Parton Distribution Functions

2024 CTEQ Summer School on QCD and Electroweak Phenomenology

Emanuele R. Nocera

Università degli Studi di Torino and INFN, Torino

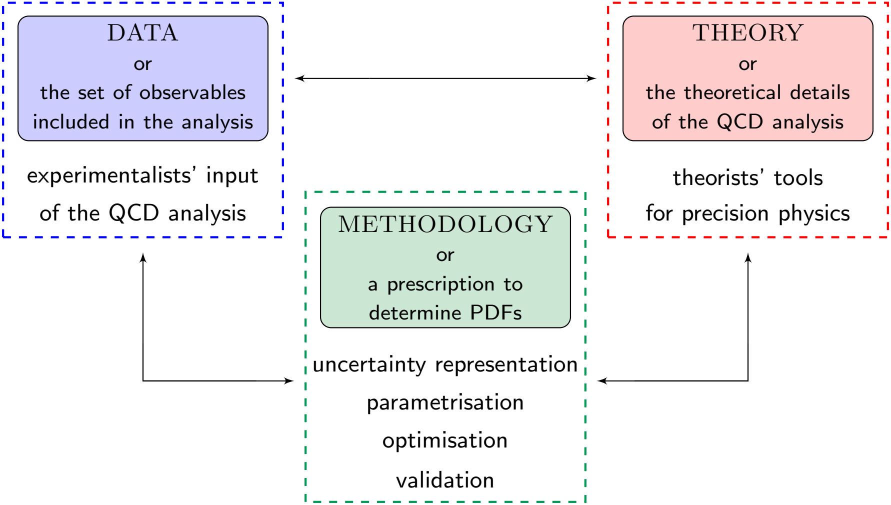
27-28 August 2024

Summary of Lecture 1

1. Parton Distribution functions are a key ingredient of the LHC program
   $\longrightarrow$ PDFs are often the dominant source of uncertainty in theoretical predictions
   $\longrightarrow$ limiting factor for precision and discovery
2. PDFs are related to physical observales via factorisation and evolution
   $\longrightarrow$ qualitative PDF features are driven by this theoretical framework
   $\longrightarrow$ valence peak follows from valence sum rules and kinematic vanishing
   $\longrightarrow$ small-x rise follows from rise of anomalous dimensions
   $\longrightarrow$ correlation of small-x rise and large-x depletion follow from momentum conservation
3. PDFs are determined from experimental data by means of parametric regression
   $\longrightarrow$ need to define data, theory, and methodology
4. Different physical observables constrain different PDF combinations
   $\longrightarrow$ fixed-target NC DIS: u and d
   $\longrightarrow$ fixed-target CC DIS: s and $\bar{s}$
   $\longrightarrow$ HERA NC and CC DIS: u, $\bar{u}$, d, $\bar{d}$, g (scaling violations and tagged DIS)
   $\longrightarrow$ fixed-target DY: u and d at large x
   $\longrightarrow$ collider DY: u, $\bar{u}$, d, $\bar{d}$, s
   $\longrightarrow$ collider DY+c: s (W) and c (Z)
   $\longrightarrow$ Zp_T, $t\bar{t}$, jets: g

Lecture 2: Theoretical and methodological accuracy in PDF determination

The ingredients of PDF determination



Each of these ingredients is a source of uncertainty in the PDF determination
Each of these ingredients require to make choices which lead to different PDF sets

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

Parton Distribution Functions

Lecture 2: Theoretical and Methodological Accuracy in PDF Determination

Outline

2.1 Can we improve the fit quality by improving the theory?
- heavy quarks and intrinsic charm
- missing higher order uncertainties
- electroweak corrections and the photon PDF

2.2 Why is the methodology important?
- parametrisation
- optimisation
- uncertainty representation
- validation of uncertainties
- PDF benchmarks

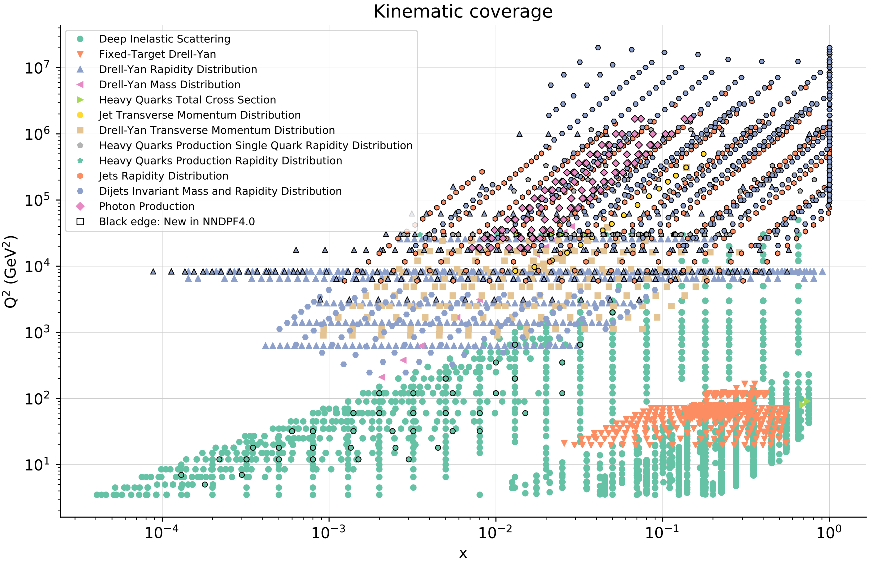
I will focus on a limited selection of recent results

I will not talk about some very interesting topics (*e.g.* aN^3LO PDFs, interplay between fitting PDFs and New Physics, non parametric regression models, ...)

See also lectures by J. Glombitza

2.1 Theory

Can we improve the fit quality by improving the theory?



$N_{\rm dat} = 4618$ $\chi^2/N_{\rm dat} \sim 1.19$ (NNLO) $1\sigma = \sqrt{2/N_{\rm dat}} \sim 0.02$

2.1.1 Heavy Quarks

Heavy Quarks in DIS

Two possible factorisation schemes for DIS structure functions

$\overline{\text{MS}}$ scheme

Heavy quarks are treated as massless (zero-mass scheme)
corrections proportional to $\ln(Q^2/m_h^2)$ are resummed to all orders by DGLAP
corrections that are $\mathcal{O}(m_h^2/Q^2)$ are neglected
This scheme is appropriate when $Q^2 \gg m_h^2$

Decoupling scheme

Heavy quarks are treated as massive (massive scheme)
corrections proportional to $\ln(Q^2/m_h^2)$ are treated at fixed order
corrections that are $\mathcal{O}(m_h^2/Q^2)$ are included
This scheme is appropriate when $Q^2 \sim m_h^2$

The third way: match the two schemes

General-mass variable-flavour number schemes (ACOT, S-ACOT, TR, FONLL, ...)

use $\overline{\text{MS}}$ for $Q^2 \gg m_h^2$ with full mass dependence retained
keep all flavour sin running DGLAP
subtract double counting terms

Intrinsic charm in QCD

What is intrinsic charm?

Do not factor charm mass singularities into operator matrix element

Choose $n_f = 3$ scheme

Charm PDF purely intrinsic, scale-independent

Intrinsic charm is charm in the $n_f = 3$ (decoupling) scheme

$$f_c^{(n_f)} = 0 \quad \rightarrow \quad f_c^{(n_f+1)} \propto \alpha_s \ln \frac{Q^2}{m_c^2} \left(P_{qg} \otimes f_g^{(n_f+1)} \right) + \mathcal{O}\left(\alpha_s^2\right)$$

3FNS charm; 4FNS charm; 4FNS gluon; **NLO matching**

How to measure intrinsic charm?

Determine PDFs from data, go to $n_f = 3$ result, look at the result

1) Parametrise PDFs in $n_f = 3$ (3FNS) and match up for fitting

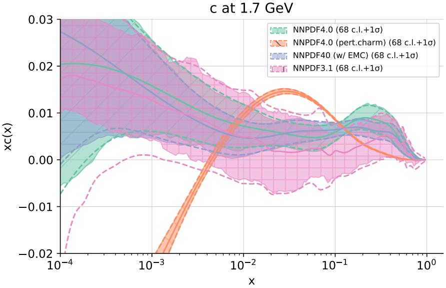
2) Parametrise PDFs in $n_f = 4$ (4FNS) and match down for determining intrinsic charm

$$c^{(n_f=4)}(x,Q) \simeq c^{(n_f=4)}_{\rm (pert)}(x,Q) + c^{(n_f=4)}_{\rm (intr)}(x,Q)$$

Extracted phenomenologically from data; *from pQCD evolution and matching*; *from intrinsic component* $c^{(n_f=3)}_{\rm (intr)}(x) \neq 0$

Large matching uncertainties [I. Bierenbaum *et al.*; J. Ablinger *et al.*]

