CHAPTER 3

DATA AND MONTE CARLO SAMPLES

The data samples used in this analysis are described in Section 3.1. The Monte Carlo (MC) simulated samples are described in Section 3.2, with the generation of the EFT samples (with the necessary EFT weights) covered in 3.2.1, and the quadratic parametrization of the weights detailed in 3.2.2. All samples used in this analysis are in the v9 NanoAOD format [10] with Ultra Legacy (UL) reconstruction [11].

3.1 Data samples and triggers

This analysis uses data from proton-proton collisions at $\sqrt{s} = 13$ TeV collected by the CMS experiment from 2016 to 2018, using the subset of lumisections that have been certified by CMS as good for physics analysis (Table A.1). The total integrated luminosity is 138 fb^{-1} with an uncertainty of 1.6% fb^{-1} [12].

The data used in this analysis are collected with a combination of single, double, and triple-lepton triggers. The p_T thresholds for the various single-lepton triggers range from 22 to 35 GeV. The p_T thresholds for the double and triple-lepton triggers are generally not as high as the single-lepton thresholds, since events with multiple leptons are more rare, so the p_T thresholds may be lowered without resulting in too high of a trigger rate; for example, the triple-muon trigger has p_T thresholds of 12, 10, and 5 GeV. Sets of related triggers are grouped into categories referred to as datasets. The datasets and associated triggers used in this analysis are listed in Table A.2, A.3, and A.4 (for 2016, 2017, and 2018, respectively).

While the triggers in a given dataset are exclusive (i.e. an event may never pass more than one trigger in a given dataset), triggers from different datasets may overlap. This overlap between must be accounted for in order to avoid double counting. The following procedure is used:

- An arbitrary order of the datasets from a given year is chosen.
- An event that is from the first dataset (dataset A) is never discarded.
- An event that is from the second dataset (dataset B) is discarded if it passes any of the triggers from dataset A (since it was already accounted for in A).
- An event that is from the third dataset (dataset C) is discarded if it passes any of the triggers from dataset A or dataset B (since it was already accounted for).
- The procedure continues for all of the datasets that are included in the given data-taking period.

The orders of the datasets listed in Tables A.2, A.3, and A.4 correspond to the order used in the overlap removal procedure implemented in this analysis, and Appendix A.1 steps through the procedure for an example event.

3.2 Monte Carlo samples

This analysis aims to study dimension-six EFT effects on processes in which one or more top quarks are produced in association with additional charged leptons; processes which lead to the same multilepton final-state signatures but are not impacted by these EFT operators are backgrounds for this analysis. The expected background contributions are estimated using a combination of simulated samples and data-driven techniques, discussed in Chapter 8 (with the simulated samples used in the background estimation listed in Appendix A).

The expected yield for a given selection is calculated as

$$\text{Expected yield} \;=\; \sigma\,\mathcal{L}\;\frac{\sum_{\text{Pass}} w}{\sum_{\text{Gen}} w}, \tag{3.1}$$

where σ is the inclusive SM cross section for the given process, $\mathcal{L}$ is the integrated luminosity, w are the event weights. Conceptually, the event weights represent how much a given event contributes to the overall cross section. The sum in the numerator ($\sum_{\text{Pass}} w$) is over the events that pass the given selection criteria, and the sum in the denominator ($\sum_{\text{Gen}} w$) is over all generated events. The ratio of these sums corresponds to the acceptance times efficiency. As will be explained in the following sections, the weights of the signal samples are functions of the WCs; The details of the signal sample generation are described in Section 3.2.1, and 3.2.2 covers the details that are specific to the EFT weights. After the EFT weights have been explained, we will revisit Eq. 3.1 (in Section 3.2.2 Eq. 3.6), discussing the the subtleties of the normalization that are specific to this analysis.

3.2.1 Monte Carlo generation of signal samples

The signal processes for this analysis are $t\bar{t}H$, $t\bar{t}l\nu$, $t\bar{t}l\bar{l}$, $tl\bar{l}q$, tHq, and $t\bar{t}t\bar{t}$. The signal samples are produced at leading order (LO) with the MadGraph [13] event generator (version 2.6.5). As discussed in Chapter 2, the dim6top UFO model [8] is used to incorporate the EFT effects. Parton showering and hadronization for the samples are performed with the Pythia generator [14], which also handles the decays of the top quark and the Higgs boson. In order to avoid overlap between the $t\bar{t}l\bar{l}$ and $t\bar{t}H$ samples, we specify in the MadGraph process card that the $t\bar{t}l\bar{l}$ process should not include an intermediate H; the same requirement is made for the $tl\bar{l}q$ process in order to avoid overlap with tHq. All simulated signal processes are normalized to theoretical SM cross sections at next-to-leading order (NLO) in QCD, as listed in Table A.5. The EFT samples produced for this analysis are stored at the Notre Dame T3. Tor reference their file paths are listed in Tables A.6, A.7, A.8, and A.9.

