Recent Advances and Challenges in Proton PDF Determination

54th International Symposium on Multiparticle Dynamics (ISMD 2026)

Emanuele R. Nocera

Università degli Studi di Torino and INFN, Torino

19 August 2026







UNIVERSITÀ DI TORINO

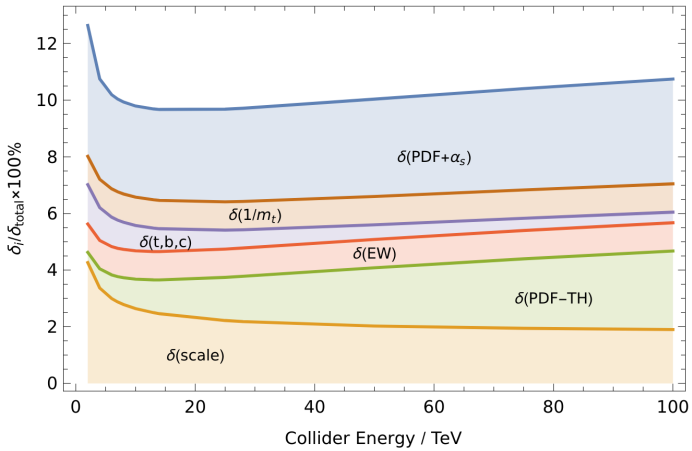
Parton Distribution Functions at the LHC

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

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

Precision

Higgs production in gluon-gluon fusion

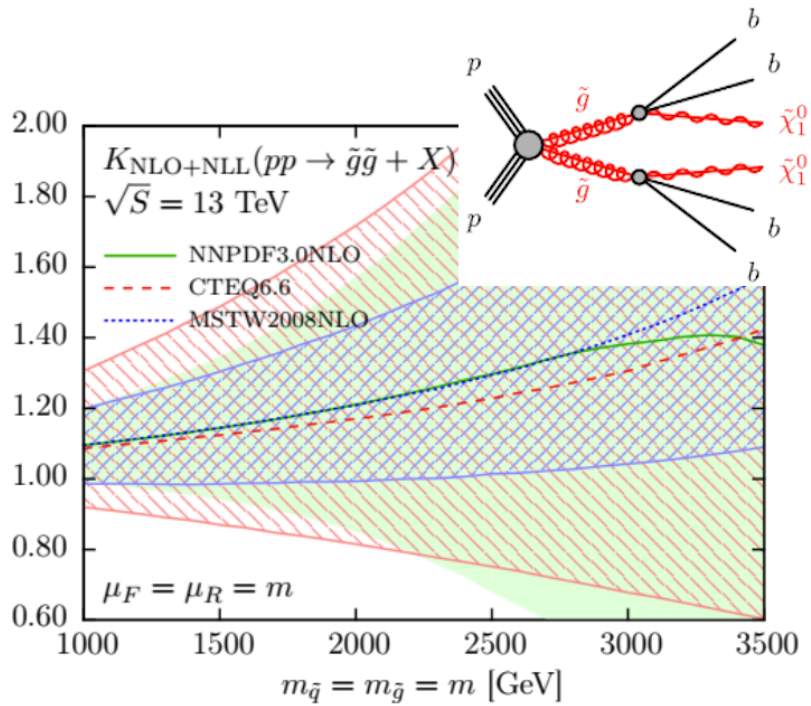


[CERN Yellow Rep.Monogr. 7 (2019) 221]

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

[EPJ C84 (2024) 1309]

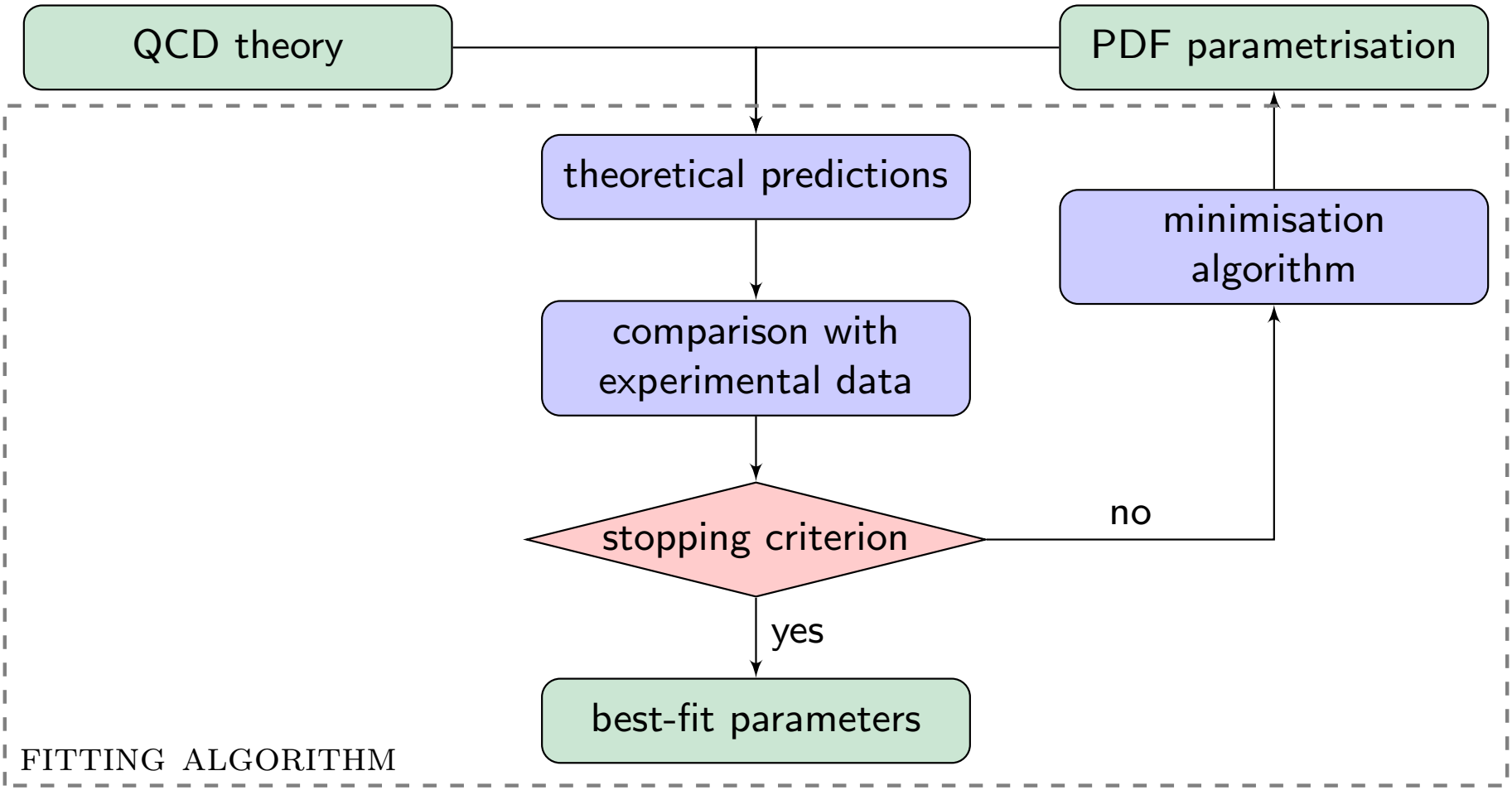
Discovery



[EPJC 76 (2016) 53]

Higgs boson characterisation, determination of SM parameters, BSM searches

Determining PDFs from (LHC) experimental data

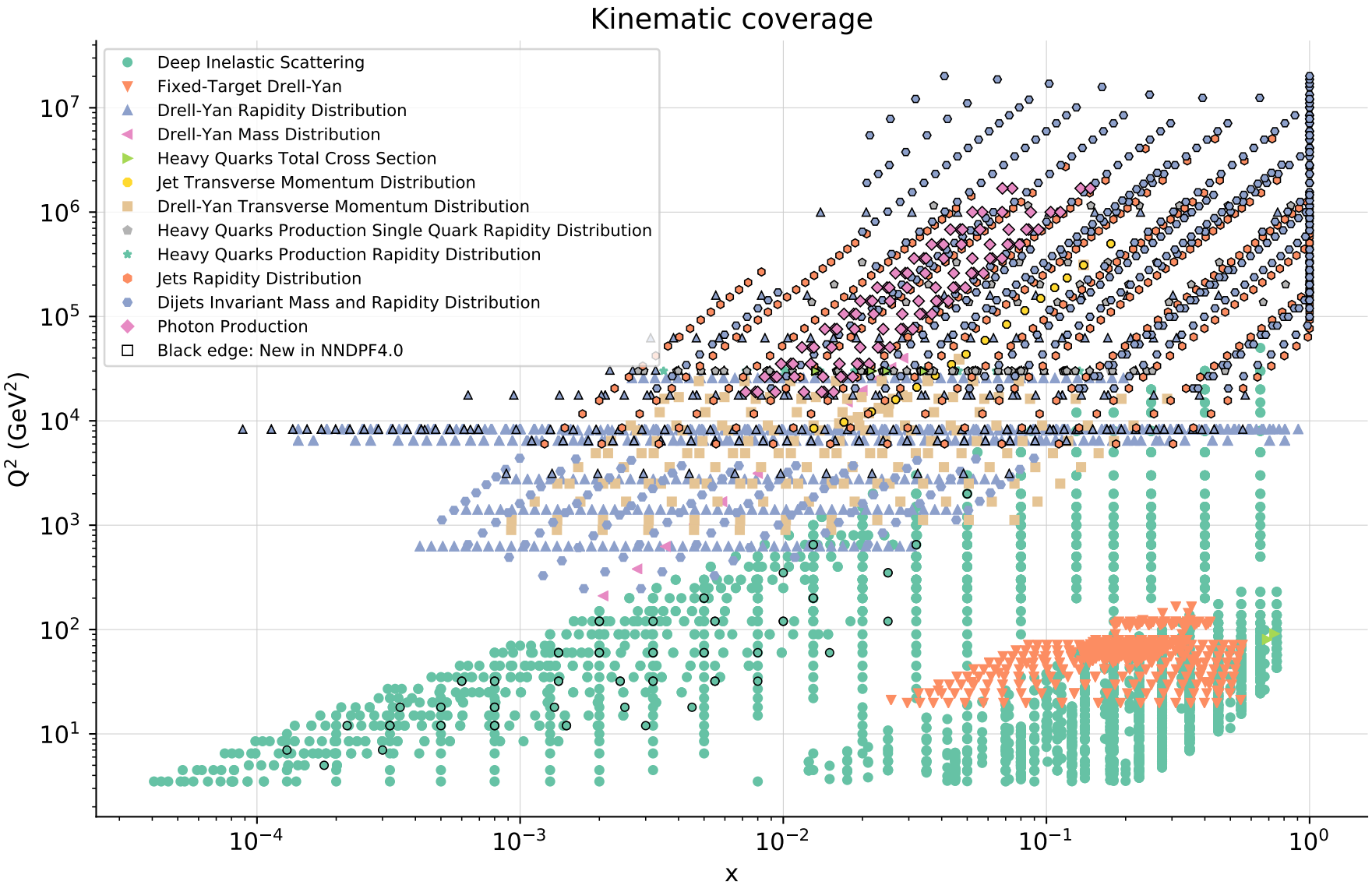


$$\chi^2 = \sum_{i,j}^{N_{\rm dat}} [T_i[\{\vec{a}\}] - D_i](\mathrm{cov}^{-1})_{ij}[T_j[\{\vec{a}\}] - D_j] \qquad \text{with } \{\vec{a}\} \text{ the set of parameters}$$

Interplay of DATA, THEORY, AND METHODOLOGY

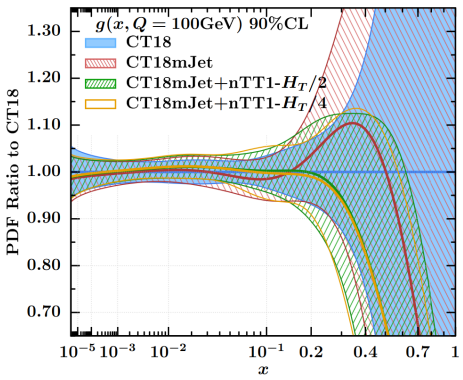
1. Data

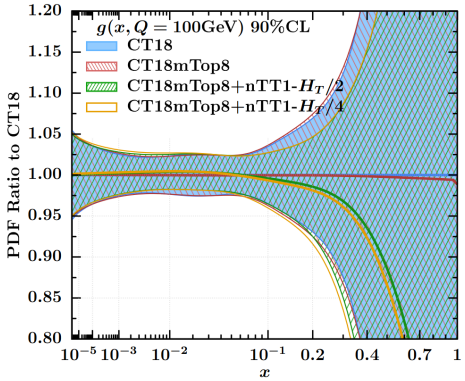
The data set



$N_{\rm dat} \sim 4618$ $N_{\rm proc} \sim 13$ $\chi^2/N_{\rm dat} \sim 1.16$ (NNLO)

Gluon





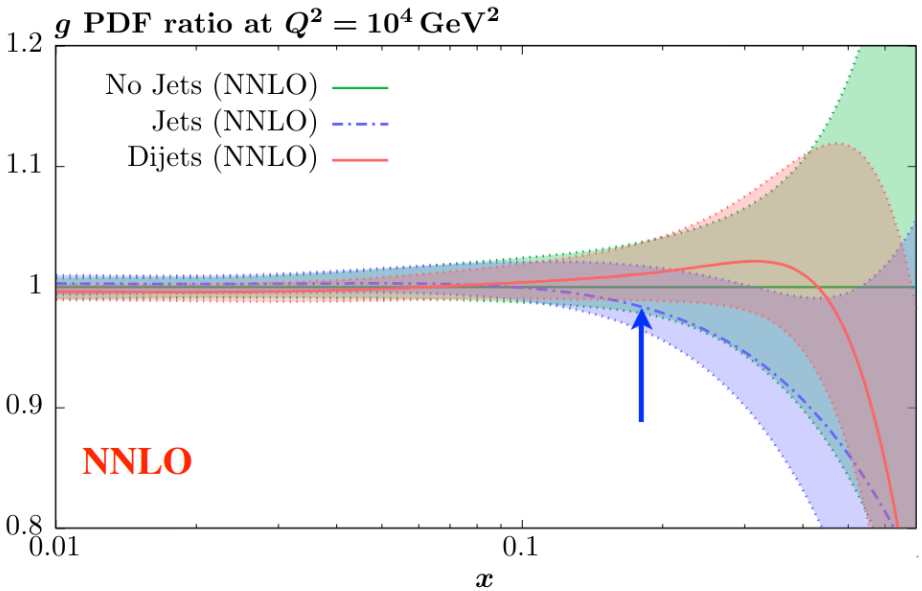
[PRD 109 (2024) 054027]

