Why Open Likelihoods? The PDF viewpoint

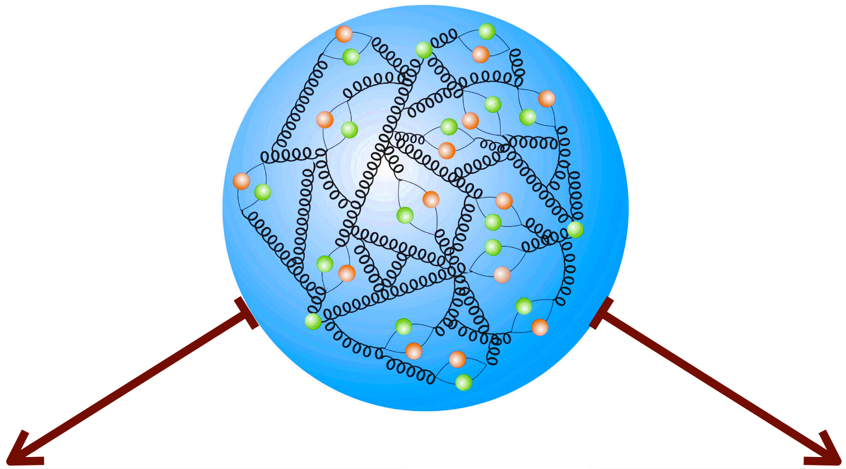
Juan Rojo
VU Amsterdam & Theory group, Nikhef

Publication of Statistical Models - Hands-on Workshop

08/11/2021

Why Parton Distributions?

Parton distributions (PDFs) are ubiquitous in particle, nuclear, and astroparticle physics



Key component of predictions for particle, nuclear, and astro-particle experiments

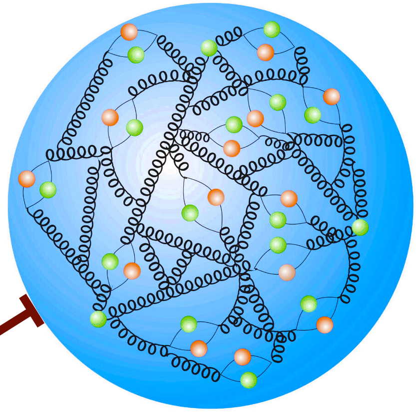
- **pp: ATLAS, CMS, LHCb, ALICE**
- ep: fixed target DIS, HERA
- neutrinos: IceCube, KM3NET, Forward Physics Facility @ LHC
- heavy ions: LHC Pb, LHC O, RHIC
- *pp (future): HL-LHC, FCC, SppS*
- ep (future): **EIC,** LHeC, FCC-eh

Address fundamental questions about Quantum Chromodynamics

- origin of mass & spin
- **heavy quark** & **antimatter content**
- 3D imaging
- gluon-dominated matter
- **nuclear modifications**
- Interplay with BSM e.g. via ``**SMEFT PDFs**”

Why Parton Distributions?

Parton distributions (PDFs) are ubiquitous in particle, nuclear, and astroparticle physics



Key component of predictions for particle, nuclear, and astro-particle experiments

- **pp: ATLAS, CMS, LHCb, ALICE**
- ep: fixed target DIS, HERA
- neutrinos: IceCube, KM3NET, Forward Physics Facility @ LHC
- heavy ions: LHC Pb, LHC O, RHIC
- *pp (future): HL-LHC, FCC, SppS*
- ep (future): **EIC,** LHeC, FCC-eh

LHC master formula

hard cross-section

$$\sigma(M,s) \propto \sum_{ij=u,d,g,\ldots} \int_{M^2}^{s} d\hat{s}\, \mathcal{L}_{ij}(\hat{s},s)\, \widetilde{\sigma}_{ij}(\hat{s},\alpha_s(M))$$

