First look at boosted $\mathbf{Z \rightarrow b\bar{b}}$ in Run-3 data

The CMS Collaboration

Afroditi Kyprianou

Supervisors: Dr. Santeri Laurila and Dr. Giovanni Petrucciani

August 2023

Abstract

Large-radius jets originating from hadronic decays of highly Lorentz-boosted objects such as W, Z, or Higgs bosons provide a powerful tool for a variety of physics analyses. In this project, the performance of such boosted objects is investigated using the Z boson as a "standard candle", analyzing proton-proton collision data that were collected in 2022 at the record-high center-of-mass energy of 13.6 TeV. The data sample corresponds to an integrated luminosity of 34.2 fb^{-1}. A sample of Z bosons decaying to a pair of b quarks is selected, and used for comparisons between simulated event samples and the recorded data. Highly Lorentz-boosted Higgs bosons decaying to $b\bar{b}$ are reconstructed as single large-radius jets, and are identified using jet substructure and a dedicated b tagging technique based on a deep neural network.

1 Introduction

The Z boson is a neutral elementary particle and like its electrically charged cousin, the W boson, the Z boson carries the weak force. Measurements of Z-boson production in pp collisions constitute an important test of the Standard Model (SM), since they allow the electroweak sector to be precisely probed. As the title suggests, the focus is on high energy collisions producing a Z boson, which decays to a pair of bottom quarks. The simplest description of this process is the so-called leading order (LO) process. In Fig. 1 a LO Feynman diagram is shown.

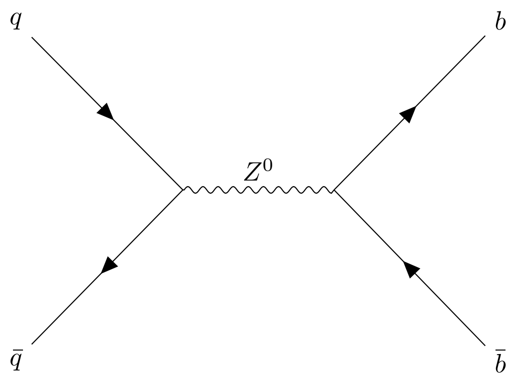

Figure 1: Leading order Feynman diagram for $Z \to b\bar{b}$ process

This decay is important for searches in final states with a quark pair. The b quarks in the decay are identified by detecting the jets they form when hadronizing. It is very hard to look directly for the b-jets due to the large number of background multi-jet events originating from quantum chromodynamics (QCD).

We want to estimate the expected sensitivity for Higgs boson decays in b quarks pairs. The inclusive search for the $Z \to b\bar{b}$ decay at the LHC is of great interest, since it enables full validation of the analysis procedure used to target Higgs bosons decaying to $b\bar{b}$. The measurement of the Higgs boson coupling to b quarks is an important test of the Standard Model.

There are four main ways that a Higgs boson can be produced: gluon-gluon fusion, vector-boson fusion, associated production with top quarks, and Higgs-strahlung, also known as associated production with a vector boson. The first two are the two dominant Higgs boson production modes at the LHC. Feynman diagrams of these two process are shown in Fig. 2.

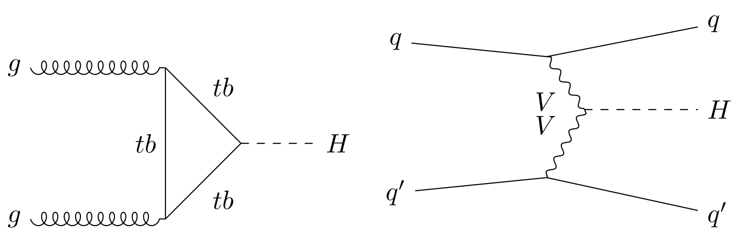

Figure 2: The left diagram is the gluon-gluon fusion and the right one the vector-boson fusion. These two are the dominant Higgs boson production modes at the LHC. Taken from [1]

2 Analysis

2.1 Jet reconstruction

In this project the jets referred are AK8 PUPPI [2] jets. AK8 jets are the jets clustered with the anti-kT algorithm [3] with a distance parameter of R=0.8. The anti-kT clustering algorithm finds two particles in the event that are closest in some distance measure: $\Delta R\ \min(p_T^1, p_T^2)^a$, where $a = -1$ for the anti-kT algorithm.

