Analysis of Higgs boson decay into four leptons at 8 TeV

Matteo Malucchi

October 5, 2022

1 Introduction

The repository HiggsToFourLeptons-8tev contains an analysis of the decay $H \to ZZ \to 4l$ using reduced NanoAOD files created by CMS Open Data with data from Run 1 at 8 TeV for a total of 11.6 fb^{-1}. The analysis follows loosely the official CMS analysis published in 2012 [Phys. Lett. B 716 (2012) 30]. The main objectives of this analysis are to reconstruct the Higgs boson candidate, measure its mass and train a DNN to classify events as signal (Figure 1a) or background (Figures 1b 1c 1d).

2 Download of the input datasets

The input datasets can be downloaded locally using either multithreading, multiprocessing or sequential computing. Since the overall time depends substantially from the internet speed, it's hard to compare how long does it take to the different strategies to complete the task.

3 Skim the datasets

The *skimming* process consists in reducing the initial samples to a dataset specific for this analysis removing all events which are not of interest for the reconstruction of Z bosons from combinations of leptons ($4e$, 4μ or $2e2\mu$), which may originate from the decay of a Higgs boson. The analysis is done by using `ROOT.RDataFrame` and by invoking `C++` functions thanks to the `ROOT` interpreter after having enabled `ROOT` implicit multithreading. The `C++` functions used in the skimming process are tested using the `unittest` framework. The various cuts applied consist of the requests that there are exactly 4 high P_t leptons the charge sum of which is equal to zero, that they are isolated in the detector and far form each other. Furthermore, they are requested to originate from the primary vertex and to be produced centrally.

All the variables used later on are defined in this step, this includes mass, P_t, η and ϕ of Z and Higgs bosons, as well as the five decay angles formed by the leptons in the final state, as described in detail in the article [Phys.Rev.D86:095031,2012], which are defined as shown in Figure 2.

Let us consider, of all the possible combinations of same-flavor opposite-sign lepton pairs, the one that has invariant mass closest to $m_Z = 91.2$ GeV and the other one formed by the remaining pair. Of the two reconstructed Z bosons that originate from such lepton pairs, the one with highest mass is called Z_1 and the another one Z_2. The three-momentum of the Z_i boson is called $\mathbf{q}_i$, while $\mathbf{q}_{i1}$ and $\mathbf{q}_{i2}$ indicate respectively the three-momenta of the lepton and anti-lepton associated with Z_i. Indicating as superscript the rest frame in which the three-momenta are taken, the definitions of the angles are:

- $\theta^* \in [0, \pi]$

$$\cos\theta^* = \frac{q^{4l}_{1z}}{|\mathbf{q}^{4l}_1|}$$

- $\Phi_1 \in [-\pi, \pi]$

$$\Phi_1 = \frac{\mathbf{q}^{4l}_1 \cdot (\hat{n}_1 \times \hat{n}_{coll})}{|\mathbf{q}^{4l}_1 \cdot (\hat{n}_1 \times \hat{n}_{coll})|} \times \arccos\left(\hat{n}_1 \cdot \hat{n}_{coll}\right)$$

where

$$\hat{n}_z = (0, 0, 1), \qquad \hat{n}_1 = \frac{\mathbf{q}^{4l}_{11} \times \mathbf{q}^{4l}_{12}}{|\mathbf{q}^{4l}_{11} \times \mathbf{q}^{4l}_{12}|}, \qquad \hat{n}_{coll} = \frac{\hat{n}_z \times \mathbf{q}^{4l}_1}{|\hat{n}_z \times \mathbf{q}^{4l}_1|}$$

- $\Phi \in [-\pi, \pi]$

$$\Phi = \frac{\mathbf{q}^{4l}_1 \cdot (\hat{n}_1 \times \hat{n}_2)}{|\mathbf{q}^{4l}_1 \cdot (\hat{n}_1 \times \hat{n}_2)|} \times \arccos\left(-\hat{n}_1 \cdot \hat{n}_2\right)$$

where

$$\hat{n}_2 = \frac{\mathbf{q}_{21}^{4l} \times \mathbf{q}_{22}^{4l}}{|\mathbf{q}_{21}^{4l} \times \mathbf{q}_{22}^{4l}|}$$

- $\theta_1 \in [0, \pi]$, $\theta_2 \in [0, \pi]$

$$\cos\theta_1 = -\frac{\mathbf{q}_2^{Z_1} \cdot \mathbf{q}_{11}^{Z_1}}{|\mathbf{q}_2^{Z_1}||\mathbf{q}_{11}^{Z_1}|}, \qquad \cos\theta_2 = -\frac{\mathbf{q}_1^{Z_2} \cdot \mathbf{q}_{21}^{Z_2}}{|\mathbf{q}_1^{Z_2}||\mathbf{q}_{21}^{Z_2}|}$$

4 Machine learning

In this step a Deep Neural Network is trained on the Monte Carlo samples in order to create a discriminant that allows the classification of events as either signal or background. The machine learning model is built thanks to the `ROOT.TMVA` library with `Keras` API, the high-level API of `TensorFlow`. The training is done using as discriminating variables the ones shown in Figure 3: the masses of the Z bosons (Mass Z_1 and Mass Z_2) and the angles $\cos\theta^*$, Φ, Φ_1, $\cos\theta_1$ and $\cos\theta_2$. The plots of the normalized signal and background show clearly how the variables used as input for the DNN are distributed differently for signal or background. The correlation matrices of such variables are shown in Figure 4. Before being used as input to the DNN, the variables are *decorellated* and *gaussianized* (Figure 4). The input data were split randomly using half the dataset for training and half for testing and were normalized using `NormMode=NumEvents`.

