Theoretical Uncertainties in Parton Distribution Functions

Rosalyn Laura Pearson

Doctor of Philosophy
The University of Edinburgh
May 2020

Abstract

We are now in the era of high precision particle physics, spurred on by a wealth of new data from the Large Hadron Collider (LHC). In order to match the precision set by modern experiments and test the limits of the Standard Model, we must increase the sophistication of our theoretical predictions. Much of the data available involves the interaction of protons, which are composite particles. Their interactions are described by combining perturbative Quantum Field Theory (QFT) with parton distribution functions (PDFs), which encapsulate the non-perturbative behvaiour. Increasing accuracy and precision of these PDFs is therefore of great value to modern particle physics.

PDFs are determined by multi-dimensional fits of experimental data to theoretical predictions from QFT. Uncertainties in PDFs arise from those in the experimental data and theoretical predictions, as well as from the methodology of the fit. At the current levels of precision theoretical uncertainties are increasingly significant, but have so far not been included in PDF fits. Such uncertainties arise from many sources, an important one being the truncation of the perturbative expansion for the theoretical predictions to a fixed order, resulting in missing higher order uncertainties (MHOUs).

In this thesis we consider how to include theory uncertainties in future PDF fits, and address several sources of uncertainties. MHOUs are estimated and included as a proof of concept in next-to-leading order (NLO) PDFs. We find that these capture many of the important features of the known PDFs at the next order above (NNLO). We then go on to investigate uncertainties from previously ignored heavy nuclear effects and higher twist effects, estimate their magnitude and assess the impact of their inclusion on the PDFs.

Declaration

I declare that this thesis was composed by myself, and details work carried out as a member of the NNPDF collaboration, and alongside my supervisor Richard Ball. Unless explicitly stated in the text, all results are mine or come from collaborative projects to which I have made a significant contribution. This work has not been submitted for any other degree or professional qualification.

Parts of this work have been published in [].

(*Rosalyn Laura Pearson, May 2020*)

Acknowledgements

Insert people you want to thank here.

Contents

0.1 Critical Dependencies

- Computing power on Eddie. I need to run PDF fits on the cluster but these often queue for a long time, and I have had many issues with fits crashing/the cluster being down. This is slowing my progress considerably. I can get around this somewhat by calling in favours with other NNPDF members at different universities to run some of these for me, but otherwise it is slow work.

0.2 List of Publications

- **Towards parton distribution functions with theoretical uncertainties**, Pearson, R. L. and Voisey, C. Nuclear and Particle Physics Proceedings, Volumes 300-302, July-September 2018, Pages 24-29, e-Print: 1810.01996 [hep-ph]
- **Nuclear Uncertainties in the Determination of Proton PDFs**, NNPDF Collaboration: Richard D. Ball et al. (Dec 21, 2018), Published in: Eur.Phys.J.C 79 (2019) 3, 282, e-Print: 1812.09074 [hep-ph]
- **A first determination of parton distributions with theoretical uncertainties**, NNPDF Collaboration: Rabah Abdul Khalek et al. (May 10, 2019), Published in: Eur.Phys.J. C (2019) 79:838, e-Print: 1905.04311 [hep-ph]
- **Uncertainties due to Nuclear Data in Proton PDF Fits**, Rosalyn Pearson, Richard Ball, Emanuele Roberto Nocera (Jun 14, 2019), Published in: PoS DIS2019 (2019) 027, Contribution to: DIS 2019, 027
- **Parton Distributions with Theory Uncertainties: General Formalism and First Phenomenological Studies**, NNPDF Collaboration:

Rabah Abdul Khalek et al. (Jun 25, 2019) Published in: Eur.Phys.J.C 79 (2019) 11, 931, e-Print: 1906.10698 [hep-ph]

0.3 Outline of chapter contents

The table of contents above outlines what will appear in the thesis. Chapter 1 will be based on my first year review, Chapters 2 and 3 will be based on th papers 1905.04311 and 1906.10698 and Chapter 4 will be based on 1812.09074. Chapters 5 and 6 detail projects yet to be completed.

Introduction

Chapter 1

Background - 60%

Over the past 100 years, following the discovery of the atomic nucleus by Rutherford in 1911, great strides have been made towards understanding subatomic structure. We now know that atoms are made up of hadrons (such as protons and neutrons) and leptons (such as the electron). Interactions between these constituents and with gravity come from force-carrying bosons. Probing hadrons with high energy photons shows that they are composed of quarks, which do not exist independently. This means that they can only be studied within a hadronic environment.

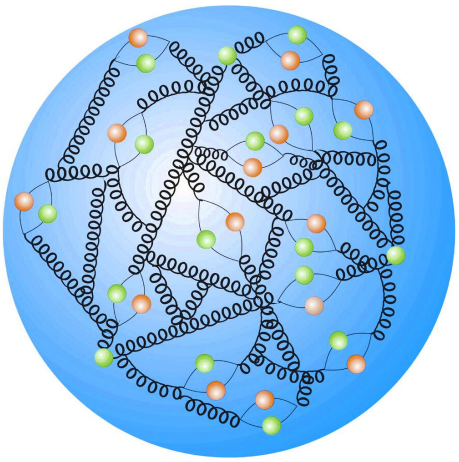

Figure 1.0.1 *A visualisation of the internal structure of the proton [1]. Quarks (the blobs) are bound together by gluons (the loopy lines).*

The Standard Model of particle physics has proven thus far to be an extremely accurate model of nature, and the current focus is on providing ever more precise experimental and theoretical results to test it and search for new physics which it can't explain.

Cutting edge particle physics experiments are currently being carried out at the Large Hadron Collider (LHC). The LHC predominantly collides protons. At a basic level we

can think of a proton as being composed of two up quarks and one down quark (uud), bound together by the strong interaction. However, the situation is vastly more complex than this.

The proton is in reality highly complicated and inaccessible to the normal perturbative calculations of Quantum Field Theory - Fig. 1.0.1. Such objects can only be treated using probabilistic methods. When two protons collide we do not know which constituents, or *partons* are interacting, or what individual properties they have, such as their momentum and spin. We need some way of relating the known properties of the proton to the unknown properties of the partons. One way of doing this is using parton distribution functions (PDFs), which to first approximation give the probability of picking out a certain type of parton with certain properties.

Confinement of the quarks means experimental data is collected at the hadronic level, whereas theoretical predictions using Quantum Field Theory are made at the partonic level. The parton model provides a link between the two; in this framework partonic predictions are convolved with corresponding PDFs, summing over all possible partonic interactions. This produces PDF-dependent hadronic predictions. For useful theoretical predictions we therefore need as precise and accurate a handle on the PDFs as possible.

PDFs are unknowns in perturbative Quantum Chromodynamics (QCD), the theory of the strong interaction. Crucially, they are process independent. This means that they can be determined in a global fit between a wealth of experimental data and theoretical predictions. Once constrained, they can then be applied to any process. Any errors in PDF determination will propagate through to future predictions. There are three places these errors can be introduced:

1. Experimental errors
2. Theoretical errors
3. Errors from fitting procedure (methodological).

Until recently, experimental errors were the dominant source of error, meaning that theoretical errors have been largely ignored in standard PDF fits. However, with the onset of higher and higher precision experiments, we need to introduce a proper treatment of theoretical errors.

For comprehensive references dealing with Quantum Field Theory and the Standard Model the reader is referred to Refs. [**?**][**?**][**?**][**?**].

1.1 Parton Distribution Functions (PDFs) in Hadroproduction

In the Standard Model of particle physics, the strong force is responsible for short range interactions which bind together quarks. The theory of the strong force is known as QCD and corresponds to the Lagrangian

$$\mathcal{L}_{QCD} = -\frac{1}{4}F^A_{\alpha\beta}F^{\alpha\beta}_A + \sum_{flavours} \bar{q}_a(i\not{D} - m)_{ab}q_b \tag{1.1.1}$$

where the field strength tensor is given by

$$F^A_{\alpha\beta} = \left[\partial_\alpha \mathcal{A}^A_\beta - \partial_\beta \mathcal{A}^A_\alpha - g f^{ABC}\mathcal{A}^B_\alpha \mathcal{A}^C_\beta\right] \tag{1.1.2}$$

(ignoring gauge-fixing and ghost terms) [**?**]. Upper case Latin letters label *colour*, lower case Latin letters label *flavour* and lower case Greek letters run over components. The first term comes from self-interacting gluons, $\mathcal{A}$, and the second term comes from quarks, q, which obey the Dirac equation.

