Distinguishing $W'$ Signals at Hadron Colliders Using Neural Networks

Spencer Chang, Ting-Kuo Chen, and Cheng-Wei Chiang

$^1$Department of Physics and Institute for Fundamental Science
University of Oregon, Eugene, Oregon 97403, U.S.A.

$^2$Department of Physics, National Taiwan University, Taipei 10617, Taiwan

Abstract

We investigate a neural-network (NN)-based hypothesis test to distinguish different $W'$ and charged scalar resonances through the $\ell + \mathcal{E}_T$ channel at hadron colliders. This is traditionally challenging due to a four-fold ambiguity at proton-proton colliders, such as the Large Hadron Collider. Of the neural network approaches we studied, we find a multi-class classifier based on a convolutional neural network (CNN) to be the best approach, where the CNN is trained on 2D histograms made from the transverse momentum $p_T$ and pseudorapidity $\eta$ of $\ell$. The CNN performance is quite impressive and can begin to distinguish between hypotheses when the signal to background ratio is above 10%, with near perfect performance for $S/B \gtrsim 60\%$. In addition, the performance is quite robust against variations in the signal such as the overall signal strength and the decay width of the resonance. As a comparison to traditional approaches, we compare our method with Bayesian hypothesis testing and discuss the pros and cons of each approach. Finally, by considering the next-to-leading order (NLO) process with an additional jet, we demonstrate that one can generalize the CNN to multi-dimensional histograms by utilizing RGB colors to represent different variable pairs. The neural network scheme presented in this paper is a powerful tool that could help investigate the properties of charged resonances and more generally can be applied to many other hypothesis testing situations.
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 di-jet channel conducted by ATLAS [3], and the 13 TeV search with $\ell + j$ [4] and $\ell + \not E_T$ [5] final states conducted by CMS. So far, the mass limits have been pushed above the TeV level (see ref. [6]), and thus future $W'$ signals are expected to occur at higher energies 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 spins in its center-of-mass (COM) frame at the LHC suffers from two ambiguities:

- **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 anti-quark 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 [7].

- **Missing longitudinal momentum**: Since the colliding frames of the incident partons are typically boosted, we need to identify the missing longitudinal momentum to correctly determine the COM angular distribution in $\cos \theta_{\text{COM}}$. From kinematics, the longitudinal momentum can be solved from a quadratic equation assuming the mediating boson to be 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 $W \rightarrow e\nu$ at the SPS $p\bar{p}$ Collider [11] and top pair production at the Tevatron [8].

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. [9-13].
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 on a variety of challenging problems in different fields. Some recent high energy physics applications include refs. [14–23].

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 multi-class 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 it in event-by-event as an array, or combine a number of events and form 2D histograms by choosing a certain pair of variables. The latter would be similar to feeding in part of the probability density function on the chosen 2D kinematic plane. We have considered the following three NN models for this problem:

- **Deep Neural Network** (DNN): We constructed a simple DNN trained upon the kinematic information of $\ell$ from each individual event.

- **Convolutional Neural Network** (CNN): We constructed a simple CNN trained upon 2D histograms made from pairs of kinematic observables of a certain number of events.

- **Transfer-Learning Network** (TLN): As a more sophisticated CNN model, a TLN has a part called the *base model*, which is modified from a publicly available pre-trained model, and another part called the *top model*, which links the output of the base model to the target output layer. We choose the VGG-19 model [24] as the base model, and import the pre-trained weights from the ImageNet database [25]. The TLN serves as a comparison with our own CNN.

The first two methods mentioned above have already been proposed and used in ref. [22] to distinguish the mono-jet and di-jet signatures of weakly interacting massive particles (WIMPs) from those of the SM and other dark matter models. During our study, we found
that the DNN could barely distinguish among the three classes, while the CNN outperformed
the TLN. The first comparison shows that changing from individual-event identification to
“global” feature identification would increase the efficiencies. The second comparison tells
that there is no need for a sophisticated NN model to solve this problem, as the features
within the base model of TLN do not seem general enough to outperform our own fully
trained CNN. Therefore, we will only demonstrate the results of the CNN in this paper.

In our study, we investigate the application of this method to the classification of samples
into the following three coupling classes\footnote{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 $\sim W'_\mu \bar{\psi} \gamma^\mu (a + b \gamma_5) \psi$. The discriminating power, of course, will depend upon how close the different coupling classes are.}:

- **Vector/Axial (VA):** This class corresponds to a $W'$ with vector-like (V) fermionic
couplings, $\sim W'_\mu \bar{\psi} \gamma^\mu \psi$, or axial-vector-like (A) fermionic couplings, $\sim W'_\mu \bar{\psi} \gamma^\mu \gamma_5 \psi$.

- **Chiral (CH):** This class corresponds to a $W'$ with left-handed (LH) fermionic couplings,
$\sim W'_\mu \bar{\psi} \gamma^\mu (1 - \gamma_5) \psi$, or right-handed (RH) fermionic couplings, $\sim W'_\mu \bar{\psi} \gamma^\mu (1 + \gamma_5) \psi$.

- **Scalar (SC):** This class corresponds to an $H$ with Yukawa-like fermionic couplings,
$\sim H \bar{\psi} \psi$.

For a $pp$ collider, we will show that for signal alone the $p_T$ and $\eta$ variables of the lepton
cannot distinguish between the V and A hypotheses or between the LH and RH hypothe-
ses. 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.

For the sake of simplicity, we only focus on positively charged resonances, $W'^+ \text{ and } H^+$,
with masses of 1 TeV, and analyze the $e^+ \nu_e$ final state, although this can be applied to
higher masses, negatively-charged final states, and to the muon final states as well. Also,
we assume that the coupling strengths and structure are universal to both the quark and
the lepton sectors (even for $H^+$), and to all generations.

We also take into consideration the effects of different boson resonance widths, varying
from truly narrow widths to sizeable ones for the 1 TeV resonance. Explicitly, we considered
widths of $\sim 100, 10, 1, \text{ and } 0.1 \text{ GeV. However, it turns out that the training outcomes upon}$
different widths are quite similar. We will focus on the samples of 10-GeV width, a choice to mimic the SM $W$ width-to-mass ratio $\Gamma_W/m_W \approx 1/40$, in most of our presentation below.

Beside the leading-order (LO) process, we have also studied the next-to-leading-order (NLO) process in which an extra jet is produced in the final state. To account for the extra information provided by the jet, we will further extend the 2D histogram inputs of our CNN to include more variable pairs by using the three RGB colors, to demonstrate that the CNN approach of Ref. [22] can be generalized to more dimensions. We will formulate a few different input schemes for these NLO histograms, although there is no major performance difference among the different schemes. To understand these 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 this, our results show that the CNN approach is more convenient than and superior to conventional methods, such as Bayesian hypothesis or $\chi^2$ tests.

