Chapter II

Parton distribution functions

1 Construction and phenomenological applications of PDF4LHC parton distributions [1]

We revisit the construction and application of combined PDF sets (PDF4LHC15) developed by the PDF4LHC group in 2015. Our focus is on the meta-analysis technique employed in the construction of the 30-member PDF4LHC15 sets, and especially on aspects that were not fully described in the main PDF4LHC recommendation document. These aspects include construction of the 30-member sets at NLO (in addition to NNLO), extension of the NLO and NNLO sets to low QCD scales, and construction of such sets for 4 active flavors. In addition, we clarify a point regarding the calculation of parton luminosity uncertainties at low mass. Finally, we present a website containing predictions based on PDF4LHC15 PDFs for some crucial LHC processes.

1.1 Introduction

To simplify applications of parton distribution functions (PDFs) in several categories of LHC experimental simulations, the 2015 recommendations [181] of the PDF4LHC working group introduce combinations of CT14 [182], MMHT2014 [183], and NNPDF3.0 [184] PDF ensembles, by utilizing the Monte Carlo (MC) replica technique [185]. The central PDF and the uncertainties of the combined set are derived from the 900 MC replicas of the error PDFs of the above three input ensembles. As the 900 error PDFs are often too many to be manageable, they are "compressed" into smaller PDF error sets using three reduction techniques [186–188]. Consequently, the final combined PDFs come in three versions, one with 30 error sets (PDF4LHC15_30), and the other two with 100 error sets (PDF4LHC15_100 and PDF4LHC15_MC). Two of these, PDF4LHC15_30 and PDF4LHC_100, are constructed in the form of Hessian eigenvector sets [189]. The PDF4LHC15_MC ensemble is constructed from MC replicas. The central sets are the same in the 900-replica prior as well as in the _100, _30, and _MC ensembles. They are equal to the average of central sets of CT14, MMHT2014, and NNPDF3.0 ensembles. The error sets of the three PDF4LHC15 ensembles are different, reflecting the specifics of each reduction technique. They are available in the LHAPDF library [190] at NLO and NNLO in QCD coupling strength α_s, with the central value of $\alpha_s(M_Z)$ equal to 0.118, and with additional sets corresponding to the α_s variations by 0.0015 around the central value.

The 30-member ensemble is constructed using the meta-parametrization technique introduced in [186]. This contribution describes additional developments in the 30-member ensemble that happened at the time, or immediately after, the release of the original PDF4LHC recommendation document. They include construction of the PDF4LHC15_30 ensemble at NLO, extension of PDF4LHC15_30 to scales below 8 GeV, and the specialized ensemble with 4 active quark flavors. These features are already incorporated in the LHAPDF distributions. We provide comparisons of PDFs and parton luminosities and introduce a website [191] illustrating essential LHC cross sections computed with the PDF4LHC15 and other ensembles, and using a variety of QCD programs.

When deciding on which of the three PDF4LHC sets to use, it is important to keep in mind that all of them reproduce well the uncertainties of the 900-replica "prior" PDF ensemble. This

[1] J. Gao, T.-J. Hou, J. Huston, P. M. Nadolsky, B. T. Wang, K. P. Xie

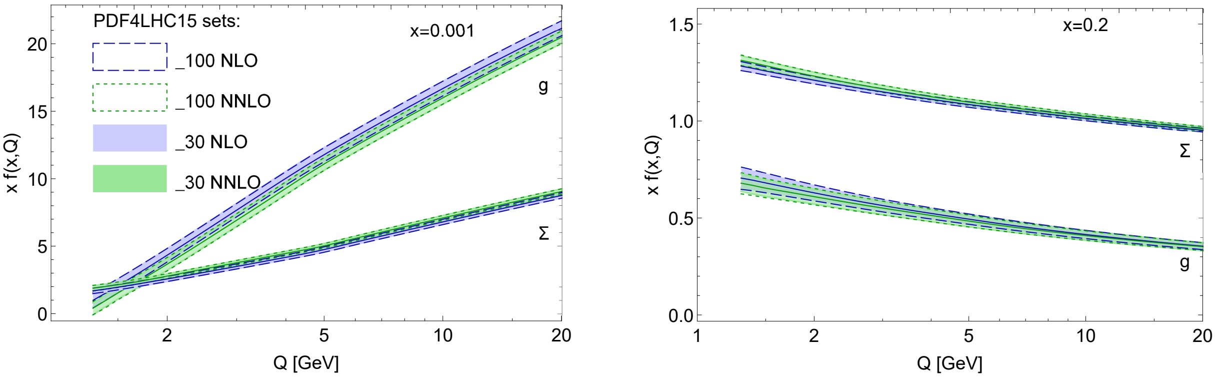

Fig. II.1: The singlet and gluon PDFs, $\Sigma(x,Q)$ and $g(x,Q)$, from 100- and 30-member PDF4LHC15 sets at NLO and NNLO, plotted vs. the QCD scale Q at $x = 10^{-3}$ (left) and 0.2 (right).

prior itself has some uncertainty both in its central value and especially in the size of the PDF uncertainty itself, reflecting differences between the central values and the uncertainty bands of CT14, MMHT2014 and NNPDF3.0, which become especially pronounced at very low x and high x. At moderate x values, contributing to the bulk of precision physics cross sections at the LHC, the agreement between the three input PDF sets is often quite better, meaning that the combined prior and the three reduced ensembles constructed from it are also known well. In general, the 30-member ensemble keeps the lowest, best-known eigenvector sets, and thus provides a slightly lower estimate for the uncertainty of the 900-replica prior, but one that is known with higher confidence than the exact uncertainty of the prior set. We will demonstrate that, across many practical applications, the 30-member error estimates are typically close both to those of the prior and of the Hessian 100 PDF error set.

1.2 QCD scale dependence of the 30-member NLO PDF4LHC ensemble

The NLO meta-parametrizations are constructed in a slightly different manner compared to the NNLO version. In Ref. [186], we have shown that the differences of the numerical implementation of DGLAP evolution at NNLO in CT10 [192], MSTW2008 [193], and NNPDF2.3 [194] PDFs are negligible compared to the intrinsic PDF uncertainties.[2] However, at NLO, the NNPDF2.3 group uses evolution that neglects some higher-order terms compared to HOPPET, which can result in deviations by up to 1 % in the small- and large-x regions, compared to the evolution used by CT10 and MSTW2008. These differences in NLO numerical DGLAP evolution, while formally allowed, also affect the most recent generation of NLO PDFs, i.e., CT14+MMHT2014 vs. NNPDF3.0. When the 900-replica prior ensemble at NLO is constructed by taking 300 replicas from each of the input CT14, MMHT14, and NNPDF3.0 ensembles, the implication is that Q-scale dependence of these replicas is not strictly Markovian. Probability regions at the low Q scale, as sampled by the MC replicas, are not exactly preserved by DGLAP evolution to a higher Q scale. This is in contrast to the consistent DGLAP evolution of a single input PDF set, which guarantees that the probability/confidence value associated with a given error set is independent of the Q scale.

Thus, the NLO prior ensemble is not inherently consistent, even though the deviations in DGLAP evolution of individual replicas are arguably small. One should apply a correction to restore the Markovian nature of the evolution. In the PDF4LHC15_30 NLO set we do this by first constructing the central PDF set at any Q by averaging the CT14, MMHT14, and

[2] CT10 PDFs use the x-space evolution provided by the program HOPPET [195].

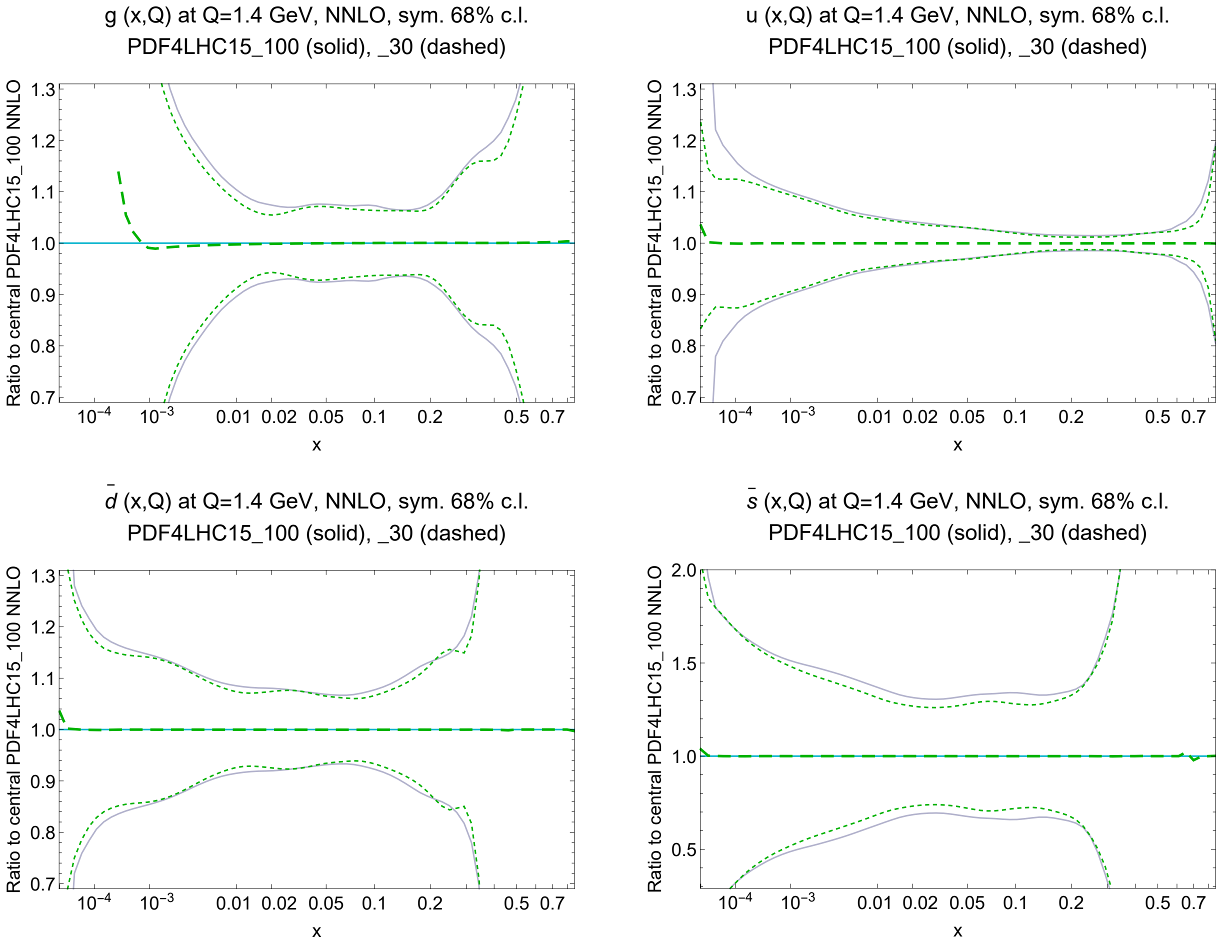

Fig. II.2: PDF central predictions and uncertainty bands for select parton flavors from the 100- and 30-member NNLO PDF4LHC15 ensembles, plotted versus x at a QCD scale $Q = 1.4$ GeV as ratios to the central PDF4LHC15_100 distributions.

