Chapter II

Parton distribution functions

1 Theoretical uncertainties and dataset dependence of parton distributions [1]

We study theoretical uncertainties on parton distributions (PDFs) due to missing higher order (MHO) corrections by determining the change of PDFs when going from next-to-leading (NLO) to next-to-next-to-leading order (NNLO) theory. Based on the NNPDF3.1 framework, we compare PDF determinations obtained from different datasets, specifically a global, a proton-only, and a collider-only dataset. We show that PDF determinations obtained from a wider input dataset exhibit greater perturbative stability, and thus are likely to be affected by smaller theoretical uncertainties from MHOs. We also show that the effect of including deuterium nuclear corrections is smaller than that of excluding the deuterium data altogether.

1.1 Parton distribution uncertainties

The accurate determination of the uncertainties on parton distribution functions (PDFs) of the proton [402] has been the main challenge in PDF studies at the LHC. For instance, PDFs represent one of the main sources of uncertainty in Higgs physics [236, 594]; they also have a significant impact on searches of physics beyond the standard model (see e.g. [595]), and on precision measurements such as the determination of the W boson mass [596]. Uncertainties on the current combined PDF4LHC15 PDF set [597] are typically of order 3-5% in the region covered by data, but the more recent NNPDF3.1 PDF set [598], which includes a wide array of LHC data, has PDF uncertainties typically between 1% and 3% in the data region (not including, in either case, the uncertainty on α_s).

This uncertainty — indeed, what is usually referred to as "PDF uncertainty" — includes the propagated uncertainty on the data used for PDF determinations, as well as further uncertainties due to the fitting methodology, but it does not include any theory error. Specifically, the current PDF uncertainty does not include a contribution accounting for the fact that PDFs are determined using fixed-order perturbative QCD, and thus surely one has to account for a missing higher order uncertainty (MHOU). As the uncertainty due to the data and methodology keeps decreasing, this theory error is bound to stop being negligible, and eventually become dominant.

So far, there has been a broad consensus that the use of the widest possible dataset for PDF determination — leading to so-called global PDF fits — is advantageous. Indeed, on the one hand, more data contain more information and thus allow for the accurate determination of the widest set of PDF in the most extended kinematic region. On the other hand, the use of multiple datasets provides a cross-check on both the theory and methodology. One may however ask whether the use of a wider dataset is also advantageous — or indeed not — in terms of theoretical uncertainties, specifically the MHOUs. In particular, it is be important to understand if MHOUs are likely to be smaller with a global dataset or with a reduced dataset.

---

[1] S. Forte, Z. Kassabov, J. Rojo, L. Rottoli

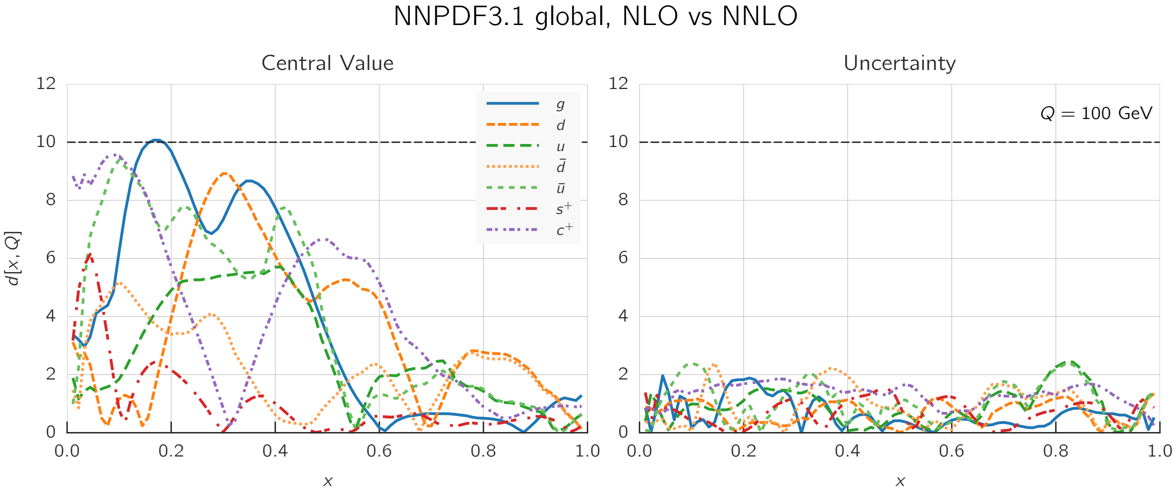

Fig. II.1: Distance between the central values (left) and the uncertainties (right) of NLO and NNLO global PDFs.