2.2 Jet mass estimation

We mainly use the ParticleNet algorithm [4] to estimate the mass of the the jets, but also the Soft Drop algorithm [5] to compare the two results. The ParticleNet algorithm is a graph neural network-based particle identification algorithm for identifying hadronic decays of Highly Lorentz-boosted quarks, W, Z, and Higgs bosons and classifying different decay modes, such as $Z \rightarrow b\bar{b}$. ParticleNet can also be used to regress the jet mass, as detailed in [6]. The Soft Drop declustering, with angular exponent β = 0 and soft radiation fraction z = 0.1 is applied to the Z boson jet candidate to remove soft and wide-angle radiation.

2.3 B-tagging

B-tagging of the jet means assigning probabilities that it originates from a process with b quarks as final state, which is done with ParticleNet. We define:

$$D_{b\bar{b}} = \frac{\mathrm{P(Xbb)}}{\mathrm{P(Xbb) + P(QCD)}} \tag{1}$$

as the XbbVsQCD score of the jets, where $P(Xbb)$ and $P(QCD)$ are the output scores of the ParticleNet jet multiclassifier.

2.4 Selection Criteria

As mentioned, Z bosons decaying to $b\bar{b}$ are reconstructed as single large-radius jets, and are identified using jet substructure and a dedicated b tagging technique based on a deep neural network. The candidate jet is required to have $p_T > 450$ GeV to satisfy restrictive trigger requirements that suppress the large background from jets produced via the strong interaction, referred to as quantum chromodynamics (QCD) multijet events. Also, the jet is required to have pseudorapidity in the range $|\eta| < 2.4$ and ParticleNet mass, $m_{PN} > 40$ GeV. We require to have at least one jet with these selections. We compare the $D_{b\bar{b}}$ scores of the selected jets and we search for the jet with the highest score. The goal is to minimize the QCD background compared to our signal.

2.5 Analysis procedure

In this entire section, Z candidates are only considered, if the jets pass the isolation and identification cuts described in Section 2.4. At first we started the analysis of data Vs Monte Carlo simulations with no requirement on b-tagging. The resulting mass distribution is binned from 0 to 200 GeV with a bin width of 20 GeV. The mass plot we got is shown in Fig. 3:

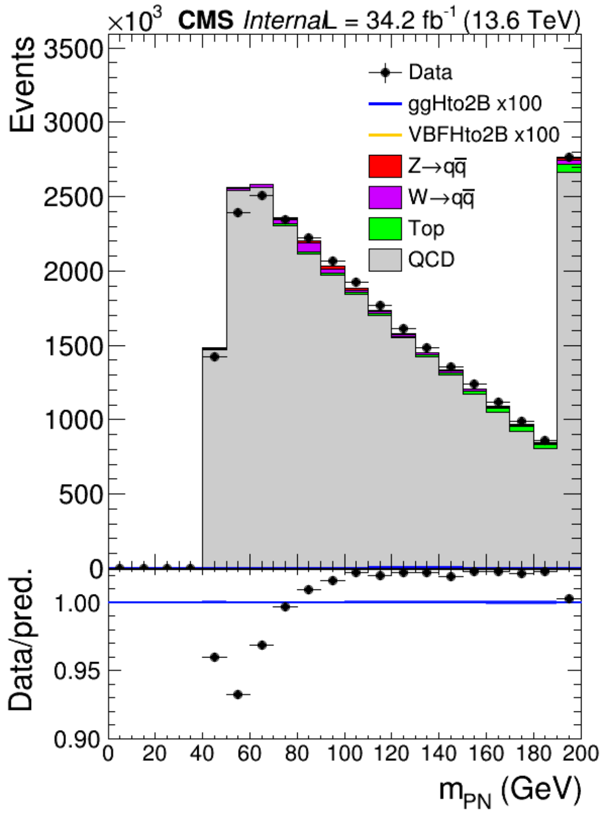

Figure 3: We present the ParticleNet mass distribution of the jet with the highest $D_{b\bar{b}}$. The bottom panel shows the difference between the data and the total background prediction, divided by the statistical uncertainty in the data.

