The Collinear Parton Distributions of the Proton: Achievements and Open Issues

Forward Physics and QCD at the LHC and EIC

Emanuele R. Nocera

Università degli Studi di Torino and INFN, Torino

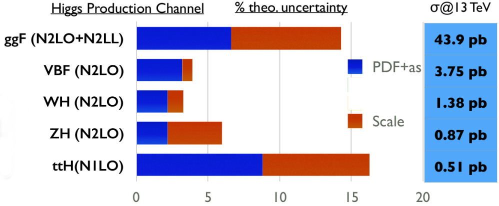
23 October 2023

PDFs at the LHC

$$\sigma(Q^2,\tau,\mathbf{k}) = \sum_{ij}\int_\tau^1 \frac{dz}{z}\mathcal{L}_{ij}(z,Q^2)\hat{\sigma}_{ij}\left(\frac{\tau}{z},\alpha_s(Q^2),\mathbf{k}\right) \qquad \mathcal{L}_{ij}(z,Q^2) = (f_i^{h_1}\otimes f_j^{h_2})(z,Q^2)$$

PDF uncertainty is often the dominant source of uncertainty in LHC cross sections

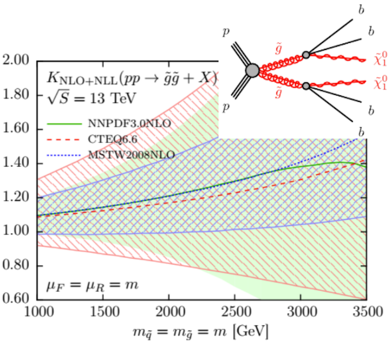
Precision



Channel	$m_{W^+} - m_{W^-}$ [MeV]	Stat. Unc.	Muon Unc.	Elec. Unc.	Recoil Unc.	Bckg. Unc.	QCD Unc.	EW Unc.	PDF Unc.	Total Unc.
$W \to e\nu$	−29.7	17.5	0.0	4.9	0.9	5.4	0.5	0.0	24.1	30.7
$W \to \mu\nu$	−28.6	16.3	11.7	0.0	1.1	5.0	0.4	0.0	26.0	33.2
Combined	−29.2	12.8	3.3	4.1	1.0	4.5	0.4	0.0	23.9	28.0

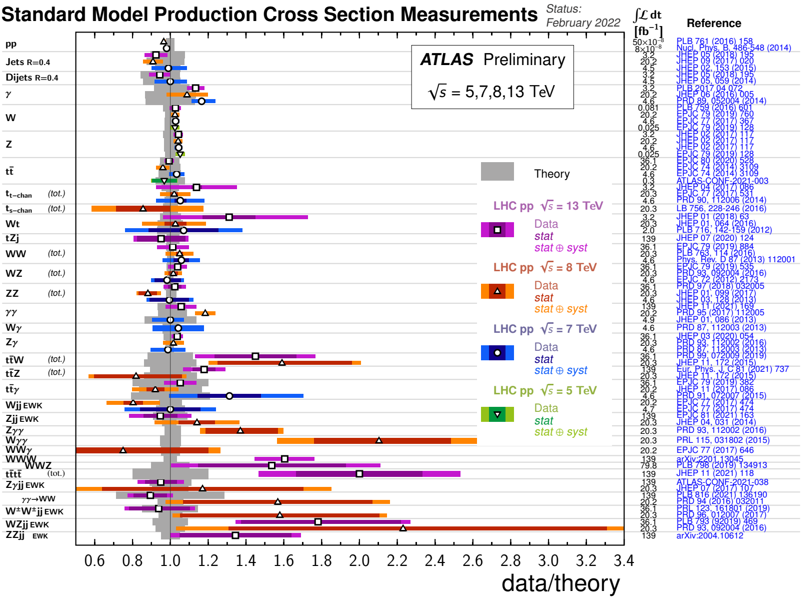
[Plot from the CERN Yellow Report 2016]

Discovery



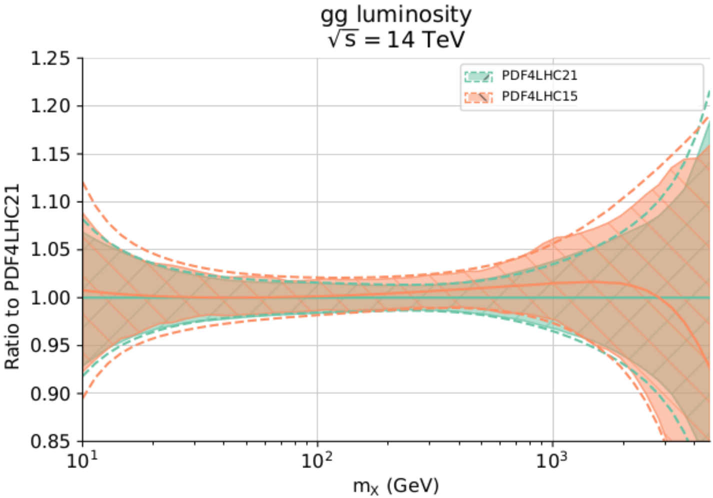
[EPJC 76 (2016) 53]

Physics at the LHC as Precision Physics



[Plot from ATLAS Collaboration web page]

