

Searching for HWW anomalous couplings with simulation-based inference

Marta Silva^{*}

*Nikhef National Institute for Subatomic Physics, Amsterdam, The Netherlands
and Laboratory of Instrumentation and Experimental Particle Physics, Lisbon, Portugal*

Ricardo Barru [†]

*Marietta Blau Institute for Particle Physics, Vienna, Austria; Instituto Superior T cnico, Lisbon, Portugal
and Laboratory of Instrumentation and Experimental Particle Physics, Lisbon, Portugal*

In s Ochoa[‡]

Laboratory of Instrumentation and Experimental Particle Physics, Lisbon, Portugal

Patricia Conde Mu  o[ ]

*Instituto Superior T cnico, Lisbon, Portugal;
and Laboratory of Instrumentation and Experimental Particle Physics, Lisbon, Portugal*



(Received 20 November 2025; accepted 11 February 2026; published 9 March 2026)

Understanding the source of the universe’s asymmetry between matter and antimatter is one of the major open questions in particle physics. In this work, the sensitivity of novel machine-learning-based inference techniques to CP -odd and CP -even HWW anomalous couplings is studied in the $WH \rightarrow \ell \nu b \bar{b}$ channel ($\ell = e, \mu$), within the standard model effective field theory framework. Two machine-learning simulation-based inference (SBI) methods are explored: a per-event likelihood-ratio estimator, which directly approximates the ratio of probability densities between competing hypotheses, is benchmarked against a per-event optimal-observable estimator optimized for sensitivity to the parameters of interest. Both approaches are also compared to traditional summary statistics, in this case histograms of kinematic and angular observables, as commonly used in experimental analyses. SBI methods provide tighter constraints than one-dimensional summary statistics, though their performance is comparable to two-dimensional histogram analysis. The optimal-observable approach remains promising for its ability to probe multiple couplings simultaneously. Restricting the analysis to a region of high S/B also enhances sensitivity to CP -odd operators while preserving sensitivity to CP -even operators, which histogram analyses often lose. Although the likelihood-ratio estimator sometimes struggles with likelihood minima and shapes, optimizations that target its robustness could make it more sensitive than both the optimal-observable estimator and the histogram method. These results underscore the potential of advanced simulation-based inference techniques, encouraging further exploration with LHC Run 3 data to surpass current ATLAS and CMS sensitivities.

DOI: [10.1103/29xq-b66r](https://doi.org/10.1103/29xq-b66r)

I. INTRODUCTION

One of the major questions left unaddressed by the Standard Model (SM) is the observed asymmetry between matter and antimatter in the Universe. Sources of

charge-parity (CP) symmetry violation in the SM are insufficient to explain this imbalance [1] and CP violation in the Higgs sector is a well-motivated way to address it. In fact, since the discovery of the Higgs boson, one of the main goals of the Large Hadron Collider (LHC) has been to precisely measure its properties.

Most analyses in experimental particle physics share a common central question: *Based on observed data, what is the most likely value of the physical parameters that describes the underlying process?* This is known as statistical inference and its fundamental key is the likelihood function. However, in particle physics this quantity is typically intractable as it involves integrating over millions of random variables [2]. Performing statistical

^{*}Contact author: femarta@cern.ch

[†]Contact author: ricardo.barrue@cern.ch

[‡]Contact author: miochoa@lip.pt

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](https://creativecommons.org/licenses/by/4.0/) license. Further distribution of this work must maintain attribution to the author(s) and the published article’s title, journal citation, and DOI. Funded by SCOAP³.

inference when the data-generating process does not have a tractable likelihood is a major challenge common to many fields, and the methods that address this intractability using simulators of the underlying physical process are collectively referred to as simulation-based inference techniques.

The traditional approach to search for beyond the SM (BSM) physics, such as anomalous couplings in the Higgs sector, is to reduce the dimensionality of the probability density function by filling histograms or templates from Monte Carlo (MC) simulations of one or two discriminating observables (summary statistics). Common examples are angular variables which are used, e.g., by the ATLAS Collaboration to constrain the presence of CP -violating components in Higgs and EW boson interactions via the $H \rightarrow WW^*$ [3] decay or in WH [4] production. However, while this is straightforward and computationally efficient, it leads to an inevitable loss of information about the process and underlying physical parameters being studied [5,6].

Other widely used methods, such as the Matrix Element Method (MEM) [7] or Optimal Observables [8], used by ATLAS and CMS e.g., in measurements of vector boson fusion (VBF) production of $H \rightarrow 4\ell$ [9,10] and $H \rightarrow \tau\tau$ [11,12] leverage the fact that some parts of the likelihood function can be computed, particularly through matrix element calculations, to compress the information from multiple kinematic variables with minimal losses in sensitivity. Nevertheless, these methods rely on approximations of the parton shower, hadronization, and detector interactions, thereby simplifying the typical pp collision simulation chain.

More recently, machine learning (ML) methods have been recognized as valuable tools for handling this problem without relying on simplifications or compromising statistical power. In particular, novel ML-inference techniques, introduced in Refs. [13–18], exploit additional information from MC simulations in order to train neural networks (NN) that can effectively estimate quantities relevant to statistical inference. An example is the likelihood ratio which, according to the Neyman-Pearson lemma, is the optimal test statistic for simple-vs-simple hypothesis testing [19].

In this work, we study the sensitivity of these innovative techniques to HWW anomalous couplings with the main goal of designing the best strategies to employ in the analysis of LHC Run 3 data and beyond, highlighting areas of potential future improvements. We focus in particular on the WH production channel, with the W boson decaying to leptons and the Higgs boson to a pair of b quarks, with anomalous couplings parametrized with the Standard Model effective field theory (SMEFT) [20] framework. In particular, we consider both CP -violating and CP -conserving contributions to the Lagrangian, in order to disentangle their contributions and provide a more complete picture of anomalous couplings to this vertex. Kinematic and angular

observables commonly used in experimental analyses are benchmarked against three ML-based methods described in the literature, SALLY [14], ALICE, and ALICES [21]. Existing work on this topic includes a study of the SALLY method on the same final state, but focusing solely on the CP -violating coupling [22], and a study of SALLY for targeting only CP -even couplings [23]. Additionally, a practical application of SBI techniques has been carried out by the CMS Collaboration in WH and ZH production [24], targeting multiple SMEFT operators. Finally, an implementation of ML-based SBI has been performed by the ATLAS Collaboration, using an off-shell Higgs boson couplings measurement as an example [25]. Ultimately, our work aims to inform analysis strategies in order to extract the best sensitivity to BSM physics from LHC data, using novel but increasingly robust and widespread SBI approaches.

The remainder of this document is structured as follows: Section II introduces simulation-based inference and the methods employed in this work. Section III outlines the effective field theory framework and the anomalous couplings studied, as well as the details of the analysis and event generation. Section IV presents the results of all the scenarios studied and, finally, in Sec. V, the main conclusions of this work are drawn.

II. SIMULATION-BASED INFERENCE

A. The likelihood-free inference problem

The central quantity in any particle physics analysis is the likelihood function $p_{\text{full}}(x|\theta)$, which quantifies the probability of observing a set of events, each characterized by a vector of observables x as a function of the parameters of interest θ . In this work, θ will parametrize the effect of the anomalous couplings, which will be introduced in Sec. III. The likelihood function is usually expressed as

$$p_{\text{full}}(x|\theta) = \text{Pois}(n|L\sigma(\theta)) \prod_i p(x_i|\theta), \quad (1)$$

where n represents the observed number of events, L the integrated luminosity, $\sigma(\theta)$ the cross section, $\text{Pois}(n|\lambda) = \lambda^n e^{-\lambda}/n!$ the probability mass function of the Poisson distribution and $p(x_i|\theta)$ is the likelihood that an event i with reconstruction-level kinematics x_i is described by a certain value of θ (henceforth simply called the *likelihood*).

Another useful quantity is the score $t(x|\theta)$, which is the first order term of the log-likelihood Taylor expansion. Evaluated at a reference point θ_{ref} , it is a sufficient statistic of the likelihood, i.e., a statistically optimal observable around that reference point, and is defined as

$$t(x|\theta)_{\theta_{\text{ref}}} = \nabla_{\theta} \log p(x|\theta)|_{\theta_{\text{ref}}}. \quad (2)$$

Particle physicists use simulations to generate “synthetic” observations according to a given choice of θ

and compare them to data in order to perform simulation-based inference. Since the MC simulation chain factorizes into the hard-scatter interaction, parton shower and hadronization, and detector simulation, the likelihood can be written as follows:

$$p(x|\theta) = \int dz_d \int dz_s \int dz_p \times \underbrace{p(x|z_d)p(z_d|z_s)p(z_s|z_p)p(z_p|\theta)}_{p(x,z|\theta)}, \quad (3)$$

where

- (i) z_p , z_s , and z_d are the latent variables describing the parton-level event (i.e., the parton-level kinematics), the parton shower and hadronization, and the interactions with the detector and event reconstruction, respectively. These variables cannot be directly measured in experiments.
- (ii) $p(x, z|\theta)$ is the joint likelihood, jointly conditioned on the reconstruction-level variables x and the latent variables of the event, $z = (z_p, z_s, z_d)$.
- (iii) $p(z_s|z_p)$, $p(z_d|z_s)$, and $p(x|z_d)$ describe the probabilistic evolution from $z_p \rightarrow x$ and $p(z_p|\theta)$ is the parton-level likelihood, given by

$$p(z_p|\theta) = \frac{1}{\sigma(\theta)} \frac{d\sigma(\theta)}{dz_p}, \quad (4)$$

where $\sigma(\theta)$ and $\frac{d\sigma(\theta)}{dz_p}$ are the total and the differential cross sections, respectively.