Perturbative vs Fitted Charm

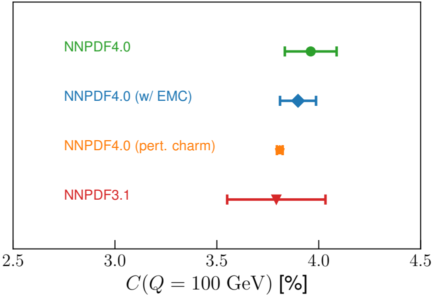


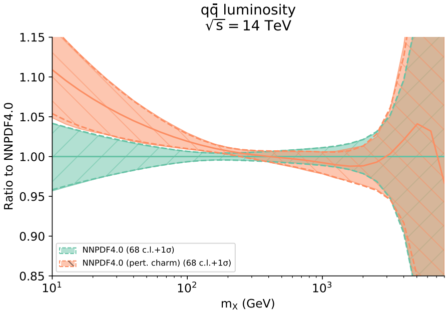
Fitting charm modifies the flavour decomposition and improves the fit

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

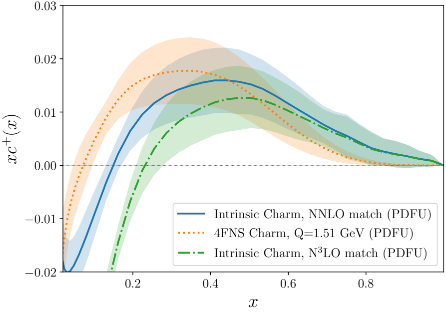
mainly due to a worsening of the LHC W, Z and top pair data sets

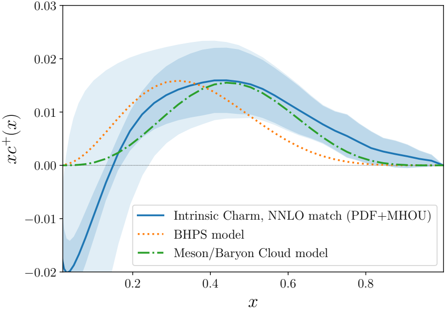
Small charm momentum fraction
$C(Q^2) = \int_0^1 dx x c^+(x, Q^2)$

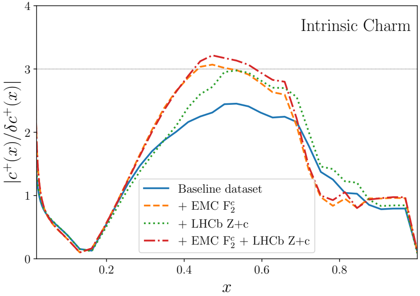




Total intrinsic charm







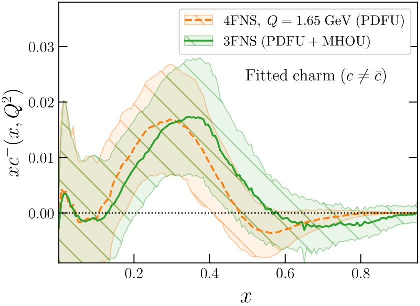
Small but nonzero
valence-like intrinsic charm (3FNS)

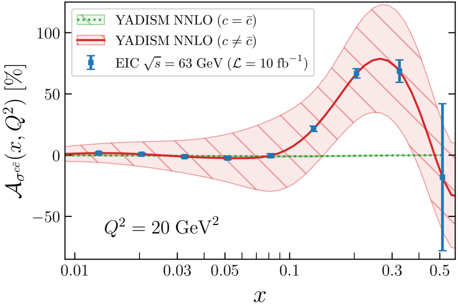
Stable upon inclusion of MHOUs
(estimated as the difference between
NNLO and N^{3}LO matching conditions)

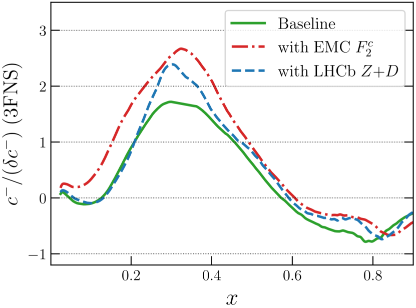
Consistence with model predicitons

2.5σ significance for baseline
3.0σ with LHCb $Z + c$ and/or EMC F_2^c

Intrinsic charm-anticharm asymmetry







Small but nonzero charm-anticharm asymmetry (3FNS)

MHOUs estimated as the difference between NNLO and N^{3}LO matching conditions

1.5σ significance for baseline
2.5σ with LHCb $Z + c$ and/or EMC F_2^c

Can be significantly improved at the EIC

$$\mathcal{A}_{\sigma^{c\bar{c}}}(x,Q^2) \equiv \frac{\sigma^{c}_{\rm red}(x,Q^2)-\sigma^{\bar{c}}_{\rm red}(x,Q^2)}{\sigma^{c\bar{c}}_{\rm red}(x,Q^2)}$$

2.1.2 Missing Higher Order Uncertainties

Perturbative Accuracy in PDF Determination

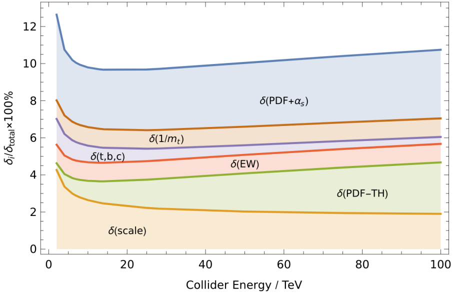
NNLO is the precision frontier for PDF determination

N3LO is the precision frontier for partonic cross sections

Mismatch between perturbative order of partonic cross sections and accuracy of PDFs may become a significant source of uncertainty

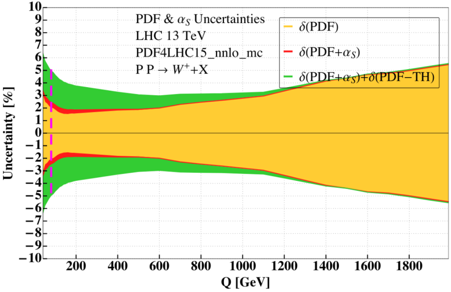
$$\hat{\sigma} = \alpha_s^p \hat{\sigma}_0 + \alpha_s^{p+1}\hat{\sigma}_1 + \alpha_s^{p+2}\hat{\sigma}_2 + \mathcal{O}(\alpha_s^{p+3}) \qquad \delta(\mathrm{PDF-TH}) = \frac{1}{2}\left|\frac{\sigma^{(2)}_{\mathrm{NNLO-PDFs}} - \sigma^{(2)}_{\mathrm{NLO-PDFs}}}{\sigma^{(2)}_{\mathrm{NNLO-PDFs}}}\right|$$

Higgs production in gluon-gluon fusion



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

W^+ boson production in CC Drell-Yan



[JHEP 11 (2020) 143]

MHOUs and Scale Variations

As an example, let us consider the NS DIS structure function

$$F_2^{\rm NS}(N,Q^2) = xC_{\rm NS}(N,\alpha_s(Q^2))\exp\left[\int_{Q_0^2}^{Q^2}\frac{d\lambda^2}{\lambda}\gamma_{\rm NS}\left(N,\alpha_s(\mu^2)\right)\right]f_{\rm NS}(Q_0^2)$$

Sources of MHOUs

$$\gamma_{\rm NS}^{\rm N^kLO}(N,\alpha_s) = \alpha_s\gamma_{\rm NS}^{(0)} + \alpha_s^2\gamma_{\rm NS}^{(1)} + \dots \alpha_s^{k+1}\gamma_{\rm NS}^{(k)}$$

$$C_{\rm NS}^{\rm N^kLO}(N,\alpha_s) = 1 + \alpha_s C_{\rm NS}^{(1)} + \dots \alpha_s^k C_{\rm NS}^{(k)}$$

Scale variations

Idea: $\alpha_s(\kappa^2\mu^2) = \alpha_s(\mu^2)[1+\mathcal{O}(\alpha_s)]$

at N$^{\rm k}$LO differences due to higher orders are related to the QCD β function up to β_k

$\bar{C}_{\rm NS}(\alpha_s(\kappa_r^2\mu^2,\kappa_r^2)) = C_{\rm NS}(\alpha_s(\mu^2))[1+\mathcal{O}(\alpha_s)]$ fixes $\bar{C}^{(k)}$ in terms of $C^{(k)}$

$\bar{\gamma}_{\rm NS}(\alpha_s(\kappa_f^2\mu^2,\kappa_f^2)) = \gamma_{\rm NS}(\alpha_s(\mu^2))[1+\mathcal{O}(\alpha_s)]$ fixes $\bar{\gamma}^{(k)}$ in terms of $\gamma^{(k)}$

$$\Delta C_{\rm NS} = \bar{C}_{\rm NS}(\alpha_s(\kappa_r^2\mu^2,\kappa_r^2)) - C_{\rm NS}(\alpha_s(\mu^2))$$

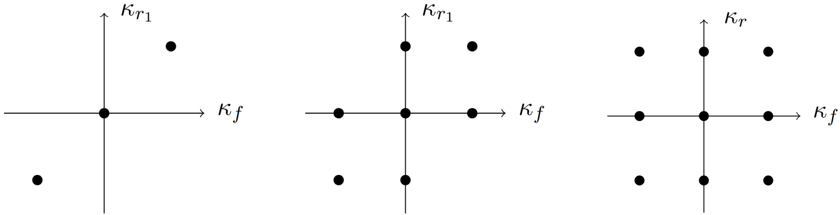
renormalisation scale (at which UV divergences are subtracted) $\mu_r = \kappa_r\mu$

$$\Delta\gamma_{\rm NS} = \bar{\gamma}_{\rm NS}(\alpha_s(\kappa_f^2\mu^2,\kappa_f^2)) - \gamma_{\rm NS}(\alpha_s(\mu^2))$$

factorisation scale (at which collinear divergences are factorised) $\mu_f = \kappa_f\mu$

