Distinguishing W' signals at hadron colliders using neural networks

Spencer Chang,[1] Ting-Kuo Chen,[2] and Cheng-Wei Chiang[2,3]

[1]*Department of Physics and Institute for Fundamental Science University of Oregon, Eugene, Oregon 97403, USA*

[2]*Department of Physics, National Taiwan University, Taipei 10617, Taiwan*

[3]*Physics Division, National Center for Theoretical Sciences, Taipei, Taiwan 10617, Republic of China*

(Received 8 August 2020; accepted 28 January 2021; published 24 February 2021)

We investigate a neural-network-based hypothesis test to distinguish different W' and charged scalar resonances through the $\ell + \not{E}_T$ channel at hadron colliders. This is traditionally challenging due to a fourfold ambiguity at proton-proton colliders, such as the Large Hadron Collider. Of the neural network approaches we study, we find a multiclass classifier based on a fully connected neural network trained upon two-dimensional histograms made from kinematic variables of the final state ℓ to be the most powerful. Furthermore, by considering the one-jet processes, we demonstrate that one can generalize to multiple two-dimensional histograms to represent different variable pairs. Finally, as a comparison to traditional approaches, we compare our method with Bayesian hypothesis testing and discuss the pros and cons of each approach. The neural network scheme presented in this paper is a powerful tool that can help probe the properties of charged resonances.

DOI: 10.1103/PhysRevD.103.036016

I. INTRODUCTION

Ever since the discovery of the W boson through the $e\nu$ decay channel in 1983 at the SPS collider [1,2], the search for W' and other charged boson resonances has continued. The latest analyses include the 13 TeV search in the dijet channel [3–5], the dijet + lepton channel [6], the $\ell + \not{E}_T$ [7,8] channel, the $\tau\nu$ channel [9], and diboson channels [10,11] conducted by ATLAS and CMS. So far, the mass limit for sequential W' has been pushed above the TeV level (see Ref. [12]), and thus future W' signals are expected to occur at higher masses in high-energy hadron colliders. One such example is the CERN Large Hadron Collider (LHC), which is the main focus of our study. In this case, the leptonic search turns out to be a favorable choice, as it avoids the large QCD background. Some of the most important properties to be identified of a W' would be the mass, decay width, and couplings to the Standard Model (SM) fermions; if we further include the study of charged scalar bosons, spin would also be important. However, determining the boson's couplings and spin in its c.m. frame at the LHC suffers from two ambiguities:

(1) *Unknown initial state*: To study the Lorentz structure of a charged-current interaction, the incident partons must be identified so as to define the forward direction (e.g., in the quark direction, not the antiquark direction.). Due to the parton distribution functions (PDFs), the best one can do is to make a reasonable guess for this from the PDF properties [13].

(2) *Missing longitudinal momentum*: Since the c.m. frame of the colliding partons is typically boosted, we need to identify the missing longitudinal momentum associated with the neutrino to correctly determine the c.m. angular distribution in $\cos\theta_{\text{c.m.}}$. From kinematics, the longitudinal momentum can be solved from a quadratic equation assuming that the mediating boson is on-shell, but there is no event-by-event information that can be used to determine which of the two quadratic solutions is correct. This ambiguity has already been pointed out in several studies involving $\not{E}_T$, such as the reconstruction of $W \to e\nu$ at the SPS $p\bar{p}$ Collider [1] and top-pair production at the Tevatron [14].

Even though the mentioned ambiguities have imposed an obstacle to such studies, several studies based on traditional approaches have still been conducted to reconstruct the information of the W', such as Refs. [15–19].

In this paper, we investigate deep-learning-based approaches to tackle the problem of determining the spin and interaction type of a heavy charged boson resonance through its leptonic decay channels. In particular, we will consider W' and H, generic spin-1 and spin-0 charged resonances, respectively. Over the past few years, neural

networks have made enormous strides in a variety of challenging problems in different fields. Some recent high-energy physics applications include Refs. [20–28].

The above ambiguities make event-by-event reconstruction by a neural network challenging, but classification based on a collection of events can still have significant distinguishing power. Bosons with different leptonic couplings and spins will manifest distinctive kinematic features which become apparent as one accumulates events. Thus, instead of trying to reconstruct the spins and couplings directly, we can use a multiclass neural network classifier that takes measured lab quantities of a set of events as input. There are two straightforward ways to input this collection of events: either simply feed them in event by event as an array, or combine a number of events and form a two-dimensional (2D) histogram of a selected pair of variables. The latter would be similar to feeding in part of the probability density function on the chosen 2D kinematic plane. Based upon these possibilities, we have considered the following three neural network (NN) models for this problem:

(1) *Fully connected neural network upon individual events* (FNNi): We constructed a fully connected neural network (FNN) trained upon the kinematic information of individual events. To utilize the scores of this network for hypothesis testing on a group of accumulated events, we use the normalized class score product of the group.
(2) *Fully connected neural network upon histograms* (FNNh): We constructed an FNN trained upon flattened 2D histograms made from pairs of kinematic observables of a certain number of events.
(3) *Convolutional neural network* (CNN): We also constructed a CNN trained upon the 2D histograms mentioned above.

These methods have already been proposed and used in Ref. [27][1] to distinguish the monojet and dijet signatures of weakly interacting massive particles from those of the SM and other dark matter models. In our study, we investigate the application of these methods to the classification of samples into the following three coupling classes:[2]

(1) *Vector/axial* (VA): This class corresponds to a W' with vector-like (V) fermionic couplings, $W'_\mu\bar{\psi}\gamma^\mu\chi$, or axial-vector-like (A) fermionic couplings, $W'_\mu\bar{\psi}\gamma^\mu\gamma_5\chi$.
(2) *Chiral* (CH): This class corresponds to a W' with left-handed (LH) fermionic couplings, $W'_\mu\bar{\psi}\gamma^\mu(1-\gamma_5)\chi$, or right-handed (RH) fermionic couplings, $W'_\mu\bar{\psi}\gamma^\mu(1+\gamma_5)\chi$.
(3) *Scalar* (SC): This class corresponds to an $H^\pm$ with Yukawa fermionic couplings, $H\bar{\psi}\chi$ and $H\bar{\psi}\gamma_5\chi$.

For a pp collider, we will show that for signal alone the p_T and η variables of the lepton cannot distinguish between the V and A hypotheses or between the LH and RH hypotheses. Interference between a W' and the SM W background could in principle break this degeneracy, yet such effects are found to be negligible for the TeV-mass bosons considered in this study. Thus, under our approximations the VA, CH, and SC hypotheses comprise three distinct signals.

We prepare the samples assuming 14 TeV pp collisions, which is the expected c.m. energy of the HL-LHC. Going beyond the signal-only hypothesis testing of Ref. [27], we will also include the SM background from the W boson. We will investigate scenarios of different S/B ratios, assuming an HL-LHC integrated luminosity of $\mathcal{L} = 3$ ab^{-1}.

To choose the masses for our study, we use Ref. [8] to determine the 95% C.L. cross section upper limits of different charged resonance masses. Since we anticipate that our technique requires $S/B \gtrsim 1$ to be effective, we consider masses where the current cross section limits allow $S/B \sim 1$ and where we can still expect to get a 5σ discovery in the HL-LHC era. These conditions force the mass to be $\geq$4.5 TeV, so we will focus on the mass 4.5 TeV. As a comparison, we will also explore 6 TeV resonances, where the signal purity can be higher but the hypothesis testing is more challenging due to low statistics.

We only study the $e\nu$ decay channel, though this method can be readily applied to the $\mu\nu$ channel and improve its efficiency. Also, we assume that the coupling strength and structure are universal to all generations in both the quark and lepton sectors (even for $H^\pm$).

We also take into consideration the effects of different boson resonance widths, using the values 500, 200, and 50 GeV for the 4.5 TeV resonances. It is observed that the training outcomes upon different widths are quite similar. We will focus on the samples with a width of 200 GeV, chosen to mimic the SM W width-to-mass ratio $\Gamma_W/m_W \approx 1/40$, in most of our presentation below. As for the 6 TeV resonances, we only study the case with a width of 300 GeV.

Beside the zero-jet process, we have also studied the one-jet process in which an extra jet is included in the final state. Since in real experiments jets can be copiously produced through either soft radiation or hard interactions, we consider all processes of jet multiplicities up to 2, and extract from them the zero-jet and one-jet samples with criteria to be mentioned in Sec. III. To make use of the extra information provided by the jet, we will further extend the 2D histogram inputs to include more variable pairs by using "RGB" colors to demonstrate that the histogram approach of Ref. [27] can be generalized to higher dimensions.

[1]We note that there are several other studies that also use ensembles of events and/or multidimensional histograms to perform machine learning. See Refs. [29–34].

[2]These are the interactions familiar to us in the SM. The proposed method can be generalized to include other interactions, such as other linear combinations of $\sim W'_\mu\bar{\psi}\gamma^\mu(a+b\gamma_5)\chi$. The discriminating power, of course, will depend upon how close the different coupling classes are.

We will formulate a few different input schemes for these one-jet histograms, although there is no major performance difference among them. To understand the results, we will also study the importance and contributions of the different variable pairs in these schemes. It is worth noting here that for situations involving more kinematic variables like the current study, our results show that the NN approach is more convenient than and superior to conventional methods, such as Bayesian hypothesis or χ^2 tests.

In the Appendix, we further provide detailed technical studies of the NN performance when the bin resolution and kinematic window are varied. In addition, we compare the performances of binary classifiers to those of the original ternary classifiers by performing a *projection* on the testing scores of the latter, which demonstrates that our ternary classifier is as capable as individual binary classifiers. Finally, we investigate the results of applying to the testing samples models trained for an incorrect assumption of significance or decay width, testing the flexibility of our methods.

This paper is organized as follows. In Sec. II, we briefly review the kinematic properties of bosons of different coupling classes. In Sec. III, we discuss the zero-jet and one-jet samples and analyze their kinematic features. In Sec. IV, we describe the details of our NN models as well as the training specifications. In Sec. V, we present and discuss the zero-jet and one-jet training results. In Sec. VI, we compare our NN method with the Bayesian hypothesis test and discuss the pros and cons. In Sec. VII, we draw conclusions and propose possible further studies. More technical details of our investigations are provided in the Appendix.

II. PARTON-LEVEL ANALYSIS OF GENERAL SINGLY CHARGED BOSONS

Consider the following processes:

$$pp \to W/W'/H \to e\nu_e. \tag{1}$$

The corresponding p_T and η differential cross sections of e are given by

$$\frac{d\sigma}{d\chi} = \sum_{q,q'} \int dxdy \frac{d\hat{\sigma}(x,y)}{d\chi} \cdot q(x,Q^2)\bar{q}'(y,Q^2) \quad (\chi = p_T, \eta), \tag{2}$$

where $q(x,Q^2)$, $\bar{q}'(y,Q^2)$ are the PDFs.

The parton-level p_T and η differential cross sections for H and W' are given, respectively, by

$$\frac{d\hat{\sigma}_H}{dp_T} = \frac{1}{2\pi} \frac{y_H^4}{(p^2 - m_H^2)^2 + m_H^2\Gamma_H^2} \frac{p_T}{\sqrt{1 - \frac{4p_T^2}{p^2}}}, \tag{3a}$$

$$\frac{d\hat{\sigma}_{W'}}{dp_T} = \frac{1}{2\pi} \frac{2(c_V^2 + c_A^2)^2(1 - \frac{2p_T^2}{p^2})}{(p^2 - m_{W'}^2)^2 + m_{W'}^2\Gamma_{W'}^2} \frac{p_T}{\sqrt{1 - \frac{4p_T^2}{p^2}}}, \tag{3b}$$

and

$$\frac{d\hat{\sigma}_H}{d\eta} = \frac{\mathrm{sech}^2\eta}{32\pi} \frac{128E_1^2E_2^2}{(p^2 - m_H^2)^2 + m_H^2\Gamma_H^2} \cdot y_H^4 \frac{F(E_1,E_2,\eta)}{G^2(E_1,E_2,\eta)}, \tag{4a}$$

$$\begin{aligned}\frac{d\hat{\sigma}_{W'}}{d\eta} = {} & \frac{\mathrm{sech}^2\eta}{32\pi} \frac{128E_1^2E_2^2}{(p^2 - m_{W'}^2)^2 + m_{W'}^2\Gamma_{W'}^2} \\ & \times \left\{ 2(c_V^2 + c_A^2)^2 \left[\frac{I(E_1,E_2,\eta)}{H(E_1,E_2,\eta)} + \frac{I(E_2,E_1,\eta)}{H(E_2,E_1,\eta)} \right] \right. \\ & \left. + 4c_V^2c_A^2 \left[\frac{J(E_1,E_2,\eta)}{H(E_1,E_2,\eta)} + \frac{J(E_2,E_1,\eta)}{H(E_2,E_1,\eta)} \right] \right\},\end{aligned} \tag{4b}$$

where $p^2 = xys$, $E_1 = \frac{x\sqrt{s}}{2}$, $E_2 = \frac{y\sqrt{s}}{2}$, $\sqrt{s} = 14$ TeV, and F, G, H, I, J are given by

$$\begin{aligned} F(A,B,\eta) &\equiv (A+B)^2 + (A-B)^2\tanh^2\eta, \\ G(A,B,\eta) &\equiv (A+B)^2 - (A-B)^2\tanh^2\eta, \\ H(A,B,\eta) &\equiv [(A+B) - (A-B)\tanh\eta]^4, \\ I(A,B,\eta) &\equiv A^2(1-\tanh\eta)^2 + B^2(1+\tanh\eta)^2, \\ J(A,B,\eta) &\equiv A^2(1-\tanh\eta)^2 - B^2(1+\tanh\eta)^2. \end{aligned} \tag{5}$$

From these parton-level differential cross sections, one can tell H and W' apart from the p_T distributions alone. However, the W' bosons of different coupling structures would give identical p_T distributions up to the normalization $(c_V^2 + c_A^2)^2$ factor in Eq. (3b). On the other hand, the second term in the curly brackets of Eq. (4b) is proportional to $c_V^2c_A^2$ and would lead to distinct η distributions for different W' coupling scenarios. Thus, combining the *parton-level* p_T and η distributions, one should be able to readily distinguish among the three classes but cannot distinguish between V and A nor between LH and RH from the shape of the distributions alone. After convoluting with the PDFs, the distribution differences among the classes become less obvious, but will still be detectable through our technique.

III. SAMPLE GENERATION AND ANALYSIS

We prepare our parton-level samples using MadGraph5_aMC@NLO v2.7.3 [35], followed by parton shower and hadronization performed with PYTHIA 8.2.44 [36,37]. To properly interface these two softwares as we include processes of jet multiplicities of 0–2, we utilize MLM matching with a jet merging scale of 30 GeV. The cuts imposed at the generator level are summarized in Table I.

TABLE I. Summary of cuts imposed on the samples at the generator level.

Basic cuts	$p_T^j > 30$ GeV; $\lvert\eta^j\rvert < 5.0$; $\lvert\eta^\ell\rvert < 4.0$
Selection cuts	$p_T^\ell, \not{E}_T > 0.3 m_{W',H}$

The selection cut is imposed to suppress the SM W background while retaining a sufficient amount of the new-physics (NP) signals below the Jacobian peak at $p_T^\ell = m_{W'}/2 = m_H/2$. This p_T cut is a practical one so that the NN training samples are not background dominated at the low end of this cut, which assists in training while allowing our p_T binning to be sufficiently high in resolution. In the Appendix, we explore how the NN performance depends on the p_T cut and show that there can be a trade-off between information loss (too high of a cut) and p_T resolution (too low of a cut).

