Identification of a bottom quark-antiquark pair in a single jet with high transverse momentum and its application

Chunhui Chen

Identification of a bottom quark-antiquark pair in a single jet
with high transverse momentum and its application

Chunhui Chen

Department of Physics and Astronomy, Iowa State University, Ames, Iowa 50011, United State

In this paper we introduce a new approach to identify a bottom quark-antiquark pair inside a single jet with high transverse momentum by using the jet substructure in the center-of-mass frame of the jet. We demonstrate that the method can be used to discriminate the boosted heavy particles decaying to a $b\bar{b}$ final state from QCD jets. Applications to searches for the standard model Higgs boson (H) decaying to $b\bar{b}$ when produced in association with a weak vector boson are also discussed.

I. INTRODUCTION

The existence of a Higgs boson like particle with a mass of around $125\,\text{GeV}$ has been firmly established by both the ATLAS and CMS experiments [1, 2] in its bosonic decay modes ($H \rightarrow \gamma\gamma$, $H \rightarrow ZZ$, and $H \rightarrow WW$). However, its decay to a bottom quark-antiquark pair ($b\bar{b}$) final state has not been observed yet. Finding such a decay signal at the Large Hadron Collider (LHC) is challenging because of the large amount background from the production of multijets containing b quarks, despite that the $H \rightarrow b\bar{b}$ decay mode is predicted in the standard model (SM) to have a branch fraction of $58\,\%$ for $m_H = 125$ GeV. As a result, such searches have been mostly performed in the $pp \rightarrow VH$ production mode, where V is either a W or a Z boson that decays leptonically, and $H \rightarrow b\bar{b}$. So far, no evidence of such a decay has yet been seen by either the ATLAS or CMS experiment [3, 4].

It has been shown that the search sensitivity of $H \rightarrow b\bar{b}$ in the VH production mode can be significantly improved by reconstructing the hadronically decaying Higgs boson with large transverse momentum in a single jet [5], especially together with the implementation of the jet substructure techniques [6–8]. Such an approach requires the identification of both b quarks decaying from the Higgs boson in a single jet, hereafter referred as a Higgs jet (H jet). While the identification of an isolated jet stemming from the hadronization of a single b quark (b-tagging) has been widely used in many experimental measurements, its application to the Higgs jets is not trivial since a Higgs jet has two b quarks inside [9, 10]. In this paper, we extend the studies presented in Refs. [11–13] to explore the identification of the $b\bar{b}$ pair inside a H jet (double b-tagging) in the center-of-mass frame of the jet. We demonstrate that the method can greatly reduce the QCD jet background while maintaining a high identification efficiency of the boosted Higgs boson even in an environment with very large numbers of multiple interactions per event (pileup), where the QCD jets are defined as those jets initiated by a non-top quark or gluon.

We organize this paper as follows: In Section. II, we describe the event sample we used in the study. Section III discusses the method to identify a bottom quark-antiquark pair in a single jet in the jet center-of-mass frame and its performance. Applications of our method to the searches of $H \rightarrow b\bar{b}$ in the VH production mode are discussed in Section IV. We conclude in Section V.

II. EVENT SAMPLE

We use the boosted H jets, from the SM process of WH production, as an benchmark to illustrate our proposed double b-tagging method. For simplicity, we only consider the background from the SM W+jets production to study the background rejection performance of the QCD jets as it is the largest background in searches for $H \rightarrow b\bar{b}$ in the WH production mode. However, our method is generic and is applicable to any boosted heavy particles decaying to a $b\bar{b}$ final state. In addition, we also generate events to simulate the SM processes of ZH, WZ, WW, ZZ, Z+jets and top quark production.

All the events used in this analysis are produced using the Pythia 8.186 event generator [14, 15] for the pp collision at $14\,\text{TeV}$ center-of-mass energy. In order to simulate the finite resolution of the Calorimeter detector at the LHC experiments, we divide the (η, ϕ) plane into 0.1×0.1 cells. The energies of particles entering each cell in each event, except for the neutrinos, are summed over and replaced with a massless pseudoparticle of the same energy, also referred as an energy cluster, pointing to the center of the cell. These pseudoparticles are fed into the FastJet 3.0.1 [16] package for jet reconstruction. The jets are reconstructed with the anti-k_T algorithm [17] with a distance parameter of $\Delta R = 0.8$. The anti-k_T jet algorithm is the default one used at the ATLAS and CMS experiments. As for the charged tracks, their momentum and vertex positions are smeared according to the expected resolutions of the ATLAS detector [18]. In order to evaluate the performance of the double b-tagging with the currently expected experimental conditions at the LHC, we generate the MC events with different average numbers of multiple interactions per event, where the beamspot is assumed to follow a Gaussian distribution with a width of 0.015 mm in the transverse beam direction, and 45 mm in the longitudinal beam direction. We then perform our studies for each scenario and com-

pare their performances.

