CHAPTER 5

OBJECT RECONSTRUCTION

This chapter describes the reconstruction and selection of the final state objects (leptons, jets, and b-tagged jets) used in this analysis. The reconstruction algorithm is explained in Section 5.1. The identification of jets and b-tagged jets is discussed in Section 5.2. The lepton selection is described in Section 5.3.

5.1 Particle flow reconstruction

In order to identify and reconstruct the particles produced in each collision, CMS uses a holistic reconstruction technique to correlate the elements from each subdetector (tracks and clusters) and construct a global picture of each event. This approach is referred to as particle flow (PF) reconstruction [6]. Since a particle generally interacts with multiple subdetectors, there are expected to be several PF elements associated with a given particle. The first step in the PF reconstruction is to link together the PF elements from the different subdetectors into sets of elements referred to as PF blocks.

In each block, the identification and reconstruction of the particles is performed in a specific order, and as each particle is identified, the associated elements are removed from the block. First, muons are identified based on tracks in the muon chambers, tracks in the inner tracker, and lack of energy deposits in the calorimeters. Next, electrons and isolated photons are reconstructed from energy clusters in the ECAL, lack of significant energy clusters in the HCAL, and tracks (or lack thereof) in the inner tracker. Finally, charged hadrons, neutral hadrons, and non-isolated photons are identified through several iterative steps using information from the inner tracker and both calorimeters. A schematic representation of PF reconstruction is shown in Figure 5.1, with example signatures displayed for each particle type (muon, electron, charged hadron, neutral hadron, and photon).

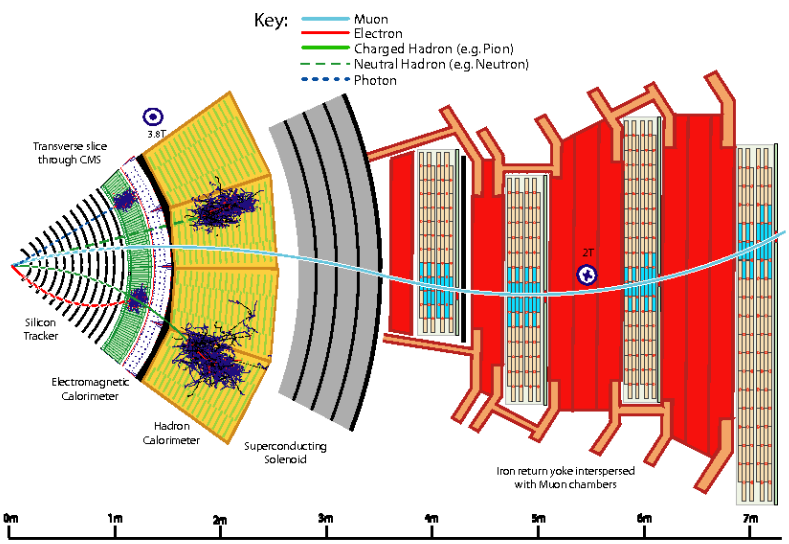

Figure 5.1. Schematic representation of the reconstruction of particles within CMS. An example signature for a muon, electron, charged hadron, neutral hadron, and a photon are shown. The PF algorithm is described in detail in Ref. [6] (from which this figure is taken) and summarized in Section 5.1.

Despite this thorough reconstruction algorithm, there may still be rare cases in which particles are misreconstructed, so a post-processing step is performed to mitigate these potential issues. The post-processing algorithm primarily aims to identify events with artificially high $p_{\mathrm{T}}^{\mathrm{miss}}$, usually due to the misidentification or misreconstruction of a high-p_T muon. The reconstruction of such muons is modified a posteriori and the $p_{\mathrm{T}}^{\mathrm{miss}}$ is recomputed; if the $p_{\mathrm{T}}^{\mathrm{miss}}$ is consequently reduced by at least a factor of two, the modified reconstruction is used.

