PROCEEDINGS OF SCIENCE

Machine learning approaches for parameter reweighting in MC samples of top quark production in CMS

Valentina Guglielmi[1,*]

Deutsches Elektronen-Synchrotron (DESY),
Notkestrasse 85, Hamburg, Germany

E-mail: valentina.guglielmi@desy.de

In particle physics, Monte Carlo (MC) event generators are needed to compare theory to the measured data. Many MC samples have to be generated to account for theoretical systematic uncertainties, at a significant computational cost. Therefore, the MC statistic becomes a limiting factor for most measurements and the significant computational cost of these programs a bottleneck in most physics analyses. In this contribution, the Deep neural network using Classification for Tuning and Reweighting (DCTR) approach is evaluated for the reweighting of two systematic uncertainties in MC simulations of top quark pair production within the CMS experiment. DCTR is a method, based on a Deep Neural Network (DNN) technique, to reweight simulations to different model parameters by using the full kinematic information in the event. This methodology avoids the need for simulating the detector response multiple times by incorporating the relevant variations in a single sample.

European Physical Society, EPS2023
20-25 August, 2023
Hamburg, Germany

[1]For the CMS Collaboration.
[*]Speaker

PoS(EPS-HEP2023)325

1. Introduction

In particle physics, Monte Carlo (MC) event generators are essential for comparing theoretical predictions with experimental data. Generating numerous MC samples is necessary to account for theoretical uncertainties, which comes at a significant computational cost. As a result, the limitations of MC statistics hinder many measurements, and the high computational expense poses a bottleneck for physics analyses. For instance, in a recent study on top quark-antiquark pair production, the main source of uncertainty in the mass of the top quark came from the MC statistics of the samples used for systematic estimation [1]. To address this issue, reweighting the MC samples can be a solution. This approach involves generating only a sample with nominal values, and then variations are obtained by reweighting this nominal sample. By doing so, the need for simulating the detector response multiple times is eliminated, reducing the MC statistics and computational cost. In contrast to traditional reweighting, which compares distributions in specific bins at a truth level, Machine Learning (ML) reweighting offers a more flexible approach. Standard reweighting is limited by the choice of binning and the number of input dimensions. ML reweighting, using a neural network, does not have such limitations. It can utilize all event information as input, improving reweighting precision. Additionally, it allows for the simultaneous reweighting of multiple MC parameters, considering their correlations. This project introduces Deep Neural Network (DNN) using Classification for Tuning and Reweighting (DCTR) [2] and tests its performance in two scenarios [3]: discrete reweighting of the h_{damp} variation in parton-level Powheg HVQ events and continuous reweighting of a b-fragmentation function parameter in particle-level Pythia8 events.

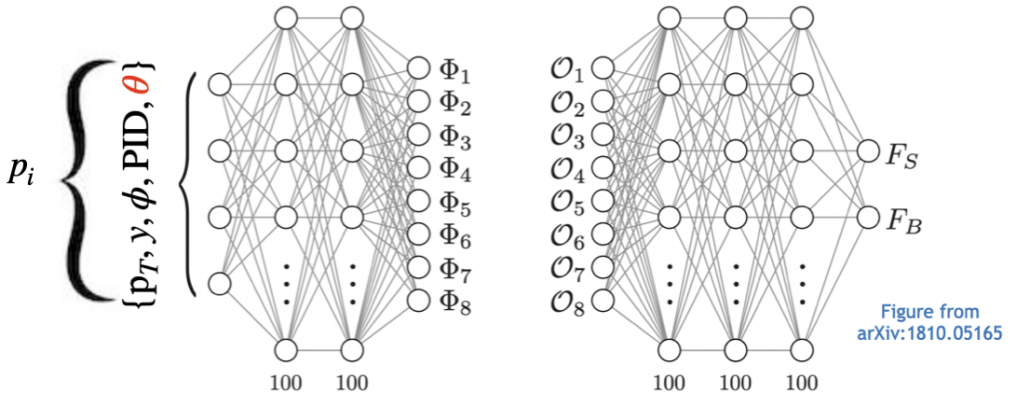

Figure 1: The PFN architecture used in the DCTR method is depicted. It parametrizes the per-particle mapping Φ (on the left) and the function F (on the right), shown for the case of a latent space of dimension $l = 8$. The latent observable is $O_{\mathrm{a}} = \sum_i \Phi_{\mathrm{n}}(p_{\mathrm{T}}, y_{\mathrm{i}}, \phi_{\mathrm{i}}, m_{\mathrm{i}}, PID_{\mathrm{i}})$ [4].

