A Comparative Machine Learning Study of Color Tagger Variables in VH(bb)

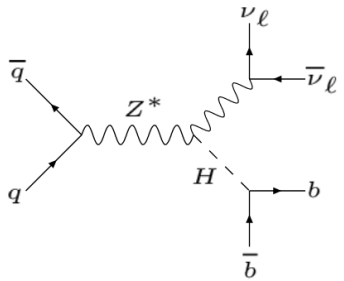
L. Cavallini, A. Coccaro, G. Manco, S. Marzani, F. Parodi,
D. Rebuzzi, A. Rescia, G. Stagnitto

Jets and their substructure from LHC data

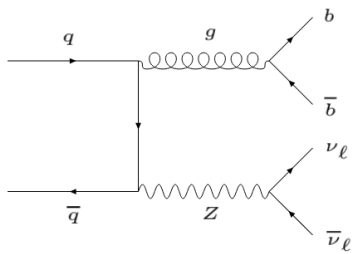
June 1, 2021

Objective

- Train machine learning algorithms to distinguish decays from color singlets vs. color octects
- Focus on two processes:
  - Signal: $pp \to ZH$, $Z \to \nu_\ell \overline{\nu}_\ell$, $H \to b\overline{b}$ (ZHbb)
  - Background: $pp \to b\overline{b}\nu_\ell \overline{\nu}_\ell$ (gbb)
- Analysis of simulated data and extraction of the **8 variables**
  - Pull vector components and pull angle $t_{\parallel a}$, $t_{\perp a}$, θ_{pa} relative to Jet J_a
  - Pull vector components and pull angle $t_{\parallel b}$, $t_{\perp b}$, θ_{pb} relative to Jet J_b
  - Color Ring (cr)
  - D_2
- Training of BDT and NNs on high-level variables

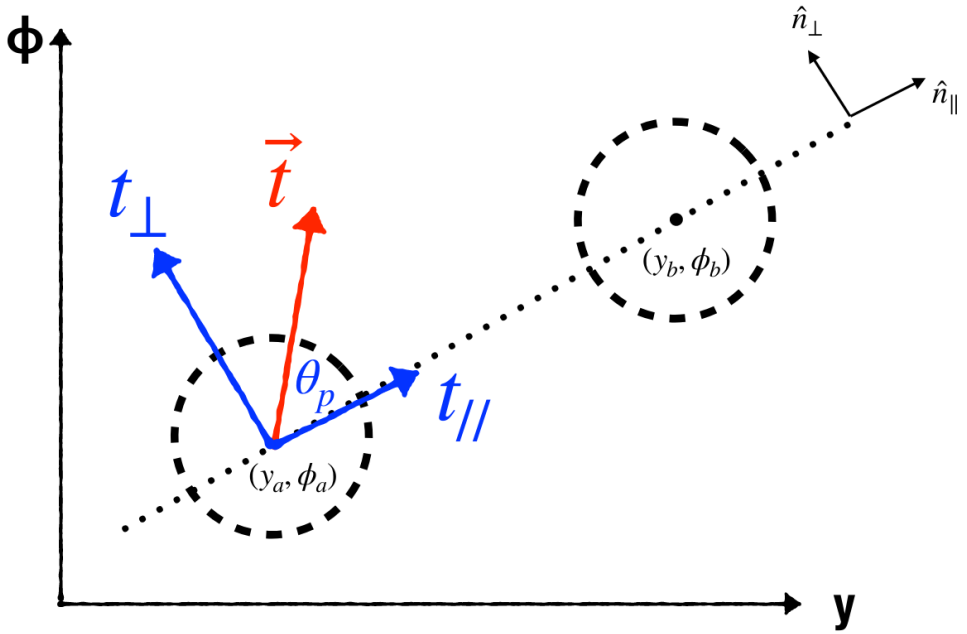


ZHbb



gbb

Jet Pull Components



Pull vector components of Jet J_a
[arXiv:1911.05090v2]

$$\vec{t} = \frac{1}{p_{t_a}} \sum_{i \in J_a} p_{t_i} |\vec{r}_i|^2 \hat{r}_i$$

$$\vec{r}_i = (y_i - y_a, \phi_i - \phi_a)$$

$$\hat{n}_\parallel = \frac{1}{\sqrt{\Delta y^2 + \Delta\phi^2}} (\Delta y, \Delta\phi)$$

$$\hat{n}_\perp = \frac{1}{\sqrt{\Delta y^2 + \Delta\phi^2}} (-\Delta\phi, \Delta y)$$

$$t_\parallel = \vec{t} \cdot \hat{n}_\parallel$$

$$t_\perp = \vec{t} \cdot \hat{n}_\perp$$

$$\theta_p = \arccos\left(\frac{t_\parallel}{|\vec{t}|}\right)$$

$t_\parallel, t_\perp$ **are IRC safe observables!**
θ_p **is Sudakov safe**

D_2

- D_2 is defined as

$$D_2^{(\beta)} = \frac{e_3^{(\beta)}}{(e_2^{(\beta)})^3}$$

- $e_n^{(\beta)}$ is the normalized n-point Energy Correlator function
- For this case,

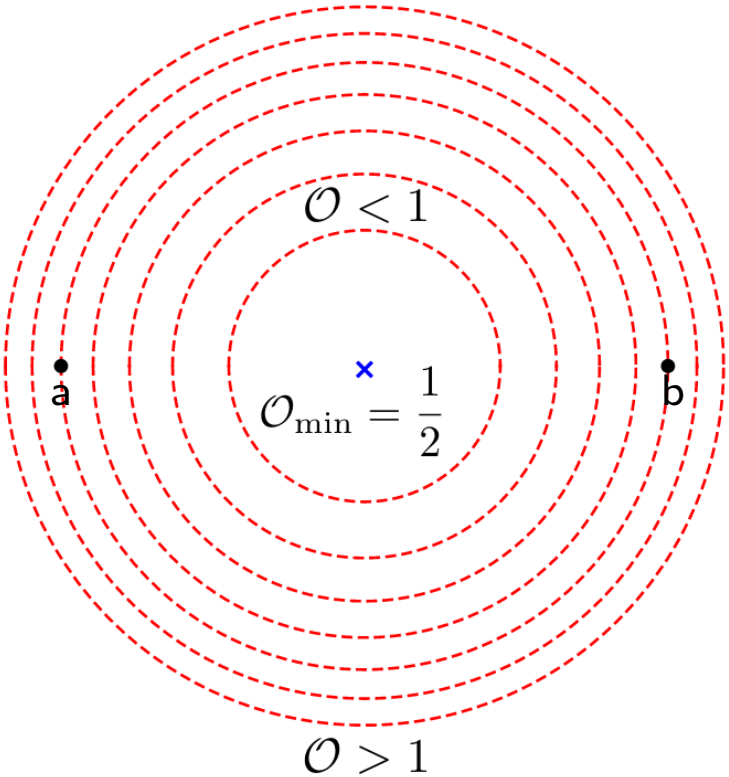
$$e_2^{(\beta)} = \frac{1}{p_{TJ}^2} \sum_{1 \leq i < j \leq n_J} p_{T_i} p_{T_j} R_{ij}^{\beta}$$

$$e_3^{(\beta)} = \frac{1}{p_{TJ}^3} \sum_{1 \leq i < j < k \leq n_J} p_{T_i} p_{T_j} p_{T_k} R_{ij}^{\beta} R_{ik}^{\beta} R_{jk}^{\beta}$$

- p_{TJ} transverse momentum of jet w.r.t. beam
- $R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$
- $\beta = 2$

[arXiv:1409.6298v1]

Color Ring



[arXiv:2006.10480v2]

$$\mathcal{O} = \frac{|\mathcal{M}_{\mathcal{B}}|^2}{|\mathcal{M}_{\mathcal{S}}|^2} \approx \frac{\theta_{ak}^2 + \theta_{bk}^2}{\theta_{ab}^2}$$

- θ_{ak} (θ_{bk}) angle between hard parton a (b) and gluon (k)
  $\Rightarrow$ Requires 3 objects!
- θ_{ab} angle between hard partons
- **IRC safe observable!**

Generation Chain

- 1 M Signal/Bkg events in MG5_aMC v2.8.3.2
- Shower in Pythia v8.235
- Run in Delphes v3.4.2 w/ modified ATLAS card
  - Extract Monte Carlo Truth
  - Extract Fast Detector Reconstruction

Truth

	Events Passed
Signal	$\sim$ 200k
Background	$\sim$ 60k

Reco

	Events Passed
Signal	$\sim$ 140k
Background	$\sim$ 60k

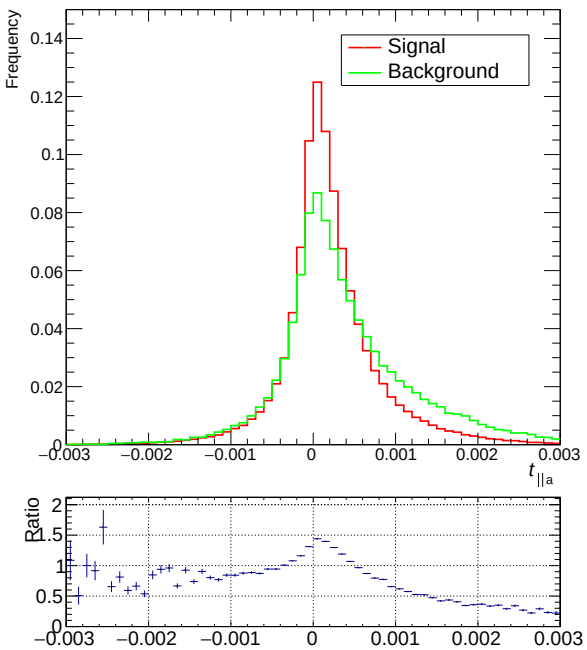
Analysis

- **Truth:** Remove ν's and cluster stable particles w/ $p_T > 0.5$ GeV into jets
- **Reco:** Cluster into jets
  - EM Calo Towers
    - $E_{min} = 0.5$ GeV
    - $S_{min} = 2.0$ (Significance)
  - Hadron Calo Towers
    - $E_{min} = 1.0$ GeV
    - $S_{min} = 2.0$ (Significance)
  - Tracks w/ $p_T > 0.5$ GeV
- Identify Large Jets and choose the hardest
  - R = 1.0
  - $p_T > 250$ GeV
  - $|y| < 1.5$
- Identify subjets
  - R = 0.2
  - $p_T > 10$ GeV
  - ΔR = 0.8 from Large Jet
- Angular b-labeling and selection of events w/ exactly 2 b-subjets
  - b-parton $p_T > 5$ GeV
  - ΔR = 0.2
  - $|\eta| < 2.5$
- Choose hardest non-b subjet for color ring
  - If not present, cr = -1

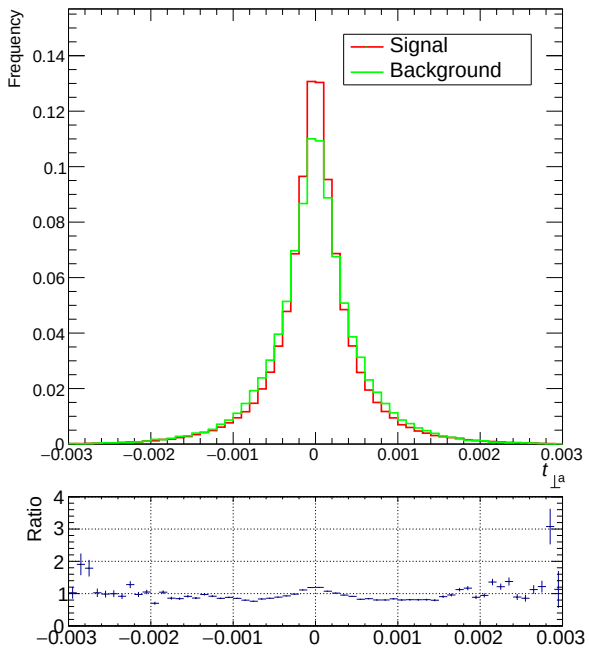
	Events Passed w/ $cr = -1$
ZHbb Truth	58%
ZHbb Reco	53%
gbb Truth	58%
gbb Reco	66%

Distributions - Signal vs. Bkg (Truth)