We are approaching 1% PDF uncertainties

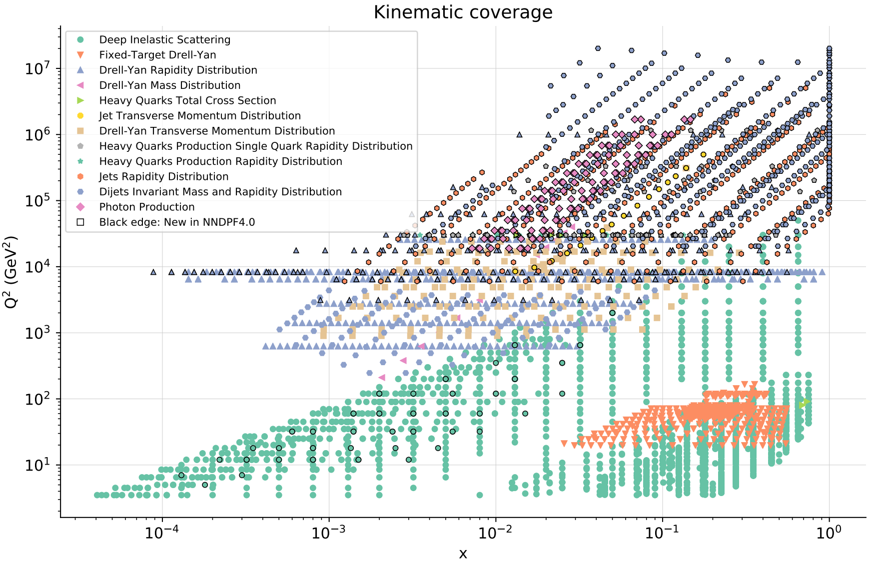


[J.Phys.G 49 (2022) 080501]

The path towards 1% PDF uncertainties goes through data, theory and methodology

1. Data

Data: kinematic coverage



NNPDF4.0 (NNLO) $N_{\rm dat} = 4618$ $\chi^2/N_{\rm dat} = 1.16$

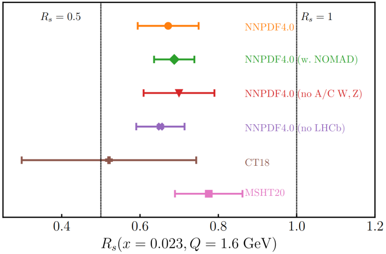
Data consistency

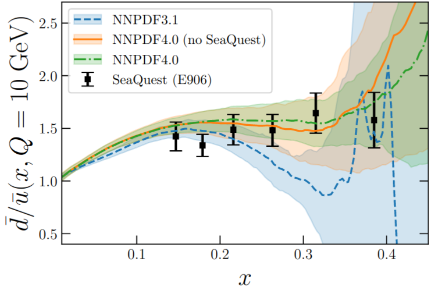
Strange in the proton [EPJ C80 (2020) 1168]

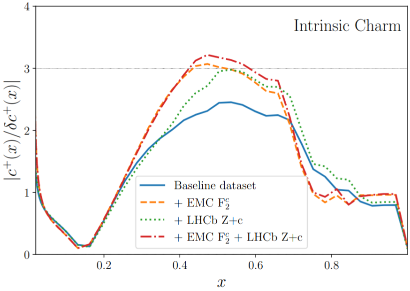
$$R_s(x,Q^2) = \frac{s(x,Q^2)+\bar{s}(x,Q^2)}{\bar{u}(x,Q^2)+\bar{d}(x,Q^2)}$$

The d/u ratio in the proton [EPJ C82 (2022) 428]

Charm in the proton [Nature 608 (2022) 7923 483]







Data inconsistency: tensions between data sets

Give more weight to a data set p

$$\chi^2 \to \chi^2 + w\chi_p^2$$

Refit: the total χ^2 will increase
Which data sets get worse? How much?

Refit: the data set χ_p^2 will decrese
Self-consistency? Inconsistency?

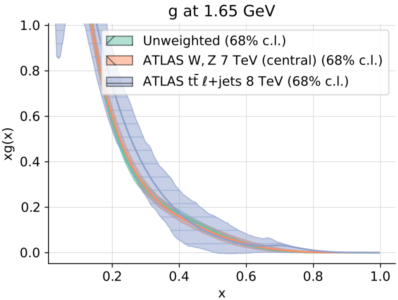
Examples: ATLAS W, Z and $t\bar{t}$

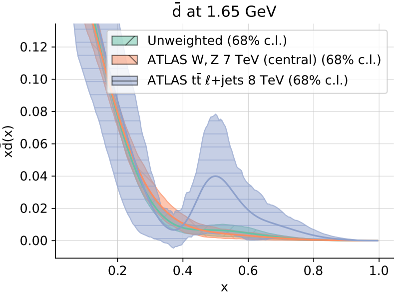
Inconsistency clearly spotted
unnatural PDF shapes appear
error in other data sets increases

Otherwise global fit quality
and PDFs remain unaltered

Data set	baseline	rw W, Z	rw $t\bar{t}$
ATLAS W, Z 7 TeV	1.86	1.23	—
ATLAS $t\bar{t}$ 8 TeV	4.11	—	1.21
Total	1.20	1.21	1.73

[EPJ C82 (2022) 428]





Data inconsistency: experimental correlations

Single inclusive jet data from ATLAS 7 TeV

default correlations: terrible χ^2

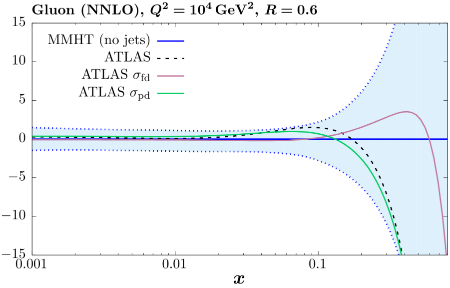
(correlations across rapidity bins)

decorrelation models: improve the fit a lot

$n_{\rm dat}$	default	part. decorr.	full decorr.
140	1.89	1.28	0.83

no significant effect on the extracted gluon

similar gluon irrespective of the rapidity bin



[EPJ C78 (2018) 248; EPJ C80 (2020) 797]

Top pair production from ATLAS 8 TeV

default correlations: terrible χ^2

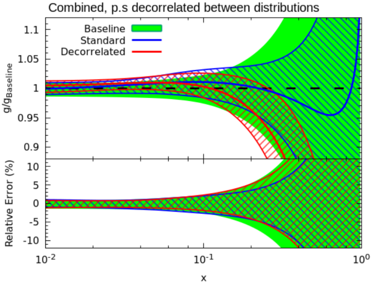
(correlations across different spectra)

decorrelation models: improve the fit a lot

$n_{\rm dat}$	default	stat. uncorr.	p.s. uncorr
25	7.00	3.28	1.80

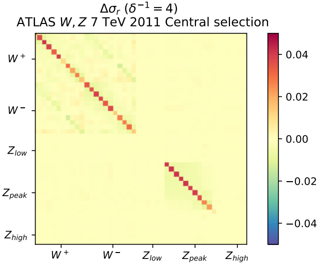
appreciable effect on the extracted gluon

