Deep neural network application: Higgs boson *CP* state mixing angle in $H \to \tau\tau$ decay and at the LHC

K. Lasocha,[1,2] E. Richter-Was,[1] M. Sadowski,[3] and Z. Was[4]

[1] *Institute of Physics, Jagellonian University, Lojasiewicza 11, 30-348 Krakow, Poland*
[2] *CERN, 1211 Geneva 23, Switzerland*
[3] *Faculty of Mathematics and Computer Science, Jagellonian University,*
ul. Lojasiewicza 6, 30-348 Kraków, Poland
[4] *Institute of Nuclear Physics Polish Academy of Sciences, PL-31342 Krakow, Poland*

(Received 17 July 2020; accepted 15 January 2021; published 3 February 2021)

The consecutive steps of cascade decay initiated by $H \to \tau\tau$ can be useful for the measurement of Higgs couplings and in particular of the Higgs boson parity. In the previous papers we have found that multidimensional signatures of the $\tau^\pm \to \pi^\pm \pi^0 \nu$ and $\tau^\pm \to 3\pi^\pm \nu$ decays can be used to distinguish between scalar and pseudoscalar Higgs state. The machine learning techniques (ML) of binary classification, offered break-through opportunities to manage such complex multidimensional signatures. The classification between two possible *CP* states: scalar and pseudoscalar, is now extended to the measurement of the hypothetical mixing angle of Higgs boson parity states. The functional dependence of $H \to \tau\tau$ matrix element on the mixing angle is predicted by theory. The potential to determine preferred mixing angle of the Higgs boson events sample including τ-decays is studied using deep neural network. The problem is addressed as classification or regression with the aim to determine the per-event: (a) probability distribution (spin weight) of the mixing angle; (b) parameters of the functional form of the spin weight; (c) the most preferred mixing angle. Performance of proposed methods is evaluated and compared.

DOI: 10.1103/PhysRevD.103.036003

I. INTRODUCTION

Machine learning (ML) techniques find increasing number of applications in high energy physics (HEP) phenomenology. Being used at Tevatron and LHC experiments, they have became an analysis standards. For the recent reviews see, e.g., [1–3]. The most common approach is via classification routines, however the impact of regression methods is not negligible as well. Let us point to two such examples in LHC experimental analysis. The measurement of polarization fractions in *WW* pair production using deep neural network (*DNN*) [4] explores both; the classification [5] and regression [6] approaches. The regression technique is also used in [7] for parton distribution functions.

In this paper we present how ML techniques can be helpful to exploit substructure of the hadronically decaying τ leptons in the measurement of the Higgs boson *CP*-state mixing angle in $H \to \tau\tau$ decay. This problem has a long, decades long, history [8,9] and was studied both for electron-positron [10,11] and for hadron-hadron [12,13] colliders. Despite these efforts, the Higgs boson *CP* state was for the first time measured at LHC, from $H \to \tau\tau$ decay only recently see preliminary document [14]. Until that reference the ML has not been even presented for the analysis design, contrary to the classical experimental analysis strategies, see, e.g., [15] for High Luminosity LHC. One of the reasons is that ML adds complexity to the data analysis. ML solutions need to be investigated in context of their suitability for work on systematic ambiguities.

In [14] only some of the τ decay modes are explored. In particular impact parameter is used for the $\tau^\pm \to \pi^\pm \nu$ and $\tau^\pm \to \mu^\pm \nu \bar{\nu}$ and secondary decays of $\tau \to 3\pi\nu$ are only in part used. This is because complexity of the signatures and its multi-dimensional nature. In principle all τ decay modes have the same sensitivity to spin [16] necessary to constrain Higgs *CP* state. It requires precise reconstruction of its all decay products, including neutrinos and knowledge of τ decay matrix elements. In [14] reconstruction of all decay products separately in $\tau \to 3\pi$ decays is not attempted for example. Relatively simple observable for $H \to \tau\tau \to \nu\rho\nu\rho$ which can be understood with the help of single optimal variable [17], for other cases evolve into multitude of Higgs *CP* sensitive variables. For the case of $H \to \tau - \tau \to \nu 3\pi - \nu 3\pi$ that would be 16 variables even if neutrino momenta were not taken into account.

That variables cannot be called optimal, but only optimal-like ones, but may be used in studies of systematic ambiguities. Already in [18] we have investigated such variables. See Fig. 3 there.

Obviously ν_τ escapes direct detection, that is why our paper documents a step of the path, this time toward the measurement of the CP mixing angle and not, as in our previous papers, toward distinguishing two hypotheses of the parity states. We include neutrino momenta, as if they were well accessible, in the feature list used to evaluate options of ML techniques in CP mixing angle measurement. This is obviously an idealistic case, but very useful for exploring potential of the ML techniques. We do not address issues of systematic ambiguities due to reconstruction of neutrino momenta, but we rely on Ref. [19] which was dedicated to the discussion on different approximations and their impact on the ML performance in discriminating between two CP mixing hypotheses. Those results indicate that approximate knowledge on the neutrino momenta can be used for that purpose too. In practice, neutrino momenta need to be reconstructed from energy-momentum conservation of the whole event (in pp case transverse components only) and of the τ-lepton and Higgs boson mass constraints. Even that turns out not to be sufficient. In Ref. [19], devoted to distinction of Higgs CP parity states, we have exploited fact that neutrinos orientation angles can be reconstructed with the help of τ decay vertex position. Even more promising, the τ lepton direction can be reconstructed from τ decay vertex position and ansatz on τ lepton time of flight and then neutrinos momenta can be inferred. The later case turned out to be the best performing of all studied approximations on ν_τ kinematics.

Identification of neutrinos orientation angles is helpful to develop intuition on the observable structure, also as it is the most difficult to grab, to discuss systematic in contributions from $\tau \to 2\pi$ and $\tau \to 3\pi$ decay modes. Such studies are of a value to cross check ML results. They may be useful in evaluation of systematic ambiguities, it is more suitable to evaluate systematic ambiguities for one dimensional distributions. Note that in Ref. [14] there were no attempts to use τ decay position vertices in case of $\tau \to 2\pi$ and $\tau \to 3\pi$ decay modes. The identification and studies of one dimensional distributions important for sensitivity, may be useful in evaluation of systematic ambiguities, it is more standard to evaluate systematic ambiguities for one dimensional distributions. It was a necessary preliminary step for our present study. Obviously they will need, as it was mentioned in Ref. [19], to be repeated with detailed detector response. Not only for momenta smearing, but also for background. The second point will become more important once statistics will grow and systematic errors will get of more concern, than in prototype studies.

On the other hand, theoretical basis for the measurement is simple, the cross section dependence on the parity mixing angle has the form of the first order single angle trigonometric polynomial. It can be implemented in the Monte Carlo simulations as per event spin weight wt, see [20] for more details. In [18,19] we have performed analysis for the three channels of the τ lepton-pair decays, respectively $\rho^\pm\nu_\tau\rho^\mp\nu_\tau$, $a_1^\pm\nu_\tau\rho^\mp\nu_\tau$, and $a_1^\pm\nu_\tau a_1^\mp\nu_\tau$ but we limited ourselves to the scalar-pseudoscalar classification case. In the scope of our interest was the kinematics of outgoing decay products of the τ leptons and geometry of decay vertices.

With these concerns in mind, in the following we extend our previous work on the physics of the Higgs CP parity scalar/pseudoscalar classification, to a measurement of scalar-pseudoscalar mixing angle ϕ^{CP} of the $H\tau\tau$ coupling. We do not intend to investigate possible extensions the Standard Model and avoid discussion on the motivations. We constrain ourselves to the measurement of the coupling and the simplest channel of $H \to \tau^+\tau^- \to \rho^+\nu_\tau\rho^-\nu_\tau \to \pi^+\pi^0\nu_\tau\pi^-\pi^0\nu_\tau$ decay. and focus on comparative studies for potential of different ML techniques.

Possible solutions are analyzed with *deep neural network* (*DNN*) algorithms [4] implemented in *Tensorflow* environment [21] which we have previously found working well for the binary classification [18,19] between scalar or pseudoscalar Higgs boson variants (which correspond to $\phi^{CP} = 0$ and $\phi^{CP} = \pi/2$). Our goals for the *DNN* algorithms are to determine per event:

(i) Spin weight as a function of the mixing angle.
(ii) Decay configuration dependent coefficients, for the known functional form of the spin weight distribution.
(iii) The most preferred mixing angle, i.e., where the spin weight is at a maximum.

These goals are complementary or even to large extent redundant, e.g., with functional form of the spin weight we can easily find the mixing angle at which it has a maximum. But the precision of predicting that value would not be necessarily the same for different methods. All three cases are studied as classification and as regression problems. By this we mean, that the underlying *DNN* cost functions is either designed for classification or regression tasks. We show quantitative comparison of the performance of *DNN* learning on the distributions which are relevant for physics analyses and then draw some conclusions.

Paper is organized as follows. In Sec. II we describe a basic phenomenology of the problem. Properties of the matrix elements and distributions at the level of final, measurable quantities as well as unmeasurable quantities are presented. In Sec. III we review lists of features (variables) used as an input to *DNN* and present samples prepared for analyses. As a straightforward extension of [18,19], still using binary classification, we analyze possibility to distinguish between scalar and arbitrary mix of scalar/pseudoscalar states. This study is covered in Sec. IV. The multi-class classification approach is covered in Sec. V. The regression approach is discussed in Sec. VI.

The comparison of the classification and regression is covered in Sec. VII. Observations relevant for the future studies of systematic errors are addressed. The summary, Sec. VIII, closes the paper.

In Appendix more technical details on the *DNN* architecture are given together with arguments supporting such a choice. In addition, we describe briefly the data preprocessing chain.

II. PHYSICS CONTENT OF THE PROBLEM

The most general Higgs boson Yukawa coupling to τ lepton pair, expressed with the help of the scalar–pseudoscalar parity mixing angle ϕ^{CP} reads as

$$\mathcal{L}_Y = N\bar{\tau}\mathrm{h}(\cos\phi^{CP} + i\sin\phi^{CP}\gamma_5)\tau, \tag{1}$$

where N denotes normalization, h Higgs field and $\bar{\tau}$, τ spinors of the τ^+ and τ^-. As we will see later, this simple analytic form translates itself into useful properties of observable distributions convenient for our goal, determination of the ϕ^{CP}. Recall of the definitions is thus justifiable, and helpful to systematize properties and correlations of the observable quantities (features).

The matrix element squared for the scalar/pseudoscalar/mix parity Higgs, with decay into $\tau^+\tau^-$ pairs can be expressed as

$$|M|^2 \sim 1 + h_+^i h_-^j R_{i,j}; \qquad i,j = \{x,y,z\} \tag{2}$$

where $h_\pm$ denote polarimetric vectors of τ decays (solely defined by τ decay matrix elements) and $R_{i,j}$ density matrix of the τ lepton pair spin state. In Ref. [22] details of the frames used for $R_{i,j}$ and $h_\pm$ definition are given. The corresponding *CP* sensitive spin weight *wt* has the form:

$$wt = 1 - h_+^z h_-^z + h_+^\perp R(2\phi^{CP}) h_-^\perp. \tag{3}$$

The formula is valid for $h_\pm$ defined in $\tau^\pm$ rest-frames, h^z stands for longitudinal and $h^\perp$ for transverse components of h. The $R(2\phi^{CP})$ denotes the $2\phi^{CP}$ angle rotation matrix around the z direction: $R_{xx} = R_{yy} = \cos 2\phi^{CP}$, $R_{xy} = -R_{yx} = \sin 2\phi^{CP}$. The $\tau^\pm$ decay polarimetric vectors h_+^i, h_-^j, in the simplest case of $\tau^\pm \to \pi^\pm\pi^0\nu$ decay, read

$$h_\pm^i = \mathcal{N}(2(q\cdot p_\nu)q^i - q^2 p_\nu^i), \tag{4}$$

where τ decay products $\pi^\pm$, π^0 and ν_τ 4-momenta are denoted respectively as $p_{\pi^\pm}$, p_{π^0}, p_ν and $q = p_{\pi^\pm} - p_{\pi^0}$. Obviously, complete *CP* sensitivity can be extracted only if p_ν is known (for $\tau^\pm \to \pi^\pm\pi^\pm\pi^\mp\nu$ formula is longer, dependence on modeling of the decay appear too [23], but is of no principle differences).

