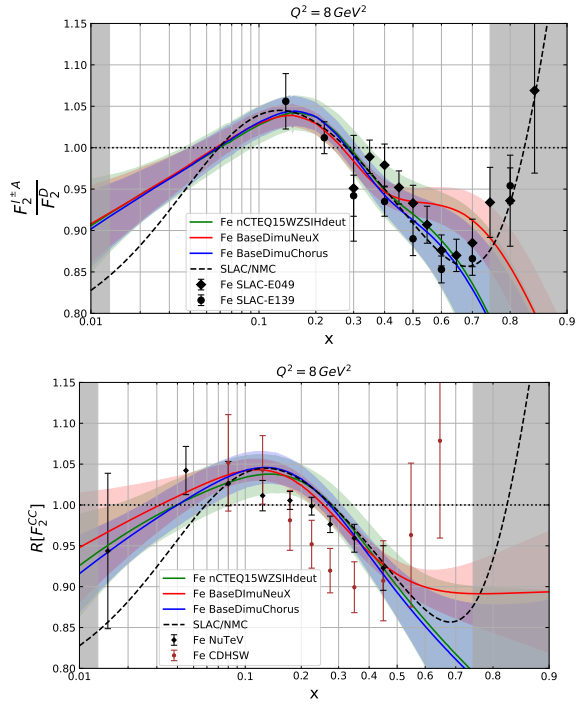


FIGURE 4.13: Neutral current nuclear ratio $F_2^{\rm Fe}/F_2^{\rm D}$ (left) and charged current nuclear ratio $R[F_2^{\rm CC}]$ as defined in Eq. (4.22) (right) using the fitted nPDFs. Note that we have applied nuclear corrections for the neutral current deuterium structure function F_2^D, but not for the charged current one.

data (S_E = 9.72 largest not shown) and by the NuTeV anti-neutrino data (S_E = 3.37). Without these data, the S_E distribution would be very similar to that of nCTEQ15WZSIHdeut.

In Figs. 4.11 and 4.12 we compare the extracted nuclear PDFs to the ones of nCTEQ15WZSIHdeut. Several points can be learned from the figures. First, the central values of BaseDimuNeuX are within the errors of the reference analysis, although the shapes are still similar to that of BaseDimuNeu. This is expected as we have $\Delta\chi^2_S \sim 45$ in the BaseDimuNeuX. Second, we see a huge reduction of PDF uncertainties, showing the constraining power of the new data. Note that for strange quark PDF, the overall uncertainty seems to be bigger than the one from the reference fit. However, in the region where the data are located ($0.1 \lesssim x \lesssim 0.3$), the strange quark uncertainty is reduced dramatically.

The predictions for the nuclear correction factors from the neutral and charged current DIS are shown in Fig. 4.13 and compared with those of the reference analysis. For the neutral current F_2^A/F_2^D at low x, we can see a perfect agreement with that of the reference fit. This is expected as no neutrino data is present in this region due to the low x cut. At high x, we see disagreement between the charged lepton data and the predictions from BaseDimuNeuX, which shows softer shadowing. The shallow shadowing in the EMC region is a known behaviour of the NuTeV data, see Fig. 4.2. Thus, tensions at high x, although less significant compared to the ones in low x, are still present. On the right panel of Fig. 4.13, we see that the structure function data from NuTeV and CDHSW are not correctly described even in the intermediate

x region. In short, while the tensions at low x in BaseDimuNeuX largely disappear, there are still ones at higher x. Overall, we see that applying the cut $x > 0.1$ to all neutrino data reduces the tensions just enough for this fit to be considered consistent. It needs to be stressed once more that this analysis can be considered the final result only if a plausible explanation for the additional kinematic cut is put forward.

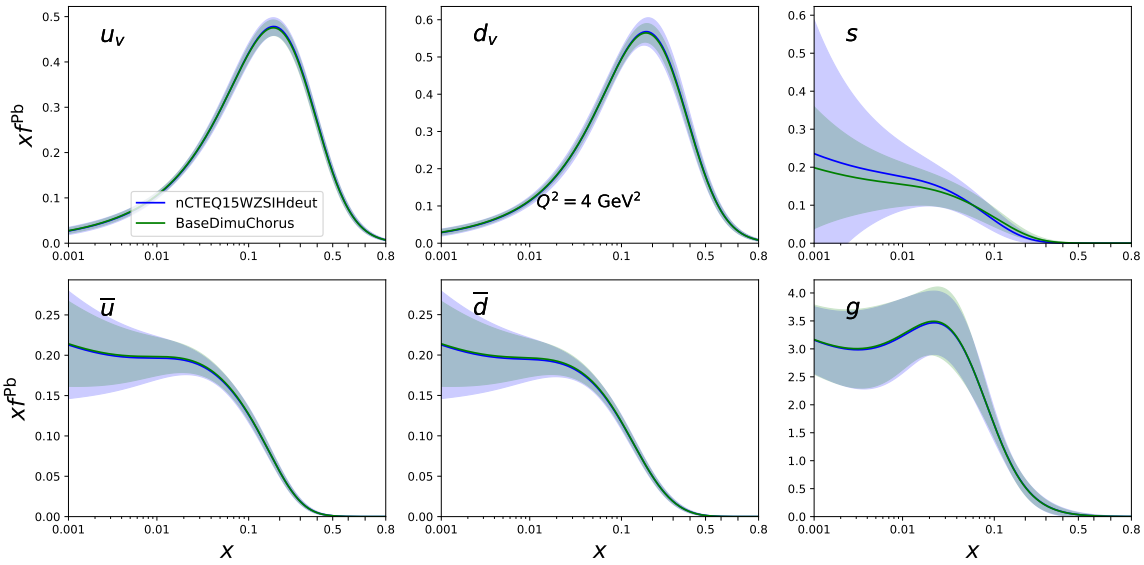


FIGURE 4.14: The full lead PDFs at $Q^2 = 4$ GeV^2. All uncertainty bands are computed using the Hessian method with $\Delta\chi^2 = 45$.

4.7.4 BaseDimuChorus

In Section 4.7.3, we have seen that by excluding low x neutrino data, the tensions between the charged lepton and neutrino data can be significantly reduced, but barely enough to pass the compatibility criteria. In this section, we follow a slightly different approach: we only include neutrino data that are compatible with the charged lepton data. By examining Fig. 4.2, it is not hard to identify Chorus data as potentially compatible with the charged lepton data, as the extracted nuclear ratio shape is similar to SLAC/NMC curve. Furthermore, from the scan in Fig. 4.10, we can see that the dimuon data prefer similar strange quark PDF as the charged lepton data, suggesting compatibility. Thus, this time, we exclude all inclusive neutrino data taken on iron, keeping only dimuon and Chorus data. The resulting combined fit will be referred to as BaseDimuChorus, with an overall χ^2/pt =0.97.

As usual, let's start by examining the compatibility criteria for this fit. Looking at statistical information for this fit in Tab. 4.5, we can see that the tensions with the charged lepton data is almost nonexistent. This fit easily passes the S- and $\bar{S}$- compatibility criteria. Examining the S_E distribution from BaseDimuChorus fit in Fig. 4.8, we can see that the resulting distribution is now much more similar to the ideal case. The fit, in general, is overfitted, with average S_E given by μ = -0.54 and the standard deviation of the distribution is larger than the one for

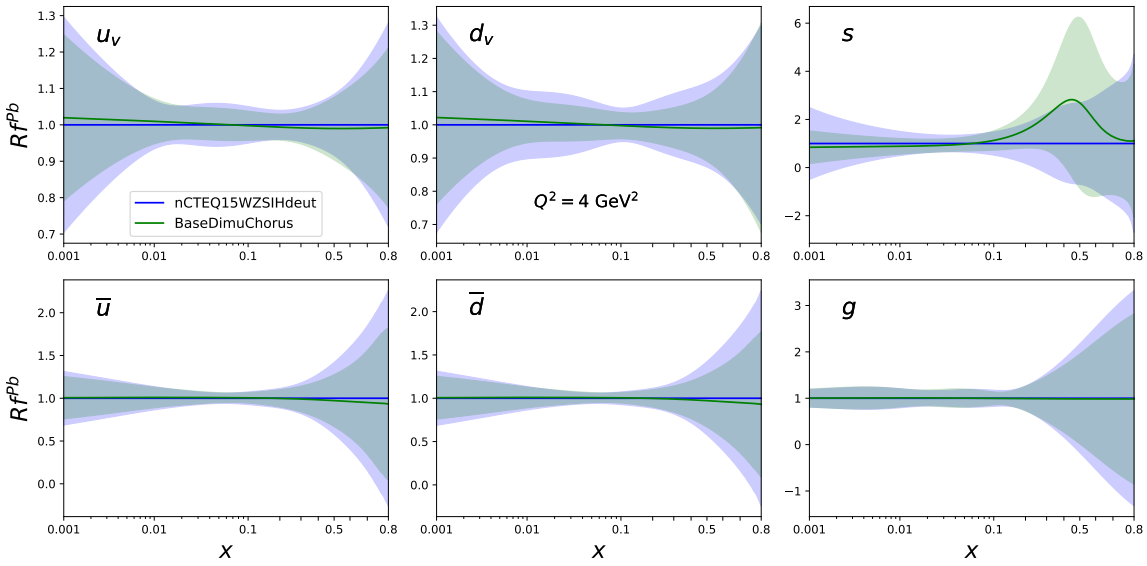


FIGURE 4.15: The fitted lead PDF ratio to nCTEQ15WZSIHdeut. All uncertainty bands are obtained using the Hessian method with $\Delta\chi^2 = 45$.

nCTEQ15WZSIHdeut ($\sigma = 1.28$). The wide S_E-distribution is mostly due to Chorus data ($\chi^2/\text{pt} = 1.27$, $S_E = 3.61$) and also the di-muon data from CCFR ($\chi^2/\text{pt} = 1.68$, $S_E = 2.70$). However, we can see that these data were not described much better in any other analysis. Given that all the other criteria do not signal inconsistencies, we can still regard the distribution as acceptable.

In Figs. 4.14 and 4.15 we show the extracted lead PDFs from this analysis and compare them to those extracted from the reference nCTEQ15WZSIHdeut analysis. We can see that the central values are almost identical, except for the strange quark PDF at high x. Looking at the PDF uncertainties, we see slight reductions for the valence and sea quark PDFs, and no reduction in the gluon PDF uncertainty. Given that we have an additional 974 data points from the neutrino data, which is slightly more than the data in reference fit, we expect that the reduction of the uncertainties would be bigger. This suggests that the Chorus data has large errors and therefore does not have the same constraining power as the other data. Nevertheless, such a very good agreement between the central values of BaseDimuChorus and the reference fit reassures us that it is indeed possible to include some of the neutrino data in a fit with the charged lepton data.

In Figs. 4.13, we compare the theory predictions for the neutral and charge current structure function ratio from BaseDimuChorus and the reference fit. We can see that they are almost identical. We can also see that the theoretical prediction from the BaseDimuChorus analysis does not describe the structure function data from NuTeV or CDHSW well. This is expected as we have omitted the inclusive neutrino data from iron. We should note that even though the normalization of the cross-section data from NuTeV was allowed to vary as a part of the fitting procedure, no shift was applied to the structure function data shown in Figs. 4.13. Shifting the NuTeV data by the normalization of 3.6% determined in the BaseDimuNeuX analysis would reduce the tensions between the data and the theory from this analysis.

4.8 Application : Comparisons with the NOMAD and CDHS Data

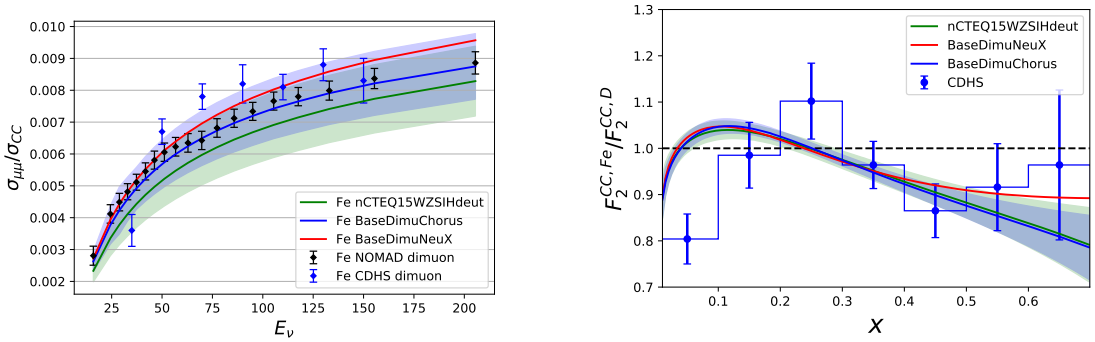


FIGURE 4.16: Comparison between the data from the NOMAD experiment [104] and our theory predictions using our fitted PDFs for the ratio of the di-muon production and the total charged current DIS cross-section.

As discussed in the previous sections, we have by now identified two fits that pass our compatibility criteria: BaseDimuNeuX and BaseDimuChorus. In this section, we will discuss how predictions from these two fits compare to neutrino data that was not included in these fits. For this, we will use dimuon production data from NOMAD[104] and the structure function ratio $F_2^{CC,Fe}/F_2^{CC,D}$ data from the old CDHS experiment[135].

The dimuon data from NOMAD collaboration[104] takes the form of cross section ratio $\sigma_{\mu\mu}/\sigma_{cc}$ as a function of E_μ, x and $\sqrt{\hat{s}} = \sqrt{Q^2(1/x - x)}$. Here, $\sigma_{\mu\mu}$ is the charm-tagged dimuon cross section and σ_{CC} is the inclusive neutrino scattering cross section, E_ν is the incoming neutrino energy and x is the Bjorken variable. To calculate the theory prediction for $\sigma_{\mu\mu}$, we compute the numerator $\sigma_{\mu\mu}$ as :

$$d\sigma_{\mu\mu}(E_\nu) = \int_{Q>Q_0} \frac{d\sigma_{\mu\mu}}{dx\,dy} dx dy \tag{4.23}$$

$$d\sigma_{\mu\mu}(x) = N\Delta x \int_{Q>Q_0} \frac{d\sigma_{\mu\mu}}{dx\,dy}\,\phi(E_\nu),\, dy\, dE_\nu \tag{4.24}$$

Here, $Q^2 = 2ME_\nu xy$, $Q_0 = 1.3$ GeV, Δx is the size of x-bin, $\phi(E_\nu)$ is the E_ν-dependent flux and $N = (\int \phi(E)dE)^{-1}$ where the integral lower bound is $E_\nu = 3$ GeV and for the upper bound $E_\nu = 297$ GeV. Although NOMAD uses $Q^2 > 1$ GeV^2 acceptance, using slightly higher value of Q^2 should have a little effect on the predicted cross section ratio, as done in [136]. The E_ν-dependent flux function was obtained from NOMAD flux data (see Appendix A in [104]) which are interpolated using the Lagrange interpolation. For the dimuon differential cross section, we assume that the charm fragmentation function $D_c^h(z) = D_c(z)$ is independent of the hadron h and thus, upon integrating out z, see (4.12) :

$$\frac{d\sigma_{\mu\mu}}{dxdy} = \frac{d\sigma_c}{d\xi dy} B_\nu \tag{4.25}$$

where a normalized to one fragmentation function $D_c(z)$ has been assumed. The E_ν dependent semi-leptonic branching ratio $B_\nu = \sum_h f_h \mathrm{Br}(h \to \mu X)$ is parametrized as $B_\nu(E_\nu) = a(1 + b/E_\nu)^{-1}$ where $a = 0.097$ and $b = 6.7$ are taken from [104].

In the left panel of Fig. 4.16, we show the theory predictions for the cross section ratio $\sigma_{\mu\mu}/\sigma_{cc}$ as a function of the incoming neutrino energy, while in the right panel, we show the predictions for the charged current nuclear ratio $F_2^{CC,Fe}/F_2^{CC,D}$. We also show the corresponding data from NOMAD[104] and CDHS dimuon[137]. We can see that the predictions from fits with neutrino data(the BaseDimuChorus and BaseDimuNeuX fits), can better describe the NOMAD data compared to the reference fit. However, all the NOMAD data are within the uncertainties of the reference fit. Furthermore, at high energy, the predictions from the BaseDimuChorus fit is in better agreement with the data compared to ones from the BaseDimuNeuX. We also observe that the uncertainty on the prediction is much larger than the experimental errors. Therefore, this data has the potential to put stronger constraints on the strange quark PDF.

In the right panel pof Fig. 4.16, we show the predictions for the old CDHS nuclear ratio data[135]. It is interesting to see here that the CDHS data has a typical nuclear ratio shape: shadowing, anti-shadowing, and EMC effect. This is drastically different from the R^ν and $R^{\bar{\nu}}$ curves in Fig. 4.2. Comparing theory predictions from our fits to the data, we can see that we generally have a good agreement for all but the lowest x. That onset of shadowing/anti-shadowing from the data is shifted to the right, making the shadowing the predictions overshoot the data at low x.

4.9 Summary

To summarize, we have investigated the long-standing tensions between the neutrino and charged lepton data within the nCTEQ nPDF fitting framework. First, we set up a reference fit, which is basically nCTEQ15WZSIH fit[84] with deuteron correction applied for the charged lepton data, to which fits with the neutrino data will be compared. Before we performed a global analysis with the neutrino data, we examined the shape of the nuclear ratio extracted from the data directly and identified the low x region as the region that generate most tensions. We then did several combined fits that represent our approaches to relax the tensions. The combined fit with all the neutrino data, the BaseDimuNeu fit, does not pass our compatibility criteria. The same is also true for the BaseDimuNeuU fit, where the NuTeV correlation is ignored.

The combined fits that pass our compatibility criteria are the BaseDimuNeuX, where low x cut is applied to *all* neutrino data, and the BaseDimuChorus, where all inclusive neutrino data on iron are excluded. These fits represent different philosophies when dealing with the tension problem. If, there is a physical motivation that suggests different shadowing mechanism in neutrino DIS, then applying low x cut is scientifically acceptable and therefore the BaseDimuNeuX represent the final combined fit with neutrino data. Alternatively, we drop all

the inclusive neutrino data in the combined fit and work only with tensionless data such as Chorus and Dimuon data. In this approach, the BaseDimuChorus is the final fit.

Without new neutrino DIS data, there is no way to decide if this inconsistency is due to a different mechanism for the neutrino-nucleus interaction or simply a sign of problems in the acquisition of the current neutrino data. Nevertheless, there is potential to obtain new crucial data from novel ideas or experiments such as the proposed Forward Physics Facility [138] at the LHC or from precise measurements of charged current DIS processes at the future Electron-Ion-Collider [139, 140].

Chapter 5

Target Mass Corrections

This chapter is based on the work in [141].

Deep inelastic scattering (DIS) of leptons has been known as a clean example of a process that can be predicted using pQCD and factorization, and hence it has been the backbone for both proton[7, 8, 10, 142–144] and nuclear[25, 60, 72, 88, 145–150] PDF determinations. This process will be at the forefront at the future Electron-Ion Collider (EIC) [140] where DIS off nucleons and various nuclear targets will be studied with high precision.

As the precision and the kinematical coverage of the DIS data has improved, it is important to include all source of corrections that enter into the standard pQCD formalism. Such corrections are electroweak radiative corrections[151], quark mass effects[152], target mass corrections (TMCs)[91, 92, 153], and higher twist effects[10, 58, 154]. In this study, the main focus will be the TMCs.

Target mass corrections arise when one takes into account the mass of the target hadron in the calculation of DIS structure functions. Generally, there are two approaches to calculate TMCs. The first is a method based on parton model and factorization. This standard approach was used in ref. [91, 92, 153, 155] to derive structure functions that include both heavy quark and target mass. The second approach, that will be the main focus here, is based on the operator product expansion (OPE)[34, 156, 157]. The use of OPE to derive TMCs at leading order of pQCD was first done by Georgi and Politzer[158]. This work was extended to next-to-leading order (NLO) QCD by DeRujula, Georgi and Politzer [159]. Kretzer and Reno employed OPE formalism to derive TMCs for charged current (CC) and weak neutral current (NC) neutrino-nucleon DIS,including NLO QCD corrections and heavy quark mass effects using modern conventions[160]. A review article by Schienbein *et al*[161] put TMCs derived using the OPE formalism in a compact form in terms of a master formula.

The purpose of this study is twofold. First, to clarify the use of master formula of TMCs, reviewed in [160, 161] in the nuclear case. Second, to show that the shapes of the TMCs are fairly independent of nPDFs and to provide parameterizations for the TMCs which can be used for fitting nPDFs.

5.1 TMCs in the OPE Formalism

In this section, we will give a brief overview about the operator product expansion (OPE) and how it can be used to derive the TMC master formula[160, 161]. OPE can be regarded as a systematic way to organize divergences appearing in a product of local operators as they approach each other. As explained in [162], product of local operators are generally divergent. An example for this is the (free) vacuum expectation value (vev) of a time ordered product of scalar field operators

$$\langle 0|T\phi(x)\phi(y)|0\rangle = -i\int \frac{d^4p}{(2\pi)^4}\frac{e^{-ip.(x-y)}}{m^2-p^2-i\epsilon}\,. \tag{5.1}$$

It is clear that when $x \to y$, the vev becomes infinite. OPE basically organized the divergence as a series of local operators, whose coefficients are divergent. The statement of OPE goes back to Kenneth Wilson[34, 163], which basically says that, for any two local operators $\hat{A}(x)$ and $\hat{B}(y)$, as $x \to y$, the product can be written as

$$\lim_{x\to y}\hat{A}(x)\hat{B}(y) = \sum_i c_i(x-y)\mathcal{O}(\frac{x+y}{2})\,. \tag{5.2}$$

The usefulness of OPE lies in the fact that the relation (5.2) are valid at operator level. This means, once the coefficients $c_i(0)$ are known from a simple process, the values are still valid for *any* states that sandwich the operators in the left hand side.

Having introduced the formalism, a question naturally arises : how is the OPE useful to compute cross sections or structure functions in the DIS process? To answer this question, recall that the cross section of DIS of a lepton scattering off a nucleon, $l + N \to l' + X$, is proportional to leptonic and hadronic tensors $L_{\mu\nu}$ and $W_{\mu\nu}$:

$$d\sigma \sim L^{\mu\nu}W_{\mu\nu}\,. \tag{5.3}$$

The leptonic tensor can be easily computed with the help of Feynman diagrams. For massless leptons with a photon exchange, one has

$$\sum_{\{\lambda\}} L^{\mu\nu}\Big|_{\text{QED}} = 4e^2\Big\{k_1^\mu k_2^\nu + k_1^\nu k_2^\mu - (k_1\cdot k_2)g^{\mu\nu}\Big\}. \tag{5.4}$$

The hadronic tensor $W_{\mu\nu}$, is proportional to the square of the scattering amplitude of the process $\gamma^* + N \to X$:

$$e^2\epsilon_\mu\epsilon_\nu^* W^{\mu\nu} = \frac{1}{2\pi}\sum_{X,spin}\int d\Pi_X\,(2\pi)^4\delta^4(q+p-p_X)\;|\mathcal{M}(\gamma^*+N\to X)|^2\,. \tag{5.5}$$

Here, $e, \epsilon_\mu, q = k - k', p, p_X$ are the charge of the incoming lepton, the polarization vector of the off-shell photon γ^*, the momentum of the photon, the momentum of the nucleon, and the total

momentum of the hadronic final states respectively. As $\mathcal{M}(\gamma^* + N \to X) = e\epsilon_\mu \langle X|J_\mu(0)|p\rangle$, one can see that the hadronic tensor can be written as

$$W_{\mu\nu}(q,p) = \frac{1}{2\pi}\sum_X \int d\Pi_X \, \langle N(p)|J_\nu(0)|X\rangle\langle X|J_\nu(0)|N(p)\rangle \, (2\pi)^4\delta^4(q+p-p_X) \tag{5.6}$$

$$= \frac{1}{2\pi}\sum_X \int d\Pi_X \, \langle N(p)|J_\nu(0)|X\rangle\langle X|J_\nu(0)|N(p)\rangle e^{i(q+p-p_X)} \tag{5.7}$$

$$= \frac{1}{2\pi}\int d^4z e^{iq.z} \, \langle N(p)|J_\mu(x)J_\nu(0)|N(p)\rangle \, , \tag{5.8}$$

where we have used an integral representation of the delta function :

$$\delta^n(y) = \frac{1}{(2\pi)^n}\int d^n \, z e^{iz.y}, \tag{5.9}$$

translation invariance :

$$\langle N(p)|J_\mu(0)|X\rangle = \langle N(p)|e^{i\hat{P}.x}J_\mu(x)e^{i\hat{P}.x}|X\rangle = e^{-i(p-p_X).x}\langle N(p)|J_\mu(x)|X\rangle, \tag{5.10}$$

as well as the fact that we are considering an inclusive DIS process :

$$\int d\Pi_X |X\rangle\langle X| = 1. \tag{5.11}$$

Alternatively, one can also rewrite (5.8) as[1]

$$W_{\mu\nu}(q,p) = \frac{1}{2\pi}\int d^4z e^{iq.z} \, \langle N(p)|[J_\mu(x),J_\nu(0)]|N(p)\rangle \, . \tag{5.12}$$

In short, we can express the hadronic tensor as a Fourier transform of a product of local operators. At this stage, one can not use OPE directly as due to the integration, the space time points, x and $y = 0$ in (5.8) do not necessarily approach each other. However, in the so-called DIS limit, where

$$Q^2 = -q^2 \to \infty \qquad \nu = \frac{p.q}{M} \to \infty, \qquad x = \frac{Q^2}{2M\nu} \leq 1 \, , \tag{5.13}$$

one can prove that the biggest contribution to the integral in (5.8) comes from the region where $z^2 \to 0$. Further arguments can be set up, as we will see, to show that, while expanding $W_{\mu\nu}$ in the short distance limit using OPE is not possible, one can use the forward amplitude $T_{\mu\nu}$ that allows the use of OPE. As $W_{\mu\nu}$ and $T_{\mu\nu}$ are related by an analytic continuation, once $T_{\mu\nu}$ is obtained using OPE, one can determine $W_{\mu\nu}$ as well.

Following [162], to prove that the biggest contribution to the integral in (5.8) comes from

[1] To prove this, one first notes that the integral $\int d^4z \, e^{iq.z} \, \langle N(p)|J_\nu(0)J_\mu(x)|N(p)\rangle = 0$. This is because the delta function $\delta^4(q-p+p_X)$ factor appearing in the integral after imposing translation invariance, see (5.6) and (5.7). For physical DIS process, one require $q^2 \leq 0$ and $W^2 \equiv p_X^2 \geq M^2$. The argument of the delta function can not be zero, as if it is zero, then $p_X = p - q \to W^2 = M^2 + q^2 - 2p.q = M^2 + q^2(1 + \frac{1}{x}) \leq M^2$, contradicting the physical region condition. As the argument of the delta function is not zero for the physical region, the integral must vanish.

the light cone region ($z^2 \to 0$), one first note that only the region where $|q.z|$ is finite gives non-zero contributions to the integral. This is because an infinite phase makes the integrand oscillate without bound. One then needs to express $q.z$ in terms of Q^2 and ν, in order to see how the DIS limit translates into further constraints on z. In the frame of the nucleon target, $\nu = q_0$. Therefore,

$$q.z = q_0 z_0 - \vec{q}\cdot\vec{z} = \nu\left[z_0 - r\sqrt{1-\frac{q^2}{\nu^2}}\right] = \nu\left[x_0 - r\sqrt{1+\frac{2Mx}{\nu}}\right] \tag{5.14}$$

$$= \nu(z_0 - r) - Mxr + \mathcal{O}(1/\nu)\,, \tag{5.15}$$

where $r = \vec{q}\cdot\vec{z}/|\vec{q}|$. Note that, at this stage, we do not require that $Q^2 \to \infty$. However, as we assume $x \leq 1$ is finite in the DIS limit, the assumption $Q^2 \to \infty$ is implicit. The linear expansion of the square root in the RHS of (5.14) is valid as $2Mx/\nu \ll 1$, which is a consequence of $\nu \to \infty$. In order for $q.z$ to be finite, then both terms in the RHS of (5.15) must be finite as well. Therefore,

$$|z_0 - r| \leq \frac{c}{\nu}, \qquad |r| \leq \frac{d}{Mx}\,. \tag{5.16}$$

Here, c and d are some finite constants. Using the inequality relation : $|a| - |b| \leq |a-b|$, the first inequality in (5.16) gives

$$z_0^2 \leq (|r| + \frac{c}{\nu})^2 \simeq |r|^2 + 2c\frac{r}{\nu} \leq \vec{z}^2 + \frac{2cd}{Mx\nu}\,. \tag{5.17}$$

Using the relation $Mx\nu = Q^2/2$, this implies

$$z^2 \leq \frac{4cd}{Q^2}\,. \tag{5.18}$$

From (5.12) and using causality argument, that $[J_\mu(x), J_\nu(0)] = 0$ for space like z ($z^2 < 0$), one can easily see that the integrand in (5.8) has support only for $z^2 \geq 0$. Thus, in the DIS limit, the only non-vanishing contribution to the hadronic tensor (5.8) comes from the light-cone region :

$$0 \leq z^2 \leq \frac{4cd}{Q^2}\,. \tag{5.19}$$

The proof of the light-cone dominance has been given. But there is still issue with the fact that $z^2 \to 0$ does not necessarily imply $z_\mu \to 0$. In the momentum space, $z_\mu \to 0$ correspond to $q_0 \to \infty$ (in the frame of target nucleon, this correspond to an infinite energy transfer $E_l - E_{l'}$). This implies $Q^2, \nu \to \infty$ while keeping Q^2/ν^2 fixed. This region is of course unphysical in the DIS case, as this implies $x \to \infty$. Thus, taking the short distance limit $z_\mu \to 0$ (which is a requirement for OPE to work) to the hadronic tensor $W_{\mu\nu}$ contradict the DIS limit, invalidating OPE to expand $W_{\mu\nu}$.

As mentioned earlier, while it is not possible to expand $W_{\mu\nu}$ directly using OPE, one can get around that by using OPE on the forward scattering amplitude $T_{\mu\nu}$. The forward scattering

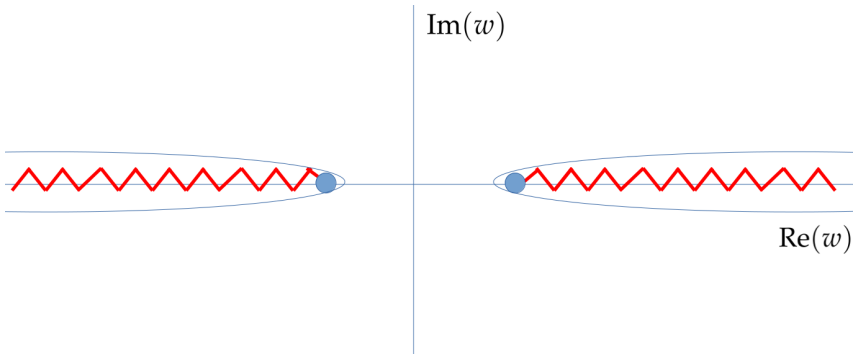


FIGURE 5.1: Contour for the integration of $T_{\mu\nu}$ to avoid branch cuts (sawtooth lines) in the complex w-plane.

amplitude is defined as

$$T_{\mu\nu} = i \int d^4z e^{iq.z} \langle N(p)|TJ_\mu(x)J_\nu(0)|N(p)\rangle \tag{5.20}$$

Note that the factor $1/2\pi$ in (5.12) is replaced with i here. With the time ordered product appearing in the inner product, it is possible to calculate $T_{\mu\nu}$ in perturbation theory using Feynman diagram. Furthermore, $T_{\mu\nu}$ is well-defined in the short distance limit (hence OPE can be used). For the following discussion, it is useful to define $w = 1/x = 2M\nu/Q^2$, as $w \to 0$ as $Q^2 \to \infty$. Due to production of on-shell intermediate states, at fixed Q^2, $T_{\mu\nu}$ is analytic in w except for when $(p \pm q)^2 = M^2 + Q^2(1 \pm w)$[2]. Thus, $T_{\mu\nu}$ has branch cuts : $w > 1$ (DIS region) and $w < 1$ (unphysical region). By decomposing $T_{\mu\nu}$ into its advanced and retarded parts, and by using the inverse Fourier transform of $W_{\mu\nu}$:

$$\langle N(p)|J_\mu(z)J_\nu(0)|N(p)\rangle = 2\pi \int \frac{d^4q'}{(2\pi)^4} e^{-iq'.z} W_{\mu\nu}(p,q')\,, \tag{5.21}$$

it is then possible to derive the relation between $W_{\mu\nu}$ and $T_{\mu\nu}$ as[164]

$$i\left(T_{\mu\nu}\big|_{w+i\epsilon} - T_{\mu\nu}\big|_{w-i\epsilon}\right) = 2\pi W_{\mu\nu}(p,q), \qquad \text{for } w > 0\,, \tag{5.22a}$$

$$i\left(T_{\mu\nu}\big|_{w+i\epsilon} - T_{\mu\nu}\big|_{w-i\epsilon}\right) = -2\pi \left[W_{\mu\nu}(p,-q)\right]^\dagger, \qquad \text{for } w < 0\,. \tag{5.22b}$$

Note that $W_{\mu\nu}(p,q) = 0$ for $w \leq 1$, but in this region, $W_{\mu\nu}(p,-q) \neq 0$.

To obtain the $T_{\mu\nu}$ for $|w| \leq 1$ (the region where it is analytic), one can expand around $w = 0$ as a Laurent series. For a Laurent series $f(z) = \sum_{n=0} a_n (z - z_0)^n$, the coefficient of the expansion can be obtained as

$$a_n = \frac{1}{2\pi i} \oint \frac{f(z)}{(z - z_0)^{n+1}} dz\,. \tag{5.23}$$

[2]The plus sign in $P \pm q$ comes from $\gamma + P \to X$. The minus sign however, comes from $p + X \to \gamma$, which is obvious if one pictures the relevant handbag diagram.

