CHAPTER 8

BACKGROUNDS

This chapter describes the backgrounds for this analysis. Backgrounds are defined as contributions that populate the signal regions (SRs) described in Chapter 6 but are not significantly impacted by the WCs enumerated in Table 2.2. The backgrounds are categorized as either irreducible or reducible.

A background is categorized as irreducible if all of the final-state leptons are prompt. The dominant irreducible backgrounds are diboson processes, but smaller contributions also arise from triboson processes and the tWZ process. It is interesting to note that the tWZ process in principle should be affected by some of the WCs included in this analysis; preliminary studies indicated that the effect is not large, so this process is not included as a signal for this analysis. However, it may be interesting to revisit this process in the future and explore if it would be possible to improve the sensitivity to this process and other rare processes. The irreducible backgrounds are modeled using MC simulation. The datasets (produced centrally by CMS) are listed in Appendix A, in Tables A.11, A.12, A.13, and A.14.

Reducible backgrounds result from the misreconstruction or misidentification of objects. The primary source of reducible backgrounds arises when objects that are not genuine prompt leptons pass the tight selection criteria defined in Section 5.3. This contribution is referred to as the "nonprompt" background, and it is estimated with a data-driven technique, which will be described in Section 8.1. In the 2ℓss category, there is also a contribution from two lepton opposite sign (2ℓos) events where the charge of one of the leptons is mismeasured. This background is referred to as the "charge flip" background. The charge flip background is also estimated with a data-driven technique, which will be described in Section 8.2.

Though the nonprompt and charge flip backgrounds represent the largest contributions to the reducible backgrounds, we also account for an additional, smaller source of reducible background. Referred to as the "conversion" background, this backgrounds arises from $\gamma \rightarrow e^+e^-$ conversions where one of the leptons carries most of the energy of the photon (meaning the other may fail to be reconstructed). The contribution of this background is modeled with MC simulation, using the datasets listed in Appendix A. In the event selection, MC truth requirements are applied for this sample to ensure that at least one lepton is associated with a conversion (the `genPartFlav` property is required to be 22).

In order to validate the handling of the various sources of backgrounds, several control regions (CRs) are studied. To examine the nonprompt and flip backgrounds, we define a 2ℓss CR. This CR is similar to the 2ℓss SR, except that it requires exactly one medium b-tagged jet and fewer jets than the 2ℓss SR (which guarantees that this region will not overlap with the SR). This CR is dominated by the nonprompt and flip backgrounds, and plots for various kinematic distributions in this CR are shown in Appendix D. A dedicated CR for the charge flips is also defined, and this CR is described in Section 8.2. In order to study the diboson background, a 3ℓ CR is defined. This CR is similar to the 3ℓ SR, except that we require exactly zero medium b tags (to guarantee that there is no overlap with the SR). This CR is dominated by diboson events, and plots for various kinematic distributions in this CR are also shown in Appendix D. We additionally define a 2ℓos CR that is dominated by Drell-Yan (DY) events, as well as a 2ℓos CR that is dominated by $\mathrm{t\bar{t}}$ events. While these processes do not represent significant backgrounds for our SRs, it is useful to study these relatively pure CRs as a cross check of the data-to-MC corrections. Plots for various kinematic distributions in these CRs are also shown in Appendix D.

8.1 Nonprompt background

The nonprompt background arises when leptons that are not prompt pass the tight selection criteria defined in Section 5.3. As discussed in Section 5.3, a prompt lepton is a lepton that is produced directly from the hard scatter event (e.g. in the decay of W boson in a $t\bar{t}W$ event); a nonprompt lepton is a lepton that is produced through the decay of a particle that is not part of the hard scatter event (e.g. in the decay of a hadron associated with a jet arising from the hadronization of a b quark). The procedure of estimating the nonprompt contribution is developed and performed by the $t\bar{t}H$ multilepton team, and the measured probabilities are shared between the $t\bar{t}H$ analysis [39] and the analysis described in this thesis [43], which also share a synchronized object selection. The nonprompt estimation involves two main steps: the measurement of the probability for nonprompt leptons to pass the tight selection, and the application of these probabilities to a set of events in a sideband of the signal region in order to estimate the contribution in the signal region.

