Determining PDFs accurately and precisely: data, theory, and methodology

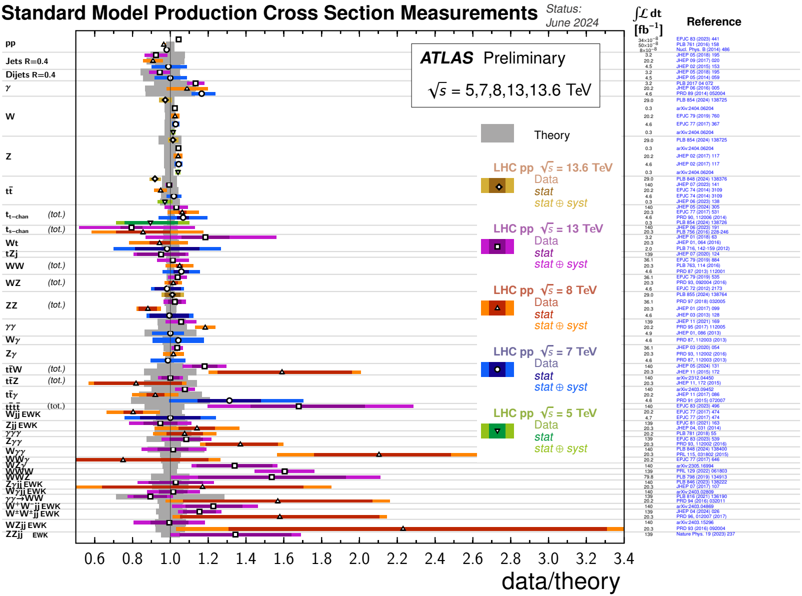
Workshop on High Luminosity LHC and Hadron Colliders

Emanuele R. Nocera

Università degli Studi di Torino and INFN, Torino

2 October 2024

Physics at the LHC as Precision Physics



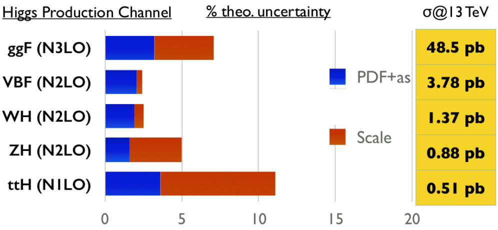
[Plot from ATLAS Collaboration web page]

Parton Distribution Functions 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

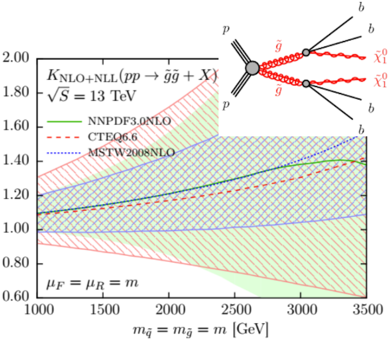
Precision



Unc. [MeV]	Total	Stat.	Syst.	PDF	A_i	Backg.	EW	e	μ	u_T	Lumi	Γ_W	PS
p_T^ℓ	16.2	11.1	11.8	4.9	3.5	1.7	5.6	5.9	5.4	0.9	1.1	0.1	1.5
m_T	24.4	11.4	21.6	11.7	4.7	4.1	4.9	6.7	6.0	11.4	2.5	0.2	7.0
Combined	15.9	9.8	12.5	5.7	3.7	2.0	5.4	6.0	5.4	2.3	1.3	0.1	2.3

[CERN Yellow Report 2016; arXiv:2403.15085]

Discovery



[EPJC 76 (2016) 53]

PDF determination in statistical language

Inverse problem

Given a set of data D, determine $p(f|D)$ in the space of functions $f : [0,1] \to \mathbb{R}$.

Solution: parametric regression

Approximate $p(f|D)$ with its projection in the space of parameters $p(\boldsymbol{\theta}|D)$

$$xf_i(x, Q_0^2) = A_{f_i}\, x^{a_{f_i}}\, (1-x)^{b_{f_i}}\, \mathscr{F}(x, \{c_{f_i}\})$$

Determine $p(\boldsymbol{\theta}|D) \propto p(D|\boldsymbol{\theta})p(\boldsymbol{\theta})$ as MAP $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|D)$

$$\chi^2 = \sum_{i,j}^{N_{\rm dat}} [T_i[\boldsymbol{\theta}] - D_i](\mathrm{cov}^{-1})_{ij}[T_j[\boldsymbol{\theta}] - D_j]$$

Use a prescription to compute expectation values and uncertainties of observables

$$E[\mathcal{O}] = \int \mathcal{D}f \mathcal{P}(f|D)\mathcal{O}(f) \qquad V[\mathcal{O}] = \int \mathcal{D}f \mathcal{P}(f|D)[\mathcal{O}(f) - E[\mathcal{O}]]^2$$

Monte Carlo: $\mathcal{P}(f|D) \longrightarrow \{f_k\}$

$E[\mathcal{O}] \approx \frac{1}{N}\sum_k \mathcal{O}(f_k)$

$V[\mathcal{O}] \approx \frac{1}{N}\sum_k [\mathcal{O}(f_k) - E[\mathcal{O}]]^2$

Maximum likelihood: $\mathcal{P}(f|D) \longrightarrow f_0$

$E[\mathcal{O}] \approx \mathcal{O}(f_0)$

$V[\mathcal{O}] \approx$ Hessian, $\Delta\chi^2$envelope, ...