Since $\sigma(\theta)$ and $\frac{d\sigma(\theta)}{dz_p}$ can be extracted from event generators for an arbitrary z and θ , both the Poisson term and $p(z_p|\theta)$ can be evaluated. However, the simulation of a single event can easily involve many millions of random latent variables. While $p(x|\theta)$ can be sampled ($x \sim p(x|\theta)$), calculating the likelihood (or any function of it, such as the score) would require integrating over all the possible histories leading to that observation ($z \rightarrow x$) [13,26]. In other words, its calculation is intractable. Historically, analyses have circumvented this problem by restricting themselves to histograms of one or two physics-motivated variables, thereby reducing their sensitivity. Alternatively, they have used matrix element calculators to build optimal observables, which compress the kinematic information in the event into a single number, but at the cost of having to approximate the simulation chain with a suitable transfer function $p(x|z_p)$, or ignoring it altogether. Examples include the CP -mixing angle used by ATLAS and CMS to study the $H\tau\tau$ coupling [27,28]; an optimal observable defined as the ratio of the matrix element interference term to the SM contribution, used by ATLAS to probe the HVV vertex in $H \rightarrow 4\ell$ [9]; and a similar matrix element likelihood approach (MELA) adopted by CMS for the same process [29].

ML-based inference methods have been introduced [30] to estimate likelihood ratios using binary classifiers, with the main drawback that they are quite computationally intensive. References [13–15] introduced the idea of using augmented data extracted from the simulator as a more efficient approach.

B. Aiding the inference with augmented data

Even though the likelihood ratio cannot be directly evaluated, the joint likelihood ratio can be calculated since the intractable parts corresponding to the showering and detector effects cancel out,

$$\begin{aligned} r(x, z|\theta_0, \theta_1) &\equiv \frac{p(x, z|\theta_0)}{p(x, z|\theta_1)} \\ &= \frac{p(x|z_d)p(z_d|z_s)p(z_s|z_p)p(z_p|\theta_0)}{p(x|z_d)p(z_d|z_s)p(z_s|z_p)p(z_p|\theta_1)} \\ &= \frac{p(z_p|\theta_0)}{p(z_p|\theta_1)} \\ &= \frac{d\sigma(z_p|\theta_0)}{d\sigma(z_p|\theta_1)} \frac{\sigma(\theta_1)}{\sigma(\theta_0)}, \end{aligned} \quad (5)$$

where $d\sigma(z_p|\theta)$ are the parton-level event weights from the event generator. The parameters θ_0 and θ_1 correspond to the BSM and SM hypotheses, respectively.

Analogously, the joint score can be calculated as follows:

$$\begin{aligned} t(x, z|\theta) &\equiv \nabla_\theta \log p(x, z|\theta) \\ &= \frac{p(x|z_d)p(z_d|z_s)p(z_s|z_p)\nabla_\theta p(z_p|\theta)}{p(x|z_d)p(z_d|z_s)p(z_s|z_p)p(z_p|\theta)} \\ &= \frac{\nabla_\theta d\sigma(z_p|\theta)}{d\sigma(z_p|\theta)} - \frac{\nabla_\theta \sigma(\theta)}{\sigma(\theta)}. \end{aligned} \quad (6)$$

The key result derived in [14,15] is that the joint likelihood ratio and joint score can be used to define loss functions that, when minimized, converge to the true likelihood ratio or the true score. This minimization procedure can be achieved by NNs trained with back-propagation and gradient descent. Therefore, a NN can be trained with a suitable loss function to regress the likelihood ratio or score.

There are several variations of this idea, differing mainly in the exact formulation of the employed loss function. The simple but effective score approximates likelihood locally (SALLY) method—a detector-level optimal observable, minimizes the mean squared error (MSE) with respect to the joint score to estimate the score

$$L_{\text{SALLY}}[\hat{t}(x)] = -\frac{1}{N} \sum_{(x_i, z_i) \sim p(x, z|\theta_{\text{ref}})} |t(x, z|\theta_{\text{ref}}) - \hat{t}(x|\theta_{\text{ref}})|^2. \quad (7)$$

Since this technique is trained to regress the score, $t(x|\theta_{\text{ref}})$ at the reference point θ_{ref} , it is only locally optimal. Furthermore, the SALLY output can be treated like any other observable for performing a likelihood fit. Therefore, this inference method is essentially an estimator of an optimal observable which instead of approximating the generation process via a transfer function, learns it implicitly.

$$L_{\text{ALICE}}[\hat{s}(x|\theta_0, \theta_1)] = -\frac{1}{N} \sum_{(x_i, z_i, \theta_{\text{choice}})} [s(x_i, z_i|\theta_0, \theta_1) \log(\hat{s}(x_i)) + (1 - s(x_i, z_i|\theta_0, \theta_1)) \log(1 - \hat{s}(x_i))]. \quad (8)$$

Events are drawn according to $\theta_{\text{choice}} = \theta_1, \theta_0$, i.e., sampling from the SM signal or BSM signal datasets (where the chosen BSM dataset is that with $\theta_{\text{benchmark}}$ closest to the value of θ_0 being sampled).

In the above formula, s is the boundary classification function modeled by a classifier trained to distinguish between two sets of observations sampled from $x \sim p(x|\theta_0)$

Another method is approximate likelihood with improved cross-entropy estimator (ALICE) [21], a per-event surrogate model of the likelihood ratio. It uses the joint likelihood ratio and takes as inputs theory parameters at two different values, in addition to the observables x . Its loss functional is given by

and $x \sim p(x|\theta_1)$, and which can be turned into an estimator for the likelihood ratio via the likelihood-ratio trick [31].

Finally, the approximate likelihood with improved cross-entropy estimator and score (ALICES) method is explored. It is an extension of L_{ALICE} since it uses not only the joint likelihood ratio but also leverages the joint score to guide the estimator. The loss function is therefore extended as

$$L_{\text{ALICES}}[\hat{s}(x|\theta_0, \theta_1)] = L_{\text{ALICE}} - \frac{1}{N} \sum_{(x_i, z_i, \theta_{\text{choice}})} \alpha(1 - y_i) \left| t(x_i, z_i|\theta_0, \theta_1) - \nabla_{\theta} \log \left(\frac{1 - \hat{s}(x_i|\theta, \theta_1)}{\hat{s}(x_i|\theta, \theta_1)} \right) \right|_{\theta_0}^2, \quad (9)$$

where α is a hyperparameter that weights the joint likelihood ratio and joint score terms in the loss function, y is a label corresponding to $y_i = 0$ for samples drawn from the BSM hypothesis, $(x_i, z_i) \sim p(x, z|\theta_0)$, and $y_i = 1$ for samples drawn from the SM hypothesis, $(x_i, z_i) \sim p(x, z|\theta_1)$. The ALICES method converges to the ALICE technique when $\alpha = 0$ [21].

The ALICE/ALICES methods are expected to exhibit superior performance to SALLY as they perform an estimation of the likelihood ratio rather than its local approximation. Since they do not rely on the assumption that the parameter θ is close to the SM and are parametrized on θ , they are optimal over the entire parameter space. Both the SALLY and ALICE(S) methods are robust against reducible and irreducible backgrounds (which have fixed values of the joint quantities) and to unobservable degrees of freedom, such as the longitudinal momentum of the neutrino in the W decay.

These simulation-based inference techniques based on data augmentation are implemented in a Python package called MadMiner [32], which was used extensively for this work.

III. CONSTRAINING HWW ANOMALOUS COUPLINGS

A. Current state of the art

The SM can be regarded as a low-energy effective theory of an underlying and unknown theory, valid up to an energy

scale Λ (also called the New Physics scale) beyond the reach of current colliders. In this work, such an effective theory is written using the SMEFT approach, which extends the SM Lagrangian with additional operators of mass dimension $d > 4$, constructed from combinations of SM fields invariant under SM $SU(3)_C \times SU(2)_L \times U(1)_Y$ transformations

$$\mathcal{L}_{\text{SMEFT}} = \mathcal{L}_{\text{SM}} + \sum_{d>4} \sum_i c_i \frac{\mathcal{O}_i^{(d)}}{\Lambda^{d-4}}, \quad (10)$$

where c_i is called the Wilson coefficient, which parametrizes the relative strength of an operator.

This study focuses on dimension-6 operators. The only dimension-5 operator, the Weinberg operator, is related to neutrino masses and violates lepton number conservation, and is therefore not considered. Using the Warsaw basis [33], two dimension-6 operators are pertinent for this work: one CP odd, which changes sign under a CP transformation, and a CP even whose sign is not altered under the same transformation. These operators are written as follows:

$$\tilde{\mathcal{O}}_{HW} = H^\dagger H \epsilon_{\mu\nu\rho\sigma} W^{I\rho\sigma} W^{I\mu\nu} \quad (CP \text{ odd}), \quad (11)$$

$$\mathcal{O}_{HW} = H^\dagger H W_{\mu\nu}^I W^{I\mu\nu} \quad (CP \text{ even}), \quad (12)$$

where H is the $SU(2)_L$ Higgs doublet, $W_{\mu\nu}^a$ the field strength tensor and $\epsilon^{\mu\nu\rho\theta}$ is the antisymmetric Levi-Civita tensor.