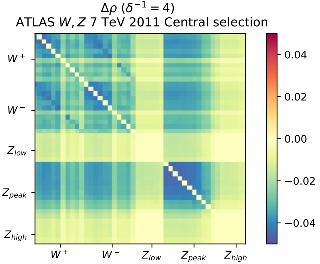
different gluon depending on the top spectrum



[EPJ C80 (2020) 1; Les Houches proceedings, 2019]

Good knowledge of experimental correlations is important





Assumptions:

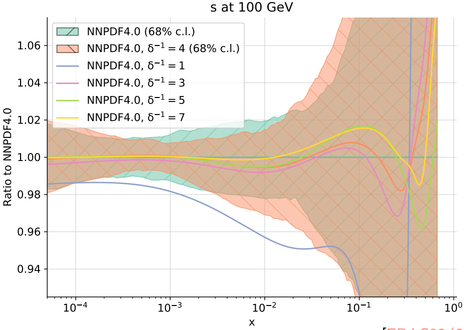
correlations are determined less precisely than variances

inaccuracy is limited to a small number of uncertainties

Regularisation procedure:

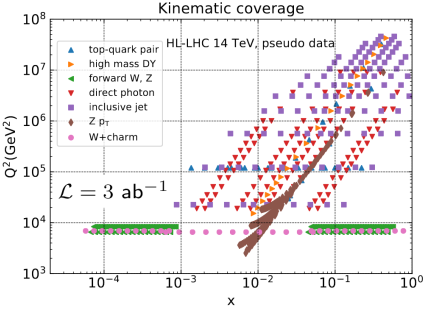
clip the singular values of the correlated part of the matrix of uncertainties to a constant δ, whenever these are smaller than that, while leaving the rest of the singular vectors unchanged

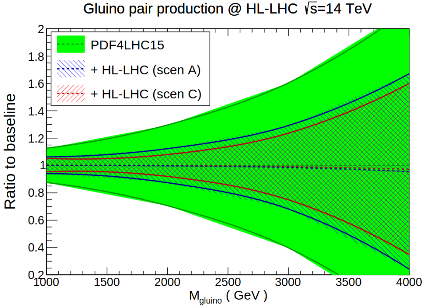
$\chi^2_{4.0}/N_{\rm dat} = 1.16 \qquad \chi^2_{\delta=4}/N_{\rm dat} = 1.11 \qquad (N_{\rm dat} = 4618)$

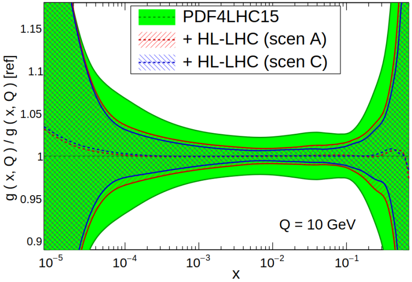


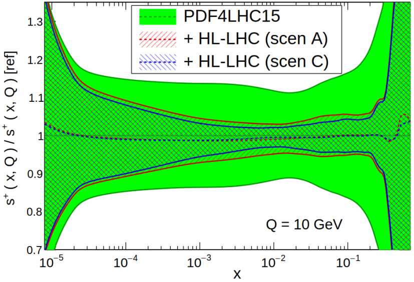
[EPJ C82 (2022) 956]

Impact of future data: HL-LHC



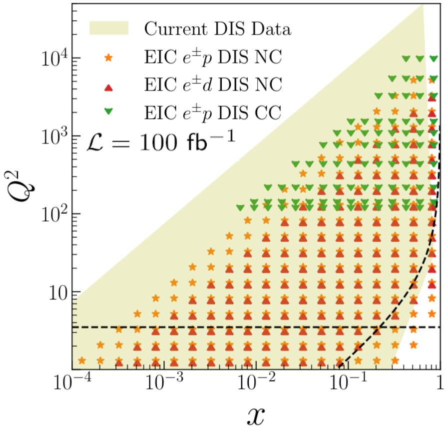






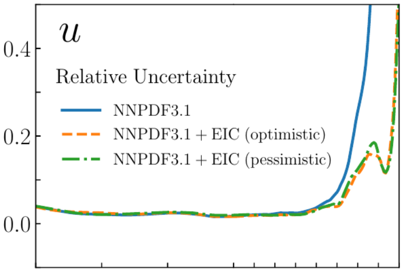
[EPJ.C 78 (2018) 962]