Interplay between DATA, THEORY, and METHODOLOGY

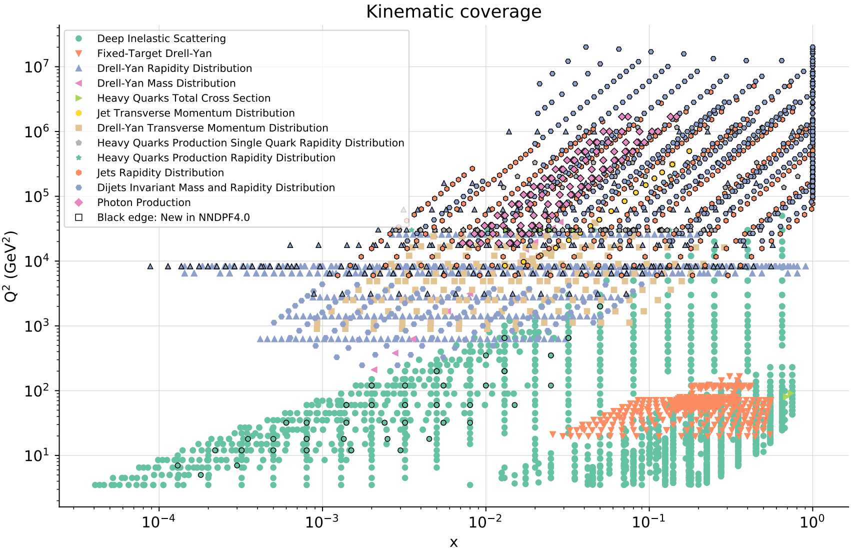
Overview of current PDF determinations

	NNPDF4.0	MSHT20	CT18	HERAPDF2.0	CJ22	ABMP16
Fixed-target DIS	☑	☑	☑	☒	☑	☑
JLAB	☒	☒	☒	☒	☑	☒
HERA I+II	☑	☑	☑	☑	☑	☑
HERA jets	☑	☒	☒	☑	☒	☒
Fixed target DY	☑	☑	☑	☒	☑	☑
Tevatron W, Z	☑	☑	☑	☒	☑	☑
LHC vector boson	☑	☑	☑	☒	☑	☑
LHC $W+c$ $Z+c$	☑	☒	☒	☒	☒	☒
Tevatron jets	☑	☑	☑	☒	☑	☒
LHC jets	☑	☑	☑	☒	☒	☒
LHC top	☑	☑	☒	☒	☒	☑
LHC single t	☑	☒	☒	☒	☒	☒
LHC prompt γ	☑	☒	☒	☒	☒	☒
statistical treatment	Monte Carlo	Hessian $\Delta\chi^2$ dynamical	Hessian $\Delta\chi^2$ dynamical	Hessian $\Delta\chi^2 = 1$	Hessian $\Delta\chi^2 = 1.645$	Hessian $\Delta\chi^2 = 1$
parametrisation	Neural Network	Chebyschev pol.	Bernstein pol.	polynomial	polynomial	polynomial
HQ scheme	FONLL	TR′	ACOT-χ	TR′	ACOT-χ	FFN
accuracy	aN^{3}LO	aN^{3}LO	NNLO	NNLO	NLO	NNLO
latest update	EPJ C82 (2022) 428	EPJ C81 (2021) 341	PRD 103 (2021) 014013	EPJ C82 (2022) 243	PRD 107 (2023) 113005	PRD 96 (2017) 014011

All PDF sets are available as (x, Q^2) interpolation grids through the LHAPDF library

1. Data

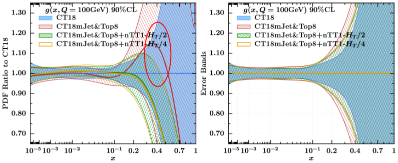
Overview of experimental data



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

Gluon

Global fit without jet and $t\bar{t}$ data vs CT18NNLO



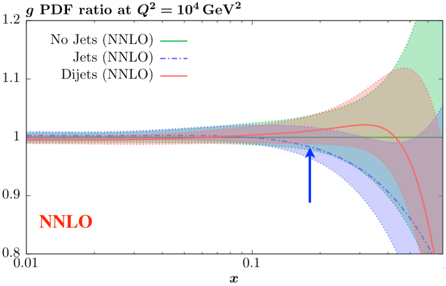
[M. Guzzi, PDF4LHC Nov. 2023]

Various processes (included in all PDF sets)
Z p_T, jets, di-jets, $t\bar{t}$

Largest impact of jets/di-jets at large x

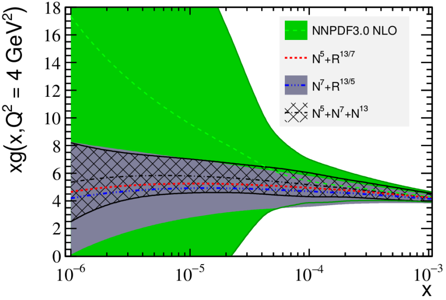
Di-jets preferred over single-inclusive jets

Forward charm production impacts small x potentially crucial for UHE neutrino-nucleus cross section measurements



[L. Harland-Lang, PDF4LHC Nov. 2023]

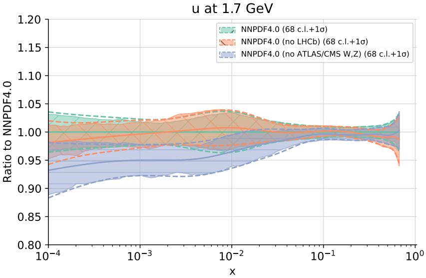
[See also EPJ C80 (2020) 797]

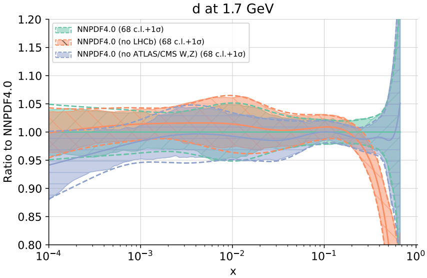


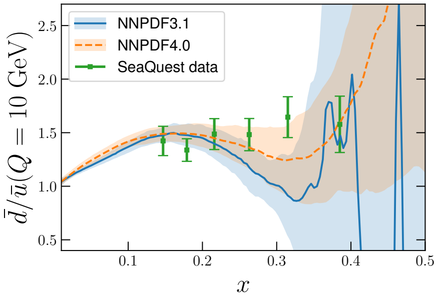
[PRL 118 (2017) 072001]