Hadroproduction at the LHC consists of processes like $pp \to X$ where p is a proton and X is some final hadronic state. This means experimental measurements are typically of some observable based on cross-section $\sigma_{pp\to X}$. Theoretical predictions are usually performed at the partonic level, for interactions between partons a and b, *i.e.* $\hat{\sigma}_{ab\to X}$. The convention is that hats apply to partonic variables and no-hats apply to hadronic variables. Factorisation theorems [32] allow the hadronic cross section to be expressed as a convolution of the partonic cross sections with relevant PDFs f:

$$\sigma_{pp\to X}(s, M_X^2) = \sum_{a,b}\int dx_1 dx_2 \; f_a(x, \mu_F^2)\; f_b(x_2, \mu_F^2)\; \hat{\sigma}_{ab\to X}(x_1 x_2 s, M_X^2, \mu_R^2, \mu_F^2) \tag{1.1.3}$$

where x is the momentum fraction of the proton associated with each parton, s

is the proton centre-of-mass frame energy and M_X is the invariant mass of the final state X.

PDFs depend on the scale μ_F at which the factorisation is applied. However, this is an artificially introduced scale so observables such as $\sigma_{pp\to X}$ must be independent of it. This observation leads to a series of coupled partial differential "renormalisation group" equations relating PDFs at different scales known as the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) equations [20][28][31]:

$$\mu_F \frac{d}{d\mu_F} f_i(y, \mu_F^2) = \sum_j \int_z^1 \frac{dz}{z} \mathcal{P}_{ij}\left(\frac{y}{z}, \alpha_s\right) f_j(z, \mu_F^2). \tag{1.1.4}$$

These show that the scale dependence of a given parton's PDF depends on all the other partons' PDFs through "splitting functions" $\mathcal{P}_{ij}$, which in turn depend on the strong coupling constant α_s.

Although PDFs may seem at first sight to be totally unknown there are some theoretical observations which we can use to constrain their form. These are known as the "sum rules" [**?**]. One thing which intuitively makes sense is that if you add up all the momenta of the partons you end up with the momentum of the proton. This enforces the condition

$$\int_0^1 dx \sum_i x f_i(x, Q^2) = 1. \tag{1.1.5}$$

The other thing we know about the proton is that it is made up of two up and one down (and no strange) "valence" quarks. So we can normalise the PDFs using the expressions

$$\int_0^1 dx \big(f_u - f_{\bar{u}}\big) = 2 \tag{1.1.6a}$$

$$\int_0^1 dx \big(f_d - f_{\bar{d}}\big) = 1 \tag{1.1.6b}$$

$$\int_0^1 dx \big(f_s - f_{\bar{s}}\big) = 0. \tag{1.1.6c}$$

Note that these conditions require that the PDFs are integrable.

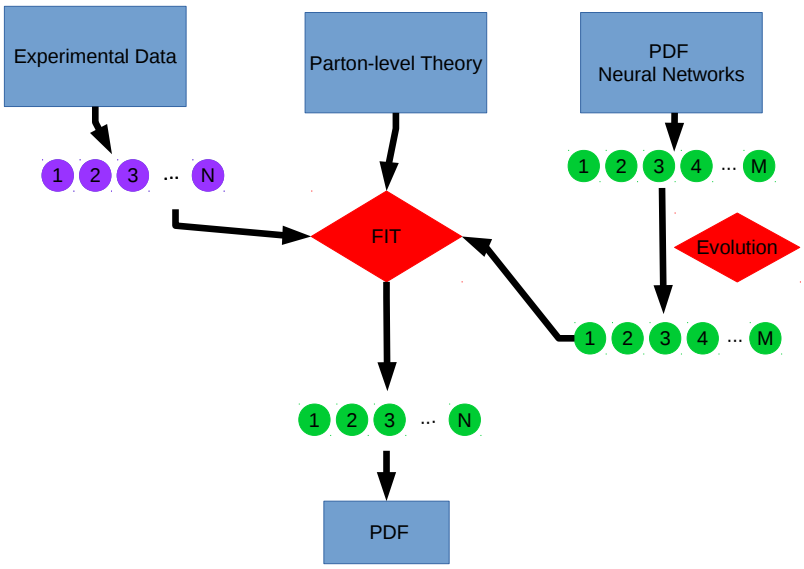

Figure 1.1.1 *NNPDF general strategy.*

1.1.1 PDF Fitting Strategy

There are a number of groups currently active in carrying out PDF fits and these include

- MSTW [5]
- CTEQ [17]
- NNPDF [15]
- HERAPDF/xFitter [27]
- ABM [34].

This section focusses on the techniques adopted by NNPDF. There are two main features of the NNPDF strategy which differ from other fitting collaborations' [18]. These are:

1. Use of Monte Carlo rather than Hessian approach to error analysis.
2. Fitting using artificial neural networks.

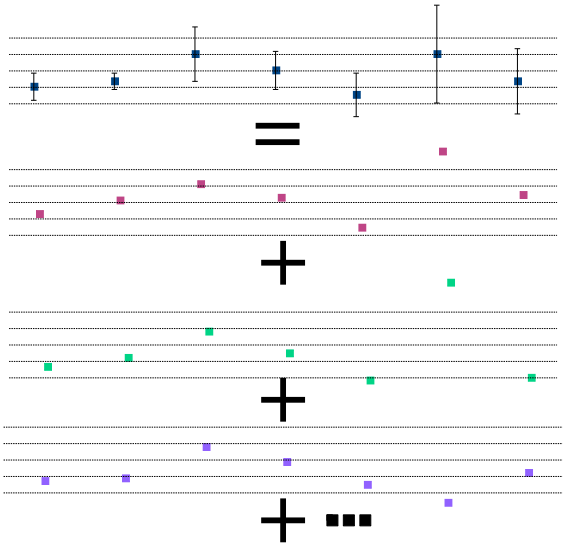

Figure 1.1.2 *Generation of Monte Carlo replicas of pseudodata using a Gaussian probability density function which has the same mean and variance as the experimental data. This means that the mean and variance of the replicas can reproduce those of the experimental data arbitrarily closely, given enough replicas.*

Fig. 1.1.1 outlines the NNPDF general strategy and Fig. **??** shows the fitting code structure. Experimental data is converted into an ensemble of N artificial Monte Carlo replicas or *pseudodata*. These are randomly generated in accordance with multi-Gaussian distributions centred around each data point, with variance given by the experimental uncertainty (see Fig. 1.1.2). Each Monte Carlo replica contains the same number of data points as the original experimental measurements. Given enough replicas, the Monte Carlo set contains complete experimental information; the experimental central value can be retrieved by taking the mean, and the experimental variance is the variance calculated over the replicas.

NNPDF uses a variety of experimental data from a number of particle colliders. These are observables such as cross sections, differential cross sections and structure functions. Fig. 1.1.3 is a plot of the (x, Q^2) range spanned by the datasets in the latest NNPDF3.1 [15] release. Here x and Q are Björken variables corresponding to the parton momentum fraction and the energy scale of the process. The majority of the data are from DIS.

