Progress in NNPDF

Valerio Bertone
CERN

Rencontres de Blois

31 May - 5 June, 2015, Blois

On behalf of the NNPDF Collaboration:
R. D. Ball, VB, S. Carrazza, C. S. Deans, L. Del Debbio, S. Forte, A. Guffanti, Z. Kassabov, N. P. Hartland, J. I. Latorre, P. G Merrild, J. Rojo, L. Rottoli, M. Ubiali

The NNPDF Methodology
Main Ingredients

- **Monte Carlo** generation of data replicas:
  - which reproduces the statistical features of the original dataset.
- **Neural networks** as interpolants:
  - unbiased basis parametrized by a very large set of parameters.
- **Genetic Algorithm** for neural network training:
  - suitable exploration of the space of parameters avoiding to fall into local minima of the χ^2.
- Determination of the best fit by **cross-validation**:
  - proper fitting avoiding over-learning.
- Fits based on a **global dataset**:
  - including collider and fixed-target DIS and hadronic data.

The NNPDF PDF Sets

The Latest Releases

- **The NNPDF2.3 sets** [arXiv:1207.1303]**:**
  - first PDF determination to include **LHC data**:
  - Used as a default stand-alone PDF set in **MadGraph5_aMC@NLO**.
- **The NNPDF2.3 QED sets** [arXiv:1308.0598]**:**
  - PDF sets with QED corrections,
  - first **real fit** of the photon PDF and relative uncertainty,
  - inclusion of measurements of **photon-initiated** processes from the LHC,
  - NNPDF2.3QED LO available as a stand-alone set in Pythia 8
  - and used in the **Monash 2013 tune** of Pythia 8.1 [arXiv:1404.5630].
- **The NNPDF3.0 sets** [arXiv:1410.8849]**:**
  - more LHC (*e.g.* top-pair and W+c) and HERA data (~1000 more points),
  - various theoretical and methodological improvements,
  - the first PDF sets ever based on **closure tests**.
  - Closure tests finally **prove** that our PDFs correctly reproduce the statistical distribution of results expected on the basis of the experimental uncertainties.

The NNPDF3.0 Sets

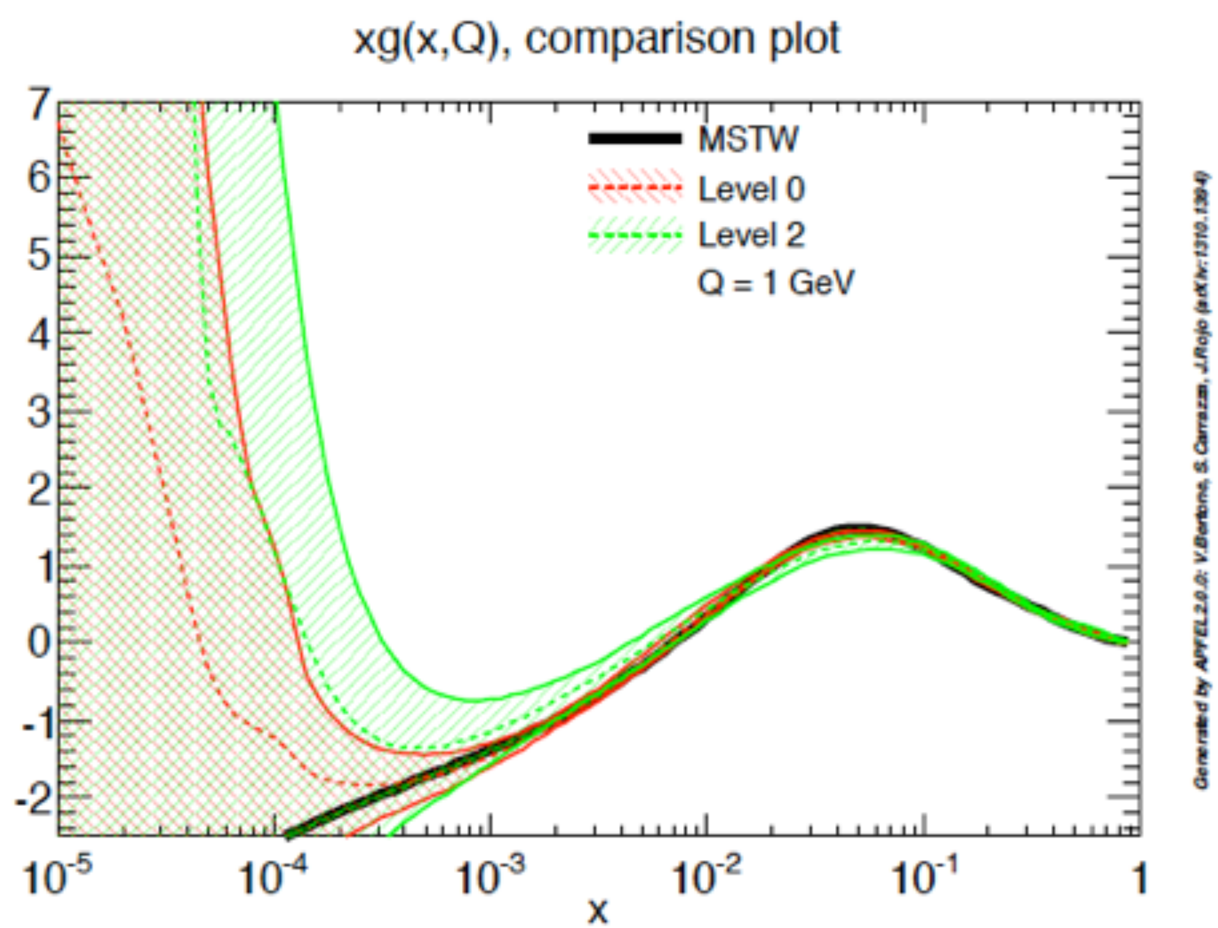
Closure Tests

- How do we know whether our minimization strategy is reliable?
  1) **Assume underlying PDFs** are known (*e.g.* MSTW).
  2) **Generate pseudo-data** with given statistical and correlated systematics.
  3) **Perform a fit** and compare to the "truth".
- Levels of closure tests:
  - **LEVEL 0**:
    - data points equal to the MSTW "true" values,
    - uncertainties assumed equal to the experimental ones,
    - we must find $\boldsymbol{\chi^2 = 0}$ and that **error on predictions tends to zero**.
  - **LEVEL 2**:
    - data points obtained as random fluctuations with given covariance matrix about the "truth",
    - generate Monte Carlo replicas of these data,
    - fit a PDF set to each Monte Carlo replica,
    - we must find $\boldsymbol{\chi^2 = \chi^2_{MSTW}}$ **(~1)** and that MSTW "true PDFs" are **within the 1-σ band**.
- If needed, use the closure tests to **tune** the fitting algorithm.

The NNPDF3.0 Sets

Closure Tests

- Both LEVEL 0 and LEVEL 2 gave a **positive result**.
  - PDF errors in **LEVEL 0** closure test reflect the **functional uncertainty**.
  - PDF errors in **LEVEL 2** closure test reflect the **data uncertainty**.