5.2 Jets and b-tagging

After the PF reconstruction has identified all particles in the event, jets (collimated sprays of particles representing the experimental signature of quarks and gluons) are reconstructed using the anti-k_{T} jet clustering algorithm [32]. As explained in [32], the algorithm first defines a distance measure between the objects (d_{ij}), and between objects and the beam (d_{iB}); if the smallest distance measure is between two objects, they are combined into a single object, while if the smallest distance measure is between an object and the beam, the object is called a jet and removed from the list of objects. The distances are then recalculated and the process continues until all objects have been clustered into jets. For the anti-k_{T} algorithm, the distance measures are defined as follows:

$$d_{ij} = \min\left(\frac{1}{p_{Ti}^2}, \frac{1}{p_{Tj}^2}\right)\frac{\Delta R_{ij}^2}{R^2} \tag{5.1}$$

$$d_{iB} = \frac{1}{p_{Ti}^2} \tag{5.2}$$

where p_{Ti} is the p_T of particle i, p_{Tj} is the p_T of particle j, and R is a radius parameter; $\Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$, where y is the rapidity, and ϕ is the azimuthal angle. The the anti-k_{T} clustering algorithm is implemented with the FASTJET package [33]. The distance parameter R in the anti-k_{T} algorithm is set to 0.4.

Charged particles arising from pileup interactions are excluded from the jet clustering algorithm using the charged hadron subtraction (CHS) technique [34]. With this approach, charged hadrons with tracks associated with pileup vertices (i.e. vertices other than the primary vertex with the highest Σp_T^2) are removed from the list of particles to be used in the object reconstruction.

In this analysis, we require all jets to have $p_T > 30$GeV and $|\eta| < 2.4$. The jets are cleaned using the loose (or better) leptons with a ΔR requirement of greater than

0.4. Jets that overlap with objects in the fakeable lepton collection (electrons and muons) are also removed. The definitions of loose and fakeable leptons are provided in Section 5.3.

In order to identify jets originating from b quarks, the DeepJet b-tagging algorithm is used [35]. The relatively long lifetimes of b hadrons can lead to displaced secondary vertices and relatively large impact parameters with respect to the primary vertex (PV). As shown in Figure 5.2, the impact parameter is defined as the distance at the closest point of approach between the track and the PV. The relatively large mass of b hadrons can lead to decay particles with relatively large p_T with respect to the jet axis. These distinguishing characteristics are leveraged in the identification of b jets with the DeepJet algorithm.

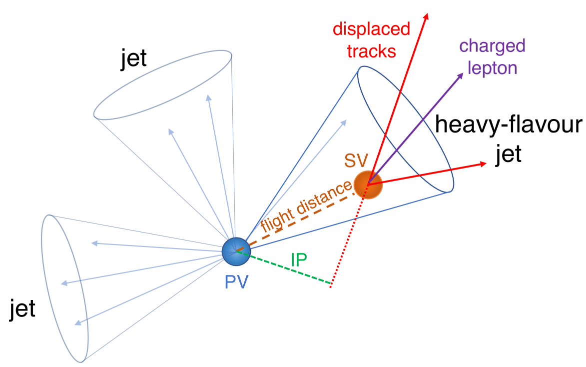

Figure 5.2. Schematic depiction of b jet, illustrating the impact parameter "IP" with respect to the PV. [35]

The b-tagging efficiency and mistag rates depend on the working point utilized (i.e. the cut on the discriminant output, which ranges from zero to one). For this analysis, we make use of the loose and medium working points of the DeepJet algorithm, as defined by the CMS b-tagging and vertexing (BTV) physics object group (POG) for the UL datasets [37–40].

5.3 Lepton object selection

This section will describe the lepton object selection used in this analysis, which is based on the selection used in the $t\bar{t}H$ multilepton analysis [41]. Section 5.3.1 provides an overview of the selection strategy and a conceptual introduction to the three stages of the lepton selection. A more detailed description of each selection stage is provided in Section 5.3.2, the requirements for which are summarized in Tables 5.1 and 5.2. Finally, the technical definitions of the observables that are used in the selection requirements are described in Section 5.3.3.

5.3.1 Conceptual overview of lepton object selection

The event selection for the signal regions of this analysis (Chapter 6) is defined primarily by the multiplicity of charged leptons (usually referred to as simply "leptons"). For this reason, it is important to ensure that the objects that are classified as leptons are likely to correspond to true leptons of interest. To this end, the lepton object selection aims to identify leptons that arise promptly from the PV.