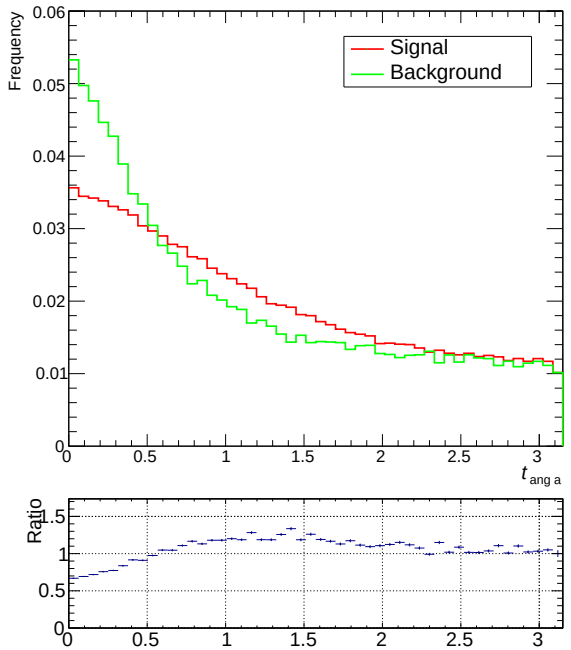
$t_{\parallel a}$



$t_{\perp a}$

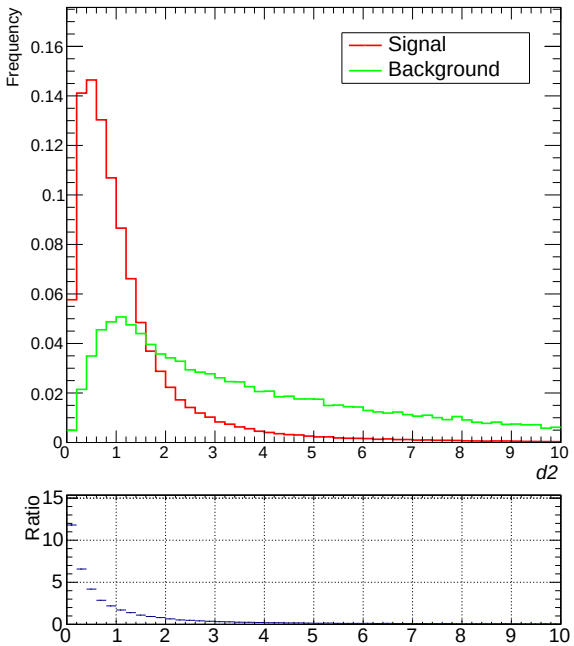


θ_{pa}

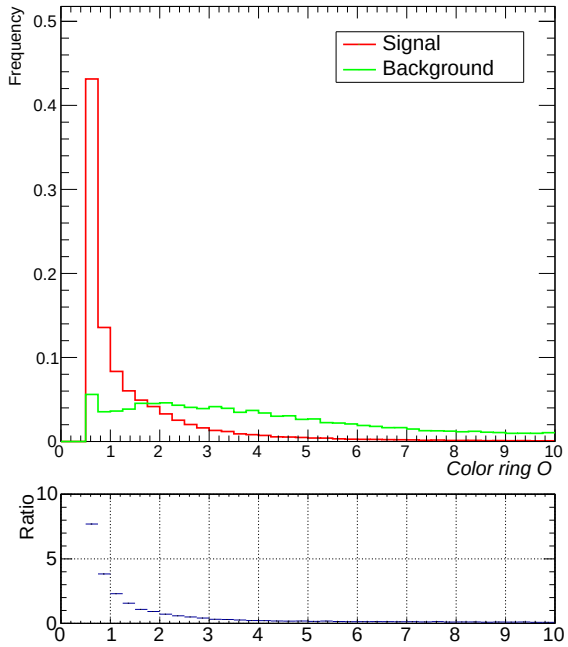


Distributions - Signal vs. Bkg (Truth)

D_2



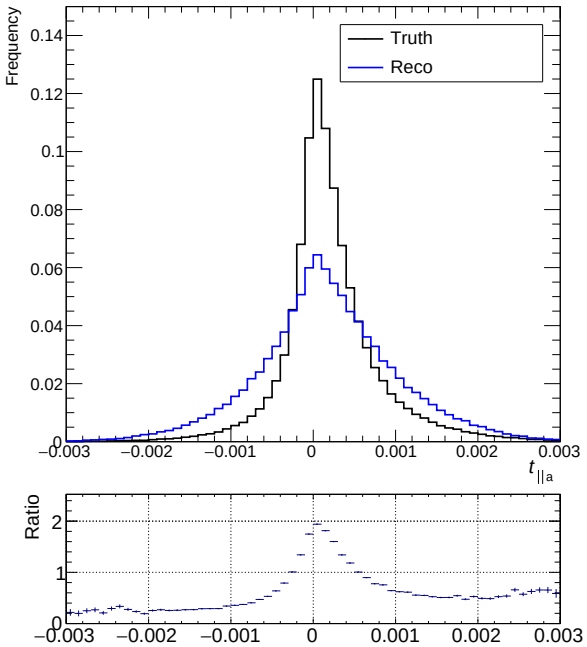
cr



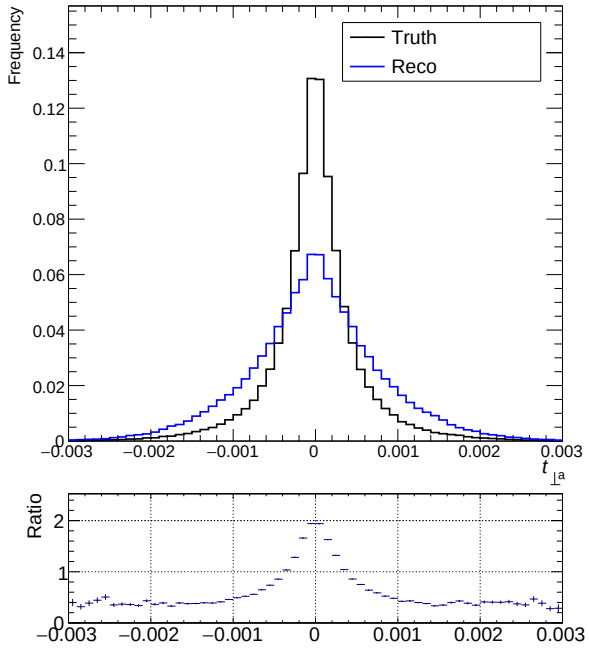
MC truth shows strong discrimination potential for the high-level variables considered

Distributions - Truth vs. Reco (ZHbb)

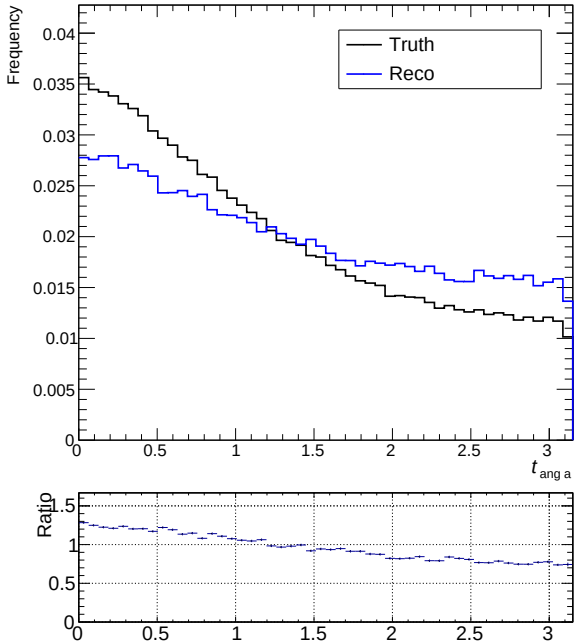
$t_{\parallel a}$



$t_{\perp a}$

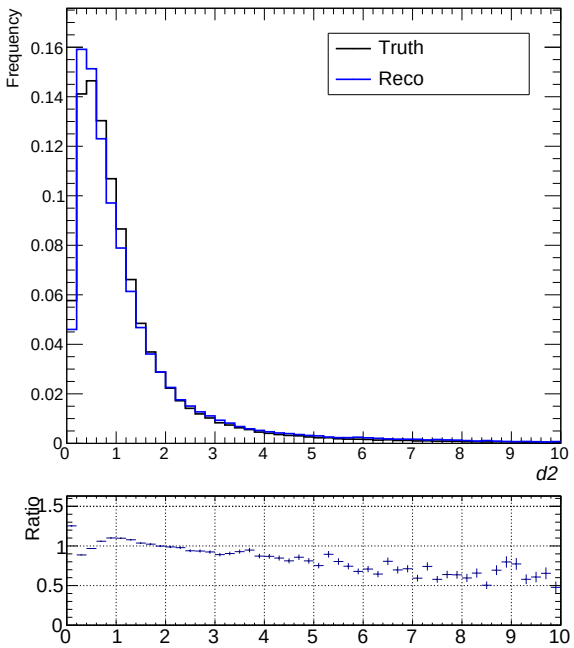


θ_{pa}

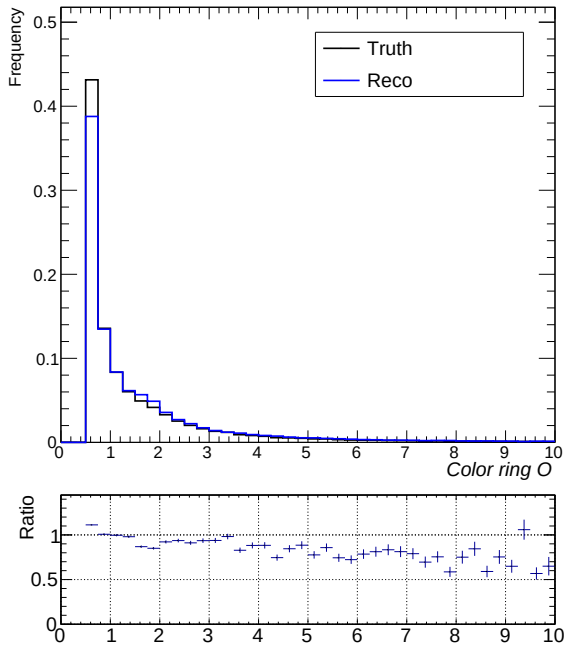


Distributions - Truth vs. Reco (ZHbb)

D_2



cr



Truth and Reco events show similar features

Machine Learning Parameters

BDT - TMVA

Parameters	Value
No. of Trees	50
Max Depth	5
MinNodeSize	2.5%
Boost Type	AdaBoost
Train/Test	60/40
No. of Cuts	80
Downsampling	No

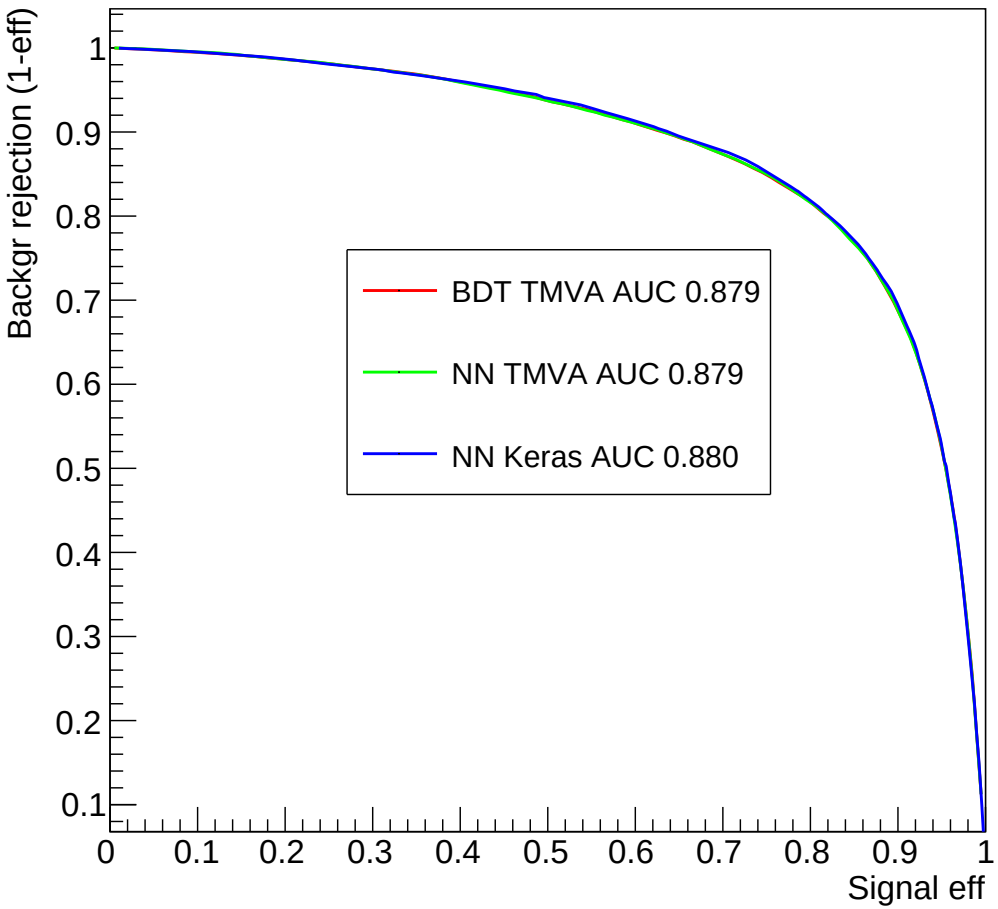
NN - TMVA

Parameters	Value
Activation	ReLU
Epochs	600
Hidden Layers	1
Neurons	9
Train/Test	60/40
Downsampling	No
Batch size	Entire Sample
Learning Rate	0.02

NN - Keras

Parameters	Value
Activation	tanh
Epochs	300
Hidden Layers	3
Neurons	9, 10, 5
Train/Test	70/30
Downsampling	Yes
Batch size	1000
Learning Rate	0.001

