Chapter 3

Event reconstruction

3.1 Introduction

This chapter deals with the way in which the simulation and the data collected at the CMS experiment are processed for use in the $\mathrm{H} \to \gamma\gamma$ analysis. This procedure, known as *reconstruction*, is performed on an event-by-event level using custom-built software. Parts of the processing common to most CMS analyses are performed using the `CMSSW_X_X_X` package; further reconstruction specific to the $\mathrm{H} \to \gamma\gamma$ analysis is performed using the `FLASHgg_Y_Y_Y` software. A short discussion on some of the common methods used to perform and validate the reconstruction of events can be found in Section 3.2. The same reconstruction algorithms are applied both to the data collected at the CMS experiment and to simulated samples. The way in which both types of sample is obtained is discussed in Section 3.3.

It has already been noted in Section 1.2.3 that $\mathrm{H} \to \gamma\gamma$ is one of the most sensitive decays with which to study the Higgs boson in the LHC environment. This is despite it having a small branching fraction ($\sim 0.2\%$) and an irreducible background of SM processes which have two photons in the final state. The channel has the benefit of having a fully reconstructible final state of two high-energy photons in the ECAL. The experimental method is therefore to look for a resonant Higgs boson peak on top of a continuous diphoton invariant mass spectrum. It can be shown that the invariant mass of the diphoton system ($m_{\gamma\gamma}$), i.e. the invariant mass of the Higgs boson, is given by:

$$m_{\gamma\gamma} = \sqrt{2E_{\gamma}^{1}E_{\gamma}^{2}(1-\cos\alpha)}, \tag{3.1}$$

where E_γ^1, E_γ^2 represent the energies of the two photons and α represents the opening angle between them. The calculation of the opening angle relies on the spatial location of the Higgs boson decay. Therefore, Equation 3.1 indicates that to study $\mathrm{H} \rightarrow \gamma\gamma$, the most important steps are reconstructing photons and locating the PV. These two steps are discussed in Section 3.4 and Section 3.5 respectively.

Higgs bosons are produced at the LHC chiefly by the mechanisms, or *modes*, described in Section 1.2.2. At leading order, in the dominant production mode ggH, the final state consists only of the two Higgs decay photons. However, for other production modes, the Higgs boson can be produced in association with other particles. These additional objects in the detector can be reconstructed and provide information on the mode in which the Higgs boson was likely to have been produced. The methods used to reconstruct such additional objects are described in Section 3.6.

3.2 Common reconstruction methods

3.2.1 Particle-flow

The Particle flow (PF) event reconstruction algorithm [44, 45] combines information from all CMS sub-detectors to reconstruct and identify individual particles. The inputs to this algorithm are the tracks reconstructed in the tracker and muon system, and the clusters of energy reconstructed in the ECAL and HCAL. The outputs of the algorithm are objects corresponding to stable particles (photons, electrons, muons, charged hadrons or neutral hadrons). These so-called PF *candidates* can then be further clustered to obtain jets and identify missing energy from unreconstructed particles. In this scheme, ECAL SCs which are not on the extrapolated trajectory of tracks from either the muon system or the tracker are identified as photon candidates. If a track in the tracker is associated to one or more ECAL SCs, then it can be identified as an electron candidate. If a track in the tracker is consistent with a track or multiple hits in the muon system, then it is identified as a muon candidate. Tracks in the tracker which are not associated with any track in the muon system or any deposit in the ECAL are interpreted as charged hadron candidates. Finally, deposits in the HCAL which are not associated with any tracks can be identified as neutral hadron candidates.

3.2.2 Boosted decision trees

Analyses in high energy physics (HEP) often use multi-variate analysis (MVA) techniques to improve their overall sensitivity. An example of an MVA technique used repeatedly in this thesis is the decision tree (DT) method, where a technique known as *boosting* is applied to produce a BDT. Problems where a BDT is of use always involve a list of items with N_{inputs} *features* or *input variables*, labelled here $\vec{x} = (x_0, ..., x_{N_{\text{inputs}}})$, and a property y, the *target variable* to be determined. The objective of a BDT is to produce a function $F(\vec{x})$ which is an estimate of the true value of y for a given set of input variable values [46]. There are two common uses for BDTs: *classification* and *regression*. For example, for a classification BDT where the items are events, y could describe whether an event contained a Higgs boson decay (signal) or not (background), based on a set of input variables $\vec{x}$ derived from properties of the PF objects in the event. In this example, as with all classification BDTs, the target variable takes discrete values (background or signal). In the case of regression BDTs, the target variable is continuous rather than discrete. For example, the energy correction for SCs in the ECAL (F_{SC} in Equation 2.4) is obtained using a regression BDT, as is described in Section 3.4.3. Unless otherwise specified, the BDTs used in this thesis were produced using the `TMVA` framework [47] as part of the `ROOT` software package.

In general, a BDT is a linear combination of DTs. A DT is obtained using a *training dataset* consisting of a list of items $(\vec{x}_m, y_m, w_m)$ for $m = 0, ..., N_{\text{items}}$, where $\vec{x}_m$ is a set of input variables values, y_m is the true value of the target variable and items can be weighted with weight w_m. In the simplest case, the value of y_m is binary: signal or background. A numerical value, 1 and -1 say, can be assigned to these two options respectively. The following description uses this binary output example, but can be generalised for y_m to take any number of discrete values for classification DTs, or continuous values in the case of regression DTs. To construct a DT, the training dataset is first split into two sub-samples by applying a selection, which will be referred to as a *cut*, on one or more of the input variables. The *purity* $p(s)$ of a sub-sample s is the proportion of signal items, given by:

$$p(s) = \frac{\sum_{m=0}^{N^s_{\text{items}}} w_m \cdot \text{Bool}(y_m = 1)}{\sum_{m=0}^{N^s_{\text{items}}} w_m}, \tag{3.2}$$

where N^s_{items} is the number of items in the sub-sample and $\text{Bool}(X)$ is equal to 1 (0) if X it true (false). The cuts on the input variables are chosen to maximise the separation of signal and background in the resulting sub-samples. This is achieved by minimizing a separation criterion. A common separation criterion is the *Gini index* $2p(s)(1-p(s))$, which has a maximum at 0.5 for sub-samples with an equal amount of signal and background items and gives 0 in cases where all items are of the same type (signal or background). Many other separation criteria exist, for instance *cross-entropy, misclassification error, statistical significance* and *average squared error*. Each sub-sample can then be further split by a new set of cuts on the target variables. This procedure is repeated iteratively for each sub-sample until either the number of iterations reaches some predefined threshold known as the *tree depth*, or if the sub-sample satisfies some predetermined requirement on the value of the separation criterion. Each sub-sample obtained after the final set of cuts has been applied is known as a *leaf*. The output score of the items in a given leaf is then $1(-1)$ if $p(s) > 0.5$ ($p(s) \leq 0.5$).

