Parton Distribution Functions in the Precision Era: Status, Challenges and New Tools

SHARP 2026 First Network Conference

Emanuele R. Nocera

Università degli Studi di Torino and INFN, Torino

5 March 2026

UNIVERSITÀ DI TORINO

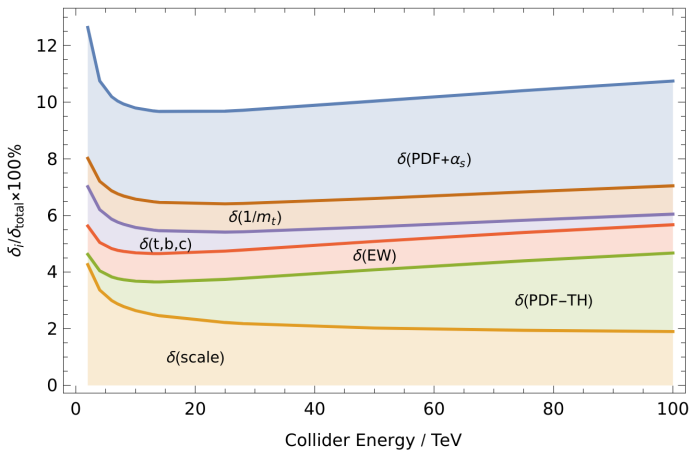
Parton Distribution Functions (at the LHC)

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

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

Precision

Higgs production in gluon-gluon fusion

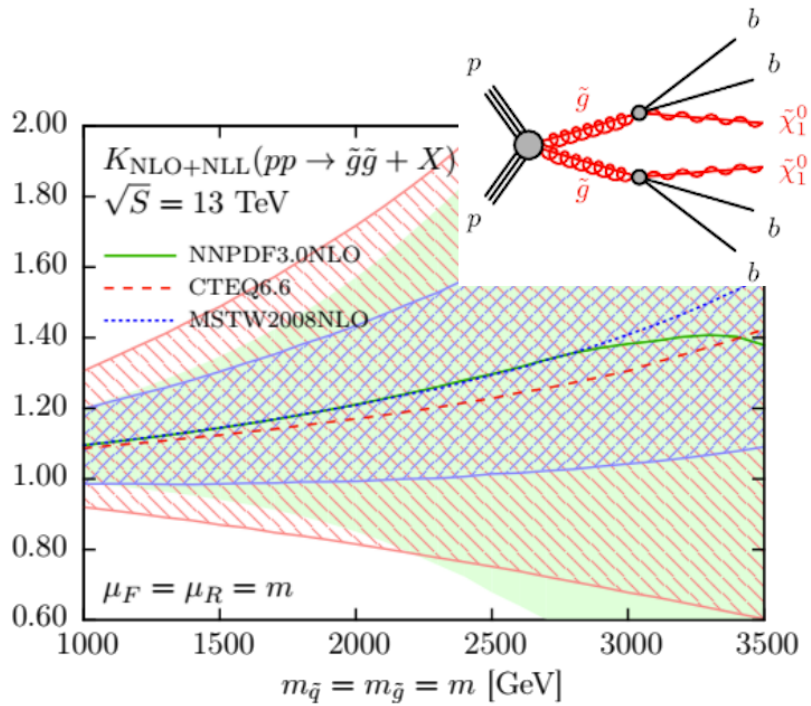


[CERN Yellow Rep.Monogr. 7 (2019) 221]

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

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

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

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

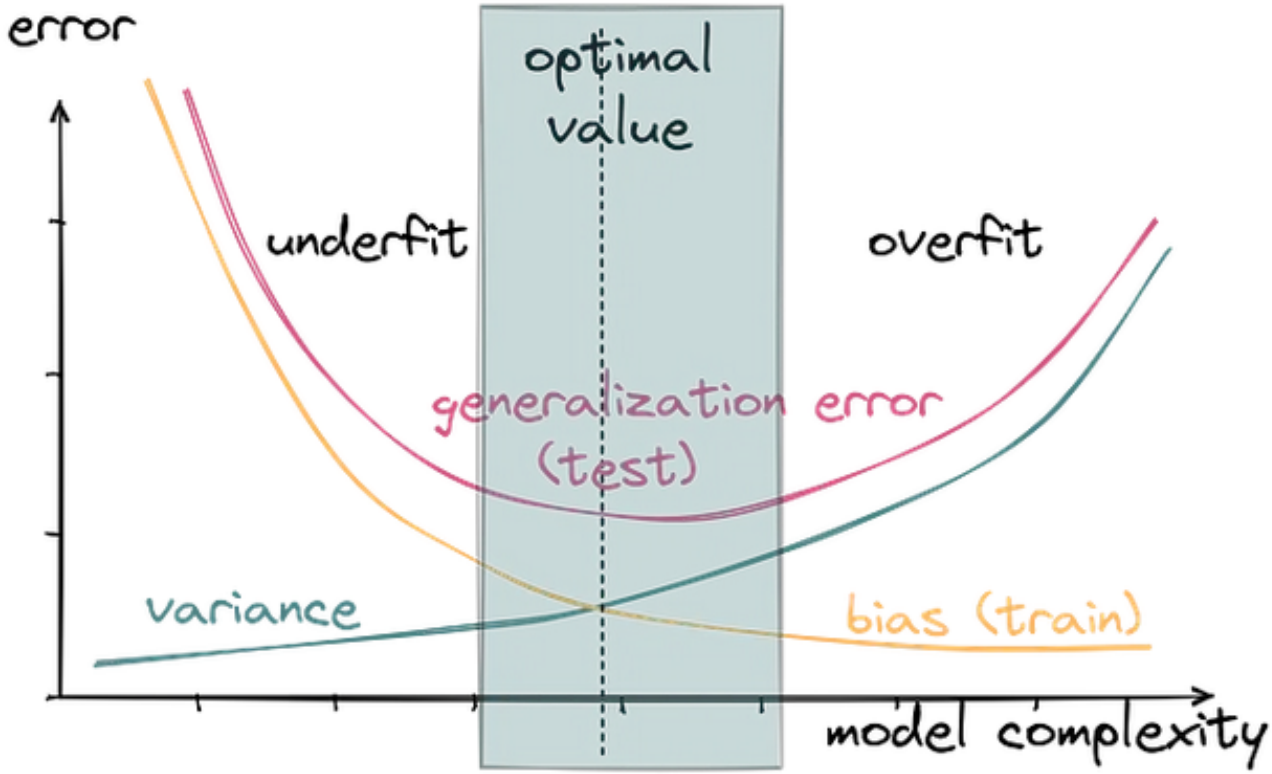
Maximum likelihood: $\mathcal{P}(f|D) \longrightarrow f_0$

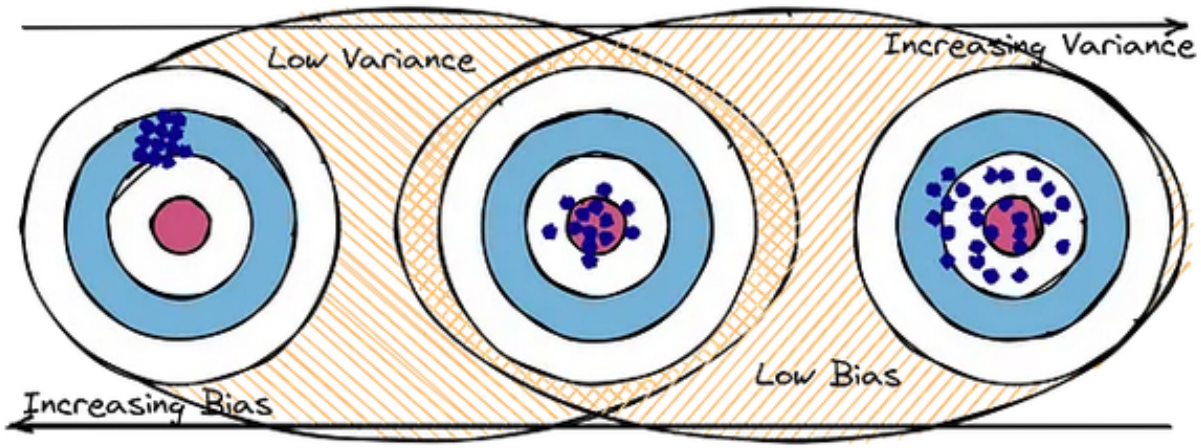
$E[\mathcal{O}] \approx \mathcal{O}(f_0)$

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

Interplay between DATA, THEORY, and METHODOLOGY

How can ML help? Generalisation!





