Closure tested parton distributions for the LHC

Christopher S. Deans

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy
to the
University of Edinburgh
July 2015

Abstract

Parton distribution functions (PDFs) provide a description of the quark and gluon content of the proton. They are important input into theoretical calculations of hadronic observables, and are obtained by fitting to a wide range of experimental data. The NNPDF approach to fitting PDFs provides a robust and reliable determination of their central values and uncertainties. The PDFs are modelled using neural networks, while the uncertainties are generated through the use of Monte Carlo replica datasets.

In this thesis I provide an in depth description of development of the latest NNPDF determination: NNPDF3.0. A number of novel adaptations to the genetic algorithm and network structure are outlined and the results of tests as to their effectiveness are shown. Centrally, the use of closure tests, where artificial data is generated according to a known theory and used to perform a fit, has been instrumental in both the development and validation of the NNPDF3.0 approach. The results of these tests, which demonstrate the ability of our methodology to reproduce a known underlying law, are investigated in detail.

Finally, results from the NNPDF3.0 PDF sets are presented. The parton distributions obtained are compared with results from other PDF collaborations, and PDFs fit to limited datasets are also discussed. Physical observables relevant for future collider runs are presented and compared to other determinations.

Declaration

This thesis was composed solely by myself, based on work performed as part of the NNPDF collaboration to which I made a substantial contribution. During my PhD, I was largely involved in the development of the new fitting methodology and code, outlined in Chapter 5, and with implementing and performing the closure tests, outlined in Chapter 6. I additionally gave more minor contributions to a wide range of other collaboration projects.

The results at the end of Chapter 3 are also shown in our paper

- NNPDF Collaboration, *Nucl.Phys.* **B867** (2013) 244–289, [`arXiv:1207.1303`]

while Chapters 4, 6 and 7 were adapted from our paper

- NNPDF Collaboration, *JHEP* **1504** (2015) 040, [`arXiv:1410.8849`].

This work has not been submitted for any other degree or professional qualification.

Christopher S. Deans
June 2015

Acknowledgements

Firstly I'd like to thank my supervisors Richard Ball and Luigi Del Debbio. Their guidance and supervision have been invaluable over the last four years.

During my PhD I had the opportunity to be part of the NNPDF collaboration. I have greatly enjoyed my time working with the group and the friendships I have made. I'd like to thank Stefano Forte, Juan Rojo and José Ignacio Latorre for their advice, Alberto Guffanti, Maria Ubiali, Stefano Carrazza and Valerio Bertone for their help with the projects we worked on, and everyone for the lively and constructive discussions. I'd also like to give special thanks to Nathan Hartland for his friendship and advice during the time we worked together in Edinburgh.

I'd like to thank everyone in the Edinburgh PPT for the friendly and supportive environment in the group. Particular thanks go to Mark, Moritz, Eliana, Sam, Øyvind, Nico, Samuel, Jack, Vlad, Brian, Arjun, Donal, Einan and Jenny.

Finally, I'd like to thank my parents and my fiancée Tirion for their support through my PhD and in all the things I do. I would not have gotten this far without their encouragement.

Lay Summary

The Large Hadron Collider (LHC) experiment is an international effort to understand the fundamental building blocks of the universe. By colliding subatomic particles called *protons* together at extremely high energies we can learn more about physics at both scales much smaller than an atom and much larger than the galaxy.

Protons are *composite* particles, made up of smaller still *partons*: the *quarks* and *gluons*. When two protons collide, it is in fact these partons that actually interact. The behaviour of the quarks and gluons in such high energy collisions can be predicted using the theoretical calculations, however we cannot calculate the amount of the proton which is made up of each type.

Parton distribution functions (PDFs) are a method to characterise the internal structure of the proton, in terms of how much of it is made up of each type of parton. As they cannot be calculated, they are instead determined from the results of previous particle physics experiments. This work is vital in order to fully understand what we see at the LHC.

My work, presented in this thesis, was to determine these parton distributions functions. Working with an international collaboration, I developed a novel method of doing so, fitting the PDFs to experimental data using *neural networks*, a form of artificial intelligence based on the structure of the brain. Our results have already been used for several LHC measurements, and will play an important role in future experimental studies.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the past hundred years, the study of elementary particles and their interactions has become an important and highly active area of physics research. Through high-energy experiments, like the currently running Large Hadron Collider, we obtain an increasing amount of information on particle physics, which informs developments in our theoretical understanding of the topic. The data from these collider experiments—and data from other sources, astronomical data for example—is currently best described by the *Standard Model* of particle physics, replete with the recently observed Higgs Boson [1, 2]. This theory is not a complete description, however, and since May of this year the LHC is being operated at a new higher energy, testing the limits of our understanding and searching for clues of a more fundamental description.

In order to make use of the data collected at the LHC, we need a framework through which to interpret them. As the particles used in the collisions are protons, this framework is provided by Quantum Chromodynamics (QCD), a quantum gauge theory which describes the strong force. Protons are composite particles, and are composed of fundamental particles, quarks, bound together by gluons, the force carriers of QCD. A calculation of a typical LHC process will typically involve two parts: the high energy (or "hard") central collision between individual quarks or gluons, and the lower energy ("soft") interactions within the proton. Much like in Quantum Electrodynamics, the quantum field theory describing the electromagnetic force, many predictions can be made with QCD by using a perturbative expansion in the coupling constant. However, one important property of QCD is *asymptotic freedom*, the phenomenon that the strength of an interaction is inversely related to the energy at which it occurs. At high energies, like those at the LHC, the coupling constant is small and so the perturbative description is accurate. On the other hand, for interactions at low energy, including those between the quarks inside the proton, the coupling constant is larger than one so the standard method to perform the calculation fails. To generate theoretical predictions of LHC processes we therefore need to use a different approach to handle the soft effects. This approach is the use of parton distributions.

Parton distributions functions, or PDFs, characterise the internal structure of the proton. There is a distribution function for each flavour of quark, and for the gluon itself, and each describes, broadly speaking, how likely it is to find a particular type of quark or gluon with a specific fraction of the proton's total momentum. To perform a calculation of an LHC cross-section, these PDFs must be convoluted with the separate calculations for the hard sub-processes and summed over all flavours. As the PDFs are related to low energy dynamics, they cannot be calculated from perturbative QCD, and must instead be determined from experimental data. Fortunately, the parton distributions are *universal*, and are the same for different experiments and process. This means that PDFs determined using data from one set of experiments can be then used in calculations for different experiments.

Early approaches to PDF determination were fairly rudimentary, based partly on theoretical models and on experimental data limited in both quality and amount. With the increasing demands placed on PDFs for precision hadronic physics, and the increasing amount of data available, PDF fitting has become a very sophisticated exercise. It is no longer sufficient to determine the best fit central values alone, and for modern applications it is necessary to also provide an accurate estimation of the uncertainties on the PDFs. The NNPDF approach seeks to determine PDFs and their associated uncertainties in a way which is accurate and unbiased. The parton distributions are parameterised by *neural networks*, while uncertainties are obtained by generating a Monte Carlo ensemble. This novel approach has been used to perform a number of successful determinations over the past decade [3–5], which have been widely used to perform many theoretical calculations.

In this thesis I will detail the work done in producing a new NNPDF parton distributions determination, NNPDF3.0 [6]. This involved the development of a new fitting code with a substantially updated and validated methodology, and the implementation of many new experimental datasets from the LHC and from the HERA electron-proton collider. Given the importance of PDFs to understanding the results from experiment, NNPDF3.0 also involves a comprehensive statistical study of the effectiveness and accuracy of our methodology. This was performed using the closure test framework, where fits are performed using artificial data, allowing us to compare our results directly to a known correct answer. Alongside validating our approach, this tool has also proven very useful in a number of other ways, from evaluating the impact of different methodological improvements to disentangling the components of the PDF uncertainties from different sources.

Chapter 2 provides a brief overview of some of the theory underlying PDFs and their determination, particularly the role of factorisation. Chapter 3 then builds on these ideas to look at PDF fitting, first in general, and then in detail for the NNPDF approach. Results from the NNPDF2.3 fit are also presented. The remainder of the thesis describes the NNPDF3.0 analysis. Chapter 4 gives information about the data included in the fit, with additional detail for the new datasets not previously used in NNPDF fits. Details about the way the data is implemented, including the theoretical tools used and the treatment of systematic uncertainties, are also provided here. Chapter 5 looks at the methodology of NNPDF3.0, describing a large number of variations to the fitting algorithm and the tests used to determine their effectiveness. While the closure testing framework is introduced in Chapter 5, Chapter 6 looks at the implementation and results of the NNPDF closure tests in much more detail, and provides evidence that the tests demonstrate the validity of our approach. Finally, Chapter 7 provides results of the NNPDF3.0 fits, looking at the PDFs themselves and the quality of fit to the experimental data. Results for fits with reduced datasets are also given, showing the impact of the new data and exploring the possibility of PDFs based on maximally consistent datasets. In addition, a brief study of NNPDF3.0 prediction of standard LHC observables and of select BSM processes is included here.

Chapter 2

The parton model and factorisation

In this chapter I will provide a brief overview of some of the key theoretical concepts related to parton distributions, specifically factorisation and the DGLAP evolution equations. It is not intended to be a thorough or extensive description of these issues (which can instead be found in any of the many good textbooks on the topic, [7] for example), only to give context for the rest of the thesis. I will start by describing the parton model and define the parton distributions themselves.

2.1 The parton model

That the proton was a composite particle, instead of an elementary one, was proposed in 1964 independently by Gell-Mann [8] and Zweig [9] based on the Eightfold Way interpretation of hadrons. Both suggested that the proton was composed of *quarks* (or *aces*) with spin $\frac{1}{2}$ and fractional charge. This picture was validated in 1968 by deep inelastic scattering experiments at SLAC. The experiment collided electrons with proton targets, and found that the inelastic cross-section had very weak dependance on the momentum-transfer (Q^2) of the interaction [10]. This scale independence, known as Bjorken scaling [11], demonstrated that the proton was composed of point-like (or almost point-like) constituents, as we expect the cross-section to depend on the ratio of the scale of the interaction to the scale of the proton's internal structure. These constituents, initially disassociated with the quark model, were named *partons* by Feynman [12].

This discovery gave rise to the parton model description of the proton, where scattering at high energies is described by virtual photons scattering incoherently off of one of the constituent partons. This is essentially moving from a picture shown by

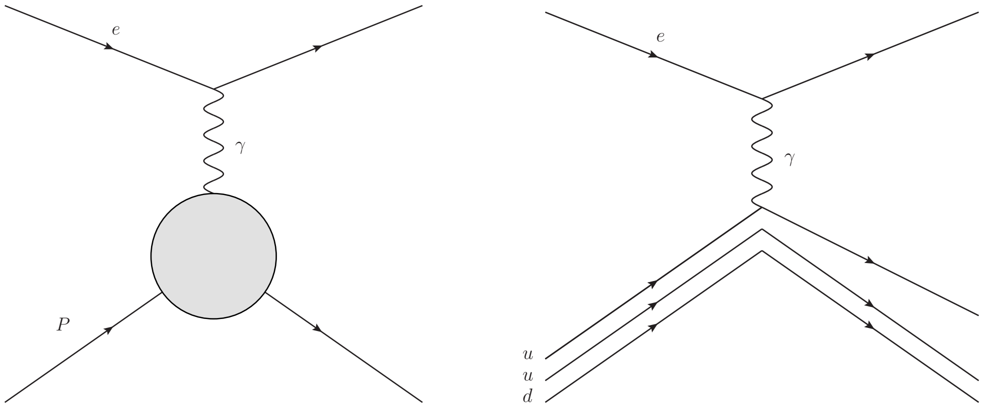


Figure 2.1: Diagrams of DIS scattering between an electron and proton, represented as interaction with the proton (left) and with a single quark in the parton model (right).

the diagram in the left side of Figure 2.1, where the proton interacts directly with the virtual photon by some effective interaction, to the right hand side where the photon interacts with an individual quark. In the parton model the quarks are characterised by distribution functions, which describe the likelihood of encountering each flavour of parton in a collision. More precisely, we define functions $f_i(x, Q^2)$ which give (at LO in QCD) the probability of finding a parton of flavour i with fraction x of the proton's total momentum. x can also be defined in terms of the DIS variables as

$$x = \frac{Q^2}{2M_P\left(E_e - E_e'\right)}, \tag{2.1}$$

where M_P is the mass of the proton and E_e and E_e' are the energies of the incoming and outgoing electron respectively, in the rest frame of the proton. DIS structure functions can then be written in terms of these functions, for example

$$F_2^{EM} = x\left(\frac{4}{9}\left(f_u + f_{\bar{u}} + f_c + f_{\bar{c}}\right) + \frac{1}{9}\left(f_d + f_{\bar{d}} + f_s + f_{\bar{s}}\right)\right). \tag{2.2}$$

In addition to the quark distributions, there is also a distribution function for the gluon. The gluon was not discovered in the same DIS experiments as the rest of the partons, and was instead seen for the first time in the PLUTO experiment at the DORIS electron collider at DESY [13]. As they have no electric charge, gluons are hard to see directly in collisions between protons and electrons, with the main effects of the gluon distribution in the proton coming from violations of the Bjorken scaling due to QCD interactions between the partons. The gluon distribution is however much more important in hadron scattering experiments like the Tevatron and the LHC, where quark-gluon and gluon-gluon scattering events are frequent.

The parton distributions are subject to a number of sum rules. From the definition, the total momentum carried by the partons cannot exceed that of the proton itself, giving the momentum sum rule

$$\sum_i \int_0^1 dx\ x\ f_i(x) = 1. \tag{2.3}$$

There are also three (or more fully, six, including sum rules for charm, bottom and top) valence sum rules which reflect the quark content of the proton, which are

$$\int_0^1 dx\ \left(u(x) - \bar{u}(x)\right) = 2; \qquad \int_0^1 dx\ \left(d(x) - \bar{d}(x)\right) = 1;$$
$$\int_0^1 dx\ \left(s(x) - \bar{s}(x)\right) = 0. \tag{2.4}$$

2.2 Factorisation

In any collision involving one or more hadrons, the are multiple scales involved, from the (generally) 'hard' scale of the central interaction to the 'soft' scale of the QCD interactions holding the hadron together. This might appear to make calculations of such collisions impossible due to the failure of perturbative QCD at low energies, resulting in divergences in the theory. However, it is possible to separate the long and short distance behaviour by a process called *factorisation.* The soft parts of the interaction are subsumed into the parton distributions functions, leaving only the hard, calculable part. This allows the cross-sections for DIS processes to be written as a convolution between a hard scattering kernel C and the factorised parton distributions, i.e.

$$\sigma(x, Q^2) = \sum_i \int_x^1 \frac{d\xi}{\xi}\ C_i\left(\frac{x}{\xi}, \alpha_S\left(Q^2\right)\right) f_i(\xi, Q^2). \tag{2.5}$$

The parton distributions themselves are now defined in terms of *bare* distribution functions f_i^0 and QCD *splitting functions*, P_{ij}, at a chosen factorisation scale μ_F, i.e.

$$f_i(x, \mu_F^2) = f_i^0(x) + \frac{\alpha_S}{2\pi}\sum_j \int_x^1 \frac{d\xi}{\xi}\ f_j^0(\xi)\left(C_{ij}\left(\frac{x}{\xi}\right) + P_{ij}\left(\frac{x}{\xi}\right)\ln\frac{\mu_F^2}{\kappa^2}\right) + \ldots \tag{2.6}$$

where κ is a small cut-off regulating the singularities (dimensional regularisation could instead be used here) and C_{ij} are finite contributions which are to some extend arbitrary, and for which different choices define different 'factorisation schemes'. The factorisation scale μ_F is an unphysical parameter which defines the boundary between what is considered hard and soft in the theory. Partons with a transverse momentum smaller than μ_F are considered long-distance and factorised into the hadron's structure. Similarly factorisation in DIS has been rigorously proved to all orders in perturbation theory [14].

One of the consequences of separating the soft and hard parts of interactions through factorisation is that the soft description (i.e. the parton distributions) are universal, and are the same independent of the details of the hard interaction. The same parton distribution functions described above can therefore be used also for hadronic process. Here, there is an equivalent expression to Eq. 2.5, now involving two parton distributions, one for each of the incident hadrons

$$\sigma = \sum_{ij} \int dx_1 dx_2 \; f_i(x_1, Q^2) f_j(x_2, Q^2) \; \hat{\sigma}_{ij}. \tag{2.7}$$

Unlike in the DIS case, there is currently no full proof that factorisation holds in hadronic processes, largely due to the introduction of possible colour correlations across the two incident hadrons. It has however, been proven for a number of inclusive processes, and the impact is expected to decrease at higher energies [7].

2.3 DGLAP evolution equations

The factorised parton distributions in Eq. 2.6 cannot be calculated perturbatively, as they depend on the long-distance interactions within the proton. Their dependance on the factorisation scale μ_F, however, can be calculated. The evolution of the parton distribution functions with the scale is given by the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) equation [15–17]

$$\mu_F^2 \frac{\partial}{\partial \mu_F^2} f(x, \mu_F^2) = \frac{\alpha_S\left(\mu_F^2\right)}{2\pi} \int_x^1 P\left(\frac{x}{\xi}, \alpha_S(\mu_F^2)\right) f(x, \mu_F). \tag{2.8}$$

More precisely, as the quarks and gluon distributions are connected by the splitting functions, the DGLAP equations take the form of a system of $2n_f + 1$ differential equations

$$\mu_F^2 \frac{\partial}{\partial \mu_F^2} f_i(x, \mu_F^2) = \frac{\alpha_S\left(\mu_F^2\right)}{2\pi} \sum_j \int_x^1 P_{ij}\left(\frac{x}{\xi}, \alpha_S(\mu_F^2)\right) f_j(x, \mu_F)\;. \tag{2.9}$$

Fortunately, SU(n_f) flavour symmetry and charge conjugation mean that this set of equations can be greatly simplified. In fact for particular combinations of the quarks the evolution equation can be written in a separated form like Eq. 2.8, specifically the non-singlet valence and triplet combinations

$$
\begin{aligned}
V &= \sum_i q_i^- \\
V_3 &= u^- - d^- \\
V_8 &= u^- + d^- - 2s^- \\
V_{15} &= u^- + d^- + s^- - 3c^- \\
V_{24} &= u^- + d^- + s^- + c^- - 4b^- \\
V_{35} &= u^- + d^- + s^- + c^- + b^- - 5t^- \\
T_3 &= u^+ - d^+ \\
T_8 &= u^+ + d^+ - 2s^+ \\
T_{15} &= u^+ + d^+ + s^+ - 3c^+ \\
T_{24} &= u^+ + d^+ + s^+ + c^+ - 4b^+ \\
T_{35} &= u^+ + d^+ + s^+ + c^+ + b^+ - 5t^+ \ ,
\end{aligned} \tag{2.10}
$$

where the quark plus and minus distributions are given by

$$q_i^\pm = q_i \pm \bar{q}_i \ . \tag{2.11}$$

For the gluon distribution and the remaining quark combination, the singlet

$$\Sigma = \sum_i q_i^+ \ , \tag{2.12}$$

evolution is still given by a pair of coupled equations, as in Eq. 2.9.

The DGLAP equations can be directly solved in x-space using numerical methods. This is the approach used many QCD tools, such as HOPPET [18] and APFEL [19], generally with some form of interpolation to improve the speed of the calculation. Another way to obtain a solution is to perform a Mellin transformation

$$M[f][N] = \int_0^1 dy\, y^{N-1} f(y) \tag{2.13}$$

of the equations, which reduces the convolutions in Eqs. 2.8 and 2.9 to multiplications. This is the approach used to produce evolution kernels in NNPDF fits, as described in [4, 20].

Chapter 3

PDF Determination

The ideas discussed in the previous chapter provide an approach we can follow to obtain parton distributions: take data from a large number of experiments, making use of universality, and use factorised theoretical calculations to fit the distributions. We can look at each each element of this in turn. The need for good quality, high precision data covering a wide kinematic range is central to the determination of PDFs. Traditionally, DIS data has formed the core of the dataset used, from experiments like SLAC and CERN SPS, and, in more recent determinations, data from the HERA experiments. However, there is now a large amount of relevant data from hadronic colliders, with data from the Tevatron and new sets released every year from the LHC.

One constraint on the data that can be included is the need to have a theoretical description of the data in order to fit it. DIS data is straightforward in this respect, but for hadronic data there are still a number of processes without a full description at NNLO, or for which there is a description but it is not implemented in a fast enough format to be used in PDF fits. A substantial amount of recent progress has been made in this respect, with a large number of new calculations and tools having been released in the last few years. Further work has been made in improving the quality of the theoretical calculations, with fits using NNLO theory being the current standard, and the possibility of N3LO fits already being discussed.

With the data and the theory, the next step is to actually fit the parton distributions. This takes the form of a standard fitting exercise, where the parameterised PDFs are modified, usually in a directed way, comparing the theoretical value to the data until the best set of parameters is obtained. Modern PDF sets include uncertainties, which are generally calculated according to the Hessian or Monte Carlo approaches, both detailed later in this chapter. The full PDF set, with uncertainties, is then provided in the common LHAPDF format [21], so that they can be easily used with the variety of particle physics codes and tools.

In this chapter I will describe the fitting procedures of general global PDF fits and the innovations used in the NNPDF approach. I will finish by describing the NNPDF2.3 fit, the first global PDF analysis to use LHC data.

3.1 Global QCD Fits: MMHT and CTEQ-TEA

With the importance of a thorough understanding of PDFs for modern collider experiments, it is unsurprising that there are a large number of different groups, each using different approaches and datasets to obtain their distributions. For instance, there are recent updates of the HERAPDF group [22], which performs PDF determinations using only data from the HERA experiments in order to ensure a fully consistent and under-control control dataset, and from ABM [23], who aim to provide as complete and transparent a theoretical treatment as possible for factors like higher-twist effects and nuclear corrections. Additionally, there are a large number of determinations of other aspects of proton structure: nuclear PDFs with extra parameters for the A and Z of the nucleus; polarised PDFs which include information on parton spin; transverse momentum PDFs which constrain also parton transversity and correlations; generalised PDFs which combine PDFs with electric form factors; double PDFs useful for multiple hadron scattering events at the LHC.

In this section I will focus on two other major series of global PDF fits produced by the MSTW/MMHT and CTEQ collaborations. Both groups have very recently released new PDF sets, MMHT2014 [24] and CT14 [25]. The MMHT PDFs are the latest update of the widely used MSTW2008 PDF sets [26], which in turn follow on from MRST and MRS PDF sets going back over 25 years [27–29]. The CT14 PDFs also derive from a long ancestry, with multiple sets of CTEQ PDF released since the CTEQ1 in 1993 [30, 31].

The MMHT and CTEQ methodologies are broadly similar. Both use essentially the same dataset, and, as I will describe over the next few subsections, perform Hessian fits using fixed functional forms. The fact that, despite the similarity of their approaches, there was some disagreement between the results they obtained was one of the pushes towards the development of the substantially different NNPDF methodology, discussed later in this chapter. However, thanks to a number of benchmarking exercises between the different sets over the last few years [32, 33], and changes to the methodologies and theory treatments as a result of these, the MMHT, CT and NNPDF PDF sets are in increasingly good agreement. This has recently culminated in efforts to produce combined PDF sets, integrating the results from the different groups in a statistically correct way [34, 35].

3.1.1 Functional Forms

A full PDF determination involves fitting each of the thirteen independent parton distributions, corresponding to the six quark and anti-quark flavours and the gluon. However, as the charm, bottom and top masses are usually larger than the scale at which perturbative QCD gives a good description of the interactions within the proton, the usual approach of generating the c, b and t distributions radiatively from the other quark distributions is generally sufficient, and they do not need to be separately parameterised. There is possibly a small non-perturbative “intrinsic” component of the charm PDF, and this is a subject of previous [36] and ongoing work [37–39]. There have also been a number of determinations of the photon PDF of the proton [40, 41].

The seven remaining flavours are generally not parameterised directly (except at LO), and instead for technical reasons combinations of the quark PDFs, often close to the parton evolution basis, are determined. As the PDFs at different scales are related by the DGLAP evolution equations (Eqs. 2.9) only the x dependence needs to be parameterised.

The standard form of parameterisation, used by the majority of approaches, for parton distribution f_i is

$$f_i(x, Q_0^2) = x^{\alpha_i}(1-x)^{\beta_i} P_i(x) \tag{3.1}$$

where α_i and β_i are parameters and $P_i(x)$ are functions which depend on x and typically other parameters. The terms proportional to x and $(1-x)$ are motivated by Regge theory and quark counting rules, and constrain the behaviour of the parton distributions as $x \to 0$ and $x \to 1$ respectively. In the past, a common choice for the form of the $P_i(x)$ were polynomials, often mixed with half-integer powers or exponentials. The latest MMHT and CTEQ-TEA releases, however, use more complicated forms for $P_i(x)$: linear combinations of Chebyshev polynomials for most partons in MMHT2014, and of Bernstein polynomials in CT14. The introduction of these more flexible approaches avoid problems seen with the more fixed parameterisations, such as the need to add extra parameters to get a good fit to new data and an increase in the PDF uncertainties when the new parameters are used [42].

3.1.2 PDF uncertainties: The Hessian approach

In both the MMHT2014 and CT14 PDF sets, PDF uncertainties are represented using the Hessian approach. The approach is based on the assumption that the probability distribution for each of the PDFs is given by a multi-Gaussian distribution in the space of the parameters. This assumption is probably sound, at least in the region well constrained by data, both for the usual Central-Limit-Theorem-esqe rationalisations and also because the experimental uncertainties are themselves generally taken to be Gaussian, though factors like positivity and sum rules may disrupt this.

The Hessian method proceeds by first obtaining the best fit PDF, for instance by finding the set of parameters which minimises a χ^2 function to the data. The relationship between the χ^2 and the probability distribution of the parameters can then be used to define confidence intervals (or here volumes) in parameter space by expanding around the minimum value of the χ^2. In the case of Gaussian uncertainties, the 68% confidence interval will be given by volume defined by $\chi^2 = \chi^2_{\text{min}} + 1$, the 90% interval by $\chi^2 = \chi^2_{\text{min}} + 2.69$ etc. This can be found by looking at the covariance matrix for the parameters in parameter space, which is equivalent to the inverse of the Hessian matrix

$$H_{ij} = \frac{1}{2} \frac{\partial^2 \chi^2}{\partial a_i \partial a_j}\bigg|_{\text{min}} \tag{3.2}$$

where i and j run over the parameters a_i, and the derivatives are evaluated at the location in parameter space which gives the minimum χ^2. This gives a convenient representation for the uncertainties which can be provided for use: the eigenvectors of the Hessian. By providing these—or at least the subset of eigenvectors sensitive to the data—the uncertainty in the PDFs and in any observable dependant on the PDF can be calculated in a straightforward manner.

In global fits, it was found [43] that this standard criteria for determining the confidence intervals was inadequate, as the obtained limits were incompatible with the results of fits to each of individual datasets. The reasons for the discrepancy are still not fully understood, but are suspected to be partly due to inconsistencies between datasets, within datasets (i.e. incorrectly estimated systematics) and also from parameterisation bias. The problem was alleviated in the MSTW and CTEQ fits by introducing a tolerance factor T so that instead of a deviation of one from the minimum χ^2, the 68% confidence interval was defined as $\chi^2 = \chi^2_{\text{min}} + T$. This tolerance factor is determined so that the spread of best fit to each experiment is compatible with the uncertainties of the global fit, and this is done dynamically in more recent analyses [26, 44]. The uncertainties obtained by the Hessian method in the CT14 fit are also checked using a separate method of determining the PDF uncertainties, the Lagrange multiplier method [25].

The PDF uncertainties calculated using the Hessian method, and using the Monte Carlo method in NNPDF fits, refers to only the propagation of the experimental uncertainties and uncertainties related to the fit (e.g. interpolation and extrapolation uncertainties). There are also a number of theoretical uncertainties which are relevant for PDF fits. These include uncertainties relating to nuclear corrections, higher twist effects, higher order corrections, and the treatment and precise value of heavy quark masses. For the most part these are dealt with by removing data for which they are particularly relevant, and by performing additional fits with various models of corrections to determine the size of the effects. The impact of several of these factors on the NNPDF3.0 is investigated in Chapter 7.

3.2 The NNPDF Approach

The NNPDF approach takes the same general shape as that used by the MMHT and CTEQ collaborations, with parameterised PDFs determined by comparing to experimental data. However, there are several key differences between the approaches. Centrally, the fixed functional forms used to parameterise the PDFs in other determinations are replaced by neural networks, which are considerably more flexible. The use of neural networks then prompts the adoption of a number of other methodological features, due to the large number of parameters, including the use of a genetic algorithm to perform the minimisation, and the generation of Monte Carlo replica PDFs to encode the PDF uncertainties. Together, this results in a consistent and successful approach which has been used to produce several PDF sets, including the first global PDF set to include LHC data in the determination, NNPDF2.3, which was released in 2012 [45]. In the remainder of this chapter I will discuss the main methodological features of the NNPDF approach and describe some of the results from the NNPDF2.3 analysis.

3.2.1 Neural Network Parameterisation

Neural Networks are processing systems with a particular structure which is based on observations of the brain. They were originally suggested as a way to mathematically model biological neural systems [46], however it is in machine learning and signal processes that neural networks have found their most important applications. One specific type, feed-forward neural networks (also sometimes called multi-layer perceptrons), are in particular very useful for modelling and pattern recognition.

A general neural network consists of a set of nodes and the connections between the nodes. The state of each node is described by a number called the *activation*, which is determined by the activations of the surrounding nodes. In feed-forward neural networks the nodes are arranged into layers, and the activation of each node only depends on state of the nodes in the previous layers. The nodes of the first layer of the network are used to provide inputs to the network (the input layer), while the activations in the final layer provide the outputs (the output layer). The other layers in the network are called hidden layers. An example of a network with two hidden layers

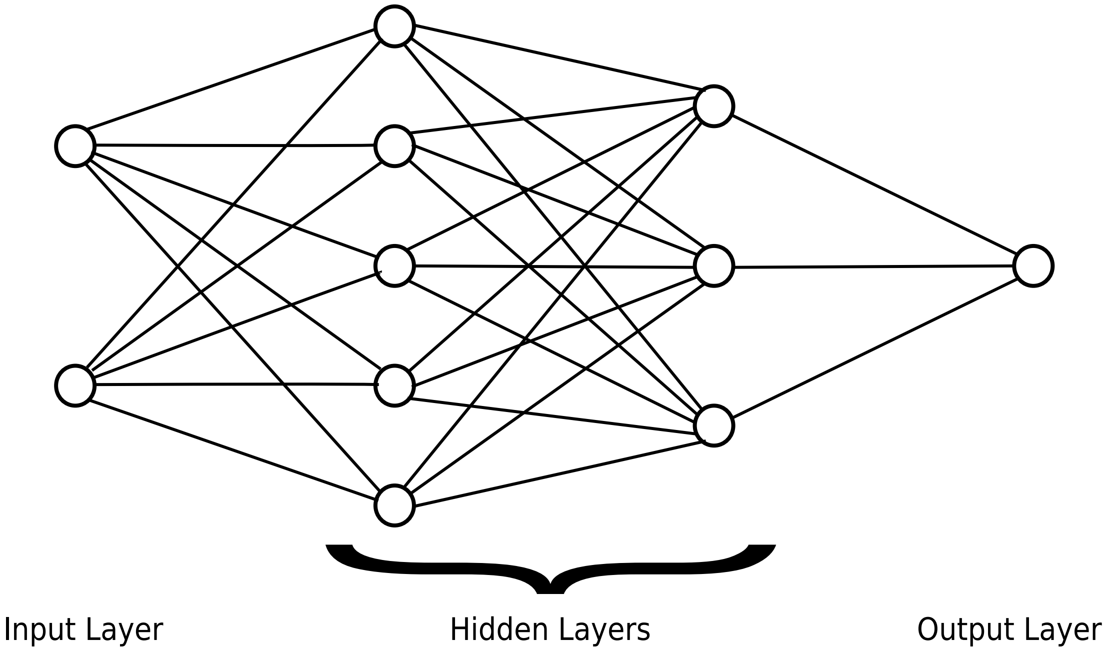


Figure 3.1: Diagram of a neural network with two nodes in the input layer, one node in the output layer, and two hidden layers with five and three nodes.

is shown in Fig. 3.1. In this way feed-forward neural networks can be used to model functions, and it has been shown that a network with a single sufficiently large hidden layer can approximate any continuous function [47].

Despite the degree of complexity they give rise to, the rules for calculating the activation of nodes in a feed-forward neural network are actually quite simple. Consider the ith node in the lth layer of the network. Its activation $\xi_i^{(l)}$ is given by

$$\xi_i^{(l)} = f\left(\tau_i^{(l)} + \sum_j w_{ij}^{(l)} \xi_j^{(l-1)}\right) , \tag{3.3}$$

where $\tau_i^{(l)}$ is a threshold term belonging to the node, $w_{ij}^{(l)}$ are the *weights* of the connections between the nodes in the $(l-1)$th and lth layers, and $f(x)$ is the activation function. There are several suitable choices for the activation function including threshold (i.e. $f(x) = 1$ if $x > 0$ and 0 otherwise), logistic ($f(x) = 1/(1 + e^{-x})$) and linear. Note that the $\tau_i^{(l)}$ term in Eq. 3.3 can be absorbed into the sum by considering it as the weight $w_{i0}^{(l)}$ between the node and an otherwise disconnected node with a constant activation of one. It is also possible to extend Eq. 3.3 to include connections from a node to other nodes two or more layer back, instead of just those in the previous layer, though in applications presented here this option is not used.

As Eq. 3.3 shows, the output of a feed-forward neural network depends on both the input parameters and the weights $w_{ij}^{(l)}$. By changing these weights, the function modelled by the neural network can be modified. In this way feed-forward neural networks can provide a flexible parametrisation for fitting, with the weights acting as parameters.

Neural networks are widely used across physics, including for many applications in particle physics. They were used extensively in Tevatron analyses, including for Higgs searches [48], for signal-background discrimination in single top production [49], and for τ and jet identification in BSM searches [50]. Neural networks were also heavily used for b-jet tagging at the Tevatron [51, 52], and continue to be used in this way at the LHC [1].

In the NNPDF analyses, we use a separate neural network to parametrise each independent PDF flavour combination at the initial fitting scale. We use deep feed-forward neural networks each with identical structure of 2-5-3-1, i.e. with two hidden layers containing five and three nodes, and for a total of 37 parameters per PDF. Fig. 3.1 shows a 2-5-3-1 network. The two input nodes are used to input x and additionally $ln(x)$, where the latter of these is included to reduce the time required to train the networks. A logistic function (shown above) is used for the activation function in order to encourage a smooth output. The suitability of these choices for NNPDF fits, including that the structure is sufficient to model the PDFs, has been investigated in the past [3], and I will also present some results on this topic from closure tests in Section 5.4.3.

For the NNPDF2.3 analysis, seven independent PDFs were parameterised:

- Gluon $g(x)$,
- Singlet $\Sigma(x) = u(x) + \bar{u}(x) + d(x) + \bar{d}(x) + s(x) + \bar{s}(x)$,
- Valence $V(x) = u(x) - \bar{u}(x) + d(x) - \bar{d}(x) + s(x) - \bar{s}(x)$,
- Triplet $T_3(x) = u(x) + \bar{u}(x) - d(x) - \bar{d}(x)$,
- Sea Asymmetry $\Delta_S(x) = \bar{u}(x) - \bar{d}(x)$,
- Strange sea $s^+(x) = s(x) + \bar{s}(x)$,
- Strange valence $s^-(x) = s(x) - \bar{s}(x)$.

This choice of basis was made in order to have a basis which was close to the full evolution basis, in order to optimise the rotation of the initial scale PDFs for evolution, but also to use combinations with some physical interpretation. However, given the flexibility of the neural network parameterisation, which specific PDF combinations are used should not affect the results of the fit.

In order to reduce the time required for the neural networks to model the data, we additionally use preprocessing terms in the definition of the PDFs. The initial scale PDFs are therefore given by

$$f_i^0(x) = A_i\, x^{-\alpha_i}(1-x)^{\beta_i}\mathrm{NN}_i(x), \tag{3.4}$$

with the extra parameters A_i, α_i and β_i. These are not treated as full parameters in the fit, and instead the A_i are used to impose the PDF sum rules, while the α_i and β_i are constants. The x and $1-x$ terms serve broadly the same role here as the equivalent terms in the MSTW parameterisation given previously, ensuring that the theoretical requirements on the large- and small-x behaviour of the PDFs are automatically satisfied. The preprocessing terms also speed up training, as the neural networks only need to model the deviations from this underlying form. However, in order to avoid biasing the fit the value of the preprocessing exponents α_i and β_i are randomly selected at the start of the fit from a pre-specified range. This range is chosen so that it is large enough not to bias the fit, but not so large that unphysical values or values which lead to a very poor fit can be selected.

While the preprocessing exponents remain constant during the fit, this is not case for several of the overall normalisations (A_i in Eq 3.4). Four of these, A_g, A_V, A_{Δ_S} and A_{s^-} are instead set in order to explicitly impose PDF sum rules. The four sum rules imposed are the total momentum sum rule

$$\int_0^1 dx\, x\,(\Sigma(x) + g(x)) = 1 \tag{3.5}$$

and the valence sum rules (in the NNPDF2.3 basis)

$$\int_0^1 dx\, V(x) = 3, \tag{3.6}$$

$$\int_0^1 dx\, (T_3(x) - 2\Delta_S(x)) = 1, \tag{3.7}$$

$$\int_0^1 dx\, s^-(x) = 0. \tag{3.8}$$

The remaining PDF normalisations A_Σ, A_{T_3} and A_{s^+} are simply set to one.

3.2.2 Genetic Algorithm

Neural networks provide a flexible, unbiased parametrisation in NNPDF analyses, however this flexibility comes at the cost of having a large number of parameters. As a result, our fitting methodology needs to be able to search through a very large parameter space. In order to do this efficiently, we use a genetic algorithm to perform the minimisation.

The basic principles of our genetic algorithm are straightforward. Each generation a large number of copies of the current best fit PDFs are generated. Each copy is then mutated by changing the parameters, to create a set of mutant PDF sets. The quality of fit to the data of each mutant is then calculated, according to some figure of merit generally based on comparison to the experimental data, and the best mutant is taken forward as the parent for the next generation. This process is then iterated until the stopping condition is satisfied or a maximum number of generations is reached.

There are many different methods which can be used to mutate the PDFs. The NNPDF methodology uses a relatively simple approach, where a small number of randomly selected parameters are changed by an amount given by

$$w \to w + \eta \frac{r_1}{N_{\mathrm{ite}}^{r_2}}, \tag{3.9}$$

where w is the parameter being mutated, N_{ite} is the number of generations which have elapsed, r_1 is a uniform random number between -1 and 1, r_2 is a second random number between 0 and 1, and η is a parameter which controls the size of the mutations. I will present tests of more complicated mutation strategies in Section 5.3.

Essentially, at each step the genetic algorithm takes a random sample from the parameter space around the current best fit, and if a better place is found the fit moves there. The η parameter in Eq. 3.9 is therefore very important as this controls the overall size of the mutations, and so the average size of the step in parameter space each generation. Large mutations allow the algorithm to quickly move through the parameter space, while small mutations allow the fit to finely manoeuvre close to the global minimum. For this reason, we use a value of η which depends inversely on the number of generations reached so far in the fit, which means that the size of the mutations decreases as the fit progresses.

The figure of merit used in the NNPDF fits is a χ^2 function of the differences between the theoretical predictions $t_i(f)$ of the data points and the experimental values d_i, i.e.

$$\chi^2\left(\boldsymbol{d}, \boldsymbol{t}(f)\right) = \frac{1}{N_{\mathrm{dat}}} \sum_{i,j} \left(d_i - t_i(f)\right) C_{ij}^{-1} \left(d_j - t_j(f)\right) \tag{3.10}$$

where C_{ij}^{-1} is the inverse of the covariance matrix of the data.

Because neural networks are very flexible, using them to fit functions introduces a significant risk of over-learning (or over-fitting). This is where the neural network starts to model the statistical fluctuations in the data, instead of just the underlying pattern. This results in a biased fit and a reduction in predictive power. In order to control over-learning we use *cross-validation* to impose a stopping condition on the fit. The dataset is split into two halves, and one half is used to train the networks (the training set), while the other is used to look for over-fitting (the validation set). If the χ^2 to the unseen validation set increases, this indicates that the fit has begun to over-learn, and the fit is stopped. In NNPDF fits, the training and validation sets are constructed using half of each dataset, randomly chosen, so each data point has a 50% chance of being in each set.

3.2.3 t_0 approach to normalisation uncertainties

One issue with the definition of the χ^2 (Eq. 3.10 above) used in the fit is the treatment of multiplicative uncertainties. These uncertainties, of which normalisation uncertainties—like the uncertainty on collider luminosity—are a particular type, are not simply a fixed value but depend on the central value obtained. One common example is the uncertainty on collider cross-sections coming from the measurement of luminosity.

In most applications, the distinction between additive and multiplicative uncertainties is unimportant and they can be treated in the same way. However, when fitting a function, multiplicative uncertainties can introduce a d'Agostini bias [53]. Since the absolute size of the uncertainty depends on the central value of the data point, the error on points with a downward statistical fluctuation is smaller than that on the data points with an upward fluctuation. The fit will therefore develop a downwards bias.

There are several different methods to remove this bias, with most involving the use of a modified error function during the fit. For the NNPDF fits, we solve the d'Agostini bias by using a modified covariance matrix in the χ^2. Instead of calculating the absolute value of the multiplicative uncertainties using the central experimental value, we instead use the central theory value from a previous fit. The resulting covariance matrix is called the t_0 covariance matrix. We then iterate, performing multiple fits, each using a covariance matrix based on the results from the last, until the PDFs converge. Fortunately this process generally only requires a small number of generations to complete. Using the theory value rather than the experimental value smooths out the fluctuations in the data, removing the bias, and has been demonstrated to be effective even in complicated situations involving multiple correlated datasets [54].

3.2.4 PDF uncertainties: the Monte Carlo approach

As I previously mentioned, it is important for modern PDF determinations to obtain accurate PDF uncertainties. In order to determine the uncertainties in NNPDF fits,

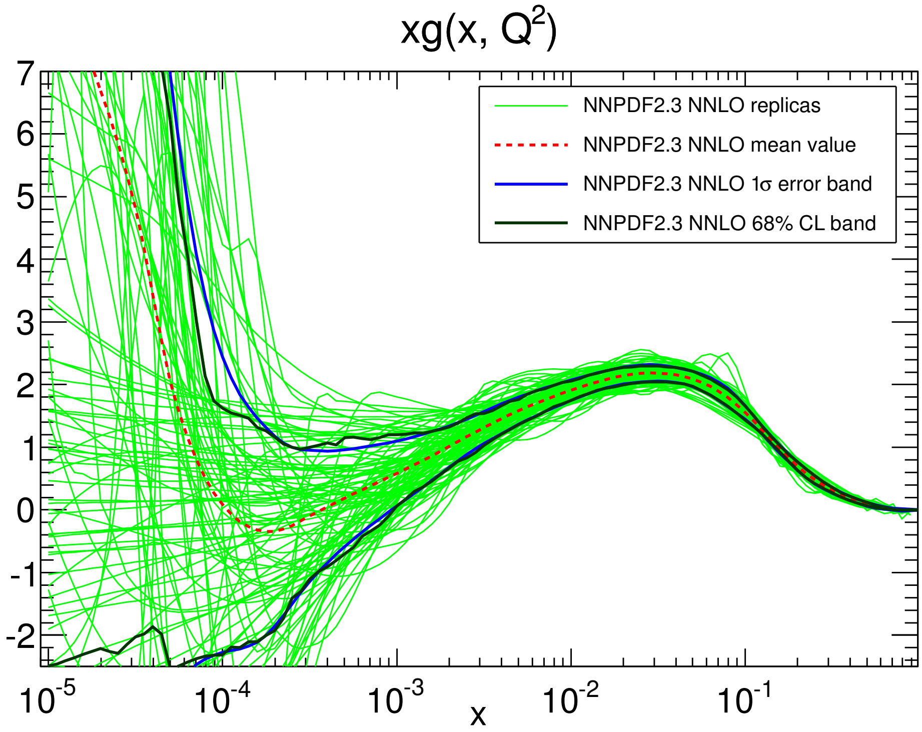


Figure 3.2: Replicas of the NNPDF2.3 NNLO gluon. Each green line shows the gluon of one replica PDF set at the initial (fitting) scale $Q^2 = 2GeV^2$. Also shown are the mean (red dashed line) and the one-sigma (blue lines) and 68% confidence (black lines) intervals.

we produce sets of Monte Carlo replica PDFs. First, replica datasets are generated by randomly fluctuating the experimental data according to its uncertainties. A separate fit is then performed to each replica dataset, and the resulting set of PDF determinations are collected as a single Monte Carlo PDF set. The idea is that this creates a sample of the probability distribution of the PDFs in function space. This sample can then be used to calculate expected values, standard deviations, correlations and any other statistical estimator is the usual way for any probabilistic sample. Unlike with the Hessian approach, described previously in this chapter, this approach can describe non-gaussian uncertainties.

An example of the resulting PDF determination is shown in Fig. 3.2, where the gluon from each of the 100 replica from the NNPDF2.3 NNLO PDF set are plotted. Each green line is the result of a separate Monte Carlo replica fit, and taken together they build up a single PDF fit with a central value given by the mean, shown by the red dashed line, and uncertainty given by the one-sigma band in blue. The replica PDFs are clustered close together at large and medium x, where the gluon is well determined, and diverge rapidly at small x. Note that while the individual replicas often have quite complicated shapes, the resulting central value is a smooth function. A set of lines showing the central 68% of replicas (i.e. the seventeenth replica from the top and bottom in this case) are also shown in black. In general, the one-sigma band and this 68% confidence interval agree well indicating that the uncertainties on this PDF are likely close to Gaussian.

Each replica data point is produced according to

$$d_i^{\mathrm{rep}} = \left[\prod_{j}^{N_{\mathrm{mult}}} \left(1 + r_j^{\mathrm{mult}} \sigma_{i,j}^{\mathrm{mult}}\right) \right] \left(d_i^{\mathrm{exp}} + \sum_{k}^{N_{\mathrm{add}}} r_k^{\mathrm{add}} \sigma_{i,j}^{\mathrm{add}} + r_i^{\mathrm{stat}} \sigma_i^{\mathrm{stat}} \right), \tag{3.11}$$

where $\sigma_{i,j}^{\mathrm{mult}}$ are the N_{mult} correlated multiplicative uncertainties for this data point, $\sigma_{i,j}^{\mathrm{add}}$ the N_{add} correlated additive uncertainties, and σ_i^{stat} is the (additive) statistical uncertainty. r_j^{mult} and r_k^{add} are unit variance Gaussian random numbers, which are shared across data-points for which each particular uncertainty is correlated, while r_i^{stat} is another Gaussian random number which is unique for each point. If a new replica data point generated according to Eq. 3.11 is negative, which can occur for points which have particularly large uncertainties, it is discarded and regenerated until a non-negative value is obtained. This results in data which essentially have a second level of fluctuations added to it, on top of the usual fluctuations between the experimental measurement and the true value.

The use of replica fits is also convenient for reducing the possibility of bias from other parts of the methodology. Along with different replica data, each replica fit is performed with different values for the preprocessing exponents and different starting values for the parameters of the neural networks. Which data are included in the training and validation sets is also varied for each replica, so rather than half of the data being completely discarded, all of it is used in at least some of the replicas.

As this method produces a probabilistic sample of the PDFs, it is important that the number of replicas used is sufficiently high. However, a large number of replicas takes a substantial amount of time to generate and, more significantly, it will also take longer to produce any theoretical calculation using the resulting set. For this reason, the majority of NNPDF sets contain 100 replicas, which balances precision with usability. A small number of 1000 replica sets are also produced, as these can be useful for some purposes (e.g. reweighting).

3.2.5 FastKernel

One of the most important technical features of the NNPDF methodology is the use of *FastKernel* (FK) tables. These combine the convolutions with both DGLAP evolution kernels and with hard scattering coefficient functions into a single calculation. This calculation can then be reduced to the product of a vector containing the initial scale PDFs at different x values and a matrix of precomputed coefficients (known as *FK tables*), which can be optimised within the NNPDF fitting code. This results in substantially faster fits, which allows for a more in depth scan of the parameter space (by increasing the number of generations, number of mutants per generation etc.) and also for easier testing of new features within the code.

The first step in the FastKernel approach is to combine the evolution and scattering coefficients convolutions. The standard form for an observable σ^I from initial scale PDFs f_i^0, using evolution kernels Γ_{ij} and scattering coefficients C_i^I, is

$$\sigma^I(x, Q^2) = \sum_{j=1}^{N_{\text{pdf}}} C_j^I(Q^2) \otimes \left(\sum_{k=1}^{N_{\text{pdf}}} \Gamma_{jk}(Q^2, Q_0^2) \otimes f_k^0(x, Q_0) \right) . \tag{3.12}$$

While this is strictly correct only for DIS observables, a similar expression can be obtained for hadronic observables by considering products of PDFs and summing to N_{pdf}^2. Using the associative and distributive properties of the convolution, we can rearrange this to get

$$\begin{aligned} \sigma^I(x, Q^2) &= \sum_{k=1}^{N_{\text{pdf}}} \left(\sum_{j=1}^{N_{\text{pdf}}} C_j^I(Q^2) \otimes \Gamma_{jk}(Q^2, Q_0^2) \right) \otimes f_k^0(x, Q_0) & (3.13) \\ &= \sum_{k=1}^{N_{\text{pdf}}} K_k^I(Q^2, Q_0^2) \otimes f_k^0(x, Q_0) & (3.14) \end{aligned}$$

with the new coefficients K_i^I. This approach saves time during the fit as the inner convolution can be performed beforehand and the results saved, rather than performing it every time an observable is calculated.

The next step is to avoid performing this, still fairly costly, single convolution during fits, and instead reduce it to a scalar product which can be calculated very efficiently. This is achieved by approximating the initial scale PDFs with a suitable set of interpolating functions, i.e. by defining N_x functions $\mathcal{I}_i^{(\alpha)}$ and setting

$$f_i^0(x) = \sum_{\alpha=1}^{N_x} f_{i\alpha}^0 \mathcal{I}_i^{(\alpha)}(x) \tag{3.15}$$

where $f_{i\alpha}^0$ are the values of the initial scale PDFs at N_x chosen x values. In the NNPDF methodology we use Hermite polynomials on a grid of 50 points in x, which has been shown to be appropriate to reproduce the full calculation with sufficient accuracy for

PDF fits [4]. With these interpolating functions we can rewrite Eq 3.14 as

$$\sigma^I(x,Q^2) = \sum_{k=1}^{N_{\text{pdf}}}\sum_{\alpha=1}^{N_x}\int_y^1 \frac{dy}{y} K_k^I(\frac{x}{y},Q^2,Q_0^2)\mathcal{I}_k^{(\alpha)}(y) f_{k\alpha}^0 \tag{3.16}$$

$$= \sum_{k=1}^{N_{\text{pdf}}}\sum_{\alpha=1}^{N_x} \mathcal{K}_{k\alpha}^I(Q^2,Q_0^2) f_{k\alpha}^0. \tag{3.17}$$

These arrays $\mathcal{K}_{k\alpha}^I$, known as FK tables, can be computed for each data point once, and then stored for use in any number of fits. Different theory settings require different K_k^I however, so different FK tables need to be generated for different pertubative orders, values of $\alpha_S(M_Z)$, choice of initial scale etc. In practice, many of the entries in each $\mathcal{K}_{k\alpha}^I$ are zero, so further improvements in performance can be obtained by performing the sums in Eq. 3.17 only over the non-zero terms. Again, a similar approach is taken for hadronic data, to produce tables $\mathcal{K}_{jk\alpha\beta}^H$ which are used to calculate observables according to

$$\sigma^H(x,Q^2) = \sum_{j=1}^{N_{\text{pdf}}}\sum_{k=1}^{N_{\text{pdf}}}\sum_{\alpha=1}^{N_x}\sum_{\beta=1}^{N_x} \mathcal{K}_{jk\alpha\beta}^H(Q^2,Q_0^2) f_{j\alpha}^0 f_{k\beta}^0. \tag{3.18}$$

More information about the NNPDF FastKernel implementation can be found in [4] and [45].

3.2.6 Positivity

In the naive parton model, the parton distributions can be directly identified as probability distributions for the quarks and gluon. Moving to the QCD-improved quark model, this stops being the case above lowest order [55]. The result is that while at LO the PDFs themselves are positive (semi-)definite, for NLO and above it is only required that all cross-sections are non-negative. Here, ‘all’ means quantities that could in principle be measured, not just those which we are actually capable of obtaining from experiments. Non-negative observables can of course be obtained by forcing the PDFs to be positive, and this approach is taken in several other determinations, often by using a positive-definite functional form. However, this could potentially bias the PDFs away from forms which are negative but still physical, and in particular result in a artificially smaller uncertainty.

In the NNPDF analyses, we instead impose positivity during the fit by penalising PDF values which produces negative values for a number of additional observables included in the fit. This is done by including an extra term in the figure of merit which is non-zero when an observable is negative. The error function which is used to rank

Dataset	Ref.	$N_{\rm dat}$	$[\eta_{min}, \eta_{\rm max}]$	$\langle\sigma_{stat}\rangle$ (%)	$\langle\sigma_{sys}\rangle$ (%)
CMS W electron asy 840 pb^{-1}	[57]	11	$[0, 2.4]$	2.1	4.7
ATLAS W & Z 36 pb^{-1}	[58]	30	$[0, 3.2]$	1.8	3.8
LHCb W 36 pb^{-1}	[59]	10	$[2, 4.5]$	4.1	10.1
ATLAS Inclusive Jets 36 pb^{-1}	[60]	90	$[0, 4.5]$	10.2	23.7

Table 3.1: Details of the LHC datasets included in the NNPDF2.3 fits.

the mutants in the genetic algorithm is then

$$E(\boldsymbol{d}, f) = \chi^2(\boldsymbol{d}, f) + \lambda_{\rm pos} \sum_{i=1}^{N_{\rm pos}} \mathcal{H}\left(-\mathcal{O}_i(f)\right) |\mathcal{O}_i(f)|, \tag{3.19}$$

where $\mathcal{O}_i$ are the $N_{\rm pos}$ positivity observables, $\mathcal{H}$ is a unit step function, and $\lambda_{\rm pos}$ is a Lagrange multiplier. Note that the amount added to the figure of merit if an observable is negative is proportional to the size of the deviation. The extra parameter $\lambda_{\rm pos}$ is set outside of the fit, and needs to be large enough to properly enforce positivity without drowning out improvements in the fit to the data with slight reductions in negativity.

For NNPDF2.3, we have included positivity observables for three processes: F_L, which constrains the small-x gluon, F_2^c, which constrains the large-x gluon, and the dimuon differential cross-section $d^2\sigma^{\nu,c}/dxdy$, which constrains the strangeness. These are only applied at the initial parameterisation scale, as for higher scales DGLAP evolution will maintain positivity if it is present at a lower scale.

3.2.7 NNPDF2.3

NNPDF2.3 is a full PDF determination at NLO and NNLO based on a global dataset and the methodology described above. It builds on the previous NNPDF2.0 and NNPDF2.1 analyses [4, 56] but with the inclusion of several early LHC datasets from ATLAS, CMS and LHCb. It also features several methodological improvements over the previous determinations, made possible by the performance increases from wider use of the FastKernel method. Since the release of the original paper [45], a additional NNPDF2.3 LO determination has been performed, and an extension of the NLO set to include QED corrections with a determination of the photon PDF has been released [40].

Features

The main feature of NNPDF2.3 is the inclusion of data from the LHC. In total four LHC datasets were added: ATLAS inclusive jet cross-sections [60] and W and Z lepton rapidity distributions [58], CMS W electron asymmetry [57], and LHCb W lepton asymmetry [59]. All of the new measurements are based on the 2010 7TeV run which saw an integrated luminosity of 36pb^{-1}, except for the CMS dataset which was based on the more substantial 2011 840pb^{-1} run. Details of the number of data points, kinematic coverage and average statistical and systematic uncertainties are given in Table 3.1. A total of 141 LHC data points were included in the NNPDF2.3 fits. While there were several other LHC results sensitive to PDFs which had been released at the time that the NNPDF2.3 analysis was performed (including some I will discuss in Chapter 4, included in the NNPDF3.0 dataset), only these four were available with full covariance matrix.

These new data were added to the existing dataset used in the NNPDF2.1 analysis [56]: fixed-target DIS data from NMC [61, 62], BCDMS [63, 64] and SLAC [65]; the combined HERA-I DIS dataset [66], HERA F_L [67] and the separated ZEUS and H1 F_2^c structure function data [68–74], some ZEUS HERA-II DIS cross-sections [75,76]; CHORUS inclusive neutrino DIS [77], and NuTeV dimuon production data [78,79]; fixed-target E605 [80] and E866 [81–83] Drell-Yan production data; CDF W asymmetry [84] and CDF [85] and D0 [86] Z rapidity distributions; CDF [87] and D0 [88] Run-II one-jet inclusive cross-sections.

In addition to the new data, NNPDF2.3 boasted several methodological improvements over the NNPDF2.1 analysis. The FastKernel method (described in Section 3.2.5, above) was introduced for hadronic processes in addition to DIS, which resulted in a substantial reduction in the time required to perform a PDF fit. This was especially important given that all of the new LHC data was for hadronic processes. This performance upgrade was traded against improvements in the effectiveness of the genetic algorithm minimisation. The parameters of the NLO fit were modified in line with the NNLO fit in order to search the parameter space more effectively, providing a larger number of mutants and mutations each generation. The length of the fits were also increased from 30000 to 50000 generations, providing the genetic algorithm with more time to find the global minimum. This increase in length was combined with a more stringent cross-validation stopping condition, in order to prevent the fit from stopping prematurely. We additionally added a post-fit check of the quality of fit to ensure that outlying replicas, possibly caused by poor starting conditions or fluctuations in validation χ^2, are not included in the final set. If the final χ^2 of a replica is more than four sigma higher than the average χ^2 it is replaced.

Results

Table 3.2.7 shows the χ^2 obtained for each of the different datasets used in the NNPDF2.3 fits for the NLO and NNLO sets of NNPDF2.1 and multiple NNPDF2.3 analyses. The first NNPDF2.3 column (‘Global’) provides the results for the central determinations using the full global dataset, while the second column (‘noLHC’) gives

		NNPDF2.1		**NNPDF2.3**					
		Global		Global		noLHC		Collider	
Experiment	N_{dat}	NLO	NNLO	NLO	NNLO	NLO	NNLO	NLO	NNLO
Total	3482(3501)	1.145	1.167	1.121	1.153	1.101	1.147	1.018	1.034
NMC-pd	132	0.97	0.93	0.93	0.94	0.93	0.94	[4.72]	[5.03]
NMC	224	1.68	1.58	1.61	1.57	1.59	1.56	[1.86]	[1.87]
SLAC	74	1.34	1.04	1.26	1.02	1.28	1.04	[1.80]	[1.48]
BCDMS	581	1.21	1.29	1.19	1.29	1.20	1.28	[1.81]	[2.08]
CHORUS	862	1.10	1.08	1.10	1.06	1.09	1.07	[1.93]	[1.81]
NTVDMN	79	0.70	0.50	0.45	0.55	0.42	0.48	[28.51]	[22.61]
HERAI-AV	592	1.04	1.04	1.00	1.01	1.01	1.03	0.97	0.98
FLH108	8	1.34	1.23	1.28	1.20	1.29	1.21	1.33	1.25
ZEUS-H2	127	1.21	1.21	1.20	1.22	1.20	1.22	1.30	1.32
ZEUS F_2^c	50(62)	0.75	0.81	0.82	0.90	0.81	0.86	0.73	0.77
H1 F_2^c	38(45)	1.50	1.44	1.58	1.52	1.58	1.49	1.34	1.30
DYE605	119	0.94	1.09	0.88	1.02	0.85	1.07	[11.12]	[4.56]
DYE886	199	1.42	1.76	1.28	1.62	1.24	1.61	[4.44]	[4.63]
CDF W asy	13	1.87	1.63	1.54	1.70	1.45	1.66	1.17	1.16
CDF Z rap	29	1.77	2.42	1.79	2.12	1.77	2.15	1.49	1.49
D0 Z rap	28	0.57	0.68	0.57	0.63	0.57	0.64	0.57	0.61
ATLAS W,Z	30	[1.58]	[2.22]	1.27	1.46	[1.37]	[1.94]	1.08	1.08
CMS W e asy	11	[2.26]	[1.45]	1.04	0.96	[1.50]	[1.37]	0.96	0.96
LHCb W,Z	10	[1.34]	[1.42]	1.21	1.22	[1.24]	[1.33]	1.22	1.29
CDF RII k_T	76	0.68	0.65	0.61	0.67	0.60	0.67	0.57	0.59
D0 RII cone	110	0.90	0.98	0.84	0.93	0.84	0.94	0.83	0.93
ATLAS jets	90	[1.65]	[1.48]	1.55	1.42	[1.57]	[1.45]	1.46	1.41

Table 3.2: χ^2 values for the different datasets included in the NNPDF2.3 analysis. The results are given for central NLO and NNLO NNPDF2.1 and 2.3 determinations, and for fits to a pair of reduced dataset. Where N_{dat} is different at NNLO it is shown in brackets.

values for a fit without the new LHC data and the third ('Collider') values for a fit only including data from colliders (i.e. data from HERA, the Tevatron and the LHC). Where a fit does not include a particular dataset the χ^2 is provided in brackets.

Some interesting results are immediately visible just looking at the χ^2 in Table 3.2.7 alone. Comparing the columns for the NNPDF2.1 and NNPDF2.3 noLHC sets tests the impact of the improved methodology, as both of these fits were performed using the same dataset. The newer methodology obtains a better total χ^2 and also significantly improved χ^2 for several individual datasets. The noLHC fit obtains a reasonably good description of the LHC data, indicating that there is little evidence of tension between the new LHC data and the existing dataset, though the Global fit description is still slightly better as would be expected. On the other hand, in the Collider-only fit, several of the excluded datasets are very poorly described, indicating that these datasets contain information which is lost by excluding them in the fit.

In general we find that the new LHC data have a small but noticeable impact on several PDFs. Fig. 3.3 compares several PDFs from the central NNPDF2.3 set with their counterparts from the noLHC fit. This shows the impact of including the new LHC datasets on these PDFs (in green) compared to leaving them out (in red). For the Singlet distribution (top left), there is an upward shift at small x of about 0.5σ

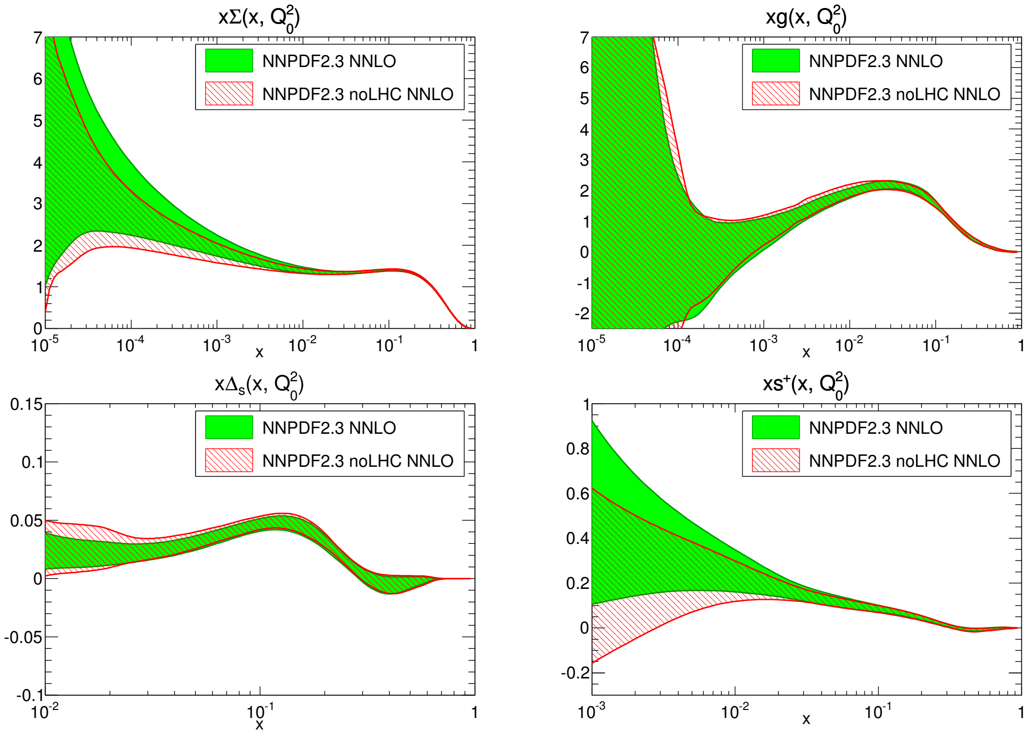


Figure 3.3: Comparison of the Singlet, gluon, sea asymmetry and strangeness PDFs from the central and noLHC NNLO NNPDF2.3 sets at the initial fitting scale ($Q^2 = 1GeV^2$). The shaded area displays the one-sigma contour.

due to these data, while the gluon is largely unaffected. There are also effects in the quark sea sector, with a matching upward shift in the strangeness and a reduction of uncertainties in the sea asymmetry, again at small x.

Fig. 3.4 presents a similar comparison for the NNLO collider-only NNPDF2.3 fit. The difference between this fit and the global fit is again the absence of datasets from the fit, in this case all of the fixed target data leaving just the data from colliders. The singlet, gluon, valence and sea asymmetry are shown for both this reduced dataset fit and the global fit. The singlet and gluon are reasonably well constrained by the collider data, though there are some significant deviations at medium x from the global fit results. On the other hand, for the valence and sea asymmetry the collider-only description is markedly poorer, resulting in a much larger uncertainty for the large x valence and an essentially featureless Δ_S. This demonstrates that a collider only dataset is not yet sufficient to properly constrain all PDFs, though as more LHC data is collected this may change.

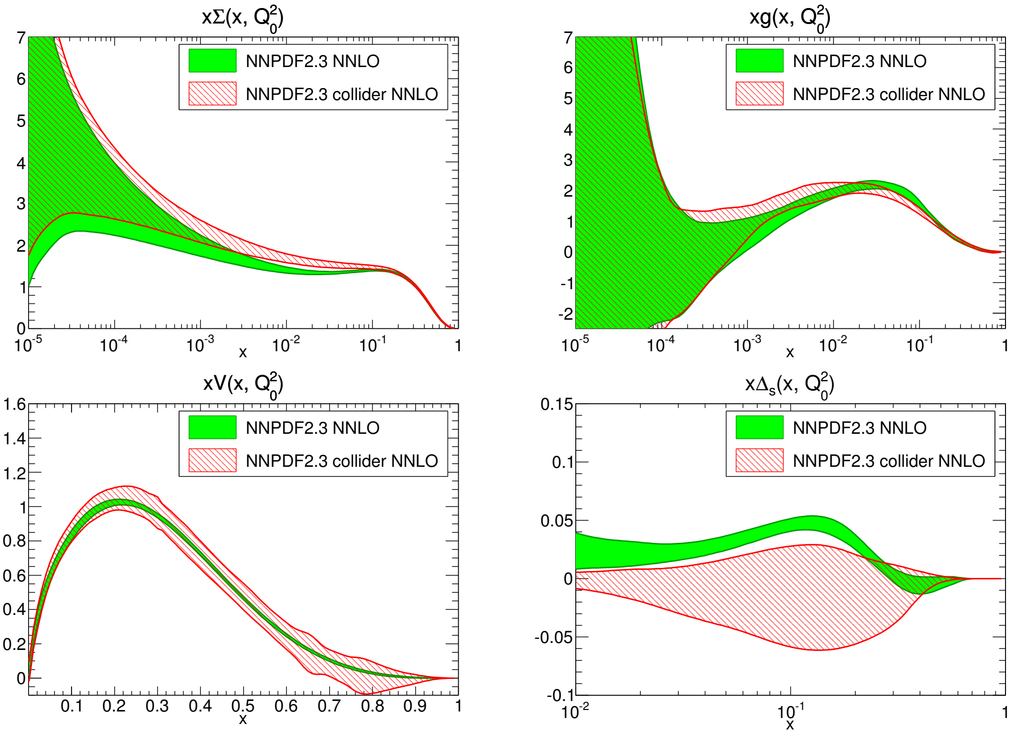


Figure 3.4: Comparison of the singlet, gluon, valence and sea asymmetry PDFs from the central and collider-only NNLO NNPDF2.3 fits at the initial scale ($Q^2 = 1GeV^2$). The valence is plotted on a linear scale in x, while the rest are plotted on a log scale.