The procedure known as boosting helps to improve the performance of a DT, for example by reducing the impact of statistical fluctuations in the training sample. Many boosting algorithms exist, but in all cases several individual DTs are produced, each trained on sub-sets or modified versions of the training dataset. The final BDT is a linear combination of the individual DTs. Supposing that there are N_{DT} individual DTs, labelled as $f_l(\vec{x}, \vec{\alpha}_l)$ where $l = 0, .., N_{\text{DT}}$ and $\vec{\alpha}_l$ is the set of cuts in the corresponding DT, the full BDT, F, is written as:

$$F(\vec{x}, \vec{\beta}, \vec{\alpha}) = \sum_{l=0}^{N_{\text{DT}}} \beta_l f_l(\vec{x}, \vec{\alpha}_l), \tag{3.3}$$

where $\vec{\beta} = (\beta_0, ..., \beta_{N_{\text{DT}}})$ is the set of coefficients applied to each DT in the BDT. The values of of $\vec{\beta}$ are determined by the boosting algorithm [47, 48]. A consequence of Equation 3.3 is that the output of the BDT is no longer a discrete value of ± 1, but instead a semi-continuous variable between -1 and 1.

Some of the most common boosting algorithms are *adaptive boosting* and *gradient boosting*. In adaptive boosting, the first DT is trained on the full training dataset as usual. The training dataset is then re-weighted so that items which were assigned an incorrect output score by the previous DT are given a larger weight and the items which were correctly classified are given a smaller weight. The next DT is then trained on the modified dataset, and the whole procedure is repeated over a large number of iterations. The final $\vec{\beta}$ for adaptive boosting algorithms is given by the logarithm of the weights applied to the dataset at each step. In gradient boosting the values of the individual components of $\vec{\alpha}$ and $\vec{\beta}$ are varied in such a way as to minimise the *loss function* $L(F, y)$, which is a measure of the deviation of the BDT output score from the true value of y across all items in the training dataset. Popular choices for the loss function include $L(F, y) = (y - F(\vec{x}, \vec{\beta}, \vec{\alpha}))^2$ and $L(F, y) = \ln(1 + e^{-2F(\vec{x}, \vec{\beta}, \vec{\alpha})y})$ [47, 48].

3.2.3 Tag-and-probe

The tag-and-probe technique [49] is a common way to determine the efficiency of a selection S in data. The technique typically exploits decays where a resonant particle decays to a pair of particles, for instance $\mathrm{Z} \to \mathrm{e}^+\mathrm{e}^-$ or $\mathrm{Z} \to \mu^-\mu^+$. The events used for the tag-and-probe method are selected such that the invariant mass of the decay products are near the mass peak of the resonant particle, thus ensuring high-purity sample. One can impose a strict identification requirement on one of the decay products, referred to as the *tag*, and then apply a very loose identification requirement for the other decay product, referred to as the *probe*. The requirements applied to the probe should be loose enough that they does not affect the efficiency of S. The efficiency of S is then the fraction of probes which satisfy S.

3.3 Samples

3.3.1 Simulation samples

Signal samples are sets of simulated events containing $\mathrm{H} \to \gamma\gamma$ decays, which are produced for each of the main Higgs boson production modes for a range of values of m_{H}. The cross-section and branching fractions used for the simulations under each m_{H} scenario are those recommended by the LHC Higgs Cross Section Working Group (LHCHXSWG) [50]. Signal samples are used to train BDTs, validate reconstruction algorithms and produce signal models, as described in Section 5.1. Signal simulations are produced at parton-level using the generator MADGRAPH5_aMC@NLO [51], which makes use of perturbative QCD at next-to-leading order (NLO). The parton-level samples are then interfaced with PYTHIA 8 [52], using the tune CUETP8M1 [53], which models the subsequent showering and hadronisation of partons. *Background samples* are sets of events which represent the reducible and irreducible backgrounds for the $H \to \gamma\gamma$ decay. Background samples are used to train BDTs, to validate reconstruction algorithms and to optimise the categorisation scheme. The *irreducible background* samples are composed of SM processes which yield two genuine photons from a p-p interaction in the final state, and are modelled using the `Sherpa` [54] generator. The *reducible background* samples represent events where some jets are incorrectly identified as isolated photons. The largest contributors to the reducible background are $\gamma + \text{jet}$ and QCD multijet events, which are modelled using the `PYTHIA 8` generator, where a filter designed to enhance the fraction of events with a large component of electromagnetic energy is applied. Drell-Yann (DY), $W\gamma$ and $Z\gamma$ samples are also used for validation purposes, and these are simulated using `MADGRAPH5_aMC@NLO`.

For all simulated samples, the detailed response of the CMS detector is modelled using `GEANT 4` [55], This modelling takes into account the effect of additional interactions in an event other than the one occurring at the PV, which are collectively referred to as pile-up (PU). The PU is modelled in the nominal bunch crossing as well as in previous and subsequent bunch crossings. The samples are re-weighted such that their PU distributions match the data before they are used in the analysis.

3.3.2 Data samples and trigger

The data sample analysed in this thesis was recorded using the CMS detector in between March and October 2016, during p-p LHC collisions at $\sqrt{s} = 13\,\text{TeV}$. It corresponds to an integrated luminosity of $32.4\,\text{fb}^{-1}$. As was described in Section 2.2.6, events recorded for analysis at CMS must pass the requirements of the triggering system.