We prepare the samples assuming 14-TeV $pp$ collisions, which is the expected COM energy of LHC Run-III. Going beyond the signal-only hypothesis testing of [22], we will also include the SM background from the $W$ boson. We will investigate scenarios of different $S/B$ and $S/\sqrt{B}$ assuming an integrated luminosity of $\mathcal{L} = 60$ fb$^{-1}$, a value comparable to the amount of the LHC Run-II data collected in 2018. This value will affect two things: the labelling of $S/\sqrt{B}$, which can be substituted by the $\mathcal{L}$-independent $S/B$; and the number of events used to form individual histograms, whose effects have also been explored in our study and provided in the Appendix. As our results will demonstrate, the CNN can start distinguishing the signal hypotheses when $S/B = 0.1$ and have nearly perfect performance for $S/B \gtrsim 0.6$. Overall, our approach can in some sense be viewed as a model-independent approach, as it does not require specific model details except for the masses, widths, and couplings of the heavy bosons. This will be justified in Sec. [11]

In the Appendix, we further provide details of technical studies of the CNN performance when varying the event numbers per histogram and corresponding total sample sizes, resolutions, and kinematic windows. In addition, we investigate the results of applying the wrong models on the testing samples. Finally, 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 the individual binary classifiers.
This paper is organized as the following. In Sec. II, we briefly review the kinematic properties of bosons of different coupling classes. In Sec. III, we discuss the LO and NLO samples and analyze their kinematic features. In Sec. IV, we describe the details of our CNN model as well as the training specifications. In Sec. V, we present and discuss the LO and NLO training results. In Sec. VI, we compare our NN method with a few traditional hypothesis tests and discuss the pros and cons. In Sec. VII, we draw conclusions and propose possible further studies. Finally, more technical details of our investigations are provided in Appendix A.

II. PARTON-LEVEL ANALYSIS OF GENERAL SINGLY-CHARGED BOSONS

Consider the following processes:

\[ p p \rightarrow W^+/W'^+ / H^+ \rightarrow e^+\nu_e . \]  

(1)

The corresponding \( p_T \) and \( \eta \) differential cross sections of \( e^+ \) are given by

\[
\frac{d\hat{\sigma}}{d\chi} = \sum_{q,q'} \int dx dy \frac{d\hat{\sigma}(x,y)}{d\chi} |V_{qq'}|^2 \cdot q(x,Q^2)\bar{q}(y,Q^2), \quad (\chi = p_T, \eta)
\]  

(2)

where \( V_{qq'} \) is the Cabibbo-Kobayashi-Maskawa (CKM) matrix element and \( q(x,Q^2), \bar{q}(y,Q^2) \) are the parton distribution functions (PDFs).

The parton-level \( p_T \) and \( \eta \) differential cross sections for \( H \) and \( W' \) are given respectively by

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

(3a)

and

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

(3b)

and

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

(4a)

\[
\frac{d\hat{\sigma}_{W'}}{d\eta} = \text{sech}^2 \eta \frac{128E_1^2E_2^2}{32\pi(p^2 - m_{W'}^2)^2 + m_{W'}^2 \Gamma_{W'}^2} \left\{ 2(c_V^2 + c_A^2)^2 \left[ \frac{I(E_1, E_2, \eta)}{H(E_1, E_2, \eta)} + \frac{I(E_2, E_1, \eta)}{H(E_2, E_1, \eta)} \right] + 4c_V^2c_A^2 \left\{ \frac{J(E_1, E_2, \eta)}{H(E_1, E_2, \eta)} + \frac{J(E_2, E_1, \eta)}{H(E_2, E_1, \eta)} \right\} \right\},
\]

(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

\[
F(A, B, \eta) \equiv (A + B)^2 + (A - B)^2 \tanh^2 \eta ,
\]

\[
G(A, B, \eta) \equiv (A + B)^2 - (A - B)^2 \tanh^2 \eta ,
\]

\[
H(A, B, \eta) \equiv [(A + B) - (A - B) \tanh \eta]^4 ,
\]

\[
I(A, B, \eta) \equiv A^2(1 - \tanh \eta)^2 + B^2(1 + \tanh \eta)^2 ,
\]

\[
J(A, B, \eta) \equiv A^2(1 - \tanh \eta)^2 - B^2(1 + \tanh \eta)^2 .
\]

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 \( c_V^2 c_A^2 \) and would lead to distinct \( \eta \) distributions for different \( W' \) coupling scenarios. Thus, combining the parton-level \( p_T \) and \( \eta \) 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 PDF’s, 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 \texttt{MadGraph5_aMC@NLO v2.7.0} \cite{26}, followed by parton shower and hadronization performed with \texttt{Pythia 8.2.44} \cite{27}. The samples are then passed to \texttt{Delphes 3.4.2} \cite{28} for detector simulation using the default CMS card. The events are reconstructed with \texttt{FastJet 3.3.2} \cite{29}. In particular, the final-state jets in the NLO processes are reconstructed using the anti-\( k_T \) clustering algorithm \cite{30} with the cone radius \( R = 0.4 \).

The processes are simulated for 14-TeV LHC collisions with the \texttt{NNPDF23_nlo_as_0119} \cite{31} PDF set. The \( W' \)- and \( H \)-mediated processes are generated respectively with the Wprime model and General 2HDM from the \texttt{FeynRules} \cite{32} model database. For concreteness, the new resonance masses \( m_{W'} = m_H = 1 \) TeV. In what follows, we describe the details of the LO and NLO samples.
A. LO samples

The LO samples are generated for the processes:

\[ pp \rightarrow W'^+ / H^+/W^+ \rightarrow e^+ \nu_e . \]

The cuts imposed at the generator level are summarized in TABLE I. The selection cut is imposed to avoid the tail of SM \( W \) background while retaining a sufficient amount of the new-physics (NP) signals below the Jacobian peak at \( p_T^\ell = m_{W'} / 2 = m_{H^+} / 2 = 500 \) GeV. This \( p_T \) cut is a practical one so that the CNN 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. However, in the Appendix, we will explore how the CNN performance depends on the \( p_T \) cut, where we show there can be a trade-off in information loss (too high of a cut) and \( p_T \) resolution (too low of a cut).

| Basic cuts | \( p_T^\ell > 10 \) GeV ; \( |\eta^\ell| < 2.5 \) |
|-------------|---------------------------------------------|
| Selection cut | \( p_T^\ell > 300 \) GeV |

