tHq(ML): $2\ell + \tau_{had}$

Lepton assignment for the 2ℓ SS

Pablo Martínez-Agulló

Monday, June 22nd, 2022

IFIC (CSIC-UV)

Introduction - Study of y_t

The discovery of a Higgs boson by the ATLAS and CMS experiments in 2012 opened a new field for exploration in the realm of particle physics. In order to better understand the Standard Model (SM) of particle physics, it is of prominent interest to understand the Yukawa coupling of the Higgs boson to the top quark (y_t), being the latter the most massive fundamental particle and, consequently, the one with the largest coupling to the Higgs boson.

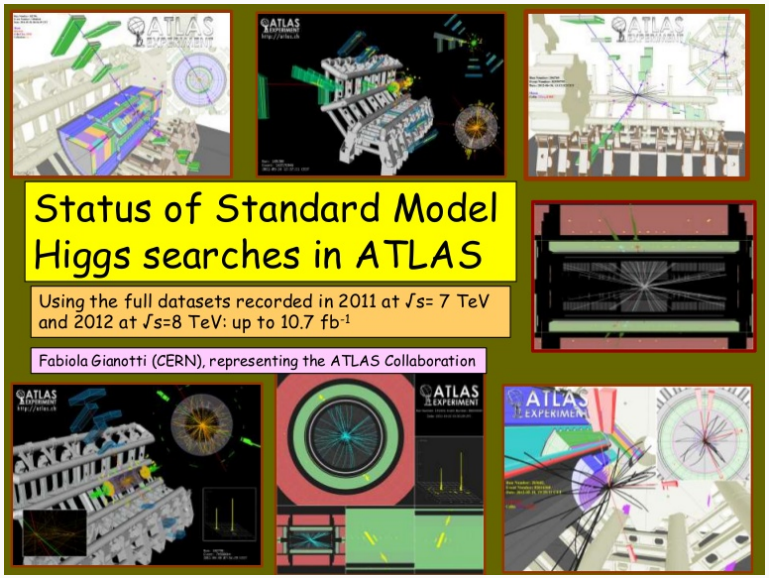


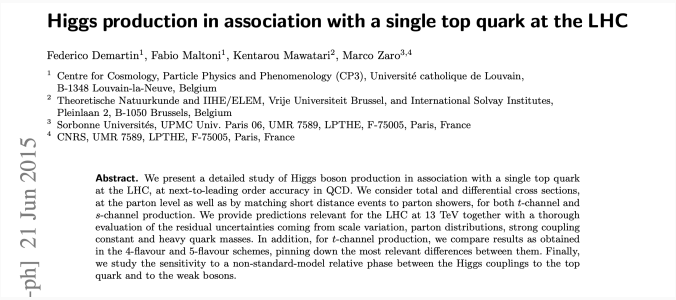
Figure 1: Presentation of Higgs discovery in Comic Sans.

Figure 2: A very polite and friendly top quark.

Introduction - Importance of *tHq*

Higgs boson asociated production with:

- a paier of top quarks ($t\bar{t}H$): Sensible to the magnitude of y_t
- a single top quark (tHq): Sensible to the magnitude and sign of y_t



Higgs production in association with a single top quark at the LHC

Federico Demartin[1], Fabio Maltoni[1], Kentarou Mawatari[2], Marco Zaro[3,4]

[1] Centre for Cosmology, Particle Physics and Phenomenology (CP3), Université catholique de Louvain, B-1348 Louvain-la-Neuve, Belgium
[2] Theoretische Natuurkunde and IIHE/ELEM, Vrije Universiteit Brussel, and International Solvay Institutes, Pleinlaan 2, B-1050 Brussels, Belgium
[3] Sorbonne Universités, UPMC Univ. Paris 06, UMR 7589, LPTHE, F-75005, Paris, France
[4] CNRS, UMR 7589, LPTHE, F-75005, Paris, France

Abstract. We present a detailed study of Higgs boson production in association with a single top quark at the LHC, at next-to-leading order accuracy in QCD. We consider total and differential cross sections, at the parton level as well as by matching short distance events to parton showers, for both t-channel and s-channel production. We provide predictions relevant for the LHC at 13 TeV together with a thorough evaluation of the residual uncertainties coming from scale variation, parton distributions, strong coupling constant and heavy quark masses. In addition, for t-channel production, we compare results as obtained in the 4-flavour and 5-flavour schemes, pinning down the most relevant differences between them. Finally, we study the sensitivity to a non-standard-model relative phase between the Higgs couplings to the top quark and to the weak bosons.

-ph] 21 Jun 2015

The only way of directly measuring the absolute magnitude and sign of y_t is via the associated Higgs production with a single top quark (tHq) [1]. This is due to the fact that the two leading order Feynman diagrams for the tHq production interfere with each other depending on the y_t sign. Current experimental constrains on y_t favour the SM predictions, but an opposite sign with respect to the expectations of the SM is not completely excluded yet [2].

[1] F. Demartin, F. Maltoni, K. Mawatari and M. Zaro, *Eur. Phys. J. C* **75** (2015) 267.
[2] CMS Collaboration, *Eur. Phys. J. C* **81** (2021) 378

Intro - LO diagrams

The destructive interference between the two LO diagrams generate a very small cross section for the *tHq* production

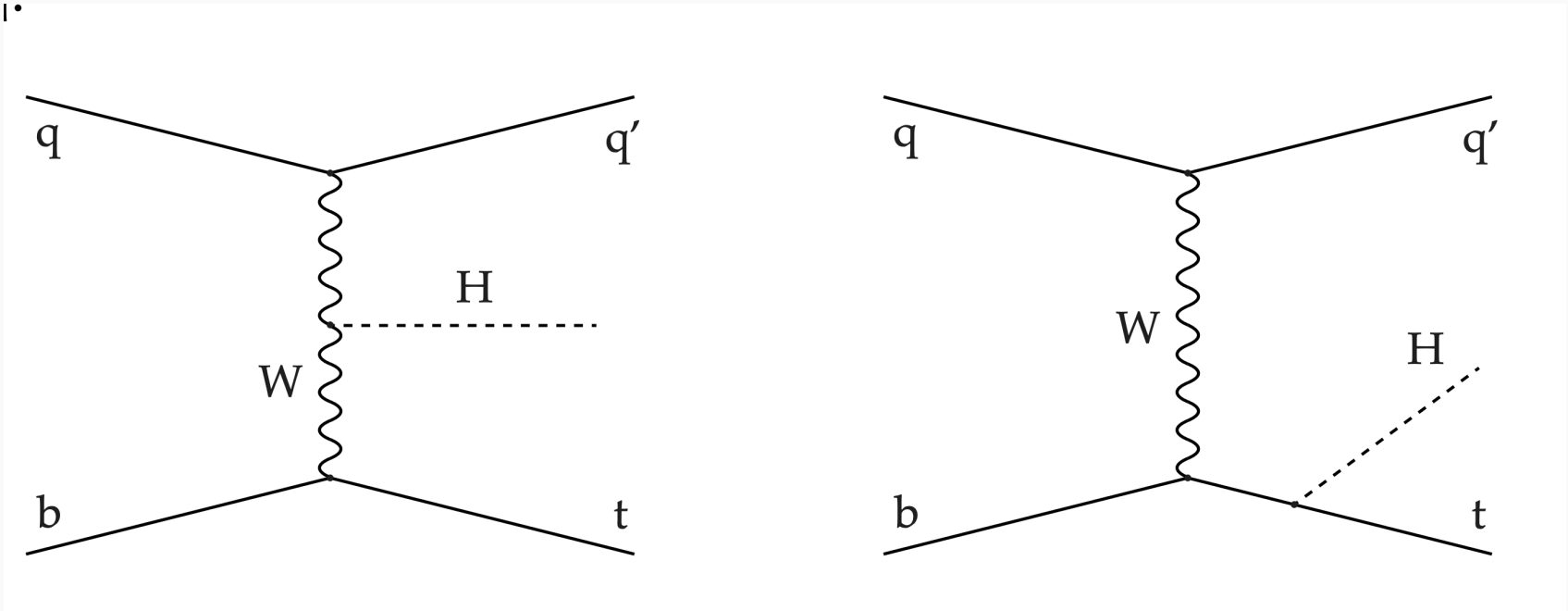


Figure 3: Leading order diagrams for *tHq* production.

Intro - Channels

The ongoing ATLAS effort on *tHq* searches is classified in channels according to the light lepton (ℓ) and hadronic tau (τ_{had}) multiplicities:

#	0 τ_{had}	1 τ_{had}	2 τ_{had}
1ℓ	tHq $(b\bar{b})$ 1ℓ		tHq $(WW/ZZ/\tau\tau)$ $1\ell + 2\tau_{had}$
2ℓ	tHq $(WW/ZZ/\tau\tau)$ 2ℓ SS	tHq $(WW/ZZ/\tau\tau)$ $2\ell + \tau_{had}$	
3ℓ	tHq $(WW/ZZ/\tau\tau)$ 3ℓ		

Table 1: Different channels for *tHq* production according to the presence of light-flavoured leptons and hadronicanlly-decaying taus in the final state.

Intro - $2\ell + \tau_{had}$ channels

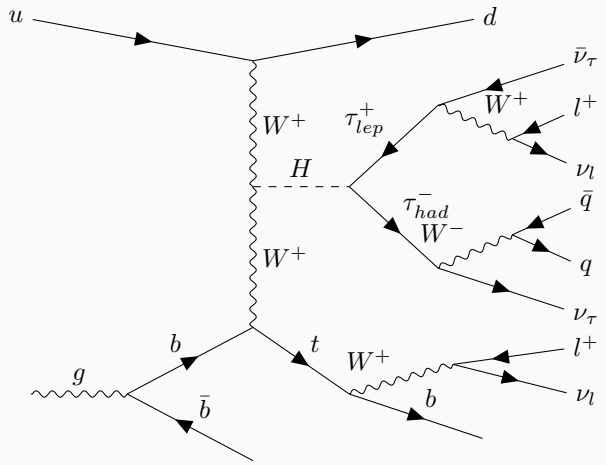


Figure 4: Representative Feynman diagram of the $2\ell + \tau_{had}$ channel.

The current ATLAS effort on *tHq* searches includes the production characterised by a final state with two light charged leptons (electrons or muons) and one hadronically decaying tau lepton (named $2\ell + \tau_{had}$ channel).

Split in two sub-channels according to the relative sign between the light leptons:

- 2ℓ OS $+ 1\tau_{had}$: The ℓ have opposite charge
- 2ℓ SS $+ 1\tau_{had}$: The ℓ have same charge

This search is exceptionally challenging due to the extremely small cross-section of the *tHq* process (70 fb). Particularly, the $2\ell + \tau_{had}$ final-state channel only accounts for a 3.5% of the total *tHq* production.

Intro - Importance of lepton assignment

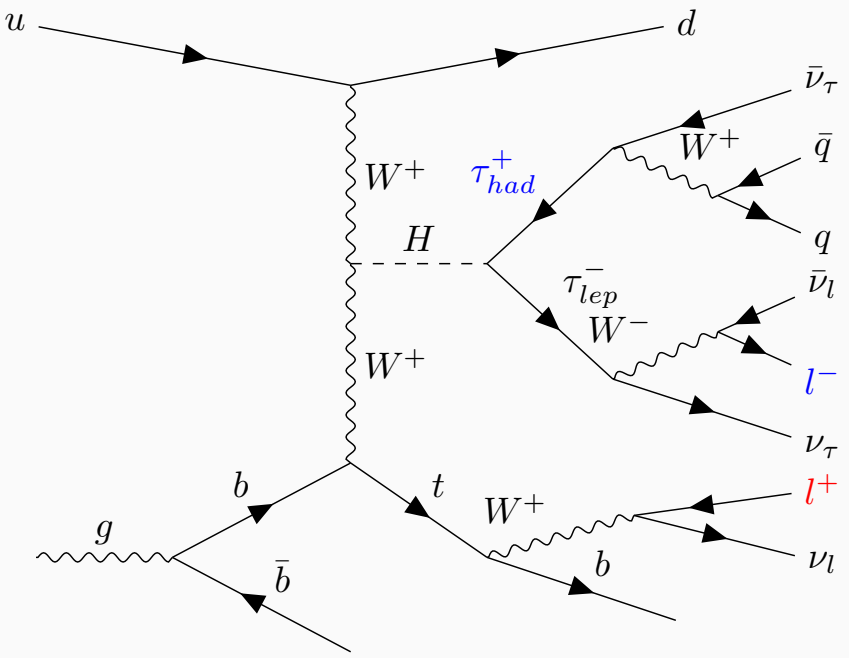
In this context, the reconstruction of the events is of primal importance in order to define variables that can be used to separate the signal events form the main backgrounds. The first step to reconstruct the event is to determine which of the light leptons comes from the Higgs boson system and which from the top quark system.

Assuming that the τ_{had} is produced from the Higgs boson decay, the assignment of the lepton's parent is unambiguous when these have opposed electric charge. Contrarily, when the light leptons have the same electric charge ($2\ell\,\mathrm{SS} + 1\tau_{had}$), **it is not possible to determine the origin a priori**.

Lepton origin of OS leptons

To assign the origin of the lepton means to determine for each lepton whether it is originated from the Higgs decay products or from the top decay chain. This is defined for the case in which the τ_{had} comes from the Higgs-boson decay (83.7%).

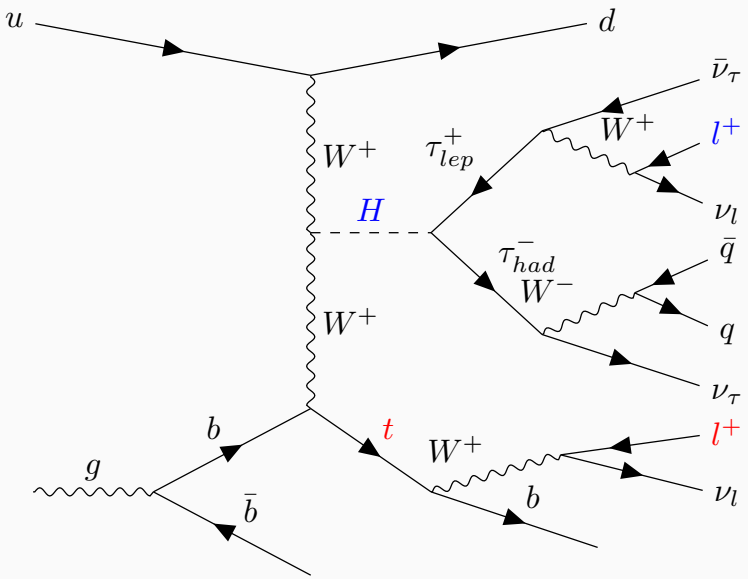
In the 2ℓ OS $+ 1\tau_{had}$ channel, the origin of the light leptons can be uniquely determined:

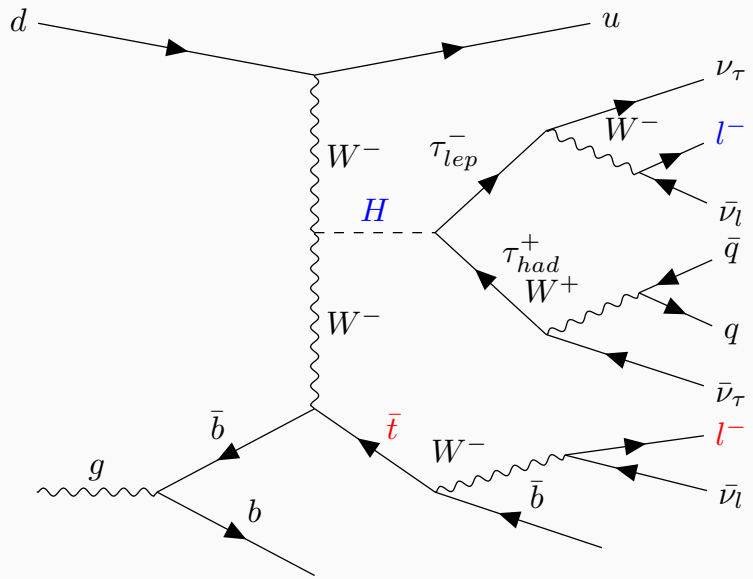


The light lepton with same charge as the τ_{had} is the one from the top-quark system while the other comes from the Higgs boson.

Lepton origin of SS leptons

When the light leptons have the same electric charge (2ℓ SS + $1\tau_{had}$), it is not possible to determine the origin a priori.





Different methods have been developed to preform the association.

- Current baseline method (see presentation) cutting directly on two mass variables ($m_{vis,H}$ and $m_{pred,t}$).
- NN based method (see presentation) based on the assumption $\Delta R(\ell_1^{reco}|\ell_{top}^{truth}) < \Delta R(\ell_2^{reco}|\ell_{top}^{truth})$.
- BDT method presented int these slides. ← Best performance.

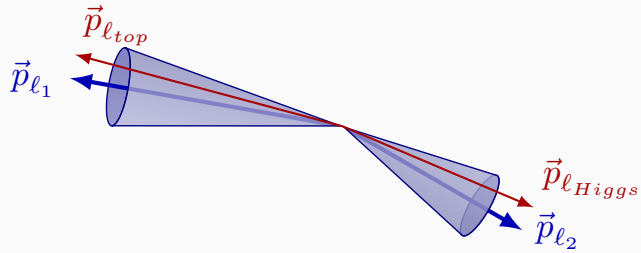
Overview of the BDT-based method

- Create labels for the training:
  - Using signal events only.
  - Parton level and reconstruction level matching via ΔR cones.
  - Two types of events:
    - Type1: ℓ_1 from Top and ℓ_2 from Higgs.
    - Type2: ℓ_1 from Higgs and ℓ_2 from Top.
- Train a BDT to classify events according to the origin of the light leptons
  - ROOT:TMVA based method
  - Optimisation of variables and hyperparameters.
  - Only positively weighted events used for training but all weights included the in test.
  - k-folding with 5 folds.
- Evaluate the BDT score for each event.
- Select the optimal threshold point for classification.

Assignment - Labeling with truth

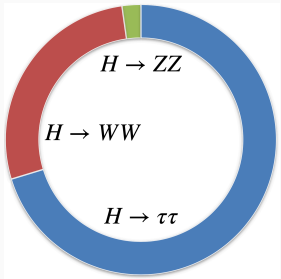
Before carrying a supervised training it is necessary to have the label (aka target variable). To do so, the truth-level and reconstruction-level information of the events is compared to determine a correspondence between the leading (ℓ_1^{reco}) and subleading (ℓ_2^{reco}) leptons and the lepton from Higgs (ℓ_{Higgs}^{truth}) and lepton from top (ℓ_{Top}^{truth}).

- Define $\Delta R \leq 0.01$ cones around the $\overrightarrow{p}_{\ell_{reco}}$.
- If within one cone a there is ℓ^{reco} and ℓ^{truth} is said that there is "a match".



  - Both leptons in 2ℓ SS $+ 1\tau_{had}$ are required to have a truth-reco match:
    - If only one the leptons matches, no label is set for the event
    - Able to assign label (89% for SS and 66% of OS) ← We only need this for SS

- Set label: `is_Lep1_from_Top`
  - `is_Lep1_from_Top = 1`: Type1 (61.1%)
  - `is_Lep1_from_Top = 0`: Type2 (38.9%)
- Done for $H \to \tau\tau$ and $H \to WW$.
  - The contribution of $H \to ZZ$ is negligible (1.9% of $2\ell + \tau_{had}$ production when the τ_{had} comes from Higgs).