partonic luminosities

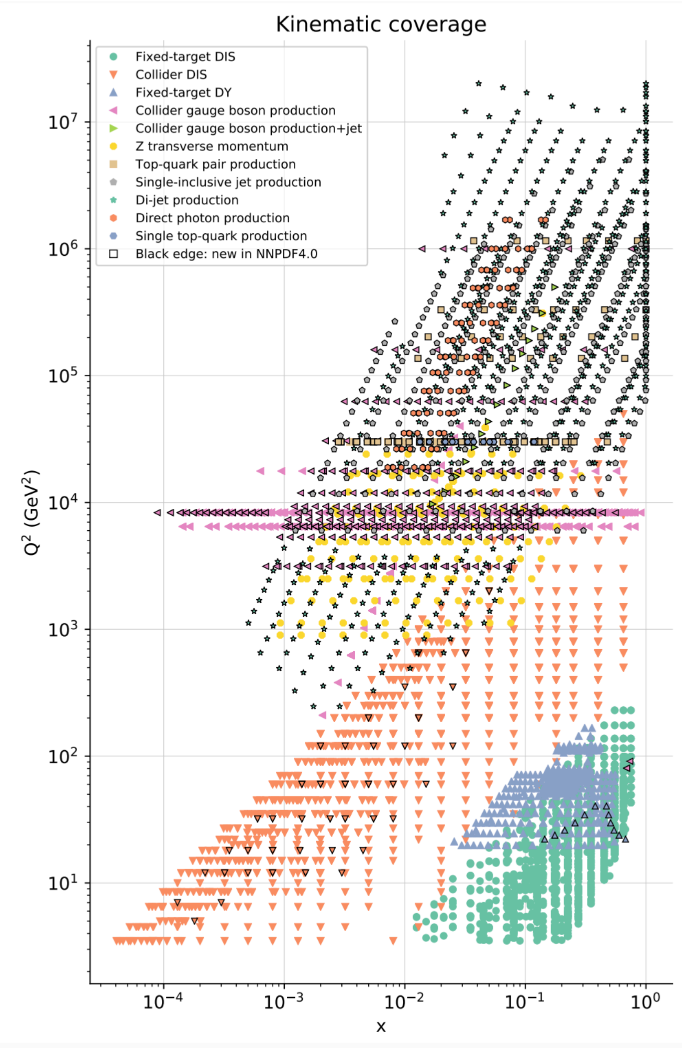
parton distributions

$$\mathcal{L}_{ij}(Q,s) = \frac{1}{s}\int_{Q^2/s}^{1} \frac{dx}{x} f_i\left(\frac{Q^2}{sx},Q\right) f_j\left(x,Q\right)$$

Global PDF analyses

PDFs are determined from a **global analysis** of hard-scattering processes

$$\chi^2(\boldsymbol{\theta}) \equiv \frac{1}{N_{\rm dat}} \sum_{i,j=1}^{N_{\rm dat}} \left(\mathcal{O}_i^{\rm (th)}\left(\boldsymbol{\theta}\right) - \mathcal{O}_i^{\rm (exp)} \right) ({\rm cov}^{-1})_{ij} \left(\mathcal{O}_j^{\rm (th)}\left(\boldsymbol{\theta}\right) - \mathcal{O}_j^{\rm (exp)} \right)$$



$$\mathcal{O}_i^{\rm (th)}\left(\boldsymbol{\theta}, Q\right) = \sum \mathtt{FK}_{ijk}(\alpha_s, Q, Q_0) \cdot f_j(x_k, Q_0, \boldsymbol{\theta})$$

pQCD ***PDF parametrisation***

All available PDF fits based on **Gaussian likelihood optimisation**

- Non-gaussianities in PDF error estimate via Monte Carlo techniques or Lagrange multipliers
- Covariance matrix includes all relevant theory and experimental uncertainties (Gaussian approx)

NNPDF4.0: O(5000) data points from O(100) different datasets

Global PDF analyses

PDFs are determined from a **global analysis** of hard-scattering processes

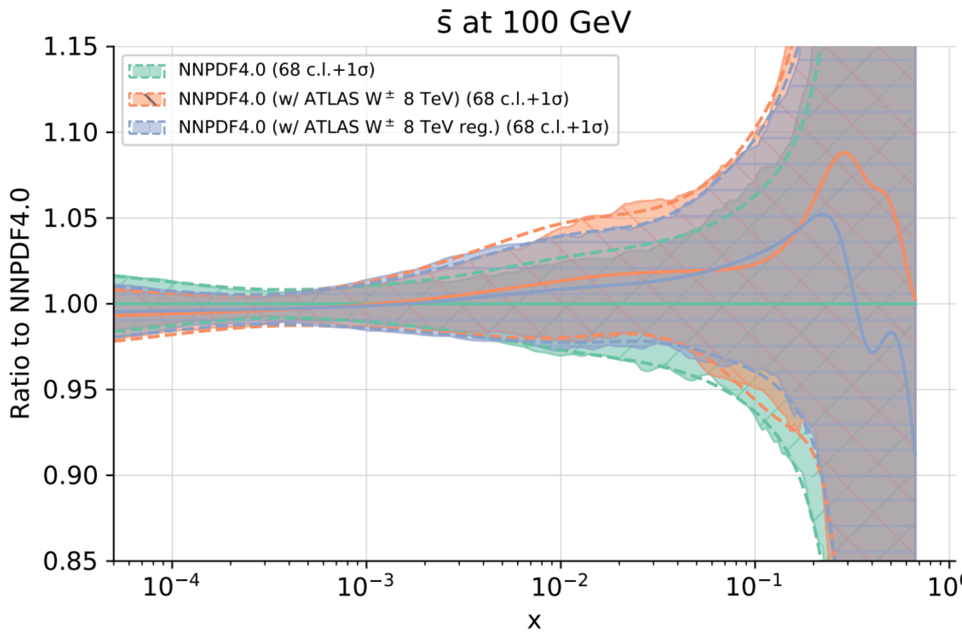
$$\chi^2(\boldsymbol{\theta}) \equiv \frac{1}{N_{\rm dat}} \sum_{i,j=1}^{N_{\rm dat}} \left(\mathcal{O}_i^{\rm (th)}(\boldsymbol{\theta}) - \mathcal{O}_i^{\rm (exp)} \right) ({\rm cov}^{-1})_{ij} \left(\mathcal{O}_j^{\rm (th)}(\boldsymbol{\theta}) - \mathcal{O}_j^{\rm (exp)} \right)$$