Validation: closure tests

Fit PDFs to pseudodata generated assuming a known underlying law

Assess the faithfulness of uncertainties

multi-closure tests

Define bias and variance

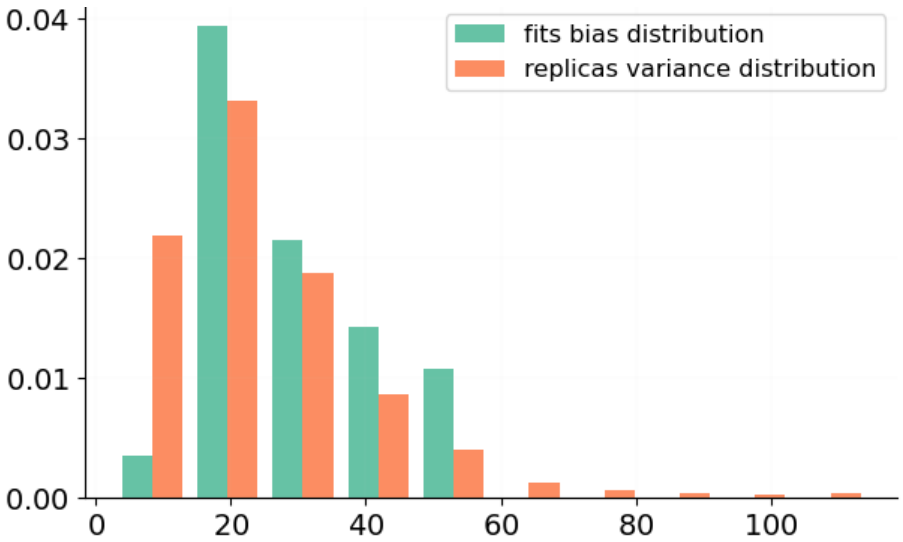
bias difference of central prediction and truth

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]

Characterise PDF ucnertainties

closure test levels

lvl 0: unfluctuated pseudodata

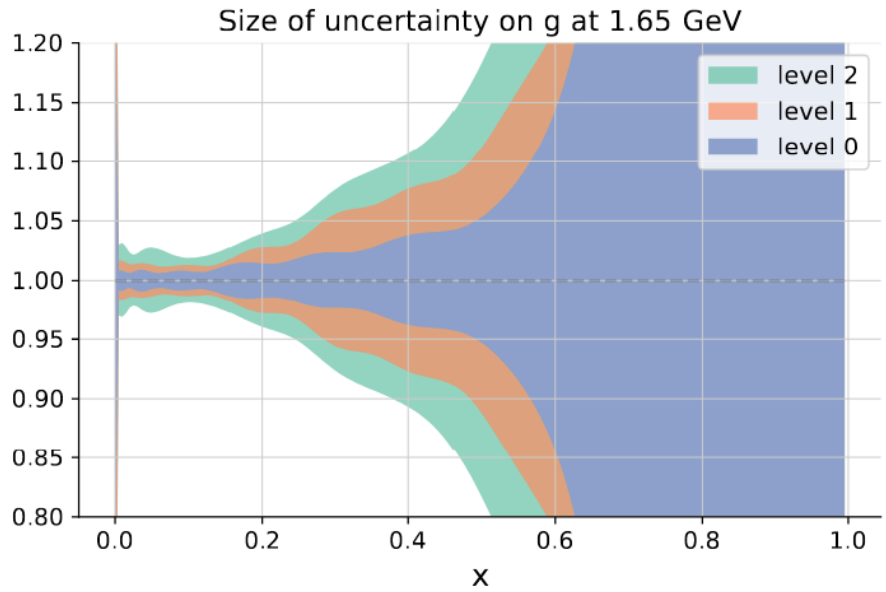
$\longrightarrow$ interpolation/extrapolation uncertainty

lvl 1: statistical noise in the pseudodata

$\longrightarrow$ functional uncertainty

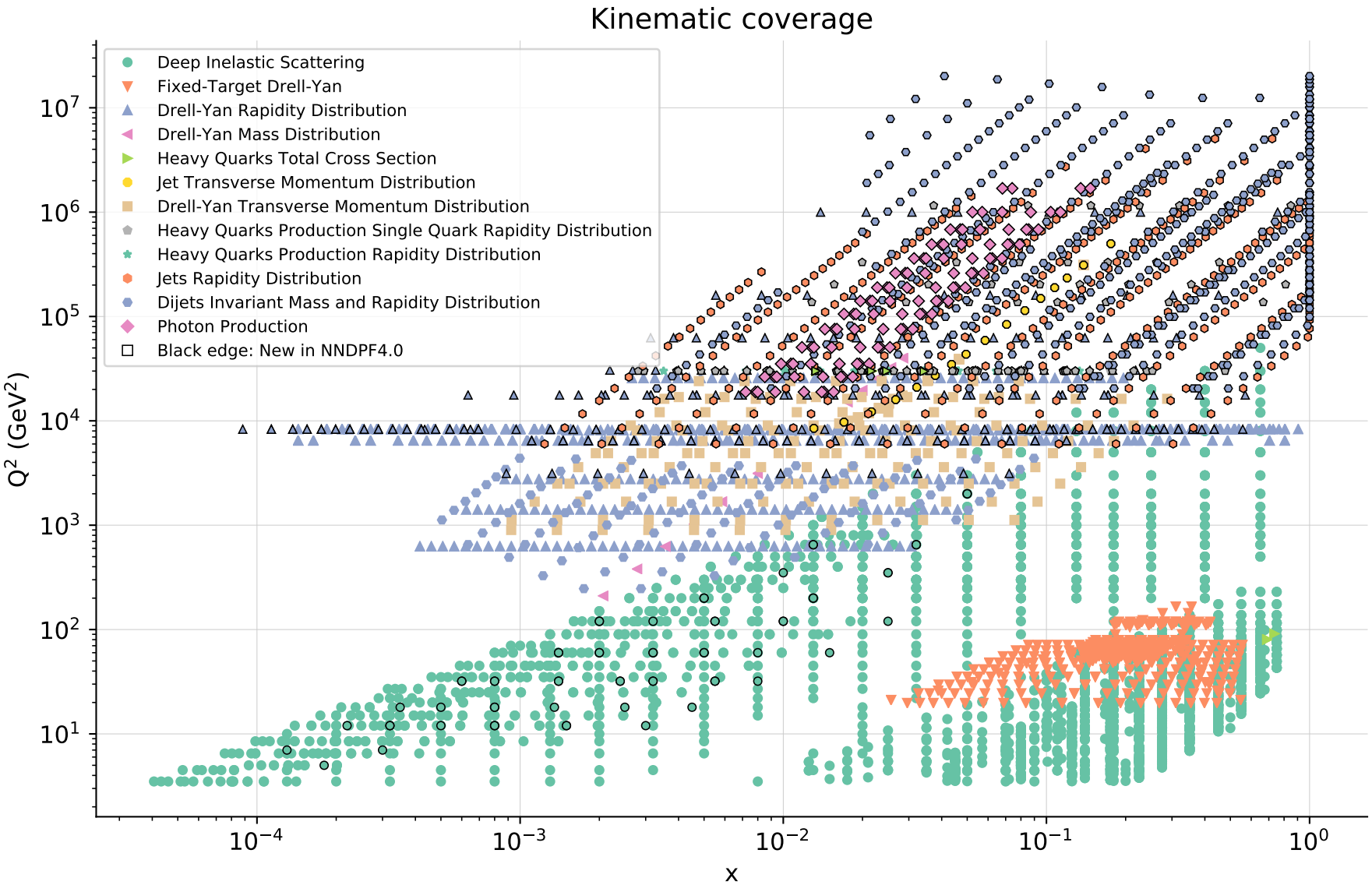
lvl 2: add replica generation

$\longrightarrow$ data uncertainty



1. Data

The data set



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

Data inconsistency: experimental correlations

Single inclusive jet data from ATLAS 7 TeV

default correlations: terrible χ^2

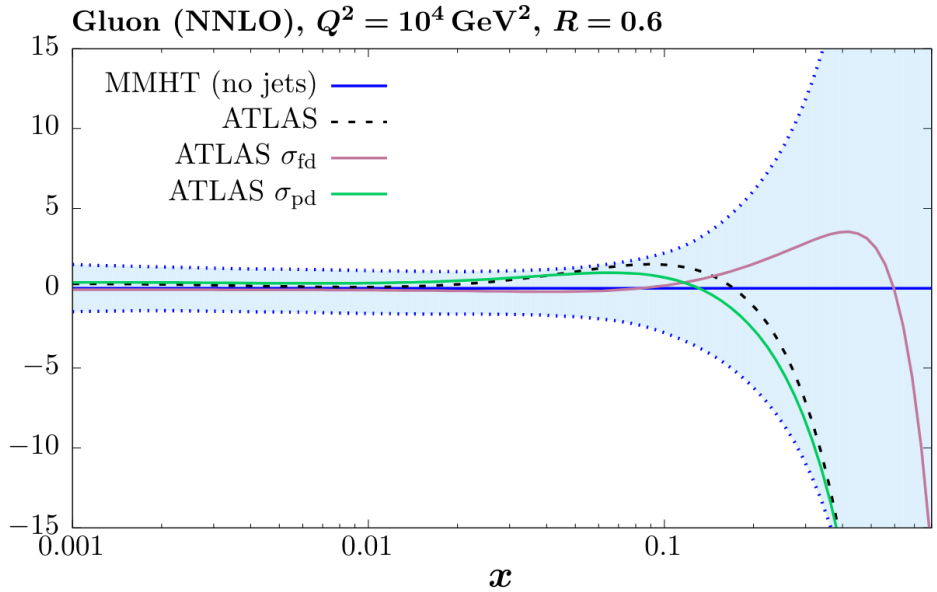
(correlations across rapidity bins)

