Summer Student Project Report

First look at new Run3 VHcc analysis

Author:
Tristan Brandes
Karlsruhe Institute of Technology

Supervisors:
Loukas Gouskos
Huilin Qu

in the

CMS Group *at the*
European Organization for Nuclear Research (CERN)

August 22, 2023

Contents

1 Introduction

Understanding the building blocks of our universe is the driving force behind the research conducted at the Large Hadron Collider (LHC) at CERN. My role as a Summer Student within the CMS physics analysis group was to take a first look at the new Run3 VHcc/bb analysis.

The study of the interactions, or couplings, of the Higgs boson with other particles, provide crucial insights into the fundamental nature of matter. Notably, the relationships between the Higgs boson and second-generation quarks, such as the charm quark, are still being explored. This uncharted territory has become a significant area of interest as derivations from the values predicted by the Standard Model (SM) can be an indicator of physics beyond SM.

The primary goal of the Run 3 version of the analysis, is among other developments, to extend the Run2 VHcc analysis, to simultaneously investigate the Higgs-charm and Higgs-bottom couplings. A notable advantage of this approach is that while we focus on quantifying the $\mathrm{H} \rightarrow \mathrm{cc}$ coupling, the $\mathrm{H} \rightarrow \mathrm{bb}$ one is not restricted to its SM value. Instead, as illustrated in Figure 1.1, it is treated as a variable parameter. This strategic approach enhances our sensitivity to potential BSM physics.

This project report outlines the initial stages of this analysis, which aimed to comprehend and assess the background related to $\mathrm{H} \rightarrow \mathrm{bb}$ in the different analysis channels.

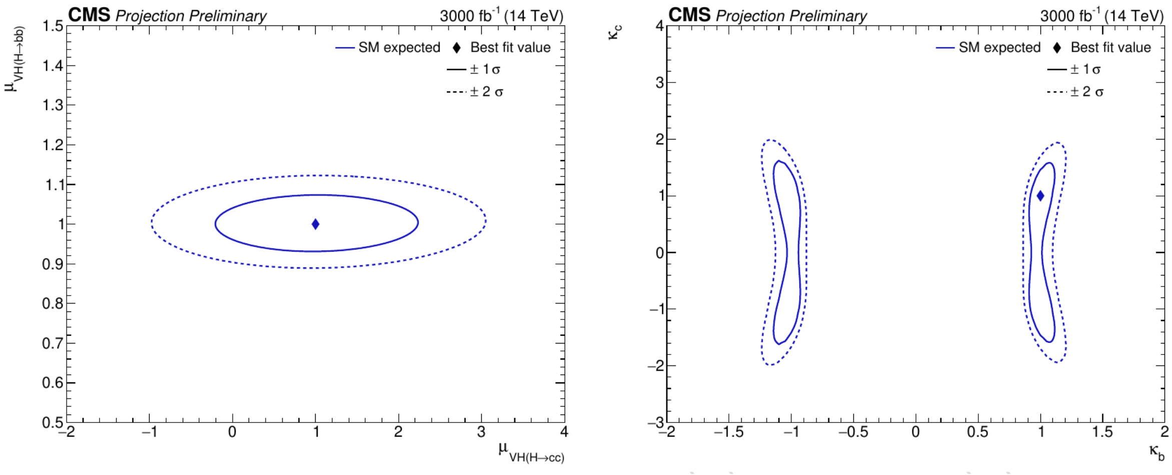


FIGURE 1.1: Profile likelihood scans are depicted as functions of $\mu_{\mathrm{VH(H \rightarrow c\bar{c})}}$ and $\mu_{\mathrm{VH(H \rightarrow b\bar{b})}}$ on the left, while κ_b and κ_c are illustrated on the right. The SM-expected best fit point is marked in black. Expected contours for one and two standard deviations are represented by solid and dashed lines, respectively.

2 Project Work

The analysis is conducted in 3 channels, which are distinguished by the leptonic decay modes of the vector boson as illustrated in Figure 2.1.

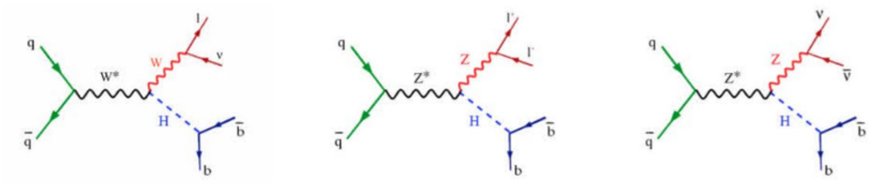


FIGURE 2.1: Illustration of various leptonic decay channels in VH production. From left to right: 1l-channel, 2l-channel, and 0l-channel.

The following plots were generated through a systematic application of cuts and selections. The analysis target Higgs bosons produced at moderate to high transverse momenta (they are "boosted"). The Higgs decay products are reconstructed as a single large-radius jet, with distance parameter $R = 1.5$. The baseline selection involved the implementation of both a baseline selection and an enhanced, boosted selection strategy, along with supplementary criteria tailored to specific channels. The baseline selection necessitated a minimal transverse momentum p_T of 150 GeV for the vector boson and 200 GeV for the Higgs candidate.

2.1 2l-Channel

In the 2l-channel, a discernible background peak originating from the Z boson is prominently present at 90 GeV, accompanied by an overlapping Z+jets background. These distinctive backgrounds, emanating from known sources, stand apart from the signal and thus pose no significant concerns.