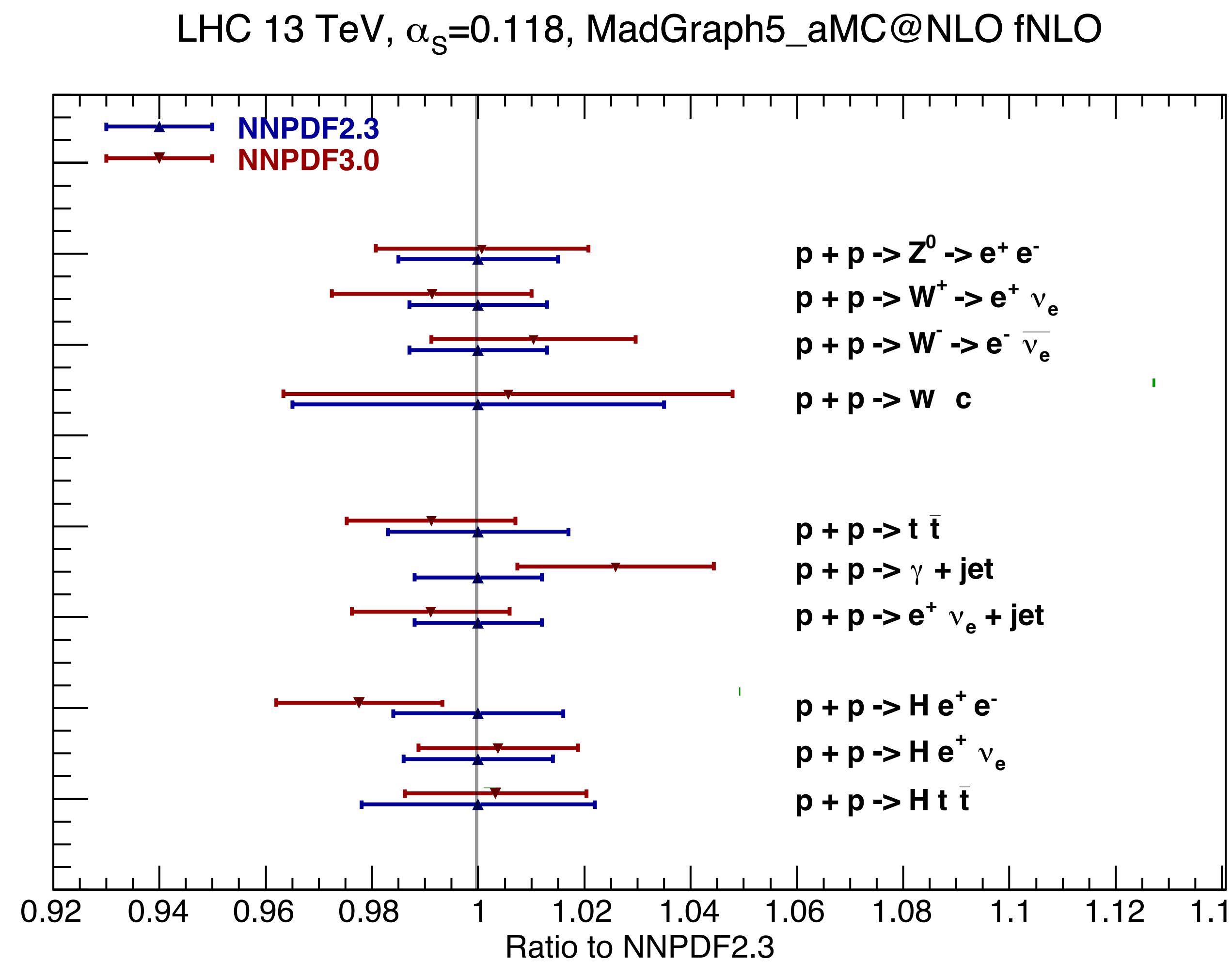


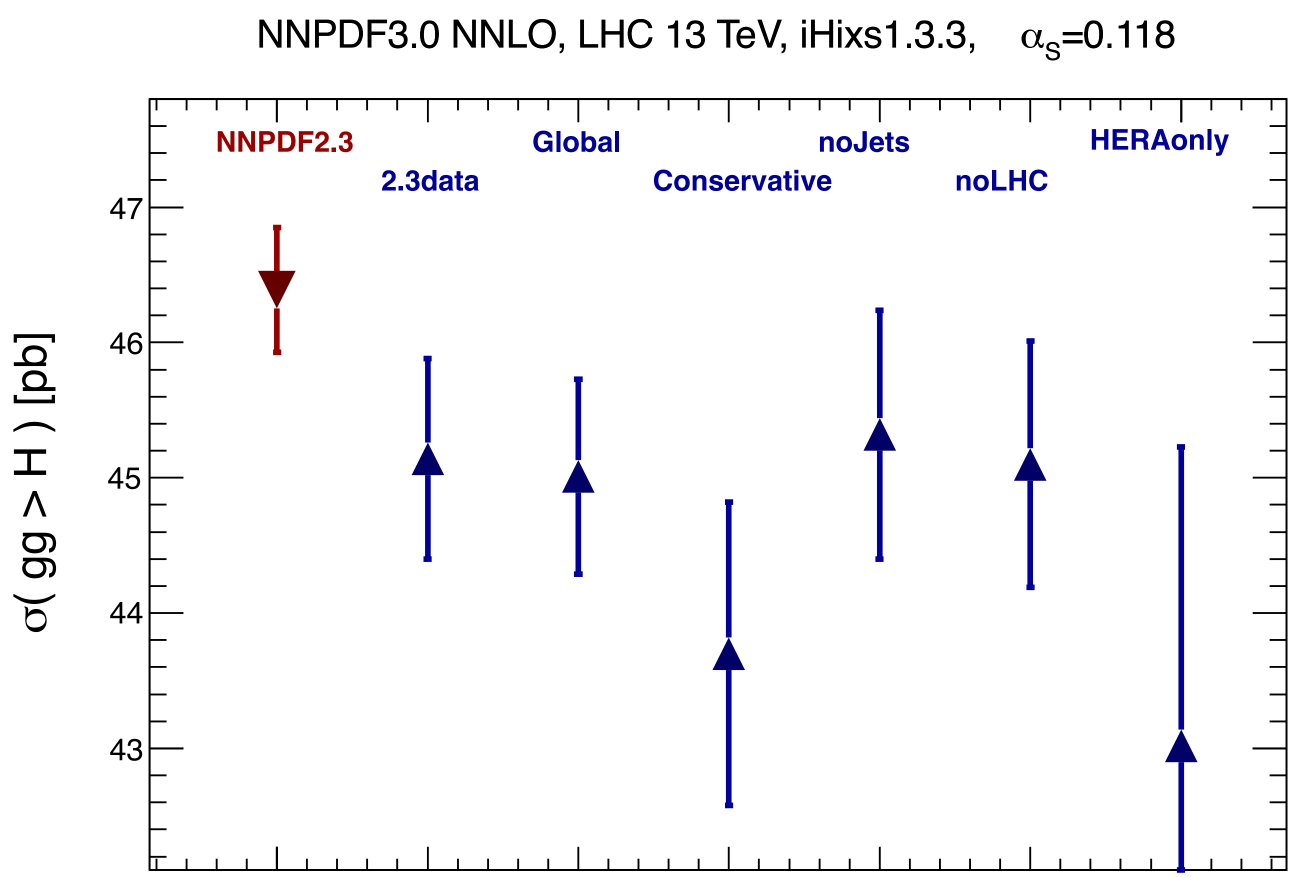
- In the **data region**: data uncertainty ≫ functional uncertainty
- In the **extrapolation region**: data uncertainty ~ functional uncertainty.
- Prove of the reliability of the NNPDF methodology.

The NNPDF3.0 Sets

Phenomenology

- Consistent results between 2.3 and 3.0 for a large number of hadronic cross sections.
- Slightly larger uncertainties in general due to improvements in methodology, better exploration of parameter space.





- Decrease of the Higgs production in gluon fusion.
- Fits to smaller datasets lead to consistent results.
- Again, effect of the improved methodology.

What is new in NNPDF3.0?

- Many new experimental data points:
  - LHC data from ATLAS, CMS and LHCb,
  - combined HERA I data and all the available HERA II data.
- Theoretical improvements:
  - improved jet data treatment in the NNLO fits,
  - inclusion of the EW corrections where needed,
  - FONLL-B in the NLO fits.
- Methodological improvements:
  - tune of the fit based on closure tests,
  - test of the PDF parametrization,
  - improved positivity,
  - rewriting in C++ and optimization of the fitting code.

NNPDF3.0
with Threshold Resummation

Threshold Resummation

- Many LHC calculations supplement fixed-order results with the **resummation of soft threshold logarithms** to all orders:
  - Higgs production, top pair production, high-mass supersymmetry…
- No PDF fit which consistently accounts for soft gluon resummation effects has ever been produced:
  - needed for producing consistent predictions with resummed codes.
- NNPDF is producing PDF sets at **NLO+NLL** and **NNLO+NNLL** in the so-called $\mathcal{N}$-soft approximation:

$$\hat{\sigma}^{(\mathrm{res})}(N,\alpha_s) = \sigma^{(\mathrm{born})}(N,Q^2,\alpha_s)\, C^{(\mathrm{res})}(N,\alpha_s),$$

$$C^{(\mathrm{res})}(N,\alpha_s) \underset{N\to\infty}{\longrightarrow} C^{(N-\mathrm{soft})}(N,\alpha_s) = g_0(\alpha_s)\exp\mathcal{S}(\ln N,\alpha_s)\,,$$

$$\mathcal{S}(\ln N,\alpha_s) = \left[\frac{1}{\alpha_s} g_1(\alpha_s \ln N) + g_2(\alpha_s \ln N) + \alpha_s g_3(\alpha_s \ln N) + \ldots\right],$$

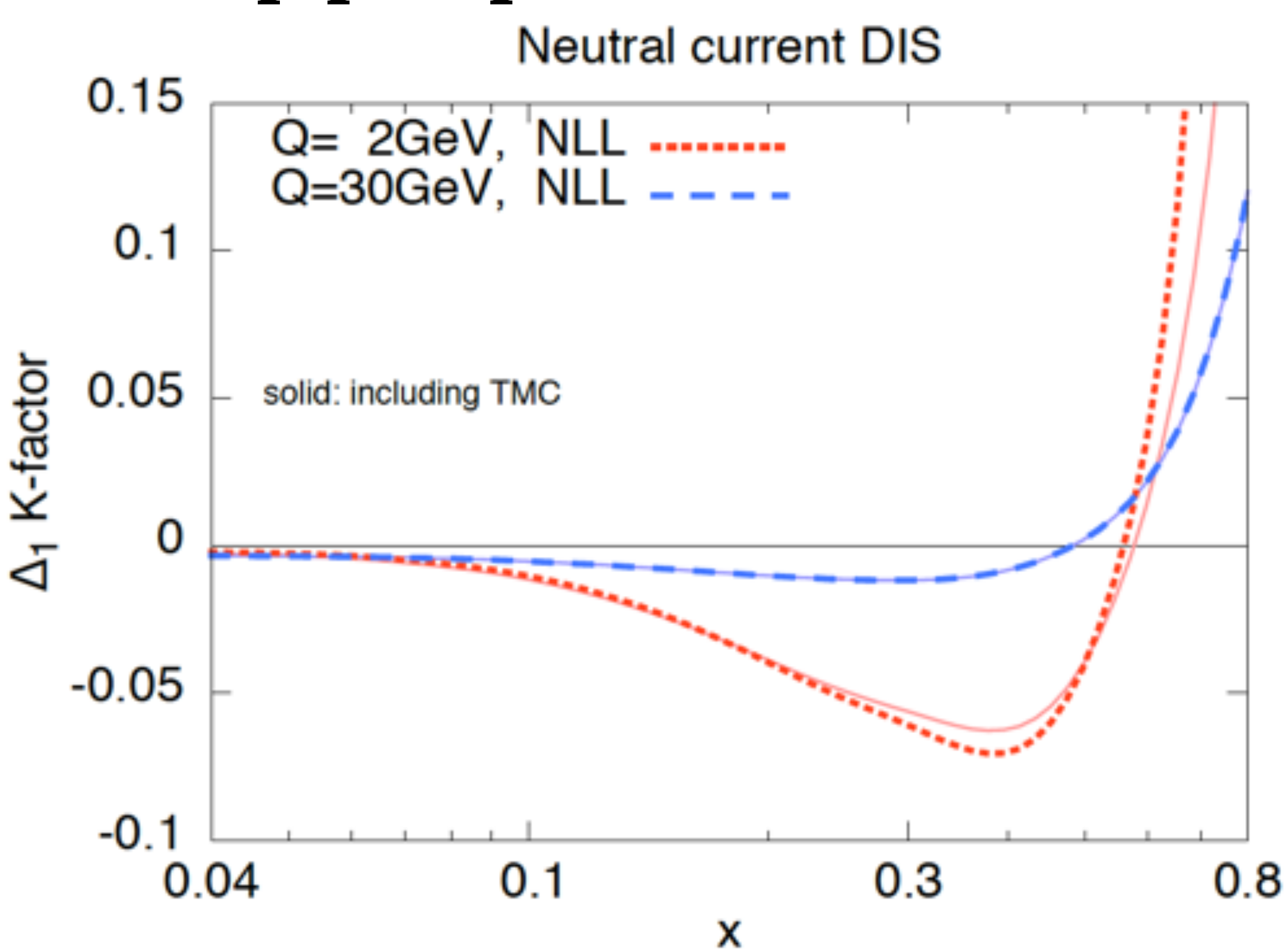
- In the $\overline{\mathrm{MS}}$ factorization scheme the threshold resummation affects only the hard cross sections with no effect on the DGLAP evolution.

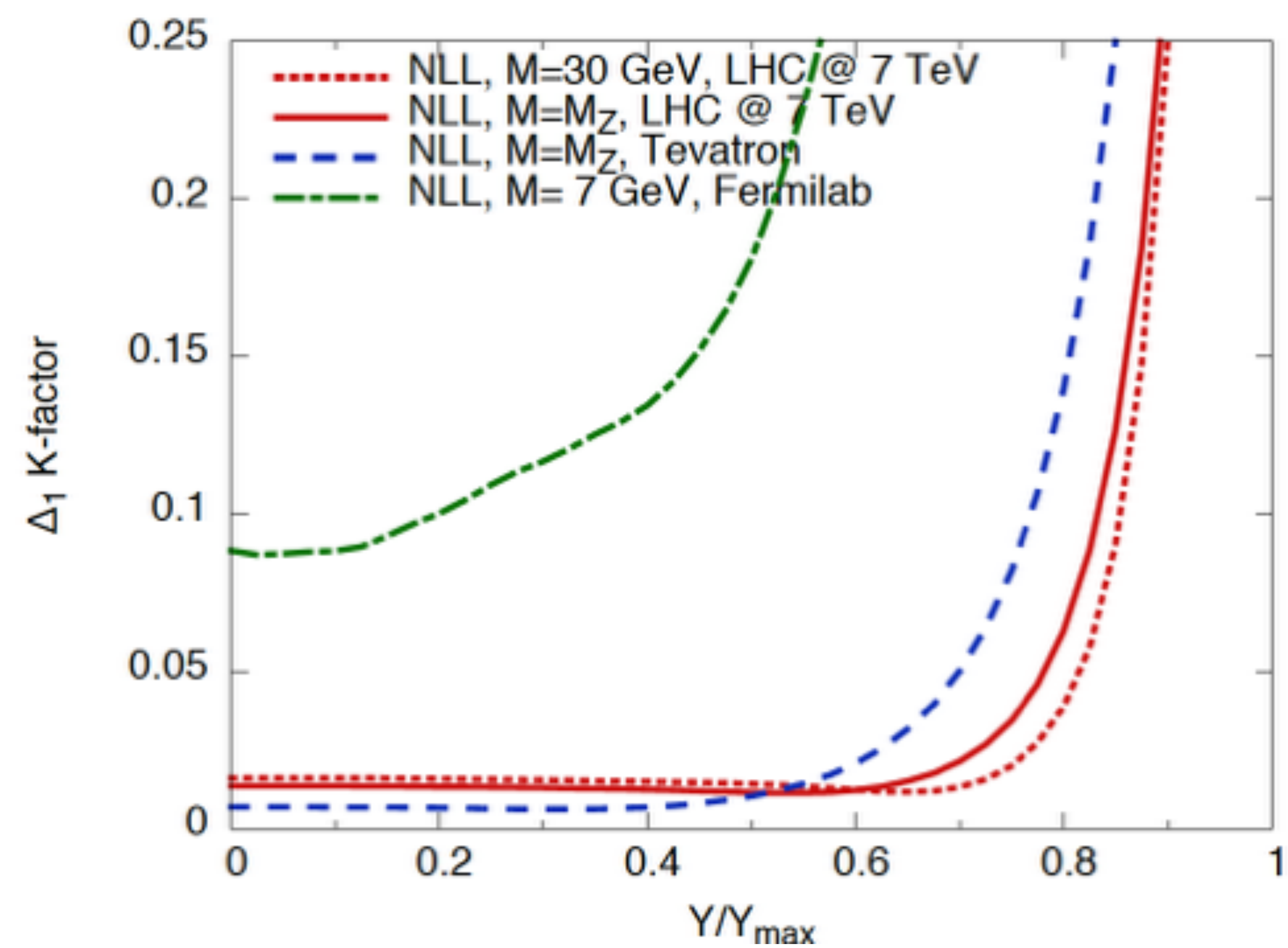
Threshold Resummation

- Soft gluon resummation effects are included by means of K-factors:

$$\sigma_{\mathrm{N}^j\mathrm{LO}+\mathrm{N}^k\mathrm{LL}} = \sigma_{\mathrm{N}^j\mathrm{LO}} + \sigma_{\mathrm{LO}} \times \Delta_j K_{\mathrm{N}^k\mathrm{LL}}.$$

- K-factors computed for:
  - **DIS** and **neutral current DY** rapidity distributions (TROLL),
  - **top-pair production** inclusive cross sections (Top++).





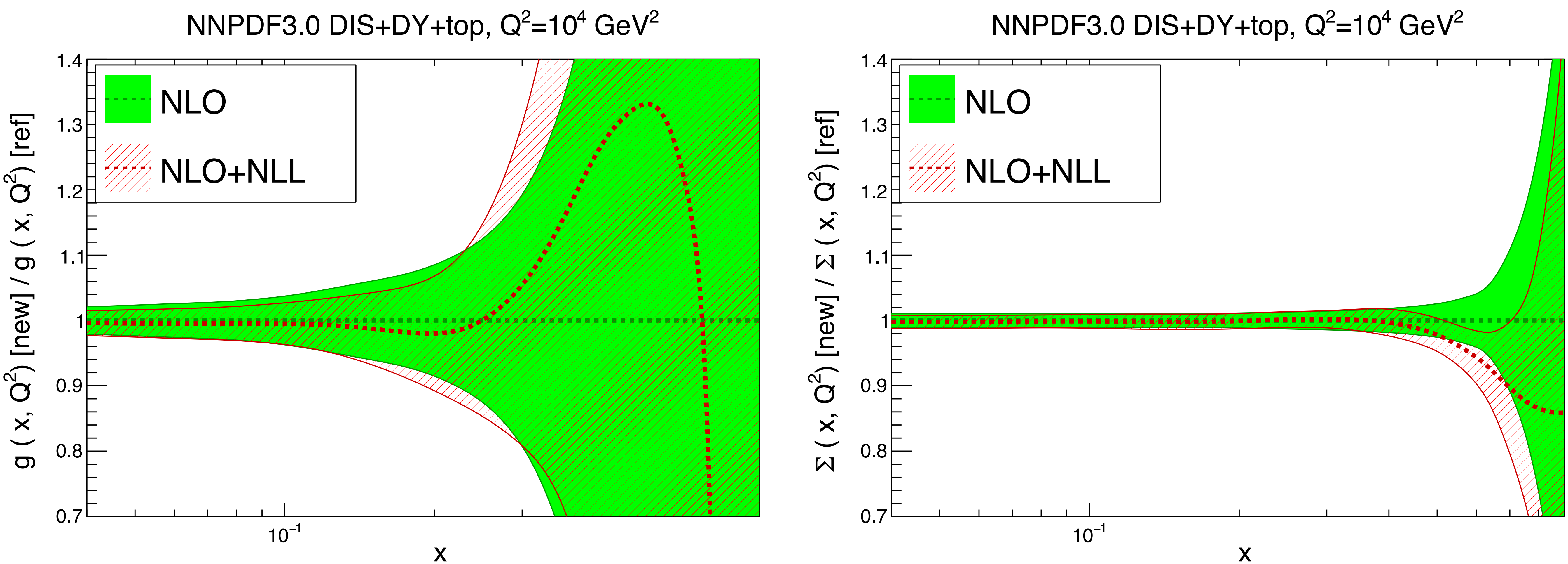
- Substantial effect at large values of x in DIS and forward rapidities in DY.

Threshold Resummation

Experiment	Observable	NNPDF3.0 (N)NLO	NNPDF3.0resLX (N)NLO+(N)NLL
NMC	$\sigma^{\rm NC}_{\rm dis}, F_2^d/F_2^p$	Yes	Yes
BCDMS	F_2^d, F_2^p	Yes	Yes
SLAC	F_2^d, F_2^p	Yes	Yes
CHORUS	$\sigma^{\rm CC}_{\nu N}$	Yes	Yes
NuTeV	$\sigma^{\rm CC,charm}_{\nu N}$	Yes	Yes
HERA-I	$\sigma^{\rm NC}_{\rm dis}, \sigma^{\rm CC}_{\rm dis}$	Yes	Yes
ZEUS HERA-II	$\sigma^{\rm NC}_{\rm dis}, \sigma^{\rm CC}_{\rm dis}$	Yes	Yes
H1 HERA-II	$\sigma^{\rm NC}_{\rm dis}, \sigma^{\rm CC}_{\rm dis}$	Yes	Yes
HERA charm	$\sigma^{\rm NC,charm}_{\rm dis}$	Yes	Yes
DY E866	$\sigma^{\rm NC}_{\rm DY,p}, \sigma^{\rm NC}_{\rm DY,d}/\sigma^{\rm NC}_{\rm DY,p}$	Yes	Yes
DY E605	$\sigma^{\rm NC}_{\rm DY,p}$	Yes	Yes
CDF Z rap	$\sigma^{\rm NC}_{\rm DY,p}$	Yes	Yes
CDF Run-II k_t jets	$\sigma_{\rm jet}$	Yes	No
D0 Z rap	$\sigma^{\rm NC}_{\rm DY,p}$	Yes	Yes
ATLAS Z 2010	$\sigma^{\rm NC}_{\rm DY,p}$	Yes	Yes
ATLAS W 2010	$\sigma^{\rm CC}_{\rm DY,p}$	Yes	No
ATLAS 7 TeV jets 2010	$\sigma_{\rm jet}$	Yes	No
ATLAS 2.76 TeV jets	$\sigma_{\rm jet}$	Yes	No
ATLAS high-mass DY	$\sigma^{\rm NC}_{\rm DY,p}$	Yes	Yes
ATLAS W p_T	$\sigma^{\rm CC}_{\rm DY,p}$	Yes	No
CMS W electron asy	$\sigma^{\rm CC}_{\rm DY,p}$	Yes	No
CMS W muon asy	$\sigma^{\rm CC}_{\rm DY,p}$	Yes	No
CMS jets 2011	$\sigma_{\rm jet}$	Yes	No
CMS $W+c$ total	$\sigma^{\rm NC,charm}_{\rm DY,p}$	Yes	No
CMS 2D DY 2011	$\sigma^{\rm NC}_{\rm DY,p}$	Yes	Yes
LHCb W rapidity	$\sigma^{\rm CC}_{\rm DY,p}$	Yes	No
LHCb Z rapidity	$\sigma^{\rm NC}_{\rm DY,p}$	Yes	Yes
ATLAS CMS top prod	$\sigma(t\bar{t})$	Yes	Yes

Threshold Resummation

- As expected, larger effects found in the large-x region.



- At NLO+NLL, effects of resummation can be large, up to 10% for quarks and 30% for the gluons at large x,

- Relevant effect for the production of new BSM particles.

