Resolving combinatorial ambiguities in dilepton $t\bar{t}$ event topologies with neural networks

Haider Alhazmi[ORCID],[1,2,*] Zhongtian Dong[ORCID],[1,†] Li Huang[ORCID],[3,4,‡] Jeong Han Kim,[5,6,7,§]
Kyoungchul Kong[ORCID],[1,‖] and David Shih[ORCID][8,¶]

[1]*Department of Physics and Astronomy, University of Kansas, Lawrence, Kansas 66045, USA*
[2]*Department of Physics, Jazan University, Jazan 45142, Saudi Arabia*
[3]*International Centre for Theoretical Physics Asia-Pacific,*
University of Chinese Academy of Sciences, 100190 Beijing, China
[4]*Taiji Laboratory for Gravitational Wave Universe,*
University of Chinese Academy of Sciences, 100049 Beijing, China
[5]*Department of Physics, Chungbuk National University, Cheongju, Chungbuk 28644, Korea*
[6]*Center for Theoretical Physics of the Universe, Institute for Basic Science, Daejeon 34126, Korea*
[7]*School of Physics, KIAS, Seoul 02455, South Korea*
[8]*NHETC, Dept. of Physics and Astronomy, Rutgers,*
The State University of NJ Piscataway, New Jersey 08854, USA

We study the potential of deep learning to resolve the combinatorial problem in supersymmetrylike events with two invisible particles at the LHC. As a concrete example, we focus on dileptonic $t\bar{t}$ events, where the combinatorial problem becomes an issue of binary classification: pairing the correct lepton with each b quark coming from the decays of the tops. We investigate the performance of a number of machine learning algorithms, including attention-based networks, which have been used for a similar problem in the fully hadronic channel of $t\bar{t}$ production, and the Lorentz Boost Network, which is motivated by physics principles. We then consider the general case when the underlying mass spectrum is unknown, and hence no kinematic end point information is available. Compared against existing methods based on kinematic variables, we demonstrate that the efficiency for selecting the correct pairing is greatly improved by utilizing deep learning techniques.

I. INTRODUCTION

Signatures with missing transverse momentum ($\vec{\mathcal{P}}_T$) are one of the most exciting classes of events at the Large Hadron Collider and future colliders. They are produced by well-motivated scenarios of physics beyond the Standard Model (BSM), including supersymmetry and dark matter. Unfortunately, events with $\vec{\mathcal{P}}_T$ are difficult to interpret and analyze due to *instrumental effects*, *unknown nature of the invisible particles*, and *incomplete kinematic information* [1].

One common approach to analyzing $\vec{\mathcal{P}}_T$ events is to hypothesize a certain event topology and design suitable event variables adapted to this interpretation [2]. Already at this stage, one faces a combinatorial problem: how to associate the reconstructed objects in the event with the elementary particles in the final state of the event topology. The most common practice in resolving the combinatorial problem is to choose the "best" assignment event by event. In this case, one tries to design an algorithm (typically involving kinematic variables) which will single out one (or maybe several) among many possible assignments as the most likely "correct" assignment. Then, the value of the kinematic variable obtained with this specific choice is used for further analysis. In the presence of $\vec{\mathcal{P}}_T$, the combinatorics problem becomes more severe due to the unknown momenta of the invisible particles. To address the combinatorial problem properly, various strategies have been proposed, depending on the length of cascade decays. We refer to Ref. [1] and references therein for existing methods.

*haider@ku.edu
†cdong@ku.edu
‡huangli@ucas.ac.cn
§jeonghan.kim@cbu.ac.kr
‖kckong@ku.edu
¶shih@physics.rutgers.edu

In this paper, we study the combinatorial problem in events with $\vec{P}_T$ using supervised machine learning (ML). As a concrete example, we will consider dileptonic $t\bar{t}$ production. In this case, the combinatorial ambiguity is simply twofold: how to correctly pair the two b quarks with the two leptons in every event. In one sense, the twofold ambiguity of dileptonic $t\bar{t}$ is the simplest combinatorial problem; yet on the other hand, the presence of two missing neutrinos brings additional challenges in reconstructing the final state. In any event, the twofold ambiguity can be mapped to a straightforward binary classification task in supervised ML.

Supervised ML methods have recently been applied with much success to many areas of high energy physics (HEP), including jet and event classification [3]. ML methods—especially those based on deep neural networks—are able to learn subtle correlations in a high-dimensional space and so often outperform more conventional methods based on physics-motivated high-level features (i.e., specially designed kinematic variables). However, despite all of this recent activity, so far, the combinatorial problem has not received much attention. Only very recently have there been studies of particular ML methods—those utilizing the permutation-invariant structure of attention-based neural networks—for the combinatorial problem in fully hadronic $t\bar{t}$ production [4–6]. Also, other final states such as $t\bar{t}h$ [4], four tops [4,7], HZ [7], and a stop pair with R-Parity violation (RPV) [8] have been studied in the fully hadronic channel using ML methods.

There are already many existing, non-ML strategies developed for the dilepton channel, such as end point methods [1,2,9–11], the hemisphere method [12–15], topness [16], the kinematic likelihood fit [17], etc. Here, we will compare a number of new ML-based methods, including the attention-based methods explored in Refs. [4–6], and benchmark them against the existing methods. We will demonstrate that the new ML approaches lead to significantly improved ability in resolving the twofold ambiguity compared to previous methods. More accurate solutions to the combinatorial problem of dileptonic $t\bar{t}$ could have many potential applications, including the following:

(1) Resolving the twofold ambiguity in dilepton top-quark pair production is of high importance for precision measurements of top-quark and Higgs properties. For example, Refs. [18–20] aim to measure the Yukawa coupling and the CP phase in the top-Higgs interaction via $t\bar{t}h$ production in the dilepton channel, utilizing the kinematic methods proposed in Refs. [1,21].

(2) Another nontrivial example is double Higgs production in the $hh \to (b\bar{b})(WW^*) \to (b\bar{b})(\ell^+\ell^-\nu\bar{\nu})$ final state, where the dominant background is the dilepton $t\bar{t}$ production. Recent studies [22,23] adopt a traditional χ^2 method (topness and Higgs-ness) to

enhance the signal sensitivity of double Higgs production. The method attempts to solve the combinatorial problem by choosing the smallest χ^2 value of all possible combinations. We expect that hybrid methods with ML would resolve the twofold ambiguity better, eventually leading to the improved signal sensitivity.

(3) The dilepton channel resembles signatures arising in various BSM scenarios, where the two missing particles could be dark matter candidates. Therefore, our results will be valuable in reducing the $t\bar{t}$ backgrounds in the search for any new physics in the same final state. Also, we will generalize our method to arbitrary mass spectra, in order to address the twofold ambiguity in the new physics scenarios.

This paper is structured as follows. We begin our discussion by describing the event generation and setup of ML methods in Sec. II. Our investigation of the performance of various machine learning algorithms is contained in Sec. III. In Sec. III F, we summarize our findings and compare various ML methods against one another as well as against existing approaches (which are briefly described in the Appendix). In Sec. IV, we study the combinatorial problem without the prior knowledge of the mass information, considering the $t\bar{t}$-like event topology. Section V is reserved for discussion.

II. SETUP

A. Details of the simulation

The event topology considered in this article is depicted in Fig. 1, where A_i, B_i, and C_i ($i = 1, 2$ to denote two sides of decays) are top quarks (t), W bosons, and neutrinos (ν), respectively. (In Sec. IV, we will assume they are particles in new physics beyond the SM, whose masses are unknown.) We assume two-body decay at each step, $A_i \to b_i B_i$ and $B_i \to c_i C_i$, where b_i and c_i are two visible particles and C_i is the invisible particle. We further assume that c_1 and c_2 (two charged leptons) are distinguishable,
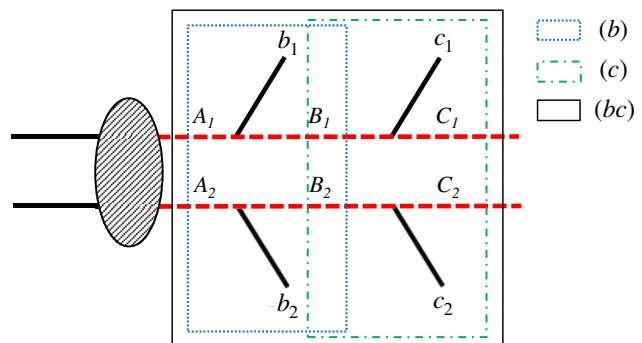

FIG. 1. The event topology considered in this paper, together with the three possible subsystems. The blue dotted, the green dot-dashed, and the black solid boxes indicate the subsystems (b), (c), and (bc), respectively. The figure is taken from Ref. [24].

while b_1 and b_2 (two b-tagged jets) are not. Therefore, there is a twofold ambiguity in pairing b_i and c_i. Figure 1 shows three possible subsystems in the blue dotted (b), the green dot-dashed (c), and the black solid boxes (bc).

For the numerical studies in this article, we generate a partonic $t\bar{t}$ dilepton sample, using the MadGraph5_aMC@NLO at the 14 TeV LHC with the default parton distribution functions [25]. All event samples are generated at leading-order accuracy in QCD. The masses of the top quark and the W boson are set to 173 GeV and 80.419, respectively. We take into account the proper finite widths, which often make the top quarks and the W bosons significantly off shell. To reduce the background, we apply the same basic cuts as those used in Refs. [1,11]. We did not include initial state radiation/final state radiation (ISR/FSR) in our study, but they should be relatively harmless for our problem, given the high b-tagging efficiency and small fake rates expected at the high luminosity (HL)-LHC [26]. This setup also allows us to make a fair comparison of our findings against existing results, as most studies in the literature did not consider ISR/FSR.

To simulate detector effects, we follow the parametrization used in the ATLAS detector performances report for the HL-LHC [27]. The energy resolution is parametrized by three terms, noise (N), stochastic (S), and constant (C) terms

$$\frac{\sigma}{E} = \sqrt{\left(\frac{N}{E}\right)^2 + \left(\frac{S}{\sqrt{E}}\right)^2 + C^2}, \qquad (2.1)$$

where in our analysis we use $N = 5.3$, $S = 0.74$, and $C = 0.05$ for jets; we use $N = 0.3$, $S = 0.1$, and $C = 0.01$ for electrons [28]; and the energy E is in GeV. The muon energy resolution is given by the Inner Detector (ID) and Muon Spectrometer (MS) resolution functions

$$\sigma = \frac{\sigma_{\mathrm{ID}}\sigma_{\mathrm{MS}}}{\sqrt{\sigma_{\mathrm{ID}}^2 + \sigma_{\mathrm{MS}}^2}}, \qquad (2.2)$$

where σ_{ID} and σ_{MS} are defined as

$$\sigma_{\mathrm{ID}} = E\sqrt{\alpha_1^2 + (\alpha_2 E)^2}, \qquad (2.3)$$

$$\sigma_{\mathrm{MS}} = E\sqrt{\left(\frac{\beta_0}{E}\right)^2 + \beta_1^2 + (\beta_2 E)^2}. \qquad (2.4)$$

We use $\alpha_1 = 0.023035$, $\alpha_2 = 0.000347$, $\beta_0 = 0.12$, $\beta_1 = 0.03278$, and $\beta_2 = 0.00014$ in our study. We prepare one million parton-level events and the corresponding one million smeared events for the $b\bar{b}\ell^+\ell^- + \vec{P}_T$ final state. We denote those smeared events by "detector-level events" in the rest of this study.

B. Setup of ML methods

From the one million events we prepared, we take random selection of 900,000 for the training/validation of all of the ML methods (with a 90/10 split), and the remaining 100,000 events for testing.

For all the ML methods in our study, unless otherwise noted, the dimension of the input data is 18, including 4-momentum of two b-tagged jets, two charged leptons, and the missing transverse momentum. We order two b-tagged jets by their p_T and label them as b_1 and b_2 and the corresponding correct lepton pairing as ℓ_1 and ℓ_2, respectively. Note that ℓ_1 and ℓ_2 are not necessarily ordered by their p_T. They are the lepton pairing corresponding to the two p_T ordered b jets, b_1 and b_2.

We prepare two datasets with labels 1 (correct b-ℓ pairing) and 0 (incorrect pairing), consisting of $(p_{b_1}, p_{\ell_1}, p_{b_2}, p_{\ell_2}, \vec{P}_T)$ and $(p_{b_1}, p_{\ell_2}, p_{b_2}, p_{\ell_1}, \vec{P}_T)$, respectively. In other words, the information about the pairing correctness is encoded in the ordering of the 4-vectors that make up each event. Note that each event is counted twice, once in the correct-pairing dataset and once in the incorrect-pairing dataset. We have checked that this reusing/double-counting of events does not affect the performance of the ML methods in any way but helps increase the size of the dataset.

The ML methods are then formulated as binary classifiers (with the usual binary cross-entropy loss and sigmoid activation) between the correct and incorrect pairing datasets. Unlike usual binary classification problems, however, here each event actually produces two scores, one for each ordering of the leptons, i.e., $f(p_{b_1}, p_\ell, p_{b_2}, p_{\ell'})$ and $f(p_{b_1}, p_{\ell'}, p_{b_2}, p_\ell)$, where f represents the sigmoid output of the ML method. Note that previously, we used ℓ_1 and ℓ_2 to denote a lepton, which should be paired with b_1 and b_2, respectively. However, in practice, this truth information is not available, and we will have to make an arbitrary pairing of a lepton and b-tagged jet. For this purpose, we used ℓ and ℓ' to denote two leptons. Therefore, one of the two scores, $f(p_{b_1}, p_\ell, p_{b_2}, p_{\ell'})$ or $f(p_{b_1}, p_{\ell'}, p_{b_2}, p_\ell)$, is close to 1 (correct pairing), and the other will be close to 0 (incorrect pairing). To the extent that the ML method is optimal, we expect that these two scores contain the exact same information; i.e., $f(p_{b_1}, p_\ell, p_{b_2}, p_{\ell'})$ and $1 - f(p_{b_1}, p_{\ell'}, p_{b_2}, p_\ell)$ are both equal to the probability that (b_1, ℓ) is the correct pairing [29]. Of course, in practice, owing to the finite training data and ML model capacity, these two scores will only be approximate estimates of the true probability. Therefore, going forward, we define an averaged score in order to incorporate information from both estimates:

$$P_{b_1,\ell} = \frac{1 + f(p_{b_1}, p_\ell, p_{b_2}, p_{\ell'}) - f(p_{b_1}, p_{\ell'}, p_{b_2}, p_\ell)}{2}. \qquad (2.5)$$

Note that, defined this way, P does have the correct behavior as a binary probability, $P_{b_1,\ell} + P_{b_1,\ell'} = 1$.