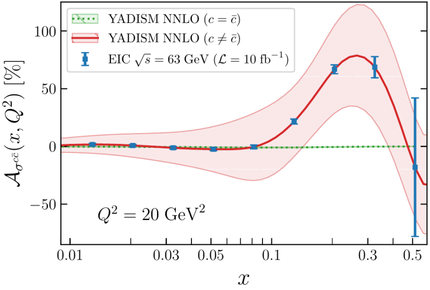
Impact of future data: EIC



$E_\ell \times E_p$ [GeV]: 18× 275; 10× 100; 5× 100

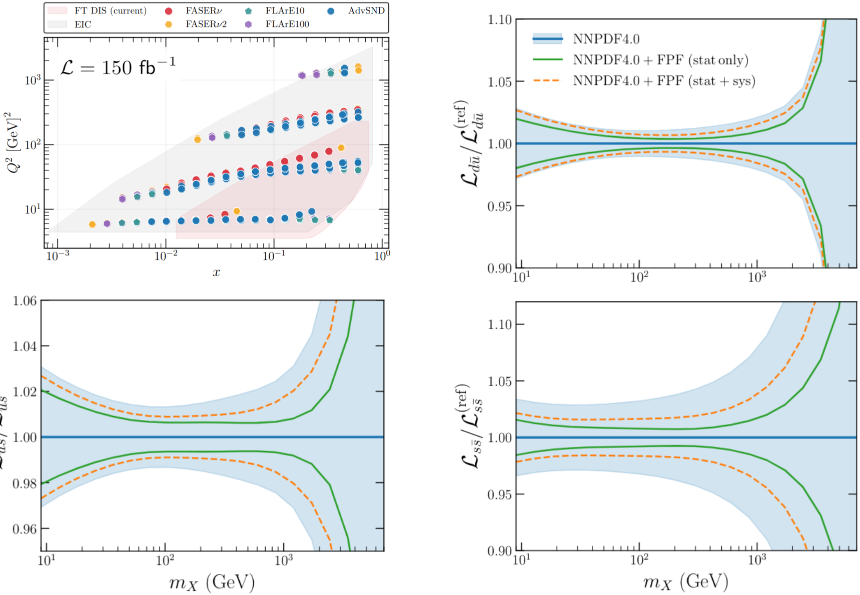
$$\mathcal{A}_{\sigma^{c\bar{c}}} = \frac{\sigma^{c}_{\rm red} - \sigma^{\bar{c}}_{\rm red}}{\sigma^{c\bar{c}}_{\rm red}}$$





[PRD 103 (2021) 096005; see also arXiv:2309.11269 and NNPDF, in preparation]

Impact of future data: FPF



[arXiv:2309.09581]

2. Theory

N^3LO QCD corrections in PDF determination

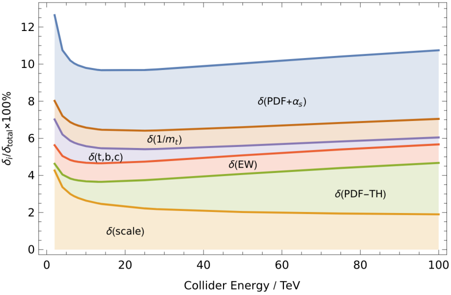
NNLO is the precision frontier for PDF determination

N3LO is the precision frontier for partonic cross sections

Mismatch between perturbative order of partonic cross sections and accuracy of PDFs is becoming a significant source of uncertainty

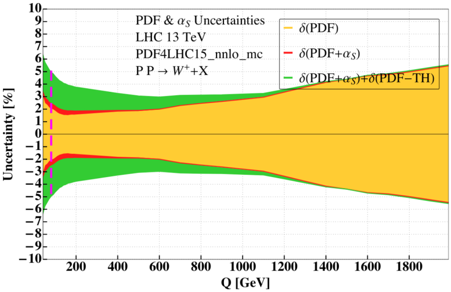
$$\hat{\sigma} = \alpha_s^p \hat{\sigma}_0 + \alpha_s^{p+1}\hat{\sigma}_1 + \alpha_s^{p+2}\hat{\sigma}_2 + \mathcal{O}(\alpha_s^{p+3}) \qquad \delta(\mathrm{PDF-TH}) = \frac{1}{2}\left|\frac{\sigma^{(2)}_{\mathrm{NNLO-PDFs}} - \sigma^{(2)}_{\mathrm{NLO-PDFs}}}{\sigma^{(2)}_{\mathrm{NNLO-PDFs}}}\right|$$

Higgs production in gluon-gluon fusion



[CERN Yellow Rep.Monogr. 7 (2019) 221]

W^+ boson production in CC Drell-Yan



[JHEP 11 (2020) 143]

N^3LO QCD corrections in PDF determination

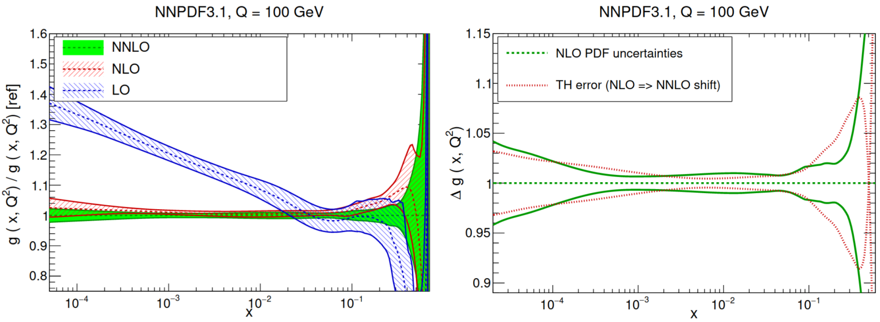
NNLO is the precision frontier for PDF determination

N3LO is the precision frontier for partonic cross sections

Mismatch between perturbative order of partonic cross sections and accuracy of PDFs is becoming a significant source of uncertainty

$$\hat{\sigma} = \alpha_s^p \hat{\sigma}_0 + \alpha_s^{p+1}\hat{\sigma}_1 + \alpha_s^{p+2}\hat{\sigma}_2 + \mathcal{O}(\alpha_s^{p+3}) \qquad \delta(\mathrm{PDF}-\mathrm{TH}) = \frac{1}{2}\left|\frac{\sigma^{(2)}_{\mathrm{NNLO-PDFs}} - \sigma^{(2)}_{\mathrm{NLO-PDFs}}}{\sigma^{(2)}_{\mathrm{NNLO-PDFs}}}\right|$$

Perturbative stability and uncertainty of the gluon PDF



[EPJ C77 (2017) 663]

N^3LO QCD corrections in PDF determination

Splitting Functions

Singlet (P_{qq}, P_{gg}, P_{gq}, P_{qg})

– large-n_f limit [NPB 915 (2017) 335; arXiv:2308.07958]

– small-x limit [JHEP 06 (2018) 145]

– large-x limit [NPB 832 (2010) 152; JHEP 04 (2020) 018; JHEP 09 (2022) 155]

– 5 (10) lowest Mellin moments [PLB 825 (2022) 136853; ibid. 842 (2023) 137944; ibid. 846 (2023) 138215]

Non-singlet ($P_{NS,v}$, $P_{NS,+}$, $P_{NS,-}$)

– large-n_f limit [NPB 915 (2017) 335; arXiv:2308.07958]