1.2 Missing higher order corrections and dataset dependence

1.2.1 Comparing NLO and NNLO PDFs

Quite in general, there is currently no way to reliably estimate the MHOUs. However, when several perturbative orders are known, we may at least study the behaviour of the perturbative expansion. Assuming reasonable convergence, the shift between, say, known NLO and NNLO results then provides a reasonable estimate of the MHOU on the NLO result.

We will thus address the problem of MHOU on PDFs by studying the way PDF change from NLO to NNLO. To this purpose, we have produced PDF determinations based essentially on the same dataset as in the NNPDF3.1 global analysis [598]. The only difference is that, while in NNPDF3.1 some jet data for which NNLO corrections were not yet available were treated approximately, here we only include both at NLO and NNLO jet data for which exact NNLO theory is available (see Sect. 4.4 of Ref. [598]), and using the exact NNLO corrections [129]. This ensures perturbative consistency.

In Fig. II.1 we show the distance between PDFs determined at NLO and at NNLO. Recall that the distance d is defined as the difference in units of standard deviation, normalized so that $d = 10$ corresponds to an one-σ shift (see Ref. [598] and references therein for a more detailed discussion). It is clear from Fig. II.1 that, whereas no distance between central values is greater than one-σ in the data region, several distances (in particular for the gluon and light quarks) are of order one-σ. But a one-σ distance means that NLO-NNLO shift is the same size as the PDF uncertainty, so we conclude that, at NLO, MHOUs on central values are comparable to PDF uncertainties. Distances between uncertainties are instead of order one, *i.e.* comparable to a statistical fluctuation. This means that PDF uncertainties at NLO and NNLO do not differ by a statistically significant amount, consistent with the expectation that they reflect the uncertainty on the data and methodology, and thus, by and large, do not systematically depend on the perturbative order.

1.2.2 Dataset dependence

In order to study the issue that we set out in the introduction, namely, the dataset dependence of the MHOUs associated to a fit of parton distributions, we have repeated the previous PDF

	NLO	NNLO	NLO/NNLO	Δ
global	1.279	1.253	1.02	1.16
proton	1.248	1.193	1.05	1.97
collider	1.181	1.114	1.06	2.07

Table II.1: Value of χ^2 per data point for the global, proton-only, and collider-only fits at NLO and NNLO. The NLO/NNLO ratio, and difference normalized to the standard deviation (see text) are also given. Note that the total number of data points $N_{\rm dat}$ in each of the three fits is different.

determination now based on two reduced datasets. First, we have produced a proton-only determination, in which we excluded all data with nuclear and deuterium targets (specifically fixed-target deep-inelastic and Drell-Yan production). Then, we have produced a collider-only determination, in which we have excluded all fixed-target DIS and DY data altogether.

These PDF determinations from smaller datasets were already discussed in Ref. [598] (as well as in previous NNPDF studies [422,599]) where it was argued that, even though these smaller datasets are in principle more consistent, the increased theoretical reliability does not make up for the great loss in accuracy: the PDF uncertainty increases monotonically when reducing the dataset, and there is no evidence of inconsistency between the data entering the global fit. With this motivation, the use of the more global dataset for the baseline PDF determination was advocated. These two PDF determinations have now been redone starting from the global dataset described in Sec. 1.2.1.

The values of the total χ^2 per data point for these NNPDF3.1-based PDF determinations are collected in Table II.1. In each case, we show the χ^2 per data point at NLO and NNLO, their ratio, and the difference $\Delta = \frac{\chi^2_{\rm NLO}-\chi^2_{\rm NNLO}}{\sqrt{2N_{\rm dat}}}$ which is a measure of the improvement of the χ^2 in units of its standard deviation. Note that the results in Table II.1 only consider the dataset that was included in the fit in each case, and consequently the total number of data points $N_{\rm dat}$ in each of the three fits is different. Clearly, the PDF fits based on smaller datasets lead to a better χ^2, due to the greater consistency of the dataset. Interestingly, however, the deterioration of the total χ^2 from NNLO to NLO, as measured both by the χ^2 ratios, and the difference in units of the standard deviation, is more severe for the fits based on a smaller datasets, and thus largest in the case of the collider-only fit. This provides a first indication that the use of a wider dataset may lead to greater perturbative stability.