To properly compare different methods, we define a common set of metrics as follows. It is natural to take the prediction of the ML method to be the pairing (b_1, ℓ) or (b_1, ℓ') that returns the higher value of P; i.e., the ML predicts (b_1, ℓ) when $P_{b_1,\ell} > 0.5$. The fraction of correct predictions provides a baseline measure of the method's performance: the purity at 100% efficiency.

The purity can be improved at the expense of efficiency, by considering a subset of data which passes certain criteria or selection cuts. For the ML methods, we can accomplish this by requiring the method to be more confident in its prediction, i.e., only keeping events for which $P_{b_1,\ell} > P_c$ or $P_{b_1,\ell'} > P_c$. Given such a selection, we define the efficiency (ϵ) and the purity (P) to be [9,10,30]

$$\epsilon = \frac{N_{\text{cuts}}}{N_{\text{total}}},$$
$$= \frac{\text{the number of events which pass the selection}}{\text{the total number of events}} \quad (2.6)$$

$$P = \frac{N_{\text{correct}}}{N_{\text{cuts}}}$$
$$= \frac{\text{the number of the correctly identified events}}{\text{the number of events that passed the selection}}. \quad (2.7)$$

Note that in the machine learning literature the purity is often referred to as precision $= \frac{\text{TP}}{\text{TP}+\text{FP}}$, where TP is the number of true positives and FP is the number of false positives. In Sec. III, we will produce purity vs efficiency curves [analogous to receiver operating characteristic (ROC) curves] by varying this threshold P_c.

C. Existing methods

If the individual momentum of the missing neutrinos (or dark matter candidates) cannot be uniquely determined, the next best approach would be to consider some sort of approximation [31]. For instance, a matrix element method can be used to find the most likely values of the invisible momenta (to be discussed in Sec. A 5). However, the method itself suffers from the combinatorial problem and is very model dependent as it requires to fully specify the underlying physics in the consideration (masses, spins, couplings, etc.).

An alternative approach would be to rely only on kinematics such as masses and event topology (without spin or coupling information) and to obtain the invisible momenta by optimizing a suitable kinematic function. But what constitutes a good target function for such an optimization? Several algorithms are proposed depending on what kind of target function is considered: kinematic end points (Sec. A 1), hemisphere method/recursive jigsaw (Sec. A 2), topness (Sec. A 3), kinematic likelihood (KL)

fitter (Sec. A 4), matrix element methods (Sec. A 5), and analytic reconstruction (Sec. A 6). In general, methods which invoke fewer assumptions are more robust and model independent but lead to rather vague conclusions with poor results, while methods with more assumptions give better results, but the methods themselves are fragile as they are not typically applicable to more general cases. This is one of the main reasons why we want to develop as many methods as possible. We also understand kinematics better by comparing how each method works.

In some of these existing kinematic methods, there are natural ways to select a subset of events with higher purity, analogous to the cut $P_{b_1,\ell} > P_c$ or $P_{b_1,\ell'} > P_c$ described in the previous subsection for the ML methods. For example, in the case of end point methods, it is known that a cut on M_{eff} or H_T reduces the number of samples but can improve the accuracy of the pairing prediction [11].

III. MACHINE LEARNING APPROACHES

The data collected by high-energy physics experiments such as the LHC are very complex and very high dimensional. Collider physics analyses could be considered as a series of dimensional reduction processes at several stages. The last stage is likely to involve the reconstructed objects. Even then, the dimension of the input data is still quite large—proportional to the number of reconstructed particles. It is a difficult task to understand the full correlations of the high-dimensional data; this has motivated the consideration of suitable kinematic variables that capture the salient features of the initial data, and it also motivates the study of ML-based methods where this kind of feature engineering is automated directly from the initial data.

In the following subsections, we will explore various algorithms such as tree-based methods (Sec. III A), deep neural networks (Sec. III B), recurrent neural networks (Sec. III C), attention-based networks (Sec. III D), and the Lorentz boost network (Sec. III E). We discuss each method very briefly without going into details. We refer to Refs. [3,32–34] for more details of various machine learning algorithms and Refs. [3,34,35] for machine learning in high-energy physics. All codes used in this paper are publicly available in Ref. [36].

A. Tree-based methods

As a rudimentary ML baseline, we consider two common tree-based methods: Random Forest and boosted decision trees. These are shallow ML methods that are popular in HEP and can handle a modest number of inputs. For more about these methods, we refer the reader to Refs. [37,38].

First, we tried the tree-based methods with 4-momentum information only (18-dimensional input). In our analysis, the forest includes 2000 trees, with each leaf having at most 0.01% of the entire training data. The fraction of training

set used to train each tree does not have a significant impact in our result. The Random Forest algorithm is implemented using the SCIKIT-learn PYTHON library [38]. By taking the combination which returns the higher score as the correct combination, we find that the Random Forest classifies 83.4% of the data correctly for the parton-level events and 81.2% for the detector-level events (for 100% efficiency).

In general, boosted decision trees (BDT) outperform most other tree-based methods. In our analysis, the implementation of the algorithm is done using the XGBoost library in PYTHON [39]. We find that the BDT classifies 86.1% of the data correctly for the parton-level events and 82.4% for the detector-level events. This is comparable to or better than the existing methods discussed in Appendix. Good performance with the detector-level data is especially notable.

We repeated the same analysis, with boosted trees and a Random Forest, including the kinematic variables ($m_{b\ell}$, M_{2Ct}, M_{2CW}, $\Delta R_{b\ell}$, topness) defined in the Appendix and obtained a higher purity 90.4% and 90.1% for parton-level events and 84.1% and 84% for detector-level events,

respectively. The fact that including kinematic variables improved the performance of these algorithms indicates that the Random Forest and BDT did not catch important kinematic features during the training—possibly a sign that shallow ML methods are insufficiently expressive to fully automate feature learning.

Finally, we also tried training Random Forests and BDTs on kinematic variables only (without four momenta). The results do not change significantly; we obtained 89.5% and 83.4% purity using boosted trees and 89.7% and 83.7% purity using a Random Forest, respectively. Evidently, the 4-vectors are not adding much to the performance beyond the kinematic variables, for these shallow ML methods.

B. Deep neural networks

Next, we consider a fully connected deep neural network (DNN). The input layer is followed by three hidden layers of 512 neurons each, with rectified linear unit (ReLU) activation functions in between. The last hidden layer is
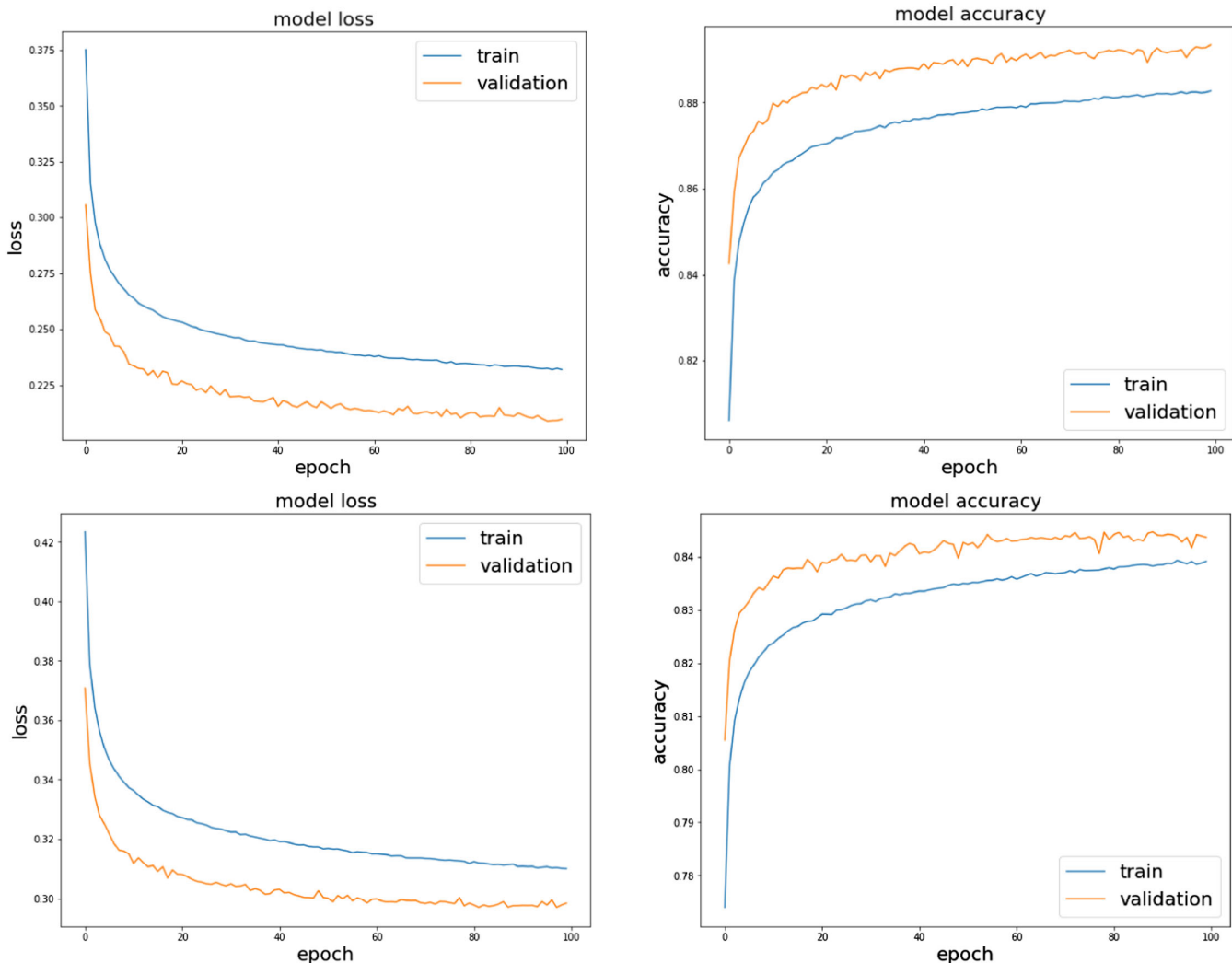

FIG. 2. DNN loss (left) and accuracy (right) for the parton-level (top) and the detector-level (bottom) events.
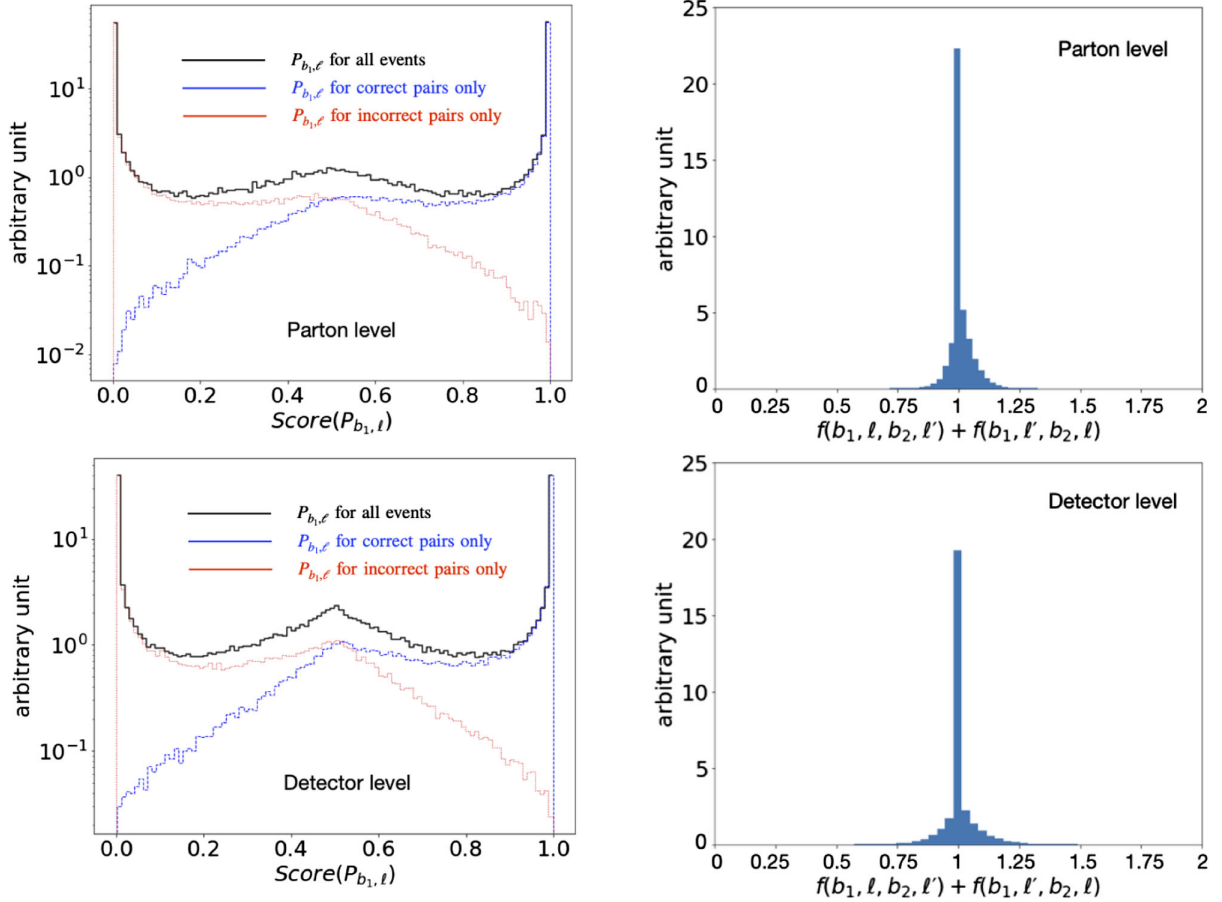
FIG. 3.   Distributions of the DNN scores (left) and the sum of two scores $[f(p_{b_1}, p_\ell, p_{b_2}, p_{\ell'}) + f(p_{b_1}, p_{\ell'}, p_{b_2}, p_\ell)]$ (right) for parton-level events (top) and detector-level events (bottom). The score $P_{b_1, \ell}$ for all events is shown in black, while $P_{b_1, \ell}$ for which b_1, ℓ is the correct (incorrect) pairing is in blue (red) and peaked close to 1 (0).

connected to a single output neuron with the sigmoid activation function, to match the target binary label of 0 (for the incorrect pair) or 1 (for the correct pair).