Chapter 4

NNPDF3.0 dataset

In this section I will look at the dataset used for the NNPDF3.0 fits [6], focusing particularly on the changes from that used in the NNPDF2.3 analysis. I will first discuss in detail each of the new experimental datasets included in the fit. I will also cover several issues of the theoretical treatment of data: the computational tools used to implement perturbative corrections, NNLO QCD corrections to jet production, electroweak corrections, and the treatment of heavy quark mass effects. Finally, I will look at details of the implementation of systematic experimental uncertainties in the fit.

4.1 Experimental data

In addition to the datasets included in the NNPDF2.3 fits, described in Section 3.2.7, a large amount of new experimental data has been added to the NNPDF3.0 fits. In this section I will describe in detail the new datasets; information about the previously included dataset can be found in the NNPDF papers [3, 4, 45].

Details of the datasets included in the NNPDF3.0 fit are provided in Table 4.1. For each dataset I have provided the corresponding published reference, the availability and treatment of systematics (further discussed in Section 4.3.2 below), the number of data points before and after cuts at NLO and NNLO (again discussed later in this chapter), and the kinematic coverage of each dataset. Information on sets removed from NNPDF3.0 is also given in Table 4.2.

The kinematical coverage of the NNPDF3.0 dataset in the $\left(x, Q^2\right)$ plane is shown in the scatter plot Fig. 4.1 (note that for hadronic data, leading-order kinematics have been assumed for illustrative purposes, as discussed in [4]).

In NNPDF3.0 we have supplemented the combined HERA-I dataset with the inclusion of all the relevant HERA-II inclusive cross-sections measurements from H1

Experiment	Dataset	Ref.	Sys. Unc.		$N_{\mathbf{dat}}$ (cut)	Kinematics
NMC	NMC d/p	[61]	add		289 (132)	$3.5 \times 10^{-3} \leq x \leq 0.47$ $0.8 \leq Q^2 \leq 61.2$ GeV^2
	NMC $\sigma^{\mathrm{NC},\mathrm{p}}$	[62]	add		211 (224)	$1.5 \times 10^{-3} \leq x \leq 0.68$ $0.2 \leq Q^2 \leq 99$ GeV^2
SLAC	SLAC p	[65]	add	a	191 (37)	$0.07 \leq x \leq 0.85$ $0.58 \leq Q^2 \leq 29.2$ GeV^2
	SLAC d	[65]	add	a	191 (37)	$0.07 \leq x \leq 0.85$ $0.58 \leq Q^2 \leq 29.1$ GeV^2
BCDMS	BCDMS p	[63]	add	b	351 (333)	$0.07 \leq x \leq 0.75$ $7.5 \leq Q^2 \leq 230$ GeV^2
	BCDMS d	[64]	add	b	254 (248)	$0.07 \leq x \leq 0.75$ $8.8 \leq Q^2 \leq 230$ GeV^2
CHORUS	CHORUS ν	[77]	add	c	572 (431)	$0.02 \leq x \leq 0.65$ $0.3 \leq Q^2 \leq 95.2$ GeV^2
	CHORUS $\bar{\nu}$	[77]	add	c	572 (431)	$0.02 \leq x \leq 0.65$ $0.3 \leq Q^2 \leq 95.2$ GeV^2
NuTeV	NuTeV ν	[78, 79]	add		45 (41)	$0.027 \leq x \leq 0.36$ $1.1 \leq Q^2 \leq 116.5$ GeV^2
	NuTeV $\bar{\nu}$	[78, 79]	add		44 (38)	$0.021 \leq x \leq 0.25$ $0.8 \leq Q^2 \leq 68.3$ GeV^2
HERA-I	NC e^+	[66]	mult	d	434 (379)	$6.2 \times 10^{-7} \leq x \leq 0.65$ $0.045 \leq Q^2 \leq 3 \times 10^4$ GeV^2
	NC e^-	[66]	mult	d	145 (145)	$1.3 \times 10^{-3} \leq x \leq 0.65$ $90 \leq Q^2 \leq 3 \times 10^4$ GeV^2
	CC e^+	[66]	mult	d	34 (34)	$8 \times 10^{-3} \leq x \leq 0.40$ $300 \leq Q^2 \leq 1.5 \times 10^4$ GeV^2
	CC e^-	[66]	mult	d	34 (34)	$0.013 \leq x \leq 0.40$ $300 \leq Q^2 \leq 3 \times 10^4$ GeV^2
HERA-II ZEUS	NC e^-	[75]	mult	e	90 (90)	$5 \times 10^{-3} \leq x \leq 0.65$ $200 \leq Q^2 \leq 3 \times 10^4$ GeV^2
	CC e^-	[76]	mult	e	37 (37)	$0.015 \leq x \leq 0.65$ $280 \leq Q^2 \leq 3 \times 10^4$ GeV^2
	NC e^+	[89]	mult	f	90 (90)	$5 \times 10^{-3} \leq x \leq 0.40$ $200 \leq Q^2 \leq 3 \times 10^4$ GeV^2
	CC e^+	[90]	mult	f	35 (35)	$7.8 \times 10^{-3} \leq x \leq 0.42$ $280 \leq Q^2 \leq 3 \times 10^4$ GeV^2
HERA-II H1	NC e^-	[91]	mult	g	139 (139)	$2 \times 10^{-3} \leq x \leq 0.65$ $120 \leq Q^2 \leq 4 \times 10^4$ GeV^2
	NC e^+	[91]	mult	g	138 (138)	$2 \times 10^{-3} \leq x \leq 0.65$ $120 \leq Q^2 \leq 4 \times 10^4$ GeV^2
	CC e^-	[91]	mult	g	29 (29)	$8 \times 10^{-3} \leq x \leq 0.40$ $300 \leq Q^2 \leq 3 \times 10^4$ GeV^2
	CC e^+	[91]	mult	g	29 (29)	$8 \times 10^{-3} \leq x \leq 0.40$ $300 \leq Q^2 \leq 3 \times 10^4$ GeV^2
	low Q^2	[92]	mult		136 (124)	$2.8 \times 10^{-5} \leq x \leq 0.015$ $1.5 \leq Q^2 \leq 90$ GeV^2
	high y	[92]	mult		55 (52)	$2.9 \times 10^{-5} \leq x \leq 5 \times 10^{-3}$ $2.5 \leq Q^2 \leq 90$ GeV^2
HERA σ^c_{NC}	σ^c_{NC}	[93]	mult		52 (47)	$3 \times 10^{-5} \leq x \leq 0.05$ $2.5 \leq Q^2 \leq 2 \times 10^3$ GeV^2

Experiment	Dataset	Ref.	Sys. Unc.		$N_{\mathbf{dat}}$ (cuts)	Kinematics
E866	DY d/p	[83]	mult		15 (15)	$0.017 \leq x \leq 0.87$ $19.8 \leq Q^2 \leq 251$ GeV^2
	DY p	[81, 82]	mult		184 (184)	$0.025 \leq x \leq 0.56$ $21.2 \leq Q^2 \leq 166$ GeV^2
E605	DY	[80]	mult		119 (119)	$0.14 \leq x \leq 0.65$ $50.5 \leq Q^2 \leq 286$ GeV^2
CDF	Z rapidity	[85]	mult	h	29 (29)	$2.9 \times 10^{-3} \leq x \leq 0.80$ $M^2 = 8315$ GeV^2
	Run-II k_t jets	[94]	mult	h	76 (76/52)	$4.6 \times 10^{-3} \leq x \leq 0.90$ $3364 \leq p_T^2 \leq 3.7 \times 10^5$ GeV^2
D0	Z rapidity	[86]	mult		28 (28)	$2.9 \times 10^{-3} \leq x \leq 0.72$ $M^2 = 8315$ GeV^2
ATLAS	W, Z 2010	[58]	mult	i	30 (30)	$0 \leq \lvert\eta_l\rvert \leq 2.5$ $M^2 = 6464,\ 8315$ GeV^2
	7 TeV jets 2010	[60]	mult	i,j	90 (90/9)	$0 \leq \lvert y\rvert \leq 4.4$ $400 \leq p_T^2 \leq 2.3 \times 10^6$ GeV^2
	2.76 TeV jets	[95]	mult	j	59 (59/3)	$0 \leq \lvert\eta^{\text{jet}}\rvert \leq 4.4$ $400 \leq p_T^2 \leq 4 \times 10^4$ GeV^2
	high-mass DY	[96]	mult		11 (5)	$116 \leq M_{ll} \leq 1500$ GeV
	W p_T	[97]	mult		11 (9/0)	$0 \leq p_T^W \leq 300$ GeV
	$\sigma(t\bar{t})$	[98, 99] [100]	mult		3 (3)	
CMS	W e asym.	[57]	mult		11 (11)	$0 \leq \lvert\eta_l\rvert \leq 2.4$
	W μ asym.	[101]	mult		11 (11)	$0 \leq \lvert\eta_l\rvert \leq 2.4$
	7 TeV jets 2011	[102]	mult		133 (133/83)	$0 \leq \lvert\eta\rvert \leq 2.5$ $114 \leq p_T^{\text{jet}} \leq 2116$ GeV
	$W + c$ total	[103]	mult		5 (5)	$0 \leq \lvert\eta_l\rvert \leq 2.1$
	$W + c$ ratio	[103]	mult		5 (5)	$0 \leq \lvert\eta_l\rvert \leq 2.1$
	2D DY 2011	[104]	mult		124 (88/110)	$20 \leq M_{ll} \leq 1200$ GeV $0 \leq \lvert\eta_{ll}\rvert \leq 2.4$
	$\sigma(t\bar{t})$	[105, 106] [107]	mult		3 (3)	
LHCb	W rapidity	[59]	mult		10 (10)	$2.0 \leq \eta_l \leq 4.5$
	Z rapidity	[108]	mult		9 (9)	$2.0 \leq \eta_l \leq 4.5$

Table 4.1: Experimental data included in the NNPDF3.0 global fits. The data is separated by experiment and dataset as they are in the code, with DIS data on the first page and hadronic data here. For each dataset the table also gives: a reference; whether the systematics are treated multiplicatively or additively; which sets it is cross-correlated with; how many datapoints it contains before and after cuts (at NLO/NNLO); and details of its kinematic coverage.

Experiment	Dataset	Ref.	$N_{\mathbf{dat}}$	Details
H1	F_2^c 2001	[72]	12	Superseded by combination
	F_2^c 2009	[73]	6	
	F_2^c 2010	[74]	26	
ZEUS	F_2^c 1999	[68]	21	Superseded by combination
	F_2^c 2003	[69]	31	
	F_2^c 2008	[70]	9	
	F_2^c 2009	[71]	8	
CDF	W asymmetry	[84]	13	Lepton-level data available from LHC
D0	Run II cone jets	[88]	110	Infrared unsafe at NNLO

Table 4.2: Experiments that were present in NNPDF2.3 but that have been excluded from NNPDF3.0. The last column provides a brief description of why each set was removed; consult the text for more information.

and ZEUS [89–92]. These data provide an improvement in the statistical and systematic precision at medium- and high-Q^2 over the HERA-I data, and thus provide valuable information on the quarks at medium and large x. We have also included low-Q^2 data from H1 that provides additional information on the small-x gluon.

From the H1 experiment, we have included the new high-Q^2 data from the HERA-II run [91], which covers the large Q^2 region $60 \leq Q^2 \leq 5\,10^4$ GeV^2, and which has improved statistical and systematic precision in comparison to Run-I. These data, taken at the default proton beam energy of $E_p = 920$ GeV used in most of the HERA-II run, have been supplemented with inclusive cross-section measurements performed at lower centre-of-mass energies [92], obtained with proton beam energies of $E_p = 575$ GeV and $E_p = 460$ GeV. These lower-energy measurements are the same ones used to determine the longitudinal structure function F_L in a dataset we had previously included in our fits. Therefore, we exclude the F_L [67] dataset from the present fit to avoid any double counting, and for the same reason we have not included any of the updated F_L extractions from HERA [109,110]. For completeness, in NNPDF3.0 we also include the high-inelasticity data that H1 extracted from their Run II measurements [92].

From the ZEUS experiment, in NNPDF2.3 we already included some of their HERA-II data for neutral- and charged-current DIS with an electron beam [75, 76]. In NNPDF3.0, we now also include neutral- and charged-current cross-sections with a positron beam [90, 111], which have since been published. As in the case of H1, ZEUS Run II inclusive cross-sections exhibit reduced statistical and systematic uncertainties in the medium- and large-Q^2 region, when compared to Run I data. For both H1 and ZEUS, we use the data averaged over lepton beam polarizations.

It is worth noting that the separate H1 and ZEUS inclusive measurements included in NNPDF3.0 have recently been supplanted by the final combined HERA dataset [112]. This set will be included in future NNPDF releases, however we expect this replacement

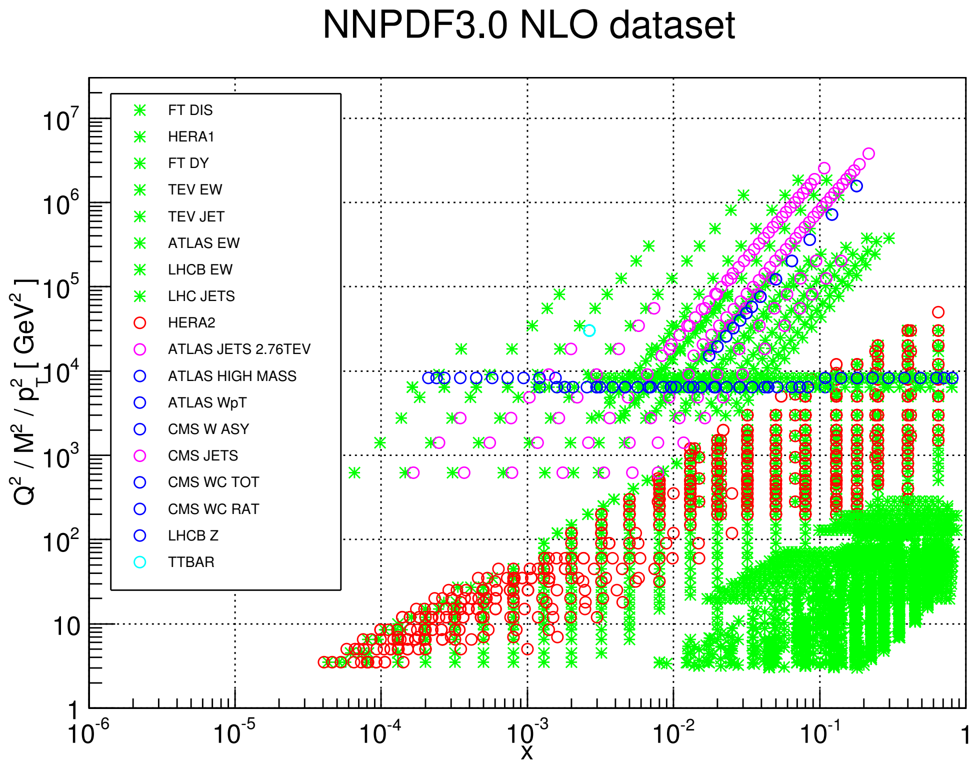


Figure 4.1: The kinematical coverage in the (x, Q^2) plane of the NNPDF3.0 dataset. For hadronic data, leading-order kinematics have been assumed for illustrative purposes. The green stars mark the data already included in NNPDF2.3, while the different coloured circles correspond to experiments that are new in NNPDF3.0.

to have a small impact on the PDFs, as the neural network fit effectively performs a dataset combination itself, something which we was partially demonstrated for the combined HERA-I dataset in the NNPDF2.0 analysis [4].

Turning to semi-inclusive measurements, in NNPDF3.0 we have replaced the separate charm structure function data F_2^c from the H1 and ZEUS experiments [68–74] with the combined HERA charm production dataset [93], which provides data for the reduced cross-section (rather than structure function), and is based on a more extensive dataset. Furthermore, cross-calibration between common systematics means that the combined data is more accurate that the separate inclusion of the individual measurements. The combined HERA charm production cross-sections offer a handle on the small-x gluon [113], provide a unique testing ground for different treatments of heavy quark mass effects, and allow one to extract the running charm quark mass $m_c(m_c)$ with competitive uncertainties [114, 115].

Turning now to the new LHC data, we have added a large amount of new LHC vector boson production data, supplementing the existing vector boson data included in

NNPDF2.3. From the ATLAS experiment, we include high-mass Drell-Yan production data from the 2011 run [96], based on an integrated luminosity of 4.9 fb^{-1}. These data are presented in terms of the invariant mass of the electron pairs produced at an invariant mass larger than the Z peak, extending to $M_{ll} = 1.5$ TeV, and can be used to constrain the large-x antiquarks. In addition, it was shown in Ref. [40] that high-mass Drell-Yan at the LHC provides an important constraint on the photon PDFs of the proton, and was indeed there used in the construction of the NNPDF2.3QED PDF set. We now also include the ATLAS measurement of the W boson transverse momentum distribution from the 2010 run of the LHC at $\sqrt{s}$ =7 TeV [97], corresponding to an integrated luminosity of 31 pb^{-1}. This data has the potential to constrain the gluon and the light quark distributions in the medium-x region [116]. There is also a 7 TeV ATLAS measurement of the Z boson transverse momentum distribution [117], however this data is not provided with all information on correlated uncertainties and so is not included in our fit.

From the CMS experiment, we include the W muon asymmetry data based on the full statistics (5 fb^{-1}) of the 7 TeV run [101], and the double-differential distributions for Drell-Yan production for dilepton masses in the range $20 \leq M_{ll} \leq 1500$ GeV, in bins of the dilepton invariant mass and rapidity, from the full 2011 dataset [104]. For the first time, we include CMS data for the production of charm quarks associated to W bosons [103], which provides important information on strangeness [118, 119]. The measurement is included as both absolute cross-sections, differential on the lepton rapidity from the W decay η_l, and as cross-section ratios W^+/W^-, also binned in η_l. The former constrains the shape and overall normalisation of the total strangeness $s + \bar{s}$ at $Q \sim M_W$ and the latter offer some handle on the strangeness asymmetry in the proton, $s - \bar{s}$. Data for the same process are available from the ATLAS Collaboration [120], but are given at the hadron level and thus cannot be directly included in our fit. It is possible that this data will be included in future fits by for example estimating a hadron-to-parton correction factor using MadGraph5_aMC@NLO.

We also include the LHCb $Z \to ee$ rapidity distributions from the 2011 dataset [108], which are more precise than the previous data from the 2010 run. The forward kinematics of this data provide constraints on PDFs at both smaller and larger values of x than the ATLAS and CMS vector boson data. Further LHCb data from the 2011 run for Z boson rapidity distributions in the $\mu\mu$ channel [121] and for low mass Drell-Yan production [122] were unavailable for NNPDF3.0 but will likely feature in future determinations.

Concerning inclusive jet production from ATLAS and CMS, we include the CMS inclusive jet production measurement at 7 TeV from the full 5 fb^{-1} dataset [102], which is provided with the full experimental covariance matrix, and which supersedes previous inclusive jet measurements from CMS [123]. This data has a large kinematical coverage, extending for instance in the central rapidity region up to jet transverse momenta of more than 2 TeV, and thus constraining the large-x quark and gluon PDFs [124, 125]. From ATLAS, we include the new inclusive cross-section measurement at $\sqrt{s} = 2.76$ TeV [95], which is provided with the full correlation matrix and with correlations to the corresponding $\sqrt{s} = 7$ TeV measurement. Including correlated measurements of jet cross-sections at two different centre of mass energies in this way enhances the impact of the data on PDFs as the experimental (particularly jet energy scale) systematic uncertainties are effectively reduced [126]. On the other hand, no LHC dijet data are included [127], since it is at present very difficult to achieve a good theoretical description of these measurements [124].

Finally, we include six independent measurements of the total top quark pair production cross-section from ATLAS and CMS, both at 7 TeV and at 8 TeV. These data provide information on the large-x gluon PDF, complementary to that provided by inclusive jet production. At 7 TeV we include the measurements in the dilepton channel, based on 0.70 fb^{-1} integrated luminosity from ATLAS [98] and on 2.3 fb^{-1} from CMS [105], and also the measurements performed using lepton+jets events from ATLAS [99] and CMS [106]. At 8 TeV we have included the dilepton channel measurement corresponding to an integrated luminosity of 2.4 fb^{-1} by CMS [128] and the ATLAS analysis of the lepton+jets final state in a dataset corresponding to an integrated luminosity of 5.8 fb^{-1} [100]. In the future top quark rapidity distribution data will also be included, but the theory for the this process was not available in a fast enough format for the release of NNPDF3.0.

In comparison to the NNPDF2.3 dataset, we have removed the Tevatron D0 Run II inclusive jet cross-section measurements [88], which were obtained with the infrared unsafe [129] midpoint algorithm and therefore are incompatible with NNLO calculations. On the other hand, the equivalent CDF set are retained in NNPDF3.0, as these are based on the k_t algorithm. We have also removed the CDF Tevatron W asymmetry data [84], since we now include cleaner and more precise data from the LHC (based on leptons rather than on the reconstructed W) which more than covers the same region in x, and since updated Tevatron W asymmetry data has recently been released [130, 131], which again will be included in future NNPDF fits.

The NNPDF3.0 dataset is therefore composed of essentially all of the relevant, high-quality data which was available at the time of the fits with full information on systematic correlations. There are a number of important datasets which have been released since, for instance the HERA combined data [112] and Tevatron legacy vector boson measurements mentioned above, as well as a large amount of remaining LHC run-I data. These sets will likely be included in an updated NNPDF analysis in the near future.

4.2 Theoretical treatment

As in previous determinations, the NNPDF3.0 PDFs are provided at LO, NLO and NNLO corresponding to fits performed using theoretical calculations at these orders of perturbative QCD. While for most of the observables included in the fit NNLO QCD corrections are known, some observables are known only up to NLO, while for others only partial contributions to the full NNLO corrections have been calculated. Specifically, NNLO corrections are totally unavailable for two processes included in the NNPDF3.0 fits: the vector boson transverse momentum distribution and the $W + c$ rapidity distribution. For the jet inclusive cross-section, only the gg-channel is fully available at NNLO, having been recently computed [132, 133], while for the full cross-section there is only an approximate NNLO prediction based on threshold resummation [134, 135]. For all other observables included in the fit the cross-sections are known up to NNLO.

The theoretical predictions for DIS observables have been implemented in the FastKernel framework and thoroughly benchmarked [5, 19]. Drell-Yan cross-sections, both for fixed target and for collider experiments, are computed at NNLO during the fit using special local C-factors computed according to the procedure described in Ref. [5], defined at the ratio of NNLO to NLO calculations but using fixed NNLO PDFs, that is

$$C^{\rm nnlo} \equiv \frac{\hat{\sigma}^{\rm nnlo} \otimes \mathcal{L}^{\rm nnlo}}{\hat{\sigma}^{\rm nlo} \otimes \mathcal{L}^{\rm nnlo}} \ , \tag{4.1}$$

where $\hat{\sigma}$ is the partonic cross-section computed at either NNLO or NLO accuracy, and $\mathcal{L}^{\rm nnlo}$ is the corresponding parton luminosity computed with a reference set of NNLO parton distributions.

Given that electroweak corrections can be relevant in the large invariant mass region covered by some of the experimental data included in our fit [136] we provide EW corrections for all LHC vector boson production data. To include these corrections in our NLO and NNLO calculation, we compute additional factors, $C^{\rm ew}$, defined analogously to Eq. (4.1), with the NNLO computation substituted by the NLO+EW one, and using NLO parton luminosities on both numerator and denominator; details on their computation and implementation are provided in Section 4.2.2.

For NNPDF3.0 effects of all-order perturbative resummation of QCD corrections are not included. These will be the object of a future separate study, likely leading to the construction of dedicated resummed sets. We do not include nuclear corrections, which are relevant for fixed-target deuterium DIS data, neutrino DIS data, and fixed-target Drell-Yan data. We will briefly assess the impact of this omission in Section 7.1.4.

4.2.1 Computational tools

The inclusion of hadronic processes in our PDF fits requires fast computation of the relevant theoretical cross-sections. Several fast interfaces have been developed to this purpose, including APPLgrid [137], which in turn provides an interface to MCFM [138, 139] and NLOjet++ [140], and FastNLO [141, 142], which can also be interfaced to NLOjet++. The MCgrid [143] package connects the Rivet [144] analysis package to APPLgrid, making use of the BlackHat/Sherpa [145] prescription for NLO reweighting. Recently, a new fast interface has become available, namely aMCfast [146], which is interfaced to MadGraph5_aMC@NLO [147], allowing matching to parton shower simulations.

Such tools have been used extensively in the present analysis. For the 7 TeV CMS jet data, we have used the FastNLO calculation with central scales $\mu_F = \mu_R = p_T^{\mathrm{jet}}$, while for the 2.76 ATLAS jet data, we have instead used an APPLgrid calculation. For consistency, we use exactly the same settings for the calculations, including the central scales, that were used for the corresponding ATLAS 7 TeV inclusive jet analysis. The CDF Run II k_t jets have also been computed using the FastNLO calculation again with compatible settings.

For all of the electroweak vector boson production data we have used the APPLgrid code interfaced to MCFM6.6, with a consistent choice of electroweak parameters. We use the G_μ scheme, with $M_Z = 91.1876$ GeV, $M_W = 80.398$ GeV and $G_F = 1.16637 \cdot 10^{-5}$ GeV^{-2} as input parameters and with α_e, $\sin\theta_W$ derived from those, and the Narrow-Width approximation turned off. For all rapidity distributions, we set $\mu_F = \mu_R = M_V$, with $V = W, Z$. For the W p_T distribution we set $\mu_F = \mu_R = M_W$, while in the case of CMS double-differential distribution we set the scales to the central value of the invariant mass bin. The MCFM6.6 calculations have been cross-checked with independent calculations of DYNNLO [148–151] and FEWZ3.1 [152, 153] at NLO finding perfect agreement in all cases.

In the NNLO fits, the NNLO $C-$factors defined in Eq. (4.1) have been computed with FEWZ3.1 and cross-checked against DYNNLO1.3. In order to achieve negligible integration errors in all data bins it was necessary to perform very high statistics runs. The C-factors were then smoothed with a high-degree polynomial interpolation, making sure that the difference between smoothed and original NNLO predictions was always within the Monte Carlo uncertainty of the code used to compute it. The NNLO QCD corrections are in several cases quite sizeable, especially for small invariant masses of

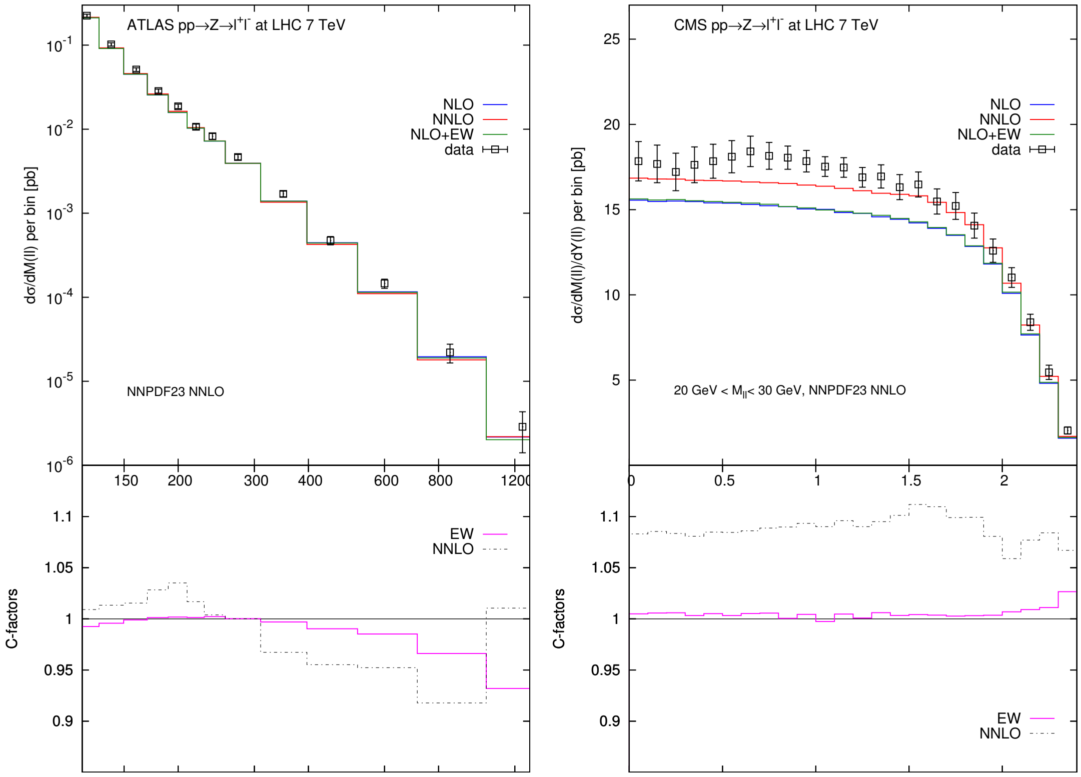


Figure 4.2: The NLO, NNLO and NLO+EW predictions compared to the ATLAS high-mass Drell-Yan distribution data as a function of the invariant mass of the dilepton system M_{ll} (left) and the CMS double-differential Drell-Yan distribution as a function of the rapidity of the lepton pair in the lowest invariant mass bin, with $20 \leq M_{ll} \leq 30$ GeV. The three curves displayed have been computed with FEWZ3.1 with the same input PDF set, namely NNPDF2.3 with $n_f = 5$ and $\alpha_s(M_Z) = 0.118$.

the produced lepton pairs. This can be seen from the right-hand plot in Fig. 4.2, where the NNLO C-factor for the CMS double-differential Drell-Yan data at low M_{ll} is around 10%, independent of the dilepton rapidity. NNLO corrections are also important for the ATLAS high-mass Drell-Yan data, again reaching almost 10% around $M_{ll} \sim 1$ TeV, as the left plot of Fig. 4.2 shows.

As previously mentioned, NNLO corrections to jet production in the gg-channel have become recently become available. [132, 133]. While the calculation of the full correction has yet to be obtained, this incomplete result can be used to gauge the accuracy of the approximate NNLO prediction based on threshold resummation which was presented in [134]. This was done recently in a systematic study [154], which found reasonable agreement between the two calculations in the high-p_T and central-y regions, with greater discrepancies going to large rapidities and small transverse momentum. In NNPDF3.0, we follow the strategy of Ref. [154] and compute approximate NNLO

C-factors, Eq. (4.1), using the threshold calculation, while restricting the fitted dataset to the region where we know the approximation to be reliable. This leads to the set of cuts outlined below in Section 4.3.

For the computation of top quark pair production data at NLO, we again used APPLgrid interfaced to MCFM6.6. The NNLO $C-$factors have been computed using the NNLO calculation of Ref. [155], as implemented in the top++ code [156]. Finally, we have used aMCfast interfaced to MadGraph5_aMC@NLO to compute the Higgs rapidity distributions in gluon fusion at NLO with an unphysical boson of mass $m_h = \sqrt{5}$ GeV. As explained in Section 5.6, this unphysical observable has been used to enforce the positivity of cross-sections that depend on the small-x gluon.

4.2.2 Electroweak corrections

Electroweak corrections, though generally small, may become large at high scales $Q^2 \gg M_V^2$. While this will certainly be an issue for future LHC data at higher centre of mass energy, already for some high-mass data included in NNPDF3.0 the high accuracy of the experimental measurements may require theoretical predictions at the percent level of precision, and so the size of the EW corrections here also needs to be carefully assessed.

The NLO EW one-loop corrections are known [157–162] and have been implemented in several public codes such as HORACE [157] and ZGRAD2 [161, 162]. In FEWZ3.1 [152, 153] the NLO EW corrections are combined to the NNLO QCD corrections using the complex mass scheme. This code allows the user to separate a gauge–invariant QED subset of the corrections from the full EW result, including initial–state QED radiation, final–state QED radiation (FSR) and the initial–final interference terms. Within the current uncertainties that affect the photon PDF, as determined in the NNPDF2.3QED analysis [40], initial–state QED corrections are compatible with zero for most of the data included in our fit, and we exclude the data from the fit for which they may be sizeable. For the final state radiation, we use data from which it has already subtracted where available, including all of the ATLAS and CMS vector boson production data.

We may thus consistently isolate and compute the weak component of the FEWZ EW corrections, and include it in our calculation via the computation of additional $C-$factors for all the electroweak gauge boson production. The size of the corrections for the ATLAS high-mass Drell-Yan and CMS double-differential Drell-Yan is displayed in Fig. 4.2. We find that the effect of the these corrections is negligible for most of the data in the Zpeak region, of the order 1% or below. For the CMS double differential distributions, in the smallest invariant mass bin, we find that the EW corrections are small, much smaller than the NNLO QCD corrections. At large invariant masses the EW corrections are rather large and negative, as expected from the results of Ref. [163]. This can be clearly seen from the ATLAS results in Fig. 4.2, where the EW corrections reach $\sim 7\%$ in the last bin of the distribution. Although the ATLAS high mass distribution is the only measurement for which we find that EW corrections are required, for consistency we include the corrections for all the Z/γ^* production data.

4.2.3 Treatment of heavy quarks

In NNPDF3.0, as in previous NNPDF analysis since NNPDF2.1, heavy quark structure functions have been computed using the FONLL general-mass variable-flavour-number (GM-VFN) scheme [164]. In this scheme, the massive fixed-order calculation (in which the heavy quark is only counted in the final state) and resummed calculation (in which the heavy quark is treated as a massless parton) are consistently matched order by order. There is some latitude in deciding at which order the fixed-order massive result is included, compared to the perturbative order at which parton evolution is treated. Specifically, in an NLO computation one may decide to include massive contributions to structure functions up to $\mathcal{O}\left(\alpha_s\right)$, (on the grounds that this is the order at which the massless structure functions are computed) or up to $\mathcal{O}\left(\alpha_s^2\right)$, (on the grounds that the massive structure function starts at $\mathcal{O}\left(\alpha_s\right)$). These approaches are called the FONLL-A and FONLL-B schemes respectively [164]. At NNLO, while in principle a similar ambiguity would exist, in practice the massive coefficient function can only be included up to $\mathcal{O}\left(\alpha_s^2\right)$ because the $\mathcal{O}\left(\alpha_s^3\right)$ massive result is not currently known (though there is progress in this direction, for instance in [165]). In analogy to the NLO schemes, this is called the FONLL-C scheme.

While the FONLL-A scheme was used for the NNPDF2.1 and NNPDF2.3 NLO PDF sets, we now adopt the FONLL-B scheme for NNPDF3.0, with FONLL-C (as before) used at NNLO. While this scheme is less systematic, in that when going to NLO to NNLO the massless computation goes up one order but the massive one does not, it has the advantage that massive terms at NLO are more accurate, thereby allowing for the inclusion of a somewhat wider set of data, as small-x and Q^2 charm production data are affected by large $\mathcal{O}\left(\alpha_s^2\right)$ corrections, and cannot be described accurately in the FONLL-A scheme. For the same reason, using FONLL-B allows for a more accurate description of the HERA inclusive cross-section data at small-x.

Heavy quark structure functions are computed using the expression which corresponds to the pole mass definition. In this paper, we use for the heavy quark pole masses,

$$m_c = 1.275 \text{ GeV}\,, \qquad m_b = 4.18 \text{ GeV}\,, \qquad m_t = 173.07 \text{ GeV}\,, \tag{4.2}$$

which correspond to the current PDG values for the $\overline{\text{MS}}$ masses. Note that these values differ from the ones used in NNPDF2.3, which are $m_c = \sqrt{2}$ GeV, $m_b = 4.75$ GeV and $m_t = 175$ GeV. A brief analysis of the effect of these changes is given in Section 7.1.4; a full investigation of heavy quark mass dependence will likely be the subject of future NNPDF work.

In NNPDF3.0 we use for our central sets the $n_f = 5$ scheme, in which the number of active flavours never exceeds $n_f = 5$ (i.e. in the fit the top quark is always treated as massive, never as a parton), though fits using $n_f = 3$, $n_f = 4$ and $n_f = 6$ schemes are also available. This is another difference from NNPDF2.3, which used the $n_f = 6$ as default. In previous determinations the distinction between $n_f = 5$ and $n_f = 6$ was relevant only for delivery, as no data used was above the top threshold, now several jet data (especially the 2011 CMS inclusive jets) are above top threshold, so the decision of which to use becomes more important. Close to top threshold use of an $n_f = 5$ scheme is advantageous because the top mass is treated exactly, while the loss of accuracy due to the fact that the $n_f = 5$ running of α_s differs from the exact $n_f = 6$ running [166] is a comparatively smaller effect, only being visible for processes which start at high order in α_s [167]. Furthermore, most of the codes which we used for NNLO computations (specifically NLOJET++ and FEWZ) use an $n_f = 5$ scheme, and the same is true for many of the codes and interfaces used in the computation of LHC processes. With an ever increasing set of LHC data, the use of an $n_f = 5$ both in fitting, and as a default for PDF delivery appear to be the better choice.

4.3 Construction of the dataset

4.3.1 Kinematic cuts

As in previous NNPDF fits, we apply a cut in Q^2 and W^2 to fixed-target DIS data, in order to avoid including data that is subject to large higher-twist corrections. The cuts used in all of the NNPDF3.0 fits are

$$Q^2 \geq 3.5 \text{ GeV}^2, \qquad W^2 \geq 12.5 \text{ GeV}^2\,. \tag{4.3}$$

The stability of the fit with respect to these choices (and in particular the explicit check that they eliminate the need for higher twists) was studied in detail in Ref. [168]. With the introduction of new HERA and LHC data, Low-scale DIS data carry less weight

Experiment	$N_{\rm dat}^{\rm cut}$	Inclusion regions in the (y, p_T) plane	
CDF Run-II k_t jets [94]	52	$1.1 < \lvert y\rvert < 1.6$	$224 \leq p_T \leq 298$ GeV
ATLAS 2.76 TeV jets [95]	3	$\lvert y\rvert < 0.3$	$p_T \geq 260$ GeV
ATLAS 7 TeV jets 2010 [60]	9	$\lvert y\rvert < 0.3$ $0.3 < \lvert y\rvert < 0.8$	$p_T \geq 400$ GeV $p_T \geq 800$ GeV
CMS jets 2011 [102]	83	$1.0 < \lvert y\rvert < 1.5$	$p_T \geq 272$ GeV

Table 4.3: Summary of the inclusion regions in jet transverse momentum p_T and rapidity $|y|$ used in the NNPDF3.0 NNLO fits for the inclusive jet production measurements. As explained in the text, these inclusion regions are determined from a cut-off in the relative difference between the exact and approximate threshold C-factors in the gluon-gluon channel [154]. $N_{\rm dat}$ in the second column is the number of experimental data points for these jet datasets that pass the selection cuts in the NNLO fits.

in our current fit than they did previously, so we expect that the impact of the precise value of the cuts is smaller than in this previous study. Note that all NNPDF fits include target-mass corrections, following the method of Ref. [3].

As discussed in Section 4.2 above, NNLO corrections are not available for the W p_T distribution or for $W + c$ production. Because of this, ATLAS W p_T distribution data are included only in the LO and NLO fits and are excluded from the NNLO fits. The CMS $W + c$ distribution data, on the other hand, is included in the NNLO dataset, with matrix elements computed up to NLO only (but α_s running at NNLO), in order to include the important constraints on the strange PDFs that these data provide.

For inclusive jet production, we include all available data in the NLO fit, while in the NNLO fit we include only the data points such that the relative difference between the exact and the approximate gg-channel NNLO C-factors differ by less than 10%, as described in the previous section. In Table 4.3 we summarise the resulting inclusion regions in the (p_T, y) plane and the number of data points points $N_{\rm dat}$ within the regions which survive the cut for each experiment.

For the ATLAS measurement of the W transverse momentum distribution, we include only those data points with $p_T^W > 25$ GeV. This cut excludes the first two bins in p_T, and is motivated by the observation that at small p_T the perturbative series is not well-behaved and all-order resummation is needed (either analytically or by matching the fixed order calculation to a parton shower).

For the neutral-current Drell-Yan measurements from ATLAS and CMS, we include only data for which the dilepton invariant mass satisfies $M_{ll} < 200$ GeV. This excludes the last six bins of the ATLAS DY invariant mass distribution, and the 12 points in the rapidity distribution corresponding to the last bin of invariant mass, $M_{ll} \in [200, 1500]$ GeV, for the CMS measurement. The reason behind this cut is that in these regions the photon-initiated contribution to the cross-section can be come sizeable (up to 20%), and this contribution is not included in the electroweak correction used in our fits. Including the photon–initiated contributions in the dilepton cross-section would require an initial photon PDF $\gamma(x, Q^2)$, which is not fitted in this analysis.

One final cut is imposed, in the NLO fit only, to the lowest invariant-mass bin of the CMS Drell-Yan double differential distributions. As can be seen from Fig. 4.2 (right plot), for the bin with invariant mass $20 \leq M_{ll} \leq 30$ GeV, the NNLO C–factors are large, around 10%, while experimental uncertainties are a few percent. It is clear that because of this it would be difficult to obtain a reasonable NLO fit to these data points, and therefore the 24 points of the $20 \leq M_{ll} \leq 30$ GeV bin are excluded from the NLO analysis.

The number of data points before and after cuts, both in the NLO and NNLO fit, are summarised for each dataset in Table 4.1. Collectively these cuts reduce the unaltered dataset of 5179 points to 4276 at NLO and 4078 at NNLO. At LO we use the same cuts as in the NLO fit, as we expect the theory uncertainties at LO to be much larger than the experimental uncertainties, so it does not make much sense to attempt to devise a set of optimised kinematical cuts specifically for the LO fits.

4.3.2 Treatment of correlated systematic uncertainties

The majority of experimental datasets included in the NNPDF3.0 fits are included with correlated systematic uncertainties. This is generally provided either as an overall covariance matrix, or as a set of nuisance parameters with the corresponding point-by-point correlations. The only exceptions to this are the SLAC, NuTeV and fixed target Drell-Yan datasets, and the top pair production total cross-sections, where only a total uncorrelated systematic uncertainty is used.

In previous NNPDF fits, systematics have been separated into two categories, general uncertainties and normalisation uncertainties, with the latter being treated differently in the fit. The normalisation uncertainties were treated "multiplicatively", i.e. they were taken to be proportional to the theoretical value of the observable in question. This poses particular problems when including them in a fit, as mentioned in Section 3.2.3, because their naive inclusion in the covariance matrix would lead to a systematically biased result [53]. In hadron collider experiments, it is not only normalisation uncertainties but most, or perhaps all, of the correlated sources of uncertainty that are multiplicative. After checking with the respective experimental collaborations, we have thus concluded that the most accurate treatment of correlated systematics is obtained by treating all systematics as multiplicative. For deep-inelastic experiments, we treat all the correlated systematics of the HERA data as multiplicative, while for fixed-target experiments the systematics are treated as before, with only the normalisation uncertainties taken as multiplicative. This information is summarised in the fourth column of Table 4.1 (page 33); normalisation uncertainties are treated multiplicatively for all experiments (even those labeled "add"). In Chapter 7 I will study the effects of a changing this treatment of systematics, and find that it is small though perhaps not completely negligible, particularly for the large-x gluon.

In order to generate Monte Carlo replica datasets, as described in Section 3.2.4, we require a breakdown of the systematics into individual sources of independent correlated uncertainty. However for some LHC experiments—both LHCb sets and all of the CMS data except the jets—this breakup is not provided and only the experimental covariance matrix is available. In those cases, we create a set of artificial systematics which are consistent with the covariance matrix. To do this, first note that the covariance matrix C_{ij} can be obtained (ignoring statistical and other uncorrelated uncertainties) from the $N_{\rm sys}$ individual systematics v_i^k using

$$\begin{aligned} C_{ij} &= \sum_k^{N_{\rm sys}} v_i^k v_j^k && (4.4)\\ &= V_{ik}V_{kj} = \left(VV^T\right)_{ij} && (4.5) \end{aligned}$$

where in the second line we have defined a matrix V whose columns are the systematics. Based on this, in order to generate artificial systematics from a given C, we need to find a matrix V which satisfies Eq. 4.5. One possibility for this, which we use for the NNPDF3.0 data, is to perform a spectral decomposition of C, to obtain

$$\begin{aligned} C_{ij} &= U_{ik}\Lambda_{kl}U^T_{lj} = \left(U\Lambda^{\frac{1}{2}}\right)_{ik}\left(U\Lambda^{\frac{1}{2}}\right)^T_{kj} && (4.6)\\ \implies V &= U\Lambda^{\frac{1}{2}} && (4.7) \end{aligned}$$

where U is the matrix of eigenvectors of C and Λ is a diagonal matrix of the eigenvalues. This then provides a set of $\rm N_{dat}$ artificial systematics which can be used to generate replica datasets in the same way as if we had the full breakup of systematics uncertainties, according to the multivariate Gaussian distribution implied by the experimental covariance matrix, and which recombine correctly to give the original covariance matrix.

Chapter 5

Study of methodology for NNPDF3.0

5.1 Introduction

For the NNPDF3.0 analysis we have performed a complete overhaul of the NNPDF fitting methodology. Many outdated elements in the code have been removed and replaced by new features empirically shown to improve the quality of the results. In this section I will provide an overview of this work, describing the improvements to our approach and showing the results of tests we performed to demonstrate their effectiveness.

There are several reasons why the lead up to NNPDF3.0 was an ideal time for this renovation. The ability to perform closure tests (described briefly in Section 5.2 and at length in Section 6) provide an ideal environment for testing new minimisation features. Also, with the increasing precision of experimental measurements it is important that we can ensure that our analysis is as good as possible. The chief reason, however, was that the work could be included as part of a previously planned complete rewrite of fitting code.

The old fitting code, used for all previous central NNPDF analyses, was written in Fortran 77 and designed for the specifications of the NNPDF1.0 analysis [3] with a limited range of observables and simple minimisation strategy. Since then, a huge amount of new data has been added to the fit, including a large number of computationally intensive hadronic datasets, as well as many improvements to the methodology. This has led to performance issues as the structure of the code is not suited to the new tasks it has to perform. For instance, theoretical calculations were performed in several different ways depending on the process, which made the code confusing to work with and difficult to optimise. There was also the fear that the adjustments to the code may have introduced so far undiscovered bugs.

The new code, written in C++, has been written completely from scratch and extensively tested to reduce the chance of bugs. The modular structure of C++ has been exploited to completely separate the parts of the code which deal with data, theory and fitting methodology, which makes it easy to make changes to one aspect without disrupting the others. The calculation of theoretical observables has been standardised, with all processes now using FKtables, described in Section 3.2.5. This has allowed for huge performance improvements by optimising the simple vector products required to use the FKtables. Another source of optimisation has come from changing the genetic algorithm to check the χ^2 of mutants between each dataset and immediately reject any mutant which is already worse than the previous generation's best. If the datasets are sensibly organised, with hadronic processes calculated last, this results in another large increase in speed while obtaining precisely the same results. Overall, with these optimisations, as well as the changes to its structure, the new code is about 5 or 6 times faster that the old version.

This section will describe the many methodological improvements tested and used in the NNPDF3.0 global fits. First I will introduce the closure testing approach which is used for all of the tests presented in this section, while the remainder of the section is split into three main subsections, corresponding to different aspects of the minimisation, and presented in roughly the order in which the features were implemented and tested. Results for all of the new techniques which were included in final NNPDF3.0 will be shown, along with a subset of the other unsuccessful features and changes we investigated.

5.2 Closure Testing

In the past, development of the NNPDF fitting methodology has been performed using standard PDF fits to real experimental data. We would often then judge a new feature to be an improvement over the existing setup if it improved the quality of fit to the experimental data. However, in doing this we run the risk of tuning the parameters of the minimisation to the specific dataset used in the fit, effectively overfitting at the level of the methodology, with negative consequences for the predictive validity of the resulting determination. In general, it is difficult to determine that a decrease in the χ^2 actually represents an improved methodology, and that the results obtained using a methodology which includes the new features are a better description of the information in the data.

These problems can be avoided by instead using closure tests to evaluate the methodology. In closure tests, instead of using the real experimental data, PDFs are determined from pseudo-data generated using a known theory. There are several key benefits to this. Centrally, because the underlying theory in the pseudo-data is known, we can directly assess the effectiveness of our determination by comparing the PDFs from the fit to the generating PDFs. This allows us to investigate a range of statistical features of our fits, from testing the validity of the PDF uncertainties, to determining which fractions of the uncertainty are from functional, extrapolation and data sources. These issues will be studied in detail in Section 6.

Closure tests also provide the perfect environment for testing new methodological features. Again, the presence of a known 'right answer' is very important, and provides new ways to assess improvements beyond just looking at the value of the χ^2. However, with closure tests we also have the ability to generate as many different sets of pseudo-data as we want, each with unique statistical fluctuations generated using different random seeds. This allows us to confirm that a new feature provides a genuine improvement in the methodology over multiple datasets, rather than just improving the ability of fitting algorithm to model one specific set of data. As well as this ability to generate different statistical fluctuations, we also have the option to generate data where this noise is absent. Together with the additional fluctuations in the data which can be introduced during replica generation at the beginning of the fit, this gives three different types of closure test fit which can be performed, corresponding to different levels of noise in the data:

- **Level 0** With this setting data is generated without any stochastic noise from either the closure test data generation or replica generation steps. The pseudo-data used in each replica fit is therefore precisely the central theory values obtained from the generating PDF set. Fitting at level 0 tests of the ability of the minimisation algorithm to fit the underlying PDFs directly in the absence of noise, with the advantage that in this situation over-learning is not an issue.

  In our level 0 fits, the definition of the χ^2 minimised in the fit and shown in the results uses the same experimental (or more precisely, t0) covariance matrix used in level 1 and level 2 fits. In principle, since the level 0 data is noiseless, a different normalisation for the figure of merit could be used. However, using the experimental covariance matrix means that the χ^2 have the same correlations and relative weights between data points as in other fits. In any case, as the data is noiseless, the ideal χ^2 we want to obtain from a fit is zero.

- **Level 1** Here, data is generated with statistical noise given by the experimental uncertainties, but replicas during the fit are not varied and instead all replicas use the same (noisy) data. It is also possible to perform a slightly different type of level 1 fit by using only the fluctuations from replica generation, which is effectively the same as each replica being a separate level 1 fit.

- **Level 2** This is the full closure test, with statistical fluctuations introduced both during pseudo-data generations and during replica generation. This creates a situation which is equivalent to real fits, and so is useful for both tests of methods for controlling over-learning and to evaluate the statistical validity of our methodology as a whole. The ideal χ^2 to the replica data is the same as in fits to the real experimental data, i.e. values of about two.

In this section we will look at results from level 0 and level 2 closure tests, as these are both useful for testing different aspects of the minimisation. The absence of over-learning in level 0 closure tests make them ideal to test improvements to the genetic algorithm, since a lower χ^2 unambiguously indicates a better fit. On the other hand, level 2 tests are necessary to look at the impact of changes in the methodology on the PDF uncertainties, as well as improvements in approaches to control overfitting. Together, these two types of closure test provide a way to objectively test all aspects of the NNPDF fitting methodology. NNPDF3.0 is in fact the first PDF determination for which the complete fitting methodology has been thoroughly tested and tuned in closure tests based on pseudo-data that have the same kinematical coverage and statistical properties as the experimental data included in the fit. The idea of using perfect pseudo-data to validate some specific aspects of a PDF fitting methodology has been previously explored in Ref. [169].

Another difference between closure test fits and fits to real data is that, because the pseudo-data is generated using the same theoretical calculations eventually used in the fit, the theory and data are always perfectly consistent. This is actually somewhat of a disadvantage, as at the end of the day we want to fit PDFs using the real data, inconsistencies and all. In the future we plan investigate introducing artificial inconsistencies into the closure test data in a controlled way, in order to study how the fit behaves. However, for testing the methodology the consistency of data and theory is advantageous, as it removes another confounding factor in analysing the results and also means that the details of the theory used in the fit are largely irrelevant. For simplicity, the closure tests presented here and in the next section use the NLO fktables, and the same theory settings (cuts to the data, coupling values etc.) used in the NLO fits to real data. To generate the pseudo-data, for the closure tests in this section we used the MSTW2008 NLO PDF set. This choice was made for several reasons. Firstly, although in principle any functions could be used for the closure tests [1], it makes sense to use a set

[1] At least, any functions which satisfy the various theoretical constraints directly imposed in the fit, e.g. sum rules, large and small-x behaviour.

of at least PDF-like functions. Secondly, we decided to use an MSTW PDF set rather than an NNPDF set to remove the possibility of bias from the methodology fitting a shape it is predisposed to produce. Closure tests using other PDF determinations, and also using more unorthodox sets, were also performed and will be presented in Section 6.

One issue which came up when performing closure tests was the issue of the PDF positivity. Initially, the closure tests were performed using the NNPDF3.0 positivity constraints. However, it was discovered that a small number of positivity points, corresponding to regions with little constraint from data, are violated by the MSTW central values. This potentially leads to a situation where the PDF uncertainty in the extrapolation region is not consistent with the generating PDF. For this reason, the majority of the closure test presented here were therefore performed without the positivity constraints. The only exceptions are some of the level 0 closure tests of different genetic algorithm features, where we are generally only interested in the central values of the PDFs in the data region, so the impact from this issue is negligible.

5.3 Tests of the Genetic Algorithm

While many aspects of the NNPDF methodology have evolved since the release of the first NNPDF set, the underlying genetic algorithm has remained largely unchanged. The development of the new C++ code and the introduction of closure tests provide an opportunity to completely reevaluate the genetic algorithm and to test a number of variations on the original format. By improving the genetic algorithm, we have the potential to both obtain a better result at the end of the fit and to do so in a shorter amount of time.

In order to develop a new minimisation algorithm we decided to start from a pared down version of the genetic algorithm from NNPDF2.3, without targeted weight training or different training phases, or (for the moment) cross validation. New features were then introduced one at a time, so that the differences in the fit results when using them could be clearly attributed to each feature individually. This has the disadvantage of missing possible improvements from specific combinations of feature and parameters, however given the huge numbers of each it is not clear how a systematic check of combinations could be performed.

To look at changes to the genetic algorithm, we evaluated each feature on the basis of results from level 0 closure tests. For each feature or setting we performed a 100 replica, 20000 generation fit. To determine whether a feature improved the methodology, it was sufficient in most cases just to look at the average χ^2 of the replicas, $\langle \chi^2_{\text{rep}} \rangle$, which describes how good each individual replica fit was, and the central χ^2 of the ensemble,

$\chi^2_{\rm cent}$, which shows the quality of the full set. The closure test fits were performed using pseudo-data based on the NNPDF2.3 dataset, with the addition of the ATLAS 2.76 TeV jet data. As the study of the fitting methodology was performed early in the development of NNPDF3.0, many of the other sets included in the final analysis had not yet been implemented, resulting in this reduced dataset. The choice of datasets should have minimal impact on the conclusions drawn here, and final closure tests performed later with the full dataset demonstrated that this was the case.

5.3.1 Nodal Mutations

Genetic algorithms have been widely used over the last few decades, and so there exist a large number of potential features which it is possible to include in any particular implementation. The majority of these are very general, applicable to most situations involving genetic algorithms, and the majority of the genetic algorithm features studied in this section fall into this category. However, the use of nodal mutations is an innovation specifically in the use of genetic algorithms to train neural networks. The general idea is to exploit the structure of the networks to mutate sets of related parameters at the same time, leading on average to more successful mutants and improved training. Results for the use of nodal mutations were presented in a contribution to the proceedings of a conference in artificial intelligence [170], which studied a range of methods for fitting neural networks with genetic algorithms. In the example they use, they found a slight advantage in performance over a more standard mutation strategy. Here I will present results for fits using nodal mutations in the NNPDF methodology.

As mentioned in Section 3.2.1, the activation of a node is calculated using the activations of nodes in the previous layer and weights for the connections between the nodes. With nodal mutations, instead of mutating each of the weights independently, all of the weights used in the calculation of the activation of a node are mutated at the same time. Each node has a fixed chance of receiving one of these multi-mutations each generation, and the size and direction of mutation is calculated independently for each weight according to Eq. 3.9 as before. For initial tests of nodal mutations, the probability of mutation was chosen as 10% uniformly for all nodes and PDFs. This results in about a 60% chance for at least one mutation in each network, and a total of about 6 node mutations on average for the whole mutant.

The first part of Table 5.1 shows the central and average replica χ^2 for a fit using nodal mutations and a fit with the standard mutation strategy and otherwise identical settings. Both fits achieve low values, close to the ideal of zero, but the fit with nodal mutations is significantly better. The individual replicas are twice as good on

Description	$\langle\chi^2_{\mathrm{rep}}\rangle$	χ^2_{cent}
Standard Mutations	0.070(5)	0.0279
Nodal Mutations	0.035(3)	0.0077
Nodal Mutations (different seed)	0.029(2)	0.0076
Stochastic number of mutations	0.056(4)	0.0228

Table 5.1: Average and central χ^2s for different genetic algorithm mutation strategies. All numbers are based on single 100 replica, 20000 generation fits to a common level 0 closure test dataset. The uncertainty on the average χ^2s are also given. The different strategies are described in the text.

average, while the central χ^2 is over 3 times better. The uncertainty on the average χ^2, given by the standard deviation across the replicas divided by the square root of the number of replicas and shown in parentheses, suggest that this different is unlikely to be due to chance. Also shown in Table 5.1 are the χ^2s for another nodal mutations fit performed using a different random seed. This gives some idea of the variability in the central χ^2, information which will be useful when evaluating fits with different mutation parameters.

This comparison between the standard and nodal mutations strategies is slightly flawed, as along with the change in the type of mutation we have also switched from using a fixed number of mutations, as used in NNPDF2.3, to a fixed probability for each element (node or weight) to be mutated. In Table 5.1, the central and average χ^2 are also shown for a fit which mutates each weight individually, as in the old strategy, but now assigns an independent probability of each weight to be mutated. The χ^2 show that this strategy of using stochastic number of mutations leads to an improvement in the fit quality, however the results are not as good as for the fits which additionally use nodal mutations.

In the fit with nodal mutations described above, the parameters controlling the size and frequency of mutations were somewhat arbitrarily chosen. Since these parameters were unlikely to be optimal, we decided to perform a study of the parameter space of these two mutation parameters to look for a better set. The approach we took to do this was to first perform separate one dimensional scans in each parameter, and then perform a final combined two dimensional scan looking at the most promising region. The results for the mutation probability tests are shown in Table 5.2. We can see that the smaller values for the probability give better χ^2, although there is some noise in the central χ^2 with the value for the 0.05 fit being anomalously large. The corresponding results for the scan of mutation size are shown in Table 5.3. Here, larger values are preferred, with the best fit using $\eta = 15$. Finally, Table 5.4 shows results from fits using probabilities of 2% and 5%, and mutation sizes of 15 and 20, with the single best set of parameters based on this study being 5% and $\eta = 15$. We can also see

Mutation Probability	Mutation Size	$\langle\chi^2_{\rm rep}\rangle$	$\chi^2_{\rm cent}$
0.02	10	0.026(3)	0.0051
0.05	10	0.027(2)	0.0064
0.07	10	0.031(3)	0.0058
0.1	10	0.035(3)	0.0077
0.2	10	0.056(4)	0.0151

Table 5.2: Same as Table 5.1, but now for variations of the mutation probability in fits with nodal mutations. The fits were performed with otherwise identical settings and identical datasets.

Mutation Probability	Mutation Size	$\langle\chi^2_{\rm rep}\rangle$	$\chi^2_{\rm cent}$
0.1	2	0.054(4)	0.0207
0.1	10	0.035(3)	0.0077
0.1	15	0.029(2)	0.0054
0.1	20	0.041(3)	0.0068
0.1	30	0.070(6)	0.0122

Table 5.3: Same as Tables 5.1 and 5.2, but now for variations of the mutation size parameter (η) in fits with nodal mutations. The fits were performed with otherwise identical settings and identical datasets.

that the differences between the fits are relatively small, which is fairly encouraging as it suggests that the dependance of the fit on the precise values of these parameters is small, on the order of the dependance of the random seed.

From these results we can conclude that nodal mutations are a fairly substantial improvement over the NNPDF2.3 algorithm. On this basis, nodal mutations form a central part of the NNPDF3.0 genetic algorithm, and will be used as the default setting for the rest of the studies in this chapter.

5.3.2 Other Mutation strategies

Crossover

Another idea commonly used in genetic algorithms is gene crossover. This is where the new mutants in each generation are constructed using parts of multiple parents. Whereas mutation searches the parameter space, crossover looks to combine the best elements from the existing mutant population. Crossover is generally performed alongside standard mutations, and there are several different forms of crossover which can be implemented in a particular algorithm. Here we will look at implementations of uniform crossover with two parents, where elements of each mutant are selected with equal probability from each parent.

In order to investigate the impact of including crossover techniques in the NNPDF genetic algorithm, I performed several level 0 closure test fits with different crossover

Mutation Probability	Mutation Size	$\langle\chi^2_{\mathrm{rep}}\rangle$	χ^2_{cent}
0.02	15	0.044(5)	0.0065
0.02	20	0.032(3)	0.0053
0.05	15	0.024(2)	0.0043
0.05	20	0.029(2)	0.0041

Table 5.4: Same as Tables 5.1-5.3, but now for further variations of both the mutation probability and mutation size in fits with nodal mutations. The fits were performed with otherwise identical settings and identical datasets.

implementations. Each fit was run using the same settings as discussed above and with nodal mutations. At the end of each generation the best two mutants were identified and carried over to the next generation, instead of just one as used normally. Then, the crossover step was used to generate new mutants in four different ways:

- **Weight crossover**: Every weight in the mutant is set as the equivalent (i.e. same position in the network of the same PDF) weight from one of the parents, chosen with equal probability and independently for each weight. This is the most basic form of uniform crossover, however also has the possibility to be very destructive unless the two parent are very similar.

- **Node crossover**: Same as above, but instead of selecting each weight independently, the weights in a node are all taken from the same parent. This has the potential to be more effective in our fits than weight crossover, especially given the good results obtained using node based mutations.

- **PDF crossover**: The complete network for each PDF is randomly selected from one of the two parents. This has the advantage of avoiding crossover within the neural networks, though there are potential problems from some of the positivity constraints which depend on combinations of the PDFs.

- **No crossover (two parents)**: Each mutant is set initially as one of the two parents, randomly chosen with equal probability, before being mutated. This could alternatively be viewed as crossover of the entire PDF set. While this is not truly an implementation of crossover, it still has the potential to improve performance on the basis that the algorithm can to some extent explore the parameter space in two directions at once (albeit half as quickly).

After the crossover step was used to initialise the mutants, they were mutated in the standard (nodal) way.

The results from the crossover closure test fits are shown in Table 5.5. The most effective strategy using crossover is either the PDF or no crossover options, however

Description	$\langle\chi^2_{\rm rep}\rangle$	$\chi^2_{\rm cent}$
Standard approach	0.024(2)	0.0043
Weight crossover	0.059(6)	0.0115
Node crossover	0.034(4)	0.0073
PDF crossover	0.028(2)	0.0054
Two parents	0.030(3)	0.0053
Fitness Proportional Selection	0.103(9)	0.0232
$N_{\rm ite}$ mutation scaling with $\gamma = 0.8$	0.028(2)	0.0053
$N_{\rm ite}$ mutation scaling with $\gamma = 0.9$	0.035(4)	0.0054
$N_{\rm ite}$ mutation scaling with $\gamma = 1.2$	0.028(3)	0.0052
$N_{\rm ite}$ mutation scaling with $\gamma = 1.5$	0.031(3)	0.0053

Table 5.5: Average and central χ^2s for implementations of various different features in the generic algorithm. All numbers are based on single 100 replica, 20000 generation fits to a common Level 0 closure test dataset and using nodal mutations. The uncertainty on the average χ^2 is also given. The specifics of the different fits are described in the text.

neither are better than the standard approach. Node crossover is only slightly worse than these two options, while using weight crossover yields considerably poorer results. While it is perhaps possible that better results could be obtained by tuning the genetic algorithm parameters for crossover, as was done for nodal mutations, we ultimately opted instead to not include crossover in the NNPDF3.0 methodology.

There is reason to suspect that crossover would be somewhat incompatible with neural networks, given their high degree of both interconnectedness, which would be disrupted by crossover, and degeneracy, which means that individual weights within the structure do not necessarily play the same role even within networks which produce similar functions. Crossover is most effective in situations where the ‘chromosome’ of parameters is made up of largely independent pieces which can be readily swapped, which is not the case here even at the PDF level.

Fitness Proportionate Selection

In the standard NNPDF minimisation, the mutant selected to be carried forward from one generation to the next is always the best out of the current set of mutants (including the parent used to generate them). One common technique in genetic algorithms is to choose stochastically which mutant to keep instead. Fitness proportionate selection (or roulette wheel selection) does this by allocating a probability of selection to each mutant based on the fitness of the mutant relative to the others. The idea is that by allowing the training to occasionally choose a mutant with a larger error function, we can potentially prevent the fit from getting stuck in a local minimum.

Defining the ‘fitness’ of a mutant as the inverse of its χ^2, so that better fitting mutants have a larger fitness, we can define the probability that a mutant is selected in any given generations by

$$P_i = \frac{f_i}{\sum\limits_{i}^{N_{\text{mut}}} f_i} = \frac{\frac{1}{\chi_i^2}}{\sum\limits_{i}^{N_{\text{mut}}} \frac{1}{\chi_i^2}}, \tag{5.1}$$

where i runs over the N_{mut} mutants in that generation. With this implementation I performed a level 0 closure test fit using the same settings outlined above, and the average and total χ^2 from this fit are shown in Table 5.5. It is clear that the fit with fitness proportionate selection was substantially worse that the standard approach, and that any advantage in avoiding local minima is outweighed by the poorer overall performance. Again, it is possible that turning the fit parameters or using a different definition for the fitness f than in Eq. 5.1 above could improve these results. For the moment, however, fitness proportionate selection will not be used in the NNDPF methodology.

Mutation scaling

The formula for the mutations applied in the genetic algorithm is (repeated from Section 3.2.2)

$$w \to w + \eta \frac{r_1}{N_{\text{ite}}^{r_2}}. \tag{5.2}$$

The only nontrivial part of this equation is the factor of $N_{ite}^{-r_2}$, which reduces the average size of mutations as the fit goes on. This feature was originally introduced for NNPDF1.0 with a fixed exponent of 1/3, and was updated to its current form with a random exponent in the NNPDF2.0 analysis. In this section I will look at the impact of extending this feature by introducing a new parameter γ to modify the scaling exponent, i.e. by mutating according to

$$w \to w + \eta \frac{r_1}{N_{\text{ite}}^{\gamma r_2}}. \tag{5.3}$$

As r_2 is a uniform random number between 0 and 1, this essentially changes the range of r_2 to being between 0 and γ. Through this, we can control how rapidly the mutation size is reduced relative to the generation number.

The results for fits with $\gamma = 0.8$, 0.9, 1.2 and 1.5 are shown in Table 5.5. The value of γ has little impact on the performance of the genetic algorithm, and the effect due to it appears to be smaller than the noise in the results. No value performs better than the standard methodology however, so for the NNPDF3.0 fits we use $\gamma = 1$, as used in

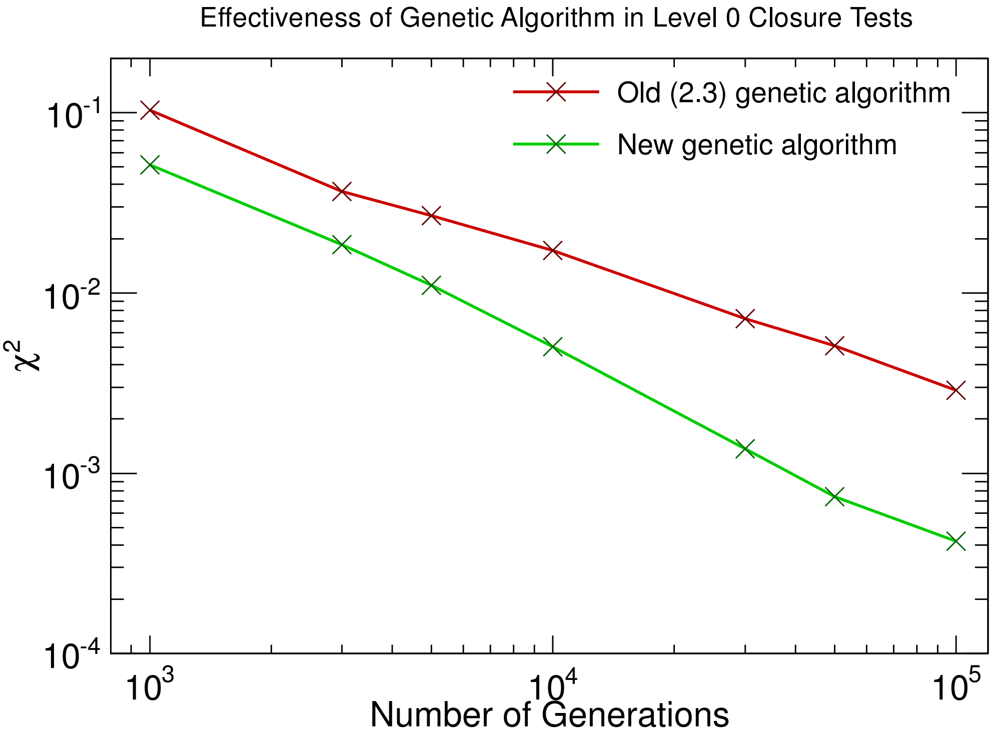


Figure 5.1: Central χ^2 calculated along the length of 100 replica level 0 closure test fits. The values for a fit using the NNPDF2.3 setting is shown in red, while those for a fit with the final nodal mutation settings is are shown in green. Both axes use a log scale.

the previous NNPDF fits.