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

We denote the new boson width by \( \Gamma_{NP} \) and consider four different values: 100, 10, 1, and 0.1 GeV. We will show in Sec. V that the width varying in this range does not affect the training outcomes much. At the generator level, we generate 1M events for each of the VA, CH, SC, and SM classes. After detector simulation, the successfully tagged event numbers are roughly as listed in TABLE II for all four widths.

| Number | Class | VA   | CH   | SC   | SM   |
|--------|-------|------|------|------|------|
|        | \( \Gamma_{NP} \approx 100 \) GeV | 717K | 736K | 719K |      |
|        | \( \Gamma_{NP} \approx 10 \) GeV   | 716K | 737K | 716K | 733K |
|        | \( \Gamma_{NP} \approx 1 \) GeV    | 716K | 738K | 717K |      |
|        | \( \Gamma_{NP} \approx 0.1 \) GeV  | 717K | 737K | 717K |      |

TABLE II: LO event numbers for each of VA, CH, SC, and SM classes and for \( \Gamma_{NP} \approx 100, 10, 1, \) and 0.1 GeV after detector simulation.
We choose to divide both $p_T^e$ and $\eta^e$ into 40 bins so as to satisfy the minimum dimension requirement of the VGG-19 model used in the TLN while our CNN still remains trainable. We only show the corresponding $p_T^e$, $\eta^e$, and $p_T^e$ vs. $\eta^e$ distributions for $\Gamma_{NP} \approx 10$ GeV in FIG. 1. As the boson width increases, the Jacobian peak of $p_T^e$ distribution would become broader, while the $\eta^e$ distribution would remain identical.

As discussed in Sec. II, ideally 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 numerical precision in the selected couplings. Since there is a much larger difference in the $\eta$ distributions, we do not expect this difference to strongly affect the training or performance. The same issue will also occur in the NLO case.

The color scheme for FIG. 1(c), 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. From FIG. 1(c), we see the Jacobian peaks at $p_T = 500$ GeV for all the NP classes, with the CH class possessing the longest tail toward low $p_T^e$, followed by the VA class and finally the SC class. Such differences in the $p_T^e$ tail and the spread in $\eta^e$ show the kinematic information that can be used to distinguish among the three classes, even after including the background.
FIG. 1: (a) $p_T^e$, (b) $\eta^e$, (c) $p_T^e$ vs. $\eta^e$ distributions for the LO samples of $\Gamma_{NP} \approx 10$ GeV. In plots (a) and (b), VA is depicted in red, CH in green, SC in blue, and SM in black. All the distributions are unit normalized. In plot (c), the color scale range goes from 0 to the maximum entry among all four classes, with the warmer/colder regions denoting more/fewer event entries. The same color scheme is applied to all the following figures.

Within the selected phase space, the fiducial cross section for the SM class is

$$\sigma_{B,LO} = 25.30 \text{ fb} .$$

Correspondingly, the number of SM events is

$$B_{LO} = \sigma_{B,LO} \times \mathcal{L} \approx 1520 .$$

Each histogram to be fed into the CNN is made up of a total of $N_{LO}$ events, where $N_{LO}$ is given by

$$N_{LO} = B_{LO} \times \left( 1 + \frac{S_{LO}}{B_{LO}} \right) ,$$

where $S_{LO}$ denotes the number of signal events and we will vary the signal-to-background ratio $S_{LO}/B_{LO}$ in our considerations.

We first study low-significance scenarios (with $S_{LO}/\sqrt{B_{LO}} = 1 - 10$), and then move on to high-significance scenarios (with $S_{LO}/B_{LO} = 0.1 - 1.5$), focusing exclusively on samples
of $\Gamma_{NP} \approx 10$ GeV. A few sample $(\eta,p_T)$ histograms for $S_{LO}/B_{LO} = 1.0$ are shown in FIG. 2. Note that it is quite challenging to distinguish them by eye at high accuracy but will be a simple job for the CNN.

![Histograms](image)

FIG. 2: Some LO input histograms for (a) VA, (b) CH, and (c) SC samples of $S_{LO}/B_{LO} = 1.0$.

In order to generate enough training histograms, we shuffle the sample events several times before grouping them into sets of $N_{LO}$, and then repeat this process until at least 15K histograms are produced for each significance scenario.

### B. NLO samples

The NLO samples are generated for the processes:

$$pp \rightarrow jW'^+/H^+/W^+, \; W'^+/H^+/W^+ \rightarrow e^+\nu_e.$$
which include three types of diagrams: initial state radiation (ISR), $gq$ $t$-channel, and $gq$ $s$-channel.

The cuts imposed at the generator level are summarized in TABLE III. The $p_T^\ell$ selection cut is imposed for the same reason as in the LO case. However, the emission of the jet allows more of the background to pass this cut, so we place an addition $E_T$ selection cut on all NLO samples.

| Basic cuts            | $p_T^\ell > 10$ GeV , $p_T^j > 20$ GeV ; $|\eta^\ell| < 2.5$ , $|\eta^j| < 5.0$ ; $\Delta R_{j\ell} > 0.4$ |
|-----------------------|--------------------------------------------------------------------------------------------------|
| Selection cuts        | $p_T^\ell > 300$ GeV , $E_T > 300$ GeV |

TABLE III: Summary of cuts imposed on the NLO samples at the generator level.

All the physical parameters are the same as in the LO $\Gamma_{NP} \approx 10$ GeV case. We also generate 1M parton-level events for each class, and the tagged event numbers after detector simulation are listed in TABLE IV.

| Class | VA  | CH  | SC  | SM  |
|-------|-----|-----|-----|-----|
| Number| 668K| 684K| 668K| 700K|

TABLE IV: NLO event numbers after selection cuts for each of VA, CH, SC, and SM classes and for $\Gamma_{NP} \approx 10$ GeV.

The corresponding SM fiducial cross section and event number within this phase space are given by

$$\sigma_{B,NLO} = 19.08 \text{ fb ,}$$

(9)

$$B_{NLO} = \sigma_{B,NLO} \times \mathcal{L} \approx 1140 .$$

(10)

As in the LO case, we also study both the low-significance and high-significance scenarios.

Since the NLO process has a three-body final state, we now have 5 physical degrees of freedom. The straightforward observables are:

- $p_T^\ell$ and $p_T^j$: transverse momenta of $e^+, j$, respectively.
- $\eta^\ell$ and $\eta^j$: pseudorapidities of $e^+, j$, respectively.
- $\Delta \phi_{ej}$: azimuthal separation between $e^+$ and $j$.