To understand which leptons are to be classified as prompt, let us consider a $t\bar{t}W$ event. Each top quark will decay to a W boson and a b quark[1]. The event thus leads to three W bosons and two b quarks. The W bosons will decay either leptonically or hadronically; for the purposes of this discussion, let us consider leptonic decays, and let us assume the leptons are electrons or muons [2]. The bottom quarks will hadronize, producing a collimated sprays of hadrons (i.e. b jets). At this point, the collision has produced three charged leptons, three neutrinos, and two b jets. From an experimental perspective, all of these particles are produced essentially instantaneously, directly at the location of the PV.

[1]While it is possible for a top quark to decay to other flavors of quarks, the branching fraction for a top quark to decay to a W and a bottom quark is nearly 100% [42]

[2]Electrons and muons that arise from the decays of prompt taus are also considered to be prompt.

The charged leptons will travel outward from the PV, eventually encountering the CMS detector (which will measure their properties as described in Section 5.1); these leptons are considered to be prompt. The b hadrons in the b jets, however, may travel some observable distance before decaying; the decays may lead to leptons, but these leptons will not have originated directly from the PV, so these leptons are considered to be nonprompt. The goal of the lepton object selection is to distinguish prompt leptons from nonprompt leptons. The leptons that are identified as prompt are the leptons we use for the event selection.

Although a $t\bar{t}W$ event was used in this example, the same logic is applied to all processes. Any lepton that arises promptly from the PV is considered to be prompt. For example, a lepton that arises from the decay of a Z boson from the decay of a Higgs boson in a $t\bar{t}H$ event is also considered to be a prompt lepton. As one more example, a lepton that arises directly from an EFT vertex (e.g. the $c_{t\ell}^{(\ell)}$ vertex in Figure 2.1) would also be considered to be a prompt lepton.

The the identification of prompt leptons is performed in three stages, referred to as the loose, fakeable, and tight selections. As illustrated in Figure 5.3, the stages apply increasingly stringent requirements, so the collection identified in each stage is a subset of the previous collection.

The loose stage aims for high efficiency; this baseline set of objects is used to veto events with leptons from light resonances (i.e. from the decays of J/ψ and Υ particles, as explained in Section 6.2) and in the training of the MVA that is used in the tight selection. Building on the requirements of the loose selection, the fakeable selection is used to identify the set of leptons that are used in the estimation of the nonprompt background (discussed in Sectoin 8.1). Finally, the tight selection is applied in order to select the final set of prompt leptons that will be used in the event selection.

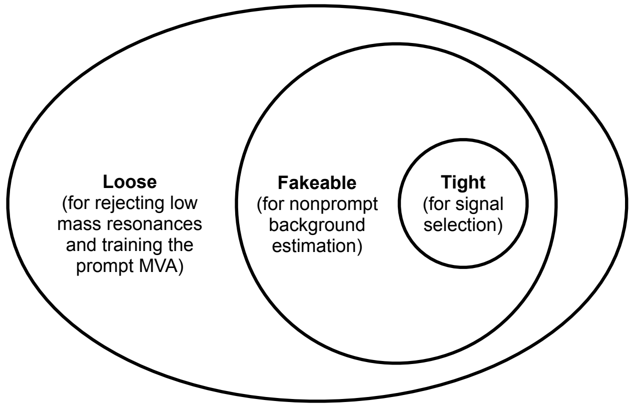

Figure 5.3. Venn diagram illustrating the three lepton object selection stages.

5.3.2 Lepton object selection stages

With the goal of separating electrons from jets, the loose stage of the electron object selection makes use of a boosted decision tree (BDT) multivariate algorithm, trained by the CMS EGamma POG [43–45]. For this loose selection, we use the loose working point (which has an effeciency of 98% [44]). Loose muons are required to pass the loose identification PF requirements defined by the CMS Muon POG [46]. The loose selection requirements for electrons and muons also include cuts on the isolation of the lepton and the impact parameter of the lepton's track with respect to the PV. Since signal leptons are expected to be relatively isolated from hadronic activity and to originate from the hard scatter process, these requirements can help to distinguish the leptons of interest from background objects. All of the requirements for the loose selection are are summarized in the "Loose" columns of Tables 5.1 and 5.2. The loose leptons are used to identify and veto events with a low-mass resonance (as described in Section 6.2), and are also used for training the prompt-e and prompt-μ MVAs described below.

