LHC data for PDF determination: challenges and prospects

(Re)interpretation of the LHC results for new physics

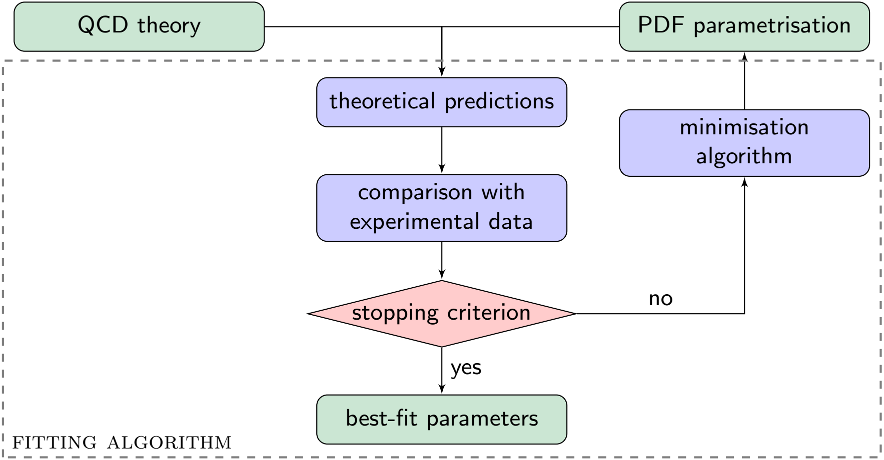
Emanuele R. Nocera

Università degli Studi di Torino and INFN — Torino

Durham University - 30 August 2023

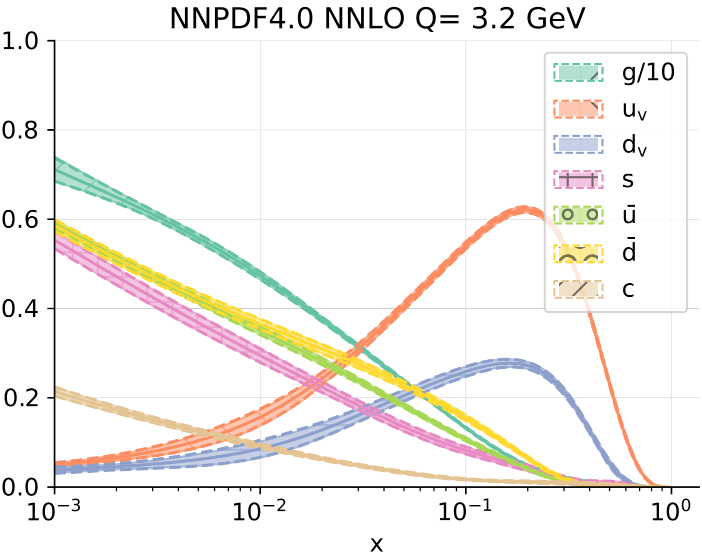
Determining PDFs from (LHC) experimental data

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



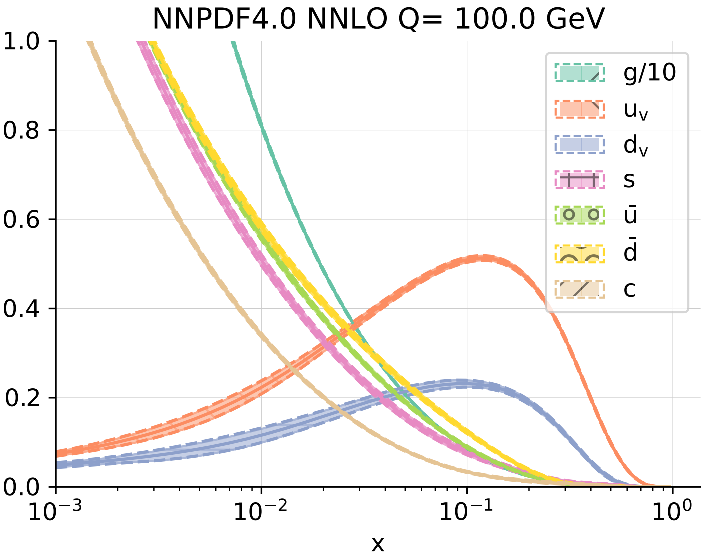
$$\chi^2 = \sum_{i,j}^{N_{\mathrm{dat}}}[T_i[\{\vec{a}\}]-D_i](\mathrm{cov}^{-1})_{ij}[T_j[\{\vec{a}\}]-D_j]$$ with $\{\vec{a}\}$ the set of parameters

A modern PDF set: NNPDF4.0 (2022)



[2022 PDG Review of Particle Physics]

A modern PDF set: NNPDF4.0 (2022)



[2022 PDG Review of Particle Physics]