The samples are then passed to DELPHES 3.4.2 [38–40] for detector simulation using the Phase-II CMS card. The events are reconstructed with FastJet 3.3.2 [41]. In particular, the final-state jets are reconstructed using the anti-k_T clustering algorithm [42] with the cone radius $R = 0.4$.

The processes are simulated for 14 TeV LHC collisions with the NNPDF23_nlo_as_0119 [43] PDF set. The W'- and H-mediated processes are generated, respectively, with the Wprime model and General 2HDM model from the FeynRules [44] model database. In what follows, we describe the details of the zero- and one-jet samples.

A. Zero-jet samples

The zero-jet samples simply include all events with an observable electron and have ≥ 0 jets. For these samples, we only make use of the electron observables and ignore all of the jet information. We denote the new boson width by $\Gamma_{\rm NP}$ and consider three different values: 500, 200, and 50 GeV for 4.5 TeV resonances. We show in Sec. V that the width varying in this range does not greatly affect the training outcomes, and thus we consider only $\Gamma_{\rm NP} \approx 300$ GeV for the heavier 6 TeV resonances. At the generator level, we generate 500 000 events for each of the VA, CH, SC, and SM classes. After detector simulation, the successfully tagged event numbers of all three NP classes (including different widths and masses) and the SM class are all roughly around 300 000.

We choose to divide both p_T^e and η^e into 60 bins so that our NNs remain trainable. We only show the corresponding unit-normalized p_T^e, η^e, and p_T^e vs η^e distributions for $\Gamma_{\rm NP} \approx 200$ GeV for 4.5 TeV (left column) and 6 TeV (right column) resonances in Fig. 1. As the boson width increases, the Jacobian peak in the p_T^e distribution becomes broader, while the η^e distribution remains identical.

As discussed in Sec. II, naively the p_T curves of the VA and CH classes should be identical in Fig. 1. However, there is a slight difference between the two due to the η^e cut mentioned in Table I. Since there is a much larger difference in the η^e distributions, we do not expect this difference to strongly affect the training or performance of our classifiers. The same issue will also occur in the one-jet case.

The color scheme for Figs. 1(e) and 1(f), and also for the remaining 2D histograms, are as follows: the coldest color (blue) denotes a 0 entry, while the warmest color (red) denotes the maximum entry among all four classes. As shown in the plots, the Jacobian peaks are at around $p_T = m_{W',H}/2$ for all of the NP classes, with the CH class possessing the longest tail toward low p_T^e, while the VA and SC classes have similar tails but with different η^e distribution. Such differences in the p_T^e tail and the η^e distribution show the kinematic information that can be used to distinguish among the three classes, even after including the background.

Within the selected phase space, the expected number of SM zero-jet events are

$$B_0 = \sigma_{B_0} \times \mathcal{L} \approx \begin{cases} 84 & \text{for 4.5 TeV,} \\ 7 & \text{for 6 TeV.} \end{cases} \tag{6}$$

Thus, the total number of events we expect to observe is

$$N_0 = B_0 \times \left(1 + \frac{S_0}{B_0}\right), \tag{7}$$

where S_0 denotes the number of signal events. We will vary the signal-to-background ratio S_0/B_0 in our considerations.

We study scenarios of different S_0/B_0 ratios within the range specified as follows: given our selection criteria, the lower bound is set by the requirement that a $\geq 5\sigma$ excess by the end of the HL-LHC is to be expected and the upper bound is set by the current upper limit on the cross section from ATLAS [8]. The corresponding range for 4.5 TeV resonances is $S_0/B_0 \in [0.6, 1.0]$, while that for 6 TeV is $S_0/B_0 \in [2.5, 5.5]$. We extend both ranges a little bit to better understand the trend of varying S_0/B_0, i.e., to [0.4, 1.2] and [1.5, 6.0], respectively. For 4.5 TeV resonances, we shuffle the samples repetitively until 15 000 histograms per class are generated. As for 6 TeV resonances, we generate 50 000 histograms per class to make up for the low event statistics in individual histograms. The same setting is also applied to one-jet scenarios.

As seen in Figs. 1(c) and 1(d), η^e is mostly confined within $[-2, 2]$. We therefore only bin the data within this range when making the histograms. The same procedure is also applied to the one-jet samples. A few p_T^e vs η^e sample histograms for 4.5 TeV resonances of $\Gamma_{\rm NP} = 200$ GeV with $S_0/B_0 = 1.0$ are shown in Fig. 2. Note that it is quite challenging to distinguish them by eye at high accuracy, but this will be manageable for the NNs.

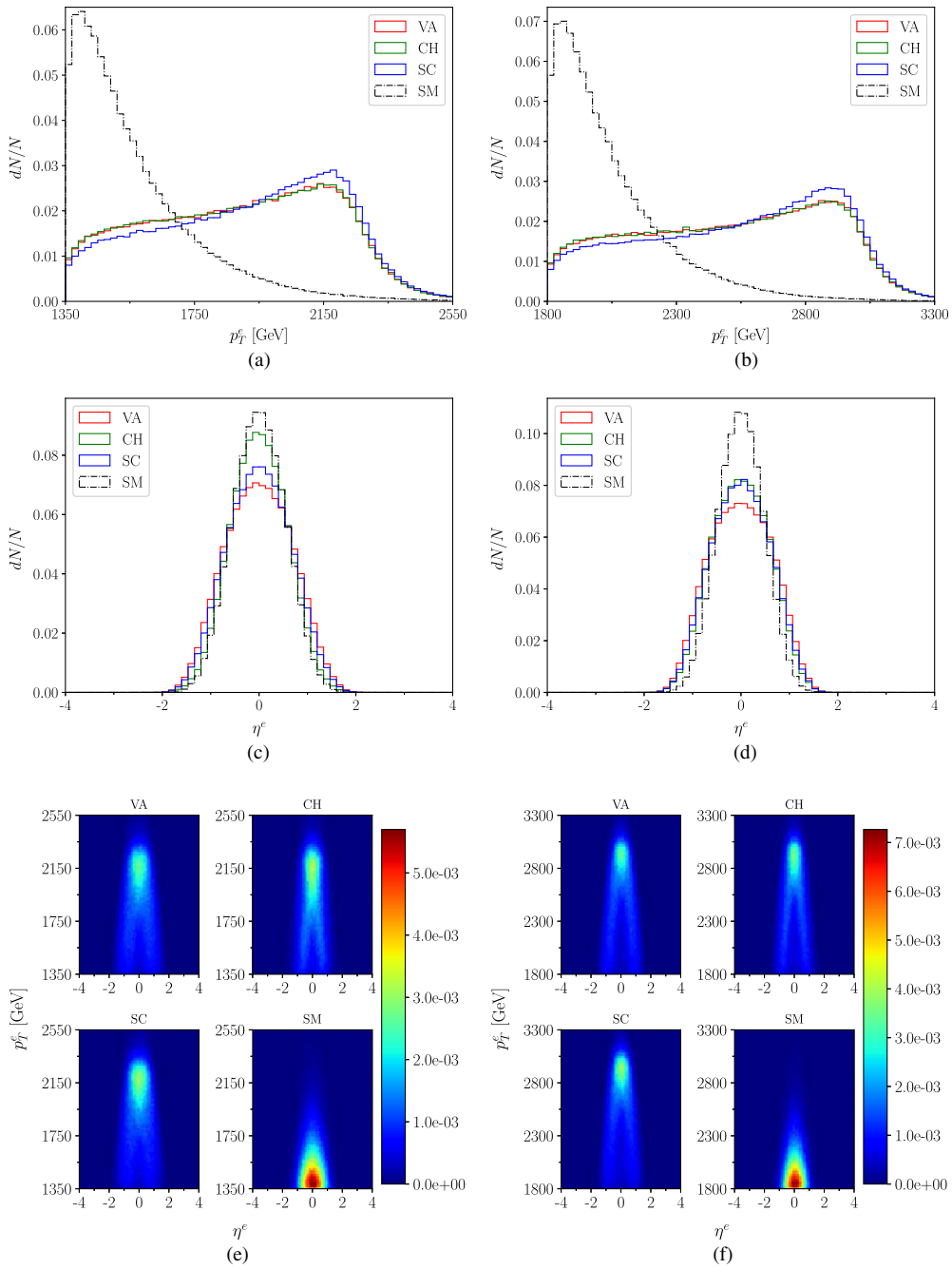

FIG. 1. (a) p_T^e, (c) η^e, and (e) p_T^e vs η^e distributions for the 4.5 TeV zero-jet samples with $\Gamma_{\rm NP} \approx 200$ GeV, and (b) p_T^e, (d) η^e, and (f) p_T^e vs η^e distributions for the 6 TeV zero-jet samples with $\Gamma_{\rm NP} \approx 300$ GeV. In plots (a), (b), (c), and (d) VA is depicted in red, CH in green, SC in blue, and SM in black. In plots (e) and (f) the color scale range goes from 0 to the maximum entry among all four classes, with the warmer/colder regions denoting more/fewer entries. The same color scheme is applied to all of the following figures. All of the distributions are normalized to unity.

B. One-jet samples

The one-jet samples include those events that have a leading jet with $p_T^j > 30$ GeV and are a subset of the zero-jet samples. Such events take up roughly 83% of all NP samples and 69% of the SM samples. For these samples, we ignore any subleading jet information. Therefore, the SM one-jet event numbers within this phase space are given by

$$B_1 = \sigma_{B_1} \times \mathcal{L} \approx \begin{cases} 58 & \text{for 4.5 TeV,} \\ 4 & \text{for 6 TeV.} \end{cases} \tag{8}$$

Hence, the corresponding S_1/B_1 are scaled up from S_0/B_0 by a factor of $0.83/0.69 \simeq 1.2$. For the convenience of an easy comparison with the zero-jet analysis, we will still label the signal-to-background ratio of one-jet samples by

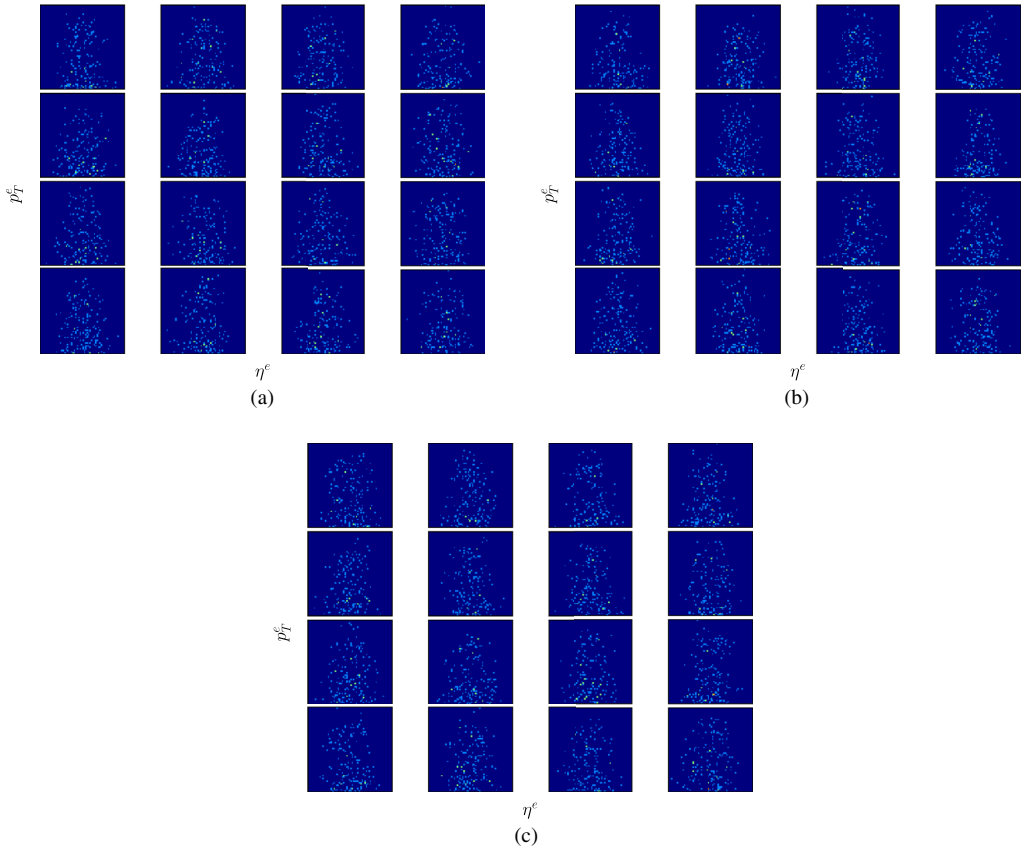

FIG. 2. Examples of zero-jet input histograms for (a) VA, (b) CH, and (c) SC samples for 4.5 TeV resonances of $\Gamma_{\rm NP} = 200$ GeV with $S_0/B_0 = 1.0$.

the zero-jet S_0/B_0 ratio, even though the true mixing ratio is S_1/B_1.

The kinematic observables of a one-jet process are as follows.

(1) p_T^e and p_T^j: transverse momenta of e and leading jet j, respectively.
(2) η^e and η^j: pseudorapidities of e and j, respectively.
(3) $\Delta\phi_{ej}$: azimuthal separation between e and j.

To form the required histograms and at the same time involve as much information as possible, we further consider three derived observables:

(1) $\not{E}_T$: missing transverse energy.
(2) $\Delta\phi_{e\not{E}_T}$ and $\Delta\phi_{j\not{E}_T}$: azimuthal separations between e and $\not{E}_T$ and between j and $\not{E}_T$, respectively.

We show the distributions of these kinematic observables for 4.5 TeV resonances in Fig. 3.

To utilize the additional information contained in these kinetic observables, we will make "RGB" histograms by choosing three pairs of variables. We propose the following four schemes.

(1) Scheme 1—*Physical relationship*: Intuitively, the kinematic information measured from a single object should manifest high correlation. Therefore, we first pair up p_T^e and η^e as well as p_T^j and η^j. Then, guessing that observables of the same mass dimension could be correlated, we choose two out of the three azimuthal separation variables, $\Delta\phi_{e\not{E}_T}$ and $\Delta\phi_{j\not{E}_T}$, to form the third pair.
(2) Scheme 2—*Principal component analysis*: Following Ref. [27], we also select another three pairs of variables by performing a principal component analysis (PCA). The results are shown in Table II. We start from the principal component (PC) with the highest variance. In each PC, we select the two variables with the highest (absolute) correlations to form a pair. Thus, from PC-1, we pair up p_T^e and $\not{E}_T$; and from PC-2, we pair up $\Delta\phi_{ej}$ and $\Delta\phi_{e\not{E}_T}$. Since $\Delta\phi_{ej}$ is already paired, we skip PC-3 and use PC-4 to pair up η^e and η^j.
(3) Scheme 3—*Common axis*: In this scheme, we investigate whether spatial correlation among the RGB channels provides better discriminating power. If we set one of the two axes of the three channels to always be p_T^e, the NN can then possibly make use of the correlations of the other variables to p_T^e, as it now becomes physically meaningful to compare the corresponding pixels with a common p_T^e coordinate.