NNPDF3.0 central sets that were evolved by their own native programs. Then, we reduce the 900-member into the 30-member set at scale $Q_0 = 8$ GeV and evolve all replicas to other Q values using HOPPET. Finally, we estimate the difference between the HOPPET evolution and native evolution of the central set, and subtract this difference at every Q from the HOPPET-evolved values of every error set. After such universal shift, the Q dependence of all error sets is practically the same as the native evolution of the central PDF. The probability regions are now independent of Q; this preserves sum rules for momentum and quark quantum numbers.

1.3 PDF4LHC15_30 PDFs at low Q

The original formulation of the meta-PDFs had a minimum Q value of 8 GeV. The relatively high lower cutoff on Q was introduced to justify the combination of PDFs obtained in different heavy-quark schemes, and it is sufficient to describe all high-Q^2 physics at the LHC. However, the extension of the _30 PDFs down to lower Q values can be useful, too as for example in the simulatation of parton showers and the underlying event in Monte-Carlo showering programs. The PDF4LHC15_30 version on LHAPDF includes such an extension down to a Q value of 1.4 GeV, obtained by backward evolution from 8 GeV using HOPPET. It should be remembered that the PDF4LHC15 combination is statistically consistent when the factorization scale in the PDFs is much higher than the bottom mass, as is typical in the bulk of LHC applications. The

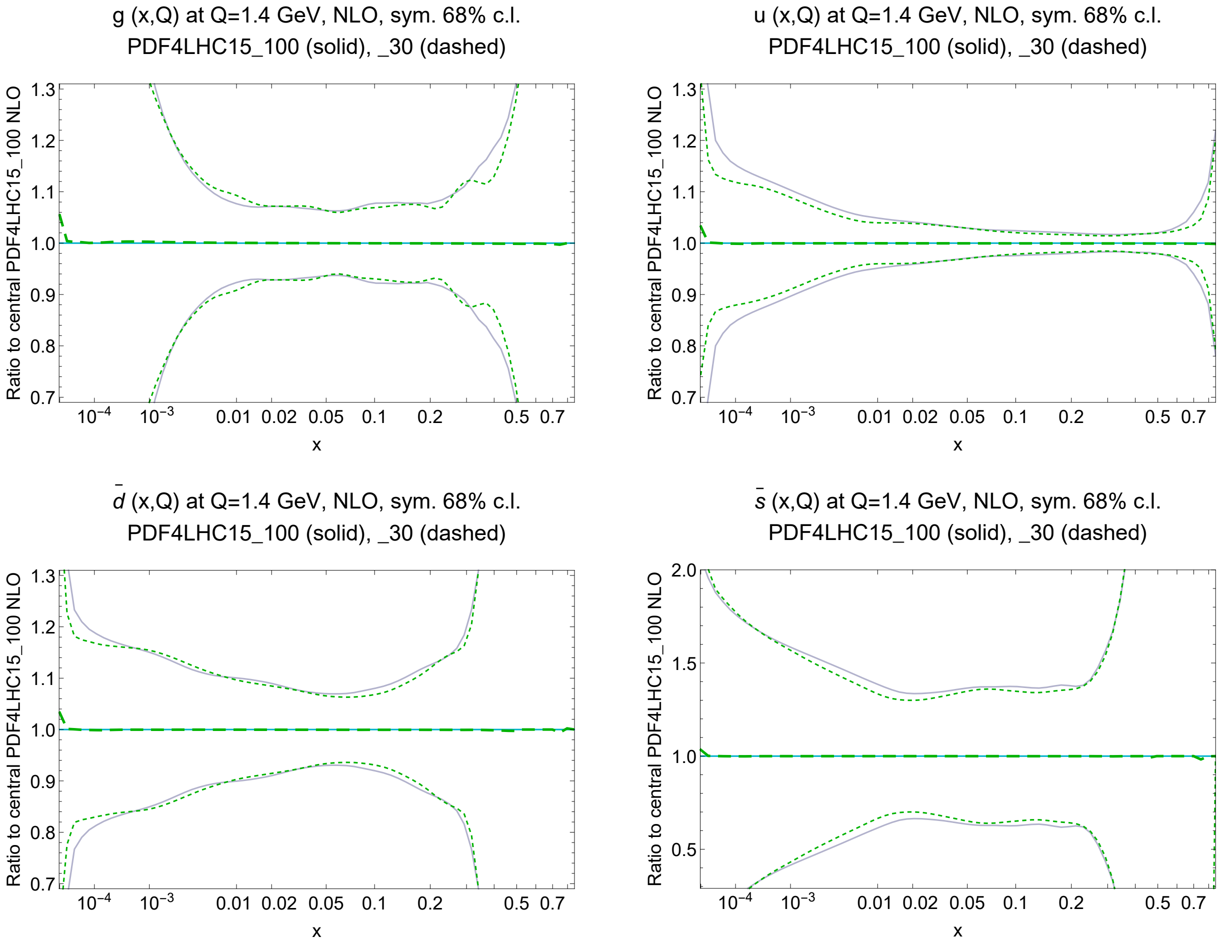

Fig. II.3: Same as Fig. II.2, for NLO PDF sets.

extension below $Q = 8$ GeV should be used in less accurate aspects of the calculation that are not sensitive to heavy-quark mass effects, such as inside the parton shower merged onto an (N)NLO fixed-order cross section.

Figure II.1 illustrates the Q dependence of singlet and gluon PDFs of the _30 and _100 ensembles at NLO and NNLO, for two select values of Bjorken x. Figs. II.2 and II.3 compare the uncertainty bands for the $g, u, \bar{d}$ and $\bar{s}$ distributions at a Q value of 1.4 GeV, at NNLO and NLO, respectively, for the PDF4LHC_30 and PDF4LHC_100 PDF sets. Good agreement between the two sets is found in all cases; the backward evolution is smooth and stable across the covered Q range, with only minor deviations observed below 2 GeV. [When examining the figures, recall that the _30 error bands can be slightly narrower for unconstrained x regions and PDF flavors at any Q].

1.4 PDF4LHC15 parton luminosities at NLO and NNLO

Even more relevant for physics applications than the PDF error bands are the parton luminosities. We have calculated the luminosities as a function of the mass of the final state, for a center-of-mass energy of 13 TeV. Comparisons of the gg and $q\bar{q}$ PDF luminosities, at NLO and NNLO, and defined as in [196], are shown in Fig. II.4 for PDF4LHC15_100, _30, and _MC sets, and in Fig. II.4 for PDF4LHC15_100, CT14, MMHT14, and NNPDF3.0 sets. Note that the size of the uncertainties shown here, and the level of agreement among the error bands, are different at low mass from those shown in the PDF4LHC document [181]. That is because,

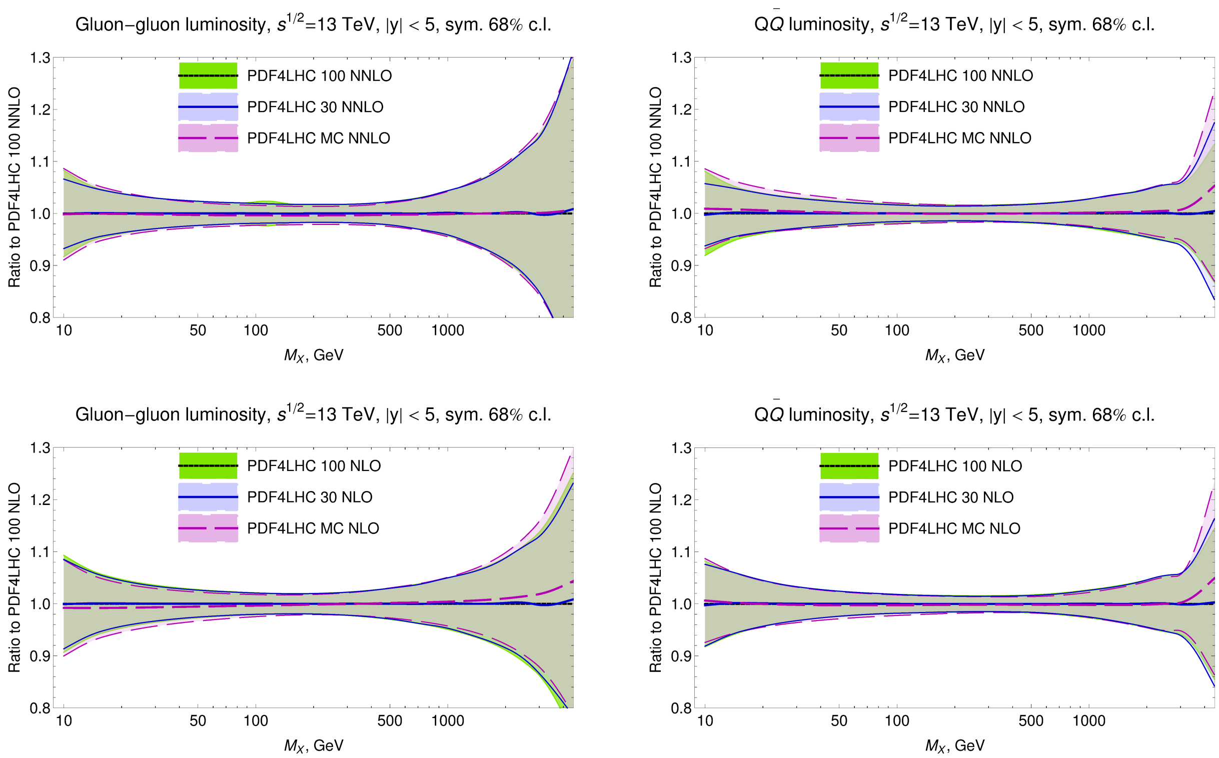

Fig. II.4: PDF4LHC15 NNLO and NLO parton luminosities at $\sqrt{s} = 13$ TeV in the experimentally accessible rapidity region $|y| < 5$.

