# Boosted Higgs boson tagging using jet substructures

Pavel Shvydkin

Moscow Institute of Physics and Technology

Supervisors: Michele de Gruttola, Devdatta Majumder

## 1 Introduction

Many beyond the standard model (BSM) theories predict massive particles that may decay to the standard model Higgs boson plus a vector boson ($W^+/W^-$ or Z), or quark. With the recent discovery of the Higgs boson [1, 2], searching for BSM particles using Higgs boson final states have now become common. The Higgs boson decays to a $\mathrm{b\overline{b}}$ pair with a branching fraction of $\sim 58\%$, and hence this is the decay mode that is often used in BSM physics searches. The W bosons or the Z boson is typically required to decay into a leptonic (electron or muon) final state, to provide a clean signature to trigger on, with relatively low backgrounds. A combination of the two allows for a full mass reconstruction of many new physics particles like resonances decaying to dibosons [3], or vector-like quarks [4].

Because the BSM particle searched for is mass, about 1 TeV or above, the decay products are highly Lorentz-boosted. This leads to high transverse momentum ($p_\mathrm{T}$) jets arising from the $\mathrm{H} \to \mathrm{b\overline{b}}$ decay, in these events. The high $p_\mathrm{T}$ results in the b quark-antiquark pair to be very close to each other, other, and merge into one jet, that is typically reconstructed using large jet sizes ($\Delta R = 0.8$). The distribution of $p_\mathrm{T}$ of the Higgs boson is shown on Fig. 1 (left). The Higgs boson has large $p_\mathrm{T}$ with the maximal value of $\frac{1}{2}m_\mathrm{X}$. On the Fig. 1 (right) there is the pseudorapidity ($\eta$) distribution. As one can note from this plot the Higgs boson is very centralized.

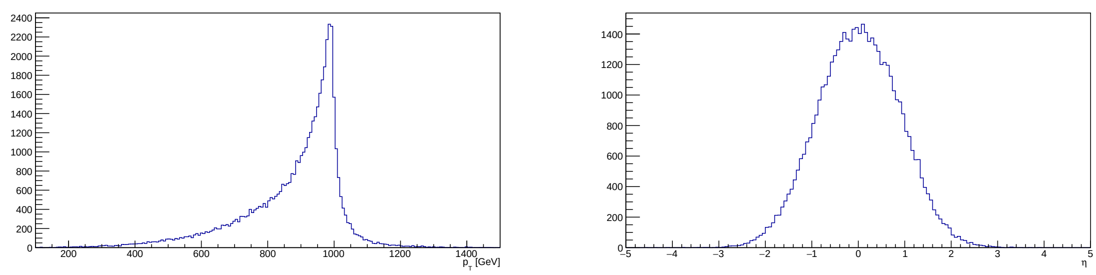


Figure 1: Left: $p_\mathrm{T}$ distribution of the Higgs boson. Right: $\eta$ distribution of the Higgs boson.

Figure 2 shows that the distance $\Delta R$ between simulated b and $\overline{\mathrm{b}}$ quarks is inversely proportional to $p_\mathrm{T}$. This work focuses on the improvement of the Higgs jet mass resolution using a regression technique. The regression makes use of boosted jets with substructure information [6], coupled with the pecularities of a b quark decay, like the presence of a soft lepton (SL) inside the jet. It has allowed to improve the resolution of the mass reconstruction and transverse momentum of the Higgs boson. The results of this investigation are shown in the report.

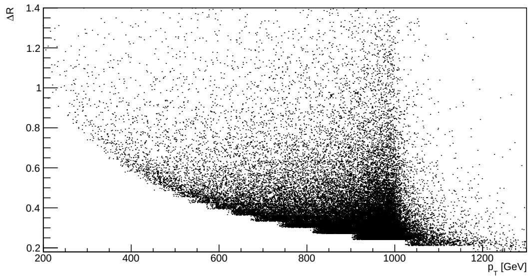


Figure 2: Distance $\Delta R$ between b and $\overline{\mathrm{b}}$ quarks depending on $p_{\mathrm{T}}$ of the Higgs boson