To measure the probability for a nonprompt lepton to pass the tight selection, we first identify a set of events that is dominated by nonprompt leptons. Referred to as the measurement region (MR), the data for this sample is collected with a set of single lepton nonisolated triggers (listed in [43]). The selected events are required to contain exactly one lepton that passes the fakeable selection criteria defined in Section 5.3 and at least one jet. The leptons in this collection are then subdivided based on whether the fakeable lepton passes or fails the tight selection criteria. If this sample were composed entirely of nonprompt leptons, the probability f for a nonprompt lepton to pass the tight selection would be simply

$$f = N_{\text{pass}}/(N_{\text{pass}} + N_{\text{fail}}), \tag{8.1}$$

where N_{pass} and N_{fail} are the number of events where the fakeable lepton passes or fails the tight selection, respectively. However, the MR also contains a small contribution from processes that produce genuine prompt leptons (e.g. W+jets). In order to account for this contamination, the sample is binned according to a variable (referred to as m_T^{fix} and defined in [43]) that is designed to discriminate between the processes of interest (the multijet contribution) and the background processes (e.g. W+jets); a fit is performed in different p_T and η regions in order to extract the N_{pass} and N_{fail} for each region. We can then use Eq. 8.1 to obtain the probability f for a nonprompt lepton to pass the tight selection.

Once the probabilities f have been measured, the next step is to use these probabilities to estimate the nonprompt contribution to the signal region. In order to obtain this estimation, the probabilities are applied to events in a sideband of the signal region, which is referred to as the application region (AR). Orthogonal to the signal region, the AR requirements are identical to signal region categories (defined in Chapter 6) except that at least one of the leptons is required to fail the tight requirements. From the number of events observed in the AR and the measured probability f for a nonprompt lepton to pass the tight selection, we can work backwards to obtain the estimation for the contribution in the signal region.

For example, let us consider the two-lepton case. The total number of events that make it into the signal region (i.e. have two tight leptons) where at least one of the leptons is nonprompt can be written as

$$N_{\text{SR}} = f N_{\text{1np}} + f^2 N_{\text{2np}}, \tag{8.2}$$

where N_{1np} is the number of events with exactly one nonprompt lepton, N_{2np} is the number of events with exactly two nonprompt leptons, and f is the probability for a nonprompt lepton to pass the tight selection. Here it is assumed that the probability for a true prompt lepton to pass the tight selection is 1. We do not know N_{1np} or

$N_{\rm 2np}$, but we do know the number of events with exactly two fakeable leptons where one of the leptons passes the tight selection (which we can call $N_{\rm tf}$) and the number of events with exactly two fakeable leptons where neither pass the tight selection (which we can call $N_{\rm ff}$), as these are the events we observe in the AR. We can then write $N_{\rm tf}$ and $N_{\rm ff}$ in terms of $N_{\rm 1np}$, $N_{\rm 2np}$, and f (which will allow us to solve for $N_{\rm 1np}$ and $N_{\rm 2np}$):

$$N_{\rm tf} = (1-f)N_{\rm 1np} + 2f(1-f)N_{\rm 2np} \tag{8.3}$$

$$N_{\rm ff} = (1-f)(1-f)N_{\rm 2np} \tag{8.4}$$

Solving Eq. 8.4 for $N_{\rm 2np}$, we can plug the result into Eq. 8.3 and solve for $N_{\rm 1np}$. Since we now have expressions for $N_{\rm 2np}$ and $N_{\rm 1np}$ in terms of known quantities ($N_{\rm tf}$, $N_{\rm ff}$, and f), we can plug these expressions into Eq. 8.2 in order to express $N_{\rm SR}$ in terms of known quantities:

$$\begin{aligned} N_{\rm SR} &= f\left(\frac{N_{\rm tf}}{1-f} - \frac{2fN_{\rm ff}}{(1-f)^2}\right) + f^2\left(\frac{N_{\rm ff}}{(1-f)^2}\right) \\ &= \frac{f}{1-f}\,N_{\rm tf} - \left(\frac{f}{1-f}\right)^2 N_{\rm ff}, \end{aligned} \tag{8.5}$$

