The VBF $H \to ZZ^* \to 4\ell$ multivariate analysis in ATLAS

Vector boson fusion is among the rarer ways to produce a Higgs boson, with its production cross section being roughly twelve times smaller than that of gluon fusion. However, its highly characteristic final state (i.e. a central Higgs boson in association with two forward jets) means that the VBF production mode is an ideal candidate for the use of multivariate techniques, allowing for the enhancement and extraction of a small VBF "signal" on top of a large ggF "background". The use of multivariate techniques in VBF searches is not a new phenomenon in ATLAS; the $H \to \gamma\gamma$ analysis group, for example, used such techniques during Run-I to disentangle VBF production from the large, irreducible diphoton backgrounds present in the channel [1]. Likewise, the $H \to ZZ^* \to 4\ell$ analysis made cursory use of a multivariate VBF discriminant in previous analyses during Run-I [2]. This chapter presents work done by the author to build an optimized multivariate discriminant which separates VBF- and ggF-mediated $H \to ZZ^* \to 4\ell$ events for Run-II of the LHC, targeting a centre-of-mass energy of $\sqrt{s} = 13\,\text{TeV}$. This chapter also details the technical implementation of the resultant tools into the larger ZZ^* analysis framework.

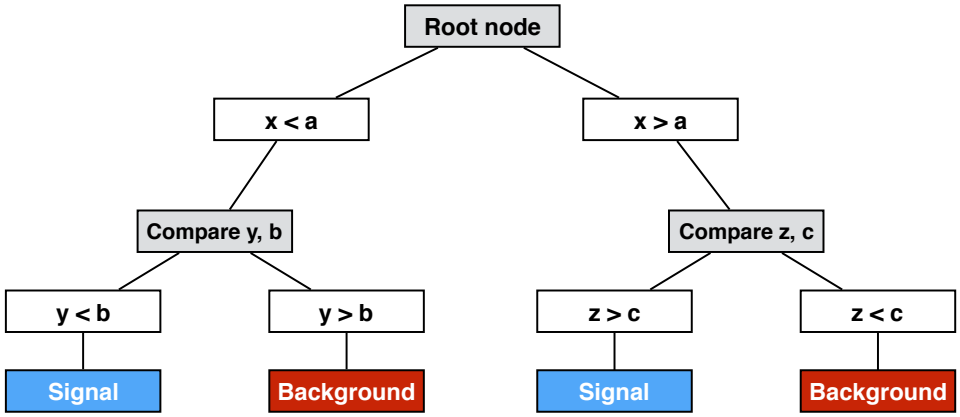


Figure 1: Schematic diagram of a binary decision tree.

Boosted decision trees

The multivariate classifier used in this analysis is derived using the boosted decision tree (BDT) method, as implemented in the TMVA package [3]. The BDT method uses multiple binary decision trees (as shown in Figure 1) to form a robust, statistically stable classifier which discriminates signal from background events in a data sample [3]. Each binary tree makes decisions using a set of user-provided discriminating variables, and is trained using distinct, non-overlapping signal and background samples. The function of each tree is to use a series of 'yes/no' decisions to classify individual events as signal-like or background-like.

The result from each tree (the *leaf node*) is combined via weighted average into a single discriminant, known as the *BDT score* [3]. This score varies nominally on [-1, +1], with background-like events assigned highly negative scores, and signal-like events assigned highly positive scores. The 'boosting' aspect derives from the fact that the weight of each tree is proportionate to its rate of misclassification (i.e. poor signal/background discrimination yields a low weight, and vice versa).

The form of boosting used in this analysis is gradient boosting [4], which treats the individual trees as a sum of terms in a function expansion approach. Considering a set of input variables $\mathbf{x} = (x_1, x_2, ..., x_n)$, gradient boosting seeks to model the BDT score as an output variable y, mapped to $\mathbf{x}$ through a function $y_i = \hat{F}(\mathbf{x}_i)$. In particular, the gradient boosting approach develops a function $F(\mathbf{x})$ that estimates $\hat{F}(\mathbf{x})$ by minimizing the expected value of a binomial log-likelihood loss function $L(y, F(\mathbf{x}))$. This boosting method was chosen for the multivariate analysis as it tends to isolate signal- and background-like events in opposite ends of the BDT score spectrum by default. Furthermore, gradient-boosted decision trees are also less susceptible to *overtraining* [3] – an inefficiency wherein the BDT learns only from a narrow region of phase space, or statistical fluctuations due to a low training sample size.

There are two interrelated tasks to consider when developing a BDT-based analysis. Firstly, signal and background training samples must be selected that both model well the underlying physics of the system under study, and are large enough to avoid overtraining effects in the BDT response. Secondly, training variables must be chosen to differentiate signal from background efficiently, and show no bias towards the quantity (or quantities) of interest in the analysis. The successful implementation of these two elements will produce an optimized, robust BDT-based analysis.

Input sample selection

In order to derive an effective BDT classifier, the input training samples must properly describe both the kinematics of the training variables, as well as the *correlations* between variables. Inadequate or incorrect modelling of the correlations between training variables can lead the BDT to make suboptimal decisions, meaning reduced signal

efficiency or purity when the MVA is applied to real data. At the same time, Monte Carlo samples used for ATLAS analyses are required to undergo both detector and pile-up simulation, to properly account for the effects of both object reconstruction efficiency and resolution on signal and background event yields. These aspects of the simulation are particularly time-consuming and computationally intensive, meaning that reconstruction-level Monte Carlo samples tend to only be as large as is necessary to provide sensible yield predictions in each analysis category, resulting oftentimes in samples with $\mathcal{O}(100\text{k})$ events or less. Accordingly, for ggF events with two jets, which constitute approximately 12% of the total cross section [5], the training sample size could be as small as $\mathcal{O}(10\text{k})$ events.

Therefore, the decision was made to consider two cases when deriving the BDT to separate ggF and VBF Higgs boson production. Firstly, the derivation and optimization of the BDT would take place using *truth-level* samples, which would allow for the efficient production of large Monte Carlo training samples, of size $\mathcal{O}(1\text{M})$ events or more. The large sample size would allow for all regions of the phase space to be adequately probed by the BDT, and provide an accurate representation of the underlying correlations between kinematic variables. In a sense, this approach also allows for a more "physics-motivated" BDT training, as no detector effects are considered in either the shapes or correlations of training variables. Secondly, once an optimized multivariate approach was derived which reflects the true underlying physics of the ggF and VBF processes, the BDT would be re-trained using fully detector-simulated, reconstruction-level samples of each process. Such an approach allows for reconstruction-level effects to be factored into the decisions made by the BDT, while also preventing the optimization process from being sensitive to detector efficiency or

luminosity conditions.

To justify such an approach, the shapes and correlations of each training variable would require comparisons at truth- and reconstruction-level to ensure that the underlying physics of each process are properly represented by the reconstruction-level Monte Carlo. To model VBF signal at both truth- and reconstruction-level, the POWHEG VBFH generator was used. Conversely, to model truth-level dijet ggF production, the POWHEG HJJ generator was used, which produces $H+2j$ events at NLO QCD, while at reconstruction-level, the NNLOPS method was used, which produces dijet events at LO. Although the kinematic differences between $H+2j$ calculations at LO and NLO QCD are relatively small (as shown at truth-level in Figure **??**), the physics-motivated nature of the BDT derivation meant that each training sample should be as theoretically accurate as possible. A comparison of the truth-level and reconstruction-level distributions of key dijet kinematic variables are shown in Figures 2 and 3, with good agreement seen between truth and reconstruction-level shapes for both VBF and ggF production.

