Published for SISSA by Springer

Received: *April 25, 2019*
Revised: *August 17, 2019*
Accepted: *August 26, 2019*
Published: *September 6, 2019*

Portraying double Higgs at the Large Hadron Collider

Jeong Han Kim,[a] Minho Kim,[b,c] Kyoungchul Kong,[a] Konstantin T. Matchev[d] and Myeonghun Park[c]

[a] *Department of Physics and Astronomy, University of Kansas,*
Lawrence, KS 66045, U.S.A.

[b] *Department of Physics, POSTECH,*
77 Cheongam-ro, Nam-gu, Pohang, 37673, Korea

[c] *Institute of Convergence Fundamental Studies and School of Liberal Arts,*
Seoultech, 232 Gongneung-ro, Nowon-gu, Seoul, 01811, Korea

[d] *Institute for Fundamental Theory, Physics Department,*
University of Florida, Gainesville, FL 32611, U.S.A.

E-mail: jeonghan.kim@ku.edu, kmhmon@postech.ac.kr, kckong@ku.edu, matchev@phys.ufl.edu, parc.seoultech@seoultech.ac.kr

Abstract: We examine the discovery potential for double Higgs production at the high luminosity LHC in the final state with two b-tagged jets, two leptons and missing transverse momentum. Although this dilepton final state has been considered a difficult channel due to the large backgrounds, we argue that it is possible to obtain sizable signal significance, by adopting a deep learning framework making full use of the relevant kinematics along with the jet images from the Higgs decay. For the relevant number of signal events we obtain a substantial increase in signal sensitivity over existing analyses. We discuss relative improvements at each stage and the correlations among the different input variables for the neutral network. The proposed method can be easily generalized to the semi-leptonic channel of double Higgs production, as well as to other processes with similar final states.

Keywords: Beyond Standard Model, Higgs Physics

ArXiv ePrint: 1904.08549

https://doi.org/10.1007/JHEP09(2019)047

Contents

1 Introduction

The discovery of the Higgs boson [1, 2] jumpstarted the comprehensive program of precision measurements of all Higgs couplings. While the Higgs boson couplings to fermions and gauge bosons are in good agreement with the Standard Model (SM) predictions [3], the Higgs self-couplings are difficult to measure experimentally [4–12]. Yet, the knowledge of those couplings is crucial for understanding the exact mechanism of electroweak symmetry breaking and the origin of mass in our universe. It is also a guaranteed physics target which can be probed at the upgraded Large Hadron Collider (LHC) or at future colliders. The resulting experimental constraints on the Higgs self-couplings will have an immediate and long-lasting impact on model-building efforts beyond the SM.

We parameterize the Higgs self-interaction as follows:

$$V = \frac{m_h^2}{2}h^2 + \kappa_3 \lambda_3^{\rm SM} v h^3 + \frac{1}{4}\kappa_4 \lambda_4^{\rm SM} h^4 \,, \tag{1.1}$$

where m_h is the mass of the SM Higgs boson (h), $v \approx 256\,\text{GeV}$ is the Higgs vacuum expectation value,

$$\lambda_3^{\rm SM} = \lambda_4^{\rm SM} = \frac{m_h^2}{2v^2}$$

are the SM values for the Higgs self-couplings, while κ_3 and κ_4 parametrize the corresponding deviations from them. In order to access κ_3 (κ_4), one has to measure the process of double (triple) Higgs boson production at the LHC, possibly with high luminosity (HL), or at future colliders.

Double Higgs (hh) production has been studied in many channels, including $b\bar{b}b\bar{b}$ [13–18], $b\bar{b}\gamma\gamma$ [6, 10, 19–31], $b\bar{b}\tau\tau$ [6, 21, 32–36], $b\bar{b}W^+W^-/ZZ$ [8, 9, 37–41], $W^+W^-W^+W^-$ [42], etc. Among the different possible final states, here we focus on hh production at the HL-LHC in the final state with two b-tagged jets, two leptons and missing transverse momentum. The signal process is $(h \to b\bar{b})(h \to W^\pm W^{*\mp} \to \ell^+\nu_\ell \ell'^- \bar{\nu}_{\ell'})$ and it suffers from large SM backgrounds, primarily due to top quark pair production ($t\bar{t}$). The few existing studies in this channel therefore employ sophisticated algorithms (neutral network (NN) [8], deep neutral network (DNN) [11], boosted decision tree (BDT) [9, 43], etc.) to increase the signal sensitivity, but show somewhat pessimistic results, with a significance no better than 1σ at the HL-LHC with 3 ab^{-1} luminosity.

The recent study in ref. [39] introduced some new ideas for reducing the SM backgrounds in this channel. For example, the new variables *Topness* and *Higgsness* were designed to test whether the event kinematics is consistent with $t\bar{t}$ or hh, respectively. The use of Topness and Higgsness already effectively reduced the $t\bar{t}$ background to a manageable level, and additional variables were then employed to handle the remaining SM background processes — e.g., the subsystem variable $M_{T2}^{(\ell)}$ is effective in eliminating background arising from τ decays. In this paper, we supplement the novel kinematic method from ref. [39] with the analysis of the jet image in the $h \to b\bar{b}$ decay, where the basic idea is to treat the detector as a camera and the streams of jets as an image [44–51]. In our case, the collimated nature of the Higgs decay will hopefully differ from the patterns obtained in SM production processes. In addition, we adopt a deep learning framework in our main analysis, since it is known that modern deep learning algorithms trained on jet images provide improved signal-to-background discrimination [45–50, 52–54].

The analysis presented in this paper contains a number of improvements in comparison to previous studies:

- Unlike the customized detector simulation performed in ref. [39], here we employ Delphes [55] to simulate detector effects such as detector resolution, reconstruction efficiency, etc., and Fastjet [56] for jet-reconstruction.

- We use deep learning framework to optimize the cuts, which further increases the significance compared to the conventional cut-and-count as performed in ref. [39].

- We exploit an enlarged set of relevant variables which consists of the 10 variables originally considered in ref. [43]: $p_{T\ell_1}$, $p_{T\ell_2}$, $\not{p}_T$, $m_{\ell\ell}$, m_{bb}, $\Delta R_{\ell\ell}$, ΔR_{bb}, p_{Tbb}, $p_{T\ell\ell}$, and $\Delta\phi_{bb,\ell\ell}$, supplemented with the six recent variables from ref. [39]: Topness, Higgsness, $M_{T2}^{(b)}$, $M_{T2}^{(\ell)}$, $\hat{s}_{\rm min}^{(\ell\ell)}$ and $\hat{s}_{\rm min}^{(bb\ell\ell)}$.

- We include a SM background process, tW production, which was missing from all previous discussions of this channel, yet it turns out to be the next dominant background once the $t\bar{t}$ background is under control.

- The fact that the Higgs boson h is a color-singlet allows us to use the jet image of the $h \to b\bar{b}$ decay for further background suppression [45–48, 50].

- We examine the effect of pile-up, which was missing from previous studies. The expected average number of pile-up $\langle \mu \rangle$ at the HL-LHC is $\mathcal{O}(200)$ collisions per bunch crossing [57]. Thus for any precision measurements, it is crucial to have a strategy in place to ensure that pile-up effects do not jeopardize the analysis. Here we choose to apply the Soft Drop algorithm [58] for QCD analyses, which is a powerful pile-up mitigation technique. In order to reduce pile-up effects on the relevant kinematic variables, we adopt the definition for a missing transverse momentum from ATLAS, which excludes contributions from soft neutral particles [59].

Our results show that the dominant $t\bar{t}$ background can be significantly reduced until it is comparable to the other subdominant backgrounds, i.e., after all cuts, we find that all SM backgrounds contribute at similar levels. This reduction can be accomplished without sacrificing too much of the signal rate, which leads to an improved signal significance. Our study indicates that the dilepton channel from $hh \to b\bar{b}W^+W^-$ could contribute to the combined significance for hh discovery on par with the other final states, making double Higgs production sooner accessible at the HL-LHC.

This paper is structured as follows. We begin our discussion of the SM backgrounds and present the details of our simulation in section 2. In the following two sections 3 and 4, we provide some basic information on the kinematic variables used later in the analysis and on jet images, respectively. Then in section 5 we discuss how we set up our analysis in a deep learning framework. Section 6 presents our results, while section 7 is reserved for the discussion and conclusions. We include a brief review on deep neural networks in appendix A.

2 Event generation and detector simulation

Parton-level signal and background events were generated using MadGraph5_aMC@NLO v2.6 [60] with the default NNPDF2.3QED parton distribution functions [61] at leading order QCD accuracy at the $\sqrt{s} = 14\,\text{TeV}$ LHC. The default dynamical renormalization and factorization scales were used. We assume 3ab^{-1} of luminosity throughout this paper. Parton-level events were generated with the following cuts: $p_{Tj} > 20\,\text{GeV}$, $p_{Tb} > 20\,\text{GeV}$, $p_{T\gamma} > 10\,\text{GeV}$, $p_{T\ell} > 10\,\text{GeV}$, $\eta_j < 5$, $\eta_b < 5$, $\eta_\gamma < 2.5$, $\eta_\ell < 2.5$, $\Delta R_{bb} < 1.8$, $\Delta R_{\ell\ell} < 1.3$, $70\,\text{GeV} < m_{jj}, m_{bb} < 160\,\text{GeV}$ and $m_{\ell\ell} < 75\,\text{GeV}$. For $jj\ell\ell\nu\bar{\nu}$, $\ell\ell bj$ and $tW + j$ backgrounds, we impose $5\,\text{GeV} < m_{\ell\ell} < 75\,\text{GeV}$ additionally. Here the angular distance ΔR_{ij} is defined by

$$\Delta R_{ij} = \sqrt{(\Delta\phi_{ij})^2 + (\Delta\eta_{ij})^2}, \tag{2.1}$$

where $\Delta\phi_{ij} = \phi_i - \phi_j$ and $\Delta\eta_{ij} = \eta_i - \eta_j$ are respectively the differences of the azimuthal angles and rapidities between particles i and j.

The double Higgs production cross-section is normalized to $\sigma_{hh} = 40.7\,\text{fb}$, the next-to-next-to-leading order (NNLO) accuracy in QCD [62]. Considering all relevant branching fractions, we obtain signal cross section $\sigma_{hh} \cdot 2 \cdot \mathrm{BR}(h \to b\bar{b}) \cdot \mathrm{BR}(h \to WW^* \to \ell^+\ell^-\nu\bar{\nu}) =$ 0.648 fb, where ℓ denotes an electron or a muon, including leptons from tau decays. The major background is $t\bar{t}$ production, whose cross section is normalized to the NNLO QCD cross-section 953.6 pb [63]. Another important background is $t\bar{t}h$, which is normalized to the next-to-leading order (NLO) QCD cross-section of 611.3 fb [64]. For the $t\bar{t}V$ ($V = W^{\pm}, Z$) background, we apply an NLO k-factor of 1.54, resulting in a cross-section of 1.71 pb [65]. We apply an NLO k-factor of 1.0 for the Drell-Yan type backgrounds $\ell\ell bj$ and $\tau\tau bb$, where j denotes partons in the five-flavor scheme. Note that a recent study indicates that $\mathrm{k}^{NNLO,DY}_{QCD\otimes QED} \approx 1$ [66]. The irreducible $jj\ell\ell\nu\nu$ background from the mixed QCD+EW process is included with $\mathrm{k}_{NLO} = 2$. Finally, we generate $tW + j$ events with up to one additional matched jet (in the five-flavor scheme), whose cross-section turns out to be 0.51 pb (after the cuts) including all relevant branching fractions. As we try to reconstruct events, off-shell effects for the top quark and W boson need to be taken care of properly. We generate parton level events with MadGraph5, which includes the proper treatment of the off-shell effects for the top quark and the W boson for both signal and all backgrounds.

Events are further processed for parton-shower/hadronization using Pythia8235 [67]. We use Delphes 3.4.1 [55] for simulating the detector effects and FastJet 3.3.1 [56] for jet-reconstruction, with modified ATLAS settings as follows.

- Jets are clustered with the anti-k_T algorithm [68] with cone-size $\Delta R = 0.4$, where ΔR is the distance (2.1) in the $(\phi,\, \eta)$ space. For the analysis, we consider jets with $p_{Tj} > 20\,\mathrm{GeV}$ and $|\eta_j| < 2.5$.

- We use the a flat b-tagging efficiency, $\epsilon_{b\to b} = 0.75$, and flat mis-tagging rates for non-b jets of $\epsilon_{c\to b} = 0.1$ and $\epsilon_{j\to b} = 0.01$ [57].

- For lepton isolation, we require $\frac{p_{T\ell}}{p_{T\ell}+\sum_i p_{Ti}} > 0.7$, where the sum is taken over the transverse momenta p_{Ti} of all final states particles i, $i \neq \ell$, with $p_{Ti} > 0.5\,\mathrm{GeV}$ and within $\Delta R_{i\ell} < 0.3$ of the lepton candidate ℓ. Leptons are also required to have $p_{T\ell} > 10\,\mathrm{GeV}$ and $|\eta_\ell| < 2.5$.

- For photon isolation, we analogously require $\frac{\sum_i p_{Ti}}{p_{T\gamma}} < 0.12$ for particles within $\Delta R_{i\gamma} < 0.3$ of the photon candidate γ. Photons are also required to have $p_{T\gamma} > 25\,\mathrm{GeV}$ and $|\eta_\gamma| < 2.5$.

- The missing transverse momentum $\vec{\not{p}}_T$ is defined as the negative vector sum of the transverse momenta of the accepted leptons, photons, jets and soft tracks as follows [59];

$$\vec{\not{p}}_T = -\left(\sum \vec{p}_{T\ell} + \sum \vec{p}_{T\gamma} + \sum \vec{p}_{Tj} + \sum \vec{p}_{T(\mathrm{track})}\right). \tag{2.2}$$

  Here the last term is added to consider unused soft tracks. These tracks are required to have $p_T > 0.4\,\mathrm{GeV}$, $|\eta| < 2.5$ and transverse (longitudinal) impact parameter $|d_0| < 1.5\,\mathrm{mm}$ ($|z_0 \sin\theta| < 1.5\,\mathrm{mm}$). To reduce effects from pile-up, we only use particles which have track information.

After particle reconstruction, we employ the following *baseline selection cuts*[1] from ref. [39]:

- the two leading jets must be b-tagged, each with $p_T > 30\,\text{GeV}$,
- exactly two isolated leptons of opposite sign, each with $p_{T\ell} > 20\,\text{GeV}$,
- $\not{p}_T = |\vec{\not{p}}_T| > 20\,\text{GeV}$ for the reconstructed missing transverse momentum,
- proximity cut of $\Delta R_{\ell\ell} < 1.0$ for the two leptons,
- proximity cut of $\Delta R_{bb} < 1.3$ for the two b-tagged jets,
- $m_{\ell\ell} < 65\,\text{GeV}$ for the two leptons,
- $95\,\text{GeV} < m_{bb} < 140\,\text{GeV}$ for the two b-tagged jets.

For those events which passed the baseline cuts, we form 16 kinematic variables, as well as jet images. As we will see later, the jet images can capture additional features which are not already contained in the 16 standard kinematic variables. Therefore one can obtain better performance by combining kinematics and jet images, which is one of the main ideas of this paper.

3 Kinematics in signal and backgrounds

In this section we introduce the 16 kinematic variables used in this analysis. Their kinematic distributions (for signal and all relevant backgrounds) are shown in figure 1 and will be discussed shortly.