To prevent overfitting during training, we consider a 25% dropout after each hidden [DNN, Long Short-Term Memory (LSTM), and Lorentz Boost Network (LBN)] layer. We find this dropout value makes the training and validation accuracy converge relatively well as shown in Fig. 2. (As is generally the case, dropout is only used during training and not during validation/testing, which causes the training accuracy/loss to appear worse than the validation accuracy/loss over the training history.) Batch normalization is applied before each dropout. We used the Adam optimizer with a learning rate of 10^{-3} to minimize the binary cross-entropy loss function for all neural networks (NNs) that are considered in this paper. The whole network structure is implemented using the KERAS library [40], and the best model is selected based on the validation loss.

Figure 3 shows the score distributions [$P_{b_1, \ell}$ in black for all events, $P_{b_1, \ell}$ in blue for which (b_1, ℓ) is the correct pairing, and $P_{b_1, \ell}$ in red for which (b_1, ℓ) is the incorrect

pairing] in the left and the sum of two raw DNN outputs $[f(p_{b_1}, p_\ell, p_{b_2}, p_{\ell'}) + f(p_{b_1}, p_{\ell'}, p_{b_2}, p_\ell)]$ in the right panel for parton-level (top) and detector-level events (bottom). As expected, the sum of the two DNN outputs is approximately (but not exactly) equal to 1, and the $P_{b_1, \ell}$ score for which (b_1, ℓ) is the correct (incorrect) pairing peaks near 1 (0). We find the DNN predicts correctly 89.5% of the time at parton level and 84.5% at the detector level, which are are very similar to the accuracy (right panel) in Fig. 2.

As a cross-check, we have tried a different approach. Instead of taking both correct and incorrect combinations (i.e., instead of using the same event twice) and producing two scores $P_{b_1, \ell}$ and $P_{b_1, \ell'}$ for each event, we have tried to produce a single score by p_T ordering the leptons in event and labeling each event according to whether this p_T ordering produced the correct pairing or not. With the single score, we find a minor difference for the parton-level purity (88% for 100% efficiency), while the detector-level purity (84%) remains very similar to the case with the two scores ($P_{b_1, \ell}$ and $P_{b_1, \ell'}$). Since our first method gives a slightly better outcome, we use the two-score scheme for the remaining NNs.

We also find that the result remains approximately the same when we exclude $\vec{p}_T$ information in the input, keeping the 4-momentum of the visible particles only. This is to be expected since the $\vec{p}_T$ in our simulation is simply negative of the sum of the transverse momentum of all visible particles, implying that the $\vec{p}_T$ does not add new information. We also notice that, unlike the tree-based methods, including more inputs such as the kinematic variables or topness value does not change the result of neural networks significantly. This observation indicates that NN learns the (high-level) kinematic features efficiently from the low-level features (four momenta), unlike tree-based methods. NN with kinematic variables only (without four momenta) leads to slightly lower but comparable purity 89.75% and 83.98% for the parton-level and the detector-level events, respectively. Finally, one can improve the purity at the cost of statistics. For example, we can cut on the value of $P_{b_1,\ell}$ (keeping the events for which $P_{b_1,\ell} > P_c$ or $P_{b_1,\ell'} > P_c$) targeting 99% purity, which leads to the efficiency of 76.2% and 54.3% for parton-level and detector-level events, respectively. More details will be discussed in Sec. III F.

C. Long Short-Term Memory

Next, we consider whether the NN can extract more information from the data by treating it as a sequence. We focus on the LSTM architecture [41], which was originally developed to overcome the vanishing gradient problem that often arises when training more traditional recurrent neural networks (RNNs). In collider physics, RNNs have been applied to study jet tagging problems, where the input is a sequence of jet kinematic information [34,42]. In dileptonic $t\bar{t}$ production, one can consider the time ordering of the final-state particles. Two b quarks would appear before two leptons or two neutrinos. Although the decay is somewhat short, it is worth investigating the performance of the LSTM and comparing against that of the DNN.

We implement the LSTM using TensorFlow [43]. The 16-dimensional input vector (without the missing transverse momentum) made up of four momenta of four particles (four sequences) is fed into two LSTM layers of 256 dimensions of the output space, followed by one DNN layer with 128 neurons. We use the same optimizer, loss function, and dropout used for DNN in Sec. III B. The number of trainable weights in the LSTM network is 1 million, which is comparable to the size of the DNN considered in the previous subsection. We obtain the purity of 89.27% and 83.95% for the parton-level and detector-level events, which are similar to those with DNN.

D. Attention-based network

Neural networks with attention are a technique that imitates cognitive attention. Attention is the ability to choose and concentrate on relevant stimuli, and respond accordingly

[44]. The effect of attention module enhances the important parts of the input data and fades out the less important parts such that the network devotes more computing power on a more relevant part of the data. Which part of the data is more important than the others depends on the problem and is learned via training. Attention-based networks are used in a wide variety of machine learning models, including in natural language processing and computer vision.

Recently, the combinatorial problem in $t\bar{t}$ production with fully hadronic top decays has been examined in detail using attention-based neural networks [4–6]. The authors showed that the performance significantly improved over the traditional kinematic methods. In this section, we apply an attention-based network to the twofold ambiguity in the dilepton production. Our NN is based on the standard self-Attention Network, which is relatively simpler than the architecture used in Refs. [4–6]. There are two reasons for this: both our input and output are simpler. For input, we only need to consider the two b-tagged jets without worrying about all the extra jets, which makes our input fixed length, compared to the variable lengths in the fully hadronic final state. As for the output, the fully hadronic channel is more complicated because one needs to identify which jets are selected as well as which jets form the top quark/W boson, while in our case, whatever label we obtain is essentially a twofold ambiguity, and it is mathematically equivalent to using a binary label.

We implemented the self-attention mechanism in our study by using the transformer encoder layer from PYTORCH [45]. The schematic diagram of the Attention Network is provided in Fig. 4. The input to this network is similar to that of the LSTM network, which is the 4-momentum of the four visible particles. Each particle momentum is first fed into dense embedding layers of dimensions 8, 32, and 64, consecutively. The embedded vectors are fed into three transformer encoder layers, where the inputs first flow through a multihead self-attention layer with four heads, and then into a feed-forward layer, with residual connection over each of the layers. The four output vectors of those layers with dimension 64 are then flattened, before feeding into another set of dense layers of dimensions 64 and 1, with the last layer being the output layer of the network that has a sigmoid activation. We obtain 89.8% purity for parton-level and 84.4% for detector-level events, which are comparable to results using DNN and LSTM.

E. Lorentz Boost Networks

Lorentz Boost Networks are motivated by particle kinematics in the rest frames of various particle combinations. The LBNs take the 4-momentum of n final-state particles as an input and create m arbitrary combinations of particles, which are boosted to the rest frame of m different combinations of particles. The set of m particle combinations and the set of m rest frames are constructed by a linear combination of input momenta determined by two trainable
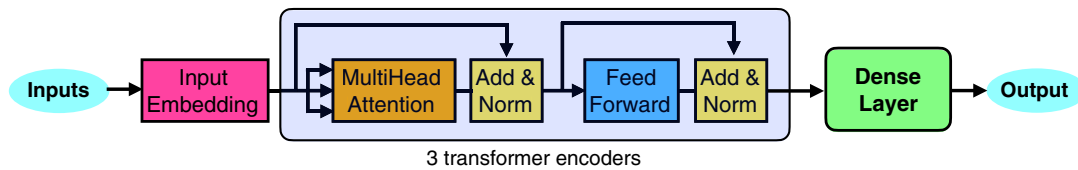
FIG. 4. Schematic diagram of the Attention Network in our analysis.
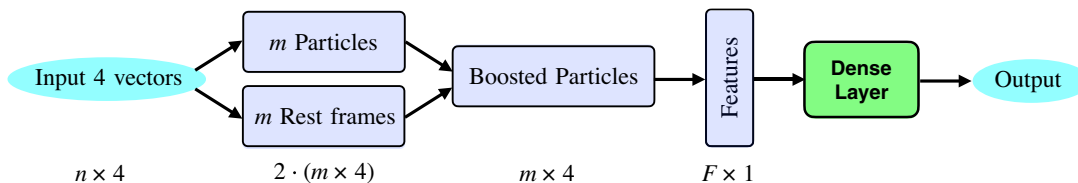

FIG. 5. Schematic diagram of the Lorentz Boost Network. See Refs. [46,47] for more details.

n by m matrices. This step creates four-momenta of m particles (boosted to the rest frames), which are used to calculate the high-level variables such as invariant mass or the angle between any particle pairs. The parameter m is taken to be a hyperparameter of the network. These high-level variables are combined with the 4-momentum of the boosted particles and fed into a simple DNN classification network. The LBNs have been applied for the semileptonic channel of $t\bar{t}h$ production [46,47] and dilepton $t\bar{t}$ production [48].

As the LBN architecture takes advantage of particle kinematics in various rest frames of composite particles, it is worth investigating the twofold ambiguity using the LBN. Since the complete kinematic information of all the final-state particles is unknown in our study, we use the momentum of the four visible particles (two b-tagged jets and two leptons) and the missing transverse momentum with zeros in the E and p_z component. Note that LBN does not use the mass information explicitly.

We use the TensorFlow implementation of the Lorentz Boost Network, which is available from https://git.rwth-aachen.de/3pia/lbn. The schematic diagram of the Lorentz Boost Network is shown in Fig. 5. We feed the 20-dimensional input momenta into the LBN layer. We set the number of particles (n) and rest-frame combinations (m) that LBN builds to $n = m = 8$, which leads to the optimal result. The LBN layer returns the $F = 69$ dimensional output made up of six kinematic features (E, p_T, η, ϕ, mass, and $\cos\theta_{ij}$, which is the decay angle of ith particle in the jth rest frame) for those eight particles, which are fed into two DNN layers with 512 neurons for each.

At parton level with 4-momentum input to the classifier, we obtain an accuracy of 89.8%. With the detector-level events, we obtain an accuracy 86%. For other values of m, we obtained a slightly worse result. For example, for $m = 4$, we obtained 88% purity at the parton level and 85% at detector level.

F. Comparison of different methods

Now, we compare the performance of each method in finding the correct and incorrect combination of a b quark and a lepton for the dilepton $t\bar{t}$ production. Table I summarizes the purity (P) and the efficiency (ϵ) for various ML approaches discussed in Sec. III and the existing methods described in Appendix. All methods use four momenta of four visible particles and the missing transverse momentum as basic inputs. Some examples use kinematic variables ($m_{b\ell}$, $M_{2CW}^{(b\ell)}$, $M_{2C t}^{(\ell)}$, topness, $\Delta R_{b\ell}$) in addition to the basic inputs, except for the cases "with kinematic variables only," where four momenta are omitted in the inputs.

Most methods lead to similar results. Especially, if BDT and a Random Forest take advantage of additional kinematic features, their performance is comparable to that of DNN. However, in the case of DNN, we have not observed such improvement even with the additional features. We believe that this is due to the flexibility and the efficiency of NN, which learns the high-level features (kinematic variables) from the lower-level data (four momenta inputs) and therefore NN does not need kinematic features as additional inputs. Even with different network structures, the results did not change significantly.

Table I shows that, when using some machine learning algorithms such as BDT or DNN, the methods with "kinematic variables only" lead to comparable results, surpassing the performance of the traditional use of the kinematic variables (topness, end points, or hemisphere). This observation tells us that ML algorithms are able to find nontrivial correlation among these kinematic variables efficiently. However, the same algorithms do not improve with both four momentum and kinematic variables as inputs. We also notice that most ML methods lead to the purity $\sim$90% and $\sim$85%, for the parton-level and detector-level events.

TABLE I. Summary of the purity (P) and the efficiency (ϵ) for various methods discussed in Sec. III and Appendix. Basic inputs are four momenta of two b-quarks, two leptons, and the missing transverse momentum. Some examples use kinematic variables ($m_{b\ell}$, $M_{2CW}^{(b\ell)}$, $M_{2Ct}^{(\ell)}$, topness, and $\Delta R_{b\ell}$) in addition to the basic inputs.