Thus for the expansion of $T_{\mu\nu}$ around $w = 0$:

$$T^{\mu\nu}(p,q) = \sum_n a_n^{\mu\nu} w^n, \tag{5.24}$$

one can calculate the Laurent series coefficient by following a contour as shown in Fig. 5.1. Following (5.23), one can calculate the coefficient a_n^i as

$$\begin{aligned} a_n^{\mu\nu} &= \frac{1}{2\pi i}\oint \frac{T^{\mu\nu}(w)}{w^{n+1}} dw \\ &= \frac{1}{2\pi i}\Big[\int_{w=1}^{\infty} dw\; w^{-(n+1)}(T^{\mu\nu}(w+i\epsilon) - T^{\mu\nu}(w-i\epsilon)) \\ &\qquad + \int_{w=-\infty}^{-1} dw\; w^{-(n+1)}(T^{\mu\nu}(w+i\epsilon) - T^{\mu\nu}(w-i\epsilon))\Big] \\ &= \frac{1}{2\pi i}\int_1^{\infty} w^{-(n+1)}(-2\pi i) W^{\mu\nu}(w) dw + \frac{1}{2\pi i}\int_{-\infty}^{-1} w^{-(n+1)}(2\pi i) W^{\mu\nu}(-w) dw \\ &= \left[-1 + (-1)^{n+1}\right]\int_0^1 x^{n-1} W^{\mu\nu}(x) dx, \end{aligned}$$

where in the last line, we have used (5.22) and we have assumed that $W_{\mu\nu}^{\dagger} = W_{\mu\nu}$, which is valid for neutral current DIS processes. It is then clear that

$$T^{\mu\nu}(p,q) = -2\sum_{2n} W^{\mu\nu,2n}\, w^{2n}\,, \tag{5.25}$$

where $W^{\mu\nu,n}$ is just the n-th Mellin moment of W_i:

$$W^{\mu\nu,n} = \int_0^1 x^{n-1} W^{\mu\nu}(x) dx, \tag{5.26}$$

It is interesting to see here that the coefficients of an odd powers of w in the Laurent series are always zero.

Note that the Mellin moment here is computed in the physical region of $W_{\mu\nu}$. (5.25) then serves as a bridge that connects $W_{\mu\nu}$, which is defined (namely, non vanishing) in the DIS region, to $T_{\mu\nu}$, which is analytic in the short distance limit, in which OPE can be performed. This also summarizes our strategy to compute $W_{\mu\nu}$. First, one evaluates the OPE of $T_{\mu\nu}$ around $w = 0$. This resulting expansion can then be matched to the Laurent series (5.25), giving the Mellin moment of $W_{\mu\nu}$. By inverting the Mellin transform, one obtains $W_{\mu\nu}$.

5.1.1 TMC Master Formula for a Nucleon Target

In this section, we present the derivation of the TMC master formula. The main references for this topic are [158, 160, 161, 164]. To derive the TMC formula, let's start with relating the hadronic tensor $W_{\mu\nu}$ to the structure functions F_i's. From the Lorentz invariant property of $W_{\mu\nu}$,

one can parameterize the dependence on p and q as

$$W_{\mu\nu}(p,q) = -g_{\mu\nu}W_1 + \frac{p_\mu p_\nu}{M^2}W_2 - i\epsilon_{\mu\nu\alpha\beta}\frac{p^\alpha q^\beta}{M^2}W_3 + \frac{q_\mu q_\nu}{M^2}W_4 + \frac{(p_\mu q_\nu \pm p_\nu q_\mu)}{M^2}W_{5,6}\,. \tag{5.27}$$

Here, the structure functions $W_i(x,Q^2)$ are real. Furthermore, $W_6 = 0$ due to time reversal invariance of QCD. These structure functions are related to the ones measured by the experimentalist F_i's as

$$\left\{F_1,\ F_2,\ F_3,\ F_4,\ F_{5\ (6)}\right\} = \left\{W_1,\ \frac{Q^2}{2xM^2}W_2,\ \frac{Q^2}{xM^2}W_3,\ \frac{Q^2}{2M^2}W_4,\ \frac{Q^2}{2xM^2}W_{5\ (6)}\right\}. \tag{5.28}$$

Similarly, one can also parameterize $T_{\mu\nu}$ in the same way :

$$T_{\mu\nu}(p,q) = -g_{\mu\nu}T_1 + \frac{p_\mu p_\nu}{M^2}T_2 - i\epsilon_{\mu\nu\alpha\beta}\frac{p^\alpha q^\beta}{M^2}T_3 + \frac{q_\mu q_\nu}{M^2}T_4 \pm \frac{(p_\mu q_\nu + p_\nu q_\mu)}{M^2}T_{5,6}\,. \tag{5.29}$$

Using OPE formalism, the forward amplitude $T_{\mu\nu}$ can be written as

$$T_{\mu\nu}(p,q) = i\int d^4z e^{iq.z}\,\langle N(p)|TJ_\mu(x)J_\nu(0)|N(p)\rangle = \sum_{i\tau,n} c^{i,\mu_1,\ldots,\mu_n}_{\tau,\mu\nu}(q)\langle N(p)|\mathcal{O}^{i,\tau}_{\mu_1\ldots\mu_n}|N(p)\rangle\,. \tag{5.30}$$

Note here that the Wilson coefficient $c^{i,\mu_1,\ldots,\mu_n}_{\tau,\mu\nu}(q)$ is a function of q alone, while the dependence of the momentum p of external nucleon resides in the expectation value $\langle N(p)|\mathcal{O}^{i,\tau}_{\mu_1\ldots\mu_n}|N(p)\rangle$. The index τ appearing (5.30) denotes the twist, to be made clear in a moment. The index i catalogs all possible operators for a given twist τ and n. Just like in EFT expansions, it is customary to work with a an operator basis which is gauge invariant and transforms as an irreducible representation of the Lorentz group, in this case, labelled by the spin s. An operator of spin s will be a symmetric, traceless tensor fo rank s. As an example, for spin 2 with quark q (instead of nucleon N) in the initial state :

$$\mathcal{O}^{0,2}_{\mu\nu} = \bar{\psi}_q\left(i\gamma_\mu\partial_\nu + i\gamma_\nu\partial_\mu - \frac{1}{2}ig_{\mu\nu}\not{\partial}\right)\psi_q\,. \tag{5.31}$$

For this case, the basis for spin-s operators is [30]

$$\mathcal{O}^{r,\tau}_{\mu_1,\ldots,\mu_s} = \bar{\psi}\gamma_\mu i\partial_{\mu_1}\ldots i\partial_{\mu_s}(-\partial^2)^r\psi_q + \text{symmetrization of } \mu_i \text{ -traces}\,. \tag{5.32}$$

Given the spin s and the mass dimension of these operators d, one can define the twist $\tau = d - s$. From a simple dimensional analysis, together with the requirement that all operators are gauge invariant, it is easy to see that $\tau \geq 2$.

To relate the OPE formalism to parton model, it is sufficient to use twist 2 operators in the OPE of $T_{\mu\nu}$. Furthermore, just like $W_{\mu\nu}$, using Lorentz invariance, one can parameterize the Wilson coefficient $c^{i,\mu_1,\ldots,\mu_n}_{\tau,\mu\nu}(q)$ in terms of scalar coefficients C_i's. The OPE expansion of $T_{\mu\nu}$

then reads [160]:

$$T_{\mu\nu}(p,q) = \mathcal{N}\sum_{k=1}^{\infty}\Big[-g^{\mu\nu}q_{\mu_1}q_{\mu_2}C_1^{2k} + g^{\mu}_{\mu_1}g^{\nu}_{\mu_2}Q^2C_2^{2k} - i\epsilon^{\mu\nu\alpha\beta}g_{\alpha\mu_1}q_\beta q_{\mu_2}C_3^{2k} + \frac{q^\mu q^\nu}{Q^2}q_{\mu_1}q_{\mu_2}C_4^{2k}$$
$$+ (g^{\mu}_{\mu_1}q^\nu q_{\mu_2} \pm g^{\nu}_{\mu_1}q^\mu q_{\mu_2})C_{5,6}^{2k}\Big]\cdot q_{\mu_3}...q_{\mu_{2k}}\frac{2^{2k}}{Q^{4k}}\langle N|\mathcal{O}^{i,\tau}_{\mu_1...\mu_k}|N\rangle \tag{5.33}$$
$$\equiv T_{1\mu\nu} + T_{2\mu\nu} + T_{3\mu\nu} + T_{4\mu\nu} + T_{5\mu\nu} + T_{6\mu\nu} \tag{5.34}$$

where N is a normalization constant to be determined later. The coefficients C_i^{2k}'s can be computed by matching the expansion to quantities calculated in pQCD. The matrix element $\langle N|\mathcal{O}^{i,\tau}_{\mu_1...\mu_n}|N\rangle$ which contains the p-dependence of the $T_{\mu\nu}$, can be written as

$$\langle N|\mathcal{O}^{i,\tau}_{\mu_1...\mu_k}|N\rangle = A^{2k}_{\tau=2}\,\Pi^{\mu_1...\mu_{2k}}(p)\,. \tag{5.35}$$

The coefficient $A^{2k}_{\tau=2}$ is a scalar, which is sometimes called the reduce matrix element. Again, using Lorentz invariant argument, the dependence of $\Pi^{\mu_1...\mu_{2k}}(p)$ can be parameterized as [158, 165]

$$\tilde{\Pi}^{\mu_1...\mu_{2k}} = \sum_{j=0}^{k}(-1)^j\,\frac{(2k-j)!}{2^j(2k)!}\,\underbrace{\{g...g\}}_{j\ g^{\mu_n\mu_{m}}\text{'s}}\ \underbrace{\{p...p\}}_{(2k-2j)\ p^{\mu_n}\text{'s}}\ (p^2)^j \tag{5.36}$$
$$= \underbrace{\{p...p\}}_{2k\ p^{\mu_m}} + \underbrace{\{p...p\}}_{(2k-2)\ p^{\mu_m}}\ \underbrace{\{g...g\}}_{1\ g^{\mu_m\mu_n}}M^2 + \underbrace{\{p...p\}}_{(2k-4)\ p^{\mu_m}}\ \underbrace{\{g...g\}}_{2\ g^{\mu_m\mu_n}}M^4 + \ldots \tag{5.37}$$

Here, $g...g\ \ p...p$ is an abbreviation for a sum over all permutation of the indices. Thus, this represents symmetrization procedure. Note there are $N_{j,k} = (2k)!/[2^j j!(2k-2j)!]$ ways to arrange the indices. One can see that the first term in (5.37) which does not contain the target mass dependence M^2, correspond to the massless parton model. The target mass effects then reside in terms with $j > 0$. As another remark, the contractions between the Wilson coefficients which are functions of q, and $\Pi^{\mu_1...\mu_k}$ result in polynomial of $(2p.q)/Q^2 = 1/x = w$. Thus, at the end, the OPE in this case is just a Taylor expansion around $w = 0$.

Now, to compute $T_i's$, one need to match the OPE expression in (5.33) to the parameterization in (5.29). To do that, one must first do the contractions between the Wilson coefficients and the matrix elements in (5.33). As an example, for the contractions in $T_{1\mu\nu}$ in (5.33) :

$$T_{1\mu\nu} = \mathcal{N}\sum_{k=1}^{\infty}\left[-g_{\mu\nu}C_1^{2k}A^{2k}_{\tau=2}\right]\times\frac{2^{2k}}{(Q^2)^{2k}}\times\left(\prod_{m=1}^{2k}q_{\mu_m}\right)\times\tilde{\Pi}^{\mu_1...\mu_{2k}}$$
$$= -\mathcal{N}g_{\mu\nu}\sum_{k=1}^{\infty}\left[C_1^{2k}A^{2k}_{\tau=2}\right]\sum_{j=0}^{k}\frac{(2k-j)!}{j!(2k-2j)!}\left(\frac{M^2}{Q^2}\right)^j w^{(2k-2j)}\,. \tag{5.38}$$

The next is the contractions in $T_{2\mu\nu}$. They are the most complicated as now we have two metric tensors contracting on a bunch of other metric tensors g and momenta p. In fact, there five

different ways for the two metrics to contract with terms in $\Pi^{\mu_1\cdots\mu_{2k}}$, each leads to different prefactors $g_{\mu\nu}, p_\mu p_\nu, q_\mu q_\nu, p_\mu q_\nu, p_\nu q_\mu$. Therefore, the coefficient C_2^{2k} will appear in the extracted T_1, T_2, T_4, T_5, T_6. Upon working the contractions, one finds

$$\begin{aligned}
T_{2\mu\nu} = {} & \mathcal{N}\frac{2p_\mu p_\nu}{Q^2}\sum_{k=1}^{\infty}\left[C_2^{2k}A_{\tau=2}^{2k}\right]\sum_{j=0}^{k}\frac{(2k-j)!(2k-2)!}{(2k)!j!(2k-2j-2)!}\left(\frac{M^2}{Q^2}\right)^j w^{(2k-2j-2)} \\
& - \mathcal{N}g_{\mu\nu}\sum_{k=1}^{\infty}\left[C_2^{2k}A_{\tau=2}^{2k}\right]\sum_{j=0}^{k}\frac{(2k-j)!(2k-2)!}{(2k)!(j-1)!(2k-2j)!}\left(\frac{M^2}{Q^2}\right)^j w^{(2k-2j)} \\
& + \mathcal{N}\frac{2q_\mu q_\nu}{Q^2}\sum_{k=1}^{\infty}\left[C_2^{2k}A_{\tau=2}^{2k}\right]\sum_{j=0}^{k}\frac{(2k-j)!(2k-4)!}{(2k)!(j-2)!(2k-2j)!}\left(\frac{M^2}{Q^2}\right)^j w^{(2k-2j)} \\
& - \mathcal{N}\frac{2(p_\mu q_\nu + p_\nu q_\mu)}{Q^2}\sum_{k=1}^{\infty}\left[C_2^{2k}A_{\tau=2}^{2k}\right]\sum_{j=0}^{k}\frac{(2k-j)!(2k-3)!}{(2k)!(j-1)!(2k-2j-1)!}\left(\frac{M^2}{Q^2}\right)^j w^{(2k-2j-1)}.
\end{aligned} \tag{5.39}$$

Just like $T_{1\mu\nu}$, the contractions in $T_{3\mu\nu}$ is easy to evaluate. The result is

$$T_{3\mu\nu} = -\mathcal{N}i\epsilon_{\mu\nu\alpha\beta}\frac{p^\alpha q^\beta}{Q^2}(-1)\sum_{k=1}^{\infty}\left[C_3^{2k}A_{\tau=2}^{2k}\right]\sum_{j=0}^{k}\frac{(2k-j)!(2k-1)!}{(2k)!j!(2k-2j-1)!}\left(\frac{M^2}{Q^2}\right)^j w^{(2k-2j-1)}. \tag{5.40}$$

Using the above results, T_i's can be extracted as

$$T_1 = \mathcal{N}\sum_{l=0}^{\infty}w^{2l}\sum_{j=0}^{\infty}\frac{(2l+j)!}{j!(2l)!}\left(\frac{M^2}{Q^2}\right)^j\left[\left(C_1^{2l+2j}A_{\tau=2}^{2l+2j}\right) + \frac{j\left(C_2^{2l+2j}A_{\tau=2}^{2l+2j}\right)}{(2l+2j)(2l+2j-1)}\right], \tag{5.41a}$$

$$T_2 = \mathcal{N}\frac{2M^2}{Q^2}\sum_{l=0}^{\infty}w^{(2l-2)}\sum_{j=0}^{\infty}\frac{(2l+j)!}{j!(2l)!}\left(\frac{M^2}{Q^2}\right)^j(2l)(2l-1)\frac{\left(C_2^{(2l+2j)}A_{\tau=2}^{(2l+2j)}\right)}{(2l+2j)(2l+2j-1)}, \tag{5.41b}$$

$$T_3 = \mathcal{N}\frac{M^2}{Q^2}\sum_{l=0}^{\infty}w^{(2l-1)}\sum_{j=0}^{\infty}\frac{(2l+j)!}{j!(2l)!}\left(\frac{M^2}{Q^2}\right)^j(2l)\frac{\left(C_3^{(2l+2j)}A_{\tau=2}^{(2l+2j)}\right)}{(2l+2j)}. \tag{5.41c}$$

Having derived the OPE expansion of T structure functions, there is still question how the Wilson coefficients C_i^k and the reduced forward amplitude $A_{\tau=2}^k$ can be calculated. To do this, let's assume the nucleon is massless. Therefore, the OPE expansions for T_i's are exactly the same as derived earlier, but with $M^2 = 0$. Note that in order this statement to be true, the reduced forward amplitude $A_{\tau=2}^{2k}$ must not depend on M. In the OPE expansions (5.41a),

(5.41b), (5.41c), setting up $M^2 = 0$ causes terms with $j > 0$ to vanish. Therefore

$$\lim_{M^2\to 0} T_1 = \mathcal{N}\sum_{l=0}^{\infty} w^{2l} \left(C_1^{2l} A_{\tau=2}^{2l}\right) , \tag{5.42a}$$

$$\lim_{M^2\to 0} \frac{Q^2}{2M^2} T_2 = \mathcal{N}\sum_{l=0}^{\infty} w^{2l-2} \left(C_2^{2l} A_{\tau=2}^{2l}\right) , \tag{5.42b}$$

$$\lim_{M^2\to 0} \frac{Q^2}{2M^2} T_3 = \mathcal{N}\sum_{l=0}^{\infty} w^{2l-1} \left(C_3^{2l} A_{\tau=2}^{2l}\right) . \tag{5.42c}$$

This shows that the coefficient of the Laurent series of T_i's are just $C_i^k A_{\tau=2}^k$.

In the previous section, we derived the Laurent series of $T_{\mu\nu}$ as

$$T^{\mu\nu}(p,q) = -2\sum_{2n} W^{\mu\nu,2n}\, w^{2n} . \tag{5.43}$$

This implies that the Laurent series of T_i around $w = 0$ is given by

$$T_i(w) = -2\sum_{2n} W_i^{2n} w^{2n} , \tag{5.44}$$

where W_i^n is just the n-th Mellin moment of W_i. Up until now, $\mathcal{N}$ is completely arbitrary. Any value of $\mathcal{N}$ can be absorbed into C_i^n. For convenience, we set $\mathcal{N} = -2$. Therefore, by inserting (5.44) into (5.42), one obtains

$$C_1^n A_{\tau=2}^n = \lim_{M^2\to 0} W_1^n = F_1^n|_{M^2=0} , \tag{5.45a}$$

$$C_2^n A_{\tau=2}^n = \lim_{M^2\to 0} \frac{Q^2}{2M^2} W_2^{n-2} = F_2^{n-1}|_{M^2=0} , \tag{5.45b}$$

$$C_3^n A_{\tau=2}^n = \lim_{M^2\to 0} \frac{Q^2}{M^2} W_3^{n-1} = F_3^n|_{M^2=0} . \tag{5.45c}$$

From (5.44), (5.41), and (5.28), one can express the n-th Mellin moment of F as

$$F_1^{2l} = \sum_{j=0}^{\infty} \frac{(2l+j)!}{j!(2l)!} \left(\frac{M^2}{Q^2}\right)^j \left[F_1^{2l+2j}|_{M^2=0} + \frac{j\, F_2^{2l+2j-1}|_{M^2=0}}{(2l+2j)(2l+2j-1)}\right] , \tag{5.46a}$$

$$F_2^{2l-1} = \sum_{j=0}^{\infty} \frac{(2l+j)!}{j!(2l)!} \left(\frac{M^2}{Q^2}\right)^j \frac{(2l)(2l-1)}{(2l+2j)(2l+2j-1)} F_2^{2l+2j-1}|_{M^2=0} , \tag{5.46b}$$

$$F_3^{2l} = \sum_{j=0}^{\infty} \frac{(2l+j)!}{j!(2l)!} \left(\frac{M^2}{Q^2}\right)^j \frac{2l}{(2l+2j)} F_3^{2l+2j}|_{M^2=0} . \tag{5.46c}$$

At this point, the derivation is almost finished. We have successfully derived the Mellin moment of the full TMC structure functions in terms of the massless ones. The last step would be inverting the Mellin transforms to obtain the original TMC structure functions. Here, we will

demonstrate derivation only for F_2. Useful identities that will be extensively used are summarized here. For any integrable function $B(y)$ over $y \in [0,1]$ and $m \geq 0$, one has

$$\frac{1}{(m+1)} \int_0^1 dy\, y^{m+1}\, B(y) = \int_0^1 dy\, y^m\, H(y), \tag{5.47a}$$

$$\frac{1}{(m+2)(m+1)} \int_0^1 dy\, y^{(m+2)}\, B(y) = \int_0^1 dy\, y^m\, G(y), \tag{5.47b}$$

$$H(y) = \int_y^1 dy'\, B(y'), \quad \text{and} \tag{5.47c}$$

$$G(y) = \int_y^1 dy' \int_{y'}^1 dy''\, B(y'') = \int_y^1 dy'\, (y'-y)\, B(y')\,. \tag{5.47d}$$

Using (5.47b), one finds

$$\frac{F_2^{2l+2j-1}|_{M^2=0}}{(2l+2j)(2l+2j-1)} = \int_0^1 dy\, y^{2l+2j-2}\, g_2(y), \tag{5.48}$$

where

$$g_2(y) \equiv \int_y^1 dy' \int_{y'}^1 dy''\, (y'')^{-2}\, F_2^0(y'')\,. \tag{5.49}$$

Here, $F_2^{(0)}$ is the massless structure function, obtained by taking the limit $M^2 \to 0$. (5.46b) then can be rewritten as

$$F_2^{(2l-1)} = \int_0^1 dy\, y^{2l-2}\, g_2(y)\, (2l)(2l-1) \sum_{j=0}^{\infty} \frac{(2l+j)!}{j!(2l)!} \left(\frac{y^2M^2}{Q^2}\right)^j \tag{5.50}$$

$$= \int_0^1 dy\, y^{-1}\, g_2(y)\, \frac{(2l)(2l-1)\, y^{2l-1}}{(1-y^2M^2/Q^2)^{2l+1}}\,. \tag{5.51}$$

In the last step, we used the following identity :

$$\sum_{j=0}^{\infty} \frac{(n+j)!}{j!n!}\, z^j = \frac{1}{(1-z)^{(n+1)}}\,. \tag{5.52}$$

Then, the inverse Mellin transform can be evaluated as

$$F_2(x) = \frac{1}{2\pi i} \int_{c-i\infty}^{c+i\infty} dN\, x^{-N}\, \tilde{F}_2^{A\,N} \tag{5.53}$$

$$= \frac{x^2}{2\pi i} \frac{d^2}{dx^2} \int_0^1 dy\, \frac{g_2(y)}{y(1-y^2M^2/Q^2)^2} \int_{c-i\infty}^{c+i\infty} dN\, \frac{(y/x)^N}{(1-y^2M^2/Q^2)^N} \tag{5.54}$$

$$= x^2 \frac{d^2}{dx^2} \int_0^1 dy\, \frac{g_2(y)}{y(1-y^2M^2/Q^2)^2}\, \delta\left[\log\left(\frac{(y/x)}{(1-y^2M^2/Q^2)}\right)\right]\,. \tag{5.55}$$

The delta function of the log can be calculated as

$$\delta\left[\log\left(\frac{(y/x)}{(1-y^2M^2/Q^2)}\right)\right] = \left[\frac{y(1-y^2M^2/Q^2)}{(1+y^2M^2/Q^2)}\right]\,\delta(\xi_A - y)\,. \tag{5.56}$$

where

$$\xi = \frac{2x}{1+r}, \qquad r \equiv \sqrt{1+4x^2M^2/Q^2} = \frac{(1+\xi^2M^2/Q^2)}{(1-\xi^2M^2/Q^2)}\,, \tag{5.57}$$

The ξ variable is often called Nachtmann variable, which reduces to Bjorken x for $M = 0$. Finally :

$$F_2(x) = x^2\,\frac{d^2}{dx^2}\int_0^1 dy\,\frac{g_2(y)}{y(1-y^2M^2/Q^2)^2}\,\left[\frac{y(1-y^2M^2/Q^2)}{(1+y^2M^2/Q^2)}\right]\,\delta(\xi - y) \tag{5.58}$$

$$= x^2\,\frac{d^2}{dx^2}\,\left[\frac{(1+r)^2}{4r}g_2(y)\right]\,. \tag{5.59}$$

Upon inserting the functional form of g_2, one obtains the F_2 structure function with target mass correction included as a function of the massless one. To clearly indicate TMC structure functions, a superscript "TMC" will be attached.

Finally, the master formula for target mass-corrected structure functions F_j^{TMC} for $j = 1,\ldots,6$ reads [160, 161]:

$$F_1^{TMC}(x) = \frac{x}{\xi r}F_1^{(0)}(\xi) + \frac{M^2x^2}{Q^2r^2}\,h_2(\xi) + \frac{2M^4x^3}{Q^4r^3}\,g_2(\xi)\,, \tag{5.60a}$$

$$F_2^{TMC}(x) = \frac{x^2}{\xi^2r^3}F_2^{A,(0)}(\xi) + \frac{6M^2x^3}{Q^2r^4}\,h_2(\xi) + \frac{12M^4x^4}{Q^4r^5}\,g_2(\xi)\,, \tag{5.60b}$$

$$F_3^{TMC}(x) = \frac{x}{\xi r^2}F_3^{A,(0)}(\xi) + \frac{2M^2x^2}{Q^2r^3}\,h_3(\xi) + 0\,, \tag{5.60c}$$

where the functions $h_i(\xi,Q^2)$ and $g_2(\xi,Q^2)$ are given by the integrals

$$h_2(\xi,Q^2) = \int_\xi^1 du\,\frac{F_2^{(0)}(u,Q^2)}{u^2}, \tag{5.61a}$$

$$h_3(\xi,Q^2) = \int_\xi^1 du\,\frac{F_3^{(0)}(u,Q^2)}{u}, \tag{5.61b}$$

$$g_2(\xi,Q^2) = \int_\xi^1 du\,h_2(u,Q^2). \tag{5.61c}$$

Additionally, the target mass-corrected longitudinal structure function is given by

$$\begin{aligned} F_L^{TMC}(x) &= r^2 F_2^{TMC}(x) - 2x F_1^{TMC}(x) \\ &= \frac{x^2}{\xi^2 r}[F_2^{(0)}(\xi) - 2\xi F_1^{(0)}(\xi)] + \frac{4M^2x^3}{Q^2r^2}\, h_2(\xi) + \frac{8M^4x^4}{Q^4r^3}\, g_2(\xi) \\ &= \frac{x^2}{\xi^2 r} F_L^{(0)}(\xi) + \frac{4M^2x^3}{Q^2r^2}\, h_2(\xi) + \frac{8M^4x^4}{Q^4r^3}\, g_2(\xi)\,. \end{aligned} \tag{5.62}$$

Note that in the derivation of the master formula, one does not assume Callan-Gross relation between F_1 and F_2. In fact, the relation is simply violated in the presence of TMC. Furthermore, it is also violated if the heavy quark masses are included in the calculation of the massless structure functions. Thus, the leading term $F_L^{A,(0)}$ will be non-zero for finite quark masses, h_2, g_2 terms give the nucleon mass M contributions.

Regarding the master formula (5.71), a few remarks are in order. First, the argument of the TMC structure functions in the LHS is the Bjorken x, while the $F_i^{(0)}$, h_2 and g_2 are evaluated at the $x = \xi$. As $r \geq 1$, this implies $\xi \leq x$. At high x, $F_i^{(0)}$ is monotonically decreasing, therefore, $F_i^{(0)}(\xi)/F_i^{(0)}(x) \geq 1$ at high x. Thus, we expect that at high x, $F_i^{TMC}(x)$ is higher than the massless $F_i^{(0)}(x)$.

Looking at the master formula, it can be seen that $F_2^{(0)}$ appears in the computation of F_1^{TMC} and F_3^{TMC}. Actually, this is the result of contractions in $T_{2\mu\nu}$. After working out the contraction, $T_{2\mu\nu}$ contains Lorentz structures of W_1, W_2, W_4, W_5, W_6 as can be seen in (5.39), with coefficients that are proportional to $C_2^{2k} A_{\tau=2}^{2k} = F_2^{2k-1}$. Therefore, the resulting TMC formula for F_1, F_2, F_4, F_5, F_6 depend on $F_2^{(0)}$.

Finally, it is worth noting that the derivation of the master formula presented above was done without using perturbation theory, hence the formula is valid to all orders in pQCD. pQCD enters when calculating the massless structure functions, which are connected to $C_i^n A_{\tau=2}^n$ (hence T_i) through (5.45). It is the Wilson coefficients C_i^n that store the dependence on the order of perturbative calculation. They are in principle calculable using perturbation theory. C_i^n can be calculated, for example, by computing $T_{\mu\nu}$ in photon-quark scattering $\gamma^* + q \to \gamma^* + q$ (instead of $\gamma^* + N \to \gamma^* + N$) and match the OPE expansion. Once we compute C_i^n to some order of the strong coupling constant α_s, the resulting C_i^n is also valid for any external state. If the external state are the nucleon, then the reduce matrix elements A_τ^n encode the non-perturbative nature of the nucleon, hence A_τ^n are directly related to parton distribution functions. Using the notation of Refs. [160, 161], the master formula can be summarized in a single equation

$$F_j^{TMC}(x, Q^2) = \sum_{i=1}^{6} A_j^i F_i^{(0)}(\xi, Q^2) + B_j^i h_i(\xi, Q^2) + C_j g_2(\xi, Q^2)\,. \tag{5.63}$$

The coefficients A_j^i, B_j^i, C_j are the same that are given in Tables I,II,III in [160]. As all the perturbativity dependences are encoded in the computation of $F_i^{(0)}$, the coefficients A_j^i, B_j^i, C_j, once computed, are valid at all orders in pQCD.

5.1.2 Comparisons of F_i^{TMC}, $F_i^{leading}$, $F_i^{(0)}$ and F_i^{acot}

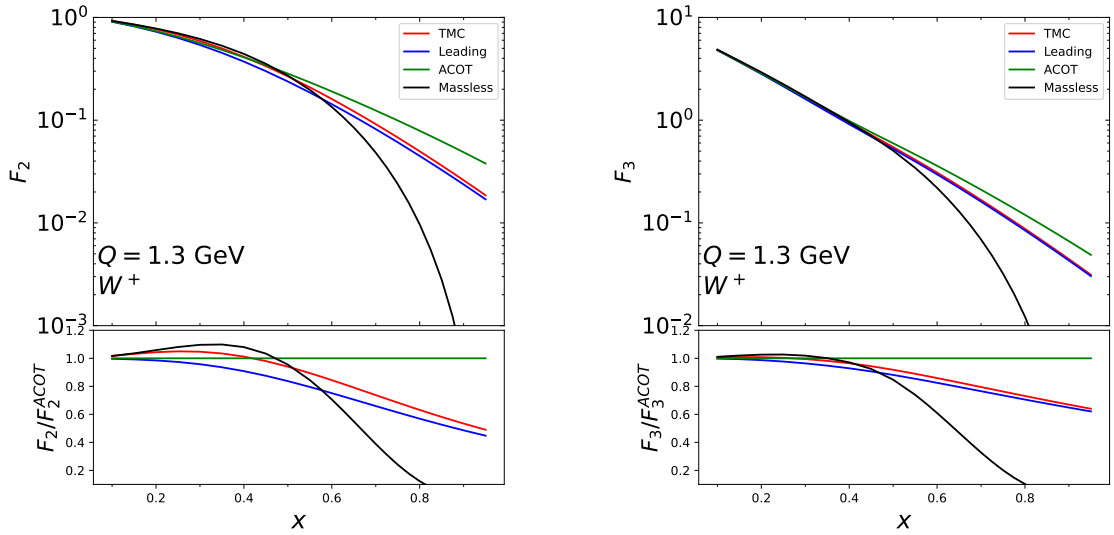


FIGURE 5.2: The F_2 (left) and F_3 (right) charged current (W^+-mediated) DIS structure functions calculated using different methods: TMC, leading TMC, ACOT and massless. The bottom panels shows the ratio to ACOT calculations. All calculations are done using proton PDFs from nCTEQ15 proton PDF baseline.

The original ACOT[91, 92, 153] formalism to compute the DIS structure functions was computed in the helicity basis within the framework of factorization and the parton model. Using the helicity basis has the advantage of having a boost-invariant polarization vector, making it easier to disentangle the partonic and hadronic structure functions. Furthermore, it elegantly encodes the dependence on heavy quark masses and target mass in the Wigner rotation matrix. As the ACOT method, which is based on the parton model, also includes target mass corrections, it is interesting to compare the resulting structure function calculation from OPE and the ACOT.

From the master formula (5.60), one can see that the biggest contributions to the calculated TMC SF comes from the first term in the RHS of (5.60). For an obvious reason, this term is called the leading TMC structure functions $F_i^{A,leading}$:

$$F_1^{leading}(x) = \frac{x}{\xi r} F_1^{(0)}(\xi)\,, \tag{5.64}$$

$$F_2^{leading}(x) = \frac{x^2}{\xi^2 r^3} F_2^{(0)}(\xi)\,, \tag{5.65}$$

$$F_3^{leading}(x) = \frac{x}{\xi r^2} F_3^{(0)}(\xi)\,. \tag{5.66}$$

To calculate these structure functions, we use the nCTEQ++ code that allows us to calculate structure functions at LO and NLO using ACOT scheme. To calculate massless structure functions $F_i^{(0)}(\xi)$ at $x = \xi$, the Nachtmann variable ξ is computed first for a given Bjorken x and hadron mass M. $F_i^{(0)}(\xi)$ is then calculated with M=0.

In top panels in Fig. 5.2, we show proton structure functions for W^+-mediated process, computed using master formula (5.60) (labelled as TMC in the figure), leading TMC (labelled as leading), ACOT-based calculation (labelled as ACOT), and massless calculation (labelled as massless). In the bottom panels, we show the ratio of these structure functions to the ACOT ones. One can see that the difference between the massless and the other methods are dramatic, especially in the larger x region. Comparing the full TMC and the leading calculations, one can see that they are quite similar for all x. The difference between these two and the ACOT is $\lesssim 20\%$ for $x \lesssim 0.6$ at $Q = 1.3$.