[See also PRD 109 (2024) 113001]

Quark flavour separation







Relative impact of ATLAS/CMS/LHCb gauge boson production

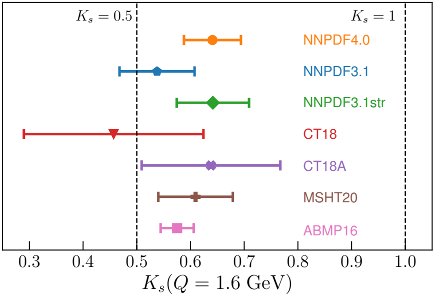
LHCb is at forward rapidity

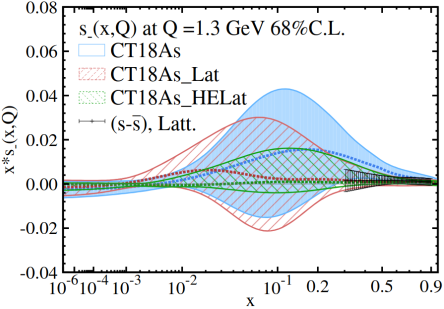
New constraint on $\bar{d}/\bar{u}$ ratio from SeaQuest

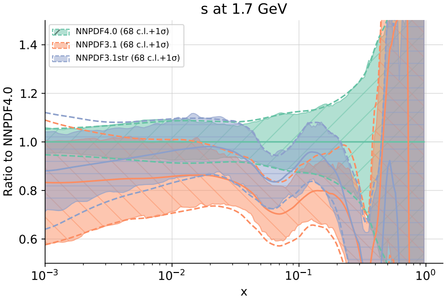
[Nature 590 (2021) 561]

Studied by CT, MSHT, NNPDF, ABMP
Some tension with NuSea found

Strange







Good consistency of K_s across PDF sets

$$K_s(Q^2) = \frac{\int_0^1 dx[s(x,Q^2)+\bar{s}(x,Q^2)]}{\int_0^1 dx[\bar{u}(x,Q^2)+\bar{d}(x,Q^2)]}$$

Effect of data and nuclear uncertainties

ATLAS W, Z and W+jet data enhance s

NOMAD data reduce uncertainties

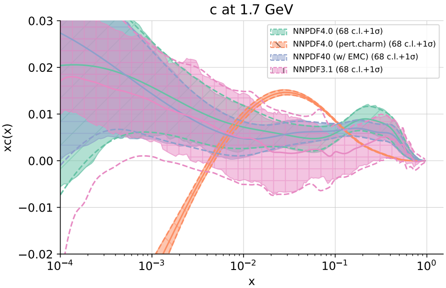
nuclear uncertainties accommodate data sets

Useful input from lattice QCD

[EPJ C80 (2020) 1168; PRD 107 (2023) 076018]

[See also PRD 91 (2015) 094002]

Charm



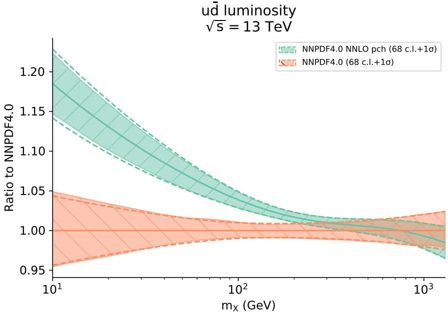
Perturbative charm alters the flavour decomposition and deteriorates the fit

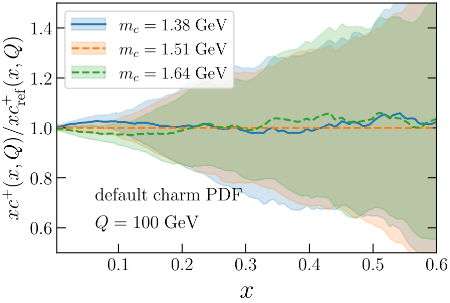
$$\chi^2_{\text{fitted charm}} = 1.17 \to \chi^2_{\text{pert. charm}} = 1.19$$

mainly due to a worsening of the LHC W, Z and top pair data sets

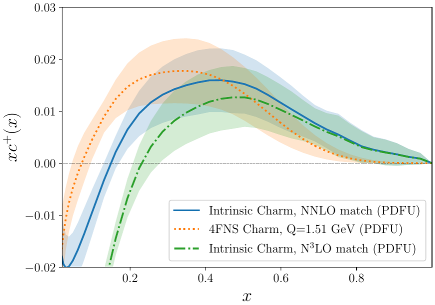
fitting charm reduces the dependence from m_c

[EPJ C76 (2016) 647; C77 (2017) 663; C82 (2022) 428]





Intrinsic Charm



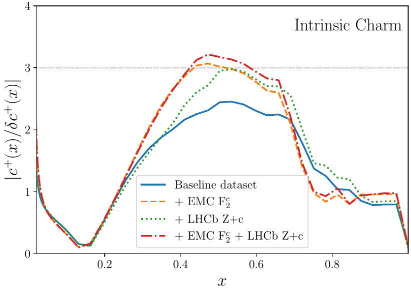
Evolve results backwards (below m_c) with N^{3}LO matching

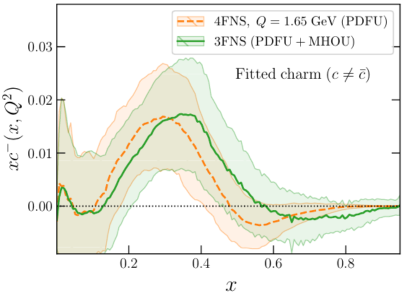
Evidence of intrinsic charm and of $c - \bar{c}$ shape compatible with models

[Nature 608 (2022) 483; arXiv:2311.00743]

Evidence enhanced by EMC F_2^c and $Z + D$

Challenged by CT18 [PLB 843 (2023) 137975]