Theoretical predictions of parton-level observables are computed using external codes such as `MCFM` [26], `DYNNLO` [16], `FEWZ` [30] and `NLOjet++` [35]. These are converted to higher orders as necessary using QCD and electroweak correction

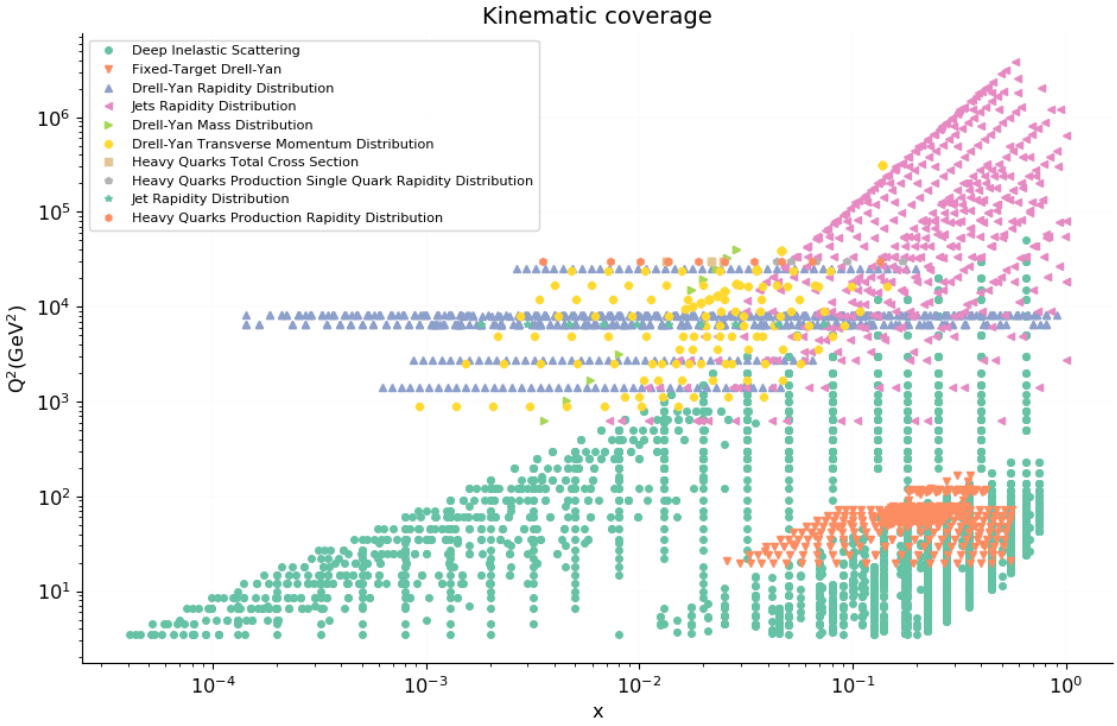

Figure 1.1.3 *Plot of the (x, Q^2) range spanned by data included in the latest NNPDF3.1 NLO fit.*

("k") factors. They are then combined with DGLAP evolution kernels, which evolve PDFs from an initial reference energy scale to the energy scale of each experiment using the DGLAP equations (Eqn. 1.1.4).

Next, M independent artificial neural networks are generated. These are combined with the partonic theory and compared to the experiments in a global fit. Here each pseudodata set is fitted against all the neural networks, which learn the functional form of the PDF. The best fit is determined based on some suitable figure of merit such as χ^2 value between the experimental observables and the theoretical observables calculated using the current iteration of PDFs.

In order to prevent overlearning, where the neural network also fits random fluctuations in the data, the pseudodata are divided into two sets, one for training and one for validation. At each step in the fit the χ^2 is computed for both sets, but minimisation is based only on the training set. When the χ^2 of the validation set stops decreasing, this signifies that overlearning has begun so any subsequent versions of the PDF can be discarded.

The whole fitting process produces N "best fit" neural networks, which act as a Monte Carlo parametrisation of the PDF (for example Fig. 1.1.4). This means

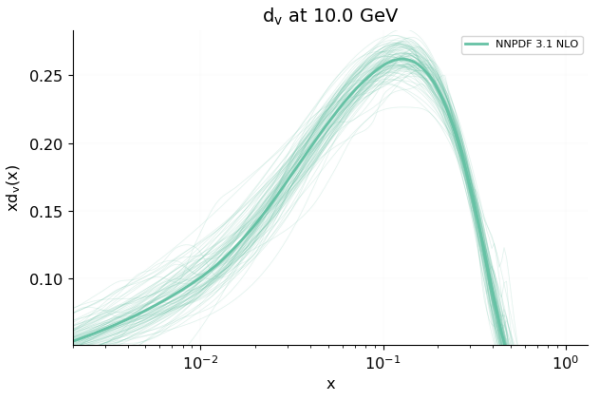

Figure 1.1.4 *The Monte Carlo replicas for the down valence quark PDF NNPDF3.1 at NLO. The scale is $Q = 10$ GeV*

that the PDF and its error can be extracted by taking the mean and standard deviation. The final PDFs are made publicly available on the LHAPDF [4] website (`https://lhapdf.hepforge.org/`).

1.2 Parametrisation and Neural Networks

PDFs at a given energy scale can be evolved to any other scale using the DGLAP equations (Eqn. 1.1.4). In order to aid the fitting process, the PDFs undergo *preprocessing*; they are parametrised as:

$$f_i(x) = A_i x^{-\alpha_i}(1-x)^{\beta_i} N_i(x) \tag{1.2.1}$$

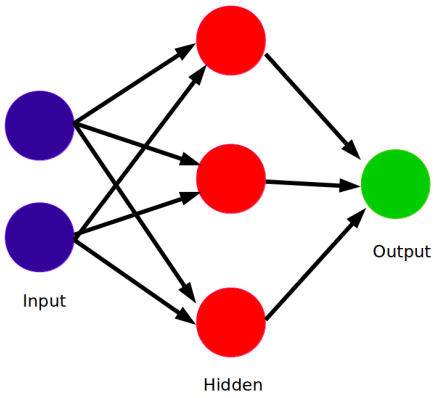

Figure 1.2.1 *Schematic depiction of an artificial neural network.*

where A_i are coefficients set by the sum rules (see earlier) and α_i and β_i are parameters to be fitted. $N_i(x)$ are artificial neural networks. This form is chosen to force the PDFs to 0 at large x.

Inspired by how the brain processes information, neural networks are composed of a collection of nodes known as *neurones*, connected together in various ways. They are trained by example, so have the capability to learn a PDF's functional form given a

set of data. The layout consists of input layers, hidden layers and output layers, as shown in Fig. 1.2.1. In a feed-forward neural network, information can only be passed in one direction through the layers (from input to output). Here, the output of a neurone in the l^{th} layer is given by

$$\xi_i^{(l)} = g\left(\sum_j^{inputs} \omega_{ij}^{(l)} \xi_j^{(l-1)} + \theta_i^{(l)} \right) \tag{1.2.2}$$

where the ωs and θs are "weights" and "thresholds"; parameters to be minimised with respect to. g is an "activation function" which is set to

$$g(a) = \begin{cases} \frac{1}{1+e^{-a}} & \text{for hidden layers} \\ a & \text{for output layer.} \end{cases} \tag{1.2.3}$$

1.3 Experimental Uncertainties

Experimental uncertainties are implemented via a *covaraince matrix* which links vectors of data points. This encapsulates the total breakdown of errors, ρ, and can be constructed:

$$\sigma_{ij} = \delta_{ij}\rho_i^{uncorr}\rho_j^{uncorr} + \sum_a^+ \rho_{i,a}\rho_{j,a} + \left(\sum_m^\times \rho_{i,m}\rho_{j,m} \right) T_i^0 T_j^0. \tag{1.3.1}$$

The notation $\sum^+$ indicates a sum over additive systematics and $\sum^\times$ over multiplicative systematics (more on this below).

Here the uncorrelated statistical uncertainties appear down the diagonal, but correlated systematic uncertainties can also appear on the off-diagonals. Correlated uncertaintes include those which link multiple data points, for example systematic uncertainties from a particular detector which will affect all of its data in a similar way.

Systematic uncertainties further divide into two types, "additive" and "multiplicative". Additive systematics are perhaps a more familiar type of error, and are independent of the datapoint values themselves. On the other hand, multiplicative systematics depend on the measured values. In the context of particle physics experiments, a common example is total detector luminosity. This is because recorded cross sections are dependent on the luminosity of the detector;

a higher luminosity means more collisions will take place so the measured cross section will be greater.

The covariance matrix is used in two places during the fit; the generation of pseudodata and calculation of χ^2.

