NNPDF3.1

Neural Networks for precision PDF determinations

Stefano Carrazza

on behalf of the NNPDF Collaboration, based on arXiv:1706.00428

ACAT 2017, 21-25 August 2017, University of Washington, Seattle

Theoretical Physics Department, CERN, Geneva.

European Research Council
Established by the European Commission

Introduction to NNPDF

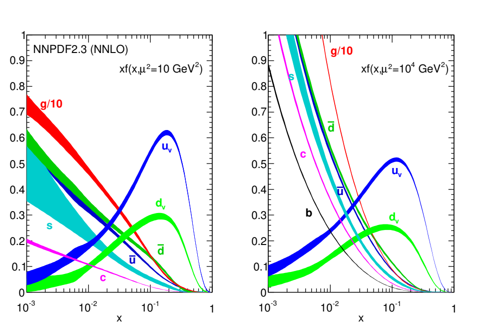
Parton distribution functions

- **PDFs** are **essential** for a **realistic computation** of hadronic particle physics **observable**, σ, thanks to the factorization theorem

$$\sigma = \hat{\sigma} \otimes f,$$

where the elementary **hard cross-section** $\hat{\sigma}$ is convoluted with f the **PDF**.

- **PDFs** are **not calculable**: reflect non-perturbative physics of confinement
- **PDFs** are **extracted** by comparing theoretical predictions to real data



The NNPDF methodology

The NNPDF (Neural Networks PDF) implements the Monte Carlo approach to the determination of a global PDF fit. We propose to:

1. **reduce** all sources of **theoretical bias**:
   - no fixed functional form
   - possibility to reproduce non-Gaussian behavior

   $\Rightarrow$ use Neural Networks instead of polynomials

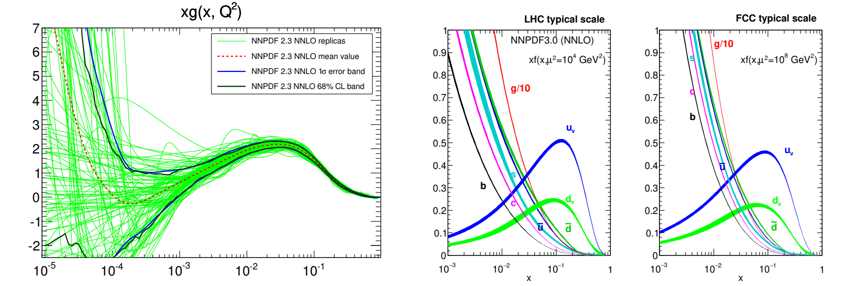
2. provide a sensible estimate of the **uncertainty**:
   - uncertainties from input experimental data
   - minimization inefficiencies and degenerate minima
   - theoretical uncertainties

   $\Rightarrow$ use MC artificial replicas from data, training with a GA minimizer

The NNPDF methodology (an oversimplification)

PDFs are extracted with:

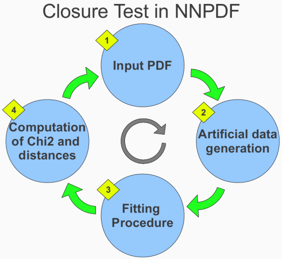
- Generate artificial Monte Carlo replicas of experimental data
- **Neural Network** parametrization (MLP 2-5-3-1, sigmoids-linear)
- Minimization driven by a **genetic algorithm**
- Optimization controlled by **training/validation method**
- **Monte Carlo** representation of results



The NNPDF3.x methodology

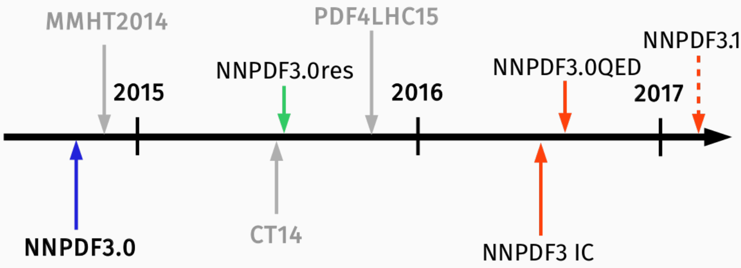
- **Parametrization of NNs for PDF fit in multiple basis:**
  - Perturbative charm: $g, \sum, V, V_3, V_8, T_3, T_8$
  - Fitted charm: $g, \sum, V, V_3, V_8, T_3, T_8, c^+$
- **Look-back as stopping criteria:**
  - minimize training loss function
  - monitor validation loss function during 30k iterations.
  - store and export parametrization at the validation minimum.
- **GA settings are validated through closure tests:**

1. Defined underlying physical law, e.g. input PDF from MMHT, etc.
2. Generate random pseudo-data for the NNPDF3.1 dataset
3. Perform a NNPDF fit
4. Verify compatibility between input and resulting PDF sets.



