PROCEEDINGS OF SCIENCE

Parton Distribution Functions at LHC and the SMPDF web-based application

Stefano Carrazza[*][†]
Theoretical Physics Department, CERN, Geneva Switzerland
E-mail: stefano.carrazza@cern.ch

Zahari Kassabov[‡]
Dipartimento di Fisica, Università di Torino and INFN, Sezione di Torino
TIF Lab, Dipartimento di Fisica, Università di Milano
E-mail: kassabov@to.infn.it

We present SMPDF Web, a web interface for the construction of parton distribution functions (PDFs) with a minimal number of error sets needed to represent the PDF uncertainty of specific processes (SMPDF).

VII Workshop italiano sulla fisica pp a LHC
16-18 Maggio 2016
Pisa, Italy

[*] Speaker.
[†] CERN-TH-2016-129
[‡] TIF-UNIMI-2016-5

Parton distribution functions at LHC The accurate determination of the parton distribution functions of the proton is crucial for precise predictions at the Large Hadron Collider (LHC). PDFs are extracted by comparing experimental data to theoretical parton level predictions obtained in perturbative QCD. Currently, modern global PDF fits include data from multiple processes such as deep-inelastic scattering (DIS), fixed target Drell-Yan and hadronic data from colliders.

In Figure 1 we show the kinematics coverage in the (x,Q^2) plane for the experimental data included in the recent NNPDF3.0 set of PDFs [4]. The DIS dataset covers a broad range in x, and is thus an important ingredient in PDF determination. Hadronic data from ATLAS, CMS and LHCb extends the Q^2 coverage. In particular, jet data improves the gluon PDF uncertainties at large x, while W production improves the ability to resolve individual quark flavours. The values in Figure 1 are however insufficient to precisely gauge the kinematic impact of a given dataset in the parton fit: For one, they have have been produced using Leading Order kinematics only. More importantly, they do not take into account the complex correlations induced on the different kinematic regions by the fitting procedure.

Here we present SMPDF Web, a tool which allows to precisely determine the kinematic impact of a given dataset on a PDF fit, by correlating the data values with each parton flavour on a grid of points in x. This information can additionally be used to produce compressed representations of the input PDFs which allow to efficiently compute predictions for a given set of input processes.

The PDF4LHC15 recommendation and tools The PDF4LHC working group has been tasked with producing a recommendation for a standard method of calculating PDF+α_S uncertainties. The most recent PDF4LHC recommendation [1] prescribes a combined PDF set (branded PDF4LHC15), composed of the statistical combination [2, 3] of the individual PDF determinations from three independent collaborations: NNPDF3.0 [4], CT14 [5] and MMHT2014 [6]. These PDF sets satisfy a set of compatibility requirements: usage of global datasets for the PDF determination, theoretical predictions and parton evolution computed in the General Mass Variable Flavour Number Scheme (GM-VFNS) and α_s set to the PDG average [7]. The recommended usage and applicability are discussed in Ref. [1].

The recommendation is implemented by first constructing a combined prior PDF set of Monte Carlo PDF *replicas* from each of the three collaborations: A probability distribution related to the PDF uncertainty of a given hadronic level observable can be computed by convolving the corresponding parton level quantity with each of the replicas. Then any statistical quantity (such as the mean and the standard deviation) can be obtained from the resulting distribution. This becomes impractical when the number of Monte Carlo replicas required to faithfully reproduce the prior distribution is high enough that a significant computational effort is required to perform all the convolutions. This is the case for the PDF4LHC prior which contains a total of 900 replicas. Therefore several compressed PDF sets are delivered together with the prior. They are based on three different algorithms: Compressed Monte Carlo (CMC-PDF) [8, 9], Monte Carlo to Hessian, MC2H [10] and Meta-PDF [11]. The latest two also transform the Monte Carlo sample into a Hessian set, which is more adequate for certain experimental analyses, and more efficient in the circumstances where one can assume that the prior set is Gaussian. The CMC methodology excels at reproducing the non-Gaussianities of the prior (particularly important for searches). A study about the accuracy of these methodologies for different observables is performed in Ref. [12]. The

