PDF4LHC15 combined PDFs, compressed PDFs, and LHAPDF

Andy Buckley for the PDF4LHC authors
University of Glasgow

ISMD 2015, Wildbad Kreuth, Oct 2015

Introduction

This talk (mostly) on behalf of the PDF4LHC15 group.
They did the work, not me!

Talk outline:

- a potted history of the notorious PDF4LHC procedure
- recent developments in the recommendation ⇒ PDF4LHC15 PDFs
- construction of "compressed PDFs"
- *status of LHAPDF*

Credit: most of what I show will be "Frankenstein'd" from Juan Rojo & Albert de Roeck at QCD@LHC last month!

PDF4LHC uncertainty

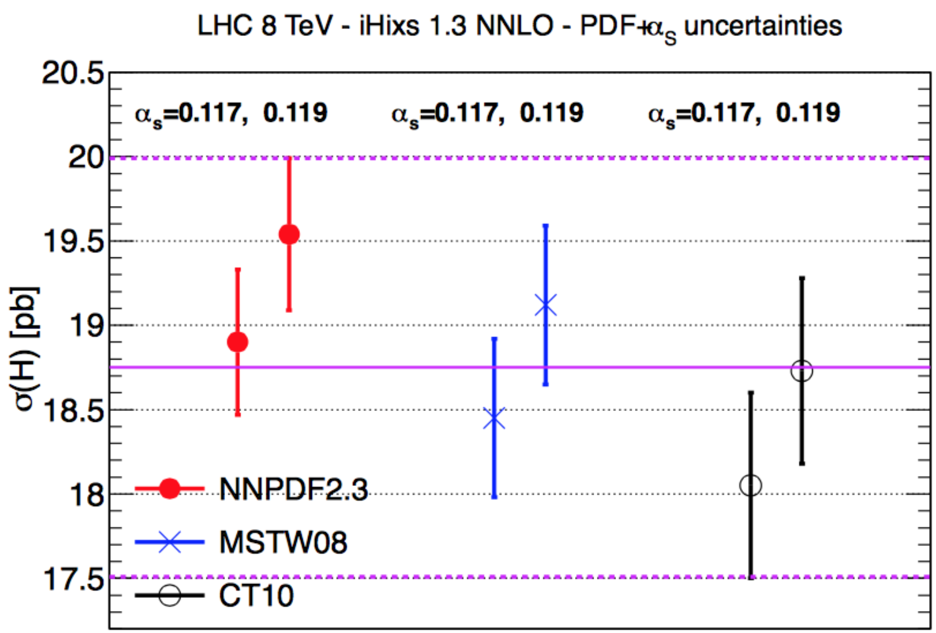
PDF4LHC…

- PDF4LHC originated from the HERA/LHC workshops organized in the period 2004-2008
- Forum for PDF discussions with all PDF fitting groups and experiments
- First recommendation in 2010: arXiv:1101.0538 (under some duress of the Higgs cross section group ☺)
- This recommendation (see later) was a bit tedious: revised and being improved now…
- New recommendation being finalized, which should be a easier for the user, without (noticable) loss of precision

2010 Recommendation

THE OLD PDF4LHC RECOMMENDATION

HIGGS IN GLUON FUSION



Using the full envelope was conservative but statically not optimal

- TAKE THE ENVELOPE OF RESULTS
- RED CONSERVATIVE WAY OF HANDLING DISCREPANCIES
- GGH: DISCREPANCY NOT UNDERSTOOD DESPITE INTENSIVE BENCHMARKING

Usages of PDFs

Note: basically two kinds of usages:

- For the assessment of the PDF uncertainty in comparing predictions to theory (top cross section, jets, W-rapidity distributions)

  Use the individual PDF sets (CT, NNPDF, MMHT, HERAPDF, ABM, JR, CJ…)

- For assessment of the PDF uncertainty as in searches, discovery, acceptance, measurements (eg Higgs, SUSY, …)

  Use the PDF4LHC prescription

History of the PDF4LHC Recommendation

- 2010: take the envelope of the 3 (most) global fits to available data: MSTW08, CTEQ6.6, NNPDF2.0

  Of these, only MSTW08 had an NNLO version
- In 2013 this was updated by taking the latest versions of the PDFs: CT10, MSTW2008 and NNPDF2.3

  All have NNLO versions
- A lot of extensive benchmarking of these and other PDF sets was done as input to these studies eg arXiv:1101.0536
- More details can be found on the PDF4LHC web

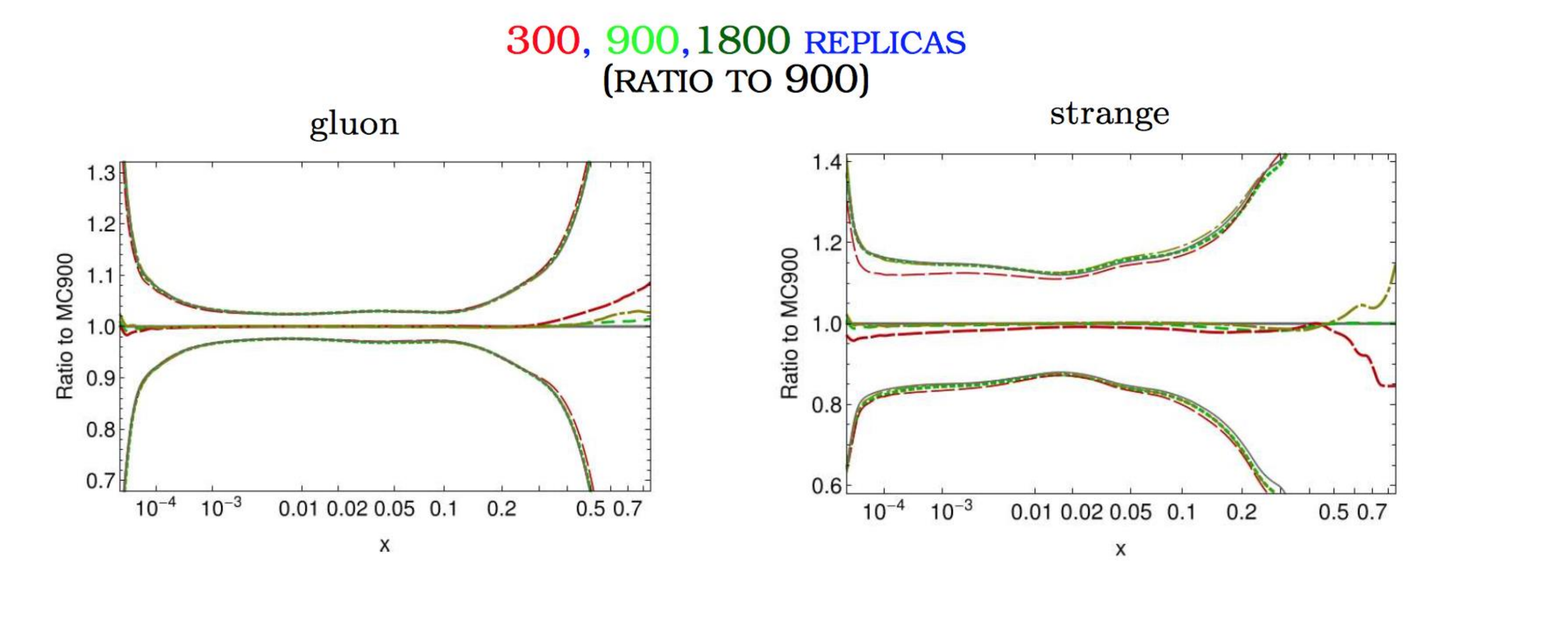
  http://www.hep.ucl.ac.uk/pdf4lhc/
- The making of the envelope was left to the user. The new recommendation will be more “user friendly”

Updated Recommendation

- Criticism: use of only 3 PDF (motivated in arXiv:1101.0538)
- Criticism: complicated for the user, having to make the envelope for each observable. Some signal programs are not adapted well to do that in an efficient way.
- Criticism: not statistically rigorous/robust
- Response: Use the transition from Run-I to Run-II for a more statistically rigorous procedure. Ideas developed in o.a. Forte arXiv:1011.5247, G Watt PDF4LHC meeting 4/2013
- Idea: create a "combined-PDF" set based on N input PDFs with its uncertainty band, using the MC replica method. This is what the user can use directly. Correlations between the PDFs should be kept. The user should have a simple procedure.
- Compress the set to reduce the number of replicas

Current Combination of PDFs

The present combination is based on Monte Carlo set combination of CT14, MMHT and NNPDF3.0
A total of 900 replicas (300 per PDF set) ensures a % level of accuracy on all quantities. We use this as a baseline



MC replicas and reduction of the replicas and getting Hessian representation discussed by J. Rojo in the previous talk

The New PDF4LHC Prescription

- Perform a Monte Carlo Combination of the underlying PDF sets (see NNPDF)
- Sets entering the combination must satisfy common requirements
- Deliver combined PDF sets in Hessian and MC representation.
- Monte Carlo: a reduced set of PDF replicas is delivered
- Hessian: central set and error sets are delivered
- α_s : PDFs are delivered for a central value of α_s of and for 1-sigma error.
- Hence the user gets combined sets he/she can directly use
- This scheme also allows to include more PDF families

Requirements for Inclusion

- PDF sets to be included in the combination must satisfy requirements which make the compatible
- Details are on the indico of the PDF4LHC workshop of April
- A few examples (see document for the full list)
  - Provide results at NLO and NNLO with NNLO α_s of evolution benchmarked against HOPPET and QCDNUM
  - Baseline set for $\alpha_s = 0.118$ and sets with α_s variation
  - Use a general-mass variable flavor number scheme with up to 5 flavors
  - For direct combination with current technology fits must be based on approximately common global sets of data. Reweighting will be necessary otherwise.
- HERAPDFs expressed interest to be included

Three Sets Proposed

From the user-community we also got feedback for the need for different sets for different applications… Hence we provide three sets

- **PDF4LHC15_mc:** a Monte Carlo PDF set with $N_{\mathrm{rep}} = 100$ replicas.
- **PDF4LHC15_30:** a symmetric Hessian PDF set with $N_{\mathrm{eig}} = 30$ eigenvectors.
- **PDF4LHC15_100:** a symmetric Hessian PDF set with $N_{\mathrm{eig}} = 100$ eigenvectors.

