Multiclassification of Production Modes of Higgs Bosons Using TMVA

Chenji Han
University of Chinese Academy of Sciences
Junior undergraduate student

- I visited the UC Berkeley at this spring semester, and I am doing some research at the ATLAS group(HGamma subgroup) at LBNL.

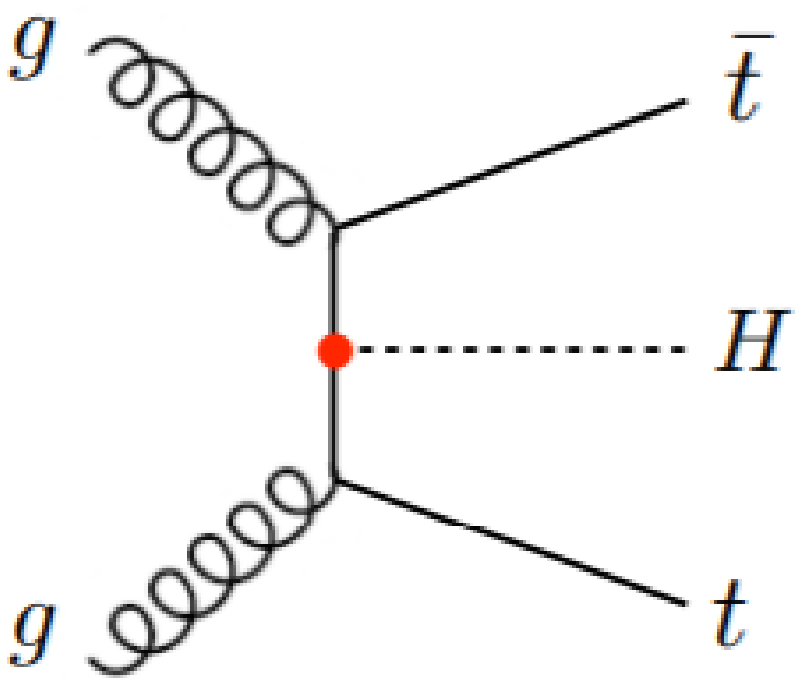
- I am mainly working on two projects:
  - The study of the CP property of Higgs bosons
  - The multiclassification of production modes of Higgs bosons

- Today, I will focus on the second project

Outline

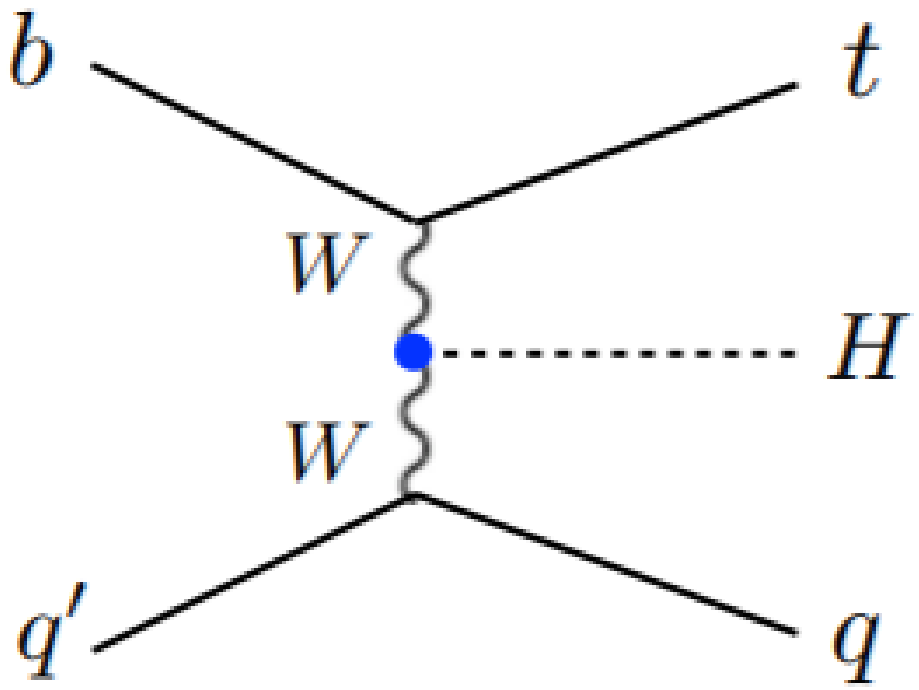
- The motivation of multiclassification of production modes of Higgs
- The brief introduction to the Decision Tree
- The brief introduction to the work flow of the multiclassification

Motivation



- The ttH production mode is preferred in the study of Higgs property:
  - 1. the b-tagging algorithm to reconstruct top quark is mature, and the reconstruction efficiency is high(60%)
  - 2. it's relatively easy to distinguish ttH with the background

Motivation

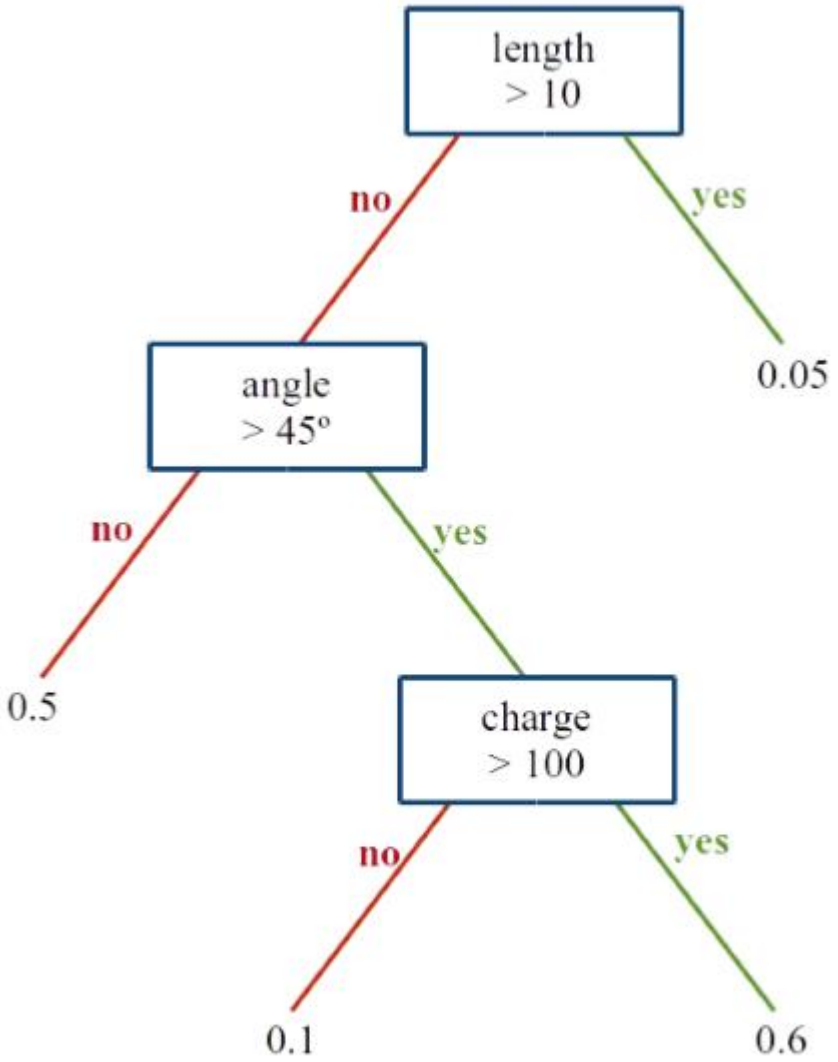


- the previous researches don’t distinguish the tH and ttH, for that:
  - 1. strictly search for ttH mode would cause low reconstruction efficiency(0.6*0.6=0.36)
  - 2. tH mode has much smaller cross section than the ttH
  - 3. in most cases, it’s not necessary

Motivation

- During the study of CP property of Higgs, it seems to be necessary to have category which is rich in ttH and category which is rich in tH
- Professor at LBNL asked me to setup a framework to multi-classify the ttH, tH and the background.

Brief Introduction to the Decision Trees



- Such trees with different structure and weight will form an ensemble.
- Each tree in the ensemble will give the probability of how likely the input is belong to each category.
- Finally, the ensemble will give the final probability of belonging to different categories according to the weight of the tree.