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

- $E_T$: missing transverse energy.

- $\Delta \phi_{eE}$ and $\Delta \phi_{jE}$: azimuthal separations between $e^+$ and $E_T$ and between $j$ and $E_T$, respectively.

The distributions of these kinematic observables are shown in FIG. 3.

(a) (b) (c) (d)
To choose the three pairs of variables for making our RGB histograms, we propose three schemes:

- **Physical Relation** (Scheme 1): Intuitively, the kinematic information measured from a single object should manifest high correlation. Therefore, we first pair up $p_T^e$ and $\eta^e$ as well as $p_T^j$ and $\eta^j$. Then, guessing that observables of the same mass dimension should be more related, we randomly choose two out of the three azimuthal separation variables, $\Delta \phi_{eE}$ and $\Delta \phi_{jE}$, to form the third pair.

- **Principal Component Analysis** (Scheme 2): Following Ref. [22], we also select another three pairs of variables by performing a principal component analysis (PCA). The
results are shown in TABLE V. 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 $\Delta \phi_{eE}$ and $\Delta \phi_{jE}$; and from PC-2, we pair up $p_T^e$ and $E_T$. Since $\Delta \phi_{ej}$ is already paired, we skip PC-3 and use PC-4 to pair up $\eta^e$ and $\eta^j$.

- **Common Axis** (Scheme 3): Even though the 5 degrees of freedom could be covered by choosing 3 different pairs of variables, each RGB channel would have no spatial correlations since they have different variables. On the other hand, if we set one of the two axes of the three channels to always be $p_T^e$, the CNN 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 correspondent pixels with a common $p_T^e$ coordinate. In light of this, we choose the following three pairs for scheme 3: $p_T^e$ and $\eta^e$, $p_T^e$ and $E_T$, and $p_T^e$ and $\Delta \phi_{ej}$.

|        | Variance | $p_T^e$ | $p_T^j$ | $E_T$ | $\eta^e$ | $\eta^j$ | $\Delta \phi_{ej}$ | $\Delta \phi_{eE}$ | $\Delta \phi_{jE}$ |
|--------|----------|---------|---------|-------|----------|----------|--------------------|--------------------|--------------------|
| PC-1   | 1.31     | 0.00    | 0.00    | 0.00  | 0.00     | 0.00     | 0.48               | 0.73               | −0.49              |
| PC-2   | 1.06     | 0.66    | 0.30    | 0.69  | 0.02     | −0.02    | 0.00               | 0.00               | −0.00              |
| PC-3   | 1.04     | −0.00   | −0.01   | 0.01  | 0.00     | −0.03    | 0.71               | 0.00               | 0.70               |
| PC-4   | 1.00     | −0.00   | −0.04   | 0.02  | 0.71     | 0.70     | 0.02               | 0.00               | 0.01               |
| PC-5   | 0.995    | −0.30   | 0.91    | −0.13 | 0.20     | −0.15    | 0.00               | 0.00               | 0.00               |
| PC-6   | 0.994    | 0.04    | −0.25   | 0.02  | 0.68     | −0.69    | −0.02              | 0.00               | −0.01              |
| PC-7   | 0.944    | 0.69    | 0.12    | −0.72 | 0.02     | 0.00     | 0.01               | 0.00               | 0.01               |
| PC-8   | 0.651    | 0.00    | 0.00    | 0.00  | 0.00     | 0.51     | −0.68              | −0.68              | −0.52              |

TABLE V: PCA result on NLO samples. The correlations indicate the linear components of the principal components (PCs). The higher the variance is in absolute value, the more significant the PC contributes to the diversity of the samples.

The corresponding 2D histograms are shown in FIG. 4. We will show in Sec. V that all three schemes turn out to give similar training results.
FIG. 4: NLO 2D histograms formed from variable pairs determined according to the three different schemes. Scheme 1: (a), (b), and (c); scheme 2: (c), (d), and (e); scheme 3: (a), (d), and (f).
IV. MODEL STRUCTURE AND TRAINING SPECIFICATIONS

In this section, we describe in detail the structure of our CNN model, which is constructed with the Keras [33] library along with TensorFlow [34] for backend implementation. We will also describe our training specifications, including the training parameters and strategies.

A. CNN structure

Our CNN is designed to read $40 \times 40$ 2D histograms of kinematic variable pairs as input, and to classify each histogram into one of the three signal classes. For the LO samples, we only input one channel: $p_T^e$ vs. $\eta^e$; while for NLO, we input three channels based on the three different schemes described above. Again, for scheme 1, we input $p_T^e$ vs. $\eta^e$, $p_T^j$ vs. $\eta^j$, and $\Delta \phi_{eE}$ vs. $\Delta \phi_{jE}$; for scheme 2, we input $p_T^e$ vs. $E_T$, $\eta^e$ vs. $\eta^j$, and $\Delta \phi_{eE}$ vs. $\Delta \phi_{jE}$; and for scheme 3, we input $p_T^e$ vs. $\eta^e$, $p_T^e$ vs. $E_T$, and $p_T^j$ vs. $\Delta \phi_{ej}$. The CNN structure is specified in TABLE VI. In finalizing these parameters, we found that increasing the complexity of the NN model easily leads to over-fitting, and would even result in higher instability and worse training outcomes.

B. Training specifications

In all trainings, we 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 over-training, we call for an early stopping if the validation loss has not improved by more than $2 \times 10^{-4}$ for over 100 epochs.

To evaluate the performance of our CNN, 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 the measure of the CNN performance.
|                  | LO | NLO |
|------------------|----|-----|
| Input            | $p_T^e$ vs. $\eta^e$ | 40 × 40 images |
|                  |    | RGB Color Schemes |
| Scheme 1: $p_T^e$ vs. $\eta^e$ |    | vs. $\eta_j^j$, $\Delta \phi_{eE}$ vs. $\Delta \phi_{jE}$ |
| Scheme 2: $p_T^e$ vs. $E_T$, $\eta^e$ vs. $\eta_j^j$, $\Delta \phi_{eE}$ vs. $\Delta \phi_{jE}$ |    | |
| Scheme 3: $p_T^e$ vs. $E_T$, $p_T^e$ vs. $E_T$, $p_T^e$ vs. $\Delta \phi_{ej}$ |    | |
| Layers           | batch normalization layer | |
|                  | convolutional 2D layer: 3-32 | |
|                  | max pooling 2D layer: 2-2 | |
|                  | convolutional 2D layer: 3-32 | |
|                  | max pooling 2D layer: 2-2 | |
|                  | flatten layer | |
|                  | dense layer: 128 | |
| Layer settings   | hidden layer activation = **relu** | |
|                  | output layer activation = **softmax** | |
| Compilation      | loss = **categorical_crossentropy** | |
|                  | optimizer = **adam** | |
|                  | metric = **accuracy** | |

This means that the filter kernel dimension is $3 \times 3$, and that there are 32 nodes in the convolutional layer.