The model used for the DNN is shown in Figure 6, where the hidden layers activation function is a `relu`, while the activation function for the output layer is a `softmax`. The loss used is `categorical_crossentropy`, the optimizer is `adam` and the metrics is `accuracy`.

The DNN was trained for 20 epochs with a batch-size of 128. The training history, the `ROC` curve and the classifier output distribution are shown respectively in Figures 7a 7b 7c. The trained DNN is evaluated on the whole dataset and the resulting values of the discriminant is used to classify the events as signal or background depending on whether the DNN Discriminant is above or below the threshold, which is set to the optimal cut value with highest $S/\sqrt{S^2 + B^2}$ ratio as seen in Figure 7d.

5 Histograms and plots

For each variable saved in the skimming process an histogram. Later, plots of these variables are produced for various sample combinations: only the signal, only the background, only the data, the combination of all the above (with signal and background stacked on top of each other) and also the combination of normalized signal and background (as in Figure 3). Each variable is also plotted for each possible final state and for their combination. Each plot is created both for all events and for only those events that have the DNN Discriminant above threshold. Some examples of plots with data, signal and background considering all the possible final states are shown in Figures 8, 9 and 10 both with and without the selection based on the DNN Discriminant. The five angles are shown in Figure 11 for the combination of all datasets.

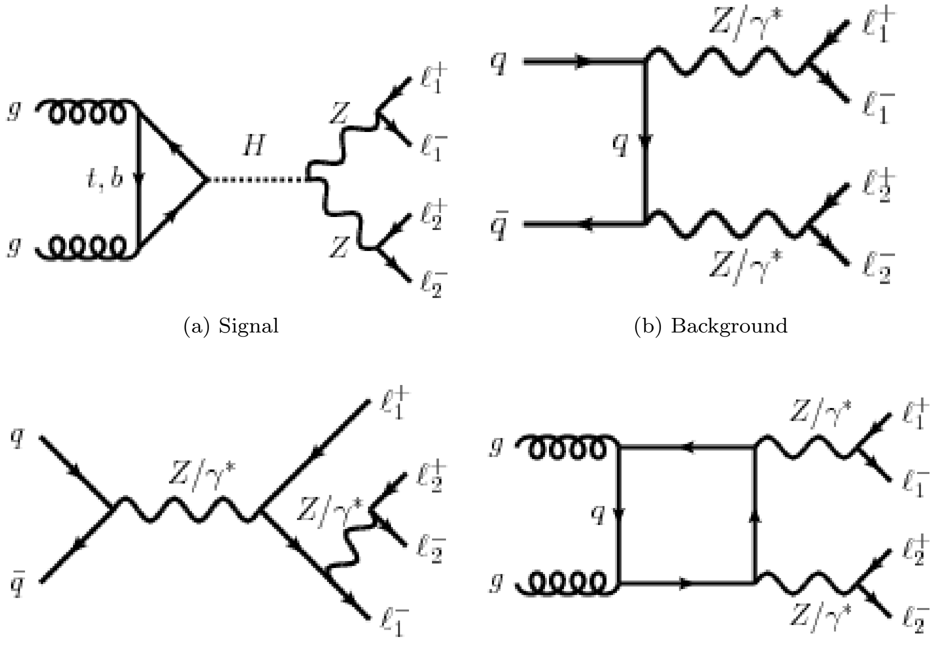
Subsequently, 2D distributions of the invariant mass of the 4 leptons system VS DNN Discriminant are created and shown in Figure 12 for the simulated signal and for the simulated background. Each plot contains both the combination of all background/signal datasets and the real data separated in the three possible final states.

6 Higgs mass fit

Finally, the invariant mass of the four leptons is fitted with a Crystal Ball using the `ROOT.RooFit` library. Once again this is done both for all events and for only those events with DNN Discriminant above threshold, as shown in Figure 13. The estimates of the mass of the Higgs boson are shown in Table 1.

	Without DNN selection	With DNN selection
Higgs mass from MC [GeV]	124.934 ± 0.016	124.938 ± 0.015
Higgs mass from data [GeV]	125.40 ± 0.26	125.39 ± 0.30

Table 1: Higgs mass values yield by the fits.



(a) Signal

(b) Background

(c) Background

(d) Background

Figure 1: Feynman diagrams of signal and background.