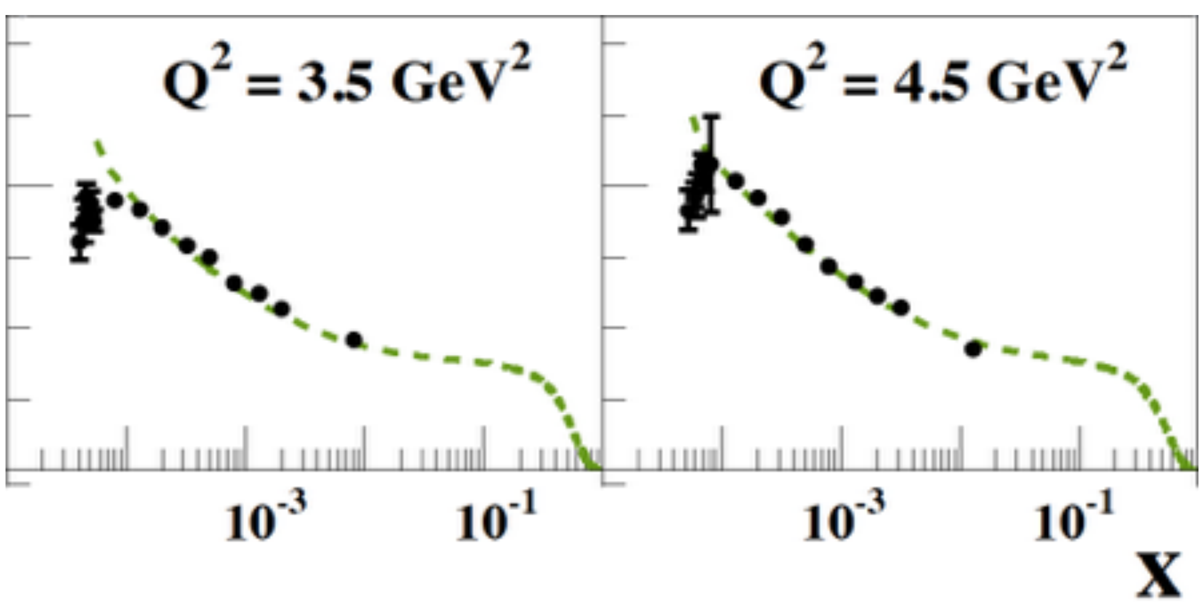
Towards NNPDF3.1

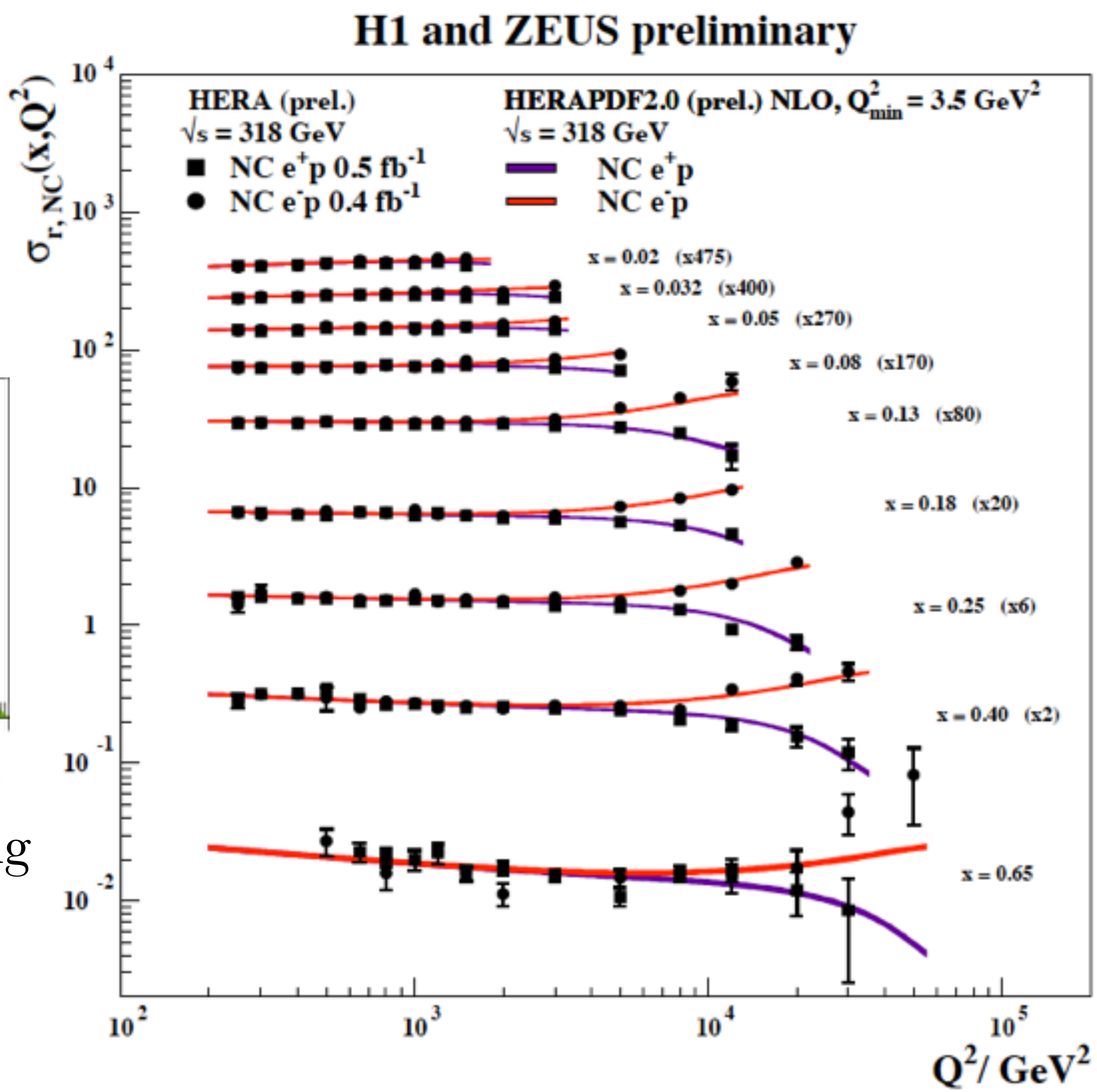
New Experimental Data
The New DIS HERA data

- The new DIS final **combined HERA II** inclusive data set will shortly become public.

From Eram Rizvi's talk at QCD@LHC 2014

- Such data will supersede all the HERA I and HERA II data.
- No big impact expected since all the individual measurement are already included NNPDF3.0.





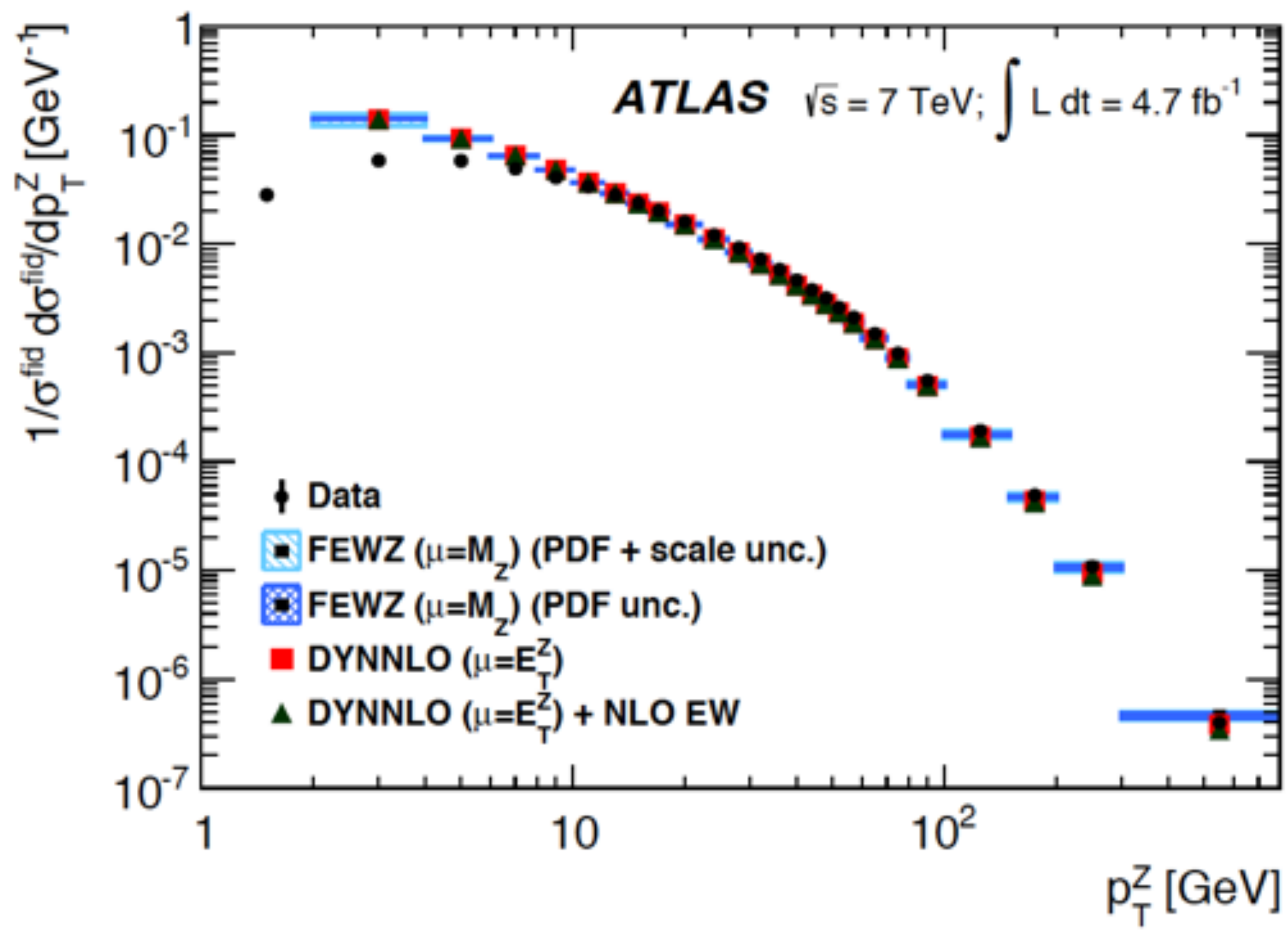
- Low x/Q^2 data harder to describe using fixed-order perturbation theory.
- Need of small-x resummation?

New Experimental Data

New Hadron-Collider Data

ATLAS low mass Drell-Yan dilepton production 7 TeV, 1.6 fb $^{-1}$
ATLAS isolated photon production 7 TeV, 4.6 fb $^{-1}$
ATLAS Z pt distribution 7 TeV, 4.7 fb $^{-1}$
ATLAS inclusive jet production 7 TeV, 5 fb $^{-1}$
ATLAS ttbar differential distributions, 7 TeV, 4.6 fb^{-1}
CMS Drell-Yan double-differential distribution 2012 dataset, 8 TeV
CMS ttbar differential distributions, 8 TeV
CMS pt and rap differential distributions 2012 dataset, 8 TeV
D0 W muon asymmetry final results
D0 W electron asymmetry final results
LHCb W-> mu nu rapidity distributions from 2011
LHCb Z > e+e- rapidity distributions from 2011 - 8 TeV

- Since the NNPDF3.0 sets have been released, many new hadron-collider data have appeared.
  - ATLAS low mass DY $\Rightarrow$ small-x PDFs,
  - ATLAS isolated photon $\Rightarrow$ medium-x gluon,
  - ATLAS Z p_T $\Rightarrow$ medium and large-x PDFs,
    - $p_T > 25$ GeV data included, low-p_T data need resummation (*e.g.* parton shower).

