Probing Higgs boson exotic decays at the LHC with machine learning

Sunghoon Jung,[1,*] Zhen Liu,[2,†] Lian-Tao Wang,[3,4,‡] and Ke-Pan Xie[1,5,§]

[1]*Center for Theoretical Physics, Department of Physics and Astronomy, Seoul National University, Seoul 08826, Korea*

[2]*School of Physics and Astronomy, University of Minnesota, Minneapolis, Minnesota 55455, USA*

[3]*Enrico Fermi Institute, University of Chicago, Chicago, Illinois 60637, USA*

[4]*Department of Physics, University of Chicago, Chicago, Illinois 60637, USA*

[5]*Department of Physics and Astronomy, University of Nebraska, Lincoln, Nebraska 68588, USA*

(Received 19 September 2021; accepted 21 January 2022; published 10 February 2022)

We study the tagging of Higgs exotic decay signals using different types of deep neural networks (DNNs) while focusing on the $W^{\pm}h$ associated production channel followed by Higgs decaying into n b quarks with $n = 4$, 6, and 8. All the Higgs decay products are collected into a fat jet, to which we apply further selection using the DNNs. Three kinds of DNNs are considered—namely, the convolutional neural network, the recursive neural network, and the particle flow network (PFN). The PFN can achieve the best performance because its structure allows one to enfold more information in addition to the four-momenta of the jet constituents, such as the particle identifier, and tracks the parameters. Using the PFN as an example, we verify that it can serve as an efficient tagger even though it is trained on a different event topology with different b multiplicity from the actual signal. The projected sensitivity to the branching ratio of Higgs decaying into n b quarks at the HL-LHC are 10%, 3%, and 1%, for $n = 4$, 6, and 8, respectively.

DOI: 10.1103/PhysRevD.105.035008

I. INTRODUCTION

Higgs exotic decay is a promising window for probing new physics. The Higgs portal can provide the most relevant coupling between the Higgs boson and new physics, while its narrow Standard Model width enhances the sensitivity to exotic decay modes. Information on the exotic decay extracted from measurements of Higgs properties [1–3], while useful, is less sensitive. At the same time, exclusive searches targeting specific channels are less efficient in casting a wide net. It would be useful to develop strategies that can bridge these two extremes.

As a benchmark scenario, we consider a dark sector consisting of multiple dark scalars [4–15]. Owing to their mixing with the Higgs boson, the final decay products would be heavy fermions, such as b jets. At the same time, there could be cascade decays among the dark scalars, resulting in a variety of final states with different b multiplicities. As such, this furnishes a good example in which a more universal "tagger" (rather than focusing on a particular final state) could be beneficial.

It is challenging to develop such a strategy. Searches based on simple cuts run the risk of being too tailored to a specific feature or being too inclusive (and hence less sensitive). It is a place where some of the machine learning techniques can shine. In this paper, we test several deep neural networks (DNNs) with the benchmark signal. These include the convolutional neural network (CNN) [16–22], the recursive neural network (RecNN) [23–25], and the particle flow network (PFN) [26]. See Refs. [27,28] and the references therein for reviews of the DNN applications in LHC physics. In addition to comparing their performance on an individual channel, we also consider the universal applicability by testing them in the channels that they are not trained on.

The rest of this paper is organized as follows. Section II describes the details of the benchmarks as well as the preparation of machine learning data. Various DNNs are built and trained to distinguish the signals from backgrounds, as described in Sec. III. Their performances are then compared in Sec. IV, where the current bound and projected reach of the exotic decay branching ratios are also given. We conclude in Sec. V.

II. THE BENCHMARK PROCESSES

We consider the following Higgs exotic decay scenarios:
(1) The $4b$ channel: $h \to a_0 a_0$, $a_0 \to b\bar{b}$.

*sunghoonj@snu.ac.kr
†zliuphys@umn.edu
‡liantaow@uchicago.edu
§kepan.xie@unl.edu

TABLE I. The cut flow table for the preselection cuts before applying deep neutral networks. For the signals, the cross sections are given under the assumption of 100% branching ratios, e.g., for the $\ell^\pm\nu 4b$ channel we assume $\mathrm{Br}(h \to 4b) = 100\%$. For the SM $W^\pm h$, we have used the branching ratio $\mathrm{Br}(h \to b\bar{b}) = 58\%$ [45].

	Higgs exotic decay			SM		
Cross section (fb)	$\ell^\pm\nu 4b$	$\ell^\pm\nu 6b$	$\ell^\pm\nu 8b$	$W^\pm + \mathrm{jets}$	$t\bar{t}$	$W^\pm h$
Boosted $\ell^\pm$	8.21	7.66	7.04	2.53×10^5	6.21×10^3	5.48
Fat jet	7.01	6.56	6.03	2.01×10^5	4.95×10^3	4.66
b veto	6.17	5.80	5.35	1.96×10^5	2.17×10^3	4.07
Mass window	3.34	3.19	2.99	5.66×10^3	400	2.08
Efficiency	1.37%	1.36%	1.34%	0.96%	0.25%	1.31%

(2) The $6b$ channel: $h \to a_0 a_0$, with one of the $a_0 \to b\bar{b}$ and the other $a_0 \to a_1 a_1$, followed by $a_1 \to b\bar{b}$.
(3) The $8b$ channel: $h \to a_0 a_0$, with both $a_0 \to a_1 a_1$, followed by $a_1 \to b\bar{b}$.

Here a_0 and a_1 are the new light neutral scalars, whose masses are chosen as $M_0 = 30$ GeV and $M_1 = 12$ GeV, respectively, as benchmarks. For the production mechanism, we consider the $W^\pm h$ process with the leptonic decay $W^\pm \to \ell^\pm\nu$. The main background is the Standard Model (SM) $W^\pm + \mathrm{jets}$ with $W^\pm \to \ell^\pm\nu$, and the $t\bar{t}$ with semi-lepton decay. We consider the case in which all of the Higgs decay products are clustered into a fat jet [29].

We write the model file of Higgs exotic decay with the FeynRules package [30]. The parton-level events of signal and backgrounds at the 14 TeV LHC are simulated using the MadGraph5_aMC@NLO [31] package. The events are matched to $+1$ jet final states and then interfaced to PYTHIA8 [32] for parton shower and hadronization, and to DELPHES3 [33] for fast detector simulation. We use the CMS detector configurations as the detector setup.[1] We modify the b-tagging efficiency (and mistag rates for c-jet, light-flavor jets) to 0.77 (and 1/6, 1/134) according to Ref. [34].