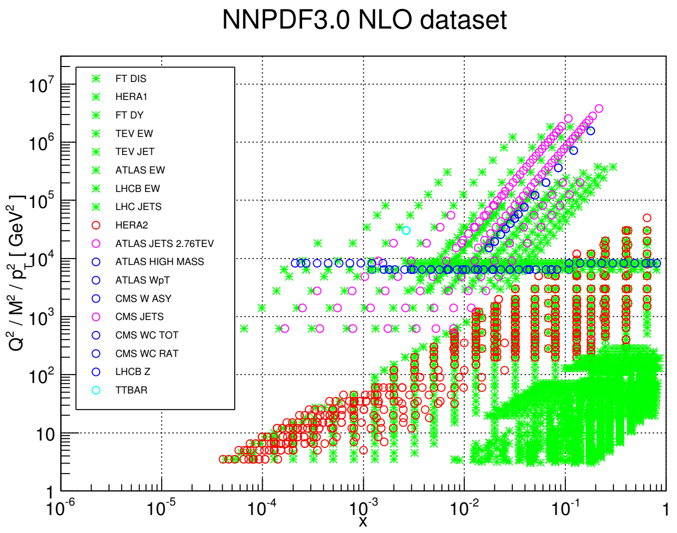


Figure 1: Leading-order kinematics coverage in the (x, Q^2) plane of the NNPDF3.0 dataset from [4]. This plot illustrates the current state of the art data included in modern PDF fits.

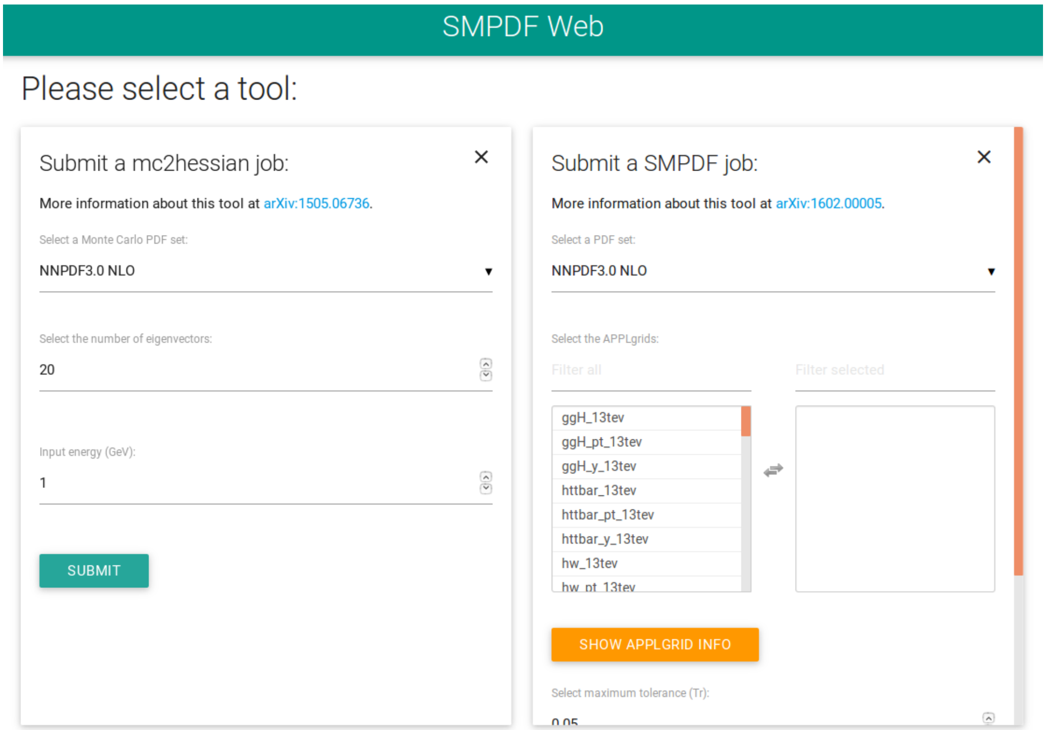
final result is made available in the LHAPDF format [13].