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

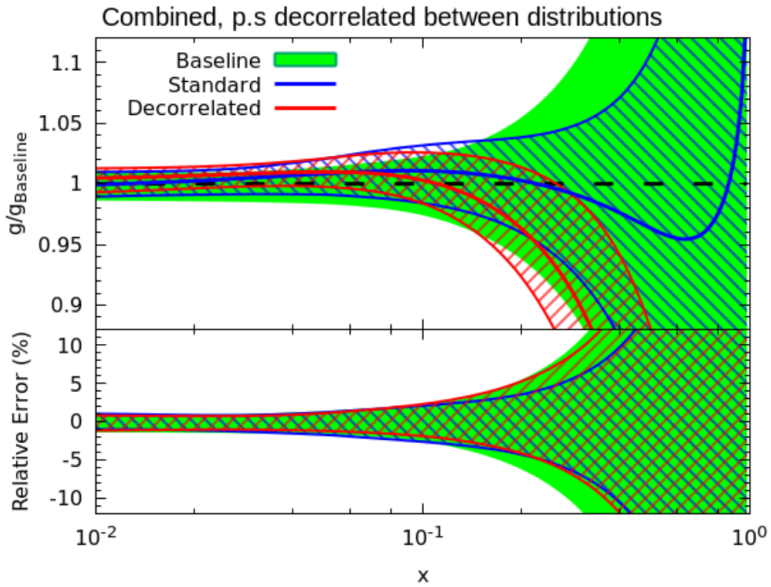
(correlations across different spectra)

decorrelation models: improve the fit a lot

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

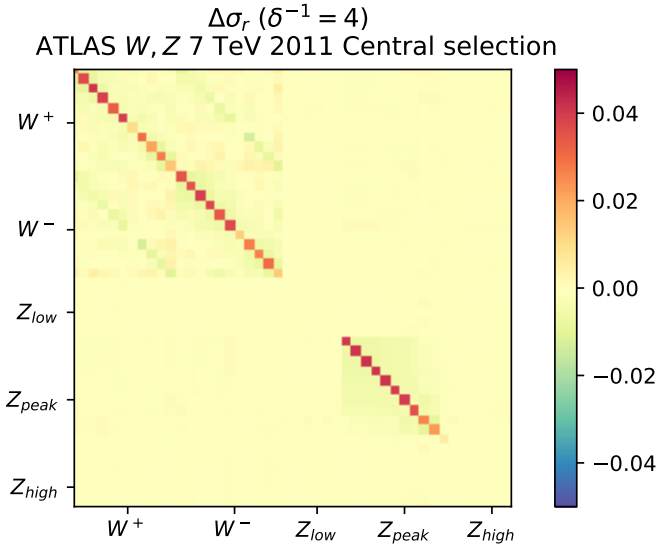
appreciable effect on the extracted gluon

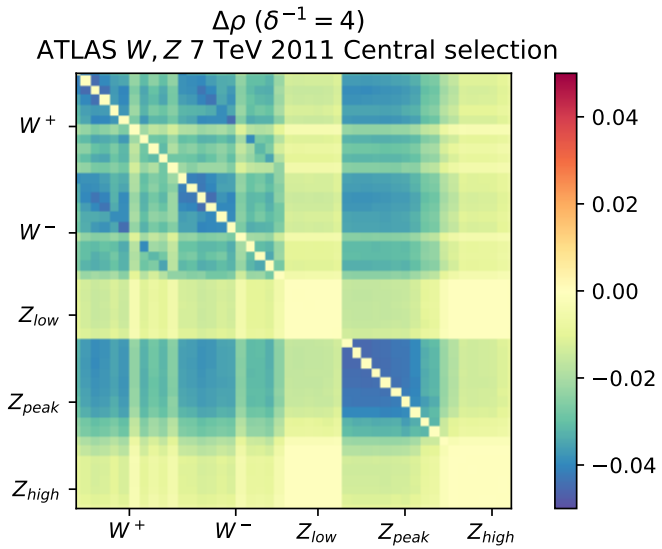
different gluon depending on the top spectrum



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

Good knowledge of experimental correlations is important





Assumptions:

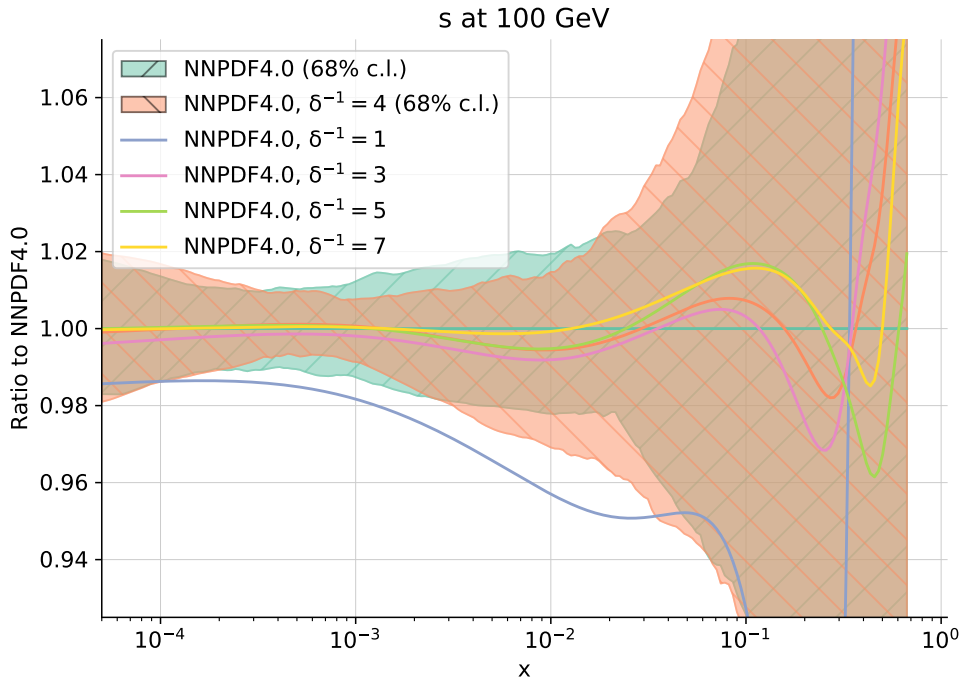
correlations are determined less precisely than variances

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]

