# CMC-PDF:

## Combination, compression and Hessian representation

Stefano Carrazza

Parton Distributions for the LHC, Benasque

February 20, 2015

University of Milan, Italy

We present a short overview of the following correlated topics:

1. **Combined MC sets of PDFs:** a practical implementation of the *PDF4LHC recommendation* using the MC representation.

   CMC-PDF

2. **Compression of MC PDFs:** a tool which reduces the size of a MC set of replicas preserving its statistical properties.

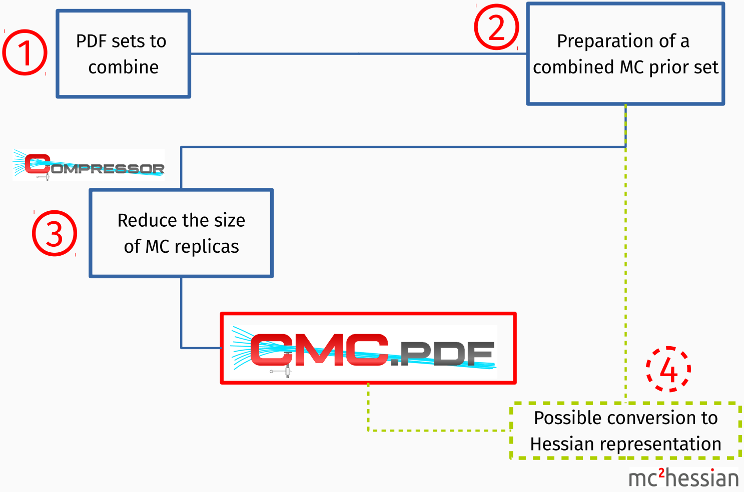
   COMPRESSOR

3. **MC to Hessian conversion:** a transformation strategy to convert any MC set of replicas into a set of eigenvectors.

   mc2hessian

The output of points 2 and 3 can be applied to *any MC set of PDFs*.

# MOTIVATION


1
PDF sets to combine
2
Preparation of a combined MC prior set
COMPRESSOR
3
Reduce the size of MC replicas
CMC.PDF
4
Possible conversion to Hessian representation
mc2hessian

# COMBINING MC SETS

## CMC-PDFs:

The aim of **CMC-PDFs** is to provide a practical implementation of the *PDF4LHC recommendation*.

$\Rightarrow$ *accurate*, *simple to use* and *computationally less intensive* $\Leftarrow$

## CMC-PDFs:

The aim of **CMC-PDFs** is to provide a practical implementation of the *PDF4LHC recommendation*.

$\Rightarrow$ *accurate*, *simple to use* and *computationally less intensive* $\Leftarrow$

## The combination strategy:

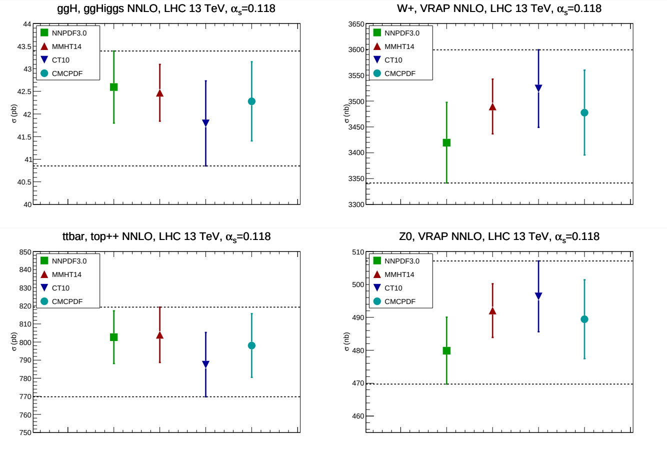
1. Transform the Hessian PDF sets to their **Monte Carlo representation** (Watt and Thorne 12) implemented in LHAPDF6
2. Combine the **same number of replicas from each of the prior sets**, assuming *equal weight* in the combination (i.e. an unweighted set)

In the next we combine $N_{\rm rep} = 100$ replicas from NNPDF3.0, CT10 and MMHT14, however any **other choice is possible**.