Specialized minimal PDFs As a follow-up of the studies that led to the PDF4LHC15 recommendations, a Hessian reduction algorithm called Specialized Minimal PDF (SMPDF) was proposed and implemented in Ref. [14]. It is a development upon the MC2H methodology, where the PDF covariance matrix is reproduced by selecting the largest eigenvectors through Principal Component Analysis (PCA) and expressing them as a linear combination of the input Monte Carlo replicas. In the SMPDF methodology we focus on providing a representation of the covariance matrix, which allows an accurate determination of the PDF uncertainties for a specific set of processes, with a minimal number of PDF error sets. This is achieved through an iterative procedure, in which one eigenvector (error set) is added at a time until the standard deviation for each of the input processes is reproduced better than a threshold selected by the user. Therefore the threshold is an upper bound for the inaccuracy on the input observables.

Different processes can be combined (either in the same SMPDF set or from independent ones) in such a way that the PDF correlation between them is reproduced, and further information can always be added. To this end, we provide explicitly the linear transformation that converts the prior set into the resulting SMPDF.

The eigenvectors are selected based on kinematic considerations (specifically the correlation between the value of a given PDF flavour and point in x with the value of the observable), which ensures that the methodology is robust upon variations in the cuts and generalizes efficiently to similar processes. Therefore one can reliably use SMPDFs to compute predictions of processes that were not explicitly given as input to the algorithm (or were given with different cuts) but hold a similar PDF dependence.

The SMPDF methodology has been explicitly validated for a number of representative Standard Model processes of particular relevance at the LHC (including Higgs, top and electroweak physics). In each case we observe a large reduction in the number of error sets while keeping an



SMPDF Web

Please select a tool:

Submit a mc2hessian job:

More information about this tool at arXiv:1505.06736.

Select a Monte Carlo PDF set: NNPDF3.0 NLO

Select the number of eigenvectors: 20

Input energy (GeV): 1

Submit a SMPDF job:

More information about this tool at arXiv:1602.00005.

Select a PDF set: NNPDF3.0 NLO

Select the APPLgrids:

Select maximum tolerance (Tr): 0.05

Figure 2: The SMPDF Web interface. The user can select the MC2H tool (left) or the SMPDF one (right)

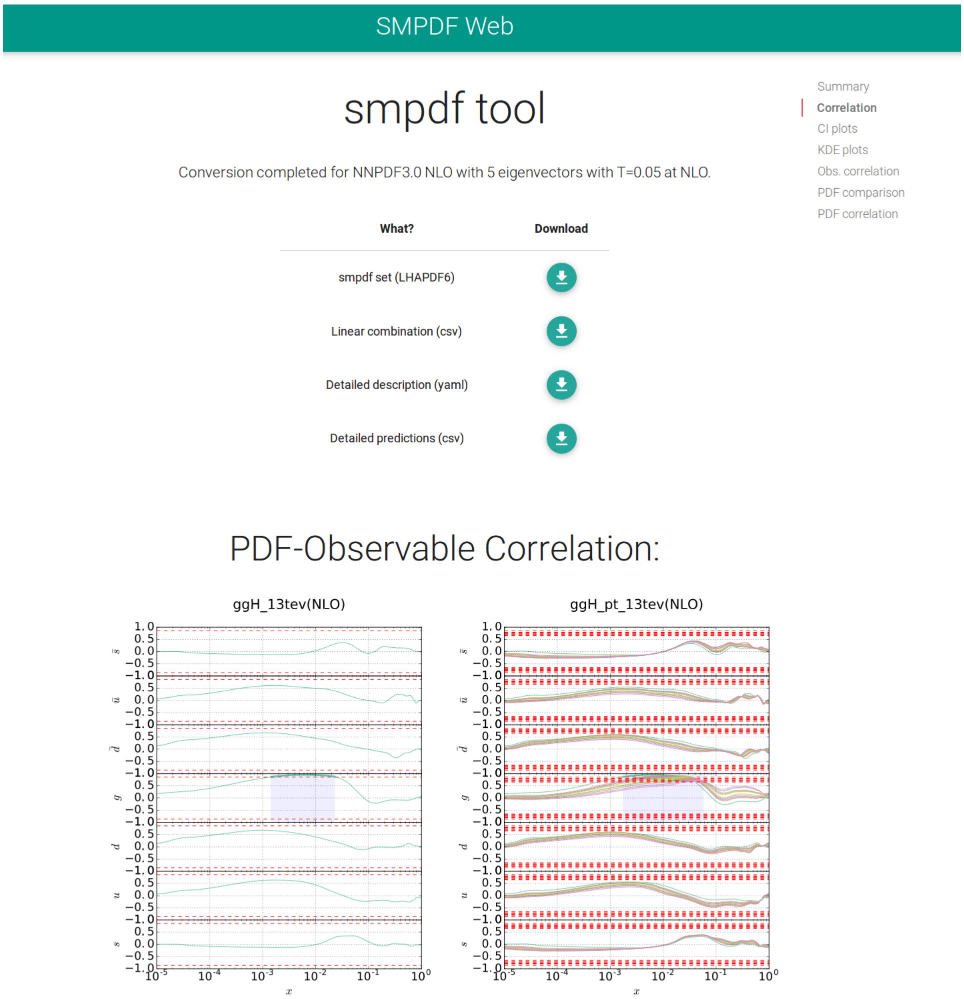
accuracy comparable to that of the MC2H reduction. For example, as discussed in Ref. [14], it is possible to reproduce the predictions of the PDF4LHC prior for the most relevant Higgs production channels with 15 error sets (to be compared with the 900 of the prior and the 100 of MC2H). If one is only interested in the gluon fusion channel, then only 4 error sets suffice.