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

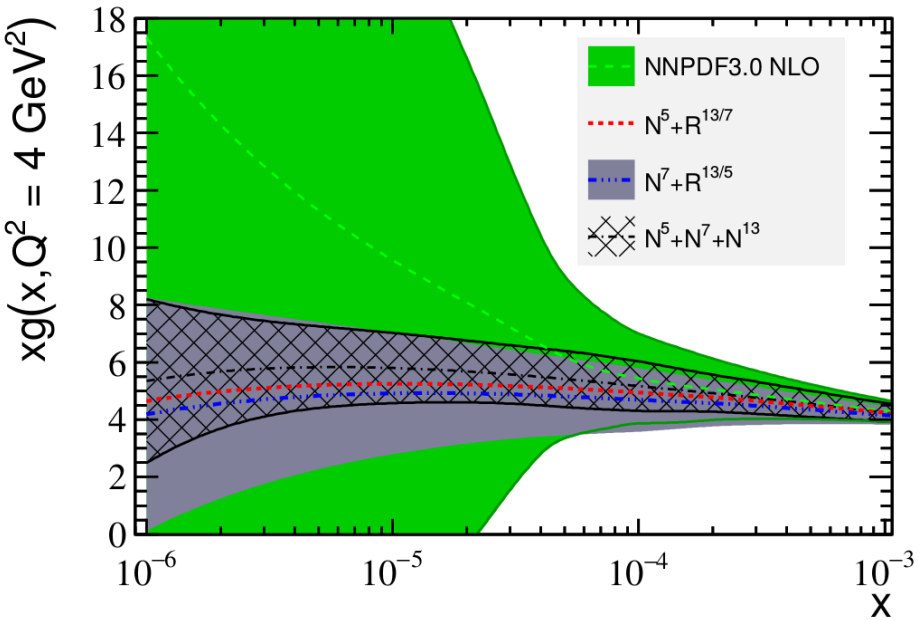
Di-jets preferred over single-inclusive jets

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



[EPJ C84 (2024) 446]

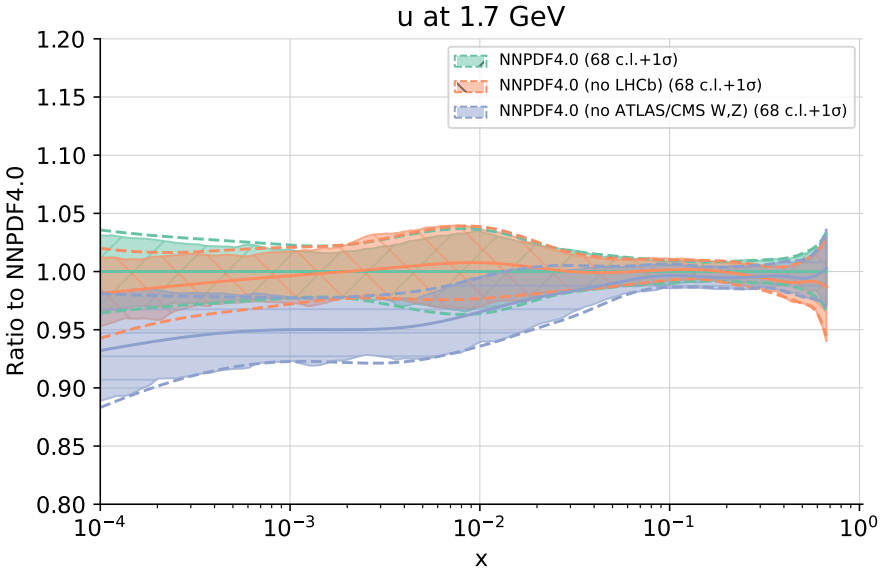
[See also EPJ C80 (2020) 797]

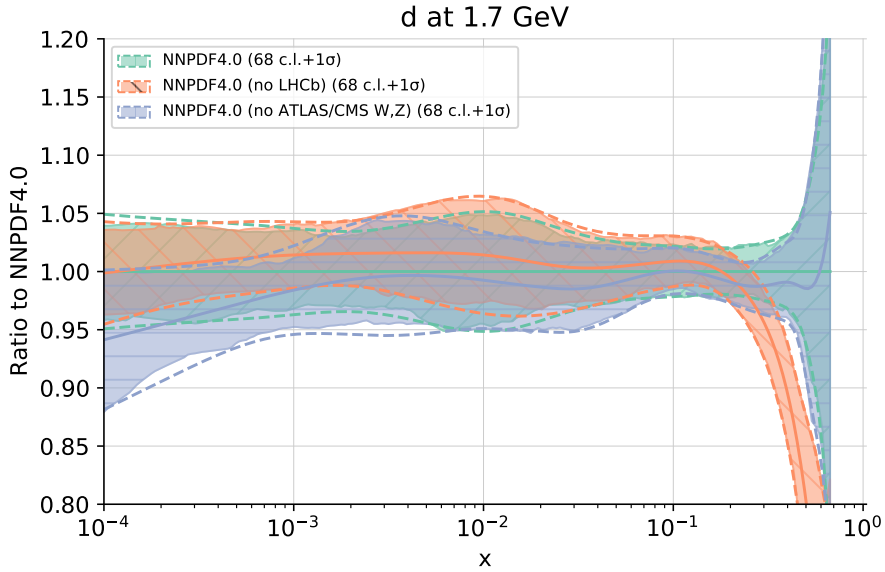


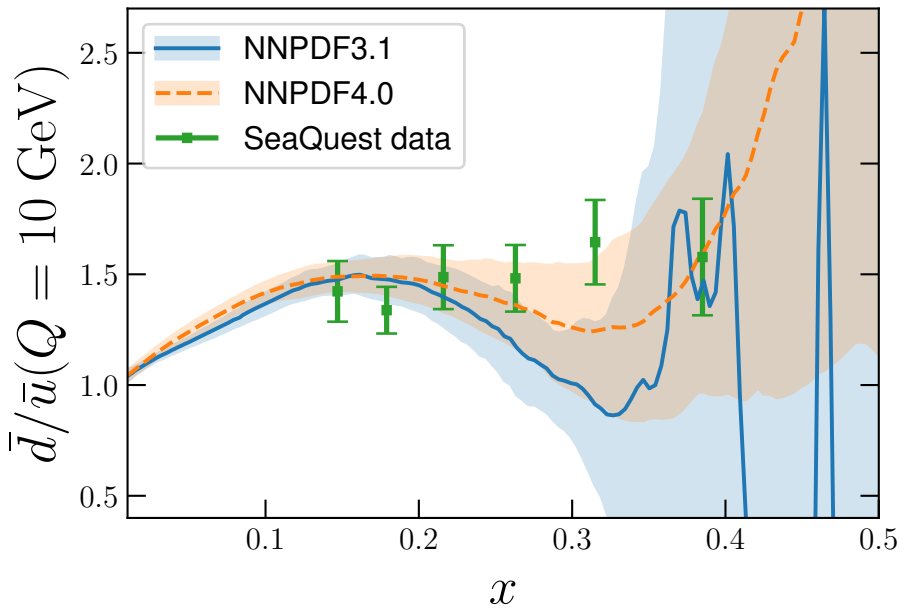
[PRL 118 (2017) 072001]

[See also 2512.23792]

Quark flavour separation







Controlled by p and d fixed-target DIS and by HERA NC and CC DIS

Relative impact of ATLAS/CMS/LHCb gauge boson production

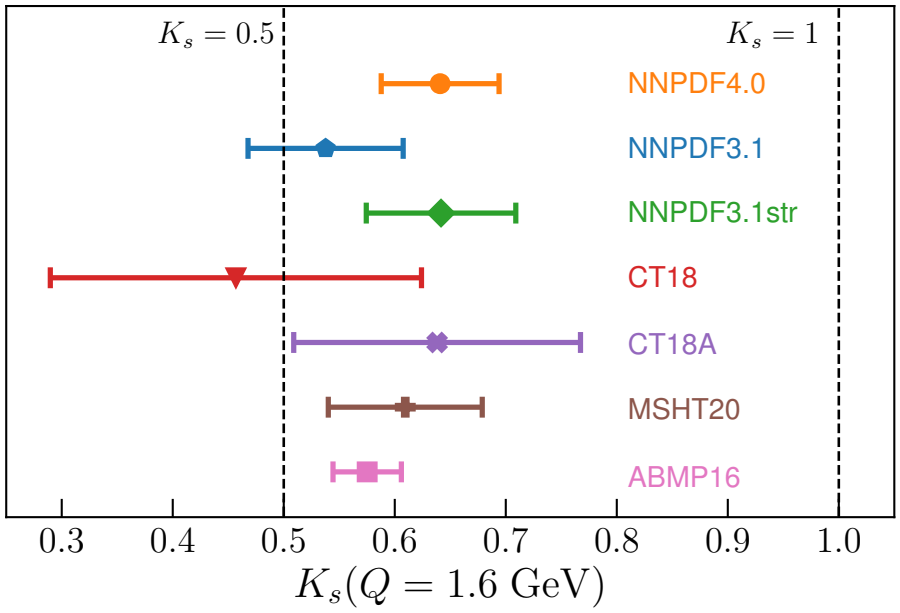
LHCb is at forward rapidity

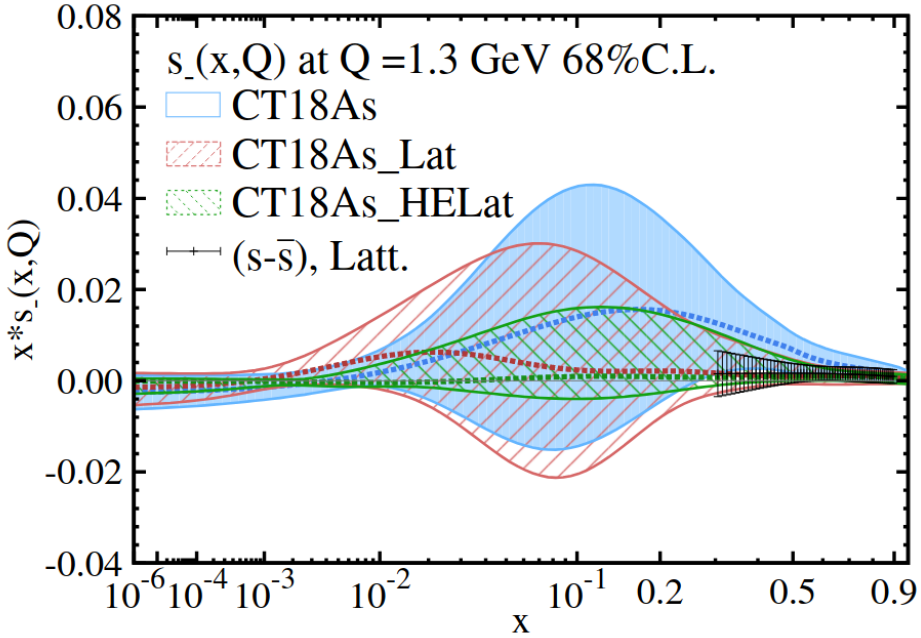
New constraint on $\bar{d}/\bar{u}$ ratio from SeaQuest

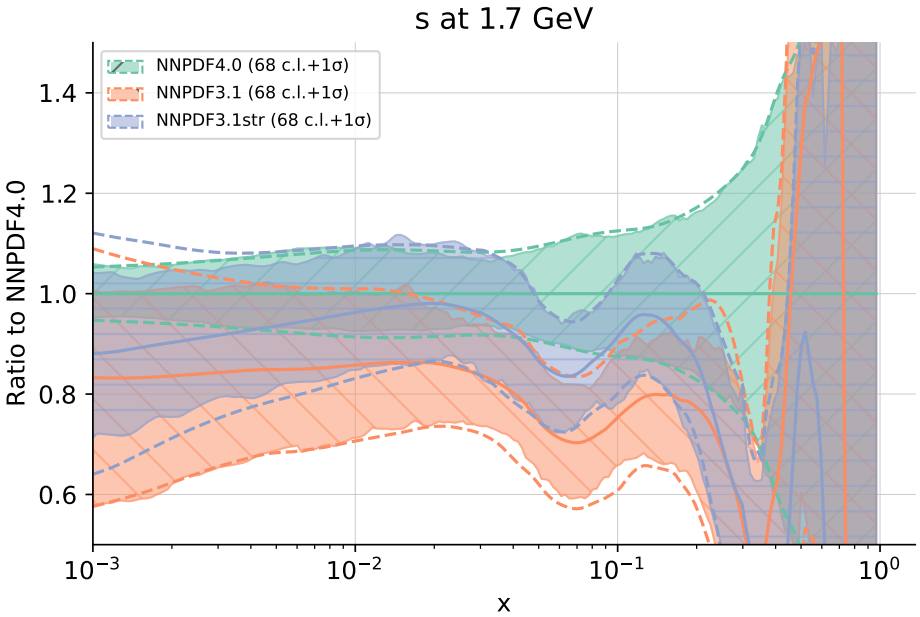
[Nature 590 (2021) 561]

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

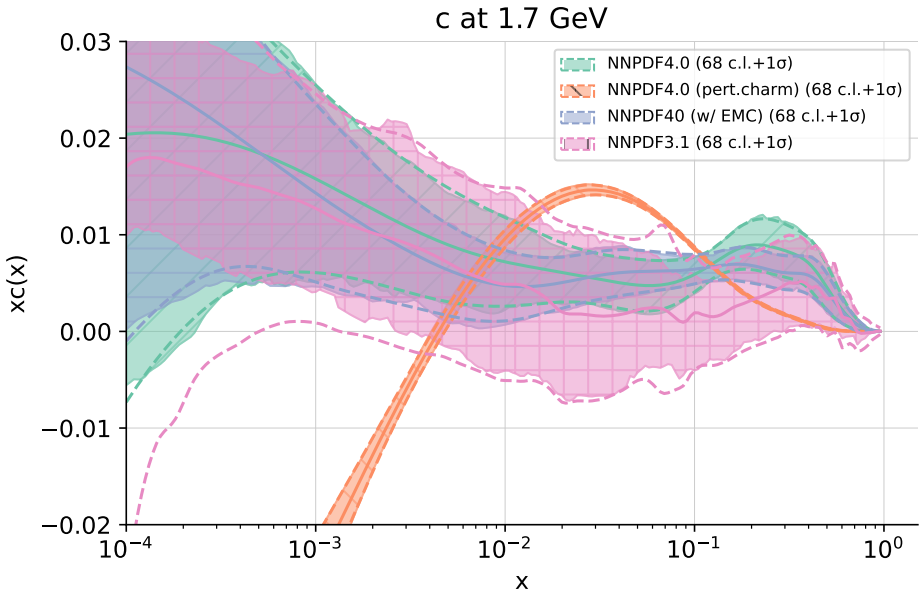
nuclear uncertainties accommodate data sets

Useful input from lattice QCD

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

[See also PRD 91 (2015) 094002]

Charm



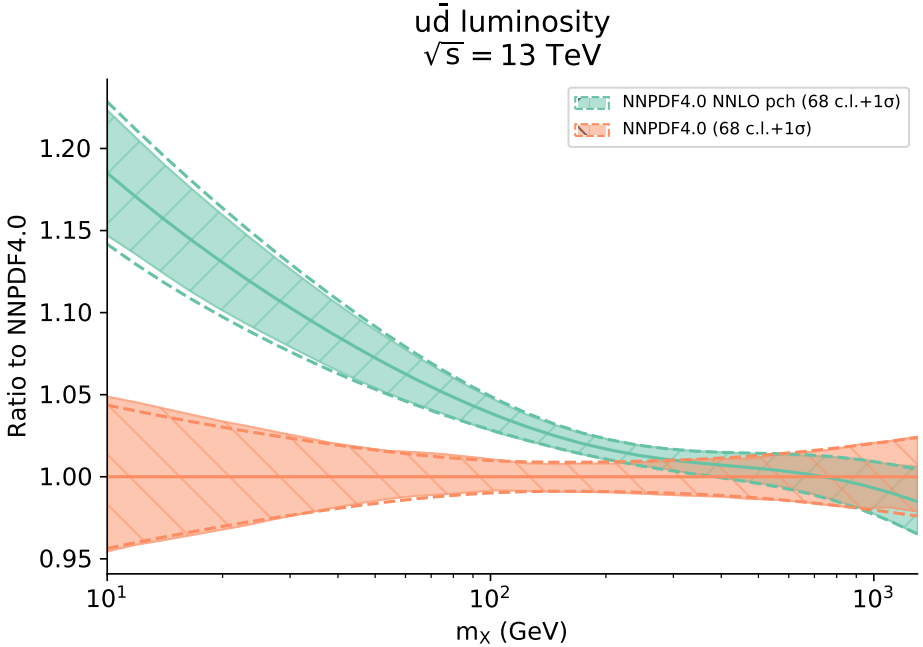
Perturbative charm alters the flavour decomposition and deteriorates the fit