Note that the spin weight *wt* is a simple first order trigonometric polynomial in a $2\phi^{CP}$ angle. This observation is valid for all τ decay channels. For the clarity of the discussion on the *DNN* results, we introduce $\alpha^{CP} = 2\phi^{CP}$, which spans over $(0, 2\pi)$ range. The $\alpha^{CP} = 0, 2\pi$ corresponds to scalar state, the $\alpha^{CP} = \pi$ to pseudoscalar one. Spin weight can be expressed as

$$wt = C_0 + C_1\cdot\cos(\alpha^{CP}) + C_2\cdot\sin(\alpha^{CP}), \tag{5}$$

where

$$\begin{aligned} C_0 &= 1 - h_+^z h_-^z, \\ C_1 &= -h_+^x h_-^x + h_+^y h_-^y, \\ C_2 &= -h_+^x h_-^y - h_+^y h_-^x, \end{aligned} \tag{6}$$

depend on the τ decays only.

Distribution of the C_0, C_1, C_2 coefficients, for the sample of $H \to \tau\tau$ events used for our numerical results is shown in Fig. 1. The C_0 spans $(0, 2)$ range, while C_1 and C_2 of $(-1, 1)$ range have a similar shape, quite different than the one of C_0.

The amplitude of the *wt* as function of α^{CP} depends on the multiplication of the length of the transverse components of the polarimetric vectors. The longitudinal component $h_+^z h_-^z$ is defining shift with respect to zero of the *wt* mean value over a full $(0, 2\pi)$ range. The maximum of the *wt* distribution is reached for $\alpha^{CP} = \measuredangle(h_+^T, h_-^T)$, the opening angle of transverse components of the polarimetric vectors.

The spin weight of formula (5) can be used to introduce transverse spin effects into the event sample when for the generation transverse spin effects were not taken into account at all. The above statement is true, independently if longitudinal spin effects were included and which τ decay channels complete cascade of $H \to \tau\tau$ decay. The shape of weight dependence on the Higgs coupling to τ parity mixing angle is preserved.

In Fig. 2 we show distribution of spin weight *wt* for five example $H \to \tau\tau$ events collected in Table I. For each event, depending on the polarimetric vectors, single value of α^{CP} is preferred (by the largest weight). For a physics model with α^{CP} the sample will be more abundantly populated with events for which the angle between polarimetric vectors, $\measuredangle(h_+^T, h_-^T)$, is close to α^{CP}. We show distributions when complete polarimetric vectors are used for spin weight *wt* and when only hadronic parts of polarimetric vectors are used. The second case is indicating easier to attain sensitivity part of observables. The α^{CP} at which spin weight has its maximum is then a bit shifted. Table I specifies values of the polarimetric vectors and the resulting coefficients C_i calculated from formulas (6) and for events of Fig. 2. It also explicitly gives $\measuredangle(h_+^T, h_-^T)$ calculated from complete polarimetric vectors and (in brackets) from their hadronic parts only.

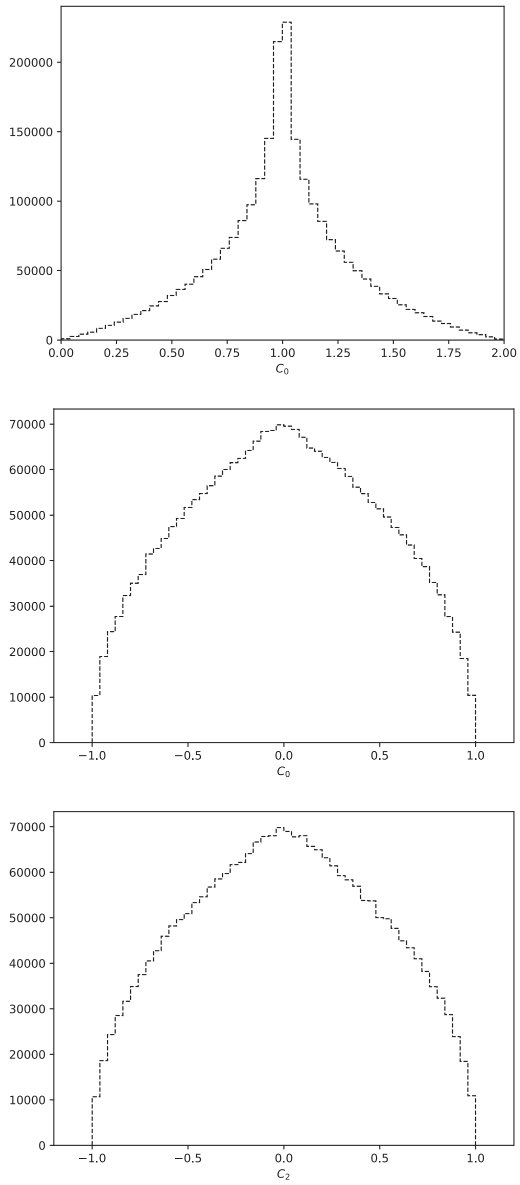

FIG. 1. Distributions of the formula (5) C_0, C_1, C_2 coefficients, for the $H \to \tau\tau$ events sample.

III. MONTE CARLO SAMPLES AND FEATURE LISTS

For compatibility with our previous publications [18,19], we use the same generated event samples, namely Monte Carlo events of the Standard Model, 125 GeV Higgs boson, produced in pp collision at 13 TeV center-of-mass energy, generated with PYTHIA 8.2 [24] and with spin correlations introduced with TAUSPINNER [20] package. For τ lepton decays we use TAUOLAPP library [25]. All spin and parity effects are implemented with the help of weight wt [26,27]. The sample is generated without spin effects, and the spin weights wt_i for few different values of CP mixing angle α_i^{CP} are stored. Spin weight, formula (3), is calculated using $R_{i,j}$ density matrix and polarimetric vectors $h_{\pm}$.

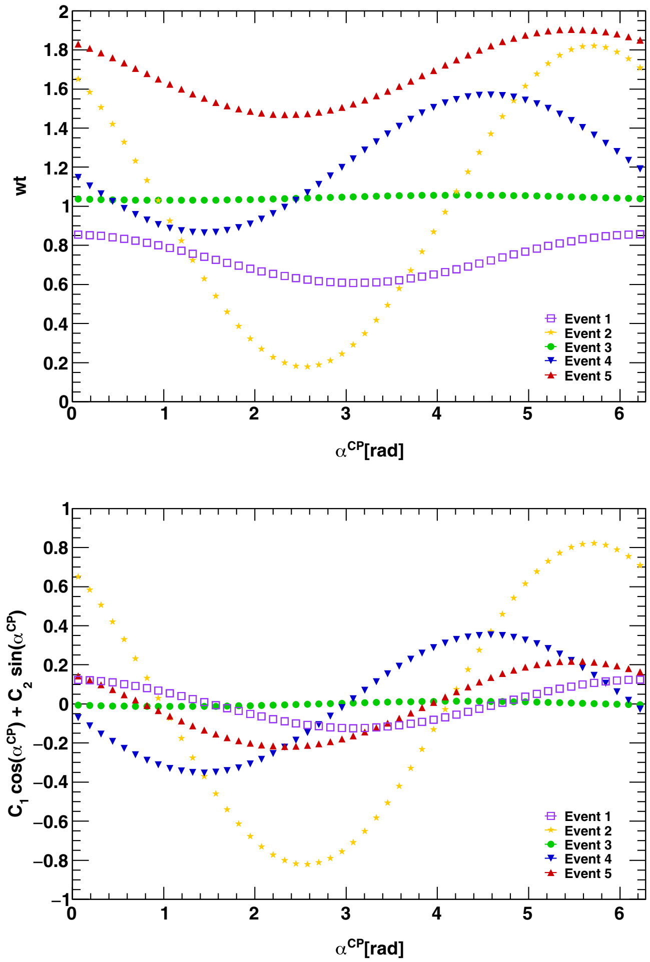

FIG. 2. The spin weight wt (top plot) and only its α^{CP} dependent component (bottom plot) for five $H \to \tau\tau$ events of Table I. Note the vertical scale change between top and bottom plots.

Later, for a given event it is possible to calculate coefficients C_0, C_1, C_2, using three α^{CP} and linear equation (5). Figure 3 shows the cross-check how well this procedure works. The functional form (orange line) and evaluated spin weights (blue dots) for two example events are shown. The C_0, C_1, C_2 coefficients for the functional form are calculated solving Eq. (5) for wt stored in the generated event samples at three values of α^{CP}.

In this paper we present results for the case when both τ's decay $\tau^{\pm} \to \rho^{\pm}\nu_{\tau}$ and about 5×10^6 simulated Higgs events are used. To partly emulate detector conditions, a minimal set of cuts is used. We require that the transverse momenta of the visible decay products combined, for each τ, are larger than 20 GeV. It is also required that the transverse momentum of each $\pi^{\pm}$ is larger than 1 GeV. In experimental conditions complex cuts will be applied.

TABLE I. Polarimetric vectors, resulting C_i coefficients of formulas (6) and angle $\measuredangle(h_+^T, h_-^T)$ between transverse components of polarimetic vectors for five example events of $H \to \tau^+\tau^-, \tau^\pm \to \rho^\pm \nu_\tau$. In brackets, angle of only hadronic part of polarimetric vector is given.

Events	Polarimetric vectors	$\lvert h_+^T \rvert \lvert h_-^T \rvert$	C_0	C_1	C_2	$\measuredangle(h_+^T, h_-^T)$ [rad] (hadronic part only)
Event 1	$h_+^{x,y,z} = (0.7547 - 0.2232 - 0.6167)$ $h_-^{x,y,z} = (-0.9093 - 0.2931 - 0.2953)$	0.7519	0.8179	0.7517	0.0183	6.2586 (6.1738)
Event 2	$h_+^{x,y,z} = (0.86170.04850.5050)$ $h_-^{x,y,z} = (-0.59590.7892 - 0.1487)$	0.8535	1.0751	0.5518	−0.6511	5.4134 (5.6307)
Event 3	$h_+^{x,y,z} = (0.34020.9377 - 0.0682)$ $h_-^{x,y,z} = (0.82620.1272 - 0.5487)$	0.8339	0.9626	−0.1619	−0.8180	5.2130 (4.1923)
Event 4	$h_+^{x,y,z} = (-0.69640.6204 - 0.3605)$ $h_-^{x,y,z} = (0.2142 - 0.3885 - 0.8962)$	0.4138	0.6769	−0.0919	−0.4035	4.4883 (4.5127)
Event 5	$h_+^{x,y,z} = (0.1115 - 0.4989 - 0.8595)$ $h_-^{x,y,z} = (-0.2347 - 0.011080.9720)$	0.1201	1.8354	0.0317	−0.1158	4.9793 (5.4300)

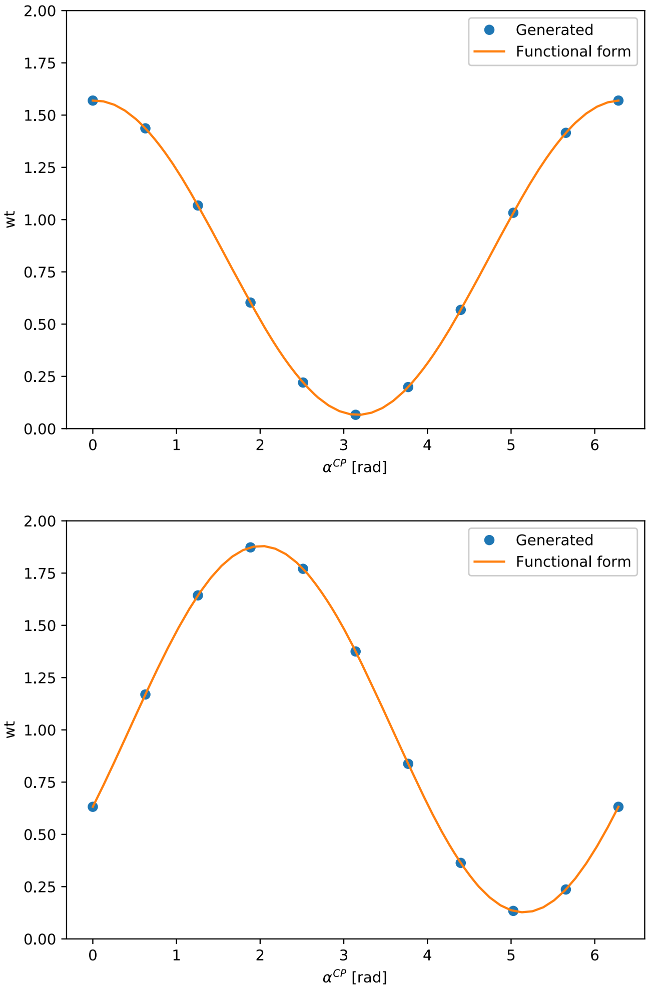

FIG. 3. Cross-check distributions of the spin weight *wt* calculated at generation (blue points) and from functional form of Eq. (5) (orange line), as a function of *CP* mixing parameter α^{CP}. For top and bottom plots two different example events were used. Coefficients C_i are reconstructed from Eq. (5) and *wt* is taken at three different α^{CP}.