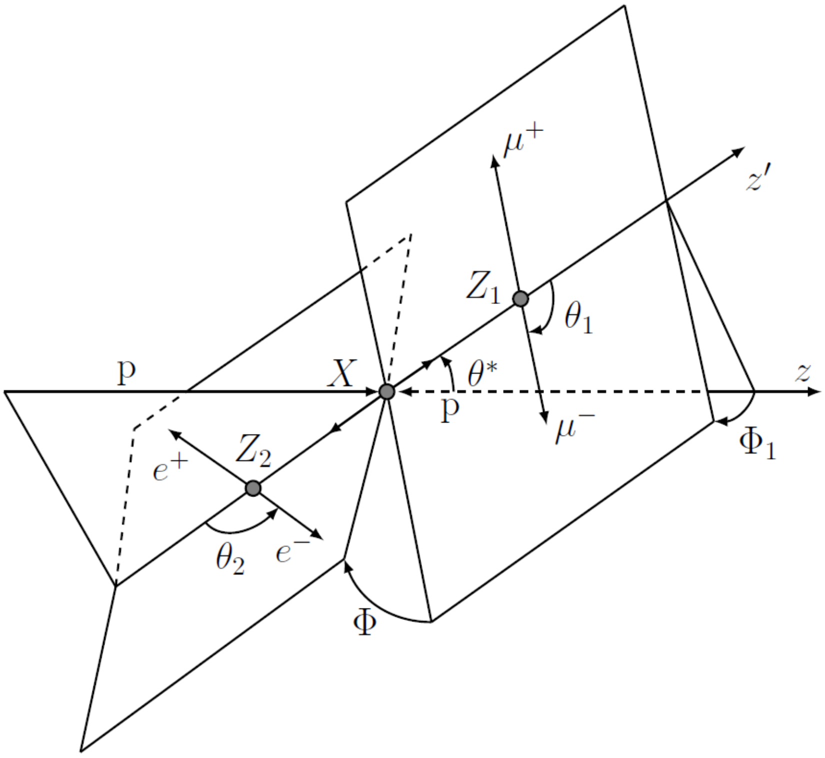
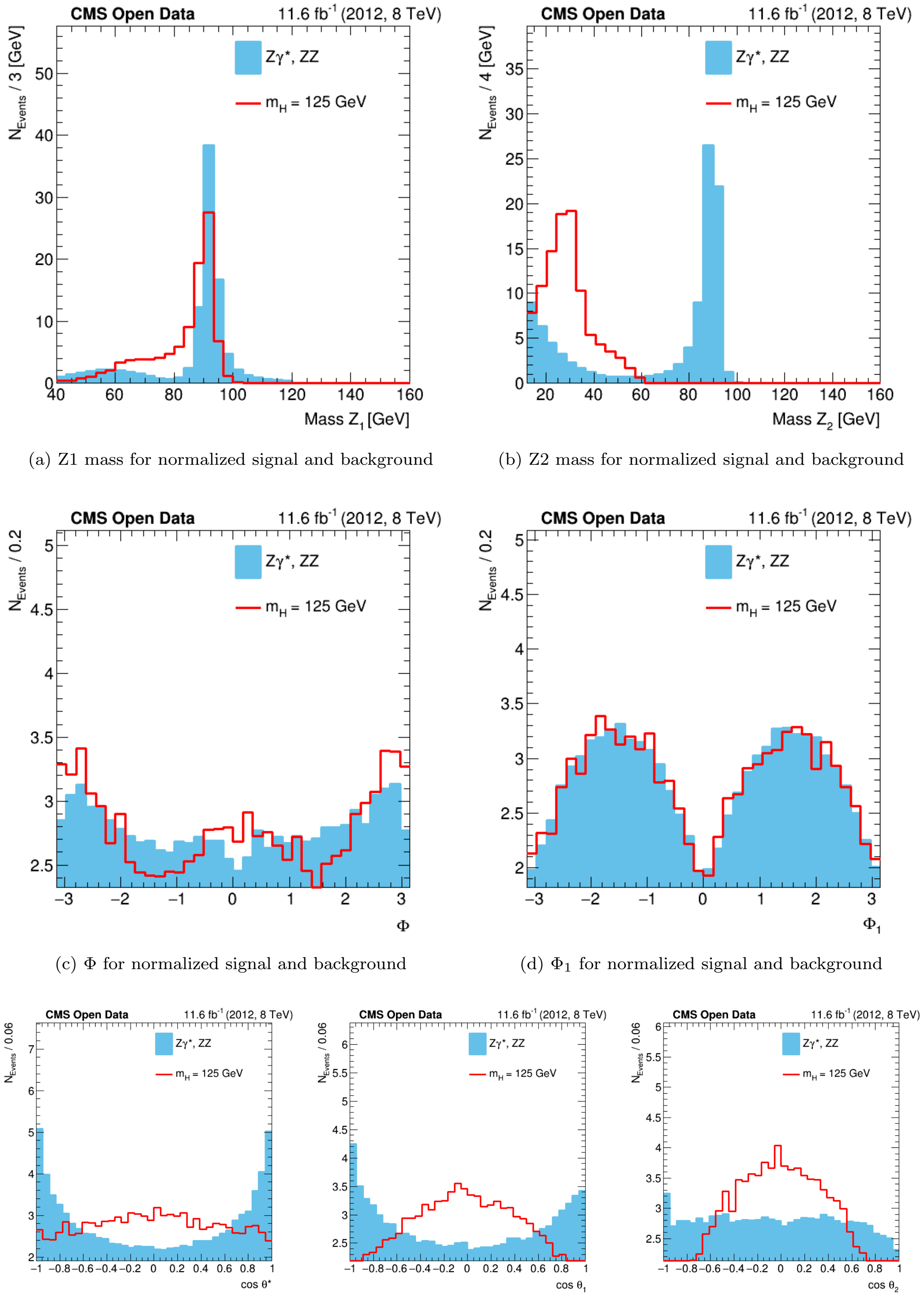


Figure 2: Definition of the decay angles formed by the final state leptons.



(a) Z1 mass for normalized signal and background