Results Truth

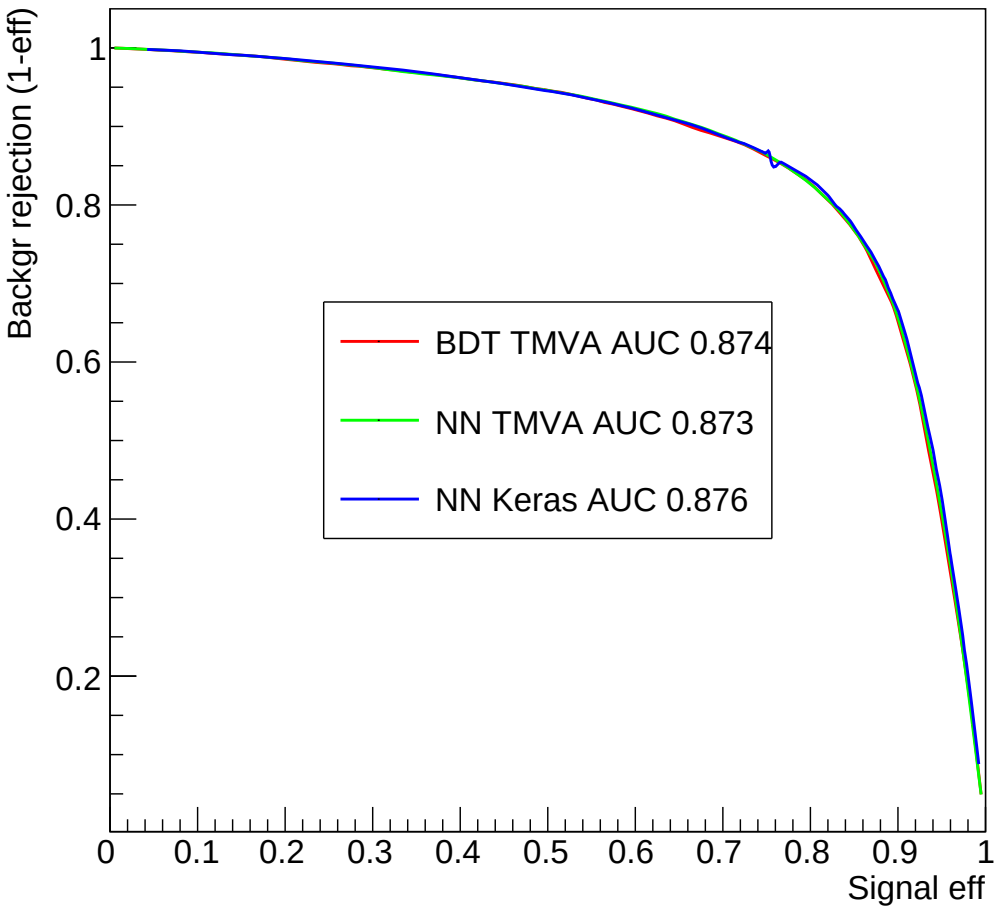


ROC curves for BDT, NN (TMVA) and NN (Keras)

Variabile Ranking - BDT

Rank	Var.	Importance
1	D_2	4.51×10^{-1}
2	cr	1.82×10^{-1}
3	$t_{\parallel a}$	6.72×10^{-2}
4	θ_{pa}	6.59×10^{-2}
5	$t_{\perp a}$	6.56×10^{-2}
6	θ_{pb}	5.90×10^{-2}
7	$t_{\parallel b}$	5.84×10^{-2}
8	$t_{\perp b}$	5.01×10^{-2}

Results Reco



ROC curves for BDT, NN (TMVA) and NN (Keras)

Variabile Ranking - BDT

Rank	Var.	Importance
1	D_2	3.29×10^{-1}
2	cr	2.29×10^{-1}
3	θ_{pb}	8.56×10^{-2}
4	$t_{\parallel b}$	8.18×10^{-2}
5	$t_{\perp b}$	7.45×10^{-2}
6	$t_{\perp a}$	6.87×10^{-2}
7	θ_{pa}	6.71×10^{-2}
8	$t_{\parallel a}$	6.38×10^{-2}

Conclusions

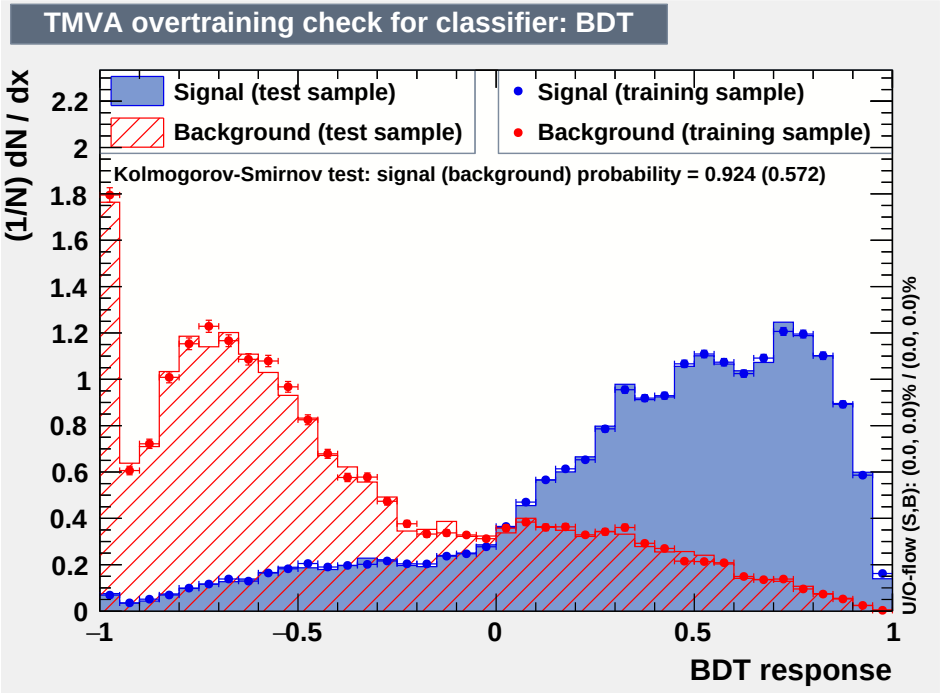
- The color-sensitive variables considered are capable of distinguishing decays from color singlets from color octects
- All ML algorithms provide excellent signal/background discirmination
- Results are comparable for Truth and Reco
- Next steps:
  - Adversarial debiasing from invariant mass dependence
  - Try using parton showers with more exact color evolution
  - Study of other possible color-sensitive variables
  - Study the interplay between these variables and others based on jet constituents (Lund-plane approach by S. Marzani and C. Kaur)

	AUC - Test Sample	
	Truth	Reco
BDT	0.879	0.874
NN - TMVA	0.879	0.873
NN - Keras	0.880	0.876