The L1T requires either a deposit in the ECAL with $p_T > 25\,\text{GeV}$ or two deposits with $p_T > 15\,\text{GeV}$ and $p_T > 10\,\text{GeV}$ respectively. Events passing the L1T are processed at the HLT, where a basic clustering is applied to the candidate photon deposits. The requirements for an events to be saved to the double-photon sample by the HLT are as follows:

- the event contains two candidate photons, with $m_{\gamma\gamma}$ above 90 GeV,
- the candidate photon with most energy satisfies $E_T > 30\,\text{GeV}$,
- the candidate photon with second-most energy satisfies $E_T > 18\,\text{GeV}$,

- both candidate photons pass a loose calorimeter-based identification using shower shape and isolation requirements.

Events passing the L1T and HLT requirements detailed are saved for further processing. The efficiency of the trigger requirements is studied using a modified tag-and-probe method in data with $\mathrm{Z} \rightarrow \mathrm{e}^{+}\mathrm{e}^{-}$ events. The L1T efficiency is found to be above 97.5% in the EB and above 92% in the EE. The efficiency of the HLT is found to be above 97% in the EB and 96% in the EE.

3.4 Photon reconstruction

3.4.1 Clustering of ECAL deposits

Photons or electrons which impact the ECAL deposit most of their energy in a single crystal, which is referred to as the *seed* crystal. However, particles can have several deposits associated with them. For instance, the electromagnetic shower which develops in the seed crystal typically spreads out into neighbouring crystals. Photons travelling towards the ECAL can interact with the tracker material and undergo pair conversion ($\gamma \rightarrow \mathrm{e}^{+}\mathrm{e}^{-}$), resulting in two nearby or overlapping showers. Electrons or positrons travelling towards the ECAL are deflected by the magnetic field, and emit photons via bremsstrahlung: in this case the radiated photons also leave deposits in the ECAL which are offset from the seed in the ϕ-direction. All the deposits associated with any individual particle from the PV must be grouped into a SC to make an accurate measurement of the particle's energy. This is achieved through a process called *clustering*.

The clustering algorithm begins by identifying seed crystals as those with the largest amount of energy in the local area, above a minimum noise threshold. The threshold represents approximately two standard deviations of the electronic noise in the ECAL. Next, clusters are obtained by iteratively grouping crystals which have a common side with a crystal already in the cluster if their energy is above the threshold. The location of a cluster is defined as the energy-weighted average position of the individual crystals which compose it. During the iterative clustering process, if a crystal could belong to different clusters, its energy is shared between them according to the distance between the crystal and each cluster, assuming a Gaussian shower profile. Finally, the clusters are dynamically merged into SCs by gathering those which lie between two parabolas extending into the ϕ-direction as a function of η. This gives the SCs a mustache-like shape, designed to recover the energy of bremsstrahlung photons and to mitigate the effect of PU.

The SCs obtained from clustering the ECAL deposits are fed into the PF algorithm described in Section 3.2.1, and are interpreted as either electron or photon candidates.

3.4.2 Common variables used for photon and electron studies

The study of photons and electrons in the CMS ECAL relies on a number variables which are used to characterise the development of the electromagnetic shower within the $PbWO_4$ crystals and mitigate the effect of PU. These variables are used in various stages of the reconstruction, and for convenience some of the most important ones are defined here.

General variables:

- ρ, the median energy density in the event, per unit area;
- *Electron veto*, a variable which returns "true" if the PF algorithm determined that a SC corresponds to an electron.

Shower shape variables:

- R_9, the energy in the 3×3 array of crystals around a seed crystal divided by the energy of the SC. The R_9 variable gives information about whether a photon converted (high values of R_9) or not (lower values of R_9);
- S_4, the energy of the most energetic 2×2 array of crystals containing the seed crystal, divided by the energy of the SC;
- n_{clusters}, the number of clusters in a SC;
- $\sigma_{i\eta i\eta}$, the width of the shower in the η-direction, measured in units of crystals;
- σ_{η}, the standard deviation of the logarithmic energy-weighted η-position of individual crystals within the SC;
- $\sigma_{i\phi i\phi}$, the width of the shower in the ϕ-direction, measured in units of crystals;
- σ_{ϕ}, the standard deviation of the logarithmic energy-weighted ϕ-position of individual crystals within the SC;
- σ_{RR}, the width of the shower in the radial direction (defined in endcaps only);

- $cov_{i\eta i\phi}$, the covariance of the single-crystal values of η and ϕ in units of crystals for the 5×5 array centered around the crystal with the most energy;
- E_{seed}/E_{SC}, $E_{\mathrm{seed}}/E_{3\times3}$, $E_{\mathrm{seed}}/E_{5\times5}$, ratios of the energy of the seed crystal and energy of the SC, the 3×3 array centered around the seed and the 5×5 array centered around the seed respectively;
- H/E, the ratio of the energy deposited in the HCAL in a cone of radius $R = 0.15$ directly behind a SC and the energy of that SC as recorded in the ECAL.

Isolation variables:

- $Iso^{\mathrm{PF}\gamma}_{R=0.3}$, PF photon isolation i.e. the sum of the transverse energy of all PF photons which lie inside a cone of radius $R = 0.3$ around the candidate photon, where the energies have been corrected according to ρ;
- $Iso^{\mathrm{PF\ ch.\ had.}}_{R=0.3}(V)$, PF charged hadron isolation i.e. the sum of the transverse energy of all PF charged hadrons associated with a particular vertex V which lie inside a cone of radius $R = 0.3$ around the candidate photon, where the energies have been corrected according to ρ. This quantity is typically calculated with respect to the selected primary vertex, and also for the vertex which has the largest isolation sum (referred to as the *worst vertex*). The latter helps to reject photons candidates which are mis-identified jets originating at a vertex other than the chosen one;
- $Iso^{\mathrm{tracker}}_{0.04<R<0.3}$ tracker hollow cone isolation: the sum of the p_{T} of all tracks falling inside a hollow cone of radius $0.04 < R < 0.3$ around the photon candidate.