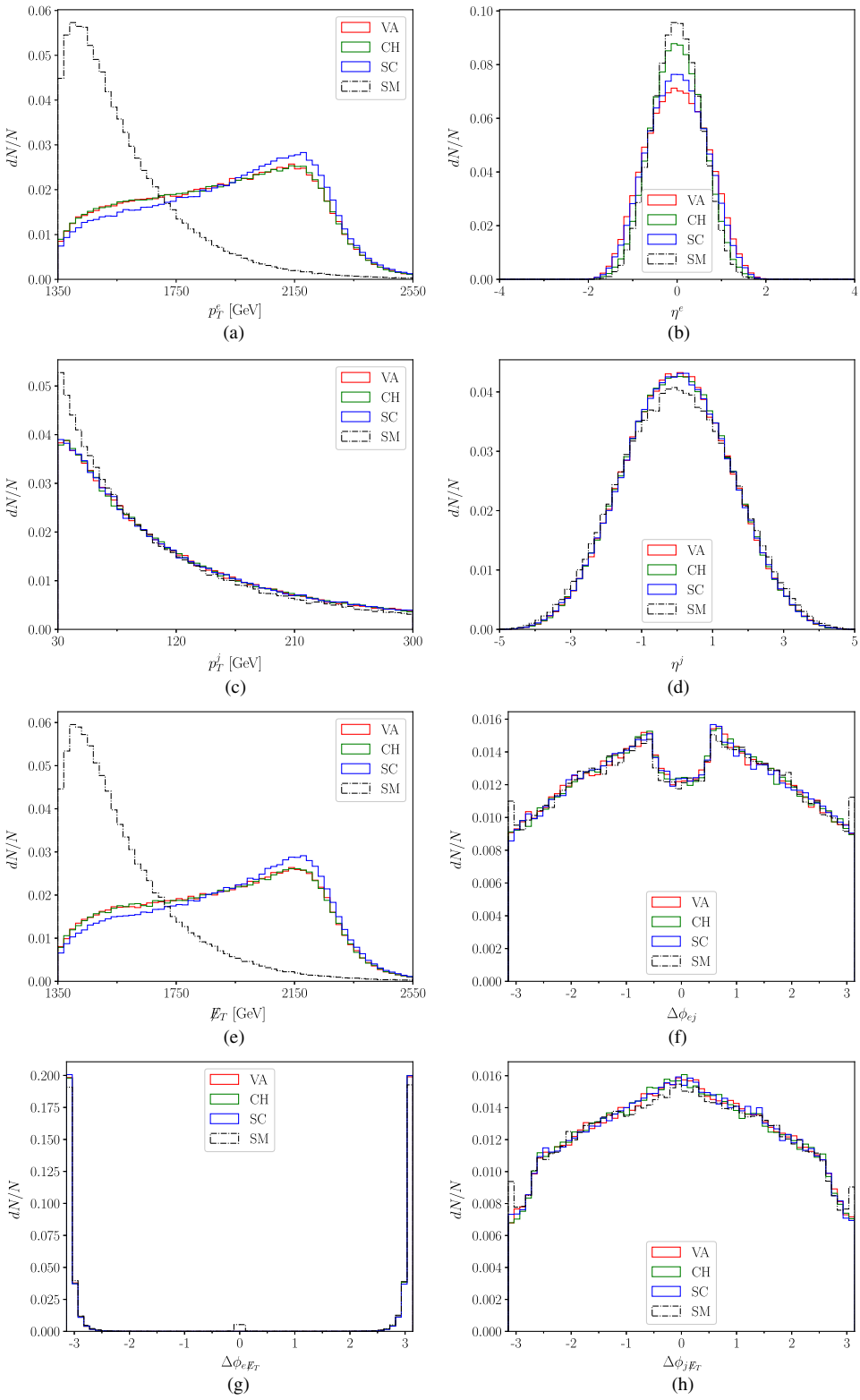

FIG. 3. Distributions of the kinematic observables (a) p_T^e, (b) η^e, (c) p_T^j, (d) η^j, (e) $\not{E}_T$, (f) $\Delta\phi_{ej}$, (g) $\Delta\phi_{e\not{E}_T}$, and (h) $\Delta\phi_{j\not{E}_T}$ for one-jet samples of mass 4.5 TeV and $\Gamma_{\mathrm{NP}} \approx 200$ GeV.

TABLE II. PCA result on one-jet samples. The correlations indicate the linear components of the PCs. The higher the variance is in its absolute value, the more significant the component contributes to the diversity of the samples.

		Correlations							
	Variance	p_T^e	η^e	p_T^j	η^j	$\not{E}_T$	$\Delta\phi_{ej}$	$\Delta\phi_{e\not{E}_T}$	$\Delta\phi_{j\not{E}_T}$
PC-1	1.78	0.707	0.001	0.040	0.003	0.706	0.009	−0.020	−0.015
PC-2	1.73	−0.019	0.001	−0.001	−0.001	−0.019	0.473	−0.760	−0.446
PC-3	1.27	0.003	−0.000	−0.001	0.000	0.004	0.695	0.011	0.719
PC-4	1.01	−0.001	−0.706	−0.014	0.708	0.000	0.000	−0.001	−0.001
PC-5	0.999	−0.011	−0.110	0.989	−0.089	−0.044	0.001	0.000	0.001
PC-6	0.991	−0.003	0.699	0.140	0.701	−0.008	0.000	0.000	0.000
PC-7	0.221	−0.707	0.000	0.024	0.000	0.706	0.000	0.000	0.000
PC-8	0.000	0.000	0.000	0.000	0.000	0.000	−0.542	−0.650	0.533

In light of this, we choose the following three pairs for this scheme: p_T^e and η^e, p_T^e and $\not{E}_T$, and p_T^e and $\Delta\phi_{ej}$.

(4) Scheme 4—*Best individuals*: After obtaining the training results of all of these individual pairs (also including the $\Delta\phi_{e\not{E}_T}$ vs $\Delta\phi_{j\not{E}_T}$ pair omitted in Scheme 2), to be shown in Sec. V, we further combine the three most powerful pairs to formulate the scheme using p_T^e vs η^e, η^e vs η^j, and p_T^e vs $\not{E}_T$.

It turns out that all four schemes give similar results. Even Scheme 4, which one naively expects to have the best efficiency, does not show noticeable superiority over the others. Since fixing one axis for all three color channels makes it easier to apply the Bayesian hypothesis test, to be discussed in Sec. VI, we will only focus on Scheme 3 in this paper. The corresponding 2D histograms for 4.5 TeV resonances are shown in Fig. 4.

IV. MODEL STRUCTURE AND TRAINING SPECIFICATIONS

In this section we describe in detail the structure of our FNNi, FNNh, and CNN models, which are constructed with the KERAS [45] library along with TensorFlow [46] for backend implementation. We will also describe our training specifications, including the training parameters and strategies.

A. FNNi structure

Our FNNi is designed to read the one-dimensional arrays of individual event observables as input, and to classify each histogram into one of the three signal classes. For the zero-jet samples we input two variables, p_T^e and η^e, while for one-jet samples we input p_T^e, p_T^j, $\eta^e\eta^j$, $\not{E}_T$, $\Delta\phi_{ej}$, $\Delta\phi_{e\not{E}_T}$ and $\Delta\phi_{j\not{E}_T}$. The FNNi structure is specified in Table III.

B. FNNh structure

Our FNNh is designed to read the flattened 60×60 2D histograms of kinematic variable pairs as input, and to classify each histogram into one of the three signal classes. For the zero-jet samples we only input one channel, p_T^e vs η^e, while for one-jet samples we input three channels based on the four different schemes described above, though only the results of Scheme 3 are presented in this paper. The FNNh structure is specified in Table IV.

C. CNN structure

Our CNN is designed to read 60×60 2D histograms of kinematic variable pairs as input with the RGB schemes mentioned previously, and to classify each histogram into one of the three signal classes. The CNN structure is specified in Table V.

D. Training specifications

In all trainings, we generate 15 000 histograms per class for 4.5 TeV and 50 000 for 6 TeV resonances. As for FNNi, we use 300 000 SM samples for the zero-jet study and 200 000 for the one-jet study, while the numbers of the NP samples are determined by S/B. We then split the dataset into three subsets—training, validation, and testing sets—in the proportion of 0.64:0.16:0.20. We set the batch size to 128 and the maximum training epoch to 1000. To avoid overtraining, we call for an early stopping if the validation loss has not improved by more than 2×10^{-4} for over 100 epochs.

To evaluate the performance of our NNs, we determine the receiver operating characteristic (ROC) curve in terms of the *one-against-all* strategy: we only consider the binary comparisons between class i and a combination of the other two classes, where i is the target class to be tested. Then, we calculate the areas under the ROC curves (AUCs) as a measure of the NN performance.

V. TRAINING RESULTS

In this section, we present the trained NN results of the zero- and one-jet processes for various zero-jet $S/B \equiv S_0/B_0$ ratios. We mainly focus on FNNh since it gives the

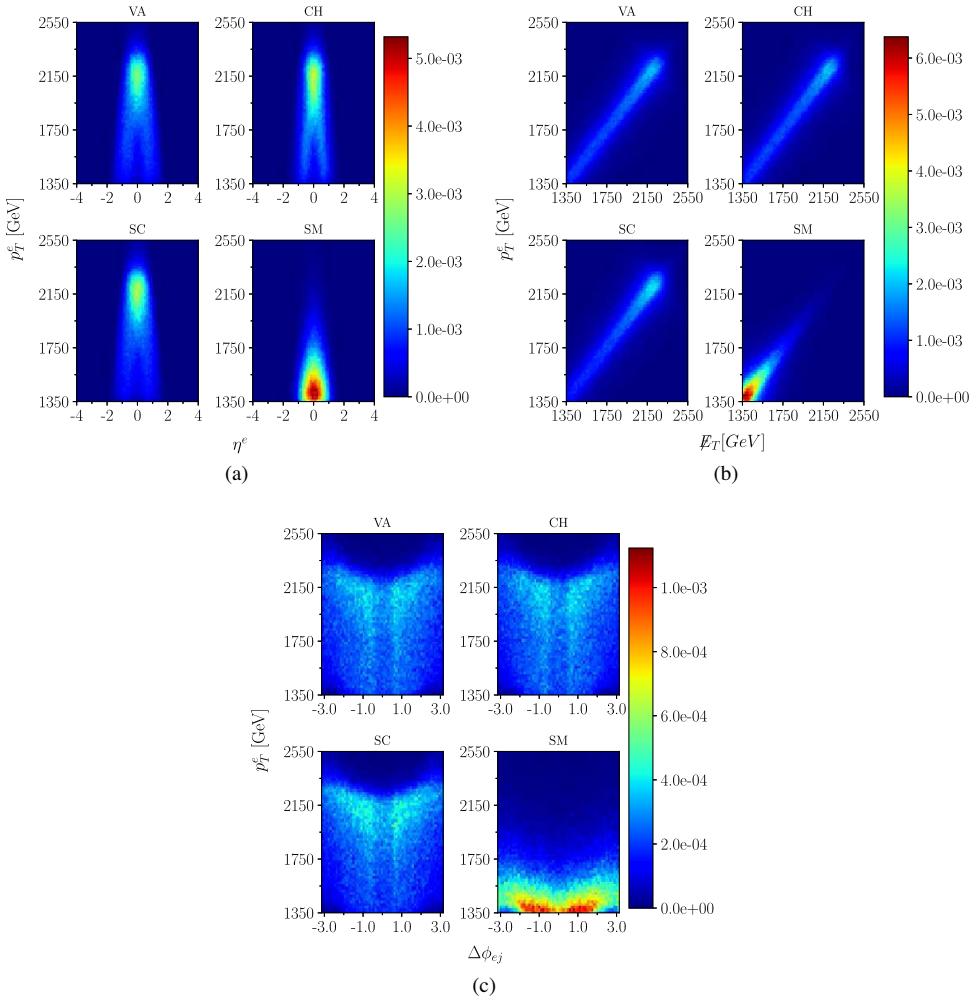

FIG. 4. One-jet 2D histograms formed from variable pairs determined according to Scheme 3 for samples of mass 4.5 TeV and $\Gamma_{\rm NP} \approx 200$ GeV. (a) p_T^e vs η^e (b) p_T^e vs $\not{E}_T$ (c) p_T^e vs $\Delta\phi_{ej}$.

best performance. We refer some more technical details of FNNh training to the Appendix. For the one-jet samples, we further investigate the importance of individual kinematic observable pairs.

TABLE III. Zero-jet and one-jet FNNi structure specifications.

	Zero jet	One jet
Input	p_T^e, η^e, ϕ^e	$p_T^e, \eta^e, p_T^j, \eta^j$ $\not{E}_T, \Delta\phi_{ej}, \Delta\phi_{e\not{E}_T}, \Delta\phi_{j\not{E}_T}$
Layers	batch normalization layer dense layer: 256[a] dense layer: 256	
Layer settings	hidden layer activation = `relu` output layer activation = `softmax`	
Compilation	loss = `categorical_crossentropy` optimizer = `adam` [47] metric = `accuracy`	

[a]This means that there are 256 nodes in the dense layer.

A. Zero-jet results

Since we only make use of electron information of the zero-jet samples, ignoring the jet information, the analysis in this section will determine how well the visible electron

TABLE IV. Zero-jet and one-jet FNNh structure specifications.

	Zero jet	One jet
Input	Flattened 60×60 images p_T^e vs η^e	p_T^e vs η^e, p_T^e vs $\not{E}_T$, p_T^e vs $\Delta\phi_{ej}$
Layers	batch normalization layer dense layer: 1024 dense layer: 256	
Layer settings	hidden layer activation = `relu` output layer activation = `softmax`	
Compilation	loss = `categorical_crossentropy` optimizer = `adam` metric = `accuracy`	

TABLE V. Zero-jet and one-jet CNN structure specifications.

	Zero jet	One jet
Input	60×60 images	
	p_T^e vs η^e	RGB colors: p_T^e vs η^e, p_T^e vs $\not{E}_T$, p_T^e vs $\Delta\phi_{ej}$
Layers	batch normalization layer convolutional 2D layer: 3-32[a] max pooling 2D layer: 2-2[b] convolutional 2D layer: 3-32 max pooling 2D layer: 2-2 flatten layer dense layer: 128 dense layer: 64	
Layer settings	hidden layer activation = `relu` output layer activation = `softmax`	
Compilation	loss = `categorical_crossentropy` optimizer = `adam` metric = `accuracy`	

[a]This means that the filter kernel dimension is 3×3, and that there are 32 nodes in the convolutional layer.

[b]This means that the max pooling kernel dimension is 2×2, and that each stride is 2 pixels.

information can distinguish the signal hypotheses. We first present the CNN, FNNh, and FNNi training outcomes of 4.5 TeV resonances with a width of 200 GeV and of 6 TeV resonances with a width of 300 GeV in Fig. 5. In the shaded regions in the figures, we denote two regions of S/B: one where HL-LHC will not achieve a 5σ excess, and one that violates the current constraint from ATLAS [8]. For 4.5 TeV resonances, all of the NNs can already start to distinguish the signal scenarios when $S/B \gtrsim 0.4$ and steadily improve with higher signal purities. At the 5σ discovery level, which corresponds to $S/B = 0.6$ in this case, both CNN and FNNh can distinguish with AUCs over 0.7 for all three classes, while FNNi just barely reaches this value for the SC class. Also, FNNh is always the best in terms of the identification of VA and SC classes. On the other hand, for 6 TeV resonances the differences among the three neural networks are milder, with FNNh still performing the best. For the 6 TeV plots, a 5σ discovery level requires $S/B = 2.5$, where the FNNh can reach AUCs ≥ 0.65 for all three classes, while at the current 95% C.L. limit $S/B = 5.5$, it can reach AUCs of around 0.75. Note that the CH class is always the easiest to identify, while VA and SC are more difficult. Moreover, even though the valid S/B values for 6 TeV resonances are much higher than those for 4.5 TeV resonances, the corresponding AUCs are significantly lower, suggesting that event statistics can be more critical than signal purity for this method. Note, however (as pointed out in Ref. [48]) that from a statistical point of view there should be no general superiority of FNNh over FNNi. One major reason for and benefit of using the FNNh approach is to enable the simplification of the model structure and training procedures. Hence, even though we identify FNNh as the best approach in our study, this fact is based upon the specific simple designs of our NN models. This argument also holds in the one-jet study.