As a brief technical aside, a BDT is derived in two fundamental stages: a *training* stage where the classifier is derived based on input signal and background samples; and a *testing* stage where the trained classifier is applied to separate testing samples to evaluate the compatibility of the resultant signal/background discriminant distributions with those of the training samples. In order to avoid inducing biases in the training, the samples used for training and testing the BDT are required to be *orthogonal*, or have no events present in both training and testing samples. The separation of input samples into training and testing components was done using the default, random splitting algorithm in TMVA, ensuring that non-overlapping sub-samples

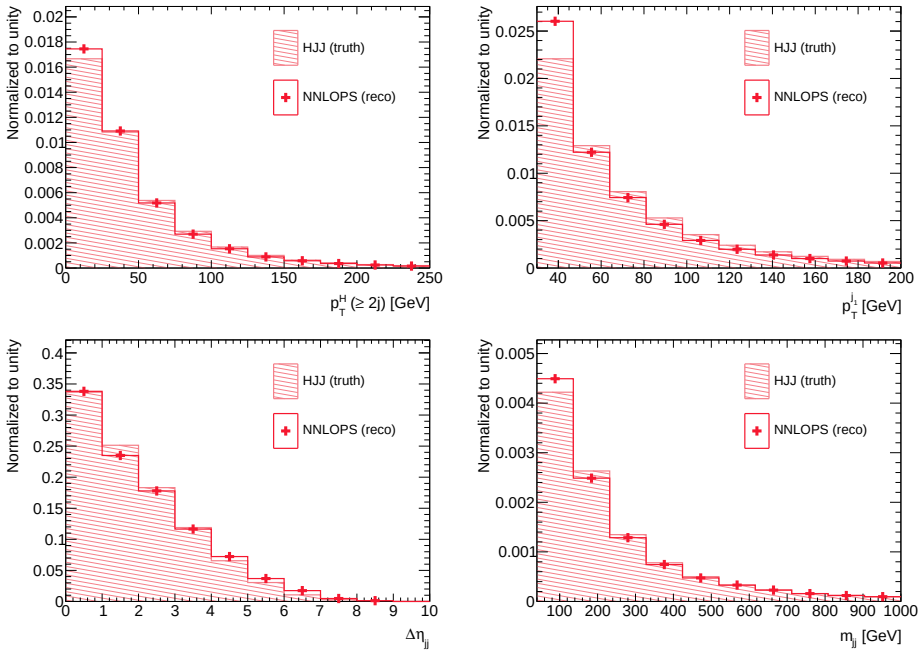


Figure 2: Comparison of truth- and reconstruction-level kinematics for ggF $H + 2j$ events. The truth-level events are generated at NLO QCD using the POWHEG HJJ generator, while the reconstruction-level events are generated at LO using the POWHEG NNLOPS generator.

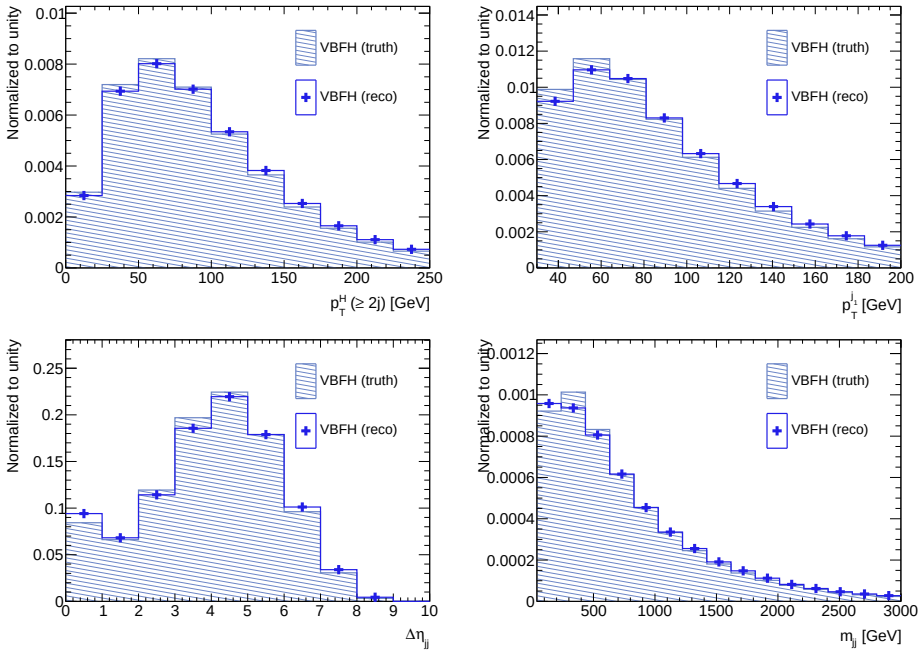


Figure 3: Comparison of truth- and reconstruction-level kinematics for VBF $H + 2j$ events. Both truth- and reconstruction-level events are generated using the POWHEG VBFH generator.

were produced, which are populated by essentially random selections of events.

Input variable selection

The VBF process provides a rather unique kinematic signature at tree-level, consisting of two final state jets which tend to have high rapidity, along with a Higgs boson that tends to be central and low-momentum. Therefore, the kinematic variables which stand to best characterize this signal process are those which exploit its unique characteristics. For example, the dijet invariant mass $m_{jj} = \sqrt{(E_{j_1} + E_{j_2})^2 - (\vec{p}_{j_1} + \vec{p}_{j_2})^2}$ tends to be quite large in VBF, owing to the fact that the two jets are highly separated in rapidity, to the point that the sum $\vec{p}_{j_1} + \vec{p}_{j_2}$ becomes nearly zero. Similarly, the transverse momentum of the combined Higgs boson - dijet system, p_{T}^{Hjj}, tends towards small values, owing to the small dijet system p_{T} from forward jets, and the low momentum of the Higgs boson itself.

Various numerical approaches can be taken to assess the discriminating power of a given training variable. In TMVA, two particular quantities are used to assess this property. The *shape separation* $\langle S^2 \rangle$ is used as an initial metric of performance, calculated as,

$$\langle S^2 \rangle = \frac{1}{2} \sum_{i=1}^{N_{\mathrm{bins}}} \frac{(s_i - b_i)^2}{s_i + b_i}. \tag{1}$$

where s and b represent binned probability density functions for the "signal" (VBF) and "background" (ggF) processes, respectively. Worth noting is that this variable is entirely shape-based, and so ignores differences in the overall normalization between the two samples, providing a highly intuitive, kinematically-motivated measure of the

separability of the two processes.

The other quantity used to assess discriminating power is the *importance*, calculated internally within TMVA after the first pass of BDT training. The importance is a BDT-specific quantity, and quantifies how frequently a variable is used to make decisions. It is computed for a given variable as the frequency of its use in node splitting, weighted by both the number of events in each node, and the separation gain-squared it achieved [3]. So, the ideal set of training variables will demonstrate high individual shape separation, and a relatively equal balance of importance in their usage within the BDT.

To begin, a number of training variables were initially considered based around the Higgs, jet, and dijet kinematics, with studies performed to shrink an initially large list using $\langle S^2 \rangle$ and the BDT importance as metrics. Shape distributions of the most performant variables of this initial list are shown in Figure 4. Aside from the more straightforward single jet and dijet-related variables, a number of variables also represent the composite Higgs-dijet system, such as η^*_{ZZ}, which is defined as,

$$\eta^*_{ZZ} = \left| \frac{\eta_{j_1} + \eta_{j_2}}{2} - \eta_{ZZ} \right|, \tag{2}$$

and the minimum $\Delta R(j_1, j_2, Z_1, Z_2)$ (or $\Delta \mathrm{R}^{\mathrm{min}}_{jZ}$), which is defined as the minimum ΔR among the following choices,

$$\Delta R(j_1, Z_1),\ \Delta R(j_1, Z_2),\ \Delta R(j_2, Z_1),\ \Delta R(j_2, Z_2).$$

To demonstrate the performance of each variable in a multivariate context, a BDT was trained using all chosen discriminating variables. The shape separation