Assignment - Labeling with truth

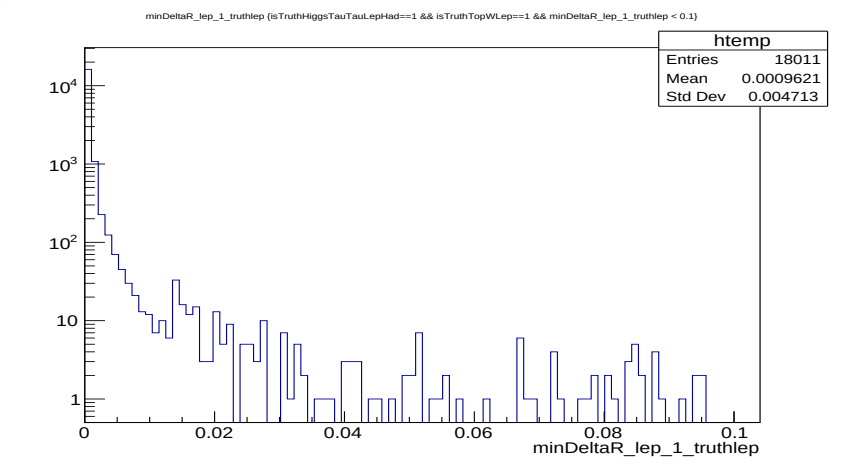


Figure 5: $min\Delta R(\ell_1^{reco}|\ell^{truth})$ in $H \rightarrow \tau\tau$

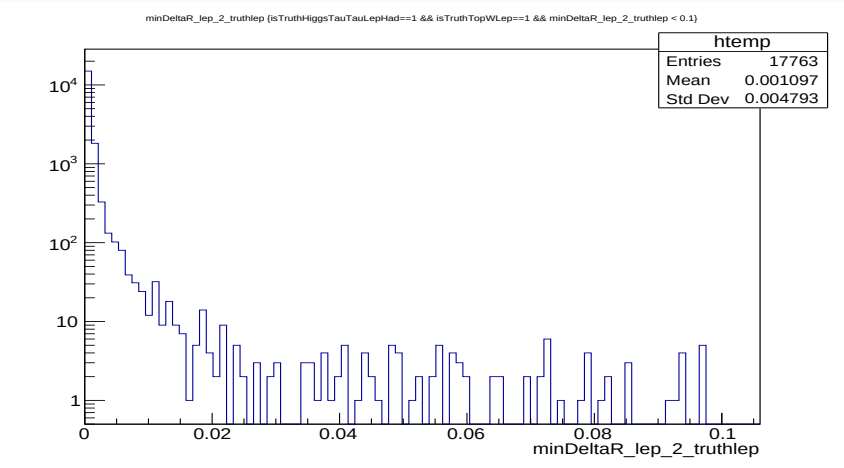


Figure 6: $min\Delta R(\ell_2^{reco}|\ell^{truth})$ in $H \rightarrow \tau\tau$

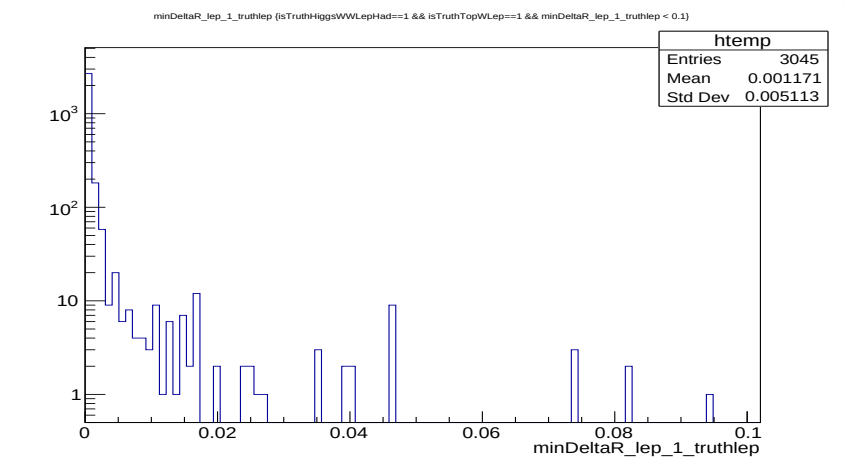


Figure 7: $min\Delta R(\ell_1^{reco}|\ell^{truth})$ in $H \rightarrow WW$

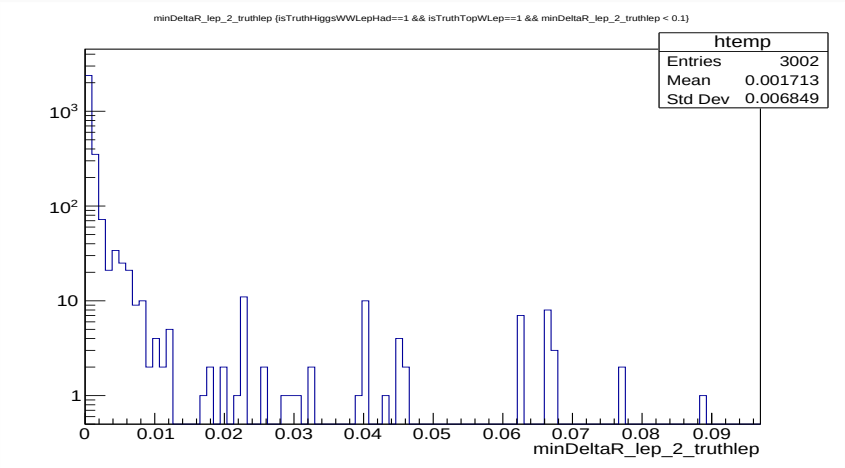


Figure 8: $min\Delta R(\ell_2^{reco}|\ell^{truth})$ in $H \rightarrow WW$

Training the MVA - Overview

- Using ROOT:TMVA library $\rightarrow$ A lot of tools included <3
- Only signal events (*346799*): Initially 37542 entries (15905 of SS).
- Requirements before training
  - Truth-reco matching (*is_Lep1_from_Top* = 1 or 0)
  - Train only over SS events[1] (*SS_LepHad* = 1)
  - τ_{had} from Higgs (*isTruthTopWLep*=1)
  - $H \rightarrow \tau\tau$ or $H \rightarrow WW$ (*isTruthHiggsTauTauLepHad*=1 OR *isTruthHiggsWWLepHad*=1)
- Using the label *is_Lep1_from_Top* as predictor
- Two types of events:
  - Type1 $\rightarrow$ is_Lep1_from_Top = 1
    Leading lep form Top and sub-leading lep from Higgs
  - Type2 $\rightarrow$ is_Lep1_from_Top = 0
    Leading lep form Higgs and sub-leading lep from Top
- Ignoring the negative weights
- Using k-folding to avoid overtraining.

[1]The bkgs are not of interest for these studies but an score is given to them.

Training the MVA - Kfold

5 Folds used: Avoids overtraining and mitigates the effect of low statistics

For each fold a BDT:

- One train
- One test ← The score which is fiannly applied is this one, never the one of the training
- Events:
  - Training: 11364 (80%)
  - Test: 2864 (20%)

Training the MVA - Optimisation - Hyperparameters

Hyperparameters: Optimised via grid search (this might be further refined)

- MaxDepth = 2 ← Maximum depth of cell tree
- Shrinkage = 0.10 ← Learning rate for GradBoost algorithm
- NTrees = 1000 ← Number of trees in method
- nCuts = 20 ← Number of grid points in variable range used in finding optimal cut in node splitting
- NegWeightTreatment[2] = IgnoreNegWeightsInTraining ← How to treat events with negative weights in the BDT training

Managed via configuration file.

[2] Discussed later

Training the MVA - Optimisation - Features

Features: TMVA automatically ranks the variables

- Several other variables have been tested (see backup).
- Correlations studied.
- TMVA ranking tool used to decide the final set.
- Keep a low number of variables: It may create conflicts in the SR and CR definition.

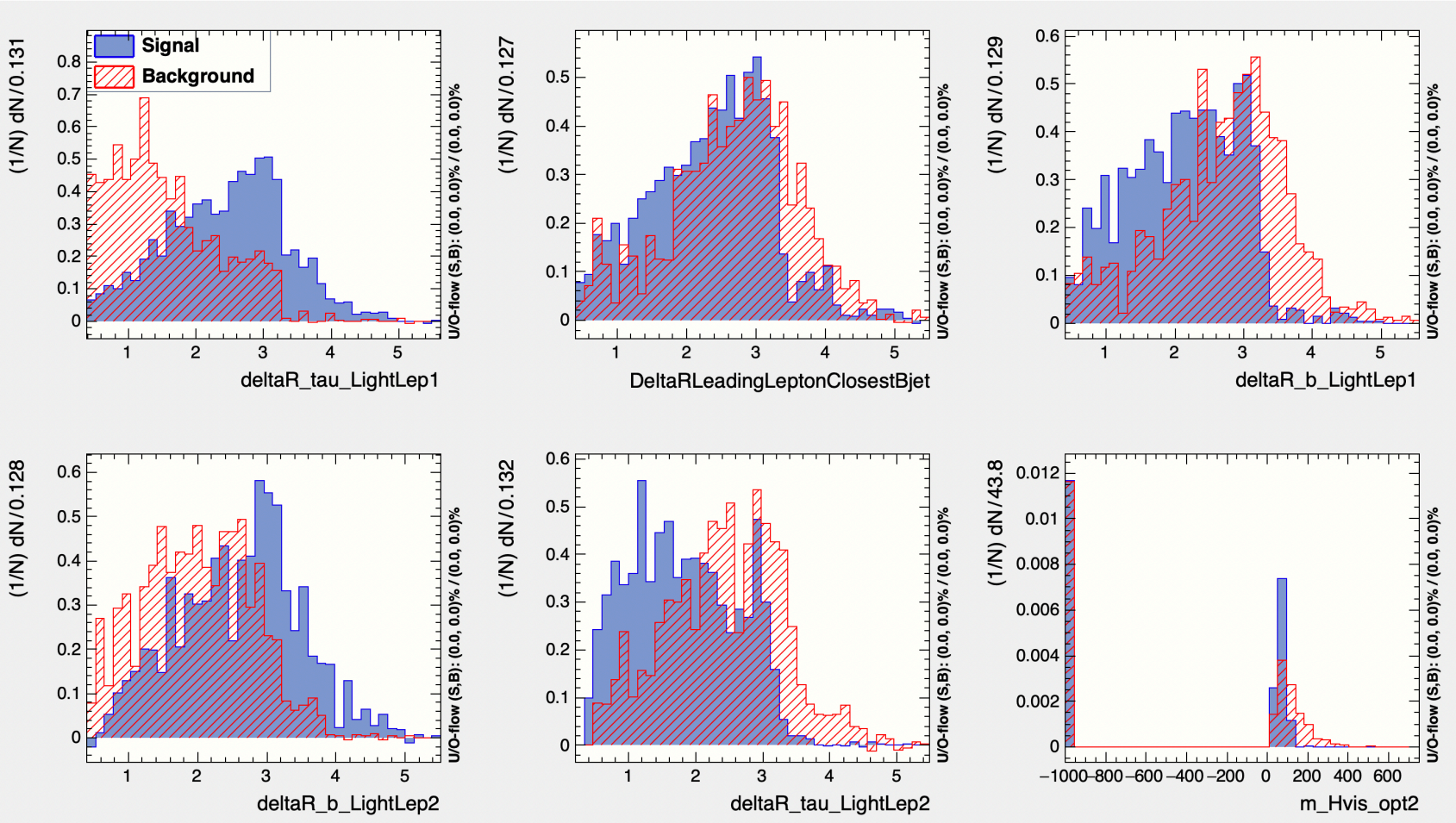
```
: ---------------------------------------
: Rank : Variable                       :
: ---------------------------------------
:    1 : deltaEta_tau_LightLep1         :
:    2 : deltaR_tau_LightLep1           :
:    3 : deltaEta_tau_LightLep2         :
:    4 : m_Hvis_opt1                    :
:    5 : deltaR_b_LightLep2             :
:    6 : deltaR_b_LightLep1             :
:    7 : deltaR_tau_LightLep2           :
:    8 : m_Hvis_opt2                    :
:    9 : DeltaRLeadingLeptonClosestBjet :
: ---------------------------------------
```

Figure 9: Ranking of variables

MVA - Variables in training

- **DeltaRLeadingLeptonClosestBjet**: ΔR between the leading lepton and the closest b-tagged jet to that lepton.
- **deltaR_b_LightLep2**: ΔR between the b-tagged jet and the sub-leading light-flavoured lepton.
- **deltaR_tau_LightLep1**: ΔR between the hadronic tau and the leading light-flavoured lepton.
- **deltaR_b_LightLep1**: ΔR between the b-tagged jet and the leading light-flavoured lepton.
- **deltaR_tau_LightLep2**: ΔR between the hadronic tau and the sub-leading light-flavoured lepton.
- **m_Hvis_opt1** : mass of the combined hadronic tau and the leading light-flavoured lepton
- **deltaEta_tau_LightLep1**: $\Delta\eta$ between the hadronic tau and the leading light-flavoured lepton.
- **deltaEta_tau_LightLep2**: $\Delta\eta$ between the hadronic tau and the sub-leading light-flavoured lepton.