## 2 Reconstruction of boosted $\mathrm{H} \to \mathrm{b}\overline{\mathrm{b}}$ jets

The importance of physics at transverse momenta $p_{\mathrm{T}} >> m_{\mathrm{Z}}$ has implications for the structure of the final state because at high transverse momenta, "signature" particles, W, Z, Higgs bosons and top quarks, have very collimated decays (due to their relativistic boost). Standard approaches for identifying these particles fail because all the decay products end up in a single jet. The work so far on identifying hadronic decays of boosted heavy particles has fallen into two broad classes: particles with two-pronged decays (the electroweak (EW) bosons), and those with three-pronged decays (top quarks). In two-pronged jets which are interested in there are two main issues of jet reconstruction resulting in jet mass error. First is mass loss because of perturbative gluon radiation for small distance parameter. Vice versa for bigger $\Delta R$ values there is significant contribution of underlying events (UE) events and pileup. So that one can understand the task of minimizing both types of contribution [5]. The jet pruning algorithm [7] is used to remove UE and pileup component of the jet energy. The data to be regressed is AK8 jet, which was defined by anti-$k_{\mathrm{T}}$ algorithm with $\Delta R$ equal to 0.8. Figure 3 shows the high efficiency of this jet definition for high $p_{\mathrm{T}}$.

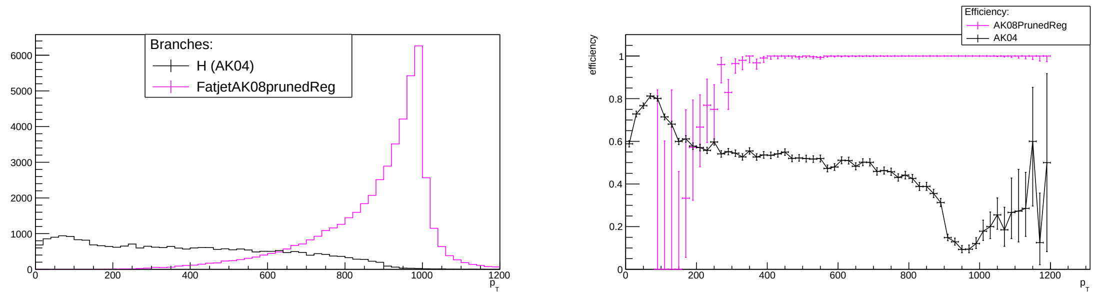


Figure 3: Left: Comparison of mass reconstruction for anti-$k_{\mathrm{T}}$ algorithm with $\Delta R = 0.8$ and $\Delta R = 0.4$. Right: Efficiency of AK8 and AK4 jets. Distribution "FatjetAK8pruned_pt" was obtained by pruning [7] of ungroomed data.

From the Fig. 3 one can see that AK8 jet's reconstructed $p_{\mathrm{T}}$ is close to simulated one and its efficiency is equal to 100% for wide range of high $p_{\mathrm{T}}$ that is much better than AK4 jet's efficiency for instance.

# 3 Mass Regression

The Toolkit for Multivariate Analysis (TMVA) [9] provides a ROOT-integrated environment for the processing, parallel evaluation and application of multivariate classification and multivariate regression techniques. All multivariate techniques in TMVA belong to the family of "supervised learning" algorithms. They make use of training events, for which the desired output is known, to determine the mapping function that either describes a decision boundary (classification) or an approximation of the underlying functional behavior defining the target value (regression). The mapping function can contain various degrees of approximations and may be a single global function, or a set of local models [9].

In this work the Boosted Decision Tree (BDT) method is used. A decision (regression) tree is a binary tree structured classifier (regressor). Repeated left/right (yes/no) decisions are taken on one single variable at a time until a stop criterion is fulfilled. The phase space is split this way into many regions that are eventually classified as signal or background, depending on the majority of training events that end up in the final leaf node. In case of regression trees, each output node represents a specific value of the target variable. The boosting of a decision (regression) tree extends this concept from one tree to several trees which form a forest. The trees are derived from the same training ensemble by reweighting events, and are finally combined into a single classifier (regressor) which is given by a (weighted) average of the individual decision (regression) trees. Boosting stabilizes the response of the decision trees with respect to fluctuations in the training sample and is able to considerably enhance the performance w.r.t. a single tree [9].

The regression seeks to improve the mass resolution of the jet. However we use as the target value, the ratio of the Higgs boson $p_{\mathrm{T}}$ from simulations to the reconstructed jet $p_{\mathrm{T}}$ $p_{\mathrm{T}}^{\mathrm{Gen.H}}/p_{\mathrm{T}}^{prunedjet}$ as the target value in performing the regression. It is expected that the pruned mass of the jet scales the same way as $p_{\mathrm{T}}^{\mathrm{Gen.H}}/p_{\mathrm{T}}^{prunedjet}$.

