Parton distributions and the LHC

Nathan Hartland
University of Oxford

The NNPDF Collaboration:
R. D. Ball, V. Bertone, S. Carrazza, C. Deans,
L. Del Debbio, S. Forte, A. Guffanti,
N.H, J.I. Latorre, J. Rojo and M. Ubiali.

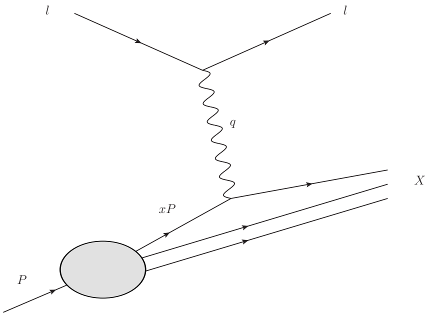
University of Göttingen
Friday 8th May 2015

What is a parton distribution function?

QCD Factorisation:

When considering a scattering process with a single hadron in the initial state, the calculation may be factorized into a soft part and a perturbatively calculable hard part.

$$\sigma_X(Q^2) = \sum_a \int_0^1 dx\, f_a(x, \mu^2)\sigma_{q_a \to X}\left(x, \frac{Q^2}{\mu^2}\right)$$



$\sigma_{q_a \to X}$ - perturbative

Hard cross section for lepton scattering off a parton of flavour a, carrying a fraction x of the parent hadron's momentum.

$f_a(x, \mu^2)$ - non-perturbative

Parton distribution function describing nonperturbative dynamics of target hadron. At LO can be interpreted as the probability of finding a parton of flavour a with momentum fraction x inside the target hadron.

What is on the market?

A dizzying array of options!

- Lots of recent activity in the 'PDF industry'.

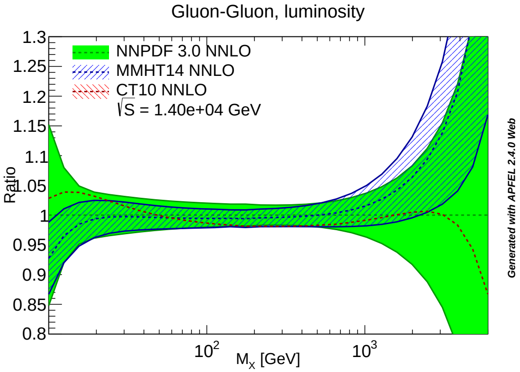
Agreement between modern global sets generally very good
(with areas of important difference)

Global sets:

- NNPDF3.0 [arxiv:1410.8849]
- MMHT14 [arxiv:1410.3989]
- CT14 [preliminary]

(more) Restrictive sets:

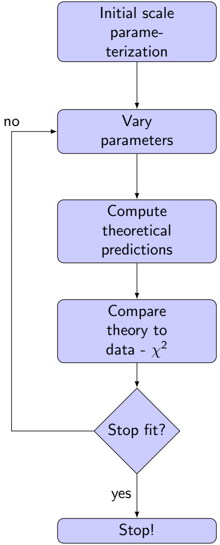
- ABM12 [arxiv:1310.3059]
- HERAPDF2.0 [preliminary]



Comprehensive benchmarking program of newer PDF sets underway in PDF4LHC

How can we determine proton PDFs?

1. Theoretical input
   - (N)NLO QCD, α_S, HQ Treatment
2. PDF Parameterization
   - What is a suitable choice of functional form?
3. Theoretical predictions
   - How can we make fast pQCD predictions for experimental data while including higher order corrections?
4. Comparison to data
   - What does (LHC) data tell us about proton structure?



In this talk $\rightarrow \chi^2[f] = \frac{1}{N_{\rm dat}} \sum_{i,j}^{N_{\rm dat}} (D_i - T_i[f])\, \sigma_{ij}^{-1}\, (D_j - T_j[f])$.

What do we know from theory?

Factorisation scale dependance:

PDFs evolve with scale according to the DGLAP equations.

$$\mu^2 \frac{\partial f(x,\mu^2)}{\partial \mu^2} = \int_x^1 \frac{dy}{y} P\left(\frac{x}{y}\right) f(y,\mu^2)$$

Where P are the perturbatively calculable splitting functions.

Theoretical constraints:

- PDF Sum Rules

$$\int_0^1 dx\, x(\Sigma(x) + g(x)) = 1, \qquad \sum_q \int_0^1 dx\, (q(x) - \bar{q}(x)) = 3$$

- Approx asymptotic behaviour

$$x \to 0: \quad f(x) \sim x^\alpha$$
$$x \to 1: \quad f(x) \sim (1-x)^\beta$$

- Positivity of *physical observables* (F, σ)
  - Beyond LO pdfs are not restricted to be positive.