To investigate this issue further, we have computed again the distance between the NLO and NNLO fits, as shown in Fig. II.1, but now for the proton-only and collider-only PDF sets. Results are shown in Fig. II.2. It is clear that while distances which were already sizable in the global fit (specifically for the gluon and down quark) are still big, now also the light quark PDFs (in particular also up and anti-up) display sizable NLO-to-NNLO shifts.

In order to achieve a fully quantitative comparison, in Fig. II.3 we display the shift of PDF central values between the NLO and NNLO fits, normalized to the NLO, for the gluon and light quarks, comparing the three PDF determinations. In order to facilitate visualization, the shifts are symmetrized about the x axis. It is clear that while for the gluon the shift is of similar size (and quite small) in the three PDF determinations, for the quarks there is a uniform hierarchy: the smallest dataset, *i.e.* the collider-only PDF set, nearly always displays the largest shifts, with very few localized exceptions.

A simple explanation of the greater perturbative stability of the more global fit seen in Table II.1 could be that hadron collider processes, which have larger perturbative corrections than deep-inelastic scattering, carry a greater weight in the fits to a reduced dataset: the global fit improves less because it is less consistent. But in this case, we would expect the global fit,

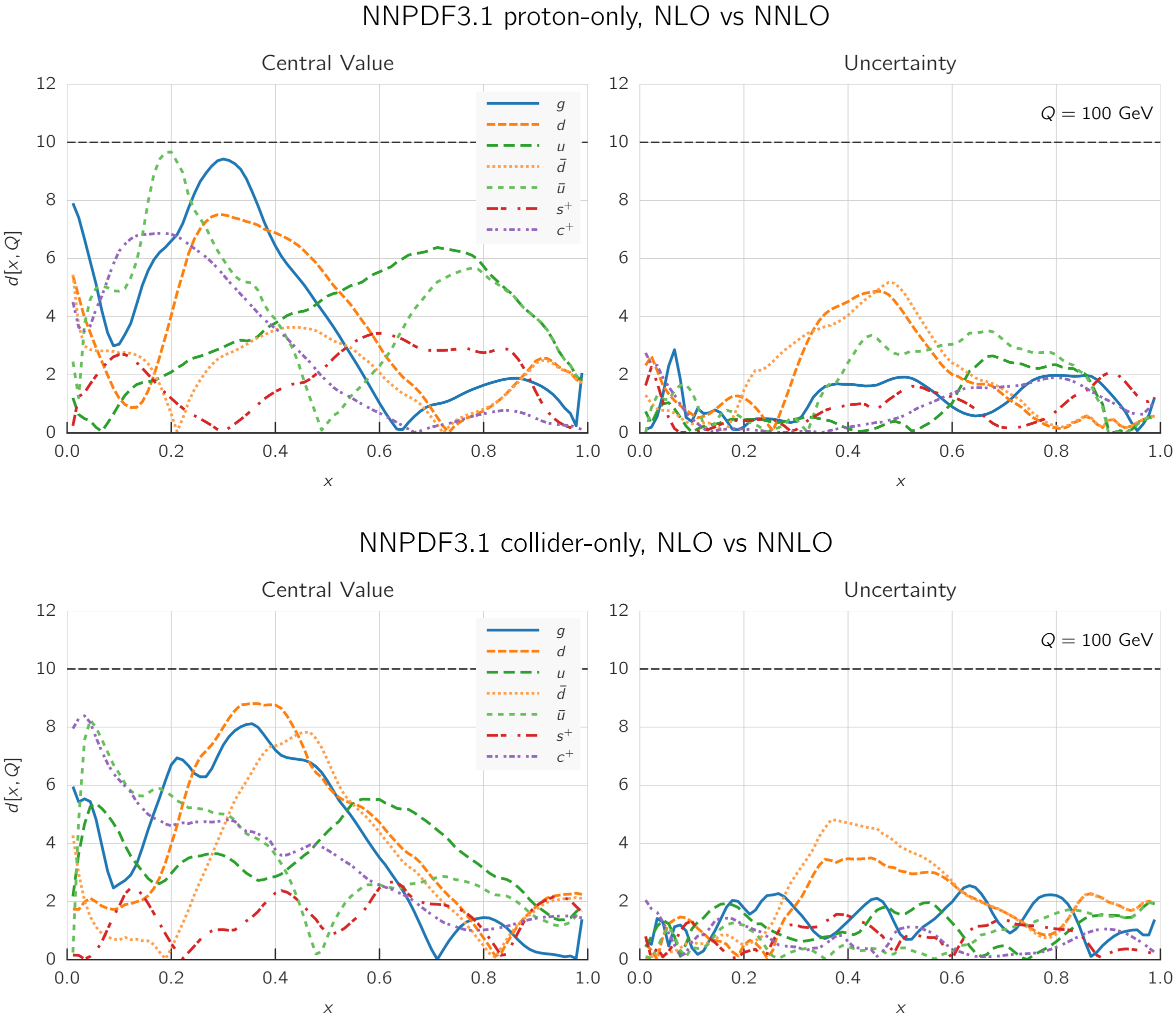

Fig. II.2: Same as Fig. II.1, but for the proton-only (top) and collider-only (bottom) PDF determinations.

due to its poorer consistency, not to show a significant improvement in PDF uncertainties, or to display a sizable change in results when going from NLO to NNLO. Instead, the opposite is the case. As extensively discussed in Ref. [598] the more global fit has significantly smaller PDF uncertainty, and as shown in Fig. II.3 it also changes less from NLO to NNLO. Hence the global fit is both less uncertain and more perturbatively stable.

An alternative explanation of the observed perturbative stability then seems more likely. Namely, that it is a consequence of the fact that missing higher order terms for different processes distort PDFs randomly by pulling them in different directions. Therefore, in a more global dataset in which the same PDF combination is determined from constraints by different processes, these uncertainties tend to average out.

1.3 Deuterium nuclear corrections

Perhaps the main advantage of the PDF fits based on smaller datasets is their greater consistency not only from the experimental, but also from the theoretical point of view. In particular, proton-only PDFs do not make use of any data that are affected by the poorly known nuclear corrections. It is then interesting to ask how the size of nuclear corrections compares to the uncertainties