5.2 Master Formula for a Nucleus Target

In this section, the TMC master formula for the DIS structure functions in $l + A \to l' + X$ collisions will be given. As now the target hadron is a nucleus, with mass significantly larger than the proton mass, it is interesting to know how the TMC structure functions scale with A(the mass number of the nucleus). This will be crucial for nPDF analysis, as many DIS nuclear data are included, with some of them have data points that goes to very high x and low Q^2 at the same time.

In the lepton-nucleus DIS, let p_A denotes the momentum of the nucleus. Then the Bjorken variable is given by $x_A = Q^2/(2p_A.q)$. Requiring the invariant mass of the hadronic final states $p_X^2 \geq M_A^2$, with M_A is the nucleus mass, giving us $x_A \leq 1$, just like in the lepton-nucleon DIS. In term of x_A, the hadronic invariant mass can be written as[3]

$$\tilde{W}_A^2 = M_A^2 + Q^2 \left(\frac{1 - x_A}{x_A} \right) \tag{5.67}$$

At this point, all results for the lepton-nucleon DIS derived in the previous section are valid also for the lepton-nucleus case, with M and x are replaced with M_A and x_A. However, some care needs to be given when trying to justify the use of OPE in the lepton-nucleus case. As discussed earlier, the OPE expansion is valid when two local operators are close to each other $z \to 0$. In the lepton-nucleon DIS, the local operators are the currents $J_\mu(z)$ and $J_\nu(0)$. It has been shown that in the DIS limit, the light cone region $z^2 \to 0$ gives the most dominating contribution to the hadronic tensor $W_{\mu\nu}$. For lepton-nucleus case, we also expect the same light-cone dominance, as the constant c and d in (5.16) should be the same in the nuclear case. In the actual OPE of $T_{\mu\nu}$, one assumes $Q^2, \nu_A^2 \to \infty$, while keeping Q^2/ν_A^2 fixed. Here,

$$\nu_A = \frac{p_A.q}{M_A} \tag{5.68}$$

In the frame of target nucleus, $\nu_A = (E_l - E_l')$, hence it does not depend on A. Similarly also for $Q^2 = -(k - k')^2$. Thus, the ratio Q_A^2/ν_A^2 only depends on the incoming and outgoing lepton, and therefore does not depend on type of the target. As the OPE in nuclear case is identical to

[3]The use of tilde in $\tilde{W}_A^2$ shows that this is a total (unrescaled) quantity. All rescaled quantities will be denoted without tilde.

that in the nucleon case, the TMC master formula for lepton-nucleus DIS is identical:

$$\tilde{F}_1^{A,TMC}(x_A) = \frac{x_A}{\xi_A r_A}\tilde{F}_1^{A,(0)}(\xi_A) + \frac{M_A^2 x_A^2}{Q^2 r_A^2}\,\tilde{h}_2^A(\xi_A) + \frac{2M_A^4 x_A^3}{Q^4 r_A^3}\,\tilde{g}_2^A(\xi_A)\,, \tag{5.69a}$$

$$\tilde{F}_2^{A,TMC}(x_A) = \frac{x_A^2}{\xi_A^2 r_A^3}\tilde{F}_2^{A,(0)}(\xi_A) + \frac{6M_A^2 x_A^3}{Q^2 r_A^4}\,\tilde{h}_2^A(\xi_A) + \frac{12M_A^4 x_A^4}{Q^4 r_A^5}\,\tilde{g}_2^A(\xi_A)\,, \tag{5.69b}$$

$$\tilde{F}_3^{A,TMC}(x_A) = \frac{x_A}{\xi_A r_A^2}\tilde{F}_3^{A,(0)}(\xi_A) + \frac{2M_A^2 x_A^2}{Q^2 r_A^3}\,\tilde{h}_3^A(\xi_A) + 0\,, \tag{5.69c}$$

where the functions $\tilde{h}_i^A(\xi_A, Q^2)$ and $\tilde{g}_2^A(\xi_A, Q^2)$ are given by the integrals

$$\tilde{h}_2^A(\xi_A, Q^2) = \int_{\xi_A}^1 du_A\,\frac{\tilde{F}_2^{A,(0)}(u_A, Q^2)}{u_A^2}, \tag{5.70a}$$

$$\tilde{h}_3^A(\xi_A, Q^2) = \int_{\xi_A}^1 du_A\,\frac{\tilde{F}_3^{A,(0)}(u_A, Q^2)}{u_A}, \tag{5.70b}$$

$$\tilde{g}_2^A(\xi_A, Q^2) = \int_{\xi_A}^1 du_A\,\tilde{h}_2^A(u_A, Q^2). \tag{5.70c}$$

When studying lepton-nucleus DIS, it is useful to rescale the structure functions with the mass number of the nucleus. This rescaling has been discussed in Section 2.6.1. In terms of the rescaled structure functions $F_i^A(x_N)$, where $x_N = Ax_A$, $M_N = M_A/A$, the master formula becomes

$$F_1^{A,TMC}(x_N) = \frac{x_N}{\xi_N r_N}F_1^{A,(0)}(\xi_N) + \frac{M_N^2 x_N^2}{Q^2 r_N^2}\,h_2^A(\xi_N) + \frac{2M_N^4 x_N^3}{Q^4 r_N^3}\,g_2^A(\xi_N)\,, \tag{5.71a}$$

$$F_2^{A,TMC}(x_N) = \frac{x_N^2}{\xi_N^2 r_N^3}F_2^{A,(0)}(\xi_N) + \frac{6M_N^2 x_N^3}{Q^2 r_N^4}\,h_2^A(\xi_N) + \frac{12M_N^4 x_N^4}{Q^4 r_N^5}\,g_2^A(\xi_N)\,, \tag{5.71b}$$

$$F_3^{A,TMC}(x_N) = \frac{x_N}{\xi_N r_N^2}F_3^{A,(0)}(\xi_N) + \frac{2M_N^2 x_N^2}{Q^2 r_N^3}\,h_3^A(\xi_N) + 0\,, \tag{5.71c}$$

where :

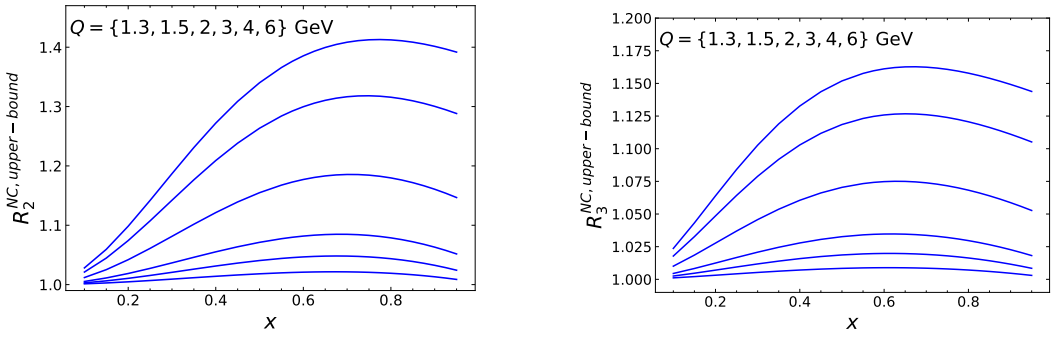
$$\xi_N = \frac{2x_N}{1+r_N}, \qquad r_N \equiv \sqrt{1+4x_N^2 M_N^2/Q^2} = \frac{(1+\xi_N^2 M_N^2/Q^2)}{(1-\xi_N^2 M_N^2/Q^2)}\,, \tag{5.72}$$

and the functions $h_i^A(\xi_N, Q^2)$ and $g_2^A(\xi_N, Q^2)$ are given by the integrals

$$h_2^A(\xi_N, Q^2) = \int_{\xi_N}^A du_N\,\frac{F_2^{A,(0)}(u_N, Q^2)}{u_N^2}, \tag{5.73a}$$

$$h_3^A(\xi_N, Q^2) = \int_{\xi_N}^A du_N\,\frac{F_3^{A,(0)}(u_N, Q^2)}{u_N}, \tag{5.73b}$$

$$g_2^A(\xi_N, Q^2) = \int_{\xi_N}^A du_N\,h_2^A(u_N, Q^2). \tag{5.73c}$$



(a) (b)

FIGURE 5.3: The upper bounds of $R_i \equiv F_i^{TMC}/F_i^{TMC-Leading}$ for a) F_2 and b) F_3 for selected $Q = \{1.3, 1.5, 2, 3, 4, 6\}$ GeV, (from top to bottom).

As a remark, note that the master formula (5.71) is very similar to (5.69) and (5.60). This a consequence of the rescaling : $x_A M_A = x_N M_N$, $x_A/\xi_A = x_N/\xi_N$, $r_N = r_A$. The upper limit of the integrals in (5.73) is $u_N = A$. In practice, however, $F_i^{A,(0)}(x_N)$ falls off very quickly for $x \geq 1$, therefore it is zero for $x_N \geq 1$. Thus, the upper limit of the integrals is effectively $u_N = 1$. Finally, nPDFs are often written as an average of bound proton and neutron PDFs :

$$f_k^A = \frac{Z}{A} f_k^{p/A} + \frac{N}{A} f_k^{n/A}, \tag{5.74}$$

where k represent parton flavor, Z and N denotes the atomic and neutron number respectively. The bound nucleon PDFs, which usually differ from the free nucleon ones by $\lesssim 20\%$ for $x_N \leq 0.8$, are the ones that are fitted to the data in an nPDF global analysis. As $F_i^{A,(0)}$ are linear functionals of f_i^A, this means

$$F_i^{A,(0)} = \frac{Z}{A} F_i^{p/A,(0)} + \frac{N}{A} F_i^{n/A,(0)}, \tag{5.75}$$

where $F_i^{p,n/A,(0)}$ are the massless structure functions computed using bound proton (neutron) PDFs. Due to the averaging (5.75), one expects that $F_i^{A,(0)}$ varies rather slowly with A, with the envelop given by the $F_i^{p/A,(0)}$ and $F_i^{n/A,(0)}$. As a result, we also expect the same for the full TMC structure functions. We will explore this universality in the following section.

5.3 TMCs for Various Nuclear Targets

As one of the main focuses of this study, the impact of TMCs on structure functions for various nuclei will be examined. First, let's see how structure functions with TMC applied differ to the massless ones. In Fig. 5.4, we show the ratio of $F_i^{A,TMC}/F_i^{A,(0)}$ computed using nCTEQ15 nPDFs for various nuclei (see Tab. 5.1) and for $Q = 1.3, 3, 6$ GeV. In the bottom panel of the figure, we show the spread of the curves at $Q = 1.3$ GeV . In the figure, the dashed and dotted lines correspond to proton and neutron respectively. We can see form the figure that the ratio

Symbol	A	Z	Symbol	A	Z	Symbol	A	Z	Symbol	A	Z
H	1	1	Be	9	4	Ca	40	20	Xe	131	54
D	2	1	C	12	6	Fe	56	26	W	184	74
^{3}He	3	2	N	14	7	Cu $_{\text{iso}}$	64	32	Au	197	79
He	4	2	Ne	20	10	Kr $_{\text{iso}}$	84	42	Au $_{\text{iso}}$	197	98.5
Li	6	3	Al	27	13	Ag $_{\text{iso}}$	108	54	Pb $_{\text{iso}}$	207	103.5
Li	7	3	Ar	40	18	Sn $_{\text{iso}}$	119	59.5	Pb	208	82

TABLE 5.1: List of nuclei considered in this work. The nuclei indicated with "iso" subscript are isoscalar nuclei; thus, the Z value can be half-integer.

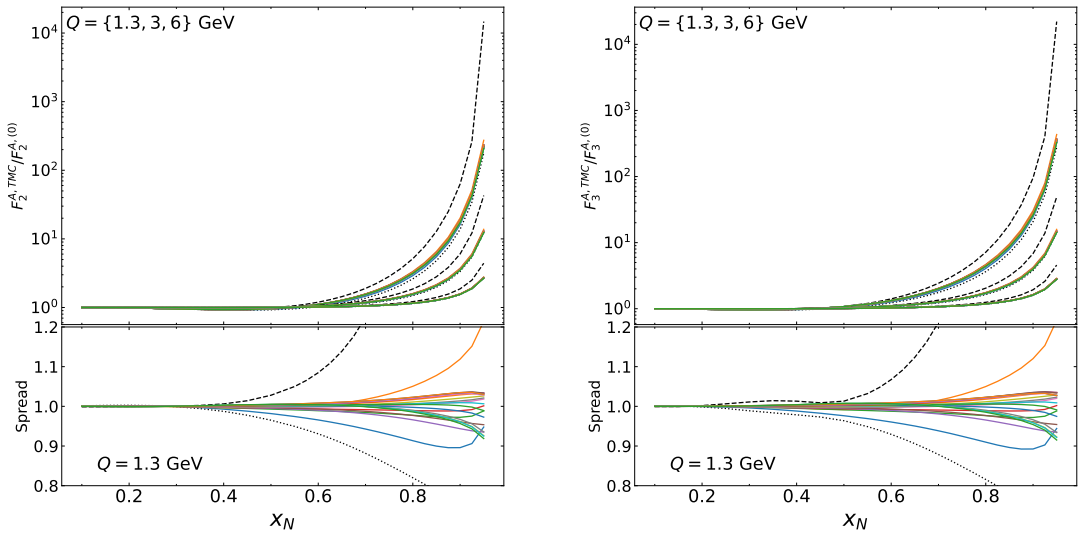


FIGURE 5.4: (Top panels) The ratio of (W^+-mediated) DIS structure function $F_i^{A,TMC}/F_i^{A,(0)}$, $i = 2$ (left) and $i = 3$ (right), for various nuclei. (Bottom panels) the spread of the ratios around their averages.

quickly deviates from unity as we go to higher x_N. Furthermore, for the same x_N, the deviation from unity is higher as Q decreases.

Note here that the argument of $F_i^{A,leading}$ is the nucleonic Bjorken x_N, while terms on the RHS are evaluated at ξ_N. Due to the suppression M^2/Q^2 for the h-term and M^4/Q^4 in the g term, $F_i^{A,leading}$ is supposed to capture most of the target mass effects. As it is proportional to $F_i^{A,(0)}(\xi_N)$, it is just as fast to calculate as in $F_i^{A,(0)}(\xi_N)$. As mentioned in the previous section, $F_i^{A,TMC}/F_i^{A,(0)}$ varied slowly with respect to A. As $F_i^{A,leading}$ already encodes most of the TMC effects, the ratio $R_i \equiv F_i^{A,TMC}/F_i^{A,leading}$ should be even more universal in A. In fact, this can be seen from the upper bound for R_i, which can be computed by assuming structure functions are monotonically decreasing (an entirely reasonable assumption in the large x region). Using

this assumption, it is easy to derive the upper bound for R_i as [161]:

$$R_2 \equiv \frac{F_2^{A,TMC}}{F_2^{A,leading}}(x_N, Q^2) \;\leq\; 1 + \left(\frac{M_N}{Q}\right)^2 \frac{6x_N\xi_N}{r_N}(1-\xi) + \left(\frac{M_N}{Q}\right)^4 \frac{12x_N^2\xi_N^2}{r_N^2}(-\ln\xi_N - 1 + \xi)$$
$$R_3 \equiv \frac{F_3^{A,TMC}}{F_3^{A,leading}}(x_N, Q^2) \;\leq\; 1 - \left(\frac{M_N}{Q}\right)^2 \frac{2x_N\xi_N}{r_N}\ln\xi_N \quad .$$

Note that there is no PDF dependence on the RHS. Here, we can see explicitly the powers of (M/Q) which drive the ratios to unity for large Q. In Fig. 5.3, we plot these bounds as a function of x_N for selected Q values. One can see that the upper bounds are quite conservative, reaching as high as $R_2 = 1.4$ for $Q = 1.3$ GeV. Keep in mind that the upper bounds are valid for all $F_i^{A,(0)}$ shapes (as long as they are monotonically decreasing), and yet, the differences between $F_2^{A,TMC}$ and $F_2^{A,leading}$ are only 40% at most. This remarkable result suggest that R_i is essentially independent of nuclei given the fact that the nPDFs for these nuclei does not differ much with each other.

To show explicitly the universality of R_i under nuclei variation, in Fig. 5.5, we show the ratio R_i for various nuclei in the top panels. One can see that, aside from an extreme non-isoscalar nuclei such as neutron and proton, the R_i curves seems to be universal for all nuclei. To better see the universality, in the bottom panels, we show the spread around the average R_i^{ave}, defined as:

$$R_i^{ave}(x_N, Q^2) \;=\; \frac{1}{n_A}\sum_k^{n_A} \frac{F_i^{A_k,\mathrm{TMC}}(x_N, Q^2)}{F_i^{A_k,leading}(x_N, Q^2)} \,. \tag{5.76}$$

One can see that, aside from neutron and proton curves, the ones from all other nuclei spread with less than 0.5% around the average at $Q = 1.3$. Note that the spread will be smaller as we go to higher Q. At $Q = 2.0$ GeV, the spread will be less than 0.1%. Looking at the proton and neutron curves, we can see that they almost always on the opposite ends. As explained in the previous section, this is a consequence of the averaging $f_i^A = Z/A f_i^{p/A} + N/A f_i^{n/A}$ and the fact that these structure functions receive major contributions from either u or d quarks. As an example, at high x, we have $F_2(\gamma/Z) \sim x_N/9(4u(x_N) + d(x_N))$, $F_2(W^-) \sim 2xu(x_N)$, $F_2(W^+) \sim 2xd(x)$ as we can neglect the gluon and sea quarks. As $d/u \ll 1$ in the large x region, this explains why the proton F_2 curve from γ/Z is higher than the one from neutron. Patterns happen in all other F_i can also be explained similarly.

Given that the ratio R_i's are almost independent of nuclei, it is then useful to have a parameterization that capture R_i-shape without needing to calculate the time consuming convolution integrals as in the master formula. The parameterization is beneficial when we need to calculate the TMC structure functions many times during nPDF fitting. In the following section, we will discuss how such parameterization can be constructed.

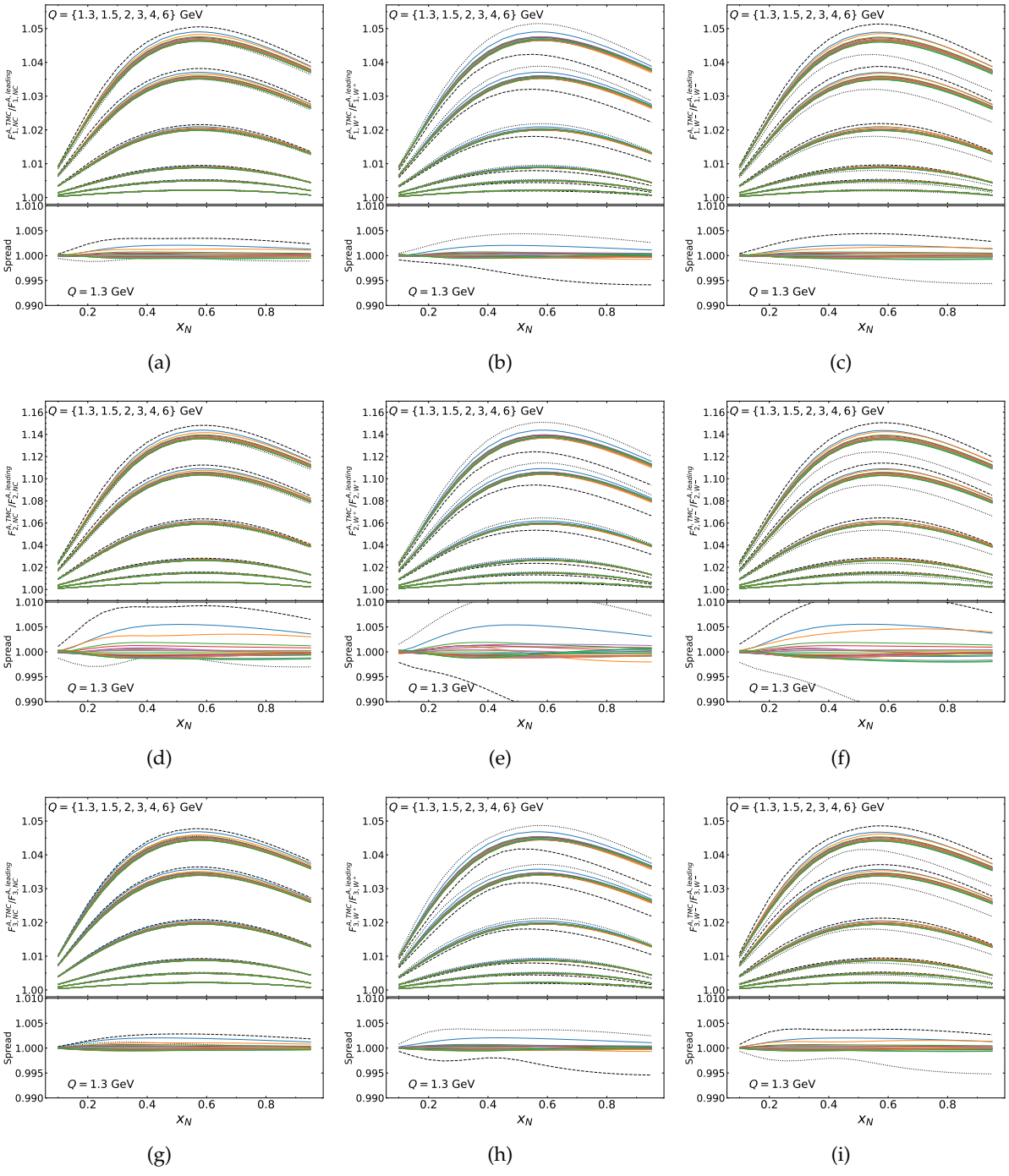


FIGURE 5.5: (top panel) The nuclear variation of the ratio $F_i^{A,TMC}/F_i^{A,TMC-Leading}$ for CC and NC processes for $Q = \{1.3, 1.5, 2, 3, 4, 6\}$ GeV (from top to bottom) calculated using nCTEQ15 nPDFs. (Bottom panel) the spread of the nuclear variation around the average Eq. (5.76) for $Q = 1.3$ GeV. In all panels, proton lines as shown as dashed black line, while the neutron lines are shown as dotted black lines. $F_i^{TMC}/F_i^{TMC-Leading}$, $i = 1, 2, 3$ are shown in top, middle and bottom rows respectively. The first columns shows $F_i^{TMC}/F_i^{TMC-Leading}$ for DIS process with γ and Z exchange. Similarly, the second and third columns shows the ratios for DIS process with W^+ and W^- exchanges respectively.

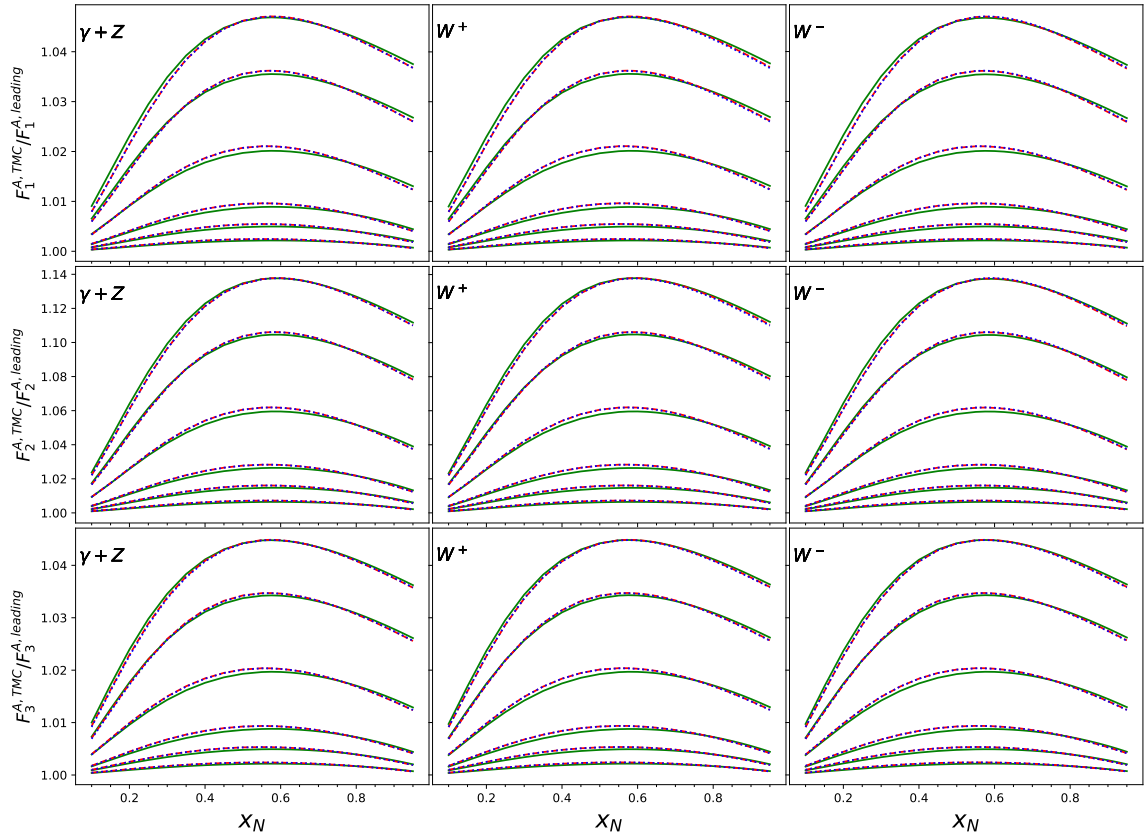


FIGURE 5.6: The comparison between the average full/leading TMC ratios at Q= {1.3, 1.5, 2, 3, 4, 6} GeV (from top to bottom) using the nCTEQ15 nPDFs (solid green) and the parameterizations based on Eqs. (5.87) (red-dashed) with parameters given in Table 5.2.

5.4 Parameterizing $F_a^{TMC}/F_a^{leading}$

In the previous section, we demonstrated that the full/leading TMC ratios were effectively insensitive to the nuclear A value. In this section, we discuss how to parameterize the R_i ratios. Starting from Eqs (5.69), we divide $F_i^{A,TMC}(x)$ by the leading term $F_i^{A,leading}(x)$:

$$\frac{F_1^{A,TMC}(x_N)}{F_1^{A,leading}(x_N)} = 1 + \frac{M_N^2}{Q^2}\frac{x_N\xi_N}{r_N}\frac{h_2^A(\xi_N)}{F_1^{A,(0)}(\xi_N)} + \frac{M_N^4}{Q^4}\frac{2x_N^2\xi_N}{r_N^2}\frac{g_2^A(\xi_N)}{F_1^{(0)}(\xi_N)}, \tag{5.77}$$

$$\frac{F_2^{A,TMC}(x_N)}{F_2^{A,leading}(x_N)} = 1 + \frac{M_N^2}{Q^2}\frac{6x_N\xi_N^2}{r_N}\frac{h_2^A(\xi_N)}{F_2^{A,(0)}(\xi_N)} + \frac{M_N^4}{Q^4}\frac{12x_N^2\xi_N^2}{r_N^2}\frac{g_2^A(\xi_N)}{F_2^{A,(0)}(\xi_N)}, \tag{5.78}$$

$$\frac{F_3^{A,TMC}(x_N)}{F_3^{A,leading}(x_N)} = 1 + \frac{M_N^2}{Q^2}\frac{2\xi_N x_N}{r_N}\frac{h_3^A(\xi_N)}{F_3^{A,(0)}(\xi_N)} \quad . \tag{5.79}$$

To parameterize these ratios, in principle one can use any universal function approximators, such as polynomial-based parameterization (Legendre, Cebyshev, Bernstein polynomials...), neural network, and many others. However, to minimize the number of open parameters, one

can utilize some physical assumptions and simpler parameterizations can be obtained, with only two open parameters needed for each structure function. First, we assume the structure functions $F_a^{A,(0)}(x)$, $a = 1, 2, 3$, vanish at $x = 1$. This assumption follows from the fact that most nPDFs vanish at $x = 1$, which is a consequence of $(1-x)^b$ factor appearing in the parameterizations. Next, we use a finite difference formula to approximate the derivatives of $F_a^{A,(0)}(x = \xi)$ as

$$F_i^{'A,(0)}(u = \xi_N) \approx -\frac{\gamma_i F_i^{A,(0)}(\xi_N)}{1-\xi_N}\,, \tag{5.80}$$

$$F_i^{j,A,(0)}(u = \xi_N) \approx \frac{0 - F_i^{A,j-1,(0)}(\xi_N)}{(1-\xi_N)} = (-1)^j \frac{\gamma_i F_i^{A,(0)}(\xi_N)}{(1-\xi_N)^j} \quad . \tag{5.81}$$

Here, $\gamma_i(x)$ is a universal (A-independent) correction factor for the first derivative which we assume to have a mild x dependence. We can then expand $F_i^{A,(0)}(u)$ about $u = \xi_N$ as

$$F_i^{A,(0)}(u) = \sum_{j=0}^{\infty} \frac{1}{j!} F_i^{j,A,(0)}(\xi_N)(u-\xi_N)^j \tag{5.82}$$

$$\approx F_i^{A,(0)}(\xi_N) \left(1 + \sum_{j=1}^{\infty} \frac{1}{j!}(-1)^j \frac{\gamma_i}{(1-\xi_N)^j}(u-\xi_N)^j\right) \tag{5.83}$$

$$\equiv F_i^{A,(0)}(\xi)\, K_i(u, \xi, \gamma_i) \quad . \tag{5.84}$$

As $K_i(u, \xi_N, \gamma_i)$ is independent of A, this implies that the ratios

$$\frac{h_i^A(\xi)}{F_i^{A,(0)}(\xi_N)} = \int_{\xi_N}^1 L_i(u) K_i(\xi_N, x_N, \gamma_i) du\,, \tag{5.85}$$

$$\frac{g_2^A(\xi)}{F_2^{A,(0)}(\xi_N)} = \int_{\xi_N}^1 \frac{u-\xi_N}{u^2} K_i(\xi_N, x_N, \gamma_i) du\,. \tag{5.86}$$

are also independent A. Here we have defined $L_1(u) = 2/u$, $L_2(u) = 1/u^2$, and $L_3(u) = 1/u$. Evaluating the explicit expression for $K_i(u, \xi_N, \gamma_i)$ in the above relations and assuming the the Callan-Gross relation $F_L = F_2 - 2xF_1$ (deviation from Callan-Gross relation can be absorbed by the fitted γ_i), we obtain :

$$\frac{h_2(\xi)}{F_2^{(0)}(\xi)} = \frac{1-\xi}{\xi} + \gamma_2(\xi)\frac{1-\xi}{\xi^2} \sum_{j=1}^{j_{\max}} \frac{(-1)^j}{j!(j+1)}\, {}_2F_1\left(2, j+1, j+2, 1-\frac{1}{\xi}\right)\,, \tag{5.87a}$$

$$\frac{h_3(\xi)}{F_3^{(0)}(\xi)} = -\ln(\xi) + \gamma_3(\xi)\frac{1-\xi}{\xi} \sum_{j=1}^{j_{\max}} \frac{(-1)^j}{j!(j+1)}\, {}_2F_1\left(1, j+1, j+2, 1-\frac{1}{\xi}\right)\,, \tag{5.87b}$$

$$\frac{h_2(\xi)}{F_1^{(0)}(\xi)} = 2\xi\left[\frac{1-\xi}{\xi} + \gamma_1(\xi)\frac{1-\xi}{\xi^2} \sum_{j=1}^{j_{\max}} \frac{(-1)^j}{j!(j+1)}\, {}_2F_1\left(2, j+1, j+2, 1-\frac{1}{\xi}\right)\right]\,, \tag{5.87c}$$

	$F_1^{TMC}/F_1^{leading}$		$F_2^{TMC}/F_2^{leading}$		$F_3^{TMC}/F_3^{leading}$	
nPDFs	λ_1	δ_1	λ_2	δ_2	λ_3	δ_3
nCTEQ15	2.352	-0.122	2.264	-0.074	2.090	0.035
EPPS16	2.222	-0.080	2.135	-0.032	2.007	0.059
nNNPDF2.0	2.240	-0.095	2.152	-0.046	2.094	0.041
TUJU19	2.441	-0.156	2.355	-0.110	2.123	0.024

TABLE 5.2: The values of λ and δ parameters that parameterize the ratio $F_i^{A,TMC}/F_i^{A,leading}$ for each structure function type. The $\{\lambda_i, \delta_i\}$ parameters are independent of the exchanged boson ($\gamma, Z, W^{\pm}$), and are relatively insensitive to the specific underlying nPDF.

$$\frac{g_2(\xi)}{F_2^{(0)}(\xi)} = -\ln(\xi) - (1-\xi) + \gamma_2(\xi)\frac{(1-\xi)^2}{\xi^2}\sum_{j=1}^{j_{\max}}\frac{(-1)^j}{j!(j+2)}\,{}_2F_1\left(2, j+2, j+3, 1-\frac{1}{\xi}\right), \quad (5.87d)$$

$$\frac{g_2(\xi)}{F_1^{(0)}(\xi)} = 2\xi\left[-\ln(\xi) - (1-\xi) + \gamma_1(\xi)\frac{(1-\xi)^2}{\xi^2}\sum_{j=1}^{j_{\max}}\frac{(-1)^j}{j!(j+2)}\,{}_2F_1\left(2, j+2, j+3, 1-\frac{1}{\xi}\right)\right], \quad (5.87e)$$

Here, ${}_2F_1(a,b,c,z)$ is a hypergeometric function. Although the summations over the index j can, in principle, go to infinity, we can truncate the series as they converge quickly due to $1/(j!)$ prefactor. Setting $j_{\max} = 4$ is sufficient to reproduce the exact results. The slowly varying γ_i can be parameterized as :

$$\gamma_i(\xi_N) = \lambda_i + \frac{\delta_i}{\ln(1+\xi_N)} \quad . \quad (5.88)$$

The values of $\{\lambda_i, \delta_i\}$ are obtained by fitting the parameterization to the exact results for each structure function type. Note that the values of $\{\lambda_a, \delta_a\}$ are independent of the type of the exchanged bosons. In Table 5.2, we show the values of $\{\lambda_i, \delta_i\}$ obtained by fitting the parameterizations to the exact results which are computed using nCTEQ15[25], EPPS16[80], nNNPDF2.0[26] and TUJU19[149]. We can see that the fitted parameters are largely insensitive to the specific nPDF sets.