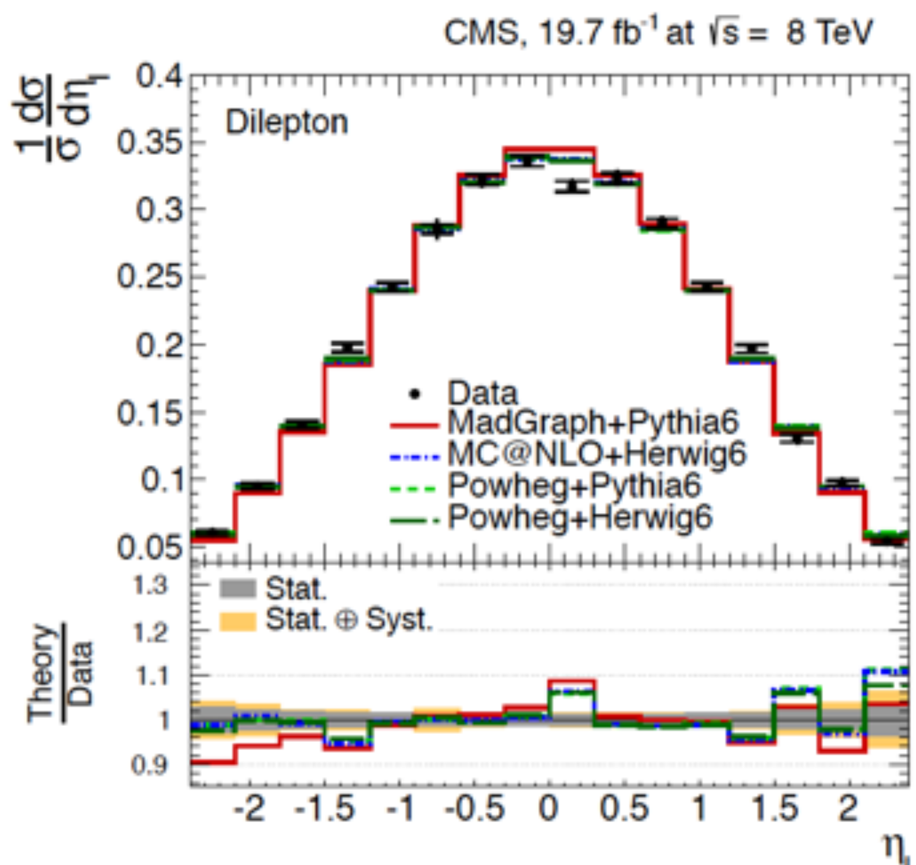


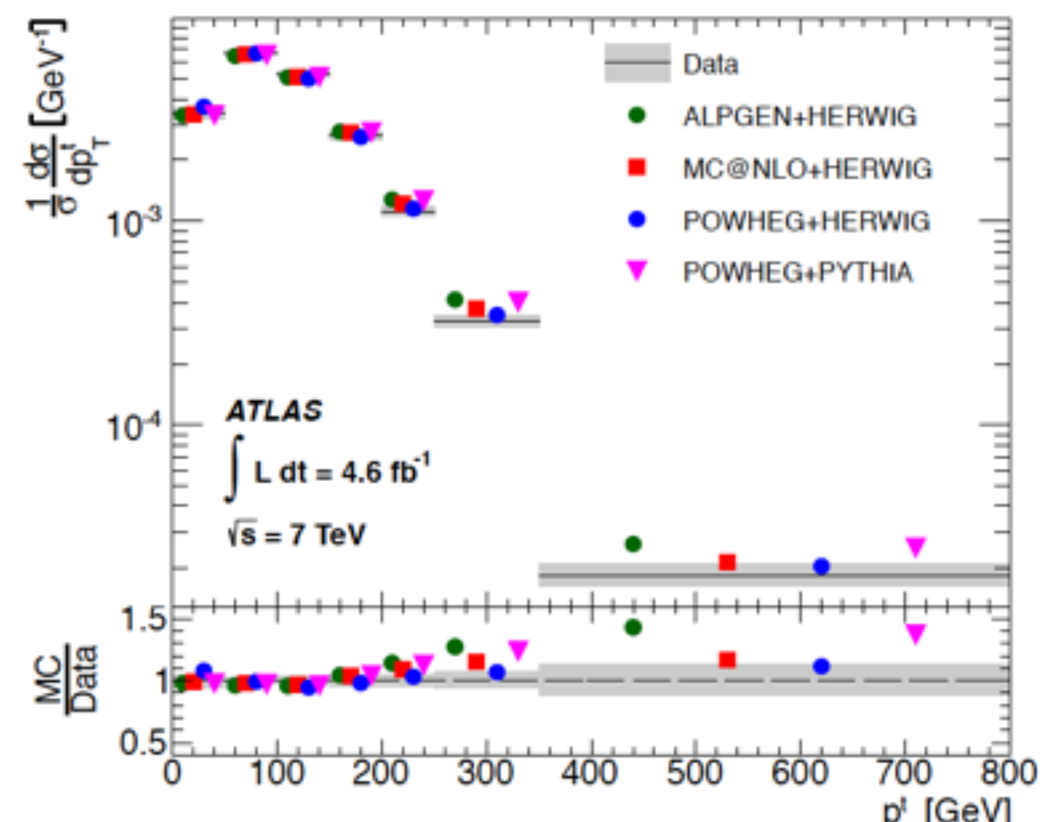
New Experimental Data

New Hadron-Collider Data

ATLAS low mass Drell-Yan dilepton production 7 TeV, 1.6 fb $^{-1}$
ATLAS isolated photon production 7 TeV, 4.6 fb $^{-1}$
ATLAS Z pt distribution 7 TeV, 4.7 fb $^{-1}$
ATLAS inclusive jet production 7 TeV, 5 fb $^{-1}$
ATLAS ttbar differential distributions, 7 TeV, 4.6 fb^{-1}
CMS Drell-Yan double-differential distribution 2012 dataset, 8 TeV
CMS ttbar differential distributions, 8 TeV
CMS pt and rap differential distributions 2012 dataset, 8 TeV
D0 W muon asymmetry final results
D0 W electron asymmetry final results
LHCb W-> mu nu rapidity distributions from 2011
LHCb Z > e+e- rapidity distributions from 2011 - 8 TeV

- ATLAS inclusive jet data ⇒ the large-x gluon,
- ATLAS and CMS top-pair differential distributions ⇒ large-x gluon,
  - exact NNLO corrections with Top++,





- LHCb forward data ⇒ flavor separation at large-x,
- legacy Tevatron W and Z production data,
- …

A New Prediction Engine

- Starting from NNPDF3.1, the **DGLAP evolution** and the **DIS structure functions** will be compute using the **public APFEL code** [arXiv:1310.1394].
  - APFEL has been fully validated against the old internal NNPDF code (FKgenerator).
- Hadronic observables are treated using the **FastKernel** method with APFEL for the DGLAP evolution:
  - combine and store hard cross sections and evolution factors:

$$
\begin{aligned}
\sigma_I &= W_{ij} \otimes f_i(Q) \otimes f_j(Q) \\
&= \underbrace{\left[W_{ij} \otimes \Gamma_{ik}(Q, Q_0) \otimes \Gamma_{jl}(Q, Q_0) \right]}_{\text{Precomputed and stored}} \otimes f_k(Q_0) \otimes f_l(Q_0)
\end{aligned}
$$

A New Prediction Engine

- Starting from NNPDF3.1, the **DGLAP evolution** and the **DIS structure functions** will be compute using the **public APFEL code** [arXiv:1310.1394].
  - APFEL has been fully validated against the old internal NNPDF code (FKgenerator).
- Hadronic observables are treated using the **FastKernel** method with APFEL for the DGLAP evolution:
  - combine and store hard cross sections and evolution factors:

Provided by APFEL

APPLgrid/FastNLO

$$\begin{aligned}\sigma_I &= W_{ij} \otimes f_i(Q) \otimes f_j(Q) \\ &= \underbrace{[W_{ij} \otimes \Gamma_{ik}(Q,Q_0) \otimes \Gamma_{jl}(Q,Q_0)]}_{\text{Precomputed and stored}} \otimes f_k(Q_0) \otimes f_l(Q_0)\end{aligned}$$

  - Observables written in terms of initial scale PDFs $\Rightarrow$ **no need of the evolution** during the fit.
  - Optimized convolution using SSE acceleration.

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)

A New Prediction Engine

- Starting from NNPDF3.1, the **DGLAP evolution** and the **DIS structure functions** will be compute using the **public APFEL code** [arXiv:1310.1394].
  - APFEL has been fully validated against the old internal NNPDF code (FKgenerator).
- Hadronic observables are treated using the **FastKernel** method with APFEL for the DGLAP evolution:
  - combine and store hard cross sections and evolution factors:

Provided by APFEL

APPLgrid/FastNLO

$$\sigma_I = W_{ij} \otimes f_i(Q) \otimes f_j(Q)$$
$$= \underbrace{[W_{ij} \otimes \Gamma_{ik}(Q,Q_0) \otimes \Gamma_{jl}(Q,Q_0)]}_{\text{Precomputed and stored}} \otimes f_k(Q_0) \otimes f_l(Q_0)$$

  - Observables written in terms of initial scale PDFs ⇒ **no need of the evolution** during the fit.
  - Optimized convolution using SSE acceleration.

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)

  - Huge **improvement in speed** as compared to the APPLgrid standard convolution.

A New Prediction Engine

- The new prediction engine, in conjunction with the rewriting in c++ of the fitting code (since NNPDF3.0), enormously **facilitates** the fitting procedure:
  - global fits (~4000 data points) completed within a **few hours**,
  - possibility to easily change the parameters of the fit (heavy quark masses, $\alpha_s(M_Z)$, perturbative order, etc.)
- The use of APFEL will give immediate access to a number of functionalities that can and will be employed for future PDF fits:
  - structure functions and DGLAP evolution in the presence of $\overline{\mathrm{MS}}$ running masses $\Rightarrow$ better **perturbative convergence** and possible determination of the heavy quark masses,
  - QED corrections to the DGLAP equations $\Rightarrow$ fit of the **photon PDF**,
  - renormalization and factorization scale variations $\Rightarrow$ estimate of the **theory error**

Conclusions and Outlook

- Inclusion of the threshold resummation effects in NNPDF3.0.
- Towards the NNPDF3.1 release:
  - Inclusion of more data:
    - final HERA II combined data,
    - most recent LHC and Tevatron data.
  - A new optimized prediction engine:
    - extremely fast convolution based on APFEL that makes fits much faster.
  - More theoretical improvements in the pipeline:
    - fit of the charm PDFs (intrinsic charm),
    - $\overline{\text{MS}}$ running masses,
    - QED corrections,
    - scale variations,
    - small-x resummation,
    - inclusion of parton shower effects,
    - …

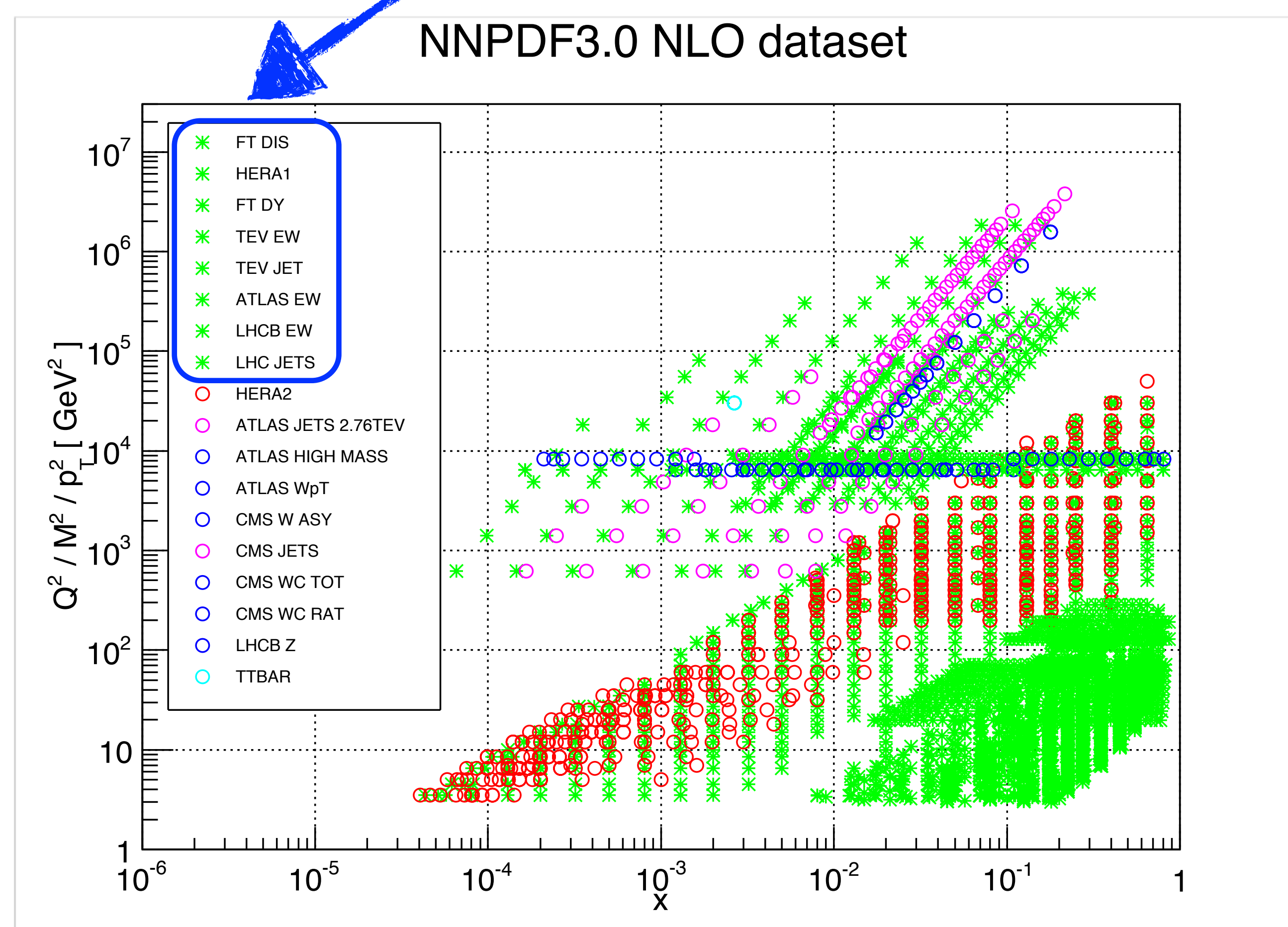
Backup Slides

The NNPDF3.0 Sets

The Dataset

- **~1000 new data points** as compared to NNPDF2.3.
- **HERA DIS data**:
  - HERA I combined charm cross sections σ_{cc}^{red}.
  - HERA II cross sections from ZEUS and H1.
- **LHC vector boson production data**:
  - Drell-Yan distributions from ATLAS and CMS,
  - W muon asymmetry from CMS,
  - forward $Z \rightarrow ee$ production from LHCb,
  - $W\ p_T$ distributions from ATLAS,
  - W + charm production from CMS.
- **LHC inclusive jet production data**
- **LHC top pair production total cross sections**