To suppress the background, we require the final state to have exactly one charged lepton with

$$p_T^\ell > 25\ \mathrm{GeV}, \qquad |\eta^\ell| < 2.5,$$
$$p_T^{\ell+\not{E}_T} > 200\ \mathrm{GeV}, \qquad M_T < 100\ \mathrm{GeV}, \tag{1}$$

where the transverse mass is defined as

$$M_T = \sqrt{2 p_T^\ell \not{E}_T (1 - \cos\Delta\phi)}, \tag{2}$$

with $\Delta\phi$ being the azimuthal angle difference between $\ell^\pm$ and $\not{E}_T$. We also demand at least one fat jet reconstructed by the anti-k_t algorithm with $\Delta R = 1.5$ and

$$200\ \mathrm{GeV} < p_T^J < 500\ \mathrm{GeV}, \qquad |\eta^J| < 2.5. \tag{3}$$

The fat jets are trimmed by $R_{\mathrm{cut}} = 0.3$ and $f_{\mathrm{cut}} = 0.05$ [35]. Next, the small-R jets are clustered using an anti-k_t algorithm with $R = 0.4$, and b-tagged ones within

$$p_T^b > 25\ \mathrm{GeV}, \qquad |\eta^b| < 2.5, \tag{4}$$

are vetoed to suppress the $t\bar{t}$ background. Finally, we require the mass of the leading fat jet to be in the Higgs mass window

$$100\ \mathrm{GeV} < m_J < 150\ \mathrm{GeV} \tag{5}$$

and treat this jet as the Higgs candidate. The cut flows for signals and backgrounds based on the leading-order (LO) cross sections are listed in Table I, where we also give the cross sections for the SM $W^\pm h \to \ell^\pm\nu b\bar{b}$.[2] In particular, the composition of the data after preselection cuts in our sample is $W^\pm + \mathrm{jets}$: $t\bar{t}$: $W^\pm h = 0.9337 : 0.0660 : 0.0003$. Beyond LO, the next-to-leading-order K factor for the high-p_T $W^\pm h$ production is $\sim$1.5 [29,36,37]. While for the main backgrounds, the K factors for the SM $W^\pm + \mathrm{jets}$ and $t\bar{t}$ backgrounds are $\sim$1.5 and $\sim$1.6, respectively, at next-to-next-to-leading order [38–44]. Therefore, the LO-based analysis is not affected too much by the higher-order corrections.

III. BUILDING DIFFERENT DNNs

Around 30,000 events for each channel are generated, with 80% and 20% of the sample devoted to the training and validation/test datasets, respectively. We build DNNs to classify a specific exotic decay signal $W^\pm h \to \ell^\pm\nu nb$ (with $n = 4$, 6, or 8) from the backgrounds. We focus on the $W^\pm + \mathrm{jets}$ and $t\bar{t}$ backgrounds for the study, and

[1] We change the isolation ΔR parameters for electron and muon to 0.2 and 0.3, respectively, for increasing the signal acceptance with high multiplicity final states.

[2] To improve the event generation efficiency, without loss of generality we require at least one final state (light) quark or gluon with $p_T > 100$ GeV and 50 GeV for the $W^\pm + \mathrm{jets}$ and $t\bar{t}$ backgrounds at the event generator level, respectively. These (light) quarks and gluons include both initial state radiation and final state radiation, as well as decay products from SM particles such as W and t.

correspondingly the DNNs have three output neurons, one for the signal and the other two for backgrounds. The input features are basically the momenta of the constituents of the Higgs-candidate jet, but the concrete representation depends on the specific DNN that we choose. We use three different types of DNNs—namely, CNN, RecNN, and PFN—that are suitable for our purposes. These DNN constructions are described one by one in the following.

A. CNN and jet images

We expand the constituents of the Higgs candidate in the η-ϕ plane to form a 35×35 pixel jet image with the granularity of 0.1×0.1. The intensity I_i of a pixel i is defined as the p_T sum of all constituents inside it. We further normalize the intensities under the L^2-norm scheme so that $\sum_i I_i^2 = 1$ [17]. To boost the learning efficiency, a series of preprocessing procedures are applied to the jet images: (i) shifting the image so that the pixel with the highest intensity is in the center, (ii) rotating the image so that the pixel with the second highest intensity is exactly below the center of the image, and (iii) reflecting the image so that the pixel with the third highest intensity always lies in the right side. An individual jet image is rather sparse in that only a few pixels are activated. To manifest the pattern of the images, in Fig. 1 we average 10,000 images for each channel. One can read from the figure that higher b multiplicity gives more complex images for the exotic decay signals.

A CNN is built using the KERAS [46] package (with TENSORFLOW [47] as the back end) to classify the jet images. The two-dimensional information of an input image is first transformed into a one-dimensional vector by a set of convolutional modules (denoted as CLs, whose content is a few Conv2D, MaxPooling, and DROPOUT layers) and then passed through the fully connected layers

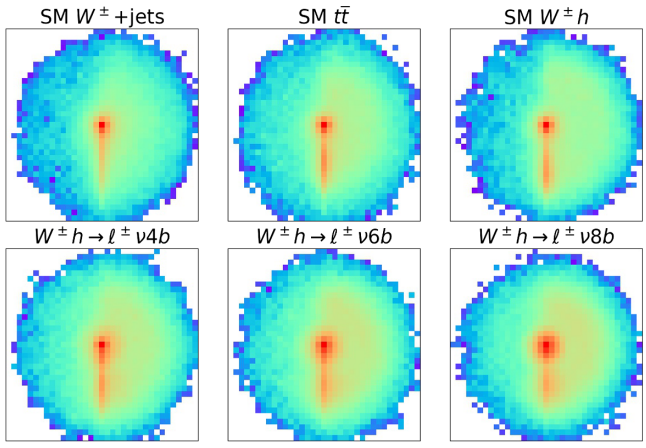

FIG. 1. The average of 10,000 jet images for CNN. From blue to red, the color represents the increasing intensities of the pixels.