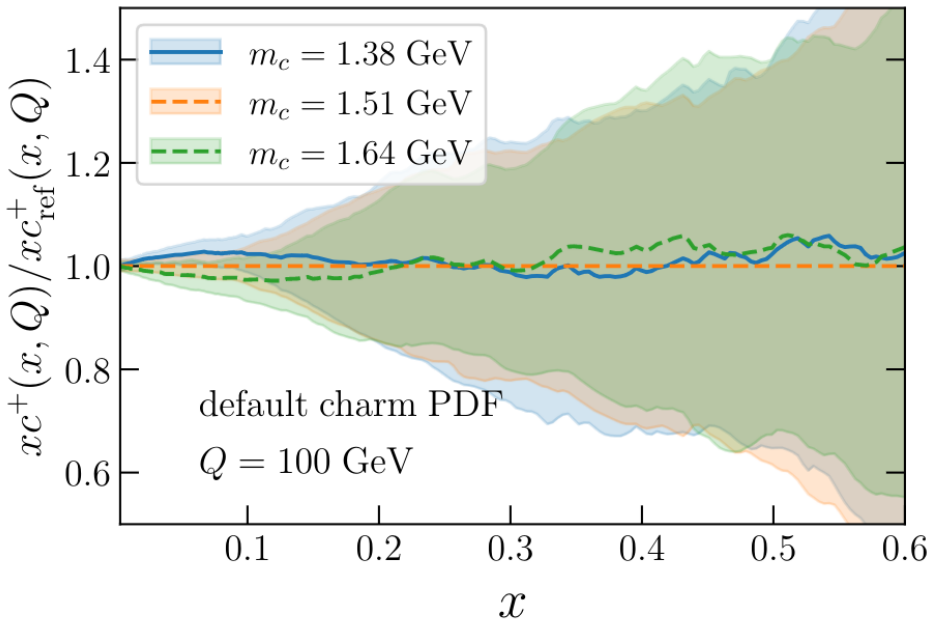
$$\chi^2_{\text{fitted charm}} = 1.17 \rightarrow \chi^2_{\text{pert. charm}} = 1.19$$

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

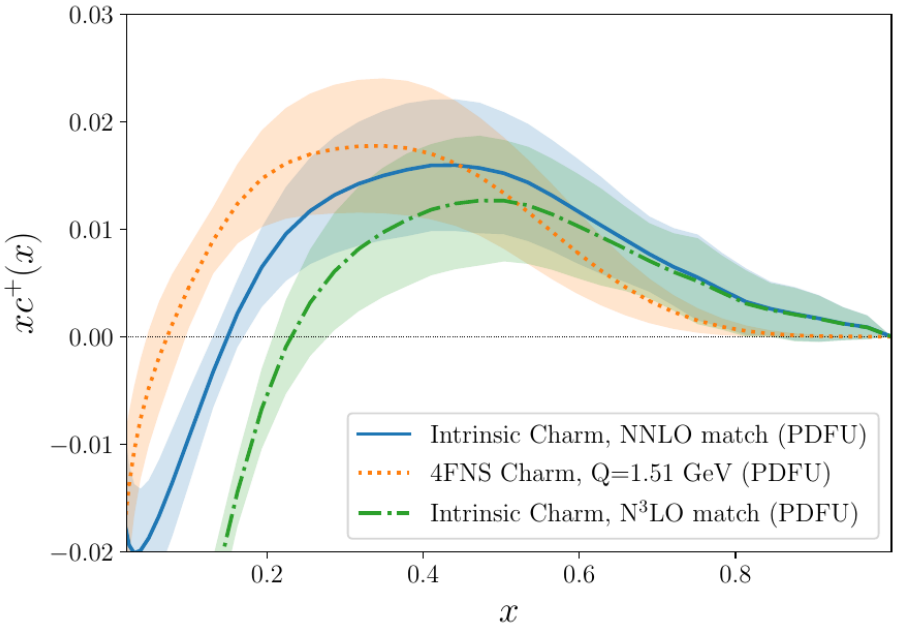
fitting charm reduces the dependence from m_c

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





Intrinsic Charm



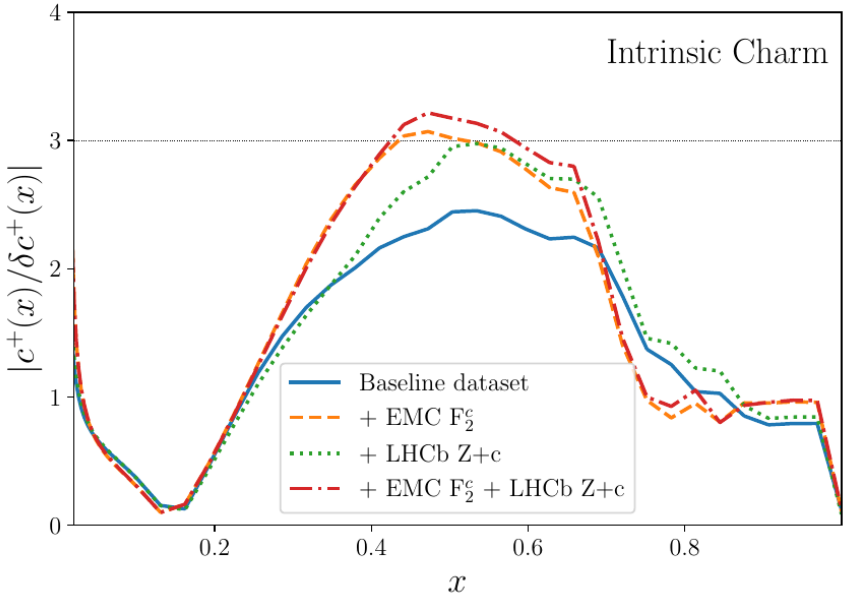
Evolve results backwards (below m_c) with N^{3}LO matching

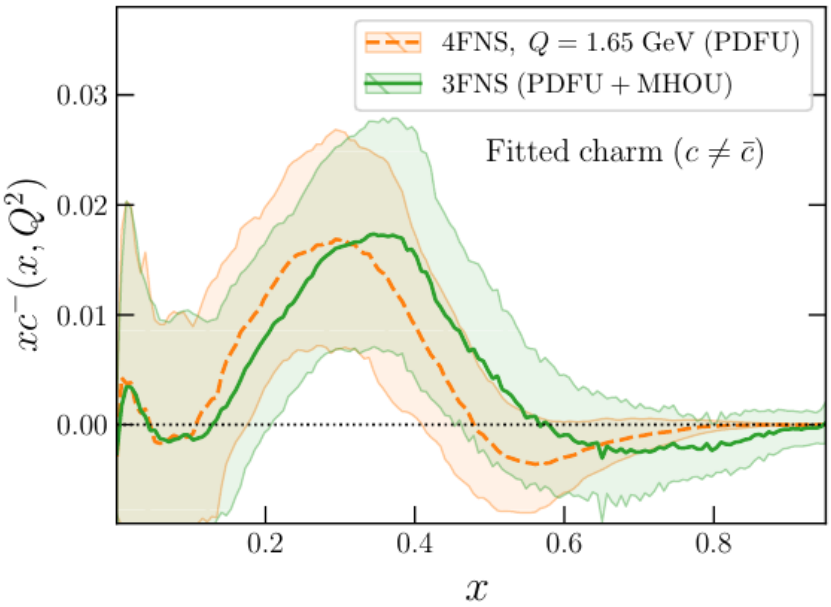
Evidence of intrinsic charm and of $c - \bar{c}$ shape compatible with models

[Nature 608 (2022) 483; PRD 109 (2024) L091501]

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

Challenged by CT18 [PLB 843 (2023) 137975]





2. Theory

Can we model theory uncertainties?

PDF determination is affected by theory uncertainties: nuclear, MHOU, PCs, α_s, ...

Idea: associate nuisance parameters to these uncertainties

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

Approach 1: determine nuisance parameters from the data

rewrite the χ^2 as a function of r_k; determine r_k so that $\partial\chi^2/\partial r_k = 0$

$$\chi^2 = \sum_i^{N_{\rm dat}} \left[\frac{1}{\alpha_i^2} \left(D_i - T_i[\{\vec{a}\}] - \sum_k r_k \beta_{ki} \right)^2 + \sum_k r_k^2 \right]$$

Approach 2: implement Bayesian parameter estimation

define a prior theoretical covariance matrix on top of the experimental covariance matrix

$$\mathrm{cov}_{ij} = \mathrm{cov}_{ij}^{\rm exp} + \mathrm{cov}_{ij}^{\rm th} \qquad \text{with} \quad \mathrm{cov}_{ij}^{\rm th} = S_{ij}$$

The two approaches are formally equivalent

Missing higher order uncertainties



$$S_{ij} = \frac{1}{N}\sum_k \Delta_i^{(k)} \Delta_j^{(k)}$$

$$\Delta_i^{(k)} = T_i^{7\text{pt}}(\mu_R^{(k)}, \mu_F^{(k)}) - T_i(\mu_R^0, \mu_F^0)$$

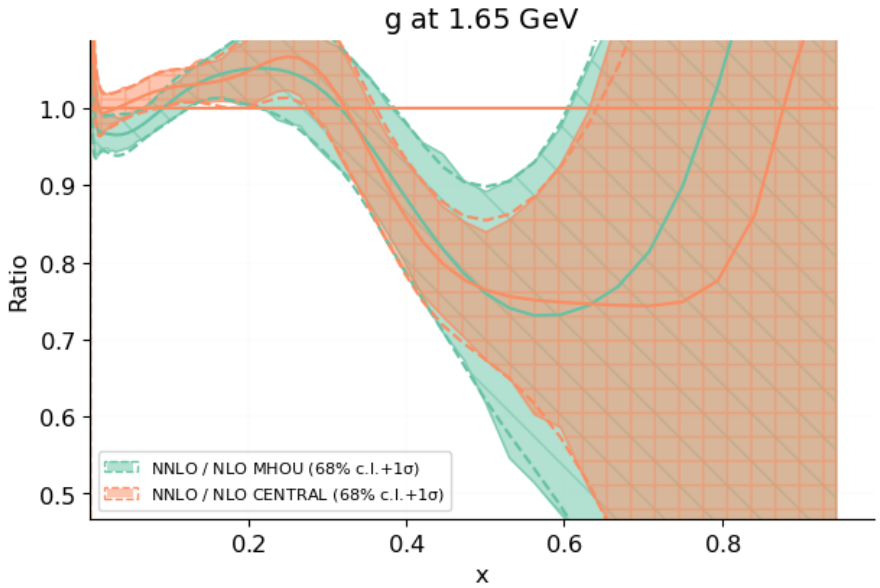
$\chi^2_{\text{tot}} = 1.23 \to \chi^2_{\text{tot}} = 1.34$ (no MHOU, NLO)

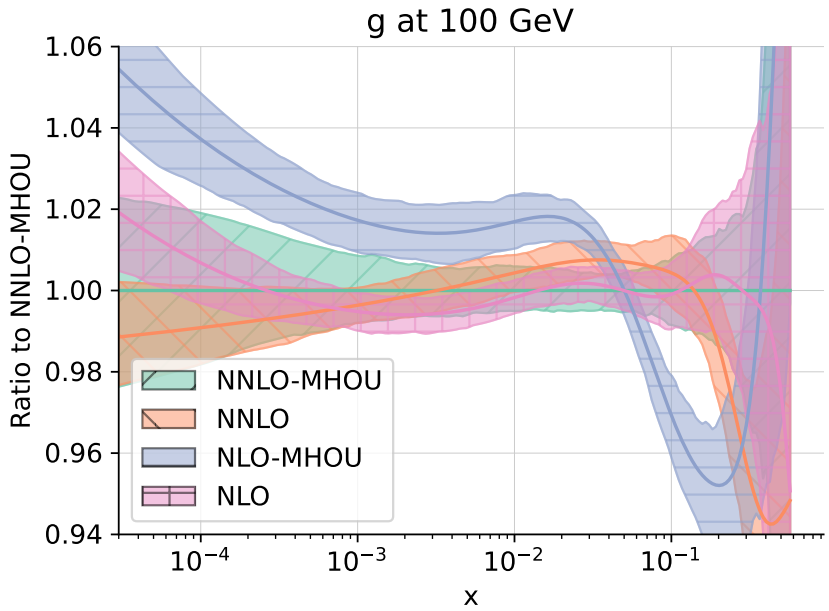
$\chi^2_{\text{tot}} = 1.13 \to \chi^2_{\text{tot}} = 1.17$ (no MHOU, NNLO)

Tensions relieved

Faster perturbative convegence

Small increase of PDF uncertainties





[EPJ C79 (2019) 282; EPJ C81 (2021) 37; EPJ C84 (2024) 517]

Nuclear, higher twist and power correction uncertainties

$$S_{ij} = \frac{1}{N}\sum_k \Delta_i^{(k)} \Delta_j^{(k)}$$

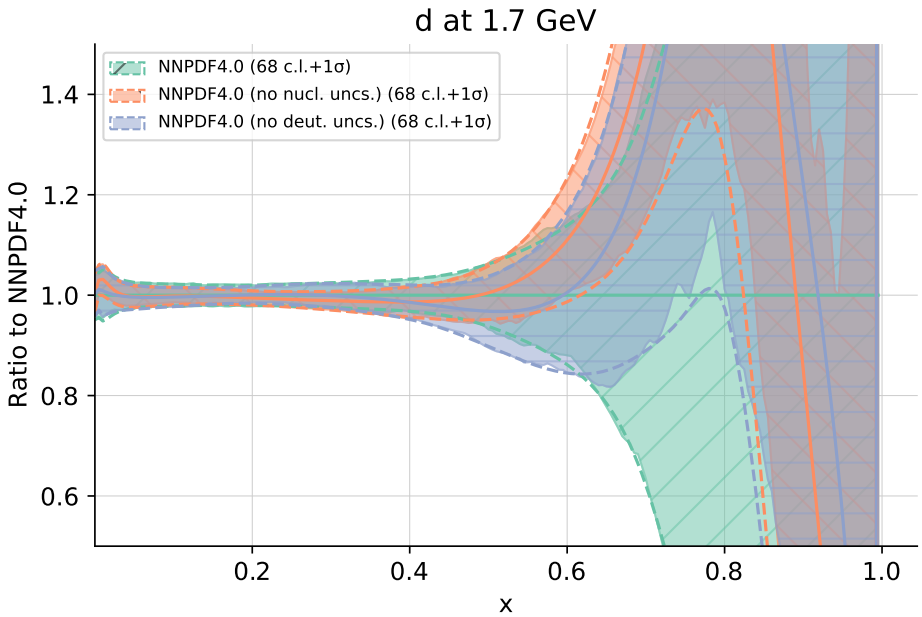
$$\Delta_i^{(k)} = T_i^N[f_N^{(k)}] - T_i^N[f_p]$$

$\chi^2_{\rm tot} = 1.17 \to \chi^2_{\rm tot} = 1.26$ (no nucl. uncs.)

$\chi^2_{\rm LHCb} = 1.54 \to \chi^2_{\rm tot} = 1.76$ (no nucl. uncs.)

