

# Parametrized classifiers for optimal EFT sensitivity

Siyu Chen,<sup>a</sup> Alfredo Glioti,<sup>a</sup> Giuliano Panico<sup>b,c</sup> and Andrea Wulzer<sup>a,d,e</sup>

<sup>a</sup>*Theoretical Particle Physics Laboratory (LTP), Institute of Physics, EPFL, Lausanne, Switzerland*

<sup>b</sup>*INFN, Sezione di Firenze, Via G. Sansone 1, 50019 Sesto Fiorentino, Italy*

<sup>c</sup>*Dipartimento di Fisica e Astronomia, Università di Firenze, Via G. Sansone 1, 50019 Sesto Fiorentino, Italy*

<sup>d</sup>*CERN, 1211 Geneva 23, Switzerland*

<sup>e</sup>*Dipartimento di Fisica e Astronomia, Università di Padova, Padova, Italy*

*E-mail:* [siyu.chen@epfl.ch](mailto:siyu.chen@epfl.ch), [alfredo.glioti@epfl.ch](mailto:alfredo.glioti@epfl.ch), [giuliano.panico@unifi.it](mailto:giuliano.panico@unifi.it), [andrea.wulzer@cern.ch](mailto:andrea.wulzer@cern.ch)

**ABSTRACT:** We study unbinned multivariate analysis techniques, based on Statistical Learning, for indirect new physics searches at the LHC in the Effective Field Theory framework. We focus in particular on high-energy ZW production with fully leptonic decays, modeled at different degrees of refinement up to NLO in QCD. We show that a considerable gain in sensitivity is possible compared with current projections based on binned analyses. As expected, the gain is particularly significant for those operators that display a complex pattern of interference with the Standard Model amplitude. The most effective method is found to be the “Quadratic Classifier” approach, an improvement of the standard Statistical Learning classifier where the quadratic dependence of the differential cross section on the EFT Wilson coefficients is built-in and incorporated in the loss function. We argue that the Quadratic Classifier performances are nearly statistically optimal, based on a rigorous notion of optimality that we can establish for an approximate analytic description of the ZW process.

**KEYWORDS:** Beyond Standard Model, Effective Field Theories

**ARXIV EPRINT:** [2007.10356](https://arxiv.org/abs/2007.10356)

---

**Contents**

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Teaching new physics to a machine</b>	<b>5</b>
2.1	The Standard Classifier	7
2.2	The Quadratic Classifier	8
<b>3</b>	<b>Fully leptonic ZW</b>	<b>11</b>
3.1	Analytic approximation	14
3.2	Monte Carlo generators	15
<b>4</b>	<b>Optimality on Toy data</b>	<b>17</b>
4.1	Results	19
4.2	MADGRAPH Leading Order	21
<b>5</b>	<b>The reach at Next-to-Leading Order</b>	<b>22</b>
5.1	Estimating the test statistics distributions	22
5.2	Results	24
<b>6</b>	<b>Neural Network implementation and validation</b>	<b>25</b>
6.1	The Quadratic Classifier	26
6.2	The Standard Classifier	27
6.3	Validation	28
<b>7</b>	<b>Conclusions and outlook</b>	<b>30</b>
<b>A</b>	<b>The general Quadratic Classifier</b>	<b>33</b>
<b>B</b>	<b>Minimization of the parametrized loss</b>	<b>34</b>

---

**1 Introduction**

The amazing richness of LHC data makes searching for new physics an extremely complex process. Three main steps can be identified, taking however into account that they are strongly interconnected and not necessarily sequential in time. First, we need a target new physics theory. In our case this is provided by the Standard Model (SM) itself, supplemented by operators of energy dimension  $d > 4$  that encapsulate the indirect effects of heavy new particles and interactions. This setup is often dubbed the SM Effective Field Theory (EFT) in the context of high-energy physics (see e.g. refs. [1–3]). However the EFT approach is extremely common and widely employed in many other domains, eminently in Flavor Physics. The methodologies discussed in this paper could thus find applications also in other areas.

The second step is to turn the new physics theory into concrete predictions. These should be sufficiently accurate, since the EFT operator effects are often a small correction to the pure SM predictions. The predictions are provided by Monte Carlo generator codes, which produce event samples that are representative of the true particles momenta distributions. Accurate simulation of the detector response are further applied in order to obtain a representation of the distribution as a function of the variables that are actually observed in the experiment. It should be mentioned that this program occasionally fails. Namely it could be impossible for the Monte Carlo codes to provide a sufficiently accurate representation of all the components of the data distribution, for instance of reducible backgrounds from misidentification. In this case, the artificial Monte Carlo event sample should be supplemented with natural data collected in some control region, which model the missing component. We will not discuss this case explicitly, however it should be emphasized that our methodology would apply straightforwardly. Namely, the “Monte Carlo” samples we refer to in what follows might well not be the output of a Monte Carlo code, but rather have (partially) natural origin.

The final step, i.e. the actual comparison of the predictions with the data, is often further split in two, by identifying suitable high-level observables (e.g. cross sections in bins) that are particularly sensitive to the EFT operators. These observables can be measured in an experimental analysis that does not target the EFT explicitly, and compared with the EFT predictions at a later stage. This is convenient from the experimental viewpoint because the results are model independent and thus potentially useful also to probe other new physics theories. If the measurements are performed at the truth (unfolded) level, this is also convenient for theorists because detector effects need not to be included in the predictions. However a strategy based on intermediate high-level observables is unavoidably suboptimal. It would approach optimality only if the fully differential distribution was measured for all the relevant variables, with sufficiently narrow binning. However there are often too many discriminating variables to measure their distribution fully differentially, and, even if this was feasible, one would not be able to predict accurately the cross section in too many bins. In this situation, the sensitivity to the presence (or absence) of the EFT operators could be strongly reduced and it could be impossible to disentangle the effect of different operators and resolve flat directions in the parameter space of the EFT Wilson coefficients. One should thus switch to the direct comparison of the EFT with the data, by employing more sophisticated unbinned multivariate data analysis techniques.

Several multivariate methods have been developed and applied to the EFT or to similar problems, including Optimal Observables [4–9], the so-called “Method of Moments” [10–13] and similar approaches (see e.g. refs. [14, 15]) based on parametrizations of the scattering amplitude. The virtue of these techniques is that they are still based on high-level observables, making data/theory comparison simpler. The disadvantage is that they are intrinsically suboptimal and not systematically improvable towards optimality. Recent proposals to employ machine learning techniques for event categorization or for the design of non-linear summary statistics for regular binned analysis strategies with performances optimized for specific new physics targets (potentially, EFTs), should also be mentioned in this context [16–18].

A potentially optimal approach, which is closely analog to the one based on Machine Learning we employ in this paper, is the “Matrix Element Method” [19–25]. The main idea behind this construction is that optimal data analysis performances are unmistakably obtained by employing the likelihood  $\mathcal{L}(c|\mathcal{D})$ , i.e. the probability density of the observed dataset “ $\mathcal{D}$ ” seen as a function of the free parameters “ $c$ ” of the probability model. In our case, the free parameters coincide with the EFT Wilson coefficients.<sup>1</sup> The LHC data consist of independent repeated measurements of the variable “ $x$ ” that describes the kinematical configuration of each observed event. Therefore the likelihood factorizes and evaluating it requires the knowledge of the distribution of  $x$ . More precisely, since we are interested in  $\mathcal{L}(c|\mathcal{D})$  only up to an overall  $c$ -independent normalization, it is sufficient to know the ratio  $r(x, c)$  between the density as a function of  $c$  (and  $x$ ) and the density at some fixed value  $c = \bar{c}$ . The SM point,  $\bar{c} = 0$ , can be conventionally chosen.

It should be emphasized that extracting  $r(x, c)$  is a highly non-trivial task, as nicely explained in ref. [26] in terms of latent variables. The Monte Carlo generator code does of course implement an analytic point-by-point representation of the density (and, in turn, of  $r$ ), which is however expressed in terms of abstract variables “ $z$ ” and not of the variables  $x$  that are actually observed. The analytic representation of  $r$  in the  $z$  variables can be used as a surrogate of  $r(x, c)$  only if there is a faithful one-to-one correspondence between  $z$  and  $x$ . This is typically the case at leading order, if showering and detector effects are small, and if there are not undetected particles. However it is sufficient to have neutrinos in the final state, or to include Next to Leading Order (NLO) corrections to spoil the correspondence between  $z$  and  $x$ . Showering, hadronization and detector effects also wash out the correspondence. In the Matrix Element Method,  $r(x, c)$  is obtained by a phenomenological parameterization of these effects in terms of transfer functions that translate the knowledge of the density at the “ $z$ ” level into the one at the “ $x$ ” level. The free parameters of the phenomenological modeling of the transfer functions are fitted to Monte Carlo samples.

The Matrix Element Method is potentially optimal and improvable towards optimality. However it is not “systematically” improvable, in the sense that a more accurate reconstruction of  $r(x, c)$  requires a case-by-case optimization of the transfer function modeling. Addressing this issue by removing the sensitivity on the transfer function, or to remove them entirely, is an active research direction for the development of the Matrix Element Method, and some strategies already exist (see e.g. [27, 28]). However the issue does not even arise with the alternative approach employed in this paper which is based on the reconstruction of  $r(x, c)$  using Machine Learning techniques rather than phenomenological modeling. Systematic improvement is possible using bigger Neural Networks and larger training sets and computational cost is the only potential bottleneck. Furthermore refining the reconstruction by including additional effects requires substantial effort and increases the computational complexity of the Matrix Element Method, while the complexity of the Machine Learning-based reconstruction is a priori independent of the degree of refinement of

---

<sup>1</sup>The notion of “optimality” can be made fully rigorous and quantitative, both when the purpose of the analysis is to measure the free parameters of the EFT and when it is to test the EFT hypothesis ( $c \neq 0$ ) against the SM ( $c = 0$ ) one, and both from a Bayesian and from a frequentist viewpoint. The case of a frequentist hypothesis test is discussed in more details below.

the simulations. Therefore it is worth investigating these novel techniques as an alternative and/or as a complement to the Matrix Element Method.

There is already a considerable literature on the reconstruction of  $r(x, c)$  using Neural Networks [26, 29–34] and several algorithms exist. Here we adopt the most basic strategy, mathematically founded on the standard Statistical Learning problem of classification (see section 2.1 for a brief review), which we improve by introducing the notion of “Quadratic Classifier”. The relation between our methodology and the existing literature, the possibility of integrating it in other algorithms and to apply it to different problems is discussed in details in section 2.2 and in the Conclusions. However it is worth anticipating that, unlike simulator-assisted techniques [32], the Quadratic Classifier only exploits Monte Carlo data samples (in the extended sense outlined above) and no other information on the data generation process. It can thus be used as it is with any Monte Carlo generator.

Apart from describing the Quadratic Classifier, the main aim of the paper is to investigate the potential impact of Machine Learning methods on LHC EFT searches, from two viewpoints.

The first question we address is if and to what extent statistically optimal sensitivity to the presence or absence of the EFT operators can be achieved. In order to answer, a rigorous quantitative notion of optimality is defined by exploiting the Neyman-Pearson lemma [35], namely the fact that the “best” (maximum power at fixed size, in the standard terminology of e.g. ref. [36]) frequentist test between two simple hypotheses is the one that employs the likelihood ratio as test statistic. By regarding the EFT at each given value of the  $c$  Wilson coefficients as a simple hypothesis, to be compared with the SM  $c = 0$  hypothesis, we would thus obtain the strongest expected 95% Confidence Level (CL) exclusion bounds on  $c$  (when the SM is true) if the true distribution ratio  $r(x, c)$  was available and used for the test. This can be compared with the bound obtained by employing the approximate ratio  $\hat{r}(x, c)$  reconstructed by the Neural Network, allowing us to quantify the approximation performances of the method in objective and useful terms.<sup>2</sup> Of course, the exact  $r(x, c)$  is not available in a realistic EFT problem, therefore the comparison can only be performed on a toy problem. In order to make it as close as possible to reality, our “Toy” problem is defined in terms of an analytical approximation of the differential cross section of the process of interest (i.e. fully leptonic ZW, see below), implemented in a dedicated Monte Carlo generator.

The second aim of the paper is to quantify the potential gain in sensitivity of Machine Learning techniques, compared with the basic approach based on differential cross section measurements in bins. The associated production of a Z and a W boson decaying to leptons, at high transverse momentum ( $p_{T,Z} > 300$  GeV) and with the total integrated luminosity of the High Luminosity LHC (HL-LHC), is considered for illustration. This final state has been selected to be relatively simple, but still described by a large enough number of variables to justify the usage of unbinned analysis techniques. Moreover it has been

---

<sup>2</sup>It should be emphasized that we adopt this specific notion of “optimality” only because the frequentist hypothesis test between two simple hypotheses is relatively easy to implement in a fully rigorous manner. The reconstructed likelihood ratio could be employed for any other purpose and/or relying on asymptotic approximations using standard statistical techniques.

studied already quite extensively in the EFT literature (see e.g. refs. [37–44]) and a number of variables have been identified, including those associated with the vector bosons decay products [41, 45, 46], with the potential of improving the sensitivity to the EFT operators.

The comparison with the binned analysis is performed on the Toy version of the problem mentioned above, on the exact tree-level (LO) modeling of the process and on NLO QCD plus parton showering Monte Carlo data. By progressively refining our modeling of the problem in these three stages, this comparison also illustrates the flexibility of the approach and the fact that increasingly sophisticated descriptions of the data are not harder for the machine to learn. This should be contrasted with the Matrix Element method, which would instead need to be substantially redesigned at each step. As an illustration, we will show that employing the analytical approximated distribution ratio, that was optimal on the Toy problem, leads to considerably worse performances than the Neural Network already at LO. At NLO the performances further deteriorate and the reach is essentially identical to the one of the binned analysis.

The rest of the paper is organized as follows. In section 2 we introduce the Quadratic Classifier as a natural improvement of the standard Neural Network classifier for cases, like the one of the EFT, where the dependence of the distribution ratio on the “ $c$ ” parameters is known. The fully leptonic ZW process, the EFT operators we aim at probing and the relevant kinematical variables, are discussed in section 3. The Toy, the LO and the NLO Monte Carlo generators employed in the analysis are also described. The first set of results, aimed at assessing the optimality of the Quadratic Classifier on the Toy data, are reported in section 4. The results obtained with the LO Monte Carlo are also discussed, showing the stability of the Neural Network performances as opposite to the degradation of the sensitivity observed with the Matrix Element and with the binned analysis methods. NLO results are shown in section 5. We will see that the Quadratic Classifier methodology applies straightforwardly at NLO in spite of the fact that negative weights are present in the NLO Monte Carlo samples. The only complication associated with negative weights, which we discuss in section 5.1, is not related with the reconstruction of the  $\hat{r}(x, c)$  function by the Neural Network, but with the calculation of the distribution of the variable  $\hat{r}(x, c)$  itself, which is needed for the hypothesis test. All the technical details on the Neural Network design and training are summarized in section 6, and our conclusions are reported in section 7. Appendices A and B contain the generalization of the Quadratic Classifier for an arbitrary number of Wilson coefficients and the proof of its asymptotic optimality.

## 2 Teaching new physics to a machine

Consider two hypotheses,  $H_0$  and  $H_1$ , on the physical theory that describes the distribution of the variable  $x$ . In the concrete applications of the following sections,  $H_0$  will be identified with the SM EFT and  $H_1$  with the SM theory. The statistical variable  $x \in X$  describes the kinematical configuration in the search region of interest  $X$ . In the following,  $x$  will describe the momenta of 3 leptons and the missing transverse momentum, subject to selection cuts. Each of the two hypotheses (after choosing, if needed, their free parameters) characterizes the distribution of  $x$  completely. Namely they contain all the information that is needed

to compute, in line of principle, the differential cross sections  $d\sigma_0(x)$  and  $d\sigma_1(x)$ . The differential cross sections describe both the probability density function of  $x$ , which is obtained by normalization

$$\text{pdf}(x|H_{0,1}) = \frac{1}{\sigma_{0,1}(X)} \frac{d\sigma_{0,1}}{dx}, \quad (2.1)$$

and the total number of instances of  $x$  (i.e. of events) that is expected to be found in the dataset, denoted as  $N(X|H_{0,1})$ . This is equal to the cross section integrated on  $X$  and multiplied by the luminosity of the experiment, namely  $N(X|H_{0,1}) = L \cdot \sigma_{0,1}(X)$ .