The variables chosen for the regression correspond to the properties of a Higgs jet, for instance the kinematic variables like the jet $p_{\mathrm{T}}$, $\eta$ and the mass (ungroomed, groomed vs. calibrated and uncalibrated), and well the properties of a B hadron decay, like the presence of a secondary vertex (SV) and soft leptons. B tagging is frequently used for identifying Higgs jets. The newly developed "double b tagger algorithm" [8] holistically identifies jets from massive resonances like the Z or the Higgs bosons decaying to a $\mathrm{b\overline{b}}$ pair. This information is also used for the regression to study the effects of flavour tagging. The baseline regression (Cat 0) uses the following variables:

1. The ungroomed AK8 jet $p_{\mathrm{T}}$;
2. The pruned AK8 jet $p_{\mathrm{T}}$ ;
3. The $p_{\mathrm{T}}$ of the pruned AK8 jet after jet energy calibrations are applied to the pruned mass;
4. The $\eta$ of the pruned AK8 jet after jet energy calibrations are applied to the pruned mass;
5. The number of tracks associated with the AK8 jet;
6. The mass of the leading SV associated to the AK8 jet;
7. The SV energy ratio to that of the AK8 jet, of the leading secondary vertex associated to the AK8 jet;
8. The SV energy ratio to that of the AK8 jet, of the subleading secondary vertex associated to the AK8 jet;
9. The number of SL associated to the AK8 jet;

10. The $p_{\mathrm{T}}^{\mathrm{rel}}$ of the leading soft lepton associated to the jet.

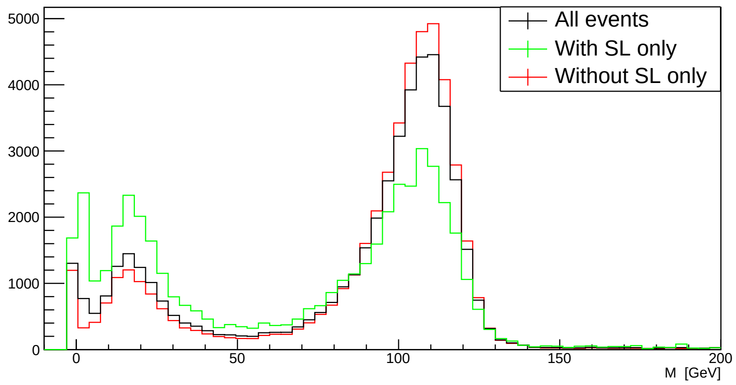


Figure 4: AK8 jet mass distribution for different SL cases: all events marked in black, events with SL is marked in green and events without SL are marked with red color

The presence of the SL from a semileptonic B hadron decay inside the jet significantly distorts the mass distribution. It can be attributed to neutrinos accompanying such a decay, which escapes undetected, thereby reducing the energy of the original parton. The influence of SL presence on reconstructed mass is shown on Fig. 4. As one can see distribution of events without SL has narrower the Higgs boson mass peak.

In addition we define two more categories of regression:

1. Cat 1 = Cat 0 + the double b tagger information;
2. Cat 2 = Cat 0 - the soft lepton information.

B tagging is important and provide us with information about whether event is signal or not. Another regression category uses different values of parameter $m_{\mathrm{X}}$

On Fig. 5 there is a fit with gaussian function. And on Fig. 6 there is a fit with Crystal Ball (CB) function on the narrower range. The gaussian function gives less $\sigma$ value because it fits only the peak of the distribution.

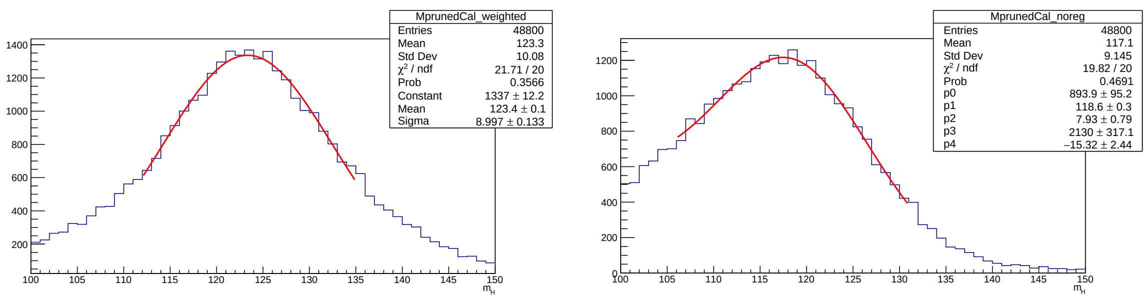


Figure 5: Right: non-regressed jet mass distribution gaussian function fit. Left: Regressed jet mass distribution gaussian function fit.

