Bootstrap and Feature Importance Studies for non-Resonant bbyy BDT Analysis

Jay Chan Alkaid Cheng Alex Wang
Sau Lan Wu Rui Zhang Chen Zhou

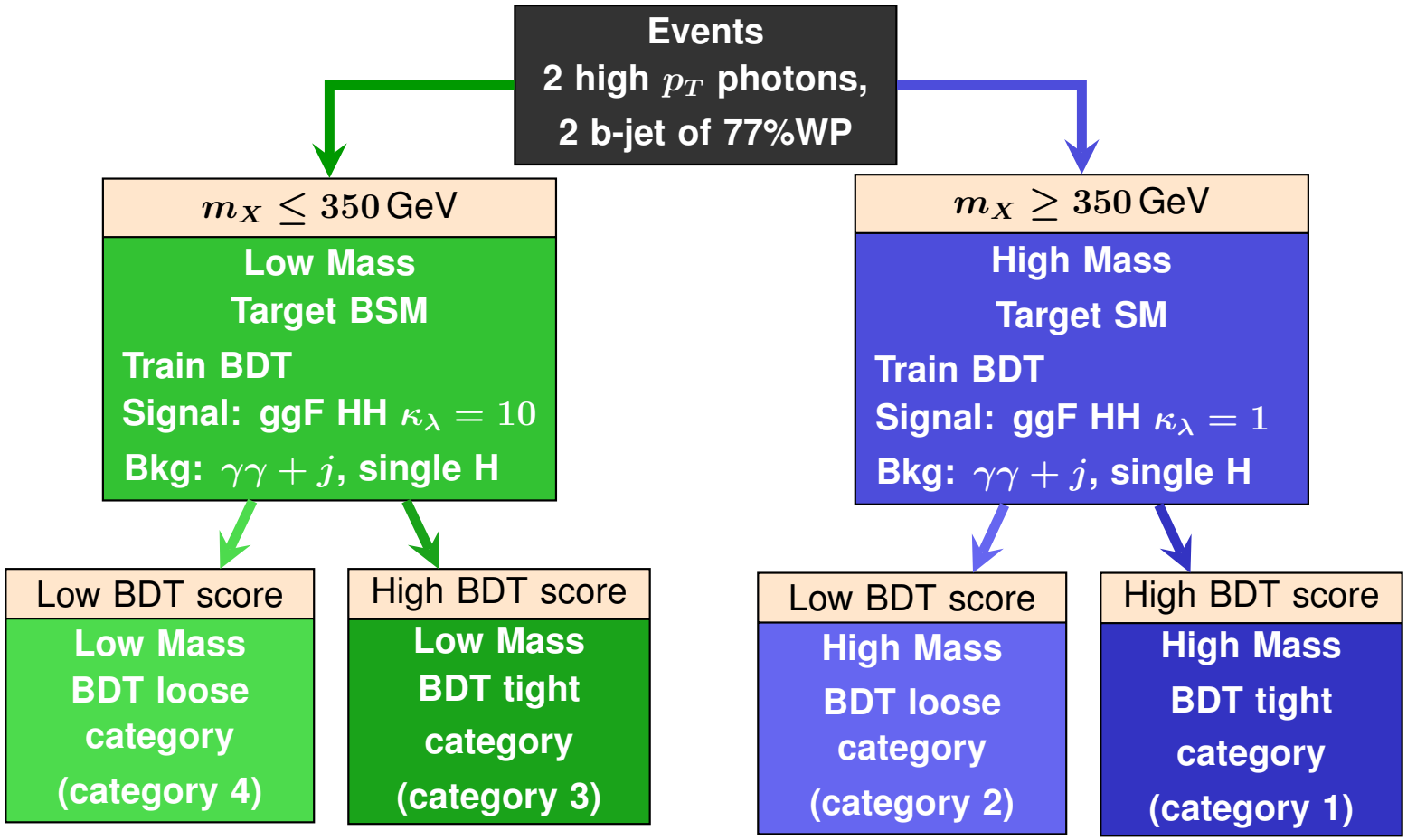
University of Wisconsin-Madison

October 7, 2020

Contents

- Review of BDT categorization
- Bootstrap Study
  - Evaluation of significance
  - Principle of the Poisson bootstrap
  - Results
- Feature Importance Study
  - Permutation Feature Importance (PFI)
  - Leave-One-Covariate-Out (LOCO) method
  - SHapley Additive exPlanations (SHAP) values

Review of BDT Categorization



- After BDT training, each event is assigned a BDT score from the BDT model and categorized acoording to the above categorization scheme.

Bootstrap Study - Introduction

- This presentation is for validation of the BDTs in h026 MxAOD
- Previous talk on hyperparameter optimization (HPO) can be found here
- From the HPO, a new BDT model with improved significance is obtained. In this talk, we want to answer the questions
  - Is the improvement in significance significant? (Suggested by Marc)
  - How will the significance fluctuate for a different data sample?
- The goal is to compute the error on the relative significance improvement $Z_{\text{imp}} = (Z' - Z)/Z$.
- This can be done using the bootstrap resampling method. A related talk that uses this method can be found here

Evaluation of Significance

- For each category c, the signal yield s_c and background yield b_c are calculated by summing over the event weights in that category.

$$\begin{aligned} s_c &= \sum_i w_{\text{sig},c,i} & b_c &= \sum_i w_{\text{bkg},c,i} \\ \delta s_c &= \sqrt{\sum_i w^2_{\text{sig},c,i}} & \delta b_c &= \sqrt{\sum_i w^2_{\text{bkg},c,i}} \end{aligned} \tag{1}$$

- The Asimov estimate for significance in each category is calculated by

$$\begin{aligned} Z_c &= \sqrt{2[(s_c + b_c)\ell - s_c]}, \quad \text{where } \ell = \ln(1 + s_c/b_c) \\ \delta Z_c &= \sqrt{(\ell/Z_c)^2(\delta s_c)^2 + [(\ell - s_c/b_c)/Z_c]^2(\delta b_c)^2} \end{aligned} \tag{2}$$

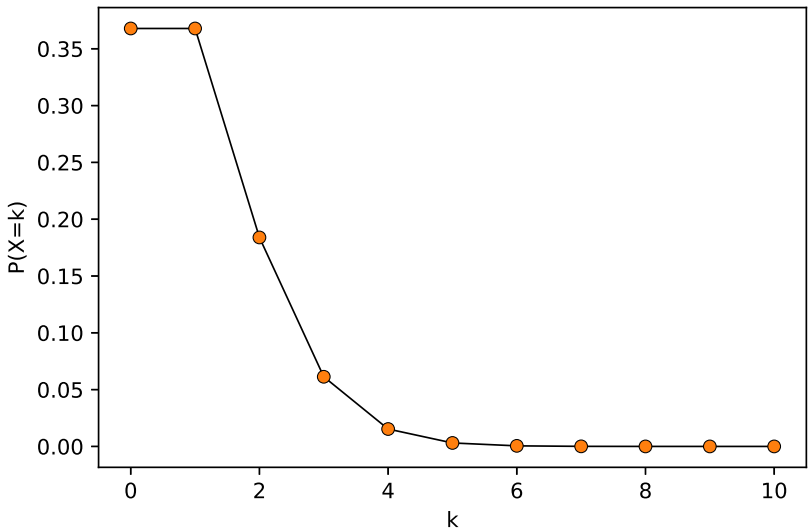
- The combined significance is

$$Z = \sqrt{\sum_c Z_c^2}, \qquad \delta Z = \sqrt{\sum_c (\delta Z_c)^2} \tag{3}$$

Note that δZ from error propagation cannot be used to compare the significance between two different trainings because of correlated events.

Poisson Bootstrap: Methodology

- In the Poisson bootstrap method, a set of replicas of the event weight distribution is generated from the original data sample by assigning each event an extra random weight distributed according to the Poisson distribution with $\mu = 1$. This Poisson weight is analogous to how many times the event from the original sample is redrawn to the new sample.
- A Poisson distribution with $\mu = 1$ has the following shape

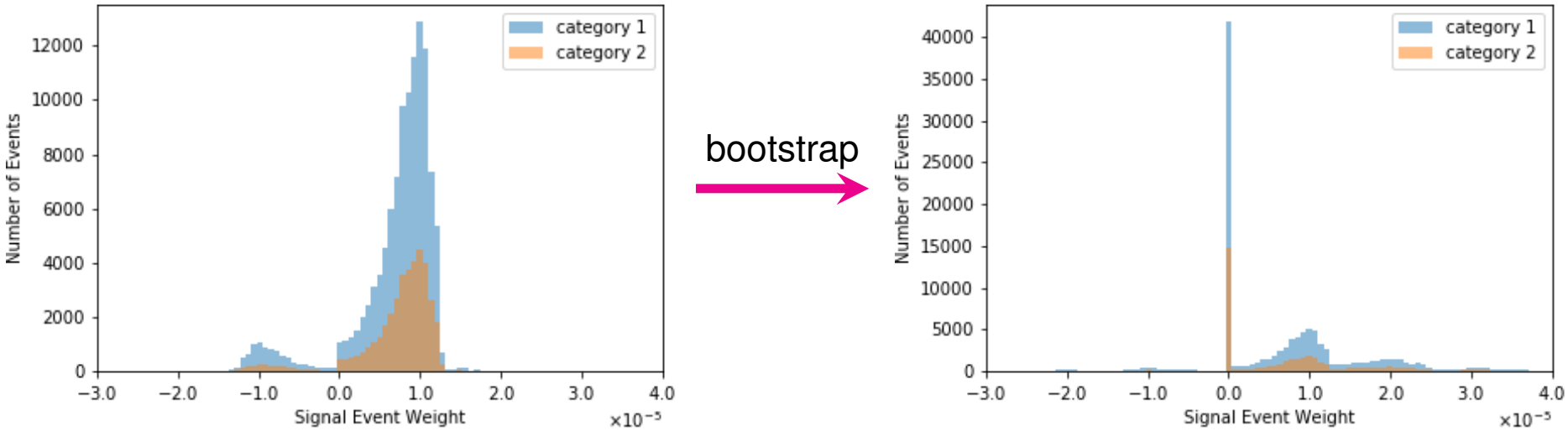


So each event will have a 36.8% chance to be drawn 0 times, 36.8% chance to be drawn once, 18.4% chance to be drawn twice and so on.

Poisson Bootstrap: Examples I

Take an example of the BDT model before HPO in the SM case (i.e. using lambda = 01 signal):

- ✦ Signal weight distributions



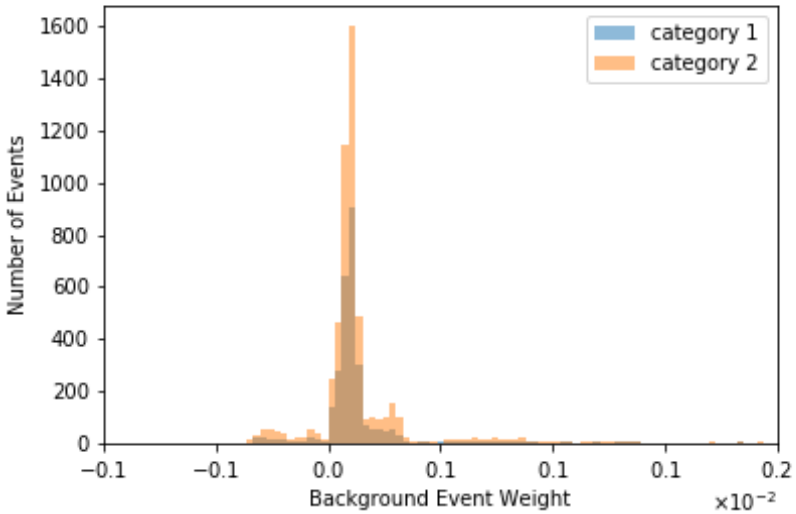
(a) Original sample (b) Bootstrapped sample

$$s_c = (0.789, 0.285) \xrightarrow{\text{bootstrap}} (0.796, 0.282) \tag{4}$$

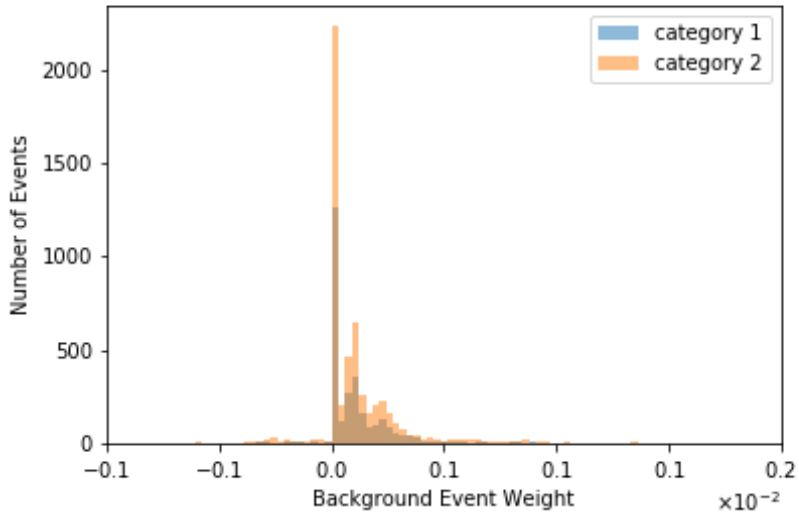
Note: The low mass categories (categories 3 and 4) are not included in the SM case because they have low sensitivity

Poisson Bootstrap: Examples II

- Background weight distributions



bootstrap



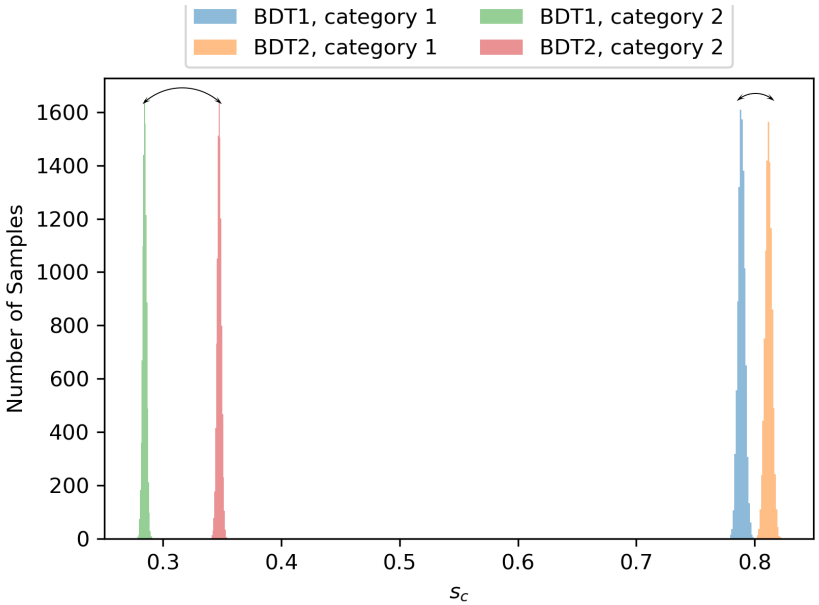
(a) Original sample (b) Bootstrapped sample

$$b_c = (2.716, 5.197) \xrightarrow{\text{bootstrap}} (2.493, 5.189)$$
$$Z_c = (0.458, 0.124) \xrightarrow{\text{bootstrap}} (0.481, 0.123) \quad (5)$$

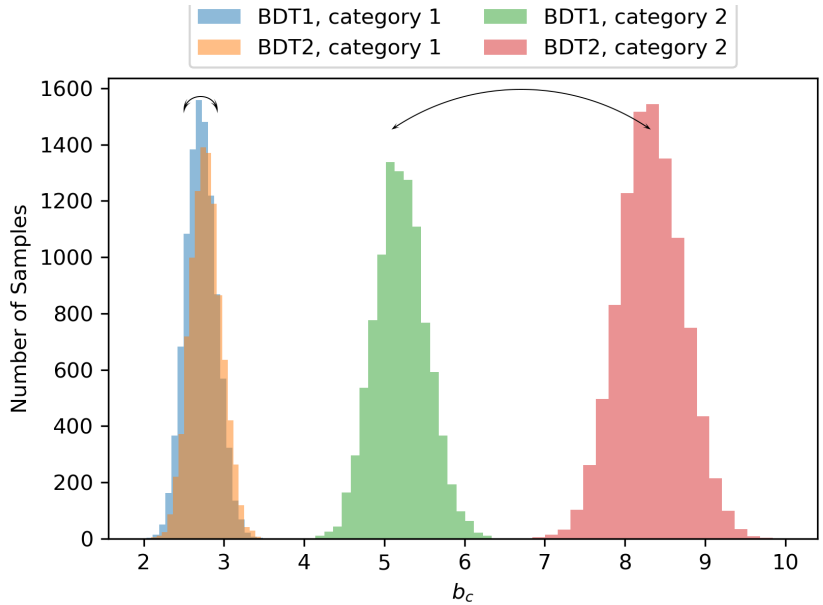
- By repeating this procedure, a distribution of the significance and its variance can be obtained.

Poisson Bootstrap: Results for SM case I

- A total of $N = 10,000$ bootstrapped samples are produced for both **BDT models before and after HPO (= BDT 1 and BDT 2)**. In each sample, the same Poisson weights are applied for both BDT models.
- In the SM case (i.e. using lambda = 01 signal), the resulting distributions of the signal and background yields for each category is shown below



(a) Signal Yields

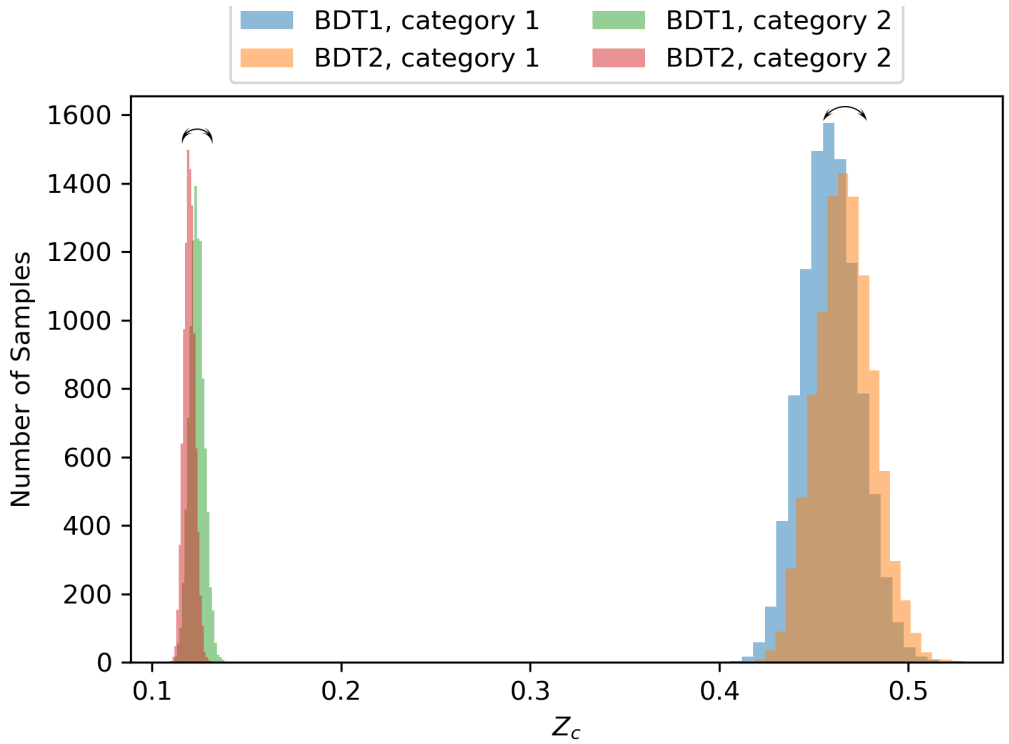


(b) Background Yields

- Distributions joined by arrows are from the same category but different BDT models.
- The shift in s/b values in each category for different BDT models is due to the change in the optimal BDT boundaries.