2. h_{damp} dicrete reweighting

The default MC sample of top pair production in CMS is generated using the Heavy Quark Process (HVQ) [5] of the event generator Powheg [6][7]. In Powheg, the resummation of the next-to-leading-order radiation is regulated by the h_{damp} variable, which enters the damping parameter

D as in Eq. 1:

$$D = \frac{h_{\text{damp}}^2}{p_{\text{T}}^2 + h_{\text{damp}}^2} \tag{1}$$

where p_{T} is the transverse momentum of the particle and h_{damp} a parameter defined as $h_{\text{damp}} = h \cdot m_{\text{t}}$, where m_{t} is the mass of the quark top (m_{t}=172.5 GeV) and h is a real number. Since the parameter h_{damp} is not physical, an arbitrary value must be chosen in the simulation and varied to calculate the associated systematic uncertainty. Two variations (down and up) from its nominal value are considered in CMS. The nominal value of h_{damp} is set to $1.379 \cdot m_{\text{t}}$, while the down (up) variation is $0.8738 \cdot m_{\text{t}}$ ($2.305 \cdot m_{\text{t}}$). Two separate neural network models are trained to reweight the nominal h_{damp} to these variations. For the training 40 million events are generated for each h_{damp} value, and parton-level information is used as input to the neural network. The Particle Flow Network (PFN) is employed for training, passing the quadrimomentum and the particle PID (p_{T}, y, ϕ, m, PID) of the top and antitop to it. The performance of reweighting is checked by applying the weights to the p_{T} and η observables of the $\text{t}\bar{\text{t}}$ system. Results, reported in Fig. 2, indicate that the reweighted and original samples agree with a method closure within 2%. In these proceedings, we present results for the down variation of h_{damp}, but similar results are obtained for the up variation as reported in [3].