We begin with ten standard kinematic variables, which were previously considered in refs. [9, 43] (their distributions are shown in the first ten panels of figure 1):

- m_{bb}, the invariant mass of the two b-tagged jets (1st plot in the 1st row). This is expected to be a good variable, since for signal events, the two b-jets originate from the decay of a narrow resonance (the Higgs boson) and would therefore reconstruct to the Higgs mass, up to resolution effects: $m_{bb} \sim m_h$. This justifies the baseline cut of $95\,\text{GeV} < m_{bb} < 140\,\text{GeV}$, as indicated with the vertical dotted lines. In contrast, no such correlations exists for backgrounds events: the two b-jets either originate from different decay chains and are uncorrelated (as in the case of $t\bar{t}$, for example), or they reconstruct to the mass of a Z-boson or an off-shell gluon, with a mass lower than m_h. The plot in figure 1, while confirming those expectations, also shows that the total background happens to peak at a value of m_{bb} which, unfortunately, is not too far away from m_h, providing the motivation to explore other variables.

[1] For the motivation behind these cuts, see figure 1 (in which the cut values are indicated with vertical dotted lines) and the related discussion in section 3 below.

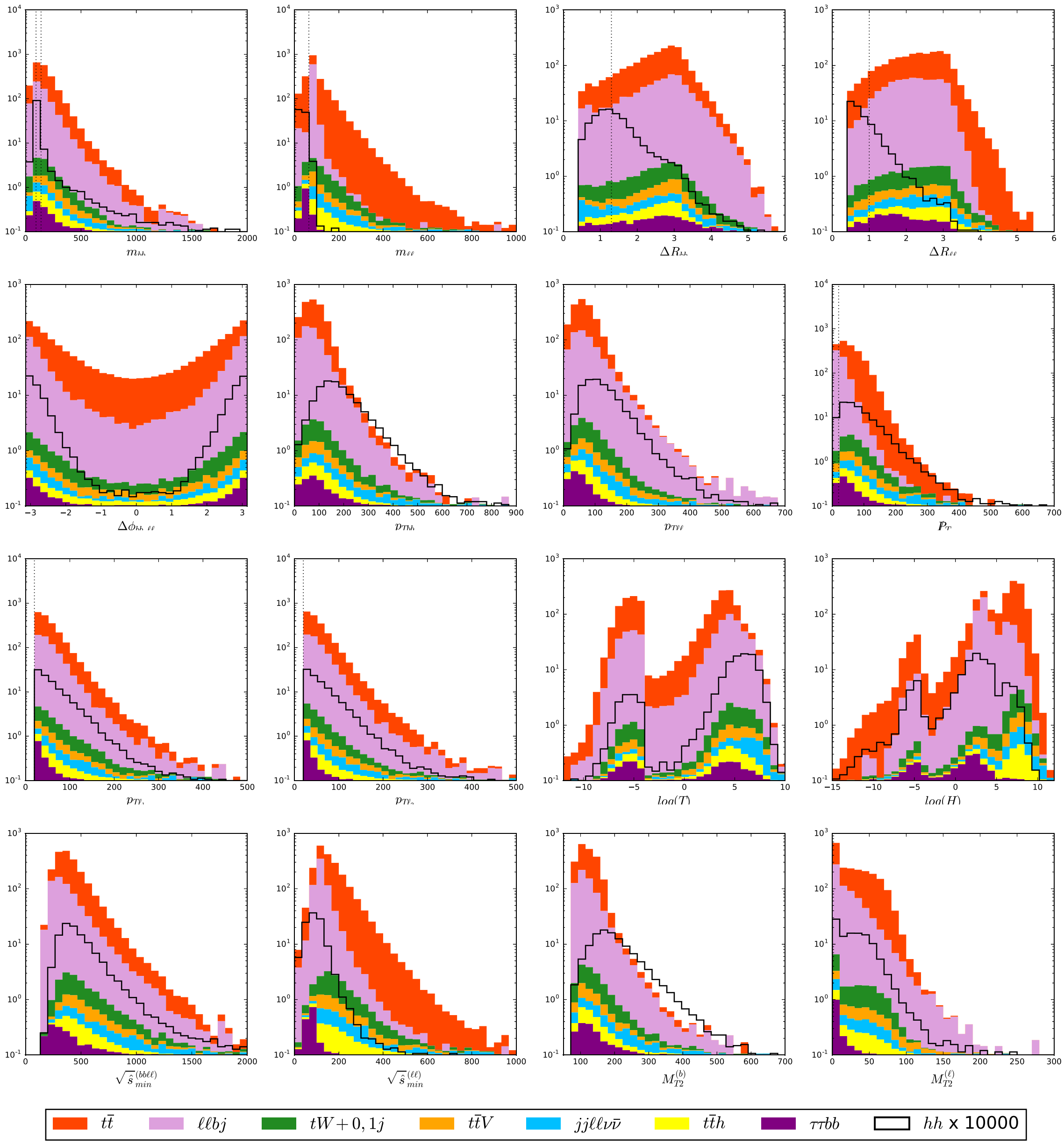


Figure 1. Distributions of the 16 kinematic variables for signal (hh) and different types of backgrounds ($t\bar{t}$, $t\bar{W}$, $t\bar{t}V$, $t\bar{t}h$, $\tau\tau bb$, $\ell\ell bj$ and $jj\ell\ell\nu\nu$) before baseline cuts. The y-axis represents the number of events for each process and all individual distributions are normalized properly according to their respective cross-sections assuming 3 ab^{-1} at the 14 TeV LHC. The dotted vertical lines indicate the baseline cuts introduced in section 2.

- $m_{\ell\ell}$, the invariant mass of the two leptons (2nd plot in the 1st row). For the case of the signal, the two leptons ultimately originate from the Higgs boson decay, and therefore their invariant mass $m_{\ell\ell}$ is bounded from above, hence the baseline cut of $m_{\ell\ell} < 65\,\text{GeV}$. Note that the $m_{\ell\ell}$ distribution (which is observable) should be the same as the distribution of $m_{\nu\nu}$ (which is unobservable).

- ΔR_{bb}, the angular separation (2.1) between the two b-tagged jets (3rd plot in the 1st row). Given the relatively low Higgs mass, the two Higgs particles in hh production have sizable transverse momentum and their respective decay products (e.g., the two b-quarks) tend to go in the same direction. This and the next four variables try to exploit this kinematic property of the signal. For example, the Higgs boost implies that ΔR_{bb} is relatively small for signal events, and this motivates the proximity cut of $\Delta R_{bb} < 1.3$.

- $\Delta R_{\ell\ell}$, the angular separation (2.1) between the two leptons (4th plot in the 1st row). Here the same arguments apply as in the case of ΔR_{bb} just discussed. The corresponding plot in figure 1 confirms that the signal $\Delta R_{\ell\ell}$ distribution peaks well below most of the background processes, prompting the baseline cut of $\Delta R_{\ell\ell} < 1.0$.

- $\Delta\phi_{bb,\ell\ell}$, the azimuthal angle in the transverse plane between the two b-jet system and the two lepton system (1st plot in the 2nd row). This is yet another way to capture the back-to-back boost of the two Higgs bosons in double Higgs production. Figure 1 shows that the signal peaks at $\Delta\phi_{bb,\ell\ell} = \pm\pi$ more sharply than the background, which could be exploited later in the neural network analysis. However, no baseline cut was applied in this case, since $\Delta\phi_{bb,\ell\ell}$ is expected to be largely correlated with ΔR_{bb} and $\Delta R_{\ell\ell}$.

- p_{Tbb}, the transverse momentum of the two b-jet system (2nd plot in the 2nd row). Like the previous three variables, this variable is motivated by the significant boost of the Higgs bosons in the signal, but no baseline cut was applied.

- $p_{T\ell\ell}$, the transverse momentum of the two lepton system (3rd plot in the 2nd row). This variable behaves similarly to p_{Tbb}, but to a lesser extent, since the two leptons come from separate Ws, while the two b-quarks are direct decay products of the Higgs boson.

- $\not{p}_T = |\vec{\not{p}}_T|$, the magnitude of the missing transverse momentum (4th plot in the 2nd row). A $\not{p}_T$ cut is routinely applied in order to fight the QCD backgrounds (not shown in figure 1). Following ref. [9], here we use a baseline cut of $\not{p}_T > 20\,\text{GeV}$.

- $p_{T\ell_1}$, the transverse momentum of the hardest lepton (1st plot in the 3rd row).

- $p_{T\ell_2}$, transverse momentum of the next-hardest lepton (2nd plot in the 3rd row). As shown in figure 1, the individual transverse momenta of the two leptons are similar for both signal and backgrounds. Therefore, the lepton p_T's may be good for triggering purposes, but not for background rejection.

We note that for the signal, many of these 10 variables are strongly correlated to each other.[2] This implies that cutting on one variable significantly reduces the power of other

[2] The strong correlation arises due to the very nature of double Higgs production — the two Higgs particles are produced with a sizable transverse momentum, which restricts the kinematics of their decay products.

variables. At the same time, while these 10 variables are among the most commonly used in high energy physics, it is not guaranteed that they fully capture all kinematic differences between signal and background. This is why we introduce six additional variables [39]: Topness, Higgsness, $\sqrt{\hat{s}}_{\rm min}^{(bb\ell\ell)}$, $\sqrt{\hat{s}}_{\rm min}^{(\ell\ell)}$, $M_{T2}^{(b)}$ and $M_{T2}^{(\ell)}$, shown in the last six panels of figure 1, which are meant to take full advantage of the kinematic differences between the signal and background event topologies.

The Topness variable measures the degree of consistency of a given event with the kinematics of dilepton $t\bar{t}$ production, where there are 6 unknowns (the three-momenta of the two neutrinos, $\vec{p}_\nu$ and $\vec{p}_{\bar{\nu}}$) and four on-shell constraints, m_t, $m_{\bar{t}}$, m_{W^+} and m_{W^-}. Here $m_t = m_{\bar{t}}$ is the mass of top or antitop quark, and $m_{W^\pm} = m_W$ is the mass of the W boson. Then the neutrino momenta can be fixed by minimizing the following quantity

$$\chi^2_{ij} \equiv \min_{\vec{\not{P}}_T = \vec{p}_{T\nu} + \vec{p}_{T\bar{\nu}}} \left[\frac{\left(m^2_{b_i \ell^+ \nu} - m_t^2\right)^2}{\sigma_t^4} + \frac{\left(m^2_{\ell^+ \nu} - m_W^2\right)^2}{\sigma_W^4} + \frac{\left(m^2_{b_j \ell^- \bar{\nu}} - m_t^2\right)^2}{\sigma_t^4} + \frac{\left(m^2_{\ell^- \bar{\nu}} - m_W^2\right)^2}{\sigma_W^4} \right], \tag{3.1}$$

subject to the missing transverse momentum constraint, $\vec{\not{P}}_T = \vec{p}_{T\nu} + \vec{p}_{T\bar{\nu}}$. The parameters σ_t and σ_W are indicative of the corresponding experimental resolutions and intrinsic particle widths. In principle, they can be treated as free parameters and one can tune them using NN, BDT, etc. In our numerical study, we shall use $\sigma_t = 5\,\mathrm{GeV}$ and $\sigma_W = 5\,\mathrm{GeV}$. Since there is a twofold ambiguity in the paring of a b-quark and a lepton, Topness is defined as the smaller of the two χ^2s [39],

$$T \equiv \min\left(\chi^2_{12}\,,\, \chi^2_{21}\right)\,. \tag{3.2}$$

The Topness distributions for both signal and backgrounds before baseline cuts are shown in figure 1 (3rd plot in the 3rd row). We observe that, as expected, T tends to have smaller values for the main background ($t\bar{t}$) than for signal.

In our signal of hh production, the two b-quarks arise from a Higgs decay ($h \to b\bar{b}$), and therefore their invariant mass m_{bb} can be used as a first cut to enhance the signal sensitivity. For the decay of the other Higgs boson, $h \to W^\pm W^{*\mp}$, Higgsness is defined as follows [39]

$$H \equiv \min_{\vec{\not{P}}_T = \vec{p}_{T\nu} + \vec{p}_{T\bar{\nu}}} \left[\frac{\left(m^2_{\ell^+\ell^-\nu\bar{\nu}} - m_h^2\right)^2}{\sigma^4_{h_\ell}} + \frac{\left(m^2_{\nu\bar{\nu}} - (m^{\rm peak}_{\nu\bar{\nu}})^2\right)^2}{\sigma^4_\nu} + \min\left(\frac{\left(m^2_{\ell^+\nu} - m_W^2\right)^2}{\sigma_W^4} + \frac{\left(m^2_{\ell^-\bar{\nu}} - (m^{\rm peak}_{W^*})^2\right)^2}{\sigma^4_{W^*}}\,, \frac{\left(m^2_{\ell^-\bar{\nu}} - m_W^2\right)^2}{\sigma_W^4} + \frac{\left(m^2_{\ell^+\nu} - (m^{\rm peak}_{W^*})^2\right)^2}{\sigma^4_{W^*}} \right) \right]. \tag{3.3}$$

It tests whether the neutrino kinematics can be compatible with having the Higgs boson and one of the W-bosons on-shell, while at the same time being consistent with the invariant

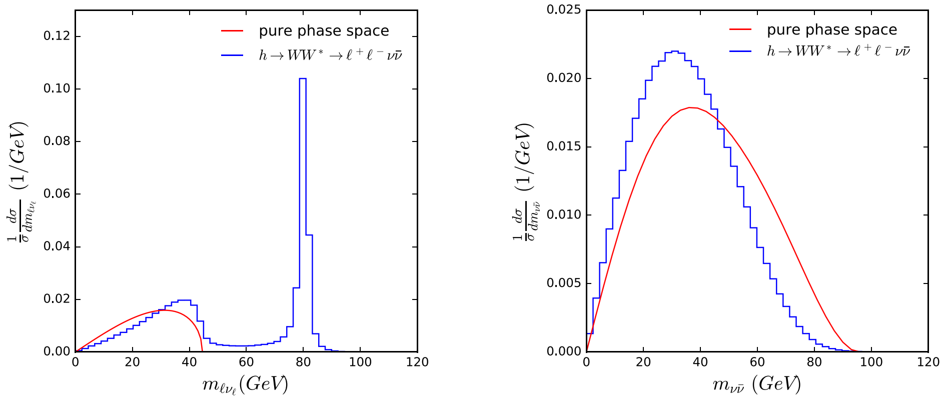


Figure 2. Unit-normalized invariant mass distribution of the the lepton-neutrino ($m_{\ell\nu}$, left) and the two neutrinos ($m_{\nu\bar{\nu}}$, right).

mass distributions expected for the off-shell W-boson, W^*, and the neutrino pair, $\nu\bar{\nu}$. The invariant mass m_{W^*} is bounded by $0 \le m_{W^*} \le m_h - m_W$ and the peak of its distribution is at

$$m_{W^*}^{\text{peak}} = \frac{1}{\sqrt{3}}\sqrt{2\left(m_h^2 + m_W^2\right) - \sqrt{m_h^4 + 14 m_h^2 m_W^2 + m_W^4}}\,. \tag{3.4}$$

The left panel of figure 2 shows the unit-normalized invariant mass distribution of the proper lepton-neutrino system ($m_{\ell\nu}$). The distribution has a bimodal shape — the narrow peak on the right near 80 GeV corresponds to the on-shell W-boson resonance, while the broader hump to the left is due to the off-shell W^*, with a clear end-point at $m_h - m_W = 45\,\text{GeV}$ and a maximum near $m_{W^*}^{\text{peak}} = 40\,\text{GeV}$ in accordance with (3.4).