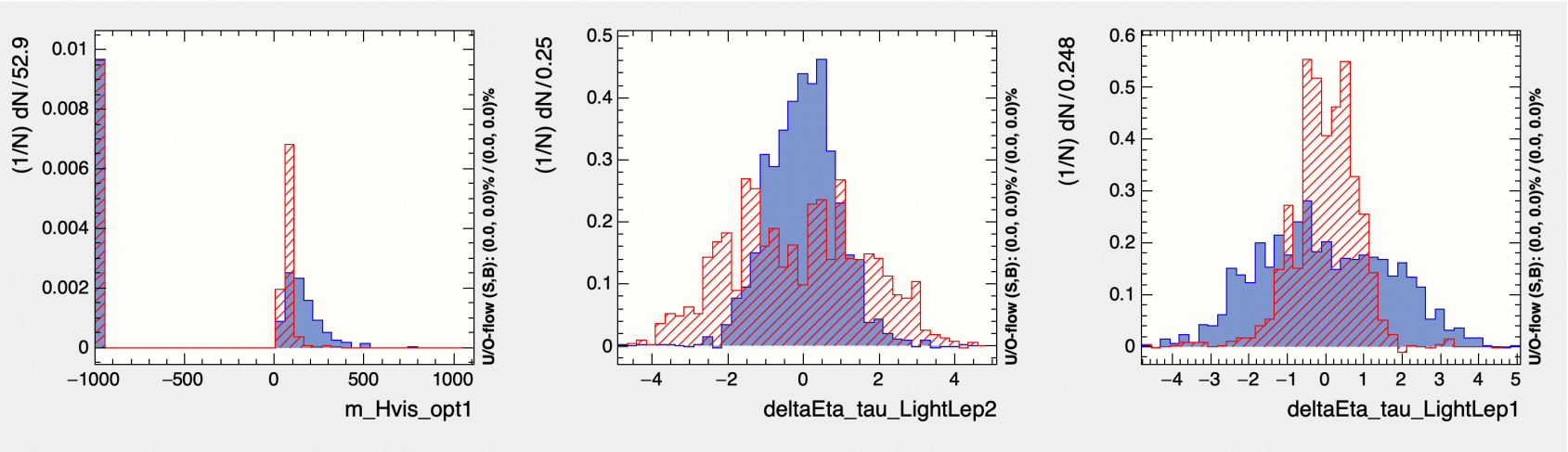
Variables in training - Distribution



This is not signal vs background but Type1 vs Type2

Only signal here

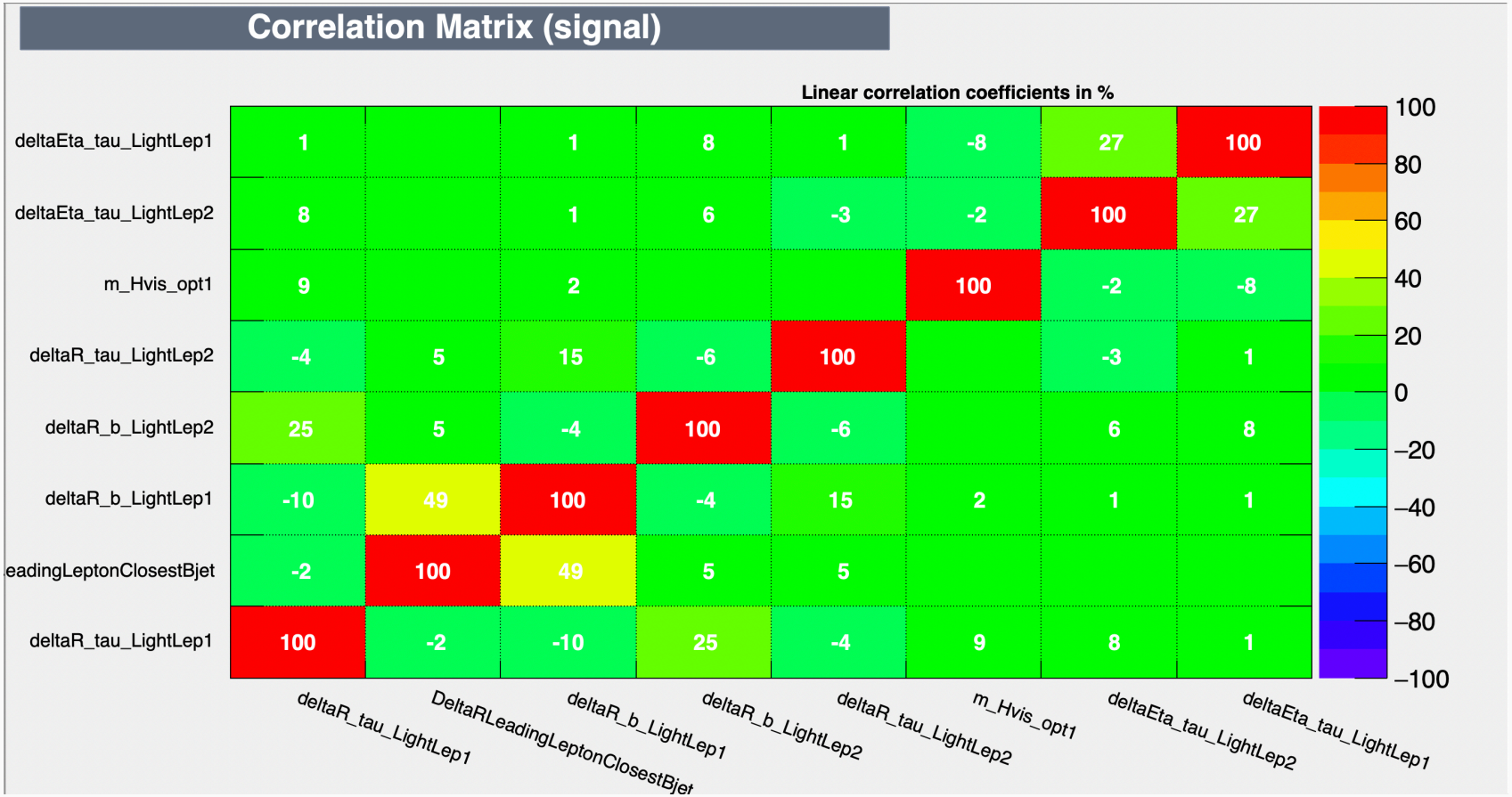
Variables in training - Distribution



This is not signal vs background but Type1 vs Type2

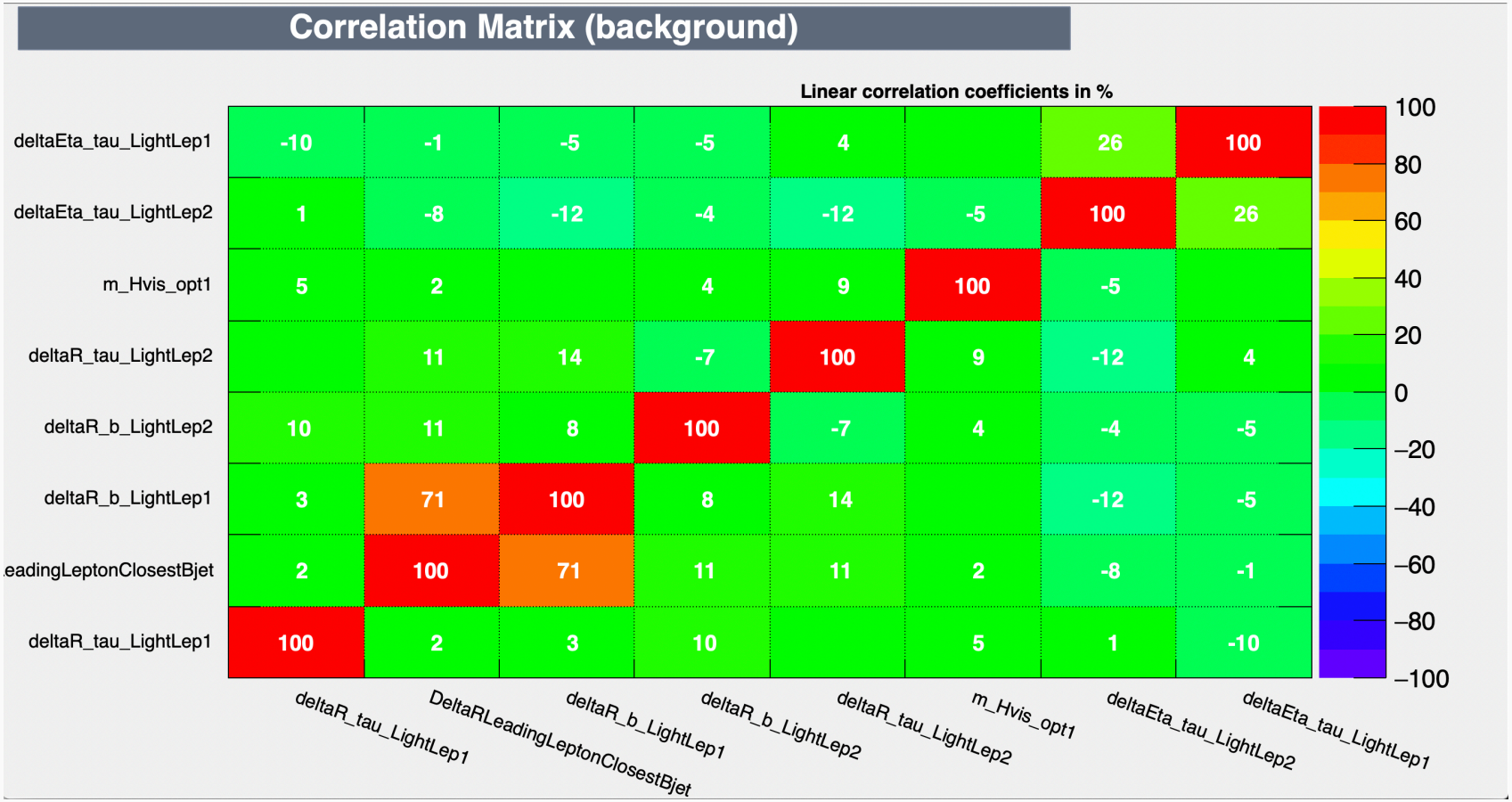
Only signal here

Variables in training - Correlations for Type 1 Events



Correlated variables have to be explored

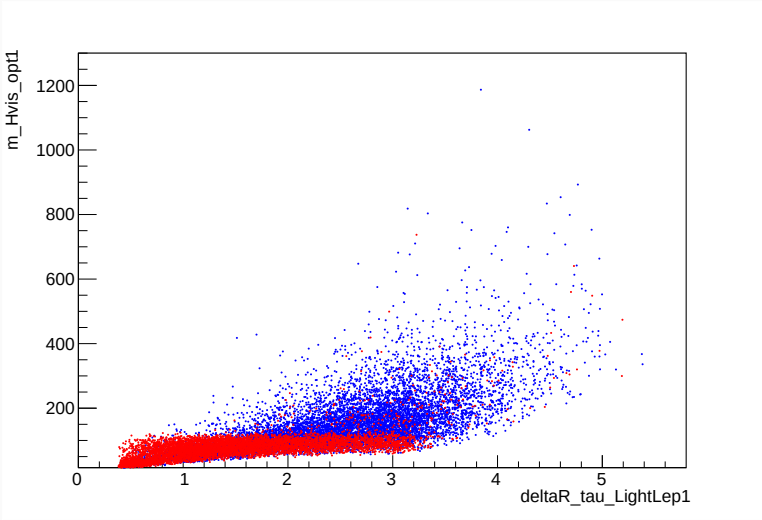
Variables in training - Correlations for Type 2 Events

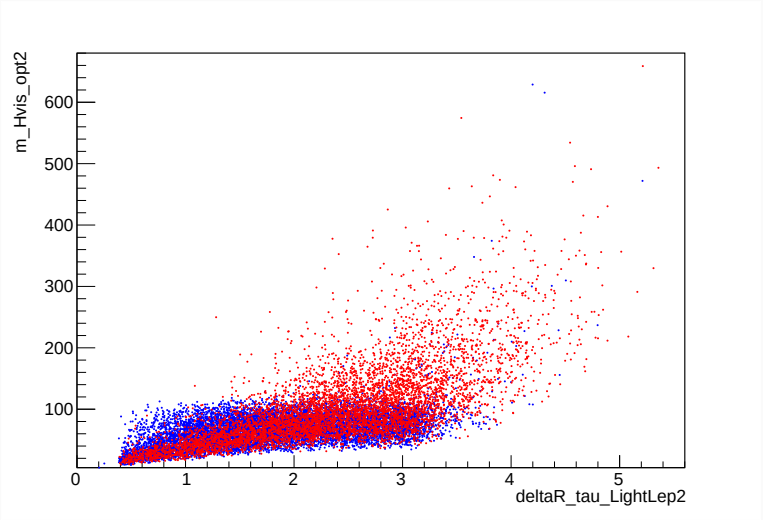


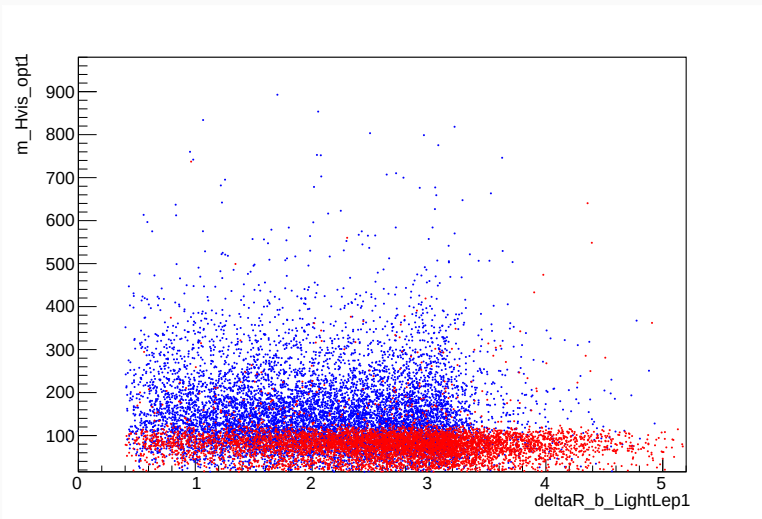
Correlated variables have to be explored

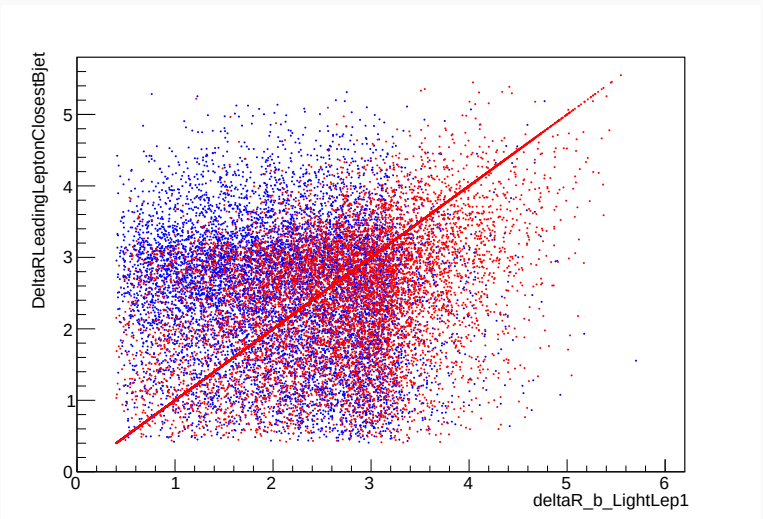
MVA - Variables in training - Correlations

The most correlated pairs of variables (according to the two previous slides) are studied distinguishing between Type1 and Type2 events:









MVA - Negative weight treatment

About a 40% of the signal entries are negatively weighted. Before rejecting these events in the training it has to be checked that the distributions are not significantly changed when doing so.

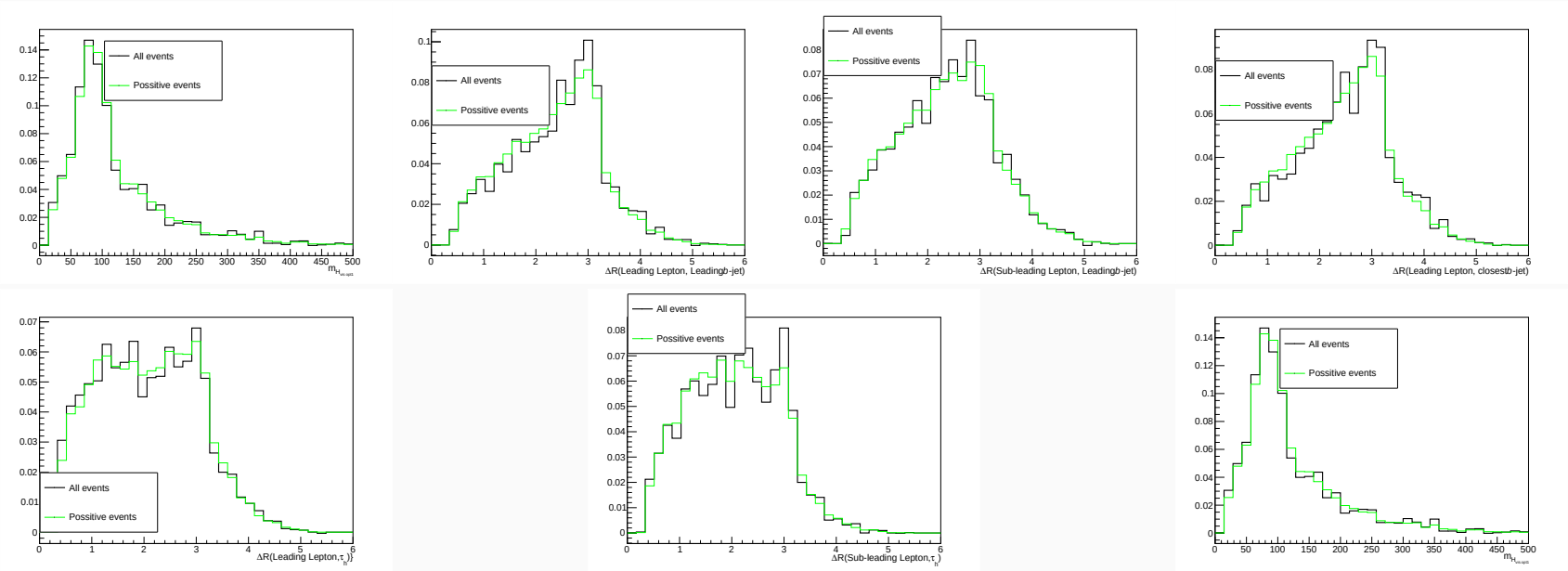
Currently using IgnoreNegWeightsInTraining

ROOT:TMVA allows several options for the NegWeightTreatment:

- IgnoreNegWeightsInTraining : Recomended one. Only trains with positively weighted events. Performance and evaluation performed with all events
- InverseBoostNegWeights: Uses the negatively weighted events for the training. The observed performance is not good.
- PairNegWeightsGlobal : Still experimental. Pair events with negative and positive weights in training sample and "annihilate´´ them.
- Pray: Not described

MVA - Negative weight treatment

We compare the shapes of the distributions with all the events to the ones with the positively weighted events only (the used for training).



Normalised. SS events only

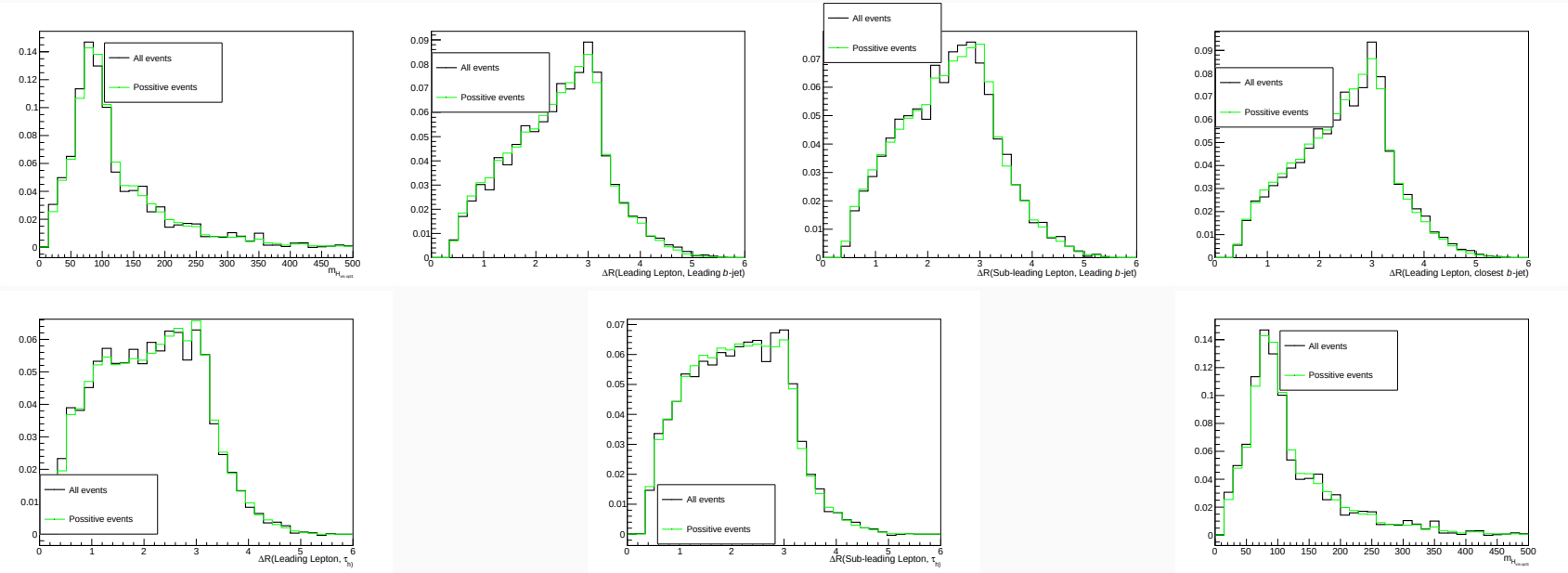
Using only the positive weights is not problematic.

Fraction of negative events: 36.1% for type 1 and 36.6% for type2.
Balanced.Removing neg weights does not bias towards type nor type2

Conclusion: Using only the positive weights is not problematic.

MVA - Negative weight treatment - (SS + OS)

We compare the shapes of the distributions with all the events to the ones with the positively weighted events only (the used for training).

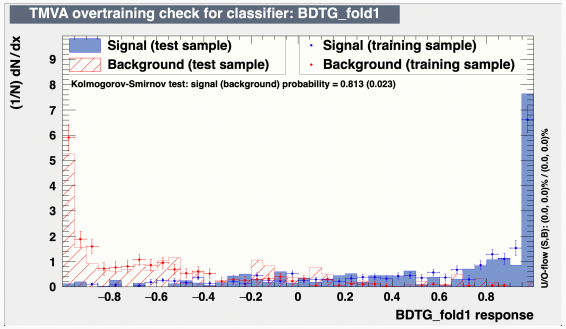


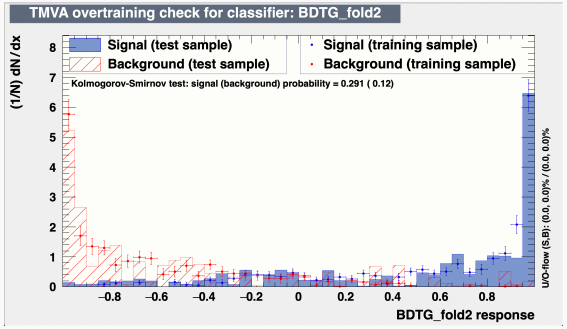
Normalised.

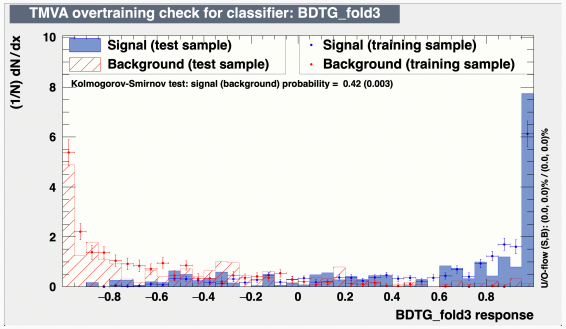
Using only the positive weights is not problematic.