# PROPERTIES OF THE COMBINED PDF SET

The resulting **combined MC set** has statistical properties which lead to smaller uncertainties than the PDF4LHC envelope.

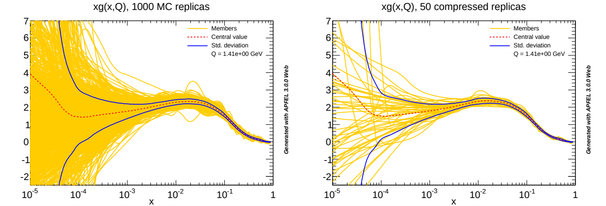
$\Rightarrow$ the envelope gives more weight to **outliers**



# COMPRESSION OF MC SETS

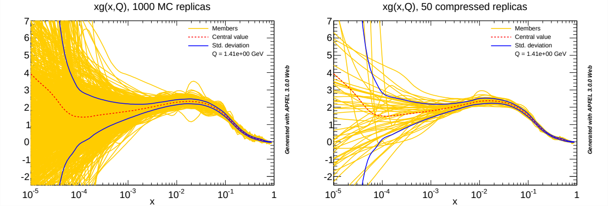
Compression idea:

**Reduce** the size of a PDF set of Monte Carlo replicas with no/minimal **loss of information**, *e.g.*:



## Compression idea:

**Reduce** the size of a PDF set of Monte Carlo replicas with no/minimal **loss of information**, *e.g.*:




Problem: **Preserve** as much as possible *the underlying statistical distribution* of the prior MC PDF set:

- Avoid bias in the extrapolation region.
- Conserve physical requirements: positivity, correlations, etc.

We define **statistical estimators** for the MC prior set:

1. **moments:** central value, variance, skewness and kurtosis
2. **statistical distances:** the Kolmogorov distance
3. **correlations:** between flavors at multiple $x$ points

We define **statistical estimators** for the MC prior set:

1. **moments:** central value, variance, skewness and kurtosis
2. **statistical distances:** the Kolmogorov distance
3. **correlations:** between flavors at multiple $x$ points

These estimators are them **compared** to subsets of replicas **interactively** driven by an *error function*, i.e.

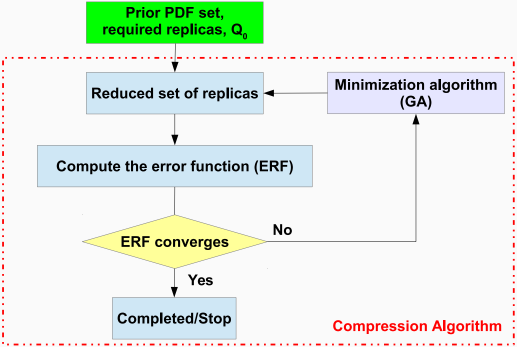
$$\mathrm{ERF}_{\mathrm{tot}} = \sum_n \frac{1}{N_n} \sum_i \left( \frac{C_i^{(n)} - O_i^{(n)}}{O_i^{(n)}} \right)^2$$

where $n$ runs over the number of statistical estimators and

- $N_i$ is a normalization factor extracted from random realizations
- $O_i^{(n)}$ is the value of the estimator for the prior
- $C_i^{(n)}$ is the corresponding value for the compressed set

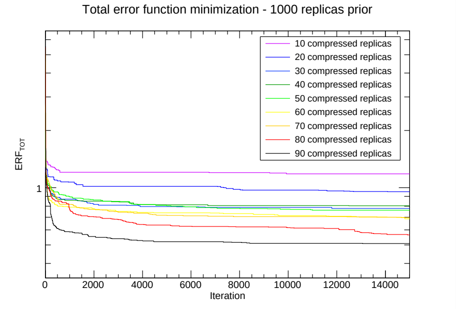
The algorithm **selects replicas** from the prior that minimize the **error function**. The minimization is driven by a *genetic algorithm*.

Validation: estimators, PDF plots, theoretical predictions, distances, $\chi^2$ to experimental data, etc.



## Test case:

Example results for the NNPDF3.0 NLO set with $N_{\mathrm{rep}} = 1000$ replicas.