Chapter 2

Theory uncertainties in PDFs - 10%

- What are theory uncertainties?
- Why are they now important?
- Types of unc - see below
- Bayesian interpretation of these uncertainties - there is one "true" value e.g. for the higher order value, so need to estimate uncertainty bc will never know the size of e.g. MHOU unless you go ahead and calc - maybe ref d'Agostini paper on Bayesian interpretation
- Will use Bayesian framework and assume Gaussianity of the expected true value of theory calc
- Show C+S in fit - plus sign because exp and th unc are independent so combine errors in quadrature. They are also on an equal footing in terms of their effect on the PDFs
- When many datasets/global fit, can have v strong theory correlations even across different experiments, because the underlying theory connects them

2.1 Fitting PDFs including theory uncertainties

Historically, experimental uncertainties have been the dominant source of error in PDF fits. In the NNPDF framework both replica generation and computation of

χ^2 are currently based entirely on these. We must now try to match the ongoing drive to increase experimental precision by including errors introduced at the theoretical level. This is especially important given recent data sets such as the Z boson transverse momentum distributions [8] [19] [29], which have very high experimental precision. Without the inclusion of theoretical errors, this has led to tension with the other datasets.

In future NNPDF fits theoretical uncertainties will be included following a procedure outlined by Ball & Desphande [22]. This hinges on a result from Bayesian statistics which applies to Gaussian errors. Namely, theory uncertainties can be included by directly adding a theoretical covariance matrix to the experimental covariance matrix prior to the fitting. A brief summary of the derivation is given below.

When determining PDFs we incorporate information from experiments in the form of N_{dat} experimental data points D_i, $i = 1, ..., N_{dat}$. The associated uncertainties and their correlations are encapsulated in an experimental covariance matrix C_{ij}. Parts of the matrix which associate two independent experiments will be populated by zeros. However we would expect there to be correlations between data points from the same detector, for example.

Each data point is a measurement of some fundamental "true" value, $\mathcal{T}_i$, dictated by the underlying physics. In order to make use of the data in a Bayesian framework, we assume that the experimental values follow a Gaussian distribution about the unknown $\mathcal{T}$. Then, assuming the same prior for D and $\mathcal{T}$, we can write an expression for the conditional probability of $\mathcal{T}$ given the known data D:

$$P(\mathcal{T}|D) = P(D|\mathcal{T}) \propto \exp\left(-\frac{1}{2}(\mathcal{T}_i - D_i)C_{ij}^{-1}(\mathcal{T}_j - D_j)\right). \tag{2.1.1}$$

However, in a PDF fit we cannot fit to the unknown true values $\mathcal{T}$, and must make do with predictions based on current theory T_i. This is the origin of theory uncertainties in PDF fits; where our theory is incomplete, fails to describe the physics well enough, or where approximations are made, we will introduce all kinds of subtle biases into the PDF fit. The theory predictions themselves also depend on PDFs, so uncertainties already present in the PDFs are propagated through. This, in particular, leads to a high level of correlation because the PDFs are universal, and shared between all the theory predictions.

We can take a similar approach when writing an expression for the conditional

probability of the true values $\mathcal{T}$ given the available theory predictions T, by assuming that the true values are Gaussianly distributed about the theory predictions.

$$P(\mathcal{T}|T) = P(T|\mathcal{T}) \propto \exp\left(-\frac{1}{2}(\mathcal{T}_i - T_i)S_{ij}^{-1}(\mathcal{T}_j - T_j)\right), \tag{2.1.2}$$

where S_{ij} is a "theory covariance matrix" encapsulating the magnitude and correlation of the various theory errors. We will need to do some work to determine S_{ij} for the different sources of error, and this will be outlined in detail in the following chapters.

When we fit PDFs we aim to maximise the probability that a PDF-dependent theory is true given the experimental data available. This amounts to maximising $P(T|D)$, marginalised over the unknown true values $\mathcal{T}$. To make this more useful for fitting purposes, we can relate this to $P(D|T)$ using Bayes' Theorem:

$$P(D|T)P(\mathcal{T}|DT) = P(\mathcal{T}|T)P(D|\mathcal{T}T), \tag{2.1.3}$$

where we note that the experimental data D do not depend on our modelled values T, so $P(D|\mathcal{T}T) = P(D|\mathcal{T})$. So we can integrate Bayes' Theorem over the possible values of the N-dimensional true values $\mathcal{T}$:

$$\int D^N\mathcal{T}P(D|T)P(\mathcal{T}|DT) = \int D^N\mathcal{T}P(\mathcal{T}|T)P(D|\mathcal{T}), \tag{2.1.4}$$

and, because $\int D^N\mathcal{T}P(\mathcal{T}|TD) = 1$ as all possible probabilities for the true values must sum to one,

$$P(D|T) = \int D^N\mathcal{T}P(\mathcal{T}|T)P(D|\mathcal{T}). \tag{2.1.5}$$

We can always write the theory predictions T in terms of their shifts Δ relative the true values $\mathcal{T}$:

$$\Delta_i \equiv \mathcal{T}_i - T_i. \tag{2.1.6}$$

Use the expression for conditional probability $P(X \cap Y) = P(X|Y)P(Y)$ and

integrate over all possible values of the true theory T (as this is an unknown):

$$\begin{aligned} \int dT\ P(T|y\cap f)P(y|f) &= \int dT\ P(y|T\cap f)P(T|f) \\ P(y|f) &= \int dT\ P(y|T\cap f)P(T|f)\,. \end{aligned} \tag{2.1.7}$$

Now assume Gaussian uncertainties for data and theory, of the form $\exp(-\frac{1}{2}\chi^2)$

$$P(y|Tf) \propto \exp\bigg(-\frac{1}{2}(y-T)^T\sigma^{-1}(y-T)\bigg) \tag{2.1.8}$$

$$P(T|f) \propto \exp\bigg(-\frac{1}{2}(T-T[f])^T s^{-1}(T-T[f])\bigg) \tag{2.1.9}$$

and substitute these into Eq. 2.1.7 to get

$$P(y|f) \propto \int dT\ \exp\bigg(-\frac{1}{2}\bigg[(y-T)^T\sigma^{-1}(y-T)+(T-T[f])^T s^{-1}(T-T[f])\bigg]\bigg)\,. \tag{2.1.10}$$

Note that the difference between the full theory T and the theory predictions $T[f]$ defines the total correction, C. Therefore the substitution $T = T[f] + C \Rightarrow dT = dT[f] + dC$ can be made, noting that $dT[f] = 0$ because $T[f]$ is the fixed output of NNPDF analysis. The overall expression then becomes

$$P(y|f) \propto \int dC\ \exp\bigg(-\frac{1}{2}\bigg[(y-T[f]-C)^T\sigma^{-1}(y-T[f]-C)+C^T s^{-1}C\bigg]\bigg) \tag{2.1.11}$$

which can be evaluated by Gaussian integration over shifted variables, leading to

$$P(y|f) \propto \exp\bigg(-\frac{1}{2}(y-T[f])^T(\sigma+s)^{-1}(y-T[f])\bigg)\,. \tag{2.1.12}$$

The final result is that you can treat theoretical errors in exactly the same was as you treat experimental errors.

2.2 Sources of Theoretical Uncertainties

The next step is to estimate the theory covariance matrix, s. This can include a number of different theoretical uncertainties that may appear in PDF fits, such as:

1. **Statistical uncertainties** such as from Monte Carlo generators. These provide diagonal entries to s.

2. **Systematic uncertainties**. These are trickier, and can be estimated by varying some fit parameter ξ from its value at the central prediction, ξ_0, and applying

$$s_{ij} = \langle (T[f;\xi] - T[f;\xi_0])_i (T[f;\xi] - T[f;\xi_0])_j \rangle \tag{2.2.1}$$

   where the angled brackets denote the averaging over a given range of ξ according to some prescription.

Including systematic uncertainties will pose the biggest challenge. We need to identify the places they are being introduced and then make a suitable choice of ξ. Examples of systematic uncertainties we have begun to address are:

- **Missing higher order uncertainties (MHOUs)**. These are a result of calculations being done only up to a certain perturbative order in the expansion of α_s. As discussed in more detail below, we can get a handle on these by varying the artificial renormalisation (μ_R) and factorisation (μ_F) scales introduced in the calculation. Here ξ can be thought of as a vector (μ_R, μ_F).