2. Theory

Perturbative accuracy in PDF determination

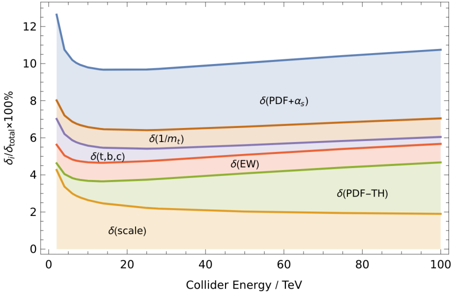
NNLO is the precision frontier for PDF determination

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

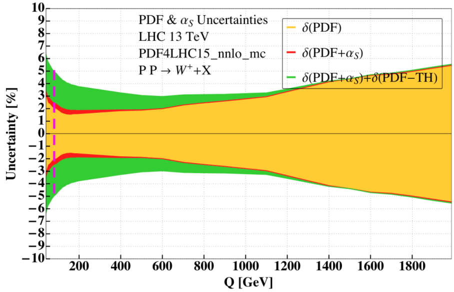
$$\hat{\sigma} = \alpha_s^p \hat{\sigma}_0 + \alpha_s^{p+1} \hat{\sigma}_1 + \alpha_s^{p+2} \hat{\sigma}_2 + \mathcal{O}(\alpha_s^{p+3}) \qquad \delta(\mathrm{PDF-TH}) = \frac{1}{2} \left| \frac{\sigma^{(2)}_{\mathrm{NNLO-PDFs}} - \sigma^{(2)}_{\mathrm{NLO-PDFs}}}{\sigma^{(2)}_{\mathrm{NNLO-PDFs}}} \right|$$

Higgs production in gluon-gluon fusion



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

W^+ boson production in CC Drell-Yan



[JHEP 11 (2020) 143]

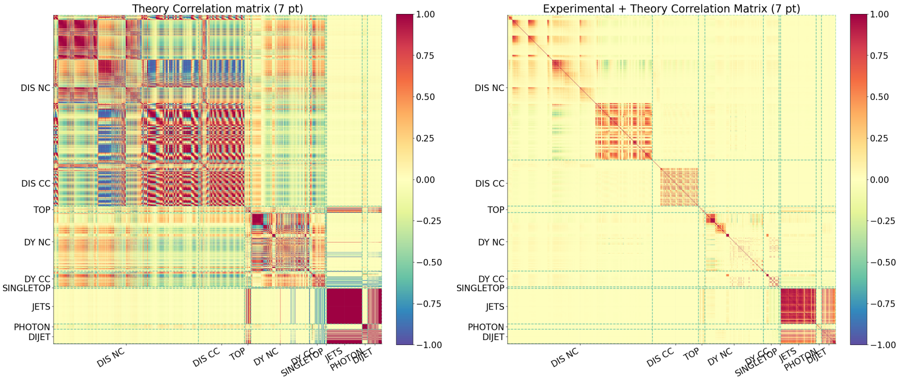
Theory uncertainties in PDF determination

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

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



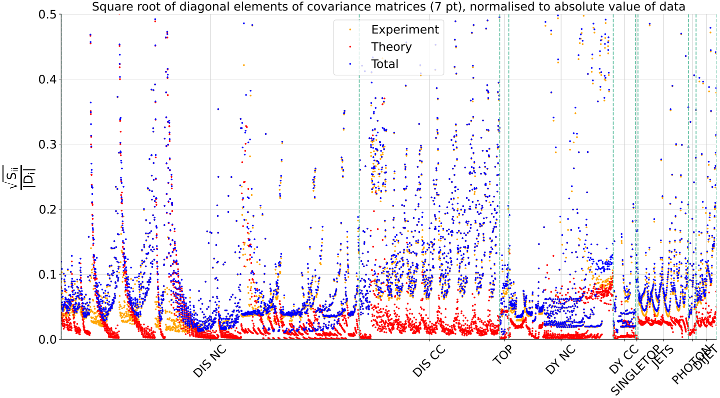
Theory uncertainties in PDF determination

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

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



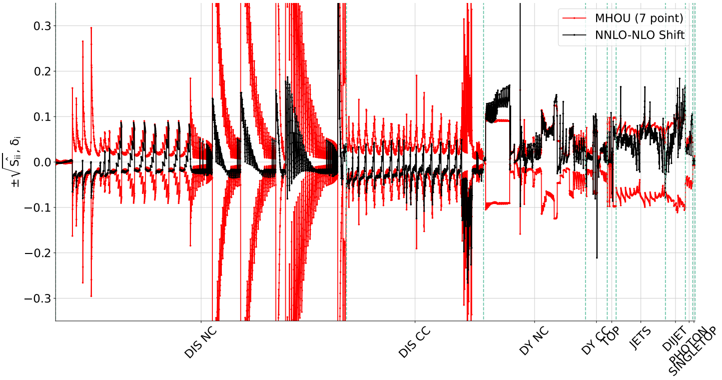
Theory uncertainties in PDF determination

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

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

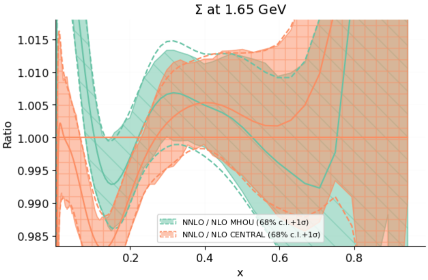
Problem reduced to estimate the th. cov. matrix, *e.g.* in terms of nuisance parameters

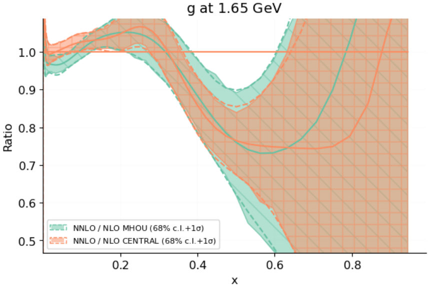
$$\Delta_i^{(k)} = T_i(\mu_R, \mu_F) - T_i(\mu_{R,0}, \mu_{F,0});\ \text{vary scales in } \frac{1}{2} \leq \frac{\mu_F}{\mu_{F,0}}, \frac{\mu_R}{\mu_{R,0}} \leq 2$$