The definition of Higgsness (3.3) also includes a term which tests for consistency with the expected invariant mass distribution $\frac{\mathrm{d}\sigma}{\mathrm{d}m_{\nu\bar{\nu}}}$ for the neutrino pair,[3] which is shown in the right panel of figure 2. The red solid curve gives the pure phase space prediction

$$\frac{\mathrm{d}\sigma}{\mathrm{d}m_{\nu\bar{\nu}}} \propto \int \mathrm{d}m_{W^*}^2 \lambda^{1/2}(m_h^2, m_W^2, m_{W^*}^2) f(m_{\nu\bar{\nu}})\,, \tag{3.5}$$

where $\lambda(x, y, z) = x^2 + y^2 + z^2 - 2xy - 2yz - 2zx$ is the two-body phase space function and $f(m)$ is the invariant mass distribution of the antler topology with $h \to WW^* \to \ell^+\ell^-\nu\bar{\nu}$:

$$f(m) \sim \begin{cases} \eta\, m\,, & 0 \le m \le e^{-\eta}E, \\ m \ln(E/m)\,, & e^{-\eta}E \le m \le E, \end{cases} \tag{3.6}$$

where the endpoint E and the parameter η are defined in terms of the particle masses as

$$E = \sqrt{m_W m_{W^*}\, e^{\eta}}\,, \tag{3.7}$$

$$\cosh\eta = \left(\frac{m_h^2 - m_W^2 - m_{W^*}^2}{2 m_W m_{W^*}}\right)\,. \tag{3.8}$$

[3] In the limit of massless leptons, the distribution $\frac{\mathrm{d}\sigma}{\mathrm{d}m_{\nu\bar{\nu}}}$ is the same as the dilepton mass distribution $\frac{\mathrm{d}\sigma}{\mathrm{d}m_{\ell^+\ell^-}}$, which is directly observable and therefore more commonly discussed in the literature [69–72].

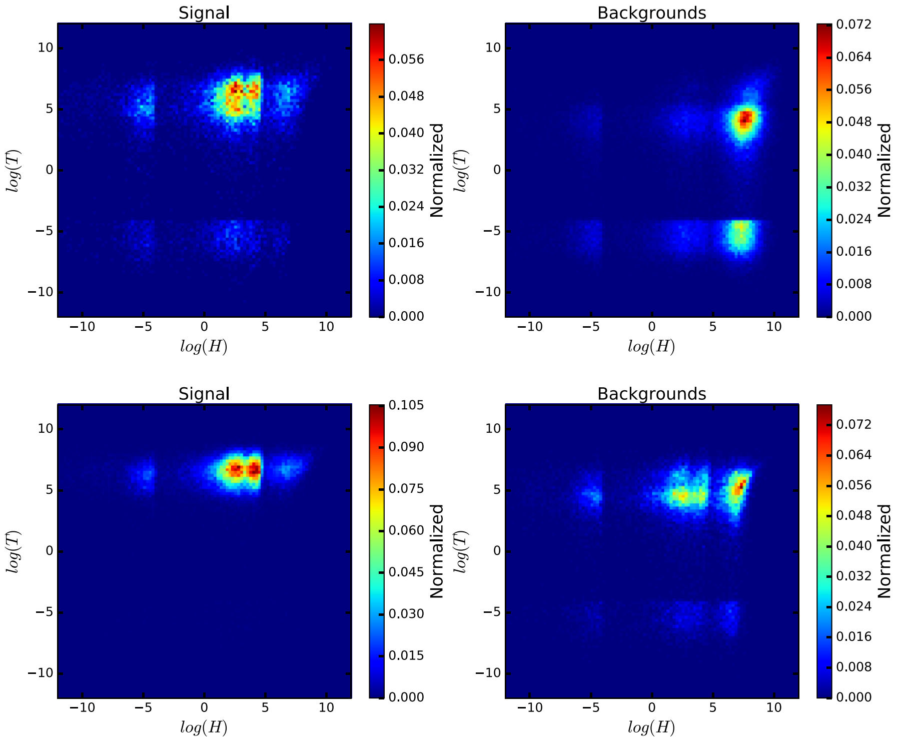


Figure 3. Two-dimensional correlation plots for Higgsness and Topness for signal (left) and backgrounds (right) before (top) and after (bottom) baseline cuts.

Note that by allowing one of the W-bosons to be on-shell, eqs. (3.5)–(3.8) generalize the results previously derived in refs. [69–72] for the purely on-shell case. The blue histogram in the right panel of figure 2 shows the actual $m_{\nu\bar{\nu}}$ distribution, whose shape is slightly different from the pure phase space result (3.5), due a helicity suppression in the W-ℓ-ν vertex. In particular, we observe that the actual peak is at $m_{\nu\bar{\nu}}^{\text{peak}} \approx 30\,\text{GeV}$, which is the value that we shall use in the definition of Higgsness (3.3).[4]

The definition of Higgsness (3.3) contains some additional resolution parameters: σ_h for the reconstructed mass of the Higgs boson, σ_{W^*} for the reconstructed mass of the off-shell W boson, and σ_ν for the $m_{\nu\bar{\nu}}$ resolution. In what follows, we shall take $\sigma_{W^*} = 5\,\text{GeV}$, $\sigma_{h_\ell} = 2\,\text{GeV}$, and $\sigma_\nu = 10\,\text{GeV}$.[5]

The Higgsness distributions for both signal and backgrounds before baseline cuts are shown in figure 1 (4th plot in the 3rd line). The two dimensional map of (Higgsness, Topness) on a log-log scale is depicted in figure 3. The Higgsness and Topness distributions in figure 1 are projections of this two dimensional scatter plot onto the x-axis and y-axis, respectively. Although the signal and the backgrounds do not exhibit a very clean

[4] We note that other variants of Higgsness are also possible — for example, instead of penalizing the function H by the distances to the peaks in the corresponding distributions, one can introduce penalty terms which take advantage of the knowledge of the exact probability distributions (the blue histograms in figure 2).

[5] We have checked that our results are not very sensitive to these choices.

separation in the individual one-dimensional projections in figure 1, their two dimensional correlation plots show some visible differences. We note that even after employing the baseline cuts, one can still see a difference in the two dimensional correlation of Higgsness and Topness (bottom row plots).

Along with Higgsness and Topness, we also consider two versions of the $\hat{s}_{\rm min}$ variable [73, 74], which is defined as

$$\hat{s}_{\rm min}^{(\rm v)} = m_{\rm v}^2 + 2\left(\sqrt{|\vec{P}_T^{\rm v}|^2 + m_{\rm v}^2}\;|\vec{\not{P}}_T| - \vec{P}_T^{\rm v}\cdot\vec{\not{P}}_T\right), \tag{3.9}$$

where (v) represents a set of visible particles under consideration, while $m_{\rm v}$ and $\vec{P}_T^{\rm v}$ are their invariant mass and transverse momentum, respectively. The variable (3.9) characterizes the system comprising of the visible particles (v) and the invisible particles (here assumed to be massless) which are responsible for the measured missing transverse momentum $\vec{\not{P}}_T$. It provides the minimum value of the Mandelstam invariant mass $\hat{s}$ for the system which is consistent with the observed visible 4-momentum vector. We shall apply (3.9) to the whole event, where $\rm v = \{bb\ell\ell\}$, or to the subsystem resulting from the decay $h \to W^\pm W^{*\mp} \to \ell^+\ell^-\nu\bar{\nu}$, where $\rm v = \{\ell\ell\}$. The distributions of the resulting variables $\hat{s}_{\rm min}^{(bb\ell\ell)}$ and $\hat{s}_{\rm min}^{(\ell\ell)}$ are shown in the left two panels on the fourth row of figure 1. The $\hat{s}_{\rm min}^{(bb\ell\ell)}$ variable represents the minimum energy required to produce the two original parent particles (the two Higgs bosons in the case of the signal and the two top quarks in the case of the major $t\bar{t}$ background). This is why one would expect the distribution to peak around the parent mass threshold, $2m_h$ for the signal and $2m_t$ for the background [73]. However, the first panel in the fourth row of figure 1 shows that while the background $\hat{s}_{\rm min}^{(bb\ell\ell)}$ distribution peaks near $2m_t$, which is expected, the signal $\hat{s}_{\rm min}^{(bb\ell\ell)}$ distribution peaks around 400 GeV, which is substantially higher than $2m_h$. This implies that the two top quarks are produced more or less at rest, while the two Higgs bosons have a sizable boost. Similarly, the variable $\hat{s}_{\rm min}^{(\ell\ell)}$ is the minimum energy required to produce the two W bosons. For the $t\bar{t}$ background, where both W bosons are on-shell, the peak is expected to occur around $2m_W$. On the other hand, the signal distribution should be softer, since one of the W bosons is off-shell, and furthermore, the peak should be located slightly below the Higgs boson mass. These kinematic differences are illustrated in the second plot on the fourth row of figure 1, and motivate the use of $\hat{s}_{\rm min}^{(\ell\ell)}$ as an analysis variable.

The last two panels in the fourth row of figure 1 show distributions of the subsystem M_{T2} variable [75] — first when it is applied to the $b\bar{b}$ visible system resulting from the $t \to bW$ decays ($M_{T2}^{(b)}$), and then when it is applied to the $\ell^+\ell^-$ visible system resulting from the $W \to \ell\nu$ decays ($M_{T2}^{(\ell)}$). In principle, M_{T2} is defined as [76]

$$M_{T2}(\tilde{m}) \equiv \min\left\{\max\left[M_{TP_1}(\vec{p}_{T\nu},\tilde{m}),\; M_{TP_2}(\vec{p}_{T\bar{\nu}},\tilde{m})\right]\right\}, \tag{3.10}$$

where the minimization over the transverse masses of the parent particles M_{TP_i} ($i = 1, 2$) is performed over the transverse neutrino momenta $\vec{p}_{\nu T}$ and $\vec{p}_{\bar{\nu}T}$, subject to the $\vec{\not{P}}_T$ constraint.[6] The parameter $\tilde{m}$ in (3.10) is the test mass for the daughter particle: in

[6] See refs. [77–85] for more information and other variants of M_{T2}.

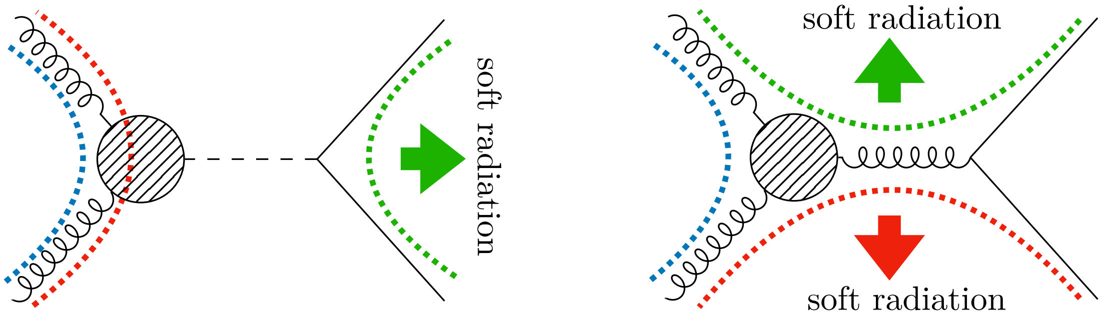


Figure 4. Color flow diagrams for a color-singlet particle (left) and a color-octet particle (right). The colored dotted lines represent QCD color-connection and arrows denote the direction of hadron concentration.

the case of $M_{T2}^{(\ell)}$ one should use $\tilde{m} = m_\nu = 0$, while in the case of $M_{T2}^{(b)}$, the daughter particles are the W bosons, and $\tilde{m} = m_W = 80\,\text{GeV}$, which leads to the lower bound $m_W \le M_{T2}^{(b)}$ visible in the plot. By construction, the M_{T2} variables are bounded by the mass of the corresponding parent particle. Indeed, the $M_{T2}^{(b)}$ distribution for $t\bar{t}$ production shows a sharp drop around $M_{T2}^{(b)} = m_t$, while the signal distribution extends well above m_t. Similarly, the $M_{T2}^{(\ell)}$ distribution for $t\bar{t}$ drops around m_W, as expected. In addition, it exhibits a peak structure in the first bin, which is due to leptonic tau decays. This suggests that $M_{T2}^{(\ell)}$ can be effective in eliminating backgrounds with τs.

This concludes our discussion of the 16 kinematic variables depicted in figure 1. The newly introduced 6 variables (Topness, Higgsness, $\hat{s}_{\text{min}}^{(bb\ell\ell)}$, $\hat{s}_{\text{min}}^{(\ell\ell)}$, $M_{T2}^{(b)}$ and $M_{T2}^{(\ell)}$) typically require a few extra steps to compute them, thus we shall refer to them as high-level kinematic variables, while the remaining 10 traditional variables will be called low-level kinematic variables. We will perform two independent analyses — one with and one without the high-level kinematic variables, in order to estimate the performance benefit from adding the additional 6 variables.

4 Color flow in signal and backgrounds

We note that the two b-quarks in the signal result from the decay of a single non-colored object, the Higgs boson. In contrast, the two b-quarks in $t\bar{t}$ production (which is the dominant background) arise from the decays of top quarks, which in turn are produced via the strong interactions from a gluon-gluon initial state. This distinction is pictorially illustrated in figure 4. The different color-flow [86] will lead to different hadronization patterns, which can be used to discriminate a color singlet particle from a color octet (or triplet) at hadron colliders such as the LHC. Since the quarks which originate from a color singlet particle are color-connected to each other, their hadronization will not involve the initial state partons. On the contrary, the quarks which originate from a color octet particle are color-connected to the annihilating partons in the initial state, and consequently their hadronization is correlated with these initial state partons, see figure 4.

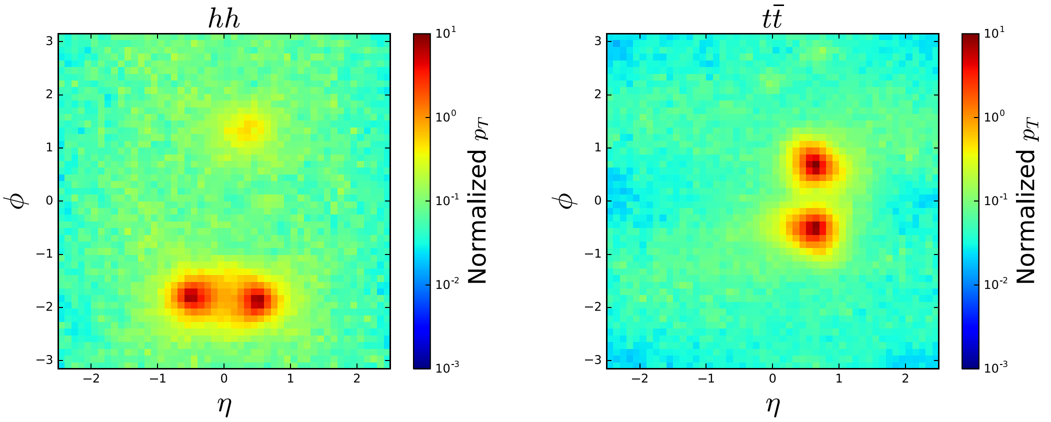
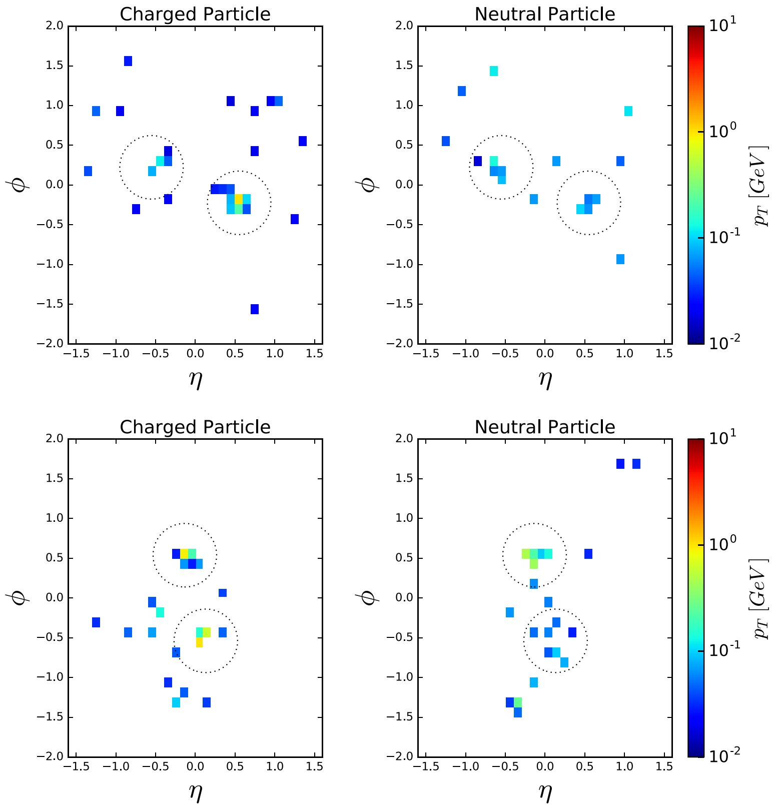


Figure 5. Cumulative p_T distributions resulting from showering 10,000 times a single partonic event for the signal (left) and $t\bar{t}$ production (right). The two b quarks from $h \to b\bar{b}$ are color-connected to each other and the soft radiation tends to fill in the region between them (left panel), while the two b quarks from $t\bar{t}$ production are not color-connected and the two clusters from their hadronization tend to be more isolated (right panel).