- **Nuclear and deuteron corrections**. Here data are taken from nuclear targets but the theoretical treatment does not account for this. We can re-calculate the observables under different nuclear models or parametrisations. In this case ξ indexes the model or parametrisation used. The use of multiple models helps to remove any systematic bias introduced by each individual one.

2.3 Renormalisation group invariance

2.4 Scale variation in partonic cross-sections

2.5 Scale variation in PDF evolution

2.6 Double scale variations

2.7 Multiple scale variations

Chapter 3

Missing higher order uncertainties 0%

- MHOUs dominate theoretical uncertainties in LHC
- To estimate MHOU, standard is scale var of μ_r amd μ_f - refs for cons/alternatives e.g. Cacciari-Houdeau, but this is widely applicable to all procs and builds in the correlations e.g. between similar kinematic regions
- Could use another method in the Bayesian framework though, using this for MHOUs
- MHOUs not yet included in PDF fits
- MHOUs used to be small, esp since NNLO PDFs emerged
- In 3.1 electroweak scale, unc down to 1% level, and QCD MHOU is at % level
- MHOUs can increase/decrease weights of experiments in the fit
- Here we formulate inclusion, include at NLO, verify against NNLO (including developing verification methods/tools) and assess impact on pheno

3.1 Prescriptions to generate the theory covariance matrix

3.2 Validating the theory covariance matrix

3.3 PDFs with missing higher order uncertainties

3.4 Impact on phenomenology

3.5 Usage and delivery

Chapter 4

Nuclear Uncertainties - 90%

4.1 Introduction

Parton distribution functions (PDFs) are universal quantities encapsulating the internal structure of the proton, and are crucial for making predictions in particle physics [25]. To maximally constrain them, PDFs are determined by fitting a range of experimental data over a wide variety of processes and kinematic regimes. Some of this data consists of measurements on nuclear targets, rather than proton targets. In this case, the surrounding nuclear environment will have an effect on the measured observables, which in turn will influence the form of the fitted PDFs. The uncertainties associated with these effects are termed "nuclear uncertainties". Such uncertainties are small [14][23] but becoming increasingly relevant with the advent of the Large Hadron Collider and the era of precision physics it has ushered in [10].

In these proceedings, we show how to use existing nuclear PDFs (nPDFs) to provide an estimate of nuclear uncertainties, and include them in future proton PDF fits within the Neural Network PDF (NNPDF) framework [24] [1]. We first review the nuclear data (Sec. 4.2), then outline the construction and form of nuclear uncertainties (Sec. 4.3). Finally, we assess the impact on the PDFs and associated phenomenology (Secs. 4.4 and 4.5).

[1] For a more detailed analysis, see [**?**].

4.2 Nuclear Data

There are three experiments with nuclear targets currently included in NNPDF analyses: charged current inclusive deep inelastic scattering (DIS) cross sections from CHORUS [9], on Pb; DIS dimuon cross sections from NuTeV [13][36] on Fe; and Drell-Yan dimuon cross sections from E605 at Fermilab [3], on Cu. After cuts, nuclear data make up 993/4285 of the data points ($\sim 23\%$). For a complete summary of the data sets, see [15].

A study of the correlation between these measurements and the fitted PDFs reveals that the CHORUS data has most impact on the up- and down-valence distributions, NuTeV data has most impact on the strange, and E605 data has most impact on the other light sea quarks: anti-up and anti-down. Therefore, we anticipate largest effects from nuclear uncertainties in these PDFs.

4.3 Determining Nuclear Uncertainties

In a PDF fit we include an experimental covariance matrix, C_{ij}, describing the breakdown of statistical and systematic errors, where i, j run over the data points. Uncertainties due to nuclear data must be considered in addition to the experimental uncertainties, and in general they can be encapsulated in a theoretical covariance matrix, S_{ij}. In a PDF fit we simply add this to C_{ij} [22], so that the nuclear uncertainties act like experimental systematics.

We adopted an empirical approach to construct the nuclear uncertainties, using nPDFs rather than appealing to nuclear models, which rely on various assumptions [21]. We compared theoretical predictions for nuclear observables made with the correct corresponding nPDFs for an isotope "N", $T_i^N[f_N^{(n)}]$, to those with proton PDFs, $T_i^N[f_p]$. Here f_p is the central value for a proton PDF and $f_N^{(n)}$ is one Monte Carlo replica in an nPDF ensemble, where $n = 1, ..., N_{rep}$ [25]. To generate such an ensemble we combined three recent nPDF sets: DSSZ12 [6], nCTEQ15 [12] and EPPS16 [11]. Note that DSSZ12 does not provide a Cu PDF, so for the case of E605 we combined just two nPDF sets.

We considered two definitions of nuclear uncertainties:

1. **Def. 1,** (a conservative approach) where the only modification is to include

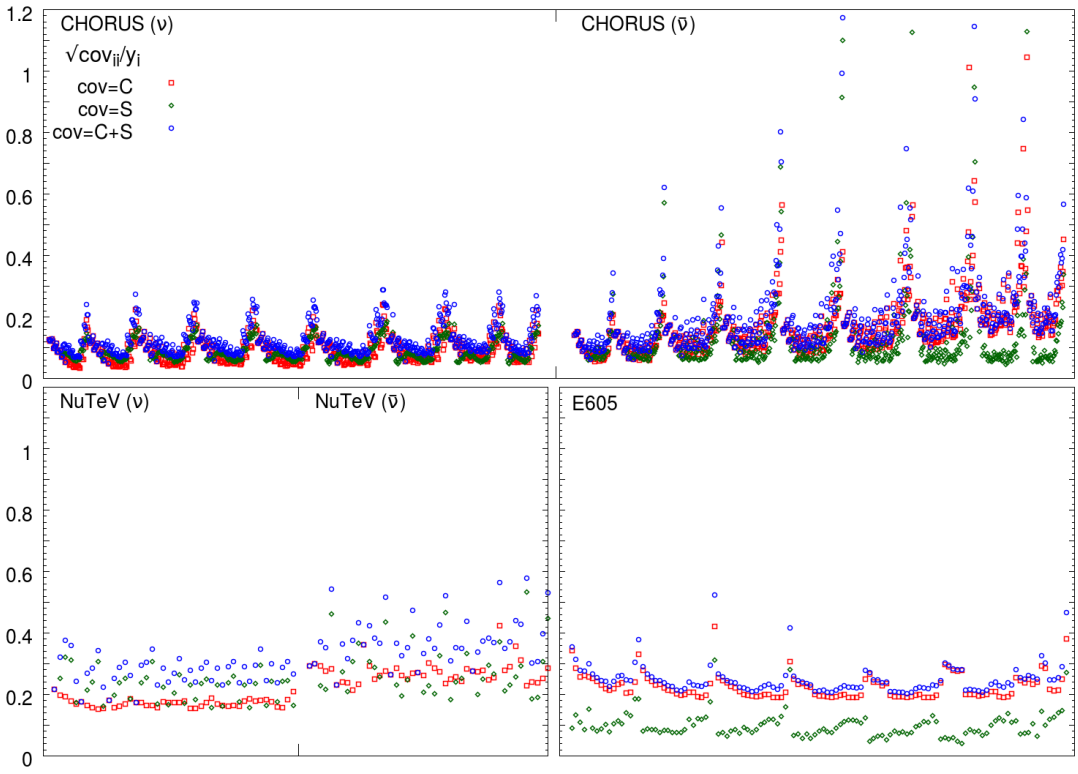

Figure 4.3.1 *The square root of the diagonal elements of the covariance matrices, normalised to corresponding data. Experimental contributions are red, theory green and the total blue. Data from CHORUS and NuTeV are split into neutrino and anti-neutrino parts. Points are binned in (anti-)neutrino beam energy E: 25, 35, 45, 55, 70, 90, 110, 120, 170 GeV. In each bin x increases from left to right, 0.045 < x < 0.65.*

nuclear uncertianties, with

$$\Delta_i^{(n)} = T_i^N[f_N^{(n)}] - T_i^N[f_p]; \tag{4.3.1}$$

2. **Def. 2,** (a more ambitious approach) where a shift,

$$\delta T_i^N = T_i^N[f_N] - T_i^N[f_p], \tag{4.3.2}$$

is also applied to the corresponding observable, meaning that the uncertainty should be defined relative to the shifted value,

$$\Delta_i^{(n)} = T_i^N[f_N^{(n)}] - T_i^N[f_N]. \tag{4.3.3}$$

Whilst Def. 1 just deweights the nuclear data sets in a PDF fit, Def. 2 also attempts to directly apply a nuclear correction. In both cases we can construct a theoretical covariance matrix as

$$S_{ij} = \frac{1}{N_{rep}} \sum_{n=1}^{N_{rep}} \Delta_i^{(n)} \Delta_j^{(n)}. \tag{4.3.4}$$

We did this separately for each experiment, which is a conservative treatment.

