PDF4BSM
Parton Distributions in the Higgs Boson Era

NNPDF Status and Plans

Juan Rojo
STFC Rutherford Fellow
Rudolf Peierls Center for Theoretical Physics
University of Oxford

The NNPDF Timeline

2008 ☑ **NNPDF1.0:** DIS only, fixed strangeness

Building on previous fits to F2NS (2002), F2 proton (2005) and qNS (2007)

2009 ☑ **NNPDF1.2**: DIS only, strangeness from neutrino dimuon data

Impact on the NuTeV anomaly, precision determination of Vcs

2010 ☑ **NNPDF2.0**: first global PDF fit with Drell-Yan, W,Z and jet production data, ZM-VFN

Included for first time combined HERA-I data, studies of combined PDF+alphas unc

01/2011 ☑ **NNPDF2.1 NLO**: heavy quark effects with FONLL, charm and bottom mass variations

Combined PDF+mc uncertainties, inclusion of HERA F2charm data

07/2011 ☑ **NNPDF2.1 NNLO and LO**: first LO and NNLO sets

FONLL-C GM-VFN scheme, LO vs LO PDFs study*

2012 ☑ **NNPDF2.3**: first global PDF fit included ATLAS, CMS and LHCb constraints

ATLAS jets and W,Z, CMS W asymmetry, LHCb W rapidity + NNPDF2.3QED ….

*Reference internal PDF set in **Pythia8** and in **MadGraph5_aMC@NLO***

10/2014 ☑ **NNPDF3.0**: major update following a **complete rewriting of the code in C++ and Python**, methodology validated on closure tests, > 1000 data points from HERA-II and LHC, approximate NNLO for jets

Quantitative understanding of the different statistical origin of PDF uncertainties

*Reference internal PDF set in **Sherpa 2.2.0***

NNPDF3.0 is really NNPDF++

☑ Completely rewritten **Fortran NNPDF fitting code** into **C++** and **Python:** massive code development effort, > 2 years to complete

☑ **Modular structure**: each dataset is now an individual object, with the associated theory encapsulated in individual **FK tables:** easy to include new measurements, to upgrade theory for existing ones

☑ This new efficient code is the key to be able to perform systematic studies of the PDF fitting methodology like **closure testing**, which were not possible with the old Fortran code

☑ **Greatly improved fitting efficiency**: main bottleneck for PDF fits is convolution between input PDFs and theory, performed here with **assembly-like structure**

☑ **Generalized PDF parametrization**: fits can now be performed in any **arbitrary input PDF basis**, with new self-consistent method to determine **preprocessing exponent ranges**

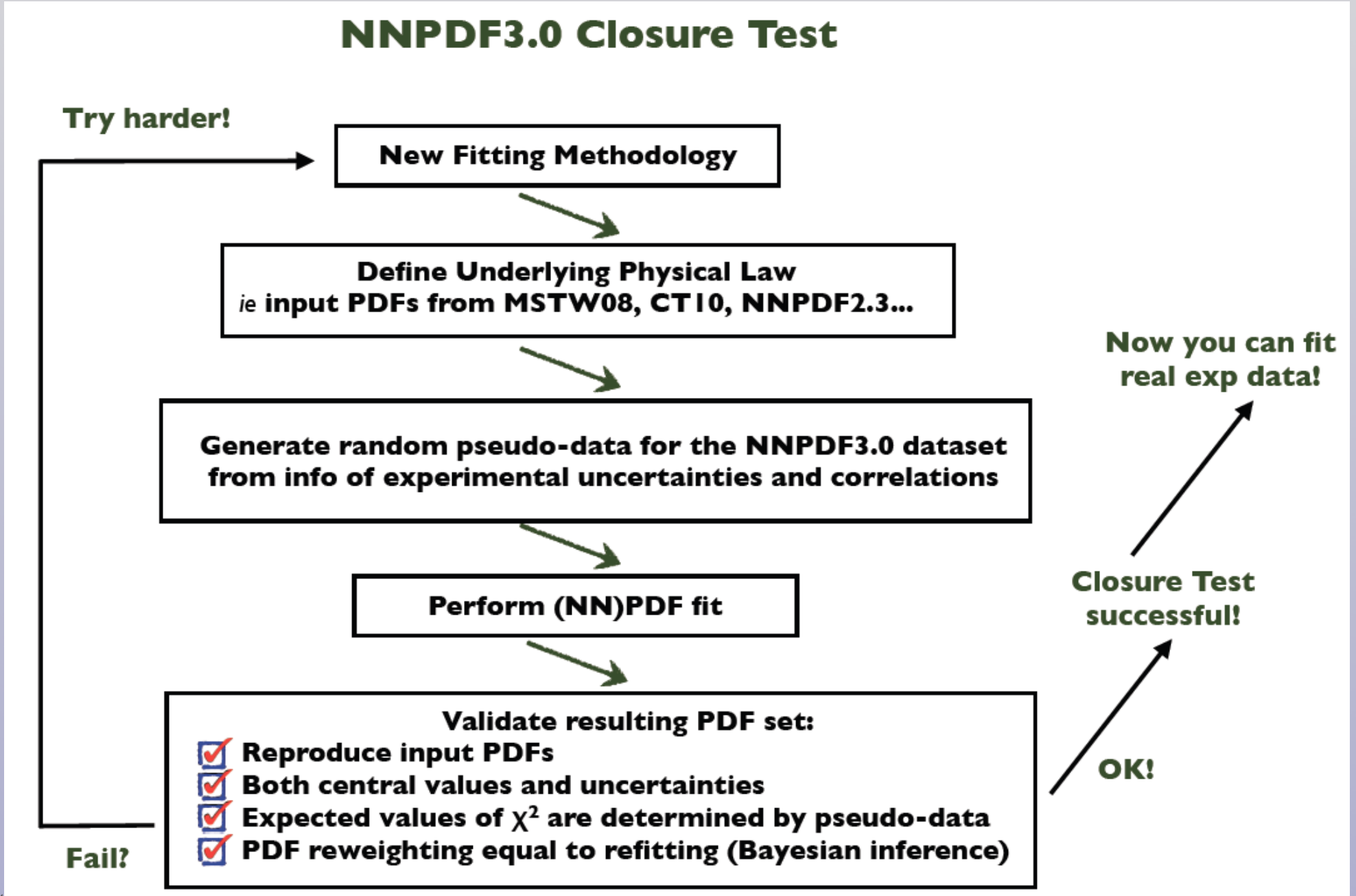
☑ Optimisation of the **generalised positivity of PDFs**: crucial for robust estimates of PDF errors in extrapolation regions

☑ **These technical and conceptual improvements** guarantee **robustness and stability for NNPDF development** in the medium and long term

☑ Now working in integrating **APFEL** in NNPDF with **APFELcomb**, to be able to include straightforwardly in the NNPDF fits **new theory developments** like resummations, intrinsic charm, scale variations, new QCD+QED fits …..

Closure Testing

Validation and optimization of fitting strategy performed on **closure test with known** underlying PDF set



Closure Testing

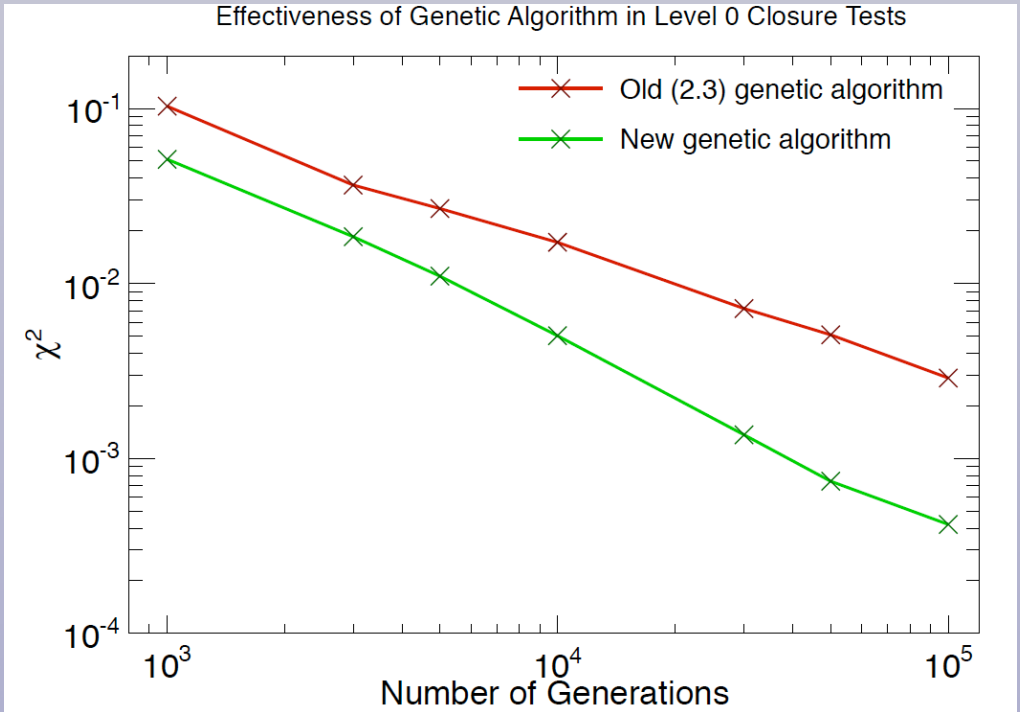
- **Level 0: no fluctuations** on pseudo-data, **no Monte Carlo** replica generation
- **Level 1**: **with fluctuations** on pseudo-data, **no Monte Carlo** replica generation
- **Level 2**: **with fluctuations** on pseudo-data, **with Monte Carlo** replica generation

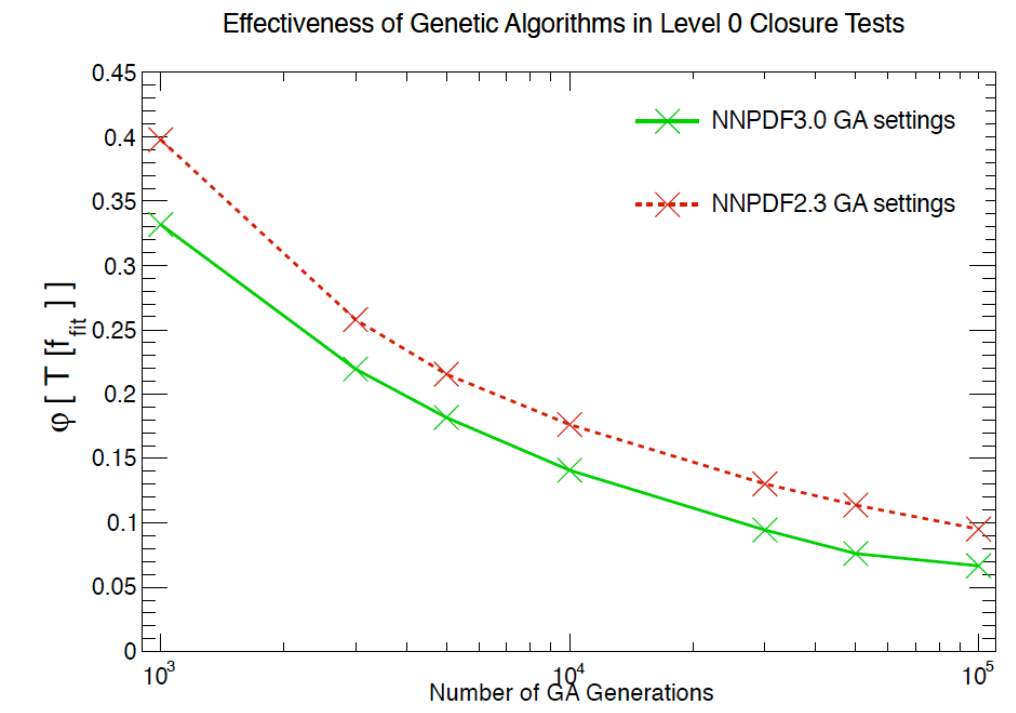
Level 0 Closure Tests:

- ☑ **Central values of input PDF** reproduced with arbitrary accuracy
- ☑ **PDF uncertainties on the fitted data points** can become arbitrarily small

Genetic Algorithms minimization efficiency substantially improved wrt NNPDF2.3

$$\varphi_{\chi^2} \equiv \sqrt{\langle \chi^2[\mathcal{T}[f_{\rm fit}], \mathcal{D}_0]\rangle - \chi^2[\langle \mathcal{T}[f_{\rm fit}]\rangle, \mathcal{D}_0]}\,.$$





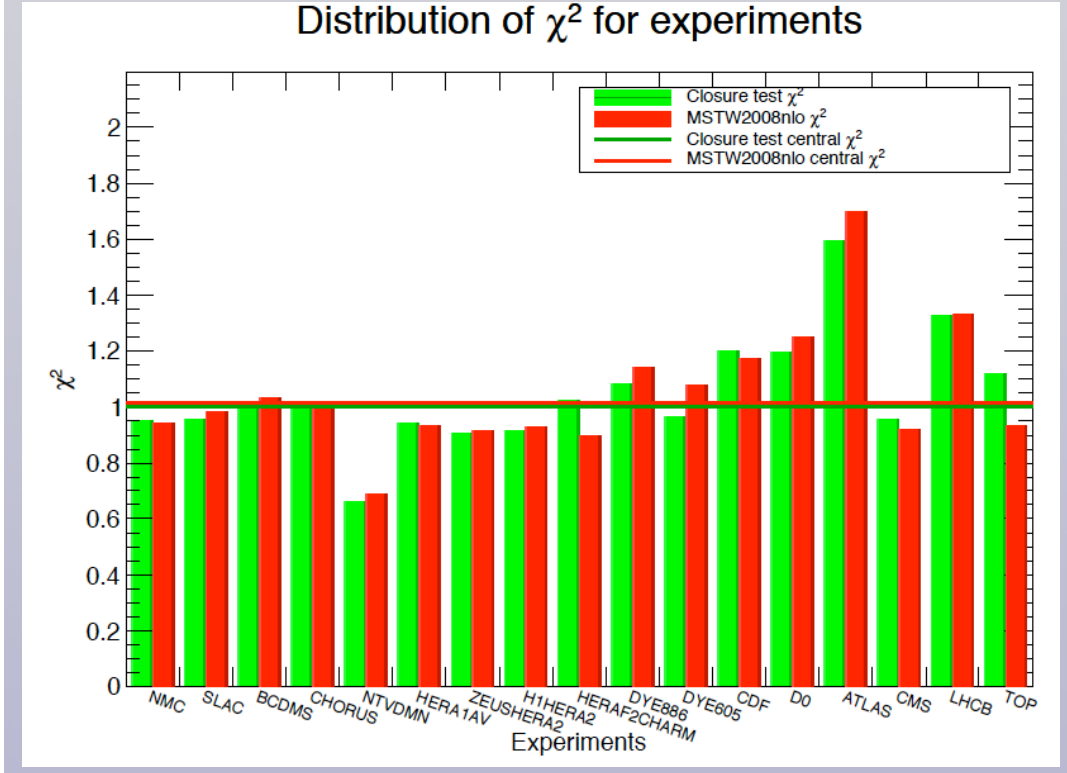
Measure of PDF uncertainties in the fitted cross-sections

Closure Testing

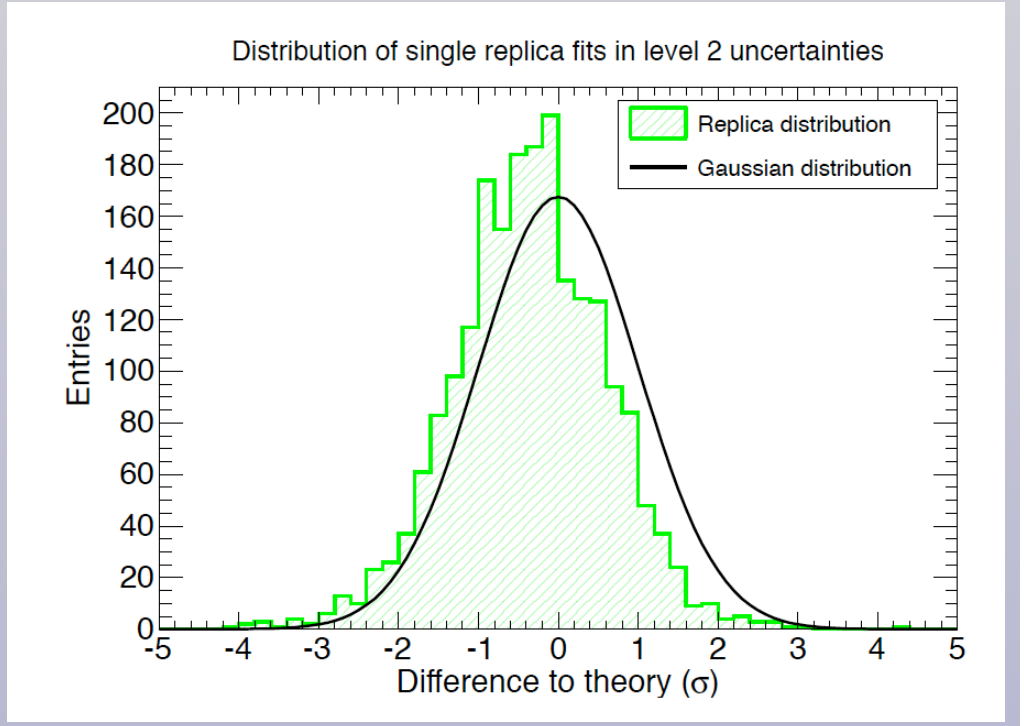
Level 2 Closure Tests:

☑ Reproduced **χ^2 of input PDF** - both total and individual experiments

☑ **Fitted PDF central values** fluctuate around input values by the same amount as expected from the **size of the PDF uncertainties**



Difference between fit and input PDFs central values in units of the PDF uncertainties



$$\xi_{n\sigma} = \frac{1}{N_{\rm PDF}}\frac{1}{N_x}\frac{1}{N_{\rm fits}}\sum_{i=1}^{N_{\rm PDF}}\sum_{j=1}^{N_x}\sum_{l=1}^{N_{\rm fits}} I_{[-n\sigma^{i(l)}_{\rm fit}(x_j),n\sigma^{i(l)}_{\rm fit}(x_j)]}\left(\langle f^{i(l)}_{\rm fit}(x_j)\rangle - f^i_{\rm in}(x_j)\right)$$

$$\xi^{(12)}_{1\sigma} = 0.699\,, \qquad \xi^{(12)}_{2\sigma} = 0.948\,,$$

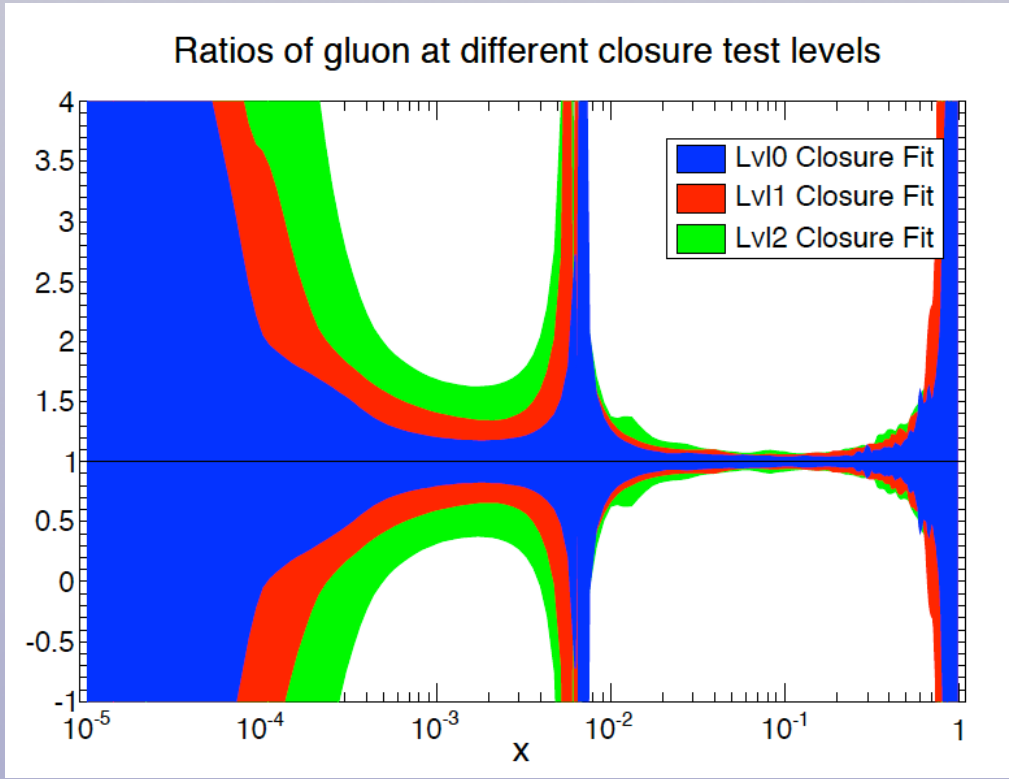
☑ The **central values of the fitted PDFs** fall in the **one(two)-sigma intervals** around **68%(95%)** of the times (average over x values and flavors)

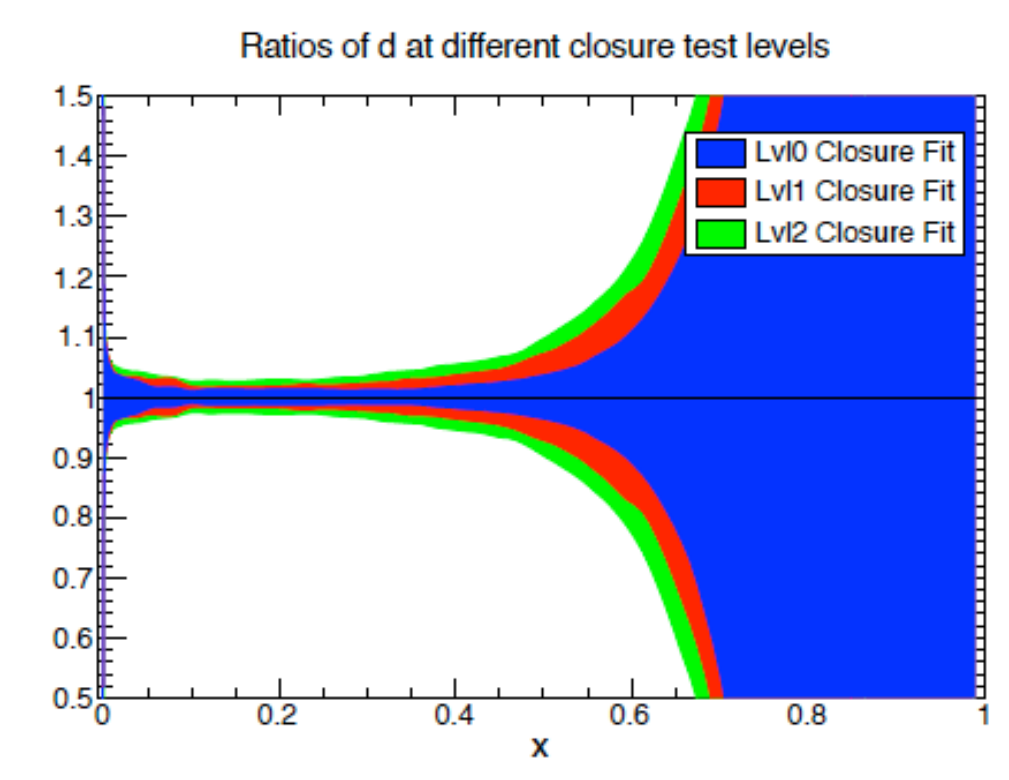
Closure Testing

Comparing **Level 0, 1 and 2 closure tests** provide a quantitative determination of the **different components of the total PDF uncertainty:**

- **Level 0:** ***Extrapolation*** **uncertainty,** due to limited kinematical coverage of experimental data. Here best-fit PDF uniquely defined - the input PDF set
- **Level 1:** ***Functional*** **uncertainty,** where a large number of different functional forms can provide an equally good fit to the pseudo-data (which now includes fluctuations)
- **Level 2:** ***Experimental*** **data uncertainty,** due to the genuine fluctuations in the experimental measurements around their true value

In regions with experimental data, the **three components are of similar size**: mandatory to include all of them for a robust estimate of PDF uncertainties





Closure Testing - What's next?

