

PAPER • OPEN ACCESS

## Machine learning challenges in theoretical HEP

To cite this article: Stefano Carrazza 2018 *J. Phys.: Conf. Ser.* **1085** 022003

View the [article online](#) for updates and enhancements.



**IOP | ebooks™**

Bringing you innovative digital publishing with leading voices to create your essential collection of books in STEM research.

Start exploring the [collection](#) - download the first chapter of every title for free.

# Machine learning challenges in theoretical HEP

**Stefano Carrazza**

Theoretical Physics Department, CERN, Geneva, Switzerland

E-mail: [stefano.carrazza@cern.ch](mailto:stefano.carrazza@cern.ch)<sup>1</sup>

**Abstract.** In these proceedings we perform a brief review of machine learning (ML) applications in theoretical High Energy Physics (HEP-TH). We start the discussion by defining and then classifying machine learning tasks in theoretical HEP. We then discuss some of the most popular and recent published approaches with focus on a relevant case study topic: the determination of parton distribution functions (PDFs) and related tools. Finally, we provide an outlook about future applications and developments due to the synergy between ML and HEP-TH.

## 1. Introduction

Over the past several years machine learning (ML) has become one of the most popular and powerful sets of techniques and tools used for multidisciplinary scientific research. Such popularity has been continuously growing in the past years thanks to the increasing number of methodological developments, the availability of faster hardware with strong computational capabilities such as modern GPUs and coprocessors, and finally, the great interest and investment from the private sector.

The recent enthusiasm has led to a new underlying code development strategy where several tools based on ML are available as open source projects, some examples are TensorFlow [1], scikit-learn [2], Keras [3], Theano [4] among others. Easy access to these tools has simplified the integration of new modern techniques in many research fields, in particular for those where a large amount of data is available.

The dissemination of the innumerable applications and developments based on ML has given way to conferences such as ICML<sup>2</sup>, NIPS<sup>3</sup> and ACAT, but also to the composition of specialized working groups, e.g the IML LHC working group at CERN<sup>4</sup> which promotes the integration of new techniques for specific requirements in experimental analysis.

In the next sections we focus the discussion on ML in theoretical High Energy Physics (HEP-TH). We start by defining the categories of applications observed in that field. Then we take as a case study the recent development achieved for the determination of parton distribution functions (PDFs), where several techniques from ML are employed. Finally, we conclude by listing the most plausible development directions for the integration of ML in HEP-TH.

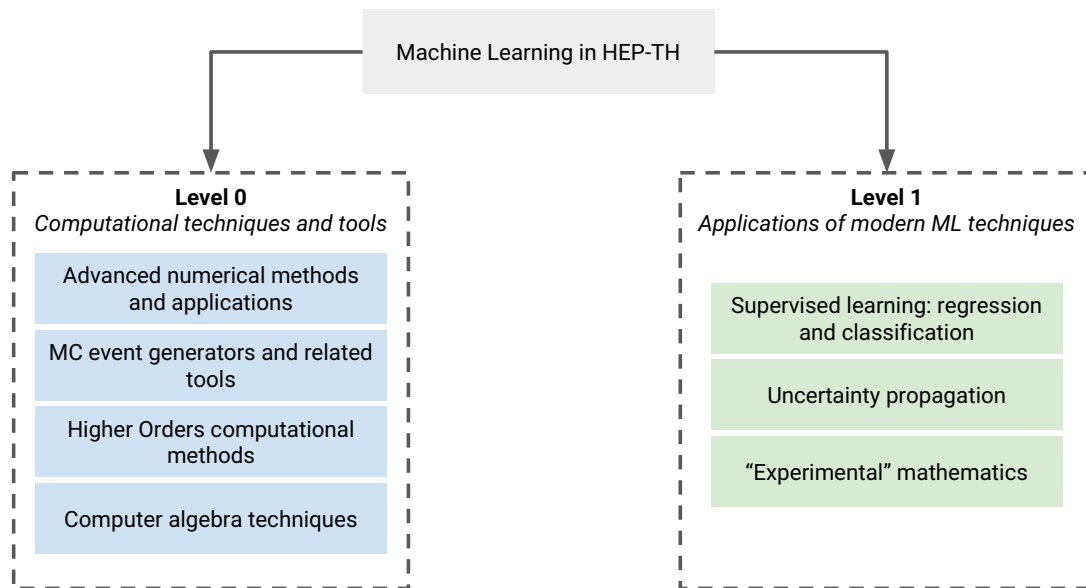
<sup>1</sup> Preprint: CERN-TH-2017-212

<sup>2</sup> <http://icml.cc>

<sup>3</sup> <http://nips.cc>

<sup>4</sup> <https://iml.web.cern.ch/>





**Figure 1.** One possible graphical schematic representation of how machine learning is identified in HEP-TH applications.