– small-x limit [JHEP 08 (2022) 135]

– large-x limit [JHEP 10 (2017) 041]

– 8 lowest Mellin moments [JHEP 06 (2018) 073]

DIS structure functions (F_L, F_2, F_3)

– DIS NC (massless) [NPB 492 (1997) 338; PLB 606 (2005) 123; NPB 724 (2005) 3]

– DIS CC (massless) [Nucl.Phys.B 813 (2009) 220]

– massive from parametrisation combining known limits and damping functions [NPB 864 (2012) 399]

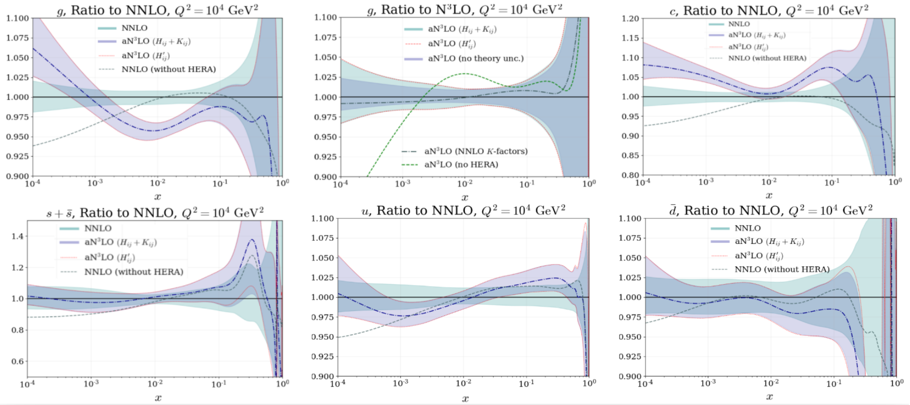
PDF matching conditions

– all known except for $a^3_{H,g}$ [NPB 820 (2009) 417; NPB 886 (2014) 733; JHEP 12 (2022) 134]

Coefficient functions for other processes

– DY (inclusive) [JHEP 11 (2020) 143]; DY (y differential) [PRL 128 (2022) 052001]

aN^{3}LO PDFs — MSHT



[EPJ C83 (2023) 185; see also T. Cridge's poster flash talk]

3-5% correction on the gluon PDF at $x \sim 10^{-2}$

larger charm PDF (perturbatively generated)

inclusion of theory uncertainties may inflate PDF uncertainties at small x

inclusion of aN^{3}LO corrections generally improve the χ^2 of HERA and LHC jets

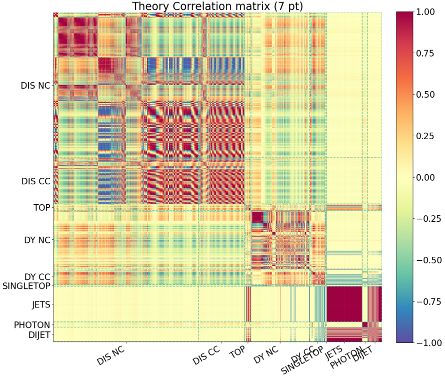
Theory uncertainties in PDF determination

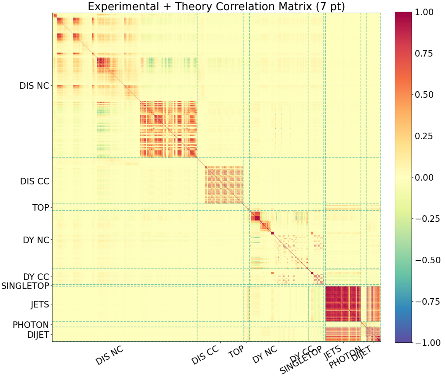
Assuming that theory uncertainties are (a) Gaussian and (b) independent from experimental uncertainties, modify the figure of merit to account for theory errors

$$\chi^2 = \sum_{i,j}^{N_{\rm dat}} (D_i - T_i)(\mathrm{cov}_{\rm exp} + \mathrm{cov}_{\rm th})^{-1}_{ij}(D_j - T_j);\ (\mathrm{cov}_{\rm th})_{ij} = \frac{1}{N}\sum_k^N \Delta_i^{(k)}\Delta_j^{(k)};\ \Delta_i^{(k)} \equiv T_i^{(k)} - T_i$$

Problem reduced to estimate the th. cov. matrix, *e.g.* in terms of nuisance parameters

$$\Delta_i^{(k)} = T_i(\mu_R, \mu_F) - T_i(\mu_{R,0}, \mu_{F,0});\ \text{vary scales in } \frac{1}{2} \leq \frac{\mu_F}{\mu_{F,0}}, \frac{\mu_R}{\mu_{R,0}} \leq 2$$





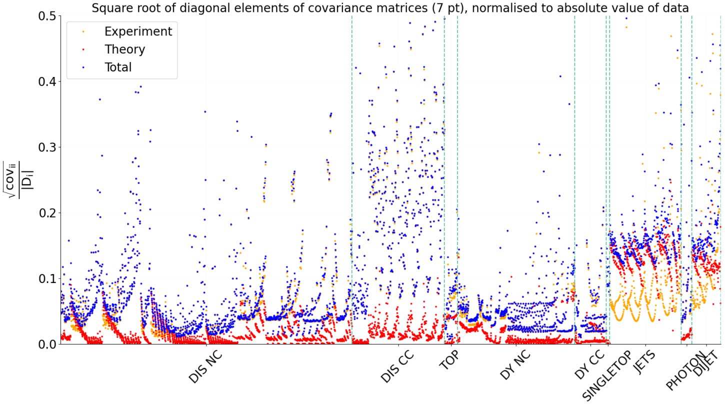
Theory uncertainties in PDF determination

Assuming that theory uncertainties are (a) Gaussian and (b) independent from experimental uncertainties, modify the figure of merit to account for theory errors