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

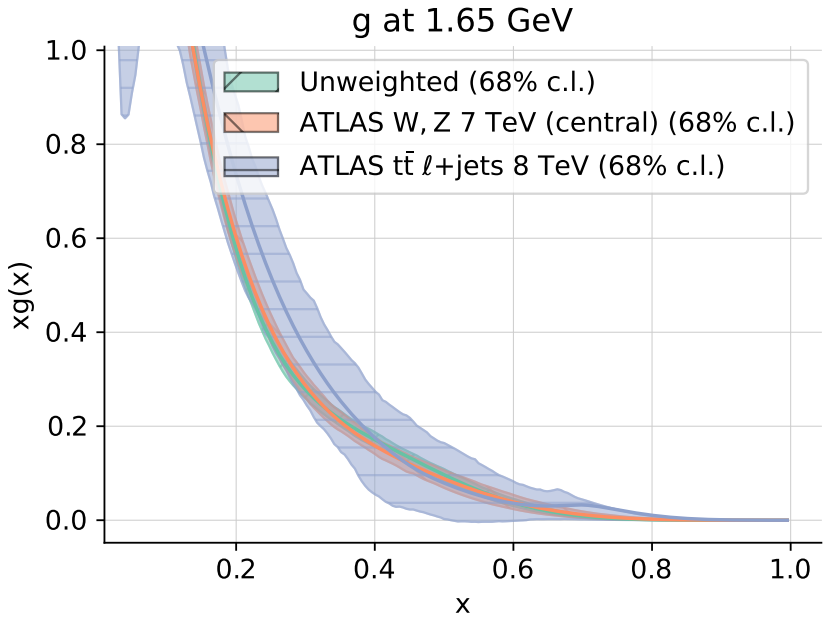
error in other data sets increases

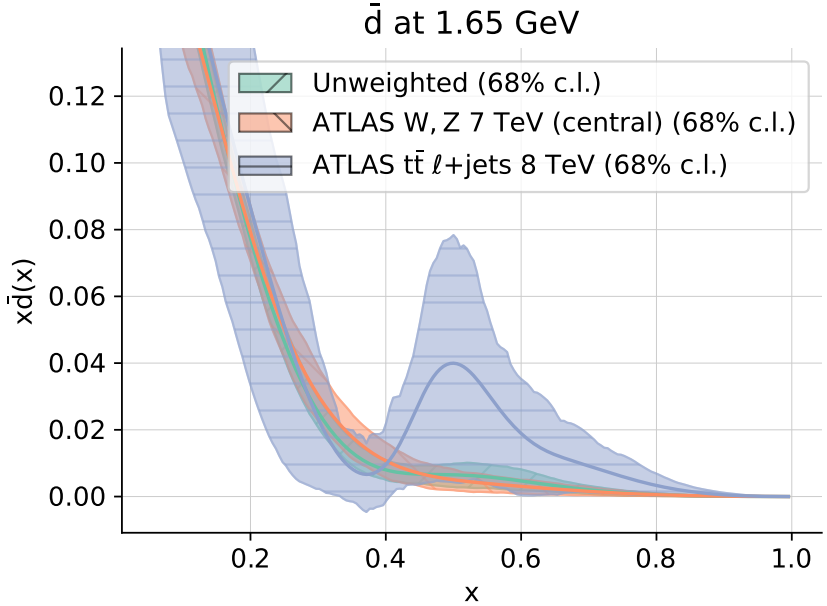
Otherwise global fit quality

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

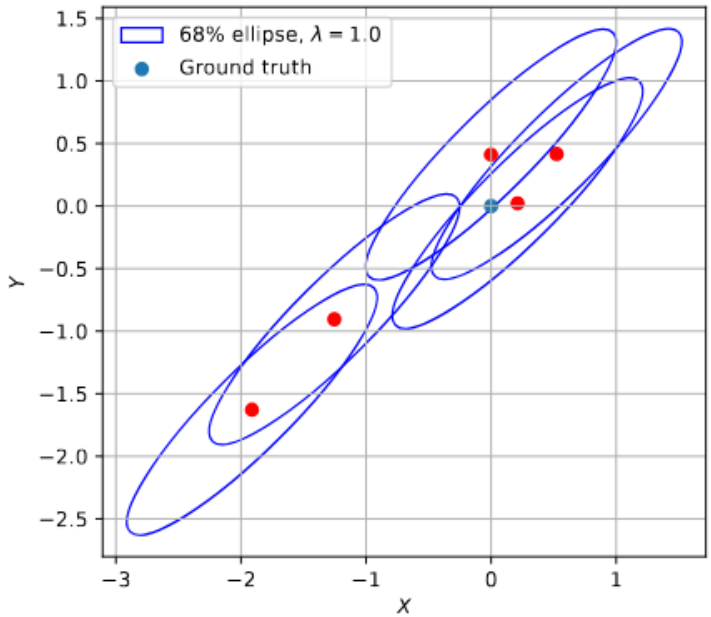
$$C = C^{\text{stat}} + C^{\text{syst}} \qquad C_{ij}^{\text{syst}} = \sum_k \Delta_i^k \Delta_j^k \qquad \Delta_i^k \text{ is the k-th sys. unc. for the i-th point}$$

Assume systematic uncertainties are underestimated and perform a fit with

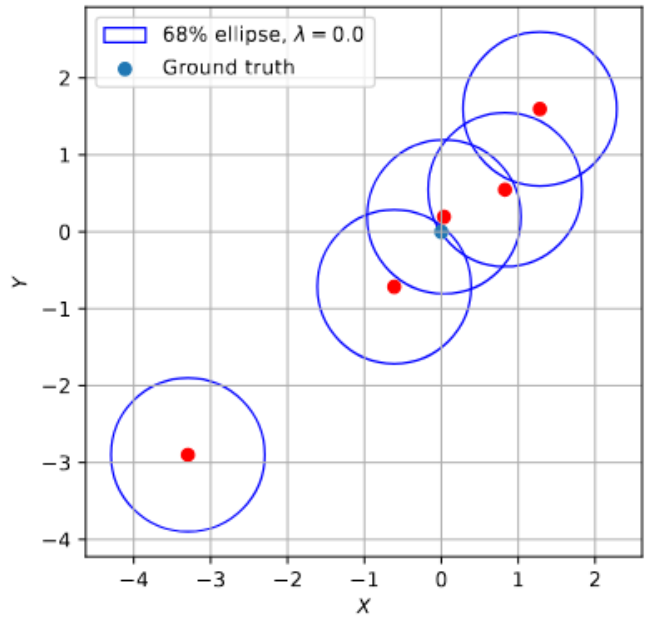
$$\Delta_i^k \to \lambda \Delta_i^k \qquad \lambda = 1 \text{ consistency} \qquad \lambda = 0 \text{ extreme inconsistency}$$

PREDICTED UNCERTAINTY ON GENERATED DATA

CONSISTENT



EXTREME INCONSISTENCY



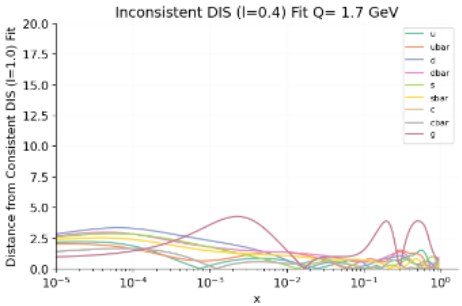
[arXiv:2503.17447]

Inconsistent closure tests

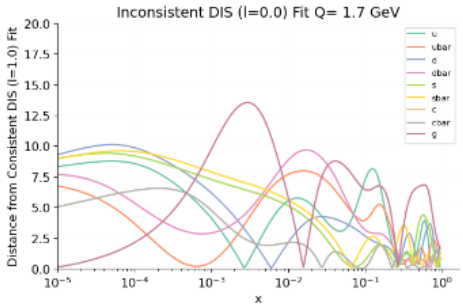
DISTANCES: $\frac{f^{\text{consistent}} - f^{\text{inconsistent}}}{\sigma^{\text{PDF}}/\sqrt{N_{\text{rep}}}}$; $d \sim 1 \Rightarrow$ STATISTICAL EQUIVALENCE