This means that the max pooling kernel dimension is $2 \times 2$, and that each stride is 2 pixels.

This means that there are 128 nodes in the dense layer.

**TABLE VI**: LO and NLO CNN structure specifications.

**V. TRAINING RESULTS**

In this section, we present the trained CNN results of the LO and NLO processes for various significances. We refer some more technical details to Appendix A. For the NLO case, we further investigate the importance of individual kinematic observable pairs.
A. LO results

FIG. 5: LO low-significance training outcomes for samples of $\Gamma_{\text{NP}} \approx$ (a) 100, (b) 10, (c) 1, and (d) 0.1 GeV. The AUCs for the NN to identify VA against non-VA are depicted red, CH against non-CH in green, and SC against non-SC in blue. The same color scheme applies to all the subsequent figures.

We present the low-significance training outcomes in FIG. 5 with different boson widths taken into consideration. Low-significance scenarios refer to those with $S/\sqrt{B} = 1, 2, \cdots, 10$. As shown in the figures, the CNN can already start to distinguish the signal scenarios when $S/B \gtrsim 0.1$ and steadily improve with higher signal purities. Interestingly, $S/B = 0.1$ is close to the discovery threshold $S/\sqrt{B} = 5$, suggesting that these technique can be used with a similar luminosity as a discovery analysis, given the caveat that we have not attempted
to optimize for discovery in our original cuts. For all the four different $\Gamma_{NP}$ samples, the AUCs are roughly consistent with one another among three three NNs, suggesting that the information of boson width 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 $\Gamma_{NP} \approx 10$ GeV in what follows. As another comparison, we found that the TLN did not give better results, which showed that this problem had features not contained in the base model and confirmed that our CNNs were properly trained with good performance.

Next, we present the high-significance results in FIG. 6. High-significance scenarios refer to those with $S/B = 0.1, 0.2, \cdots, 1.5$ which all have $S/\sqrt{B} \geq 4$. All the AUCs are steadily increasing with $S/B$ and reach nearly perfect identification rates for $S/B \gtrsim 0.6$. In both the low-significance and high-significance scenarios, CH class is always the easiest to be identified, while VA and SC are more difficult.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig6.png}
\caption{LO high-significance training outcomes samples of $\Gamma_{NP} \approx 10$ GeV.}
\end{figure}

To give another more interpretable metric, we present in FIG. 7 the “accuracies” (ACCs) of our CNN. The accuracy here (and the NLO case below) is to be understood as the class-wise 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. Although the curves are less stable in low-significance scenarios, the CNN can achieve $\sim 80\%$ accuracy.
for all classes at $S/B \approx 0.3$, and $\sim 90\%$ at $S/B \approx 0.5 - 0.6$, showing a similar performance turn on as the AUC curves. Even though accuracy is more interpretable, we continue to focus on AUC as the more conventional metric to compare performance of our classifiers.

![AUC Curve](image)

**FIG. 7:** LO low- and high-significance accuracies for samples of $\Gamma_{NP} \approx 10$ GeV.

### B. NLO results

For the NLO case, $\Gamma_{NP}$ is also set at $\approx 10$ GeV for reasons stated before. We first present the low-significance training outcomes in FIG. 8. Comparing FIG. 8 with FIG. 5, we see that the NLO performance is roughly at the same level as the LO performance. This suggests that our technique can generalize to higher-dimensional approaches, thus broadening the range of viable channels to be studied. Moreover, there is no significant difference among the three different schemes. Scheme 1 is slightly better at picking out the CH model at the expense of the performances of the other two in comparison with Schemes 2 and 3. Due to their similarity, we only focus on Scheme 3 in the high-significance training. The high-significance training outcomes are shown in FIG. 9. Compared to FIG. 6, we see that the NLO AUC curves saturate later at $S/B \approx 1.0$, but it is difficult to make a fair comparison as there are more events in the LO histograms for a given $S/B$. 

21
FIG. 8: NLO low-significance training outcomes for (a) Scheme 1, (b) Scheme 2, and (c) Scheme 3 for samples of $\Gamma_{NP} \approx 10$ GeV. (a) Scheme 1 utilizes $p_T^e$ vs. $\eta^e$, $p_T^j$ vs. $\eta^j$, and $\Delta \phi_{eE}$ vs. $\Delta \phi_{jE}$. (b) Scheme 2 utilizes $p_T^e$ vs. $E_T^e$, $\eta^e$ vs. $\eta^j$, and $\Delta \phi_{eE}$ vs. $\Delta \phi_{jE}$. (c) Scheme 3 utilizes $p_T^e$ vs. $E_T^e$, $p_T^e$ vs. $E_T^j$, and $p_T^j$ vs. $\Delta \phi_{ej}$.

FIG. 9: NLO high-significance training outcomes for Scheme 3 for samples of $\Gamma_{NP} \approx 10$ GeV.
We also present in FIG. 10 the accuracies for NLO. The curves are unstable at low-significance levels, as we saw in the LO accuracies, but the CNN can achieve $\sim 80\%$ accuracy for all classes at $S/B \approx 0.4$, and $\sim 90\%$ at $S/B \approx 0.6 - 0.7$.

![Graph showing NLO low- and high-significance accuracies for samples of $\Gamma_{NP} \approx 10$ GeV.](image)

**FIG. 10:** NLO low- and high-significance accuracies for samples of $\Gamma_{NP} \approx 10$ GeV.

To understand the importance of the different variable pairs in the NLO CNN, we have also trained the CNN on single pair histograms. The individual training outcomes using a CNN trained on the individual histograms in the three NLO schemes are shown in FIG. 11. Clearly, $p_T^e$ vs. $\eta^e$ plays the most important role in the class discrimination, while $p_T^j$ vs. $\eta^j$ and $\Delta\phi_{eE}$ vs. $\Delta\phi_{jE}$ barely contribute. This is physically understandable as we expect the angular and coupling information of the leptonic decay to be preserved mostly in $e^+$, which is a direct decay product of the new charged bosons, rather than in $j$. Following $p_T^e$ vs. $\eta^e$ are $\eta^e$ vs. $\eta^j$, $p_T^{\Delta e}$ vs. $\Delta\phi_{ej}$, and $p_T^{\Delta j}$ vs. $E_T^j$, with the first two best at identifying the CH class and the latter two identifying the SC class. However, in all cases, VA is always the most difficult to be identified. Compared to FIG. 9 we see that combining different channels does lead to better overall performances, thus demonstrating that the multi-dimensional CNN can successfully utilize the additional information in these channels.
FIG. 11: NLO high-significance training outcomes using individual channels: (a) $p_T^e$ vs. $\eta^e$, (b) $p_T^j$ vs. $\eta^j$, (c) $\Delta \phi_{eE}$ vs. $\Delta \phi_{jE}$, (d) $p_T^e$ vs. $E_T$, (e) $\eta^e$ vs. $\eta^j$, and (f) $p_T^e$ vs. $\Delta \phi_{ej}$. 