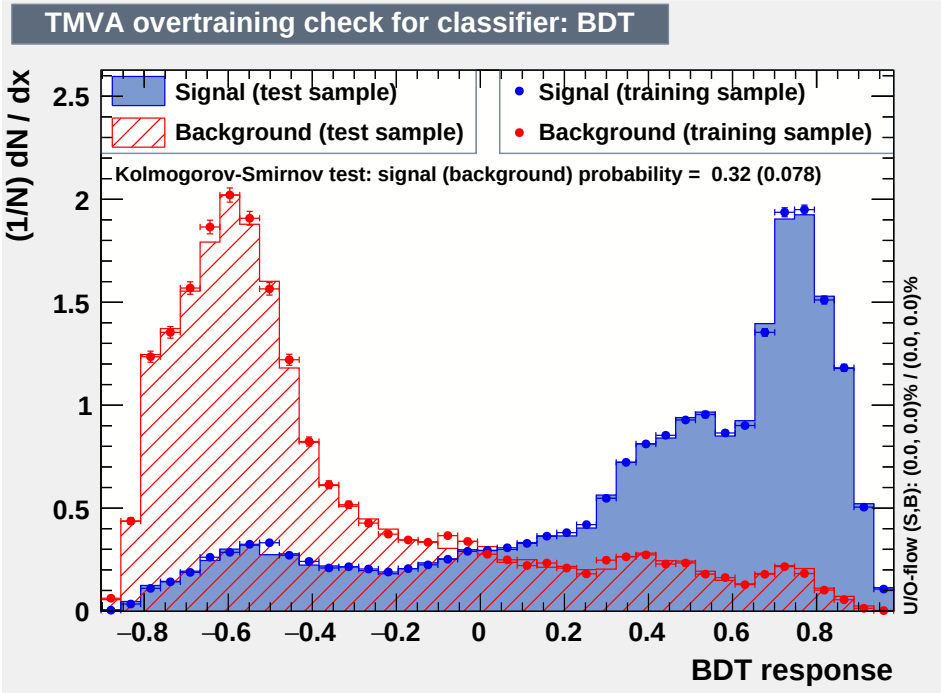
Backup

BDT - Overtraining Check

Truth

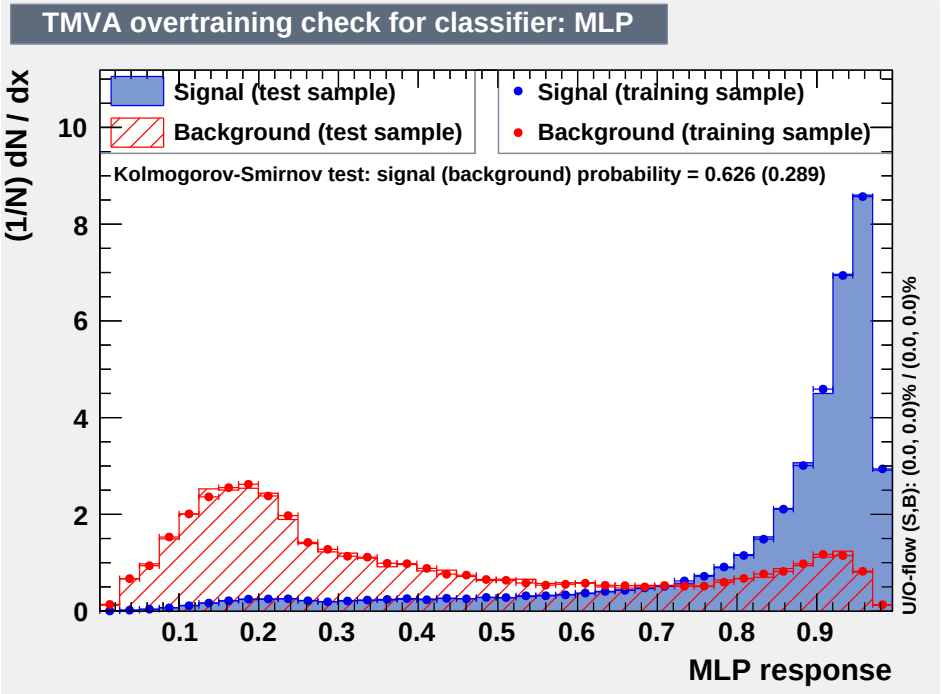


Reco

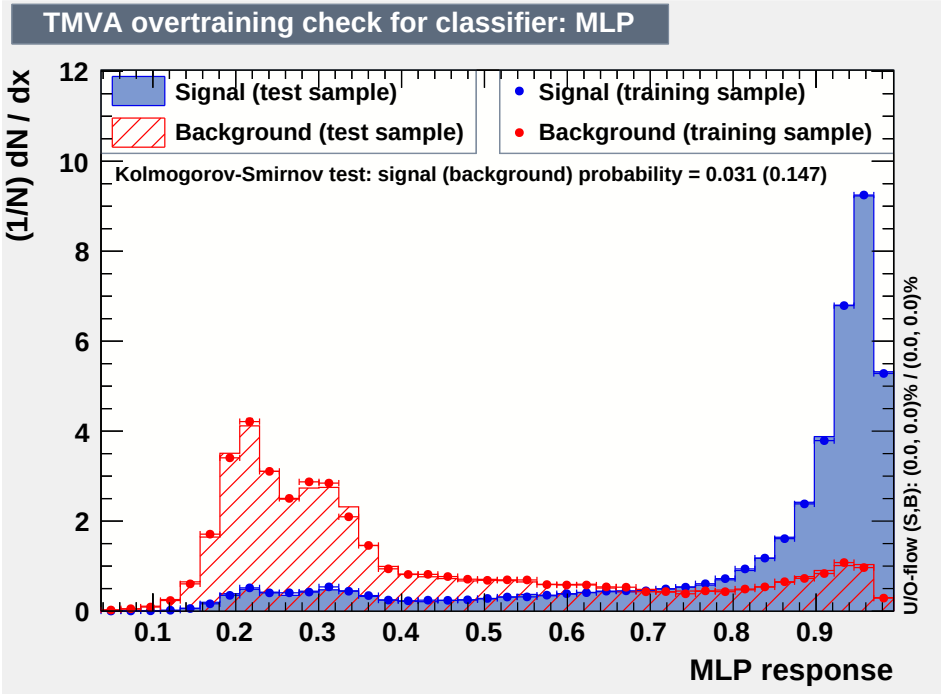


NN TMVA - Overtraining Check

Truth

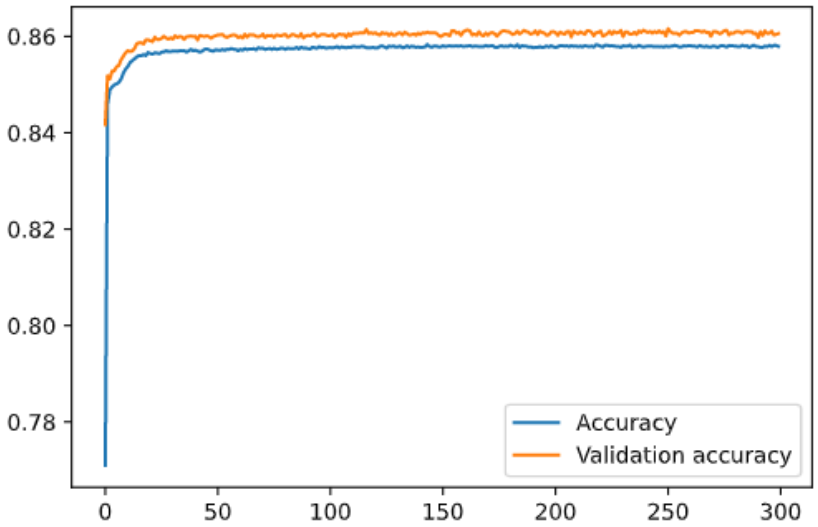


Reco

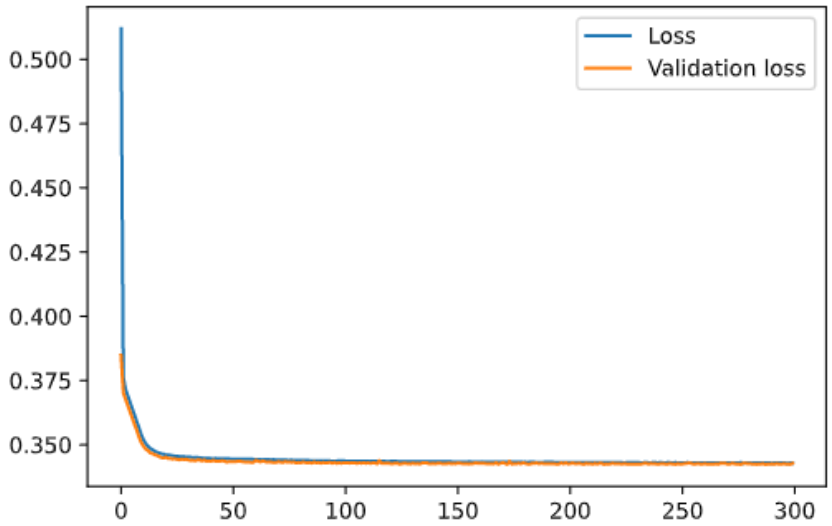


Accuracy & Loss - Keras NN (Truth)

Accuracy

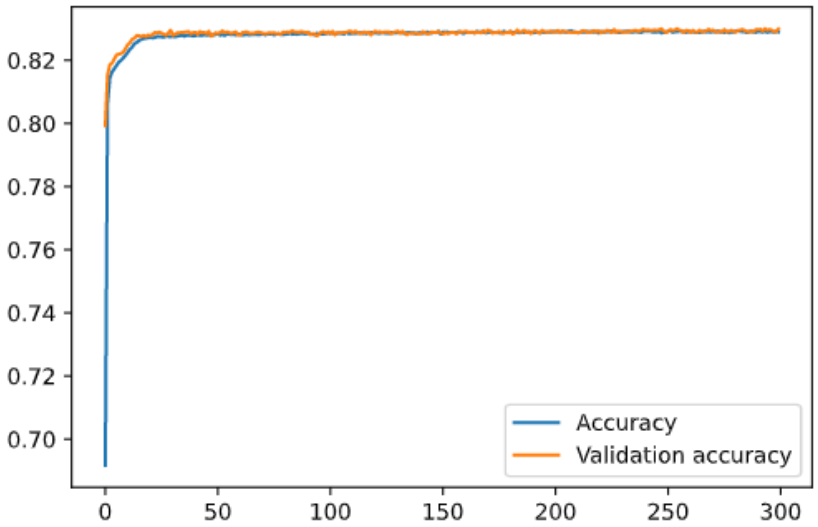


Loss

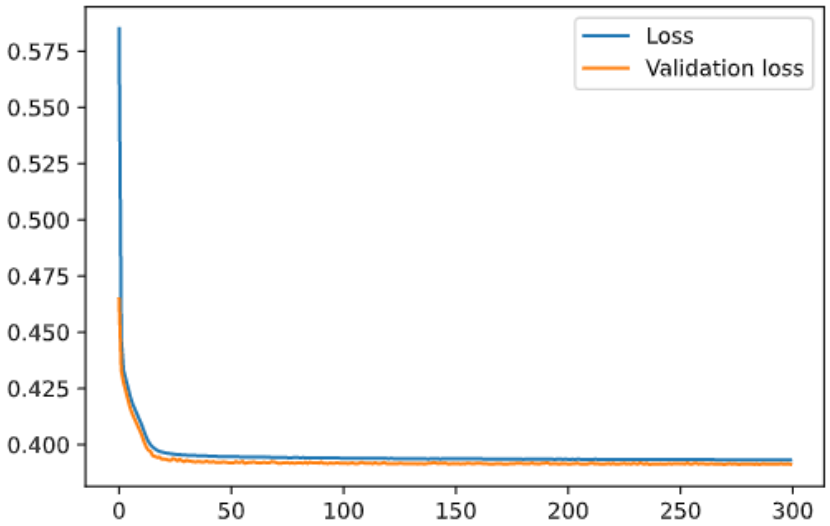


Accuracy & Loss - Keras NN (Reco)