- **Closure tests** introducing **artificial inconsistencies between datasets**: does the interpretation of PDF uncertainties change? Can we understand this way the values of the **tolerances** used by CT and MMHT?

- **Closure tests** including **theory systematics**: can we test their statistical interpretation? What happens if different probability distributions for the theory (gaussian, flat, …) are used?

- Closure tests to optimize **experimental measurements**. Very clean tool to diagnose the best way to **present data**, in terms of binning, breakup of errors etc….

- For instance we found that **finner binning** is always preferable to **smaller uncertainties**, since the latter is more sensitive to **functional and extrapolation PDF uncertainties** (similar remarks by Pavel)

- Closure tests might provide also a unique, very clean, framework for **PDF benchmarking exercises**

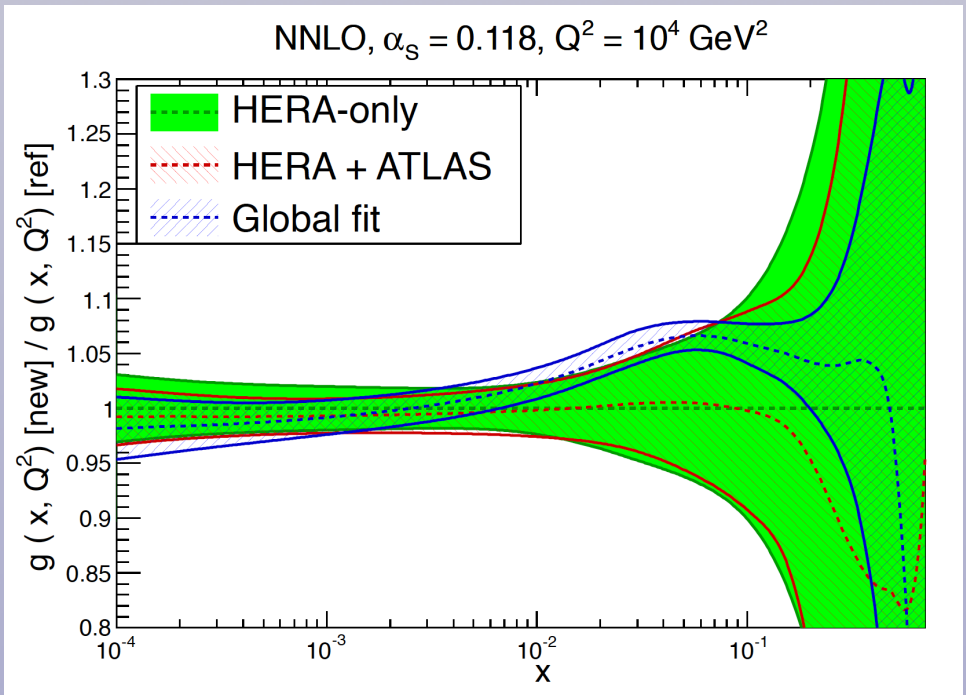
NNPDF Plans - Experimental Data

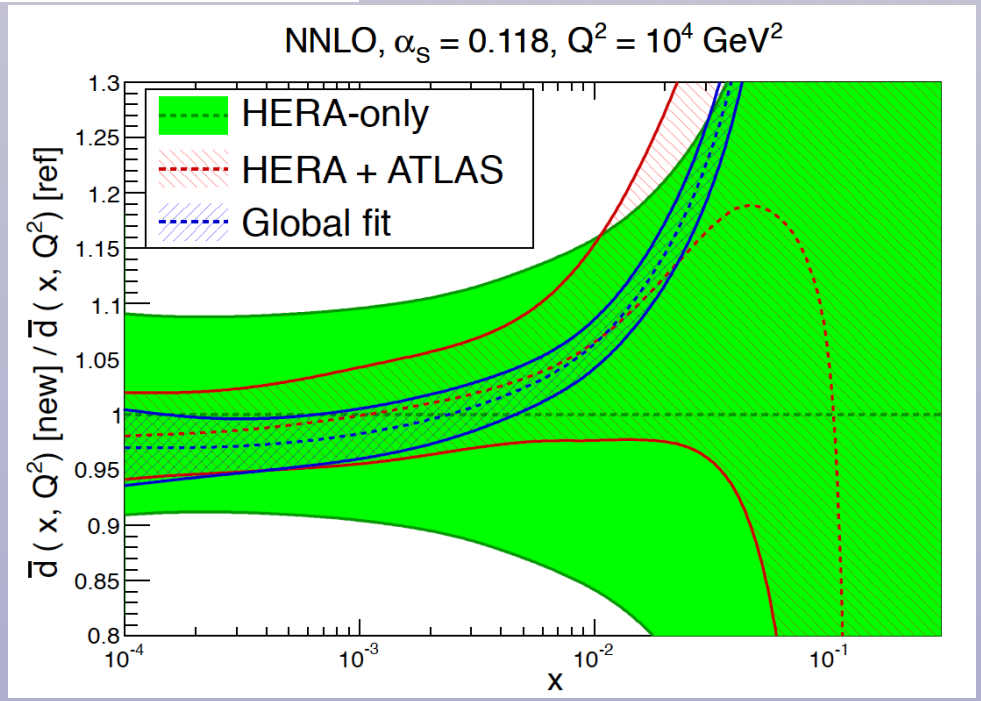
☑ First of all, we plan to include the final **HERA legacy dataset (March 2015)** as soon as it is released (but note that NNPDF3.0 already includes all published HERA-II inclusive data!)

☑ We keep also including LHC measurements: **ATLAS 2011 inclusive jets and dijets, ATLAS/CMS top quark pair differential distributions, ATLAS W+charm, CMS 2012 Drell-Yan,** LHCb Z rap distributions, direct photon production, **Z pt distributions,** ……

☑ With available data, **collider-only PDFs** are still subject to **rather larger uncertainties than in global fits**, but as more and more data becomes available, the differences will be reduced

"ATLAS Global fit" based on NNPDF3.0



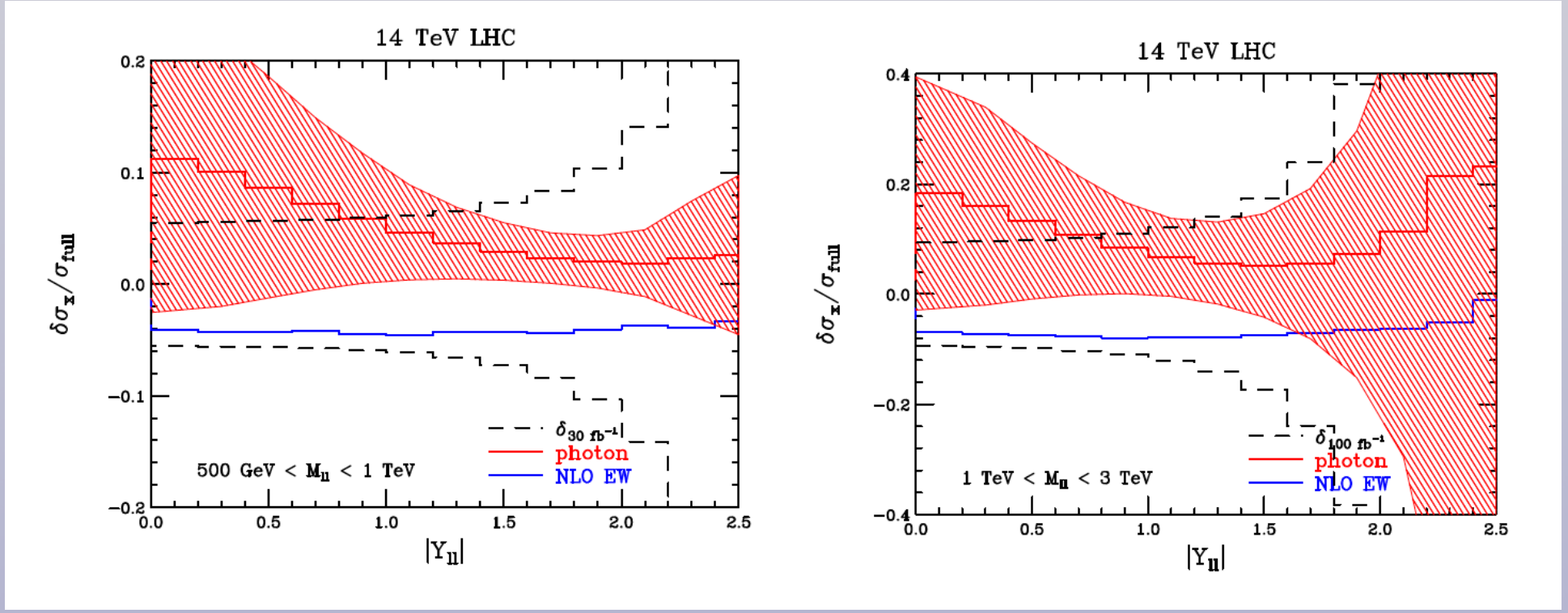


Pinning down the photon PDF

At **Run II,** some interesting **EWK measurements** that have been proposed to constrain the **photon PDF**

For instance, the triple differential measurement of Drell-Yan cross-sections in **invariant mass, rapidity and lepton transverse momentum** allows to neatly disentangle photon PDF effects from other EWK effects

Boughezal et al, arXiv:1312.3972



We plan to produce a NNPDF3.0QED fit using the **aMCfast** interface to the upcoming **QED+EW capabilities of MadGraph5_aMC@NLO** and updated LHC data sensitive to the photon PDF