24
VI. COMPARISON WITH BAYESIAN HYPOTHESIS TEST

Finally, to give context for our CNN approach, we compare the LO results with a standard hypothesis test, the Bayesian hypothesis test (BH test)\[2\] In the Bayesian approach, for a specific observed dataset $D$, the probability for it to suggest a specific hypothesis $H_k$ is given by

$$P(H_k|D) = \frac{P(D|H_k) \times P(H_k)}{\sum_k P(D|H_k) \times P(H_k)},$$

where $P(H_k)$ denotes the prior that the hypothesis $H_k$ is correct, and $P(D|H_k)$ gives the conditional probability to obtain the dataset $D$ given the fact that $H_k$ is correct. In our study, we assume that it is equally likely for all 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 are then given by

$$P(D|H_k) = \prod_{m,n} f(h_{mn}^D, H_{mn}^k)$$

where $f(h_{mn}^D, H_{mn}^k)$ denotes the Poisson probability for an observed number of counts $h_{mn}^D$ at the pixel $(m, n)$ in the 2D histogram, assuming an expectation value of $H_{mn}^k$. From the definition of Eq. (12), we could see that there would be a problem if any $H_{mn}^k = 0$ because an observed count in this pixel would have an extremely high weight in determining the hypothesis. To overcome this problem, we first symmetrize $H_k$ across the $\eta^e$ axis and then perform locally non-uniform binning to take into account small numbers of events at the pseudorapidity edges $|\eta^e| \sim 2.5$.

The low- and high-significance results for both CNN and BH tests are shown in FIG. 12. The plots indicate that the CNN approach is able to produce the same or better level of performance as the BH test for $S/B \geq 0.3$. We note that the binning strategy used in the BH test could be more difficult to resolve 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 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.

2 We have also tried to contrast with the $\chi^2$ test, but it suffers from serious issues given 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. 12: LO AUCs for samples of $\Gamma_{NP} \approx 10$ GeV, obtained using the CNN, with the histogram dimension $40 \times 40$, and using the BH tests, with the histogram dimension $40 \times 40$ and adjusted with locally non-uniform binning, for both low- and high-significance scenarios.

FIG. 13: NLO AUCs for samples of $\Gamma_{NP} \approx 10$ GeV, obtained using the CNN, with histogram dimension $40 \times 40$, and using the BH tests, with the histogram dimension $20 \times 20$, for both low- and high-significance scenarios.

In fact, we encounter such an issue when we turn to NLO. To compare to the NLO
Scheme 3 CNN, we perform a BH test using

$$P(D|H^k) = \prod_{a=1}^{3} P(D_a|H^k_a)$$  \hspace{1cm} (13)$$

where $a$ represents the three input channels. This time, it is harder to resolve the zero-bin issue properly, and hence we simply reduce the histogram dimension to $20 \times 20$. The results are shown in FIG. 13. We can see that the CNN outperforms the BH test once $S/B \gtrsim 0.2$.

As these two comparisons show, the CNN compares favorably in performance with the standard Bayesian hypothesis test. To summarize the pros and cons compared to the BH test, the CNN has the advantage that it automatically takes care of binning issues, does not require a large sample to approximate the probability density functions, and easily generalizes to higher dimensions while it has the normal neural network disadvantage of proper training and validation.

VII. CONCLUSIONS

In this paper, we have investigated the ability of a neural network to distinguish different resonances in the $pp \rightarrow W'/H \rightarrow \ell\nu\ell$ process at the LHC. We showed that the original event-by-event ambiguities in the coupling differentiation problem could be tackled with neural network classifiers with a convolutional neural network architecture using binned histograms as input images. The predicted $p_T$ distributions allow the discrimination between $H$ and $W'$, and because of the boosted parton collision frame, $W'$ with different couplings further manifest distinct $\eta$ distributions.

Extending previous signal-only analyses \cite{22}, we demonstrated that a simple CNN could start distinguishing the signals even with low signal-to-background ratio ($S/B$), saturating to nearly perfect performance at $S/B \gtrsim 0.6$. As our NLO schemes show, the 2D approach of \cite{22} can also be generalized to higher dimensions, where we took into account the extra information of the jet by using RGB channels to represent different kinematic variable pairs. These NLO performances were roughly as good as those in the LO case even with fewer number of events at the same $S/B$. Performance differences resulting from different boson

3 In principle, a 4D version of the BH test could be done with the full knowledge of the probability density function in the 4D phase space of $(p_T, E_T, \eta, \Delta \phi_{ej})$. This is challenging computationally, but it would be interesting to compare with either a 4D CNN or a 6-color 2D CNN taking in all the variables.
widths and the three 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 NLO CNNs, and found out that they had different discrimination power for the three signal classes, with some variable pairs being more suited to picking out certain classes.

Out of all the variable pairs, the CNN still relied on the information of the charged lepton the most, 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 1-TeV mass for the new charged resonance, it can be readily extended to other mass ranges that better meet the current experimental constraints. Moreover, more general studies can also be considered, such as including lepton universality violation, non-universal couplings to quark and lepton sectors, different channels other than the LO and NLO process presented here, etc. More generally, the technique we have explored can easily be applied to any hypothesis testing scenario. For possible future explorations, inputs of more than 3 channels and higher-dimensional “super-images” would be interesting directions to extend the current approach.

ACKNOWLEDGMENTS

The work of SC was supported in part by the U.S. Department of Energy under Grant Number DE-SC0011640. The works of TKC and CWC were supported in part by the Ministry of Science and Technology (MOST) of Taiwan under Grant Number MOST-108-2112-M-002-005-MY3. We appreciate the support of NVIDIA Corporation with the donation of a 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. SC 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.
Appendix A: Technical studies

To better understand the technical details of this method, we investigate the dependence of the LO CNN on variables such as the resolution and kinematic window. We also confirm the consistency between binary and ternary classifiers by introducing a projection of scores in the latter case, which has also been studied in ref. [23]. Finally, we demonstrate how robust the performance of the CNN is even when applied to testing samples from a distribution it is not trained on with different $S/B$ ratios and decay widths.