Accuracy

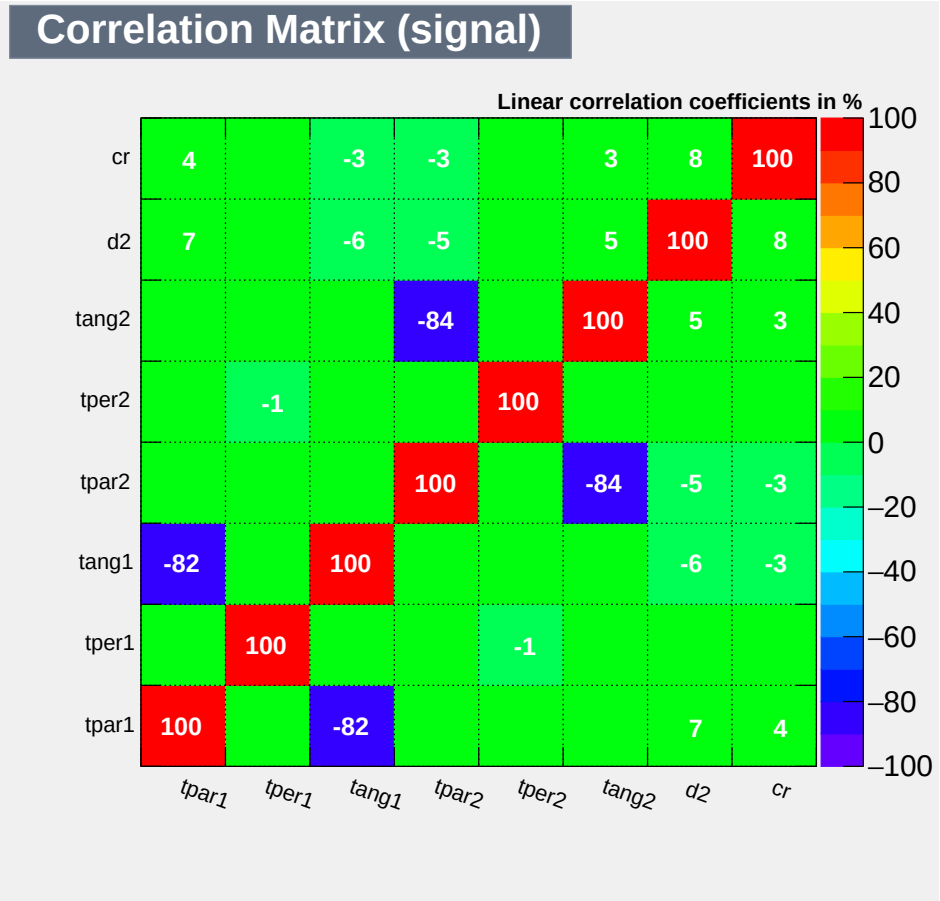


Loss

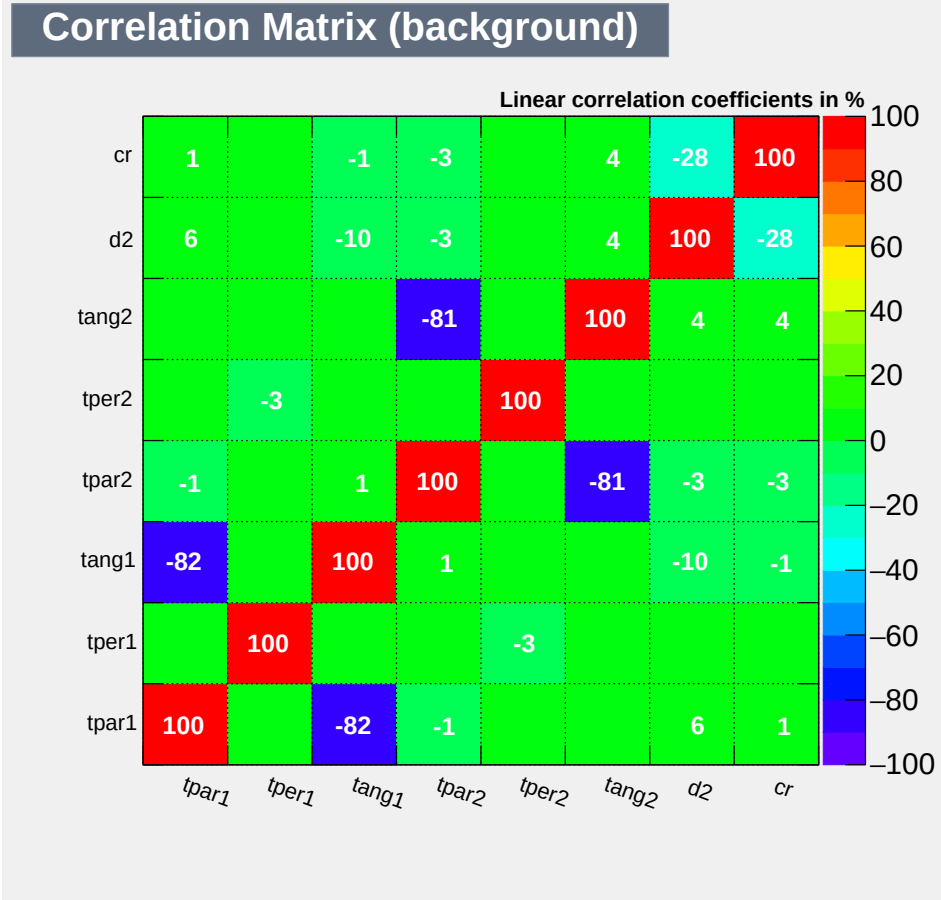


Correlation Matrix - Reco

ZHbb

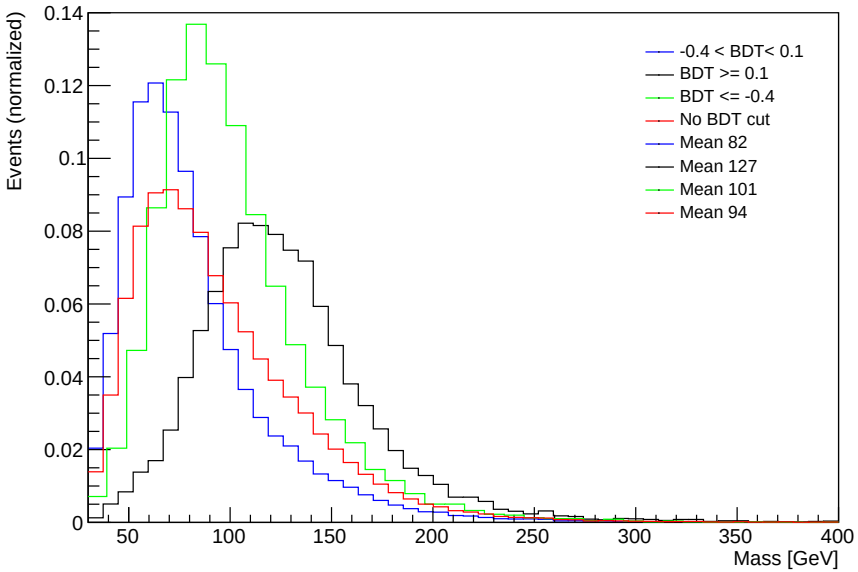


gbb

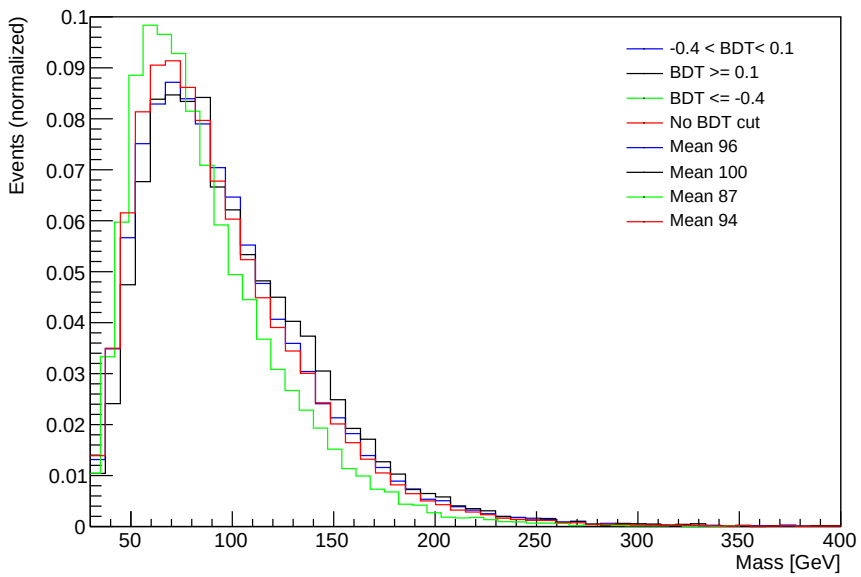


Mass Bias - Truth

All Variables



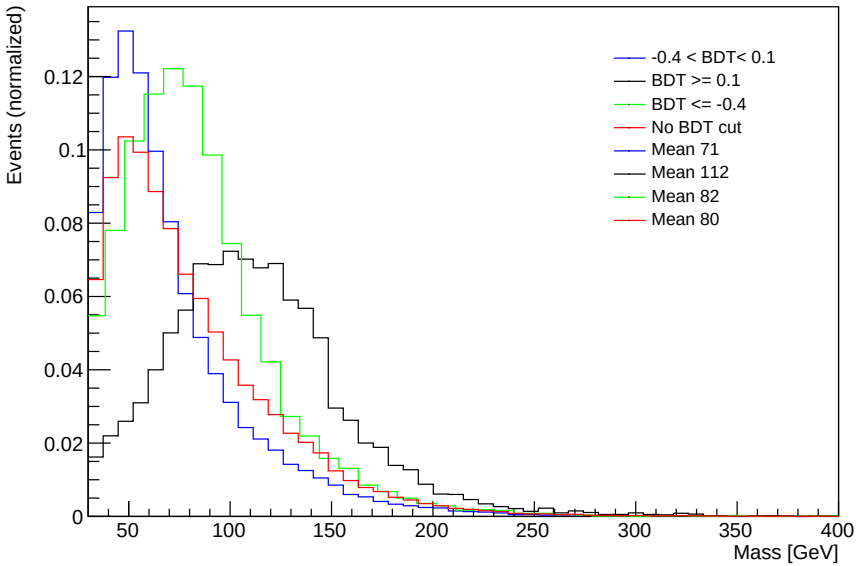
No D_2



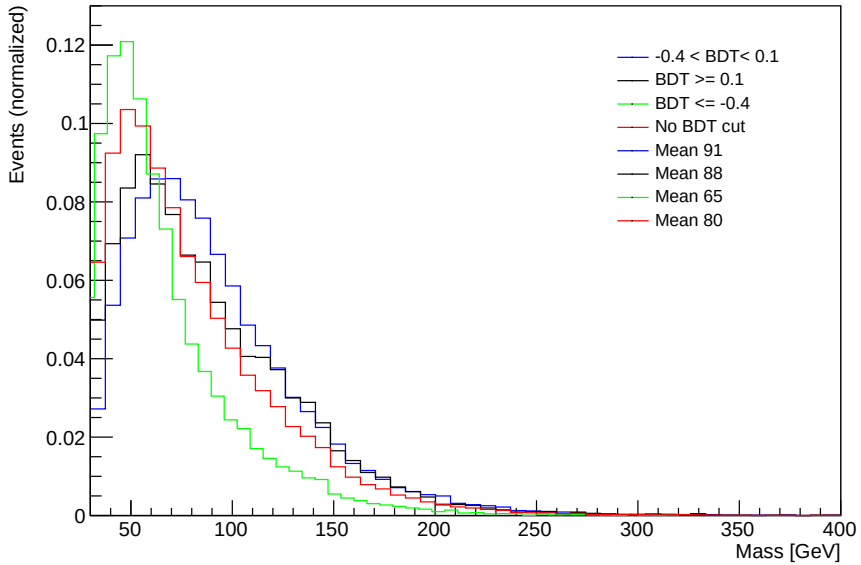
	AUC w/o D_2
BDT	0.775
NN - TMVA	0.773

Mass Bias - Reco

All Variables



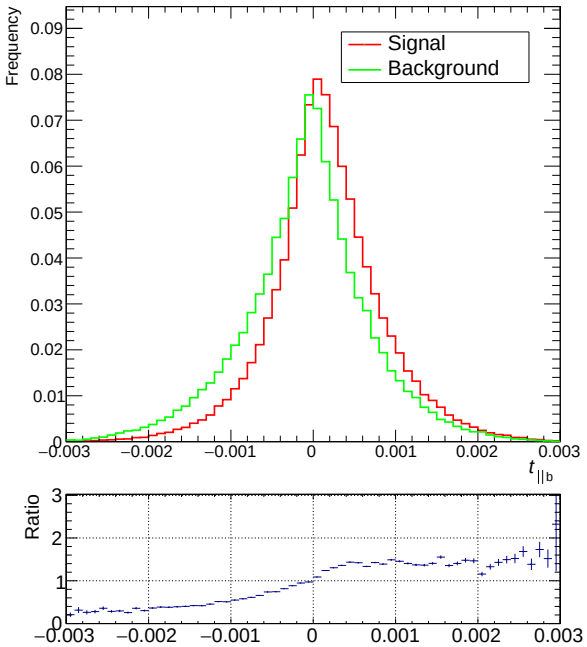
No D_2



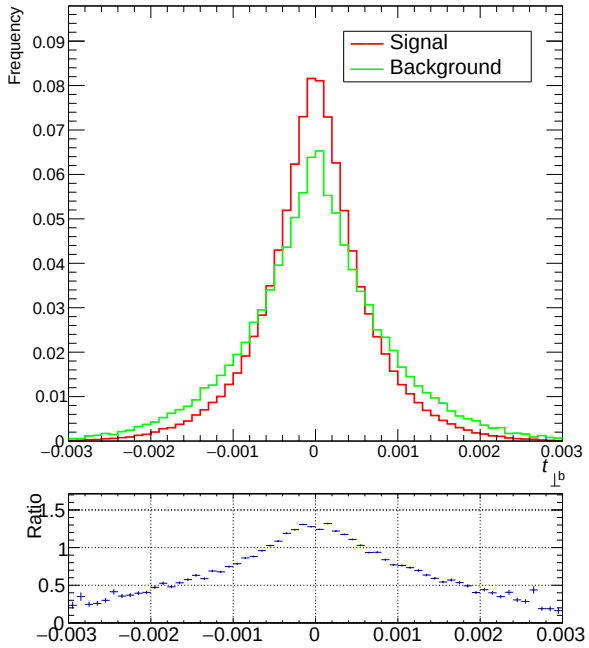
	AUC w/o D_2
BDT	0.773
NN - TMVA	0.772

Signal vs. Bkg (Truth) - Jet b Pull Variables

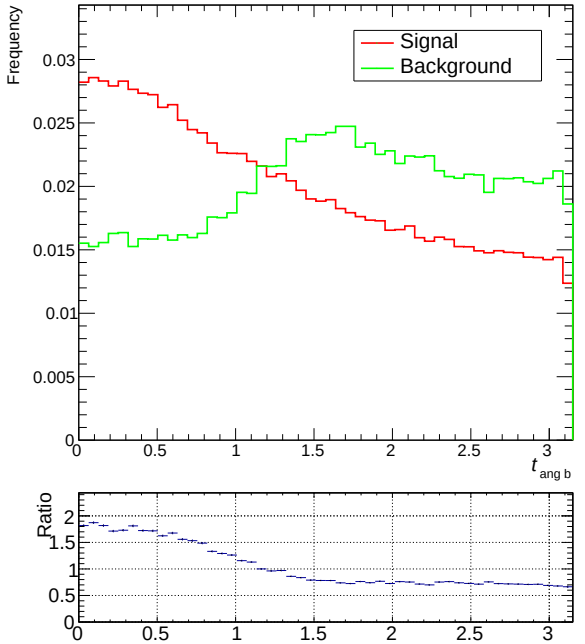
$t_{||b}$



$t_{\perp b}$

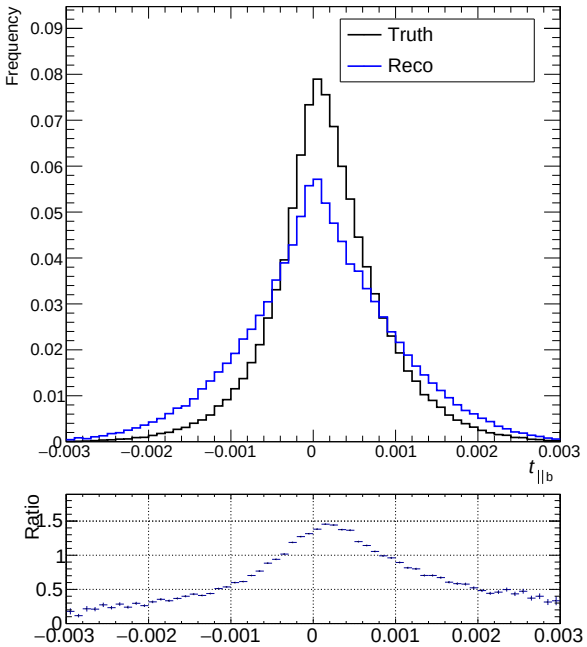


θ_{pb}

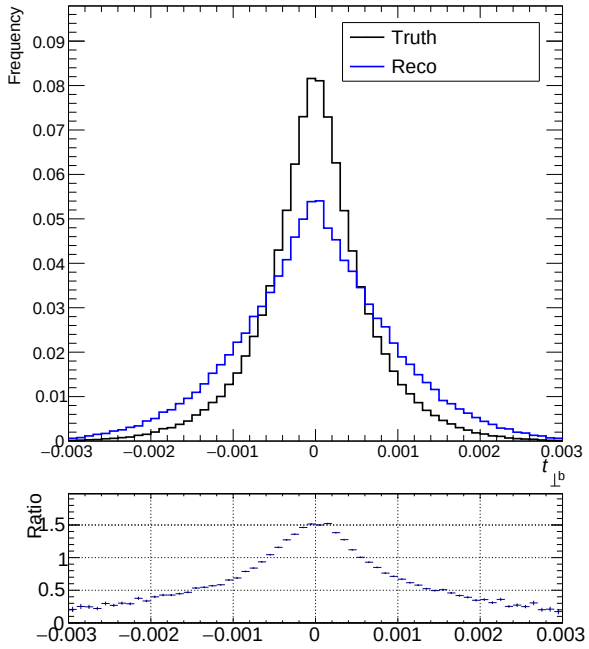


Truth vs. Reco (ZHbb) - Jet b Pull Variables

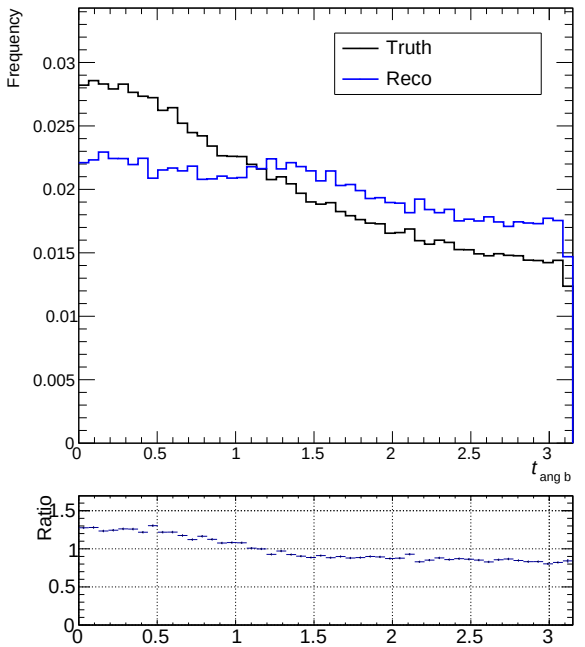
$t_{\| b}$



$t_{\perp b}$

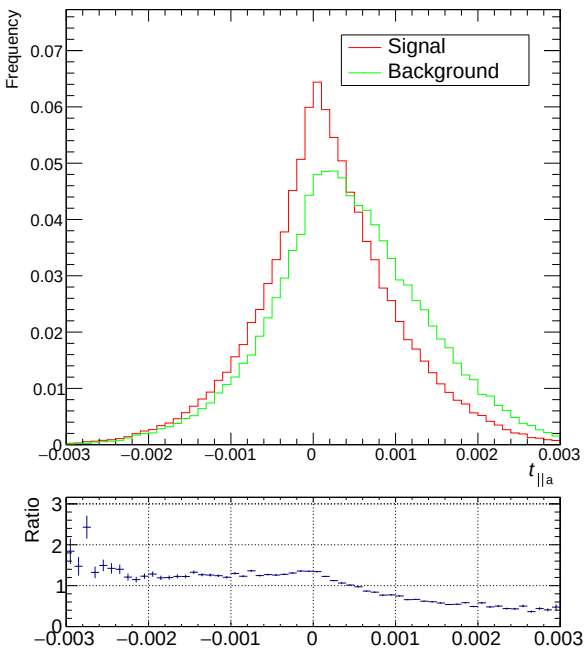


θ_{pb}

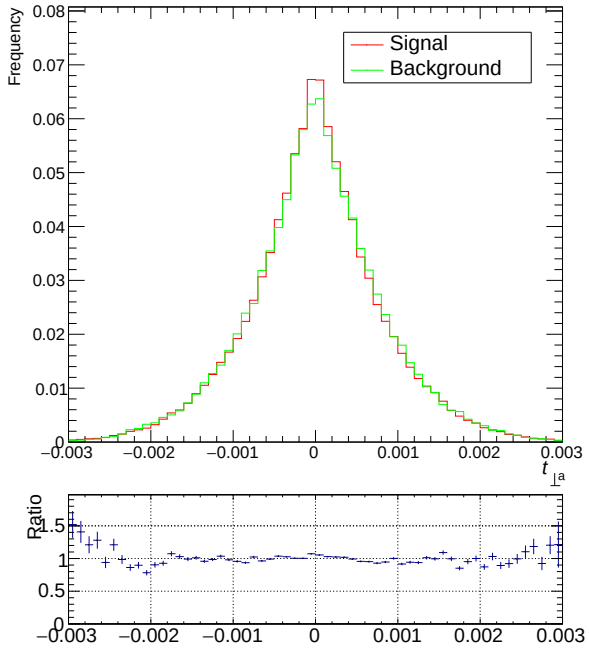


Distributions - Signal vs. Bkg (Reco)