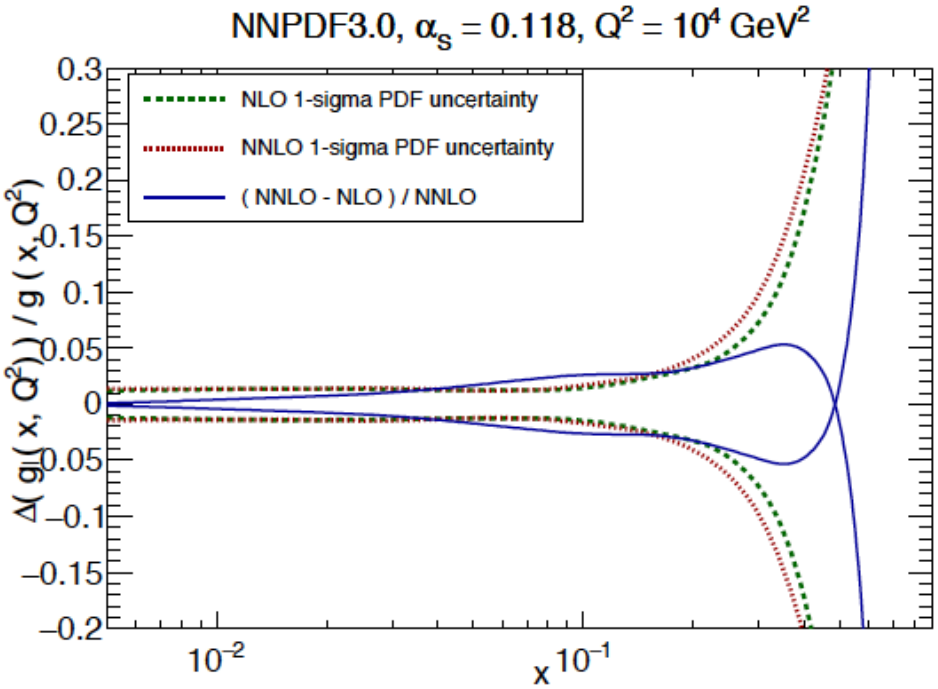
We also plan to **systematically include EWK corrections** in NNPDF using aMCfast

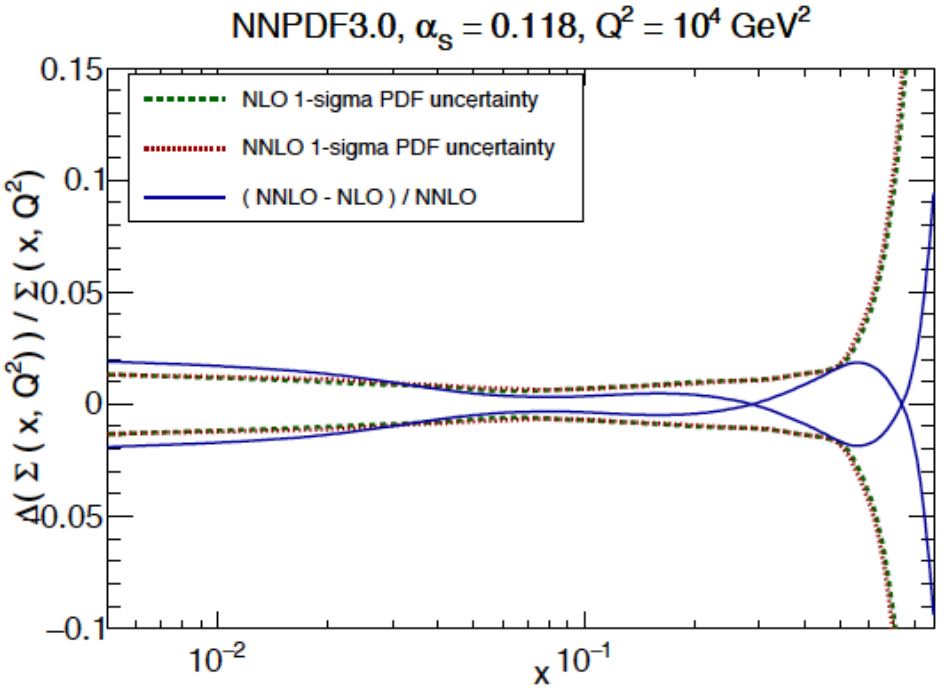
Theoretical uncertainties on PDFs

☑ Perhaps one of the most urgent topics that PDF fitters need to tackle as soon as possible is that of the **theoretical uncertainties on PDFs**

☑ Assessing TH uncertainties on PDFs is really crucial to establish the robustness of PDF predictions at the LHC and ensure that there are no hidden systematics

☑ A very crude estimate of how large TH uncertainties might be can be provided by comparing PDFs at **different perturbative orders**





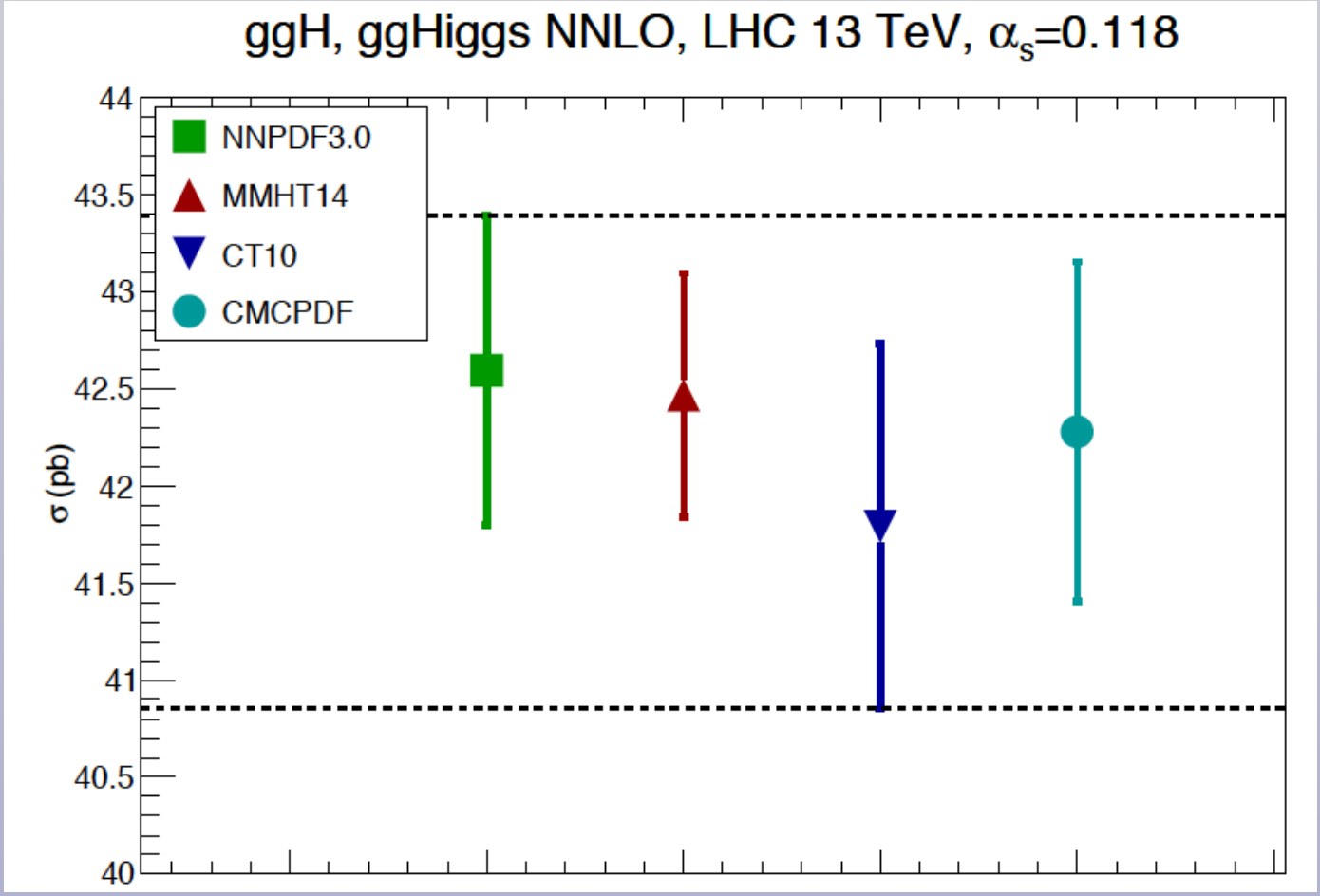
☑ Other options that should be explored are scale variations, the Cacciari-Houdeu method ….

☑ **Challenging task**, but we cannot avoid it for longer ….

Theoretical uncertainties on PDFs

The recent updates of the three global PDF sets **agree reasonably well for ggHiggs**

But is the picture **robust once we include theoretical uncertainties in the PDF fits**? What about other **theory systematics like heavy quark masses**? This is a really important issue for our community!



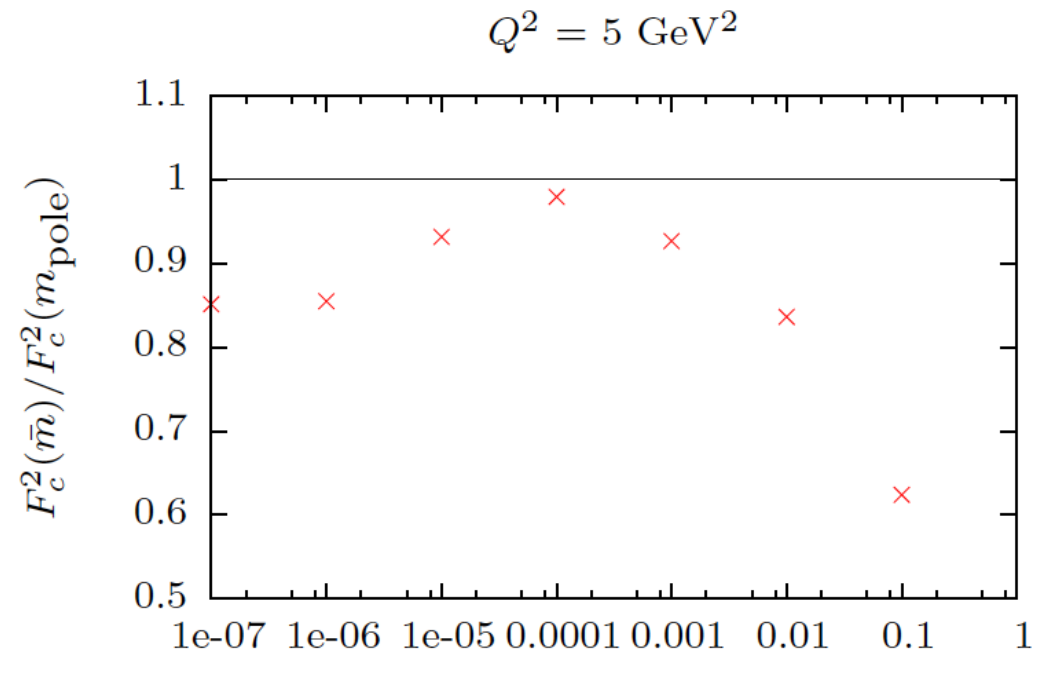
Improving heavy quark treatment

☑ We are working on extending the **FONLL GM-VFN scheme** to include **MSbar running masses** and the possibility of **intrinsic charm**