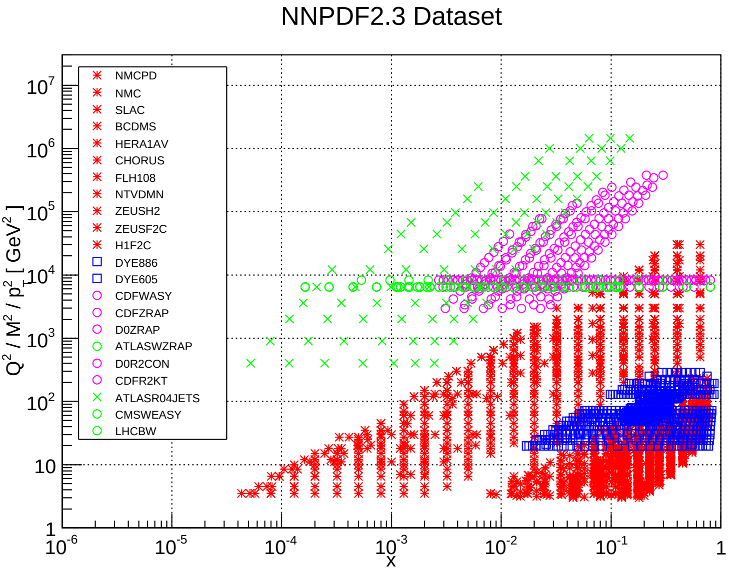
Aside from these constraints,
x dependence must be determined by fitting to experimental data!

What do we know from experiment?

Many sources of precise experimental information on PDFs

- The **backbone** - DIS data
  - Large fixed-target DIS datasets available from SLAC/BCDMS/NMC.
  - Precise, clean data from HERA.
  - NC data constrains quark singlet, gluon distributions.
    CC data gives a handle on light flavour separation.
  - ν-DIS data (still) provides most of the constraints upon strange distributions.
- **Flavour separation** - Drell-Yan data
  - Low-energy fixed-target data from FNAL
  - Tevatron and LHC data providing important constraints.
- **Large-x gluon** - Inclusive jet data (also t-production)
  - Substantial constraints from Tevatron/LHC inclusive jet measurements.
  - $t\bar{t}$ data becoming more important.

Dataset selection - kinematic coverage



Parton distribution fitting - initial scale parameterization

- PDFs at the initial scale are parametrised by some functional form

Typical Parameterisations

- MSTW08 $\sim$ 28 total PDF parameters

$$f_v(x) \sim ax^b(1-x)^c(1+d\sqrt{x}+ex),$$

- CT10 $\sim$ 26 total PDF parameters

$$f_v(x) \sim ax^b(1-x)^c \exp\left(dx + ex^2 + f\sqrt{x}\right).$$

- HERAPDF $\sim$ 10 total PDF parameters

$$f_v(x) \sim ax^b(1-x)^c \exp\left(1 + dx + ex^2\right).$$

NNPDF functional form

- NNPDF $\sim$ 259 total PDF parameters

$$f(x) \sim ax^b(1-x)^c\mathrm{NN}(x).$$

- Attempt to minimise figure of merit by varying $(a..f)$.
- Choice of functional form: Parameterization *bias*

NNPDF Strategy

- Minimise bias by choosing extremely flexible functional form
- Each PDF parametrized by a 2-5-3-1 Neural Network
- 259 Free parameters $\rightarrow$ **massively redundant** parameterization
- $b, c \rightarrow$ randomised preprocessing .

Theoretical Predictions

Calculate theoretical predictions for comparison with experimental data.
Evolve to required scale and perform convolution with hard coefficients.

DIS data: cross sections parametrized in terms of **structure functions**:

$$F_i(x, Q^2) = \int \frac{dy}{y} C_i^j(y, \alpha_s(Q^2)) f_j\left(\frac{x}{y}, Q^2\right)$$

Hadron Collider data: perform double convolution over PDFs

$$\sigma_X = \sum_{a,b} \int_0^1 dx_1 dx_2 f_a(x_1, Q^2) f_b(x_2, Q^2) \sigma_{q_a q_b \to X}\left(x_1, x_2, Q^2\right)$$

Hadronic data dependant upon PDFs through parton-parton luminosities:

$$\Phi_{ij}(\tau, M_X^2) = \frac{1}{s} \int_\tau^1 \frac{dx_1}{x_1} f_i(x_1, M_X^2) f_j(\tau/x_1, M_X^2)$$

Minimisation and Stopping in NNPDF

Minimisation by genetic algorithms

Problem: Very large parameter space, $\chi 2$ highly nonlocal.

- Minimisation is challenging.

Solution: Genetic Algorithms (GA)

- Generate mutations of fit parameters.
- Select those mutations that minimise figure of merit.