The difference in color flow will be reflected in the resulting hadron distributions. Hadrons coming from a color-singlet object will tend to be closer to the direction of the original mother particle, and as a result, the soft radiation will tend to populate the region between the two b quarks. On the other hand, hadrons from the decay of a color-octet particle will not be so narrowly focused, due to the influence of the initial state partons. These features are illustrated in figure 5, where we show the cumulative p_T distributions in the (η, ϕ) plane after showering the same partonic event 10,000 times. In the left panel we used a signal event, while in the right panel we used an event from $t\bar{t}$ production. We see that the b-jet clusters in the right panel tend to be better defined and more isolated, since they are not color-correlated among themselves. On the other hand, in the left panel we observe quite a bit of soft radiation in the region between the two b jets, due to the existing color connection between them.

Of course, the results in figure 5 are only valid in the statistical sense, since we took the same parton-level event and hadronized it multiple times. In reality, only one instance of this hadronization will be realized, as illustrated in figure 6. The top row of plots shows the hadronization patterns for charged particles (left panel) and neutral particles (right panel) in the case of one signal event, while the bottom row shows the same, but for one $t\bar{t}$ event. The parton-level event information is quoted (in GeV) to the right of each row of panels, and then each event is translated in the (η, ϕ) plane until the origin is aligned with the direction of the b-quark pair. The color scheme indicates the total p_T in each pixel, while the dotted circles represent the $\Delta R = 0.4$ cones for reconstruction of the corresponding b jets.

As an alternative to figure 5, in figures 7 and 8, we illustrate the effects of color-connection by showing the average of the jet images for the signal and the different background processes before and after the baseline cuts, respectively (some basic generation-



$$
\begin{aligned}
p_{b_1} &= (234.5, 15.0, -207.8, 106.7)\\
p_{b_2} &= (54.0, 22.3, -39.0, -29.2)\\
p_{\ell_1} &= (30.8, -11.1, 28.5, 2.8)\\
p_{\ell_2} &= (136.6, 0.9, 129.7, 42.8)\\
\vec{\not{p}}_T &= (-23.6, 71.6)\\
p_{j_1} &= (174.9, -33.8, 7.9, -171.4)\\
(\eta_{b_1}, \phi_{b_1}) &= (0.49, -1.50)\\
(\eta_{b_2}, \phi_{b_2}) &= (-0.61, -1.05)
\end{aligned}
$$

$$
\begin{aligned}
p_{b_1} &= (273.5, -111.7, -29.5, 247.6)\\
p_{b_2} &= (281.8, -63.7, 68.1, 265.6)\\
p_{\ell_1} &= (57.7, 22.4, -18.2, 49.9)\\
p_{\ell_2} &= (63.9, 30.0, -16.3, 54.0)\\
\vec{\not{p}}_T &= (153.3, -70.2)\\
p_{j_1} &= (51.6, 29.6, 15.3, 38.6)\\
(\eta_{b_1}, \phi_{b_1}) &= (1.51, -2.88)\\
(\eta_{b_2}, \phi_{b_2}) &= (1.77, 2.32)
\end{aligned}
$$

Figure 6. Transverse momentum distribution of charged particles (left) and neutral particles (right) for one chosen signal event (top row) and one chosen $t\bar{t}$ event (bottom row), where the origin is taken to be the center of the b-quark pair. The dotted circles represent the ΔR=0.4 cones for reconstructing the corresponding b jets. The four-momentum information of each event is given to the right of each panel row.

level cuts were imposed on the events in figure 7). The origin of the (η, ϕ) plane plane is taken to be the center of the b quark pair and the color scheme indicates the total p_T in each pixel. The black dotted line delineates the region $1.6 \leq \eta \leq 1.6$ and $-2.01 \leq \phi \leq 2.01$ used in the analysis. One can observe a striking difference in density between signal and background events in figure 7 — the two b quarks tend to be more collimated in the signal and more spread out in the background.

Unfortunately, after imposing the baseline cuts introduced in section 2, this distinction tends to be washed out and the backgrounds start mimicking the signal: one can see a similar structure emerging in all panels in figure 8, albeit with some subtle differences. Although one may find it difficult to discriminate signal from backgrounds simply by looking at a particular event, the patterns in the average jet images are different, and have been used actively for signal versus background separation [45, 47, 87]. In this paper, instead of quantifying the difference (e.g., with a pull vector [45]) we will use the images themselves on deep neural networks (DNNs), along with the 16 kinematic variables introduced in the previous section.

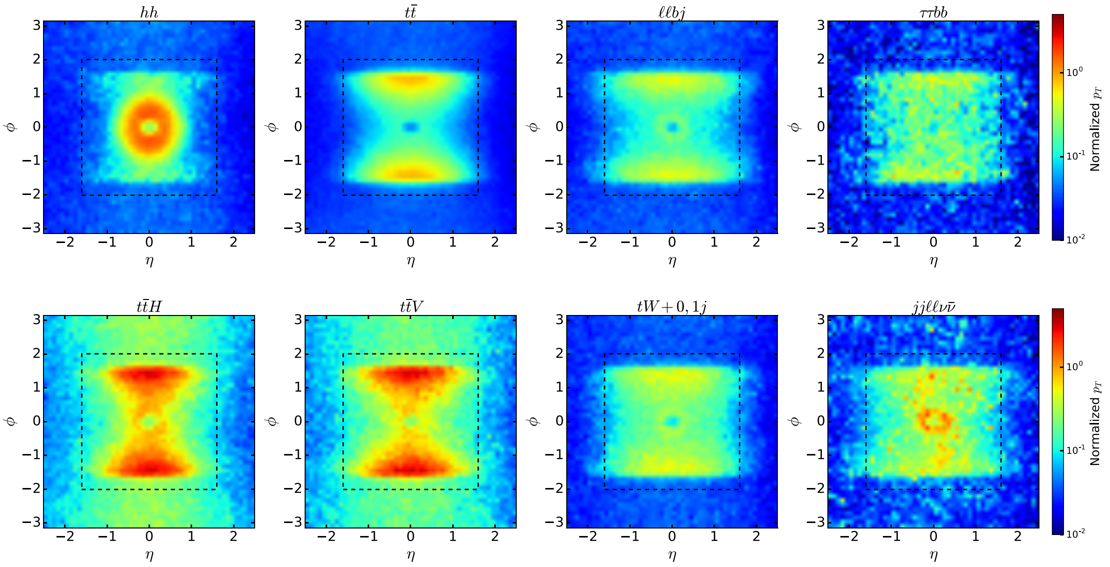


Figure 7. The cumulative average of the jet images for the signal and the different background processes before the baseline cuts (basic cuts at the event generation stage were still imposed). The origin of the (η, ϕ) plane is taken to be the center of the b quark pair and the color scheme indicates the total p_T in each pixel. The black dotted line delineates the region $1.6 \leq \eta \leq 1.6$ and $-2.01 \leq \phi \leq 2.01$ used in the analysis.

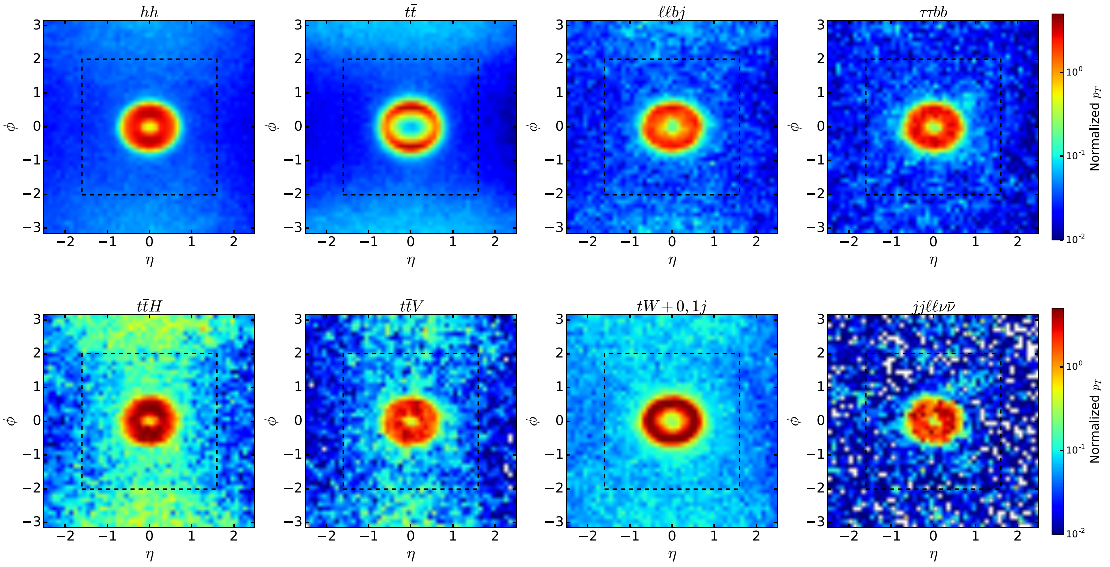


Figure 8. The same as Figure 7, but after imposing the baseline cuts introduced in section 2.

5 Analysis using deep learning

DNN is known to be very efficient and powerful in image recognition [88, 89] and the particle physics community has used it for various applications.[7] For instance, one can map the information about the direction and the energy (or transverse momentum) of a particle onto a pixel in an image. DNN then provides excellent classification between signal and background in the jet image [48, 50, 53, 54]. It also shows performance gains in multivariate analyses over traditional cut-and-count analyses or BDTs [52, 93]. In this section, we describe how we organize our analysis in a DNN framework. In the following three subsections, we address the issues of data pre-processing, DNN architecture and training of the NN.

5.1 Data pre-processing

In order to achieve the improved DNN learning performance and to minimize the error, it is important to properly process signal and backgrounds events before feeding them into a DNN framework. For each event passing the baseline cuts, the jet images are processed as follows.

1. *Input data:* we use the particle flow for our input data [94].

2. *Particle classification:* we divide the particle flow into two groups: neutral particles and charged particles. Neutral particles include photons and neutral hadrons, while charged particles include charged hadrons.

3. *Lepton removal:* if there is a lepton, we remove it.

4. *Shift:* we shift all particle coordinates in the (η, ϕ) plane with respect to the center of the reconstructed b-quark pair, i.e., we set $(\frac{\eta_b+\eta_{\bar{b}}}{2}, \frac{\phi_b+\phi_{\bar{b}}}{2})$ as the new origin, (0,0).

5. *Pixelization:* we discretize the rectangular region in the (η, ϕ) plane defined by $-2.5 \leq \eta \leq 2.5$ and $-\pi \leq \phi \leq \pi$ into a grid of 50×50 pixels for each particle classification (charged particle set and neutral particle set). In each pixel, we record the total transverse momentum as the pixel's intensity (in case of more than one particle, we add the transverse momenta and record the total sum). We refer to this 50×50 discrete image as the jet image [48].

6. *Normalization:* we rescale each jet image intensity as $I^{ij} \to I^{ij}/I^{\max}$, where $i, j = 1, 2, \ldots, 50$, and I^{ij} represents the intensity value in the (i, j) pixel. $I^{\max}$ is defined to be the largest value of pixel intensity found in the two 50×50 pixel images.

7. *Cropping:* we crop the jet image to 32×32 pixels, by further restricting to the (η, ϕ) rectangular range of $-1.6 \leq \eta \leq 1.6$ and $-2.01 \leq \phi \leq 2.01$.

[7] The use of neural networks for data analysis in high energy physics can be traced back to the pioneering work by R. Field and his students in the mid-nineties [90–92].

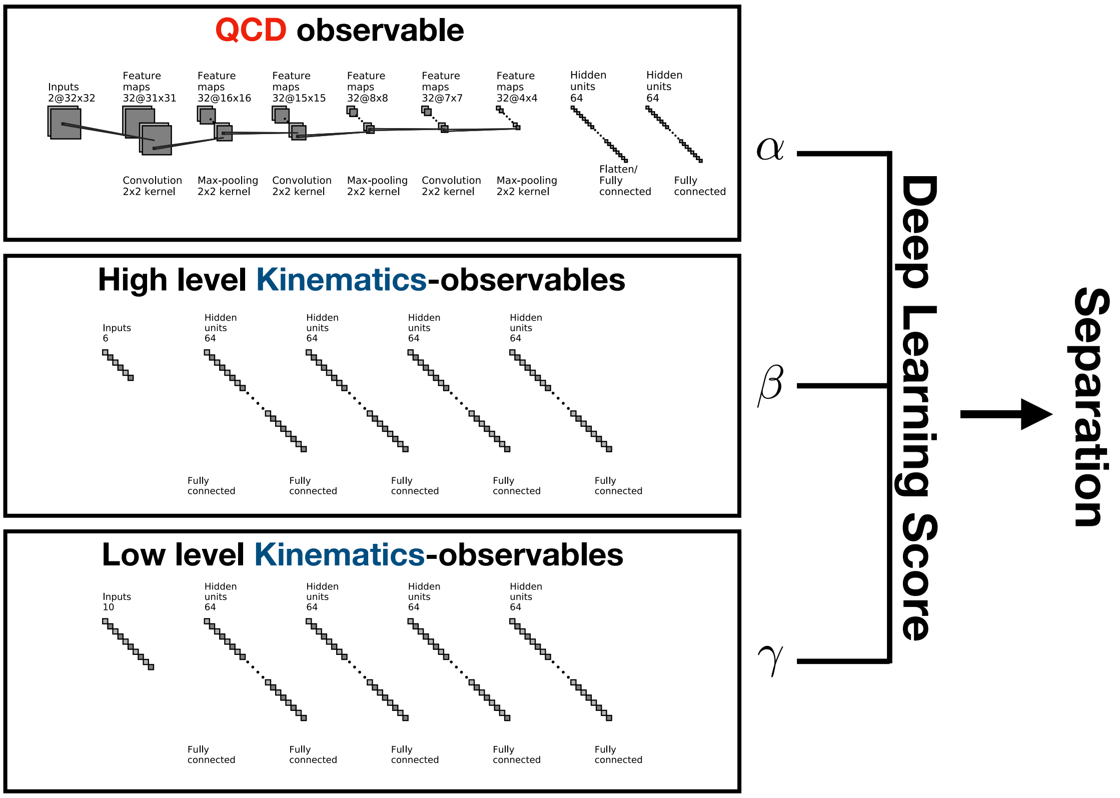


Figure 9. Illustration of the concept of combined Deep Learning.[8]

The final jet image has dimension $2 \times 32 \times 32$ and is comprised of one charged particle channel with dimension $1 \times 32 \times 32$ and a neutral particle channel with dimension $1 \times 32 \times 32$. This pre-processed jet-image is the input to the DNN. We note that Figures 7 and 8 showed the combined $1 \times 50 \times 50$ jet-image obtained by adding the neutral and charged particle layers. The black dotted rectangular area in those figures showed the restricted $1 \times 32 \times 32$ pixel area.