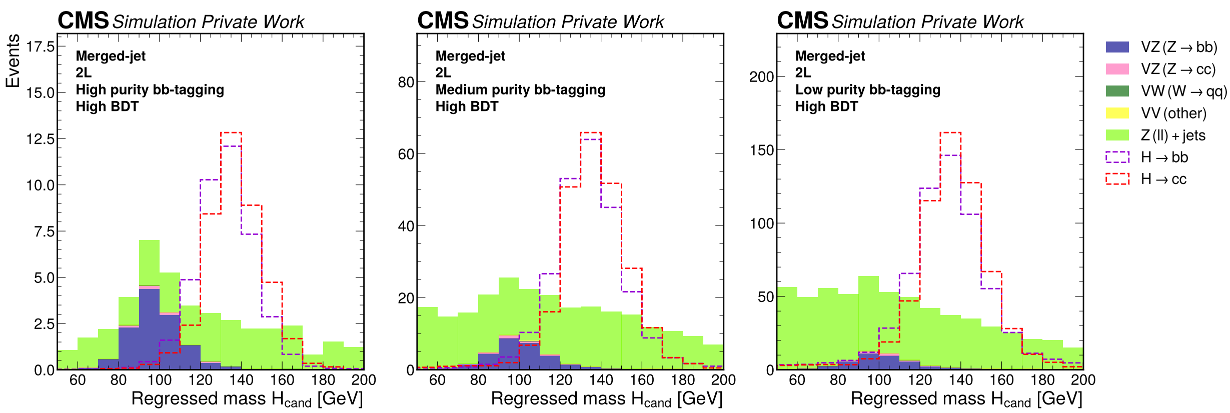
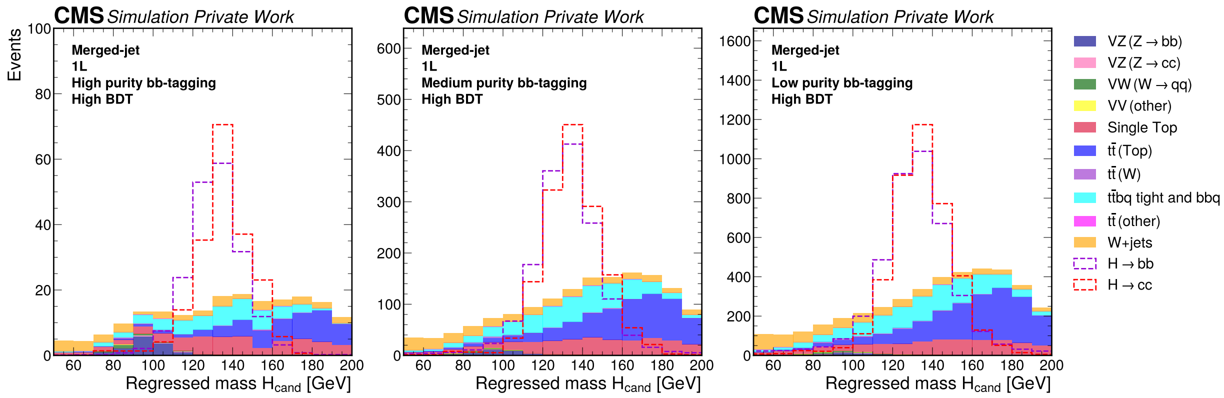
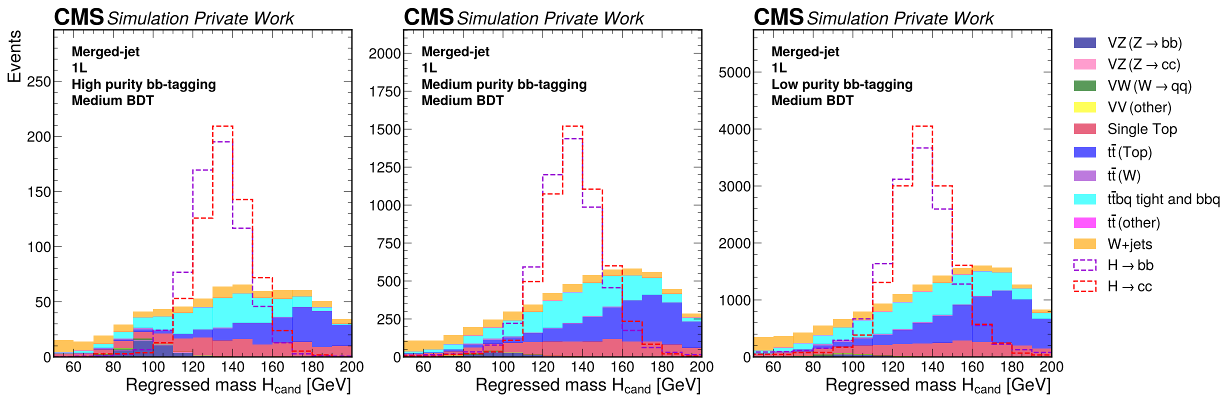


FIGURE 2.2: The Higgs candidate mass distribution for background (filled histograms) and $H \to cc$, $H \to bb$ signal events (dashed lines) in the 2l-channel. The $H \to cc$, $H \to bb$ signal is normalized to the total bkg.



(A) High BDT > 0.7 selection.



(B) Medium BDT > 0.55 selection.

FIGURE 2.3: The Higgs candidate mass distribution for background (filled histograms) and H → cc, H → bb signal events (dashed lines) for the 1l-channel. The H → cc, H → bb signal is normalized to the total bkg.