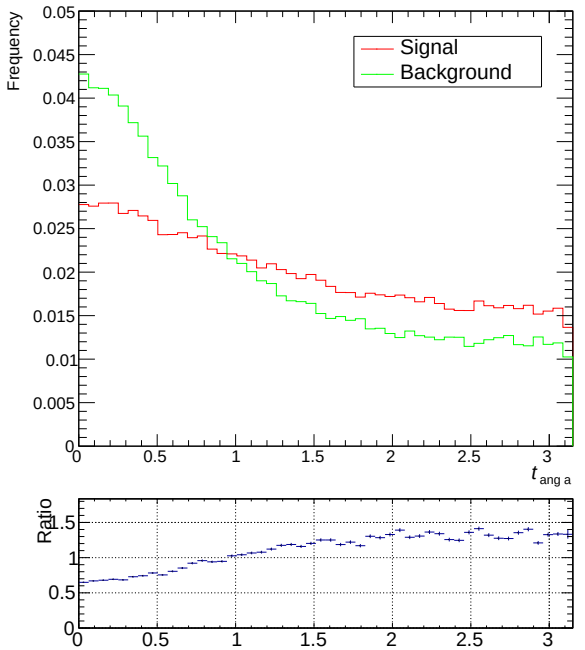
$t_{\parallel a}$



$t_{\perp a}$

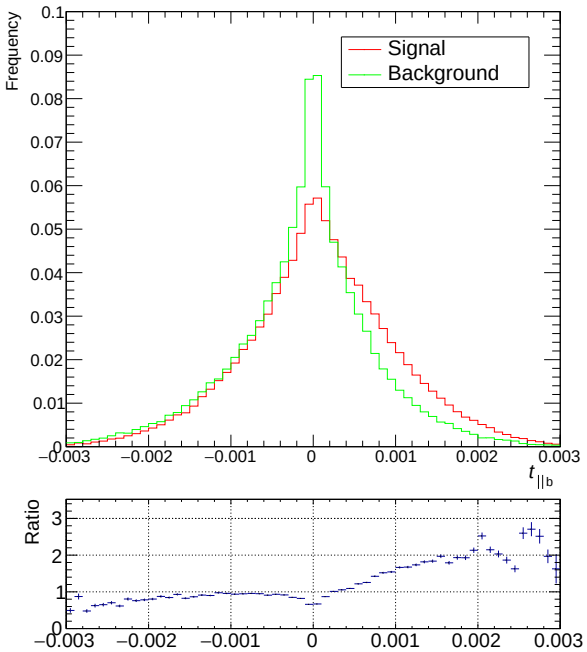


θ_{pa}

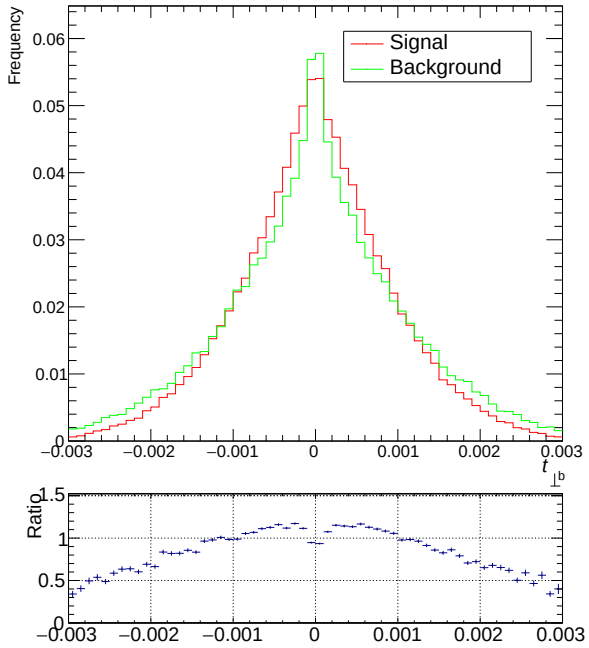


Distributions - Signal vs. Bkg (Reco)

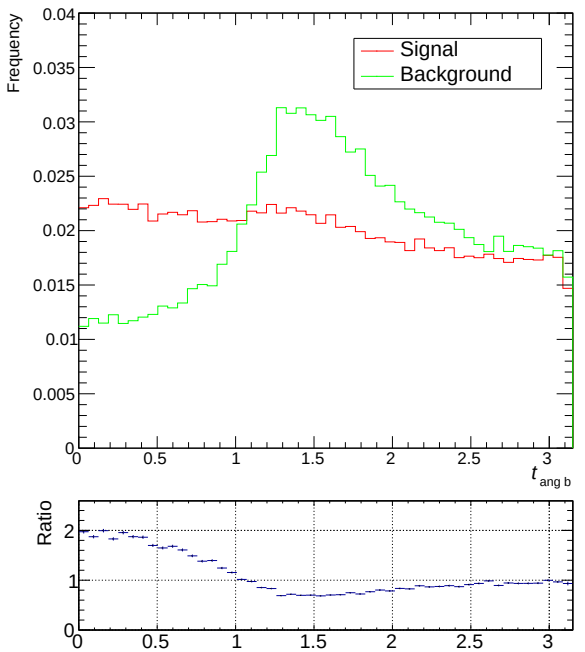
$t_{\| b}$



$t_{\perp b}$

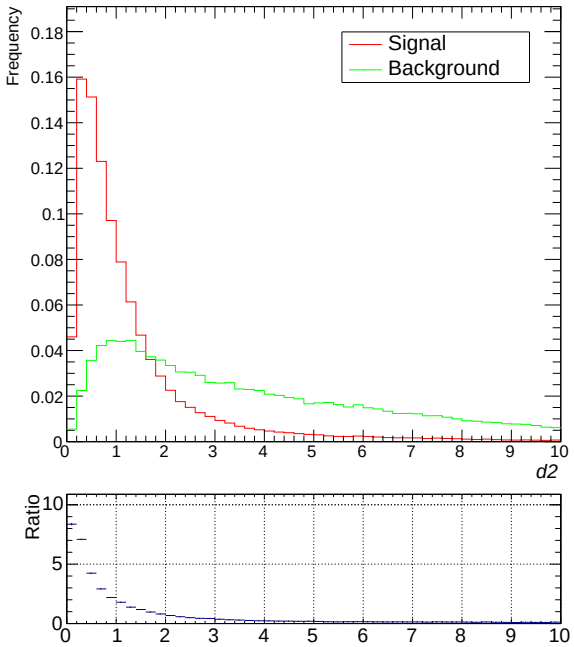


θ_{pb}

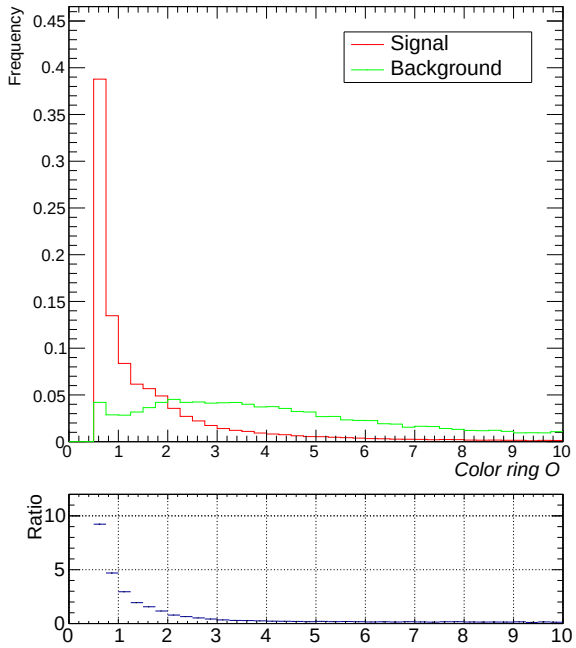


Distributions - Signal vs. Bkg (Reco)

D_2

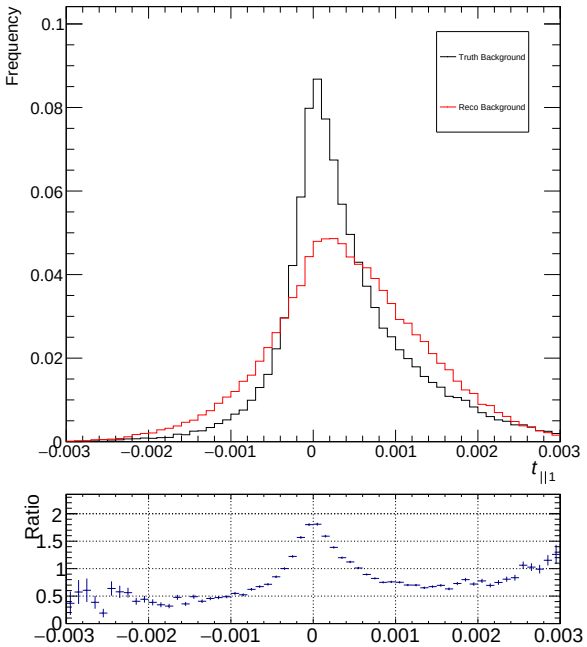


cr

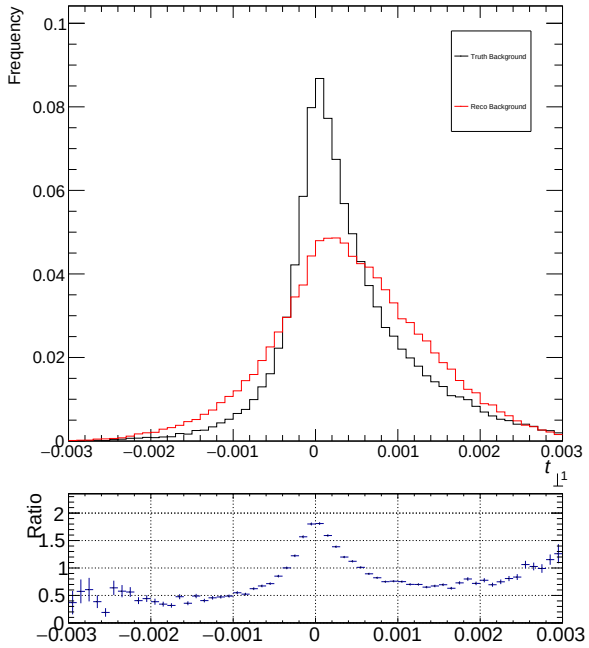


Distributions - Truth vs. Reco (gbb)