Since FNNh gives the best results of the three or comparable results to the other two in all scenarios, we further present the results for 4.5 TeV resonances with $\Gamma_{\rm NP} \approx 500$ and 50 GeV using FNNh in Fig. 6. For all three different $\Gamma_{\rm NP}$ samples, the AUCs are roughly consistent with one another, suggesting that the boson width information does not affect the NN performance very much. This is believed to be mainly due to the fact that only the p_T distribution is changed by the width, only making it harder to distinguish between W' and the H' hypotheses. Thus, we will focus exclusively on the samples of $\Gamma_{\rm NP} \approx 200$ GeV for 4.5 TeV resonances in what follows.

To give a more interpretable metric, we now present the "accuracies" (ACCs) of our FNNh. The ACC here (and in the one-jet case below) is to be understood as the classwise true positive rate. For this, we associate each testing histogram to the class for which it gets the highest score, and then calculate the true positive rate for each class. We also calculate the average ACC curves, defined as the global true positive rate. Notice that although the average ACC curves [as shown in Figs. 7(a) and 7(c)] are stably improving, the classwise ACCs are rather unstable. This is mainly due to model biases. When evaluating the ACCs we only pick the best class score of each event, and thus the relation between different class scores is in some sense ignored. Unlike the AUCs which are evaluated using a sliding threshold, the ACCs are therefore more sensitive to model biases. Thus, to improve the stability we further apply a tenfold cross validation (CV) to better address this issue, with the results shown in Figs. 7(b) and 7(d). As expected, CV helps to stabilize the classwise accuracies and does not significantly alter the average. For the sake of comparison, we also show in Fig. 8 the AUCs after applying the tenfold CV. Notice that the resulting AUCs are only at most 2% and the ACCs at most 3% better than those without the CV, meaning that it does not matter much in the zero-jet case. However, the CV does a nice job of stabilizing both the 4.5 and 6 TeV testing performance.

Focusing once more on the average ACCs with the CV applied, we find that the ACCs for 4.5 TeV resonances are all above 0.6, and can reach 0.75 at $S/B = 1.2$; on the other hand, the ACCs for 6 TeV resonances are around 0.5 for $S/B \lesssim 2.5$, and can reach almost 0.6 at $S/B = 6.0$. All of these numbers improve significantly compared to a random guess with ACC = 0.33. Even though the ACC metric is more interpretable, we will continue to focus on the AUC as a more conventional metric to compare the performance of our classifiers.

Finally for FNNh, we analyze the confidence level at which it can rule out alternative hypotheses. For this, we split the ternary scores and analyze the following three

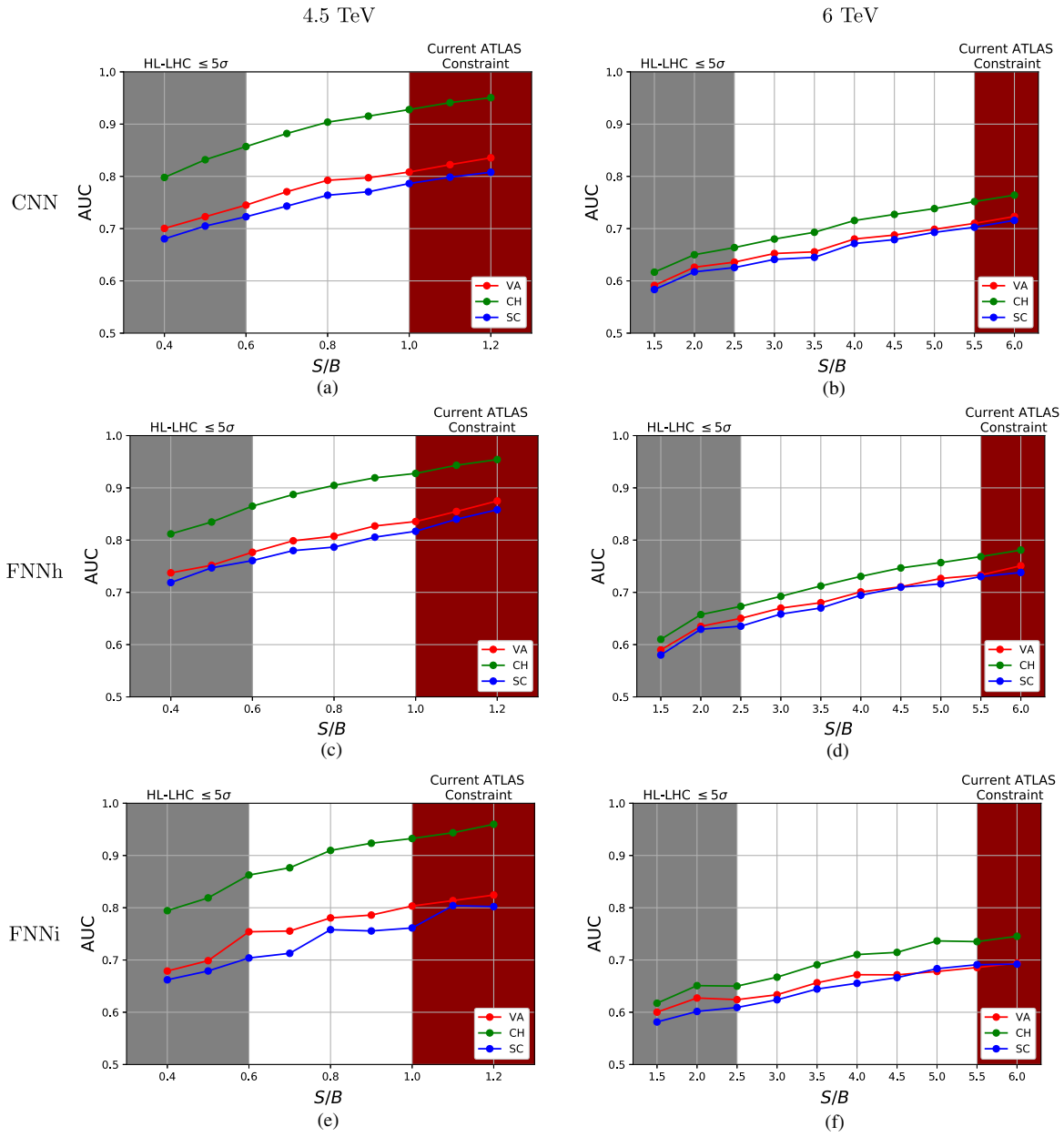

FIG. 5. AUC as a function of the S/B ratio for zero-jet samples. The left column is for a 4.5 TeV resonance with $\Gamma_{\rm NP} \approx 200$ GeV, while the right column is for a 6 TeV resonance with $\Gamma_{\rm NP} \approx 300$ GeV. The first row uses CNN, the second row FNNh, and the third row FNNi. The AUCs for the NNs to identify VA against non-VA are shown in red, CH against non-CH in green, and SC against non-SC in blue. The shaded regions denote S/B values where HL-LHC will not achieve a 5σ excess (gray) and the ATLAS constraint [8] is violated (red). The same color scheme applies to all of the subsequent figures.

cases separately: VA vs non-VA, CH vs non-CH, and SC vs non-SC. Note that these mirror the one-against-all strategy, allowing comparisons with the earlier AUC/ACC results. For the VA vs non-VA case at a fixed S/B, we assume that the VA hypothesis is true and use the VA score as the test statistic to constrain the non-VA hypothesis. We take the median value for the VA hypothesis and use it to determine the median expected p-value for the non-VA hypothesis, $p_{\rm med}$, which then gives a median expected exclusion for the alternative hypothesis at a confidence level of C.L. $= 1 - p_{\rm med}$. The modification for the other two cases requires swapping the assumed true and alternative hypotheses. We plot these C.L.'s against the S/B values for both 4.5 and 6 TeV resonances in Fig. 9. These C.L.'s are correlated but not directly related to our AUC and ACC metrics, since the latter are derived with varying thresholds. For example, one can see that the C.L.'s are higher (lower) than the AUCs for the 4.5 (6) TeV mass and

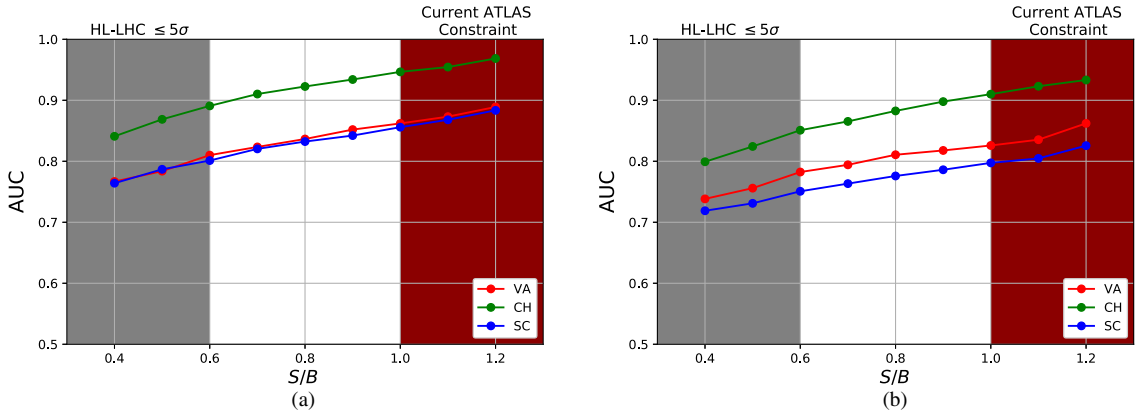

FIG. 6. FNNh training outcomes for zero-jet samples of a 4.5 TeV resonance with $\Gamma_{\rm NP} \approx$ (a) 500 and (b) 50 GeV, using the same color scheme as in Fig. 5.

that a C.L. value may correspond to very different corresponding AUC values. For both 4.5 and 6 TeV resonances, all of the alternative classes can be excluded at a C.L. $> 80\%$ in the S/B region of our interest, with the CH class always surpassing the other two, as expected from the previous AUC/ACC results. In particular, only the non-CH class can be excluded at $> 95\%$ C.L. in the allowed S/B range.

B. One-jet results

In this section, we include the information of the leading jet in addition to the visible lepton and show how such additional information helps to compensate for the lower event statistics. We will only present the result using Scheme 3 for 4.5 TeV resonances with $\Gamma_{\rm NP} \approx 200$ GeV and 6 TeV resonances with $\Gamma_{\rm NP} \approx 300$ GeV for the reasons

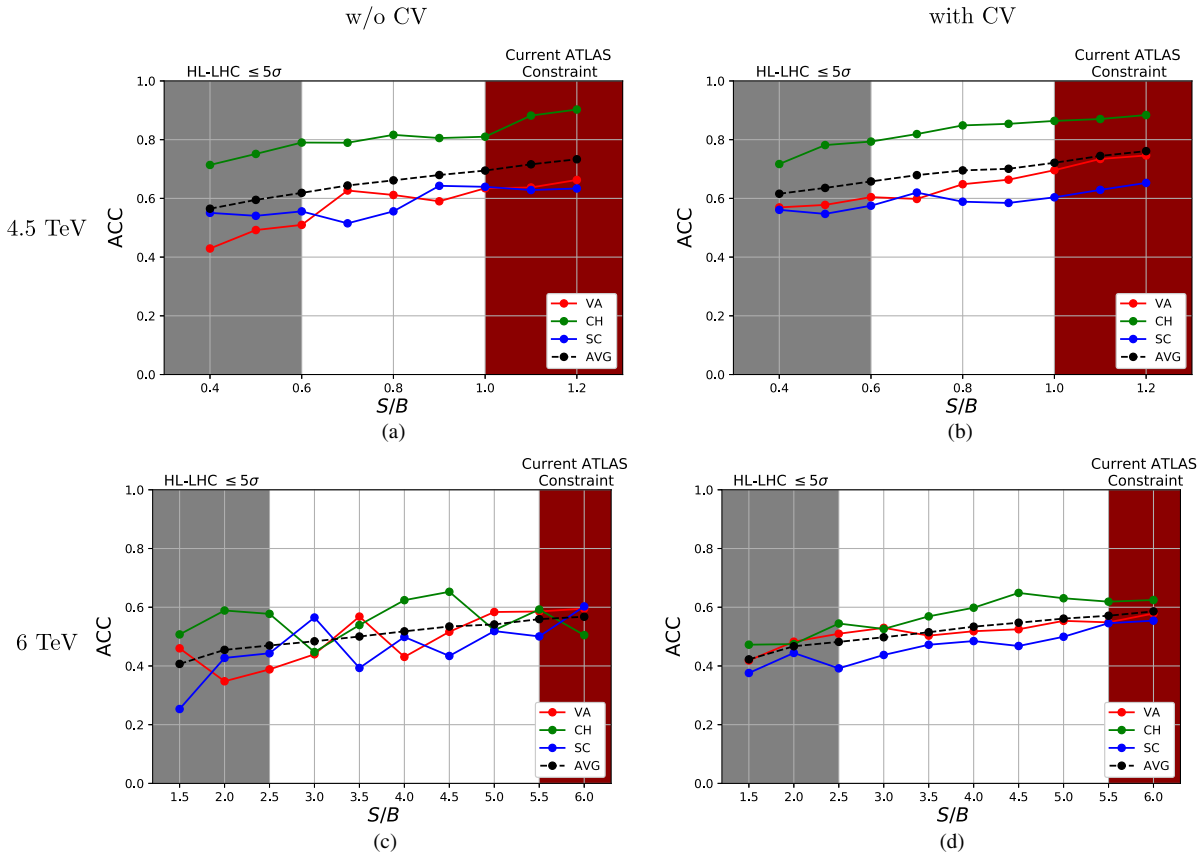

FIG. 7. Zero-jet ACCs for samples of (a) 4.5 TeV and (c) 6 TeV resonances using FNNh without CV, and (b) 4.5 TeV and (d) 6 TeV resonances with the tenfold CV applied. Color scheme is the same as in Fig. 5.

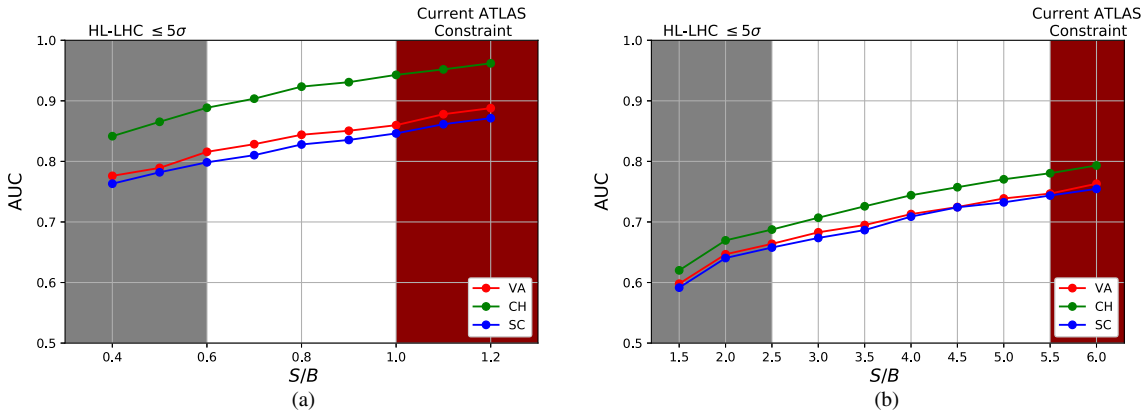

FIG. 8. Zero-jet AUCs for samples of (a) 4.5 TeV and (b) 6 TeV resonances using FNNh with the tenfold CV applied. Color scheme is the same as in Fig. 5.