The bulk of the effect is due to nuclear uncertainties for heavy nuclei at large x

Reduced tension between DIS and LHC data



[EPJ C79 (2019) 282; EPJ C81 (2021) 37]

$$S_{ij} = \frac{1}{N}\sum_k \Delta_i^{(k)} \Delta_j^{(k)}$$

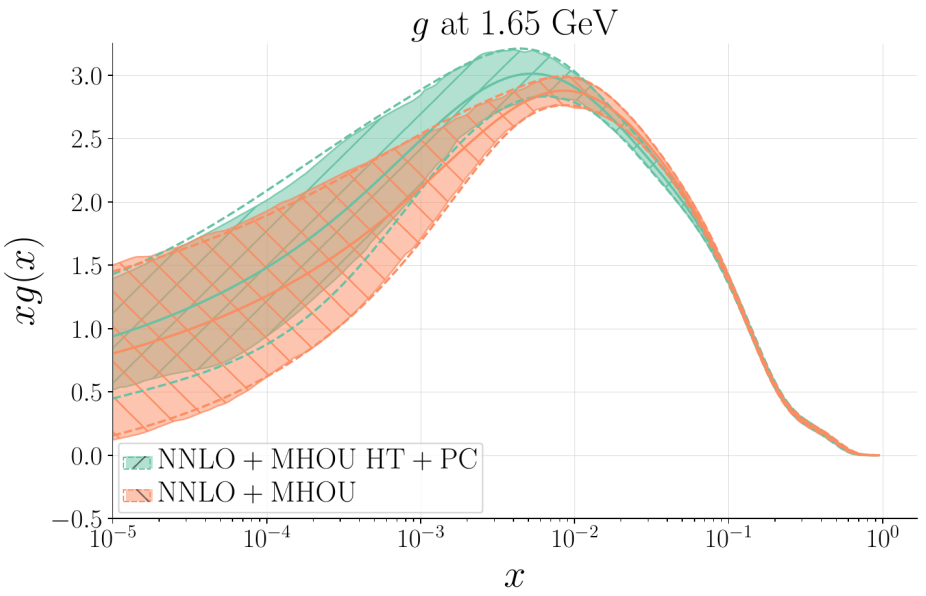
$$\Delta_i^{(k)} = T_i^{\rm HT,PC(k)} - T_i^{\rm LT}$$

$$T_i^{\rm HT(k)} = T_i^{\rm LT}\left(1 + \frac{H^{(k)}(x)}{Q^2}\right)$$

$$T_i^{\rm PC(k)} = T_i^{\rm LT}\left(1 + \frac{H^{(k)}(\eta)}{p_T}\right)$$

$\chi^2_{\rm tot} = 1.11 \to \chi^2_{\rm tot} = 1.13$ (no HT/PC NNLO)

Tensions relieved, shift of PDF central value



[EPJ C86 (2026) 281; ibid. 115]

The photon PDF and QED corrections

Photon PDF à la LuxQED

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

Fit quality unaltered: $\chi^2/N_{\rm dat} = 1.13$

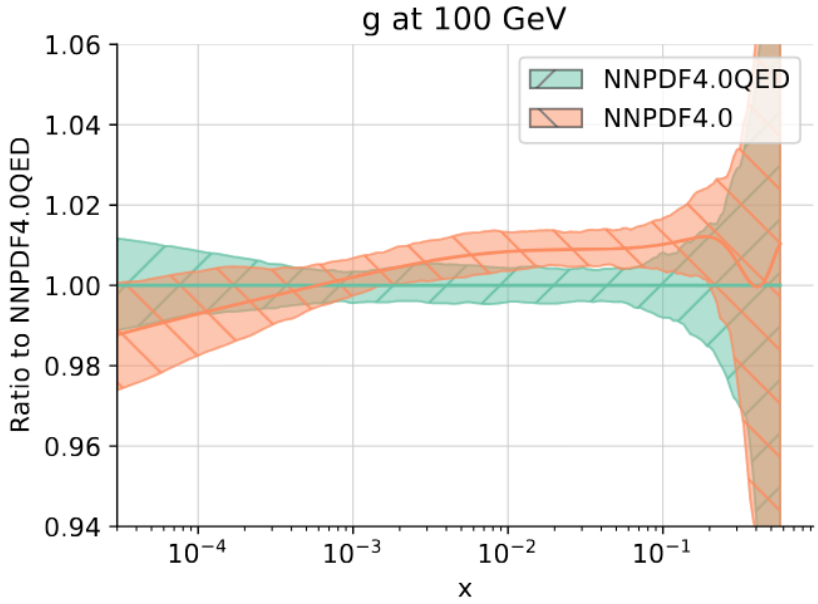
Small (0.5%) momentum shift from g to γ

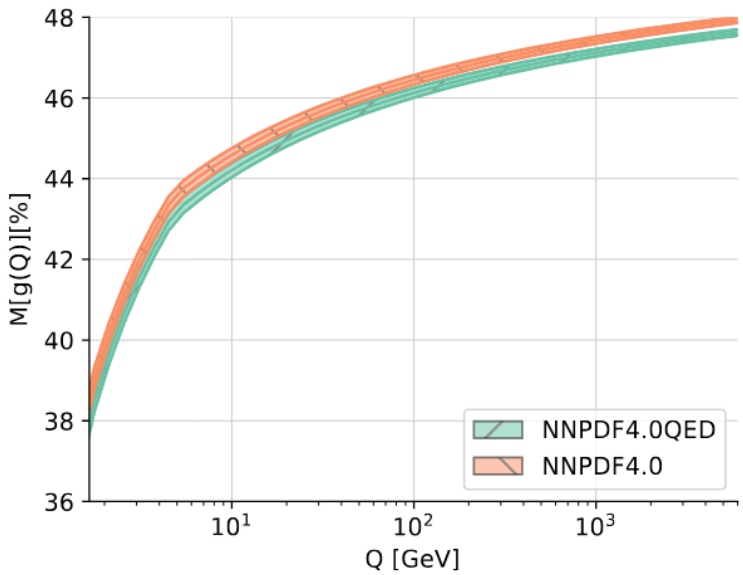
Small (1%) suppression of the gluon PDF

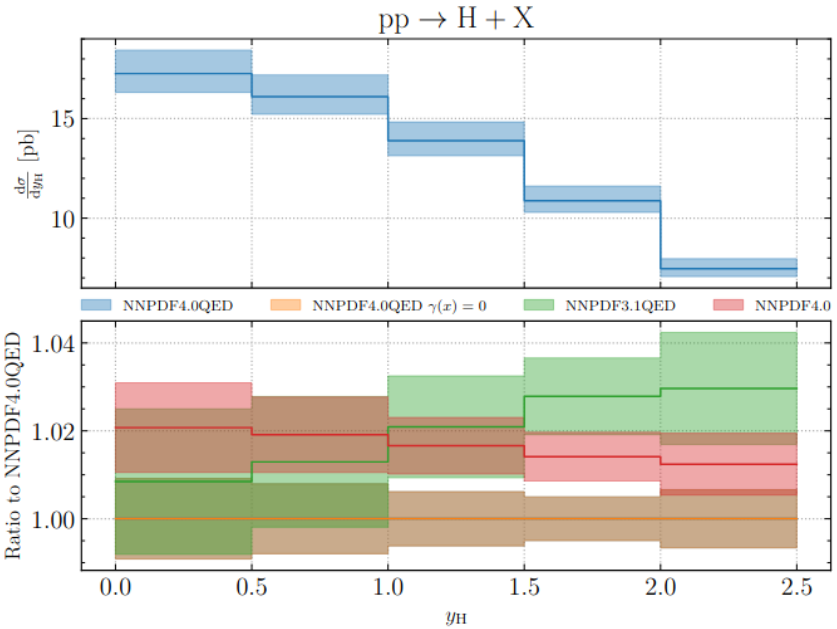
1-2% suppression in ggH cross section

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

Similar results by CT and MSHT







The gluon PDF and aN^{3}LO corrections

Incorporate all available N^{3}LO computations

Model incomplete and missing higher orders with a covariance matrix

Fit quality unaltered: $\chi^2/N_{\rm dat} = 1.13$

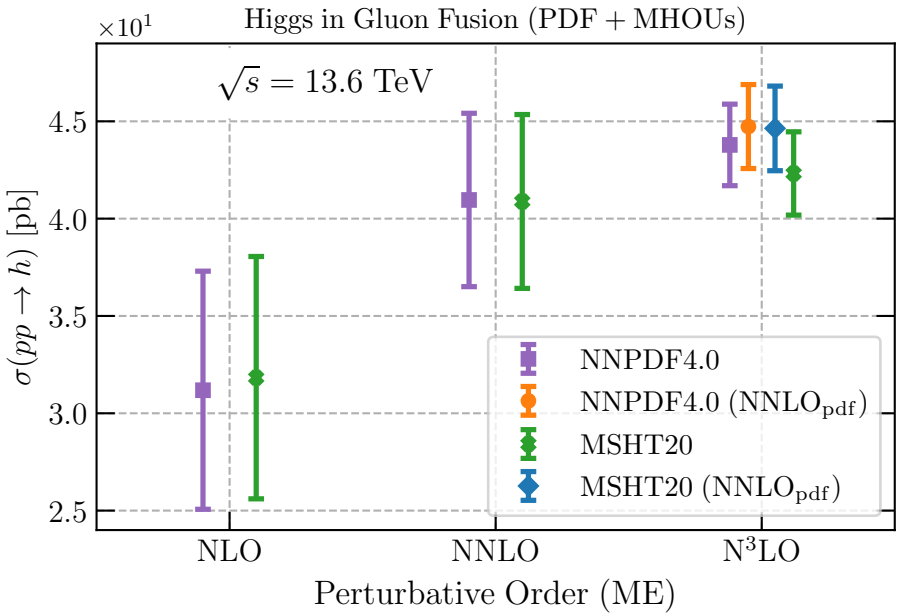
Small (2%) suppression of the gluon PDF

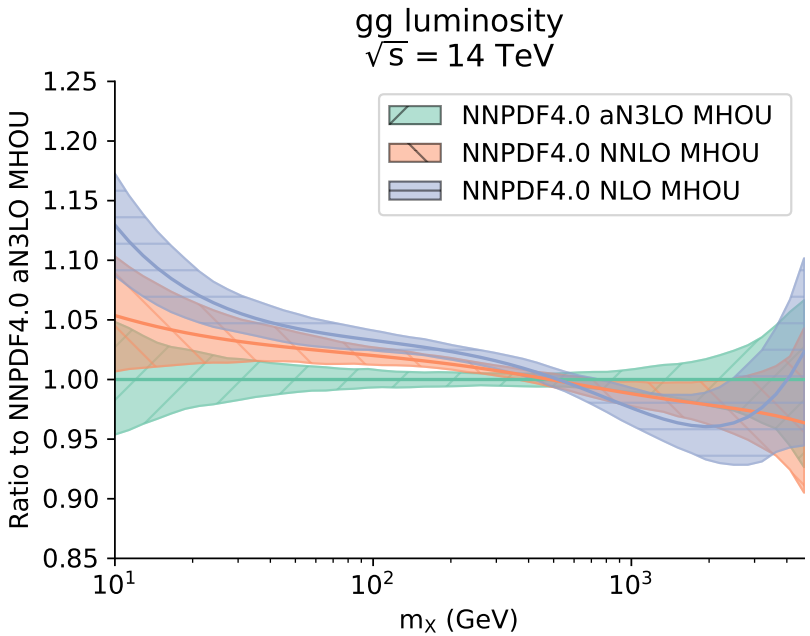
2-3% suppression in ggH cross section

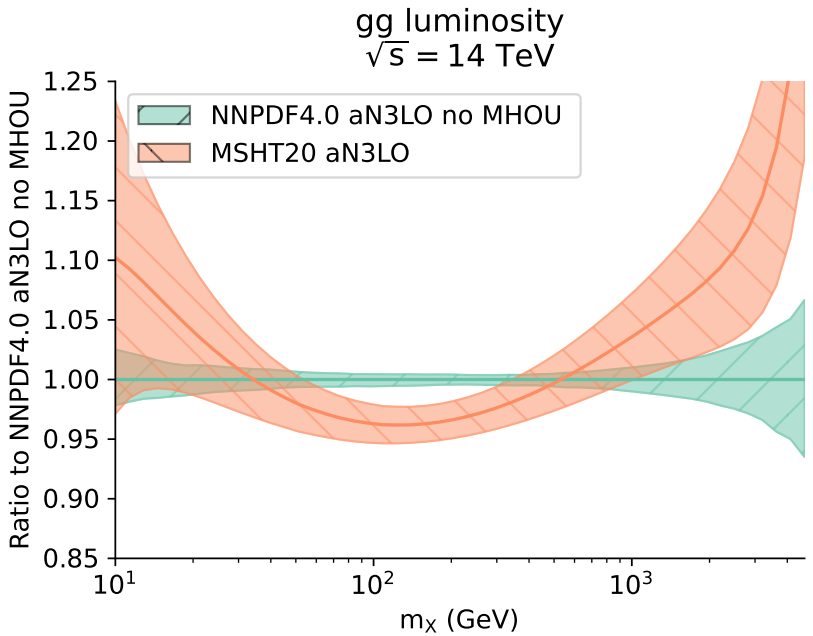
5% discrepancy w.r.t. MSHT20

[See, *e.g.* EPJ C84 (2024) 859]

Ongoing PDF4LHC benchmark with CT







3. Methodology

Experimental correlations matter (a lot)

Single inclusive jet data from ATLAS 7 TeV

default correlations: terrible χ^2

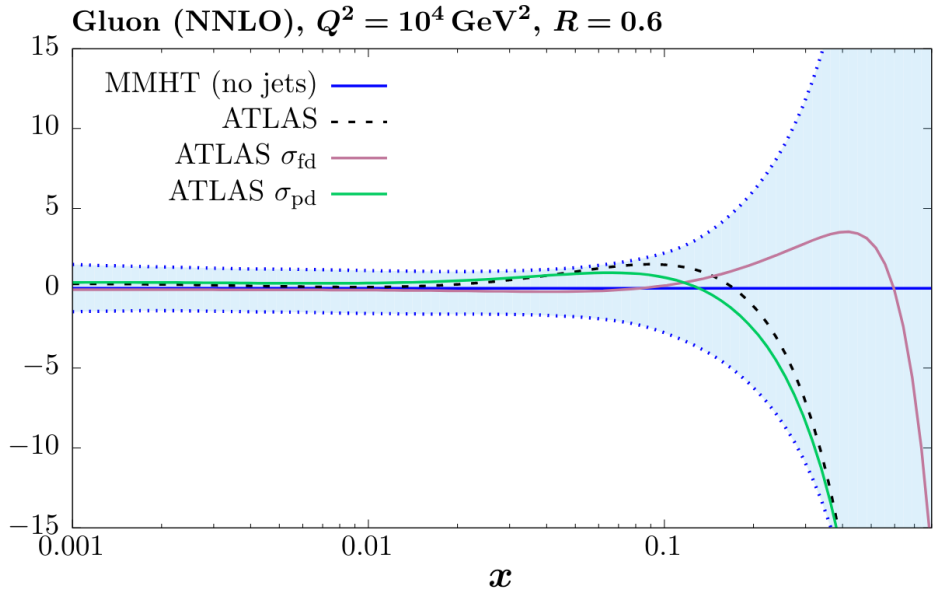
(correlations across rapidity bins)