5.3.3 Final genetic algorithm settings

Of all the features tested in this section, only nodal mutations demonstrated any significant improvement over the old algorithm. Therefore, the final genetic algorithm, which will be used for the rest of the studies in this chapter and for the final NNPDF3.0 fits, is given by the simple settings described at the start of this subsection combined with the nodal mutation approach using the mutation parameters we found to give the best results. Table 5.6 describes the new mutation settings and compares them to the settings used for NNPDF2.3. Fig. 5.1 compares the central χ^2 at different points in a long fit for the final genetic algorithm to the same quantity in a fit using the NNPDF2.3-like genetic algorithm. We can see that the new algorithm is better early on in the fit and continues to outperform the old approach in the later stages of the fit, widening the relative gap in χ^2.

NNPDF2.3		
Single Parameter Mutation		
PDF	N_{mut}	η
$\Sigma(x)$	2	10, 1
$g(x)$	3	10, 3, 0.4
$T_3(x)$	2	1, 0.1
$V(x)$	3	8, 1, 0.1
$\Delta_S(x)$	3	5, 1, 0.1
$s^+(x)$	2	5, 0.5
$s^-(x)$	2	1, 0.1

NNPDF3.0		
Nodal Mutation		
PDF	P_{mut}	η
$\Sigma(x)$	5% per node	15
$g(x)$	5% per node	15
$V(x)$	5% per node	15
$V_3(x)$	5% per node	15
$V_8(x)$	5% per node	15
$T_3(x)$	5% per node	15
$T_8(x)$	5% per node	15

Table 5.6: Comparison of genetic algorithm parameters between the NNPDF2.3 and NNPDF3.0 fits. The mutation parameters are shown for the two determinations in terms of their respective fitting bases. For the NNPDF3.0 fit the mutation probability is now set at 5% per network node, and the mutation size is set to a consistent $\eta = 15$, while in NNPDF2.3 there were a fixed number of mutations for each PDF each generation, with different sizes.

5.4 Parameterisation and Neural Network Structure

5.4.1 Variations of network design

Another underlying issue in the NNPDF methodology is the structure of neural network used in the fit. This is related to the genetic algorithm, in that a better performing algorithm can capitalise on a more complicated parameterisation. However, it is also connected to the issue of overfitting, as this becomes a greater danger as the size of the networks increase. In this section I will present some results of tests of the neural network size and structure in closure tests.

Size

Neural networks are widely used because of their characteristics as flexible unbiased interpolators, capable of modelling any continuous function given an infinite number of nodes [47]. However, since an infinite sized neural networks is impractical, we must make do with a finite sized neural network which is large enough for what we need to model. The typical way to ensure that the networks used are sufficiently large is to look at networks which are larger or smaller than your chosen size and examine the effects of using them in fits. For a reasonably sized network, the results should be largely independent of small changes in size.

This test has been performed for the NNPDF methodology in the past, as work leading to NNPDF1.0 [3]. However, the dataset used in the fit has greatly increased since then, in particular with the inclusion of hadronic data, so it is not clear whether these previous results will still hold. In addition, the previous test only looked at

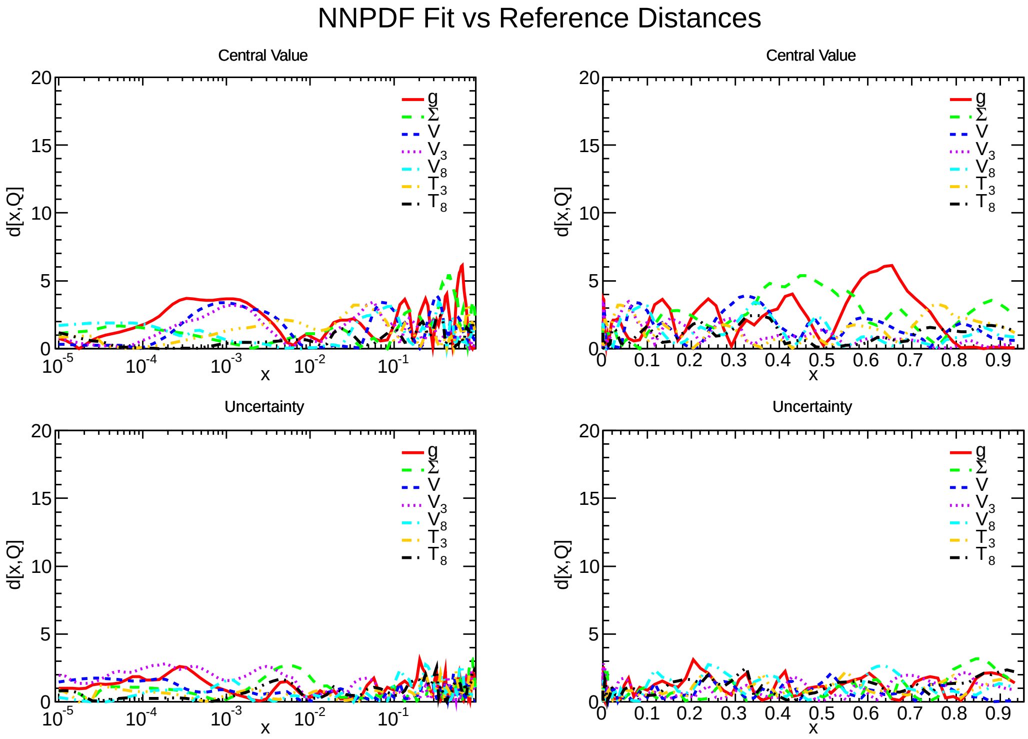


Figure 5.2: Distances between PDF central values and uncertainties of level 2 NNPDF3.0 closure tests using 2-5-3-1 and 2-20-15-1 neural networks. The distances are shown for the evolution basis PDFs in both log and linear scales for x, and for the fitting scale of $Q^2 = 1\text{GeV}^2$. The definition of PDF distances is given in Appendix A.

results for a smaller network than used in the main fits, and only one node smaller (2-4-3-1, compared to the usual 2-5-3-1, which has 6 fewer parameters). Also, this test was performed with the real experimental data, rather than with closure test pseudo-data as we use here, so it will suffer from some of the problems associated with this mentioned previously.

In order to look at the dependence of the NNPDF3.0 fits on the size of neural network used, we performed a level 2 closure test fit with the final NNPDF3.0 dataset and settings (including parametrisation and stopping settings discussed later in this section), and with an extremely large neural network. Instead of the standard 2-5-3-1 structure, the PDFs were parametrised using 2-20-15-1 networks, which have over ten times the number of parameters. The distances (see Appendix A) between this fit and a fit using the standard sized neural networks are shown in Fig. 5.2. The distances are reasonably small, below 5 at all values of x for most PDFs, which is only slightly higher than the general standard for the same fit performed with two different seeds. The largest discrepancy is for the large-x gluon, where the central values have a distance of about 6 between $x = 0.6$ and 0.7. The gluons for the two fits over this region are

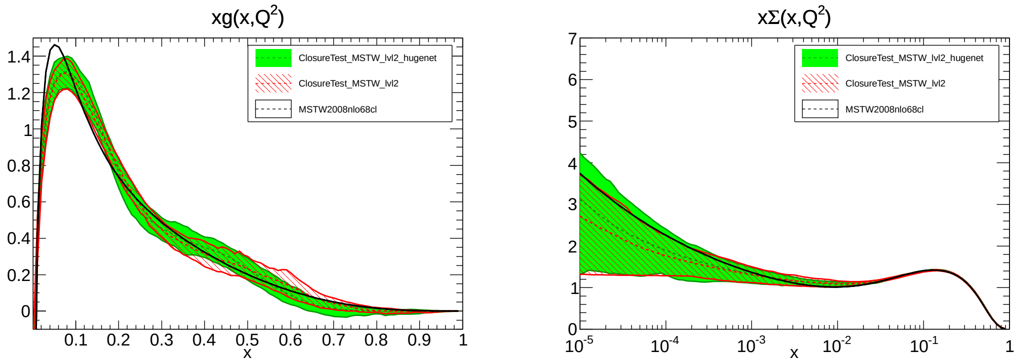


Figure 5.3: Comparison of gluon and singlet PDFs for level 2 NNPDF3.0 closure test fits using 2-5-3-1 (red) and 2-20-15-1 (green) neural networks. The central value for each PDF are shown by the dotted lines, while the bands give the 68% confidence intervals. The central values of MSTW2008 NLO, the PDF set used to generate the closure test data used in the fits, is included with the black curve. The gluon PDFs are plotted with a linear scale in x in order to highlight the differences at large x, while the singlets are shown on a log scale.

pictured in the left hand plot of Fig. 5.3. While we can see that the central value of the gluon from the huge network fit (dotted green line) is somewhat far from that of the standard network fit (in red), it is further away from the 'correct' result of the MSTW PDF (in black). This may indicate that the discrepancy is caused by overfitting, with the increased flexibility of the network allowing for better fit to the data but a poorer description of the underlying theory. The singlet PDFs for the two fits are also shown in Fig. 5.3, but here the two fits more or less agree, as indicated by the distances plot.

These results show that the fit, while there are some slightly significant differences from hugely increasing the complexity of the neural networks, the results are largely consistent and even possibly slightly worse. This demonstrates that the current 2-5-3-1 network structure is sufficient to successfully model the data, and justifies using this structure in the NNPDF3.0 analysis.

Structure

In addition to the size of the neural networks, another property which may have an effect on our fits is the structure of the networks, in particular the number of hidden layers. NNPDF analyses in the past have used networks with two hidden layers between the input and output layers. Neural networks with a single hidden layer were more common in the past, but multi–hidden layer or deep networks have recently seen an increase in popularity [171].

Fig. 5.4 shows the distances between a level 0 closure test fit performed using the standard two hidden layer networks and another using instead only one hidden layer. The neural networks with one hidden layer used the structure 2-9-1 in order to

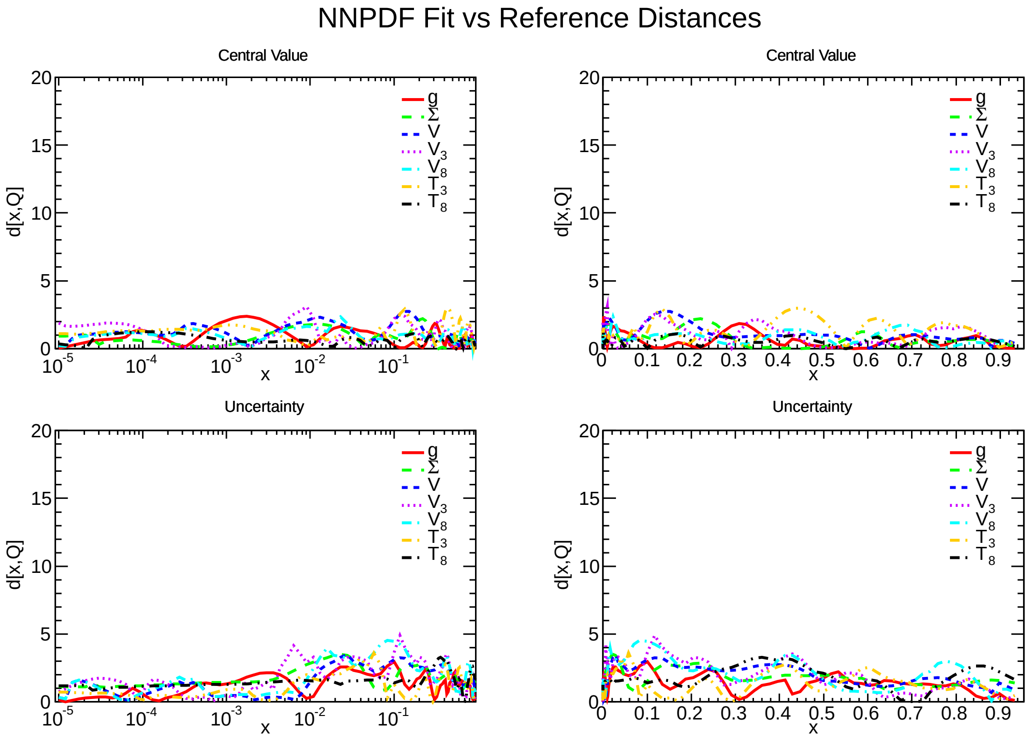


Figure 5.4: Same as Fig. 5.2, but for level 0 closure test fits with one and two hidden layers in the neural networks.

maintain the same total number of parameters (37) as the standard 2-5-3-1 networks. The distances in the central values are small for all PDFs, essentially consistent with statistical fluctuations, from which we conclude that number of nodes has a very minor impact on PDF fitting.

5.4.2 Input normalization

When using neural networks it is common to normalise the input values provided to the first layer of the network. This preconditioning can result in improved training, though it is not strictly necessary as the neural network should be able to adjust automatically to the scale of the input. A typical approach is to rescale the inputs so that they are all within one or two of zero, for instance by subtracting the average and dividing by the standard deviation.

The neural networks used in the NNPDF methodology contain two input nodes, one for x and another for $\ln(x)$. The x input takes values between 10^{-7} and 1, which means that the $\ln(x)$ input receives values in the range -7 to 0. In previous NNPDF fits, input normalisation was applied to both inputs, so the actual values provided to

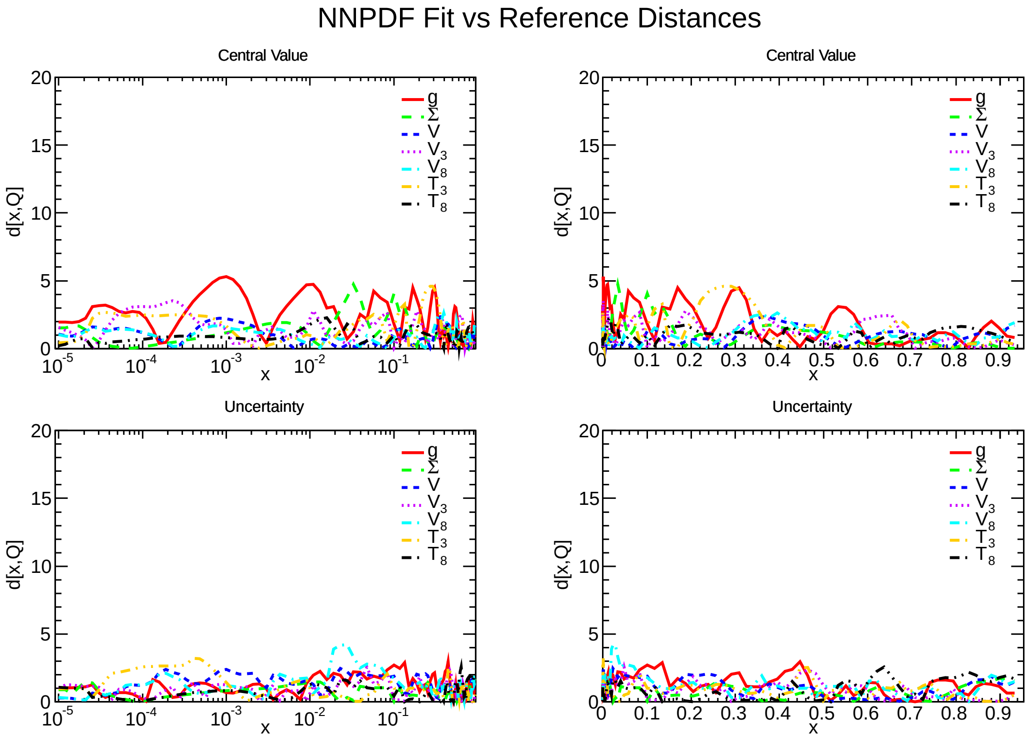


Figure 5.5: Same as Fig. 5.2, but for level 0 closure test fits with and without input normalisation.

the neural networks were

$$\xi_1 = 0.8\frac{x-\delta}{1-\delta} + 0.1 \tag{5.4}$$

$$\xi_2 = 0.8\frac{\ln(x)-\ln(\delta)}{-\ln(\delta)} + 0.1 \tag{5.5}$$

$$\tag{5.6}$$

where δ is a small constant set to 0.001 for all PDFs. This normalising function leaves the x input mostly unchanged, but rescales the $\ln(x)$ input to the range $[-1, 0.9]$. This potentially decreases the training time required by reducing scale disparity between the two inputs.

With the development of the new NNPDF3.0 genetic algorithm, we decided to investigate what impact including this input normalisation has and whether it is necessary with the updated approach. We performed a set of level 0 closure test fits, one with the input normalisation and another without, where x and $\ln(x)$ are given directly to the networks. We can see from the PDF distances in Fig. 5.5 that the normalisation only has a small impact for most PDFs, with a larger effect in the singlet and especially the gluon.

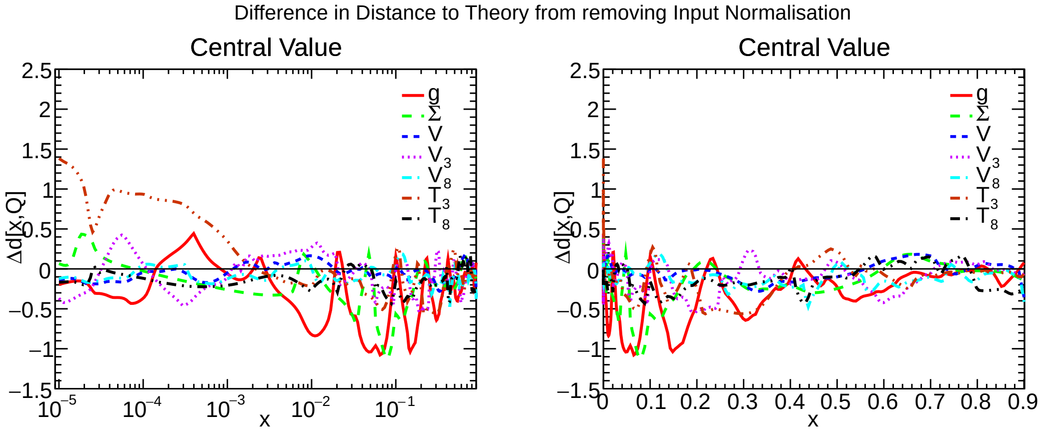


Figure 5.6: Difference in closure test distance between fits with and without input normalisation. The value indicates how much the distance to the MSTW PDFs used to generate the closure test data changed due to removing input normalisation. The values were calculated at the initial fitting scale. The definition of closure test distance (which is slightly different from the distance shown in Fig 5.5 etc.) is given in Appendix A.

The distances show the magnitude of the difference between the two fits, but not its direction, i.e. whether removing input normalisation results in a fit closer or further away to the underlying MSTW PDFs. In order to look at this we can compare the distances of each fit to the MSTW PDFs (which must be defined in a slightly different way, see Appendix A. Fig. 5.6 shows the change in the distance to the MSTW PDFs from turning off input normalisation. A negative value indicated that the PDF fit without input normalisation is closer to the MSTW PDF, while positive values show it is further away. The values are generally negative, especially in the data region, which indicates that the fit without input normalisation is closer to the ideal value.

We can also study the PDFs from the two fits more directly. Fig. 5.7 shows ratio

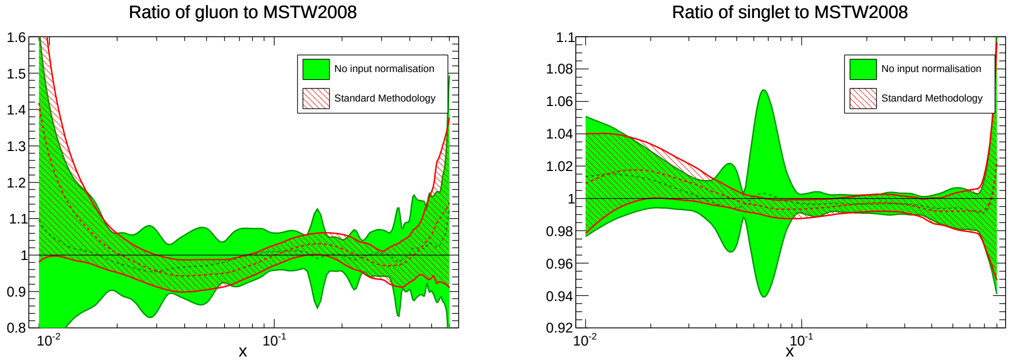


Figure 5.7: Ratios of the gluon and singlet PDFs for level 0 closure test fits with (red) and without (green) input normalisation to the central values of MSTW2008. The ratios were calculated at the initial fitting scale of $Q^2 = 1\text{GeV}^2$.

plots for the gluon and singlet of the fits, in the region around 10^{-2} and 10^{-1} where the differences are large. The PDFs are plotted as a ratio to the MSTW2008 set used to generate the closure test data, so the underlying theory that the fit is trying to reproduce is given by the line at one. We can see for the gluon that the fit with input normalisation (shown in red) oscillates considerably around the ideal value, while the central value for the fit without (the dotted green line) is more consistantly closer to the theory. These plots also show that the standard deviation for the fit without input normalisation, given by the green band, is much less smooth than the red band of the fit with input normalisation. This indicates that the input normalisation does help with the network training, as without it there are still a significant number of undertrained replicas which pull the standard deviation away from the average, creating the large bumps. This is is alleviated by increasing the training length, and the PDFs produced using the final NNPDF3.0 are much smoother.

Overall, input normalisation mostly has a small impact on the fit, and while slight improvement can been seen in terms of having fewer outliers, it appear to introduce a bias in the determination of the central values for the gluon and more generally gives a worse reconstruction of the underlying law. On this basis, we have removed input normalisation from the NNPDF methodology.

5.4.3 PDF basis

As previously mentioned in Section 3.2.1, all of the previous NNPDF fits have been performed using a modified PDF evolution basis with explicit parametrisation of the up-down quark sea asymmetry Δ_S instead of the valence triplet V_3. However, we now realise that the combination of this basis and our treatment of PDF sum rules imposes an unphysical restriction on the range of preprocessing values which can be used for the Δ_S and T_3 distributions. In the NNPDF2.3 basis, the second valence sum rule is imposed during the fit by setting the overall normalisation of Δ_S according to the integral of a combination of T_3 and Δ_S (see Eq 3.7). In order for this give sensible results the integral must converge, however in practice the only way to achieve integrability for the combination is to have integrability of both Δ_S and T_3 separately. In particular, we need a good chance of the PDFs to be integrable even for a neural network with random parameters in order to find a reasonable starting point for the fit. So, to get around this we must limit the range of preprocessing exponents we assign to these T_3 and Δ_S to only those which are finite when integrated down to $x = 0$.

For the NNPDF3.0 fits, we have solved this problem by making two main changes. Instead of using the PDF basis utilised in previous fits, we have moved to using the evolution basis as the standard parameterisation basis in all fits. The second valence

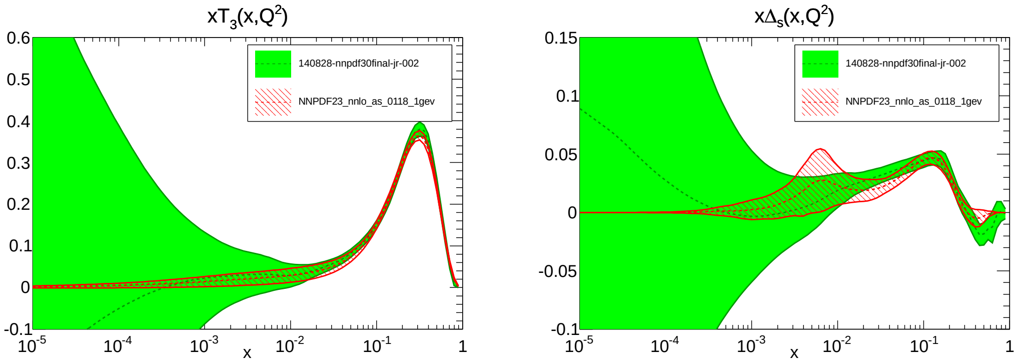


Figure 5.8: T_3 and Δ_S PDFs for NNPDF2.3 (red) and NNPDF3.0 (green) showing the impact of the extended preprocessing ranges. The dotted line shows the central value, while the band gives the 68% confidence interval. The PDFs are plotted at the initial fitting scale $Q^2 = 1\text{GeV}^2$.

sum rule is now imposed as a condition on the normalisation of the V_3 distribution, according to

$$\int_0^1 dx\ V_3 = 1, \tag{5.7}$$

with the preprocessing exponent for V_3 chosen in order to impose integrability, as in the other valence distributions. Using the evolution basis also means we now model the V_8 and T_8 distributions directly instead of s_+ and s_-. Secondly, we have decoupled the neural network and preprocessing bases, with the later always being applied in the evolution basis. For the evolution basis, nothing changes, but in any other bases the neural network bases are rotated before preprocessing. This allows us to the use the NNPDF2.3 basis or any other basis without worrying about integrability, which is useful to test that the results are independent of this choice. This also has the advantage that we do not need to redetermine preprocessing exponents when changing bases.

The impact of the new small-x preprocessing ranges for T_3 and Δ_S can be seen in Fig. 5.8, where the two PDFs are shown from fits using the NNPDF2.3 and NNPDF3.0 settings. The uncertainties on the PDFs now blow up after about $x = 10^{-2}$ as the replicas are no longer forced to go to zero for small x. The new preprocessing range has also removed the odd bump seen in Δ_S at around $x = 0.05$.

One unexpected side effect of preprocessing in evolution basis was that the method we used in the past to ensure PDF positivity at LO no longer worked. At LO, the flavour basis PDFs themselves are positive definite, instead of just physical observables. In the past we have imposed this exact positivity by performing the fit in the flavour basis squaring the output of the networks before the preprocessing stage. As the preprocessing term can only change the overall sign of the PDF (and the data will constrain at least some of each flavour to be positive), this guaranteed that the resulting PDF was non-negative. However, if we perform the preprocessing in a different basis from the networks, squaring the network outputs will no longer be sufficient to ensure positivity. This is because the actual flavour basis PDFs are defined in terms of the preprocessed evolution basis PDFs, while the neural networks only set the unpreprocessed evolutions basis PDFs. This means that in general the final flavour basis PDFs cannot be written as some function times a single neural network, and are instead combinations of multiple neural networks. Depending on the relative sizes of the different neural networks this can lead particular PDFs to become negative. For the NNPDF3.0 LO fits, instead of squaring the network outputs we impose positivity using the same positive observables used in the NLO and NNLO fits, though with a much larger Lagrange multiplier in order to generate a stricter constraint.

In addition to these changes to the application of the valence and total momentum sum rules to our fits, we have also added new checks on a number of related PDFs sums. The up, down and strange total momentum fractions, defined by

$$F_u = \int_0^1 dx\; x\; (u(x) + \bar{u}(x)) \tag{5.8}$$

$$F_d = \int_0^1 dx\; x\; (d(x) + \bar{d}(x)) \tag{5.9}$$

$$F_s = \int_0^1 dx\; x\; (s(x) + \bar{s}(x)), \tag{5.10}$$

do not have an associated sum rule, in the sense that their values are not constrained by theory and are instead they are determined by the fit. However, we do know that they must be integrable and so in NNPDF3.0 we enforce this as a condition during the fit, and reject any mutant for which the above integrals to not converge.

The changes to the parametrisation basis described above could potentially affect the results of the fit. Given the flexibility of the neural networks used in NNPDF fits, we do not expect this effect to be large. Fig 5.9 shows the distances between a pair of level 2 closure test fits using the evolution and NNPDF2.3 bases for the neural networks, with preprocessed in the evolution basis in both cases. The distances for the central values are all below 4, indicating a high level of consistency. The largest distances are seen in the strange and anti-strange distributions, which is unsurprising as here the basis has changed from modelling $s_\pm$ to V_8 and T_8.

5.4.4 Preprocessing

Apart from the changes to the basis described above, the preprocessing in NNPDF3.0 fits is done in much the same was as in previous NNPDF fits, using the same form for the prefactor given in Eq. 3.4 with randomised exponents for each PDF. As mentioned

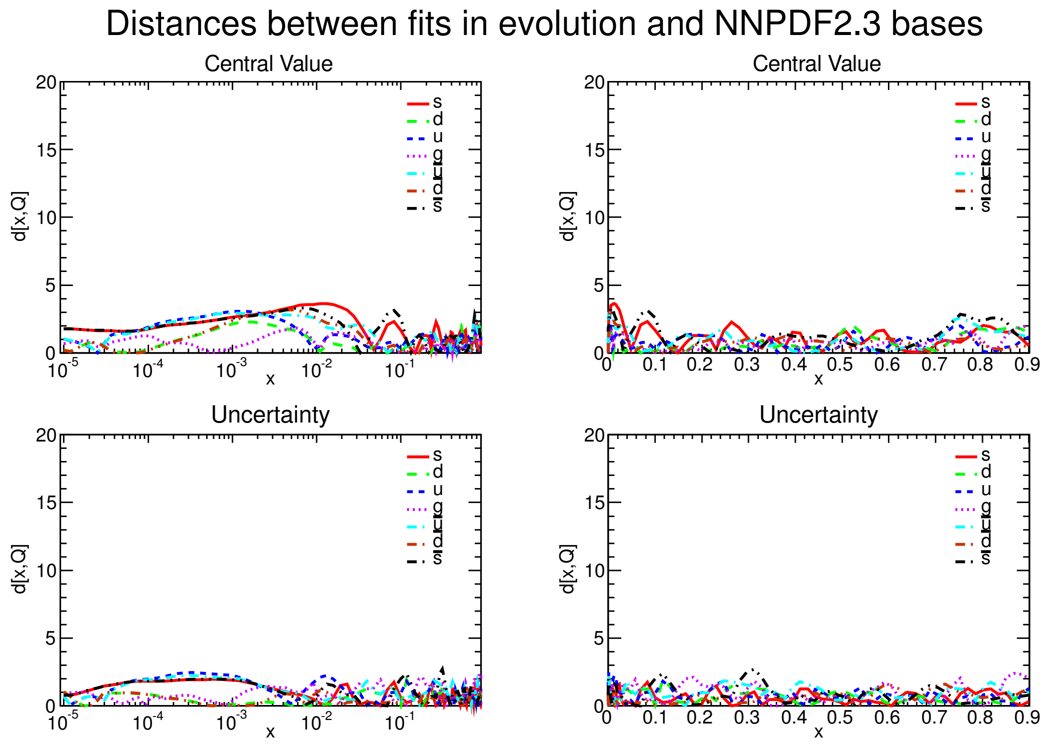


Figure 5.9: PDF distances between level 2 closure test fits using the evolution and NNPDF2.3 bases for the neural networks.

in Section 3.2.1, the range these exponents are chosen from must be selected in order to speed up the fits without biasing the results. Unlike in earlier NNPDF fits, where this range was determined based on a stability analysis of the results of multiple fits, we now generate the range using an automatic self-consistency procedure.

The new procedure works iteratively, generating the range for each fit based on an earlier fit with the same settings. First, we calculate effective asymptotic exponents for the initial scale PDFs of the first fit using

$$\alpha_{\text{eff},i}(x) = \frac{\ln f_i(x)}{\ln 1/x} \tag{5.11}$$

$$\beta_{\text{eff},i}(x) = \frac{\ln f_i(x)}{\ln(1-x)}. \tag{5.12}$$

Other definitions for the effective exponents would be possible, such as

$$\alpha_{\text{eff},i}(x) = \frac{d\ln f_i(x)}{d\ln 1/x} \qquad \beta_{\text{eff},i}(x) = \frac{d\ln f_i(x)}{d\ln(1-x)}, \tag{5.13}$$

and give quantitively similar results. The effective exponents are calculated for each PDF of each replica at a number of points in x: $x = 0.65$ and 0.95 for all of the

PDF	NLO		NNLO	
	$[\alpha_{\rm min}, \alpha_{\rm max}]$	$[\beta_{\rm min}, \beta_{\rm max}]$	$[\alpha_{\rm min}, \alpha_{\rm max}]$	$[\beta_{\rm min}, \beta_{\rm max}]$
Σ	[1.06, 1.22]	[1.31, 2.68]	[1.02, 1.33]	[1.31, 2.74]
g	[0.96, 1.37]	[0.28, 5.45]	[1.05, 1.53]	[0.85, 5.20]
V	[0.54, 0.70]	[1.20, 2.91]	[0.54, 0.70]	[1.18, 2.80]
V_3	[0.29, 0.58]	[1.31, 3.42]	[0.29, 0.61]	[1.36, 3.73]
V_8	[0.54, 0.73]	[0.80, 3.09]	[0.55, 0.72]	[1.06, 3.07]
T_3	[-0.17, 1.36]	[1.58, 3.14]	[-0.25, 1.41]	[1.64, 3.20]
T_8	[0.54, 1.25]	[1.30, 3.42]	[0.54, 1.27]	[1.33, 3.23]

Table 5.7: Ranges from which the small- and large-x preprocessing exponents in Eq. 3.4 are randomly chosen for each PDF. For each replica, a value is chosen from these ranges assuming a flat probability distribution. Shown are the values used for the global NLO and NNLO NNPDF3.0 fits. The two sets of ranges, obtained at each perturbative order, are determined independently using an iterative procedure, as explained in the text.

large-x β exponents; $x = 10^{-6}$ and 10^{-3} for the small-x α exponents, except for α_Σ and α_g where only the first value is used due to their increased structure at x values around the second. The new preprocessing ranges are then defined for each exponent as the envelope of twice the 68% confidence interval for each x value, where by "twice the 68% confidence interval" we mean a value twice as far above or below the mean value than the upper and lower limit. This condition provides a range which is large enough to easily cover the range of variation seen in the replicas, ensuring that the preprocessing exponents are not drawn from too narrow a range. The second points in x, at 10^{-3} and 0.65, help convergence of the criterion by considering the exponents in the non-asymptotic region where the existing preprocessing has less impact. The process is iterated until the new range matches or lies within the old one. Reassuringly, convergence is typically very fast, with generally only one iteration needed to achieve stability in most cases.

This procedure was used to generate the preprocessing exponent ranges for all NNPDF3.0 fits, and so there is no single set of ranges for the analysis as a whole. Instead, each individual fit has a different set of ranges, with for example fits to reduced datasets requiring considerably wider ranges than the standard dataset fits due to the reduced constraint from the experimental data. Table 5.7 shows the range for each exponent of the final central NLO and NNLO NNPDF3.0 fits, while Fig 5.10 shows the preprocessing range and the calculated effective exponents for the gluon and singlet from the 100 and 1000 replica global NNPDF3.0 NLO fit. The dashed lines show the double 68% confidence envelope used to generate the preprocessing ranges.

This procedure has also been used in the generation of a set of polarized neural network PDFs [172, 173]).

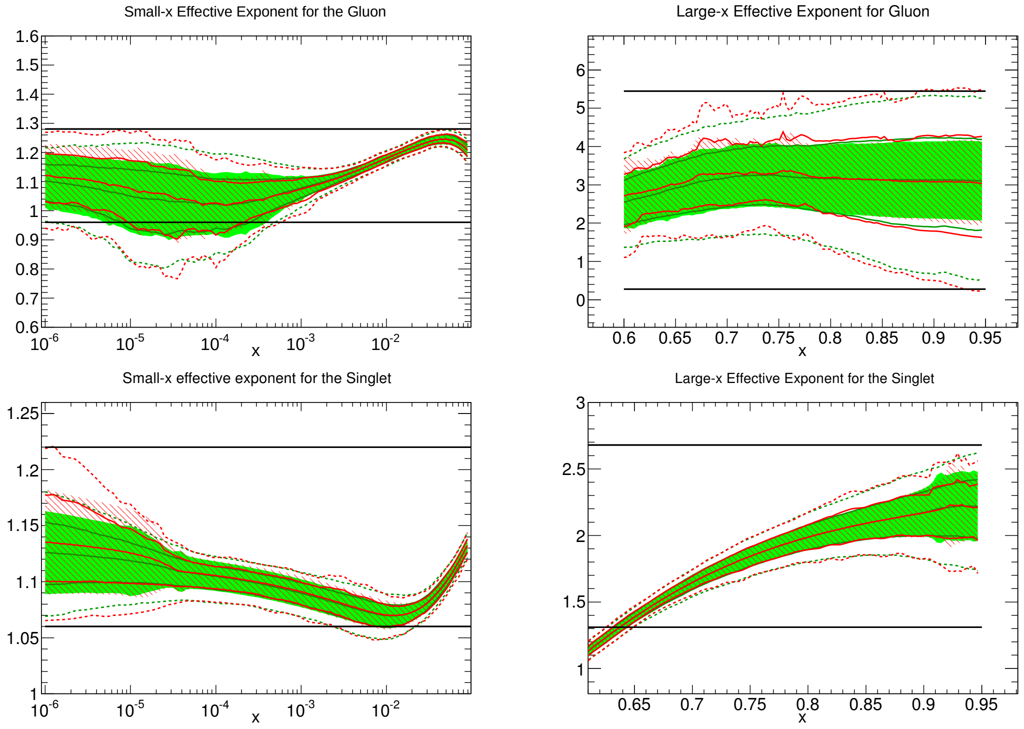


Figure 5.10: The small-x and large-x effective asymptotic exponents, α_{eff}, Eq. 5.11 (left) and β_{eff}, Eq. (5.12) (right), for the gluon (top) and singlet (bottom) in the global NNPDF3.0 NLO sets, for 100 replicas in red and 1000 replicas in green. The solid lines give the central values and 68% confidence intervals, while the dashed line is double the 68% confidence interval (compared to the mean). The black solid horizontal lines provide the range for the preprocessing exponents used in the fit.

5.5 Controlling Overfitting

As mentioned in Section 3.2.2, the flexibility of neural networks introduces a significant possibility of overfitting, which is where the fitted functions model not just the underlying law but also the statistical noise in the data. With the improvement to the genetic algorithm described above, it is important that we also investigate ways of improving the methodology for preventing overfitting. In this section I will describe three different methods to deal with the problem of over-learning, look-back cross-validation, weight penalty training and weight decay. The first is a variation on the cross-validation technique used in previous NNPDF fits, while the other two work by a very different principle.

As in Section 5.3 we will again look at results of various closure test fits, but now because we want to look at overfitting we will always use level 2 closure tests. The fits described in this section were performed using all of the new features described

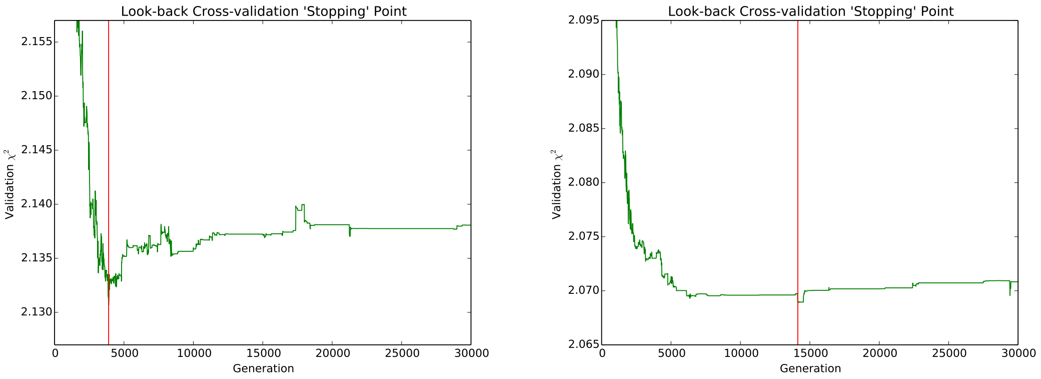


Figure 5.11: Validation χ^2 against generation for individual replicas from a level 2 closure test fit. The red line indicates the point at which the look-back cross-validation method chooses to 'stop' the fit. The plot is based on values taken every ten generations, so in both cases the green points do not display the lowest value itself.

previously in this section, with pseudo-data generated using MSTW2008 NLO PDF set and based on the final NNPDF3.0 dataset. Here we will focus specifically on the impact of introducing the various techniques for controlling overfitting; a more general analysis of the validity of the PDF uncertainties themselves is presented in Section 6.

5.5.1 Look-back cross-validation

The general idea of cross-validation is described at the end of Section 3.2.2. In short, the data is split into two subsets and the networks are trained using only one, while the other subset is used to detect overfitting. During correct training, the quality of fit to both the training and validation datasets decreases. When the quality of fit to the validation set increases, this indicates that over-learning is occurring and the fit is stopped.

While the method itself is quite simple, complications arise from the fact that the quality of fit to the validation set is often noisy from generation to generation. This means that an unsophisticated stopping condition, based solely on the validation χ^2 in one generation being higher than that of the previous generations, will stop the fit while authentic training is occurring. What we want to look instead is the general trend of the quality of fit to the validation set. In previous NNPDF fits, this method used to do this was to average the χ^2 over several hundred generations, smearing out the noise in the figure, and to tolerate small increases in the figure of merit. However, this introduced a number of extra fitting parameters which needed to be tuned in order to obtain a balance between stopping too early and too soon.

For NNPDF3.0, we have taken a different approach to cross-validation. We have discarded the concept of stopping the fit, and instead let every replica run until some pre-specified long cutoff. We then select at the end of the fit the generation where the validation χ^2 was lowest. The idea is that, instead of trying to determine whether the validation χ^2 is increasing, it is better to retrospectively identify the global minimum. The final PDFs will therefore be the best fit to the unseen dataset that was obtained during the fit. The two plots in Fig. 5.11 show the new method in action. In each plot the validation χ^2 for each generation is shown in green, while the red line shows the point in the fit chosen by look-back cross-validation. In some cases the point chosen can be quite early in the fit, as is the case on the left, while in other is can be much later on or even right at the end.

The obvious disadvantage of this approach is that all of the fits need to be run for a long time, rather than stopping early. The whole process of generating a complete set of replicas will therefore require appreciably larger computer time. However, because we perform many Monte Carlo replica fits for individual PDF determination, and under the old stopping criterion some replicas failed to stop, the real world time required to perform a full fit is largely unchanged.

Results

Look-back cross-validation was tested using level 2 closure test fits to data generated using the MSTW 2008 NLO PDF set with the settings of the final NNPDF3.0 fits. Fig. 5.12 shows the PDF distances between a fit using the new cross-validation method and a separate fit using the same random seed without cross-validation or the training-validation split of the dataset. Cross-validation has only a minor effect on the fit, and for both the central values and uncertainties the distances are mostly below 3, with a small number of spikes to at most a distance of 6.

Fig. 5.13 shows the change in distance to the MSTW PDFs caused by introducing look-back cross-validation compared to the fixed length fit. For the V_8 and large-x valence PDFs cross-validation improves the closure test fit, while for the others the results are mixed. The only major change is in the gluon at about 0.05, where there is a relatively large spike. Comparing to the distances in Fig. 5.12, it is not immediately obvious where this seemingly large discrepancy comes from. It turns out that this is at a point where the gluon for this particular closure test is far from the MSTW PDF, and the small increase in uncertainty in the gluon in this region, shown by a tiny bump on the distance plot, is enough to cause the large difference in closure test distance.

In general though, these results tell us that that the introduction of look-back cross-validation has a very small impact on the fit. This indicates that whatever overfitting is present in the NNPDF fits is small, possibly because of we use a large dataset with

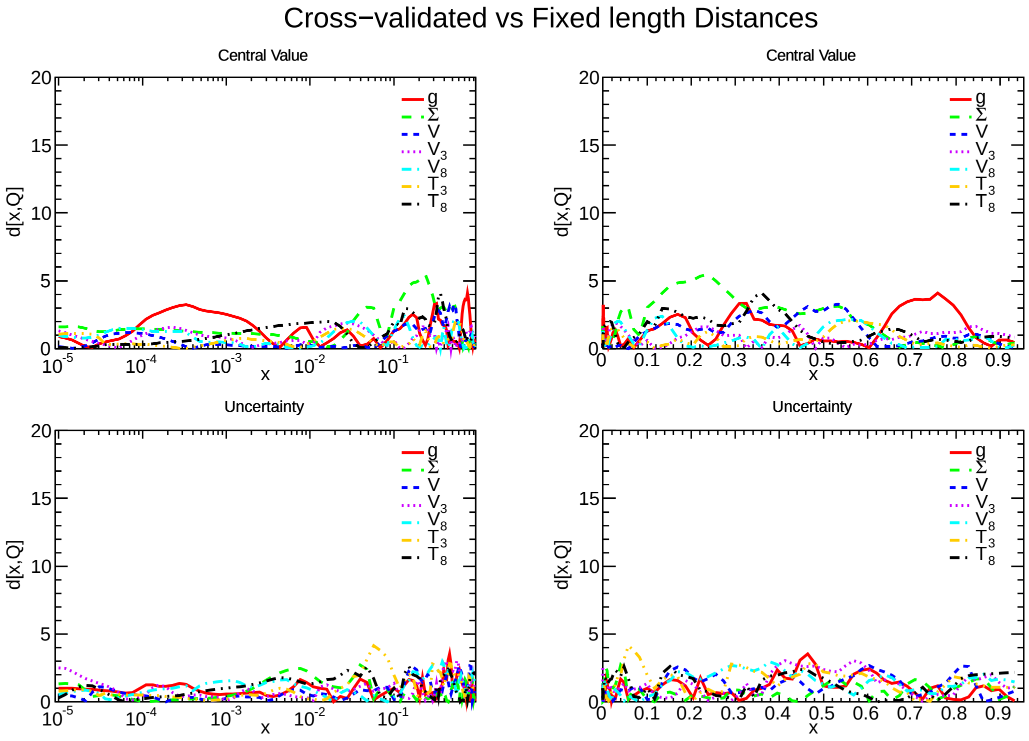


Figure 5.12: PDF distances between level 2 closure test with and without look-back cross-validation. The fit without cross-validation was performed without training-validation splitting, i.e. using 100% of the data.

a high level of redundancy. However, there are still several reasons to include cross-validation in the final NNPDF3.0 settings. It is possible that the tests we have used are not precise or comprehensive enough to detect all over-learning, and some could remain in the fit. It is also useful to be able to produce fits to radically different datasets with as close to the same methodology as possible, including fits to reduced datasets. For instance, in the case of fits to a HERA-only dataset the level of redundancy in the dataset is greatly reduced, so overfitting becomes a much bigger concern. Considering these issues, and the fact that the overall impact in most fits will be very small, we incorporated look-back cross-validation in the standard NNPDF3.0 methodology.

Dependance on maximum number of generations

One of the advantages of look-back cross-validation over the stopping-oriented cross-validation used in previous NNPDF fits is that it has much fewer fitting parameters. The only parameter it uses is the maximum number of generations, N_{gen}, i.e. the number of iterations that the fit is performed for and so that the global minimum is determined from. Provided that the maximum number of genetic algorithm generations

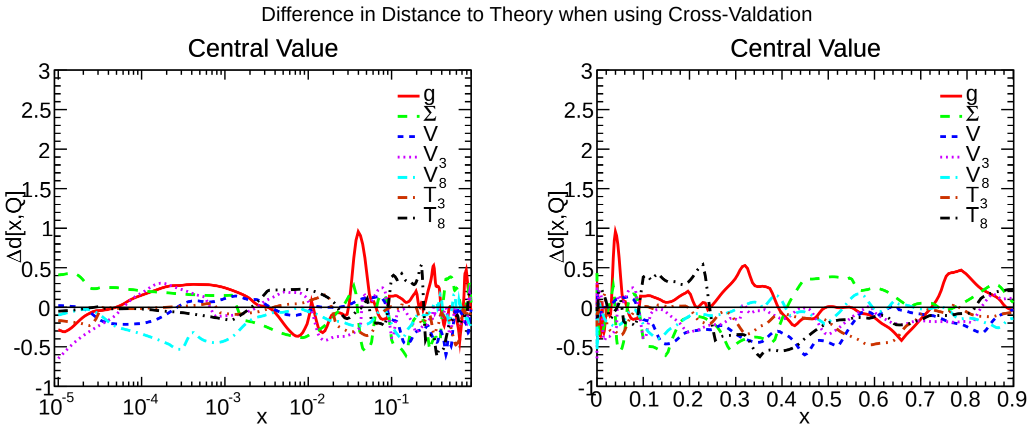


Figure 5.13: Difference in closure test distance between a fixed length fit and a fit with look-back cross-validation. The value indicates how much the distance changed in the cross-validated fit compared to the distance for the fixed length fit. The distances in both cases are to the MSTW2008 NLO PDFs which were used to generate the closure test data, and were calculated at the initial fitting scale. The definition of closure test distance (which is slightly different from the distance shown in Fig 5.12 etc.) is given in Appendix A.

is large enough, we expect results based on the look-back method to be independent of its precise value. For this to be the case, N_{gen} needs to be large enough that the fit has passed its global minimum value of the validation χ^2, or is sufficiently close to it that running for more generations would not greatly change the final PDFs.

For the fits described above, N_{gen} was set to 30000 generations. To verify that this is sufficiently large, a second level 2 closure test fit with look-back cross-validation was performed using the same settings and random seed, but with an increased maximum length of 80000 generations. From the distances shown in Fig. 5.14, it can be seen that both the central values and uncertainties for all PDFs are unchanged by running for a longer time. We can therefore rule out any sizeable dependence on the total training length in our current results, and we can stick to a baseline maximum number of generations of 30000 for the fits to real data.

Fig. 5.15 shows the distribution of stopping points for the 100 replicas in the 30000 and 80000 generation closure test fits. In both cases the stopping points are fairly evenly spread about the fit, but with a build-up towards the end, which indicates that many replicas do actually improve during the longer fitting length, though from the distances in Fig. 5.14 we know it is not by much. On the other hand, the spike at the beginning of the 80000 generation plot contains roughly the same number of replicas as are in the equivalent span in the 30000 generation plot, indicating that these twenty or so replicas have properly stopped, and likely have χ^2 profiles similar to the left-hand plot in Fig. 5.11.

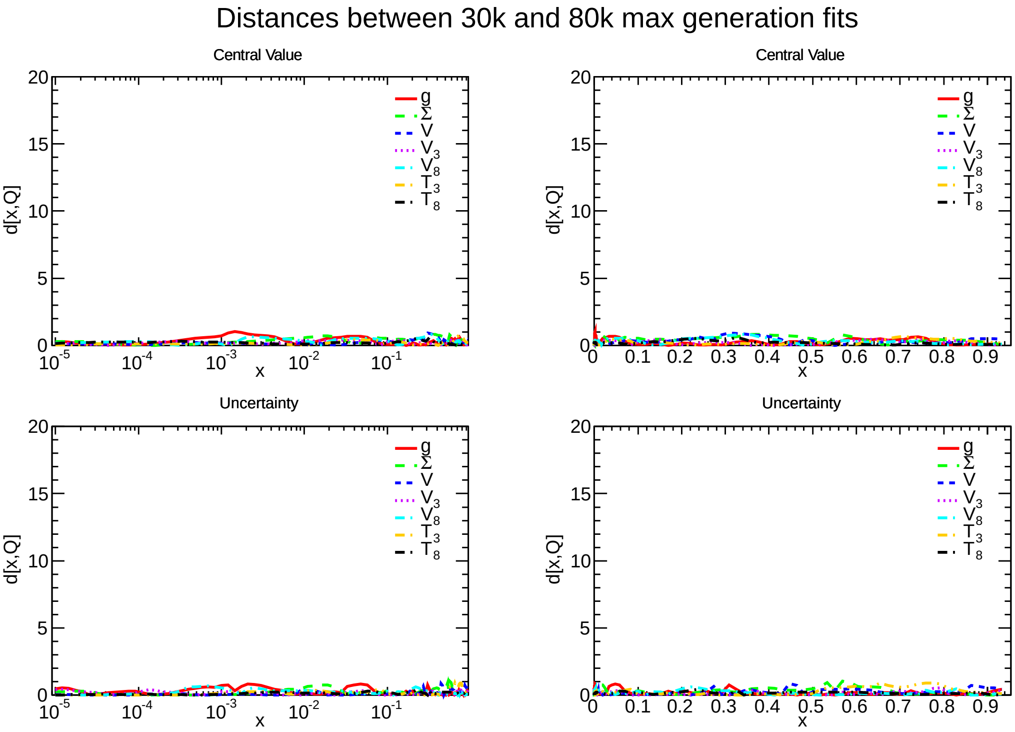


Figure 5.14: Distances between central values and uncertainties of two level 2 closure test fits with maximum numbers of generations set to 30k and 80k, with all other fit and dataset settings identical.

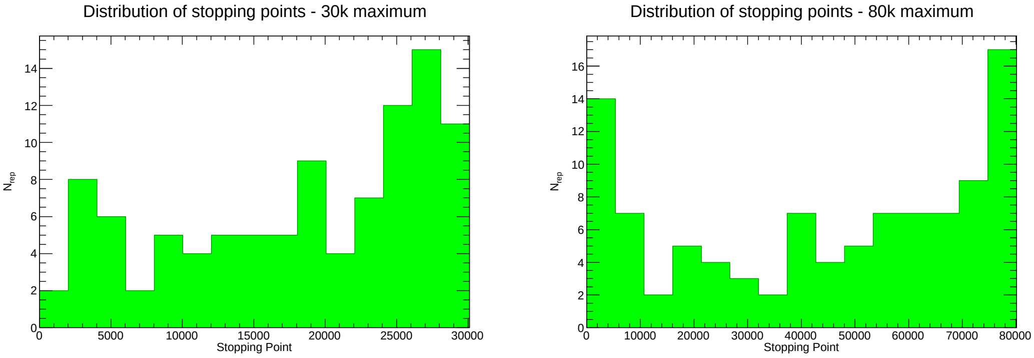


Figure 5.15: Distribution of the final generation chosen by look-back cross-validation for 30k and 80k generation 100 replica level 2 closure test fits. Note that the number of bins is the same in each plot, but the 80k generation fit has a much larger range, so in the right plot each bin covers more generations.

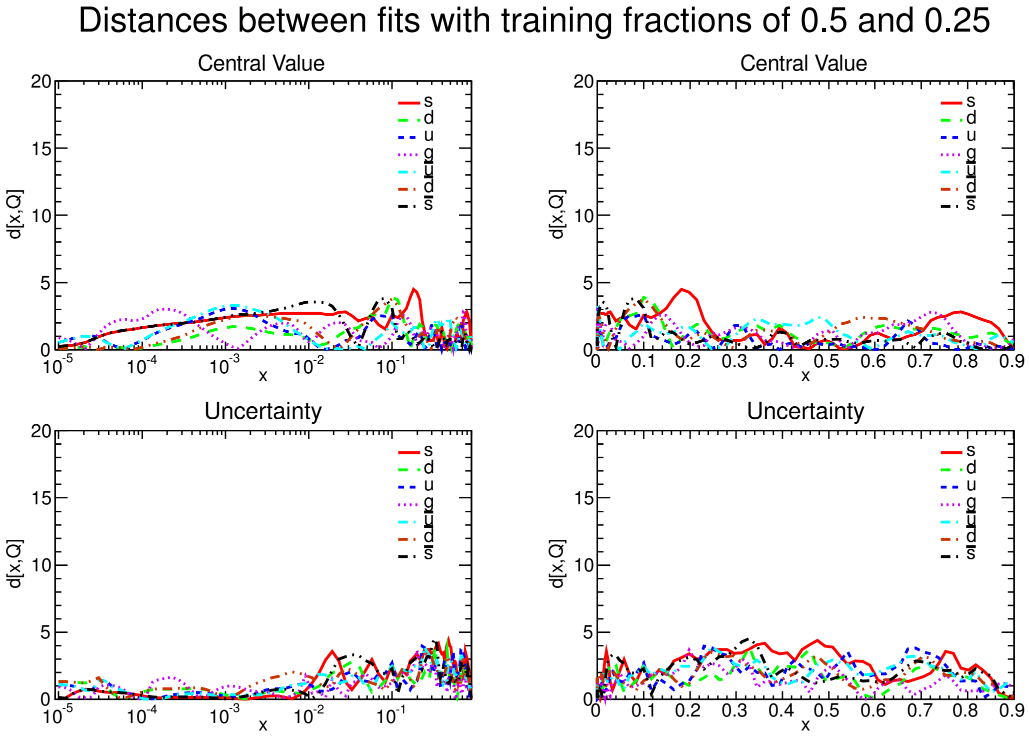


Figure 5.16: Same as Fig. 5.14 for the level 2 closure fits based on training fractions of 25% and 50%.

Dependence on the training fraction

As mentioned previously, cross-validation requires us to separate the fitted dataset into two disjoint subsets: the training set and the validation set. In the fits shown so far the fraction of data in the training and validation sets was set to one half of the total. Obviously by only training half the data we lose information in each replica fit, so it's possible that our results might change if this fraction is varied from 50%. In particular, it is important to check both that the 50% training fraction is enough to retain all the relevant information contained in the original dataset, and that using a smaller value will negatively impact the fit.

In order to study the impact on the fitted PDFs of the use of a different training fraction, I have produced a pair of level 2 closure fits with identical settings to the standard 50% fit except for the size of the training fraction: one with a smaller fraction of 25% and another with a larger fraction of 75%. The distances comparing these fits with alternative values of the training fraction to the standard fit are shown in Figs. 5.16 and 5.18. The first set of distances indicate that when the training fraction is reduced to 25% the central values of the PDFs are more or less the same, while the uncertainties on the fitted PDFs slightly increase. This suggests that, for the NNPDF3.0 dataset,

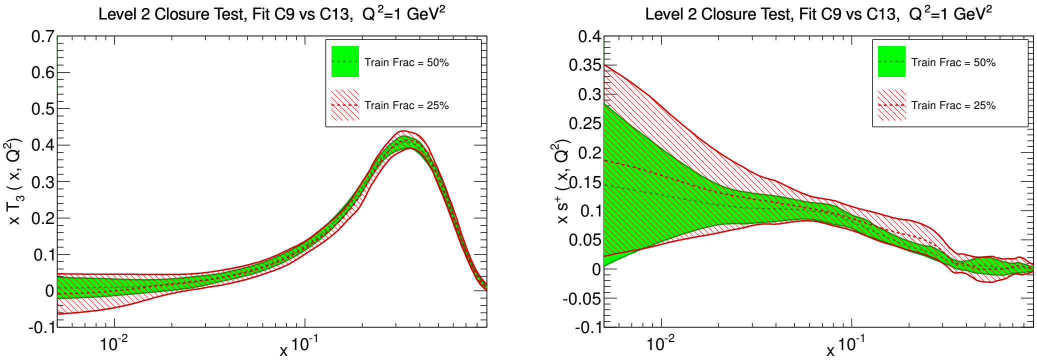


Figure 5.17: Comparisons between PDFs from the Level 2 closure fits performed using training fractions of 25% and 50%. They shown the quark triplet $xT_3(x, Q_0^2)$ (left plot) and the total strangeness $xs^+(x, Q_0^2)$ (right plot) at the initial parameterisation scale of $Q_0^2 = 1$ GeV^2.

some of the data removed when using the smaller training fraction is not redundant, and information is lost. The effect of the reduced training fraction can be identified more directly by looking at the PDFs in the two fits, shown in Fig. 5.17, where the increase in the PDF uncertainties is clear.

On the other hand, as can be seen from the distances in Fig. 5.18, the fits with training fractions of 50% and 75% are much more alike, and are effectively statistically indistinguishable. We can thus conclude that the loss of information due to the splitting of the dataset required by the cross-validation procedure is small provided the training fraction is above 50%, but using smaller training fractions has a larger impact.

5.5.2 Weight Penalty

Cross-validation is a common and straightforward approach to controlling overfitting in neural network training, however it is not the only approach, nor is it always the most effective. A number of others exist, many of which avoid the major problem of only being able to fit part of dataset. In this section I will describe a technique called *Weight Penalty* training[2], which aims to prevent overfitting during the training by penalising networks which model more complicated functions. This is essentially the same as setting a prior distribution for the probability of the parameters of the network, and so the method works in a similar way as Bayesian model selection.

[2] This approach is more usually called *Weight Decay* in much of the machine learning literature. Confusingly, however, the method I will describe after this (in Section 5.5.3) is also often described as Weight Decay, so here I will use Weight Penalty instead.

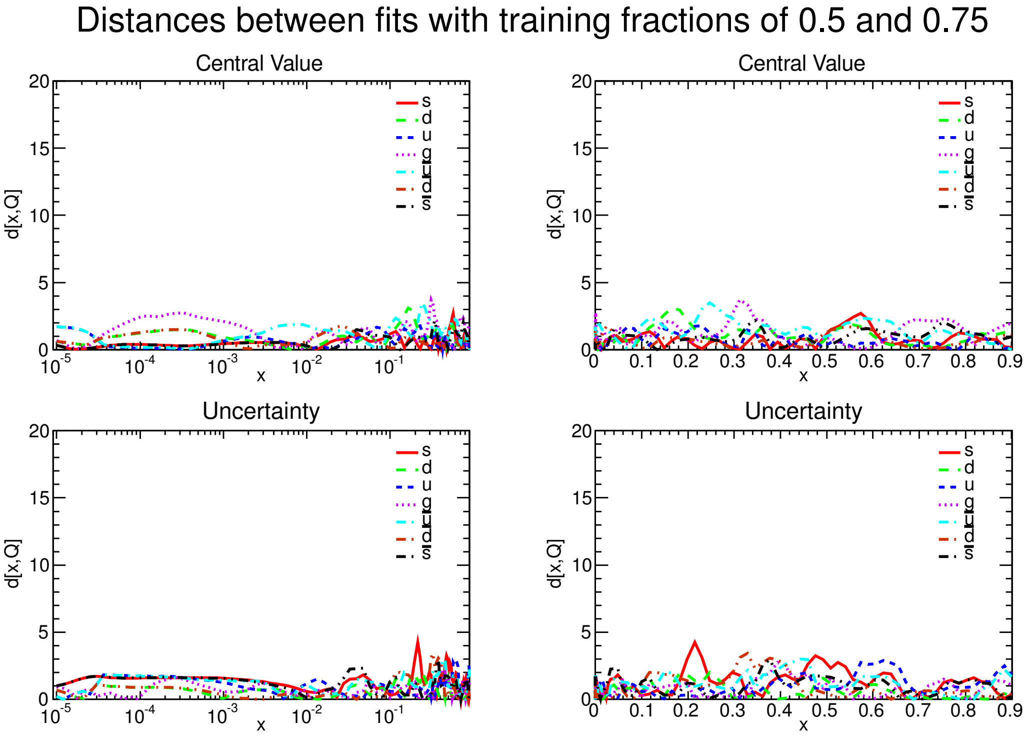


Figure 5.18: Same as Fig. 5.16 but training fractions of 50% and 75%.

Theory

The general idea of the weight penalty method is to include an extra term in the goodness of fit function minimised during the training which depends on the size of the neural network weights. Instead of evaluating the mutants in the genetic algorithm according to just the χ^2, we instead use

$$E_{\text{tr}}(\boldsymbol{d},\boldsymbol{t}(\boldsymbol{w})) = \chi^2(\boldsymbol{d},\boldsymbol{t}(\boldsymbol{w})) + \alpha\Delta(\boldsymbol{w}), \tag{5.14}$$

where $\boldsymbol{d}$ are the data points, $\boldsymbol{t}$ are the theoretical predictions, $\boldsymbol{w}$ are the network parameters they depend on, α is an external parameter which controls the strength of the penalty and $\Delta(\boldsymbol{w})$ is a selected function of the neural network weights (see below). Including this extra term encourages the training to eliminate weights in the network not being used to fit the data, which reduces the effective number of parameters and the complexity of the produced function. This results in a trade-off during the fit between the closeness of predictions to the data-points and the smoothness of the functions, and so can prevent overfitting.

There are multiple reasonable choices for the penalty function $\Delta(\boldsymbol{w})$ used in the fit. One very common one, which we will use here, is based on the idea of a Gaussian prior distribution for the weights:

$$\Delta(\boldsymbol{w}) = \sum_{i}^{N_w} w_i^2, \tag{5.15}$$

where N_w is the total number of weights. The penalty term in Eq. 5.14 also includes an overall size α, which must be chosen in order to achieve a balance between reducing complexity and fitting the data. In many implementations this tuning is done by hand in an ad hoc way; we will instead follow the *Bayesian Regulation* approach outlined in [174], where α is automatically determined based on the results of previous fits.

The idea of the automated process is that the most probable value for α can be found by taking the derivative of the probability distribution of α given the data, i.e. by setting

$$\frac{\partial}{\partial \alpha} P(\alpha|\boldsymbol{d}) = 0. \tag{5.16}$$

Using Bayes' theorem we can then write $P(\alpha|\boldsymbol{d})$ as

$$P(\alpha|\boldsymbol{d}) = P(\boldsymbol{d}|\alpha)\frac{P(\alpha)}{P(\boldsymbol{d})}. \tag{5.17}$$

If we assume that the prior probabilities for α and the data are uniform—so that $P(\alpha)$ and $P(\boldsymbol{d})$ are constant—then $P(\alpha|\boldsymbol{d})$ is proportional to $P(\boldsymbol{d}|\alpha)$, and the condition for the most probable α with $P(\alpha|\boldsymbol{d})$ in Eq. 5.16 applies equally well with $P(\boldsymbol{d}|\alpha)$.

We can expand $P(\boldsymbol{d}|\alpha)$ as

$$P(\boldsymbol{d}|\alpha) = \int d^{N_w}w \; P(\boldsymbol{d}|\boldsymbol{w}\alpha) \; P(\boldsymbol{w}|\alpha). \tag{5.18}$$

From this we can identify $P(\boldsymbol{d}|\boldsymbol{w}\alpha)$ as the probability of the data given the parameters and α, which can be written as

$$P(\boldsymbol{d}|\boldsymbol{w}\alpha) \equiv P(\boldsymbol{d}|\boldsymbol{w}) = Z_D \; e^{-\frac{1}{2}\chi^2(\boldsymbol{d},\boldsymbol{w})}, \tag{5.19}$$

where we have used the fact that the probability is independent of α. The normalisation Z_D is given by

$$Z_D^{-1} = \int D\boldsymbol{d} \; e^{-\frac{1}{2}\chi^2(\boldsymbol{d},\boldsymbol{w})}, \tag{5.20}$$

which we cannot evaluate in the general case, but we can see is independent of α. Likewise, $P(\boldsymbol{w}|\alpha)$ is just the prior probability of the weights for a given value of α, and can be written as

$$P(\boldsymbol{w}|\alpha) = Z_w \; e^{-\frac{1}{2}\alpha\Delta(\boldsymbol{w})} \tag{5.21}$$

where this time the normalisation Z_w can be evaluated using Eq. 5.15 to give

$$Z_w = \int d^{N_w}w\; e^{-\frac{1}{2}\alpha \sum\limits_i^{N_w} w_i^2} = \left(\int dw\; e^{-\frac{1}{2}\alpha w^2}\right)^{N_w} = \left(\frac{\alpha}{2\pi}\right)^{\frac{N_w}{2}} . \tag{5.22}$$

Taking these definitions together with Eqs. 5.16 and 5.18 we can see that for the most probable value of α

$$\frac{\partial}{\partial\alpha}P(\boldsymbol{d}|\alpha) = \int d^{N_w}w\; P(\boldsymbol{d}|\boldsymbol{w}\alpha)\;\frac{\partial}{\partial\alpha}P(\boldsymbol{w}|\alpha) = 0 \tag{5.23}$$

$$\implies \int d^{N_w}w\; P(\boldsymbol{d}|\boldsymbol{w}\alpha)\;\left(\frac{N_w}{2\alpha} - \frac{1}{2}\Delta(\boldsymbol{w})\right) P(\boldsymbol{w}|\alpha) = 0 \tag{5.24}$$

$$\implies \frac{N_w}{2\alpha}P(\boldsymbol{d}|\alpha) = \frac{1}{2}\int d^{N_w}w\; \Delta(\boldsymbol{w})\; P(\boldsymbol{d}|\boldsymbol{w}\alpha)\; P(\boldsymbol{w}|\alpha) \tag{5.25}$$

$$\implies \alpha_{best} = \frac{N_w}{\langle\Delta(\boldsymbol{w})\rangle}, \tag{5.26}$$

where we used Bayes' theorem again between the third and fourth lines. Here $\langle X\rangle$ means the expected value of X for the given distribution of weights, i.e. $\langle X\rangle = \int d^{N_w}wXP(\boldsymbol{w}|\boldsymbol{d}\alpha)$, which for our PDF fits is represented by averaging over X calculated for each replica. Since the expected value depends on the value of α itself, we will need to iterate until the value calculated using the replica PDFs at the end of the fit matches the starting α. Note that this condition is the same as saying that at the end of the fit we want the numerical contribution of the penalty term to the error function to be equal to the number of parameters in the fit.