Our largely optimistic ones can be used in feasibility estimations though.

The emphasis of the paper is to explore different ML approaches to the problem, and we discuss only the case of the `Variant-All` feature list from paper [19]. It contains the four-momenta of *all* decay products of τ leptons defined in the rest frame of intermediate resonance pairs, and with sum of hadronic decay products aligned with z-axis are. This represents an ideal benchmark case scenario, for performance monitoring.

IV. BINARY CLASSIFICATION

The use of the *DNN* for binary classification have been discussed in our previous papers [18,19]. The focus was on discriminating between *CP*-scalar ($\mathcal{H}_0$ hypothesis) and *CP*-pseudoscalar ($\mathcal{H}_1$ hypothesis).

Now we apply the same procedure but with alternative hypothesis ($\mathcal{H}_{\alpha^{CP}}$) representing the scalar-pseudoscalar mixed state of mixing parameter α^{CP}. To quantify performance for Higgs *CP* state classification the weighted area under curve (AUC) [28,29] is used again. For each simulated event we know also Bayes optimal probability that it is sampled from $\mathcal{H}_0$ or $\mathcal{H}_{\alpha^{CP}}$ hypothesis, see more detailed description in Appendix. This forms the so called *oracle predictions*, i.e., ultimate discrimination for this problem. We calculate oracle predictions and evaluate the results of *DNN*. This is a straightforward extension of the method used in [18,19]. That is why, simple attempt on future discussion of systematic error may follow that suggested in [19]: variations within expected range of detector response can be easily introduced and biases studied.

The oracle predictions for discriminating between $\mathcal{H}_0$ and $\mathcal{H}_{\alpha^{CP}}$ hypotheses is increasing with α^{CP} and reach AUC $= 0.78$ for $\alpha^{CP} = \pi$. The performance of *DNN* is following similar pattern, reaching maximum at $\alpha^{CP} = \pi$

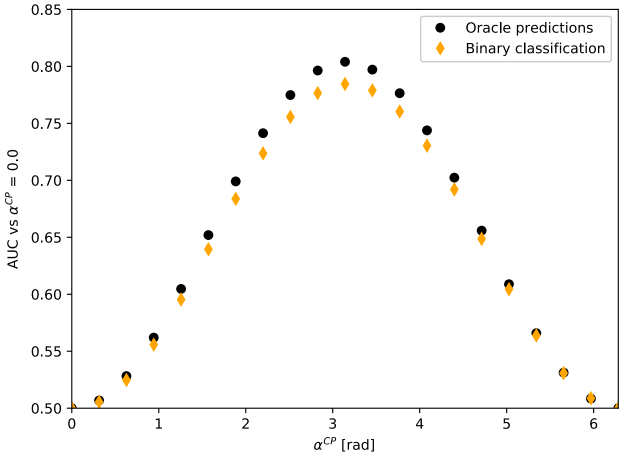

FIG. 4. The AUC score for binary classification between $\mathcal{H}_0$ and $\mathcal{H}_{\alpha^{CP}}$ hypotheses and corresponding oracle predictions.

(pure pseudoscalar case). It decreases for smaller or larger α^{CP}, where admixture of the scalar component appear. In case of complete feature list, it is almost achieving the performance of oracle predictions. In Fig. 4, the AUC values are plotted for full α^{CP} range. The distributions are (almost) symmetric around $\alpha^{CP} = \pi$. Note that the functional form of spin weight wt, Eq. (5), encapsulating sensitivity to α^{CP} is not symmetric, see Fig. 3. In Table II we show numerical results for few α^{CP}.

V. MULTICLASS CLASSIFICATION

The binary classification discussed in the previous section is easy to generalize to the multiclass case. The *DNN* is learning to provide per-event probabilities to associate with each class. Single class represents either discrete point or a specific range in 1-dimensional parameter space. We explore three approaches, each providing complementary physics information, but all allowing to quantify, on the per-event basis, which is the preferred mixing angle of the studied Higgs sample:

TABLE II. The AUC scores for discriminating between Higgs *CP* states. Results from oracle predictions and binary classification for discriminating between $\mathcal{H}_0$ hypothesis that Higgs *CP* is a scalar (*CP*-mixing angle $\alpha^{CP} = 0.0$ or 2π) and $\mathcal{H}_{\alpha^{CP}}$ hypothesis, when Higgs *CP* is of a parity mixed state, are shown. *CP*-mixing angle $\alpha^{CP} = \pi$ corresponds to the pseudoscalar case. Note that for small α^{CP} classification strength grows with its square.

CP-mixing angle α^{CP} (units of π)	Oracle predictions	Binary classification
0.2	0.528	0.525
0.4	0.605	0.595
0.6	0.699	0.684
0.8	0.775	0.756
1.0	0.804	0.784

(i) The *DNN* classifier is learning per-event spin weight as a function of mixing angle α^{CP}. The range of mixing angle $(0, 2\pi)$ is discretized into equally spaced points called classes. This approach is described in Sec. V A, and used for the figures labeled with: `Classification:wt`.

(ii) The *DNN* classifier is learning per-event coefficients C_0, C_1, C_2. The allowed range of coefficients is split into several equal size ranges (classes), single class represents a range for a coefficient value. The *DNN* is trained for each coefficient separately. This approach is described in Sec. V B and used for the figures labeled with: `Classification:` C_0, C_1, C_2.

(iii) The *DNN* classifier is learning per-event most probable mixing angle $\alpha^{CP}_{\max}$, i.e., value of α^{CP} at which spin weight is maximal. The range of mixing angle $(0, 2\pi)$ is split into several equally spaced points (classes). This approach is described in Sec. V C and used for the figures labeled with: `Classification:` $\alpha^{CP}_{\max}$.

We monitor performance of the learning process in a standard manner, with the loss function on the training and validation sets. Respective distributions are shown in Fig. 20 of Appendix. Note that the loss function, the `tf.nn.softmax_cross_entropy_with_logits` of the `Tensorflow`, allows to predict probabilities of the class labels, and not the actual value of the observable at a given class. In case of predicting spin weight distribution, only the normalized to unity shape is predicted. In case of predicting values of C_i coefficients or $\alpha^{CP}_{\max}$, vector of probabilities is returned, and the one-hot encoding transformation selecting most probable class is then applied to retrieve actual predicted value of the parameter.

A. Learning spin weight *wt*

The *DNN* classifier is trained with per-event feature list and as a label normalized to unity N_{class}-dimensional vector of spin weights [30] $wt_i^{\mathrm{norm}} = wt_i / \sum_{i=1}^{i=N_{\mathrm{class}}} wt_i$ is given, each component of $wt^{\mathrm{norm}}(\alpha^{CP})$ vector corresponds to the ith discrete value of mixing angle α_i^{CP}. N_{class} denotes number of points to which range $(0, 2\pi)$ was discretized. The number of classes is kept odd, to assure that $\alpha^{CP} = 0, \pi, 2\pi$, corresponding respectively to scalar/pseudoscalar/scalar cases, are always represented as a separate class. Training of *DNN* is performed with N_{class} varying from 3 to 51. This is to understand the trade-off between the better approximation given by high number of classes and smaller complexity of the low-class system.

We quantify the *DNN* performance for classification problem in the context of physics relevant criteria. The first question is how well *DNN* is able to reproduce per-event shape of the spin weight wt^{norm}. For two example events,

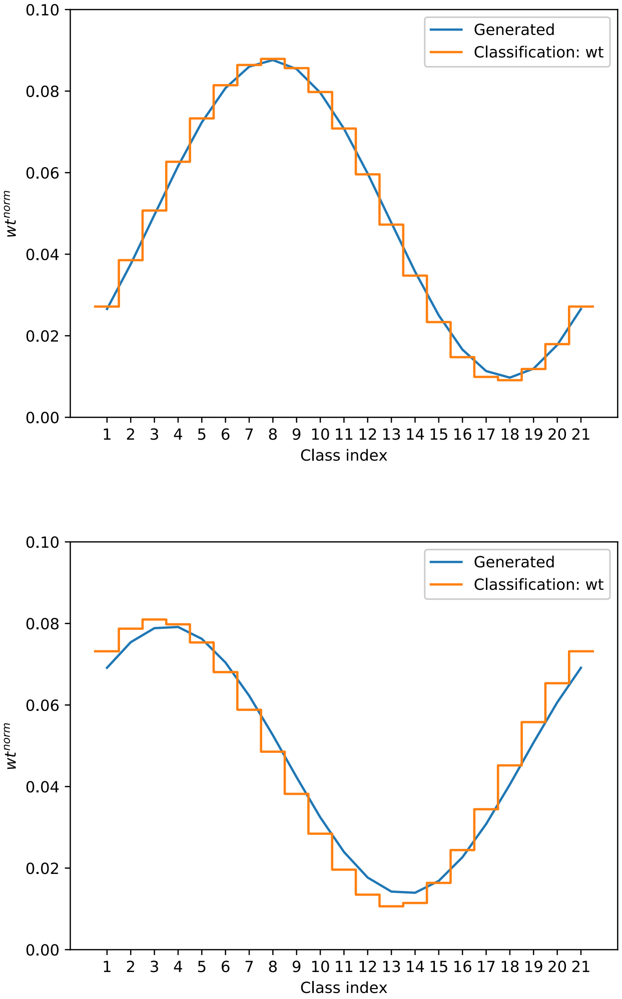

FIG. 5. Normalized to probability spin weight wt^{norm}, predicted (orange steps) and true (blue line), as a function of α_i^{CP} for two example events (top and bottom plots). *DNN* was trained with $N_{\mathrm{class}} = 21$ spanning range $(0, 2\pi)$.

true and predicted spin weight wt^{norm} distribution with $N_{\mathrm{class}} = 21$ is shown in Fig. 5 as a function of either continuous mixing parameter α_i^{CP} or class index i (representing discretized mixing parameter α_i^{CP}). Blue line denote true weights while orange steps denote weights predicted by *DNN* classifier. In overall, predicted weights follow smoothly true shape of linear $\cos(\alpha^{CP})$ and $\sin(\alpha^{CP})$ combination. This is encouraging, because the loss function is not correlating explicitly nearby classes. The *DNN* is discovering this pattern in the process of learning.

To quantify those observations, performance of *DNN* is monitored on the statistical basis with l_2 norm. The l_2 norm is defined as a square root of the integral of squared difference between predicted p_k and true wt_k^{norm} over the whole interval $(0, 2\pi)$. It then averaged over the number of events N_{evt}. Although p_k and wt_k^{norm} are functions of α^{CP}, we shall usually skip the argument for the notation brevity.

$$l_2 = \sum_{k=1}^{N_{\mathrm{evt}}} \frac{\sqrt{\int_0^{2\pi} (wt_k^{\mathrm{norm}}(\alpha^{CP}) - p_k(\alpha^{CP}))^2 d\alpha^{CP}}}{N_{\mathrm{evt}}}. \quad (7)$$

The p_k corresponds to the kth event and is represented as a step function, with step levels given by a N_{class}-dimensional output of DNN. For true weights, represented as continuous function (5), we scale them in such a way that $\int_0^{2\pi} wt^{\mathrm{norm}} d\alpha^{CP} = 1$, to enable the comparison. Distribution of l_2 norm is shown in Fig. 6, as a function of class multiplicity N_{class}. With increasing number of classes, l_2 decreases. The slope remains very steep up to $N_{\mathrm{class}} = 21$, and seems to flatten around $N_{\mathrm{class}} = 51$. These two values of N_{class} we have chosen as representative for the rest of the paper.