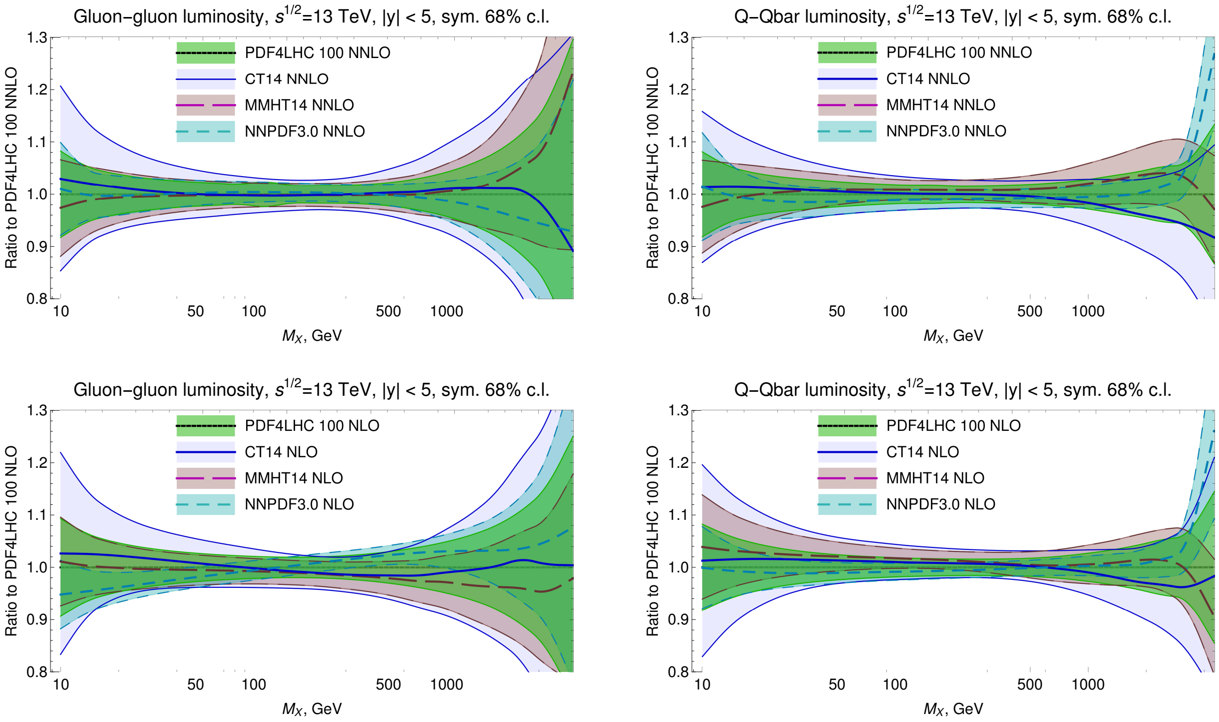

Fig. II.5: NNLO and NLO parton luminosities for PDF4LHC15_100, CT14, MMHT14, and NNPDF3.0 ensembles at $\sqrt{s} = 13$ TeV in the experimentally accessible rapidity region $|y| < 5$.

in our plots, a restriction has been applied on the x values of the PDFs to correspond to a

Process	Order	Type of calculation
$p + p \to Z + X$	NLO	aMCFast/APPLgrid
$p + p \to W^+ + X$	NLO	aMCFast/APPLgrid
$p + p \to W^- + X$	NLO	aMCFast/APPLgrid
$p + p \to W + X$, $A_{ch,W}$	NLO	aMCFast/APPLgrid
CMS $p + p \to W(l\nu) + X$, $A_{ch,l}$	NLO	aMCFast/APPLgrid
$p + p \to W^+\bar{c} + X$	NLO	aMCFast/APPLgrid
$p + p \to W^- c + X$	NLO	aMCFast/APPLgrid
$p + p \to t\bar{t} + X$	NLO	aMCFast/APPLgrid
$p + p \to t\bar{t}\gamma\gamma + X$	NLO	aMCFast/APPLgrid
ATLAS inclusive jets	NLO	NLOJET++/APPLgrid
ATLAS inclusive dijets	NLO	NLOJET++/APPLgrid
$p + p \to H(\gamma\gamma) + X$	LO, NLO	MCFM
$p + p \to H(\gamma\gamma) + jet + X$	LO, NLO	MCFM

Table II.1: Processes, QCD orders, and computer codes employed for comparisons of PDFs in the online gallery [191].

rapidity cut of $|y| < 5$ on the produced state. Without such a cut, the luminosity integral at masses below 40 GeV receives contributions from extremely low x of less than 10^{-5}, where (a) the uncertainties are larger, (b) the LHAPDF grids provided for the 30 PDF sets are outside of their tabulated range, and (c) the final state is produced in the forward region outside of the experimental acceptance of the LHC detectors. Without the constraints on the x range, the comparisons of parton luminosities at low mass are less relevant to LHC measurements.

1.5 4-flavor PDF4LHC15_30 sets

The nominal _30 ensemble has been generated for a maximum number of quark flavors of up to $N_f = 5$. An alternative _30 ensemble have been now provided for a maximum quark flavors of $N_f = 4$ at NLO, based on the same prescription as for the $N_f = 5$ sets, except that they are combined at an initial scale of 1.4 GeV in order to avoid backward evolution. We choose $\alpha_S(M_Z, N_f = 4) = 0.1126$ based on matching to $\alpha_S(M_Z, N_f = 5) = 0.118$ with a pole mass of 4.56 GeV for the bottom quark (equal to the average of masses of 4.75 and 4.18 GeV from the CT14, MMHT14, and NNPDF3.0 ensembles, and consistent with the PDG pole mass value).

1.6 PDF4LHC15 predictions for QCD observables

The PDF4LHC recommendation document [181] contains detailed guidelines to help decide which individual or combined PDFs to use depending on the circumstances.

To assist in this decision, predictions for typical LHC QCD observables have been calculated for an assortment of PDF sets. In Ref. [197], PDF4LHC15 predictions were made with the APPLgrid fast interface [3] for published LHC measurements within the fiducial region. To provide a complementary perspective, at a gallery website [191], we present LHC cross sections for processes listed in Table II.1 at 7, 8, and 13 TeV, and computed with no or minimal experimental cuts. The three (N)NLO ensembles of the PDF4LHC15 family (_100, _30, _MC [181]) are compared to those of ABM12 [198], CT14 [182], HERA2.0 [199], MMHT14 [183], and NNPDF3.0 [184]. The cross sections are calculated using (N)LO hard matrix elements either by a fast convolution of the PDFs with the tabulated parton-level cross section in the APPLgrid format [3], or by direct Monte-Carlo integration in MCFM [200]. Default $\alpha_s(M_Z)$

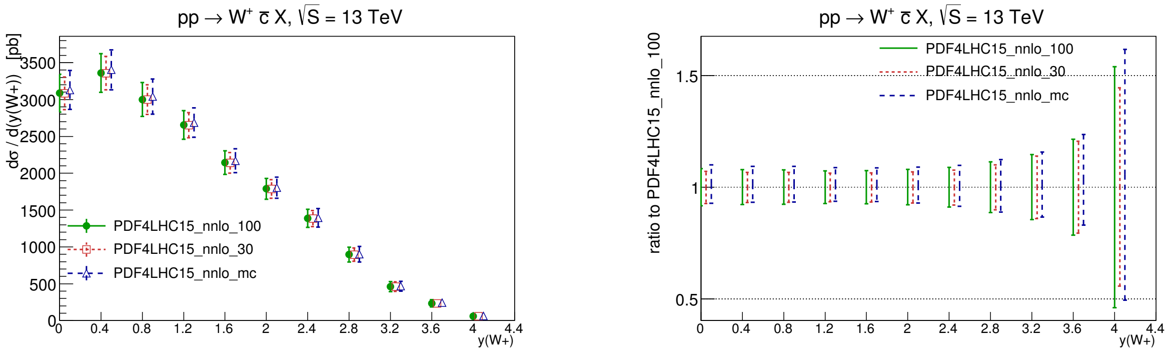

Fig. II.6: NLO predictions of $d\sigma/dy(W^+)$ in the process $pp \to W^+\bar{c}$ at the LHC 13 TeV, computed with APPLGRID.

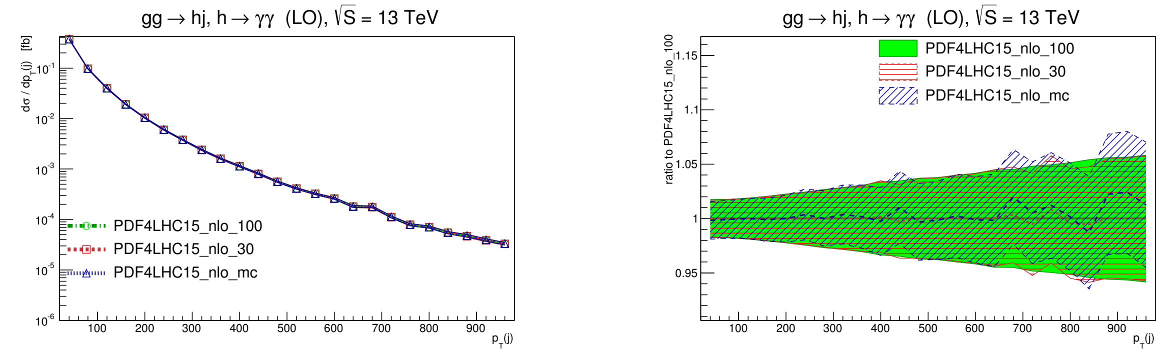

Fig. II.7: $d\sigma/dp_T(j)$ in the process $gg \to H(\gamma\gamma) + jet$ at the LHC 13 TeV, and its relative PDF uncertainties.

values are used with each PDF set. The APPLGRID files used in the computations are linked to the website. In the APPLGRID calculations, the hard cross sections are the same for all PDFs, while the MCFM-produced cross sections are sensitive to Monte-Carlo integration fluctuations that vary depending on the PDF ensemble, as will be discussed below.