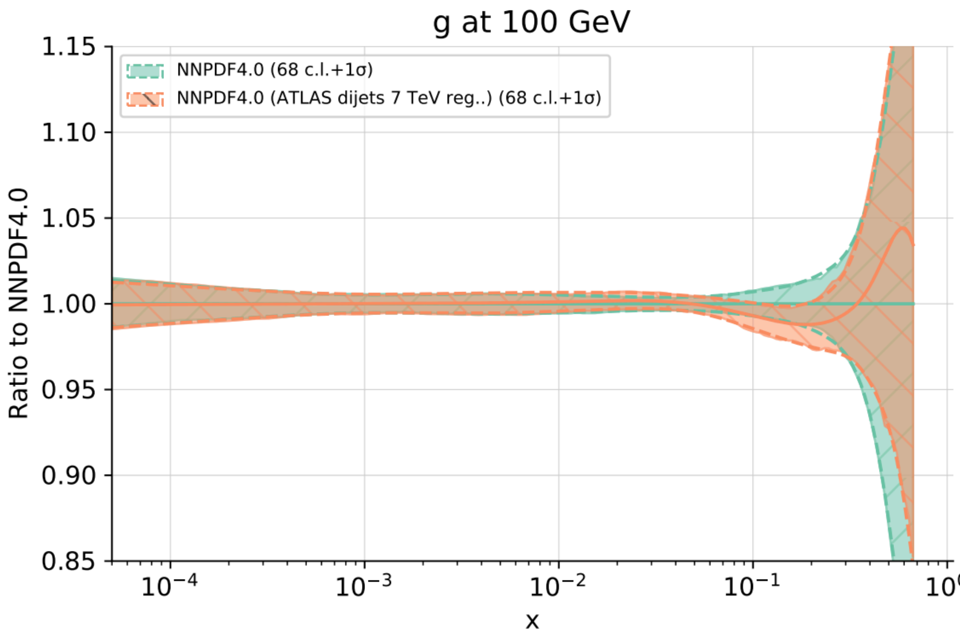
- Adopting a Gaussian likelihood is forced upon us due to **lack of full statistical model available** from the experimental measurements, leading to several issues:
  - Lack of information on correlations
  - Non-positive-definite covariance matrices
  - Lack of breakdown of correlated systematic sources
  - Presence of systematic sources (e.g. modelling) which might not be Gaussian
  - Different naming for systematic sources, which complicates combining processes
  - Data not available from HepData, multiple iterations with conveners necessary
  - Correlations between processes and/or experiments often not available

n.b. we are forced to adopt incorrect assumptions e.g. modelling/theory errors are certainly not Gaussian!

ill-defined covariance matrices

One particularly worrisome consequence of the **lack of open likelihoods** is that results may become very sensitive to **small variations of the official correlation model**





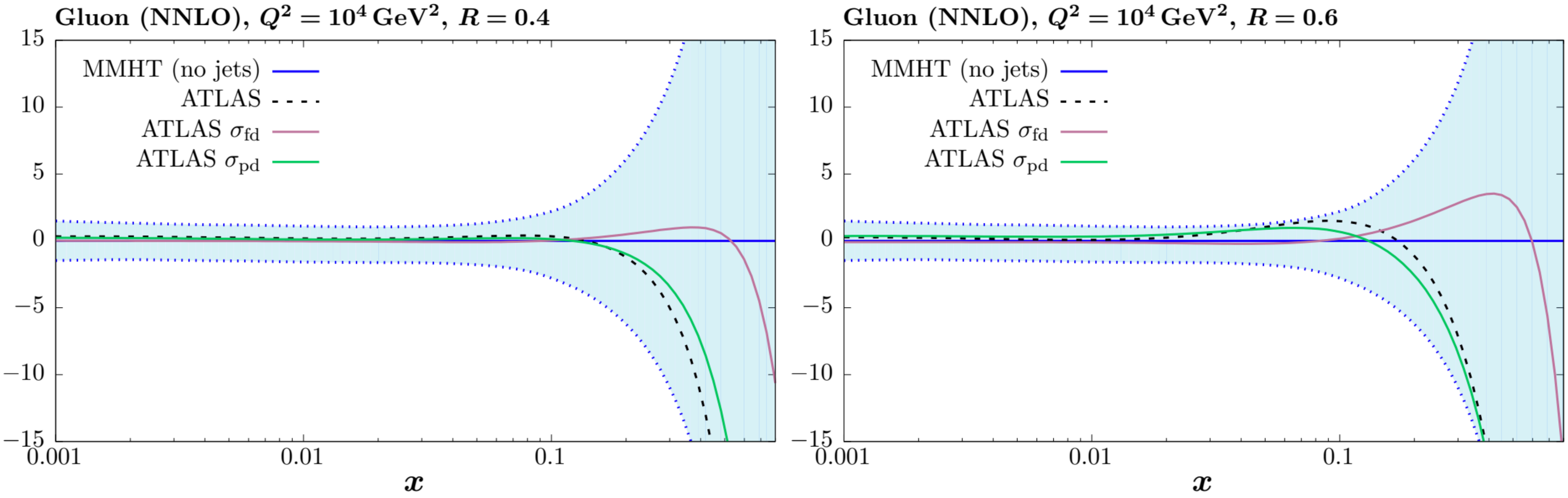
Assess impact in fit by transforming the original covariance matrix into a matrix with the **same eigenvectors** but with **clipped eigenvalues** below some cut-off: **stable PDFs** with much lower **χ^2**