TABLE I. Summary of the most stringent to date expected 68% confidence intervals from ATLAS and CMS for the CP -odd $c_{H\tilde{W}}$ and CP -even c_{HW} Wilson coefficients in the Warsaw basis of SMEFT. The results presented consider the linear plus quadratic effects of new physics.

Wilson coefficient	ATLAS (68% CL)	CMS (68% CL)
$c_{H\tilde{W}}$	$[-0.22, 0.22]$ [11]	$[-1.11, 1.11]$ [10]
c_{HW}	$[-0.0072, 0.0072]$ [34]	$[-0.28, 0.39]$ [10]

$c_{H\tilde{W}}$ and c_{HW} are the corresponding Wilson coefficients of the two operators and Λ is assumed to be 1 TeV.

The two operators contribute to the SMEFT amplitude via interference terms with the SM, linearly proportional to c_i/Λ^2 , and via quadratic terms, proportional to c_i^2/Λ^4 . It is worth pointing out that only the linear term of the CP -odd operator is CP odd. Other higher-order operators can also contribute at the order of $1/\Lambda^4$, but are not considered here for simplicity. Since the true UV theory parametrized by SMEFT is unknown, these coefficients (more particularly, the ratio c_i/Λ^2) are free parameters and need to be constrained experimentally.

In this work, different ML-based inference techniques are employed to probe the sensitivity to c_{HW} and $c_{H\tilde{W}}$, aiming to improve the state-of-the-art constraints by ATLAS and CMS, summarized in Table I. The CMS results shown correspond to those available in the Warsaw basis. Other analyses exist in different formalisms, such as the recent simulation-based inference study from CMS [24].

B. Overview of the $WH \rightarrow \ell \nu b \bar{b}$ channel

This analysis targets the associated WH production channel with the W boson decaying leptonically, $W \rightarrow \ell \nu$ ($\ell = e, \mu$) and the Higgs boson decaying to a pair of b quarks, $H \rightarrow b \bar{b}$. The final state, illustrated by the leading-order (LO) Feynman diagram in Fig. 1, is characterized by two jets originating from b quarks, missing transverse energy from the neutrino and either an electron or muon,

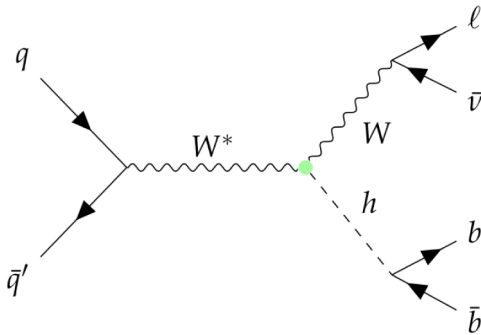
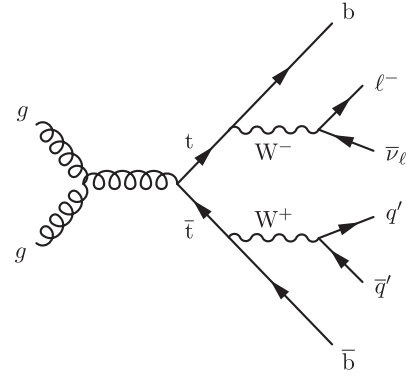
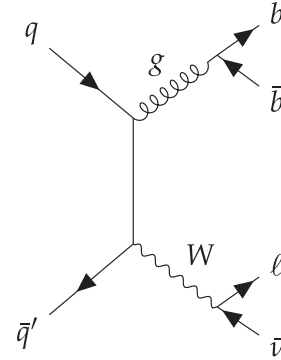


FIG. 1. Feynman diagram illustrating the LO WH associated production in the $\ell \nu b \bar{b}$ final state. The green circle corresponds to the vertex of interest.



(a) $t\bar{t}$



(b) $W + b\text{-jets}$

FIG. 2. Leading order Feynman diagram illustrating the most relevant backgrounds.

which enable effective triggering and the removal of most QCD multijet background.

The main background processes correspond to top-pair ($t\bar{t}$) production in the semileptonic decay channel and associated production of a W boson and b jets. Additional contributions from single top production in the s channel are also considered. The leading-order Feynman diagrams illustrating the most relevant background processes are shown in Fig. 2.

C. Energy-dependent and angular observables

The BSM quadratic components in the amplitude typically lead to more pronounced modifications in the distributions of observables that depend on the partonic center of mass energy. This effect is evident in the distributions of the total transverse mass of the event, $m_T^{\ell \nu b \bar{b}}$, and the transverse momentum of the W boson, p_T^W [22,23]. As illustrated in Fig. 3 for $m_T^{\ell \nu b \bar{b}}$, there is an increase in the signal-to-background ratio in the high-energy region compared to the SM prediction, indicating that these observables are sensitive to $c_{HW} \neq 0$. The same is also true for $c_{H\tilde{W}} \neq 0$.

However, since these modifications arise primarily from the quadratic term in the amplitude, there is limited sensitivity to the sign of the Wilson coefficients, especially

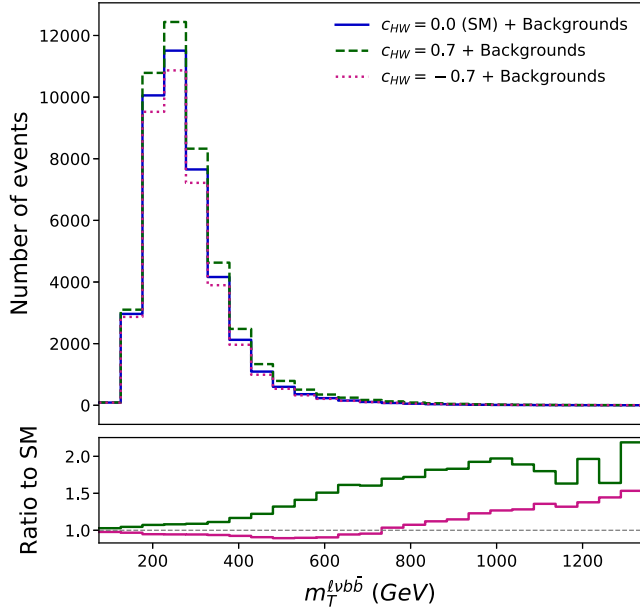


FIG. 3. Distributions of the transverse mass of the $\ell\nu b\bar{b}$ system for the SM signal plus backgrounds (blue), compared to alternative BSM signal samples plus backgrounds, with $c_{HW} = 0.7$ (dashed green) and $c_{HW} = -0.7$ (dotted pink), obtained using the morphing technique. The lower panel shows the ratio to the SM prediction. The samples and selection cuts are described in Sec. III E.

for the CP -odd operator. Therefore, two angular observables [22,35], particularly sensitive to the CP -odd interference component, were also considered

$$Q_\ell \cos \delta^+ = Q_\ell \times \frac{\mathbf{p}_\ell^{(W)} \cdot (\mathbf{p}_H \times \mathbf{p}_W)}{|\mathbf{p}_\ell^{(W)}| |\mathbf{p}_H \times \mathbf{p}_W|}, \quad (13)$$

$$Q_\ell \cos \delta^- = Q_\ell \times \frac{\mathbf{p}_W \cdot (\mathbf{p}_\ell^{(-)} \times \mathbf{p}_\nu^{(-)})}{|\mathbf{p}_W| |\mathbf{p}_\ell^{(-)} \times \mathbf{p}_\nu^{(-)}|}, \quad (14)$$

where $\mathbf{p}_W$ and $\mathbf{p}_H$ denote the momenta of the W and Higgs bosons, respectively, $\mathbf{p}_\ell^{(W)}$ the momentum of the charged lepton in the W boson rest frame, and $\mathbf{p}_\ell^{(-)}$ and $\mathbf{p}_\nu^{(-)}$ represent the momenta of the charged lepton and neutrino in the Higgs boson rest frame (with $\mathbf{p}_H \rightarrow -\mathbf{p}_H$). The longitudinal momentum of the neutrino, p_z^ν is reconstructed as described in Ref. [22]. The distributions of $Q_\ell \cos \delta^+$ are shown in Fig. 4 for SM and BSM signals plus backgrounds. It is clear that while the distribution remains symmetric around zero for the SM Higgs hypothesis, an asymmetry is introduced for values of $c_{H\bar{W}} \neq 0$, different for a positive or negative value. On the other hand, a positive (negative) value for the CP -even coefficient results in an increase (decrease) of the overall cross section but no change to the asymmetry of the distribution.