NNPDF3.1 milestone

NNPDF Timeline



NNPDF3.1 motivation

Why NNPDF3.1? (differences since last 3.0 release):

- the inclusion of new LHC data (2011-2012) with high-precision PDF-sensitive measurements
- the recent progress in NNLO QCD calculations
- introduction of fitted charm, following the NNPDF3 IC studies (NNPDF Collaboration, arXiv:1605.06515)

NNPDF3.1 motivation

Why NNPDF3.1? (differences since last 3.0 release):

- the inclusion of new LHC data (2011-2012) with high-precision PDF-sensitive measurements
- the recent progress in NNLO QCD calculations
- introduction of fitted charm, following the NNPDF3 IC studies (NNPDF Collaboration, arXiv:1605.06515)

NNPDF3.1 results are published in arXiv:1706.00428

NNPDF3.1 code framework

NNPDF3.1 uses and provides public tools for the community:

- Fast theory computation with **APFELgrid**
  (Bertone et al., arXiv:1605.02070)
- Delivery of reduced sets using:
  - **CMC-PDFs**: compress more PDF information in smaller sets of replicas, e.g. 1000 into 100. (SC et al., arXiv:1504.06459)
  - **mc2hessian**: converts MC sets into Hessian sets, suitable for nuisance parameter variation, and Gaussian approximation.
    (SC et al., arXiv:1504.06736)
  - **SMPDF**: identifies the most relevant replicas for a given process.
    (SC et al., arXiv:1601.00005)

Example: NNPDF3.1 fast theory computation

In NNPDF3.1 theoretical predictions are stored in **APFELgrid** tables.

APFELgrid (Bertone et al., arXiv:1605.02070) converts interpolated weight tables provided by APPLgrid in an efficient format for PDF fitting, *e.g.*

$$\sigma = \sum_{i,j}^{n_f} \sum_{\alpha,\beta}^{n_x} W_{ij\alpha\beta} f_i(x_\alpha, Q_0^2) f_j(x_\beta, Q_0^2)$$

where grids are pre-convoluted with PDF evolution kernels from **APFEL**.
(Bertone et al., arXiv:1310.1394)

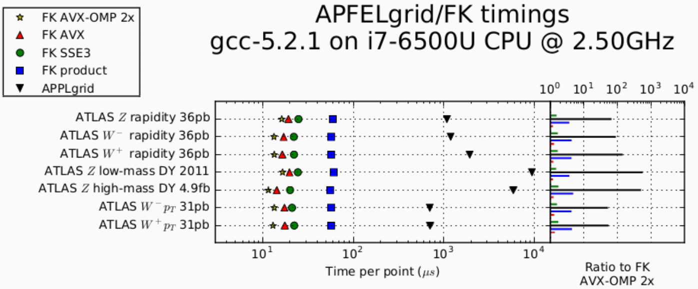
Example: NNPDF3.1 fast theory computation

In NNPDF3.1 theoretical predictions are stored in **APFELgrid** tables.

APFELgrid (Bertone et al., arXiv:1605.02070) converts interpolated weight tables provided by APPLgrid in an efficient format for PDF fitting, *e.g.*

$$\sigma = \sum_{i,j}^{n_f} \sum_{\alpha,\beta}^{n_x} W_{ij\alpha\beta} f_i(x_\alpha, Q_0^2) f_j(x_\beta, Q_0^2)$$

where grids are pre-convoluted with PDF evolution kernels from **APFEL**.
(Bertone et al., arXiv:1310.1394)



Public code: https://github.com/nhartland/APFELgrid

NNPDF3.1 settings

NNPDF3.1 new datasets

List of new data on top of the NNPDF3.0 datasets:

Measurement	Date	Constrain
ATLAS inclusive W, Z rap. at 7 TeV	2011	strangeness
ATLAS inclusive jets 7 TeV	2011	large-x gluon
ATLAS low-mass DY at 7 TeV	2010+2011	small-x quarks
ATLAS double diff. Z p_T 7,8 TeV	2011+2012	medium-x quarks/gluon
CMS 2.76 TeV jets Ratio	2012	medium/large-x gluon
CMS low+high mass DY at 8 TeV	2012	small/large-x quarks
CMS double diff. Z p_T at 8 TeV	2012	medium-x quarks/gluon
CMS W asymmetry at 8 TeV	2012	quark flavor separation
ATLAS and CMS $t\bar{t}$ diff. 8 TeV	2012	large-x gluon
LHCb W, Z rapidity dists. 7,8 TeV	2011+2012	large-x quarks
D0 legacy W asymmetry	Run II	quark flavor separation

- Extra ~300 data points on top of 3.0.