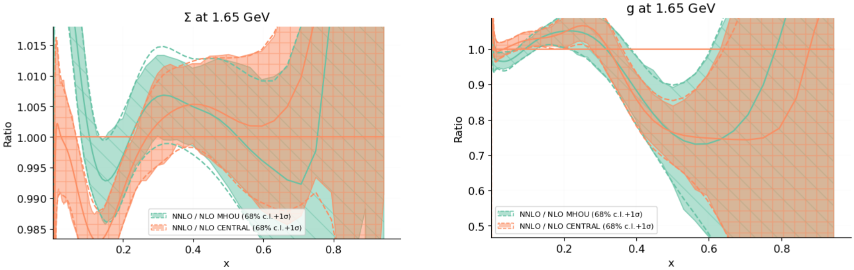
$$\chi^2 = \sum_{i,j}^{N_{\rm dat}} (D_i - T_i)(\mathrm{cov}_{\rm exp} + \mathrm{cov}_{\rm th})^{-1}_{ij}(D_j - T_j);\ (\mathrm{cov}_{\rm th})_{ij} = \frac{1}{N}\sum_k^N \Delta_i^{(k)}\Delta_j^{(k)};\ \Delta_i^{(k)} \equiv T_i^{(k)} - T_i$$

Problem reduced to estimate the th. cov. matrix, *e.g.* in terms of nuisance parameters

$$\Delta_i^{(k)} = T_i(\mu_R, \mu_F) - T_i(\mu_{R,0}, \mu_{F,0});\ \text{vary scales in } \frac{1}{2} \leq \frac{\mu_F}{\mu_{F,0}}, \frac{\mu_R}{\mu_{R,0}} \leq 2$$



Theory uncertainties in PDF determination



Faster perturbative convergence when MHOU are incorporated into PDFs

Overall (rather small) increase in uncertainties

Increase in PDF uncertainties due to replica generation
is counteracted by extra correlations in fitting minimisation

Tensions relieved: improvement in χ^2
exp only: $\chi^2/N_{\rm dat} = 1.21$ exp+th: $\chi^2/N_{\rm dat} = 1.20$

Data whose theoretical description is affected by large scale uncertainties
are deweighted in favour of more perturbatively stable data

[EPJ C79 (2019) 838; ibid. 931; NNPDF in preparation]

What happens at N3LO?

Incomplete higher order uncertainties

Approximate N^{3}LO splitting functions as

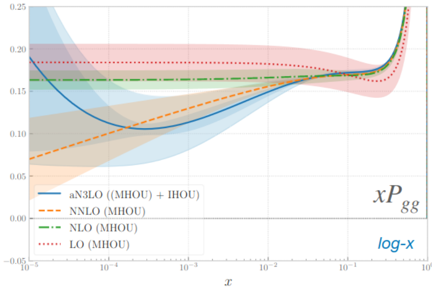
$$\gamma_{ij}^{(3)} = \gamma_{ij,n_f^3}^{(3)} + \gamma_{ij,N\to\infty}^{(3)} + \gamma_{ij,N\to 0}^{(3)} + \tilde{\gamma}_{ij}^{(3)}$$

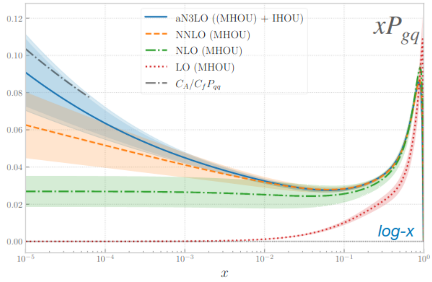
Parametrise $\tilde{\gamma}_{ij}^{(3)} = \sum_l a_{ij}^{(l)} G_l(N)$

– G_1 for the leading unknown large-N term
– G_2 for the leading unknown small-N term
– 3 or 8 G_l for the sub-leading unknown small- and large-N contributions
– vary the functions G_l to generate a variety of approximations and estimate IHOU
– determine the coefficients $a_{ij}^{(l)}$ with known moments and momentum conservation

Adopted basis function for $\tilde{\gamma}_{qg}^{(3)}$

$G_1(N)$	$\mathcal{M}[(1-x)\ln^2(1-x)]$
$G_2(N)$	$-\frac{1}{(N-1)^2} + \frac{1}{N^2}$
$G_3(N)$	$\frac{1}{N^4},\ \frac{1}{N^3},\ \mathcal{M}[(1-x)\ln(1-x)]$ $\mathcal{M}[(1-x)^2\ln(1-x)^2],\ \frac{1}{N-1}-\frac{1}{N},\ \mathcal{M}[(1-x)\ln(x)]$
$G_4(N)$	$\mathcal{M}[(1-x)(1+2x)],\ \mathcal{M}[(1-x)x^2],$ $\mathcal{M}[(1-x)x(1+x)],\ \mathcal{M}[(1-x)]$





[arXiv:2306.15294; NNPDF, in preparation]

Incomplete higher order uncertainties

Approximate N^{3}LO splitting functions as

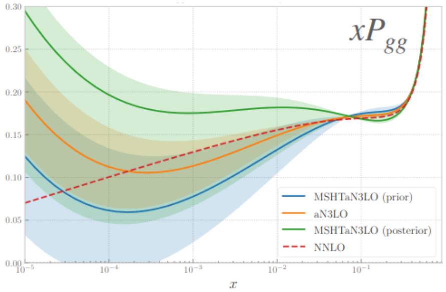
$$\gamma_{ij}^{(3)} = \gamma_{ij,n_f^3}^{(3)} + \gamma_{ij,N\to\infty}^{(3)} + \gamma_{ij,N\to 0}^{(3)} + \tilde{\gamma}_{ij}^{(3)}$$

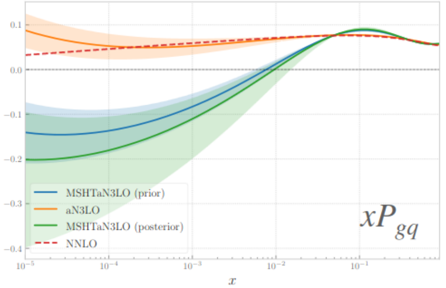
Parametrise $\tilde{\gamma}_{ij}^{(3)} = \sum_l a_{ij}^{(l)} G_l(N)$