(b) Z2 mass for normalized signal and background

(c) Φ for normalized signal and background

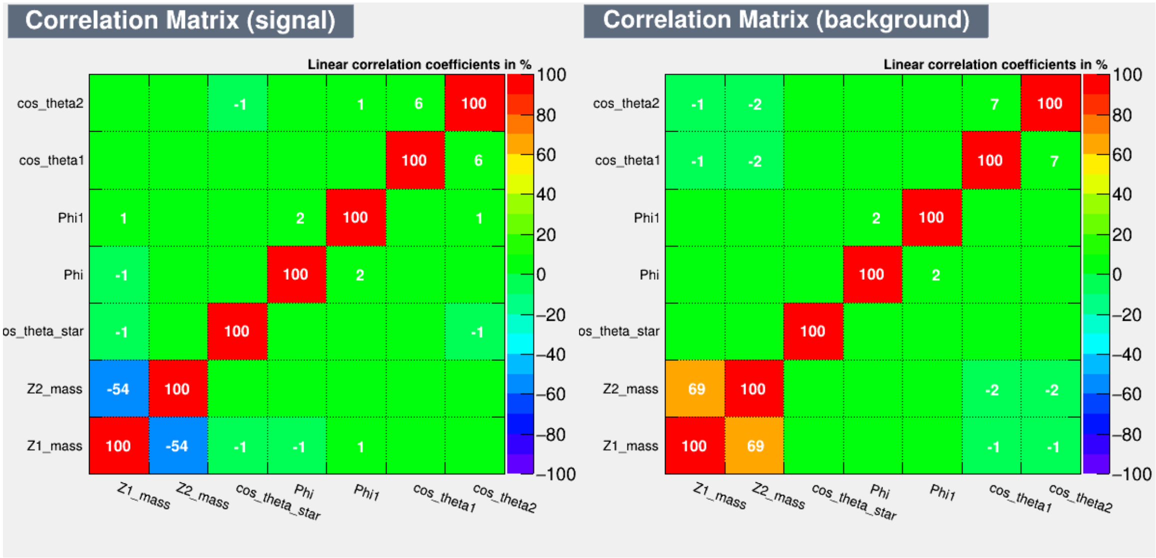
(d) Φ_1 for normalized signal and background