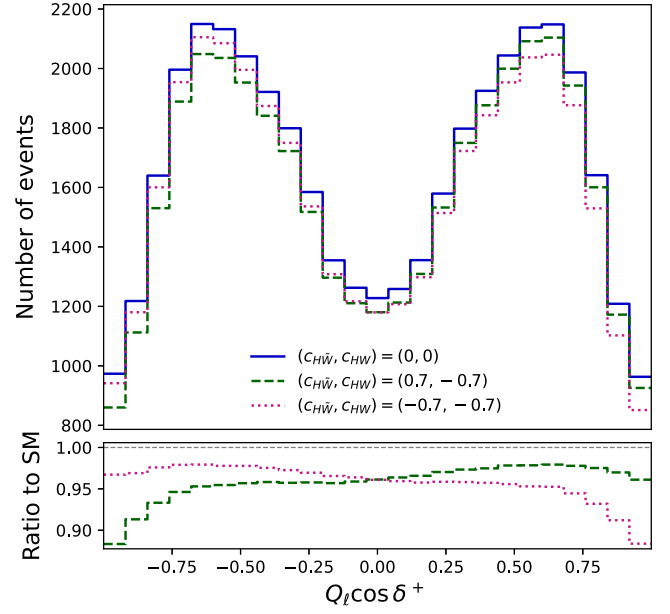


FIG. 4. Distributions of the angular variable $Q_\ell \cos \delta^+$ for the SM signal plus backgrounds (blue), compared to alternative BSM signal samples plus backgrounds, with $c_{H\bar{W}}, c_{HW} = (0.7, -0.7)$ (dashed green) and $c_{H\bar{W}}, c_{HW} = (-0.7, -0.7)$ (dotted pink), obtained using the morphing technique. The lower panel shows the ratio to the SM prediction. The samples and selection cuts are described in Sec. III E.

D. EFT scenarios studied

In this work, the sensitivity of the ALICES, ALICE, and SALLY methods to the CP -odd ($\tilde{\mathcal{O}}_{HW}$) and CP -even ($\mathcal{O}_{HW}$) dimension-6 operators was examined across three different scenarios, in order to evaluate how these inference techniques perform under distinct levels of complexity and sensitivity to BSM couplings.

1. Validation at parton level

These methods were first validated in a setup where the likelihood ratio and score can be precisely calculated. By omitting the effects of parton showers, parton distribution functions, and detector responses, all latent variables directly correspond to the observables (i.e., $x = z$). As a result, the likelihood ratio $r(x)$ becomes equivalent to the joint likelihood ratio $r(x, z)$, and the score $t(x)$ matches the joint score $t(x, z)$. Therefore, it is possible to directly compare the neural network predictions with the ground truth. In addition, for the neural network to effectively learn the matrix element, all initial and final state particles need to be fixed, so a simplified parton-level process was considered

$$u\bar{d} \rightarrow W^+ H \rightarrow \mu^+ \nu_\mu b\bar{b}. \quad (15)$$

In this scenario, only the $\tilde{\mathcal{O}}_{HW}$ operator was considered to assess the closure of the methods.

2. Simultaneous 2D fits

Following the parton-level validation, a more realistic phenomenological analysis is performed by incorporating parton shower and detector effects. Specifically, the aforementioned simulation-based inference methods were trained with CP -odd and CP -even coefficients as inputs, in order to investigate their sensitivity to both operators simultaneously. Two kinematic regimes were considered:

- (1) *Inclusive* p_T^W . Using the full simulated event sample without additional kinematic cuts.
- (2) *High- p_T^W* . Selecting events with $p_T^W > 150$ GeV, to probe the enhanced sensitivity to BSM couplings in the high p_T regime and to benefit from a higher signal-to-background ratio in this phase-space region.

For comparison, a set of baseline results were established by performing histogram-based fits using energy-dependent and angular observables, namely: a 1D fit to $m_T^{\ell\nu b\bar{b}}$, a 1D fit to $Q_\ell \cos \delta^+$, and a 2D fit to $Q_\ell \cos \delta^+ \otimes p_T^W$.

All three scenarios were conducted using MadMiner v0.9.3 [32], with distinct simulation setups for each, as detailed in the next sections.

E. Monte Carlo samples and event selection

The signal and background samples are generated following Ref. [22]. Signal samples were simulated at 13 TeV with MadGraph_aMC@NLO [36] at leading order¹ (LO) in QCD, employing the PDF4LHC15 PDF set. Modifications to the interaction vertex were incorporated using the SMEFT_{sim3} [38] UFO model, with the new physics scale set to $\Lambda = 1$ TeV and assuming $U(3)^5$ flavor symmetry.

Signal events were generated at the SM point $c_{H\bar{W}}, c_{HW} = (0, 0)$ and reweighted to benchmark BSM points, optimized with the MadMiner morphing functionality. Table II summarizes the chosen benchmark points (morphing basis) for all scenarios considered. The Wilson coefficients were limited to a range of $|c_{H\bar{W}}| < 1.2$ and $|c_{HW}| < 1.0$, slightly less stringent than the current experimental constraints.

To mitigate the large statistical fluctuations that can arise from reweighting events to distant points in parameter space, one-fifth of the signal samples were generated directly at the BSM points [32].

The background events were generated at LO in QCD using the default MadGraph_aMC@NLO SM UFO model. No reweighting or morphing was applied to these samples since the CP -odd and CP -even operators do not affect the background processes. The total number of events generated for each scenario is detailed in Table III.

¹Next-to-leading order corrections for CP -odd SMEFT operators in VH production have been studied in Ref. [37], but these are not widely available in event generation tools yet, and are therefore not included in this work.

TABLE II. Optimized morphing basis points for the validation and 2D fit scenarios. For the 2D case, the same points were used for both the inclusive and high- p_T^W studies.

Coefficient	Validation	2D	
	$c_{H\bar{W}}$	$c_{H\bar{W}}$	c_{HW}
Benchmark 1	1.15	−0.90	0.42
Benchmark 2	−1.04	−0.23	0.97
Benchmark 3	...	−1.12	−0.76
Benchmark 4	...	0.72	−0.87
Benchmark 5	...	1.15	0.63

TABLE III. Number of generated events for the validation and 2D fit scenarios. For the 2D case, the same event samples were used for both the inclusive and high- p_T^W studies.

	Validation	2D
SM signal	1.0×10^6	8.0×10^6
Backgrounds	...	12.0×10^6
BSM signal	0.2×10^6	8.0×10^6
Total	1.2×10^6	28.0×10^6

A set of generator-level selection criteria, originally detailed in Refs. [22,23] was employed, defining the *inclusive* sample. Additionally, *high- p_T^W* samples were generated with an extra generator-level cut of $p_T^W > 125$ GeV. This requirement corresponds to cumulative efficiencies of 8.93% for the SM signal, and 0.07%, 0.04%, and 1.06% for the $t\bar{t}$, W +(b)-jets, and single top backgrounds, respectively. Efficiencies for the other generator-level requirements are reported in Ref. [22].

Parton showering and hadronization were simulated using Pythia8 [39] with the A14 tune and MLM merging technique, with a merging scale of $Q_{\text{cut}} = 20$ GeV. Detector effects and object reconstruction were simulated using Delphes [40], with the ATLAS default card. For both the inclusive and high- p_T^W sample, events with at least two reconstructed anti- k_t jets with radius parameter of 0.6 and $p_T > 20$ GeV are considered, followed by a requirement that the two leading jets are b -tagged with a 70% efficiency working point, retaining 24% of the total S + B events. For the high- p_T^W sample, an additional cut of $p_T^W > 150$ GeV was applied at reconstruction level, after the inclusive selection, resulting in a cumulative retention of 10% of the generated S + B events.

The generated cross sections at different selection stages, namely before any cuts, after generator-level cuts, and after reconstruction-level cuts, together with the corresponding expected event yields for a luminosity of 300 fb^{-1} , are presented in Tables IV and V for the WH signal and the main background processes, for the inclusive and high- p_T^W samples, respectively.

TABLE IV. Cross sections at different selection stages and final expected event yields for a luminosity of 300 fb^{-1} , for the inclusive p_T^W sample.

Stage	Signal (SM)	$t\bar{t}$	W + jets	Single top	Total
No cuts [fb]	214.00	69310.00	52340.00	666.00	...
After gen-level cuts [fb]	73.16	194.24	239.50	75.74	...
After reco-level cuts [fb]	18.29	36.97	63.04	19.54	...
Final events @ 300 fb^{-1}	5485.83	11092.41	18911.46	5861.14	41350.84

TABLE V. Cross sections at different selection stages and final expected event yields for a luminosity of 300 fb^{-1} , for the high- p_T^W sample.

Stage	Signal (SM)	$t\bar{t}$	W + jets	Single top	Total
No cuts [fb]	214.00	69310.00	52340.00	666.00	...
After gen-level cuts [fb]	19.07	49.57	22.15	7.06	...
After reco-level cuts [fb]	3.11	1.98	3.31	0.30	...
Final events @ 300 fb^{-1}	934.30	593.92	993.98	91.31	2613.51

F. Sample unweighting and augmentation

The loss functions described in Sec. II B are not well suited for weighted data. Therefore, an unweighted training sample is created by drawing N events (x_i, z_i) from the original MadGraph5 samples with probabilities given by their corresponding weights. This is done via inverse transform sampling, where uniformly distributed random variables are converted into samples following $p(x_i, z_i|\theta)$ using the inverse of the cumulative distribution function (CDF). The unweighted sample is then augmented by calculating the joint score and/or the joint likelihood ratio for each event.