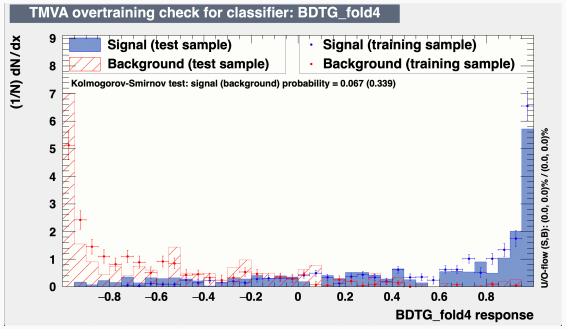
Further test: Do Type1 and Type2 have the same fraction of negative weights? Yes, 36.4% for type 1 and 36.0% for type2. Removing neg weights does not bias towards type nor type2

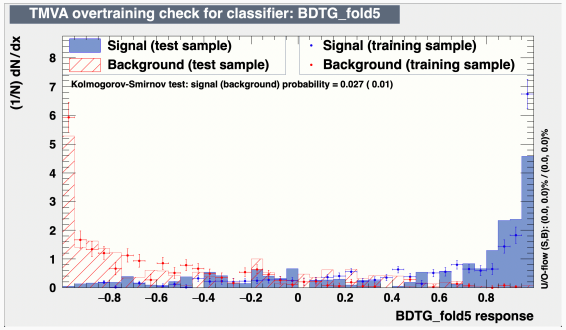
BDT result


TMVA overtraining check for classifier: BDTG_fold1
Signal (test sample)
Signal (training sample)
Background (test sample)
Background (training sample)
Kolmogorov-Smirnov test: signal (background) probability = 0.813 (0.023)
(1/N) dN/dx
BDTG_fold1 response
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.0)%


TMVA overtraining check for classifier: BDTG_fold2
Signal (test sample)
Signal (training sample)
Background (test sample)
Background (training sample)
Kolmogorov-Smirnov test: signal (background) probability = 0.291 (0.12)
(1/N) dN/dx
BDTG_fold2 response
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.0)%


TMVA overtraining check for classifier: BDTG_fold3
Signal (test sample)
Signal (training sample)
Background (test sample)
Background (training sample)
Kolmogorov-Smirnov test: signal (background) probability = 0.42 (0.003)
(1/N) dN/dx
BDTG_fold3 response
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.0)%


TMVA overtraining check for classifier: BDTG_fold4
Signal (test sample)
Signal (training sample)
Background (test sample)
Background (training sample)
Kolmogorov-Smirnov test: signal (background) probability = 0.067 (0.339)
(1/N) dN/dx
BDTG_fold4 response
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.0)%

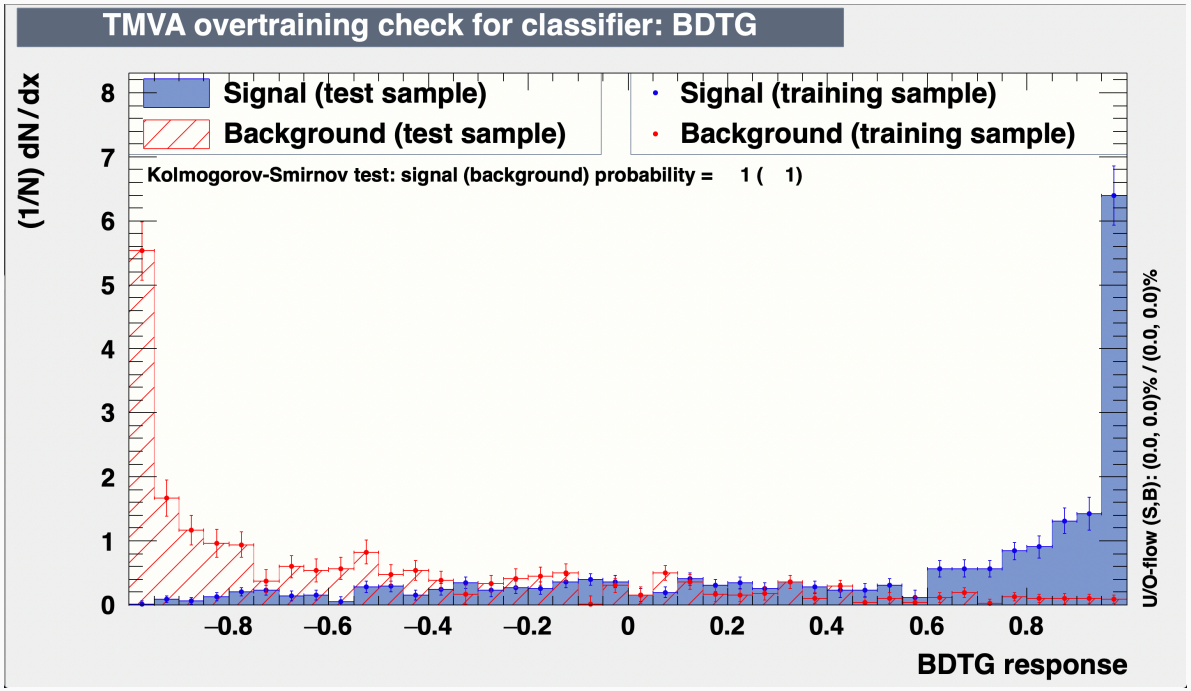

TMVA overtraining check for classifier: BDTG_fold5
Signal (test sample)
Signal (training sample)
Background (test sample)
Background (training sample)
Kolmogorov-Smirnov test: signal (background) probability = 0.027 (0.01)
(1/N) dN/dx
BDTG_fold5 response
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.0)%

Applying the MVA

- Once the Gradient BDT is trained it is included into the analysis software framework (tHqLoop) $\rightarrow$ The result of the training is stored in the weights.xml file.
- tHqLoop executed to produce processed samples with two new variables:
  - $\mathrm{LepAssign\_PredictedBDT\_val}$ :: Returns the MVA response for given event :: Between -1 (leading lepton perdicted to be from Higgs) and 1 (leading lepton predicted to be from top).
  - $\mathrm{LepAssign\_PredictedBDT\_err}$ :: Returns the uncertainty on MVA response for given event (not working, see root forum thread).
- Using $\mathrm{LepAssign\_PredictedBDT\_val}$ to assign the lepton origin in tHqLoop_LepHadReco.cxx.
- Already included in Merge Request 127.
- Next step: Select the "not-a-cut" point in the BDT score to optimise the lepton-association efficiency

Application - Optimal threshold search

The optimal point in the BDT to perform the classification is found with a linear search. Best: around 0.0



(entries %) weight %

```
Cut = -0.45     => Efficiency = (83.0%) 93.04%
Cut = -0.4      => Efficiency = (84.0%) 93.39%
Cut = -0.35     => Efficiency = (84.0%) 93.77%
Cut = -0.33     => Efficiency = (84.0%) 93.92%
Cut = -0.32     => Efficiency = (84.0%) 93.9%
Cut = -0.31     => Efficiency = (84.0%) 93.78%
Cut = -0.3      => Efficiency = (84.0%) 94.02%
Cut = -0.29     => Efficiency = (84.0%) 94.31%
Cut = -0.28     => Efficiency = (84.0%) 94.31%
Cut = -0.27     => Efficiency = (84.0%) 94.29%
Cut = -0.25     => Efficiency = (84.0%) 94.59%
Cut = -0.2      => Efficiency = (84.0%) 94.93%
Cut = -0.15     => Efficiency = (84.0%) 94.95%
Cut = -0.1      => Efficiency = (84.0%) 95.18%
Cut = -0.05     => Efficiency = (85.0%) 95.28%
Cut = 0.0       => Efficiency = (85.0%) 95.14%
Cut = 0.05      => Efficiency = (85.0%) 95.31%
Cut = 0.1       => Efficiency = (85.0%) 95.11%
Cut = 0.15      => Efficiency = (85.0%) 94.6%
Cut = 0.2       => Efficiency = (85.0%) 94.26%
Cut = 0.25      => Efficiency = (85.0%) 94.59%
Cut = 0.3       => Efficiency = (85.0%) 94.28%
Cut = 0.35      => Efficiency = (85.0%) 94.04%
Cut = 0.4       => Efficiency = (85.0%) 93.69%
Cut = 0.45      => Efficiency = (85.0%) 93.71%
```

Note that there is not signal and background but Type1 and Type2 -> We do not reject events but classify them

- In most SS of events the leading lepton comes from the Higgs boson system.
- Tool assign the origin with $\geq$ 95% precision (the current baseline method had 84.49%)
- Already implemented in software framework.

Back up

Assignment - Labeling with truth - More numbers

- $2\ell + \tau_{had}$ entries: 37542 (15905 SS and 21637 OS)
- From the 15905 2ℓ SS $+ 1\tau_{had}$:
  - τ_{had} from Higgs: 15057 (94.7% of total SS)
  - Successful labeling: 14176 (89.1% of total SS)
    - Labeled as type1: 8666 (61.1% of successfully labeled events)
    - Labeled as type2: 5510 (38.9% of successfully labeled events)
- From the 21637 2ℓ OS $+ 1\tau_{had}$:
  - τ_{had} from Higgs: 15324 (70.8% of total OS)
  - Successful labeling: 14282 (66.0% of total OS)
- Since for OS the true origin is known, it is possible to test the ability of the labeling method

Lepton Origin assignment - Prediction with MVA - Other considerations

- No GPU implementation for Gradient BDT in ROOT:TMVA
- Ideally, one should use samples with the fakes corrected before training the BDT. Nevertheless, the fakes SF affect mostly the bkg dominated by fakes and should not have a relevant impact on the signal samples. Therefore, the re-training of the BDT with the new SF is not so crucial.
- From 37542 tHq events total. From the ones that match and :
  - 18859 is_Lep1_from_Top = 1 (SS 9320)
  - 18683 is_Lep1_from_Top = 0 (SS 6584)

- Have to explore correlations
- Take into account the charge flip
- Tau SF not used
- With the old baseline method
  - Total efficiency: 84.49 %
  - Type1 efficiency: 87.82 %
  - Type2 efficiency: 79.59 %

Lepton Origin assignment - Other methods

In order to perform this association for the $2\ell\,\mathrm{SS} + 1\tau_{had}$ several methods relying in the truth-level information have been explored.:

- First method (Cyrus): Assume that the leading lepton was originated from the top
- Second method (Mathias): Cut in two variables
  - $m_{vis,H}(lep(t)) - m_{vis,H}(lep(H)) > 57.0, ,\mathrm{GeV}$
  - $m_{pred,t}(lep(H)) - m_{pred,t}(lep(t)) > 0.0\,\mathrm{GeV}$
- BDT based method presented in this work

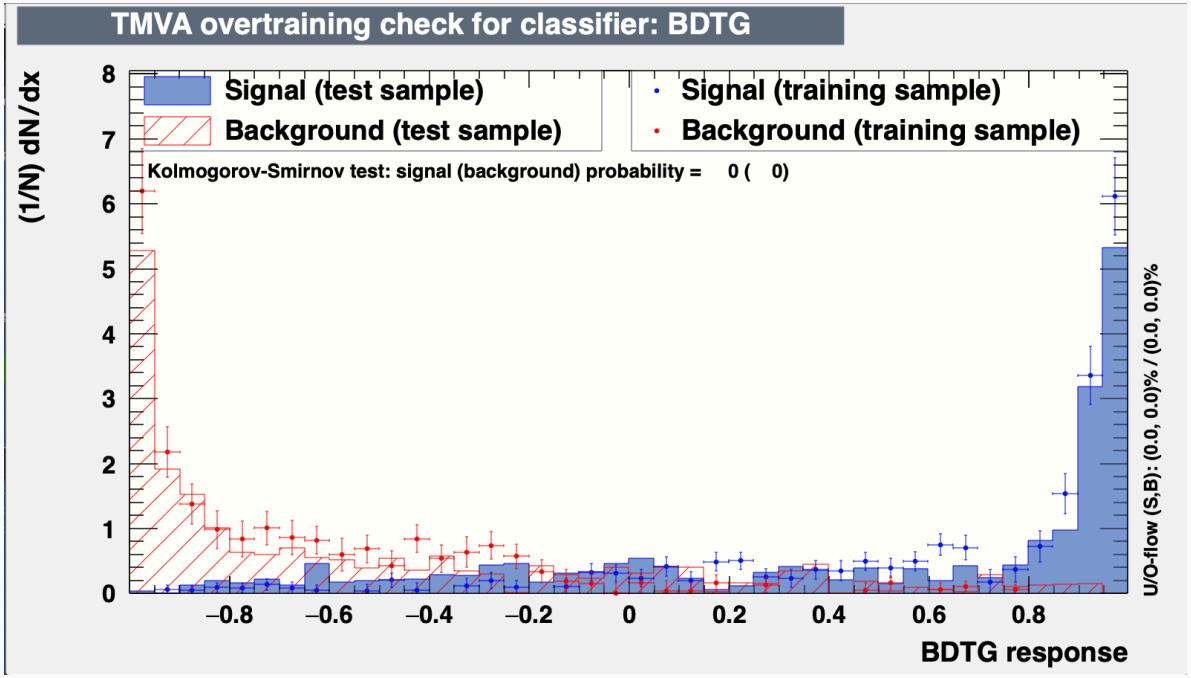
Lepton Origin assignment - Prediction with MVA - Other considerations

The variables $\Delta R(lepN, \tau)$ and $\Delta\eta(lepN, \tau)$ can be related

- deltaEta_tau_LightLep1 vs deltaR_tau_LightLep1
- deltaEta_tau_LightLep2 vs deltaR_tau_LightLep2

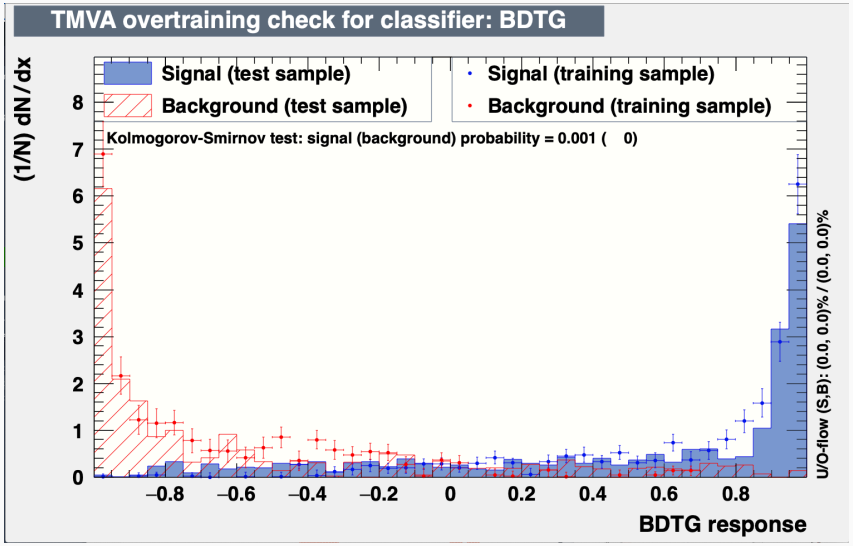
If the $\Delta\eta$ variables are removed from the training:

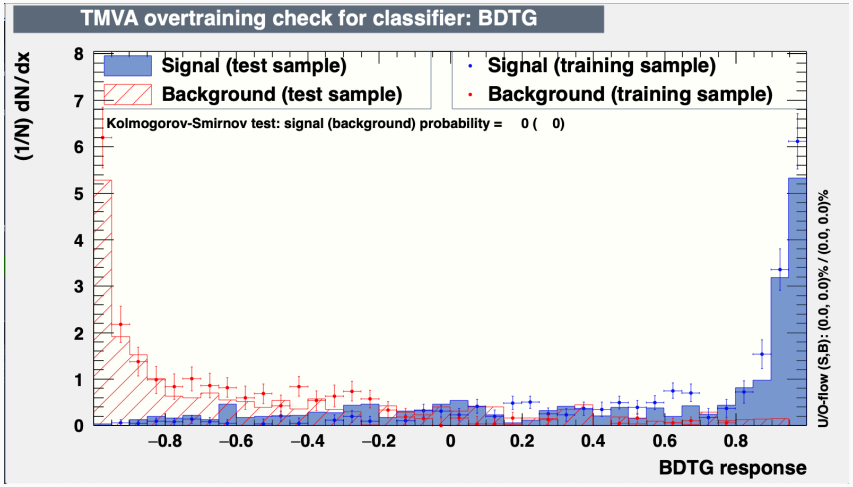
- The integrated ROC decreases from 0.911 to 0.904
- The BDT response distribution still looks good
- For further tests $\rightarrow$ add the BDT without the $\Delta\eta$ vars in tHqLoop




```
nk : Variable
-----------------------------------
 1 : DeltaRLeadingLeptonClosestBjet
 2 : eta_Tpred_opt2
 3 : deltaR_tau_LightLep1
 4 : deltaR_b_LightLep2
 5 : deltaR_tau_LightLep2
 6 : deltaR_b_LightLep1
 7 : m_Hvis_opt1
 8 : m_Hvis_opt2
```

Lepton Origin assignment - Prediction with MVA - Other considerations



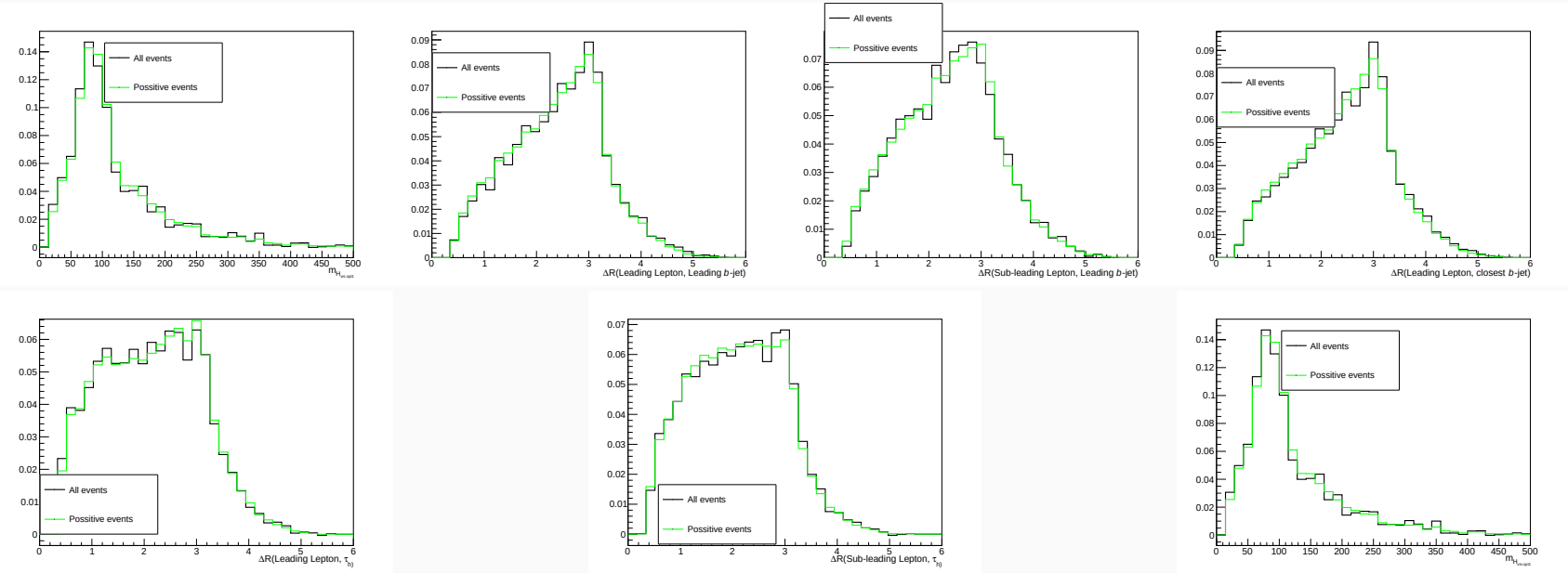


Above using deltaEta_tau_LightLep1 and deltaEta_tau_LightLep1, and below without these vars.

kfold

MVA - Negative weight treatment - (SS + OS)

We compare the shapes of the distributions with all the events to the ones with the positively weighted events only (the used for training).



Normalised. SS + OS events

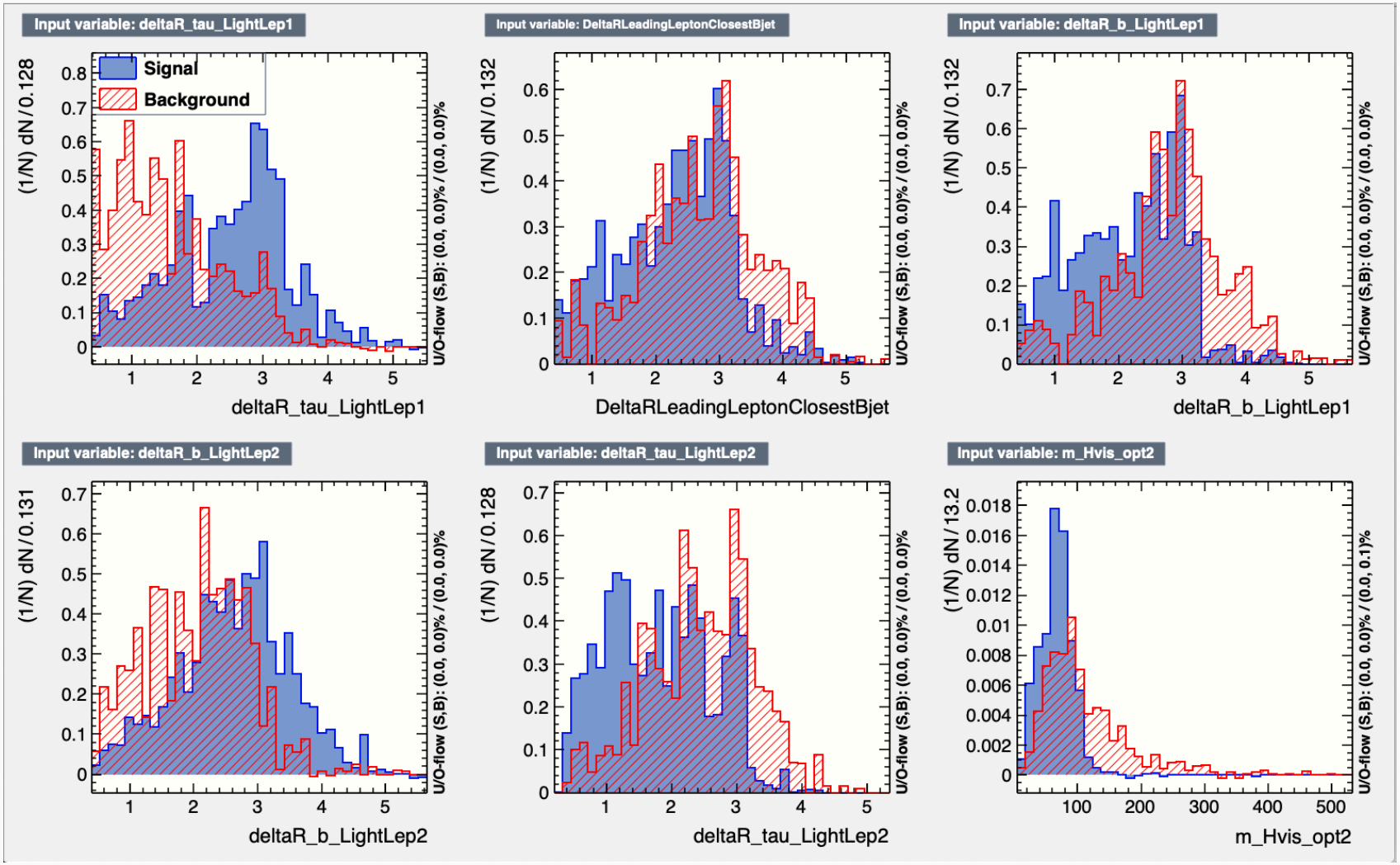
Using only the positive weights is not problematic.

