technische universität
dortmund

Arbeit zur Erlangung des akademischen Grades
Bachelor of Science

Studies for the sensitivity to dimension six operators in the context of effective field theories in single-top-quark production with a photon

Jan Lukas Späh
geboren in Witten

2019

Lehrstuhl für Experimentelle Physik IV
Fakultät Physik
Technische Universität Dortmund

Erstgutachter:	Prof. Dr. Kevin Kröninger
Zweitgutachter:	Prof. Dr. Dr. Wolfgang Rhode
Abgabedatum:	15. Juli 2019

Abstract

In this thesis, single-top-quark production with a photon is investigated in Monte Carlo simulations with an effective field theory approach. Samples with Wilson coefficients that influence the electroweak couplings of the top-quark, a Standard Model sample and a sample of top-quark-antiquark pair production with a photon are generated. The two latter samples are defined as background processes B while the samples with effective field theory contributions are defined as signal S. The samples undergo a simulation of parton showering and hadronisation and a detector simulation is performed. It is established that boosted decision trees allow a classification of signal and background based on kinematic input variables. The transverse momentum of the leading photon is the feature with the highest discrimination power compared to single-top-quark production with a photon in the Standard Model. The outputs of the classifiers are compared to the distributions of the most discriminating features using bin-wise expected significances $S_i/\sqrt{B_i}$. The combination of multiple variables into a single discriminant is shown to be superior to cuts on single variables as the highest achievable expected significances range from 17.6 to 35.6 for the classifier outputs compared to 4.5 to 11.4 for the kinematic variable distributions. It is expected that these values will be significantly lower when considering the background processes neglected in this thesis.

Kurzfassung

In dieser Arbeit wird die Produktion von einzelnen Top-Quarks mit einem Photon in Monte-Carlo-Simulationen mit einem Ansatz basierend auf effektiven Feldtheorien untersucht. Stichproben mit Wilson-Koeffizienten, welche die elektroschwachen Kopplungen des Top-Quarks verändern, eine Standardmodell-Stichprobe und eine Stichprobe von Top-Quark-Antiquark-Paarproduktion mit einem Photon werden generiert. Die beiden letztgenannten Stichproben werden als Untergrundprozesse B definiert, während die Stichproben mit Beiträgen von effektiven Operatoren als Signal S definiert werden. Die Stichproben werden einer Simulation von Partonschauern und Hadronisierung unterzogen und eine Detektorsimulation wird durchgeführt. Es wird festgestellt, dass *boosted decision trees* eine Klassifizierung von Signal und Hintergrund basierend auf kinematischen Eingangsgrößen ermöglichen. Der Transversalimpuls des führenden Photons ist das Merkmal mit der größten Trennkraft gegenüber der Einzelproduktion von Top-Quarks mit einem Photon im Standardmodell. Die Ausgaben der Klassifikatoren werden mit den Verteilungen der diskriminierendsten Merkmale unter Verwendung der erwarteten Signifikanzen $S_i/\sqrt{B_i}$ pro Bin verglichen. Die Kombination mehrerer Variablen zu einer einzigen Diskriminante erweist sich als überlegen gegenüber Schnitten auf einzelne Variablen, da die höchstmöglichen erwarteten Signifikanzen bei 17,6 bis 35,6 für die Klassifikatorausgaben im Vergleich zu 4,5 bis 11,4 für die Verteilungen der kinematischen Variablen liegen. Es wird erwartet, dass diese Werte bei der Berücksichtigung der Untergrundprozesse, die in dieser Arbeit vernachlässigt wurden, deutlich geringer sein werden.

Contents

1 Introduction

The Standard Model (SM) is widely regarded as the most successful theory describing elementary particles and their interactions. Still, there remain phenomena, such as dark matter, that the SM is not able to explain. Together with conceptual problems, e.g. its incompatibility with general relativity, these limitations may lead to the belief that the SM is only valid at the energy regions probed by current and past experiments and is only the low-energy limit of a superordinate theory. Recently, there has been increasing interest in the interpretation of results from top-quark physics through model-independent parameterisations, one of which is effective field theory (EFT). The idea of an EFT is that new physics, such as new massive particles, might only manifest themselves directly at an energy scale much larger than the current centre-of-mass energies at collider experiments. EFTs introduce operators modifying existing couplings or introducing new vertices and are able to predict their effects quantitatively without making assumptions about the specific underlying mechanisms.