Algorithm	Section	Parton level		Detector level	
		P	ϵ	P	ϵ
End points method I ($m_{b\ell}$, $M_{2CC}^{(b\ell)}$, $M_{2CC}^{(\ell)}$)	A 1	0.816	1	0.789	1
End points method II ($\star$) ($m_{b\ell}$, $M_{2CW}^{(b\ell)}$, $M_{2Ct}^{(\ell)}$)	A 1	0.957	0.769	0.874	0.742
Hemisphere method	A 2	0.78	1	0.77	1
Recursive jigsaw	A 2	0.762	1	0.757	1
Topness method ($\blacklozenge$)	A 3	0.869	1	0.814	1
KLFitter ($\blacktriangle$)	A 4	0.866	1	0.776	1
Matrix element method	A 5	0.847	1	0.817	1
Boosted decision tree	III A	0.861	1	0.824	1
BDT with kinematic variables	III A	0.904	1	0.841	1
BDT with kinematic variables only	III A	0.895	1	0.834	1
Random Forest	III A	0.834	1	0.812	1
Random Forest with kinematic variables	III A	0.901	1	0.840	1
Random Forest with kinematic variables only	III A	0.897	1	0.837	1
DNN	III B	0.895	1	0.845	1
DNN ($\times$)	III B	0.990	0.721	0.990	0.543
DNN with kinematic variables	III B	**0.907**	1	0.846	1
DNN with kinematic variables only	III B	0.898	1	0.839	1
Long Short-Term Memory	III C	0.893	1	0.839	1
Attention Network	III D	0.898	1	0.844	1
Lorentz Boost Network	III E	0.898	1	**0.860**	1

Figure 6 summarizes the performance of each method in the purity-efficiency plot for the parton-level events (left) and the detector-level events (right), taking four momenta as inputs (except for the two end point methods, which use the kinematic variables as inputs.). Note that the purity is defined as precision, which is the ratio of the number of true positives to the number of events that pass the selection cuts [Eq. (2.7)]. The efficiency is the ratio of the number of
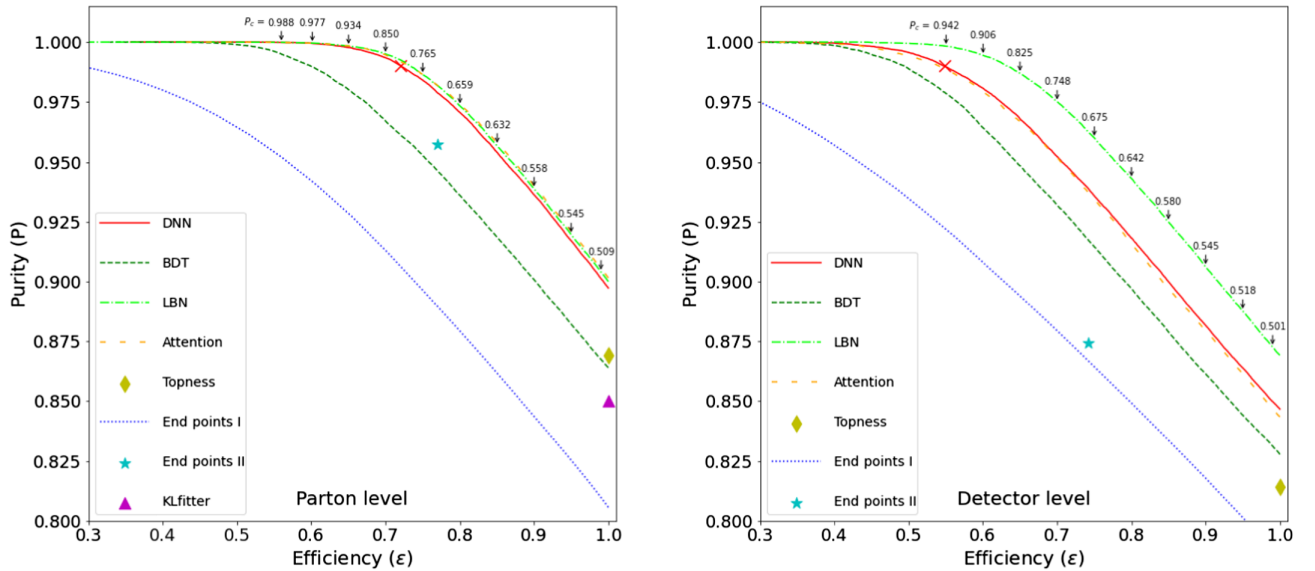

FIG. 6. Purity (P) vs efficiency (ϵ) curves for parton-level events (left) and detector-level events (right). For illustration, we show the values of P_c [for $\epsilon \in (0.55, 0.99)$ with 0.05 interval], which are used when cutting on the LBN score $P_{b_1,\ell} > P_c$ or $P_{b_1,\ell'} > P_c$.

events that pass the selection cuts to the total number of events [Eq. (2.6)]. The selection criteria that define the various purity-efficiency curves are described in Sec. II. Results using conventional methods such as the topness, end points method II, and KL fitter methods are marked as ◆, ★, and ▲, respectively. (The efficiency of end point method II is reduced due to the presence of unresolved events.) The × mark is shown as a reference to 99% purity for DNN. For the end point method I ($m_{b\ell}$, $M_{2CC}^{(b\ell)}$, $M_{2CC}^{(\ell)}$ without mass information), we cut on the transverse mass ($M_T^{t\bar{t}} = \sqrt{\hat{s}}_{\min}$) of the entire $t\bar{t}$ system to make the purity-efficiency curve [2,11,49,50].

DNN (solid red), Attention Network (long-dashed orange), and LBN (dot-dashed bright green) show the best performance with LBN being better for detector-level events. We did not find any difference between attention-based network and DNN. Results using BDT and end point method I are shown in the dark green-dashed and blue dotted curve, respectively.

Taking 99% (95%) as a benchmark purity, the efficiencies of ML methods are [0.284, 0.599, 0.721, 0.732, 0.734] ([0.560, 0.758, 0.861, 0.873, 0.867]) for parton level and [0.173, 0.495, 0.548, 0.541, 0.634] ([0.449, 0.645, 0.707, 0.0705, 0.779]) for detector level for end point I, BDT, DNN, Attention Network, and LBN, respectively. We see that deep learning and the LBN method in particular brings impressive efficiency gains at these high levels of purity, which could have major benefits for physics analyses that rely on resolving the combinatorial ambiguity in dileptonic $t\bar{t}$.

IV. FINDING THE CORRECT PARTITION WITHOUT ANY MASS INFORMATION

We have investigated various methods to resolve the twofold ambiguity in the dilepton $t\bar{t}$ production. Some methods use explicit mass information, while others do not. In this section, we consider the twofold ambiguity in the same topology as in Fig. 1 but with the unknown mass spectrum. For an illustration, we fix the mass of A, $m_A = 500$ GeV and scan over two-dimensional mass parameters, (m_B, m_C) with the mass constraint, $0 < m_C < m_B < m_A$. We will use the end point method I, hemisphere method, and neural networks. For the end point method I, we compute ($m_{b\ell}$, $M_{2CC}^{(b\ell)}$, $M_{2CC}^{(\ell)}$) without using explicit masses in the minimization. Here, the M_{2CC} is defined as

$$M_{2CC} \equiv \min_{\vec{q}_1, \vec{q}_2} \{\max[M_{P_1}(\vec{q}_1, \tilde{m}), M_{P_2}(\vec{q}_2, \tilde{m})]\},$$

$$\vec{\slashed{P}}_T = \vec{q}_{1T} + \vec{q}_{2T},$$

$$M_{A_1} = M_{A_2},$$

$$M_{B_1} = M_{B_2}, \tag{4.1}$$

where $M_{2CC}^{(b\ell)}$ and $M_{2CC}^{(\ell)}$ are the M_{2CC} variable applied to $(b\ell)$ and (ℓ) subsystems, respectively. They are similar to $M_{2CW}^{(b\ell)}$ and $M_{2Ct}^{(\ell)}$, but only the mass-equality conditions ($M_{A_1} = M_{A_2}$ and $M_{B_1} = M_{B_2}$) are imposed. However, no numerical values are used during minimization. Note that these variables satisfy $M_{T2} = M_2 \leq M_{2CC} \leq M_{2CC}$ (with mass input) $\leq \max(M_{P_1}, M_{P_2}) = m_A$, where M_{P_i} is the mass of the mother particle in the ith side, and M_2 is the same as Eq. (4.1) without the two mass constraints. These kinematic variables for the correct combination are bounded by their maximum end point, while the incorrect combination can vary and could violate the kinematic end point. Therefore, the mass variables for the correct combination tend to be smaller than values for the incorrect combination. We can resolve the twofold ambiguity without using mass information, taking the partition that gives more "smaller" values as the "correct" one. Note that there are no unresolved events, leading to 100% efficiency.

For DNN and LBN, we randomly choose 100 mass points in (m_B, m_C), covering the triangular parameter space. These points are a good representation of the two-dimensional mass space that we are interested in. We prepare 100,000 events for each chosen mass point and mixed them all before feeding them into NN. Although we do not feed numerical values of the chosen masses, NN learns how to resolve the twofold ambiguity by suitable interpolation in the entire two-dimensional mass space [51,52].

Figure 7 summarizes the results of the four methods (end point method I in the first row, hemisphere method in the second row, DNN in the third row, and LBN in the last row) for the parton-level (left) and for the detector-level (right) events. In the case of the end point and hemisphere methods, our results for the parton-level events are consistent with those in Ref. [1]. We have added similar plots for detector-level events as well in the right column.

Analogous to Fig. 6 for the top-quark decay, we see here that NNs outperform the traditional methods for the case of unknown mass spectrum as well. This is clearly depicted in Fig. 8, which shows the purity difference between LBN and end point method I (top) and between LBN and DNN (bottom) for parton-level (left) and detector-level (right) events. LBN and DNN provide very similar performance in the wide range of the mass parameters (bottom panel), while LBN surpasses the end point method I (top panel), whose performance drops in the degenerate spectrum, $m_B \approx m_A$ (in the right side of the figure) or $m_B \approx m_C$ (along the diagonal), for both parton level and detector level, and $m_B, m_C \ll m_A$ (in the left-bottom corner) for parton level.
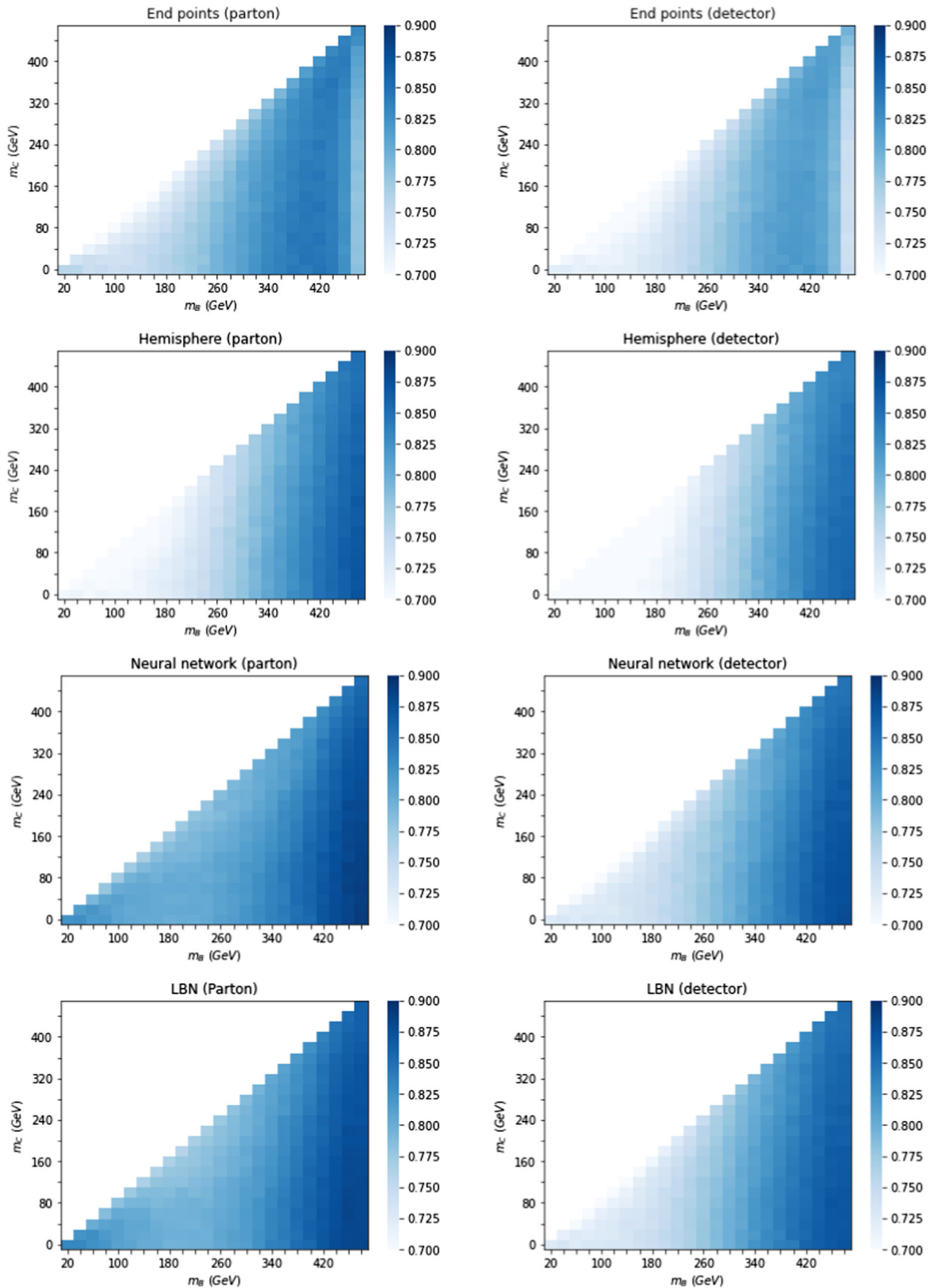
FIG. 7. Purity for choosing the correct partitioning with the end point method I (top), the hemisphere method (second row), DNN (third row), and LBN (bottom) in the two-dimensional (m_B, m_C) mass space for $m_A = 500$ GeV and $m_B > m_C$. The purity in the left (right) panels are obtained for parton-level (detector-level) events.