For the SALLY method, the sampling process is straightforward: events are sampled at the SM point, and the joint score is evaluated only at this point. In contrast, the ALICE(S) method is more complex, as it involves computing the joint likelihood ratio for multiple pairs of $\theta = (\theta_0, \theta_1)$ values, with θ_0 sampled from either a Gaussian or a flat distribution.

To reduce statistical fluctuations from large event weights, only events generated at the nearest benchmark point were used in the sampling. The sampling setup for each scenario is summarized in Table VI.

G. Training settings

In this study, the ALICES, ALICE, and SALLY methods were trained for each of the three scenarios using the

unweighted training data. The training dataset was defined as 80% of the total generated samples. This dataset was then further split into 75% for training and 25% for validation.

Two sets of variables were used as inputs to the neural networks. The first set includes kinematic information on the b jets, the charged lepton and missing transverse energy (same variables included in Ref. [22]). The second set is extended to include also the aforementioned angular observables, $Q_\ell \cos \delta^+$ and $Q_\ell \cos \delta^-$, as well as the longitudinal momentum of the neutrino, p_z^ν .

The dense neural network employed in the 2D simultaneous (validation) fit setup consisted of two (one) hidden layers with 100 nodes each (50 nodes) and with Tanh (ReLU) activation functions. The training was performed using PyTorch [41], with a batch size of 128, using the AMSGrad [42] optimizer with an initial learning rate of 0.001, later reduced to 0.0001. Input variables were standardized to zero mean and unit variance. Early stopping was employed, based on the validation dataset.

To improve the robustness of the result, an ensemble of five neural networks was trained for each setup, each using a separate unweighted dataset obtained through independent sampling from the same initial weighted dataset. In the case of ALICES, the additional hyperparameter entering the loss function α had to be defined and was set to 5 for the validation and high- p_T^W scenarios and 10 for inclusive p_T^W . These values were found to reduce the

TABLE VI. Sampling setup for the validation and 2D fit scenarios. For the 2D case, the same setup was used for both the inclusive and high- p_T^W studies.

	Validation	2D
n_{samples}	1.0×10^6	11.5×10^6
Distribution	Gaussian ($\mu = 0, \sigma = 0.4$) or uniform ($[-1.2, 1.2]$)	Gaussian ($\mu = 0, \sigma = 0.4$) or Gaussian ($\mu = 0, \sigma = 0.3$)
n_{θ_0}	1000 or 10000	10000 per θ

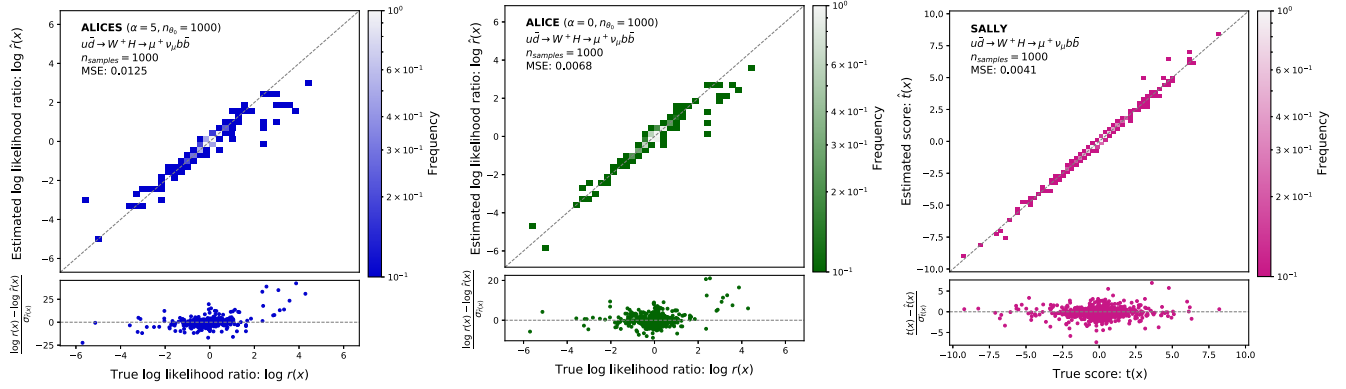


FIG. 5. Estimated vs true log-likelihood ratio or score for the ALICES (left), ALICE (middle), and SALLY (right) techniques.

variance of the individual estimators entering the ensemble, as later discussed.

H. Limit setting

To extract the 68% and 95% confidence levels (CL) of the $c_{H\tilde{W}}$ and c_{HW} coefficients, the p values associated with the negative-log-likelihood-ratio test statistic were calculated in the asymptotic limit, across the θ parameter space, and using an “Asimov dataset” [43].

For ALICE(S), calculating the p values is straightforward, as it involves evaluating the NNs at various parameter points to obtain the likelihood ratio. For the SALLY technique, inference is performed similarly to histograms of summary statistics. The $c_{H\tilde{W}}$ coefficient was scanned over a parameter range of $[-1.2, 1.2]$ and c_{HW} over $[-1.0, 1.0]$, using 35 points in each direction. The Asimov datasets (the template histograms) were built with 50% (100%) of the dataset.

The template histograms for $m_T^{\ell\nu b\bar{b}}$, $Q_\ell \cos \delta^+$, and $Q_\ell \cos \delta^+ \otimes p_T^W$ were built following the binning of Ref. [22]. For SALLY, 8×8 bins were used in the two-dimensional parameter space. A luminosity of 300 fb^{-1} was assumed for the analysis.

IV. RESULTS

A. Validation at parton-level

To validate the three methods, estimated versus true likelihood ratio/score curves were plotted, and the Mean Squared Error (MSE) was computed to quantify the degree of agreement between the two quantities. The values obtained were 0.0125 for ALICES, 0.0068 for ALICE and 0.0041 for SALLY. As illustrated in Fig. 5, the estimated quantities from each method closely align with the true values, confirming that these inference techniques yield reliable results, at least within the truth-level scenario.

In addition, the resulting MSE values for the ALICE and ALICES methods are consistently higher when trained with a uniform distribution, since it leads to a sparser sampling across the θ parameter space. Therefore, a

Gaussian distribution, which is more sample efficient, was chosen for all subsequent studies.

Regarding the individual negative-log-likelihood scans, the ALICES and ALICE estimator exhibit significant variance among their predictions for the maximum likelihood estimate (MLE). As prescribed in Ref. [14], constructing an ensemble proved essential to extract the correct minimum when using the ALICE(S) techniques. For example, the standard deviation in the central values of 5 ALICE(S) estimators was found to be 0.6 (0.2) compared to 0.001 from SALLY.

Upon some investigation of the potential sources of the ALICE(S) variance, it was found that a portion of it can be attributed to the different θ_0 populations within each dataset used for training. This was observed to be slightly mitigated by increasing the number of θ_0 drawn from the Gaussian distribution, specifically from 1000 to 10000. A non-negligible variance remained, which can in part be attributed to the increased complexity of the ALICE(S) loss functions compared to SALLY: by learning the θ dependence, these models become more susceptible to outliers and fluctuations. The value of the hyperparameter α was also found to have some impact on the resulting variance. Finally, an additional source of variance is the contribution of events from a phase-space regions with low signal to background (S/B) ratio. This is further discussed in Sec. V.

B. 2D simultaneous fit

SALLY, ALICE, and ALICES were trained with both the kinematic-only and extended sets of input features. It was observed that including the angular observables and the longitudinal momentum of the neutrino helped the estimation tasks. It was also observed that the ALICE method failed to learn the likelihood function, exhibiting a double minima behavior with very large variance between the estimators.²

²This behavior was observed even in the ideal scenario of signal-only inference.