Making predictions with PDFs

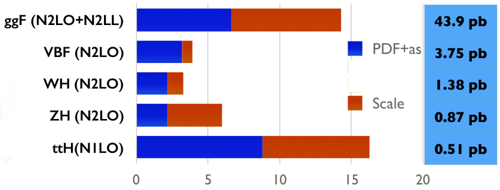
PDF uncertainty is often the dominant source of uncertainty in LHC cross sections

Higgs boson characterisation

Determination of SM parameters, such as the mass of the W boson

Searches for beyond SM physics at large invariant mass of the final state

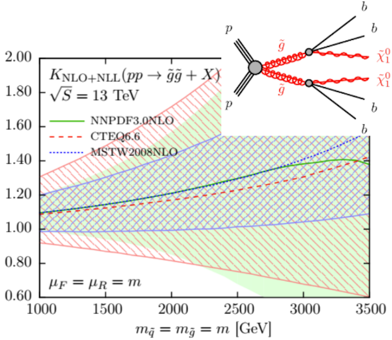
Precision



Channel	$m_{W^+} - m_{W^-}$ [MeV]	Stat. Unc.	Muon Unc.	Elec. Unc.	Recoil Unc.	Bckg. Unc.	QCD Unc.	EW Unc.	PDF Unc.	Total Unc.
$W \to e\nu$	−29.7	17.5	0.0	4.9	0.9	5.4	0.5	0.0	24.1	30.7
$W \to \mu\nu$	−28.6	16.3	11.7	0.0	1.1	5.0	0.4	0.0	26.0	33.2
Combined	−29.2	12.8	3.3	4.1	1.0	4.5	0.4	0.0	23.9	28.0

[Plot from the CERN Yellow Report 2016]

Discovery



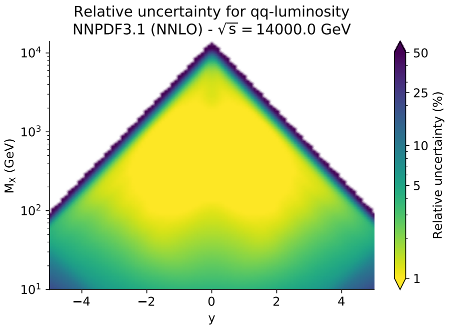
[EPJC 76 (2016) 53]

How large are PDF uncertainties?

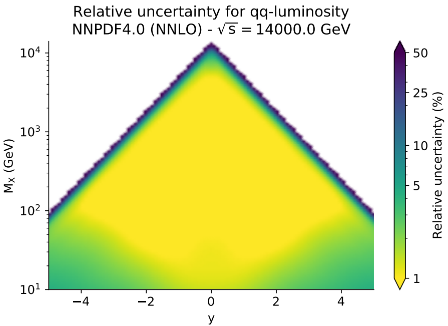
$$\mathcal{L}_{ij}(M_X, y, \sqrt{s}) = \frac{1}{s} f_i\left(\frac{M_X e^{y}}{\sqrt{s}}, M_X\right) f_j\left(\frac{M_X e^{-y}}{\sqrt{s}}, M_X\right)$$

SINGLET

NNPDF3.1 (NNLO) **2017**



NNPDF4.0 (NNLO) **2022**



Steady progress towards 1% relative uncertainties on $\mathcal{L}_{ij}$ in a broad kinematic range

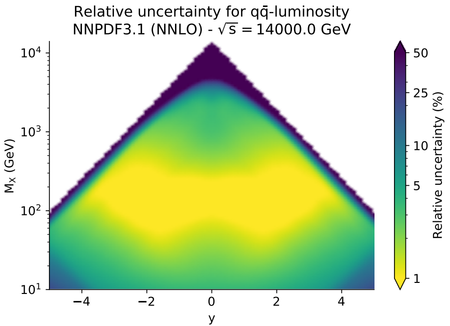
How are the data getting us there?

How large are PDF uncertainties?

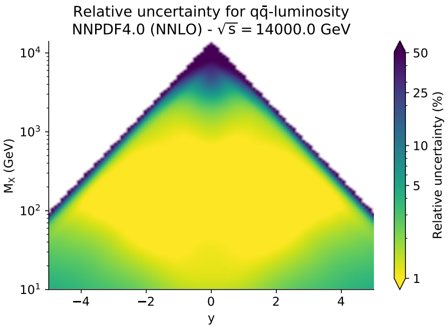
$$\mathcal{L}_{ij}(M_X, y, \sqrt{s}) = \frac{1}{s} f_i\left(\frac{M_X e^y}{\sqrt{s}}, M_X\right) f_j\left(\frac{M_X e^{-y}}{\sqrt{s}}, M_X\right)$$

SINGLET

NNPDF3.1 (NNLO) **2017**



NNPDF4.0 (NNLO) **2022**



Steady progress towards 1% relative uncertainties on $\mathcal{L}_{ij}$ in a broad kinematic range

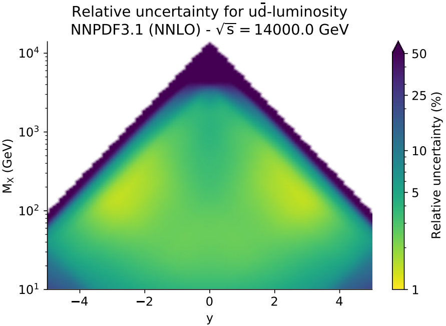
How are the data getting us there?