From physics perspectives, learning the shape of *wt* distribution as function of α^{CP}, is equivalent to learning components of the polarimetric vectors. But, because only the shape, not the normalization, is available the C_i coefficients cannot be fully retrieved from formula (5). It is not necessary the aim anyway. The physics interest is more to learn α^{CP} which is preferred by events of the analyzed sample, i.e., value at which *wt* distribution has its maximum. This corresponds to determining *CP* mixing angle of the analyzed sample.

The second criterium is the difference between most probable predicted class and most probable true class, denoted as Δ_{class}. When calculating difference between class indices, periodicity of the functional form (5) is taken into account. Class indices represent discrete values of α^{CP}, in range $(0, 2\pi)$. The distance between the first and the last class is zero. We take the distance which corresponds to the smaller angle difference and we take the sign according to clockwise orientation vs class index at which true *wt* has its maximum.

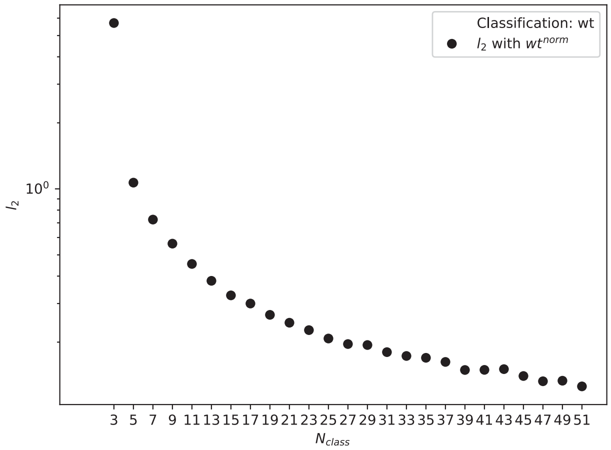

FIG. 6. The l_2 norm, quantifying difference between true and predicted spin weight wt^{norm}, as a function of class multiplicity N_{class}.

Let $idp_{\max}$ denote the index of most probable predicted class, $idc_{\max}$ be index of true most probable class. The distance $|\Delta_{\text{class}}|$ is defined as:

$$|\Delta_{\text{class}}| = \min(|idp_{\max} - idc_{\max}|, (N_{\text{class}} - 1) - |(idp_{\max} - idc_{\max})|), \tag{8}$$

and the sign is attributed

$$\Delta_{\text{class}} = \text{sign}(idp_{\max} - idc_{\max})|\Delta_{\text{class}}|, \tag{9}$$

if $(|idp_{\max} - idc_{\max}|) < ((N_{\text{class}} - 1) - |(idp_{\max} - idc_{\max})|)$, or

$$\Delta_{\text{class}} = \text{sign}(idc_{\max} - idp_{\max})|\Delta_{\text{class}}|, \tag{10}$$

otherwise.

In Fig. 7 distributions of Δ_{class} for $N_{\text{class}} = 21$ (top) and 51 (bottom) are shown. The shapes are Gaussian-like and centered around zero. The mean $\langle\Delta_{\text{class}}\rangle = -0.006$ [rad] in both cases and this we can interpret as the bias of the method. The standard deviation of per-event distribution is $\sigma_{\Delta_{\text{class}}} = 0.165$ [rad] for $N_{\text{class}} = 21$ and $\sigma_{\Delta_{\text{class}}} = 0.126$ [rad] for $N_{\text{class}} = 51$. As we can see, the performance has not improved significantly with N_{class} exceeding 21.

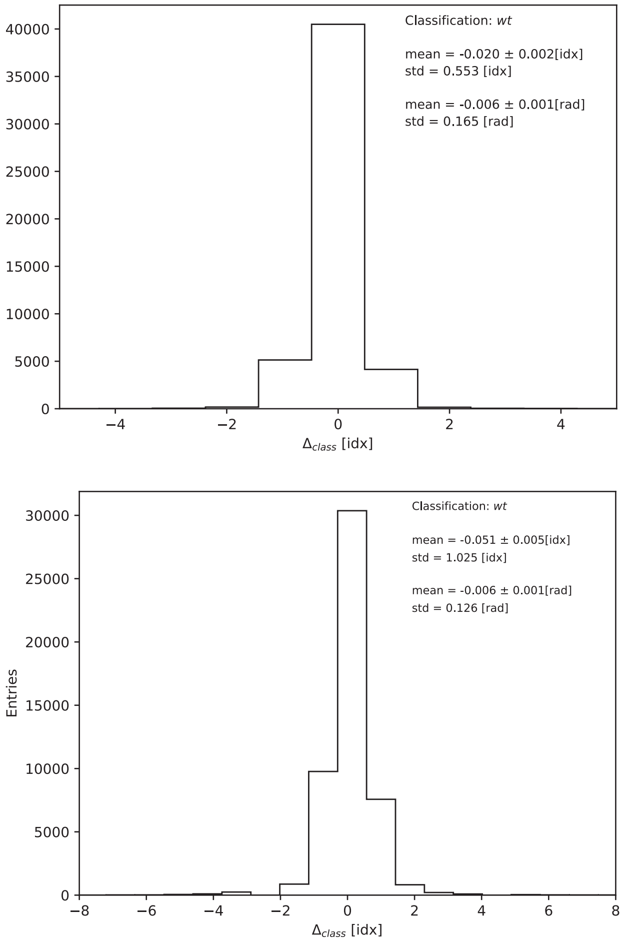

FIG. 7. Distribution of $\Delta^{\max}_{\text{class}}$ between predicted most probable class and true most probable class for $N_{\text{class}} = 21$ (top) and 51 (bottom). The `mean` and `std` are calculated in units of class index [idx] or units of radians [rad].

The *DNN* classifier which is predicting normalized spin weight wt^{norm}, provides enough information to identify the most probable mixing angle $\alpha^{CP}_{\max}$ with high precision. The information is not sufficient though to reconstruct complete set of C_i coefficients and the polarimetric vectors.

B. Learning C_0, C_1, C_2 coefficients

The second approach is to learn formula (5) coefficients C_0, C_1, C_2 for the spin weight wt. They can be then used to predict not only normalized wt^{norm}, but also original wt. Coefficients C_0, C_1, C_2 represent physical observables, products of longitudinal and transverse components of polarimetric vectors, as shown in formulas (6).

The classification technique using *DNN* is configured to learn each of the C_i with separate training. The allowed range is well known, the C_0 spans the range (0.0, 2.0) and C_1, C_2 the range $(-1.0, 1.0)$, see Fig. 1. The allowed range is binned into N_{class} and as a label, the N_{class}-dimensional vector with one-hot encoded value of the C_i parameter is associated with each event. Therefore in this case, a single class represents range of the C_i coefficient. During training, the *DNN* is learning per-event association between feature list and the class labels. The output is a probability N_{class}-dimensional vector, which is then converted to one-hot encoded representation, i.e., the most probable class is chosen as a predicted value of the C_i coefficient.

Distributions of the difference between true and predicted C_i coefficients are shown in Fig. 8. In that case, as there is no periodicity involved, $\Delta_{\text{class}} = idp - idc$ where idp, idc denote respectively true and predicted class index. Mean of ΔC_i is close to zero and standard deviation is of 0.038–0.051, which is less than 5% of the range. Precision with which C_i coefficients are predicted is clearly limited by the N_{class}.

We use the true and predicted C_0, C_1, C_2 coefficients to calculate wt distribution of (5). It is then discretized with N_{class} points (the N_{class} could be different than the one used for learning coefficients), and the $\alpha^{CP}_{\max}$ is determined from the class of maximal weight. The difference between true and predicted $\alpha^{CP}_{\max}$ is shown in Fig. 9 for $N_{\text{class}} = 21$ and 51. The Gaussian-like shape of those distributions, centered around zero, clearly demonstrated that method works as expected. The mean and standard deviation of the distributions are close to those obtained with `Classification:wt` approach, of Fig. 7.

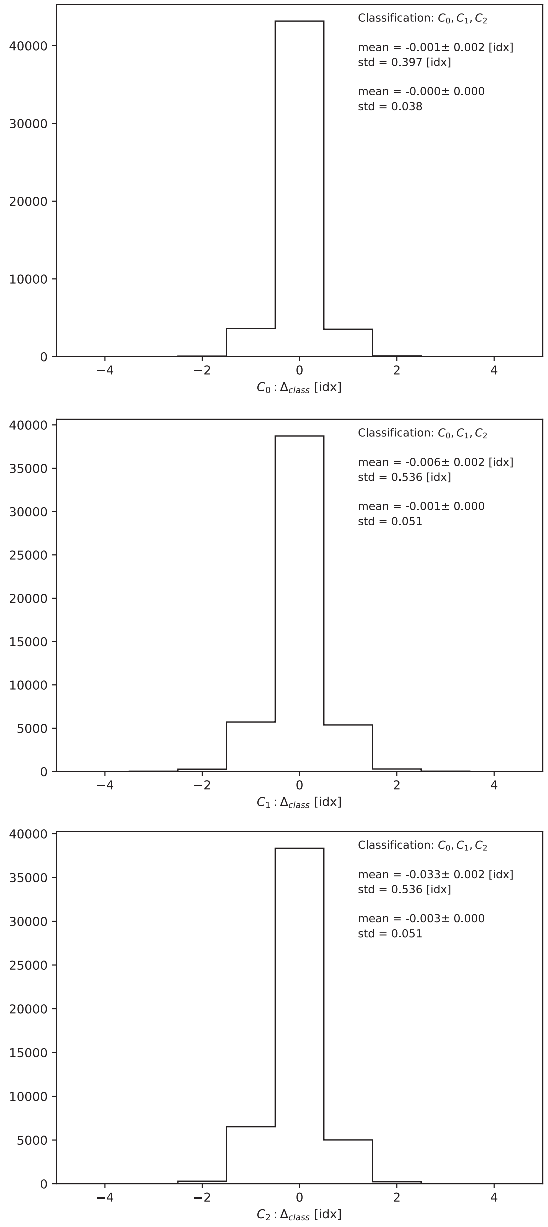

FIG. 8. Difference between true and predicted coefficients C_0, C_1, C_2 of formula (5). For *DNN* training the granularity of $N_{\mathrm{class}} = 21$ was used.

Finally, as sanity check we have compared the true distributions of C_0, C_1, C_2 with the predicted ones. As we can see in Fig. 10, both distributions match very well for all C_i.

C. Learning the $\alpha^{CP}_{\mathrm{max}}$

The third approach is to directly learn per-event most preferred mixing angle, $\alpha^{CP}_{\mathrm{max}}$. The allowed range $(0, 2\pi)$ is again binned into N_{class} classes, where single bin represents discrete α^{CP}. For training, for every event we take the one-hot encoded vector of N_{class}-dimension as a label. The *DNN* is returning N_{class}-dimensional vector of probabilities, which is then transformed into a single number, that is the class of the highest probability $\alpha^{CP}_{\mathrm{max}}$. With this approach, neither spin weight nor C_i coefficients are predicted.