The predictions were computed according to the following procedure. For production of $W^\pm$, Z^0, $t\bar{t}$, $t\bar{t}\gamma\gamma$, $W^+\bar{c}\,(W^-c)$, we use MadGraph_aMC@NLO [45], combined with aMC-fast [201] to generate APPLGRID files for different rapidities of the final-state particle. The renormalization and factorization scales are $\mu_R = \mu_F = M_W$, M_Z, $H_T/2$, $H_T/2$, M_W, respectively. H_T is the scalar sum of transverse masses $\sqrt{p_T^2 + m^2}$ of final-state particles. For $W^+\bar{c}\,(W^-c)$ production, we neglect small contributions with initial-state c or b quarks. For NLO single-inclusive jet and dijet production, we use public APPLGRID files [202] in the bins of ATLAS measurements [203], created with the program NLOJET++ [204, 205]. Similarly, the W charge asymmetry in CMS experimental bins [206, 207] is computed with APPLGRID from [208].

For cross sections of the Standard Model Higgs boson and Higgs boson+jet production via gluon fusion, with subsequent decay to $\gamma\gamma$, we use MCFM in the heavy-top quark approximation. Minimal cuts are imposed on the photons; the QCD scales are $\mu_F = \mu_R = m_H$.

The PDF uncertainties shown are symmetric, computed according to the prescriptions provided with each PDF ensemble, except for the HERA2.0 predictions, which are shown with asymmetric uncertainties, including contributions from both the eigenvector sets and the variation sets.

For each scattering process, our gallery shows plots of differential cross sections and ratios of PDF uncertainties to the central prediction based on PDF4LHC15_100. Figs. II.6 and II.7 provide two examples of comparisons presented on the website. When computed with APPLGRID, the cross sections reflect genuine differences in the PDFs; the hard cross sections are the same with all PDF sets. Thus we observe, for instance, in $W^+\bar{c}$ production in Fig. II.6 that the uncertainties of _100, _30, and _MC ensembles are very close across the central-rapidity range for most processes, with the _30 uncertainty being only slightly smaller (as expected), and with the differences that can be nearly always eliminated by slightly scaling the _30 uncertainty up by a constant factor (e.g., by multiplying it by ≈ 1.05 in Fig. II.6). The differences between the PDF4LHC15 ensembles grow at rapidities above 2-3, where the cross sections also are rapidly decreasing. The PDF uncertainties fluctuate more in the forward regions, reflecting paucity of experimental constraints on the PDFs.

Another perspective is glanced from H and $H+$jet production cross sections calculated by MCFM, cf. Fig. II.7. [Additional comparisons can be viewed on the website.] These illustrate that often the differences between the PDF4LHC15 reduced ensembles will be washed out by Monte-Carlo integration errors, save exceptionally precise calculations. To start, although the LHAPDF grids for the _100, _30, and _MC *central* sets are just independent tabulations of the *same* prior central set (they are equivalent up to roundoff errors), they will produce different fluctuations during the Monte-Carlo integration in MCFM or alike program. This is exemplified in the right frame in Fig. II.7, where the Higgs boson production cross sections are slightly different for the three LHAPDF tabulations of the central set solely because of MC fluctuations. In this figure, the cross sections were evaluated with 10^6 Monte-Carlo samplings and with PDF reweighting of events turned on. The events are exactly the same for all PDF sets within a given ensemble, and the event sequences are not the same among the ensembles because of the different roundoffs of the central LHAPDF grids. Even when the event reweighting is on, the PDF error bands fluctuate together with their respective central predictions.

The MCFM example touches on broader questions. The MC fluctuations can be suppressed by increasing the number of events or by using coarser binning for the cross sections. These adjustments tend to either lengthen the calculations, especially with the _100-replica ensembles, or to wash out the already small differences between the three PDF4LHC15 ensembles. There are several ways for “averaging” the input central PDF sets, e.g., because they use different evolution codes or round-offs. Each of these will lead to a different pattern of MC fluctuations. Finally, if the MC integration is done without PDF event reweighting, MC fluctuations will vary independently replica-by-replica. Using the combined PDF4LHC ensemble with fewer members may turn out to be preferable in such situations.

Acknowledgements

Work at ANL is supported in part by the U.S. Department of Energy under Contract No. DE-AC02-06CH11357. Work at SMU was supported by the U.S. Department of Energy under Grant DE-SC0010129. Work at MSU was supported by the National Science Foundation under Grant PHY-1410972.

2 On the accuracy and Gaussianity of the PDF4LHC15 combined sets of parton distributions [3]

We perform a systematic study of the combined PDF4LHC15 sets of parton distributions which have been recently presented as a means to implement the PDF4LHC prescriptions for the LHC Run II. These combined sets reproduce a prior large Monte Carlo (MC) sample in terms of

[3] S. Carrazza, S. Forte, Z. Kassabov, J. Rojo

either a smaller MC replica sample, or a Gaussian (Hessian) representation with two different number of error sets, and obtained using two different methodologies. We study how well the three reduced sets reproduce the prior for all the $N_\sigma \simeq 600$ hadronic cross-sections used in the NNPDF3.0 fit. We then quantify deviations of the prior set from Gaussianity, and check that these deviations are well reproduced by the MC reduced set. Our results indicate that generally the three reduced sets perform reasonably well, and provide some guidance about which of these to use in specific applications.

2.1 Introduction

Recently, the PDF4LHC Working Group [209] has presented updated recommendations and guidelines [181] for the usage of Parton Distribution Functions (PDFs) for LHC calculations at Run II. These recommendations are specifically based on the usage of combined PDF sets, which are obtained using the Monte Carlo (MC) method [185, 210], constructed from the combination of $N_{\rm rep} = 300$ MC replicas from the NNPDF3.0 [184], MMHT2014 [183] and CT14 [182] PDF sets, for a total number of $N_{\rm rep} = 900$ replicas. The combination has been performed both at NLO and at NNLO, and versions with $n_f = 4$ and $n_f = 5$ maximum number of active quark flavors are available. The impact of LHC measurements from Run I on PDF determinations has been discussed in a companion PDF4LHC report [211].

From this starting prior set, three reduced sets, two Hessian and one MC, are delivered for general usage. The reduced sets, constructed using different compression strategies, are supposed to reproduce as much as possible the information contained in the prior, but in terms of a substantially smaller number of error PDFs. The reduced Monte Carlo set, `PDF4LHC15_mc`, is constructed using the CMC-PDF method [187], and the two reduced Hessian sets, `PDF4LHC15_100` and `PDF4LHC15_30`, are constructed using the MC2H [188] and META-PDF [186] techniques, respectively. The PDF4LHC15 combined sets are available from `LHAPDF6` [190], and include additional PDF member sets to account for the uncertainty due to the value of the strong coupling constant, $\alpha_s(m_Z) = 0.1180 \pm 0.0015$.

The PDF4LHC 2015 report [181] presented general guidelines for the usage of the reduced sets, and some comparisons between them and the prior at the level of PDFs, parton luminosities, and LHC cross-sections, while referring to a repository of cross-sections on the PDF4LHC server [212] for a more detailed set of comparisons. It is the purpose of this contribution to make these comparisons more systematic and quantitative, in order to answer questions which have been frequently asked on the usage of the reduced sets. Specifically, we will perform a systematic study of the accuracy of the PDF4LHC15 reduced sets, by assessing the relative accuracy of uncertainties determined using each of them instead of the prior, for all hadronic observables included in the NNPDF3.0 PDF determination [184]. We will also compare the performance of the PDF4LHC15 reduced sets with that of the recently proposed SM-PDF sets [213]: specialized PDF sets which strive to minimize the number of PDF error sets which are needed for the description of a particular class of processes. We will then address the issue of the validity of the Gaussian approximation to PDF uncertainties by testing for gaussianity of the distribution of results obtained using the prior PDF set for a very wide variety of observables, and then assessing the performance and accuracy of both the Monte Carlo sets (which allows for non-Gaussian behaviour) and the Hessian compressed sets (which do not, by construction).

2.2 Validation of the PDF4LHC15 reduced PDF sets on a global dataset

We wish to compare the performance of the three reduced NLO sets, the two Hessian sets, `PDF4LHC15_nlo_30` and `PDF4LHC15_nlo_100`, and the Monte Carlo set `PDF4LHC15_nlo_mc`, for all the hadronic cross-sections included in the NNPDF3.0 global analysis [184]. These cross-sections have computed at $\sqrt{s} = 7$ TeV using NLO theory with `MCFM` [214], `NLOjet++` [204] and

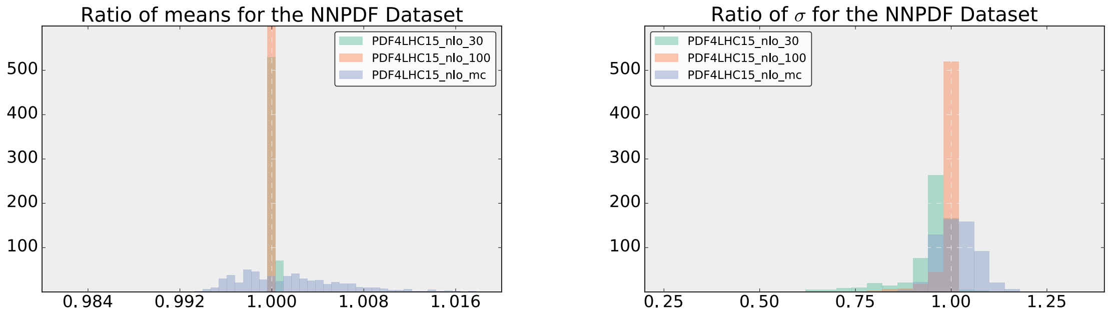

Fig. II.8: small Distribution of the ratio to the prior of means Eq. (II.1) (left) and of standard deviations Eq. (II.2) (right) computed for each of the $N_\sigma \simeq 600$ hadronic cross-sections included in the NNPDF3.0 global analysis, using each of the three reduced sets.

aMC@NLO [45, 201] interfaced to APPLgrid [3]. They include $N_\sigma \simeq 600$ independent observables for a variety of hadron collider processes such as electroweak gauge boson, jet production and top quark pair production, covering a wide region in the (x, Q) kinematical plane. In this calculation, the PDF4LHC15 combined sets are obtained from the LHAPDF6 interface.