Dataset	$N_{\rm dat}$	$Z_{\rm orig}$	$\chi^2_{\rm orig}$	$\chi^2_{\rm reg}$
ATLAS W, Z 7 TeV CC ($\mathcal{L} = 4.6$ fb^{-1})	46	9.01	1.89	0.93
ATLAS W 8 TeV (*)	22	11.28	3.50	1.15
CMS dijets 7 TeV	54	4.70	1.81	1.73
ATLAS dijets 7 TeV	90	9.93	2.14	0.92
CMS 3D dijets 8 TeV (*)	122	4.47	1.50	0.92

minimal modification of correlation model, large impact in fit quality, PDFs stable

ill-defined covariance matrices

One particularly worrisome consequence of the **lack of open likelihoods** is that results may become very sensitive to **small variations of the official correlation model**



Harland-Lang & Thorne 17

Impact of ATLAS inclusive jet data in MMHT14 varies quite markedly upon the use of different prescriptions for the decorrelation models of systematic uncertainties

These observations are **not** meant to criticise the experimentalists: they emphasise that for **systematic-dominated measurements**, open likelihoods are crucial to fully exploit and understand the information contained in the data

Beyond the χ^2

- Availability of **open likelihoods** would make possible PDF determinations where statistical and systematic uncertainties are always accounted for by means of the appropriate model

$$\mu_i = \bar{\mu}_i \pm \sigma_{\rm poiss,i} \pm \sigma_{\rm syst,gauss,i} \pm \sum \sigma_{\rm syst,gauss,ij} \pm \sum \sigma_{\rm theo,ij} \pm \sum \sigma_{\rm model,ij}$$

central value	stat error (poisson)	syst error (uncorrelated, gaussian)	syst error (correlated, gaussian)	theory error (gaussian? uniform?)	model error (correlated, not gaussian)

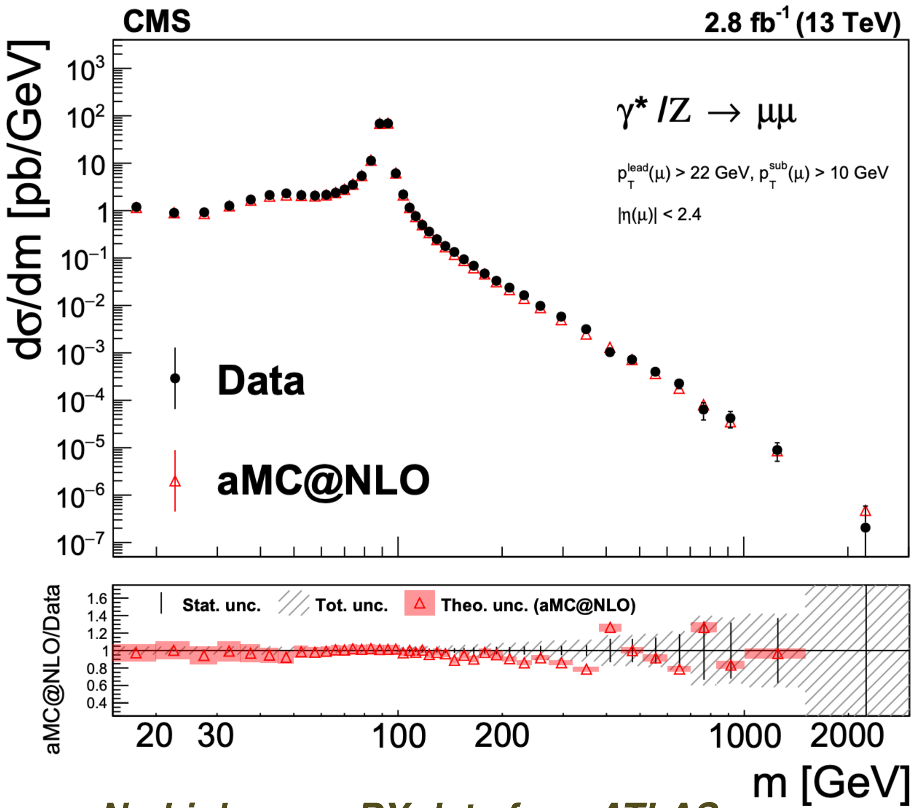
$$\max_{\boldsymbol{\theta}} \mathscr{L}\left(\mu_i, \boldsymbol{\theta}\right) = \mathscr{L}_{\rm gauss} \times \mathscr{L}_{\rm poiss} \times \mathscr{L}_{\rm uniform} \times \mathscr{L}_{\rm theory} \times \ldots$$