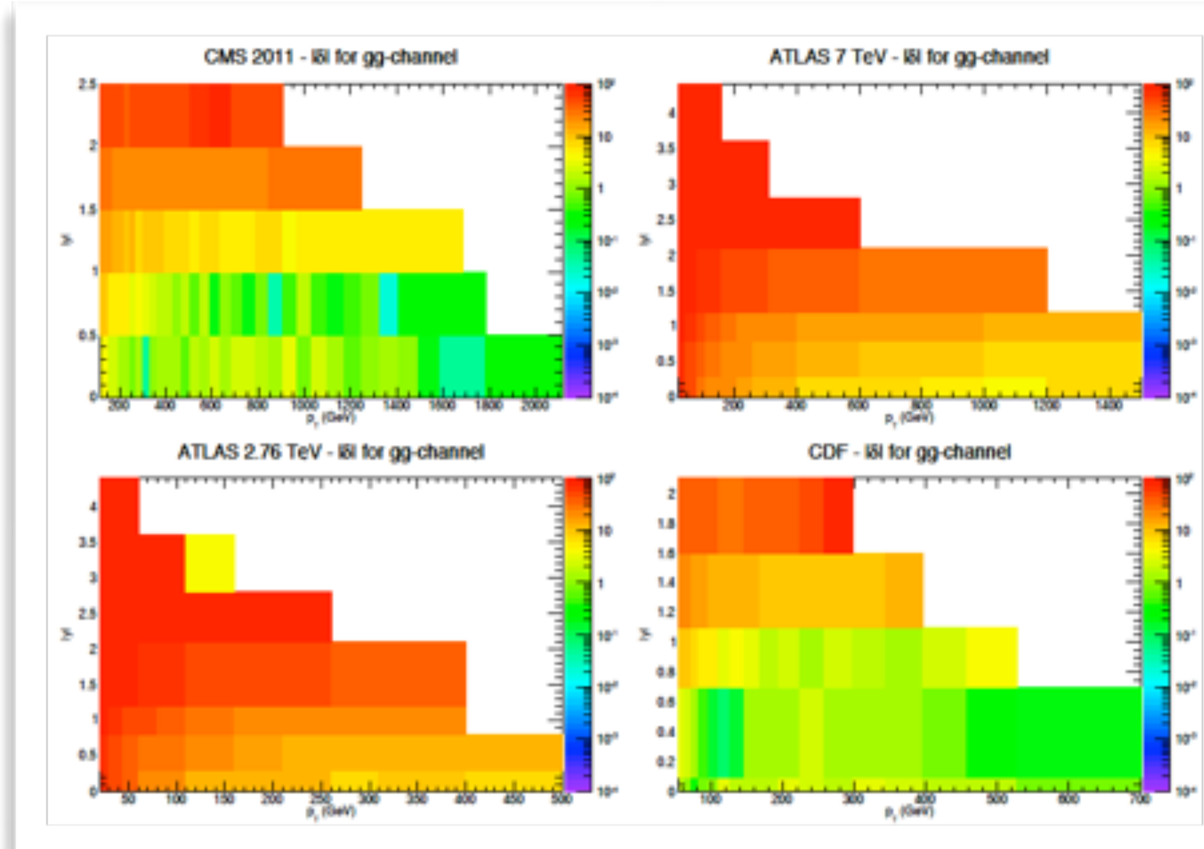
NNPDF2.3 dataset



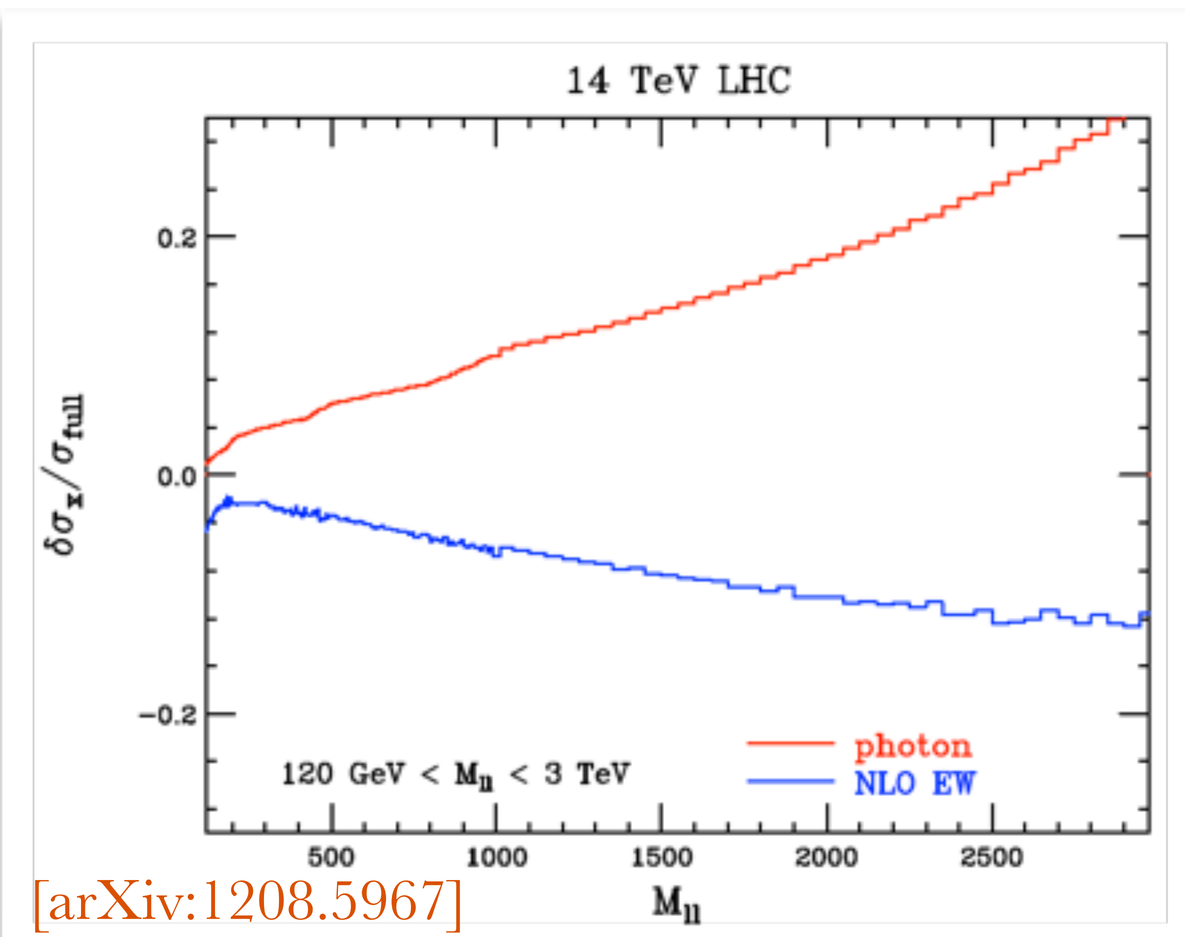
The NNPDF3.0 Sets

Some Theoretical Improvements

S. Carrazza and J. Pires [arXiv:1407.7031]



- Inclusion of the approximated **NNLO corrections to jet** production based threshold resummation:
  - supposed to be reliable at large p_T values,
  - only data for which the *gg* channel agrees with existing exact calculation are included in the fit.



[arXiv:1208.5967]

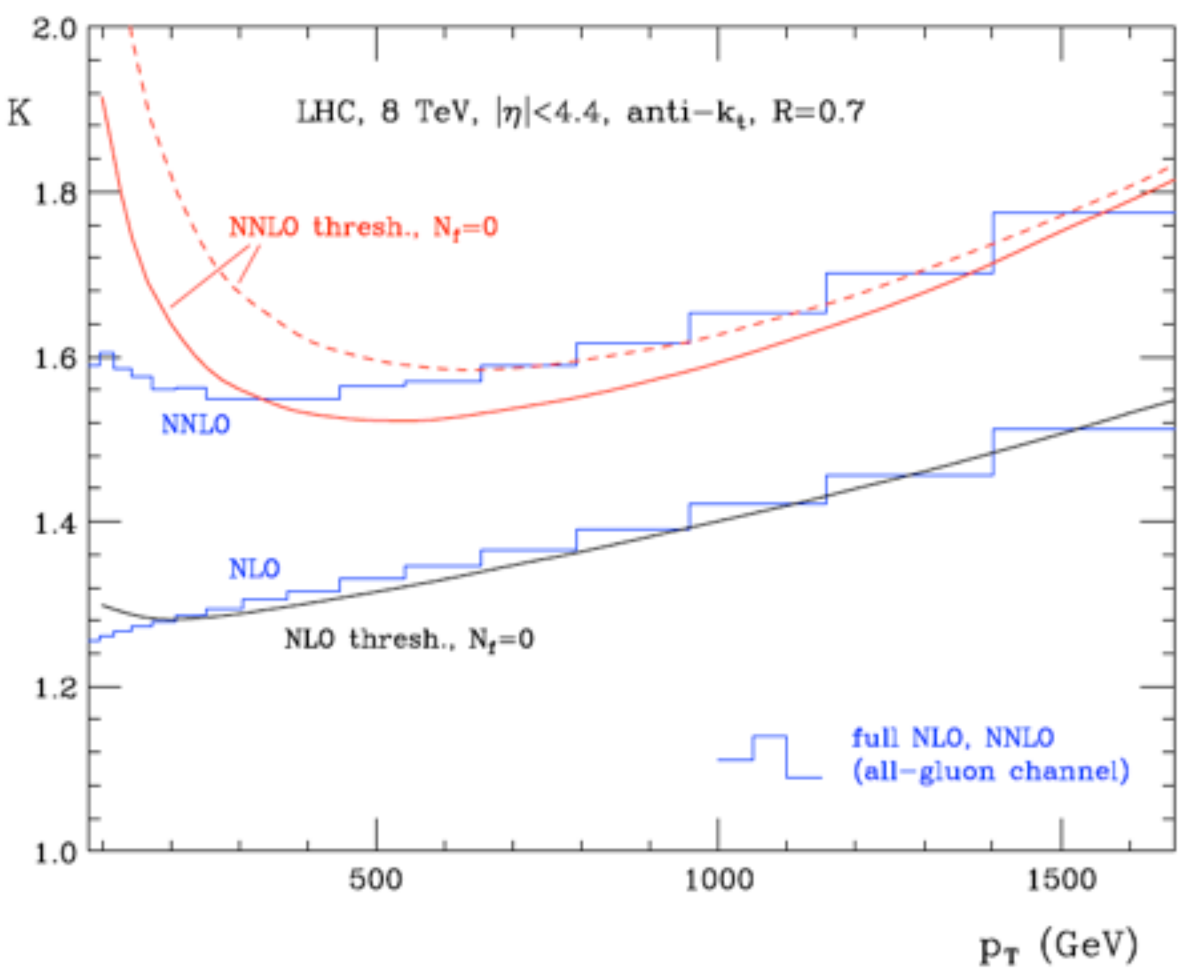
- **Electroweak corrections** to W and Z production:
  - Pure NLO electroweak corrections for all neutral current DY data computed with FEWZ3.1.
  - QED corrections affected by large uncertainty induced from uncertainty on photon PDF: not included here.