Propagate ΔC and $\Delta\gamma$ into Δf

Scale Variations: Prescriptions



Vary μ_r and μ_f about μ_0

Pick a set of possible variations

3-points: $\mu_r = \mu_f$, $\kappa_{r,f} = 2, 1/2$

7-points: μ_r, μ_f varied independently, $\kappa_{r,f} = 2, 1/2$, remove $\mu_r/\mu_f = 4$

9-points: μ_r, μ_f varied independently, $\kappa_{r,f} = 2, 1/2$

To estimate MHOUs, take the envelope,
i.e. the difference between the largest and smallest predictions

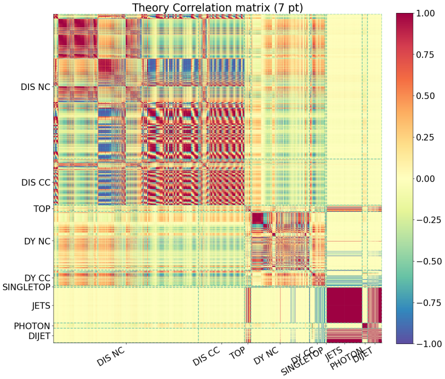
A Theory Covariance Matrix

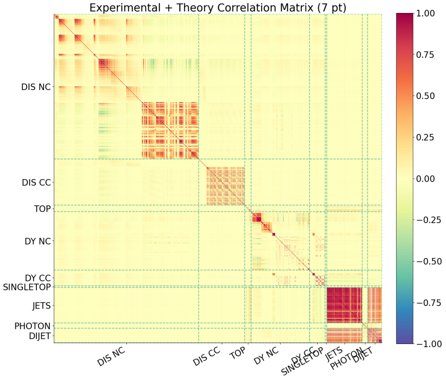
Assuming that theory uncertainties are (a) Gaussian and (b) independent from experimental uncertainties, modify the figure of merit to account for theory errors

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

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

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





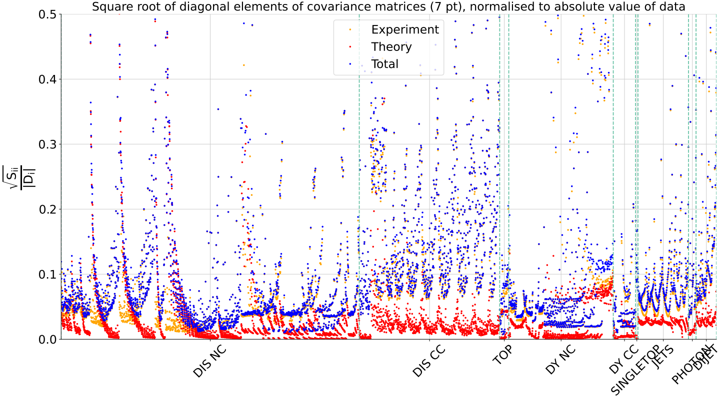
A Theory Covariance Matrix

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

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



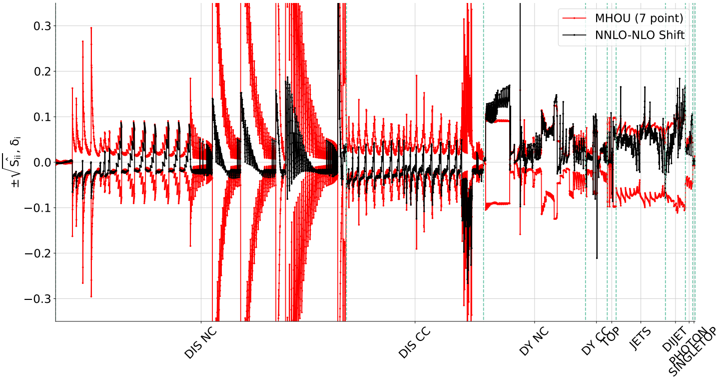
A Theory Covariance Matrix

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

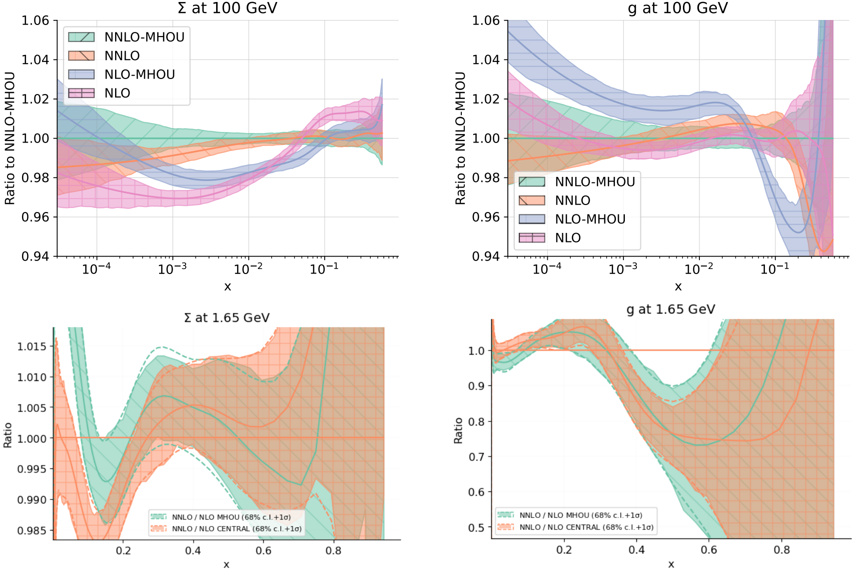
$$\chi^2 = \sum_{i,j}^{N_{\rm dat}} (D_i - T_i)(\mathrm{cov}_{\rm exp} + \mathrm{cov}_{\rm th})^{-1}_{ij}(D_j - T_j);\ (\mathrm{cov}_{\rm th})_{ij} = \frac{1}{N}\sum_k^N \Delta_i^{(k)}\Delta_j^{(k)};\ \Delta_i^{(k)} \equiv T_i^{(k)} - T_i$$

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

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



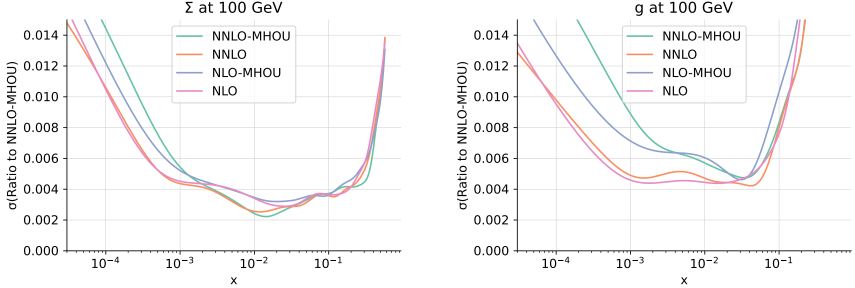
Impact on Parton Distributions



Faster perturbative convergence when MHOU are incorporated into PDFs

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

Impact on Uncertainties and Fit Quality



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

Overall (rather small) variation of uncertainties. Tensions relieved: improvement in χ^2

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

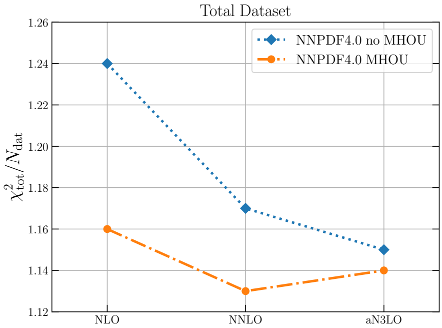
What Happens at aN^3LO?

Dataset	$N_{\rm dat}$	NLO no MHOU	NLO MHOU	$N_{\rm dat}$	NNLO no MHOU	NNLO MHOU	$N_{\rm dat}$	aN^3LO no MHOU	aN^3LO 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

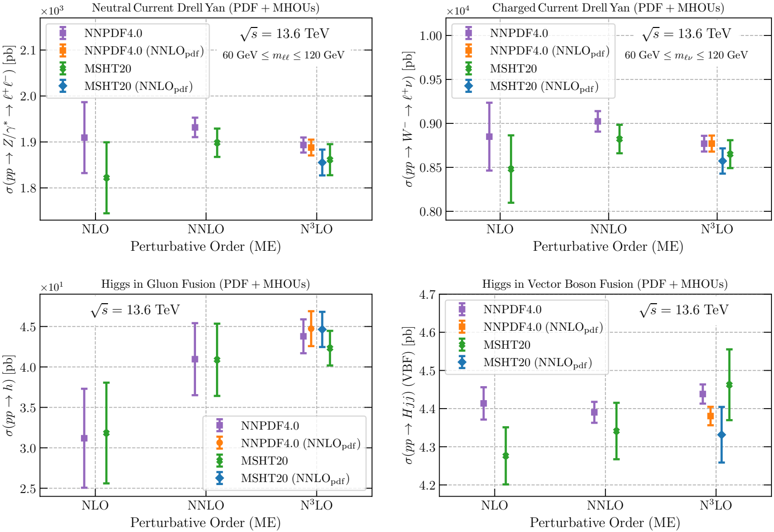
Fit quality improves with perturbative order

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]

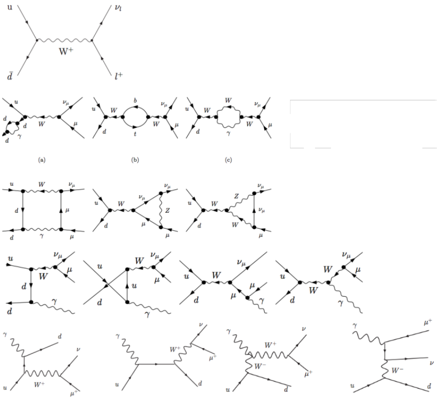
2.1.3 The photon PDF

Beyond QCD

So far we have considered only the expansion in α_s

But there exist QED and electroweak corrections to partonic cross sections

Because $\alpha(M_Z) \sim \alpha_s(M_Z)/10$ we expect NLO EW corrns $\sim$ NNLO QCD corrns



Virtual EW corrections

Real EW corrections - quark initiated

Real EW corrections - photon initiated

[Slide by courtesy of M. Ubiali]

Let us restrict ourselves to QED corrections

How should we incorporate QED in our framework?