For the $t\bar{t}X$ processes ($t\bar{t}H$, $t\bar{t}l\nu$, and $t\bar{t}l\bar{l}$), we include an additional final state parton in the matrix element (ME) generation. The inclusion of the additional parton

can improve the modeling at high jet multiplicities, and can also significantly impact the dependence of the $t\bar{t}X$ processes on the WCs [9]. The primary factors contributing to the modification of the cross section's EFT dependence are related to the new quark-gluon initiated diagrams that become available when an additional final state parton is included in the ME calculation. For example, without an extra parton, $c_{\varphi t}$ can only contribute to $t\bar{t}H$ via quark-anti-quark initiated diagrams (e.g. the diagram in the lefthand side of Figure 3.1); however, when an extra parton is included, $c_{\varphi t}$ can contribute via quark-gluon initiated diagrams like the one shown in the righthand side of Figure 3.1. Other factors, such as the chiral and color structure of the operator, can also play an important role.

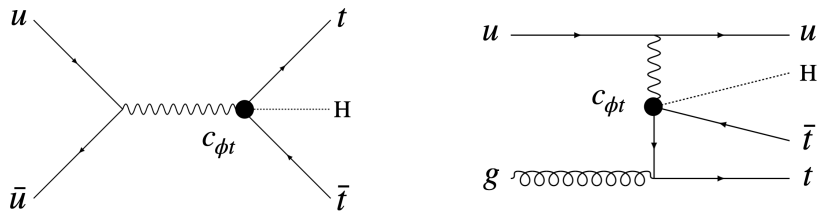

Figure 3.1. Example $c_{\varphi t}$ diagrams for $t\bar{t}H$ without and with an extra parton.

The single top processes ($t\bar{l}\bar{l}q$ and tHq) and the $t\bar{t}t\bar{t}$ sample are not produced with an additional parton. The single top processes have technical complications associated with correctly performing the jet matching between the ME and the parton shower (PS) for t-channel single top processes that currently do not allow a valid matched sample to be produced. In the case of $t\bar{t}t\bar{t}$, an additional parton is not included because the generation of the MadGraph gridpack is very computationally expensive. It would not be feasible to produce enough $t\bar{t}t\bar{t}$ samples to perform a thorough validation of the starting point and matching parameters validation. How-

ever, the effect of an additional parton for $t\bar{t}t\bar{t}$ is not expected to be very significant, since the $t\bar{t}t\bar{t}$ process already naturally populates the high-jet multiplicity bins of the analysis. Furthermore, $t\bar{t}t\bar{t}$ is already dominated by gluon-gluon initiated diagrams, so quark-gluon initiated diagrams would be expected to have a smaller impact.

Since we are unable to include an additional parton for the single top samples, and in these cases the extra parton may potentially have a significant effect on the high jet multiplicity categories (since these single top processes would not generally produce as many jets as our other signal processes), we apply an additional uncertainty to these processes, described in Chapter 9. This uncertainty is determined by comparing the jet multiplicity distribution of our EFT samples (reweighted to the SM) against SM NLO samples, listed in Table A.10.

For the samples produced with an additional parton, a matching procedure must be applied to account for the overlap in phase space between the contributions of the ME and parton shower (PS). For this analysis, the matching is implemented using the MLM scheme [15], an event-rejection based approach that matches ME partons to jets clustered by Pythia, discarding events in which the jets are not successfully matched to partons in order to avoid double counting.

It should be noted that the matching procedure can lead to complications when applied to EFT samples; since EFT effects are included in the ME contribution, but not in the PS contribution, it is possible that an inconsistency could arise. Specifically, if an EFT vertex produces a significant soft and collinear contribution, the events removed by the matching procedure will never be replaced by corresponding events generated by the PS, causing this contribution to be missed. However, of the WCs considered in this study, the operator associated with the c_{tG} WC is the most prone to these effects, and its contributions to the soft and collinear regime are suppressed; thus, the phase space overlap with the SM contribution from the PS is small, and the effects of this potential issue are negligible [9].