Further test: Do Type1 and Type2 have the same fraction of negative weights? Yes, 36.4% for type 1 and 36.0% for type2. Removing neg weights does not bias towards type nor type2

MVA - Variables in training - Ranked List - Tested

- **DeltaRLeadingLeptonClosestBjet**: ΔR between the leading lepton and the closest b-tagged jet to that lepton.
- **deltaR_b_LightLep2**: ΔR between the b-tagged jet and the sub-leading light-flavoured lepton.
- **deltaR_tau_LightLep1**: ΔR between the hadronic tau and the leading light-flavoured lepton.
- **deltaR_b_LightLep1**: ΔR between the b-tagged jet and the leading light-flavoured lepton.
- **eta_Tpred_opt2**: η of the top quark when reconstructed as the sum of the b-jet, sub-leading leptons and the reconstructed neutrino.
- **deltaR_tau_LightLep2**: ΔR between the hadronic tau and the sub-leading light-flavoured lepton.
- **m_Hvis_opt1** : mass of the combined hadronic tau and the leading light-flavoured lepton
- **deltaEta_tau_LightLep1**: $\Delta\eta$ between the hadronic tau and the leading light-flavoured lepton.
- **deltaEta_tau_LightLep2**: $\Delta\eta$ between the hadronic tau and the sub-leading light-flavoured lepton.
- **m_Hvis_opt2** : mass of the combined hadronic tau and the sub-leading light-flavoured lepton

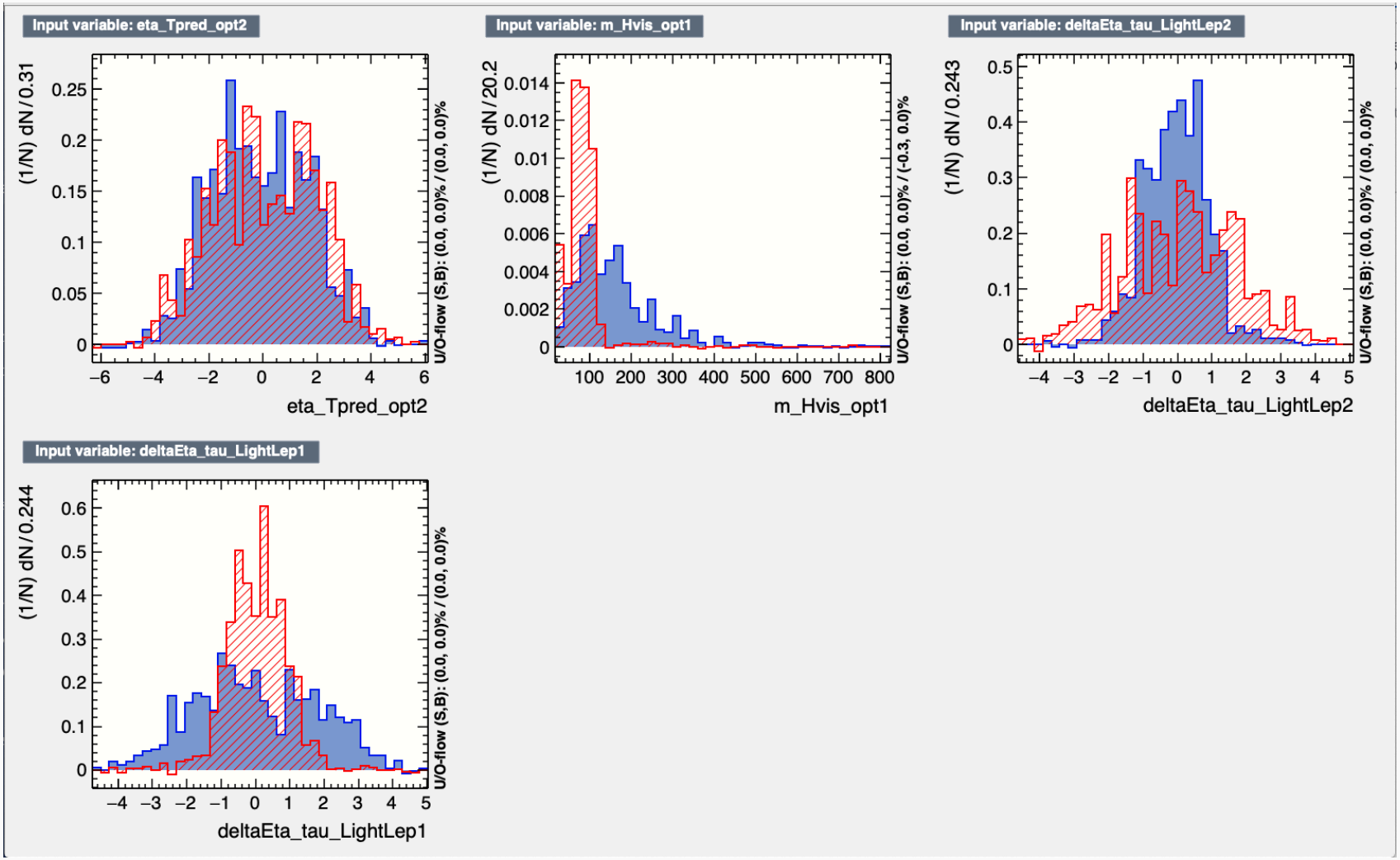
MVA - Variables tested for training - Distribution



This is not signal vs background but Type1 vs Type2

Only signal here

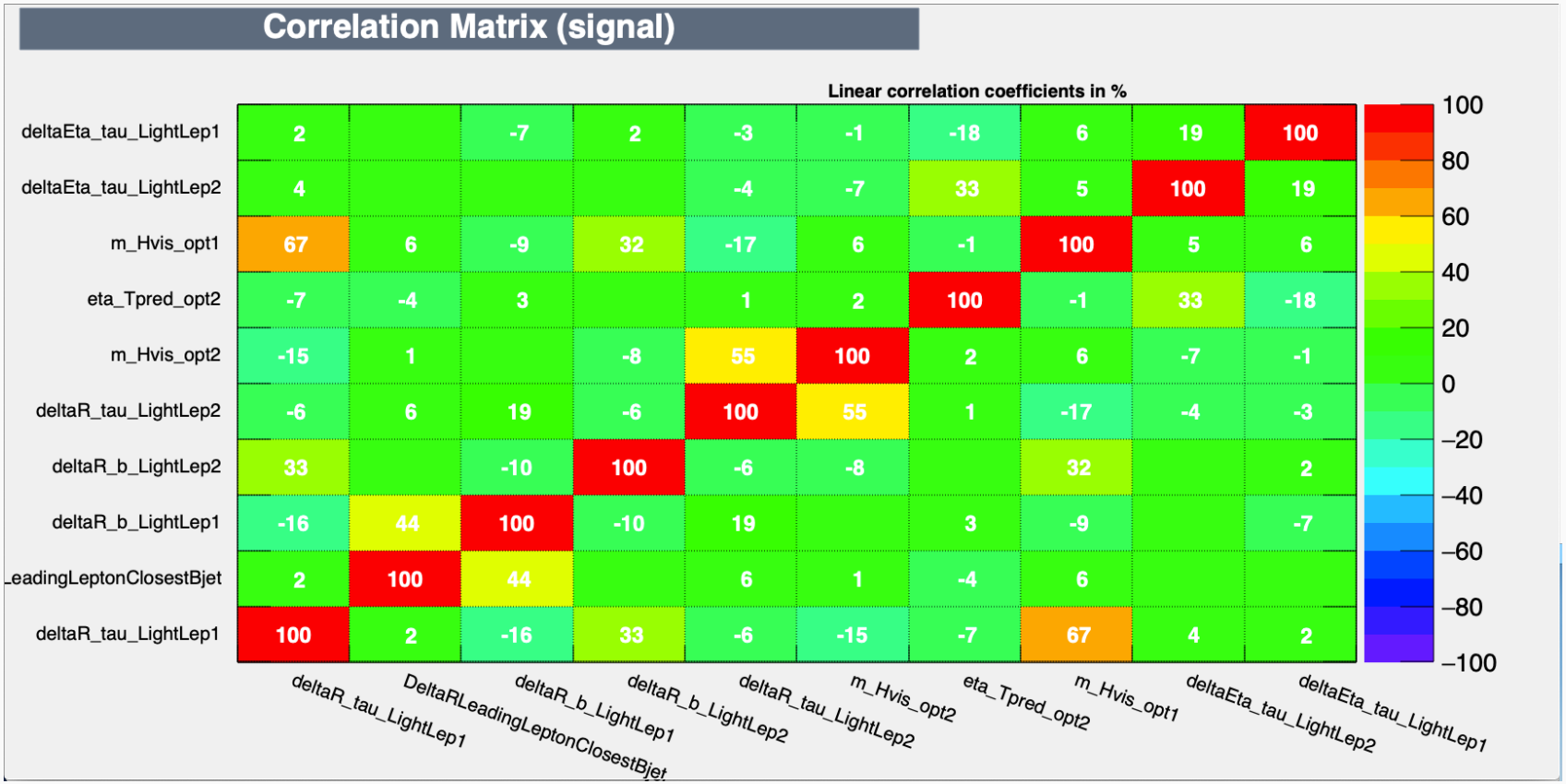
MVA - Variables in training - Distribution



This is not signal vs background but Type1 vs Type2

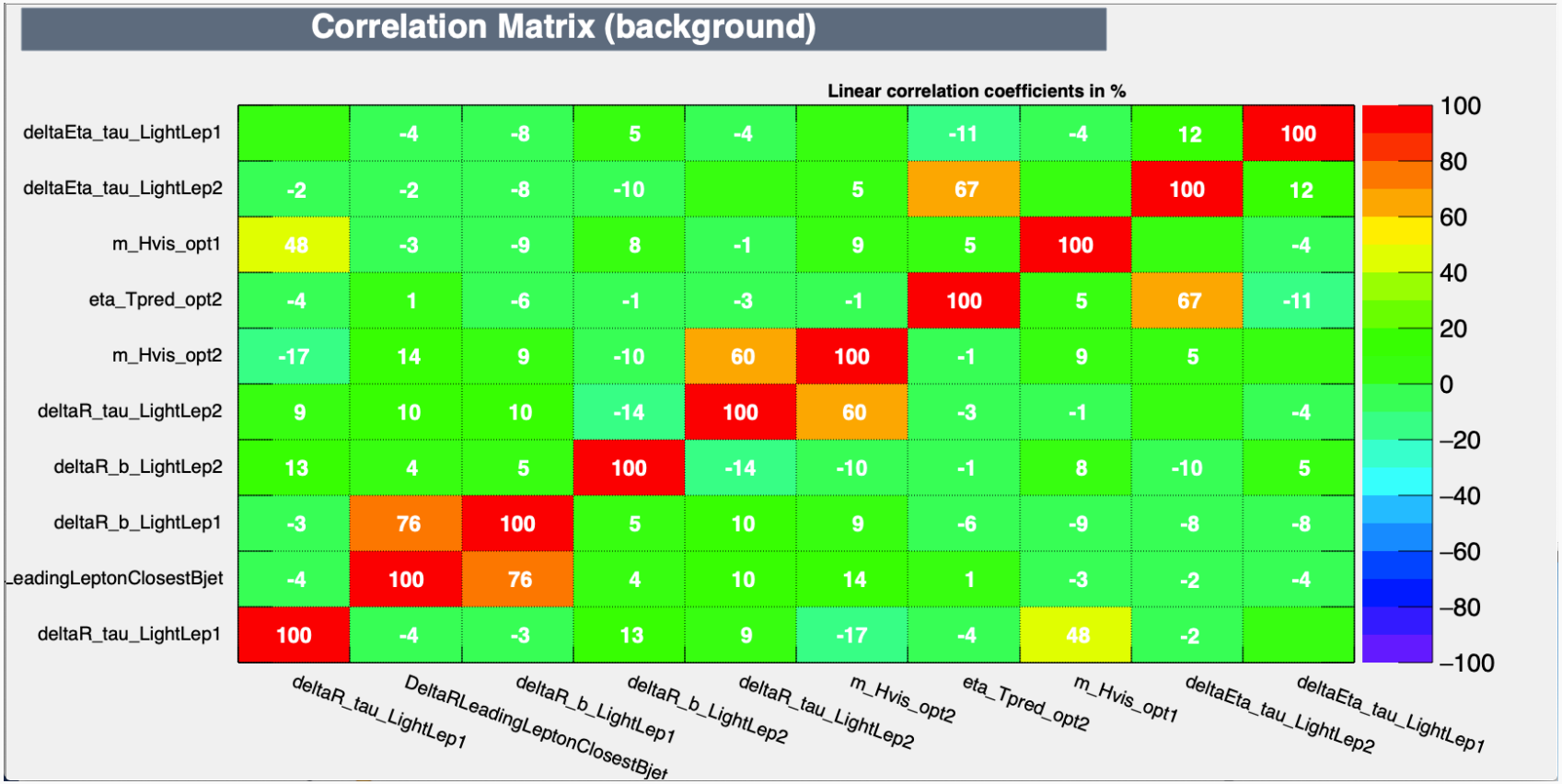
Only signal here

MVA - Variables tested for training - Correlations for Type 1 Events



Correlated variables have to be explored

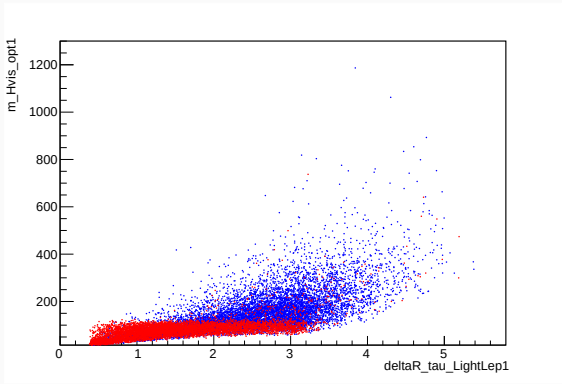
MVA - Variables tested for training - Correlations for Type 2 Events

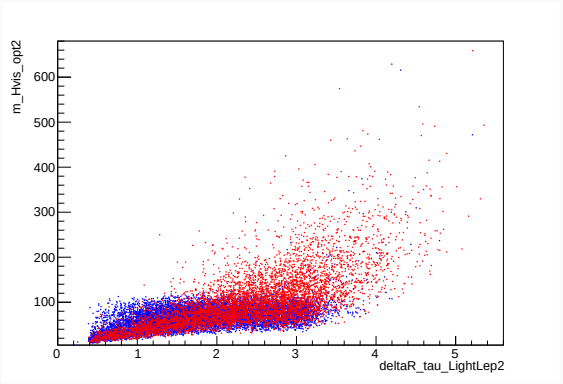


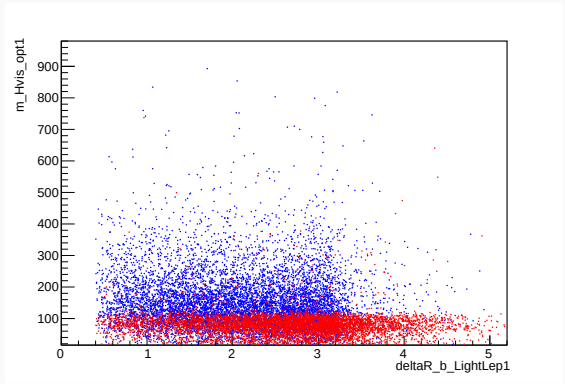
Correlated variables have to be explored

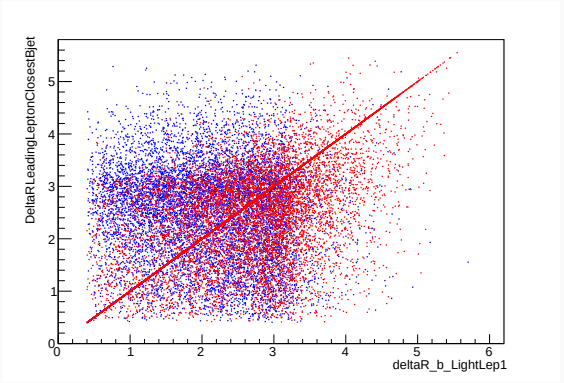
MVA - Variables tested for training - Correlations

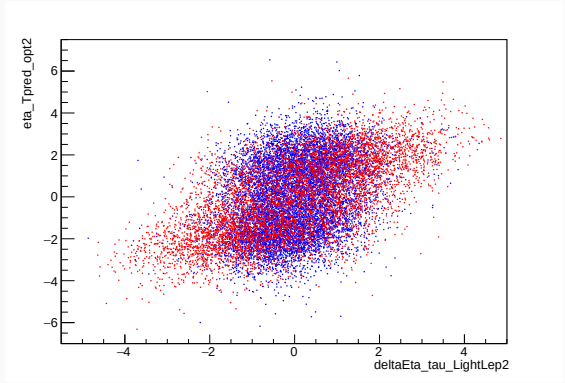
The most correlated pairs of variables (according to the two previous slides) are studied distinguishing between Type1 (blue) and Type2 (red) events:











Lepton Origin assignment - Variables used

- deltaR_b_LightLep1, deltaR_b_LightLep2: These variables are filled in tHqLoop_LepHadReco.cxx. In the case of several *b*-tagged jets they referrer to the one with highest p_T.
  - The function $\mathrm{FillJetPtVector}()$ of tHqLoopAlgDiscVar.cxx orders the vector $\mathrm{m\_jets}$ by p_T, from grater to lower.
  - tHqLoop_LepHadReco.cxx takes the *b*-tagged jets of $\mathrm{m\_jets}$ according to the *b*-tagging alg.
  - The internal variable $\mathrm{dileptau\_bjet\_Top}$ stores the *b*-tagged jet with highest p_T
- DeltaRLeadingLeptonClosestBjet: The closest *b*-tagged jet to the leading lepton. Stored internally in $\mathrm{m\_deltas\_leadinglepton\_closestbjet}$ and defined in $\mathrm{LeadingLeptonClosestBjet}()$ function of tHqLoopAlgDiscVar.cxx

Lepton Origin assignment - Variables used

- eta_Tpred_opt2 = η of the top quark when reconstructed as the sum of the *b*-jet, sub-leading leptons and the reconstructed neutrino.
  - dileptau_eta_Tpred_opt2 = (dileptau_bjet_Top + dileptau_lep_2 + ss_reco_nu_t_lep_2).Eta();
  - ss_reco_nu_t_lep_2.SetPtEtaPhiM(ss_reco_nu_Pt_lep_2, dileptau_lep_2.Eta(), ss_reco_nu_TVector2_lep_2.Phi(), 0.);
  - ss_reco_nu_Pt_lep_2 derived from ss_reco_nu_TVector2_lep_2 = Predict_nu_Top(dileptau_bjet_Top, dileptau_lep_2, conventional);
  - Details of the $\mathrm{Predict\_nu\_Top}$ in the tHqLoop_LepHadReco.cxx
  - $\mathrm{Predict\_nu\_Top}$: The nu_Top is defined using the already existing assignment therefore it should be removed
- mass_Hvis_option_1 = (dileptau_lep_1 + dileptau_tau).M();
- mass_Hvis_option_2 = (dileptau_lep_2 + dileptau_tau).M();

Lepton Origin assignment - Other variables explored

Apart from the already mentioned, other variables have been explored but not used in the final set of features

- pt_Tpred_opt1
- m_Tpred_opt1
- pt_Hvis_opt1
- m_Tpred_opt2
- pt_Tpred_opt2
- DeltaEtaLeadingLeptonClosestBjet
- eta_Tpred_opt1
- eta_Hvis_opt1
- pt_lep1

The highlighted vars correspond to those used in the current baseline method.