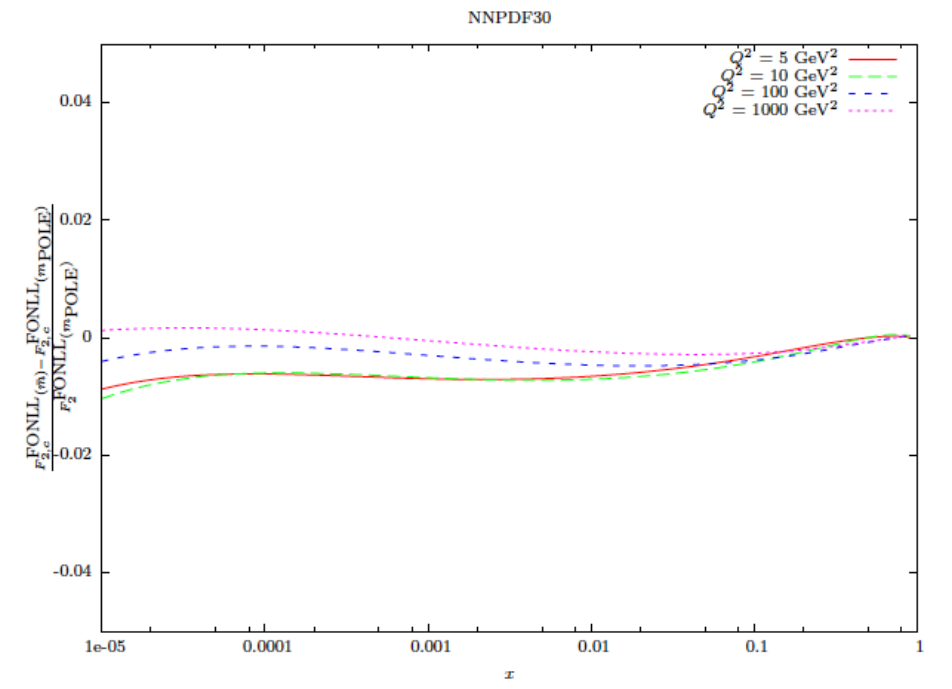
☑ The effect of the **pole -> MSbar scheme change on F2charm**, keeping the same numerical value of the charm mass fixed, is moderate, <10% in the relevant region but **not negligible** (1% effect at the inclusive level)

☑ In our next release, NNPDF3.1, we will provide a **range of variations in the charm and bottom masses**, important to estimate the **combined PDF+mc/mb uncertainties**

F2charm NNLO pole / F2charm NNLO MSbar



(F2charm_pole - F2charm_MSbar)/F2total



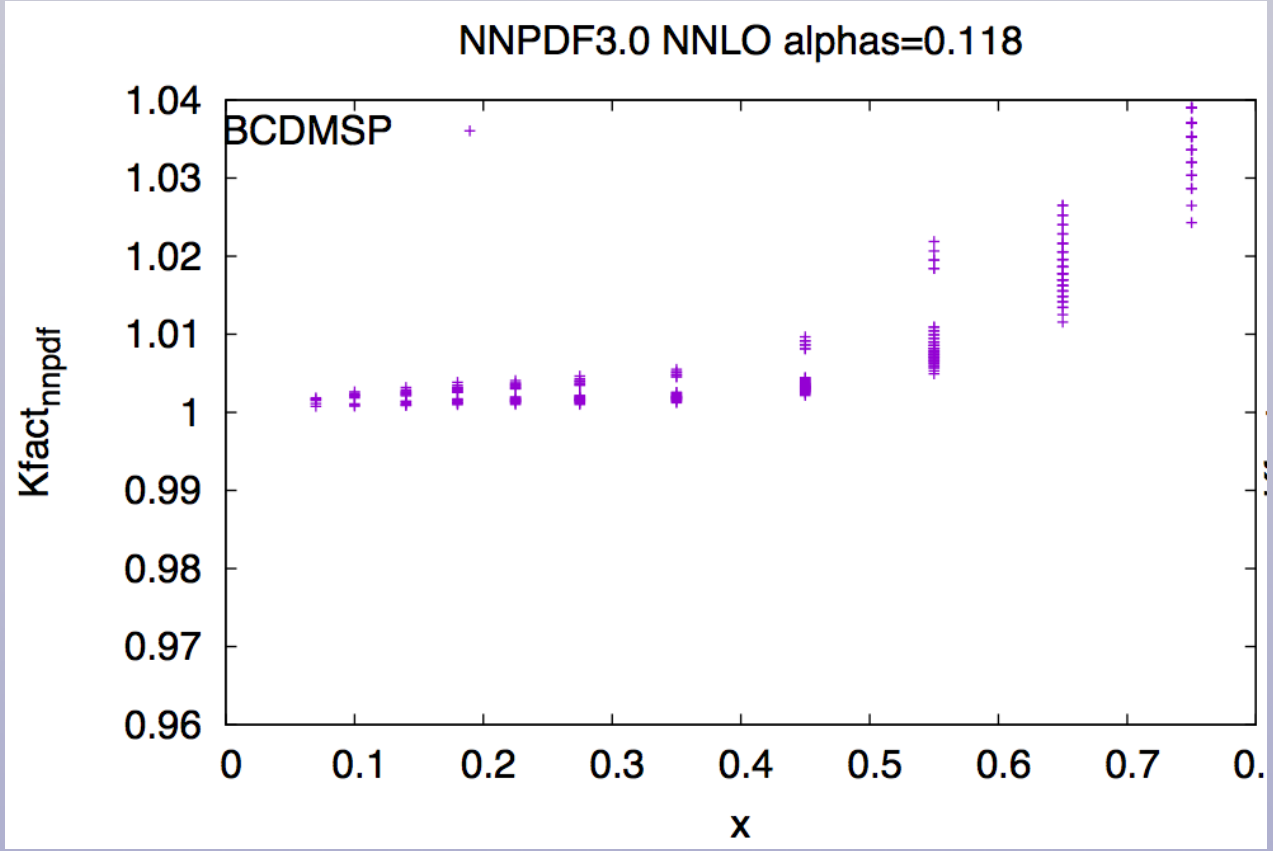
Beyond fixed-order QCD

☑ We are also working on PDF fits beyond fixed-order QCD, with **large-x (threshold) resummations** and **small-x (high-energy) resummation**

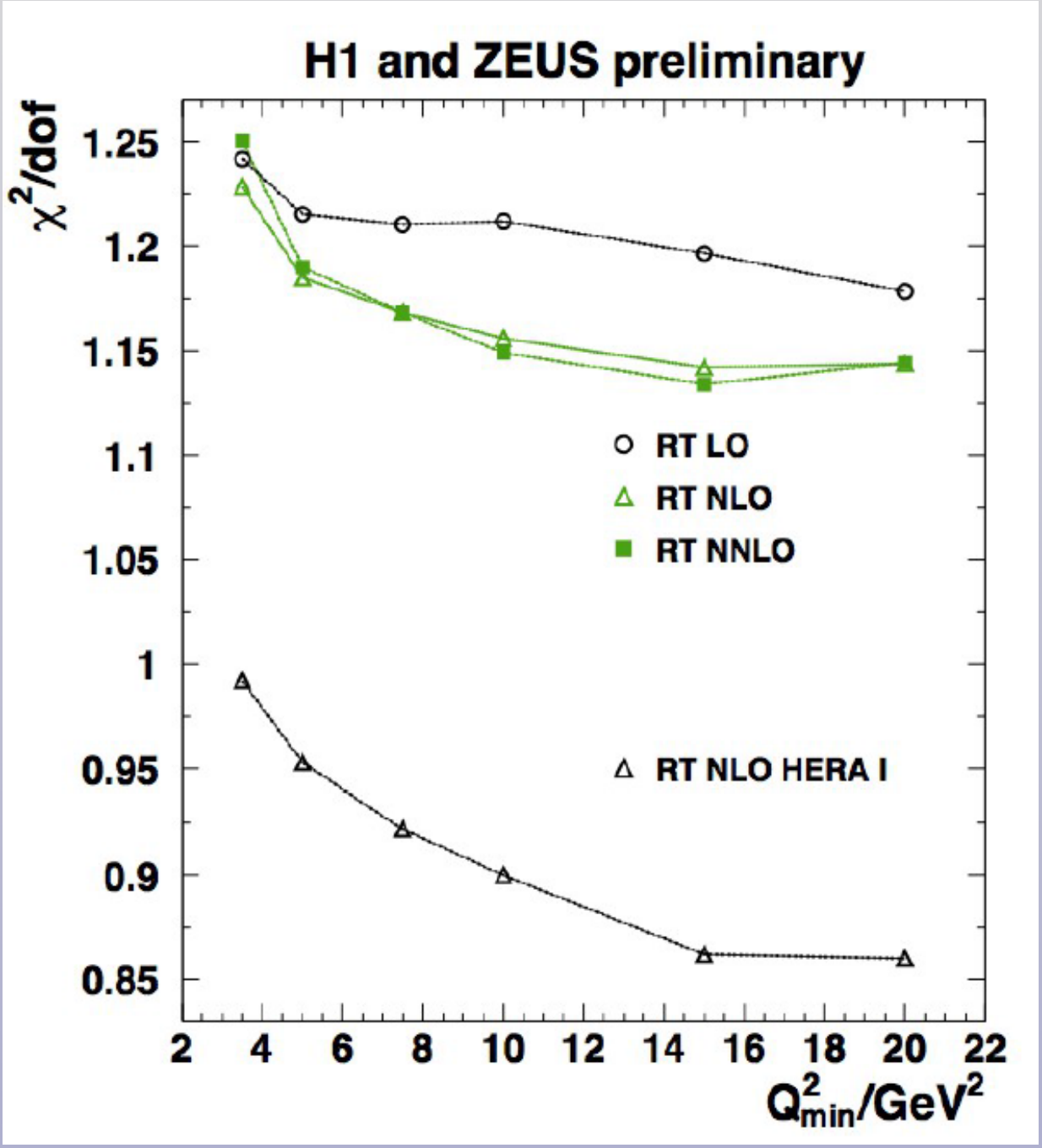
☑ Threshold resummation could important for **large-x DIS (fixed-target data)** and for **Drell-Yan at large rapidities**, although in these processes perturbative convergence is rather good

☑ Plan to produce for the first time **NNLO+NNLL resumed PDFs** (which can be used consistently with resumed calculations for Higgs, ttbar, etc) and **approximate N3LO PDFs**

Ratio of F2 NNLO+NNLL over F2 NNLO for BCDMS



High-energy resummation



☑ Possible departures from **linear DGLAP evolution** at small-x due to BFKL (high-energy) resummation, saturation or non-linear QCD effects, still need to be robustly identified

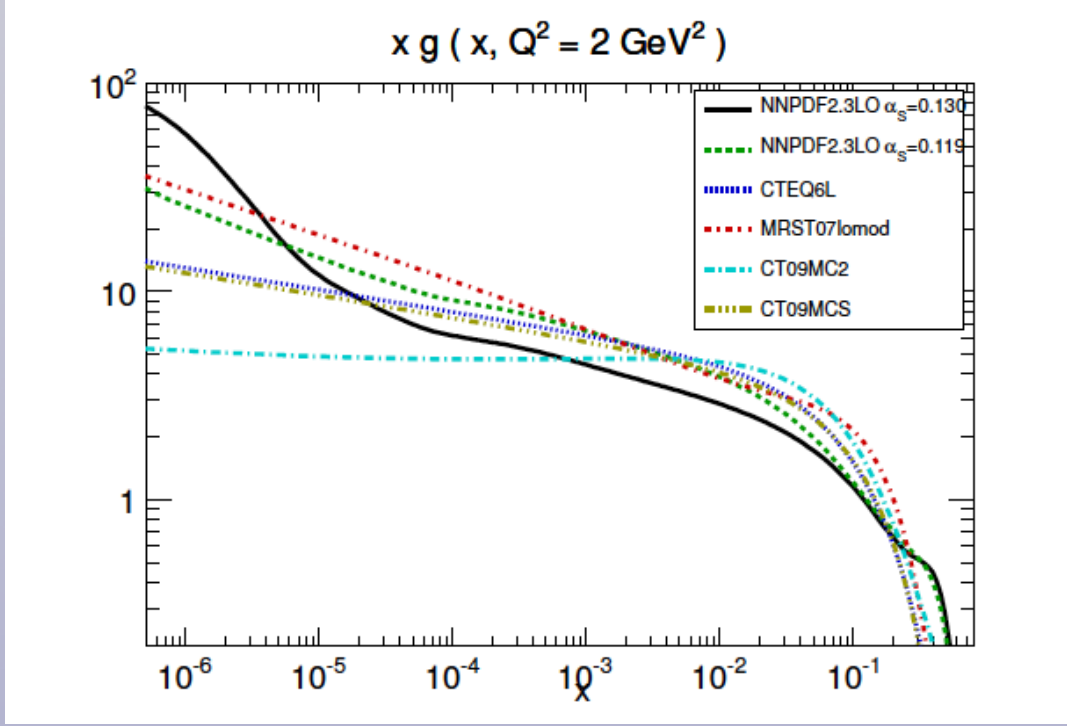
☑ The recent **HERAPDF2.0 analysis**, using the HERA legacy data, shows a strong trend toward linear DGLAP might have trouble to describe the low Q2 data