In addition to the theoretical justification outlined above, we can validate the matching procedure empirically by examining differential jet rate (DJR) distributions for the simulated samples. The resulting DJR distributions provide further evidence that the matching is working properly. The details of the validation of the DJR distributions may be found in Appendix B.1, and a more detailed discussion of the validation of matched $\mathrm{t\bar{t}X}$ samples is presented in [9].

As an additional form of validation, our EFT signal samples are reweighted to the SM and compared against SM samples that are centrally produced by the CMS collaboration. The details of this comparison are presented in Appendix **??**. The comparisons show that the level of agreement is generally good, providing further evidence that the reweighting is working properly and that the LO modeling (and associated uncertainties) are sufficient for this analysis.

3.2.2 Parameterization of the predicted yields in terms of the WCs

This section will describe the method through which the predicted yields are parameterized in terms of the WCs. In order to write the predicted yields as a function of the WCs, it is first necessary to understand how the cross section depends on the WCs. Starting with the ME, we can write the amplitude for a given process as the sum of the SM and new physics components:

$$\mathcal{M} = \mathcal{M}_{\mathrm{SM}} + \sum_{i} \frac{c_i}{\Lambda^2} \mathcal{M}_i, \tag{3.2}$$

where $\mathcal{M}_{\mathrm{SM}}$ is the SM ME, $\mathcal{M}_i$ are the MEs corresponding to the new physics components, and c_i are the WCs. Since the cross section (inclusive or differential) is proportional to the square of the ME, it will depend quadratically on the WCs:

$$\sigma \propto |\mathcal{M}|^2 \propto s_0 + \sum_{j}^{N} s_{1j} \frac{c_j}{\Lambda^2} + \sum_{j}^{N} s_{2j} \frac{c_j^2}{\Lambda^4} + \sum_{j \neq k}^{N} s_{3jk} \frac{c_j}{\Lambda^2} \frac{c_k}{\Lambda^2}, \tag{3.3}$$

where s_{jk} are structure constants of the N-dimensonal quadratic function for N WCs. The number of structure constants (K) required to describe an N-dimensional quadratic can be written as the following:

$$K = \frac{(N+1)\cdot(N+1)-(N+1)}{2} + (N+1). \tag{3.4}$$

This analysis considers 26 WCs, so by Eq. 3.4, there are 378 structure constants required to fully describe the 26-dimensional quadratic. In principle, we could solve for these structure constants if the cross section at 378 points in the 26-dimensional WC space were known. However, this would require generating 378 unique simulated samples at 378 unique points in the 26-dimensional WC space. In practice, it would not feasible to generate this many simulated samples.

Instead of attempting to determine the parametrization for the inclusive cross section, we parametrize each event's weight in terms of the WCs. Since each weight corresponds to the event's contribution to the inclusive cross section, the event weight essentially represents a differential cross section, which can be described by a 26-dimensional quadric in terms of the WCs, as written in equation 3.3. In order to determine the 378 structure constants of the event weight's quadratic parameterization, we need to know the event weight at 378 distinct points in the 26 dimensional space. This is feasible to do using the MadGraph event reweighting [16] procedure.

Given an event generated under a specific theoretical scenario, the MadGraph event reweighing procedure computes additional weights associated with the same event under alternative theoretical scenarios. In the case of EFT reweighting, the original theoretical scenario corresponds to a particular point in the 26-dimensional WC space, provided to MadGraph by the user. We refer to this as the "starting point" for the sample. The alternative theoretical scenarios correspond to other distinct points in the 26-dimensional WC space (i.e. other sets of values for the 26 WCs),

also provided to MadGraph by the user. From the matrix-element computations, MadGraph calculates the weight at the starting point and at each of the additional reweight points. With at least 378 weights corresponding to 378 independent points in the 26-dimensional WC space, we can solve for the 26 structure constants, and fully determine the 26-dimensional quadric function that describes the event's weight in terms of the WCs.