- These are available at NLO and NNLO and for standard $\alpha_s = 0.118$ and $\delta\alpha_s = \pm 0.0015$ variations
- Available in LHAPDF repository

- Which ones to use when: Part of the recommendation being documented eg
  - _MC set: contains also non-gaussian effects. Important for searches at high masses. ..
  - _30 set: very good precision and probably usable for most experimental needs
  - _100 superior precision, eg for TH Higgs cross section uncertainty determinations…

PDF compression methods

Outline

To estimate the **total PDF uncertainty** associated to LHC processes, such as Higgs cross-sections or New Physics searches, a **combination of the results from individual PDF sets** is required

PDF4LHC11 → **Envelope Method**

- Cumbersome to implement: **a posteriori combination** of cross-sections from different PDF sets
- **Statistical interpretation** unclear. Over-conservative: too much weight given to **outliers**

PDF4LHC15 → **Monte Carlo Combination**

- Statistically meaningful combination

MC reduction → **CMC-PDFs**

Hessian reductions → **MC2H**, **Meta-PDFs**

MC2H → **SM-PDFs**

Combination tailored to **specific processes**

Monte Carlo Combination of PDF sets

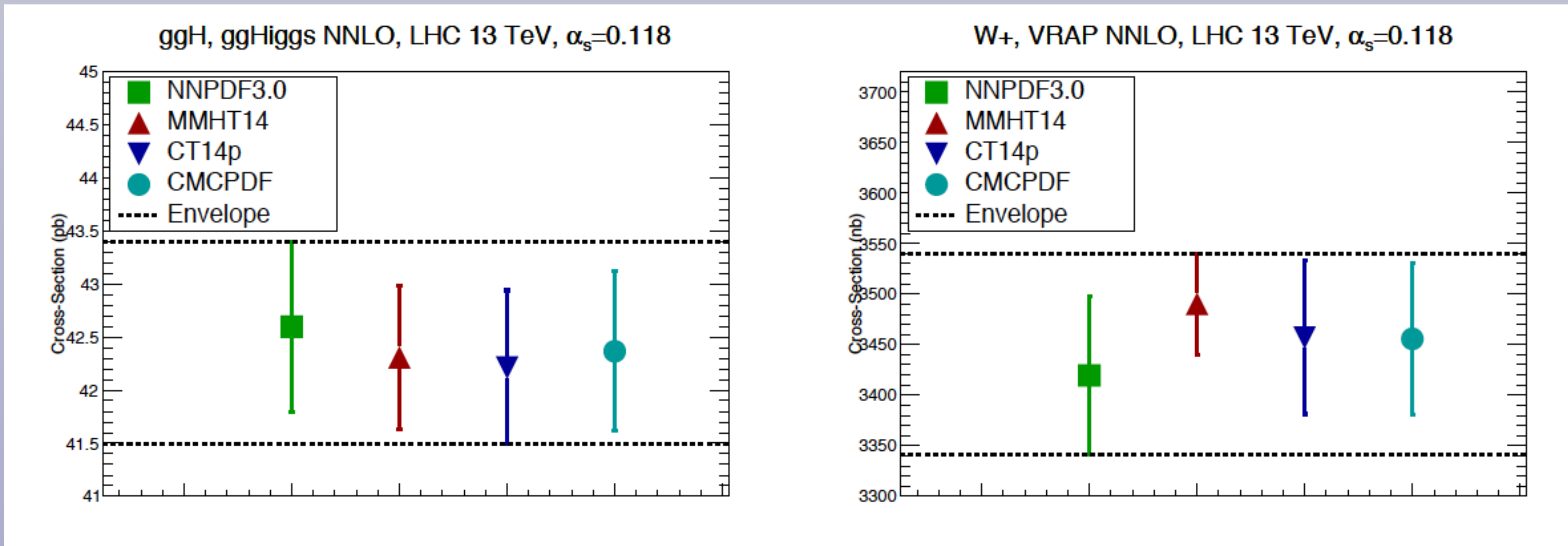
Watt and Thorne, arXiv:1205.4024
Forte, arXiv:1011.5247
Forte and Watt, arXiv:1301.6754
Carrazza, Latorre, J. R., Watt, arXiv:1504.06459

Monte Carlo combination

- First of all, select the **PDF sets that enter the combination**. Must be reasonably consistent among them for a meaningful combination
- Transform the **Hessian PDF sets into their Monte Carlo representation:**

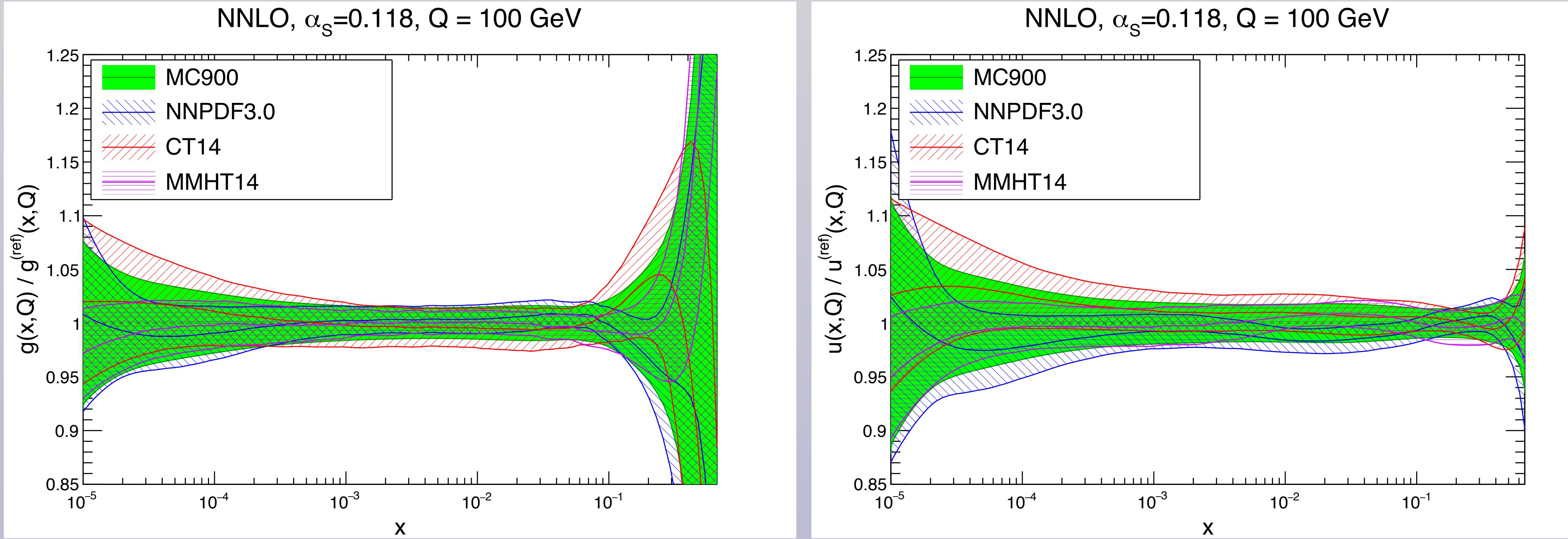
$$F^k = F(q_0) + \frac{1}{2}\sum_{j=1}^{N_{\rm eig}} \left[F(q_j^+) - F(q_j^-) \right] R_j^k \,, \quad k = 1, \ldots, N_{\rm rep} \,,$$

- Assume equal likelihood for all input sets: combine the **same number of replicas**
- The resulting Monte Carlo ensemble has a **robust statistical interpretation.** Similar results, with smaller uncertainties, compared to the **original PDF4LHC envelope** (proper treatment of outliers)



The combined Monte Carlo PDF set

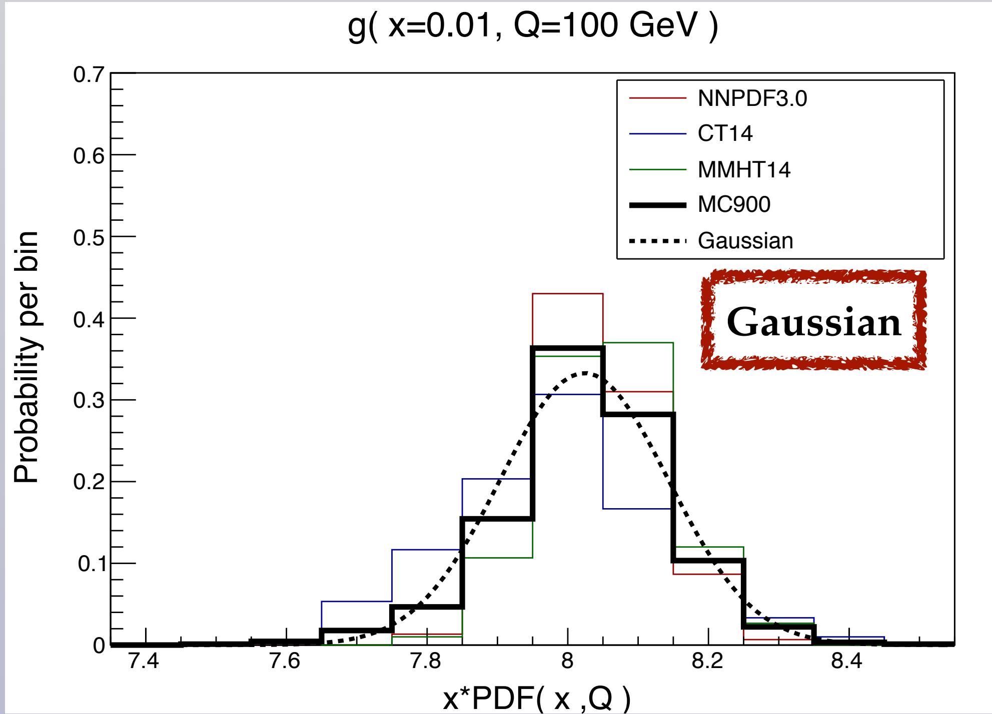
- Combine most recent releases of the three global sets: **NNPDF3.0, CT14, MMHT14**