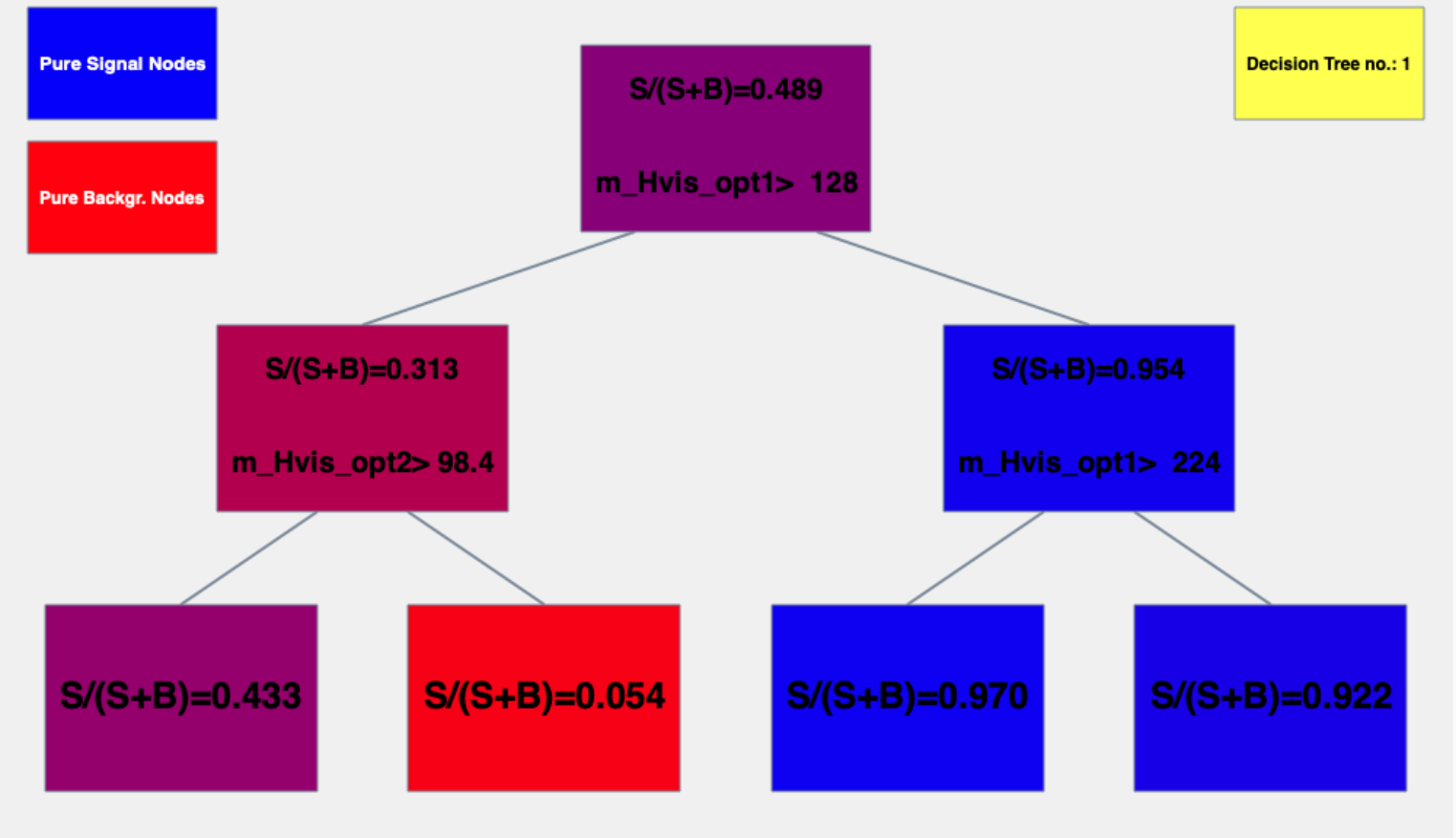
Lepton Origin assignment - Systematics associated to the BDT

How to evaluate the systematic uncertainties associated to an MVA method?

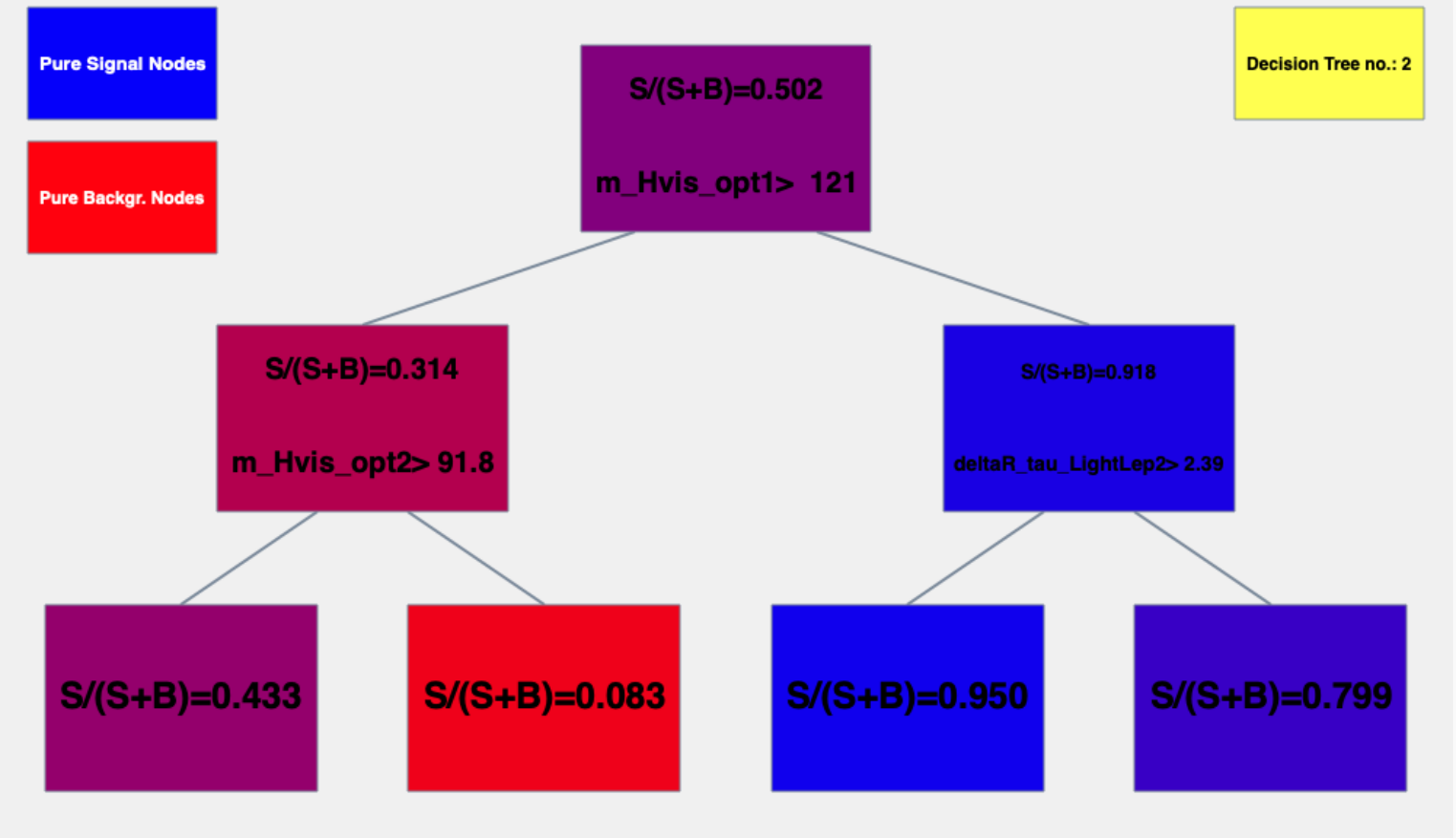
1. Find a CR[3] to compare data to MC.
2. If there is a good agreement, ATLAS does not require to provide any BDT systematic uncertainty. (In this channel the main uncertainties are those of tau fakes, not the BDT)
3. If there is not a good agreement we need to evaluate the systematics.
4. To evaluate them, a variable is removed from the list of features and a new BDT is trained (BDT_{syst}). This new BDT_{syst} is applied to the tHqLoop output.
5. The difference in the yields and distributions when using BDT_{syst} or the original BDT give us the systematic uncertainty.