In Figure 5.6, we show the comparison between our parametrization with the exact results obtained using nCTEQ15 nPDFs. We can see that our parameterization works very well to reproduce the exact results with $\lesssim 0.2\%$ agreement level. In the appendix A, we show the comparison between the exact and parameterized results for EPPS21, nNNPDF2.0, and TUJU19 nPDFs. With the parameter values given by table 5.2, similar agreement levels are obtained.

5.5 Summary

Lepton-nucleus DIS is one of the backbones of nuclear PDF analysis. As the precision and the kinematic coverage of the DIS data have improved, it is important to include all sources of corrections that enters into pQCD calculations. TMC is one of those which is significant at

high x and low Q^2. In this work, TMC master formula for lepton-nucleon DIS was derived using OPE formalism. Extending to lepton-nucleus case is straightforward due to rescaling. The main finding of this study is that the ratio $F_i^{A,TMC}/F_i^{A,leaidng}$ is fairly independent of the nucleus A, therefore, open a possibility to parameterize the ratio. The parameterization allows us to calculate TMC structure functions without evaluating the time-consuming convolutions in (5.69), hence will be useful for nPDF fitting. The parameterization is given in (5.87), which can be derived using basic properties of massless structure functions at large x. By fitting two parameters (λ and δ) for each structure function type to the exact results, we obtain an impressive agreement at $\lesssim 0.2\%$ level with nCTEQ15 nPDFs.

Chapter 6

Global Analysis with the CMS Dijet Data

The dijet (a pair of the two most energetic jets) production data from proton-lead collisions has been known to provide strong constraints on the nuclear gluon PDFs[166]. Compared to the dijet production in a lead-lead collision system, the final state effects from the quark-gluon plasma are negligible and hence it is perfect for nPDF fits. Furthermore, if the ratio to the proton-proton reference spectra is taken, scale uncertainties from missing higher order terms can also be minimized. In this chapter, we investigate the viability of including the dijet production data[167] from CMS, measured in proton-lead collisions at $\sqrt{s} = 5$ TeV. We start by discussing the pp data and compare the predictions from CJ15 and CT18 NLO PDFs. We then discuss the pPb spectra and the spectra ratio pPb/pp data. In the appendix B, we have set up a reference fit (the HIXNEU-CJ2 fit), to which the pPb and pPb/pp dijet data will be compared. This HIXNEU-CJ2 fit represents a global analysis that includes data sets used in the nCTEQ15HIX[60] and BaseDimuChorus[72] analyses, along with some improvements in the fitting methodology, such as the use of the CJ15 NLO PDFs as the new proton PDF baseline, new nPDF parameterizations, and better treatments of target mass, higher twist and deuteron corrections. At the end, we extend the HIXNEU-CJ2 fit to include the dijet pPb/pp data and discuss the results.

6.1 The CMS Dijet Data

The pPb and pp dijet data from the CMS[167] experiment is currently the most recent nuclear jet data from LHC. In this experiment, the center of mass energy is $\sqrt{s_{NN}} = 5.02$ TeV with the corresponding integrated luminosities of $35 \pm 1\ nb^{-1}$ and 27.4 ± 0.6 pb^{-1} respectively. The data is binned in $p_T^{ave} = (p_{T,1} + p_{T,2})/2$ and $\eta_{dijet} = (\eta_1 + \eta_2)/2$, where $p_{T,i}$ and η_i are the transverse momentum and the pseudorapidity of the i-th jet respectively. Jets are reconstructed using the anti-k_T clustering algorithm with jet radius $R = 0.3$. For the p-Pb system, only jets with $|\eta_{lab}| < 3.0$ are selected.

As said, the importance of the dijet data is that it provides a strong constraint to gluon PDFs, not only at low x ($x \approx 0.001$), but also at medium and high x. From the leading order

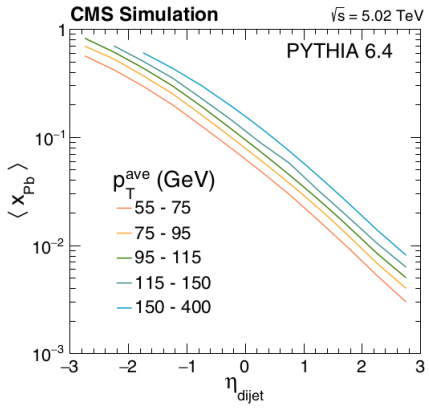


FIGURE 6.1: The average of the probed lead's partonic momentum fraction from the dijet productions at CMS. The figure is taken from [167].

kinematics for dijet production in the p-Pb system, we have

$$\eta_{dijet} = \frac{1}{2} \ln \left(\frac{x_p}{x_{Pb}} \right) , \tag{6.1}$$

where x_p and x_{Pb} are the momentum fractions of participating partons from proton and lead. To really study the sensitivity of the dijet data to a specific region of x_{Pb}, one needs to disentangle x_{Pb} from the ratio in (6.1). For this, one could perform a simulation study with a Monte-Carlo generator such as Pythia. With Pythia, it is possible to turn on initial (ISR) and final (FSR) state radiations to see the impact of gluon radiations to the probed parton momentum. Fig. 6.1, which was taken from [167], shows the average x_{Pb} probed as function of η_{dijet} for each p_T^{ave} bin. One can see that the probed x_{Pb} goes from as low as $x_{Pb} = 0.003$ to as high as $x_{Pb} = 0.9$. The shadowing, anti-shadowing and EMC effect region are probed by the dijet events with $\eta_{dijet} \gtrsim 1.5$, $-0.5 \lesssim \eta_{dijet} \lesssim 1.5$, and $\eta_{dijet} \lesssim -0.5$ respectively.

The dijet production data from CMS is measured in terms of dijet pseudorapidity spectra, defined as $1/N^{ij}_{dijet} dN_{dijet}/d\eta_{dijet}$. Here, $dN_{dijet}/d\eta_{dijet}$ represents the number of dijet events that fall within the specific ΔP_T^{dijet} and $\Delta\eta_{dijet}$, divided by $\Delta\eta_{dijet}$. N_{dijet} represents the number of all dijet events falling in ΔP_T^{dijet} with no restriction in η_{dijet}. The theoretical prediction for the spectra, in practice, is calculated by computing ratio of cross sections as follows:

$$\frac{1}{N^{ij}_{dijet}} \frac{dN^{ij}_{dijet}}{d\eta_{dijet,j}} = \left[\sum_k \frac{d\sigma_{ik}}{d\eta_{dijet,k}} \Delta\eta_{dijet,k} \right]^{-1} \frac{d\sigma_{ij}}{d\eta_{dijet,j}} . \tag{6.2}$$

The theory prediction for the dijet production can be computed using NLOJET++ program[168], which is based on Catani-Seymour substraction method[169] and interfaced with FASTJET[170] for the jet clustering. For fast calculation, one can use a gridding technique such as the one implemented in FASTNLO[171] or APPLGRID[172]. This gridding technique allows us to store the convolution of PDFs with the Wilson coefficients ahead of time and the cross section is computed by a simple sum. To calculate theory predictions for the CMS dijet data, we

use FASTNLO instead of APPLGRID. As by default FASTNLO does not support different types of hadron in the initial state, we modified some part of the code and integrated the modified FASTNLO code into our code base.

Another program capable of calculating jet production cross sections is NNLOJET[173]. It is a Monte Carlo parton-level event generator that handles divergences using antenna substraction method[174]. It provides calculations of various jet production processes at both NLO and NNLO. To make sure that our NLOJET++ setup and calculations are correct, we benchmarked the calculated dijet spectra with the result from NLOJET++ calculation. We found that both theory predictions generally agree well within the Monte Carlo uncertainties of NNLOJET calculations. We also benchmarked the integrated cross section $d\sigma/d\eta_{dijet}$, and we found good agreement within less than 1‰.

6.2 The pp Dijet Data

For nPDF fitting, the pp dijet data is very important as it serves as a basis to which the pPb data will be compared. As we do not fit proton PDFs, but rather fix them to some specific PDFs taken from other analysis, it is essential to check if the theory predictions from the proton PDF baseline can well-describe the data.

To see if the theory predictions from modern proton PDFs, such as the ones from CT18 and CJ15, can describe the data, in Fig. 6.2, we show the ratio theory/data from NLOJET++ calculations. During the calculations, we set the factorization and renormalization scales to be the same, namely at $\mu_r = \mu_f = p_T^{ave}$[1]. We can see from the figure that the theory predictions can not describe the data well. In fact, the χ^2/N is 8.81 and 8.28 for CT18 and CJ15 respectively, showing an extreme discrepancy between the data and theory.

In Table 6.1 and 6.2, we show the χ^2/N for the dijet pp data using various modern proton PDFs. We can see that the worrisome χ^2 value for this pp data is not only unique to the CT18 and CJ15 NLO PDFs, but rather to all modern PDFs in the market. Considering most of these PDFs included some jet data[176–178] from LHC in their analyses, the high χ^2/N can be interpreted as tensions between the CMS dijet data and the other jet data. However, we note here that the correlated systematic uncertainties are not available in the CMS dijet data. Therefore, such tensions might be artificial.

PDFs	MSHT20	NNPDF4.0	CT18	CJ15	CT14	ABMP16	NNPDF31
χ^2/N	6.10	9.43	8.81	8.28	7.69	3.52	4.17

TABLE 6.1: χ^2/N for CMS pp dijet data, with theory predictions calculated using several modern NLO PDFs[7–10, 179–181].

Going back to the CT18 and CJ15 NLO PDFs, to see how these PDFs should be modified in order to better describe the dijet data, we perform a Bayesian reweighting analysis (discussed

[1]Note that it is possible to choose another scale choice, for example, the invariant mass of the dijet M_{dijet}, which has been shown to improve perturbative convergence up to leading color NNLO precision[175]. However, Refs. [166] showed that using $\mu_F = M_{dijet}$ leads to similar theory predictions as when $\mu_F = p_T^{ave}$ is used.

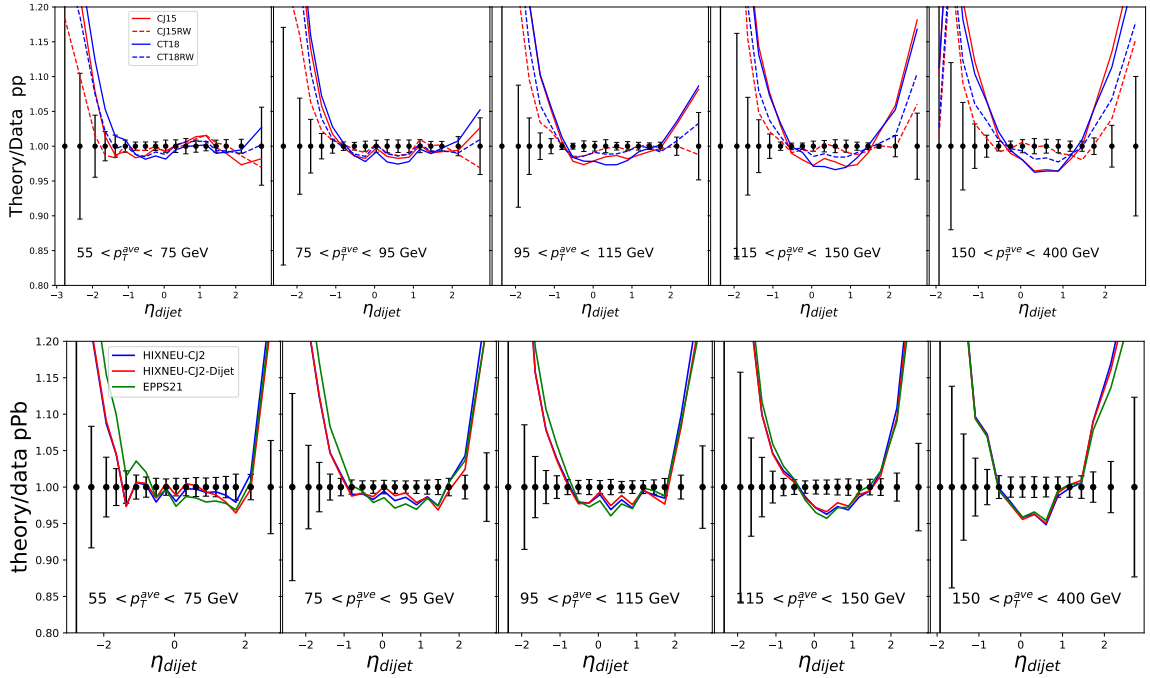


FIGURE 6.2: Theory/data for the dijet productions at CMS in proton-proton (top panel) and proton-lead (bottom panel) collisions at $\sqrt{s} = 5$ TeV.

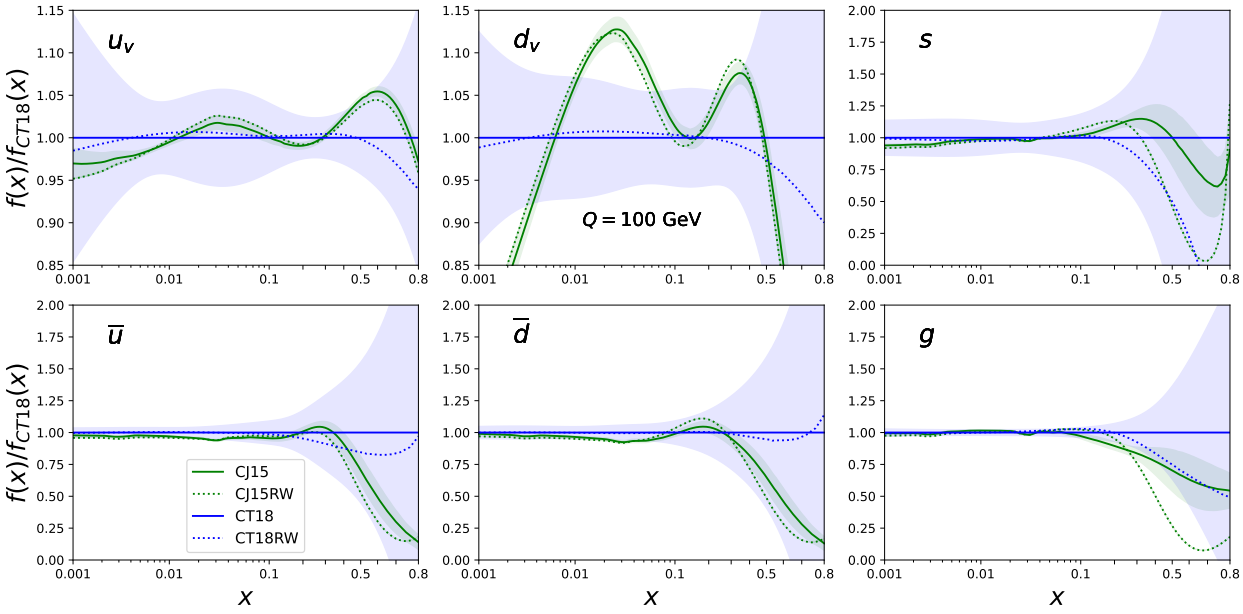


FIGURE 6.3: The shape of CT18 and CJ15 PDFs before and after reweighting procedure with the pp dijet data from CMS.

PDFs	nCTEQ15	nNNPDF3.0	EPPS21
χ^2/N	18.87	10.56	8.70

TABLE 6.2: χ^2/N for CMS pp dijet data, with theory predictions calculated using the proton PDF baselines from nCTEQ15[25], nNNPDF3.0[75] and EPPS21[24].

in Section 3.3). The data-theory comparison after the reweighting is shown in the top panel of Fig. 6.2. We can see that the predictions from the reweighted PDFs have now better agreement with the data. The χ^2/N values after reweighting are 1.61 and 2.97 for CJ15-reweighted (or CJ15RW for short) and CT18-reweighted (CT18RW) respectively[2]. In Fig. 6.3, we show the ratio of the reweighted CJ15 (CJ15RW) and CT18 (CT18RW) PDFs to the original CT18 PDFs. One can see that the CMS dijet data prefers softer PDFs at high x. While less pronounced, this reduction trend can also be seen for the valence quark PDFs at low x. As the high x region corresponds to a negative rapidity region, we can understand this as the way the fit try to adapt to the data at negative rapidity, where the original theory predictions using CJ15 and CT18 severely overshoot the data.

We have seen that it is very difficult to reproduce the CMS dijet data the latest proton PDFs from different PDF fitting groups. Reweighting the CJ15 and CT18 PDFs with the pp data improves the data description by a significant margin, but it still can not satisfactorily reproduce the data in edges of η_{dijet} region. One might argue the going to higher order of pQCD is necessary to better describe the data. However, as shown in [182], by performing proton PDF fit at NLO and NNLO with this dijet data as well with other jet data[177, 178], large χ^2/N can still be observed ($\chi^2/N = 2.51$ and $\chi^2/N = 6.91$ for the NLO and NNLO fit respectively). It is worth noting that the NNLO fit in [182] was performed using K-factors. Therefore, it is not a full NNLO fit. If we regard the K-factor fit from [148] as a good approximation to the full NNLO fit, this means that inability to satisfactorily describe the CMS dijet data is likely not because of missing higher order terms. Similar conclusions were also drawn in a study by the EPPS group[166].

6.3 The pPb and pPb/pp Data

From the previous section, we have shown that the pp data can not be reproduced by the CJ15 and the other proton PDFs. It turns out, similar to the pp data, the pPb spectra can also not be well-described by the recent EPPS21 nPDFs and the HIXNEU fits from appendix B. In the second column of Tab. 6.3, we show the χ^2/N for the pPb data, calculated using nPDFs from HIXNEU fits and EPPS21. We can see that all these nPDFs gives very large χ^2/N for the pPb data.

[2] At first, it is rather surprising that, with large errors of CT18 PDFs, the reweigted PDF should be able to find points in the parameter space preferred by the dijet data easier. However, one should note that the CT18 PDFs use a large $\Delta\chi^2 \sim 100$ tolerance. Thus, for each PDF replica of the CT18, the reweighting put a scale factor $1/100$ to the χ^2 of the dijet data, leading to larger final χ^2.

Fit	pPb	pPb / pp
HIXNEU-CTEQ	[15.82]	[7.50]
HIXNEU-CJ1	[7.00]	[3.59]
HIXNEU-CJ2	[7.36]	[2.68]
HIXNEU-DeuCJ2	[6.13]	[2.67]
HIXNEU-CJ2-Dijet	[6.45]	2.04
EPPS21	[8.80]	1.62

TABLE 6.3: χ^2/N for the dijet production spectra (pPb column) and sepctra ratio (pPb/pp column) data from proton-lead collisions at CMS. The values of χ^2/N with the brackets indicate that the corresponding data is not used in the analysis.

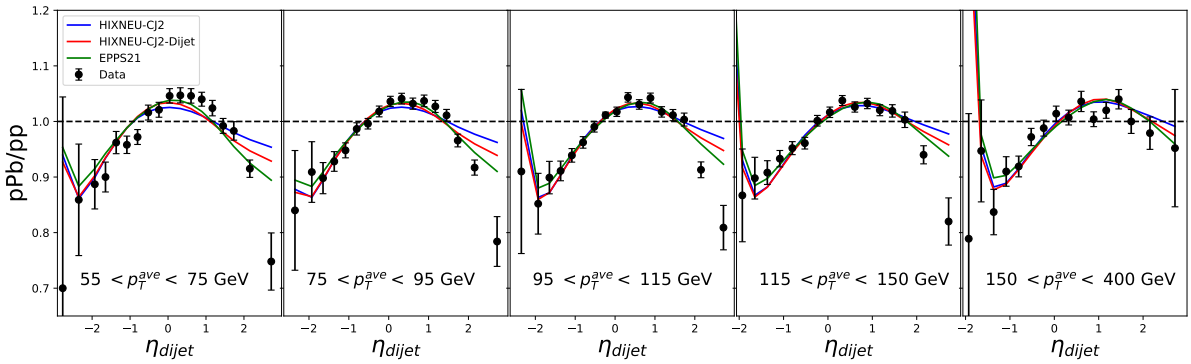


FIGURE 6.4: Data-theory comparison for the ratio pPb/pp dijet spectra data from CMS.

In the bottom panels of Fig. 6.2, we show the theory/data calculated using nPDFs from the HIXNEU-CJ2, HIXNEU-CJ2-Dijet (to be discussed in the next section) and EPPS21 fits. We can see that for $\eta_{dijet} \leq 2$, the theory/data ratios for pPb from all these three fits are very similar to the ones in the pp case (shown in the top panels of Fig. 6.2). This means that all these fits should have a good agreement with the $\eta_{dijet} \leq 2$ pPb/pp data. This expectation is confirmed in Fig. 6.4 where we show data-theory comparison for the ratio data pPb/pp. In fact, if we remove the data with $\eta_{dijet} > 2$, we obtain χ^2/N of 1.27, 1.18, and 1.14 for the HIXNEU-CJ2, HIXNEU-CJ2-Dijet, and EPPS21, respectively. This shows that these fits can reproduce the dijet ratio data quite well for $\eta_{dijet} \leq 2$. As the forward rapidity region correspond the low x lead PDFs, where the gluon dominates the process, we expect that stronger gluon shadowing is required to have a better data description.

Now, given that both the pPb and pPb/pp data can not be satisfactorily described by the HIXNEU-CJ2 fit, it is natural to ask if one should include these data in an nPDF fits. As the data correlations are not available, the large χ^2 of the pp and pPb data may actually be caused by the missing data correlations. In the pPb/pp data, however, we expect some degree of cancellations of the correlated systematic uncertainties. Therefore, the pPb/pp data is less affected by the missing correlation. This can explain why the χ^2/N value for the pPb/pp data is vastly better than the pPb one. Thus, assuming that the large χ^2/N values for both pp and pPb data are caused by the missing correlations, the pPb/pp data then can be safely included

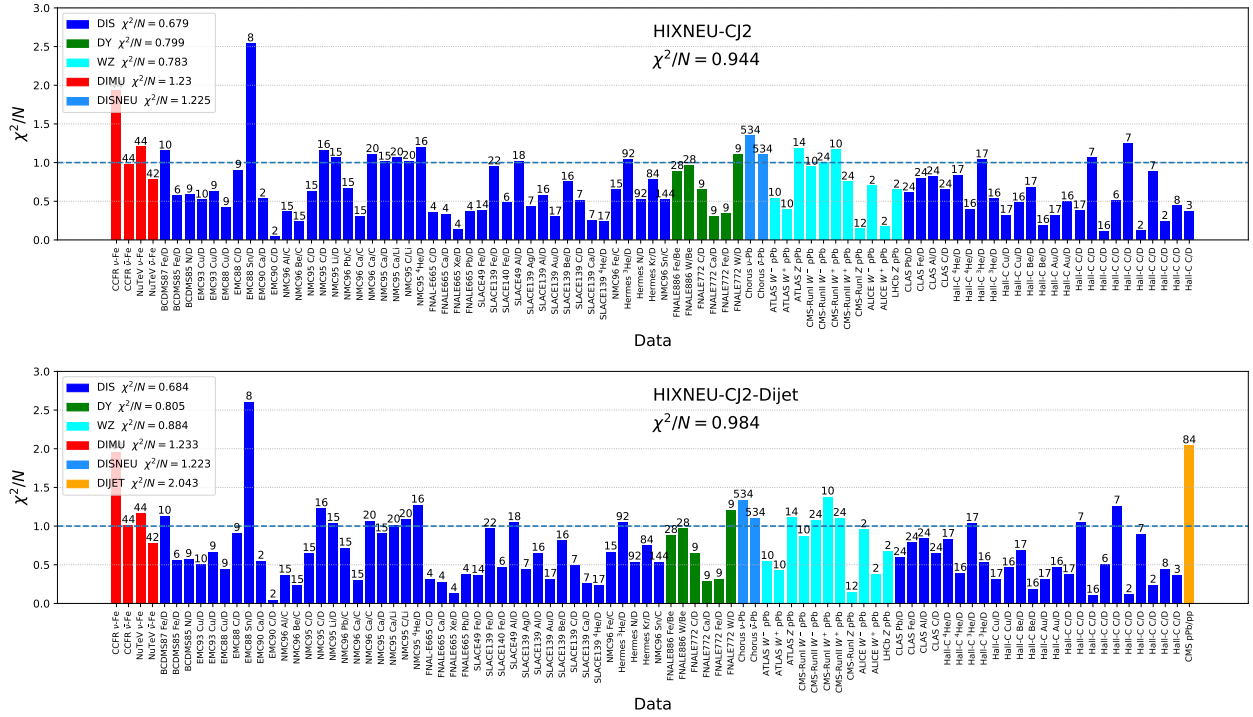


FIGURE 6.5: The distribution of χ^2/N per-experiment in the HIXNEU-CJ2 and HIXNEU-CJ2-Dijet fits.

in an nPDF fit.

However, given the fact that all proton PDFs can not describe well the pp spectra, such a better χ^2/N for the pPb/pp data may actually be caused by accidental cancellations of the data-theory discrepancies occurring in the numerator and denominator of the ratio pPb/pp. In this case, including the pPb/pp data can potentially cause the fitted nPDFs to absorb the discrepancies in the denominator. This leads to an inaccuracy of the fitted PDFs. Another issue that can arise from including the pPb/pp data is the potential tensions with the W and Z data from LHC. As the dijet data is expected to prefer deeper shadowing at low x, this contradicts with the preference of the W and Z data. Given that the W and Z data does not suffer from a proton baseline issue, the tensions between the pPb/pp and W and Z data are likely originated from the denominator issue, rather from a real physics.

Regardless of whether the cancellations of data-theory discrepancies in the numerator and denominator of pPb/pp ratios are accidental or due to cancellations of correlated systematic uncertainties, one can always try to include the pPb/pp data in an nPDF fit to see if the impact of including this data to the fitted nPDFs. After all, two latest nPDFs in the market, EPPS21[24] and nNNPDF3.0[75], also include this data in their analyses. In the following section, we will discuss the inclusion of the dijet data in an nPDF fit that extend our baseline fit (HIXNEU-CJ2).

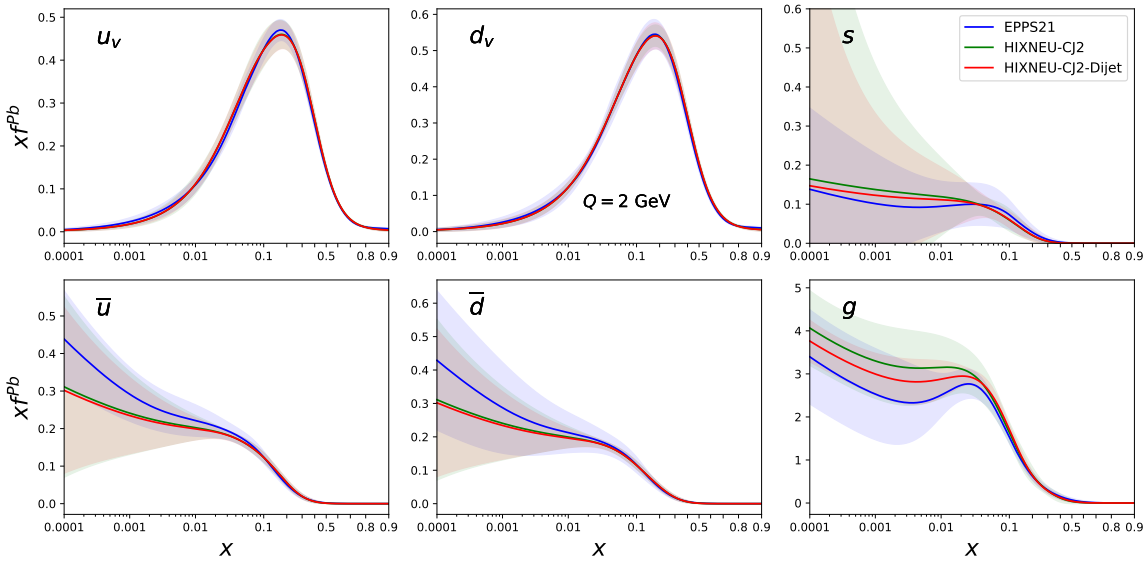


FIGURE 6.6: The full lead PDFs from the HIXNEU-CJ2, HIXNEU-CJ2-Dijet and EPPS21[24] analyses.

6.4 nPDF Fit with pPb/pp Data

The fit that extends HIXNEU-CJ2 by including the dijet ratio data from CMS is referred to as HIXNEU-CJ2-Dijet fit. In the bottom panel of Fig. 6.5, we show the χ^2/N per experiment from this fit. For comparison, we also show the χ^2/N from the reference fit, the HIXNEU-CJ2, in the top panel of Fig. 6.5. Comparing the two fits, we can see that there is almost no difference in the χ^2 values of the charged lepton DIS, DY, neutrino DIS and dimuon data. A notable, yet acceptable, increase of χ^2/N can be seen in W and Z data from LHC. Furthermore, the dijet data still has a large $\chi^2/N = 2.04$, signaling a rather poor description of the dijet data. From data-theory comparison displayed in Fig. 6.4, we can see major improvements of the data description in the central rapidity region, but the data in the forward rapidity region is still poorly described.

In Fig. 6.6, we show the extracted lead PDFs from the HIXNEU-CJ2-Dijet fit. We can see that the PDFs are identical to that of the HIXNEU-CJ2 fit for all flavors except the gluon. The HIXNEU-CJ2-Dijet fit shows a softer gluon PDF at low x, which correspond to an increased shadowing. In Fig. 6.7, we show the extracted nuclear correction. One can clearly see that the gluon PDF from the HIXNEU-CJ2-Dijet fit is more shadowed than the one from the HIXNEU-CJ2, and the EPPS21 gluon PDF has even stronger shadowing.

In the high x region ($x \sim 0.55$), one can see the tendency to have a shallower shadowing for the gluon PDF. However, it is not obvious whether this is due to the momentum sum rule or the preference of the data. After all, the data in the backward rapidity region, which corresponds to this high x region, can already be described well by the HIXNEU-CJ2 fit. Furthermore, in this region, the contributions from the valence quark increase dramatically. Therefore, the correspondence between nuclear correction of the gluon at high x and the pPb/pp data is

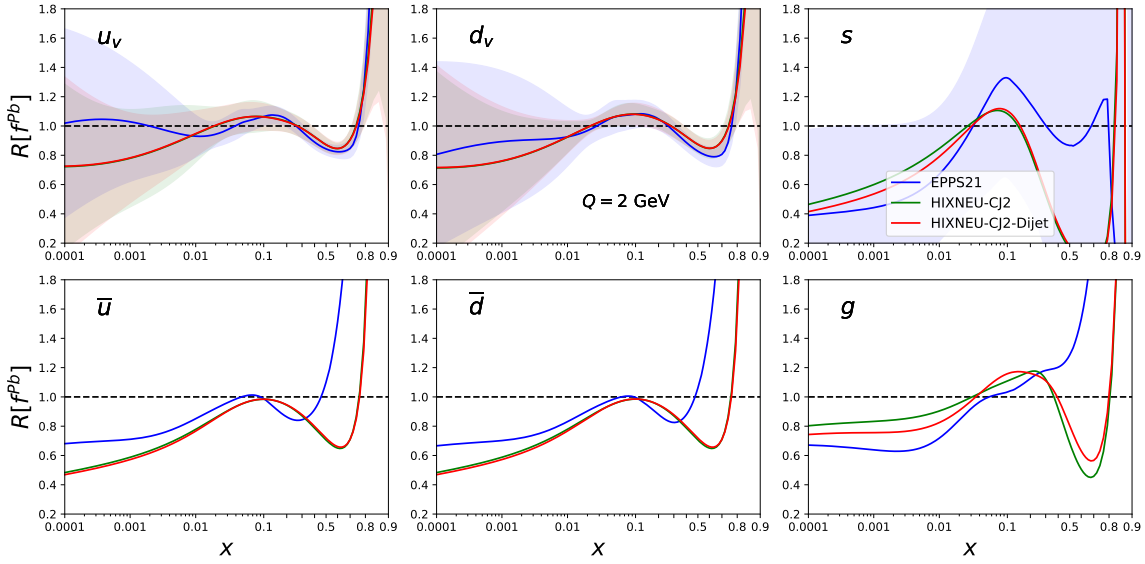


FIGURE 6.7: The extracted nuclear correction ratio from the HIXNEU-CJ2 and HIXNEU-CJ2-Dijet fits for lead nucleus. For comparison, we also show the ratio from the EPPS21 nPDFs[24]. In all plots, the denominator of the nuclear ratio is computed using the CJ15 NLO PDFs.

weaker.

Finally, comparing the PDF uncertainties from the two fits in Fig. 6.7, we can see that the uncertainties in the HIXNEU-CJ2-Dijet are very similar as in the HIXNEU-CJ2 fit for all flavor but the gluon. For the gluon PDF, we can see dramatic reductions of uncertainties, showing the constraining power of the dijet data to the gluon PDF.

6.5 Summary

In this study, we have investigated the viability of including the CMS dijet data in an nPDF global analysis with the other nuclear data. The inclusion of this data is not straightforward due to the inability of modern proton PDFs to well-describe the pp. Therefore, extractions of nPDFs using either pPb or pPb/pp data can lead to an inaccuracy, as the fitted nPDFs can absorb the data-theory discrepancies of the pp data.

There are at least two approaches that one can adopt to view this problem. First, if the large values of χ^2/N for the pp and pPb data are caused by the missing correlations, then one can safely include the pPb/pp data in an nPDF fit. This is because some correlated systematic uncertainties from the numerator and denominator are expected to cancel to some degree, making the ratio pPb/pp data less sensitive to the missing correlations. In this approach, the fact that all modern proton and nuclear PDFs can not well-describe the pp and pPb data is no longer an issue, as the discrepancies due to the missing correlations will cancel in the ratio.