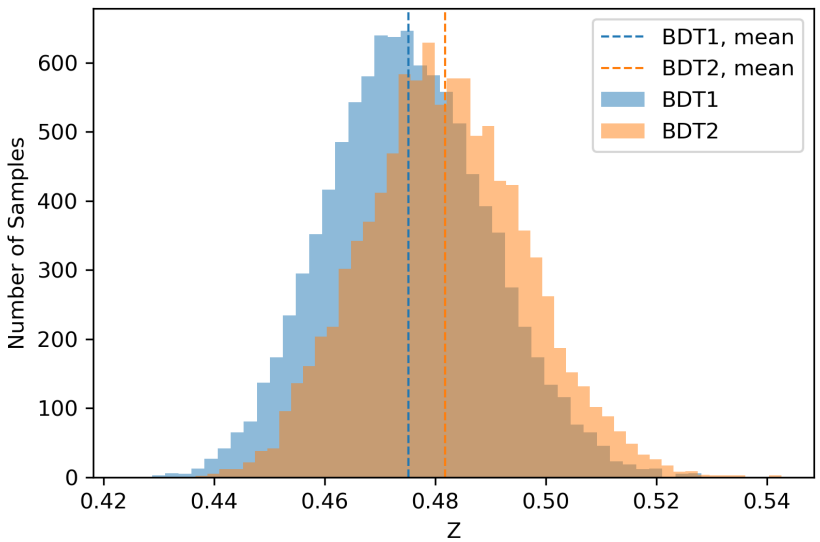
Poisson Bootstrap: Results for SM case II

- The resulting distribution of the significance for each category is

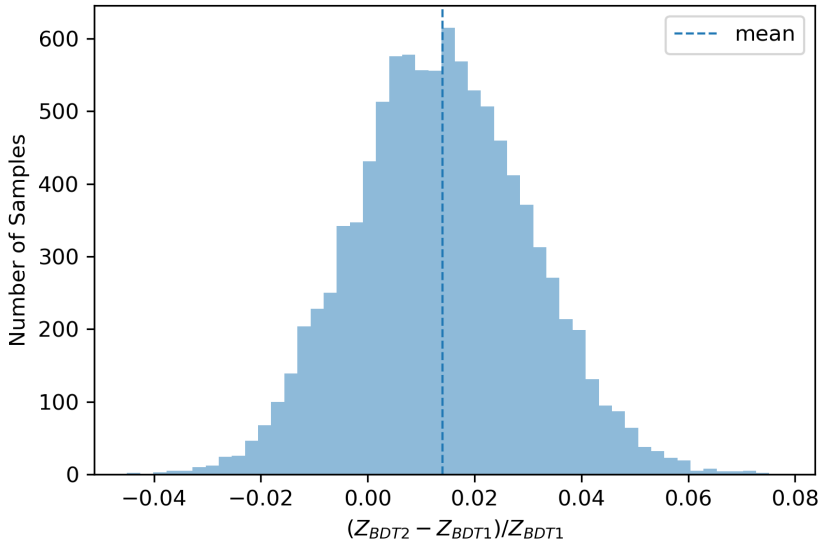


Poisson Bootstrap: Results for SM case III

- The resulting distributions of the combined significance Z and the relative improvement in significance $Z_{\text{imp}} = (Z_{\text{BDT2}} - Z_{\text{BDT1}})/Z_{\text{BDT1}}$ are shown below



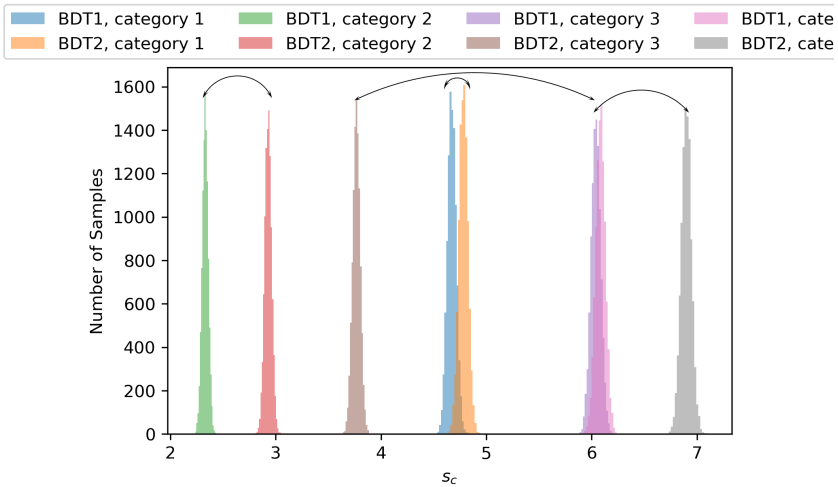
(a) Combined Significance



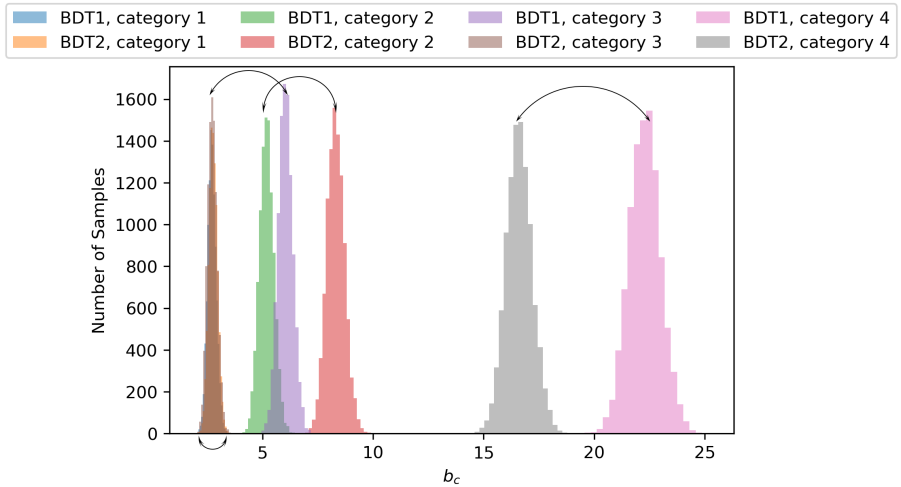
(b) Relative improvement in significance

Poisson Bootstrap: Results for BSM case I

- A total of $N = 10,000$ bootstrapped samples are produced for both **bdt models before and after HPO (= BDT 1 and BDT 2)**. In each sample, the same Poisson weights are applied for both bdt models.
- In the BSM case (i.e. using lambda = 10 signal), the resulting distributions of the signal and background yields for each category is shown below



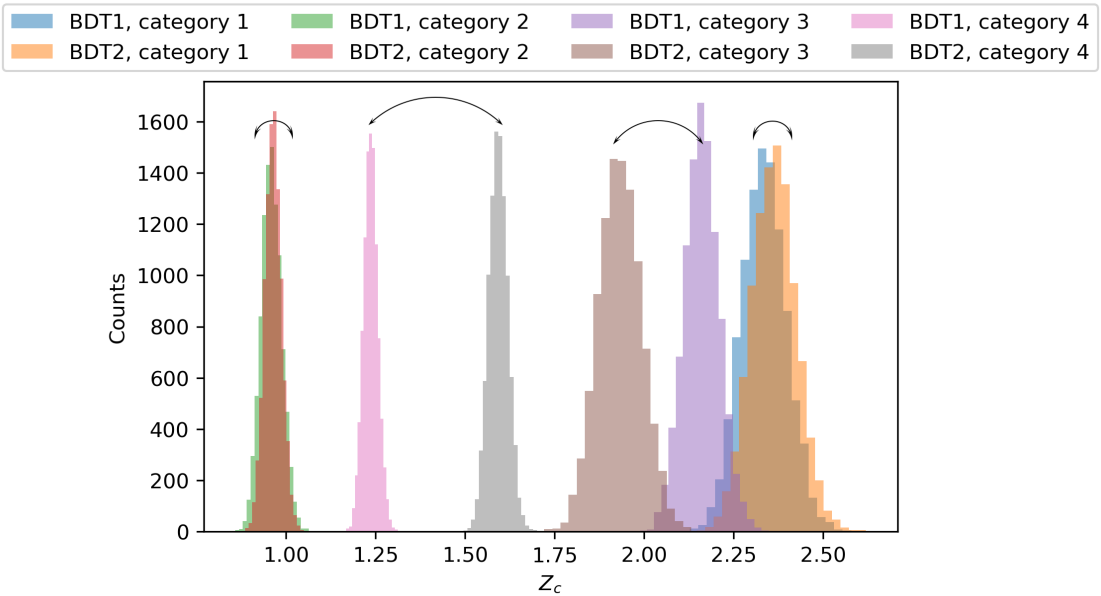
(a) Signal Yields



(b) Background Yields

Poisson Bootstrap: Results for BSM case II

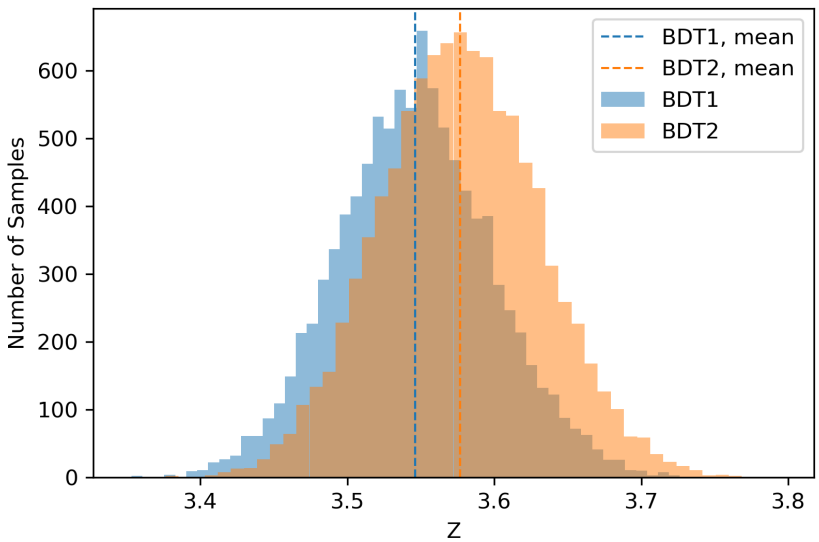
- The resulting distribution of the significance for each category is



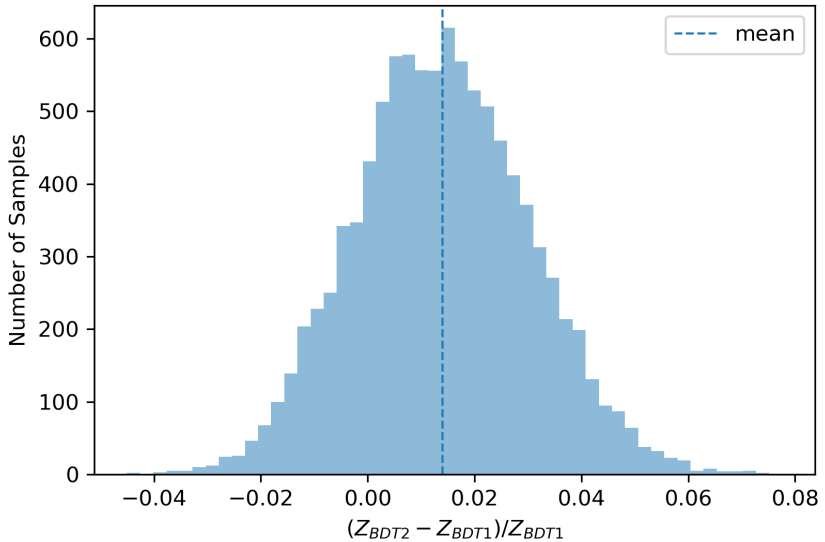
Significance

Poisson Bootstrap: Results for BSM case III

- The resulting distributions of the combined significance Z and the relative improvement in significance $Z_{\text{imp}} = (Z_{\text{BDT2}} - Z_{\text{BDT1}})/Z_{\text{BDT1}}$ are shown below



(a) Combined Significance



(b) Relative improvement in significance

Poisson Bootstrap: Summary of Results

SM Case	BDT1	BDT2
Z	0.4752 ± 0.0147	0.4818 ± 0.0146
$Z_{\text{BDT2}} - Z_{\text{BDT1}}$	0.00657 ± 0.00774	
Z_{imp}	0.0140 ± 0.00163	
Linear Correlation	0.861	

BSM Case	BDT1	BDT2
Z	3.546 ± 0.0525	3.577 ± 0.0544
$Z_{\text{BDT2}} - Z_{\text{BDT1}}$	0.0306 ± 0.0304	
Z_{imp}	0.0087 ± 0.0086	
Linear Correlation	0.839	

- The improvement in significance for the BDT model after HPO is on one sigma level for both SM and BSM cases.
- The linear correlation between the BDT model before and after HPO is very high because both models are trained using the same algorithm (boosted decision tree) on the same training and validation data with the only difference in the hyperparameters used.

Feature Importance Study - Introduction

- Currently the BDT model is trained with 22 variables (features):
  - Jets: pT, eta, phi and pseudo-continuous b-tagging score of the first 2 jets
  - Photons: pT/$m_{\gamma\gamma}$, eta, phi of the 2 photons
  - MET: magnitude and phi
  - H(bb): (formed by the first 2 jets): pT, eta, phi, mass
  - Extra variables: HT(jets), topness
- We want to know the importance of these variables in the training.
- In this talk, we present three different methods for studying feature importance:
  - Permutation Feature Importance (PFI)
  - Leave-One-Covariate-Out (LOCO) method
  - SHapley Additive exPlanation (SHAP) values

Permutation Feature Importance - Introduction

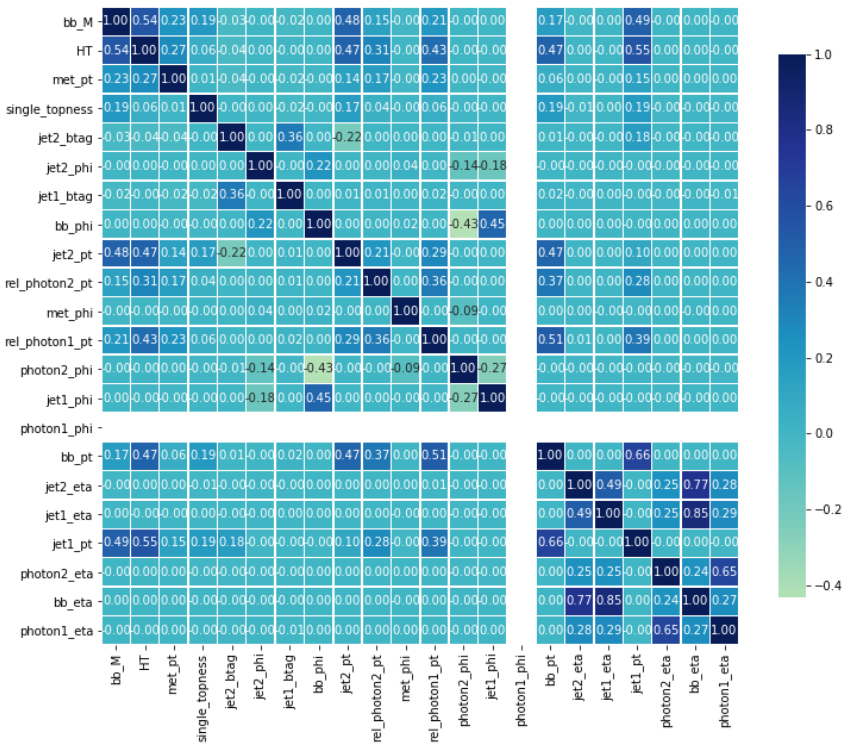
- Permutation feature importance (PFI) measures the importance of a feature by permuting its values in the data and examining the corresponding drop in the performance on a model built with the original data.
- The approach can be described in the following steps:
  1. Train the baseline model and record the metric (i.e. auc) by passing the test dataset.
  2. Re-shuffle values from one feature in the dataset, pass the dataset to the model again to calculate the metric for this modified dataset. The feature importance is the difference between the benchmark score and the one from the modified (permuted) dataset.
  3. Repeat 2. for all features in the dataset.
- More information about this method can be found here.

Permutation Feature Importance - Limitations

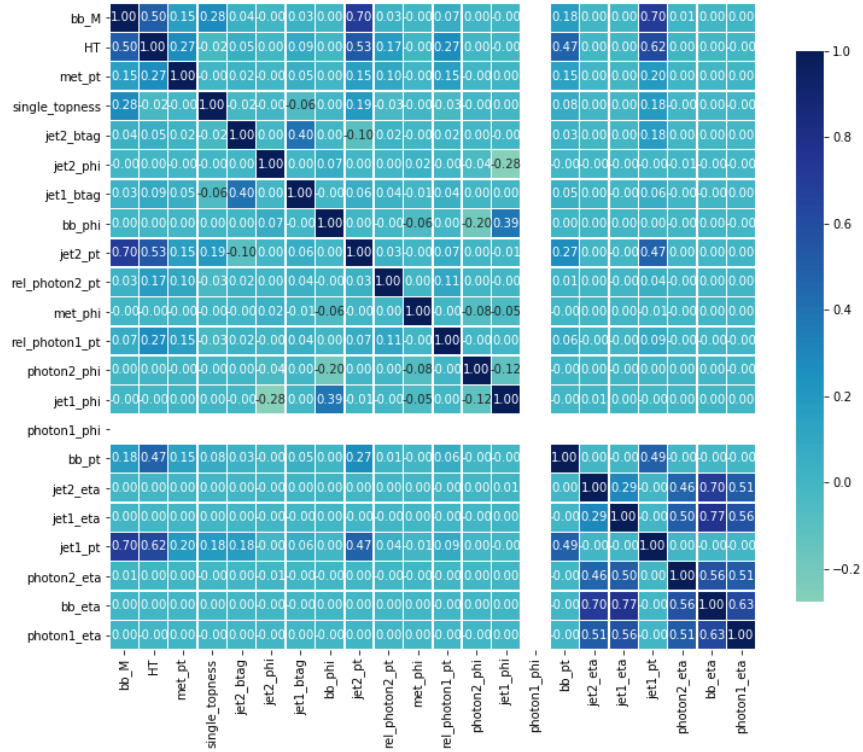
- Permutation feature importance is accurate only when the features are uncorrelated (which is usually false in reality). There are a few issues with correlated features:
  - The permutation of features produces unlikely or unrealistic data instances when two or more features are correlated (e.g. the momentums no longer add up to the invariant mass) since the correlation is broken during the shuffling step of a feature. In this case, the permutation feature importance can be biased by the unrealistic data instances.
  - Adding a correlated feature can decrease the importance of the associated feature by splitting the importance between both features. For example, if two identical variables are given, the model may think both are equality important and hence the overall importance of each is decreased.

Feature Correlations

- The plots below show the correlation matrix for the variables in the training dataset:



(a) SM case



(b) BSM case

- Variables like bb_M, HT are highly correlated with many other variables. Similar variables such as the phi variables are also highly correlated with each other.
- A major caveat is that a zero correlation coefficient does not necessary mean no correlation especially with symmetric variables. For instance, two variables can be positively correlated and negatively correlated half of the time which gives a zero overall correlation coefficient.

Permutation Feature Importance - Results

- The permuation feature importance plot is shown below

Weight	Feature
0.1891 ± 0.0016	bb_M
0.0215 ± 0.0006	HT
0.0156 ± 0.0009	met_pt
0.0093 ± 0.0014	single_topness
0.0019 ± 0.0004	jet2_btag
0.0018 ± 0.0001	jet2_phi
0.0016 ± 0.0003	jet1_btag
0.0016 ± 0.0005	bb_phi
0.0008 ± 0.0009	jet2_pt
0.0007 ± 0.0003	rel_photon2_pt
0.0007 ± 0.0004	met_phi
0.0006 ± 0.0008	rel_photon1_pt
0.0004 ± 0.0011	photon2_phi
0.0004 ± 0.0004	jet1_phi
0 ± 0.0000	photon1_phi
-0.0016 ± 0.0011	bb_pt
-0.0020 ± 0.0004	jet2_eta
-0.0027 ± 0.0005	jet1_eta
-0.0027 ± 0.0008	jet1_pt
-0.0029 ± 0.0003	photon2_eta
-0.0035 ± 0.0008	bb_eta
-0.0040 ± 0.0005	photon1_eta

(a) SM case

Weight	Feature
0.1152 ± 0.0026	bb_M
0.0085 ± 0.0012	HT
0.0068 ± 0.0009	jet1_btag
0.0067 ± 0.0012	met_pt
0.0030 ± 0.0007	single_topness
0.0030 ± 0.0007	jet2_btag
0.0025 ± 0.0007	rel_photon1_pt
0.0019 ± 0.0007	rel_photon2_pt
0.0017 ± 0.0003	jet1_pt
0.0017 ± 0.0002	jet2_pt
0.0012 ± 0.0006	photon2_phi
0.0009 ± 0.0005	photon2_eta
0.0005 ± 0.0001	bb_phi
0.0005 ± 0.0003	jet1_phi
0.0002 ± 0.0002	jet2_phi
0.0002 ± 0.0007	met_phi
0 ± 0.0000	photon1_phi
-0.0004 ± 0.0011	bb_pt
-0.0010 ± 0.0003	jet2_eta
-0.0013 ± 0.0006	bb_eta
-0.0014 ± 0.0005	jet1_eta
-0.0019 ± 0.0003	photon1_eta

(b) BSM case

The importance i_j for feature f_j is defined as

$$i_j = s - \frac{1}{K}\sum_{k=1}^{K} s_{k,j} \tag{6}$$

where s is the baseline auc, $s_{k,j}$ is the auc for the k-th trial with the j-th feature permuted, K is total number of trials

Note: SM and BSM cases refer to training for the $m_X < 350$ category and the $m_X > 350$ category respectively

- The variable bb_M seems to be the most important which is expected.
- A negative value means permuting the feature gives a better performance. Most of the negative values come from the eta variables. This may be caused by unrealistic data instances from permuting the feature or some unknown interferences among the variables or (unlikely) statistical fluctuation .

The Leave-One-Covariate-Out (LOCO) method - Introduction

- The basic principle for the Leave-One-Covariate-Out (LOCO) method is straightforward:

  Removing a feature from the data and examining the drop in performance when a new model is learned without it.

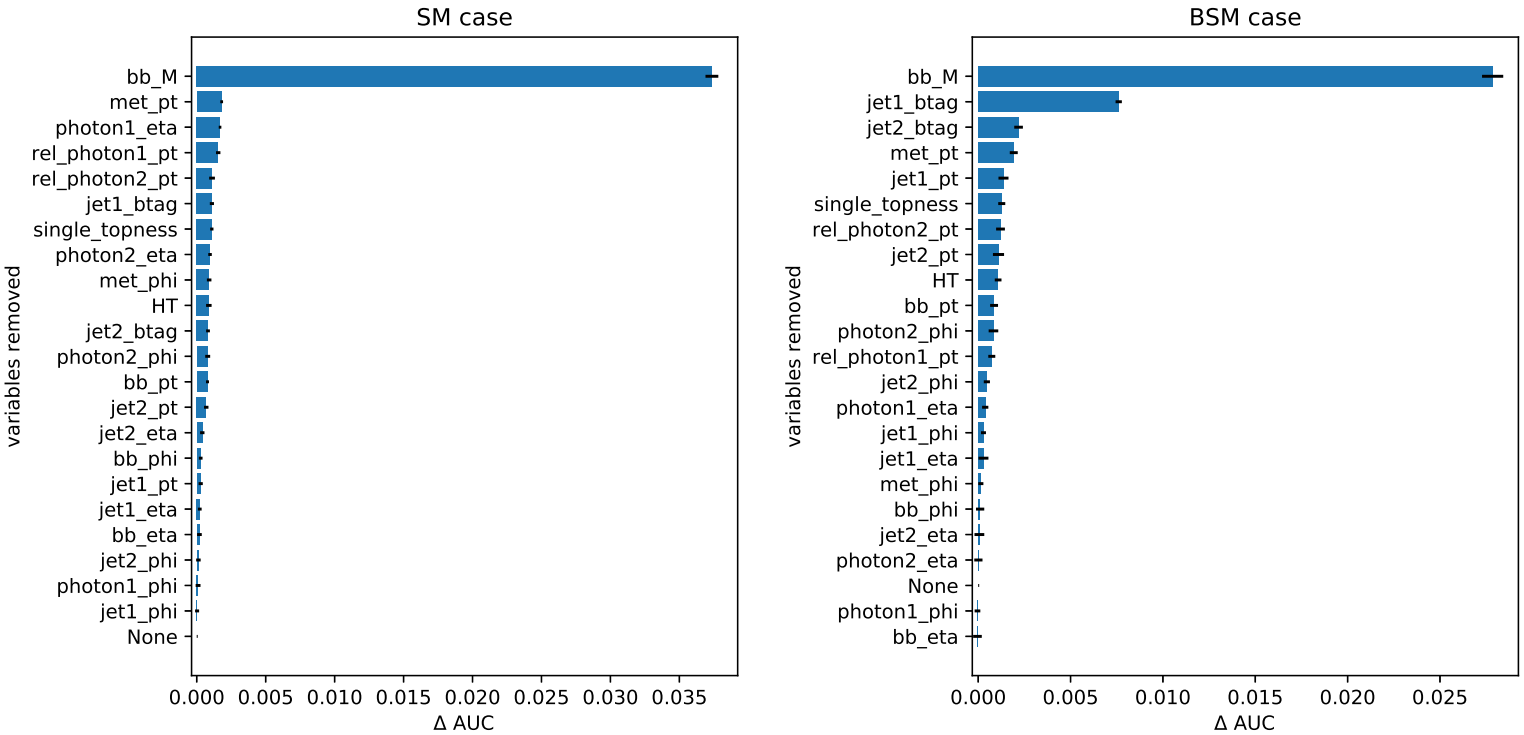
- More information about this method can be found here.

