Adversarial Neural Network-based data-simulation corrections for jet-tagging at CMS

Martin Erdmann, Benjamin Fischer, Dennis Noll, Yannik Alexander Rath, Marcel Rieger and David Josef Schmidt on behalf of the CMS Collaboration

RWTH Aachen University

E-mail: benjamin.fischer@rwth-aachen.de

Abstract. Variable-dependent scale factors are commonly used in HEP to improve shape agreement of data and simulation. The choice of the underlying model is of great importance, but often requires a lot of manual tuning e.g. of bin sizes or fitted functions. This can be alleviated through the use of neural networks and their inherent powerful data modeling capabilities. We present a novel and generalized method for producing scale factors using an adversarial neural network. This method is investigated in the context of the bottom-quark jet-tagging algorithms within the CMS experiment. The primary network uses the jet variables as inputs to derive the scale factor for a single jet. It is trained through the use of a second network, the adversary, which aims to differentiate between the data and rescaled simulation.

1. Introduction

In high energy physics collider experiments jets originating from a bottom quark can be identified by the formation of a secondary displaced decay vertex due to the comparatively long lifetime of the intermediately created B-meson. This and other characteristics are used in specialized algorithms, so called bottom quark taggers (b-taggers), to identify these jets by scoring their bottom quark-ness on a continuous scale. The algorithms are constructed using simulated events, which are known to not exactly replicate the recorded data. The resulting differences between data and simulation are reduced by applying event weight factors, commonly referred to as scale factors (SFs), which are derived from dedicated measurements. The event scale factors are a product of the corresponding object scale factors, one for each jet within the respective event. These object scale factors can either be derived for specific working points, in case of threshold like usage of the b-tagger score (b-tag discriminant), or such that the entire score distribution is corrected - with the latter being of interest here.

This article first introduces the currently employed traditional approach for calculation of such scale factors. Then a new neural network-based approach is motivated and detailed, followed by a discussion of the preliminary results. This study is conducted within the CMS experiment [1].

2. Current Scale Factor Method

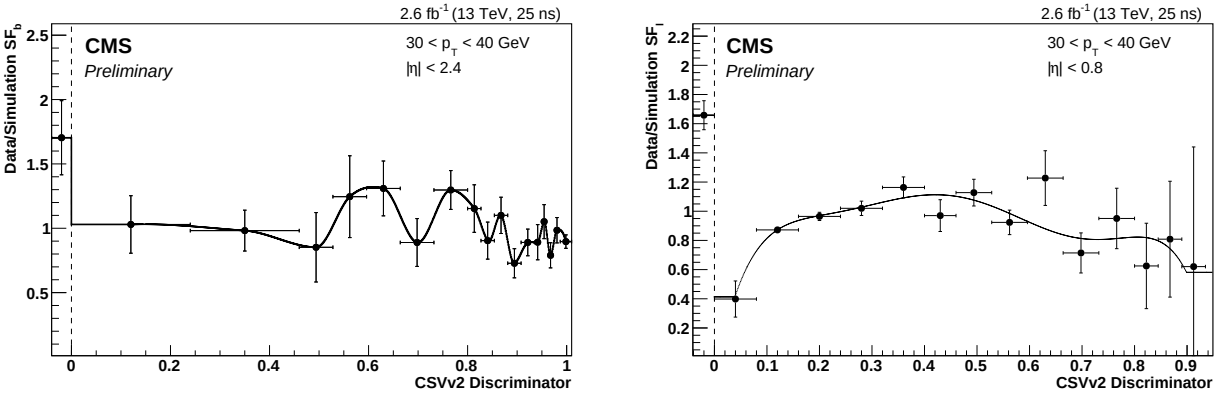
The currently employed method for the production of shape-changing scale factors for the b-tagger score distribution is described in detail in [2, 3] and only a brief outline will be given here.

A tag-and-probe method is applied to events with two leptons and exactly two additional jets. Using the tag-jet, two regions are established by a corresponding cut on the b-tagger score value: one containing events enriched in light flavor (u, d, s; LF) jets and the other containing events enriched in heavy flavor (b; HF) jets. The regions primarily consist of Drell-Yan processes and top quark pair production, respectively. These flavor regions are additionally divided into several kinematic regions based on transverse jet momentum (p_T) and, in the case of light flavor jets, absolute pseudorapidity ($|\eta|$).