The second approach to view this problem is that the cancellations of the discrepancies in the numerator and denominator are merely accidental, hence they can not be used to justify the

inclusion of pPb/pp data in an nPDF fit. This view is supported by the observation that the W and Z data from LHC, which does not suffer from proton baseline issue, prefer different shape of the gluon PDF at low x than the pPb/pp data.

Adopting the first approach, we included the dijet pPb/pp data in an global analysis that extend the HIXNEU-CJ2 fit, which represents a global analysis with the data sets used in nCTEQ15-HIX[60] and BaseDimuChorus[72]. The resulting fit is called HIXNEU-CJ2-Dijet, which has been shown to have much smaller gluon PDF uncertainties and stronger gluon shadowing at low x.

Chapter 7

Conclusions and Outlook

Parton distribution functions (PDFs) are essential to make predictions based on the QCD factorization theorem. The data-driven approach is usually adopted to determine PDFs as first principle determinations of PDFs based on lattice QCD is not yet reliable and precise enough. The determination of PDFs then can be regarded as an inverse problem: given a set of data, a set of functions must be inferred from it. The data-driven approach is reliable if: 1) all partonic flavors are sufficiently constrained by the data, ideally in all x-region, 2) the theory (model) is not misspecified. That is, the model, given the functional space of the PDFs, can describe the data, 3) reliable fitting methodology. This thesis is aimed to improve the reliability of nCTEQ nPDF fits, by improving each of these aspects.

When it comes to improving flavor separations, new data needs to be included. Given the fact that charged lepton DIS data has been included in all nPDF analyses, it is then natural to include neutrino DIS data. The inclusion of this data will result in better constraints for the valence and down-type quark PDFs. With the inclusion of the data from dimuon production data, the strange quark PDF can be further pinned down. That being said, the inclusion of neutrino data has not been straightforward due to tensions with the charged lepton data. The tensions, in the context of nPDF fitting, were first studied by the nCTEQ collaboration and also by the other nPDF fitting groups, such as EPPS, TUJU and DSSZ. In this work, we revisited this issue and carefully treated small effects such as deuteron nuclear effects, normalization uncertainties, and opened more parameters to reduce parameterization bias. We also included the CCFR and CDHSW data in our analysis, which were not included in the past nCTEQ analyses. We found that the tensions are maximal at low x, where the neutrino data seems to prefer no shadowing, while the charged lepton prefers otherwise. Based on this information, we ultimately studied four fits: the BaseDimuNeu, BaseDimuNeuX, BaseDimuNeuU, and BaseDimuChorus. Among these fits, only the BaseDimuChorus clearly passes the compatibility criteria, while the BaseDimuNeuX barely passes. We then compared the prediction of BaseDimuChorus and BaseDimuNeuX to the Dimuon data from NOMAD and showed that the results from these fits are in much better agreement with the data compared to the nCTEQ15WZSIHdeut fit. We also compared the charged current nuclear ratio predictions from all these fits to the CDHS data, finding good agreement.

When it comes to improving the model, in chapter 5, we discussed target mass corrections, which improves theory predictions for DIS structure functions in the large x and low Q^2 region.

Several prescriptions exist in the literature for the TMCs. Here, the one that we follow is based on the master formula in[160, 161], which was derived using OPE formalism. The master formula in [160, 161] was derived specifically for lepton-nucleon DIS. For the lepton-nucleus case, thanks to the rescaling method, the master formula in [160, 161] is still valid, with the Bjorken x replaced with $x_N = Ax_A \leq A$. Using the master formula, we also found that the ratio $F_i^{A,TMC}/F_i^{A,leading}$ is fairly independent of A due to the averaging procedure when constructing full nuclear PDFs from the bound nucleon ones. This suggests a shortcut to calculate the full target mass corrected structure functions from leading ones by multiplying with universal (A-independent) corrections from the ratio $F_i^{A,TMC}/F_i^{A,leading}$. We parameterized the ratio and fit the parameterizations to the exact results. To optimize the number of parameters, we used ${}_2F_1$-based parameterizations (5.87), which only need two parameters for each structure function type. The fitted parameterization works really well to reproduce the data with $\lesssim 0.2\%$ level of agreement.

In this work, we also investigated the viability of doing an nPDF fit with the recent dijet data from CMS. This data is very appealing as it can provide a better handle on the gluon PDFs from $x \sim 0.01$ to $x \sim 1$. As typically only the low x part of the gluon PDF is properly constrained by the data, the ability to constrain the gluon PDF at high x is therefore very valuable. However, it seems that using this data is not straightforward due to the inability of modern proton PDFs to describe the proton-proton (pp) dijet data. Furthermore, the nPDF from the EPPS21 fit, which included the dijet ratio pPb/pp data, can also not describe the pPb data. Surprisingly, the pPb/pp data in the central and backward rapidity region ($\eta_{dijet} \leq 2$) can be described well even by the HIXNEU-CJ2 fit (which does not include the dijet data). We propose two views to approach this puzzle. On the one hand, if the large values of χ^2/N for the pp and pPb data are caused by the missing data correlations, then we can still use the ratio pPb/pp data in an nPDF analysis without worrying too much on the data-theory discrepancies occurring in both numerator and denominator. This is because the correlated systematic uncertainties are expected to cancel to some degree in the ratio, making the pPb/pp data less sensitive to the missing correlation problems. Indeed, the reasonable value of χ^2 for the pPb/pp data with $\eta_{dijet} \leq 2$ confirms the cancellation of the data-theory discrepancies. On the other hand, one can also regard the cancellations of discrepancies as accidental. As the correlated systematic uncertainties will not 100% cancel out in the ratio, the remaining discrepancies can still be absorbed in the nPDFs. Regardless of whether the cancellation of discrepancies are accidental or not, to study the impact of the dijet pPb/pp data, we performed a global fit with the ratio data and showed that the dijet data prefers stronger shadowing than our reference fit. We also noticed dramatic reductions in the nuclear gluon PDF uncertainties.

All the works presented in this thesis are part of the efforts to better determine nPDFs, by improving the three aspects of the data-driven method. However, these are by no means finished endeavors. For example, to have a more consistent determination of nPDFs, a combined proton+nuclear PDF analysis is necessary. With more precise nuclear data coming from LHC, going to NNLO is also needed in the future. Furthermore, with the influx of more precise

data from the planned EIC, the future of nPDF research becomes even more promising, which requires advancements not only in the theory side, but also in the methodology and fitting technology.

Acknowledgements

First of all, I would like express my sincerest gratitude to Prof. Michael Klasen for giving me the opportunity to work in this topic and for supports, helps, advices, and discussions. I would like to thank Dr. Karol Kovarík, for supervising me and for endless chats and discussions. I am indebted to all the members of the nCTEQ collaboration, in particular : Prof. Fredrick Olness, Dr. Aleksander Kusina, Dr. Tomas Ježo, and Prof. Ingo Schienbein. I am grateful to all my friends at the ITP WWU, in particular Pit Duwentäster and Peter Risse.

Last but not least, I would like to thank my family and parents, especially my wife, Briskha, who always supports me during these three years.

Appendix A

Supplementary Materials for Chapter 5

In this appendix, we show the comparisons between the fitted ${}_2F_1$-parameterizations given by (5.87) and the exact results computed using the EPPS16[80], nNNPDF2.0[26], and TUJU19[149] nPDFs.

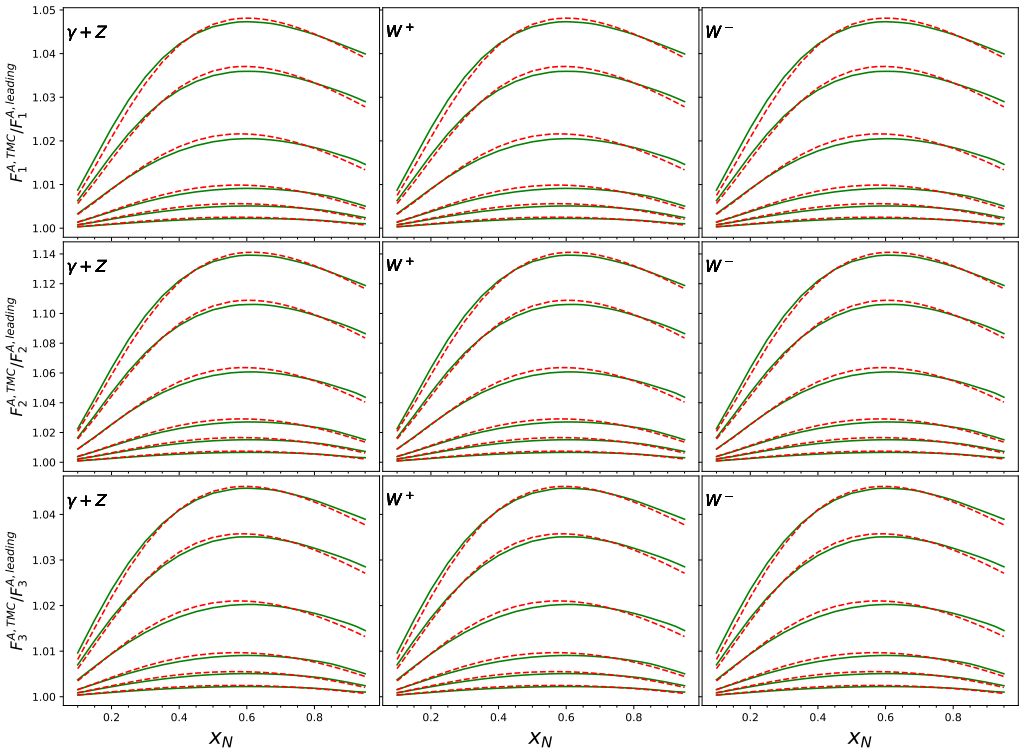


FIGURE A.1: The same as Fig. 5.6, but the nPDFs are from EPPS16[80].

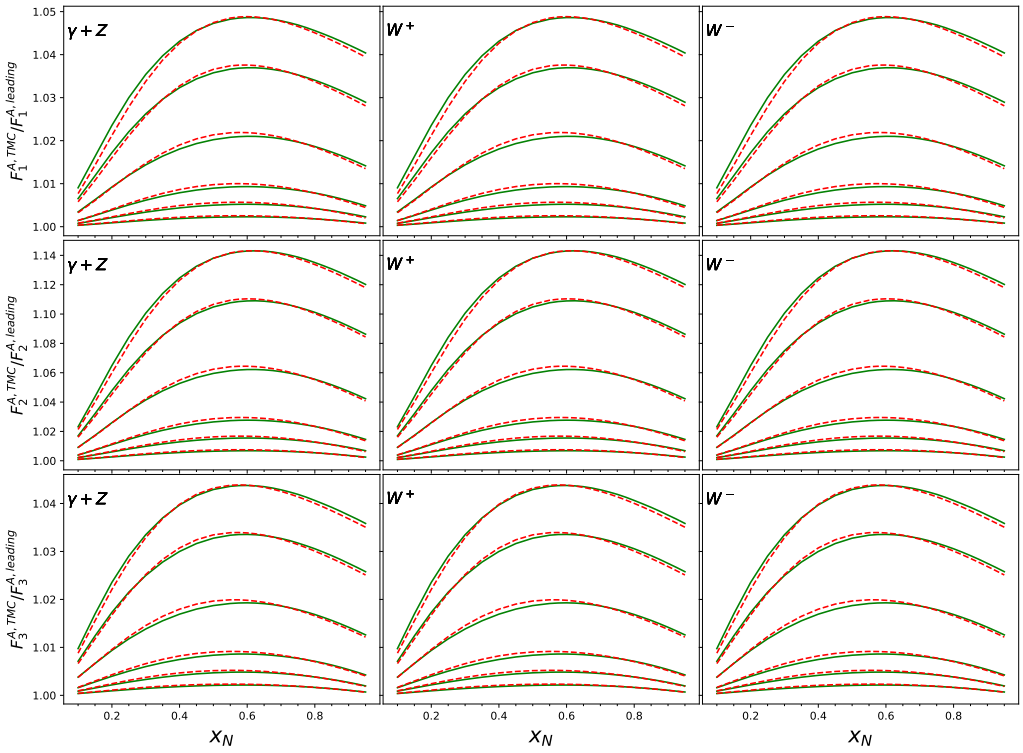


FIGURE A.2: The same as Fig. 5.6, but the nPDFs are from nNNPDF2.0[26].

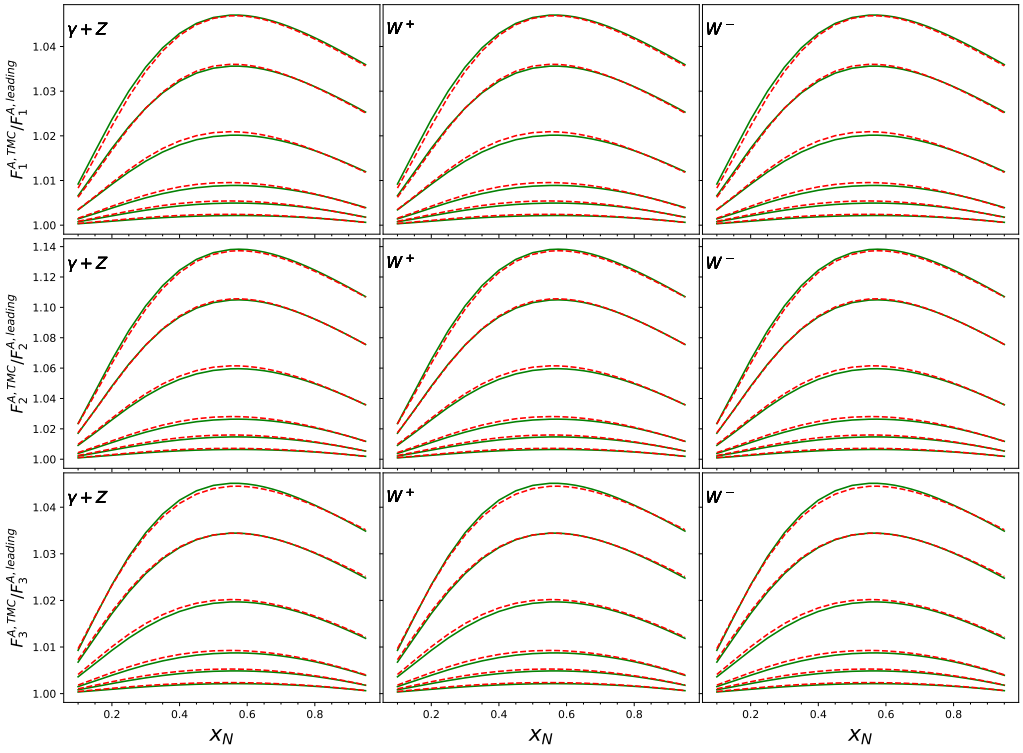


FIGURE A.3: The same as Fig. 5.6, but the nPDFs are from TUJU19[149].

Appendix B

The Combined nCTEQ15HIX and Neutrino Analyses

In this appendix, we present a discussion on our efforts to combine the nCTEQ15HIX[60] and Neutrino[72] analyses. We will improve on several aspects of the fitting methodology, such as the use of the CJ15 NLO PDFs as the proton PDF baseline, new nPDF parameterizations, as well as better treatment of target mass, higher twist, and deuteron corrections. The resulting combined fit serves as a basis to which the dijet data from CMS[167] will be compared.

B.1 Methodological Improvements

B.1.1 Proton PDF Baseline

Ideally, nPDFs are best determined together with proton PDFs using both proton and nuclear data. However, this is currently not possible as this would require a significant upgrade to the current code base. In an nPDF fit where the proton PDF is not fitted, as in the case of nCTEQ nPDFs, the choice of proton PDFs is important for two reasons. First, it provides the boundary condition at $A = 1$ and $Z = 1$. Second, as some nuclear data are presented as ratios of nuclear structure functions F_2^A / F_2^D or cross sections σ^A / σ^p, the proton PDFs are needed to calculate the denominators.

Two important aspects characterize proton PDFs as a good baseline. First, proton PDF analysis should include the largest possible set of data sets, to guarantee a good flavor separation. This is typically the case for modern proton PDFs such as CT18[7], NNPDF[9], and MSHT[8]. Second, the data sets used in the proton PDF analysis should contain as little nuclear data as possible. This is crucial to avoid double counting of nuclear data used in the nPDF analysis and to minimize the bias of applying some nuclear correction to the nuclear data. Of course, if the nuclear data used in the proton PDF analysis has large uncertainties and small nuclear corrections, then one can still argue that such proton PDFs are still a good baseline. As an example, both the NNPDF4.0[9] and CT18[7] PDFs used neutrino-induced charm-dimuon production data from NuTeV, which came from a neutrino-iron DIS experiment. However, it has

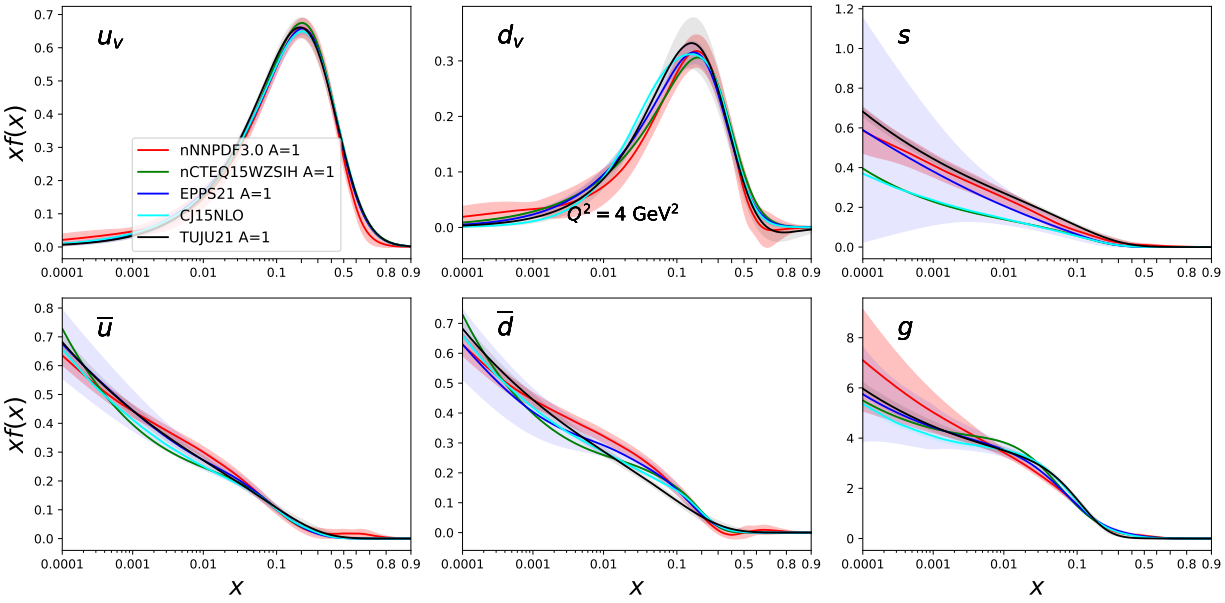


FIGURE B.1: Proton PDF baselines of the recent nPDFs available in the literature.

been argued in [136] that the nuclear correction for this process is much less than the data uncertainties. Furthermore, as this type of data provides one of the strongest constraints to the strange quark PDF, including this data is therefore beneficial for flavor separations.

Here, we summarizes the recent nPDFs with their proton PDF baseline. The nNNPDF3.0 analysis[75] used their own proton PDFs as the baseline. The baseline is an extension of the NNPDF3.1 fit[143], with all the deuteron data removed. All the omitted deuteron data are then included in the nPDF fits. In the EPPS21 analysis[24], the baseline is CT18A[7], which deviates from the standard CT18 PDFs by the addition of ATLAS W/Z rapidity distribution data, which is known to have tensions with some of the DIS data. In the TUJU21 analysis[74], the proton baseline is fitted with the DIS, DY, and W and Z production data from LHC. In nCTEQ15 and its recent iterations [64, 72, 84], a variant of CTEQ6 PDFs[65] were used as the baseline. In Fig. B.1, we show the proton PDF baselines of nCTEQ15WZSIH, nNNPDF3.0, TUJU21, and EPPS21. The figure shows that these nPDFs use a similar proton baseline, and the difference on this should not cause too much concern.

In this study, the CJ15[10] proton PDFs will be used as the proton PDF baseline. There are several reasons for this. First, aside from the deuteron DIS data, no nuclear data was used in the CJ15 analysis. While the deuteron data is technically nuclear data, its theory predictions are computed using a phenomenological model whose parameters are fitted together with the proton PDFs. Thus, assuming that the phenomenological model used in the CJ15 analysis is correct, then the extracted CJ15 PDFs are bias-free from absorbing deuteron nuclear effects. Second, as we aim to relax the kinematic cuts to allow CLAS/JLAB nuclear data[19, 20], it is then desirable to have proton PDF baseline whose analysis used the same cuts. As shown in [60], relaxing the DIS kinematical cuts to $Q^2 \geq 1.3$ GeV^2 and $W^2 \geq 1.7$ GeV^2 leads to an inclusion of data points with a very high Bjorken x ($x \approx 0.9$). If the high x region of the proton

PDF baseline is less well-constrained, due to lacks of proton data in this region, the resulting theory predictions for F_2^A/F_2^D or σ^A/σ^p will be inaccurate. Furthermore, the inaccuracy can potentially be absorbed into the fitted nPDFs. The CJ15 analysis used the same relaxed cuts as the nCTEQHIX analysis[60]. Therefore, it is a good choice in this sense. Third, the initial-scale parameterizations of CJ15 PDFs are simpler than, say, CT18. Therefore, if one uses the CJ15 parameterization to parameterize the nPDFs, the minimizer algorithm should find the minimum easier. We note here that while having a smaller number of parameters is desirable in terms of minimization, it also means that it is potentially less flexible. This will be discussed further in the following section.

B.1.2 Parameterization

As in the previous nCTEQ analysis, the full nPDFs f_i^A are written in terms of the effective bound proton $f_i^{p/A}$ and neutron $f_i^{n/A}$ PDFs as

$$f_i^A(x,Q) = \frac{Zf_i^{p/A}(x,Q) + Nf_i^{n/A}(x,Q)}{A}, \tag{B.1}$$

where $A, Z, N = A - Z$ are the mass, atomic and neutron numbers respectively. Note that, once (B.1) is true for some initial scale Q_0, it will be true for all Q as the DGLAP evolution is linear. The bound neutron PDFs can be obtained from the bound proton ones by assuming isospin symmetry, namely :

$$\begin{aligned} &f_u^{n/A} \leftrightarrow f_d^{p/A}, \\ &f_{\bar{u}}^{n/A} \leftrightarrow f_{\bar{d}}^{p/A}, \\ &f_i^{n/A} = f_i^{p/A}, \;\; i \neq \{u, d, \bar{u}, \bar{d}\}. \end{aligned}$$

The bound proton PDFs is parameterized at an initial scale $Q_0 = 1.3$ GeV in a way such that for $A = 1$, the CJ15 PDFs are recovered for all x. The strategy to allow this to happen is by first choosing a sufficiently flexible x-dependent PDF parameterization form, and then for each parameter, assign an A, Z-dependent function that reduces to parameters that reproduce the CJ15 PDFs. Of course, the natural choice for the x-dependent paramaterizations will be just the ones used in the original CJ15 analysis. The same strategy was actually employed in the pas nCTEQ analyses. This time, however, to improve the flexibility, we extend the CJ15 parameterizations to include additional five parameters. We will refer this procedure as the extended CJ15 parameterization (or CJ15 extended for short). As a reminder, the CJ15 parameterization is given by

$$xf_i(x,Q_0) = c_0 x^{c_1}(1-x)^{c_2}(1 + c_3\sqrt{x} + c_4 x), \quad i = u_v, g, \bar{u} + \bar{d}, s + \bar{s}, \tag{B.2a}$$

$$d_v(x,Q_0) = c_0\left[x^{c_1}(1-x)^{c_2}(1 + c_3\sqrt{x} + c_4 x) + c_5 x^{c_6} u_v\right], \tag{B.2b}$$

$$\frac{\bar{d}}{\bar{u}} = c_0 x^{c_1}(1-x)^{c_2} + 1 + c_3 x(1-x)_4^c. \tag{B.2c}$$

Note here that u_v enters in the d_v parameterization. This is to ensure at $x \to 1$, we have a limit $d_v/u_v \to c_0 c_5 < \infty$, which is motivated by several non-perturbative models of hadron structure[183–186], see also a discussion in [187].

The five additional parameters in the CJ15 extended enter in u_v, g, $s+\bar{s}$, $\bar{u}+\bar{d}$, and $\bar{d}/\bar{u}$. These parameters generally serve as the coefficients of the higher older polynomials in $\sqrt{x}$. The extended CJ15 parameterization reads

$$xf_i(x,Q_0) = c_0 x^{c_1}(1-x)^{c_2}(1+c_3\sqrt{x}+c_4 x+c_5 x^{3/2}), \quad i = u_v, g, \bar{u}+\bar{d}\,, \tag{B.3a}$$

$$d_v(x,Q_0) = c_0\left[x^{c_1}(1-x)^{c_2}(1+c_3\sqrt{x}+c_4 x)+c_5 x^{c_6}u_v\right]\,, \tag{B.3b}$$

$$x(s+\bar{s})(x,Q_0) = c_0 x^{c_1}(1-x)^{c_2}(1+c_3\sqrt{x}+c_4 x)e^{c_5\sqrt{x}}\,, \tag{B.3c}$$

$$\frac{\bar{d}}{\bar{u}} = c_0 x^{c_1}(1-x)^{c_2}(1+c_5 x)+1+c_3 x(1-x)^{c_4}\,. \tag{B.3d}$$

Note here that $x(s+\bar{s})$ parametrization is extended by including an exponential $e^{c_5\sqrt{x}}$, instead of simply adding a polynomial of higher degree in $\sqrt{x}$. We have checked that including an exponential function in the $x(s+\bar{s})$ leads to a better fit to CT18 NLO strange quark PDF, which will be discussed further below.

As a comparison, this is the CTEQ6 parameterization used in the nCTEQ15 analysis :

$$xf_i^{p/A}(x,Q_0) = c_0 x^{c_1}(1-x)^{c_2}e^{c_3 x}(1+e^{c_4}x)^{c_5}, \quad i = u_v, d_v, g, \bar{u}+\bar{d}, s+\bar{s}\,, \tag{B.4a}$$

$$\frac{\bar{d}}{\bar{u}} = c_0 x^{c_1}(1-x)^{c_2}+(1+c_3 x)(1-x)^{c_4}\,. \tag{B.4b}$$

Note here the appearance of exponential functions $e^{c_3 x}$ and e^{c_4}. In this case, these are used to enforce positivity to the PDFs.

For all of these parameterizations, the valence number and the momentum sum rules are evaluated in the same way. The sum rules are used to determine the normalization coefficient of the d_v, u_v, and $\bar{u}+\bar{d}$ PDFs. Let $\tilde{f}_i(x,Q_0)$ be the unnormalized PDF, which corresponds to $c_0^i = 1$. The number and momentum sum rules are evaluated as follows.

1. From the d-valence sum rule, the coefficient $c_0^{d_v}$ is determined as

$$c_0^{d_v} = \frac{1}{\int \tilde{d}_v(x,Q_0)dx}\,. \tag{B.5}$$

2. Similarly, from the u-valence sum rule, the coefficient $c_0^{u_v}$ is determined as

$$c_0^{u_v} = \frac{2}{\int \tilde{u}_v(x,Q_0)dx}\,. \tag{B.6}$$

3. The next is momentum sum rule evaluation. The total momentum of all parton flavors can be written as

$$M = \sum_i \int x f_i(x) dx = M_{u_v} + M_{d_v} + M_g + M_{s+\bar{s}} + 2M_{\bar{u}+\bar{d}}\,, \tag{B.7}$$

where $i = t, b, c, s, u, d, g, \bar{d}, \bar{u}, \bar{s}, \bar{c}, \bar{b}, \bar{t}$. In (B.7), the momentum of an individual parton basis $j = u_v, d_v, g, s+\bar{s}$, and $\bar{u}+\bar{d}$ is given by

$$M_j = \int x f_j(x) dx\,. \tag{B.8}$$

After one calculates the coefficient $c_0^{u_v}$ and $c_0^{d_v}$, one can immediately compute the total momentum M_{u_v} and M_{d_v} carried by the valence quarks.

4. Next, one calculates the gluon momentum M_g. In the nCTEQ++ code, the total gluon momentum is set to be a free parameter which is fitted to the data. Thus given the total momentum M_g, one can determine the normalization coefficient by

$$c_0^g = \frac{M_g}{\int x\tilde{g}(x) dx}\,. \tag{B.9}$$

Operationally, one can decompose $c_0^g = c_0^{g,fitted} c_0^{g,elim}$, and hence $c_0^{g,fitted} = M_g$ and $c_0^{g,elim} = 1/\int x\tilde{g}(x)dx$.

5. Given that the total momentum must be unity, the remaining momentum R carried by $s+\bar{s}$ and $\bar{u}+\bar{d}$ must be

$$R_{u_v,d_v,g} = 1 - M_{u_v} - M_{d_v} - c_0^{g,fitted}\,. \tag{B.10}$$

6. Similar to the gluon case, the momentum of $s+\bar{s}$ is user settable by choosing an appropriate normalization coefficient :

$$c_0^{s+\bar{s}} = \frac{M_{s+\bar{s}}}{\int x\tilde{f}_{s+\bar{s}}(x) dx}\,. \tag{B.11}$$

In the nCTEQ++ code, the momentum $M_{s+\bar{s}}$ is set as $M_{s+\bar{s}} = \frac{c_0^{s+\bar{s},fitted} R_{u_v,d_v,g}}{3}$, where $c_0^{s+\bar{s},fitted}$ is fitted. Thus by writing $c_0^{s+\bar{s}} = \frac{1}{3} c_0^{s+\bar{s},fitted} R_{u_v+d_v+g} c_0^{s+\bar{s},elim}$, we have $c_0^{s+\bar{s},elim} = 1/\int x\tilde{f}_{s+\bar{s}}(x)dx$.

7. The remaining momentum that must be carried by $\bar{u}+\bar{d}$ is then given by

$$\begin{aligned} R_{u_v,d_v,g,s+\bar{s}} &= 1 - M_{u_v} - M_{d_v} - c_0^{g,fitted} - M_{s+\bar{s}} \\ &= \left(1 - M_{u_v} - M_{d_v} - c_0^{g,fitted}\right)\left(1 - \frac{1}{3} c_0^{s+\bar{s},fitted}\right)\,. \end{aligned} \tag{B.12}$$

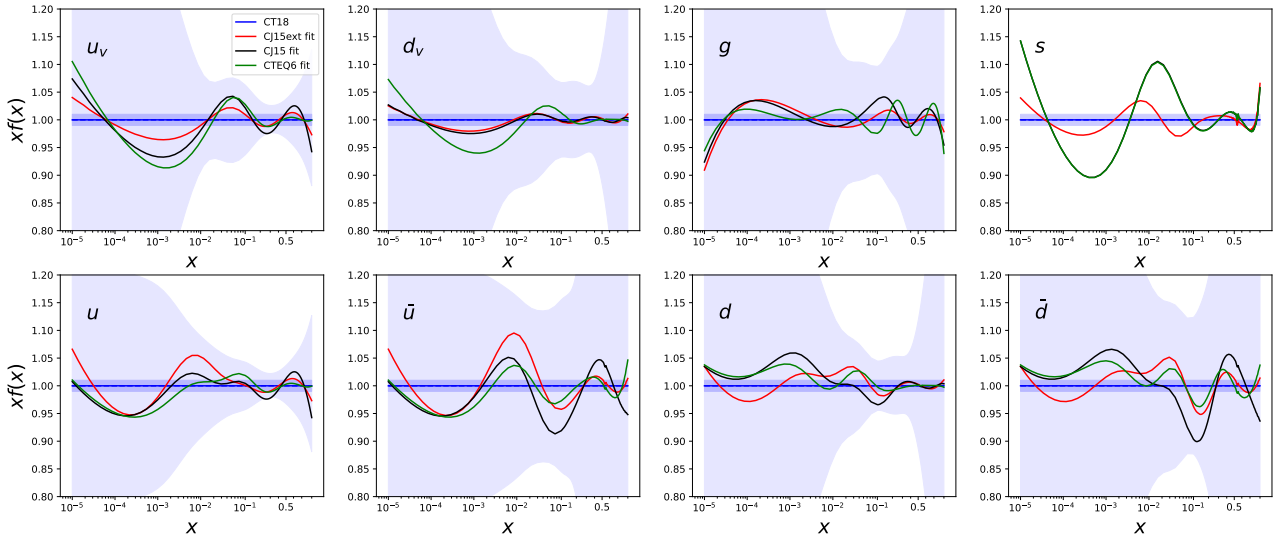


FIGURE B.2: The ratio of the fitted CJ15 (black lines), CJ15 extended (red lines) and CTEQ6 (green lines) parameterizations to the CT18 PDFs at the input scale $Q_0 = 1.3$ GeV. The light blue bands show the actual CT18 PDF uncertainty, while the dark blue shows 1% uncertainty bands.

Thus, this must be equal to $2M_{\bar{u}+\bar{d}}$, see (B.7). The coefficient $c_0^{\bar{u}+\bar{d}}$ is then given by

$$c_0^{\bar{u}+\bar{d}} = \frac{R_{u_v,d_v,g,s+\bar{s}}}{2\int x\tilde{f}_{\bar{u}+\bar{d}}(x)dx}\,. \tag{B.13}$$

To test if the additional parameters in the CJ15 extended improve the flexibility of the parameterization, we fit the standard CJ15, extended CJ15, and CTEQ6 parameterizations to the CT18 NLO PDFs at $Q = 1.3$ GeV. During the fitting loop, the number and momentum sum rules are evaluated using the steps described above. Only CT18 PDFs in the region $10^{-5} \leq x \leq 0.8$ is fitted, as generally LHAPDF interpolation in the very high x region $x > 0.8$ is unstable due to very small values of PDFs in this region. To fit CT18 PDFs, the weighted least square method was used, with the uncertainty of the data (the CT18 PDFs) set to be 1% of the central values. We use 70 data points for each flavor (hence a total of $N = 560$ data points) and we get χ^2/N of 9.14, 3.24, and 5.72 for CJ15, CJ15 extended, and CTEQ6 respectively. This shows that for reproducing the CT18 PDFs, the extended CJ15 parameterization has more flexibility than the original CJ15 and CTEQ6 parameterizations. In terms of the fitted PDFs, as shown in Fig. B.2, the same conclusion can be drawn.