☑ To verify if this is a genuine BFKL effect, we plan to produce **NNPDF fits with small-x resummation**

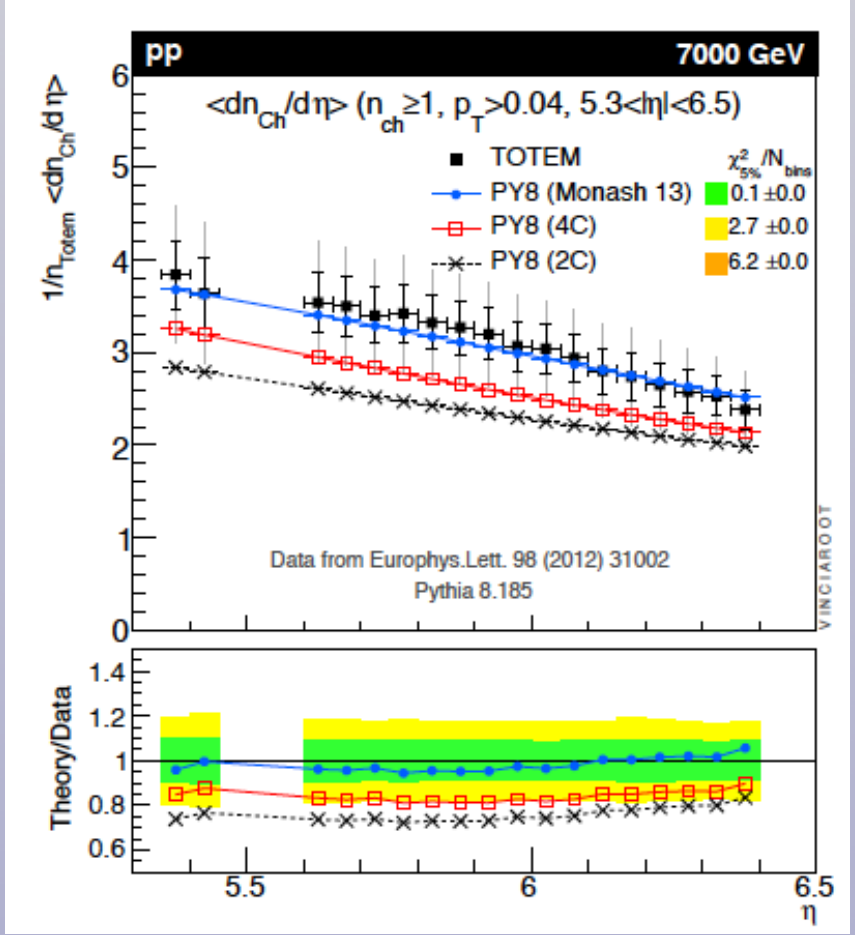
☑ If this trend cured by resummation, this would provide the **first unambiguous evidence for BKFL dynamics at small-x**, with important implications at the LHC

PDFs and Monte Carlo generators

- **PDFs** are an essential ingredient for the **tuning of soft and semi-hard physics** in LO Monte Carlo event generators like **Pythia8**, **Herwig++** or **Sherpa**
- Most updated tune of **Pythia8**, the **Monash 2013 Tune,** is based on the **NNPDF2.3LO** set. Same for the **dedicated ATLAS14 (A14) tune**
- The **harder small-x gluon in NNPDF2.3LO** helps to improve the description of the LHC **forward data**



Small-x behaviour of gluon determines soft and semi-hard physics at the LHC



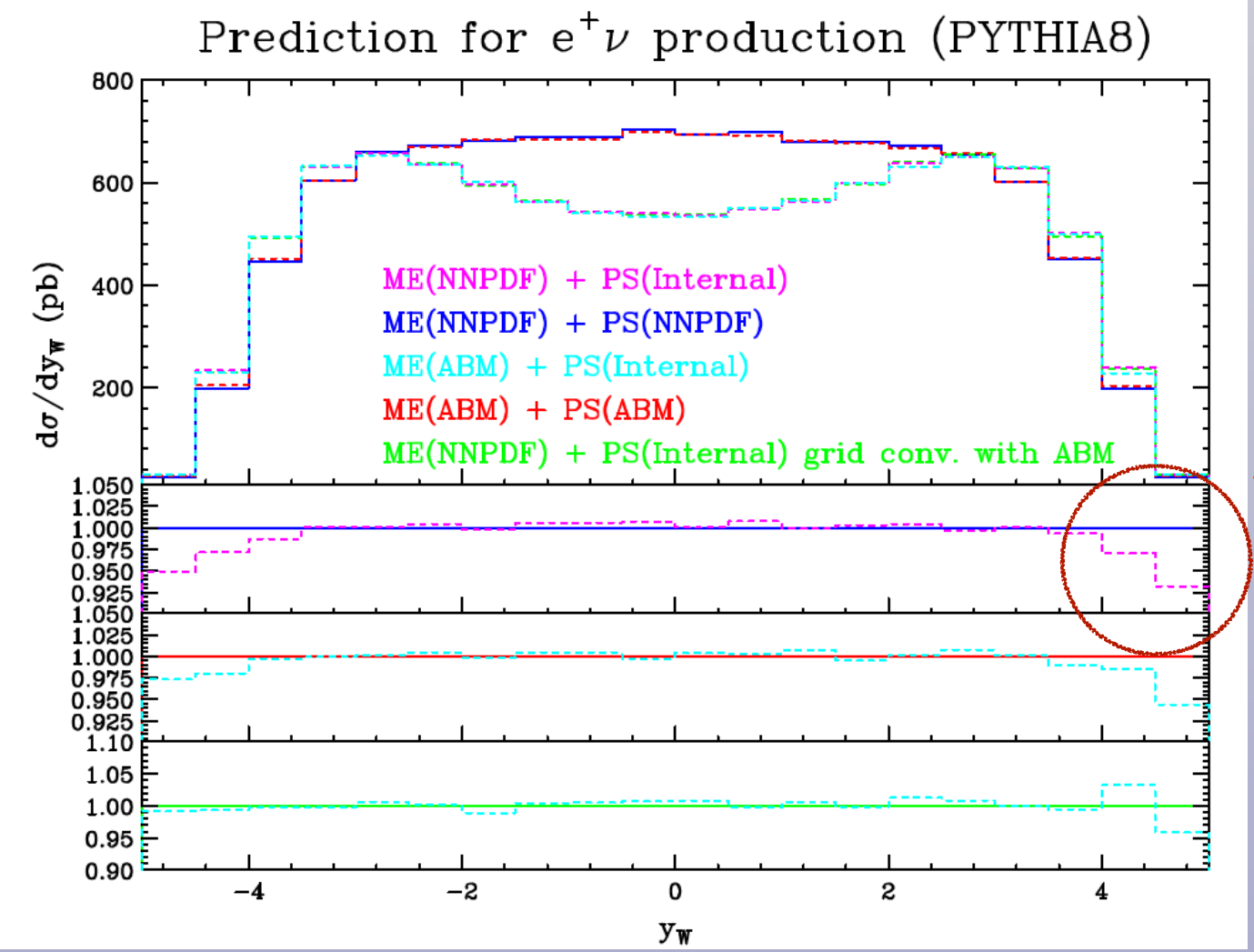
Skands, Carrazza, J.R. 14

PDF fits at NLO+PS accuracy

- **NLO+PS** is current standard for LHC event simulation, and **improves in many directions over fixed-order NLO results**: improved pert. behaviour, direct relation with measured quantities, less need for kin cuts …

- Using **NLO+PS calculations** in global PDF fits should have many important applications, like for the **W mass** among others, and is now technically possible thanks to **aMCfast,** the fast interface to **MadGraph5_aMC@NLO** based on the **applgrid** library

aMCfast: Bertone, Frixione, Frederix, J.R., Sutton, arXiv:1406.7693 (for NLO), NLO+PS in preparation



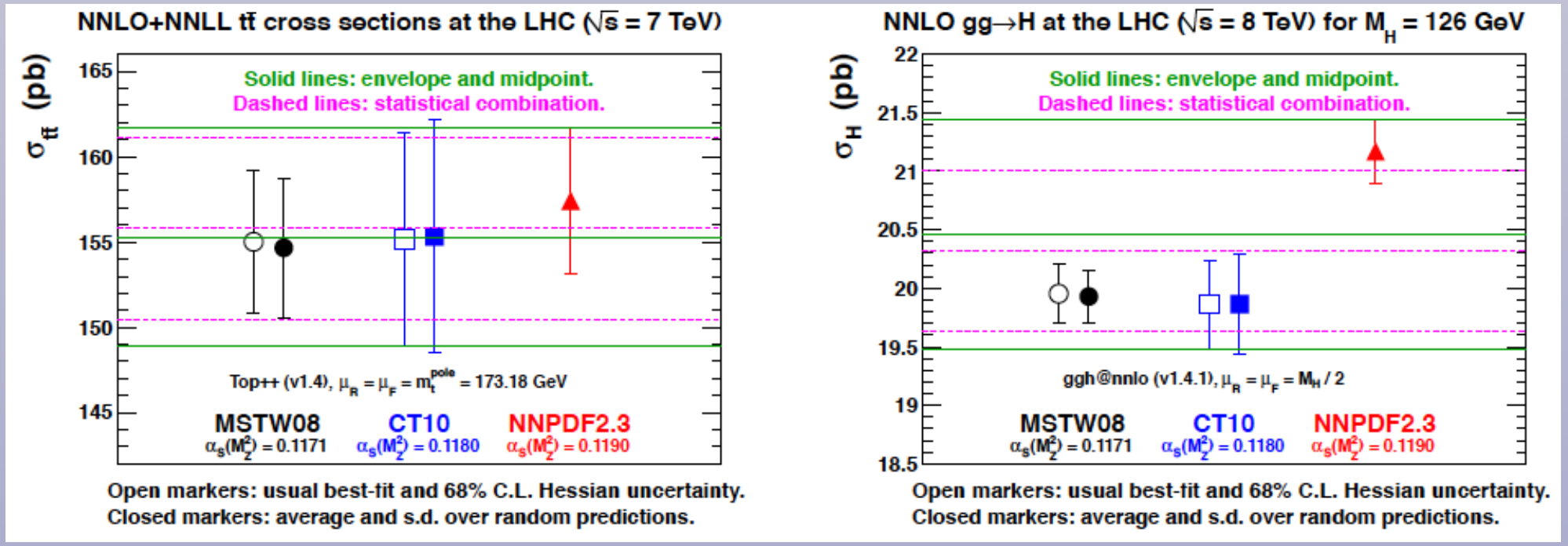
- One crucial aspect to explore is the **role of the PDF used by the MC shower**, since this is fixed even in the fast NLO+PS grid