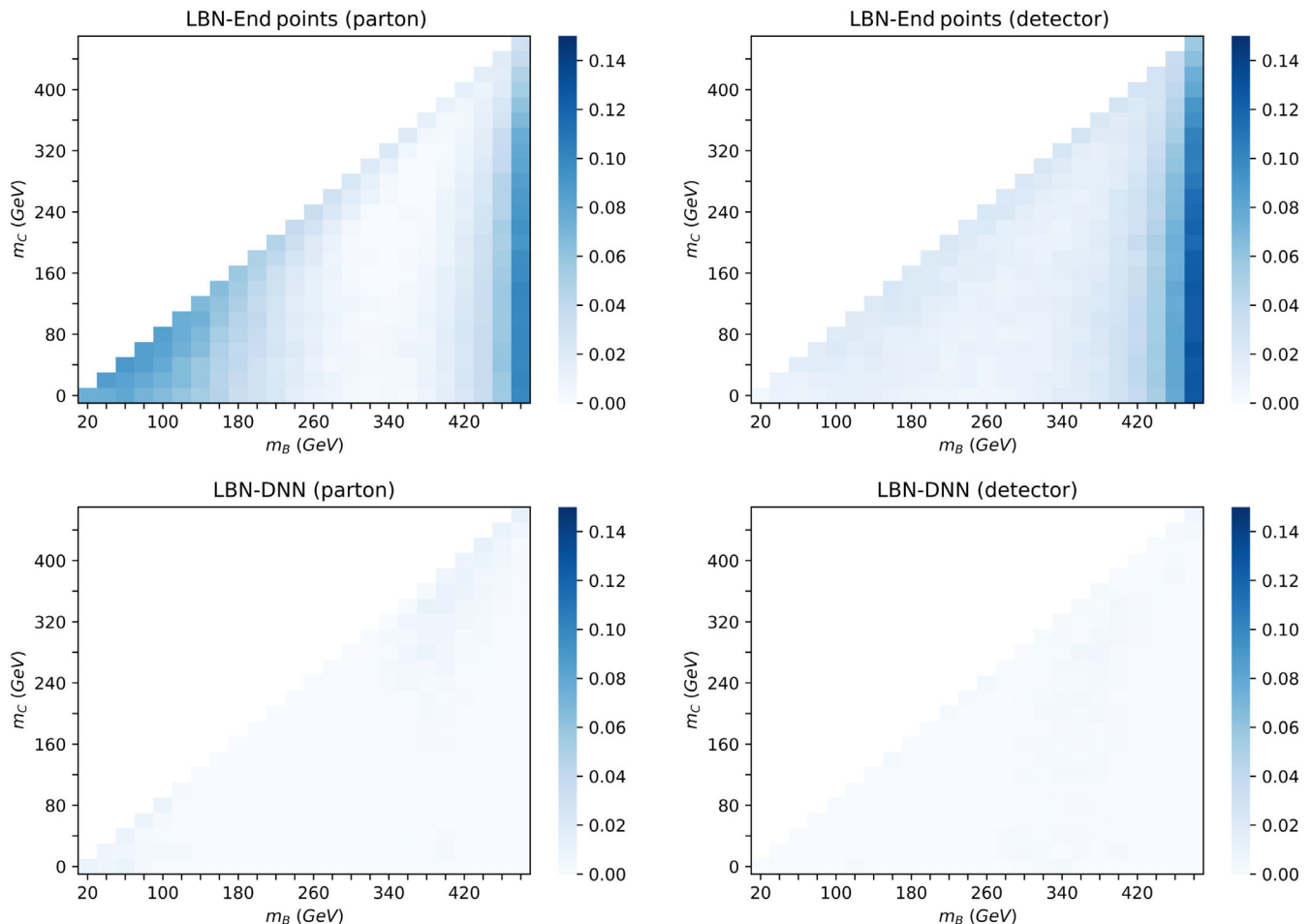FIG. 8. The purity-difference between LBN and end point method I (top) and between LBN and DNN (bottom) for parton-level (left) and detector-level (right) events.

V. CONCLUSION

Resolving the combinatorial problem in collider experiments is crucial for the discovery of new physics and precision measurements. The simplest combinatorial problem is the twofold ambiguity, which arises in the top-quark pair production. As the most massive fundamental particle in the standard model, the top quark is the only quark that decays before hadronization. The most precise measurements of top-quark properties such as top-quark mass are typically performed in the lepton-jet or dilepton channels. Therefore, it is important to find the correct assignment for the reconstructed objects (b-tagged jet and a lepton), as the incorrect assignment will reduce the precision in the measurements of the top-quark properties.

In the dileptonic $t\bar{t}$ events, we do not know the complete kinematic information due to the two neutrinos since they do not leave a trace in the detector. In particular, we do not know the individual momentum of each missing particle; we only know the total missing transverse momentum $\vec{P}_T$. We are unable to reconstruct the missing momenta exactly,

which poses the difficulty in assigning a lepton and a reconstructed b-quark pair.

In this paper, we have revisited with machine learning algorithms the combinatorial problem in the dileptonic $t\bar{t}$-like event topology. We have compared the performance of various algorithms against that of existing methods. In particular, we investigated the performance of attention-based networks, which has been found useful in the fully hadronic channel, and the Lorentz Boost Network, which is motivated by underlying physics principles. We found that most of the machine learning methods outperform the existing approaches based on kinematic variables. We then generalized the mass spectrum in consideration of new physics where the underlying mass spectrum is unknown, and therefore no kinematic end point information is available. We showed that the purity for selecting the correct partition is greatly improved by utilizing the machine learning techniques, especially in the regions where the particle spectrum is degenerate.

A specific application of our study would be measurement of the top-quark Yukawa coupling ($y_{t,SM}$) in the $t\bar{t}h$ production. As discussed in Refs. [18,19], top-quark

reconstruction in the dilepton channel plays an important role. For example, Table I shows that the end points method gives 74.2% efficiency and 87.4% purity at detector level (marked as ⋆ in Fig. 6), which is comparable to results presented in Refs. [18,19]. Using the M_2 reconstruction and the sideband subtraction, the uncertainty ($\delta\kappa_t$) on the top-quark Yukawa coupling ($\kappa_t = y_t/y_{t,\mathrm{SM}}$, which is the deviation from the SM value) is calculated as $\delta\kappa_t \sim$ 0.096 [19] at the HL-LHC. On the other hand, our results show that the same purity (87.4%) can be obtained at a higher efficiency using DNN (with 92% efficiency) and LBN (with nearly 100% efficiency), which would lead to a gain of 24% and 35% more events, respectively. Naive rescaling indicates that a gain α in the number of (both signal and background) events would lead to the $\alpha/2$ reduction in the precision on $\delta\kappa_t$. Following more a accurate procedure in Ref. [19], we estimate the uncertainty to be reduced to $\delta\kappa_t \sim 0.086$ and $\delta\kappa_t \sim 0.082$ using DNN and LBN, respectively.

Another application of our study would be testing Bell's inequality with top-quark pair production [53,54]. It is well known that the leptonic final state is maximally correlated to top-quark polarization, which motivates the measurement of violation of Bell's inequality performed in the dilepton channel. One of main goals is to reconstruct the spin density matrix in the center of the momentum frame, which requires full reconstruction of two neutrinos, resolving the twofold ambiguity. We anticipate that ML methods would help such a measurement [55].

Finally, we would like to make brief comments on impacts of ISR/FSR and backgrounds, postponing detailed analysis in a future study. Our goal in this paper was to investigate different NN architecture in depth to resolve the combinatorial problem (twofold ambiguity) without worrying about ISR/FSR and backgrounds so that we can try various ML methods and make fair comparison against existing approaches. We took the $t\bar{t}$ dilepton production as a specific example and then applied the ML methods to the case with arbitrary mass spectrum, keeping an application in BSM searches in mind. Therefore, we intentionally ignored ISR/FSR and backgrounds, as their impacts are model dependent in a sense that the hardness of ISR depends on the mass scale of new particles and backgrounds depend on the mass splitting.

Although we have not investigated the effects of ISR/ FSR using NNs, one can get a rough idea using the existing methods for $t\bar{t} + X$ production. References [18,19] studied the twofold ambiguity in the $t\bar{t}h$ production with $h \to b\bar{b}$, where parton-shower and hadronization are simulated with PYTHIA. For example, Ref. [19] finds the efficiency of end point II to be 78%, while our detector-level efficiency in this paper is 74%. The small difference is due to different set of cuts in data preparation. In Refs. [18,19], two hardest b-tagged jets are chosen, which effectively rejects

contamination arising from ISR/FSR, resulting in similar performance in the absence ISR/FSR.

The effect of ISR for top-quark production is studied in details in Ref. [10], using the M_{T2} and $m_{b\ell}$ method, where ISR/FSR and hadronization and detector effects are simulated with PYTHIA and pretty good simulation (PGS), respectively. Taking two leading b-tagged jets as b-quark candidates from the top decay, they find an efficiency of 51.7% with a purity of 94.9% at the 7 TeV LHC, which is comparable to end point I in the right panel of Fig. 6. This comparison indicates that effects of ISR/FSR and hadronization are mild or negligible when resolving the twofold ambiguity for the dilepton top-quark production.

Effects of ISR in reconstruction of new particle masses for new physics processes have been studied in Ref. [56]. They developed a novel technique to reduce ISR effects, taking gluino pair production and its decay to two jets and neutralino, $\tilde{g}\tilde{g}+j \to 5\mathrm{jets} + \tilde{\chi}_1^0\tilde{\chi}_1^0$, as an example. With five jet candidates, they compute $M_{T2}(i)$, excluding the ith jet ($i = 1, \ldots, 5$) and define $M_{T2}^{\min} = \min_{i=1,\ldots,5}(M_{T2}(i))$. Therefore, by construction, $M_{T2}^{\min} < M_{T2}^{\mathrm{end\ point}}$, and the $i_{\min}$th jet that satisfies $M_{T2}(i_{\min}) = M_{T2}^{\min}$ is considered to be the ISR jet. Surprisingly, this simple kinematic method gives reasonably good efficiency in identifying the ISR jet. For a given mass spectrum ($m_{\tilde{g}} = 685$ GeV and $m_{\tilde{\chi}_1^0} = 101.7$ GeV), they identify the ISR jet among five jets correctly 29% of time. This number increases up to 44% with a cut $\min(M_{T2}) > 500$ GeV. A similar method is used to distinguish $t\bar{t}$ and tW production [57].

Similarly, the importance of backgrounds also depends on the details of new physics including mass spectrum. For example, one can consider a new physics scenario, where the mass difference between B and C is much larger than m_W ($m_B - m_C \gg m_W$), which would lead to two very high p_T leptons with little backgrounds. On the other hand, if $m_B - m_C \lesssim m_W$, we would suffer from the $t\bar{t}$ background. Therefore, to better estimate their impacts, it is appropriate to consider a specific new physics case with a fixed mass spectrum.

Even within SM, the effects of ISR/FSR and backgrounds depend on what other particles are produced along the two top quarks. Note that all discussion in this paper is valid for any $t\bar{t} + X$ processes, as long as X is reconstructable. For example, one can consider $t\bar{t}h$ with $h \to \gamma\gamma$ or $t\bar{t}h$ with $h \to b\bar{b}$, where dominant backgrounds are different. Specifically, for top-quark production, ISR/FSR should be relatively harmless for our problem given the high b-tagging efficiency and small fake rates expected at the HL-LHC [10,26]. Therefore, in the case of $t\bar{t}$ production, background is not really a big concern. One can further reduce the backgrounds at the cost of statistics, utilizing various kinematic methods, as those methods themselves are optimized for signal ($t\bar{t}$) and backgrounds processes

would violate the end point structures [18,19]. An interesting question now is how well one can improve these results using NNs, the details of which we reserve for a future study.

So far, we have focused on resolving the twofold ambiguity. One of advantage of using the kinematic methods is their byproduct, i.e., some methods provide ansatz for the momentum of the missing particles, which can be used to reconstruct the full final state approximately. A similar study with neural networks has been done utilizing Lorentz structure of the four momenta [48]. Generalizing such a method for arbitrary mass spectrum would be useful in search for new physics beyond the standard model. We expect that NN-inspired reconstruction will help to expedite discovery as well as the precision measurement.

The source code can be found from [58].

ACKNOWLEDGMENTS

APPENDIX: OVERVIEW OF EXISTING METHODS

There are several methods to resolve the twofold ambiguity in the dilepton $t\bar{t}$ production. Although each method is introduced elsewhere, we did not find detailed description on the topic. Moreover, there is no comprehensive study on the twofold ambiguity, which compares the performance of all different methods using the same set of events. We find it valuable to provide a brief review on various existing methods in this Appendix, including the performance of each method, and proper comparison. Although we report the purity and efficiency of the individual method in resolving the twofold ambiguity, a method with a higher efficiency or a higher purity does not necessarily mean a better method. Each method is unique,

has different motivations, and is based on different set of assumptions. Depending on the underlying event topology and the target study point, they may show different levels of performance, hence it is prudent to keep as many tools as possible in the analysis toolbox.

1. End point methods

For events with two missing particles, the on-shell constrained M_2 variable [24,31] provides a good estimate for the unmeasured invisible momenta and thus can be useful to discriminate combinatorial ambiguities [1,2,31]. It is defined as a $(3 + 1)$-dimensional version of M_{T2} [59–62],

$$M_2(\tilde{m}) \equiv \min_{\vec{q}_1, \vec{q}_2}\{\max[M_{P_1}(\vec{q}_1, \tilde{m}), M_{P_2}(\vec{q}_2, \tilde{m})]\},$$
$$\vec{P}_T = \vec{q}_{1T} + \vec{q}_{2T}, \tag{A1}$$

where the *actual* parent masses, M_{P_i}, are considered instead of their transverse masses, M_{TP_i} ($i = 1, 2$). The $\tilde{m}$ is the test mass, which we take to be zero in our study. Note that the minimization is performed over the 3-component momentum vectors $\vec{q}_1$ and $\vec{q}_2$ of the two missing particles [2], assuming the missing transverse momentum constraint as shown in Eq. (A1). At this point, M_{T2} and M_2 are known to be equivalent, in the sense that the resulting two variables will have the same numerical value $M_2 = M_{T2} \leq \max(M_{P_1}, M_{P_2})$ [2,24,63]. Figure 9 shows M_2 distribution of the correct (red, solid) and incorrect (blue, dotted) pair for parton-level events (left) and detector-level events (right). The correct pairing respects the mass bound, while the incorrect paring goes beyond the expected end point, which is the mass of the top quark in this example.