- Similar to the Permutation Feature Importance method, it also suffers from the problem of correlated features. There can be a huge drop in importance value or ranking number for highly correlated features since the model can still learn well using the rest of the correlated variables if one of them are removed.

- For the results shown in the next two pages, each trial (removal of a particular variable) is repeated 10 times (data is shuffled) to obtain the average value and the standard deviation

The Leave-One-Covariate-Out (LOCO) method - Results I

- The plot for the difference in AUC compared to the baseline is shown below

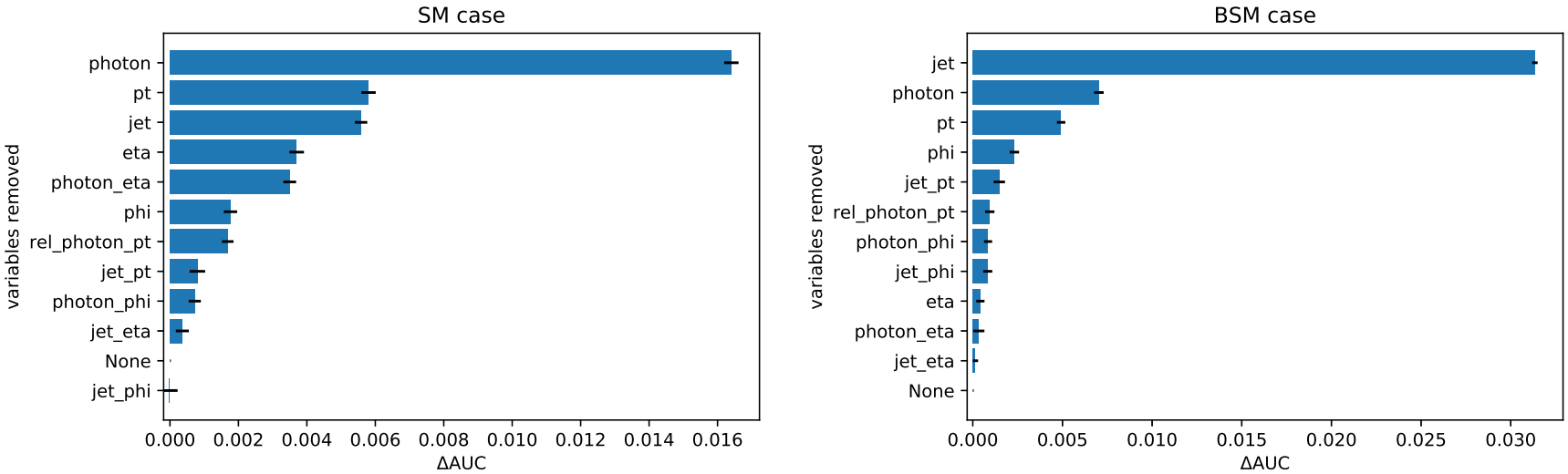


(a) SM case (b) BSM case

- Some of the eta and phi variables seem to have little effect in the training when removed. This can either be that they are not important or that other correlated variables replaced their role.

The Leave-One-Covariate-Out (LOCO) method - Results II

- Alternatively, one can remove a group of similar variables. The result is shown below



(a) SM case (b) BSM case

- Here 'pt', 'phi', 'eta','photon','jet' means all variables with 'pt', 'phi', 'eta','photon','jet' in their names are removed from the data.
- Removing all the pt variables seem to have the greatest impact compared to removing all the eta or phi variables.

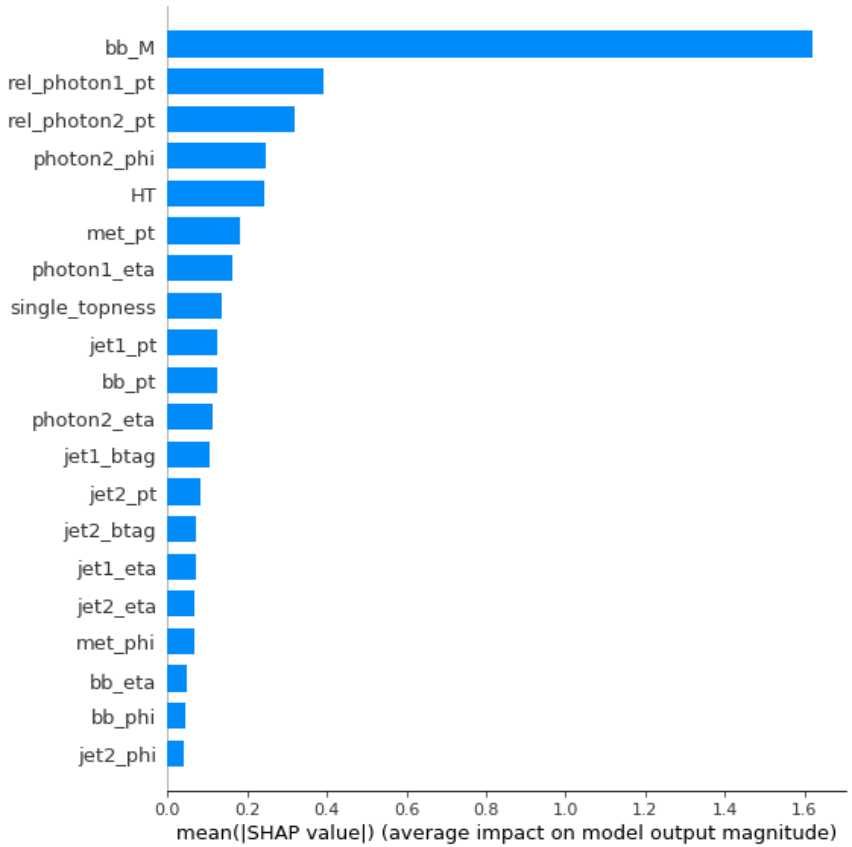
The SHAP (SHapley Additive exPlanation) values - Introduction

- The SHAP values quantifies the average marginal contribution of a feature to the prediction made by the model (i.e. BDT score) across all possible combinations.
- For example, to find the marginal contribution of the variable bb_M, the set of combinations of variables to be considered are:

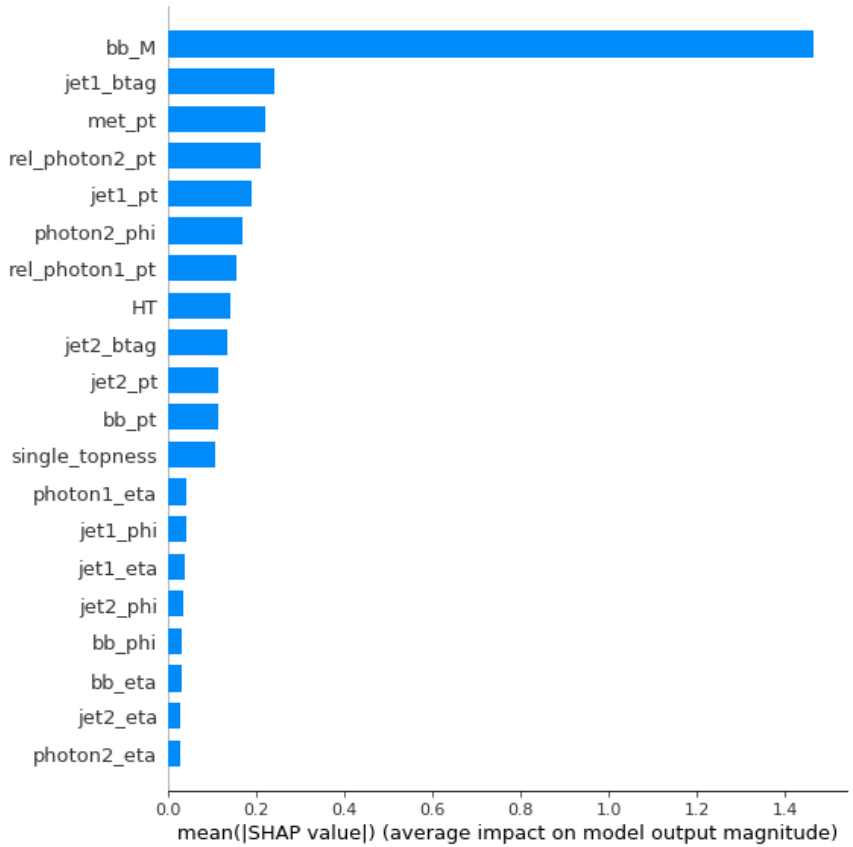
$$\{bb_M\}, \{bb_M, HT\}, \{bb_M, met_{pt}\}, \ldots, \{bb_M, HT, met_{pt}\}, \ldots, \{\text{all variables}\}$$

  There are a total of 2^F combinations for data with F number of features.
- Instead of evaluating 2^F models, some approximation and sampling methods are used to simplify the caculation.
- More information on how the SHAP values are calculated can be found here
- Like the PFI method, the SHAO value method suffers from inclusion of unrealistic data instances when features are correlated.

The SHAP (SHapley Additive exPlanation) values - Results I



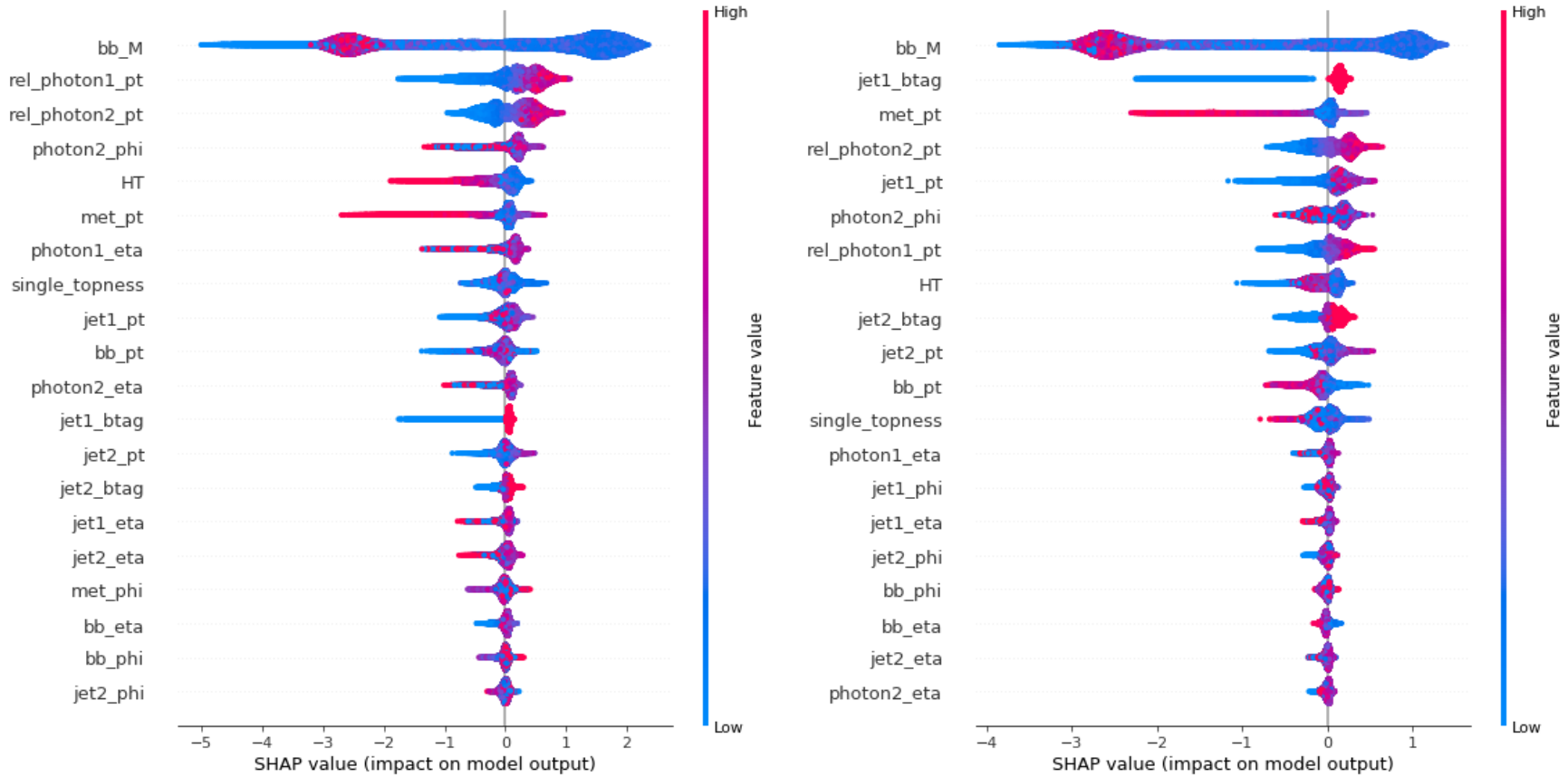
(a) SM case



(b) BSM case

- The global importance for a feature is obtained by averaging the absolute Shapley values per feature across the data. It is a measure of the average absolute impact of the feature on the model ouput (i.e. BDT score).

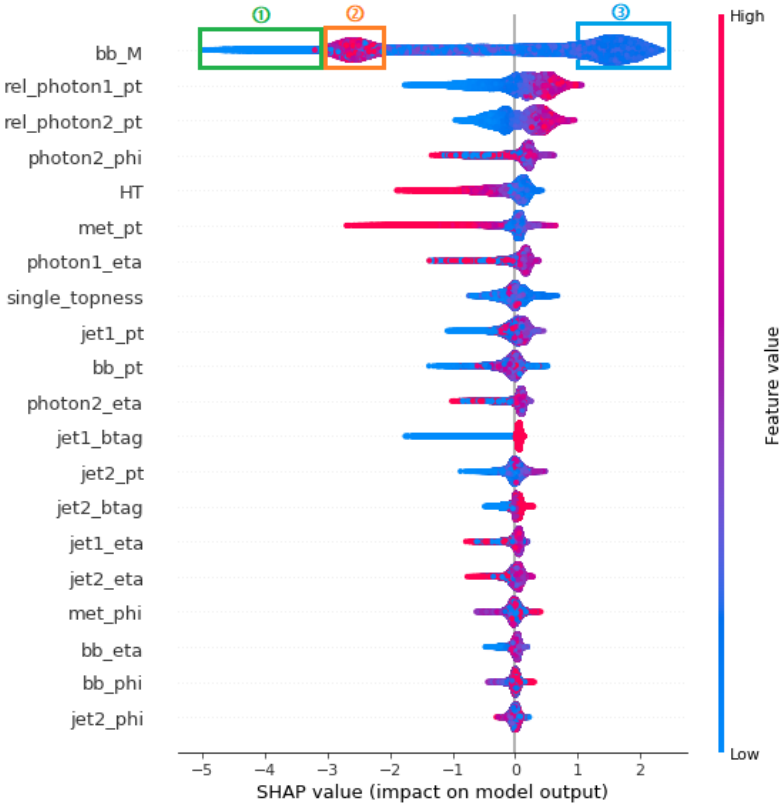
The SHAP (SHapley Additive exPlanation) values - Results II

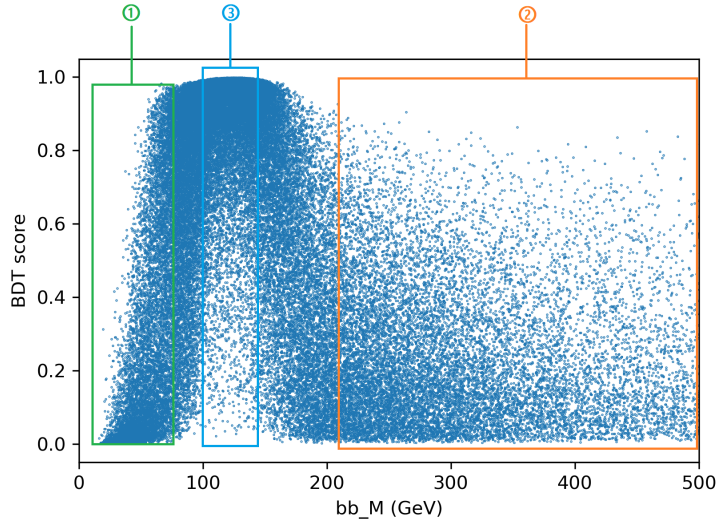


(a) SM case (b) BSM case

- The summary plots above combines feature importance with feature effects. Each point on the plot is a Shapley value for a feature and an instance (a training using a subset of features that includes the target feature). The color represents the value of the feature from low to high. A 'blob' indicates a large number of instances end up at some particular SHAP values.

The SHAP (SHapley Additive exPlanation) values - Explanation of Result





- The SHAP value measues the impact of a feature on the model output (BDT score). A more positive SHAP value means more signal like whereas a more negative SHAP value means more background like.
- For the feature bb_M, a very small value (green rectangle) and a very large value (orange rectangle) result in a lower BDT score since they correspond to the majority of the background events. Values around 125 GeV result in a higher BDT score since they correspond to the majority of the signal events.

Feature Importance Study: Summary

- The importance of the variables used in the training of BDT models are studied using three different methods:
  FPI, LOCO and SHAP values.
- All three methods shows that bb_M is the most important variable which is expected.
- All three methods shows that the phi and eta variables are among the least important variables. However, sometimes a few of the eta and/or phi variables have a moderately high importance ranking depending on the method used.
- Other variables such as HT, single topness, btag and the pt variables usually have a moderate importance ranking which goes up and down depending on the method used.
- There are variables that are highly correlated with each other (such as the eta variables), the importance ranking may not be very accurate but not necessarily useless. We should be caucious about making conclusions based on these importance rankings.

Backup Slides - Q & A

- How to resolve the problem of correlated features?

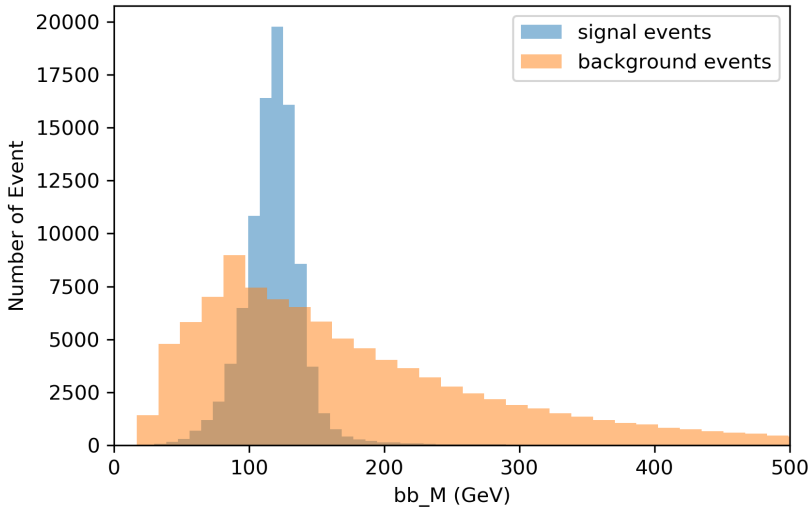
  In case of strong correlation of, for example, two features, it is sensible to permute these together, meaning the building of a group such that the correlation is still present in the calculation of PFI (Parr et al. 2018).

- Whether we should use test or training data to calculate permutation feature importance? Answer here

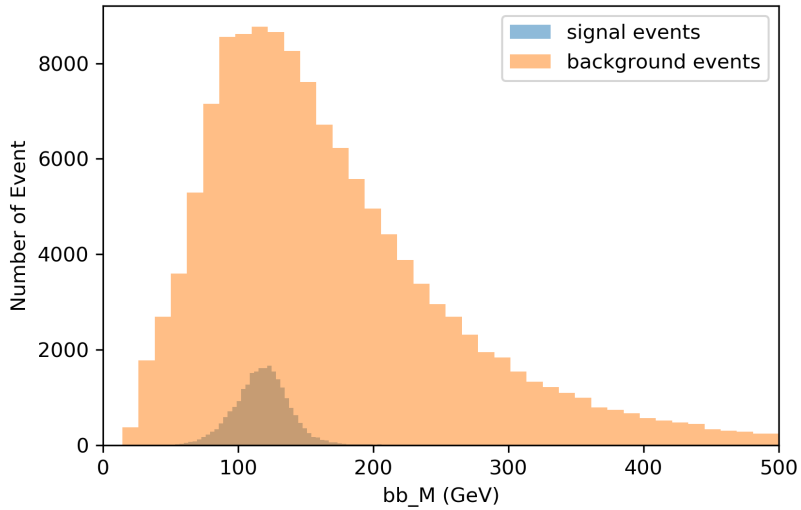
- Alternative methods for finding feature imporrtance?