## 2. Identifying machine learning applications in HEP-TH

While classifying ML applications in experimental physics may seem simple because the experimental analysis usually requires model regression, classification and noise filtering to extract useful information from large datasets. It becomes challenging when trying to determine applications from the point of view of theoretical physics.

When talking about ML techniques in theoretical High Energy Physics (HEP-TH) we may naively imagine that the usage of such tools is not ideal. In fact, we can argue that the theoretical physicists are trained to decode nature by constructing conjectures and building theoretical frameworks which remap the complexity of the measured observable in conceptual model rules. Therefore, this approach is in contrast with a model determination through a ML black box model, of which a physical interpretation is very difficult or nearly impossible to achieve.

However, nowadays such misleading interpretation is disappearing and machine learning is becoming a set of useful tools that helps achieving practical results in situations where the theory has a strong computational requirement or when it requires the determination of free parameters. So, before continuing our discussion, we think it is important to define and classify how ML is translated in HEP-TH. In Figure 1 we summarize one plausible graphical representation of the two branches of applications that we will discuss now in turn: *Level-0* and *Level-1*.

### 2.1. *Level-0: computational techniques and tools*

The first level contains machine learning in terms of computational techniques and tools. In this category we find the great majority of topics presented in the Track 3 section of the ACAT conference. These applications contain the most robust implications of computing in theoretical physics. Due to the specific nature of the problems addressed in HEP-TH these tools should be considered as part of ML applications because their development has contributed in exhaustive manner to the development of new techniques and methods in this field.

A non-exhaustive list of topics involved in *Level-0* are summarized in the points below:

- Advanced numerical methods and applications: some examples are algorithms for Monte Carlo and Quasi Monte Carlo integration, techniques for subtraction schemes and regularization of Feynman integrals, e.g. [5], resummation techniques [6].
- Monte Carlo event generators: in this category we have the methodological developments established by MC codes such as POWHEG [7], MadGraph\_aMG5 [8], Pythia [9], Herwig [10], MCFM [11] and several others. The possibility to reuse events independently from the parton distribution functions thanks to reweighting and weight storage techniques such as APPLgrid [12] and FastNLO [13].
- Higher order computational methods: numerical techniques for  $N$ -loop integrals as OneLoop [14], QCDLoop [15], LoopTools [16]; parton level generators at NNLO, e.g. DYNLO [17] and the more recent N3LO [18].
- Computer algebra techniques: some examples of algebra systems developed for the HEP-TH community as FORM [19] and QGRAF [20] for the translation of Feynman diagram rules into compact analytic expressions for its numerical evaluation.

### 2.2. Level-1: applications of ML modern techniques

This second level contains the ML applications in “sensu stricto” i.e. using ML modern techniques used in data sciences. This kind of application usually requires hybrid projects where experimental data and theory are included together. Nevertheless, during the last few years there has been a strong development of very successful tools based on these techniques.

Some examples for this category are:

- Supervised learning, such as regression and classification: parton distribution [21] and fragmentation functions [22] determination, Monte Carlo tunes [23], reweighting techniques, jet discrimination through deep convolutional neural networks [24].
- Techniques for uncertainty estimation and combination: in this category we find several tools from the PDF4LHC15 recommendation [25], and some recent methods to provide a reliable MC uncertainty for simulations, e.g. by modeling jet predictions [26].
- “Experimental” mathematics: applications based on machine learning optimization algorithms which may lead to the determination of multivariate densities [27, 28] and sampling for integral evaluation [29].