How large are PDF uncertainties?

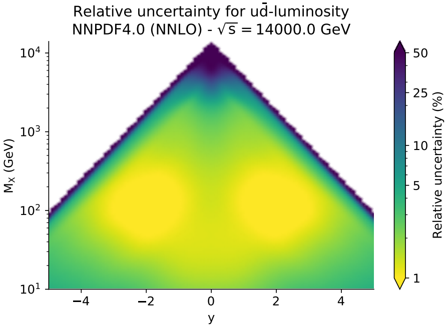
$$\mathcal{L}_{ij}(M_X, y, \sqrt{s}) = \frac{1}{s} f_i\left(\frac{M_X e^y}{\sqrt{s}}, M_X\right) f_j\left(\frac{M_X e^{-y}}{\sqrt{s}}, M_X\right)$$

FLAVOURS

NNPDF3.1 (NNLO) **2017**



NNPDF4.0 (NNLO) **2022**



Steady progress towards 1% relative uncertainties on $\mathcal{L}_{ij}$ in a broad kinematic range

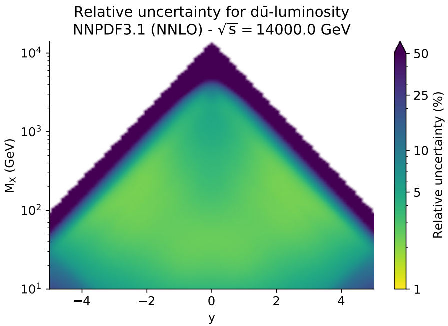
How are the data getting us there?

How large are PDF uncertainties?

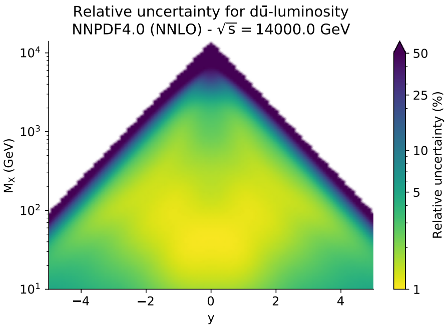
$$\mathcal{L}_{ij}(M_X, y, \sqrt{s}) = \frac{1}{s} f_i\left(\frac{M_X e^{y}}{\sqrt{s}}, M_X\right) f_j\left(\frac{M_X e^{-y}}{\sqrt{s}}, M_X\right)$$

FLAVOURS

NNPDF3.1 (NNLO) **2017**



NNPDF4.0 (NNLO) **2022**



Steady progress towards 1% relative uncertainties on $\mathcal{L}_{ij}$ in a broad kinematic range

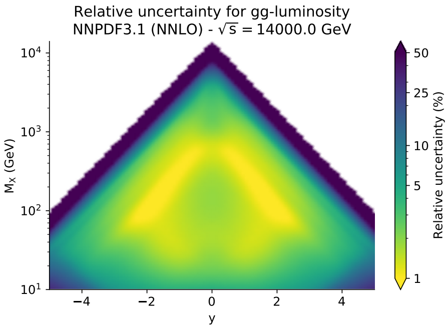
How are the data getting us there?

How large are PDF uncertainties?

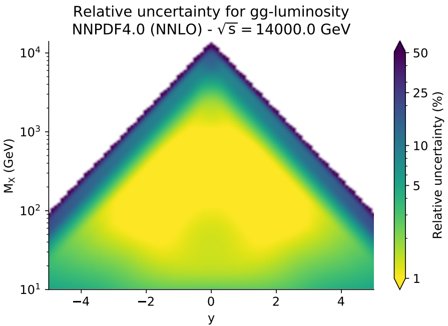
$$\mathcal{L}_{ij}(M_X, y, \sqrt{s}) = \frac{1}{s} f_i\left(\frac{M_X e^y}{\sqrt{s}}, M_X\right) f_j\left(\frac{M_X e^{-y}}{\sqrt{s}}, M_X\right)$$

GLUON

NNPDF3.1 (NNLO) **2017**



NNPDF4.0 (NNLO) **2022**



Steady progress towards 1% relative uncertainties on $\mathcal{L}_{ij}$ in a broad kinematic range

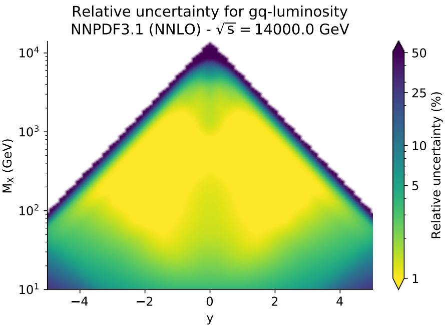
How are the data getting us there?

How large are PDF uncertainties?