In the literature on this method, it is suggested that different values of alphas are used for the separate layers of the network, and for weights on the connections from each input node [175]. We follow that approach in our implementation, in addition to using separate α values for each PDF. The actual error function minimised during neural network training is therefore

$$E_{\rm tr}(\boldsymbol{d},\boldsymbol{w}) = \chi^2(\boldsymbol{d},\boldsymbol{w}) + \sum_i^{N_{\rm pdf}}\sum_j^{N_{\rm cat}}\left(\alpha_{ij}\sum_k^{N_w} w_{(ij)k}^2\right), \tag{5.27}$$

where $N_{\rm cat}$ is the number of different categories of weights, which for our networks is four.

Results

With this setup, we performed a number of level 2 closure test fits using a weight penalty in the error function. Each fit was performed using the final NNPDF3.0 settings

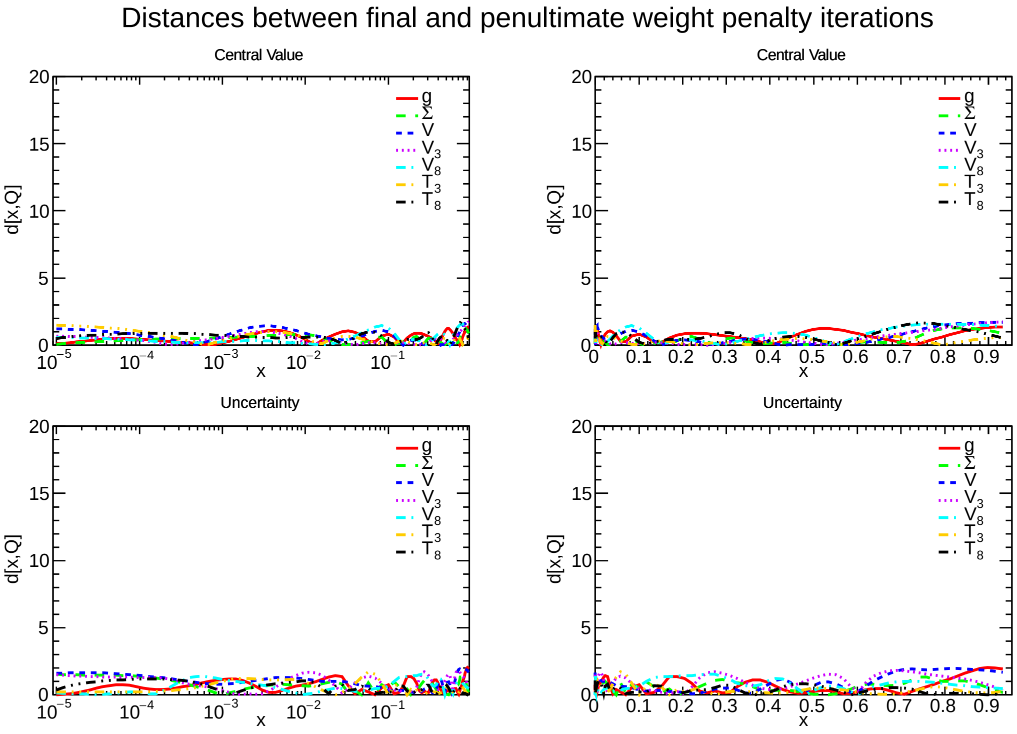


Figure 5.19: PDF distances for central values and uncertainties between the final and penultimate weight penalty iterations. Each fit was a level 2 closure test fit to data generated using MSTW PDFs. The distances were calculated at the initial fitting scale of $Q^2 = 1$ GeV^2.

(without cross-validation), with a maximum length of 30000 generations. As mentioned above, the values of α_{ij} used in the fits needed to be iterated, with each final fit shown here using values calculated from the results of a chain of previous fits.

The first thing it was necessary to establish was that the α_{ij} would actually converge, and how rapidly this would occur. Starting from values calculated from a fit with the penalties set to 0, i.e. without the penalty, fits were performed sequentially. We found that convergence was initially very fast, but slowed as the values approached their fixed point. Fortunately, we also discovered that the fit is largely insensitive to the precise value of α_{ij}, and that similar results are obtained for values within about a factor of two. For this reason we choose our condition for determining convergence to be quite broad, allowing for a change of at most 40% in the α_{ij} values between generations, as long as differences between the PDFs themselves were also small. The distances between fits with the final and penultimate α_{ij} settings are shown in Fig. 5.19. For both uncertainties and central values the distances are uniformly below 2, indicating that the fits are statistically equivalent with differences about the size we would expect from changing genetic algorithm random seed. The final values for α_{ij} are shown in Table 5.8, along with the average size of weight for each PDF and category. The

	Input (x)		Input ($ln(x)$)		Hidden		Output		Total
PDF	α	$\overline{\lvert w\rvert}$	α	$\overline{\lvert w\rvert}$	α	$\overline{\lvert w\rvert}$	α	$\overline{\lvert w\rvert}$	$\overline{\lvert w\rvert}$
g	0.045	3.39	0.102	2.28	0.054	2.99	0.042	3.27	2.98
Σ	0.083	2.60	0.893	0.84	0.121	2.05	0.057	3.10	2.08
V	0.079	2.75	0.282	1.51	0.114	2.14	0.050	2.83	2.22
V_3	0.148	2.01	0.629	1.04	0.148	1.88	0.088	2.51	1.85
V_8	0.132	2.13	0.833	0.90	0.199	1.63	0.088	2.53	1.70
T_3	0.065	2.96	0.435	1.23	0.159	1.76	0.065	2.85	1.97
T_8	0.121	2.14	0.354	1.35	0.106	2.19	0.045	3.34	2.20

Table 5.8: Final values for penalty strength α used in last iteration of weight penalty closure test fits to level 2 data generated using MSTW PDFs. Values are shown by PDF and weight category. The average weight magnitude (i.e. ignoring the sign) $\overline{|w|}$ is also shown in each case, as well as for each PDF in total.

gluon has the largest weights and smallest α_{ij}, which is unsurprising given that it has somewhat more structure than the other PDFs.

Having obtained the final α settings, we can then look at the overall effect of including the penalty term in closure test fits. Fig. 5.20 shows the distances between a fit with weight penalty and a fixed length fit without it but otherwise using equivalent settings. We can see that the distances in the central value are mostly below about 6, showing that the weight penalty fit is largely consistent with the fixed length fit, though with some significant discrepancies for instance in the V_3 distribution at medium-x and V_8 at small-x. This reaffirms the conclusion that we drew from the results on cross-validated fits, that there is only minimal over-learning during the fit. Compared to the equivalent plot for look-back cross-valdiation (Fig. 5.12) we can see that the weight penalty method has a much larger impact on the uncertainties of the fit, with distances of over five in places. Looking further into the results it becomes clear that the weight penalty uncertainties are generally smaller that the fixed length determination. We can understand this by interpreting the weight penalty as an additional constraint that the PDFs should be smooth, which provides additional information to the fit about the PDFs. Fig. 5.21 shows the two of the PDFs themselves for the closure tests fits with and without weight penalty. The reduction in uncertainties suggested by the distance plots can be seen by comparing the size of the red and green bands.

We can also look at the difference in distances to the underlying MSTW PDFs used to generate the closure test data, as we did for the cross-validated fit. These are shown for the weight penalty fit in Fig. 5.22. Compared to the equivalent plot for look-back cross-validation in Fig. 5.13, weight penalty has a considerably larger impact on the quality of the prediction. In some places it provides an improvement, especially at large x, but there are also regions where the weight penalty fit reproduces

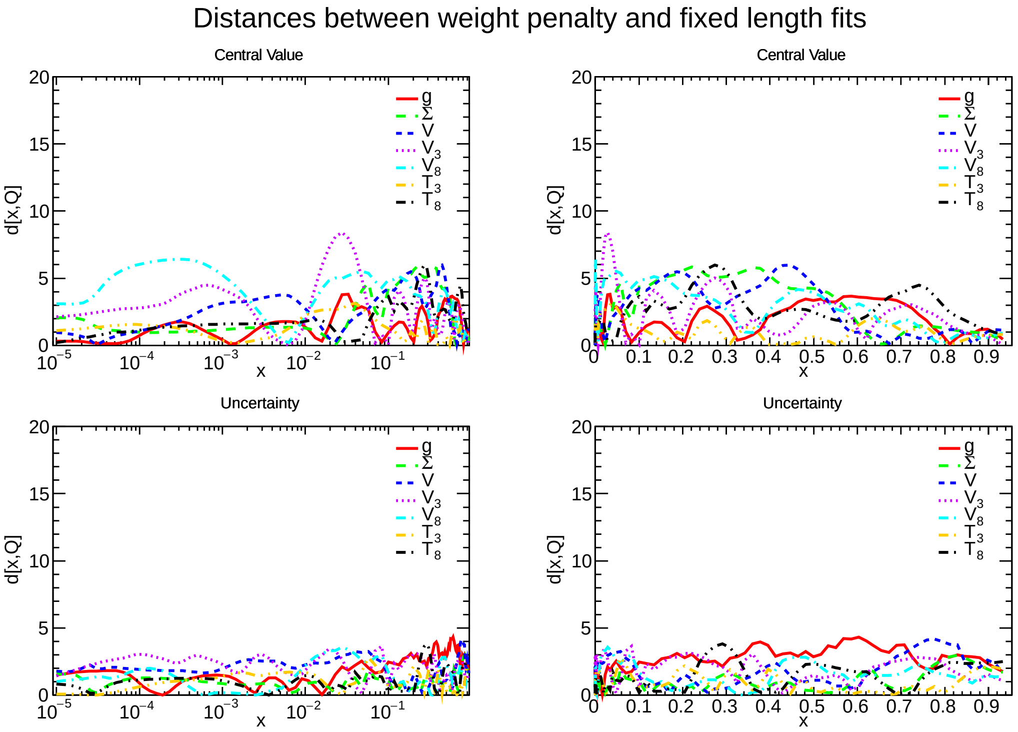


Figure 5.20: Same as Fig. 5.19 but between the final weight penalty iteration and a fixed length fit (i.e. a fit without weight penalty) with otherwise identical settings.

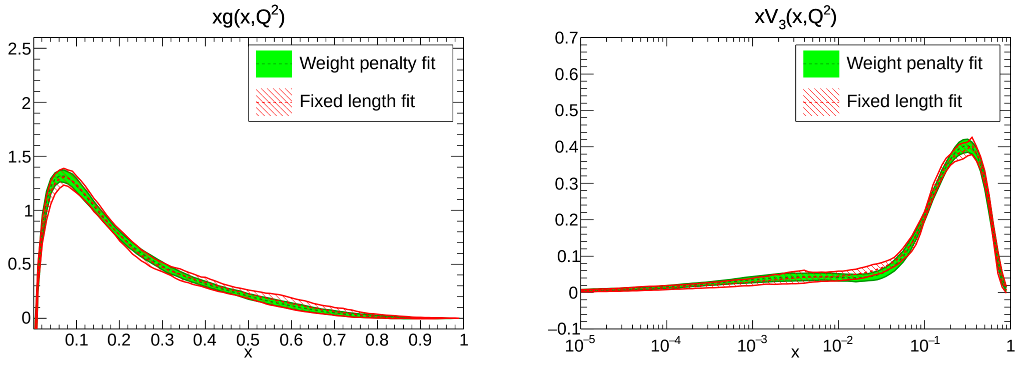


Figure 5.21: PDFs (multiplies by x) for the gluon and V_3 distributions of the final weight penalty iteration and fixed length level 2 closure tests.

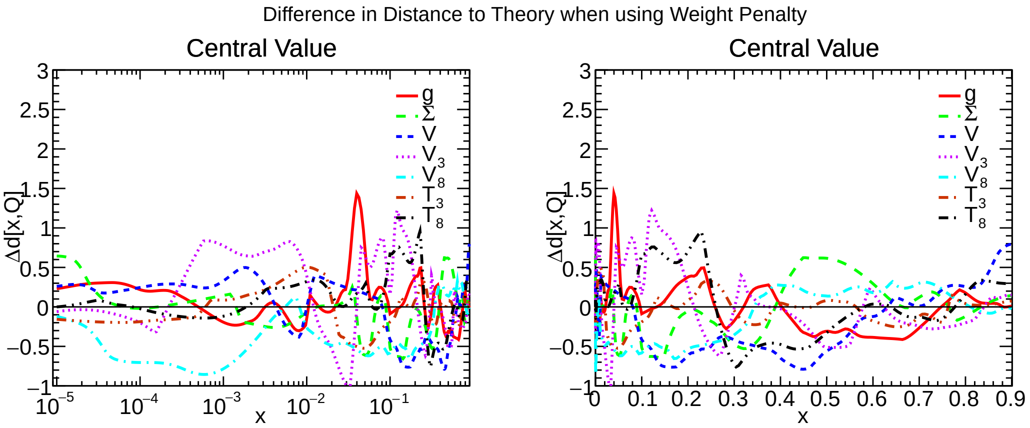


Figure 5.22: Same as Fig. 5.13 for fit with weight penalty.

the generating PDFs much more poorly than the fixed length fit.

We can also directly compare the PDFs from the weight penalty fit to their counterparts from the fit with look-back cross-validation. From Fig. 5.23 we can see that while the results of using a weight penalty are somewhat different to those obtained through cross-validation, they are closer to this fit than to the fit without any mechanism to stop over-learning (Fig. 5.20). This indicates that, while weight penalty and cross-validation work by very different principles, they have the same effect of controlling the small amount of overfitting present in the fit. However, the weight penalty method does not outperform cross-validation, and given the fact that it is the more complicated and ambiguous of the two methods we opted to use cross-validation in the final NNPDF3.0 fits.

5.5.3 Weight Decay

There are other methods which can be employed to obtain the same effect as the weight penalty approach outlined above. One such method, which I will describe as *Weight Decay* (see footnote on pg. 77), involves periodically shrinking all of the network weights by a small amount. The idea is that weights which are necessary to describe the data will be restored to their previous value by the minimisation, while weights which are superfluous will be naturally eliminated. If successful, this will therefore obtain the same effect as using a weight penalty, but without the need to perform multiple fits in order to iterate the α parameters. However, modifying the neural network parameters indiscriminately can easily disrupt the fit if the weights are decayed to greatly or too often.

Weight decay is straightforward to implement. It simply requires a step in the genetic algorithm after selection where all of the parameters are universally multiplied

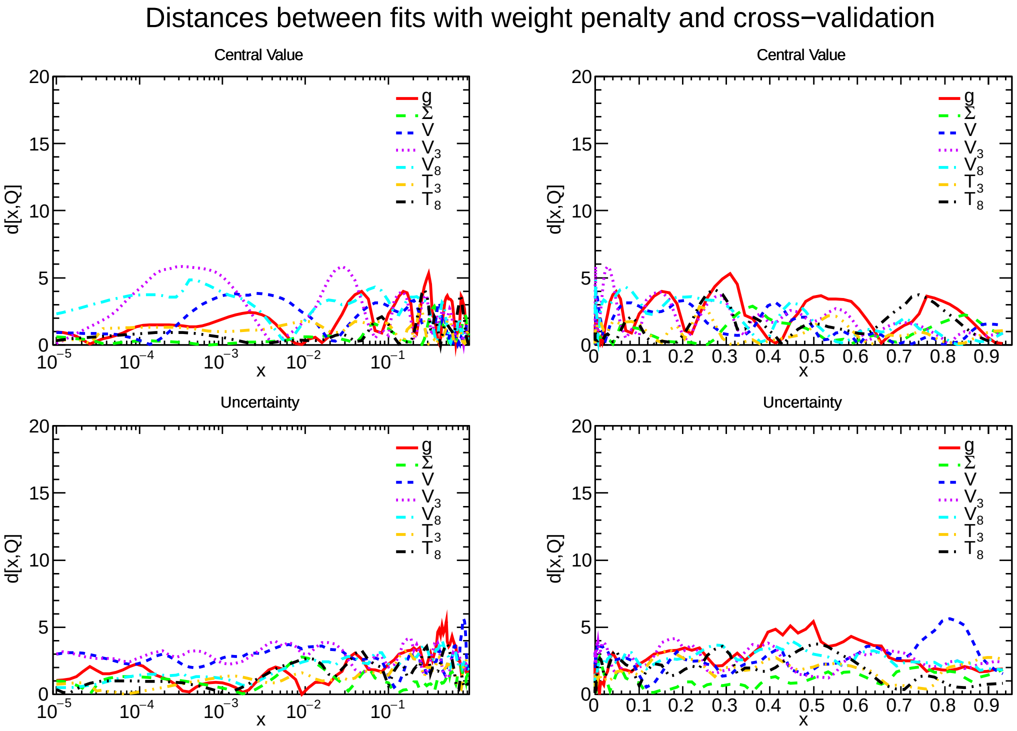


Figure 5.23: Same as Fig. 5.19 but between the final weight penalty iteration fit and a fit using cross-validation.

by a number slightly less than one. There are two main parameters in the approach: the frequency of decay and its strength. Early tests demonstrated that decaying the weights every generation give very poor results, even if the size of the decay is set to be relatively small. Instead, a more reasonable approach is to decay much less often but by a sizeable amount. Here I will present results using a decay of 2% (i.e. multiplying the weights by 0.98) every 100 generations. This corresponds to a reduction of the weights by more than 95% over 15000 generations, if mutations are ignored. In order to prevent unwanted bias from the final results, weight decay will not be used during the last 2000 generations.

In order to test the weight decay method, as for previous features I performed a level 2 closure test fit with pseudo-data generated using the MSTW2008 NLO PDF set. Fig. 5.24 shows the distances between the weight decay fit and a standard-settings fit without decay. We can see that the introduction of weight decay has a relatively small, though not negligible, impact on the fit results, with all distances below 5. However, looking at Fig. 5.25, which shows the change in distance to the MSTW PDFs from introducing the method, it is clear that weight decay does not improve accuracy of the prediction of the underlying distributions at large-x, and substantially worsens

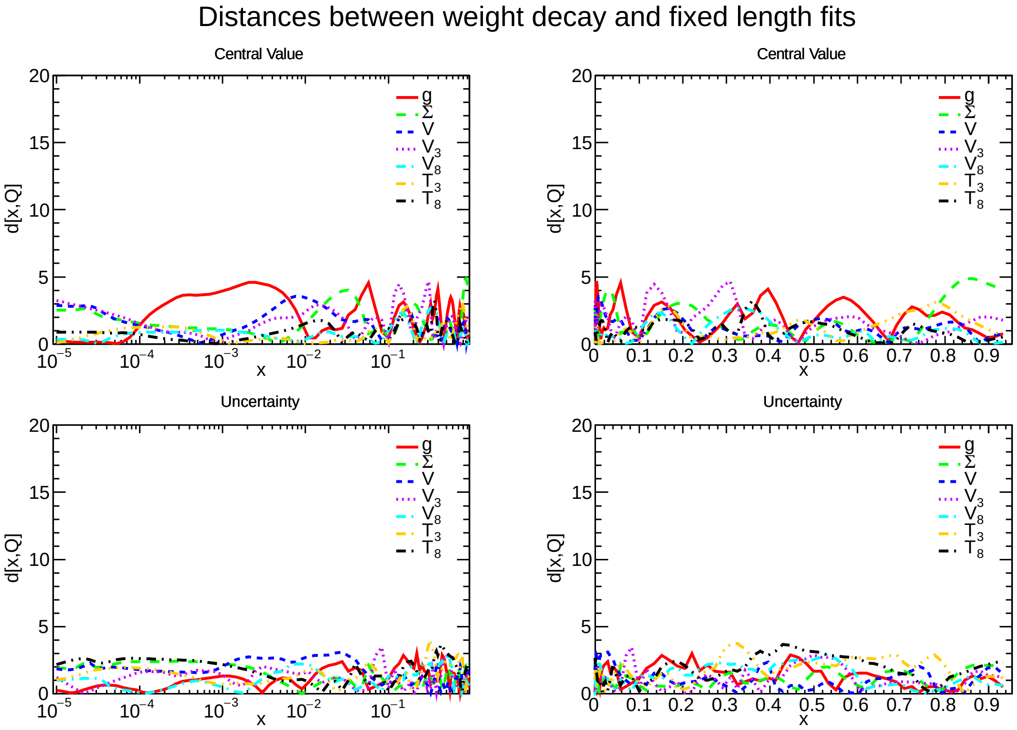


Figure 5.24: Same as Fig. 5.20 but for a fit using weight decay.

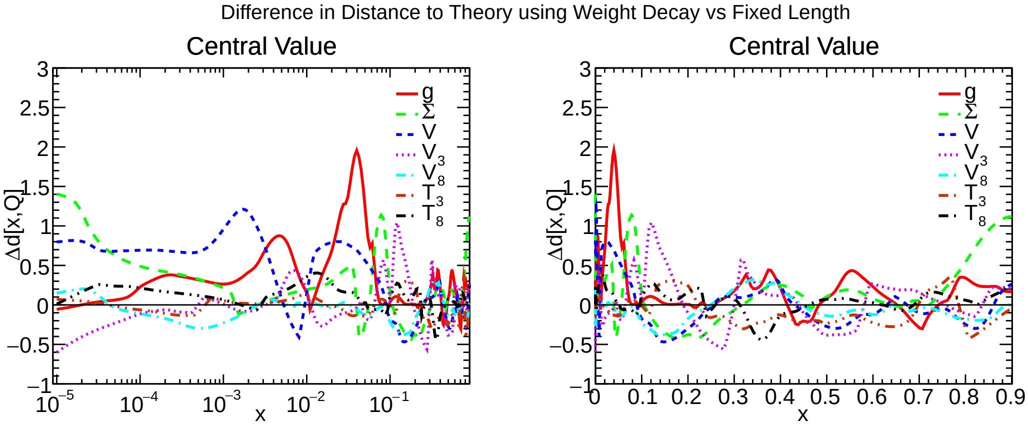


Figure 5.25: Same as Fig. 5.22 but for a fit using weight decay.

it across almost all PDFs at small-x. On this basis, weight decay can only provide another check that overfitting is small and that the results obtained are relatively robust against changes in methodology, but is inappropriate to use as part of the central fitting methodology. It is possible that tuning the size and frequency of the decay could improve the quality of the results, however it is not clear that this is case, and working on this would take time away from investigating more promising approaches.

5.6 Positivity constraints

As described in Section 3.2.6, in previous NNPDF fits we enforced PDF positivity by imposing constraints on the deep-inelastic structure functions F_L, F_2^c and of the neutrino charm production ("dimuon") cross-section. However, while these conditions were sufficient to guarantee the positivity of most physical observables, in order to ensure positivity of all observables for NNPDF3.0 we have increased the number and kinematic coverage of positivity constraints used in the fit. In particular, we have chosen to impose positivity of some pseudo-observables which must respect positivity for reasons of principle, but which are not measurable in practice. We choose the three tagged deep-inelastic structure functions F_2^u, F_2^d and F_2^s and the three flavour Drell-Yan rapidity distributions, $d\sigma^{\mathrm{DY}}_{u\bar{u}}/dy$, $d\sigma^{\mathrm{DY}}_{d\bar{d}}/dy$ and $d\sigma^{\mathrm{DY}}_{s\bar{s}}/dy$, to enforce generalised positivity of the quark and anti-quark distributions, and the light contribution to the longitudinal stucture function, F_L^l supplemented with the rapidity distribution $d\sigma^H_{gg}/dy$ for the production in gluon-gluon fusion of a Higgs-like scalar with mass $m_H^2 = 5$ GeV^2 to constrain the gluon. All these positivity constraints are imposed at $Q^2_{\mathrm{pos}} = 2$ GeV^2, and for $x \in [10^{-7}, 1]$, which, because of the structure of QCD evolution, ensures positivity at all higher scales. In practice we computed the observables at 20 points in the given x range, equally spaced on a log scale for $x < 0.1$ (ten points) and on a linear scale for $x \geq 0.1$.

As well as imposing the positivity constraints during the minimisation by means of a Lagrange multiplier, we have introduced a further constraint that the final fit result is negative in any pseudo-observable by at most 25% of an absolute value calculated using a fixed reference set, discarding any replica for which this is not the case. This condition is necessary for cross-sections which are very close to zero (e.g. close to kinematic boundaries, like the rapidity tails of Drell-Yan distributions) where the Lagrange multiplier strategy is not effective.

This strategy is used in both the NLO and NNLO fits, with the NNLO fits using also the NLO pseudo-observables, since at the low Q^2 values at which the positivity pseudo-observables are computed large unresummed NNLO corrections lead to perturbatively unstable predictions at large and small x. There is also some evidence that the resummed result is closer to NLO than to NNLO, see for example Ref. [176] for the case of deep inelastic structure functions. As described in Section 5.4.3, in the LO fits, where PDFs are strictly positive-definite, we use the same strategy with pseudo-observables now computed at LO and with a larger Lagrange multiplier. We have verified this is sufficient to ensure positive-definite PDFs, and also use a flag in the LHAPDF6 NNPDF3.0LO grids to force a positive-definite output.

The impact of the positivity constraints on the final PDFs is looked at in Section 7.3.2 where I compare two NNPDF3.0 NLO fits with and without the positivity constraints, and discuss further a posteriori checks of the implementation of the positivity conditions. I will also explore the impact of the improved positivity constraints on searches for high-mass new physics.

Chapter 6

Closure testing the NNPDF3.0 methodology

The previous chapter looked at studies of various new features in the NNPDF methodology using the closure testing technique. In this chapter I will subsequently focus on tests of the statistical properties of the final methodology, and how well generally our approach satisfies the closure test. Benchmarking the methodology is especially important now due to the substantial increase in experimental data included in NNPDF3.0, and the increased precision of the resulting PDFs. As data become more precise and their kinematic coverage increases, it becomes more and more important to eliminate as far as possible methodological uncertainties, and to verify that our results are statistically valid. As discussed in Section 5.2 NNPDF3.0 is the first PDF determination performed using a methodology based on closure tests, and here we will further look at using closure tests to study the statistical properties of the resulting PDFs.

The basic idea of the closure test, described briefly in Section 5.2, is simple [177,178]: we take a given assumed form for the PDFs (for example MSTW2008), a given theoretical model (for example NLO pQCD), and with them generate a set of global pseudo-data with known but realistic statistical properties (by using the covariance matrices of the real datasets that together make up, for example, the NNPDF3.0 dataset). These pseudo-data are then 'perfect', in the sense that they have known statistical properties, no internal inconsistencies, and are also entirely consistent with the theoretical model used to produce them. Thus if we then use our fitting methodology to perform a fit to these pseudo-data, we should reproduce the central value of the assumed underlying PDF, within correctly determined uncertainties.

First I will introduce notations for the various closure test quantities which will be used in this section, and look at the different options for pseudo-data production which are available. I will next study the efficacy of the final training methodology by looking at level 0 closure test fits. This will be followed by a study of PDF uncertainties, and at ways of investigating the validity of the obtained values and at contributions from different sources (data, functional and extrapolation). I will then look at the results from a full closure test—level 2 data fit with the final NNPDF3.0 methodology—specifically in terms of how well it reproduces the underlying law. All of the closure test fits in the previous chapter, and the majority in this chapter, have been performed using pseudo-data generated with the MSTW2008 PDFs, so in the final section of this chapter I will show results from closure tests using a range of different PDF sets.

6.1 NNPDF closure testing

The new framework used in NNPDF3.0 for the computation of observables provides us with the ideal tool to successfully implement closure tests. In particular, the clean separation between theoretical assumptions and input PDFs allows us to generate pseudo-data using a given set of PDFs and the experimental covariance matrix as an input, and to perform a fit to this pseudo-data using exactly the same theoretical settings (encoded in the FK tables) that were used for generating them.

Throughout this chapter I shall refer to the parton distributions used to generate the pseudo-data as the *input* PDFs, and denote them by f_{in}. Any PDF set available through the LHAPDF interface [179] can be used as an input set to generate the pseudo-data. Most of the closure tests described here will be performed using MSTW2008 NLO PDFs, though I will present some results from test with other input PDFs at the end of the chapter. We denote the set of pseudo-data by $\mathcal{D} = \{d_i\}$; the dependence of the pseudo-data on the input PDFs f_{in} and experimental covariance matrix will be left implicit.

The outcome of the closure test fits is then a set of fitted PDFs f_{fit}, which we will compare to the input PDFs in order to study the statistical precision and possible systematic biases in the fitting methodology. For any PDF set f, whether input or fitted, the FastKernel framework delivers a set of theoretical predictions, $\mathcal{T}(f) = \{t_i(f)\}$, based on a particular theoretical model. As in the previous chapter we will in general use NLO perturbative QCD, precisely as implemented in the NNPDF3.0 fits to real data, with the same parameter choices (e.g. quark masses, α_S) and so on. Also, we will continue to omit the positivity constraints for the reasons outlined previously.

In terms of these definitions, this χ^2 minimised during the fit can be written as

$$\chi^2[\mathcal{T}(f), \mathcal{D}] = \frac{1}{N_{\mathcal{D}}} \sum_{i,j} (t_i(f) - d_i)\, C^{-1}_{ij}\, (t_j(f) - d_j)\,. \tag{6.1}$$

In this expression, C_{ij} is the (t_0) covariance matrix of the data and $N_{\mathcal{D}}$ is the total number of data points of the dataset. Note that when fitting the pseudo-data we use exactly the same procedure (with exactly the same code) as for fits to real data, with the only difference being the values of the data points. Since in the closure test fits the 'correct' solution is known—and are the PDFs f_{in} used as input—the result f_{fit} of the fit should ideally reproduce the input PDFs within the statistical uncertainties of f_{fit} as determined by the fit.

As mentioned in the previous chapter, we can introduce three distinct categories of closure tests depending on the amount of stochastic noise added to the pseudo-data points generated from the initial PDFs. In order to make these tests as realistic as possible, this stochastic noise is generated using the complete information in the experimental covariance matrix, so the fluctuations and correlations of the pseudo-data reproduce precisely those of the real experimental data. For the closure tests presented in this section, the pseudo-data is in one-to-one correspondence with the experimental data used in the global fit, i.e. we have generated pseudo-data for every point in the NNPDF3.0 global dataset described in Table 4.1.

The three levels of closure test that we will study, which we call level 0, level 1 and level 2 are set up as follows:

- **Level 0**. Pseudo-data $\mathcal{D}_0 = \{d_i^0\}$ are generated without adding any stochastic noise. We then perform N_{rep} fits, each to exactly the same set of pseudo-data (i.e. without generating replica datasets), but using different random seeds for the initialisation of the random numbers used in the minimisation. This yields an ensemble of PDF replicas $\{f_{\text{fit}}^k\}$, where $k = 1, \ldots, N_{\text{rep}}$.

  Note that the error function which we minimise (given by Eq. 6.1) is still computed using the covariance matrix of the data, even though the level 0 pseudo-data are precisely the theory value and so have no uncertainty. While this will effect the way the parameter space is seen by the genetic algorithm, as it essentially gives the data points different weights, the minimum will be unchanged (as it is at $\chi^2 = 0$), and including the experimental correlations means that the level 0 pseudo-dataset contains the same total amount of independent information as at other levels and with real data.

  It should be clear from its definition that in Level 0 closure tests, the fit quality can be arbitrarily good, provided we use a sufficiently flexible PDF parametrisation and a sufficiently efficient minimisation algorithm. Indeed, since by construction the pseudo-data does not have any stochastic noise, and there are no inconsistencies, there exist *perfect* fits to the Level 0 pseudo-data that have a vanishing χ^2. Note that the best fit is not unique, and there will be an infinity of fits which lead to vanishing χ^2 by going through all data points, but differ in the way they interpolate between data points. These optimal solutions to the minimisation problem reproduce precisely the predictions of the set of PDFs used as input in the generation of the pseudo-data at each of the experimental data points. With our genetic algorithm, however, we are unlikely to generate a perfect solution exactly. Instead we expect that as the fit proceeds, the best fit PDFs should approach the ideal solution, and the value of the error function should approach zero.

- **Level 1**. Here we add one set of stochastic fluctuations on top of the level 0 pseudo-data, similarly to how replica datasets are generated at the beginning of normal fits (Eq. 3.11):

$$d_i^1 = \left[\prod_j^{N_{\text{mult}}} \left(1 + r_j^{\text{mult}} \sigma_{i,j}^{\text{mult}}\right)\right] \left(d_i^0 + \sum_k^{N_{\text{add}}} r_k^{\text{add}} \sigma_{i,j}^{\text{add}} + r_i^{\text{stat}} \sigma_i^{\text{stat}}\right), \tag{6.2}$$

where again $\sigma_{i,j}^{\text{add}}$, $\sigma_{i,j}^{\text{mult}}$ and σ_i^{stat} are the additive and multiplicative systematic, and statistical uncertainties for each data point, and the random numbers $r_{i,j}^{\text{add}}$, $r_{i,j}^{\text{mult}}$ and r_i^{stat} are generated according to unit variance normal distributions. These shifted data points represent the measured values of hypothetical experiments with the same statistical and systematic uncertainties as the real data.

From its definition, with one level of stochastic fluctuation, we expect that in level 1 closure tests the error function (which as at Level 0 coincides with the t_0 χ^2 per degree of freedom, i.e. $\chi^2[\mathcal{T}(f), \mathcal{D}_1]$) of the best fit will be around one. There is also an 'ideal' value that we want to obtain from a level 1 fit, which is the value of the error function calculated with the input PDFs. In practice the fitted PDFs will have a slightly lower value than the input PDFs as, depending on the random seed, it is likely that there will exist functions which are more likely given that particular set of pseudo-data.

Adding the single layer of stochastic fluctuations to the pseudo-data can be performed at two stages in the fit: the pseudo-data generation and the artificial replica dataset generation. In the first case, a single set of level 1 data is generated and used (as in level 0 fits) for all replicas, whereas in the second each replica is run with a different level 1 dataset. The former type of level 1 closure test is useful for looking at the error propagation, as the difference between this level and level 2 fits is that the experimental uncertainties are not propagated through to the PDF uncertainties. In these tests we therefore expect that the PDF uncertainties will be underestimated, and can be compared to the uncertainties obtained in level 2 fits. The other level 1 case is also useful for a particular estimator of uncertainties, as it can be used to approximate a set of central values of level 2 fits, as will be described later.

- **Level 2.**

  At this level stochastic fluctuations are added both during pseudo-data and replica generation, i.e. starting from the shifted pseudo-data in Eq. 6.2, we generate $N_{\rm rep}$ Monte Carlo replicas datasets $\mathcal{D}_2^l = \{d_i^{2,l}\}$ with

$$d_i^{2,l} = \left[\prod_j^{N_{\rm mult}} \left(1 + r_j^{{\rm mult},l}\sigma_{i,j}^{\rm mult}\right)\right]\left(d_i^1 + \sum_k^{N_{\rm add}} r_k^{{\rm add},l}\sigma_{i,j}^{\rm add} + r_i^{\rm stat}\sigma_i^{{\rm stat},l}\right), \qquad (6.3)$$

  for $l = 1, \ldots, N_{\rm rep}$, with different random numbers for each replica. From the practical point of view, once we have generated a set of level 1 pseudo-data Eq. 6.2, the level 2 $N_{\rm rep}$ Monte Carlo pseudo-data replicas Eq. 6.3 are obtained using exactly replica generation process as is used for the fits to real data.

  In level 2 fits, each Monte Carlo replica represents a fluctuation around the level 1 pseudo-data, and the procedure should correctly propagate the fluctuations in the pseudo-data, due to the experimental statistical and systematic uncertainties, into the fitted PDFs. The fit to each data replica yields a PDF replica $f_{\rm fit}^l$, and the ensemble of PDF replicas then contains all the information on PDF uncertainties and correlations. We expect the final error function of a Level 2 fit to be two, since each replica dataset has been fluctuated twice, while the χ^2 per degree of freedom of the replica PDFs to the original pseudo-data (i.e. $\chi^2[\mathcal{T}[f_{\rm fit}^l], \mathcal{D}_1]$) will be close to one. Again the actual 'ideal' value will not be precisely these values due to random fluctuations, and will be given by the error function for the input PDFs to the pseudo-data. Moreover, for a correctly determined set of fitted PDFs, we expect the input PDFs $f_{\rm in}$ to lie within the one-sigma band of the fitted PDFs with a probability of 68%.

6.2 Validation of the training efficiency: Level 0 closure tests

Here I will present the results of a number of level 0 closure tests using the final methodology settings given in Chapter 5, and use them to assess the training efficiency of the NNPDF3.0 minimisation. In level 0 fits there exists a number of optimal solutions for the minimisation where the error function to the pseudo-data is reduced to zero. With level 0 closure tests we can therefore perform tests of different approaches to the

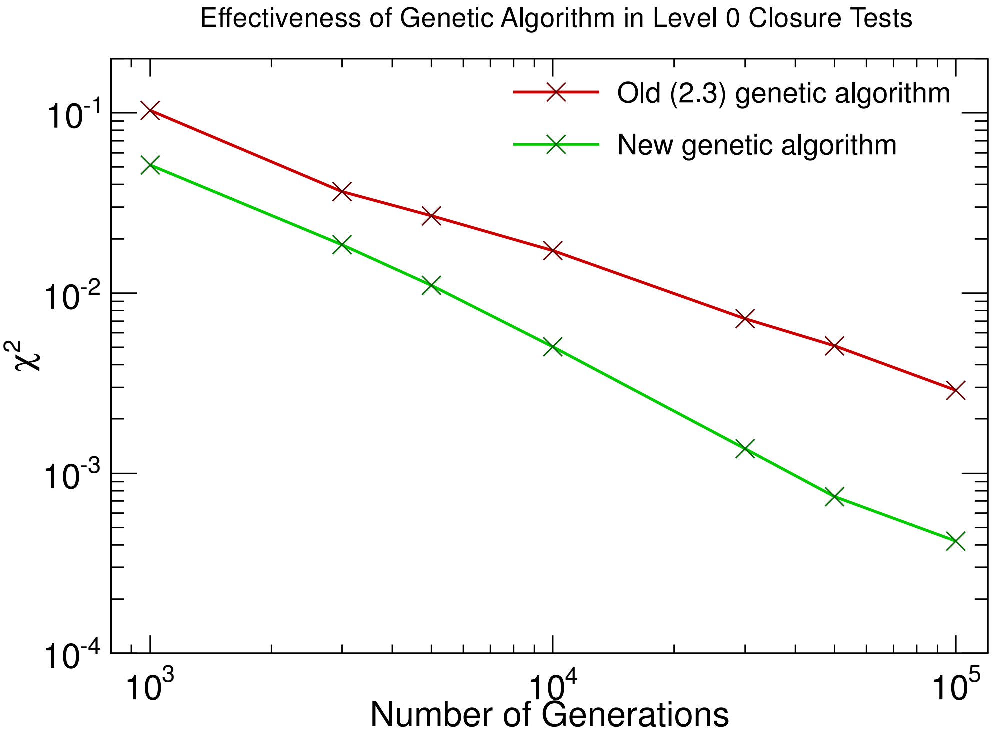


Figure 6.1: The normalised central χ^2 of level 0 closure tests, Eq. 6.1, for the old and new genetic algorithms as a function of the length of the genetic algorithms minimisation (repeated from Fig. 5.1).

minimisation, as in the previous chapter, and investigate the power of the final settings, as we will do here.

The two main ingredients of our fitting methodology that can be tested in level 0 closure tests are the adequacy of the neural network architecture and the efficiency of the genetic algorithm minimization. Since at level 0 no stochastic fluctuations are added in the generation of the pseudo-data, the ideal χ^2 is zero, and as the length of the training is increased we expect the fitted PDFs to get closer and closer to the input ones. In order to verify that this is the case, we have performed a number of fixed length fits to the full dataset, and studied the dependence on the training length of the χ^2 calculated over the ensamble, i.e. $\chi^2[\langle\mathcal{T}[f_{\rm fit}]\rangle\,,\mathcal{D}_0]$ where the theory values are averaged over the replicas. These fits were performed with identical settings apart from the training length, which was varied between 1000 and 100000 genetic algorithm generations. As the fits used level 0 pseudodata, no cross-validation was used.

The dependence of the χ^2 on the training length for these level 0 closure tests is plotted in Fig. 5.1 from the previous chapter and repeated here in Fig. 6.1. In Section 5.3.3 we used this plot to compare the updated genetic algorithm used in NNPDF3.0 with the genetic algorithm used in the NNPDF2.3 fit, demonstrating that the newer methodology was a significant improvement over the previous approach. Here, I want to highlight the fact that from the figure we can see that the χ^2 of the fit decreases as the fit length is increased, with a behaviour that is approximately described by a power law with a power of about -1.1. We can herefore see that given enough time, the genetic algorithm can obtain results arbitrarily close to ideal, though as we approach the minimum an increasing amount of time is required for improvement[1].

Given that the χ^2 is tending towards zero, we expect almost perfect agreement between the fitted and input PDFs. We can look at this by looking at the resulting PDFs from the fits themselves, as shown in Figs. 6.2 and 6.3. In these plots we show compare the PDFs obtained from the level 0 fit with the longest training length (100k generations) to the MSTW input PDFs. The central values of our fitted PDFs, shown on the figures by the dotted green lines, are computed as the average over replicas in the usual way:

$$\langle f_{\rm fit} \rangle = \frac{1}{N_{\rm rep}} \sum_{k=1}^{N_{\rm rep}} f^k_{\rm fit} \, , \tag{6.4}$$

the angled brackets denoting the average over replicas, and the band on the plots is here the one-sigma uncertainty[2] given by:

$$\sigma_{\rm fit} = \sqrt{\langle (f^2_{\rm fit} - \langle f_{\rm fit} \rangle)^2 \rangle} = \left(\frac{1}{N_{\rm rep}} \sum_{k=1}^{N_{\rm rep}} \left(f^k_{\rm fit} - \langle f_{\rm fit} \rangle \right)^2 \right)^{\frac{1}{2}} \, . \tag{6.5}$$

It is clear from the plots that the input PDFs are reproduced to a very high standard. It is interesting to observe that PDFs for which there is a large amount of experimental information, such as for example the up quark in the valence region, are perfectly reproduced with essentially no uncertainty. PDFs for which information is more sparse or indirect, such as for example the gluon, have an uncertainty even when the χ^2 at the data points is essentially zero. This is likely due to both a larger amount of freedom in interpolating between data points for these regions and PDFs, and also that they have a smaller relative weight in the figure of merit, so are less well trained. On this basis, if we look at the combination of the PDFs which corresponds directly to a experimental measurement, it should have smaller uncertainties than the individual PDFs. For instance, the bottom row of plots in Fig. 6.3 show the PDF dependence of

[1] This is a well known behaviour of genetic algorithms. In this particular case, once we are close enough to the absolute minimum it might be more useful to switch to other strategies like steepest descent. However, in actual fits the issue of over-learning mean that the actual minimum of the training χ^2 is not necessarily the best fit, so this is unnecessary in practice

[2] This one of the two ways to generate an uncertainty from a Monte Carlo replica set, with the other being to calculate the central 68% confidence interval, i.e. the interval which contains 68% of the replicas, with 16% above and below.

the leading-order expression of the structure function F_2^p, namely $\frac{4}{9}\left(u+\bar{u}+c+\bar{c}\right)+\frac{1}{9}\left(d+\bar{d}+s+\bar{s}\right)$, which is directly probed by the HERA data, the uncertainty on it at small x in the HERA data region $10^{-4}<x<10^{-3}$ is significantly smaller.

This effect on the PDF uncertainties is however not necessarily the case in the extrapolation regions, where we expect large PDF uncertainties, and moreover uncertainties which are essentially independent of the training length. This is due to the fact that, as by definition there is little data in these regions, the functional forms taken by the neural networks can vary substantially without changing the fitted figure of merit, $\chi^2[\mathcal{T}[f_{\rm fit}],\mathcal{D}_0]$. These two phenomena, very small PDF uncertainties in the data region, and large PDF uncertainties in the extrapolation regions, in particular at small and large x are clearly visible in the plots in Fig. 6.2, even at the end of the 100k-generation training. These results provide a way of quantifying the *extrapolation* uncertainty on the PDFs, caused by the lack of direct constraints in these regions. This is a source of PDF uncertainty that can only be reduced if new data is provided, and that accounts for the majority of the PDF uncertainties in the extrapolation regions. The extrapolation uncertainty will be studied in more detail in Section 6.3.

Additional interesting information can be extracted from the set of level 0 fits by looking at the PDF uncertainties of the resulting fits, computed as the standard deviation over the sample of $N_{\rm rep}=100$ fitted replicas either at the level of parton distributions or at the level of physical observables. Given that the level 0 input pseudo-data do not fluctuate, and that only difference between replicas at this closure test level is the random seed for the minimisation, we expect that the cross-sections computed from the fitted PDFs should converge to the input values for each replica as the training length is increased; i.e. that the uncertainty on the predicted value for all the observables included in the fit must go to zero.

To verify this expectation, we can use the φ_{χ^2} estimator defined in Appendix B. In Fig. 6.4 we show φ_{χ^2} for the level 0 fits as a function of the length of the genetic algorithms minimisation (the equivalent for φ_{χ^2} of Fig. 6.1). We can see that as we increase the training length the spread of the theoretical predictions at the data points for different replicas decreases monotonically. Again, here we can also observe the improvement from the more efficient minimisation strategy in NNPDF3.0.

6.3 PDF uncertainties: data, functional and extrapolation components

In the previous section we looked at some results from level 0 closure tests, in which the fit quality can become arbitrarily good and the PDF uncertainties arise largely due

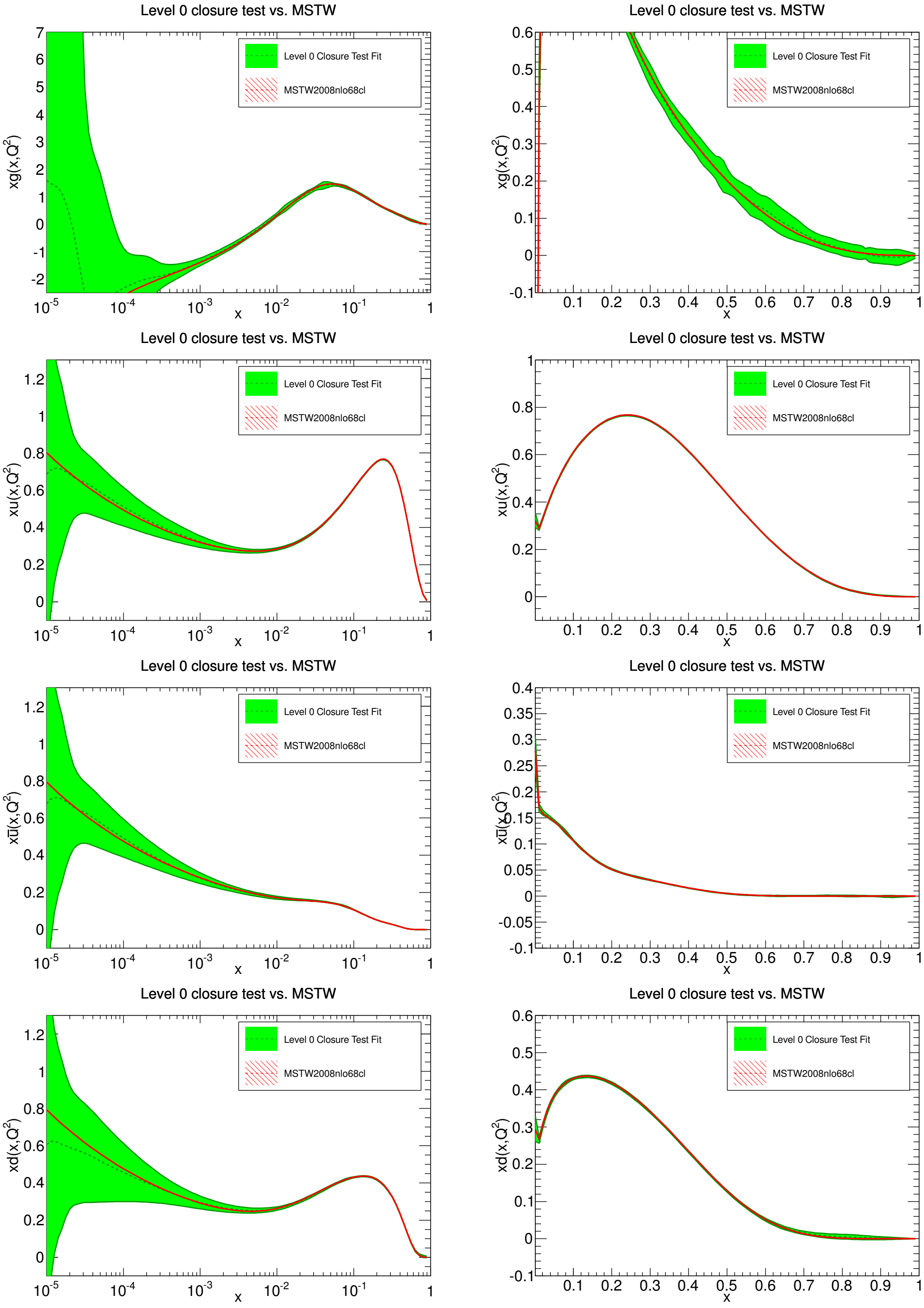


Figure 6.2: Comparison between the results of the level 0 closure fit with 100k GA generations and the corresponding input PDF set, the central value of MSTW2008 NLO PDF set. The green band shows the one-sigma interval computed over the sample of $N_{\rm rep} = 100$ replicas, with the green dotted line showing the mean value. The plots show the gluon, u, $\bar{u}$ and d PDFs on both linear (right hand side) and logarithmic (left) scales in x, at the scale $Q^2 = 1$ GeV^2 where the PDFs are parametrized.

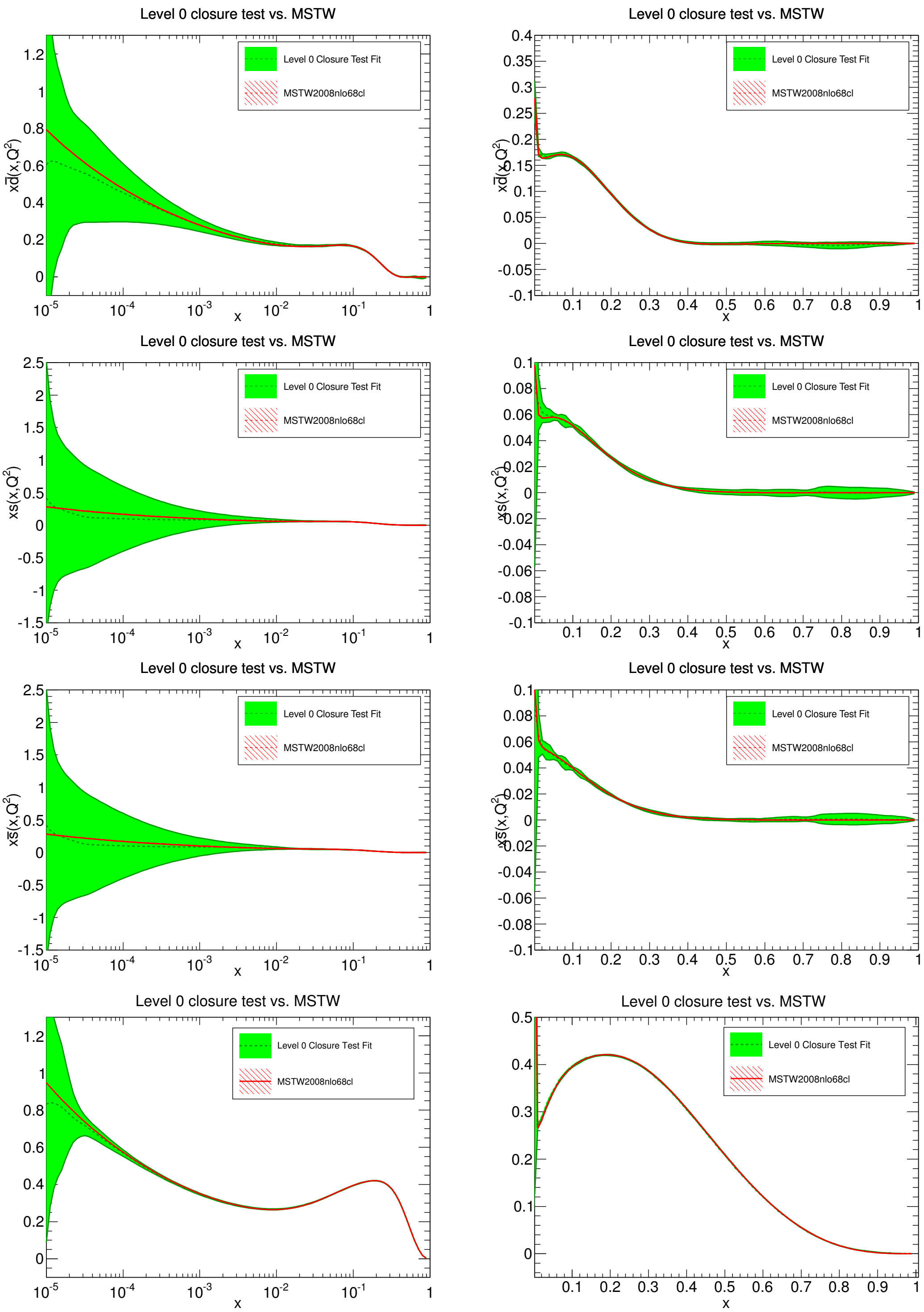


Figure 6.3: Same as Fig. 6.2 for the $\bar{d}$, s and $\bar{s}$ PDFs, and for the combination of PDFs which corresponds to the leading-order expression of F_2^p.

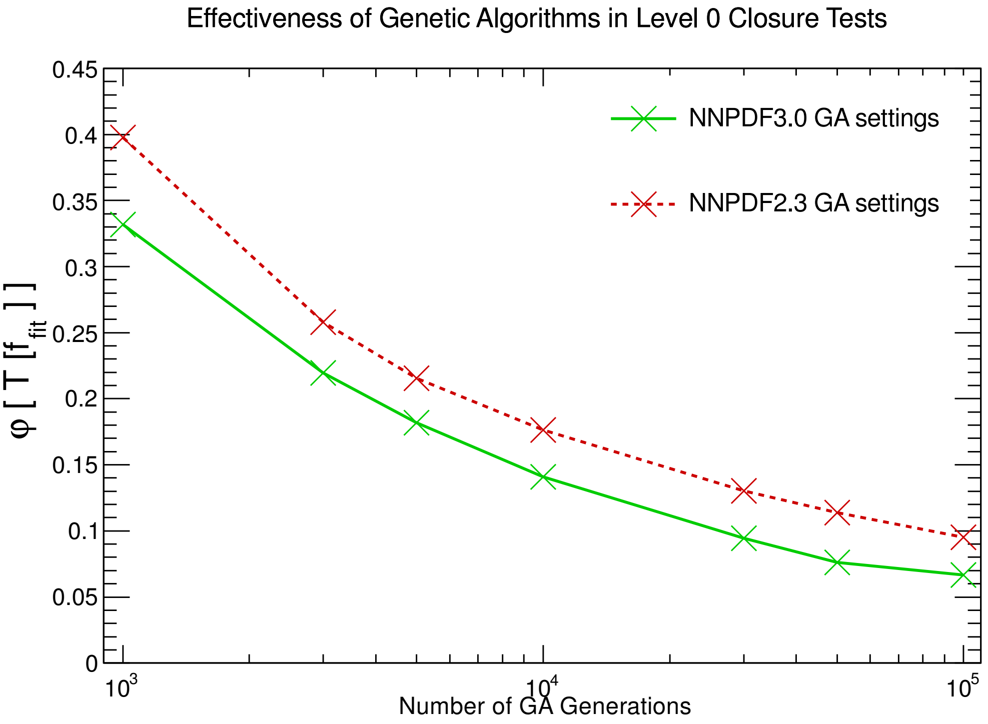


Figure 6.4: The estimator φ_{χ^2}, given in App. B, as a function of the length of the genetic algorithms minimisation. The results for both the NNPDF3.0 and NNPDF2.3 GAs are shown, with the actual closure test results marked by crosses.

to the fact that the experimental data used in the fit has finite kinematical coverage. Now I will turn to level 1 and level 2 closure tests in order to shed some light, in the cleanly controlled environment of closure testing, on the various contributions to PDF uncertainties, specifically those due to the uncertainty of the experimental data, to the choice of functional form, and to the (previously mentioned) interpolation and extrapolation uncertainties due to the finite coverage of the data.

A sophisticated understanding of the various sources that form the total PDF uncertainties can be obtained in the context of closure tests by comparing level 0, level 1, and level 2 fits. This is because in each of these different levels the PDF uncertainty band has different components. In level 0 fits, the only significant component is the interpolation and extrapolation uncertainty (which I will collectively refer to as extrapolation uncertainty for short); in level 1 fits, fluctuations in the data mean it is now possible to overfit, so here we also have uncertainty due to the selection of a specific function for the PDF; and in level 2 finally one also adds the uncertainties propagated from the experimental data. Therefore, by comparing the results obtained in level 0, level 1 and level 2 closure fits we can analyse how much of the total PDF uncertainties is from the data, functional and extrapolation uncertainties.

Let us begin with the extrapolation uncertainty. As discussed in the previous section, in a level 0 closure test, the genetic algorithm error function should go to zero for all replicas as the training length is increased. This implies that PDF uncertainties should also decrease monotonically as a function of the training length wherever data are available. However, in between data (interpolation) and outside the data region (extrapolation) PDFs can fluctuate, as these regions are not directly constrained by the error function. We refer to this residual uncertainty, which would remain even with infinite training length, as the *extrapolation* uncertainty. Note that, given the highly non-trivial dependence of PDFs on the measured cross-sections—including that these cross-sections typically depend on multiple PDFs—and the wide range of observables included in the fit, it is very difficult to determine precisely how this extrapolation region is defined in our fits. While a non-negligible extrapolation component is expected for all PDFs at small enough and large enough values of x, it is also possible, though perhaps unlikely, that significant uncertainties due to interpolation could also be present at intermediate x. Also, it is worth noting that for finite length fits there will be a spurious component of the uncertainty at level 0 due to the non-convergence of the fit.

In a level 1 fit, the central values of the data have been fluctuated around the theoretical prediction, and therefore $f_{\rm fit} = f_{\rm in}$ no longer provides an absolute minimum for the χ^2, and instead gives a value of $\chi^2[\mathcal{T}[f_{\rm fit}], \mathcal{D}_1] \approx 1$. However, instead of a single best fit, because of this there will be a number of possible functions with roughly the same goodness of fit, corresponding to equally likely possibilities for the underlying law which cannot be distinguished on the basis of just this data. Therefore, in Level 1 closure fits, on top of the extrapolation component, the total PDF uncertainty will include a new component which we refer to as *functional* uncertainty.

This functional uncertainty is a consequence of the fact that the optimal χ^2 in the presence of data fluctuations is not the absolute minimum of the χ^2. Indeed, provided the PDF parameterisation is flexible enough, it will be possible to find functional forms with a χ^2 much smaller than one, but which will not be optimal as they will provide poor predictions of future data. In a closure test, the optimal result corresponds to the true underlying functional form, and thus the optimal χ^2 is the one of the level 1 pseudo-data, whose value is approximately one, depending on the fluctuations in the data. For an infinite-dimensional space of functions, this χ^2 value can be obtained in an infinite number of different ways, and the spread of these possibilities provides the functional uncertainty. This source of uncertainty is in many ways similar to the extrapolation uncertainty, but whereas the latter is due to the finite range of the data points, the functional uncertainty is caused by the loss of information about the exact values of the data points due to the fluctuations.

In level 2 fits, the starting point is again the level 1 pseudo-data generated by adding a Gaussian fluctuation over the predictions obtained from the input PDFs based on the quoted experimental uncertainties. However, now there is a second step, where an additional set of fluctuations are applied separately to each replica dataset. This yields an ensemble of fitted PDFs $\{f^k_{\text{fit}}\}$ with statistical properties which faithful propagate the uncertainties of the underlying dataset. The increase in the uncertainty from level 1 to level 2 fits is the *data* uncertainty. It is worth noting that this uncertainty is separate from the functional uncertainty in the PDFs which are due to the actual fluctuations in the level 1 dataset. Indeed, it is possible to perform a closure test fit with the standard data fluctuations but much larger or smaller uncertainties, which produces PDFs which then have the same size functional uncertainty but larger or smaller data uncertainty.

Figs. 6.5 and 6.6 shows the size of the uncertainties in comparable level 0, level 1 and level 2 closure test fits as a ratio of the central value in each case. Each fit was performed using the MSTW2008 NLO set as input PDFs f_{in}, and a maximum (or total, in the level 0 fit) training length of 30k generations. The level 1 and level 2 fits were performed with look-back cross-validation, and the level 0 fit was performed without a training-validation split. Results are provided for the PDFs in the flavour basis at the input parameterisation scale of $Q^2 = 1$ GeV^2.

From the descriptions above, it is possible to understand the features that we can observe in Fig. 6.5 and 6.6. Firstly, we see that level 0 uncertainties (the blue bands) are generally smaller than the level 1, and in turn these are generally smaller than those at level 2. This confirms the expectation that at each level we are adding a new component of the total PDF uncertainty, extrapolation, functional and data components, respectively.

We also observe that in the small-x and large-x regions it is the extrapolation uncertainty that dominates, given that the level 2 PDF uncertainties are already reasonably reproduced by those of level 0 closure fits. However, we can see that the level 0 uncertainty is also not negligible in some medium x regions where there is more constraint from experimental data. This could be due to valid sources of uncertainty such as interpolation or degeneracies, or simply due to the failure of some replicas to converge.

By comparing the level 1 results to the level 0, we see that the functional uncertainty (shown by the difference between the red and blue bands) is generally sizeable, and is the dominant component for several PDFs at large x on the boundary between the data and extrapolation regions, for example $\bar{d}$ and $\bar{u}$ at $x = 0.3$. The data uncertainty, shown by the difference between the level 2 uncertainties in green and the level 1 uncertainties in red, is also significant in the data region and less so outside this, as we would expect.

Interestingly, in regions where we have a rather reasonable coverage from available data, the three components on the uncertainty are roughly of similar size. Take for

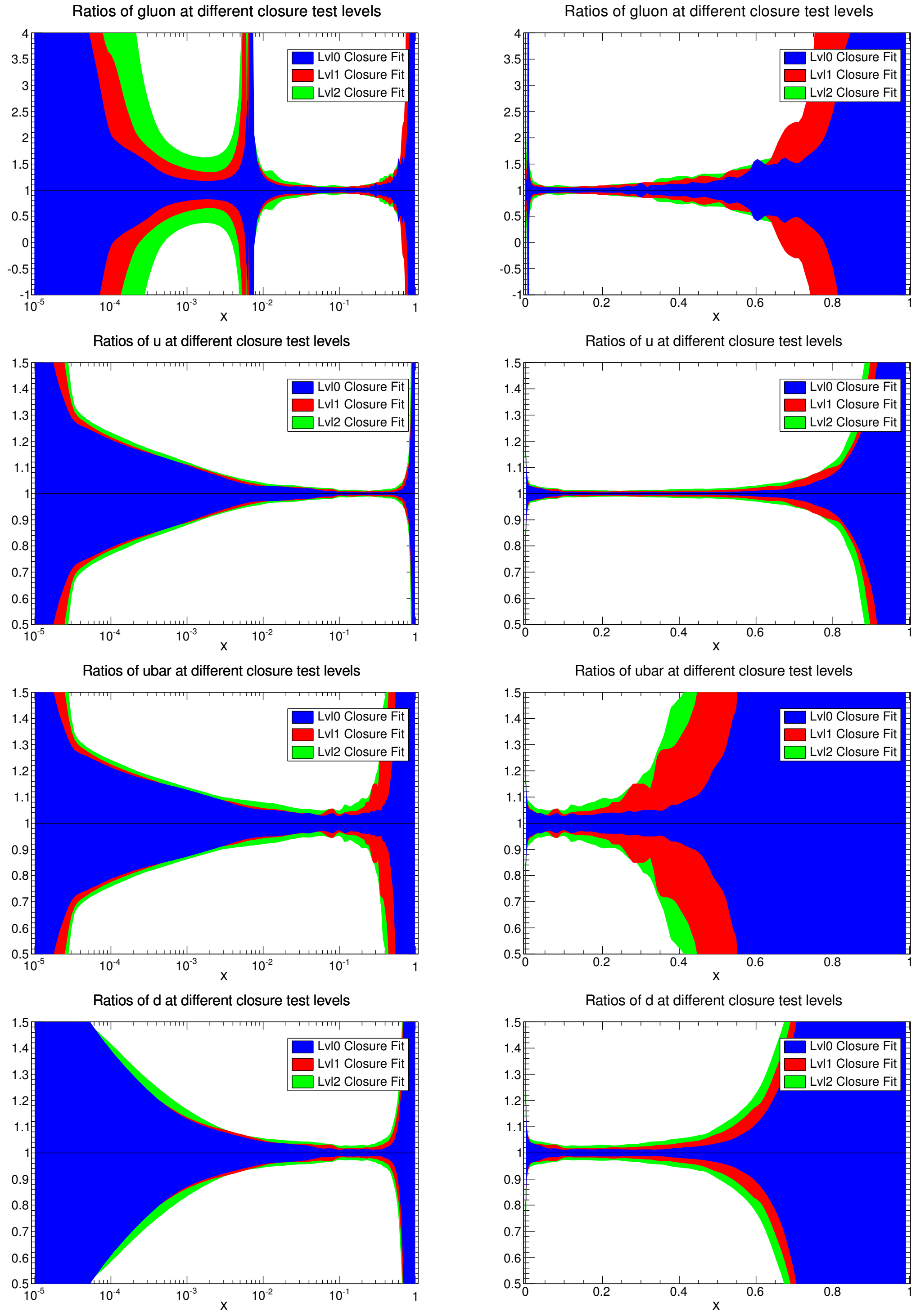


Figure 6.5: Comparison of relative PDF uncertainties obtained from level 0 (red), level 1 (blue) and level 2 (green) closure test fits with MSTW2008 NLO as input set. The PDFs are shown as a ratio to their own central value. Results for the gluon, u, $\bar{u}$ and d PDFs are shown on this page, and for $\bar{d}$, s and $\bar{s}$ PDFs on the next page. All ratios are plotted at the input parameterisation scale of $Q^2 = 1$ GeV^2, both in logarithmic (left) and in linear (right) scales.

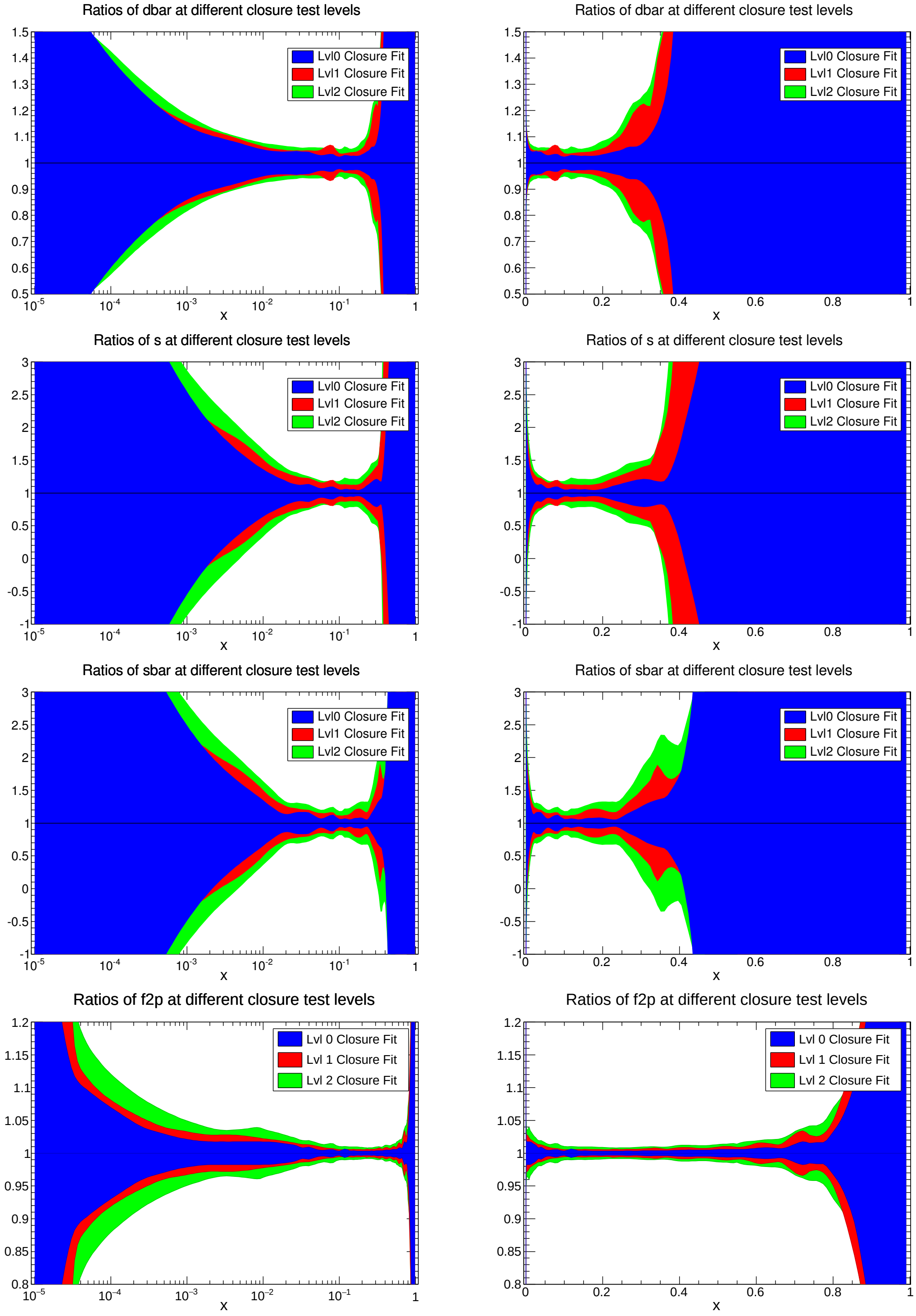


Figure 6.6: Continued from Fig. 6.5, relative uncertainties in level 0, 1 and 2 closure test fits for the $\bar{d}$, s and $\bar{s}$ PDFs, and for the combination of PDFs which corresponds to the leading-order expression of F_2^p.

example the gluon around $x \sim 10^{-3}$, which is well constrained by the high-precision HERA measurements, or the PDF component of leading order expression of F_2^p. We see that the functional and data uncertainties are of similar size to the level 0 uncertainty. This also applies for other PDF flavours, such as for example strangeness for $x <\sim 0.01$ (with abundant constraints from neutrino DIS and LHC data) or the up and down quarks at medium and large-x (with many DIS and LHC datasets providing information).