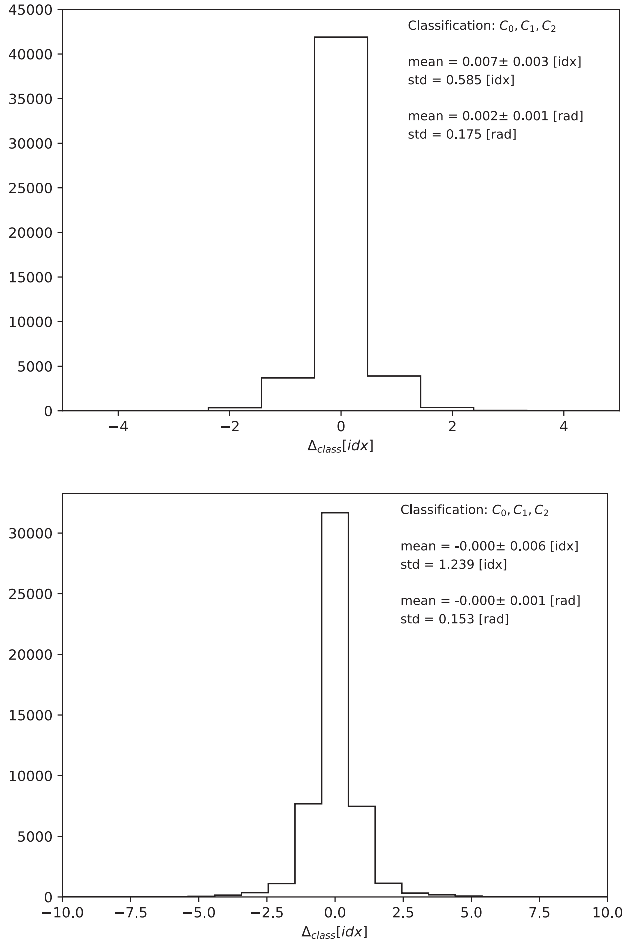

FIG. 9. The difference between true and predicted most probable mixing angle $\alpha^{CP}_{\mathrm{max}}$, calculated using formula (5) and coefficients C_0, C_1, C_2 learned with classification method. The granularity of $\alpha^{CP}_{\mathrm{max}}$, $N_{\mathrm{class}} = 21$ and 51 was used respectively for top and bottom plot.

As the event sample is generated without any *CP* mixture favored, the distribution of the $\alpha^{CP}_{\mathrm{max}}$ is expected to be uniform, and such sanity check is demonstrated in the top plot of Fig. 11. The *DNN* is well reproducing this behavior. The $\Delta\alpha^{CP}_{\mathrm{max}}$, the difference between true and predicted value of the $\alpha^{CP}_{\mathrm{max}}$ is shown in the bottom plot of Fig. 11. In the case of $N_{\mathrm{class}} = 21$, it has a Gaussian-like shape with the mean $\langle\Delta\alpha^{CP}_{\mathrm{max}}\rangle = 0.003 \pm 0.001$ [rad] and standard deviation 0.139 [rad]. Results are again comparable with the ones obtained with the previously discussed approaches.

VI. REGRESSION

The ML regression is not so commonly used in the high energy physics analyses. The main feature is, that contrary

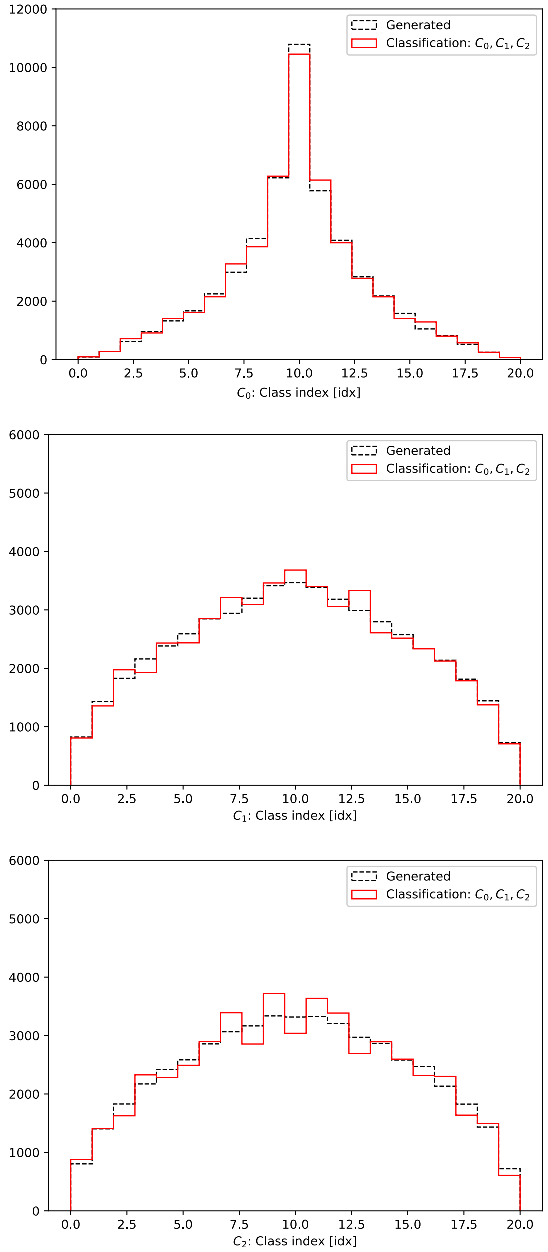

FIG. 10. Distributions of true and predicted coefficients C_0, C_1, C_2 of formula (5). For *DNN* training the granularity of $N_{\mathrm{class}} = 21$ was used.

to the classification case, we get a continuous parameter (or set of parameters) as a *DNN* output. We explore three approaches, defined similarly as in Sec. V

(i) The *DNN* is learning to predict per-event spin weight as a function of mixing angle α^{CP}. The range of mixing angle $(0, 2\pi)$ is split into discrete points of α^{CP} at which value of spin weight is learned. This approach is described in Sec. VI A and used for the figures labeled with: `Regression:wt`.

(ii) The *DNN* is learning to predict per-event value of the coefficients C_0, C_1, C_2 of the functional form (5). The *DNN* is trained for all coefficients simultaneously. This approach is described in Sec. VI B and used for the figures labeled with: `Regression:`C_0, C_1, C_2.

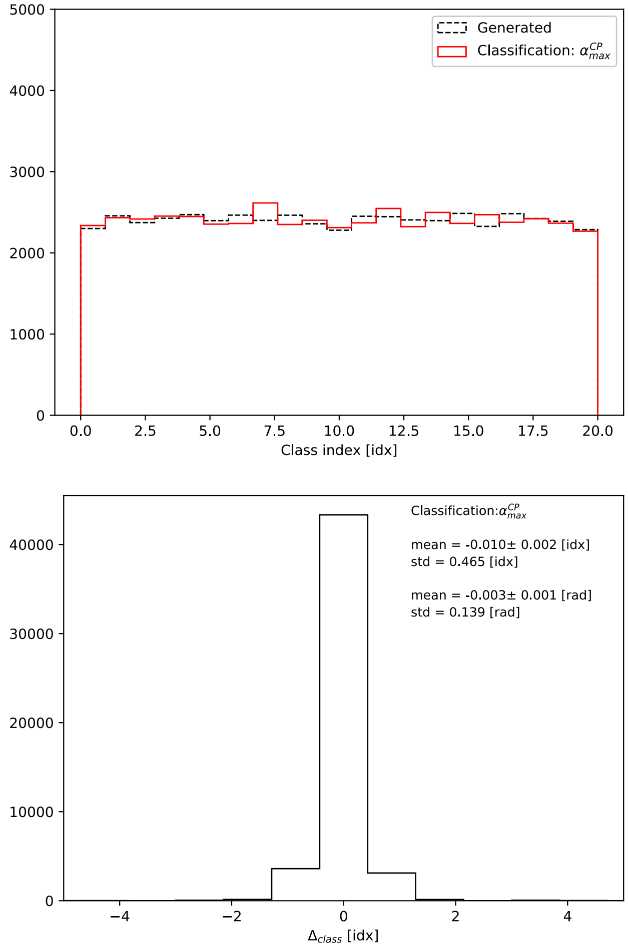

FIG. 11. Distributions (top plot) of true and predicted most preferred mixing angle α^{CP}. The distribution of per-event difference of the two is shown on the bottom plot. The granularity of $N_{\mathrm{class}} = 21$ was used for training *DNN*.

(iii) The *DNN* is learning to predict per-event most probable mixing angle $\alpha^{CP}_{\mathrm{max}}$, i.e., where α^{CP} spin weight has maximum. This approach is described in Sec. VI C and used for the figures labeled with: `Regression:` $\alpha^{CP}_{\mathrm{max}}$.

We continue with the TENSORFLOW package, but now with *tf.losses.mean_squared_error* function as a *loss* in the training procedure of Secs. VI A, VI B and self-defined function in the training procedure of Sec. VI C. Mentioned self-defined function is discussed in the Appendix.

A. Learning spin weight *wt*

Similarly as in the classification case, the *DNN* regression is trained on an input information consisting of per-event feature list. As a training output we provide a vector of the spin weight wt_i for the discrete values of α^{CP}. Training is performed for different granularities of

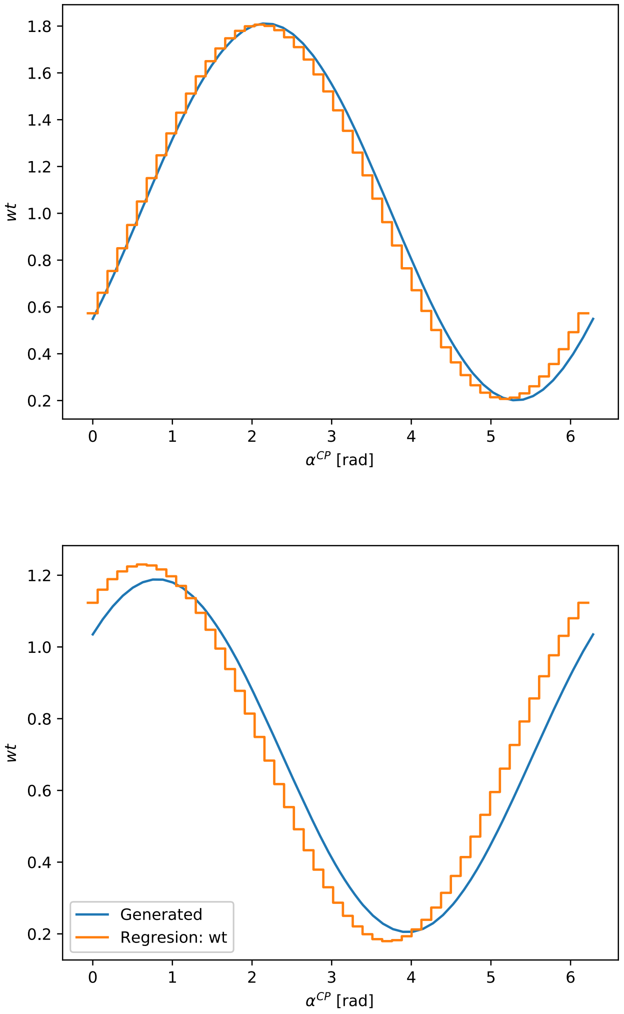

FIG. 12. Example plots with *DNN* regression results: the spin weight *wt*, predicted (orange steps) and true (blue line), as a function of α_i^{CP} for two example events (top and bottom plots). *DNN* was trained with $N_{\text{class}} = 51$ spanning range $(0, 2\pi)$.

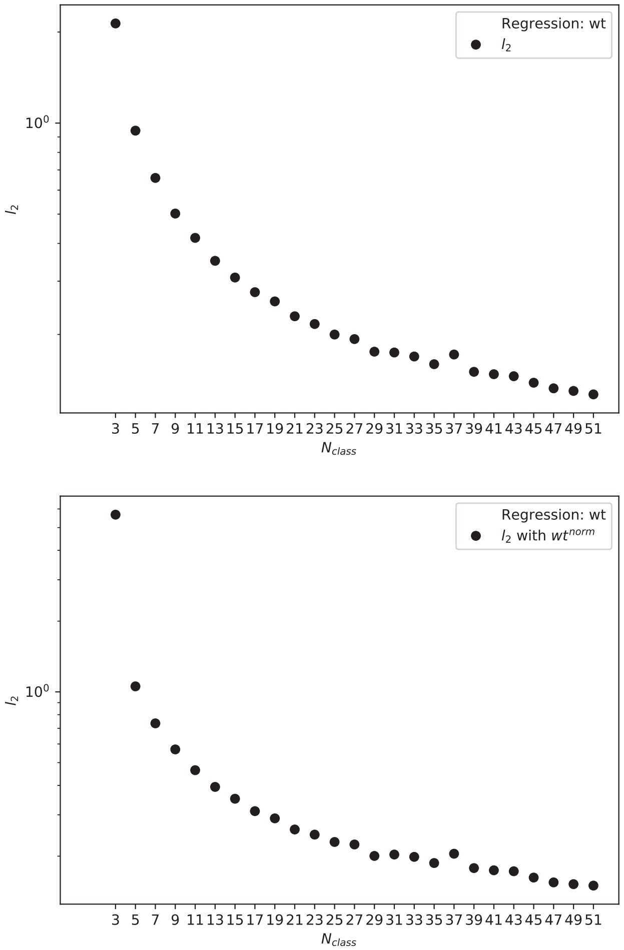

FIG. 13. The l_2 norm for predicted spin weight *wt* (top) and wt^{norm} (bottom) as a function of N_{class}.