- This is not only **technically correct**, but also allows including much more information on PDF fits as compared to what is currently available e.g. by means of **search data**

Beyond the χ^2

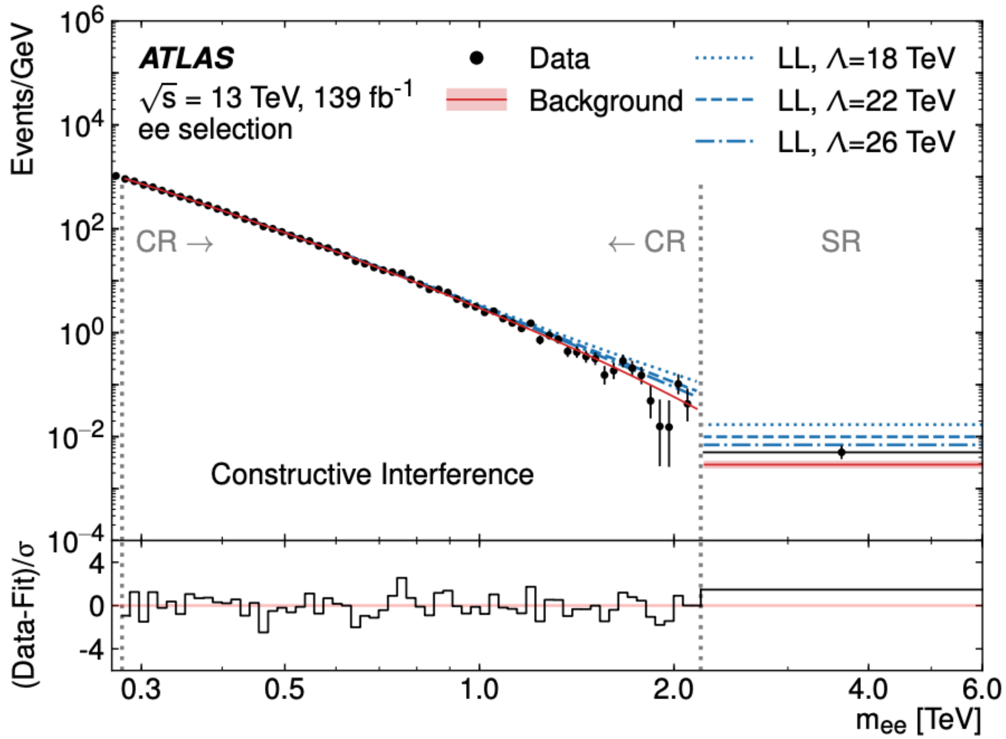
How can we ensure to fully exploit the PDF constraining potential of Run III data?

Most updated SM measurement



No high-mass DY data from ATLAS

One of the recent BSM searches



ATLAS non-resonant dilepton search

SM measurements often lag behind **searches**, but on the longer term their impact is bigger.

Search data with open likelihoods: new handle for PDF determinations!