This provides an important general conclusion that the data uncertainties are not dominant, and including the extrapolation and functional components is important to correctly estimate the overall PDF uncertainty. This conclusion is consistent with that of previous, less sophisticated, NNPDF work on this topic, such as that in [180]. It also is natural to conjecture that the tolerance method [181] which is used in Hessian fits, provides an effective way of supplementing the data uncertainty obtained through the Hessian method with these extra necessary components of the uncertainty.

We can also get a more quantitative assessment of the contributions to PDF uncertainties by means of the estimator φ_{χ^2} introduced in the previous section and described in Appendix B. This provides a measure of the average size of the PDF uncertainties on the data points, in units of the experimental uncertainties. Note that as φ_{χ^2} is calculated exclusively at the data points it cannot show the extrapolation uncertainties, as these are only present away from the data points. For the three levels of closure test fit, we obtain

$$\varphi_{\chi^2}^{\mathrm{lvl0}} = 0.095\,, \qquad \varphi_{\chi^2}^{\mathrm{lvl1}} = 0.173\,, \qquad \varphi_{\chi^2}^{\mathrm{lvl2}} = 0.254\,. \tag{6.6}$$

If we assume that the functional and data uncertainties are added in quadrature, we can calculate from these values the fraction of the total uncertainty from these sources. Doing this we obtain

$$\varphi_{\chi^2}^{\mathrm{func}} = 0.145\,, \qquad \varphi_{\chi^2}^{\mathrm{data}} = 0.186\,. \tag{6.7}$$

This suggests that the functional and data uncertainties are roughly equally sized, confirming what we see in Fig. 6.5 and 6.6, though the data uncertainty makes up a larger proportion of the total uncertainty: 53%, compared to 33%. As the extrapolation uncertainty is not captured by this measure, $\varphi_{\chi^2}^{\mathrm{lvl0}}$ should be zero, however our level 0 fit still has a non-zero value, suggesting that while smaller than the functional and data uncertainties is still a substantial fraction of the total (14%). This is likely due to replicas failing to converge to the global minimum during training, and we can see from Fig. 6.4 that the value of $\varphi_{\chi^2}^{\mathrm{lvl0}}$ falls as the training length is increased. This indicates that the improved fitting methodology is still quite far from ideal, and so is something which could be improved further in future work.

6.4 Validation of the closure test fits

So far, I have used closure tests to study the effectiveness of new methodological features and to investigate contributions to the PDF uncertainties. However, I have not looked at the main use of closure tests: to statically validate the results of our fits. In this section I will demonstrate that our methodology can successfully reproduce the input PDFs in closure test fits, and has a number of other important statistical features. First, I show how similar the PDFs and χ^2 values from the input and fitted sets are, both for the total dataset and for individual experiments. Then I will discuss a quantitative validation of the PDF uncertainties obtained in the closure tests, using the estimators defined in Appendix B. Finally I will look at closure test fits using different input PDFs, including to NNPDF3.0 (giving a 'true' closure test) and to a set of PDFs with an unrealistic degree of complexity and structure.

6.4.1 Central values

To evaluate the effectiveness of our methodology in reproducing the underlying law, we performed a level 2 closure test fit with the final NNPDF3.0 setting, and a pseudo-dataset based on the final NNPDF3.0 dataset and generated using the MSTW2008 NLO PDFs. The results shown here are based on a single 100 replica PDF set, but we also performed multiple equivalent with different seeds—one of which is looked at in Section 6.4.3—to verify that the chosen set is representative.

One indicator of the quality of a closure test fit is provided by the values of the central χ^2 to the pseudo-data, calculated using the average value of the observables over the replicas. If the test is successful, this should reproduce the central χ^2 obtained using the generating PDFs. We can look at this directly by considering the Δ_{χ^2} estimator defined in Appendix B. For our level 2 closure test we obtain

$$\Delta_{\chi^2} = -0.011\,, \qquad (6.8)$$

which shows that the fitted PDF set reproduces the χ^2 of the input PDF set at the 1% level.

This level of agreement is achieved not only for the total χ^2, but also for the central χ^2 for the individual experimental datasets included in the fit. This is important to demonstrate, since it provides a more fine-grained test that the fitted PDFs are reproducing the underlying law across all kinematic regions and PDFs. Fig. 6.7 shows

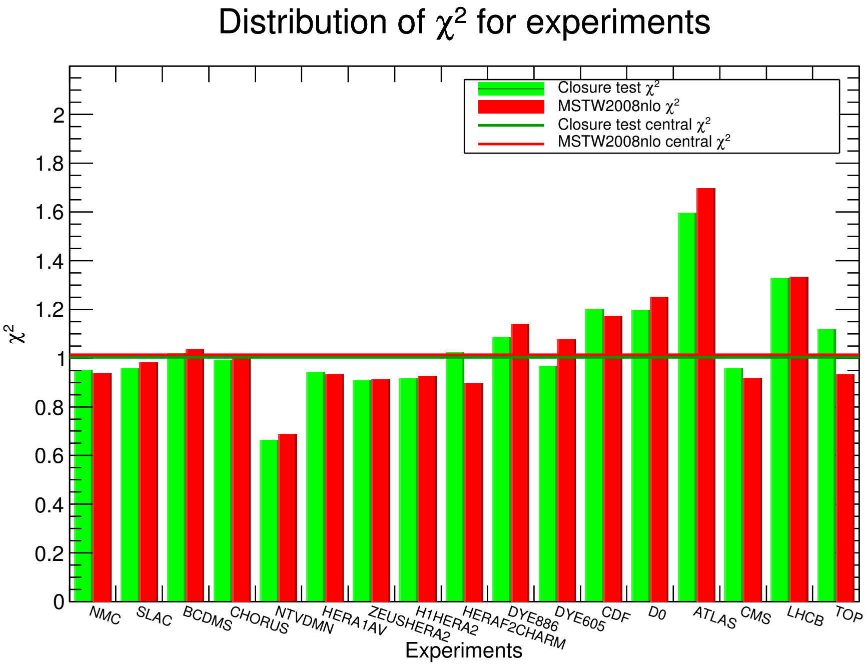


Figure 6.7: Comparison of the central χ^2 to the closure test data obtained with the input (red) and with the fitted (green) PDFs, for a level 2 closure test fit based on MSTW2008 pseudo-data, for individual datasets included in the fit. The horizontal bars show the total central χ^2 for the two PDF sets. The datasets shown are the same used in the baseline NNPDF3.0 global fit, see Tables 4.1 in Section 4.

the central χ^2 for the closure test fit to the pseudo-data generated for each individual experiment, compared to the corresponding values for the input PDFs. The horizontal lines show the total central χ^2s for the two PDF sets, which are effectively averages of the individual experiment values weighted by the number of points in each dataset. Note that the χ^2s obtained for each dataset can be quite different from one, as they depend on the specific fluctuations added to the pseudo-data. We can see from this figure that the NNPDF methodology successfully reproduces the χ^2 of the input PDFs not only for the total dataset but also experiment by experiment, and does so even when the target χ^2 is far from one. Fig. 6.7 therefore provides strong evidence that, at least at the level of central values, the level 2 closure test is successful.

We can also look at the agreement in the PDFs themselves by plotting the distance between the fitted PDFs and the input MSTW2008 PDFs in units of the standard deviation of the fit PDFs, as defined in Appendix A. These are shown in Fig. 6.8. The plots show that the fitted and input PDFs are in good agreement, generally at the level of one sigma or better, and with deviations to about two sigma in some places for some PDFs as one would expect if the underlying distribution was roughly Gaussian. In the extrapolation regions, at small and large x, the distances between input and fitted

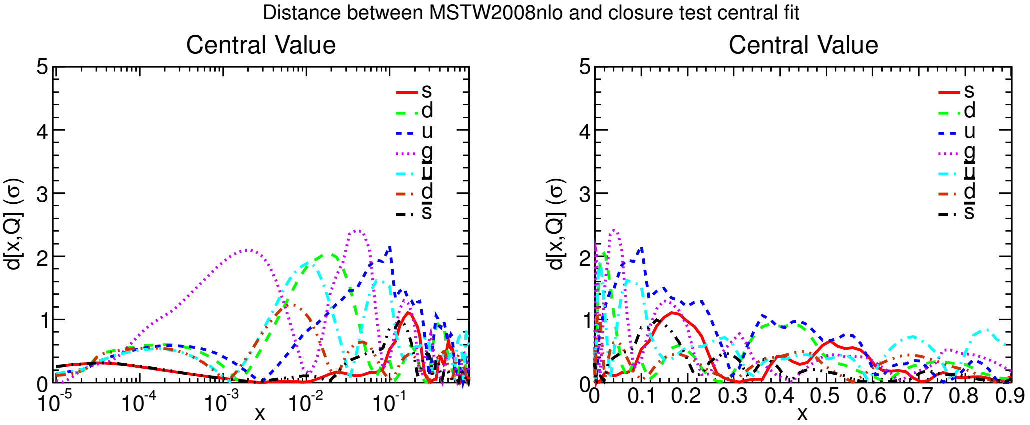


Figure 6.8: Distances between the central values of the fitted PDFs from a level 2 closure test and the MSTW2008 PDFs, which were used as input to generate the pseudo-data. The values are normalised to the standard deviations of the fitted PDFs, as described in Appendix A. The values are computed at the input parametrisation scale of $Q^2 = 1$ GeV^2.

PDFs become smaller because of the large extrapolation uncertainties in these regions.

From the distances in Fig. 6.8 we can see that at the qualitative level the closure test is successful, since the fitted PDFs fluctuate around the truth by an amount which is compatible with statistical expectations. More insight on this comparison is provided by plotting the ratio between the fitted and input PDFs, $f_{\rm fit}/f_{\rm in}$, for all PDF flavours. This comparison is shown in Fig. 6.9 and 6.10 on both linear and logarithmic scales in x. It is clear from these plots that the NNPDF methodology reproduces successfully the input PDFs, with deviations from the input functions by two standard deviations at most. This comparison provides initial evidence that PDF uncertainties are properly estimated in Level 2 closure tests, in that the deviations of central value of the fitted PDFs from the truth are consistent with the size of the PDF errors.

While the deviations of two sigma we can see in Fig. 6.9 and 6.10—for example in the gluon at x=10^{-3} and $\bar{d}$ at x=10^{-2}—can be explained by the statistical fluctuations in the pseudo-data, it is still slightly off-putting, and it is possible that they are evidence of a bias. Fortunately, with closure tests it is easy to test whether this is the case: we simply need to perform a second closure test with the same settings but using a different random seed to generate the pseudo-data. If the two-sigma differences are just fluctuations, they should not appear in the results of the second fit. Fig. 6.11 superimposes the ratios for the gluon and $\bar{d}$ PDFs from a fit with a different set of psuedo-data, over the previous results from Fig. 6.9 and 6.10. We can see that the disagreements in the first fit are not present in the second, indicating that the differences between the fitted PDFs and the input PDFs were indeed just due to the particular set of pseudo-data used, not due to the methodology.

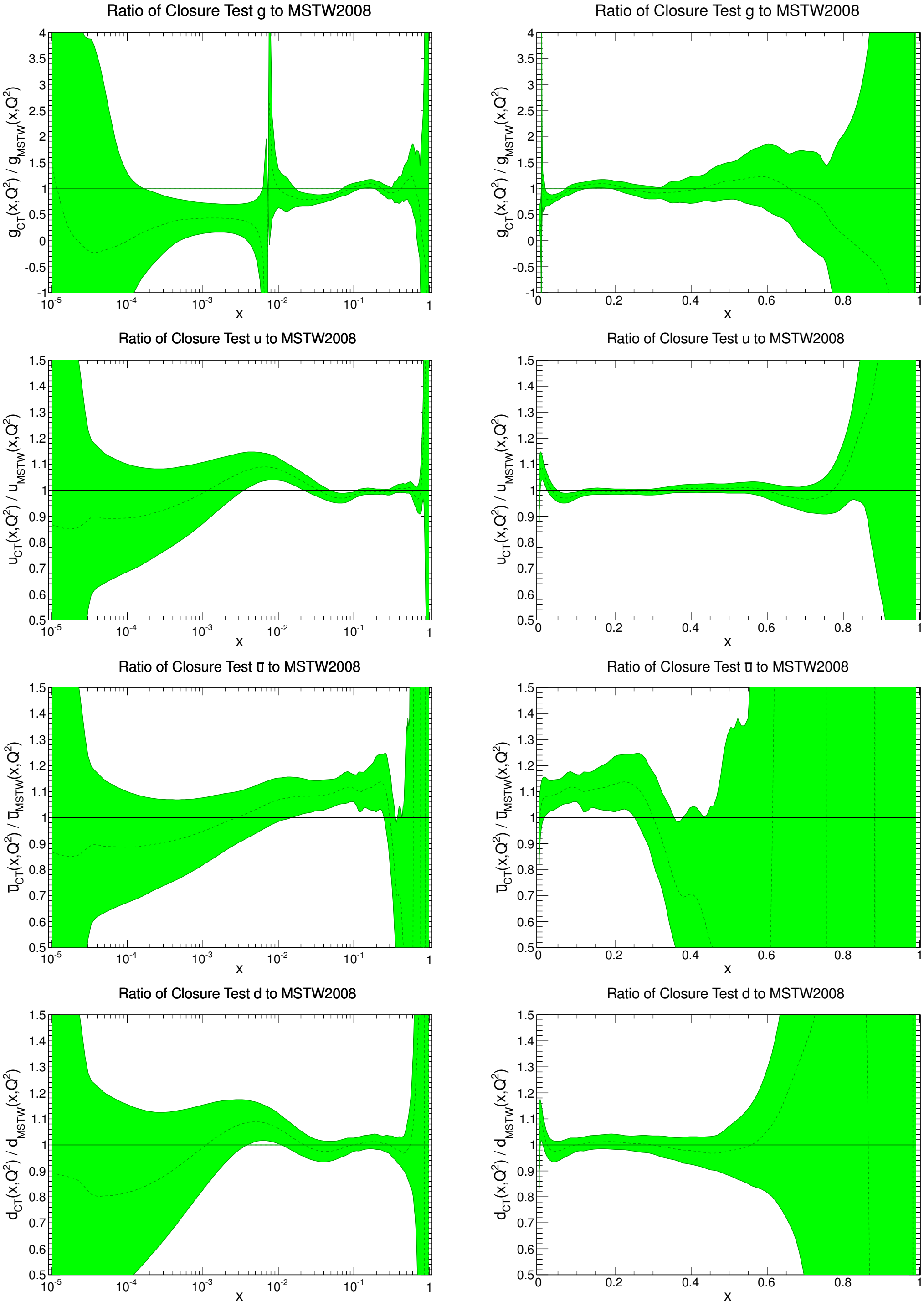


Figure 6.9: Ratio of the PDFs obtained from a level 2 closure test which uses MSTW2008 PDFs as input, with respect to the input MSTW2008 PDFs themselves. The green band shows the one-sigma interval of the fitted PDFs, while the green dotted line is the corresponding mean. The plots for the gluon, u, $\bar{u}$ and d PDFs, on both linear (right hand side) and logarithmic (left) scales in x, are shown here, and the equivalent plots for the $\bar{d}$, s and $\bar{s}$ PDFs are on the next page. The comparison is performed at the fitting scale of $Q^2 = 1$ GeV^2.

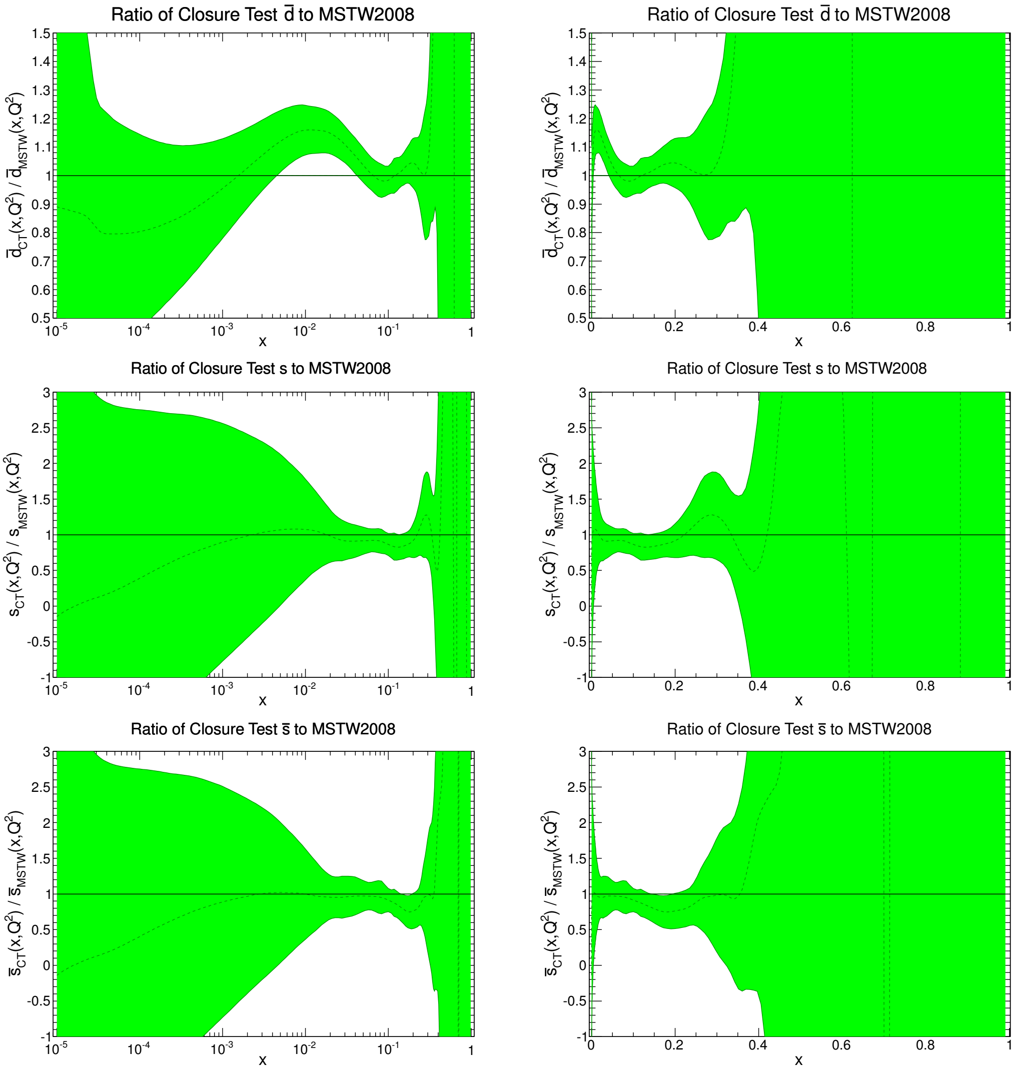


Figure 6.10: Continued from Fig. 6.9, ratio of fitted and input PDFs for $\bar{d}$, s and $\bar{s}$ PDFs.

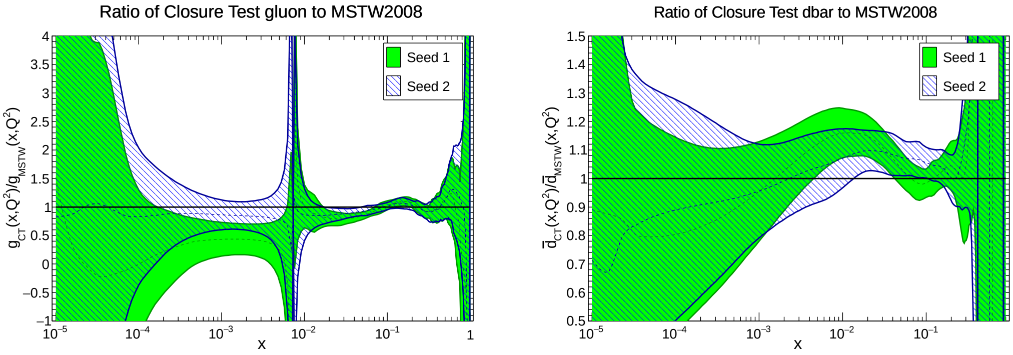


Figure 6.11: Same as Fig. 6.9 but with the ratios for a different set of psuedo-data (in blue) superimposed over the original results.

6.4.2 PDF uncertainties

While it is straightforward, as the previous section showed, to use closure tests to validate the central values obtained from our PDF fits, it is less clear how they can be used to demonstrate that the PDF uncertainties are also valid. With the central values we can compare directly to the ideal results, the input PDFs, while the uncertainties are not related to the input set so we do not have a 'correct' answer to compare with. However, there are a number of techniques we can use to obtain information about the validity of our uncertainties from closure tests. We have already seen some evidence from Fig. 6.9 and 6.10 that the uncertainties obtained are consistent with the size of deviations from the theory values, which suggests that the uncertainties are reasonably sized. In this section I will discuss a way of more quantitively demonstrating that the PDF uncertainties we obtain are valid.

PDF uncertainties, by definition, should give the probability that the true value for the theory is some particular value, given the data used in the fit. In particular, the theory value should have a 68% chance of lying within one sigma of the PDF central value (assuming Gaussianity). In principle this is something which could be tested using closure test fits, however generating a large number of PDF theory values according to a particular distribution is very complicated. Instead, we can invert this relationship and test the number of times the theory value is within one sigma (say) of the central value of a large number of closure test fits each performed using a statistically different set of pseudo-data. This is the idea behind the ξ_σ estimator described in detail in Appendix B. Essentially we generate a large number of closure test fit central values with different pseudo-data and perform the check described above averaging over multiple different PDFs and points in x, using uncertainties from a full closure test fit. For the level 2 closure test described above, we obtain

$$\xi_\sigma^{(12)} = 0.699\,, \qquad \xi_{2\sigma}^{(12)} = 0.948\,, \tag{6.9}$$

to be compared with the theoretical expectations of 0.683 and 0.955. This excellent agreement confirms that the PDF replicas obtained by our fitting methodology provide a faithful representation of the probability distribution for the PDFs given the data used in the fit.

To verify that this agreement is not accidental, or a fluke of the definition of the estimator ξ_σ, but rather a robust feature of our analysis, we can compute again the ξ_σ estimators but instead using the uncertainties from a level 1 closure test fit. While the central values are the same, we know that in level 1 closure tests PDF uncertainties are underestimated, as they lack the component of uncertainty coming from the data as described in Section 6.3, and therefore there will be inconsistency between the spread of

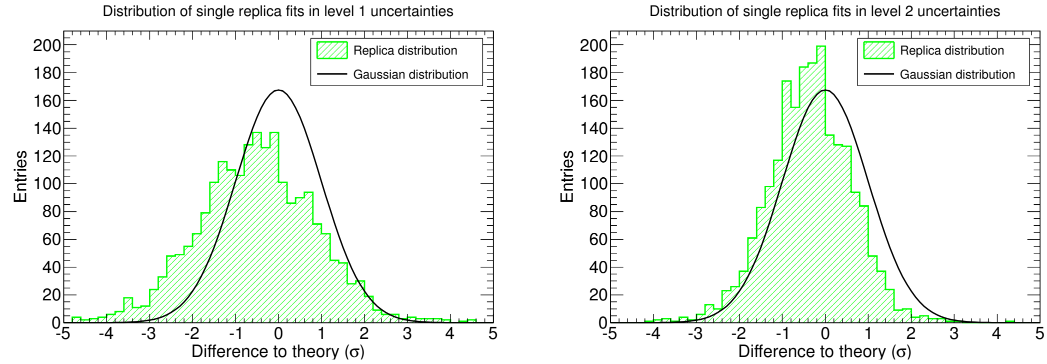


Figure 6.12: Histograms for the difference between the input PDF and multiple fitted PDF central values obtained from different sets of closure test pseudo data, in units of the standard deviation of separate level 1 (left) and level 2 (right) closure test fits. An appropriately scaled Gaussian distribution is shown for comparison.

central values and the uncertainties. Based on this, in level 1 closure tests we expect the ξ_σ estimators to be somewhat smaller than the theoretical expectations above. Indeed, computing ξ_σ and $\xi_{2\sigma}$ at level 1 this is precisely what we find:

$$\xi_{\sigma}^{(l1)} = 0.512\,, \qquad \xi_{2\sigma}^{(l1)} = 0.836\,, \tag{6.10}$$

which shows that indeed the level 1 closure tests fail, in the sense that level 1 fits underestimate the PDF uncertainties, and strengthens the results we obtained from the level 2 fits.

We can look at this statistic in more detail by looking at the distribution of the multiple PDF central values that we generated to calculate the ξ_σ estimators above. This tests not only the one- and two-sigma confidence intervals, but the shape of the whole distribution of deviations between the prediction and the truth. Fig. 6.12 shows the histograms of the differences between $\langle f_{\rm fit}\rangle$ obtained using different closure test datasets (that is, pseudo-data generated with different random seeds) and the central value $f_{\rm in}$ of the MSTW input PDFs, in units of the standard deviation of the fitted PDFs. The distribution for the level 1 closure test is shown on the left and for the level 2 fits on the right. The histogram is generated using the values at $x = 0.05$, 0.1 and 0.2 for each PDF, as a representative sampling. The resulting distribution is close to a Gaussian distribution with a standard deviation of one when using the level 2 uncertainties, but is considerably wider using the level 1 uncertainties as we would expect from the values of ξ_σ above. In both cases, however, the distribution appears to be offset from zero; it's not clear what the cause of this is and it may be worth investigating in future closure test work.

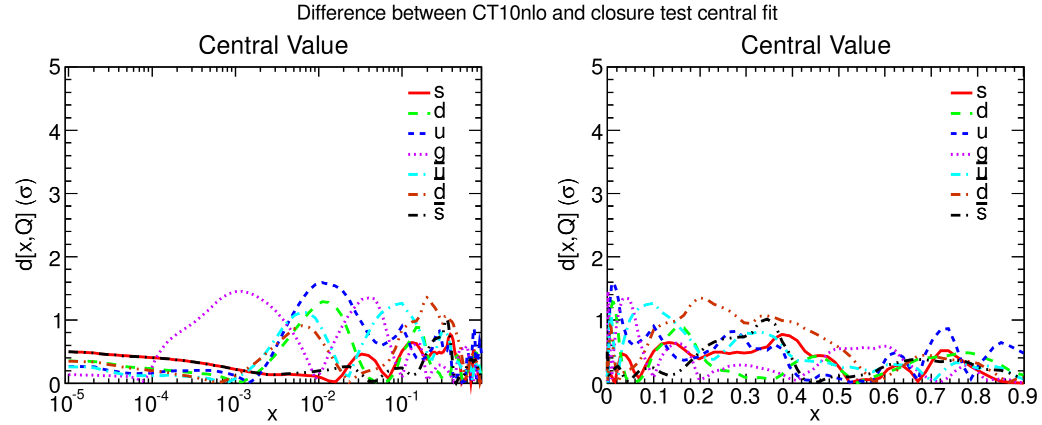


Figure 6.13: Distances (same as Fig. 6.8) between the central values of the fitted PDFs and the input CT10 PDFs for a level 2 closure test, in units of the standard deviation of the fitted PDFs.

6.4.3 Tests with different input PDFs

So far in this chapter we have looked only at the results of closure test fits using pseudo-data generated using MSTW2008 input PDFs. However, it is important to verify that there is nothing special in using this particular input set, and that our methodology is flexible enough so that similarly successful results are obtained using other PDF sets as input. In particular, we want to explicitly verify that 'true' closure test is successful, i.e. that we can correctly reproduce NNPDF3.0, a PDF set generated using the same methodology and dataset, in closure test fits. I will also demonstrate that the closure test is successful even when using a comparatively more complicated set of PDFs than the relatively simple form used for the MSTW2008 parametrisation.

Input set	$\chi^2_{\rm input}$	$\chi^2_{\rm fit}$	Δ_{χ^2}
MSTW2008	1.013	1.002	-0.011
MSTW2008 (seed 2)	0.956	0.947	-0.010
CT10	1.036	1.028	-0.007
NNPDF3.0	0.976	0.976	0.0007
NNPDF3.0 (w/ positivity)	0.976	0.976	-0.0001
NNPDF3.0, replica 22	1.055	1.056	0.002

Table 6.1: χ^2s to the closure test pseudo-data for the input and fitted PDFs in fits in cases with different input PDFs. The Δ_{χ^2} statistic described in Appendix B is also shown.

Fig. 6.13 shows the distances between the fitted and the input PDFs for the closure test fit which uses the CT10 NLO PDF set to generate the data. These are the equivalent of the corresponding results obtained using MSTW08 as input shown in Fig. 6.8. We observe that, just as we found for MSTW2008, the fitted PDFs are in

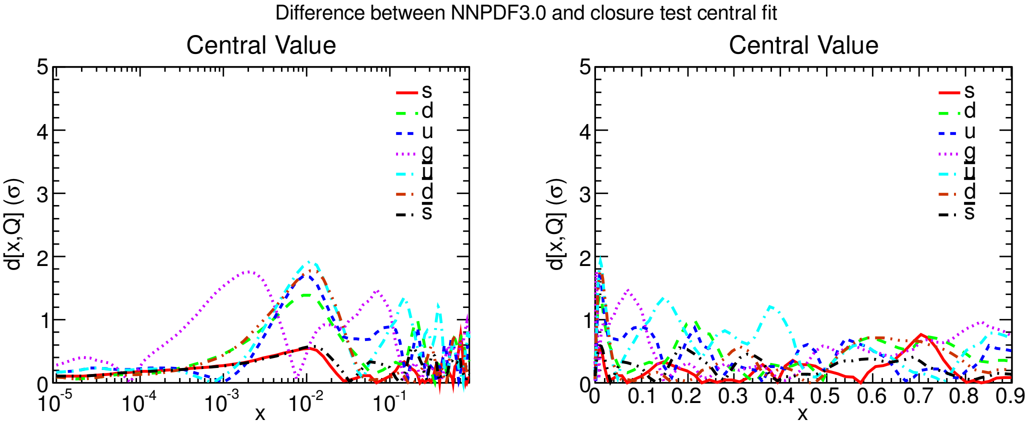


Figure 6.14: Same as Fig. 6.13 for a closure test based on NNPDF3.0 as input PDFs, without positivity constraints.

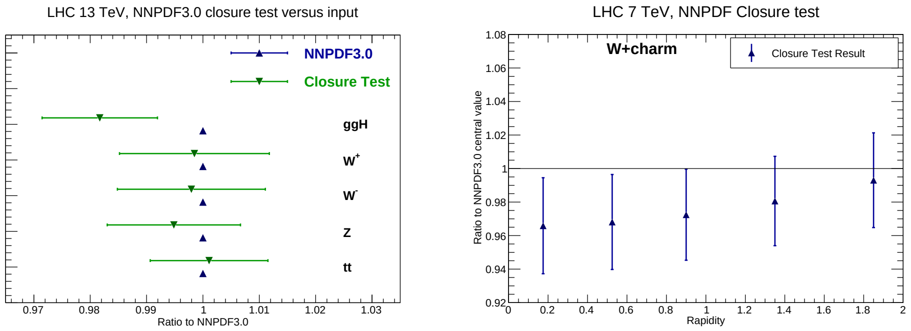


Figure 6.15: (left) Comparison of closure test based on NNPDF3.0 pseudo-data to NNPDF3.0 for various 13 TeV LHC processes. (right) Same but for $W + c$ at different rapidity values.

good agreement with the input mostly at the one-sigma level and with a few larger deviations. In this respect, the closure test based on CT10 is as successful as that based on MSTW08. We calculated Δ_{χ^2} with the CT10 pseudo-data, and this is shown in Table 6.1. It is very close to zero, in fact slightly closer than the equivalent figure for MSTW, indicating that the central values of the data points are reproduced to a very high standard.

The distances for a closure test using the NNPDF3.0 NLO PDFs are shown in Fig. 6.14. Again, the agreement is as good as that obtained using other PDFs. Perhaps because this is a self-closure test, the Δ_{χ^2} for this test, also shown in Table 6.1, is an order of magnitude smaller than was found for MSTW and CT10 closure tests.

For a closure test to NNPDF3.0 PDFs we can also look at how well LHC observables are reproduced in closure tests. This is not possible for the other sets due to the difference treatment of PDF evolution, heavy quark masses etc. used by the other PDF collaborations, which introduce deviation between the input and closure results.

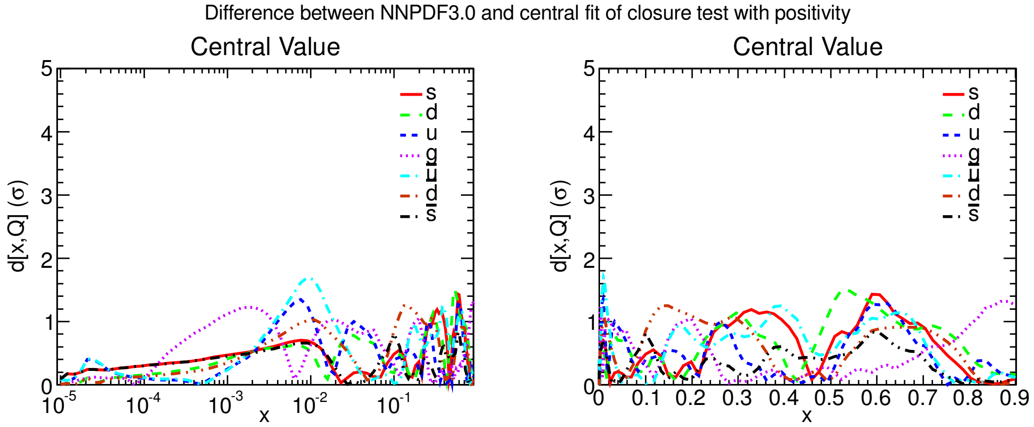


Figure 6.16: Same as Fig. 6.14 for a NNPDF3.0 level 2 closure test including positivity constraints during the minimisation. See text for more details.

Fig. 6.15 shows calculations of a variety of LHC observables using a closure test fit based on NNPDF3.0-derived pseudo-data, compared to similar values calculated with the NNPDF3.0 PDFs themselves. The left-hand plot compares inclusive cross-sections for vector-boson production (computed with Vrap [182]), top pair production (top++ [156]), and Higgs production by gluon-gluon fusion (ggHiggs [183]), while the right-hand plot shows the differential cross-section for $W^+ + \bar{c}$ production. Here we can again see the good reproduction of the input PDFs, with the closure test results generally being consistent with the NNPDF3.0 values at the one-sigma level. The largest difference is seen for the ggH cross-section, where the closure test is about two standard deviations from the NNPDF3.0 value, though this is perfectly consistent with a statistical fluctuation.

All of the closure test fits shown so far have been performed without the positivity constraints used in the fits to real data, described in Section 3.2.6. The motivation for this is that some of the input PDFs used in the closure tests, in particular MSTW08, do not satisfied all of the constraints, and therefore including them would potentially introduce tension between the generated pseudo-data and the positivity constraints during the minimisation. However, as the NNPDF3.0 PDFs are produced with the constraints, they satisfy them by construction (this is also verified in Section 7.3.2). Therefore if we use NNPDF3.0 as input PDF we can include positivity in the closure test and expect it to have no effect on the results. Fig. 6.16 shows the distances for another closure test using the NNPDF3.0 NLO PDFs, now with the positivity constraints imposed during the closure test fit. We indeed find that the level of agreement is similar to the first NNPDF3.0 closure test, and this is confirmed by the very essentially identical χ^2 shown in Table 6.1.

Finally, Fig. 6.17 shows the distances for a closure test in an extreme case, where

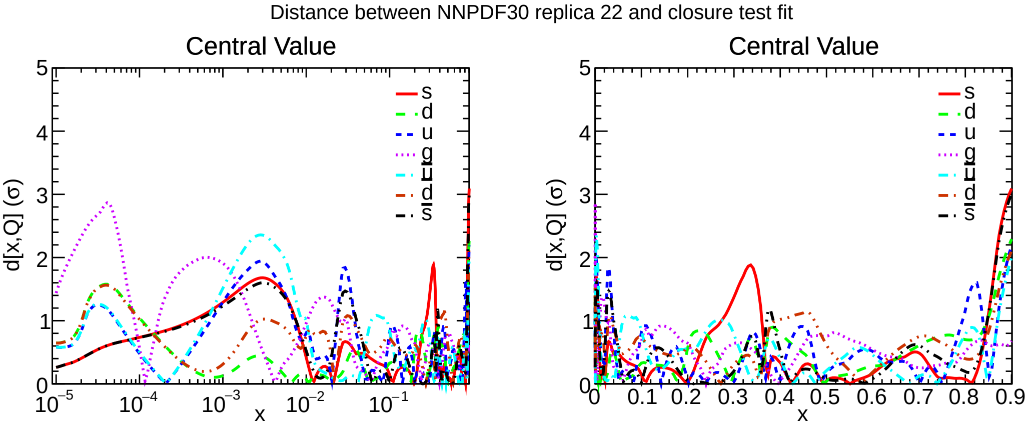


Figure 6.17: Distances (same as Fig. 6.13) between the fitted and input PDFs for a closure test fit to pseudo-data generated using replica 22 of the NNPDF3.0 NLO PDF set.

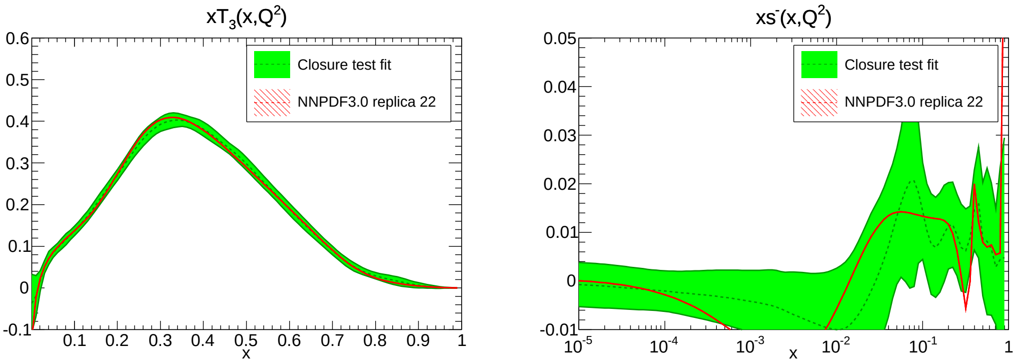


Figure 6.18: Comparison of the fitted (green) and input (red) PDFs from a closure test to replica 22 of the NNPDF3.0 NLO set. The T_3 distribution is shown with a linear scale in x on the left, while the s^- distributions is plotted with a logarithmic scale on the right.

a single replica from the NNPDF3.0 NLO set to generate the pseudo-data. While the central value of NNPDF sets are roughly as smooth as MSTW and CT PDFs, the individual replicas are in general fluctuate a lot more. As the distances show, even with these irregular PDFs our methodology can successfully reproduce the input PDFs, especially at medium x. The large distances at very small and large x are due to the unpredictable behaviour of the input PDFs in the extrapolation region where there is little data. Fig. 6.18 compares the fitted PDF to the central value of the input replica PDFs for the large-x T_3 and small x s^- distributions, showing both the erratic shape of the input PDFs and the excellent closure test reproduction.

6.5 Conclusions

In this chapter, I have demonstrated that the updated NNPDF methodology passes the closure test, i.e. that it is capable of successfully reproducing a known correct result to a very high standard. I have also provided evidence that the PDF uncertainties we obtain through our Monte Carlo approach have the many of the statistical features that they are required to have. We can therefore have a great deal of confidence in our fits applying the same methodology to the real experimental data. In the next chapter I will present results for the final NNPDF3.0 PDF sets.

Chapter 7

NNPDF3.0 Results

In this chapter I will present results from the NNPDF3.0 LO, NLO and NNLO global fits. First, I will discuss quality of fit to the experimental data and the dependence of the χ^2 on its exact definition, and on details of the treatment of systematic and normalisation uncertainties. This will be followed by results for PDFs themselves, where I will compare the new sets with previous NNPDF2.3 results and with other existing PDF sets.

In the second section I will explore the dependence of the NNPDF3.0 PDFs on the choice of experimental dataset. I will study a wide range of variations of the fitted dataset, including fits without LHC data, fits without jet data, and fits using only HERA and LHC data. Fits to reduced datasets will also be used to study the impact of jet data on the global fit, and to look at the strangeness of the proton, something which has been the object of various recent studies. I will also present a conservative PDF set, based on a dataset defined by an assessment of the consistency of an individual dataset with the global fit.

I then turn to an assessment of the stability of the NNPDF3.0 results upon variations in the fitting methodology. I will repeat some of the tests in Chapter 5 with fits to the real experimental data, and also look at other aspects which couldn't be studied in closure tests. I will first look at NNPDF3.0 fits based on the NNPDF2.3 dataset, which provide a way of disentangling the data and methodology changes in NNPDF3.0. I will then look at a range of issues including the impact of positivity constraints, the stability upon change of fitting basis, and the dependence on whether the systematic experimental uncertainties are treated as additive or multiplicative.

Finally, I will study the implications of NNPDF3.0 for LHC phenomenology, including for PDF luminosities, standard model and Higgs cross-sections, and searches for massive BSM particles.

7.1 The NNPDF3.0 set of parton distributions

7.1.1 Fit quality

Table 7.1 shows the results for the fit quality of the global LO, NLO and NNLO NNPDF3.0 sets. The values shown are calculated for a common value of $\alpha_s(M_Z) = 0.118$. Both the central t_0 and experimental χ^2 per data point are given for all sets, along with the number of data points used in the fit. Note that the precise definition of the t_0 χ^2 varies with the perturbative order, as it depends on the theoretical values of the cross-sections included in the fit.

As mentioned in Section 3.2.3, the t_0 χ^2 is used during the minimisation as it corresponds to an unbiased maximum-likelihood estimator even in the presence of multiplicative uncertainties. The experimental χ^2, on the other hand, is based on the experimental covariance matrix released by the experimental collaborations, and while it cannot be used for minimisation, it is best suited for benchmarking as it only depends on publicly available results (the final PDFs and the experimental covariance matrix).

The overall fit quality in the NLO and NNLO fits is good, with a central experimental χ^2 of 1.23 at NLO and 1.29 at NNLO. The LO fit in contrast has a much poorer fit quality, as we would expect due to the missing and relatively large NLO corrections. Exploring further into the table, while for some experiments like CHORUS, SLAC, ATLAS high-mass Drell-Yan, the W lepton asymmetry or top quark pair production, the χ^2 improves when going from NLO to NNLO, for most of the experiments it remains either very similar or gets slightly worse. This is also the case for the new HERA-II datasets. For the jet data the fit quality is quite similar at NLO and NNLO using the t_0 definition, but note that the kinematical cuts in the two cases are often very different (see Section 4.3.1). This is also the case for the CMS double differential Drell-Yan data: the χ^2 is slightly worse at NNLO but this is because at NLO we impose kinematical cuts that remove the region with large NNLO corrections. Without such cuts, the χ^2 at NLO is substantially poorer.

Another interesting feature that one can observe from Table 7.1 is that the numerical differences between the two definitions of the χ^2 can be substantial. This effect is particularly acute for experiments where multiplicative systematic uncertainties dominate, as we would expect, and emphasises the crucial role of careful estimation of systematic errors in PDF fitting. One such example is provided by the CMS inclusive jet data, where for the NNLO fit the central χ^2 is 1.90 for the experimental definition and 1.07 for the t_0. These large differences may at first glance appear quite alarming, however I will show in Section 7.3.4 that these differences in the value of the χ^2 do

	LO			NLO			NNLO		
	N_{dat}	χ^2_{exp}	$\chi^2_{\mathrm{t_0}}$	N_{dat}	χ^2_{exp}	$\chi^2_{\mathrm{t_0}}$	N_{dat}	χ^2_{exp}	$\chi^2_{\mathrm{t_0}}$
Total	4258	2.42	2.17	4276	1.23	1.25	4078	1.29	1.27
NMC d/p	132	1.41	1.09	132	0.92	0.92	132	0.93	0.93
NMC	224	2.83	3.3	224	1.63	1.66	224	1.52	1.55
SLAC	74	3.29	2.96	74	1.59	1.62	74	1.13	1.17
BCDMS	581	1.78	1.78	581	1.22	1.27	581	1.29	1.35
CHORUS	862	1.55	1.16	862	1.11	1.15	862	1.09	1.13
NuTeV	79	0.97	1.03	79	0.70	0.66	79	0.86	0.81
HERA-I	592	1.75	1.51	592	1.05	1.16	592	1.04	1.12
ZEUS HERA-II	252	1.94	1.44	252	1.40	1.49	252	1.48	1.52
H1 HERA-II	511	3.28	2.09	511	1.65	1.65	511	1.79	1.76
HERA σ^{c}_{NC}	38	1.80	2.69	47	1.27	1.12	47	1.28	1.20
E886 d/p	15	2.04	1.10	15	0.53	0.54	15	0.48	0.48
E886 p	184	0.98	1.64	184	1.19	1.11	184	1.55	1.17
E605	119	0.67	1.07	119	0.78	0.79	119	0.90	0.72
CDF Z rapidity	29	2.02	3.88	29	1.33	1.55	29	1.53	1.62
CDF Run-II k_t jets	76	1.51	2.12	76	0.96	1.05	52	1.80	1.20
D0 Z rapidity	28	1.35	2.48	28	0.57	0.68	28	0.61	0.65
ATLAS W, Z 2010	30	5.94	3.20	30	1.19	1.25	30	1.23	1.18
ATLAS 7 TeV jets 2010	90	2.31	0.62	90	1.07	0.52	9	1.36	0.85
ATLAS 2.76 TeV jets	59	3.88	0.61	59	1.29	0.65	3	0.33	0.33
ATLAS high-mass DY	5	13.0	15.6	5	2.06	2.84	5	1.45	1.81
ATLAS W p_T	-	-	-	9	1.13	1.28	-	-	-
CMS W electron asy	11	10.9	0.95	11	0.87	0.79	11	0.73	0.70
CMS W muon asy	11	76.8	2.25	11	1.81	1.80	11	1.72	1.72
CMS jets 2011	133	1.83	1.74	133	0.96	0.91	83	1.9	1.07
CMS $W + c$ total	5	11.2	25.8	5	0.96	1.30	5	0.84	1.11
CMS $W + c$ ratio	5	2.04	2.17	5	2.02	2.02	5	1.77	1.77
CMS 2D DY 2011	88	4.11	12.8	88	1.23	1.56	110	1.36	1.59
LHCb W rapidity	10	3.17	4.01	10	0.71	0.69	10	0.72	0.63
LHCb Z rapidity	9	5.14	6.17	9	1.10	1.34	9	1.59	1.80
$\sigma(t\bar{t})$	6	42.1	115	6	1.43	1.68	6	0.66	0.61

Table 7.1: The values of the χ^2 per data point for the LO, NLO and NNLO central fits of the NNPDF3.0 family with $\alpha_s(M_Z) = 0.118$, obtained using both the experimental and the t_0 definitions.

not convert into a large impact on the PDFs, which are rather stable upon changes of the χ^2 definition. The dependence of the χ^2 on its definition is weaker for fixed target experiments and DIS data, for which statistical uncertainties are dominant.

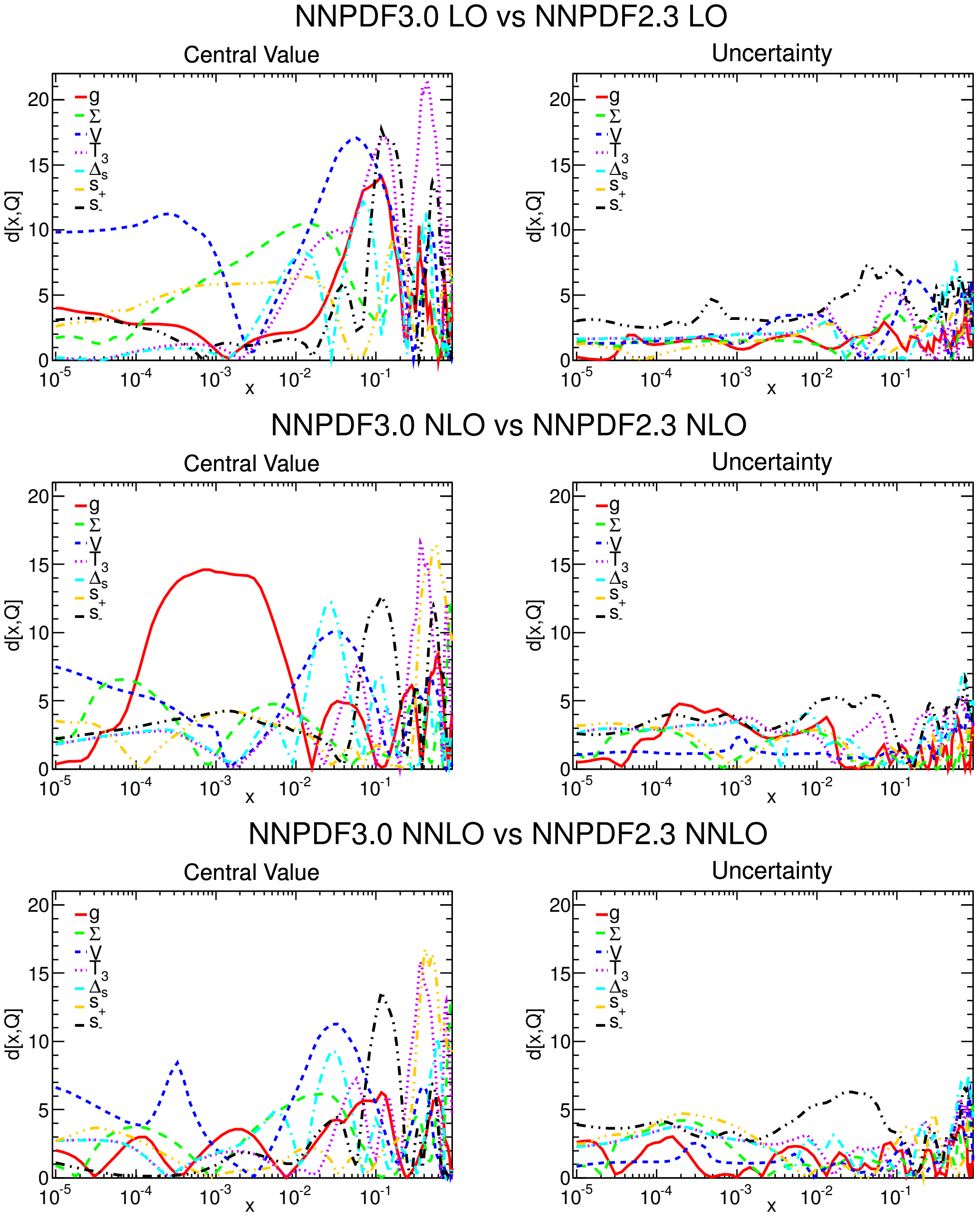
7.1.2 Parton distributions

Fig. 7.1 shows the distances (see Appendix A) between the parton distributions of the NNPDF3.0 and NNPDF2.3 sets for each of the three perturbative orders, LO, NLO and NNLO. As mentioned in Appendix A, when comparing two sets of 100 replicas, $d < 2$ means that the two sets are statistically indistinguishable (they have differences on the level of two different sets of replicas extracted from the same underlying probability distribution), while $d \sim 10$ means that the sets correspond to PDFs that disagree at the one-sigma level. The distances shown here, and in the rest of this chapter, are computed at a scale of $Q^2 = 2$ GeV^2, and are produced using the $\alpha_s(M_Z) = 0.118$ sets. For the LO plot, the $\alpha_s(M_Z) = 0.119$ NNPDF2.3 is used instead, as $\alpha_s(M_Z) = 0.118$ is not available for NNPDF2.3 LO. This has a minor effect on the comparison.

As Fig. 7.1 demonstrates, the size and character of the differences between the NNPDF3.0 and 2.3 PDFs vary significantly with the perturbative order. At LO, the gluon is in very good agreement between the two sets for $x <\sim 0.01$. This suggests that Monte Carlo tunes, which strongly depend on the small-x gluon, based on NNPDF2.3LO—such as the Monash 2013 tune of Pythia8 [184]—should also work reasonably well with NNPDF3.0LO. On the other hand, larger differences, going up to about two sigma, are found for both the quarks and the gluon at medium and large x. It is worth noting that at LO theory uncertainties dominate over PDF uncertainties, so the actual impact of these differences will likely be quite small.

At NLO and NNLO, NNPDF2.3 and NNPDF3.0 are typically in agreement at the one-sigma level, with occasionally somewhat larger distances of order 1.5–sigma. In particular, while the total quark singlet PDF is relatively stable, there are larger differences for individual quark flavours, especially at medium and large-x. Significant differences can also been seen for the gluon PDF, especially at NLO, though here it should be noted that NNPDF2.3 used the FONLL-A treatment of heavy quarks, while NNPDF3.0 uses FONLL-B (see Section 4.2.3). This comparison also shows that PDF uncertainties change at the level of one sigma: this is to be expected, as a consequence of the constraints coming from new data, and the improved fitting methodology.

A direct comparison of NNPDF2.3 and NNPDF3.0 NLO PDFs can be seen in Fig. 7.2, where the gluon, singlet PDF, isospin triplet and total valence PDFs from the two sets are plotted, again with $\alpha_s(M_Z) = 0.118$ at $Q^2 = 2$ GeV^2. We can see that in the NNPDF3.0 NLO set the central value of the gluon remains positive, even



NNPDF3.0 LO vs NNPDF2.3 LO

Central Value

Uncertainty

NNPDF3.0 NLO vs NNPDF2.3 NLO

Central Value

Uncertainty

NNPDF3.0 NNLO vs NNPDF2.3 NNLO

Central Value

Uncertainty

Figure 7.1: Distances between NNPDF2.3 and NNPDF3.0 at LO (top), NLO (center) and NNLO (bottom) PDFs, computed between sets of $N_{\text{rep}} = 100$ replicas at $Q^2 = 2\text{GeV}^2$. All PDFs use $\alpha_s(M_Z) = 0.118$, except the LO NNPDF2.3 set which has $\alpha_s(M_Z) = 0.119$.

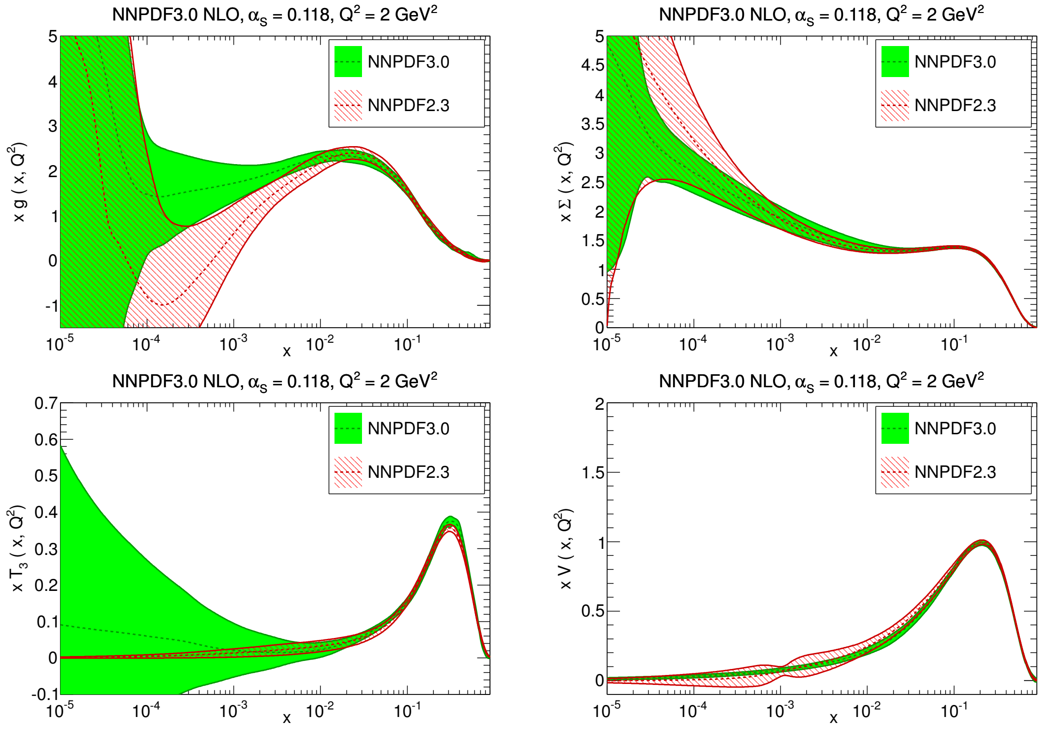


Figure 7.2: Comparison of NNPDF2.3 and NNPDF3.0 NLO PDFs at $Q^2 = 2$ GeV^2 with $\alpha_s(M_Z) = 0.118$. From top to bottom and from left to right the gluon, singlet, isospin triplet and total valence are shown.

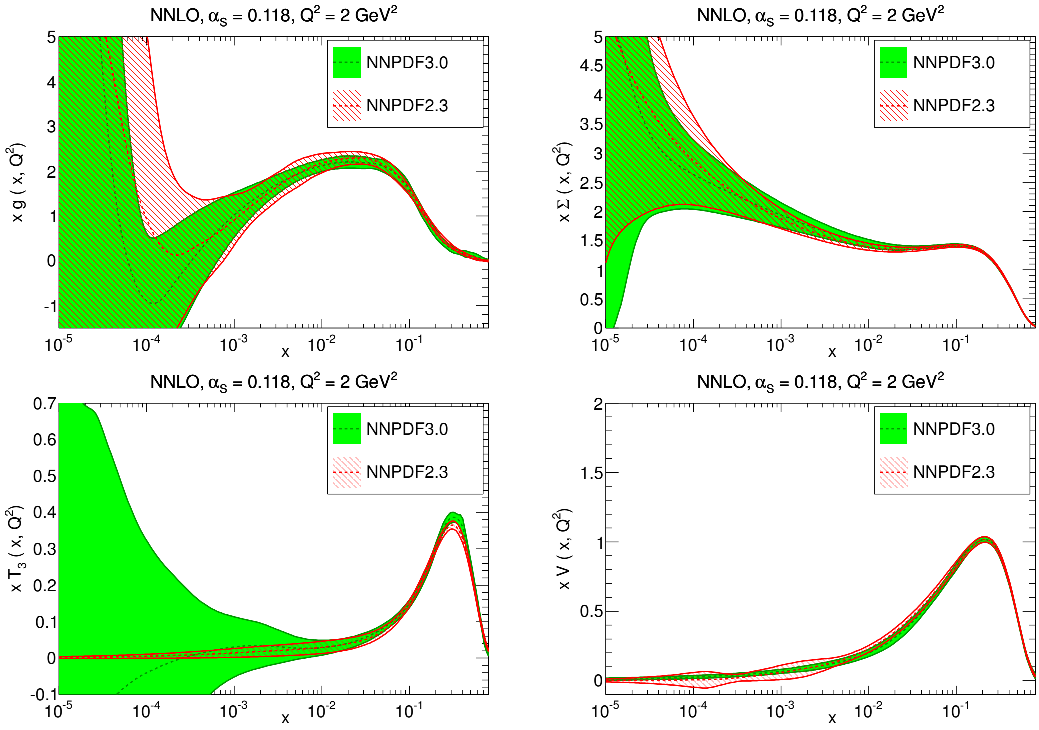


Figure 7.3: Same as Fig. 7.2 but at NNLO.

at small-x. It is flat down to $x \sim 10^{-4}$ and then it begins to grow, within its large uncertainty, always remaining above its NNPDF2.3 counterpart. This difference can be understood, as mentioned before, to be a consequence of moving to the FONLL-B heavy quark scheme, and also due to the more stringent positivity constraints that are imposed in the new set. For the total quark singlet on the other hand there is good agreement between 2.3 and 3.0. For the quark triplet we see two interesting features: the larger uncertainty at small x, due to the changes to the preprocessing and sum rules described in Section 5.4.3, and also a difference at large x, where the 3.0 result is larger than from 2.3, especially in the region where the PDF peaks.

The same comparison are shown for the NNLO sets in Fig. 7.3. In this case, we can observe good consistency for the gluon PDF, with a reduction in the PDF uncertainties at small-x. Note that unlike at NLO, here both the 3.0 and 2.3 fits use the same FONLL-C GM-VFN scheme. For the quark singlet and triple PDFs the situation at NNLO is much the same as it was at NLO, with broad agreement for the singlet and the specific differences mentioned for the triplet.

It is interesting to also perform a comparison of the NNPDF2.3 and 3.0 sets at the higher scale of $Q^2 = 10^4$ GeV^2, typical of LHC processes. Results for this comparison

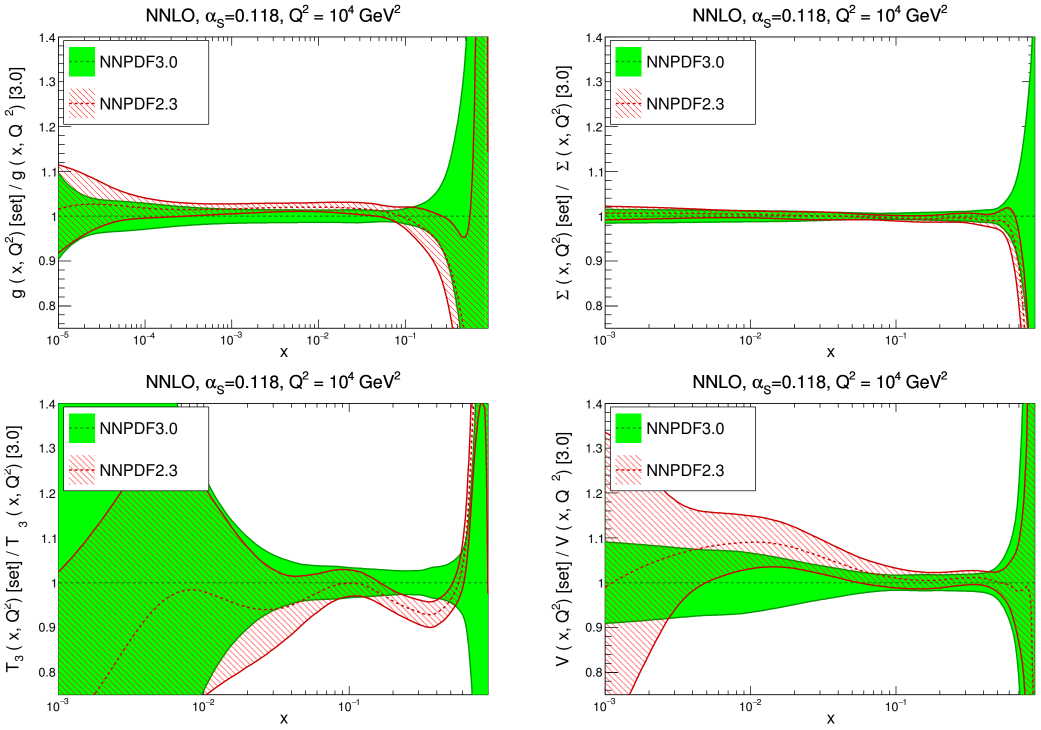


Figure 7.4: Same as Fig. 7.3, but at $Q^2 = 10^4$ GeV^2, and with results shown as ratios to the NNPDF3.0 central value.

at NNLO are shown Fig. 7.4, in this case as ratios to the NNPDF3.0 central value. We see that the two PDF sets agree typically at the one-sigma level or better, with a small number of exceptions. The NNPDF3.0 gluon is somewhat softer than in NNPDF2.3, in particular in the region around $x \sim 0.01$ which is important for the Higgs gluon fusion cross-section. There is very good agreement in the quark singlet, as we would expect from the low scale results above. For the triplet there is good agreement, except near $x \sim 0.3$ where the NNPDF2.3 and 3.0 fits disagree at about the two-sigma level, and again at $x \sim 0.02$ where there is about a one sigma difference. For the total valence PDF there is a reasonable agreement at large x, with disagreement going to smaller values of x, growing to a maximum of about 1.5 sigma at $x \sim 10^{-2}$.

Another set of comparisons useful for evaluating the phenomenological impact of these changes is the parton luminosities. Following Ref. [185], we define the parton luminosity for the ij initial state as

$$\Phi_{ij}\left(M_X^2\right) = \frac{1}{s}\int_{\tau}^{1} \frac{dx_1}{x_1} f_i\left(x_1, M_X^2\right) f_j\left(\tau/x_1, M_X^2\right) \ , \tag{7.1}$$

where $f_i(x, M_X^2)$ is the PDF for the ith parton, $\tau \equiv M_X^2/s$ and M_X is the invariant

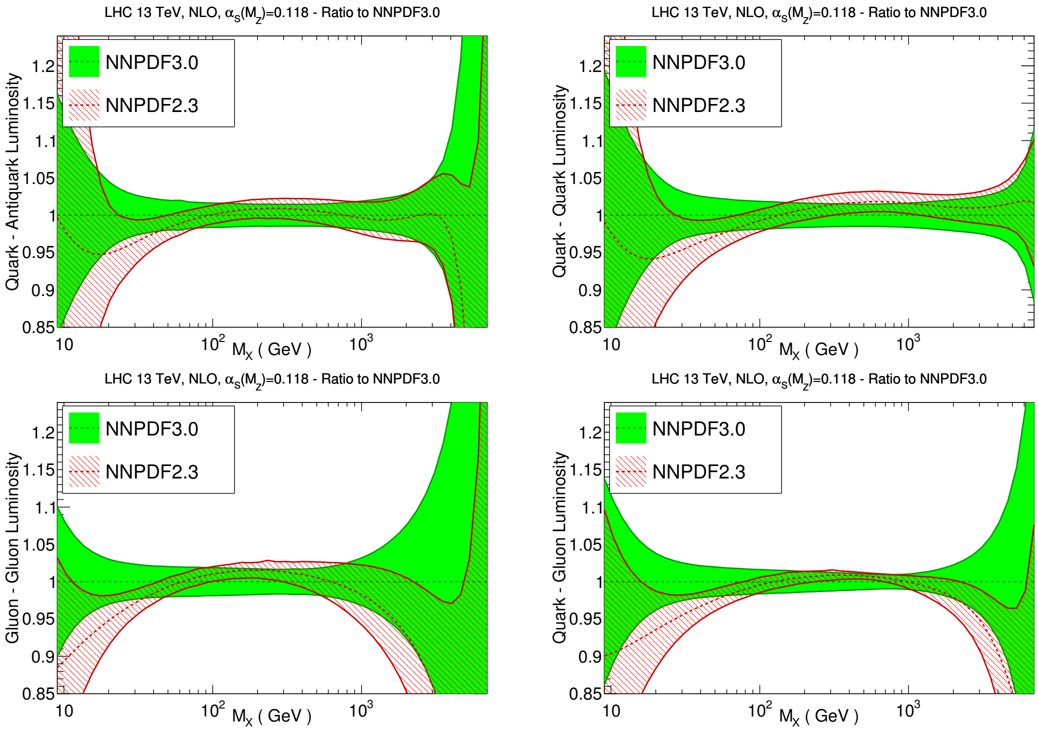


Figure 7.5: Parton luminosities, Eq. 7.1 computed using NNPDF2.3 and NNPDF3.0 NLO PDFs with $\alpha_s(M_Z) = 0.118$, as a function of the invariant mass of the final state M_X. Results are shown as ratios to NNPDF3.0. From top to bottom and from left to right the $q\bar{q}$, qq, qq and qg luminosities are shown.

mass of the final state.