  Partial Importance (PI) and Individual Conditional Importance (ICI)

bb_M distributions

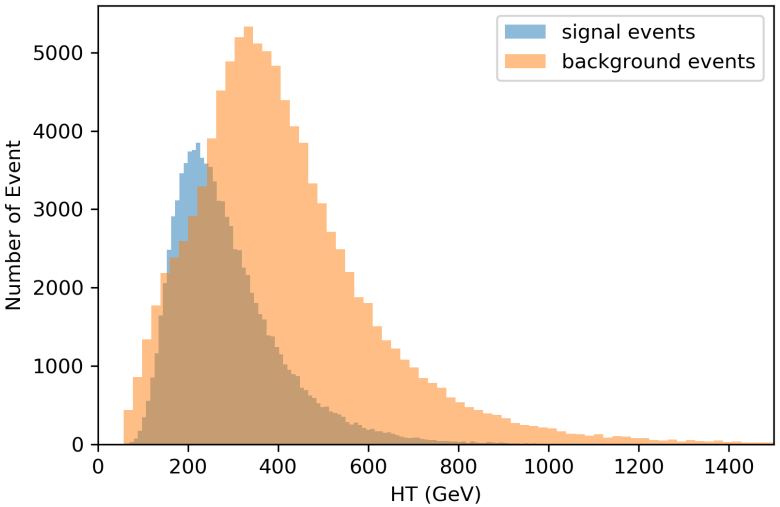


(a) SM case

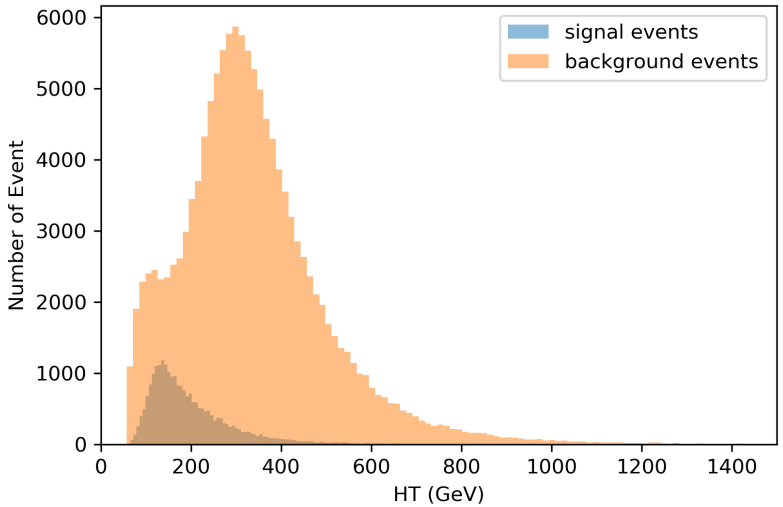


(b) BSM case

HT

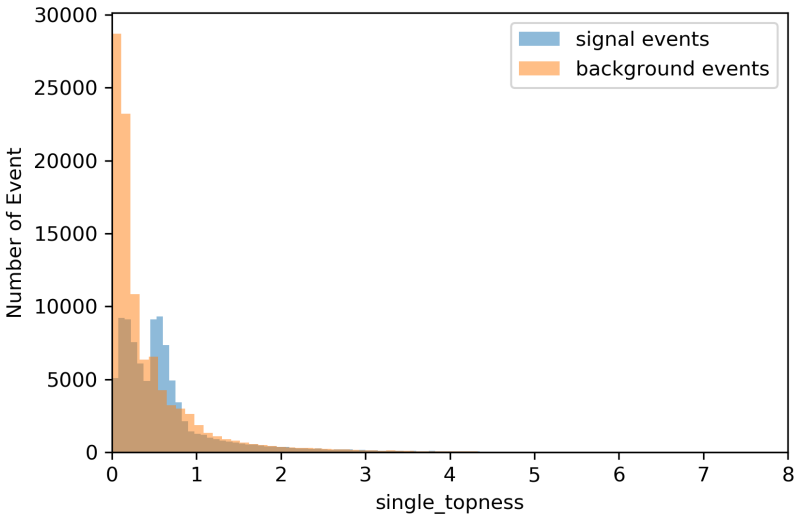


(a) SM case

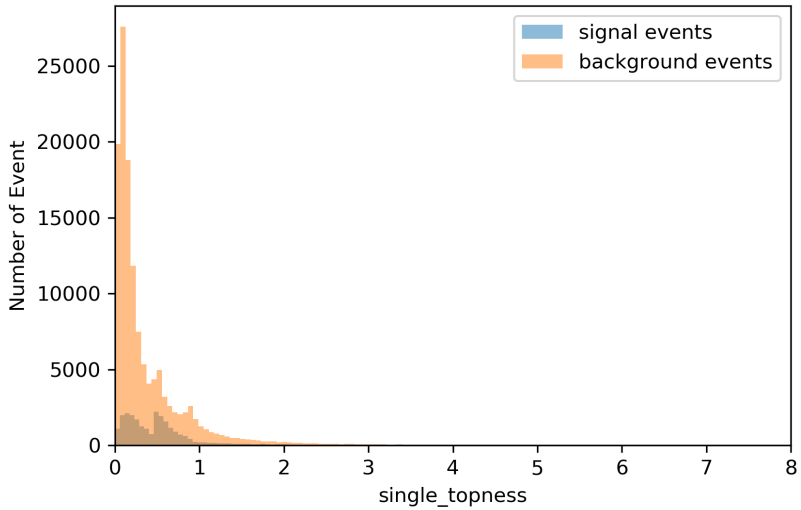


(b) BSM case

single_topness

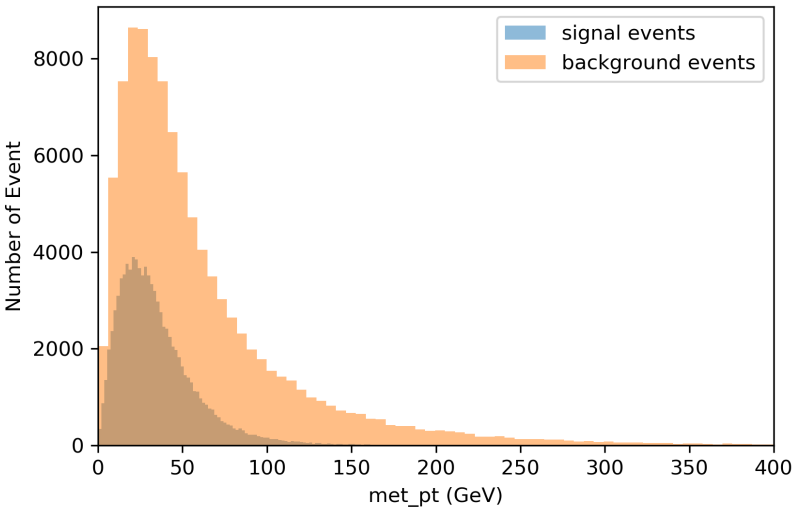


(a) SM case

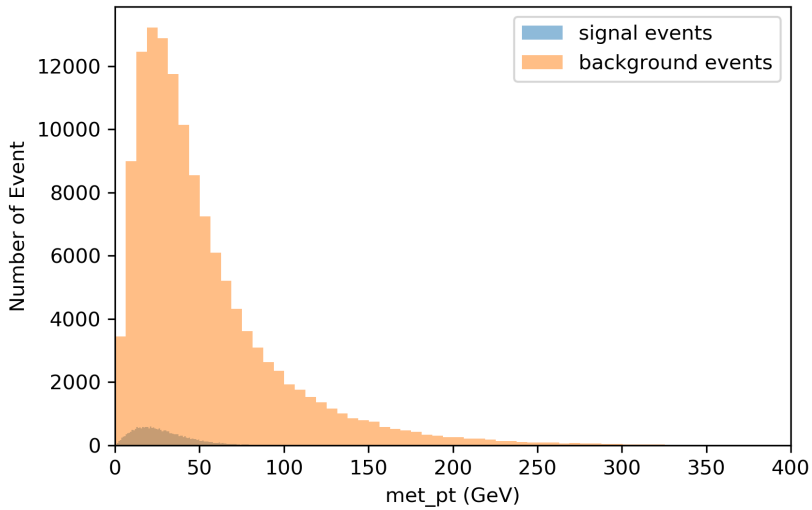


(b) BSM case

met_pt

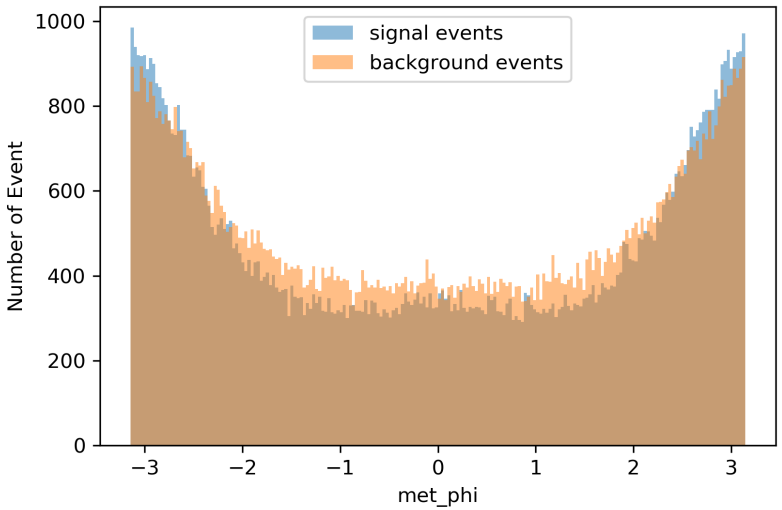


(a) SM case

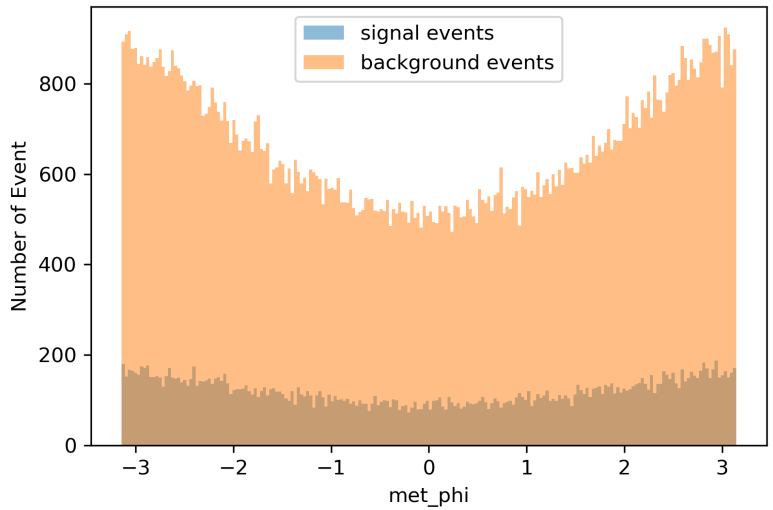


(b) BSM case

met_phi

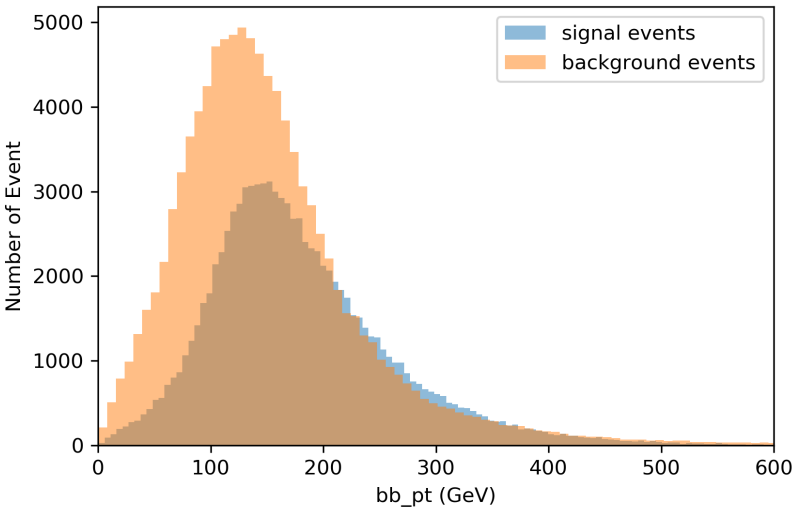


(a) SM case

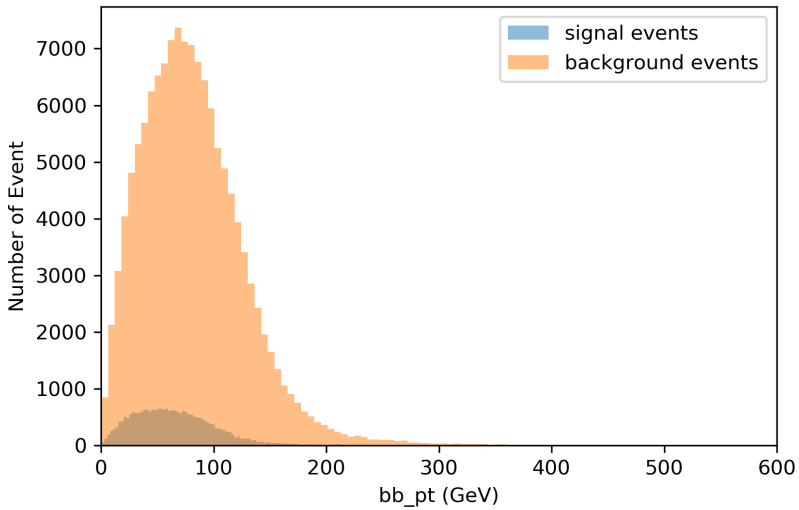


(b) BSM case

bb_pt

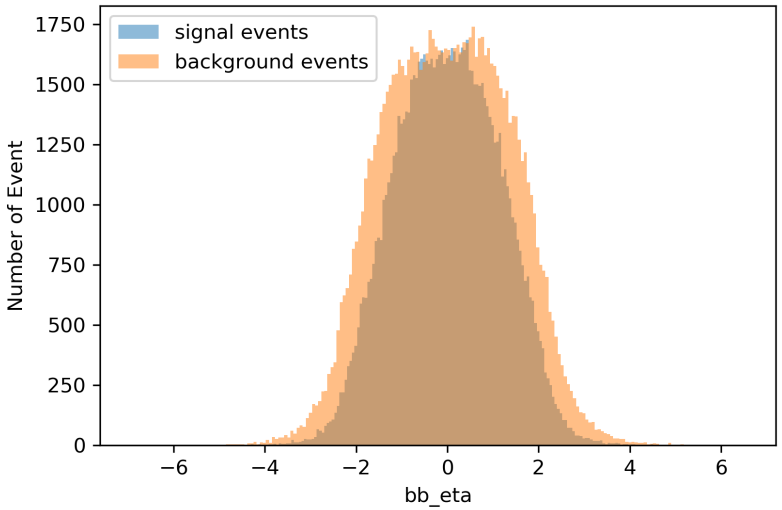


(a) SM case

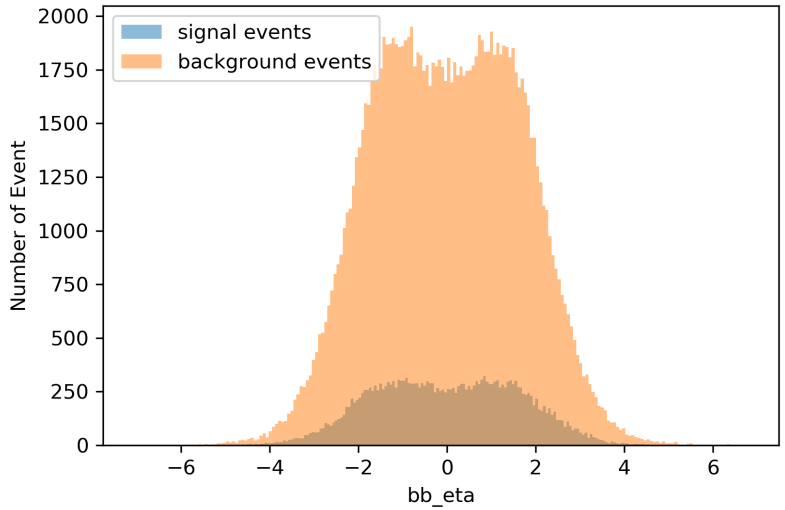


(b) BSM case

bb_eta

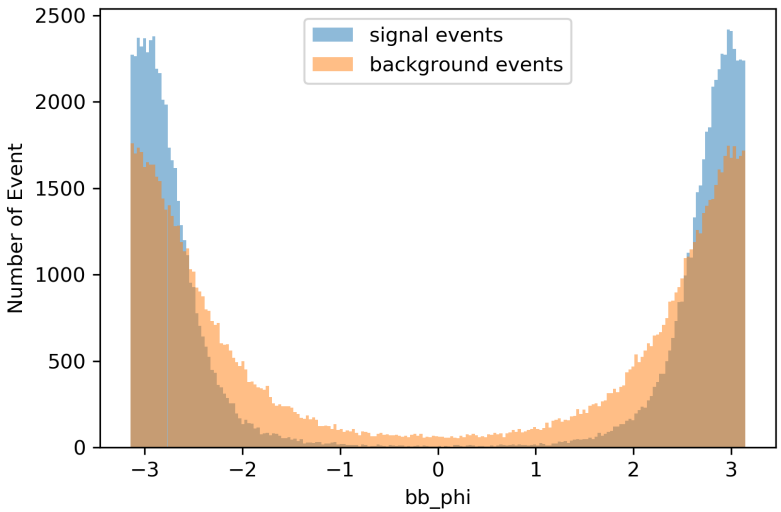


(a) SM case

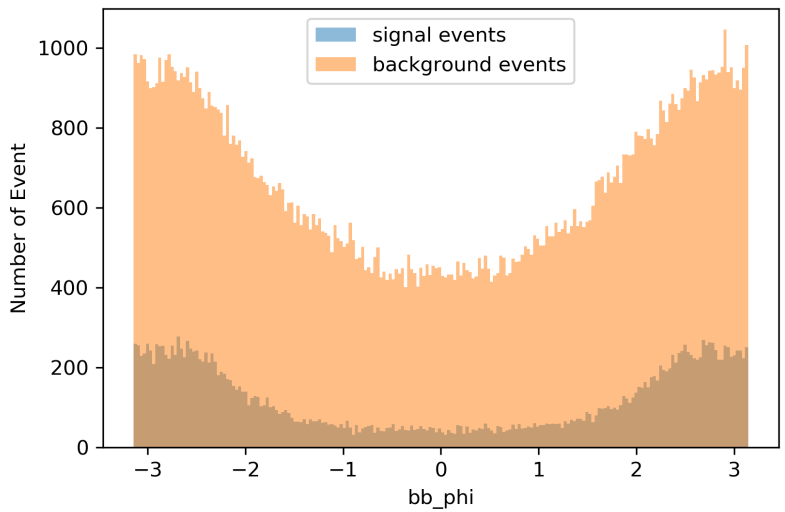


(b) BSM case

bb_phi

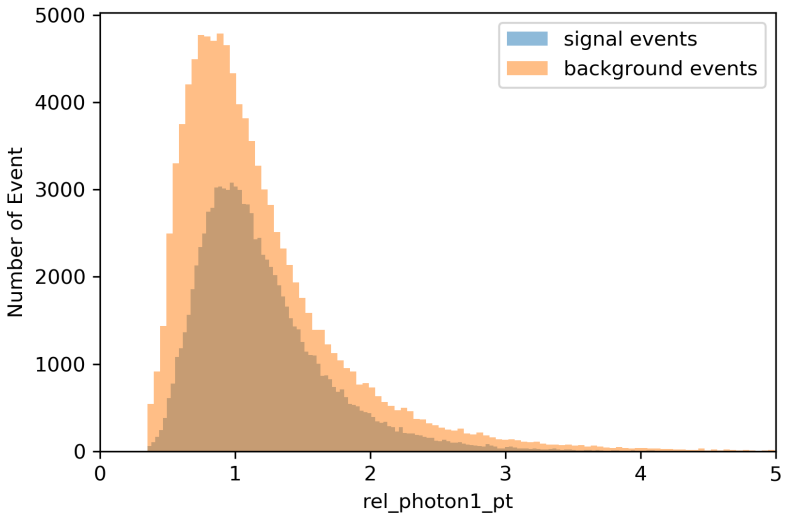


(a) SM case

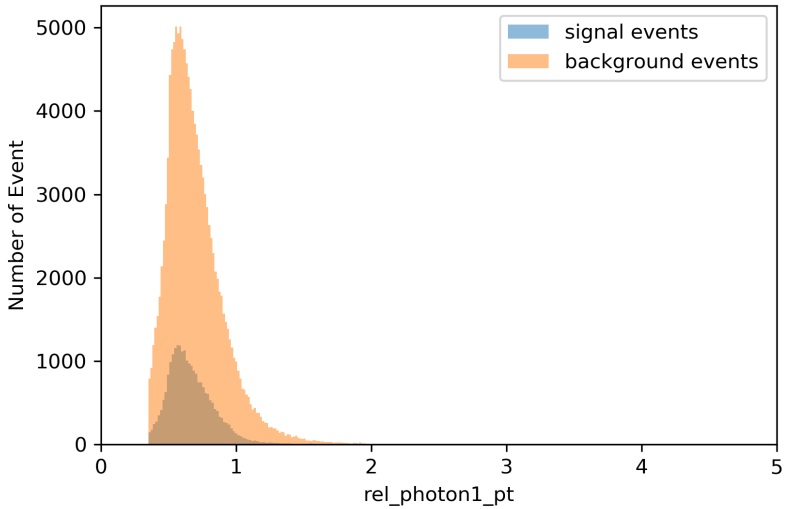


(b) BSM case

rel_photon1_pt

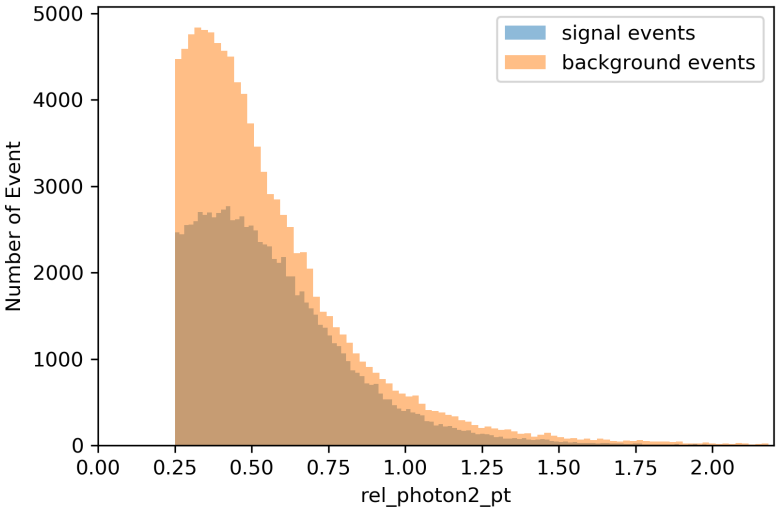