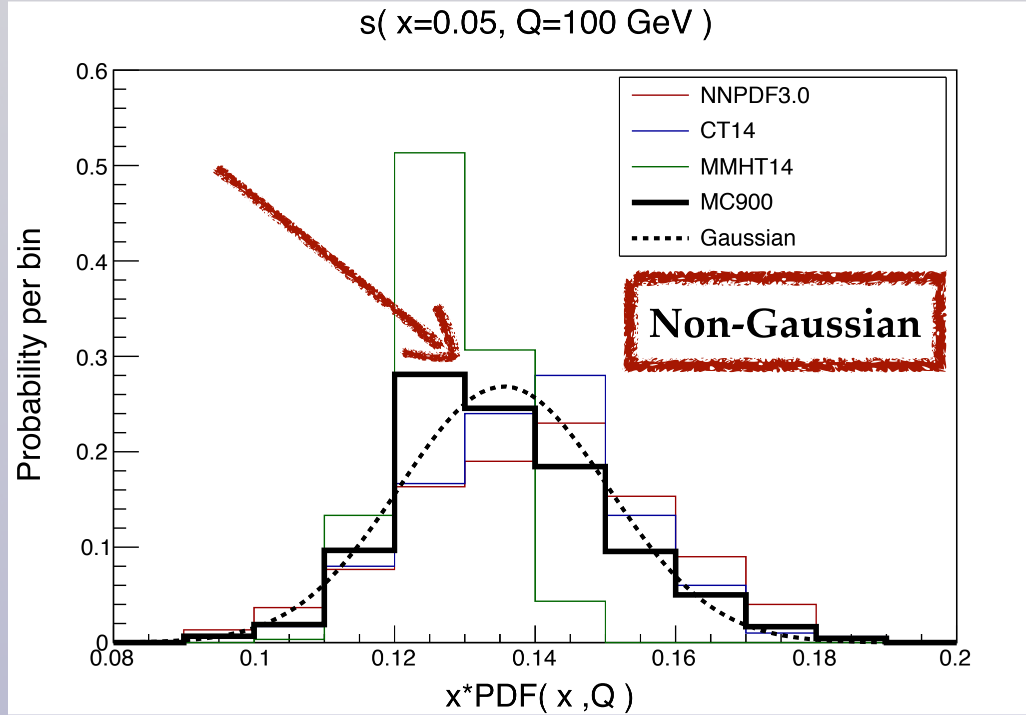


- **900 MC replicas** required to stabilise the combination: these define the **MC900 prior**
- Too many for practical applications: need to find a strategy to **reduce the number of replicas** in which the MC combination is based
- In addition, for many important application, a **Hessian representation** would be required

The combined Monte Carlo PDF set

- Combine most recent releases of the three global sets: **NNPDF3.0, CT14, MMHT14**





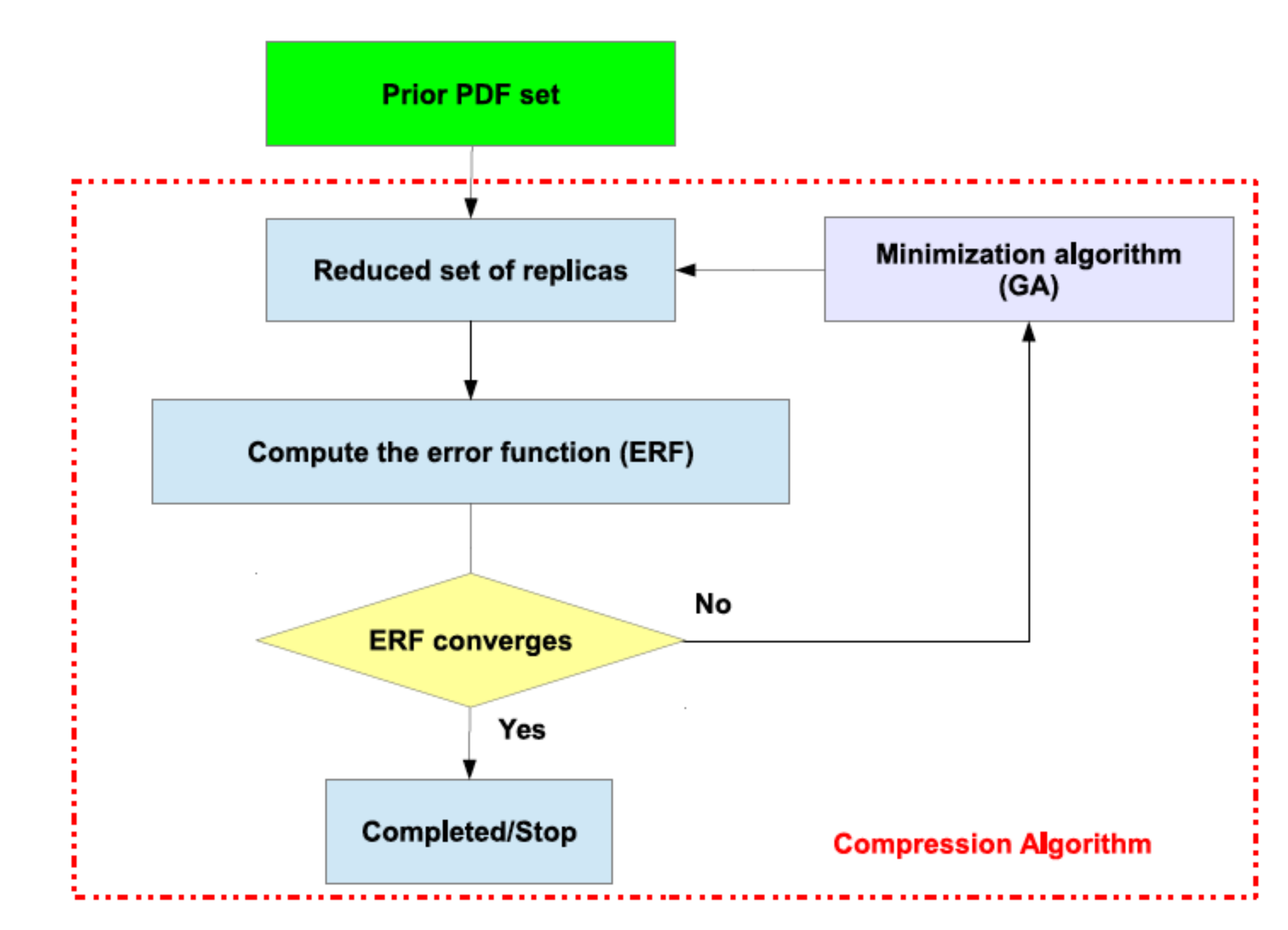
- The **MC combination** is usually **Gaussian** (in the data region) but in many cases **non-Gaussian features** are observed (in extrapolation regions, or for poorly constrained flavors)
- Specially important for **BSM searches**, which rely on PDFs in regions where **PDF errors are large**

CMC-PDFs

Carrazza, Latorre, J. R., Watt, arXiv:1504.06459

Compressed Monte Carlo PDFs

Compress the MC900 prior down to a **smaller number of replicas**, in a way that all the relevant estimators (mean, variances, correlations, higher moments) for the PDFs are fully reproduced



$$\mathrm{ERF} = \sum_k \frac{1}{N_k} \sum_i \left(\frac{C_i^{(k)} - O_i^{(k)}}{O_i^{(k)}} \right)^2 ,$$

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

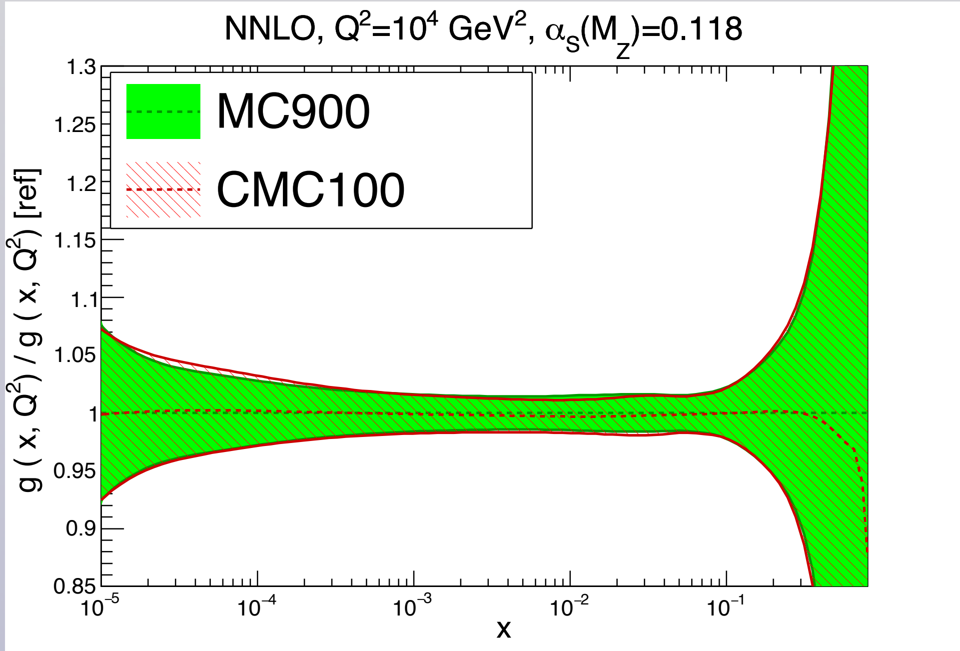
$$f_i^{\mathrm{CV}}(x_j, Q) = \frac{1}{N_{\mathrm{rep}}} \sum_{r=1}^{N_{\mathrm{rep}}} f_i^r(x_j, Q)$$