Within each kinematic region the events are histogrammed in the b-tag discriminant, with varying bin-widths to account for the uneven distribution of the events. The scale factor for each flavor f is then calculated from the $Data_f/Sim_f$ ratio in every bin. Since there is no jet flavor generator information in the data, $Data_f$ is modeled by subtracting all simulation of other flavors ($Sim_{\neg f}$) from the entirety of the data events ($Data$). The simulation events $Sim_{\neg f}$ are each weighted according to their event weights w_i which are arising from the application of the respective b-tagger scale factors SF_{f_j} for all the event's jets j. This constitutes a circular dependency, which is resolved in an iterative manner with all scale factors of the first iteration ($SF_{f,i=0}$) set to 1. These iterations i are repeated until sufficient convergence is reached. Consequently the scale factors are determined as:

$$SF_{f,i+1} = \frac{Data_f}{Sim_f} = \frac{Data - w_i \odot Sim_{\neg f}}{Sim_f} \quad , \quad w_i = \prod_{jets}^{j} SF_{f_j,i} \quad , \quad SF_{f,0} = 1 \tag{1}$$

Finally, for each kinematic region and flavor, the determined scale factors are smoothed by fitting an appropriately chosen function. Figure 1 shows two examples from an earlier data taking period.



(a) Heavy flavor jets (b) Light flavor jets

Figure 1: Examples of scale factors for heavy (a) and light (b) flavored jets for the CSVv2 bottom quark tagger as used for the 2015 data taking period. The measured data/simulation ratio is shown as points with the corresponding statistical uncertainties. The solid line represents the scale factors as described by the chosen fitted function for each particular region, a spline (a) and polynomial (b) respectively. [2]

3. Neural Network-based Approach

3.1. Concept

The studied neural network approach intends to simplify the procedure by which the scale factors are produced while maintaining the performance and increasing flexibility. In particular,

it proposes the use of one neural network per flavor, which directly receives p_T, $|\eta|$, and b-tag discriminant as input and produces the according scale factor. In effect, these neural networks replace the binned fit of explicitly chosen smoothing functions while the remainder of the traditional approach remains unchanged.

This has the following advantages: first, there is no need to bin and optimize the binning of the input data, neither for kinematic regions using p_T or $|\eta|$ nor along the b-tag discriminant; additionally, this enables the addition of further input variables as the network is able to cope with the curse of dimensionality; furthermore, there is no need to explicitly choose the type and parameters of the fitted function, as a sufficiently large neural network can model practically any reasonable function [4].

3.2. Training Setup

For a supervised training of the scale factor network (SF-net) the target output values are not available on event by event basis since the objective of the training is to optimize a summary statistic, that is to minimize the difference in the density distributions of data and simulation by reweighting the latter. To make the training possible, this summary statistic, in form of a density distribution ratio, is encoded into the data-simulation discrimination network (Discriminator), which is trained to differentiate between data and simulation. Then, the SF-net can be trained through adversarial feedback from the Discriminator.

In detail, the Discriminator is trained until convergence to differentiate between data and the simulation reweighed by the SF-net output, thus capturing the residual differences between the two. Then, the SF-net is briefly trained to minimize the square of these residual differences. This process is repeated until the SF-net has converged, which means that its performance does not improve for several training steps. A scheme of this training setup is outlined in Fig. 2.