Define a photon PDF and include it in DGLAP

$$Q^2\frac{\partial}{\partial Q^2}g(x,Q^2)=\sum_{q,\bar{q},g}P_{ga}(x,\alpha_s(Q^2))\otimes f_a(x,Q^2)+P_{g\gamma}(x,\alpha_s(Q^2))\otimes\gamma(x,Q^2),$$

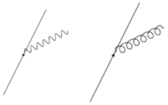
$$Q^2\frac{\partial}{\partial Q^2}q(x,Q^2)=\sum_{q,\bar{q},g}P_{qa}(x,\alpha_s(Q^2))\otimes f_a(x,Q^2)+P_{q\gamma}(x,\alpha_s(Q^2))\otimes\gamma(x,Q^2),$$

$$Q^2\frac{\partial}{\partial Q^2}\gamma(x,Q^2)=P_{\gamma\gamma}\otimes\gamma(x,Q^2)+\sum_{q,\bar{q},g}P_{\gamma a}(x,\alpha_s(Q^2))\otimes f_a(x,Q^2).$$

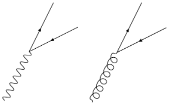
Expand the solution in α_s, α and $\alpha_s\alpha$

$$P_{ij}=\sum_{m,n}\left(\frac{\alpha_S}{2\pi}\right)^m\left(\frac{\alpha}{2\pi}\right)^n P_{ij}^{(m,n)}$$

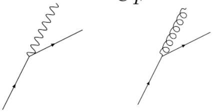
$$P_{qq}^{(0,1)}=\frac{e_q^2}{C_F}P_{qq}^{(1,0)}$$



$$P_{q\gamma}^{(0,1)}=\frac{e_q^2}{T_R}P_{qg}^{(1,0)}$$



$$P_{\gamma q}^{(0,1)}=\frac{e_q^2}{C_F}P_{gq}^{(1,0)}$$



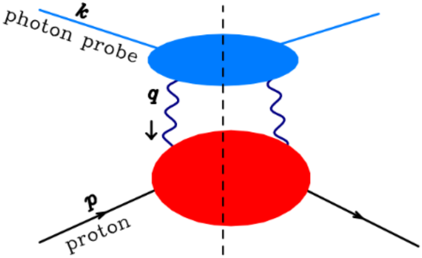
Determine the photon PDF (from data?)

The LuxQED photon PDF

LUXQED [PRL 117 (2016) 242002]

View the $ep \to e + X$ process as an electron scattering off the photon field of the proton

Consider a BSM process, *e.g.* production of a heavy supersymmetric lepton L in ep collision, write the cross section in terms of structure functions and of f_γ, and equate the two to obtain f_γ



$$\sigma = c_0 \sum_a \int_x^1 \frac{dz}{z}\, \hat{\sigma}_a(z,\mu^2)\, \frac{M^2}{zs} f_{a/p}\left(\frac{M^2}{zs},\mu^2\right)$$

$$\sigma = \frac{c_0}{2\pi} \int_x^{1-\frac{2xm_p}{M}} \frac{dz}{z} \int_{Q^2_{\rm min}}^{Q^2_{\rm max}} \frac{dQ^2}{Q^2} \alpha^2_{\rm ph}(-Q^2) \Bigg[\bigg(2-2z+z^2 + \frac{2x^2m_p^2}{Q^2} + \frac{z^2Q^2}{M^2} - \frac{2zQ^2}{M^2} - \frac{2x^2Q^2m_p^2}{M^4}\bigg) F_2(x/z,Q^2) + \left(-z^2 - \frac{z^2Q^2}{2M^2} + \frac{z^2Q^4}{2M^4}\right) F_L(x/z,Q^2)\Bigg]$$

$$x f_{\gamma/p}(x,\mu^2) = \frac{1}{2\pi\alpha(\mu^2)} \int_x^1 \frac{dz}{z} \Bigg\{ \int_{\frac{x^2m_p^2}{1-z}}^{\frac{\mu^2}{1-z}} \frac{dQ^2}{Q^2}\alpha^2(Q^2) \left[\left(z p_{\gamma q}(z) + \frac{2x^2m_p^2}{Q^2}\right) F_2(x/z,Q^2) - z^2 F_L\left(\frac{x}{z},Q^2\right)\right] - \alpha^2(\mu^2) z^2 F_2\left(\frac{x}{z},\mu^2\right) \Bigg\}$$

Iterate a QCD fit including f_γ in DGLAP and in the momentum sum rule

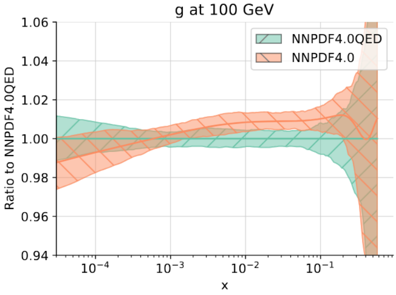
Implications for PDFs and LHC processes

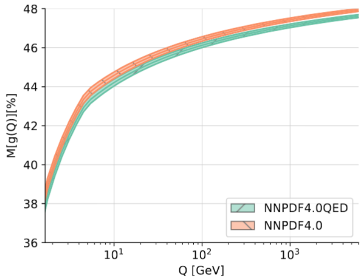
Fit quality unaltered: $\chi^2/N_{\rm dat} = 0.17$

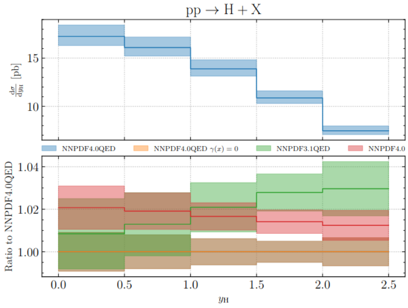
Small (0.5%) momentum shift from g to γ

Small (1%) suppression of the gluon PDF

1-2% suppression in ggH cross section

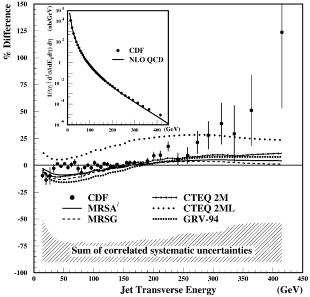


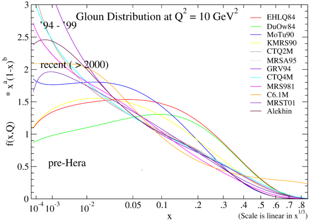


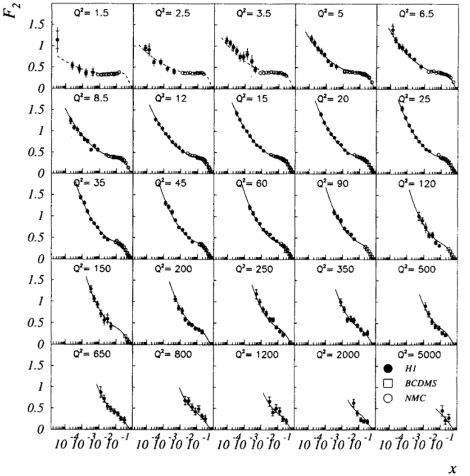


2.2 Methodology

Why is the methodology important?