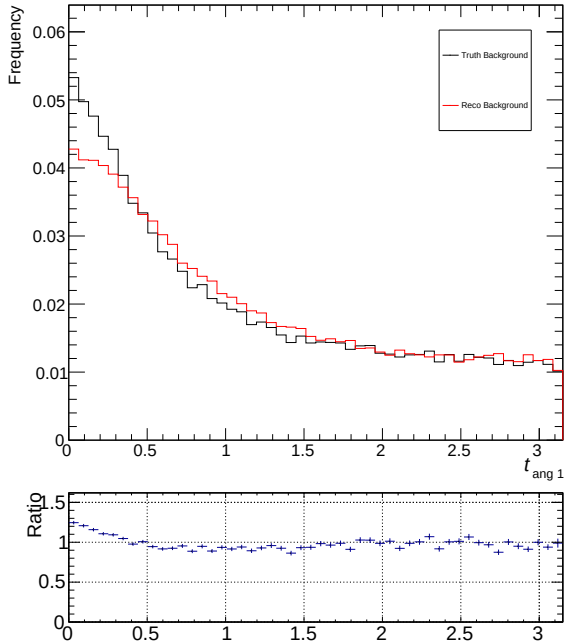
$t_{\parallel a}$



$t_{\perp a}$

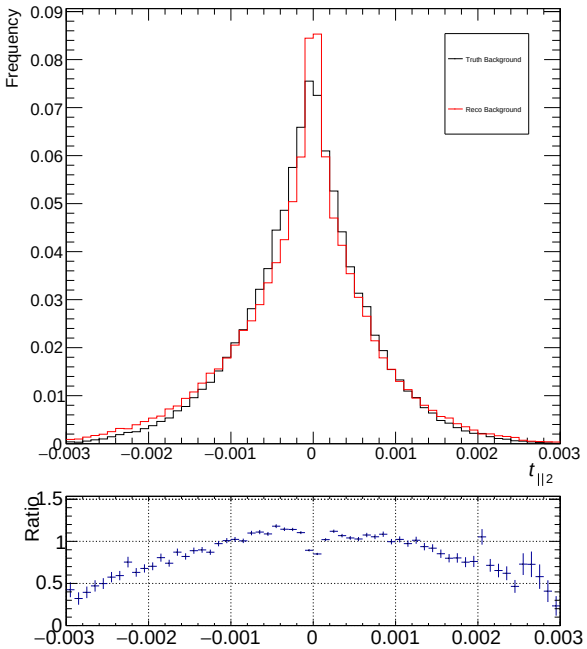


θ_{pa}

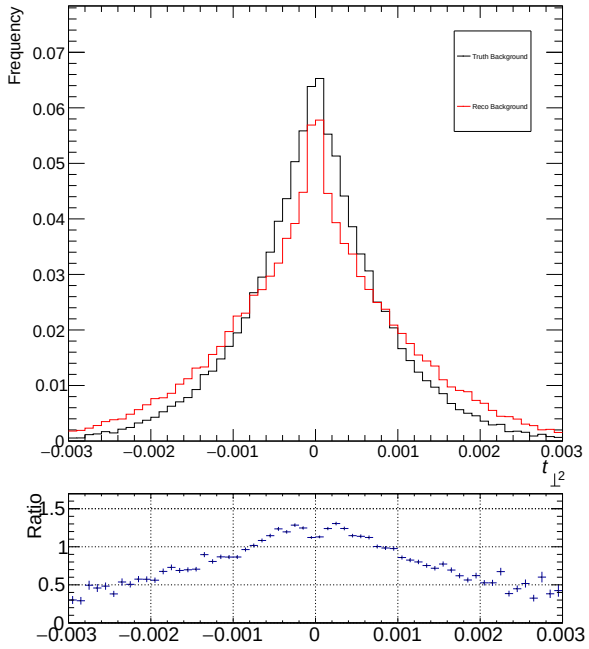


Distributions - Truth vs. Reco (gbb)

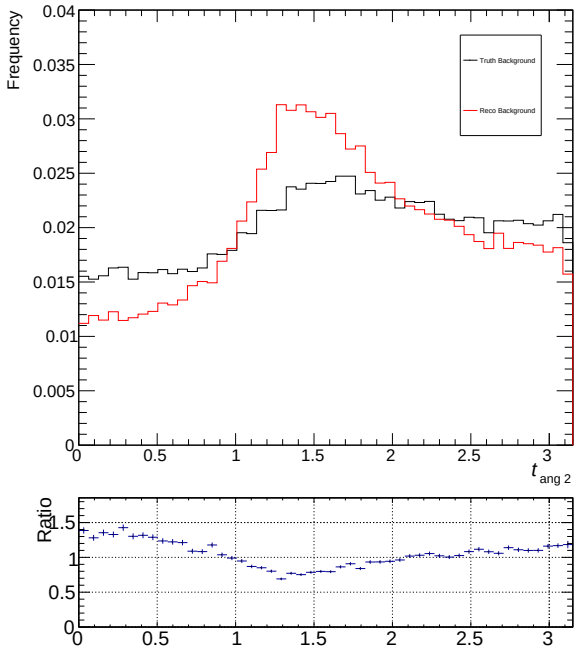
$t_{||b}$



$t_{\perp b}$

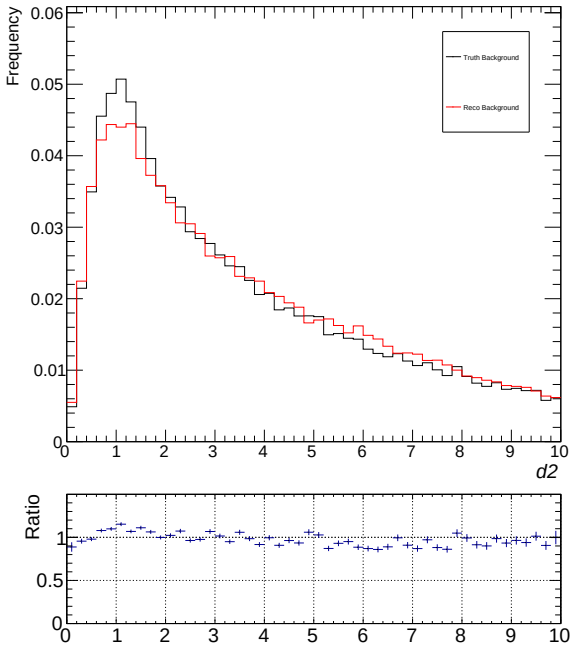


θ_{pb}

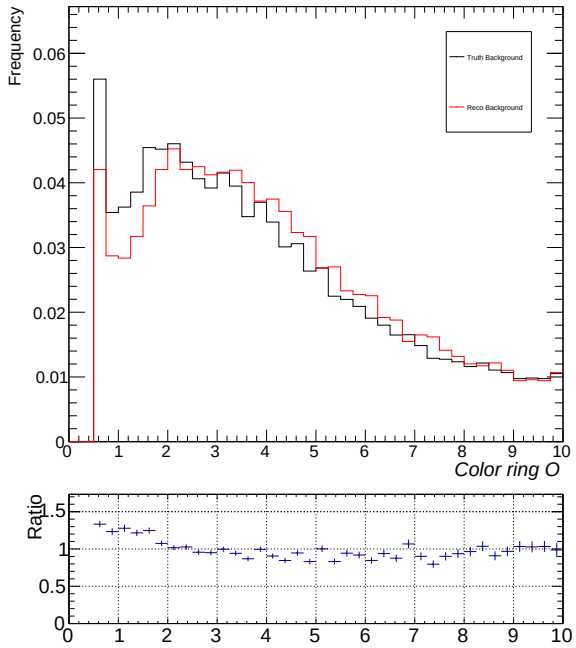


Distributions - Truth vs. Reco (gbb)

D_2

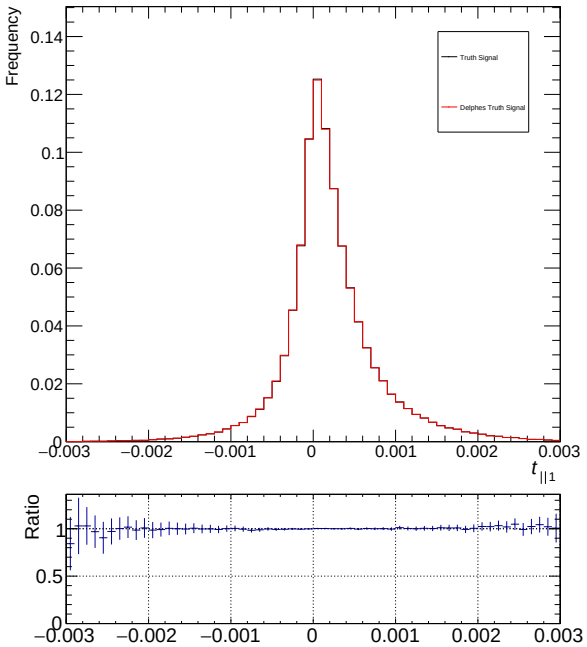


cr

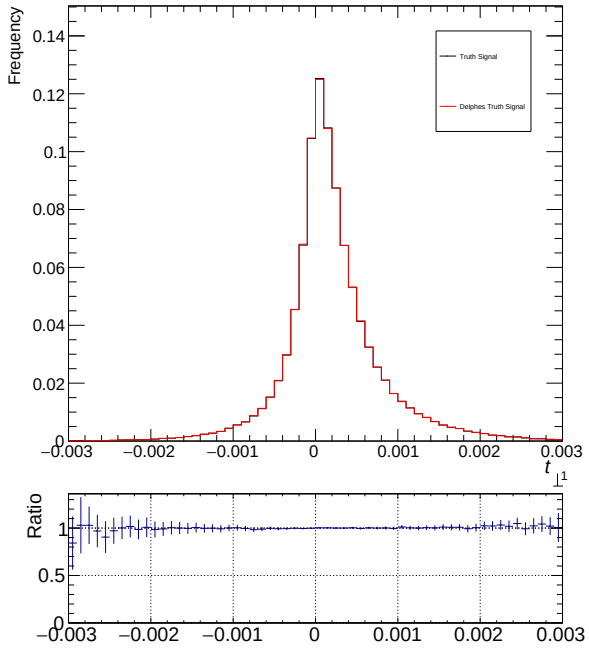


Distributions - Pythia vs. Delphes Truth (ZHbb)

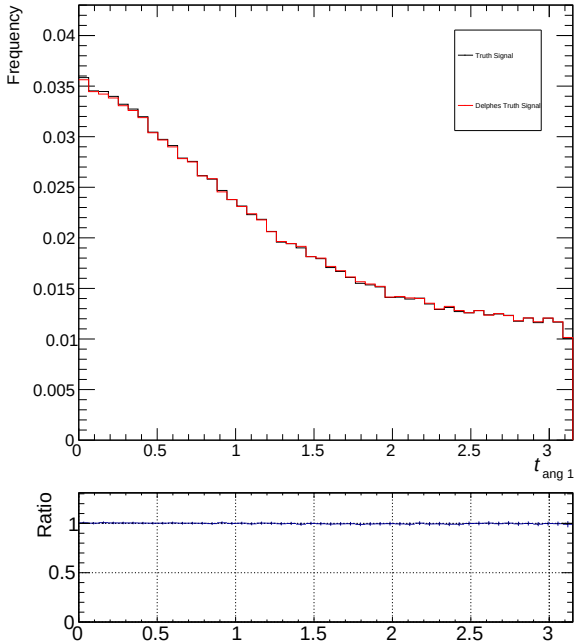
$t_{\parallel a}$



$t_{\perp a}$

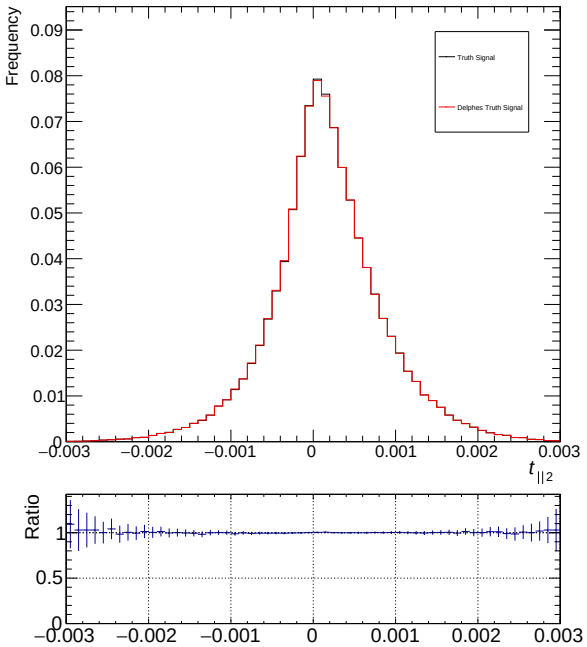


θ_{pa}

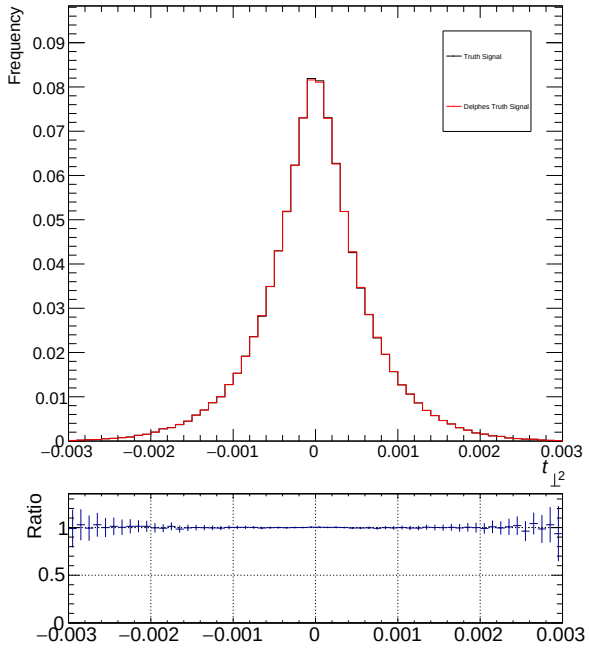


Distributions - Pythia vs. Delphes Truth (ZHbb)