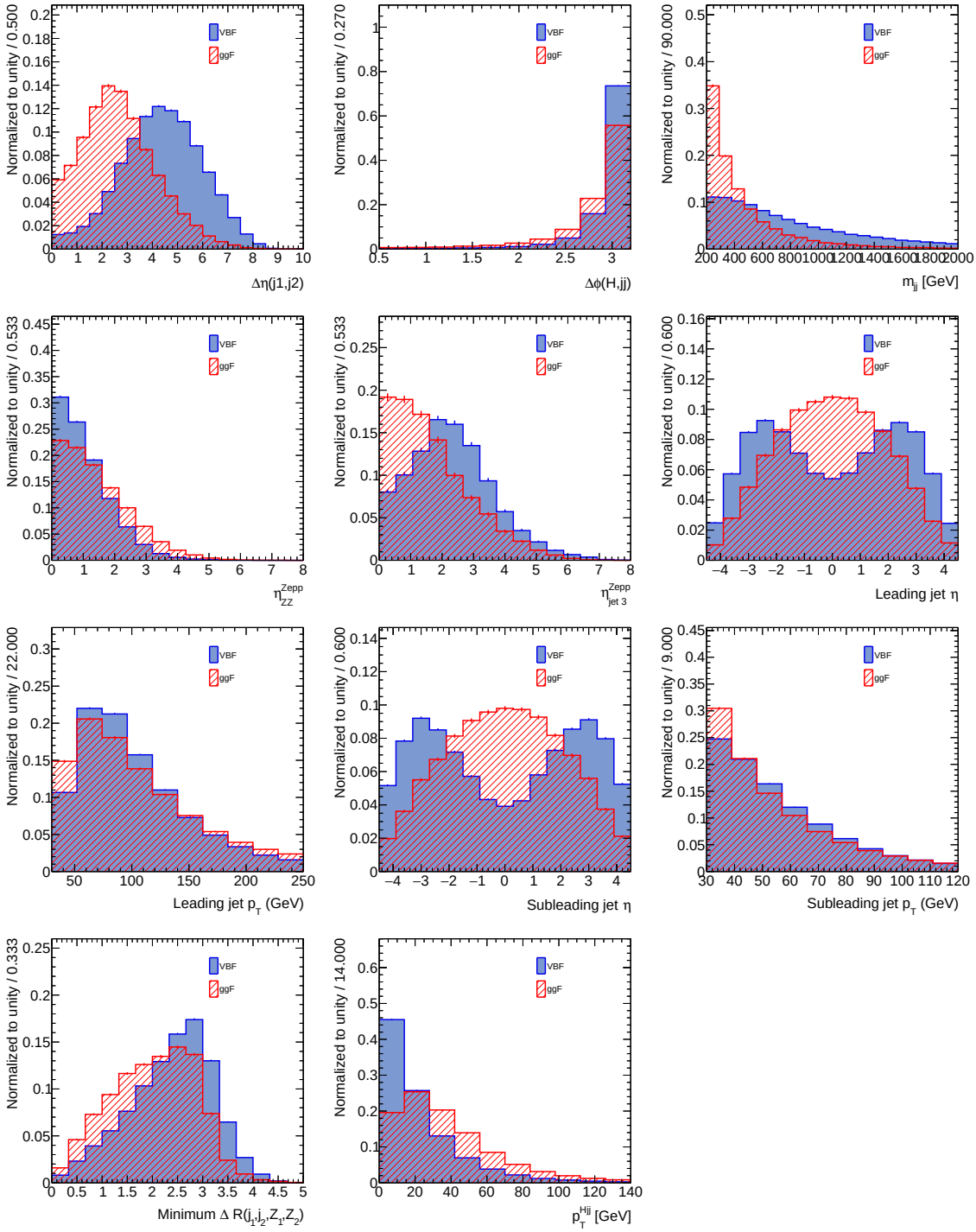


Figure 4: Discriminating variables considered for use in training the BDT to separate VBF and ggF production in the $H \rightarrow ZZ^* \rightarrow 4\ell$ channel.

and importance for each variable is given in Table 1, with the variables ranked in order of decreasing shape separation. The most obvious conclusion from Table 1 is that $\Delta\eta_{\mathrm{jj}}$ is both the most separating variable, and used most frequently by the BDT. Such a result is not particularly surprising; considering that the tell-tale sign of VBF is two high energy, forward jets in conjunction with a central Higgs boson, it should be expected that this characteristic is exploited prominently by the BDT training.

However, it can also be gathered from this table that variables with high shape separation can in practice have low importance to the BDT. For instance, p_{T}^{Hjj} is the third most separating training variable, but is only sixth in terms of importance, implying that it is either less frequently used by the BDT to make decisions, or that much of the decisions using this variable have a relatively small effect on the final signal purity. Likewise, m_{jj} is the fourth most important variable, a result which is likely due to its high correlation with $\Delta\eta_{\mathrm{jj}}$, conferring it a slight degree of redundancy.

Conversely, the second most important variable is found to be η^{*}_{ZZ}, which is ranked ninth in terms of separating power. The utility of this variable derives from its correlations with more separating variables, as its definition implicity carries information about the presence of additional radiation beyond the leading and subleading jets. In particular, tight cuts on η^{*}_{ZZ} are correlated with a VBF-like phase space containing two high-rapidity jets and a central Higgs boson, as this limiting scenario corresponds to $\eta^{*}_{ZZ} \rightarrow 0$. This behaviour demonstrates how the shape separation and importance are complimentary metrics, and that an optimized BDT will rely on training variables which are both intrinsically separating, and which embed higher-order information through correlations with other variables.

Variable	Separation $\langle S^2 \rangle$	BDT importance	Importance ranking
$\Delta\eta_{\mathrm{jj}}$	0.24	0.14	1
m_{jj}	0.23	0.097	4
p_{T}^{Hjj}	0.11	0.087	6
η_{j_2}	0.086	0.10	3
η_{j_1}	0.053	0.067	11
$\Delta\phi_{Hjj}$	0.050	0.079	8
$\Delta\mathrm{R}_{jZ}^{\mathrm{min}}$	0.046	0.084	7
$\eta_{j_3}^{*}$	0.037	0.089	5
η_{ZZ}^{*}	0.036	0.11	2
$p_{\mathrm{T}}^{j_1}$	0.016	0.078	9
$p_{\mathrm{T}}^{j_2}$	0.0095	0.073	10

Table 1: Table of the shape separation $\langle S^2 \rangle$ and BDT importance for a number of kinematic variables with power to separate VBF and ggF Higgs production. Variables are ranked in order of descending $\langle S^2 \rangle$, with the corresponding rankings for BDT importance shown.

Optimization of input variables

Given the conclusions of the previous section, it is worth investigating whether a subset of the full list of variables can be derived to optimize the overall performance of the BDT. In principle, the set of training variables chosen should be that which maximizes efficacy and efficiency, in the sense that a highly performant discriminant should be developed, trained using as few variables as possible to avoid "black box" scenarios where the BDT makes decisions that cannot be easily physically intuited.

In order to determine the optimal set of input training variables, the choice was made to perform an iterative scan over each possible permutation of the training variables shown in Figure 4 to find the configuration which provides the most effective discrimination with the least number of variables. That is to say, given the initial full list of $n = 11$ training variables, the following steps were performed:

1. For k training variables, iterate over all $n!/k!(n-k)!$ possible permutations
2. For each iteration j, perform a BDT training using the chosen input variables
3. Assess the maximal VBF signal significance with respect to ggF for the j^{th} iteration
4. Determine the iteration which fully maximizes the VBF signal significance out of all $n!/k!(n-k)!$ permutations

In lieu of performing a computationally-expensive signal significance calculation using e.g. a likelihood fit-based approach [6], this quantity is approximated using the *median significance*,

$$Z_0 \approx \sqrt{2\left((s+b)\ln\left(\frac{s+b}{b}\right)-s\right)} \tag{3}$$

which is an asymptotic expression for the commonly-used log-likelihood test statistic [6], valid when the number of background events, b, is large with respect to the number of signal events, s. The median significance is calculated using an integrated luminosity of $\int \mathcal{L} = 1\,\text{fb}^{-1}$ for simplicity, as in practice, the median significance will scale with the square root of the integrated luminosity.

Within each iteration of step (3) defined above, a second scan procedure is performed to determine an optimal cut to place on the trained BDT discriminant. An iterative series of cuts in steps of 0.05 are placed on the BDT score, and at each step, the VBF signal significance is recomputed for events passing the cut. The chosen discriminant cut is that which maximizes the median significance for a given training iteration. The purpose of this procedure is to ensure that the performance of each

k	Optimal variable set	Optimal significance	Optimal BDT cut
2	m_{jj}, $\Delta\eta_{\mathrm{jj}}$	0.373	0.35
3	m_{jj}, $\Delta\eta_{\mathrm{jj}}$, p_{T}^{Hjj}	0.419	0.5
4	m_{jj}, $\Delta\eta_{\mathrm{jj}}$, p_{T}^{Hjj}, η^*_{ZZ}	0.430	0.5
5	m_{jj}, $\Delta\eta_{\mathrm{jj}}$, p_{T}^{Hjj}, η^*_{ZZ}, $\eta^*_{j_3}$	0.434	0.5
6	m_{jj}, $\Delta\eta_{\mathrm{jj}}$, p_{T}^{Hjj}, η^*_{ZZ}, $\eta^*_{j_3}$, $p_{\mathrm{T}}^{j_2}$	0.437	0.5
7	(Same as 6) + $p_{\mathrm{T}}^{j_1}$	0.440	0.6
8	(Same as 7) + $\Delta\mathrm{R}_{jZ}^{\mathrm{min}}$	0.443	0.6
9	(Same as 8) + $\Delta\phi_{Hjj}$	0.444	0.55
10	m_{jj}, $\Delta\eta_{\mathrm{jj}}$, η_{j_1}, η_{j_2}, p_{T}^{Hjj}, $\Delta\mathrm{R}_{jZ}^{\mathrm{min}}$, η^*_{ZZ}, $\eta^*_{j_3}$, $p_{\mathrm{T}}^{j_2}$, $\Delta\phi_{Hjj}$	0.443	0.6
11	(All variables)	0.445	0.6
Run-I variables	m_{jj}, $\Delta\eta_{\mathrm{jj}}$, $p_{\mathrm{T}}^{j_1}$, $p_{\mathrm{T}}^{j_2}$, η_{j_1}	0.375	0.0

Table 2: Optimal training variable configurations for each choice of k variables from the full set of $n = 11$ possible variables.

training permutation is fairly assessed, as in principle measurements will be made in bins of the BDT score, with the most "VBF-enriched" bin providing the largest contribution to the overall significance. Accordingly, calculating the median significance only in the region above the discriminant threshold provides a straightforward approximation of the behaviour expected in the final analysis.