The total number of observed events follows a Poisson distribution. Hence for a given observed dataset  $\mathcal{D} = \{x_i\}$ , with  $\mathcal{N}$  observed events, the  $H_1/H_0$  log likelihood ratio reads

$$\lambda(\mathcal{D}) \equiv \log \frac{\mathcal{L}(H_1|\mathcal{D})}{\mathcal{L}(H_0|\mathcal{D})} = N(X|H_0) - N(X|H_1) - \sum_{i=1}^{\mathcal{N}} \log \frac{d\sigma_0(x_i)}{d\sigma_1(x_i)}. \quad (2.2)$$

The statistic  $\lambda(\mathcal{D})$  is known as the “extended” log likelihood ratio [47], and it is the central object for hypothesis testing ( $H_0$  versus  $H_1$ ) or for measurements (if  $H_0$  contains free parameters), both from a Frequentist and from a Bayesian viewpoint. The “N” terms in eq. (2.2) can be computed as Monte Carlo integrals. What is missing in order to evaluate  $\lambda$  is thus the cross section ratio

$$r(x) \equiv \frac{d\sigma_0(x)}{d\sigma_1(x)}. \quad (2.3)$$

This should be known locally in the phase space as a function of  $x$ .

The physical knowledge of the  $H_0$  and  $H_1$  models gets translated into Monte Carlo generator codes, which allow us to estimate  $\sigma_{0,1}(X)$  and to produce samples,  $S_{0,1}$ , of artificial events following the  $\text{pdf}(x|H_{0,1})$  distributions. More precisely, the Monte Carlo generates weighted events  $e = (x_e, w_e)$ , with  $x_e$  one instance of  $x$  and  $w_e$  the associated weight. If the  $w_e$ ’s are not all equal,  $x_e$  does not follow the pdf of  $x$  and the expectation value of the observables  $O(x)$  has to be computed as a weighted average. We choose the normalization of the weights such that they sum up to  $\sigma_{0,1}(X)$  over the entire sample

$$\sum_{e \in S_{0,1}} w_e = \sigma_{0,1}(X). \quad (2.4)$$

With this convention, the weighted sum of  $O(x_e)$  approximates the integral of  $O(x) \cdot d\sigma_{0,1}(x)$  on  $x \in X$ . Namely

$$\sum_{e \in S_{0,1}} w_e O(x_e) \xrightarrow{\text{LS}} \int_{x \in X} d\sigma_{0,1}(x) O(x) = \sigma_{0,1}(X) \text{E}[O|H_{0,1}], \quad (2.5)$$

in the Large Sample (LS) limit where  $S_{0,1}$  are infinitely large. We will see below how to construct an estimator  $\hat{r}(x)$  for  $r(x)$  (or, in short, to fit  $r(x)$ ) using finite  $S_0$  and  $S_1$  samples.

For tree-level Monte Carlo generators the previous formulas could be made simpler by employing unweighted samples where all the weights are equal. However radiative corrections need to be included for sufficiently accurate predictions, at least up to NLO in

the QCD loop expansion. NLO generators can only produce weighted events, and some of the events have a negative weight. Therefore the NLO Monte Carlo samples cannot be rigorously interpreted as a sampling of the underlying distribution. However provided they consistently obey the LS limiting condition in eq. (2.5), they are equivalent to ordinary samples with positive weights for most applications, including the one described below.

## 2.1 The Standard Classifier

The estimator  $\hat{r}(x)$  can be obtained by solving the most basic Machine Learning problem, namely supervised classification with real-output Neural Networks (see ref. [48] for an in-depth mathematical discussion). One considers a Neural Network acting on the kinematical variables and returning  $f(x) \in (0, 1)$ . This is trained on the two Monte Carlo samples by minimizing the loss function

$$L[f(\cdot)] = \sum_{e \in S_0} w_e [f(x_e)]^2 + \sum_{e \in S_1} w_e [1 - f(x_e)]^2, \quad (2.6)$$

with respect to the free parameters (called weights and biases) of the Neural Network. The trained Neural Network,  $\hat{f}(x)$ , is in one-to-one correspondence with  $\hat{r}(x)$ , namely

$$\hat{f}(x) = \frac{1}{1 + \hat{r}(x)} \Leftrightarrow \hat{r}(x) = \frac{1}{\hat{f}(x)} - 1. \quad (2.7)$$

The reason why  $\hat{r}(x)$ , as defined above, approximates  $r(x)$  is easily understood as follows. If the Monte Carlo training data are sufficiently abundant, the loss function in eq. (2.6) approaches its Large Sample limit and becomes

$$L[f(\cdot)] \xrightarrow{\text{LS}} \int_{x \in X} d\sigma_0(x) [f(x)]^2 + \int_{x \in X} d\sigma_1(x) [1 - f(x)]^2. \quad (2.8)$$

Furthermore if the Neural Network is sufficiently complex (i.e. contains a large number of adjustable parameters) to be effectively equivalent to an arbitrary function of  $x$ , the minimum of the loss can be obtained by variational calculus. By setting to zero the functional derivative of  $L$  with respect to  $f$  one immediately finds

$$\hat{f}(x) \simeq \frac{d\sigma_1(x)}{d\sigma_1(x) + d\sigma_0(x)} = \frac{1}{1 + r(x)} \Rightarrow \hat{r}(x) \simeq r(x). \quad (2.9)$$

The same result holds for other loss functions such as the standard Cross-Entropy, which has been found in ref. [31] to have better performances for EFT applications, or the more exotic “Maximum Likelihood” loss [49], which is conceptually appealing because of its connection with the Maximum Likelihood principle. We observed no strikingly different performances with the various options, but we did not investigate this point in full detail. In what follows we stick to the quadratic loss in eq. (2.6).

The simple argument above already illustrates the two main competing aspects that control the performances of the method and its ability to produce a satisfactory approximation of  $r(x)$ . One is that the Neural Network should be complex in order to attain a configuration that is close enough to the (absolute) minimum,  $f(x) = 1/(1 + r(x))$ , of



the loss functional in eq. (2.8). In ordinary fitting, this is nothing but the request that the fit function should contain enough adjustable parameters to model the target function accurately. On the other hand if the Network is too complex, it can develop sharp features, while we are entitled to take the Large Sample limit in eq. (2.8) only if  $f$  is a smooth enough function of  $x$ . Namely we need  $f$  to vary appreciably only in regions of the  $X$  space that contain enough Monte Carlo points. Otherwise the minimization of eq. (2.6) brings  $f$  to approach zero at the individual points that belong to  $S_0$  sample, and to approach one at those of the  $S_1$  sample. This phenomenon, called overfitting, makes that for a given finite size of the training sample, optimal performances are obtained by balancing the intrinsic approximation error of the Neural Network against the complexity penalty due to overfitting. A third aspect, which is extremely important but more difficult to control theoretically, is the concrete ability of the training algorithm to actually reach the global minimum of the loss function in finite time. This requires a judicious choice of the minimization algorithm and of the Neural Network activation functions.

The problem of fitting  $r(x)$  is mathematically equivalent to a classification problem. A major practical difference however emerges when considering the level of accuracy that is required on  $\hat{r}(x)$  as an approximation of  $r(x)$ . Not much accuracy is needed for ordinary classification, because  $\hat{r}(x)$  (or, equivalently,  $\hat{f}(x)$ ) is used as a discriminant variable to distinguish instances of  $H_0$  from instances of  $H_1$  on an event-by-event basis. Namely, one does not employ  $\hat{r}(x)$  directly in the analysis of the data, but a thresholded version of  $\hat{r}(x)$  that isolates regions of the  $X$  space that are mostly  $H_0$ -like ( $r$  is large) or  $H_1$ -like ( $r$  is small). Some correlation between  $\hat{r}(x)$  and  $r(x)$ , such that  $\hat{r}(x)$  is large/small when  $r(x)$  is large/small, is thus sufficient for good classification performances. Furthermore the region where  $r(x) \simeq 1$  is irrelevant for classification.

The situation is radically different in our case because the EFT operators are small corrections to the SM. The regions where the EFT/SM distribution ratio is close to one cover most of the phase-space, but these regions can contribute significantly to the sensitivity if they are highly populated in the data sample. Mild departures of  $r(x)$  from unity should thus be captured by  $\hat{r}(x)$ , with good accuracy relative to the magnitude of these departures. Obviously the problem is increasingly severe when the free parameters of the EFT (i.e. the Wilson coefficients “ $c$ ”) approach the SM value  $c = 0$  and  $r(x)$  approaches one. On the other hand it is precisely when  $c$  is small, and the EFT is difficult to see, that a faithful reconstruction of  $r(x)$  would be needed in order to improve the sensitivity of the analysis.

## 2.2 The Quadratic Classifier

Barring special circumstances, the EFT prediction for the differential cross section is a quadratic polynomial in the Wilson coefficients.<sup>3</sup> If a single operator is considered, so that a single free parameter  $c$  is present and the SM corresponds to the value  $c = 0$ , the EFT

---

<sup>3</sup>The only exception is when the relevant EFT effects are modifications of the SM particles total decay widths. Also notice that the cross section is quadratic only at the leading order in the EFT perturbative expansion, which is however normally very well justified since the EFT effects are small. Higher orders could nevertheless be straightforwardly included as higher order polynomial terms.

differential cross section reads

$$d\sigma_0(x; c) = d\sigma_1(x) \{ [1 + c\alpha(x)]^2 + [c\beta(x)]^2 \}, \quad (2.10)$$

where  $\alpha(x)$  and  $\beta(x)$  are real functions of  $x$ . An estimator  $\hat{r}(x, c)$  for the distribution ratio in the entire Wilson coefficients parameters space could thus be obtained as

$$\hat{r}(x, c) = [1 + c\hat{\alpha}(x)]^2 + [c\hat{\beta}(x)]^2, \quad (2.11)$$

from estimators  $\hat{\alpha}(x)$  and  $\hat{\beta}(x)$  of the coefficient functions  $\alpha(x)$  and  $\beta(x)$ . Notice that eq. (2.10) parametrizes, for generic  $\alpha(x)$  and  $\beta(x)$ , the most general function of  $x$  and  $c$  which is quadratic in  $c$ , which is always positive (like a cross section must be) and which reduces to the SM cross section for  $c = 0$ . The equation admits a straightforward generalization for an arbitrary number of  $c$  parameters, which we work out in appendix A.

The estimators  $\hat{\alpha}(x)$  and  $\hat{\beta}(x)$  are obtained as follows. We first define a function  $f(x; c) \in (0, 1)$ , in terms of two neural networks  $n_\alpha$  and  $n_\beta$  with unbounded output (i.e.  $n_{\alpha,\beta} \in (-\infty, +\infty)$  up to weight-clipping regularization), with the following dependence on  $c$ :

$$f(x, c) \equiv \frac{1}{1 + [1 + c n_\alpha(x)]^2 + [c n_\beta(x)]^2}. \quad (2.12)$$

Next, we consider a set  $\mathcal{C} = \{c_i\}$  of values of  $c$  and we generate the corresponding EFT Monte Carlo samples  $S_0(c_i)$ . At least two distinct values of  $c_i \neq 0$  need to be employed, however using more than two values is beneficial for the performances. Monte Carlo samples are also generated for the  $H_1$  (i.e.  $c = 0$ ) hypothesis, one for each of the  $S_0(c_i)$  samples. These are denoted as  $S_1(c_i)$  in spite of the fact that they are all generated according to the same  $c = 0$  hypothesis. The samples are used to train the  $n_{\alpha,\beta}$  Networks, with the loss function

$$L[n_\alpha(\cdot), n_\beta(\cdot)] = \sum_{c_i \in \mathcal{C}} \left\{ \sum_{e \in S_0(c_i)} w_e [f(x_e, c_i)]^2 + \sum_{e \in S_1(c_i)} w_e [1 - f(x_e, c_i)]^2 \right\}. \quad (2.13)$$

We stress that in the second term in the curly brackets, the function  $f$  is evaluated on the same value of  $c = c_i$  that is employed for the generation of the  $S_0(c_i)$  Monte Carlo sample which we sum over in the first term.

By taking the Large Sample limit for the loss function as in eq. (2.8), differentiating it with respect to  $n_\alpha$  and  $n_\beta$  and using the quadratic condition (2.10), it is easy to show that the trained Networks  $\hat{n}_\alpha$  and  $\hat{n}_\beta$  approach  $\alpha$  and  $\beta$ , respectively. Namely

$$\hat{\alpha}(x) \equiv \hat{n}_\alpha(x) \simeq \alpha(x), \quad \hat{\beta}(x) \equiv \hat{n}_\beta(x) \simeq \beta(x). \quad (2.14)$$

More precisely, by taking the functional derivative one shows that the configuration  $n_\alpha = \alpha$  and  $n_\beta = \beta$  is a local minimum of the loss in the Large Sample limit. It is shown in appendix B that this is actually the unique global minimum of the loss.

It is simple to illustrate the potential advantages of the Quadratic Classifier, based on the analogy with the basic binned histogram approach to EFT searches. In that approach,

the  $X$  space is divided in bins and the likelihood ratio is approximated as a product of Poisson distributions for the countings observed in each bin. Rather than  $\hat{r}(x, c)$ , the theoretical input required to evaluate the likelihood are estimates  $\hat{\sigma}_0(\mathbf{b}; c)$  for the cross sections integrated in each bin “ $\mathbf{b}$ ”. Employing the Standard Classifier approach to determine  $\hat{r}(x, c)$  would correspond in this analogy to compute  $\hat{\sigma}_0(\mathbf{b}; c)$  for each fixed value of  $c$  using a dedicated Monte Carlo simulation. By scanning over  $c$  on a grid,  $\hat{\sigma}_0(\mathbf{b}; c)$  would be obtained by interpolation. Every EFT practitioner knows that this is a very demanding and often unfeasible way to proceed. Even leaving aside the computational burden associated with the scan over  $c$ , the problem is that the small values of  $c$  (say,  $c = \bar{c}$ ) we are interested in probing typically predict cross sections that are very close to the SM value and it is precisely the small relative difference between the EFT and the SM predictions what drives the sensitivity. A very small Monte Carlo error, which in turn requires very accurate and demanding simulations, would be needed in order to be sensitive to these small effects. In the Standard Classifier method, the counterpart of this issue is the need of generating very large samples for training the Neural Network. Furthermore, this should be done with several values of  $c$  for the interpolation. This approach is computationally demanding even when a single Wilson coefficient is considered, and it becomes rapidly unfeasible if  $c$  is a higher-dimensional vector of Wilson coefficients to be scanned over.

The strategy that is normally adopted in standard binned analyses is closely analog to a Quadratic Classifier. One enforces the quadratic dependence of  $\sigma_0(\mathbf{b}; c)$  on  $c$  as in eq. (2.10), and estimates the three polynomial coefficients (i.e. the SM cross section and the analog of  $\alpha$  and  $\beta$ ) in each bin by a  $\chi^2$  fit to  $\hat{\sigma}_0(\mathbf{b}; c)$ , as estimated from the Monte Carlo simulations for several values of  $c$ . The values of  $c$  used for the fit are much larger than the reach of the experiment  $c = \bar{c}$ , so that their effects are not too small and can be captured by the Monte Carlo simulation. The Quadratic Classifier works in the exact same way. It can learn  $\hat{\alpha}(x)$  and  $\hat{\beta}(x)$  using training samples generated with large values of  $c$ , for which the difference between the  $S_0(c)$  and  $S_1(c)$  is sizable. The training can thus recognize this difference, producing accurate estimates of  $\hat{\alpha}(x)$  and  $\hat{\beta}(x)$ . This accurate knowledge results in an accurate estimate of  $\hat{r}(x, c)$  and of its departures from unity even at the small value  $c = \bar{c}$ , because our method exploits the exact quadratic relation in eq. (2.10).