III. DOUBLE b-TAGGING AND JET SUBSTRUCTURE IN THE REST FRAME

A. Event Selection

In this section we describe the method to identify a bottom quark-antiquark pair in a single jet using the substructure in the center-of-mass frame of the jet in order to distinguish a boosted hadronically decaying Higgs boson from QCD jets.

The study is done using the MC simulated events of the WH and W+jets productions, where the W boson decays leptonically ($W \to e\nu, \mu\nu$). We select events with one isolated lepton (electron or muon) with $p_{\mathrm{T}} > 20$ GeV and $|\eta| < 2.4$, where p_{T} and η are the transverse momentum and psudorapidy of the lepton. For the jet reconstruction, studies show that its energy and invariant mass (m_{jet}) can significantly shift to higher values due to the presence of additional energy depositions from underlying events and pileups. We employ a jet area correction technique [19] to take into account the effects on the event-by-event basis. For each event, a distribution of transverse energy densities is calculated for all jets with $|\eta| < 2.1$ and its median is taken as an estimate of the energy density of the pileup and underlying events. We subsequently correct each jet by subtracting the product of the transverse energy density and the jet area, which is determined with the "active" area calculation technique [19]. This method results in a modified jet four-momentum that are used throughout the paper unless explicitly stated otherwise. The jets with $p_{\mathrm{T}} \geq 300$ GeV, $|\eta| \leq 1.7$ and 40 GeV $\leq m_{\mathrm{jet}} \leq 240$ GeV in an event are selected as the H jet candidates for further analysis.

For b-tagging, only charged tracks with $p_{\mathrm{T}} > 1$ GeV and $|\eta| < 2.5$ are considered. They are also required to satisfy the criteria that $|d_0| < 1$ mm and $|z_0 - z_{\mathrm{pv}}| \sin\theta < 1.5$ mm, where d_0 and z_0 are the transverse and longitudinal impact parameters of the charged track, z_{pv} is the longitudinal position of the primary vertex, and θ is the polar angle of the charged track. A charged track is considered to be associated with a jet if the distance parameter of $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ is less than 0.8, where $\Delta\eta$ and $\Delta\phi$ are defined as the differences in psudorapidity and azimuthal angle between the charged track and the jet, respectively.

B. Center-of-mass frame of a jet

We define the center-of-mass frame (rest frame) of a jet as the frame where the four momentum of the jet is equal to $p_\mu^{\mathrm{rest}} \equiv (m_{\mathrm{jet}}, 0, 0, 0)$. A jet consists of its constituent particles. The distribution of the constituent particles of a boosted Higgs jet in its center-of-mass frame looks like a back-to-back di-jet event with one b quark in each of the

subjets. On the other hand, a QCD jet acquires its mass through gluon radiation and it is not a closed system. The constituent particle distribution of a QCD jet in the rest frame is more likely to be random, as illustrated in Figure 1.
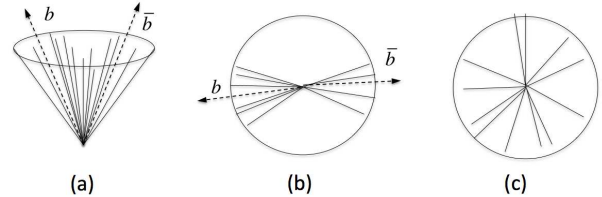

FIG. 1: Illustration of the constituent particle distribution of a jet. (a) H jet in the lab frame. (b) H jet of in the jet rest frame. (d) QCD jet in its rest frame.