Dynamical fit stopping by cross-validation

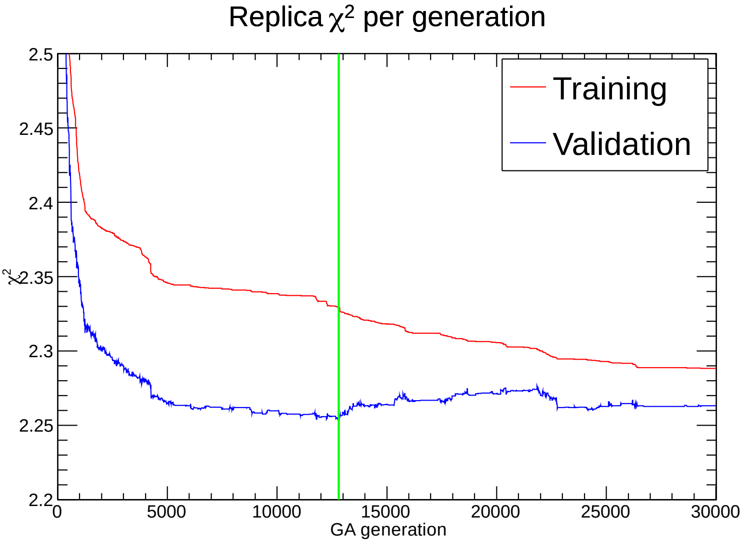
Problem: extremely flexible parameterisations are prone to *overfitting*.

- Fit has so many parameters, the minimum χ^2 corresponds to a fit not only to the data, but also statistical noise.

Solution: dynamical stopping by *Cross Validation*.

- Split the dataset into a training set and a validation set.
- Use the training set for minimisation, monitor the χ^2 to the validation set.
- Stop the fit when the χ^2 to the validation set starts to increase while the χ^2 to training set is still decreasing.

Cross-validation stopping



Standard approach to parton fitting - uncertainties

How to propagate uncertainties from the experimental data to the PDFs?
Standard Approach: Linear propagation of uncertainties by Hessian Method.

- For a set of fit parameters $\{a\}$ define a tolerance in χ^2:

$$\Delta\chi^2(a) \equiv \chi^2(a) - \chi^2(a^{\min}) = \sum_{i,j=1}^{n} H_{ij}(a_i - a_i^{\min})(a_j - a_j^{\min}).$$

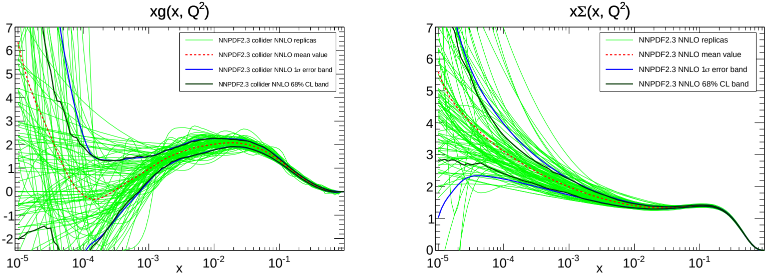
- Determine PDFs on surface of constant $\Delta\chi^2 = T$ in parameter space.
  - Numerical difficulties: Need to use rescaled eigenvectors of H.
- Obtain $2n$ PDF sets $S_i^{\pm}$, where n is the number of free parameters in the fit.

- Uncertainty in an observable $\mathcal{O}$ given by:

$$\mathrm{Var}[\mathcal{O}] = \frac{1}{2}\sum_{i=0}^{n}(\mathcal{O}[S_i^+] - \mathcal{O}[S_i^-])^2.$$

Monte Carlo uncertainty determination

- Form an ensemble of N artificial data 'replicas' by importance sampling the original data set.
- The ensemble of artificial data replicas forms a representation of the probability distribution in data.
- Perform a separate fit to each data replica, obtain an ensemble of PDF replicas.



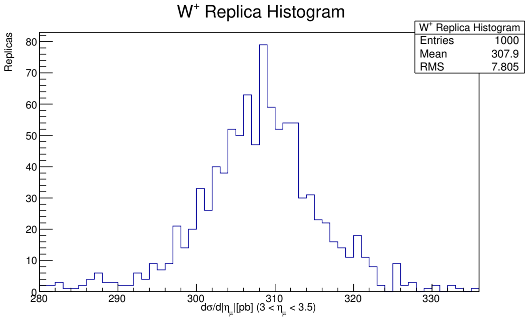
NNPDF User Guide

Central value predictions

$$\langle \mathcal{O} \rangle = \frac{1}{N} \sum_{k=1}^{N} \mathcal{O}[f_k] \, .$$

Uncertainties

$$\mathrm{Var}[\mathcal{O}] = \frac{1}{N} \sum_{k=1}^{N} \left(\mathcal{O}[f_k] - \langle \mathcal{O} \rangle \right)^2 .$$

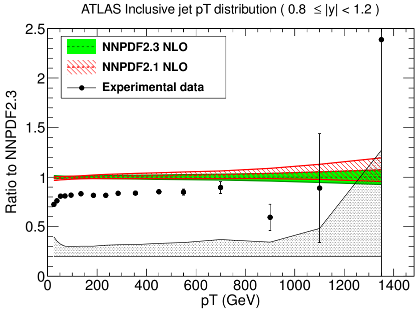


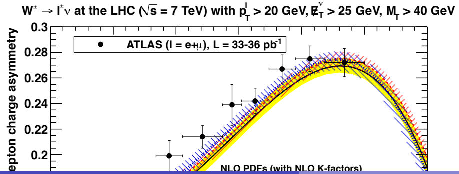
Uncertainties in the PDFs faithfully represent the experimental data.
No tolerance criterion is required.