It should be noted that the “Quadratic Classifier” introduced in eq. (2.12) is “Parametrized” in the sense that it encapsulates the dependence on the  $c$  parameters, but it is the exact opposite of the Parametrized Neural Network (or Parametrized Classifier) of ref. [30]. In that case, the Wilson Coefficient  $c$  is given as an input to the Neural Network, which acts on an enlarged  $(x, c)$  features space. The purpose is to let the Neural Network learn also the dependence on  $c$  of the distribution ratio in cases where this is unknown. Here instead we want to enforce the quadratic dependence of the distribution ratio on  $c$ , in order to simplify the learning task.

An alternative strategy to exploit the analytic dependence on  $c$  is the one based on “morphing” [33]. Morphing consists in selecting one point in the parameter space for each of the coefficient functions that parametrize  $d\sigma_0(x; c)$  as a function of  $c$ , and expressing  $d\sigma_0(x; c)$  as a linear combination of the cross-sections computed at these points. For instance, a total of 3 “morphing basis points”,  $c_{1,2,3}$ , are needed for a single Wilson coefficient and

quadratic dependence, and  $d\sigma_0(x; c)$  is expressed as a linear combination of  $d\sigma_0(x; c_{1,2,3})$ . This rewriting can be used to produce two distinct learning algorithms.

The first option is to learn the density ratios  $d\sigma_0(x; c_{1,2,3})/d\sigma_1(x)$  individually (one-by-one or simultaneously), by using training data generated at the morphing basis points  $c_{1,2,3}$ , and to obtain  $\hat{r}(x, c)$  using the morphing formula. In the analogy with ordinary binning, this would correspond to extracting the dependence on  $c$  of the cross-sections by a quadratic interpolation of  $\hat{\sigma}_0(b; c_{1,2,3})$  at the selected points. Of course it is possible to reconstruct the cross sections accurately also by using 3 very accurate simulations, rather than fitting less accurate simulations at several points. However a judicious choice of the values of  $c_{1,2,3}$  is essential for a proper reconstruction of the quadratic and of the linear term of the polynomial. For the former, it is sufficient to take  $c$  very large, but for the latter a value of  $c$  should be selected that is neither too large, such that the quadratic term dominates by too many orders of magnitude, nor too small such that the constant SM term dominates. Notice that the optimal  $c$  depends on the analysis bin because the EFT effects relative to the SM (and the relative importance of the quadratic and linear terms) can be vastly different in different regions of the phase space. With “plain” morphing as described above, we are obliged to employ only few values of  $c$ , which might not be enough to cover the entire phase space accurately. With the Quadratic Classifier instead, all values of  $c$  that are useful to learn the distribution in some region of the phase space (see e.g. eq. (4.4)) can be included simultaneously in the training set.

Alternatively, one can use the morphing formula in place of eq. (2.10), producing a different parameterization of the classifier than the one in eq. (2.12), to be trained with values of the parameters that are unrelated with the morphing basis points. The parametrization employed in the Quadratic Classifier is arguably more convenient, as it is simpler, universal and bounded to  $f \in (0, 1)$  interval owing to the positivity of eq. (2.10). Importantly, also the condition  $\hat{r}(x, 0) \equiv 1$  is built-in in the Quadratic Classifier. However this could be enforced in the morphing parameterization as well by selecting  $c = 0$  as one of the basis points. If this is done, we do not expect<sup>4</sup> a degradation of the performances if employing the morphing-based parametrization rather than ours. Indeed, we believe that the key of the success of the Quadratic Classifier that we observe in this paper stems from the appropriate choice of the values of  $c$  used for training, and not from the specific parametrization we employ. The non-optimal performances of the morphing strategy observed in ref. [33] (on a different process than the one we study) are probably to be attributed to a non-optimal choice. Further investigations on this aspect are beyond the scope of the present paper.

### 3 Fully leptonic ZW

Consider ZW production at the LHC with leptonic decays, namely  $Z \rightarrow \ell^+ \ell^-$  and  $W \rightarrow \ell \nu$ , where  $\ell = e, \mu$ . As explained in the Introduction, this is arguably the simplest process, of established EFT relevance, where a multivariate approach is justified and potentially improves the sensitivity. We focus on the high-energy tail of the process, with a selection

---

<sup>4</sup>Provided that the possibility of having  $f$  outside the  $(0, 1)$  interval is not a problem when training, for instance, with the cross-entropy loss function.

cut on the transverse momentum of the Z-boson,  $p_{T,Z} > 300$  GeV, because of two independent reasons. First, because at high energy we can approximate the differential cross section analytically and define a realistic enough Toy problem to assess the optimality of the method. Second, because at high-energy the statistics is sufficiently limited (less than  $5 \times 10^3$  expected events at the HL-LHC, including both W charges) to expect systematic uncertainties not to play a dominant role. The reach we will estimate in the  $p_{T,Z} > 300$  GeV region, on purely statistical bases, should thus be nearly realistic.

The high-energy regime, in spite of the relatively limited statistics, is the most relevant one to probe those EFT operators that induce energy-growing corrections to the SM amplitudes. There are only two CP-preserving and flavor-universal operators in the ZW channel that induce quadratically energy-growing terms, namely<sup>5</sup>

$$\mathcal{O}_{\varphi q}^{(3)} = G_{\varphi q}^{(3)} (\bar{Q}_L \sigma^a \gamma^\mu Q_L) (iH^\dagger \overleftrightarrow{D}_\mu H), \quad \mathcal{O}_W = G_W \varepsilon_{abc} W_\mu^{a\nu} W_\nu^{b\rho} W_\rho^{c\mu}. \quad (3.1)$$

We thus focus on these operators in our analysis. The BSM relevance of these operators is well established in the literature. Searching for  $\mathcal{O}_{\varphi q}^{(3)}$  in di-boson final states is among the most promising probes of Higgs compositeness at the HL-LHC [40]. The operator  $\mathcal{O}_W$  is generated with loop-suppressed coefficient in perturbative BSM models and in composite Higgs models of SILH [2] type. However it does not experience loop suppression for strongly-coupled BSM and is even potentially enhanced by the strong coupling in particular scenarios [50].

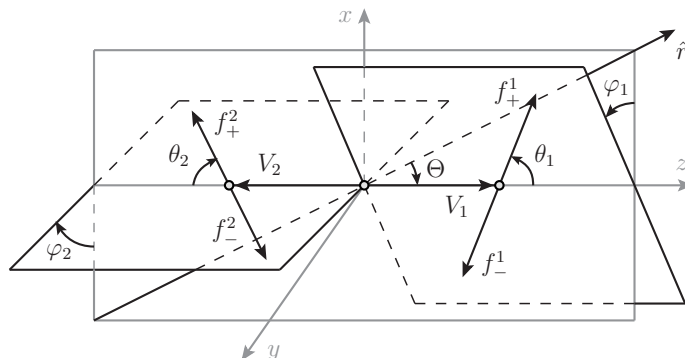
Both  $\mathcal{O}_{\varphi q}^{(3)}$  and  $\mathcal{O}_W$  contribute to the ZW production amplitudes with quadratically energy-growing terms of order  $G \cdot s$ , where  $s$  is the center-of-mass energy squared of the diboson system. However the way in which this energy growth manifests itself in the cross section is rather different for the two operators (see e.g. refs. [40, 41]). The  $\mathcal{O}_{\varphi q}^{(3)}$  operator mainly contributes to the “00” helicity amplitude, in which the gauge bosons are longitudinally polarized. The SM amplitude in this channel is sizable and has a constant behavior with  $s$  at high energy. As a consequence, a sizable quadratically-growing interference term between the SM and the BSM amplitudes is present in the cross section. This happens even at the “inclusive” level, i.e. when only the hard scattering variables describing ZW production (and not the decay ones) are measured.

On the contrary, the  $\mathcal{O}_W$  operator induces quadratically-growing contributions only in the transverse polarization channels with equal helicity for the two gauge bosons (namely, ++ and --). In the SM this channel is very suppressed at high energy, since its amplitude decreases as  $m_W^2/s$ . Therefore in inclusive observables the interference between  $\mathcal{O}_W$  and the SM does not grow with the energy and is very small. In order to access (or “resurrect” [41]) the interference, which is the dominant new physics contribution since the Wilson coefficient of the operator is small, the vector bosons decay variables must be measured. We thus expect that the sensitivity to  $\mathcal{O}_W$  will benefit more from an unbinned multivariate analysis technique than the one on  $\mathcal{O}_{\varphi q}^{(3)}$ .

The relevant kinematical variables that characterize the four-leptons final state are defined as in ref. [41] and depicted in figure 1, where  $V_1$  is identified with the Z and  $V_2$

---

<sup>5</sup>We use the definition  $H^\dagger \overleftrightarrow{D}_\mu H = H^\dagger D_\mu H - (D_\mu H)^\dagger H$ .



**Figure 1.** The kinematical variables in the “special” coordinate frame [41].

with the W boson. The figure displays the kinematics in the rest frame of the ZW system, obtained from the lab frame by a boost along the direction of motion (denoted as  $\hat{r}$  in the figure) of the ZW pair, followed by a suitable rotation that places the Z along the positive  $z$  axis and  $\hat{r}$  on the  $x$ - $z$  plane with positive  $x$  component. The “inclusive” variables associated with ZW production are the center-of-mass energy squared  $s$  and  $\Theta \in [0, \pi]$ , which is defined as the angle between  $\hat{r}$  and the Z-boson momentum. The decay kinematics is described by the polar and azimuthal decay angles  $\theta_{1,2}$  and  $\varphi_{1,2}$ . The latter angles are in the rest frame of each boson and they are defined as those of the final fermion or anti-fermion with helicity  $+1/2$  (e.g. the  $\ell^+$  in the case of a  $W^+$  and the  $\bar{\nu}$  for a  $W^-$ ), denoted as  $f_{\pm}^{1,2}$  in the figure.<sup>6</sup> The remaining variables that are needed to characterize the four leptons completely are weakly sensitive to the presence of the EFT operators and can be ignored, with the exception of the total transverse momentum of the ZW system,  $p_{T,ZW}$ , which is a useful discriminant at NLO [40].

The variables described above are useful for the theoretical calculation of the cross section, but they cannot be used for our analysis because they are not experimentally accessible. The “measured” variables we employ are defined as follows. First, since we do not measure the neutrino (longitudinal) momentum, this needs to be reconstructed by imposing the on-shell condition for the W. The reconstructed neutrino momentum, rather than the true one, is used to define the kinematical variables and in particular  $s$  and  $\Theta$ . Moreover, since we do not measure the helicity of the fermions but only their charge, the decay angles of the Z boson, denoted as  $\theta_Z$  and  $\varphi_Z$ , are defined in terms of the charge-plus lepton rather than of the helicity plus lepton. Depending on the (unobserved) leptons helicities these angles are either equal to  $\theta_1$  and  $\varphi_1$ , or to  $\pi - \theta_1$  and  $\varphi_1 + \pi$ , respectively. The W decay angles, defined in terms of the lepton or the reconstructed neutrino depending on the charge of the W as previously explained, are denoted as  $\theta_W$  and  $\varphi_W$ . In summary, the variables we employ in the analysis are

$$\{s, \Theta, \theta_W, \varphi_W, \theta_Z, \varphi_Z, p_{T,ZW}\}, \tag{3.2}$$

where of course  $p_{T,ZW}$  is non-vanishing only at NLO.

<sup>6</sup>The correct definition of  $\varphi_2$  appears in version four of ref. [41].

The on-shell condition for the W boson has no real solution if the W-boson transverse mass is larger than the W pole mass  $m_W$ . The neutrino is reconstructed in this case by assuming that the neutrino rapidity is equal to the one of the lepton. If instead the transverse mass is smaller than  $m_W$ , the condition has two distinct real solutions, each of which produces a different reconstructed kinematics. For our analysis we picked one of the two solutions at random on an event-by-event basis, while for the analysis of the actual data it would be arguably more convenient to duplicate the kinematical variables vector by including both solutions. Nothing changes in the discussion that follows if this second option is adopted.

### 3.1 Analytic approximation

At the tree-level order, and based on the narrow-width approximation for the decays, it is easy to approximate the cross section analytically in the high-energy regime. The crucial simplification is that the reconstructed 3-momentum of the W boson (with any of the two solutions for the neutrino) becomes exact when the W is boosted, so that the reconstructed  $\Theta$  and  $s$  variables approach the “true” ones of figure 1. Notice that  $\Theta$  is the angle between the Z and the direction of motion of the ZW system in the lab frame, which corresponds at tree-level to the direction of motion of the most energetic incoming parton. In the kinematical region we are interested in, the (valence) quark is more energetic than the anti-quark in more than 80% of the events. Therefore we can identify  $\Theta$  as the angle between the Z and the  $u$  quark or the  $d$  quark in the  $u\bar{d} \rightarrow ZW^+$  and  $d\bar{u} \rightarrow ZW^-$  processes, respectively.

With these identifications, the non-vanishing on-shell helicity amplitudes  $\mathcal{M}_{h_z h_w}$  for the hard scattering process  $u\bar{d} \rightarrow ZW^+$ , at leading order in the high-energy expansion, read

$$\begin{aligned} \mathcal{M}_{00} &= -\frac{g^2 \sin \Theta}{2\sqrt{2}} - \sqrt{2}G_{\varphi q}^{(3)} s \sin \Theta, & \mathcal{M}_{++} = \mathcal{M}_{--} &= \frac{3gc_w G_W s \sin \Theta}{\sqrt{2}}, \\ \mathcal{M}_{-+} &= -\frac{g^2(s_w^2 - 3c_w^2 \cos \Theta)}{3\sqrt{2}c_w} \cot \frac{\Theta}{2}, & \mathcal{M}_{+-} &= \frac{g^2(s_w^2 - 3c_w^2 \cos \Theta)}{3\sqrt{2}c_w} \tan \frac{\Theta}{2}, \end{aligned} \quad (3.3)$$

where  $g$  is the  $SU(2)_L$  coupling,  $c_w$  and  $s_w$  are the cosine and the sine of the Weak angle. An overall factor equal to the cosine of the Cabibbo angle has not been reported for shortness. The amplitudes for the  $d\bar{u} \rightarrow ZW^-$  process can be obtained from the ones above with the formal substitutions  $\Theta \rightarrow -\Theta$  and  $s_w^2 \rightarrow -s_w^2$ . The amplitudes are non-vanishing only for left-chirality initial quarks. Notice that the above formulas depend on the conventions in the definition of the wave-function of the external particles, and that these conventions must match the ones employed in the decay amplitude for the consistency of the final results. The wave-function reported in ref. [51] are employed.

Let us now turn to the vector bosons decays. The decay amplitudes assume a very simple form in terms of the  $\theta = \theta_{1,2}$  and  $\varphi = \varphi_{1,2}$  variables, namely

$$\mathcal{A}_h = -\sqrt{2}g_V m_V e^{ih\varphi} d_h(\theta), \quad (3.4)$$

where  $h$  is the helicity of the decaying vector boson ( $V = V_{1,2} = Z, W$ ) and  $d_h(\theta)$  are the Wigner  $d$ -functions. The overall coupling factor  $g_V$  depends on the nature of the boson and, in the case of the  $Z$ , on the electric charge of the helicity-plus fermion it decays to. Specifically,  $g_W = g/\sqrt{2}$  for the  $W$ ,  $g_Z = g_L = -g(1 - 2s_w^2)/2c_w$  if the  $Z$  decays to an helicity-plus  $\ell^+$  and  $g_Z = g_R = g s_w^2/c_w$  if the  $Z$  decays to an helicity-plus  $\ell^-$ . The two options for the helicity (which are physically distinct) correspond to two terms in the cross section. In the first one the  $Z$  decay amplitude is evaluated with the  $g_L$  coupling, with  $\theta = \theta_1 = \theta_Z$  and  $\varphi = \varphi_1 = \varphi_Z$ . In the second one we employ  $g_R$ ,  $\theta = \theta_1 = \pi - \theta_Z$  and  $\varphi = \varphi_1 = \varphi_Z + \pi$ . There is no helicity ambiguity in the  $W$ -boson decay angles. However the reconstruction of the azimuthal decay angle is exact in the high-energy limit only up to a twofold ambiguity [41]. Namely the reconstructed  $\varphi_W$  approaches  $\varphi_1$  on one of the two solutions for the neutrino (and we do not know which one), and  $\pi - \varphi_1$  on the other. Since we are selecting one solution at random, we should average the cross section over the two possibilities  $\varphi = \varphi_2 = \varphi_W$  and  $\varphi = \varphi_2 = \pi - \varphi_W$  for the  $W$  azimuthal angle. The polar angle is instead  $\theta = \theta_2 = \theta_W$  in both cases.