C. Reclustering

We recluster the energy clusters of a jet to reconstruct subjets in the jet rest frame using a modified e^+e^- Cambridge jet reconstruction algorithm [20] in the FastJet 3.0.1 [16] package by replacing the distance parameter with a new choice of the distance parameter, Ω, where Ω is defined as the angle between two pseudoparticles in the jet rest frame. The algorithm performs a sequential recombination of the pair of psedoparticles that is closest in angle Ω, except for $\Omega > 0.8$. The reconstructed subjets are required to have energy $E_{\mathrm{subjet}} > 10$ GeV in the H jet rest frame. We then boost all the tracks associated with the H jet candidate back to the center-of-mass frame of the jet. A charged track is considered to be associated with a subjet only if their angular separation is less than 0.8 in the jet rest frame. By doing so, we separate the charged tracks that originate from different partons of the Higgs boson decay and reject many charged tracks from underlying events and pileup. This allows a straightforward identification of the b quarks inside the H jets by applying the existing b-tagging algorithms on the charged tracks associated with each subjets. In our analysis, we only retain the jets if their two subjets with the highest energy (leading subjet) have at least one associated charged track. Those two subjets are considered as the b and $\bar{b}$ subjet candidates of the H jet.

D. Double b-tagging

In this paper, we illustrate the double b-tagging in the jet rest frame with a tagging algorithm based on the charged track impact parameters since this algorithm is widely used in many experiments and is easy to implement. It is also among the official b-tagging methods used by the ATLAS experiment [21]. The impact parameters of tracks are computed with respect to the primary vertex
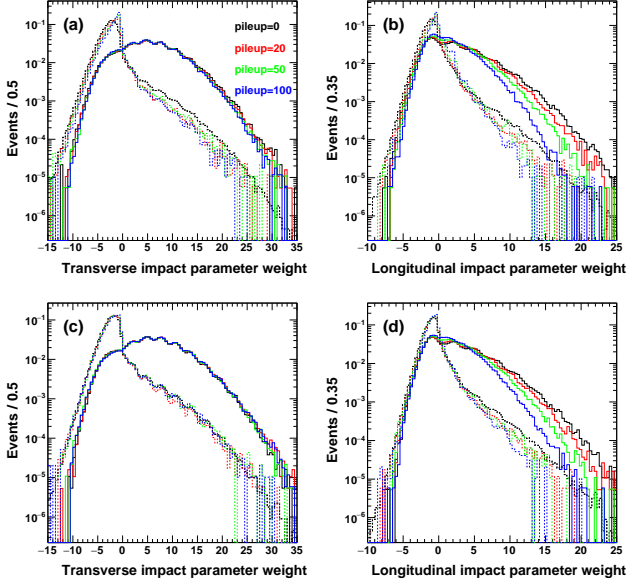
FIG. 2: The subjet weight distributions of the transverse and longitudinal impact parameter weight under different pileup conditions. In (a) and (b), the solid (dashed) lines represent the distributions of the charged tracks associated with the subjets that have the highest energy in the H (QCD) jet rest frame. In (c) and (d), the solid (dashed) lines represent the distributions of the charged tracks associated with the subjets that have the second highest energy in the H (QCD) jet rest frame. All the distributions are normalized to unity.
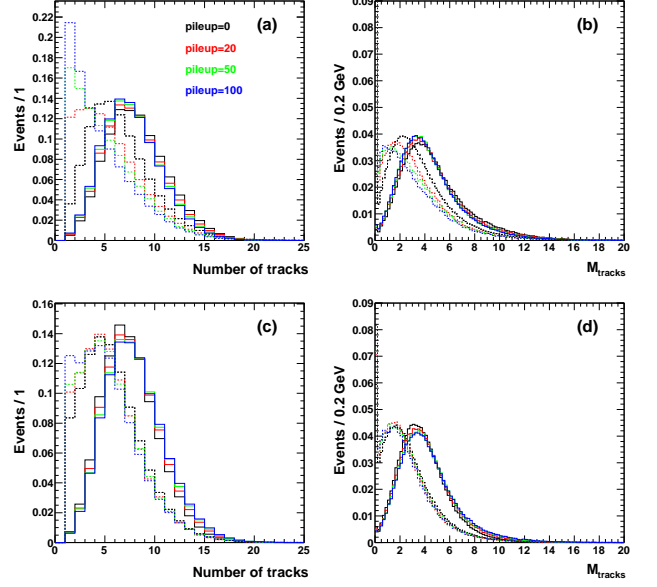
FIG. 3: The distribution of the number and invariant mass of the charged tracks that are associated with the subjets in the rest frame of the H (solid line) and QCD (dashed line) jets under different pileup conditions. In (a) and (b), the distributions are from the subjets with the highest energy in the jet rest frame. In (c) and (d), the distribution are from the subjets with the second highest energy in the jet rest frame. All the distributions are normalized to unity.