Figs. 7.5 and 7.6 compare the gg, qq, $q\bar{q}$ and qg luminosities obtained using NNPDF2.3 and 3.0 PDF sets for $\sqrt{s}$=13 TeV and $\alpha_s(M_Z) = 0.118$ (where for quarks a sum over light flavours is understood). The NLO comparisons are shown in Figs. 7.5 and NNLO in 7.6. At NLO, we generally find agreement at the one-sigma level or below in all cases, with slightly more disagreement below about $M_X \sim 40$ GeV and in the qg channel above 1 TeV, where the luminosity is rather larger in NNPDF3.0 than in NNPDF2.3. Note that in the gg channel in the region around 100-200 GeV the NNPDF3.0 luminosity is somewhat softer than in NNPDF2.3, though always in agreement within PDF uncertainties.

At NNLO, in the qq and $q\bar{q}$ channels there is generally good agreement, with differences within one sigma. For $q\bar{q}$, the NNPDF3.0 luminosity is slightly larger at high invariant masses, while for qq around 500 GeV NNPDF3.0 is somewhat lower. More significant differences are found in the gg and qg channels, where in both cases the luminosity at medium invariant masses is smaller by more than one sigma in NNPDF3.0

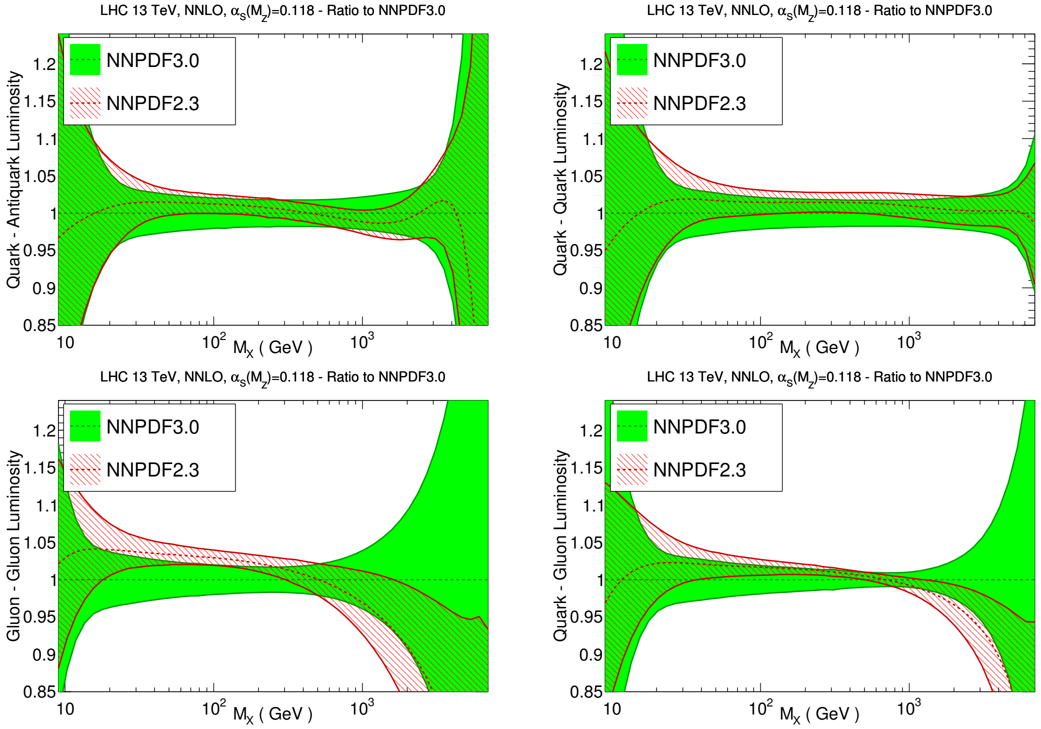


Figure 7.6: Same as Fig. 7.5 at NNLO.

than in NNPDF2.3. In particular, for 30 GeV $\leq M_X \leq$ 300 GeV, the gg one-sigma bands barely overlap. This has important consequences for gluon-initiated processes such as inclusive Higgs production, as will be shown in Section 7.4.3 below. As discussed in Section 7.1.2, these differences stem from a combination of the improved fitting methodology and the new constraints from HERA and LHC data.

7.1.3 Perturbative stability

In the NNPDF approach the same methodology is used at all orders, with only the underlying QCD theory (and to a small extent the dataset) changing from one order to the next. Comparing the results at different orders is therefore a meaningful comparison. Fig. 7.7 shows the distances between the NNPDF3.0 pairs of fits at consecutive orders: LO vs. NLO and NLO vs. NNLO. In the former case, the main variation is as would be expected seen in the gluon PDF, which is very different at LO. There are also significant differences in the large-x quarks. As mentioned in the previous section, at LO theory uncertainties completely dominate over the PDF uncertainty, which depends on the data and is roughly the same at all orders, as the right hand plots in these comparisons show.

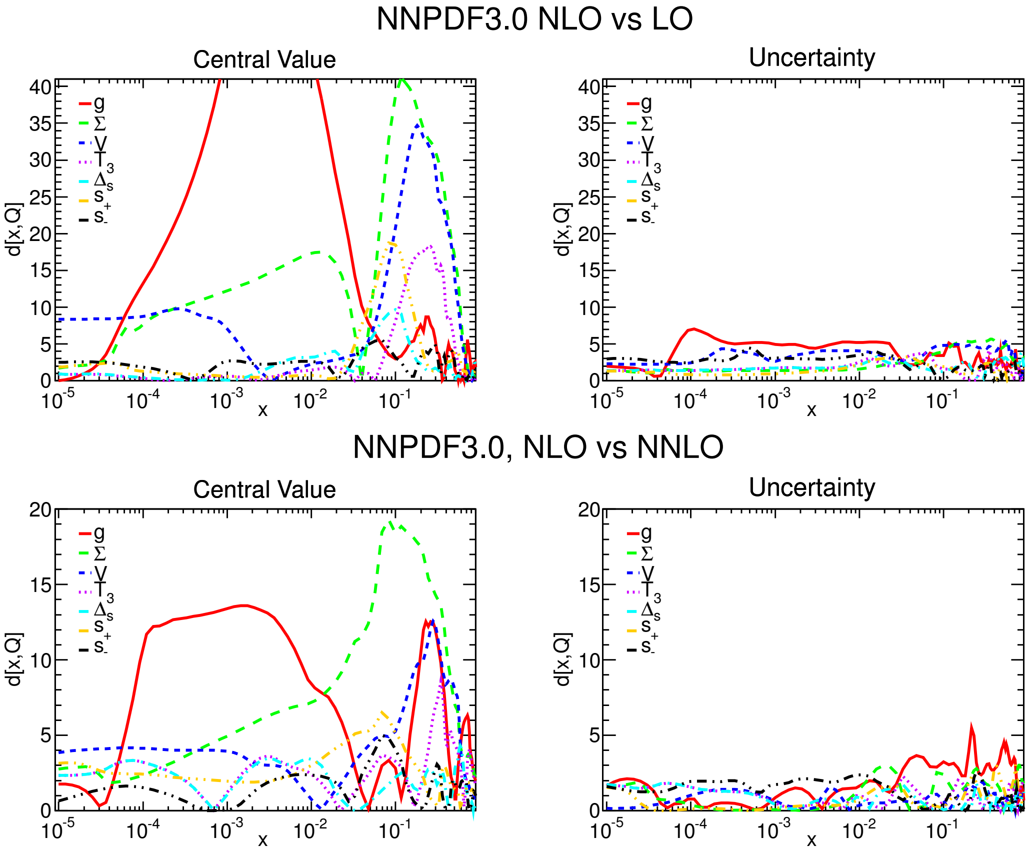


Figure 7.7: Same as Fig. 7.1, but now comparing NNPDF3.0 LO vs NLO (top) and NLO vs. NNLO (bottom).

Distances are generally smaller when comparing NLO to NNLO (note the difference in the y-axis scale). For central values, the main differences are seen in the gluon PDF, both at small x and at large x, and in the medium- and large-x quarks, in particular the total quark singlet. Uncertainties are again quite stable, with the exception on the large-x gluon, where the PDF uncertainties are larger at NNLO because of the additional cuts applied to the jet data for this order (see Section 4.3.1). These differences in the details of the jet dataset used in the fit also impact the central values of the two fits.

Next, Fig. 7.8 compares directly the LO, NLO and NNLO NNPDF3.0 parton distributions at $Q^2 = 2$ GeV^2. The large shift in the gluon between LO and NLO described above, and its subsequent stability at NNLO, is clearly seen. Specifically, the LO gluon is very large, compensating for missing NLO terms in the DIS splitting functions and anomalous dimensions. At NLO, the small-x gluon is rather flatter than the NNLO one, which goes almost negative at small-x. This relatively unstable perturbative behaviour of the small-x gluon might be related to unresummed small-x

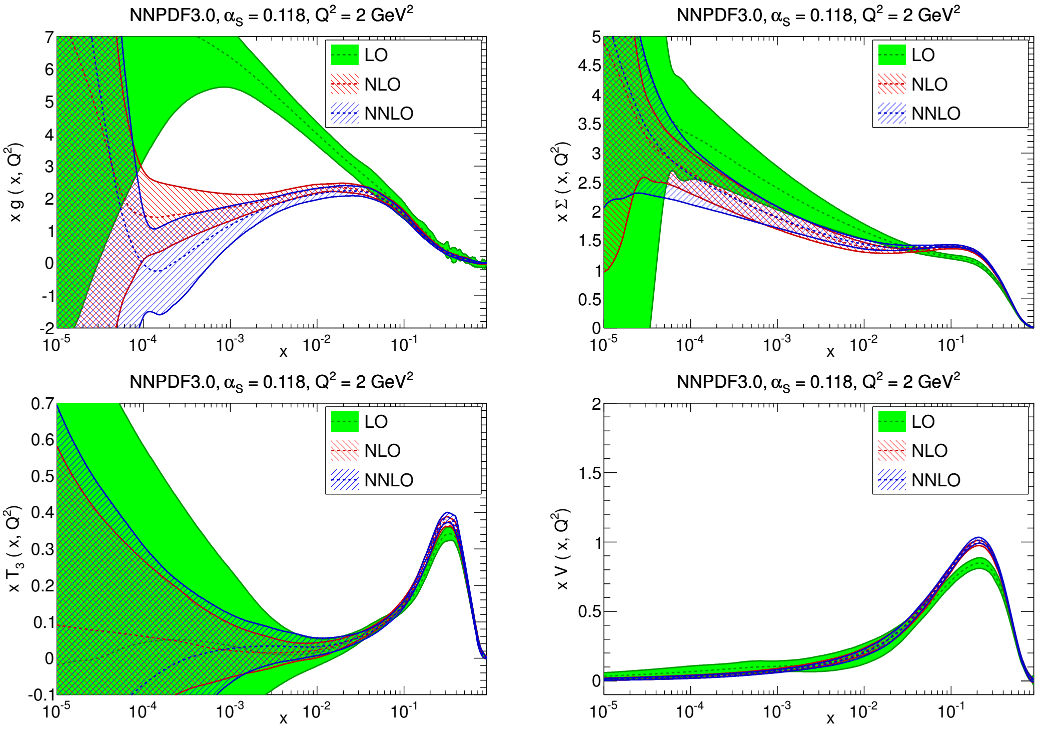


Figure 7.8: Same as Fig. 7.2, but now comparing NNPDF3.0 LO, NLO and NNLO PDFs.

perturbative corrections [186]. Quark PDFs are generally quite stable, with NNLO and NLO mostly in agreement at the one-sigma level, though sizeable shifts are seen in the singlet in the region around $x \sim 0.1$ when going from LO to NLO and NLO to NNLO.

7.1.4 Model uncertainties

While uncertainties related to higher order corrections are perhaps the largest source of uncertainty not included in the standard PDF uncertainty, there are a few more sources of uncertainty which are also not part of the current PDF uncertainty, and which might become relevant as the precision of the data increases. One source is to do with further approximations which are made in the theoretical description of the data, which here I will refer to as “model” uncertainties. This section will discuss some, likely dominant, sources of model uncertainties, namely those related to deuteron nuclear corrections and those related to the treatment of heavy quarks.

Several fixed-target data included in the NNPDF3.0 PDF determination are based on scattering on nuclear targets. This includes all of the neutrino deep-inelastic scattering data (CHORUS, NuTeV), the data for charged-lepton deep-inelastic scattering from deuteron targets (NMC, BCDMS, SLAC), and the data for Drell-Yan

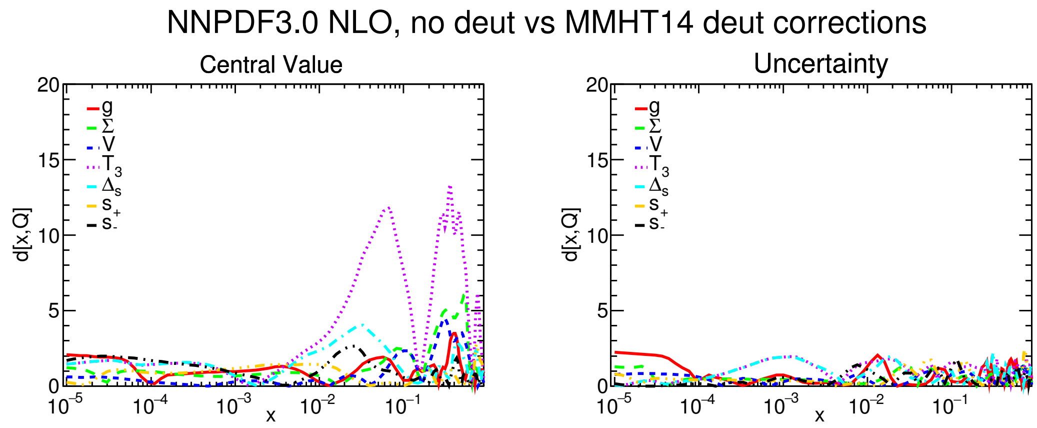


Figure 7.9: Same as Fig. 7.1, but now comparing the NNPDF3.0 NLO PDFs with and without deuteron nuclear corrections.

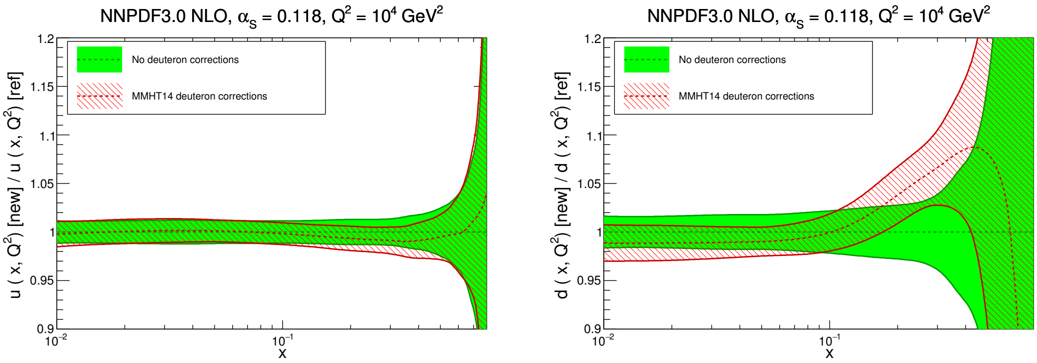


Figure 7.10: Same as Fig. 7.4, but now comparing the NNPDF3.0 NLO fit with and without deuterium nuclear corrections. From left to right the up and down quark PDFs are shown.

production on a deuterium target in the DY E866 dataset. The impact of nuclear corrections on the NNPDF2.3 PDF determination has previously been discussed in Ref. [168]. There, the NNPDF2.3 fit was repeated introducing deuteron nuclear corrections according to a number of models, and found non-negligible (up to about one and a half sigma) but very localized, impact on the down distribution at large x.

To look at the impact of deuteron corrections on the NNPDF3.0 PDF determination, we have again repeated the fit, but now including deuterium corrections according to the recent model of Ref. [24], which supersedes the previous treatment of higher twist corrections of Ref. [187], considered in Ref. [168]. The distances between resulting PDFs with deuteron corrections and the standard PDFs are shown in Fig. 7.9. The pattern of deviations here is very similar to that seen Ref. [168], but with a somewhat more moderate impact, as one might expect given the larger dataset used in NNPDF3.0.

Essentially only the up and down quark distributions are affected, and by comparing the PDFs in Fig. 7.10 it is apparent that the effect is always below one sigma. In view of the theoretical uncertainty involved in the modeling of these corrections, we prefer not to include them in the fit, as it is unclear that the uncertainty on them is significantly smaller than their actual size. Nuclear corrections to neutrino data are likely to be yet smaller, with the possible exception of the strange distribution [24].

Another important potential source of theoretical uncertainty is related to the treatment of heavy quarks. As discussed in Section 4.2.3, we use a computational scheme, the FONLL scheme, which ensures that all available perturbative information is included. However, there are also aspects that go beyond perturbation theory. In particular, the dependence on the quark mass itself, and the possible presence of an intrinsic heavy quark component [188].

The dependence of PDFs on the values of the heavy quark masses was previously studied in Ref. [56] within the context of the NNPDF2.1 PDF determination, where the values of m_c and m_b were varied, in the absence of intrinsic heavy quark PDFs. The main result of this study was that in such a case the value of the heavy quark mass mostly affects the threshold for generating the heavy quark by perturbative evolution, with a lower mass value corresponding to a larger PDF at a given scale, due to a longer evolution. However this also suggests that, while for the b quark this dependence on the quark mass value is likely to be physical, for charm, which is at the boundary of the perturbative region and might have a non-negligible intrinsic component, the dependence on the mass is unphysical, and would be reabsorbed by an intrinsic PDF.

As mentioned in Section 4.2.3, Eq. 4.2, the heavy quark mass values used in the current NNPDF3.0 PDF determination differ from the values previously used in the NNPDF 2.3 determination, as we now use the $\overline{\mathrm{MS}}$ PDG mass values, while the previous values were essentially the pole mass values. This shift is larger than the current uncertainty on $\overline{\mathrm{MS}}$ masses. In order to assess the impact of this change, and thus also of the dependence on heavy quark masses, we have repeated the NNPDF3.0 PDF NLO determination using the same heavy quark mass values that were used for the NNPDF2.3 set. Fig. 7.11 compares the parton luminosities at $Q^2 = 10^4$ GeV^2 for the two sets of masses. Results are in agreement with the findings of Ref. [56], where a similar effect due to changes of the charm mass was found. The effect is not entirely negligible, however, as mentioned, it is likely that most of this dependence would be absorbed into an intrinsic charm PDF. At NLO, $\overline{\mathrm{MS}}$ and pole mass-scheme expressions coincide, with a small correction at NNLO, hence it seems more appropriate to use the more accurate $\overline{\mathrm{MS}}$ mass value. The shifts seen in Fig. 7.11 should be taken as an upper bound to the size of the uncertainty related to the charm mass value, the exact assessment of which will only be possible once an intrinsic charm component is

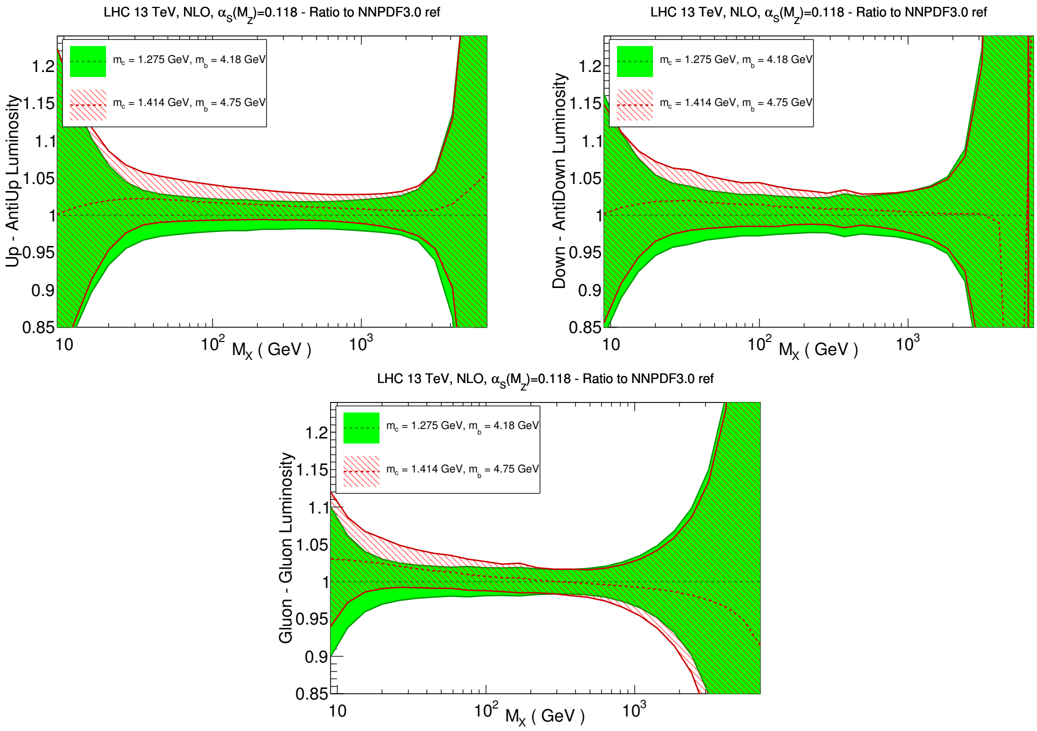


Figure 7.11: Dependence on the value of the heavy quark masses of parton luminosities Eq. 7.1 computed using NNPDF3.0 NLO PDFs with $\alpha_s(M_Z) = 0.118$. Results are shown as ratios to the default set. The up-antiup, down-antidown and gluon-gluon luminosities are shown.

introduced in our PDF fits.

Finally, it is also worth mentioning that further model uncertainties are expected to come from the treatment of electroweak interactions, both from the choice of parameters, and from the treatment of higher order terms (including mixed strong-electroweak corrections [189]). Though these are generally smaller than the uncertainties discussed here, they could become significant in particular kinematic regions or for specific processes, such as for instance high-mass production of W pairs.

7.2 Dependence on the dataset

7.2.1 Conservative PDFs from a consistent dataset

Inconsistencies between data which enter a global PDF determination can distort the statistical interpretation of PDF uncertainties. Inconsistency of any individual dataset with the bulk of the global fit may suggest that our understanding of it, either from the theoretical or experimental point of view, is not complete, and in these cases exclusion from the fit might be advantageous, despite the loss of information from doing so. In order to minimise such inconsistencies, "conservative" PDFs have been suggested, for example by introducing restrictive kinematic cuts which remove potentially dangerous regions [190], or by picking data which one might expect to be more reliable. One example of the latter are collider-only fits, for instance the NNPDF2.3 collider-only fit [45], which are based on the expectation that collider data, because of their higher energy, should be more reliable than fixed-target data.

For NNPDF3.0 we developed a new objective definition of a conservative dataset based on a measure of consistency between datasets introduced in Ref. [191, 192]. This idea is based on observing that lack of compatibility can always be viewed as an underestimate of the covariance matrix: if the covariance matrix is inflated by a factor α^2, then compatibility can always be attained if α^2 if large enough (crudely speaking, if uncertainties are all multiplied by a factor α). It is then possible to measure compatibility by assuming that the prior knowledge is given by all experiments in the global dataset but the given one, and using Bayes' theorem to study how this prior is modified when the excluded experiment is added. One can then compute the a posteriori probability $P\left(\alpha\right)$ that the covariance matrix of the given experiment should be rescaled by a factor α. Compatibility corresponds to the case in which $P(\alpha)$ peaks around $\alpha \sim 1$, while if the most likely value is at $\alpha_0 > 1$, this means that compatibility is only achieved when uncertainties are inflated by α_0. The probability distribution $P(\alpha)$ is calculated based on the weight penalty method described in Ref. [191, 192]. The t_0 definition of the χ^2, which is used for minimisation (see Section 7.1.1), is also used in the determination of $P(\alpha)$.

To generate the "conservative" dataset, we first computed the probability distribution of the rescaling variable α, $P(\alpha)$, for each dataset included in the global fit. In practice, for simplicity we compute the probability $P(\alpha)$ without excluding the given experiment from the global fit. This provides a conservative estimate of the compatibility (which is clearly increased by including the experiment under investigation in the prior) without requiring us to construct a new set of replicas for each experiment. We then exclude from the conservative fit all experiments for which at least two of the $P(\alpha)$ mean, median and mode are above a chosen threshold value, denoted by $\alpha_{\max}$. We discard all datasets for which the criterion fails either at NLO or at NNLO (or both), which corresponds to the most conservative choice of only retaining experiments which are well described at all perturbative orders, and has the obvious advantage that the resulting "conservative" dataset does not depend on the perturbative order. Because of this choice, we also exclude the ATLAS W p_T data, for which no NNLO prediction is available. The values of the mean, median and mode computed for all the experiments in the NNPDF3.0 global fits at NLO and NNLO are

	NLO global fit			NNLO global fit		
Experiment	mean	mode	median	mean	mode	median
NMC d/p	1.04	1.01	1.03	1.04	1.01	1.03
NMC $\sigma^{\mathrm{NC},p}$	1.32	1.31	1.27	1.27	1.26	1.27
SLAC	1.31	1.27	1.30	1.13	1.09	1.12
BCDMS	1.17	1.16	1.17	1.20	1.19	1.20
CHORUS	1.11	1.10	1.11	1.10	1.09	1.09
NuTeV	1.04	0.90	0.98	1.06	0.92	1.00
HERA-I	1.09	1.09	1.10	1.10	1.09	1.09
ZEUS HERA-II	1.23	1.22	1.23	1.25	1.24	1.25
H1 HERA-II	1.30	1.3	1.31	1.35	1.34	1.34
HERA σ_{NC}^{c}	1.10	1.06	1.09	1.14	1.11	1.13
E886 d/p	1.00	0.88	0.96	1.01	0.88	0.96
E886 p	1.13	1.11	1.12	1.15	1.14	1.15
E605	0.97	0.94	0.96	0.94	0.91	0.93
CDF Z rapidity	1.34	1.28	1.32	1.39	1.32	1.36
CDF Run-II k_t jets	1.09	1.06	1.08	1.15	1.12	1.14
D0 Z rapidity	1.34	1.28	1.32	0.86	0.82	0.85
ATLAS W, Z 2010	1.20	1.15	1.18	1.17	1.12	1.15
ATLAS 7 TeV jets 2010	0.76	0.74	0.75	1.09	0.92	1.02
ATLAS 2.76 TeV jets	0.86	0.83	0.85	1.07	0.57	0.83
ATLAS high-mass DY	2.22	1.68	2.03	1.82	1.34	1.63
CMS W electron asy	1.05	0.91	0.99	1.00	0.87	0.95
CMS W muon asy	1.62	1.42	1.54	1.60	1.40	1.53
CMS jets 2011	1.01	0.97	0.99	1.09	1.07	1.08
CMS $W{+}c$ total	1.60	1.17	1.42	1.50	1.09	1.33
CMS $W{+}c$ ratio	1.93	1.43	1.74	1.88	1.39	1.69
CMS 2D DY 2011	1.27	1.25	1.27	1.28	1.27	1.28
LHCb W, Z rapidity	1.10	1.02	1.07	1.20	1.12	1.17
$\sigma(t\bar{t})$	1.65	1.24	1.49	1.09	0.75	0.95

Table 7.2: The mean, mode and median of the $P(\alpha)$ distributions [191,192] (see text) for all the experiments in the NNPDF3.0 global fits, both at NLO (left) and at NNLO (right).

collected in Table 7.2.

Here results for “conservative” patrons obtained with values of $\alpha_{\max} = 1.1$, 1.2 and 1.3 will be presented. Table 7.3 gives the χ^2 (for ease of comparison we show results obtained using the experimental definition, see Section 7.1.1) for the PDF fits to these datasets. To facilitate the comparison with the global fit, we also provide its χ^2 values in the same table, taken from Table 7.1.

The improvement in global fit quality as $\alpha_{\max}$ is lowered is apparent, though perhaps unsurprising, with the most conservative option leading to an essentially perfect χ^2 of order one. It is interesting to observe that NMC proton data, which are known to have internal inconsistencies [193], as well as other datasets such as the H1 HERA-II data, the ATLAS high-mass Drell-Yan data, and the CMS $W{+}c$ data are excluded even from the least conservative set. On the other hand, the CMS inclusive jet data is included for all values of $\alpha_{\max}$ (note that for this dataset, and for several of the jet datasets,

	$\alpha_{\rm max}=1.1$		$\alpha_{\rm max}=1.2$		$\alpha_{\rm max}=1.3$		Global fit	
	$\chi^2_{\rm nlo}$	$\chi^2_{\rm nnlo}$	$\chi^2_{\rm nlo}$	$\chi^2_{\rm nnlo}$	$\chi^2_{\rm nlo}$	$\chi^2_{\rm nnlo}$	$\chi^2_{\rm nlo}$	$\chi^2_{\rm nnlo}$
Total	0.96	1.01	1.06	1.10	1.12	1.16	1.23	1.29
NMC d/p	0.91	0.91	0.89	0.89	0.88	0.89	0.92	0.93
NMC $\sigma^{\rm NC,p}$	-	-	-	-	-	-	1.63	1.52
SLAC	-	-	-	-	1.77	1.19	1.59	1.13
BCDMS	-	-	1.11	1.15	1.12	1.16	1.22	1.29
CHORUS	-	-	1.06	1.02	1.09	1.07	1.11	1.09
NuTeV	0.35	0.34	0.62	0.64	0.70	0.70	0.70	0.86
HERA-I	0.97	0.98	1.02	1.00	1.02	0.99	1.05	1.04
ZEUS HERA-II	-	-	-	-	1.41	1.48	1.40	1.48
H1 HERA-II	-	-	-	-	-	-	1.65	1.79
HERA $\sigma^{\rm c}_{\rm NC}$	-	-	1.21	1.32	1.20	1.31	1.27	1.28
E886 d/p	0.30	0.30	0.43	0.40	0.44	0.46	0.53	0.48
E886 p	-	-	1.18	1.40	1.27	1.53	1.19	1.55
E605	1.04	1.10	0.74	0.83	0.75	0.88	0.78	0.90
CDF Z rapidity	-	-	-	-	-	-	1.33	1.53
CDF Run-II k_t jets	-	-	1.01	2.01	1.04	1.84	0.96	1.80
D0 Z rapidity	0.56	0.61	0.62	0.71	0.60	0.69	0.57	0.61
ATLAS W,Z 2010	-	-	1.19	1.13	1.19	1.17	1.19	1.23
ATLAS 7 TeV jets 2010	0.96	1.65	1.08	1.58	1.10	1.54	1.07	1.36
ATLAS 2.76 TeV jets	1.03	0.38	1.38	0.36	1.35	0.35	1.29	0.33
ATLAS high-mass DY	-	-	-	-	-	-	2.06	1.45
ATLAS W p_T	-	-	-	-	-	-	1.13	-
CMS W electron asy	0.98	0.84	0.82	0.72	0.85	0.73	0.87	0.73
CMS W muon asy	-	-	-	-	-	-	1.81	1.72
CMS jets 2011	0.90	2.09	0.96	2.09	0.99	2.10	0.96	1.90
CMS $W+c$ total	-	-	-	-	-	-	0.96	0.84
CMS $W+c$ ratio	-	-	-	-	-	-	2.02	1.77
CMS 2D DY 2011	-	-	-	-	1.20	1.30	1.23	1.36
LHCb W rapidity	-	-	0.69	0.65	0.74	0.69	0.71	0.72
LHCb Z rapidity	-	-	1.23	1.78	1.11	1.58	1.10	1.59
$\sigma(t\bar{t})$	-	-	-	-	-	-	1.43	0.66

Table 7.3: The experimental χ^2 values at NLO and NNLO for NNPDF3.0 fits using conservative datasets with three different values of the threshold $\alpha_{\rm max}$ (see text). In each case, the χ^2 is shown for the datasets which pass the conservative cut. The values for the global fit (same as in Table 7.1) are also shown for ease of comparison.

	NNLO global fit			NNLO cons. fit $\alpha_{\max} = 1.1$		
Experiment	mean	mode	median	mean	mode	median
NMC $\sigma^{\mathrm{NC},p}$	1.27	1.26	1.27	1.50	1.45	1.48
SLAC	1.13	1.09	1.12	1.61	1.37	1.48
BCDMS	1.20	1.19	1.20	2.02	1.86	1.92
CHORUS	1.10	1.09	1.09	2.55	1.69	2.32
ZEUS HERA-II	1.25	1.24	1.25	1.38	1.33	1.36
H1 HERA-II	1.35	1.34	1.34	1.51	1.47	1.49
HERA σ^{c}_{NC}	1.14	1.11	1.13	1.13	1.09	1.12
E886 p	1.15	1.14	1.15	2.18	1.62	2.03
CDF Z rapidity	1.39	1.32	1.36	1.56	1.40	1.50
CDF Run-II k_t jets	1.15	1.12	1.14	1.25	1.18	1.22
ATLAS W, Z 2010	1.17	1.12	1.15	1.38	1.25	1.32
ATLAS high-mass DY	1.00	1.34	1.63	1.63	1.19	1.45
CMS W muon asy	1.60	1.40	1.53	2.90	2.48	2.81
CMS W+c total	1.50	1.09	1.33	1.85	1.37	1.67
CMS W+c ratio	2.00	1.39	1.69	2.12	1.58	1.94
CMS 2D DY 2011	1.28	1.27	1.28	1.29	1.28	1.29
LHCb	1.20	1.12	1.17	1.58	1.22	1.48

Table 7.4: The mean, mode and median of the $P(\alpha)$ distributions at NNLO for the experiments *excluded* from the conservative fit with $\alpha_{\max} = 1.1$, either when the prior is the global fit (same as Table 7.2) or when using as prior the conservative set itself.

the experimental χ^2 shown in Table 7.3 is significantly worse than the t_0 value used for the actual determination of $P(\alpha)$). The maximally consistent dataset, found with $\alpha_{\max} = 1.1$, includes the NMC d/p data, the NuTeV and HERA-I DIS data, the Drell-Yan data from E866 and E605, the D0 Z rapidity, the ATLAS and CMS inclusive jets and the CMS W electron asymmetry.

In Table 7.4, we furthermore compare the mean, mode and median of the $P(\alpha)$ distributions for the experiments excluded from the NNLO conservative fit with $\alpha_{\max} = 1.1$ when the global fit is used as prior (i.e. the same numbers for the corresponding entries in Table 7.2), to the same quantities computed using as a prior the conservative fit itself. All the peak values of $P(\alpha)$ deteriorate when using the conservative set as a prior, as we would expect. Clearly, this deterioration will be maximal for datasets which are internally consistent, but inconsistent with the rest, and more moderate for experiments which are affected by internal inconsistencies, so that a rescaling of uncertainties is needed in order to describe them regardless of what one takes as a prior. This is the case for instance for the NMC $\sigma^{\mathrm{NC},p}$ which are affected by internal inconsistencies as already mentioned.

The distances between the conservative sets and the baseline NNPDF3.0 NNLO global fit are show in Fig. 7.12, while the PDFs are compared directly in Fig. 7.13, where the NNLO conservative fits with $\alpha_{\max} = 1.1$ and 1.2 and the reference fit are shown. All of the conservative sets are consistent with the global fit, with PDFs that differ at most

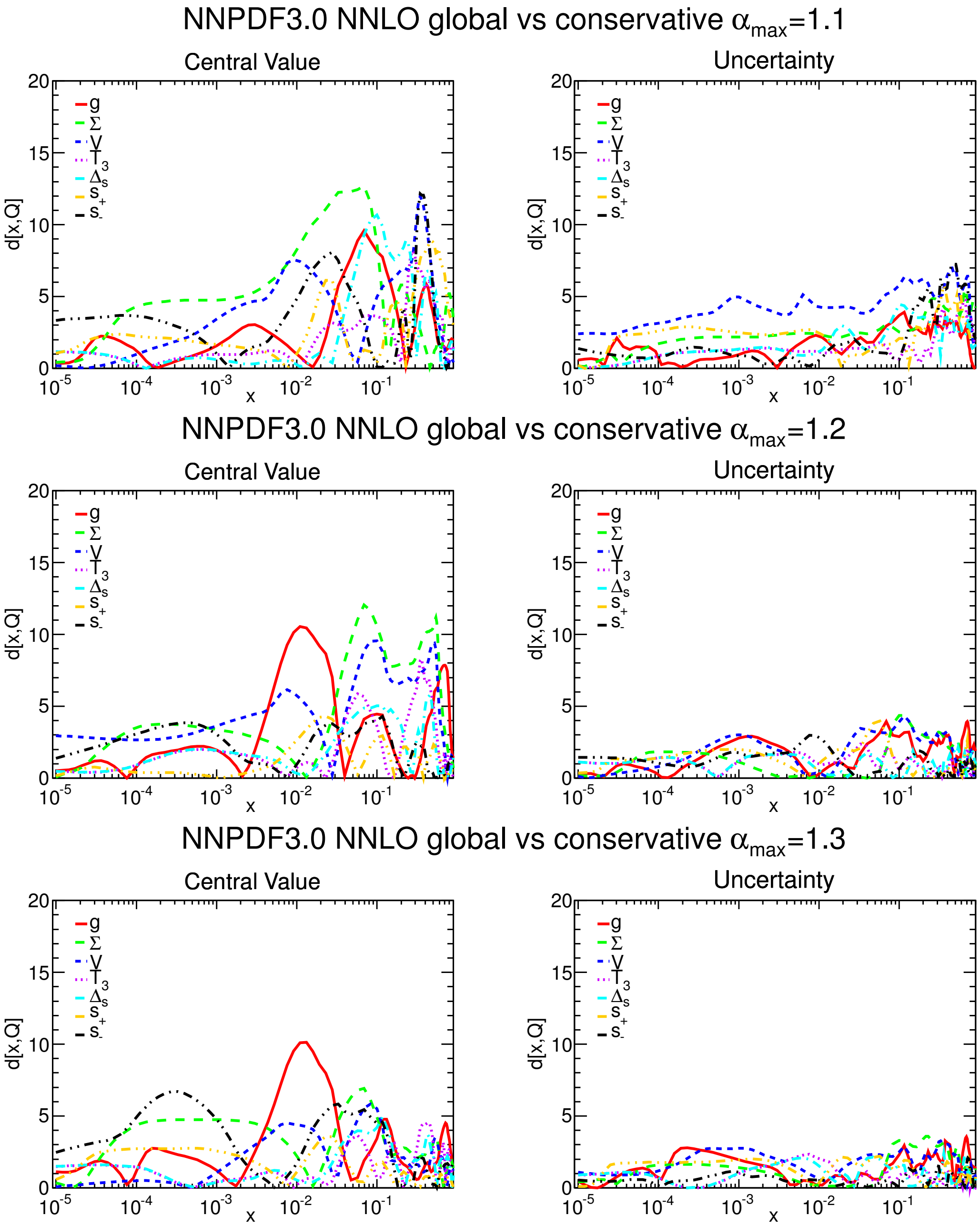


Figure 7.12: Same as Fig. 7.1, but now comparing the baseline NNPDF3.0 global fit to the conservative fits obtained using three difference values of $\alpha_{\max}$: $\alpha_{\max} = 1.1$ (top), $\alpha_{\max} = 1.2$ (center), $\alpha_{\max} = 1.3$ (bottom).

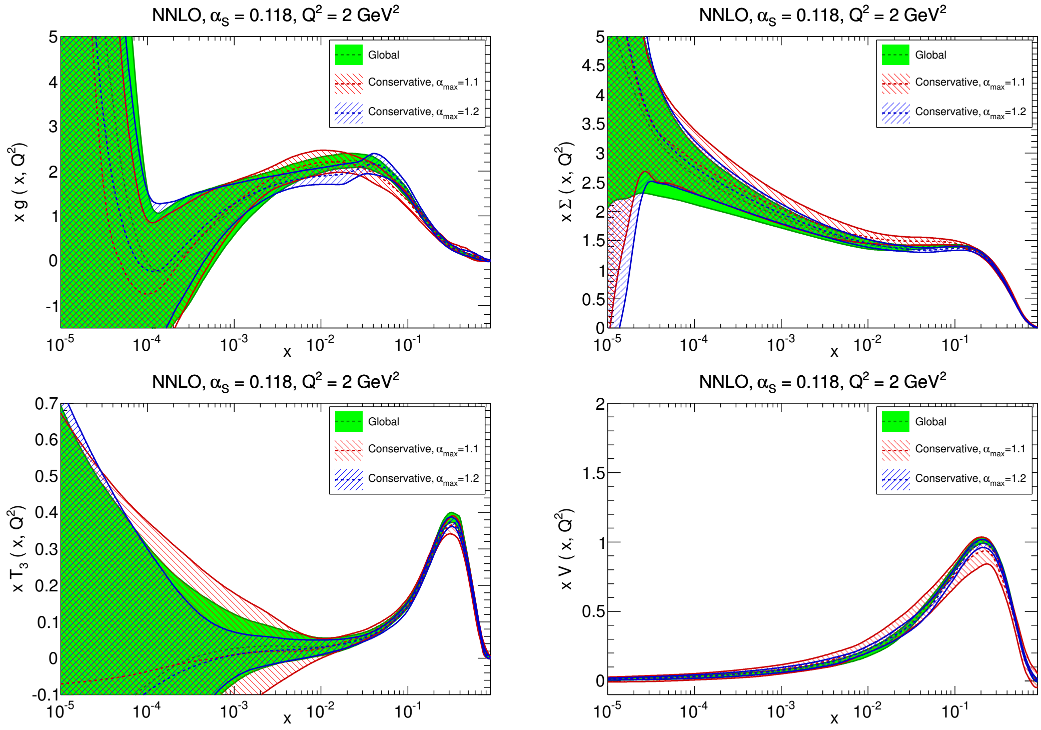


Figure 7.13: Same as Fig. 7.8, but now comparing the default global NNLO fit to the two conservative fits with $\alpha_{\max} = 1.1$ and $\alpha_{\max} = 1.2$.

at the one-sigma level, thereby confirming the consistency of the procedure, though of course PDF uncertainties are larger in the conservative fits due to their reduced datasets. At small-x, the gluon is similar in all cases as it is driven by the HERA-I data, while there is more dependence on the choice of $\alpha_{\max}$ at medium and large x. Interestingly, in the region relevant for Higgs production in gluon fusion the gluon is significantly affected by the choice of $\alpha_{\max}$, though not beyond the one-sigma level. The quarks are in good agreement, with the main differences seen at medium x.

One use for these conservative parton sets is studies aimed at assessing how individual datasets affect LHC observables by looking at their effect on a maximally self-consistent dataset, such as was performed in Ref. [194]. In the future, as more and more data will become available, this approach could also be used in determining an optimal dataset on which a global fit should be based.

7.2.2 Impact of the new HERA and LHC data

In this section I will examine in detail the impact of the new HERA and LHC data in the NNPDF3.0 fits. This will be done first both by looking at their impact in the global fit, and also in fits with substantially reduced datasets. While the former is more

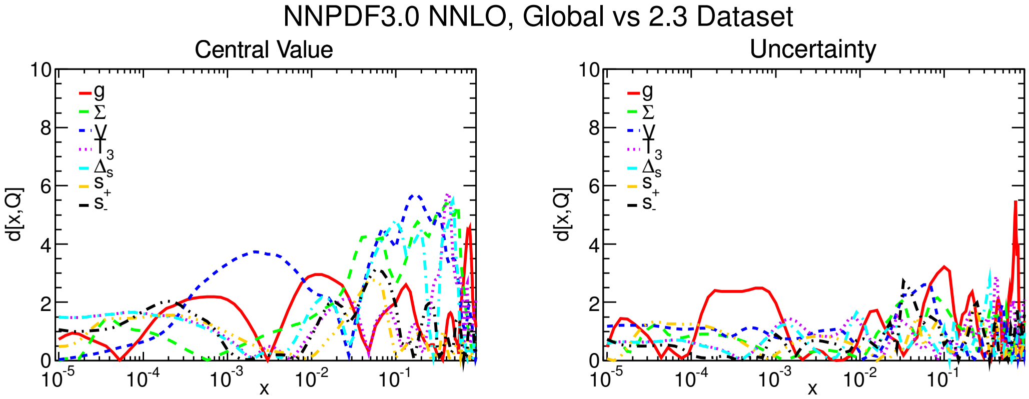


Figure 7.14: Same as Fig. 7.1, but now comparing the default NNLO set to a set obtained using the same methodology but an NNPDF2.3-like dataset.

realistic, the latter allow for an assessment of the specific impact of individual datasets (though of course will over-estimate their impact in a realistic setting). The impact of jet data will be specifically discussed in Section 7.2.3. In all these fits, precisely the same theory and methodology of the standard NNPDF3.0 fit will be used, with only the dataset changing, so that the impact of the dataset can be isolated. This will eventually allow us to provide a quantitative assessment of the dependence of our results on the dataset.

In order to obtain a first overall assessment of the impact of the new data, we have produced a variant of the NNPDF3.0 fit using the same methodology, but using a dataset very similar to that used in NNPDF2.3. We excluded all datasets which were not used in NNPDF2.3, however the resulting set is not quite identical to the NNPDF2.3 dataset, as we use slightly different cuts to the data in NNPDF3.0, and also because a small number of sets from NNPDF2.3 were not included in NNPDF3.0 (H1 and ZEUS uncombined F_2^c, CDF W asymmetry and D0 jet data).

The distances between PDFs from this fit and their NNPDF3.0 counterparts are shown in Fig. 7.14, while the PDF ratios, for $Q^2 = 10^4$ GeV^2, are compared in Fig. 7.15. Is clear that the addition of the new data affects moderately all PDFs, with central values varying by at most half a sigma in terms of the PDF uncertainties. This is unsurprising, as the NNPDF2.3 PDFs already described the new experimental data rather well, so the main impact of the new data is to reduce the uncertainties. Indeed, the PDF comparison shows that the change in uncertainties, seen in the distance plot again at a half-sigma level, generally corresponds to a reduction in uncertainty. This demonstrates the conclusion stated in Section 7.1.2, that the increase in uncertainty seen when comparing the NNPDF3.0 and 2.3 PDF sets is due to the changes to the

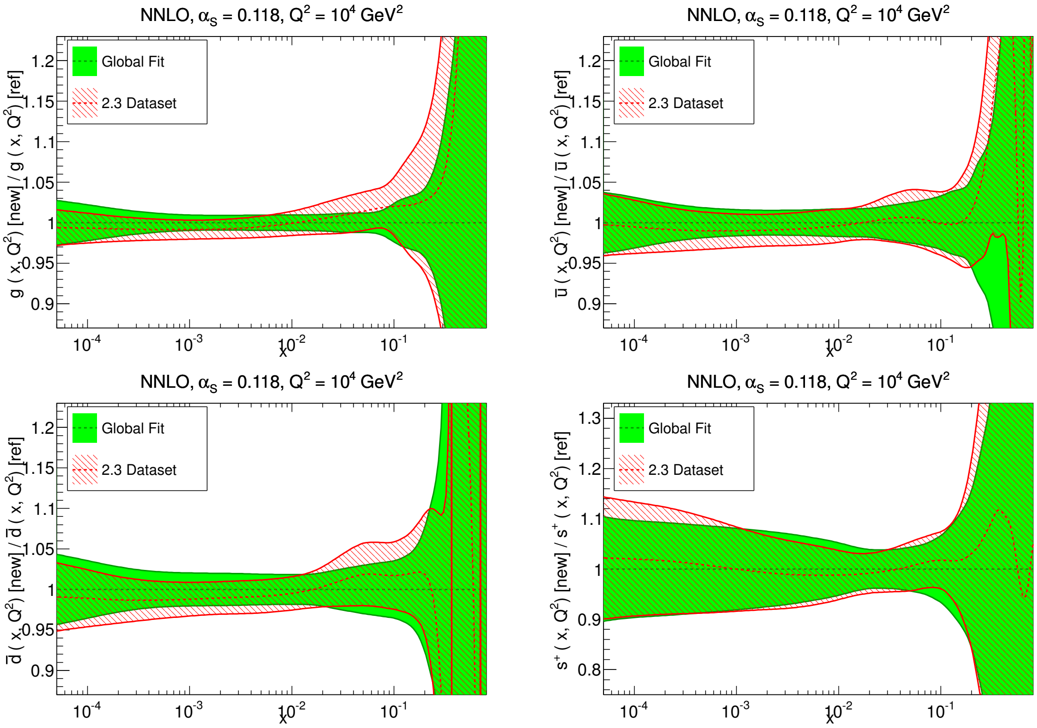


Figure 7.15: Comparison of NNPDF3.0 NNLO PDFs at $Q^2 = 10^4$ GeV^2 to PDFs obtained using an NNPDF2.3-like dataset. Results are shown as ratio to the default set. From top to bottom and from left to right the gluon, anti-up, anti-down quarks and total strangeness are shown.

methodology, as here we can see when using a consistent methodology the new data acts to reduce the overall PDF uncertainties.

Looking at the impact in more detail, the largest effect on central values is seen for the large- and medium-x quarks, followed by the gluon in the same region. The small-x gluons and quarks are quite stable since there is no new data that affects them in this region. Uncertainties mostly decrease for the gluon PDF, both at large x due to the new LHC jet and top quark data, and at medium and small x due to the new HERA-II data. The new data appear to favour a rather softer gluon at large x in comparison to the NNPDF2.3-like dataset, though the differences here are always within the PDF uncertainties. For the antiquark sea there is a visible improvement, especially at medium x, where the bulk of the new LHC electroweak vector boson production data is. Finally, there are some improvements in strangeness; the role of the LHC data in pinning down $s(x, Q)$ will be discussed in more detail in Section 7.2.4.

Focusing specifically on the new LHC data included in the NNPDF3.0 analysis, we produced a fit excluding all the LHC data from the dataset, and keeping all the

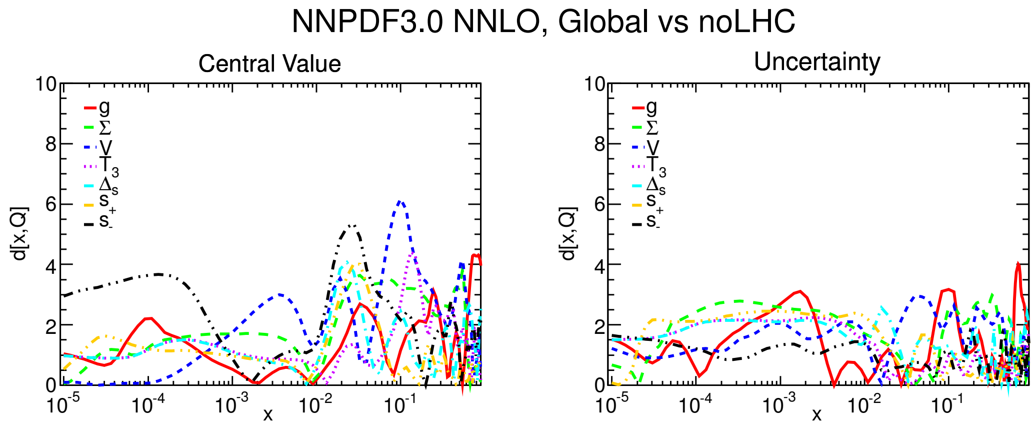


Figure 7.16: Same as Fig. 7.1, but now comparing the default NNLO set to PDFs obtained with all LHC data excluded.

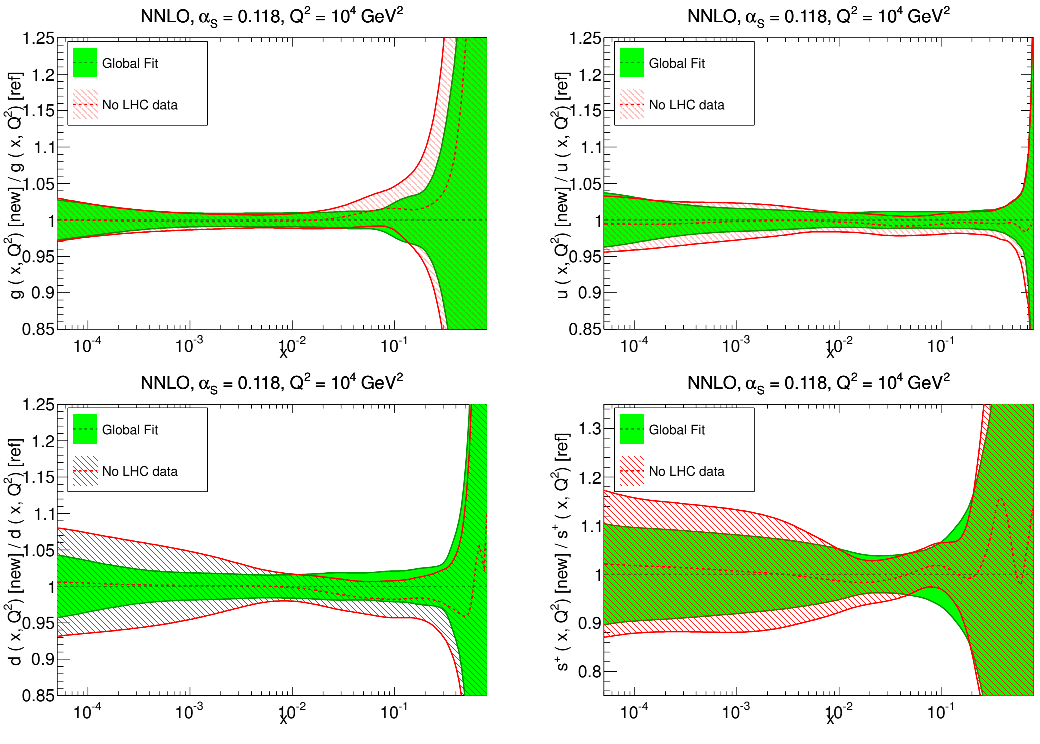


Figure 7.17: Same as Fig. 7.15 but now comparing the default NNLO set to PDFs obtained with all LHC data excluded.

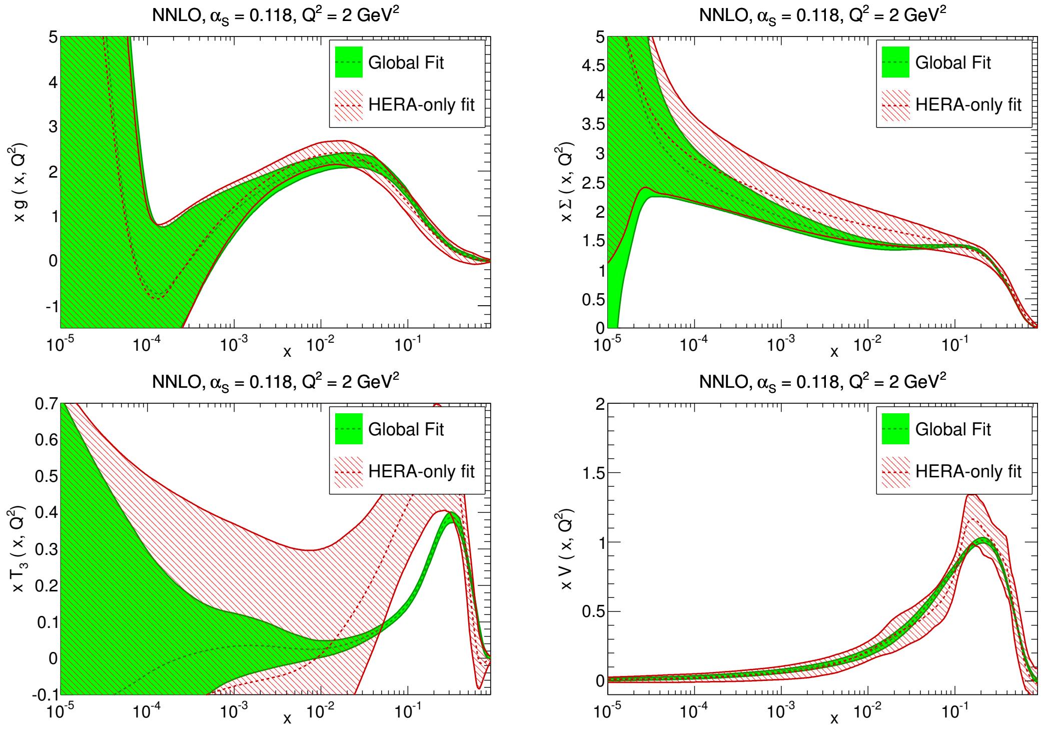


Figure 7.18: Same as Fig. 7.13 but now comparing the default NNLO set to PDFs obtained using only HERA data.

other data. By comparing this reduced dataset fit with the standard NNPDF3.0 fit including the LHC data we can evaluate the effect of their inclusion in the global dataset. Distances between this and the standard fit including LHC data are shown in Fig. 7.16, while PDF ratios are shown in Fig. 7.17. The impact of including the LHC data is seen to be moderate, at a half-sigma level, both for central values and for uncertainties, but always leads to a reduction in PDF uncertainties. The central values for the quarks at medium and large x are most affected, with a smaller effect for the gluon.

Reassuringly, PDFs without LHC data are always within the one-sigma uncertainty bands of the global fit PDFs, demonstrating the consistency of the results from fits with and without the LHC data. The gluon at medium and small x is already well constrained by HERA and Tevatron data, but the LHC improves uncertainties for $x \geq 0.02$, largely due to (as mentioned in the previous section) the ATLAS and CMS inclusive jet data and top quark production data. The down quark and strange PDFs are also affected, especially in the small-x region, but also at medium x.

The previous tests looked at the impact of adding the datasets of interest to a large global dataset. We can also perform the same test but instead using as a base

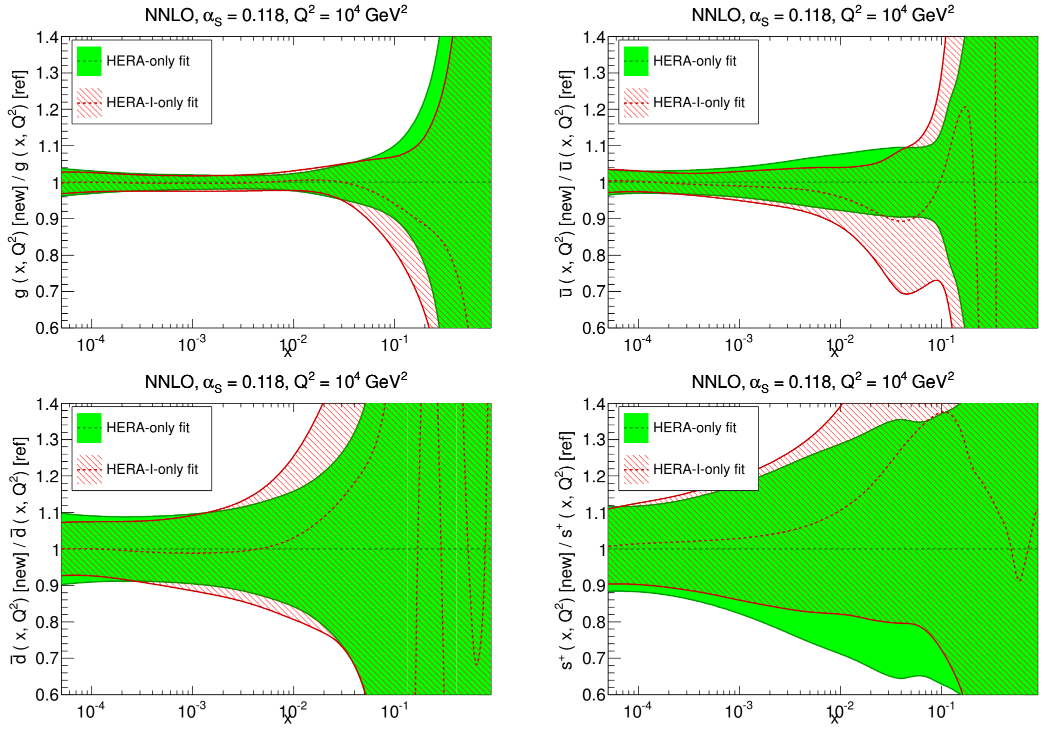


Figure 7.19: Same as Fig. 7.15 but now comparing HERA-only and HERA-I-only PDFs (see text).

a substantially restricted dataset. While the tests with the global dataset provide the best estimate of the role of the new data in the NNPDF3.0 fits, the tests I will show here have the advantage that it will be easier to assess the overall constraint of the data on PDFs.

To begin with, Fig. 7.18 shows the PDFs obtained from a fit using only the HERA data. These results will then provide a baseline to compare to fits including further data. Clearly, most of the PDFs, except perhaps the small-x gluon, have much larger uncertainties than in the global fit. Specifically, the quark flavour separation and the large-x gluon are very poorly constrained in the HERA-only fit, demonstrating that this is not competitive with a global fit for phenomenology applications.

However, the HERA dataset has widened considerably with the addition of the complete HERA-II inclusive data from H1 and ZEUS and combined HERA charm production data. In order to study the impact of this new data, we have also produced a version of the HERA-only fit in which we have kept only the combined HERA-I data, i.e. a HERA-I-only fit. The NNLO PDF ratios of the HERA-only and HERA-I-only fits are compared at $Q^2 = 10^4$ GeV^2 in Fig. 7.19. The additional information provided by HERA-II has a moderate impact: the gluon is mostly unchanged, while the PDF

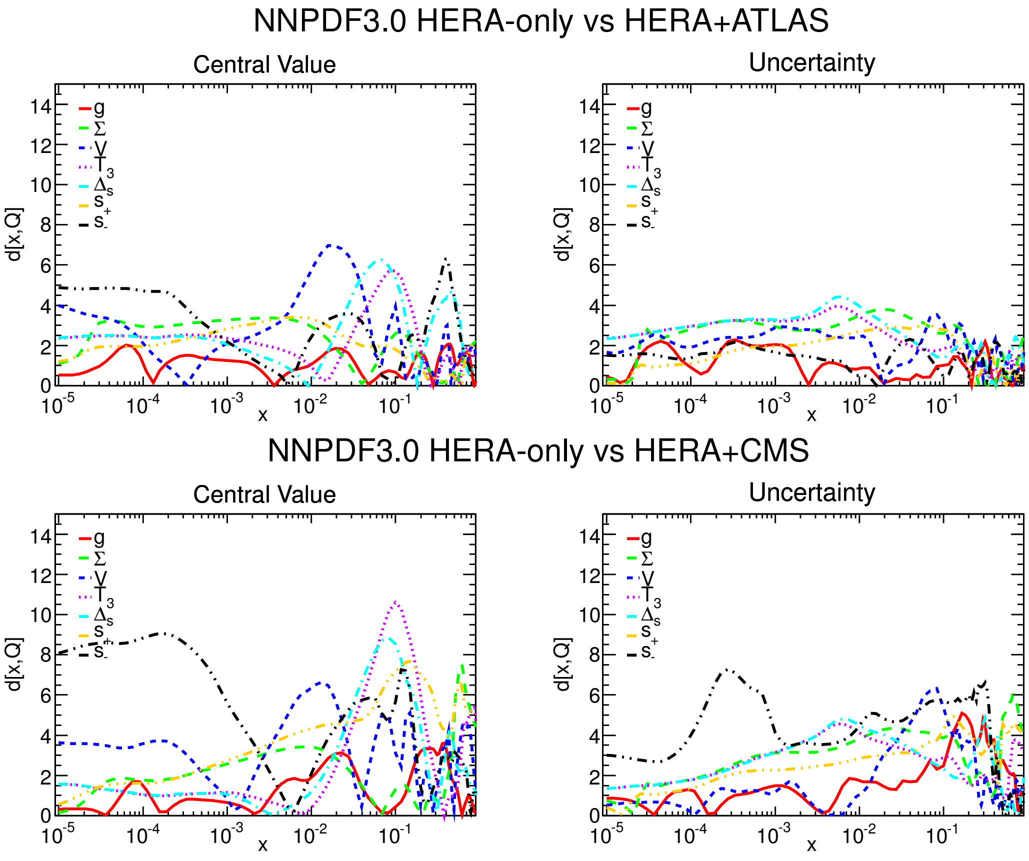


Figure 7.20: Same as Fig. 7.1, but now comparing the NNLO HERA-only fit to the HERA+ATLAS (top) and HERA+CMS (bottom) fits.

uncertainties on the medium- and large-x up antiquarks and (to a lesser extent) on the down antiquarks are moderately reduced. We conclude that, while certainly beneficial, the new HERA-II data does not change substantially the known fact that HERA-only fits obtain large PDF uncertainties.

Having looked at the HERA-only fit itself, we can now study the response of HERA-only fit to the addition of various other datasets. In particular, we have produced two fits: one which adds all of the ATLAS data included in the NNPDF3.0 global fit, and another which adds all of the CMS data. Specifically, in the HERA+CMS fit the HERA data is supplemented with data on jet production, W asymmetries, Drell-Yan differential distributions, $W+c$ production and top quark total cross-sections, while in the HERA+ATLAS fit, the HERA measurements are supplemented with W, Z rapidity distributions from the 2010 dataset, inclusive jet data at 7 TeV and 2.76 TeV, and high-mass Drell-Yan production.

The distances between the HERA-only fit and the HERA+ATLAS and HERA+CMS

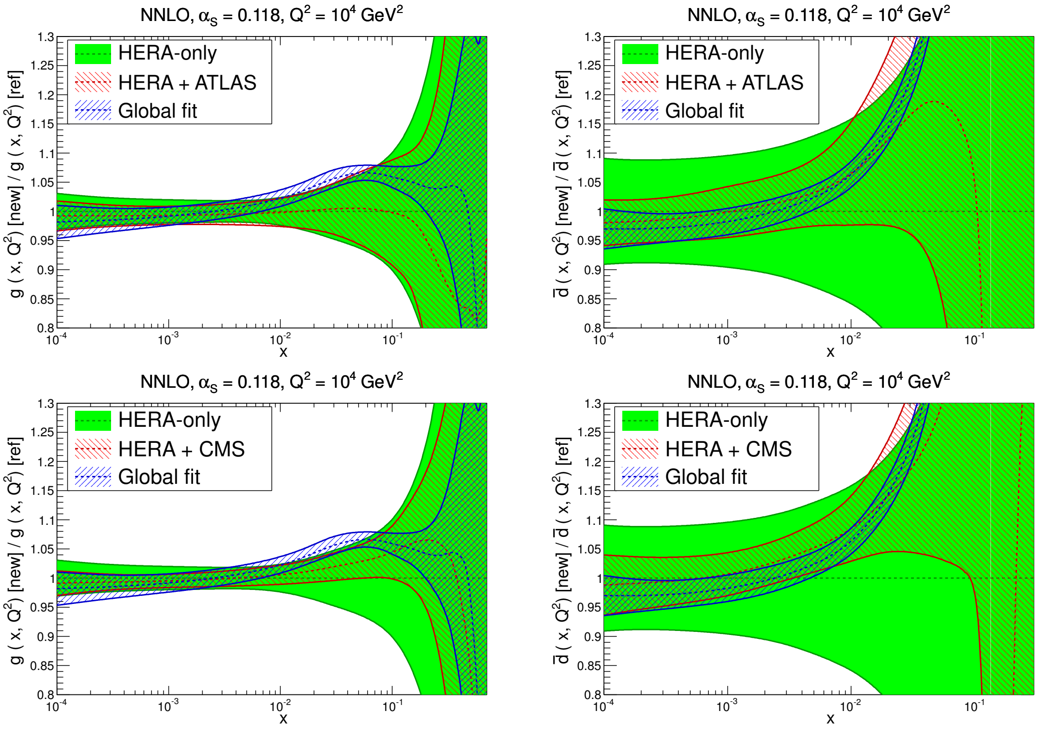


Figure 7.21: Comparison of the gluon and antidown NNLO PDF at $Q^2 = 10^4$ GeV^2 of the HERA-only and HERA+ATLAS sets (top) or the HERA-only and HERA+CMS sets (bottom), shown as rations to the HERA-only PDFs. For reference, the PDFs from the default NNPDF3.0 global set are also shown.

fits are shown in Fig. 7.20, while the gluon and $\bar{d}$ PDFs are shown in each case in Fig. 7.21, along with the global fit result. The impact of the LHC data is apparent, in particular for PDF combinations which are poorly constrained in the HERA-only fit, like the valence and triplet distributions. Note that the CMS data provides more stringent constraints on the gluon at large x since it uses the 2011 inclusive jet data, which for ATLAS is still not available. ATLAS and CMS have a similar constraining power for the medium and large-x quarks, with CMS slightly superior for the strangeness PDFs thanks to the availability of the $W+c$ measurements, and also for flavour separation (and thus for $\bar{d}$) due to the fact that the CMS electroweak dataset is somewhat more extensive. On the other hand, the comparison to the global fit shows that neither the HERA+ATLAS nor the HERA+CMS fits are currently competitive.