(e) $\cos\theta^*$ for normalized signal and background

(f) $\cos\theta_1$ for normalized signal and background

(g) $\cos\theta_2$ for normalized signal and background

Figure 3: Distributions of the variables used for the training of the DNN shown for the normalized signal and background samples.



(a) Signal correlation matrix

(b) Background correlation matrix

Figure 4: Correlation matrices of the DNN input variables for signal and background.

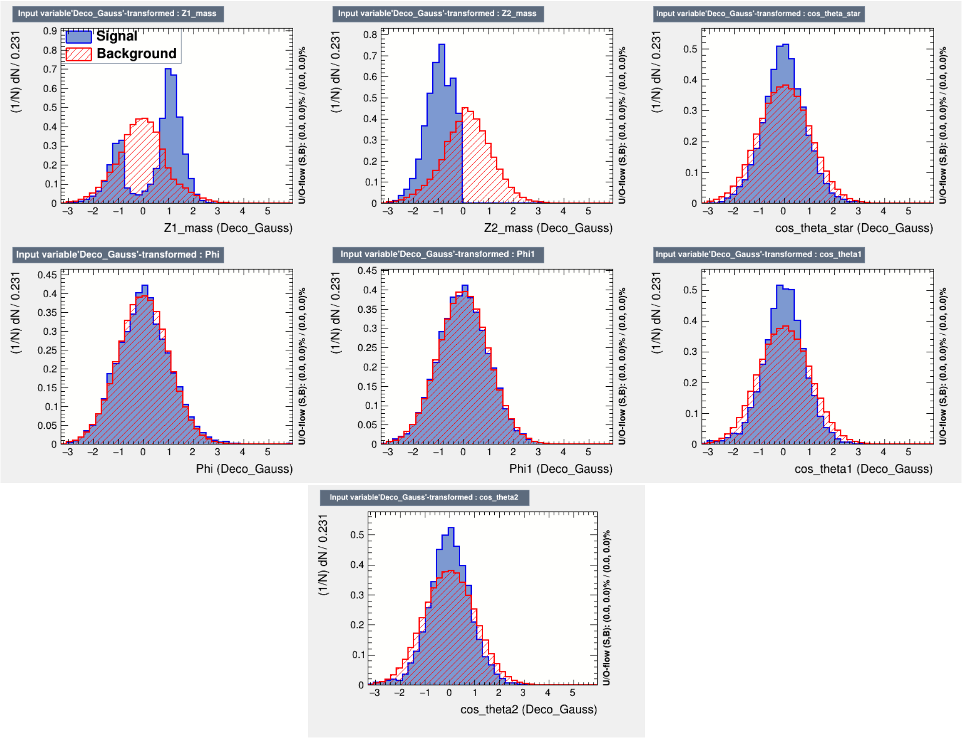
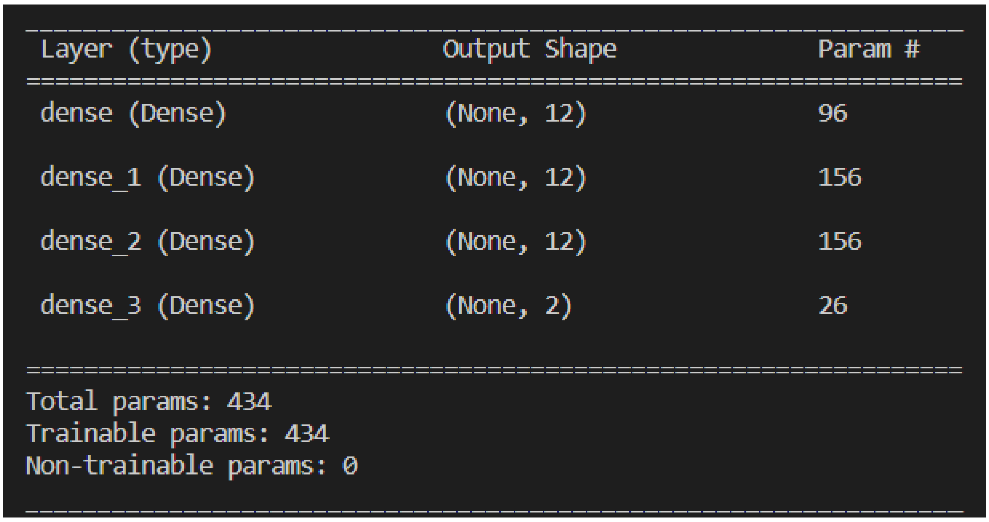
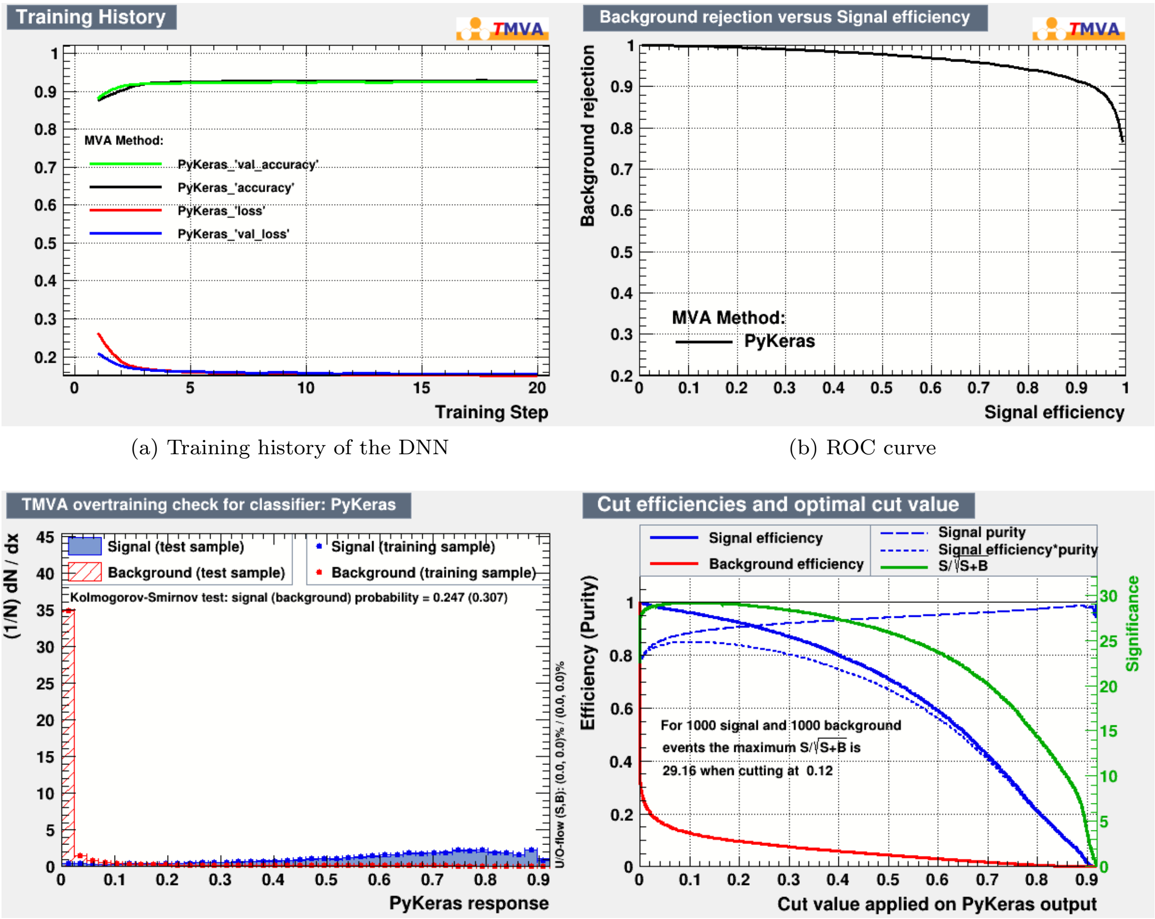