Parton distributions in the LHC era

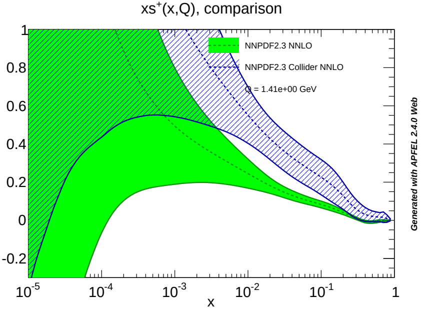
Run-I LHC data has provided a *wealth* of information on parton distributions.

- **NNPDF2.2** (2011) → First (trial) inclusion of LHC data into PDF fits
- **NNPDF2.3** (2012) → First comprehensive analysis including all available pdf-sensitive LHC data.





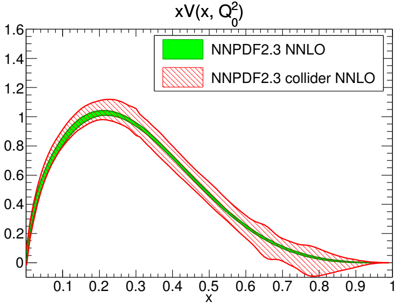
What can the LHC tell us?



LHC Constraints available upon almost all PDF combinations.

Key areas

- Valence distributions
- Light flavour separation
- Strangeness



Parton distributions ready for Run-II

Are our results and methodology good enough?

- Can we take *full advantage* of the available experimental datasets?
- Do we have the computational tools required to include a large (and ever-expanding) LHC dataset?
- How can we best verify that our methodology is robust?

For NNPDF3.0 every point of the methodology has been revisited.

- Fitting code re-written from scratch in modern, modular C++.
- Make use of efficient computational tools to include **all relevant and available LHC data**.
- Thorough examination of fitting procedure through the lens of *closure tests*.

Computation tools for PDF fits

How can we efficiently include LHC data into a full fit?

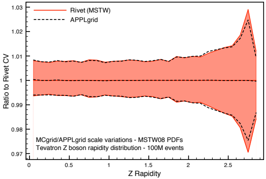
Tools: APPLgrid/FastNLO projects

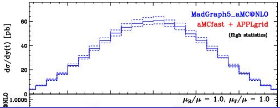
- Precompute and store MC weights on an interpolation grid in x and Q^2:

$$\sigma = \sum_p \sum_l^{N_{\rm sub}} \int_0^1 dx_1 dx_2 \, \hat{\sigma}^{(p)(l)}\left(x_1, x_2, Q^2\right) F^{(l)}\left(x_1, x_2, Q^2\right) \rightarrow$$

$$\sigma = \sum_p \sum_l^{N_{\rm sub}} \sum_{\alpha,\beta}^{N_x} \sum_{\tau}^{N_Q} W^{(p)(l)}_{\alpha\beta\tau} F^{(l)}\left(x_\alpha, x_\beta, Q^2_\tau\right) \qquad (1)$$

- Interpolating interfaces to **automated** NLO codes have recently arisen.





MCgrid [Del Debbio et al.]

SHERPA → APPLgrid/FastNLO

aMCfast [Bertone et al.]

aMC@NLO → APPLgrid

Computation tools for PDF fits → FastKernel

$$\sigma = \sum_{p}^{N_{\rm sub}} \sum_{l}^{N_x} \sum_{\alpha,\beta}^{N_Q} \sum_{\tau} W^{(p)(l)}_{\alpha\beta\tau} \, F^{(l)} \left(x_\alpha, x_\beta, Q^2_\tau \right) \qquad (2)$$

APPLgrid/FastNLO are fast, but **not fast enough** for NNPDF fits.
Idea: Combine weight grids with evolution grids for best performance.

$$f_i(x_\alpha, Q^2_\tau) = \sum_{\beta}^{N_x} \sum_{j}^{N_{\rm pdf}} A^\tau_{\alpha\beta ij} N^0_j(x_\beta) \quad \rightarrow \quad \sigma = \sum_{\alpha,\beta}^{N_x} \sum_{i,j}^{N_{\rm pdf}} \sigma_{\alpha\beta ij} N^0_i(x_\alpha) N^0_j(x_\beta)$$

- Precomputing all Q^2 dependence leads to extremely efficient calculations.

Speed per datapoint of convolution methods

Observable	APPLGRID	FK	optimized FK
W^+ production	1.03 ms	0.41 ms (2.5x)	0.32 ms (3.2x)
Inclusive jet production	2.45 ms	20.1 μs (120x)	6.57 μs (370x)

Dataset for NNPDF3.0

With these computational tools, we could considerably expand the dataset.

HERA-II

- H1 large Q^2 data (NC+CC).
- H1 low Q^2, large y data (NC).
- ZEUS positron beam data (NC+CC).
- HERA combined F_c^2 data.

LHCb

- $Z \to ee$ large-y.