where the quantity $\frac{f}{1-f}$ is usually referred to as F. We have thus written the estimation for the nonprompt contribution to the signal region in terms of the probability f for a nonprompt lepton to pass the tight selection (which is measured in the MR as described above) and the number of data events observed in the AR. In this example calculation, we have made the simplifying assumption that all leptons have the same f, and have only considered the two-lepton case. The expressions resulting from the full calculation (for the two-lepton and 3-lepton cases) are shown in Eqn. 13 and 14 of [43]; the weights expressed in these equations are then applied to the events in the AR in order to estimate the nonprompt contribution to each of the SR categories.

In order to account for the fact that the probability of a prompt lepton to pass the tight selection is not actually one, the size of the effect is estimated with MC samples and subtracted from the nonprompt contribution. As described in Section 6.2, MC truth information is utilized to require that the leptons in simulated samples are prompt; events from these samples that fall into the AR thus represent an estimation of the contribution of events containing prompt leptons that fail the tight selection criteria. The contribution of these samples in the AR is subtracted from the nonprompt yield in a procedure that is known as prompt subtraction.

8.2 Charge misidentification background

When the charge of one of the leptons in 2ℓos event is mismeasured, the event can enter the 2ℓss signal region, contributing to a reducible background referred to as the "charge flip" background. The charges of particles traversing through the detector are determined by the curvature of the tracks, so charge flips occur when the curvature of the track is incorrectly identified. One reason why this may occur is if a radiated photon converts into an electron-positron pair, complicating the reconstruction and possibly resulting in an incorrectly reconstructed charge. Since the charge misidentification rates for muons are much smaller than electrons, the charge flip background is only relevant for electrons.

The charge flip probabilities are expected to be larger for electrons of higher p_T, since the tracks are straighter, making it more difficult to determine the direction of curvature. The charge flip rates are also expected to be larger in the endcap than in the barrel region. For these reasons, we measure the charge flip probabilities in bins of p_T and $|\eta|$, following the approach outlined in Ref. [60].

The measurement of the charge flip probability is performed with MC DY and $\mathrm{t\bar{t}}$ samples. We count the number of electrons that pass the tight requirements (defined in section 5.3) that have had their charges mismeasured (according to the MC truth information). In addition to the tight lepton requirement, we also apply the tight charge requirement (as discussed in 6.2.1), since this requirement is applied to all electrons in our 2ℓss categories. The measurement is performed for each UL period separately, using the DY and $t\bar{t}$ samples listed in tables A.11, A.12, A.13, and A.14 of Appendix A. The measured charge flip probabilities for each year are shown in figure 8.1.