The optimal configuration of training variables for each choice of k variables from $k = 2$ to 11 is shown in Table 2, along with the optimal discriminant cut, and the corresponding median significance. As a baseline for the optimization, results are also shown for a training made with the variables used to build the Run-I BDT discriminant – namely, m_{jj}, $\Delta\eta_{\mathrm{jj}}$, $p_{\mathrm{T}}^{j_1}$, $p_{\mathrm{T}}^{j_2}$, and η_{j_1}. It is evident that the change in efficacy of the BDT grows asymptotically small as more variables are added; while the change in optimal significance from $k = 2$ to $k = 6$ is $\sim 17\%$, the change in significance from $k = 6$ to $k = 11$ is less than 2%. Furthermore, the following variables all appear

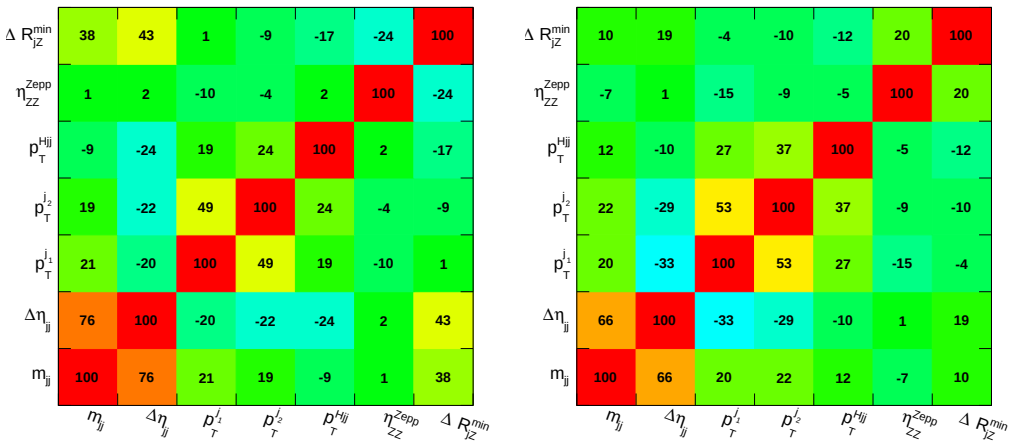
repeatedly among the optimal variable list permutations:

$$m_{\mathrm{jj}},\ \Delta\eta_{\mathrm{jj}},\ p_{\mathrm{T}}^{Hjj},\ p_{\mathrm{T}}^{j_2},\ p_{\mathrm{T}}^{j_1},\ \eta_{ZZ}^{*},\ \Delta\mathrm{R}_{jZ}^{\mathrm{min}}$$

In a sense, the configuration of variables shown above should not come as a surprise. As previously mentioned, both m_{jj} and p_{T}^{Hjj} strongly characterize the VBF production process, along with the dijet pseudorapidity separation $\Delta\eta_{\mathrm{jj}}$, which is highly correlated to m_{jj}. Furthermore, the distributions of the leading and subleading jet p_{T} differ enough between VBF and ggF production to provide some additional discriminating power. Lastly, the distributions of η_{ZZ}^{*} and $\Delta\mathrm{R}_{jZ}^{\mathrm{min}}$ can indicate the presence of additional radiation in the central region of the detector, but do not explicitly act as vetoes with tight cuts.

While the variable $\eta_{j_3}^{*}$ does also reappear a few times in the optimal variable lists, tests done using simulated high-luminosity conditions (described in Appendix **??**) demonstrated that the variable held no discriminating power at high levels of pile-up. Furthermore, the modelling of the third subleading jet in ggF is highly generator-dependent; while the use of an NLO QCD calculation provides a more accurate result derived from the matrix element calculation, a LO implementation would necessitate the modelling of the third jet through the SMC generator, which can lead to poorly modelled behaviour if the jet is not soft. Therefore, $\eta_{j_3}^{*}$ was removed from consideration to reduce the overall number training variables used in the BDT.

Matrices of the linear correlation coefficients (LCCs) for the remaining discriminating variables are found in Figure 5. The strong correlation between m_{jj} and $\Delta\eta_{\mathrm{jj}}$ is evident in the matrices corresponding to both VBF and ggF production, although the degree of correlation differs between processes. Furthermore, the differences in



(a) LCCs (VBF) **(b)** LCCs (ggF)

Figure 5: Linear correlation coefficients for the optimized list of training variables, produced using *truth-level* samples of VBF and ggF $H \to 4\ell$ events, generated at $\sqrt{s} = 14\,\text{TeV}$ using the POWHEG generator.

correlation strengths for p_{T}^{Hjj} and other training variables in VBF and ggF production may explain the consistent presence of these three variables among the lists of optimal permutations.

Given that the above list of variables consistently provides a superior discrimination of VBF and ggF production, with the addition of more variables bringing only marginal performance improvements, it was decided that only these seven variables would be used to perform the BDT training. Therefore, this configuration of variables fulfils the expectation of both maximal efficacy and efficiency in the BDT training. Any further tests or modifications to the training methodology would be done using samples of *reconstruction-level* events, as the physics-motivated aims of the optimization were fulfilled, with the remaining aims targeted towards optimizing experimental signal efficiency and purity, while minimizing measurement uncertainties.

BDTs and theory uncertainties in $H + 2j$ events

One important consideration when training a discriminant to separate physical processes is whether the signal and background process cross sections are well-defined in all regions of phase space probed by the discriminant. In general, igher-order QCD corrections to the ggF cross section calculation are highly non-trivial, and encompass both virtual and real corrections to the tree-level diagram. These corrections lead to theoretical complications when cuts are placed which restrict the number of final state jets in ggF, such as tight cuts on p_{T}^{Hjj} which create two-jet-enriched regions reminiscent of tree-level VBF Higgs production.

In general, the calculation of an observable (e.g. a cross section) in perturbative QCD can be written as,

$$O = \phi \otimes \hat{O} + P_O \tag{4}$$

where $\hat{O} = \sum_n c_n \alpha_s^n$ is the partonic version of the observable, with the series calculated to some finite order, and ϕ represents the parton distribution / fragmentation functions. The terms in P_O represent corrections of the form $p_n(\Lambda_{\mathrm{QCD}}/Q)$, where Q is the energy scale of the process, and Λ_{QCD} is the QCD energy scale. For ggF cross sections with ≥ 1 final state jets, the imposition of cuts which suppress Feynman diagrams featuring the real emission of soft (or "infrared / IR") gluons leads to the presence of large *Sudakov logarithms* in c_n at each order of the perturbative expansion $\hat{O}$. For example, placing an exclusive cut on the jet transverse momentum $p_{\mathrm{T}}^{\mathrm{cut}}$, such that events are separated into N and $\geq N + 1$-jet bins, induces Sudakov logarithms of the form $L = \ln(p_{\mathrm{T}}^{\mathrm{cut}}/Q)$ [7]. For very tight cuts, the logarithms become $\mathcal{O}(1)$, and

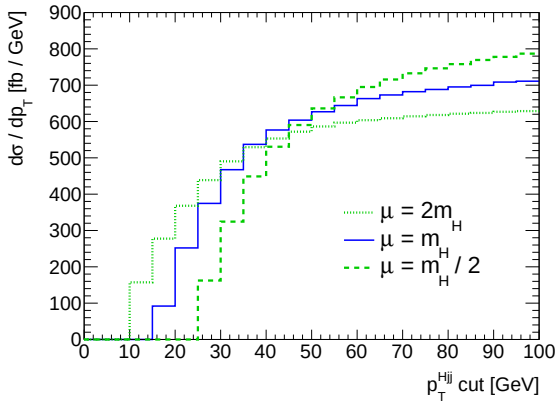


Figure 6: Effect of cuts on p_{T}^{Hjj} on the parton-level ggF $H + 2j$ differential cross section $d\sigma/dp_{\mathrm{T}}$, along with the scale variation error bands defined by $\mu_{\mathrm{R}} = \mu_{\mathrm{F}} = m_H/2$ and $2m_H$. All distributions are derived using MCFM with $m_H = 125\,\mathrm{GeV}$, requiring jet p_{T} above 30 GeV, and $m_{\mathrm{jj}} > 120\,\mathrm{GeV}$.

dominate the perturbative series, leading to "IR-sensitivity" in the cross section prediction. If the logarithms become large enough that they overcome the α_s suppression necessary for perturbative QCD, then the perturbative expansion completely breaks down, leading to meaningless cross section estimates [7].