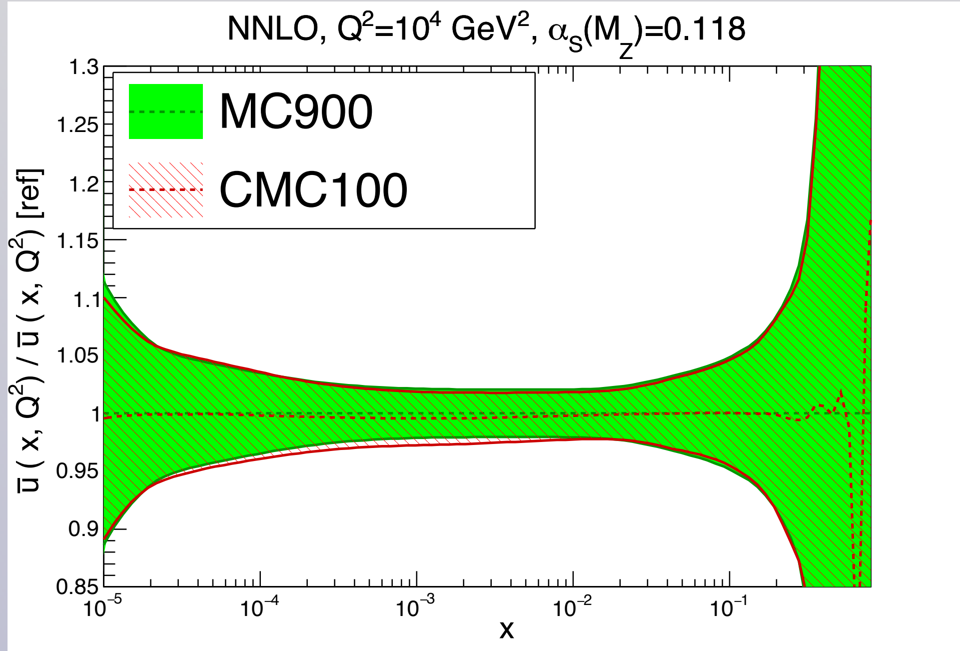
Similar for **variances, correlations, skewness, kurtosis** and the **Kolmogorov** distance

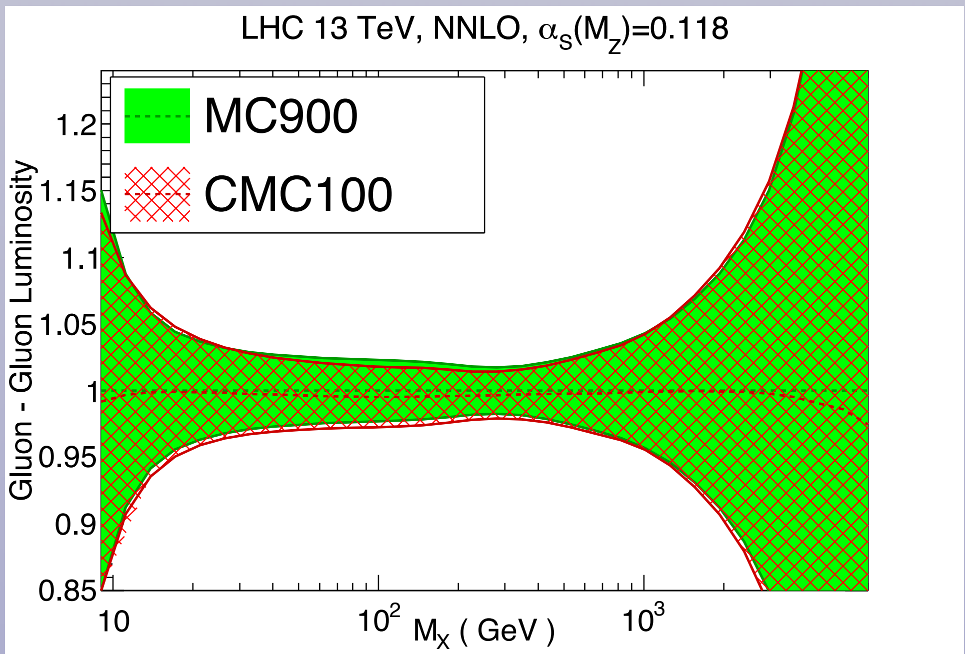
Minimise **distance between prior and compressed sets** using Genetic Algorithms applied to an **Error Function,** which ensures the reproduction of all statistical properties of the prior

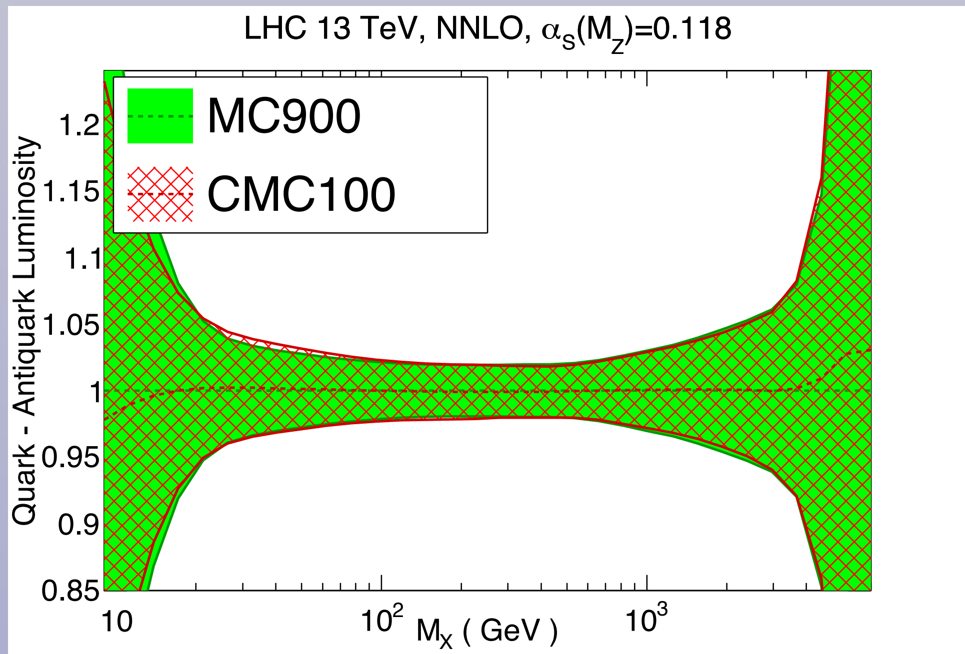
CMC-PDFs

Good agreement between **MC900** and **CMC** from a number of compressed replicas **N$_{rep}$ >50**