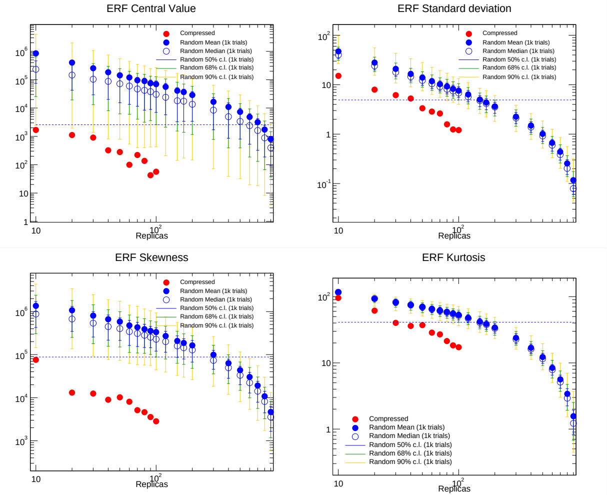



| Compressor v1.0.0 | |
|---|---|
| GA Parameters | |
| $N_{\mathrm{gen}}^{\mathrm{max}}$ | 15000 |
| $N_{\mathrm{mut}}$ | 5 |
| $N_{\mathrm{x}}$ | 50 |
| $x_{\mathrm{min}}$ | $1.1 \cdot 10^{-5}$ |
| $x_{\mathrm{max}}$ | $5.4 \cdot 10^{-1}$ |
| $n_f$ | 3 |
| $N_{\mathrm{x}}^{\mathrm{corr}}$ | 3 |
| $N_{\mathrm{rand}}$ | 1000 |

- The algorithm reaches the **stability plateau** after 2k iterations.
- A large prior of MC replicas **increases** the possible combinations.

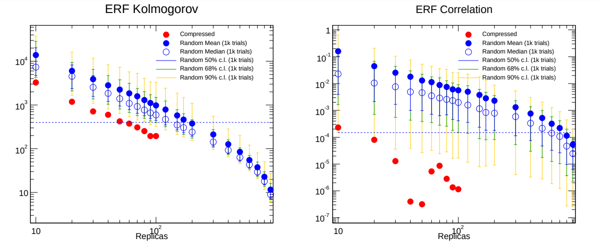
# COMPRESSION OF NATIVE MC PDF SETS

Moment estimators for the compression and random selections. [1]



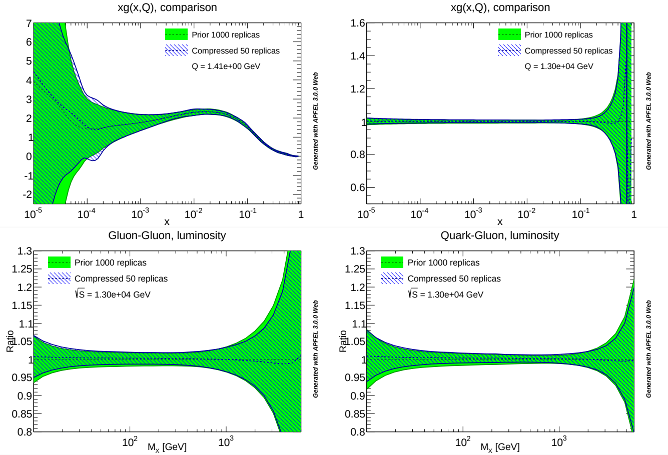

[1] Horizontal dashed line: lower limit 68% c.l. range for random selections, $N_{\text{rep}} = 100$.

Other estimators used in the error function:

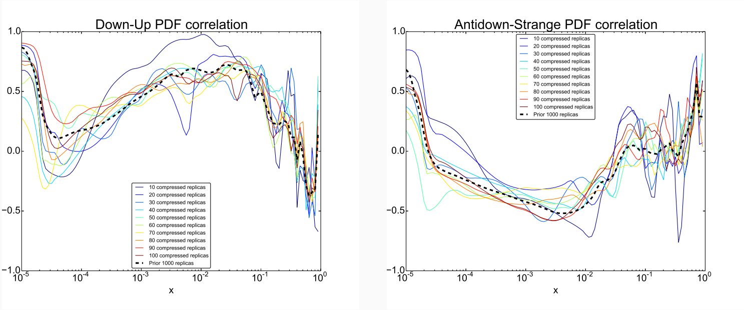



- Substantial **improvements** as compared to random selections.
- Compression is able to successfully reproduce **higher moments** and **correlations**.
- Results with $N_{\mathrm{rep}} = 50$ are equivalent to MC fits with 100 replicas.