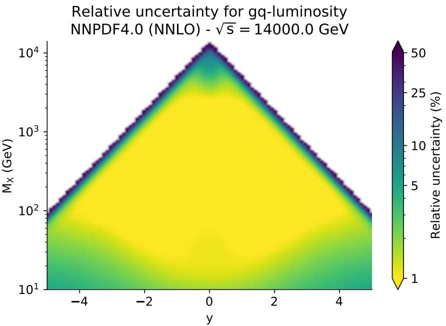
$$\mathcal{L}_{ij}(M_X, y, \sqrt{s}) = \frac{1}{s} f_i\left(\frac{M_X e^y}{\sqrt{s}}, M_X\right) f_j\left(\frac{M_X e^{-y}}{\sqrt{s}}, M_X\right)$$

GLUON

NNPDF3.1 (NNLO) **2017**



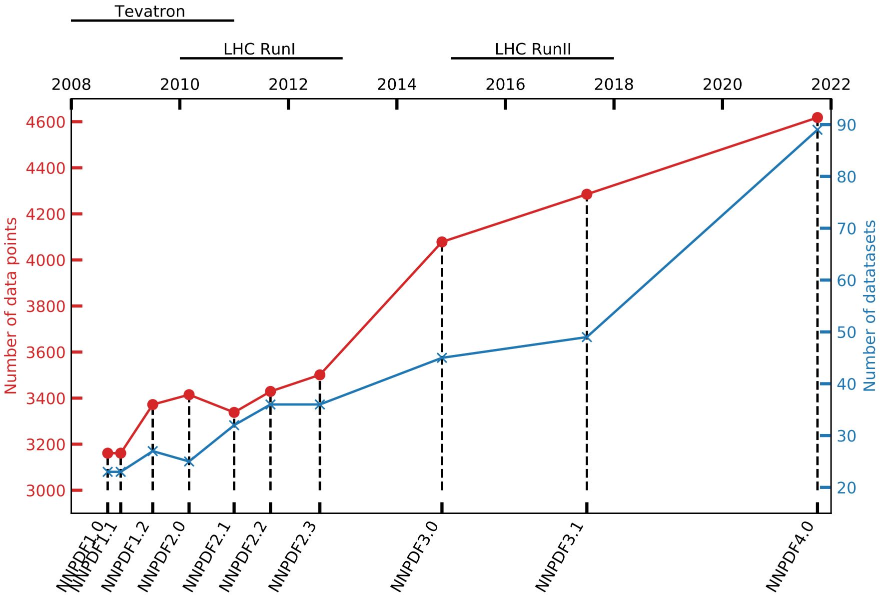
NNPDF4.0 (NNLO) **2022**



Steady progress towards 1% relative uncertainties on $\mathcal{L}_{ij}$ in a broad kinematic range

How are the data getting us there?

Overview of (NNPDF) experimental data: 2008–2022



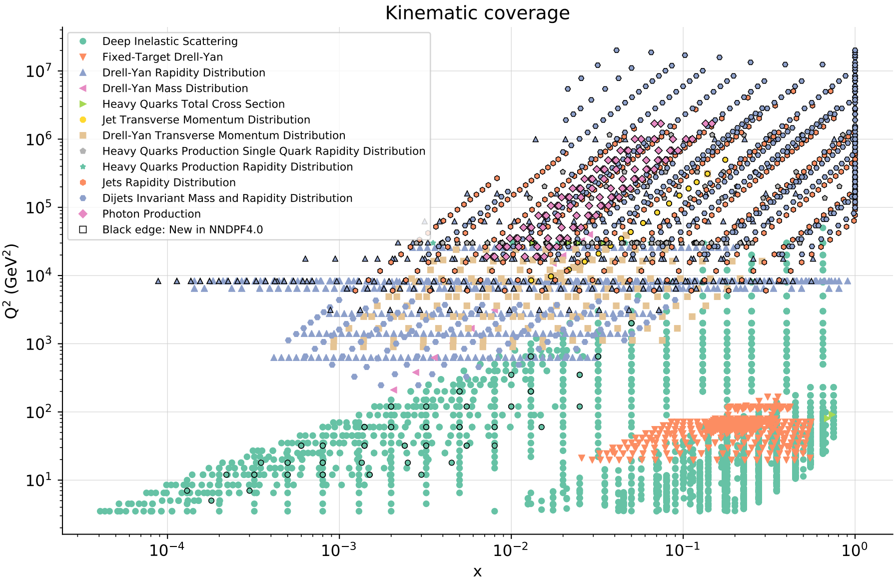
Data implementation and selection

There are general considerations to make before considering a new data set

1. Which observables?
   So far we focused on largely inclusive observables (with limited exceptions).
   Consider more exclusive observables? Which ones?
2. Which measurements?
   LHC RunII full luminosity. What else?
   Look at Hepdata and experimental collaboration pages
3. What's a good data set?
   Consistent.
   Accuracy. How can we test/improve consistency?
   Constraining.
   Precision. Should this be a criterion not to implement/include a data set?
   Redundant.
   Robustness. In-sample vs out-of-sample; K-folding.