(a) SM case

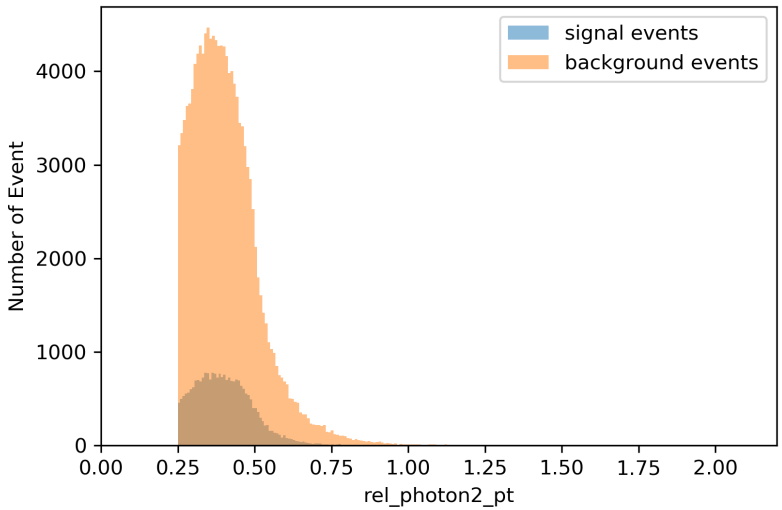


(b) BSM case

rel_photon2_pt

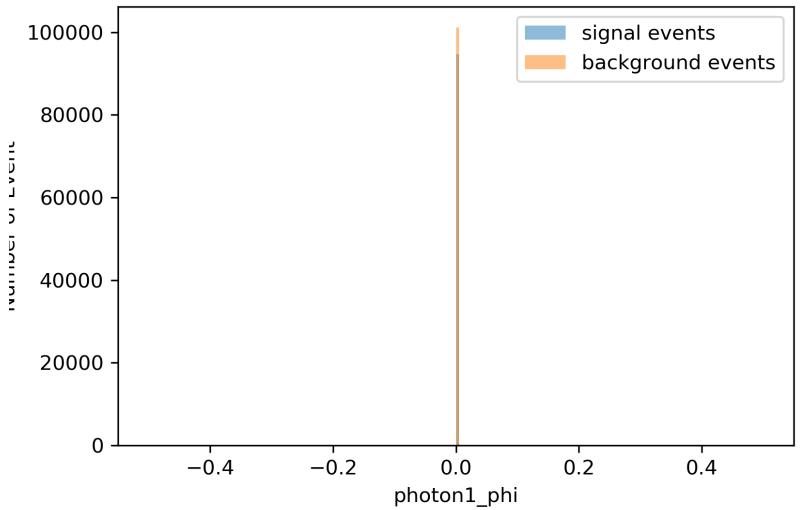


(a) SM case

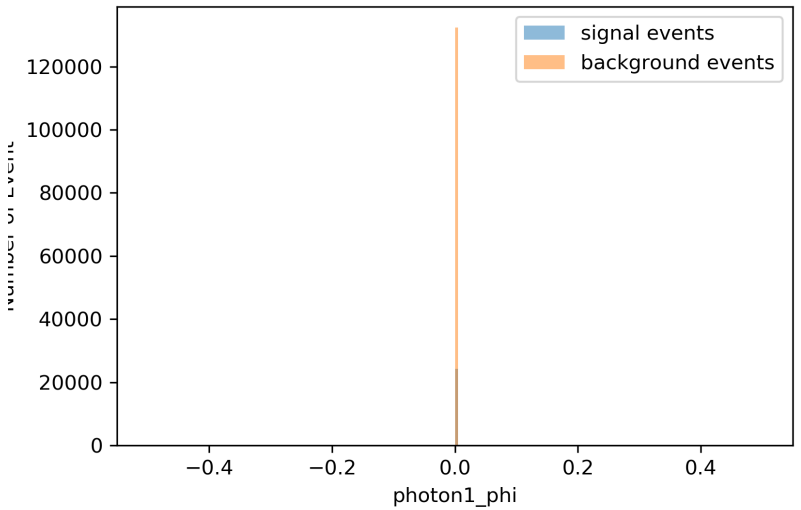


(b) BSM case

photon1_phi

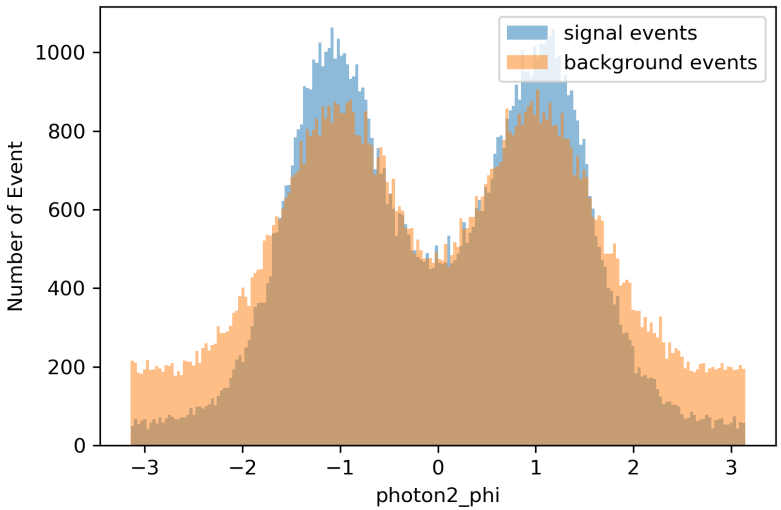


(a) SM case

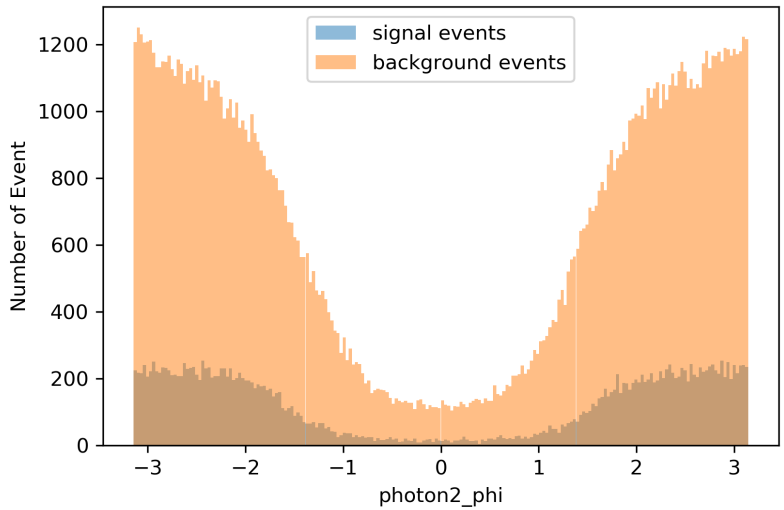


(b) BSM case

photon2_phi

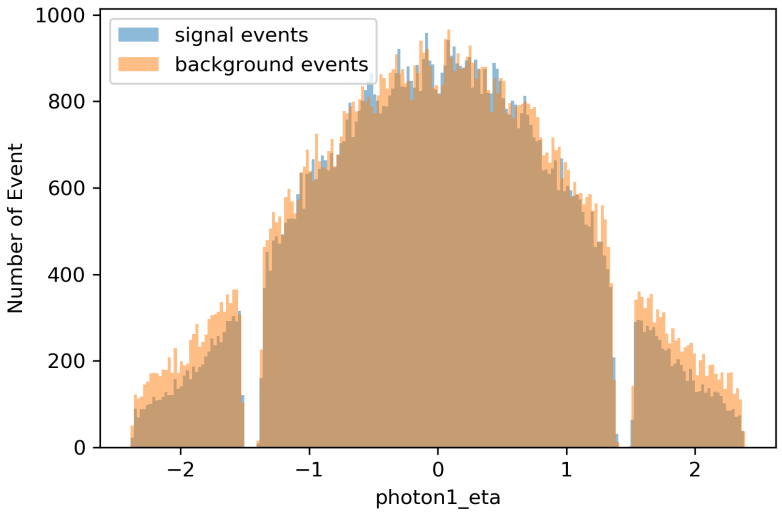


(a) SM case

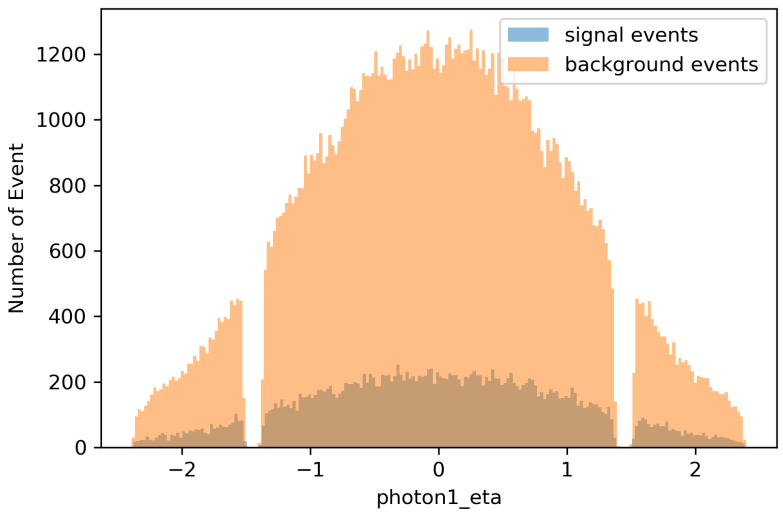


(b) BSM case

photon1_eta

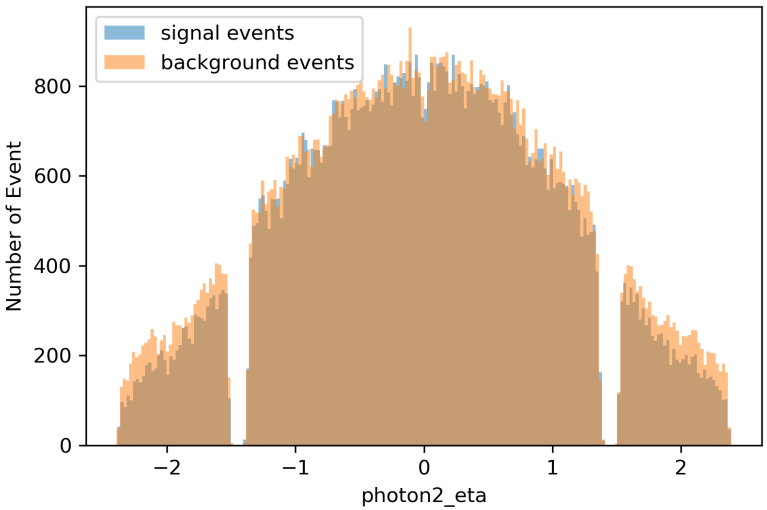


(a) SM case

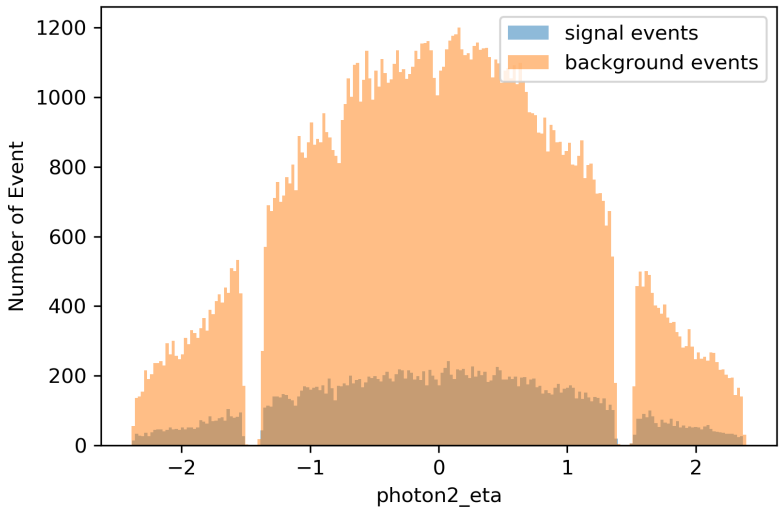


(b) BSM case

photon2_eta



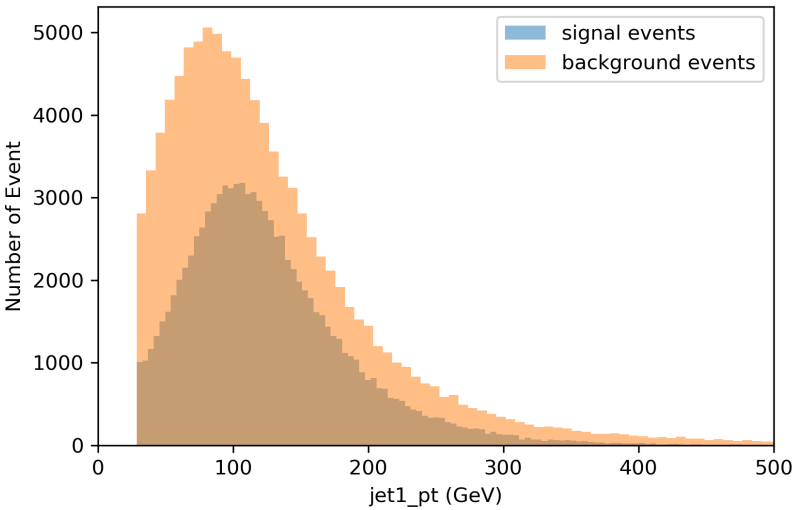
(a) SM case



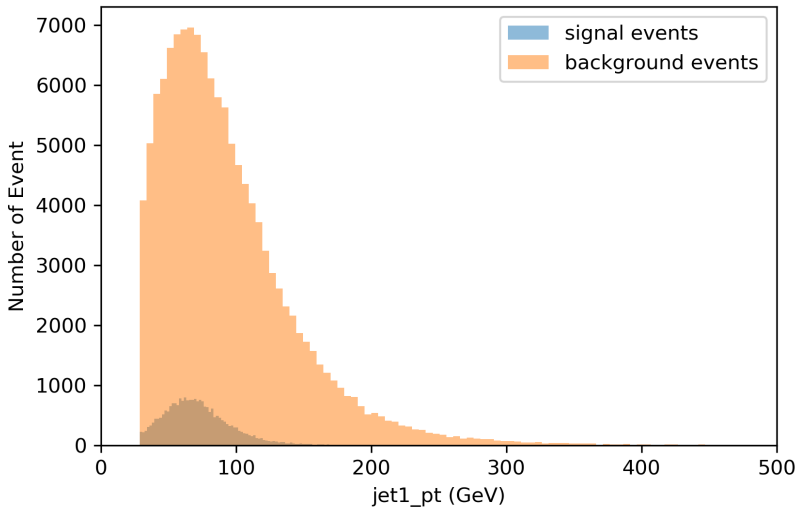
(b) BSM case

Backup Slides - Kinematics Distributions for Signal and Background Events

jet1_pt

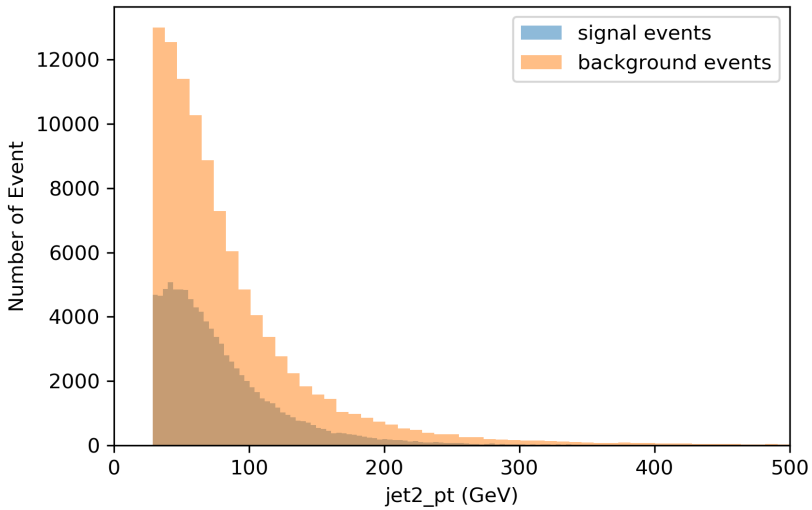


(a) SM case

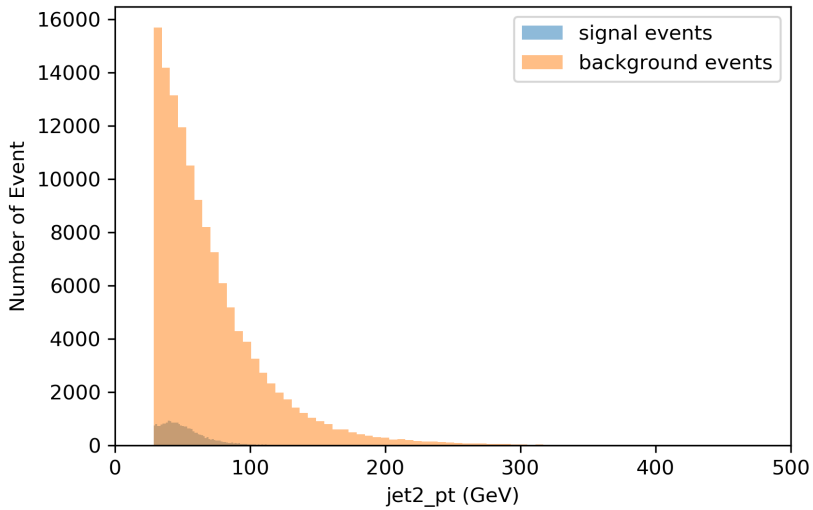


(b) BSM case

jet2_pt

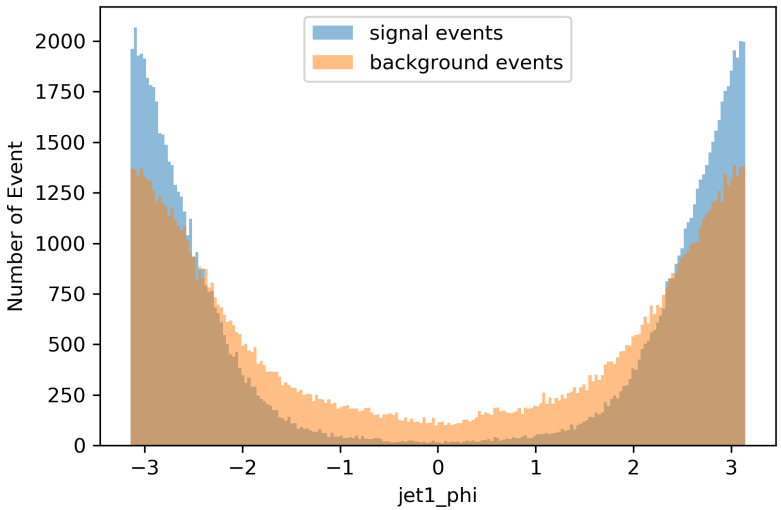


(a) SM case

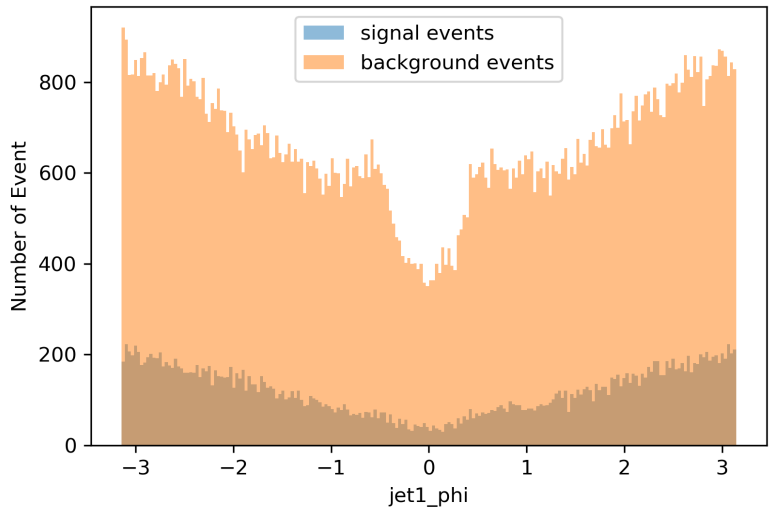


(b) BSM case

jet1_phi

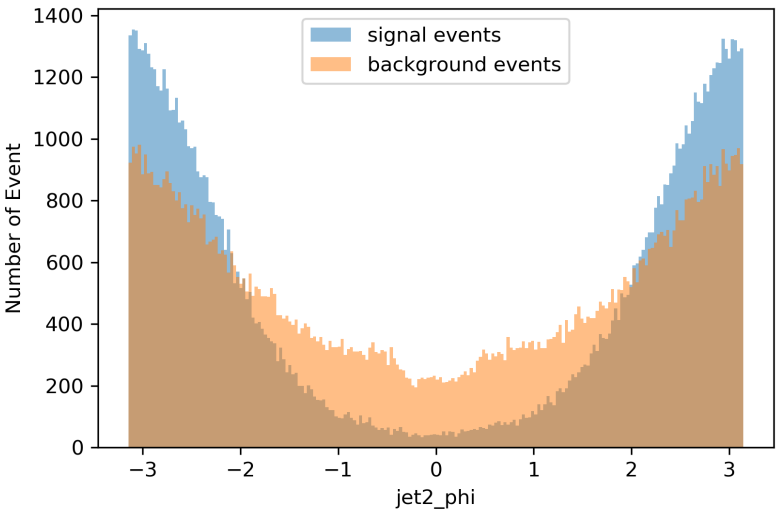


(a) SM case