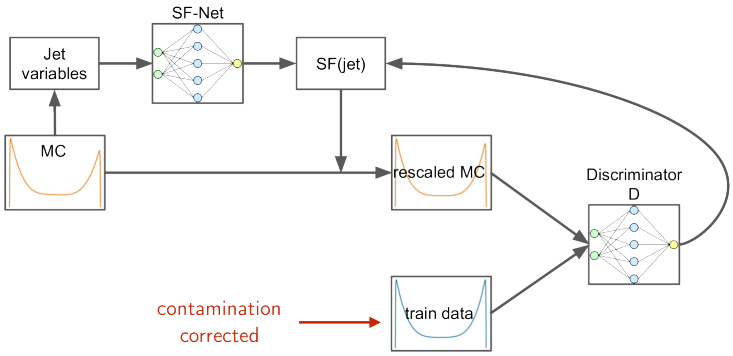


Figure 2: Scheme of neural network setup. The jet variables of the simulation samples are used by the scale factor network (SF-net) to produce scale factors and subsequently a reweighed simulation. These samples together with contamination-corrected data samples are processed by the data-simulation discrimination network (Discriminator). Using its output the SF-net is trained in an adversarial fashion to minimize the differences between data and simulation [5].

As with the traditional approach, the data also contain a contamination of unidentifiable events of the wrong flavor, which are accounted for by subtracting the corresponding simulation events through an event weight sign flip. Again, this necessitates an iterative approach of the two flavors, since these subtracted events will need to incorporate the corresponding scale factors, as was described in Section 2.

3.3. Implementation

Both the SF-net and Discriminator consist of a 10 layer 128 unit densely connected network with LeakyReLU ($\alpha = 0.01$) activations. The input for both are the p_T, $|\eta|$, and b-tag discriminant of the probe jet, with the according event weights folded into the training loss. The Discriminator has a Sigmoid output activation and minimizes the binary cross-entropy. The SF-net uses a squared difference loss between its output SF, which has a Softplus activation ($\log(1+e^x)$), and the constant target value of $SF \cdot R$. The data-simulation ratio $R = D/(1-D)$ is calculated using the Discriminator output D. During the alternating training the number of consecutive training steps is limited to 12 for the Discriminator and a single one for the SF-net. Both networks are trained in batches of 4096 samples until convergence, which typically occurs for the SF-net after a total of about 50 epochs. Additionally, regularization is done through a weight decay with strength 3×10^{-5}. None of these parameters were significantly tuned for this application. The training is also monitored using an independent validation data set which is about a fourth of the size of the training data set, in order to verify the absence of overtraining. Both flavors are handled in the exact same manner.

Additionally, the entire procedure, including the iterative handling of the contamination subtraction, is repeated 25 times with different random seeds and the final scale factors are combined in an unweighted ensemble via their mean. The fluctuations within this ensemble can then be used to gauge the stability of the scale factors for any particular phase space point, but should not be equated with systematic or statistical uncertainties although a strong correlation with the latter is expected. The implementation is available at [6].

The data of the 2017 (Run II) data taking period amount to $41.3\,\mathrm{fb}^{-1}$ of integrated luminosity and yield about 2.4 million samples with about 12.5 million samples of corresponding simulation. These samples are prepared equivalently to those used for the measurement of the traditional scale factors, as is described in [2]. The commonly used $\mathrm{DeepCSV_{b+bb}}$ bottom quark tagger is the chosen subject of study, which operates on jets clustered with the anti-$k_\mathrm{T}$algorithm with $R = 0.4$ and applied charged hadron subtraction.

4. Results

The performance of the neural network-based approach is inspected by comparing the resulting scale factors to those of the traditional approach.