Searches for physics beyond the SM (BSM) often involve the top-quark, the most massive elementary particle. It couples to all fundamental interactions considered in the SM and the study of its decay properties allows detailed investigation of the weak interaction in particular. Large numbers of top quarks can be produced singly or in quark-antiquark pairs in proton-proton collisions at the Large Hadron Collider, currently the world's most powerful particle accelerator. The former process, referred to as single-top-quark production, was discovered in 2009 at the Tevatron [1–3]. A photon can be radiated in this process as well, which is then referred to as $tq\gamma$, making it possible to investigate both the coupling of the top-quark to the W boson and the photon. Evidence for this process was published in December 2018 by the CMS Collaboration [4].

The Monte Carlo (MC) simulation framework MadGraph5_aMC@NLO [5] is used to generate the samples and multivariate analysis (MVA) methods are employed to study the discrimination between EFT enriched $tq\gamma$ samples and exemplary SM background processes. The goal is to identify the effects of the EFT operators on the distributions of kinematic variables through feature importances, to show that the classification is possible in principle and to compare the accuracies of boosted decision trees (BDT) and linear discriminant analysis (LDA) for this task. The significance of the results shall be discussed with regard to the assumptions made in this thesis, e.g. the omission of some underground processes.

2 Brief overview of the Standard Model and effective field theories in the top-quark sector

In this chapter, the Standard Model (SM) of particle physics is briefly summarised in order to lay the theoretical foundations for this thesis. After that, some limitations of the SM are presented in order to motivate the search for physics beyond the Standard Model (BSM) with effective field theories (EFT), which will be introduced subsequently. Afterwards, the role of the top-quark in the search for BSM phenomena through an EFT approach is emphasised. Finally, the process investigated in this thesis, the single-top-quark production in association with a photon, is described and connected to EFTs.

2.1 A brief summary of the Standard Model

The SM is currently the most successful theory of elementary particles and three fundamental interactions between them: the electromagnetic, the weak and the strong interactions. It is formulated mathematically as a quantum field theory (QFT), is able to explain measurements convincingly and predicted a number of phenomena which were confirmed by experiments.

An overview of the elementary particles and their interactions in the SM is shown in Figure 2.1. The Particles are distinguished by their quantum numbers, one of which is spin. The spin allows the classification of particles into fermions, which have half-integer spin ($s = 1/2, 3/2, \ldots$), and bosons with integer spin ($s = 0, 1, \ldots$). The elementary fermions are spin-1/2 particles and are known as matter-particles. The elementary spin-1 bosons are also called gauge bosons and mediate the fundamental interactions of the SM.

Fermions are further divided into leptons and quarks and each fermion is assigned an antiparticle with opposite charges. Both are classified into three generations with increasing masses and decreasing lifetimes. Each charged lepton is assigned an uncharged neutrino, which is assumed to be massless in the mathematical formulation of the SM. Charged leptons participate in the electromagnetic interaction and can be converted into each other by the weak interaction.

Quarks interact through all three fundamental interactions of the Standard Model. They carry the charge of the strong interaction which is called colour charge. The strong force is mediated by eight massless gluons that can also couple to themselves like the weak bosons because they carry colour charge themselves. The characteristic property of this interaction is that only colour-neutral particles are observable since the force between two colour-charged particles increases linearly with their distance. This phenomenon is called confinement. Thus, energetic unbound quarks only exist for a short time. Quark-antiquark pairs are created out of the vacuum while the quark is slowed down. Additionally, the quarks may fragment and emit gluons due to these violent accelerations. These processes lead to bound states of quarks and antiquarks (hadrons). A narrow stream of particles called jet is the result of this hadronisation.

The photon is the gauge boson of the electromagnetic interaction described by quantum electrodynamics (QED). It is massless and couples to electrically charged particles.
The weak interaction is mediated through three gauge bosons, the electrically charged $W^{\pm}$ bosons and the electrically neutral Z boson. The W bosons are the only particles capable of changing the so-called flavour of particles. Flavour is a quantum number of leptons and quarks and corresponds to the type of particle present. The transition probability between two quarks i and j is given by $|V_{ij}|^2$, where the V_{ij} are the elements of the quark mixing

Cabibbo–Kobayashi–Maskawa (CKM) matrix. The Z boson cannot change the flavour. It is present in weak elastic scattering processes and commonly decays to fermion-antifermion pairs.