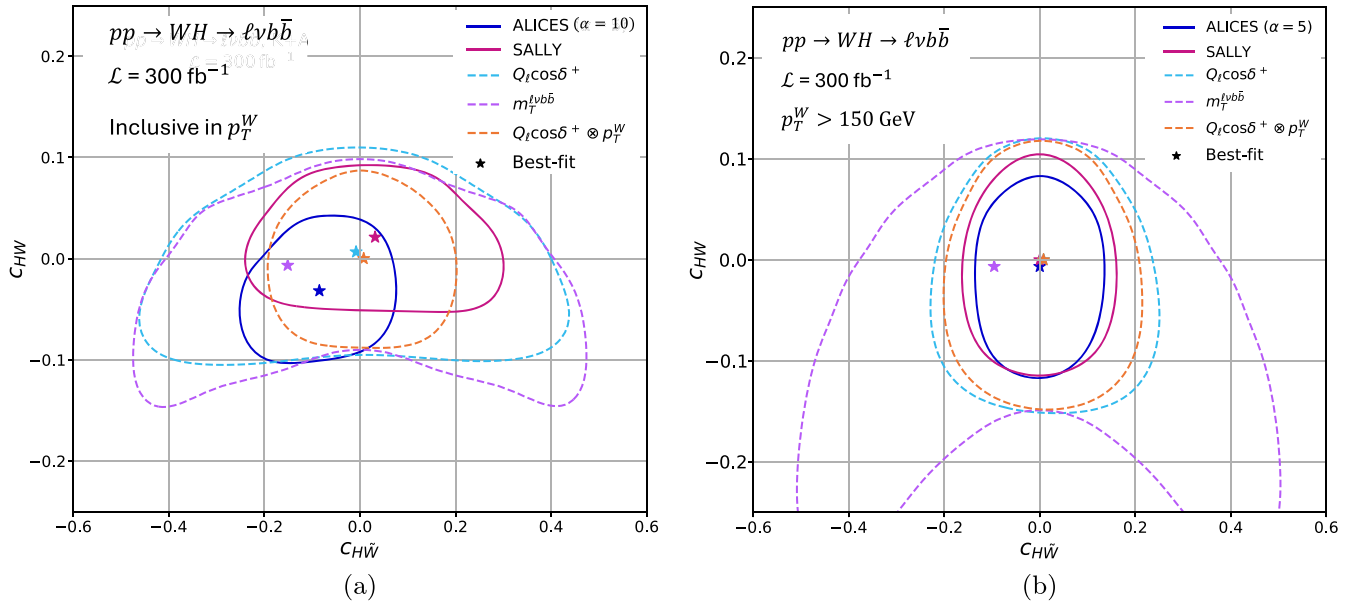


FIG. 6. Likelihood ratio contours at 95% CL in a simultaneous fit of the Wilson coefficients $c_{H\bar{W}}$ and c_{HW} . The scans compare the ALICES (blue) and SALLY (pink) methods, as well as the histogram-based results with $Q_\ell \cos \delta^+$, $m_T^{\ell \nu b \bar{b}}$ and $Q_\ell \cos \delta^+ \otimes p_T^W$. The best-fit points are marked with a star. The models are trained and evaluated in (a) an inclusive phase-space and in (b) a subset of events, where $p_T^W > 150$ GeV.

This highlights the importance of including the joint score information, as is done in ALICES.

The results are shown in Fig. 6. Overall, SALLY and ALICES yielded tighter limits than those obtained with the best-performing one-dimensional summary statistic ($Q_\ell \cos \delta^+$ for $c_{H\bar{W}}$ and $m_T^{\ell \nu b \bar{b}}$ for c_{HW}). However, the constraints from SALLY were nearly identical to those obtained using the two-dimensional histogram of $Q_\ell \cos \delta^+ \otimes p_T^W$. This is illustrated in Fig. 6(a), which shows the 95% CL contours in the $c_{H\bar{W}}-c_{HW}$ plane. Nevertheless, SALLY remains particularly promising compared to this 2D histogram, as it can be straightforwardly extended to probe more couplings simultaneously, i.e., without the need for defining new observables.

In order to understand the role of the low S/B regions in the performance of the methods, the training and evaluation were repeated with events in the $p_T^W > 150$ GeV region. From the results shown in Fig. 6(b), it is evident that focusing on the phase-space with higher S/B is helpful in reducing the variance of the ML methods and producing the correct minimum. This is especially true for ALICES, as shown in Fig. 7, where the ALICES contours are plotted for each estimator of the ensembles, comparing the inclusive and the high- p_T^W regions. In terms of sensitivity to the CP -odd operator, for both ALICES and SALLY focusing on the high- p_T^W regions produces tighter constraints. On the other hand, focusing on this region results in a lower sensitivity to the c_{HW} coefficient, for all methods, but especially the $m_T^{\ell \nu b \bar{b}}$ histogram. This behavior is explained by the distinct impact of the CP -even and CP -odd operators as a function

of p_T^W : higher absolute values of $c_{H\bar{W}}$ lead to increasingly higher S/B as p_T^W increases, independently of the sign of the coefficient; on the other hand, the impact of c_{HW} on the p_T^W distribution is less pronounced and depends on its sign. A positive CP -even coefficient impacts mostly the signal normalization, while a negative coefficient can increase (lower) the S/B in the high (low) p_T^W region. This is the same behavior observed in the histogram of the (highly correlated) variable $m_T^{\ell \nu b \bar{b}}$ (see Fig. 3). Finally, it is interesting to observe that the sensitivities of the ML methods to the CP -even coefficient are less susceptible to this narrowing of the phase space: this is potentially another advantage of further exploring them. The obtained 95% individual intervals, calculated by fixing one Wilson coefficient to zero while profiling over the other, are shown in Tables VII and VIII for all methods and summarized in Fig. 8. We see that in the high- p_T^W region, ALICES (SALLY) produces limits on $c_{H\bar{W}}$ that are 1.6 (1.3) times better than the 2D summary statistics. For c_{HW} , SALLY produces the best limits in the high- p_T^W region, 1.3 times tighter than the 2D histogram (1.2 times for ALICES). For the inclusive scenario, leaving out ALICES given its less robust results, we see that the 2D summary statistic outperforms SALLY by a 1.2 factor for $c_{H\bar{W}}$, whereas SALLY outperforms the 2D histograms by a factor of 1.3 for the c_{HW} coefficient.

The strongest limits on $c_{H\bar{W}}$ and c_{HW} reported above can be used to derive 95% CL lower limits on the scale of new physics, Λ , corresponding to 38 and 53 TeV, respectively.

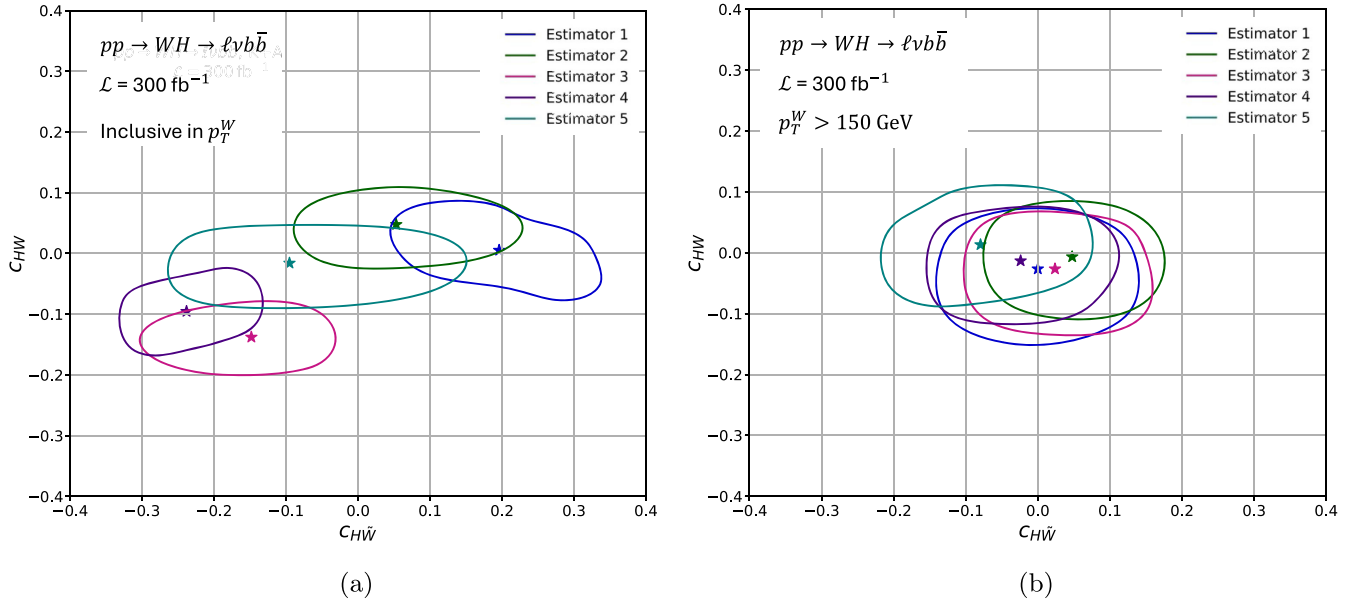


FIG. 7. Likelihood ratio contours at 95% CL in a simultaneous fit of the Wilson coefficients $c_{H\bar{W}}$ and c_{HW} . The scans compare the independent ALICES estimators in (a) an inclusive phase space and in (b) a subset of events where $p_T^W > 150 \text{ GeV}$. The hyperparameter α is set to 10 and 5, respectively. The best-fit points are marked with a star.

C. Discussion

The ALICES estimators learn the correct marginalization of $r(x, z|\theta_0, \theta_1)$ according to the joint density $p(x, z)$. Therefore, in order to better understand the observed

performance, it is useful to plot the joint likelihood ratio for several values of θ_0 drawn from a Gaussian distribution. This is effectively a snapshot of the augmented data the network uses as input. Figure 9 shows the histograms of the

TABLE VII. Summary of the 95% CLs for the c_{HW} Wilson coefficient, obtained using the ALICES and SALLY methods, as well as the best limits derived from 1D and 2D summary statistics: $Q_\ell \cos \delta^+$ and $Q_\ell \cos \delta^+ \otimes p_T^W$. The results are shown for the fits performed in an inclusive phase space and in a subset of events where $p_T^W > 150 \text{ GeV}$.

	Inclusive in p_T	$p_T^W > 150 \text{ GeV}$
ALICES	$[-0.18, 0.11]$	$[-0.11, 0.10]$
SALLY	$[-0.19, 0.20]$	$[-0.13, 0.13]$
$Q_\ell \cos \delta^+$	$[-0.35, 0.34]$	$[-0.18, 0.19]$
$Q_\ell \cos \delta^+ \otimes p_T^W$	$[-0.15, 0.17]$	$[-0.16, 0.17]$

TABLE VIII. Summary of the 95% CLs for the c_{HW} Wilson coefficient, obtained using the ALICES and SALLY methods, as well as the best limits derived from 1D and 2D summary statistics: $m_T^{\ell \nu b \bar{b}}$ and $Q_\ell \cos \delta^+ \otimes p_T^W$. The results are shown for the fits performed in an inclusive phase space and in a subset of events where $p_T^W > 150 \text{ GeV}$.