2.2 1l-Channel

As depicted in Figure 2.3, the 1l-channel presents a notable challenge posed by the top background. Several factors contribute to this challenge, with a primary reason being the strong resemblance between the decay of the top quark, illustrated in Figure 2.4, and the Higgs boson (H → bb) signal. Since the top quark predominantly decays into b quarks, distinguishing between H → bb and $t \rightarrow Wb \rightarrow q\bar{q}b$ becomes inherently intricate for the tagger.

Hence, it is crucial to gain a comprehensive understanding of the distinct contributions from $t\bar{t}$ background. To achieve this, a decomposition of the $t\bar{t}$ background into various subcategories has been undertaken. These subcategories are showcased in Figure 2.5. This breakdown facilitates a more nuanced comprehension of the individual components comprising the $t\bar{t}$ background.

Furthermore, it's essential to acknowledge that large-R jets with varying flavor content and substructure manifest divergent responses to the ParticleNet tagger. Consequently, they might require different simulation-to-data scale factors.

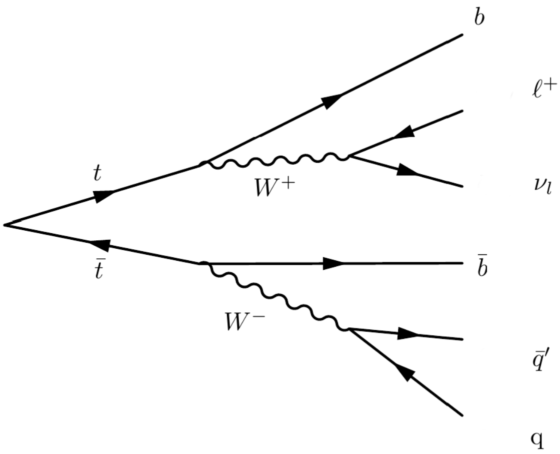
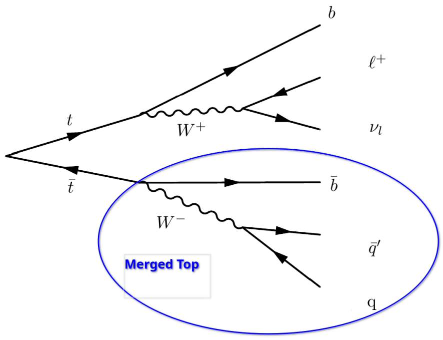
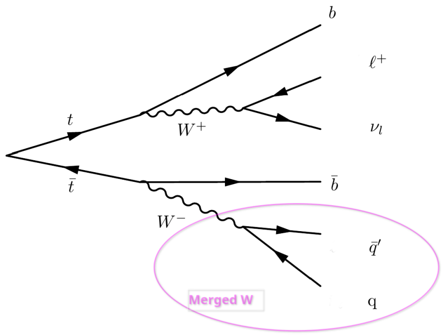


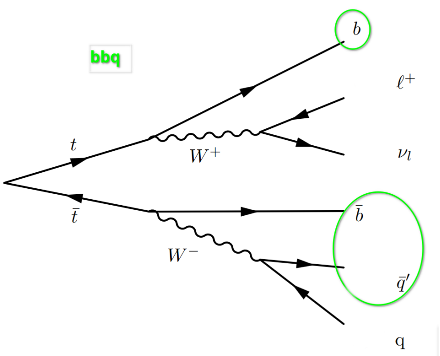
FIGURE 2.4: This Feynman Diagram illustrates the Top decay.



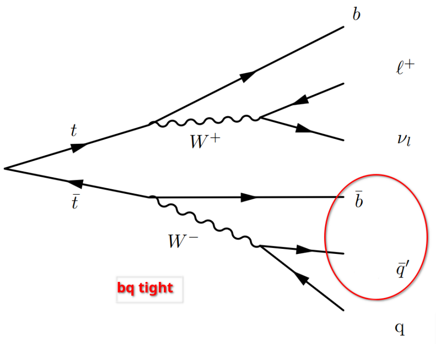
(A) Illustration of the ak15 Merged Top selection.



(B) Illustration of the ak15 Merged W selection.



(C) Illustration of the ak15 bbq selection.

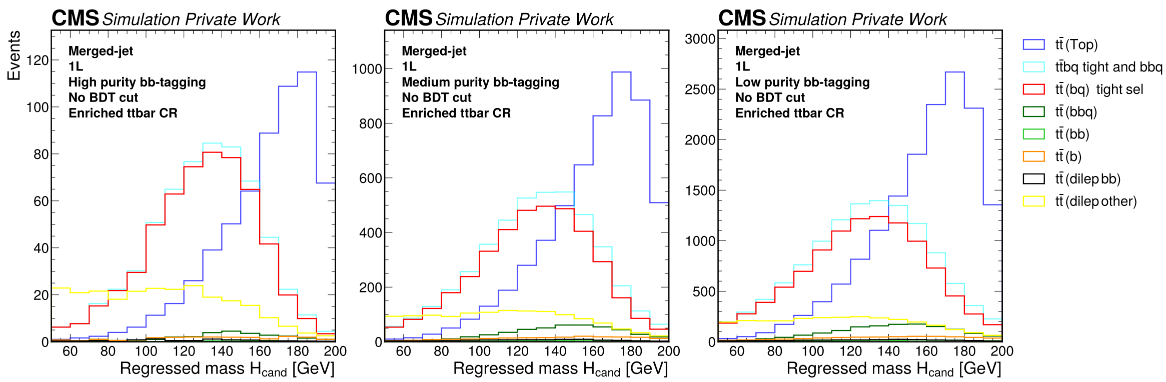


(D) Illustration of the ak15 bq tight selection.