Brief Introduction to the Decision Trees

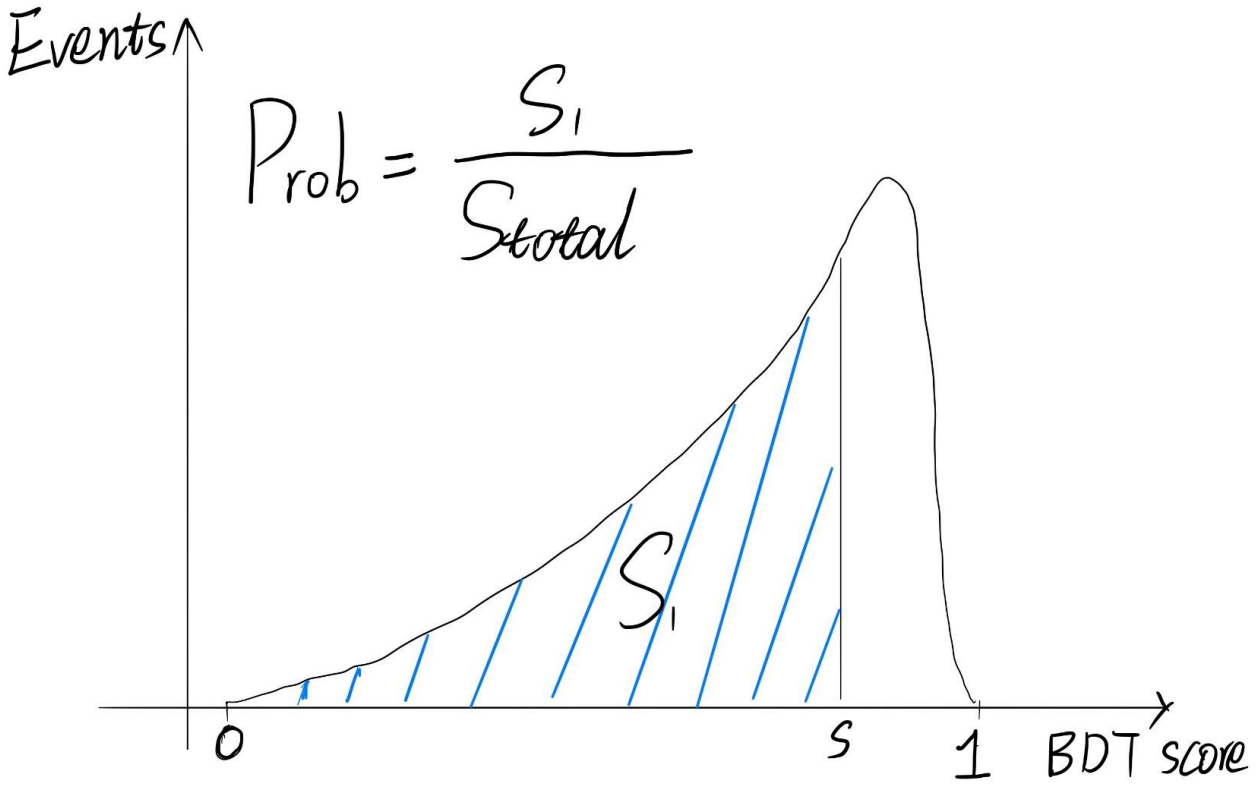
- The decision tree is based on the variables that would suggest the difference between different categories that we want to classify.
- The essence of decision trees is that different categories are clustered at different positions of the hyper-space whose basis is the variables

Work Flow

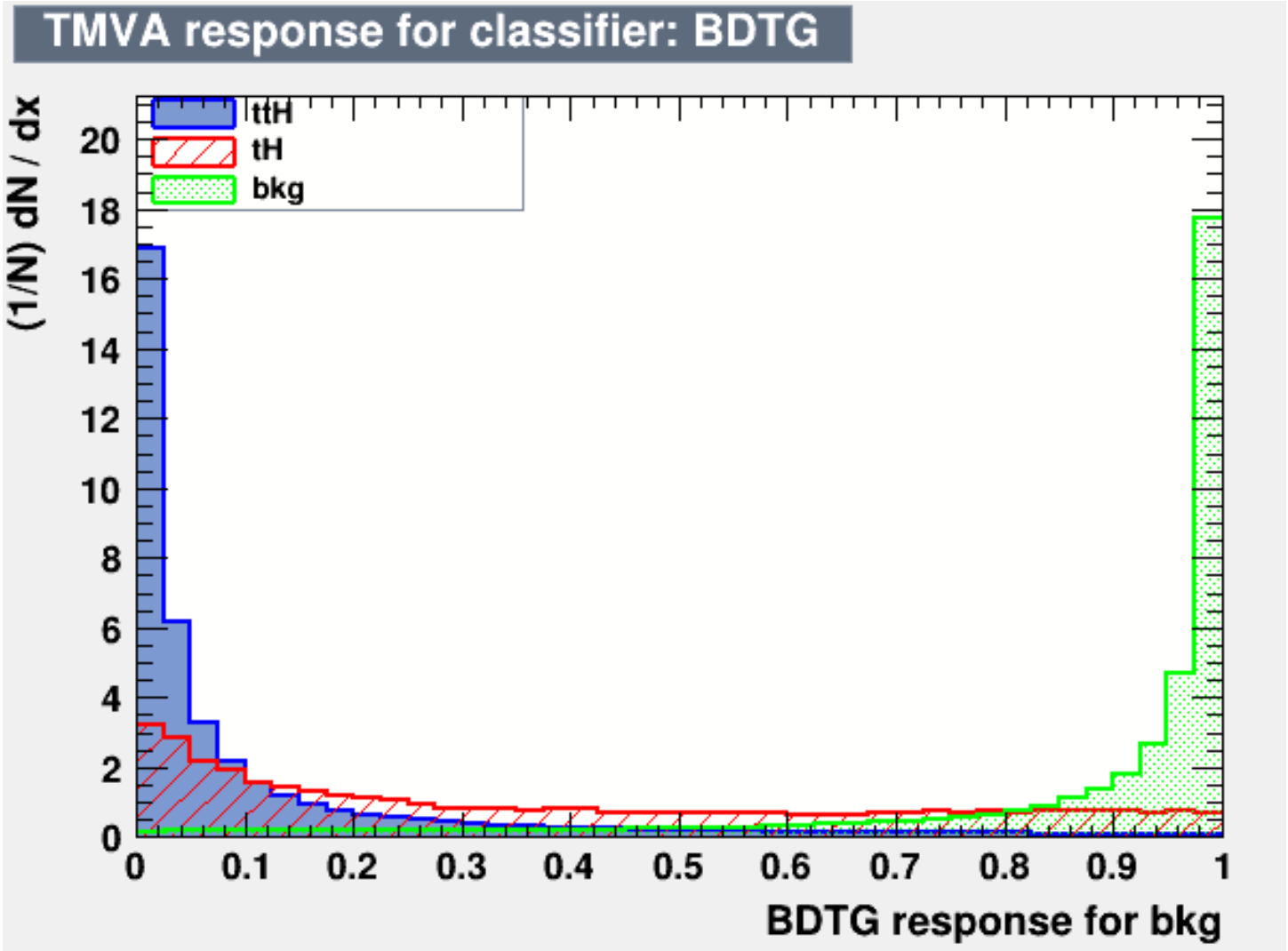
ROOT file of certain category		
50% validation	25% testing	25% training

Work Flow

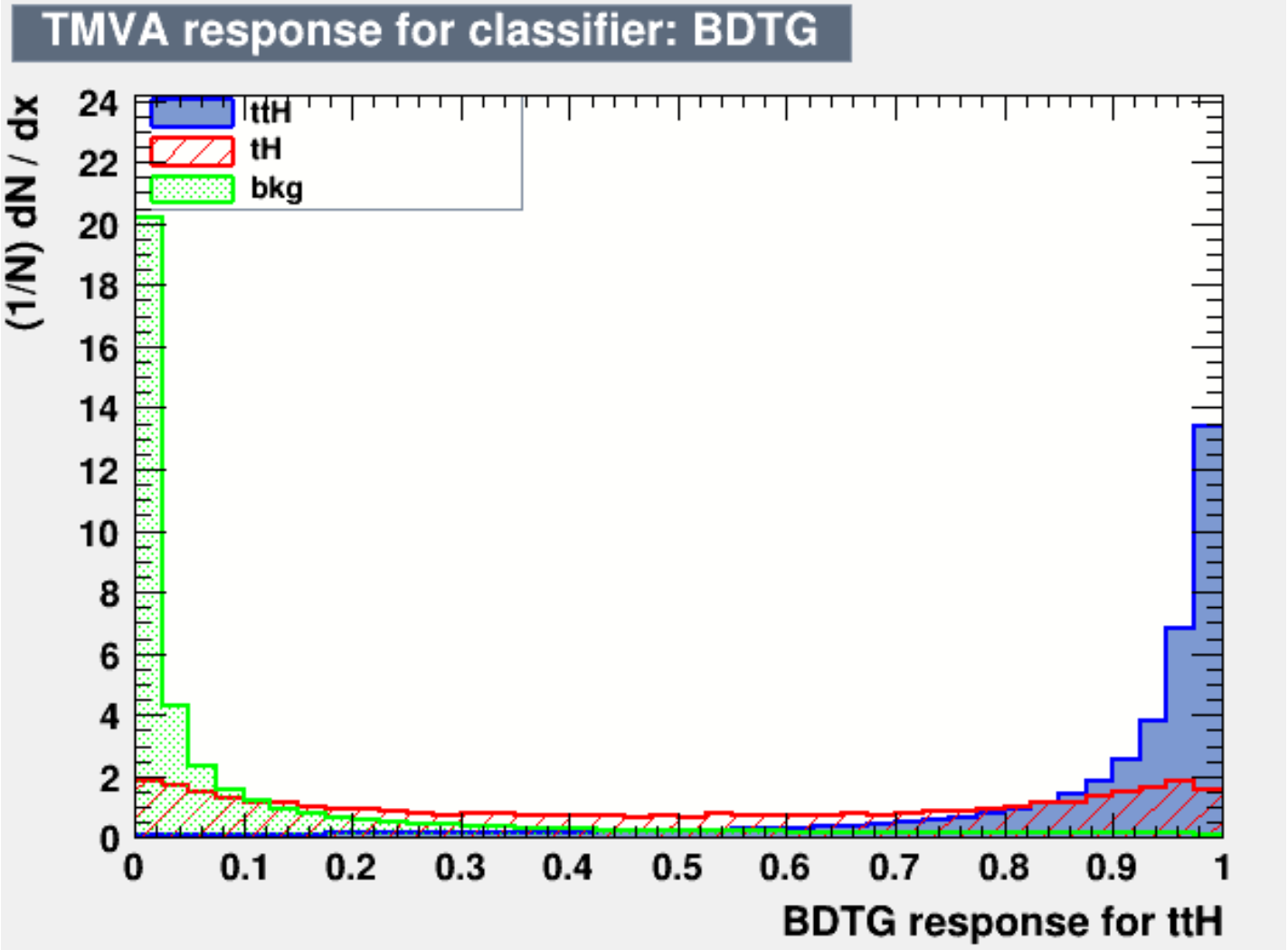
- The method to calculate the probability based on the BDT score distribution