(denoted as FLs, which consist of a few DENSE and DROPOUT layers) to the output layer. The activation functions are ReLU except for the output layer, where the SOFTMAX function is used so that the three neuron outputs $r_{0,1,2}$ satisfy

$$0 < r_{0,1,2} < 1, \qquad r_0 + r_1 + r_2 = 1, \tag{6}$$

and hence can be interpreted as the probabilities of the signal and the two backgrounds, respectively. For a well-trained CNN, the output of signal events should form a peak at around $r_0 = 1$.

We train one CNN for each signal channel. The ADAM optimization is adopted, and the batch size is set at 1000. The POOL_SIZE of MaxPooling is chosen as (2, 2), the step size of filtering is 1, and the parameter *padding* is *valid*, which means that the size of the image will reduce as the convolution applies. To choose the best configuration, we tune the hyperparameters as follows.

(1) For the CL module, we use two `Conv2D` layers, with the filter numbers 32 or 64 per layer. The filter sizes are chosen as [11, 11], [13, 13], [15, 15], [11, 7], or [15, 13].
(2) For the FL part, we try the structures of [256, 128], [256, 128, 64], or [512, 256, 128].
(3) Different initial learning rates (0.0001, 0.0005, 0.001, 0.005, or 0.01) and dropout rates (0.1, 0.3, or 0.5) are also tested.

The training is stopped early if the accuracy of the validation data does not increase for ten epochs. This early stopping can efficiently avoid overfitting.

B. RecNN and natural languages

The application of RecNN in classifying particle physics data is inspired by natural language processing. In this paradigm, the momenta are analogous to words, while clustering the constituents to reconstruct a jet is analogous to parsing a sentence [23]. The jet clustering history is a full binary tree whose root is the reconstructed jet, while the leaves are the jet constituents.

The jet binary tree can be embedded in a vector as follows [23]: First, for each node k we attach a q-dimensional state vector $\mathbf{u}_k$ by

$$\mathbf{u}_k = \sigma(W_u \mathbf{v}_k + b_u), \tag{7}$$

where $\mathbf{v}_k$ is the node's kinetic information (which we choose as the seven observables $p, \eta, \theta\phi, E, p_T$, and E/E_J), W_u is a $q \times 7$ weight matrix, b_u is a $q \times 1$ bias vector, and σ is the ReLU activation function. Second, for each node we further define a q-dimensional embedding vector $\mathbf{h}_k$ recursively from the leaves to the root of the binary tree

$$\mathbf{h}_k = \begin{cases} \mathbf{u}_k, & \text{if } k \text{ is a leaf}, \\ \sigma\left(W_h \begin{bmatrix} \mathbf{h}_{k_L} \\ \mathbf{h}_{k_R} \\ \mathbf{u}_k \end{bmatrix} + b_h\right) & \text{otherwise}, \end{cases} \tag{8}$$

where W_h and b_h are $q \times 3q$ weight matrix and $q \times 1$ bias vector, respectively, while h_{k_L} and h_{k_R}, respectively, denote the embedded vector of the left and right children of node k (where we require the left child is more energetic by data preprocessing). Finally, we denote the embedding of the root as $\mathbf{h}_{\text{jet}}$ and interface it to a fully connected module with two layers and q hidden neurons per layer to classify the signal and backgrounds. The activation of the output layer is again SOFTMAX for the purpose of a probability interpretation. Through this procedure, we have built up a RecNN with event-dependent tree structures; however, they share the same weight matrices W_u, W_h and bias vectors b_u, b_h, which will be optimized automatically when the RecNN are trained. On the other hand, the embedding dimension q is the hyperparameter that needs to be chosen before training.

As suggested by Ref. [23], we apply the k_t algorithm to the constituents of the Higgs candidate for a second step reconstruction and get the clustering history. The illustrations of signals and backgrounds are shown in Fig. 2. The RecNN is built and implemented using the PYTHON codes provided in Ref. [23]. To obtain the best configuration, we vary the embedding dimension q in 40, 80, the initial learning rate in 0.001, 0.01, and batch size in 100, 1000 and compare the learning accuracies. As the training of RecNN is rather time consuming, we stop the training when the accuracy of the validation data does not increase for three epochs.

C. PFN and tracks information

Denoting the observables of a jet constituent as a d-dimensional vector $p = \{\xi_1, \xi_2, \ldots, \xi_d\}$, then most high-level jet observables (such as jet mass, multiplicity, track mass, momentum dispersion, etc.) can be written in the form

$$\mathcal{O} = F\left(\sum_i \Phi(p_i)\right), \tag{9}$$

where $\Phi(p)$ and $F(x)$, respectively, are $\mathbb{R}^d \to \mathbb{R}^\ell$ and $\mathbb{R}^\ell \to \mathbb{R}$ functions determined by the observable $\mathcal{O}$, while the summation index i runs over all constituents of the jet. For example, for the jet mass m_J, we have $\Phi(p) \equiv p^\mu \equiv (E, p_x, p_y, p_z)$ and $F(x) = x^\mu x_\mu$. More examples can be found in Ref. [26]. Inspired by this, Komiske *et al.* [26] proposed the PFN, which is based on Eq. (9) but treated the $\Phi(p)$ and $F(x)$ as unknown functions to be constructed through the machine learning training process. More specifically, $\Phi(p)$ $[F(x)]$ is represented by a set of fully connected layers with N_Φ (N_F) hidden layers and neuron number n_Φ (n_F) per layer. We generally use the RELU function as the activation, except for the output layer of $F(x)$, where we use SOFTMAX.