However, for the $t\bar{t}$ production considered in this paper (in general, $t\bar{t} + X$, where X is fully reconstructed), the value of the W-boson mass m_W is experimentally known, and therefore we can introduce the following variable in the $(b\ell)$ subsystem:

$$M_{2CW}^{(b\ell)} \equiv \min_{\vec{q}_1, \vec{q}_2}\{\max[M_{t_1}(\vec{q}_1, \tilde{m}), M_{t_2}(\vec{q}_2, \tilde{m})]\},$$
$$\vec{P}_T = \vec{q}_{1T} + \vec{q}_{2T},$$
$$M_{t_1} = M_{t_2},$$
$$M_{W_1} = M_{W_2} = m_W. \tag{A2}$$

Here, the second constraint $M_{t_1} = M_{t_2}$ requires the equality of two parent masses without use of a specific numerical value, while the true W mass is used in the third constraint $M_{W_1} = M_{W_2} = m_W$. Similarly, taking the top-quark mass m_t in the minimization, we can define a new variable in the (ℓ) subsystem:
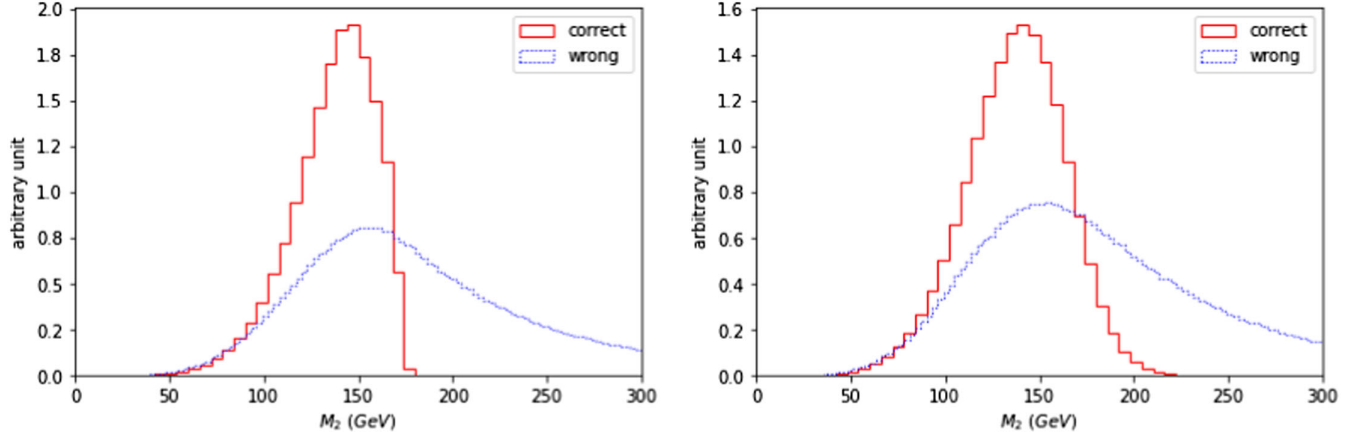
FIG. 9. M_2 distribution of the correct (red, solid) and incorrect (blue, dotted) pair for parton-level events (left) and detector-level events (right).

$$M_{2Ct}^{(\ell)} \equiv \min_{\vec{q}_1, \vec{q}_2}\{\max\left[M_{W_1}(\vec{q}_1, \tilde{m}), M_{W_2}(\vec{q}_2, \tilde{m})\right]\},$$

$$\vec{P}_T = \vec{q}_{1T} + \vec{q}_{2T},$$

$$M_{W_1} = M_{W_2},$$

$$M_{t_1} = M_{t_2} = m_t. \tag{A3}$$

By construction, $M_{2Ct}^{(\ell)} \leq m_W$ and $M_{2CW}^{(b\ell)} \leq m_t$.

Figure 10 shows M_{2CW} (top) and M_{2Ct} (bottom) distributions of the correct (red, solid) and incorrect (blue, dotted) pair for parton-level events (left) and detector-level events (right), respectively. We note that M_{2CW} shows a sharper distribution compared to M_2 distribution, while
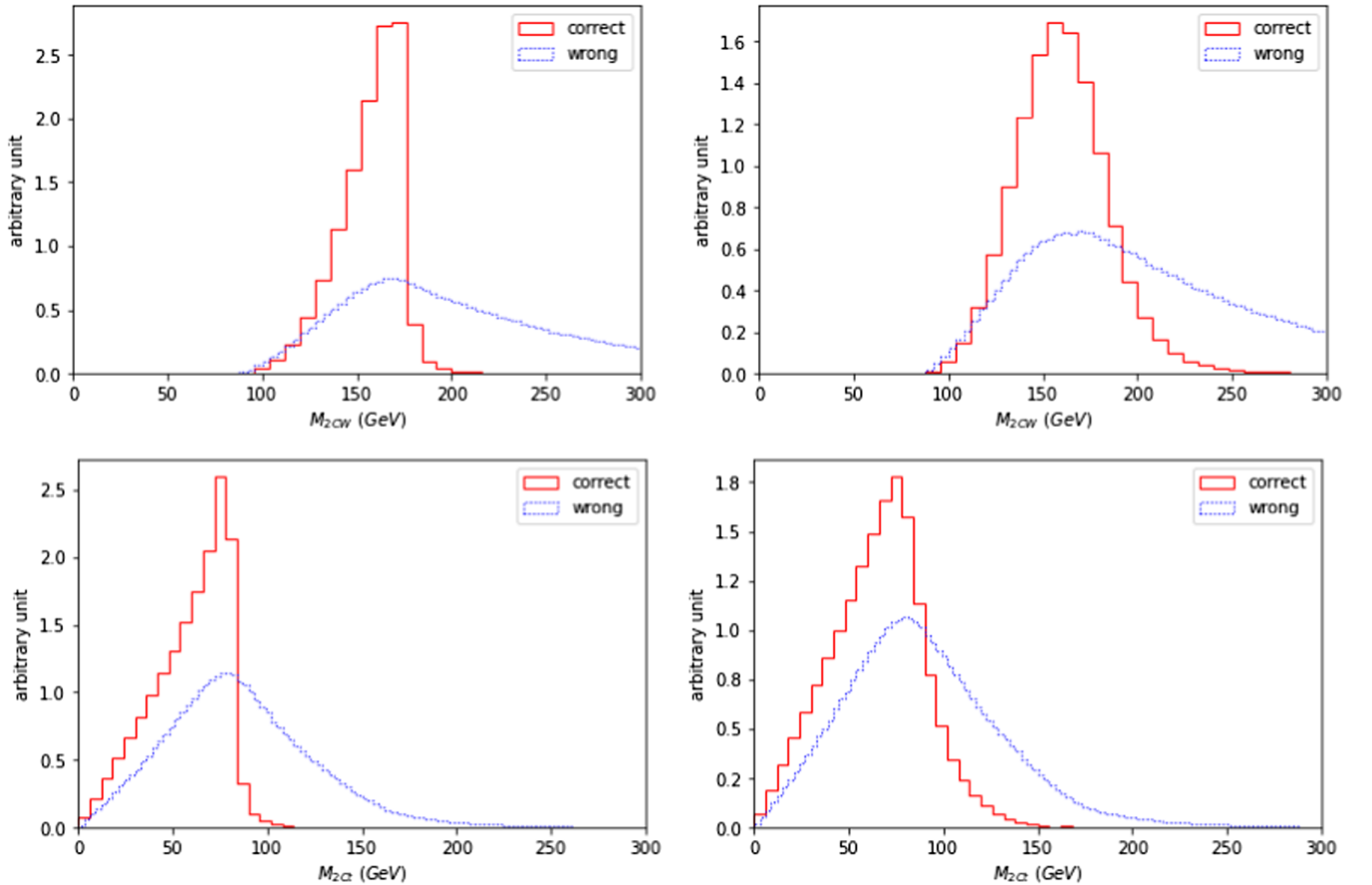

FIG. 10. M_{2CW} [top, Eq. (A2)] and M_{2Ct} [bottom, Eq. (A3)] distributions of the correct (red, solid) and incorrect (blue, dotted) pair for parton-level events (left) and detector-level events (right).
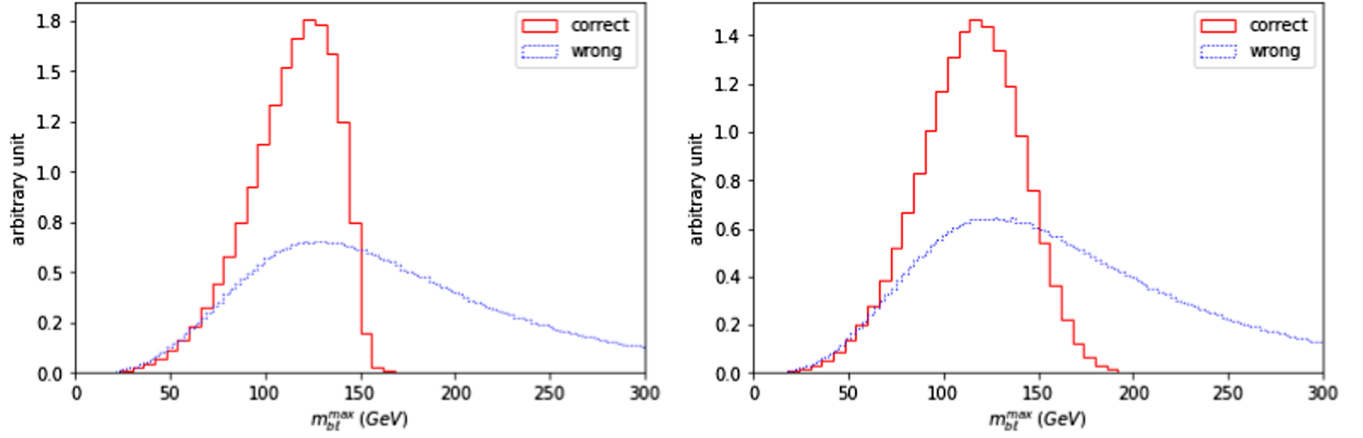
FIG. 11. $\max\{m_{b\ell}^{(i)}\}$ distribution of the correct (red, solid) and incorrect (blue, dotted) pair for parton-level events (left) and detector-level events (right).

both respect the same end point (top-quark mass). For both cases, the input mass (or test mass) is zero. However, M_{2CW} distribution begins from m_W since the numerical value of the W-boson mass is imposed during the minimization as shown in Eq. (A2).

For the minimization of M_2, M_{2Ct}, and M_{2CW}, we use OPTIMASS [64]. While these mass-constraining variables are proposed for mass measurement originally, one could use them for other purposes such as measurement of spins and couplings [1,10]. In our study, we use these variables to fully reconstruct the final state of our interest, with the unknown neutrino momenta obtained via minimization procedure. These momenta may or may not be true particle momenta, but they provide important nontrivial correlations with other visible particles in the final state, which help reconstruction.

The other useful kinematic variable is the invariant mass $m_{b\ell}^{(i)}$ of b and ℓ in ith pairing ($i = 1, 2$). The invariant mass distribution for the correct pairing is bounded by the minimum and maximum values ($m_{b\ell}^{\min} \leq m_{b\ell} \leq m_{b\ell}^{\max}$), which are given by

$$(m_{b\ell}^{\max/\min})^2 = \frac{1}{2}\left(m_t^2 - m_W^2 + m_b^2 \right.$$
$$\left. \pm \sqrt{((m_t - m_b)^2 - m_W^2)((m_t + m_b)^2 - m_W^2)} \right),$$
$$(A4)$$

which become $m_{b\ell}^{\max} = \sqrt{m_t^2 - m_W^2}$ and $m_{b\ell}^{\min} = 0$ in the $m_b \to 0$ limit. Figure 11 shows the $m_{b\ell}$ distribution of the correct (red, solid) and incorrect (blue, dotted) pair for parton-level events (left) and detector-level events (right). We take the larger ($\max\{m_{b\ell}^{(i)}\}$) of two possible invariant masses for each choice of partitioning.

Now, we follow the procedure described in Ref. [1] to resolve the twofold ambiguity. For each event, we compute

the following three-dimensional vector for both correct and incorrect pairings:

$$(m_{b\ell}^{\max} - \max_i\{m_{b\ell}^{(i)}\}, m_t - M_{2CW}^{(b\ell)}, m_W - M_{2Ct}^{(\ell)}). \quad (A5)$$

The correct combination would respect the anticipated end points of $m_{b\ell}$, $M_{2CW}^{(b\ell)}$, and $M_{2Ct}^{(\ell)}$, leading to positive components of above three-dimensional mass space. On the other hand, the incorrect pairing could give either sign. Therefore, by requiring that the partition which gives more "plus" signs as the "correct" one, we can resolve the twofold ambiguity. Then, we treat the corresponding momenta of the two missing particles, which are obtained via the minimization procedure, as "approximate" momenta of the two missing neutrinos. If both partitions give the same numbers of positive and negative signs, we discard such events, since they are "unresolved cases." For parton-level events, we find such unresolved events are 23%, while it is 26% for detector-level events.

With the definitions of efficiency and purity in Eqs. (2.6) and (2.7), selecting the resolved events only, the end point method leads to 77% efficiency and 96% purity for parton-level events and 74% efficiency and 87% purity for detector-level events, respectively. We call this method end point method II, which is shown as (⋆) in Table I and Fig. 6. Similar methods were considered in the literature, using $(m_{b\ell}, p_t)$ [9] or $(m_{b\ell}, M_{T2})$ [10,11]. We find that the latest study with $(m_{b\ell}, M_{2Ct}, M_{2CW})$ in Ref. [1] gives the best result concerning the combinatorial problem.

Although the use of the mass spectrum resolves the twofold ambiguity more accurately, it is instructive, and perhaps necessary in some examples, to repeat a similar analysis without using mass information explicitly. We use $(m_{b\ell}, M_{2CC}^{(b\ell)}, M_{2CC}^{(\ell)})$ without prior knowledge of mass spectrum. For each event, there are two possible values for each component of $(m_{b\ell}, M_{2CC}^{(b\ell)}, M_{2CC}^{(\ell)})$. Choosing the
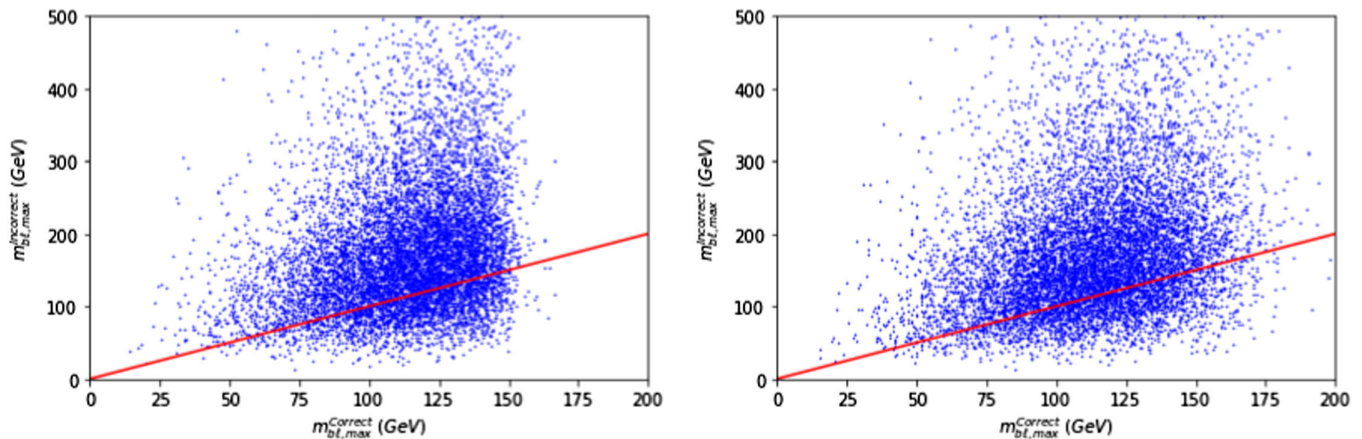
FIG. 12.   Scatter distribution of $(m_{b\ell}^{\text{correct}}, m_{b\ell}^{\text{incorrect}})$ for parton-level events (left) and detector-level events (right). The red-diagonal line represents $m_{b\ell}^{\text{correct}} = m_{b\ell}^{\text{incorrect}}$.

combination that gives more smaller components as the correct one, we obtain 81.6% purity for parton-level events and 78.9% purity for detector-level events. Since there are three quantities that we compare, in this case, there is no unresolved event, and therefore the efficiency is 100%. This is denoted as end point method I in Table I and Fig. 6.