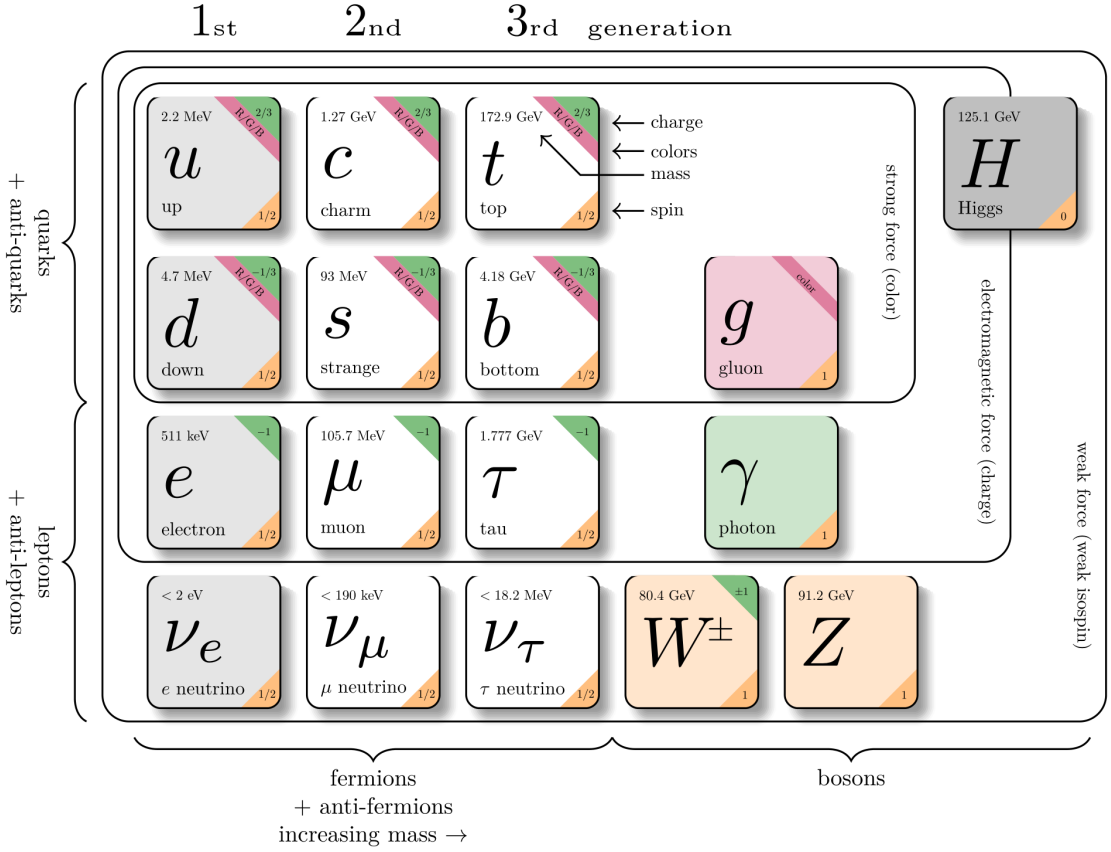


Figure 2.1: Overview of the elementary particles, their properties, and interactions in the Standard Model [6]. The masses are quoted from [7] and the direct measurement of the top-quark mass is quoted.

The dynamics of the particle fields are described by the Lagrangian density of the QFT $\mathcal{L}$. Each QFT is connected to a symmetry group and $\mathcal{L}$ needs to be invariant under a local gauge transformation under that group. The gauge bosons of the three fundamental interactions arise from this requirement. In addition to that, these bosons turn out to be massless. This is consistent with the photon and the gluons being massless. However, this result contradicts experimental evidence that the W and Z bosons are massive. Also, all fermions should be massless as mass terms in the Lagrangian are not gauge invariant.

This apparent problem is solved through the Higgs mechanism. The mass generation coincides with the breaking of the electroweak symmetry, meaning that the weak bosons and all fermions acquire mass. The weak bosons acquire mass through their interaction with the Higgs field while fermions gain mass through so-called Yukawa interactions between the fermion fields and the Higgs field in the Lagrangian. The four gauge bosons of the electroweak interaction are denoted as B, W^1, W^2 and W^3. These gauge eigenstates are related to the

physical mass eigenstates γ, Z and $W^{\pm}$ through both a linear combination and rotation by the Weinberg angle θ_{W}:

$$\begin{aligned} \gamma &= \sin\theta_{\mathrm{W}} W^3 + \cos\theta_{\mathrm{W}} B\,, \\ Z &= \cos\theta_{\mathrm{W}} W^3 - \sin\theta_{\mathrm{W}} B\,, \\ W^{\pm} &= \frac{1}{\sqrt{2}}(W^1 \pm iW^2)\,. \end{aligned} \tag{2.1}$$

The Higgs mechanism also predicts a spin-0 boson, called the Higgs boson. A particle consistent with the predictions for the Higgs boson was finally discovered in 2012 at the Large Hadron Collider (LHC) at CERN by the ATLAS and CMS Collaborations [8, 9]. Investigations of its properties lead to the conclusion that it is the Higgs boson with very high confidence, marking a significant success for the SM as its particle content is completely observed.