As shown in Ref. [12], by selecting a general enough set of input observables (constructing what we call Ladder SMPDF), one can achieve the same accuracy as Meta PDF, for a large set of Standard Model processes, with about half as many error members and with the possibility to trade better accuracy for more error sets, by decreasing the threshold parameter and/or increasing the number of input processes. This shows that SMPDFs can be advantageous in situations where both computational performance and accuracy in the computation of PDF uncertainties are needed.

In the outlook of the original paper we speculated about the integration of the SMPDF software in a web-based application such as APFEL Web [15, 16, 17]. Here we describe such an application: SMPDF Web. It is available at:

http://smpdf.mi.infn.it/

With this application one can generate customized SMPDF set, for the desired prior PDF and observables, as well as a Hessian representation of a given Monte Carlo PDF obtained with the MC2H algorithm. The resulting PDF sets (in the LHAPDF6 format) can be downloaded from the result page, together with complementary information about the procedure (in particular the resulting linear transformations) and validation plots. We have computed a selection of parton level observables to be used for the SMDPF algorithm.



SMPDF Web

smpdf tool

Conversion completed for NNPDF3.0 NLO with 5 eigenvectors with T=0.05 at NLO.

What?	Download
smpdf set (LHAPDF6)	⬇
Linear combination (csv)	⬇
Detailed description (yaml)	⬇
Detailed predictions (csv)	⬇

PDF-Observable Correlation:

Figure 3: The result page for the SMPDF tool. The user is presented with the option to download the resulting PDF set in the LHAPDF format, as well as other relevant outputs and validation plots (in the image, figures highlighting the kinematic PDF dependence of the ggH process).

SMPDF Web is based on the public SMPDF code (where also the MC2H algorithm is implemented), available at:

https://github.com/scarrazza/smpdf

In Figure 2 we show the front page of the website. The user can select one of the two tools by clicking on the corresponding thumbnail. Then a form is presented asking for the required parameters. For the MC2H tool these are:

- The input PDF set, from a list of installed options.
- The number of desired eigenvectors.
- The energy scale at which the PDFs are evaluated, in GeV.