Production and decay are conveniently combined using the density matrix notation. We define the hard density matrix

$$d\rho_{h_z h_w h'_z h'_w}^{\text{hard}} = \frac{1}{24s} \mathcal{M}_{h_z h_w} (\mathcal{M}_{h'_z h'_w})^* d\Phi_{ZW}, \quad (3.5)$$

where  $d\Phi_{ZW}$  is the two-body phase space and the factor  $1/24s$  takes care of the flux and of the averages over the colors and the helicities of the initial quarks. The decay processes are instead encoded into decay density matrices. The one for the  $Z$ -boson, including the sum over the  $\ell^\pm$  helicities as previously explained, reads

$$d\rho_{h_z h'_z}^Z = \frac{1}{2m_Z \Gamma_Z} \left[ \mathcal{A}_{h_z} \mathcal{A}_{h'_z}^* \Big|_{g_L, \theta_Z, \varphi_Z} + \mathcal{A}_{h_z} \mathcal{A}_{h'_z}^* \Big|_{g_R, \pi - \theta_Z, \varphi_Z + \pi} \right] d\Phi_{\ell^+ \ell^-}, \quad (3.6)$$

where  $\Gamma_Z$  is the  $Z$  decay width. For the  $W$ , since we average on the neutrino reconstruction ambiguity, we have

$$d\rho_{h_w h'_w}^W = \frac{1}{2m_W \Gamma_W} \frac{1}{2} \left[ \mathcal{A}_{h_w} \mathcal{A}_{h'_w}^* \Big|_{\frac{g}{\sqrt{2}}, \theta_W, \varphi_W} + \mathcal{A}_{h_w} \mathcal{A}_{h'_w}^* \Big|_{\frac{g}{\sqrt{2}}, \theta_W, \pi - \varphi_W} \right] d\Phi_{\ell\nu}. \quad (3.7)$$

The complete partonic differential cross section is finally simply given by

$$d\hat{\sigma} = 4 \sum d\rho_{h_z h_w h'_z h'_w}^{\text{hard}} d\rho_{h_z h'_z}^Z d\rho_{h_w h'_w}^W, \quad (3.8)$$

where the sum is performed on the four helicity indices and the factor of 4 takes into account the decay channels into electrons and muons.

### 3.2 Monte Carlo generators

For our analysis we use three Monte Carlo generators, of increasing accuracy.

The first one is the Toy generator that implements the analytic approximation of the cross section in eq. (3.8), with the hard amplitudes expanded up to order  $G \cdot s$  in the EFT



contribution and up to order  $s^0$  in the SM term, as in eq. (3.3). This implies, in particular, that in the Toy Monte Carlo all the mixed transverse/longitudinal helicity channels vanish exactly, that only the  $\pm\mp$  and  $00$  channels are retained in the SM and that new physics is just in the  $00$  and  $\pm\pm$  channels for  $\mathcal{O}_{\varphi q}^{(3)}$  and  $\mathcal{O}_W$ , respectively. The Toy Monte Carlo employs a simple fit to the ( $u\bar{d}$  or  $d\bar{u}$ ) parton luminosities obtained from the nCTEQ15 [52] PDF set (implemented through the `ManeParse` [53] Mathematica package). The variable  $s$  is sampled according to the parton luminosity, while all the other variables are sampled uniformly. The cut  $p_{T,Z} = \sqrt{s}/2 \sin \Theta > 300$  GeV is implemented at generation level. Since the analytical distribution is extremely fast to evaluate, this basic approach is sufficient to obtain accurate Monte Carlo integrals and large unweighted event samples in a very short time.

The second generator is MADGRAPH [54] at LO, with the EFT operators implemented in the UFO model of ref. [55]. We simulate the  $2 \rightarrow 4$  process  $pp \rightarrow \mu^+ \mu^- e \nu_e$ , with the Z and the W decaying to opposite flavor leptons for a simpler reconstruction, and multiply the resulting cross section by 4. The cut on  $p_{T,Z}$ , defined as the sum of the  $\mu^+$  and  $\mu^-$  momenta, is imposed at generation level, as well as the cuts

$$m_{T,ev} \leq 90 \text{ GeV}, \quad 70 \text{ GeV} \leq m_{\mu\mu} \leq 110 \text{ GeV}, \quad (3.9)$$

on the transverse mass of the virtual W and the invariant mass of the virtual Z. These are needed to suppress non-resonant contributions to the production of the 4 leptons. Standard acceptance cuts on the charged leptons are also applied. The unweighted events obtained with MADGRAPH are further processed to compute the kinematical variables in eq. (3.2) after neutrino reconstruction, as detailed at the beginning of this section.

The MADGRAPH LO generator is slightly more accurate than the Toy one. It contains all the ZW helicity amplitudes and no high-energy approximation. Furthermore, it describes non-resonant topologies and off-shell vector bosons production, which affects the reconstruction of the neutrino and in turn the reconstruction of the Z and W decay variables [41]. Nevertheless on single-variable distributions the Toy Monte Carlo and the LO one agree reasonably well, at around 10%.

The third and most refined generator is MADGRAPH at NLO in QCD, interfaced with PYTHIA 8.244 [56, 57] for QCD parton showering. The complete  $2 \rightarrow 4$  process is generated like at LO, but no cuts could be applied at generation level apart from default acceptance cuts on the leptons and the lower cut on  $m_{\mu\mu}$  in eq. (3.9). At NLO, the cut on  $p_{T,Z}$  needs to be replaced with the cut  $p_{T,V} > 300$  GeV, with  $p_{T,V} = \min[p_{T,Z}, p_{T,W}]$ . This cut suppresses soft or collinear vector boson emission processes, which are insensitive to the EFT. In order to populate the  $p_{T,V} > 300$  GeV tail of the distribution with sufficient statistics, events were generated with a bias. The bias function was equal to one for  $p_{T,V}$  above 290 GeV, and equal to  $(p_{T,V}/290 \text{ GeV})^4$  below. The momenta of the charged leptons and the transverse momentum of the neutrino in the generated events were read with MADANALYSIS [58] and the kinematical variables in eq. (3.2) reconstructed like at LO. The cut  $p_{T,V} > 300$  GeV and the remaining cuts in eq. (3.9) were imposed on the reconstructed events. The total cut efficiency on the Monte Carlo data, thanks to the bias, was large enough (around 17%) to allow for an accurate prediction of the cross section and for the generation of large enough event samples.

Even if ours is an electroweak process, it is known that NLO QCD corrections can in principle affect significantly the sensitivity to the EFT operators. Relevant effects are related with the tree-level zero [59] in the transverse amplitude, which is lifted at NLO, and with the appearance of same-helicity transverse high-energy amplitudes due to real NLO radiation [60]. All these effects are properly modeled by the MADGRAPH NLO generator.

#### 4 Optimality on Toy data

Our goal is to reconstruct the EFT-over-SM cross section ratio  $r(x, c)$  as accurately as possible using the methods introduced in section 2. Since  $r$  is known analytically for the Toy problem, a simple qualitative assessment of the performances could be obtained by a point-by-point comparison (see figures 7 and 8) of  $r(x, c)$  with its approximation  $\hat{r}(x, c)$  provided by the trained Neural Network. However a point-by-point comparison is not quantitatively relevant, since the level of accuracy that is needed for  $\hat{r}(x, c)$  can be vastly different in different regions of the phase-space, depending on the volume of expected data and on the discriminating power of each region (i.e. on how much  $r$  is different from one).

The final aim of the entire construction is to obtain an accurate modeling of the extended log-likelihood ratio in eq. (2.2), to be eventually employed in the actual statistical analysis. A quantitative measure of the  $r$  reconstruction performances is thus best defined in terms of the performances of the final analysis that employs  $\hat{r}$ , instead of  $r$ , in the likelihood ratio. Among all possible statistical analyses that could be carried out, frequentist tests to the EFT hypothesis  $H_0(c)$  (regarded as a simple hypothesis for each given value of  $c$ ), against the SM one,  $H_1$ , are considered for the illustration of the performances.

Four alternative test statistic variables are employed. One is the standard Poisson binned likelihood ratio (see below). The others are unbinned and take the form

$$t_c(\mathcal{D}) = N(\mathbf{X}|H_0) - N(\mathbf{X}|H_1) - \sum_{i=1}^{\mathcal{N}} \tau_c(x_i), \tag{4.1}$$

where  $\tau_c(x)$  is either equal to the exact  $\log[r(x, c)]$  or to  $\log[\hat{r}(x, c)]$ , as reconstructed either with the Standard Classifier or with the Quadratic Classifier described in section 2.1 and 2.2, respectively. In each case the probability distributions of  $t$  in the two hypotheses are computed with toy experiments (or with the simpler strategy of section 5.1), and used to estimate the expected (median) exclusion reach on  $c$  at 95% Confidence Level if the SM hypothesis is true. In formulas, the 95% reaches ( $c_{2\sigma}$ ) we quote in what follows are solutions to the implicit equation

$$p(t_{\text{med}}(c_{2\sigma}); c_{2\sigma}) = 0.05, \text{ with } t_{\text{med}}(c) = \text{Median}[t_c(\mathcal{D})|H_1], \tag{4.2}$$

where the  $p$ -value is defined as

$$p(t_c; c) = \int_{t_c}^{\infty} dt'_c \text{pdf}(t'_c|H_0(c)). \tag{4.3}$$

The two Wilson coefficients  $c = G_{\varphi q}^{(3)}$  and  $c = G_W$  are considered separately. Therefore the results that follow are single-operator expected exclusion reaches.

Summarizing, the four methodologies we employ are

i) *Matrix Element (ME)*

In this case we set  $\tau_c(x) = \log[r(x, c)]$  in eq. (4.1), with  $r$  computed analytically using eq. (3.8). Therefore  $t$  coincides with the log-likelihood ratio  $\lambda$  in eq. (2.2), which in turn is the optimal discriminant between  $H_0$  and  $H_1$  due to the Neyman-Pearson lemma [35]. Namely, a straightforward application of the lemma guarantees that by employing  $t = \lambda$  as test statistic we will obtain the optimal (smallest)  $c_{2\sigma}$  reach, better than the one we could have obtained using any other variable. The Matrix Element Method is thus optimal in this case, and the optimality of the other methods can be assessed by comparing their  $c_{2\sigma}$  reach with the one of the Matrix Element.

ii) *Standard Classifier (SC)*

The second method consists in setting  $\tau_c(x) = \log[\hat{r}(x, c)]$  in eq. (4.1), with  $\hat{r}$  reconstructed by the Standard Classifier as in section 2.1. Notice that a separate training is needed to reconstruct  $\hat{r}(x, c)$  for each value of the Wilson Coefficient. Therefore computing  $c_{2\sigma}$ , as defined in eq. (4.2), requires scanning over  $c$ , performing first the Neural Network training and next the calculation of the distributions of  $t$  by toy experiments. For the Quadratic Classifier (and for the Matrix Element Method), the first step is not needed. The details on the Neural Network architecture and training, and of its optimization, will be discussed in section 6.

iii) *Quadratic Classifier (QC)*

The third approach is to employ  $\hat{r}(x, c)$  as reconstructed by the Quadratic Classifier of section 2.2. Implementation details are again postponed to section 6, however it is worth anticipating that the key for a successful reconstruction is to train using values for the Wilson coefficients that are significantly larger than the actual reach. Specifically, we used

$$\begin{aligned}
 G_{\varphi q}^{(3)} &: \{ \pm 50, \pm 20, \pm 5 \} \times 10^{-2} \text{ TeV}^{-2}, \\
 G_W &: \{ \pm 20, \pm 10, \pm 5 \} \times 10^{-2} \text{ TeV}^{-2}.
 \end{aligned}
 \tag{4.4}$$

These values have been selected as those that induce order one departures from the SM cross section in the low, medium and high regions of the  $p_{T,Z}$  distribution. If willing to compute cross-section in each  $p_{T,Z}$  region by quadratic interpolation, using the values selected with this criterion can be shown to maximize the accuracy on the reconstruction of the linear term, while still allowing for a good determination of the quadratic term. We expect this choice to be optimal for the Quadratic Classifier training as well. Also notice that the total number of training Monte Carlo events is the same one (6M, see section 6) employed for each of the separate trainings performed on the Standard Classifier.

iv) *Binned Analysis (BA)*

Finally, in order to quantify the potential gain of the unbinned strategy, we also perform a binned analysis. The test statistic in this case is provided by the sum over the bins of the log-ratio of the SM over EFT Poisson likelihoods, with the expected countings as a function of the Wilson coefficients computed from Monte Carlo simulations. The test statistic distributions, and in turn the reach by eq. (4.2), are computed with toy experiments like for the other methods and no asymptotic formulas are employed. For both  $G_{\varphi q}^{(3)}$  and  $G_W$  we considered 3 bins in  $p_{T,Z}$ , with the following boundaries

$$p_{T,Z}[\text{GeV}] : \quad \{300, 500, 750, 1200\} \text{ GeV} . \quad (4.5)$$

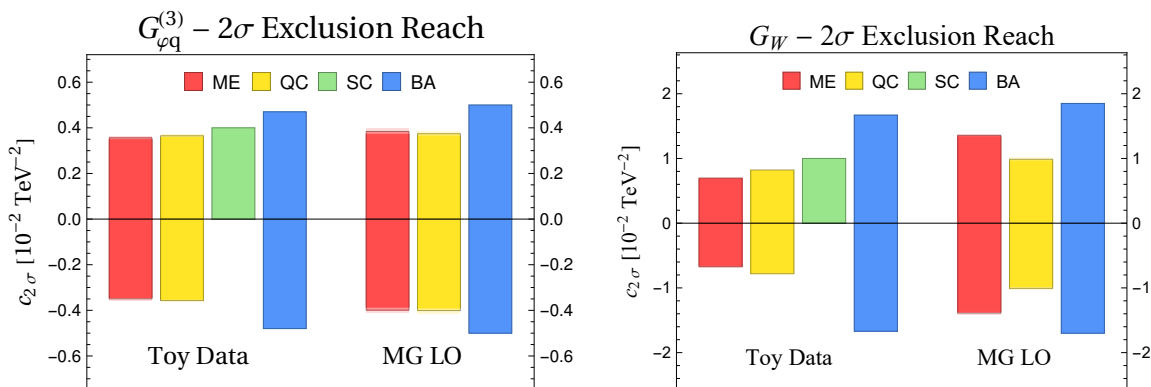
The  $p_{T,Z}$  variable is an extremely important discriminant because it is sensitive to the energy growth induced by the EFT. The three bins are selected based on the studies in refs. [40, 41], and a narrower binning has been checked not to improve the sensitivity significantly. A cut  $\cos \Theta < 0.5$  is imposed in the analysis targeting  $G_{\varphi q}^{(3)}$ , because this helps [40] in isolating the longitudinal helicity channel thanks to the amplitude zero in the transverse SM amplitudes. For  $\mathcal{O}_W$ , no  $\cos \Theta$  cut is performed, and each  $p_{T,Z}$  is split in two bins, for  $\cos 2\varphi_W$  larger and smaller than zero. This is sufficient to partially capture the leading EFT/SM interference term as discussed in ref. [41].

Most likely the binned analysis could be improved by considering more (and/or better) variables and a narrower binning. However it should be noticed that the simple strategies described above already result from an optimization, targeted to the specific operators at hand, and that the reach we obtain is consistent with the sensitivity projections available in the literature.