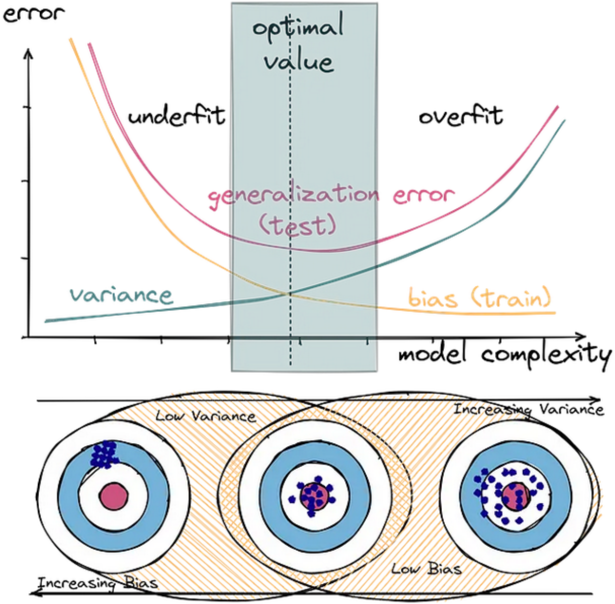




circa 1995: small-x rise of HERA F_2^p and CDF jet discrepancy

The methodology is crucial if we aim at percent-level accurate PDFs

Accuracy vs precision or bias vs variance



What are the ingredients of a fitting methodology?

parametrisation

polynomials/neural network(s)
is there a bias due to the parametrisation?

optimisation

(adaptive) gradient descent
is the parameter space explored efficiently?

uncertainty representation

Hessian/bootstrap of experimental uncertainties
what is the statistical meaning of uncertainties?

validation

closure tests (what happens if I know in advance the underlying law that I am fitting?)
are interpolation and extrapolation uncertainties statistically faithful?

benchmark

PDF4LHC working group
are PDFs obtained independently by various groups equivalent?

2.2.1 Parametrisation

Parametrisation: general features

Problem projected onto the finite-dimensional space of parameters

Choose a parametrisation at an initial scale Q_0^2
for each independent parton i (or a combination of them)

$$xf_i(z, Q_0^2) = A_i\, x^{a_i}\, (1-x)^{b_i}\, \mathscr{F}_i(x, \{c_{f_i}\})$$

small x

$$xf_i(x, Q_0^2) \xrightarrow{x\to 0} x^{a_i}$$

$\mathscr{F}_i(x, \{c_{f_i}\}) \xrightarrow[x\to 1]{x\to 0}$ finite

$\longrightarrow$ interpolation in between

large x

$$xf_i(x, Q_0^2) \xrightarrow{x\to 1} (1-x)^{b_i}$$

The problem is reduced to the determination of the finite set of parameters $\{c_{f_i}\}$

The interpolating function $\mathscr{F}_i(x, \{c_{f_i}\})$ should be sufficiently

GENERAL (the range of PDF behaviours in the space of functions should not be limited)

SMOOTH (PDFs are implicitly assumed ot be smooth functions)

FLEXIBLE (it should be able to adapt to a variety of data and processes)

to describe the data with minimal bias

Parametrisation: two alternative choices

1. Polynomial (Bernstein, Chebyschev) parametrisation, *e.g.*

$$\mathscr{F}_i = 1 + \sum_{i=1}^{n} a_i T_i^{\mathrm{Ch}}\left(y(x)\right) \qquad y = 1 - 2\sqrt{x}$$

in terms of a (relatively) small set of parameters ($\mathcal{O}(30)$ per PDF set)

$$\{\mathbf{a}\} = \{a_i, b_i, \gamma_i, \delta_i\}$$

$\Rightarrow$ **smooth behavior (a desirable feature for a PDF)**
$\Rightarrow$ **potential source of bias if the parametrisation is too rigid**

2. Redundant parametrisation, *e.g.*

a neural network

in terms of a huge set of parameters ($\mathcal{O}(200)$ per PDF set)

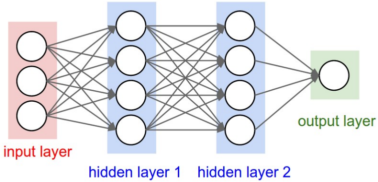
$$\{\mathbf{a}\} = \{\omega_{ij}^{(L-1),f_i}, \theta_i^{(L),f_i}\}$$

$\Rightarrow$ **potentially non-smooth**
$\Rightarrow$ **bias due to the parametrisation reduced as much as possible**

Parametrisation: what a neural network exactly is?

A convenient functional form providing a flexible parametrization
used as a generator of random functions in the PDF space

EXAMPLE: MULTY-LAYER FEED-FORWARD PERCEPTRON



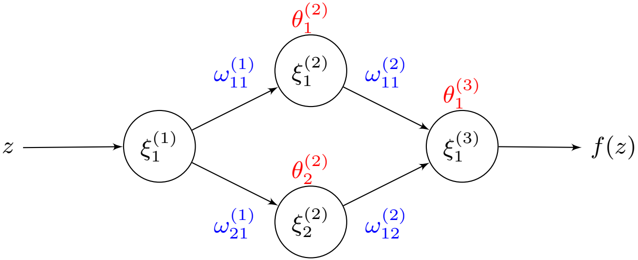
$$\xi_i^{(l)} = g\left(\sum_j^{n_l-1} \omega_{ij}^{(l-1)} \xi_j^{(l-1)} - \theta_i^{(l)}\right)$$

$$g(y) = \frac{1}{1+e^{-y}}$$

- made of neurons grouped into layers (define the architecture)
- each neuron receives input from neurons in the preceding layer (feed-forward NN)
- activation $\xi_i^{(l)}$ determined by a set of parameters (weights and thresholds)
- activation determined according to a non-linear function (except the last layer)

Parametrisation: what a neural network exactly is?

EXAMPLE: THE SIMPLEST 1-2-1 MULTI-LAYER FEED-FORWARD PERCEPTRON

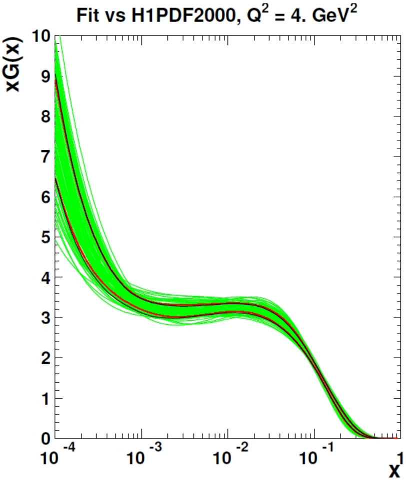


$$f(z) \equiv \xi_1^{(3)} = \left\{ 1 + \exp\left[\theta_1^{(3)} - \frac{\omega_{11}^{(2)}}{1 + e^{\theta_1^{(2)} - x\omega_{11}^{(1)}}} - \frac{\omega_{12}^{(2)}}{1 + e^{\theta_2^{(2)} - x\omega_{21}^{(1)}}} \right] \right\}^{-1}$$

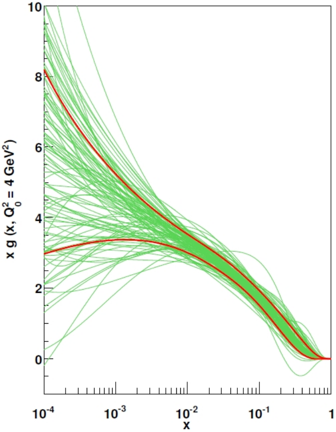
Recall: $\xi_i^{(l)} = g\left(\sum_j^{n_l - 1} \omega_{ij}^{(l-1)} \xi_j^{(l-1)} - \theta_i^{(l)} \right)$; $g(z) = \frac{1}{1 + e^{-z}}$

Parametrisation: standard *vs* redundant

HERA-LHC 2009 PDF benchmark



simple parametrization



redundant parametrization (NN)

2.2.2 Optimisation

Fit quality

1. Define the fit quality (the χ^2 function)

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

with the experimental covariance matrix

$$(\mathrm{cov})_{ij} = \delta_{ij} s_i^2 + \left(\sum_{\alpha}^{N_c} \sigma_{i,\alpha}^{(c)} \sigma_{j,\alpha}^{(c)} + \sum_{\alpha}^{N_{\mathcal{L}}} \sigma_{i,\alpha}^{(\mathcal{L})} \sigma_{j,\alpha}^{(\mathcal{L})} \right) D_i D_j$$

s_i are N_{dat} uncorrelated uncertainties (statistic + uncorrealted systematic ucnertainties)
$\sigma_{i,\alpha}^{(c)}$ are $N_{\mathrm{dat}} \times N_c$ additive correlated uncertainties
$\sigma_{i,\alpha}^{(\mathcal{L})}$ are $N_{\mathrm{dat}} \times N_{\mathcal{L}}$ multiplicative uncertainties

2. Find the best-fit configuration of parameters $\{\mathbf{a_0}\}$ which minimise the χ^2

3. Treat conveniently

   - uncorrelated/correlated uncertainties
     need not to overestimate uncertainties and to let the χ^2 be faithful
   - additive/multiplicative uncertainties
     need to avoid the D'Agostini bias

Parameter optimisation: general framework

Optimisation usually performed by means of simple gradient descent:
compute and minimise the gradient of the fit quality with respect to the fit parameters

$$\frac{\partial \chi^2}{\partial a_i}\,, \qquad \text{for } i = 1, \ldots, N_{\rm par}$$

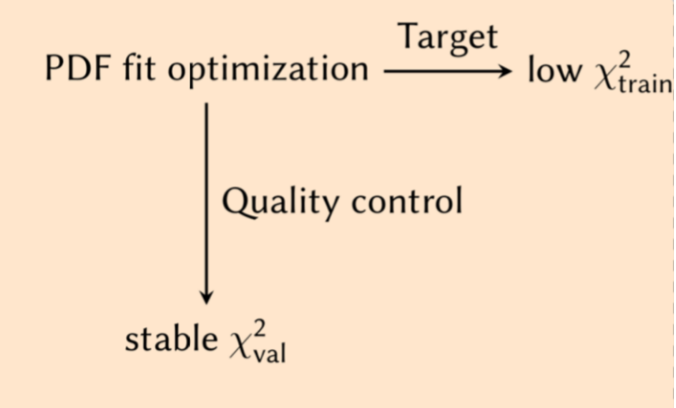
Optimisation should minimise the noise in the χ^2 driven by noisy experimental data

Additional complications in case of a redundant parametrisation (huge parameter space)

1. need to explore the parameter space as uniformly as possible
(in order to avoid stopping the fit in a local minimum)
2. need for a computationally efficient minimisation
(non-trivial relationship between FFs and observables via convolution)
3. need to define a criterion for minimisation stopping
(avoid learning statistical fluctuations of the data)

Alternative algorithms:
genetic algorithms, adaptive algorithms, ...