In general, the use of fixed-order QCD calculations also comes with associated uncertainties due to the fact that there are higher-order corrections "missing" from the perturbative expansion. These uncertainties are typically estimated by varying the renormalization (μ_{R}) and factorization (μ_{F}) scales in the cross section calculation, often in a correlated fashion using a single scale $\mu = \mu_{\mathrm{F}} = \mu_{\mathrm{R}} = \mu_0/2$ or $2\mu_0$, where μ_0 is the central scale choice (typically m_H or $m_H/2$). So, beyond the estimation of the cross section itself, the use of tight or exclusive cuts on real emissions poses difficulties for estimating the uncertainties resulting from missing higher-order QCD corrections, as well. An illustration of the effect on the scale variation uncertainties is shown in Figure 6 for a parton-level ggF $H + 2j$ calculation at NLO QCD, made using the MCFM Monte Carlo generator [8]. In this distribution, the x-axis shows

the p_{T}^{Hjj} cut that events are required to pass, and the y-axis shows the differential cross section $d\sigma/dp_{\mathrm{T}}$ and scale variation error bands for the region passing the cut.

For the case of an inclusive $\geq$ 2-jet calculation, it is generally expected that the error band defined by the $\mu = \mu_0/2$ scale choice should provide a continuous upper bound for the perturbative uncertainty, with $\mu = 2\mu_0$ providing the lower bound. However, Figure 6 shows the effects induced by large Sudakov logarithms when a "2-jet-enriched" phase space is created using tight cuts on p_{T}^{Hjj}. In particular, the scale variation uncertainties shrink from $\mathcal{O}\,(30\%)$ to $\mathcal{O}\,(5\%)$ in certain regions, and the lower and upper error bands cross over at roughly $p_{\mathrm{T}}^{Hjj} < 50\,\mathrm{GeV}$.

In practice, these effects can be understood as events "migrating" back and forth between the 2 and $\geq$ 3-jet-enriched bins in a manner dependent on the scale choice, as the renormalization and factorization scales also dictate the hardness of the additional radiation in the parton-level process. In particular, the effects of migration are anti-correlated with respect to scale choice, with a smaller scale choice driving migration from the 2-jet to $\geq$ 3-jet bins, and a larger scale choice doing the opposite. Accordingly, any estimate of the perturbative QCD uncertainty in the "2-jet-enriched" region must incorporate not only cross section variation effects, but also migration effects as a consequence of the presence of Sudakov logarithms.

To this end, the *Stewart-Tackmann* (S-T) *method* can be employed to produce more sensible, symmetric perturbative uncertainties for the 2-jet-enriched cross section when placing cuts on IR-sensitive variables such as p_{T}^{Hjj}. The full implementation of the method is described in detail in Refs. [9] and [10]. The main conclusion of this formulation is that, when placing an exclusive cut on an IR-sensitive variable such as p_{T}^{Hjj}, the uncertainty on the cross section of the 2-jet-enriched phase space ($\Delta\sigma_2$)

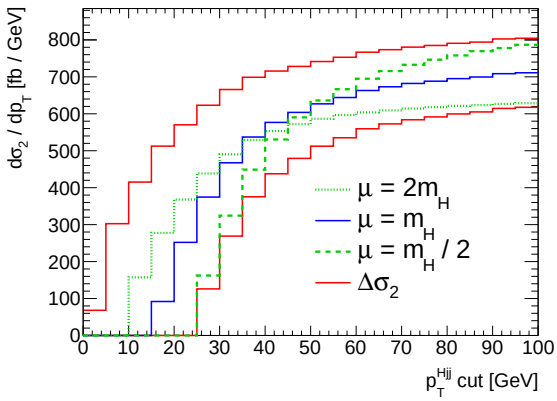


Figure 7: Effect of cuts on p_{T}^{Hjj} on the parton-level ggF $H+2j$ differential cross section $d\sigma/dp_{\mathrm{T}}$, along with the scale variation error bands defined by $\mu_{\mathrm{R}} = \mu_{\mathrm{F}} = m_H/2$ and $2m_H$. Also shown are the symmetrized error bands computed using the Stewart-Tackmann method. All distributions are derived using MCFM with $m_H = 125\,\mathrm{GeV}$, requiring jet p_{T} above 30 GeV, and $m_{\mathrm{jj}} > 120\,\mathrm{GeV}$.

can be estimated as,

$$\Delta\sigma_2^2 = \Delta\sigma_{\geq 2}^2 + \Delta\sigma_{\geq 3}^2, \tag{5}$$

where $\Delta\sigma_{\geq 2}$ is the cross sectional uncertainty for the inclusive $\geq$ 2-jet cross section, and $\Delta\sigma_{\geq 3}$ is the cross sectional uncertainty for the $\geq$ 3-jet-enriched region, i.e. that which fails the cut on p_{T}^{Hjj}. To estimate these uncertainties, parton-level NLO QCD $H+2j$ cross sections were computed using the MCFM Monte Carlo generator assuming $m_H = 125\,\mathrm{GeV}$, as well as requiring jet p_{T} above 30 GeV, and $m_{\mathrm{jj}} > 120\,\mathrm{GeV}$. Figure 7 shows the estimated S-T uncertainties as a function of the p_{T}^{Hjj} cut threshold, with the corresponding relative uncertainties given in Table 3. It is clear that the S-T method produces much more sensible estimates of the scale variation uncertainty, as the error bands both continuously envelop the nominal cross section predictions, and for loose cuts on p_{T}^{Hjj}, they also converge to the error bands created by the $\mu_0/2$ and $2\mu_0$ scale variations.

p_{T}^{Hjj} cut (GeV)	$\Delta\sigma_2/\sigma_2$ (%)
20	126
25	66.4
30	42.4
40	24.1
50	18.3
60	15.6
70	14.4
80	13.8

Table 3: Relative uncertainties for the 2-jet enriched region, $\Delta\sigma_2/\sigma_2$, in bins of rectangular cuts on p_{T}^{Hjj}, computed using the Stewart-Tackmann method.

0.1 Flattening the distribution of $p_{\mathbf{T}}^{Hjj}$

Given that p_{T}^{Hjj} was included as a training variable in the BDT based on its ability to separate VBF and ggF Higgs production, it became crucial that the BDT was trained in a way that prevented cuts on p_{T}^{Hjj} which could induce large theoretical uncertainties like those seen in Table 3. To prevent the BDT from making such cuts, the choice was made to use a modified version of the variable that is "flattened" below some threshold value, such that,

$$p_{\mathrm{T}}^{'Hjj} = \begin{cases} p_{\mathrm{T}}^{Hjj} & \text{if } p_{\mathrm{T}}^{Hjj} > X\,\mathrm{GeV} \\ X\,\mathrm{GeV} & \text{if } p_{\mathrm{T}}^{Hjj} < X\,\mathrm{GeV} \end{cases}. \tag{6}$$

With this modification, a tight cut on the BDT classifier could induce a QCD scale variation uncertainty that is, at most, equivalent to the Stewart-Tackmann uncertainty for an exclusive cut on p_{T}^{Hjj} at X GeV. An example of the shape of p_{T}^{Hjj} in ggF $H+2j$ events, before and after "flattening" the distribution, is given in Figure 8

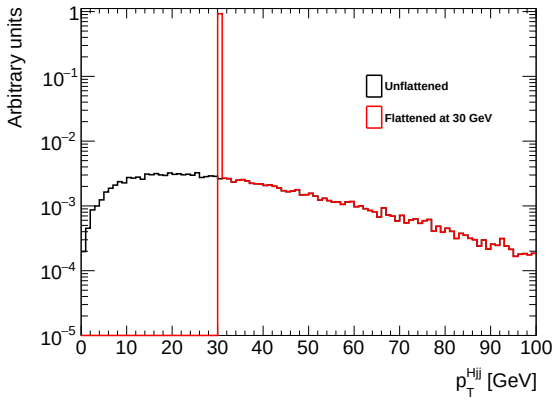


Figure 8: A comparison of the shape of p_{T}^{Hjj} in ggF $H + 2j$ events, before and after "flattening" the distribution to create $p_{\mathrm{T}}^{'Hjj}$, for a threshold value of 30 GeV.

for a threshold value of 30 GeV.

In order to balance the discriminating power of the BDT with the associated perturbative uncertainty size, an iterative scan was performed to select the optimal flattening threshold for $p_{\mathrm{T}}^{'Hjj}$. The scan proceeded through nine iterations, with the threshold incrementing in 10 GeV steps between values of 20 GeV and 100 GeV [1]. At the i^{th} iteration, the distribution of p_{T}^{Hjj} was "flattened" below X_i GeV, and the BDT was re-trained with the newly-modified $p_{\mathrm{T}}^{'Hjj}$ in place of p_{T}^{Hjj}.