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

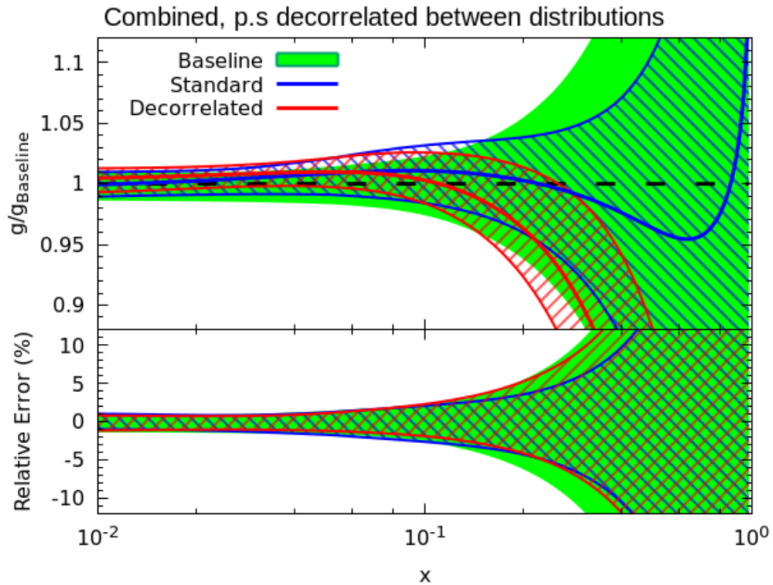
(correlations across different spectra)

decorrelation models: improve the fit a lot

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

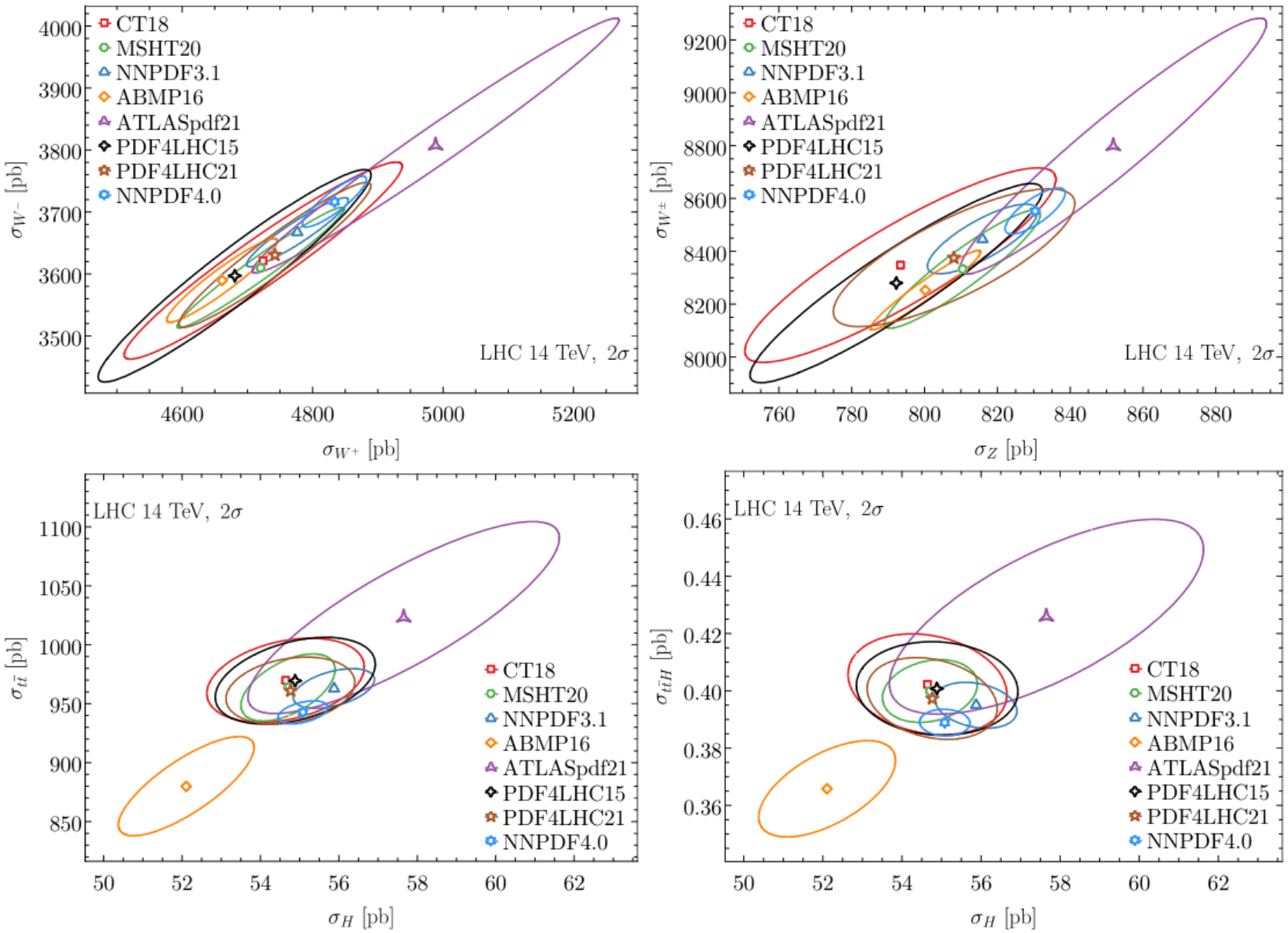
appreciable effect on the extracted gluon

different gluon depending on the top spectrum



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

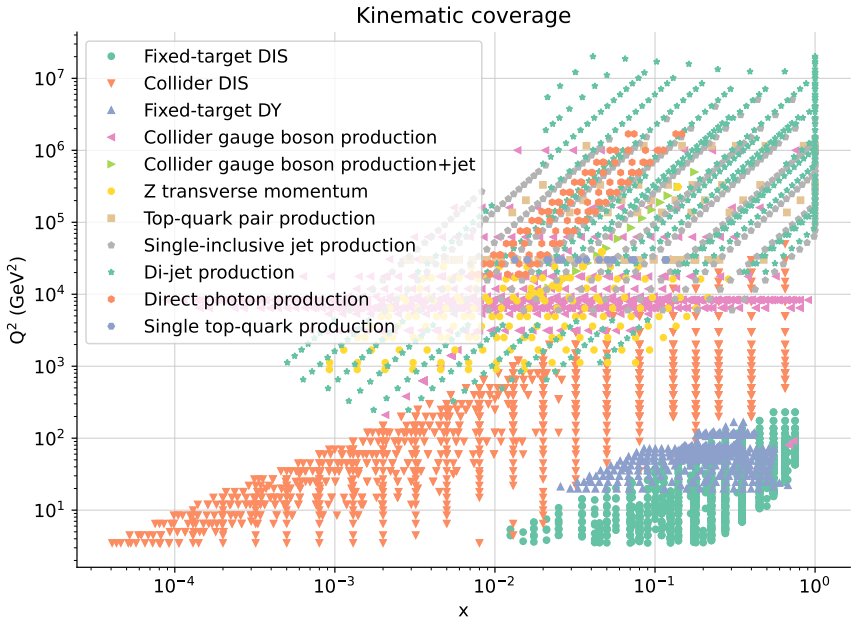
Making predictions with PDFs



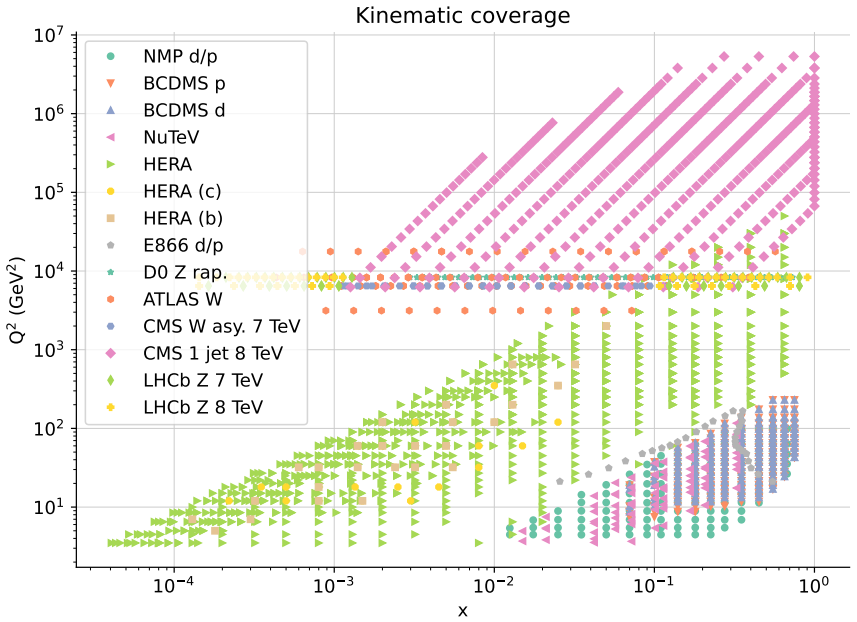
[Acta Phys.Polon.B 53 (2022) 12]

The PDF4LHC21 benchmark

Global data set ($N_{\rm dat} \sim 5000$)



Reduced data set ($N_{\rm dat} \sim 2000$)



Same reduced data set

Same heavy quark mass values: $m_c = 1.4$ GeV; $m_b = 4.75$ GeV; $\alpha_s(M_Z) = 0.118$

Same strong coupling value: $\alpha_s(M_Z) = 0.118$

No strangeness asymmetry at input scale: $(s - \bar{s})(Q_0) = 0$

Charm perturbatively generated

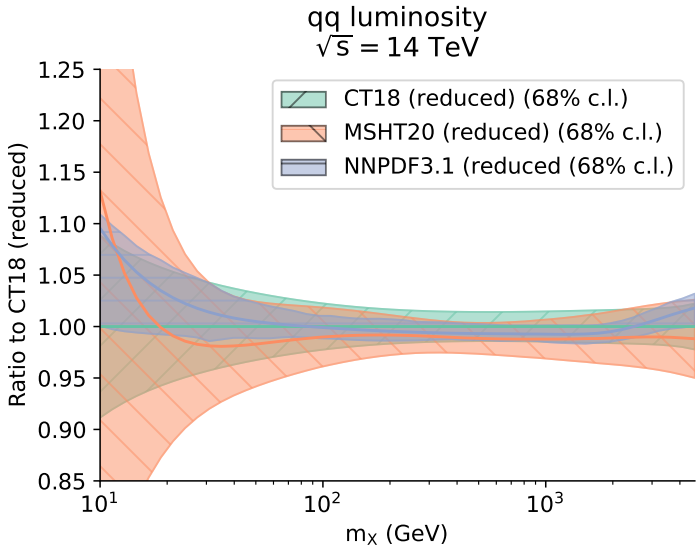
Positive-definite quark distributions

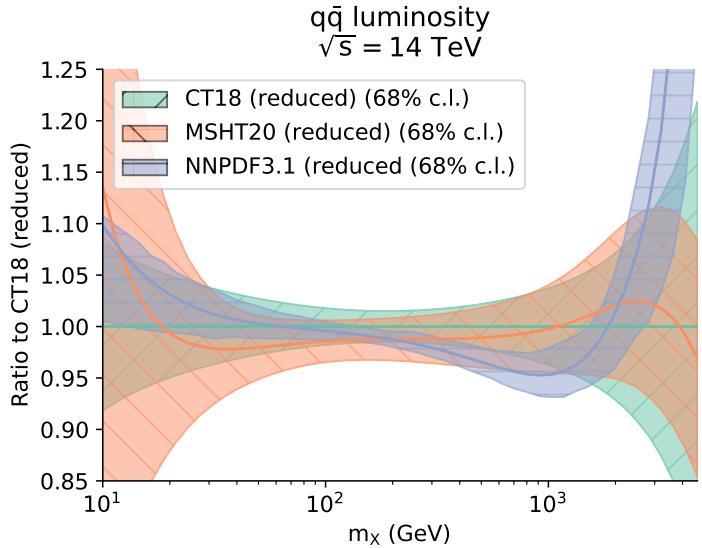
No deuteron or nuclear corrections to analyse νN cross sections

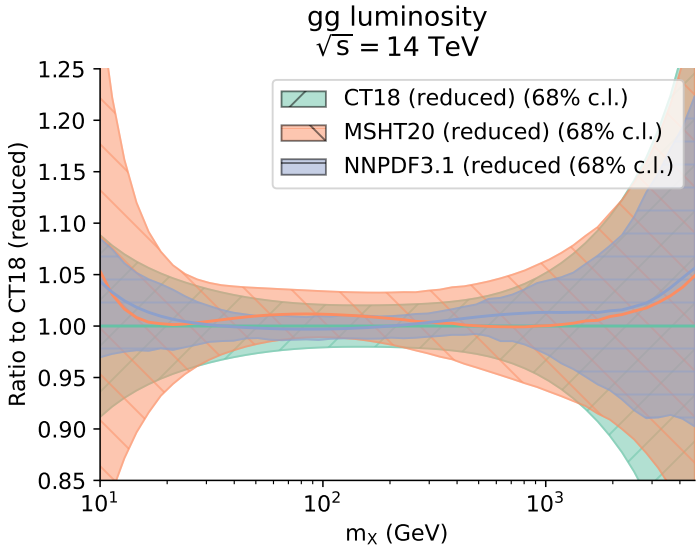
Fixed branching ratio for charm hadrons to muons; NNLO corrections for dimuon data

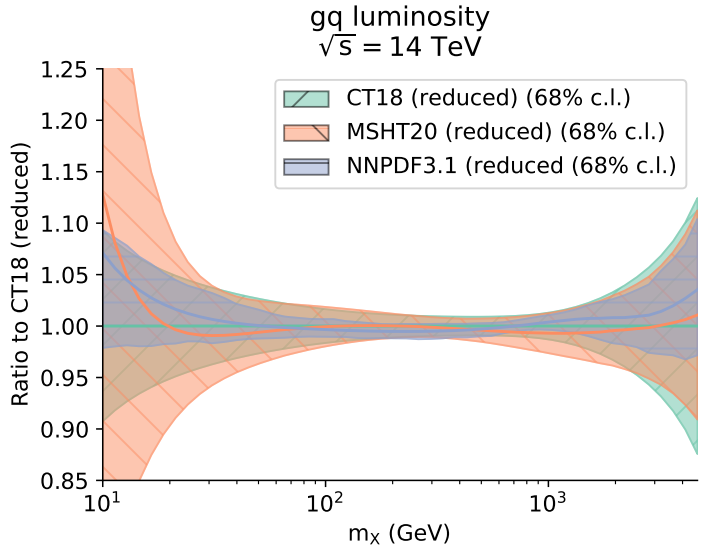
[JPG 49 (2022) 080501]

The PDF4LHC21 benchmark









Very good agreement within uncertainties; similar size of uncertainties in the data region
Remaining differences reflect methodological choices

[JPG 49 (2022) 080501]

The methodology matters

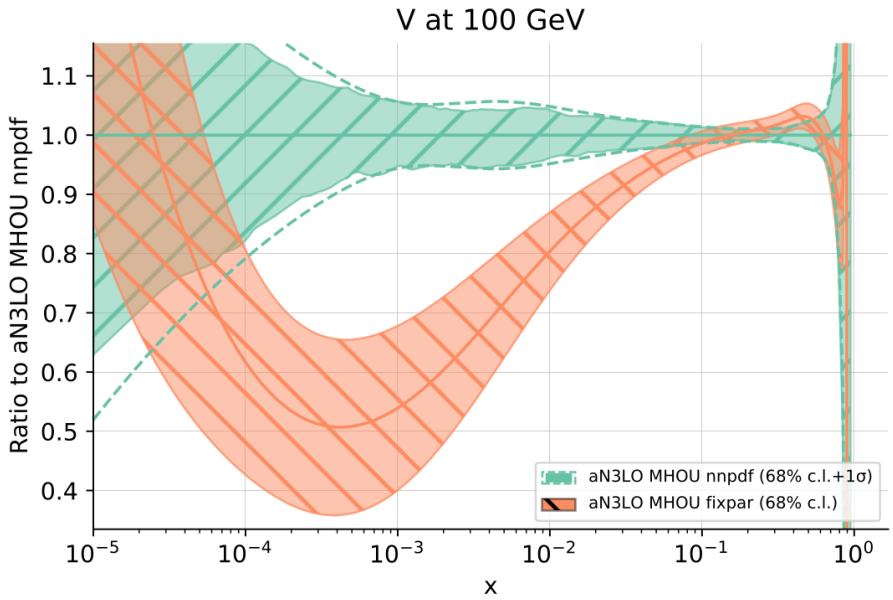
Which differences are driven by theory and which are driven by methodology?