All indicators show a **good description** of the prior set with only 50 replicas, e.g. luminosity and PDF comparison plots:



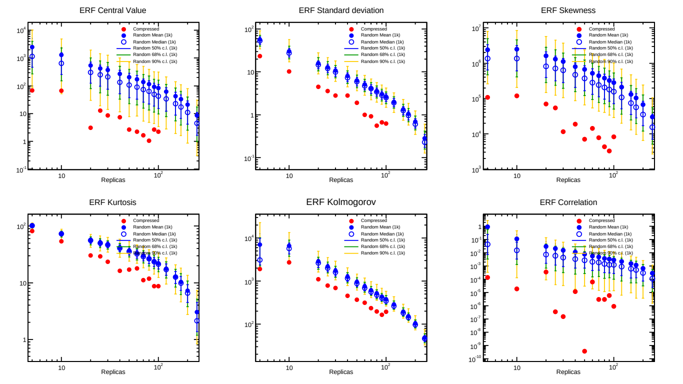
# COMPRESSION OF NATIVE MC PDF SETS




Central values and variances well reproduced, but also higher moments and correlations.

# COMPRESSION OF COMBINED PDF SETS

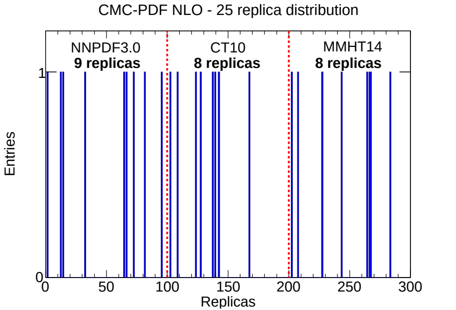
We apply **compression** to the combined PDF set with $N_{\mathrm{rep}} = 300$, composed by CT10, MMHT14 and NNPDF3.0.




The inclusion of higher moments are necessary because in this condition working in **Gaussian** approximation might not be **reliable**.

We have tested for multiple compressions sizes by computing a very large number of processes for **LHC observables**.

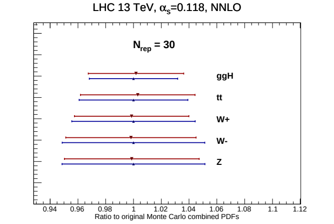
We found that $N_{\mathrm{rep}} = 20, 30$ are enough for **phenomenology**.

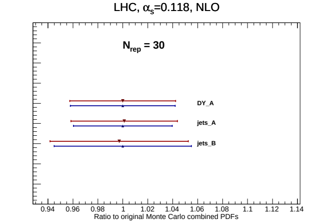


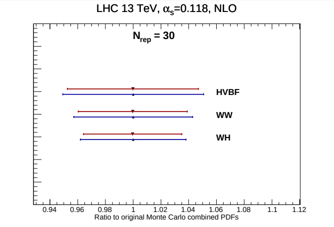

On average, the **same number** of replicas from each of the three sets is automatically selected by the compression algorithm.

# CMC-PDF PHENOMENOLOGY TESTS

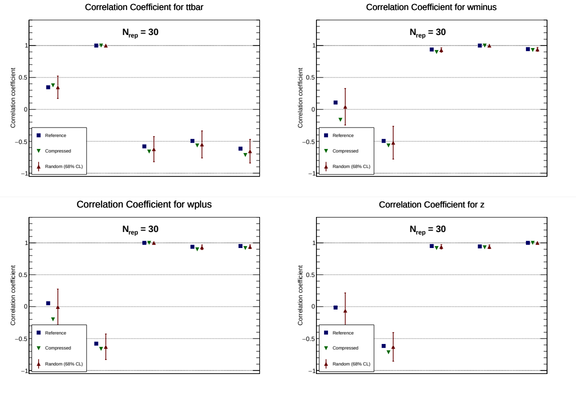
Good agreement for a large number of processes at **inclusive and differential level**.