Impact on fit quality and PDFs

Dataset	$N_{\rm dat}$	NLO no MHOU	NLO MHOU	NNLO no MHOU	NNLO MHOU
DIS NC	2100	1.30	1.22	1.23	1.20
DIS CC	989	0.92	0.87	0.90	0.90
DY NC	736	2.01	1.71	1.20	1.15
DY CC	157	1.48	1.42	1.48	1.37
Top pairs	64	2.08	1.24	1.21	1.43
Single-inclusive jets	356	0.84	0.82	0.96	0.81
Dijets	144	1.52	1.84	2.04	1.71
Prompt photons	53	0.59	0.49	0.75	0.67
Single top	17	0.36	0.35	0.36	0.38
Total	4616	1.34	1.23	1.17	1.13





[EPJ C79 (2019) 838; ibid. 931; EPJ C84 (2024) 517]

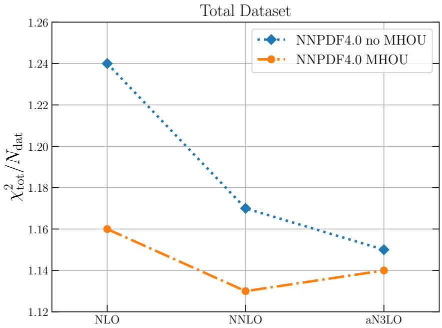
What happens at aN^{3}LO?

Dataset	NLO $N_{\rm dat}$	NLO no MHOU	NLO MHOU	NNLO $N_{\rm dat}$	NNLO no MHOU	NNLO MHOU	aN^{3}LO $N_{\rm dat}$	aN^{3}LO no MHOU	aN^{3}LO MHOU
DIS NC	1980	1.30	1.22	2100	1.22	1.20	2100	1.22	1.20
DIS CC	988	0.92	0.87	989	0.90	0.90	989	0.91	0.92
DY NC	667	1.49	1.32	736	1.20	1.15	736	1.17	1.16
DY CC	193	1.31	1.27	157	1.45	1.37	157	1.37	1.36
Top pairs	64	1.90	1.24	64	1.27	1.43	64	1.23	1.41
Single-inclusive jets	356	0.86	0.82	356	0.94	0.81	356	0.84	0.83
Dijets	144	1.55	1.81	144	2.01	1.71	144	1.78	1.67
Prompt photons	53	0.58	0.47	53	0.76	0.67	53	0.72	0.68
Single top	17	0.35	0.34	17	0.36	0.38	17	0.35	0.36
Total	4462	1.24	1.16	4616	1.17	1.13	4616	1.15	1.14

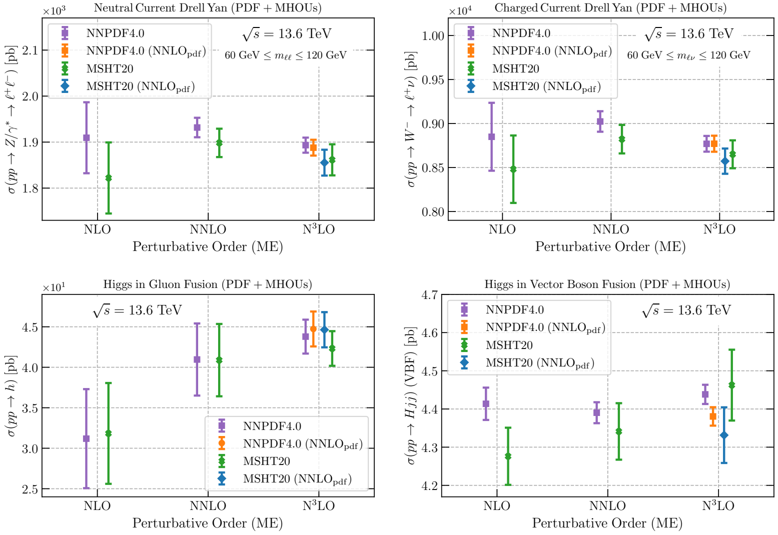
Fit quality improves with perturbative order

Fit quality almost independent from perturbative order when MHOU are included

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



Impact on Inclusive Cross Sections



Effect of using aN^{3}LO PDFs instead of NNLO PDFs in N^{3}LO predictions is small

Good consistency between NNPDF4.0 [EPJ C84 (2024) 659] and MSHT20 [EPJ C83 (2023) 185]

The photon PDF and QED corrections

Photon PDF à la LuxQED

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

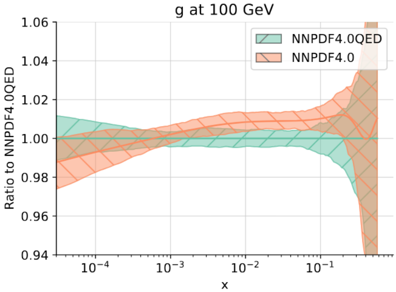
Fit quality unaltered: $\chi^2/N_{\rm dat} = 0.17$