[3] A region in which we know that the BDT performs correctly

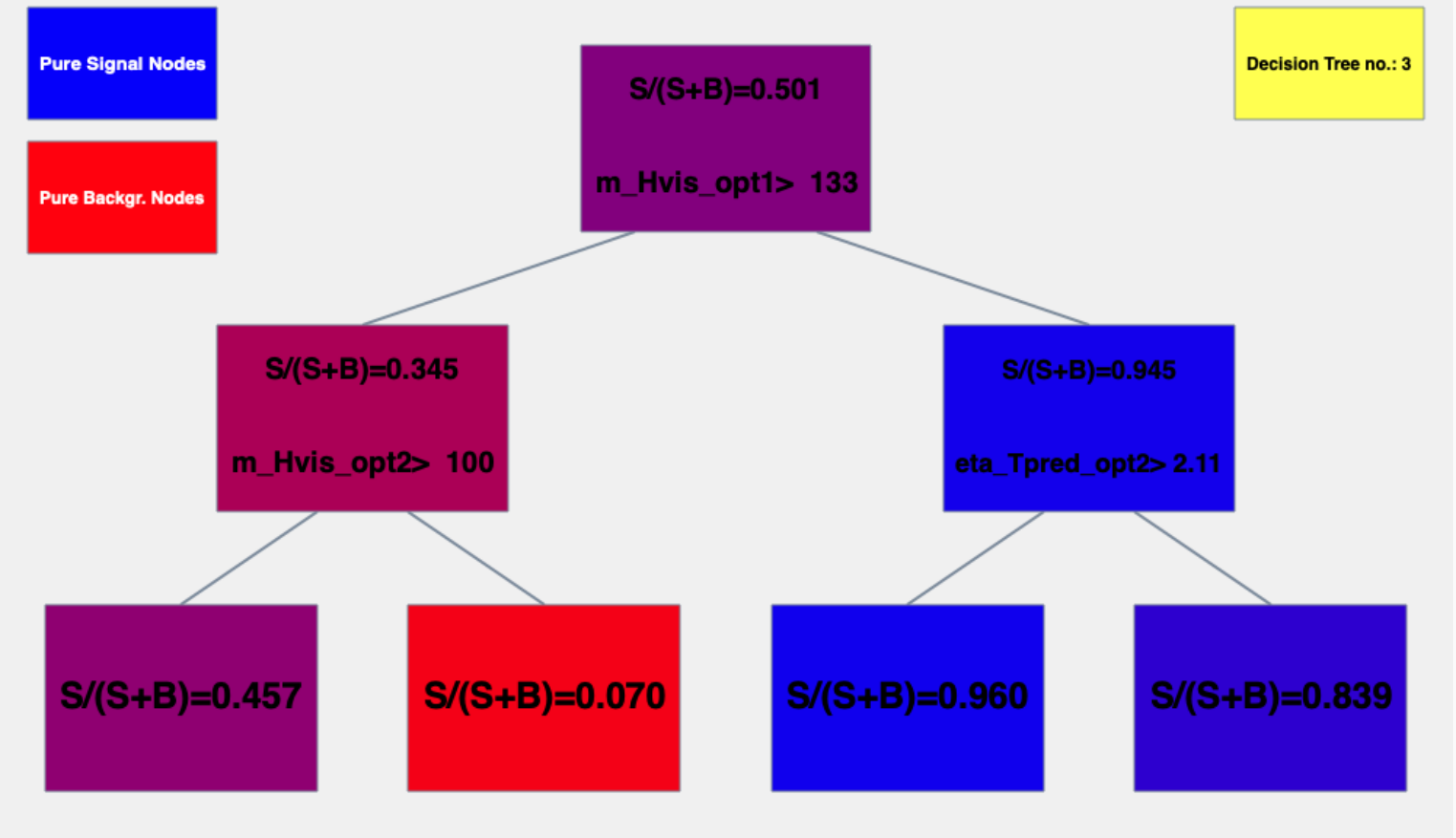
Lepton Origin assignment - BDTs



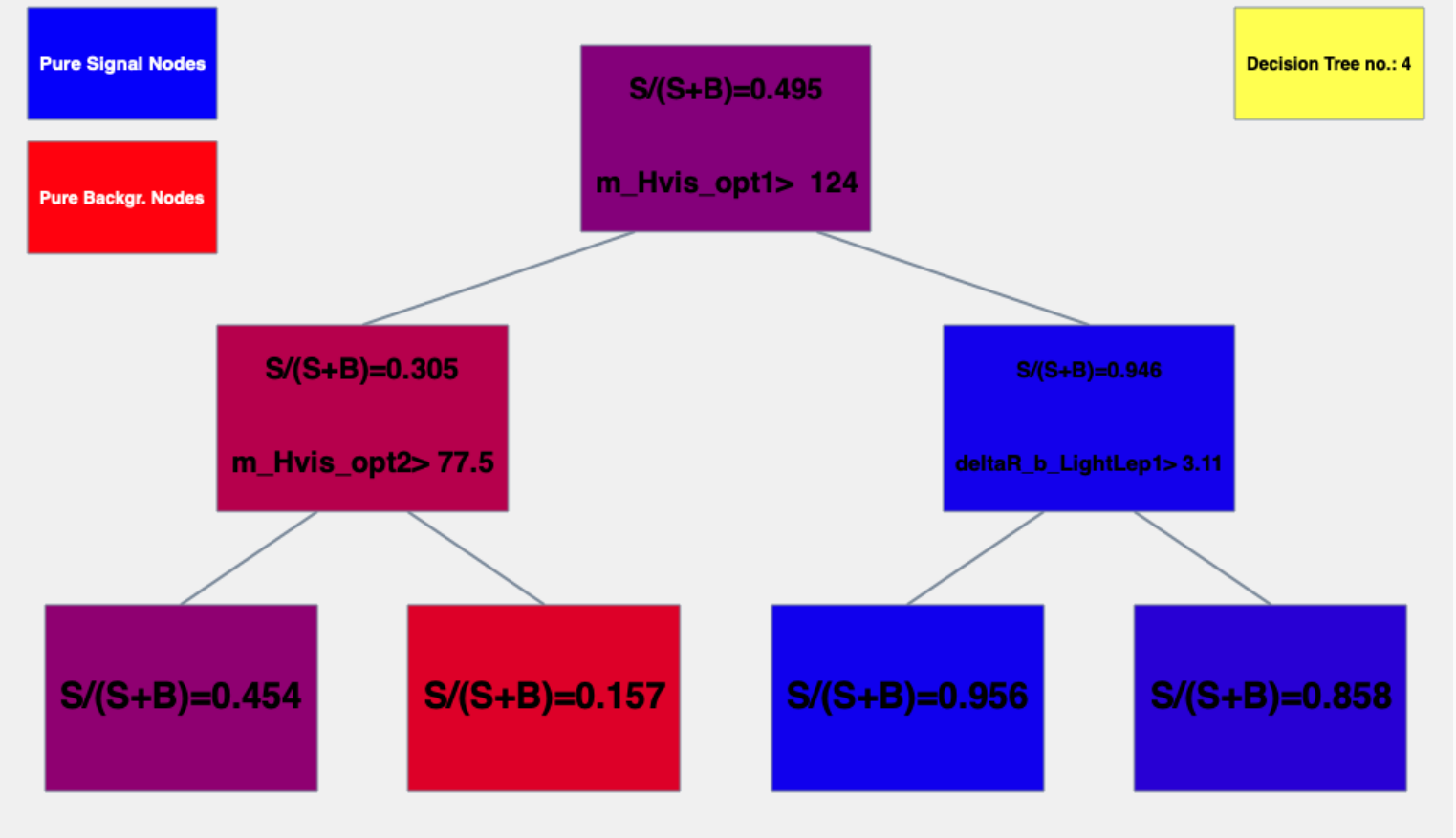
Lepton Origin assignment - BDTs



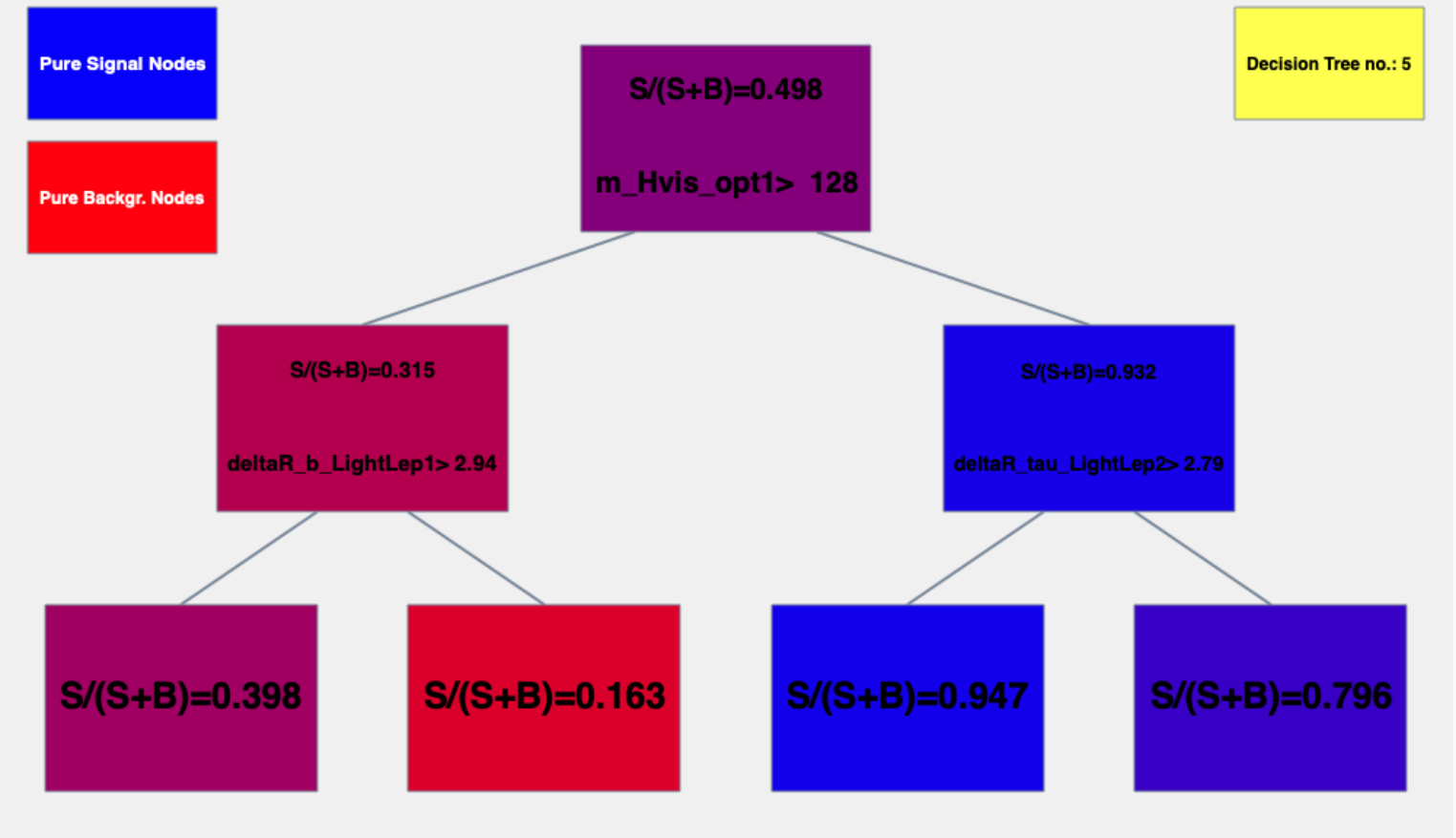
Lepton Origin assignment - BDTs



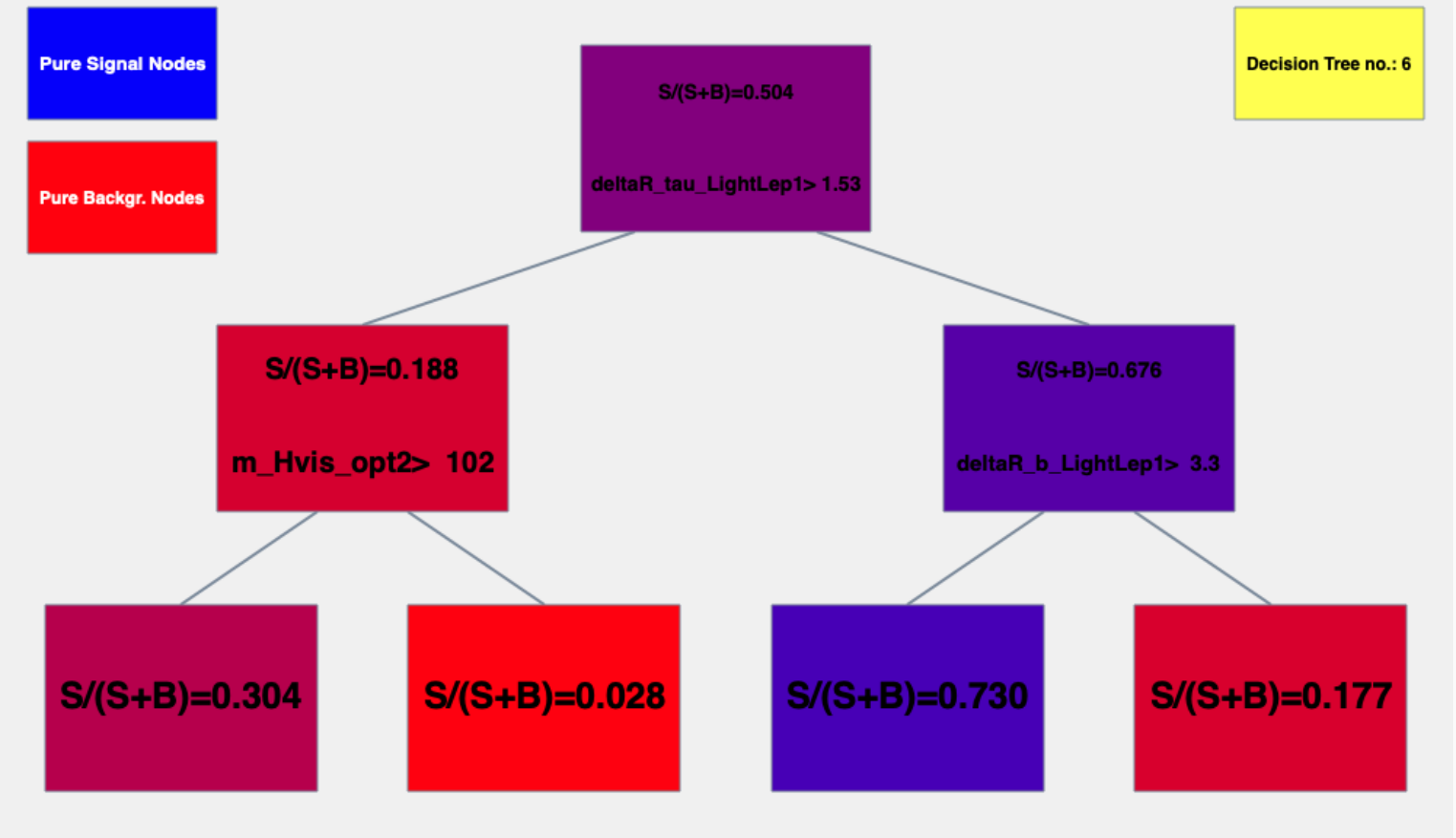
Lepton Origin assignment - BDTs



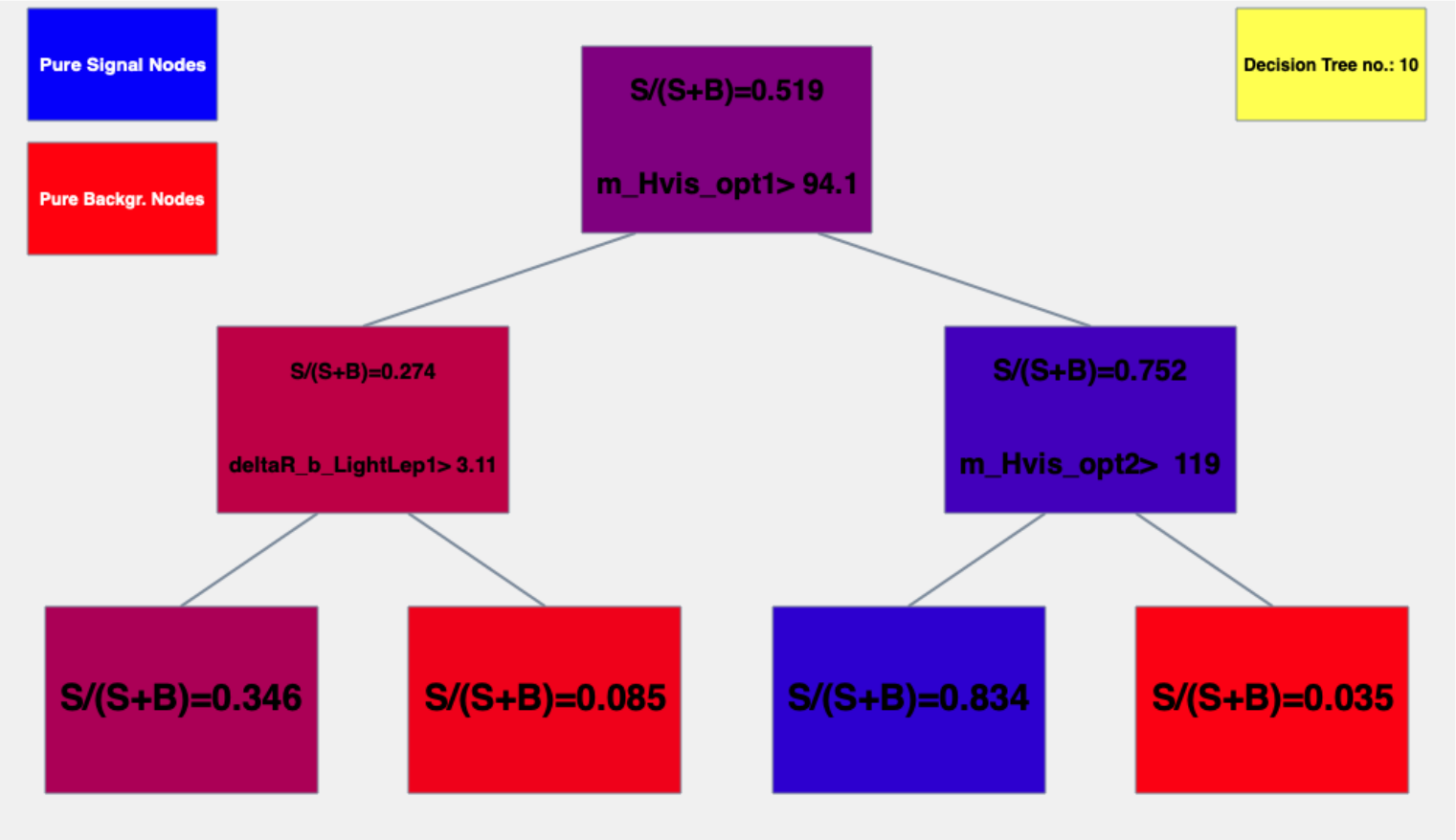
Lepton Origin assignment - BDTs



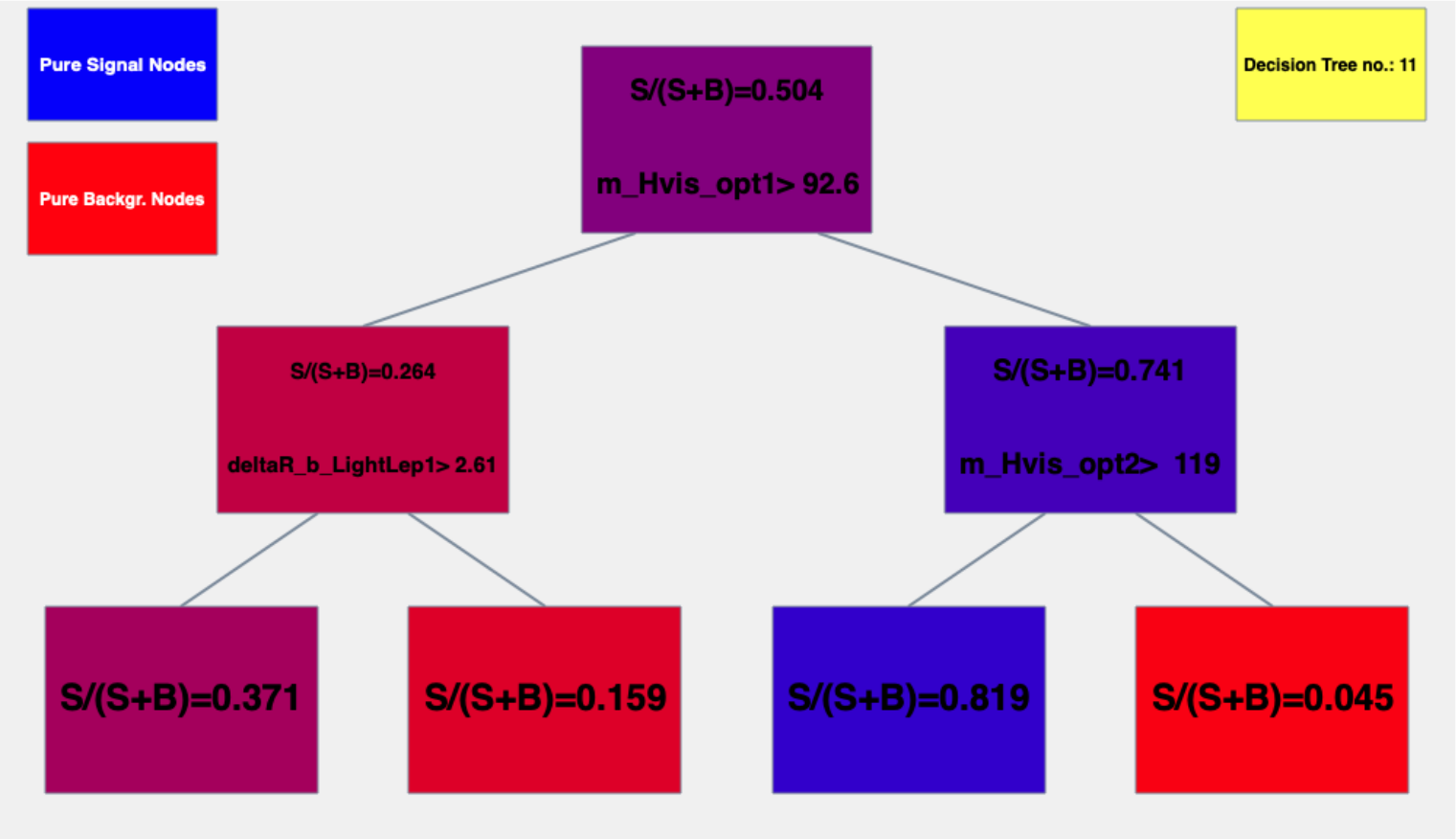
Lepton Origin assignment - BDTs



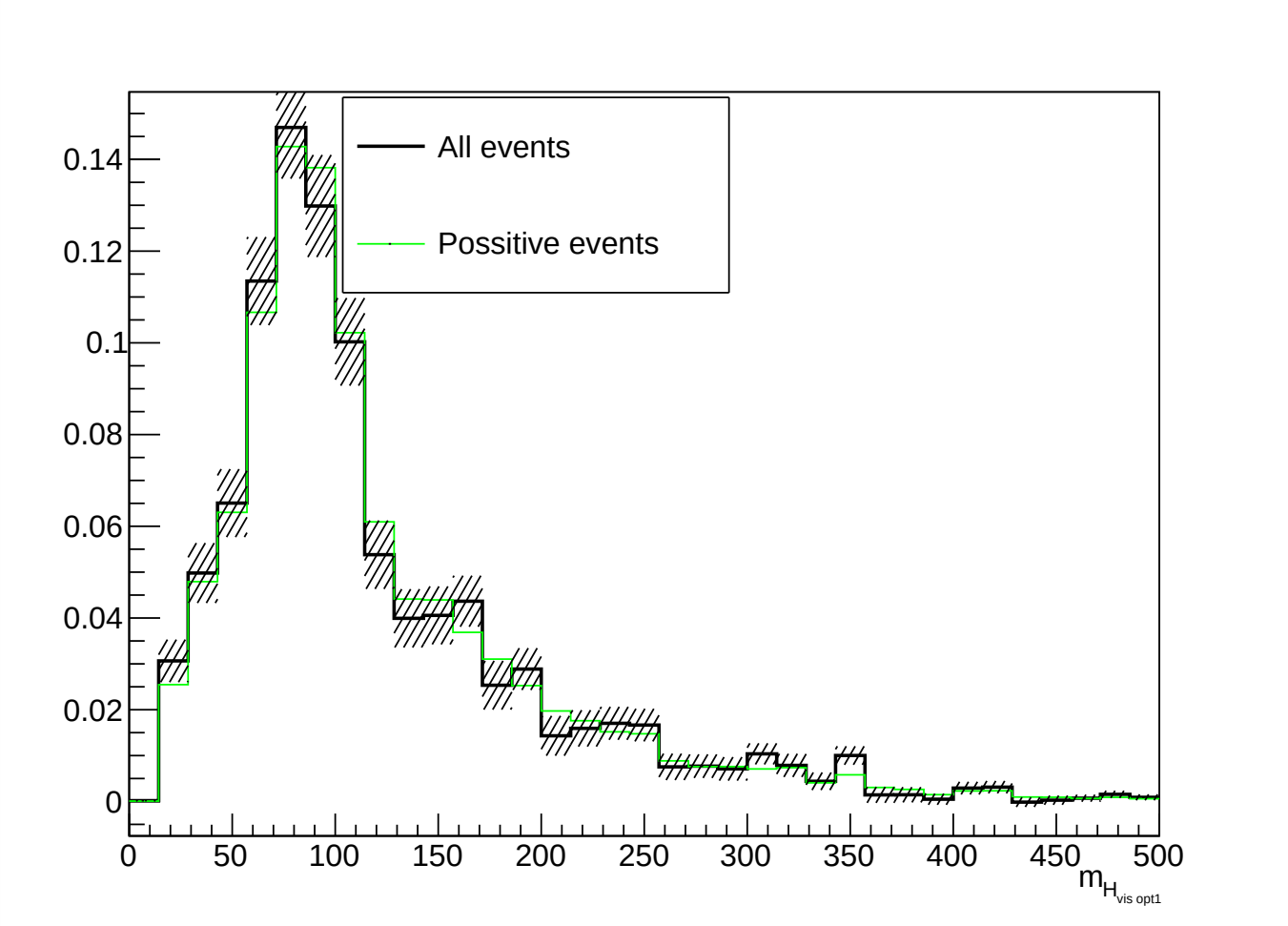
Lepton Origin assignment - BDTs

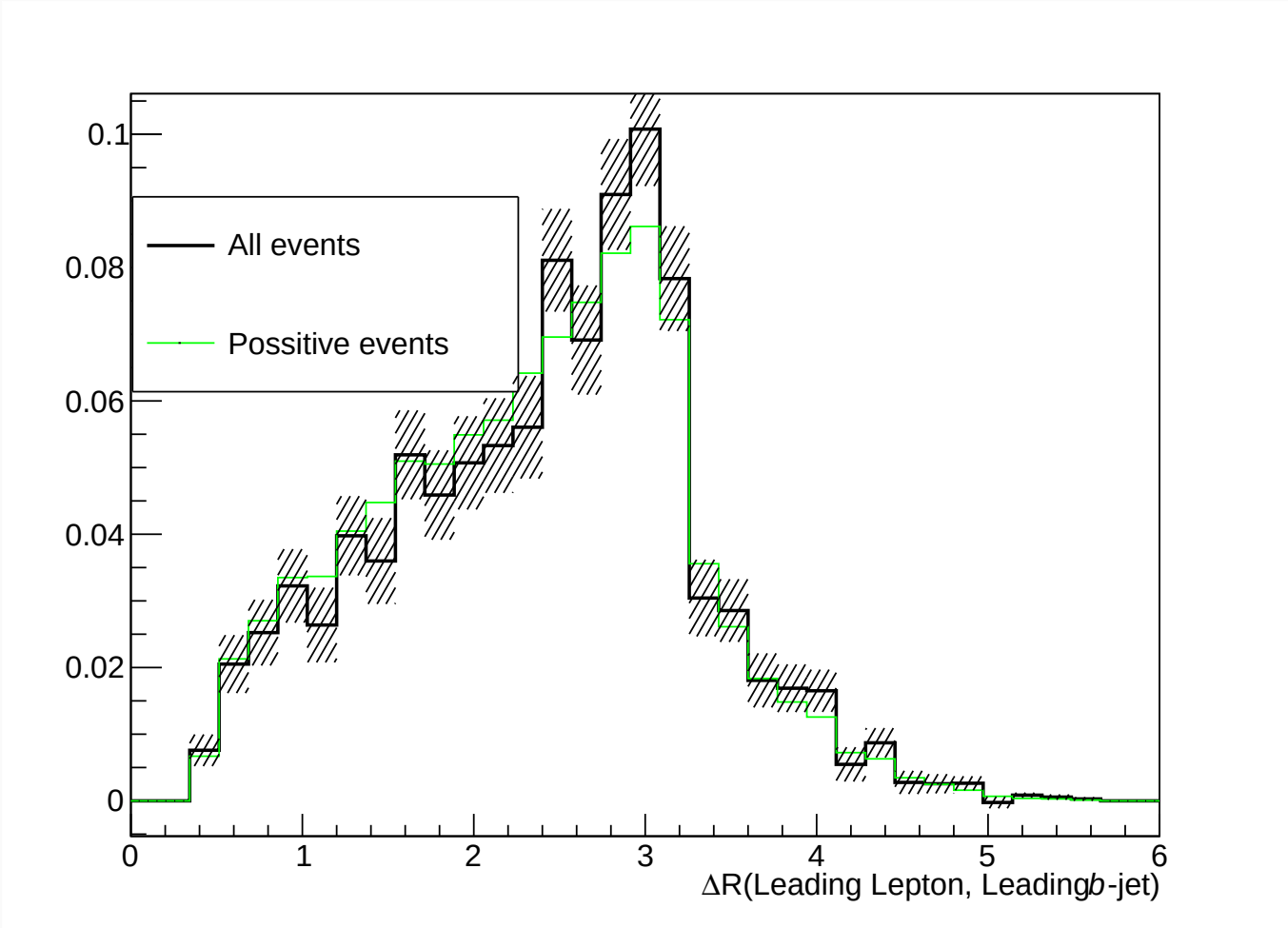


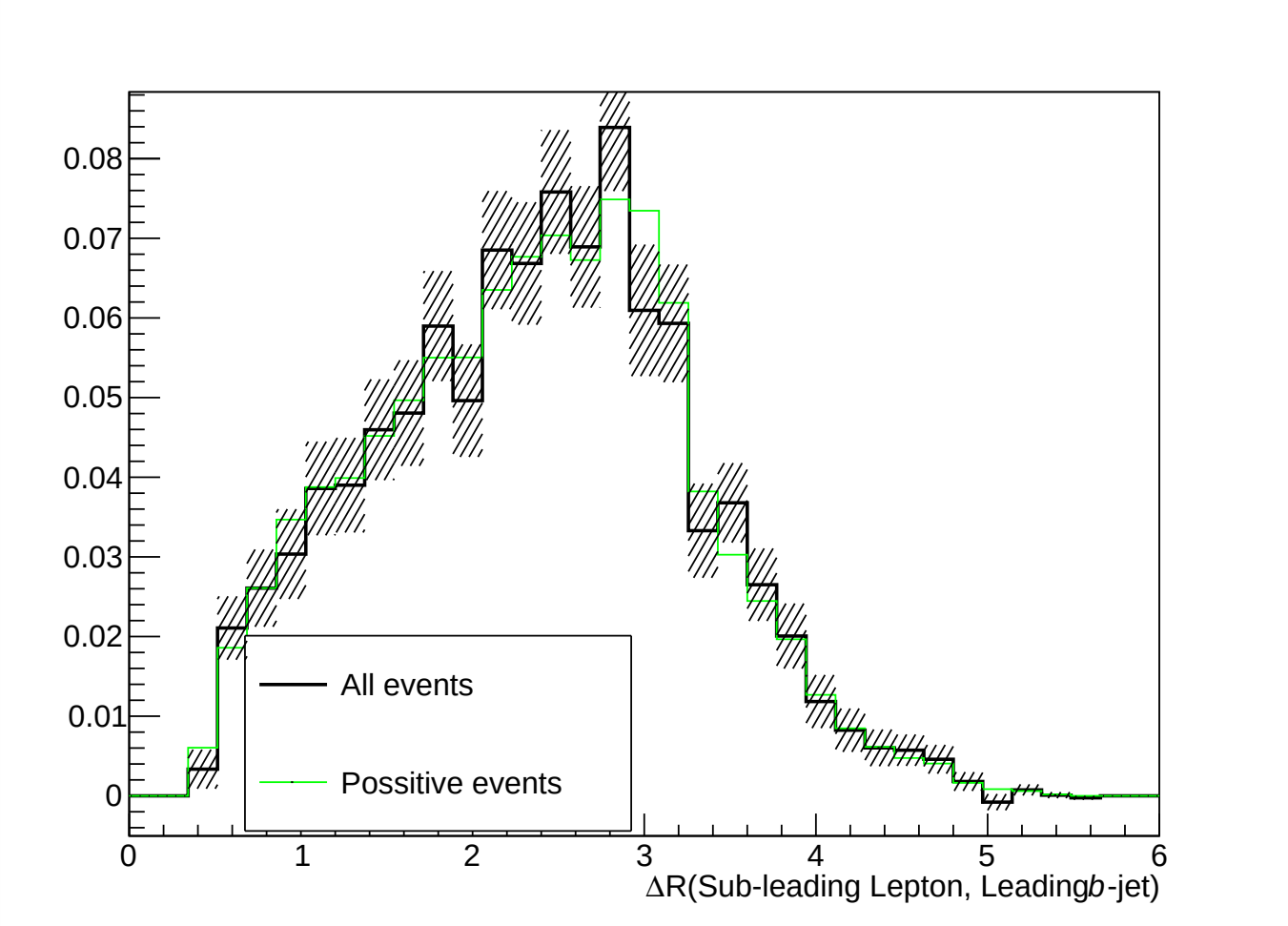
Lepton Origin assignment - BDTs

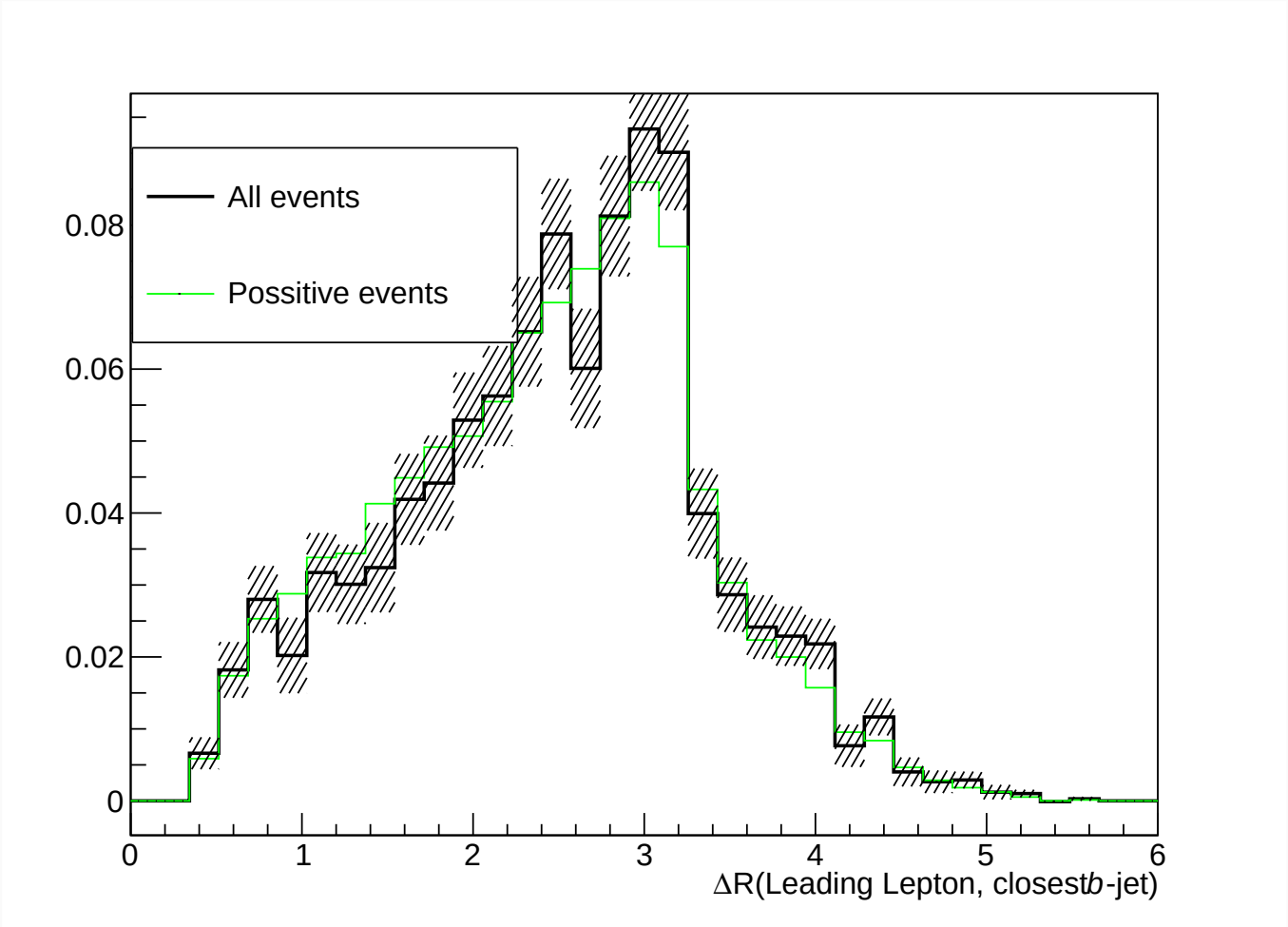


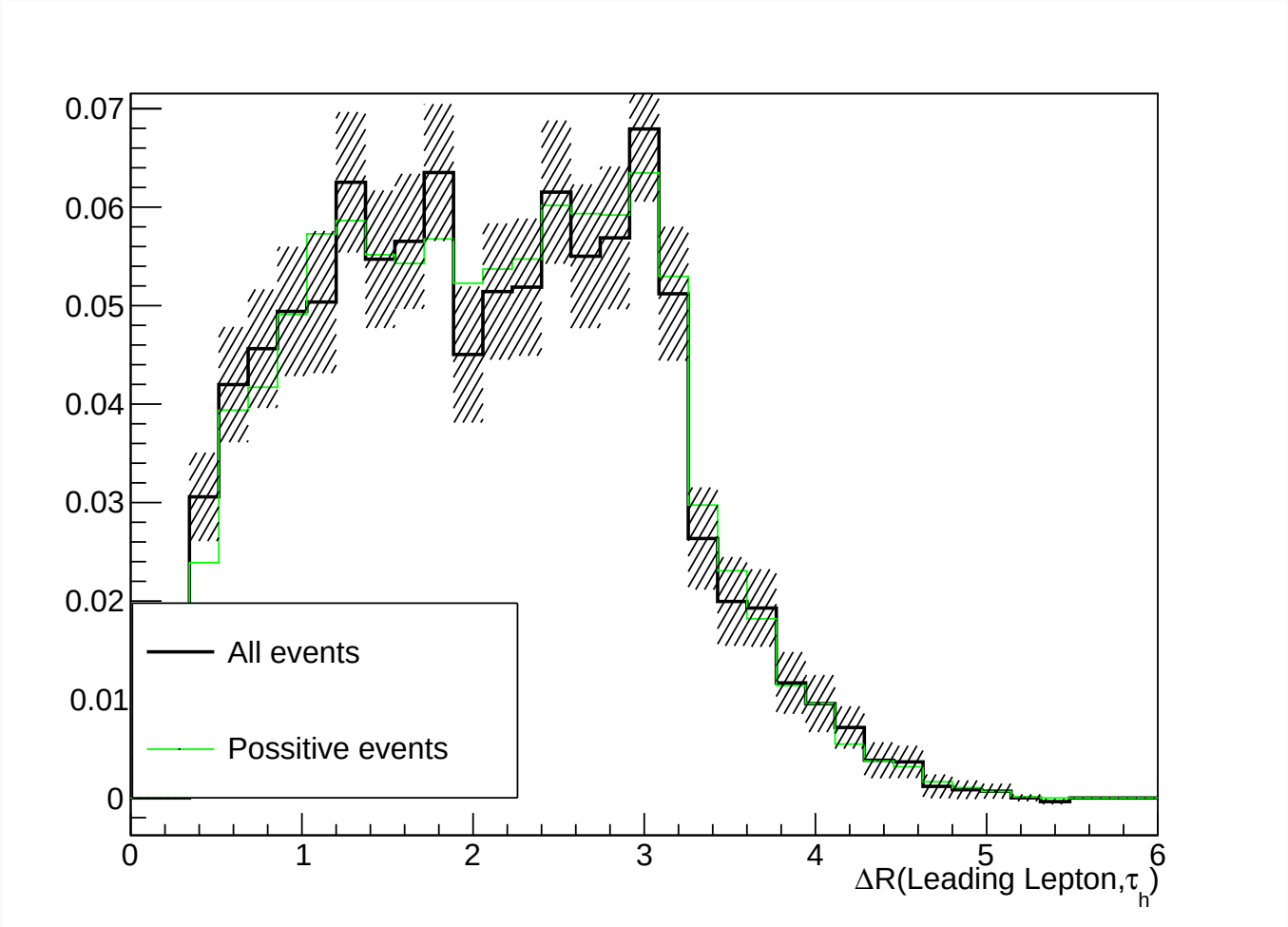
The gradient BDT gives a prediction model in the form of an ensemble of weak prediction models, which are decision trees