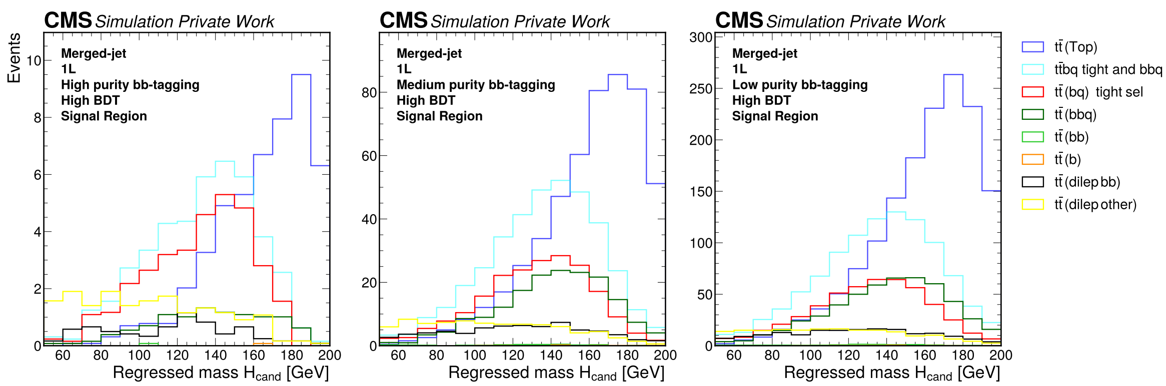
FIGURE 2.5: Shown in this figure are various selection categories applied to the ak15 jet in order to distinguish between the different ttbar background fractions.

2.2.1 Absolute yields

As demonstrated in Figure 2.3, it becomes apparent that the $t\bar{t}$ (Top) background does not present a notable concern. This is primarily due to its peak manifesting around the top mass value of approximately $m_{\mathrm{Top}} \approx 170\,\mathrm{GeV}$, a considerable distance from the signal region.
The $t\bar{t}$ (W) background exhibits both small magnitude and a distinctive peak centered at $m_{\mathrm{W}} \approx 80\,\mathrm{GeV}$. However, the $t\bar{t}$ (bq tight) and $t\bar{t}$ (bbq) backgrounds emerge as challenging elements. These backgrounds sustain a consistent contribution spanning from 100 GeV to 180 GeV, thereby presenting a significant background presence around the peak of the Higgs mass.
For further investigation, the absolute yields of the above mentioned $t\bar{t}$ components were examined for the signal and enriched $t\bar{t}$ control region. The control region is later utilized to derive simulation-to-data correction factors. Therefore, it is desirable that the overall shape and relative fractions between contributions for signal and control regions do not change.
However, as illustrated in Figure 2.6, a notable shift in the relative fraction between the bq tight and bbq components becomes apparent when transitioning from the control region (Figure 2.6a) to the signal region (Figure 2.6b). Specifically, within the control region (Figure 2.6a), the bbq contribution (depicted in dark green) is notably minor in comparison to the bq tight selection (depicted in red). Intriguingly, this balance alters in the signal region, where the bbq contribution escalates to match the same relative fraction as the bq tight selection.



(A) $t\bar{t}$ Control Region with $N_{\mathrm{jets}} \geq 2$ and no BDT cut applied.



(B) Signal Region with $N_{\mathrm{jets}} < 2$ and High BDT cut applied.

FIGURE 2.6: Absolute yields of the different $t\bar{t}$ bkg. components in the signal (lower) and control (upper) region.

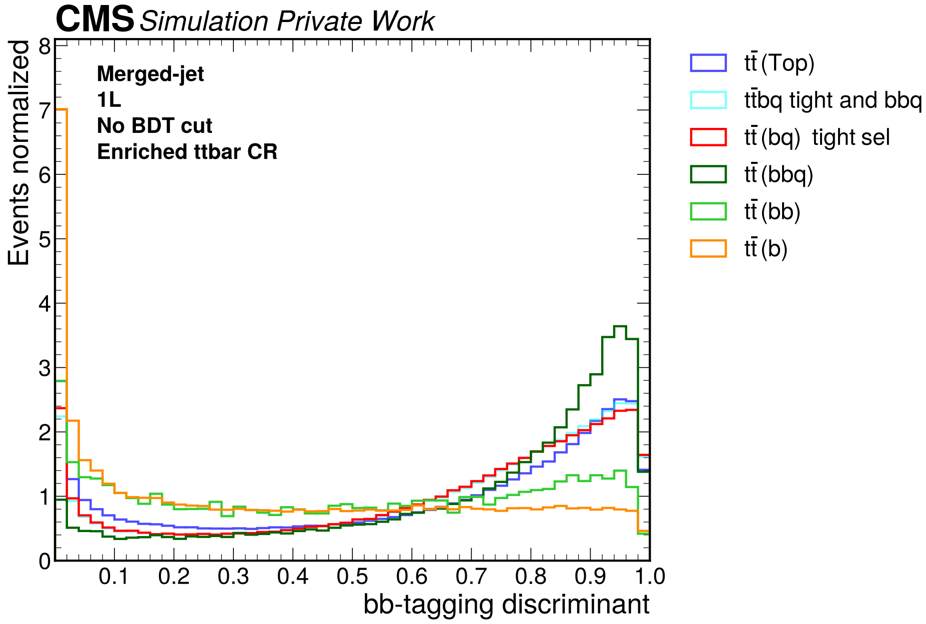
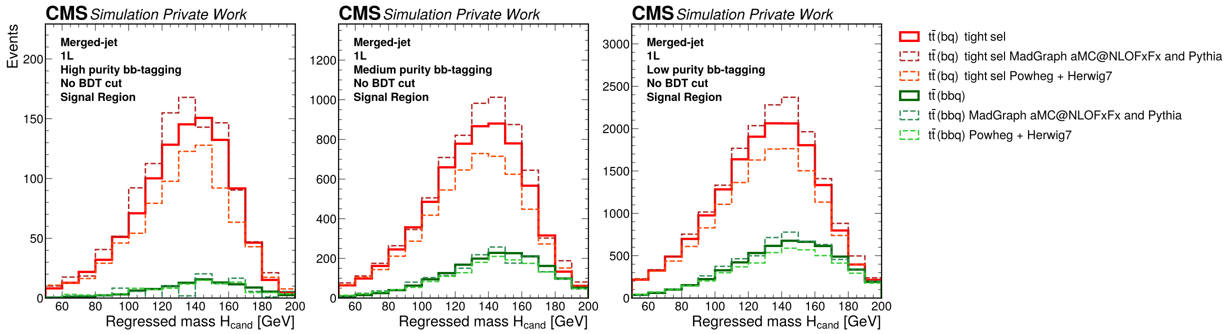