Unfortunately a data set has to be implemented (in the NNPDF framework) in order to test whether it is a good data set or not.

Overview of (NNPDF) experimental data: 2022



Precision of the data of the order of percent; mostly from correlated systematic uncertainties

Data distribution: asymmetric uncertainties

Assumption: Uncertainties are well-behaved Gaussian errors
Sometimes they are NOT

$Y =$ '*best value*'$^{+\Delta_+}_{-\Delta_-}$ Δ_+ and Δ_- can be positive or negative

Possible origins of asymmetric uncertainties in LHC data:
non-parabolic χ^2 or log-likelihood curves
non-linear error propagation
systematic uncertainties (example: two-point systematic uncertainties)

Let us indicate with $\mathbf{X}$ the set of quantities that concur to construct Y, *i.e.* $Y = Y(\mathbf{X})$

Typically, $\mathbf{X}$ is unknown (to the final user)

In a Bayesian framework, it can be shown that [physics/0403086]

$$E(Y) \approx Y(E[X]) + \sum_i \delta_i$$
$$\sigma^2(Y) \approx \sum_i \bar{\Delta}_i^2 + 2\sum_i \delta_i^2$$

with $\delta_i = \frac{\Delta_+ - \Delta_-}{2}$ and $\bar{\Delta} = \frac{\Delta_+ + \Delta_-}{2}$

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

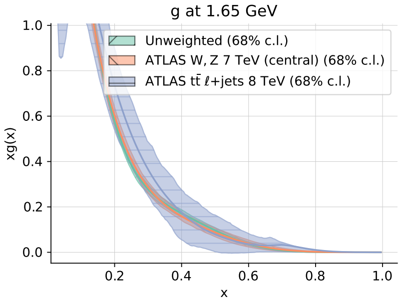
unnatural PDF shapes appear

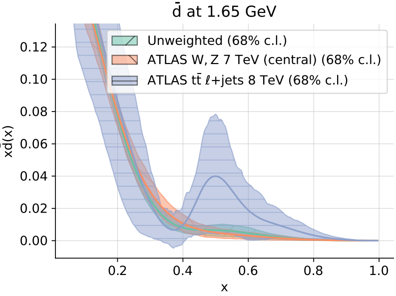
error in other data sets increases

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





Data inconsistency: experimental correlations

Single inclusive jet data from ATLAS 7 TeV

default correlations: terrible χ^2

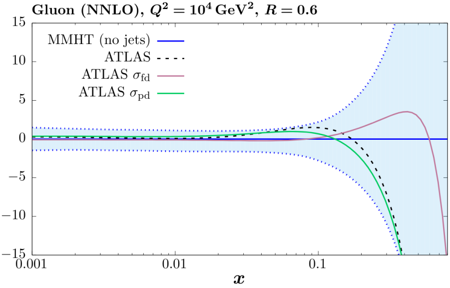
(correlations across rapidity bins)

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

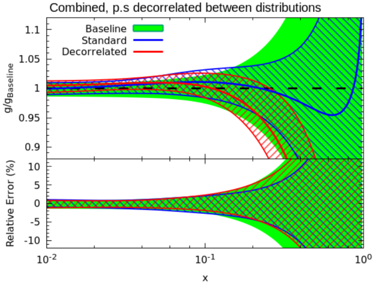
(correlations across different spectra)

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]

Data inconsistency: experimental correlations

1. What is correlated with what?
   Correlations between data points in a data set
      Easy (clear). Identify the various sources (~ 300) of uncertainty.
   Between data sets in the same experiment
      Medium (usually clear). Put in correspondence uncertainties with the same name.
   Between different experiments
      Difficult (typically obscure). Usually not clear how to match uncertainties.
2. How much are uncertainties correlated? Assumption: 100%.
   Sometimes this is NOT realistic. There exist decorrelation models.
3. Do experimentalists release complete information to properly treat correlations?
   Information on correlation/decorrelation provided years after publication.