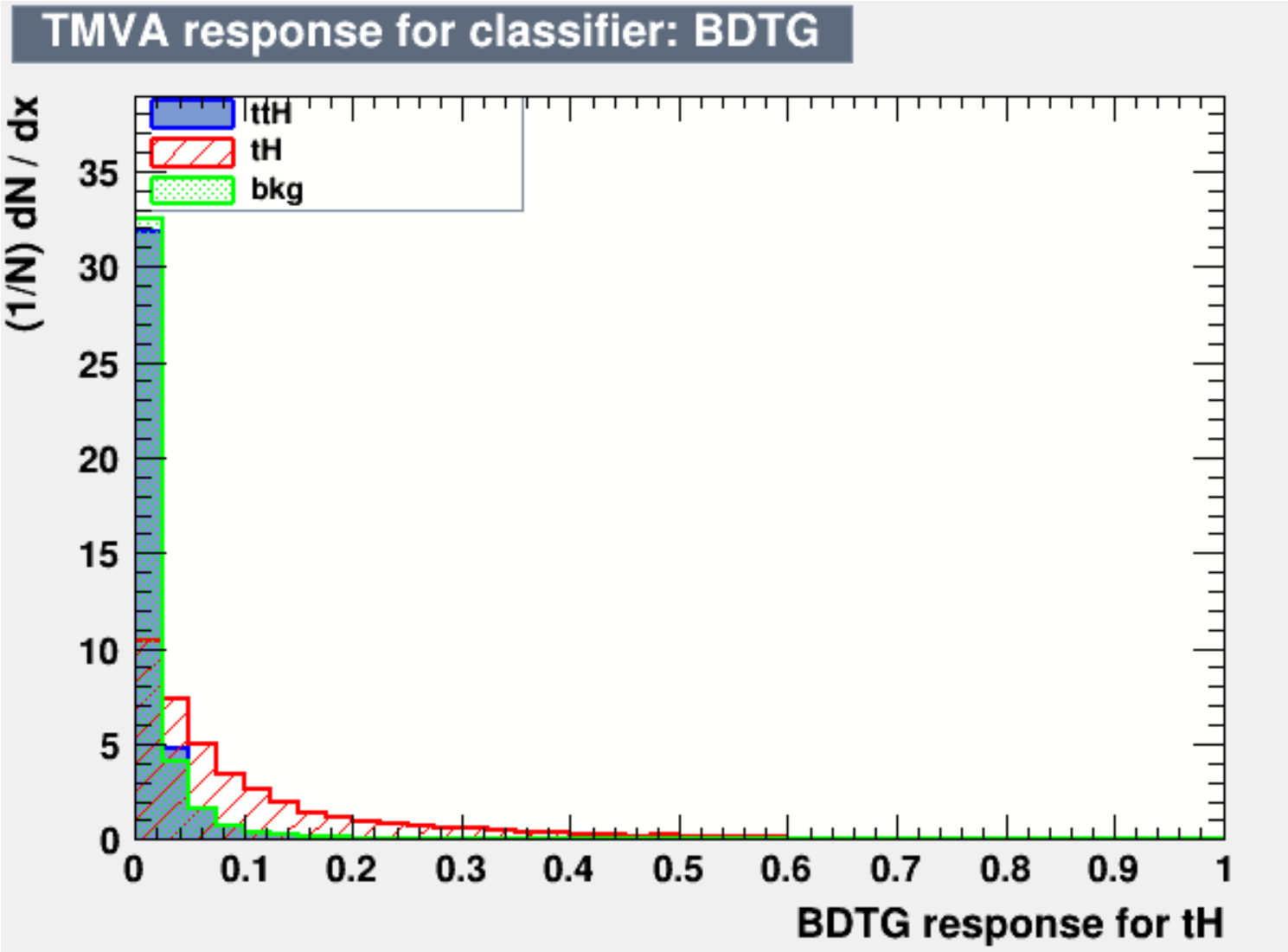
The Current Result



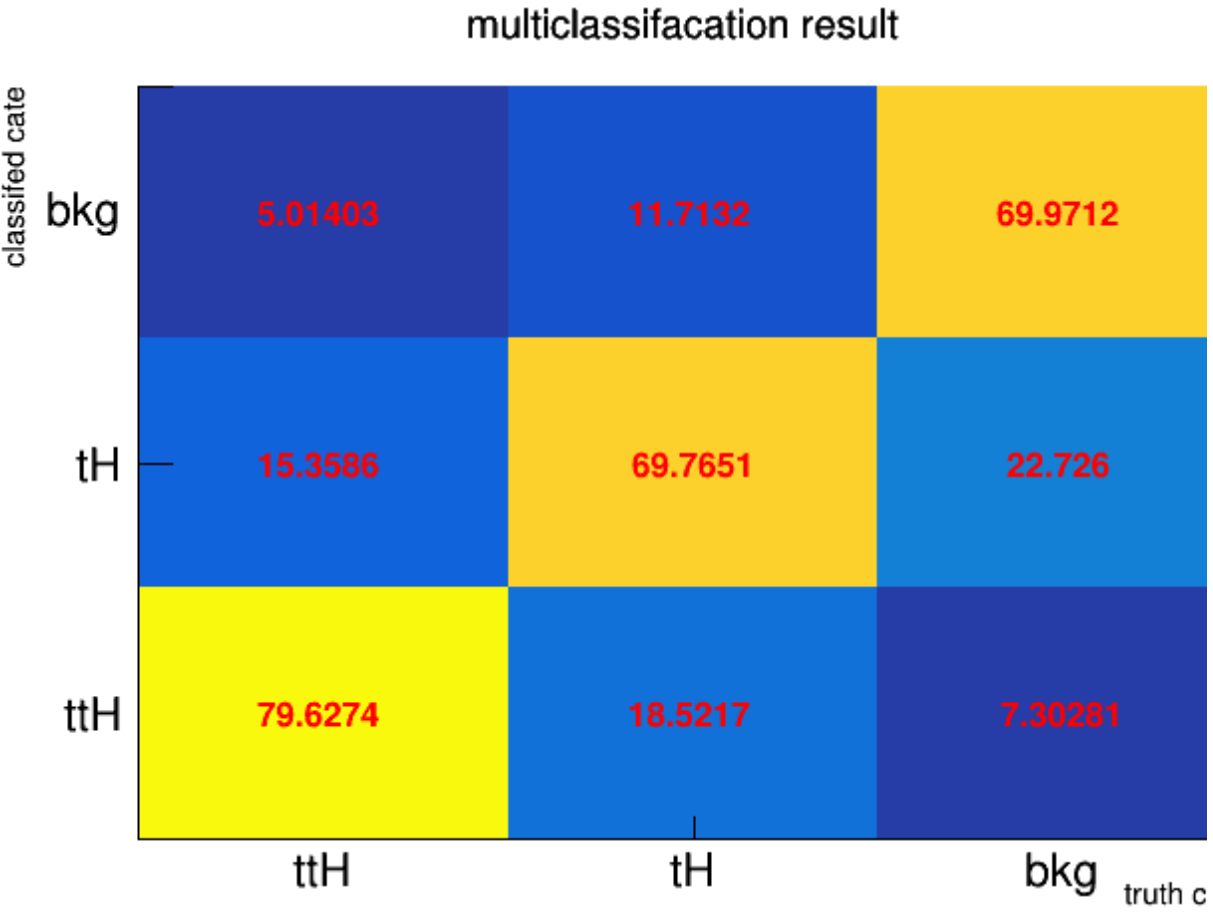
The Current Result

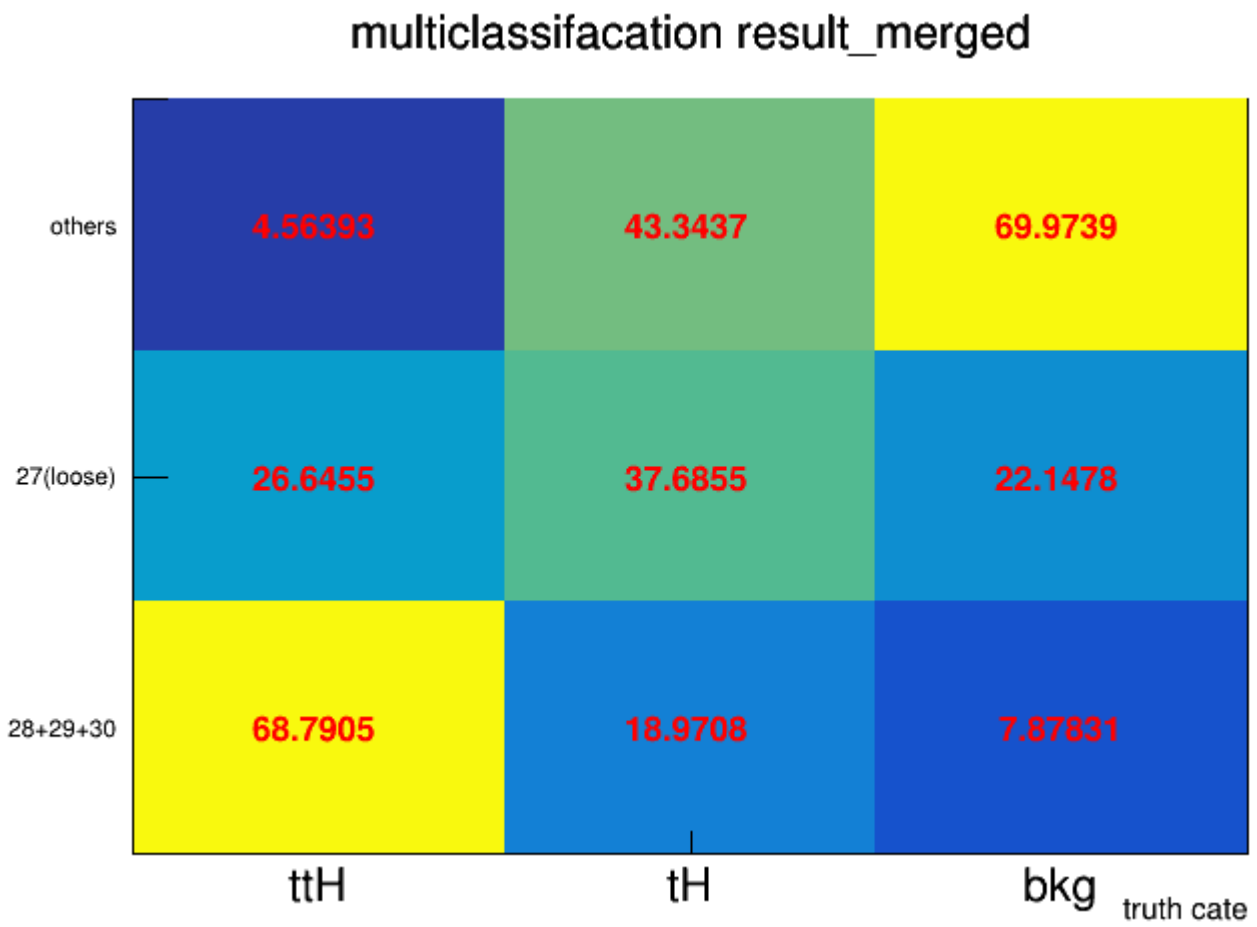