α^{CP} discretization, to monitor performance sensitivity. Again in this case we use odd number of equally spaced points α_i^{CP}, so the $\alpha^{CP} = 0, \pi, 2\pi$ coincide with a single point. It is worth noting, that in case of regression, both shape and normalization of the *wt* are learned by the *DNN*.

For two example events in Fig. 12, true continuous spin weight *wt* distribution as well as step-function prediction is shown as a function of mixing parameter α^{CP}. In overall, predicted weights follow smoothly expected shape of linear $\cos(\alpha^{CP})$ and $\sin(\alpha^{CP})$ combination, even if no attempt to regularize for such smooth behavior was made.

Distributions of l_2 norm, defined in the same way as in the classification case, as a function of N_{class} (granularity for discretising α^{CP}) is shown in Fig. 13. For more compatibility with the classification case of Sec. V A we present results for original *wt*, as well as normalized to unity wt^{norm}. The results are comparable, with a visible flattening of l_2 for higher values of N_{class}.

In Fig. 14 distributions of Δ_{class} for $N_{\text{class}} = 21$ and 51 used to train *DNN* regression are respectively shown. The shape is Gaussian-like and as expected centered around $\Delta_{\text{class}} = 0$.

B. Learning C_0, C_1, C_2 coefficients

Regression approach allows us to predict C_0, C_1, C_2 coefficients directly, without any need of discretization. The differences between true and predicted ones are shown in Fig. 15. On average, all three coefficients are predicted reasonably well. Consistent are the statistical summaries of ΔC_i: means remain in the range ±0.004 and standard deviations in range (0.029–0.042). Coefficients C_i are then used to calculate predicted spin weight *wt* of formula (5).

We have investigated also, how well predicted C_0, C_1, C_2 can be used to estimate the most preferred mixing angle, α_{max}^{CP}. For consistency, we evaluate it using the

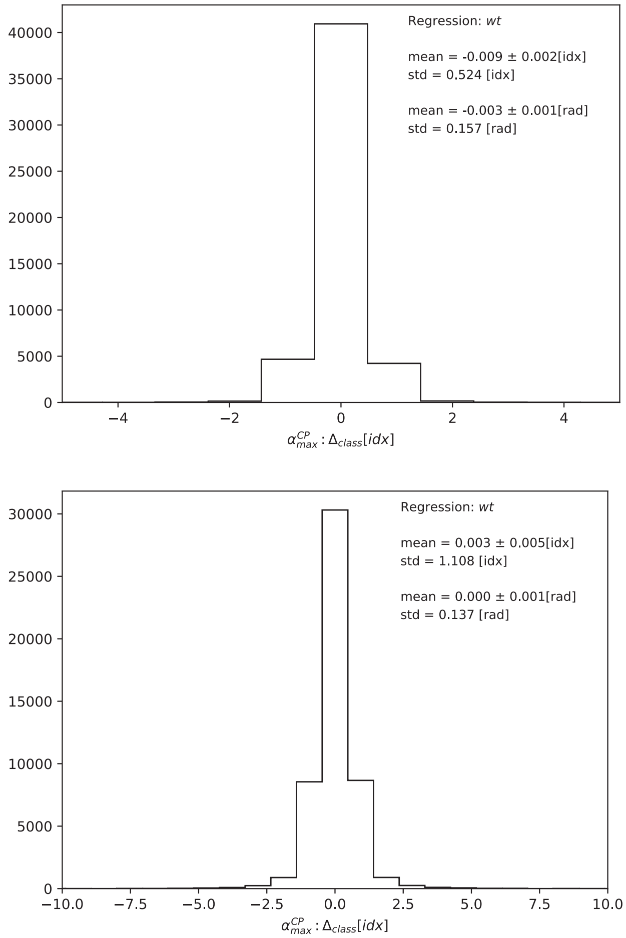

FIG. 14. Distribution of Δ_{class} between most probable predicted class and true most probable class. The $N_{\mathrm{class}} = 21$ and 51 are used for respectively top and bottom plot. The `mean` and `std` standard deviation are calculated in units of class index [idx] and units of radians [rad].

same criteria as for classification approaches. This is achieved by using coefficients C_0, C_1, C_2 to calculate spin weight wt, and then turning it into discrete predictions for wt and wt^{norm} in the N_{class} points. As in Sec. V for classification approach, we use Δ_{class}, defined by formulas (8)–(10).

The distributions of the true and predicted most probable class, $\alpha^{CP}_{\mathrm{max}}$ and their difference are shown in Fig. 16 for the $N_{\mathrm{class}} = 51$. We expect the distributions to be flat as sample was generated without any polarization correlation (carrier of CP effects) included, and this sanity check seems to be positive. The difference between true and predicted $\alpha^{CP}_{\mathrm{max}}$ forms a narrow peak with the mean value $\langle \Delta\alpha^{CP}_{\mathrm{max}} \rangle = -0.001 \pm 0.001$ [rad] and standard deviation 0.138 [rad].

Finally, as a sanity check, we have compared the true overall distribution of C_0, C_1, C_2 with the predicted one. As we can see in Fig. 17, both distributions match very well.

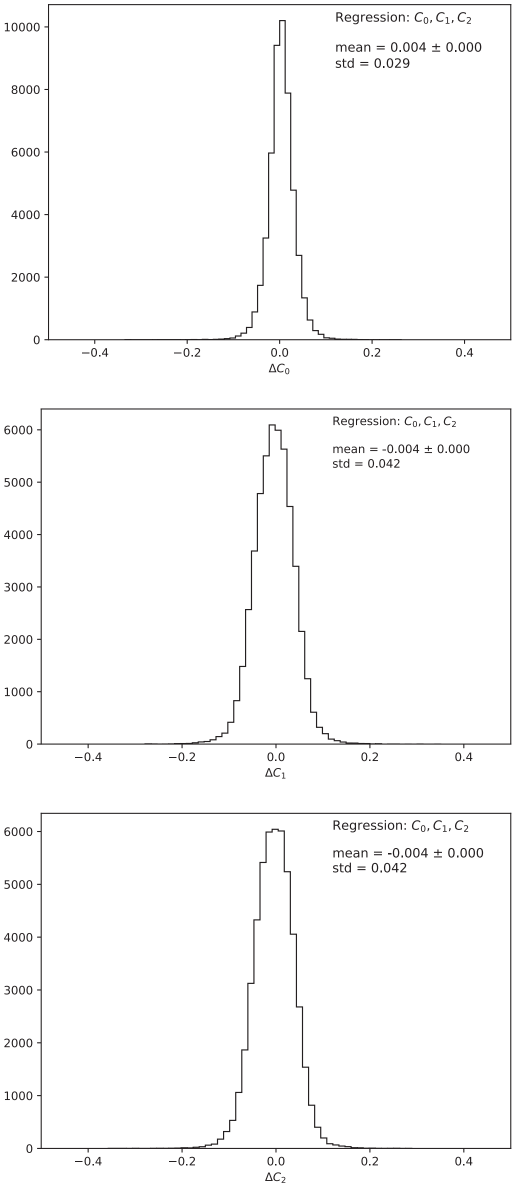

FIG. 15. Difference between true and predicted coefficients C_0, C_1, C_2 of formula (5).

C. Learning the $\alpha^{CP}_{\mathrm{max}}$

As was in the previous subsection, the implementation of the regression method allows a direct, nondiscrete estimation of continuous parameters. This is also desired with the most preferred mixing angle $\alpha^{CP}_{\mathrm{max}}$.

The distributions of the true and predicted most probable class, $\alpha^{CP}_{\mathrm{max}}$ and their difference are shown in Fig. 18 for the $N_{\mathrm{class}} = 51$. We expect the distributions to be flat as sample was generated without any polarization correlation (carrier of CP effects) included, and this sanity check seems to be positive. As the used event sample is generated without any polarization, the distribution of the $\alpha^{CP}_{\mathrm{max}}$ is expected to be uniform, see the top plot of Fig. 18. The DNN is reproducing this feature well. The difference between true

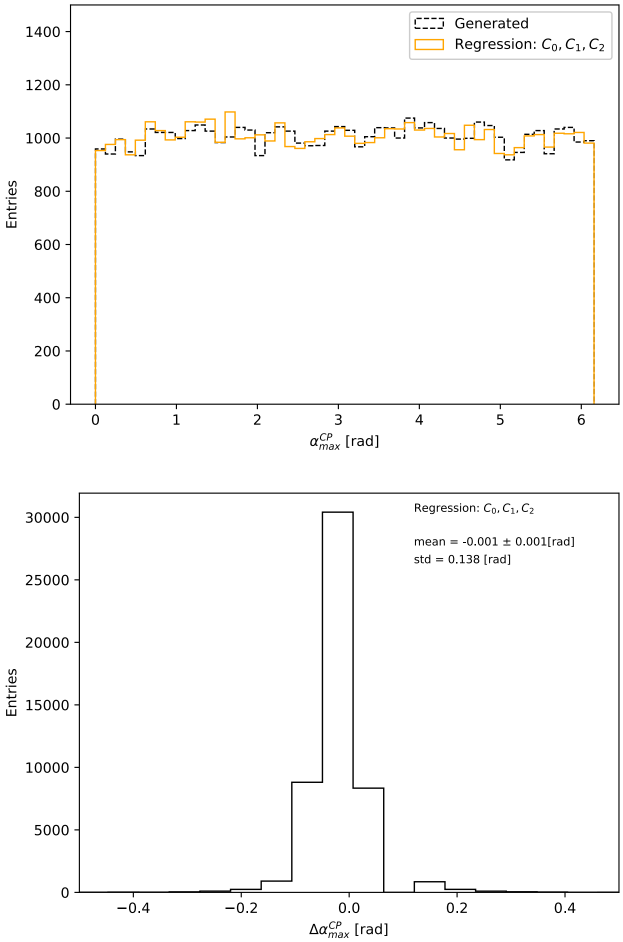

FIG. 16. Distributions (top plot) of true (black dashed line) and predicted (orange line) most preferred mixing angle α^{CP}. The prediction was based on coefficients C_0, C_1, C_2. The distribution of per-event difference of the two is shown on the bottom plot.

and predicted $\alpha^{CP}_{\max}$ forms a narrow peak with the mean $\langle\Delta\alpha^{CP}_{\max}\rangle = 0.020 \pm 0.003$ [rad] and standard deviation 0.458 [rad].

VII. CLASSIFICATION OR REGRESSION: COMPARISON AND COMPLEMENTARITY

In this section we shortly compare classification and regression approaches. In Table III we collect the mean and standard deviation for difference between true and predicted with classification and regression methods C_i. There is no clear winner, both methods give predictions of similar precision, with only C_0 being better predicted with regression.

In Table IV we compare the difference between true and predicted $\alpha^{CP}_{\max}$ obtained with different methods. With the classification method comparable performance is achieved when learning spin weight wt, coefficients C_0, C_1, C_2 or directly $\alpha^{CP}_{\max}$. For the regression method learning directly $\alpha^{CP}_{\max}$ is significantly worse performing. Otherwise, there is no clear winner between different methods.