Considering the diagonal elements of the covariance matrices (Fig. 4.4.2), we see that the nuclear uncertainty has the largest impact on the NuTeV data, where the nuclear uncertainties dominate the data uncertainties. This is mirrored in the off-diagonal elements (Fig. 4.4.1). Given the high correlation of NuTeV observables with the s and $\bar{s}$ PDFs, the effect of including the uncertainties ought to be greatest for these PDFs.

4.4 The Impact on Global PDFs

We compared four different PDF fits:

- **Baseline,** based on NNPDF3.1, with small improvements [22];
- **NoNuc,** Baseline with nuclear data removed;
- **NucUnc,** Baseline with nuclear uncertainties according to Def. 1.
- **NucCor,** Baseline with nuclear uncertainties and a nuclear correction according to Def. 2.

Table 4.4.1 shows the variation in χ^2 for selected data sets [2]. All of the fits show reduced χ^2 compared to Baseline, highlighting tension due to nuclear data. However, the strange-sensitive ATLAS W/Z at 7 TeV (2011) measurements [2] still have a poor χ^2, indicating that possible tensions with NuTeV were unlikely responsible for this; in any case, the data sets occupy different kinematic regions. The best fit is obtained for NucUnc, which has the largest uncertainties.

Fig. 4.5.1 shows the light sea quark PDFs for NucUnc compared to Baseline. These are the distributions with greatest impact, but there is little appreciable

[2]For a full break-down see [22].

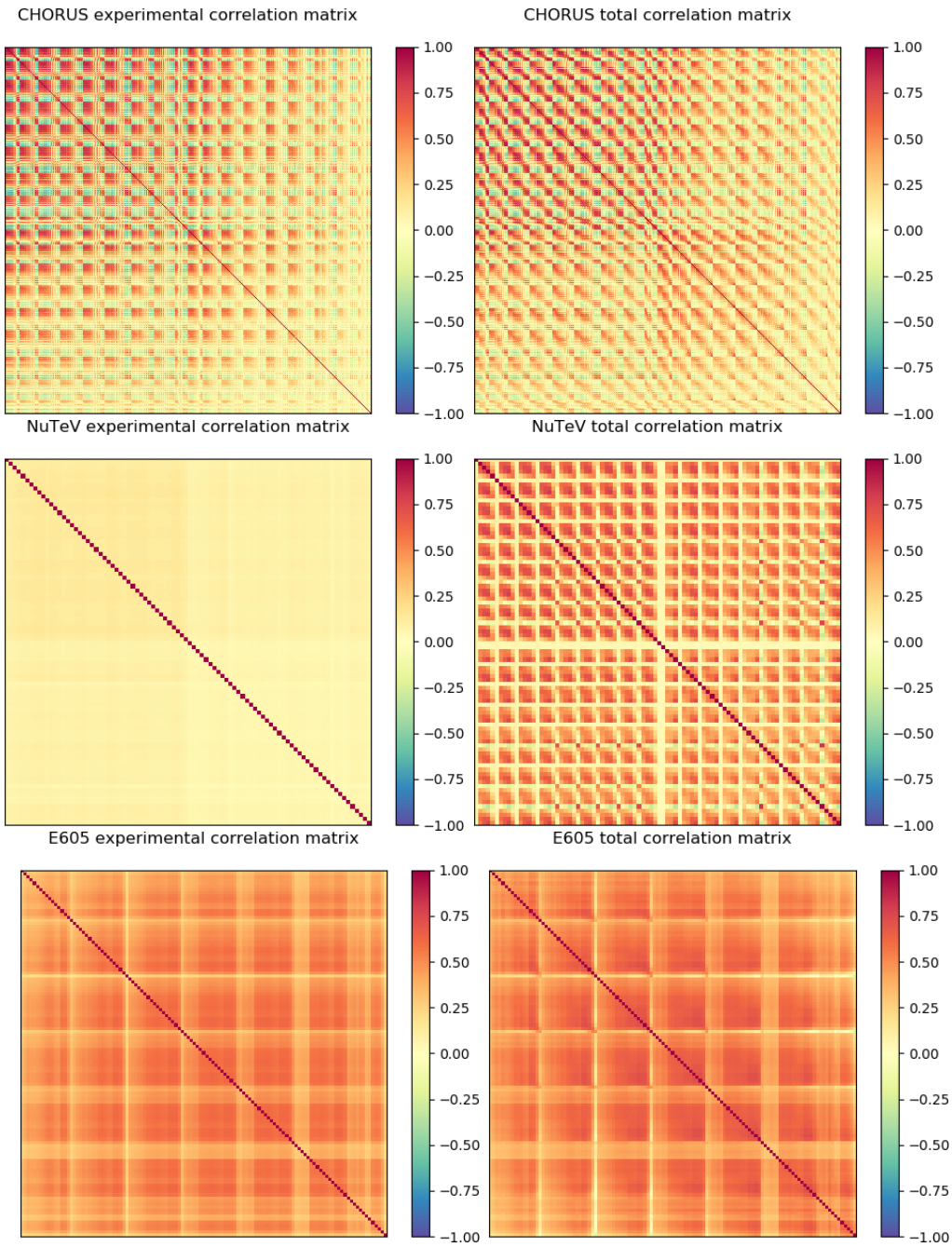

Figure 4.4.1 *Correlation matrices, $\rho_{ij}^{cov} = \frac{cov_{ij}}{\sqrt{cov_{ii}cov_{jj}}}$, before (left) and after (right) including nuclear uncertainties. Data are binned the same as in Fig. 4.4.2. The top row corresponds to CHORUS, middle row to NuTeV and bottom row to E605. Results are displayed for Def. 1 but are qualitatively similar for Def. 2.*

Experiment	N_{dat}	Baseline	NoNuc	NucUnc	NucCor
CHORUS ν	416	1.29	–	0.97	1.04
CHORUS $\bar{\nu}$	416	1.20	–	0.78	0.83
NuTeV ν	39	0.41	–	0.31	0.40
NuTeV $\bar{\nu}$	37	0.90	–	0.62	0.83
E605 σ^p	85	1.18	–	0.85	0.89
ATLAS W/Z (2011)	34	1.97	1.78	1.87	1.94
ATLAS	360	1.08	1.04	1.04	1.05
CMS	409	1.07	1.07	1.07	1.07
LHCb	85	1.46	1.27	1.32	1.37
Total	4285	1.18	1.14	1.07	1.09

Table 4.4.1 *χ^2 per data point for selected data sets. The final row shows results for the full fitted data.*

change other than a small shift in the central value and increase in uncertainties. NucCor behaves similarly. Overall, the nuclear uncertainties are small compared to the global experimental uncertainty.

4.5 Phenomenology

Given the changes to the light sea quark PDFs, it is interesting to examine the impact on relevant phenomenological quantities, namely: the sea quark asymmetry, $\bar{u}/\bar{d}$; strangeness fraction, $R_s = (s+\bar{s})/(\bar{u}+\bar{d})$; and strange valence distribution, $xs^- = x(s-\bar{s})$ (Fig. 4.6.1). In all cases it is clear that removing the nuclear data has a significant effect, emphasising the need to retain this data in proton PDF fits. Adding nuclear uncertainties, however, makes very little difference. In particular, the known tension between ATLAS W/Z + HERA DIS data and NuTeV data, which is apparent in the strangeness fraction [7], is not relieved with the addition of nuclear uncertainties.

We found no appreciable difference between using NucUnc versus NucCor, so opt to incorporate uncertainties using NucUnc (Def. 1) as this is the more conservative option.