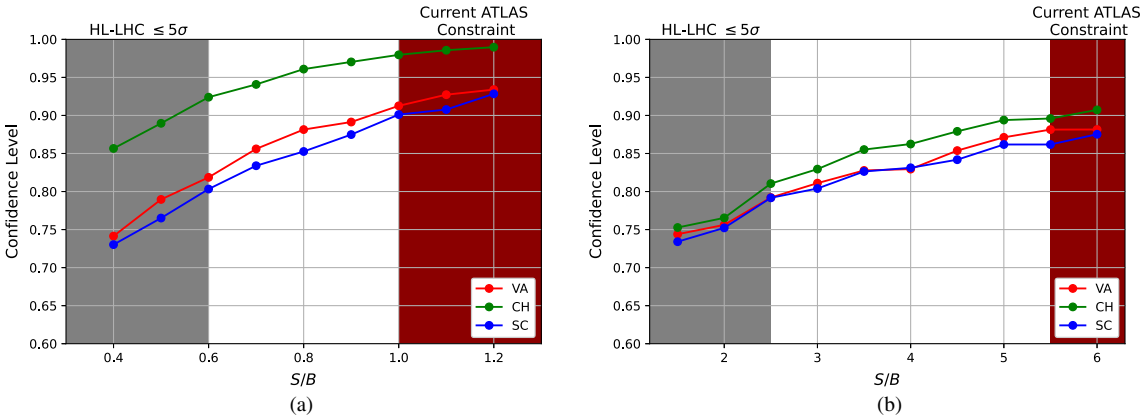

FIG. 9. Median zero-jet confidence levels at which the non-VA (red), non-CH (green), and non-SC (blue) hypotheses are excluded by the trained FNNh for samples of (a) 4.5 TeV and (b) 6 TeV resonances when assuming the VA, CH, and SC hypotheses are true, respectively.

stated before. We show the CNN, FNNh, and FNNi training outcomes in Fig. 10. First of all, we see again that for 4.5 TeV resonances FNNh outperforms the other two. There is an intriguing trend in the 6 TeV results: as CNN is consistently better than FNNi, FNNh is only slightly better than both of them at $S/B = 1.5$. As soon as S/B reaches 2.0, FNNh makes a sudden jump and significantly outperforms the other two thereafter.

Comparing Fig. 10(c) with Fig. 5(c), we see that the one-jet FNNh performance for 4.5 TeV resonances is much better than that of the zero-jet performance in terms of the VA and SC classes, both of which can reach AUCs of 0.8 even at $S/B = 0.4$, while CH seems to be only slightly better. A comparison between Fig. 10(d) and Fig. 5(d) shows an even more interesting trend for 6 TeV resonances: all three classes can be better classified using the one-jet strategy except for $S/B = 1.5$, and can even reach AUCs of 0.8 for $S/B \gtrsim 5.0$. This shows that even with the drop in statistics by going to one-jet events, there is improved discriminating power over the zero-jet analysis. Thus, this proves that this technique is promising for higher-dimensional histograms, thus broadening the range of viable channels to be studied and even potentially granting better distinguishing power.

We present in Fig. 11 the FNNh ACCs without (left column) and with (right column) CV for one-jet processes, and in Fig. 12 the FNNh AUCs with the CV applied. Compared to the zero-jet results, the CV does an even better job of stabilizing the one-jet results. For 4.5 TeV resonances, the AUCs are on the average 4–5% better and the average ACCs are 6–7% better than those without the CV. On the other hand, for the 6 TeV resonances the AUCs are boosted by 1–2% and the average ACCs by 2–3%. Comparing the ACCs to those of the zero-jet study, we can see that they are much better except for the 6 TeV case when $S/B = 1.5$, indicating that the one-jet strategy is more powerful in distinguishing different interaction hypotheses.

To understand the importance of each individual variable pair in the one-jet FNNh training, we have also trained the

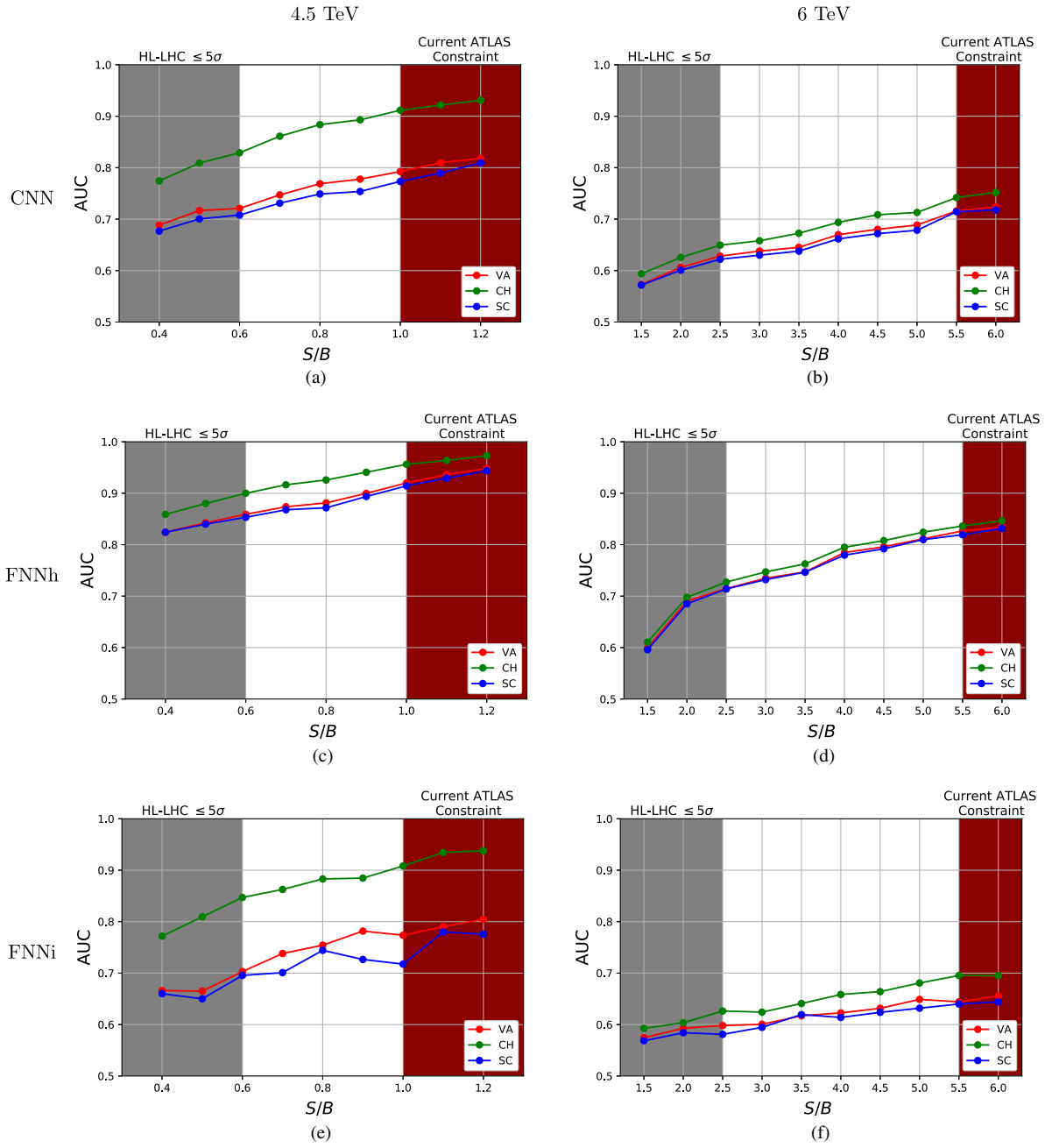

FIG. 10. Training outcomes of 4.5 TeV resonances using one-jet samples in Scheme 3 for (a) CNN, (c) FNNh, and (e) FNNi, and training outcomes of 6 TeV resonances using one-jet samples in Scheme 3 for (b) CNN, (d) FNNh, and (f) FNNi. Color scheme is the same as in Fig. 5.

FNNh on single pair histograms for 4.5 TeV resonances. The FNNh training outcomes for the most powerful individual histograms mentioned in all four one-jet schemes are shown in Fig. 13. We dropped the results of p_T^j vs η^j, $\Delta\phi_{ej}$ vs $\Delta\phi_{e\not{E}_T}$, and $\Delta\phi_{e\not{E}_T}$ vs $\Delta\phi_{j\not{E}_T}$ here as they barely have any distinguishing power. Clearly, p_T^e vs η^e plays the most important role in the class discrimination. This is physically understandable as we expect the angular and coupling information of the leptonic decay to be preserved mostly in the charged lepton, which is a direct decay product of the new charged bosons, rather than in j. Following p_T^e vs η^e are η^e vs η^j, p_T^e vs $\not{E}_T$, and p_T^e vs $\Delta\phi_{ej}$, with the first two being best at identifying the CH class and the latter two at identifying the SC class. Compared to Fig. 10, we see that combining different channels does lead to a better overall performance, thus demonstrating that the multidimensional FNNh can successfully utilize the additional information in these channels.

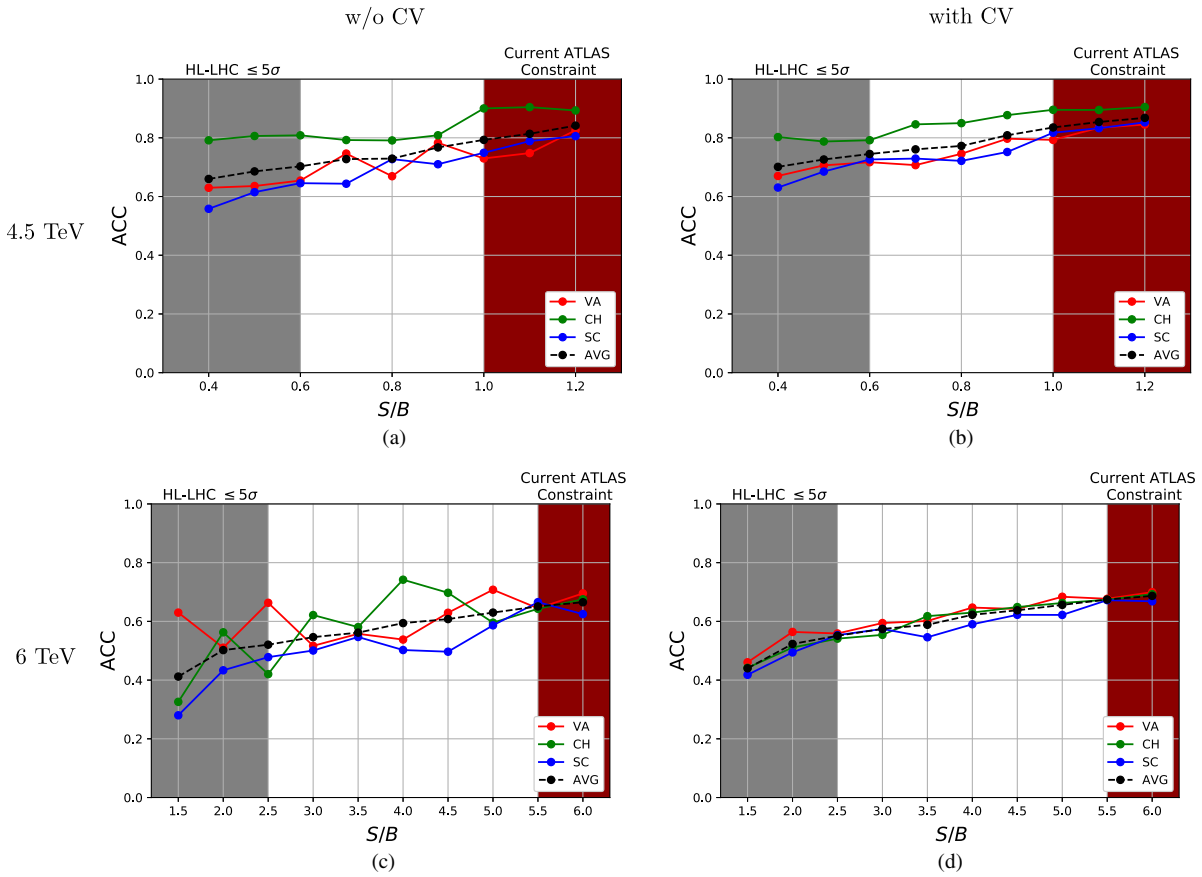

FIG. 11. One-jet (a) 4.5 and (c) 6 TeV resonances FNNh Scheme 3 ACCs without CV, and (b) 4.5 and (d) 6 TeV resonances with the tenfold CV applied. Color scheme is the same as in Fig. 5.

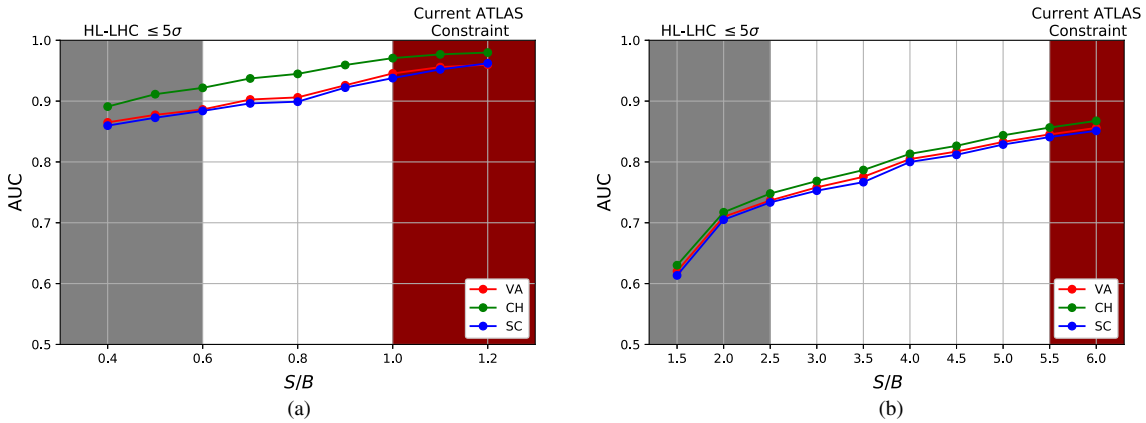

FIG. 12. One-jet AUCs for samples of (a) 4.5 and (b) 6 TeV resonances using FNNh in Scheme 3 with the tenfold CV applied. Color scheme is the same as in Fig. 5.

Finally, we also analyze the C.L. at which the FNNh can rule out alternative hypotheses. We plot these C.L.'s against the S/B values for both 4.5 and 6 TeV resonances in Fig. 14. For 6 TeV resonances, all of the alternative classes can again be excluded at a C.L. $\gtrsim 80\%$ in the S/B region of our interest, while for 4 TeV resonances they can reach C.L.'s of over 95%. Both of these are better than their zero-jet counterparts, providing us with yet another metric to highlight the improvement.