5.2 DNN architecture

Our DNN architecture consists of three sub-architectures, which will merge later, as illustrated in figure 9. Combined deep learning (DL) is not yet very common,[9] but recently there have been several studies in particle physics [50], as well as in other areas [96, 97], which showed improved results over simple DL. In this subsection, we provide some details of our DNN layer architecture as follows:

0. *Initialization.* Since DNN has a lot of parameters, it is important to give non-biased initial values for the (weight, bias) before running DNN with all input data. We use the He uniform initialization method as in ref. [98], among several other algorithms for parameter initialization [98–100].

[8] Parts of figure 9 are generated using the Python script in https://github.com/gwding/draw_convnet.

[9] Combined DL is similar to ensemble learning [95].

1. *Jet images.* They are represented by the top panel in figure 9.

   (a) *Input data:* we use pre-processed jet images as inputs.

   (b) *Convolutional neural networks layers:* we use three layers of convolutional neural networks (CNN). Each layer has a $32 \times 2 \times 2$ filter with no stride and no padding. We proceed with the batch normalization process after filtering [101], using the ReLU function as our activation function [102]. After activation, we introduce the max pooling layer which has a 2×2 shape with 2×2 strides and padding.

   (c) *Dense layers:* we feed the output of the CNN into two fully connected 1×64 dense layers, using ReLU as the activation function.

2. *The 6 high level variables.* Those are illustrated by the middle panel in figure 9.

   (a) *Input data:* $\sqrt{\hat{s}}_{\rm min}^{(bb\ell\ell)}$, $\sqrt{\hat{s}}_{\rm min}^{(\ell\ell)}$, $M_{T2}^{(b)}$, $M_{T2}^{(\ell)}$, Higgsness and Topness.

   (b) *Dense layers:* we introduce four fully connected 1×64 dense layers with the ReLU activation function. All four layers have the batch normalization process before activation.

3. *The 10 low level variables*: those are illustrated by the bottom panel in figure 9.

   (a) *Input data:* $p_{T\ell_1}$, $p_{T\ell_2}$, $\not{p}_T$, $m_{\ell\ell}$, m_{bb}, $\Delta R_{\ell\ell}$, ΔR_{bb}, p_{Tbb}, $p_{T\ell\ell}$, $\Delta\phi_{bb,\ell\ell}$.

   (b) *Dense layers:* we follow the same procedure as in the case with the 6 high level variables above.

4. *Combination*

   (a) *Merge:* we apply three single (1×1) dense layers to the jet image, the 6 high level variables and the 10 low level variables. These layers are denoted as α, β, and γ, respectively, as shown in figure 9. To merge the three sub-architectures, we introduce the final dense layer of dimension 1×3 without an activation function.

   (b) *Final output:* to distinguish signal from backgrounds, we apply a layer of dimension 1×2 without an activation function.

5.3 DNN training

We now proceed with deep learning on the DNN architecture described in section 5.2, using the pre-processed input data. We use Microsoft CNTK [103] as the main DNN library on GPU with an Nvidia CUDA platform. We use the Adam optimizer [104] with cross entropy with SoftMax loss function and classification error function. The sizes of the training data set and the testing data set are about 40k and 17k, respectively. The size of the mini-batch is 128 and that of the epoch is 30.

For each event, we prepare the jet images and the 16 variables. The dimension of the final output is 1×2, $(\mathcal{P}_{\rm sig}, \mathcal{P}_{\rm bknd} = 1 - \mathcal{P}_{\rm sig})$. If the deep learning score is equal to 1, i.e., $\mathcal{P}_{\rm sig} = 1$ $(\mathcal{P}_{\rm bknd} = 0)$, the corresponding event is taken to be a signal event. If $\mathcal{P}_{\rm sig} = 0$ $(\mathcal{P}_{\rm bknd} = 1)$, the event is considered to be background.

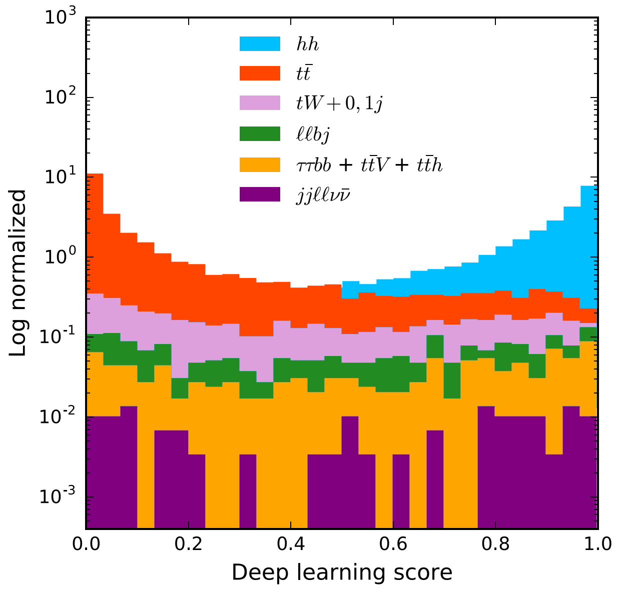


Figure 10. Deep learning score for the signal and the individual backgrounds.

6 Results

In this section we present our results. First we validate our framework by repeating the analysis performed in ref. [39] under similar assumptions.[10] We obtained consistent results for the conventional cut-and-count method with Delphes detector simulation. When we added deep learning, the signal significance improved slightly by 5-10%.

Now considering *all* relevant backgrounds and using *all 16 variables and jet images*, we show the deep learning score for the signal and the individual background processes in figure 10. The signal should peak near $\mathcal{P}_{\rm sig} = 1$ by construction, and indeed this is what is observed in the figure. Note that the $t\bar{t}$ and tW processes are well separated from the signal and both peak near $\mathcal{P}_{\rm sig} = 0$. This is direct consequence of the improvements made in our analysis — introducing the proper kinematic variables and jet images, which were meant to target the dominant background ($t\bar{t}$ production), as evidenced in figures 1, 3 and 8. Although the subdominant backgrounds are also reduced in this process, they remain rather flat in figure 10.

The deep learning score shown in in figure 10 can now be used as a signal-to-background discriminator. By placing a lower cut and counting the number of surviving signal and

[10] Our current analysis has several notable improvements over the one carried out in ref. [39]. First, the detector simulation is different — in the current study, we use Delphes, which assumes (on average) $\sim 90\%$ ($\sim 80\%$) reconstruction efficiency for leptons (b-jets), while ref. [39] assumed 100% reconstruction efficiency for both. In addition, the Delphes detector resolution itself is slightly different from one used in ref. [39]. In particular we find that the resolution of the missing transverse momentum is worse in Delphes and hence our current results are more conservative (if not more realistic). Finally, as mentioned earlier, we are now including $tW + j$ production, which turned out to be the next dominant background, yet was missing from all previous studies. These effects should be kept in mind when comparing our results here to previous results in the literature.

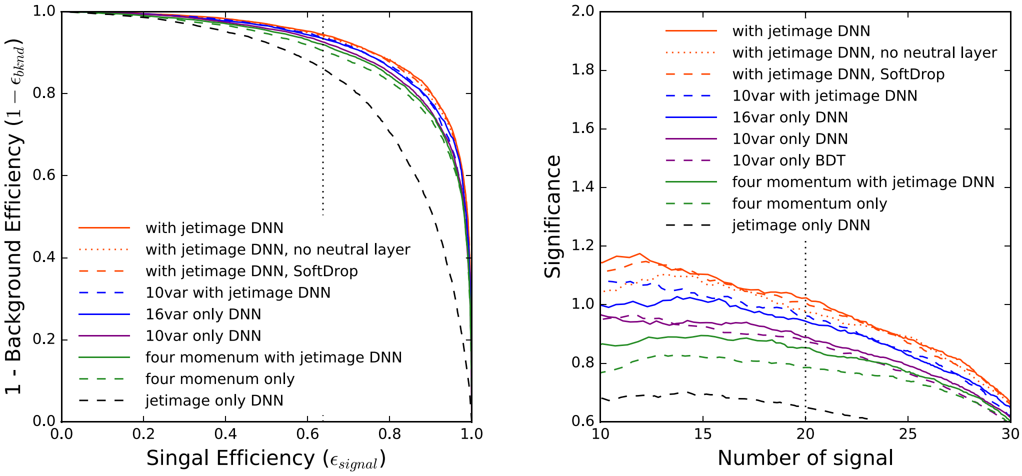


Figure 11. A ROC curve (left panel) and signal significance as a function of the number of signal events (right panel). The vertical lines mark $N = 20$ signal events, which corresponds to $\epsilon_{\text{signal}} = 0.64$.

background events, one obtains the efficiency curve (also known as a receiver operating characteristic (ROC)) shown in the left panel of figure 11. The curve contains several independent runs of deep learning and shows the signal efficiency (ϵ_{signal}) versus the fraction of rejected background events, i.e., $1 - \epsilon_{\text{bknd}}$, where ϵ_{bknd} is the background efficiency. The efficiency corresponding to the results in figure 10 is shown with the red solid curve labeled "with jetimage DNN". The other two solid lines show the efficiencies which would be obtained if we were to remove the jet images from the analysis: the purple solid curve (labelled "10var only DNN") is obtained with the help of the 10 low-level kinematic variables, while the blue solid curve (labelled "16var only DNN") shows the improvement when we add the 6 high-level variables and use the full set of 16 variables from section 3, but still without jet images. The black dotted curve (labeled "jet image only DNN") shows the result when we use jet images alone, with no help from any of the 16 kinematic variables. Finally, the blue dashed line (labelled "10var with jetimage DNN") shows the result from an analysis combining jet images with the 10 low-level kinematic variables only. The corresponding signal significances are shown as a function of the number of events in the right panel of figure 11. Note that the right panel contains an additional curve (the purple dashed line labeled "10var only BDT") where we use the 10 low-level variables and adopt a BDT algorithm using the TMVA tool kit [105]. The comparison of the latter line against the "10var only DNN" result (purple solid line) reveals the relative performance of DNN versus BDT.

In order to examine the effects of pile-up, we use several methods as follows. In the first method, we use the Soft Drop algorithm [58] to remove soft jet activity which is exacerbated by pile-up. We set $\beta = 0$ and $z_{\text{cut}} = 0.1$ with $R = 1.2$ anti-k_T clustered fatjets. Then we select the closest fatjet to the $b\bar{b}$ momentum in the η-ϕ plane and replace the particle flow data with the charged and neutral jet constituents of the selected fatjet. Soft Drop does not affect the jet images and retains the same shapes as in figure 8. In second

	Signal	$t\bar{t}$	$t\bar{t}h$	$t\bar{t}V$	$\ell\ell bj$	$\tau\tau bb$	$tw+j$	$jj\ell\ell\nu\nu$	σ	S/B
Baseline cuts: $\not{p}_T > 20$ GeV, $p_{T,\ell} > 20$ GeV, $\Delta R_{\ell\ell} < 1.0$, $p_{T,b} > 30$ GeV, $\Delta R_{bb} < 1.3$, $m_{\ell\ell} < 65$ GeV, $95 < m_{bb} < 140$ GeV	0.01046	1.8855	0.0269	0.0179	0.0697	0.0250	0.2209	0.0113	0.38	0.0046
jet-image DL	0.00667	0.1855	0.0147	0.00731	0.0243	0.0128	0.0626	0.00786	0.65	0.021
10 low-level variables DL	0.00668	0.0738	0.0132	0.00529	0.0184	0.00842	0.0424	0.00516	0.89	0.040
16 variables DL	0.00668	0.0676	0.0109	0.00454	0.0163	0.00689	0.0376	0.00418	0.94	0.045
10 variables + jet-image DL	0.00667	0.0630	0.00964	0.00429	0.0194	0.00791	0.0343	0.00393	0.96	0.047
16 variables + jet-image DL	0.00668	0.0602	0.00914	0.00252	0.0133	0.00689	0.0299	0.00344	1.0	0.053

Table 1. Signal and background cross sections in fb after baseline cuts (first row) and at different stages of analysis, using a combination of kinematic variables and jet images while requiring $N = 20$ signal events. The significance σ is calculated using the log-likelihood ratio for a luminosity of 3 ab^{-1} at the 14 TeV LHC.

method, we remove the neutral jet image layer in the analysis. Unlike charged particles, which can be cleaned up from pile-up relatively easily by checking the longitudinal vertex information [106], neutral particles cannot be treated the same way and suffer from non-removable pile-up effects. The corresponding results with these two pile-up mitigation methods are also shown in figure 11 with the red dotted line labelled “16var with jetimage DNN, SoftDrop” and the red, dashed line labelled “16var with jetimage DNN, no neutral layer”, respectively.

We also examine the performance of the DNN with four momentum information as input. The corresponding results are shown in figure 11, where the green-dashed (green-solid) curve represents the significance with four momentum information only (four momentum information plus jet images). The inputs are 18 real numbers, i.e. the four momenta of the two leptons and the two b-tagged jets and the missing transverse momentum. For this exercise, we use a 4 × 128 dense layer instead of a 4 × 64 dense layer. We notice that the DNN performance with kinematic variables is better. This is because, in general, the use of four momenta requires a large training sample in order to be effective, while the kinematic variables already perform efficiently with a smaller data set. If the architecture is deep enough with a large amount of data, the DNN performance with four momentum information would be comparable (or better) to that with kinematic variables only. This exercise illustrates the importance of the appropriate use of kinematic variables.

In summary, figure 11 demonstrates that jet images (which capture the effects of color flow) can improve performance over the baseline selection cuts. At the same time, DL with jet image substructure alone does not show the best performance, and becomes fully effective (and still stable under pile-up) only when it is combined with the full set of 16 variables, including the high-level ones.

Table 1 summarizes the signal and background cross sections in fb at different stages of the analysis for the case of $N = 20$ signal events. The last two columns show the signal significance σ and the signal-to-background ratio S/B. The significance is calculated using the log-likelihood ratio for a luminosity of 3 ab^{-1} at the 14 TeV LHC.

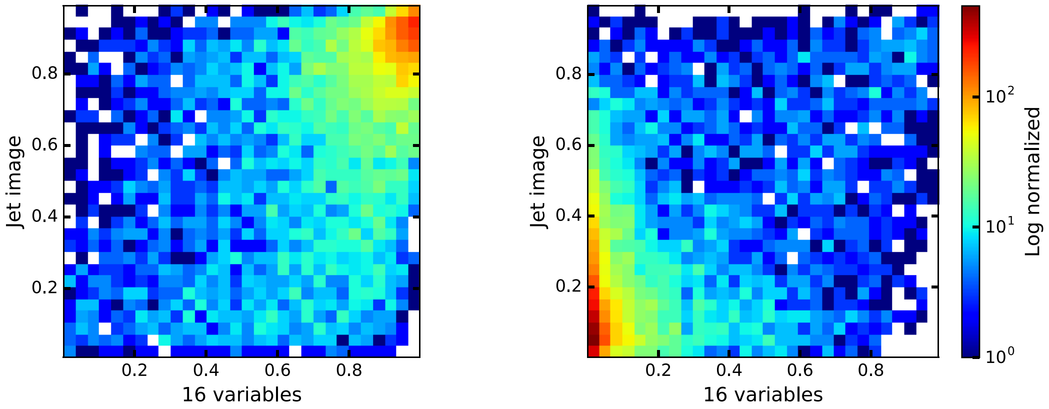


Figure 12. Correlation of the deep learning scores obtained in independent DL analyses using the 16 kinematic variables only (x-axis) and jet images only (y-axis) for signal (left panel) and background (right panel).

In order to understand the correlation between jet images and the 16 kinematic variables, we performed two independent runs with "jet images only; no kinematic variables" and "16 kinematic variables; no jet images". The corresponding results are shown in figure 12. Since the two DLs are trained separately, both the x-axis and the y-axis are normalized to unity. As expected, figure 12 reveals a degree of correlation between the jet images and the 16 kinematic variables, which is somewhat stronger for the signal and less so for the background.