3.4.3 Photon energy reconstruction BDT

The energy of a SC is given by Equation 2.4, where F_{SC} is a correction to the SC energy which takes into account second order effects such as how well a SC is contained within a crystal. This correction is obtained using a per-SC regression BDT, referred to here as $BDT_{\gamma\ \mathrm{E}}$, assuming that the SC corresponded to a photon deposit. A similar but separate BDT is used to correct the energy of the SC under the assumption that the energy was deposited by an electron.

The target variable for the $BDT_{\gamma\ \mathrm{E}}$ is the ratio of the true photon energy E_{true} and the raw energy of the SC E_{SC}. The $BDT_{\gamma\ \mathrm{E}}$ is trained separately for SCs in the barrel and endcaps. The set of input variables is listed below:

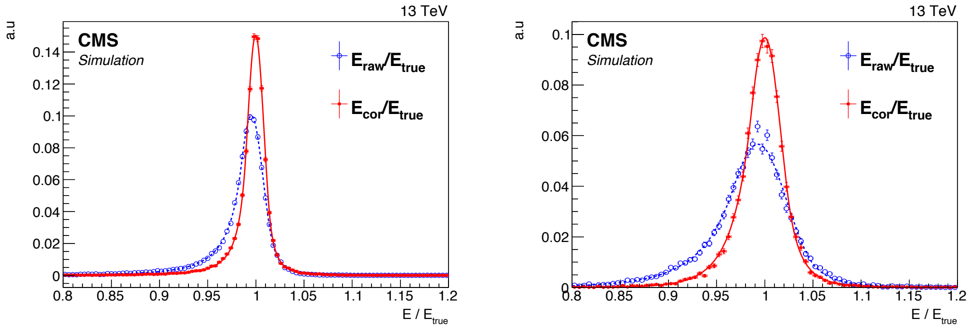
- Shower shape variables such as those defined in Section 3.4.2, which provide information about whether a photon converted, the extent to which it began showering before hitting the ECAL, and the extend to which the shower was contained within the ECAL;
- Position variables i.e. the position of the seed crystal in terms of crystal indices $i\eta$ and $i\phi$, the distance between the positions of SC and the seed crystal and the distance between the seed crystal and the boundaries between ECAL modules; These variables provide information about energy lost through gaps in between the detector modules and between crystals;
- Noise variables i.e. variables related to PU such as the number of reconstructed vertices in the event and the median energy density ρ in the ECAL as a function of position.

The $BDT_{\gamma\,\mathrm{E}}$ is trained on a simulated sample of double-photon events re-weighted to flatten the p_{T} and η distributions of the individual photons. The training is done using a *semi-parametric likelihood* technique. In this technique, the distribution of the target variable from the training sample is fitted to a functional form, in this case a double Crystal Ball (DCB) function [56]. The DCB was chosen as it is typically a good fit for processes modelling detector resolutions. It consists of a Gaussian core and independent power-law tails of each side, with 6 independent parameters: the width σ_{DCB} and mean μ_{DCB} of the core, the left tail cutoff and power parameters $\alpha^L_{DCB}, n^L_{DCB}$ and the right tail cutoff and power parameters $\alpha^R_{DCB}, n^R_{DCB}$. The fitting is performed by minimizing the negative log-likelihood (NLL):

$$-\ln\mathcal{L} = -\sum_{\text{photons}} \ln P_{DCB}(\frac{E_{true}}{E_{SC}}|\mu_{DCB}, \sigma_{DCB}, \alpha^L_{DCB}, n^L_{DCB}, \alpha^R_{DCB}, n^R_{DCB}), \tag{3.4}$$

where P_{DCB} is the DCB probability distribution and all the DCB parameters are functions of $\vec{x}$, the set of input variables for $BDT_{\gamma\,\mathrm{E}}$. The non-parametric dependence of the DCB parameters on $\vec{x}$ is obtained using separate BDTs, where a new tree is produced for each iteration of the NLL fit. The most probable value of E_{true} is then given by $\mu_{DCB}(\vec{x}) \times E_{SC}$ and the relative energy resolution for each photon is approximated by $\sigma_{DCB}(\vec{x})/\mu_{DCB}(\vec{x})$.

The performance of the $BDT_{\gamma\,\mathrm{E}}$ is tested on a simulated sample of $\mathrm{H} \to \gamma\gamma$ photons where $m_{\mathrm{H}} = 125\,\mathrm{GeV}$, as is show on Figure 3.1.



(a) EB

(b) EE

Figure 3.1: The ratio of the SC energy, shown before the $BDT_{\gamma\ \mathrm{E}}$ correction is applied (E_{raw}) and after the correction is applied (E_{cor}), and true energy (E_{true}) of simulated photons in a sample of $\mathrm{H} \to \gamma\gamma$ decays where $m_{\mathrm{H}} = 125\,\mathrm{GeV}$, separately for the EB and the EE. All distributions have been fitted to DCB functions.

3.4.4 Photon pre-selection

The photon candidates considered in the $\mathrm{H} \to \gamma\gamma$ analysis are required to satisfy certain requirements on their kinematics, shower shapes and isolation. All photon candidates are first required to have an electron veto value of “false”, and are the grouped into *diphotons* by determining all possible pairs of photons in the event. For each diphoton, the photon with the largest p_{T} (the *leading* photon) must satisfy $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ while the photon with the second-largest p_{T} (the *sub-leading* photon) must satisfy $p_{\mathrm{T}} > 20\,\mathrm{GeV}$. Additional requirements are made on a per-photon basis on the following variables: $\sigma_{i\eta i\eta}$, H/E, $Iso^{\mathrm{PF}\gamma}_{R=0.3}$, $Iso^{\mathrm{PF\ ch.\ had.}}_{R=0.3}$, $Iso^{\mathrm{tracker}}_{0.04<R<0.3}$ and R_9.

The pre-selection is designed to be more stringent than the triggering requirement described in Section 3.3.2, such that the data (which must pass the trigger) and the simulation (for which no triggers are defined) samples inhabit a common phase space.

The pre-selection efficiency, for all requirements aside from the electron veto, is measured in data and simulation using $\mathrm{Z} \to \mathrm{e}^{+}\mathrm{e}^{-}$ events, using a tag-and-probe technique for photons in different regions of η and R_9. The results can be seen in Table 3.1. The efficiency of the electron veto was measured separately using a sample of $\mathrm{Z} \to \mu\mu\gamma$ events, and found to be between 96% and 100%.