FIGURE 2.7: This figure shows the ParticleNet tagger response shapes for the different $t\bar{t}$ bkg. components.

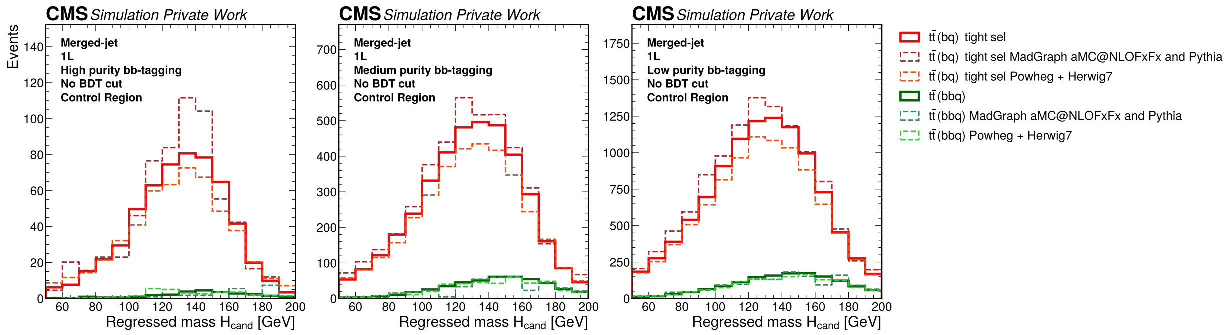
2.2.2 Tagger response

This observation gives rise to two primary concerns. Firstly, it suggests the possibility of a shift in the overall cyan shape (comprising bbq and bq tight), indicating a change in the aggregate behavior. Secondly, as evidenced by Figure 2.7, discrepancies exist in the ParticleNet tagger responses for "bbq" and "bq tight" meaning that in addition to the relative fractions changing, they can also shift due to the different tagger response, introducing additional issues when attempting to derive corrections from the control regions to apply them in the signal region.

After conducting a more in-depth investigation, it was discovered that the trained Boosted Decision Tree (BDT) exhibited a clear preference for selecting events with a characteristic of $N_{\text{jets}} = 0$. This observation aligns with the expectation that events featuring zero additional jets possess a higher likelihood of being signal events. However, an unintended consequence of this preference was an increase in background contributions from the bbq process, which also tends to produce events with zero associated additional jets. To address this issue, an attempt was made to retrain the BDT while excluding N_{jets} as one of the input variables. Unfortunately, this adjustment had only a marginal impact on the contribution from the bbq background, and the relative shapes of the distributions still exhibited the same changes in relative fractions following the application of the BDT selection. In summary, despite efforts to mitigate the influence of events with zero additional jets, the bbq background contribution remained largely unaffected by the exclusion of N_{jets} from the BDT training. The underlying patterns in the relative distributions persisted even after this modification.



(A) Signal Region with $N_{\text{jets}} < 2$ and no BDT cut applied.



(B) $t\bar{t}$ Control Region with $N_{\text{jets}} \geq 2$ and no BDT cut applied.

FIGURE 2.8: The figures display the absolute yields in both Signal and $t\bar{t}$ Control Region, where the solid line corresponds to the results obtained from the original MC simulation, while the dotted lines depict the outcomes derived from MC simulations conducted using distinct generators and matrix elements.

2.2.3 Uncertainty estimation

To achieve a more comprehensive assessment of the uncertainty associated with the bq tight and bbq components, a systematic interchange of generators, matrix element, and shower calculations was performed across various Monte Carlo (MC) simulations. The analysis of the fluctuations in the bq tight and bbq components for the absolute yields, as depicted in Figure 2.8, indicates that the bbq component exhibits minimal impact, whereas the bq tight component displays pronounced fluctuation, particularly around its peak at 140 GeV. This outcome is promising, given that the bbq background, with its shifting fraction, represents the primary challenge.

However, upon examination of the tagger shape uncertainty in Figure 2.9, a contrasting pattern emerges, with the bbq component demonstrating more significant effect. This discrepancy poses a significant complication, as the tagger response is inherently uncertain. This lack of precise a priori knowledge concerning the tagger response contributes to substantial associated uncertainties for the bbq component, further complicating the matter. In summary, the uncertainty analysis underscores the feasibility but not the desirability of merely assigning an uncertainty value to the bbq component and proceeding with the analysis. Therefore, additional steps have to be undertaken to further help reduce the bbq contributions.