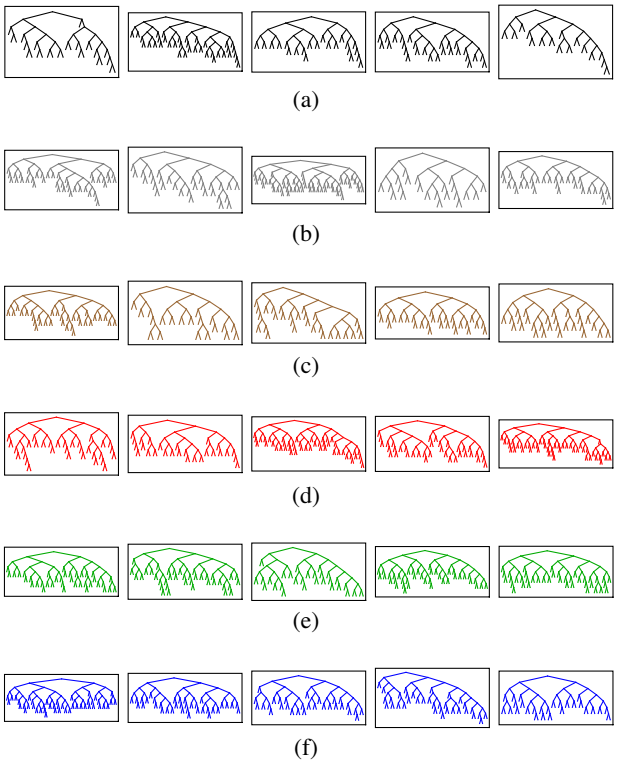

FIG. 2. Illustrations of the jet clustering histories from signals and backgrounds for RecNN. (a) The clustering history of fat jets from SM $W^\pm$ + jets. (b) The clustering history of fat jets from SM $t\bar{t}$. (c) The clustering history of fat jets from SM $W^\pm h$. (d) The clustering history of fat jets from $\ell^\pm \nu 4b$. (e) The clustering history of fat jets from $\ell^\pm \nu 6b$. (f) The clustering history of fat jets from $\ell^\pm \nu 8b$.

The "primary" setup for a PFN is to use the momenta of the jet constituents as input observables, i.e., (p_T, η, ϕ), and hence $d = 3$ for a single particle. However, the advantage of PFN is that it can enfold extra information besides the four-momentum. For example, a jet constituent's particle identifier (PID) can also be added as a component of p, and then $d = 4$. Technically, we use the so-called float PID function mapping the PID onto a real number between 0.1 and 1.1 [26].[3] Such a PFN is denoted as PFN_PID. Since the final state of the signal contains multiple b jets, the track information could help distinguish them from the backgrounds. Therefore, we build an extended PFN (denoted as "PFN_tracks") with the track impact parameters d_0 and z_0

[3]We use different numbers only to label charged particles such as electrons, muons, protons, pions, and kaons. The identification of charged particles can be done well, especially for electrons, muons, and charged pions. We did not model the difference between neutral hadrons such as K_L^0 and neutrons.

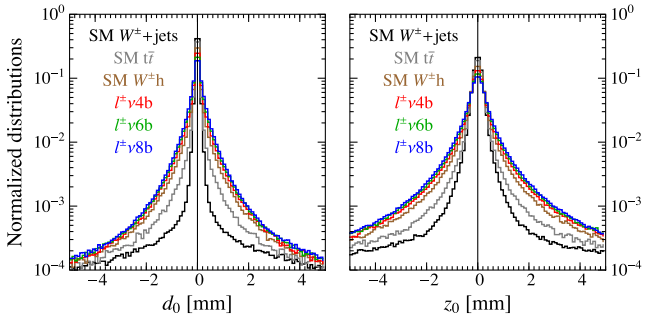

FIG. 3. The impact parameter distributions d_0 and z_0 for the signals and backgrounds for PFN tracks.

as two additional p components, and hence $d = 6$. As shown in Fig. 3, the impact parameter distributions have a longer tail for the events with higher b multiplicities. The smearing effect is important for the impact parameter reconstruction, and we use the smearing procedure suggested by Wildauer [48].

The PFN classifiers are built using the PyTorch package [49]. For simplicity, we fix

$$N_\Phi = 2, \qquad N_F = 3, \qquad n_\Phi = n_F = 100 \quad (10)$$

as a benchmark and vary the single-particle embedding dimension ℓ in 32, 64, 128, and the batch size in 100, 200 to find the best configurations. During training, if there are three continuous epochs in which the accuracy of the validation data does not increase, we stop the training to prevent overfitting.

IV. DNN RESULTS

A. Performance of different DNNs

After training, we apply the DNNs to the test data and get the neuron output distributions. Owing to the SOFTMAX function of the output layer, the zeroth output neuron r_0 peaks at around 1 for the signal process and peaks at around 0 for the two backgrounds. Adding a cut of $r_0 > r_c$ helps to enhance the signal significance. By varying the cut threshold r_c from 0 to 1, we can get the signal efficiency versus background-rejection curves. While the signal efficiency is defined as

$$\epsilon_S \equiv \frac{N_S^{(r_0>r_c)}}{N_S^{(\text{total})}}, \quad (11)$$

the background efficiency is defined as the weighted sum of the two backgrounds,

$$\epsilon_B \equiv \frac{\sigma_{B1} N_{B1}^{(r_0>r_c)} + \sigma_{B2} N_{B2}^{(r_0>r_c)}}{\sigma_{B1} N_{B1}^{(\text{total})} + \sigma_{B2} N_{B2}^{(\text{total})}}, \quad (12)$$

where $B1$ and $B2$ represent the $W^{\pm}$ + jets and $t\bar{t}$ backgrounds, respectively. We have checked to ensure that adding SM $W^{\pm}h$ as an extra neuron output to the DNNs results in almost no change to ϵ_B.

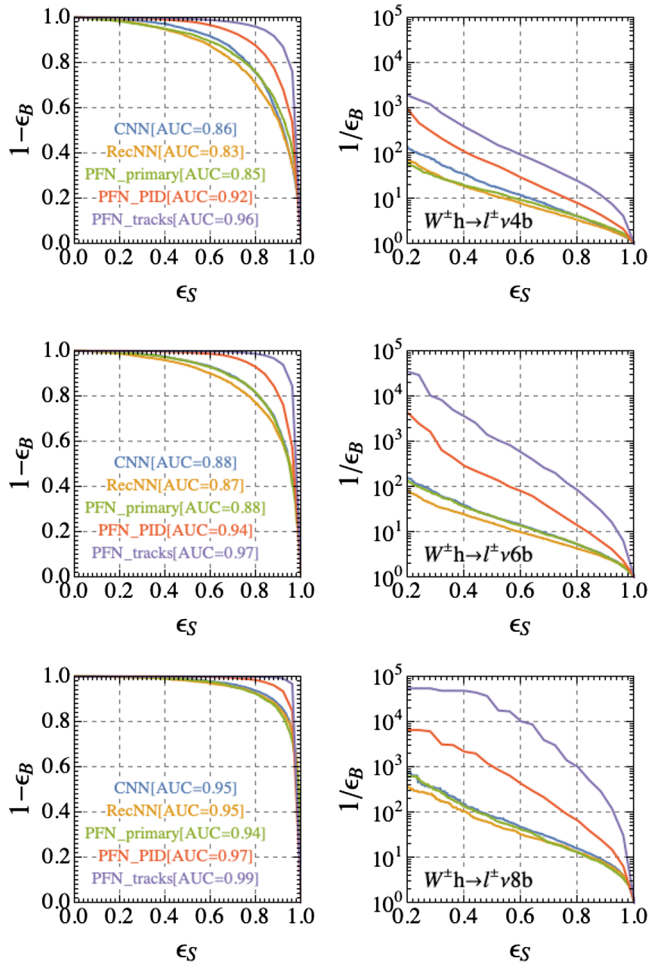

FIG. 4. The signal efficiencies versus background rejections for the three signal benchmarks.