VI. COMPARISON WITH BAYESIAN HYPOTHESIS TEST

Finally, to give context for our NN approach, we compare the zero- and one-jet FNNh 4.5 TeV resonance

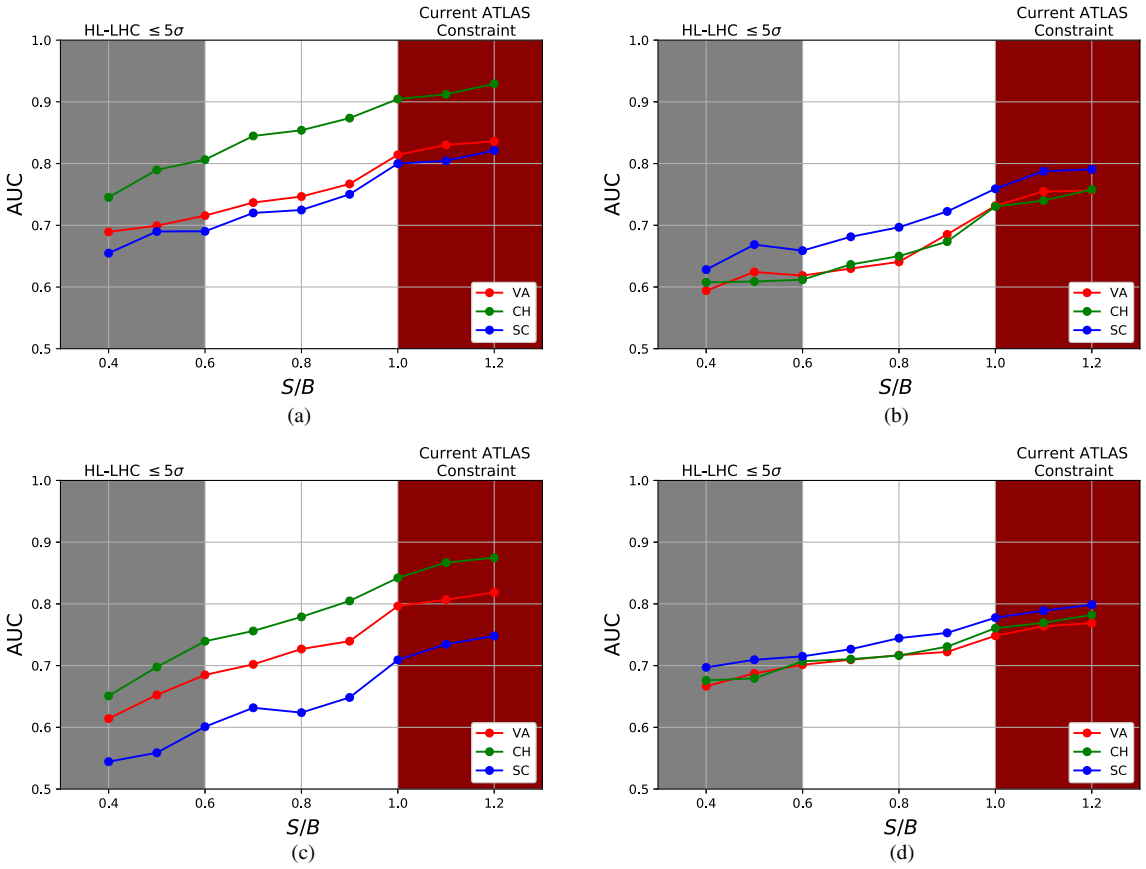

FIG. 13. One-jet FNNh training outcomes for 4.5 TeV resonances using individual channels: (a) p_T^e vs η^e, (b) p_T^e vs $\not{E}_T$, (c) η^e vs η^j, and (d) p_T^e vs $\Delta\phi_{ej}$. Color scheme is the same as in Fig. 5.

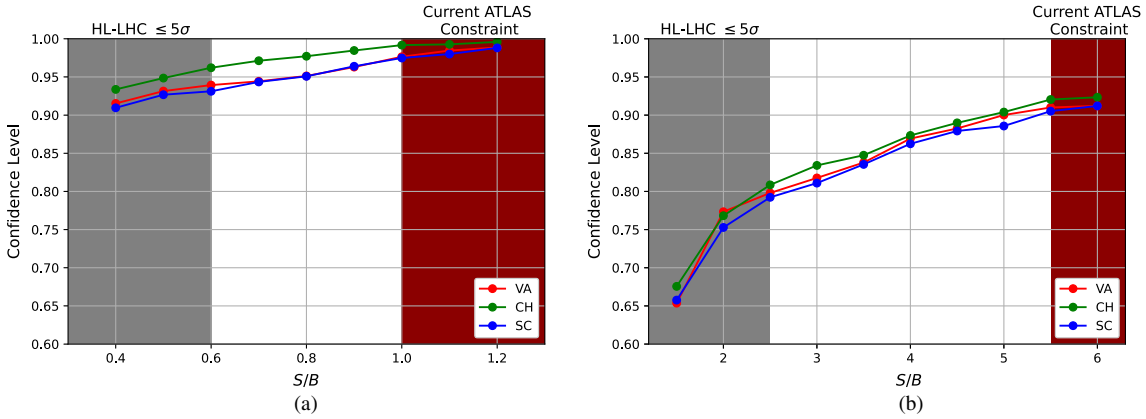

FIG. 14. Median one-jet confidence levels at which the non-VA (red), non-CH (green), and non-SC (blue) hypotheses are excluded by the trained FNNh for samples of (a) 4.5 TeV and (b) 6 TeV resonances when assuming the VA, CH, and SC hypotheses are true, respectively.

results with a standard hypothesis test: the Bayesian hypothesis (BH) test.[3] In the Bayesian approach, for a specific observed dataset D, the probability for it to suggest a specific hypothesis H^k is given by

$$P(H^k|D) = \frac{P(D|H^k) \times P(H^k)}{\sum_k P(D|H^k) \times P(H^k)}, \tag{9}$$

where $P(H^k)$ denotes the prior that the hypothesis H^k is correct, and $P(D|H^k)$ gives the conditional probability to

[3]We have also tried to compare with the χ^2 test. However, it suffers from serious issues in the presence of bins with small or zero expected events, which has to be resolved through coarser binning and decomposing the ternary test to multiple binary tests.

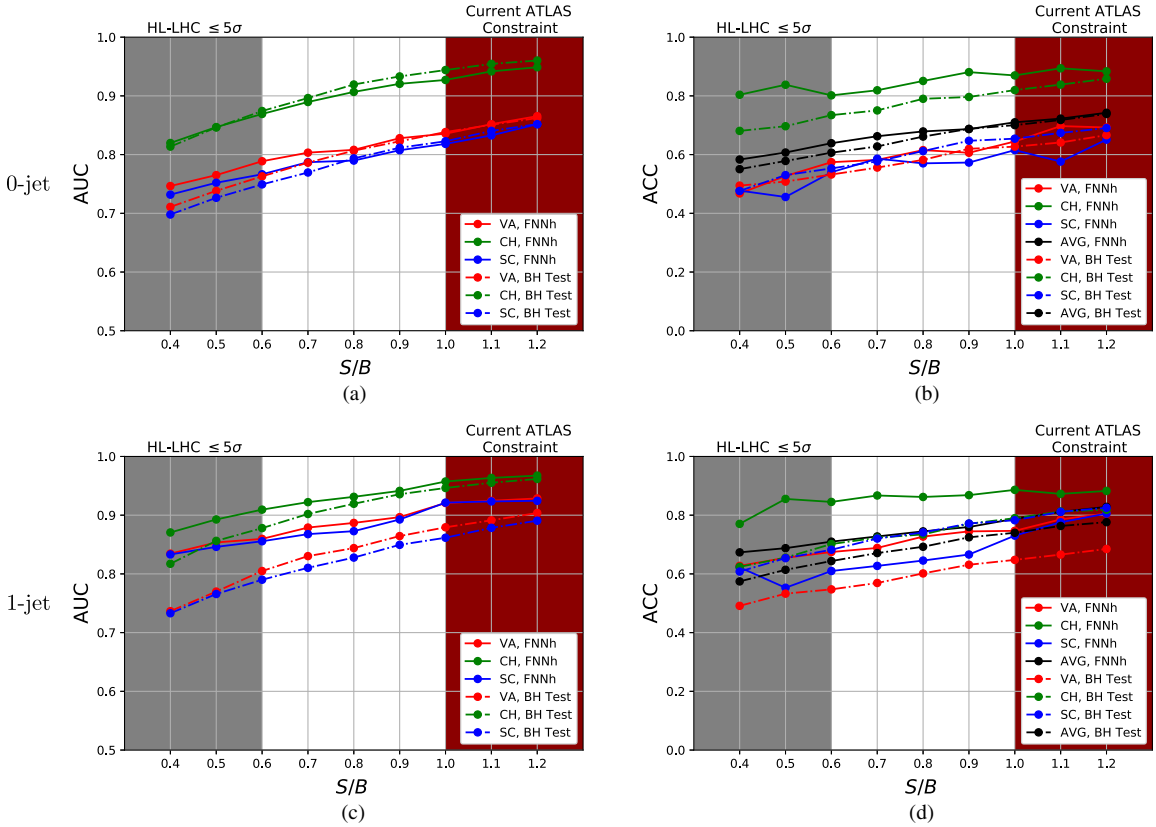

FIG. 15. (a) Zero-jet AUCs, (b) zero-jet ACCs, (c) one-jet Scheme 3 AUCs, and (d) one-jet Scheme 3 ACCs for 4.5 TeV resonances, obtained using FNNh with the tenfold CV and the BH test. The histogram dimension is set to 60×60 and the binning configuration mentioned in the main text is applied. Color scheme is the same as in Fig. 5.

obtain dataset D given the fact that H^k is correct. In our study, we assume that it is equally likely for all of the hypotheses ($k =$ VA, CH, and SC) to be correct, and hence $P(H^k) = 1/3$. We assume Poisson distributions for all individual bin counts, and the conditional probabilities $P(D|H^k)$ are then given by

$$P(D|H^k) = \prod_{m,n} f(h^D_{mn}, H^k_{mn}), \tag{10}$$

where $f(h^D_{mn}, H^k_{mn})$ denotes the Poisson probability for an observed number of counts h^D_{mn} at the pixel (m, n) in the 2D histogram, assuming an expectation value of H^k_{mn}. From the definition of Eq. (10), one can see that there would be a problem if any $H^k_{mn} = 0$ because an observed count in this pixel would have an extremely high weight in determining the hypothesis. An even worse case is when $H^k_{mn} = 0$ while $H^{k'}_{mn} \neq 0$ for $k' \neq k$, as any sample histogram with $h^D_{mn} \neq 0$ will definitely have zero probability to be identified as class k. This results from the fact that this approach does not take into account possible systematic or statistical errors, and hence it cannot be trusted around these low-statistics regions. To overcome this problem, we first symmetrize H^k with respect to the $\eta^e = 0$ axis and then exclude the problematic bins (those where $H^k_{mn} = 0$ while $H^{k'}_{mn} \neq 0$). To fairly compare the BH test with the FNNh, we apply the same binning configuration to the training and testing samples with the tenfold CV.

The zero-jet AUCs and ACCs for both FNNh with the tenfold CV (solid lines) and BH tests (dashed lines) are shown in Figs. 15(a) and 15(b), respectively. The VA and SC AUCs as well as the VA, CH, and average ACCs indicate that the FNNh approach is able to produce the same level of performance as the BH test for $S/B \gtrsim 0.8$, and even performs better when $S/B \lesssim 0.8$. One exception happens for the CH AUCs when $S/B \gtrsim 0.6$, although the difference is less than 1.5%. Another exception shows up in the SC ACCs, where the BH test is better than FNNh.

This kind of binning strategy used in the BH test could be more difficult to implement in other cases. For example, for a higher-mass resonance the p_T range to be studied would be wider, with more chances to get empty bins. Therefore, either more events need to be generated, more bins need to be excluded, or the bins should be made coarser; otherwise, the BH test cannot be applied properly. Another complication could occur if more kinematic variables are needed. As the dimension of the phase space to be studied increases, proper binning will become more

challenging. In fact, we encountered such an issue when we turned to one-jet samples. To compare with the one-jet FNNh in Scheme 3, we performed a BH test using

$$P(D|H^k) = \prod_{a=1}^{3} P(D_a|H_a^k), \tag{11}$$

where a denotes the three input channels.[4] The resulting AUCs and ACCs are shown in Figs. 15(c) and 15(d). We can see that for the one-jet case, FNNh is consistently better than the BH test in terms of all metrics except for the SC ACC. Thus, we can conclude that the FNNh generally has a better ability to adapt to this binning issue, as well as a stronger power to combine information from different channels, except when it comes to the identification of SC signals. However, as the previous results have shown, FNNh does not suffer from the binning issue and thus can in general improve.

As these two comparisons illustrate, the FNNh compares favorably in performance with the standard BH test in most cases. To summarize the pros and cons compared to the BH test, the FNNh has the advantages that it automatically takes care of the binning issue, does not require a large sample to approximate the probability density functions, and easily generalizes to higher dimensions, while it has the usual neural network disadvantages of proper training and validation.

VII. CONCLUSIONS

In this paper, we have investigated the ability of using deep neural networks to distinguish different resonances in the $pp \to W'/H \to \ell\nu_\ell$ process at the HL-LHC. We showed that the original event-by-event ambiguities in the coupling differentiation problem could be tackled by classifiers with an NN architecture that takes binned histograms as the input. The predicted p_T^e distributions allow a discrimination between H and W', and because of the boosted parton collision frame, W' bosons with different couplings further manifest different η distributions.

Extending previous signal-only analyses [27], we demonstrated that simple NNs could start to distinguish the signals even with a low signal-to-background ratio, S/B. Of the three NN approaches we studied, the best was a fully connected neural network whose inputs were flattened histograms of kinematic variables, though FNNi could potentially compete with it if more careful fine-tunings were applied, as pointed out in Ref. [48]. We found that the FNNh could achieve AUCs over 0.80 when $S/B \gtrsim 0.8$ for 4.5 TeV resonances, and over 0.60 when $S/B \gtrsim 3.0$ for 6 TeV resonances. As our one-jet schemes showed, the 2D approach of Ref. [27] could also be generalized to higher dimensions, where we took into account the extra information of the jet by using the "RGB" channels to represent different kinematic variable pairs. This additional jet information compensated for the drop in the event statistics, generally leading to better performance for both 4.5 and 6 TeV resonances. We also investigated the usefulness of cross validation, and discovered that it helped less for the zero-jet study, but boosted the performance and stability greatly for the one-jet study. Performance differences resulting from different boson widths and the four pairing schemes were also investigated, and it was concluded that there was no major difference among the training results.

Finally, we studied the importance of each individual variable pair in the one-jet FNNhs, and found that they had different discriminating power for the three signal classes, with some variable pairs being more suited to picking out certain classes. Out of all of the variable pairs, the FNNh still relied mostly on the information of the charged lepton, although our results showed that the RGB color scheme successfully combined multiple channels to produce a better overall performance. As a final comparison, we also showed that this technique was as good or better than the conventional Bayesian hypothesis testing procedure, without having to worry about binning issues or how to generalize to higher dimensions.

Even though this study is based upon the specific choice of 4.5 and 6 TeV masses for the new charged resonance, it can be readily extended to other mass ranges at future colliders, in which case sufficient event statistics apparently is a critical factor for the success of the NN technique. Moreover, more general studies can also be considered, such as modifying the hypotheses (e.g., spin and couplings), analyzing channels other than the zero- and one-jet processes presented here, or constructing NNs with inputs of more than three channels and higher-dimensional "super images."

ACKNOWLEDGMENTS

The work of S. C. was supported in part by the U.S. Department of Energy under Grant No. DE-SC0011640. The works of T.-K. C. and C.-W. C. were supported in part by the Ministry of Science and Technology (MOST) of Taiwan under Grant No. MOST-108-2112-M-002-005-MY3. We appreciate the support of the NVIDIA Corporation with the donation of the Titan Xp GPU used in this study. We also thank Kai-Feng Chen for the suggestions about hypothesis tests and Yu-Chen Janice Chen for support and discussion about NNs. S. C. thanks the hospitality of the Physics Department of National Taiwan University and the sabbatical support of the MOST of Taiwan when this project was initiated.