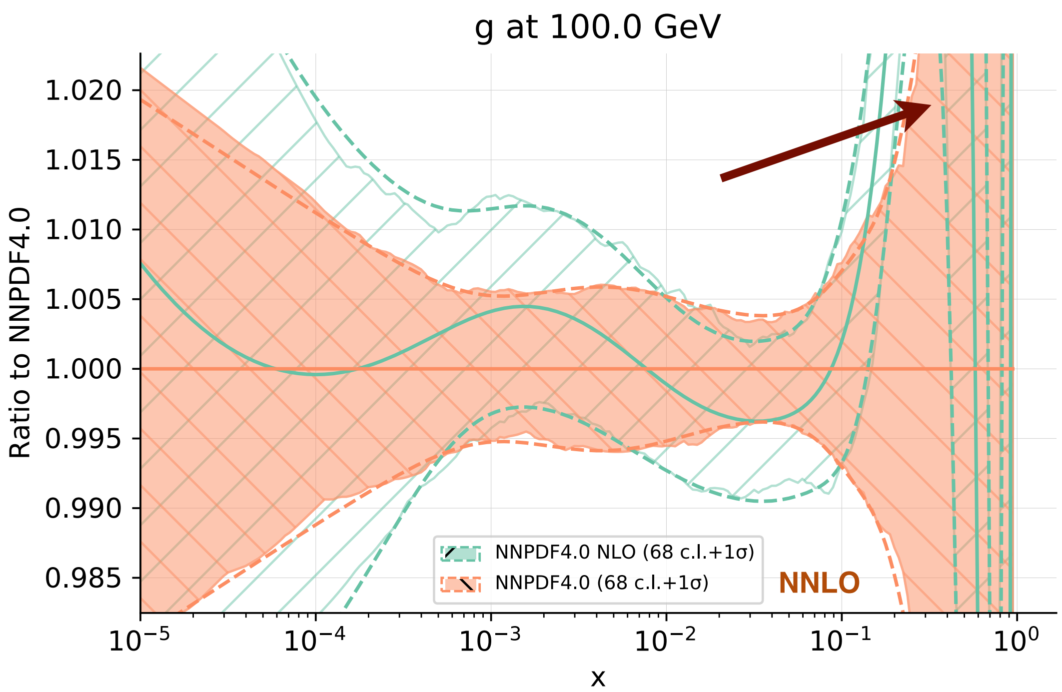
QCD uncertainties in PDF fits

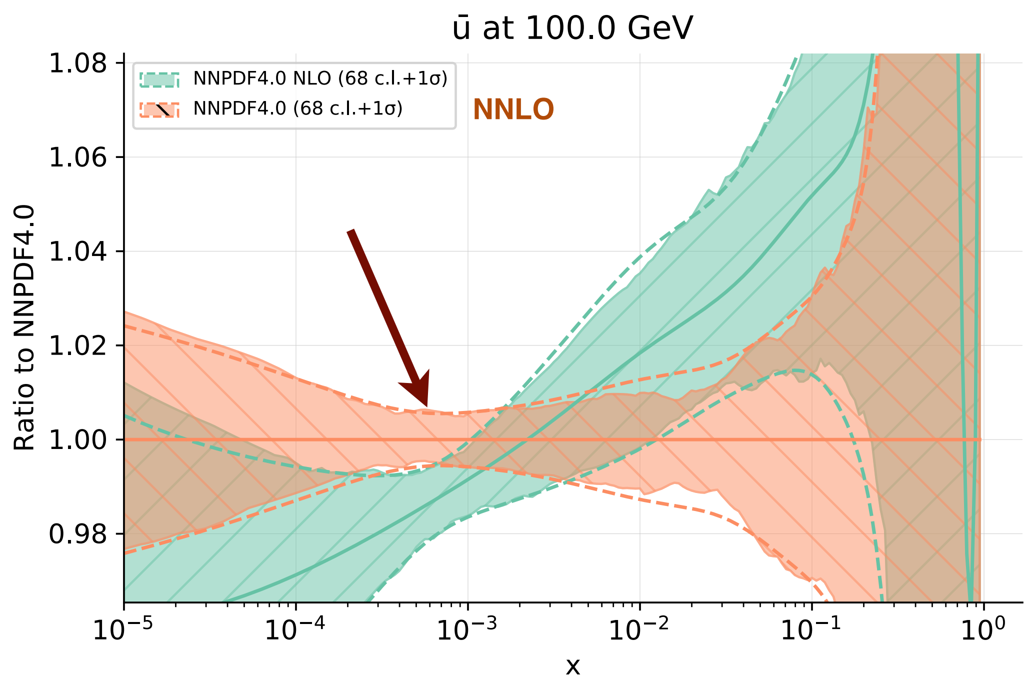
Standard global PDF fits are based on **fixed-order QCD calculations**

$$\sigma = \alpha_s^{p}\sigma_0 + \alpha_s^{p+1}\sigma_1 + \alpha_s^{p+2}\sigma_2 + \mathcal{O}(\alpha_s^{p+3})$$

The truncation of the perturbative series has associated a theoretical uncertainty: **Missing Higher Order (MHO)** uncertainty

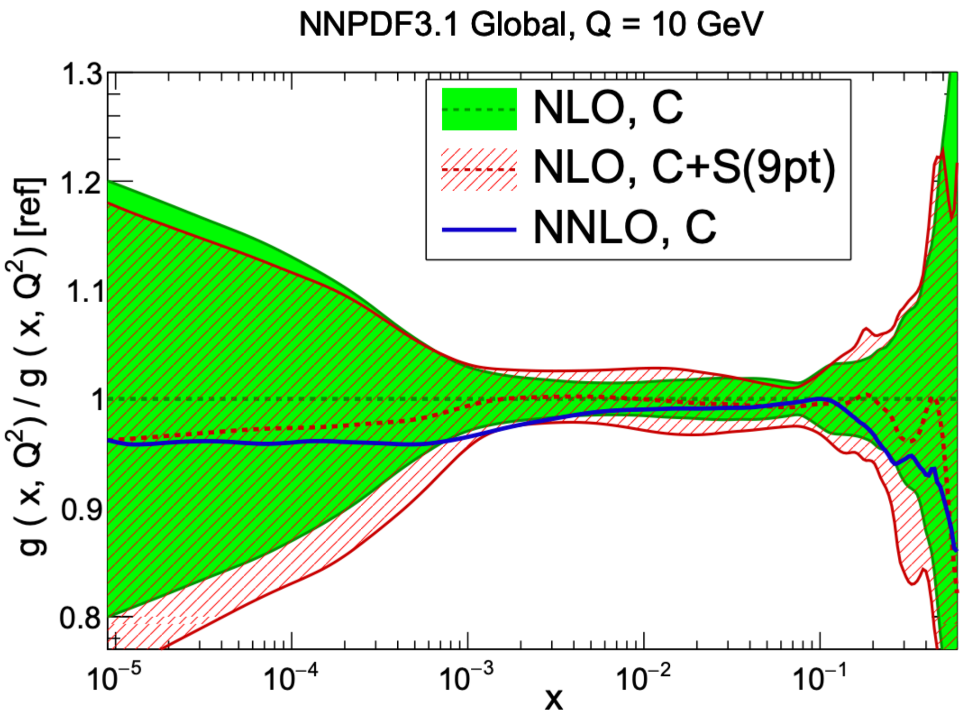
How severe is **ignoring MHOUs** in modern global PDFs fits?