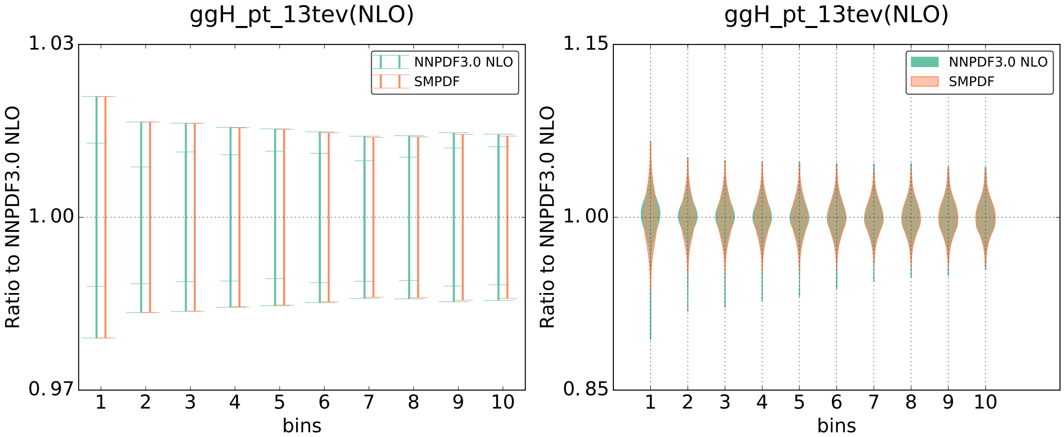


Figure 4: Example of confidence interval (left) and kernel density estimate (right) predictions for the SMPDF for *ggH* predictions.

MC2H works best when the number of eigenvectors is high enough that the whole PDF covariance matrix can be reproduced (a plot is shown as part of the output). In this case the energy scale makes little impact as long as it is above the scale at which the evolution of the input PDF can be considered reliable.

In turn, the SMPDF tool requires the following input:

- An input PDF set (MC or Hessian).
- A list of observables.
- The threshold parameter (tolerance), indicating the maximum allowable deviation in the reproduction of uncertainties.
- The perturbative order (LO/NLO) at which the observables are computed.

The observables we provide in the web interface are in the APPLgrid format [18] and include those analyzed in the SMPDF paper [14]). The SMPDF code also allows hadron level predictions entered directly in a text format. We find that a tolerance of 5% or 10% results in a good compromise between accuracy and the resulting number of error sets, as long as the SMPDF set is used for observables compatible with those given as input.

After the successful completion of the respective forms, a new page is rendered, from which the user can download the resulting PDF set, as well as other useful results (such as the linear transformations involved in both algorithms), and showing several validation plots. Figure 3 shows an example of result page with the summary information and the PDF-observable correlation plots for each process requested by the user. These plots provide a direct handle for the sensitivity of that observable to the different kinematical regions of the PDF, and correspondingly shows which regions would improve the most if that observable was to be included in the fit.

Figures 4 and 5 show examples of validation plots from SMPDF Web. Specifically we have generated an SMPDF set using NNPDF3.0 NLO as a prior PDF and the ggH p_T distribution at NLO as input, with 5% tolerance. We obtain a Hessian set with $N_{\rm eig} = 3$. Figure 4 left compares

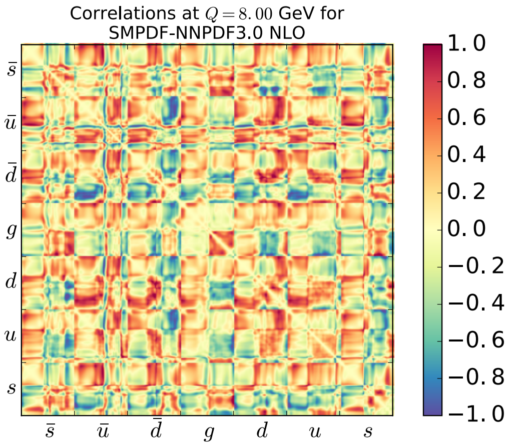


Figure 5: Example of PDF-PDF correlations for SMPDFs optimized for Higgs production from gluon fusion. The most relevant correlations for this process (gluon at small x) are reproduced at percent level

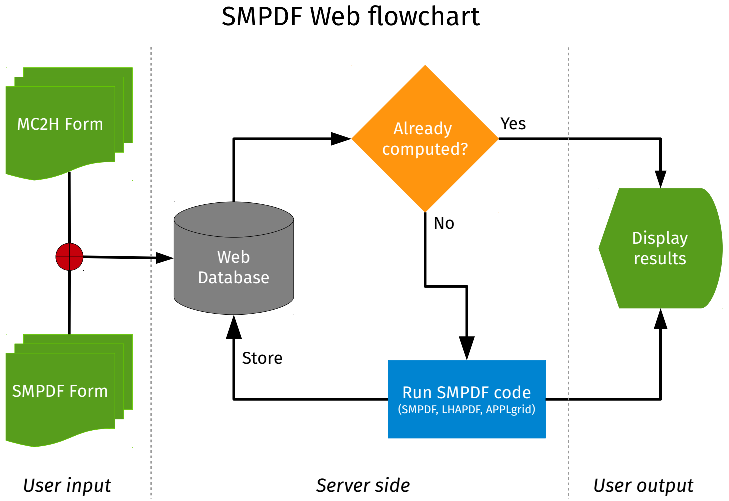


Figure 6: The SMPDF web application flowchart.