The fit results are presented for different regression trainings in Tables 1 and 2. Table 1 is for the gaussian function fit and Table 2 is for the CB one.

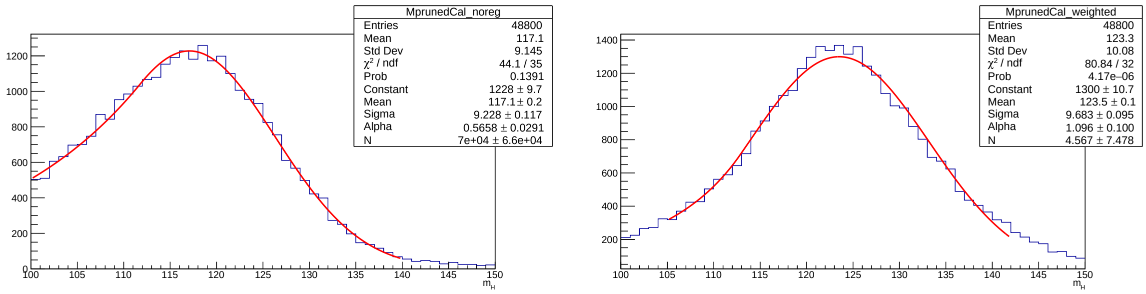


Figure 6: Left: non-regressed jet mass distribution CB fit. Right: regressed jet mass distribution CB function fit

Table 1: The gaussian function fit of different categories of regression

| training | mean | $\sigma$ | $\sigma/\mu$ |
|---|---|---|---|
| non-regressed | 118.10 | 8.79 | 0.0744 |
| Cat 0 | 123.35 | 9.00 | 0.0729 |
| Cat 1 | 123.36 | 9.05 | 0.0734 |
| Cat 2 | 123.32 | 9.16 | 0.0743 |
| varied $m_X$ | 125.05 | 10.74 | 0.0859 |

Table 2: The CB function fit of different categories of regression

| training | mean | $\sigma$ | $\sigma/\mu$ |
|---|---|---|---|
| non-regressed | 117.07 | 9.23 | 0.0788 |
| Cat 0 | 123.51 | 9.68 | 0.0784 |
| Cat 1 | 123.49 | 9.62 | 0.0779 |
| Cat 2 | 123.45 | 9.64 | 0.0781 |
| varied $m_X$ | 125.16 | 10.10 | 0.0807 |

# 4 Summary

We consider boosted $\mathrm{H} \to \mathrm{b\bar{b}}$ jets which were defined by AK8 method. We apply regression technique to jet mass reconstruction. In regression we use several different methods with several different sets of training variables.

This application results in improvement of the mass reconstruction by 3-4 percent.

There are several options to improve regression results. First is more careful pileup rejection. Second is using several different $m_X$ values. And third is considering other decays for example with diboson or H + t production.

# Acknowledgements

I would like to the CERN summer programme for giving me the opportunity to work at CERN over the summer. This work is under the supervision of Michele de Gruttola and Devdatta Majumder of the CMS Collaboration at CERN. I would like to thank them a lot for consistent explanation of all stages of the work, for clarification of complicated issues. Without the detailed instructions and help on lots of technical issues in code-writing I would not have progressed

with the work. Also I am grateful to Anna Sfyrla and Helge Voss, the lecturers of Summer Student Program, who said about related things and answered my questions, that helped me to understand the whole picture of the work better.

## References


[1] G. Aad *et al.* [ATLAS Collaboration], Phys. Lett. B **716** (2012) 1 doi:10.1016/j.physletb.2012.08.020 [arXiv:1207.7214 [hep-ex]].

[2] S. Chatrchyan *et al.* [CMS Collaboration], Phys. Lett. B **716** (2012) 30 doi:10.1016/j.physletb.2012.08.021 [arXiv:1207.7235 [hep-ex]].

[3] CMS Collaboration, CMS-PAS-B2G-16-003.

[4] CMS Collaboration, CMS-PAS-B2G-16-005.

[5] G. P. Salam, Eur. Phys. J. C **67** (2010) 637 doi:10.1140/epjc/s10052-010-1314-6 [arXiv:0906.1833 [hep-ph]].