in the lab frame. They typically have significant nonzero values for the charged tracks from the b hadron decays because of its long lifetime. The impact parameter is signed to further discriminate the tracks from b-hadron decay from tracks originating from the primary vertex based on the fact that the decay position of the b hadron lies along its flight path. The sign of transverse impact parameter d_0 is determined using the subjet momentum $\vec{p}_{\mathrm{subjet}}$, the track momentum $\vec{p}_{\mathrm{trk}}$ at the point of the closest approach $\vec{x}_{\mathrm{trk}}$ [21] to the primary vertex position $\vec{x}_{\mathrm{pv}}$:

$$\mathrm{sign}(d_0) = (\vec{p}_{\mathrm{subjet}} \times \vec{p}_{\mathrm{trk}}) \cdot (\vec{p}_{\mathrm{trk}} \times (\vec{x}_{\mathrm{pv}} - \vec{x}_{\mathrm{trk}})). \quad (1)$$

The sign of longitudinal impact parameter z_0 is measured by the sign of $(\eta_{\mathrm{subjet}} - \eta_{\mathrm{trk}}) \times z_{0,\mathrm{trk}}$, where η_{subjet} is the pseudorapidity of the subjet, and η_{trk} and $z_{0,\mathrm{trk}}$ are the pseudorapidity and longitudinal impact parameters of the track at the position $\vec{x}_{\mathrm{trk}}$, respectively. All the quantities in the computation of the signed impact parameters are the ones defined in the lab frame.

We form a likelihood of the charged tracks associated with a subjet. The measured impact parameter significance S_i of the ith track in a subjet is compared to the predefined functions for both b subjet and non-b subjet hypothesis, $b(S_i)$ and $u(S_i)$, where $b(S)$ and $u(S)$ are the smoothed and normalized distributions of the charged tracks that are associated with the b subjets in the signal H jets and the subjets in the QCD jets, respectively. The

ratio of the probabilities $b(S_i)/u(S_i)$ defines a weight W_i. A subjet weight W_{subjet} is then calculated as the sum of the W_i from all the charged tracks associated with the subjet. In case there are no charged tracks associated with a subjet, its subjet weight is assigned to be zero. The distributions of the subjet weights of the H and QCD jets are shown in Fig 2. It shows a clear separation between the signal and background distributions.

To further help identify subjets that are originated from a b quark, we explore two additional properties of the subjet: the number and the invariant mass of the charged tracks that are associated with the subjet. Their distributions of the H and QCD jets are shown in Fig. 3.

The final double b-tagging variable is constructed using a boosted decision tree (BDT) algorithm with the subjet weights of the first two leading subjets in the jet rest frame, the numbers and invariant masses of the charged tracks associated with the first two leading subjets. The signal efficiency of H jets by identifying bottom quark and antiquark inside vs. the background rejectn of QCD jets for the BDT variable is shown in Fig 4. Note that the performance of the double b-tagging is slightly better with higher pileup at certain signal efficiencies. This is an effect that is caused by the selection of the jets used in the evaluation of the double b-tagging performance. In our analysis, we only use jets that have $p_{\mathrm{T}} > 300$ GeV, $40 \leq m_{\mathrm{jet}} \leq 250$ GeV and at least two subjets with
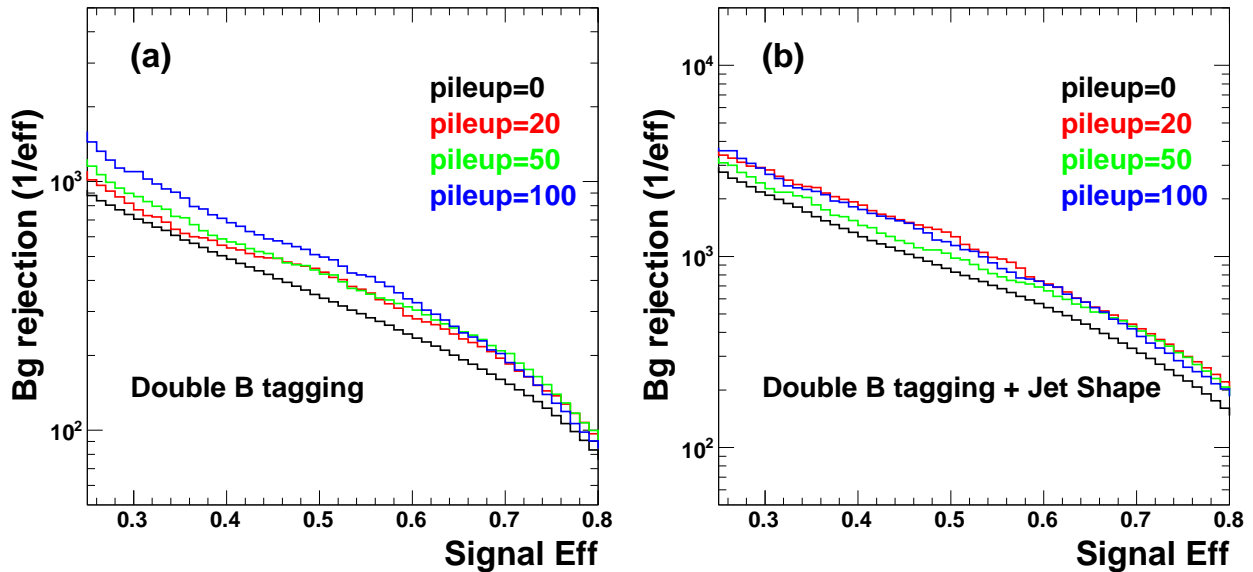
FIG. 4: The background rejection of QCD jets vs. the signal efficiency of H jets in different pileup conditions based on the double b-tagging without (a) and with (b) using the jet substructure information in its rest frame.