	Inclusive in p_T	$p_T^W > 150 \text{ GeV}$
ALICES	$[0.05, 0.14]$	$[-0.10, 0.07]$
SALLY	$[-0.05, 0.06]$	$[-0.09, 0.07]$
$m_T^{\ell \nu b \bar{b}}$	$[-0.08, 0.07]$	$[-0.12, 0.09]$
$Q_\ell \cos \delta^+ \otimes p_T^W$	$[-0.07, 0.07]$	$[-0.12, 0.09]$

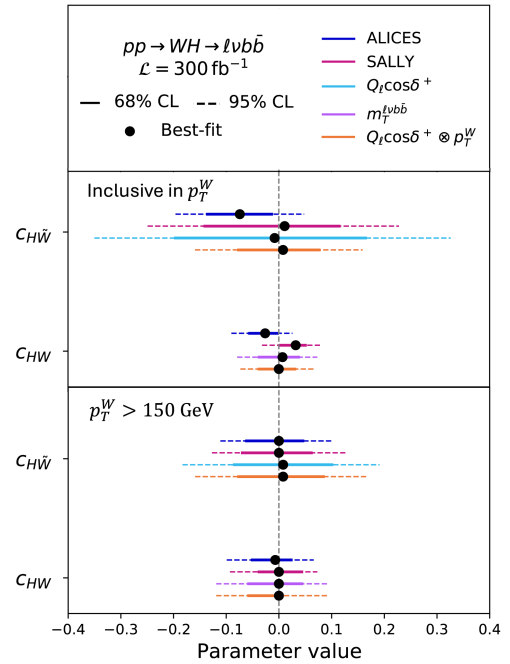


FIG. 8. Summary of the 68% and 98% CLs obtained for $c_{H\bar{W}}$ and c_{HW} Wilson coefficients. The results are shown for the fits performed in an inclusive phase space (top) and in a subset of events, where $p_T^W > 150 \text{ GeV}$ (bottom).

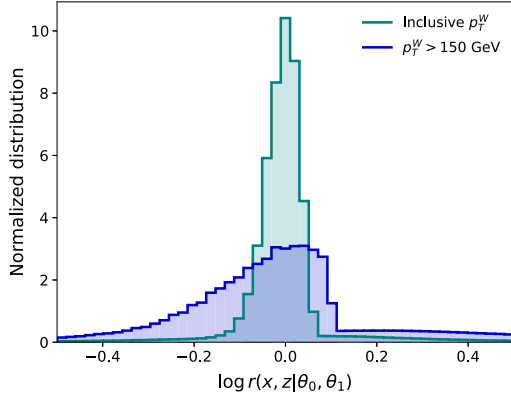


FIG. 9. Histograms of the joint log-likelihood ratio for the inclusive (teal) and high- p_T^W (blue) scenarios.

joint likelihood ratio for both the inclusive and high- p_T^W scenarios.

From Fig. 9, it is clear that the joint quantity takes extremely small values, particularly for the inclusive p_T^W scenario, confirming how kinematically alike the BSM parameter points are. Therefore, any bounds derived from this quantity will be highly sensitive to minor variations in the network output. One way to tackle this is naturally to reduce the impact of the backgrounds and increase the sensitivity to the BSM couplings, as is done by moving to higher p_T^W regions.

The noticeable improvement in performance in the high- p_T^W scenario, compared to the inclusive one, can be understood by examining the 2D density plots of the joint log-likelihood ratio over the $(c_{H\tilde{W}}, c_{HW})$ parameter space, shown in Fig. 10. For the inclusive case [Fig. 10(a)], the joint log-likelihood ratio is nearly flat across the parameter

space, making it extremely challenging for the NN to distinguish between different values of the Wilson coefficients. This explains the observed spread in the MLEs and the significant variance among the individual estimators.

In contrast, for the high- p_T^W region [Fig. 10(b)] a clear radial spread is visible: the magnitude of the joint log-likelihood ratio increases with the absolute values of the Wilson coefficients, reflecting the enhanced sensitivity to BSM effects. The two main populations in the plot also become more pronounced: the green region, corresponding to points where the linear term in c_{HW} dominates and, being negative, leads to cross sections smaller than the SM prediction; and the purple region, where the cross section increases relative to the SM. Near the SM point, the joint log-likelihood ratio remains close to zero, as expected from the small impact of the BSM parameters. The enhanced discriminating power across the parameter space in the high- p_T^W case enables the estimator to more reliably distinguish between SM and BSM scenarios, improving the overall sensitivity of the method.

It is noteworthy that the SALLY estimator, on the other hand, learns the score function, which is an optimal observable that encapsulates all the information about the parameters near the SM and, therefore, can effectively capture the differences between SM and BSM points, even for small values of the Wilson coefficients.

From a practical perspective, the computational cost of implementing these ML-based methods can be substantial, particularly for ALICES, where training and evaluation can take several hours, in contrast to the few minutes required for SALLY. In addition, the sampling time for SALLY is also significantly shorter, since it only requires events generated at the SM point. A possible way to improve

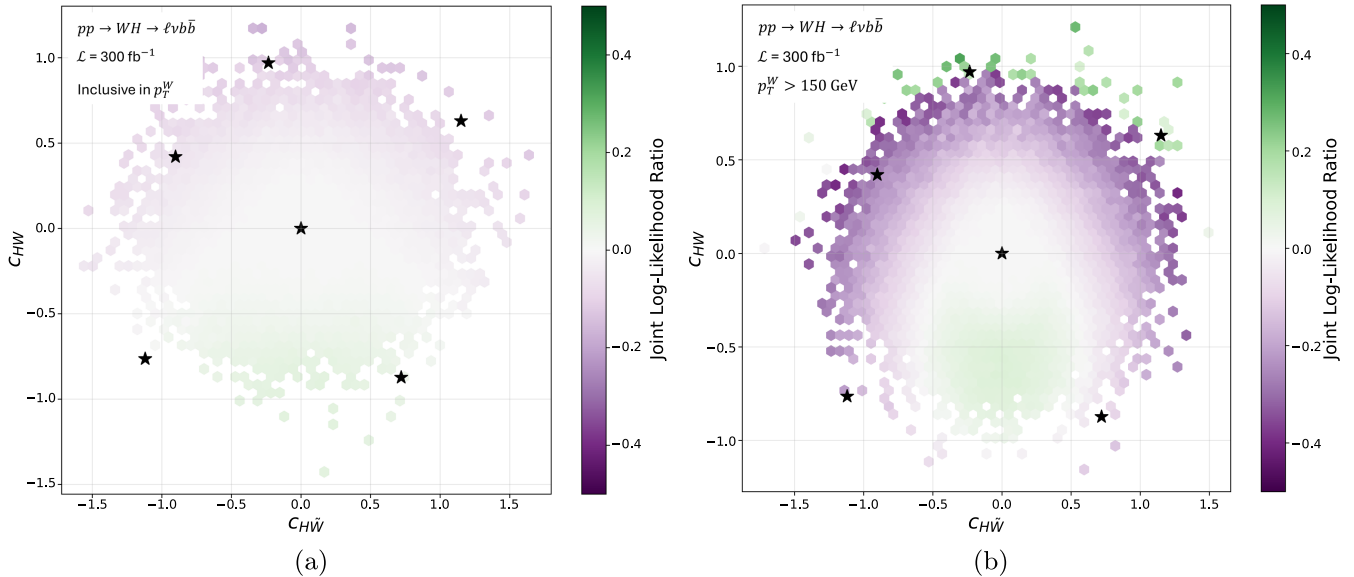


FIG. 10. Density plots of the joint log-likelihood ratio for the ALICES method in the $(c_{H\tilde{W}}, c_{HW})$ parameter space. The plots are shown for (a) an inclusive phase space and (b) a subset of events where $p_T^W > 150$ GeV.

the sampling efficiency of ALICES would be to calculate a binned cumulative distribution function or to explore alternative sampling strategies. These approaches are left for future work.

Regarding the robustness of the methods of ALICE(S), in addition to ensembling, calibration methods can also be explored, with the general aim of ensuring better estimators of the true likelihood ratios than the raw output of the machine learning models. In particular, the expectation calibration method described in Ref. [14] is a type of closure calibration to ensure that evaluating the likelihood ratio estimator on a sample of events drawn according to the SM hypothesis returns an expectation value of 1. This was studied but found to have no significant impact on the results. Another calibration method is to ensure, e.g., that the classifier $\hat{s}(x|\theta_0, \theta_1)$ has a direct probabilistic interpretation and corresponds to the optimal decision function as expected by the likelihood ratio trick. A challenge with pursuing such a calibration in this case is that it has to be implemented for the various θ points. This and other possible strategies for improving robustness have been left for future work.