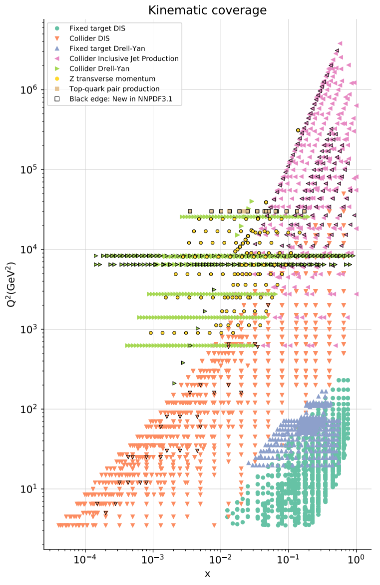
NNPDF3.1 new datasets

The total number of data points for the default PDF determination is

- 4175 at LO, 4295 at NLO and 4285 at NNLO.

Theoretical predictions:

- DIS and FTDY computed with APFEL.
- Hadronic data included using APPLgrid/FastNLO interfaced to MCFM/aMC@NLO/NLOjet++, supplemented by bin-by-bin NNLO/NLO K-factors obtained separately for each specific process.



NNPDF3.1 NNLO treatment

In 3.1 the NNLO correction are applied through K-factors.

The NNLO treatment in 3.1:

- NNLO top pair production (total and differential):
  - Czakon, Heines, Mitov, arXiv:1511.00549
  - Czakon, Hartland, Mitov, Nocera, Rojo, arXiv:1611.08609
- Z p_T double differential distributions:
  - Boughezal et al., e.g arXiv:1602.08140
  - Gehrmann-De Ridder et al., arXiv:1605.04295
  - adding extra uncorrelated theory errors for K-factors (arXiv:1704.00471)
- Inclusive jet cross-section:
  - adding extra correlated theory uncertainties (scale variations)
  - Currie et al. arXiv:1611.01460

NNPDF3.1 results and phenomenology

NNPDF3.1 fit quality

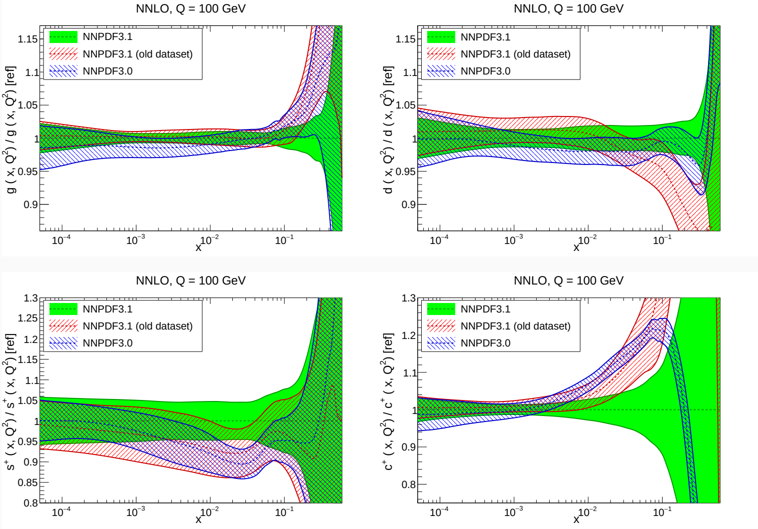
Summary of NNPDF3.1 fit quality in terms of χ^2/N_{dof} :

	NNLO		NLO	
	Fitted Charm	Pert. Charm	Fitted Charm	Pert. Charm
HERA	1.16	1.21	1.14	1.15
ATLAS	1.09	1.17	1.36	1.45
CMS	1.06	1.09	1.20	1.2
LHCb	1.47	1.48	1.62	1.77
Total	**1.148**	**1.187**	**1.168**	**1.197**

- Good description of new collider measurements included in NNPDF3.1
- The global PDF analysis with fitted charm provides a slightly better fit quality when compared to perturbative charm fits.