2.2 Limitations of the Standard Model

Currently, there are no experimental results considered to be contradicting the predictions of the SM by a significance level of more than five sigma[1]. This fact may lead to the belief that our current understanding of particle physics is the ultimate theory, able to explain every phenomenon observed and observable. However, the numerous precision tests of the SM only confirm its validity for the energy ranges and processes studied in the experiments. These considerations contribute to justified doubts about the validity of the SM at higher energies than the ones studied so far. The following points are often stated as important limitations and problems of the SM:

- The amount of visible baryonic mass is not sufficient to explain a number of phenomena, e.g. the rotation curves of galaxies or the stability of galaxy clusters [10]. So-called dark matter, which is measured to make up approximately 25% of the matter content of the observable universe, is postulated to solve these problems. No model is currently able to describe this consistently.

- The observable universe consists almost exclusively of matter while antimatter is present only in minute quantities. Based on current measurements of the SM parameters it seems unlikely that the generation of this asymmetry is possible within the framework of the SM [11].

- A successful, consistent incorporation of gravity into the SM has not yet been discovered. The unification of general relativity with the SM in the form of quantum gravity would be a big success, but this seems to be far away as the attempts to formulate general relativity as a QFT have been unsuccessful until today.

The search for explanations of these phenomena and problems often entails the direct search for new particles, fundamental interactions or anomalous couplings. One example is the search for supersymmetric particles. Supersymmetry is an extension of SM and is considered theoretically attractive as it would establish more connections between fermions and bosons. The lightest supersymmetric particle would be a candidate for dark matter. However, the

[1] By the convention of high energy physics (HEP), this is the significance which is needed to claim a discovery. This high limit is needed to exclude most of the statistical fluctuations, which are bound to occur from time to time as a high number of experiments are performed to test the SM.

direct search for these particles at the LHC has not yet revealed any statistically significant deviations from the SM predictions [7]. A disadvantage of these direct searches is their model dependence since the absence of a signature for a new particle predicted by a particular theory falsifies only one of many possible theories. Because of that, more general possibilities to search for BSM phenomena can be adopted, one of which is presented in the following subchapter.

2.3 Effective field theories as a model-independent approach to new physics

EFTs are a model-independent approach to describe processes in particle physics that do not describe the basic mechanisms of phenomena and are legitimised by the agreement of the experimental results with their quantitative predictions. A particular motivation for EFT is that the mechanisms of a superordinate theory, e.g. the exchange of new massive particles, may not be directly visible at the typical energy E of a measurable process that is much smaller than the typical energy scale Λ of the complete theory. The EFT then claims to describe the phenomenology in the energy range $E \ll \Lambda$ with sufficient accuracy.

To illustrate the concept with an example, the four-fermion interaction firstly described by Enrico Fermi [12] is worth mentioning. It describes the radioactive beta decay, where a neutron n decays into a proton p, an electron e^- and an electron-antineutrino $\overline{\nu}_e$. The mechanism of this decay was not known at the time Fermi developed the theory so that a pointlike interaction with an effective coupling G_F between all involved fermions was postulated. This is shown in Figure 2.2.

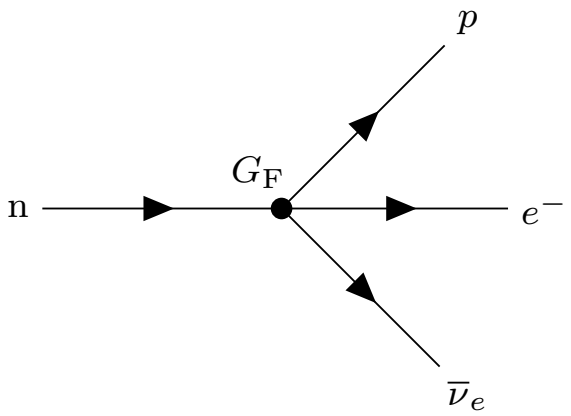


Figure 2.2: Fermi's four-fermion interaction as an effective theory of beta decay. The dot indicates that the resolution of the underlying structure of the vertex is unknown at $E \ll \Lambda$.