DIS: BULK INCONSISTENCY

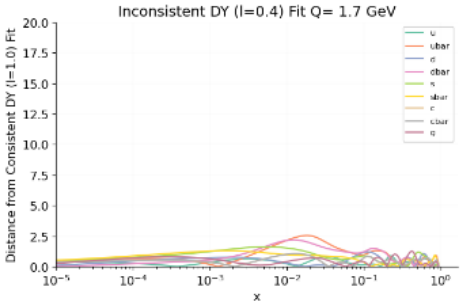
$\lambda = 0.4$: MODEL CORRECTS

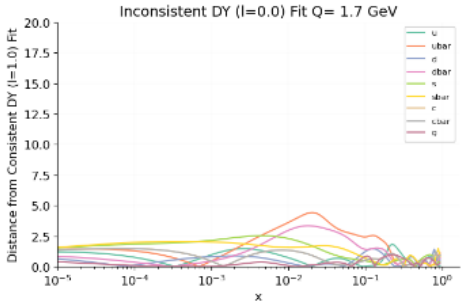


$\lambda = 0$: MODEL FAILS

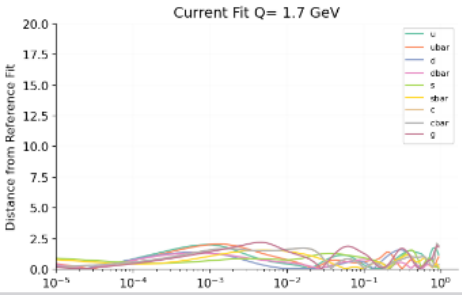


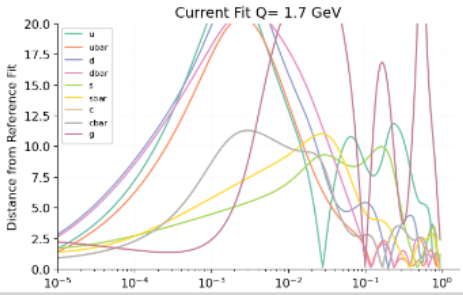
DY: SINGLE DATASET INCONSISTENCY





JETS: HIGH-IMPACT INCONSISTENCY





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

2. Theory

Theory uncertainties: the covariance matrix method

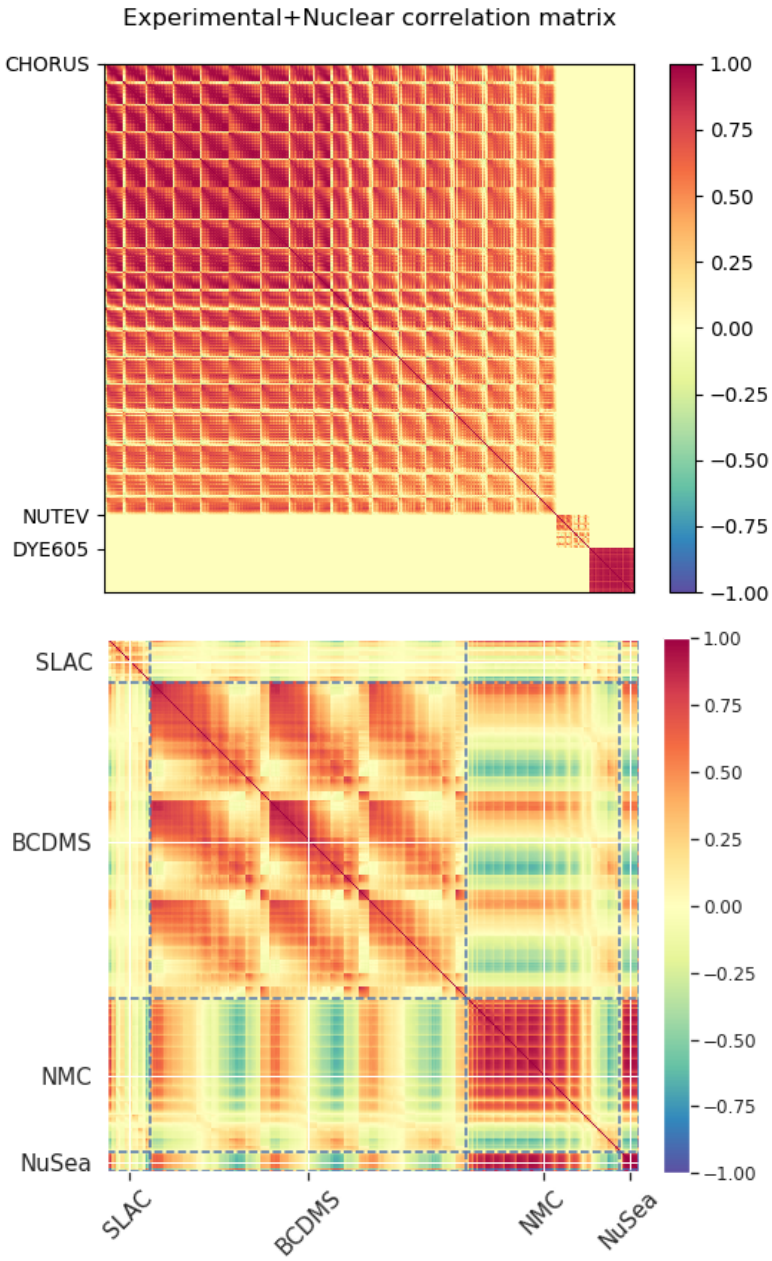
PDF determination is affected by theory uncertainties: nuclear, MHOU, PCs, α_s, ...
Idea: associate nuisance parameters to these uncertainties and
implement Bayesian parameter estimation by means of a theoretical covariance matrix

$S_{ij} = \beta_i \beta_j$ (sampling & fitting)	correlated uncertainty covariance matrix
$T \to T + \lambda\beta$	nuisance parameter λ on prediction T
$P(T\|D,\lambda) \propto \exp\left[-\frac{1}{2}(T+\lambda\beta-D)^T C^{-1}(T+\lambda\beta-D)\right]$	Probability of prediction T given the data D and the nuisance parameter λ
$P(\lambda) \propto \exp\left(-\frac{\lambda^2}{2}\right)$	Uncertainty on nuisance parameter
$P(T\|D) \propto \exp\left[-\frac{1}{2}(T-D)^T (C+S)^{-1}(T-D)\right]$	Probability of prediciton T given the data
$P(\lambda\|T,D) \propto \exp\left[-\frac{1}{2} Z^{-1}(\lambda - \bar{\lambda}(T,D))^2\right]$	Posterior distribution of the nuisance parameter λ

with $\bar{\lambda}(T,D) = \beta^T (C+S)^{-1}(D-T)$, $Z = 1 - \beta^T (C+S)^{-1}\beta$, and C exp. cov. mat.

[EPJ C81 (2021) 830]

Nuclear uncertainties



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

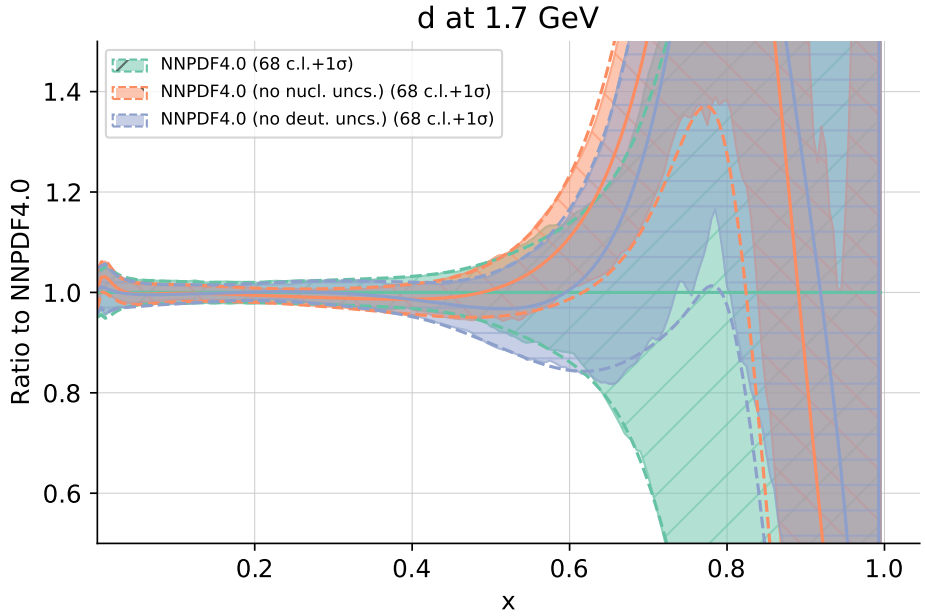
$$\Delta_i^{(k)} = T_i^N[f_N^{(k)}] - T_i^N[f_p]$$

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

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

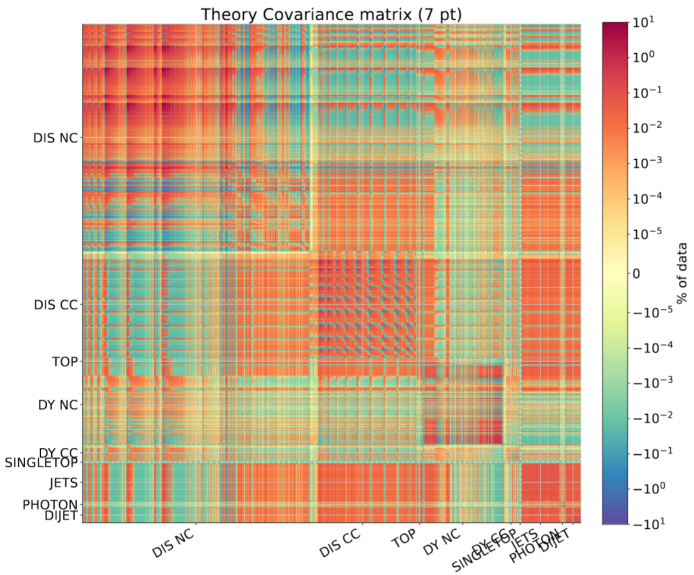
The bulk of the effect is due to nuclear uncertainties for heavy nuclei at large x

Reduced tension between DIS and LHC data



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

Missing higher order uncertainties



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

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

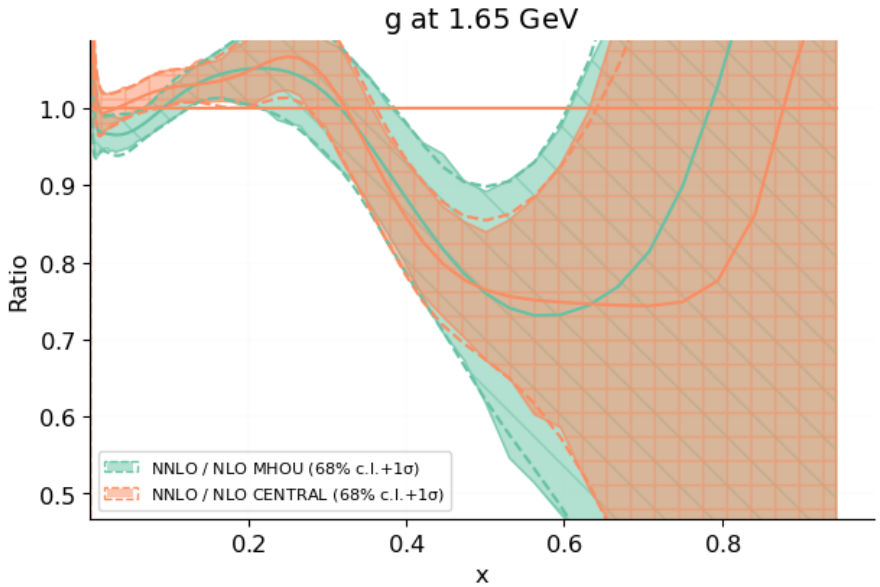
$\chi^2_{\text{tot}} = 1.23 \to \chi^2_{\text{tot}} = 1.34$ (no MHOU, NLO)