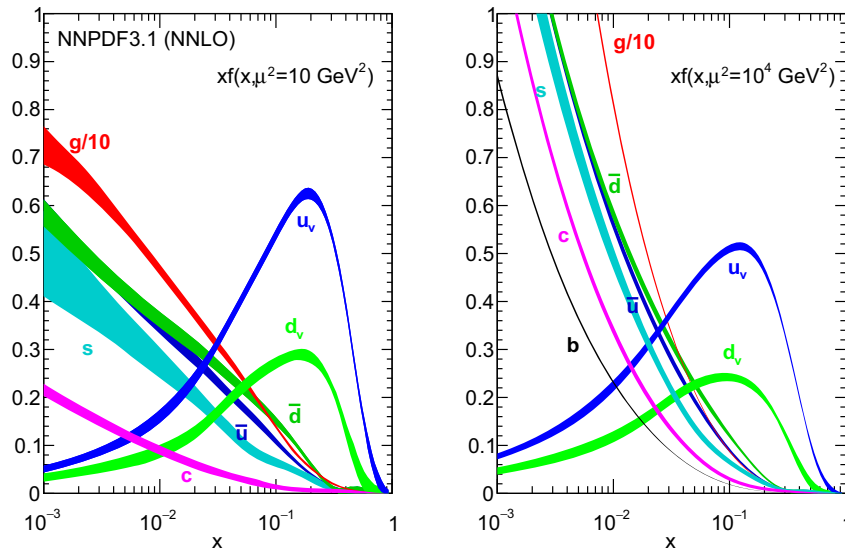
In the next section we describe the recent innovative achievements obtained by the parton distribution function community in HEP-TH using ML methods.

## 3. Case study: the proton structure determination

The Quantum Chromodynamics (QCD) theory describes the proton structure in terms of partons, e.g. quarks and gluons, but due to the non-perturbative regime of confinement we are unable to evaluate from QCD the momentum fraction of the proton carried by each parton. However, in order to avoid our lack of knowledge we introduce the concept of parton distribution functions (PDFs). These PDFs are then inferred from data of all relevant processes, together with the theoretical knowledge on how they affect the cross section which can be calculated approximately in perturbation theory.

### 3.1. The NNPDF approach

The NNPDF collaboration uses ML techniques to obtain a PDF determination. In contrast to other problems in ML, where finding an accurate and fast algorithm is enough, we are not only interested in the best PDF fit, but also in obtaining an uncertainty estimate of the PDF determination. The procedure we employ is described in full details in [30] and can be summarized as:



**Figure 2.** The NNPDF3.1 NNLO PDFs, evaluated at  $\mu^2 = 10 \text{ GeV}^2$  (left) and  $\mu^2 = 10^4 \text{ GeV}^2$  (right) from [21].

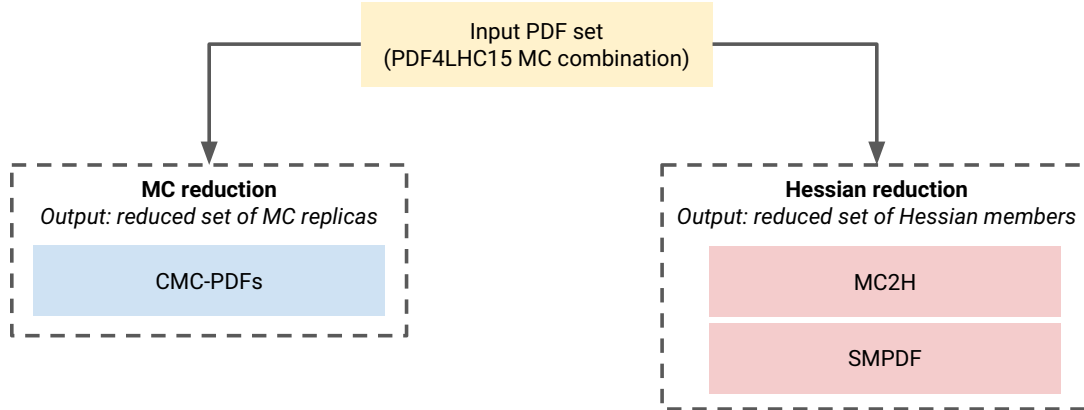
- Monte Carlo generation of artificial data. Experimental data, with central values, errors and their correlations are used to generate further artificial data, consistent with the covariance matrix provided by the experiment.
- Neural network fit to artificial data. A genetic algorithm is used to fit each artificial data set to a neural network representing the PDF (more details in [31]).
- Predictions are later obtained by computing statistical estimators (such as means, quantiles and standard deviations) over the set of neural networks. Figure 2 illustrates the current state of the art PDFs, NNPDF3.1 [21], obtained with the NNPDF framework.