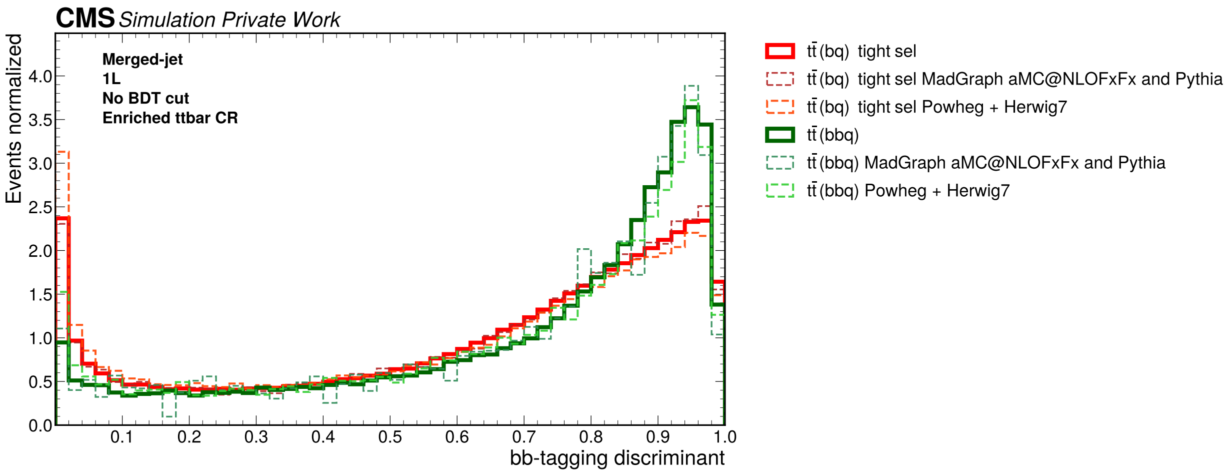


FIGURE 2.9: This figure portrays the ParticleNet tagger response within the $t\bar{t}$ Control Region, focusing on the bbq and bq tight components. The solid line represents the results derived from the original MC simulation, while the dotted lines illustrate outcomes obtained through MC simulations utilizing different generators and matrix elements.

2.2.4 Ak15 jet substructure

One approach involves leveraging jet substructure to discern between the various components. The characterization of jet prongs is facilitated by the $\tau_{N,\,N-1}$ - ratio, defined as:

$$\tau_{N,\,N-1} = \frac{\tau_N}{\tau_{N-1}} \,. \tag{2.1}$$

In a jet with N prongs, one expects $\tau_1, \ldots, \tau_{N-1}$ to be large and $\tau \geq N$ to be small. In our context, particular attention is given to the $\tau_{3,\,2}$ - ratio, which effectively distinguishes between 2-prong and 3-prong jets. A $\tau_{3,\,2}$ - ratio close to 1 signifies a 2-prong jet, while a lower ratio indicates a 3-prong jet.
Displayed in Figure 2.10, the $\tau_{3,\,2}$ - ratio demonstrates the distinct characteristics of various $t\bar{t}$ components as well as the H → bb signal. Notably, the bbq component exhibits a 3-prong substructure, with its ratio differing from that of 2-prong ak15 jets belonging to $t\bar{t}$ (W) or H → bb signal. However, imposing a cut at for example $\tau_{3,\,2} = 0.75$ or 0.8 would still result in sizable signal loss.

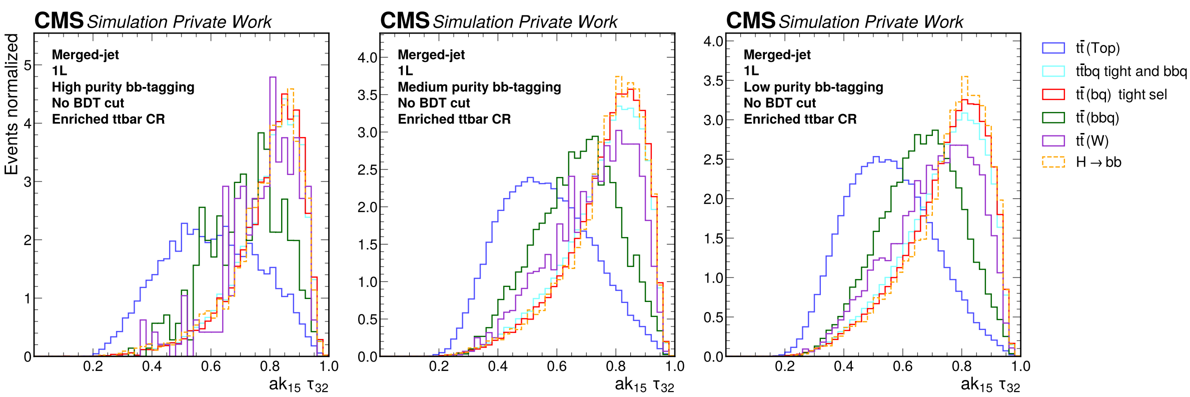


FIGURE 2.10: This figure displays the shapes of the $\tau_{3,\,2}$ - ratio for different $t\bar{t}$ components as well as the H → bb signal.