Once the x-dependent parameterizations have been fixed, the next step will be choosing the A-dependence. For this, two parameterizations are investigated :

- *Amode-1* parameterization :

$$c_k(A,Z) = p_k + a_k(1 - A^{-b_k})\,. \tag{B.14}$$

Note that this is the same as in the previous nCTEQ analyses.

- *Amode-2* parameterization :

$$c_k(A,Z) = p_k + a_k \ln(A) + b_k \ln^2(A)\,. \tag{B.15}$$

In both cases, each p_k is fixed to a value which reproduces CJ15 PDFs for $A = 1$. When fitting nPDFs, only the parameters a_k and b_k are varied.

Our reasons for proposing the *Amode-2* parameterization are :

1. As c_k is linearly related to a_k and b_k, fitting these parameters should be easier as it becomes less likely to get stuck in local minima. The linear relation also gives more confidence when using the Hessian method[1].

2. The nucleus mass number A covers rather large range : $1 \leq A \leq 208$. Therefore, using *Amode-1* increase the correlation between a light nucleus with a heavy one. This means, tuning PDFs for $A = 208$ requires significant changes of PDFs for $A = 3$. Using $ln(A)$ as in *Amode-2* minimize this issue.

3. If either a_k or b_k is zero, using *Amode-1*, one always has $c_k(A,Z) = p_k$ for all A. This means, in order to have non-flat $c_k(A,Z)$, both a_k and b_k can not be zero. This is an issue if one wants to fix one of the a_k or b_k parameters to zero to reduce the number of open parameters in an nPDF fit. This issue does not exist in *Amode-2*.

4. It is easy to extend the flexibility of the *Amode-2* parameterization, for instance, by adding a higher degree polynomial in $\ln(A)$. With *Amode-1*, it is less obvious how to do this.

5. Finally, as shown in [41], the size of shadowing and EMC effect of the F_2^A DIS structure function seems to be described well by a linear function in $\ln(A)$, therefore modifying the PDF parameters by a polynomial of $\ln(A)$ seems to be more natural.

To improve the flexibility of the parameterizations at $x \gtrsim 0.7$, in particular to reproduce a steep rise of nuclear ratios from Fermi motion, we modify the CJ15 extended as

$$x f_i^{p/A}(x,Q_0) = c_0 x^{c_1}(1-x+\eta)^{c_2}(1 + c_3\sqrt{x} + c_4 x + c_5 x^{3/2}), \quad i = u_v, g, \bar{u}+\bar{d}\,, \tag{B.16a}$$

$$d_v^{p/A}(x,Q_0) = c_0 \left[x^{c_1}(1-x+\eta)^{c_2}(1 + c_3\sqrt{x} + c_4 x) + c_5 x^{c_6} u_v\right]\,, \tag{B.16b}$$

$$x(s+\bar{s})^{p/A}(x,Q_0) = c_0 x^{c_1}(1-x+\eta)^{c_2}(1 + c_3\sqrt{x} + c_4 x) e^{c_5\sqrt{x}}\,, \tag{B.16c}$$

$$\frac{\bar{d}^{p/A}}{\bar{u}^{p/A}} = c_0 x^{c_1}(1-x)^{c_2}(1+c_5 x) + 1 + c_3 x(1-x)^{c_4}\,. \tag{B.16d}$$

where

$$\eta(x,A) = \epsilon\, x^{\kappa} \ln(A)\,. \tag{B.17}$$

[1]For example, an absolute (non ratio) cross section or structure function can be written as a convolution of Wilson coefficient and PDFs. Thus, it can be regarded as a linear operator acting on PDFs. In the CJ15 extended parameterizations, all the PDFs are linear in c_k, except for c_1 and c_2. The PDFs depend on c_1 and c_2 via $x f_i(x) \propto x^{c_1}$ and $x f_i(x) \propto (1-x)^{c_2}$. These monotonic functions behave like linear functions if one looks at a small neighborhood. All in all, the cross sections or structure functions then behave as a linear function of the PDF parameters a_k and b_k.

Note that η is the same for all parton flavors. This modification was inspired by the x-rescaling in [60], which was shown to improve the description of the very high x DIS data. In this work, however, a total rescaling as discussed in [60] is not used, but is rather imposed partially to only $(1-x)^{c_2}$. It is important to note here that (B.16) implies : $f_i^A(x=1) \neq 0$. This should not raise an issue as nPDFs can, in principle, have non-zero values for $1 \leq x \leq A$. However, in this region, the values of the PDFs are very small, hence this can be ignored for practical purposes.

B.1.3 Corrections From Deuteron Nucler Effects

In the nCTEQ15 analysis[25], the deuteron structure function F_2^D is computed as an isoscalar combination: $F_2^D = F_2^{ISO} = F_2^p + F_2^n$, where F_2^p and F_2^n are the free proton and neutron structure functions. However, as shown in Fig. 2.8, the deuteron structure function slightly deviates from the isoscalar one by less than 1% at $x \lesssim 0.6$ and quickly increases at higher x. In the nCTEQ15HIX[60] analysis, the deuteron nuclear effects are taken into account by modifying the data :

$$\left.\frac{F_2^A}{F_2^D}\right|_{data} \to \left.\frac{F_2^A}{F_2^D}\right|_{data} \times \frac{F_2^{D,CJ}}{F_2^{P,CJ}}\,, \tag{B.18}$$

$$F_2^D|_{data} \to F_2^D|_{data} \times \frac{F_2^{P,CJ}}{F_2^{D,CJ}}\,. \tag{B.19}$$

Here, $F_2^{P,CJ}$ and $F_2^{D,CJ}$ are the proton and deuteron structure functions fitted in the CJ15 analysis. Thus, this method converts F_2^A/F_2^D and F_2^D data into F_2^A/F_2^p and F_2^p data. One should note that in the nCTEQ15HIX analysis, only the central values of the data are multiplied by the $CJ15$ corrections, while the data uncertainties are not modified. This leads to a slightly overestimation of the converted data uncertainties.

A more consistent way to treat deuteron nuclear effects was used in BaseDimuChorus analysis[72]. Instead of modifying the data, one modifies the theory prediction as

$$F_2^D = F_2^p \times \frac{F_2^{D,CJ}}{F_2^{P,CJ}} \qquad (PCJ \text{ Method}) \tag{B.20}$$

We will refer this as the PCJ method. Alternatively, one can also construct the deuteron structure function prediction as

$$F_2^D = F_2^{ISO} \times \frac{F_2^{D,CJ}}{F_2^{ISO,CJ}} \qquad (ISOCJ \text{ Method}) \tag{B.21}$$

$$F_2^D = F_2^{D,CJ} \qquad (DCJ \text{ Method}) \tag{B.22}$$

Here, $F_2^{ISO,CJ} = F_2^{P,CJ} + F_2^{N,CJ}$ is the isocalar structure function from the CJ15 analysis. Theoretically speaking, these three approaches for calculating F_2^D should be the equivalent if one

Method	NMC F_2^D (275 pts)	BCDMS F_2^D (253 pts)
Isoscalar	1.25	1.26
DCJ	1.33	1.17
ISOCJ	1.21	1.15
PCJ	1.21	1.15
PCJ (BaseDimuChorus[72])	1.47	1.06
FIT (HIHXNEU-DEUCJ2)	1.11	1.10

TABLE B.1: χ^2/N for BCDMS[188] and NMC[12] deuteron data calculated using different deuteron correction treatments.

uses the same PDFs and the theory setup[2] as in the CJ15 analysis to calculate F_2^p and F_2^{ISO}. In practice, however, even if one uses the CJ15 PDFs, some difference is expected due to different theory setup. Then the most pragmatic way to treat deuteron nuclear corrections is by using the method that gives the best χ^2 for deuteron F_2^D data. In Table B.1, we show the χ^2/N for BCDMS[188] and NMC[12], with theory predictions given by isoscalar combination of free proton and neutron, *DCJ*, *PCJ*, and *ISOCJ*. We note here that kinematic cuts as in the CJ15 analysis $Q \geq 1.3$ GeV and $W \geq 1.7$ GeV have been imposed to the data. Furthermore, in the calculations of F_2^p, F_2^n and F_2^{ISO} in (B.20) and (B.21), we use CJ15 NLO PDFs, with target mass corrections discussed in the section 5.4 and higher twist correction as in the CJ15 analysis have been applied on top. The table shows that while *PCJ* and *ISOCJ* are perfectly equivalent and give the best χ^2/N, they differ quite a bit with the *DCJ* and *isoscalar* methods. Based on the values of χ^2/N from the table, it is clear that *PCJ* and *ISOCJ* are preferred. In this work, therefore, we use the *PCJ* method to compute F_2^D.

There is another, arguably more consistent, way to compute deuteron structure functions. Given the nPDF fitting framework, one can fit bound proton PDFs $f_i^{p/D}$ of deuteron directly to the data, in the same way as we fit bound proton PDFs of other nuclei. This method is more consistent than the previously mentioned approaches in the sense that now the full deuteron PDFs follow the same continuous A dependence as the nPDFs from other nuclei. Furthermore, this method reduces the dependency of nPDF analysis on the choice of proton PDF baseline as F_2^D is fitted directly to the data. However, fitting $f_k^{p/D}$ to the deuteron data could potentially raise several issues. First, the shape of $f_k^{p/D}$ can be driven by non-deuteron data via the assumed A-dependence, especially in the kinematical region with no precise deuteron data in it. Furthermore, there is not enough flavor separation constrained by the deuteron data, as most of the deuteron data are measured as F_2^D, F_2^D/F_2^P. Lastly, the A-dependent parameterization of the bound nucleon PDFs in an nPDF fit is chosen based on guesses, rather than a physical model, a methodological bias can significantly influence the determination of $f_i^{p/D}$. The deuteron correction from CJ15 analysis has less of this issue as it was determined from a well-established phenomenological model.

We note here that fitting deuteron data in nPDF fit has been done by the other nPDF fitting

[2]Here, we mean the same code for the α_S and DGLAP evolution as well as the same target mass and higher twist corrections.

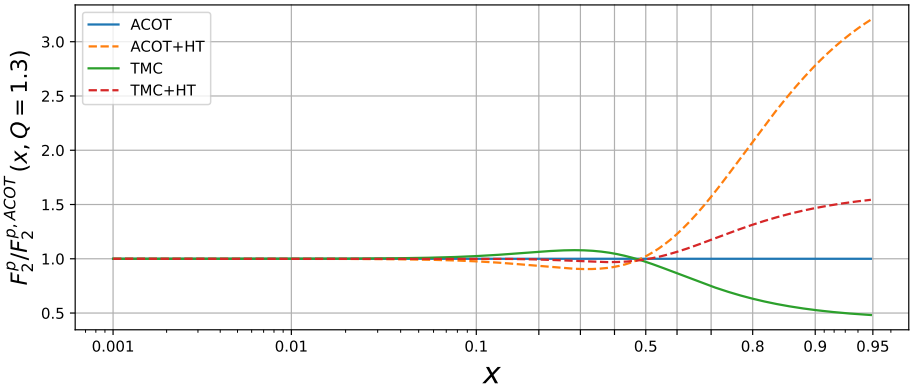


FIGURE B.3: Ratio of F_2 structure function computed using different methods to F_2^{ACOT}. All structure function calculations are computed using the CJ15 NLO PDFs.

groups as well. For example, the latest nPDF releases from NNPDF group (NNPDF3.0[75]) and TUJU (TUJU21[74]). In EPPS21 analysis[24], F_2^D is calculated as an isoscalar structure function. The argument was, that as the EPPS21 proton PDF baseline (CT18ANLO[7]) already used some deuteron data, then the proton baseline essentially absorbs some of the deuteron nuclear effects making the baseline equivalent to the bound PDFs of the deuteron.

In this work, for completeness and future reference, treatment of deuteron nuclear effects by fitting the deuteron PDFs will also be investigated.

B.1.4 Target Mass and Higher Twist Corrections

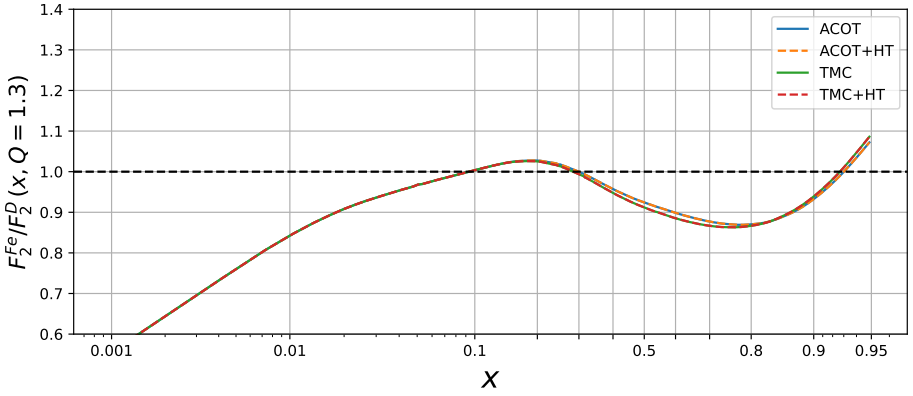


FIGURE B.4: The nuclear ratio F_2^{Fe}/F_2^D predictions using the nCTEQ15HIX nPDFs with different combination of TMC and HT corrections.

In this study, the same theory setup as in the previous nCTEQ analysis is used, except for the DIS structure function calculations, where now OPE-based target mass corrections are applied by default. The TMCs are calculated using ${}_2F_1$-parameterization as discussed in detail in Section 5.4. Furthermore, to deal with the remaining power correction from higher twist

operators, a further multiplicative correction from CJ15 analysis is also applied :

$$F_2^{TMC+HT} = F_2^{TMC}\left(1 + \frac{h_0 x^{h_1}(1+h_2 x)}{Q^2}\right) \tag{B.23}$$

Here, the parameter h_0 controls the overall scale of the corrections, while h_1 controls the well-known rise of the corrections at large x. The parameter h_2 allows for the possibility of negative higher twists at smaller x.

In Fig.B.3, we show the ratio of proton structure function F_2^p computed using different treatment of TMC and higher twist (HT) corrections to the same F_2^p computed using the ACOT scheme (which already includes target mass corrections in the parton model approach). All calculations are done using CJ15 NLO PDFs with $Q = 1.3$ GeV. One can see that, all the four calculations agree well for $x \leq 0.1$. At $0.1 \leq 0.5$, the differences can go as high as 10%. As expected, the largest difference between the four happens in the high x region ($x \geq 0.5$). In this region, the OPE TMC prediction undershoots the ACOT one, while adding HT on top of OPE TMC, the prediction overshoot it.

It is worth pointing out that, although the difference between the combinations of higher twist and TMC are large at high x, it will drop off when one calculates the structure function ratio $F_2^{A_1}/F_2^{A_2}$. This is because the higher twist correction (B.23) are multiplicative and A-independent. Similarly, as the OPE TMC structure function is computed by multiplying leading TMC $F_2^{leading}$ by an A-independent correction factor, it will also cancel out in the ratio. The difference between ACOT and TMC+HT in this case then must come from the difference between leading OPE TMC and ACOT. in Fig. B.4, we show theory predictions for nuclear ratio F_2^{Fe}/F_2^D, calculated using nCTEQ15HIX iron PDFs. For this plot, the deuteron structure function is computed using *PCJ* method. It can be observed that the difference between different combinations of TMC and HT are very small, even at high x.

B.2 The Combined Fits

The main purpose of our study reported in this appendix is to set up combined fits with all the data sets (except the single inclusive hadron data) used in the nCTEQ15HIX[60] and BaseDimuChorus[72] analysis. Specifically, we include :

- Charged lepton DIS data from [12–18, 106–115],
- Drell-Yan lepton pair productions data from [116, 117],
- Drell-Yan W and Z production data from LHC [125–132],
- Neutrino DIS from Chorus[71],
- Dimuon data from CCFR and NuTeV[99],
- Charged lepton DIS data from JLab[19, 20],

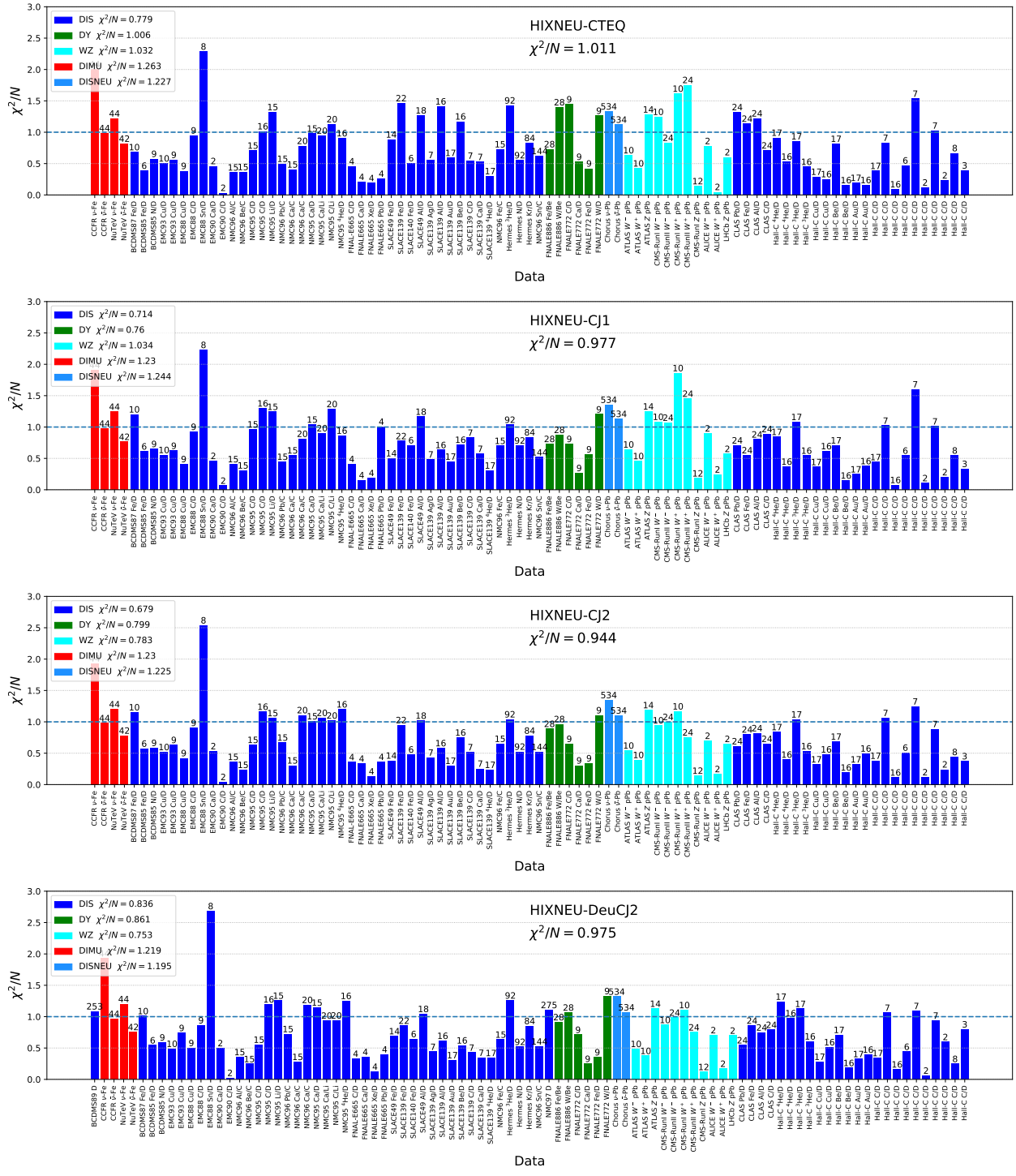


FIGURE B.5: The distribution of χ^2/N per-experiment in the HIXNEU fits. In the lowest panel, we also show the χ^2/N of the BCDMS F_2^D [188] (the left-most bar) and NMC[12] (the right-most blue bar before the green bars).

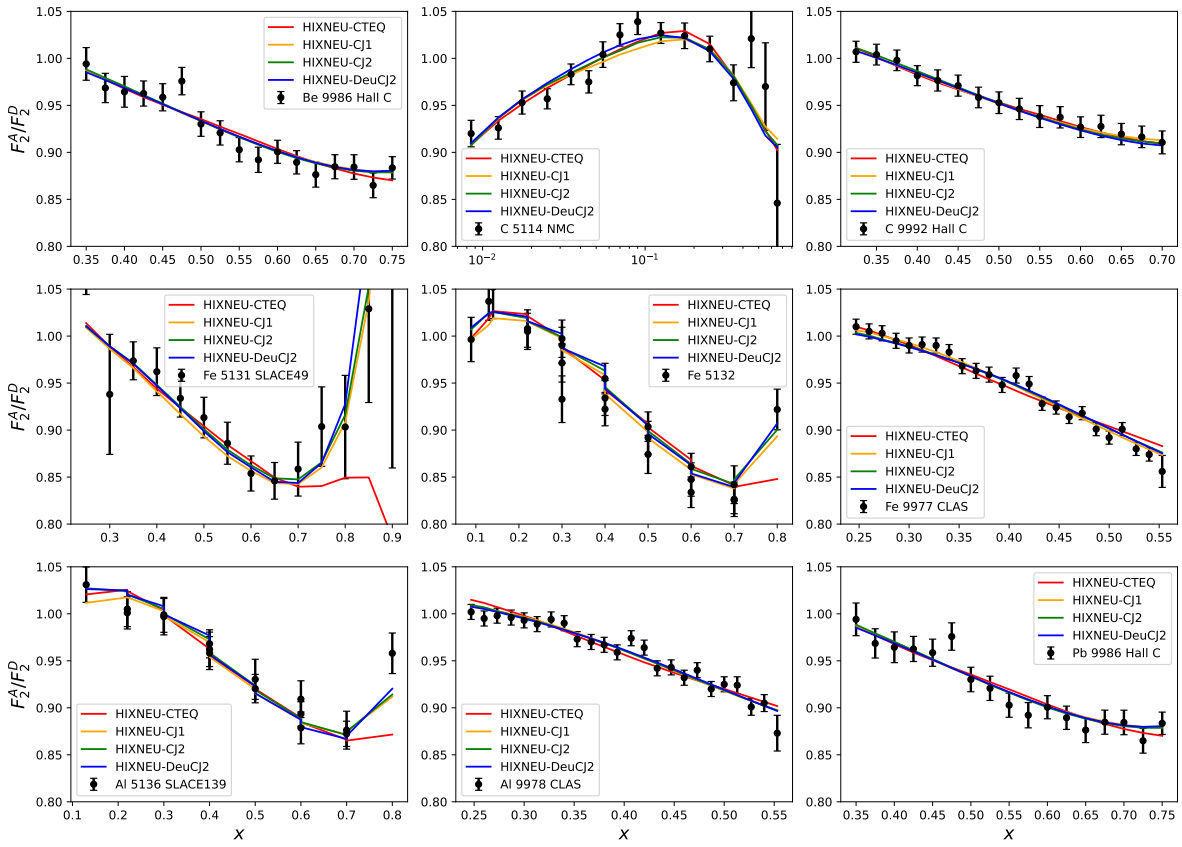


FIGURE B.6: Data-theory comparison for selected data sets, with theory predictions computed using nPDFs from the HIXNEU fits.

- Deuteron F_2^D data from BCDMS[188] and NMC[12] (only in the HIXNEU-DEUCJ2 fit, see discussion below).

For all DIS data (which includes the charged lepton DIS, neutrino DIS, and dimuon data), the following kinematical cuts are applied : $Q_2 \geq 1.3$ and $W^2 \geq 1.7$. These are the same cuts as in the nCTEQ15HIX analysis.

Regarding the fitting methodology, the new fits feature: the updated proton PDF baseline (the CJ15 NLO), the new initial scale parameterization, an improved and more consistent treatment of the deuteron nuclear effects, and the inclusion of OPE TMCs and HT corrections for the DIS structure function calculations. All theory calculations are performed at NLO of pQCD.

Having mentioned the included data sets and the improved methodology, to assess the impact of these imprvements, we do several fits :

- **HIXNEU-CTEQ** : a fit that use the nCTEQ15 PDF parameterizations, with CTEQ6[65] as the proton baseline. All data sets mentioned above, except the deuteron F_2^D data from BCDMS and NMC, are included. The total number of data points is 2651 pts. For this fit, 31 free parameters are fitted.

- **HIXNEU-CJ1** : a fit that use the CJ15 extended parameterization, with an A-dependent given by *Amode1*. The total number of data points is 2651 pts. a total 42 parameters are open during the fitting procedure.
- **HIXNEU-CJ2** : the same as the HIXNEU-CJ1 fit, but *Amode2* is used for the A-dependence.
- **HIXNEU-DeuCJ2** : The same as HIXNEU-CJ2 fit, but the deuteron data from NMC and BCDMS are now included and the deuteron PDFs are fitted. The total number of data points is 3179 pts and a total 42 free parameters are fitted.

Collectively, we will call these fits as HIXNEU fits. The fits are done by minimizing a loss function explained in section 3.1. For the minimization algorithms, Simplex and Migrad are used. They are provided by the Minuit package[66]. To avoid premature stop due to saddle point trap, those minimizers are chained, so that the initial point of the subsequent minimizer is obtained from the earlier one. A total of six or seven minimizers are usually used. To avoid getting trapped inside a local minimum, the fits are performed as follows. First, all 42 parameters are open. Normally, it is challenging for Minuit minimizers to converge if the number of parameters is more than 20. However, with the new *Amode2* parameterization, the fits are easier to converge. After a fit with 42 parameters, we redo the fit by opening a subset of all these parameters, while keeping the others fixed to the same values as obtained from the previous fit. This step is redone several times until the absolute best χ^2 is obtained. Sometimes, some data sets are initially given large weight to force the initial fit to go into the direction preferred by these data sets. This technique is beneficial if one wants to check if a large χ^2 for some data sets is actually caused by a local minimum trap, or due to some other cause (for example, tensions with the other data).

B.2.1 Fit Quality

In terms of overall χ^2/N, as shown in Fig. B.5, all the fits considered here are excellent, as the χ^2/N are very close to the expected value from the ideal case. One can also see that the HIXNEU-CTEQ has the most unequal χ^2/N distribution per experiment. Using the CJ15 extended parameterization with the same A-dependence as in the HIXNEU-CJ1 fit, one can see that the fit has a slightly improved overall χ^2/N. By using the *Amode2* for the A-dependence instead of *Amode1*, in the HIXNEU-CJ2 fit, we can see further improvement in the overall description of the data. The HIXNEU-DeuCJ2 yields similar χ^2/N distribution as in the HIXNEU-CJ2 fit, however, the neutrino data is slightly better described.

To further investigate if the HIXNEU fits can reproduce the data, in Fig. B.6, we show data-theory plots for selected charged lepton DIS data sets. The plots show that all these HIXNEU fits can satisfactorily reproduce the data at low and medium x. Thanks to the x-shifting procedure, the HIXNEU-CJ1, HIXNEU-CJ2, and HIXNEU-DeuCJ2 can better reproduce the high x data compared to the HIXNEU-CTEQ fit.

B.2.2 Deuteron Nuclear Correction

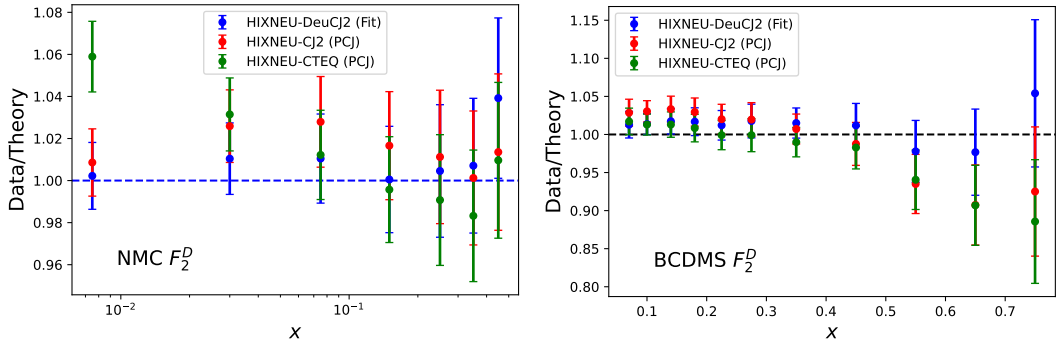


FIGURE B.7: The weighted average of the data/theory of the NMC[12] (left) and the BCDMS[188] data (right).

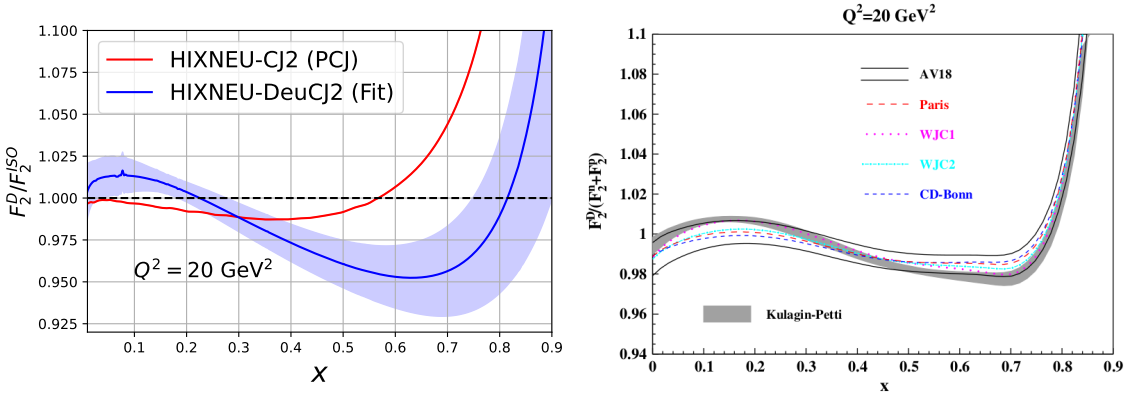


FIGURE B.8: (Left) Deuteron nuclear correction ratio at $Q^2 = 25$ GeV^2 computed using *PCJ* method (red) and the fitted deuteron PDFs (blue). (Right) F_2^D nuclear corrections from [189].

Treatment of deuteron nuclear effects is the main differentiating factor between the HIXNEU-CJ2 and the HIXNEU-DeuCJ2 fits. As mentioned before, in the HIXNEU-CJ2 fit, *PCJ* method is used to calculate the deuteron structure function and the PDF of the bound proton inside deuteron is exactly the same as the free proton ones. This is different in the BaseDeuCJ fit, where the bound proton PDFs of deuteron are fitted in the same way as those of other nuclei. To assess the quality of data description, the χ^2 of NMC[12] and BCDMS[188] deuteron data are shown in table B.1. One can see that predictions from the BaseDeuCJ fit give better χ^2 compared to the standard *PCJ* method. Surprisingly, *PCJ* method with the CTEQ6 proton PDFs as in HIXNEU-CTEQ fit gives a better χ^2 for the BCDMS data, but significantly worse χ^2 for the NMC data.

In Fig. B.7, we show the weighted average of data/theory for both NMC and BCDMS F_2^D data. Let $R_i = D_i/T_i$ be the data-theory ratio for the i-th data point. Then, the weighted average

can be computed as :

$$\mathcal{R}(x) = \sum_i w_i R_i, \tag{B.24}$$

$$\Delta\mathcal{R}(x) = \sqrt{N}\left(\sum_i w_i^2 (\Delta R_i^\sigma)^2\right)^{1/2}. \tag{B.25}$$

In (B.25), a factor $\sqrt{N}$ is introduced to compensate for the averaging error, which goes like $1\sqrt{N}$, where N is the number of data points that are being averaged. The error, $\Delta\mathcal{R}$, defined in this way, represents the actual spread of R_i, while without $\sqrt{N}$ factor, the error represent the averaging errors when the experiment is repeated many times. The weight w_i is defined as

$$w_i = \left(\sum_j \frac{1}{(\Delta R_j^\sigma)^2}\right)^{-1} \frac{1}{(\Delta R_i^\sigma)^2}, \tag{B.26}$$

which is derived from maximum likelihood estimation. Fig. B.7 shows that for the NMC data, predictions from the HIXNEU-DeuCJ2 fit have better agreement with the data than the ones from the HIXNEU-CJ2 and the HIXNEU-CTEQ fits. For the BCDMS data, the predictions from the HIXNEU fits generally agrees well with each other, although the predictions of the HIXNEU-CTEQ fit at the lowest x are much higher than the ideal values. At higher x however, the predictions diverge. At high x, the *PCJ* method from the HIXNEU-CTEQ and the HIXNEU-CJ2 fits leads to too high theory predictions. In contrast, the ones from the HIXNEU-DeuCJ2 fit undershoot the data, but with a less deviation from the ideal value compared to HIXNEU-CJ2 and HIXNEU-CTEQ.

In Fig. B.8, the extracted shape of deuteron nuclear correction is shown. It is interesting to see that predictions from the HIXNEU-DeuCJ2 fit follow a typical nuclear ratio curve: shadowing at low x, anti-shadowing, EMC dip, and Fermi motion. Overall, this curve is similar to the one from Kulagin-Petti model[189], as shown in the right panel of Fig. B.8. On the other hand, the *PCJ*-based prediction from the HIXNEU-CJ2 shows a similar nuclear ratio curve from CJ15 analysis, with no anti-shadowing at mid-x and very steep Fermi motion rise at high x.