$E_{\text{subjet}} > 10$ GeV and nonzero charged tracks in its rest frame. As a result, when pileup increases, some QCD jets that otherwise would not pass the jet selection criteria can be selected. We repeat the studies of double b-tagging for jets selected with higher p_T requirement (up to 1 TeV) and observe no degrading of the performance. For a given signal efficiency, the background rejection is actually slightly higher for jets with higher transverse momentum. This is primarily due to the fact that the displaced decaying vertices of b hadrons in higher p_T jets are further away from the beamspot as they have larger Lorentz boosts.

E. Jet Substructure

The jet substructure information can be used to improve the identification of the boosted $H \to b\bar{b}$ in addition of the double b tagging. Here we demonstrate it by combining the double b-tagging with the jet substructure variables defined in the jet rest frame, introduced in Ref. [11]. They are: thrust, thrust-minor, Sphericity, Aplanarity, and Fox-Wolfram Moments R_2. Those variables are designed to identify a boosted two-body decay heavy particle whose final decaying products are reconstructed in a single jet. They have been successfully implemented by the ATLAS experiment to make the first observation of the boosted hadronically decaying vector boson reconstructed as a single jet from the SM W/Z+jets production [22]. In addition, we introduce another variable $\cos\Theta$, where Θ is defined as the angle between the direction of the thrust axis of a jet in its rest frame [11] and the jet momentum direction.

We form a BDT variable using the jet substructure variables described above with the variable used in the double b tagging in Sec. III D. Studies show that the jet substructure variables calculated based on the energy clusters have a great dependence on the pileup condition because of the additional energy depositions from the pileup and underlying events. In order to minimize the effect, the jet substructure variables in this paper are all computed using the charged tracks associated with the jet. Their distributions for the H jets and the QCD jets under different pileup conditions can be found in Appendix A. As shown in Fig. 4, the background rejection achieved by combining the double b-tagging and the jet substructure variables in the jet rest frame is a factor of 2 to 3 better compared to the one that only relies on the double b tagging.

IV. APPLICATION

In this section, we study two examples of the application of the double b tagging algorithm in searches for $H \to b\bar{b}$ in the VH production modes, where the W/Z boson decays leptonically. For simplicity, we only consider the kinematic region of the VH production where the Higgs boson has a relatively high transverse momentum so that its hadronically decay products can be reconstructed in a single jet. In both examples, we assume that the average pileup at the LHC is 50.