The points above summarize the main difficulties in PDF fits. The first consists in the inclusion of multiple experimental data which introduces indirect constraints on the PDFs. The data included in such fits are based on several datasets measured during the past decades and based on different physical processes: deep-inelastic scattering, fixed target Drell-Yan and hardronic data. This data is obtained through diverse experimental techniques and statistical analyses, therefore yielding possible inconsistencies and tensions among themselves. In order to limit these effects the NNPDF methodology propagates the uncertainty of the experimental data on the PDFs by performing the Monte Carlo artificial replica generation based on the covariance matrix provided by each experiment.

The second most relevant problem consists in the choice of an unbiased functional form able to adapt and allow the propagation of data uncertainties into PDF errors while keeping under control physical requirements such as momentum sum rule conservation and positivity constraints. The NNPDF collaboration has successfully employed neural networks based on feed-forward multilayer perceptron architecture to model each single PDF flavor entering the fit procedure. The large number of parameters from these PDFs are then trained through a genetic optimization algorithm.

NNPDF has released several PDF sets of global unpolarized determinations in the last years<sup>5</sup>,

<sup>5</sup> <http://nnpdf.mi.infn.it/>



**Figure 3.** Graphical representation of the reduction algorithms used in the PDF4LHC15 recommendation [25].

together with the more recent polarized PDFs [32] and Fragmentation Functions [22]. One of the current most important tasks of the NNPDF collaboration is to improve the performance and quality of the optimization having in mind that in the next years the number of the new measurements from LHC will increase.

### 3.2. PDF4LHC15 tools for LHC Run II

Another issue concerning PDF determination consists in deciding which set of PDF should be used for the calculation of a PDF-dependent quantity. The question addressed here is how to obtain the best combined PDF uncertainty from individual PDF sets. This kind of issue is studied by the PDF4LHC working group which releases a recommendation document usually every three years. The latest one published in 2015 [25] aims to release combined PDF sets based on methods with clear statistical interpretation.

The PDF4LHC15 prescription starts by constructing a prior Monte Carlo combined set of PDFs derived from global determinations, namely MMHT14 [33], CT14 [34] and NNPDF3.0 [30]. These sets satisfy requirements such as: similar dataset, DGLAP solution and  $\alpha_s$ . The prior construction is obtained by converting the Hessian sets to MC replicas following the procedure described in [35]. After that we apply algorithms to reduce the redundant information stored in the prior and deliver sets of PDFs with the same properties of the prior but smaller number of replicas.

#### *Compression algorithm*

The CMC-PDFs [36] implement the compression algorithm of MC replicas. This algorithm is designed to extract a subset of replicas which preserves as much as possible the underlying statistical distribution of the prior MC PDF set. This algorithm pre-computes for the input PDF the moments (central value, variance, skewness and kurtosis), the statistical distance (Kolmogorov distance) and the correlations for each flavor in a grid of  $x$  points. These estimators are then compared to subsets of replicas selected by a genetic algorithm. The procedure terminates when a tolerance value for the error function comparing these estimators is flat.

From a practical point of view in the context of the PDF4LHC15 the initial  $N_{\text{rep}} = 900$  replicas of the prior set have been reduced to 100 replicas. Further technical investigations based on clustering algorithms have also been performed in [37].

### *MC2H*

The MC2H algorithm [38] was first introduced to convert MC PDF sets into a Hessian representation. This algorithm performs the principal component analysis (PCA) of the PDF covariance matrix in a predefined grid of  $x$  nodes for all flavors at a given initial scale. The eigenvectors obtained from the PCA are the basis of the Hessian representation. This representation consists in simple linear combinations of the input MC replicas. The MC2H procedure also allows the reduction of the number of required replicas of the input PDF set because we can reject the eigenvectors associated to small eigenvalues therefore obtaining a reduced set of Hessian replicas.

### *SMPDF*

Starting from the MC2H algorithm we have developed the Specialized Minimal PDFs (SMPDF) [39] which consists in obtaining the smallest set of Hessian replicas for a given physical process. The SMPDF algorithm performs an interactive PCA reduction on top of the PDF covariance matrix computed in a grid of  $x$  points determined as the region where the PDF-process correlation is maximal. We have also provided a public web-interface for the construction of SMPDFs available and documented in [40].