prediction uncertainties between the prior (green) and the resulting SMPDF (orange), showing an almost perfect agreement, and the right figure represents the same quantities as a kernel density estimate. Figure 5 shows the PDF correlation difference between the prior and the final SMPDF set. As expected, we observe small differences for the gluon channel.

For the time being we provide a limited number of input PDF sets and processes; however users are invited to submit upload requests directly from the website.

SMPDF Web implements a simple caching mechanism which stores configuration and results. In this way we provide instantaneous results for jobs which have been already computed, and the generated URL can be used to reference and share the output. In Figure 6 we show a diagram summarizing the layout of the web application.

Acknowledgements S.C. is supported by the HICCUP ERC Consolidator grant (614577). Z.K. is supported by the Executive Research Agency (REA) of the European Commission under the Grant Agreement PITN-GA-2012-316704 (HiggsTools).

References

[1] J. Butterworth *et al.*, J. Phys. G **43** (2016) 023001 arXiv:1510.03865.

[2] , S. Forte, Acta Phys. Polon. **B41** (2010), 2859-2920 arxiv:1011.5247

[3] G. Watt and R. S. Thorne, JHEP **1208** (2012) 052 arXiv:1205.4024.

[4] R. D. Ball *et al.* [NNPDF Collaboration], JHEP **1504** (2015) 040 arXiv:1410.8849.

[5] S. Dulat *et al.*, Phys. Rev. D **93** (2016) no.3, 033006 arXiv:1506.07443.

[6] L. A. Harland-Lang, A. D. Martin, P. Motylinski and R. S. Thorne, Eur. Phys. J. C **75** (2015) no.5, 204 arXiv:1412.3989.

[7] K. A. Olive *et al.* [Particle Data Group Collaboration], Chin. Phys. C **38** (2014) 090001.

[8] S. Carrazza, J. I. Latorre, J. Rojo and G. Watt, Eur. Phys. J. C **75** (2015) 474 arXiv:1504.06469.

[9] S. Carrazza and J. I. Latorre, arXiv:1605.04345 [hep-ph].

[10] S. Carrazza, S. Forte, Z. Kassabov, J. I. Latorre and J. Rojo, Eur. Phys. J. C **75** (2015) no.8, 369 arXiv:1505.06736.

[11] J. Gao and P. Nadolsky, JHEP **1407** (2014) 035 arXiv:1401.0013.

[12] J. R. Andersen *et al.*, arXiv:1605.04692 [hep-ph].

[13] A. Buckley, J. Ferrando, S. Lloyd, K. Nordstrom, B. Page, M. Rufenacht, M. Schonherr and G. Watt, Eur. Phys. J. C **75** (2015) 132 arXiv:1412.7420.

[14] S. Carrazza, S. Forte, Z. Kassabov and J. Rojo, Eur. Phys. J. C **76** (2016) no.4, 205 arXiv:1602.00005.

[15] S. Carrazza, A. Ferrara, D. Palazzo and J. Rojo, J. Phys. G **42** (2015) no.5, 057001 doi:10.1088/0954-3899/42/5/057001 [arXiv:1410.5456 [hep-ph]].

[16] V. Bertone, S. Carrazza and J. Rojo, Comput. Phys. Commun. **185** (2014) 1647 doi:10.1016/j.cpc.2014.03.007 [arXiv:1310.1394 [hep-ph]].

[17] V. Bertone, S. Carrazza and N. P. Hartland, arXiv:1605.02070 [hep-ph].

[18] T. Carli, D. Clements, A. Cooper-Sarkar, C. Gwenlan, G. P. Salam, F. Siegert, P. Starovoitov and M. Sutton, Eur. Phys. J. C **66** (2010) 503 doi:10.1140/epjc/s10052-010-1255-0 [arXiv:0911.2985 [hep-ph]].