A. Search for $H \to b\bar{b}$ in the WH production mode

In this search channel, the leptonically-decaying W boson is reconstructed by requiring exactly one isolated lepton with $p_T > 20$ GeV, $|\eta| < 2.5$ and more than 25 GeV of missing transverse energy (E_T^{miss}) in an event. We then select jets with $p_T > 300$ GeV, $|\eta| < 1.7$ and $40 \leq m_{jet} \leq 240$ GeV in the event as the hadonically decaying Higgs boson candidates. The jet is reconstructed with the anti-k_T algorithm with a distance parameter $\Delta R = 0.8$. In order to reduce the large amount of background from the SM W+jets production, a selection on the BDT variable based on the double b tagging and the jet substructure information as described in Sec. III E is applied. We optimize the selection cut on the BDT variable by maximizing $S/\sqrt{B}$, where S and B are the numbers of the signal and background events within 20 GeV of the Higgs boson mass. In addition, we reject an event if it has a b jet that is not overlapping with the selected H jet candidate. This selection significantly reduces the background from the SM $t\bar{t}$ production. The jet mass distribution of the H jet candidates after all the event selection applied is shown in Fig. 5. The significance of $S/\sqrt{B}$ in the signal window is about 4 assuming $400\,\text{fb}^{-1}$ of LHC data at 14 TeV center-of-mass energy.

B. Search for $H \to b\bar{b}$ in the ZH production mode

In the search channel of the ZH production mode, the boosted $H \to b\bar{b}$ is reconstructed in a single jet that is based on the anti-k_T algorithm with a distance parameter $\Delta R = 0.8$. We require that the jets to have $p_T > 300$ GeV, $|\eta| < 1.7$ and $40 \leq m_{jet} \leq 200$ GeV. The Z boson is reconstructed in the final states of $Z \to \ell\ell$, ($\ell = e, \mu$) and $Z \to \nu\bar{\nu}$. Candidates of $Z \to \ell\ell$ decays are selected by combining isolated, oppositely charged pair of electron or muon tracks and require the deletion invariant mass to be within 20 GeV of the Z boson mass. The leptons are also required to have $p_T > 20$ GeV and $|\eta| < 2.5$. The identification of $Z \to \nu\bar{\nu}$ decays is done by requiring $E_T^{miss} \geq 300$ GeV and $\Delta\phi(E_T^{miss}, \text{jet}) > 3$, where $\Delta\phi(E_T^{miss}, \text{jet}) > 3$ is the azimuthal angle between the directions of the E_T^{miss} and the momentum of the selected H jet candidate. The events with a b jet that is not overlapping with the selected H jet candidate are rejected. After applying the above selection criteria, the dominant background left is the events from the SM Z+jets production, where $Z \to \ell\ell, \nu\bar{\nu}$ and the recoiled jet is misidentified as a H jet candidate. This background is greatly reduced by using the BDT variable based on the double b tagging and the jet substructure variables as described in Sec. III E. We optimize the selection cut on the BDT variable by maximizing the signal significance of $S/\sqrt{B}$.

The signal yield is extracted by a binned likelihood fit to the m_{jet} distribution of the selected H jet candidates, as shown in Fig. 6. The probability density func-
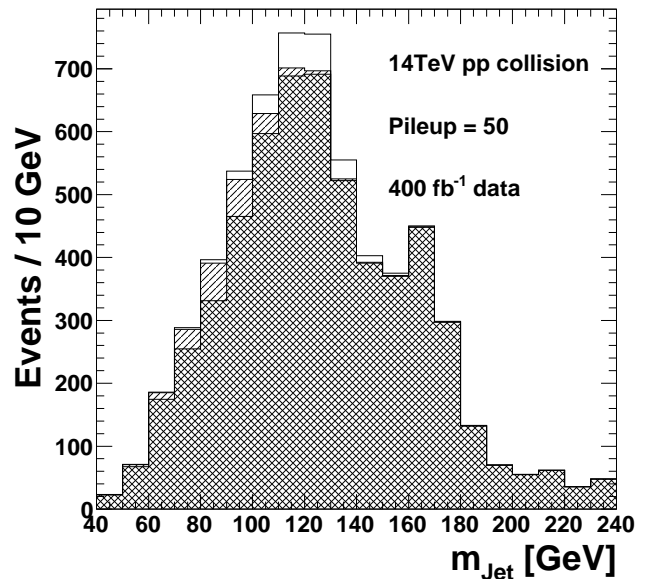

FIG. 5: Jet mass distribution of the selected H jet candidates in the MC simulated event sample that is equivalent to $400\,\text{fb}^{-1}$ of LHC data at 14 TeV center-of-mass energy after all the event selection criteria are applied. The open histogram represents the expected contribution of the WH signal events. The left hashed histogram represents the expected contribution of the WZ production, where $Z \to b\bar{b}$ is also reconstructed as a single jet. The right hashed histogram shows the expected background that is dominated by the top productions ($> 80\,\%$) with also significant contribution from the W+jets production. The peaking structure around 160 GeV is from the hadronically decaying top quark from the SM top production.