The end point method is very general and can be extended easily to different event topologies. However, it also has a few issues. First, the end points are sensitive to the detector effects and get smeared significantly. A proper deconvolution procedure (using the transfer function) is required for a better performance. Second, the finite widths of intermediate particles (top quark and W in this case) also affect the shape of the kinematic distributions, in which case the correct pairing could violate the expected end point. Finally, there is quite large number of unresolved events, which are discarded, when mass information is imposed. These issues motivate us to explore different methods to maximize both efficiency and purity.

2. Hemisphere method and recursive jigsaw reconstruction

Partitioning reconstructed particles into two decay chains is often addressed by the so-called hemisphere algorithm, developed originally within CMS [12] and later adopted in many phenomenological studies [13–15]. Using the standard hemisphere method, we cluster the visible particles into two groups by keeping the invariant mass of each group to a minimum. For the $t\bar{t}$-like topology considered in this paper, it is straightforward to see that the hemisphere method is nothing but a variation of invariant mass method without relying on the end point. The pairing whose invariant mass is smaller is chosen to be the correct pair, and the other is chosen to be the incorrect one. Since the hemisphere method does not use the numerical value of

the end point, there is no violation of the end point, and therefore we do not discard any event, keeping 100% efficiency. We obtain the 78% purity for parton-level events and 77% purity for detector-level events. These events are ones in the left-upper corner of the $m_{b\ell}^{\text{correct}} = m_{b\ell}^{\text{incorrect}}$ line (red, solid) in Fig. 12. An advantage of this method is that one could obtain relatively good purity with 100% efficiency without using the mass spectrum via fast computation. The results are not very sensitive to the detector effects. However, if a high purity sample is required, the method must be extended at the cost of statistics. For example, as we discussed in the previous method, one can further improve on the hemisphere algorithm by suitable cuts on the invariant mass and either the jet p_T [9] or M_{T2} [10,11].

A similar idea is discussed in the Recursive Jigsaw Reconstruction method. It is a technique for analyzing reconstructed particles in the presence of kinematic unknowns arising from the unmeasured particles and the combinatoric unknowns associated with indistinguishable particles [21]. The method provides a very general framework, which can be applied to various processes at collider experiments. In particular, the dileptonic $t\bar{t}$ production is one of the examples discussed in Ref. [21]. Because of the simple nature of the twofold ambiguity, the Recursive Jigsaw Reconstruction method becomes very similar to the hemisphere method. It takes the smaller of the two squared mass sum, $\min(m_{b_1\ell^+}^2 + m_{b_2\ell^-}^2, m_{b_2\ell^+}^2 + m_{b_1\ell^-}^2)$, as the correct pair. In other words, it chooses the combination where the sum of 4-vector inner products is smallest, effectively pairing particles flying closer together as expected from a common decay source. This algorithm gives 76.2% of purity for parton-level events and 75.7% for detector-level events. We also used RESTFRAMES [21,65] and obtained a similar purity. These results are very comparable to what we have obtained using the hemisphere method.
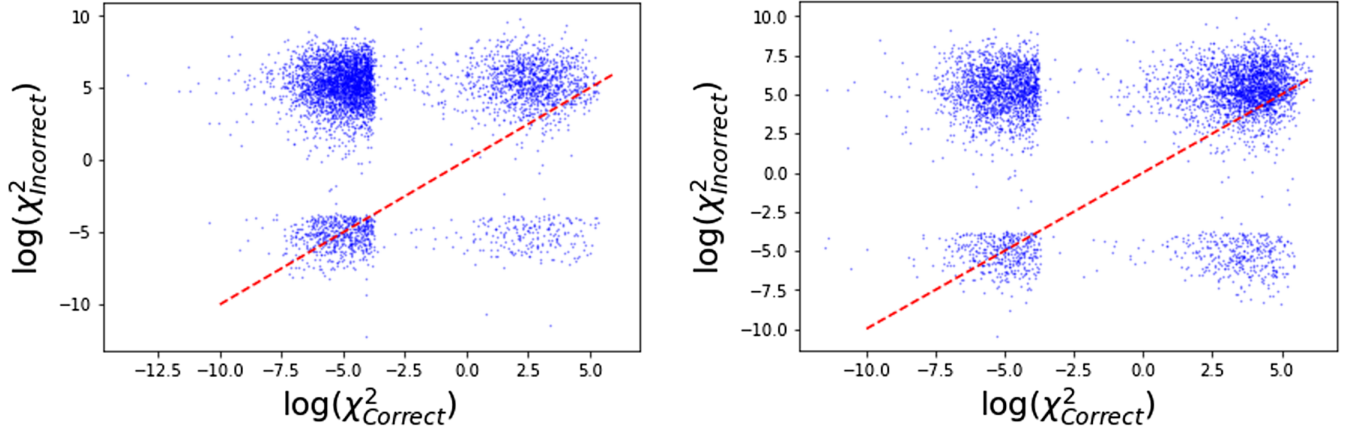
FIG. 13. χ^2 value of the correct (x axis) and the incorrect pair (y axis) for parton-level events (left) and detector-level events (right). The red-diagonal line represents $\chi^2_{12} = \chi^2_{21}$.

3. Topness

The topness (T) was originally proposed in search for supersymmetry [16], and then later a modified version was used in search for the double Higgs production [22,23]. It is nothing but a simple mass fitting in the $t\bar{t}$-like topology with a given mass spectrum.

The definition of the χ^2 statistic is given by

$$\chi^2_{ij} \equiv \min_{\vec{\not{p}}_T = \vec{p}_{\nu T} + \vec{p}_{\bar{\nu} T}} \left[\frac{(m^2_{b_i \ell^+ \nu} - m^2_t)^2}{\sigma^4_t} + \frac{(m^2_{\ell^+ \nu} - m^2_W)^2}{\sigma^4_W} \right.$$
$$\left. + \frac{(m^2_{b_j \ell^- \bar{\nu}} - m^2_t)^2}{\sigma^4_t} + \frac{(m^2_{\ell^- \bar{\nu}} - m^2_W)^2}{\sigma^4_W} \right], \quad \text{(A6)}$$

where b_1 and b_2 are the p_T-ordered b-tagged jets. The topness T is defined as $T = \min(\chi^2_{12}, \chi^2_{21})$. We find the momentum information of the two neutrinos via minimization algorithm over the missing transverse momentum constraint, $\vec{\not{p}}_T = \vec{p}_{\nu T} + \vec{p}_{\bar{\nu} T}$. Then, the correct pair is identified as the combination of b and ℓ, which gives the smaller of the two χ^2 values, being more consistent with the top-quark and W-boson masses. The correctly identified events reside in the left-upper corner of the diagonal line in Fig. 13. We obtain the 85% purity for parton-level events and 81% purity for detector-level events, keeping 100% efficiency, as shown as ($\blacklozenge$) Table I and Fig. 6. Note that in our minimization we choose $\sigma_t = \sigma_W = 5$ GeV following Refs. [22,23]. We find that the purity is not very sensitive to the choice of the σ parameters.

4. Kinematic likelihood fitter

The Kinematic Likelihood Fitter (KLFitter) [17,58] is a library for kinematic fitting using a likelihood approach developed for the top-quark reconstruction. The reconstruction of dileptonic $t\bar{t}$ events utilizes the neutrino-weighting method to solve the underconstrained kinematic system in the final state with two b-tagged jets, two charged leptons, and the missing transverse momentum. The likelihood consists of three parts as shown in Eq. (A7):

$$\mathcal{L} = \prod_{i=x,y} \mathcal{G}(E^{\text{miss}}_i | p^{\nu_1}_i, p^{\nu_2}_i, \sigma^{\text{miss}}_i(m_t, m_W, \eta_{\nu_1}, \eta_{\nu_2}))$$
$$\times \prod_{i=1}^{2} \mathcal{G}(\eta_{\nu_i} | m_t) \times (m_{\ell_1, q_1} + m_{\ell_2, q_2})^\alpha$$
$$\times \prod_{i=1}^{2} \mathcal{W}_{\text{jet}}(p^{\text{detector}}_{\text{jet},i} | p^{\text{parton}}_{\text{jet},i})$$
$$\times \prod_{i=1}^{2} \mathcal{W}_\ell(p^{\text{detector}}_{\ell,i} | p^{\text{parton}}_{\ell,i}). \quad \text{(A7)}$$

The Gaussian distribution $\mathcal{G}(\cdots)$ in the first line of Eq. (A7) is two dimensional and attempts to fix the neutrino momenta via the missing transverse momentum constraint. The second line contains two one-dimensional Gaussian distributions multiplied by the inverse of the invariant masses. α is a tuning parameter of the likelihood, and we use the default value $\alpha = -2$. This choice is consistent with choosing the smaller value of invariant masses, which could increase the likelihood. The last two lines include the transfer functions for the two charged leptons ($\mathcal{W}_\ell$) and the two jets ($\mathcal{W}_{\text{jet}}$), which are defined in Eqs. (A8) and (A9), respectively. The transfer function contains the response of the detector and is the conditional probability to observe a detector-level event for a given parton-level configuration. We use the transfer function introduced in Refs. [66–69],

$$\mathcal{W}_\ell(p^{\text{detector}}_{\ell,i} | p^{\text{parton}}_{\ell,i}) = 1 \quad \text{for } p^{\text{detector}}_{\ell,i} = p^{\text{parton}}_{\ell,i}, \quad \text{(A8)}$$
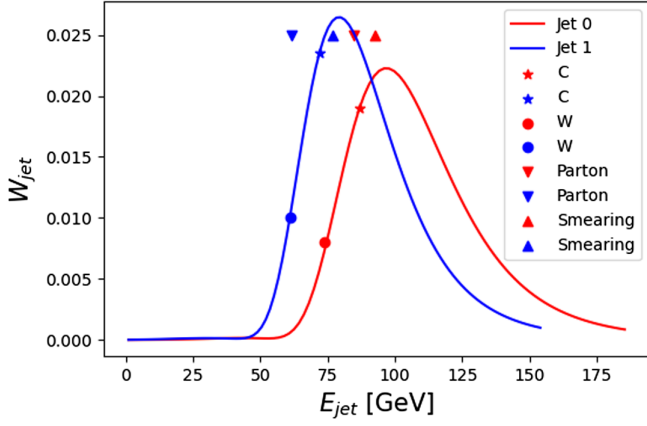
FIG. 14. Distribution of the transfer function as a function of $E_{\rm jet}$ for the two b-quark candidates in a sample event. True MC inputs are marked as filled upward triangle for parton-level energy (without detector effects) and as filled downward triangle for smeared energy (with detector effects). The likelihood method fixes the most probable $E_{\rm jet}$, which are denoted by filled star for correct pairing and filled circle for incorrect paring.

$$
\mathcal{W}_{\rm jet}(p_{{\rm jet},i}^{\rm detector}|p_{{\rm jet},i}^{\rm parton}) = \frac{1}{\sqrt{2\pi}(f_2 + f_3 f_5)}
$$
$$
\times \left(\exp^{-\frac{(\Delta - f_1)^2}{2f_2^2}} + f_3 \exp^{-\frac{(\Delta - f_4)^2}{2f_5^2}} \right),
$$
(A9)

where $\Delta = p_{{\rm jet},i}^{\rm detector} - p_{{\rm jet},i}^{\rm parton}$ and $f_j = a_j + p_{{\rm jet},i}^{\rm parton} b_j$. The parameters a_j and b_j are determined from fully simulated $t\bar{t}$ events which are given in Refs. [66–68]. We use $\mathcal{W}_\ell(p_{\ell,i}^{\rm detector}|p_{\ell,i}^{\rm parton}) = 1$ for $p_{\ell,i}^{\rm detector} = p_{\ell,i}^{\rm parton}$ (otherwise, $\mathcal{W}_\ell = 0$) and $\mathcal{W}_{\rm jet}(p_{{\rm jet},i}^{\rm detector}|p_{{\rm jet},i}^{\rm parton}) = 1$ for $p_{{\rm jet},i}^{\rm detector} = p_{{\rm jet},i}^{\rm parton}$ (otherwise, $\mathcal{W}_{\rm jet} = 0$) for the parton-level analysis. The likelihood distinguishes between neutrinos and anti-neutrinos, and the charged leptons are paired accordingly. We refer to Ref. [17] for further details.

Figure 14 illustrates how the above procedure works. Two solid curves represent the distribution of the transfer function as a function of $E_{\rm jet}$ for the two b-quark candidates in a sample event. True Monte Carlo (MC) inputs are marked as filled downward triangle for parton-level energy (without detector effects) and as filled upward triangle for smeared energy (with detector effects). The likelihood method fixes the most probable $E_{\rm jet}$, which are denoted by filled star for correct pairing and filled circle for incorrect paring. Note that these two points (either filled star or filled circle) do not coincide with the maximum location of the two curves [$\mathcal{W}(E_{\rm jet})$]. This is because the two points are obtained by maximizing the *total* likelihood Eq. (A7), including both transfer function, Gaussian distributions and invariant masses.

To distinguish the correct and incorrect pairings, we use the ratio of two likelihoods for a given event x, $\frac{\mathcal{L}(x|C)}{\mathcal{L}(x|W)}$. Therefore, the constant coefficients in Eq. (A7) cancel out. Requiring $\frac{\mathcal{L}(x|C)}{\mathcal{L}(x|W)} > 1$, we obtain the 86.6% purity for parton-level events and 77.6% purity for detector-level events, keeping 100% efficiency. Although the likelihood analysis is well motivated, we find in practice that the obtained neutrino momenta are not always sufficiently close to "true" momenta, which would encourage the use of the matrix element method to be discussed later. It is also difficult to generalize the method, when the masses of particles are unknown.