Figure 5: *Decorellated-gaussianized* variable distributions.



```
______________________________________________________________________________
 Layer (type)                Output Shape              Param #
==============================================================================
 dense (Dense)               (None, 12)                96

 dense_1 (Dense)             (None, 12)                156

 dense_2 (Dense)             (None, 12)                156

 dense_3 (Dense)             (None, 2)                 26

==============================================================================
Total params: 434
Trainable params: 434
Non-trainable params: 0
______________________________________________________________________________
```

Figure 6: DNN model used for the training.



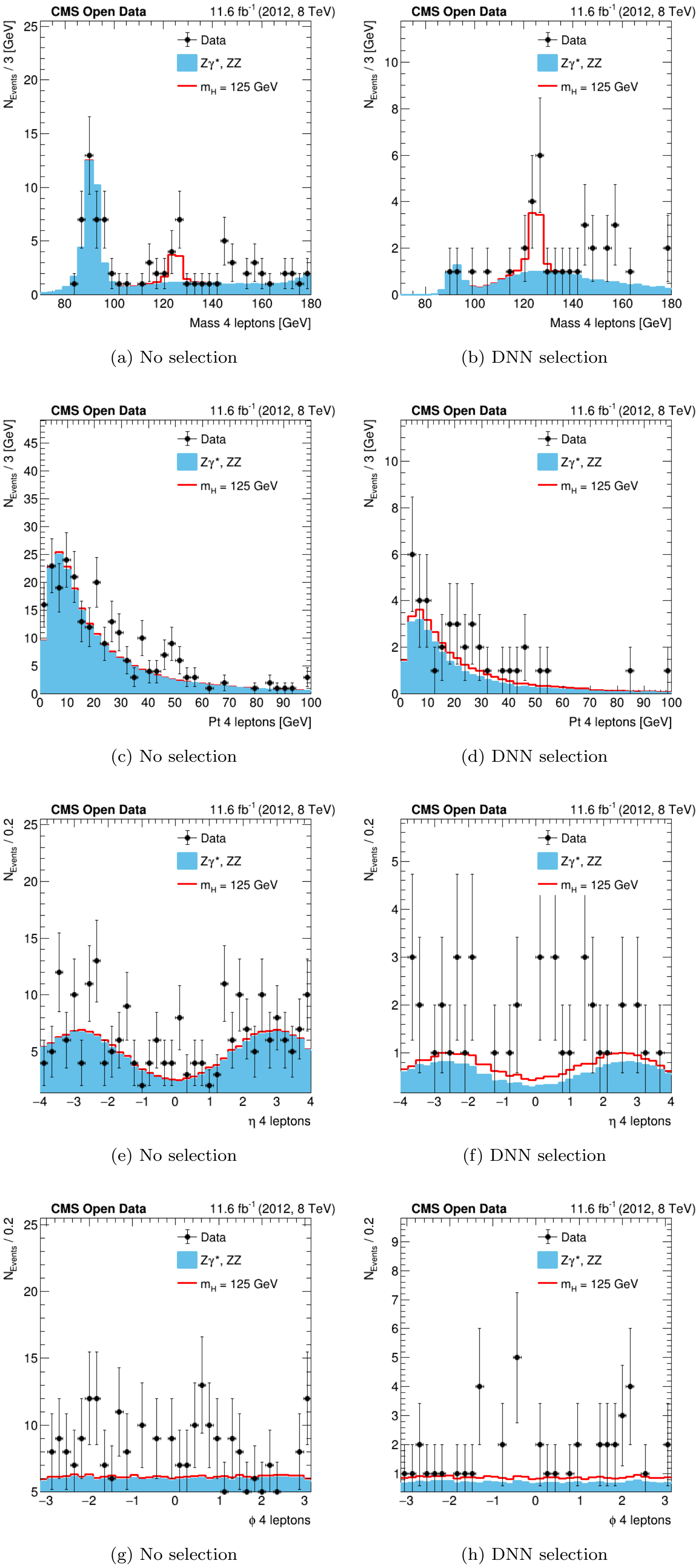
(a) Training history of the DNN

(b) ROC curve

(c) Classifier output distribution

(d) Cut efficiencies

Figure 7: Plots created during the training of the DNN.



(a) No selection

(b) DNN selection

(c) No selection

(d) DNN selection

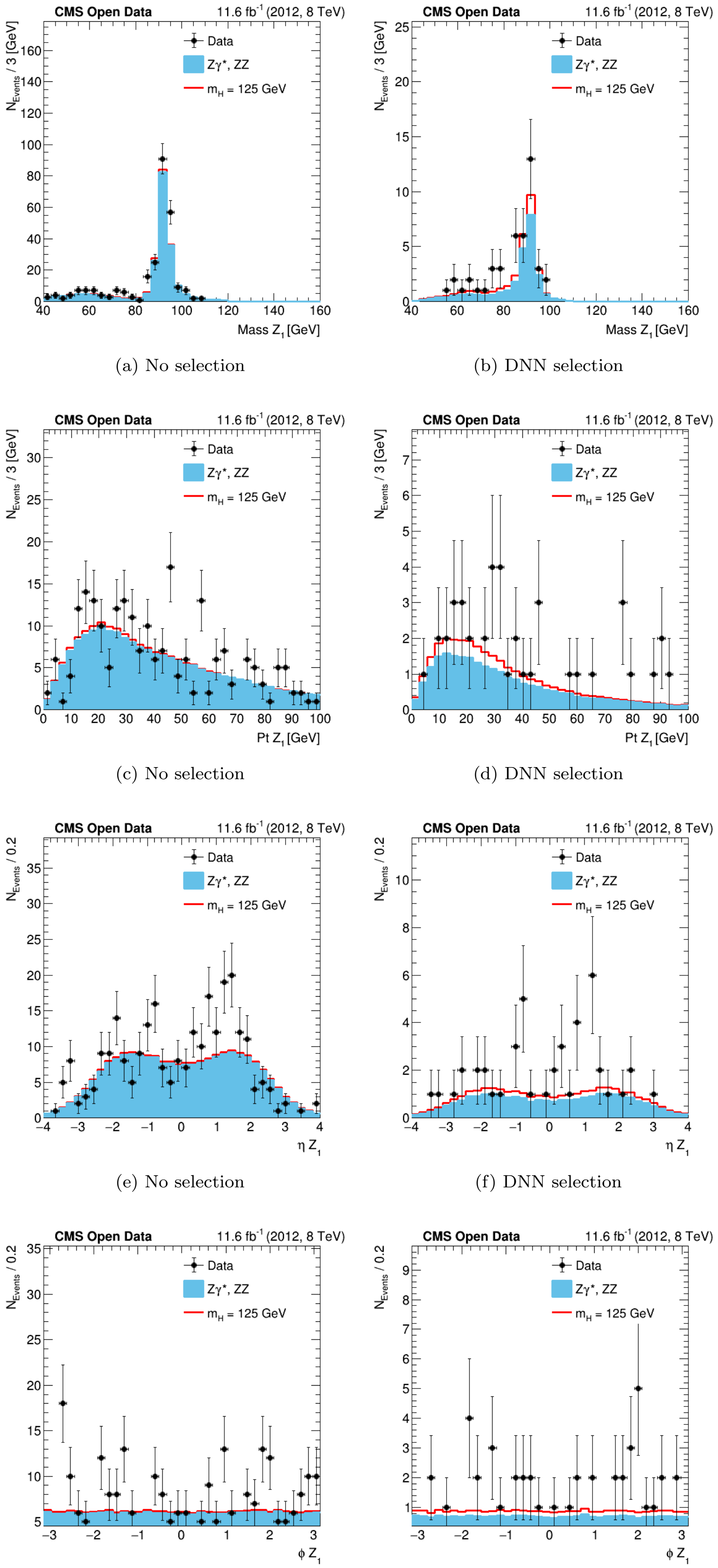
(e) No selection

(f) DNN selection

(g) No selection

(h) DNN selection

Figure 8: Distributions for the Higgs candidate with and without the DNN selection.