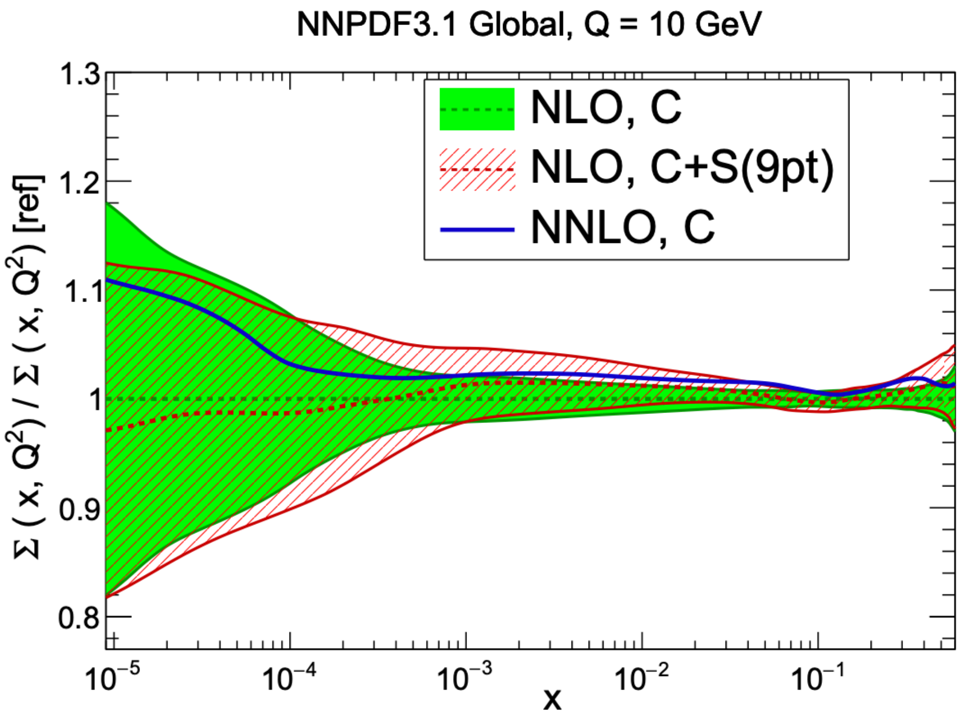




MHOUs are comparable with PDF errors certainly at NLO, likely also at NNLO

QCD uncertainties in PDF fits





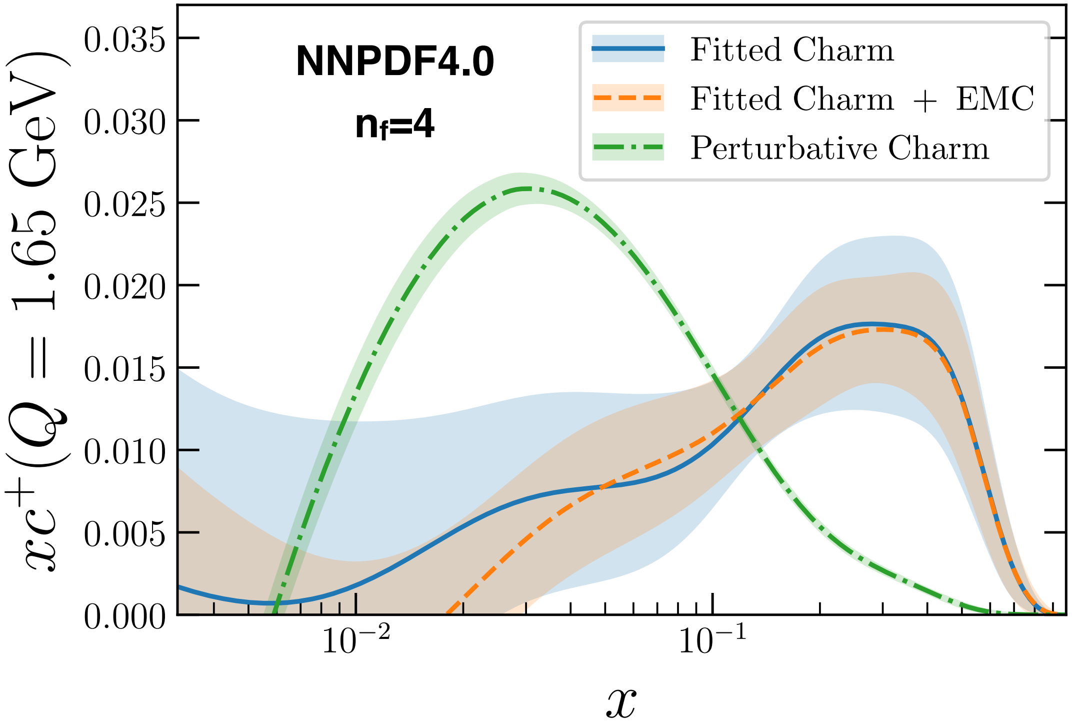
$$\chi^2_{\mathrm{C+S}} = \frac{1}{n_{\mathrm{dat}}} \sum_{ij} \left(T_i - D_i\right) \left(\mathrm{cov}^{\mathrm{exp}} + \mathrm{cov}^{\mathrm{mhou}}\right)^{-1}_{ij} \left(T_j - D_j\right)$$