In our main analysis, we performed simultaneous runs as shown in the deep learning architecture in figure 9. Before calculating our final deep learning score, we obtain three intermediate values, α, β, and γ, which represent the DL scores for the respective substructure corresponding to the jet images, the 6 high level variables and the 10 low level variables. The first 6 panels in figure 13 show the pair-wise correlations between these three intermediate scores for the signal (top row) and the background (middle row). The bottom three panels in the figure show the one-dimensional distributions of the intermediate scores for signal (blue histograms) and background (red histograms). We observe that the score from jet images (α) is relatively uncorrelated to the kinematic variables scores β and γ, which motivates the simultaneous training on jet images and kinematic variables together.

Finally, in figure 14 we scan over different values of the triple Higgs coupling κ_3 and show the discovery significance (left panel) and precision (middle panel) as a function of κ_3. Both the significance σ and the precision $\Delta\chi^2$ are calculated fixing DL cuts that would give a certain number of signal events ($N = 15, 20, 25, 30$) for the SM at $\kappa_3 = 1$ (marked with the dotted vertical line). For the significance, we used the log-likelihood-ratio

$$\sigma_{dis} \equiv \sqrt{-2\,\ln\left(\frac{L(B|S+B)}{L(S+B|S+B)}\right)} \quad \text{with} \quad L(x|n) = \frac{x^n}{n!}e^{-x}\,, \tag{6.1}$$

where S and B are the expected number of signal and background events, respectively. We

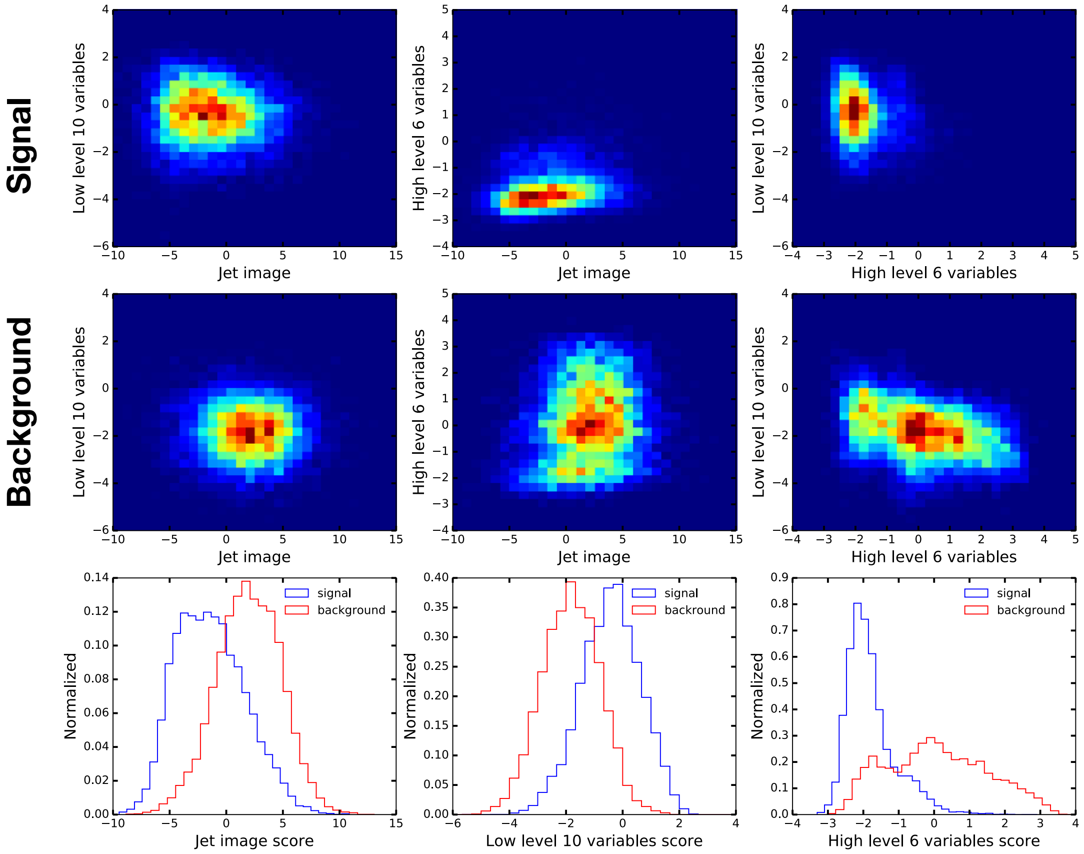


Figure 13. Correlations among the intermediate DL substructure scores for jet images, the 6 high level variables and the 10 low level variables. The top (middle) row shows the correlations for signal (background) events and the bottom row shows the corresponding distributions for each individual substructure score.

define $\Delta\chi^2$ as

$$\Delta\chi^2 = \left(\frac{\Big(S(\kappa_3)+B\Big) - \Big(S(\kappa_3=1)+B\Big)}{\sqrt{S(\kappa_3=1)+B}}\right)^2 . \tag{6.2}$$

The shape of the significance roughly follows the cross section ratios between the case of $\kappa_3 \neq 1$ to the case of $\kappa_3 = 1$. This is illustrated in the rightmost panel of figure 14, which shows the cross section scaled as $\sigma(\kappa_3)/\min\big(\sigma(\kappa_3)\big)$, i.e., normalized with respect to the minimum cross section for each curve. The blue curve represents the double Higgs production cross section before cuts, and in this case we find the minimum of the cross section somewhere between $\kappa_3 = 2$ and $\kappa_3 = 3$. After baseline cuts (the red solid line), the minimum shifts to around $\kappa_3 \sim 4$, and after DL cuts (the green solid line), the minimum shifts even further out to around $\kappa_3 \sim 5$. In the latter case, we observe that the signal cross sections for $\kappa_3 = 1$ and $\kappa_3 = 8$ are numerically very close, as indicated by the two vertical dotted lines in the right panel. This provides an explanation for the double dip structure seen in the middle panel of figure 14.

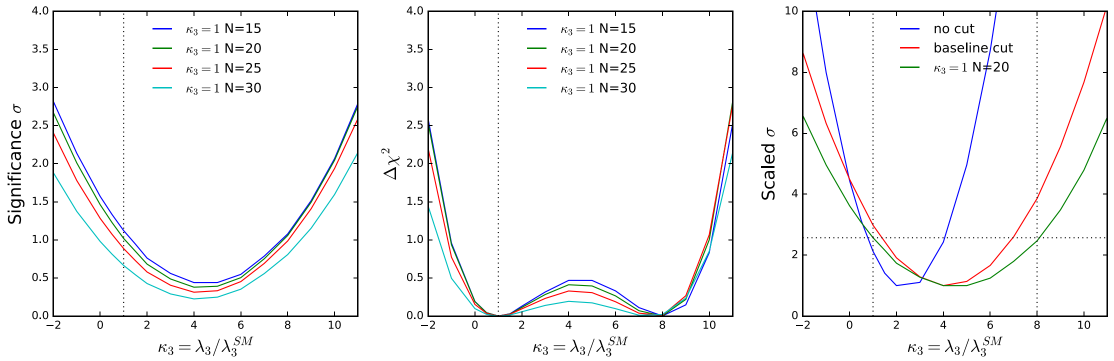


Figure 14. The discovery significance (left panel) and precision (middle panel) as a function of κ_3. Both the significance and the precision are calculated fixing DL cuts that would give a certain number of signal events ($N = 15, 20, 25, 30$) for the SM at $\kappa_3 = 1$ (marked with the dotted vertical line). The shape of the significance roughly follows the cross section ratios between the case of $\kappa_3 \neq 1$ to the case of $\kappa_3 = 1$. This is illustrated in the right panel, which shows the scaled cross section, computed as $\sigma(\kappa_3)/\min\big(\sigma(\kappa_3)\big)$ for each curve.

As demonstrated in the rightmost panel of figure 14, the analysis cuts modify the signal cross section so that the location of its minimum shifts to higher values of κ_3. This can be understood as follows. At leading order, the Higgs pair production cross section is given by

$$\sigma_{gg\to hh}(\hat{s}) = \frac{\alpha_s^2}{2^{15}v^4\pi^2\hat{s}^2}\int d\hat{t}(|F_1|^2 + |F_2|^2) \approx c_\triangle\,\kappa_3^2 + c_{\triangle,\Box}\,\kappa_3 + c_\Box\,, \tag{6.3}$$

before convoluting with the parton distribution functions [107, 108]. Here F_1 represents a parity-even triangle and box diagram contribution, while F_2 is a parity-odd box diagram contribution. Now F_1 can be rewritten as $F_1 = \kappa_3 F_\triangle + F_\Box$, where $F_\triangle$ is the triangle diagram contribution and $F_\Box$ is the box diagram contribution. Therefore the cross section can be parameterized as a quadratic function of κ_3, where the c coefficients are related to contributions from $\triangle$ and $\Box$ diagrams.

The observation that the baseline cuts and the DL cut shift the minimum cross section to a larger κ_3 value implies that the effects of the cuts are stronger on $c_\triangle$ than $c_{\triangle,\Box}$. In other words, our cuts are more likely to affect the triangle diagram which contains the triple Higgs coupling. Unlike the box diagram, the triangle diagram includes an off-shell Higgs in the s-channel. Since it is harder to produce a Higgs pair from an s-channel off-shell Higgs, the Higgs pair generated from the triangle diagram is not as energetic as the one coming from the box diagram, and will therefore tend to have lower transverse momentum. As discussed in section 4, several of the cuts on our kinematic variables, namely, ΔR_{bb}, $\Delta R_{\ell\ell}$, $\Delta\phi_{bb,\ell\ell}$, p_{Tbb} and $p_{T\ell\ell}$, rely on the fact that the Higgs bosons are produced with a significant boost. Consequently, the effect of the cuts will be to suppress the $c_\triangle$ term and enhance the box diagram contribution, which in turn shifts the location of the minimum to a larger value of κ_3.

Note that the results for the significance and the precision in figure 14 do not change dramatically when we require a different number of signal events at the SM point. This

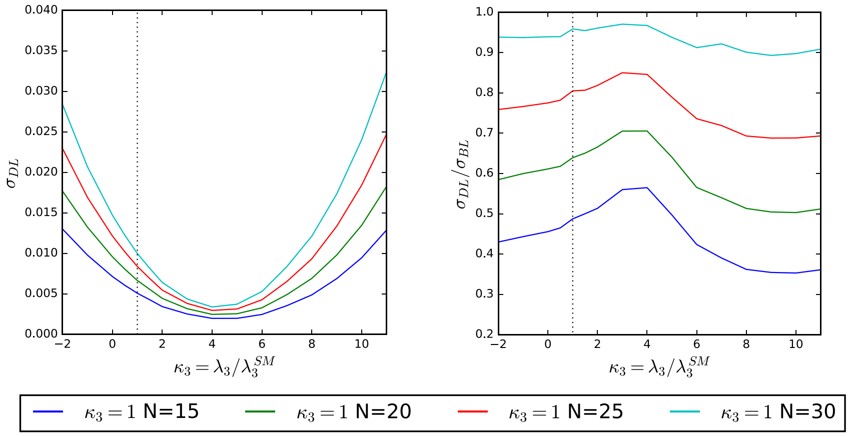


Figure 15. Cross section $\sigma_{DL}(pp \to hh \to bb\ell^+\ell^-\nu\bar{\nu})$ in pb after cutting on the DL score (left panel) and the ratio $\sigma_{DL}/\sigma_{\rm baseline}$ between the cross section σ_{DL} after the DL cut and the cross section $\sigma_{\rm baseline}$ after baseline cuts (right panel).

means that the dependence on the DL cut is relatively mild, since the kinematics remains similar when we vary κ_3, so that the dependence on the cross section is more important. This is illustrated in figure 15, which shows the cross section (in pb) after cutting on the DL score, $\sigma_{DL}(pp \to hh \to bb\ell^+\ell^-\nu\bar{\nu})$, (left panel) and the ratio $\sigma_{DL}/\sigma_{\rm baseline}$ between the cross section σ_{DL} after the DL cut and the cross section $\sigma_{\rm baseline}$ after baseline cuts (right panel).

7 Discussion

In this paper, we investigated double Higgs production in the $hh \to bbWW^* \to bb\ell\ell + \vec{\not{P}}_T$ final state. It is known to be one of the difficult channels due to the large backgrounds, $\sigma_{\rm bknd}/\sigma_{\rm hh} \sim 10^5$. We performed a detailed analysis by adopting a deep learning framework and successfully combining new kinematic variables and jet image information. As a result, we obtained a sizable increase in signal sensitivity and an improved signal-to-background ratio compared to the existing analyses.

Our results showed that the dominant $t\bar{t}$ background can be brought down to the level of the other remaining backgrounds, without sacrificing too much in the signal rate. This is mostly due to the use of Higgsness, Topness and the subsystem variable $M_{T2}^{(b)}$. Other backgrounds like $bb\tau\tau$ can be reduced further by the use of $M_{T2}^{(\ell)}$. Finally, additional improvements are possible with the use of jet images. After all cuts, we find that all backgrounds contribute at similar levels.

We find from recent CMS and ATLAS analyses with 36 fb^{-1} of LHC data at 13 TeV that the 95% confidence level observed (expected) upper limit on the production cross section is 22.2 (12.8) times the standard model value [7] for CMS and 6.7 (10.4) times the predicted Standard Model cross-section [109] for ATLAS. The leading channel in CMS is $bb\gamma\gamma$ followed by $bb\tau\tau$, while the leading channels in ATLAS are $bb\tau\tau$ and $bbbb$, followed by $bb\gamma\gamma$. The main difference arises due to the superior b-tagging efficiency for the ATLAS detector [14]. In both studies, the $bbWW^*$ channel was largely overlooked due to the expected poor significance. However, our study suggests that double Higgs production may be probed in the dilepton $bbWW^*$ channel as well, and would contribute to the combined analysis on par with the other final states, increasing the overall significance. For example, in ref. [13], the ATLAS collaboration showed that the combined significance of $hh \to bbbb$, $hh \to bb\tau\tau$, and $hh \to bb\gamma\gamma$ is 3.5 (3.0) without (with) systematic uncertainties at the 14 TeV LHC with 3 ab^{-1}. Their individual significance is 1.4, 2.5 and 2.1 (0.61, 2.1 and 2.0), respectively without (with) systematics. They did not combine with the $hh \to bbWW^*$ channel but a naive estimate shows that when including our channel, the combined significance would be about 3.7.

We urge the experimental collaborations to consider the ideas presented in this paper and test them in the LHC data. We would also like to mention that the proposed method can be easily generalized to the semi-leptonic channel from $hh \to bbWW^*$ production, as well as to other processes with similar final states.

Acknowledgments

This work is supported in part by United States Department of Energy (DE-SC0010296, DE-SC0017988, DE-SC0017965, DE-SC0019474) and Korea NRF-2018R1C1B6006572. We thank KIAS for providing computing resources. We thank Georgios Anagnostou for valuable comments and Anja Butter, Tilman Plehn and Chris Rogan for general discussion on machine learning. KK is grateful to the Mainz Institute for Theoretical Physics, which is part of the DFG Cluster of Excellence PRIMA$^+$ (Project ID 39083149), for its hospitality and its partial support during the completion of this work.

A Deep Neural Network

The artificial neural network (ANN) is one of the most popular approaches to pattern recognition in machine learning algorithms. The structure of an ANN is defined by a succession of non-linear and linear transformations between nodes or artificial neurons, which are located on input, output or hidden layers. A hidden layer which uses an ordinary one-dimensional layer is called a dense layer or a fully connected layer.

The linear operation consists of weights and bias:

$$O[i] = \sum_{j=0}^{n_I-1} \left(\mathcal{W}[i,j]I[j] + \mathcal{B}[i]\right), \quad i = 0, \cdots, n_O - 1, \tag{A.1}$$

where $I[j]$ is the value of the j-th neuron (input) in the prior layer, $O[i]$ the value of the i-th neuron (output) in the subsequent layer, $\mathcal{W}[i,j]$ are the weights, and $\mathcal{B}[i]$ the bias. The index i (j) takes the values $0, \cdots, n_O - 1$ ($0, \cdots, n_I - 1$) and n_O (n_I) is the dimension of the output (input). The input initially can be given in more than one dimension. For example, if the input results from a convolution and has dimension $n \times n$, it may be rearranged as follows:

$$I[j] = (I[1,1] \cdots I[1,n] \cdots I[n,1] \cdots I[n,n])\,, \tag{A.2}$$

where the corresponding dimension of the input would be $n_I = n^2$.