#### 4.1 Results

The results of the four methods are shown in figure 2 (see also table 1), together with the ones obtained with the MADGRAPH LO description of the ZW process, to be discussed in section 4.2. The  $2\sigma$  sensitivities reported in the figure are obtained by interpolating the median  $p$ -value as a function of the Wilson Coefficient  $c$ , in the vicinity of the reach, and computing  $c_{2\sigma}$  by solving eq. (4.2). Further details on this procedure, and the associated error, are given in section 5.1.

The figure reveals a number of interesting aspects. First, by comparing the Matrix Element reach with the one of the Binned Analysis we can quantify the potential gain in sensitivity offered by a multivariate strategy. The improvement is moderate (around 30%) for  $G_{\varphi q}^{(3)}$ , but it is more than a factor of 2 (of 2.4) in the case of the  $G_W$  operator coefficient. The different behavior of the two operators was expected on physical grounds, as discussed in details below. The figure also shows that the Quadratic Classifier is nearly optimal. More precisely, the reach is identical to the one of the Matrix Element for  $G_{\varphi q}^{(3)}$ , and  $< 20\%$  worse for  $G_W$ . We will see in section 6 that the residual gap for  $G_W$  can be eliminated with more training points than the ones used to produce figure 2. Suboptimal performances are shown in the figure in order to outline more clearly, in section 6, that our method is systematically improvable as long as larger and larger Monte Carlo samples are available.



**Figure 2.** Expected exclusion reach on  $G_{\varphi q}^{(3)}$  (left) and on  $G_W$  (right). The results are also reported in table 1. Light-color stacked bars represent the errors.

Finally, we see in the figure that the Standard Classifier is slightly less sensitive than the Quadratic one, but still its performances are not far from optimality. This is reassuring in light of possible applications of Statistical Learning methodologies to different problems, where the dependence of the distribution ratio on the new physics parameters is not known and the Quadratic Classifier approach cannot be adopted. On the other hand, the Standard Classifier method is rather demanding. First, because it requires separate trainings on a grid of values of  $c$ , out of which the reach should be extracted by interpolation. In turn, this requires a much larger number of training points than the Quadratic Classifier, since at each point of the grid we use as many training points as those the Quadratic Classifier needs in total for its training. Second, because we observed hyperparameters optimization depends on the specific value of  $c$  that is selected for training. Because of these technical difficulties, we only report sensitivity estimates for the positive Wilson coefficients reach. Furthermore these estimates (see table 1) are based on the  $p$ -value obtained at a given point of the  $c$  grid without interpolation. For the same reason, we did not try to apply the Standard Classifier methodology to the LO and to the NLO data and we focus on the Quadratic Classifier in what follows.

Let us discuss now the physical origin of the different behaviors observed for the  $\mathcal{O}_{\varphi q}^{(3)}$  and for the  $\mathcal{O}_W$  operator. The point is that the new physics effects due to  $\mathcal{O}_{\varphi q}^{(3)}$  have very distinctive features which can be easily isolated even with a simple binned analysis with few variables. Indeed  $\mathcal{O}_{\varphi q}^{(3)}$  (see eq. (3.3)) only contributes to the 00 polarization amplitude, which is non-vanishing in the SM as well and proportional to  $\sin \Theta$ . The squared 00 amplitude thus contributes to the cross section with a sizable interference term, which is peaked in the central scattering region  $\cos \Theta \sim 0$ . The other helicity channels play the role of background, and are peaked instead in the forward region. They are actually almost zero (at LO) at  $\cos \Theta \simeq 0$ . Therefore a binned analysis targeting central scattering (this is why we imposed the cut  $\cos \Theta < 0.5$ ) is sufficient to isolate the effects of  $\mathcal{O}_{\varphi q}^{(3)}$  at the interference level and thus to probe  $G_{\varphi q}^{(3)}$  accurately. By including the decay variables as in the multivariate analysis we gain sensitivity to new terms in the cross section, namely to the interference between the 00 and the transverse amplitudes, however these new terms

are comparable with those that are probed already in the Binned Analysis and thus they improve the reach only slightly.

The situation is very different for the  $\mathcal{O}_W$  operator. It contributes to the  $++$  and  $--$  helicity channels, that are highly suppressed in the SM and set exactly to zero in the Toy version of the problem we are studying here. The  $p_{T,Z}$  (and  $\Theta$ ) distribution depends only at the quartic level on  $G_W$ , i.e. through the square of the BSM amplitude, because the interference between different helicity channels cancels out if we integrate the cross section in eq. (3.8) over the  $ZW$  azimuthal decay angles. Our Binned Analysis is sensitive to the interference term through the binning in  $\varphi_W$ , however this is not enough to approach the optimal reach because all the other decay variables (and  $\Theta$  as well) do possess some discriminating power, from which we can benefit only through a multivariate analysis. More specifically, one can readily see by direct calculation that the dependence on all our kinematical variables of the  $G_W$  interference contribution to the differential cross section is different from the SM term. By integrating on any of this variables we partially lose sensitivity to this difference, and this is why the multivariate analysis performs much better than the binned one.

## 4.2 MADGRAPH Leading Order

The analyses performed for the Toy dataset can be easily replicated for the MADGRAPH LO Monte Carlo description of the process, obtaining the results shown in figure 2.

The most noticeable difference with what was found with the Toy Monte Carlo is the strong degradation of the Matrix Element reach, and the fact that it gets weaker than the one of the Quadratic Classifier. As usual, the effect is more pronounced for the  $\mathcal{O}_W$  operator. This is not mathematically inconsistent because the analytic ratio  $r(x, c)$  we employ for the Matrix Element test statistic is not equal anymore to the ratio of the true distributions according to which the data are generated. Therefore it is not supposed to give optimal performances. On the other hand the observed degradation is quantitatively surprising for  $G_W$ , especially in light of the fact that the MADGRAPH LO Monte Carlo distributions seem quite similar to the ones of the Toy data at a superficial look. The degradation is not due to the high-energy approximation in the  $ZW$  production amplitude, indeed the results we are reporting are obtained with the exact tree-level helicity amplitudes, which are employed in eq. (3.8) in place of the ones in eq. (3.3). It is due to the other approximations we performed in the calculation of the cross section, namely to the assumption that the initial quark is always more energetic than the anti-quark, which allows us to interpret  $\Theta$  as the angle between the quark and the  $Z$ , and to the one of a perfect reconstruction (up to the ambiguity) of the neutrino momentum. We verified that this is the case by repeating the Matrix Element analysis using the true neutrino momentum and the actual direction of motion of the quark in the Monte Carlo events. In this case the reach on  $G_W$  gets closer to the one obtained with the Toy data.

The degradation of the Matrix Element reach should be contrasted with the relative stability of the Quadratic Classifier method. Notice that the method is applied on the MADGRAPH LO data in the exact same way as on the Toy data, namely the architecture is the same, as well as the number of training point and the values of the Wilson coefficients

in eq. (4.4) used for training. The computational complexity of the distribution ratio reconstruction is thus identical in the two cases, in spite of the fact that the MADGRAPH LO Monte Carlo offers a slight more complete (or “complex”) description of the data. The total computational cost is somewhat higher in the MADGRAPH LO case, but just because the process of Monte Carlo events generation is in itself more costly. Similar considerations hold at NLO, where the cost of event generation substantially increases.

## 5 The reach at Next-to-Leading Order

Including NLO QCD corrections is in general essential for an accurate modeling of the LHC data. Therefore it is imperative to check if and to what extent the findings of the previous section are confirmed with the MADGRAPH NLO Monte Carlo description of the process, introduced in section 3.2. As far as the reconstruction of  $\widehat{r}(x, c)$  is concerned, using MADGRAPH NLO does not pose any conceptual or technical difficulty, provided of course the (positive and negative) Monte Carlo weights are properly included in the loss function as explained in section 2. Computing the distribution of the test statistic variable that we obtain after the reconstruction (or of the one we employ with the Matrix Element method, for which the exact same issue is encountered) is instead slightly more complicated than with the Toy and MADGRAPH LO data. This point is discussed in the following section, while the illustration of the results is postponed to section 5.2.

### 5.1 Estimating the test statistics distributions

As soon as  $\tau_c(x)$  is known, either as an analytic function in the case of the Matrix Element or as a (trained) Neural Network in the case of the Quadratic Classifier, the test statistic  $t_c(\mathcal{D})$ , as defined in eq. (4.1), is fully specified. Namely we can concretely evaluate it on any dataset  $\mathcal{D} = \{x_i\}$ , consisting of  $\mathcal{N}$  repeated instances of the variable  $x$ , for each given value of  $c$ . However in order to perform the hypothesis test, and eventually to estimate the reach  $c_{2\sigma}$ , we also need to estimate the probability distribution of  $t_c(\mathcal{D})$  under the  $H_0$  and under the  $H_1$  hypotheses. This is the problematic step at NLO, after which the evaluation of  $c_{2\sigma}$  proceeds in the exact same way as for the Toy and for the LO data. Specifically, once we are given with

$$\text{pdf}(t_c|H_0(c)) \text{ and } \text{pdf}(t_c|H_1), \tag{5.1}$$

we obtain the  $p$ -value as a function of  $t_c$  and  $c$  as in eq. (4.3) from the former, while from the latter we compute the median value of  $t_c$  and in turn

$$p_{\text{med}}(c) \equiv p(t_{\text{med}}(c); c), \tag{5.2}$$

as a function of  $c$ . After scanning over  $c$  and interpolating  $p_{\text{med}}(c)$  in the vicinity of the reach (actually we interpolate the logarithm of  $p_{\text{med}}(c)$ , using three points in  $c$  and quadratic interpolation), we can solve the equation  $p_{\text{med}}(c_{2\sigma}) = 0.05$  and obtain the reach as defined in eq. (4.2). Given the error on  $p_{\text{med}}(c)$  at the three points used for the interpolation, the error on the estimate of  $c_{2\sigma}$  is obtained by error propagation.

It is conceptually trivial (but numerically demanding) to estimate the distributions if artificial instances of the dataset  $\mathcal{D}$  (aka “toy” datasets) are available. In this case one can

simply evaluate  $t_c(\mathcal{D})$  on many toy datasets following the  $H_0(c)$  and the  $H_1$  hypotheses and estimate the distributions. More precisely, one just needs the empirical cumulative in  $H_0(c)$  and the median of  $t_c$  in  $H_1$ . Toy datasets are readily obtained from unweighted Monte Carlo samples by throwing  $\mathcal{N}$  random instances of  $x$  from the sample, with  $\mathcal{N}$  itself thrown Poissonianly around the total expected number of events. This is impossible at NLO because the events are necessarily weighted, therefore they are not a sampling of the underlying distribution of the variable  $x$ . As emphasized in section 2, NLO Monte Carlo data can only be used to compute expectation values of observables  $O(x)$  as in eq. (2.5). For instance we can compute the cross section in any region of the X space, and the mean or the higher order moments of the variable of interest,  $\tau_c(x)$ .

This suggests two options to estimate the distributions of the test statistic at NLO. The first one is to compute the distribution of  $\tau_c(x)$  by means of a (weighted) histogram with many and very narrow bins. By knowing the cross section of each bin in  $\tau_c$ , we know how many events are expected to fall in that bin and generate toy datasets for  $\tau_c$  accordingly. This procedure is quite demanding, and it relies on a careful choice of the  $\tau_c$  binning, which can only be performed on a case-by-case basis. It is still useful to validate the strategy we actually adopt, described below.

The second option is to approximate the distribution of  $t_c$  in a “nearly Gaussian” form, based on the Central Limit theorem. Namely we notice that  $t_c$  is in a trivial linear relation (see eq. (4.1)) with the variable

$$\mathcal{T}_c(\mathcal{D}) \equiv \frac{1}{N} \sum_{i=1}^{\mathcal{N}} \tau_c(x_i), \tag{5.3}$$

where  $\mathcal{N}$  is Poisson-distributed with expected  $N$ , with  $N = N(X|H)$  and  $H = H_0$  or  $H = H_1$ . The  $x_i$ 's are independent and sampled according to pdf( $x|H$ ). The cumulant-generating function of  $\mathcal{T}_c$  (which is a so-called “compound” Poisson variable [61]) is readily computed

$$K_{\mathcal{T}_c}(\xi) \equiv \log \left\{ \mathbb{E} \left[ e^{\xi \mathcal{T}_c} \mid H \right] \right\} = N \mathbb{E} \left[ e^{\frac{\xi}{N} \tau_c} \mid H \right] - N, \tag{5.4}$$

by first taking the expectation on the  $x_i$ 's conditional to  $\mathcal{N}$  and next averaging over the Poisson distribution of  $\mathcal{N}$ . Therefore the cumulants of  $\mathcal{T}_c$ ,

$$\kappa_{\mathcal{T}_c}^n \equiv \left. \frac{d^n K_{\mathcal{T}_c}(\xi)}{d\xi^n} \right|_{\xi=0} = N^{1-n} \mathbb{E} [\tau_c^n \mid H], \tag{5.5}$$

are increasingly suppressed with  $N$  for larger and larger  $n > 1$ . Since  $N$  is of the order of several thousands in our case, neglecting all cumulants apart from the first and the second one, i.e. adopting a Gaussian distribution for  $\mathcal{T}_c$ , might be a good approximation.

Actually it turns out that in order to model properly the 5% tail of the distribution, which we need to probe for the exclusion limit, non-Gaussianity effects can be relevant. These are included by using a skew-normal distribution for  $\mathcal{T}_c$ , which contains one more adjustable parameter than the Gaussian to model the skewness. The mean, standard deviation and skewness of  $\mathcal{T}_c$  are immediately obtained from eq. (5.5)

$$\mu(\mathcal{T}_c) = \langle \tau_c \rangle, \quad \sigma(\mathcal{T}_c) = \frac{1}{\sqrt{N}} \sqrt{\langle \tau_c^2 \rangle} \quad \mu_3(\mathcal{T}_c) = \frac{1}{\sqrt{N}} \frac{\langle \tau_c^3 \rangle}{\langle \tau_c^2 \rangle^{3/2}}, \tag{5.6}$$



		Toy Data	LO	NLO
$G_{\varphi q}^{(3)}$	ME	$[-0.350(6), 0.356(8)]$	$[-0.399(13), 0.384(12)]$	$[-0.55(4), 0.464(14)]$
	SC	$\gtrsim 0.4 (p = 0.077(5))$	—	—
	QC	$[-0.357(6), 0.365(8)]$	$[-0.401(12), 0.374(10)]$	$[-0.426(22), 0.401(21)]$
	BA	$[-0.48, 0.47]$	$[-0.50, 0.50]$	$[-0.58, 0.55]$
$G_W$	ME	$[-0.673(14), 0.697(11)]$	$[-1.390(21), 1.357(22)]$	$[-1.51(7), 1.93(14)]$
	SC	$\lesssim 1 (p = 0.038(3))$	—	—
	QC	$[-0.781(13), 0.822(13)]$	$[-1.007(27), 0.987(26)]$	$[-0.99(4), 1.08(12)]$
	BA	$[-1.67, 1.67]$	$[-1.70, 1.85]$	$[-1.63, 1.98]$

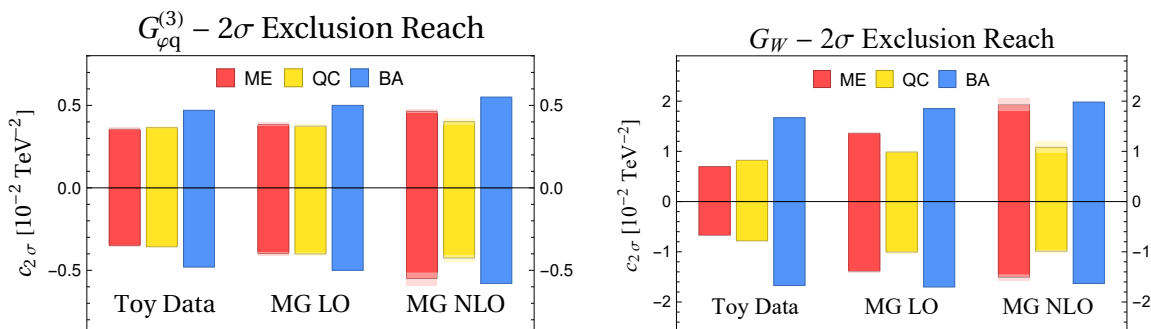
**Table 1.** Bounds on the  $G_{\varphi q}^{(3)}$  and  $G_W$  coefficients obtained for the Toy, LO and NLO datasets. The rows correspond to the Matrix Element (ME), Standard Classifier (SC), Quadratic Classifier (QC) and Binned Analysis (BA) approach. Notice that the errors on the Binned Analysis bounds are negligible. The results are given in  $10^{-2} \text{ TeV}^{-2}$  units.

where  $\langle \cdot \rangle$  is used to denote expectation for brevity. By computing the expectation values of  $\tau_c$ ,  $\tau_c^2$  and  $\tau_c^3$  using the Monte Carlo data, we thus find the parameters of the skew-normal distribution for  $\mathcal{T}_c$  and in turn the distribution of  $t_c$ . We finally obtain the median  $p$ -value from the definition in eq. (5.2). The errors on the expectation values of  $\tau_c$  are estimated from the fluctuations in the means on subsets of the entire Monte Carlo sample. These errors are propagated to the  $p$ -value and eventually to the  $c_{2\sigma}$  estimated reach as previously explained. Accurate results (see table 1) are obtained with relatively small Monte Carlo samples. Namely, 500k event were used at NLO, 1M at LO and 3M for the Toy data.