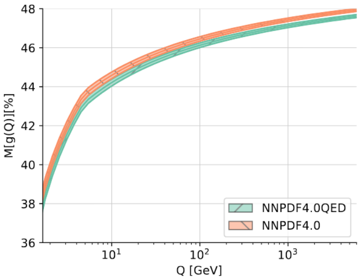
Small (0.5%) momentum shift from g to γ

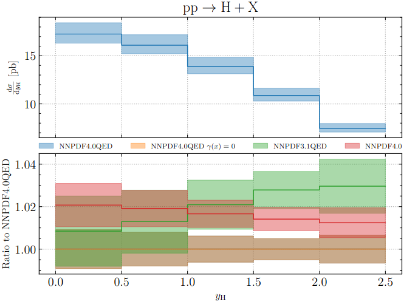
Small (1%) suppression of the gluon PDF

1-2% suppression in ggH cross section

See, *e.g.* EPJ C84 (2024) 540

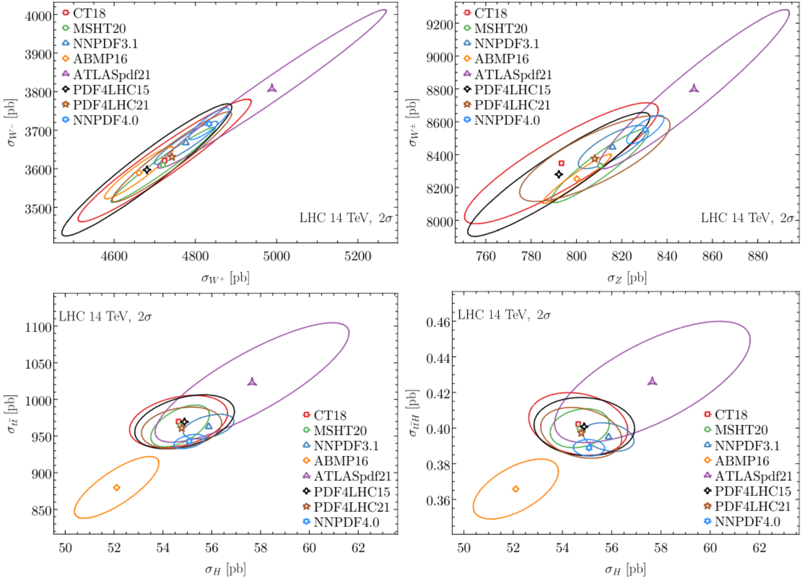






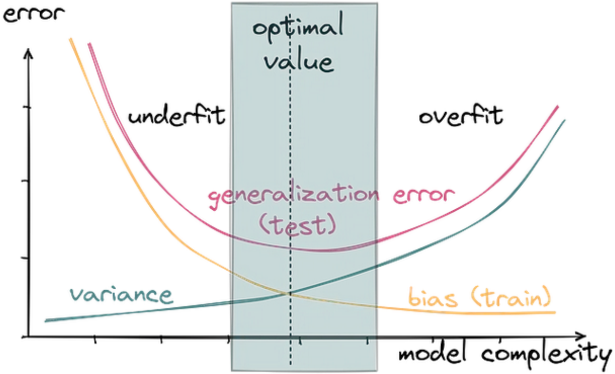
3. Methodology

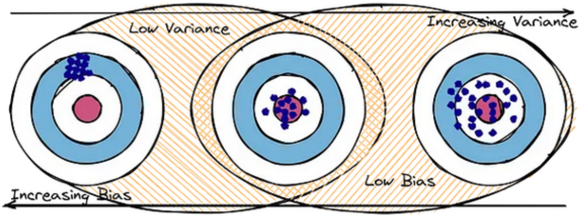
Making predictions with PDFs



[Acta Phys.Polon.B 53 (2022) 12]

Accuracy vs precision or bias vs variance





Validation of PDF uncertainties

Data region: closure tests

Fit PDFs to pseudodata generated assuming a known underlying law

Define bias and variance

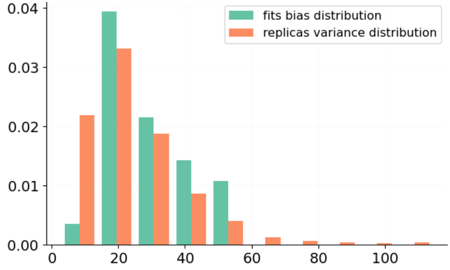
bias difference of central prediction and truth

variance uncertainty of replica predictions

If PDF uncertainty faithful, then

E[bias] = variance

25 fits, 40 replicas each



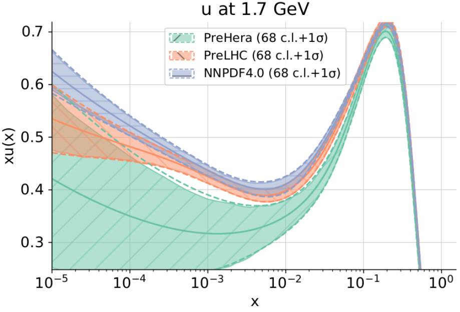
[EPJ C77 (2017) 663; EPJ C82 (2022) 330]

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]

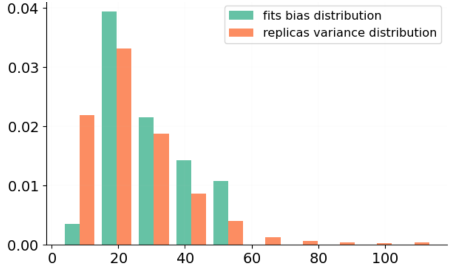
Validation of PDF uncertainties

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



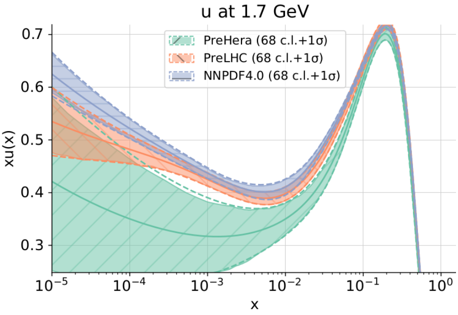
[EPJ C77 (2017) 663; EPJ C82 (2022) 330]