Systematic uncertainty	8 TeV W + jets	8 TeV Z + jets	8 TeV $t\bar{t}$ lepton + jets	13 TeV $t\bar{t}$ lepton + jets	8 TeV inclusive jets
Jet flavour response	JetScaleFlav2	Flavor Response	flavres-jes	JET29NP JET Flavour Response	syst JES Flavour Response*
Jet flavour composition	JetScaleFlav1Known	Flavor Comp	flavcomp-jes	JET29NP JET Flavour Composition	syst JES Flavour Comp
Jet punchthrough	JetScalepunchT	Punch Through	punch-jes	-	syst JES PunchThrough MC15
Jet scale	JetScalePileup2	PU OffsetMu	pileoffmu-jes	-	syst JES Pileup MuOffset
	-	PU Rho	pileoffrho-jes	JET29NP JET Pileup RhoTopology	syst JES Pileup Rho topology*
	JetScalePileup1	PU OffsetNPV	pileoffnpv-jes	JET29NP JET Pileup OffsetNPV	syst JES Pileup NPVOffset
	-	PU PtTerm	pileoffpt-jes	JET29NP JET Pileup PtTerm	syst JES Pileup Pt term
Jet JVF selection	JetJVFcut	JVF	jetvxfrac	-	syst JES Zjets JVF
B-tagged jet scale	-	btag-jes	JET29NP JET BJES Response	-	-
Jet resolution	-	jeten-res	JET JER SINGLE NP	-	-
Muon scale	-	-	mup-scale	MUON SCALE	-
Muon resolution	-	-	muonms-res	MUON MS	-
Muon identification	-	-	muid-res	MUON ID	-
Diboson cross section	-	-	dibos-xsec	Diboson xsec	-
Z + jets cross section	-	-	zjet-xsec	Zjets xsec	-
Single-t cross section	-	-	singletop-xsec	st xsec	-

[EPJ C82 (2022) 438]

Good knowledge of experimental correlations is important

Let us call A the $N_{\rm dat} \times N_{\rm err}$ matrix of uncertainties, such that cov$=AA^t$

If the theory is known, fixed and correct:

$$\langle \chi^2_{\rm true} \rangle = \|A^+ A\|_F = N_{\rm dat}$$

If we know $\bar{A}$ instead of A:

$$\langle \bar{\chi}^2 \rangle = \|\bar{A}^+ A\|_F$$

The χ^2 is *stable* if:

$$\langle \bar{\chi}^2 \rangle - \langle \chi^2 \rangle = \|\bar{A}^+ A\|_F - N_{\rm dat} < \sqrt{2N_{\rm dat}}$$

If not, define $A_{\rm reg}$ by clipping the singular values of the correlated part of $\bar{A}$ to δ, whenever these are smaller than δ; the rest of the singular vectors are left unchanged

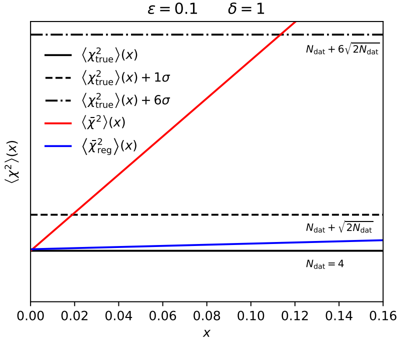
$$\langle \chi^2_{\rm reg} \rangle = \|A^+_{\rm reg} A\|_F$$

Assumptions:

correlations are determined less precisely than variances and inaccuracy is limited to a small number of uncertainties

$$A(x) = \begin{pmatrix} \epsilon & 0 & 0 & 0 & 1 & 0 \\ 0 & \epsilon & 0 & 0 & 1 & 0 \\ 0 & 0 & \epsilon & 0 & 1 & 0 \\ 0 & 0 & 0 & \epsilon & 1-x & \sqrt{1-(1-x)^2} \end{pmatrix}$$

$$\bar{A} = \begin{pmatrix} \epsilon & 0 & 0 & 0 & 1 & 0 \\ 0 & \epsilon & 0 & 0 & 1 & 0 \\ 0 & 0 & \epsilon & 0 & 1 & 0 \\ 0 & 0 & 0 & \epsilon & 1 & 0 \end{pmatrix}$$

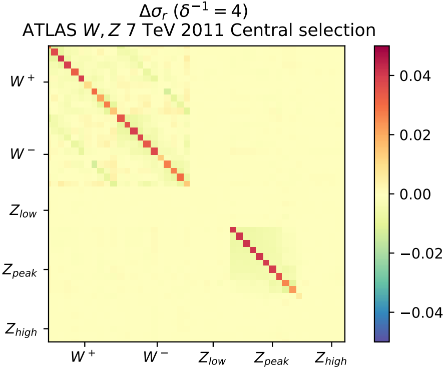


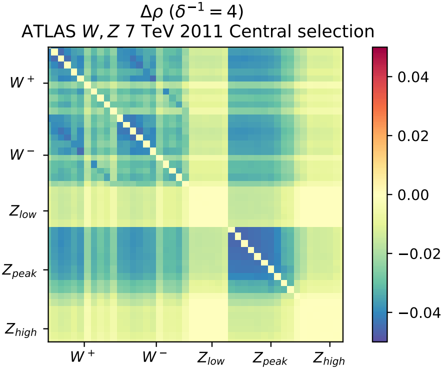
[EPJ C82 (2022) 956]

Regularising the NNPDF4.0 data set [EPJ C82 (2022) 956]

Let us test the regularisation procedure on the NNPDF4.0 data set

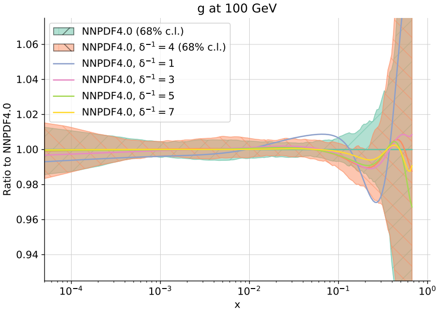
As an example, let us focus on a specific data set:
ATLAS W, Z 7 TeV 2011 central selection [EPJ C77 (2017) 367]