We cross-checked the above procedure in multiple ways. First, it reproduces within errors the LO and Toy  $p$ -values obtained with the toy experiments. Second, we validated it against the approach based on  $\tau_c$  binning on NLO data, as previously mentioned. We also verified that including the skewness changes the results only slightly, with respect to those obtained in the Gaussian limit. Further improving the modeling of the non-Gaussianity with more complex distributions than the skew-normal, with more adjustable parameters in order to fit higher order moments of  $\mathcal{T}_c$ , is therefore not expected to affect the results.

## 5.2 Results

Our results with the MADGRAPH NLO Monte Carlo are reported in figure 3 and in table 1. They essentially confirm the trend we already observed in the transition from the Toy to the MADGRAPH LO data. The Matrix Element keeps losing sensitivity because the analytic distribution ratio is now even more faraway from the actual distribution ratio since it does not include NLO QCD effects. The reach of the Binned Analysis deteriorates less, so that it becomes comparable to the one of the Matrix Element. The Quadratic Classifier reach is remarkably stable. Actually it slightly improves with respect to the LO one for  $G_W$ . This is



**Figure 3.** Expected exclusion reach on  $G_{\varphi q}^{(3)}$  (left) and on  $G_W$  (right) with the various methodologies described in the text. The results are also reported in table 1.

probably due to the appearance of same-helicity SM transverse amplitudes (see section 3.2) and of the corresponding interference term for the  $\mathcal{O}_W$  operators.

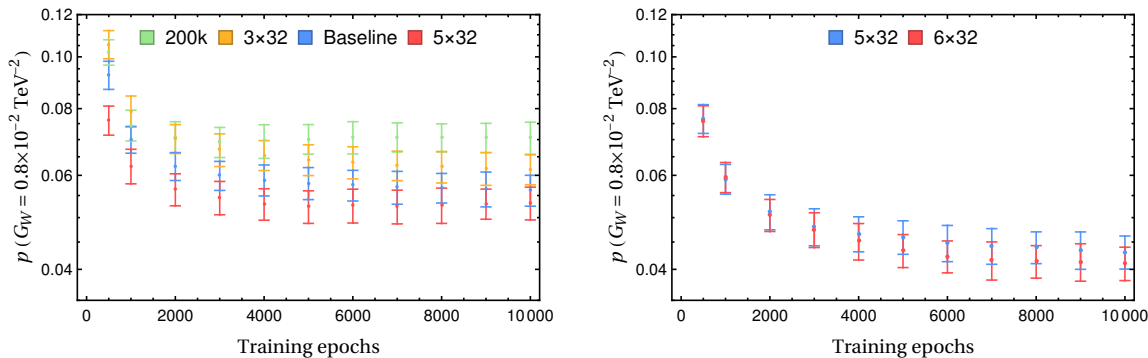
Notice few minor differences in the implementation of the Quadratic Classifier and of the Binned Analysis at NLO. The Quadratic Classifier now also employs the variable  $p_{T,ZW}$ , as discussed in section 3. The Binned Analysis for  $G_{\varphi q}^{(3)}$  employs  $p_{T,ZW}$  as well, through a cut  $p_{T,ZW}/p_{T,V} < 0.5$ . This improves the reach [40] because it helps recovering (partially) the background suppression due to the zero of the transverse amplitudes in the central region.

## 6 Neural Network implementation and validation

The strategies described in section 2 were implemented in Pytorch [62] and run on NVIDIA GeForce GTX 1070 graphics card. Fully connected feedforward deep Neural Networks were employed, acting on the features vector

$$x = \{s, \Theta, \theta_W, \theta_Z, p_{T,ZW}, p_{T,Z}, \sin \varphi_W, \cos \varphi_W, \sin \varphi_Z, \cos \varphi_Z\}, \quad (6.1)$$

for a total of 10 features. Each feature is standardized with a linear transformation to have zero mean and unit variance on the training sample. For the Quadratic Classifier training, the Wilson coefficient employed in the parametrization (2.12) were scaled to have unit variance on the training sample. Employing the redundant variables (i.e.,  $p_{T,Z}$ , and the cosines and sines of  $\varphi_{W,Z}$ ) is helpful for the performances, especially the angular ones, which enforce the periodicity of the azimuthal angular variables. The “baseline” results presented in figures 2, 3 and in table 1 were all obtained with the features vector above and employing a total of 6 million training Monte Carlo points for each of the two Wilson coefficients. Training was always performed with a single batch (which was found to perform better in all cases), even if in practice the gradients calculation was split in mini-batches of 100k points in order to avoid saturating the memory of the GPU. Apart from these common aspects, the optimization of the Neural Network design and of the training strategy is rather different for the Quadratic and for the Standard Classifier methods. They are thus discussed separately in what follows.

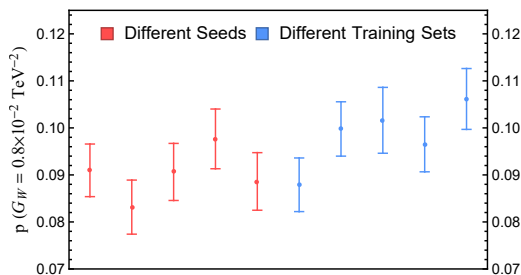


**Figure 4.** Evolution of the  $p$ -value for different architectures and training sample sizes. On the left plot we compare the baseline setup with the baseline architecture Network trained with 200k points per value of  $c$  (for a total of 2.4M points), and with the baseline number of training points (500k, times 12) on architectures with one less (“ $3 \times 32$ ”) and one more (“ $5 \times 32$ ”) hidden layer. On the right plot, a similar analysis is performed, but with 3M points per value of  $c$ .

### 6.1 The Quadratic Classifier

For the Quadratic Classifier, best performances were obtained with ReLU activation functions and with the Adam Pytorch optimizer. The initial learning rate (set to  $10^{-3}$ ) does not strongly affect the performances. Other attempts, with Sigmoid activation and/or with SGD optimizer, produced longer execution time and worse performances. The baseline architecture for the two Neural Networks  $n_\alpha$  and  $n_\beta$  in eq. (2.12) consists of 4 hidden layers with 32 neurons, namely the architecture  $\{10, 32, 32, 32, 32, 1\}$ , including the input and the output layers. Weight Clipping was implemented as a bound on the  $L_1$  norm of the weights in each layer, but found not to play a significant role. The total training time, for  $10^4$  training epochs, is around 5 hours for the baseline architecture and with the baseline number (6 million) of training points.

The Neural Network architecture was selected based on plots like those in figure 4. The left panel shows the evolution with the number of training epochs of the median  $p$ -value (see eq. (5.2)) on Toy data for  $c = G_W = 0.8 \times 10^{-2} \text{ TeV}^{-2}$ , with the baseline and with larger and smaller Networks. We see that adding or removing one hidden layer to the baseline architecture does not change the performances significantly. The plot also shows that  $10^4$  epochs are sufficient for the convergence and that no overfitting occurs. The degradation of the performances with less training point is also illustrated in the plot. Of course, the  $p$ -value is evaluated using independent Monte Carlo samples, not employed for training. The errors on the  $p$ -value are estimated from the error on the skew-normal distribution parameters as explained in section 5.1. In the baseline configuration we used 500k EFT Monte Carlo training points for each of the 6 values of  $G_W$  in eq. (4.4), plus 500k for each associated SM sample. Each sample consists instead of 3M points in the extended configuration employed on the right panel of figure 4, for a total of 36M. The same value of  $G_W = 0.8 \times 10^{-2} \text{ TeV}^{-2}$  is employed. The baseline architecture becomes insufficient, and best results are obtained with the 6 hidden layers of 32 neurons each.



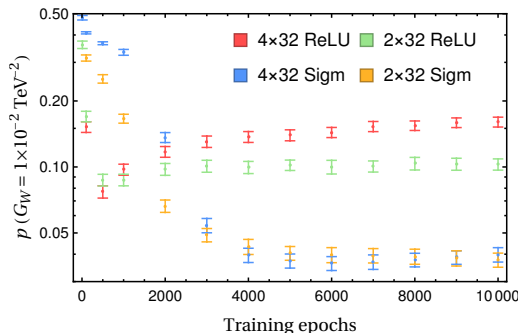
**Figure 5.** Results of 5 different trainings of the same architecture (Baseline architecture trained with 2.4M points) using: the same training data but different initialization seeds (red points) and the same initialization but different training data samples (blue points).

The figure also demonstrates that the method is systematically improvable towards optimality. The value of  $G_W$  considered in the figure was not within the 95% CL reach with the baseline setup, while it becomes visible with the extended configuration. All the reaches reported in table 1 would expectedly improve with the extended configuration. The  $G_W$  reach on Toy data becomes  $[-0.732(9), 0.764(14)] 10^{-2} \text{ TeV}^{-2}$ , which is now only less than 10% worse than the optimal Matrix Element reach. Training takes around 30 hours with the extended configuration, while generating and processing the required training points with MADGRAPH NLO (which is the most demanding generator) would take around 10 days on a 32-cores workstation. We could thus try to improve also the NLO reach even with limited computing resources.

For the reproducibility of our results we also study how the performances depend on the Neural Network initialization and on the statistical fluctuations of the Monte Carlo training sample. This analysis is performed in a reduced setup, with a total of 2.4 million training point, and for  $G_W = 0.8 \times 10^{-2} \text{ TeV}^{-2}$ . We see in figure 5 that the  $p$ -value fluctuates by varying the random seed used for training at a level comparable with the error on its determination. Similar results are observed by employing different independent Monte Carlo training samples. Notice that these fluctuations should not be interpreted as additional contributions to the error on the  $p$ -value. Each individual Neural Network obtained from each individual training defines a valid test statistic variable, on which we are allowed to base our statistical analysis. Since the fluctuations are comparable to the  $p$ -value estimate errors, our sensitivity projections were obtained by randomly selecting one of the seed/training set configuration.

## 6.2 The Standard Classifier

Hyperparameters optimization is rather different for the Standard Classifier. We see in figure 6 that Networks with ReLU activation like those we employed for the Quadratic Classifier displays overfitting, and Sigmoid activations need to be employed. The results in figures 2 and in table 1 were obtained with 2 hidden layers with 32 neurons each and Sigmoid activation. The figure shows that increasing the complexity does not improve the performances.



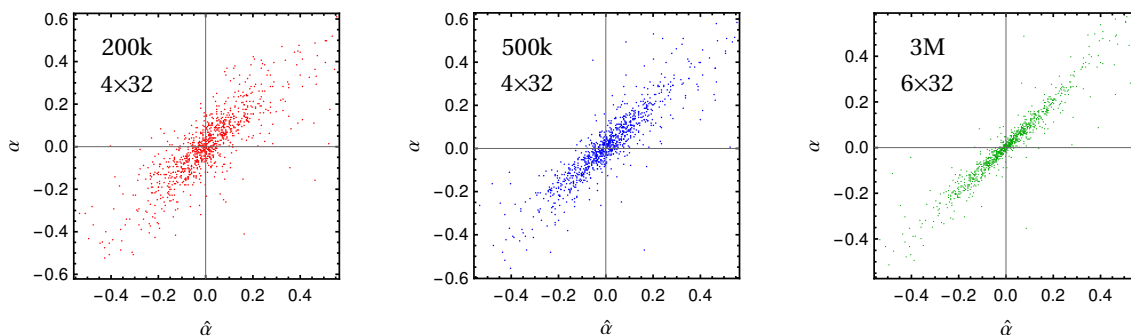
**Figure 6.** The  $p$ -value evolution during training for the Standard Classifier using different architectures and activation functions. The value  $G_W = 1 \times 10^{-2} \text{ TeV}^{-2}$  is employed.

This different behavior of the Standard Classifier compared with the Quadratic one is probably due to the fact that training is performed on small Wilson coefficient EFT data, whose underlying distribution is very similar to the one of the SM data sample. Therefore there is not much genuine difference between the two training sets, and the Network is sensitive to statistical fluctuations in the training samples. The Quadratic Classifier instead is trained with large values of the Wilson coefficients. The optimizer thus drives the Neural Networks towards the deep minimum that corresponds to a proper modeling of the distribution ratio, which is more stable against statistical fluctuations of the training samples.

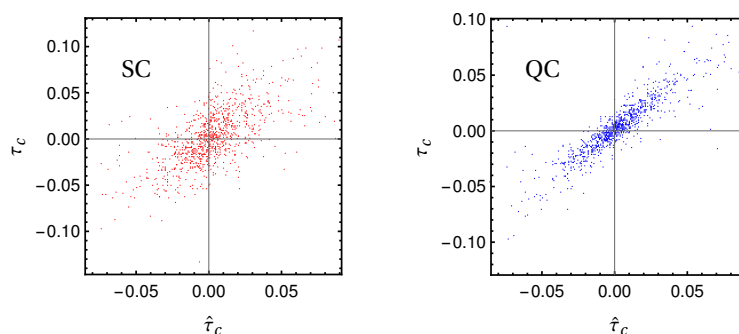
### 6.3 Validation

An important question is how to validate as “satisfactory” the outcome of the hyperparameters optimization described above. This is straightforward for the Toy version of the problem, because we have to our disposal a rigorous notion of statistical optimality, through the Neyman-Pearson lemma, and we do have direct access to the true distribution ratio through which the data are generated. Therefore we know that we can stop optimization as soon as the reach of the Neural Network becomes sufficiently close to the one of the Matrix Element method. We can also rely on a more naive validation test, based on comparing point-by-point the distribution ratio learned by the Neural Network with the true one, which is known analytically. For instance in figure 7 we compare the true linear term  $\alpha(x)$  in eq. (2.10) (for the  $\mathcal{O}_W$  operator) with its estimator  $\hat{\alpha}(x) \equiv \hat{n}_\alpha(x)$  provided by the trained Neural Network. The baseline architecture is employed, with increasing number of training points. While it is impossible to extract quantitative information, a qualitative comparison between the three scatter plots confirms that more training points improve the quality of the reconstruction. We also show, in figure 8, the correlation between the true and the reconstructed ratios (for  $G_W = 1 \times 10^{-2} \text{ TeV}^{-2}$ , which corresponds to the Standard Classifier 95% reach) obtained with the Quadratic and with the Standard Classifier. The reconstruction obtained with the Quadratic Classifier is more accurate as expected.

Validation is of course less easy if, as it is always the case on real problems, the true distribution ratio is not known. One option is to proceed like we did in the present paper.



**Figure 7.** Comparison between the reconstructed ( $\hat{\alpha}$ ) and true ( $\alpha$ ) linear term of the distribution ratio for the  $G_W$  operator.