In Fig. II.8 we show the distribution of the ratios

$$R_{\langle\sigma_i\rangle} \equiv \frac{\langle\sigma_i\rangle\,(\text{reduced})}{\langle\sigma_i\rangle\,(\text{prior})}\,, \quad i = 1\ldots, N_\sigma\,, \tag{II.1}$$

$$R_{s_i} \equiv \frac{s_i(\text{reduced})}{s_i(\text{prior})}\,, \quad i = 1\ldots, N_\sigma\,. \tag{II.2}$$

between respectively the means $\langle\sigma_i\rangle$, and the standard deviations s_i from each of the three reduced sets and the PDF4LHC15 prior, computed for all hadronic observables included in the NNPDF3.0 global analysis. For the Hessian sets the central value coincides with that of the prior, so the ratio of means is supposed to equal one by construction, with small deviations only due to rounding errors and interpolation in the construction of the LHAPDF grids, while for the MC set the mean is optimized by the CMC construction to fluctuate due to the finite size of the replica sample much less than expected on purely statistical grounds. Indeed, the histograms shows agreement of central values at the permille level. For standard deviations (i.e. PDF uncertainties) Fig. II.8 shows that using the PDF4LHC15_nlo_100 set they are reproduced typically with better than 5% accuracy. Differences are somewhat larger for the PDF4LHC15_nlo_mc and PDF4LHC15_nlo_30 sets.

In order to investigate the accuracy of PDF uncertainties in a more detailed quantitative way, we define the relative difference between the standard deviation, $s_i^{(\text{red})}$, of the cross-section σ_i computed with the reduced sets, and that of the prior, $s_i^{(\text{prior})}$:

$$\Delta_i \equiv \frac{\left|s_i^{(\text{prior})} - s_i^{(\text{red})}\right|}{s_i^{(\text{prior})}}\,, \quad i = 1, \ldots, N_\sigma\,. \tag{II.3}$$

In Figs. II.9 and II.10 the relative differences Δ_i are shown using NLO and NNLO PDFs for all hadronic observables which enter the NNPDF3.0 fit as a scatter plot in the (x, Q^2) kinematic plane, at the point corresponding to each observable using leading order kinematics [215], both for all observables, and for the 10% of observables exhibiting the largest relative differences. The x value corresponding to the the parton with highest x is always plotted, and for one-jet inclusive cross-sections, for which the x values of the two partons are not fixed even at leading

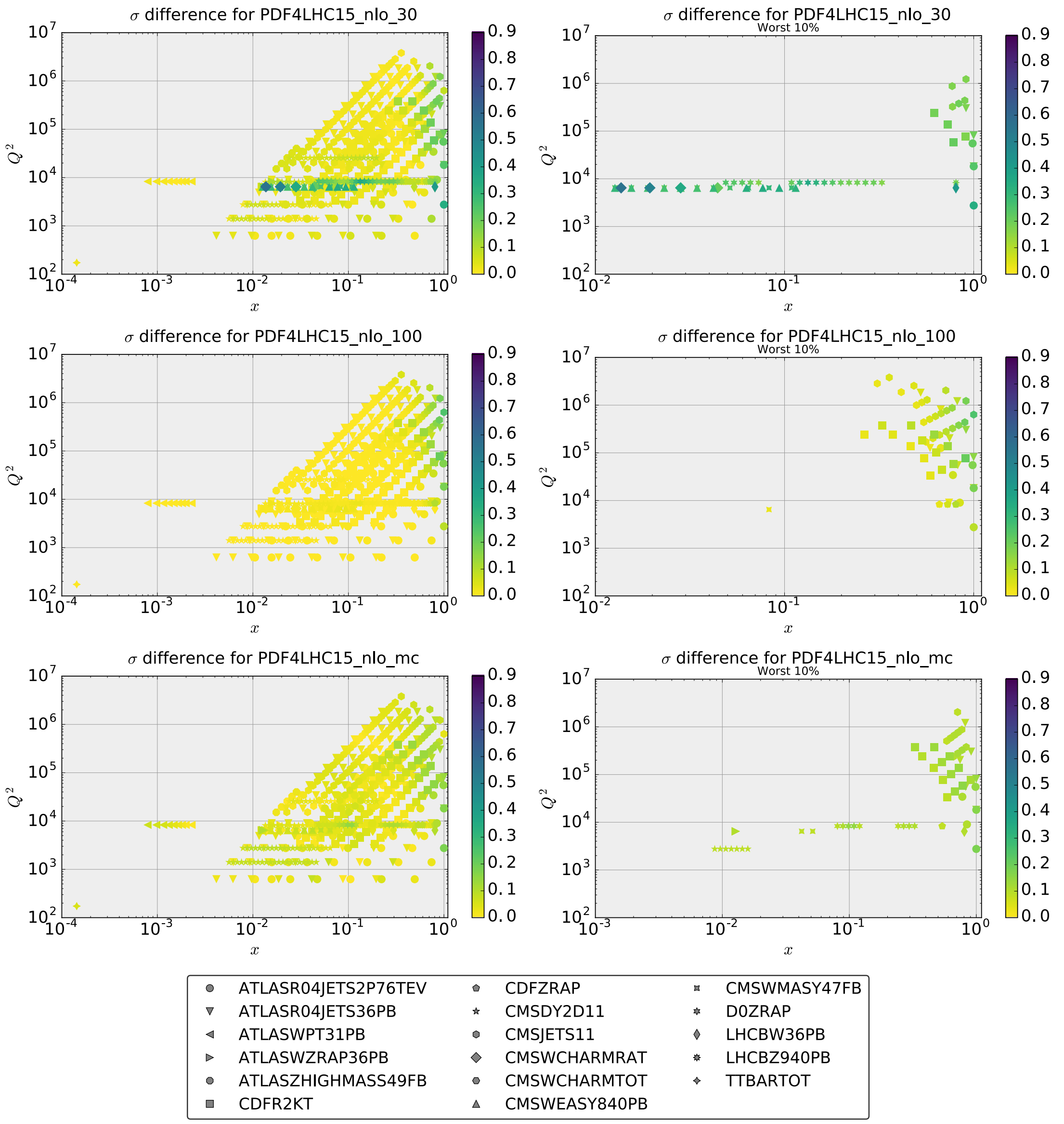

Fig. II.9: Relative difference Eq. (II.3), between the PDF uncertainties computed using the reduced set and the prior computed for all hadronic observables included in the NNPDF3.0 fit, shown as a scatter plot in the (x, Q^2) at the corresponding point, determined using leading-order kinematics. From top to bottom results for PDF4LHC_nlo_30, PDF4LHC_nlo_100 and PDF4LHC_nlo_mc are shown. In the left plots, all points are shown, in the right plots only the 10% of points with maximal deviation.

order, the largest accessible x which corresponds to the rapidity range of each data point is plotted. Of course, these x values should be only taken as indicative, and it should be born in mind that for most of the processes considered when one of the two partons involved is at large x, the other is at rather smaller x. In this comparison, NLO theory is used throughout.

From Figs. II.9 and II.10 we see that for PDF4LHC15_nlo_100 deviations are generally small, and concentrated in regions in which experimental information is scarce and PDF uncertainties are largest, such as the region of large x and large Q. For PDF4LHC15_nlo_mc and PDF4LHC15_nlo_30 the deviations are somewhat larger but still moderate in most cases, with a few outliers. No significant difference is observed between NLO and NNLO, consistent with the expectation that PDF uncertainties are driven by data, not by theory, and thus are very similar at NLO and NNLO.

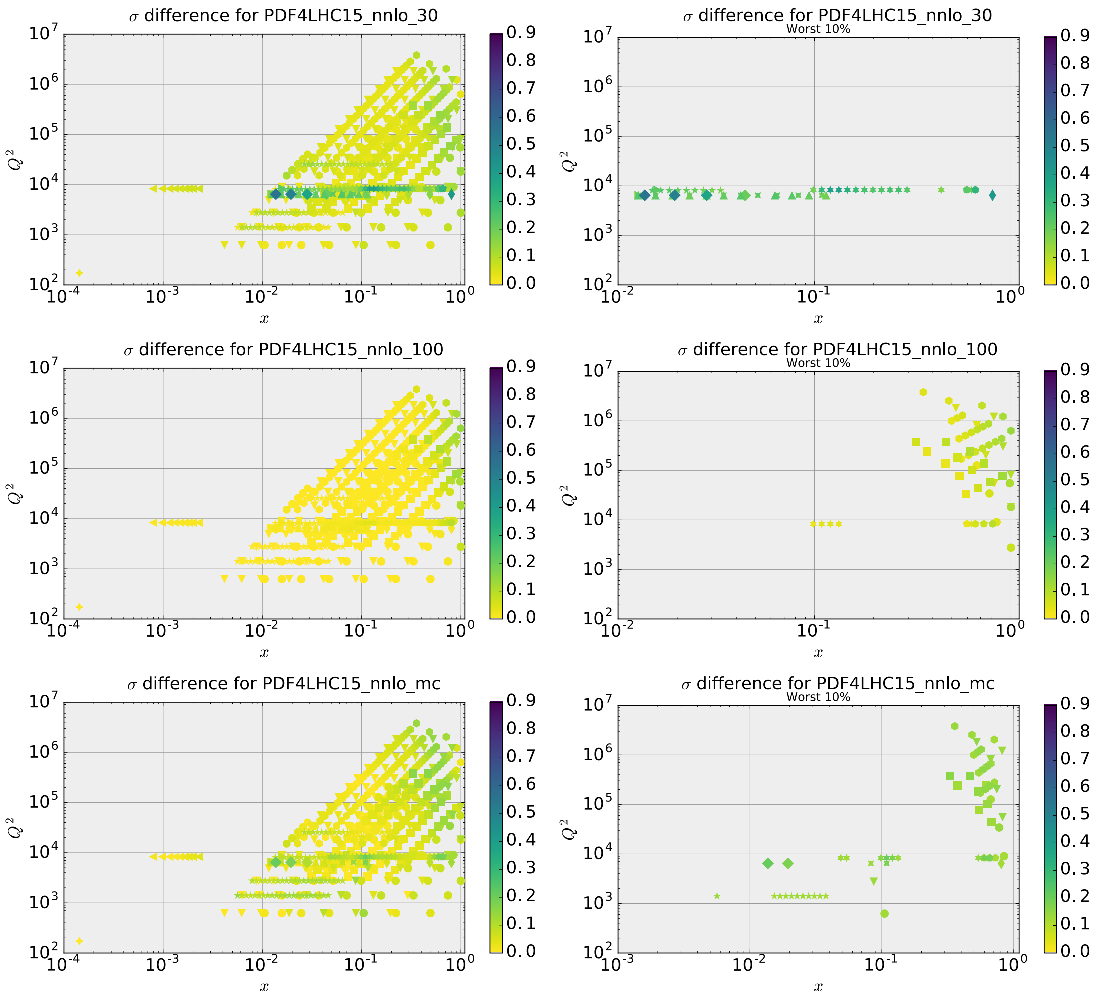

Fig. II.10: Same as Fig. II.9, but using the NNLO PDF sets.