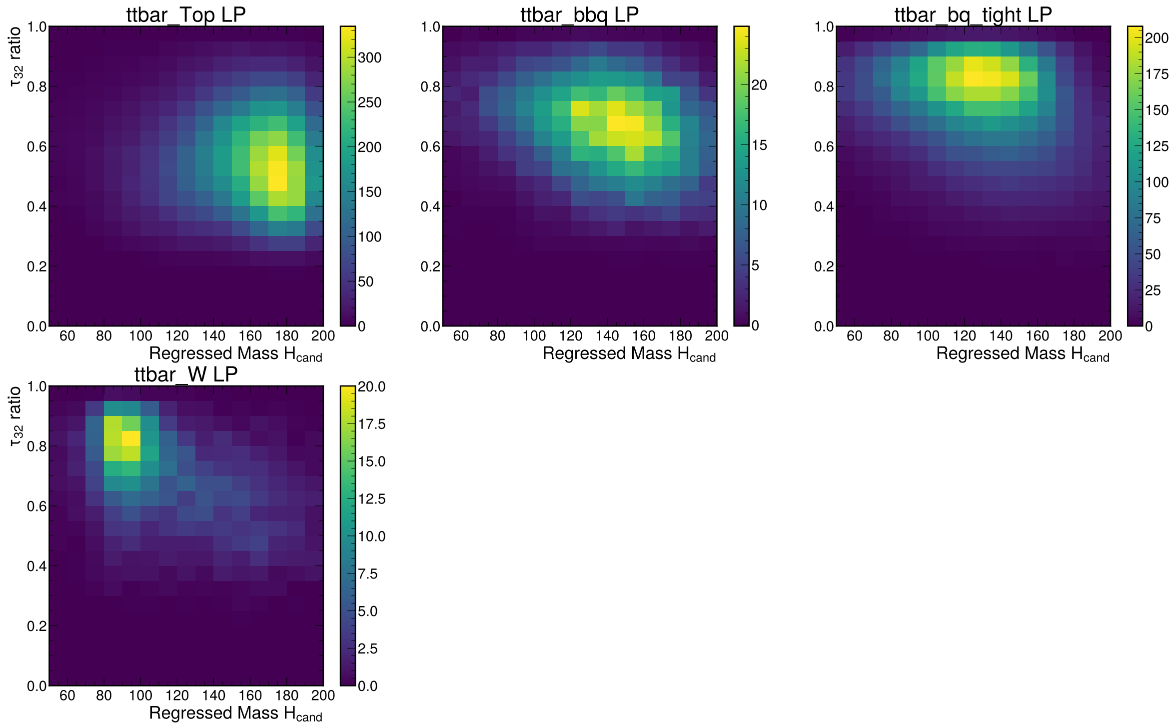
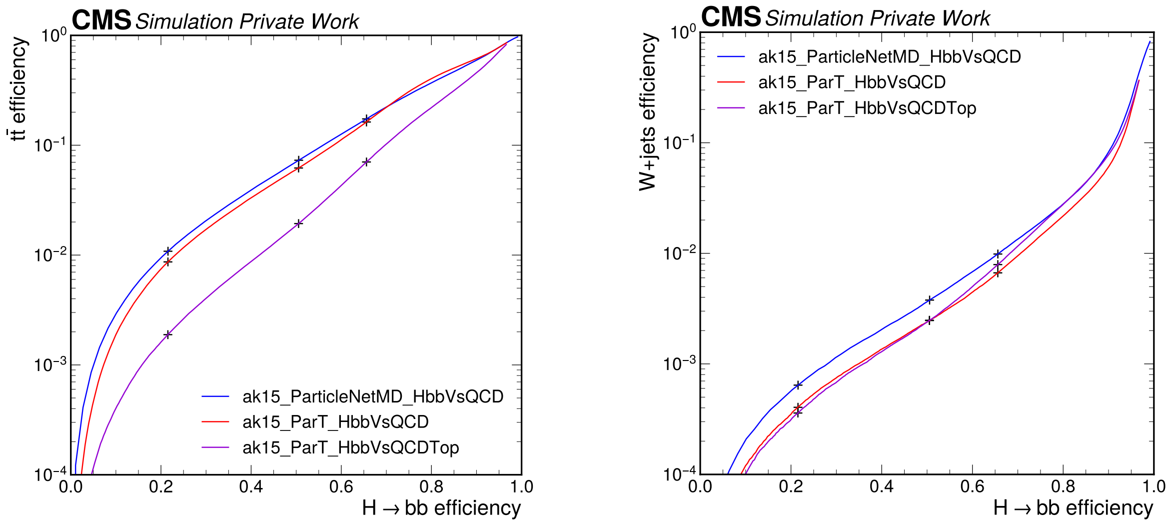


FIGURE 2.11: These 2D-hisograms portray the relation between mass and the $\tau_{3,2}$ - ratio for different $t\bar{t}$ components in the $t\bar{t}$ Control Region with $N_{\text{jets}} \geq 2$, no BDT cut applied, and in the low purity tagger selection.

A potential solution could entail leveraging the distinct masses of the various components by constructing a 2D histogram utilizing both the $\tau_{3,2}$ - ratio and mass. This concept is demonstrated in Figure 2.11, where the Top and W components manifest distinct separations based on their characteristic masses of 170 GeV and 80 GeV, respectively. It is also possible to make a cut between the bbq and bq tight components, although this strategy retains an element of uncertainty, as evidenced by the plot.



(A) Illustration of the $t\bar{t}$ background vs $H \to bb$ signal ROC-Curve.

(B) Illustration of the W+jets background vs $H \to bb$ signal ROC-Curve.

FIGURE 2.12: Shown in this figure are ROC-Curves comparing ParticleNet (blue) and ParticleTransformer as well as different ParticleTransformer discriminants (red and magenta) in terms of signal vs $t\bar{t}$ or W+jets efficiency.