Theoretical Improvements
NNLO Corrections to Jet Production

- The full NNLO corrections to jet production are **still missing**.
  - Recently, the full NNLO corrections to the ***gg* channel** [O(20-25%) enhancement wrt NLO] and the leading color contribution to the $q\bar{q}$ channel have been computed.

- Approx. NNLO calculation based on **threshold resummation** exists:
  - supposed to be reliable at large p_T values but to break down at small p_T values.

[arXiv:1310.7192]

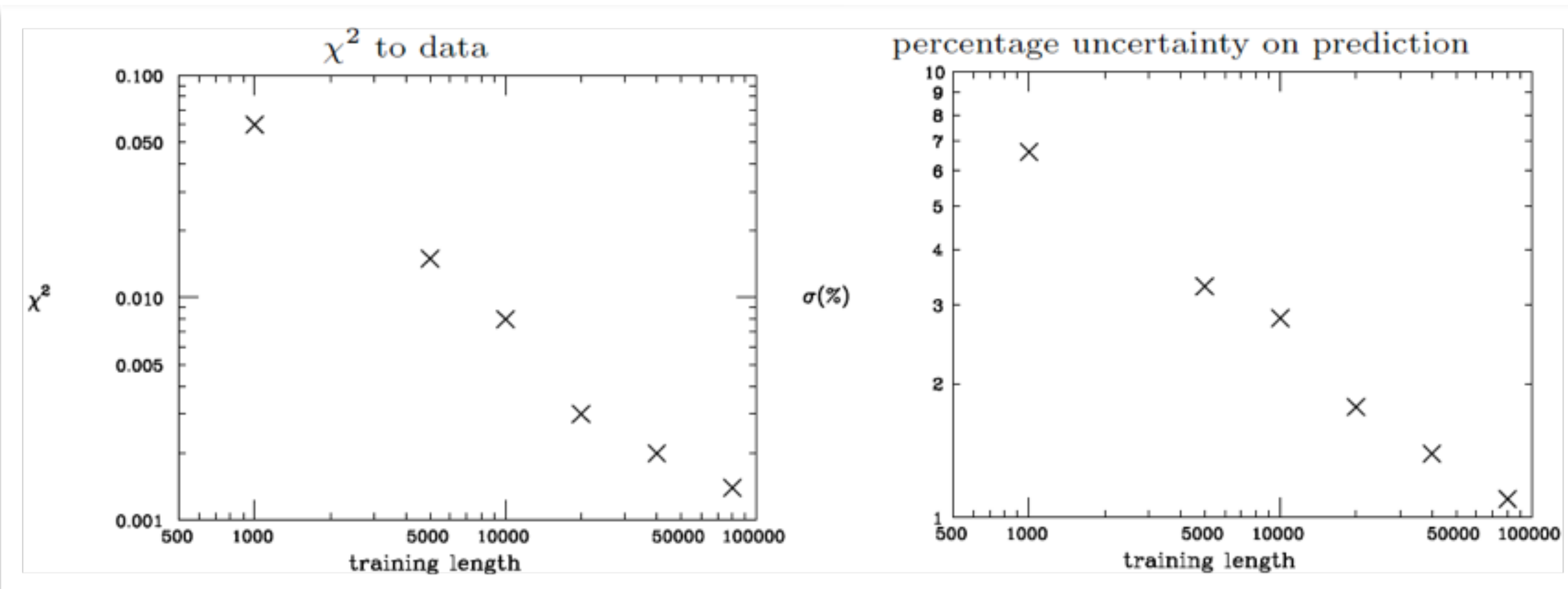


- Compute NNLO corrections using the threshold approximation.
- But how do we know whether we can trust them?
- Use the *gg* channel as a discriminant to include data in the fit.

Methodological Improvements

Closure Tests: LEVEL 0

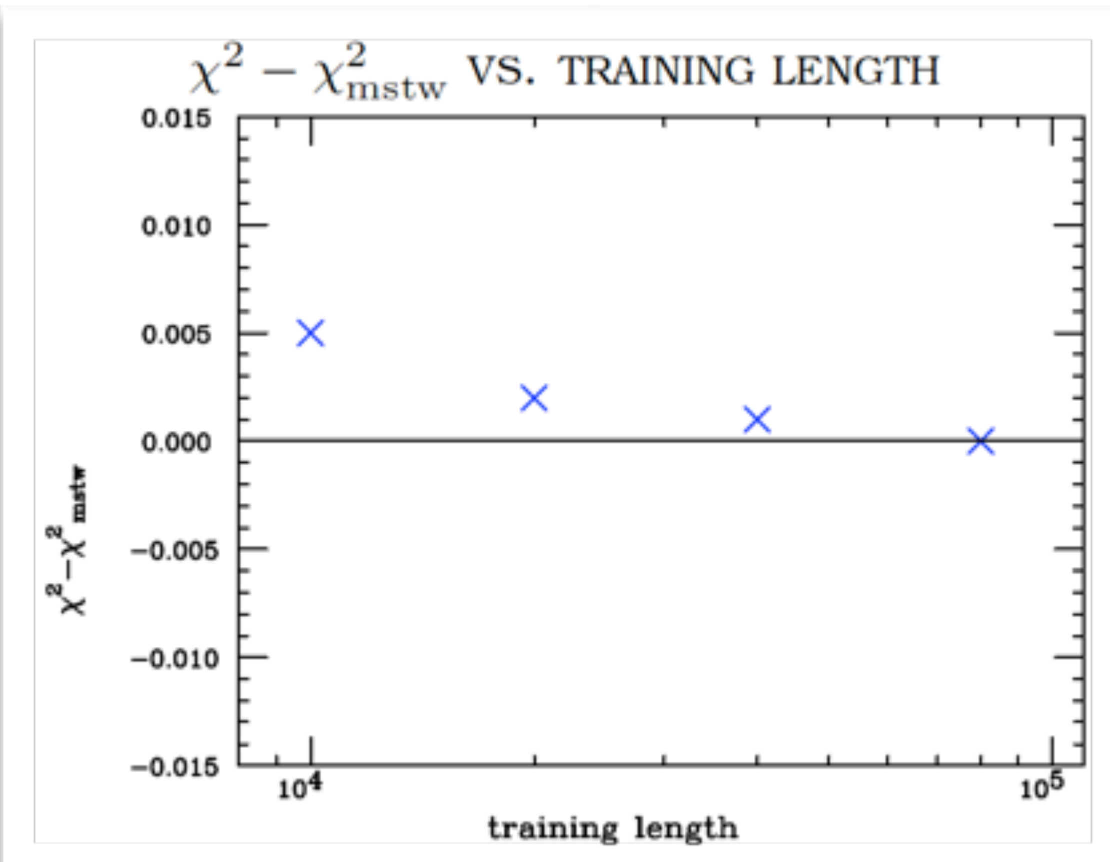
- Fits produced with increasing (fixed) training length.
- All fits on the same dataset (NNPDF2.3 dataset).

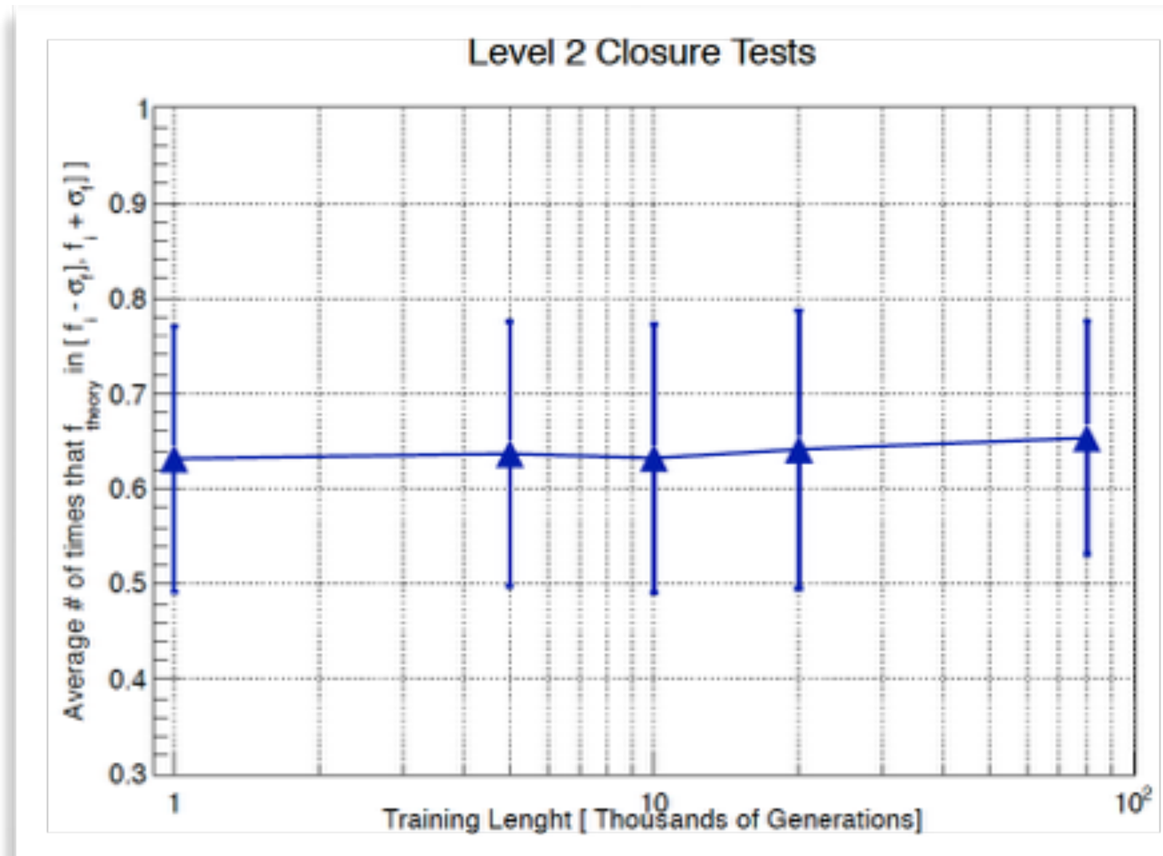


- χ^2 to data and PDF uncertainty on predictions **tend to zero** as the training length increases.
- LEVEL 0 closure test **works**!

Methodological Improvements
Closure Tests: LEVEL 2

- Fits produced with increasing (fixed) training length.
- All fits on the same dataset (NNPDF2.3 dataset).





- χ^2 tends to χ^2_{MSTW} (0.96) and the MSTW "true PDFs" are 68% of times within the 1-σ error band.
- LEVEL 2 closure test **works**!