(a) No selection

(b) DNN selection

(c) No selection

(d) DNN selection

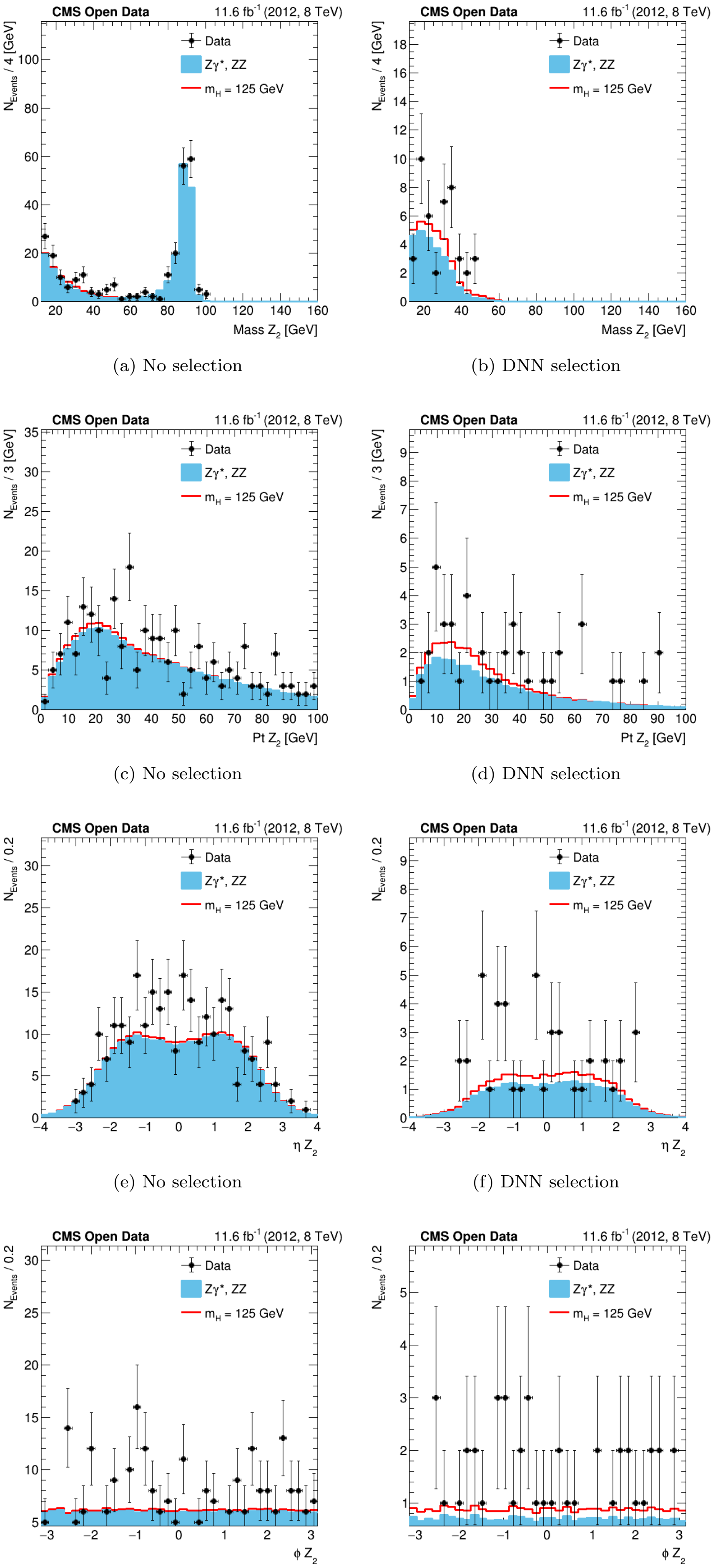
(e) No selection

(f) DNN selection

(g) No selection

(h) DNN selection

Figure 9: Distributions for the Z1 candidate with and without the DNN selection.