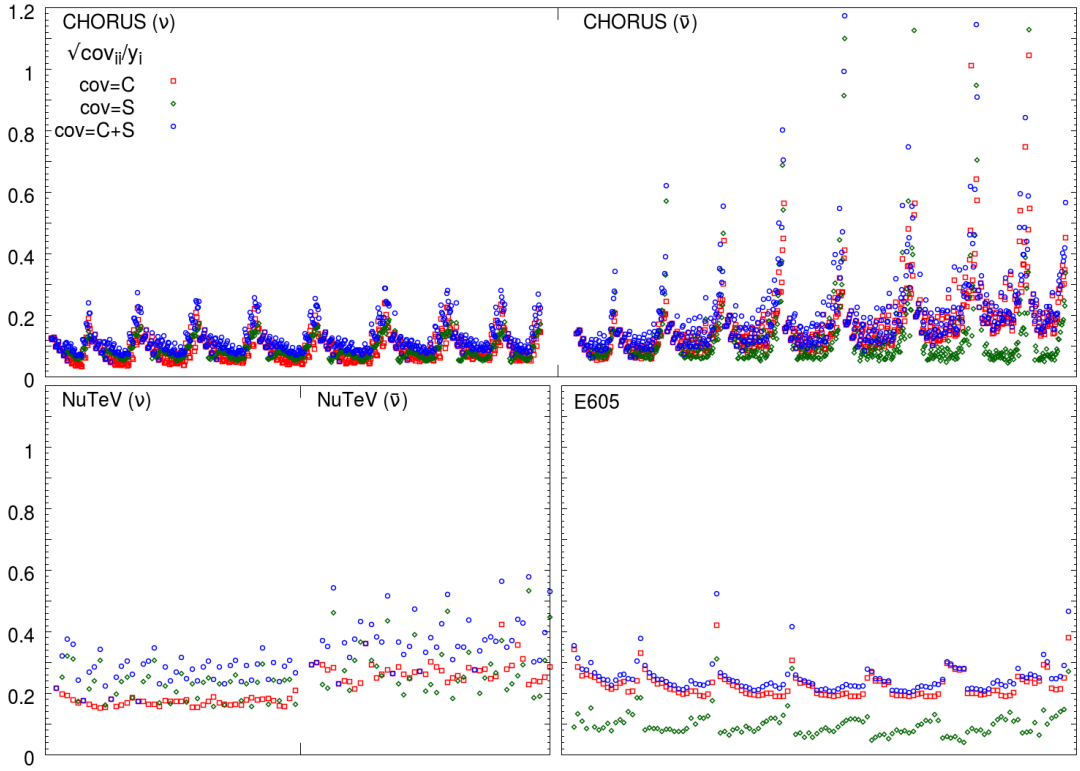

Figure 4.4.2 *The square root of the diagonal elements of the covariance matrices, normalised to corresponding data. Experimental contributions are red, theory green and the total blue. Data from CHORUS and NuTeV are split into neutrino and anti-neutrino parts. Points are binned in (anti-)neutrino beam energy E: 25, 35, 45, 55, 70, 90, 110, 120, 170 GeV. In each bin x increases from left to right, $0.045 < x < 0.65$.*

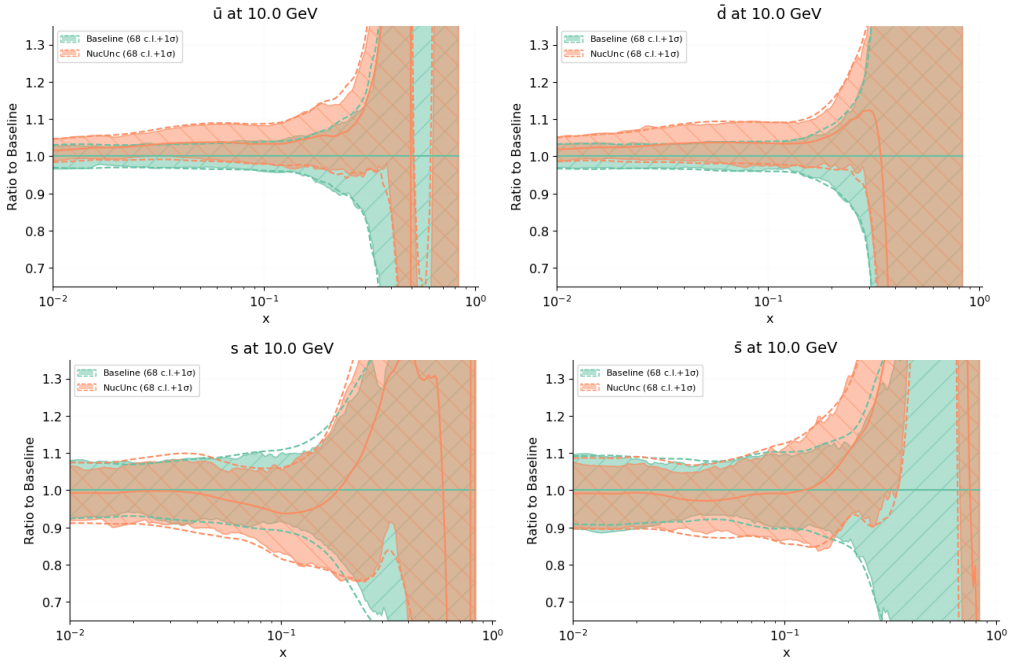

Figure 4.5.1 *NucUnc fits with nuclear uncertainties (orange) compared to Baseline (green) for PDFs at 10 GeV. Clockwise from top left: $\bar{u}$, $\bar{d}$, s and $\bar{s}$ PDFs. Error bands are 1 σ; results are normalised to Baseline fit.*

4.6 Conclusions

We studied the role of nuclear data in proton PDF fits, and adopted an empirical approach to determine the nuclear uncertainties due to this data. We based our analysis on recent nPDF fits DSSZ12, nCTEQ15 and EPPS16. Using a theoretical covariance matrix, we included these uncertainties in proton PDF fits, and found that the fit quality was improved, with the largest effect on the light sea quark distributions [3]. We found no significant impact on associated phenomenology.

We will extend this analysis to deterium data, and in the future we will be able to use nuclear PDFs from NNPDF [33] to estimate uncertainties. Furthermore, these methods can be applied to other sources of theoretical uncertainties, such as higher twist effects, fragmentation functions, and missing higher order uncertainties [**?**].

[3]The PDF sets from this analysis are available upon request from the authors in LHAPDF format [4].