The non-linear transformation is often called activation function, which imitates the action potential of biological neurons. Similar to how each neuron adjusts how much signal it needs to deliver to the next neuron using an electric action potential, the activation function determines the output of a particular neuron for a set of given inputs from neurons on the previous layer, and the output is then used as input for the next artificial neuron. The commonly used activation functions are

$$\mathrm{ReLU}(x[i]) = \max(0, x[i])\,, \tag{A.3}$$

$$\mathrm{Sigmoid}(x[i]) = \frac{1}{1+e^{-x[i]}}\,, \tag{A.4}$$

$$\mathrm{SoftMax}(x[i]) = \frac{e^{x[i]}}{\sum_i e^{x[i]}}\,, \tag{A.5}$$

where $x[i]$ represents the value of the i-th neuron.

If the neural network has sufficiently many hidden layers, the network is called deep neural network (DNN). DNN can learn from the input data to obtain the desirable output by adjusting the parameters in the hidden layers. We note that the proper normalization of the input data helps improve convergence during training. The goal of the training is to determine the parameters (weights and biases) by minimizing the loss, which represents the difference between the target output and the actual DNN output. There are various algorithms for optimization of the parameters [104, 110, 111]. Some well known loss functions are

$$\text{Mean Square Error} = \frac{1}{n}\sum_{i=1}^{n}(x[i]-t[i])^2\,, \tag{A.6}$$

$$\text{Cross Entropy} = -\sum_{i=1}^{n} t[i]\log(x[i])\,, \tag{A.7}$$

$$\text{Cross Entropy with SoftMax} = -\sum_{i=1}^{n} t[i]\log(\mathrm{SoftMax}(x[i]))\,, \tag{A.8}$$

where $\{t[i]\}$ is the true answer (either 1 or 0 in our current study), $\{x[i]\}$ is the DNN final output, and n is the number of neurons in the output layer.

Instead of feeding the entire data into the DNN all at once, one splits the input data into several subsets with random selection and takes one subset, called mini-batch, for a given iteration, which helps avoid the over-fitting problem [112, 113]. When the full training set is used, the cycle is called epoch, and one uses several epochs to obtain a well-trained DNN. When training DNN with a mini-batch, the corresponding loss is defined by the sum of all losses over the mini-batch or by their average.

Once the training is over, for testing one uses a different data set from the one used in the DNN training, in order to avoid the over-fitting problem. In order to test the trained

DNN model one can use either the loss function or the classification error function. If the number of test events is n, the classification error function is defined by

$$\text{Classification Error} = \frac{1}{n}\sum_{j=1}^{n}\delta[\text{ArgMax}(\{t[i]\}_j), \text{ArgMax}(\{x[i]\}_j)], \tag{A.9}$$

where $\text{ArgMax}(\{y[i]\})$ gives the position $i_{\max}$ where the value of $\{y[i]\}$ is maximized. j represents the j-th test event, the δ is Kronecker delta function.

Often one takes additional steps such as dropout for reducing over-fitting in neural networks [114] and batch normalization for improving the performance and stability of artificial neural networks [101]. Dropout makes a random drop of units (both hidden and visible) in a neural network and is considered an efficient way of performing model averaging. The batch normalization procedure normalizes the input layer by adjusting and scaling the activations:

$$O[i] = \gamma\hat{I}[i] + \beta\,, \tag{A.10}$$

$$\hat{I}[i] = \frac{I[i]-\mu[i]}{\sqrt{(\sigma[i])^2+\epsilon}}\,, \tag{A.11}$$

$$\mu[i] = \frac{1}{n}\sum_{\alpha=1}^{n} I[i]_\alpha\,, \tag{A.12}$$

$$\Big(\sigma[i]\Big)^2 = \frac{1}{n}\sum_{\alpha=1}^{n}\Big(I[i]_\alpha - \mu[i]\Big)^2, \tag{A.13}$$

where α represents the α-th input in a mini-batch and n is the size of the mini-batch. The dimensions of input and output are the same. Note that (γ, β) are the learned parameters during the training and ϵ is a parameter added to avoid a divergence in the denominator. The batch normalization allows each layer of a network to learn by itself independently of the other layers.

A convolutional neural network (CNN) is a class of DNN, most commonly used to analyze images. CNN utilizes filters made of a set of neurons with a fixed size. The value of parameters in each filter is learned during the training process. By varying the position of the filters on the input and learning the values of different filters, CNN can find local features of the input data. This process is called convolution and a hidden layer which uses convolution is called a convolutional layer. With n'_f filters whose size is $(n_{fs}\times n_{fs})$, the convolution is defined as follows

$$O[i,j,k] = \sum_{\gamma=0}^{n_f-1}\sum_{\alpha,\beta}\Big(\mathcal{W}[\alpha,\beta,\gamma,k]I[\alpha,\beta,\gamma] + \mathcal{B}[k]\Big)\,, \tag{A.14}$$

where the dimension of the input is $n_f\times(n\times n)$ and the dimension of output is $n'_f\times(n'\times n')$. The corresponding ranges of the parameters are $k=\{0,\cdots,n'_f-1\}$, $\alpha=\{i,\cdots,i+n_{f_s}\}$, $\beta=\{j,\cdots,j+n_{f_s}\}$, $i,j=\{0,n_s,2n_s,\cdots,n'\}$, $n'=n/n_s-n_{fs}+n_s$, and n_s is called the stride.

Since each filter has a finite size, the output size decreases, after applying the convolution (A.14) on the input or on the output from a previous layer. In order to prevent the size reduction, CNN incorporates the padding process:

$$I = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} \rightarrow O = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & \alpha & \beta & 0 \\ 0 & \gamma & \delta & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} , \tag{A.15}$$

which increases the size of the original input by adding zeros around it. Usually the padding is used before applying convolution or pooling.

CNN may include local or global pooling layers (often called sub-sampling), which combine the output of several neurons at one layer to a single neuron in the next layer. For example, max (average) pooling takes the maximum (average) value from a set of neurons at the previous layer and passes it to next layer. For a pooling dimension n_p, the relation between the output with dimension $n_f \times (n' \times n')$ and the input with dimension $n_f \times (n \times n)$ is given by

$$O[i,j,k] = \text{Max (Average)}(\{I[\alpha,\beta,k]\}) , \tag{A.16}$$

where $\alpha = \{i, \cdots, i+n_p\}$, $\beta = \{j, \cdots, j+n_p\}$, $k = \{0, \cdots, n_f - 1\}$, $i, j = \{0, n_s, 2n_s, \cdots, n'\}$, $n' = n/n_s - n_p + n_s$, and n_s is the stride.

Another beneficial feature of a CNN is the reduction of the number of parameters via convolution and pooling, which effectively increases the learning speed in deep neutral networks. A typical DNN architecture consists of a combination of convolutional layers and dense layers, which provides better performance compared to a NN with only one type of layers [115].

References

[1] ATLAS collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. **B 716** (2012) 1 [arXiv:1207.7214] [INSPIRE].

[2] CMS collaboration, *Observation of a new boson at a mass of* 125 *GeV with the CMS experiment at the LHC*, Phys. Lett. **B 716** (2012) 30 [arXiv:1207.7235] [INSPIRE].

[3] ATLAS, CMS collaboration, *Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at* $\sqrt{s} = 7$ *and* 8 *TeV*, JHEP **08** (2016) 045 [arXiv:1606.02266] [INSPIRE].

[4] ATLAS collaboration, *Study of the double Higgs production channel* $H(\rightarrow b\bar{b})H(\rightarrow \gamma\gamma)$ *with the ATLAS experiment at the HL-LHC*, ATL-PHYS-PUB-2017-001 (2017).

[5] ATLAS collaboration, *Projected sensitivity to non-resonant Higgs boson pair production in the* $b\bar{b}b\bar{b}$ *final state using proton–proton collisions at HL-LHC with the ATLAS detector*, ATL-PHYS-PUB-2016-024 (2016).

[6] J.H. Kim, Y. Sakaki and M. Son, *Combined analysis of double Higgs production via gluon fusion at the HL-LHC in the effective field theory approach*, Phys. Rev. **D 98** (2018) 015016 [arXiv:1801.06093] [INSPIRE].

[7] CMS collaboration, *Combination of searches for Higgs boson pair production in proton-proton collisions at $\sqrt{s} = 13$ TeV*, Phys. Rev. Lett. **122** (2019) 121803 [arXiv:1811.09689] [INSPIRE].

[8] CMS collaboration, *Higgs pair production at the High Luminosity LHC*, CMS-PAS-FTR-15-002 (2015).

[9] CMS Collaboration, *Projected performance of Higgs analyses at the HL-LHC for ECFA 2016*, CMS-PAS-FTR-16-002 (2017).

[10] J. Baglio et al., *The measurement of the Higgs self-coupling at the LHC: theoretical status*, JHEP **04** (2013) 151 [arXiv:1212.5581] [INSPIRE].

[11] CMS collaboration, *Search for resonant and nonresonant Higgs boson pair production in the $b\bar{b}\ell\nu\ell\nu$ final state in proton-proton collisions at $\sqrt{s} = 13$ TeV*, JHEP **01** (2018) 054 [arXiv:1708.04188] [INSPIRE].

[12] HL/HE WG2 GROUP collaboration, *Higgs Physics at the HL-LHC and HE-LHC*, arXiv:1902.00134 [INSPIRE].

[13] ATLAS collaboration, *Measurement prospects of the pair production and self-coupling of the Higgs boson with the ATLAS experiment at the HL-LHC*, ATL-PHYS-PUB-2018-053 (2018).

[14] ATLAS collaboration, *Search for pair production of Higgs bosons in the $b\bar{b}b\bar{b}$ final state using proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, JHEP **01** (2019) 030 [arXiv:1804.06174] [INSPIRE].

[15] CMS Collaboration, *Search for non-resonant Higgs pair-production in the $b\bar{b}b\bar{b}$ final state with the CMS detector*, CMS-PAS-HIG-17-017 (2018).

[16] D.E. Ferreira de Lima, A. Papaefstathiou and M. Spannowsky, *Standard model Higgs boson pair production in the $(b\bar{b})(b\bar{b})$ final state*, JHEP **08** (2014) 030 [arXiv:1404.7139] [INSPIRE].

[17] D. Wardrope et al., *Non-resonant Higgs-pair production in the $b\bar{b}$ $b\bar{b}$ final state at the LHC*, Eur. Phys. J. **C 75** (2015) 219 [arXiv:1410.2794] [INSPIRE].

[18] J.K. Behr et al., *Boosting Higgs pair production in the $b\bar{b}b\bar{b}$ final state with multivariate techniques*, Eur. Phys. J. **C 76** (2016) 386 [arXiv:1512.08928] [INSPIRE].

[19] CMS collaboration, *Search for Higgs boson pair production in the $\gamma\gamma b\bar{b}$ final state in pp collisions at $\sqrt{s} = 13$ TeV*, Phys. Lett. **B 788** (2019) 7 [arXiv:1806.00408] [INSPIRE].

[20] ATLAS collaboration, *Search for Higgs boson pair production in the $\gamma\gamma b\bar{b}$ final state with 13 TeV pp collision data collected by the ATLAS experiment*, JHEP **11** (2018) 040 [arXiv:1807.04873] [INSPIRE].

[21] CMS collaboration, *Higgs pair production at the High Luminosity LHC*, CMS-PAS-FTR-15-002 (2015).

[22] ATLAS collaboration, *Prospects for measuring Higgs pair production in the channel $H(\to \gamma\gamma)H(\to b\bar{b})$ using the ATLAS detector at the HL-LHC*, ATL-PHYS-PUB-2014-019 (2014).

[23] F. Kling, T. Plehn and P. Schichtel, *Maximizing the significance in Higgs boson pair analyses*, Phys. Rev. **D 95** (2017) 035026 [arXiv:1607.07441] [INSPIRE].

[24] U. Baur, T. Plehn and D.L. Rainwater, *Probing the Higgs selfcoupling at hadron colliders using rare decays*, Phys. Rev. **D 69** (2004) 053004 [hep-ph/0310056] [INSPIRE].

[25] P. Huang, A. Joglekar, B. Li and C.E.M. Wagner, *Probing the electroweak phase transition at the LHC*, Phys. Rev. **D 93** (2016) 055049 [arXiv:1512.00068] [INSPIRE].

[26] A. Azatov, R. Contino, G. Panico and M. Son, *Effective field theory analysis of double Higgs boson production via gluon fusion*, Phys. Rev. **D 92** (2015) 035001 [arXiv:1502.00539] [INSPIRE].

[27] Q.-H. Cao, B. Yan, D.-M. Zhang and H. Zhang, *Resolving the degeneracy in single Higgs production with Higgs pair production*, Phys. Lett. **B 752** (2016) 285 [arXiv:1508.06512] [INSPIRE].

[28] Q.-H. Cao et al., *Double Higgs production at the* 14 *TeV LHC and a* 100 *TeV pp collider*, Phys. Rev. **D 96** (2017) 095031 [arXiv:1611.09336] [INSPIRE].

[29] A. Alves, T. Ghosh and K. Sinha, *Can we discover double Higgs production at the LHC?*, Phys. Rev. **D 96** (2017) 035022 [arXiv:1704.07395] [INSPIRE].

[30] V. Barger, L.L. Everett, C.B. Jackson and G. Shaughnessy, *Higgs-pair production and measurement of the triscalar coupling at LHC(8,14)*, Phys. Lett. **B 728** (2014) 433 [arXiv:1311.2931] [INSPIRE].

[31] J. Chang et al., *Higgs-boson-pair production $H(\to b\bar{b})H(\to \gamma\gamma)$ from gluon fusion at the HL-LHC and HL-100 TeV hadron collider*, arXiv:1804.07130 [INSPIRE].

[32] ATLAS collaboration, *Search for resonant and non-resonant Higgs boson pair production in the $b\bar{b}\tau^+\tau^-$ decay channel in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Rev. Lett. **121** (2018) 191801 [*Erratum ibid.* 122 (2019) 089901] [arXiv:1808.00336] [INSPIRE].

[33] CMS collaboration, *Search for Higgs boson pair production in events with two bottom quarks and two τ leptons in proton–proton collisions at $\sqrt{s} = 13$ TeV*, Phys. Lett. **B 778** (2018) 101 [arXiv:1707.02909] [INSPIRE].

[34] U. Baur, T. Plehn and D.L. Rainwater, *Examining the Higgs boson potential at lepton and hadron colliders: a comparative analysis*, Phys. Rev. **D 68** (2003) 033001 [hep-ph/0304015] [INSPIRE].

[35] F. Goertz, A. Papaefstathiou, L.L. Yang and J. Zurita, *Higgs boson pair production in the $D = 6$ extension of the SM*, JHEP **04** (2015) 167 [arXiv:1410.3471] [INSPIRE].

[36] M.J. Dolan, C. Englert and M. Spannowsky, *Higgs self-coupling measurements at the LHC*, JHEP **10** (2012) 112 [arXiv:1206.5001] [INSPIRE].

[37] ATLAS collaboration, *Search for Higgs boson pair production in the $b\bar{b}WW^*$ decay mode at $\sqrt{s} = 13$ TeV with the ATLAS detector*, JHEP **04** (2019) 092 [arXiv:1811.04671] [INSPIRE].

[38] CMS Collaboration, *Search for resonant and non-resonant Higgs boson pair production in the $b\bar{b}l\nu l\nu$ final state at $\sqrt{s} = 13$ TeV*, CMS-PAS-HIG-17-006 (2017).

[39] J.H. Kim, K. Kong, K.T. Matchev and M. Park, *Probing the triple Higgs self-interaction at the Large Hadron Collider*, Phys. Rev. Lett. **122** (2019) 091801 [arXiv:1807.11498] [INSPIRE].