- Quite small effect in most observables, except **extreme kinematics** like forward rapidities

aMCfast makes possible to include easily **hadron-level measurements** directly into PDF fits

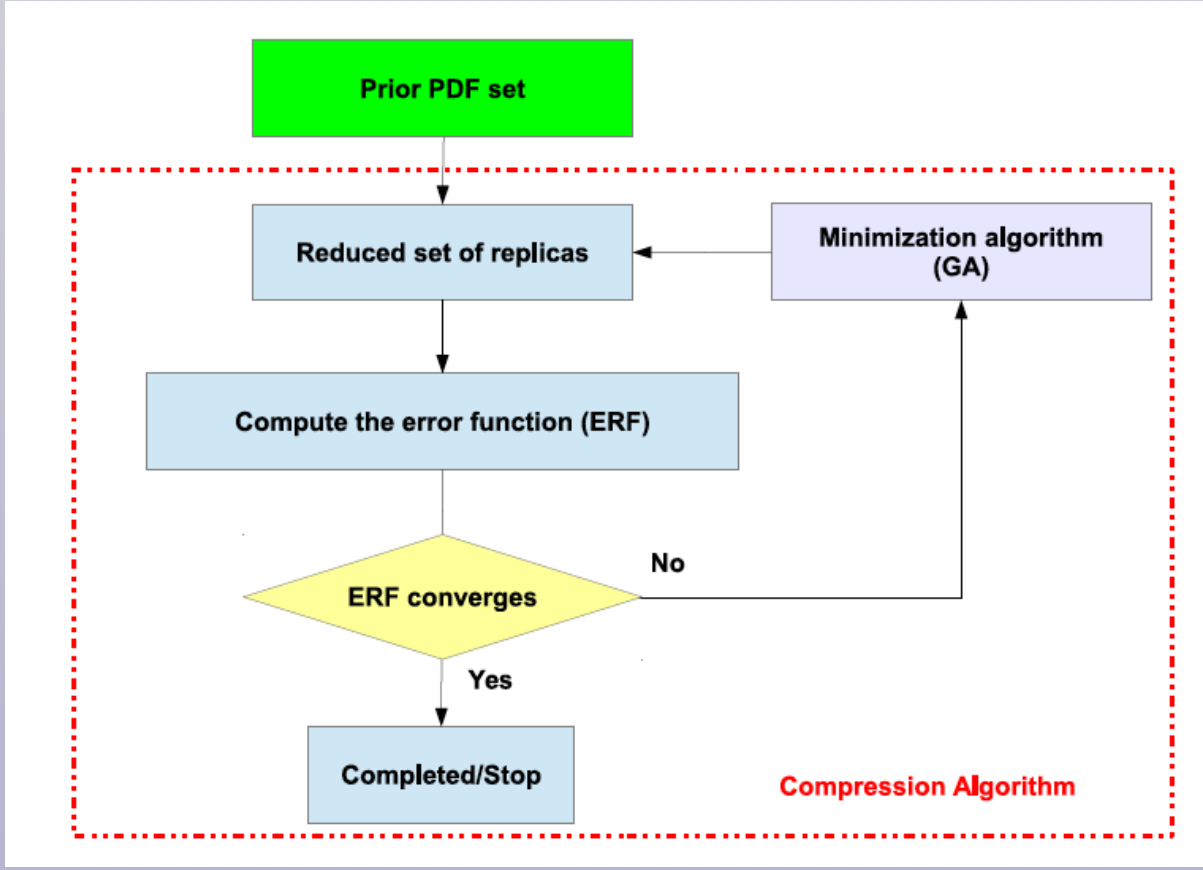
Compressed Monte Carlo PDFs

- Motivation: provide a practical implementation of the **PDF4LHC recommendation**, **easy to use** by the experiments and **computationally less intensive** that the original prescription
- Having a **single combined PDF sets** (even with large number of eigenvector/replicas) would already be useful since widely-used tools like MadGraph5_aMC@NLO, POWHEG or FEWZ provide the **PDF uncertainties without any additional cost**
- But this is not true for all theory tools used at the LHC, so there is still a good motivation to be able to use a combined PDF set with a **small number of eigenvectors/replicas**
- CMC-PDFs based on the **Monte Carlo statistical combination** of different PDF sets, followed by a **compression algorithm** to end up with a reduced number of replicas
- The Monte Carlo combination has a **robust statistical interpretation,** and in many cases leads to similar results, with somewhat smaller uncertainties, compared to the **original PDF4LHC envelope**.



Basic strategy

- The MC combination of PDF sets is easy, but **number of MC replicas still too large**
- **Compress the original probability distribution** to one with a **smaller number of replicas**, in a way that all the relevant estimators (mean, variances, correlations etc) for the PDFs are reproduced



- The compression is applied at **Q = 2 GeV,** though the results are robust wrt other choices
- Various options about how the **error function** to be minimised can be defined, ie., to reproduce central values add a term

$$ERF_{CV} = \frac{1}{N_{\rm CV}} \sum_{i=-n_f}^{n_f} \sum_{j=1}^{N_x} \left(\frac{f_i^{\rm CV}(x_j,Q) - g_i^{\rm CV}(x_j,Q)}{g_i^{\rm CV}(x_j,Q)} \right)^2$$

$$f_i^{\rm CV}(x_j,Q) = \frac{1}{N_{\rm rep}} \sum_{r=1}^{N_{\rm rep}} f_i^r(x_j,Q)$$

- Same for variances, correlations and higher moments
- At the end, **optimal choice** decided by the resulting phenomenology

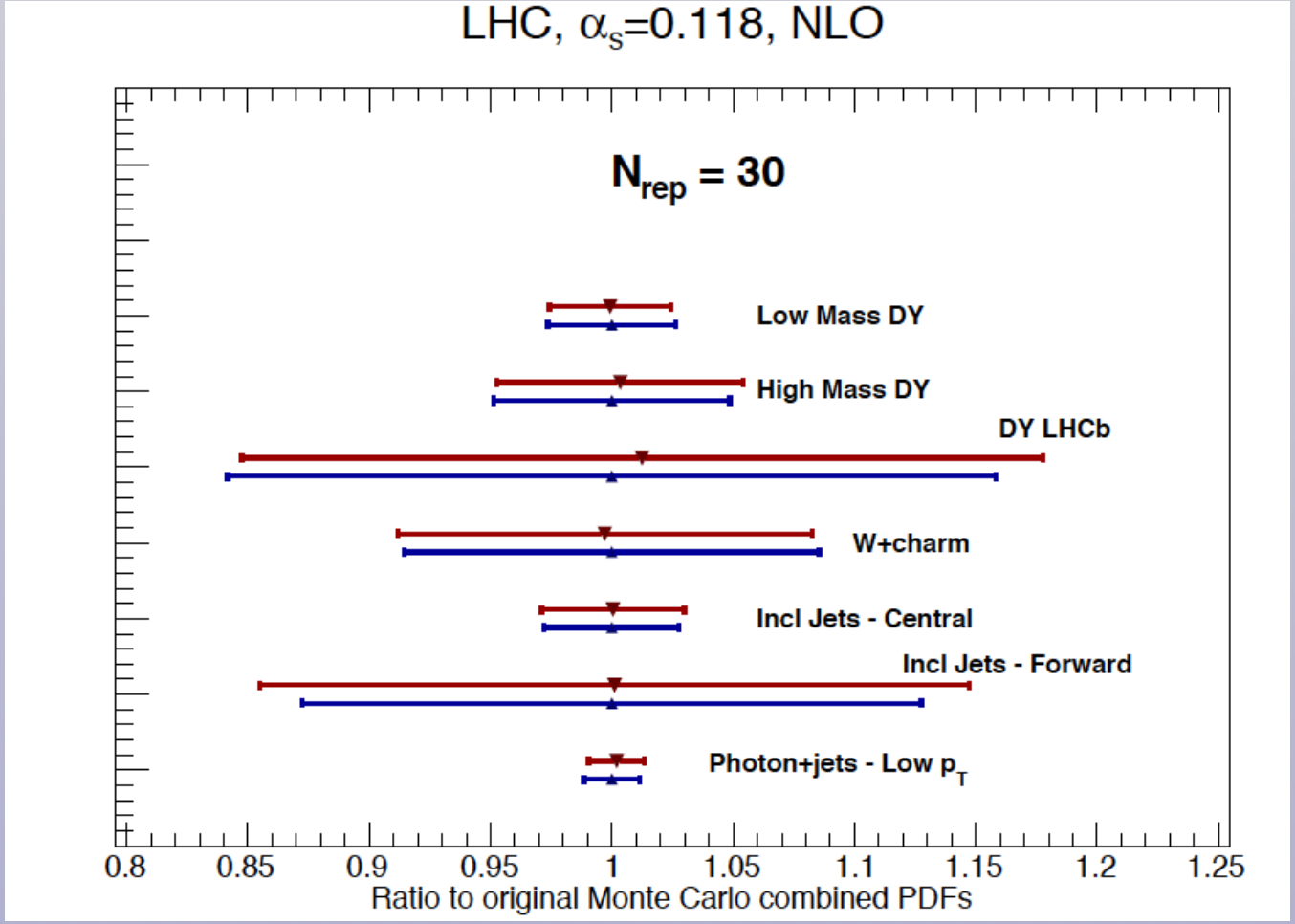
$$ERF_{KOL} = \frac{1}{N_{KOL}} \sum_{i=-n_f}^{n_f} \sum_{j=1}^{N_x} \sum_{k=1}^{6} \left(\frac{F_i^k(x_j,Q) - G_i^k(x_j,Q)}{G_i^k(x_j,Q)} \right)^2$$

$$N_{KOL} = \frac{1}{N_{\rm rand}} \sum_{d=1}^{N_{\rm rand}} \sum_{i=-n_f}^{n_f} \sum_{j=1}^{N_x} \sum_{k=1}^{6} \left(\frac{R_i^k(x_j,Q) - G_i^k(x_j,Q)}{G_i^k(x_j,Q)} \right)^2$$

- The algorithm also minimises the **Kolmogorov distance** between the original and compressed distributions

LHC Phenomenology

- The ultimate validation is of course to check that the **compressed set reproduces the original combination** for a wide variety of observables
- We have tested a very large number of processes, both at the **inclusive and differential level** and always found that $\mathbf{N_{rep}=20\text{-}30}$ replicas are enough for **phenomenology**



Compression also works for **fully differential distributions**

Tested on a large number of processes: jets, Drell-Yan, WW, W+charm, Z+jets, ….