The left panel of Fig. 4 shows the $\epsilon_S - (1 - \epsilon_B)$ curves, which are known as the receiver operating characteristic curves. The area under curves (AUCs) is also shown in the figure. Another representation of the curves, i.e., the signal efficiency ϵ_S versus background rejection $1/\epsilon_B$, is shown in the right panel of Fig. 4. For the $\ell^{\pm}\nu 6b$ and $\ell^{\pm}8b$ channels in the PFN_tracks case, when the cut on r_0 is very close to 1, we might reach a background-free region caused by limited statistics. In this case, we take a conservative estimate by replacing the background-free region with a fixed sample event number $N_B^{\min} = 0.1$.

From Fig. 4, we can conclude that CNN, RecNN, and PFN_primary (i.e., with only four-momentum information for the jet constituents) can all serve as effective signal-selection and background-rejection taggers. Moreover, they have quite similar performances. For example, in the $4b$ channel, for a signal efficiency $\epsilon_S \sim 0.6$, we can have a rejection ~ 10. Since all DNNs with only constituent four-momentum inputs have similar performances, it is

indicative of the fact that all the four-momentum information can be effectively and faithfully extracted by each of them. For all the DNNs, the performance is better for higher b multiplicities. For example, for the $\ell^{\pm}\nu 8b$ sample, we can have a rejection of 40/50 for a signal efficiency $\epsilon_S \sim 0.6$. This is expected since the signal with higher b multiplicities differs more from the major background from (sub)jet multiplicities and distributions.

Adding more information about the signal indeed leads to better performances. As discussed in Sec. III C, we can incorporate the particle ID and track information into the PFN. The results are also shown in Fig. 4. A gain of about a factor of 10 is achievable by adding the particle ID. As one would expect, the displacement from the b mesons imprints on the track impact parameters and can help improve the PFN performance. The improvement with additional track information depends on the multiplicity of b jets, with much better rejection for higher multiplicities (similar to the observation we made for other DNNs).

In reality, Higgs exotic decay signals would be a combination of different processes and b multiplicities. Without our knowing the composition of the signal, this represents a more inclusive tagger for a class of final states. Motivated by this, we should aim for a tagger which can be applied efficiently even without being trained on precisely the "right" signal sample. To this end, we further test the universality of the DNNs by applying a DNN trained on a specific signal channel to a mixed $4b + 6b + 8b$ sample with an equal amount of event numbers for each channel. To be concrete, we use PFN_track for this test. The result of the signal efficiency on the mixed sample is shown in Fig. 5. The DNNs trained on an exclusive sample still have a good performance on the mixed sample. For example, with $\epsilon_S \sim 0.6$, they have a rejection $1/\epsilon_B \sim 500$. The performance is better than that on the exclusive $\ell^{\pm}\nu 4b$ sample while worse than that on the $\ell^{\pm}\nu 8b$ sample. This result is expected due to the general improvement of the performance with the b multiplicity observed earlier. We also observe here the improvement associated with b multiplicity for the DNNs trained on the "wrong" sample.

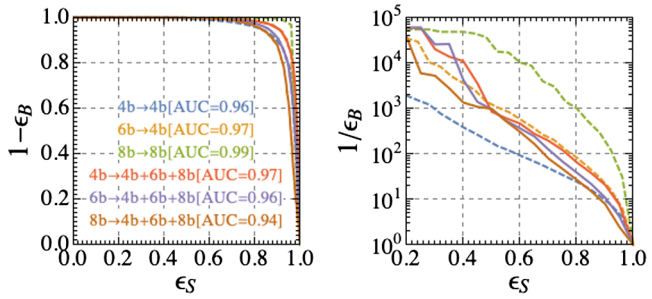

FIG. 5. The universality of the PFN_tracks. The notation "$A \to B$" in the figure means "trained on process A but tested on process B."

TABLE II. Testing the universality of the PFN with track information.

Classification accuracies		SM	$M_0 = 30$ GeV $M_1 = 12$ GeV		
		Trained on			
Tested on		$h \to b\bar{b}$	$\ell^{\pm}\nu 4b$	$\ell^{\pm}\nu 6b$	$\ell^{\pm}\nu 8b$
SM	$h \to b\bar{b}$	67.1%	61.4%	58.1%	56.5%
$M_0 = 30$ GeV	$\ell^{\pm}\nu 4b$	69.3%	73.1%	69.7%	68.1%
$M_1 = 12$ GeV	$\ell^{\pm}\nu 6b$	72.3%	77.0%	76.5%	74.9%
	$\ell^{\pm}\nu 8b$	74.4%	79.4%	79.9%	79.4%
	$4b + 6b + 8b$	⋯	76.4%	74.7%	73.6%

Moreover, DNNs trained with different exclusive samples have similar performances on the mixed sample.

In Table II, we show the classification accuracies of DNNs trained on an exclusive sample and applied to different samples, both exclusive and mixed.[4] It is interesting to note that the DNNs trained on lower b-multiplicity samples perform better when they are applied to higher multiplicity samples. For example, the DNN trained on $\ell^{\pm}\nu 4b$ sample (with an accuracy of 73.1%) has an accuracy of 79.4% on the $8b$ sample. Again, this observation implies the DNN trained in $4b$ samples relies on the b (sub)jet information. Note that the DNNs also tag the SM $h \to b\bar{b}$ events with 55%/61% efficiency, implying that the b jets and Higgs masses played important roles in the signal and background separation. When conducting a search for the exotic decays, one can apply other well-trained and optimized taggers for the $h \to b\bar{b}$ process, and hence we do not take this SM Higgs process as background when deriving the limits in the next subsection.