tions (PDF) of the $H \to b\bar{b}$ and $Z \to b\bar{b}$ are modeled as two Gaussian functions. The combinatorial background PDF is parameterized by a bifurcated Gaussian function that has different widths on the left and right side of the mean. The existence of the $Z \to b\bar{b}$ signal peak from the SM ZZ production in this search channel provides an excellent calibration sample to constrain the $H \to b\bar{b}$ PDF parameters. In actual data analysis at the LHC experiments, the parameters of the $Z \to b\bar{b}$ PDF can be precisely determined from data by studying the boosted hadronically decaying Z boson from the Z+jets production [22]. The background PDF can be also constrained using the events from the multijet production. In the default fit, the means of the Gaussian functions are allow to float with a constant difference that is fixed to the MC predicted mass difference between the Z and Higgs bosons. The widths of two Gaussian functions are set to the values predicted by MC simulation. The mean of the bifurcated Gaussian function is allow to be free in the fit, while the widths are fixed to the MC predicted values. The fit result for the MC simulated events sample that is equivalent to $400\,\text{fb}^{-1}$ of LHC data at 14 TeV center-
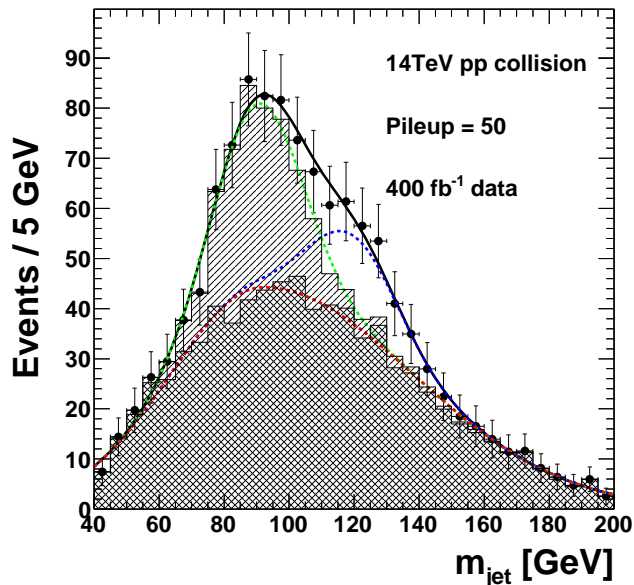
FIG. 6: Jet mass distribution of the selected H jet candidates in the MC simulated event sample that is equivalent to $400\,\mathrm{fb}^{-1}$ of LHC data at 14 TeV center-of-mass energy after all the event selection criteria are applied. The open histogram represents the expected contribution of the ZH signal events. The left hashed histogram represents the expected contribution of the ZZ production, where $Z \to b\bar{b}$ is also reconstructed in a single jet. The right hashed histogram shows the expected combinatorial background that is dominated by the Z+jets production. The solid back curve shows the final fit to the MC data. The dashed lines show each of the PDF components: signal (blue), ZZ production (green) and the combinatorial background (red).

of-mass energy is shown in Fig. 6. The fit yields more than $5\,\sigma$ of significance for both the $H \to b\bar{b}$ and $Z \to b\bar{b}$ signals.

V. CONCLUSION

In this paper we study the identification of a bottom quark-antiquark pair inside in a single jet with high transverse momentum by using the jet substructure in the center-of-rest frame of the jet. We demonstrate that the method can significantly reduce the QCD jet background while maintain a high identification efficiency of the boosted Higgs boson decaying to a $b\bar{b}$ pair even under a very large pileup condition. The study shows a good prospective on searches for $H \to b\bar{b}$ decay in the VH production mode for the LHC experiments at 14 TeV center-of-mass energy, and it is complementary to the existing searches [3, 4] in which each of the b quarks decayed from the Higgs boson is reconstructed as an indi-

vidual jet. The proposed technique can be also used to search for new physics phenomena beyond the SM, such as possible dark matter candidates produced in association with the SM Higgs boson [23].