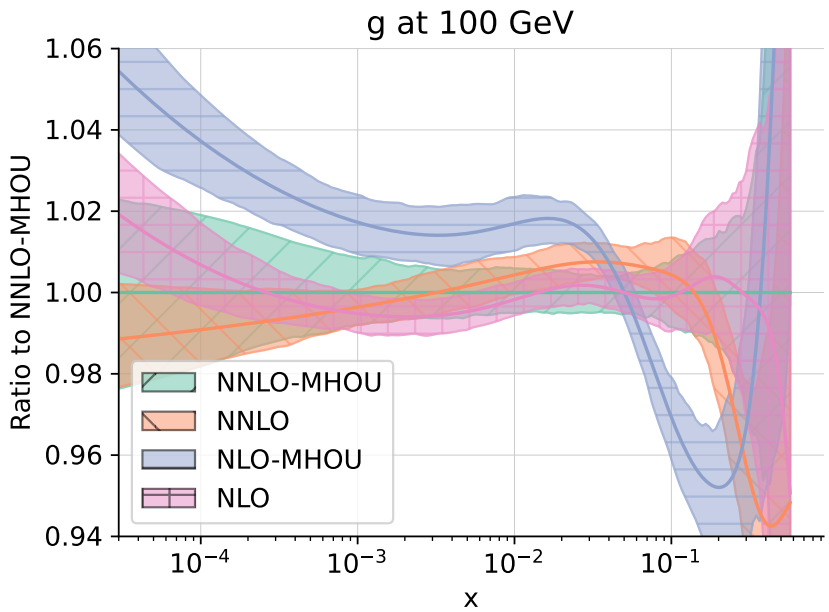
$\chi^2_{\text{tot}} = 1.13 \to \chi^2_{\text{tot}} = 1.17$ (no MHOU, NNLO)

Tensions relieved

Faster perturbative convegence

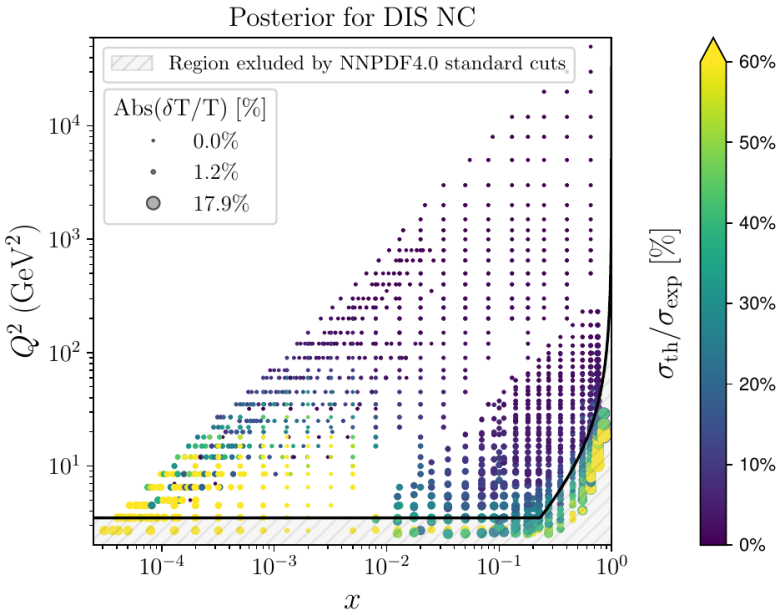
Small increase of PDF uncertainties

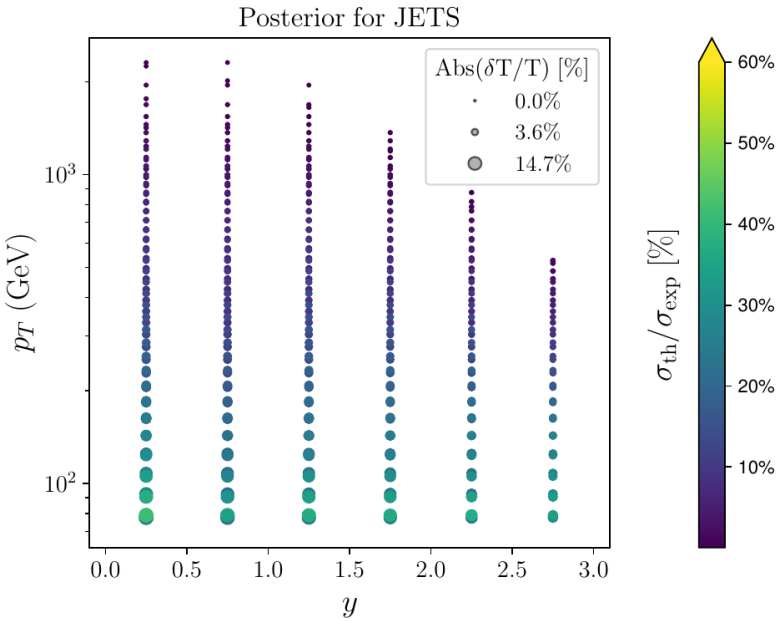




[EPJ C79 (2019) 282; EPJ C81 (2021) 37; EPJ C84 (2024) 517]

Higher twist and power corrections





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

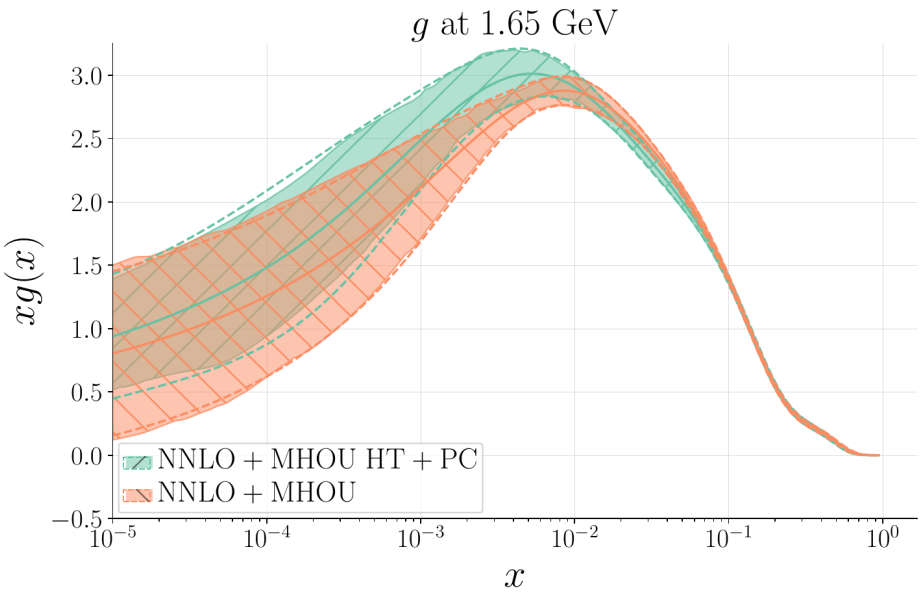
$$\Delta_i^{(k)} = T_i^{\rm HT,PC(k)} - T_i^{\rm LT}$$

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

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

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

Tensions relieved, shift of PDF central value



[arXiv:2511.14387]

3. Methodology

The methodology matters: PDFs

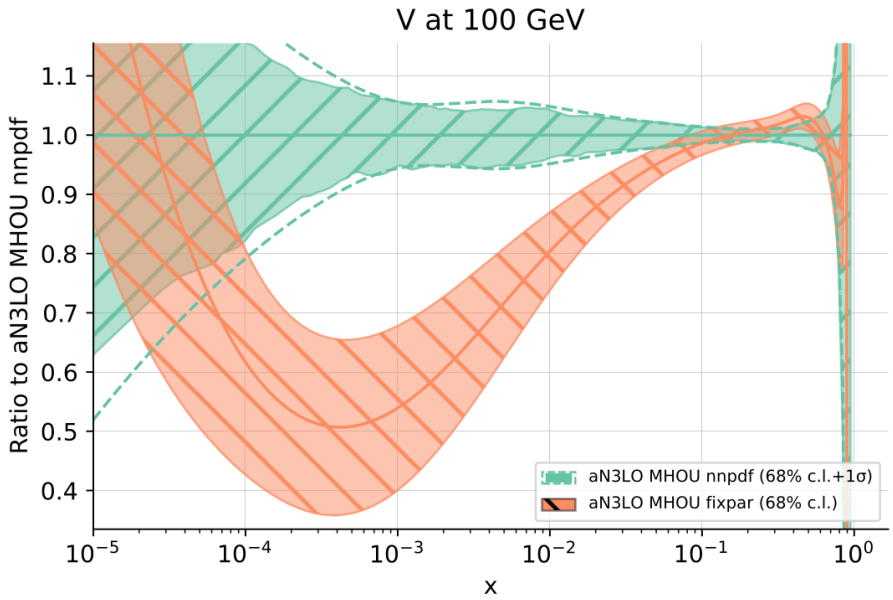
Which differences are driven by theory and which are driven by methodology?