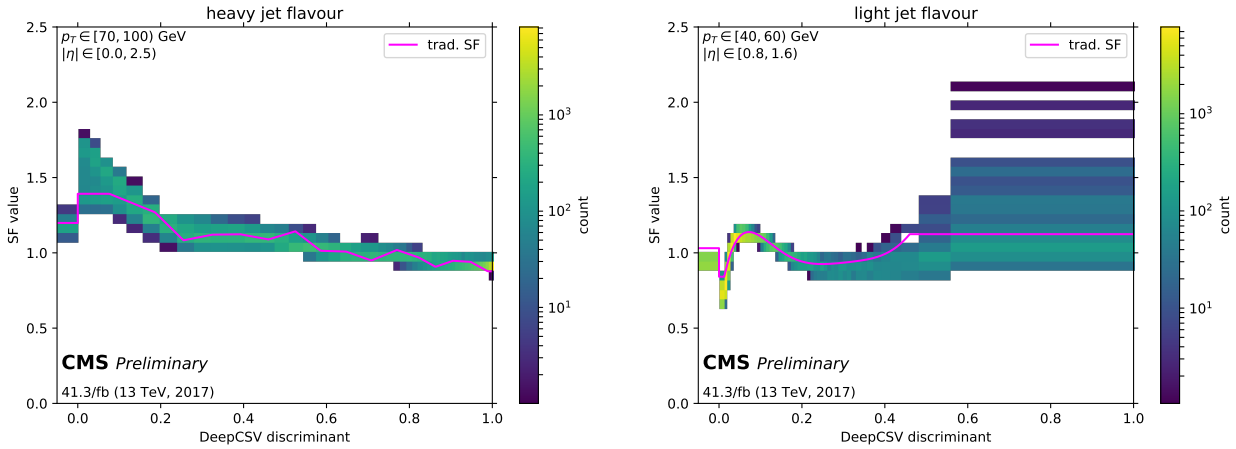


Figure 3: Scale factor (SF) values vs. b-tag discriminant, each in an example kinematic region for heavy (left) and light (right) flavored probe jets. The traditional scale factors (trad. SFs) are shown as a pink curve. The DeepSFs are shown as a 2D histogram as they can have different values due to their direct p_T and η dependence. The bin widths are adapted such that regions of high statistics (yellow areas) are finely binned [5].

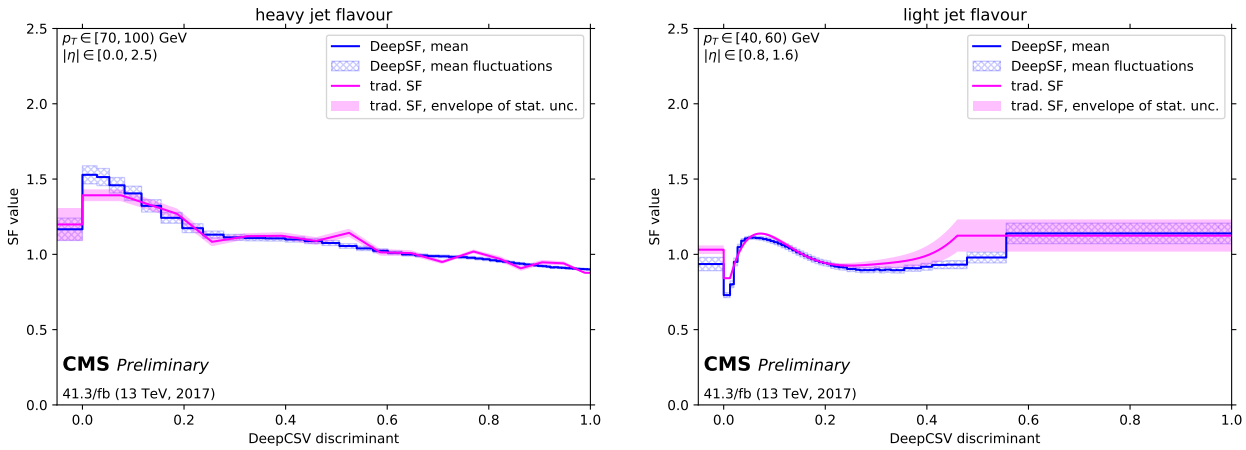


Figure 4: Scale factor (SF) values vs. b-tag discriminant, each in an example kinematic region for heavy (left) and light (right) flavored probe jets. The traditional scale factors (trad. SFs) and their statistical uncertainties are shown as a pink curve and pink shaded area. The DeepSFs are represented by a blue curve indicating the mean value for each b-tag discriminant bin. Additionally, the mean size of the ensemble fluctuations (standard error on mean) is indicated by the blue hatched area [5].

As can be seen in Figs. 3 and 4, the DeepSFs generally match the shapes of the traditional scale factors quite well. In particular, they smoothly follow complex shapes such as those seen for the light jet flavor scale factors in Fig. 4. On the other hand, they are sensitive to kinematic variables (p_T and $|\eta|$) where demanded by the objective, which is visible in the heavy jet flavor example for b-tag discriminant values between 0 and 0.1.