[6] V. Khachatryan *et al.* [CMS Collaboration], JHEP **1412** (2014) 017 doi:10.1007/JHEP12(2014)017 [arXiv:1410.4227 [hep-ex]].

[7] S. D. Ellis, C. K. Vermilion and J. R. Walsh, Phys. Rev. D **81** (2010) 094023 doi:10.1103/PhysRevD.81.094023 [arXiv:0912.0033 [hep-ph]].

[8] CMS Collaboration [CMS Collaboration], CMS-PAS-BTV-15-002.

[9] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss, "TMVA: Toolkit for Multivariate Data Analysis," PoS A CAT 040 (2007) [physics/0703039].


## Appendix. All the other fits of regressed mass distirbution

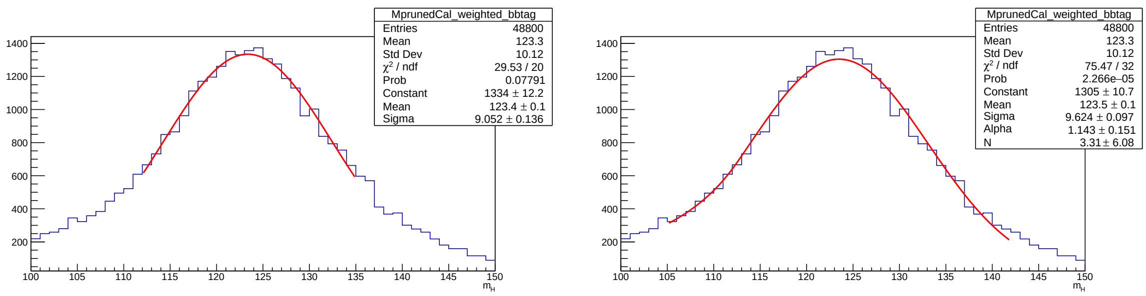


Figure 7: Cat 1 regression. Left: gaussian function fit. Right: CB function fit

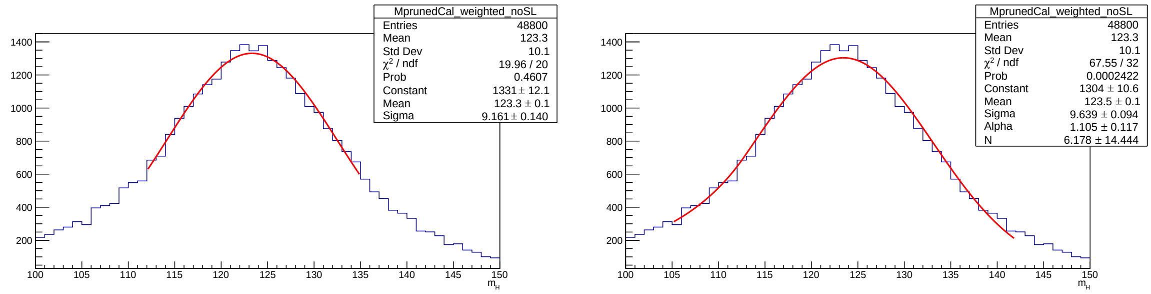


Figure 8: Cat 2 regression. Left: gaussian function fit. Right: CB function fit

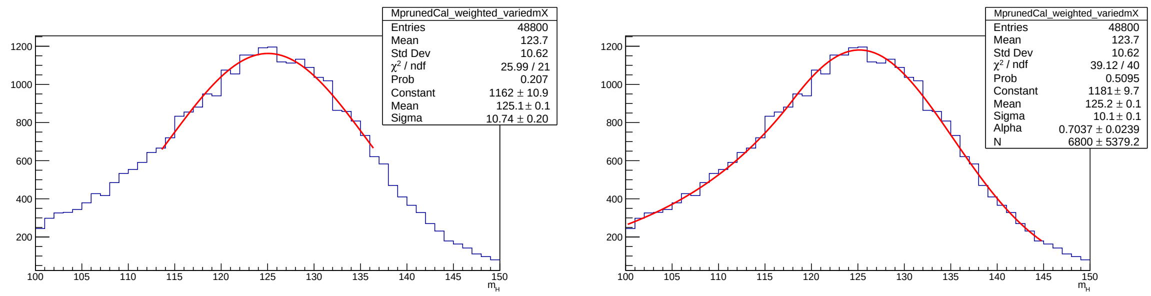


Figure 9: Regression using different $m_{\mathrm{X}}$. Left: gaussian function fit. Right: CB function fit