	DATA			Simulation		Ratio	
	Eff.	Stat. Unc.	Syst. Unc.	Eff.	Stat. Unc.	Eff.	Unc.
ECAL Barrel; R_9 >0.85	0.9451	0.0006	0.0192	0.9374	0.0007	1.0080	0.0192
ECAL Barrel; R_9 <0.85	0.8255	0.0012	0.0119	0.8258	0.0009	0.9960	0.0120
ECAL Endcap; R_9 >0.90	0.9099	0.0008	0.0212	0.9127	0.0010	0.9969	0.0212
ECAL Endcap; R_9 <0.90	0.4993	0.0018	0.0249	0.5024	0.0016	0.9938	0.0250

Table 3.1: Photon pre-selection efficiency (using all requirements aside from the electron veto) measured using $\mathrm{Z} \rightarrow \mathrm{e}^{+}\mathrm{e}^{-}$ events in data and simulation with a tag-and-probe technique.

3.4.5 Photon identification

In order to separate *prompt* photons (which were produced at the PV) from *fakes* such as misidentified jets, a per-photon BDT referred to as $BDT_{\gamma\ \mathrm{ID}}$ is applied to photon candidates which pass the pre-selection described in Section 3.4.4. The $BDT_{\gamma\ \mathrm{ID}}$ is trained on a $\gamma+$jet sample where the signal items are photon candidates which are geometrically matched to a generator-level photon from a p-p interaction. The background items are composed of th photon candidates which have no generator-level photon match, and are therefore likely to have resulted from a misidentified neutral hadron or jet. In both cases, photon candidates are required to pass the event pre-selection. To reduce the dependence of the $BDT_{\gamma\ \mathrm{ID}}$ on the kinematics of the photon, the signal items are re-weighted such that their p_{T} and η distributions match those of the background items. The input variables for the $BDT_{\gamma\ \mathrm{ID}}$ are $\sigma_{i\eta i\eta}$, $cov_{i\eta i\phi}$, S_4, R_9, σ_η, σ_ϕ, σ_{RR}, $Iso_{R=0.3}^{\mathrm{PF}\gamma}$, $Iso_{R=0.3}^{\mathrm{PF\ ch.\ had.}}$(selected vertex), $Iso_{R=0.3}^{\mathrm{PF\ ch.\ had.}}$(wrong vertex), ρ, η_{SC} and E_{SC}.

A loose requirement on the output of the $BDT_{\gamma\ \mathrm{ID}}$ is applied to all photons considered in the analysis, such that 99% of the signal photon candidates are kept while a large proportion of background photon candidates are removed. The $BDT_{\gamma\ \mathrm{ID}}$ output score of each photon is then used as a measure of the "quality" of each photon, and used as an input for the classification BDT described in Section 4.1. The $BDT_{\gamma\ \mathrm{ID}}$ is validated by comparing the output score for data and simulation for diphoton events (Figure 3.2) and for $\mathrm{Z} \rightarrow \mathrm{e}^{+}\mathrm{e}^{-}$ events (Figure 3.3) where the electron veto requirement is inverted.

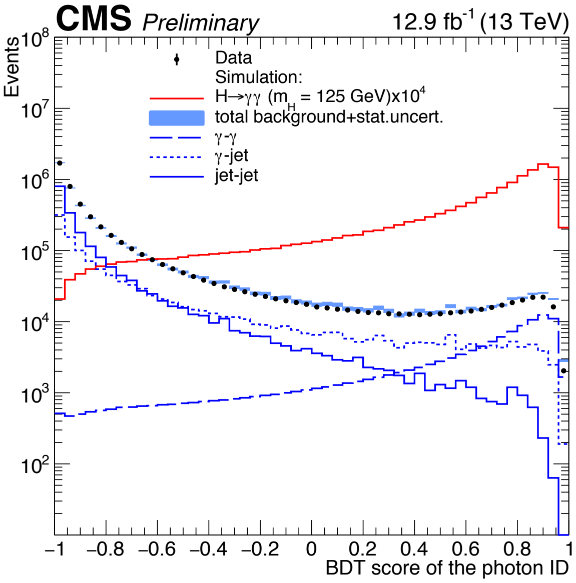


Figure 3.2: The distribution of the $BDT_{\gamma\ \mathrm{ID}}$ output score for the lower-scoring photon in each diphoton pair in the range $100 < m_{\gamma\gamma} < 180\,\mathrm{GeV}$ for data and simulation. The simulation is composed of signal ($\mathrm{H} \to \gamma\gamma$ photons with $m_{\mathrm{H}} = 125\,\mathrm{GeV}$) and background, which has been split into prompt-prompt (γ-γ), prompt-fake (γ-jet) and fake-fake (jet-jet) components. The sum of the background components has been scaled to match the number of events in data.

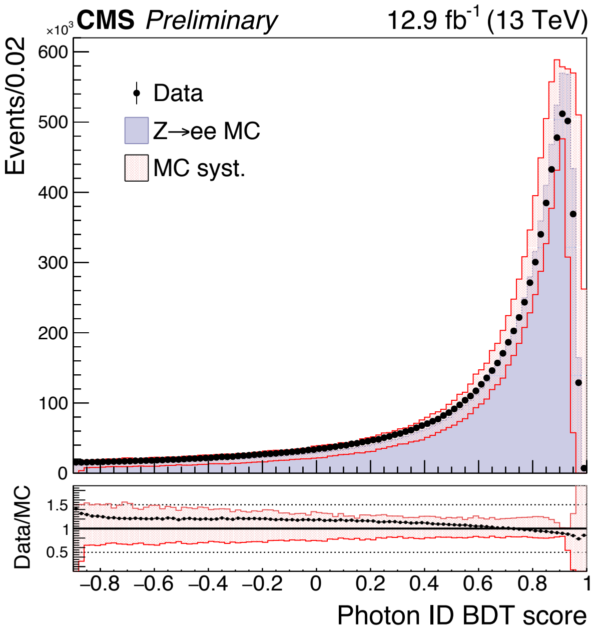


Figure 3.3: The BDT$_{\gamma\ \mathrm{ID}}$ output score for $\mathrm{Z} \to \mathrm{e}^+\mathrm{e}^-$ events in data and simulation (labelled MC), where the electrons are reconstructed as photons with the electron veto requirement inverted.