Today, the full theory describes beta decay through the decay of an down-quark inside the neutron to an up-quark. This happens through the weak interaction and the exchange of a virtual W boson that decays to an electron and an electron-antineutrino. Thus, the process is suppressed by the propagator of the massive W boson, which depends on the square of the W boson mass $m_W \approx 80\,\mathrm{GeV}$ that can be identified with the energy scale Λ. Typical energies E in the beta decay are at most a few MeV, where the predictions of the four-fermion interaction theory agree well with the experiment. The relation between the propagator and the effective coupling is

$$\frac{g^2}{p^2 - m_W^2} \approx \frac{g^2}{m_W^2} \approx G_\mathrm{F}\,, \tag{2.2}$$

with the weak coupling constant g and momentum transfer p from the decaying down-quark to the virtual W boson.

This theory is an example of the so-called bottom-up approach. The superordinate theory is not known in this case. It is also possible to approach this process in the opposite way. The calculation of the beta decay does not require any knowledge about the exchange of the virtual W boson. The substructure of the interaction is hidden on these energy scales so that the four-fermion theory can also be derived from the weak interaction as a low-energy limes. This is the so-called top-down-approach, which represents a possibility to parameterise and simplify mechanisms of known phenomena. This thesis pursues a bottom-up approach with effective operators in the top-quark sector in preliminary studies on BSM effects.

The usual way to formulate an EFT is to introduce effective operators $\mathcal{O}_i^{(D)}$ into the Lagrangian of the SM $\mathcal{L}_{\text{SM}}$. These operators allow new vertices or modify existing ones. An important concept in this context is the mass dimension D of an operator. The mass dimension is the inverse of the dimensions of length and time in natural units. Since the Lagrangian density is integrated over $d = 4$ space-time dimensions in order to obtain the dimensionless action, it must have mass dimension four. The coupling strength of an operator is given by the ratio of the so-called Wilson Coefficient c_i and the energy scale Λ. Therefore, the product of the effective operator and a power of the energy scale must result in a mass dimension of four. Thus, the extension of $\mathcal{L}_{\text{SM}}$ with effective operators reads

$$\mathcal{L} = \mathcal{L}_{\text{SM}} + \mathcal{L}_{\text{EFT}} = \mathcal{L}_{\text{SM}} + \sum_i \frac{c_i}{\Lambda^{D_i-4}} \mathcal{O}_i^{(D)} \,. \tag{2.3}$$

This thesis only considers operators respecting the conservation laws of the Standard Model. With the explanations in Ref. [13] and the exclusive consideration of operators with leading mass dimension, only dimension six operators remain. Beyond that only one operator is considered at the same time, so that the individual EFT contribution of an effective operator $\mathcal{O}_i$ to the Lagrangian density is $\frac{c_i}{\Lambda^2}\mathcal{O}_i$, where the mass dimension superscript is dropped. The Wilson coefficients and the energy scale are dual to one another, e.g a halved Wilson coefficient would require a Λ four times as high to get the same coupling strength. An energy scale of $\Lambda = 1\,\text{TeV}$ is assumed in the simulations so that a dimensionless, rescaled Wilson coefficient $C \equiv c/\Lambda^2$ can be understood as a coupling in units of $1/\text{TeV}^2$. The cross section σ of a process in the combined SM and EFT theory consists of the SM cross section σ_{SM}, which corresponds to a Wilson coefficient of zero, an interference term between SM and EFT diagrams described by σ_{int} and a pure EFT contribution σ_{BSM}. The dependence of the cross section on the Wilson coefficient C is quadratic under the given assumptions and reads

$$\sigma = \sigma_{\text{SM}} + \sigma_{\text{int}} C + \sigma_{\text{BSM}} C^2 \,. \tag{2.4}$$

2.4 Effective field theories in top-quark physics

The top-quark is of fundamental importance for tests of the SM and the search for BSM phenomena. It couples to all fundamental interactions and has a special relationship to the Higgs boson as it is the only fermion with a Yukawa coupling that is of the order of unity. It is also the only quark to decay before its hadronisation, making a direct study of a quark and its decay properties possible. These characteristics motivated countless BSM studies of processes involving top-quarks. The top-quark also offers possibilities to search for BSM phenomena

using EFT. Two commonly considered dimension six operators in top-quark processes are

$$\begin{aligned} \mathcal{O}_{tB} &= (\bar{q}_3 \sigma^{\mu\nu} t_{\mathrm{R}}) \widetilde{\varphi}\, B_{\mu\nu}\,, \\ \mathcal{O}_{tW} &= (\bar{q}_3 \sigma^{\mu\nu} \tau^I t_{\mathrm{R}}) \widetilde{\varphi}\, W^I_{\mu\nu}\,, \end{aligned} \tag{2.5}$$