– G_1 for the leading unknown large-N term
– G_2 for the leading unknown small-N term
– 3 or 8 G_l for the sub-leading unknown small- and large-N contributions
– vary the functions G_l to generate a variety of approximations and estimate IHOU
– determine the coefficients $a_{ij}^{(l)}$ with known moments and momentum conservation

Adopted basis function for $\tilde{\gamma}_{qg}^{(3)}$

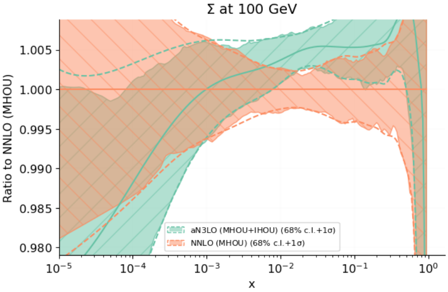
$G_1(N)$	$\mathcal{M}[(1-x)\ln^2(1-x)]$
$G_2(N)$	$-\frac{1}{(N-1)^2} + \frac{1}{N^2}$
$G_3(N)$	$\frac{1}{N^4}$, $\frac{1}{N^3}$, $\mathcal{M}[(1-x)\ln(1-x)]$ $\mathcal{M}[(1-x)^2\ln(1-x)^2]$, $\frac{1}{N-1}-\frac{1}{N}$, $\mathcal{M}[(1-x)\ln(x)]$
$G_4(N)$	$\mathcal{M}[(1-x)(1+2x)]$, $\mathcal{M}[(1-x)x^2]$, $\mathcal{M}[(1-x)x(1+x)]$, $\mathcal{M}[(1-x)]$

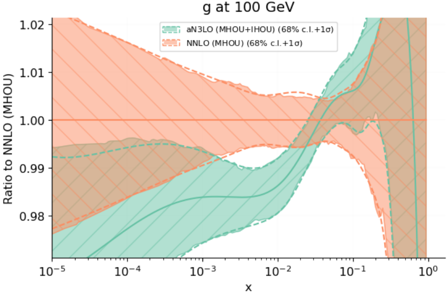




[arXiv:2306.15294; NNPDF, in preparation]

aN^{3}LO PDFs — NNPDF PRELIMINARY



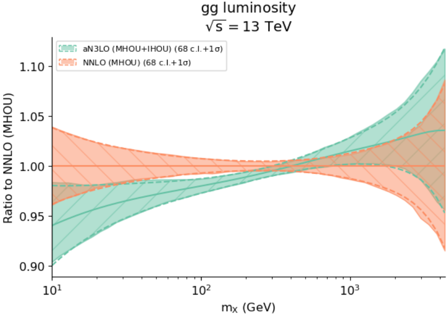


IHOU incorporated into
an independent covariance matrix
where nuisance parameters are averaged
over parametrisation variations

$\chi^2/N_{\rm dat} = 1.20$ (NNLO (MHOU))
$\chi^2/N_{\rm dat} = 1.19$ (aN^{3}LO (MHOU+IHOU))

PDFs only affected at small x

largest effect: 2% suppression in $\mathcal{L}_{gg}$
around the Higgs mass



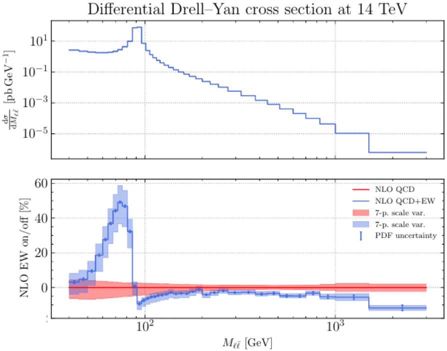
NLO EW corrections in PDF determination

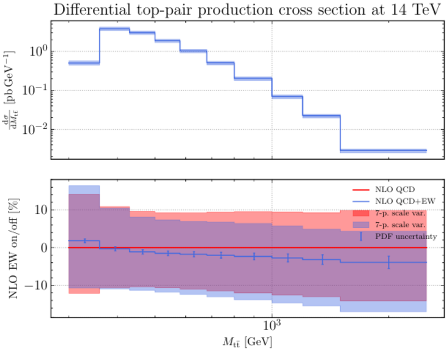
If we aim to PDF accurate to 1% NLO EW corrections do matter

especially as higher invariant mass and transverse momentum regions are accessed

Different approaches taken in general-purpose PDF fits

NLO EW K-factors (MSHT20); no NLO EW corrections by default (NNPDF4.0)





QED corrections in DGLAP evolution

[Com.Phys.Commun. 185 (2014) 1647]

Photon PDF

[PRL 117 (2016) 242002; JHEP 12 (2017) 046]

Photon PDF fits à la LuxQED

[SciPost Phys. 5 (2019) 1; JHEP 79 (2019) 10]

Automation of NLO EW corrections

[JHEP 07 (2018) 185]

Fast interpolation grids: PineAPPL

[JHEP 12 (2020 108]

Careful scrutiny of data

(no FSR nor photon-initiated subtraction)

3. Methodology

Validation of PDF uncertainties

Data region: closure tests

Fit PDFs to pseudodata generated
assuming a known underlying law

Define bias and variance

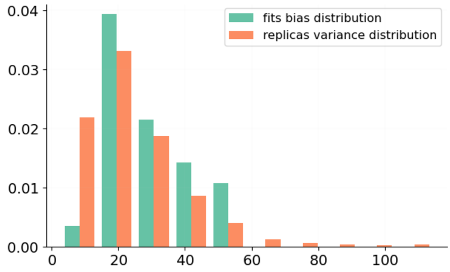
bias difference of central prediction and truth

variance uncertainty of replica predictions

If PDF uncertainty faithful, then

E[bias] = variance

25 fits, 40 replicas each



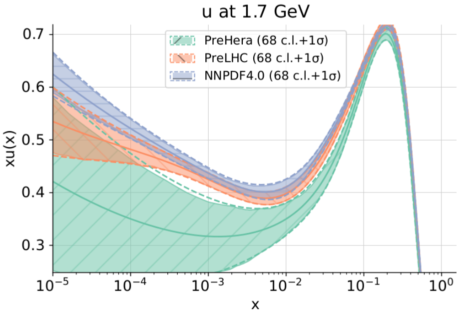
[EPJ C77 (2017) 663; EPJ C82 (2022) 330]