3.5 Vertex reconstruction

3.5.1 Vertex identification

As was discussed in Section 3.1, the determination of the location of Higgs decay is an important step in the reconstruction of $\mathrm{H} \to \gamma\gamma$ events, as it impacts the calculation of the invariant mass of diphoton system. If the selected vertex is within 1 cm of the true vertex location in the z-direction, then the impact of the opening angle on the mass resolution is negligible. Conversely, failing to identify the vertex within 1 cm leads to a degradation of the mass resolution.

Since the CMS ECAL is composed of a single layer of crystals, it cannot be used to point towards the vertex location. None the less, it is possible to exploit particle candidates recoiling from the diphoton system and the tracks of any electrons resulting from pair conversion to help determine the location of the vertex.

The first step is to produce a list of candidate vertex locations by considering all the tracks recorded in the tracker and grouping them into common points of origin by determining their closest point of approach to the beam-line. Next, a per-vertex BDT is used to determine which of the candidate vertices is most likely to be the point of origin of the Higgs boson decay. This BDT is refereed to as $BDT_{\mathrm{VTX\ ID}}$. The set of input variables is listed below, where N^{vtx}_{tracks} is the number of charged PF candidates associated with a given vertex, $\vec{p_{\mathrm{T}}}^{i}$ is the transverse momentum of the i^{th} candidate and $\vec{p_{\mathrm{T}}}^{\gamma\gamma}$ is the transverse momentum of the diphoton system :

- the sum of squared transverse momenta of all tracks, $\sum_{i=0}^{N^{vtx}_{tracks}} |\vec{p_{\mathrm{T}}}^{i}|^2$;
- the transverse momentum balance, $\sum_{i=0}^{N^{vtx}_{tracks}}(-\vec{p_{\mathrm{T}}}^{i} \cdot \frac{\vec{p_{\mathrm{T}}}^{\gamma\gamma}}{|\vec{p_{\mathrm{T}}}^{\gamma\gamma}|})$;
- the transverse momentum asymmetry, $\frac{(|\sum_{i=0}^{N^{vtx}_{tracks}} \vec{p_{\mathrm{T}}}^{i}|-|\vec{p_{\mathrm{T}}}^{\gamma\gamma}|)}{(|\sum_{i=0}^{N^{vtx}_{tracks}} \vec{p_{\mathrm{T}}}^{i}|+|\vec{p_{\mathrm{T}}}^{\gamma\gamma}|)}$.

Two additional variables are also considered if one of the two photons has converted into an $\mathrm{e}^{+}\mathrm{e}^{-}$ pair, where additional information is available to help identify the vertex:

- the number of converted photon candidates in the event;
- the pull $|z_{vertex} - z_{conv}|/\sigma_{z_{conv}}$, where z_{vertex} and z_{conv} are the z-components of the positions of the reconstructed vertex under consideration and the position of the vertex extrapolated from the conversion tracks respectively, and $\sigma_{z_{conv}}$ is the uncertainty on the extrapolated vertex position.

The $BDT_{\mathrm{VTX\ ID}}$ was trained using simulated Higgs boson events with $m_{\mathrm{H}} = 126\,\mathrm{GeV}$, where events from each production mode were weighted by their respective cross-section. The signal items are the vertices where a Higgs boson decay occurs at generator level. Any vertex in the sample not associated with a Higgs boson decay is treated as a background item. The training samples are re-weighted to account for the fact that the width of the *beamspot* (the distribution of the number of reconstructed vertices as a function of longitudinal position) in data and simulation is not the same: this width is modelled as 5.1 cm in simulation but is measured to be 3.6 cm in the data samples considered in this analysis. The re-weighting is performed as a function of the distance Δz between the true vertex and the selected vertex, such that the width of the distribution of Δz in simulation after the re-weighting matches the width of the Δz distribution in data, which is the beamspot width multiplied by a factor of $\sqrt{2}$.

The $BDT_{\mathrm{VTX\ ID}}$ was validated for unconverted photons using $\mathrm{Z} \to \mu^{-}\mu^{+}$ events in data and simulation. After determining the true vertex of the decay from the muon tracks, the events were re-reconstructed removing the muon tracks to mimic $\mathrm{H} \to \gamma\gamma$ system. For converted photons, the $BDT_{\mathrm{VTX\ ID}}$ was validated using a similar technique with $\gamma + \mathrm{jet}$ samples, where the true vertex was obtained from the tracks associated with the jet, and

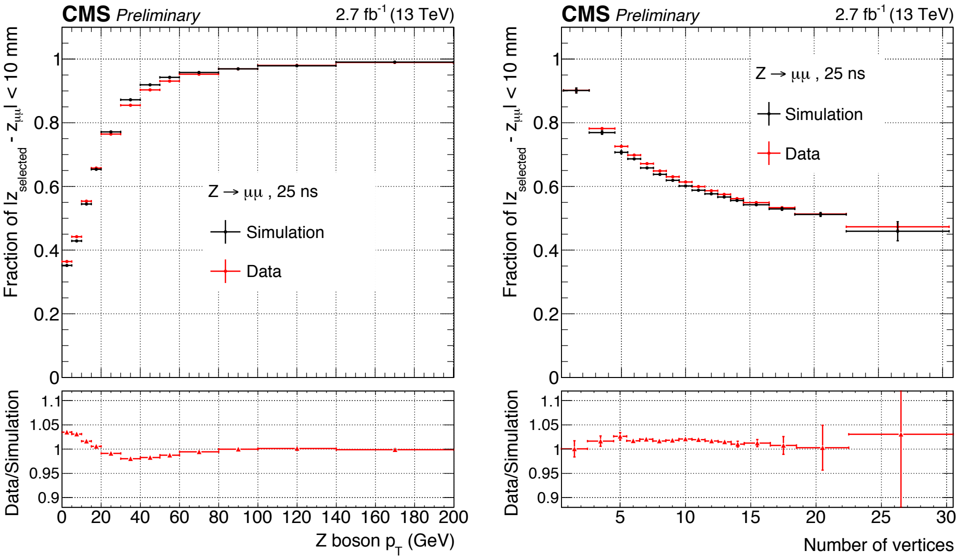


Figure 3.4: The efficiency of selecting a vertex within 1 cm of the true vertex in $Z \to \mu^- \mu^+$ events, as a function of the p_T of the Z-boson(left) and as a function of the number of vertices (right) in the event.