where $\bar{q}_3$ is the left-handed third generation quark doublet and t_{R} is the right-handed top-quark. The Pauli Matrices are denoted by τ^I, $\sigma^{\mu\nu} = \frac{i}{2}[\gamma^\mu,\gamma^\nu]$ and $\widetilde{\varphi}$ is related to the Higgs field. The electroweak bosons B, W^1, W^2 and W^3 are present with their fields $B_{\mu\nu}$ and $W^I_{\mu\nu}$. These operators are given in a basis before electroweak symmetry breaking. This choice of basis is in principle arbitrary. It is also possible to express these operators in the physical states after electroweak symmetry breaking using Equations (2.1). There are also other dimension six operators affecting the top-quark, which are summarised and explained along with a general EFT analysis strategy in Ref. [14]. This thesis only considers the two given operators because they are expected to influence the cross section and the kinematics of single-top-quark production considerably. They also respect both the theoretical as well as the empirical conservation laws of the SM and thus allow for conservative parameterisation of SM extensions.

A process that allows the study of the top-photon coupling is top-quark-antiquark pair production ($t\bar{t}$) in association with a photon, denoted by $t\bar{t}\gamma$. This process was discovered in 2015 by the ATLAS Collaboration [15] and interpreted with an EFT approach by Ghneimat in 2018 [16]. In summary, the top-quark offers opportunities for model-independent searches for new physics through an EFT approach. This motivates the investigation of single-top-quark production in which the top-quark is involved and which will be explained in the following subchapter.

2.5 Single-top-quark production with a photon and its connection with effective field theories

Single-top-quark production (tq) was discovered at the Tevatron in proton-antiproton collisions separately by the D0 [1] and CDF [2] Collaborations in 2009 and combined into a single result in Ref. [3]. This process has a substantially lower cross section than $t\bar{t}$, especially at the LHC, where the gluon-gluon fusion to $t\bar{t}$ is much more likely than weak interactions between individual quarks resulting in single top-quarks. Single top-quarks can be produced in three different ways: t-channel, s-channel and W-associated single-top-quark production (tW).

In this thesis, the t-channel production is investigated with an additional photon in the final state and is called $tq\gamma$. One possible Feynman diagram for this process is shown in Figure 2.3. Single-top-quark production with a photon is not observed as of yet. The CMS Collaboration reported evidence in December 2018 with a significance of 4.4σ [4].

In the QCD leading-order (LO)[2] $tq\gamma$ process an initial up-type quark, denoted as q_i, exchanges a W boson with an initial bottom-quark. It becomes a down-type quark q_f, while the bottom-quark is converted into a top-quark[3]. The initial up-type quark can also be an antiquark of down-type that becomes an antiquark of up-type through the exchange of the W boson. Single-top-quark production with a photon is sensitive to the operators $\mathcal{O}_{tB}$ and $\mathcal{O}_{tW}$ from Equations (2.5) because three vertices are present that can be influenced by them. While

[2] LO in QCD refers to one strong vertex being present in this process.

[3] It is also possible to produce single anti-top-quarks. In general, charge conjugated processes are meant to be included.

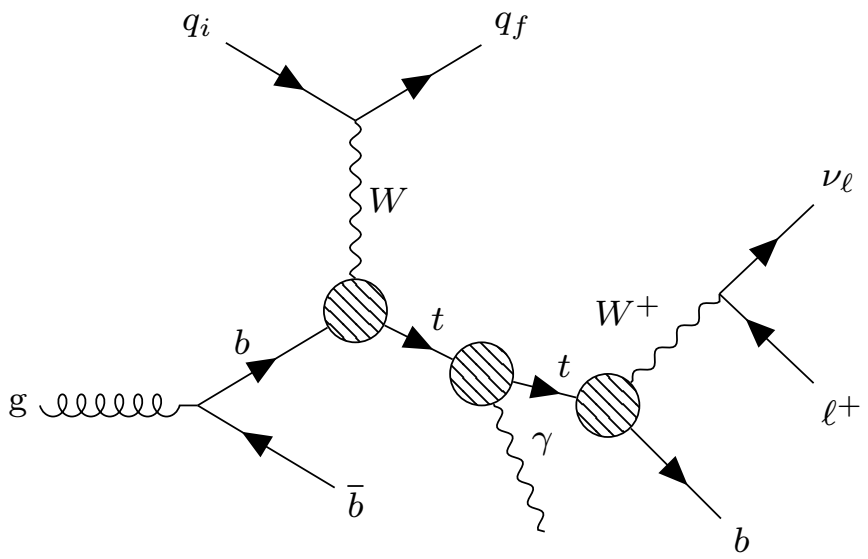


Figure 2.3: Leading-order Feynman diagram for t-channel single-top-quark production in association with a photon with leptonic W boson decay. While the operator $\mathcal{O}_{tW}$ can influence all three of the marked vertices, $\mathcal{O}_{tB}$ may only influence the photon radiation from the top-quark.