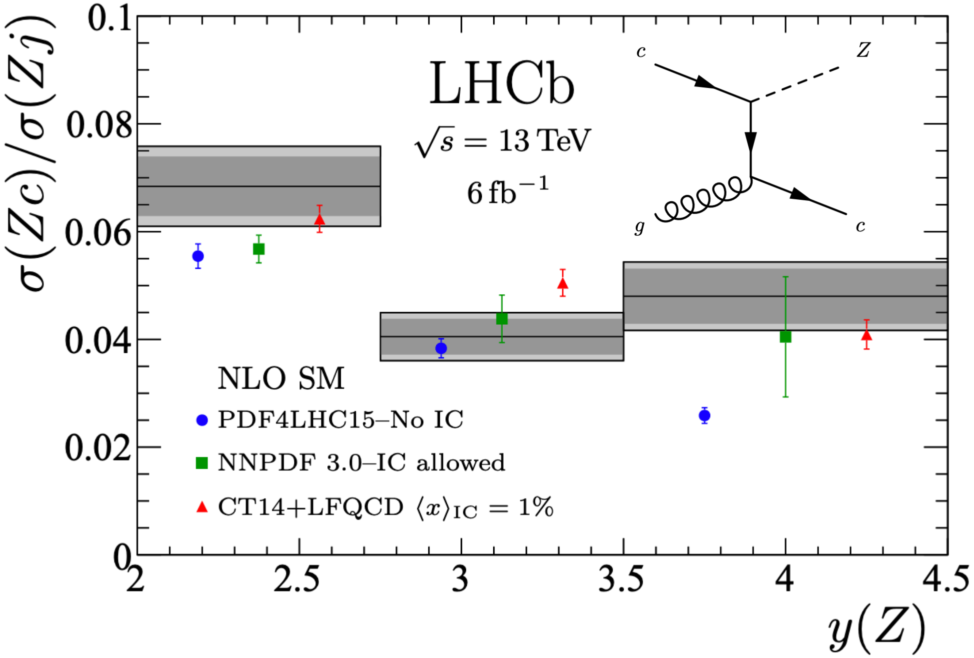
experimental uncertainties — **theory uncertainties from MHOUs**

A NLO global fit with MHOUs highlights how these cannot be neglected, both in terms of **accurac**y and of **precision**

Developing a PDF determination framework based on full likelihoods makes also possible a better treatment of theory errors

Intrinsic Charm





- Increasing evidence for **non-perturbative charm component** within the proton
- Bulk of constraints provided by new **precision LHC data**, complemented by fixed-target DIS
- Consistent with recent LHCb measurement of **forward *Z+D* production**, directly sensitive to the (large-*x*) charm content of the nucleon

however…

Numerical results are provided in Table 3. The statistical uncertainties are uncorrelated between $y(Z)$ intervals, whereas the systematic uncertainties are approximately completely correlated. Since many of the systematic uncertainties are correlated between $y(Z)$ intervals, numerical results are also provided for the ratios of $\mathcal{R}_j^c$ values between pairs of intervals in Table S1.

would benefit from release of full statistical model

Summary and outlook

- The requirements of precision physics at the LHC in general, and of global fits of PDFs in particular, require the careful assessment of the consistency (or lack thereof) between **experimental data and theory calculations**

- Several issues affect modern PDF fits due to restrictions of the available information, from having to drop datasets altogether due to **ill-defined covariance matrices** to implementing ***ad-hoc*** **decorrelation models** and the impossibility to account for **constraints from search data**

PDF interpretations of the HL-LHC data may become seriously hampered, or even impossible altogether, unless **experiments release their full statistical models**