B.2.3 The Fitted nPDFs

Given that all the fits, with all the methodological differences, give very good χ^2/N, it is instructive to look at the resulting nPDFs. To study the impact of the added data to the PDFs and compare the resulting nPDFs to the ones from the previous nCTEQ analyses (nCTEQ15[25], nCTEQ15HIX[60], and BaseDimuChorus[72]), In Fig. B.9, we show the predictions for the nuclear ratio curves, defined as

$$R[f_i^A] = \frac{f_i^A}{f_i^{A,free}}, \qquad f_i^{A,free} = \frac{Z}{A} f_i^p + \frac{N}{A} f_i^n. \tag{B.27}$$

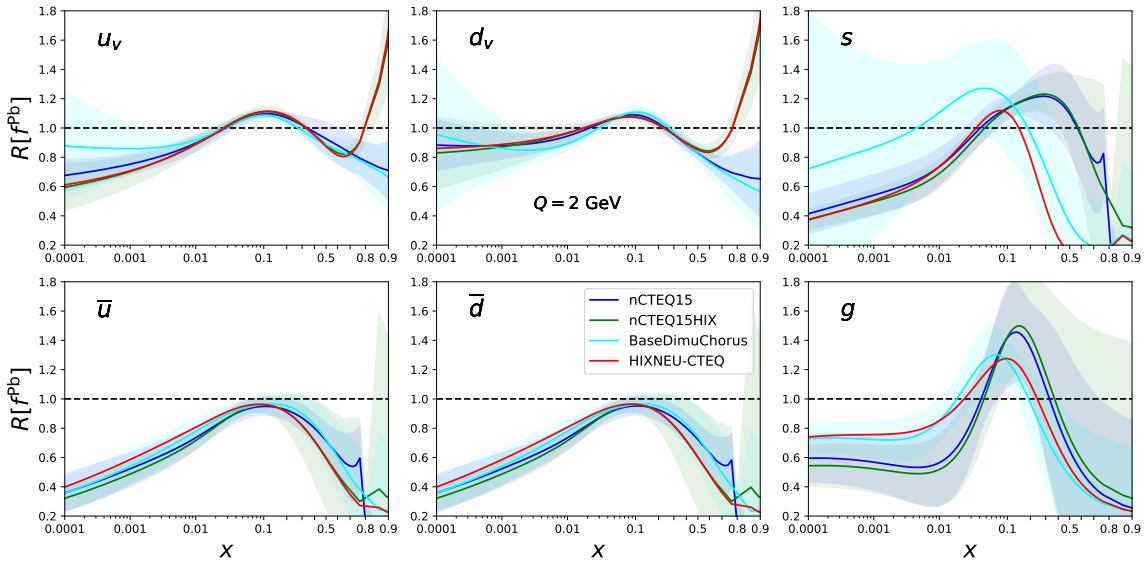


FIGURE B.9: Nuclear corrections of lead PDFs from the previous nCTEQ analyses and the HIXNEU-CTEQ fit. The denominator for the nuclear ratio is computed using the CTEQ6 proton PDFs[65].

Here, $f_i^{A,free}$ is computed using the CTEQ6 PDFs[65]. We can see that the valence quark PDFs of the HIXNEU-CTEQ fit are identical in all x regions to the ones from the nCTEQ15HIX. This is expected, as both fits used exactly the same charged lepton DIS data, which are sensitive to the valence quark PDFs. The $\bar{u}$ and $\bar{d}$ PDFs. The $\bar{u}$ and $\bar{d}$ PDFs are also very similar in all these fits. Switching to the gluon PDF, one can see that the HIXNEU-CTEQ gluon PDF is very similar to that of the BaseDimuChorus fit. The W and Z data are known to place a very strong constraints to the gluon PDF at low x ($x \lesssim 0.1$)[72]. Therefore, it is expected that the HIXNEU-CTEQ gluon is similar to that of BaseDimuChorus, as the W and Z data are both present in both fits. Examining the strange quark PDF, one can see that the HIXNEU-CTEQ agrees well at high x with the BaseDimuChorus, and at low x with the nCTEQ15 and nCTEQ15HIX. Nevertheless, the overall shape of the strange quark PDF from the HIXNEU-CTEQ fit is similar to the one from the BaseDimuChorus up to a normalization factor, suggesting that the difference might come from the momentum sum rule. In short, the HIXNEU-CTEQ fit results in reasonable nPDFs and some differences from previous nCTEQ analyses can be attributed to the addition of the new data.

Lets's now compare the full nPDFs from the HIXNEU fits. In Fig. B.10, we show the extracted nPDFs from the HIXNEU fits. The uncertainties of the nPDFs shown in the figure are obtained by using χ^2 tolerance of $T^2 = 95$, which correspond to 90% percentile of a χ^2 distribution with $N = 2651$ degrees of freedom. We can see that, in general, all the HIXNEU fits are in good agreement within uncertainties. To better see the difference, we rescale the PDFs with the free "lead" PDFs computed using the CJ15 PDFs, to obtain nuclear corrections ratios. The resulting ratios are displayed in Fig. B.11.

Let's first compare the nuclear ratios from the HIXNEU-CTEQ and HIXNEU-CJ1 fits. The

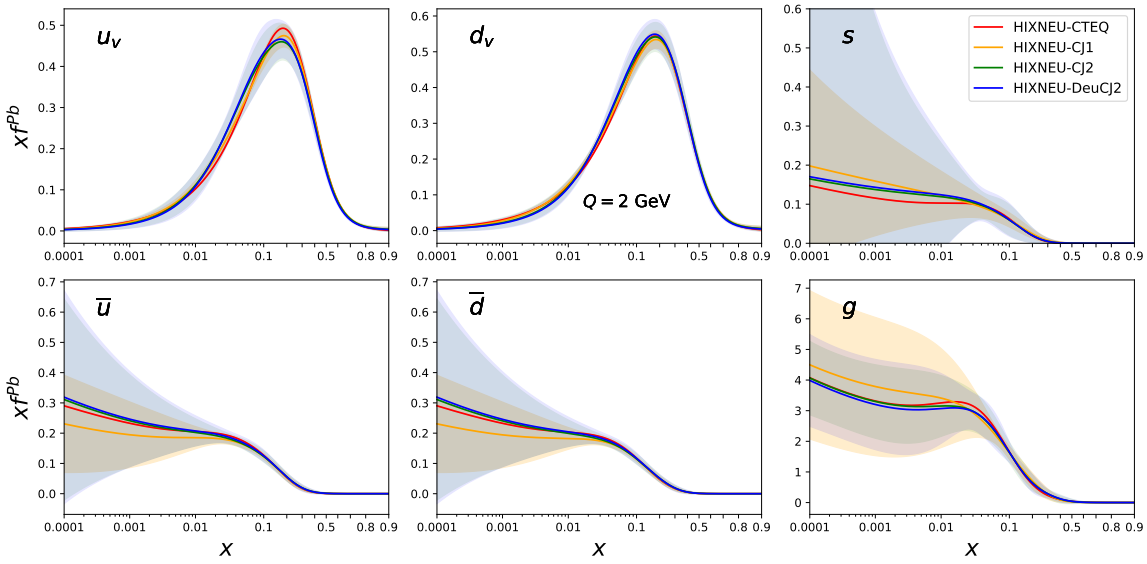


FIGURE B.10: The full lead nPDFs at $Q = 2$ GeV from the HIXNEU fits.

nuclear ratios from these fits are similar overall, except $\bar{u}$ and $\bar{d}$ in the high x region, where the HIXNEU-CJ1 shows a steep rise. Looking at the HIXNEU-CJ1 u_v and d_v curves and comparing them to the ones from the HIXNEU-CJ2, one immediately see some difference the low x ($x \lesssim 0.1$): the HIXNEU-CJ1 fit shows softer shadowing and anti-shadowing. Comparing the ratio curves from the HIXNEU-CJ2 and HIXNEU-DeuCJ2 fits, one can see that they are almost identical, although the HIXNEU-DeuCJ2 generally prefers a slightly stronger size of nuclear ratios for all x.

Looking at the PDF uncertainties, we can see that HIXNEU-CJ1 has much smaller uncertainties than the other fits despite having the same tolerance $T^2 = 95$. This is likely because the *Amode1* is highly non-linear and therefore the Hessian method underestimate the uncertainties. It is interesting to see here that the uncertainties from the HIXNEU-CJ2 are actually smaller than the ones from the HIXNEU-DeuCJ2 fits, despite having less data points. In the HIXNEU-DeuCJ2 fit, F_2^D data from NMC and BCDMS were included, adding a total of 528 data points. As the deuteron PDFs are fitted to the data instead of freezing it to values from other analyses, the uncertainties of the denominator of F_2^A/F_2^D for the charged lepton DIS data and ratio of dimuon yield Y_{pA}/Y_{pD} for the Drell-Yan data are essentially propagated to the HIXNEU-DeuCJ2 PDFs, leading to larger overall uncertainties despite having more data points.

To see the extracted nuclear correction from the HIXNEU fits as a function of A, we plot the ratio $R[f_i^A]$ in Fig. B.12 for 5 nuclei : ^{2}D, ^{4}He, ^{12}C, ^{56}Fe, ^{108}Ag, and ^{208}Pb. For comparison, we also the nuclear corrections from the EPPS21 fit. For the current discussion, we will focus on the valence quark PDFs. In the shadowing region, we see that the heavier nuclei tend to have a stronger shadowing for all parton flavors shown in the figure. This is, of course, aligned with the observation that the size of the shadowing increases with A. However, we can see a striking difference between the HIXNEU-CJ2 and HIXNEU-DeuCJ2 fits with the other fits: the

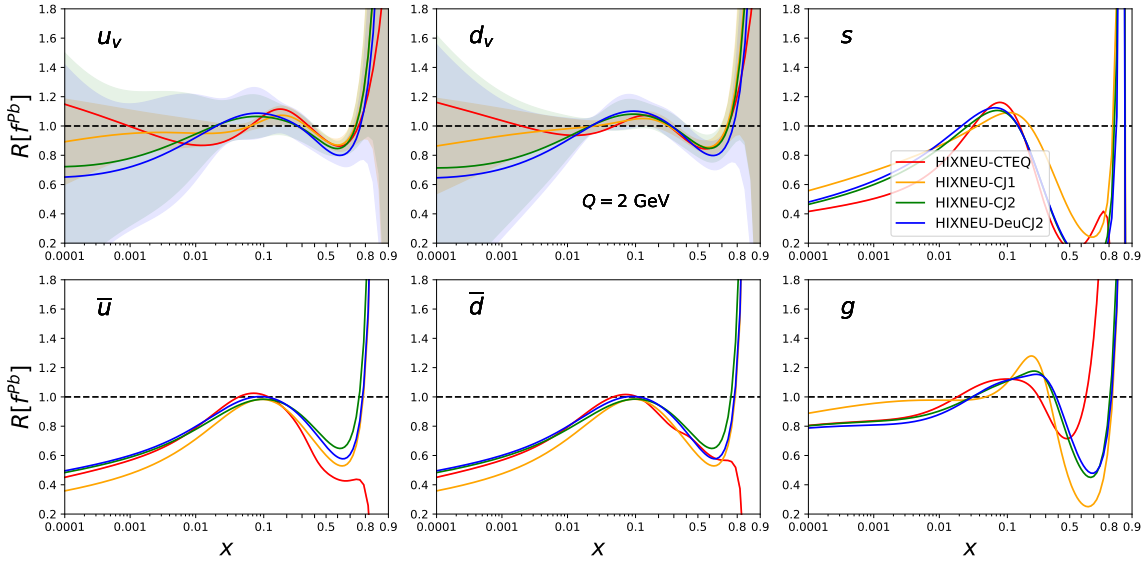


FIGURE B.11: Nuclear ratio for lead at $Q = 2$ GeV from the HIXNEU fits. The denominator for the ratio is computed using the CJ15 NLO PDFs.

crossing points (a point where $R[f] = 0$), which mark the transition from shadowing to anti-shadowing, vary quite rapidly with A. This is especially true for light nuclei. These varying crossing points are not observed in the HIXNEU-CTEQ, HIXNEU-CJ1, and EPPS21 nPDFs. Moving to the anti-shadowing region, we can see that the HIXNEU fits show monotonically increasing anti-shadowing strength as A increases. This pattern is not observed in the EPPS21 d_v PDFs. Moving to the EMC and Fermi motion regions, we see stronger nuclear corrections as A increases for all the fits.

Considering that all these fits have a good χ^2/N with a decent χ^2/N distribution per data sets, the nuclear data is not precise nor numerous enough to justify a *strong* preference of one A-dependent parametrization over the other[3]. Nevertheless, the HIXNEU-CJ2 fit, with *Amode2* parameterization, has slightly better overall χ^2 (the total χ^2 in HIXNEU-CJ2 fit is 80 points, or 3%, smaller than that of HIXNEU-CJ1 fit). Therefore, in this respect, *Amode2* parameterization is slightly better.

[3] As an example here, the d_v in the shadowing region from HIXNEU-CTEQ, with *Amode1* for the A-dependence, has rather unusual $R[d_v]$ shapes : lead ^{208}Pb is less shadowed than ^{4}He, while in the other fits, it is the opposite.

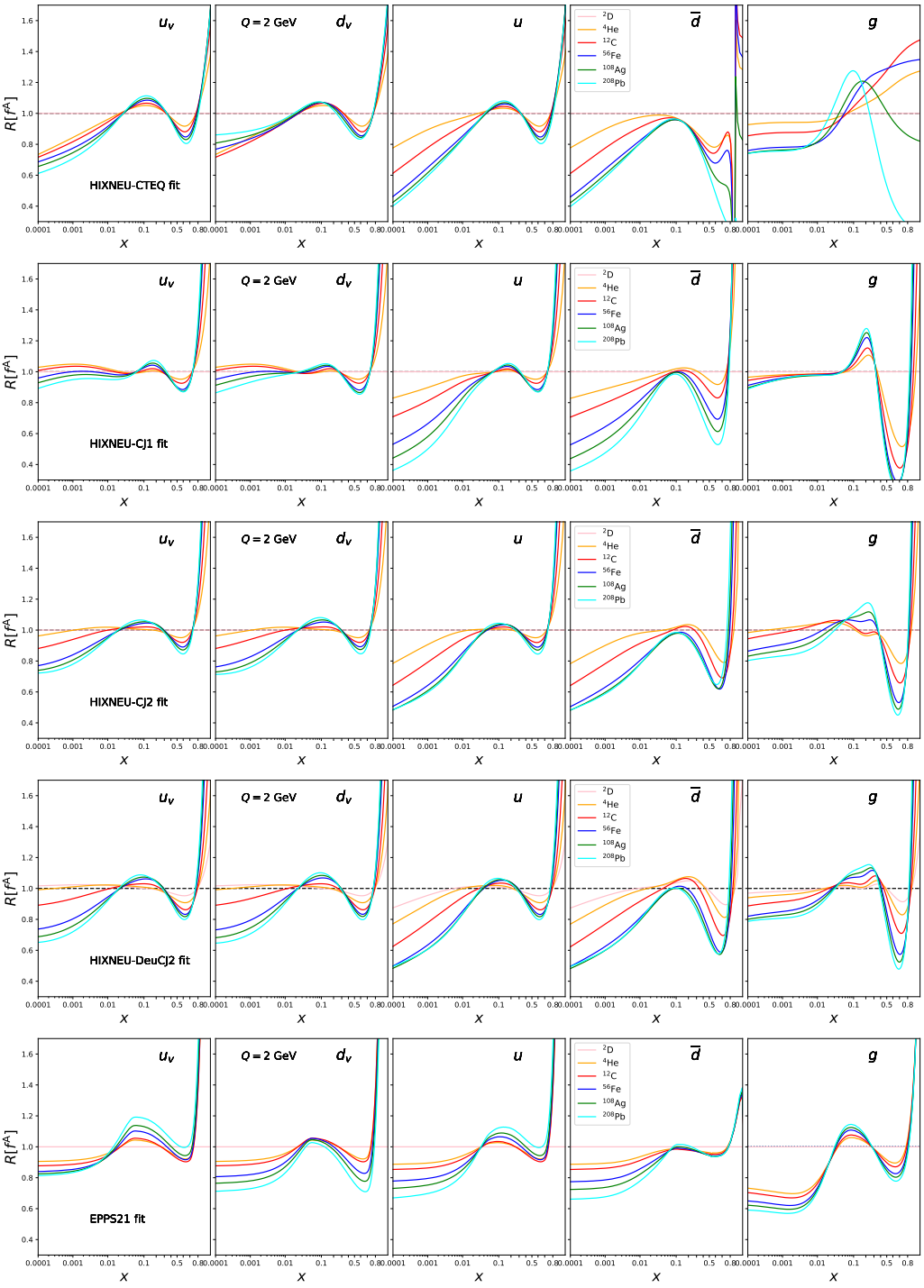


FIGURE B.12: Nuclear correction ratios for ^{2}D, ^{4}He, ^{12}C, ^{56}Fe, ^{108}Ag nuclei computed using the HIXNEU and EPPS21 (the lowest panel) fits. Due to isospin symmetry, $u \sim d$ and $\bar{u} \sim \bar{d}$. Therefore, one should keep in mind that the nuclear ratio for $\bar{u}$ and d are similar to that of u and $\bar{d}$ respectively.

Bibliography

[1] S. L. Glashow. "Partial Symmetries of Weak Interactions". In: *Nucl. Phys.* 22 (1961), pp. 579–588. DOI: 10.1016/0029-5582(61)90469-2.

[2] Steven Weinberg. "A Model of Leptons". In: *Phys. Rev. Lett.* 19 (1967), pp. 1264–1266. DOI: 10.1103/PhysRevLett.19.1264.

[3] Abdus Salam. "Weak and Electromagnetic Interactions". In: *Conf. Proc. C* 680519 (1968), pp. 367–377. DOI: 10.1142/9789812795915_0034.

[4] Murray Gell-Mann. "A Schematic Model of Baryons and Mesons". In: *Phys. Lett.* 8 (1964), pp. 214–215. DOI: 10.1016/S0031-9163(64)92001-3.

[5] Murray Gell-Mann. "The Eightfold Way: A Theory of strong interaction symmetry". In: (Mar. 1961). DOI: 10.2172/4008239.

[6] John C. Collins, Davison E. Soper, and George F. Sterman. "Factorization of Hard Processes in QCD". In: *Adv. Ser. Direct. High Energy Phys.* 5 (1989), pp. 1–91. DOI: 10.1142/9789814503266_0001. arXiv: hep-ph/0409313.

[7] Tie-Jiun Hou et al. "New CTEQ global analysis of quantum chromodynamics with high-precision data from the LHC". In: (2019). arXiv: 1912.10053 [hep-ph].

[8] S. Bailey et al. "Parton distributions from LHC, HERA, Tevatron and fixed target data: MSHT20 PDFs". In: *Eur. Phys. J. C* 81.4 (2021), p. 341. DOI: 10.1140/epjc/s10052-021-09057-0. arXiv: 2012.04684 [hep-ph].

[9] Richard D. Ball et al. "The path to proton structure at 1% accuracy". In: *Eur. Phys. J. C* 82.5 (2022), p. 428. DOI: 10.1140/epjc/s10052-022-10328-7. arXiv: 2109.02653 [hep-ph].

[10] A. Accardi et al. "Constraints on large-x parton distributions from new weak boson production and deep-inelastic scattering data". In: *Phys. Rev. D* 93.11 (2016), p. 114017. DOI: 10.1103/PhysRevD.93.114017. arXiv: 1602.03154 [hep-ph].

[11] J. J. Aubert et al. "The ratio of the nucleon structure functions $F_2{}^N$ for iron and deuterium". In: *Physics Letters B* 123.3-4 (Mar. 1983), pp. 275–278. DOI: 10.1016/0370-2693(83)90437-9.

[12] M. Arneodo et al. "Measurement of the proton and deuteron structure functions, F2(p) and F2(d), and of the ratio sigma-L / sigma-T". In: *Nucl. Phys. B* 483 (1997), pp. 3–43. DOI: 10.1016/S0550-3213(96)00538-X. arXiv: hep-ph/9610231.

[13] M. Arneodo et al. "The Structure Function ratios F2(li) / F2(D) and F2(C) / F2(D) at small x". In: *Nucl. Phys. B* 441 (1995), pp. 12–30. DOI: 10.1016/0550-3213(95)00023-2. arXiv: hep-ex/9504002.

[14] A. Bodek et al. "Electron Scattering from Nuclear Targets and Quark Distributions in Nuclei". In: *Phys. Rev. Lett.* 50 (1983), p. 1431. DOI: 10.1103/PhysRevLett.50.1431.

[15] J. Gomez et al. "Measurement of the A-dependence of deep inelastic electron scattering". In: *Phys. Rev. D* 49 (1994), pp. 4348–4372. DOI: 10.1103/PhysRevD.49.4348.

[16] S. Dasu et al. "Measurement of kinematic and nuclear dependence of R = sigma-L / sigma-t in deep inelastic electron scattering". In: *Phys. Rev. D* 49 (1994), pp. 5641–5670. DOI: 10.1103/PhysRevD.49.5641.

[17] A. C. Benvenuti et al. "Nuclear Effects in Deep Inelastic Muon Scattering on Deuterium and Iron Targets". In: *Phys. Lett. B* 189 (1987), pp. 483–487. DOI: 10.1016/0370-2693(87)90664-2.

[18] G. Bari et al. "A Measurement of Nuclear Effects in Deep Inelastic Muon Scattering on Deuterium, Nitrogen and Iron Targets". In: *Phys. Lett. B* 163 (1985), p. 282. DOI: 10.1016/0370-2693(85)90238-2.

[19] J. Seely et al. "New measurements of the EMC effect in very light nuclei". In: *Phys. Rev. Lett.* 103 (2009), p. 202301. DOI: 10.1103/PhysRevLett.103.202301. arXiv: 0904.4448 [nucl-ex].

[20] B. Schmookler et al. "Modified structure of protons and neutrons in correlated pairs". In: *Nature* 566.7744 (2019), pp. 354–358. DOI: 10.1038/s41586-019-0925-9. arXiv: 2004.12065 [nucl-ex].

[21] Richard D. Ball et al. "The Path to Proton Structure at One-Percent Accuracy". In: (Sept. 2021). arXiv: 2109.02653 [hep-ph].

[22] Marina Walt, Ilkka Helenius, and Werner Vogelsang. "A QCD analysis for nuclear PDFs at NNLO". In: *PoS* DIS2019 (2019), p. 039. DOI: 10.22323/1.352.0039. arXiv: 1908.04983 [hep-ph].

[23] Rabah Abdul Khalek, Jacob J. Ethier, and Juan Rojo. "Nuclear parton distributions from lepton-nucleus scattering and the impact of an electron-ion collider". In: *Eur. Phys. J.* C79.6 (2019), p. 471. DOI: 10.1140/epjc/s10052-019-6983-1. arXiv: 1904.00018 [hep-ph].

[24] Kari J. Eskola et al. "EPPS21: a global QCD analysis of nuclear PDFs". In: *Eur. Phys. J. C* 82.5 (2022), p. 413. DOI: 10.1140/epjc/s10052-022-10359-0. arXiv: 2112.12462 [hep-ph].

[25] K. Kovařík et al. "nCTEQ15 - Global analysis of nuclear parton distributions with uncertainties in the CTEQ framework". In: *Phys. Rev.* D93.8 (2016), p. 085037. DOI: 10.1103/PhysRevD.93.085037. arXiv: 1509.00792 [hep-ph].

[26] Rabah Abdul Khalek et al. "nNNPDF2.0: Quark Flavor Separation in Nuclei from LHC Data". In: (June 2020). arXiv: 2006.14629 [hep-ph].

[27] O. W. Greenberg. "Spin and Unitary-Spin Independence in a Paraquark Model of Baryons and Mesons". In: *Phys. Rev. Lett.* 13 (20 1964), pp. 598–602. DOI: 10.1103/PhysRevLett.13.598. URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.598.

[28] M. Y. Han and Y. Nambu. "Three-Triplet Model with Double SU(3) Symmetry". In: *Phys. Rev.* 139 (4B 1965), B1006–B1010. DOI: 10.1103/PhysRev.139.B1006. URL: https://link.aps.org/doi/10.1103/PhysRev.139.B1006.

[29] H. Lehmann, Kurt Symanzik, and Wolf Zimmermann. "Zur Formulierung quantisierter Feldtheorien". In: *Il Nuovo Cimento (1955-1965)* 1 (1955), pp. 205–225.

[30] Matthew D. Schwartz. *Quantum Field Theory and the Standard Model*. Cambridge University Press, Mar. 2014. ISBN: 978-1-107-03473-0, 978-1-107-03473-0.

[31] Michael E. Peskin and Daniel V. Schroeder. *An Introduction to quantum field theory*. Reading, USA: Addison-Wesley, 1995. ISBN: 978-0-201-50397-5.

[32] Siegfried Bethke. "The 2009 World Average of alpha(s)". In: *Eur. Phys. J. C* 64 (2009). Ed. by Douglas H. Beck, Dieter Haidt, and John W. Negele, pp. 689–703. DOI: 10.1140/epjc/s10052-009-1173-1. arXiv: 0908.1135 [hep-ph].

[33] J. D. Bjorken and E. A. Paschos. "Inelastic Electron-Proton and γ-Proton Scattering and the Structure of the Nucleon". In: *Physical Review* 185.5 (Sept. 1969), pp. 1975–1982. DOI: 10.1103/PhysRev.185.1975.

[34] Kenneth G. Wilson. "Nonlagrangian models of current algebra". In: *Phys. Rev.* 179 (1969), pp. 1499–1512. DOI: 10.1103/PhysRev.179.1499.

[35] R. P. Feynman. *High Energy Collisions: International Conference Proceedings*. Gordon & Breach, 1970. ISBN: 978-0-677-13950-0.

[36] Guido Altarelli and G. Parisi. "Asymptotic Freedom in Parton Language". In: *Nucl. Phys. B* 126 (1977), pp. 298–318. DOI: 10.1016/0550-3213(77)90384-4.

[37] V. N. Gribov and L. N. Lipatov. "Deep inelastic e p scattering in perturbation theory". In: *Sov. J. Nucl. Phys.* 15 (1972), pp. 438–450.

[38] Yuri L. Dokshitzer. "Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics." In: *Sov. Phys. JETP* 46 (1977), pp. 641–653.

[39] Hannu Paukkunen. "Global analysis of nuclear parton distribution functions at leading and next-to-leading order perturbative QCD". Other thesis. June 2009. arXiv: 0906.2529 [hep-ph].

[40] Huey-Wen Lin et al. "Parton distributions and lattice QCD calculations: a community white paper". In: *Prog. Part. Nucl. Phys.* 100 (2018), pp. 107–160. DOI: 10.1016/j.ppnp.2018.01.007. arXiv: 1711.07916 [hep-ph].

[41] Michele Arneodo. "Nuclear effects in structure functions". In: *Phys. Rept.* 240 (1994), pp. 301–393. DOI: 10.1016/0370-1573(94)90048-5.

[42] Nestor Armesto. "Nuclear shadowing". In: *J. Phys. G* 32 (2006), R367–R394. DOI: 10.1088/0954-3899/32/11/R01. arXiv: hep-ph/0604108.

[43] I. Schienbein et al. "PDF Nuclear Corrections for Charged and Neutral Current Processes". In: *Phys. Rev. D* 80 (2009), p. 094004. DOI: 10.1103/PhysRevD.80.094004. arXiv: 0907.2357 [hep-ph].

[44] Stanley J. Brodsky and Hung Jung Lu. "Shadowing and Antishadowing of Nuclear Structure Functions". In: *Phys. Rev. Lett.* 64 (1990), p. 1342. DOI: 10.1103/PhysRevLett.64.1342.

[45] Stanley J. Brodsky, Ivan Schmidt, and Jian-Jun Yang. "Nuclear antishadowing in neutrino deep inelastic scattering". In: *Phys. Rev. D* 70 (2004), p. 116003. DOI: 10.1103/PhysRevD.70.116003. arXiv: hep-ph/0409279.

[46] Stanley J. Brodsky, Valery E. Lyubovitskij, and Ivan Schmidt. "Novel corrections to the momentum sum rule for nuclear structure functions". In: *Phys. Lett. B* 824 (2022), p. 136812. DOI: 10.1016/j.physletb.2021.136812. arXiv: 2110.13682 [hep-ph].

[47] Donald F. Geesaman, K. Saito, and Anthony William Thomas. "The nuclear EMC effect". In: *Ann. Rev. Nucl. Part. Sci.* 45 (1995), pp. 337–390. DOI: 10.1146/annurev.ns.45.120195.002005.

[48] S. V. Akulinichev, Sergey A. Kulagin, and G. M. Vagradov. "The Role of Nuclear Binding in Deep Inelastic Lepton Nucleon Scattering". In: *Phys. Lett. B* 158 (1985), pp. 485–488. DOI: 10.1016/0370-2693(85)90799-3.

[49] S. V. Akulinichev et al. "LEPTON NUCLEUS DEEP INELASTIC SCATTERING". In: *Phys. Rev. Lett.* 55 (1985), pp. 2239–2241. DOI: 10.1103/PhysRevLett.55.2239.

[50] Edmond L. Berger, F. Coester, and Robert B. Wiringa. "Pion Density in Nuclei and Deep Inelastic Lepton Scattering". In: *Phys. Rev. D* 29 (1984), p. 398. DOI: 10.1103/PhysRevD.29.398.

[51] Magda Ericson and Anthony William Thomas. "Pionic Corrections and the EMC Enhancement of the Sea in Iron". In: *Phys. Lett. B* 128 (1983), pp. 112–116. DOI: 10.1016/0370-2693(83)90085-0.

[52] C. E. Carlson and T. J. Havens. "Quark Distributions in Nuclei". In: *Phys. Rev. Lett.* 51 (1983), p. 261. DOI: 10.1103/PhysRevLett.51.261.

[53] R. L. Jaffe. "Quark Distributions in Nuclei". In: *Phys. Rev. Lett.* 50 (1983), p. 228. DOI: 10.1103/PhysRevLett.50.228.

[54] F. E. Close, R. G. Roberts, and Graham G. Ross. "The Effect of Confinement Size on Nuclear Structure Functions". In: *Phys. Lett. B* 129 (1983), pp. 346–350. DOI: 10.1016/0370-2693(83)90679-2.

[55] O. Nachtmann and H. J. Pirner. "Color Conductivity in Nuclei and the Emc Effect". In: *Z. Phys. C* 21 (1984), p. 277. DOI: 10.1007/BF01577042.

[56] Claudio Ciofi Degli Atti and S. Liuti. "On the Effects of Nucleon Binding and Correlations in Deep Inelastic Electron Scattering by Nuclei". In: *Phys. Lett. B* 225 (1989), pp. 215–221. DOI: 10.1016/0370-2693(89)90808-3.

[57] Claudio Ciofi degli Atti and S. Liuti. "Realistic microscopic approach to deep inelastic scattering of electrons off few nucleon systems". In: *Phys. Rev. C* 41 (1990), pp. 1100–1114. DOI: 10.1103/PhysRevC.41.1100.

[58] Sergey A. Kulagin and R. Petti. "Global study of nuclear structure functions". In: *Nucl. Phys. A* 765 (2006), pp. 126–187. DOI: 10.1016/j.nuclphysa.2005.10.011. arXiv: hep-ph/0412425.

[59] M. Sajjad Athar, I. Ruiz Simó, and M.J. Vicente Vacas. "Nuclear medium modification of the F2(x,Q2) structure function". In: *Nuclear Physics A* 857.1 (2011), pp. 29–41. ISSN: 0375-9474. DOI: https://doi.org/10.1016/j.nuclphysa.2011.03.008. URL: https://www.sciencedirect.com/science/article/pii/S0375947411002430.

[60] E. P. Segarra et al. "Extending nuclear PDF analyses into the high-x , low-Q^2 region". In: *Phys. Rev. D* 103.11 (2021), p. 114015. DOI: 10.1103/PhysRevD.103.114015. arXiv: 2012.11566 [hep-ph].

[61] R. B. Wiringa, V. G. J. Stoks, and R. Schiavilla. "Accurate nucleon-nucleon potential with charge-independence breaking". In: *Phys. Rev. C* 51 (1 1995), pp. 38–51. DOI: 10.1103/PhysRevC.51.38. URL: https://link.aps.org/doi/10.1103/PhysRevC.51.38.

[62] R. Machleidt. "High-precision, charge-dependent Bonn nucleon-nucleon potential". In: *Phys. Rev. C* 63 (2 2001), p. 024001. DOI: 10.1103/PhysRevC.63.024001. URL: https://link.aps.org/doi/10.1103/PhysRevC.63.024001.

[63] Franz Gross and Alfred Stadler. "Covariant spectator theory of *np* scattering: Phase shifts obtained from precision fits to data below 350 MeV". In: *Phys. Rev. C* 78 (1 2008), p. 014005. DOI: 10.1103/PhysRevC.78.014005. URL: https://link.aps.org/doi/10.1103/PhysRevC.78.014005.

[64] P. Duwentäster et al. "Impact of heavy quark and quarkonium data on nuclear gluon PDFs". In: (Apr. 2022). arXiv: 2204.09982 [hep-ph].

[65] J. F. Owens et al. "The Impact of new neutrino DIS and Drell-Yan data on large-x parton distributions". In: *Phys. Rev.* D75 (2007), p. 054030. DOI: 10.1103/PhysRevD.75.054030. arXiv: hep-ph/0702159 [HEP-PH].

[66] F. James and M. Winkler. "C++ MINUIT User's Guide". Part of the CERN ROOT package. URL: https://root.cern.ch/root/htmldoc/guides/minuit2/Minuit2.pdf.

[67] Jack Sherman and Winifred J. Morrison. "Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix". In: *The Annals of Mathematical Statistics* 21.1 (1950), pp. 124 –127. DOI: 10.1214/aoms/1177729893. URL: https://doi.org/10.1214/aoms/1177729893.

[68] G. D'Agostini. "On the use of the covariance matrix to fit correlated data". In: *Nucl. Instrum. Meth. A* 346 (1994), pp. 306–311. DOI: 10.1016/0168-9002(94)90719-6.

[69] D. Stump et al. "Uncertainties of predictions from parton distribution functions. 1. The Lagrange multiplier method". In: *Phys. Rev. D* 65 (2001), p. 014012. DOI: 10.1103/PhysRevD.65.014012. arXiv: hep-ph/0101051.