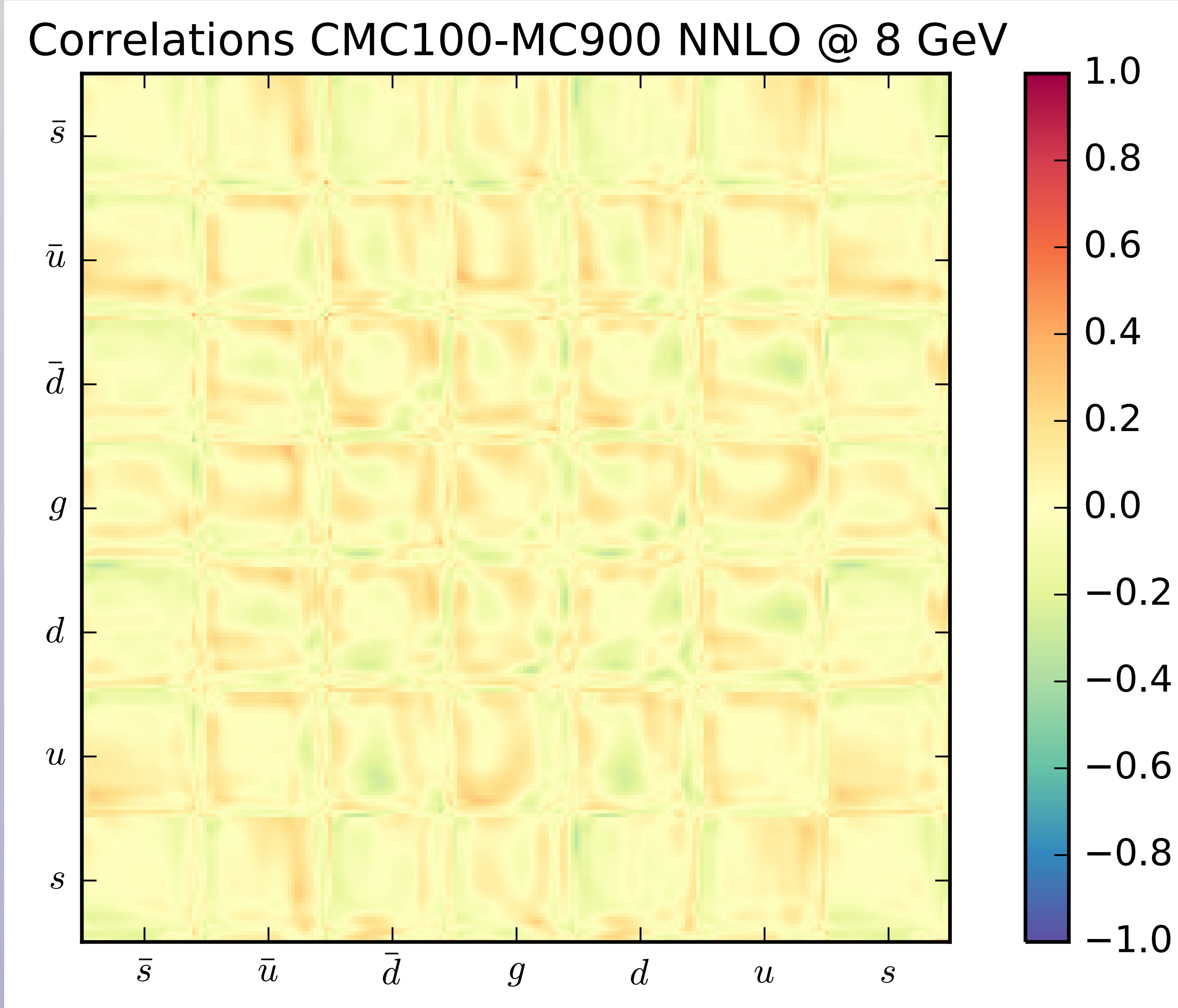






CMC-PDFs

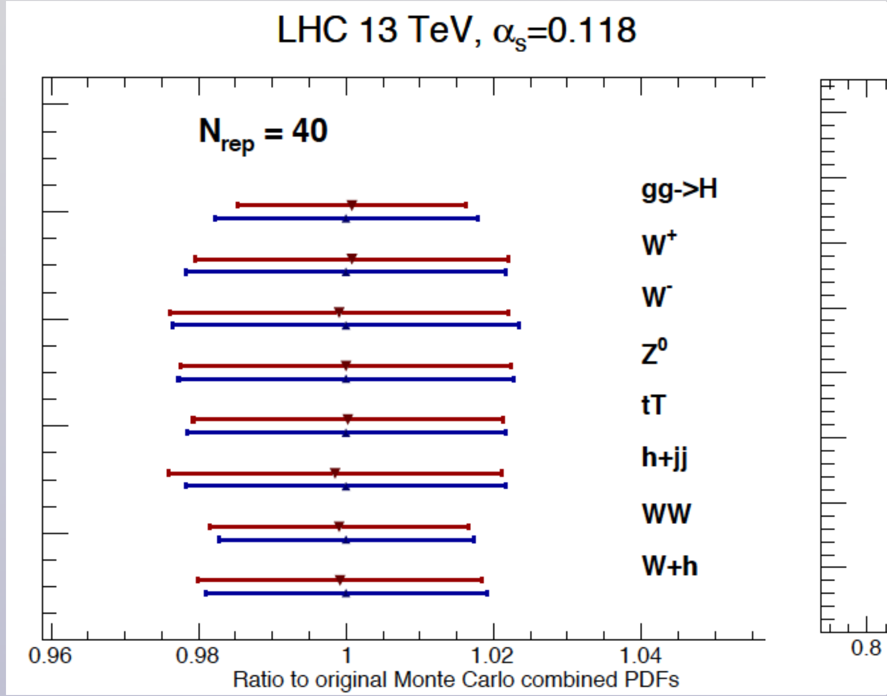
Reasonable agreement as well for the **correlations between different PDF flavours**

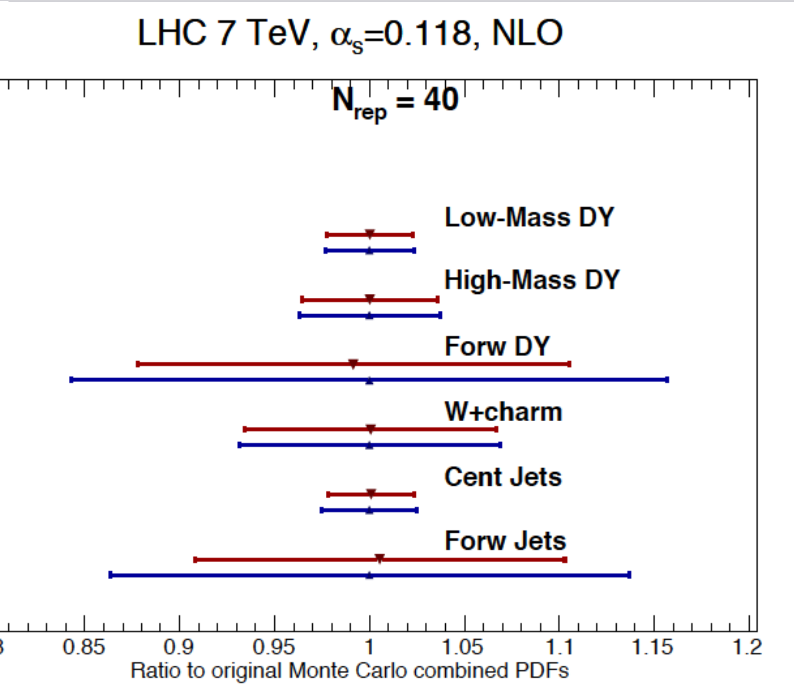


In this contour plot, the **differences in the correlation coefficients of MC900 and CMC100** between different PDF flavours are represented

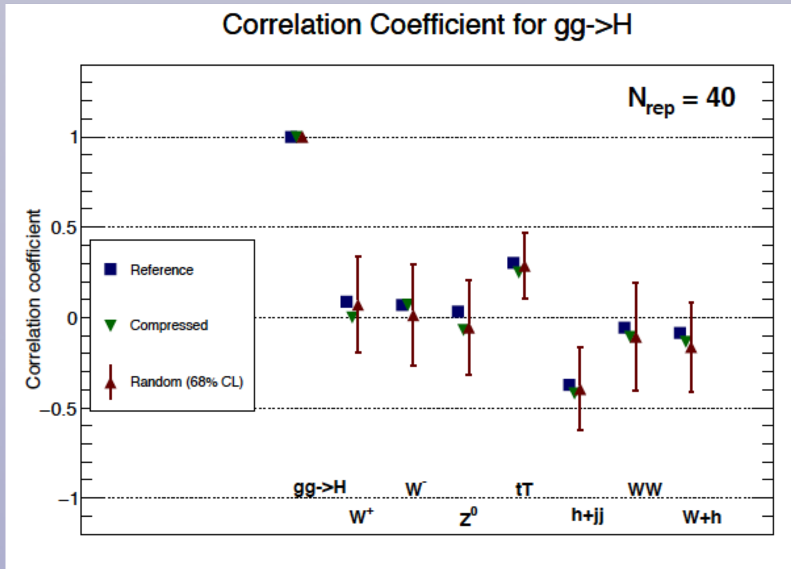
Correlation coefficients are computed at **Q = 8 GeV**, and in the **range of x of [10^{-5},0.9]** logarithmically spaced

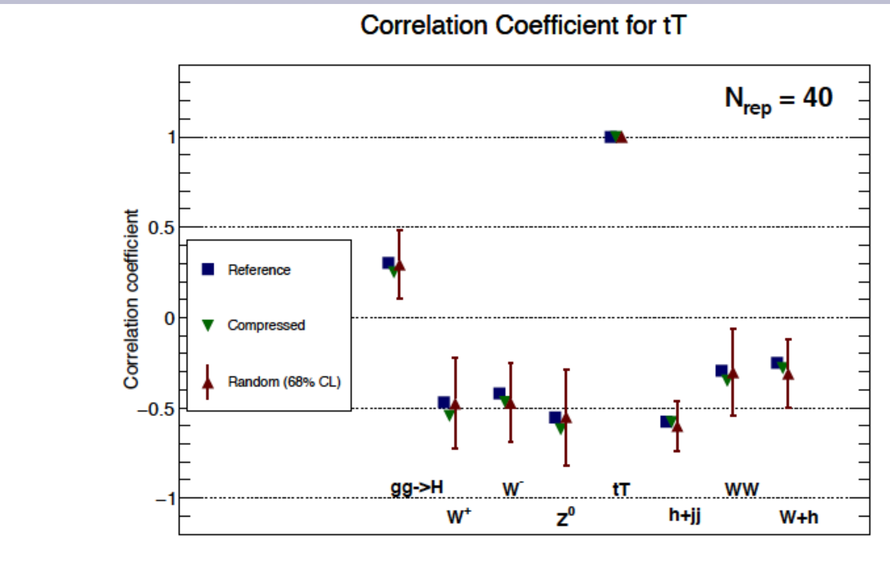
LHC Phenomenology





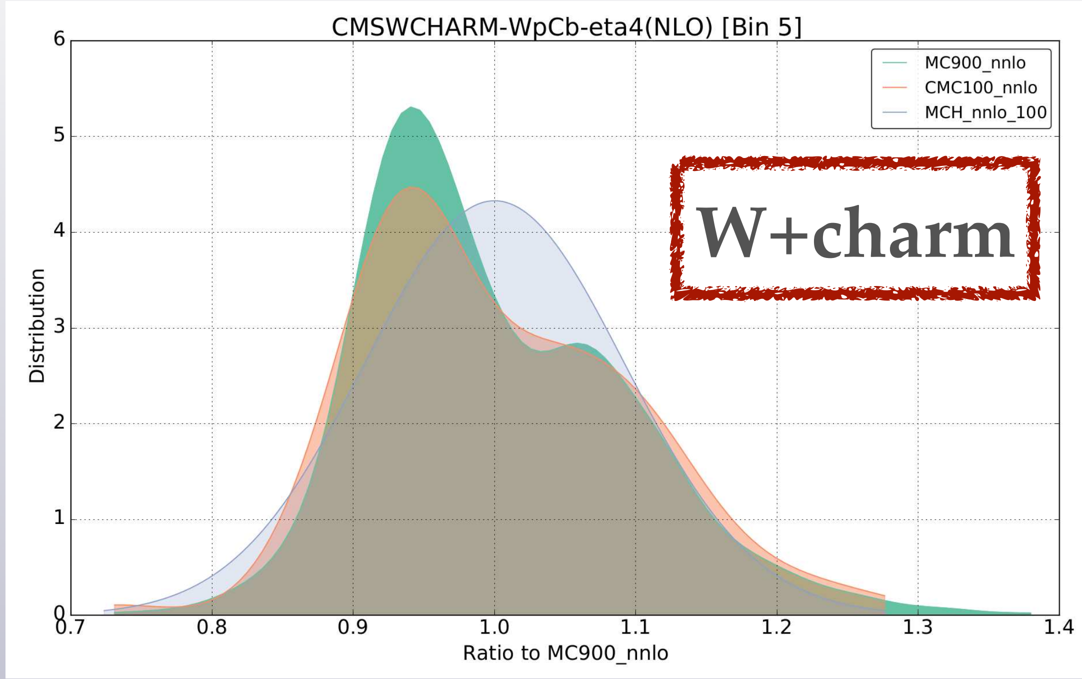
As expected from good agreement at the PDF level, CMC-PDFs also validated for **LHC inclusive cross-sections and diff distributions,** including **correlations**