$\mathcal{O}_{tW}$ influences all three electroweak top-quark vertices, $\mathcal{O}_{tB}$ only affects photon radiation from the top quark.

In principle, it would also be possible for down- and strange-quarks to be converted into a top-quark. However, these processes are highly suppressed because of $|V_{\mathrm{tb}}|^2 \gg |V_{\mathrm{ts}}|^2 + |V_{\mathrm{td}}|^2$. In fact, this thesis assumes $|V_{\mathrm{tb}}| = 1$ and neglects these contributions entirely. The initial b-quark is shown to originate from gluon splitting in Figure 2.3, although it is also possible that the b-quark originates from an initial proton or antiproton as a sea quark [4]. The latter possibility is also neglected in this thesis. A so-called four-flavour-scheme (4FS) is adopted, meaning that it is assumed that only the four lightest quarks are present as sea quarks. A five-flavour-scheme (5FS) would include the bottom-quark as a sea quark. Also, only leptonic decays of the W boson are considered. However, the event selection in Chapter 3.6 aims to suppress events without leptonic W boson decays so that this is not a problematic approximation.

All in all, the final state of the diagram consists of a light quark, a lepton, a neutrino, two b-quarks and a photon under these assumptions. This is also referred to as matrix element or truth level.

[4] Sea quarks are virtual quarks and antiquarks in protons. Together with gluons, they often form the initial state for the production of particles in hard-scattering processes at high center-of-mass energies, such as at the LHC.

3 Event generation, validation and selection

In this chapter, the generation and processing of events and the design of the event selection is described. Checks are performed to see if the implementation of the EFT model and the processes is consistent. This is important because single-top-quark production has not been studied with an EFT approach yet so that no results for an independent validation are available.

3.1 Relevant background processes in single-top-quark production with a photon

Background processes commonly considered in the analysis of $tq\gamma$ and the expected event yields of signal and background are summarised in Table A.1, taken from Ref. [4] (Table 1). These numbers may only be understood as a rough estimation because they are calculated with a complete event selection specific to the muon channel CMS analysis. In general, background processes for $tq\gamma$ can be divided into two classes based on the origin of the photon in the event: The main background processes with a genuine photon [1] are $t\bar{t}\gamma$, $W\gamma+\text{jets}$ and $Z\gamma+\text{jets}$. In general, these notations refer to the final state of the hard-scattering process, including possible fragmentation contributions. For example, a W boson is produced in association with multiple jets and a photon in the process termed $W\gamma+\text{jets}$. The background processes containing genuine photons contribute significantly because their event signature is similar to that of $tq\gamma$. The second category consists of fake photon processes where a jet or an electron is misidentified as a photon by the detector algorithms. These processes include $t\bar{t}$, $W+\text{jets}$, where a jet is misidentified as a photon, and $Z+\text{jets}$, where a lepton from the Z boson decay is misidentified as a photon. It is difficult to estimate the contribution of fake photons with only simulations. In order to achieve a full fake photon estimation it would be necessary to perform a full detector simulation and validate the results with empirical data, which would be beyond the scope of this thesis. Thus, this category is left out in this study. As the goal is to study the influence of EFT operators and to investigate the discrimination between EFT enriched samples and SM background, samples generated with all Wilson coefficients set to zero are treated as background, while $tq\gamma$ with one non-zero coefficient is defined as signal in this analysis. The two background processes considered in the simulations are $t\bar{t}\gamma$ and $tq\gamma$, which is judged to be a reasonable tradeoff between computation time and connection to an analysis of empirical data. Only semileptonic and dileptonic decay channels for $t\bar{t}\gamma$ are simulated as a simulation of the all-hadronic decay channel would require a dedicated estimation of the event yield coming from fake leptons. Fake Leptons are defined as jets misidentified as reconstructed leptons (mostly electrons).