[4] In principle, a four-dimensional (4D) version of the BH test could be done with the full knowledge of the probability density function in the 4D phase space of $(p_T^e, \not{E}_T, \eta^e, \Delta\phi_{ej})$. This is computationally challenging, but it would be interesting to compare with either a 4D CNN or a six-color 2D CNN taking into account all of the variables.

APPENDIX: TECHNICAL STUDIES

To better understand the technical details of our method, we investigate the dependence of the zero-jet FNNh on variables such as the resolution and kinematic window. We also confirm the consistency between zero- and one-jet binary and ternary classifiers by introducing a projection of scores in the latter case, which has also been studied in Ref. [28]. Finally, we demonstrate how robust the performance of the FNNh is even when applied to testing samples from a distribution that it is not trained on with different S/B ratios and decay widths. All of the following studies are based on 4.5 TeV resonances.

1. Kinematic window and resolution

We expect the performances of the NNs to be better if we extend the phase space from $p_T^e > 1350$ GeV to a lower p_T^e minimum as it would include more information about the signal. However, there are two problems associated with an unchecked extension of this lower bound:

(1) First, when p_T^e gets closer to $m_W/2$, the number of NP signals will be overwhelmed by the number of SM signals around the W boson Jacobian peak. Therefore, including information from this region would contribute little to nothing. What is even worse is that the excess of SM signals may confuse the NNs and reduce their efficiency.

(2) Second, if one were to maintain the same p_T resolution for the histogram bins, the required NN complexity and computational resources for training would increase rapidly as $p_{T,\min}$ decreases. Yet, if one wants to maintain the same level of input bins for the NNs, the resolution in p_T^e would be compromised.

As a result, we expect a "sweet window" that balances these issues. We base our study on samples of $S/B = 0.4$ and 1.2 under the cuts given in Table I. To extend p_T^e to lower regions, we first define the following parameters:

$B, B'(k)$: numbers of SM events for $p_T^e \geq 1350$ GeV, $p_{T,\min}$, respectively.

$S_c, S_c'(k)$: numbers of class c events for $p_T^e \geq 1350$ GeV, $p_{T,\min}$, respectively.

We generate another set of samples based upon the same settings as before, but change the selection cut from $p_T^e > 1350$ GeV to $p_T^e > 750$ GeV as we are setting $p_{T,\min}$ to 750. Introducing the ratios $r_B \equiv B'/B$ and $r_{S_c} \equiv S_c'/S_c$, the mixing ratio between the NP and SM events should then be modified to

$$\frac{S_c'}{B'} = \frac{r_{S_c}}{r_B}\frac{S_c}{B}. \tag{A1}$$

In general, $r_{S_{\mathrm{VA}}}$, $r_{S_{\mathrm{CH}}}$, and $r_{S_{\mathrm{SC}}}$ are all different. This would lead to histograms with different numbers of events. Instead, we define $r_S \equiv \sum_c r_{S_c}/3$ and mix the new samples of all three classes according to

$$\frac{S'}{B'} = \frac{r_S}{r_B}\frac{S}{B}. \tag{A2}$$

This procedure is carried out for $p_{T,\min} = 750, 950, \ldots,$ 2150 GeV, respectively, and the corresponding histograms of dimension 60×60 are then made from the mixed samples. Note that in these studies we fix the bin size of η^e.

The AUCs of FNNh trained on zero-jet histograms of different $p_{T,\min}$ are plotted in Fig. 16. It is clear that there exists a "sweet window" for the cut at 1350 GeV for $S/B = 0.4$ and within [950, 1150] for $S/B = 1.2$. The performance deteriorates for $p_{T,\min}^e$ either below or above the window boundaries. The windows for the two different S/B ratios are different because with higher S/B, it is more likely to get "useful" signals as $p_{T,\min}$ decreases, and hence

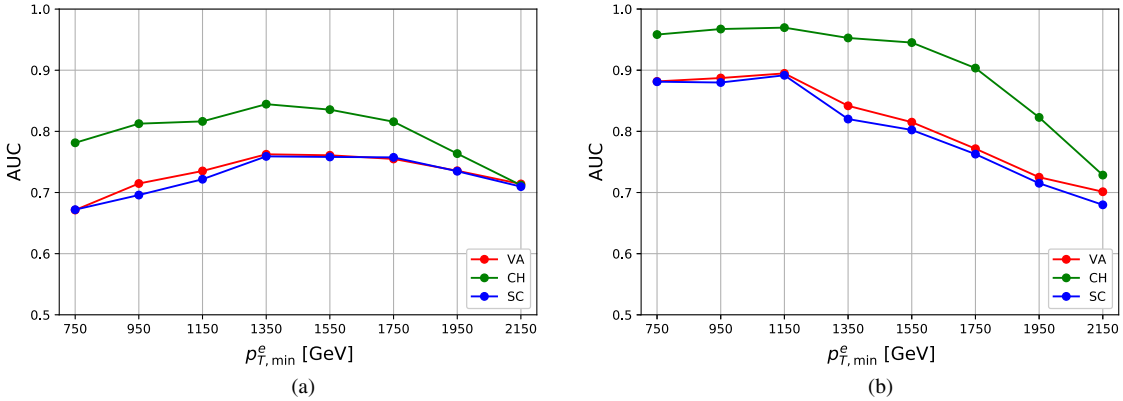

FIG. 16. Zero-jet AUCs of training on histograms of different $p_{T,\min}$, with their dimensions fixed to 60×60 and covering the entire p_T range. The histograms are made from samples of $S/B =$ (a) 0.4 and (b) 1.2. The η^e resolution remains the same as the value implemented in previous training.

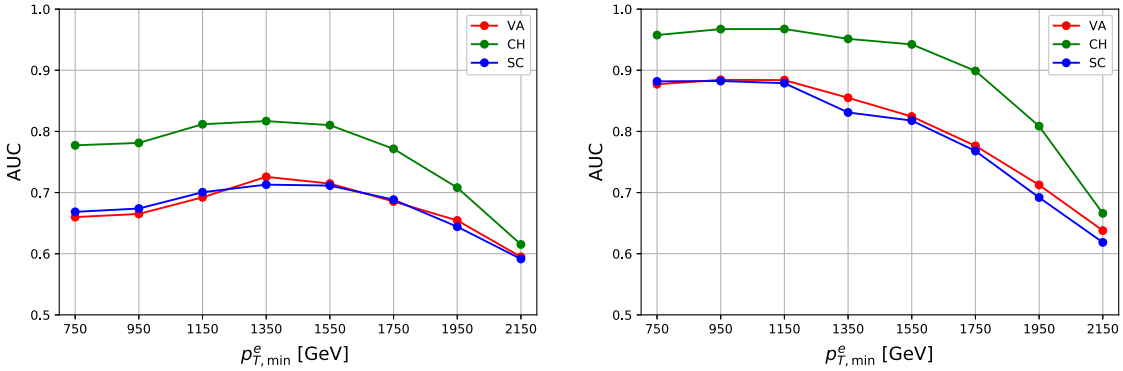

FIG. 17. Zero-jet AUCs of training on histograms of different $p_{T,\min}$, with their dimensions fixed to 60×60, the p_T bin size fixed to 30 GeV, and the uncovered bins left empty. The histograms are made from samples of $S/B = (a)\,0.4$ and (b) 1.2. The η^e bin number is fixed at 60.

the optimal p_T cut that balances the previous issues should naturally lie somewhat lower.

To pin down whether the effect of $p_{T,\min}$ is due to resolution, we also fix the p_T bin size to 30 GeV. The bins outside the p_T cut are then filled with zeros so as to retain a uniform structure for our NNs. The results are given in Fig. 17. The overall trend suggests that the reduced performance of a lower $p_{T,\min}$ is mainly due to incomplete training rather than the resolution, which appears to be more important for the $S/B = 0.4$ scenario due to lower event numbers.

We further study the effect of p^e_T resolution in the following way: we only use events with $p^e_T \in [1350, 2550]$ GeV and partition them into 1, 2, 3, 4, 5, 10, 20, 40, 60 bins, respectively. Samples of $S/B = 0.4$ and 1.2 are again used. To retain the same NN structure, we fill in null bins so that the histograms are still of dimension 60×60. The training outcomes are shown in Fig. 18. The AUCs apparently drop as the bin number decreases, but only when there are five or fewer bins, confirming that the p_T resolution does play a role in the NN performance but only when the binning is extremely coarse. As one increases the number of bins, the AUCs nearly saturate their maximum values way before $N_{\text{bin}} = 60$. Consequently, we can infer that as long as the η^e resolution remains sufficiently high, the p^e_T resolution does not need to be maximized to obtain the optimal NN performance.

2. Consistency between binary and ternary classifiers

Even though we are dealing with a three-class problem, one alternative other than training a ternary classifier to tag a specific sample set is to test it with multiple binary classifiers. If the NNs are all properly trained, we should expect a consistency in their performances. Therefore, we compare the two methods in the following way.

After each individual testing sample is tested by a trained ternary NN classifier, it will be assigned a three-component score array, (P_1, P_2, P_3), denoting its "probabilities" of belonging to one of the three classes. Suppose we are trying to compare a ternary NN's performance with that of a binary NN concerning the discrimination between class i

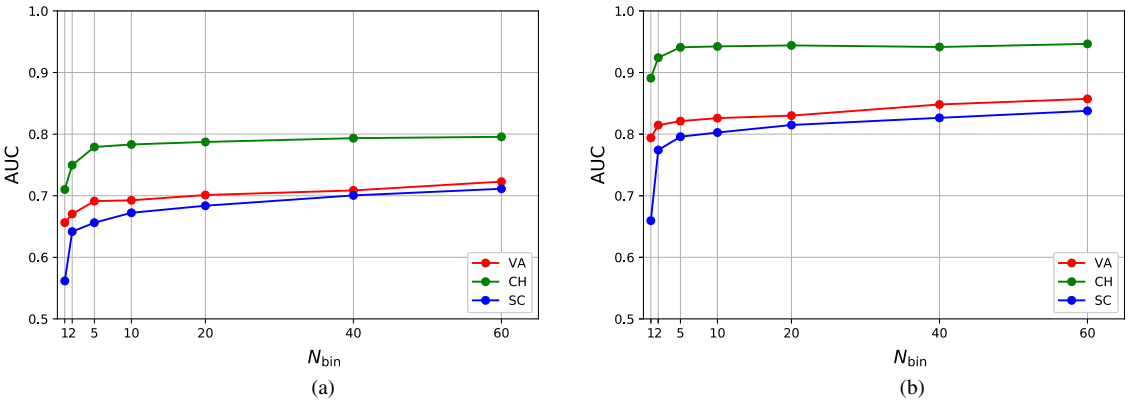

FIG. 18. Zero-jet AUCs of trainings on histograms in which the p^e_T range [1350, 2550] GeV is binned into 1, 2, 3, 4, 5, 10, 20, 40, 60 bins with uncovered bins left empty. Samples of $S/B = (a)\,0.4$ and (b) 1.2 are used.

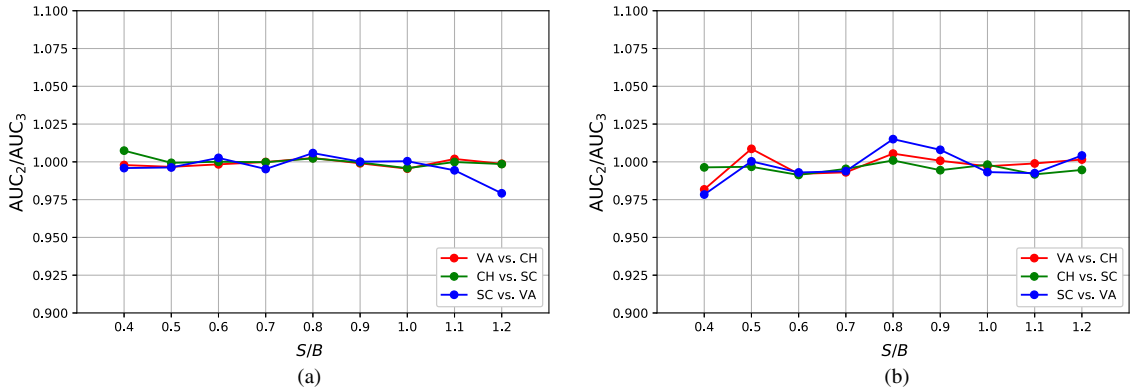

FIG. 19. AUCs ratios of the binary to the projected ternary of (a) zero-jet and (b) one-jet dedicated to VA vs CH, CH vs SC, and SC vs VA for different S/B ratios. The AUCs for VA vs CH are depicted in red, CH vs SC in green, and SC vs VA in blue.

and class j. We *project* the score components of the ternary by defining

$$P'_k = \frac{P_k}{P_i + P_j}, \qquad k = i, j. \tag{A3}$$

We then go on to compare the projected zero- and one-jet AUCs (AUC_3) with the AUCs given by the true binary classifier (AUC_2) dedicated to classes i and j in terms of the ratio $\mathrm{AUC}_2/\mathrm{AUC}_3$. Figure 19 shows these ratios dedicated to VA vs CH, CH vs SC, and SC vs VA for different S/B ratios. The plots show that the projected ternary AUCs are consistent with the binary AUCs, implying that the ternary classifier gives the same level of performance in binary classifications as the dedicated binary classifiers.

3. Applying the wrong models

Another interesting question is what would happen if the *wrong* models are applied to a set of testing samples. There are two variables to test this in our analysis: wrong significance and wrong decay widths. In the following, we show the two corresponding tests.

The first test is to use the models trained on zero-jet samples of 4.5 TeV resonances with $\Gamma_{\mathrm{NP}} \approx 500$ GeV to test the samples of $\Gamma_{\mathrm{NP}} \approx 200$ GeV at a fixed S/B and vice versa, as well as between samples of $\Gamma_{\mathrm{NP}} \approx 200$ GeV and $\Gamma_{\mathrm{NP}} \approx 50$ GeV, and samples of $\Gamma_{\mathrm{NP}} \approx 500$ GeV and $\Gamma_{\mathrm{NP}} \approx 50$ GeV. We then calculate the ratios of the "wrong AUCs" (AUC) to the "correct AUCs" (AUC_0) with respect to different significances. To compare with Fig. 1(e), we show the p_T^e vs η^e distributions for $\Gamma_{\mathrm{NP}} \approx 500$, 50 GeV in Fig. 20. The training results are shown in Fig. 21. We can see that applying models of the wrong widths still has some discriminating power, yet they are consistently worse than applying the correct models. This indicates the importance of getting the right order of magnitude for Γ_{NP} before setting up the trainings, and shows that even an incorrectly trained NN still has an AUC within ∼10–25% of the correctly trained model.