[40] A. Papaefstathiou, L.L. Yang and J. Zurita, *Higgs boson pair production at the LHC in the $b\bar{b}W^+W^-$ channel*, Phys. Rev. **D 87** (2013) 011301 [arXiv:1209.1489] [INSPIRE].

[41] T. Huang et al., *Resonant di-Higgs boson production in the $b\bar{b}WW$ channel: probing the electroweak phase transition at the LHC*, Phys. Rev. **D 96** (2017) 035007 [arXiv:1701.04442] [INSPIRE].

[42] ATLAS collaboration, *Search for Higgs boson pair production in the $WW^{(*)}WW^{(*)}$ decay channel using ATLAS data recorded at $\sqrt{s} = 13$ TeV*, JHEP **05** (2019) 124 [arXiv:1811.11028] [INSPIRE].

[43] A. Adhikary et al., *Revisiting the non-resonant Higgs pair production at the HL-LHC*, JHEP **07** (2018) 116 [arXiv:1712.05346] [INSPIRE].

[44] B. Bhattacherjee, S. Mukherjee and R. Sengupta, *Discrimination between prompt and long-lived particles using convolutional neural network*, arXiv:1904.04811 [INSPIRE].

[45] J. Gallicchio and M.D. Schwartz, *Seeing in color: jet superstructure*, Phys. Rev. Lett. **105** (2010) 022001 [arXiv:1001.5027] [INSPIRE].

[46] J. Gallicchio et al., *Multivariate discrimination and the Higgs + W/Z search*, JHEP **04** (2011) 069 [arXiv:1010.3698] [INSPIRE].

[47] A. Hook, M. Jankowiak and J.G. Wacker, *Jet dipolarity: top tagging with color flow*, JHEP **04** (2012) 007 [arXiv:1102.1012] [INSPIRE].

[48] J. Cogan, M. Kagan, E. Strauss and A. Schwartzman, *Jet-images: computer vision inspired techniques for jet tagging*, JHEP **02** (2015) 118 [arXiv:1407.5675] [INSPIRE].

[49] L. de Oliveira et al., *Jet-images — Deep learning edition*, JHEP **07** (2016) 069 [arXiv:1511.05190] [INSPIRE].

[50] J. Lin, M. Freytsis, I. Moult and B. Nachman, *Boosting $H \to b\bar{b}$ with machine learning*, JHEP **10** (2018) 101 [arXiv:1807.10768] [INSPIRE].

[51] L. de Oliveira, M. Paganini and B. Nachman, *Learning particle physics by example: location-aware generative adversarial networks for physics synthesis*, Comput. Softw. Big Sci. **1** (2017) 4 [arXiv:1701.05927] [INSPIRE].

[52] P. Baldi, P. Sadowski and D. Whiteson, *Searching for exotic particles in high-energy physics with deep learning*, Nature Commun. **5** (2014) 4308 [arXiv:1402.4735] [INSPIRE].

[53] P.T. Komiske, E.M. Metodiev and M.D. Schwartz, *Deep learning in color: towards automated quark/gluon jet discrimination*, JHEP **01** (2017) 110 [arXiv:1612.01551] [INSPIRE].

[54] G. Kasieczka, T. Plehn, M. Russell and T. Schell, *Deep-learning top taggers or the end of QCD?*, JHEP **05** (2017) 006 [arXiv:1701.08784] [INSPIRE].

[55] DELPHES 3 collaboration, *DELPHES 3, a modular framework for fast simulation of a generic collider experiment*, JHEP **02** (2014) 057 [arXiv:1307.6346] [INSPIRE].

[56] M. Cacciari, G.P. Salam and G. Soyez, *FastJet user manual*, Eur. Phys. J. **C 72** (2012) 1896 [arXiv:1111.6097] [INSPIRE].

[57] ATLAS collaboration, *Expected performance of the ATLAS detector at the High-Luminosity LHC*, ATL-PHYS-PUB-2019-005 (2019).

[58] A.J. Larkoski, S. Marzani, G. Soyez and J. Thaler, *Soft drop*, JHEP **05** (2014) 146 [arXiv:1402.2657] [INSPIRE].

[59] ATLAS collaboration, *Performance of missing transverse momentum reconstruction with the ATLAS detector using proton-proton collisions at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. **C 78** (2018) 903 [arXiv:1802.08168] [INSPIRE].

[60] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections and their matching to parton shower simulations*, *JHEP* **07** (2014) 079 [arXiv:1405.0301] [INSPIRE].

[61] NNPDF collaboration, *Parton distributions with QED corrections*, *Nucl. Phys.* **B 877** (2013) 290 [arXiv:1308.0598] [INSPIRE].

[62] J. Grigo, K. Melnikov and M. Steinhauser, *Virtual corrections to Higgs boson pair production in the large top quark mass limit*, *Nucl. Phys.* **B 888** (2014) 17 [arXiv:1408.2422] [INSPIRE].

[63] M. Czakon, P. Fiedler and A. Mitov, *Total top-quark pair-production cross section at hadron colliders through $O(\alpha_S^4)$*, *Phys. Rev. Lett.* **110** (2013) 252004 [arXiv:1303.6254] [INSPIRE].

[64] LHC Higgs Cross Section Working Group collaboration, *Handbook of LHC Higgs cross sections: 1. Inclusive observables*, arXiv:1101.0593 [INSPIRE].

[65] LHC Higgs Cross Section Working Group collaboration, *Handbook of LHC Higgs cross sections: 4. Deciphering the nature of the Higgs sector*, arXiv:1610.07922 [INSPIRE].

[66] D. de Florian, M. Der and I. Fabre, *QCD⊕QED NNLO corrections to Drell-Yan production*, *Phys. Rev.* **D 98** (2018) 094008 [arXiv:1805.12214] [INSPIRE].

[67] T. Sjöstrand et al., *An introduction to PYTHIA 8.2*, *Comput. Phys. Commun.* **191** (2015) 159 [arXiv:1410.3012] [INSPIRE].

[68] M. Cacciari, G.P. Salam and G. Soyez, *The anti-k_t jet clustering algorithm*, *JHEP* **04** (2008) 063 [arXiv:0802.1189] [INSPIRE].

[69] T. Han, I.-W. Kim and J. Song, *Kinematic cusps: determining the missing particle mass at colliders*, *Phys. Lett.* **B 693** (2010) 575 [arXiv:0906.5009] [INSPIRE].

[70] T. Han, I.-W. Kim and J. Song, *Kinematic cusps with two missing particles II: cascade decay topology*, *Phys. Rev.* **D 87** (2013) 035004 [arXiv:1206.5641] [INSPIRE].

[71] T. Han, I.-W. Kim and J. Song, *Kinematic cusps with two missing particles I: antler decay topology*, *Phys. Rev.* **D 87** (2013) 035003 [arXiv:1206.5633] [INSPIRE].

[72] W.S. Cho, D. Kim, K.T. Matchev and M. Park, *Probing resonance decays to two visible and multiple invisible particles*, *Phys. Rev. Lett.* **112** (2014) 211801 [arXiv:1206.1546] [INSPIRE].

[73] P. Konar, K. Kong and K.T. Matchev, *$\sqrt{\hat{s}}_{\rm min}$: a global inclusive variable for determining the mass scale of new physics in events with missing energy at hadron colliders*, *JHEP* **03** (2009) 085 [arXiv:0812.1042] [INSPIRE].

[74] P. Konar, K. Kong, K.T. Matchev and M. Park, *RECO level $\sqrt{s}_{\rm min}$ and subsystem $\sqrt{s}_{\rm min}$: improved global inclusive variables for measuring the new physics mass scale in $\not{E}_T$ events at hadron colliders*, *JHEP* **06** (2011) 041 [arXiv:1006.0653] [INSPIRE].

[75] M. Burns, K. Kong, K.T. Matchev and M. Park, *Using subsystem MT2 for complete mass determinations in decay chains with missing energy at hadron colliders*, *JHEP* **03** (2009) 143 [arXiv:0810.5576] [INSPIRE].

[76] C.G. Lester and D.J. Summers, *Measuring masses of semiinvisibly decaying particles pair produced at hadron colliders*, *Phys. Lett.* **B 463** (1999) 99 [hep-ph/9906349] [INSPIRE].

[77] A.J. Barr et al., *Guide to transverse projections and mass-constraining variables*, *Phys. Rev.* **D 84** (2011) 095031 [arXiv:1105.2977] [INSPIRE].

[78] D. Kim, K.T. Matchev, F. Moortgat and L. Pape, *Testing invisible momentum ansatze in missing energy events at the LHC*, *JHEP* **08** (2017) 102 [arXiv:1703.06887] [INSPIRE].

[79] W.S. Cho et al., *On-shell constrained M_2 variables with applications to mass measurements and topology disambiguation*, *JHEP* **08** (2014) 070 [arXiv:1401.1449] [INSPIRE].

[80] P. Konar, K. Kong, K.T. Matchev and M. Park, *Superpartner mass measurement technique using 1D orthogonal decompositions of the Cambridge transverse mass variable M_{T2}*, *Phys. Rev. Lett.* **105** (2010) 051802 [arXiv:0910.3679] [INSPIRE].

[81] P. Konar, K. Kong, K.T. Matchev and M. Park, *Dark matter particle spectroscopy at the LHC: generalizing M_{T2} to asymmetric event topologies*, *JHEP* **04** (2010) 086 [arXiv:0911.4126] [INSPIRE].

[82] P. Baringer, K. Kong, M. McCaskey and D. Noonan, *Revisiting Combinatorial Ambiguities at Hadron Colliders with M_{T2}*, *JHEP* **10** (2011) 101 [arXiv:1109.1563] [INSPIRE].

[83] D. Kim and K. Kong, *Kinematic discrimination of tW and $t\bar{t}$ productions using initial state radiation*, *Phys. Lett.* **B 751** (2015) 512 [arXiv:1503.03872] [INSPIRE].

[84] D. Goncalves, K. Kong and J.H. Kim, *Probing the top-Higgs Yukawa CP structure in dileptonic $t\bar{t}h$ with M_2-assisted reconstruction*, *JHEP* **06** (2018) 079 [arXiv:1804.05874] [INSPIRE].

[85] D. Debnath et al., *Resolving combinatorial ambiguities in dilepton $t\bar{t}$ event topologies with constrained M_2 variables*, *Phys. Rev.* **D 96** (2017) 076005 [arXiv:1706.04995] [INSPIRE].

[86] F. Maltoni, K. Paul, T. Stelzer and S. Willenbrock, *Color flow decomposition of QCD amplitudes*, *Phys. Rev.* **D 67** (2003) 014026 [hep-ph/0209271] [INSPIRE].

[87] ATLAS collaboration, *Measurement of colour flow using jet-pull observables in $t\bar{t}$ events with the ATLAS experiment at $\sqrt{s} = 13$ TeV*, *Eur. Phys. J.* **C 78** (2018) 847 [arXiv:1805.02935] [INSPIRE].

[88] A. Krizhevsky, I. Sutskever and G.E. Hinton, *Imagenet classification with deep convolutional neural networks*, in *Advances in neural information processing systems 25*, F. Pereira et al. eds., Curran Associates Inc., U.S.A. (2012).

[89] Y. Lecun, Y. Bengio and G. Hinton, *Deep learning*, *Nature* **521** (2015) 436.

[90] R.D. Field, Y. Kanev, M. Tayebnejad and P.A. Griffin, *Using neural networks to enhance the Higgs boson signal at hadron colliders*, *Phys. Rev.* **D 53** (1996) 2296 [INSPIRE].

[91] R.D. Field, Y. Kanev and M. Tayebnejad, *A Topological analysis of the top quark signal and background at hadron colliders*, *Phys. Rev.* **D 55** (1997) 5685 [INSPIRE].

[92] R. Field, *Genetic algorithms and neural networks as tools in particle physics*, talk given at the Tevatron University, Fermilab, May 21 (1998).

[93] P. Baldi et al., *Jet substructure classification in high-energy physics with deep neural networks*, *Phys. Rev.* **D 93** (2016) 094034 [arXiv:1603.09349] [INSPIRE].

[94] CMS collaboration, *Particle-flow event reconstruction in CMS and performance for jets, taus and MET*, CMS-PAS-PFT-09-001 (2009).

[95] T.G. Dietterich, *Ensemble methods in machine learning*, in *Multiple classifier systems*, N.C. Oza et al eds., Springer, Berlin Germany (2000).

[96] T. Majtner, S. Yildirim-Yayilgan and J.Y. Hardeberg, *Combining deep learning and hand-crafted features for skin lesion classification*, talk given at the *Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA 2016)*, December 12–15, Oulu, Finland (2016).

[97] E. Park, X. Han, T.L. Berg and A.C. Berg, *Combining multiple sources of knowledge in deep cnns for action recognition*, talk given at the *EEE Winter Conference on Applications of Computer Vision (WACV 2016)*, March 7–9, Lake Placid, U.S.A. (2016).

[98] K. He, X. Zhang, S. Ren and J. Sun, *Delving deep into rectifiers: surpassing human-level performance on imagenet classification*, arXiv:1502.01852 [INSPIRE].

[99] X. Glorot and Y. Bengio, *Understanding the difficulty of training deep feedforward neural networks*, *J. Mach. Learn. Res. — Proc. Track* **9** (2010) 249.

[100] A.M. Saxe, J.L. McClelland and S. Ganguli, *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks*, arXiv:1312.6120.

[101] S. Ioffe and C. Szegedy, *Batch normalization: accelerating deep network training by reducing internal covariate shift*, arXiv:1502.03167 [INSPIRE].

[102] X. Glorot, A. Bordes and Y. Bengio, *Deep sparse rectifier neural networks*, in the proceedings of the 14th *International Conference on Artificial Intelligence and Statisitics (AISTATS 2011)*, April 11–13, Ft. Lauerdale, U.S.A. (2011).

[103] D. Yu et al., *An introduction to computational networks and the computational network toolkit*, technical report (2014).

[104] D.P. Kingma and J. Ba, *Adam: a method for stochastic optimization*, arXiv:1412.6980 [INSPIRE].

[105] A. Hocker et al., *TMVA — Toolkit for multivariate data analysis*, physics/0703039 [INSPIRE].

[106] D. Bertolini, P. Harris, M. Low and N. Tran, *Pileup per particle identification*, *JHEP* **10** (2014) 059 [arXiv:1407.6013] [INSPIRE].

[107] E.W.N. Glover and J.J. van der Bij, *Higgs boson pair production via gluon fusion*, *Nucl. Phys.* **B 309** (1988) 282 [INSPIRE].

[108] S. Borowka et al., *Full top quark mass dependence in Higgs boson pair production at NLO*, *JHEP* **10** (2016) 107 [arXiv:1608.04798] [INSPIRE].

[109] ATLAS collaboration, *Combination of searches for Higgs boson pairs in pp collisions at 13 TeV with the ATLAS experiment*, ATLAS-CONF-2018-043 (2018).

[110] J. Duchi, E. Hazan and Y. Singer, *Adaptive subgradient methods for online learning and stochastic optimization*, *J. Mach. Learn. Res.* **12** (2011) 2121.

[111] T. Tieleman and G. Hinton, *Lecture 6.5 — rmsprop, coursera: neural networks for machine learning*, technical report (2012).

[112] R. Ge, F. Huang, C. Jin and Y. Yuan, *Escaping from saddle points — online stochastic gradient for tensor decomposition*, arXiv:1503.02101.

[113] D. Masters and C. Luschi, *Revisiting small batch training for deep neural networks*, arXiv:1804.07612.

[114] N. Srivastava et al., *Dropout: A simple way to prevent neural networks from overfitting*, *J. Mach. Learn. Res.* **15** (2014) 1929.

[115] A. Krizhevsky, I. Sutskever and G.E. Hinton, *Imagenet classification with deep convolutional neural networks*, *Neural Inf. Proc. Syst.* **25** (2012).