B. Branching ratio upper limits for the exotic decay

As an application of the techniques studied here, we derive a projection for the sensitivity to the Higgs exotic decays at the LHC and other future hadron colliders.

The projections are obtained as follows. Given a cut threshold r_c, we collect the event numbers of the signal and background samples that pass the cut, i.e., $N_S^{(r_0>r_c)}$ and $N_{B1}^{(r_0>r_c)} + N_{B2}^{(r_0>r_c)}$. They can be interpreted as the cross sections σ_S and σ_B after the cut. Therefore, given an integrated luminosity $\mathcal{L}$ at the LHC, the corresponding expected signal and background event numbers are[5]

[4]The classification accuracy is defined as the ratio of "correct predictions" to the length of the test dataset, where the "prediction" for a given event is defined as the neuron with maximal output. For instance, we count this event's classification as a signal if the neurons have outputs with $r_0 = 0.5$, $r_1 = 0.1$, and $r_2 = 0.3$. By this definition, the accuracy of a random prediction is 33.3% for a three-neuron output DNN.

[5]To avoid confusion, we always use $N_{S,B}$ to denote the number of events in the training and validation/test samples, while we use S, B to denote the events normalized to a given integrated luminosity.

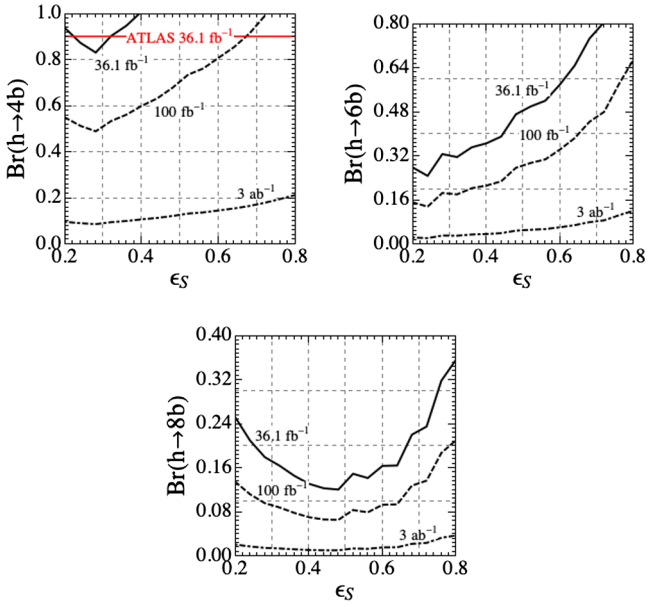

FIG. 6. Branching ratio upper limits for the exotic decays at different luminosities, derived by the PFN_tracks trained on the corresponding channel. For the $\ell^{\pm}\nu 4b$ channel, the ATLAS result [50] is also plotted as a reference.

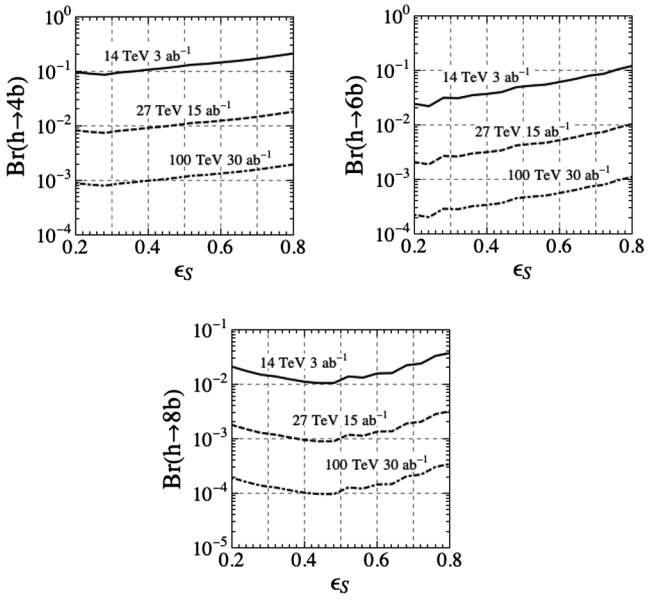

FIG. 7. Branching ratio upper limits for the exotic decays at different colliders driven by a naive rescaling of the 14 TeV results in PFN_tracks.

$$S = \sigma_S \times \mathrm{Br} \times \mathcal{L}, \qquad B = \sigma_B \times \mathcal{L}, \tag{13}$$

with Br being the branching ratio of the exotic decay channel. If no excess is obtained at the LHC, then the signal event number upper limit $S_{\max}$ is determined by

$$\sqrt{2\left[S_{\max} - B\ln\left(1+\frac{S_{\max}}{B}\right)\right]} = 2 \tag{14}$$

at 95% confidence level. Combining Eqs. (13) and (14) gives the branching ratio upper limit that can be achieved at a specific luminosity. The results for PFN_tracks at different luminosities are shown in Fig. 6. For the $\ell^{\pm}\nu 4b$ channel, the ATLAS Collaboration has a search on the nonboosted region at 36.1 fb^{-1} [50],[6] which is also shown in the figure; our reach is comparable to that of ATLAS. We can observe the maximal performance working point of signal efficiency around 0.2/0.3 for $4b$ and $6b$ samples, and around 0.45 for the $8b$ samples.

In Fig. 7, we have also shown the projection for a couple of proposed options of future hadron colliders, based on a simple rescaling by requiring the same signal events as in the HL-LHC. We can see that our semi-inclusive DNN taggers can help probe up to 10^{-4} of the Higgs exotic decays into multiple b jets, which is comparable to what one can achieve at future lepton collider Higgs factories or lepton-hadron colliders with a much lower background [51,52].

V. CONCLUSION

In this article we study the possibility of probing the Higgs exotic decays via deep learning methods. We focus on the $W^{\pm}h$ associated production with Higgs decaying to $4b$, $6b$, or $8b$ final states. Such decays can go through a cascade of some intermediate scalars the intermediate light scalars. As a benchmark, we consider two such scalars $a_{0,1}$, for which the decay $a_0 \to a_1 a_1$ is possible. The decay products of the Higgs are collected into a fat jet whose kinetic information is fed to the DNNs for machine learning.