Considering the relatively smaller number of iterations with respect to the scan done in Section , this optimization scan could employ a full likelihood-based fit, implemented using the ROOT-based HISTFACTORY [11] and ROOSTATS [12] statistical analysis software. In particular, a negative log-likelihood minimization technique [6] was applied across 10 bins of the BDT discriminant between $[-1, 1]$, where the expected SM signal event yields in each bin were used as pseudo-data. The log-likelihood method tests two hypotheses: the nominal case of a VBF "signal" present on top of

[1] No thresholds below 20 GeV were tested, as beyond this point, the parton-level $H + 2j$ cross section was found to be negative.

the ggF "background"; and a "background-only" hypothesis where the VBF signal strength is zero, with the excess yield assumed to be caused by a change in the dijet ggF cross section. The irreducible $qqZZ$ process was also used as a fixed background source to better emulate real experimental conditions.

The primary metric of improvement in this scan was the VBF signal significance Z_0^{VBF} over the ggF and $qqZZ$ yields, which was computed using the background-only p-value (p_0), derived from the test statistic q_μ,

$$q_\mu = -2 \ln \frac{\mathcal{L}(\mu_{\mathrm{VBF}}, \theta)}{\mathcal{L}(\hat{\mu}_{\mathrm{VBF}}, \hat{\theta})}, \tag{7}$$

such that q_0 represents the background-only hypothesis with $\mu_{\mathrm{VBF}} = 0$. Here, μ_{VBF} is the VBF signal strength parameter $\mu_{\mathrm{VBF}} = \sigma_{\mathrm{VBF}}/\sigma_{\mathrm{VBF,SM}}$, while θ represents the set of nuisance parameters for signal and background. The terms with single circumflexes ($\hat{\mu}$ and $\hat{\theta}$) denote the unconditional maximum likelihood estimates of each parameter. The value of p_0 is defined to be the probability to obtain a value of q_0 larger than the observed value under the background-only hypothesis. In particular, the value of p_0 can be expressed as,

$$p_0 = \int_{q_{0,\mathrm{obs}}}^{\infty} f(q_0|0, \hat{\theta}_0)\, dq_0. \tag{8}$$

where the test statistic shape $f(q_0|0, \hat{\theta}_0)$ is represented by a χ^2 distribution with one degree of freedom. The p_0 value is converted to the corresponding number of standard deviations (σ) in a one-sided Gaussian test, yielding the VBF significance Z_0^{VBF}.

The error on the signal strength $\Delta\mu_{\mathrm{VBF}}$ was also used as a metric of improvement, as in practice, one always retains $\mu_{\mathrm{VBF}} = 1$ from the log-likelihood fit when not testing

the background-only hypothesis. So, the relevant concern is to optimize the VBF signal strength measurement precision by minimizing the size of $\Delta\mu_{\mathrm{VBF}}$, if possible. To limit the effects of statistical uncertainty in the likelihood fit, the likelihood fits were made assuming an integrated luminosity of $\mathcal{I} = 100\,\mathrm{fb}^{-1}$, which is roughly the amount of data recorded by the LHC up to the end of 2017.

In general, performing a cut on a multivariate discriminant produces non-linear cuts on the constituent training variables. As it pertains to IR-sensitive variables such as p_{T}^{Hjj}, placing a loose cut on the BDT score should not induce the same level of uncertainty as a tight cut, as the phase space isolated by each cut will contain different fractions of exclusively 2-jet and $\geq$ 2-jet events. To account for this effect in the iterative scan, the effective ggF perturbative uncertainty in a given BDT score bin was computed using a weighted average of S-T uncertainties for different exclusive cuts on p_{T}^{Hjj}. In particular, the relative uncertainty on the ggF yield in a given BDT bin was calculated as,

$$\left(\frac{\Delta\sigma_2^{\mathrm{tot}}}{\sigma_2}\right)^2 = \sum_{i=1}^{N}\left(\frac{\Delta\sigma_2^i}{\sigma_2^j} \times h(x_i)\right)^2 \tag{9}$$

Here, i is the histogram bin index, h is an N-bin probability density histogram of p_{T}^{Hjj} built from events in the specified BDT score range, and x_i is the central value of the i^{th} histogram bin. In addition, σ_2^i is the 2-jet-enriched cross section for events passing a cut $p_{\mathrm{T}}^{Hjj} < x_i$, and $\Delta\sigma_2^i$ is the corresponding S-T uncertainty for such a cut, with both values being derived from samples generated with MCFM, as described above. The above weighted sum operates as follows: In the limit that the BDT score cut induces tight restrictions on p_{T}^{Hjj}, the shape of the p_{T}^{Hjj} histogram will approach a delta function at the flattening threshold, hence the effective uncertainty

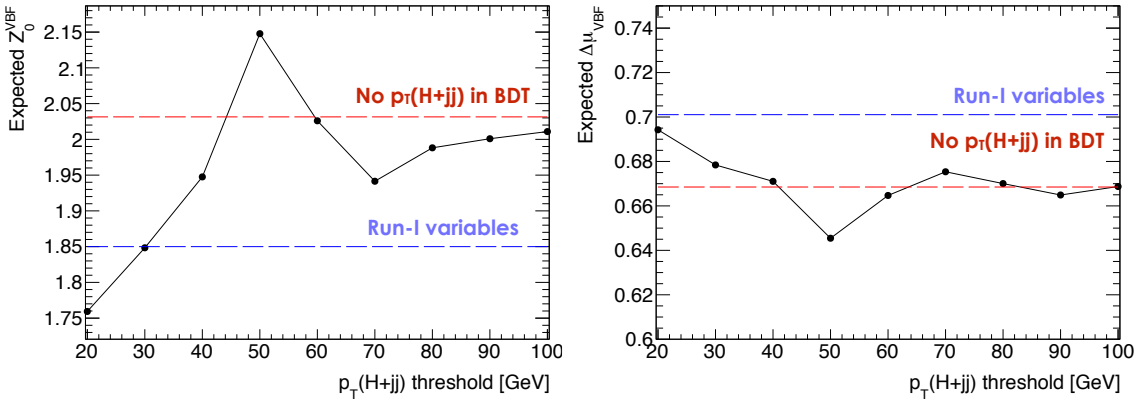


Figure 9: Distribution of expected VBF significance (left) and signal strength uncertainty (right) using a likelihood fit across 10 BDT bins in $[-1, 1]$, under different flattening thresholds of p_{T}^{Hjj}. Also shown are analogous values for BDTs trained using the Run-I variables (blue line), and a BDT trained without p_{T}^{Hjj} (red line).

will approach the S-T uncertainty for a rectangular cut at said threshold. Conversely, when the BDT score range of interest produces only loose restrictions on p_{T}^{Hjj}, the total uncertainty will shrink as the bins with high p_{T}^{Hjj} are more populated.

The intention of this scan was to find a threshold for flattening p_{T}^{Hjj} which maximized Z_0^{VBF}, while limiting the effect that the (potentially large) perturbative uncertainty has on $\Delta\mu_{\mathrm{VBF}}$. The resultant distribution of the expected VBF signal significance at each threshold choice is shown in Figure 9, along with the expected signal strength uncertainty, $\Delta\mu_{\mathrm{VBF}}$. Also shown on these plots are the expected Z_0^{VBF} and $\Delta\mu_{\mathrm{VBF}}$ for two alternate scenarios: a BDT trained using the list of "optimized" variables derived in Section , minus p_{T}^{Hjj}; and the same variables used to train the Run-I BDT discriminant.

It is evident from this plot that a balance of discriminating power is found at a threshold of 50 GeV, where Z_0^{VBF} is maximized. For thresholds below this point, the presence of larger S-T uncertainties for lower p_{T}^{Hjj} allows the ggF component

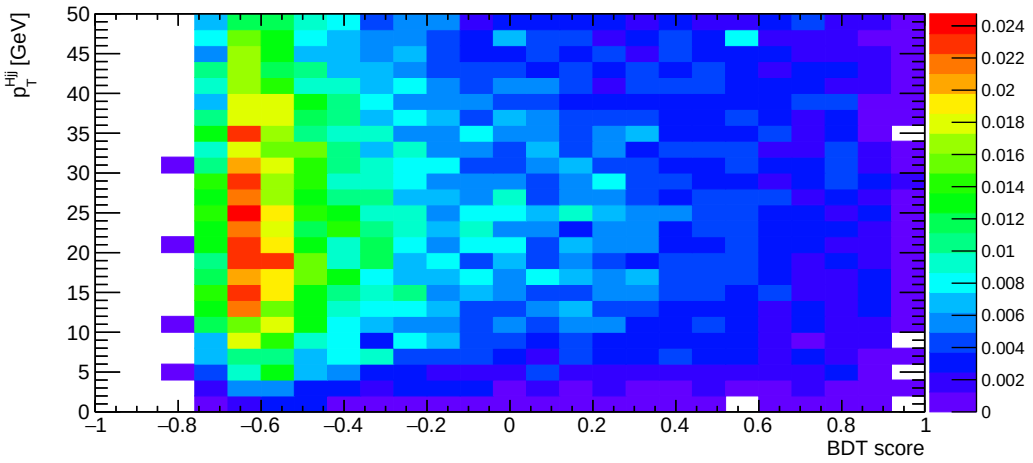


Figure 10: Distribution of p_{T}^{Hjj} vs. BDT score in ggF $H \to ZZ^* \to 4\ell$ events with $m_{\mathrm{jj}} > 120\,\mathrm{GeV}$ and $p_{\mathrm{T}}^{Hjj} < 50\,\mathrm{GeV}$, indicating that the trained BDT does not have a visible correlation with p_{T}^{Hjj} in this region of phase space.

to progressively "absorb" the VBF contribution, shrinking the signal significance, and increasing the signal strength uncertainty. Conversely, higher thresholds reduce the utility of p_{T}^{Hjj}, reducing it effectively to a delta function, such that it no longer has power to discriminate VBF and ggF production. Consequently, the variable is used successively less often to make decisions in the BDT. In this limit, the results asymptotically reach a point which is equivalent to a BDT training made without the use of p_{T}^{Hjj}, as one might expect.