Impact of new data

Perform a fit using the NNPDF3.0 dataset in order to disentangle the effect of new data and methodology improvement (charm PDF):

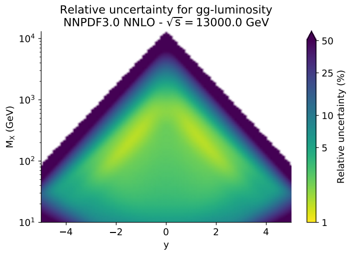


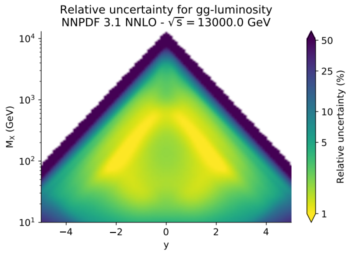
We observe a reduction of PDF uncertainties due to the new data.

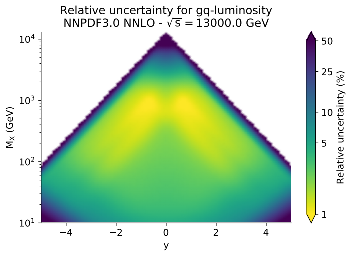
NNPDF3.1 vs NNPDF3.0: 2D luminosity

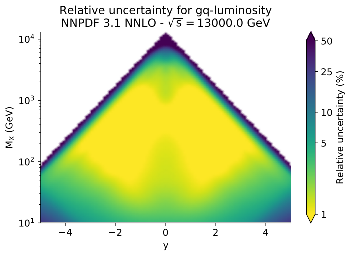
NNPDF3.0 NNLO

NNPDF3.1 NNLO





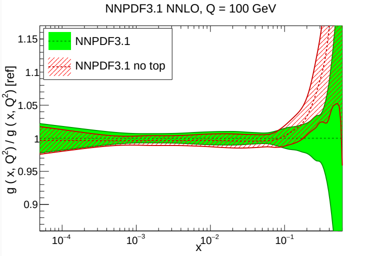


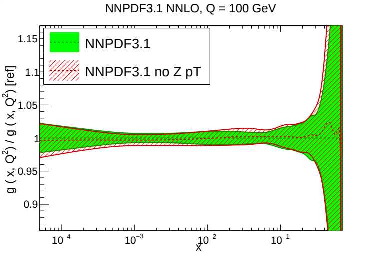


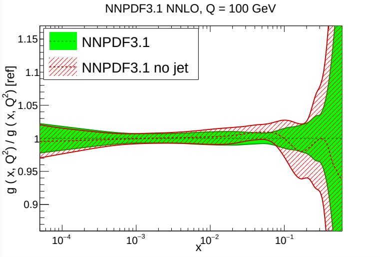
All PDFs exhibit reduced PDF uncertainties at large-x region.

Impact on the gluon

LHC Z p_T, top and inclusive jets data provide direct information on the gluon of NNPDF3.1. The constraints from each process is consistent among them:







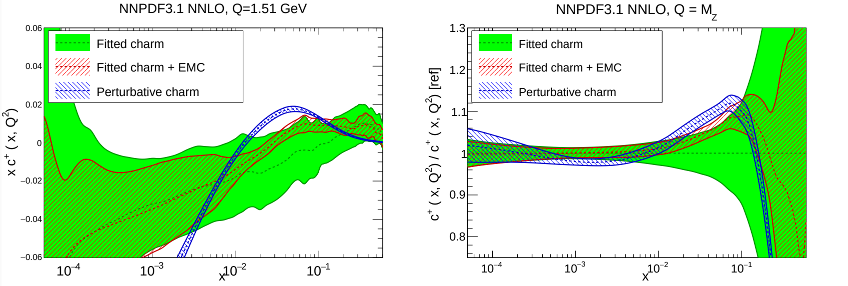
Charm content in NNPDF3.1

In NNPDF3.1 the fitted charm is constrained by the new datasets, in particular top, and LHCb and ATLAS EW boson production.