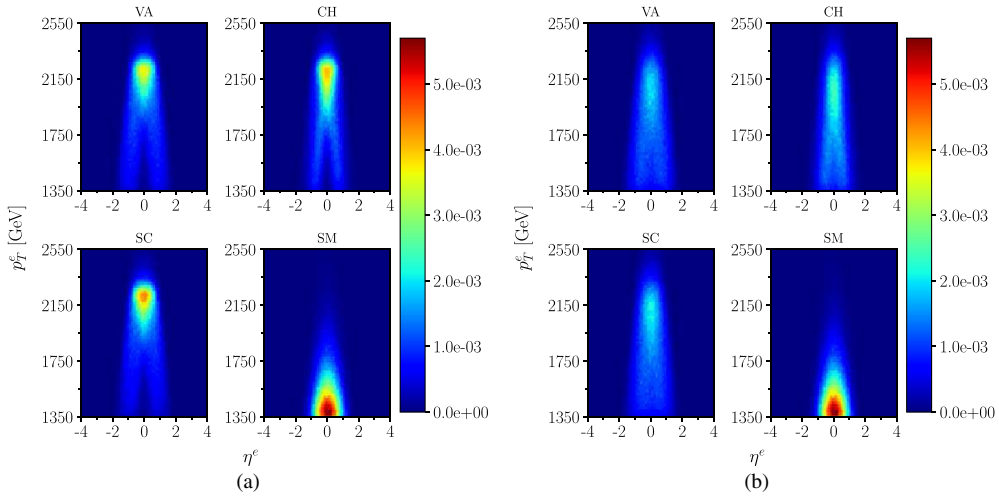

FIG. 20. p_T^e vs η^e distributions for 4.5 TeV resonances with $\Gamma_{\mathrm{NP}} \approx$ (a) 50 and (b) 500 GeV.

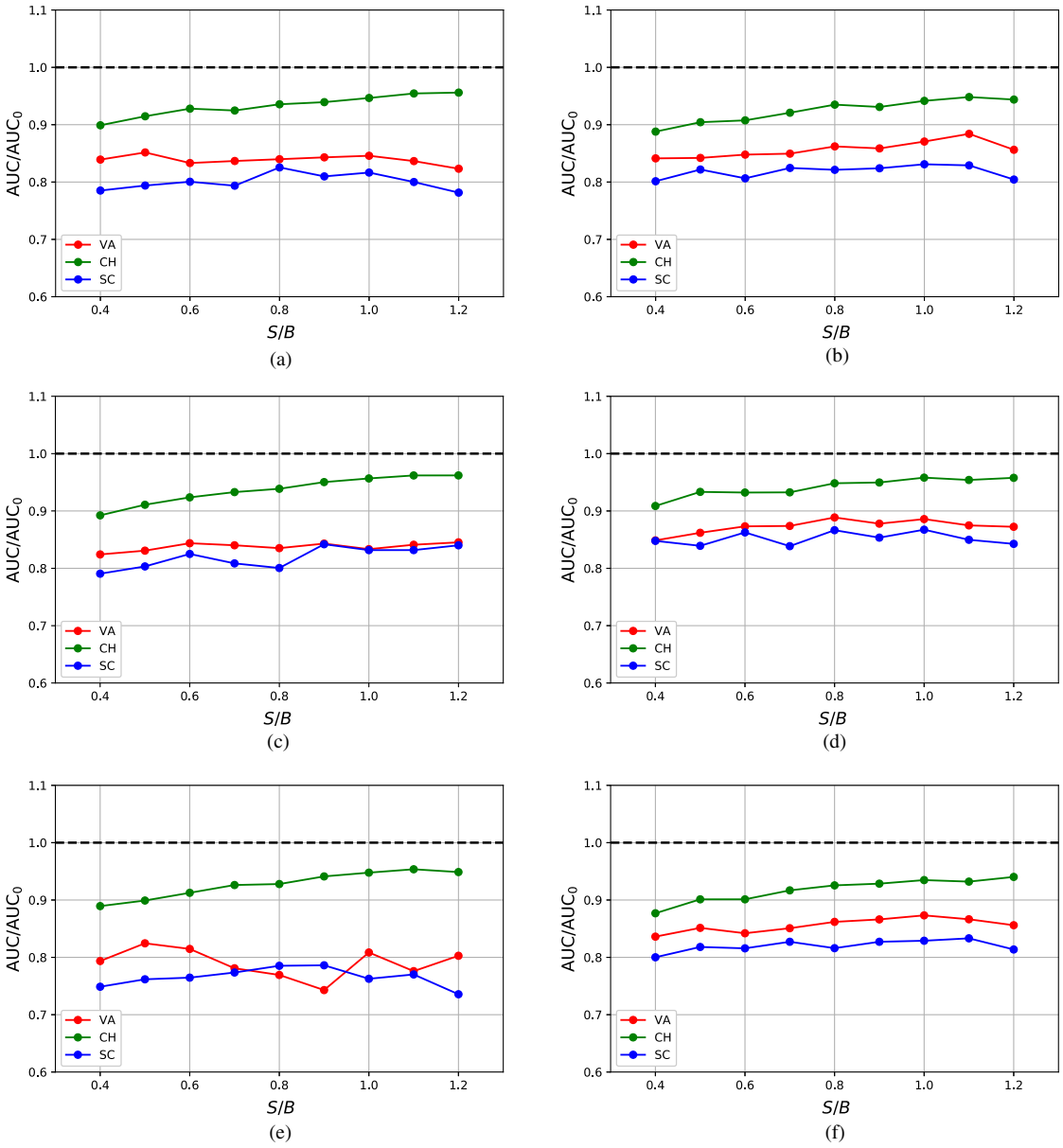

FIG. 21. Ratios of the zero-jet AUCs from the tests with (a) $\Gamma_{\rm NP} \approx 500$ GeV models applied to $\Gamma_{\rm NP} \approx 200$ GeV samples, (b) $\Gamma_{\rm NP} \approx 200$ GeV models applied to $\Gamma_{\rm NP} \approx 500$ GeV samples, (c) $\Gamma_{\rm NP} \approx 200$ GeV models applied to $\Gamma_{\rm NP} \approx 50$ GeV samples, (d) $\Gamma_{\rm NP} \approx 50$ GeV models applied to $\Gamma_{\rm NP} \approx 200$ GeV samples, (e) $\Gamma_{\rm NP} \approx 500$ GeV models applied to $\Gamma_{\rm NP} \approx 50$ GeV samples, and (f) $\Gamma_{\rm NP} \approx 50$ GeV models applied to $\Gamma_{\rm NP} \approx 500$ GeV samples, to the AUCs using the correct models in the low-significance scenarios.

The second test is to use the models trained on zero-jet samples of $S/B = 0.4$, 0.8, 0.12 to test the samples of $S/B \in [0.4, 1.2]$ for 4.5 TeV resonances with fixed $\Gamma_{\rm NP} \approx 200$ GeV. We also calculate the ratios of the "wrong AUCs" (AUC) to the "correct AUCs" (AUC$_0$) for different significances and show them in Fig. 22. The plots show that the wrong models are still able to yield reasonable results in the vicinity of the trained significance level. This result shows that some deviation from the correct significance is all right if one is satisfied with performance within 10%.

These two comparisons indicate that when applying our analysis to the parameter space of the signal hypotheses, even a coarse set of FNNhs covering the allowed parameter space will still have reasonable performance for a model with a decay width or significance different than the ones used for the set of FNNhs, allowing a reduction of computing resources with the trade-off of a small drop in performance.

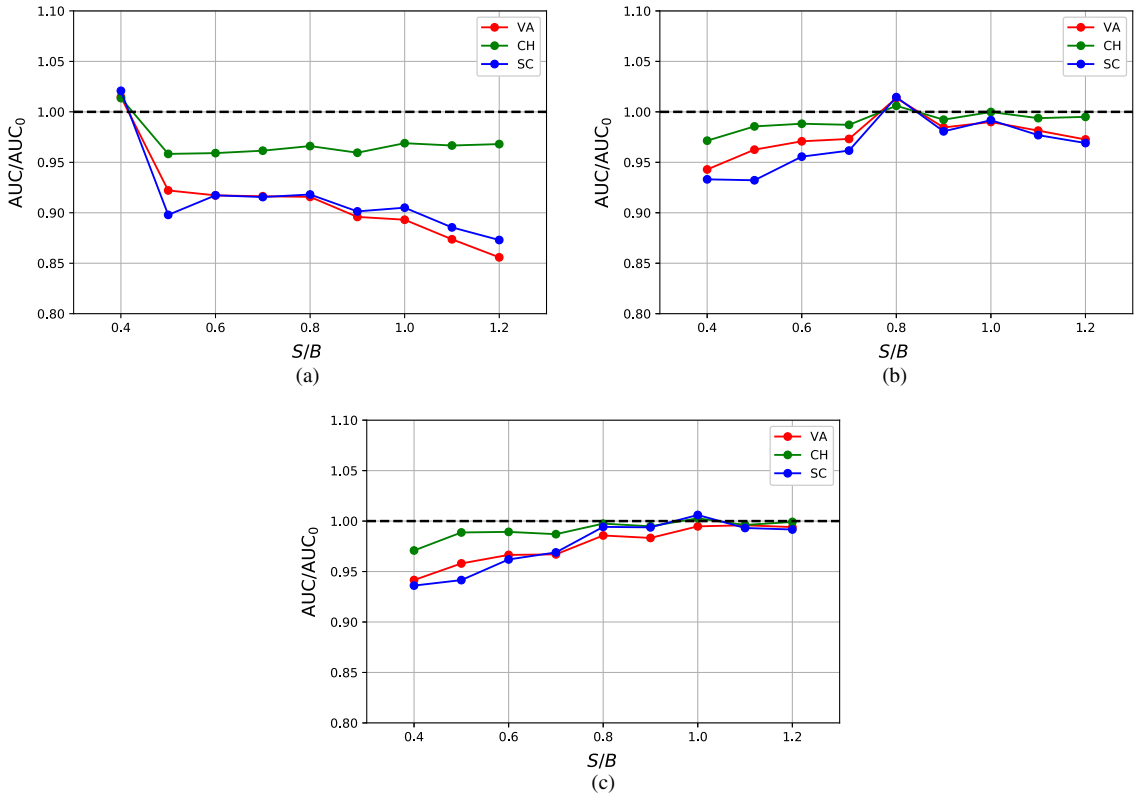

FIG. 22. Ratios of the zero-jet AUCs from the tests with (a) $S/B = 0.4$, (b) $S/B = 0.8$, and (c) $S/B = 0.12$ models applied to samples of different S/B ratios to the correct AUCs, using 4.5 TeV resonance samples with $\Gamma_{\mathrm{NP}} \approx 200$ GeV.

[1] G. Arnison *et al.* (UA1 Collaboration), Phys. Lett. **122B**, 103 (1983).
[2] M. Banner *et al.* (UA2 Collaboration), Phys. Lett. B **122**, 476 (1983).
[3] G. Aad *et al.* (ATLAS Collaboration), J. High Energy Phys. 03 (2020) 145.
[4] A. M. Sirunyan *et al.* (CMS Collaboration), J. High Energy Phys. 05 (2020) 033.
[5] G. Aad *et al.* (ATLAS Collaboration), arXiv:2005.02983.
[6] G. Aad *et al.* (ATLAS Collaboration), J. High Energy Phys. 06 (2020) 151.
[7] A. M. Sirunyan *et al.* (CMS Collaboration), J. High Energy Phys. 06 (2018) 128.
[8] G. Aad *et al.* (ATLAS Collaboration), Phys. Rev. D **100**, 052013 (2019).
[9] A. M. Sirunyan *et al.* (CMS Collaboration), Phys. Lett. B **792**, 107 (2019).
[10] A. M. Sirunyan *et al.* (CMS Collaboration), Eur. Phys. J. C **80**, 237 (2020).
[11] G. Aad *et al.* (ATLAS Collaboration), J. High Energy Phys. 09 (2019) 091; 06 (2020) 042.
[12] P. Zyla *et al.* (Particle Data Group), Prog. Theor. Exp. Phys. **2020**, 083C01 (2020).
[13] P. Langacker, R. W. Robinett, and J. L. Rosner, Phys. Rev. D **30**, 1470 (1984).
[14] F. Abe *et al.* (CDF Collaboration), Phys. Rev. Lett. **74**, 2626 (1995).
[15] T. G. Rizzo, J. High Energy Phys. 05 (2007) 037.
[16] L.-T. Wang and I. Yavin, Int. J. Mod. Phys. A **23**, 4647 (2008).
[17] S. Gopalakrishna, T. Han, I. Lewis, Z.-g. Si, and Y.-F. Zhou, Phys. Rev. D **82**, 115020 (2010).
[18] O. Eboli, C. S. Fong, J. Gonzalez-Fraile, and M. Gonzalez-Garcia, Phys. Rev. D **83**, 095014 (2011).
[19] C.-W. Chiang, N. D. Christensen, G.-J. Ding, and T. Han, Phys. Rev. D **85**, 015023 (2012).
[20] G. Kasieczka, T. Plehn, M. Russell, and T. Schell, J. High Energy Phys. 05 (2017) 006.
[21] J. Brehmer, K. Cranmer, G. Louppe, and J. Pavez, Phys. Rev. Lett. **121**, 111801 (2018).
[22] J. Brehmer, K. Cranmer, G. Louppe, and J. Pavez, Phys. Rev. D **98**, 052004 (2018).
[23] U. Simola, B. Pelssers, D. Barge, J. Conrad, and J. Corander, J. Instrum. **14**, P03004 (2019).
[24] G. Kasieczka, N. Kiefer, T. Plehn, and J. M. Thompson, SciPost Phys. **6**, 069 (2019).

[25] F. F. Freitas, C. K. Khosa, and V. Sanz, Phys. Rev. D **100**, 035040 (2019).
[26] G. Kasieczka, T. Plehn, A. Butter, K. Cranmer, D. Debnath, B. M. Dillon, M. Fairbairn, D. A. Faroughy, W. Fedorko, C. Gay *et al.*, SciPost Phys. **7**, 014 (2019).
[27] C. K. Khosa, V. Sanz, and M. Soughton, arXiv:1910.06058.
[28] Y.-C. J. Chen, C.-W. Chiang, G. Cottin, and D. Shih, Phys. Rev. D **101**, 053001 (2020).
[29] Y. S. Lai, arXiv:1810.00835.
[30] Y.-L. Du, K. Zhou, J. Steinheimer, L.-G. Pang, A. Motornenko, H.-S. Zong, X.-N. Wang, and H. Stöcker, Eur. Phys. J. C **80**, 516 (2020).
[31] A. Mullin, H. Pacey, M. Parker, M. White, and S. Williams, arXiv:1912.10625.
[32] F. Flesher, K. Fraser, C. Hutchison, B. Ostdiek, and M. D. Schwartz, arXiv:2011.04666.
[33] M. Lazzarin, S. Alioli, and S. Carrazza, arXiv:2010.02213.
[34] Y. S. Lai, D. Neill, M. Płoskoń, and F. Ringer, arXiv:2012.06582.
[35] 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.
[36] T. Sjostrand, S. Mrenna, and P. Z. Skands, J. High Energy Phys. 05 (2006) 026.
[37] T. Sjstrand, 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).
[38] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lematre, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), J. High Energy Phys. 02 (2014) 057.
[39] M. Selvaggi, J. Phys. Conf. Ser. **523**, 012033 (2014).
[40] A. Mertens, J. Phys. Conf. Ser. **608**, 012045 (2015).
[41] M. Cacciari, G. P. Salam, and G. Soyez, Eur. Phys. J. C **72**, 1896 (2012).
[42] M. Cacciari, G. P. Salam, and G. Soyez, J. High Energy Phys. 04 (2008) 063.
[43] R. D. Ball *et al.*, Nucl. Phys. **B867**, 244 (2013).
[44] A. Alloul, N. D. Christensen, C. Degrande, C. Duhr, and B. Fuks, Comput. Phys. Commun. **185**, 2250 (2014).
[45] F. Chollet *et al.*, Keras, https://keras.io, 2015.
[46] M. Abadi *et al.*, TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow.org.
[47] D. P. Kingma and J. Ba, arXiv:1412.6980.
[48] B. Nachman and J. Thaler, arXiv:2101.07263.