As we can see, there's a lot of QCD multijet background that we need to minimize in order to observe our signal.

Moreover, the p_T distribution plots are shown in Fig. 4.

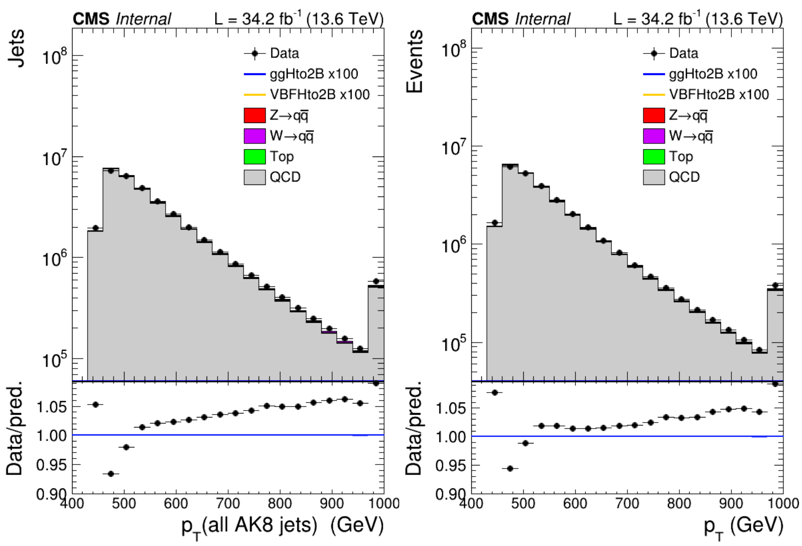

Figure 4: We present the jets' p_T distribution. On the left we show the p_T distribution of all the jets and on the right the distribution for the jet with the highest $D_{b\bar{b}}$. The bottom panel shows the difference between the data and the total background prediction, divided by the statistical uncertainty in the data.

As mentioned, we have to use the b-tagging technique in order to minimize our QCD background. In the b-tagging distribution plot (Fig. 5) we can see the QCD background dominating over the signal.

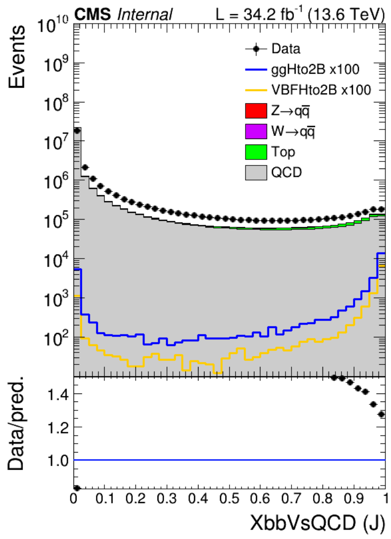

Figure 5: We present the b-tagging distribution plot of the jet with the highest $D_{b\bar{b}}$, which is considered as the $Z \to b\bar{b}$.

Then, we placed a stringent requirement on b-tagging and noticed the peak of $Z \to b\bar{b}$ becoming clearer. The jet with the highest bb-tagging score is selected as the boosted $Z/H \to b\bar{b}$ candidate jet. With $D_{b\bar{b}} > 0.99$ now the plots we get are as shown in Fig. 6:

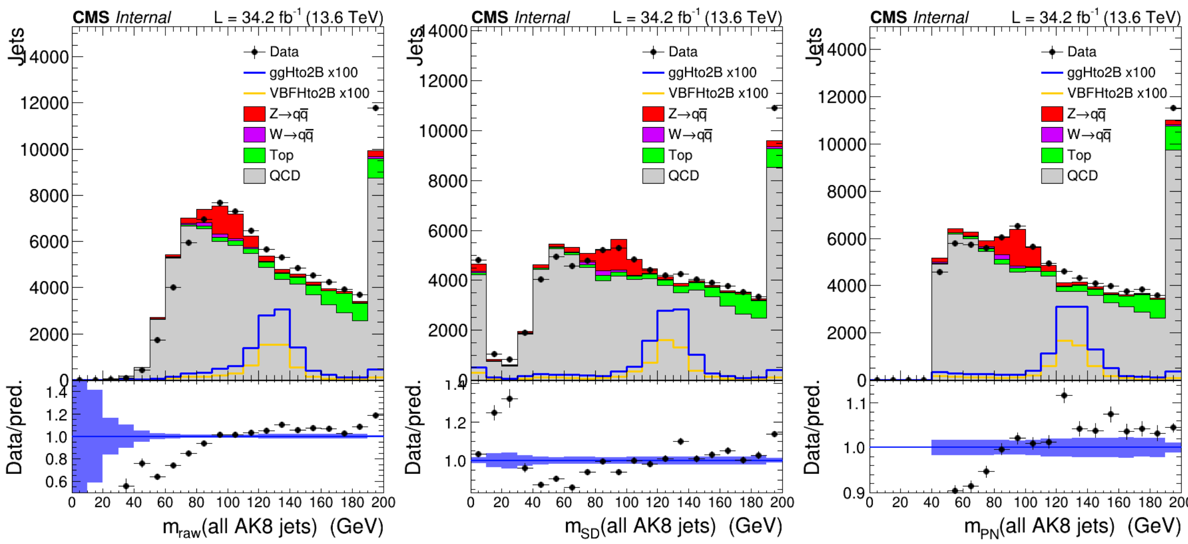

Figure 6: We present the jets' mass distribution. On the left we show the raw mass, the middle the Soft Drop mass and on the right the ParticleNet mass. The bottom panel shows the difference between the data and the total background prediction, divided by the statistical uncertainty in the data.

From the plots in Fig. 6 we can compare the two distributions from the two algorithms and conclude that the ParticleNet algorithm gives a clearer, sharper peak of our signal.

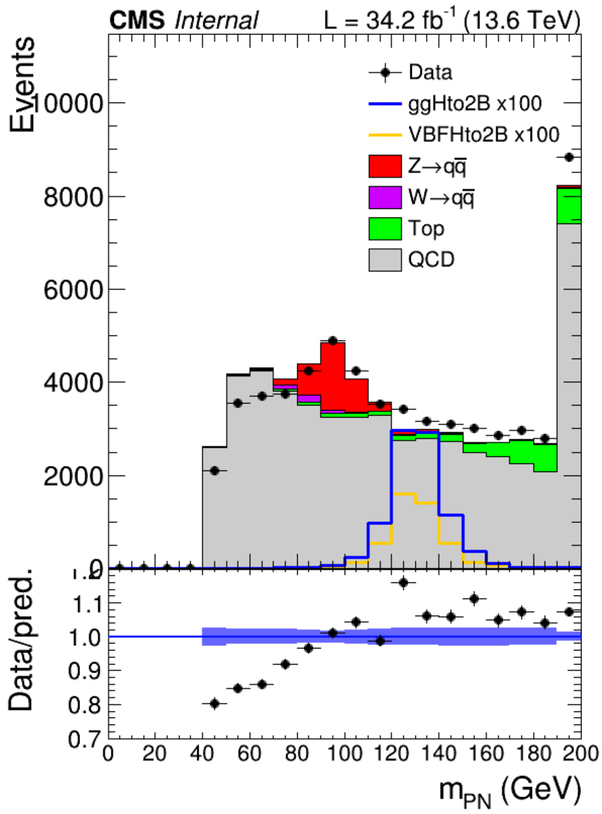

Figure 7: We present the ParticleNet mass distribution of the jet with the highest $D_{b\bar{b}}$. The bottom panel shows the difference between the data and the total background prediction, divided by the statistical uncertainty in the data.

Moreover, the p_T distribution plots are shown in Fig. 8.