Extrapolation regions: future test

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

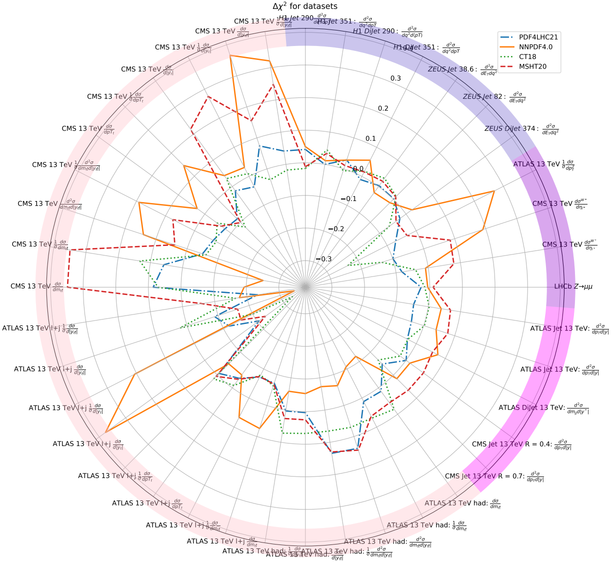
Data set	NNPDF4.0	pre-LHC	pre-HERA
pre-HERA			0.86
pre-LHC		1.17	1.22
NNPDF4.0	1.12	1.30	1.38

Exp+PDF cov. matrix



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

Are all PDF sets equally accurate?



$$\Delta\chi^2 = \frac{\chi^{2,(i)}_{\rm exp+mhou+pdf} - \left\langle \chi^2_{\rm exp+mhou+pdf} \right\rangle}{\left\langle \chi^2_{\rm exp+mho+pdf} \right\rangle}$$

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 the interplay between data, theory, and methodology into PDF uncertainties.

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 the interplay between data, theory, and methodology into PDF uncertainties.

Refine the theoretical accuracy of a PDF determination.

Represent theory uncertainties into PDF uncertainties.

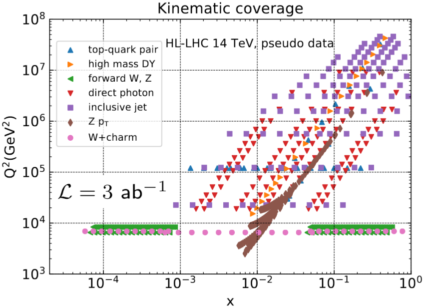
Deploy a robust fitting methodology and good statistical tests of it.

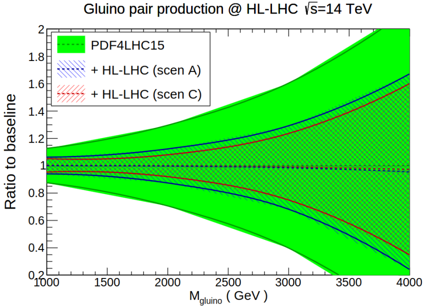
Benchmark efforts may benefit from public releases of PDF codes and inputs.

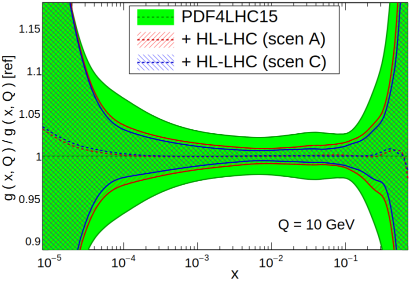
Thank you

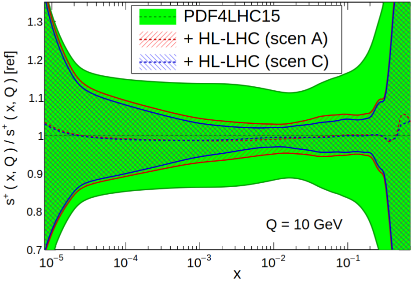
Appendix

Impact of future data: HL-LHC



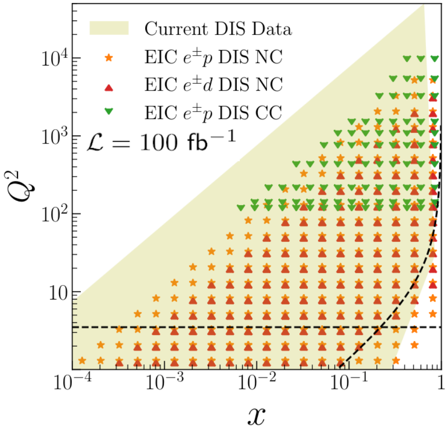


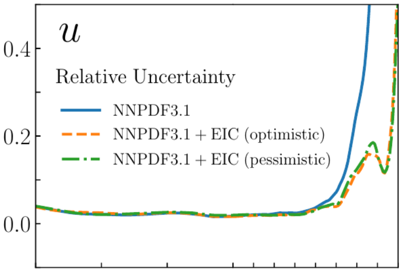


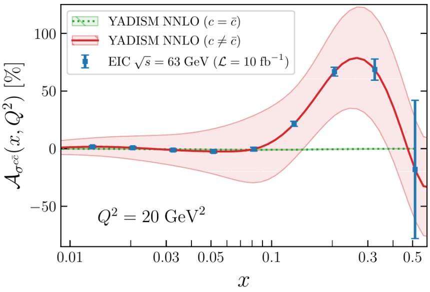


[EPJ.C 78 (2018) 962]