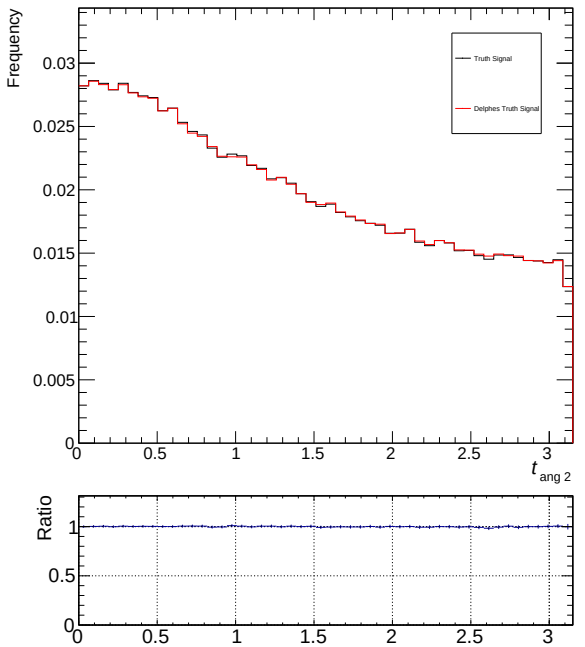
$t_{||b}$



$t_{\perp b}$

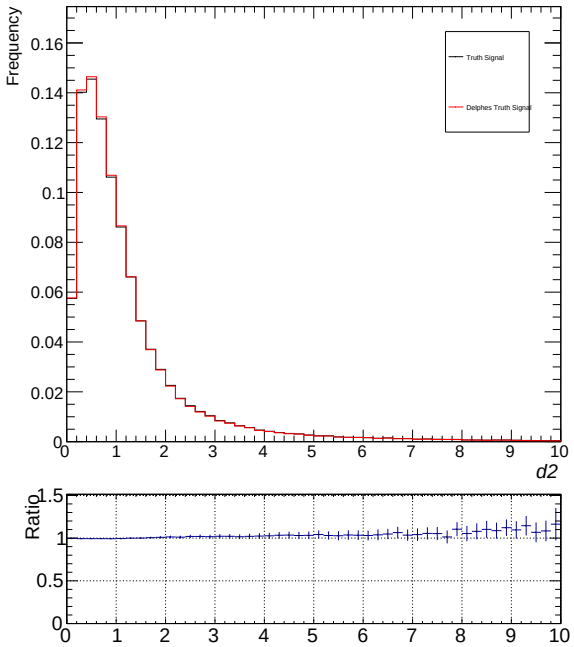


θ_{pb}

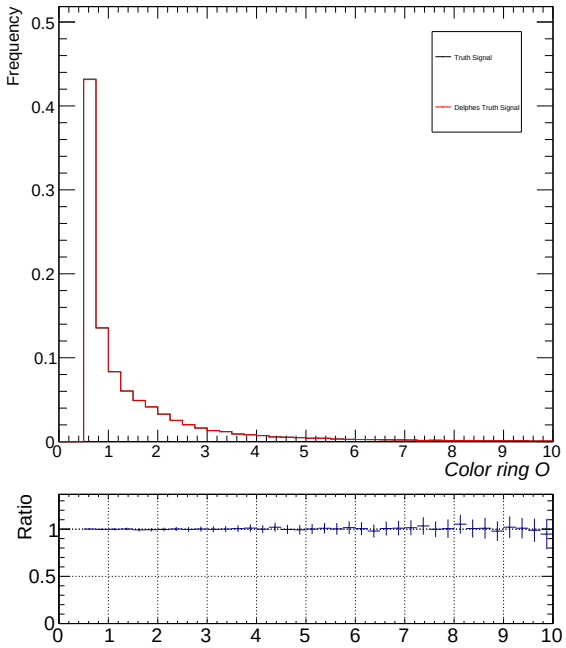


Distributions - Pythia vs. Delphes Truth (ZHbb)

D_2

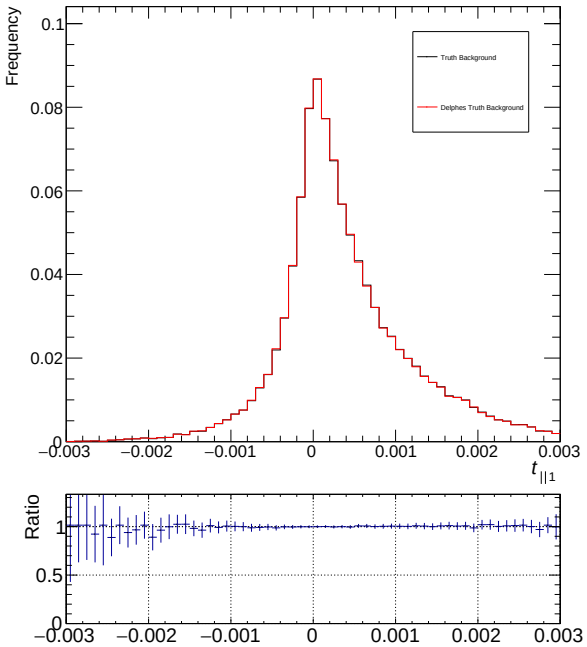


cr

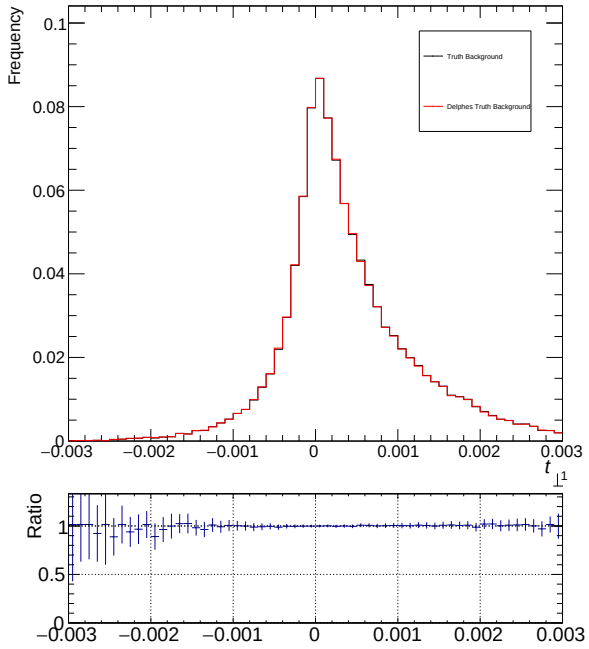


Distributions - Pythia vs. Delphes Truth (gbb)

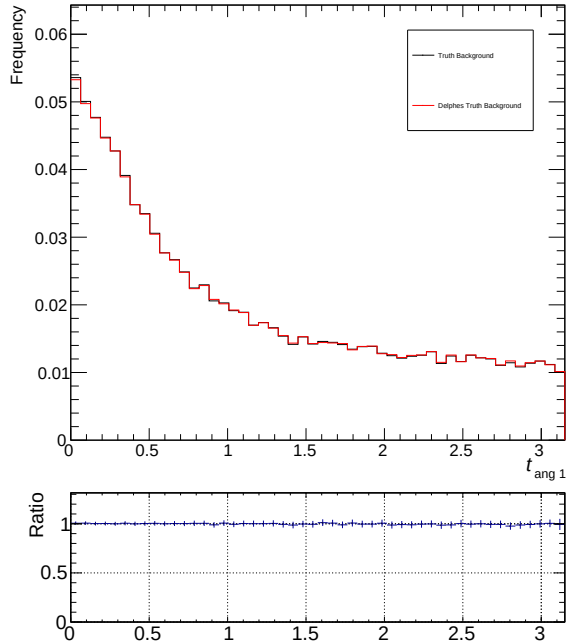
$t_{||a}$



$t_{\perp a}$

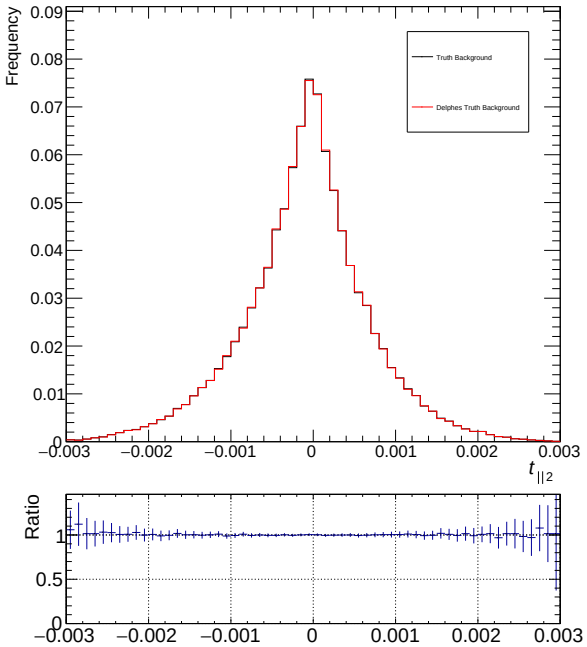


θ_{pa}

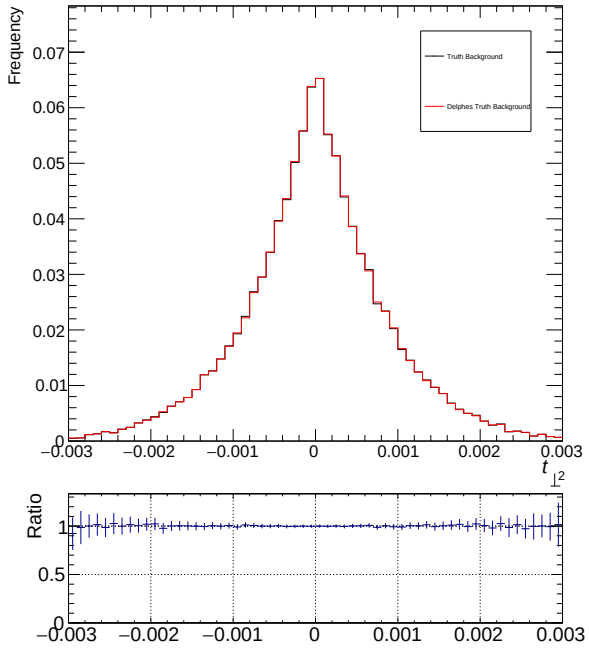


Distributions - Pythia vs. Delphes Truth (gbb)

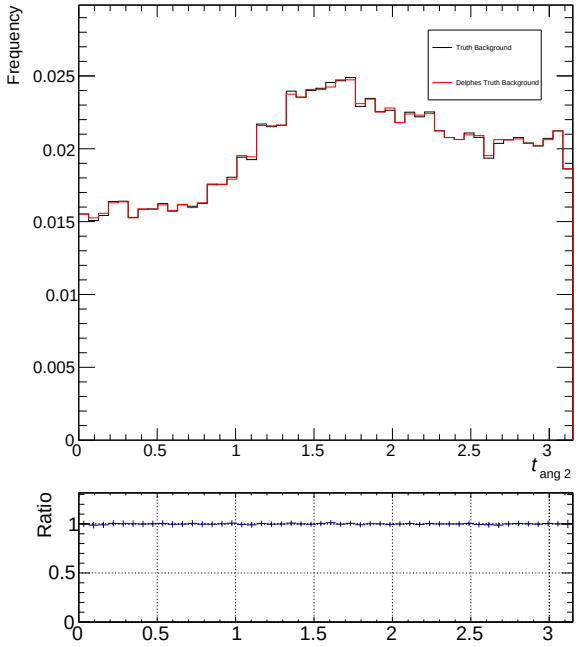
$t_{||b}$



$t_{\perp b}$

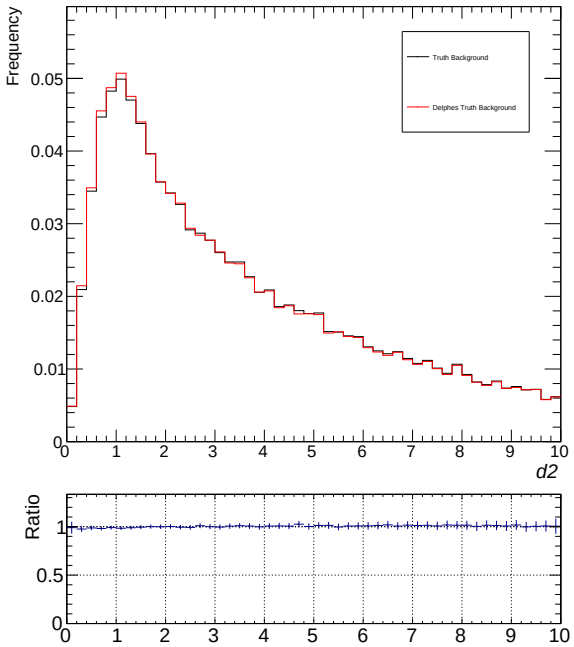


θ_{pb}



Distributions - Pythia vs. Delphes Truth (gbb)

D_2



cr