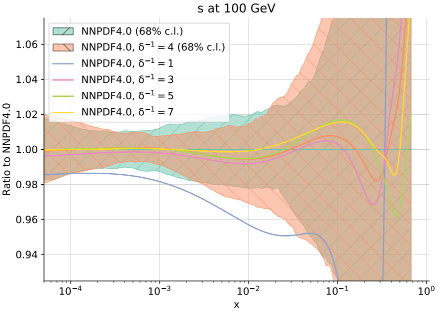




Data set	Z	$\delta^{-1}=1$ $\Delta\sigma_r$	$\delta^{-1}=1$ $\lvert\Delta\rho\rvert$	$\delta^{-1}=2$ $\Delta\sigma_r$	$\delta^{-1}=2$ $\lvert\Delta\rho\rvert$	$\delta^{-1}=3$ $\Delta\sigma_r$	$\delta^{-1}=3$ $\lvert\Delta\rho\rvert$	$\delta^{-1}=4$ $\Delta\sigma_r$	$\delta^{-1}=4$ $\lvert\Delta\rho\rvert$	$\delta^{-1}=5$ $\Delta\sigma_r$	$\delta^{-1}=5$ $\lvert\Delta\rho\rvert$	$\delta^{-1}=7$ $\Delta\sigma_r$	$\delta^{-1}=7$ $\lvert\Delta\rho\rvert$
ATLAS W, Z 7 TeV	9.0	94.4	0.50	21.9	0.19	8.63	0.09	4.15	0.05	2.12	0.02	0.50	0.01

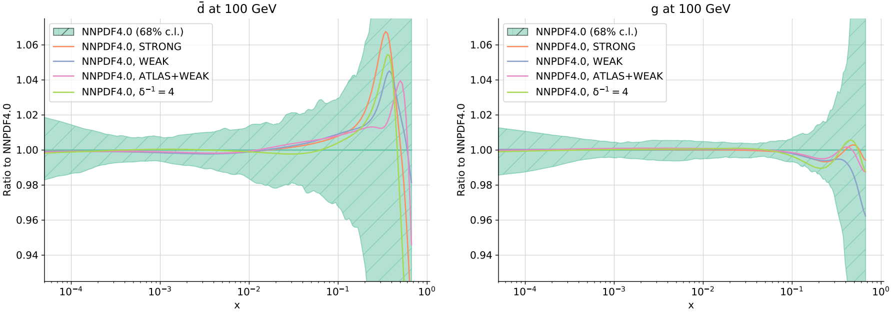
Regularising the NNPDF4.0 data set [EPJ C82 (2022) 956]





		$\chi^2/N_{\rm dat}$						
Data set	$N_{\rm dat}$	NNPDF4.0	$\delta^{-1}=1$	$\delta^{-1}=2$	$\delta^{-1}=3$	$\delta^{-1}=4$	$\delta^{-1}=5$	$\delta^{-1}=7$
Deep-inelastic scattering	3089	1.12	0.64	1.02	1.09	1.11	1.12	1.12
Fixed-target Drell-Yan	195	0.98	0.48	0.90	0.96	0.97	0.97	0.99
Tevatron Drell-Yan	65	1.11	0.48	0.71	0.85	0.93	1.02	1.10
ATLAS total	679	1.24	0.50	0.84	0.97	1.04	1.10	1.19
W, Z 7 TeV CC	46	1.92	0.31	0.74	0.94	1.21	1.47	1.76
CMS total	474	1.31	0.39	0.83	1.08	1.21	1.26	1.28
LHCb total	116	1.55	0.73	1.41	1.53	1.56	1.55	1.55
Total	4618	1.16	0.58	0.97	1.07	1.11	1.13	1.15

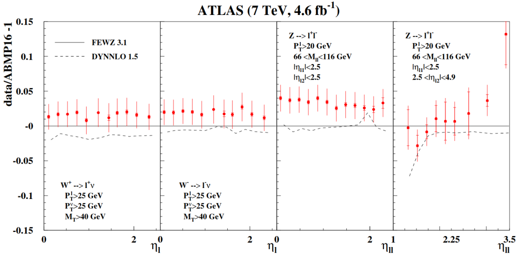
Regularising the NNPDF4.0 data set [EPJ C82 (2022) 956]