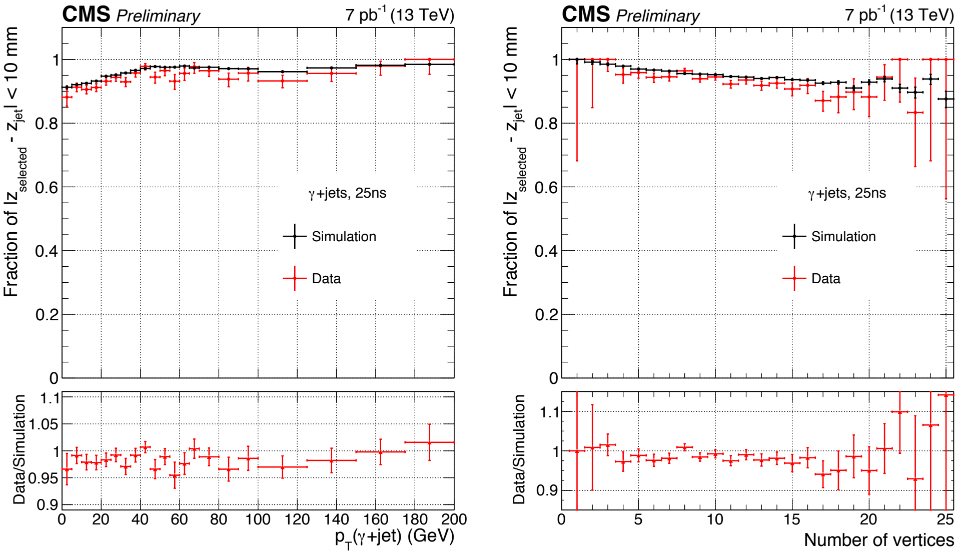


Figure 3.5: The efficiency of selecting a vertex within 1 cm of the true vertex in $\gamma +$jet events, as a function of the p_T of the $\gamma +$jet system (left) and as a function of the number of vertices (right) in the event.

the events were re-reconstructed removing the tracks associated with the jet to mimic a diphoton system. The vertex-finding efficiencies as a function of p_T and as a function of the number of vertices in the event can be seen in Figure 3.6 and Figure 3.4.

The efficiency of the $BDT_{\mathrm{VTX\ ID}}$ to select the right vertex within 1 cm of the true one was estimated with simulated signal events where $m_{\mathrm{H}} = 125\,\mathrm{GeV}$. The results are shown as a function of the number of vertices in the event and as a function of p_{T}. These can be seen on Figure 3.6. The average efficiency is of the order of 80%.

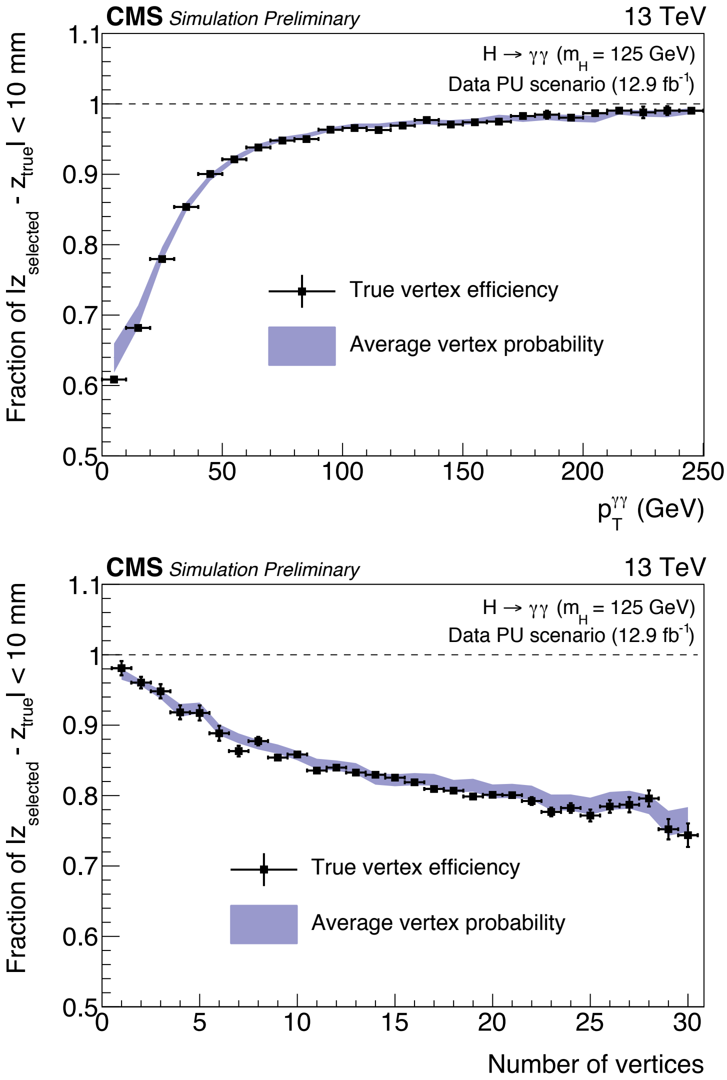


Figure 3.6: The efficiency to select a vertex within 1 cm of the true vertex in simulated $\mathrm{H} \to \gamma\gamma$ events as a function p_{T} and the number of vertices in the event. The estimated probability that the vertex was chosen within 1 cm is super-imposed. The uncertainty on the vertex-finding probability was determined using $\mathrm{Z} \to \mu^{-}\mu^{+}$ events. The simulation was re-weighted such that the distribution if the number of vertices and the width of the interaction region matched in data and simulation.