The Current Result

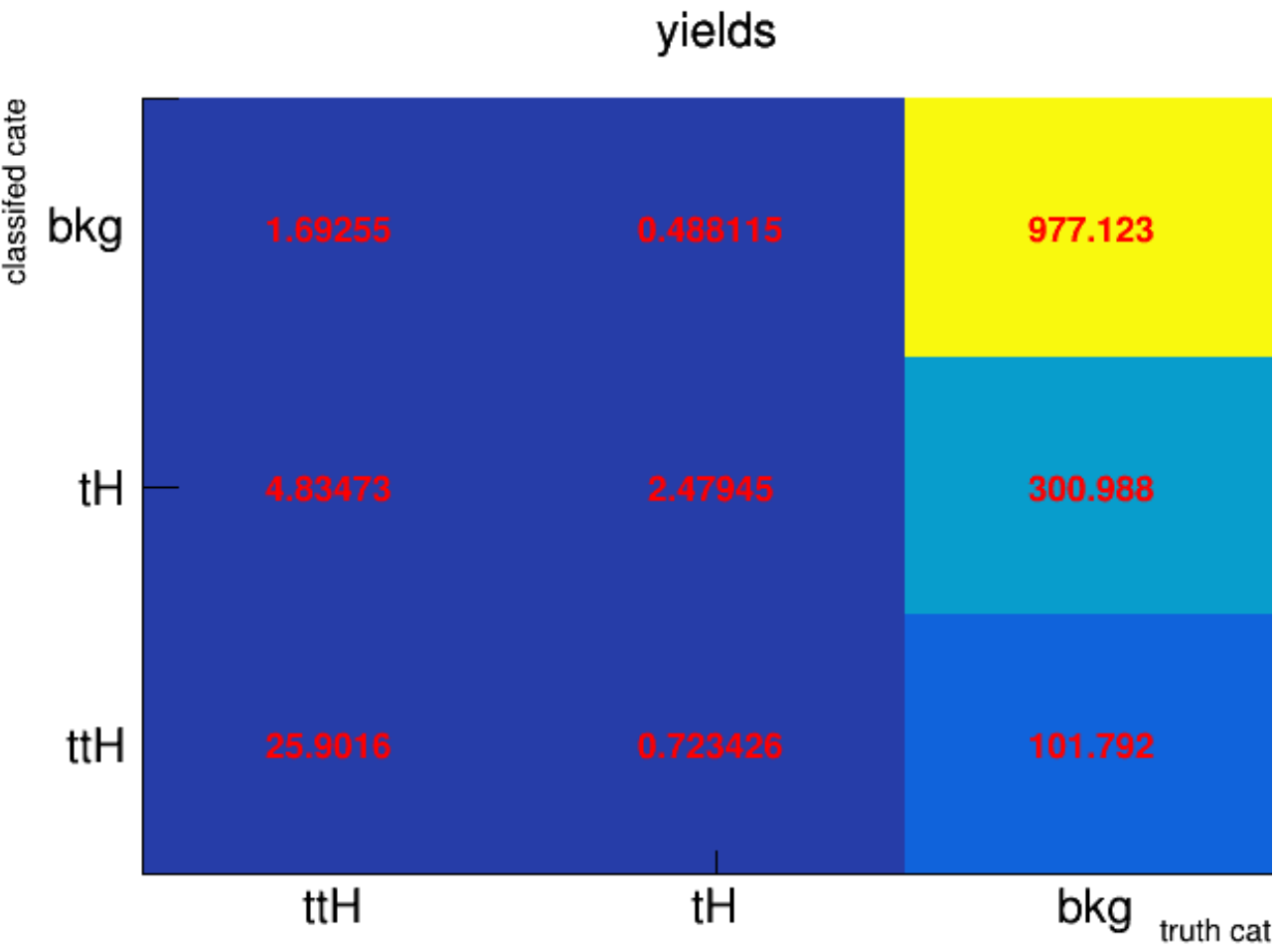


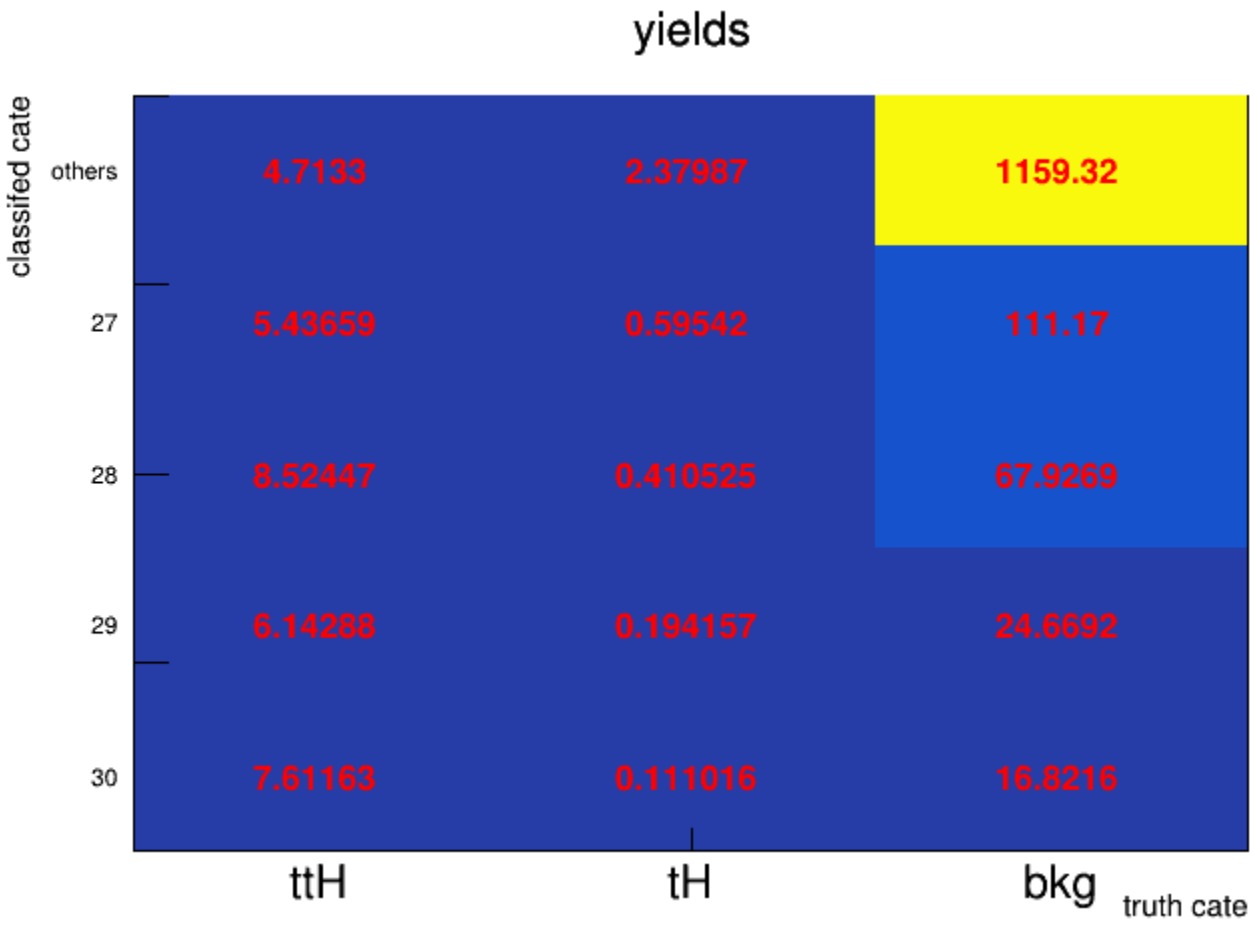
The Current Result





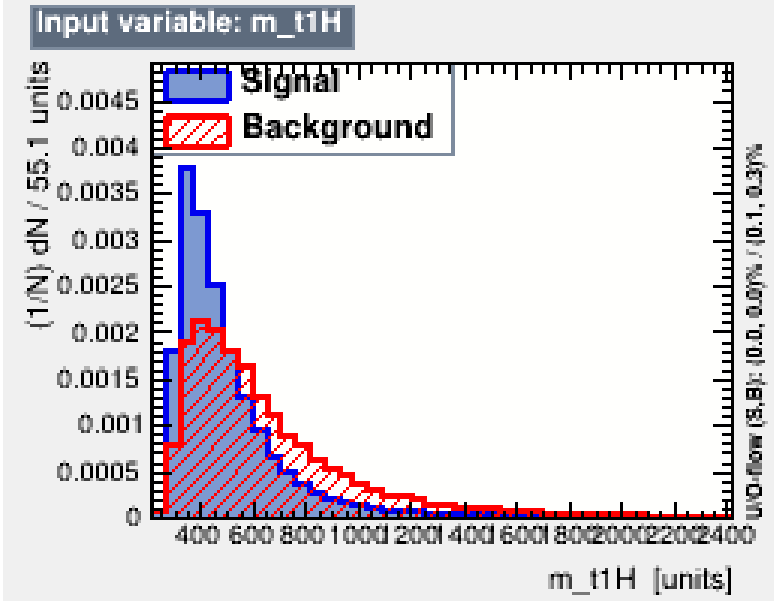