The next stage in the object selection is referred to as the fakeable selection. Building on the loose collection, the objects that pass the fakeable requirements are a subset of the objects that pass the loose selection. Listed in the "Fakeable" columns of Tables 5.1 and 5.2, these selection requirements result in a collection of leptons that are very close to the final selection of signal leptons; this collection is used for the estimation of the nonprompt background contribution (as described in Section 8.1).

The final stage in the lepton object selection is the tight selection, which is used to identify the set of leptons that are considered to be signal leptons. Building on the fakeable and loose selections described above, the objects that pass the tight selection are a subset of the objects that pass the fakeable selection. The purpose of the tight selection is to identify leptons that arise promptly from the hard scatter, e.g. a lepton that is produced in the decay of a W boson in a $t\bar{t}W$ event. These leptons are considered to be the signal leptons for this analysis, and are referred to as "prompt" leptons. The tight selection criteria likewise aims to reject leptons that are produced in other ways (e.g. leptons arising from the decays of hadrons produced in the hadronization of a b jet), which are referred to as "nonprompt" leptons. This separation of prompt from nonprompt leptons is accomplished with a BDT that is trained by the $t\bar{t}H$ multilepton group. The BDT is described in detail in [41, 47, 48], a brief overview of which is provided below.

The training of the BDT is performed with simulated $t\bar{t}H$ and $t\bar{t}$ samples; the electrons and muons used in the training are required to pass the loose selection criteria defined in Tables 5.1 and 5.2. A separate BDT is trained for electrons and for muons. For electrons, MC samples with detector conditions corresponding to each each ultra-legacy (UL) period (UL16APV, UL16, UL17, and UL18) were used. For muons, the training from Ref. [48] (which was performed with end-of-year samples) was used, as there was no improvement in performance observed with the ultra-legacy datasets. The variables used in the training include the p_T of the lepton, the

η of the lepton, the impact parameters, the isolation of the lepton with respect to other charged particles, and the output of BDT trained by the CMS EGamma POG. The BDTs for electrons and for muons are referred to as the "prompt-e MVA" and "prompt-μ MVA", respectively.

TABLE 5.1

MUON OBJECT SELECTION REQUIREMENTS SUMMARY

Observable	Loose	Fakeable	Tight
p_T	> 5 GeV	> 10 GeV[4]	> 10 GeV
$\|\eta\|$	< 2.4	< 2.4	< 2.4
$\|d_{xy}\|$	< 0.05 cm	< 0.05 cm	< 0.05 cm
$\|d_z\|$	< 0.1 cm	< 0.1 cm	< 0.1 cm
d/σ_d	< 8	< 8	< 8
I_μ	$< 0.4 \times p_T$	$< 0.4 \times p_T$	$< 0.4 \times p_T$
PF muon	>WP-loose[1]	>WP-loose[1]	>WP-medium[1]
Deep Jet of nearby jet	–	<WP-interp. (<WP-medium)[2]	<WP-medium[2]
Jet relative isolation[3]	–	<0.5 (–) †	–
Prompt-μ MVA	–	< 0.85 (> 0.85)	> 0.85

[1] WPs as defined by Muon POG (see Section 5.3.2).
[2] Upper cut on the Deep Jet score defined with a linear interpolation from Deep Jet WP-medium at cone-p_T 20 GeV to Deep Jet WP-loose at cone-p_T 45 GeV, taking the Deep Jet WPs as defined by JetMET POG (see Section 5.2).
[3] Defined as 1/`jetPtRatio`-1 if the muon is matched to a jet within $\Delta R < 0.4$ or as the PF relative isolation with ΔR=0.4 otherwise.
[4] Here cone-p_T is used.
† Fails (passes) the requirement prompt-μ MVA > 0.85.