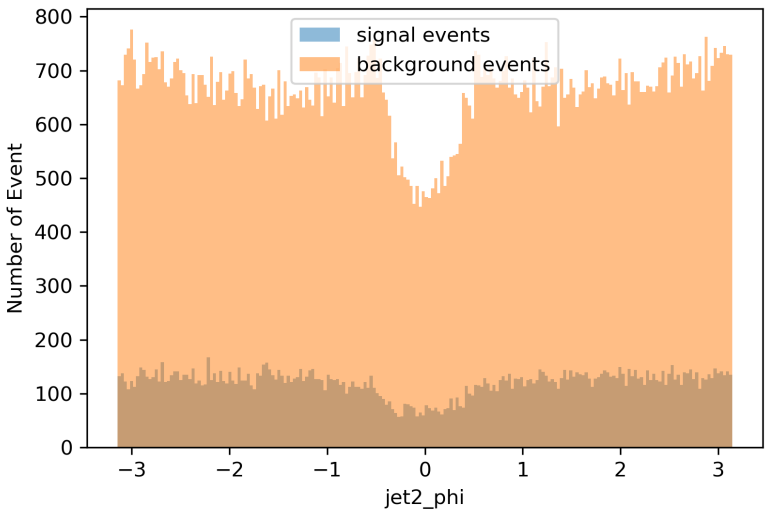


(b) BSM case

jet2_phi

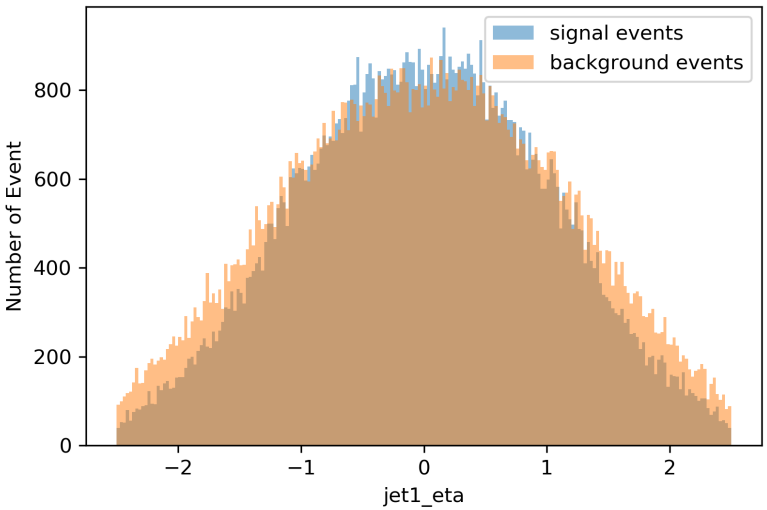


(a) SM case

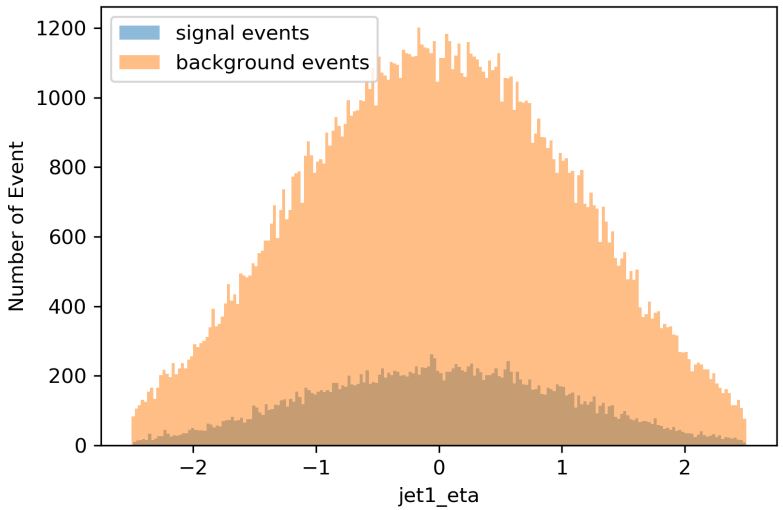


(b) BSM case

jet1_eta

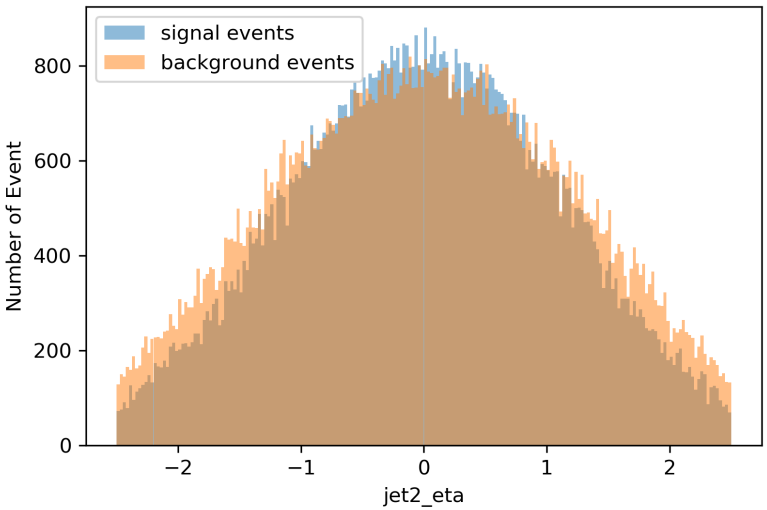


(a) SM case

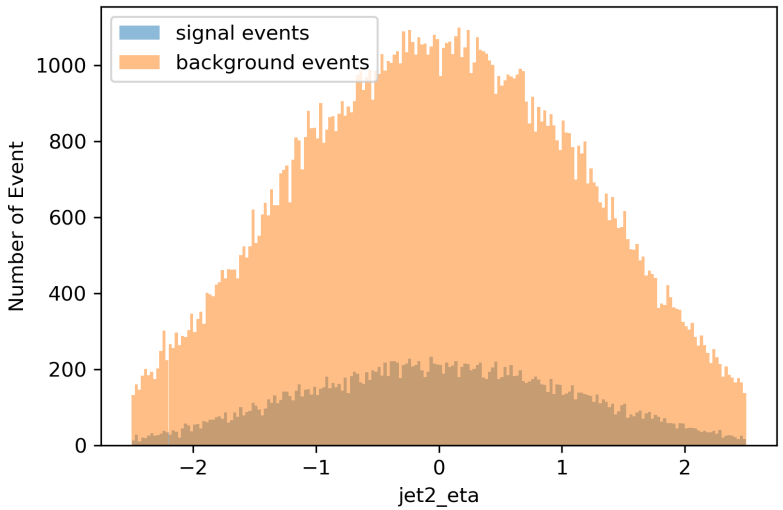


(b) BSM case

jet2_eta

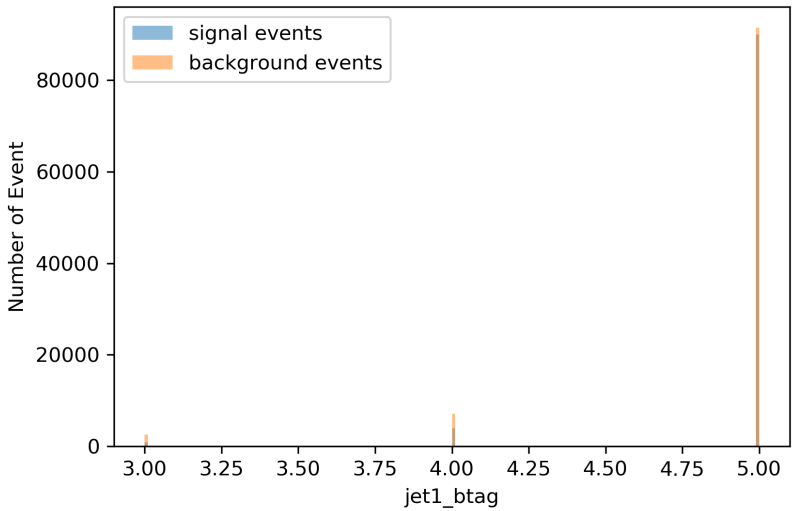


(a) SM case

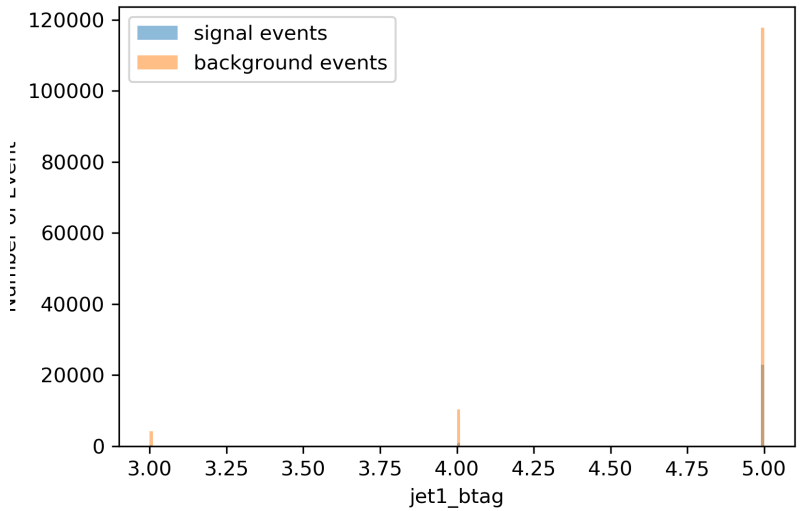


(b) BSM case

jet1_btag

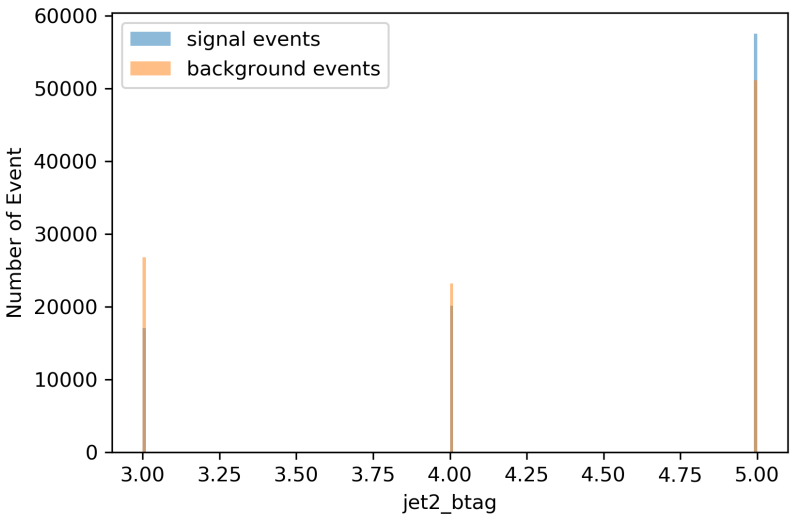


(a) SM case

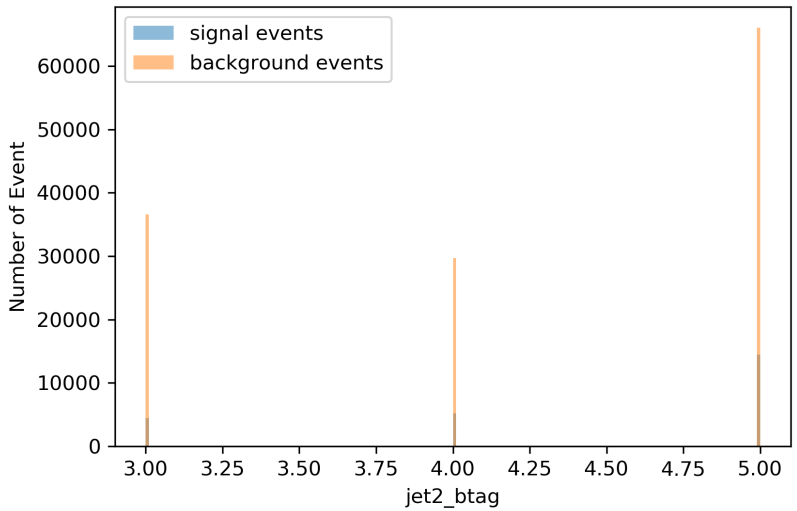


(b) BSM case

jet2_btag

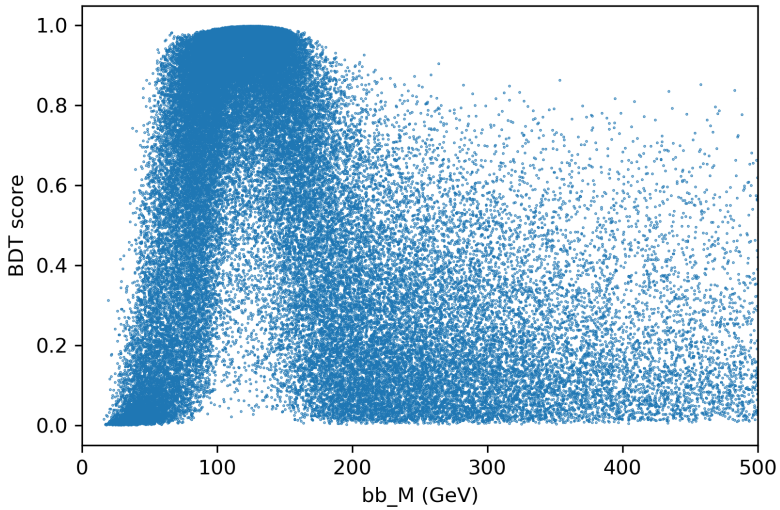


(a) SM case

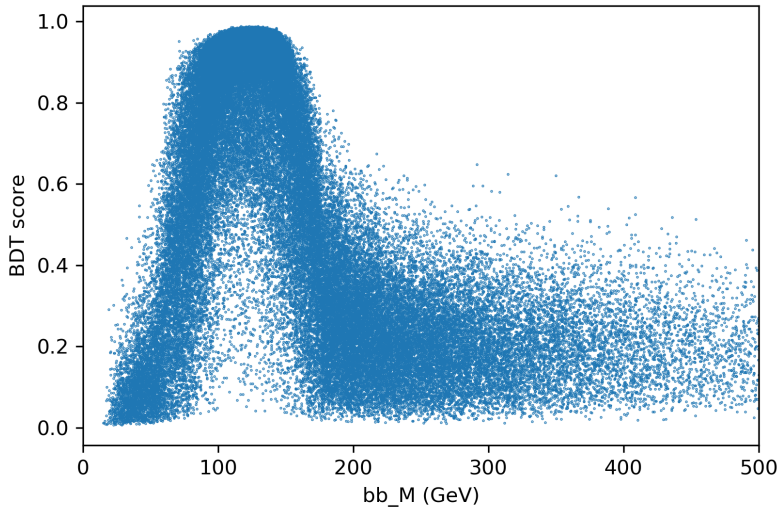


(b) BSM case

bb_M distributions

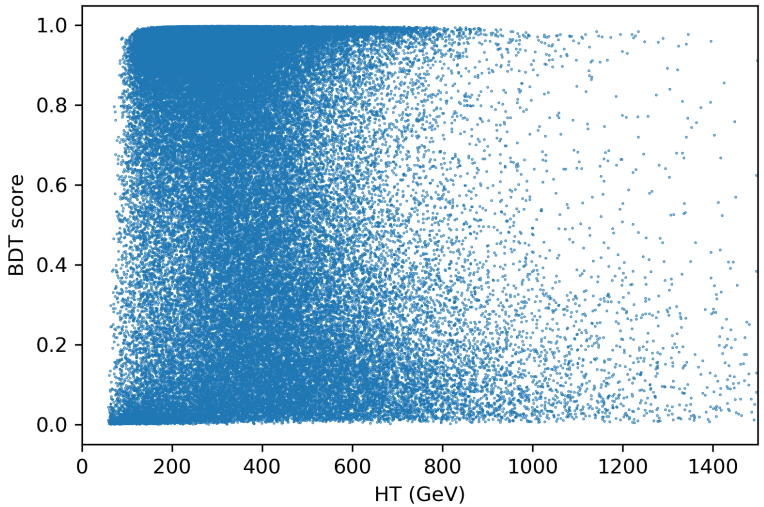


(a) SM case

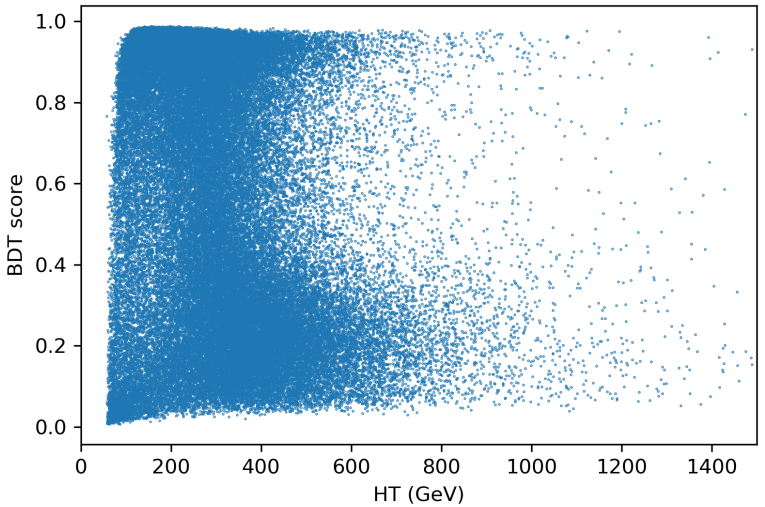


(b) BSM case

HT

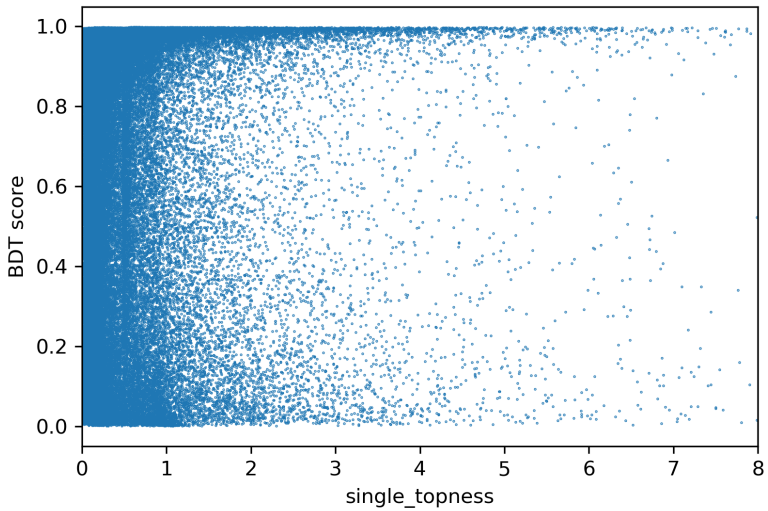


(a) SM case

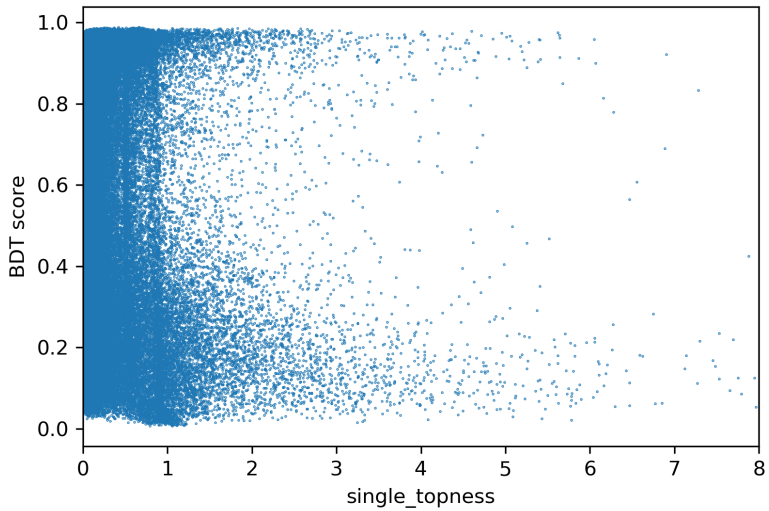


(b) BSM case

single_topness

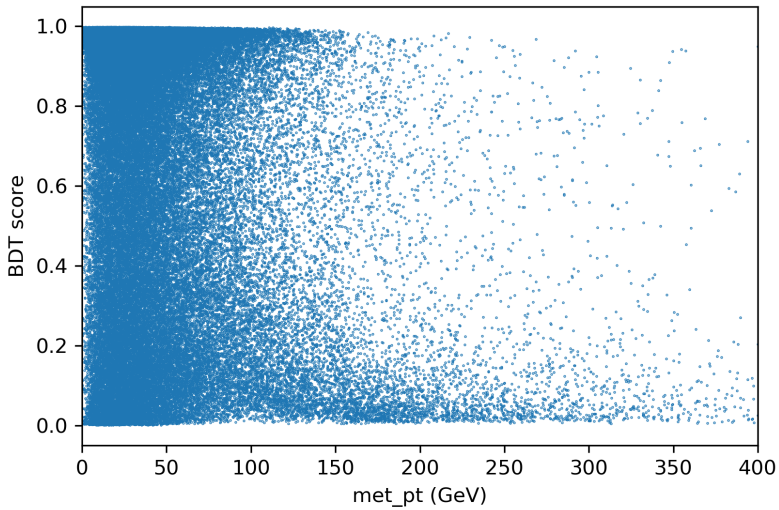


(a) SM case

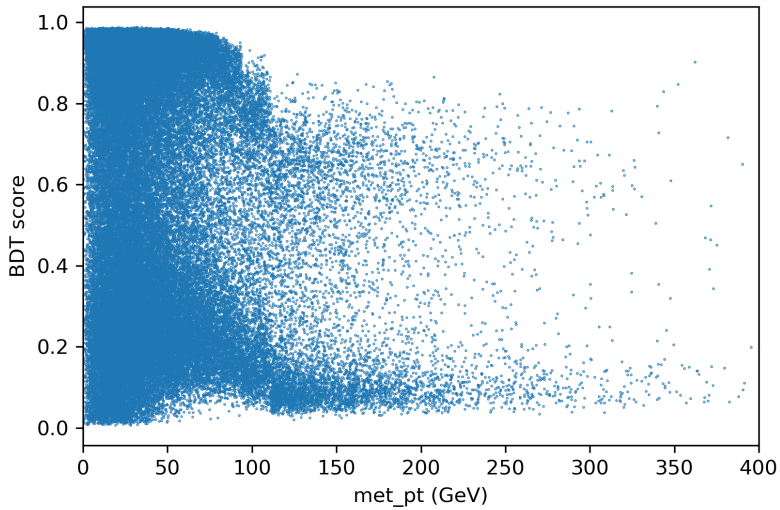


(b) BSM case

met_pt

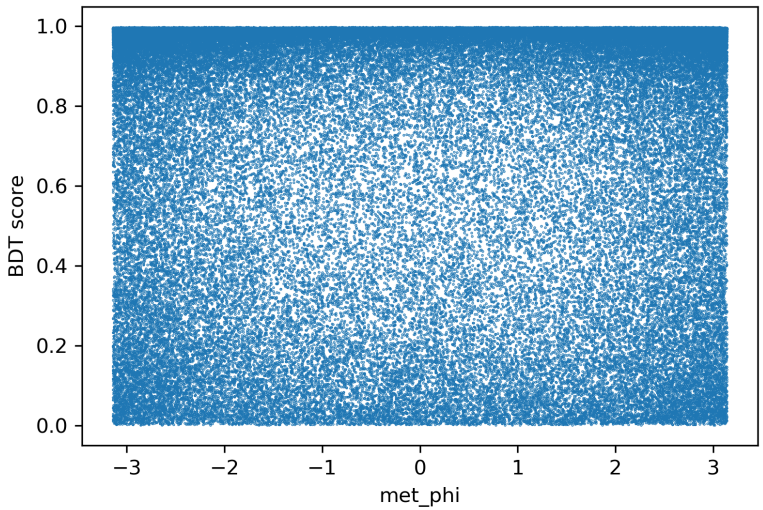


(a) SM case