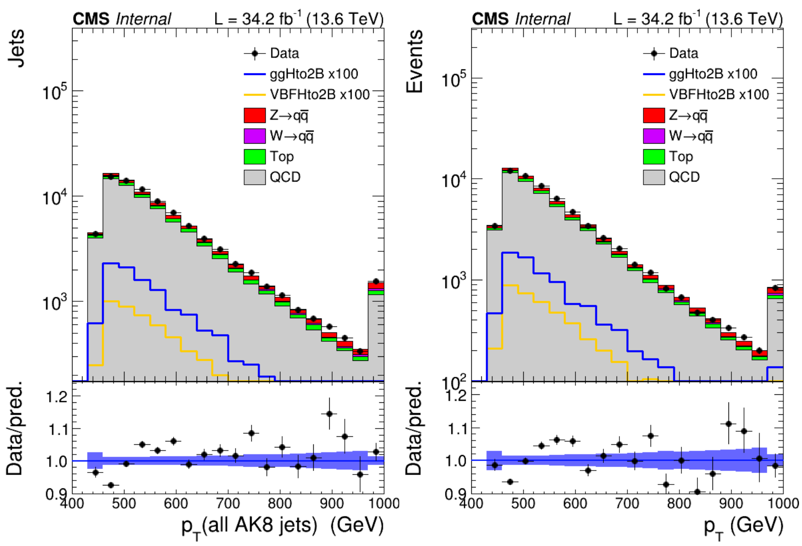

Figure 8: We present the jets' p_T distribution. On the left we show the p_T distribution of all the jets and on the right the distribution for the jet with the highest $D_{b\bar{b}}$. The bottom panel shows the difference between the data and the total background prediction, divided by the statistical uncertainty in the data.

As we notice in the plot of the ParticleNet mass distribution of the jet with the highest $D_{b\bar{b}}$ (Fig. 7), we observe the $Z \to b\bar{b}$ decay, since the corresponding peak is shown clearly. In Fig. 7 we can see a trend in Data/pred., especially in the first few bins. The difference is the highest there, because that region is very pure in QCD, and this suggests that the modeling of the QCD multijet background in simulation is not sufficiently good. Therefore, the next step is to replace this simulated background with a more accurate data-driven estimate.

2.6 Estimation of QCD multijet background

Events where the selected AK8 jet passes the ParticleNet constitute the “passing” or "signal" region, while events failing ParticleNet form the “failing”, or "control" region.

The "control" region ($D_{b\bar{b}} < 0.9$) is used to estimate the QCD multijet background in the signal region ($D_{b\bar{b}} > 0.99$), so we will now compare the different results of b-tagging. The corresponding plot is shown in Fig. 9:

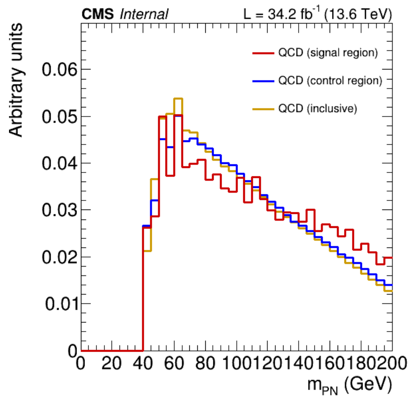

Figure 9: We present the ParticleNet mass distribution of QCD process of the jet with the highest $D_{b\bar{b}}$ with the different b-tagging. The red line is the QCD for the "signal" region, meaning $D_{b\bar{b}} > 0.99$, the blue line is the QCD for the "control" region, meaning $D_{b\bar{b}} < 0.9$ and the orange line is the QCD used for the comparison. This plot is a shape comparison (each curve is normalized to unity) and it shows simulated QCD events.

Now, for the control region the b-tagging distribution plot is shown in Fig. 10:

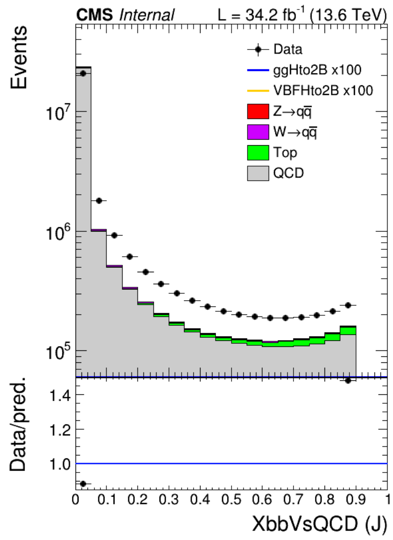

Figure 10: We present the b-tagging distribution plot of the jet with the highest $D_{b\bar{b}}$ for the control region. The bottom panel shows the difference between the data and the total background prediction, divided by the statistical uncertainty in the data

To be able to estimate the QCD background we plotted the ratio of the "signal" to the "control" region using three different p_T categories: the jet with the highest $D_{b\bar{b}}$ to have $450 < p_T < 550$ GeV, $550 \leq p_T < 700$ GeV and $p_T \geq 700$ GeV as shown in Fig. 11:

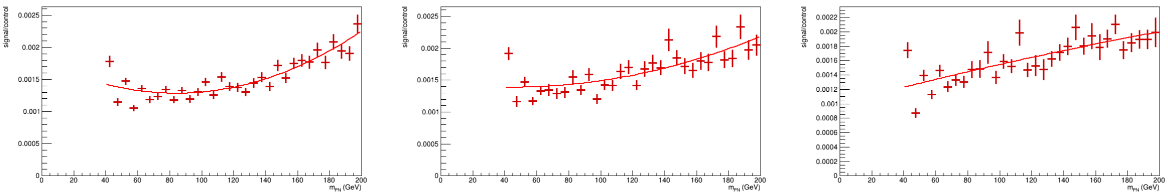

Figure 11: We present the ratio of the "signal" to the "control" region using three different p_T cuts of the jet with the highest $D_{b\bar{b}}$. The left is the one corresponding to $450 < p_T < 550$ GeV. The middle one is the one corresponding to $550 \leq p_T < 700$ GeV and the right one $p_T \geq 700$ GeV.

A binned fit is performed on these ratios with second-order polynomial.

Fig. 9 shows that the mass shape is different between the "signal" region and the "control" region, thus the transfer function must depend on the jet mass to account for this effect.

3 Results

We use the fitting parameters as a transfer function for adding weight to the QCD background shapes obtained from the control region and we get the corresponding plots of the mass distribution of the jet with the highest $D_{b\bar{b}}$:

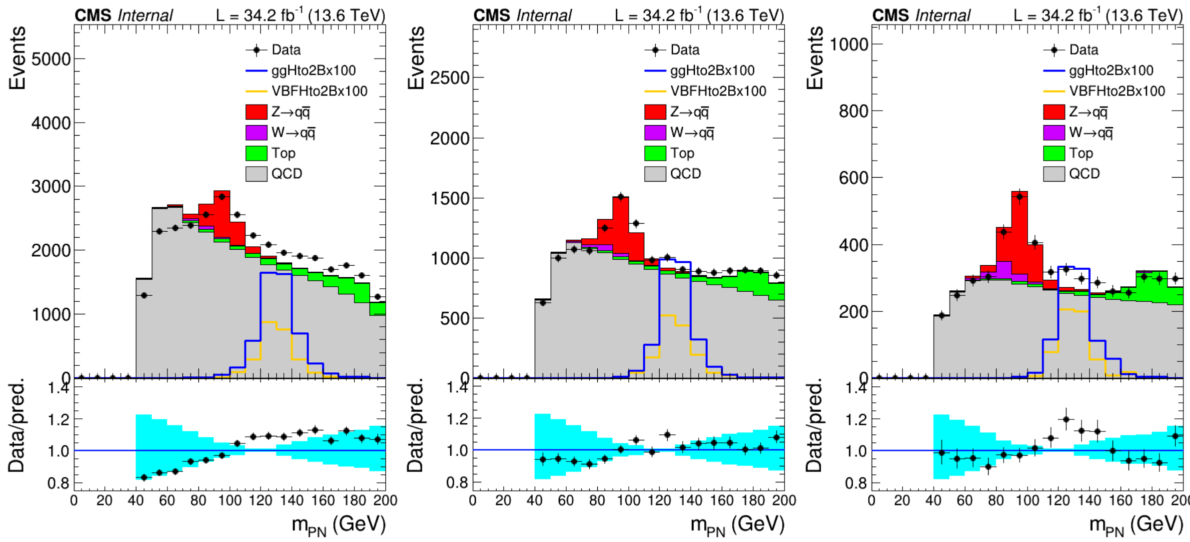

Figure 12: We present the ParticleNet mass distributions of the jet with the hightest $D_{b\bar{b}}$ using the three mentioned p_T cuts. The left is the one corresponding to $450 < p_T < 550$ GeV. The middle one is the one corresponding to $550 \leq p_T < 700$ GeV and the right one $p_T \geq 700$ GeV. The bottom panel shows the difference between the data and the total background prediction, divided by the statistical uncertainty in the data.