(a) No selection

(b) DNN selection

(c) No selection

(d) DNN selection

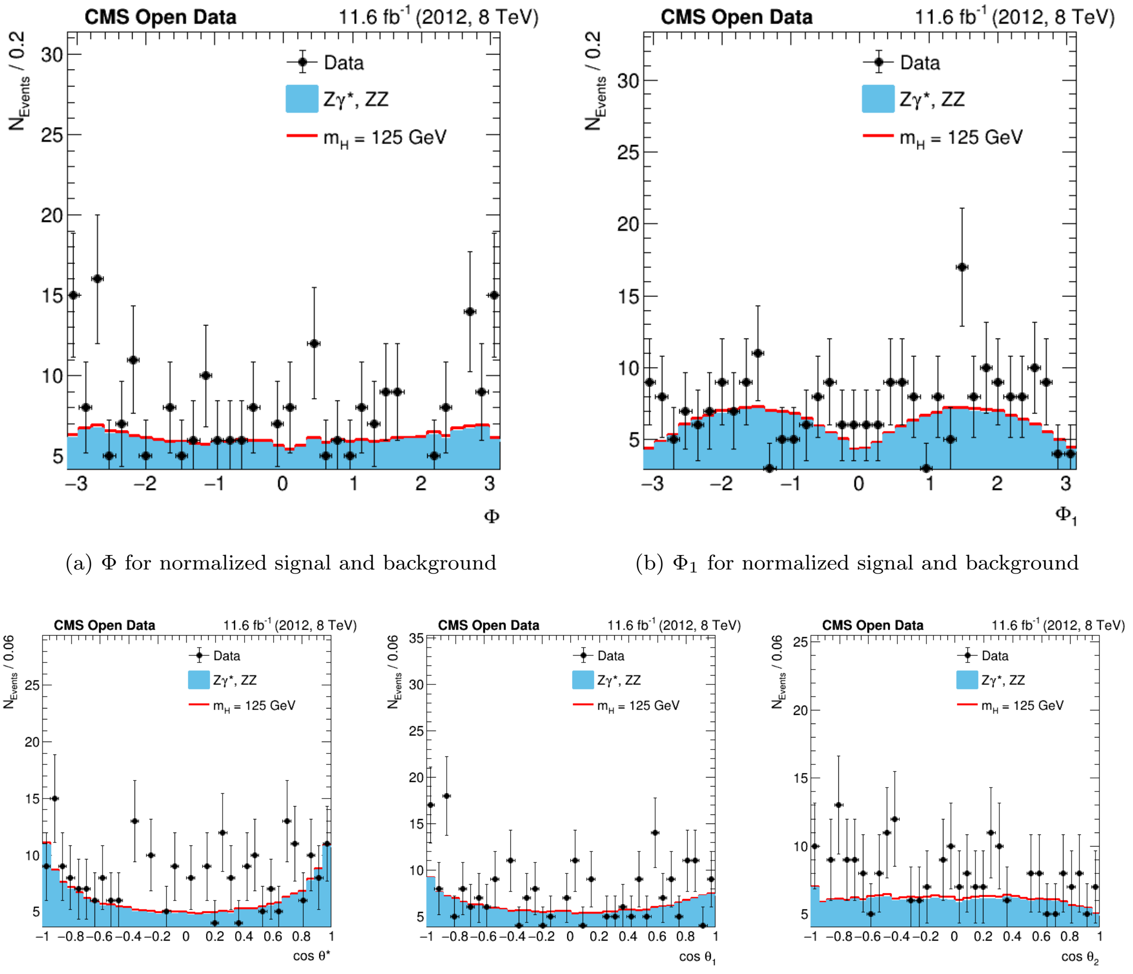
(e) No selection

(f) DNN selection

(g) No selection

(h) DNN selection

Figure 10: Distributions for the Z2 candidate with and without the DNN selection.



(a) Φ for normalized signal and background

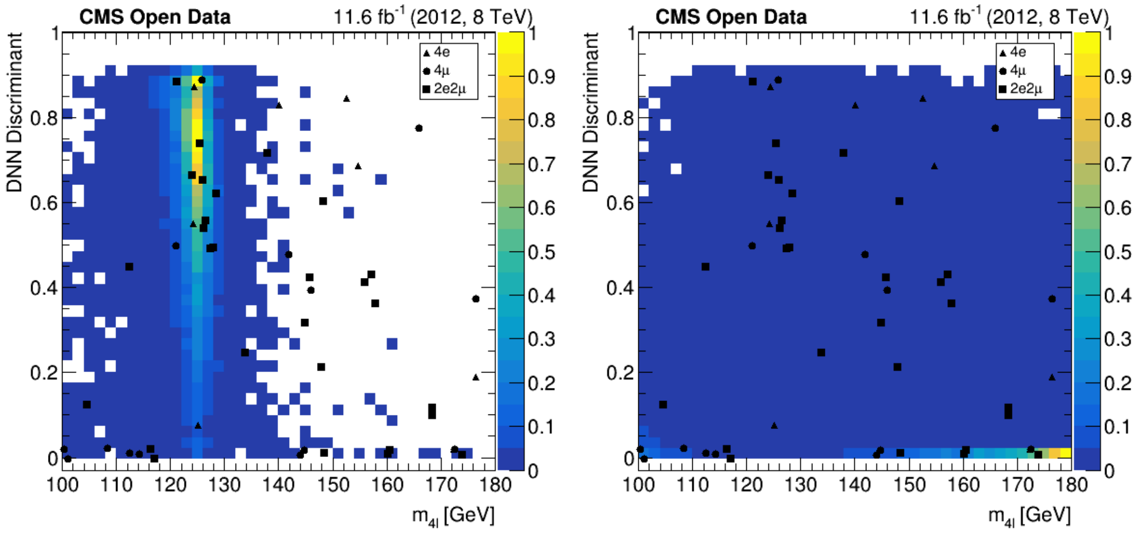
(b) Φ_1 for normalized signal and background

(c) $\cos\theta^*$ for normalized signal and background

(d) $\cos\theta_1$ for normalized signal and background

(e) $\cos\theta_2$ for normalized signal and background

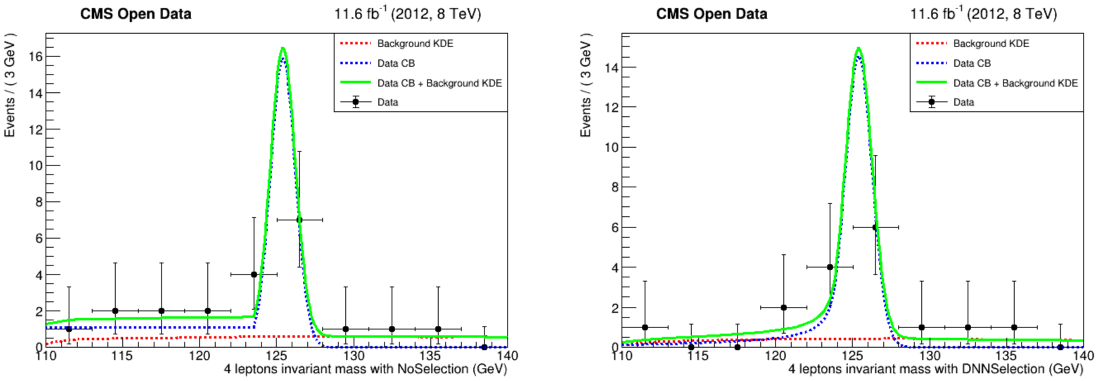
Figure 11: Distributions of the decay angles without the DNN selection.



(a) Distribution for the signal

(b) Distribution for the background

Figure 12: Distributions of m_{4l} VS DNN Discriminant for signal and background with the data points overlaid.



(a) Higgs mass fit without DNN selection

(b) Higgs mass fit with DNN selection

Figure 13: Higgs mass fit.