ATLAS

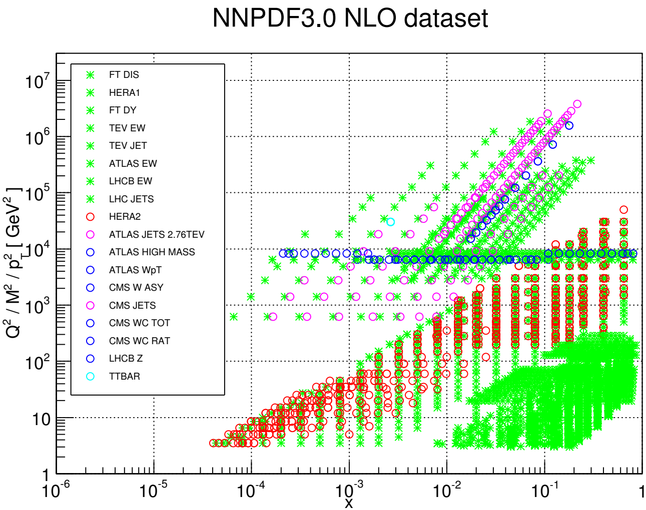
- $\sqrt(s) = 2.76$ TeV inclusive jets.
- High mass Drell-Yan.

CMS

- Inclusive jets at $\sqrt{s} = 7$ TeV.
- Double diff. Drell-Yan.
- W μ charge asymmetry.
- $W + c$.

i.e all relevant data available (with correlation information) at the time

Dataset for NNPDF3.0



4276 total datapoints (**471** from the LHC).

Methodology for NNPDF3.0

How do we ensure that our fit minimises *bias*?
Related studies by Thorne-Watt [arXiv:1205.4024]

Perform a **Closure Test**:

- **Generate artificial pseudo-data based upon a known PDF distribution.**
  Pseudodata generated according to NLO pQCD. Dataset is therefore free of internal inconsistencies.

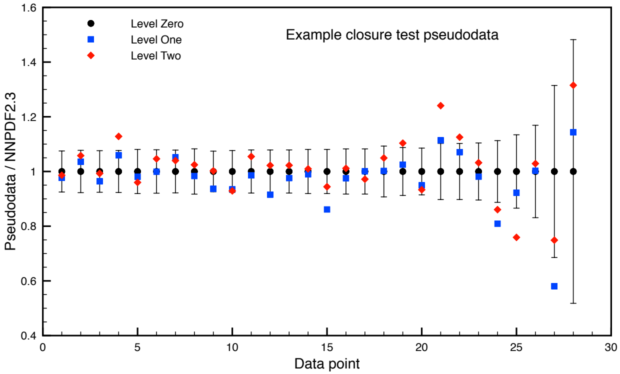
- **Simulate experimental noise in the pseudodata.**
  Data points perturbed according to multi-gaussian distribution defined by the experimental covariance matrix.

- **Perform a full PDF fit to the pseudo-dataset.**
  Closure fit should recover generating PDF up to the level of experimental uncertainty. Reproduction must be (reasonably) independent of generating PDF.

Methodology for NNPDF3.0

Pseudodata categorised into three 'levels' corresponding to the χ^2 of a perfect fit.

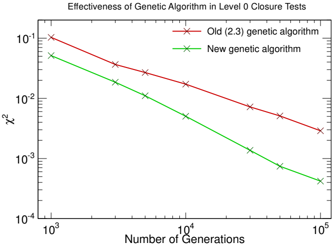
- Level zero - perfect pseudodata with no fluctuations.
- Level one - pseudodata with fluctuations according to σ.
- Level two - pseudodata with fluctuations **and** MC replicas.

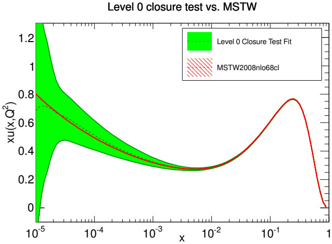


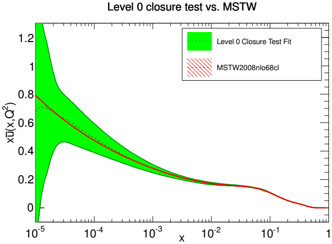
Level zero closure tests

- Level zero tests - $\chi^2 \sim 0$
- Determines fitting flexibility
- Tests *extrapolation uncertainty*

Ideal environment for tuning minimisation.







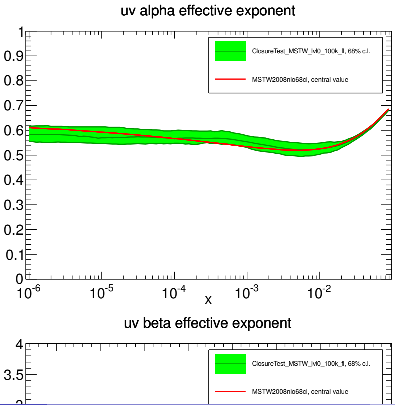
Level zero closure tests - preprocessing

Tests provided important information of the sensitivity of our fits to *preprocessing*

Recall : $f(x) \sim x^{-\alpha}(1-x)^{\beta}P(x).$

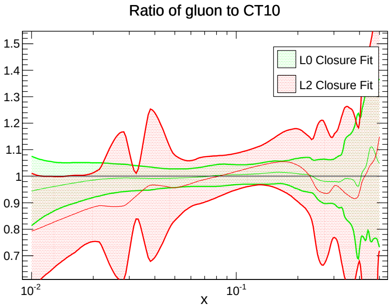
- NNPDF2.3 $\rightarrow$ preprocessing randomised within a fixed range of values.
- NNPDF3.0 $\rightarrow$ range is iteratively improved on a fit-by-fit basis.