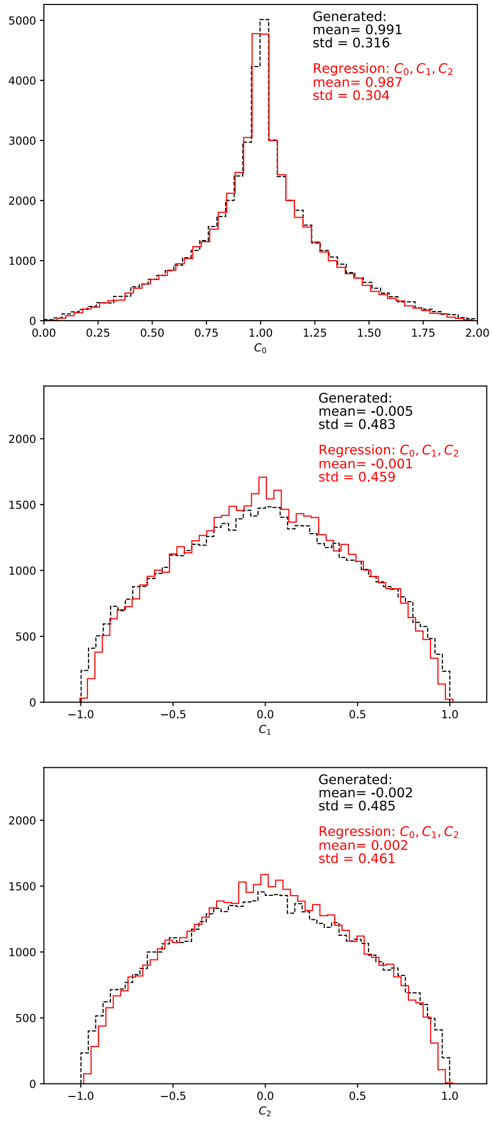

FIG. 17. Distributions of true and predicted coefficients C_0, C_1, C_2 of formula (5).

In Ref. [18] we have compared performance of the NN method with the one which can be deduced from the set of one dimensional optimal-variable-like distributions. Assuming partial correlations the classification strength was comparable, somewhat smaller though. This conclusion holds now as well. Note that optimal-like subvariables may turn to be useful for evaluation of systematic ambiguities. For one dimensional distributions established strategies are ready to be used.

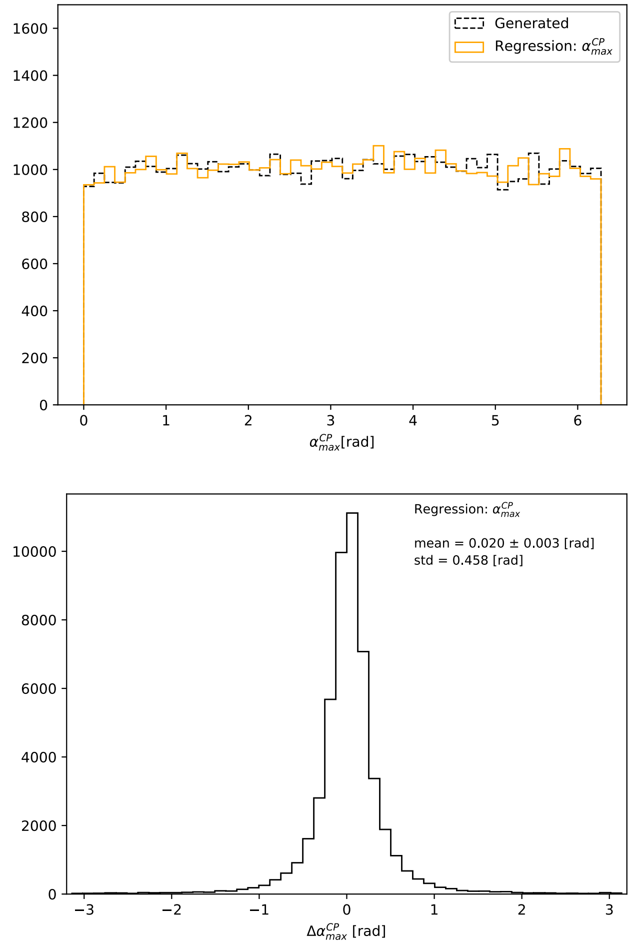

FIG. 18. Distributions (top plot) of true (black dashed line) and predicted (orange line) most preferred mixing angle α^{CP}. The distribution of per-event difference of the two is shown on the bottom plot.

VIII. SUMMARY

We have performed a proof-of-concept for the *DNN* methods in the measurement of Higgs boson $H \to \tau\tau CP$ mixing angle dependent coupling. That extends work of Refs. [18,19] of classification between scalar and pseudoscalar Higgs *CP* state. Several solutions of classification and of regression types were prepared and numerical results were collected. For the measurement we have studied approaches where; (i) spin weights, (ii) coefficients for the functional form of the spin weight (iii) directly the mixing angle at which the weight has its maximum, were targeted. In cases (i) and (ii) the classification approach seemed comparable to the regression, but the comparisons relied on the discretized and normalized quantities due to classification limitations. The regression approach seems more natural for continuous observables and does not have such limitations. On the other hand, regression approach has performed much worse in the case of direct $\alpha^{CP}_{\mathrm{max}}$ prediction.

TABLE III. The mean and standard deviations of ΔC_i, the difference between generated and predicted C_i, obtained from *DNN* with classification and regression methods for $N_{\mathrm{class}} = 51$.

Coefficients	Classification	Regression

ΔC_0	mean = 0.000
	std = 0.038
		mean = 0.004
		std = 0.029

ΔC_1	mean = 0.001
	std = 0.051
		mean = −0.004
		std = 0.042

ΔC_2	mean = −0.003
	std = 0.051
		mean = −0.04
		std = 0.042

TABLE IV. The mean and standard deviation of $\Delta\alpha^{CP}_{\mathrm{max}}$, the difference between true and predicted $\alpha^{CP}_{\mathrm{max}}$, obtained from *DNN* with classification and regression methods.

Method	Classification	Regression

Using *wt*	mean = −0.006 ± 0.001 [rad]
	std = 0.126 [rad]
		mean = 0.000 ± 0.001 [rad]
		std = 0.137 [rad]

Using C_0, C_1, C_2	mean = 0.000 ± 0.001 [rad]
	std = 0.153 [rad]
		mean = −0.001 ± 0.001 [rad]
		std = 0.138 [rad]

Direct	mean = −0.003 [rad]
	std = 0.139 [rad]
		mean = 0.020 [rad]
		std = 0.458 [rad]

For the feature list we have chosen idealistic case, assuming that complete set of τ decay products 4-momenta is known, including challenging to reconstruct neutrinos. We have exploited then the $\tau \to \rho\nu$ decay mode. The results are encouraging, the understanding of environment for future discussion of measurement ambiguities was not compromised with respect to what was achieved in previous publications for scalar/pseudoscalar classifications.

The mean value of the preferred mixing angle $\alpha^{CP}_{\mathrm{max}}$ can be constrained by the trained *DNN* with per-event resolution better than 0.15 [rad] using a classification approach. Both classification and regression approaches allow to learn spin weight with uncertainties (average l_2 norm) better than 15%. Both approaches allow also to learn coefficients C_0, C_1, C_2 of the functional spin weight form. The coefficients are directly related to the polarimetric vectors of decaying $\tau^{\pm}$ leptons. This provides interesting possibility for the future studies of experimental ambiguities with samples of the $Z \to \tau\tau$ decays, much more abundant and available for the LHC measurements. Departure from SM predictions on $Z\tau\tau$ coupling can reveal itself in the observables build from polarimetric vectors of decaying $\tau^{\pm}$ leptons too.

We plan, following [18,19], to extend our studies to more realistic feature lists and other decay modes. Already now, the variety of ML methods for the determination of most preferred *CP* state mixing angle, demonstrated potential and robustness for future experimental analyses and measurements with the LHC data.

While analyzing the realistic feature list, one can, in parallel, construct one dimensional distributions which feature partial sensitivity to *CP*. This can be used to monitor, which improvements are of importance, and to understand why it is the case. Also, by hand combination of such partial significance can be helpful to understand performance of ML solution, even if important correlations between distributions are then ignored. On the other hand, it can help to understand consequences of background contaminations, which may mimic the signatures. In Ref. [27] we have discussed potential consequences of such contaminations, originating from Drell-Yan background, but to exhaust, more complete studies with detector simulation details are to be applied, and not only to optimal-like distributions. In that paper, we have concentrated on a tool, potentially useful for background impact on multidimensional distributions.

Background contamination is a valid concern for the discussed signatures. It may even mimic the signal. That it is not the case needs to be checked. In Ref. [20] we have studied the transverse spin effects of the Drell Yan process; a good example of the background for *CP* sensitive observables of Higgs to τ decays. It was found, see Fig 4. there, that because sign of transverse spin correlation depends on orientation with respect to reaction plane, spanned on incoming quarks and outgoing τ leptons momenta, the effect cancels out if averaged over the sample. That observation trivially extends to the case of other τ decays and signatures prepared with the help of ML techniques. Only very specific cuts could change that. For that however very detailed knowledge of the detector responses and cut implementation would be needed. This is to be achieved only with the help of collaborations detector response software. Our intuition is that such cuts are rather not an issue. The background will just contribute with events of, on average, no *CP*-sensitive signatures of any sort. That should be the case of any processes mediated by intermediate spin 1 state, and τ-pairs of no common origin as well. In presence of such background, we expect sensitivity of observables based on optimal variables and ML techniques alike, to be reduced in proportion to possible background fluctuations.

ACKNOWLEDGMENTS

We would like to thank J. Kurek and P. Winkowska for help with technical implementation and testing of the analysis code used for this paper preparation. This project was supported in part from funds of Polish National Science Centre under decisions No. DEC-2017/27/B/ST2/01391. The majority of the numerical calculations were performed at the PLGrid Infrastructure of the Academic Computer Centre CYFRONET AGH in Krakow, Poland.

APPENDIX: DEEP NEURAL NETWORK

The structure of the simulated data and the *DNN* architecture follows what was published in our previous papers [18,19]. It is prepared for TENSORFLOW [31], an open-source machine learning library.

We consider $H \to \tau\tau$ channel of both $\tau^{\pm} \to \rho^{\pm}\nu$ decay. The data point is thus an event of the Higgs boson production and τ lepton pair decay products. The structure of the event is represented as follows:

$$x_i = (f_{i,1}, \ldots, f_{i,D}), \qquad w_{a_i}, w_{b_i}, \ldots, w_{m_i}. \qquad \text{(A1)}$$

The $f_{i,1}, \ldots, f_{i,D}$ represent numerical features and $w_{a_i}, w_{b_i}, w_{m_i}$ are weights proportional to the likelihoods that an event comes from a class $A, B, \ldots, M$, each representing different α^{CP} mixing angle. The $\alpha^{CP} = 0, 2\pi$ corresponds to scalar CP state and $\alpha^{CP} = \pi$ to pseudoscalar CP state. The weights calculated from the quantum field theory matrix elements are available and stored in the simulated data files. This is a convenient

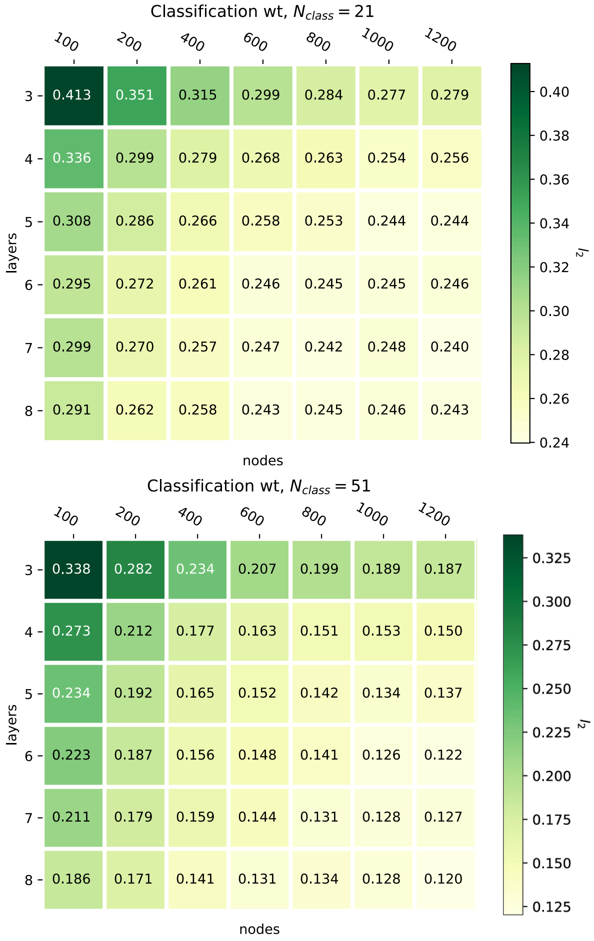

FIG. 19. Performance of *wt* fitting (l_2) for different number of layers and nodes, assuming $N_{\rm class} = 21$ (top) and $N_{\rm class} = 51$ (bottom).

situation, which does not happen in many other cases of ML classification. The $A, B, \ldots M$ distributions highly overlap in the $(f_{i,1}, \ldots, f_{i,D})$ space, the more detailed discussion in case of two hypotheses only, scalar and pseudoscalar, can be found in [18].