Once we have obtained each event's 26-dimensional quadratic parametrization $w_i(\vec{c}/\Lambda^2)$, we can find the dependence of any observable bin on the WCs by summing the quadratic parameterizations for each of the events that passes the selection criteria for the given bin. The sum of the weights for the passing events ($\sum_{\text{Pass}} w$ from Eq. 3.1) can thus be written follows:

$$\begin{aligned} \sum_{\text{Pass}} w &= \sum_i w_i \left(\frac{\vec{c}}{\Lambda^2} \right) \\ &= \sum_i \left(s_{0i} + \sum_j s_{1ij} \frac{c_j}{\Lambda^2} + \sum_j s_{2ij} \frac{c_j^2}{\Lambda^4} + \sum_{j \neq k} s_{3ijk} \frac{c_j}{\Lambda^2}, \frac{c_k}{\Lambda^2} \right), \end{aligned} \tag{3.5}$$

where the sum over i corresponds to the sum over all of the events that pass the selection criteria for the given bin. Performing a similar sum over all events in the sample, we can obtain the $\sum_{\text{Gen}} w$ term from Eq. 3.1. Since the sum of multiple quadratic functions is also quadratic, both $\sum_{\text{Pass}} w$ and $\sum_{\text{Gen}} w$ are quadratic in terms of the WCs.

In this analysis, we evaluate $\sum_{\text{Gen}} w$ at the SM (effectively canceling the LO cross section). We then normalize the predicted yield to a more precise (i.e. NLO or better) SM cross section calculation. Rewriting Eq. 3.1 with these nuances included, the expected yield in any bin can be expressed as a function of the WCs as follows:

$$\text{Expected yield } (\vec{c}) \;=\; \sigma_{SM}\, \mathcal{L}\; \frac{\sum_{\text{Pass}} w(\vec{c})}{\sum_{\text{Gen}} w(SM)}, \tag{3.6}$$

Where $\vec{c}$ are the WCs, σ_{SM} is the inclusive cross section from an NLO (or better) calculation, $\sum_{\text{Pass}} w(\vec{c})$ is the sum of the event weight parameterizations of the passing events, and $\sum_{\text{Gen}} w(SM)$ is the sum of the event weight parameterizations for all generated events (evaluated at the SM).

Since we are thus able to write the predicted yield of any observable bin as a function of the 26 WCs, we can obtain detector-level predictions at any arbitrary point in the 26-dimensional EFT space. This is the key enabling concept of this analysis, as it allows for all EFT effects across all analysis bins to be simultaneously accounted for when performing the likelihood fitting with the statistical framework (which will be described in Chapter 10).

We generate all of our signal processes using the reweighting procedure described in this section. However, we do not include all WCs for all processes (since some of the WCs do not impact some of the processes), so the number of reweight points included in the MC generation varies by sample. The $t\bar{t}t\bar{t}$ process incorporates the full set of 26 WCs. By Eq. (3.4), a total of 378 weights are required to fully determine the 26-dimensional quadratic parameterization. However, in order to ensure that a good fit can be found, we over-constrain the fit by including approximately 20% more points than the minimum number required, for a total of 454 reweight points. As discussed in Section 2.2, the other five signal samples have a negligible dependence on the four four-heavy operators to which $t\bar{t}t\bar{t}$ is sensitive, so these samples incorporate only 22 WCs. This means a minimum of 276 reweight points are required to determine the 22-dimensional quadratic fit, but we again ensure the fit is over-constrained by generating additional reweight points, for a total of 332 reweight points for each event.

The MadGraph reweighting procedure is powerful because it allows different regions of EFT phase space to be probed with a single MC sample; however, there is an important caveat to the procedure that should be highlighted. Since MadGraph

produces unweighted samples of events, the events generated by MadGraph mainly correspond to phase space occupied by original event. Thus, the reweighting procedure does not work unless the original point in phase space (i.e. the starting point) and the alternative points in phase space (i.e. the reweight point) have some overlap. EFT operators lead to new diagrams that may populate areas in phase space that are not present in the SM, therefore the SM cannot be used as a valid starting point for the reweighting procedure. Instead, a point that is relatively far from the SM should be chosen.

Nevertheless, even for non-SM starting points, there is still no guarantee that the chosen point will allow MadGraph to properly reweight to all areas of relevant phase space. Therefore, it is important to validate reweighted samples to ensure that they are able to be consistently reweighted to as much of the relevant phase space as possible. For example, we check that the samples are able to be consistently reweighted to other points in EFT phase space (by comparing against dedicated samples produced at the given point in phase space), as well as checking the distribution of event weights for samples generated at different starting points. Details regarding the validation checks performed for this analysis are provided in Appendix B.2.