In Table 7.5 we show φ_{χ^2} for the global NNPDF3.0 NLO and NNLO fits, as well as for the fits based on reduced datasets described above. As described in Appendix B this provides a measure of the size of the PDF uncertainties in terms of the average experimental uncertainty, i.e. how much the overall uncertainty on each data point is

Dataset	φ_{χ^2} LO	φ_{χ^2} NLO	φ_{χ^2} NNLO
Global	0.512	0.291	0.302
HERA-I	-	0.453	0.439
HERA	-	0.375	0.343
HERA+ATLAS	-	0.391	0.318
HERA+CMS	-	0.315	0.345
Conservative	-	0.422	0.478
no LHC	-	0.312	0.316

Table 7.5: The value of the fractional uncertainty φ_{χ^2} (defined in App. B) for the default NNPDF3.0 NLO and NNLO fits compared to that obtained in various fits to reduced datasets. At LO, only the value for the global fit is available as the reduced dataset fits were not performed at this order. The result for the conservative set refers to the fit with $\alpha_{\max} = 1.1$.

reduced by fitting the combined dataset. In general we expect this to fall as more data is included in the fit, unless the data is inconsistent. The result for the conservative set refers to the fit with threshold $\alpha_{\max} = 1.1$.

For the global fits, we find $\varphi_{\chi^2} = 0.291$ and 0.302 for the NLO and NNLO sets respectively, to be compared with the corresponding value at LO, $\varphi_{\chi^2} = 0.512$. The improvement between LO and NLO, almost by a factor of two in terms of the reduction of the PDF uncertainties on the fitted data points, is clear evidence of the better consistency of the NLO fit in comparison to the LO one. On the other hand, the NNLO fit is very similar to the NLO one in this respect (perhaps marginally worse), consistent with the observation that the quality of the NNLO fit is not significantly better than that of the NLO fit, which is also reflected by the values of the χ^2, see Table 7.1. The decreasing trend seen in the values of φ_{χ^2} for the fits to reduced datasets, from HERA-I to HERA-all, to HERA+ATLAS or HERA+CMS, to the global fit, shows the expected uncertainty reduction as more data are combined.

7.2.3 Impact of jet data on the global fit

In this section I will explore the impact of jet data in the NLO and NNLO NNPDF3.0 fits, with the motivation of investigating the possible bias which could result from theoretical limitations in the description of jet data, in particular the current lack of full knowledge of NNLO corrections.

In order to study this, we produced an NNPDF3.0 PDF fit in which all jet data are removed from the global dataset, the gluon from which is compared to that from the default global fit in Fig. 7.22. It is clear that removing jet data from the global fit leads to a substantial increase of the PDF uncertainties on the gluon at medium and large x, both at NLO and NNLO. Also, note that, when jet data are included, the uncertainties are very similar at NLO and NNLO, which is reassuring as it is consistent with the

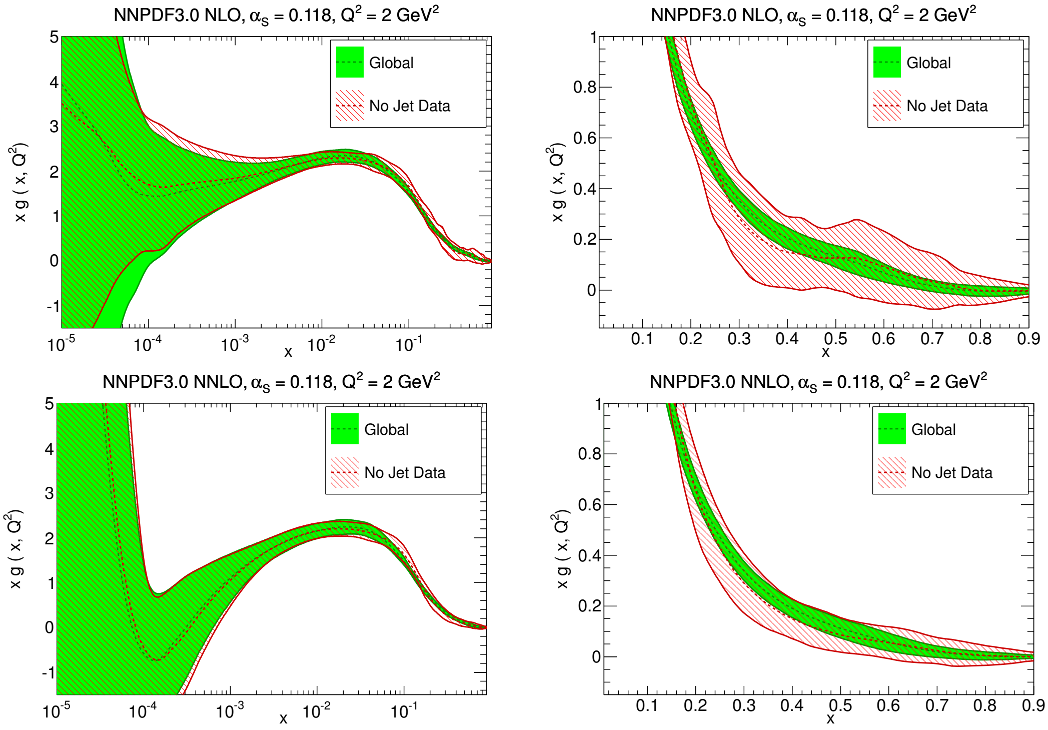


Figure 7.22: Comparison of the gluon in a fits using a datasets with and without jet data at NLO (top) and NNLO (bottom), plotted at $Q^2 = 2$ GeV^2 vs. x on a logarithmic (left) and linear (right) scale.

expectation that no instabilities are introduced by jet data in the NNLO fit despite potentially large perturbative corrections. Other PDFs are essentially unchanged upon removing jet data.

Further evidence for the lack of inconsistency in the NNLO jet data can be seen by looking at the χ^2 given in table Table 7.6. Here I compare the χ^2 to the collider jet data at NLO and NNLO, both for the reference NNPDF3.0 fit and in the fit without jet data. The description of jet data turns out to be reasonably good even when they are not included in the fit. Also shown is the value of the χ^2 for top pair production, which is sensitive to the gluon. The fact that this value changes very little upon inclusion of jet data is also evidence for general consistency.

From this we can conclude that not including jet data (or not including them at NNLO) would not lead to a significant change of the central value of the extracted gluon distribution but it would lead to a deterioration of its uncertainty. Given our conservative treatment of NNLO perturbative corrections, described in Section 4.2.1, and in the absence of indications of instability or inconsistency, we believe that the determination of the gluon is most reliable when jet data are kept in the dataset, as we

	NLO			
	Exp χ^2		t_0 χ^2	
Dataset	Global	No Jets	Global	No Jets
CDF Run II	0.95	*1.51*	1.05	*1.62*
ATLAS 7 TeV + 2.76 TeV	1.58	*1.88*	0.86	*0.96*
CMS 7 TeV 2011	0.96	*1.32*	0.90	*1.17*
Top quark pair-production	1.43	1.26	1.67	1.49

	NNLO			
	Exp χ^2		t_0 χ^2	
Dataset	Global	No Jets	Global	No Jets
CDF Run II	1.84	*1.85*	1.20	*1.58*
ATLAS 7 TeV + 2.76 TeV	1.17	*1.00*	0.72	*0.65*
CMS 7 TeV 2011	1.91	*2.23*	1.07	*1.37*
Top quark pair-production	0.73	0.43	0.61	0.42

Table 7.6: The χ^2 to jet data, computed using the default NNPDF3.0 PDFs and PDFs from a fit to a dataset without jet data. Values in italics correspond to data which have not been included in a particular fit. χ^2 calculated using both the experimental and t_0 definition are provided (see Section 7.1.1). The value for top data (included in all fits) is also shown.

do for our default fit.

7.2.4 Nucleon strangeness

Recently the size of the strange distribution has been the object of experimental and phenomenological debate. In global fits, the strange PDF is mostly constrained by the neutrino-induced deep-inelastic scattering data, such as CHORUS, NuTeV and NOMAD [78, 79, 195, 196]. While inclusive data is also sensitive to strangeness, the strongest constraint has come from the so-called dimuon process: charm production in charged-current DIS. However, the theoretical treatment of this data is affected by various sources of theoretical uncertainty, such as the need to model charm fragmentation, the treatment of charm quark mass effects at low scales, and effects related to the use of nuclear targets. Recently, LHC data which constrain the strange PDF have become available: on top of inclusive W and Z production, W production in association with charm quarks which directly probes strangeness at leading order.

In PDF global fits, with the strange PDF largely determined from neutrino data, the strange sea is typically smaller than the up and down quark sea by a factor of order $\sim \frac{1}{2}$. In 2012, a QCD analysis of the ATLAS data on W and Z rapidity distributions at 7 TeV [197] suggested that the size of the strange sea was comparable to that of the other light quarks, at least for $x \sim 0.01$. This analysis was revisited in the NNPDF2.3 framework [45], with the conclusion that while the ATLAS data in isolation

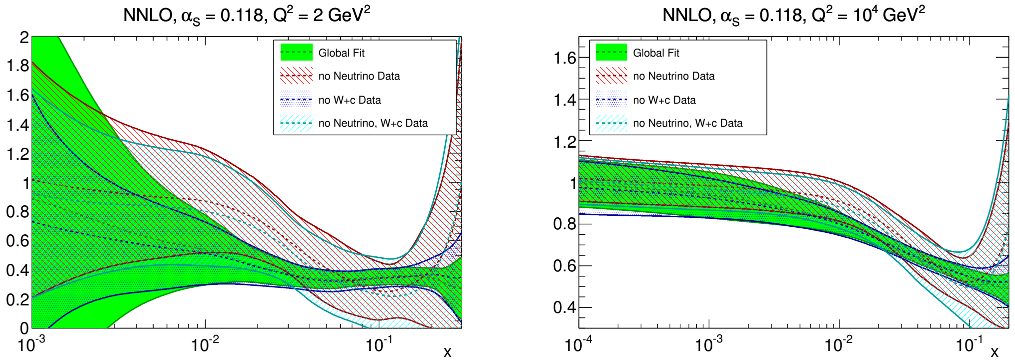


Figure 7.23: The strangeness ratio r_s (given by Eq. 7.2) in NNPDF3.0 NNLO fits to sets which alternately include and exclude the neutrino and W+c datasets included in the global NNPDF3.0 dataset. r_s is shown for $\alpha_S(M_Z) = 0.118$ and at $Q^2 = 2$ GeV^2 (left) and $Q^2 = 10^4$ GeV^2

do favour a central value of $s(x, Q^2)$ similar in size to $\bar{u}_S(x, Q^2)$ and $\bar{d}_S(x, Q^2)$, the uncertainties involved are so large that it was difficult to make a clear-cut statement, and in particular the central value of the strangeness fraction in the global NNPDF2.3 fit was compatible with that of a HERA+ATLAS fit at the one-sigma level. Also, it was found that including the ATLAS data in the global fit would have little impact on this, and strangeness would still be suppressed.

In NNPDF3.0 we have also included CMS data for W+c [103], which directly constrain the strange distribution. This dataset has recently been used in a QCD analysis [101], together with HERA data, to show that the strange PDF $s(x, Q^2)$ from collider-only data can be determined with a precision comparable to that of global fits which include neutrino data. The CMS data favours a suppressed strangeness, consistent with the indications from the neutrino data. ATLAS W+c data (which is not included in NNPDF3.0 because the data are only available at the hadron level) appear instead to favour an enhanced strangeness [120]. Fits including LHC W, Z and W+c data together with fixed target deep-inelastic scattering and Drell-Yan data have also been studied in Ref. [119], with the conclusion that a good fit to all these datasets can be obtained, and again finding suppressed strangeness.

We can study this issue in the NNPDF3.0 global PDF determination by constructing PDF sets fit to datasets which include or exclude in turn various pieces of experimental information which are sensitive to strangeness. Specifically we have produced PDF sets based on reduced datasets: a fit excluding all neutrino data (CHORUS and NuTeV; a fit excluding all CMS W+c data (but including the neutrino data); and a fit in which both the neutrino and W+c data are excluded.

Of particular interest in the comparison of these sets is the strangeness fraction r_s,

	$\chi^2_{\rm exp}$			
	Global	No neutrino	No W+c	No neutrino/W+c
CHORUS	1.13	*3.87*	1.09	*3.45*
NuTeV	0.62	*4.31*	0.66	*6.45*
ATLAS W, Z 2010	1.21	1.05	1.24	1.08
CMS W+c 2011	0.86	0.50	*0.90*	*0.61*

Table 7.7: Values of the χ^2 (using the experimental definition) to different datasets sensitive to strangeness, using as input PDFs obtained from fits in which these data are included or excluded in turn. Values in italics denotes cases in which the particular data was not included in the particular fit.

defined as

$$r_s = \frac{s+\bar{s}}{\bar{u}+\bar{d}}\,. \tag{7.2}$$

In Fig. 7.23 r_s is shown for the default NNPDF3.0 fit and the three fits described above, all plotted as a function of x, and for both $Q^2 = 2$ GeV^2 and $Q^2 = 10^4$ GeV^2.

First, we observe the remarkable compatibility of the various fits (with, as usual, smaller uncertainty at a higher scale due to asymptotic freedom), and for all fits and all x the one-sigma PDF uncertainty bands overlap. The global fit in general has the smallest uncertainties, though at very low $x <\sim 10^{-3}$ the global fit uncertainty is the largest, likely just due to statistical fluctuations. While removing neutrino data results in a dramatic increase in the uncertainty, removing the W+c data has very little impact, with only a moderate uncertainty reduction for $x \geq 0.05$ when it is included with the neutrino data. The fits without neutrino data also have a higher central value of r_s in the region of $x \sim 0.01$, though with the larger uncertainties in these fits this is only a one-sigma deviation, consistent with statistical fluctuations.

The χ^2 for the relevant experiments in these various fits are collected in Table 7.7, allowing us to compare how well each experiment is described by fits with include or exclude it. We see that W+c data are well described regardless of whether they are included in the fit or not, while the neutrino data are very poorly described if they are excluded from the fit. This agrees with the results in Fig. 7.23, and reinforces the conclusion that the impact of the W+c is moderate.

From these results we conclude that the W+c data alone are not yet competitive with the neutrino data for determining strangeness, and that their inclusion does not significantly modify the assessment of the size of the nucleon strangeness in previous global fits, a suppression of strangeness at low scales by a factor of between two and three compared to the light quarks.

7.3 Stability

In this section I will study the dependence and stability of our results upon our variations of our methodology. Some of these issues have previously been investigated in Chapter 5 in the context of closure tests, but here I will look at their impact in fits to the real experimental data; others, for instance the impact of extended positivity, can only be studied using the real data.

Specifically, I will look at the impact on the NNPDF3.0 results of the new minimization and stopping methodology discussed in Sections 5.3 and 5.5 in comparison to that previously used in NNPDF2.3; the impact of the improved treatment of positivity discussed in Section 5.6; the differences in one-sigma and 68% confidence interval definitions of PDF uncertainty; the impact of a multiplicative vs. additive treatment of systematic uncertainties (see Section 4.3.2); and finally reassess the independence of results on the choice of fitting basis.

7.3.1 Impact of the NNPDF3.0 methodology

In Section 7.2.2 I introduced an NNPDF3.0 fit performed using only the data included in the NNPDF2.3 analysis. There I used it to look at the impact of the new data by comparing it to the global NNPDF3.0 fit, but it can also be used to quantify the impact of the new NNPDF3.0 methodology and theory settings by comparing it instead to the NNPDF2.3 global fit. With this comparison we can fully disentangle the effects of the new experimental data in NNPDF3.0 from that of the improved fitting methodology and the new theoretical settings.

The distances between the original NNPDF2.3 PDFs of Ref. [45] and the NNPDF3.0 fit with NNPDF2.3 data are shown in Fig. 7.24 both at NLO and NNLO, while the NNLO PDFs are compared in Fig. 7.25. In the NLO fit the new methodology and theory settings have an impact on the small-x gluon and large-x quarks at the one and a half–sigma level. The differences in the gluon can be understood as a consequence of having switched from the FONLL-A heavy quark scheme used in NNPDF2.3 to the more accurate FONLL-B adopted in NNPDF3.0, while the differences seen for quarks are necessarily a consequence of the more efficient methodology and extended positivity constraints (see Section 7.3.2 below). At NNLO the non-insignificant differences seen in all PDFs reflect the improved methodology and positivity, there were no significant changes in the NNLO theory between 2.3 and 3.0. At high scale the most noticeable difference is the softening of the small-x gluon seen in Fig. 7.25.

Another way to compare NNPDF2.3 and 3.0 is in terms of quality of fit to their common datasets. Table 7.8 shows the χ^2s for the global NNPDF2.3 NNLO fit (taken from the original paper [45]), and for the NNPDF3.0 NNLO fits to the full dataset and

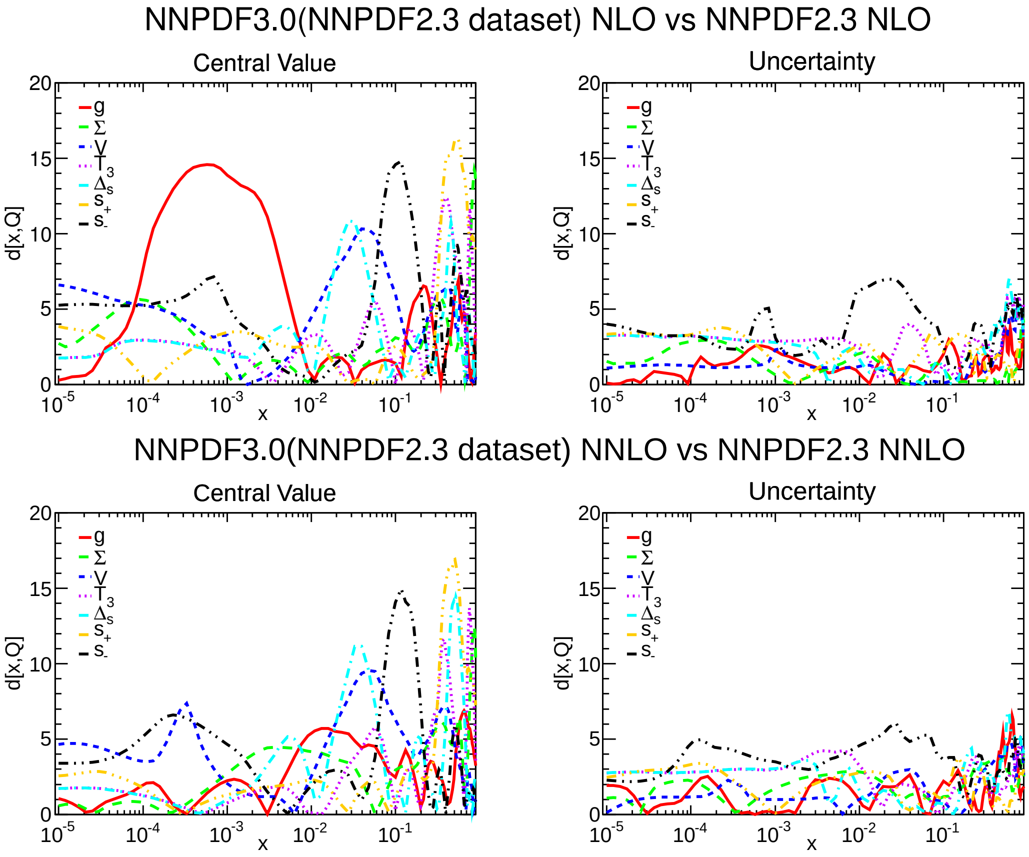


Figure 7.24: Same as Fig. 7.14, but now comparing the PDFs obtained from an NNPDF2.3-like dataset with NNPDF3.0 methodology and theory to the published NNPDF2.3 sets at NLO(top) and NNLO (bottom) [45].

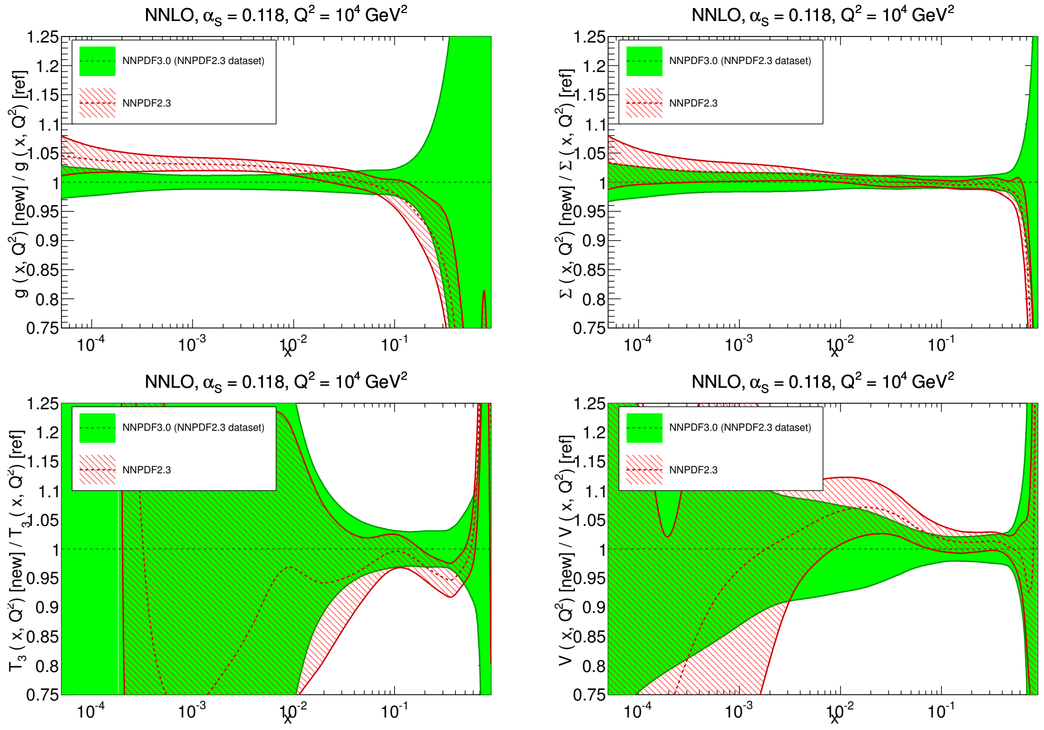


Figure 7.25: Same as Fig. 7.15, but now comparing the PDFs obtained from an NNPDF2.3-like dataset with NNLO NNPDF3.0 methodology and theory to the published NNPDF2.3 NNLO set.

to the reduced NNPDF2.3-like dataset. The changes in the treatment of the theory and uncertainties, described in Chapter 4, mean that for many datasets the NNPDF2.3 and 3.0 χ^2s are not directly comparable. For this reason, the χ^2 in the fit to the NNPDF2.3-like dataset are for some datasets very different to the original NNPDF2.3 values. This is particularly true for hadronic data, especially the jet data, where we now treat the majority of systematic uncertainties as multiplicative. Comparing the χ^2s for the two NNPDF3.0 fits, the values in the global fit are in general slightly worse, suggesting either that the global fit to these data is sub-optimal, or possibly that there is some tension between these data and the new datasets.

The main conclusion of this comparison is that a significant part of the difference between the final NNPDF2.3 and NNPDF3.0 sets, as seen specifically at high scale in Fig. 7.4 and at low scale in Fig. 7.3, are due to the improved methodology (minimisation and generalised positivity), or possibly due to changes in the treatment of data. This is consistent with the conclusion of Section 7.2.2 (see in particular Figs. 7.14-7.15), that the new data added in NNPDF3.0 generally have only a moderate impact.

	NNPDF2.3	NNPDF3.0 NNLO	
	NNLO	NN23 data	Global fit
Dataset	χ^2_{exp}	χ^2_{exp}	χ^2_{exp}
NMC d/p	0.95	0.9	0.93
NMC	1.59	1.53	1.52
SLAC	1.00	1.21	1.13
BCDMS	1.28	1.26	1.29
CHORUS	1.07	1.07	1.09
NuTeV	0.56	0.75	0.86
HERA-I	1.01	0.99	1.04
E886	1.04	0.89	0.9
E605	1.58	1.42	1.47
CDF Z rapidity	2.03	1.5	1.53
CDF Run-II k_t jets	0.68	1.82	1.8
D0 Z rapidity	0.61	0.61	0.61
ATLAS W, Z 2010	1.43	1.15	1.23
ATLAS 7 TeV jets 2010	0.94	1.47	1.36
CMS W electron asy	0.81	0.72	0.73
LHCb W rapidity	0.83	0.7	0.72

Table 7.8: The values of the experimental χ^2 per data point for the NNPDF2.3 NNLO central fit and the NNPDF3.0 NNLO fits to the global dataset and an NNPDF2.3-like dataset. Due to changes in the treatment of the data, the NN2.3 and NN3.0 values are not directly comparable.

7.3.2 Constraints from positivity

As explained in Section 5.6, in NNPDF3.0 we adopt a more extensive set of positivity constraints, in order to ensure not only positivity of the observables used in PDF fitting, but also of potential new observables such as cross-sections for new physics processes used in searches. In order to quantify the impact of these positivity constraint, we have produced a variant of the NNPDF3.0 NNLO in which positivity constraints are removed. The distances between the default fit and the fit without positivity are shown in Fig. 7.26, while some of the PDFs where the effect is largest are compared in Fig. 7.27.

The impact of positivity is relatively mild apart from for the small-x gluon and the large-x strangeness, for which there is little direct experimental information. For all other PDFs and x ranges the impact of positivity is below the one-sigma level. Note that even so, strict positivity is often necessary if one wishes to obtain meaningful predictions, e.g. in new physics searches. For the gluon, the effects of the positivity can be noticed already at $x <\sim 10^{-3}$, while at even smaller x the gluon would become much more negative if positivity were not imposed. For the strangeness asymmetry, interestingly, the dip-bump structure seen in the global fit is seen to be a consequence of positivity.

As a test of the efficiency of the Lagrange multiplier method we use to impose positivity, we have explicitly checked a posteriori that physical cross-sections at NLO

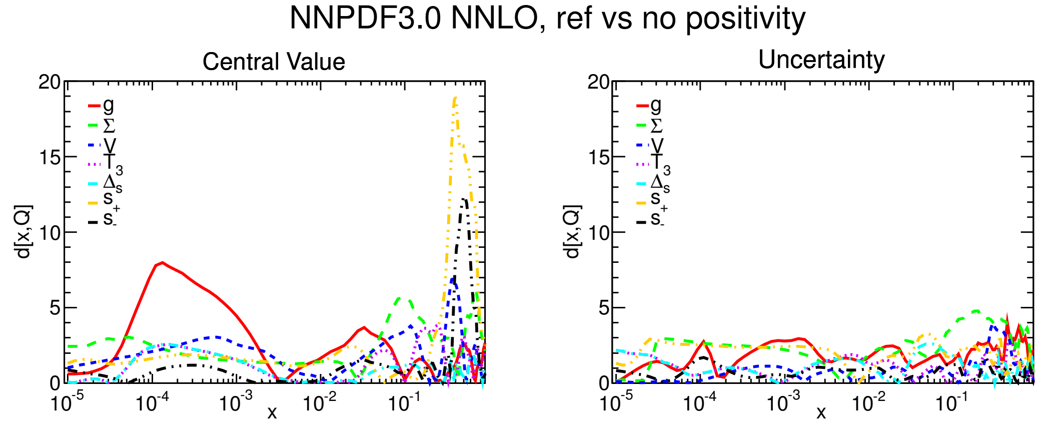


Figure 7.26: Same as Fig. 7.14, but now comparing fits with and without generalised positivity constraints.

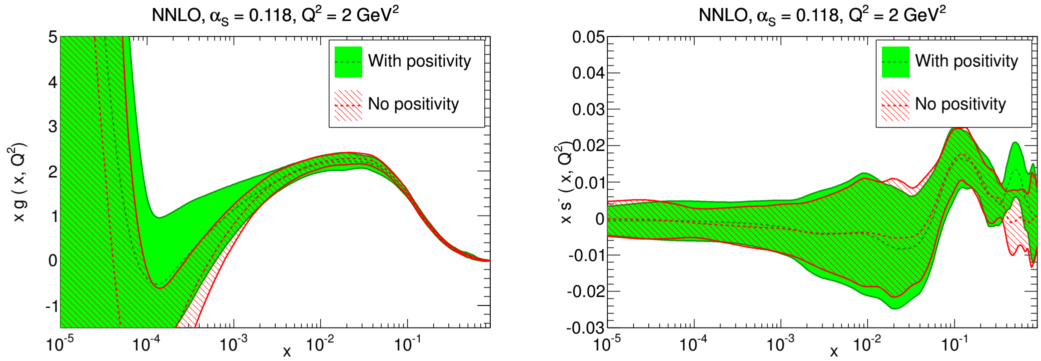


Figure 7.27: Comparison of the default NNPDF3.0 NNLO PDFs at $Q^2 = 2$ GeV^2 with $\alpha_s(M_Z) = 0.118$ to their counterpart obtained without imposing positivity. The gluon (left) and strangeness asymmetry (right) are shown.

and NNLO are indeed non-negative. This is illustrated in Fig. 7.28, where we plot two of the pseudo-observables used in the fits, namely the light component of F_L, and the $s\bar{s}$ Drell-Yan rapidity distribution. Individual replicas are shown in green dashed curves compared to the central values for the reference set used in the positivity implementation (see Section 5.6). The effectiveness of positivity is clearly seen, especially for the Drell-Yan distribution.

7.3.3 Definition of PDF uncertainties

As we generate a set of Monte Carlo replicas in our determination, it is possible to define several different measures of the PDF uncertainties. One can, as normal, calculate them using the standard deviation of the sample PDFs, however this will ignore any non-

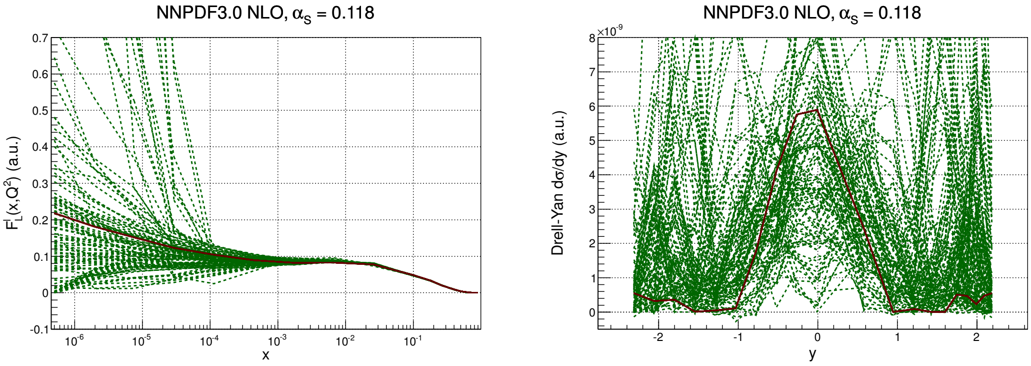


Figure 7.28: The light quark contribution to F_L (left), and the $s\bar{s}$ Drell-Yan rapidity distribution (right) plotted in arbitrary units at $Q^2 = 5$ GeV^2 for individual replicas in the NNPDF3.0 NLO set (dashed green lines). The reference set used in the positivity implementation (see Section 5.6) is also shown (red line).

Gaussianity, due to for instance positivity, in the distribution of replicas. A different approach is to take the central 68% of the replicas at every point (i.e. by dropping the top and bottom 16%), which provides by definition a 68% confidence interval. This gives a better description of the distribution in places where it deviates from normal, but is more dependant on the particular sample.

Fig. 7.29 compares these measures for the global NNPDF3.0 NLO fit with $N_{\rm rep} = 1000$ replicas. It is clear that apart from at small x (below about 10^{-3}), the standard deviation agrees with the central 68% range, suggesting that the distribution of replicas is largely Gaussian here. On the other hand, as we move to smaller x, there are significant difference in many of the PDFs, with the one-sigma contour being substantially larger. There are a few reasons for this. This difference in the extrapolation region is largely caused by individual replicas which become very large (positive or negative) in this region, as there is no direct experimental constraints on them here.

7.3.4 Additive versus multiplicative systematics

As discussed in Section 4.3.2, there is a certain ambiguity in the treatment of correlated systematics, in particular whether each one should be treated additively or multiplicatively. In order to test the impact of the additive vs. multiplicative treatment of systematics, we have produced two modified version of the NNPDF3.0 fit, which only differ in the treatment of the systematics. In the first one we treat all systematics (except for normalisation uncertainties) as additive, and in the second the setting for each systematic (again except for normalisation) is randomised, treated

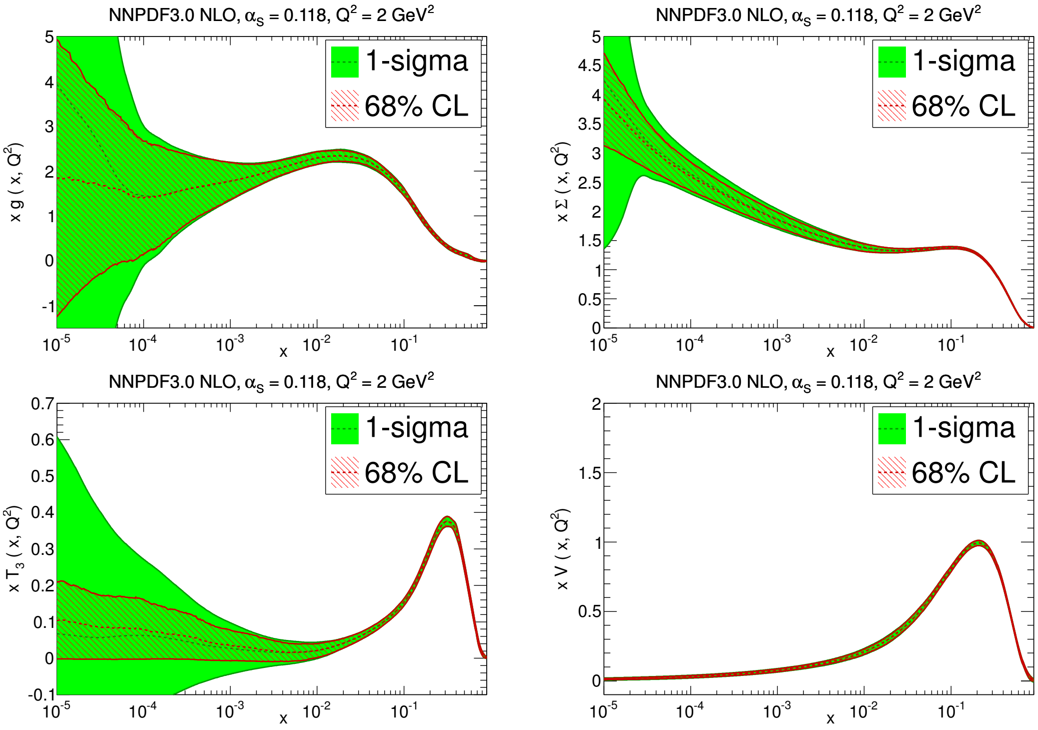


Figure 7.29: Comparison of one-sigma uncertainty bands and central 68% range for the NNPDF3.0 NLO set with $\alpha_s(M_z) = 0.119$ at $Q^2 = 2$ GeV^2. The set with $N_{\text{rep}} = 1000$ replicas has been used. From top to bottom and from left to right the gluon, singlet, isospin triplet and total valence are shown.

as either additive or multiplicative at random for each replica. The default treatment (multiplicative or additive) of systematics is given in Table 4.1 (fourth column).

The distances between these two fits and the default are shown in Fig. 7.30, while the gluon and singlet, for which the effect of the change is most noticeable, are shown in Fig. 7.31. The impact of the treatment of systematics turns out to be essentially indistinguishable from statistical fluctuations for all PDFs except the large-x gluon, for which it has an effect of at most one-sigma. This can be understood as a consequence of the fact that the gluon depends strongly on jet data, which are affected by large, multiplicative systematics. The location of the largest difference, above $x = 0.1$ also supports the conclusion that it is largely due to the jet data. The impact on the gluon is explicitly shown in Fig. 7.31. The singlet is also shown: in this case, the change in uncertainty at small x is compatible with a statistical fluctuation. When systematics are randomised the effect is diluted and the changes are always compatible with statistical fluctuations.

We conclude that the treatment of systematics, while an issue in principle, in

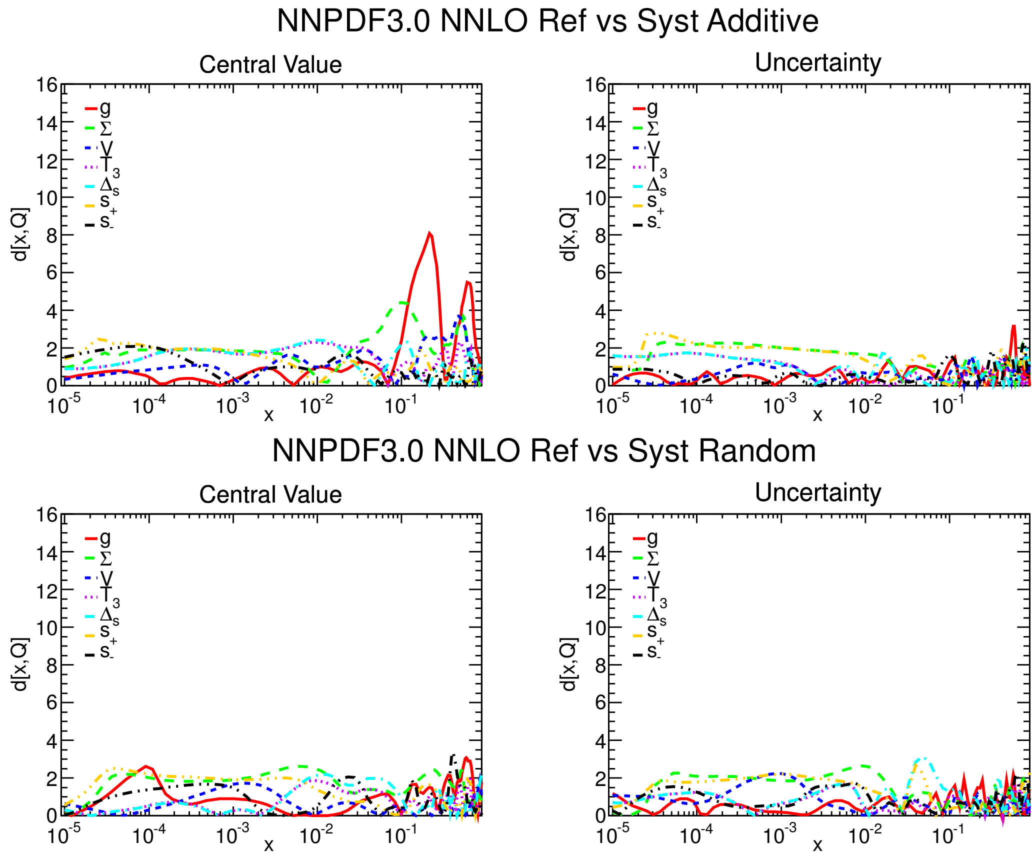


Figure 7.30: Same as Fig. 7.14, but now comparing the default set to its counterpart in which all systematics (except normalisation) are treated as additive (top) or in which the treatment of each systematic (again except normalisation) is randomised (bottom).

practice has an effect which is of comparable size as statistical fluctuations. Even when all systematics are treated as additive, which is certainly an extreme case, only the gluon changes significantly, where the effect is largely caused by the jet data for which there is less ambiguity in the treatment of systematics [198, 199]. The default treatment of systematics in the NNPDF3.0 fit thus appears to be both reliable and robust.

7.3.5 Independence of the PDF fitting basis

In Section 5.4.3 I looked at the impact of changing the PDF fitting basis has on the fit results. Fig. 7.32 shows the distances for a similar test now using the real experimental data. Distances are shown between the default NNPDF3.0 NNLO PDFs and the same fit but using the NNPDF2.3 fitting basis instead of the default NNPDF3.0 basis (see Section 5.4.3 for details) for the parametrisation of input PDFs. Results are consistent

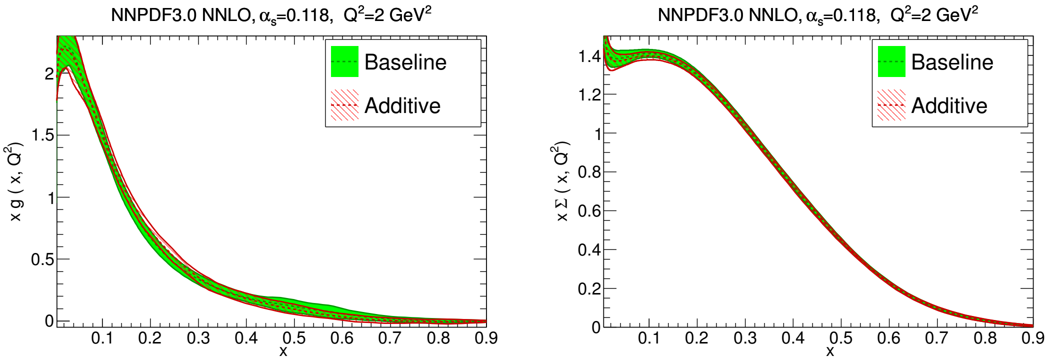


Figure 7.31: Comparison of the NNLO gluon (left) and singlet (right) for a fit in which all systematic uncertainties except normalisation are treated as additive, and for the baseline NNPDF3.0 fit, where systematic uncertainties are treated as specified in Table 4.1. The PDFs are plotted at $Q^2 = 2$ GeV^2 with a linear scale in order to highlight the large-x region.

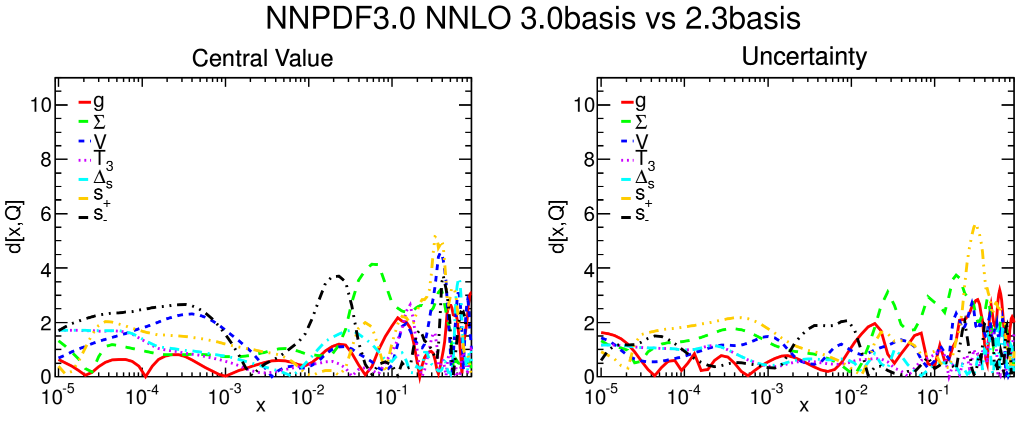


Figure 7.32: Same as Fig. 5.9, comparing fits using the NNPDF2.3 and NNPDF3.0 PDF fitting bases, but now in fits to real data.

to what was found in the closure test, with distances which are mostly compatible with statistical equivalence, and only strangeness at the valence peak deviating at the half-sigma level (slightly above the threshold of statistical indistinguishability). Note that the dip-bump structure in s^- seen in Fig. 7.27 (and related to positivity) is perfectly reproduced in NNPDF2.3 basis fit, where s^- is parameterised directly.

7.4 Implications for LHC phenomenology

In the final section of this chapter, I will provide a brief investigation of the impact of the changes in NNPDF3.0 on LHC phenomenology. I will start by comparing the parton luminosities shown in Section 7.1.2 to the same quantities calculated with the

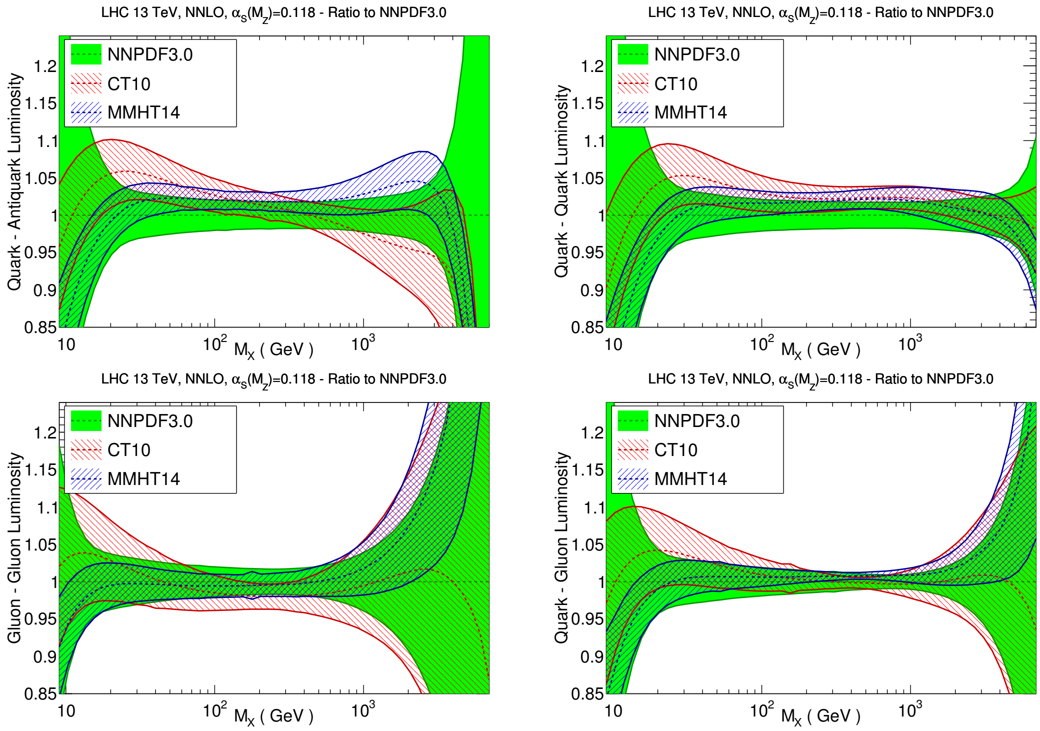


Figure 7.33: Same as Fig. 7.6 but now comparing NNPDF3.0, MMHT14 and CT10 NNLO (all with $\alpha_s(M_Z) = 0.118$). Results are shown as ratios to the NNPDF3.0 values.

CT10 and MMHT14 PDFs. I will then present predictions for a variety of LHC cross-sections at 13 TeV, specifically vector boson, top production and Higgs production, and compare results obtained using NNPDF3.0 PDFs to those of the previous NNPDF2.3 set. I will also spend some time discussing the implications of NNPDF3.0 PDFs for the dominant Higgs production channel at the LHC, gluon-fusion, including a study on the dependence of results on the datasets used in the PDF determination. Finally, I will look at the production of high-mass states, close to the LHC kinematic threshold, relevant for searches for massive New Physics at the energy frontier.

7.4.1 PDF luminosities

Fig. 7.33 compares the PDF luminosities obtained using the NNPDF3.0 set, previously discussed in Section 7.1.2 to the luminosities from the CT10 and MMHT14 sets. The three sets agree very well within their uncertainties, especially for the gg and gq cases. For the gg luminosity in the region relevant for Higgs production, the agreement between the three sets has substantially improved in comparison to the previous benchmarks using NNPDF2.3 [33]. Note that this comparison does not use the imminent CT14 PDFs; preliminary results with this set indicate that agreement will further improve with this new set.

7.4.2 Implications for $\sqrt{s}$=13 TeV LHC processes

In this section I will look at calculations of several LHC processes using the NNPDF3.0 PDFs. Unless otherwise stated, the results use the NLO sets with $\alpha_s = 0.118$, in the $N_f = 5$ variable-flavour-number scheme with massless bottom quarks, and computed using the MadGraph5_aMC@NLO program [147], version 2.1.2, interfaced to LHAPDF6. The NLO results are sufficient to assess the PDF dependence of these observables, as typically the NNLO/NLO K–factors have only a weak dependence on the PDFs. In addition to results with the NNPDF3.0 global fit, I also look at predictions with the conservative parton set with $\alpha_{\max} = 1.1$ as an illustration of results found using a maximally consistent dataset (see [200]).

The cross-sections presented here have been computed at the fiducial level, including resonance decays for several processes, and using realistic generation cuts. Jets are reconstructed with the anti-k_T algorithm [201] with radius $R = 0.5$, and the following cuts are applied to all jets in the final state:

$$|\eta_{\rm jet}| \leq 4.5\,, \quad p_{T,{\rm jet}} \geq 25 \text{ GeV}\,. \tag{7.3}$$

For final-state leptons, the following cuts are applied:

$$|\eta_l| \leq 2.5\,, \quad p_{T,l} \geq 25 \text{ GeV}\,, \quad m_{l^+l^-} \geq 30 \text{ GeV}\,. \tag{7.4}$$

Finally, for photons we impose

$$|\eta_\gamma| \leq 2.5\,, \quad p_{T,\gamma} \geq 25 \text{ GeV}\,, \tag{7.5}$$

and use the Frixione isolation criterion [202] with $\epsilon_\gamma = 1.0$ and $n = 1$ and an isolation cone radius $R_0 = 0.4$. No further analysis cuts are applied. Renormalisation and factorisation scales are set dynamically on an event by event basis to $\mu_f = \mu_r = H_T/2$, where H_T is the scalar sum of the transverse energies of all the final-state particles. Within each run, PDF and scale uncertainties in MadGraph5_aMC@NLO are obtained at no extra cost using the reweighting technique introduced in Ref. [203].

Results for NNPDF2.3, and NNPDF3.0 global and conservative sets are collected in Table 7.9, where the processes are grouped into three subsets: processes which are sensitive to quark and antiquarks, processes which are sensitive to the gluon PDF, and Higgs production processes. Gluon fusion is not included as it is discussed in the

Process	NNPDF2.3	NNPDF3.0	RelDiff	$\alpha_{\rm max} = 1.1$
$pp \to e^+e^-$	1.403 nb ($\pm$1.5%)	1.404 nb ($\pm$2.0%)	+0.1%	1.450 nb ($\pm$2.0%)
$pp \to e^+\nu_e$	10.30 nb ($\pm$1.3%)	10.21 nb ($\pm$1.9%)	-0.9%	10.29 nb ($\pm$2.3%)
$pp \to e^-\bar{\nu_e}$	7.67 nb ($\pm$1.3%)	7.75 nb ($\pm$1.9%)	+1.1%	7.96 nb ($\pm$1.9%)
$pp \to W^-\bar{c}$	2.665 nb ($\pm$3.5%)	2.680 nb ($\pm$4.2%)	+0.56%	2.807 nb ($\pm$8.8%)
$pp \to t\bar{t}$	678 pb ($\pm$1.7%)	672 pb ($\pm$1.6%)	-0.9%	655 pb ($\pm$3.3%)
$pp \to \gamma + \text{jet}$	62.24 nb ($\pm$1.2%)	63.85 nb ($\pm$1.8%)	+2.6%	61.51 nb ($\pm$1.9%)
$pp \to e^+v_e + \text{jet}$	2.353 nb ($\pm$1.2%)	2.332 nb ($\pm$1.5%)	-0.9%	2.325 nb ($\pm$1.6%)
$pp \to He^+\nu_e$	0.134 pb ($\pm$1.6%)	0.131 pb ($\pm$1.6%)	-2.2%	0.137 pb ($\pm$2.6%)
$pp \to He^+e^-$	26.48 fb ($\pm$1.4%)	26.58 fb ($\pm$1.5%)	+0.4%	27.07 fb ($\pm$2.3%)
$pp \to Ht\bar{t}$	0.458 pb ($\pm$2.2%)	0.460 pb ($\pm$1.7%)	+0.6%	0.459 pb ($\pm$4.0%)

Table 7.9: Cross-sections for LHC at 13 TeV, computed at NLO using MadGraph5_aMC@NLO with the NNPDF2.3 and NNPDF3.0 NLO PDFs, and with $N_f = 5$ and $\alpha_s(M_Z) = 0.118$. In each case, central values and the one-sigma PDF uncertainty (in parenthesis) are given. We also show the percentage difference between the central values using the two PDF sets relative to the NNPDF2.3 values, and the prediction using the conservative PDF with $\alpha_{\rm max} = 1.1$ (see Section 7.2.1).

next section. The results of Table 7.9 are also shown in Fig. 7.34, normalised to the NNPDF2.3 values.

For all of the shown cross-sections we observe stability between the NNPDF2.3 and NNPDF3.0 values, with all results varying by no more that the size of the corresponding PDF uncertainty. For top-quark pair production, going from NNPDF2.3 to NNPDF3.0 the cross-section decreases by about 1%, about half the PDF uncertainty. This can be understood recalling that the NNPDF3.0 gg luminosity is slightly softer than is NNPDF2.3 counterpart for $M_X \sim 400$ GeV. Note that NNPDF2.3 already gave a very good description of all available ATLAS and CMS 7 TeV and 8 TeV production data [204], though they were not included in that fit. For W production in association with charm quarks, we use a $N_F = 3$ scheme in order to retain the full charm mass dependence. Again, for this observable results are very stable when moving from NNPDF2.3 to NNPDF3.0.

Looking now at the Higgs production observables, for ttH the NNPDF3.0 result is about 1% larger than the NNPDF2.3 prediction, consistent with the expectation from the gg luminosity comparisons in Fig. 7.5 for $M_X \sim 500 = 700$ GeV. For the associated production channels, HW and HZ, driven by the $q\bar{q}$ luminosities, differences are well within one sigma, as would be expected from the luminosities in Fig. 7.5.

The results for the conservative PDF set included in the last column of Table 7.9 show that prediction obtained using this fit are generally consistent at the one-sigma level with the global results, with occasional differences up to around the two-sigma level, such as for example in $He^+\nu_e$. The predictions from conservative set are, of course, generally affected by larger PDF uncertainties due to the smaller dataset used,

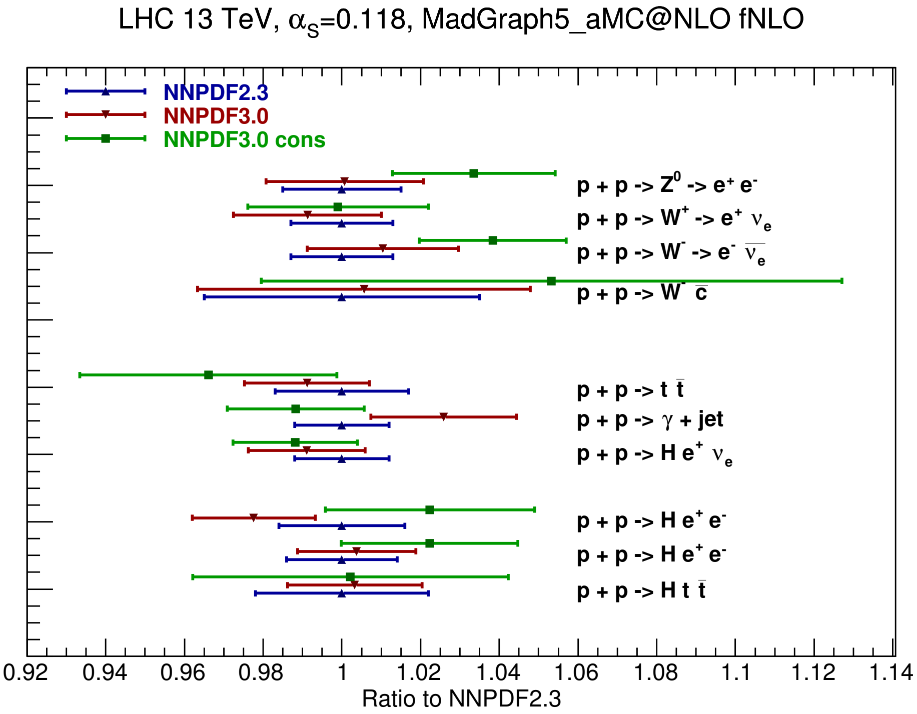


Figure 7.34: Graphical comparison the results of Tab 7.9. Results are shown normalised to the NNPDF2.3 central value.

though in several cases they are only slightly less precise that the global fit results, for example in inclusive W and Z production. On the other hand, for processes that depend on strangeness (like W+c) or that are gluon-driven (like $t\bar{t}$ and $t\bar{t}H$) the PDF uncertainties are substantially larger for the conservative PDFs than for the global fit.

7.4.3 Higgs production in gluon fusion

Following the general overview of LHC observables in the previous section, I will now focus specifically on Higgs production in gluon fusion, which is the dominant channel at the LHC and for which theoretical uncertainties are a limiting factor in the determination of Higgs properties. I will present predictions for the total cross-section at NLO and NNLO for the LHC at 13 TeV, comparing the default NNPDF3.0 set to NNPDF2.3 and to the various sets based on alternative datasets described in Section 7.2. The uncertainties shown are purely the PDF uncertainties with $\alpha_s(M_Z) = 0.118$, i.e. the α_s uncertainty is not included. The inclusive cross-sections are computed using iHixs 1.3.3, with $m_h = 125$ GeV, with renormalisation and factorisation scales set to $\mu_r = \mu_f = m_h$ and with the infinite top mass (effective theory) approach. The predictions here are therefore not meant to be realistic, however

	σ_{ggh} (pb) NLO	Pull	σ_{ggh} (pb) NNLO	Pull
NNPDF2.3	34.72 ± 0.33	-	46.39 ± 0.46	-
NNPDF3.0 with 2.3 data	34.06 ± 0.57	1.0	45.14 ± 0.74	1.4
NNPDF3.0 global	33.96 ± 0.61	1.1	45.01 ± 0.72	1.6
NN3.0 conservative $\alpha_{\max} = 1.1$	33.31 ± 0.54	2.2	43.70 ± 1.12	2.2
NNPDF3.0 no Jets	34.56 ± 1.04	0.2	45.32 ± 0.92	1.0
NNPDF3.0 noLHC	34.12 ± 0.80	0.7	45.10 ± 0.91	1.3
NNPDF3.0 HERA-only	31.96 ± 3.03	0.9	43.02 ± 2.21	1.5

Table 7.10: The total cross-section for Higgs production in gluon fusion at the LHC 13 TeV at NLO (left) and NNLO (right) for $\alpha_s(M_z) = 0.118$. Values are shown for the central NNPDF2.3 fits, and for NNPDF3.0 fits using different datasets (see Section 7.2). The pull P, defined in Eq. 7.6, is also given.

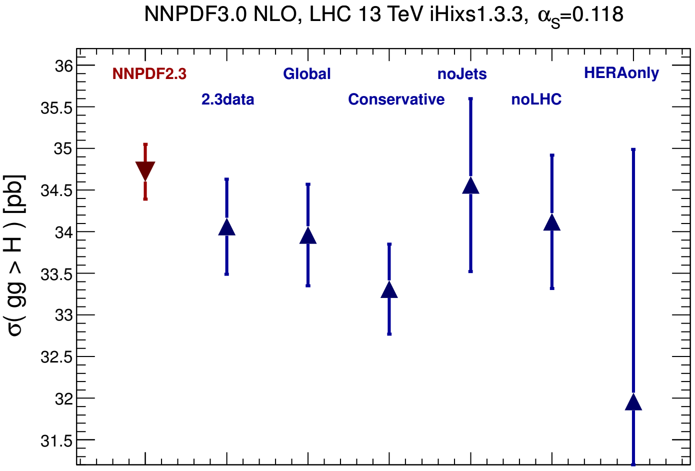


Figure 7.35: Graphical representation of the NLO results of Table 7.10.

many of the effects which are not included (such as electroweak corrections, or finite top, bottom and charm mass contributions) have a negligible PDF dependence, while α_s uncertainties are generally considered to be completely independent of the PDF uncertainty, given that the PDF and α_s uncertainties combine in quadrature even when correlated [205]. Hence the results here do provide an accurate assessment of the PDF dependence of the cross-section and its uncertainty.

The values of the Higgs gluon fusion cross-section are shown in Table 7.10 at NLO and NNLO for a variety of sets, and the results are also summarised graphically in Figs. 7.35 and 7.36. In Table 7.10 the pull of each prediction compared to the

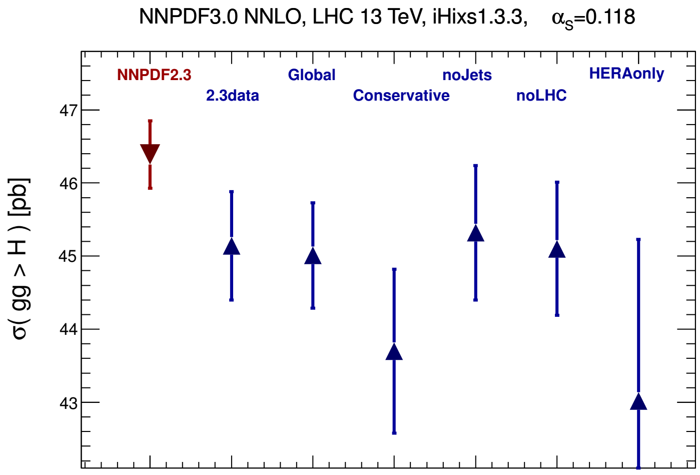


Figure 7.36: Graphical representation of the NNLO results of Table 7.10.

NNPDF2.3 result is also given, and is defined as

$$P \equiv \frac{\sigma_{ggh}(2.3) - \sigma_{ggh}(3.0)}{\sqrt{\Delta\sigma^2_{ggh}(2.3) + \Delta\sigma^2_{ggh}(3.0)}} \,, \tag{7.6}$$

where $\Delta\sigma_{ggh}$ is the one-sigma PDF uncertainty. This is similar to the concept of the distance used for PDFs, but here for a physical observable.

As expected from the comparison of the gluon-gluon luminosities in Fig. 7.6, the NNLO cross-section decreases by about two-sigma (in terms of the single set uncertainty) when going from NNPDF2.3 to NNPDF3.0, while the PDF uncertainty increases substantially. At NLO the effect is less marked, with the NNPDF2.3 and NNPDF3.0 in agreement at almost the one-sigma level. The results for the NNPDF3.0 PDFs based on a 2.3-like dataset are very similar to that of the global fit, so we must conclude that this change is largely due to methodological improvements (validated by the closure tests in Chapter 6), rather than the inclusion of new data.

The results for sets based on alternative datasets are generally consistent with each other and with the global fit at the one-sigma level, with the conservative PDFs leading to a lower result and the fit with no jet data to a slightly higher one. Uncertainties are of course larger for the sets, with statistical fluctuations, for instance for the NLO conservative set. The lowest central value is found for the HERA-only set, which is however affected by a PDF uncertainty which is a factor of two to three larger than the others. The pulls for these sets tell the same story, being of similar size in general, between 0.7 and 1.1 at NLO and between 1.3 and 1.6 at NNLO, with the fit without jet data giving better agreement with NNPDF2.3 and the conservative set giving worse agreement. On the whole there is little evidence of tension between datasets, and the differences seen are broadly consistent with statistical fluctuations.

7.4.4 New Physics particle production at high masses

New heavy particles at the TeV scale are a busy area of study as they are included in a wide variety of BSM scenarios and may be within the reach of the upgraded LHC. Production of such very massive particles probes PDFs at large x, where they are currently poorly known due to the lack of direct experimental information, and so the corresponding predictions for these particles are affected by substantial PDF uncertainties (see for example Refs. [206, 207]). Consequently, PDFs can be a limiting factor in the determination of exclusion regions for heavy particles, and so an accurate assessment of the PDFs uncertainties is therefore crucial for these searches. The NNPDF approach is advantageous in this respect in that it leads to uncertainty estimates which are not biased by assumptions on the functional form of PDFs. The only significant constraint on PDFs close to threshold comes from positivity, which is now implemented in an improved way, as discussed in Section. 5.6.

As an example, I will show here results for high-mass dilepton production and the pair production of supersymmetric particles. The first of these, high-mass dilepton production, is frequently used to search for new physics that couples to the electroweak sector, and thus it is important to provide precise predictions for the SM production mechanisms. We have computed the dilepton invariant mass distribution in $pp \to \gamma^*/Z \to l^+l^-$ events at the LHC at 14 TeV with NNPDF3.0 using FEWZ. Recall from Section 5.6 that positivity is always imposed at NLO, so it is not entirely trivial that it will also be fully constrained at NNLO. Results are shown in Fig. 7.37, in different M_{ll} bins, with each of the $N_{\rm rep} = 100$ Monte Carlo replicas given by a separate green dashed line, together with the resulting central values and one-sigma intervals. Both at NLO and NNLO all of the replicas are positive even in the highest invariant mass bins.

Predictions for the pair production of supersymmetric particles at the LHC 14 TeV are shown in Fig. 7.38. The computation has been performed using Prospino [208,209] with the NNPDF3.0 and NNPDF2.3 NLO global fits, and using settings as close as possible to those of Refs. [206, 207], though the only relevant physical input for this illustrative study are the sparticle masses. For these processes NNLO calculations are not available. We have produced results for gluino-gluino and squark-antisquark production, for three different values of the sparticle masses: 1, 2 and 3 TeV. This figure again shows the predictions for the $N_{\rm rep} = 100$ Monte Carlo replicas, this time

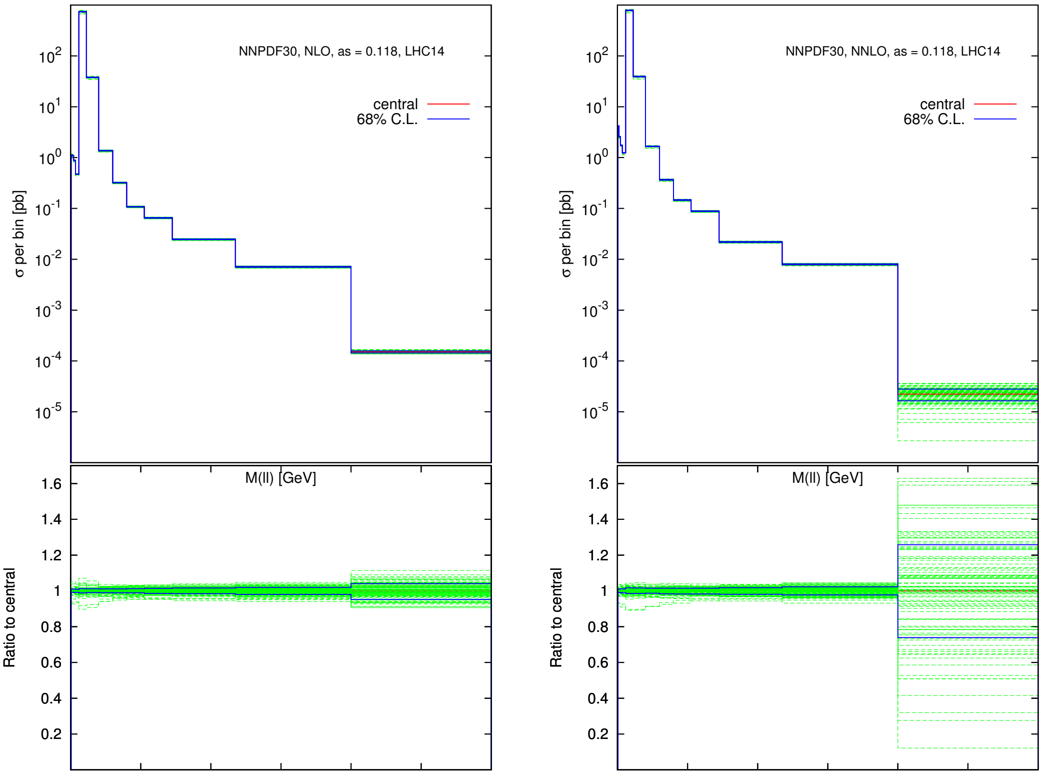


Figure 7.37: The dilepton invariant mass distribution in $pp \to \gamma^*/Z \to l^+l^-$ at the LHC 14 TeV with NNPDF3.0 at NLO (left) and NNLO (right) using FEWZ. Each green dashed line is the result for a single Monte Carlo replica PDF set, while the solid red line is the resulting average and the solid blue lines are the 68% confidence interval. Both the absolute result (top) and the ratio to the central value (bottom) are shown.

shown as dots with the central values and 68% confidence intervals given by lines.

In the case of gluino-gluino production, all replicas are strictly positive for $m_{\tilde{g}} < 3$ TeV. At 3 TeV, some replicas do lead to slightly negative cross-sections, though the number has improved in the new set: 15 in NNPDF2.3, and only 3 in NNPDF3.0. The small number of negative replicas means that they can be set to zero without impacting the central value or uncertainty of the distribution.