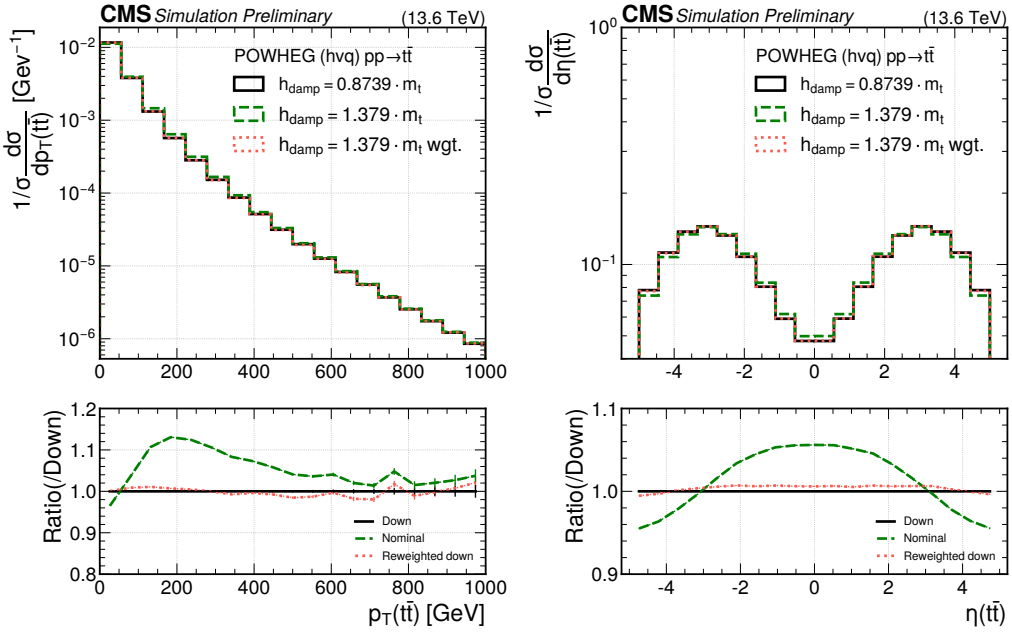

Figure 2: The normalised differential cross section as a function of the p_{T} (left) or η (right) of the $\text{t}\bar{\text{t}}$ system at parton level for a MC simulated $\text{t}\bar{\text{t}}$ sample generated with the POWHEG hvq program, [6] [7], for the CMS down variation of h_{damp} ($0.8739 \cdot m_{\text{t}}$) in black and the nominal value of h_{damp} ($1.379 \cdot m_{\text{t}}$) in green. The red line shows the nominal sample reweighted to the down h_{damp} variation using the DCTR method [2]. The bottom pad shows the ratio of the nominal sample to the h_{damp} down variation before and after the reweighting. The error bars represent the statistical uncertainties due to the limited statistics of the MC.

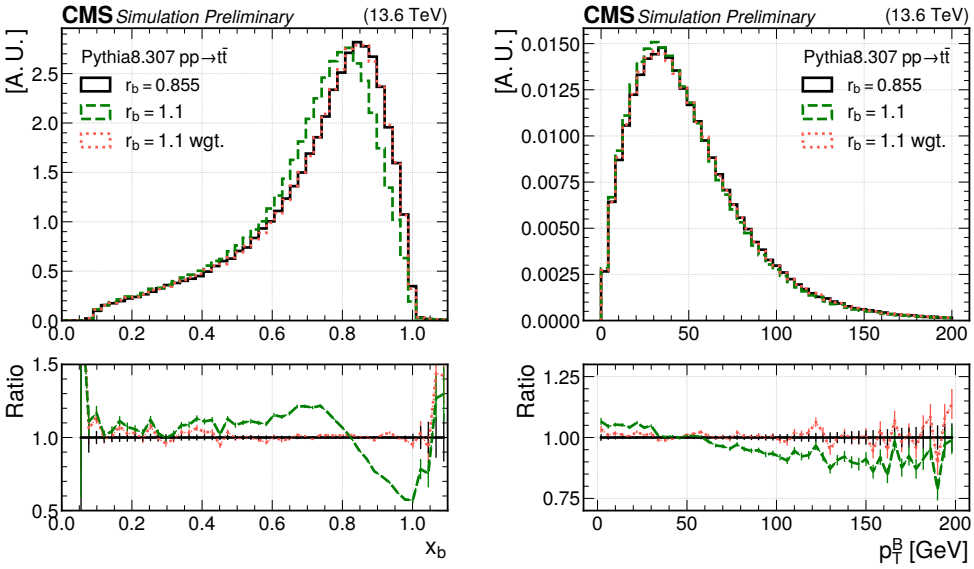

Figure 3: The distribution of x_b (left) or p_T^B of the B-hadron (right) of the $t\bar{t}$ system for a MC simulated $t\bar{t}$ sample generated with Pythia 8 [8], for the CMS nominal variation of r_b (0.855) in black and a second value of r_b (1.1) in green. The red line shows the sample with the variation of r_b reweighted to the sample generated with the nominal r_b using the DCTR method, [2]. The bottom pad shows the ratio of the sample with the variation of r_b to the sample generated with the nominal r_b before and after the reweighting. The error bars represent the statistical uncertainties due to the limited statistics of the MC.