The shape of the transfer function is found to depend on p_T, so the three categories are defined according to jet p_T, and the background estimation is performed separately in each category.

In summary, the control region ($D_{b\bar{b}} < 0.9$) is used to obtain the mass shape, and then it is normalized to $D_{b\bar{b}} > 0.99$ signal region with MC-based transfer factors in order to get our final results. In Fig. 12 we can see that the data and prediction agree within the statistical uncertainty, but there is a trend in the region of $40 - 80$GeV, where there seems to be a systematic effect. This systematic effect may originate from the fact that the transfer functions come from simulation and not from data, and thus they may not be fully accurate.

Binned maximum likelihood fit was performed simultaneously in the three categories using as input the jet mass distributions for each process from Fig. 12, with systematic uncertainties included as nuisance parameters. For each process, 30% normalization uncertainties are assigned to cover for various sources of theoretical and experimental systematic uncertainties (such as uncertainties on ParticleNet tagging efficiency). Fit uncertainties in the data-driven QCD estimate included as a shape uncertainty.

The fitted signal strength is found to be: $\mu_Z = 1.05^{+0.32}_{-0.25}$, where the uncertainties correspond to 68% confidence level.

Expected sensitivity for $H \to b\bar{b}$ (with ggF and VBF combined), obtained using the Asimov dataset, corresponds to the significance of 0.9 standard deviations.

4 Summary

An inclusive search for the Z boson decaying to a bottom quark-antiquark pair and reconstructed as a single large-radius jet with transverse momentum $p_T > 450$ GeV has been presented. The search uses a data sample of proton-proton collisions at $\sqrt{s} = 13.6$ TeV, corresponding to an integrated luminosity of 34.2 fb^{-1}. The production of a Z boson and jets is used to validate the analysis procedure that can later be used for the Standard Model (SM) Higgs boson decay to a bottom quark-antiquark pair.

We were able to observe the $Z \to b\bar{b}$ decay with the improved b-tagging technique and estimate the QCD multijet background with the various p_T categories. We calculated $\mu_Z = 1.05^{+0.32}_{-0.25}$ and the exepected significance for $H \to b\bar{b}$ decay: 0.9σ.

References

[1] Armen Tumasyan et al. "A portrait of the Higgs boson by the CMS experiment ten years after the discovery". In: *Nature* 607.7917 (2022), pp. 60–68. DOI: `10.1038/s41586-022-04892-x`. arXiv: `2207.00043 [hep-ex]`.

[2] Daniele Bertolini et al. "Pileup Per Particle Identification". In: *JHEP* 10 (2014), p. 059. DOI: `10.1007/JHEP10(2014)059`. arXiv: `1407.6013 [hep-ph]`.

[3] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The anti-k_t jet clustering algorithm". In: *JHEP* 04 (2008), p. 063. DOI: `10.1088/1126-6708/2008/04/063`. arXiv: `0802.1189 [hep-ph]`.

[4] Huilin Qu and Loukas Gouskos. "ParticleNet: Jet Tagging via Particle Clouds". In: *Phys. Rev. D* 101.5 (2020), p. 056019. DOI: `10.1103/PhysRevD.101.056019`. arXiv: `1902.08570 [hep-ph]`.

[5] Andrew J. Larkoski et al. "Soft Drop". In: *JHEP* 05 (2014), p. 146. DOI: `10.1007/JHEP05(2014)146`. arXiv: `1402.2657 [hep-ph]`.

[6] "Mass regression of highly-boosted jets using graph neural networks". In: (2021). URL: `https://cds.cern.ch/record/2777006`.

[7] "Identification of highly Lorentz-boosted heavy particles using graph neural networks and new mass decorrelation techniques". In: (2020). URL: `https://cds.cern.ch/record/2707946`.

[8] "Calibration of the mass-decorrelated ParticleNet tagger for boosted $b\bar{b}$ and $c\bar{c}$ jets using LHC Run 2 data". In: (2022). URL: `https://cds.cern.ch/record/2805611`.

[9] *Search for boosted Higgs bosons produced via vector boson fusion in the $H \to b\bar{b}$ decay mode using LHC proton-proton collision data at $\sqrt{s} = 13$ TeV*. Tech. rep. Geneva: CERN, 2023. URL: `https://cds.cern.ch/record/2866501`.