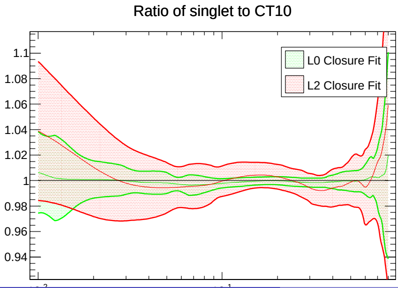
$$\alpha_{\rm eff}(x) = \ln f(x)/\ln 1/x \qquad \beta_{\rm eff}(x) = \ln f(x)/\ln(1-x)$$



Level two closure tests

- Ideal pseudodata $\rightarrow$ simulate noise $\rightarrow$ generate artificial data.

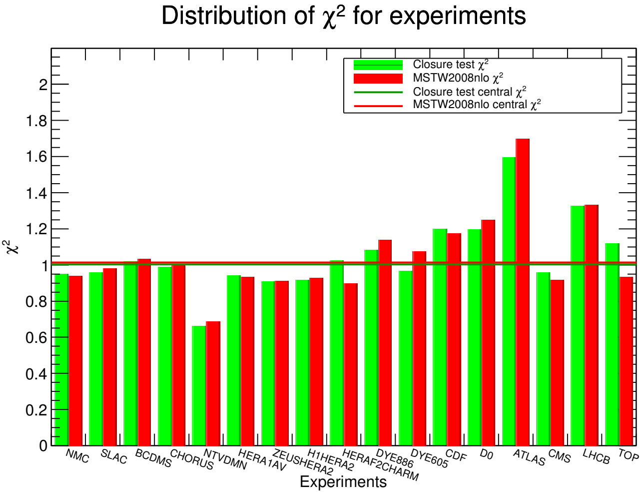




Level two closure tests - χ^2 reproduction

How do we do when it comes to the **data description?**

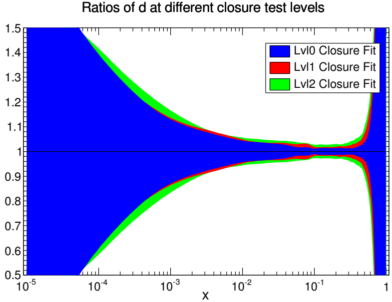
- Compare the χ^2 of CT fit to the χ^2 of the underlying PDF.

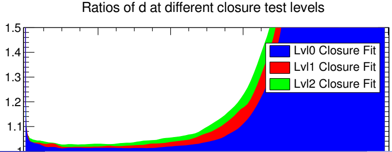


Breakdown of uncertainties

Closure tests can provide information on the breakdown of uncertainties

- Level zero - inter/extrapolation uncertainties
- Level one - functional uncertainties
- Level two - experimental data uncertainties





PDFs for the second run of the LHC

Improvements for NNPDF3.0

Closure tests provided us with plenty to learn from

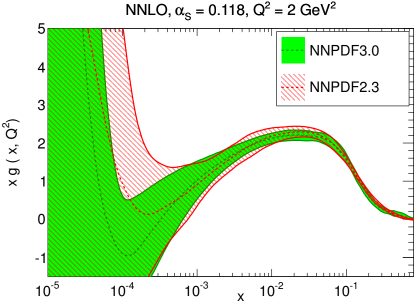
- Newer genetic algorithm - *Faster and more effective fits*.
- Improved preprocessing - *Iterative method minimises bias*
- Simpler fitting procedure

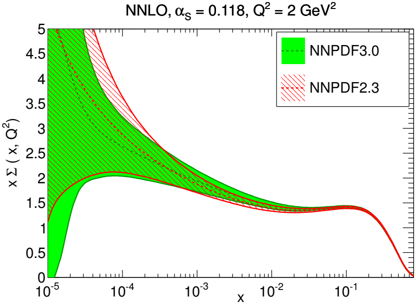
Provided a detailed validation of fitting efficiency, and PDF uncertainties.