Finally, a possible simplification of the EFT problem would be to factorize out the dependence on the Wilson coefficients and implement one estimator per contribution in the squared matrix element. With this modification, a parametrized classifier is no longer necessary. ALICES would still be a good candidate to derive the resulting estimators, given the documented improvement in computational efficiency when compared to a method like CARL based on a standard cross-entropy estimator [21].

Overall, these results demonstrate that such ML-based techniques hold significant promise for application in realistic LHC analyses. However, this study highlights the importance of further investigating ways to improve their robustness, understanding their behavior across different physics scenarios, and optimizing their implementation for both accuracy and computational efficiency.

V. CONCLUSIONS

In this work, the sensitivity to the $c_{H\bar{W}}$ and c_{HW} coefficients was explored using ML-based inference methods, in the $WH \rightarrow \ell \nu b \bar{b}$ channel ($\ell = e, \mu$).

As the LHC Run 3 advances, the application of ML-based inference methods, such as ALICES and SALLY, shows great promise in probing HWW anomalous couplings with higher sensitivity and precision. In the $WH \rightarrow \ell \nu b \bar{b}$ channel, and focusing on the high- p_T^W region, we observe that ALICES (SALLY) produces limits on $c_{H\bar{W}}$ that are 1.6 (1.3) times better than the 2D summary statistics. When it comes to the CP -even operator c_{HW} , it is SALLY that produces the best limits in this region, 1.3 times tighter than the 2D histogram (1.2 times for

ALICES). These techniques offer the potential to improve upon the traditional methods and current results from the ATLAS and CMS Collaborations.

The advantages of these techniques come with the trade-off of increased complexity and resource demands. In particular, large amounts of training data are needed to effectively train neural networks, and it can be difficult for them to converge to the true likelihood ratio when BSM signals are similar to the SM ones or in low S/B scenarios. In the particular channel studied in this work, a cut on $p_T(W)$ was essential to improve the robustness of the ALICES estimators.

Despite these complexities, their potential for simultaneously probing multiple SMEFT operators will make them highly valuable tools for Run 3 and beyond. Therefore, this work highlights the importance of addressing the shortcomings of these techniques, such as training stability and computational efficiency, to fully realize their potential.

More broadly, an increase in the sensitivity to the studied Wilson coefficients is directly connected to an increase in the sensitivity to models with extended scalar sectors such as the two Higgs doublet model (2HDM) and the complex two Higgs doublet model (C2HDM), which can generate non-zero values for c_{HW} [44] and $c_{H\bar{W}}$ [45], respectively. Reference [45] shows that a value of $c_{H\bar{W}} \approx 10^{-3}$ can be generated in the C2HDM, which may be attainable at future colliders.

Finally, beyond c_{HW} and $c_{H\bar{W}}$, the strength of these methods in probing simultaneously a large number of operators will bring sensitivity to a wider range of new physics such as models with scalar leptoquarks or vector-like fermions, which can generate nonzero values for a large range of CP -even [44] and CP -odd [46] operators, respectively.

ACKNOWLEDGMENTS

The authors would like to thank Johann Brehmer and Aishik Ghosh for helpful discussions and feedback on the manuscript. M. S. is supported by Nikhef. I. O. is supported by the Portuguese funding agency FCT (Fundação para a Ciência e a Tecnologia) within Projects No. 2023.00042.RESTART and No. 2023.06052.CEECIND/CP2838/CT0001. R. B. is supported by the Austrian Science Fund (FWF) under Grant No. PAT7453824 and received support from FCT within Projects No. SFRH/BD/150792/2020 and No. 2023.00042.RESTART. The authors would also like to acknowledge the FCT Project No. 2024.00227.CERN.

DATA AVAILABILITY

The data that support the findings of this article are openly available [47].

- [1] A. D. Sakharov, *Pis'ma Zh. Eksp. Teor. Fiz.* **5**, 32 (1967).
- [2] J. Brehmer, *Nat. Rev. Phys.* **3**, 305 (2021).
- [3] G. Aad *et al.* (ATLAS Collaboration), [arXiv:2504.07686](#).
- [4] ATLAS Collaboration, Technical Report No. ATL-PHYS-PUB-2025-022 (2025).
- [5] J. Brehmer, K. Cranmer, I. Espejo, F. Kling, G. Louppe, and J. Pavez, *J. Phys. Conf. Ser.* **1525**, 012022 (2020).
- [6] N. Castro, K. Cranmer, A. V. Gritsan, J. Howarth, G. Magni, K. Mimasu, J. Rojo, J. Roskes, E. Vryonidou, and T. You, [arXiv:2211.08353](#).
- [7] K. Kondo, *J. Phys. Soc. Jpn.* **60**, 836 (1991).
- [8] M. Diehl and O. Nachtmann, *Z. Phys. C* **62**, 397 (1994).
- [9] G. Aad *et al.* (ATLAS Collaboration), *J. High Energy Phys.* **05** (2024) 105.
- [10] A. M. Sirunyan *et al.* (CMS Collaboration), *Phys. Rev. D* **104**, 052004 (2021).
- [11] ATLAS Collaboration, [arXiv:2506.19395](#).
- [12] A. Tumasyan *et al.* (CMS Collaboration), *Phys. Rev. D* **108**, 032013 (2023).
- [13] J. Brehmer, K. Cranmer, G. Louppe, and J. Pavez, *Phys. Rev. Lett.* **121**, 111801 (2018).
- [14] J. Brehmer, K. Cranmer, G. Louppe, and J. Pavez, *Phys. Rev. D* **98**, 052004 (2018).
- [15] J. Brehmer, G. Louppe, J. Pavez, and K. Cranmer, *Proc. Natl. Acad. Sci. U.S.A.* **117**, 5242 (2020).
- [16] R. Mastandrea, B. Nachman, and T. Plehn, [arXiv:2405.15847](#).
- [17] H. Bahl, V. Bresó, G. De Crescenzo, and T. Plehn, [arXiv:2410.07315](#).
- [18] A. Ghosh, M. Griese, U. Haisch, and T. H. Park, [arXiv:2507.02032](#).
- [19] J. Neyman and E. S. Pearson, *Phil. Trans. R. Soc. A* **231**, 289 (1933).
- [20] I. Brivio and M. Trott, *Phys. Rep.* **793**, 1 (2019).
- [21] M. Stoye, J. Brehmer, G. Louppe, J. Pavez, and K. Cranmer, [arXiv:1808.00973](#).
- [22] R. Barrué, P. Conde-Muñoz, V. Dao, and R. Santos, *J. High Energy Phys.* **04** (2024) 014.
- [23] J. Brehmer, S. Dawson, S. Homiller, F. Kling, and T. Plehn, *J. High Energy Phys.* **11** (2019) 034.
- [24] V. Chekhovsky *et al.* (CMS Collaboration), *J. High Energy Phys.* **03** (2025) 114.
- [25] G. Aad *et al.* (ATLAS Collaboration), *Rep. Prog. Phys.* **88**, 067801 (2025).
- [26] J. Brehmer and K. Cranmer, [arXiv:2010.06439](#).
- [27] ATLAS Collaboration, Technical Report No. ATLAS-CONF-2022-032 (2022).
- [28] CMS Collaboration, Technical Report No. CMS-PAS-HIG-20-006 (2020).
- [29] CMS Collaboration, Technical Report No. CMS-PAS-HIG-20-007 (2021).
- [30] K. Cranmer, J. Pavez, and G. Louppe, [arXiv:1506.02169](#).
- [31] M. Sugiyama, T. Suzuki, and T. Kanamori, *Density Ratio Estimation in Machine Learning* (Cambridge University Press, Cambridge, England, 2012).
- [32] J. Brehmer, F. Kling, I. Espejo, and K. Cranmer, *Comput. Software Big Sci.* **4**, 3 (2020).
- [33] B. Grzadkowski, M. Iskrzynski, M. Misiak, and J. Rosiek, *J. High Energy Phys.* **10** (2010) 085.
- [34] G. Aad *et al.* (ATLAS Collaboration), *J. High Energy Phys.* **07** (2023) 088.
- [35] R. M. Godbole, D. J. Miller, K. A. Mohan, and C. D. White, *J. High Energy Phys.* **04** (2015) 103.
- [36] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, *J. High Energy Phys.* **07** (2014) 079.
- [37] A. N. Rossia and E. Vryonidou, *J. High Energy Phys.* **11** (2024) 142.
- [38] I. Brivio, [arXiv:2012.11343](#).
- [39] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, *Comput. Phys. Commun.* **191**, 159 (2015).
- [40] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), *J. High Energy Phys.* **02** (2014) 057.
- [41] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, [arXiv:1912.01703](#).
- [42] S. J. Reddi, S. Kale, and S. Kumar, [arXiv:1904.09237](#).
- [43] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, [arXiv:1007.1727](#).
- [44] S. D. Bakshi, J. Chakraborty, S. Prakash, S. U. Rahaman, and M. Spannowsky, [arXiv:2103.11593](#).
- [45] D. Huang, A. P. Morais, and R. Santos, *J. High Energy Phys.* **01** (2021) 168.
- [46] W. Naskar, S. Prakash, and S. U. Rahaman, *J. High Energy Phys.* **08** (2022) 190.
- [47] M. F. Silva, Hww Anomalous Couplings Analysis Code, https://github.com/martafsilva31/master_thesis_HWW_anomalous_couplings (2025), gitHub repository.