3.5.2 Correct vertex probability

If the chosen vertex is over 1 cm away from the true one, the invariant mass resolution is dominated by the uncertainty on the vertex position. It is therefore desirable to have a per-event estimate of how likely it is that the vertex was chosen within 1 cm of the true one. This is referred to as the *correct vertex probability.* This information is used to categorise events by sensitivity, as described in Section 4.1.

The estimate of the per-event vertex probability is obtained using a BDT, labelled $BDT_{\text{VTX PROB}}$, trained on simulated $\mathrm{H} \to \gamma\gamma$ events. The input variables for this BDT are:

- the number of reconstructed vertices in the event;
- the p_{T} of the diphoton system;
- the output scores of the three vertices ranked highest by the $BDT_{\text{VTX ID}}$;
- the distance in the z-direction between the first- and second-highest ranked vertices;
- the distance in the z-direction between the first- and third-highest ranked vertices;
- the number converted photons in the diphoton.

The correct vertex probability is parametrized by a 4^{th}-order polynomial as a function of $BDT_{\text{VTX ID}}$ output score. This is done separately for converted and unconverted photons. The estimated correct vertex identification probability is shown alongside the vertex efficiency measured in simulation of Figure 3.6. The $BDT_{\text{VTX PROB}}$ was validated using $\mathrm{Z} \to \mu^{-}\mu^{+}$ and $\gamma + \text{jet}$ events analogously to the $BDT_{\text{VTX ID}}$.

3.6 Other objects

3.6.1 Electrons

In the study of the $\mathrm{H} \to \gamma\gamma$ decay, electrons are used in two ways. Firstly, they are used to validate reconstruction algorithms using $\mathrm{Z} \to \mathrm{e}^{+}\mathrm{e}^{-}$ events. Secondly, they are used for the categorisation of $\mathrm{H} \to \gamma\gamma$ events where the Higgs boson events was produced by the ZH or WH mechanism. Candidate PF electrons are reconstructed either starting from ECAL deposits which are matched to tracks (in which case they are called *ECAL-driven* electrons), or starting from tracks which are matched to ECAL deposit (these are *tracker-driven* electrons). Typically, energetic and isolated electron candidates will be reconstructed as ECAL-driven, while low-energy ($p_{\mathrm{T}} < 10\,\mathrm{GeV}$) electrons will be reconstructed as tracker-driven. Electrons from both seeding algorithms are eventually grouped together to form the set of PF electron candidates. The electrons used in this thesis originate from $\mathrm{W}^{\pm}$ or Z decays, and hence are mostly ECAL-driven.

The ECAL-seeded electrons are obtained via a procedure analogous to that described for photons in Section 3.4, but with the additional step of associating a track based on geometrical requirements. Candidate tracks are obtained by from tracker hits within some window in z and ϕ around the SC position. The tracks are fitted with a special algorithm which accounts for changes in direction caused by the emission of bremsstrahlung. The SC is associated to the track whose extrapolated position in the ECAL is nearest to the energy-weighted position of the SC, but requiring that the distance in the η-direction (ϕ-direction) be no more than 0.02 (0.15). The energy of electrons is obtained from the SC energy, where the final energy correction F_{SC} is obtained using a BDT method analogous to that described in Section 3.4.3, but specially trained for electron candidates.

3.6.2 Muons

In this analysis, muons are used for the validation of reconstruction algorithms using $\mathrm{Z} \to \mu\mu\gamma$ events, and also for the selection of Higgs boson which were produced by the ZH or WH mechanism. Muons are constructed by geometrically matching tracks reconstructed independently in the tracking system and in the muon chambers. Muon candidates must have some hits in both sub-detectors to qualify as PF muons: this helps to avoids cases where cosmic rays or muons produced in jets are mis-reconstructed as prompt muons from the PV [42].

3.6.3 Jets

Jets are collections of collimated particles originating from the decay of quarks or gluons. In the $\mathrm{H} \to \gamma\gamma$ analysis, jets are used to identify events where the Higgs boson was produced by the VBF process. Jets are reconstructed using the anti-k_t clustering algorithm [57] from PF candidates, using a cone of radius $R = 0.4$. Jets originating from PU can sometimes overlap with jets which originate from particles produced at the PV. To mitigate this effect, PF charged hadron subtraction (PFCHS) is used. In this scheme, the PF charged hadron candidates associated to a vertex other than the vertex selected by the procedure described in Section 3.5 are ignored during the clustering. Since the tracker acceptance is $|\eta| < 2.5$, no PF charged hadron candidates are available outside this range, so PFCHS has no effect. For the jets reconstructed outside this region, but still in acceptance, a different PU mitigation technique is used using a selection on the width of the jet. The width of the jet is described by the variable $\sigma_{\mathrm{RMS}} = \sum_{\mathrm{PF\ candidates}} p_{\mathrm{T}}^{2} \Delta R^{2} / \sum_{\mathrm{PF\ candidates}} p_{\mathrm{T}}^{2}$, where ΔR is the distance between the PF candidate and the jet axis from the clustering cone. Jets must have $\sigma_{\mathrm{RMS}} < 0.03$ to pass the PU mitigation requirement. Finally, all jets are required to be within $|\eta| < 4.7$.

Parametric corrections to the energy of the jets are made to account for the following effects:

- the additional energy of PF neutral hadrons from PU which are clustered into jets;
- the non-uniformity of the detector response;
- the difference between data and simulation in $\gamma + \mathrm{jet}$ and $\mathrm{Z} + \mathrm{jet}$ events.

3.6.4 Missing energy

Particles which do not leave any deposits in the detector, such as neutrinos, still need to be reconstructed to identify decays from $\mathrm{W}^{\pm}$ bosons, for example when identifying Higgs boson decays originating from the WH production mode. Such particles are reconstructed by identifying the amount of energy taken away by these particles as they leave the detector, labelled as $\not{E}_{\mathrm{T}}$. The $\not{E}_{\mathrm{T}}$ is calculated by considering the magnitude and direction of p_{T} required to balance all the jets and PF objects in an event.