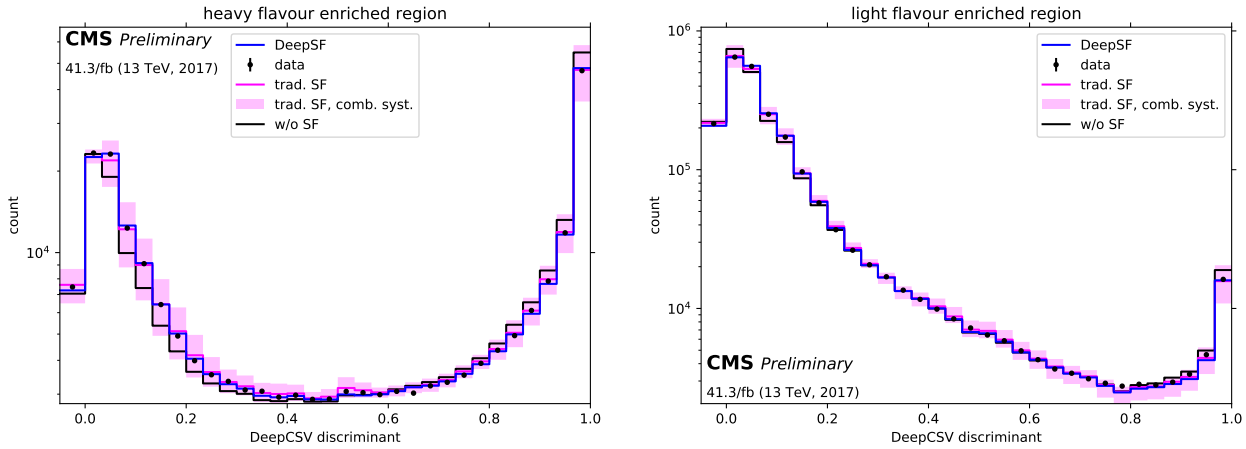


Figure 5: The plots show the distribution of the b-tag discriminant of all probe jets for the heavy (left) and light (right) flavor enriched regions. The data are shown as black points. The simulation is displayed as solid histograms with the different colors representing different types of scale factors applied: no scale factors in black, the traditional scale factors in pink, and the DeepSF in blue. Additionally, the pink shaded area shows the impact of all the combined systematic uncertainties of the traditional scale factors [5].

When inspecting the b-tag discriminant distribution shown in Fig. 5 with scale factors applied, it can be seen that both the traditional approach and the neural network-based approach fulfill their intended purpose of reducing the differences between data and uncorrected simulation.

5. Conclusion

The study presented in this article shows that a neural network-based approach is able to produce shape-changing scale factors which are comparable to those of the currently used traditional approach. The notable difference being that neither tuning of binning or fitted function types and parameters nor any other artificial constraints are necessary to achieve this. Last, the neural network-based approach is inherently capable of being extended further to incorporate more input variables or flavor enriched regions. As such, it can be considered a promising alternative which warrants more detailed investigation.

References

[1] CMS Collaboration 2008 *JINST* **3** S08004. 361 p URL https://cds.cern.ch/record/1129810

[2] CMS Collaboration 2016 Identification of b quark jets at the CMS Experiment in the LHC Run 2 Tech. Rep. CMS-PAS-BTV-15-001 CERN Geneva URL https://cds.cern.ch/record/2138504

[3] CMS Collaboration 2013 Search for Higgs Boson Production in Association with a Top-Quark Pair and Decaying to Bottom Quarks or Tau Leptons Tech. Rep. CMS-PAS-HIG-13-019 CERN Geneva URL https://cds.cern.ch/record/1564682

[4] Lu Z, Pu H, Wang F, Hu Z and Wang L 2017 *Advances in Neural Information Processing Systems 30* pp 6231–6239 (*Preprint* 1709.02540)

[5] CMS Collaboration 2019 Adversarial neural network-based data-simulation corrections for heavy-flavour jet-tagging Tech. Rep. CMS-DP-2019-003 URL http://cds.cern.ch/record/2666647

[6] DeepSF [software repository] URL https://gitlab.cern.ch/aachen-3a-cms/jet-tagging-sf/tree/deepSF