5. Matrix element method

All methods that we discuss in this paper have one thing in common. They all attempt to calculate a good variable for distinguishing different hypotheses. Often, these hypotheses are signal plus background and background alone. The Neyman-Pearson lemma suggests that the likelihood ratio is the optimal variable to distinguish hypotheses [32,33]. The likelihood and the probability are the same function with a different choice of dependent and independent variables, so in particle physics, the likelihood could be given by the differential cross section normalized by the total cross section,

$$
\mathcal{P}(\vec{p}_i^{\rm vis}|\vec{\theta}) = \frac{1}{\sigma}\int dx_1 dx_2 \frac{f_1(x_1)f_2(x_2)}{2sx_1x_2}
$$
$$
\times \left[\prod_{j\in{\rm final}} \int \frac{d^3 p_j}{(2\pi)^3 2E_j} \right] |\mathcal{M}_{\vec{\theta}}(p_j)|^2 \prod_{j\in{\rm vis}} \mathcal{W}(\vec{p}_j, \vec{p}_j^{\rm vis}),
$$
(A10)

where $|\mathcal{M}_{\vec{\theta}}(p_j)|^2$ is the squared matrix element for a given set of parameters $\vec{\theta}$. This is where the Matrix Element Method gets its name. The x_i is the momentum fraction of each parton ($i = 1, 2$) participating in the collision, $f_i(x_i)$ is the parton distribution function [70], and s is the center-of-mass energy of the collider. The transfer function $\mathcal{W}$ parametrizes the detector resolution (as discussed in Sec. A 4), and the integration is performed over all final-state particle momenta (over entire phase space.). For the visible final-state particles, we integrate over transfer functions. For the invisible final-state particles, we integrate over the missing momenta.

The Matrix Element Method is a type of multivariate analysis and provides an optimal variable. However, it can be very challenging to integrate over transfer functions (and accurately parametrizing the detector response in terms of transfer functions) and invisible particle momenta. In practice, it may be much easier to get a pretty good variable by using machine learning techniques on Monte Carlo data. Another challenge is how to incorporate the effects of

additional radiation and/or other higher-order corrections properly [71,72]. Unfortunately, the Matrix Element Method requires the full knowledge of the underlying process including masses, spins, and couplings of particles. It is difficult to consider a model-independent analysis.

Nevertheless, the biggest motivation for using Matrix Element Method (beyond Neyman-Pearson optimality) is physical transparency. It is easy to understand where the sensitivity comes from when the discriminating variable is calculated explicitly. To get the basic idea on how well the Matrix Element Method could resolve the twofold ambiguity, let us consider the $gg \to t\bar{t} \to W^+ W^- b\bar{b} \to b\bar{b}\ell^+\ell^-\nu\bar{\nu}$ process at the parton level. We reconstruct the two top quarks, taking both correct and incorrect combination with *true* neutrino momenta. We find that 93.8% of the time the squared matrix element for the correct combination is larger. This result implies that it would be difficult to improve beyond 93.8% using any methods that we are developing. A similar exercise gives 86.2% for detector-level events.

For a more realistic investigation, we use MadGraph5_ aMC@NLO to generate the squared matrix element for $gg \to t\bar{t} \to W^+ W^- b\bar{b} \to b\bar{b}\ell^+\ell^-\nu\bar{\nu}$. By computing the ratio, $\frac{\mathcal{P}(\vec{p}^{\text{vis}}|\text{correct})}{\mathcal{P}(\vec{p}^{\text{vis}}|\text{incorrect})}$ with MOMEMTA (a modular toolkit for the Matrix Element Method at the LHC) [73], we obtain 84.7% for the parton-level events and 81.7% for the detector-level events.

6. Analytic reconstruction

One can in principle solve up to a fourfold ambiguity for the neutrino momentum in the dilepton production using on-shell conditions of the top quark and the W boson [74–77]. This analytic approach naturally solves the two-fold ambiguity, when trying to reconstruct the final state [75,76]. However, the method is very sensitive to the exact value of the intermediate particle masses (the top-quark and the W-boson masses in this case), and the off-shell effects could result in no solution (or imaginary solution). We will not further investigate the features of analytical reconstruction in our current study, since these masses are unknown *a priori* when applying the method to new physics beyond the SM, and it is difficult to generalize unlike other methods. Finally, we refer to Refs [78–83] for readers who are interested in the singularity variables.

[1] D. Debnath, D. Kim, J. H. Kim, K. Kong, and K. T. Matchev, Phys. Rev. D **96,** 076005 (2017).
[2] A. J. Barr, T. J. Khoo, P. Konar, K. Kong, C. G. Lester, K. T. Matchev, and M. Park, Phys. Rev. D **84,** 095031 (2011).
[3] M. Feickert and B. Nachman, arXiv:2102.02770.
[4] A. Shmakov, M. J. Fenton, T.-W. Ho, S.-C. Hsu, D. Whiteson, and P. Baldi, arXiv:2106.03898.
[5] M. J. Fenton, A. Shmakov, T.-W. Ho, S.-C. Hsu, D. Whiteson, and P. Baldi, arXiv:2010.09206.
[6] J. S. H. Lee, I. Park, I. J. Watson, and S. Yang, arXiv:2012.03542.
[7] M. Kim, P. Ko, J.-h. Park, and M. Park, arXiv:2111.07806.
[8] A. Badea, W. J. Fawcett, J. Huth, T. J. Khoo, R. Poggi, and L. Lee, arXiv:2201.02205.
[9] A. Rajaraman and F. Yu, Phys. Lett. B **700,** 126 (2011).
[10] P. Baringer, K. Kong, M. McCaskey, and D. Noonan, J. High Energy Phys. 10 (2011) 101.
[11] K. Choi, D. Guadagnoli, and C. B. Park, J. High Energy Phys. 11 (2011) 117.
[12] G. L. Bayatian *et al.* (CMS Collaboration), J. Phys. G **34,** 995 (2007).
[13] S. Matsumoto, M. M. Nojiri, and D. Nomura, Phys. Rev. D **75,** 055006 (2007).
[14] W. S. Cho, K. Choi, Y. G. Kim, and C. B. Park, J. High Energy Phys. 02 (2008) 035.
[15] M. M. Nojiri, Y. Shimizu, S. Okada, and K. Kawagoe, J. High Energy Phys. 06 (2008) 035.
[16] M. L. Graesser and J. Shelton, Phys. Rev. Lett. **111,** 121802 (2013).
[17] J. Erdmann, S. Guindon, K. Kroeninger, B. Lemmer, O. Nackenhorst, A. Quadt, and P. Stolte, Nucl. Instrum. Methods Phys. Res., Sect. A **748,** 18 (2014).
[18] D. Gonçalves, K. Kong, and J. H. Kim, J. High Energy Phys. 06 (2018) 079.
[19] D. Gonçalves, J. H. Kim, K. Kong, and Y. Wu, J. High Energy Phys. 01 (2022) 158.
[20] R. K. Barman, D. Gonçalves, and F. Kling, Phys. Rev. D **105,** 035023 (2022).
[21] P. Jackson and C. Rogan, Phys. Rev. D **96,** 112007 (2017).
[22] J. H. Kim, K. Kong, K. T. Matchev, and M. Park, Phys. Rev. Lett. **122,** 091801 (2019).
[23] J. H. Kim, M. Kim, K. Kong, K. T. Matchev, and M. Park, J. High Energy Phys. 09 (2019) 047.
[24] W. S. Cho, J. S. Gainer, D. Kim, K. T. Matchev, F. Moortgat, L. Pape, and M. Park, J. High Energy Phys. 08 (2014) 070.
[25] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer, and T. Stelzer, J. High Energy Phys. 06 (2011) 128.
[26] ATLAS Collaboration, Technical Report No. CERN-LHCC-2017-021, No. ATLAS-TDR-030, CERN, Geneva, 2017.
[27] ATLAS Collaboration, Technical Report No. ATL-PHYS-PUB-2013-004, CERN, Geneva, 2013.
[28] G. Aad *et al.* (ATLAS Collaboration), Eur. Phys. J. C **74,** 3071 (2014).
[29] J. Neyman and E. S. Pearson, Phil. Trans. R. Soc. A **231,** 289 (1933).

[30] Y.-C. Chen (CDF, D0 Collaborations), in *Proceedings of the 43rd Rencontres de Moriond on Electroweak Interactions and Unified Theories* (2008), pp. 189–196.

[31] D. Kim, K. T. Matchev, F. Moortgat, and L. Pape, J. High Energy Phys. 08 (2017) 102.

[32] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics (Springer New York Inc., New York, NY, USA, 2001).

[33] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006).

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

[35] G. Karagiorgi, G. Kasieczka, S. Kravitz, B. Nachman, and D. Shih, arXiv:2112.03769.

[36] H. Alhazmi, C. Dong, L. Hung, J. H. Kim, K. Kong, and D. Shih, Resolving combinatorial ambiguities in dilepton $t\bar{t}$ event topologies with neural networks, https://github.com/ZhongtianD/dilepton_combinatorics.

[37] A. Hocker *et al.*, arXiv:physics/0703039.

[38] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, J. Mach. Learn. Res. **12,** 2825 (2011).

[39] T. Chen and C. Guestrin, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, New York, NY, USA, 2016), KDD '16, pp. 785–794, ISBN 978-1-4503-4232-2.

[40] F. Chollet *et al.*, Keras, https://keras.io (2015).

[41] S. Hochreiter and J. Schmidhuber, Neural Comput. **9,** 1735 (1997).

[42] R. T. de Lima, arXiv:2102.06128.

[43] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard *et al.*, in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)* (2016), pp. 265–283.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, in *Advances in Neural Information Processing Systems*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Curran Associates, Inc., 2017), Vol. 30.

[45] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, in *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett (Curran Associates, Inc., 2019), pp. 8024–8035.

[46] M. Erdmann, E. Geiser, Y. Rath, and M. Rieger, J. Instrum. **14,** P06006 (2019).

[47] B. Idaszek, bachelor thesis, RWTH Aachen, 2019.

[48] P. Jung, bachelor thesis, RWTH Aachen, 2019.

[49] P. Konar, K. Kong, K. T. Matchev, and M. Park, J. High Energy Phys. 06 (2011) 041.

[50] P. Konar, K. Kong, and K. T. Matchev, J. High Energy Phys. 03 (2009) 085.

[51] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson, Eur. Phys. J. C **76,** 235 (2016).

[52] D. Kim, K. Kong, K. T. Matchev, M. Park, and P. Shyamsundar, arXiv:2105.10126.

[53] M. Fabbrichesi, R. Floreanini, and G. Panizzo, Phys. Rev. Lett. **127,** 161801 (2021).

[54] C. Severi, C. D. E. Boschi, F. Maltoni, and M. Sioli, Eur. Phys. J. C **82,** 285 (2022).

[55] Z. Dong, D. Gonçalves, K. Kong, and A. Navarro Serratos (to be published).

[56] J. Alwall, K. Hiramatsu, M. M. Nojiri, and Y. Shimizu, Phys. Rev. Lett. **103,** 151802 (2009).

[57] D. Kim and K. Kong, Phys. Lett. B **751,** 512 (2015).

[58] J. Erdmann, S. Guindon, K. Kroeninger, B. Lemmer, O. Nackenhorst, A. Quadt, and P. Stolte, (2014), https://github.com/KLFitter/KLFitter.

[59] C. G. Lester and D. J. Summers, Phys. Lett. B **463,** 99 (1999).

[60] A. Barr, C. Lester, and P. Stephens, J. Phys. G **29,** 2343 (2003).

[61] M. Burns, K. Kong, K. T. Matchev, and M. Park, J. High Energy Phys. 03 (2009) 143.

[62] P. Konar, K. Kong, K. T. Matchev, and M. Park, J. High Energy Phys. 04 (2010) 086.

[63] G. G. Ross and M. Serna, Phys. Lett. B **665,** 212 (2008).

[64] W. S. Cho, J. S. Gainer, D. Kim, S. H. Lim, K. T. Matchev, F. Moortgat, L. Pape, and M. Park, J. High Energy Phys. 01 (2016) 026.

[65] C. Rogan, Restframes, http://restframes.com.

[66] P. Artoisenet, V. Lemaitre, F. Maltoni, and O. Mattelaer, J. High Energy Phys. 12 (2010) 068.

[67] P. Artoisenet and O. Mattelaer, *Proc. Sci.*, CHARGED2008 (2008) 025.

[68] E. Akyazi, Matrix element method & madweight, https://indico.cern.ch/event/68178/attachments/1023900/1457817/memmw.pdf.

[69] V. M. Abazov *et al.* (D0 Collaboration), Phys. Rev. D **84,** 032004 (2011).

[70] A. Buckley, J. Ferrando, S. Lloyd, K. Nordström, B. Page, M. Rüfenacht, M. Schönherr, and G. Watt, Eur. Phys. J. C **75,** 032004 (2015).

[71] J. Alwall, A. Freitas, and O. Mattelaer, Phys. Rev. D **83,** 074010 (2011).

[72] J. M. Campbell, R. K. Ellis, W. T. Giele, and C. Williams, Phys. Rev. D **87,** 073005 (2013).

[73] S. Brochet, C. Delaere, B. François, V. Lemaître, A. Mertens, A. Saggio, M. Vidal Marono, and S. Wertz, Eur. Phys. J. C **79,** 126 (2019).

[74] B. A. Betchart, R. Demina, and A. Harel, Nucl. Instrum. Methods Phys. Res., Sect. A **736,** 169 (2014).

[75] L. Sonnenschein, Phys. Rev. D **72,** 095020 (2005).

[76] L. Sonnenschein, Phys. Rev. D **73,** 054015 (2006); **78,** 079902(E) (2008).

[77] R. H. Dalitz and G. R. Goldstein, Phys. Rev. D **45,** 1531 (1992).

[78] K. T. Matchev and P. Shyamsundar, J. High Energy Phys. 04 (2020) 027.

[79] I.-W. Kim, Phys. Rev. Lett. **104,** 081601 (2010).

[80] C. B. Park, J. High Energy Phys. 11 (2021) 042.

[81] A. Rujula and A. Galindo, J. High Energy Phys. 08 (2011) 023.

[82] C. B. Park, J. High Energy Phys. 07 (2020) 089.

[83] A. De Rujula and A. Galindo, J. High Energy Phys. 06 (2012) 091.