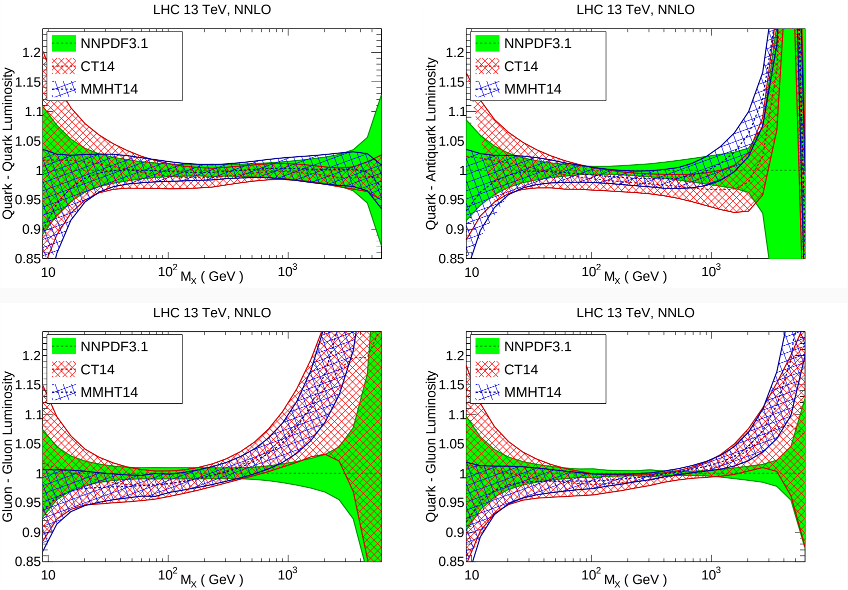
The charm momentum fraction is:

$$C(Q^2) \equiv \int_0^1 dx \left(xc(x, Q^2) + x\bar{c}(x, Q^2) \right)$$

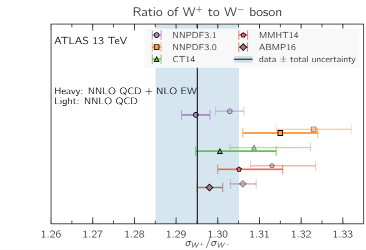
PDF set	$C(m_c)$	$C(m_Z)$
NNPDF3.1	$(0.26 \pm 0.42)\%$	$(3.8 \pm 0.3)\%$
NNPDF3.1 with pert. charm	$(0.176 \pm 0.005)\%$	$(3.73 \pm 0.02)\%$
NNPDF3.1 with EMC data	$(0.34 \pm 0.14)\%$	$(3.8 \pm 0.1)\%$

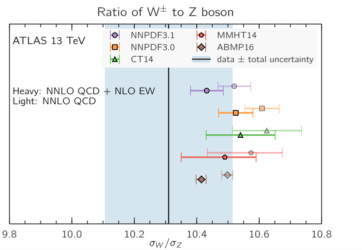


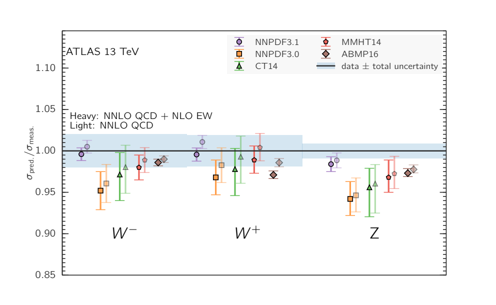
PDF luminosity



W and Z production cross-sections at LHC 13 TeV

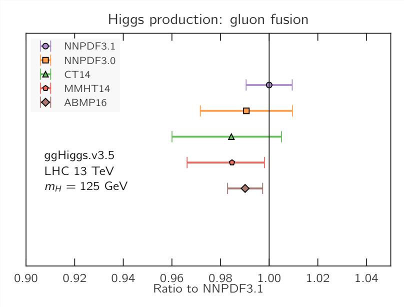


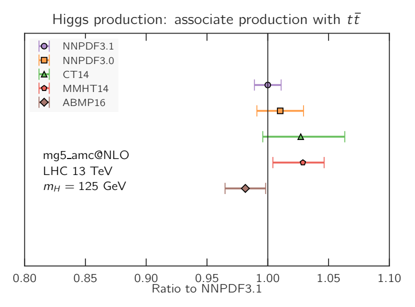


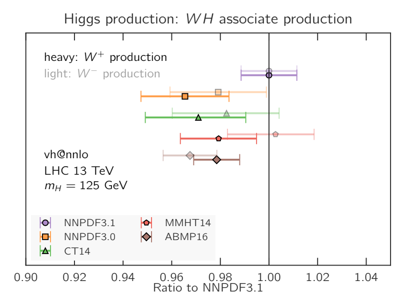


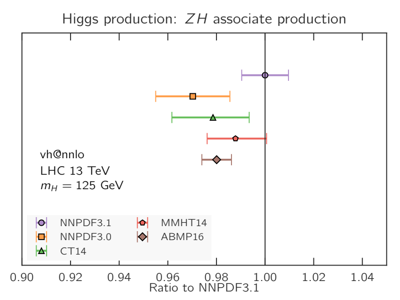
NNPDF3.1 have smaller PDF uncertainties than NNPDF3.0.