3. B-fragmentation continous reweighting

Another significant source of uncertainty in top physics is the B-fragmentation. It involves the decay of a top quark into a b quark and the subsequent formation of a B-hadron through a process described by the Lund string model in Pythia [8]. The probability distribution for heavy quarks during this process is given by Eq. 2:

$$f_B(z) \propto \frac{1}{z^{1+br_b m_b^2}}(1-z)^a exp(-bm_t^2/z) \tag{2}$$

where z is the quark longitudinal mometum, m_t the top quark mass and a and b free parameters to be tuned to experimental data. For b quarks, the Bowler modification is also considered, involving the b quark mass m_b and the r_b parameter in Pythia. To assess systematic uncertainties, CMS recommends varying the r_b parameter during MC event generation. In this project, the DCTR technique is employed to estimate the uncertainty related to B-fragmentation, which affects physical observables like x_b and p_T^B representing the energy fraction transferred from the b quark to the B-hadron and the transverse momentum of the B-hadron, respectively. In this scenario, a continous reweighting with 10 different values of r_b parameter in Pythia in the range [0.6, 1.4] has been performed training a single DNN model to reweight the samples generated with the different r_b values to the nominal sample with r_b (0.855). For the nominal r_b sample 5M events and 5k events for each r_b variation sample are passed as inputs to the DNN for a total amount of 10M events to

train. The x_{b} variable, defined in Eq 3, is passed as the only input to the DNN:

$$x_b = \frac{2p_B \cdot q}{m_t^2} \cdot \frac{1}{1 - m_w^2/m_t^2} \tag{3}$$

where p_{B} and q are the four vector of the B-hadron and of the top quark and m_{W} and m_{t} the W boson and top quark mass, respectively. The performance of reweighting is checked by applying the weights obatined in the training to the x_{b} and $p_{\mathrm{T}}^{\mathrm{B}}$ observables of the $\mathrm{t\bar{t}}$ system. The results for one of the 9 values of r_{b} (1.1) trained are shown in Fig. 2. The original sample and the reweighted one agree also in this case with a method closure within 2%. Similar results are obtained for all the other 8 values of r_{b} tested, as reported in [3].

4. Results and Conclusions

In these proceedings, the DCTR method is evaluated in two different scenarios: discrete reweighting of the h_{damp} variation in Parton-level Powheg HVQ events and continuous reweighting of a b-fragmentation function parameter in particle-level Pythia events. The method is found to work very well in both scenarios, with a method clusure within 2%. The method is implemented into CMS software framework for both cases. This approach has the potential to be applied to various other intriguing cases in the field of top physics and, more broadly, in other physics areas.

References

[1] CMS collaboration, *Measurement of the* $\mathrm{t\bar{t}}$ *production cross section, the top quark mass, and the strong coupling constant using dilepton events in pp collisions at* $\sqrt{s}$ *= 13 TeV*, *EPJC* **79** (2019) no.5, 368 [arXiv:1812.10505]

[2] A. Andreassen and B. Nachman, *Neural Networks for Full Phase-space Reweighting and Parameter Tuning*, *PRD* **101** (2020) no.9, 091901 [arXiv:1907.08209].

[3] CMS collaboration, *Machine learning approaches for parameter reweighting in MC samples of top quark production in CMS* *CMS-DP-2023-031*

[4] P. T. Komiske, E. M. Metodiev and J. Thaler, *Energy Flow Networks: Deep Sets for Particle Jets*, *JHEP* **01** (2019) 121 [arXiv:1810.05165]

[5] S. Frixione, P. Nason and G. Ridolfi, *A positive-weight next-to-leading-order Monte Carlo for heavy flavour hadroproduction*, *JHEP* **09** (2007) 126 [arXiv:0707.3088]

[6] S. Frixione, P. Nason and C. Oleari, *Matching NLO QCD computations with Parton Shower simulations: the POWHEG method*, *JHEP* **11** (2007), 070 [arXiv:0709.2092]

[7] S. Alioli, P. Nason, C. Oleari and E. Re, *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, *JHEP* **06** (2010), 043 [arXiv:1002.2581]

[8] C. Bierlich, S. Chakraborty, N. Desai, L. Gellersen, I. Helenius, P. Ilten, L. Lönnblad, S. Mrenna, S. Prestel and C. T. Preuss, *et al. A comprehensive guide to the physics and usage of PYTHIA 8.3*, *arXiv:2203.11601*