Comparing the results from CNN, RecNN, and PFN_primary, we find that although different DNNs use different representations to enfold the kinetic information, they yield very similar performances as long as the input observables include only the four-momentum of the jet constituents. This implies that each DNN has efficiently extracted such information. In addition, by adding particle ID and track information as input observables, the PFN_PID and PFN_tracks can achieve better performances. In addition, for a given type of DNN, signals with higher b multiplicity have better performances. We also test the universality of the DNNs by applying a PFN_tracks model trained on a given signal channel (e.g., $\ell^{\pm}4b$) to another signal sample (e.g., $\ell^{\pm}6b$ or a mixed sample $4b + 6b + 8b$). Similar performances are found, implying that a well-trained tagger for a specific

[6]M. Aaboud *et al.* [50] studied both the $W^{\pm}h$ and Zh associated productions. For a proper comparison, here we extract only the single-lepton channel results.

channel can also efficiently probe other exotic decay channels.

The training results from PFN_tracks are used to derive the projected reach for the Higgs exotic decay branching ratios. We find that for the $4b$ channel, the reach is comparable to the existing ATLAS searches in the resolved (i.e., nonboosted) kinetic region. At the HL-LHC, $\mathrm{Br}(h \to 4b)$ can be probed up to $\sim 10\%$. For the $6b$ and $8b$ channels, the expected reach at the HL-LHC is $\sim 3\%$ and $\sim 1\%$, respectively. By a simple rescaling, we also obtain the probe limits at future colliders such as the 27 TeV HE-LHC and 100 TeV SppC/FCC-hh.

In order for our results to be as independent of the production channel as possible, we require the Higgs boson to have large boost, $p_T > 200$ GeV, so that the decay products of the Higgs boson are relatively far away from other objects in the event. Hence, we expect the performance to be similar in other production channels. As an example, we also try the Zh production with $Z \to \ell^+\ell^-$ and $\nu\bar{\nu}$ and obtain similar classification accuracies. Future work is needed to derive the precise performance in the other channels.

Our study shows that a semi-inclusive Higgs exotic decay DNN can be constructed with good performance. One can further explore such semi-inclusive BSM DNNs for other processes and decay modes. For instance, Higgs exotic decays into multiple scalars can also provide an admixture of b jets and light jets (or kaons [53]), and it would be useful to extend the search in this direction. Moreover, such a semi-inclusive approach particularly balances the broadness of the signal space and the use of kinematics to suppress the background, which will complement the traditional exclusive searches and possible more model-agnostic searches. This strategy can be followed in many different channels, such as $W/Z/h$ decay to multiple sterile neutrinos [54], a continuum dark sector [55–57], and generally hidden strong dynamics. It is also useful to extend the study of off-shell heavy Standard Model particle decays, as well as the search of beyond the Standard Model particle decays such as W', Z', top partners, and heavy Higgs.

ACKNOWLEDGMENTS

We thank Taoli Cheng, Yandong Liu, Gilles Louppe, Yongcheng Wu, Daneng Yang, and Rui Zhang for the useful discussions. We are especially grateful to Gilles Louppe for communication on RecNN and for sharing the datasets. K.-P. X. would like to thank Guang-Ze Fu for the great help on coding. The work of Z. L. is supported in part by the U.S. Department of Energy under Grant No. DE-SC0022345. S. J. and K.-P. X. are supported by Grant Korea No. NRF-2019R1C1C1010050. S. J. is also supported by a POSCO Science Fellowship, and K.-P. X. by the University of Nebraska–Lincoln. K.-P. X. would like to thank the University of Chicago, where part of this work was performed, for the hospitality. Z. L. and L.-T. W. acknowledge the Aspen Center of Physics for its hospitality during the final phase of this study, which is supported by National Science Foundation Grant No. PHY-1607611.

APPENDIX: PERFORMANCE OF THE PFN ON DIFFERENT MASS BENCHMARKS

To test the universality of our tagger for different mass choices, we set up another benchmark with $M_0 = 50$ GeV and $M_1 = 20$ GeV and investigate the performance of PFN_tracks. The results are shown in Table III. We can see that the DNN works well on this benchmark; moreover, the accuracies decrease only slightly when training on one mass benchmark and testing on another one. Therefore, a universal tagger with a weak dependence on the unknown masses of light scalars is plausible. One can explore this direction further.

TABLE III. The classification accuracies for two different mass benchmarks, extending Table II.

Classification accuracies		SM	$M_0 = 30$ GeV, $M_1 = 12$ GeV			$M_0 = 50$ GeV, $M_1 = 20$ GeV		
			Trained on					
Tested on		$h \to b\bar{b}$	$\ell^\pm\nu 4b$	$\ell^\pm\nu 6b$	$\ell^\pm\nu 8b$	$\ell^\pm\nu 4b$	$\ell^\pm\nu 6b$	$\ell^\pm\nu 8b$
SM	$h \to b\bar{b}$	67.1%	61.4%	58.1%	56.5%	56.3%	55.8%	52.3%
$M_0 = 30$ GeV	$\ell^\pm\nu 4b$	69.3%	73.1%	69.7%	68.1%	67.4%	66.6%	62.0%
$M_1 = 12$ GeV	$\ell^\pm\nu 6b$	72.3%	77.0%	76.5%	74.9%	72.6%	73.1%	70.8%
	$\ell^\pm\nu 8b$	74.4%	79.4%	79.9%	79.4%	76.1%	77.6%	77.0%
$M_0 = 50$ GeV	$\ell^\pm\nu 4b$	62.4%	69.7%	67.1%	67.0%	72.9%	73.9%	70.7%
$M_1 = 20$ GeV	$\ell^\pm\nu 6b$	64.6%	73.9%	73.1%	73.1%	76.8%	77.3%	76.6%
	$\ell^\pm\nu 8b$	66.5%	77.2%	76.6%	77.5%	79.4%	80.2%	80.1%