Thanks to similar *DNN* architecture, we have prepared three implementations for measuring Higgs boson *CP* state: binary classification, multiclass classification and regression:

(i) For binary classification the aim is to discriminate between two hypothesis, $\mathcal{H}_0$ and $\mathcal{H}_{\alpha^{CP}}$.

(ii) For multiclass classification, the aim is to simultaneously learn weights (probabilities) for several $\mathcal{H}_{\alpha^{CP}}$ hypotheses; learn coefficients of the weight functional form or directly learn the mixing angle at which spin weight has its maximum, $\alpha^{CP}_{\max}$. A single class can be either single discretized α^{CP} or a range

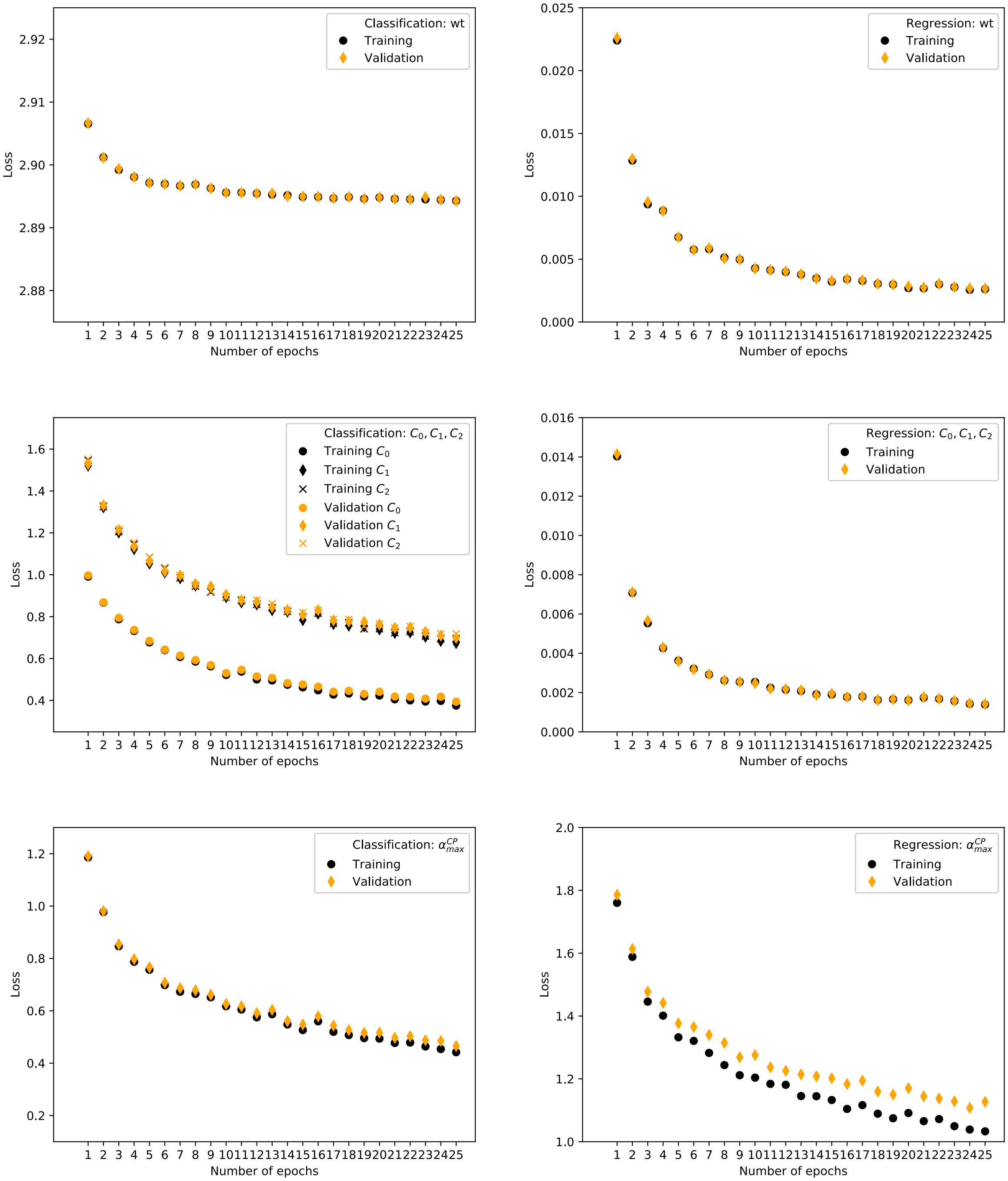

FIG. 20. The *DNN* loss for classification (left-side) and regression (right-side), as function of number of epochs used for training. It is shown for learning spin weight (top plots), C_i coefficients (middle plots) and most likely mixing angle $\alpha^{CP}_{\max}$ (bottom plots). For the classification, $N_{\rm class} = 21$ was used.

for the C_i parameters. The system is learning probabilities for classes to associate with the event.

(iii) For the regression case, the aim is similar as for multiclass classification case, but now the problem is defined as a continuous case. The system is learning value to associate with the event. The value can be a vector of spin weights for a set of $\mathcal{H}_{\alpha^{CP}}$ hypotheses, set of C_i coefficients or $\alpha^{CP}_{\max}$.

The network architecture consists of 6 hidden layers, 1000 nodes each with ReLU activation functions and is initialized with random weights. Such architecture has been found as a good trade-off between the performance and computation time, what can be seen in Fig. 19. Learning procedure is optimized using a variant of stochastic gradient descent algorithm called Adam [32] and batch normalization [33].

The last layer is specific to the implementation case, different is dimension of the output vector, activation function and a loss function. In the following, we will describe details.

Classification: The loss function used in stochastic gradient descent is a cross entropy of valid values and neural network predictions [4]. It is a common choice in case of binary or multiclass classification models. The loss function for sample of N_{evt} events and classification for N_{class} reads as follows:

$$\text{Loss} = \sum_{k=1}^{N_{\text{evt}}} \sum_{i=1}^{N_{\text{class}}} y_{i,k} \log(p_{i,k}), \tag{A2}$$

where k stands for consecutive event and i for class index. The $y_{i,k}$ represents neural-network predicted probability for event k being of class i while $p_{i,k}$ represents true probability used in supervised training.

Regression: In case of predicting *wt* the last layer of *DNN* is N dimensional output (granularity with which we want to discretize it). For predicting C_0, C_1, C_2 the last layer of *DNN* is $N = 3$ dimensional output, i.e., values of C_0, C_1, C_2. Activation of this layer is a linear function. Loss functions is defined as mean squared error (MSE) between true and predicted parameters

$$\text{Loss} = \sum_{k=1}^{N_{\text{evt}}} \sum_{i=1}^{i=N} (y_{i,k} - p_{i,k})^2, \tag{A3}$$

where k stands for event index and i for index of function form parameter. The $y_{i,k}$ represents predicted value of C_ith parameter for event k while $p_{i,k}$ represents true value. For predicting the $\alpha^{CP}_{\max}$ the last layer of *DNN* is $N = 1$ dimensional output, i.e., values of $\alpha^{CP}_{\max}$.

The `tf.reduce_mean` method of TENSORFLOW is used, with the loss function

$$\text{Loss} = \sum_{k=1}^{N_{\text{evt}}} (1 - \cos(y_k - p_k)), \tag{A4}$$

where y_k, p_k denotes respectively predicted and true value of $\alpha^{CP}_{\max}$.

In Fig. 20, for all problems considered, distributions of the loss functions on the training and validation samples, as a function of number of epochs used for training are shown. Left plots are for the classification and right plots for the corresponding regression. The values of the loss are case specific and should not be directly compared, their shape is monitoring the training process. For all cases the loss is decreasing with number of epochs, both on training and validation samples. It is overlapping for all cases except [`Regression`: $\alpha^{CP}_{\max}$] (bottom right plot), for that single case one small loss in performance is observed for validation sample compared to training sample. Training with 25 epochs seems sufficient for both classification and regression for all presented scenarios. We have not observed any gain of interest in any of extended to larger number of epochs Fig. 20 plots.

[1] D. Guest, K. Cranmer, and D. Whiteson, Annu. Rev. Nucl. Part. Sci. **68**, 161 (2018).

[2] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, and L. Zdeborová, Rev. Mod. Phys. **91**, 045002 (2019).

[3] K. Albertsson *et al.*, J. Phys. Conf. Ser. **1085**, 022008 (2018).

[4] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, Cambridge, MA, 2017).

[5] J. Lee, N. Chanon, A. Levin, J. Li, M. Lu, Q. Li, and Y. Mao, Phys. Rev. D **99**, 033004 (2019).

[6] J. Searcy, L. Huang, M.-A. Pleier, and J. Zhu, Phys. Rev. D **93**, 094033 (2016).

[7] S. Forte, L. Garrido, J. I. Latorre, and A. Piccione, J. High Energy Phys. 05 (2002) 062.

[8] M. Kramer, J. H. Kuhn, M. L. Stong, and P. M. Zerwas, Z. Phys. C **64**, 21 (1994).

[9] G. R. Bower, T. Pierzchala, Z. Was, and M. Worek, Phys. Lett. B **543**, 227 (2002).

[10] A. Rouge, Phys. Lett. B **619**, 43 (2005).

[11] K. Desch, Z. Was, and M. Worek, Eur. Phys. J. C **29**, 491 (2003).

[12] S. Berge and W. Bernreuther, Phys. Lett. B **671**, 470 (2009).

[13] S. Berge, W. Bernreuther, and S. Kirchner, Phys. Rev. D **92**, 096012 (2015).

[14] CMS Collaboration, Report No. CMS-PAS-HIG-20-006.

[15] ATLAS Collaboration, Report No. ATL-PHYS-PUB-2019-008.

[16] J. H. Kuhn, Phys. Rev. D **52**, 3128 (1995).

[17] M. Davier, L. Duflot, F. Le Diberder, and A. Rougé, Phys. Lett. B **306**, 411 (1993).

[18] R. Jozefowicz, E. Richter-Was, and Z. Was, Phys. Rev. D **94**, 093001 (2016).

[19] K. Lasocha, E. Richter-Was, D. Tracz, Z. Was, and P. Winkowska, Phys. Rev. D **100**, 113001 (2019).

[20] T. Przedzinski, E. Richter-Was, and Z. Was, Eur. Phys. J. C **74**, 3177 (2014).

[21] A. Martın *et al.*, Software available from tensorflow.org (2015).

[22] K. Desch, A. Imhof, Z. Was, and M. Worek, Phys. Lett. B **579**, 157 (2004).

[23] E. Barberio, B. Le, E. Richter-Was, Z. Was, D. Zanzi, and J. Zaremba, Phys. Rev. D **96**, 073002 (2017).

[24] 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).

[25] N. Davidson, G. Nanava, T. Przedzinski, E. Richter-Was, and Z. Was, Comput. Phys. Commun. **183**, 821 (2012).

[26] Z. Czyczula, T. Przedzinski, and Z. Was, Eur. Phys. J. C **72**, 1988 (2012).

[27] T. Przedzinski, E. Richter-Was, and Z. Was, Eur. Phys. J. C **79**, 91 (2019).

[28] A. P. B. Bradley, Pattern Recognit. **30**, 1145 (1997).

[29] T. Fawcett, Pattern Recogn. Lett. **27**, 861 (2006).

[30] The wt_i remains in the (0, 4) range, as explained in [27].

[31] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, Software available from tensorflow.org **1** (2015).

[32] D. Kingma and J. Ba, arXiv:1412.6980.

[33] S. Ioffe and C. Szegedy, arXiv:1502.03167.