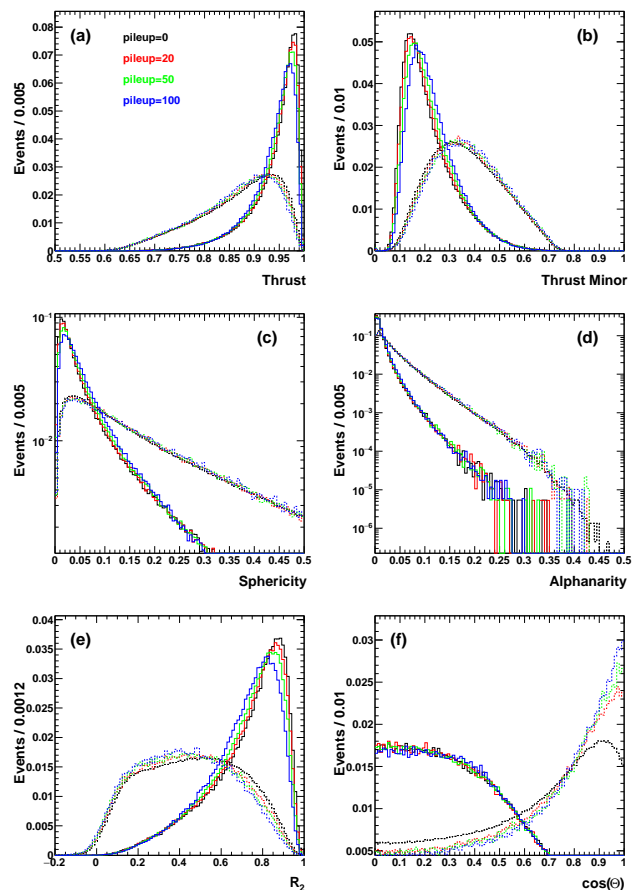## Appendix A: Jet substructure distribution

FIG. 7: The distribution of the jet shape variables in the center of mass frame of the jet: (a) Thrust, (b) Thrust-major, (c) Sphericity, (d) Aplanarity, (e) R_2 and (f) $\cos\Theta$ for the H jet signal and QCD jet background. All the distributions are normalized to unity.

[1] ATLAS Collaboration, Phys. Lett. B **716**, 1 (2012).

[2] CMS Collaboration, Phys. Lett. B **716**, 30 (2012).

[3] CMS Collaboration, Phys. Rev. D **89**, 012003 (2014).

[4] ATLAS Collaboration, J. High Energy Phys. **1501**, 069 (2015).

[5] J. M. Butterworth, A. R. Davison, M. Rubin and G. P. Salam, Phys. Rev. Lett. **100**, 242001 (2008).

[6] D. E. Soper and M. Spannowsky, JHEP **1008**, 029 (2010).

[7] J. H. Kim, Phys. Rev. D **83**, 011502 (2011).

[8] J. M. Butterworth, I. Ochoa and T. Scanlon, Eur. Phys. J. C **75**, 366 (2015).

[9] ATLAS Collaboration, ATL-PHYS-PUB-2014-013.

[10] ATLAS Collaboration, ATL-PHYS-PUB-2015-035.

[11] C. Chen, Phys. Rev. D **85**, 034007 (2012).

[12] C. Chen, Phys. Rev. D **87**, 074007 (2013).

[13] C. Chen, Phys. Rev. D **88**, 074009 (2013).

[14] T. Sjöstrand, S. Mrenna, P. Z. Skands, J. High Energy Phys. **0605**, 026 (2006).

[15] T. Sjostrand, S. Mrenna and P. Z. Skands, Comput. Phys. Commun. **178**, 852 (2008).

[16] M. Cacciari and G. P. Salam, Phys. Lett. **641**, 57 (2006).

[17] M. Cacciari, G. P. Salam and G. Soyez, J. High Energy Phys. **0804**, 063 (2008).

[18] [ATLAS Collaboration], J. of Instrumentation **3**, S08003 (2008).

[19] M. Cacciari and G. P. Salam, Phys. Lett. B **659**, 119 (2008).

[20] Y. L. Dokshitzer, G. D. Leder, S. Moretti and B. R. Webber, J. High Energy Phys. **9708**, 001 (1997).

[21] G. Aad *et al.* [ATLAS Collaboration], arXiv:0901.0512 [hep-ex].

[22] [ATLAS Collaboration], New J. Phys. **16**, 113013 (2014).

[23] L. Carpenter, A. DiFranzo, M. Mulhearn, C. Shimmin, S. Tulin and D. Whiteson, Phys. Rev. D **89**, 075017 (2014).