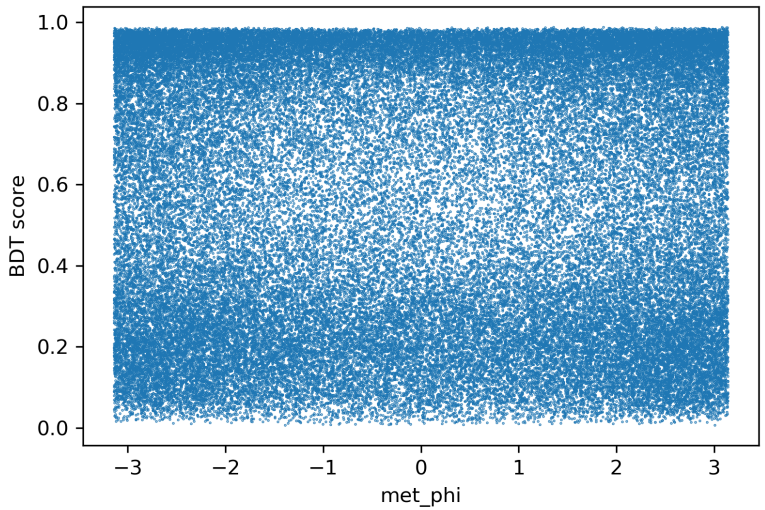


(b) BSM case

met_phi

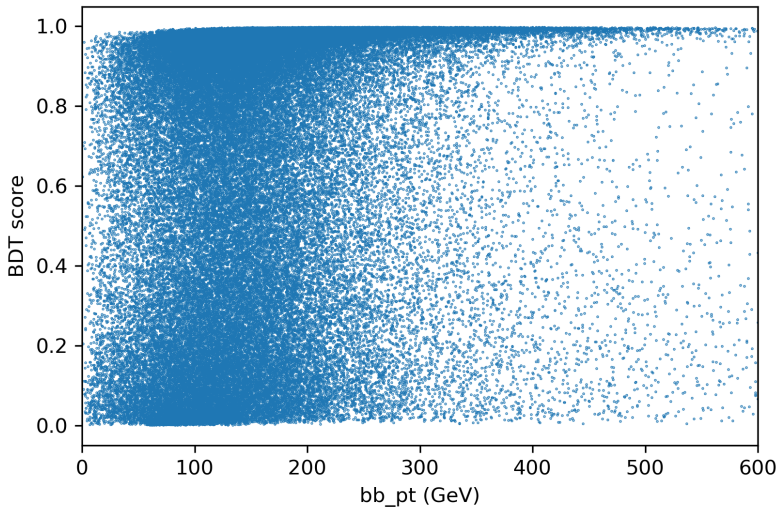


(a) SM case

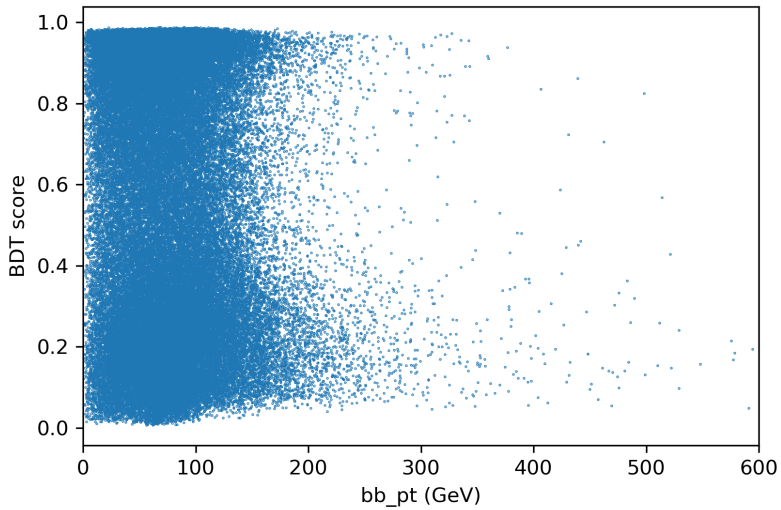


(b) BSM case

bb_pt

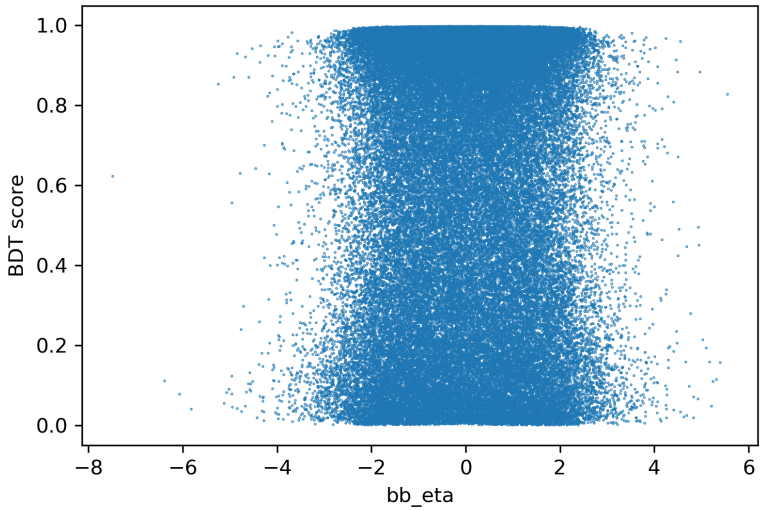


(a) SM case

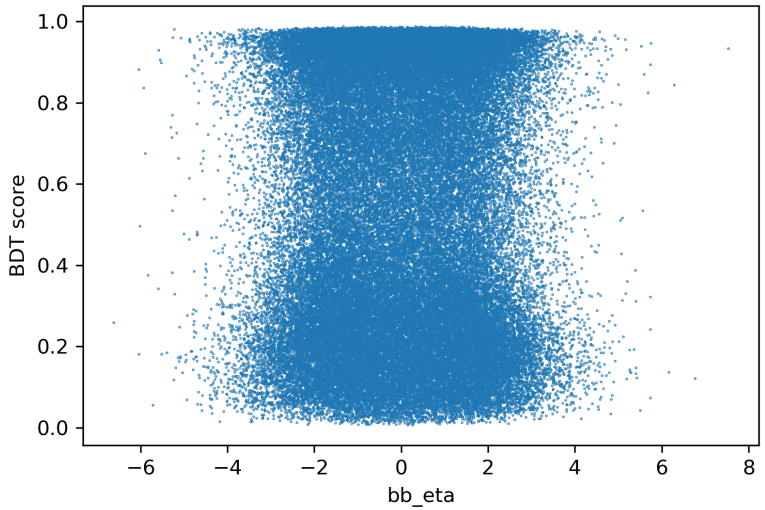


(b) BSM case

bb_eta

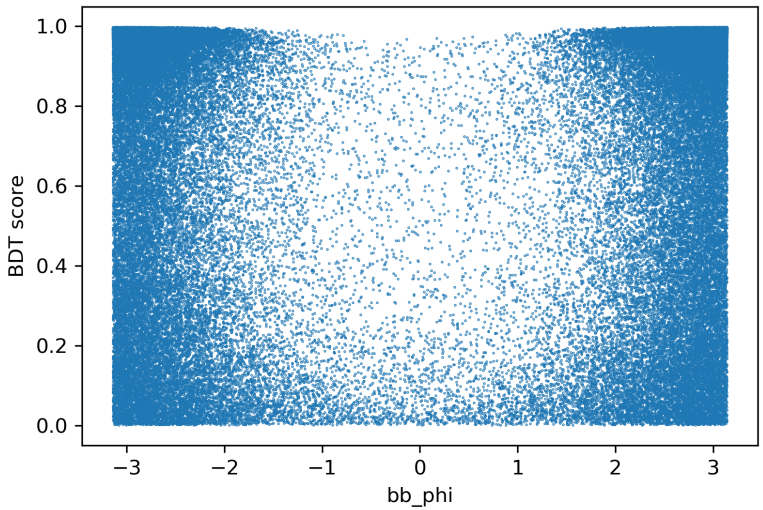


(a) SM case

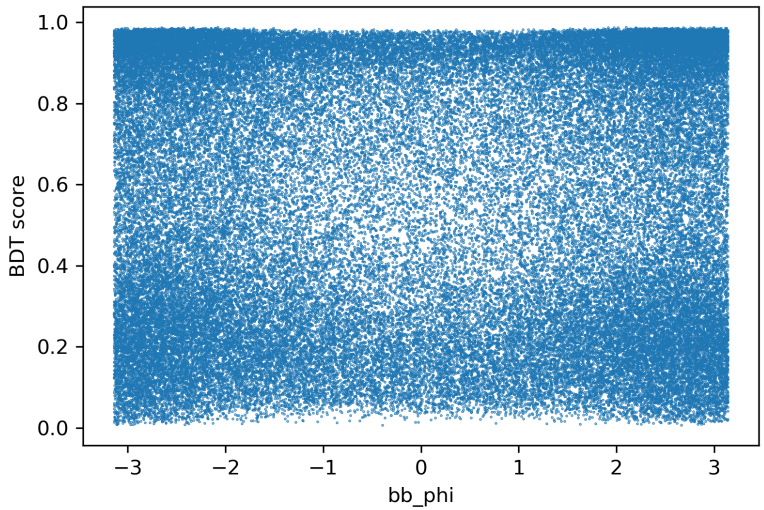


(b) BSM case

bb_phi

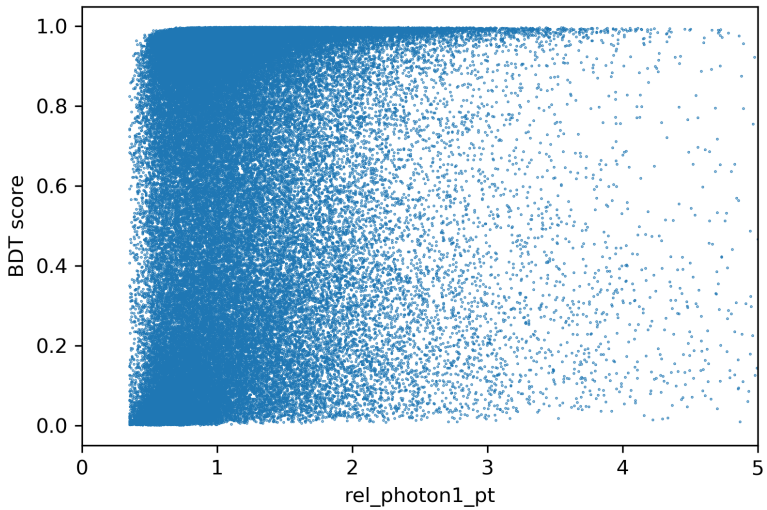


(a) SM case

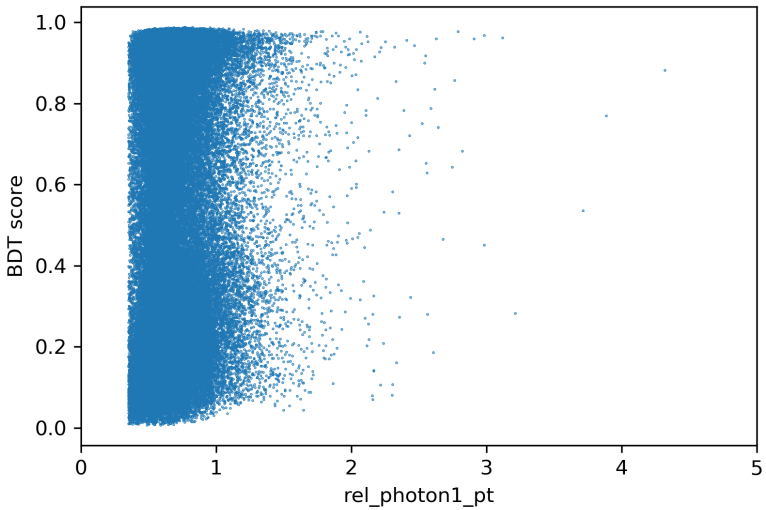


(b) BSM case

rel_photon1_pt

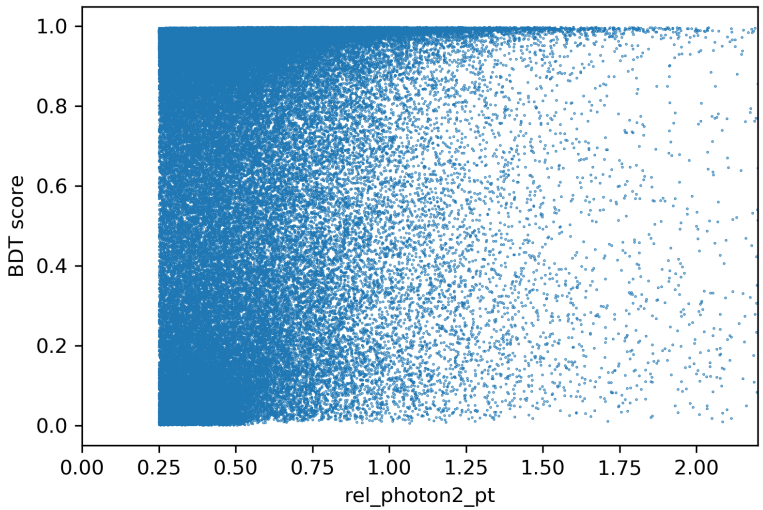


(a) SM case

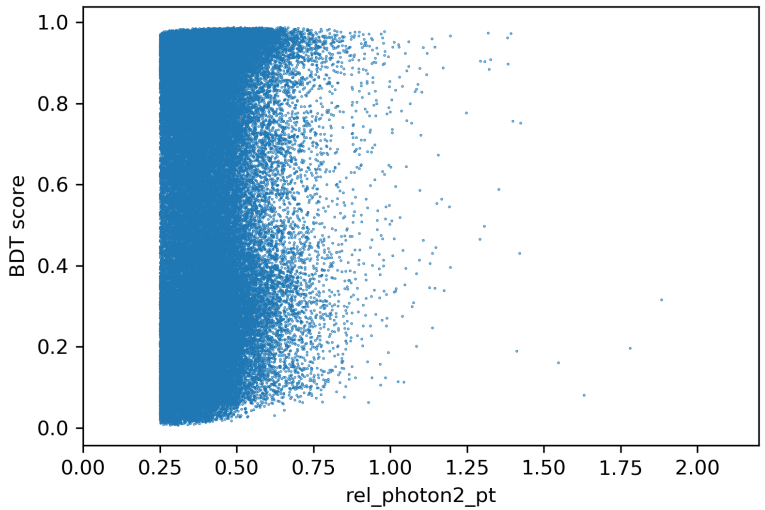


(b) BSM case

rel_photon2_pt

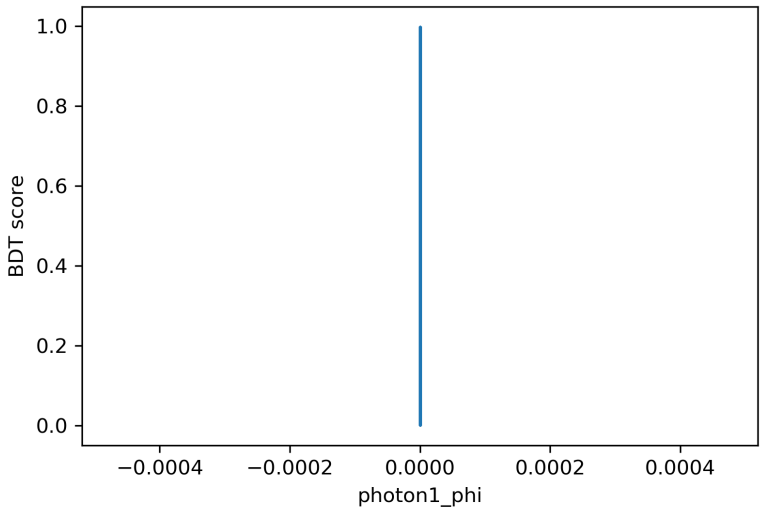


(a) SM case

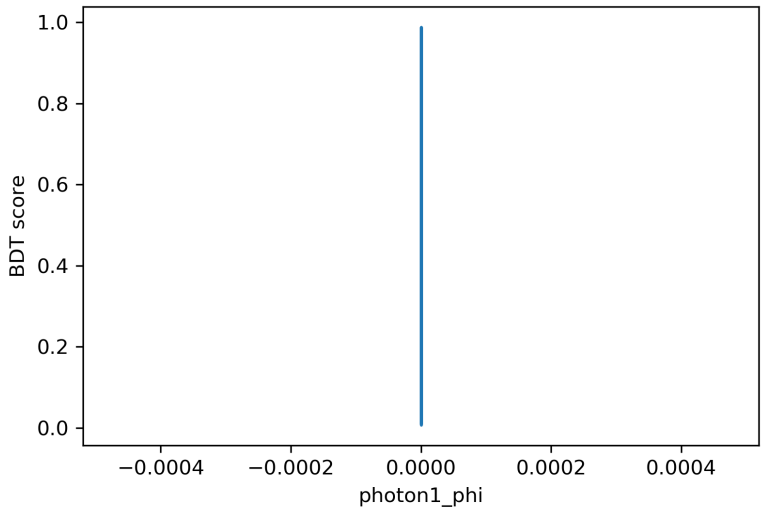


(b) BSM case

photon1_phi

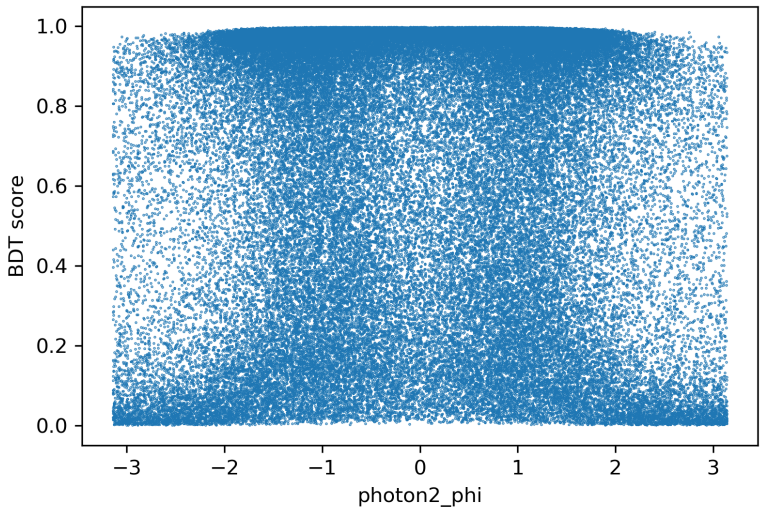


(a) SM case

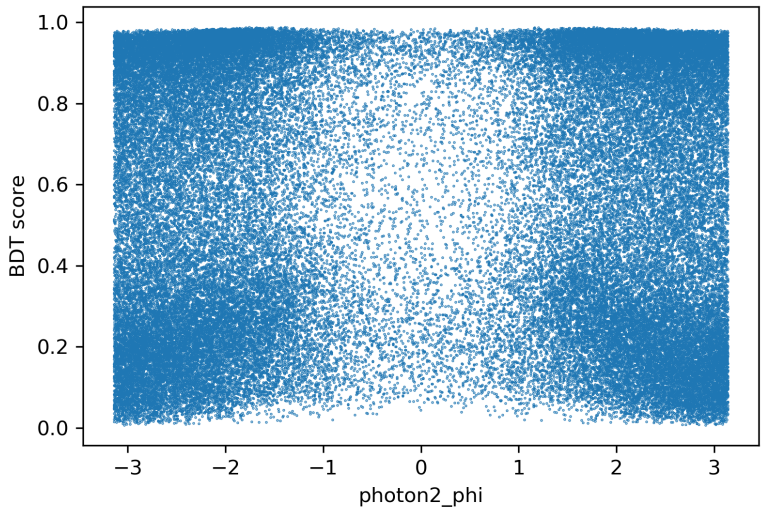


(b) BSM case

photon2_phi

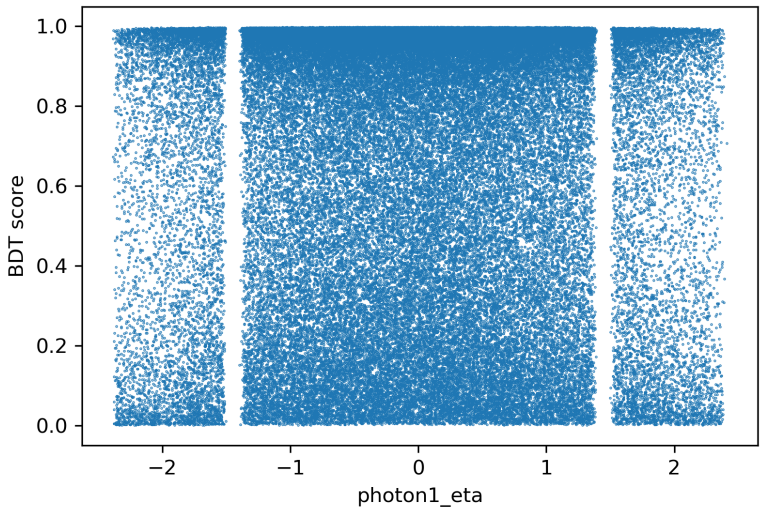


(a) SM case

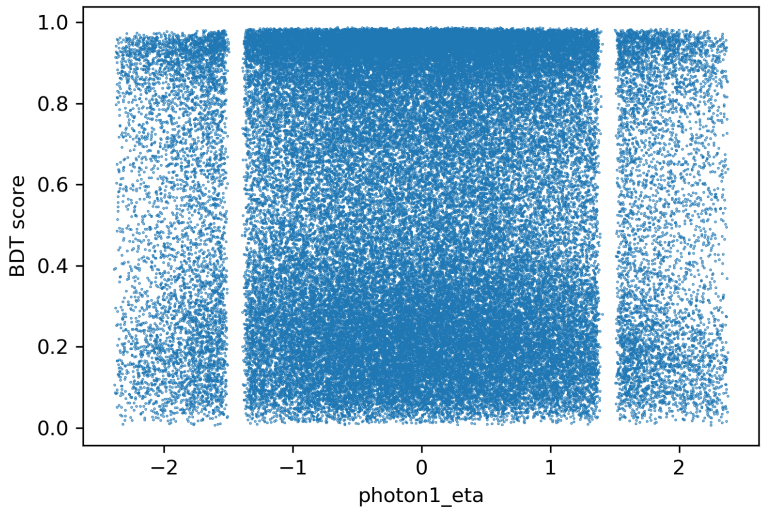


(b) BSM case

photon1_eta

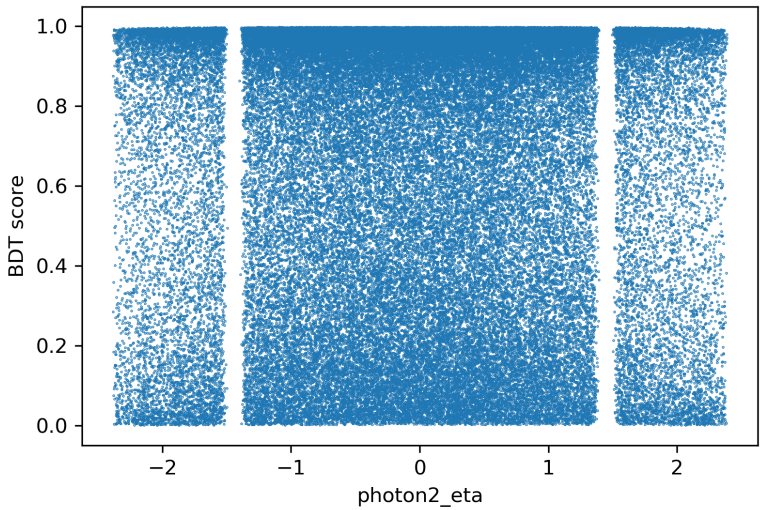