1. Kinematic window and resolution

We expect the performances of the NNs to be better if we extend the phase space from $p_T^e > 300$ 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:

- 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 none. What is even worse is that the excess of SM signals may confuse the NNs and reduce its efficiency.

- Second, if one were to maintain the same $p_T$ resolution for the histogram bins, the needed NN complexity and computational resources for training would increase rapidly as $p_{T,\text{min}}$ lowers. 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 spot” that balances these issues. We base our study upon samples of $S/B = 0.1$ and $0.2$ under the cuts given in TABLE I. To extend $p_T^e$ to lower regions, we first define the following parameters:

- $B, B'$: numbers of SM events for $p_{T,\text{min}} = 300, 100$ GeV, respectively.
- $S_c, S'_c$: numbers of class $c$ events for $p_{T,\text{min}} = 300, 100$ GeV, respectively.

We generate another set of samples based upon the same settings as before, but change the selection cut from $p_T^e > 300$ GeV to $p_T^e > 100$ GeV. 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
\[
\frac{S'_c}{B'} = \frac{r_{S_c} S_c}{r_B B}.
\] (A1)

In general, \( r_{S_{VA}} \), \( r_{S_{CH}} \), and \( r_{S_{SC}} \) are all different. This would lead to histograms with different number 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:
\[
\frac{S'B'}{B''} = \frac{r_S S}{r_B B},
\] (A2)
where the ratio \( r_S/r_B \approx 0.040 \) for the example of \( p_{T,\text{min}} = 100 \) GeV. The same procedure is carried out for \( p_{T,\text{min}} = 150, \ldots, 500 \) GeV, respectively, and the corresponding histograms of dimension \( 40 \times 40 \) are then made from the mixed samples. Note that in these studies, we fix the bin size of \( \eta' \).

The AUCs of CNN trained upon LO histograms of different \( p_{T,\text{min}} \) are plotted in FIG. 14. It is clear that there exists a “sweet window” for the cut in the range of \([150, 250]\) GeV. The performance deteriorates for \( p_{T,\text{min}} \) either lower or higher than the window boundaries.

![FIG. 14: LO AUCs of training upon histograms of different \( p_{T,\text{min}} \), with their dimensions fixed to \( 40 \times 40 \) and covering the entire \( p_T \) range. The histograms are made from samples of \( S/B = (a) 0.1 \) and (b) 0.2.](image)

To pin down whether the effect of \( p_{T,\text{min}} \) is due to resolution or training, we resize the histograms to dimension \( 60 \times 60 \), and fix the \( p_T \) bin width to 10 GeV. The bins outside the cut are then filled with zeros so as to retain a uniform structure for our NNs. The results are
given in FIG. 15. One sees that the overall bell-shaped trend still manifests, which suggests that the reduced performance of a lower $p_{T,\text{min}}$ is due to incomplete training rather than the resolution.

FIG. 15: LO AUCs of training upon histograms of different $p_{T,\text{min}}$, with their dimensions fixed to $60 \times 60$, $p_T$ bin width to 10 GeV, and the uncovered bins left empty. The histograms are made from samples of $S/B = (a) \, 0.1$ and (b) $0.2$.

We further study the effect of $p_T^e$ resolution in the following way: we only use events with $p_T^e \in [300, 600]$ GeV, and slice them into 1, 2, 3, 4, 5, 10, 20, 40 bins, respectively. Samples of $S/B = 0.1$ and 0.2 are again used. To retain the same NN structure, we fill in null bins so that the histograms are still of dimension $40 \times 40$. The training outcomes are shown in FIG. 16. 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 once the binning is extremely coarse. As one increase the bins, the AUCs have nearly saturated their maximum values way before $N_{\text{bin}} = 40$. Consequently, we could infer that as long as the $\eta^e$ resolution remains sufficiently high, the $p_T^e$ resolution does not need to be maximized to obtain the optimal NN performance.
FIG. 16: LO AUCs of trainings upon histograms in which the $p_T$ range $[300, 600]$ GeV is binned into 1, 2, 3, 4, 5, 10, 20, 40 bins with uncovered bins left empty. Samples of $S/B = (a) 0.1$ and (b) 0.2 are used.

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 with 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 $i$ and class $j$, we project the score components of the ternary by defining

$$P'_k = \frac{P_k}{P_i + P_j}, \quad k = i, j.$$ (A3)

We then go on to compare the projected AUCs with the AUCs given by the true binary classifier dedicated to class $i$ and $j$. FIG. 17 shows the AUCs of the projected ternary (left) and binary (right) scores dedicated to VA vs. CH, CH vs. SC, and SC vs. VA for different $S/B$ ratios. As one can see, the projected ternary AUCs are consistent with the binary AUCs. This implies that the ternary classifier we trained gave the same level of
performances in terms of binary classifications as the dedicated binary classifiers did.

FIG. 17: LO AUCs of (a) projected ternary and (b) binary scores 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.

Also, it is interesting to see that as long as the CH class is included, the performances are all above 0.8 for $S/B \geq 0.1$, and would already reach 1.0 at $S/B = 0.3$, meaning that it is relatively easy to identify the CH class from the other two; on the contrary, it is much harder for the NN to distinguish the VA class from the SC class.

3. Applying the wrong models

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

The first is to use the models trained upon LO samples of $\Gamma_{NP} \approx 100$ GeV to test the samples of $\Gamma_{NP} \approx 0.1$ GeV at a fixed $S/\sqrt{B}$ and vice versa, as well as between samples of $\Gamma_{NP} \approx 10$ GeV and $\Gamma_{NP} \approx 1$ GeV. We then calculate the ratios of the “wrong AUCs” to the “correct AUCs” with respect to different significances. To compare with FIG. 1(c), we show the $p_T^e$ vs. $\eta^e$ distributions for $\Gamma_{NP} \approx 100, 1, 0.1$ GeV in FIG. 18. The training results are shown in FIG. 19. We can see that applying models of the wrong widths still yields some distinguishability, yet they are much worse than applying the correct models.
This indicates the importance of getting the right order of magnitude for $\Gamma_{NP}$ before setting up the trainings, and shows that even an incorrectly trained NN still has an AUC within $\sim 20 - 30\%$ of the correctly trained model.