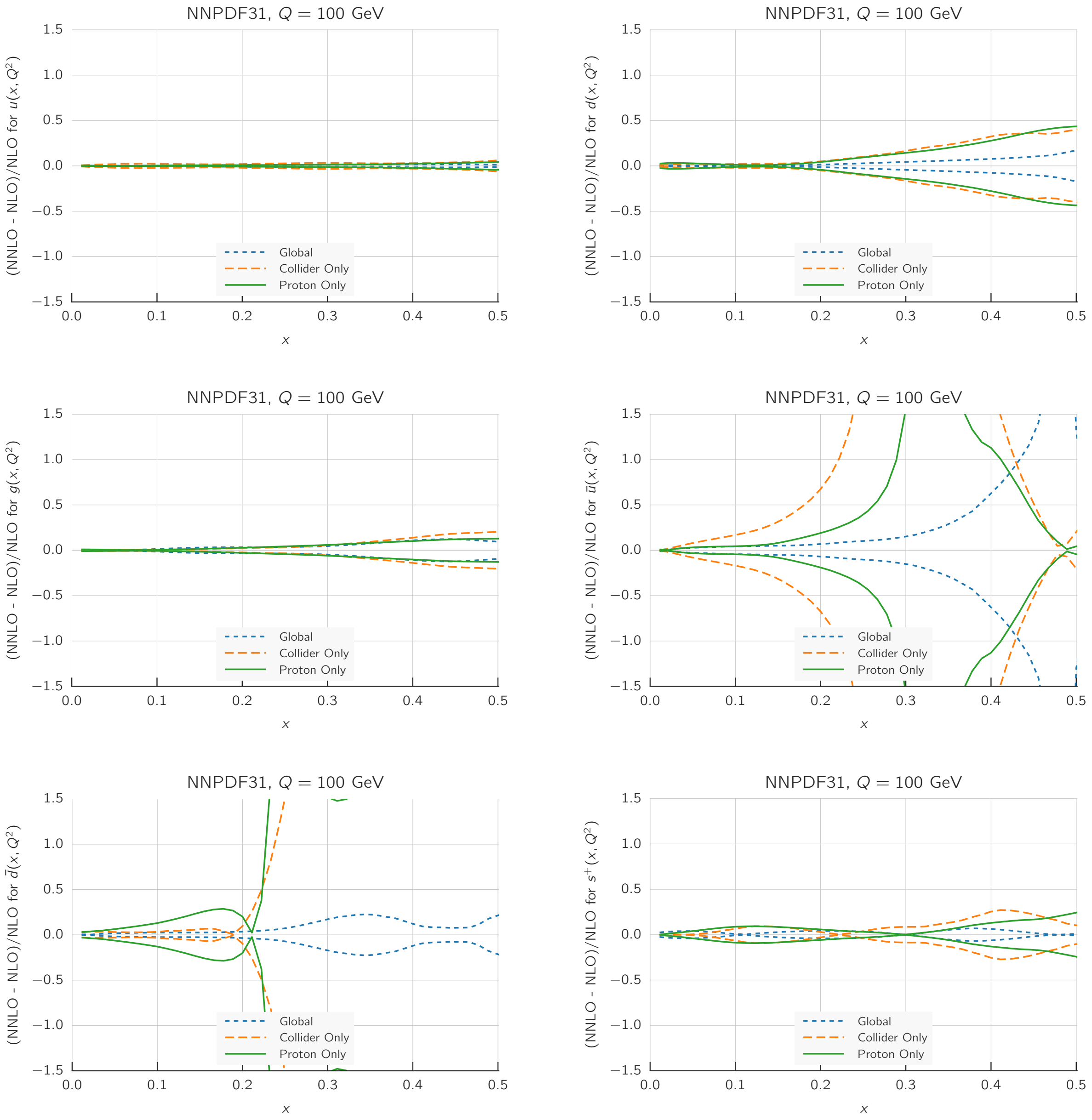

Fig. II.3: The relative shift between the central values of the NLO to NNLO PDFs normalized to the NLO, for the global, collider-only, and proton-only PDF determinations. Results are shown for, from left to right and from top to bottom, the gluon, up, down, anti-up, anti-down and total strange+antistrange PDFs. To facilitate visualization, the shifts are symmetrized about the x axis.

that we have discussed so far. Whereas existing determinations of heavy nuclear corrections are affected by large uncertainties [600], we may at least compare PDF determinations in which a model of nuclear effects for deuterium is included.

In order to isolate this effect, we have thus produced a PDF determination in which data using heavier nuclear targets have been excluded (hence specifically neutrino deep-inelastic scattering data) but data with deuterium targets (both deep-inelastic and fixed-target Drell-Yan) are kept. We then compare this fit either to the proton-only fit, or to itself with deuterium corrections included using the MMHT14 best-fit model of Ref. [601] (with default settings).