Higgs production cross-sections

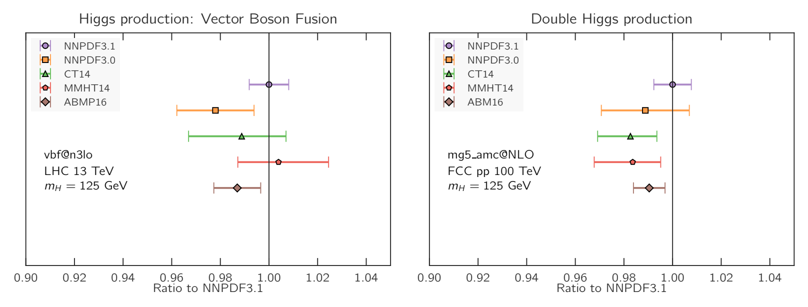








Higgs production cross-sections



- Reduced PDF uncertainties for predictions with 3.1
- Good agreement between 3.1 and 3.0 for gluon-initialed procecess
- Higher cross-sections for quark-initiated process, the new collider data pulls towards higher cross-sections.

Summary and outlook

Summary and outlook

Summary:

- New datasets included, from Tevatron legacy data to precision LHC electroweak production measurements, the $Z\ p_T$ data, and top quark production differential distributions.
- Improved fit quality once the charm PDF is fitted.
- **Reduction of the large-x PDF uncertainties** and an improved quark flavour separation

Delivery:

- NNPDF3.1 is available in LHAPDF since June 2017
  http://pcteserver.mi.infn.it/∼nnpdf/nnpdf31/
- Gallery of plots for NNPDF3.1:
  http://pcteserver.mi.infn.it/∼nnpdf/nnpdf31-gallery/

Thank you.

Data treatment

We construct the **covariance matrix** for each **experiment**. Considering the measurement of two **observables** O_I and O_J we define

$$\mathrm{cov}_{ij} = O_{I,i} O_{J,j} \left(\sum_{l=1}^{N_c} \sigma_{i,l}\sigma_{j,l} + \sum_{n=1}^{N_a} \sigma_{i,n}\sigma_{j,n} + \sum_{n=1}^{N_r} \sigma_{i,n}\sigma_{j,n} + \delta_{ij}\sigma_{i,s}^2 \right) ,$$

where i and j run over the experimental points, and the various uncertainties given as relative values, are:

- $\sigma_{i,l}$, the N_c correlated systematic uncertainties,
- $\sigma_{i,n}$ the N_a absolute and N_r relative normalization uncertainties,
- $\sigma_{i,s}$, the statistical uncertainty.

We then **generate** N_{rep} **artificial replicas** of the **original data points** by shifting with a **multi-Gaussian distribution** defined as

$$O_{I,i}^{(\mathrm{art})(k)} = O_{I,i} \left(1 + \sum_{l=1}^{N_c} r_{i,l}^{(k)} \sigma_{i,l} + r_i^{(k)} \sigma_{i,s} \right) \prod_{n=1}^{N_a} \left(1 + r_{i,n}^{(k)} \sigma_{i,n} \right) \prod_{n=1}^{N_r} \sqrt{1 + r_{i,n}^{(k)} \sigma_{i,n}} \, ,$$

PDF parametrization

NNPDF uses a **multilayer feed-forward** neural network model:

$$\xi_i^{(l)} = g\left(\sum_j w_{ij}^{(l)} \xi_j^{(l-1)} + \theta_i^l\right), \quad g(a) = \frac{1}{1+e^{-a}}$$

with a linear activation function $g(a) = a$ in the last layer and

- $\xi_i^{(l)}$ is the **activation** of the i-th node in the l-th layer
- $w_{ij}^{(l)}$ are the **weights** from that **node** to the node in the **previous layer**
- θ_i^l is the **threshold** for that **node**

In NNPDF we use the **architecture 2-5-3-1** (37 parameters per ANN).