The squark-antisquark case is similar, with again all replicas giving positive values for the cross-section with $m_{\tilde{q}} < 3$ TeV, with negative values appearing for $m_{\tilde{q}} = 3$ TeV. For NNPDF2.3, a large number of replicas gave negative values, resulting in a central value which itself was negative. With the improved positivity prescription used in NNPDF3.0, the central value is now positive and only a small part of the 68% confidence level range is in the negative region. Some replicas are still negative, though this is to be expected, partly because positivity is imposed with a Lagrange multiplier which carries a large but finite penalty, but also because it is only imposed for a finite

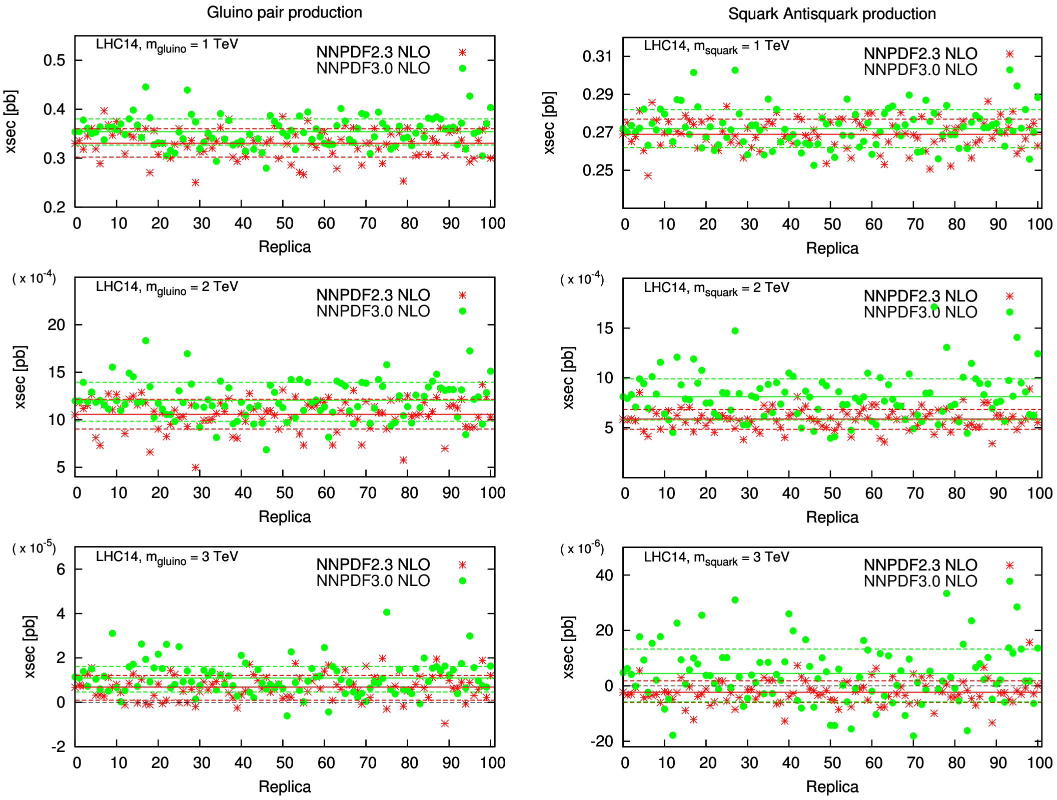


Figure 7.38: Cross-sections for NLO gluino-gluino (left) and squark-antisquark (right) pair production at the 14 TeV LHC with NNPDF3.0 (green) and NNPDF2.3 (red), for sparticle masses of 1 TeV (top), 2 TeV (middle) and 3 TeV (bottom). In each case, we show the predictions for the $N_{\rm rep} = 100$ Monte Carlo replicas as well as the average result and the 68% confidence interval.

number of standard model processes, and not for all possible processes. Note however the very large PDF uncertainties at the largest masses, which is around +200%,-100% for squark-antisquark production with $m_{\tilde{q}} = 3$ TeV.

From these results we conclude that the new implementation of positivity used in NNPDF3.0 provides significant improvement over NNPDF2.3, giving results that are generally positive even in the case of very heavy BSM particles.

Chapter 8

Conclusions and Outlook

In this thesis, I have stressed the importance of the precise and accurate determination of PDFs in order to understand the results of high energy particle physics experiments. The NNPDF methodology provides an effective way to fit PDFs, and with the NNPDF3.0 analysis we introduce new data and methodological features, to produce an updated set of PDFs suitable for use in future analyses. We have also for the first time demonstrated the validity of our methodology using closure tests.

The NNPDF3.0 analysis is in many ways a significant step up over the previous NNPDF sets. In addition to the new data described in Chapter 4, the methodology has been upgraded, and I have shown the extent of the testing we performed in order to determine the optimal genetic algorithm settings. The new features we developed have resulted in both an improvement in fitting speed and in fit quality, and so give a better overall PDF determination. I also demonstrated that a large number of methodological settings—like the fitting basis, the structure of the neural networks, or the presence of cross-validation—have only a minor impact on the fit, providing confidence in the stability of our results. Alongside the actual gains in the fit, Chapter 5 also highlights the power of the closure test framework in evaluating changes to the methodology, by providing simple, unbiased estimators of quality. The methodological development made for NNPDF3.0 will also have a substantial impact on our future work. The improved fitting speed makes it significantly easier both to perform large fits, of thousands of replicas or more, and also to perform many different fits, to test different features or to study combinations of datasets for example.

The closure test technique itself has provided interesting results. In Chapter 6 I demonstrated that the PDFs our approach generates match the correct answer to the degree we would expect. This is shown to be true both looking at agreement at the level of the central value of observables and PDFs, and also in measures of the PDF uncertainties. In the future it may be possible to develop more advanced and comprehensive measures of agreement, to further show the effectiveness of our methodology. There are also a large number of other tests which could be performed, for instance fits to inconsistent pseudo-data, or closure tests using a fixed functional form.

For the central NNPDF3.0 PDF sets, we found good quality of fit to the experimental data. The new PDFs are broadly consistent with the previous NNPDF2.3 PDFs, with some changes in the central values of the gluon and individual quark flavours of between 1 and 1.5 sigma. A separate series of reduced dataset fits demonstrate that a substantial fraction of this deviation is due to the improved methodology, though the introduction of the new LHC and HERA data do have an impact in shifting the central values and reducing the PDF uncertainties. Other tests demonstrate the stability of our results on changes of the treatment of systematic uncertainties, the fitting basis and the inclusion of nuclear corrections, while fits using a maximally consistent dataset indicate that data inconsistency should only have a minor effect on the fit. Comparisons at the level of LHC observables show a similar level of agreement with NNPDF2.3 as seen in the PDFs, with some more significant differences in, for example, the Higgs production cross-section.

Looking towards future NNPDF releases, there will likely be an updated set using the same methodology but including a number of important datasets published since NNPDF3.0. There is large amount of LHC run-I data which has not yet been included, and further in the future the first 13 TeV sets will be released. There are also a number of new releases from previous colliders, including the Tevatron legacy muon and electron asymmetry and the final combined HERA dataset, which have the potential to provide significant constraint on PDFs. With advances in theory, we can also start to include data from new processes like top quark differential distributions and prompt photon production. We are also currently working on a number of theoretical developments in parallel. Providing a direct determination of the charm quark, with an *intrinsic* non-zero component below threshold, has been a long term goal, and with the release of NNPDF3.0 has become more of a priority. Members of the collaboration have also very recently released a preliminary analysis of the first global PDF set to include large-x resummation effects, at NLO+NLL and NNLO+NNLL. In both cases the work is being done using largely the same NNPDF3.0 methodology described here.

NNPDF3.0 is a marked improvement over NNPDF2.3, both in terms of the amount of data included and the sophistication and reliability of the methodology used. Since its release NNPDF3.0 has already been used in a number of analyses, and with data collection already begun for the LHC run-II, it is likely that NNPDF3.0 PDFs will also be widely used there for both comparison of data to theory and for Monte Carlo simulations used to estimate uncertainties. The work presented here will therefore have a substantial impact on this next stage of high energy physics research.

Appendix A

Distances

We define the *distance* between the central values of two PDFs by

$$d_{\mathrm{CV}}[f_1, f_2](x) = \sqrt{N_{\mathrm{rep}}} \frac{|f_1(x) - f_2(x)|}{\sqrt{\sigma_1^2(x) + \sigma_2^2(x)}}, \tag{A.1}$$

where $f_i(x)$ is the central value of each PDF at the point x, and $\sigma_i(x)$ is the corresponding uncertainty. This gives the absolute difference between the central values in units of the combined PDF uncertainty on the mean. Note that as it is the mean of the distribution that is being compared, it is the uncertainty on the mean, not on the distribution itself, which is used, signified by the factor of $\sqrt{N_{\mathrm{rep}}}$ at the front.

Similarly we define the distance between the uncertainties of the two PDFs by

$$d_{\mathrm{SD}}[\sigma_1, \sigma_2](x) = \frac{|\sigma_1(x) - \sigma_2(x)|}{\sqrt{s_{\sigma,1}^2(x) + s_{\sigma,2}^2(x)}}, \tag{A.2}$$

where $s_\sigma(x)$ is the uncertainty on the uncertainty, given by

$$s_\sigma(x) = \sqrt{\frac{1}{N_{\mathrm{rep}}} \left(m_4(x) - \frac{N_{\mathrm{rep}} - 3}{N_{\mathrm{rep}} - 1} \sigma^4(x) \right)} \tag{A.3}$$

where $m_4(x)$ is the fourth central moment of the PDFs. As for the central value distance, this distance provides a measure of the difference between the uncertainties of the two PDFs, in meaningful units.

Fig. A.1 provides an example of a plot of distances between two PDF sets, here for the NNPDF2.3 and 3.0 NNLO fits. In order to evaluate a plot like this, it is important to understand what our expectation of the distances are. For a single point, we expect the mean to vary statistically according to a gaussian distribution, and so the distance to be below one 68% of the time etc. From PDF fits using identical settings and data,

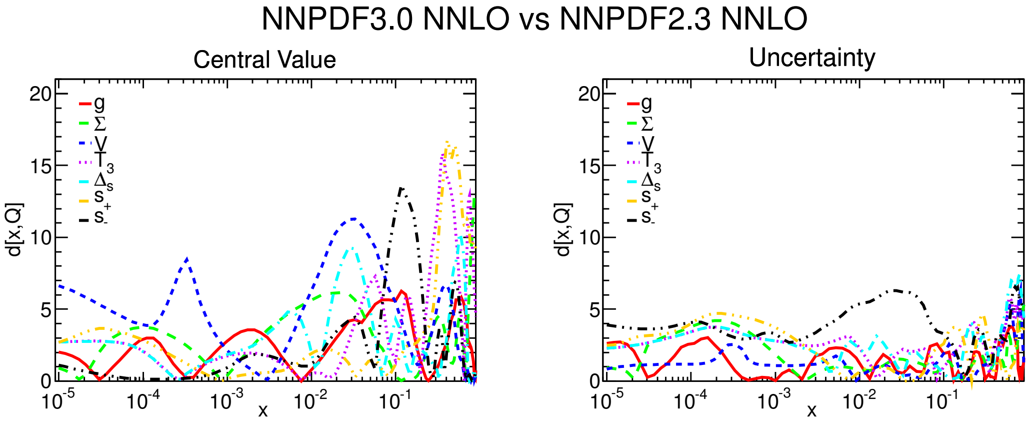


Figure A.1: Distances between NNPDF2.3 and NNPDF3.0 NNLO PDFs. Same as bottom of Fig. 7.1.

but with different random seed, we observe that distances are generally below three or four. This therefore provide our criteria for a significant difference between two PDF sets.

For distances between the central values of closure test PDFs and the input PDF, we use a slightly modified version of Eq. A.1,

$$d_{\mathrm{CV}}[f_c, f_t](x) = \frac{|f_c(x) - f_t(x)|}{\sigma_c(x)}, \tag{A.4}$$

as the uncertainty of the input PDF is irrelevant for this purpose. Note that the factor of $\sqrt{N_{\mathrm{rep}}}$ has also been dropped; this is mostly an aesthetic choice, based on the idea that here we are not looking at whether the central values themselves are compatible, but rather that the fit PDF is consistent within uncertainties with the input values.

Appendix B

Closure test estimators

In this appendix I will give details of various indicators and estimators used in Chapters 5 and 6. These include the different ways of defining the central χ^2 of a set of replicas as well as indicators of how successful the closure test fit is.

As mentioned in the main text, there are multiple different ways to define an 'average' χ^2 for a set of replicas, based on which level the average is taken at. Note that in general by χ^2 I will refer to what is actually the χ^2 per degree of data point. First, we can take the mean of the χ^2s calculated for the individual replicas PDFs f_k, i.e.

$$\langle \chi^2_{\rm rep} \rangle = \frac{1}{N_{\rm rep}} \sum_k \left(\frac{1}{N_{\rm dat}} \sum_{i,j} \left(d_i - t_i[f_k]\right) C^{-1}_{ij} \left(d_j - t_j[f_k]\right) \right) \ , \tag{B.1}$$

where d_i is the ith data point, $t_i[f_k]$ is the corresponding theory prediction calculated with replica k, and C_{ij} is the covariance matrix of the experimental data.

The average replica χ^2 given above describes how well on average the minimisation can fit the replica datasets. On the other hand, the central χ^2 provides a measure of how well the whole ensemble of replicas fits the data. There are two ways to define a central χ^2. Here, by central χ^2 I will always mean the χ^2 calculated using the average value of each observable

$$\chi^2_{\rm cent} = \frac{1}{N_{\rm dat}} \sum_{i,j} \left(d_i - \langle t_i \rangle\right) C^{-1}_{ij} \left(d_j - \langle t_j \rangle\right) \ , \tag{B.2}$$

where $\langle X \rangle$ is the average over the replicas, as above. This way of calculating the χ^2 uses the central values of each observable based on our full PDF set. We can also define another central χ^2 as the χ^2 to the average PDF $f_0 = \langle f_k \rangle$, which I will denote χ^2_0

$$\chi^2_0 = \frac{1}{N_{\rm dat}} \sum_{i,j} \left(d_i - t_i(f_0)\right) C^{-1}_{ij} \left(d_j - t_j(f_0)\right) \ . \tag{B.3}$$

For the majority of our fits, χ^2_{cent} and χ^2_0 are very similar, while $\langle\chi^2_{\mathrm{rep}}\rangle$ is generally slightly larger.

There are a number of additional ways we can use the χ^2 definitions provide a better best estimation of successfulness in a closure test. In Level 1 and Level 2 closure fits, we expect that the χ^2_{cent} of the fitted PDFs should, if the closure test is successful, reproduce the one computed using the input PDFs, i.e $\chi^2_{\mathrm{cent}}[f_{\mathrm{fit}}, \mathcal{D}_1] \approx \chi^2_{\mathrm{cent}}[f_{\mathrm{in}}, \mathcal{D}_1]$, where as in Chapter 6 by $\mathcal{D}_1$ we indicate that that we use the Level 1 pseudo-data. We can test whether this is the case for a particular fit by formally defining the statistical estimator

$$\Delta_{\chi^2} = \frac{\chi^2_{\mathrm{cent}}[f_{\mathrm{fit}}, \mathcal{D}_1] - \chi^2_{\mathrm{cent}}[f_{\mathrm{in}}, \mathcal{D}_1]}{\chi^2_{\mathrm{cent}}[f_{\mathrm{in}}, \mathcal{D}_1]}\,, \tag{B.4}$$

that is, the difference between the χ^2_{cent} of the closure test fit and the χ^2_{cent} of the input PDF set, both computed with respect to the same closure test dataset. This estimator is therefore a measure of how close the closure test fit reproduces the theoretical predictions of the input PDF. In particular, $\Delta_{\chi^2} > 0$ corresponds to underlearning (the optimal χ^2 has not been reached yet) and $\Delta_{\chi^2} = 0$ corresponds to perfect learning of the underlying law. $\Delta_{\chi^2} < 0$ can be connected to overlearning, though in practice a value slightly smaller than zero is acceptable, as for a particular set of pseudo-data there may be a set of PDFs which is more probably than the underlying law.

It is also convenient to use the χ^2 to define an indicator which measures the standard deviation over the replica sample in units of the data uncertainty. This can be defined as

$$\varphi_{\chi^2} \equiv \sqrt{\langle\chi^2_{\mathrm{rep}}\rangle - \chi^2_{\mathrm{cent}}}\,. \tag{B.5}$$

To see that this does what we want, we can multiply out the definition of the $\langle\chi^2_{\mathrm{rep}}\rangle$ given in Eq. B.1, we find that

$$\begin{aligned} N_{\mathrm{dat}}\,\langle\chi^2_{\mathrm{rep}}\rangle = &\sum_{i,j}\Big\langle t_i[f_k]\, C^{-1}_{ij}\, t_j[f_k]\Big\rangle - \sum_{i,j}\langle t_i[f_k]\rangle\, C^{-1}_{ij}\, d_j - \sum_{i,j} d_i\, C^{-1}_{ij}\, \langle t_j[f_k]\rangle \\ &+ \sum_{i,j} d_i\, C^{-1}_{ij}\, d_j\,. \end{aligned} \tag{B.6}$$

Likewise

$$\begin{aligned} N_{\mathrm{dat}}\,\chi^2_{\mathrm{cent}} = &\sum_{i,j}\langle t_i[f_k]\rangle\, C^{-1}_{ij}\, \langle t_j[f_k]\rangle - \sum_{i,j}\langle t_i[f_k]\rangle\, C^{-1}_{ij}\, d_j - \sum_{i,j} d_i\, C^{-1}_{ij}\, \langle t_j[f_k]\rangle \\ &+ \sum_{i,j} d_i\, C^{-1}_{ij}\, d_j\,. \end{aligned} \tag{B.7}$$

Taking the two expressions together

$$\langle \chi^2_{\rm rep} \rangle - \chi^2_{\rm cent} = \frac{1}{N_{\rm dat}} \sum_{i,j} \Big(\langle t_i[f_k] C^{-1}_{ij} t_j[f_k] \rangle - \langle t_i[f_k] \rangle C^{-1}_{ij} \langle t_j[f_k] \rangle \Big) . \tag{B.8}$$

Thus in terms of the covariance matrix of the theoretical predictions, defined as

$$t_{ij} \equiv \langle t_i[f_k] t_j[f_k] \rangle - \langle t_i[f_k] \rangle \langle t_j[f_k] \rangle , \tag{B.9}$$

we have

$$\varphi^2_{\chi^2} \equiv \frac{1}{N_{\rm dat}} \sum_{i,j} C^{-1}_{ij} t_{ij} , \tag{B.10}$$

i.e. the average over all the data points of the uncertainties and correlations of the theoretical predictions, t_{ij}, normalised according to the corresponding uncertainties and correlations of the data as expressed through the covariance matrix C_{ij}. If the covariance matrix was diagonal, i.e. in the absence of correlations, this would just be the variance of the predictions divided by the experimental variance averaged over data points. $\varphi^2_{\chi^2}$ is therefore the generalisation of this idea to the case with correlations. Note that this estimator can be calculated for any Monte Carlo PDF fit, not just closure test fits.

The final closure test estimator I will introduce here is ξ_σ, which describes the fraction of possible PDFs central values within one standard deviation of the theory value. Unlike the estimators described above, ξ_σ is calculated at the level of the PDFs rather than the level of the data. The central idea here is that, for a correctly determined PDF set, the PDF uncertainties should describe the probability that the true theory value for the PDFs can take a particular value, and that there should be a 68% probability that the theory lies within one sigma of the central value (assuming that the uncertainties are Gaussian). We can turn this around, and say that a given input PDF should be within the one-sigma band of 68% of fits to pseudo-data generated from that input. On this basis we can define

$$\xi_\sigma = \frac{1}{N_{\rm fits}} \sum_{l=1}^{N_{\rm fits}} I_{[-\sigma^l_{\rm fit}, \sigma^l_{\rm fit}]} \Big(\langle f^l_{\rm fit} \rangle - f_{\rm in} \Big) , \tag{B.11}$$

where $f_{\rm fit}$, $f_{\rm in}$ and $\langle X \rangle$ are defined as above, $\sigma_{\rm fit}$ is the standard deviation over the replicas of the fitted PDF, l runs over the$N_{\rm fits}$ closure test fits making up the sample, each with different pseudo-datasets. $I_A(x)$ denotes the indicator function of the interval A, that is, it is only non-zero if its argument lies in the interval A, and one otherwise.

In practice we make a few modifications to the way ξ_σ is defined, in order to generate a large enough sample. Firstly, we average ξ_σ over the PDFs, and over several points in x for each distribution. The values quoted in Section 6.4.2 are generated using a sample of the PDFs at 20 points in x between 10^{-5} and 1, half of them log spaced below 0.1 and the rest linearly spaced. This means that the actual definition of ξ_σ I use in this thesis is given (generalising to n standard deviations) by

$$\xi_{n\sigma} = \frac{1}{N_{\mathrm{PDF}}}\frac{1}{N_x}\frac{1}{N_{\mathrm{fits}}}\sum_{i=1}^{N_{\mathrm{PDF}}}\sum_{j=1}^{N_x}\sum_{l=1}^{N_{\mathrm{fits}}} I_{[-n\sigma_{\mathrm{fit}}^{i(l)}(x_j),n\sigma_{\mathrm{fit}}^{i(l)}(x_j)]}\left(\langle f_{\mathrm{fit}}^{i(l)}(x_j)\rangle - f_{\mathrm{in}}^{i}(x_j)\right) . \quad \text{(B.12)}$$

The estimators $\xi_{1\sigma}$, $\xi_{2\sigma}$, ... provide the fraction of those fits for which the input PDF falls within one sigma, two sigma, etc. of the central PDF $\bar{f}_{\mathrm{fit}}^{i(l)}$, averaged over PDF flavours and values of x. In a successful closure test we must thus have that $\xi_{1\sigma} \approx 0.68$, $\xi_{2\sigma} \approx 0.95$, etc.

The second modification is that in practice, instead of generating a large number of closure test fits—something which would take a huge amount of time and resourses—we can instead approximate the mean PDFs of each fit, $\langle f_{\mathrm{fit}}^{i(l)}\rangle$, by fitting a single replica to each set of closure test data at Level 1, i.e. without additional replica fluctuations. We can then replace the individual values of $\sigma^{i(l)}$ in Eq. B.12 with the corresponding values from a single 100 replica fit, making use of the fact that the variation in the PDF uncertainties between different closure test fits is small.

Bibliography

[1] **ATLAS** Collaboration, G. Aad et al., *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, *Phys.Lett.* **B716** (2012) 1–29, [arXiv:1207.7214].

[2] **CMS** Collaboration, S. Chatrchyan et al., *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, *Phys.Lett.* **B716** (2012) 30–61, [arXiv:1207.7235].

[3] **The NNPDF** Collaboration, R. D. Ball et al., *A determination of parton distributions with faithful uncertainty estimation*, *Nucl. Phys.* **B809** (2009) 1–63, [arXiv:0808.1231].

[4] **The NNPDF** Collaboration, R. D. Ball et al., *A first unbiased global NLO determination of parton distributions and their uncertainties*, *Nucl. Phys.* **B838** (2010) 136–206, [arXiv:1002.4407].

[5] **The NNPDF** Collaboration, R. D. Ball et al., *Unbiased global determination of parton distributions and their uncertainties at NNLO and at LO*, *Nucl.Phys.* **B855** (2012) 153–221, [arXiv:1107.2652].

[6] **NNPDF** Collaboration, R. D. Ball et al., *Parton distributions for the LHC Run II*, *JHEP* **1504** (2015) 040, [arXiv:1410.8849].

[7] R. K. Ellis, W. J. Stirling, and B. Webber, *QCD and collider physics*, *Camb.Monogr.Part.Phys.Nucl.Phys.Cosmol.* **8** (1996) 1–435.

[8] M. Gell-Mann, *A Schematic Model of Baryons and Mesons*, *Phys.Lett.* **8** (1964) 214–215.

[9] G. Zweig, *An SU(3) model for strong interaction symmetry and its breaking. Version 1*, .

[10] M. Breidenbach, J. I. Friedman, H. W. Kendall, E. D. Bloom, D. Coward, et al., *Observed Behavior of Highly Inelastic electron-Proton Scattering*, *Phys.Rev.Lett.* **23** (1969) 935–939.

[11] J. D. Bjorken, *CURRENT ALGEBRA AT SMALL DISTANCES*, *Conf. Proc.* **C670717** (1967) 55–81.

[12] R. Feynman, *Partons*, in *The Past Decade in Particle Theory* (E. Sudarshan and Y. Ne'eman, eds.), pp. 773–813. Gordon and Breach, London, 1971.

[13] **PLUTO** Collaboration, C. Berger et al., *Jet Analysis of the Υ (9.46) Decay Into Charged Hadrons*, *Phys.Lett.* **B82** (1979) 449.

[14] J. C. Collins, D. E. Soper, and G. F. Sterman, *Factorization for Short Distance Hadron - Hadron Scattering*, *Nucl. Phys.* **B261** (1985) 104.

[15] V. Gribov and L. Lipatov, *Deep inelastic e p scattering in perturbation theory*, *Sov.J.Nucl.Phys.* **15** (1972) 438–450.

[16] G. Altarelli and G. Parisi, *Asymptotic Freedom in Parton Language*, *Nucl.Phys.* **B126** (1977) 298.

[17] Y. L. Dokshitzer, *Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics.*, *Sov.Phys.JETP* **46** (1977) 641–653.

[18] G. P. Salam and J. Rojo, *A Higher Order Perturbative Parton Evolution Toolkit (HOPPET)*, *Comput. Phys. Commun.* **180** (2009) 120–156, [`arXiv:0804.3755`].

[19] V. Bertone, S. Carrazza, and J. Rojo, *APFEL: A PDF Evolution Library with QED corrections*, *Comput.Phys.Commun.* **185** (2014) 1647–1668, [`arXiv:1310.1394`].

[20] **The NNPDF** Collaboration, L. Del Debbio, S. Forte, J. I. Latorre, A. Piccione, and J. Rojo, *Neural network determination of parton distributions: The nonsinglet case*, *JHEP* **03** (2007) 039, [`hep-ph/0701127`].

[21] A. Buckley, J. Ferrando, S. Lloyd, K. Nordstrm, B. Page, et al., *LHAPDF6: parton density access in the LHC precision era*, *Eur.Phys.J.* **C75** (2015), no. 3 132, [`arXiv:1412.7420`].

[22] **ZEUS, H1** Collaboration, H. Abramowicz et al., *Combination of Measurements of Inclusive Deep Inelastic $e^{\pm}p$ Scattering Cross Sections and QCD Analysis of HERA Data*, `arXiv:1506.0604`.

[23] S. Alekhin, J. Bluemlein, and S. Moch, *The ABM parton distributions tuned to LHC data*, *Phys.Rev.* **D89** (2014), no. 5 054028, [`arXiv:1310.3059`].

[24] L. Harland-Lang, A. Martin, P. Motylinski, and R. Thorne, *Parton distributions in the LHC era: MMHT 2014 PDFs*, `arXiv:1412.3989`.

[25] **CTEQ-TEA** Collaboration, C.-P. Yuan, "Progress in CTEQ-TEA PDF Analysis." Talk given at DIS2015, Dallas, Texas, April, 2015, [Slides].

[26] A. D. Martin, W. J. Stirling, R. S. Thorne, and G. Watt, *Parton distributions for the LHC*, *Eur. Phys. J.* **C63** (2009) 189–285, [`arXiv:0901.0002`].

[27] A. D. Martin, R. G. Roberts, and W. J. Stirling, *Structure Function Analysis and psi, Jet, W, Z Production: Pinning Down the Gluon*, *Phys. Rev.* **D37** (1988) 1161.

[28] A. D. Martin, R. Roberts, W. J. Stirling, and R. Thorne, *Parton distributions: A New global analysis*, *Eur.Phys.J.* **C4** (1998) 463–496, [`hep-ph/9803445`].

[29] A. D. Martin, R. G. Roberts, W. J. Stirling, and R. S. Thorne, *Physical gluons and high-E(T) jets*, *Phys. Lett.* **B604** (2004) 61–68, [`hep-ph/0410230`].

[30] **CTEQ** Collaboration, J. Botts et al., *CTEQ parton distributions and flavor dependence of sea quarks*, *Phys.Lett.* **B304** (1993) 159–166, [`hep-ph/9303255`].

[31] P. M. Nadolsky et al., *Implications of CTEQ global analysis for collider observables*, *Phys. Rev.* **D78** (2008) 013004, [`arXiv:0802.0007`].

[32] R. D. Ball, L. Del Debbio, J. Feltesse, S. Forte, A. Glazov, et al., *Benchmarking of parton distributions and their uncertainties*, .

[33] R. D. Ball, S. Carrazza, L. Del Debbio, S. Forte, J. Gao, et al., *Parton Distribution Benchmarking with LHC Data*, *JHEP* **1304** (2013) 125, [`arXiv:1211.5142`].

[34] S. Carrazza, J. I. Latorre, J. Rojo, and G. Watt, *A compression algorithm for the combination of PDF sets*, `arXiv:1504.0646`.

[35] J. Gao and P. Nadolsky, *A meta-analysis of parton distribution functions*, *JHEP* **1407** (2014) 035, [`arXiv:1401.0013`].

[36] J. Pumplin, H. L. Lai, and W. K. Tung, *The charm parton content of the nucleon*, *Phys. Rev.* **D75** (2007) 054029, [`hep-ph/0701220`].

[37] S. Dulat, T.-J. Hou, J. Gao, J. Huston, J. Pumplin, et al., *Intrinsic Charm Parton Distribution Functions from CTEQ-TEA Global Analysis*, *Phys.Rev.* **D89** (2014), no. 7 073004, [`arXiv:1309.0025`].

[38] P. Jimenez-Delgado, T. J. Hobbs, J. T. Londergan, and W. Melnitchouk, *New limits on intrinsic charm in the nucleon from global analysis of parton distributions*, *Phys. Rev. Lett.* **114** (2015), no. 8 082002, [`arXiv:1408.1708`].

[39] R. D. Ball, V. Bertone, M. Bonvini, S. Forte, P. G. Merrild, J. Rojo, and L. Rottoli, *Intrinsic charm in a matched general-mass scheme*, `arXiv:1510.0000`.

[40] **NNPDF** Collaboration, R. D. Ball et al., *Parton distributions with QED corrections*, *Nucl.Phys.* **B877** (2013) 290–320, [`arXiv:1308.0598`].

[41] A. D. Martin, R. G. Roberts, W. J. Stirling, and R. S. Thorne, *Parton distributions incorporating QED contributions*, *Eur. Phys. J.* **C39** (2005) 155–161, [`hep-ph/0411040`].

[42] A. Martin, A. T. Mathijssen, W. Stirling, R. Thorne, B. Watt, et al., *Extended Parameterisations for MSTW PDFs and their effect on Lepton Charge Asymmetry from W Decays*, *Eur.Phys.J.* **C73** (2013), no. 2 2318, [`arXiv:1211.1215`].

[43] J. Pumplin et al., *New generation of parton distributions with uncertainties from global QCD analysis*, *JHEP* **07** (2002) 012, [`hep-ph/0201195`].

[44] H.-L. Lai et al., *New parton distributions for collider physics*, *Phys. Rev.* **D82** (2010) 074024, [`arXiv:1007.2241`].

[45] **NNPDF** Collaboration, R. D. Ball, V. Bertone, S. Carrazza, C. S. Deans, L. Del Debbio, et al., *Parton distributions with LHC data*, *Nucl.Phys.* **B867** (2013) 244–289, [`arXiv:1207.1303`].

[46] W. McCulloch and W. Pitts, *A logical calculus of the ideas immanent in nervous activity*, *The bulletin of mathematical biophysics* **5** (1943), no. 4 115–133.

[47] B. D. Ripley and N. L. Hjort, *Pattern Recognition and Neural Networks.* Cambridge University Press, New York, NY, USA, 1st ed., 1995.

[48] **CDF, D0** Collaboration, T. Aaltonen et al., *Combination of Tevatron searches for the standard model Higgs boson in the W+W- decay mode*, *Phys.Rev.Lett.* **104** (2010) 061802, [`arXiv:1001.4162`].

[49] **D0** Collaboration, V. Abazov et al., *Search for single top quark production in $p\bar{p}$ collisions at $\sqrt{s}$ = 1.96-TeV*, *Phys.Lett.* **B622** (2005) 265–276, [`hep-ex/0505063`].

[50] **D0** Collaboration, V. Abazov et al., *Search for associated production of charginos and neutralinos in the trilepton final state using 2.3 fb^{-1} of data*, *Phys.Lett.* **B680** (2009) 34–43, [`arXiv:0901.0646`].

[51] J. Freeman, J. Lewis, W. Ketchum, S. Poprocki, A. Pronko, et al., *An Artificial neural network based b jet identification algorithm at the CDF Experiment*, *Nucl.Instrum.Meth.* **A663** (2012) 37–47, [`arXiv:1108.4738`].

[52] **D0** Collaboration, V. Abazov et al., *b-Jet Identification in the D0 Experiment*, *Nucl.Instrum.Meth.* **A620** (2010) 490–517, [`arXiv:1002.4224`].

[53] G. D'Agostini, *On the use of the covariance matrix to fit correlated data*, *Nucl.Instrum.Meth.* **A346** (1994) 306–311.

[54] **The NNPDF** Collaboration, R. D. Ball et al., *Fitting Parton Distribution Data with Multiplicative Normalization Uncertainties*, *JHEP* **05** (2010) 075, [`arXiv:0912.2276`].

[55] G. Altarelli, S. Forte, and G. Ridolfi, *On positivity of parton distributions*, *Nucl. Phys.* **B534** (1998) 277–296, [`hep-ph/9806345`].

[56] **The NNPDF** Collaboration, R. D. Ball et al., *Impact of Heavy Quark Masses on Parton Distributions and LHC Phenomenology*, *Nucl. Phys.* **B849** (2011) 296–363, [`arXiv:1101.1300`].

[57] **CMS** Collaboration, S. Chatrchyan et al., *Measurement of the electron charge asymmetry in inclusive W production in pp collisions at $\sqrt{s} = 7$ TeV*, *Phys.Rev.Lett.* **109** (2012) 111806, [`arXiv:1206.2598`].

[58] **ATLAS** Collaboration, G. Aad et al., *Measurement of the inclusive $W^{\pm}$ and Z/γ^* cross sections in the electron and muon decay channels in pp collisions at $\sqrt{s}= 7$ TeV with the ATLAS detector*, *Phys.Rev.* **D85** (2012) 072004, [`arXiv:1109.5141`].

[59] **LHCb Collaboration** Collaboration, R. Aaij et al., *Inclusive W and Z production in the forward region at $\sqrt{s} = 7$ TeV*, *JHEP* **1206** (2012) 058, [`arXiv:1204.1620`].

[60] **ATLAS** Collaboration, G. Aad et al., *Measurement of inclusive jet and dijet production in pp collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector*, *Phys. Rev.* **D86** (2012) 014022, [`arXiv:1112.6297`].

[61] **New Muon** Collaboration, M. Arneodo et al., *Accurate measurement of F_2^d/F_2^p and $R_d - R_p$*, *Nucl. Phys.* **B487** (1997) 3–26, [`hep-ex/9611022`].

[62] **New Muon** Collaboration, M. Arneodo et al., *Measurement of the proton and deuteron structure functions, F_2^p and F_2^d, and of the ratio σ_L/σ_T*, *Nucl. Phys.* **B483** (1997) 3–43, [`hep-ph/9610231`].

[63] **BCDMS** Collaboration, A. C. Benvenuti et al., *A high statistics measurement of the proton structure functions $f_2(x, q^2)$ and r from deep inelastic muon scattering at high q^2*, *Phys. Lett.* **B223** (1989) 485.

[64] **BCDMS** Collaboration, A. C. Benvenuti et al., *A high statistics measurement of the deuteron structure functions $f_2(x, q^2)$ and r from deep inelastic muon scattering at high q^2*, *Phys. Lett.* **B237** (1990) 592.

[65] L. W. Whitlow, E. M. Riordan, S. Dasu, S. Rock, and A. Bodek, *Precise measurements of the proton and deuteron structure functions from a global analysis of the SLAC deep inelastic electron scattering cross-sections*, *Phys. Lett.* **B282** (1992) 475–482.

[66] **H1 and ZEUS** Collaboration, F. Aaron et al., *Combined Measurement and QCD Analysis of the Inclusive $e^{\pm}p$ Scattering Cross Sections at HERA*, *JHEP* **1001** (2010) 109, [`arXiv:0911.0884`].

[67] **H1** Collaboration, F. D. Aaron et al., *Measurement of the Proton Structure Function F_L at Low x*, *Phys. Lett.* **B665** (2008) 139–146, [`arXiv:0805.2809`].

[68] **ZEUS** Collaboration, J. Breitweg et al., *Measurement of $D^{*\pm}$ production and the charm contribution to F_2 in deep inelastic scattering at HERA*, *Eur. Phys. J.* **C12** (2000) 35–52, [`hep-ex/9908012`].

[69] **ZEUS** Collaboration, S. Chekanov et al., *Measurement of $D^{*\pm}$ production in deep inelastic $e^{\pm}p$ scattering at HERA*, *Phys. Rev.* **D69** (2004) 012004, [`hep-ex/0308068`].

[70] **ZEUS** Collaboration, S. Chekanov et al., *Measurement of $D^{\pm}$ and D^0 production in deep inelastic scattering using a lifetime tag at HERA*, *Eur. Phys. J.* **C63** (2009) 171–188, [arXiv:0812.3775].

[71] **ZEUS** Collaboration, S. Chekanov et al., *Measurement of charm and beauty production in deep inelastic ep scattering from decays into muons at HERA*, *Eur. Phys. J.* **C65** (2010) 65–79, [arXiv:0904.3487].

[72] **H1** Collaboration, C. Adloff et al., *Measurement of $D^{*\pm}$ meson production and F_2^c in deep inelastic scattering at HERA*, *Phys. Lett.* **B528** (2002) 199–214, [hep-ex/0108039].

[73] **H1** Collaboration, F. D. Aaron et al., *Measurement of the $D*$ Meson Production Cross Section and F_2^c, at High Q^2, in ep Scattering at HERA*, *Phys. Lett.* **B686** (2010) 91–100, [arXiv:0911.3989].

[74] **H1** Collaboration, F. D. Aaron et al., *Measurement of the Charm and Beauty Structure Functions using the H1 Vertex Detector at HERA*, *Eur. Phys. J.* **C65** (2010) 89–109, [arXiv:0907.2643].

[75] **ZEUS** Collaboration, S. Chekanov et al., *Measurement of high-Q^2 neutral current deep inelastic e^-p scattering cross sections with a longitudinally polarised electron beam at HERA*, *Eur. Phys. J.* **C62** (2009) 625–658, [arXiv:0901.2385].

[76] **ZEUS** Collaboration, S. Chekanov et al., *Measurement of charged current deep inelastic scattering cross sections with a longitudinally polarised electron beam at HERA*, *Eur. Phys. J.* **C61** (2009) 223–235, [arXiv:0812.4620].

[77] **CHORUS** Collaboration, G. Onengut et al., *Measurement of nucleon structure functions in neutrino scattering*, *Phys. Lett.* **B632** (2006) 65–75.

[78] **NuTeV** Collaboration, M. Goncharov et al., *Precise measurement of dimuon production cross-sections in $\nu_\mu Fe$ and $\bar{\nu}_\mu Fe$ deep inelastic scattering at the Tevatron*, *Phys. Rev.* **D64** (2001) 112006, [hep-ex/0102049].

[79] D. A. Mason, *Measurement of the strange - antistrange asymmetry at NLO in QCD from NuTeV dimuon data*, . FERMILAB-THESIS-2006-01.

[80] G. Moreno et al., *Dimuon production in proton - copper collisions at $\sqrt{s}$ = 38.8-GeV*, *Phys. Rev.* **D43** (1991) 2815–2836.

[81] **NuSea** Collaboration, J. C. Webb et al., *Absolute Drell-Yan dimuon cross sections in 800-GeV/c p p and p d collisions*, hep-ex/0302019.

[82] J. C. Webb, *Measurement of continuum dimuon production in 800-GeV/c proton nucleon collisions*, hep-ex/0301031.

[83] **FNAL E866/NuSea** Collaboration, R. S. Towell et al., *Improved measurement of the anti-d/anti-u asymmetry in the nucleon sea*, *Phys. Rev.* **D64** (2001) 052002, [hep-ex/0103030].

[84] **CDF** Collaboration, T. Aaltonen et al., *Direct Measurement of the W Production Charge Asymmetry in $p\bar{p}$ Collisions at $\sqrt{s} = 1.96$ TeV*, *Phys. Rev. Lett.* **102** (2009) 181801, [arXiv:0901.2169].

[85] **CDF** Collaboration, T. A. Aaltonen et al., *Measurement of $d\sigma/dy$ of Drell-Yan e^+e^- pairs in the Z Mass Region from $p\bar{p}$ Collisions at $\sqrt{s} = 1.96$ TeV*, *Phys. Lett.* **B692** (2010) 232–239, [arXiv:0908.3914].

[86] **D0** Collaboration, V. M. Abazov et al., *Measurement of the shape of the boson rapidity distribution for $p\bar{p} \to Z/\gamma^* \to e^+e^- + X$ events produced at $\sqrt{s}$=1.96-TeV*, *Phys. Rev.* **D76** (2007) 012003, [hep-ex/0702025].

[87] **CDF** Collaboration, T. Aaltonen et al., *Measurement of the Inclusive Jet Cross Section at the Fermilab Tevatron p-pbar Collider Using a Cone-Based Jet Algorithm*, *Phys. Rev.* **D78** (2008) 052006, [arXiv:0807.2204].

[88] **D0** Collaboration, V. M. Abazov et al., *Measurement of the inclusive jet cross-section in $p\bar{p}$ collisions at $\sqrt{s}$=1.96-TeV*, *Phys. Rev. Lett.* **101** (2008) 062001, [arXiv:0802.2400].

[89] **ZEUS** Collaboration, A. Cooper Sarkar, *Measurement of high-Q^2 neutral current deep inelastic e+p scattering cross sections with a longitudinally polarised positron beam at HERA*, arXiv:1208.6138.

[90] **ZEUS** Collaboration, H. Abramowicz et al., *Measurement of high-Q^2 charged current deep inelastic scattering cross sections with a longitudinally polarised positron beam at HERA*, *Eur.Phys.J.* **C70** (2010) 945–963, [arXiv:1008.3493].

[91] **H1** Collaboration, F. Aaron et al., *Inclusive Deep Inelastic Scattering at High Q^2 with Longitudinally Polarised Lepton Beams at HERA*, *JHEP* **1209** (2012) 061, [arXiv:1206.7007].

[92] **H1** Collaboration, F. Aaron et al., *Measurement of the Inclusive $e^{\pm}p$ Scattering Cross Section at High Inelasticity y and of the Structure Function F_L*, *Eur.Phys.J.* **C71** (2011) 1579, [arXiv:1012.4355].

[93] **H1 , ZEUS** Collaboration, H. Abramowicz et al., *Combination and QCD Analysis of Charm Production Cross Section Measurements in Deep-Inelastic ep Scattering at HERA*, *Eur.Phys.J.* **C73** (2013) 2311, [arXiv:1211.1182].

[94] **CDF - Run II** Collaboration, A. Abulencia et al., *Measurement of the Inclusive Jet Cross Section using the k_{T} algorithm in $p\bar{p}$ Collisions at $\sqrt{s}$=1.96 TeV with the CDF II Detector*, *Phys. Rev.* **D75** (2007) 092006, [hep-ex/0701051].

[95] **ATLAS** Collaboration, G. Aad et al., *Measurement of the inclusive jet cross section in pp collisions at $\sqrt{s}$=2.76 TeV and comparison to the inclusive jet cross section at $\sqrt{s}$=7 TeV using the ATLAS detector*, *Eur.Phys.J.* **C73** (2013) 2509, [arXiv:1304.4739].

[96] **ATLAS** Collaboration, G. Aad et al., *Measurement of the high-mass Drell–Yan differential cross-section in pp collisions at $\sqrt{s}$=7 TeV with the ATLAS detector*, *Phys.Lett.* **B725** (2013) 223–242, [arXiv:1305.4192].

[97] **ATLAS** Collaboration, G. Aad et al., *Measurement of the Transverse Momentum Distribution of W Bosons in pp Collisions at $\sqrt{s} = 7$ TeV with the ATLAS Detector*, *Phys.Rev.* **D85** (2012) 012005, [arXiv:1108.6308].

[98] **ATLAS** Collaboration, G. Aad et al., *Measurement of the cross section for top-quark pair production in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector using final states with two high-pt leptons*, *JHEP* **1205** (2012) 059, [arXiv:1202.4892].

[99] **ATLAS** Collaboration, G. Aad et al., *Measurement of the $t\bar{t}$ production cross-section in pp collisions at sqrts = 7 TeV using kinematic information of lepton+jets events*, ATLAS-CONF-2011-121, ATLAS-COM-CONF-2011-132.

[100] **ATLAS** Collaboration, G. Aad et al., *Measurement of the $t\bar{t}$ production cross-section in pp collisions at $\sqrt{s} = 8$ TeV using $e\mu$ events with b-tagged jets*, ATLAS-CONF-2013-097, ATLAS-COM-CONF-2013-112.

[101] **CMS** Collaboration, S. Chatrchyan et al., *Measurement of the muon charge asymmetry in inclusive pp to WX production at $\sqrt{s}$ = 7 TeV and an improved determination of light parton distribution functions*, *Phys.Rev.* **D90** (2014) 032004, [arXiv:1312.6283].

[102] **CMS** Collaboration, S. Chatrchyan et al., *Measurements of differential jet cross sections in proton-proton collisions at $\sqrt{s} = 7$ TeV with the CMS detector*, *Phys.Rev.* **D87** (2013) 112002, [arXiv:1212.6660].

[103] **CMS** Collaboration, S. Chatrchyan et al., *Measurement of associated W + charm production in pp collisions at $\sqrt{s}$ = 7 TeV*, arXiv:1310.1138.

[104] **CMS Collaboration** Collaboration, S. Chatrchyan et al., *Measurement of the differential and double-differential Drell-Yan cross sections in proton-proton collisions at $\sqrt{s}$ = 7 TeV*, *JHEP* **1312** (2013) 030, [arXiv:1310.7291].

[105] **CMS** Collaboration, S. Chatrchyan et al., *Measurement of the $t\bar{t}$ production cross section in the dilepton channel in pp collisions at $\sqrt{s}$ = 8 TeV*, *JHEP* **1402** (2014) 024, [arXiv:1312.7582].

[106] **CMS** Collaboration, S. Chatrchyan et al., *Measurement of the $t\bar{t}$ production cross section in the dilepton channel in pp collisions at $\sqrt{s} = 7$ TeV*, *JHEP* **1211** (2012) 067, [arXiv:1208.2671].

[107] **CMS** Collaboration, S. Chatrchyan et al., *Measurement of the $t\bar{t}$ production cross section in pp collisions at $\sqrt{s} = 7$ TeV with lepton + jets final states*, *Phys.Lett.* **B720** (2013) 83–104, [arXiv:1212.6682].

[108] **LHCb** Collaboration, R. Aaij et al., *Measurement of the cross-section for $Z \to e^+e^-$ production in pp collisions at $\sqrt{s} = 7$ TeV*, *JHEP* **1302** (2013) 106, [arXiv:1212.4620].

[109] **H1 Collaboration** Collaboration, V. Andreev et al., *Measurement of inclusive ep cross sections at high Q^2 at $\sqrt{s}$ = 225 and 252 GeV and of the longitudinal proton structure function F_L at HERA*, *Eur.Phys.J.* **C74** (2014) 2814, [arXiv:1312.4821].

[110] **ZEUS Collaboration** Collaboration, S. Chekanov et al., *Measurement of the Longitudinal Proton Structure Function at HERA*, *Phys.Lett.* **B682** (2009) 8–22, [arXiv:0904.1092].

[111] **ZEUS** Collaboration, H. Abramowicz, *Measurement of high-Q^2 neutral current deep inelastic e^+p scattering cross sections with a longitudinally polarised positron beam at HERA*, arXiv:1208.6138.

[112] **ZEUS s** Collaboration, H1, *Combination of Measurements of Inclusive Deep Inelastic $e^{\pm}p$ Scattering Cross Sections and QCD Analysis of HERA Data*, arXiv:1506.0604.

[113] V. Bertone and J. Rojo, *Parton Distributions with the Combined HERA Charm Production Cross Sections*, *AIP Conf.Proc.* **1523** (2012) 51–54, [arXiv:1212.0741].

[114] S. Alekhin, J. Blümlein, K. Daum, K. Lipka, and S. Moch, *Precise charm-quark mass from deep-inelastic scattering*, *Phys.Lett.* **B720** (2013) 172–176, [arXiv:1212.2355].

[115] J. Gao, M. Guzzi, and P. M. Nadolsky, *Charm quark mass dependence in a global QCD analysis*, *Eur.Phys.J.* **C73** (2013) 2541, [arXiv:1304.3494].

[116] S. A. Malik and G. Watt, *Ratios of W and Z cross sections at large boson p_T as a constraint on PDFs and background to new physics*, *JHEP* **1402** (2014) 025, [arXiv:1304.2424].

[117] **ATLAS** Collaboration, G. Aad et al., *Measurement of the transverse momentum distribution of $Z/\gamma*$ bosons in proton-proton collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector*, *Phys.Lett.* **B705** (2011) 415–434, [arXiv:1107.2381].

[118] W. Stirling and E. Vryonidou, *Charm production in association with an electroweak gauge boson at the LHC*, *Phys.Rev.Lett.* **109** (2012) 082002, [arXiv:1203.6781].

[119] S. Alekhin, J. Blümlein, L. Caminadac, K. Lipka, K. Lohwasser, et al., *Determination of Strange Sea Quark Distributions from Fixed-target and Collider Data*, arXiv:1404.6469.

[120] **ATLAS** Collaboration, G. Aad et al., *Measurement of the production of a W boson in association with a charm quark in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector*, *JHEP* **1405** (2014) 068, [arXiv:1402.6263].

[121] **LHCb** Collaboration, *Measurement of the cross-section for $Z \to \mu\mu$ production with 1 fb^{-1} of pp collisions at $\sqrt{s}$=7 TeV* , LHCb-CONF-2013-007.

[122] **LHCb** Collaboration, *Inclusive low mass Drell-Yan production in the forward region at $\sqrt{s} = 7$ TeV*, LHCb-ANA-2012-029.

[123] **CMS** Collaboration, S. Chatrchyan et al., *Measurement of the Inclusive Jet Cross Section in pp Collisions at $\sqrt{s} = 7$ TeV*, *Phys.Rev.Lett.* **107** (2011) 132001, [arXiv:1106.0208].

[124] B. Watt, P. Motylinski, and R. Thorne, *The Effect of LHC Jet Data on MSTW PDFs*, *Eur.Phys.J.* **C74** (2014) 2934, [arXiv:1311.5703].

[125] **CMS Collaboration** Collaboration, V. Khachatryan et al., *Constraints on parton distribution functions and extraction of the strong coupling constant from the inclusive jet cross section in pp collisions at $\sqrt{s} = 7$ TeV*, arXiv:1410.6765.

[126] M. L. Mangano and J. Rojo, *Cross Section Ratios between different CM energies at the LHC: opportunities for precision measurements and BSM sensitivity*, *JHEP* **1208** (2012) 010, [arXiv:1206.3557].

[127] **ATLAS Collaboration** Collaboration, G. Aad et al., *Measurement of dijet cross sections in pp collisions at 7 TeV centre-of-mass energy using the ATLAS detector*, *JHEP* **1405** (2014) 059, [arXiv:1312.3524].

[128] **CMS** Collaboration, S. Chatrchyan et al., *Top pair cross section in dileptons*, CMS-PAS-TOP-12-007.

[129] G. P. Salam and G. Soyez, *A practical Seedless Infrared-Safe Cone jet algorithm*, *JHEP* **05** (2007) 086, [arXiv:0704.0292].

[130] **D0** Collaboration, V. M. Abazov et al., *Measurement of the muon charge asymmetry in $p\bar{p} \to W$+$X \to \mu\nu$ + X events at $\sqrt{s}$=1.96??TeV*, *Phys.Rev.* **D88** (2013) 091102, [arXiv:1309.2591].

[131] **D0** Collaboration, V. M. Abazov et al., *Measurement of the electron charge asymmetry in $\boldsymbol{p\bar{p} \to W + X \to e\nu + X}$ decays in $\boldsymbol{p\bar{p}}$ collisions at $\boldsymbol{\sqrt{s} = 1.96}$ TeV*, *Phys.Rev.* **D91** (2015), no. 3 032007, [arXiv:1412.2862].

[132] J. Currie, A. Gehrmann-De Ridder, E. Glover, and J. Pires, *NNLO QCD corrections to jet production at hadron colliders from gluon scattering*, *JHEP* **1401** (2014) 110, [arXiv:1310.3993].

[133] A. Gehrmann-De Ridder, T. Gehrmann, E. Glover, and J. Pires, *Second order QCD corrections to jet production at hadron colliders: the all-gluon contribution*, *Phys.Rev.Lett.* **110** (2013) 162003, [arXiv:1301.7310].

[134] D. de Florian, P. Hinderer, A. Mukherjee, F. Ringer, and W. Vogelsang, *Approximate next-to-next-to-leading order corrections to hadronic jet production*, *Phys.Rev.Lett.* **112** (2014) 082001, [arXiv:1310.7192].

[135] N. Kidonakis and J. F. Owens, *Effects of higher-order threshold corrections in high-E(T) jet production*, *Phys. Rev.* **D63** (2001) 054019, [hep-ph/0007268].

[136] J. Campbell, K. Hatakeyama, J. Huston, F. Petriello, J. R. Andersen, et al., *Working Group Report: Quantum Chromodynamics*, arXiv:1310.5189.

[137] T. Carli, D. Clements, A. Cooper-Sarkar, C. Gwenlan, G. P. Salam, et al., *A posteriori inclusion of parton density functions in NLO QCD final-state calculations at hadron colliders: The APPLGRID Project*, *Eur.Phys.J.* **C66** (2010) 503–524, [arXiv:0911.2985].

[138] J. Campbell and R. K. Ellis, *Next-to-leading order corrections to W + 2jet and Z + 2jet production at hadron colliders*, *Phys. Rev.* **D65** (2002) 113007, [hep-ph/0202176].

[139] *Mcfm*, http://mcfm.fnal.gov.

[140] Z. Nagy, *Next-to-leading order calculation of three-jet observables in hadron hadron collision*, *Phys. Rev.* **D68** (2003) 094002, [hep-ph/0307268].

[141] T. Kluge, K. Rabbertz, and M. Wobisch, *Fast pQCD calculations for PDF fits*, hep-ph/0609285.

[142] **fastNLO** Collaboration, M. Wobisch, D. Britzger, T. Kluge, K. Rabbertz, and F. Stober, *Theory-Data Comparisons for Jet Measurements in Hadron-Induced Processes*, arXiv:1109.1310.

[143] L. Del Debbio, N. P. Hartland, and S. Schumann, *MCgrid: projecting cross section calculations on grids*, *Comput.Phys.Commun.* **185** (2014) 2115–2126, [arXiv:1312.4460].

[144] A. Buckley, J. Butterworth, L. Lonnblad, D. Grellscheid, H. Hoeth, et al., *Rivet user manual*, *Comput.Phys.Commun.* **184** (2013) 2803–2819, [arXiv:1003.0694].

[145] Z. Bern, L. Dixon, F. Febres Cordero, S. Hche, H. Ita, et al., *Ntuples for NLO Events at Hadron Colliders*, *Comput.Phys.Commun.* **185** (2014) 1443–1460, [arXiv:1310.7439].

[146] V. Bertone, R. Frederix, S. Frixione, J. Rojo, and M. Sutton, *aMCfast: automation of fast NLO computations for PDF fits*, *JHEP* **1408** (2014) 166, [arXiv:1406.7693].

[147] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, *JHEP* **1407** (2014) 079, [arXiv:1405.0301].

[148] *Dynnlo*, http://theory.fi.infn.it/grazzini/dy.html.

[149] S. Catani, G. Ferrera, and M. Grazzini, *W Boson Production at Hadron Colliders: The Lepton Charge Asymmetry in NNLO QCD*, *JHEP* **1005** (2010) 006, [arXiv:1002.3115].

[150] S. Catani and M. Grazzini, *An NNLO subtraction formalism in hadron collisions and its application to Higgs boson production at the LHC*, *Phys.Rev.Lett.* **98** (2007) 222002, [hep-ph/0703012].

[151] S. Catani, L. Cieri, G. Ferrera, D. de Florian, and M. Grazzini, *Vector boson production at hadron colliders: a fully exclusive QCD calculation at NNLO*, *Phys.Rev.Lett.* **103** (2009) 082001, [arXiv:0903.2120].

[152] R. Gavin, Y. Li, F. Petriello, and S. Quackenbush, *W Physics at the LHC with FEWZ 2.1*, *Comput.Phys.Commun.* **184** (2013) 208–214, [arXiv:1201.5896].

[153] Y. Li and F. Petriello, *Combining QCD and electroweak corrections to dilepton production in FEWZ*, *Phys.Rev.* **D86** (2012) 094034, [arXiv:1208.5967].

[154] S. Carrazza and J. Pires, *Perturbative QCD description of jet data from LHC Run-I and Tevatron Run-II*, `arXiv:1407.7031`.

[155] M. Czakon, P. Fiedler, and A. Mitov, *The total top quark pair production cross-section at hadron colliders through $O(\alpha_S^4)$*, *Phys.Rev.Lett.* **110** (2013) 252004, [`arXiv:1303.6254`].

[156] M. Czakon and A. Mitov, *Top++: A Program for the Calculation of the Top-Pair Cross-Section at Hadron Colliders*, *Comput.Phys.Commun.* **185** (2014) 2930, [`arXiv:1112.5675`].

[157] C. Carloni Calame, G. Montagna, O. Nicrosini, and A. Vicini, *Precision electroweak calculation of the production of a high transverse-momentum lepton pair at hadron colliders*, *JHEP* **0710** (2007) 109, [`arXiv:0710.1722`].

[158] F. A. Berends and R. Kleiss, *Hard Photon Effects in $W^{\pm}$ and Z^0 Decay*, *Z.Phys.* **C27** (1985) 365.

[159] F. A. Berends, R. Kleiss, J. Revol, and J. Vialle, *QED Radiative Corrections and Radiative Decays of the Intermediate Weak Bosons Produced in Proton - Anti-proton Collisions*, *Z.Phys.* **C27** (1985) 155.

[160] S. Dittmaier and M. Huber, *Radiative corrections to the neutral-current Drell-Yan process in the Standard Model and its minimal supersymmetric extension*, *JHEP* **1001** (2010) 060, [`arXiv:0911.2329`].

[161] U. Baur, S. Keller, and W. Sakumoto, *QED radiative corrections to Z boson production and the forward backward asymmetry at hadron colliders*, *Phys.Rev.* **D57** (1998) 199–215, [`hep-ph/9707301`].

[162] U. Baur, O. Brein, W. Hollik, C. Schappacher, and D. Wackeroth, *Electroweak radiative corrections to neutral current Drell-Yan processes at hadron colliders*, *Phys.Rev.* **D65** (2002) 033007, [`hep-ph/0108274`].

[163] R. Boughezal, Y. Li, and F. Petriello, *Disentangling radiative corrections using high-mass Drell-Yan at the LHC*, *Phys.Rev.* **D89** (2014) 034030, [`arXiv:1312.3972`].

[164] S. Forte, E. Laenen, P. Nason, and J. Rojo, *Heavy quarks in deep-inelastic scattering*, *Nucl. Phys.* **B834** (2010) 116–162, [`arXiv:1001.2312`].

[165] J. Ablinger, A. Behring, J. Blümlein, A. De Freitas, A. von Manteuffel, et al., *The 3-Loop Pure Singlet Heavy Flavor Contributions to the Structure Function $F_2(x, Q^2)$ and the Anomalous Dimension*, `arXiv:1409.1135`.

[166] F. Demartin, S. Forte, E. Mariani, J. Rojo, and A. Vicini, *The impact of PDF and α_s uncertainties on Higgs Production in gluon fusion at hadron colliders*, *Phys. Rev.* **D82** (2010) 014002, [`arXiv:1004.0962`].

[167] F. Cascioli, P. Maierhoefer, N. Moretti, S. Pozzorini, and F. Siegert, *NLO matching for ttbb production with massive b-quarks*, `arXiv:1309.5912`.

[168] **The NNPDF** Collaboration, R. D. Ball et al., *Theoretical issues in PDF determination and associated uncertainties*, *Phys.Lett.* **B723** (2013) 330–339, [`arXiv:1303.1189`].

[169] G. Watt and R. Thorne, *Study of Monte Carlo approach to experimental uncertainty propagation with MSTW 2008 PDFs*, *JHEP* **1208** (2012) 052, [`arXiv:1205.4024`].

[170] D. J. Montana and L. Davis, *Training Feedforward Neural Networks Using Genetic Algorithms*, in *Proceedings of the 11th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'89, (San Francisco, CA, USA), pp. 762–767, Morgan Kaufmann Publishers Inc., 1989.

[171] J. Schmidhuber, *Deep learning in neural networks: An overview*, *CoRR* **abs/1404.7828** (2014).

[172] **The NNPDF** Collaboration, R. D. Ball et al., *Unbiased determination of polarized parton distributions and their uncertainties*, *Nucl.Phys.* **B874** (2013) 36–84, [arXiv:1303.7236].

[173] **NNPDF Collaboration** Collaboration, E. R. Nocera, R. D. Ball, S. Forte, G. Ridolfi, and J. Rojo, *A first unbiased global determination of polarized PDFs and their uncertainties*, *Nucl.Phys.* **B887** (2014) 276–308, [arXiv:1406.5539].

[174] D. Mackay, *Bayesian Interpolation*, *Neural Comp.* **4** (1992) 415–447.

[175] D. Mackay, *Probable networks and plausible predictions ? a review of practical Bayesian methods for supervised neural networks*, *Network-Comp.Neural.* **6** (1995) 469–505.

[176] G. Altarelli, R. D. Ball, and S. Forte, *Small x Resummation with Quarks: Deep-Inelastic Scattering*, *Nucl. Phys.* **B799** (2008) 199–240, [arXiv:0802.0032].

[177] L. Demortier, *Proceedings, PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN,Geneva, Switzerland 17-20 January 2011*, ch. Open Issues in the Wake of Banff 2011. 2011.

[178] N. P. Hartland and C. S. Deans, *Towards closure testing of parton determinations*, *PoS* **DIS2013** (2013) 043, [arXiv:1307.2046].

[179] D. Bourilkov, R. C. Group, and M. R. Whalley, *LHAPDF: PDF use from the Tevatron to the LHC*, hep-ph/0605240.

[180] **NNPDF Collaboration** Collaboration, R. D. Ball et al., *Parton Distributions: Determining Probabilities in a Space of Functions*, arXiv:1110.1863.

[181] J. Pumplin, *Parametrization dependence and $\Delta\chi^2$ in parton distribution fitting*, *Phys.Rev.* **D82** (2010) 114020, [arXiv:0909.5176].

[182] C. Anastasiou, L. J. Dixon, K. Melnikov, and F. Petriello, *High precision QCD at hadron colliders: Electroweak gauge boson rapidity distributions at NNLO*, *Phys. Rev.* **D69** (2004) 094008, [hep-ph/0312266].

[183] M. Bonvini, R. D. Ball, S. Forte, S. Marzani, and G. Ridolfi, *Updated Higgs cross section at approximate N^3LO*, *J.Phys.* **G41** (2014) 095002, [arXiv:1404.3204].

[184] P. Skands, S. Carrazza, and J. Rojo, *Tuning PYTHIA 8.1: the Monash 2013 Tune*, *European Physical Journal* **74** (2014) 3024, [arXiv:1404.5630].

[185] J. M. Campbell, J. W. Huston, and W. J. Stirling, *Hard interactions of quarks and gluons: A primer for LHC physics*, *Rept. Prog. Phys.* **70** (2007) 89, [hep-ph/0611148].

[186] F. Caola, S. Forte, and J. Rojo, *HERA data and DGLAP evolution: Theory and phenomenology*, *Nucl.Phys.* **A854** (2011) 32–44, [arXiv:1007.5405].

[187] A. D. Martin, A. J. Mathijsseen, W. J. Stirling, R. S. Thorne, B. J. A. Watt, and G. Watt, *Extended Parameterisations for MSTW PDFs and their effect on Lepton Charge Asymmetry from W Decays*, arXiv:1211.1215.

[188] S. J. Brodsky, P. Hoyer, C. Peterson, and N. Sakai, *The Intrinsic Charm of the Proton*, *Phys. Lett.* **B93** (1980) 451–455.

[189] L. Barze, G. Montagna, P. Nason, O. Nicrosini, F. Piccinini, et al., *Neutral current Drell-Yan with combined QCD and electroweak corrections in the POWHEG BOX*, *Eur.Phys.J.* **C73** (2013), no. 6 2474, [arXiv:1302.4606].

[190] A. Martin, R. Roberts, W. Stirling, and R. Thorne, *MRST partons and uncertainties*, hep-ph/0307262.

[191] **The NNPDF** Collaboration, R. D. Ball et al., *Reweighting NNPDFs: the W lepton asymmetry*, *Nucl. Phys.* **B849** (2011) 112–143, [arXiv:1012.0836].

[192] R. D. Ball, V. Bertone, F. Cerutti, L. Del Debbio, S. Forte, et al., *Reweighting and Unweighting of Parton Distributions and the LHC W lepton asymmetry data*, *Nucl.Phys.* **B855** (2012) 608–638, [arXiv:1108.1758].

[193] S. Forte, L. Garrido, J. I. Latorre, and A. Piccione, *Neural network parametrization of deep-inelastic structure functions*, *JHEP* **05** (2002) 062, [hep-ph/0204232].

[194] S. Forte and J. Rojo, "Section II.2 in: J. Butterworth et al., "Les Houches 2013: Physics at Tev Colliders: Standard Model Working Group Report"." arXiv:1405.1067, 2014.

[195] D. Mason et al., *Measurement of the Nucleon Strange-Antistrange Asymmetry at Next-to-Leading Order in QCD from NuTeV Dimuon Data*, *Phys. Rev. Lett.* **99** (2007) 192001.

[196] **NOMAD** Collaboration, O. Samoylov et al., *A Precision Measurement of Charm Dimuon Production in Neutrino Interactions from the NOMAD Experiment*, *Nucl.Phys.* **B876** (2013) 339–375, [arXiv:1308.4750].

[197] **ATLAS** Collaboration, G. Aad et al., *Determination of the strange quark density of the proton from ATLAS measurements of the W, Z cross sections*, *Phys.Rev.Lett.* (2012) [arXiv:1203.4051].

[198] A. Glazov, "Private communication, on behalf of the H1-ZEUS combination and ATLAS.".

[199] M. Gouzevitch, "Private communication, on behalf of CMS.".

[200] J. Rojo, *Parton distributions based on a maximally consistent dataset*, arXiv:1409.3029.

[201] M. Cacciari, G. P. Salam, and G. Soyez, *The Anti-k(t) jet clustering algorithm*, *JHEP* **0804** (2008) 063, [arXiv:0802.1189].

[202] S. Frixione, *Isolated photons in perturbative QCD*, *Phys.Lett.* **B429** (1998) 369–374, [hep-ph/9801442].

[203] R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, R. Pittau, et al., *Four-lepton production at hadron colliders: aMC@NLO predictions with theoretical uncertainties*, *JHEP* **1202** (2012) 099, [arXiv:1110.4738].

[204] M. Czakon, M. L. Mangano, A. Mitov, and J. Rojo, *Constraints on the gluon PDF from top quark pair production at hadron colliders*, *JHEP* **1307** (2013) 167, [arXiv:1303.7215].

[205] H.-L. Lai et al., *Uncertainty induced by QCD coupling in the CTEQ global analysis of parton distributions*, *Phys. Rev.* **D82** (2010) 054021, [arXiv:1004.4624].

[206] M. Kramer, A. Kulesza, R. van der Leeuw, M. Mangano, S. Padhi, et al., *Supersymmetry production cross sections in pp collisions at $\sqrt{s} = 7$ TeV*, arXiv:1206.2892.

[207] C. Borschensky, M. Krmer, A. Kulesza, M. Mangano, S. Padhi, et al., *Squark and gluino production cross sections in pp collisions at $\sqrt{s}$ = 13, 14, 33 and 100 TeV*, arXiv:1407.5066.

[208] W. Beenakker, R. Hopker, M. Spira, and P. Zerwas, *Squark and gluino production at hadron colliders*, *Nucl.Phys.* **B492** (1997) 51–103, [`hep-ph/9610490`].

[209] T. Plehn. Publicly available from `http://www.thphys.uni-heidelberg.de/~plehn`.

Publications

R. D. Ball, V. Bertone, S. Carrazza, C. S. Deans, L. Del Debbio, et al., *Parton distributions with LHC data*, *Nucl.Phys.* **B867** (2013) 244–289, [arXiv:1207.1303].

C. S. Deans, *Progress in the NNPDF global analysis*, *Proceedings of Moriond QCD 2013* (2013) 353-356, [arXiv:1304.2781].

N. P. Hartland and C. S. Deans, *Towards closure testing of parton determinations*, *Proceeding of DIS2013* (2013) 043, [arXiv:1307.2046].

C. S. Deans, *Closure testing the NNPDF methodology*, *Proceedings of QCD14* (2014) 15-18, [arXiv:1409.4283].

R. D. Ball, V. Bertone, S. Carrazza, C. S. Deans, L. Del Debbio, et al., *Parton distributions for the LHC Run II*, *JHEP* **1504** (2015) 040, [arXiv:1410.8849].

C. S. Deans, *Closure testing NNPDF3.0 with LHC observables*, *Proceedings of DIS2015* (2015), [arXiv:1506.07357].