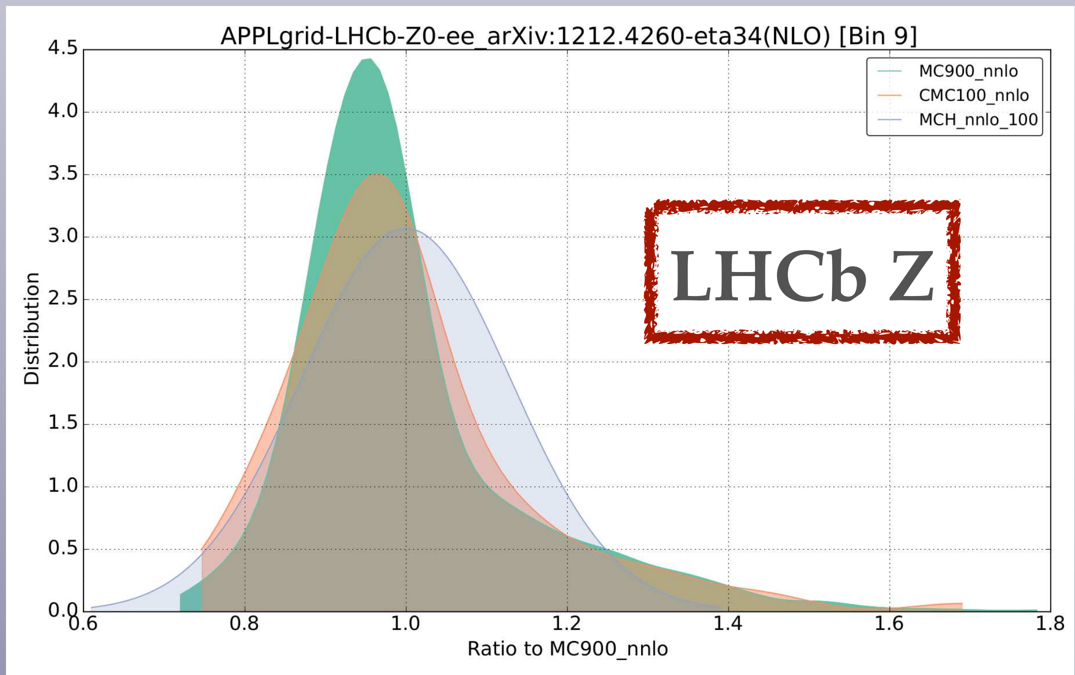


Agreement between MC900 and CMC significantly better than **random selection of same number of replicas**

Non-Gaussian features in LHC cross-sections



CMC100 reproduces properly **non-Gaussian features in MC900**, such as a **double-hump distribution** (in W+charm) and a **skewed distribution** (for forward Z)



Any **Hessian reduction fails** by construction to reproduce such features mc²hessian

MC2Hessian and Meta-PDFs

Meta-PDFs: Gao and Nadolsky, arXiv:1401.0013

MC2H: Carrazza, Forte, Kassabov, Latorre and J. R., arXiv:1505.06736

Motivation for Hessian reduction strategies

- In regions where **MC900** is approximately Gaussian, a **Hessian representation** is by definition more efficient than **CMC-PDFs**, since it needs to reproduce **only central values and covariances**
- Moreover, a Hessian representation has several advantages, such as the use of **PDF uncertainties as nuisance parameters**, the applicability of **Hessian profiling**, or the **possibility of further reductions** when applied to specific processes
- Two Hessian reduction methods have been developed. Essentially common idea, differ in the choice of **linear expansion basis**

MC900 → **Hessian reductions** → **MC2H**

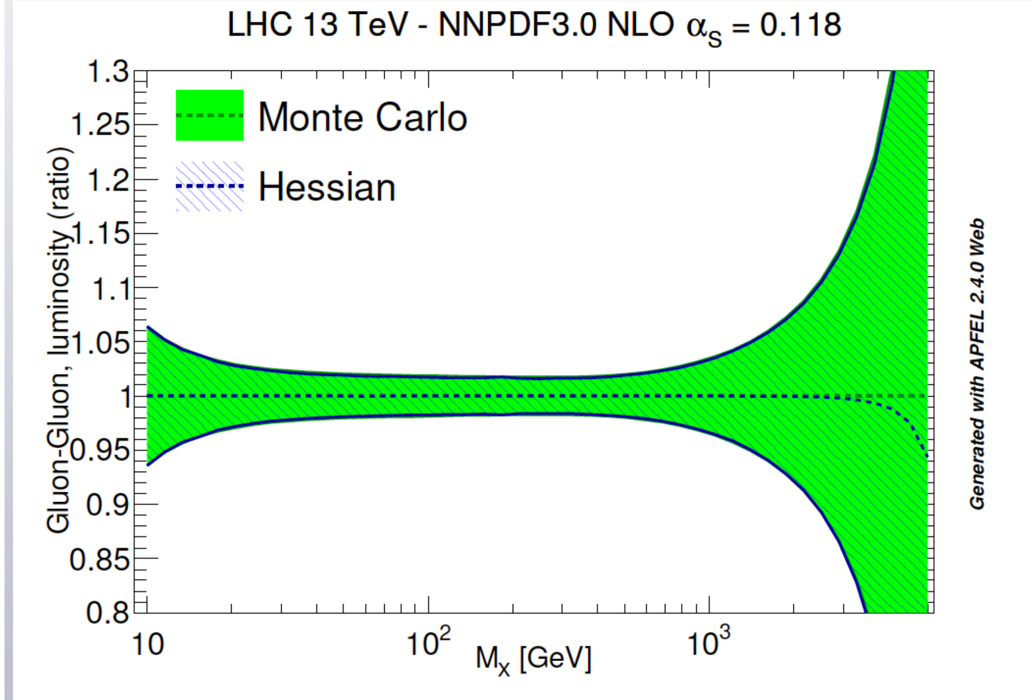
- **MC replicas** as expansion basis
- **Principal Component Analysis (PCA)** to reproduce PDF covariance matrix with arbitrary precision
- Standard linear algebra problem: **Singular Value Decomposition**

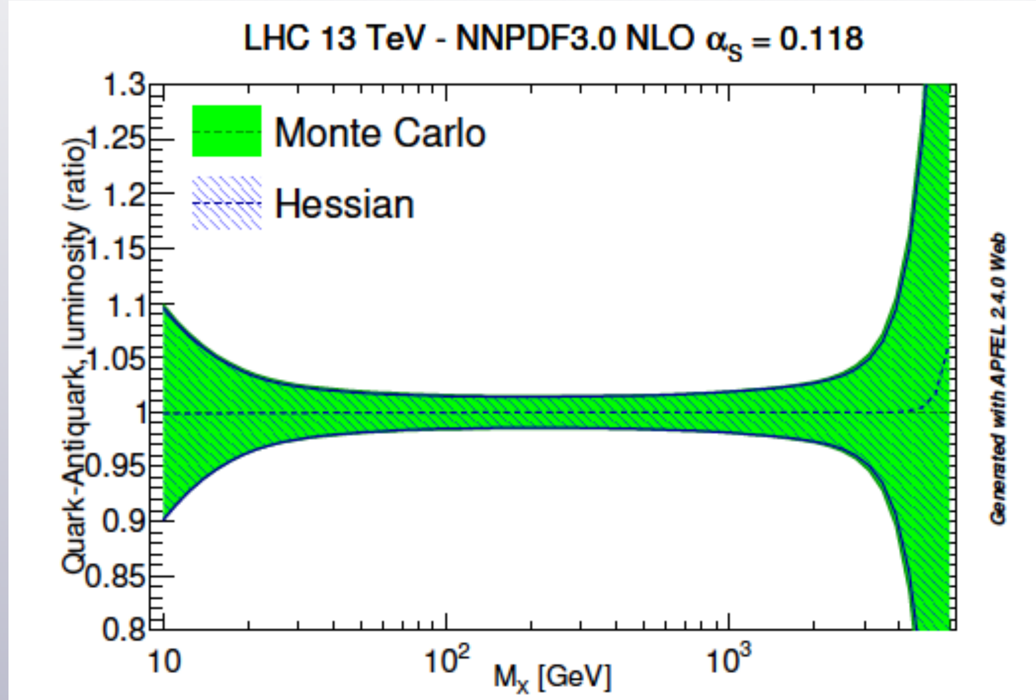
MC900 → **Hessian reductions** → **Meta-PDFs**

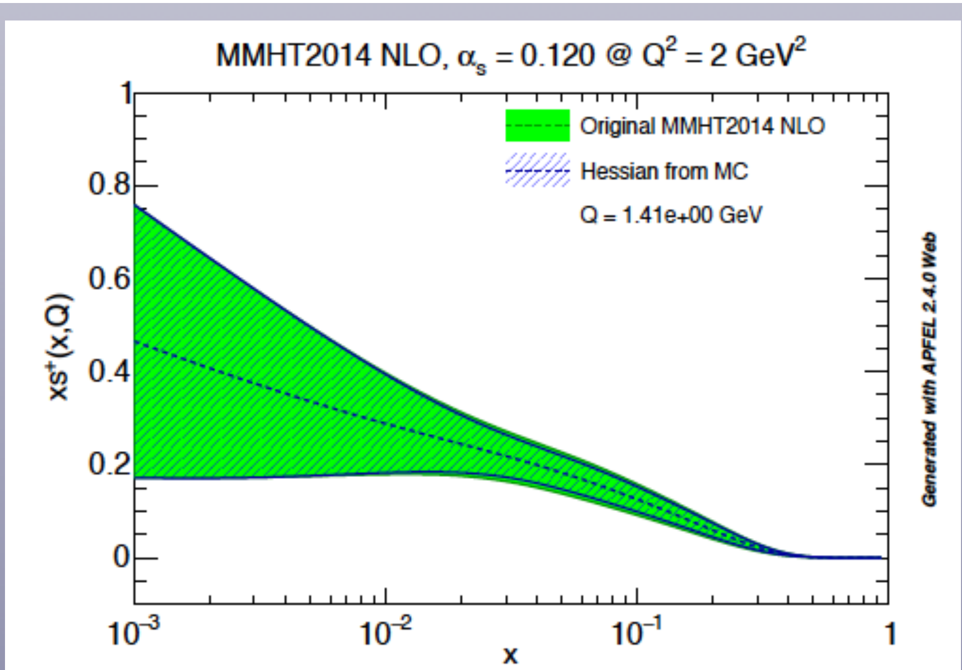
- **MC replicas** are fitted to a common **meta-parametrization**
- **MC900** defines the **68% CL of the parameters** of the fitted function form
- Linear expansion in these fit parameters