2.2.5 Quick look at ParticleTransformer

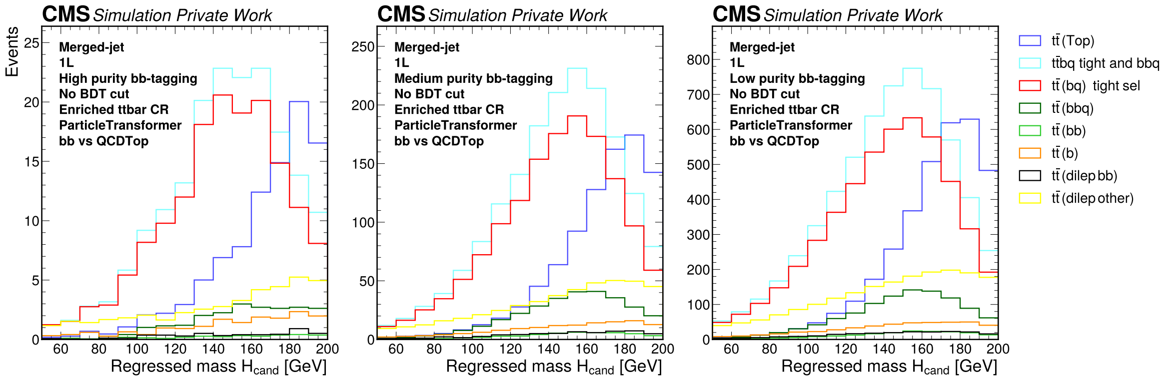
It was also investigated what potential impact of a new jet tagger, ParticleTransformer, on mitigating the ttbar background, and subsequently, the bbq contribution could have. The previously established bb-tagging discriminant, as defined in Equation 2.2, served to differentiate signal from QCD background:

$$\text{Old bb-tagging discriminant} = \frac{\text{score } (H \to b\bar{b})}{\text{score } (H \to b\bar{b}) + \text{score } (\text{QCD})} \,. \tag{2.2}$$

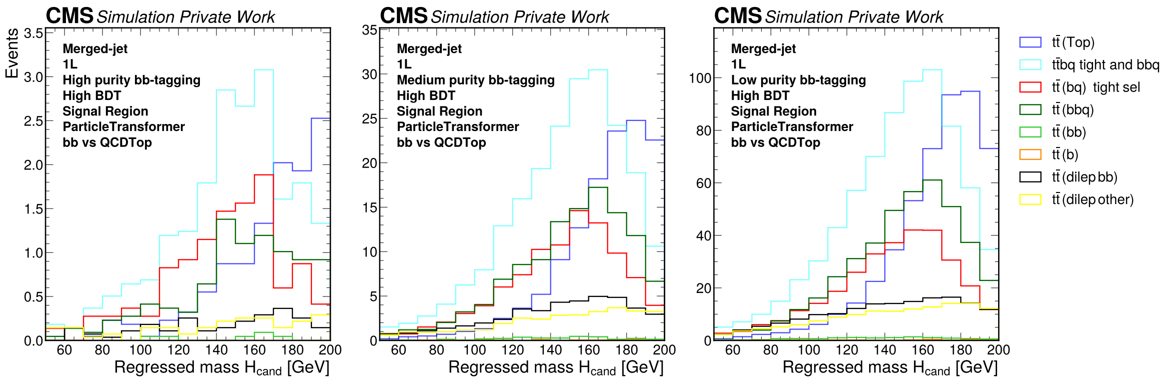
In contrast, the novel ParticleTransformer bb-tagging discriminant introduced an additional category for Top events:

$$\text{New bb-tagging discriminant} = \frac{\text{score } (H \to b\bar{b})}{\text{score } (H \to b\bar{b}) + \text{score } (\text{QCD}) + \text{score } (\text{Top})} \,. \tag{2.3}$$

The incorporation of this novel category was made possible through the inclusion of additional $t\bar{t}$ samples in the training of ParticleTransformer.
The Receiver Operating Characteristic (ROC) curves in Figure 2.12 highlight that the new ParticleTransformer surpasses the former ParticleNet performance within the W+jets samples. Remarkably, the inclusion of the Top category in the new tagging discriminant empowers it to outperform ParticleNet considerably in the $t\bar{t}$ samples.
Observing the absolute yields within both the Control Region and the Signal Region in Figure 2.13, it becomes evident that the contribution from $t\bar{t}$ (Top) has undergone a substantial reduction compared to that depicted in Figure 2.6b. However, the persistent challenge presented by the bbq contribution remains relatively unaffected. While the potential effectiveness of introducing a $t\bar{t}$ (W), $t\bar{t}$ (bq), or $t\bar{t}$ (bq) category to the tagging discriminant requires further investigation, the promising outcomes from discriminating against $t\bar{t}$ (Top) suggest it as a plausible strategy.



(A) $t\bar{t}$ Control Region with $N_{\text{jets}} \geq 2$ and no BDT cut applied.



(B) Signal Region with $N_{\text{jets}} < 2$ and no BDT cut applied.

FIGURE 2.13: The figures display the absolute yields in both Signal and $t\bar{t}$ Control Region, obtained from the new ParticleTransformer jet tagger.

3 Conclusion and Outlook

The new ParticleTransformer shows great promise in aiding to further reduce the $t\bar{t}$ bkg., with this improvement the bbq and bq tight problem should be properly mitigated as such that the analysis can estimate it's contributions correctly.
If not, the 2d-histogram approach and some additional cuts can still be applied to mitigate the bbq contribution.
There is still much to be done :)