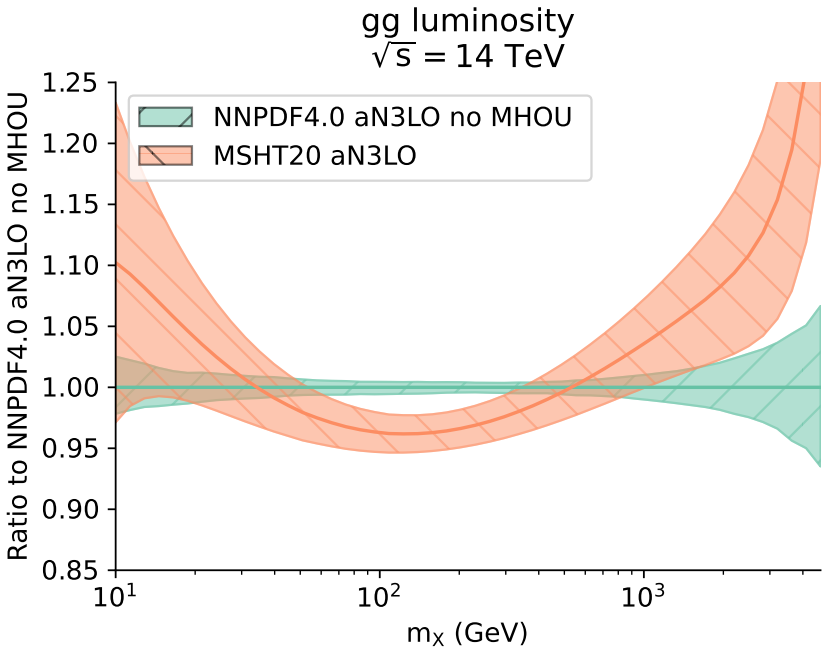
Freeze theory and data (NNPDF) and change methodology (NNPDF vs MSHT)

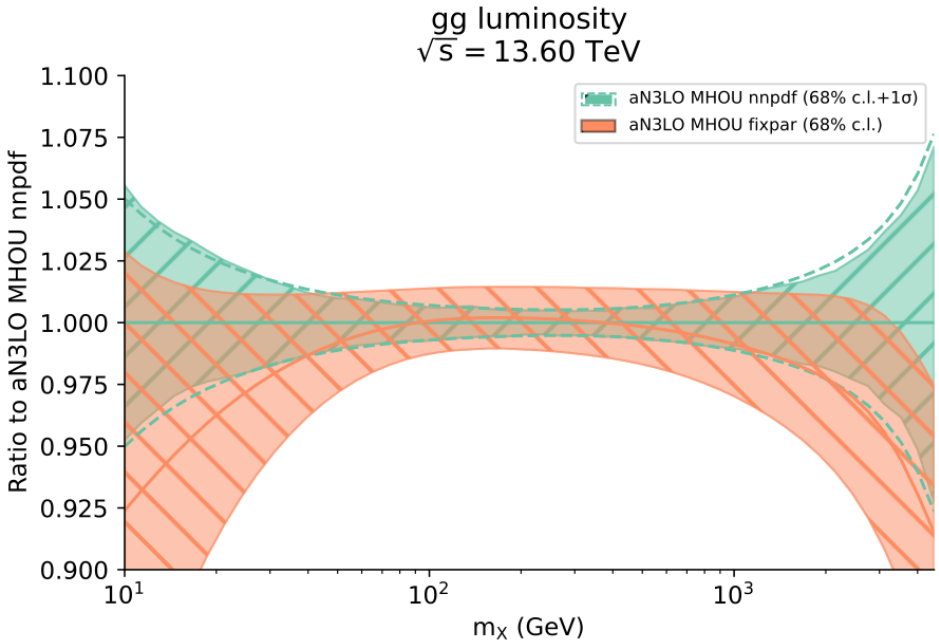
Some are due to theory (gluon)

Some are due to methodology (valence)

[JHEP 07 (2026) 185]







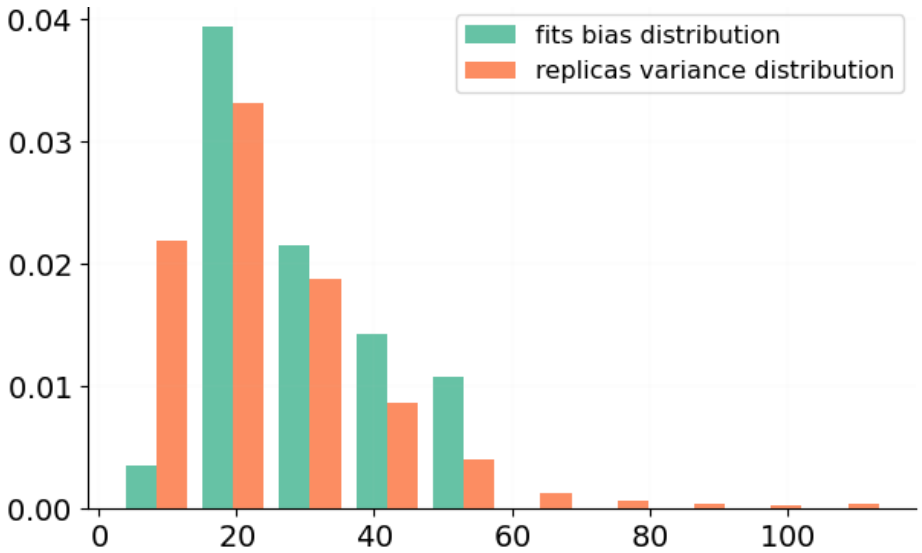
Validate PDF uncertainties: closure tests

Fit PDFs to pseudodata generated assuming a known underlying law

NNPDF

bias difference of central prediction and truth
variance uncertainty of replica predictions

If model complexity optimal,
generalisation error minimum
$(\text{E[bias]})^2 = \text{variance}$
25 fits, 40 replicas each

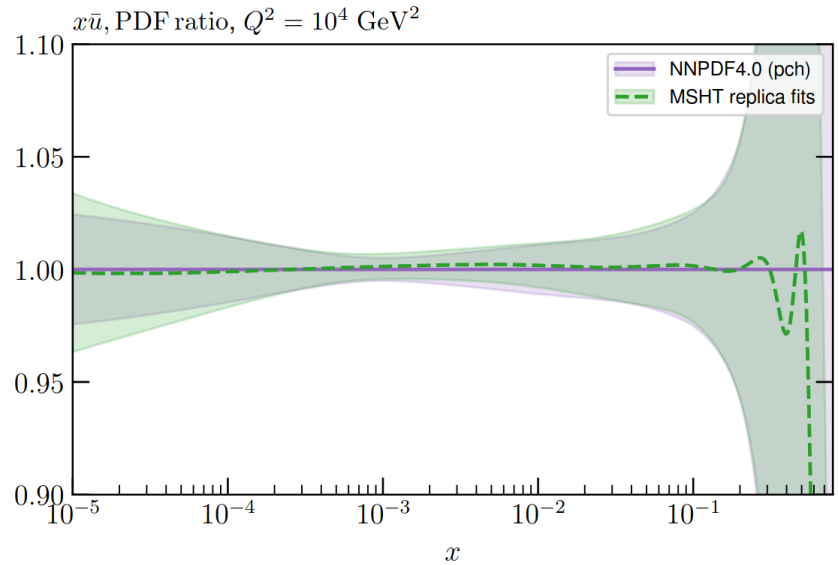


[EPJ C77 (2017) 663; EPJ C82 (2022) 330]

MSHT

parameterisation flexible enough to give
faithful description of global pseudodata

NNPDF *vs* MSHT comparison
similar fit quality
similar errors, with MSHT unit tolerance



[EPJ C85 (2025) 316]

4. Summary

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.

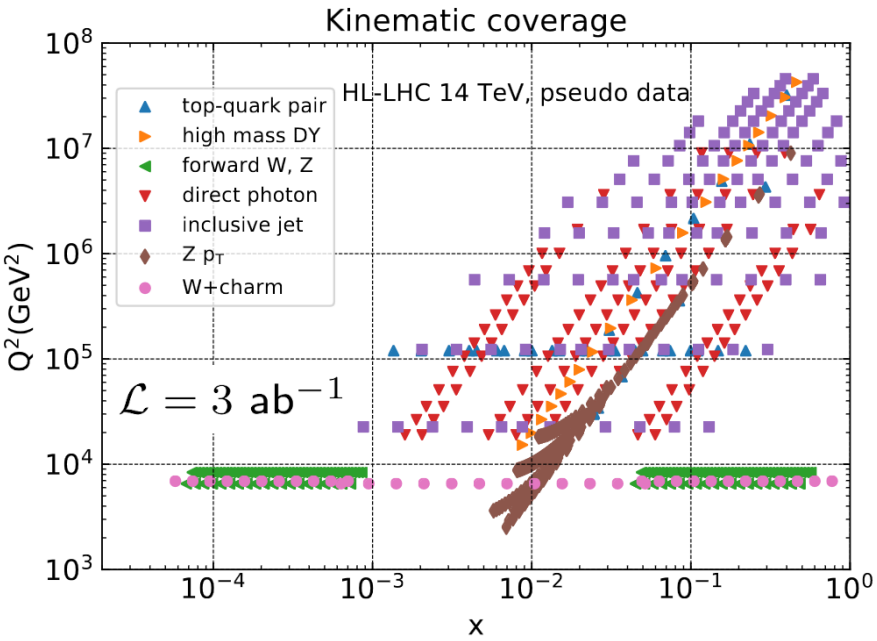
Represent theory uncertainties into PDF uncertainties.

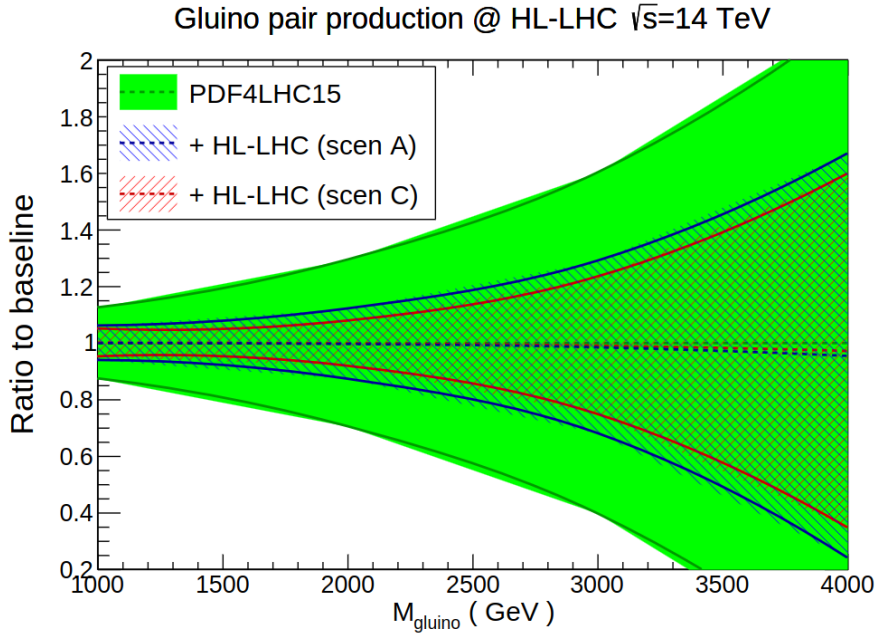
Deploy a robust fitting methodology and good statistical tests of it.

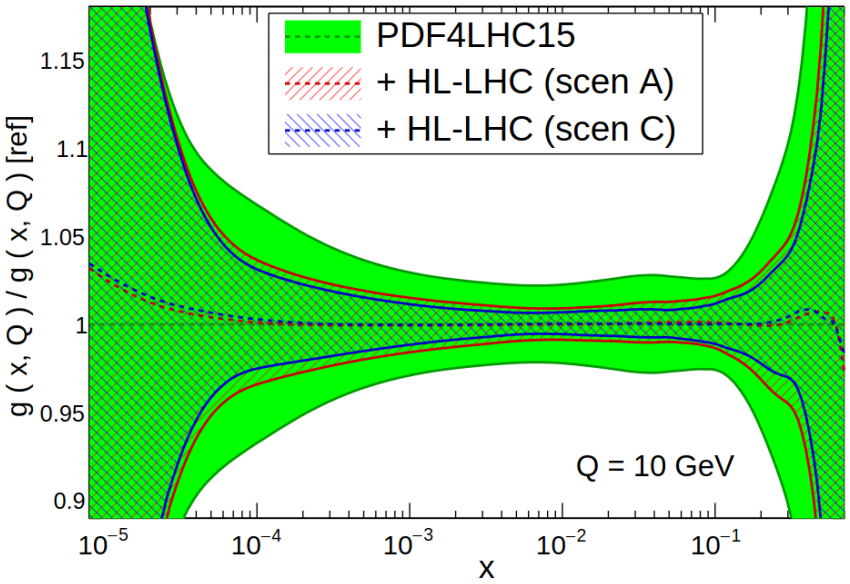
Benchmark efforts may benefit from public releases of PDF codes and inputs.