Optimisation: training and validation



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

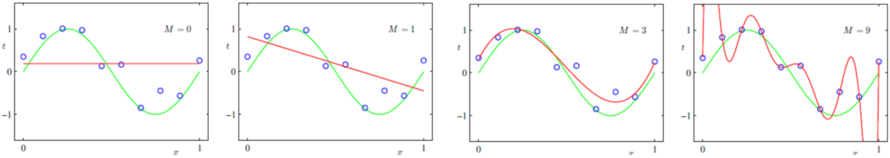
$$(\mathrm{cov})_{ij} = \delta_{ij} s_i^2 + \left(\sum_{\alpha}^{N_c} \sigma_{i,\alpha}^{(c)} \sigma_{j,\alpha}^{(c)} + \sum_{\alpha}^{N_{\mathcal{L}}} \sigma_{i,\alpha}^{(\mathcal{L})} \sigma_{j,\alpha}^{(\mathcal{L})} \right) D_i D_j$$

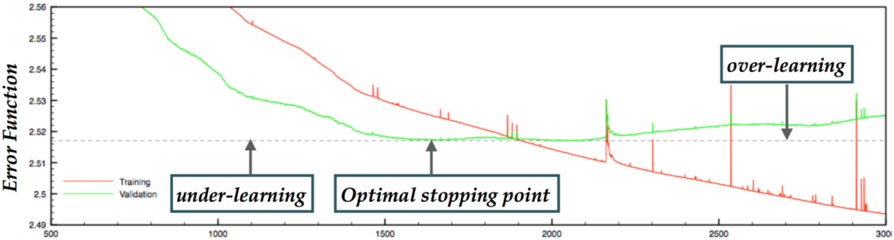
Optimisation: stopping criterion

Divide the data into two subsets (training & validation)

Train the NN on the training subset and compute χ^2 for each subset

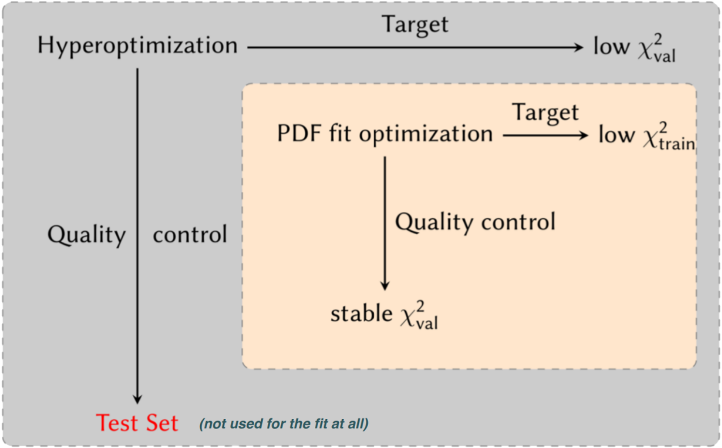
Stop when the training loss reaches the absolute minimum





The best fit does not coincide with the absolute minimum of the χ^2

Hypertoprimisation: fitting the methodology



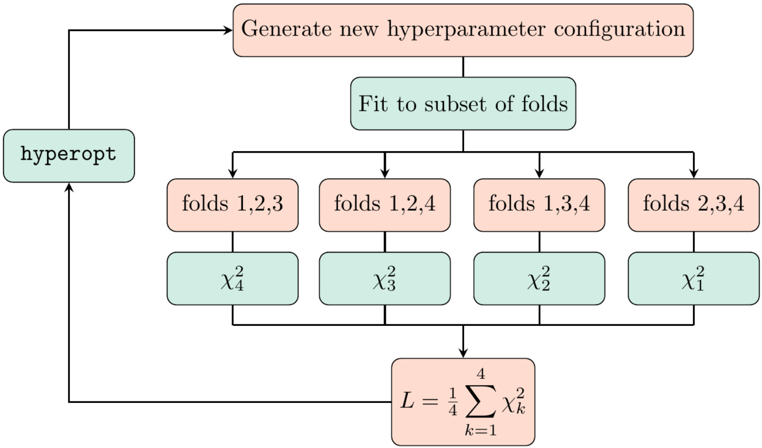
Compare to a Test Set (new set of data previously not used at all)

Who picks the Test Set? Automatic generalisation based on K foldings

Divide the data into n representative sets, fit $n-1$ sets and use the n-th set as test set

Hyperoptimise on mean and standard deviation of $\chi^2_{\mathrm{test},i}$, $i = 1 \ldots n$

Hyperoptimisation: K-folding



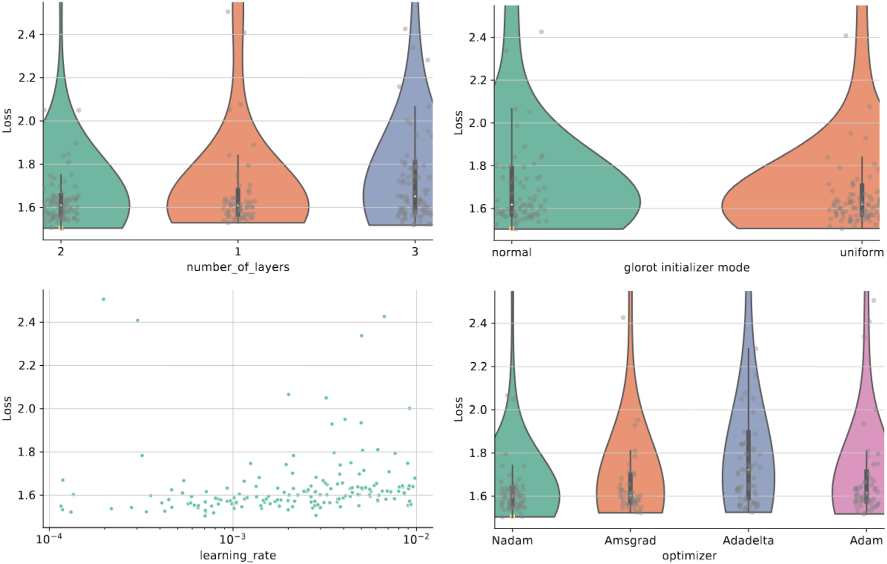
Compare to a Test Set (new set of data previously not used at all)

Who picks the Test Set? Automatic generalisation based on K foldings

Divide the data into n representative sets, fit $n-1$ sets and use the n-th set as test set

Hyperoptimise on mean and standard deviation of $\chi^2_{\mathrm{test},i}$, $i = 1 \ldots n$

Hyperparameters



2.2.3 Uncertainty representation

The Hessian method: general strategy

1. Expand the χ^2 about its global minimum at first (nontrivial) order

$$\chi^2\{\mathbf{a}\} \approx \chi^2\{\mathbf{a_0}\} + \delta a^i H_{ij} \delta a^j, \qquad H_{ij} = \frac{1}{2}\left.\frac{\partial^2 \chi^2\{\mathbf{a}\}}{\partial a_i \partial a_j}\right|_{\{\mathbf{a}\}=\{\mathbf{a_0}\}}$$

2. Assume linear error propagation for any observable $\mathcal{O}$ depending on $\{\mathbf{a}\}$

$$\langle \mathcal{O}\{\mathbf{a}\}\rangle \approx \mathcal{O}\{\mathbf{a_0}\} + a_i \left.\frac{\partial \mathcal{O}\{\mathbf{a}\}}{\partial a_i}\right|_{\{\mathbf{a}\}=\{\mathbf{a_0}\}} \qquad \sigma_{\mathcal{O}\{\mathbf{a}\}} \approx \sigma_{ij} \left.\frac{\partial \mathcal{O}\{\mathbf{a}\}}{\partial a_i}\,\frac{\partial \mathcal{O}\{\mathbf{a}\}}{\partial a_j}\right|_{\{\mathbf{a}\}=\{\mathbf{a_0}\}}$$

3. Determine σ_{ij} from H_{ij} from maximum likelihood (under Gaussian hypothesis)

$$\sigma_{ij}^{-1} = \left.\frac{\partial^2 \chi^2\{\mathbf{a}\}}{\partial a_i \partial a_j}\right|_{\{\mathbf{a}\}=\{\mathbf{a_0}\}} = H_{ij}$$

4. A C.L. about the best fit is obtained as the volume (in parameter space) about $\chi^2\{\mathbf{a_0}\}$ that corresponds to a fixed increase of the χ^2; for Gaussian uncertainties:

$$68\%\ \text{C.L.} \Longleftrightarrow \Delta\chi^2 = \chi^2\{\mathbf{a}\} - \chi^2\{\mathbf{a_0}\} = 1$$