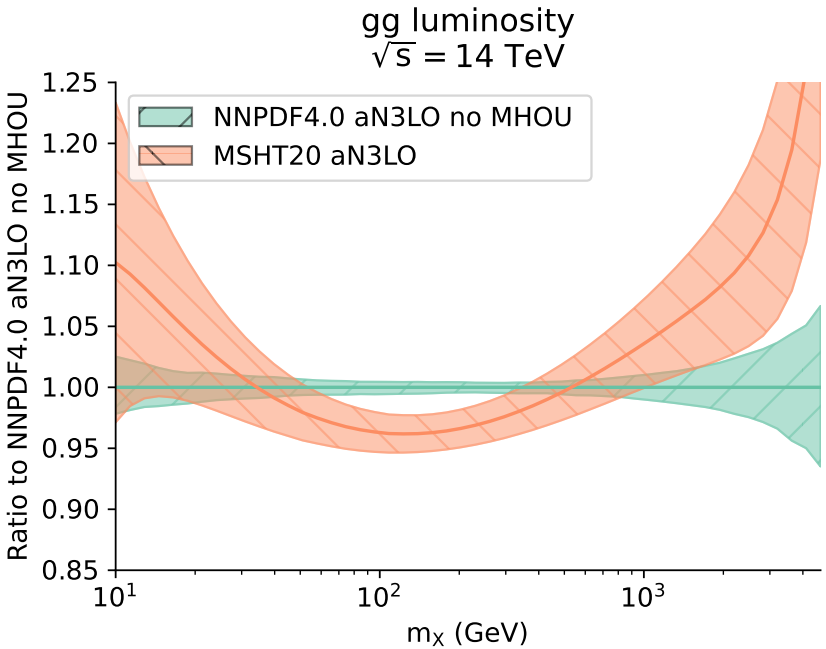
Freeze theory and data (NNPDF) and change methodology (NNPDF vs MSHT)

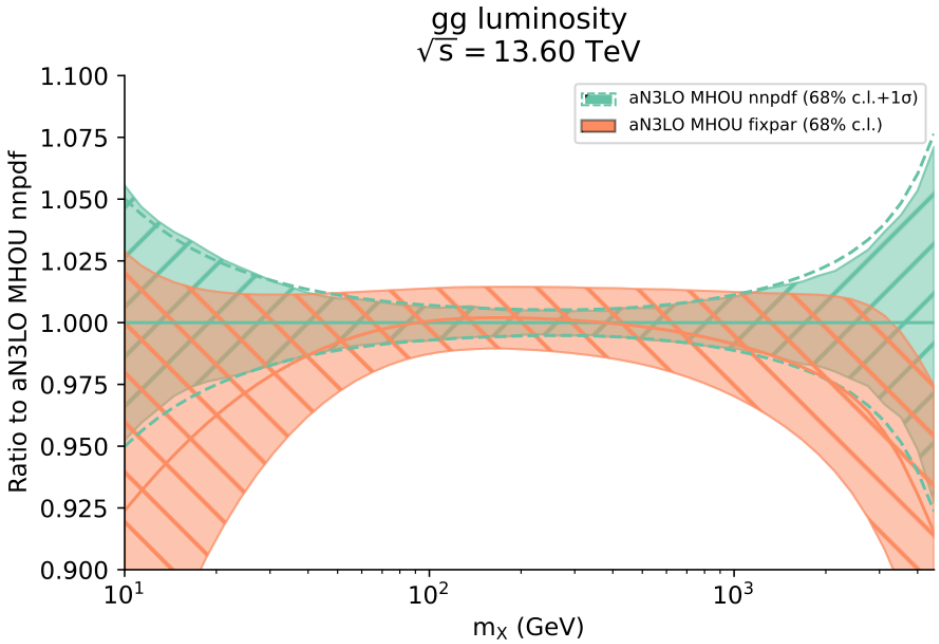
Some are due to theory (gluon)

Some are due to methodology (valence)

[arXiv:2602.07118: see also EPJ C85 (2025) 316]







The methodology matters: TMDs

Consider three TMD models:

PV19 [JHEP 07 (2020) 117], MAP22 [JHEP 10 (2022) 127], MAPNN [PRL 135 (2025) 021904]

Roughly speaking: rigid, semi-rigid, flexible

Consider the subset of Drell-Yan data fitted in the three determinations

Generate pseudodata with each model and fit it (not necessarily with the same model)

Perform a multi-closure test (50 fits of 100 replicas each)

generated	fitted	R_{bv}	$\xi_{1\sigma}$	comments
PV19	PV19	1.577 ± 0.068	0.486 ± 0.012	underfitting, uncertainties too small
MAP22	MAP22	0.978 ± 0.045	0.688 ± 0.010	optimal fitting, uncertainties faithful
MAPNN	MAPNN	0.970 ± 0.097	0.714 ± 0.043	optimal fitting, uncertainties faithful
MAP22	PV19	—	—	closure test lvl 0 fails
PV19	MAP22	—	—	closure test lvl 0 fails
MAP22	MAPNN	1.032 ± 0.077	0.699 ± 0.050	optimal fitting, uncertainties faithful
PV19	MAPNN	0.510 ± 0.195	0.675 ± 0.051	overfitting, uncertainties faithful

[K. Laurent, E.R. Nocera, A. Signori, in preparation]

With great power comes great responsibility.

Using a neural network to parametrise the nonperturbative part of TMDs is possibly advantageous, however one MUST carefully optimise hyperparameters.

See talk by Juan Cruz-Martinez

4. Summary

Conclusions

Status

PDF determination is rapidly getting close to 1%. This opens up some challenges

Challenge 1: data set inconsistency

tools: regularisation of experimental correlations, weighted fits

The challenge and tools can be beneficial to TMD fits as well

Challenge 2: theory uncertainties

tools: Bayesian estimate via theory covariance matrix method

The challenge and tools can be usful for TMDs (*e.g.* b^* prescription)

Challenge 3: methodological robustness

tools: benchmarks, closure tests, and many other tools in Juan's talk

The challenge and tools are increasingly relevant for TMD fits

Conclusions

Status

PDF determination is rapidly getting close to 1%. This opens up some challenges

Challenge 1: data set inconsistency

tools: regularisation of experimental correlations, weighted fits

The challenge and tools can be beneficial to TMD fits as well

Challenge 2: theory uncertainties

tools: Bayesian estimate via theory covariance matrix method

The challenge and tools can be usful for TMDs (*e.g.* b^* prescription)