The Current Result

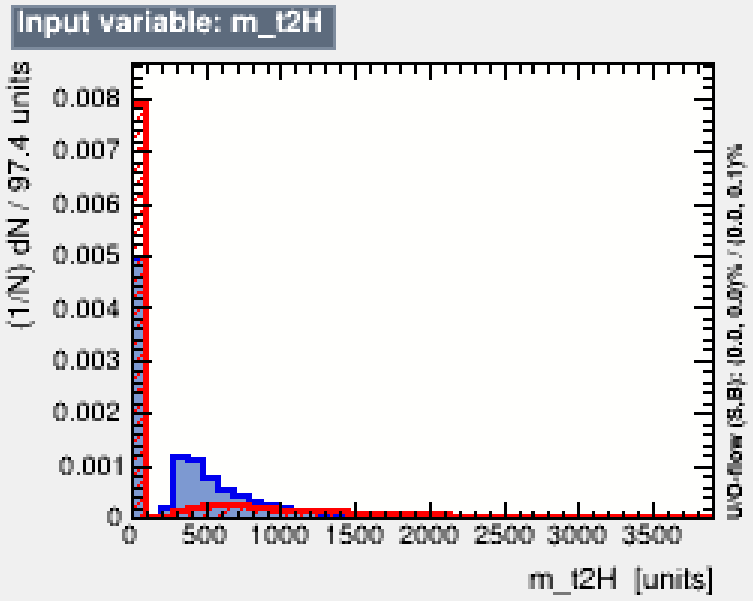


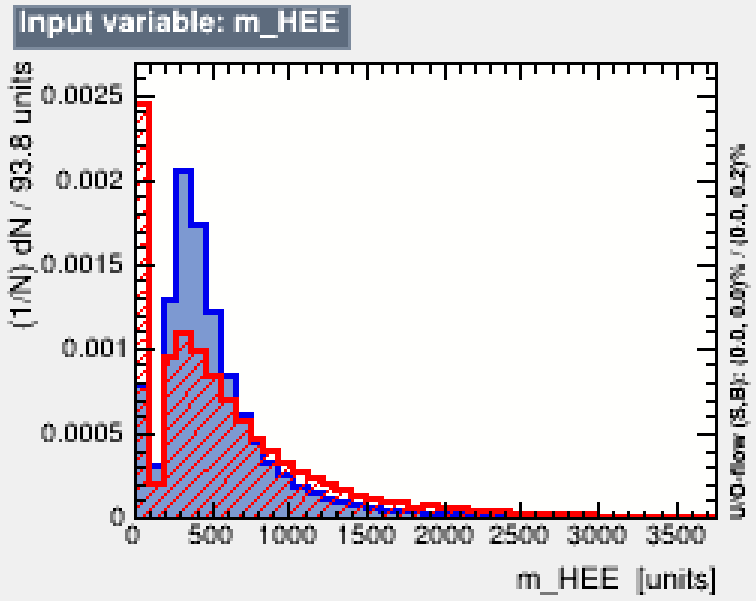


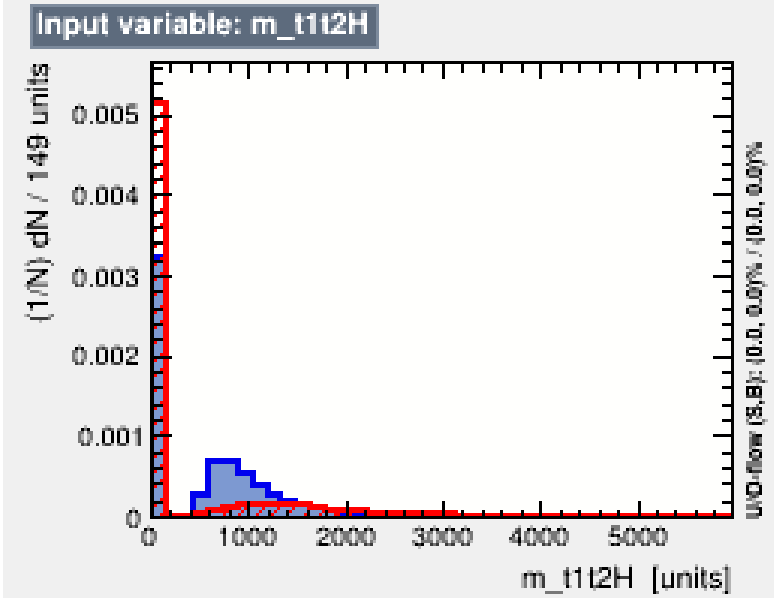
Thanks!