This exercise shows that about $N_{\rm eig} = 100$ Hessian eigenvectors are necessary for a good accuracy general-purpose PDF set. On the other hand, we have recently argued [213] that a much smaller set of Hessian eigenvectors is sufficient in order to accurately reproduce a subset of cross-sections, and presented a technique to construct such specialized minimal sets, dubbed SM-PDFs. In order to test and validate this claim, we have constructed two such SM-PDF sets, using the methodology of Ref. [213], and starting from the PDF4LHC15 NLO prior:

– `SM-PDF-ggh`: this SM-PDF set, with $N_{\rm eig} = 4$ symmetric eigenvectors, reproduces the inclusive cross-section and the p_T and rapidity distributions of Higgs production in gluon fusion at $\sqrt{s} = 13$ TeV.

– `SM-PDF-Ladder`: this SM-PDF set, with has $N_{\rm eig} = 17$ symmetric eigenvectors, reproduces all the observables listed in Table II.2, which include a wide variety of LHC processes at $\sqrt{s} = 13$ TeV.

The `APPLgrid` grids for the processes in Table II.2 have been computed using `aMC@NLO` interfaced to `aMCfast`.

In Fig. II.11 we show again the relative difference Δ_i Eq. (II.3), which was shown in Fig. II.9, but now comparing to the prior these two SM-PDF sets. In the case of the `SM-PDF-ggh` set, we find good agreement with the prior for all cross-sections on the region $x \simeq 0.01$ and $Q \simeq$

process	distribution	$N_{\rm bins}$	range
$gg \to h$	$d\sigma/dp_t^h$	10	[0,200] GeV
	$d\sigma/dy^h$	10	[-2.5,2.5]
VBF hjj	$d\sigma/dp_t^h$	5	[0,200] GeV
	$d\sigma/dy^h$	5	[-2.5,2.5]
hW	$d\sigma/dp_t^h$	10	[0,200] GeV
	$d\sigma/dy^h$	10	[-2.5,2.5]
hZ	$d\sigma/dp_t^h$	10	[0,200] GeV
	$d\sigma/dy^h$	10	[-2.5,2.5]
$ht\bar{t}$	$d\sigma/dp_t^h$	10	[0,200] GeV
	$d\sigma/dy^h$	10	[-2.5,2.5]

process	distribution	$N_{\rm bins}$	range
Z	$d\sigma/dp_t^{l^-}$	10	[0,200] GeV
	$d\sigma/dy^{l^-}$	10	[-2.5,2.5]
	$d\sigma/dp_t^{l^+}$	10	[0,200] GeV
	$d\sigma/dy^{l^-}$	10	[-2.5,2.5]
	$d\sigma/dp_t^Z$	10	[0,200] GeV
	$d\sigma/dy^Z$	5	[-4,4]
	$d\sigma/dm^{ll}$	10	[50,130] GeV
	$d\sigma/dp_t^{ll}$	10	[0,200] GeV

process	distribution	$N_{\rm bins}$	range
$t\bar{t}$	$d\sigma/dp_t^{\bar{t}}$	10	[40,400] GeV
	$d\sigma/dy^{\bar{t}}$	10	[-2.5,2.5]
	$d\sigma/dp_t^t$	10	[40,400] GeV
	$d\sigma/dy^t$	10	[-2.5,2.5]
	$d\sigma/dm^{t\bar{t}}$	10	[300,1000]
	$d\sigma/dp_t^{t\bar{t}}$	10	[20,200]
	$d\sigma/dy^{t\bar{t}}$	12	[-3,3]

process	distribution	$N_{\rm bins}$	range
W	$d\sigma/d\phi$	10	[0,200] GeV
	$d\sigma/dE_t^{\rm miss}$	10	[-2.5,2.5]
	$d\sigma/dp_t^l$	10	[0,200] GeV
	$d\sigma/dy^l$	10	[-2.5,2.5]
	$d\sigma/dm_t$	10	[0,200] GeV
	$d\sigma/dp_T^W$	5	[-4,4]
	$d\sigma/y^W$	10	[50,130] GeV

Table II.2: LHC processes and the corresponding differential distributions used as input in the construction of the `SM-PDF-Ladder` set. In each case we indicate the range spanned by each distribution and the number of bins $N_{\rm bins}$. All processes have been computed for $\sqrt{s} = 13$ TeV. Higgs bosons and top quarks are stable, while weak gauge bosons are assumed to decay leptonically. No acceptance cuts are imposed with the exception of the leptons from the gauge boson decay, for which we require $p_T^l \geq 10$ GeV and $|\eta^l| \leq 2.5$.

100 GeV, relevant for Higgs production in gluon fusion. On the other hand, as we move outside this region, the accuracy rapidly deteriorates. This exemplifies the virtues and limitations of the SM-PDF approach: a very small number of eigenvectors is sufficient to reproduce a reasonably small set of observables, but if one tries to stretch results to too many processes there is accuracy loss. The `SM-PDF-Ladder` set, on the other hand, exhibits a similar performance as the `PDF4LHC15_nlo_30` set.

2.3 Non-Gaussianities in the PDF4LHC combination

As discussed in the PDF4LHC15 report [181], the Monte Carlo combination of individual PDF set in general is not Gaussian. This is both because one of the three sets entering the combination, NNPDF3.0, allows for non-Gaussian behaviour, and also because in general the combination of Gaussian sets is not necessarily Gaussian itself. We will now study in a more systematic way the degree of non-Gaussianity of the prior set, and specifically correlate the comparison of the reduced sets to the prior with the degree of non-Gaussianity of the prior. This has the threefold purpose of determining how much the accuracy of the Hessian set deteriorates in the presence of non-Gaussianities, of checking that the reduced MC set correctly reproduces the non-Gaussianity of the prior, and of providing guidance on when the MC set should be favored over the Hessian sets in order to reproduce the non-Gaussianity.

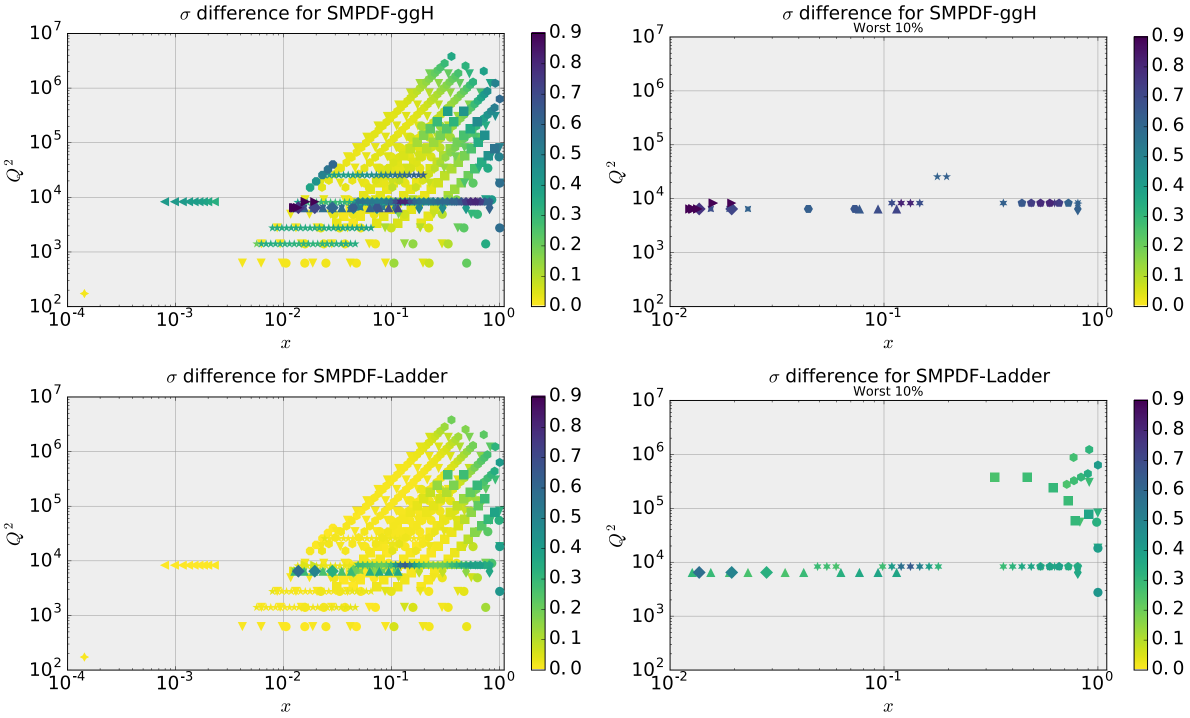

Fig. II.11: Same as Fig. II.9, this time for the two SM-PDF sets, `SM-PDF-ggh` (upper plots) and `SM-PDF-Ladder` (lower plots).