FIG. 18: $p_T$ vs. $\eta^e$ distributions for $\Gamma_{NP} \approx$ (a) 100, (b) 1, and (c) 0.1 GeV.
FIG. 19: Ratios of the LO AUCs from the tests with (a) $\Gamma_{NP} \approx 100$ GeV models applied on $\Gamma_{NP} \approx 0.1$ GeV samples, (b) $\Gamma_{NP} \approx 10$ GeV models applied on $\Gamma_{NP} \approx 1$ GeV samples, (c) $\Gamma_{NP} \approx 1$ GeV models applied on $\Gamma_{NP} \approx 10$ GeV samples, and (d) $\Gamma_{NP} \approx 0.1$ GeV models applied on $\Gamma_{NP} \approx 100$ GeV samples, to the AUCs using the correct models in the low-significance scenarios.

The second is to use the models trained upon LO samples of $S/\sqrt{B} = 3, 5, 8$ to test the samples of $S/\sqrt{B} = 1, 2, \cdots, 10$ for a fixed $\Gamma_{NP} \approx 10$ GeV. We also calculate the ratios of the “wrong AUCs” to the “correct AUCs” for different significances and show them in FIG. 20. We observe 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 5%.

These two comparisons indicate that when applying our analysis to the parameter space
of the signal hypotheses, even a coarse set of CNNs 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 CNNs, allowing a reduction of computing resources with a trade off of a small drop in performance.

![Graph](image)

**FIG. 20:** Ratios of the LO AUCs from the tests with (a) $S/\sqrt{B} = 3$, (b) $S/\sqrt{B} = 5$, and (c) $S/\sqrt{B} = 3$ models applied on samples of different $S/\sqrt{B}$ ratios to the correct AUCs, using samples of $\Gamma_{NP} \approx 10$ GeV.

---

[1] G. Arnison et al. (UA1), *Phys. Lett. B* 122, 103 (1983).
[2] M. Banner et al. (UA2), *Phys. Lett. B* 122, 476 (1983).
[3] M. Aaboud et al. (ATLAS), *Phys. Rev. D* 96, 052004 (2017), arXiv:1703.09127 [hep-ex].
[4] V. Khachatryan et al. (CMS), Tech. Rep. CMS-PAS-B2G-17-010 (CERN, Geneva, 2017).
[5] V. Khachatryan et al. (CMS), Phys. Lett. B 770, 278 (2017), arXiv:1612.09274 [hep-ex].
[6] M. Tanabashi et al. (Particle Data Group), Phys. Rev. D 98, 030001 (2018).
[7] P. Langacker, R. W. Robinett, and J. L. Rosner, Phys. Rev. D 30, 1470 (1984).
[8] F. Abe et al. (CDF), Phys. Rev. Lett. 74, 2626 (1995), arXiv:hep-ex/9503002.
[9] T. G. Rizzo, JHEP 05, 037 (2007), arXiv:0704.0235 [hep-ph].
[10] L.-T. Wang and I. Yavin, Int. J. Mod. Phys. A 23, 4647 (2008), arXiv:0802.2726 [hep-ph].
[11] S. Gopalakrishna, T. Han, I. Lewis, Z.-g. Si, and Y.-F. Zhou, Phys. Rev. D 82, 115020 (2010), arXiv:1008.3508 [hep-ph].
[12] O. Eboli, C. S. Fong, J. Gonzalez-Fraile, and M. Gonzalez-Garcia, Phys. Rev. D 83, 095014 (2011), arXiv:1102.3429 [hep-ph].
[13] C.-W. Chiang, N. D. Christensen, G.-J. Ding, and T. Han, Phys. Rev. D 85, 015023 (2012), arXiv:1107.5830 [hep-ph].
[14] S. Caron, J. S. Kim, K. Rolbiecki, R. Ruiz de Austri, and B. Stienen, Eur. Phys. J. C 77, 257 (2017), arXiv:1605.02797 [hep-ph].
[15] G. Kasieczka, T. Plehn, M. Russell, and T. Schell, JHEP 05, 006 (2017), arXiv:1701.08784 [hep-ph].
[16] J. Brehmer, K. Cranmer, G. Louppe, and J. Pavez, Phys. Rev. Lett. 121, 111801 (2018), arXiv:1805.00013 [hep-ph].
[17] J. Brehmer, K. Cranmer, G. Louppe, and J. Pavez, Phys. Rev. D 98, 052004 (2018), arXiv:1805.00020 [hep-ph].
[18] U. Simola, B. Pelssers, D. Barge, J. Conrad, and J. Corander, JINST 14, P03004 (2019), arXiv:1810.09930 [astro-ph.IM].
[19] G. Kasieczka, N. Kiefer, T. Plehn, and J. M. Thompson, SciPost Phys. 6, 069 (2019), arXiv:1812.09223 [hep-ph].
[20] F. F. Freitas, C. K. Khosa, and V. Sanz, Phys. Rev. D 100, 035040 (2019), arXiv:1902.05803 [hep-ph].
[21] A. Butter et al., SciPost Phys. 7, 014 (2019), arXiv:1902.09914 [hep-ph].
[22] C. K. Khosa, V. Sanz, and M. Soughton, (2019), arXiv:1910.06058 [hep-ph].
[23] Y.-C. J. Chen, C.-W. Chiang, G. Cottin, and D. Shih, Phys. Rev. D 101, 053001 (2020), arXiv:1908.08256 [hep-ph].
[24] K. Simonyan and A. Zisserman, arXiv:1409.1556 [cs.CV].

[25] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, in CVPR09 (2009).

[26] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro, JHEP 07, 079 (2014), arXiv:1405.0301 [hep-ph].

[27] 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), arXiv:1410.3012 [hep-ph].

[28] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lematre, A. Mertens, and M. Selvaggi (DELPHES 3), JHEP 02, 057 (2014), arXiv:1307.6346 [hep-ex].

[29] M. Cacciari, G. P. Salam, and G. Soyez, Eur. Phys. J. C 72, 1896 (2012), arXiv:1111.6097 [hep-ph].

[30] M. Cacciari, G. P. Salam, and G. Soyez, JHEP 04, 063 (2008), arXiv:0802.1189 [hep-ph].

[31] R. D. Ball et al., Nucl. Phys. B 867, 244 (2013), arXiv:1207.1303 [hep-ph].

[32] A. Alloul, N. D. Christensen, C. Degrande, C. Duhr, and B. Fuks, Comput. Phys. Commun. 185, 2250 (2014), arXiv:1310.1921 [hep-ph].

[33] F. Chollet et al., “Keras,” https://keras.io (2015).

[34] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, “TensorFlow: Large-scale machine learning on heterogeneous systems,” (2015), software available from tensorflow.org.