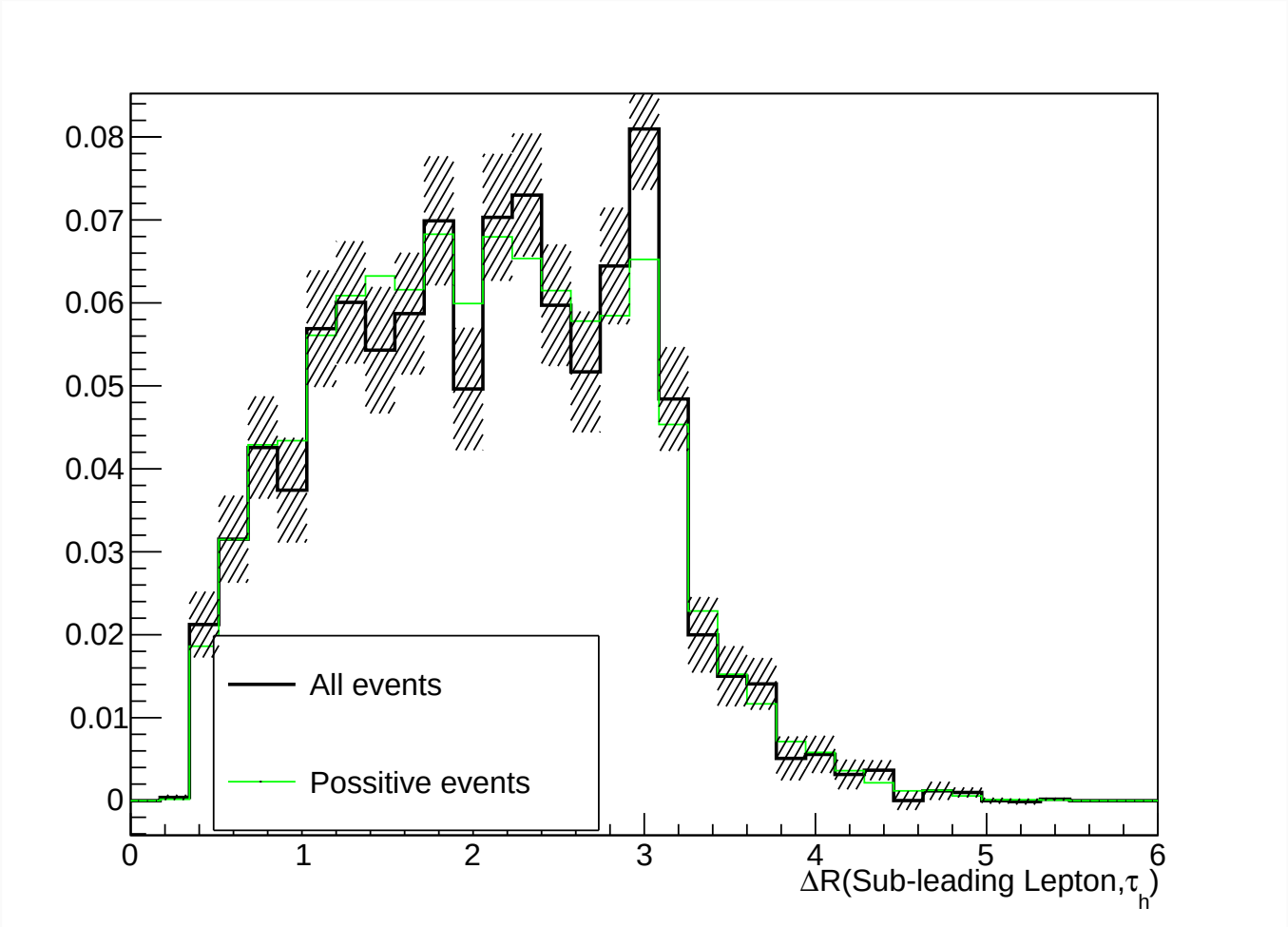

All events
Possitive events
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
50
100
150
200
250
300
350
400
450
500
$m_{H_{vis\ opt1}}$

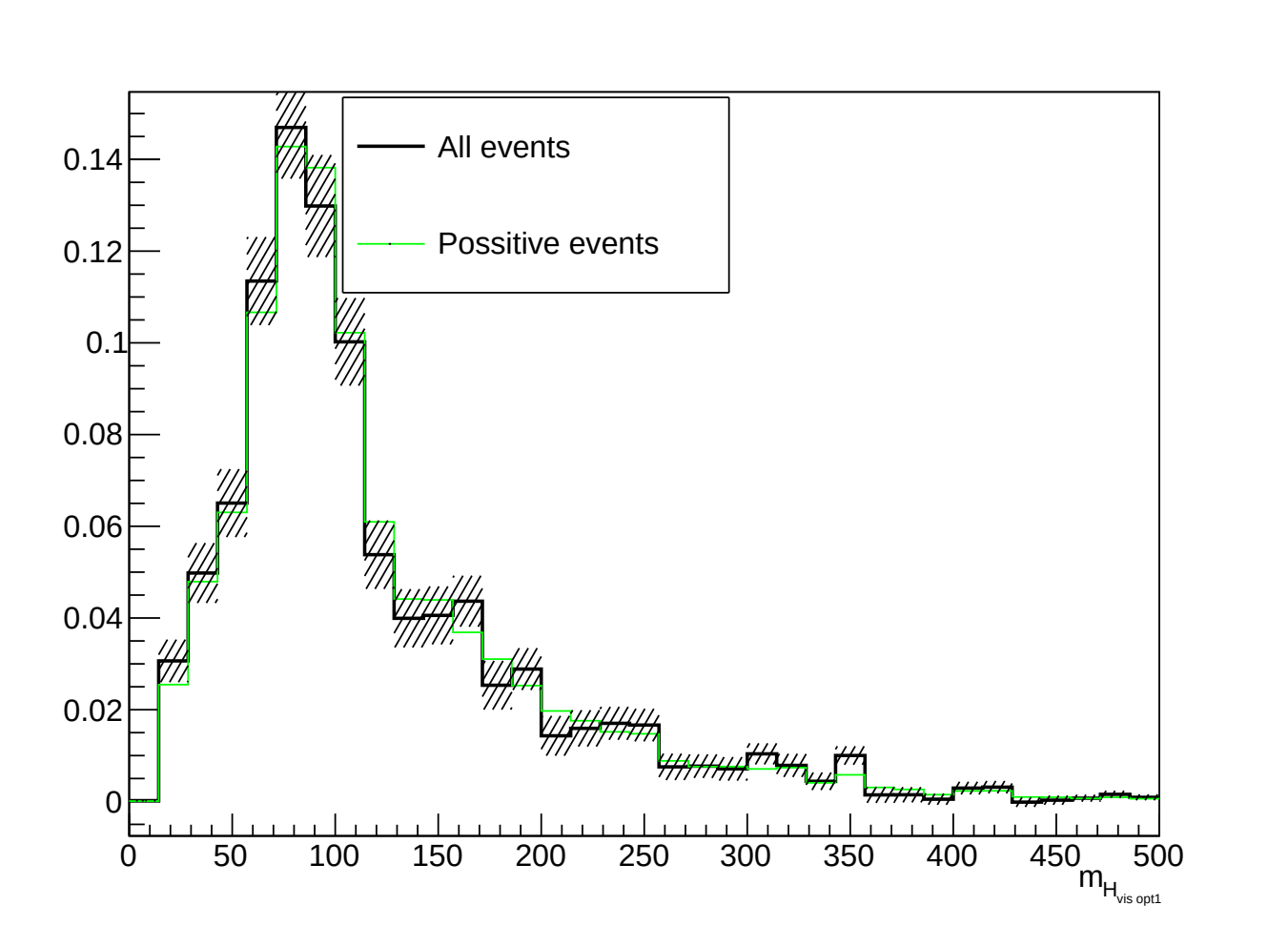

All events
Possitive events
0.1
0.08
0.06
0.04
0.02
0
0
1
2
3
4
5
6
ΔR(Leading Lepton, Leading b-jet)


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
1
2
3
4
5
6
All events
Possitive events
ΔR(Sub-leading Lepton, Leading*b*-jet)


All events
Possitive events
0.08
0.06
0.04
0.02
0
0
1
2
3
4
5
6
ΔR(Leading Lepton, closest b-jet)


0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
1
2
3
4
5
6
All events
Possitive events
ΔR(Leading Lepton,τh)


0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0
1
2
3
4
5
6
All events
Possitive events
ΔR(Sub-leading Lepton,τh)


All events
Possitive events
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
0
50
100
150
200
250
300
350
400
450
500
$m_{H_{vis\ opt1}}$