In order to study non-Gaussianity, we proceed in two steps. First, we turn the histogram, as obtained from a Monte Carlo representation, into a continuous probability distribution. Then, we compare this probability distribution to a Gaussian with the same mean and standard deviation. The first step is accomplished using the Kernel Density Estimate (KDE) method. The second, using the KullbackâĂŞLeibler (KL) divergence as a measure of the difference between two probability distributions (for a brief review of both methods see e.g. Ref. [216]).

The KDE method consists of constructing the probability distribution corresponding to a histogram as the average of kernel functions K centered at the data from which the histogram would be constructed. In our case, given $k = 1, \ldots, N_{\rm rep}$ replicas of the i-th cross-section $\{\sigma_i^{(k)}\}$, the probability distribution is

$$P(\sigma_i) = \frac{1}{N_{\rm rep}} \sum_{k=1}^{N_{\rm rep}} K\left(\sigma_i - \sigma_i^{(k)}\right) , \quad i = 1, \ldots, N_\sigma \, . \tag{II.4}$$

We specifically choose

$$K(\sigma - \sigma_i) \equiv \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{(\sigma - \sigma_i)^2}{h}\right) , \tag{II.5}$$

where the parameter h, known as bandwidth, is

$$h = \hat{s}_i \left(\frac{4}{3N_{\rm rep}}\right)^{\frac{1}{5}} , \tag{II.6}$$

where $\hat{s}_i$ is the standard deviation of the given sample of replicas. This choice is known as *Silverman rule*, and, if the underlying probability distribution is Gaussian, it minimizes the integral of the square difference between the ensuing distribution and this underlying Gaussian [217]. Once turned into continuous distributions via the KDE method, the prior and reduced Monte Carlo sets can be compared to each other, to a Gaussian, and to the Hessian sets. The comparison can be performed using the KullbackâĂŞLeibler (KL) divergence, which measures the information loss when using a probability distribution $Q(x)$ to approximate a prior $P(x)$, and is given by

$$D_{\rm KL}^{(i)}(P|Q) = \int_{-\infty}^{+\infty} \left(P(x) \cdot \frac{\log P(x)}{\log Q(x)}\right) dx. \, , \tag{II.7}$$

As a first example, in Fig. II.12 we select a data bin in which the distribution of PDF replicas is clearly non-Gaussian, namely the most forward rapidity bin in the LHCb $Z \to \mu\mu$ 8 TeV measurement [218], and we compare the distribution obtained using the PDF4LHC15 prior to those found using the `PDF4LHC15_nlo_mc` and `PDF4LHC15_nlo_100` reduced sets. The continuous distribution shown is obtained from the prior and reduced MC samples using the KDE method discussed above. For the `PDF4LHC15_nlo_100` set the distribution shown is a Gaussian with with and central value using the standard procedure, based on linear error propagation, which is used to obtain predictions from Hessian sets: namely, the central set provides the mean, and the standard deviation is the sum in quadrature of the deviations obtained using each of the error sets.

The KL divergence between the prior and a Gaussian is equal to $D_{KL} = 0.153$, while the divergence between the prior and its reduced MC representation is $D_{KL} = 0.055$, and finally between the prior and Hessian set it is $D_{KL} = 0.19$. This shows that the reduced MC representation of the prior is much closer to it than the prior is to a Gaussian, while the Hessian representation differs from it even more. In order to facilitate the interpretation these values of the KL divergence, in Fig. II.13 we plot the value of the KL divergence between two Gaussian with different width, as a function of the ratio of their width: the plot shows that $D_{KL} \sim 0.05$

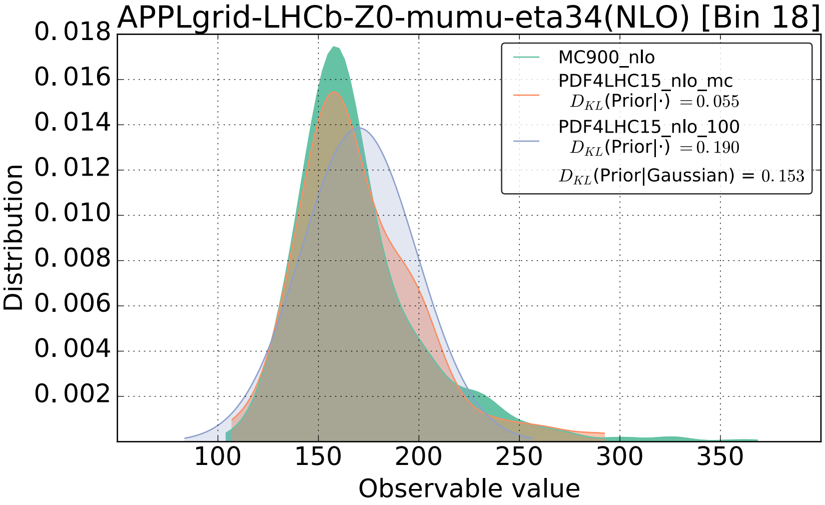

Fig. II.12: The probability distribution for the most forward bin in the LHCb $Z \to \mu\mu$ 8 TeV measurement obtained using the PDF4LHC15 prior and the `PDF4LHC15_nlo_mc` and `PDF4LHC15_nlo_100` reduced sets. The value of the KL divergence D_{KL} Eq. (II.7) between the prior and a Gaussian, and between each of the two reduced sets and the prior for this distribution, are also given.

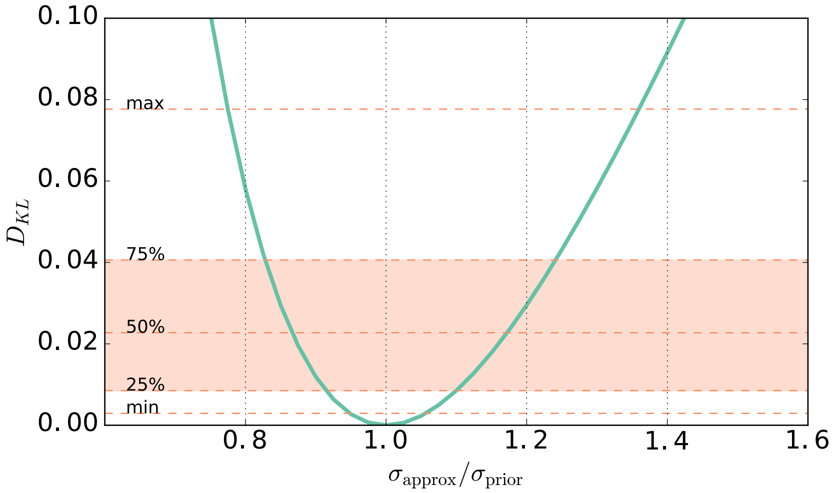

Fig. II.13: The KL divergence D_{KL} Eq. (II.7) between two Gaussian distributions with the same mean but different widths, as a function of the ratio of their standard deviations. We also show (horizontal lines) the highest value, lowest value, and the edges of the quartiles of the distribution of D_{KL} values between the prior and a Gaussian approximation to it, for all observables listed in Table II.2.

corresponds to distorting the width of a Gaussian by about 20%. In this figure we also show as horizontal lines the minimum and maximum values that we obtained, as well as the edges of the four quartiles of the distribution of results.

We have extended the type of comparisons shown in Fig. II.12 into a systematic study including all the cross-sections listed in Table II.2. As discussed in Sect. 2.2, this is a reasonably representative set of observables, since it is possible to construct a PDF set, the `SM-PDF-Ladder`, which is adequate to describe them and is also accurate to describe all the hadronic cross-section from the NNPDF3.0 fit (see Fig. II.11). Specifically, for each cross-section we have determined the probability distribution from the prior using the KDE method Eq. (II.4), and also a Gaussian

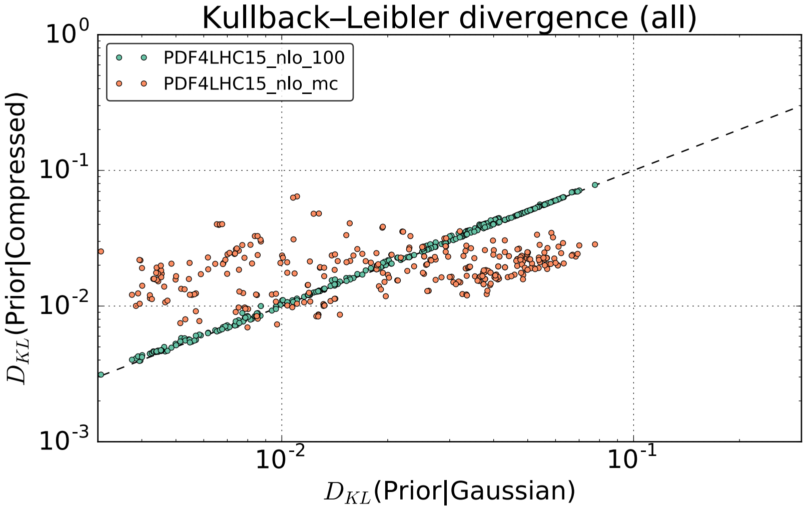

Fig. II.14: The KL divergence, Eq. (II.7) between the prior and each of its two reduced representations `PDF4LHC15_nlo_prior` (Monte Carlo) and `PDF4LHC15_nlo_mc` (Hessian) vs. the divergence between the prior and its Gaussian approximation, computed for all observables listed in Table II.2.

approximation to it, defined as the Gaussian with the same mean and standard deviation as the prior. We have computed the KL divergence between the prior distribution and this Gaussian approximation. It is clear that the vast majority of observables exhibits Gaussian behaviour to good approximation, with extreme cases such as shown in Fig. II.12 happening in a small fraction of the first quartile.

We have then computed for each observable the KL distance between the prior and the `PDF4LHC15_nlo_mc` and `PDF4LHC15_nlo_100` combined sets. Results are collected in Fig. II.14 for all processes, while in Fig. II.15 we show a breakdown for the four classes of processes of Table II.2: Higgs, top, W and Z production. For each cross-section there are two points on the plot, one corresponding to `PDF4LHC15_nlo_mc` and the other to `PDF4LHC15_nlo_100`. The points are plotted with on the x axis the KL divergence between the prior and its gaussian approximation, and on the y axis the same quantity now evaluated between the prior and the compressed set. For the `PDF4LHC15_nlo_100` all points cluster on the diagonal: this means that the reduced Hessian set only deviates from the prior inasmuch as the prior deviates from a Gaussian — only for a more extreme deviation from Gaussian such as shown in Fig. II.12 does the reduced Hessian deviate more. The `PDF4LHC15_nlo_mc` points instead approximately fall within a horizontal band: this means that the quality of the approximation to the prior of the reduced MC does not depend on the degree of non-Gaussianity of the prior itself.