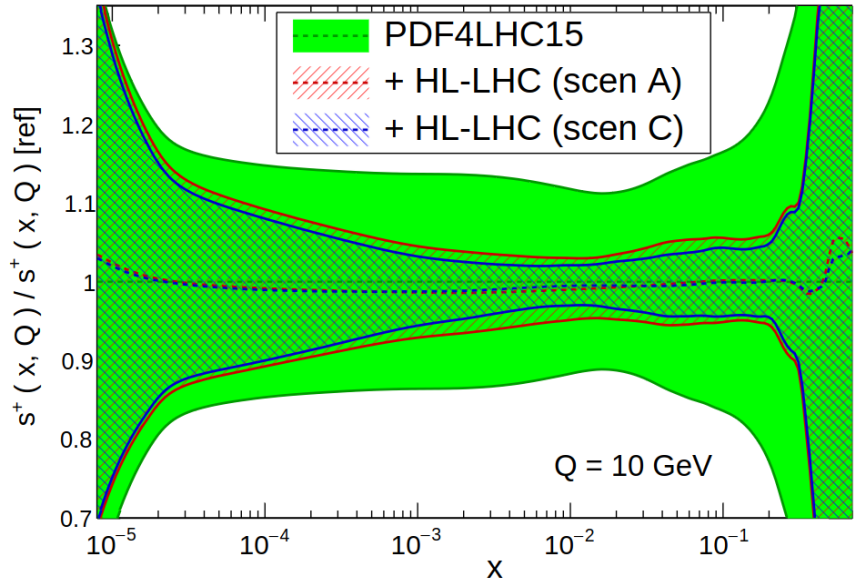
Thank you

Impact of future data: HL-LHC



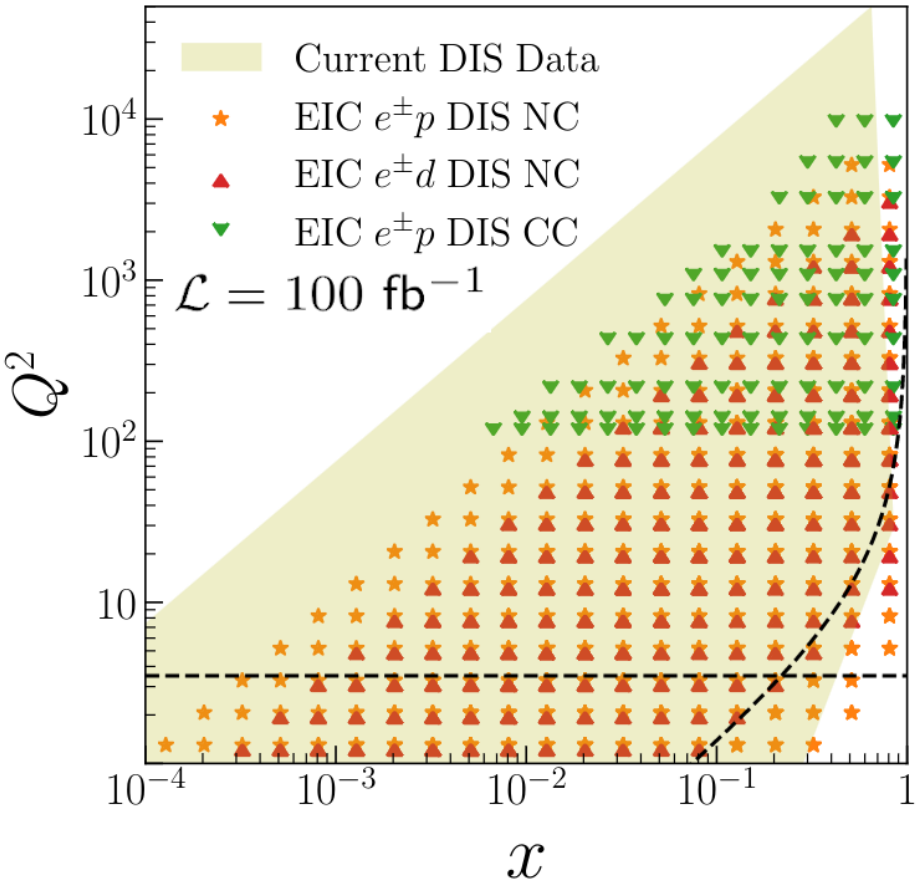






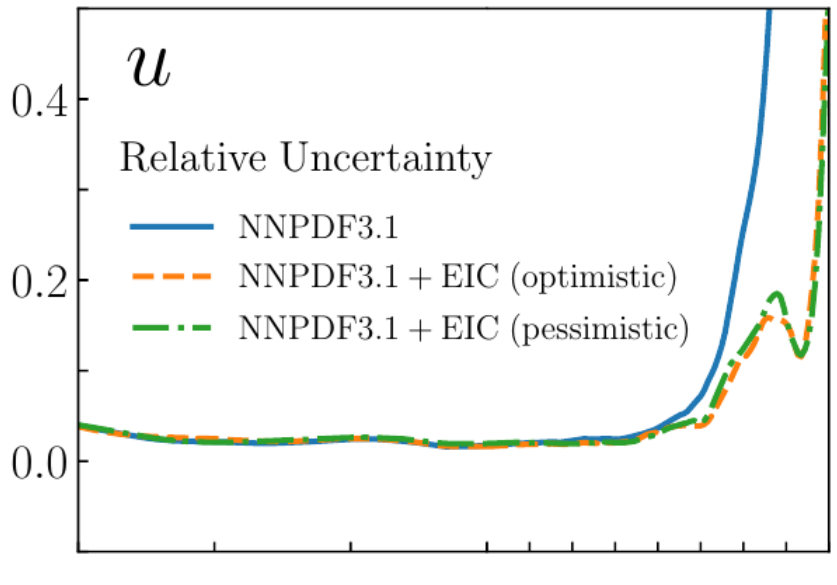
[EPJ.C 78 (2018) 962]

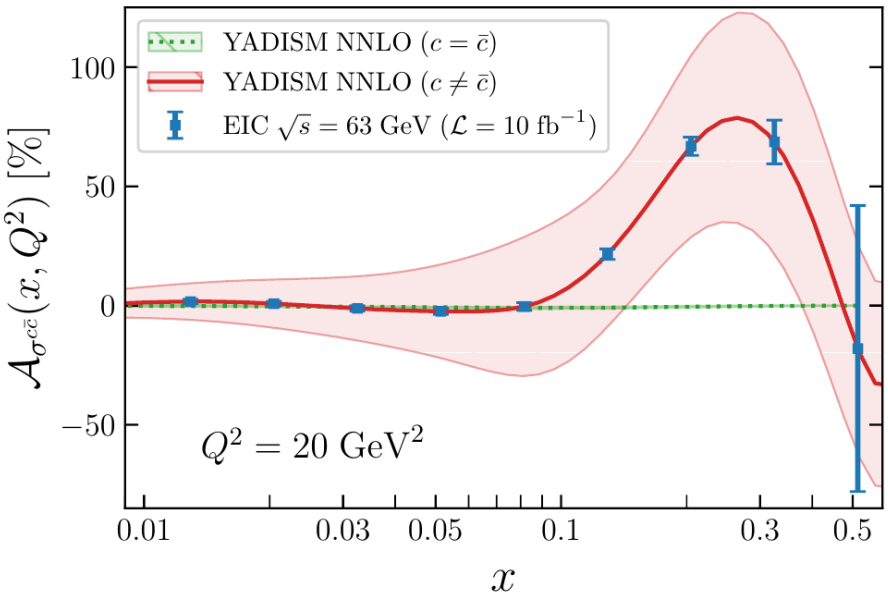
Impact of future data: EIC



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

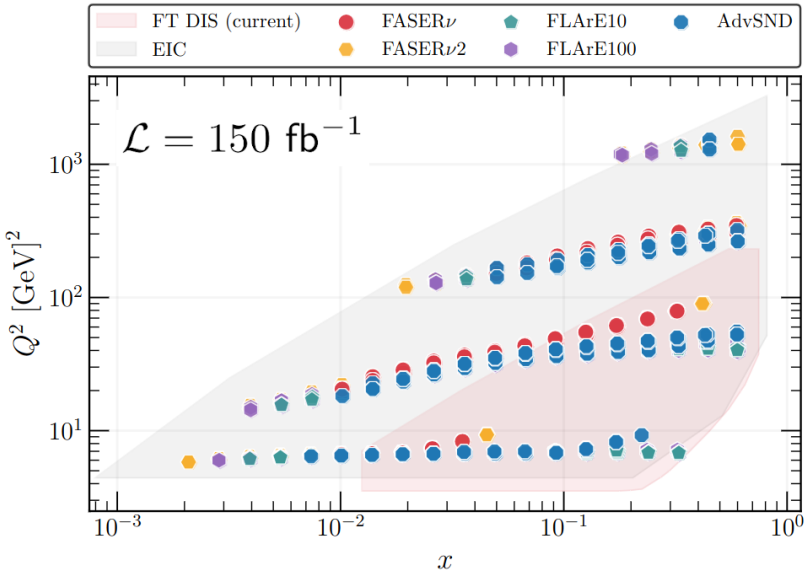
$$\mathcal{A}_{\sigma^{c\bar{c}}} = \frac{\sigma^{c}_{\rm red} - \sigma^{\bar{c}}_{\rm red}}{\sigma^{c\bar{c}}_{\rm red}}$$

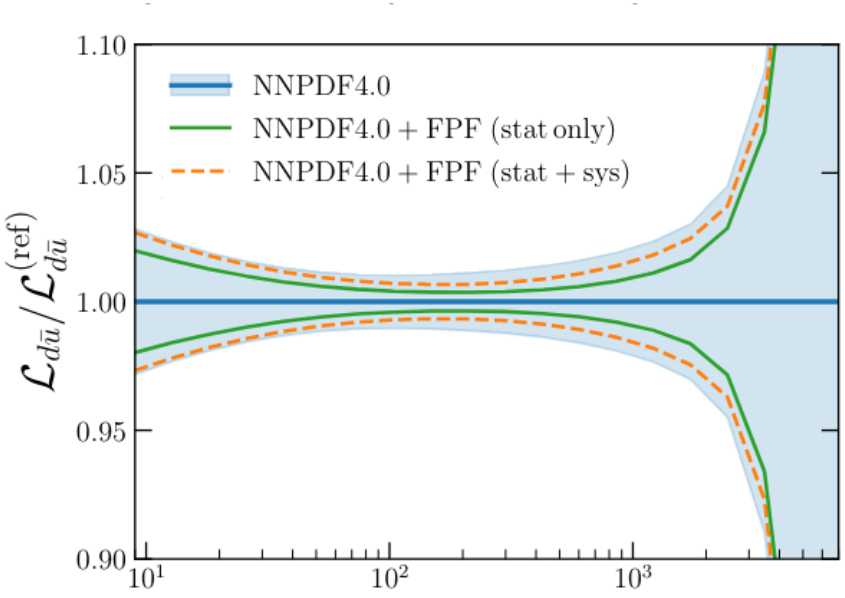


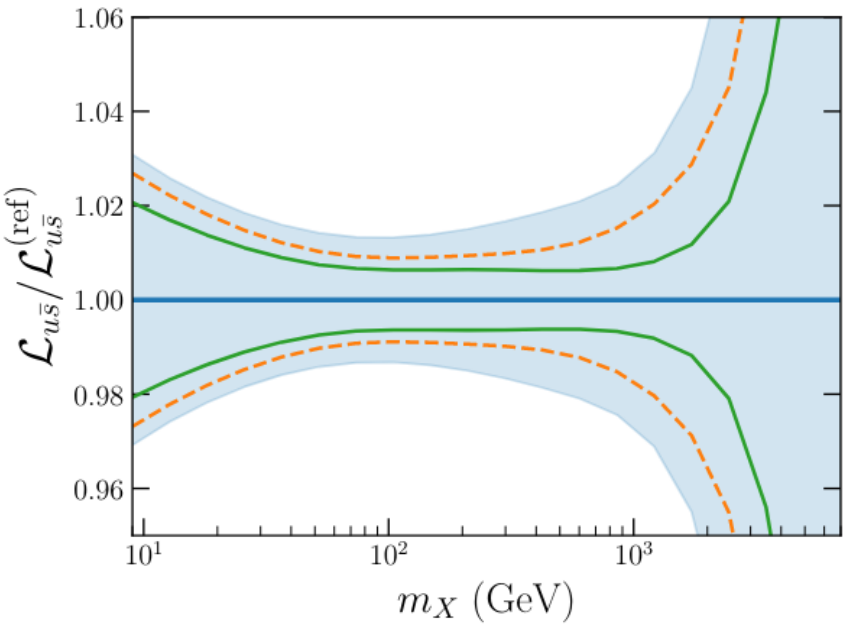


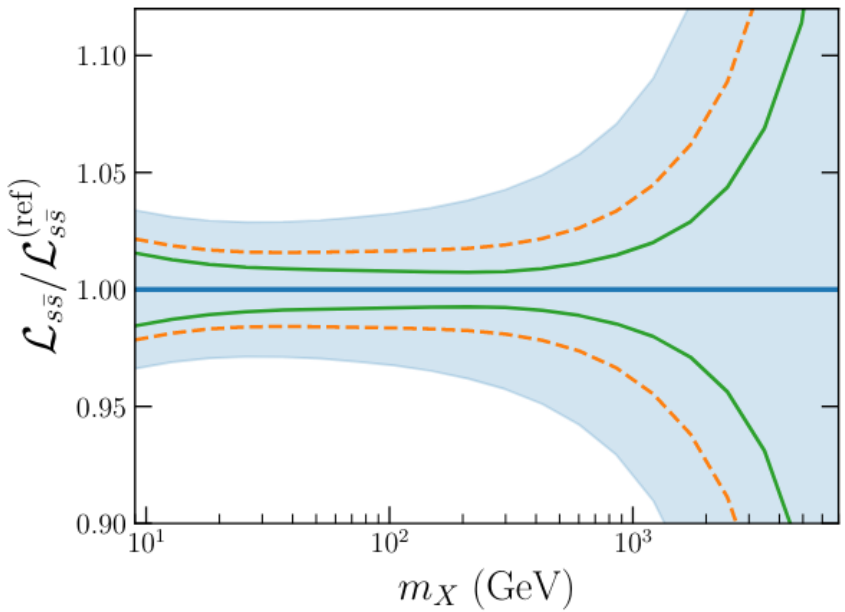
[PRD 103 (2021) 096005; see also arXiv:; arXiv:2311.00743]

Impact of future data: FPF



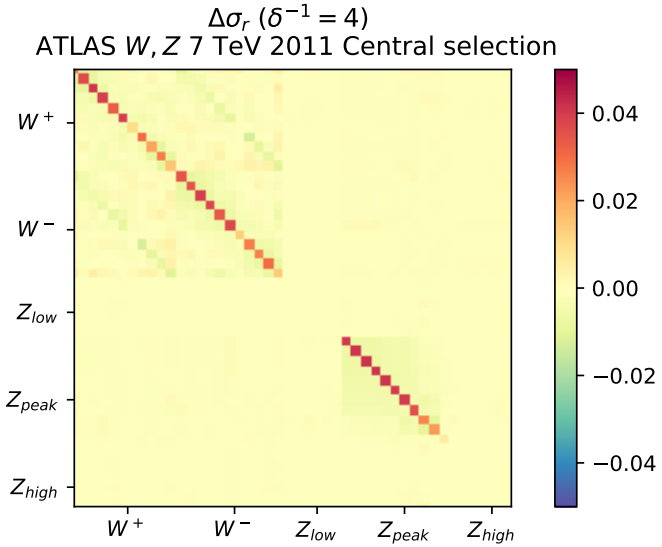


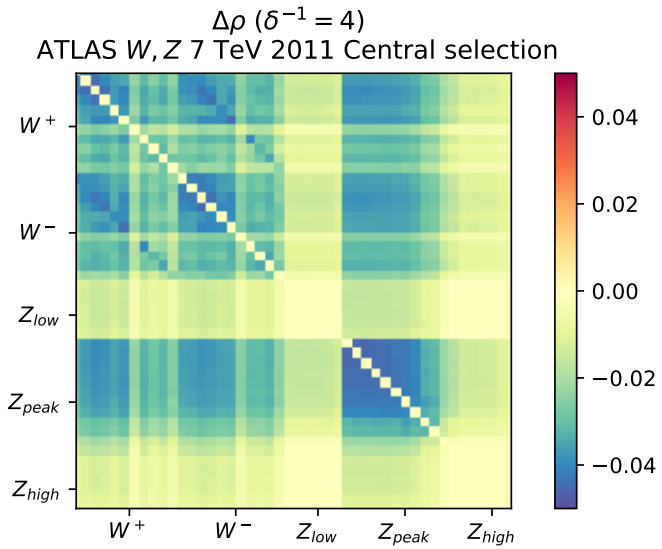




[arXiv:2309.09581; see T. Mäkelä's talk]

Good knowledge of experimental correlations is important





Assumptions:

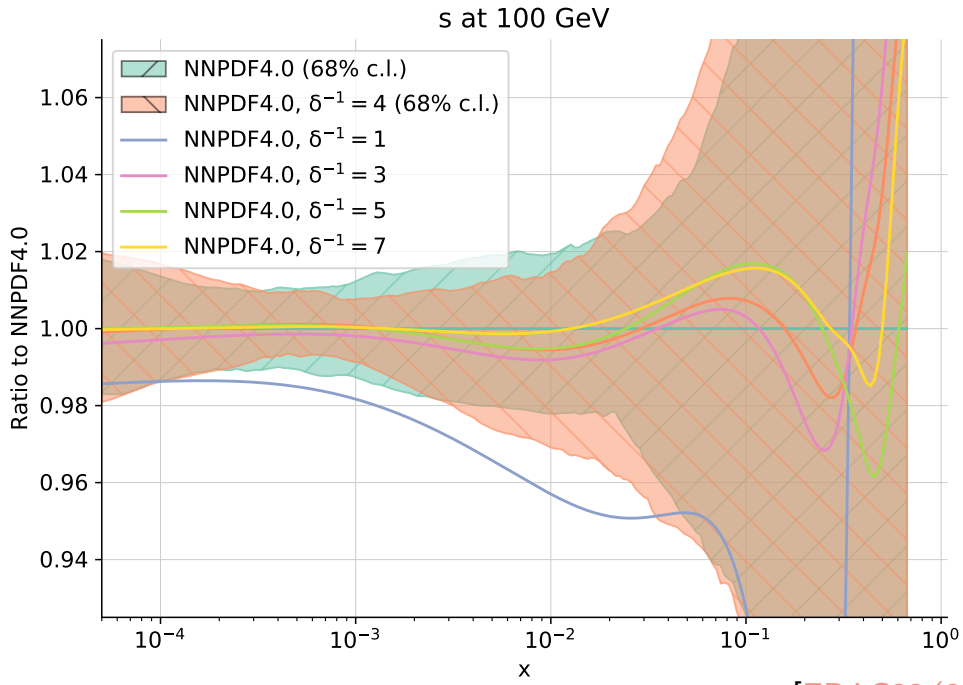
correlations are determined less precisely than variances

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]

Data inconsistency: tensions between data sets

Give more weight to a data set p

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

Refit: the total χ^2 will increase

Which data sets get worse? How much?