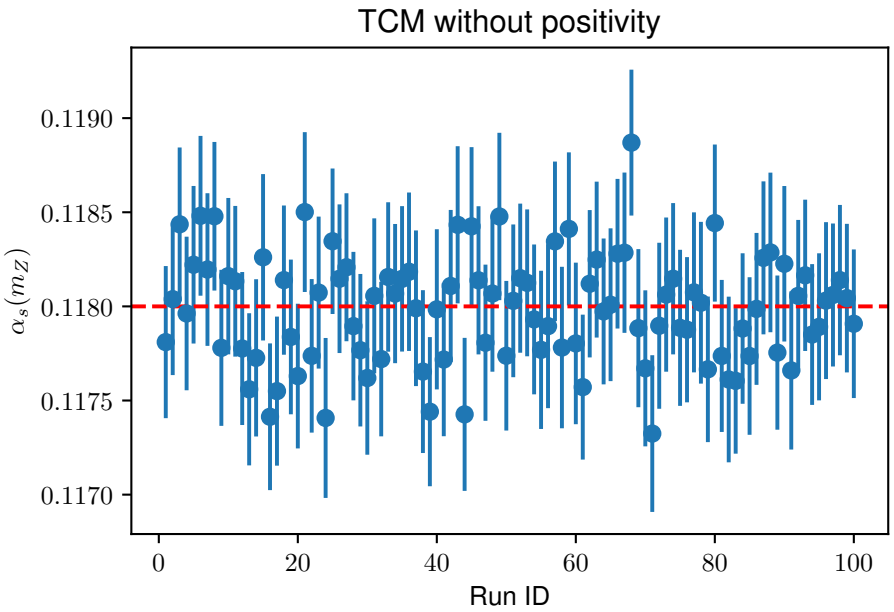
Challenge 3: methodological robustness

tools: benchmarks, closure tests, and many other tools in Juan's talk

The challenge and tools are increasingly relevant for TMD fits

Thank you

Strong coupling



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

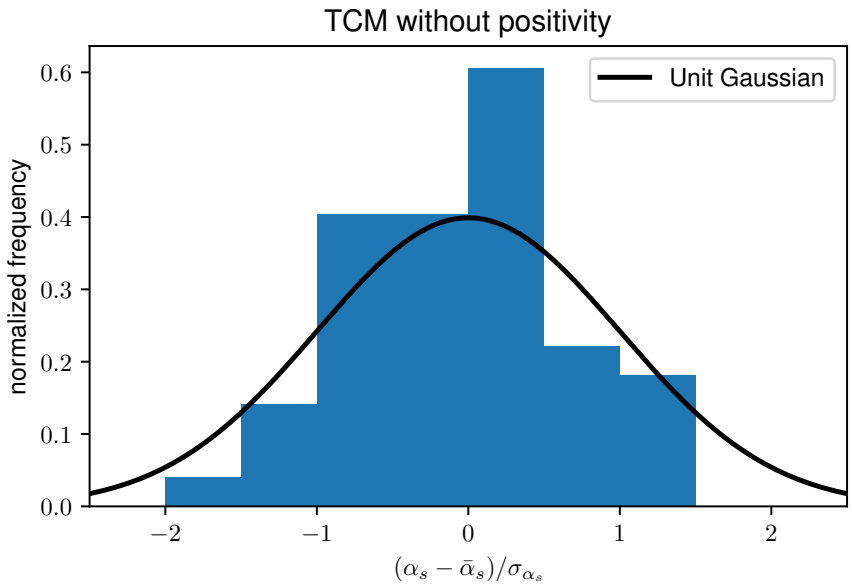
$$\Delta_i^{(k)} = T_i^{\rm 3pt}(\alpha_s^{(k)}) - T_i(\alpha_s^0)$$

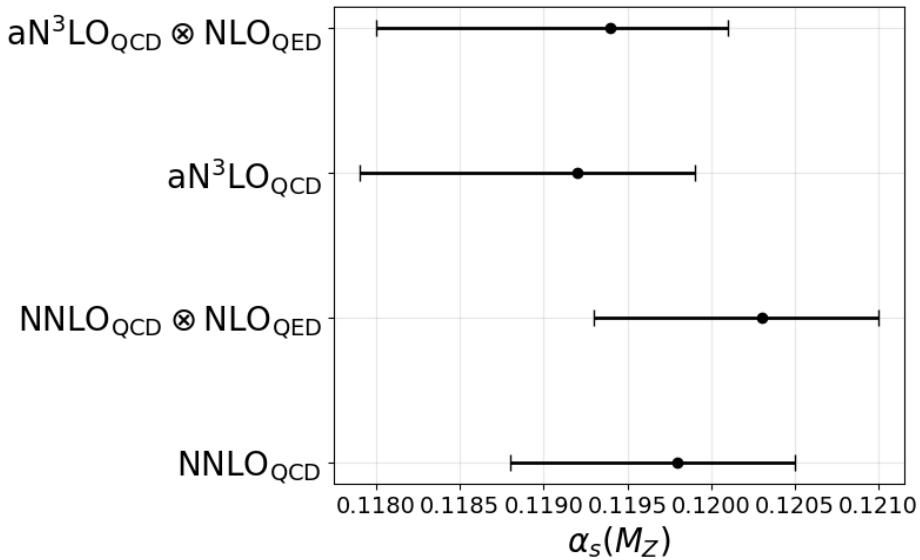
$$\alpha_s^0 = 0.118 \quad \alpha_s^+ = 0.122 \quad \alpha_s^- = 0.114$$

closure test: $\alpha_s = 0.118029 \pm 0.000077$

detect bias from positivity and multiplicative ucnertainties

fit:$\alpha_s = 0.1194^{+0.0007}_{-0.0014}$





[EPJ C85 (2025) 1001]

The photon PDF and QED corrections

Photon PDF à la LuxQED

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

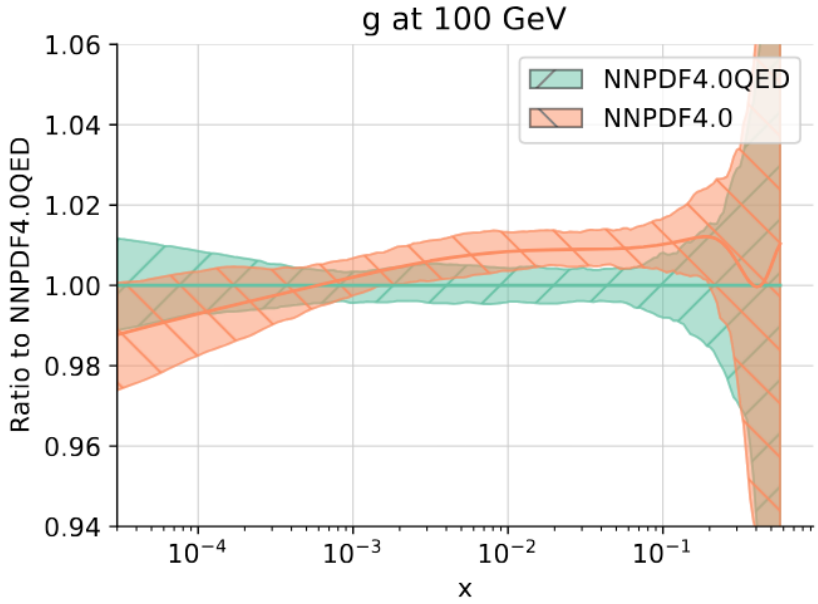
Fit quality unaltered: $\chi^2/N_{\rm dat} = 1.13$

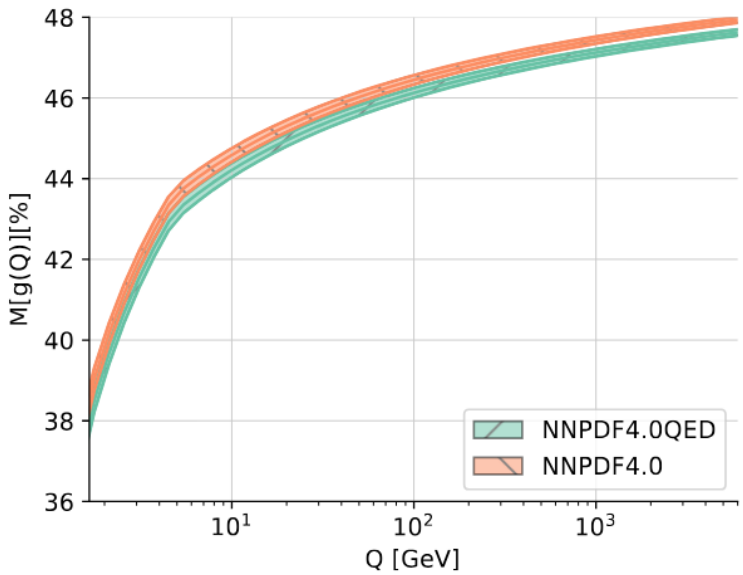
Small (0.5%) momentum shift from g to γ

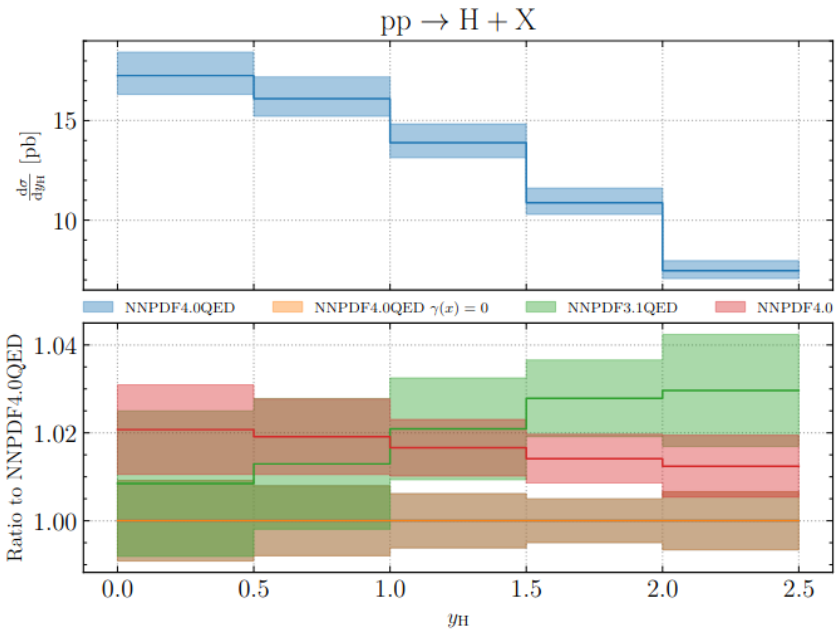
Small (1%) suppression of the gluon PDF

1-2% suppression in ggH cross section

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







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]