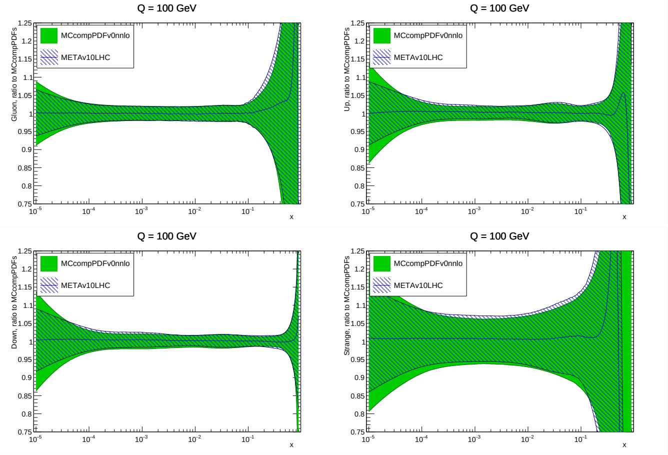



The compression algorithm reproduces the **correlation** between **physical observables**.




Correlation order: ggH, $t\bar{t}$, W+, W-, Z.

**Reasonable agreement** for central values and variances using a CMC-PDF based on MSTW08, CT10 and NNPDF2.3.



# HESSIAN REPRESENTATION OF MC SETS

The MC2Hessian idea consists in representing any MC PDF $f'_i$ as:

$$f'_i = f_0 + \sum_j a_{ij} \cdot (f_j - f_0),$$

a **linear combination** of a basis of replicas $(f_j, f_0)$.

The MC2Hessian idea consists in representing any MC PDF $f_i'$ as:

$$f_i' = f_0 + \sum_j a_{ij} \cdot (f_j - f_0),$$

a **linear combination** of a basis of replicas $(f_j, f_0)$.

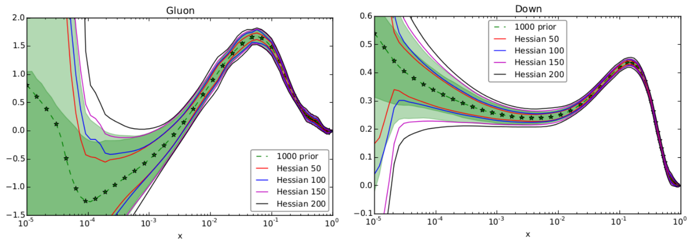
Strategy:

1. For each replica $k$ of the prior MC set we **solve** the linear system:

$$0 = \sum_{\alpha,\beta}^{n_f} \sum_{i,j}^{n_x} [f_k'^{\alpha}(a_k, x_i) - f_k'^{\alpha}(x_i)] \sigma^{-1}_{ij,\alpha,\beta}, [f_k'^{\beta}(a_k, x_j) - f_k'^{\beta}(x_j)]$$

2. **Build** the covariance matrix of $a_{ij}$: $\sigma$.
3. **Diagonalize** $\sigma^{-1}$ and determine the eigenvectors.
4. **Construct** the final **eigenvectors**.

Test case: Conversion of 1000 replicas of NNPDF3.0 NLO.




We observe that 150 **eigenvectors** reproduce well the uncertainty of the prior set.

Clear possibility to **generalize** this procedure for any MC set including CMC-PDFs.

The compression algorithm might provide a good initial basis for the hessian conversion.

# SUMMARY AND OUTLOOK

- **CMC-PDFs** provide a well defined and efficient solution for modern multi-PDF combination.
- **Flexibility** of representations: MC vs. Hessian approximation.

## Code & grid delivery:

Both codes will be publicly available soon, together with a paper where systematic studies are performed:

- **Compressor:** as a C++ program, dependencies: LHAPDF6, ROOT, GSL
- **MC2Hessian:** as a Python script, deps: LHAPDF6, numpy, numba.
- **CMC-PDFs** grids will be released soon to LHAPDF6.

## Questions?