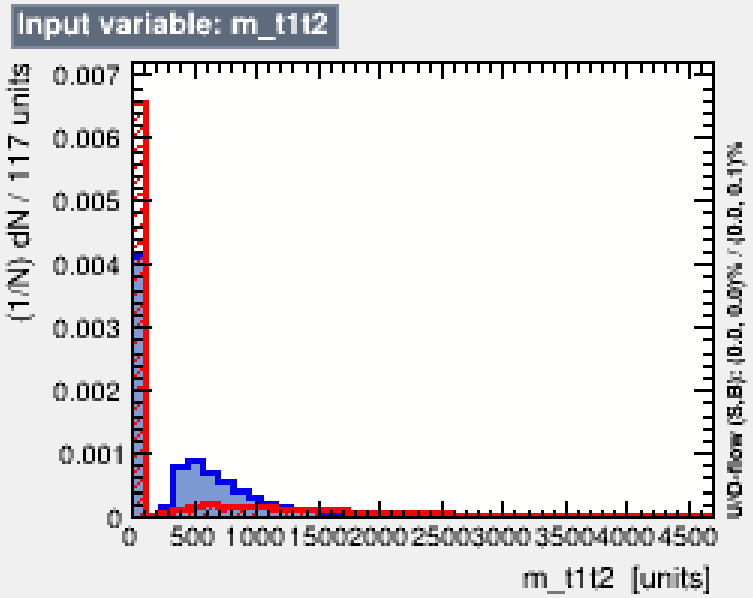
Back up

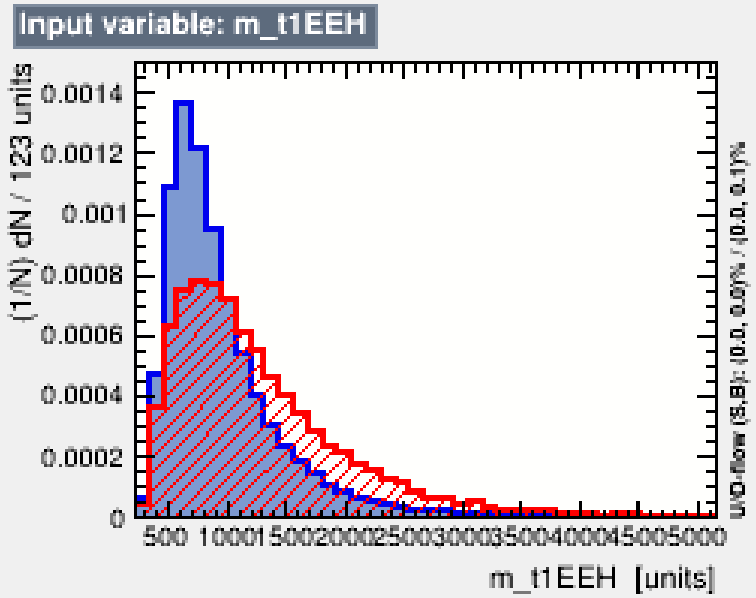

Input variable: m_t1H
Signal
Background
(1/N) dN / 55.1 units
m_t1H [units]
U/O-flow (S,B): (0.0, 0.0)% / (0.1, 0.3)%

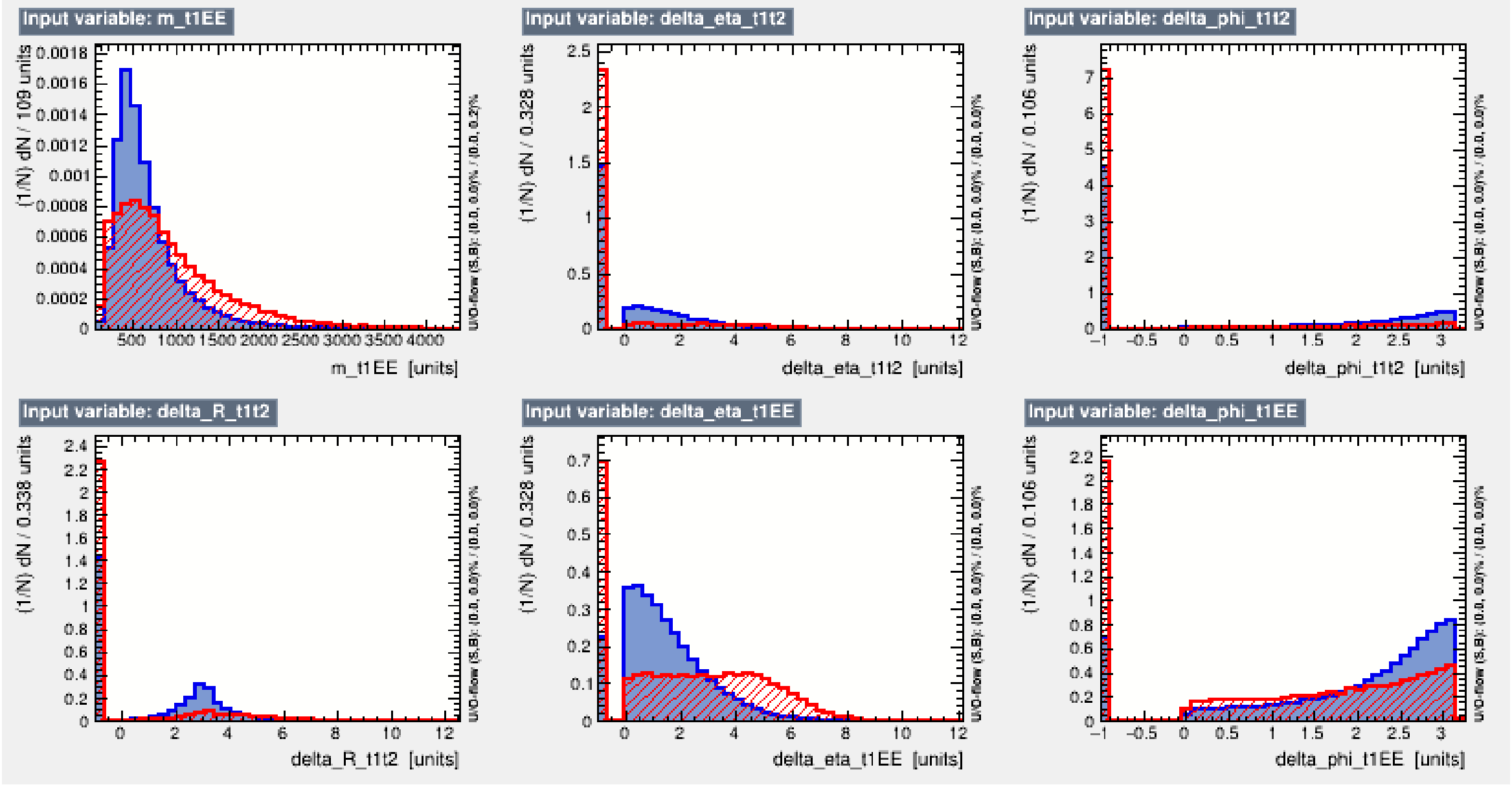

Input variable: m_t2H
(1/N) dN / 97.4 units
m_t2H [units]
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.1)%