Impact of future data: EIC





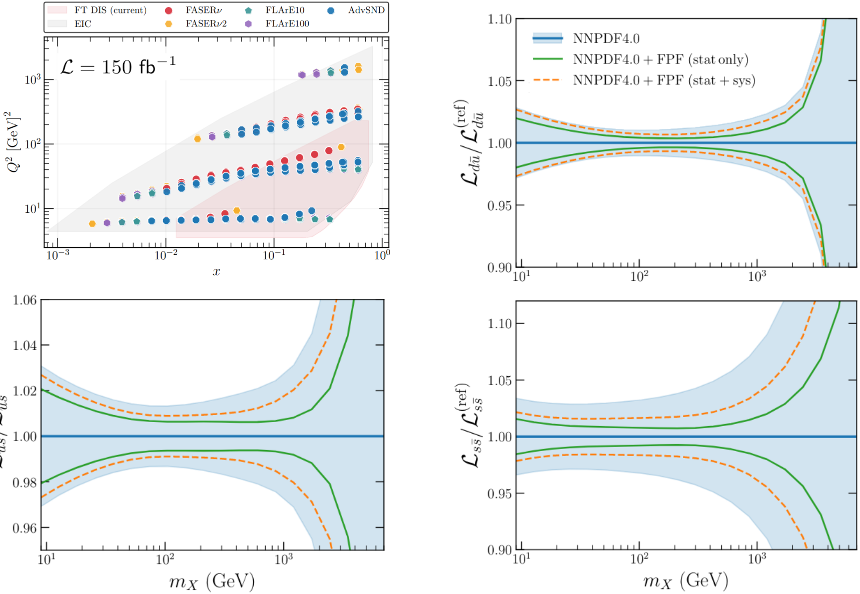


$E_\ell \times E_p$ [GeV]: 18× 275; 10× 100; 5× 100

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

[PRD 103 (2021) 096005; see also arXiv:; arXiv:2311.00743]

Impact of future data: FPF



[arXiv:2309.09581; see T. Mäkelä's talk]

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]

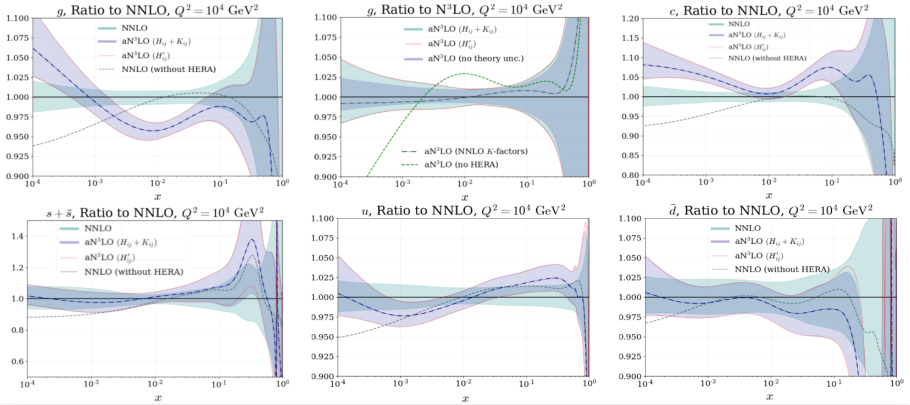
PDF matching conditions

– 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 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

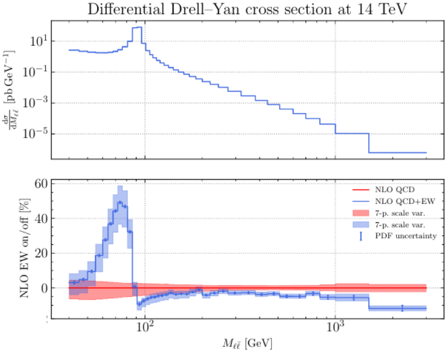
NLO EW corrections in PDF determination

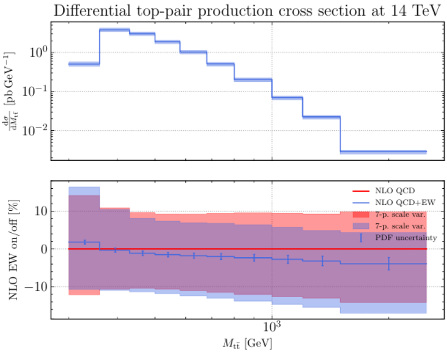
If we aim to PDF accurate to 1% NLO EW corrections do matter

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]

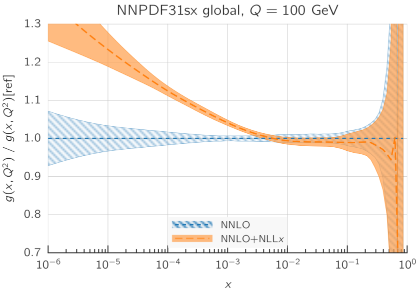
Photon PDF fits à la LuxQED
[SciPost Phys. 5 (2019) 1; JHEP 79 (2019) 10]

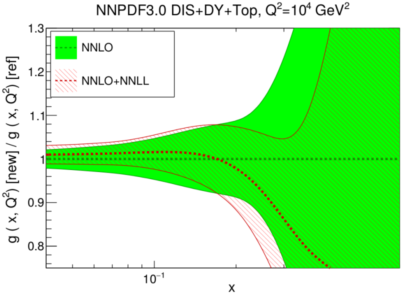
Automation of NLO EW corrections
[JHEP 07 (2018) 185]

Fast interpolation grids: PineAPPL
[JHEP 12 (2020 108]

Careful scrutiny of data
(no FSR nor photon-initiated subtraction)

Beyond fixed-order accuracy





small x: $\frac{1}{x}\ln^k x$

high-energy gluon emission: single logs

large x: $\left(\frac{\ln^k(1-x)}{(1-x)}\right)_+$

soft gluon emission: double logs

Large logs $\alpha_s \ln \sim 1$ spoil the convergence of the perturbative series

PDFs with threshold resummation [JHEP 1509 (2015) 191] (only DIS, DY Z/γ, total $t\bar{t}$ + evol.)

suppression in PDFs partially or totally compensates enhancements in partonic cross-sections

accuracy of the resummed fit competitive with the fixed-order fit, except for the large-x gluon

PDFs with high-energy (BFKL) resummation [EPJ C78 (2018) 321] (only DIS + evol.)

Resummed PDFs enhanced at small x, uncertainties reduced, fit quality improves

Large effects for future colliders, or b production at LHC

High-densitiy effects modelled in CT18X; similar outcome on PDFs and fit quality

Fitting away New Physics

DIS [PRL 123 (2019) 132001]

DY tails [JHEP 07 (2021) 122]

DIS/DY [JHEP 08 (2022) 088]

Jet/top [JHEP 05 (2023) 003]

Jets [JHEP 02 (2022) 142]

Many more analyses by ATLASs, CMS, ...