Other improvements

- Expanded set of positivity observables.
- EW corrections to LHC DY
- Improved understanding of Jet data at NNLO

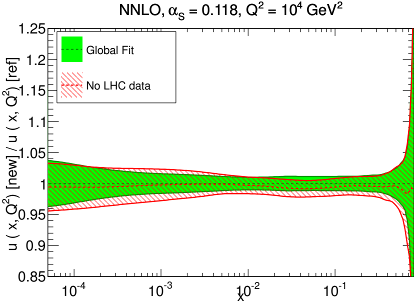
The NNPDF3.0 PDF set

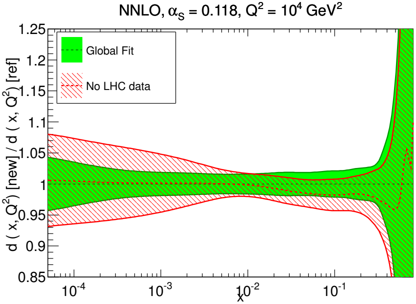




NNLO, α_S = 0.118, Q^2 = 2 GeV^2

Impact of LHC data



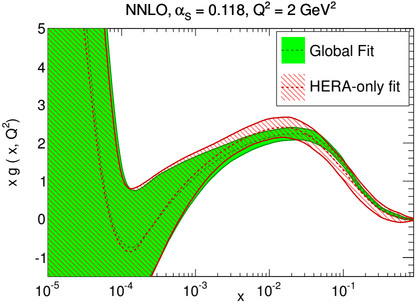


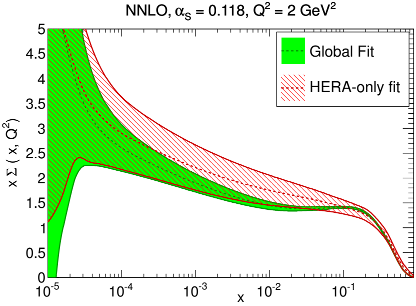
NNLO, α_S = 0.118, Q^2 = 10^4 GeV^2

NNPDF3.0 data description

	NLO			NNLO		
	N_{dat}	χ^2_{exp}	χ^2_{t0}	N_{dat}	χ^2_{exp}	χ^2_{t0}
Total	4276	1.23	1.25	4078	1.29	1.27
ATLAS W, Z 2010	30	1.19	1.25	30	1.23	1.18
ATLAS 7 TeV jets 2010	90	1.07	0.52	9	1.36	0.85
ATLAS 2.76 TeV jets	59	1.29	0.65	3	0.33	0.33
ATLAS high-mass DY	5	2.06	2.84	5	1.45	1.81
ATLAS W p_T	9	1.13	1.28	-	-	-
CMS W electron asy	11	0.87	0.79	11	0.73	0.70
CMS W muon asy	11	1.81	1.80	11	1.72	1.72
CMS jets 2011	133	0.96	0.91	83	1.9	1.07
CMS $W + c$ total	5	0.96	1.30	5	0.84	1.11
CMS $W + c$ ratio	5	2.02	2.02	5	1.77	1.77
CMS 2D DY 2011	88	1.23	1.56	110	1.36	1.59
LHCb W rapidity	10	0.71	0.69	10	0.72	0.63
LHCb Z rapidity	9	1.10	1.34	9	1.59	1.80
$\sigma(t\bar{t})$	6	1.43	1.68	6	0.66	0.61

NNPDF HERA only fits





NNLO, α_S = 0.118, Q^2 = 2 GeV^2

Conclusion

Precise and reliable PDFs are vital for physics at the LHC

- Robust fitting methodology is essential.
- An understanding of the basis of PDF uncertainties is important.
- Fits to a large, global dataset (including LHC data) remain crucial.

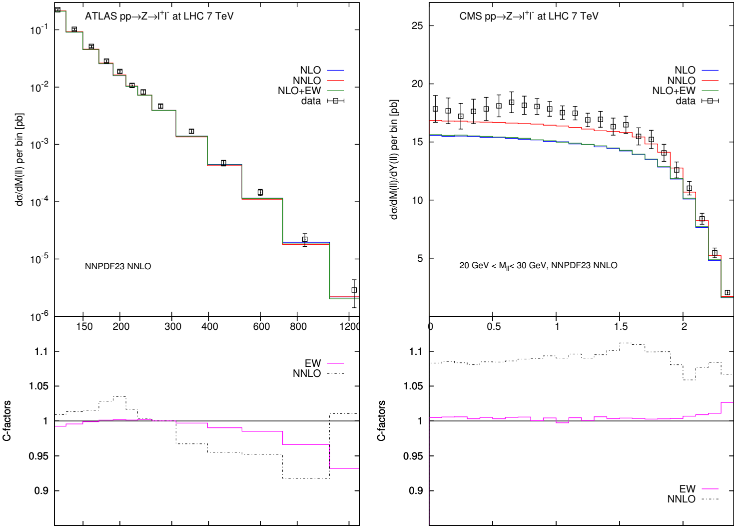
Closure tests enabled detailed validation of methodology in NNPDF3.0.
Fast interfaces, and the FK method allow an enlarged dataset.

LHC data has the potential to answer a lot of open questions in proton structure

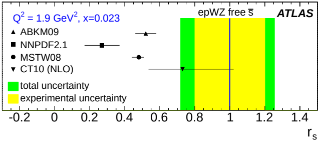
- How important are strange quark distributions?
- Are low-energy, fixed target datasets consistent?
- How reliable are deuteron/nuclear corrections?