Results: NNPDF3.0 and MMHT14

- The MC2H algorithm succeeds in **producing a Hessian representation of a native MC set,** NNPDF3.0

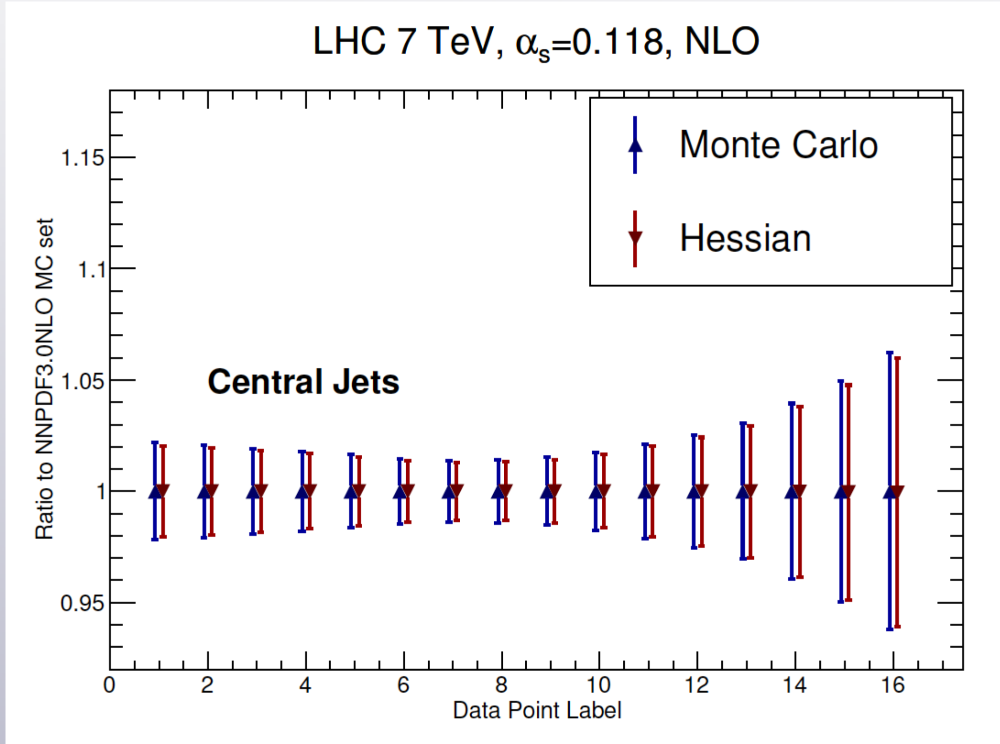


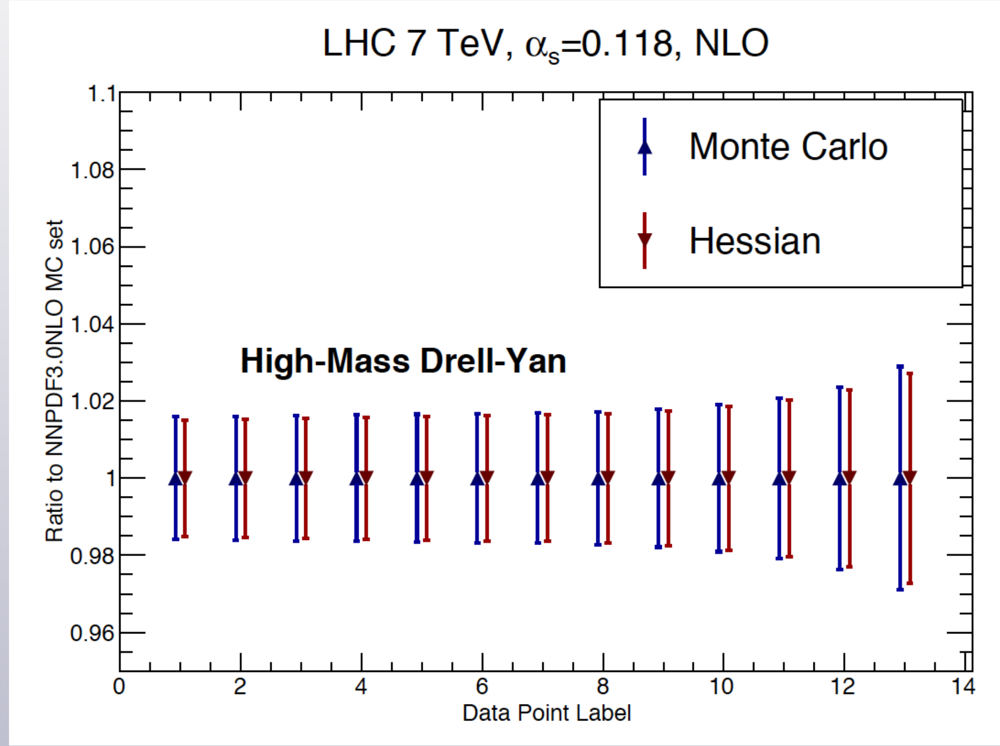


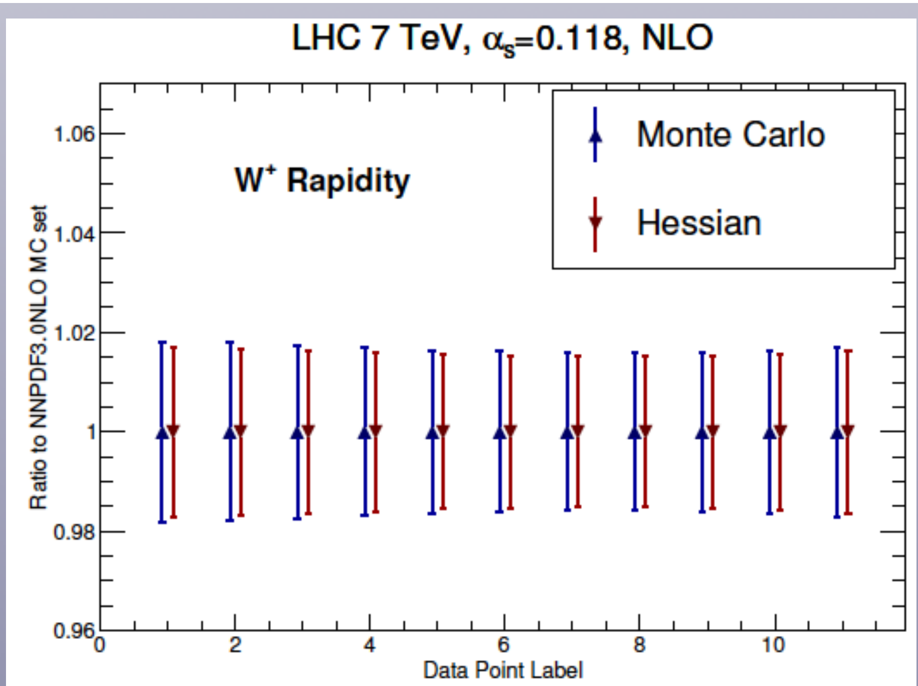


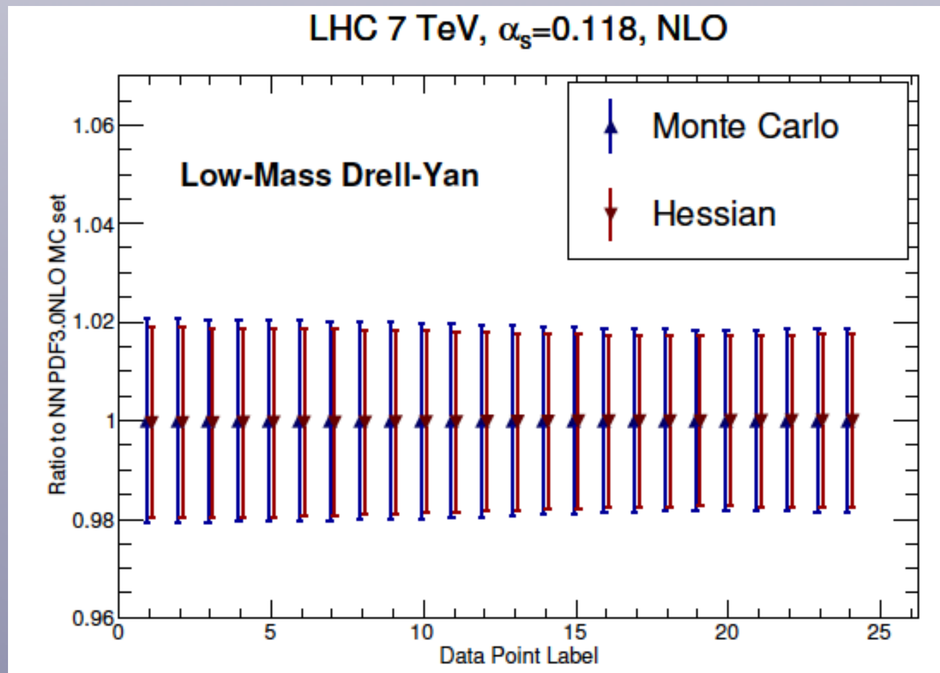
- Also verified that we can produce a **Hessian representation of a native Hessian set, MMH14**, via an **intermediate Monte Carlo representation**