Hence, the reduced MC set does reproduce well the non-Gaussian features of the prior, when they are present, and it will be advantageous to use it for points where the center of the band corresponding to `PDF4LHC15_nlo_mc` is below the diagonal. Figure II.15 shows that this happens for a significant fraction of the W and Z production cross-sections, but not for top and Higgs production. This is consistent with the expectation that non-Gaussian behaviour is mostly to be found in large x PDFs, which are probed by gauge boson production at high rapidity, but not by Higgs and top production which are mostly sensitive to the gluon PDF at medium and small x.

In order to further elucidate the dominant non-Gaussian features, we have performed a comparison of the mean and the standard deviation of each probability distribution with

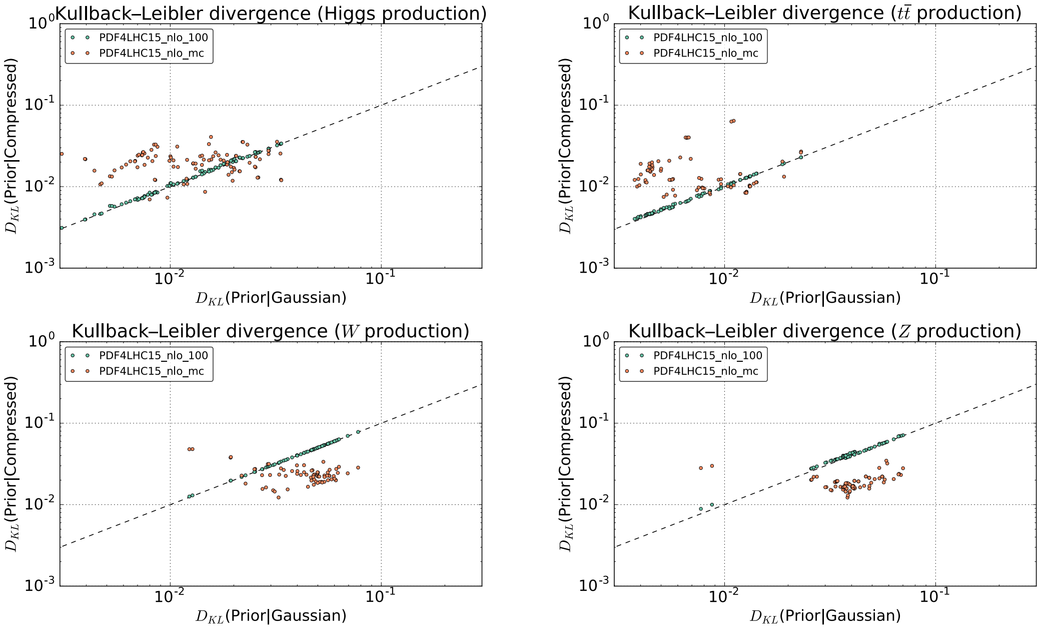

Fig. II.15: Same as Fig. II.14, now separating the contributions of the different classes of processes of Table II.2: Higgs production (top left), top quark pair production (top right), W production (bottom left) and Z production (bottom right).

respectively the median and the minimum 68% confidence interval R, defined as

$$R = \frac{1}{2}\min\{[x_{\min}, x_{\max}]; \qquad \int_{x_{\min}}^{x_{\max}} P(x) = 0.683\,. \tag{II.8}$$

The deviation of the median from the mean is a measure of the asymmetry of the distribution, while the deviation R from the standard deviation is a measure of the presence of outliers. We then define two estimators, one for the deviation of the mean from the median and for the deviation of the standard deviation s from R:

$$\Delta_\mu = \frac{\text{median} - \mu}{s}\,, \tag{II.9}$$

$$\Delta_s = \frac{R - s}{s}\,. \tag{II.10}$$

Both Δ_μ and Δ_s would vanish for a Gaussian in the limit of infinite sample size.

In Fig. II.16 we represent these two estimators in a scatter plot, with Δ_μ and Δ_s respectively on the x and y axis, computed for all the cross-sections of Table II.2. In addition, we show a color code with the KL divergence between the prior and respectively its Gaussian approximation and its two reduced MC and Hessian representations. From this comparison, it is clear that the shift in the median is only weakly correlated to the degree of non-Gaussianity (top plot), and also weakly correlated to the shift is standard deviation, which instead is strongly correlated to non-Gaussianity.

In the presence of outliers, $R \leq s$, and indeed R is seen to be always negative. We expect asymmetries related to non-Gaussian behaviour to be due to the fact that in some cases PDFs are bounded from below by positivity, but not from above where outliers may be present. Indeed in the non-gaussian region Δ_μ tends to be negative, but with large fluctuations in its value. The same correlations are seen with the KL divergence between prior and Hessian, again

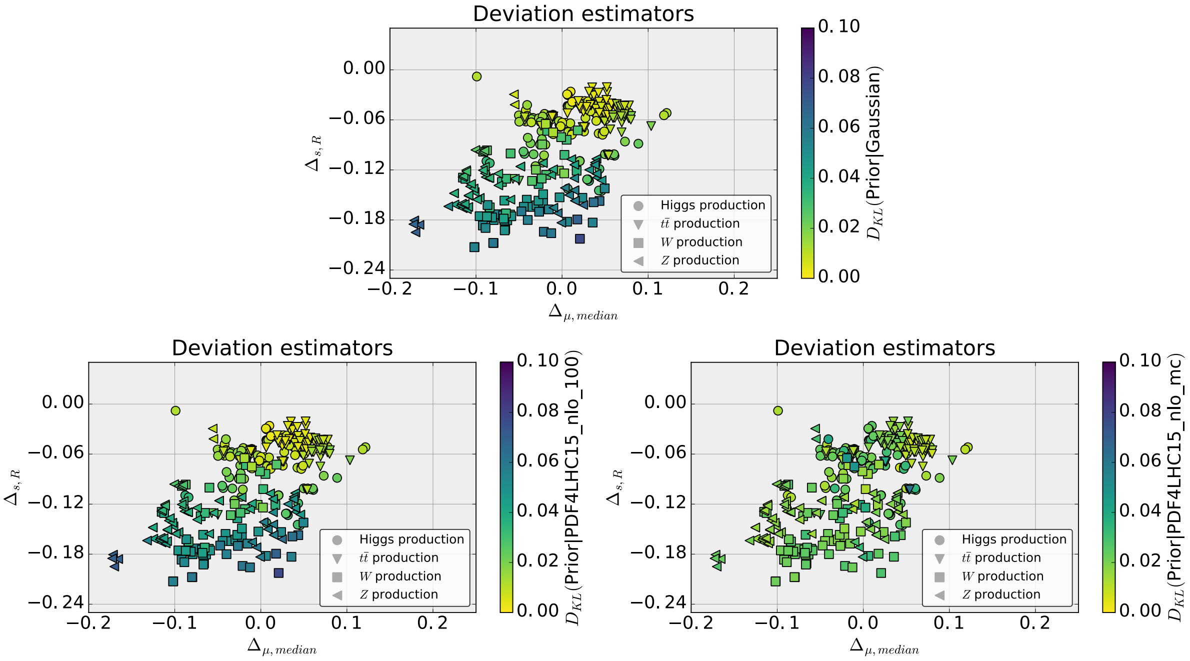

Fig. II.16: Scatter plot of indicators of deviation from gaussianity for all the cross-sections of Table II.2. For each observable, the shift between the median and the mean Eq. (II.9) is shown in the horizontal axis, while the shift between standard deviation and the 68% interval Eq. (II.10) is represented on the vertical axis. The color code shows the KL divergence between the prior and either a Gaussian (top) or the two reduced sets (bottom): Hessian (left) and MC (right).

showing that this is dominated by non-gaussian behaviour. On the other hand, no correlation is observed from the divergence between prior and reduced MC, consistent with our conclusion that the performance of the compressed MC set is independent of the degree of non-Gaussianity.

2.4 Summary and outlook

In this contribution we have performed a systematic comparison of the three reduced PDF4LHC15 PDF sets with the prior distribution they have been constructed from, with particular regard to non-Gaussian features, by comparing predictions for a wide variety of LHC cross-sections. Our general conclusion is that the three sets all perform as expected. We have specifically verified that the `PDF4LHC15_nlo_100` Hessian set provides generally the most accurate representation of the mean and standard deviation of the probability distribution, while the `PDF4LHC15_nlo_mc` and `PDF4LHC15_nlo_30` sets are less accurate though still quite good. We have also verified that specialized SM-PDF [213] sets can give an equally accurate representation, but with a smaller number of error-sets, at the price of not being suited for all possible processes, but with the option of combining them with other more accurate sets. We have then verified that in the presence of substantial deviations from Gaussianity, the `PDF4LHC15_nlo_mc` set is the most accurate. By providing a breakdown of our comparisons by type of process, we have verified that both deviations from Gaussianity and loss of accuracy of the smaller Hessian set are more marked in regions which are sensitive to poorly known PDFs, such as the anti-quarks at large x.

The results for the $N_\sigma \simeq 600$ cross-sections used for the calculations in Figs. II.9–II.11 are available from the link

`http://pcteserver.mi.infn.it/~nnpdf/PDF4LHC15/gall`

from where they can be accessed in HTML, CSV and ODS formats.

Acknowledgements

S. C. and S. F. are supported in part by an Italian PRIN2010 grant and by a European Investment Bank EIBURS grant. S. C. is supported by the HICCUP ERC Consolidator grant (614577). S. F. and Z. K. are supported by the Executive Research Agency (REA) of the European Commission under the Grant Agreement PITN-GA-2012-316704 (HiggsTools). J. R. is supported by an STFC Rutherford Fellowship and Grant ST/K005227/1 and ST/M003787/1 and by an European Research Council Starting Grant *"PDF4BSM"*.