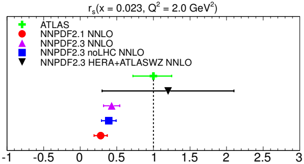
BACKUPS

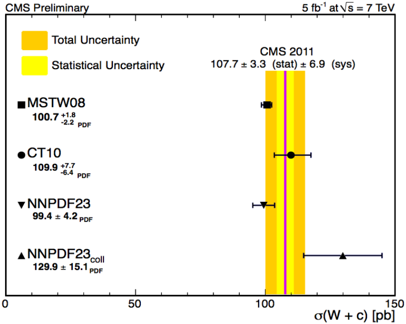
Electroweak corrections in NNPDF3.0



The strange content of the proton.



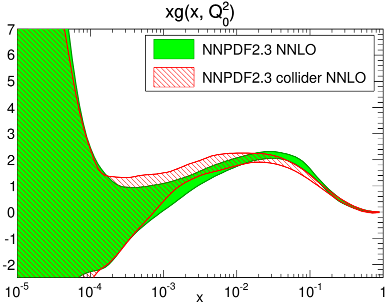


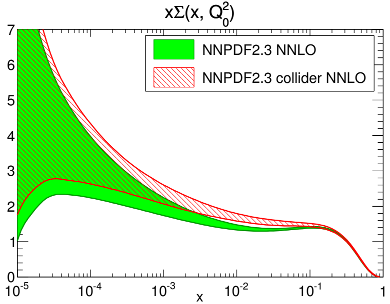


- NNPDF fit to HERA and ATLAS-WZ data finds central value consistent with ATLAS[1] determination of $r_s(x) = (s(x) + \bar{s}(x))/2d(x)$ within a large uncertainty.
- Recent CMS[2] measurement of $W + c$ consistent with strangeness in global fits. Slightly disfavours the larger strange sea in NNPDF2.3 Collider only, but consistent within uncertainties.

[1] arXiv:1203.4051
[2] CMS-SMP-12-002

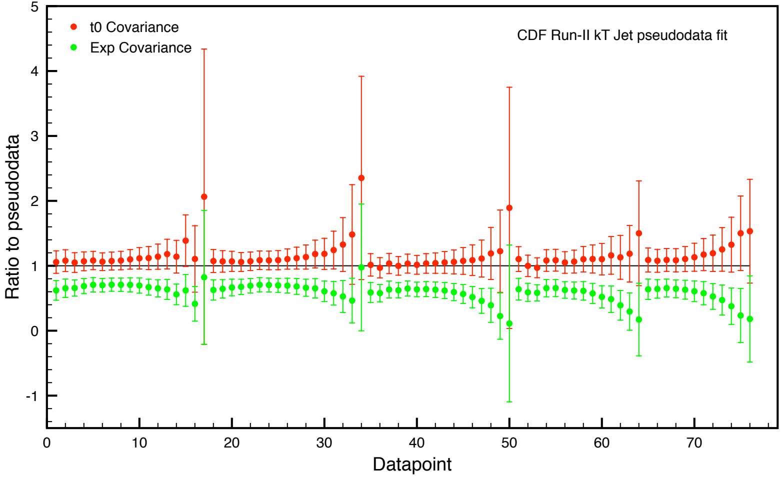
NNPDF2.3 Collider only vs NNPDF2.3





$xT_3(x, Q_0^2)$

t0 method - closure tests



Including new experimental data

How can we add new data to an existing parton set?

- Perform a new fit.
  (Difficult if a fast implementation of the theoretical prediction is not available).
- Reweight existing Monte Carlo parton set. Giele, Keller [hep-ph/9803393]

If the new data is statistically independent of the data in the prior set:

$$\mathcal{P}_{\rm new}(f) = \mathcal{N}_\chi \mathcal{P}(\chi^2|f)\, \mathcal{P}_{\rm old}(f),$$

$$\langle \mathcal{O} \rangle_{\rm new} = \int \mathcal{O}[f]\, \mathcal{P}_{\rm new}(f)\, Df = \frac{1}{N}\sum_{k=1}^{N} w_k \mathcal{O}[f_k].$$

Weights determined by statistical inference

$$w_k = \mathcal{N}_\chi \mathcal{P}(\chi^2|f_k) = \frac{(\chi^2_k)^{(n-1)/2} e^{-\frac{1}{2}\chi^2_k}}{\frac{1}{N}\sum_{k=1}^{N}(\chi^2_k)^{(n-1)/2} e^{-\frac{1}{2}\chi^2_k}}\,.$$

R. D. Ball *et al.* Nucl. Phys. B **849** 112 [arXiv:1012.0836].

Error rescaling parameter

Useful tool for analysing experimental uncertainties. Differentiates between inconsistent and constraining data when N_{eff} is small.

- Rescale uncertainties by a factor α.
- Compute a new weight $w_k(\alpha)$ with these uncertainties.
- Average over all replicas $\to$ probability of rescaling uncertainties by α.

$$\chi^2_{k,\alpha} = \chi^2_k/\alpha^2,$$

$$w_k(\alpha) = (\chi^2_{k,\alpha})^{(n-1)/2} \mathrm{e}^{-\chi^2_{k,\alpha}/2},$$

$$\mathcal{P}(\alpha|\chi^2) \propto \frac{1}{\alpha} \sum_{k=1}^{N} w_k(\alpha).$$