[1] A. M. Sirunyan *et al.* (CMS Collaboration), Eur. Phys. J. C **79**, 421 (2019).
[2] G. Aad *et al.* (ATLAS Collaboration), Phys. Rev. D **101**, 012002 (2020).
[3] M. Cepeda *et al.*, CERN Yellow Rep. Monogr. **7**, 221 (2019).
[4] M. J. Strassler and K. M. Zurek, Phys. Lett. B **651**, 374 (2007).
[5] T. Han, Z. Si, K. M. Zurek, and M. J. Strassler, J. High Energy Phys. 07 (2008) 008.
[6] P. Draper, T. Liu, C. E. M. Wagner, L.-T. Wang, and H. Zhang, Phys. Rev. Lett. **106**, 121805 (2011).
[7] D. Curtin *et al.*, Phys. Rev. D **90**, 075004 (2014).
[8] N. Craig, A. Katz, M. Strassler, and R. Sundrum, J. High Energy Phys. 07 (2015) 105.
[9] N. Arkani-Hamed, T. Cohen, R. T. D'Agnolo, A. Hook, H. D. Kim, and D. Pinner, Phys. Rev. Lett. **117**, 251801 (2016).
[10] P. Meade and H. Ramani, Phys. Rev. Lett. **122**, 041802 (2019).
[11] I. Baldes and G. Servant, J. High Energy Phys. 10 (2018) 053.
[12] A. Glioti, R. Rattazzi, and L. Vecchi, J. High Energy Phys. 04 (2019) 027.
[13] N. Arkani-Hamed, R. Tito D'Agnolo, and H. D. Kim, Phys. Rev. D **104**, 095014 (2021).
[14] M. Carena, C. Krause, Z. Liu, and Y. Wang, Phys. Rev. D **104**, 055016 (2021).
[15] S. Knapen, J. Shelton, and D. Xu, Phys. Rev. D **103**, 115013 (2021).
[16] J. Cogan, M. Kagan, E. Strauss, and A. Schwartzman, J. High Energy Phys. 02 (2015) 118.
[17] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman, and A. Schwartzman, J. High Energy Phys. 07 (2016) 069.
[18] P. Baldi, K. Bauer, C. Eng, P. Sadowski, and D. Whiteson, Phys. Rev. D **93**, 094034 (2016).
[19] J. Barnard, E. N. Dawe, M. J. Dolan, and N. Rajcic, Phys. Rev. D **95**, 014018 (2017).
[20] G. Kasieczka, T. Plehn, M. Russell, and T. Schell, J. High Energy Phys. 05 (2017) 006.
[21] J. Guo, J. Li, T. Li, F. Xu, and W. Zhang, Phys. Rev. D **98**, 076017 (2018).
[22] G. Li, Z. Li, Y. Wang, and Y. Wang, Phys. Rev. D **100**, 116013 (2019).
[23] G. Louppe, K. Cho, C. Becot, and K. Cranmer, J. High Energy Phys. 01 (2019) 057.
[24] T. Cheng, Comput. Software Big Sci. **2**, 3 (2018).
[25] K. Fraser and M. D. Schwartz, J. High Energy Phys. 10 (2018) 093.
[26] P. T. Komiske, E. M. Metodiev, and J. Thaler, J. High Energy Phys. 01 (2019) 121.
[27] D. Guest, K. Cranmer, and D. Whiteson, Annu. Rev. Nucl. Part. Sci. **68**, 161 (2018).
[28] M. Abdughani, J. Ren, L. Wu, J. M. Yang, and J. Zhao, Commun. Theor. Phys. **71**, 955 (2019).
[29] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, Phys. Rev. Lett. **100**, 242001 (2008).
[30] A. Alloul, N. D. Christensen, C. Degrande, C. Duhr, and B. Fuks, Comput. Phys. Commun. **185**, 2250 (2014).
[31] 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.
[32] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands, Comput. Phys. Commun. **191**, 159 (2015).
[33] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens, and M. Selvaggi (DELPHES 3 Collaboration), J. High Energy Phys. 02 (2014) 057.
[34] M. Aaboud *et al.* (ATLAS Collaboration), J. High Energy Phys. 12 (2018) 039.
[35] D. Krohn, J. Thaler, and L.-T. Wang, J. High Energy Phys. 02 (2010) 084.
[36] R. Ellis and S. Veseli, Phys. Rev. D **60**, 011501 (1999).
[37] J. M. Campbell, R. Ellis, and D. L. Rainwater, Phys. Rev. D **68**, 094021 (2003).
[38] J. Lindert *et al.*, Eur. Phys. J. C **77**, 829 (2017).
[39] M. Czakon and A. Mitov, Comput. Phys. Commun. **185**, 2930 (2014).
[40] M. Czakon, P. Fiedler, and A. Mitov, Phys. Rev. Lett. **110**, 252004 (2013).
[41] M. Czakon and A. Mitov, J. High Energy Phys. 01 (2013) 080.
[42] M. Czakon and A. Mitov, J. High Energy Phys. 12 (2012) 054.
[43] P. Bärnreuther, M. Czakon, and A. Mitov, Phys. Rev. Lett. **109**, 132001 (2012).
[44] M. Cacciari, M. Czakon, M. Mangano, A. Mitov, and P. Nason, Phys. Lett. B **710**, 612 (2012).
[45] M. Tanabashi *et al.* (Particle Data Group), Phys. Rev. D **98**, 030001 (2018).
[46] F. Chollet *et al.*, computer code KERAS, https://keras.io, 2015.
[47] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, https://www.tensorflow.org (2015).
[48] A. Wildauer, http://cds.cern.ch/record/2224523 (2006).
[49] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, in *Proceedings of Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS 2019), Vancouver, British Columbia, Canada, 2019*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Red Hook, NY, 2019), pp. 8024–8035, http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.
[50] M. Aaboud *et al.* (ATLAS Collaboration), J. High Energy Phys. 10 (2018) 031.
[51] Z. Liu, L.-T. Wang, and H. Zhang, Chin. Phys. C **41**, 063102 (2017).
[52] K. Cheung, O. Fischer, Z. S. Wang, and J. Zurita, J. High Energy Phys. 02 (2021) 161.
[53] X. C. Vidal, Y. Tsai, and J. Zurita, J. High Energy Phys. 01 (2020) 115.
[54] Z. Chacko, P. J. Fox, R. Harnik, and Z. Liu, J. High Energy Phys. 03 (2021) 112.
[55] R. Contino, A. Podo, and F. Revello, J. High Energy Phys. 02 (2021) 091.
[56] C. Csáki, S. Hong, G. Kurup, S. J. Lee, M. Perelstein, and W. Xue, arXiv:2105.07035.
[57] C. Csáki, S. Hong, G. Kurup, S. J. Lee, M. Perelstein, and W. Xue, arXiv:2105.14023.