## 4. Outlook

The advantages and results obtained thanks to ML have been essential for several developments of this research field from PDF determination to Monte Carlo event generation. In the next months and years new applications will be achieved. It is already possible to summarize the two main directions for these developments.

The first consist in the development of tools for the estimation and propagation of uncertainties. New methods to deal with uncertainty will possibly improve the determination of PDFs among other applications. This point also includes the new ideas about reweighting techniques in the context of higher order calculations.

The second development branch consists in the construction of new modern neural network architectures for problem specific applications together with efficient new gradient based methods. Such tools will open the possibility to obtain better methods for multidimensional density estimation. Consequently, we will obtain better sampling algorithms relevant for improved and faster integration algorithms in Monte Carlo event generators and similar tools.

In these proceedings we provided a short overview of successful applications of ML in HEP-TH but this is just the prelude of a new era where ML and HEP-TH are in synergy producing innovative and unique results.

## Acknowledgements

S. C. is supported by the HICCUP ERC Consolidator grant (614577) and by the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 740006).

## References

- [1] M. Abadi, et al. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
- [2] Pedregosa, F. *et al.* Journal of Machine Learning Research, volume 12, 2825–2830, 2011
- [3] F. Chollet. keras. <https://github.com/fchollet/keras>, 2015.

- [4] Theano Development Team [arXiv:1605.02688], May 2016,
- [5] R. Boughezal, X. Liu and F. Petriello, *JHEP* **1703**, 160 (2017) doi:10.1007/JHEP03(2017)160 [arXiv:1612.02911 [hep-ph]].
- [6] G. Altarelli, R. D. Ball and S. Forte, *Nucl. Phys. B* **799** (2008) 199 doi:10.1016/j.nuclphysb.2008.03.003 [arXiv:0802.0032 [hep-ph]].
- [7] P. Nason, *JHEP* **0411** (2004) 040 doi:10.1088/1126-6708/2004/11/040 [hep-ph/0409146].
- [8] J. Alwall *et al.*, *JHEP* **1407** (2014) 079 doi:10.1007/JHEP07(2014)079 [arXiv:1405.0301 [hep-ph]].
- [9] T. Sjostrand, S. Mrenna and P. Z. Skands, *JHEP* **0605** (2006) 026 doi:10.1088/1126-6708/2006/05/026 [hep-ph/0603175].
- [10] M. Bahr *et al.*, *Eur. Phys. J. C* **58** (2008) 639 doi:10.1140/epjc/s10052-008-0798-9 [arXiv:0803.0883 [hep-ph]].
- [11] R. Boughezal, J. M. Campbell, R. K. Ellis, C. Focke, W. Giele, X. Liu, F. Petriello and C. Williams, *Eur. Phys. J. C* **77** (2017) no.1, 7 doi:10.1140/epjc/s10052-016-4558-y [arXiv:1605.08011 [hep-ph]].
- [12] 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]].
- [13] T. Kluge, K. Rabbertz and M. Wobisch, doi:10.1142/9789812706706\_0110 hep-ph/0609285.
- [14] A. van Hameren, *Comput. Phys. Commun.* **182** (2011) 2427 doi:10.1016/j.cpc.2011.06.011 [arXiv:1007.4716 [hep-ph]].
- [15] S. Carrazza, R. K. Ellis and G. Zanderighi, *Comput. Phys. Commun.* **209** (2016) 134 doi:10.1016/j.cpc.2016.07.033 [arXiv:1605.03181 [hep-ph]].