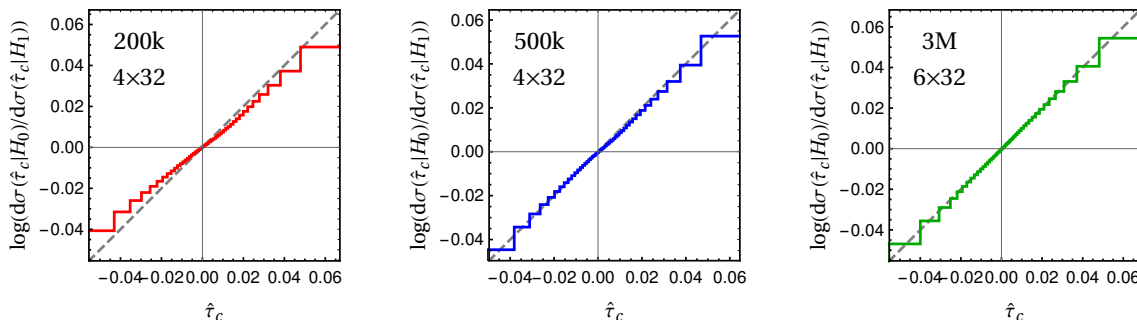
**Figure 8.** Comparison between the reconstructed ( $\hat{\tau}_c$ ) and true ( $\tau_c$ ) distribution log-ratio for  $G_W = 1 \times 10^{-2} \text{ TeV}^{-2}$ . The Standard Classifier and the Quadratic one are considered in the left and right panel of the figure, respectively.

Namely to identify a Toy version of the problem that is sufficiently close to the real one and for which the distribution ratio is known. Since it is unlikely that the true distribution is much harder to learn than the Toy distribution, and since we can establish optimality on the Toy data using a certain architecture and training dataset size, we can argue heuristically that the same configuration will be optimal also with a more refined Monte Carlo description.

Finally, one can monitor heuristically how accurately the distribution ratio is reconstructed, as follows. The true distribution log-ratio  $\tau_c(x) = \log r(x, c)$ , seen as a statistical variable for each fixed value of  $c$ , obeys, by definition, the equation

$$\frac{d\sigma_0}{d\tau_c} = e^{\tau_c} \frac{d\sigma_1}{d\tau_c}. \quad (6.2)$$

Therefore if we computed the distribution of  $\tau_c$  (if it was known) in the EFT hypothesis  $H_0(c)$  and in the SM hypothesis  $H_1$ , and take the log-ratio, the result would be a straight line as a function of  $\tau_c$ . By computing the same distributions for the reconstructed distribution log-ratio  $\hat{\tau}_c = \log \hat{r}(x, c)$ , we can thus get an indication of how closely  $\hat{r}(x, c)$  approximates  $r(x, c)$ . While no quantitative information can be extracted from these plots, they clearly illustrate the improvement achieved by enlarging the size of the training sample and the Neural Network architecture, as figure 9 shows.



**Figure 9.** Distribution log-ratio for  $\hat{\tau}_c$ , for  $c = G_W = 0.8 \times 10^{-2} \text{ TeV}^{-2}$ . The accurate determination displayed in the plots is obtained by the reweighting of a single Toy SM Monte Carlo sample. The same approach, based on reweighting, could have been adopted to assess the quality of the distribution ratio reconstruction on MADGRAPH Monte Carlo data, using MADWEIGHT.

## 7 Conclusions and outlook

We studied the potential gain in sensitivity of EFT searches at the LHC from multivariate analysis techniques. The results reported in figure 3 show that a considerable improvement is possible, especially for operators (like  $\mathcal{O}_W$ ) with a complex interference pattern that is difficult to capture with a Binned Analysis.

Multivariate analyses based on Statistical Learning techniques are particularly promising, and should be considered as an alternative to the more standard (though not yet employed for EFT LHC searches) Matrix Element method. The advantage is eminently practical, because the Matrix Element method is optimal in principle, as much as the Statistical Learning approach. However the Matrix Element method needs to be designed case-by-case, and re-designed for each new effect one is willing to add for a more accurate modeling of the distribution ratio. It already required some effort to compute the approximate distribution in section 3.1, which in turn provides the simplest modeling of the distribution ratio to be employed in the Matrix Element approach, and we saw that this modeling is inadequate to describe the LO and even less adequate at NLO. In order to improve the modeling in the case at hand one should model the neutrino reconstruction more accurately, for instance by performing the integral over the neutrino momentum point-by-point in the space of the observed kinematical variables. The integral on the radiation should be also performed if willing to add QCD NLO effects. The predictions should be further refined including transfer functions for the detector effects, if the method has to be employed on real data.

The situation is radically different with the Statistical Learning approach. We saw that the exact same computational effort is required to reconstruct the distribution ratio at the Toy level, at LO and at NLO. Furthermore the accuracy of the reconstruction can be systematically improved using more training points and bigger Networks. The limiting factor is not reconstructing the distribution ratio by the Neural Network training. That step takes a quite small fraction of the computing time. The most time-consuming part of the procedure is the generation of the Monte Carlo training data, which becomes increasingly demanding as the sophistication of the Monte Carlo code increases. Even if we are still far

from the limit for our analysis, it would be worth investigating improvements on this aspect based on Monte Carlo reweighting techniques.

It should be emphasized that Machine Learning methodologies are useful for EFT studies not only in view of the possible application to the analysis of the real data. After the conceptual and technical framework is in place, it is very easy to run the Machine Learning algorithm on the specific EFT problem at hand, and to get a feeling of the potential improvement of the reach compared with other methods. For instance our results show that the Binned Analysis we employed is inadequate for  $G_W$ , and that even for  $G_{\varphi q}^{(3)}$  it could be improved. Furthermore they provide a target for the sensitivity such improvements should attain. Similarly, the results outline the importance of neutrino reconstruction modeling and of NLO QCD corrections being implemented in the Matrix Element method, if one is willing to adopt that strategy.

When it comes to the direct applicability of the method to the data, of the ZW process for instance, two additional steps are needed. The first one is to further improve the level of detail of the simulation. Detector effects could be added very easily with DELPHES [63]. However the reliability of the DELPHES description of the detectors should be cross-checked with a complete simulation by the experimental collaborations, and the DELPHES simulation replaced with a full detector simulation, which is much more demanding, if needed.

The second aspect is to include systematic uncertainties of theoretical and experimental origin. It should be stressed that this is not more problematic in the Machine Learning framework than it is in the Matrix Element or any other multivariate approach. In particular it should be noticed that one has full control on the choice of the input variables that are given to the Neural Network and from which the sensitivity emerges. For instance in our case these would be the kinematical variables of the high-level reconstructed leptons, better if including photon recombination, in order to reduce the sensitivity to detector effects and showering, which might not be modeled accurately enough. Similarly if jets were used in the final state, high level IR-safe observables would be employed to be insensitive to hadronization, exactly like one would do for the Matrix Element method. It should also be stressed, as explained in the Introduction, that our method can be employed also in the presence of reducible backgrounds that must be extracted from the data because no reliable Monte Carlo generator is available.

The simplest strategy to deal with uncertainties is to merely quantify their impact on the sensitivity, using as discriminating variable the distribution ratio reconstructed from the nominal Monte Carlo generator that does not incorporate uncertainties. This is suboptimal, but sufficient to obtain conservative (i.e. correct) results, and to identify the irrelevant sources of uncertainties. For better results one can include the uncertainties in the likelihood (i.e. in the reconstructed distribution ratio) in the form of nuisance parameters. This is perfectly compatible with the Machine Learning approach, and already implemented in MADMINER [34] through morphing. Actually the Quadratic Classifier we employ in this paper could be useful also for this task. We will return to this point at the end of the section. While conceptually straightforward, it is quantitatively important to assess the impact of uncertainties on the sensitivities we obtained in figure 3 on purely statistical grounds. This is left to future work.



One interesting technical element of the present paper is the Quadratic Classifier, introduced in section 2.2. We have found that it performs better than the Standard Classifier, as expected since it is designed to be sensitive to the small departures from the SM due to the EFT by exploiting the exact knowledge of the (quadratic) functional dependence of the distribution ratio on the Wilson coefficients. Furthermore it is computationally much more convenient and thus feasible also when several EFT operators are considered simultaneously and the scan over the Wilson coefficients becomes unfeasible. The Quadratic Classifier has been found to be nearly optimal, with a rigorous notion of optimality based on the Neyman-Pearson lemma.

We described in the body of the paper the connection between the Quadratic Classifier and other techniques based on Statistical Learning available in the literature, but we did not yet discuss the relation with the most sophisticated such techniques, namely the ones that exploit “hidden” information from the Monte Carlo simulator [32]. The basic idea is that the simulator does contain the analytic information on the underlying distribution, and so it does contain a representation of the EFT/SM distribution ratio in terms of latent variables. One can incorporate this information in the loss function, so that the machine does not need to learn the likelihood ratio from scratch, but only the distortions of the likelihood ratio due to the transition between the latent and the true variables. The Quadratic Classifier trick is orthogonal to this interesting idea, and it could be straightforwardly implemented in the simulator-assisted methods by modifying the loss function in close analogy with eq. (2.13). The advantages of parametrization in that context could be the same we observed here.

On the other hand, simulator-assisted methods have also potential limitations, in two respects. First, because there is a clear benefit from exploiting the latent-space distribution ratio if the latter is similar to the one in the space of observables, but this is not necessarily the case. For instance in ZW we saw that a proper modeling of the neutrino reconstruction is crucial for the performances, and this is not captured by the latent-variables ratio that involves the true neutrino momentum. This can be a problem for the validation of the approach, due to the fact that any additional effect we include in the simulation, which further distorts the observed ratio, might be more and more difficult for the machine to learn. For instance a simulator-assisted method should be trivially optimal on the Toy data, where the latent space coincides with the observed space and thus the likelihood ratio employed in training coincides with the true one and the machine has nothing to learn. However this does not mean that it will work on the LO data (using the appropriate LO latent-variables ratio) because now the machine has the non-trivial task to integrate out the neutrino. Instead for our method, that learns the distribution ratio using no information from the Monte Carlo apart from the event sample itself, it is arguably equally difficult to model the distribution ratio on the Toy, on the LO and on the NLO data. Therefore the optimality on Toy data, which we can establish rigorously because we know the exact distribution ratio, heuristically indicates that the algorithm is optimal at LO and NLO as well. The second problem of simulator-assisted method is that the required information on the latent-space distribution ratio might not be made available by the Monte Carlo code. In light of this, it is reassuring to have an alternative method that does not rely on latent-space information, that is feasible and optimal, at least in the case at hand.

Finally, it should be noticed that the parametrization trick is not specific of the EFT and it could be applied to any situation where the functional dependence of the distribution on the parameters is either exactly or approximately known. One should just replace the quadratic dependence of eq. (2.12) on  $c$  with the appropriate (polynomial or not) functional form. This could be useful to include the effect of nuisance parameters in the likelihood. Nuisance parameters effects on the distribution can be normally modeled linearly (or with an exponential, to avoid negative distributions) to good approximation because their effects are small. However if they are too small (but still potentially competitive with the EFT ones) it could be difficult for the machine to learn them using simulations where the nuisances are varied within their one-sigma interval. If the analytic dependence on the nuisance parameters is incorporated in the classifier, we could ameliorate the situation by training with larger values of the parameters like we did in this paper to reconstruct the small EFT effects. Exploring this direction is left to future work.

### Acknowledgments

We thank J. Brehmer and K. Cranmer for useful discussions. The work of S.C. was supported by the Swiss National Science Foundation under contract 200021-178999. The Swiss National Science Foundation supported the work of A.G. under contracts 200020-169696. G.P. was supported in part by the MIUR under contract 2017FMJFMW (PRIN2017).

### A The general Quadratic Classifier

Any quadratic-order real polynomial of  $n-1$  variables  $c_i, i = 1, \dots, n-1$ , with arbitrary constant, linear and quadratic terms, can be written as a quadratic form in the  $n$ -dimensional variable

$$v(c) = (1, c_1, \dots, c_{n-1})^T. \tag{A.1}$$

Namely, we write the polynomial as

$$P(c) = v^T(c) A v(c), \tag{A.2}$$

with  $A$  a generic  $n$ -dimensional real symmetric square matrix.

If  $P(c)$  is non-negative for any value of  $c$ , it is easy to show that the matrix  $A$  must be positive semi-definite. Being real, symmetric and positive semi-definite, it is possible to use the Cholesky decomposition for  $A$ , and write it as

$$A = L^T L, \tag{A.3}$$

where  $L$  is a upper-triangular (i.e.,  $L_{ij} = 0$  for  $j < i$ ) real matrix. Therefore the most general positive quadratic order polynomial reads

$$P(c) = v^T(c) L^T L v(c) = \sum_{i=1}^n \left( \sum_{j=1}^n L_{ij} v_j(c) \right)^2 = \sum_{i=1}^n \left( L_{i1} + \sum_{j=2}^n L_{ij} c_{j-1} \right)^2, \tag{A.4}$$

which is manifestly non-negative because it is the sum of square terms. Moreover for  $c = 0$ , since  $L_{i1} = L_{11}\delta_{i1}$ , we have  $P(0) = L_{11}^2$ . The Cholesky decomposition is unique up to sign flips of the rows of  $L$ . Rather than resolving this ambiguity, for instance by choosing the diagonal entries of  $L$  to be positive, we adopt eq. (A.4) without further constraints as the most general (though redundant) parametrization of  $P(c)$ .

The EFT differential cross section is a positive quadratic polynomial in the Wilson Coefficient  $c_i$  at each phase-space point  $x$ , and it reduces to the SM cross section for  $c = 0$ . It must therefore take the form

$$d\sigma_0(x; c) = d\sigma_1(x) \sum_{i=1}^n \left[ \delta_{i1} + \sum_{j=2}^n \lambda(x)_{ij} c_{j-1} \right]^2, \quad (\text{A.5})$$

with  $\lambda(x)$  an upper-triangular matrix of real functions. If only one  $c$  parameter is present (i.e.,  $n = 2$ ), this reduces to eq. (2.10) with the identifications

$$\lambda(x)_{12} = \alpha(x) \quad \lambda(x)_{22} = \beta(x). \quad (\text{A.6})$$

The Quadratic Classifier that generalizes eq. (2.12) is thus defined as

$$f(x, c) \equiv \frac{1}{1 + \sum_{i=1}^n \left[ \delta_{i1} + \sum_{j=2}^n n(x)_{ij} c_{j-1} \right]^2}, \quad (\text{A.7})$$

in terms of an upper-triangular matrix  $n(x)$  of real-output Neural Networks.

## B Minimization of the parametrized loss

In the Large Sample limit, the loss function in eq. (2.13) becomes

$$L[n(\cdot)] \stackrel{\text{LS}}{=} \sum_{c \in \mathcal{C}} \left\{ \int d\sigma_0(x; c) [f(x, c)]^2 + \int d\sigma_1(x) [1 - f(x, c)]^2 \right\}, \quad (\text{B.1})$$

with the Quadratic Classifier  $f$  defined in eq. (A.7). By simple algebraic manipulations, this can be rewritten as

$$L[n(\cdot)] \stackrel{\text{LS}}{=} \sum_{c \in \mathcal{C}} \left\{ \int \frac{d\sigma_1(x) d\sigma_0(x; c)}{d\sigma_1(x) + d\sigma_0(x; c)} + \int [d\sigma_1(x) + d\sigma_0(x; c)] \left[ f(x, c) - \frac{1}{1 + r(x, c)} \right]^2 \right\}, \quad (\text{B.2})$$

with  $r(x, c) = d\sigma_0(x; c)/d\sigma_1(x)$ . The first integral is independent of  $f$  and thus it is irrelevant for the minimization of the loss. The second one is the integral of a non-negative function of  $x$  which attains its global minimum (i.e., it vanishes) if and only if

$$f(x, c) = f_{\min}(x, c) = \frac{1}{1 + r(x, c)}, \quad \forall c \in \mathcal{C}. \quad (\text{B.3})$$

By using eq. (A.5), and comparing with eq. (A.7), we immediately conclude that the configuration  $n(x)_{ij} = \lambda(x)_{ij}$  is a global minimum of the loss and that this minimum is unique

provided the set  $\mathcal{C}$  contains at least two distinct non-vanishing elements. More precisely, this holds only up to sign ambiguities, associated with those of the Cholesky decomposition. However this is irrelevant because the ambiguity cancels out in  $f$ , and in turn it cancels out in the reconstructed distribution ratio  $\widehat{r}(x, c) = 1/\widehat{f}(x, c) - 1$ .