Data set	$N_{\rm dat}$	$\chi^2/N_{\rm dat}$ NNPDF4.0	STRONG	WEAK	ATLAS	ATLAS+WEAK	$\delta^{-1}=4$
Deep-inelastic scattering	3089	1.12	1.12	1.12	1.12	1.12	1.11
Fixed-target Drell-Yan	195	0.98	1.00	0.99	0.99	0.99	0.97
Tevatron Drell-Yan	65	1.11	1.10	1.09	1.09	1.10	0.93
ATLAS total	679	1.24	1.24	1.24	1.23	1.24	1.04
CMS total	474	1.31	1.31	1.31	1.31	1.30	1.21
LHCb total	116	1.55	1.56	1.54	1.55	1.55	1.56
Total	4618	1.16	1.16	1.16	1.16	1.16	1.11

Benchmarks: PDFs

Benchmark of the theory

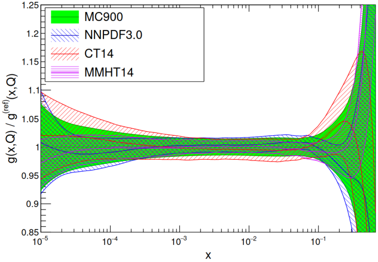


Be careful about the use of different NNLO codes for DY production in particular when experiments use non-optimal fiducial cuts [EPJ C81 (2021) 573]

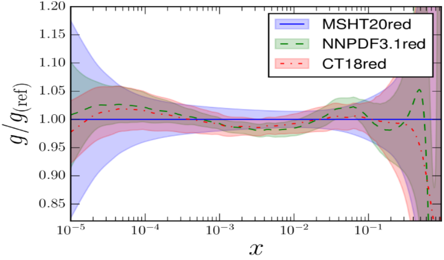
NNLO corrections usually implemented via K-factors

NNLOJet/ApplFast provide NNLO lookup tables for a limited set of data

Benchmark of PDF sets



[PDF4LHC15 benchmark, JPG 43 (2016) 023001]



[PDF4LHC21 benchmark, JPG 49 (2022) 080501]

Public codes: PDFs [EPJ C81 (2021) 958]

NNPDF

Tests passing | DOI 10.5281/zenodo.6542572

NNPDF: An open-source machine learning framework for global analyses of parton distributions

The NNPDF collaboration determines the structure of the proton using Machine Learning methods. This is the main repository of the fitting and analysis frameworks. In particular it contains all the necessary tools to reproduce the NNPDF4.0 PDF determinations.

Documentation

The documentation is available at https://docs.nnpdf.science/

Install

See the NNPDF installation guide for the conda package, and how to build from source.

Please note that the conda based workflow described in the documentation is the only supported one. While it may be possible to set up the code in different ways, we won't be able to provide any assistance.

We follow a rolling development model where the tip of the master branch is expected to be stable, tested and correct. For more information see our releases and compatibility policy.

`https://github.com/NNPDF`

NNPDF

Search docs

- Getting started
- Fitting code: `n3fit`
- Code for data: validphys
- Handling experimental data: Buildmaster
- Storage of data and theory predictions
- Theory
- Chi square figures of merit
- Contributing guidelines and tools
- Releases and compatibility policy
- Continuous integration and deployment
- Servers
- External codes
- Tutorials

The NNPDF Commondata format

A framework to standardise the input experimental information
tailored to (NNPDF) PDF determination

A framework based on HepData

A framework developed to ensure Flexibility, (Re)producibility and Scalability

A framework to help experimentalists in their analyses

Name	Last commit message	Last commit da...
..		
rawdata	added HEPdata tables for alternative scenario	3 weeks ago
data.yaml	added ATLAS_1JET_8TEV_R06 folder	last month
filter.py	finalized ATLAS_1JET	3 weeks ago
filter_utils.py	finalized ATLAS_1JET	3 weeks ago
kinematics.yaml	added ATLAS_1JET_8TEV_R06 folder	last month
metadata.yaml	this is how variants are included in metadataa.yaml	3 weeks ago
uncertainties.yaml	finalized ATLAS_1JET	3 weeks ago
uncertainties_decorrelated.yaml	finalized ATLAS_1JET	3 weeks ago

STAY TUNED!

Conclusions

Collider measurements are reducing PDF uncertainties to few percent.

This is key to make precision and discovery physics.

This opens up some challenges, among others, in the interpretation of the data.

Understand experimental systematic uncertainties and their correlations:
measure the stability of covariance matrices/provide stable covariance matrices
provide information on correlation models and/or regularise the available information.

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

The NNPDF Collaboration is developing a Commondata framework
to standardise the input experimental information tailored to PDF determination.
The framework is based on HepData and aims at fostering
the cross-talk between PDF experimentalists and phenomenologists.

Conclusions

Collider measurements are reducing PDF uncertainties to few percent.

This is key to make precision and discovery physics.

This opens up some challenges, among others, in the interpretation of the data.

Understand experimental systematic uncertainties and their correlations:
measure the stability of covariance matrices/provide stable covariance matrices
provide information on correlation models and/or regularise the available information.

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

The NNPDF Collaboration is developing a Commondata framework
to standardise the input experimental information tailored to PDF determination.
The framework is based on HepData and aims at fostering
the cross-talk between PDF experimentalists and phenomenologists.

Thank you