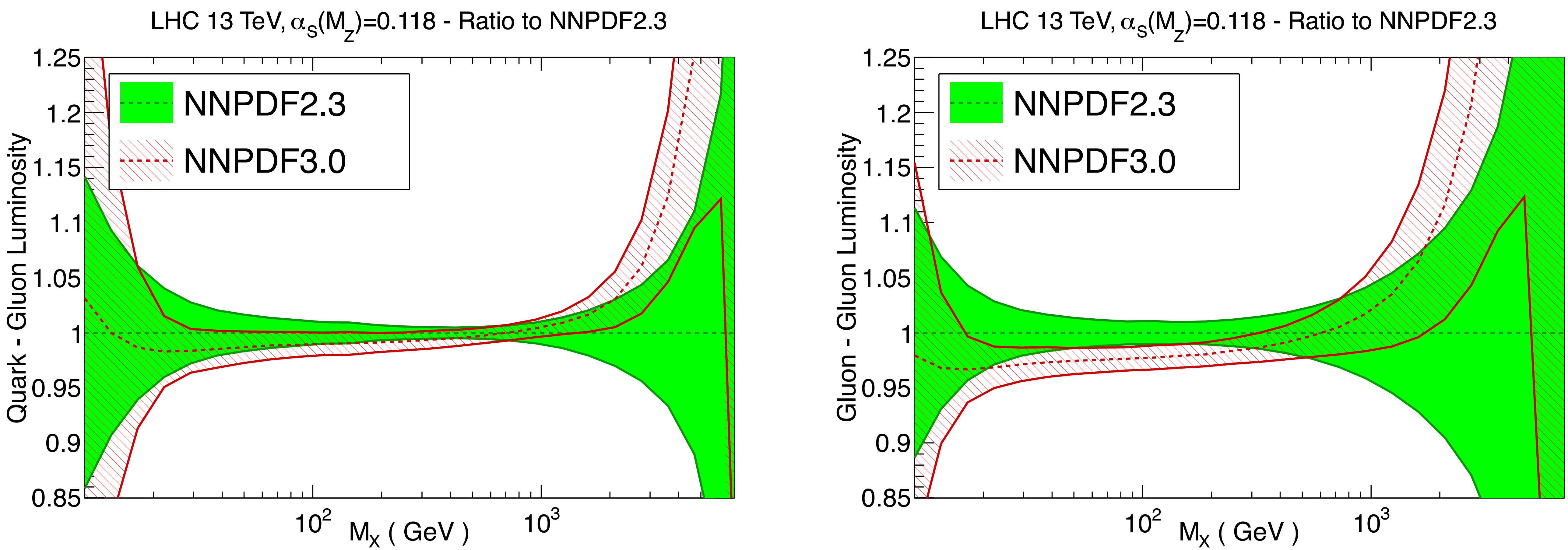
NNPDF

Results

NNPDF3.0 vs. NNPDF2.3: Luminosities

- The parton luminosity is relevant for **collider observables**:

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



- Important differences for quark-gluon and gluon-gluon luminosities.

- 1-σ difference in the range $30 \leq M_X \leq 300$ GeV which is relevant for gluon-initiated processes like **Higgs production**.

Results
The Strangeness

- In global PDF fits, the strange PDFs are mostly constrained by the **neutrino structure function** data (*e.g.* NuTeV and CHORUS):

  - at low scale they typically find $r_s \sim 0.5$, where:

$$r_s(x,Q^2) = \frac{s(x,Q^2)+\bar{s}(x,Q^2)}{\bar{d}(x,Q^2)+\bar{u}(x,Q^2)}\,, \qquad Q^2 \simeq 2\ \mathrm{GeV}^2$$

  - The suppressed strangeness has been supported by the recent analyses by the CMS Collaboration [arXiv:1312.6283] and Alehkin *et al.* [arXiv:1404.6469].

  - An analysis based on ATLAS W/Z and HERA data found $r_s \sim 1.0$ [arXiv:1109.5141]

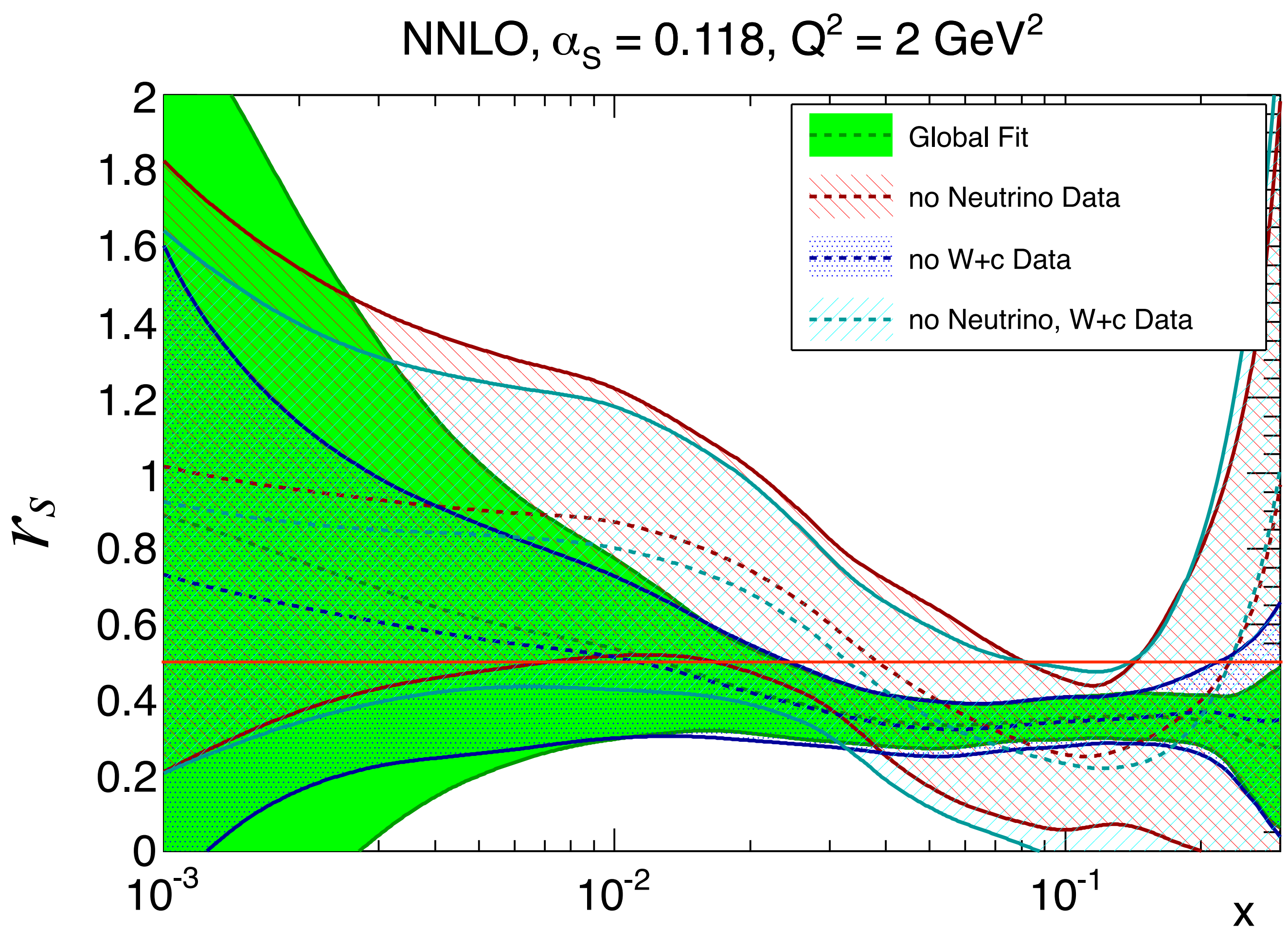
  - Result confirmed in a more recent analysis including $W + c$ data from ATLAS [arXiv:1402.6263].

  - Such an enhancement of the strange PDFs was confimed in the NNPDF2.3 framework but substantial uncertainties were found such to make the result compatible with the suppressed strangeness [arXiv:1207.1303].

Preliminary Results

The Strangeness

- Study the impact on the strangeness exlcuding one or both of the following datasets from the NNPDF3.0 fit:
  - Neutrino data: CHORUS and NuTeV data.
  - W + c data: CMS data.



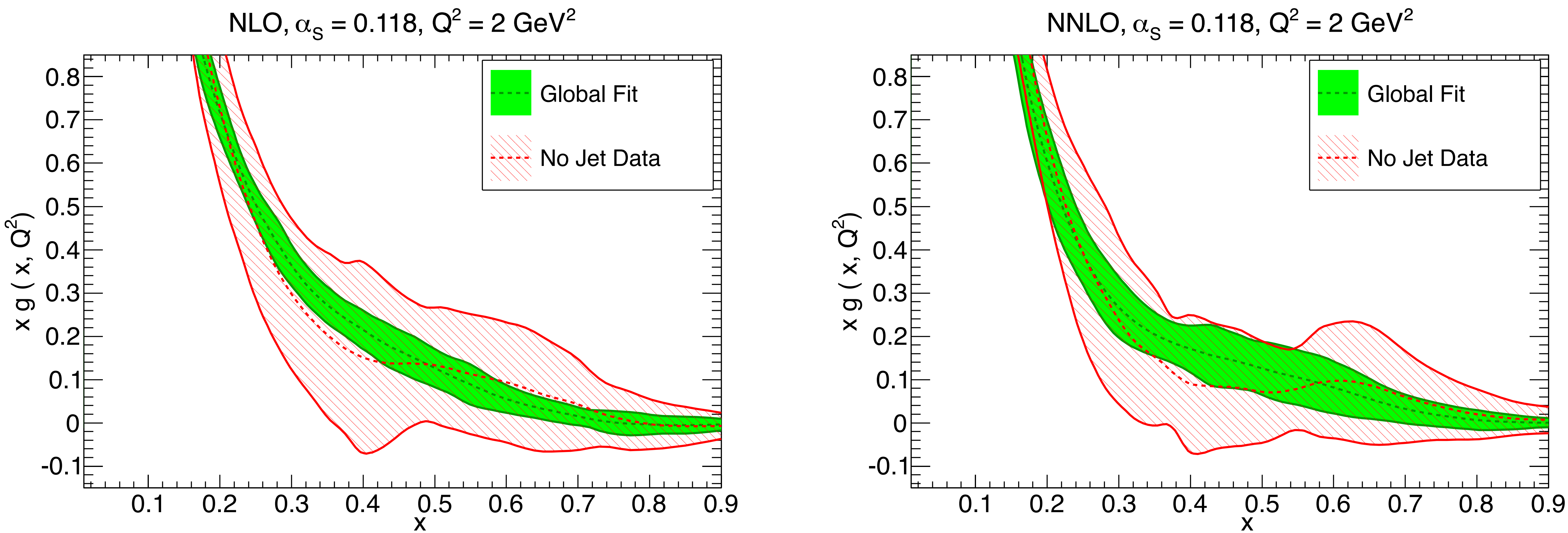
- **Overall compatibility**: all the fits are in agreement within 1-σ band.
- Neutrino data: stronger impact in the large-x unc. and tendency to suppress r_s.
- W+c data: moderate impact but also help in reducing the unc. in the large-x region.

- **No** evidence of **tension** between neutrino and LHC data.
- The **suppression** of the strange PDFs is **confirmed** at the level of central value ($r_s \sim 0.5$), but with a sizable uncertainty.

Preliminary Results

Impact of Jet Data

- Given that the exact NNLO calculation for jet productions is not available yet, we explored the possibility to remove all jet data from the fit.



- Quark PDFs and small-x gluon PDF mostly unaffected.
- Big impact on the large-x gluon PDF:
  - NNLO fit more affected because many jet data are excluded.
  - The central curves are very close ⇒ reliability of our strategy to include jet data.

Results

Conservative Partons

- Attempt to construct a **maximally consistent** PDF set by excluding from the fit the most inconsistent data sets (based and the $\mathcal{P}(\alpha)$ analysis).
- Exclude experimental data sets whose $\mathcal{P}(\alpha)$ distribution:

$$\mathcal{P}(\alpha) = \frac{1}{\alpha} \sum_{k=1}^{N} w_k(\alpha)\,, \quad w_k(\alpha) \propto \left(\frac{\chi_k^2}{\alpha^2} \right)^{\frac{1}{2}(n-1)} \exp\left[-\frac{\chi_k^2}{2\alpha^2} \right]$$

has two out of mean, median and mode bigger than $\alpha_{\max} = 1.20$.

- The $\mathcal{P}(\alpha)$ distribution measures how consistent a particular data set is with the global fit when rescaling all the uncertainties by a factor α.