Calculations use **fast NLO interfaces:**

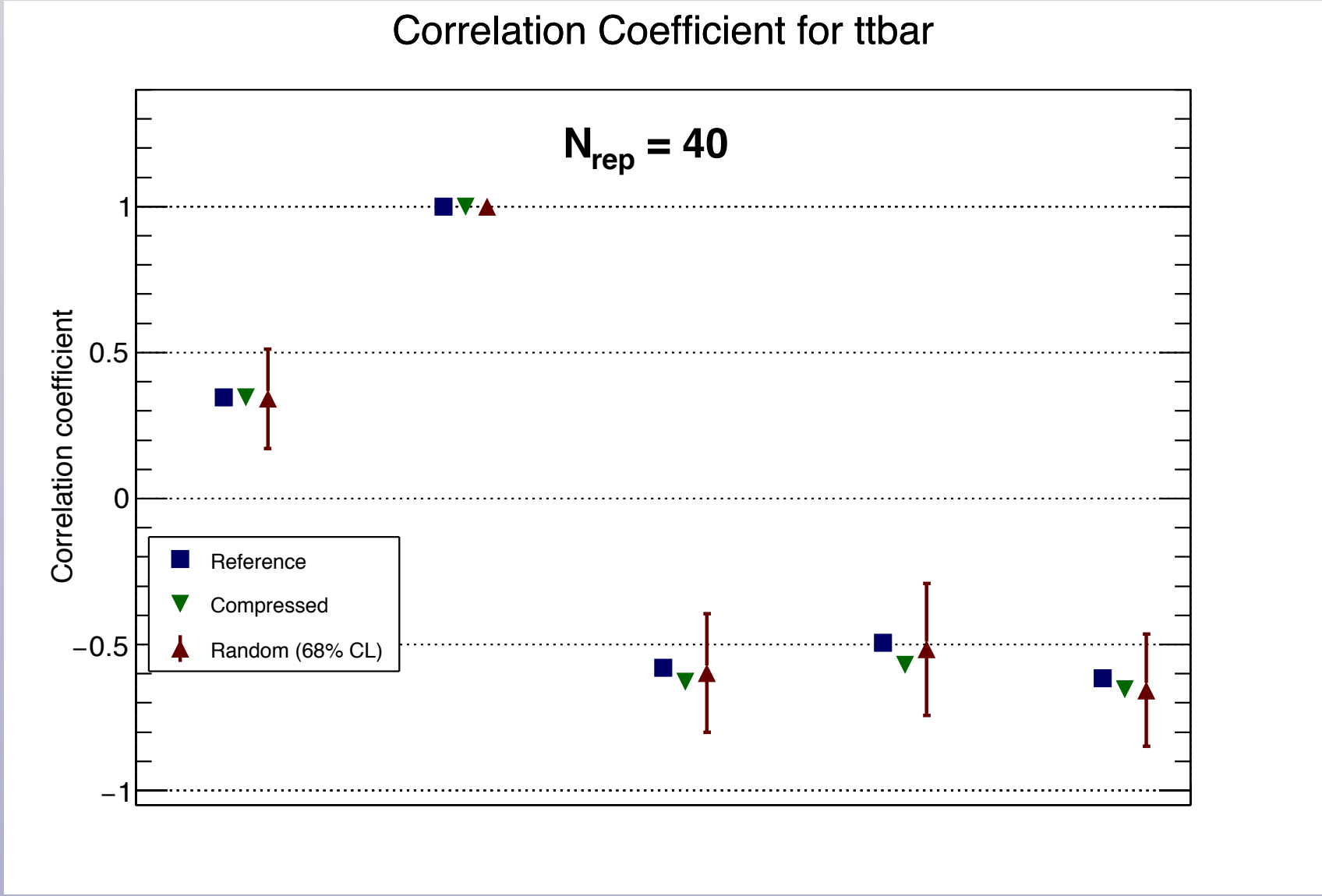
1. **aMCfast/applgrid** for **MadGraph5_aMC@NLO**
2. **applgrid** for **MCFM/ NLOjet++**

Very flexible to **redo validation** for any other compressed set

CMC-PDFs: Carrazza, Latorre, J.R. Watt, in prep

Correlations

- Agreement between original and CMC-PDFs also holds at the level of **cross-section correlations**
- Not an accident: s**electing replicas at random fails** to reproduce the correlation accurately enough



MC2Hessian

- For many important applications, a **linearised Hessian version of Monte Carlo sets**, based on **orthogonal eigenvectors**, would be useful
- A **MC2Hessian algorithm** would allow to use **NNPDF** and **CMC-PDF sets** for **profiling**, as **nuisance parameters**, construct sets with reduced number of eigenvectors for specific applications like **W mass** ….
- …. while keeping also the crucial possibility of t**esting for the potential deviations from Gaussianity** of the underlying probability distribution, **quantifying the range of validity of linear approximation**
- Various options possible, for example a la **Meta-PDF,** fit a **functional form** to each of the MC replicas

$$f(x,Q_0;\{a\}) = e^{a_1}x^{a_2}(1-x)^{a_3}e^{\sum_{i\geq 4}a_i\left[T_{i-3}(y(x))-1\right]}.$$

- But this is subject to the usual **functional bias**.
- A more robust choice is to use also **MC replicas as linear expansion basis:**

$$g^{(k)}(x,Q^2) = \sum_{i=1}^{\widetilde{N}_{\rm rep}} a_i^{(k)}\widetilde{g}^{(i)}(x,Q^2)$$

- The eigenvectors can then be determined from the a **χ^2 defined in the space of PDFs:**

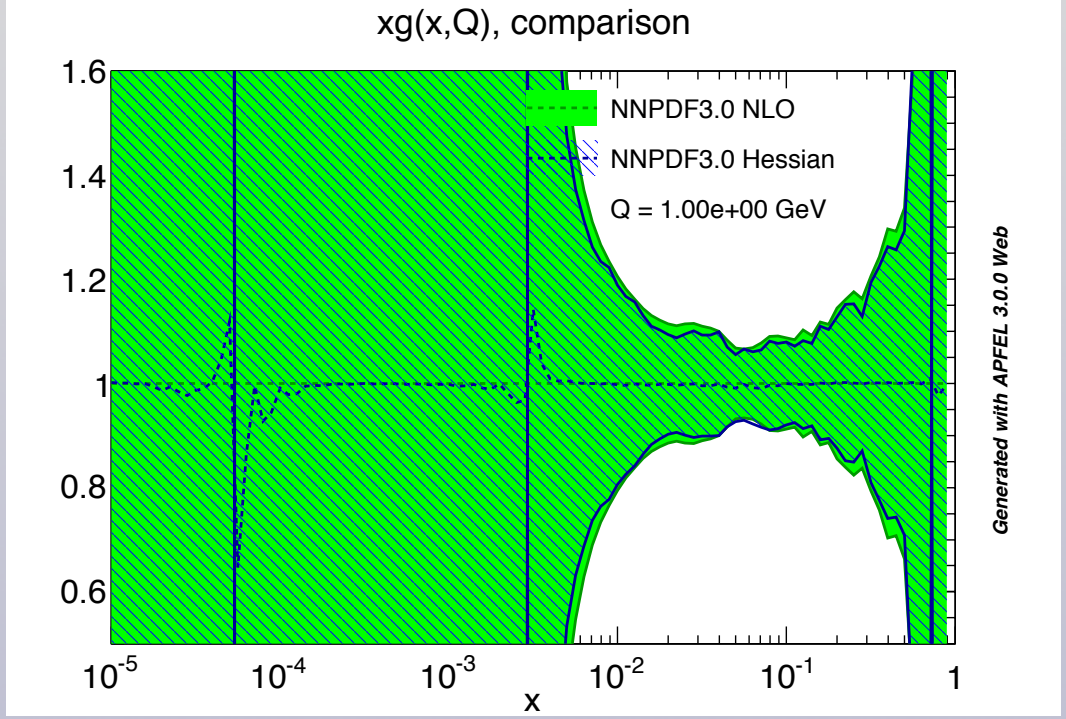
$$\chi^2_{\rm pdf} \equiv \sum_{x=1}^{N_x}\left(g^{(k)}(x_i,Q_0^2) - \left\langle g^{(k)}(x_i,Q_0^2)\right\rangle\right)(\mathrm{cov}_{ij})^{-1}\left(g^{(k)}(x_j,Q_0^2) - \left\langle g^{(k)}(x_j,Q_0^2)\right\rangle\right)$$

Covariance Matrix in the space of PDFs

MC2Hessian - Preliminary results

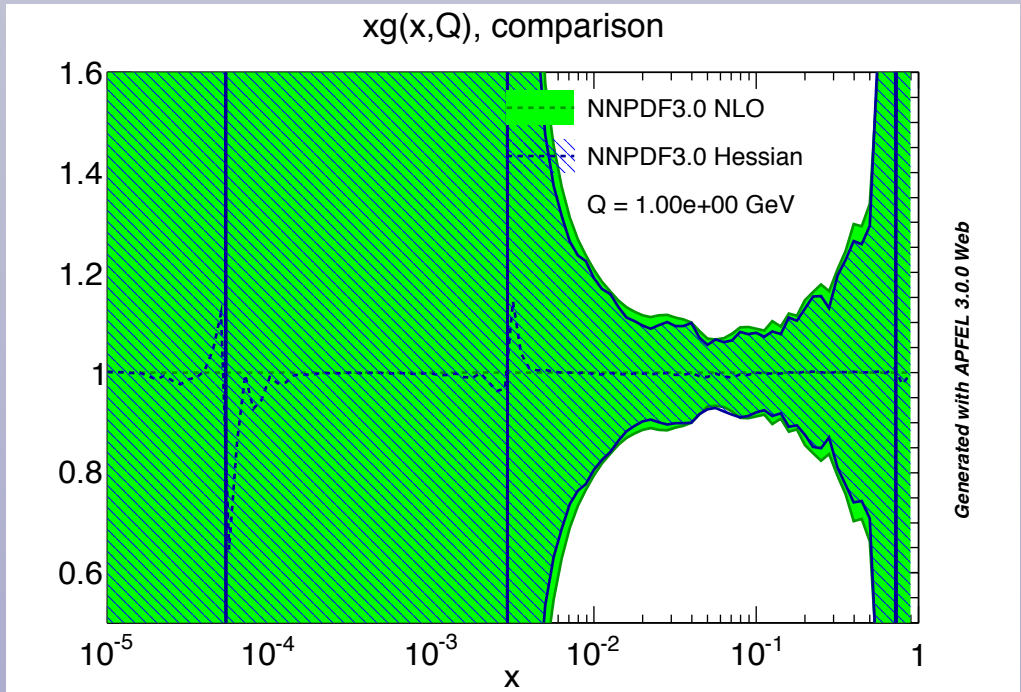
- Preliminary results validate this strategy: it is possible to **efficiently construct a Hessian representation of NNPDF3.0,** or in general of any **MC PDF set**

- In particular, the **CMC-PDFs** will also be available in a **Hessian** representation



- The comparison between the **original MC representation** of a PDF set and its **Hessian representation** allow to determine the **range of validity** of the latter

- For instance, in important cases like **BSM searches at high-mass**, it is know that the **Gaussian approximation** is not adequate



The NNPDF (Future?) Timeline

☑ **NNPDF3.0Hessian** using the **MC2Hessian** conversion algorithm

☑ **NNPDF3.1:** includes **legacy HERA data**, **LHC data** on jets, top quark differential distributions, W,Z rapidity, Drell-Yan. **MSbar running masses**, provide **range of m_{charm} and m_{bottom} variations**, implications at the LHC

☑ **NNPDF3.x with large-x resummation**

☑ **NNPDF3.x with small-x resummation**

☑ **NNPDF3.x for NLO event generators** using **aMCfast** and **aMC@NLO**

☑ **NNPDF3.x with intrinsic charm**

☑ **NNPDF3.xQED**: precision determination of the **photon PDF** from LHC data. Also **EWK corrections** systematically included in all LHC measurements with aMCfast

☑ Also: determination of the strong coupling, large-x PDFs and impact for searches, comparisons with non-perturbative models of the proton structure …..

☑ **NNPDF4.0 (?)** with theoretical uncertainties on PDFs …….