Refit: the data set χ^2_p will decrese

Self-consistency? Inconsistency?

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

Inconsistency clearly spotted

unnatural PDF shapes appear

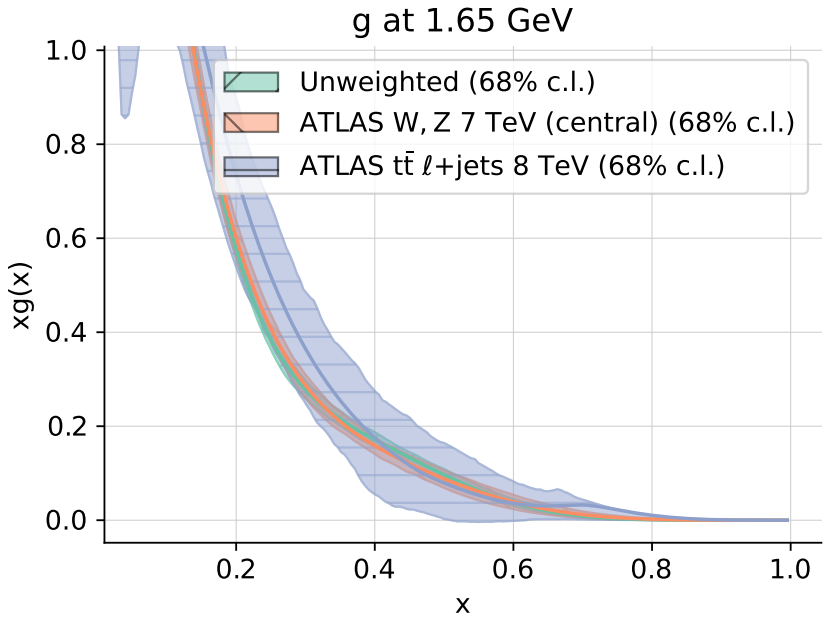
error in other data sets increases

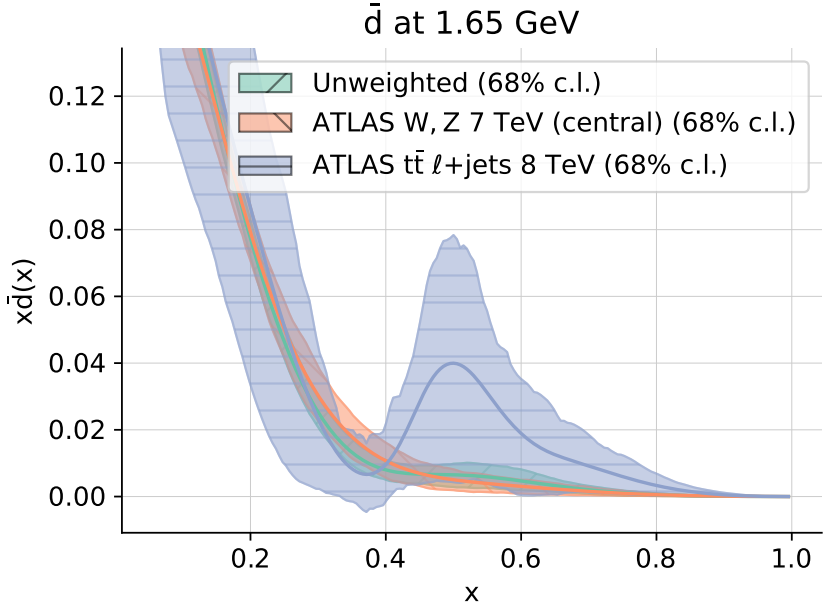
Otherwise global fit quality

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]





Inconsistent closure tests

Generate pseudodata with statistical and systematic uncertainties

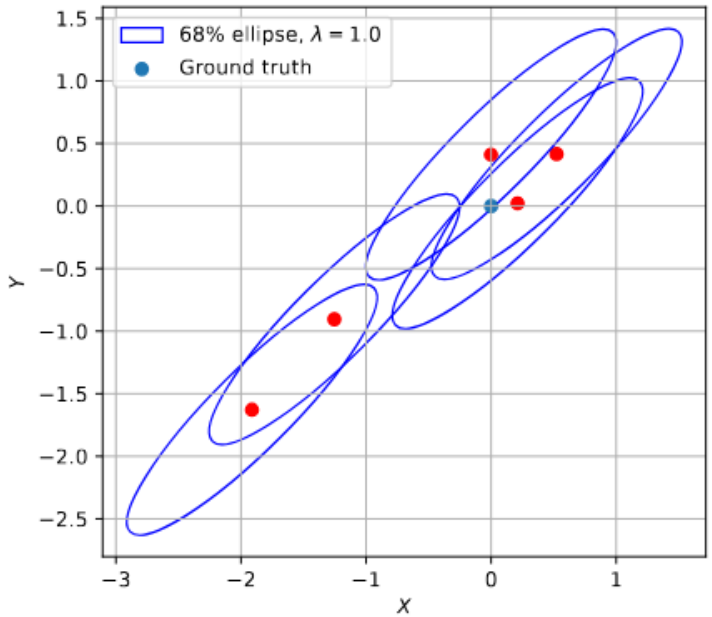
$$C = C^{\text{stat}} + C^{\text{syst}} \qquad C_{ij}^{\text{syst}} = \sum_k \Delta_i^k \Delta_j^k \qquad \Delta_i^k \text{ is the k-th sys. unc. for the i-th point}$$

Assume systematic uncertainties are underestimated and perform a fit with

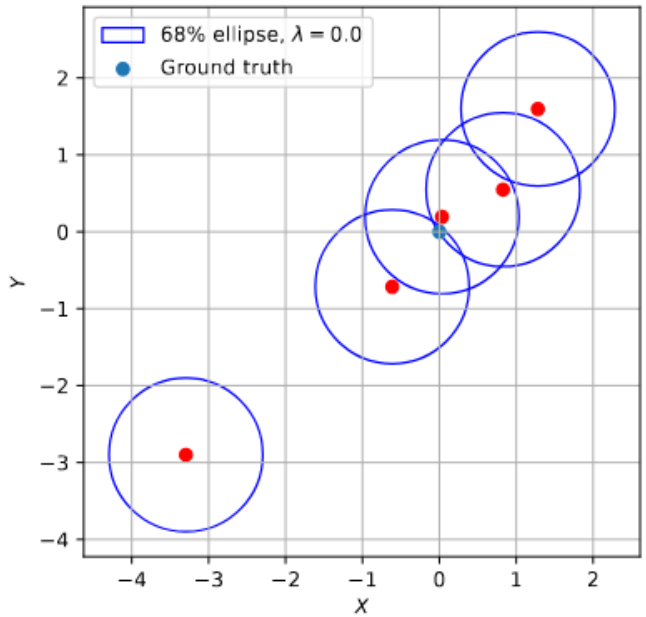
$$\Delta_i^k \to \lambda \Delta_i^k \qquad \lambda = 1 \text{ consistency} \qquad \lambda = 0 \text{ extreme inconsistency}$$

PREDICTED UNCERTAINTY ON GENERATED DATA

CONSISTENT



EXTREME INCONSISTENCY



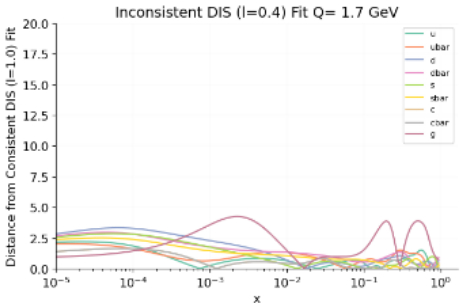
[arXiv:2503.17447]

Inconsistent closure tests

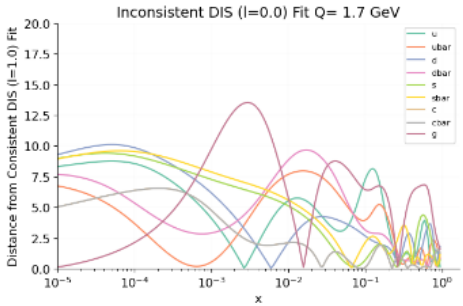
DISTANCES: $\frac{f^{\text{consistent}} - f^{\text{inconsistent}}}{\sigma^{\text{PDF}}/\sqrt{N_{\text{rep}}}}$; $d \sim 1 \Rightarrow$ STATISTICAL EQUIVALENCE

DIS: BULK INCONSISTENCY

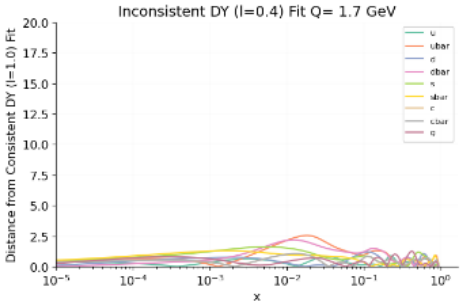
$\lambda = 0.4$: MODEL CORRECTS

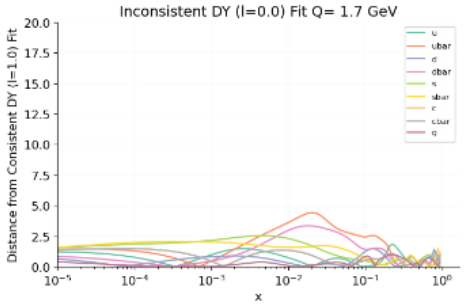


$\lambda = 0$: MODEL FAILS

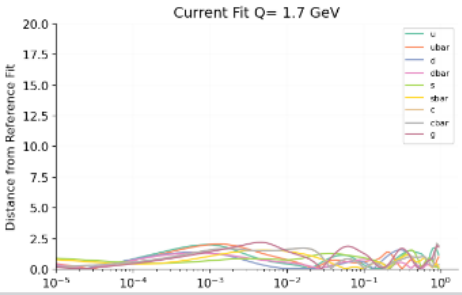


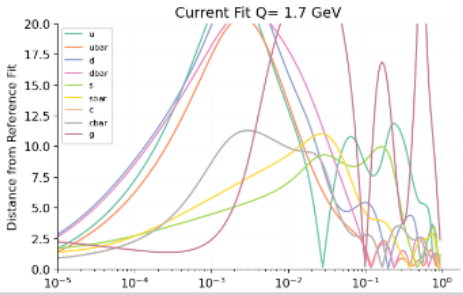
DY: SINGLE DATASET INCONSISTENCY





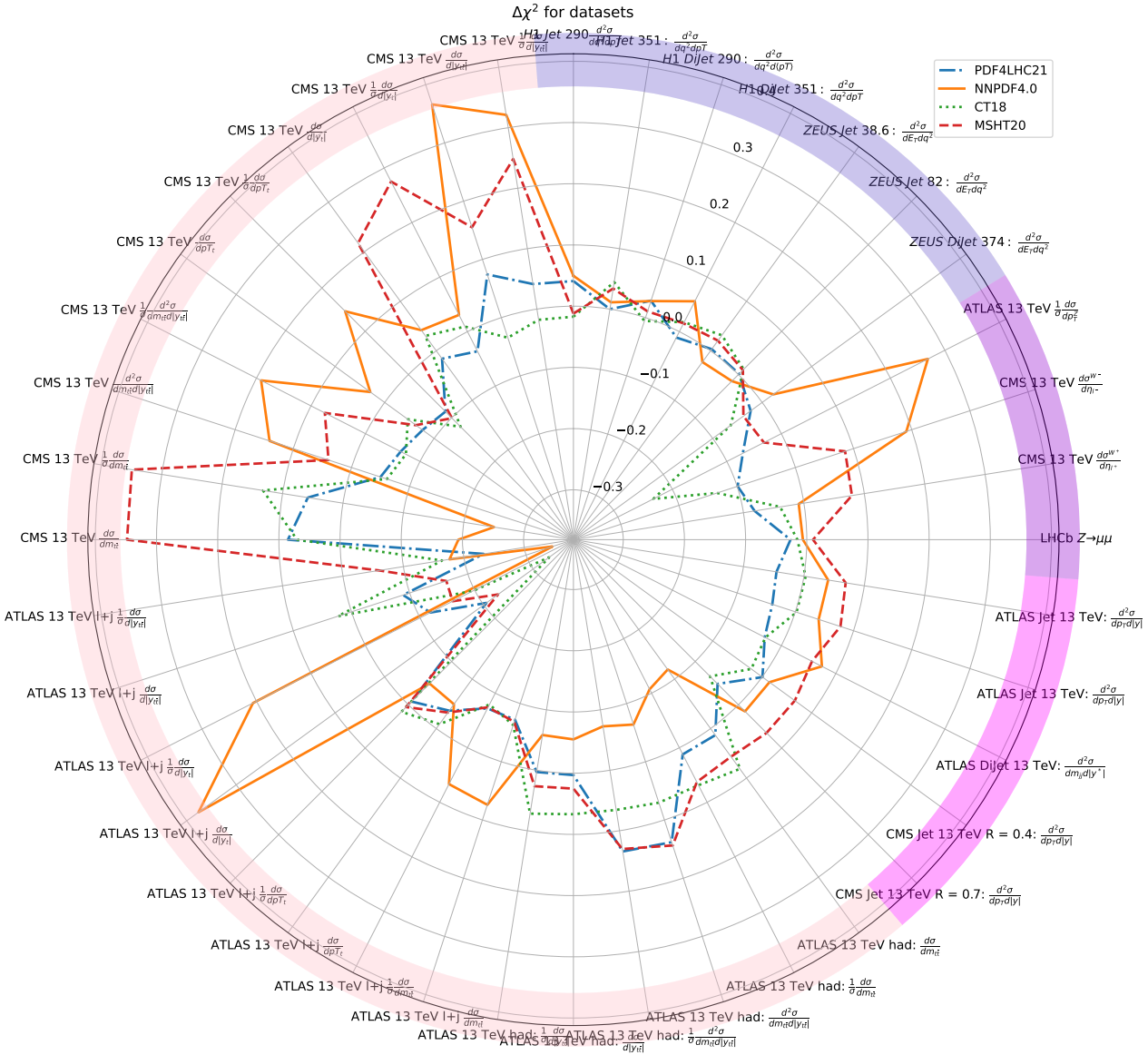
JETS: HIGH-IMPACT INCONSISTENCY





The ML model corrects for inconsistency except in extreme cases [arXiv:2503.17447]

How do PDFs generalise on unseen data?



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

[JHEP 07 (2025) 067]