However, naively binning this variable below the given threshold may not totally eliminate this problematic behaviour, as correlations with other training variables may still induce implicit cuts on p_{T}^{Hjj}. To validate that the BDT is not entering this problematic region of phase space, the shape of p_{T}^{Hjj} below 50 GeV was plotted against the full range of the BDT score, as shown in Figure 10. If the classifier is sculpting the distribution of p_{T}^{Hjj} to favour small p_{T} values at high BDT scores, then one would expect the 2D histogram of p_{T}^{Hjj} vs. BDT score to show a distinct anti-

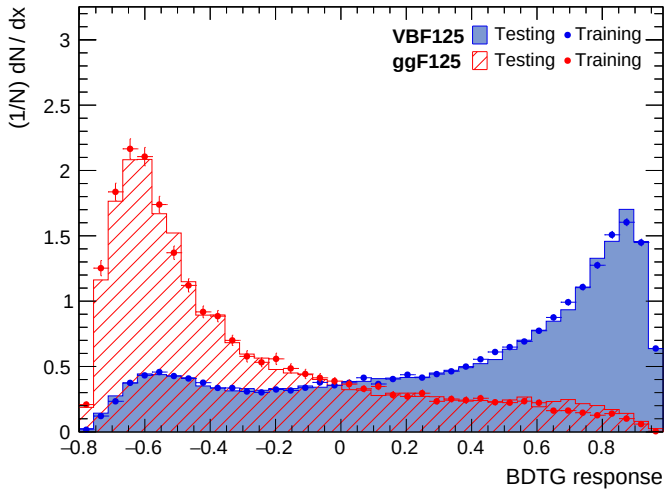


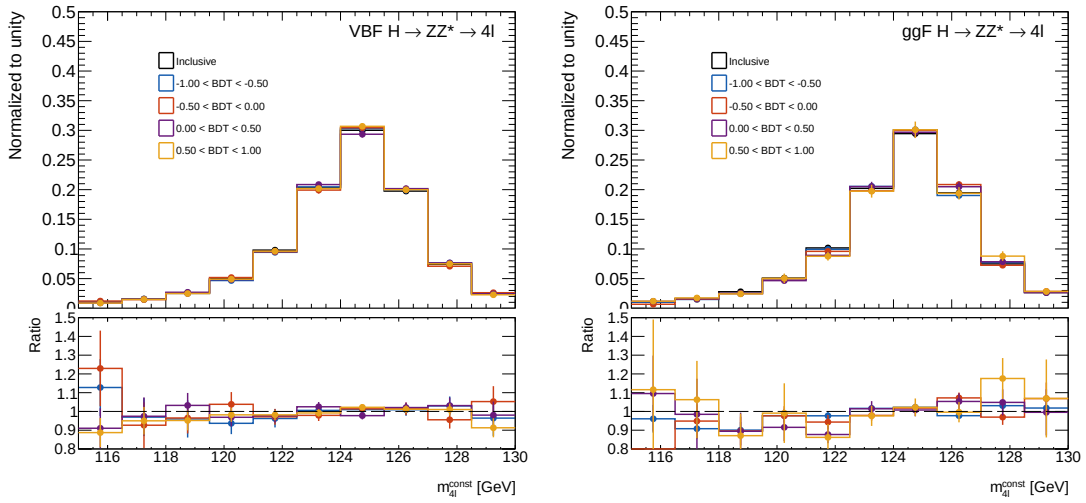
Figure 11: Superimposition of the BDT response for the training and testing samples of VBF and ggF $H \to 4\ell$ events.

correlation. However, it is clear from this distribution that no meaningful correlation is observed between the two variables. Therefore, the BDT does not induce cuts in the "IR-sensitive" region of the p_{T}^{Hjj} phase space, avoiding the induction of large theoretical uncertainties.

Overtraining and bias tests

Another consideration when training a BDT is that it should not bias or "sculpt" any parameter of interest in the analysis – for example, the invariant mass of the four lepton system $m_{4\ell}$, which is reconstructed as the Higgs boson mass. Furthermore, the classifier should be robust enough that it does not favour training events which are not properly represented in the testing samples – a behaviour known as "overtraining".

Superimposed distributions of the BDT response for the training and testing samples are shown in Figure 11 for both signal and background events. It is evident that

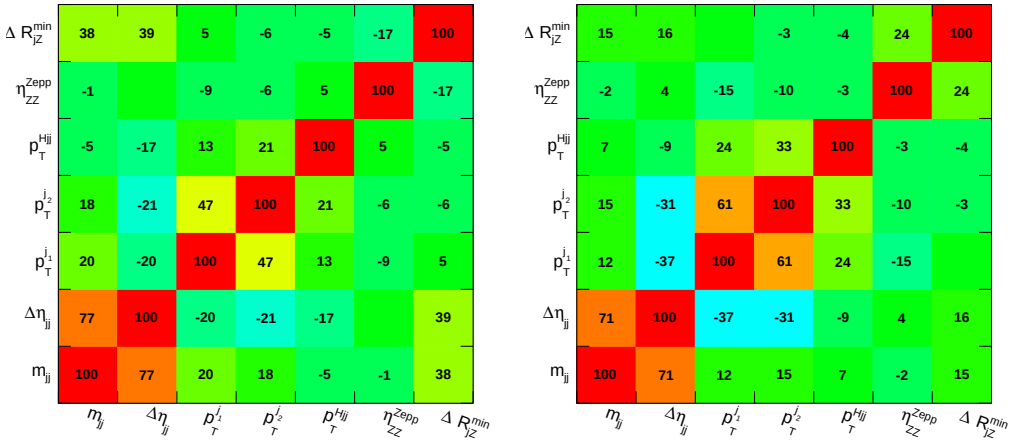


(a) VBF $H \rightarrow ZZ^* \rightarrow 4\ell$ **(b)** ggF $H \rightarrow ZZ^* \rightarrow 4\ell$

Figure 12: Shape of $m_{4\ell}$ in different regions of the BDT discriminant for both VBF (left) and ggF (right) production.

there is good agreement in the discrimiant shapes of the training and testing samples for both signal and background, allowing for some fluctuations due to the finite training sample sizes. To check that no sculpting or bias was induced on the shape or peak position of $m_{4\ell}$ through use of the BDT training, distributions of the constrained mass spectra for VBF and ggF $H \rightarrow ZZ^* \rightarrow 4\ell$ events were compared in different regions of the BDT score. In the ideal scenario where there is no correlation between the BDT discriminant and $m_{4\ell}$, the shapes of the mass distributions in each region should be approximately identical, allowing for bin-by-bin fluctuations in regions of limited event yields. The resultant plots of $m_{4\ell}$ are shown in Figure 12 for each production type, in BDT score regions delineated by bin edges of $[-1, -0.5, 0, 0.5, 1]$. Ultimately, it was found that there is no meaningful effect on the shape of the $m_{4\ell}$ spectrum from placing cuts on the BDT score.

One can also compare the LCCs between training variables at reconstruction-



(a) LCCs (VBF) **(b)** LCCs (ggF)

Figure 13: Linear correlation coefficients for the BDT training produced using fully-simulated samples of VBF and ggF $H \to 4\ell$ events.

level for the VBF and ggF processes, to ensure that the conclusions found from the truth-level discriminating variable optimization carry over once detector effects are factored in. While the magnitudes of the correlations will differ, owing particularly to reconstruction inefficiencies and the presence of pile-up jets, the relative degrees and signs of the correlations are consistent in each training with the findings at truth-level. Worth noting is that the POWHEG NNLOPS sample used to model ggF production is LO for $\geq$ 2-jet events, so the agreement in the correlation structure between the reconstruction-level NNLOPS sample and the truth-level, NLO QCD $H + 2j$ sample (generated via the POWHEG HJJ generator) indicates a good degree of robustness in the choice of training variables.