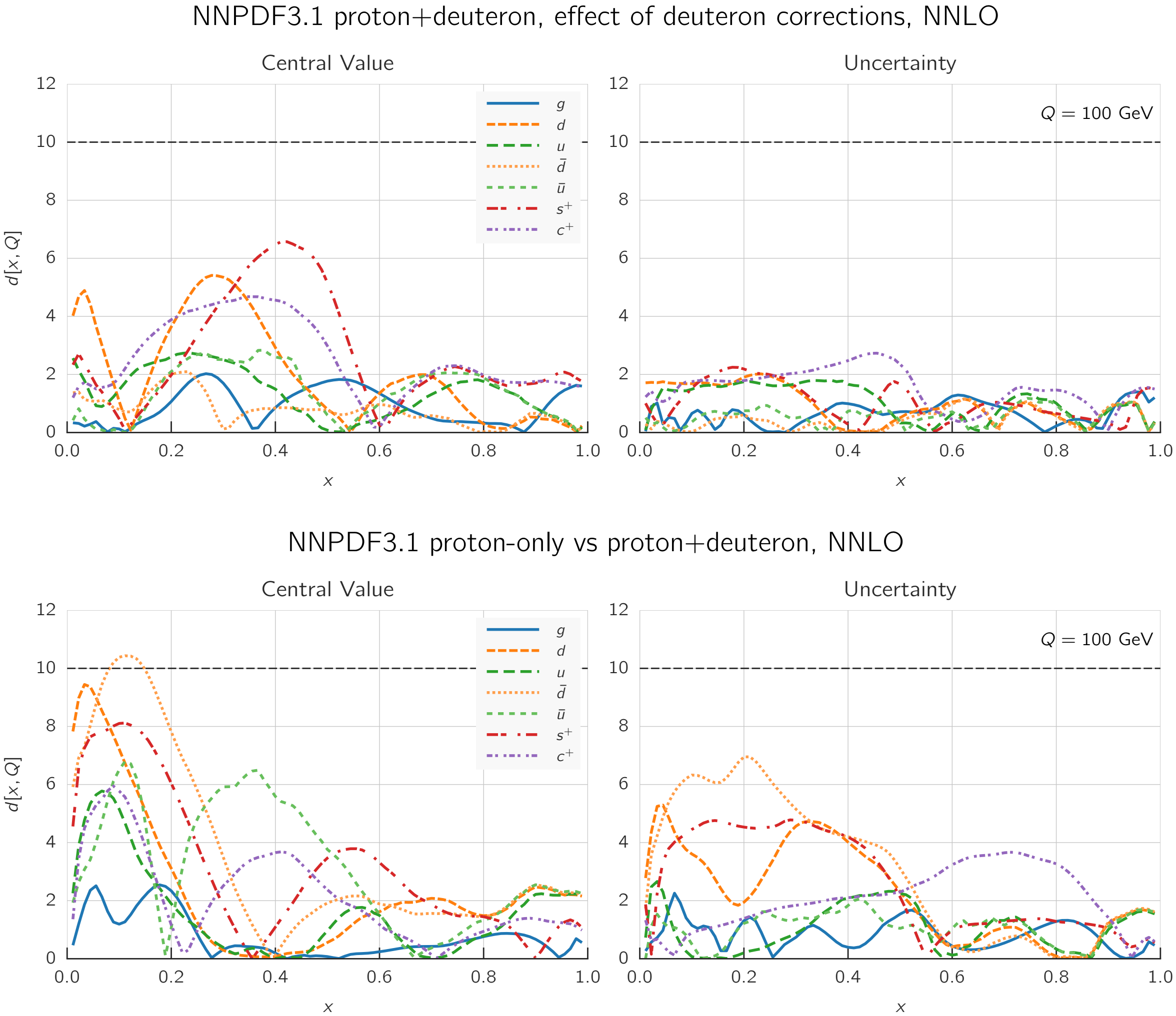

Fig. II.4: Distances between a fit with proton and deuterium target data and the proton-only fit (bottom) compared to the distances between a fit with proton and deuterium target data with and without linear corrections.

The corresponding distances between these fits are shown in Fig. II.4. It is clear that the effect of including the deuterium nuclear corrections is rather smaller than that of the deuterium data itself — whose impact is instead comparable to that of the NNLO corrections. We can thus conclude that the inclusion of deuterium data in the global PDF fit appears to be currently advantageous, in the sense that the impact of this data on the PDFs is greater than the likely size of their uncertainty due to missing nuclear corrections. On the other hand, if these data are included, some estimate of the associated theoretical uncertainty, arising both from deuterium nuclear corrections and from MHOs, should be performed, in view of the small size of the ensuing PDF uncertainty.

1.4 Conclusions

We have demonstrated that PDF determinations based on a wider dataset are characterized by greater perturbative stability, and thus are most likely to exhibit smaller theoretical uncertainties related to missing higher orders. In addition, we have shown that the inclusion of data taken with nuclear target, and specifically with deuterium ones, appears to be advantageous at present, in that the uncertainty related to the modeling of deuterium nuclear corrections appears to be rather smaller than both the MHOU and of the impact of this data on PDF uncertainties.

These results provide further evidence that the inclusion of theoretical uncertainties in PDF uncertainties, specifically those related to missing higher orders, but also to nuclear corrections, is now one of the highest priorities. An interesting observation in this respect is that a possible way to approach the determination of MHOU on PDFs might be to study the way they vary between different sets of experimental measurements. Indeed, our study suggests that these uncertainties tend to compensate when combining different classes of datasets. This, in turn, hints at the fact that the size of the MHOUs might be estimated by performing dataset variations and studying the ensuing distribution of the best-fit PDFs. These topics are the subject of ongoing and forthcoming investigations.

Acknowledgements

Stefano Forte thanks Daniel Maitre for interesting discussions on PDF uncertainties during the workshop. Stefano Forte is supported by the European Research Council under the European Union's Horizon 2020 research and innovation Programme (grant agreement ERC-AdG-740006). Juan Rojo and Luca Rottoli are supported by an European Research Council Starting Grant "PDF4BSM". The research of Juan Rojo is also partially supported by the Netherlands Organization for Scientific Research (NWO).