(a) SM case

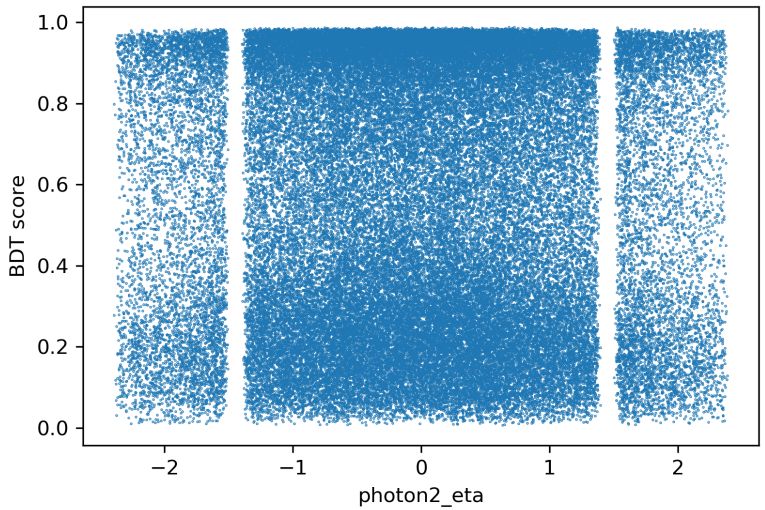


(b) BSM case

photon2_eta

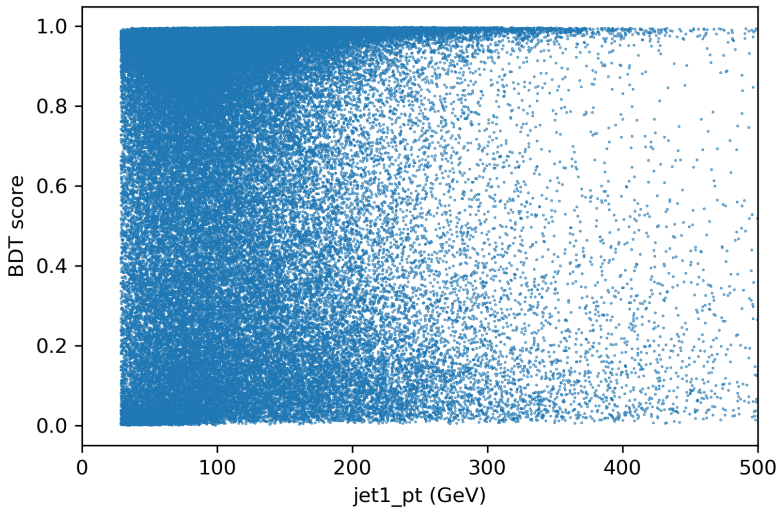


(a) SM case

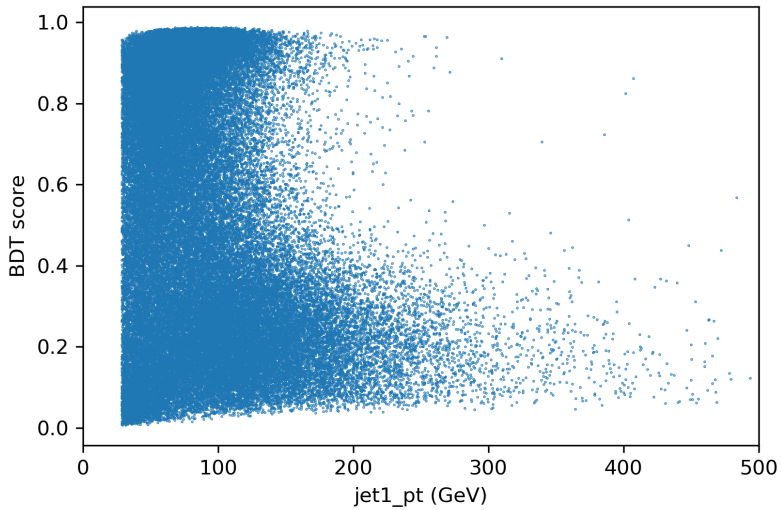


(b) BSM case

jet1_pt

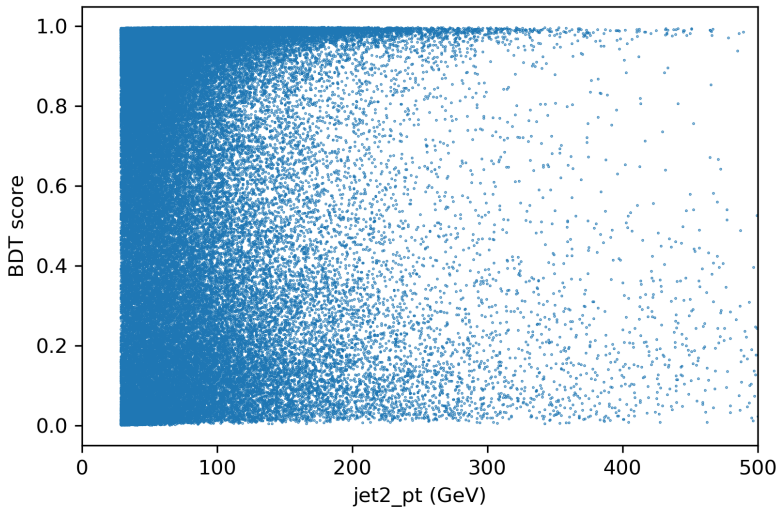


(a) SM case

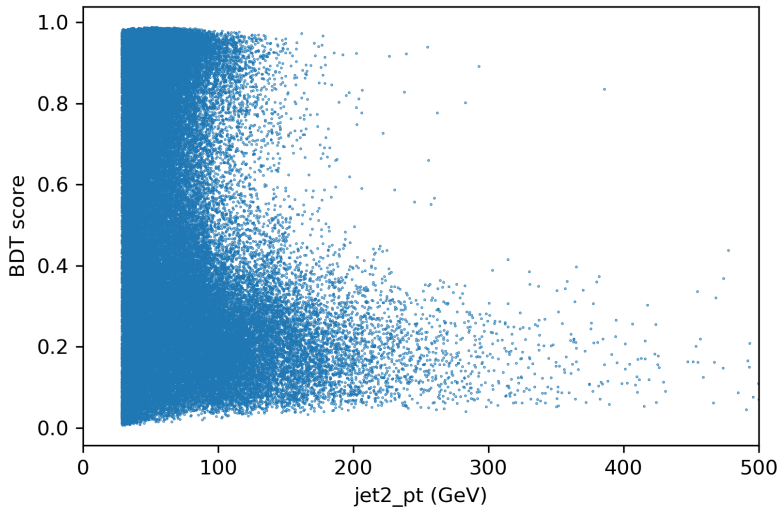


(b) BSM case

jet2_pt

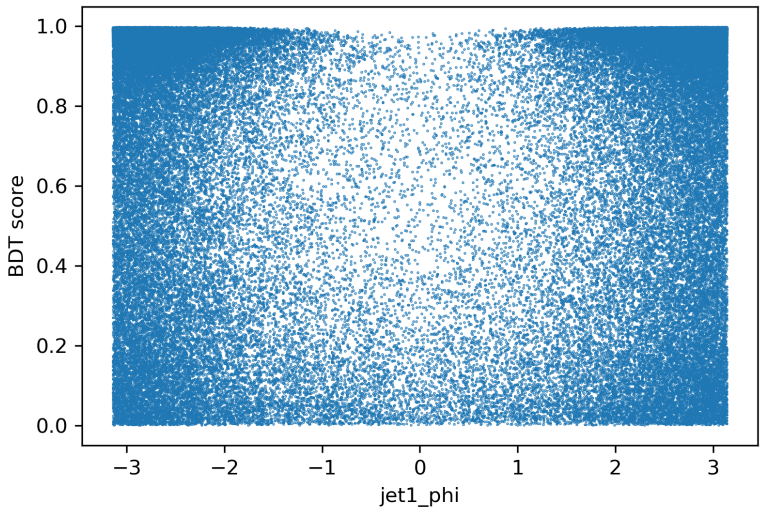


(a) SM case

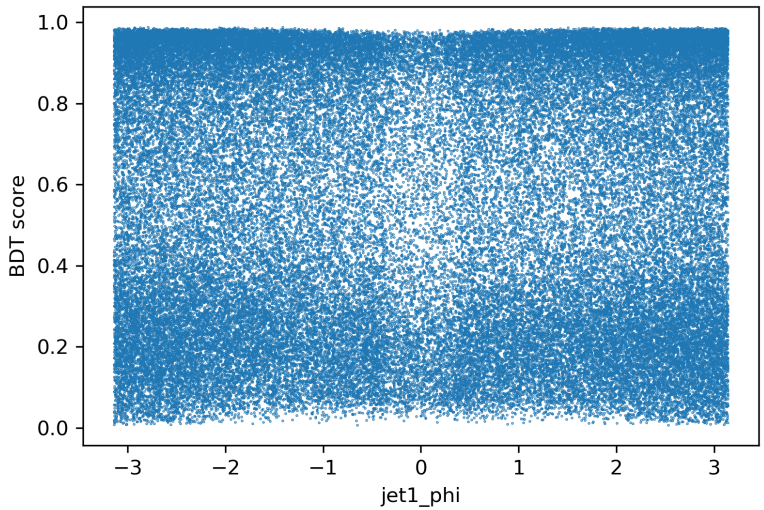


(b) BSM case

jet1_phi

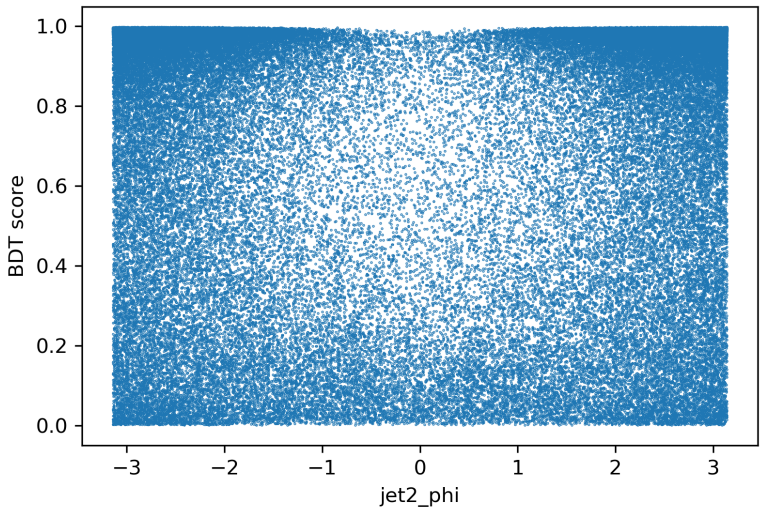


(a) SM case

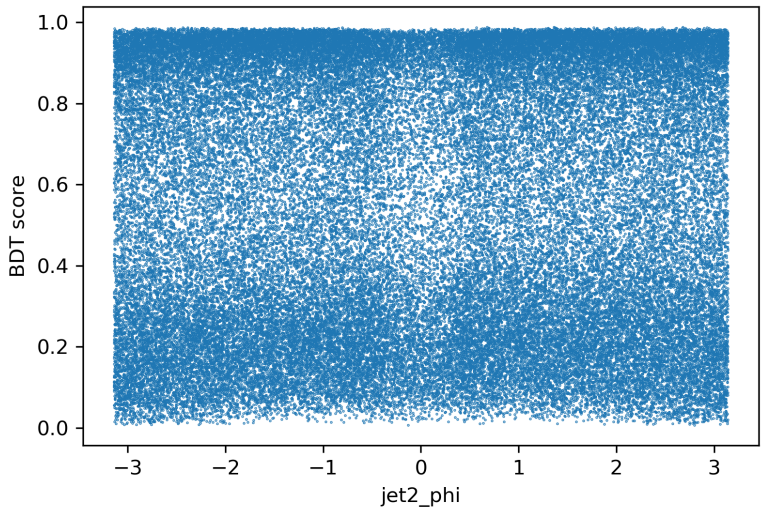


(b) BSM case

jet2_phi

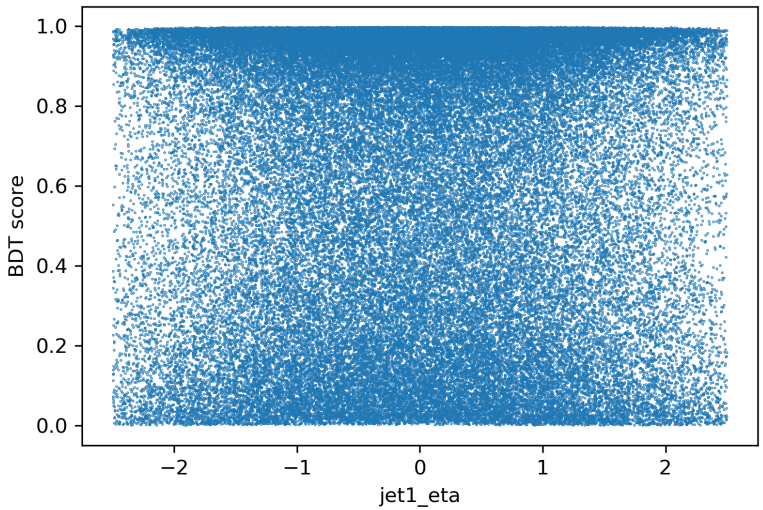


(a) SM case

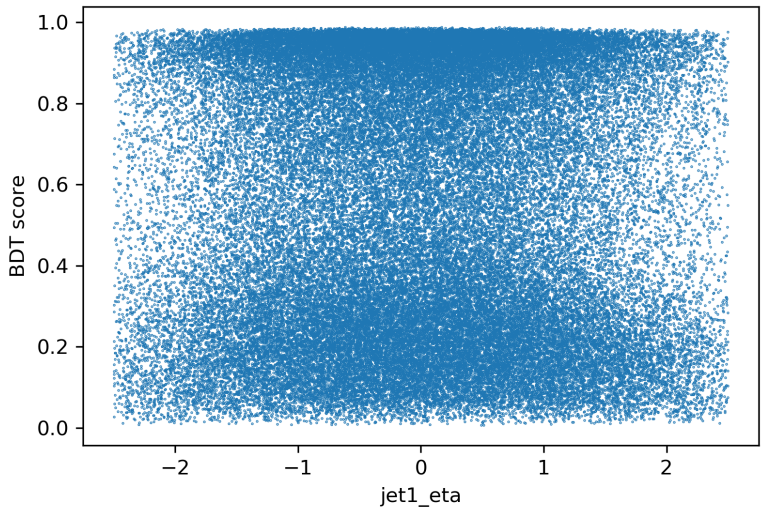


(b) BSM case

jet1_eta

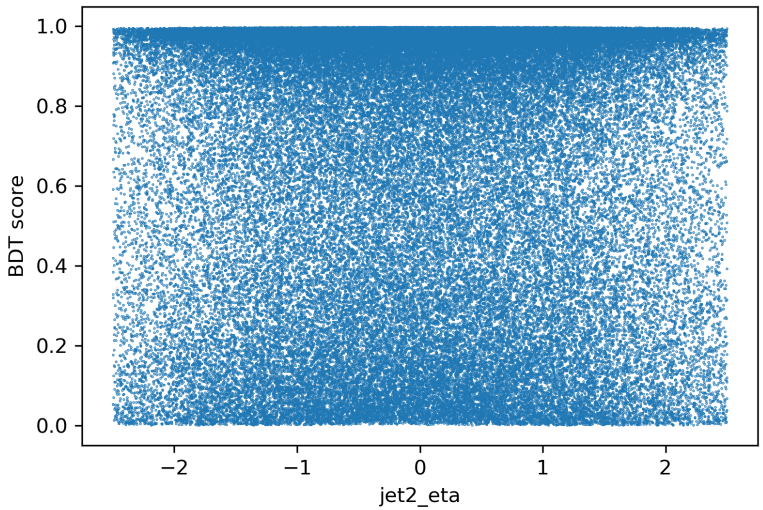


(a) SM case

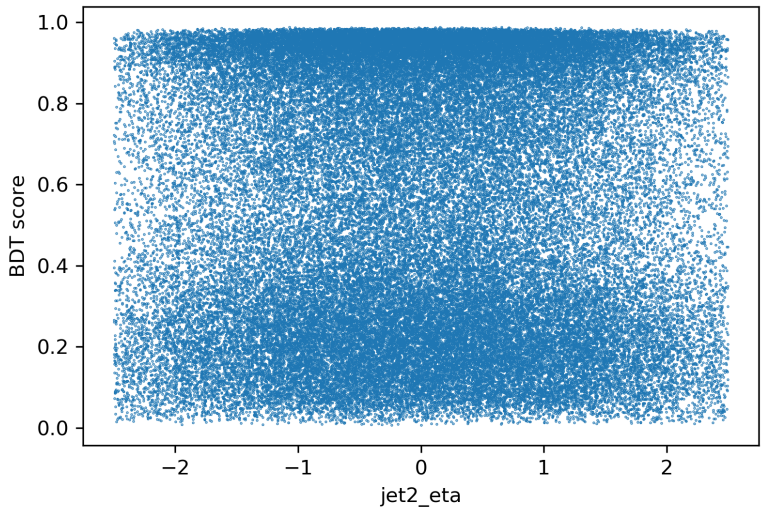


(b) BSM case

jet2_eta

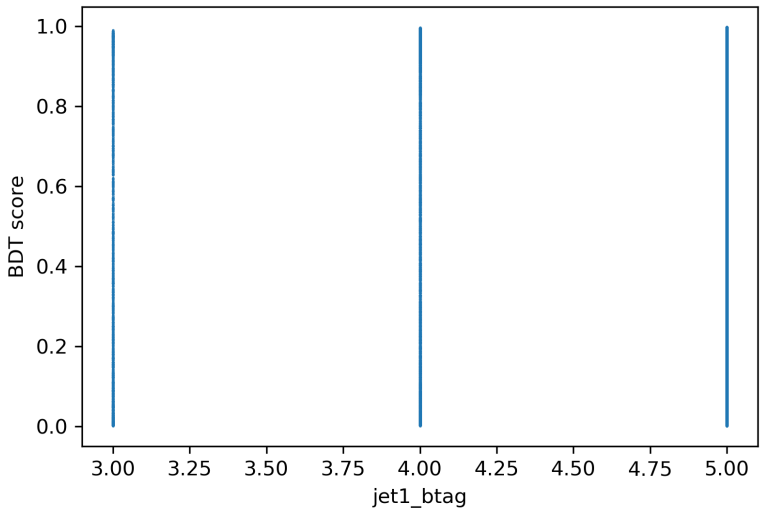


(a) SM case

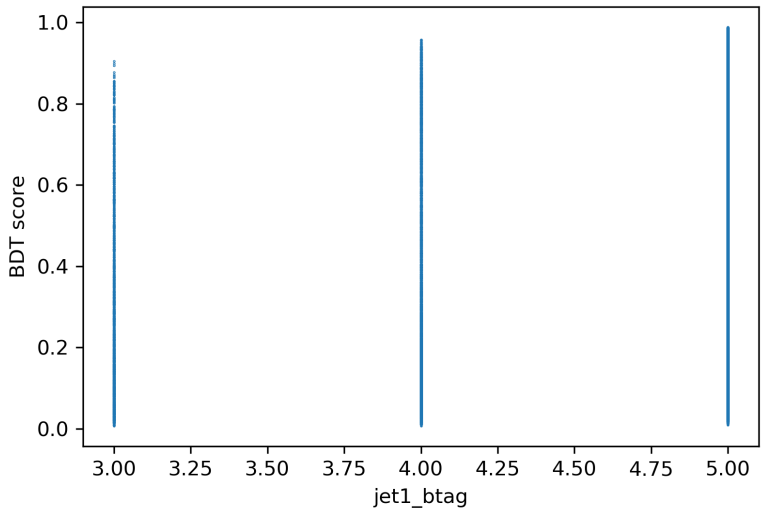


(b) BSM case

jet1_btag

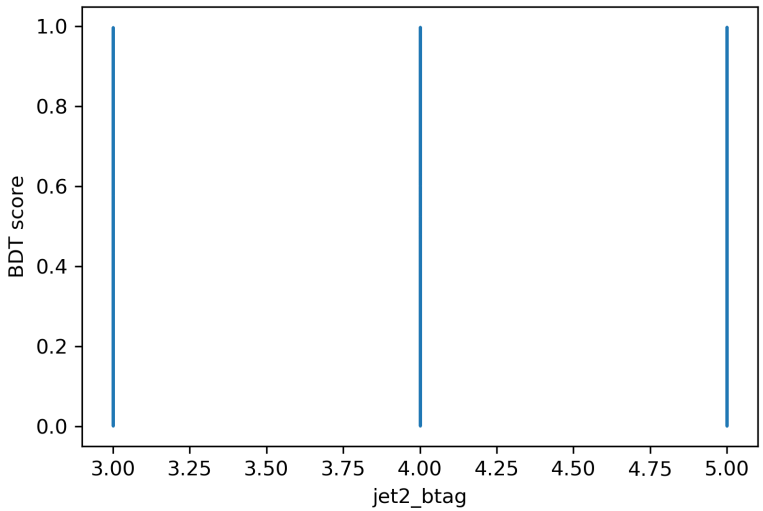


(a) SM case

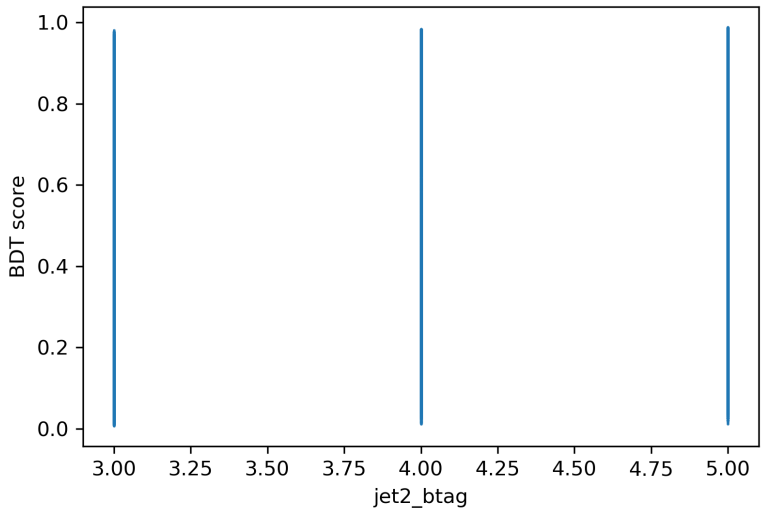


(b) BSM case

jet2_btag



(a) SM case



(b) BSM case