- [16] T. Hahn and M. Perez-Victoria, *Comput. Phys. Commun.* **118** (1999) 153 doi:10.1016/S0010-4655(98)00173-8 [hep-ph/9807565].
- [17] S. Catani, L. Cieri, G. Ferrera, D. de Florian and M. Grazzini, *Phys. Rev. Lett.* **103** (2009) 082001 doi:10.1103/PhysRevLett.103.082001 [arXiv:0903.2120 [hep-ph]].
- [18] C. Anastasiou, C. Duhr, F. Dulat, F. Herzog and B. Mistlberger, *Phys. Rev. Lett.* **114** (2015) 212001 doi:10.1103/PhysRevLett.114.212001 [arXiv:1503.06056 [hep-ph]].
- [19] B. Ruijl, T. Ueda and J. Vermaseren, arXiv:1707.06453 [hep-ph].
- [20] P. Nogueira, *J. Comput. Phys.* **105** (1993) 279. doi:10.1006/jcph.1993.1074
- [21] R. D. Ball *et al.* [NNPDF Collaboration], *Eur. Phys. J. C* **77** (2017) no.10, 663 doi:10.1140/epjc/s10052-017-5199-5 [arXiv:1706.00428 [hep-ph]].
- [22] V. Bertone *et al.* [NNPDF Collaboration], *Eur. Phys. J. C* **77** (2017) no.8, 516 doi:10.1140/epjc/s10052-017-5088-y [arXiv:1706.07049 [hep-ph]].
- [23] P. Skands, S. Carrazza and J. Rojo, *Eur. Phys. J. C* **74** (2014) no.8, 3024 doi:10.1140/epjc/s10052-014-3024-y [arXiv:1404.5630 [hep-ph]].
- [24] P. T. Komiske, E. M. Metodiev and M. D. Schwartz, *JHEP* **1701**, 110 (2017) doi:10.1007/JHEP01(2017)110 [arXiv:1612.01551 [hep-ph]].
- [25] J. Butterworth *et al.*, *J. Phys. G* **43** (2016) 023001 doi:10.1088/0954-3899/43/2/023001 [arXiv:1510.03865 [hep-ph]].
- [26] S. Carrazza, *Acta Phys. Polon. B* **48**, 947 (2017) doi:10.5506/APhysPolB.48.947 [arXiv:1704.00471 [hep-ph]].
- [27] Likas, Aristidis Computer physics communications, Vol. 135, 2, 167–175, 2001
- [28] D. Krefl, S. Carrazza, B. Hagherat and J. Kahlen, arXiv:1712.07581 [stat.ML].
- [29] J. Bendavid, arXiv:1707.00028 [hep-ph].
- [30] R. D. Ball *et al.* [NNPDF Collaboration], *JHEP* **1504** (2015) 040 doi:10.1007/JHEP04(2015)040 [arXiv:1410.8849 [hep-ph]].
- [31] S. Carrazza and N. P. Hartland, arXiv:1711.09991 [hep-ph].
- [32] E. R. Nocera *et al.* [NNPDF Collaboration], *Nucl. Phys. B* **887** (2014) 276 doi:10.1016/j.nuclphysb.2014.08.008 [arXiv:1406.5539 [hep-ph]].
- [33] L. A. Harland-Lang, A. D. Martin, P. Motylinski and R. S. Thorne, *Eur. Phys. J. C* **75** (2015) no.5, 204 doi:10.1140/epjc/s10052-015-3397-6 [arXiv:1412.3989 [hep-ph]].
- [34] S. Dulat *et al.*, *Phys. Rev. D* **93** (2016) no.3, 033006 doi:10.1103/PhysRevD.93.033006 [arXiv:1506.07443 [hep-ph]].
- [35] G. Watt and R. S. Thorne, *JHEP* **1208** (2012) 052 doi:10.1007/JHEP08(2012)052 [arXiv:1205.4024 [hep-ph]].
- [36] S. Carrazza, J. I. Latorre, J. Rojo and G. Watt, *Eur. Phys. J. C* **75** (2015) 474 doi:10.1140/epjc/s10052-015-3703-3 [arXiv:1504.06469 [hep-ph]].
- [37] S. Carrazza and J. I. Latorre, arXiv:1605.04345 [hep-ph].
- [38] S. Carrazza, S. Forte, Z. Kassabov, J. I. Latorre and J. Rojo, *Eur. Phys. J. C* **75** (2015) no.8, 369 doi:10.1140/epjc/s10052-015-3590-7 [arXiv:1505.06736 [hep-ph]].
- [39] S. Carrazza, S. Forte, Z. Kassabov and J. Rojo, *Eur. Phys. J. C* **76** (2016) no.4, 205 doi:10.1140/epjc/s10052-016-4042-8 [arXiv:1602.00005 [hep-ph]].
- [40] S. Carrazza and Z. Kassabov, *PoS PP @LHC2016* (2016) 020 [arXiv:1606.09248 [hep-ph]].