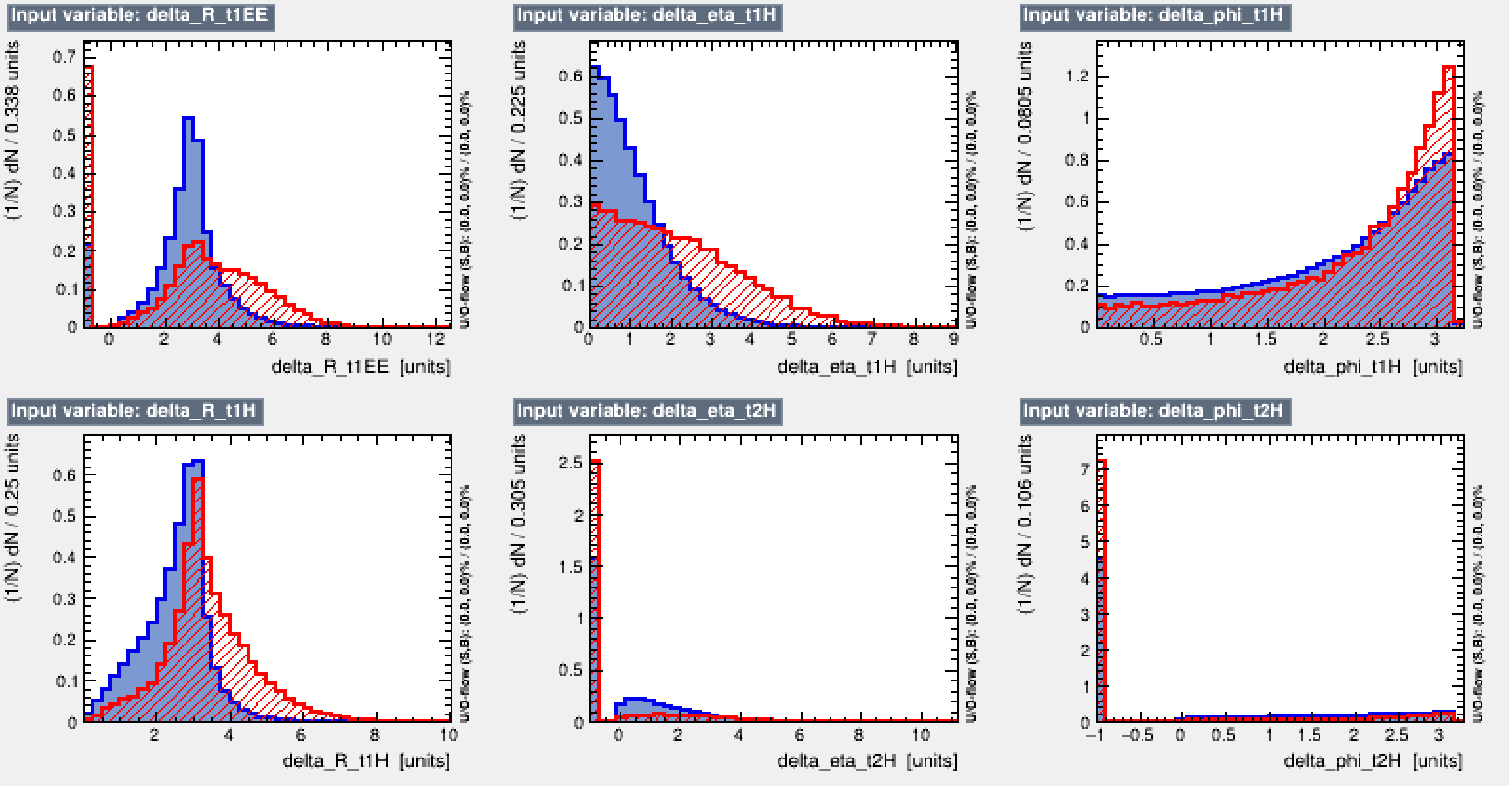

Input variable: m_HEE
(1/N) dN / 93.8 units
m_HEE [units]
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.2)%

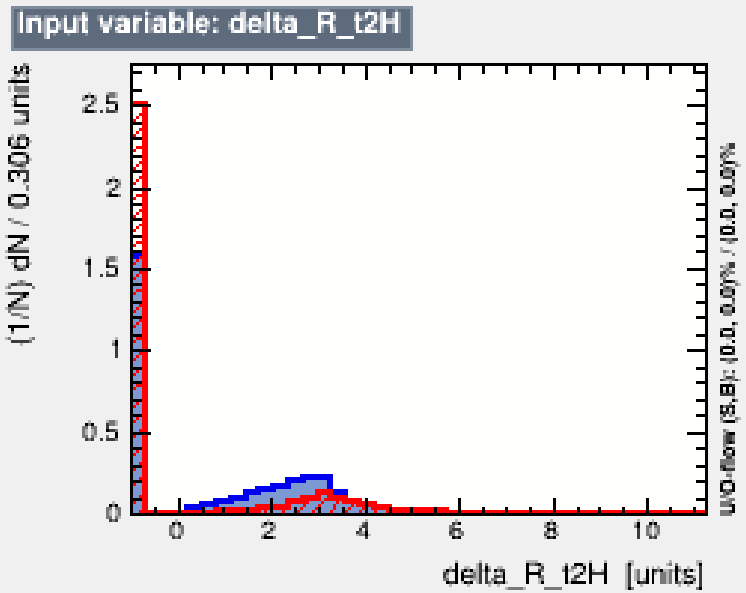

Input variable: m_t1t2H
(1/N) dN / 149 units
m_t1t2H [units]
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.0)%

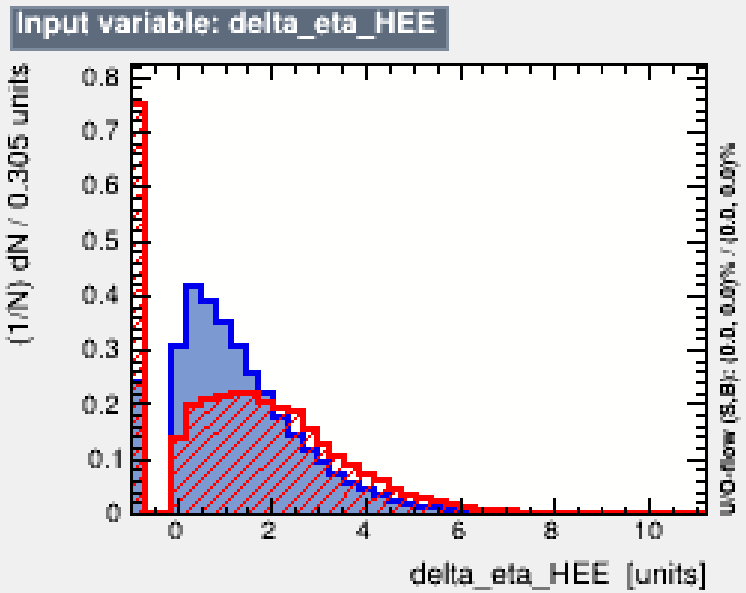

Input variable: m_t1t2
(1/N) dN / 117 units
m_t1t2 [units]
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.1)%

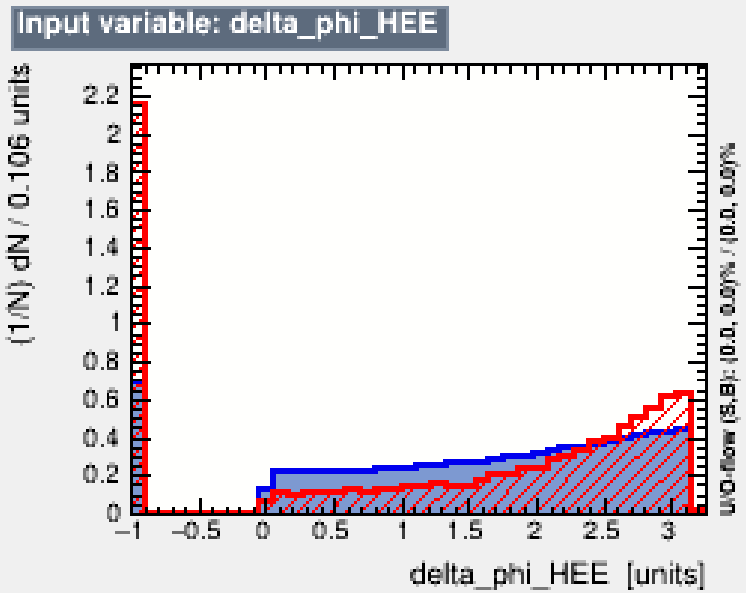

Input variable: m_t1EEH
(1/N) dN / 123 units
m_t1EEH [units]
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.1)%

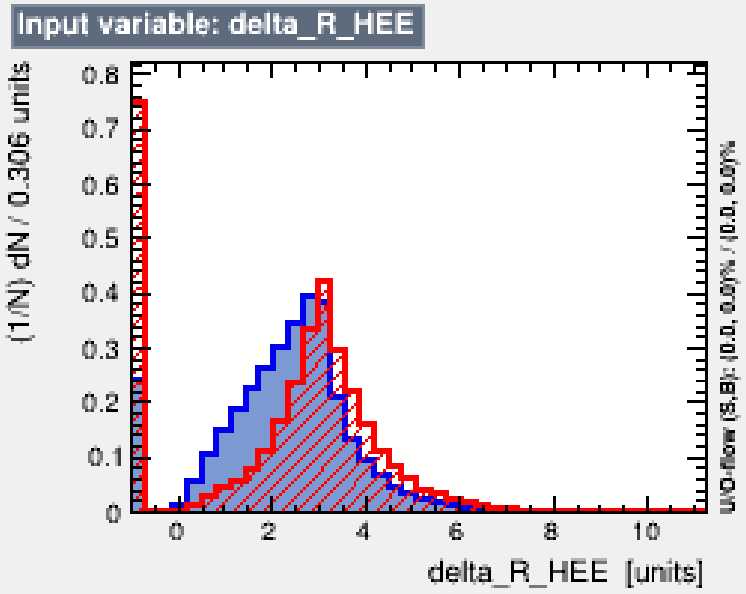

Input variable: m_t1EE
(1/N) dN / 109 units
m_t1EE [units]
Input variable: delta_eta_t1t2
(1/N) dN / 0.328 units
delta_eta_t1t2 [units]
Input variable: delta_phi_t1t2
(1/N) dN / 0.106 units
delta_phi_t1t2 [units]
Input variable: delta_R_t1t2
(1/N) dN / 0.338 units
delta_R_t1t2 [units]
Input variable: delta_eta_t1EE
(1/N) dN / 0.328 units
delta_eta_t1EE [units]
Input variable: delta_phi_t1EE
(1/N) dN / 0.106 units
delta_phi_t1EE [units]