MC2H LHC phenomenology



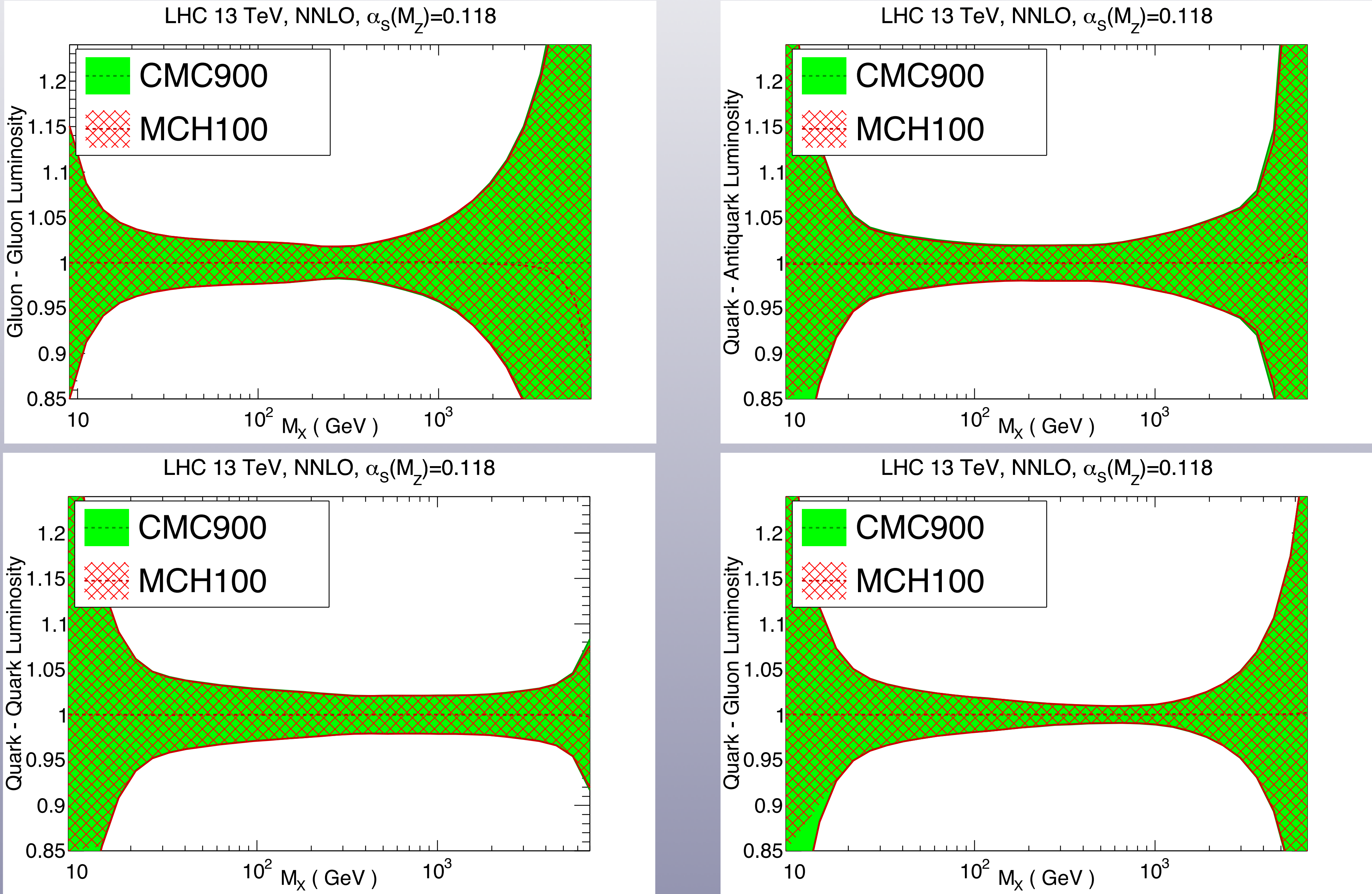






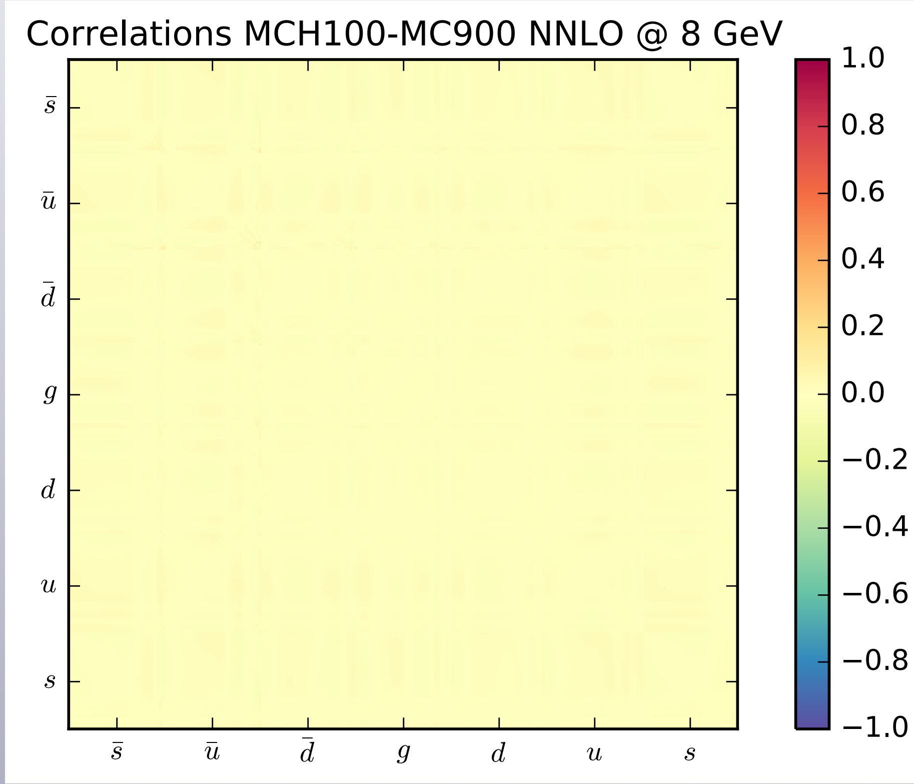
MC2H from MC900

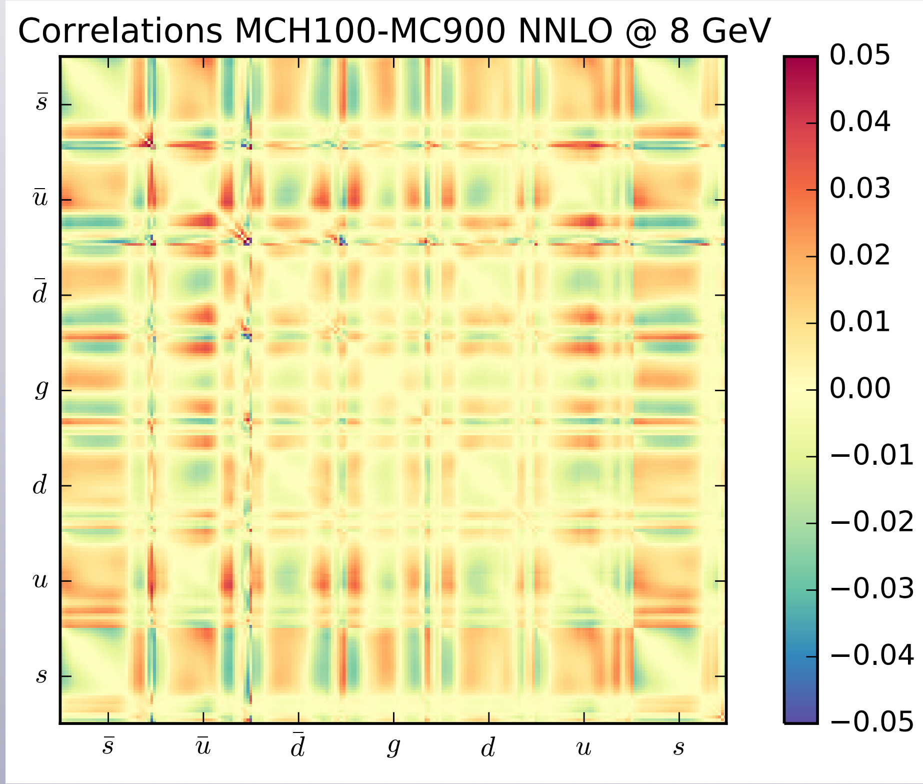
- A Hessian representation of the **MC900 PDF combination** has been constructed using **MC2H(PCA)**
- **Virtually perfect agreemen**t obtained for PDFs and luminosities as compared with MC900



Correlation matrix

Since the method is essentially exact, one expects that the **PDF correlation matrix** is **perfectly reproduced** (modulo the eigenvector reduction step)





Perfect?

Zooming in ...

Almost!

Tiny residual differences at the level of few perfect of the correlation coefficient at most, **irrelevant for LHC phenomenology**

LHAPDF

Status of LHAPDF

LHAPDF evolved from PDFLIB in 2004. Learn from experience, avoid bulky data grids in library: generate evolution on the fly.

Over 10 years, this model fell apart: almost everything as grids, with hard coded evolution arrays. *Huge* memory footprint (2GB!) and correctness issues (especially in multiset mode).

Rewritten 2012-2014: LHAPDF6 rewritten in C++, ∼every design problem solved. [arXiv:1412.7420]

Memory negligible, speed-ups, arbitrary member IDs in PDG scheme, sets now pure data & release decoupled, powerful metadata, ID code management, new universal interpolation functions (including α_S), member is now central object, compatibility (and new) interfaces, and built-in uncertainty calculation.

Latest developments: not much! *The model is working…*

Upcoming features: new Fortran API for arbitrary flavours, composite PDFs for nuclear corrections, + extended errors.

Upcoming new PDFs: PDF4LHC15, as discussed + MMHT14 and PDF4LHC15 fixed flavour sets + HERAPDF 2.0 + nCTEQ15

Summary & outlook

- PDF4LHC procedure standard for computation of PDF uncertainties in *searches*
- But requires several sets, with distinct uncertainty schemes: cumbersome & expensive
- $\Rightarrow$ new proposal: use combined PDFs based on MC replicas of $\sim$compatible standard fits
- Compression to fewer MC replicas or to Hessian as standard: different compressions for different applications
- PDF4LHC15 sets will be released very soon, via LHAPDF6. Built-in uncertainty calculator to be extended, and more new sets coming.