Discriminants for other categories

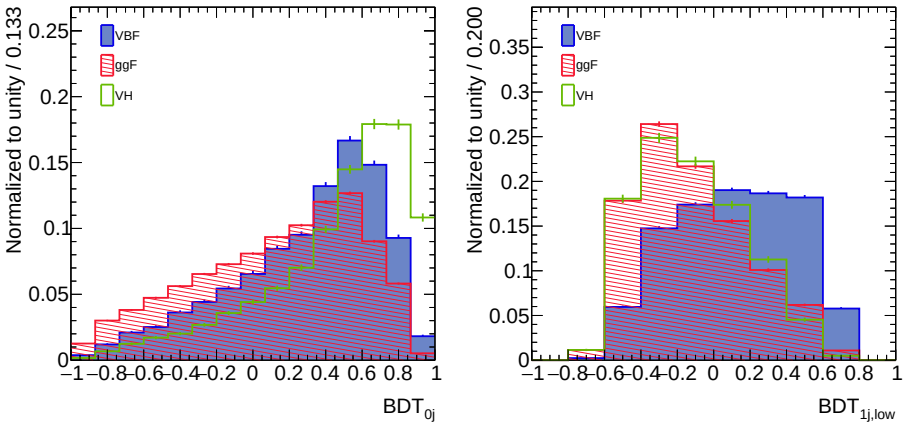
Although the focus of this chapter was on the VBF $H \to ZZ^* \to 4\ell$ BDT discriminant, the statistical analysis methodology in the $H \to ZZ^* \to 4\ell$ channel also relies on additional observables built from multivariate discriminants which are sensitive to the different SM Higgs boson production mechanisms, and applied in categories delineated by exclusive jet bins. The discriminants are used either to separate Higgs signal from the irreducible $qqZZ$ background, or to distinguish Higgs boson production modes in categories with small experimental backgrounds. Along with the classifier described in this chapter, the following discriminants are used:

- **0-jet category:** A BDT is trained to separate ggF production from $qqZZ$ background
- **1-jet, low and medium p_{T}^H categories:** A BDT is trained to separate VBF Higgs signal from background + non-VBF Higgs signal
- $\geq$ 2**-jet, VH-hadronic-enriched category:** A BDT is trained to separate hadronic VH Higgs signal from background + non-hadronic VH Higgs signal

The 0-jet discriminant is trained using p_{T}^H and η_H, as well as a kinematic discriminant built from LO matrix element calculations of the $qqZZ$ and ggF processes. The discriminant for the VH-enriched category is trained using the same input variables as the VBF discriminant, with the exception of p_{T}^{Hjj}. Lastly, the 1-jet discriminant for the low and medium p_{T}^H categories are built from the leading jet p_{T} and pseudo-rapidity, as well as ΔR between the leading jet and ZZ^* system.

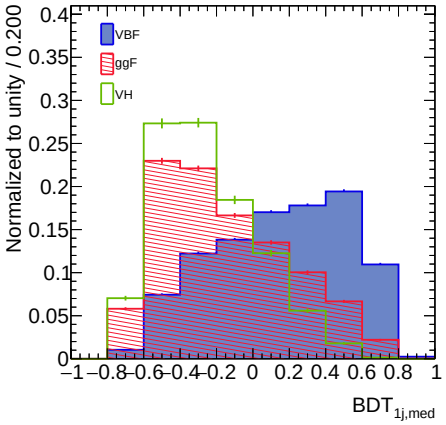
For each signal process, one-dimensional, binned probability density functions are constructed using the BDT discriminant shapes in each analysis category. Event

counting is used in the 1-jet category with $p_{\mathrm{T}}^{H} > 120\,\mathrm{GeV}$, as well as the VBF-enriched category with $p_{\mathrm{T}}^{j_1} > 200\,\mathrm{GeV}$, and the leptonic VH-enriched and ttH-enriched categories. To form the PDF in the 0-jet category, 15 bins with even bin width are used spanning a range from -1 to 1. In all the other categories, 10 bins with even bin width are used. The shapes of the per-category discriminants are shown in Figure 14 for each Higgs boson production mode used in the various trainings.

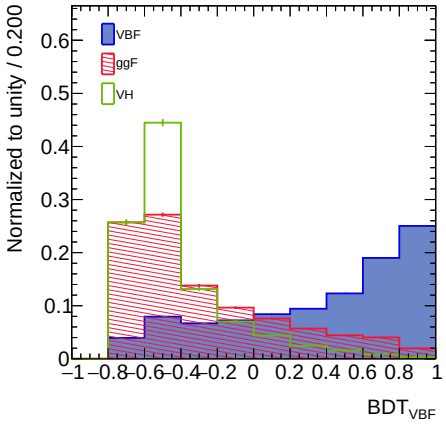


(a) 0-jet

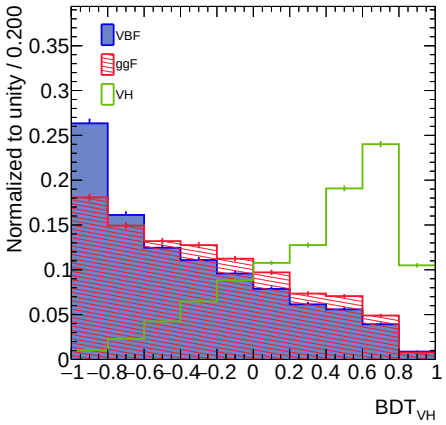
(b) 1-jet, $p_{\mathrm{T}}^{H} < 60\,\mathrm{GeV}$



(c) 1-jet, $60 < p_{\mathrm{T}}^{H} < 120\,\mathrm{GeV}$



(d) $\geq$ 2-jet, VBF-enriched



(e) $\geq$ 2-jet, VH-enriched

Figure 14: Per-category discriminants used in the $H \to ZZ^* \to 4\ell$ analysis, shown for each major considered Higgs boson production mode.

References

[1] ATLAS Collaboration, *Measurement of Higgs boson production in the diphoton decay channel in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector*, Phys. Rev. **D90** (2014) no. 11, 112015, arXiv:1408.7084 [hep-ex].

[2] ATLAS Collaboration, *Measurements of Higgs boson production and couplings in the four-lepton channel in pp collisions at center-of-mass energies of 7 and 8 TeV with the ATLAS detector*, Phys. Rev. **D91** (2015) no. 1, 012006, arXiv:1408.5191 [hep-ex].

[3] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, H. Voss, et al., *TMVA: Toolkit for Multivariate Data Analysis*, Proceedings of Science **ACAT** (2007) 040, arXiv:physics/0703039.

[4] J. Friedman, *Stochastic gradient boosting*, Computational Statistics & Data Analysis **38** (2002) no. 4, 367 – 378.

[5] LHC Higgs Cross Section Working Group Collaboration, *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*, arXiv:1610.07922 [hep-ph].

[6] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. **C71** (2011) 1554, 1–19, arXiv:1007.1727 [physics.data-an].

[7] S. Gangal, *Precision Predictions of Exclusive Jet Cross Sections at the LHC*. PhD thesis, U. Hamburg, Dept. Phys., Hamburg, 2015. http://bib-pubdb1.desy.de/search?cc=Publication+Database&of=hd&p=reportnumber:DESY-THESIS-2015-042.

[8] J. M. Campbell and R. Ellis, *MCFM for the Tevatron and the LHC*, Nucl. Phys. Proc. Suppl. **205-206** (2010) 10–15, arXiv:1007.3492 [hep-ph].

[9] S. Gangal and F. J. Tackmann, *Next-to-leading-order uncertainties in Higgs+2 jets from gluon fusion*, Phys. Rev. **D87** (2013) no. 9, 093008, arXiv:1302.5437 [hep-ph].

[10] I. W. Stewart and F. J. Tackmann, *Theory Uncertainties for Higgs and Other Searches Using Jet Bins*, Phys. Rev. **D85** (2012) 034011, arXiv:1107.2117 [hep-ph].

[11] ROOT Collaboration, K. Cranmer, G. Lewis, L. Moneta, A. Shibata, and W. Verkerke, *HistFactory: A tool for creating statistical models for use with RooFit and RooStats*, CERN-OPEN-2012-016, New York U., New York, Jan, 2012. https://cds.cern.ch/record/1456844.

[12] L. Moneta, K. Cranmer, G. Schott, and W. Verkerke, *The RooStats project*, in *Proceedings of the 13th International Workshop on Advanced Computing and Analysis Techniques in Physics Research*. 2010. arXiv:1009.1003.