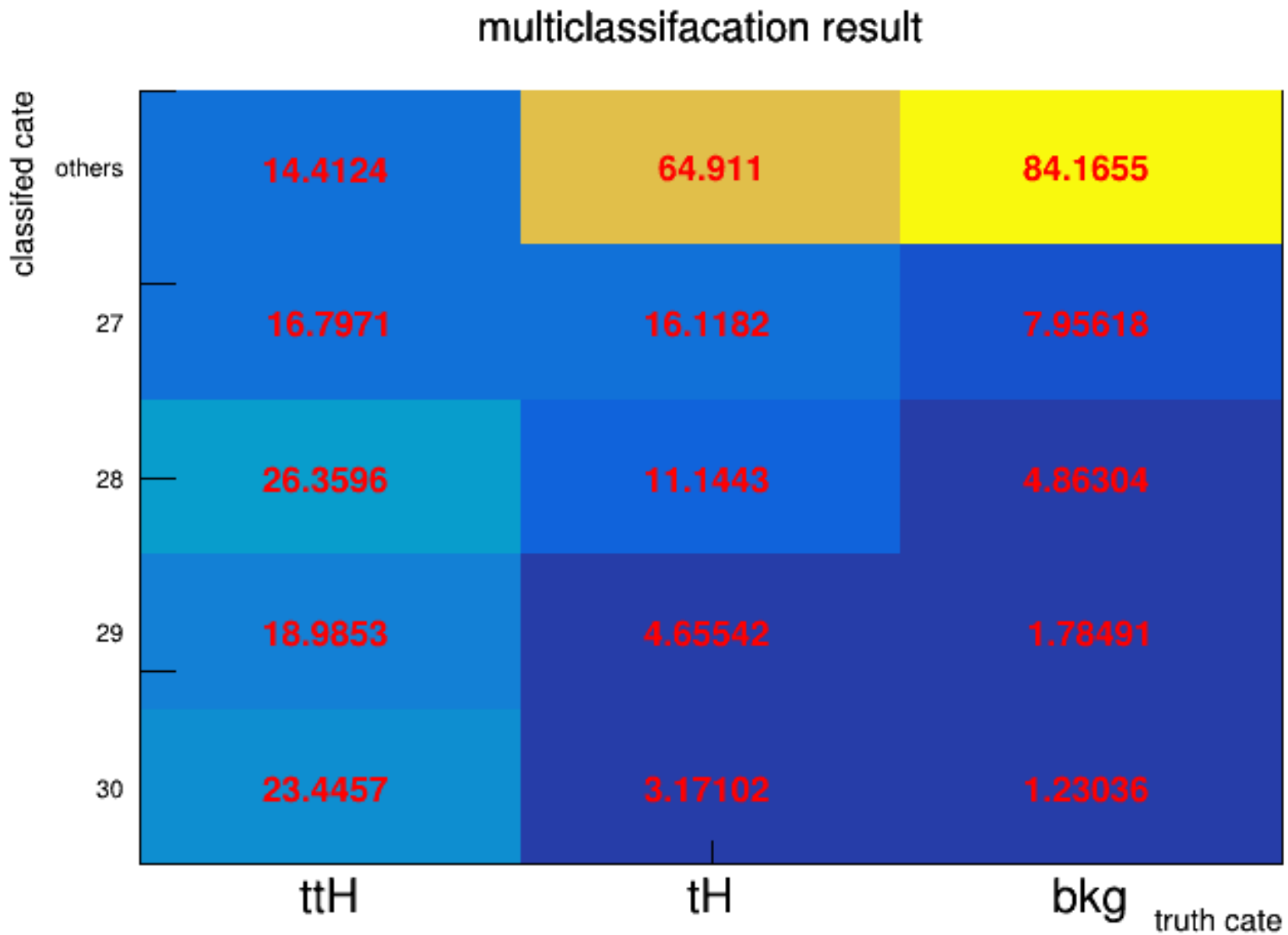

Input variable: delta_R_t1EE
(1/N) dN / 0.338 units
delta_R_t1EE [units]
Input variable: delta_eta_t1H
(1/N) dN / 0.225 units
delta_eta_t1H [units]
Input variable: delta_phi_t1H
(1/N) dN / 0.0805 units
delta_phi_t1H [units]
Input variable: delta_R_t1H
(1/N) dN / 0.25 units
delta_R_t1H [units]
Input variable: delta_eta_t2H
(1/N) dN / 0.305 units
delta_eta_t2H [units]
Input variable: delta_phi_t2H
(1/N) dN / 0.106 units
delta_phi_t2H [units]
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.0)%

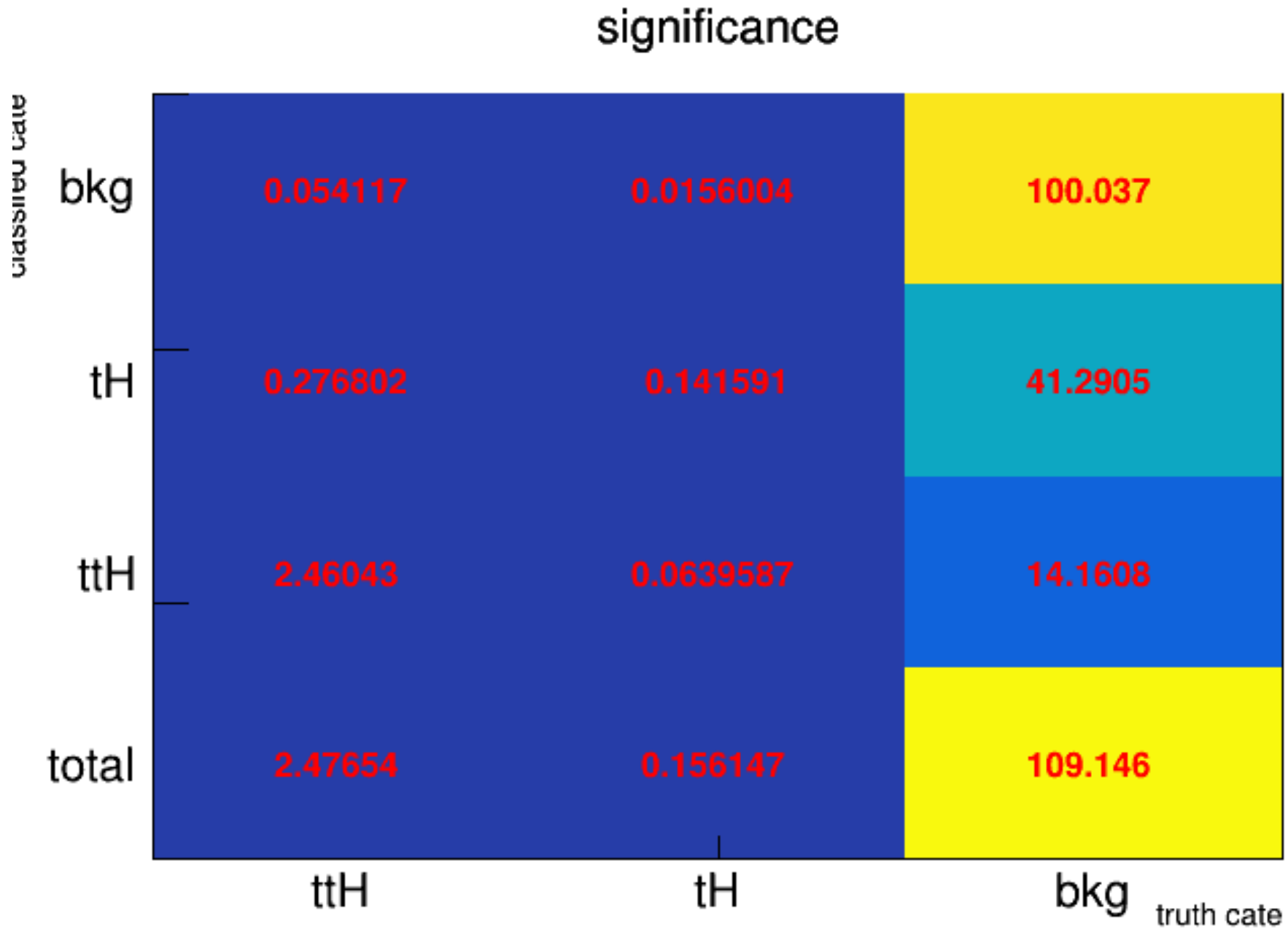

Input variable: delta_R_t2H
(1/N) dN / 0.306 units
delta_R_t2H [units]
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.0)%


Input variable: delta_eta_HEE
(1/N) dN / 0.305 units
delta_eta_HEE [units]
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.0)%


Input variable: delta_phi_HEE
(1/N) dN / 0.106 units
delta_phi_HEE [units]
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.0)%


Input variable: delta_R_HEE
(1/N) dN / 0.306 units
delta_R_HEE [units]
U/O-flow (S,B): (0.0, 0.0)% / (0.0, 0.0)%


multiclassifacation result
classifed cate
others
27
28
29
30
14.4124
64.911
84.1655
16.7971
16.1182
7.95618
26.3596
11.1443
4.86304
18.9853
4.65542
1.78491
23.4457
3.17102
1.23036
ttH
tH
bkg
truth cate


significance
bkg
tH
ttH
total
0.054117
0.0156004
100.037
0.276802
0.141591
41.2905
2.46043
0.0639587
14.1608
2.47654
0.156147
109.146
ttH
tH
bkg
truth cate