We have shown that the Quadratic Classifier reconstructs the distribution ratio exactly (in the Large Sample limit and for infinitely complex Neural Network) at the global minimum of the loss, and that this minimum is unique. Notice however that we could not show that the Large Sample limit loss does not possess additional local minimums, as it is instead readily proven for the standard classifier of section 2.1 by variational calculus.

**Open Access.** This article is distributed under the terms of the Creative Commons Attribution License ([CC-BY 4.0](https://creativecommons.org/licenses/by/4.0/)), which permits any use, distribution and reproduction in any medium, provided the original author(s) and source are credited.

## References

- [1] W. Buchmüller and D. Wyler, *Effective Lagrangian Analysis of New Interactions and Flavor Conservation*, *Nucl. Phys. B* **268** (1986) 621 [[INSPIRE](#)].
- [2] G.F. Giudice, C. Grojean, A. Pomarol and R. Rattazzi, *The Strongly-Interacting Light Higgs*, *JHEP* **06** (2007) 045 [[hep-ph/0703164](#)] [[INSPIRE](#)].
- [3] B. Grzadkowski, M. Iskrzynski, M. Misiak and J. Rosiek, *Dimension-Six Terms in the Standard Model Lagrangian*, *JHEP* **10** (2010) 085 [[arXiv:1008.4884](#)] [[INSPIRE](#)].
- [4] D. Atwood and A. Soni, *Analysis for magnetic moment and electric dipole moment form-factors of the top quark via  $e^+e^- \rightarrow t\bar{t}$* , *Phys. Rev. D* **45** (1992) 2405 [[INSPIRE](#)].
- [5] M. Diehl and O. Nachtmann, *Optimal observables for the measurement of three gauge boson couplings in  $e^+e^- \rightarrow W^+W^-$* , *Z. Phys. C* **62** (1994) 397 [[INSPIRE](#)].
- [6] G. Durieux, M. Perelló, M. Vos and C. Zhang, *Global and optimal probes for the top-quark effective field theory at future lepton colliders*, *JHEP* **10** (2018) 168 [[arXiv:1807.02121](#)] [[INSPIRE](#)].
- [7] J. Pretz and F. Müller, *Extraction of Azimuthal Asymmetries using Optimal Observables*, *Eur. Phys. J. C* **79** (2019) 47 [[arXiv:1811.09452](#)] [[INSPIRE](#)].
- [8] B. Bortolato, J.F. Kamenik, N. Košnik and A. Smolkovič, *Optimized probes of CP-odd effects in the  $t\bar{t}h$  process at hadron colliders*, *Nucl. Phys. B* **964** (2021) 115328 [[arXiv:2006.13110](#)] [[INSPIRE](#)].
- [9] D.A. Faroughy, J.F. Kamenik, N. Košnik and A. Smolkovič, *Probing the CP nature of the top quark Yukawa at hadron colliders*, *JHEP* **02** (2020) 085 [[arXiv:1909.00007](#)] [[INSPIRE](#)].
- [10] I. Dunietz, H.R. Quinn, A. Snyder, W. Toki and H.J. Lipkin, *How to extract CP-violating asymmetries from angular correlations*, *Phys. Rev. D* **43** (1991) 2193 [[INSPIRE](#)].
- [11] A.S. Dighe, I. Dunietz and R. Fleischer, *Extracting CKM phases and  $B_s - \bar{B}_s$  mixing parameters from angular distributions of nonleptonic B decays*, *Eur. Phys. J. C* **6** (1999) 647 [[hep-ph/9804253](#)] [[INSPIRE](#)].
- [12] G. Durieux and Y. Grossman, *Probing CP-violation systematically in differential distributions*, *Phys. Rev. D* **92** (2015) 076013 [[arXiv:1508.03054](#)] [[INSPIRE](#)].

- [13] S. Banerjee, R.S. Gupta, J.Y. Reiness, S. Seth and M. Spannowsky, *Towards the ultimate differential SMEFT analysis*, *JHEP* **09** (2020) 170 [[arXiv:1912.07628](#)] [[INSPIRE](#)].
- [14] I. Anderson et al., *Constraining Anomalous HVV Interactions at Proton and Lepton Colliders*, *Phys. Rev. D* **89** (2014) 035007 [[arXiv:1309.4819](#)] [[INSPIRE](#)].
- [15] A.V. Gritsan, R. Röntsch, M. Schulze and M. Xiao, *Constraining anomalous Higgs boson couplings to the heavy flavor fermions using matrix element techniques*, *Phys. Rev. D* **94** (2016) 055023 [[arXiv:1606.03107](#)] [[INSPIRE](#)].
- [16] P. De Castro and T. Dorigo, *INFERN0: Inference-Aware Neural Optimisation*, *Comput. Phys. Commun.* **244** (2019) 170 [[arXiv:1806.04743](#)] [[INSPIRE](#)].
- [17] S. Wunsch, S. Jörger, R. Wolf and G. Quast, *Optimal statistical inference in the presence of systematic uncertainties using neural network optimization based on binned Poisson likelihoods with nuisance parameters*, *Comput. Softw. Big Sci.* **5** (2021) 4 [[arXiv:2003.07186](#)] [[INSPIRE](#)].
- [18] K.T. Matchev and P. Shyamsundar, *ThickBrick: optimal event selection and categorization in high energy physics. Part I. Signal discovery*, *JHEP* **03** (2021) 291 [[arXiv:1911.12299](#)] [[INSPIRE](#)].
- [19] K. Kondo, *Dynamical Likelihood Method for Reconstruction of Events With Missing Momentum. 1: Method and Toy Models*, *J. Phys. Soc. Jap.* **57** (1988) 4126 [[INSPIRE](#)].
- [20] P. Artoisenet, V. Lemaitre, F. Maltoni and O. Mattelaer, *Automation of the matrix element reweighting method*, *JHEP* **12** (2010) 068 [[arXiv:1007.3300](#)] [[INSPIRE](#)].
- [21] F. Fiedler, A. Grohsjean, P. Haefner and P. Schieferdecker, *The Matrix Element Method and its Application in Measurements of the Top Quark Mass*, *Nucl. Instrum. Meth. A* **624** (2010) 203 [[arXiv:1003.1316](#)] [[INSPIRE](#)].
- [22] Y. Gao, A.V. Gritsan, Z. Guo, K. Melnikov, M. Schulze and N.V. Tran, *Spin Determination of Single-Produced Resonances at Hadron Colliders*, *Phys. Rev. D* **81** (2010) 075022 [[arXiv:1001.3396](#)] [[INSPIRE](#)].
- [23] J.M. Campbell, W.T. Giele and C. Williams, *The Matrix Element Method at Next-to-Leading Order*, *JHEP* **11** (2012) 043 [[arXiv:1204.4424](#)] [[INSPIRE](#)].
- [24] T. Martini and P. Uwer, *Extending the Matrix Element Method beyond the Born approximation: Calculating event weights at next-to-leading order accuracy*, *JHEP* **09** (2015) 083 [[arXiv:1506.08798](#)] [[INSPIRE](#)].
- [25] T. Martini and P. Uwer, *The Matrix Element Method at next-to-leading order QCD for hadronic collisions: Single top-quark production at the LHC as an example application*, *JHEP* **05** (2018) 141 [[arXiv:1712.04527](#)] [[INSPIRE](#)].
- [26] J. Brehmer, K. Cranmer, G. Louppe and J. Pavez, *Constraining Effective Field Theories with Machine Learning*, *Phys. Rev. Lett.* **121** (2018) 111801 [[arXiv:1805.00013](#)] [[INSPIRE](#)].
- [27] D.E. Soper and M. Spannowsky, *Finding physics signals with event deconstruction*, *Phys. Rev. D* **89** (2014) 094005 [[arXiv:1402.1189](#)] [[INSPIRE](#)].
- [28] S. Prestel and M. Spannowsky, *HYTREES: Combining Matrix Elements and Parton Shower for Hypothesis Testing*, *Eur. Phys. J. C* **79** (2019) 546 [[arXiv:1901.11035](#)] [[INSPIRE](#)].
- [29] K. Cranmer, J. Pavez and G. Louppe, *Approximating Likelihood Ratios with Calibrated Discriminative Classifiers*, [arXiv:1506.02169](#) [[INSPIRE](#)].

- [30] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski and D. Whiteson, *Parameterized neural networks for high-energy physics*, *Eur. Phys. J. C* **76** (2016) 235 [[arXiv:1601.07913](#)] [[INSPIRE](#)].
- [31] M. Stoye, J. Brehmer, G. Louppe, J. Pavez and K. Cranmer, *Likelihood-free inference with an improved cross-entropy estimator*, [arXiv:1808.00973](#) [[INSPIRE](#)].
- [32] J. Brehmer, G. Louppe, J. Pavez and K. Cranmer, *Mining gold from implicit models to improve likelihood-free inference*, *Proc. Nat. Acad. Sci.* **117** (2020) 5242 [[arXiv:1805.12244](#)] [[INSPIRE](#)].
- [33] J. Brehmer, K. Cranmer, G. Louppe and J. Pavez, *A Guide to Constraining Effective Field Theories with Machine Learning*, *Phys. Rev. D* **98** (2018) 052004 [[arXiv:1805.00020](#)] [[INSPIRE](#)].
- [34] J. Brehmer, F. Kling, I. Espejo and K. Cranmer, *MadMiner: Machine learning-based inference for particle physics*, *Comput. Softw. Big Sci.* **4** (2020) 3 [[arXiv:1907.10621](#)] [[INSPIRE](#)].
- [35] J. Neyman and E.S. Pearson, *On the Problem of the Most Efficient Tests of Statistical Hypotheses*, *Phil. Trans. Roy. Soc. Lond. A* **231** (1933) 289.
- [36] PARTICLE DATA GROUP collaboration, *Review of Particle Physics*, *Phys. Rev. D* **98** (2018) 030001 [[INSPIRE](#)].
- [37] A. Falkowski, M. Gonzalez-Alonso, A. Greljo and D. Marzocca, *Global constraints on anomalous triple gauge couplings in effective field theory approach*, *Phys. Rev. Lett.* **116** (2016) 011801 [[arXiv:1508.00581](#)] [[INSPIRE](#)].
- [38] D.R. Green, P. Meade and M.-A. Pleier, *Multiboson interactions at the LHC*, *Rev. Mod. Phys.* **89** (2017) 035008 [[arXiv:1610.07572](#)] [[INSPIRE](#)].
- [39] A. Butter, O.J.P. Éboli, J. Gonzalez-Fraile, M.C. Gonzalez-Garcia, T. Plehn and M. Rauch, *The Gauge-Higgs Legacy of the LHC Run I*, *JHEP* **07** (2016) 152 [[arXiv:1604.03105](#)] [[INSPIRE](#)].
- [40] R. Franceschini, G. Panico, A. Pomarol, F. Riva and A. Wulzer, *Electroweak Precision Tests in High-Energy Diboson Processes*, *JHEP* **02** (2018) 111 [[arXiv:1712.01310](#)] [[INSPIRE](#)].
- [41] G. Panico, F. Riva and A. Wulzer, *Diboson interference resurrection*, *Phys. Lett. B* **776** (2018) 473 [[arXiv:1708.07823](#)] [[INSPIRE](#)].
- [42] A. Azatov, J. Elias-Miro, Y. Reyimuaji and E. Venturini, *Novel measurements of anomalous triple gauge couplings for the LHC*, *JHEP* **10** (2017) 027 [[arXiv:1707.08060](#)] [[INSPIRE](#)].
- [43] A. Azatov, D. Barducci and E. Venturini, *Precision diboson measurements at hadron colliders*, *JHEP* **04** (2019) 075 [[arXiv:1901.04821](#)] [[INSPIRE](#)].
- [44] J. Baglio, S. Dawson and S. Homiller, *QCD corrections in Standard Model EFT fits to WZ and WW production*, *Phys. Rev. D* **100** (2019) 113010 [[arXiv:1909.11576](#)] [[INSPIRE](#)].
- [45] M.J. Duncan, G.L. Kane and W.W. Repko, *A New Standard Model Test for Future Colliders*, *Phys. Rev. Lett.* **55** (1985) 773 [[INSPIRE](#)].
- [46] K. Hagiwara, R.D. Peccei, D. Zeppenfeld and K. Hikasa, *Probing the Weak Boson Sector in  $e^+e^- \rightarrow W^+W^-$* , *Nucl. Phys. B* **282** (1987) 253 [[INSPIRE](#)].
- [47] G. Cowan, *Statistical data analysis*, Oxford University Press, U.S.A. (1998).

- [48] M. Anthony and P.L. Bartlett, *Neural Network Learning: Theoretical Foundations*, Cambridge University Press (1999) [DOI].
- [49] R.T. D’Agnolo and A. Wulzer, *Learning New Physics from a Machine*, *Phys. Rev. D* **99** (2019) 015014 [arXiv:1806.02350] [INSPIRE].
- [50] D. Liu, A. Pomarol, R. Rattazzi and F. Riva, *Patterns of Strong Coupling for LHC Searches*, *JHEP* **11** (2016) 141 [arXiv:1603.03064] [INSPIRE].
- [51] G. Cuomo, L. Vecchi and A. Wulzer, *Goldstone Equivalence and High Energy Electroweak Physics*, *SciPost Phys.* **8** (2020) 078 [arXiv:1911.12366] [INSPIRE].
- [52] A. Kusina et al., *nCTEQ15 — Global analysis of nuclear parton distributions with uncertainties*, *PoS DIS2015* (2015) 041 [arXiv:1509.01801] [INSPIRE].
- [53] D.B. Clark, E. Godat and F.I. Olness, *ManeParse: A Mathematica reader for Parton Distribution Functions*, *Comput. Phys. Commun.* **216** (2017) 126 [arXiv:1605.08012] [INSPIRE].
- [54] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, *JHEP* **07** (2014) 079 [arXiv:1405.0301] [INSPIRE].
- [55] C. Degrande, G. Durieux, F. Maltoni, K. Mimasu, E. Vryonidou and C. Zhang, *Automated one-loop computations in the SMEFT*, arXiv:2008.11743 [INSPIRE] [<https://feynrules.irmp.ucl.ac.be/wiki/SMEFTatNLO>].
- [56] T. Sjöstrand, S. Mrenna and P.Z. Skands, *PYTHIA 6.4 Physics and Manual*, *JHEP* **05** (2006) 026 [hep-ph/0603175] [INSPIRE].
- [57] T. Sjöstrand, S. Mrenna and P.Z. Skands, *A Brief Introduction to PYTHIA 8.1*, *Comput. Phys. Commun.* **178** (2008) 852 [arXiv:0710.3820] [INSPIRE].
- [58] E. Conte, B. Fuks and G. Serret, *MadAnalysis 5, A User-Friendly Framework for Collider Phenomenology*, *Comput. Phys. Commun.* **184** (2013) 222 [arXiv:1206.1599] [INSPIRE].
- [59] U. Baur, T. Han and J. Ohnemus, *Amplitude zeros in  $W^\pm Z$  production*, *Phys. Rev. Lett.* **72** (1994) 3941 [hep-ph/9403248] [INSPIRE].
- [60] L.J. Dixon and Y. Shadmi, *Testing gluon selfinteractions in three jet events at hadron colliders*, *Nucl. Phys. B* **423** (1994) 3 [Erratum *ibid.* **452** (1995) 724] [hep-ph/9312363] [INSPIRE].
- [61] *Poisson distribution*, in *Encyclopedia of Mathematics*, [https://encyclopediaofmath.org/wiki/Poisson\\_distribution](https://encyclopediaofmath.org/wiki/Poisson_distribution).
- [62] A. Paszke et al., *Pytorch: An imperative style, high-performance deep learning library*, in *Advances in Neural Information Processing Systems 32*, H. Wallach et al. eds., pp. 8024–8035, Curran Associates, Inc. (2019) [<http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>] [arXiv:1912.01703].
- [63] DELPHES 3 collaboration, *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*, *JHEP* **02** (2014) 057 [arXiv:1307.6346] [INSPIRE].