[70] M. Tzanov et al. "Precise measurement of neutrino and anti-neutrino differential cross sections". In: *Phys. Rev.* D74 (2006), p. 012008. DOI: 10.1103/PhysRevD.74.012008. arXiv: hep-ex/0509010 [hep-ex].

[71] G. Onengut et al. "Measurement of nucleon structure functions in neutrino scattering". In: *Phys. Lett. B* 632 (2006), pp. 65–75. DOI: 10.1016/j.physletb.2005.10.062.

[72] K. F. Muzakka et al. "Compatibility of Neutrino DIS Data and Its Impact on Nuclear Parton Distribution Functions". In: (Apr. 2022). arXiv: 2204.13157 [hep-ph].

[73] Richard D. Ball et al. "Fitting Parton Distribution Data with Multiplicative Normalization Uncertainties". In: *JHEP* 05 (2010), p. 075. DOI: 10.1007/JHEP05(2010)075. arXiv: 0912.2276 [hep-ph].

[74] Ilkka Helenius, Marina Walt, and Werner Vogelsang. "NNLO nuclear parton distribution functions with electroweak-boson production data from the LHC". In: *Phys. Rev. D* 105.9 (2022), p. 094031. DOI: 10.1103/PhysRevD.105.094031. arXiv: 2112.11904 [hep-ph].

[75] Rabah Abdul Khalek et al. "nNNPDF3.0: evidence for a modified partonic structure in heavy nuclei". In: *Eur. Phys. J. C* 82.6 (2022), p. 507. DOI: 10.1140/epjc/s10052-022-10417-7. arXiv: 2201.12363 [hep-ph].

[76] Yudi Pawitan. "In All Likelihood: Statistical Modelling and Inference Using Likelihood". In: (Jan. 2006).

[77] Frederick James. *Statistical Methods in Experimental Physics*. 2nd. WORLD SCIENTIFIC, 2006. DOI: 10.1142/6096. eprint: https://www.worldscientific.com/doi/pdf/10.1142/6096. URL: https://www.worldscientific.com/doi/abs/10.1142/6096.

[78] D. R. Cox. *Principles of Statistical Inference*. Cambridge University Press, 2006. DOI: 10.1017/CBO9780511813559.

[79] A. Kusina et al. "Impact of LHC vector boson production in heavy ion collisions on strange PDFs". In: *Eur. Phys. J. C* 80.10 (2020), p. 968. DOI: 10.1140/epjc/s10052-020-08532-4. arXiv: 2007.09100 [hep-ph].

[80] Kari J. Eskola et al. "EPPS16: Nuclear parton distributions with LHC data". In: *Eur. Phys. J.* C77.3 (2017), p. 163. DOI: 10.1140/epjc/s10052-017-4725-9. arXiv: 1612.05741 [hep-ph].

[81] Walter T. Giele and Stephane Keller. "Implications of hadron collider observables on parton distribution function uncertainties". In: *Phys. Rev. D* 58 (1998), p. 094023. DOI: 10.1103/PhysRevD.58.094023. arXiv: hep-ph/9803393.

[82] Hannu Paukkunen and Pia Zurita. "PDF reweighting in the Hessian matrix approach". In: *JHEP* 12 (2014), p. 100. DOI: 10.1007/JHEP12(2014)100. arXiv: 1402.6623 [hep-ph].

[83] Karol Kovařík, Pavel M. Nadolsky, and Davison E. Soper. "Hadron structure in high-energy collisions". In: (May 2019). arXiv: 1905.06957 [hep-ph].

[84] P. Duwentäster et al. "Impact of inclusive hadron production data on nuclear gluon PDFs". In: (May 2021). arXiv: 2105.09873 [hep-ph].

[85] I. Schienbein et al. "Nuclear parton distribution functions from neutrino deep inelastic scattering". In: *Phys. Rev. D* 77 (2008), p. 054013. DOI: 10.1103/PhysRevD.77.054013. arXiv: 0710.4897 [hep-ph].

[86] H. Abramowicz et al. "A Parametrization of sigma-T (gamma* p) above the resonance region Q**2 >= 0". In: *Phys. Lett. B* 269 (1991), pp. 465–476. DOI: 10.1016/0370-2693(91)90202-2.

[87] K. Kovarik et al. "Nuclear Corrections in Neutrino-Nucleus DIS and Their Compatibility with Global NPDF Analyses". In: *Phys. Rev. Lett.* 106 (2011), p. 122301. DOI: 10.1103/PhysRevLett.106.122301. arXiv: 1012.0286 [hep-ph].

[88] Daniel de Florian et al. "Global Analysis of Nuclear Parton Distributions". In: *Phys. Rev.* D85 (2012), p. 074028. DOI: 10.1103/PhysRevD.85.074028. arXiv: 1112.6324 [hep-ph].

[89] Hannu Paukkunen and Carlos A. Salgado. "Agreement of Neutrino Deep Inelastic Scattering Data with Global Fits of Parton Distributions". In: *Phys. Rev. Lett.* 110.21 (2013), p. 212301. DOI: 10.1103/PhysRevLett.110.212301. arXiv: 1302.2001 [hep-ph].

[90] Narbe Kalantarians, Cynthia Keppel, and M. Eric Christy. "Comparison of the Structure Function F2 as Measured by Charged Lepton and Neutrino Scattering from Iron Targets". In: *Phys. Rev.* C96.3 (2017), p. 032201. DOI: 10.1103/PhysRevC.96.032201. arXiv: 1706.02002 [hep-ph].

[91] M. A. G. Aivazis, Frederick I. Olness, and Wu-Ki Tung. "Leptoproduction of heavy quarks. 1. General formalism and kinematics of charged current and neutral current production processes". In: *Phys. Rev. D* 50 (1994), pp. 3085–3101. DOI: 10.1103/PhysRevD.50.3085. arXiv: hep-ph/9312318.

[92] M. A. G. Aivazis et al. "Leptoproduction of heavy quarks. 2. A Unified QCD formulation of charged and neutral current processes from fixed target to collider energies". In: *Phys. Rev. D* 50 (1994), pp. 3102–3118. DOI: 10.1103/PhysRevD.50.3102. arXiv: hep-ph/9312319.

[93] L.W. Whitlow et al. "A precise extraction of R=σL/σT from a global analysis of the SLAC deep inelastic e-p and e-d scattering cross sections". In: *Physics Letters B* 250.1 (1990), pp. 193–198. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(90)91176-C. URL: https://www.sciencedirect.com/science/article/pii/037026939091176C.

[94] Alan D. Martin et al. "Parton distributions: A New global analysis". In: *Eur. Phys. J. C* 4 (1998), pp. 463–496. DOI: 10.1007/s100520050220. arXiv: hep-ph/9803445.

[95] Un-Ki Yang. "A Measurement of Differential Cross Sections in Charged Current Neutrino Interactions on Iron and a Global Structure Functions Analysis". In: (2001). DOI: 10.2172/1421409.

[96] C. Boros, J. T. Londergan, and Anthony William Thomas. "Evidence for charge symmetry violation in parton distributions". In: *Phys. Rev. D* 59 (1999), p. 074021. DOI: 10.1103/PhysRevD.59.074021. arXiv: hep-ph/9810220.

[97] Stanley J. Brodsky and Bo-Qiang Ma. "The Quark / anti-quark asymmetry of the nucleon sea". In: *Phys. Lett. B* 381 (1996), pp. 317–324. DOI: 10.1016/0370-2693(96)00597-7. arXiv: hep-ph/9604393.

[98] J.P. Berge et al. "A Measurement of Differential Cross-Sections and Nucleon Structure Functions in Charged Current Neutrino Interactions on Iron". In: *Z. Phys. C* 49 (1991), pp. 187–224. DOI: 10.1007/BF01555493.

[99] M. Goncharov et al. "Precise Measurement of Dimuon Production Cross-Sections in ν_μ Fe and $\bar{\nu}_\mu$ Fe Deep Inelastic Scattering at the Tevatron." In: *Phys. Rev.* D64 (2001), p. 112006. DOI: 10.1103/PhysRevD.64.112006. arXiv: hep-ex/0102049 [hep-ex].

[100] J. Mousseau et al. "Measurement of partonic nuclear effects in deep-inelastic neutrino scattering using MINERvA". In: *Phys. Rev. D* 93 (7 2016), p. 071101. DOI: 10.1103/PhysRevD.93.071101. URL: https://link.aps.org/doi/10.1103/PhysRevD.93.071101.

[101] Q. Wu et al. "A Precise measurement of the muon neutrino-nucleon inclusive charged current cross-section off an isoscalar target in the energy range 2.5 $<$ E(nu) $<$ 40-GeV by NOMAD". In: *Phys. Lett. B* 660 (2008), pp. 19–25. DOI: 10.1016/j.physletb.2007.12.027. arXiv: 0711.1183 [hep-ex].

[102] R. Petti. "Cross-section measurements in the NOMAD experiment". In: *Nucl. Phys. B Proc. Suppl.* 159 (2006). Ed. by A. I. Studenikin, pp. 56–62. DOI: 10.1016/j.nuclphysbps.2006.08.026. arXiv: hep-ex/0602022.

[103] A. Kayis-Topaksu et al. "Leading order analysis of neutrino induced dimuon events in the CHORUS experiment". In: *Nucl. Phys. B* 798 (2008), pp. 1–16. DOI: 10.1016/j.nuclphysb.2008.02.013. arXiv: 0804.1869 [hep-ex].

[104] O. Samoylov et al. "A Precision Measurement of Charm Dimuon Production in Neutrino Interactions from the NOMAD Experiment". In: *Nucl. Phys. B* 876 (2013), pp. 339–375. DOI: 10.1016/j.nuclphysb.2013.08.021. arXiv: 1308.4750 [hep-ex].

[105] A. Accardi et al. "Deuterium scattering experiments in CTEQ global QCD analyses: a comparative investigation". In: *Eur. Phys. J. C* 81.7 (2021), p. 603. DOI: 10.1140/epjc/s10052-021-09318-y. arXiv: 2102.01107 [hep-ph].

[106] A. Airapetian et al. "Measurement of R = sigma(L) / sigma(T) in deep inelastic scattering on nuclei". In: (Oct. 2002). arXiv: hep-ex/0210068.

[107] P. Amaudruz et al. "A Reevaluation of the nuclear structure function ratios for D, He, Li-6, C and Ca". In: *Nucl. Phys. B* 441 (1995), pp. 3–11. DOI: 10.1016/0550-3213(94)00023-9. arXiv: hep-ph/9503291.

[108] M. R. Adams et al. "Shadowing in inelastic scattering of muons on carbon, calcium and lead at low x(Bj)". In: *Z. Phys. C* 67 (1995), pp. 403–410. DOI: 10.1007/BF01624583. arXiv: hep-ex/9505006.

[109] J. Ashman et al. "Measurement of the Ratios of Deep Inelastic Muon - Nucleus Cross-Sections on Various Nuclei Compared to Deuterium". In: *Phys. Lett. B* 202 (1988), pp. 603–610. DOI: 10.1016/0370-2693(88)91872-2.

[110] M. Arneodo et al. "Measurements of the nucleon structure function in the range $0.002 - GeV^2 < x < 0.17 - GeV^2$ and $0.2 - GeV^2 < q^2 < 8 - GeV^2$ in deuterium, carbon and calcium". In: *Nucl. Phys. B* 333 (1990), pp. 1–47. DOI: 10.1016/0550-3213(90)90221-X.

[111] A. Bodek et al. "A Comparison of the Deep Inelastic Structure Functions of Deuterium and Aluminum Nuclei". In: *Phys. Rev. Lett.* 51 (1983), p. 534. DOI: 10.1103/PhysRevLett.51.534.

[112] J. Ashman et al. "A Measurement of the ratio of the nucleon structure function in copper and deuterium". In: *Z. Phys. C* 57 (1993), pp. 211–218. DOI: 10.1007/BF01565050.

[113] M. R. Adams et al. "Saturation of shadowing at very low x_{BJ}". In: *Phys. Rev. Lett.* 68 (1992), pp. 3266–3269. DOI: 10.1103/PhysRevLett.68.3266.

[114] M. Arneodo et al. "The A dependence of the nuclear structure function ratios". In: *Nucl. Phys. B* 481 (1996), pp. 3–22. DOI: 10.1016/S0550-3213(96)90117-0.

[115] M. Arneodo et al. "The Q**2 dependence of the structure function ratio F2 Sn / F2 C and the difference R Sn - R C in deep inelastic muon scattering". In: *Nucl. Phys. B* 481 (1996), pp. 23–39. DOI: 10.1016/S0550-3213(96)90119-4.

[116] D. M. Alde et al. "Nuclear dependence of dimuon production at 800-GeV. FNAL-772 experiment". In: *Phys. Rev. Lett.* 64 (1990), pp. 2479–2482. DOI: 10.1103/PhysRevLett.64.2479.

[117] M. A. Vasilev et al. "Parton energy loss limits and shadowing in Drell-Yan dimuon production". In: *Phys. Rev. Lett.* 83 (1999), pp. 2304–2307. DOI: 10.1103/PhysRevLett.83.2304. arXiv: hep-ex/9906010.

[118] S.S. Adler et al. "Centrality dependence of pi0 and eta production at large transverse momentum in s(NN)**(1/2) = 200-GeV d+Au collisions". In: *Phys. Rev. Lett.* 98 (2007), p. 172302. DOI: 10.1103/PhysRevLett.98.172302. arXiv: nucl-ex/0610036.

[119] A. Adare et al. "Spectra and ratios of identified particles in Au+Au and d+Au collisions at $\sqrt{s_{NN}} = 200$ GeV". In: *Phys. Rev. C* 88.2 (2013), p. 024906. DOI: 10.1103/PhysRevC.88.024906. arXiv: 1304.3410 [nucl-ex].

[120] B. I. Abelev et al. "Inclusive π^0, η, and direct photon production at high transverse momentum in $p + p$ and d+Au collisions at $\sqrt{s_{NN}} = 200$ GeV". In: *Phys. Rev. C* 81 (2010), p. 064904. DOI: 10.1103/PhysRevC.81.064904. arXiv: 0912.3838 [hep-ex].

[121] J. Adams et al. "Identified hadron spectra at large transverse momentum in p+p and d+Au collisions at $\sqrt{s_{NN}}$ = 200 GeV". In: *Phys. Lett. B* 637 (2006), pp. 161–169. DOI: 10.1016/j.physletb.2006.04.032. arXiv: nucl-ex/0601033.

[122] S. Acharya et al. "Neutral pion and η meson production in p-Pb collisions at $\sqrt{s_{\mathrm{NN}}} =$ 5.02 TeV". In: *Eur. Phys. J. C* 78.8 (2018), p. 624. DOI: 10.1140/epjc/s10052-018-6013-8. arXiv: 1801.07051 [nucl-ex].

[123] J. Adam et al. "Multiplicity dependence of charged pion, kaon, and (anti)proton production at large transverse momentum in p-Pb collisions at $\sqrt{s_{\mathrm{NN}}}$ = 5.02 TeV". In: *Phys. Lett. B* 760 (2016), pp. 720–735. DOI: 10.1016/j.physletb.2016.07.050. arXiv: 1601.03658 [nucl-ex].

[124] S. Acharya et al. "Nuclear modification factor of light neutral-meson spectra up to high transverse momentum in p-Pb collisions at $\sqrt{s_{NN}}$ = 8.16 TeV". In: (Apr. 2021). arXiv: 2104.03116 [nucl-ex].

[125] Georges Aad et al. "Measurement of $W^{\pm}$ boson production in Pb+Pb collisions at $\sqrt{s_{\mathrm{NN}}} =$ 5.02 TeV with the ATLAS detector". In: *Eur. Phys. J. C* 79.11 (2019), p. 935. DOI: 10.1140/epjc/s10052-019-7439-3. arXiv: 1907.10414 [nucl-ex].

[126] Vardan Khachatryan et al. "Study of W boson production in pPb collisions at $\sqrt{s_{\mathrm{NN}}} =$ 5.02 TeV". In: *Phys. Lett. B* 750 (2015), pp. 565–586. DOI: 10.1016/j.physletb.2015.09.057. arXiv: 1503.05825 [nucl-ex].

[127] Albert M Sirunyan et al. "Observation of nuclear modifications in $\mathrm{W}^{\pm}$ boson production in pPb collisions at $\sqrt{s_{\mathrm{NN}}} = 8.16$ TeV". In: *Phys. Lett. B* 800 (2020), p. 135048. DOI: 10.1016/j.physletb.2019.135048. arXiv: 1905.01486 [hep-ex].

[128] Jaroslav Adam et al. "W and Z boson production in p-Pb collisions at $\sqrt{s_{\mathrm{NN}}}$ = 5.02 TeV". In: *JHEP* 02 (2017), p. 077. DOI: 10.1007/JHEP02(2017)077. arXiv: 1611.03002 [nucl-ex].

[129] Kgotlaesele Senosi. "Measurement of W-boson production in p-Pb collisions at the LHC with ALICE". In: *PoS* Bormio2015 (2015), p. 042. DOI: 10.22323/1.238.0042. arXiv: 1511.06398 [hep-ex].

[130] Georges Aad et al. "Z boson production in p+Pb collisions at $\sqrt{s_{NN}} = 5.02$ TeV measured with the ATLAS detector". In: *Phys. Rev. C* 92.4 (2015), p. 044915. DOI: 10.1103/PhysRevC.92.044915. arXiv: 1507.06232 [hep-ex].

[131] Vardan Khachatryan et al. "Study of Z boson production in pPb collisions at $\sqrt{s_{NN}} = 5.02$ TeV". In: *Phys. Lett. B* 759 (2016), pp. 36–57. DOI: 10.1016/j.physletb.2016.05.044. arXiv: 1512.06461 [hep-ex].

[132] R. Aaij et al. "Observation of Z production in proton-lead collisions at LHCb". In: *JHEP* 09 (2014), p. 030. DOI: 10.1007/JHEP09(2014)030. arXiv: 1406.2885 [hep-ex].

[133] Hannu Paukkunen and Carlos A. Salgado. "Compatibility of neutrino DIS data and global analyses of parton distribution functions". In: *JHEP* 07 (2010), p. 032. DOI: 10.1007/JHEP07(2010)032. arXiv: 1004.3140 [hep-ph].

[134] B. Z. Kopeliovich, J. G. Morfin, and Ivan Schmidt. "Nuclear Shadowing in Electro-Weak Interactions". In: *Prog. Part. Nucl. Phys.* 68 (2013), pp. 314–372. DOI: 10.1016/j.ppnp.2012.09.004. arXiv: 1208.6541 [hep-ph].

[135] H. Abramowicz et al. "Measurement of ν and $\bar{\nu}$ structure functions in hydrogen and iron". In: *Z. Phys. C* 25 (1984). Ed. by A. Meyer and E. Wieczorek, pp. 29–43. DOI: 10.1007/BF01571954.

[136] Ferran Faura et al. "The Strangest Proton?" In: *Eur. Phys. J. C* 80.12 (2020), p. 1168. DOI: 10.1140/epjc/s10052-020-08749-3. arXiv: 2009.00014 [hep-ph].

[137] H. Abramowicz et al. "Experimental Study of Opposite Sign Dimuons Produced in Neutrino and anti-neutrinos Interactions". In: *Z. Phys. C* 15 (1982), p. 19. DOI: 10.1007/BF01573422.

[138] Luis A. Anchordoqui et al. "The Forward Physics Facility: Sites, Experiments, and Physics Potential". In: (Sept. 2021). arXiv: 2109.10905 [hep-ph].

[139] A. Accardi et al. "Electron Ion Collider: The Next QCD Frontier: Understanding the glue that binds us all". In: *Eur. Phys. J. A* 52.9 (2016). Ed. by A. Deshpande, Z. E. Meziani, and J. W. Qiu, p. 268. DOI: 10.1140/epja/i2016-16268-9. arXiv: 1212.1701 [nucl-ex].

[140] R. Abdul Khalek et al. "Science Requirements and Detector Concepts for the Electron-Ion Collider: EIC Yellow Report". In: (Mar. 2021). arXiv: 2103.05419 [physics.ins-det].

[141] R. Ruiz, K.F. Muzakka, C. Leger, A. Accardi, P. Duwentäster, T.J. Hobbs,T. Ježo, C. Keppel, M. Klasen, K. Kovařík , A. Kusina, J.G. Morfín, F.I. Olness, J.F. Owens, M.H. Reno, P. Risse, I. Schienbein, J.Y. Yu. "Target Mass Corrections in Lepton-Nucleus DIS Revisited". In: *In preparation* (2022).

[142] H. Abramowicz et al. "Combination of measurements of inclusive deep inelastic $e^{\pm}p$ scattering cross sections and QCD analysis of HERA data". In: *Eur. Phys. J. C* 75.12 (2015), p. 580. DOI: 10.1140/epjc/s10052-015-3710-4. arXiv: 1506.06042 [hep-ex].

[143] Richard D. Ball et al. "Parton distributions from high-precision collider data". In: *Eur. Phys. J.* C77.10 (2017), p. 663. DOI: 10.1140/epjc/s10052-017-5199-5. arXiv: 1706.00428 [hep-ph].

[144] S. Alekhin et al. "Parton distribution functions, α_s, and heavy-quark masses for LHC Run II". In: *Phys. Rev. D* 96.1 (2017), p. 014011. DOI: 10.1103/PhysRevD.96.014011. arXiv: 1701.05838 [hep-ph].

[145] M. Hirai, S. Kumano, and T. H. Nagai. "Determination of nuclear parton distribution functions and their uncertainties in next-to-leading order". In: *Phys. Rev. C* 76 (2007), p. 065207. DOI: 10.1103/PhysRevC.76.065207. arXiv: 0709.3038 [hep-ph].

[146] K.J. Eskola, H. Paukkunen, and C.A. Salgado. "EPS09: A New Generation of NLO and LO Nuclear Parton Distribution Functions". In: *JHEP* 04 (2009), p. 065. DOI: 10.1088/1126-6708/2009/04/065. arXiv: 0902.4154 [hep-ph].

[147] A. Kusina et al. "Vector boson production in pPb and PbPb collisions at the LHC and its impact on nCTEQ15 PDFs". In: *Eur. Phys. J.* C77.7 (2017), p. 488. DOI: 10.1140/epjc/s10052-017-5036-x. arXiv: 1610.02925 [nucl-th].

[148] Rabah Abdul Khalek et al. "nNNPDF3.0: Evidence for a modified partonic structure in heavy nuclei". In: (Jan. 2022). arXiv: 2201.12363 [hep-ph].

[149] Marina Walt, Ilkka Helenius, and Werner Vogelsang. "Open-source QCD analysis of nuclear parton distribution functions at NLO and NNLO". In: *Phys. Rev.* D100.9 (2019), p. 096015. DOI: 10.1103/PhysRevD.100.096015. arXiv: 1908.03355 [hep-ph].

[150] Hamzeh Khanpour et al. "Nuclear parton distribution functions with uncertainties in a general mass variable flavor number scheme". In: *Phys. Rev. D* 104.3 (2021), p. 034010. DOI: 10.1103/PhysRevD.104.034010. arXiv: 2010.00555 [hep-ph].

[151] Dmitri Yu. Bardin et al. "QED and electroweak corrections to deep inelastic scattering". In: *Acta Phys. Polon. B* 28 (1997). Ed. by S. Jadach, M. Skrzypek, and Z. Was, pp. 511–528. arXiv: hep-ph/9611426.

[152] K. P. O. Diener, S. Dittmaier, and W. Hollik. "Electroweak radiative corrections to deep inelastic neutrino scattering: Implications for NuTeV?" In: *Phys. Rev. D* 69 (2004), p. 073005. DOI: 10.1103/PhysRevD.69.073005. arXiv: hep-ph/0310364.

[153] R. Keith Ellis, W. Furmanski, and R. Petronzio. "Unraveling Higher Twists". In: *Nucl. Phys. B* 212 (1983), p. 29. DOI: 10.1016/0550-3213(83)90597-7.

[154] S. I. Alekhin, S. A. Kulagin, and R. Petti. "Nuclear effects in the deuteron and global QCD analyses". In: *Phys. Rev. D* 105.11 (2022), p. 114037. DOI: 10.1103/PhysRevD.105.114037. arXiv: 2203.07333 [hep-ph].

[155] Riccardo Barbieri et al. "MASS CORRECTIONS TO SCALING IN DEEP INELASTIC PROCESSES". In: *Nucl. Phys.* B117 (1976), p. 50.

[156] Richard A. Brandt and Giuliano Preparata. "Operator product expansions near the light cone". In: *Nucl. Phys. B* 27 (1971), pp. 541–567. DOI: 10.1016/0550-3213(71)90265-3.

[157] Norman H. Christ, B. Hasslacher, and Alfred H. Mueller. "Light cone behavior of perturbation theory". In: *Phys. Rev. D* 6 (1972), p. 3543. DOI: 10.1103/PhysRevD.6.3543.

[158] Howard Georgi and H. David Politzer. "Freedom at Moderate Energies: Masses in Color Dynamics". In: *Phys. Rev. D* 14 (1976), p. 1829. DOI: 10.1103/PhysRevD.14.1829.

[159] Alvaro De Rujula, Howard Georgi, and H. David Politzer. "Demythification of Electroproduction, Local Duality and Precocious Scaling". In: *Annals Phys.* 103 (1977), p. 315. DOI: 10.1016/S0003-4916(97)90003-8.

[160] S. Kretzer and M. H. Reno. "Target mass corrections to electro-weak structure functions and perturbative neutrino cross sections". In: *Phys. Rev.* D69 (2004), p. 034002. eprint: hep-ph/0307023.

[161] Ingo Schienbein et al. "A Review of Target Mass Corrections". In: *J. Phys.* G35 (2008), p. 053101. DOI: 10.1088/0954-3899/35/5/053101. arXiv: 0709.1775 [hep-ph].

[162] Taizo Muta. "Foundations of Quantum Chromodynamics: An Introduction to Perturbative Methods in Gauge Theories, (2nd ed.)" In: World scientific Lecture Notes in Physics 57 (1998).

[163] K. G. Wilson and W. Zimmermann. "Operator product expansions and composite field operators in the general framework of quantum field theory". In: *Commun. Math. Phys.* 24 (1972), pp. 87–106. DOI: 10.1007/BF01878448.

[164] John C. Collins. *Renormalization: An Introduction to Renormalization, The Renormalization Group, and the Operator Product Expansion*. Vol. 26. Cambridge Monographs on Mathematical Physics. Cambridge: Cambridge University Press, 1986. ISBN: 978-0-521-31177-9, 978-0-511-86739-2. DOI: 10.1017/CBO9780511622656.

[165] Howard Georgi and H. David Politzer. "Precocious Scaling, Rescaling and xi Scaling". In: *Phys. Rev. Lett.* 36 (1976). [Erratum: Phys.Rev.Lett. 37, 68 (1976)], p. 1281. DOI: 10.1103/PhysRevLett.36.1281.

[166] Kari J. Eskola, Petja Paakkinen, and Hannu Paukkunen. "Non-quadratic improved Hessian PDF reweighting and application to CMS dijet measurements at 5.02 TeV". In: *Eur. Phys. J. C* 79.6 (2019), p. 511. DOI: 10.1140/epjc/s10052-019-6982-2. arXiv: 1903.09832 [hep-ph].

[167] Albert M Sirunyan et al. "Constraining gluon distributions in nuclei using dijets in proton-proton and proton-lead collisions at $\sqrt{s_{\mathrm{NN}}} = 5.02$ TeV". In: *Phys. Rev. Lett.* 121.6 (2018), p. 062002. DOI: 10.1103/PhysRevLett.121.062002. arXiv: 1805.04736 [hep-ex].

[168] Zoltan Nagy. "Next-to-leading order calculation of three jet observables in hadron hadron collision". In: *Phys. Rev. D* 68 (2003), p. 094002. DOI: 10.1103/PhysRevD.68.094002. arXiv: hep-ph/0307268.

[169] S. Catani and M. H. Seymour. "A General algorithm for calculating jet cross-sections in NLO QCD". In: *Nucl. Phys. B* 485 (1997). [Erratum: Nucl.Phys.B 510, 503–504 (1998)], pp. 291–419. DOI: 10.1016/S0550-3213(96)00589-5. arXiv: hep-ph/9605323.

[170] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "FastJet User Manual". In: *Eur. Phys. J. C* 72 (2012), p. 1896. DOI: 10.1140/epjc/s10052-012-1896-2. arXiv: 1111.6097 [hep-ph].

[171] Daniel Britzger et al. "New features in version 2 of the fastNLO project". In: *20th International Workshop on Deep-Inelastic Scattering and Related Subjects*. 2012, pp. 217–221. DOI: 10.3204/DESY-PROC-2012-02/165. arXiv: 1208.3641 [hep-ph].

[172] Tancredi Carli et al. "A posteriori inclusion of parton density functions in NLO QCD final-state calculations at hadron colliders: The APPLGRID Project". In: *Eur. Phys. J.* C66 (2010), pp. 503–524. DOI: 10.1140/epjc/s10052-010-1255-0. arXiv: 0911.2985 [hep-ph].

[173] A. Gehrmann-De Ridder et al. "Triple Differential Dijet Cross Section at the LHC". In: *Phys. Rev. Lett.* 123.10 (2019), p. 102001. DOI: 10.1103/PhysRevLett.123.102001. arXiv: 1905.09047 [hep-ph].

[174] A. Gehrmann-De Ridder, T. Gehrmann, and E. W. Nigel Glover. "Antenna subtraction method for jet calculations at NNLO". In: *Nucl. Phys. B Proc. Suppl.* 157 (2006). Ed. by J. Fujimoto, J. Kodaira, and T. Uematsu, pp. 32–36. DOI: 10.1016/j.nuclphysbps.2006.03.006. arXiv: hep-ph/0601145.

[175] James Currie et al. "Precise predictions for dijet production at the LHC". In: *Phys. Rev. Lett.* 119.15 (2017), p. 152001. DOI: 10.1103/PhysRevLett.119.152001. arXiv: 1705.10271 [hep-ph].

[176] Vardan Khachatryan et al. "Measurement and QCD analysis of double-differential inclusive jet cross sections in pp collisions at $\sqrt{s} = 8$ TeV and cross section ratios to 2.76 and 7 TeV". In: *JHEP* 03 (2017), p. 156. DOI: 10.1007/JHEP03(2017)156. arXiv: 1609.05331 [hep-ex].

[177] Georges Aad et al. "Measurement of dijet cross sections in pp collisions at 7 TeV centre-of-mass energy using the ATLAS detector". In: *JHEP* 05 (2014), p. 059. DOI: 10.1007/JHEP05(2014)059. arXiv: 1312.3524 [hep-ex].

[178] Albert M Sirunyan et al. "Measurement of the triple-differential dijet cross section in proton-proton collisions at $\sqrt{s} = 8$ TeV and constraints on parton distribution functions". In: *Eur. Phys. J. C* 77.11 (2017), p. 746. DOI: 10.1140/epjc/s10052-017-5286-7. arXiv: 1705.02628 [hep-ex].

[179] Sayipjamal Dulat et al. "New parton distribution functions from a global analysis of quantum chromodynamics". In: *Phys. Rev. D* 93.3 (2016), p. 033006. DOI: 10.1103/PhysRevD.93.033006. arXiv: 1506.07443 [hep-ph].

[180] S. Alekhin, J. Blümlein, and S. Moch. "NLO PDFs from the ABMP16 fit". In: *Eur. Phys. J. C* 78.6 (2018), p. 477. DOI: 10.1140/epjc/s10052-018-5947-1. arXiv: 1803.07537 [hep-ph].

[181] Richard D. Ball et al. "Parton distributions from high-precision collider data". In: *Eur. Phys. J. C* 77.10 (2017), p. 663. DOI: 10.1140/epjc/s10052-017-5199-5. arXiv: 1706.00428 [hep-ph].

[182] Rabah Abdul Khalek. "Exploring the substructure of nucleons and nuclei with machine learning". PhD thesis. Vrije U., Amsterdam, Vrije U., Amsterdam, 2021. arXiv: 2110.01924 [hep-ph].

[183] W. Melnitchouk and Anthony William Thomas. "Neutron / proton structure function ratio at large x". In: *Phys. Lett. B* 377 (1996), pp. 11–17. DOI: 10.1016/0370-2693(96)00292-4. arXiv: nucl-th/9602038.

[184] Glennys R. Farrar and Darrell R. Jackson. "Pion and Nucleon Structure Functions near $x = 1$". In: *Phys. Rev. Lett.* 35 (21 1975), pp. 1416–1419. DOI: 10.1103/PhysRevLett.35.1416. URL: https://link.aps.org/doi/10.1103/PhysRevLett.35.1416.

[185] F. E. Close and W. Melnitchouk. "Symmetry breaking and quark-hadron duality in structure functions". In: *Phys. Rev. C* 68 (3 2003), p. 035210. DOI: 10.1103/PhysRevC.68.035210. URL: https://link.aps.org/doi/10.1103/PhysRevC.68.035210.

[186] Craig D. Roberts, Roy J. Holt, and Sebastian M. Schmidt. "Nucleon spin structure at very high-x". In: *Phys. Lett. B* 727 (2013), pp. 249–254. DOI: 10.1016/j.physletb.2013.09.038. arXiv: 1308.1236 [nucl-th].

[187] A. Accardi et al. "Uncertainties in determining parton distributions at large x". In: *Phys. Rev. D* 84 (2011), p. 014008. DOI: 10.1103/PhysRevD.84.014008. arXiv: 1102.3686 [hep-ph].

[188] A. C. Benvenuti et al. "A High Statistics Measurement of the Proton Structure Functions F(2) (x, Q**2) and R from Deep Inelastic Muon Scattering at High Q**2". In: *Phys. Lett. B* 223 (1989), pp. 485–489. DOI: 10.1016/0370-2693(89)91637-7.

[189] S. I. Alekhin, S. A. Kulagin, and R. Petti. "Nuclear Effects in the Deuteron and Constraints on the d/u Ratio". In: *Phys. Rev. D* 96.5 (2017), p. 054005. DOI: 10.1103/PhysRevD.96.054005. arXiv: 1704.00204 [nucl-th].