The **weights** and **threshold** are the parameters in the fit which are changed during the **Genetic Algorithm minimization**.

The minimization algorithm

The **GA algorithm** performs three main steps:

- mutation, evaluation and selection

The **minimization procedure** of each PDF replica is **completely independent** from each other, so the procedure can be **parallelized** on **multiple machines**.

Starting from a large number of mutants, the **goodness** of the fit to data for each mutant is then calculated with

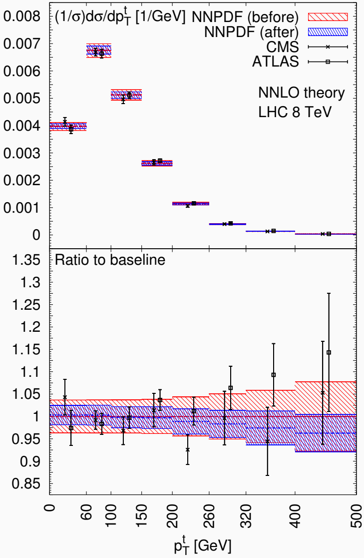
$$\chi^{2(k)} = \frac{1}{N_{\mathrm{dat}}} \sum_{i,j=1}^{N_{\mathrm{dat}}} \left(O_{I,i}^{(\mathrm{art})(k)} - O_{I,i}^{(\mathrm{NN})(k)} \right) \left(\mathrm{cov}_{t_0}\right)_{i,j}^{-1} \left(O_{J,j}^{(\mathrm{art})(k)} - O_{J,j}^{(\mathrm{NN})(k)} \right) ,$$

Stopping criterion for the GA is the cross-validation method:

- **training set:** fitted, the GA minimizes the χ^2_{tr}
- **validation set:** not fitted, stopping based on the increase of the χ^2_{val}

Top data

- NNPDF3.1 includes top quark differential distributions from:
  - ATLAS 8 TeV (arXiv:1511.04716)
  - CMS 8 TeV (arXiv:1505.04480)
- Tension between ATLAS and CMS data $\Rightarrow$ some difficulty in fitting both at the same time.

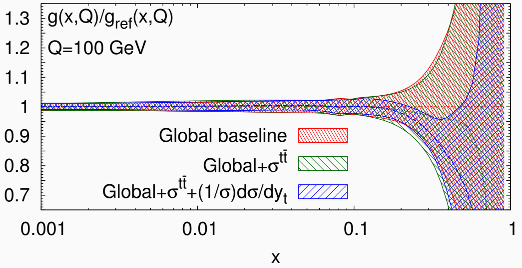


dataset HERA only +	Fit ID 1	2	3	4	5	6	7	8	9	10
ATLAS $d\sigma/dp_T^t$	2.30	2.48	**0.73**	3.16	3.46	2.04	1.34	3.28	4.88	2.89
ATLAS $d\sigma/dy_t$	0.82	1.14	1.21	**1.06**	0.75	1.04	1.31	0.59	0.75	0.74
ATLAS $d\sigma/dy_{t\bar{t}}$	1.12	1.90	2.40	2.83	**0.45**	4.43	1.96	1.88	0.40	1.49
ATLAS $d\sigma/dm_{t\bar{t}}$	4.27	2.93	2.41	2.81	4.33	**1.53**	2.70	2.88	4.37	5.09
ATLAS $(1/\sigma)d\sigma/dp_T^t$	3.47	2.60	3.80	2.92	3.15	3.91	**1.46**	3.31	3.98	4.01
ATLAS $(1/\sigma)d\sigma/dy_t$	1.21	6.07	3.32	5.95	1.34	2.24	4.27	**1.48**	1.58	1.61
ATLAS $(1/\sigma)d\sigma/dy_{t\bar{t}}$	3.11	12.8	5.09	8.34	0.72	7.04	4.95	3.60	**0.53**	2.60
ATLAS $(1/\sigma)d\sigma/dm_{t\bar{t}}$	8.14	3.07	6.53	4.94	5.42	20.5	6.44	5.61	4.40	**3.03**
ATLAS $\sigma_{t\bar{t}}$	3.88	**0.35**	3.38	0.63	1.58	1.29	**0.87**	**0.37**	**0.42**	**0.66**
CMS $d\sigma/dp_T^t$	2.04	2.29	**0.82**	3.29	2.99	1.52	1.44	2.81	4.16	2.32
CMS $d\sigma/dy_t$	3.38	2.48	2.91	**1.75**	3.51	3.47	2.32	3.03	3.48	4.81
CMS $d\sigma/dy_{t\bar{t}}$	1.00	1.58	2.29	1.68	**1.08**	3.05	1.51	1.34	1.07	1.85
CMS $d\sigma/dm_{t\bar{t}}$	3.96	5.85	4.81	4.70	4.23	**1.73**	4.46	4.23	4.71	3.74
CMS $(1/\sigma)d\sigma/dp_T^t$	2.78	4.86	1.78	5.23	4.05	2.84	**1.57**	4.69	5.29	3.40
CMS $(1/\sigma)d\sigma/dy_t$	5.73	3.15	4.10	2.35	5.04	4.88	3.13	**1.94**	4.60	6.71
CMS $(1/\sigma)d\sigma/dy_{t\bar{t}}$	1.68	2.27	2.62	2.11	1.40	3.42	1.78	1.49	**1.20**	1.98
CMS $(1/\sigma)d\sigma/dm_{t\bar{t}}$	5.30	10.3	7.83	8.24	7.06	2.71	7.45	7.41	8.06	**6.26**
CMS $\sigma_{t\bar{t}}$	6.95	**1.04**	6.17	1.59	3.24	2.75	**1.02**	**1.09**	**1.17**	**1.64**