Extrapolation regions: future test

Test PDF uncertainties on data sets
not included in a given PDF fit
that cover unseen kinematic regions

Data set	NNPDF4.0	pre-LHC	pre-HERA
pre-HERA	1.09	1.01	0.90
pre-LHC	1.21	1.20	23.1
NNPDF4.0	1.29	3.30	23.1

Only exp. cov. matrix



[Acta Phys.Polon. B52 (2021) 243]

Validation of PDF uncertainties

Data region: closure tests

Fit PDFs to pseudodata generated assuming a known underlying law

Define bias and variance

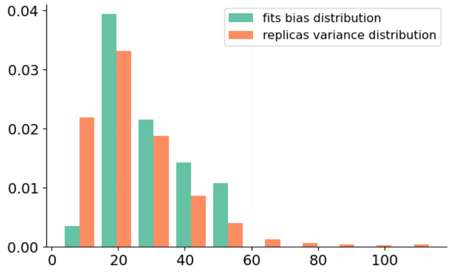
bias difference of central prediction and truth

variance uncertainty of replica predictions

If PDF uncertainty faithful, then

E[bias] = variance

25 fits, 40 replicas each



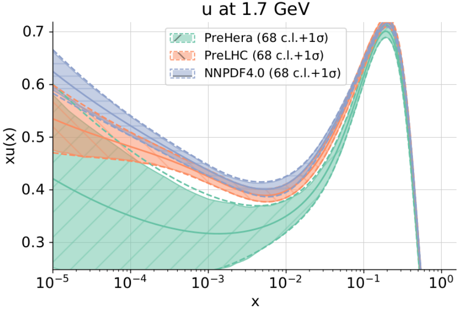
[EPJ C77 (2017) 663; EPJ C82 (2022) 330]

Extrapolation regions: future test

Test PDF uncertainties on data sets not included in a given PDF fit that cover unseen kinematic regions

Data set	NNPDF4.0	pre-LHC	pre-HERA
pre-HERA			0.86
pre-LHC		1.17	1.22
NNPDF4.0	1.12	1.30	1.38

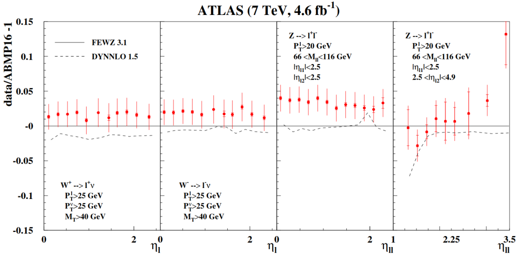
Exp+PDF cov. matrix



[Acta Phys.Polon. B52 (2021) 243]

Benchmarks

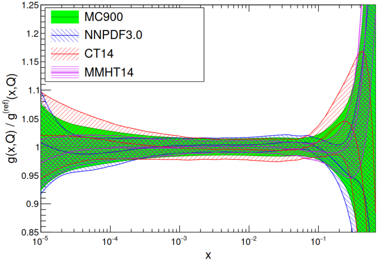
Benchmark of the theory



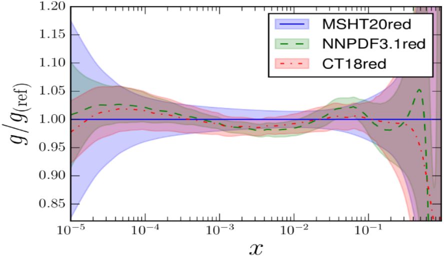
Be careful about the use of different NNLO codes for DY production in particular when experiments use non-optimal fiducial cuts [EPJ C81 (2021) 573]

NNLO corrections usually implemented via K-factors
NNLOJet/ApplFast provide NNLO lookup tables for a limited set of data

Benchmark of PDF sets



[PDF4LHC15 benchmark, JPG 43 (2016) 023001]



[PDF4LHC21 benchmark, JPG 49 (2022) 080501]

Tools

Progress has been possible thanks to the development of computational tools

Many of them are publicly available (and some are unfortunately not)

Monte Carlo generators
madgraph, MCFM, MATRIX, NNLOjet, ...

PDF evolution codes and DIS codes
Hoppet, Pegasus, QCDnum, APFEL, APFEL++, EKO, YADISM, ...

Fast-interpolation grid formats
APPLgrid, FastNLO, APPLfast, PineAPPL

Statistical tools for PDF manipulation
Monte Carlo compression, Hessian and Monte Carlo conversion

Tools to study the PDF sensitivity to data sets
L2 sensitivity, Hessian profiling, Bayesian reweighting, SMPDF

Fitting codes
XFitter, NNPDF

4. Conclusions

Summary

A precise and accurate determination of PDFs is key to do precision phenomenology.

LHC measurements are being instrumental to reduce PDF uncertainties to few percent.

The goal of achieving PDF determinations accurate to 1% opens up some challenges.

Understand experimental systematic uncertainties and their correlations.

Refine the theoretical accuracy of a PDF determination.

Represent theory uncertainties into PDF uncertainties.

Deploy a robust fitting methodology and good statistical tests of it.

Benchmark efforts may benefit from public releases of PDF codes and inputs.

Summary

A precise and accurate determination of PDFs is key to do precision phenomenology.

LHC measurements are being instrumental to reduce PDF uncertainties to few percent.

The goal of achieving PDF determinations accurate to 1% opens up some challenges.

Understand experimental systematic uncertainties and their correlations.

Refine the theoretical accuracy of a PDF determination.

Represent theory uncertainties into PDF uncertainties.

Deploy a robust fitting methodology and good statistical tests of it.

Benchmark efforts may benefit from public releases of PDF codes and inputs.

Thank you