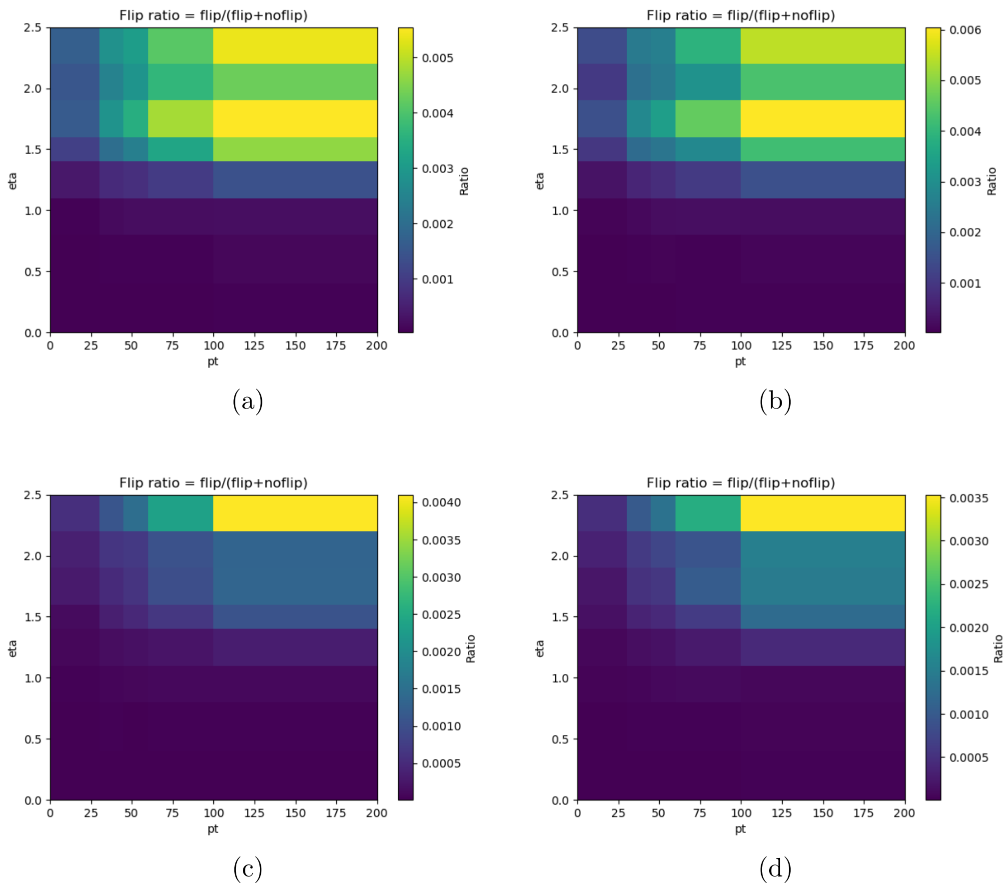

Figure 8.1. Charge flip probabilities calculated with DY and $t\bar{t}$ samples using MC truth information for UL16APV samples (a), UL16 samples (b), UL17 samples (c), and UL18 samples (d). The rates are binned according to the p_T and $|\eta|$.

Once the charge flip probabilities have been measured, we assess their validity in a dedicated charge flip CR. The charge flip CR is designed to be dominated by charge flip events. We require two tight electrons within 30 GeV of the Z peak. We make no requirement on the number of b tags, but we require fewer than four jets (to maintain orthogonality to the 2ℓss SR). The charge flip contribution is determined by scaling opposite-signed events in this region by the measured charge flip probabilities. The charge flip factor for the event corresponds to the probability that the charge of the first electron is mismeasured (and that the charge of the second electron is not mismeasured), or that the charge of the second electron is mismeasured (and the charge of the first electron is not mismeasured). We therefore have the following event weight w:

$$w = p_1(1 - p_2) + p_2(1 - p_1), \tag{8.6}$$

where p_1 is the probability that the charge of the first electron is mismeasured, and p_2 is the probability that the charge of the second electron is mismeasured. Assuming that the probabilities are small enough that terms of order $p_i \cdot p_j$ may be neglected, the charge flip probability for the event becomes the following:

$$w = p_1 + p_2. \tag{8.7}$$

The charge flip probabilities p_i for electrons are taken from the measurements described above. Since the charge flip probabilities for muons are assumed to be negligible, the flip probabilities are taken to be zero for all muons. In principle, some same-sign events should also migrate into the opposite-sign categories, but since the number of opposite-sign events is much large than same-sign events, this contribution may be neglected.

The prediction we obtain from applying the charge flip rates to opposite-sign events can then be compared to the actual same-sign data in the charge flip CR. For completeness, we include all MC backgrounds as well as the nonprompt background in the comparison. Similar to what was observed in [60], we see that our charge flip estimation over-predicts in the UL16 and UL16APV periods, and under-predicts in the UL17 and UL18 periods. To account for these differences, we apply per-year scaling factors to the charge flip probabilities (as was similarly implemented in [60]). The scaling factors are shown in Table 8.1:

TABLE 8.1

CHARGE FLIP SCALING FACTORS.

Year	Scaling factor
UL16APV	0.79
UL16	0.81
UL17	1.22
UL18	1.12

Plots from the charge flip CRs (after applying the scaling factors in Table 8.1) are shown in Appendix D in Figures D.18, D.17, D.19, and D.20. To account for the uncertainty in the measurement of the charge flip contribution, we apply a 30% flat rate uncertainty on the charge flip contribution.