The Hessian method: some remarks

1. Compact representation and computation of observables and their uncertainties

$$\langle \mathcal{O}[f(x,Q^2)]\rangle = \mathcal{O}[f_0(x,Q^2)]$$

$$\sigma_{\mathcal{O}}[f(x,Q^2)] = \frac{1}{2}\left[\sum_{i=1}^{N_{\rm par}}\left(\mathcal{O}[f_i(x,Q^2)] - \mathcal{O}[f_0(x,Q^2)]\right)^2\right]^{1/2}$$

2. Parameters can always be adjusted so that all eigenvalues of H_{ij} are equal to one (diagonalise H_{ij} and rescale the eigenvectors by their eigenvalues)

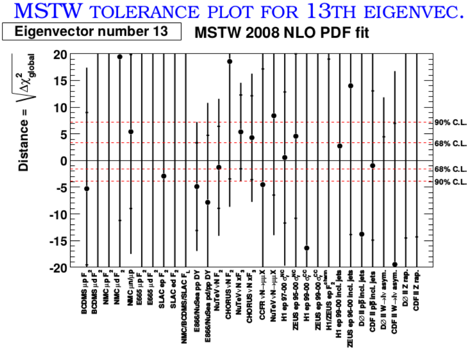
$$\delta a_i H_{ij} \delta a_j = \sum_{i=1}^{N_{\rm par}} \left[a_i'(a_i)\right]^2 \Longleftrightarrow \sigma_{\mathcal{O}\{\mathbf{a}'\}} = \left|\nabla' \mathcal{O}\{\mathbf{a}'\}\right|$$

   The total contribution to the uncertainty due to two different sources (possibly correlated) is obtained by simply adding them in quadrature

3. Any rotation in the space of parameters preserves the gradient (one can diagonalise a chosen observable without spoiling the result)
4. Unmanageable Hessian matrix if the numer of parameters is huge

The Hessian method: limitations

Uncertainties obtained with $\Delta\chi^2 = 1$ might be unrealistically small
(inadequacy of the linear approximation)



uncertainties tuned to the distribution of deviations from best-fits for single experiments

for each eigenvector in parameter space

determine the CL for the distribution of best-fits of each experiment

rescale to the $\Delta\chi^2 = T$ interval such that correct confidence intervals are reproduced

no statistically rigorous interpretation of T (tolerance)

The Monte Carlo method: general strategy

1. Generate (art) replicas of (exp) data according to the distribution

$$\mathcal{O}_i^{(art)(k)} = \mathcal{O}_i^{(exp)} + r_i^{(k)}\sigma_{\mathcal{O}_i}\,, \qquad i=1,\dots N_{\rm dat}\,, \qquad k=1,\dots,N_{\rm rep}$$

   where $r_i^{(k)}$ are (Gaussianly distributed) random numbers for each k-th replica ($r_i^{(k)}$ can be generated with any distribution, not neccesarily Gaussian)

2. Perform a fit for each replica $k=1,\dots,N_{\rm rep}$
3. Compact computation of observables and their uncertainties (PDF replicas are equally probable members of a statistical ensemble)

$$\langle\mathcal{O}[f(x,Q^2)]\rangle = \frac{1}{N_{\rm rep}}\sum_{k=1}^{N_{\rm rep}}\mathcal{O}[f^{(k)}(x,Q^2)]$$

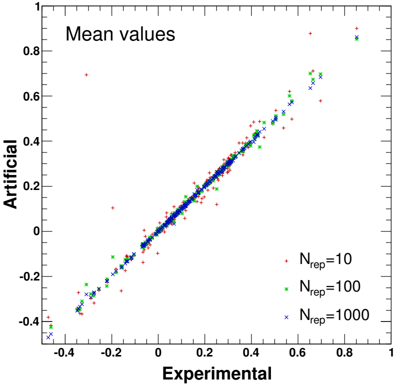
$$\sigma_{\mathcal{O}}[f(x,Q^2)] = \left[\frac{1}{N_{\rm rep}-1}\sum_{k=1}^{N_{\rm rep}}\Big(\mathcal{O}[f^{(k)}(x,Q^2)] - \langle\mathcal{O}[f(x,Q^2)]\rangle\Big)^2\right]^{1/2}$$

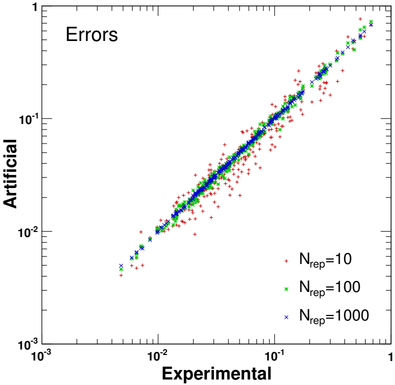
$\Rightarrow$ **no need to rely on linear approximation**

$\Rightarrow$ **computational expensive: need to perform $N_{\rm rep}$ fits instead of one**

The Monte Carlo method: determining the sample size

Require that the average over the replicas reproduces the central value of the original experimental data to a desired accuracy (the standard deviation reproduces the error and so on)

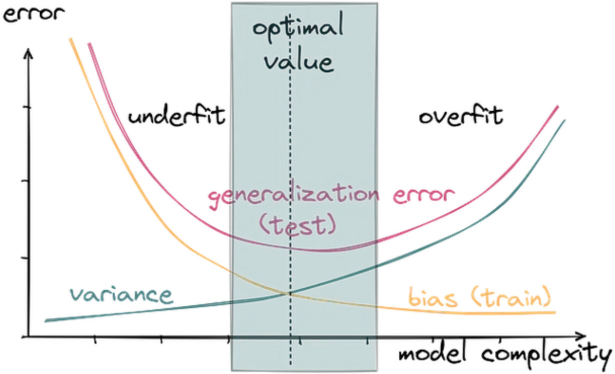


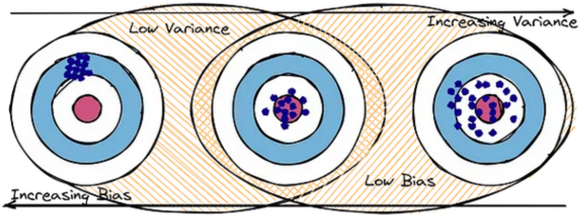


Accuracy of few % requires ~ 100 replicas

2.2.4 Validation

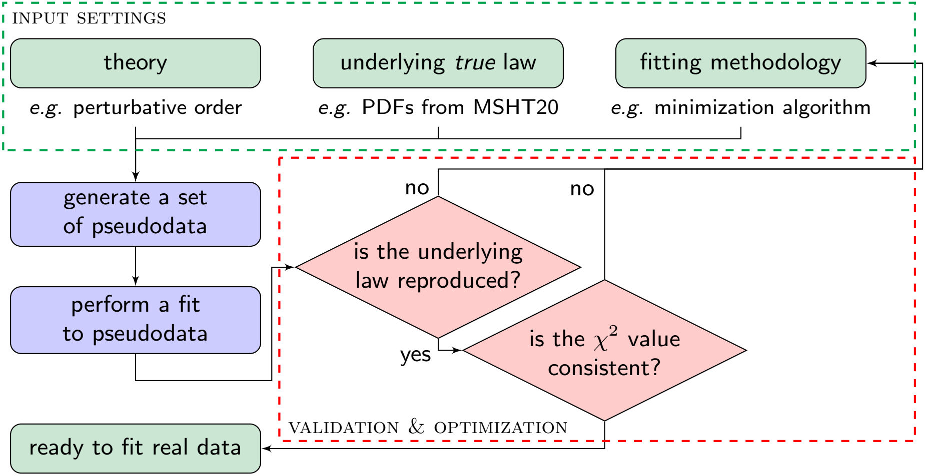
Accuracy vs precision or bias vs variance





Closure tests: general idea [JHEP 1504 (2015) 040]

Validation and optimisation of the fitting strategy with known underlying physical law

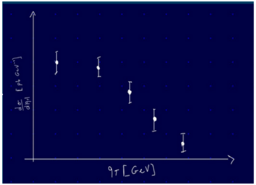


Full control of procedural uncertainties

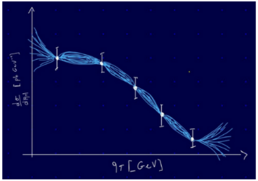
Closure Tests: Levels

Level 0

no fluctuations

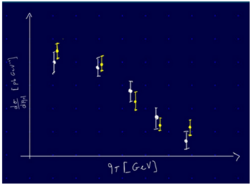


interpolation uncertainty

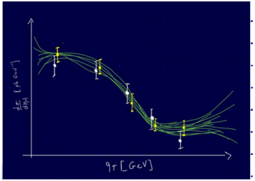


Level 1

Gaussian fluctuation

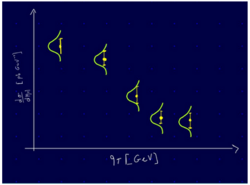


fuctional uncertainty

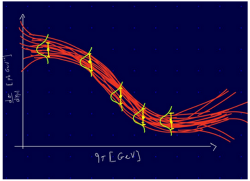


Level 2

Monte Carlo replicas



data uncertainty



Closure tests at work

Data region: closure tests

Fit PDFs to pseudodata generated assuming a known underlying law

Define bias and variance

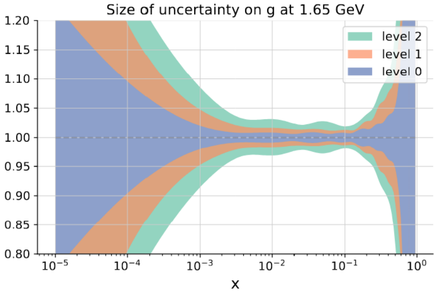
bias difference of central prediction and truth