TABLE 5.2

ELECTRON OBJECT SELECTION REQUIREMENTS SUMMARY

Observable	Loose	Fakeable	Tight
p_T	> 7 GeV	> 10 GeV[5]	> 10 GeV
$\|\eta\|$	< 2.5	< 2.5	< 2.5
$\|d_{xy}\|$	< 0.05 cm	< 0.05 cm	< 0.05 cm
$\|d_z\|$	< 0.1 cm	< 0.1 cm	< 0.1 cm
d/σ_d	< 8	< 8	< 8
I_e	$< 0.4 \times p_T$	$< 0.4 \times p_T$	$< 0.4 \times p_T$
$\sigma_{i\eta i\eta}$	–	$<$ { 0.011 / 0.030 }[1]	$<$ { 0.011 / 0.030 }[1]
H/E	–	< 0.10	< 0.10
1/E - 1/p	–	> -0.04	> -0.04
Conversion rejection	–	✓	✓
Missing hits	≤ 1	$= 0$	$= 0$
EGamma POG MVA	>WP-loose[2]	>WP-90 (>WP-loose)[2]†	>WP-loose[2]
Deep Jet of nearby jet	–	<WP-interp. (<WP-medium)[2]	<WP-medium[2]
Jet relative isolation[4]	–	< 1.0 (–) †	–
Prompt-e MVA	–	< 0.90 (> 0.90)	> 0.90

[1] Barrel / endcaps.
[2] WPs as defined by EGamma POG (see Section 5.3.2).
[3] WPs as defined by BTV POG (see Section 5.2).
[4] Defined as $1/p_T^{ratio}$-1 if the electron is matched to a jet within $\Delta R < 0.4$ or as the PF relative isolation with ΔR=0.4 otherwise.
[5] Here cone-p_T is used.
† Fails (passes) the requirement prompt-e MVA > 0.80.

5.3.3 Definitions of variables used in lepton object selection

The loose, fakeable, and tight lepton selection is summarized in Tables 5.2 and 5.1. This section will define and describe the observables that are used in each stage of the selection. Where relevant, we specify parenthetically the name of the property (in the NanoAOD) corresponding to the given variable.

- p_T (`pt`): The transverse momentum of the lepton, as defined in 4.2.1. For the fakeable selection, the cone-p_T is used. The cone-p_T is designed to provide a characterization of the p_T of the parton that led to the nonprompt lepton. The lepton isolation and the p_T of the nearest jet are incorporated into the cone-p_T definition, which is provided in [47]. For fakeable leptons, the cone-p_T generally exceeds the reconstructed p_T of the lepton; for leptons that pass the tight selection criteria, the p_T and cone-p_T are equal.
- η (`eta`): The lepton's pseudorapidity, as defined in 4.2.1.
- $|d_{xy}|$ (`dxy`): This corresponds to the lepton track's transverse impact parameter with respect to the PV.
- $|d_z|$ (`dx`): This corresponds to the lepton track's longitudinal impact parameter with respect to the PV.
- d/σ_d (`sip3d`): This refers to the signed 3-dimensional impact parameter (with respect to the PV) divided by its uncertainty.
- I_e, I_μ (`miniPFRelIso_all`): This is a measure of the isolation of the lepton (corresponding to the sum of the p_T of the objects reconstructed within a cone centered on the lepton direction, where cone size is scaled inversely with the p_T, which also helps to mitigate the effects of PU).
- $\sigma_{i\eta i\eta}$ (`sieie`): The $\sigma_{i\eta i\eta}$ of the supercluster in the ECAL, a measure of the energy distribution within the crystal cluster.
- H/E (`hoe`): A measure of the energy deposited in the HCAL to the energy deposited in the ECAL.
- 1/E - 1/p (`eInvMinusPInv`): This corresponds to the difference between the reciprocal of the electron cluster energy and the reciprocal of its track momentum.
- PF muon (`looseId`): This requires the muon to pass the loose requirements specified by the Muon POG [46].
- Conversion rejection (`convVeto`): Requires the electron's `convVeto` property to be `True`.

- Missing hits (lostHits): The number of missing hits in the tracker.

- Jet relative isolation (jetRelIso): In the case where there is a matched jet, this corresponds to the relative isolation, defined as the difference between the matched jet p_T and the lepton p_T, with respect to the lepton-p_T. In the case where there is not a matched jet, the jetRelIso is equal to pfRelIso04_all.

- DeepJet of nearby jet: The output of the DeepJet discriminant (btagDeepFlavB) for the nearest jet (matched_jet).

- EGamma POG MVA: The output of the BDT trained by the EGamma POG. As described in Section 5.3.2, this helps to distinguish real electrons from jets.

- Prompt-e MVA: The output of the prompt lepton MVAs trained by the $t\bar{t}H$ multilepton team, as discussed in Section 5.3.2.