(Czakon et al., arXiv:1611.08609)

Top data

- Only one distribution per experiment should be added at the same time in the global fit.
- Need to determine the combination which maximizes the contains on the large-x gluon.



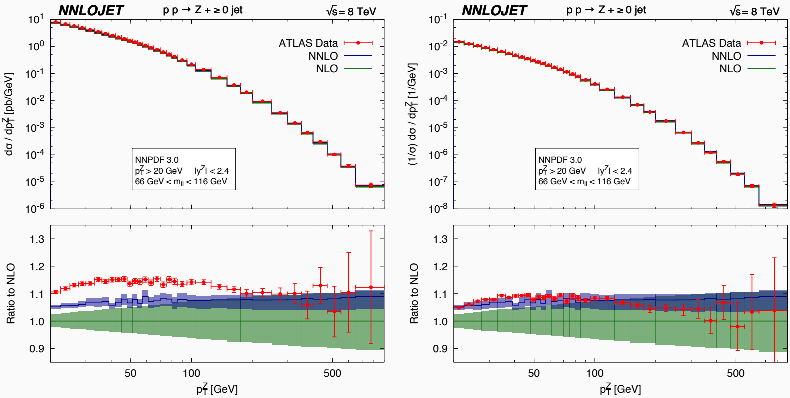
More constraining combination:

- total inclusive cross-section + normalized distribution

(Czakon et al., arXiv:1611.08609)

LHC Z p_T data

- NNPDF3.1 includes Z p_T distributions from:
  - ATLAS 7 TeV and 8 TeV (arXiv:1406.3660, 1512.02192)
  - CMS 8 TeV (arXiv:1504.03511)

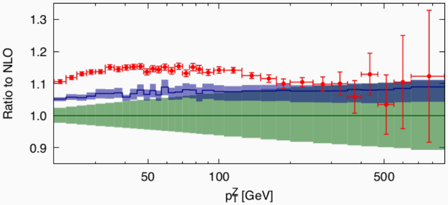


- **NNLO** predictions **improve** agreement between data and theory
- Clear **tension** for the unnormalized distribution

(Gehrmann-De Ridder et al. arXiv:1605.04295)

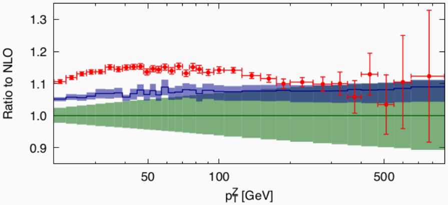
LHC Z p_T data

- The data is hugely dominate by **correlated systematic** uncertainties.
- Experimental precision $< 1\%$ up to $p_T \sim 200$ GeV.



LHC Z p_T data

- The data is hugely dominate by **correlated systematic** uncertainties.
- Experimental precision $< 1\%$ up to $p_T \sim 200$ GeV.



- Interesting case-study to probe current theory-experiment frontier.
- Uncertainties provided by MC simulation are **reliable**?
- How to estimate **unbiased uncertainties** for c-factors?