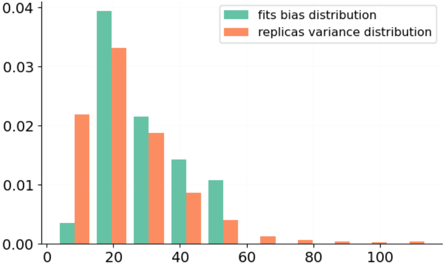
variance uncertainty of replica predictions

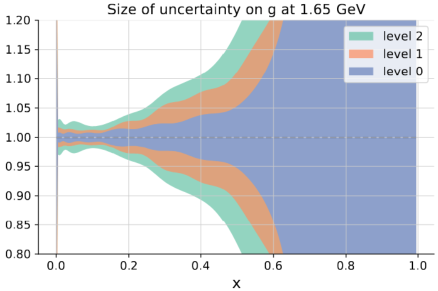
If PDF uncertainty faithful, then

E[bias] = variance

25 fits, 40 replicas each

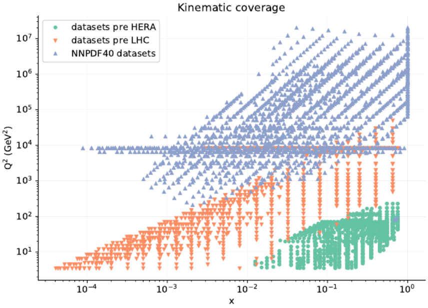


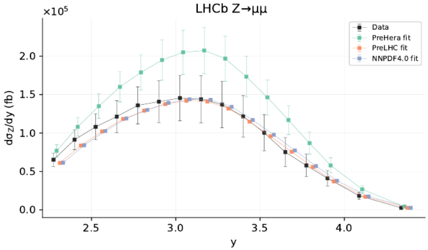




[EPJ C77 (2017) 663; EPJ C82 (2022) 330]

Future tests



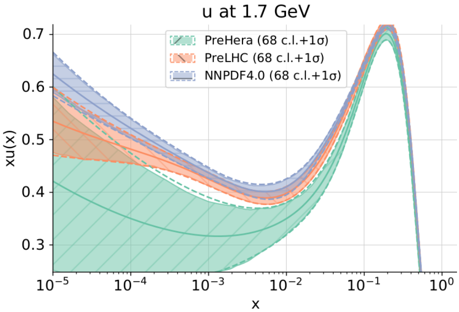


Extrapolation regions: future test

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

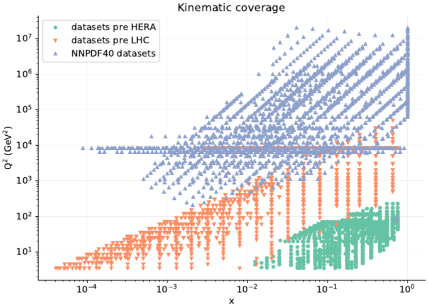
Data set	NNPDF4.0	pre-LHC	pre-HERA
pre-HERA	1.09	1.01	0.90
pre-LHC	1.21	1.20	23.1
NNPDF4.0	1.29	3.30	23.1

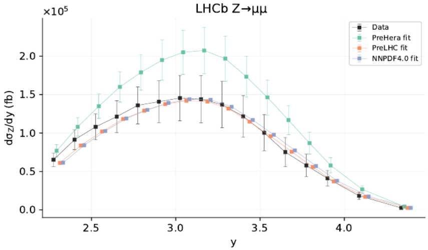
Only exp. cov. matrix



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

Future tests



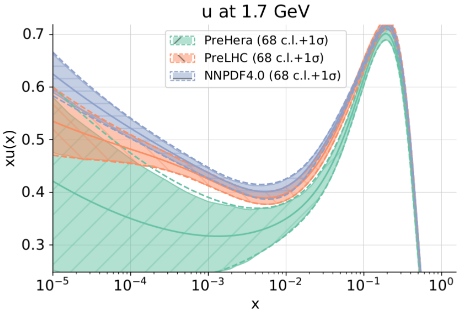


Extrapolation regions: future test

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]

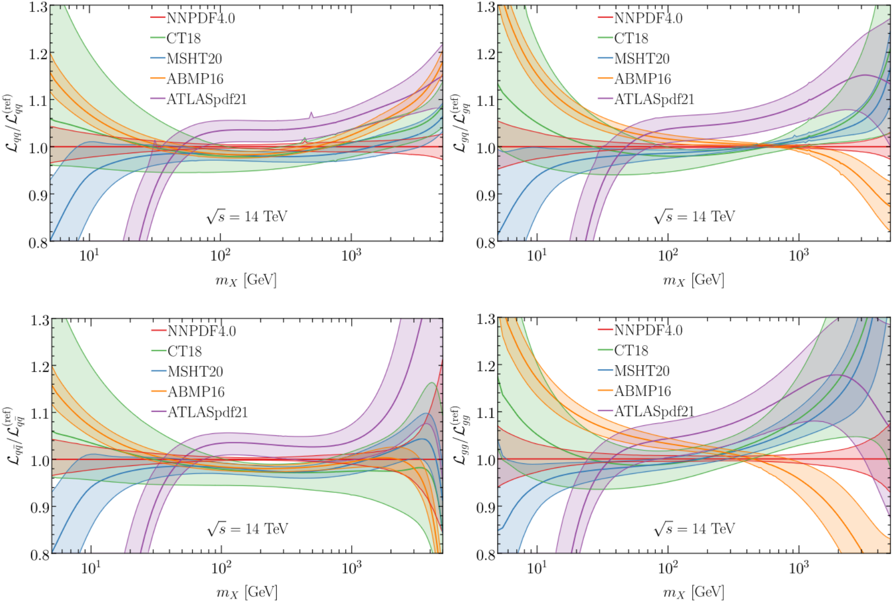
2.2.5 Benchmarks

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

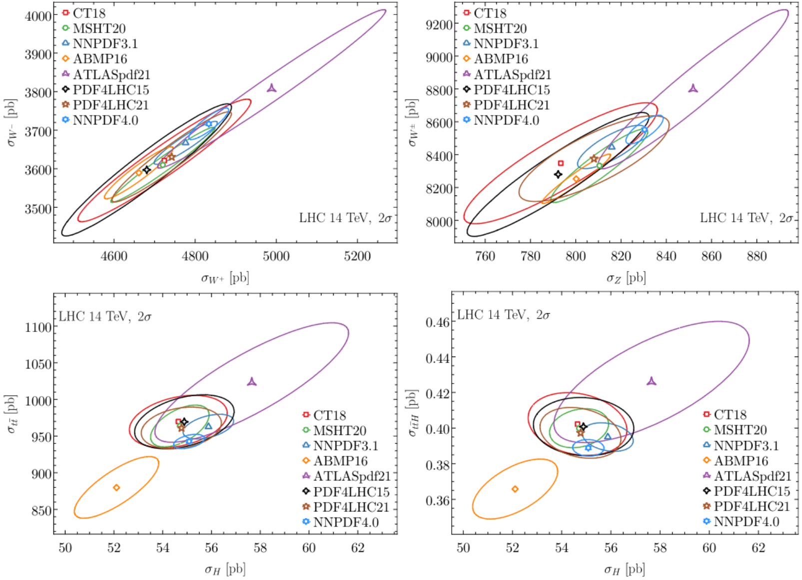
All PDF sets are available as (x, Q^2) interpolation grids through the LHAPDF library

Comparing PDF sets



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

Making predictions with PDFs



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

2.3 Summary of Lecture 2

Summary of Lecture 2

1. PDF accuracy can be improved by improving the theory
   $\longrightarrow$ proper treatment of heavy quarks is mandatory to describe DIS data
   $\longrightarrow$ evidence for intrinsic charm in the proton
   $\longrightarrow$ MHOUs can be estimated by scale variations
   $\longrightarrow$ inclusion of MHOUs stabilises fit quality
   $\longrightarrow$ electroweak corrections modify DGLAP equations
   $\longrightarrow$ the photon PDF is determined very precisely
   $\longrightarrow$ inclusion of photon PDFs impacts the gluon PDF
2. Devising the methodology is essential to minimise bias and variance
   $\longrightarrow$ bias is a measure of accuracy, variance is a measure of precision
   $\longrightarrow$ choices of parametrisation (polynomial vs neural network)
   $\longrightarrow$ choices of uncertainty representation (Hessian vs Monte Carlo)
   $\longrightarrow$ are all sources of bias and variance
   $\longrightarrow$ closure tests are a way to validate PDF uncertainites in the *seen* region
   $\longrightarrow$ future tests validate the generalisation power in the *unseen* region

Thank you