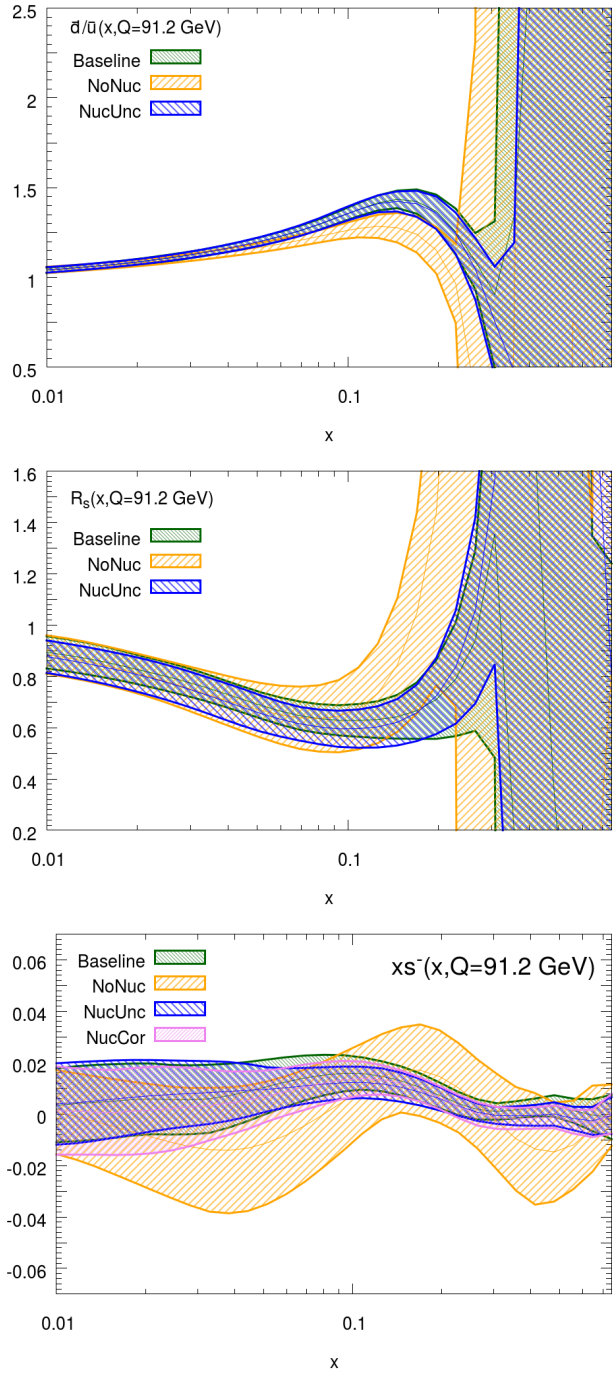

Figure 4.6.1 *Effect of including nuclear uncertainties on phenomenology. From left to right: sea quark asymmetry, strangeness fraction, strange valence distribution. Distributions correspond to the use of different PDF fits: Baseline (green), NoNuc (yellow), NucUnc (blue) and NucCor(pink). $Q = 91.2$ GeV. In the left two plots, NucCor are indistinguishable from NucUnc so are omitted for readability.*

Chapter 5

Deuteron Uncertainties - 0%

Chapter 6

Higher Twist Uncertainties - 0%

6.1 The role of higher twist data in PDFs

6.2 Ansatz for a higher twist correction

6.3 Using a neural network to model higher twist

6.3.1 Model architecture

6.3.2 Training and validating the neural network

6.4 Form of the higher twist correction

6.5 The higher twist covariance matrix

6.6 PDFs with higher twist uncertainties

Chapter 7

Conclusion - 0%

Bibliography

[1] https://www.desy.de/news/news_search/index_eng.html?openDirectAnchor=829.

[2] Aad, G., et al. "Measurement of the inclusive $W^{\pm}$ and Z/gamma cross sections in the electron and muon decay channels in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector." *Phys. Rev.* D85: (2012) 072,004.

[3] Aaltonen, T., et al. "Measurement of the Inclusive Jet Cross Section at the Fermilab Tevatron p anti-p Collider Using a Cone-Based Jet Algorithm." *Phys. Rev.* D78: (2008) 052,006. [Erratum: Phys. Rev.D79,119902(2009)].

[4] et al., A. B. "LHAPDF6: parton density access in the LHC precision era." *Eur. Phys. J.* C75: (2015) 132.

[5] et al., A. D. M. "Uncertainties on alpha(S) in global PDF analyses and implications for predicted hadronic cross sections." *Eur. Phys. J.* C64: (2009) 653–680.

[6] de Florian et al., D. "Global Analysis of Nuclear Parton Distributions." *Phys. Rev.* D85: (2012) 074,028.

[7] et al., G. A. "Determination of the strange quark density of the proton from ATLAS measurements of the $W \to \ell\nu$ and $Z \to \ell\ell$ cross sections." *Phys. Rev. Lett.* 109: (2012) 012,001.

[8] ———. "Measurement of the Z/γ^* boson transverse momentum distribution in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector." *JHEP* 09: (2014) 145.

[9] et al., G. O. "Measurement of nucleon structure functions in neutrino scattering." *Phys. Lett.* B632: (2006) 65–75.

[10] et al., J. G. "The Structure of the Proton in the LHC Precision Era." .

[11] et al., K. J. E. "EPPS16: Nuclear parton distributions with LHC data." *Eur. Phys. J.* C77, 3: (2017) 163.

[12] et al., K. K. "nCTEQ15 - Global analysis of nuclear parton distributions with uncertainties in the CTEQ framework." *Phys. Rev.* D93, 8: (2016) 085,037.

[13] et al., M. G. "Precise measurement of dimuon production cross-sections in muon neutrino Fe and muon anti-neutrino Fe deep inelastic scattering at the Tevatron." *Phys. Rev.* D64: (2001) 112,006.

[14] et al., R. D. B. "Precision determination of electroweak parameters and the strange content of the proton from neutrino deep-inelastic scattering." *Nucl. Phys.* B823: (2009) 195–233.

[15] ———. "Parton distributions from high-precision collider data." *Eur. Phys. J.* C77, 10: (2017) 663.

[16] et al., S. C. "Vector boson production at hadron colliders: a fully exclusive QCD calculation at NNLO." *Phys. Rev. Lett.* 103: (2009) 082,001.

[17] et al., S. D. "New parton distribution functions from a global analysis of quantum chromodynamics." *Phys. Rev.* D93, 3: (2016) 033,006.

[18] et al., S. F. "Neural network parametrization of deep inelastic structure functions." *JHEP* 05: (2002) 062.

[19] et al., V. K. "Measurement of the Z boson differential cross section in transverse momentum and rapidity in proton–proton collisions at 8 TeV." *Phys. Lett.* B749: (2015) 187–209.

[20] Altarelli, G., and G. Parisi. "Asymptotic Freedom in Parton Language." *Nucl. Phys. B* .

[21] Arneodo, M. "Nuclear effects in structure functions." *Phys. Rept.* 240: (1994) 301–393.

[22] Ball, R. D., and A. Deshpande. "The Proton Spin, Semi-Inclusive processes, and a future Electron Ion Collider." https://inspirehep.net/record/1648159/files/arXiv:1801.04842.pdf.

[23] Ball, R. D., V. Bertone, L. Del Debbio, S. Forte, A. Guffanti, J. Rojo, and M. Ubiali. "Theoretical issues in PDF determination and associated uncertainties." *Phys. Lett.* B723: (2013) 330–339.

[24] Ball, R. D., L. Del Debbio, S. Forte, A. Guffanti, J. I. Latorre, A. Piccione, J. Rojo, and M. Ubiali. "A Determination of parton distributions with faithful uncertainty estimation." *Nucl. Phys.* B809: (2009) 1–63. [Erratum: Nucl. Phys.B816,293(2009)].

[25] Butterworth, J., et al. "PDF4LHC recommendations for LHC Run II." *J. Phys.* G43: (2016) 023,001.

[26] Campbell, J., and R. Ellis. "An update on vector boson pair production at hadron colliders." *Phys. Rev.* .

[27] Cooper-Sarkar, A. M. "PDF Fits at HERA." *PoS* EPS-HEP2011: (2011) 320.

[28] Dokshitzer, Y. L. *Sov. Phys.* .

[29] G. Aad, G. e. a. "Measurement of the transverse momentum and ϕ^*_η distributions of Drell–Yan lepton pairs in proton–proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector." *Eur. Phys. J.* C76, 5: (2016) 291.

[30] Gavin, R., Y. Li, F. Petriello, and S. Quackenbush. "FEWZ 2.0: A code for hadronic Z production at next-to-next-to-leading order." *Comput. Phys. Commun.* 182: (2011) 2388–2403.

[31] Gribov, L. N., and V. N. Lipatov. *Sov. J. Nucl. Phys* .

[32] J.C. Collins, D. S., and G. Sterman. "Factorisation of Hard Processes in QCD." *arXiv:hep-ph/0409313* .

[33] Khalek, R. A., J. J. Ethier, and J. Rojo. "Nuclear Parton Distributions from Neural Networks." In *Diffraction and Low-x 2018 (Difflowx2018) Reggio Calabria, Italy, August 26-September 1, 2018*. 2018.

[34] S. Alekhin, J. B., and S. Moch. "The ABM parton distributions tuned to LHC data." *Phys. Rev.* D89, 5: (2014) 054,028.

[35] S. Catani, S., and M. H. Seymour. "A General algorithm for calculating jet cross-sections in NLO QCD." *Nucl. Phys.* B485: (1997) 291–419. [Erratum: Nucl. Phys.B510,503(1998)].

[36] Tzanov, M., et al. "Precise measurement of neutrino and anti-neutrino differential cross sections." *Phys. Rev.* D74: (2006) 012,008.