3.2 Sample generation

The framework MadGraph5_aMC@NLO [5], denoted by MG5, is used for the Monte Carlo (MC) simulations of the considered processes. It is a matrix element generator capable of the generation of events and the calculation of cross sections while also allowing the interfacing with tools for further simulation and analysis.
A central element of event generation with MG5 is the chosen model. It includes the particles,

[1] A photon is called genuine if it is emitted at truth level or radiated in the fragmentation process.

their interactions, and calculation schemes. While MG5 is commonly used to simulate signal and background events for dedicated SM analyses, it can also be used to explore BSM phenomenology through the generation of events with BSM models. The DIM6TOP_LO_UFO model [14], referred to as DIM6TOP model, is employed in this thesis. It consists of a number of dimension six operators modifying interactions involving top-quarks. Only the two coefficients C_{tB} and C_{tW} are varied in the study of $tq\gamma$. The DIM6TOP model only allows setting the coefficients C_{tZ} and C_{tW}. The coefficients are related by the formula

$$C_{tZ} = -\sin\theta_{\mathrm{W}} C_{tB} + \underbrace{\cos\theta_{\mathrm{W}} C_{tW}}_{=0}, \tag{3.1}$$

where the second term vanishes because only diagrams with at most one EFT vertex at a time are generated. This relation is based on Equations (2.1) and is used to switch between the two bases. Every result in this thesis is given in terms of C_{tB} by convention.

The photon is shown as being radiated from the top-quark in Figure 2.3. Since the top-photon coupling is to be investigated, the study of exactly this diagram is desired. However, it is possible that the photon is emitted by any electrically charged particle in this process, including initial and final state particles. These types of radiation are difficult to distinguish from the direct radiation from the top quark in the detector and are therefore simulated as well. It is not possible to generate EFT events without the σ_{SM} contribution while allowing the possibility of photon radiation during the decay of the top-quark and by its decay products. Non-zero Wilson coefficients include diagrams with EFT vertices in the sample, resulting in an EFT+SM sample which is called EFT enriched or referred to with the value of the Wilson coefficient used to generate it. No possibility has been found to separate events with EFT contributions from SM events at matrix element level. However, this would be beneficial in order to study the properties of the EFT impacts in a more direct way.

Requirements on kinematic variables, called cuts, are employed in the sample generation to define the phase space of the simulation. These are set to typical values for LHC experiments in order to define a phase space similar to that in the experiment. These cuts are also applied to decay products of the decay chain. Photons need to have a transverse momentum $p_{\mathrm{T}} > 10\,\mathrm{GeV}$ and the pseudorapidities $|\eta|$ of both photons and leptons are required to be less than 5 to account for the pseudorapidity coverages of the CMS and ATLAS detectors [2]. These are $|\eta| < 5$ [17] and $|\eta| < 4.9$ [18], respectively. The angular separation of photons to leptons and quarks needs to be larger than 0.2. This isolation criterion is typically applied in analyses of empirical data to suppress photons from hadron decays in jets. This requirement respects this on matrix element level. The angular separation between two reconstructed objects is defined as $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ with the differences in their pseudorapidities $\Delta\eta$ and azimuthal angles $\Delta\phi$, respectively. Samples are generated at LO accuracy at a centre-of-mass energy of $\sqrt{s} = 13\,\mathrm{TeV}$. The NNPDF3.0NLO parton distribution function (PDF) set within Les Houches Accord PDF (LHAPDF) [19] for 4FS is used to generate the $tq\gamma$ samples, while the NNPDF2.3LO for 5FS is used for the $t\bar{t}\gamma$ sample. The top-quark mass m_t is set to 172.5 GeV. The QED coupling constant α is defined through $\alpha^{-1} = 132.3489$ and the weak coupling constant g_w can be inferred from the value of $G_{\mathrm{F}} = 1.166\,37 \cdot 10^{-5}$ set in MG5, while α_{S} is set to $1.184 \cdot 10^{-1}$.

Each generated event is assigned an MC weight $w_{\mathrm{MC},i}$ in the simulation at truth level which accounts for the cross section σ of the process. An integrated luminosity of $\mathcal{L} = 140\,\mathrm{fb}^{-1}$ is

[2] A right-handed coordinate system with polar coordinates r and ϕ in the plane transverse to the collision axis with the z-axis being parallel to the collision axis is being used in this thesis. The pseudorapidity is defined by $\eta = -\ln\tan(\theta/2)$ with the polar angle θ.

assumed as the full run 2 data collected with the ATLAS Detector amounts to $139\,\mathrm{fb}^{-1}$ [20]. In general, the events are weighted according to the formula

$$w_i = \frac{w_{\mathrm{MC},i}}{\mathrm{N}} \mathcal{L}\,, \tag{3.2}$$

where N is the number of generated events. The MC weights may be reweighted by showering, hadronisation and detector simulation programs to account for the statistical nature of parton radiation, hadronisation processes, identification efficiencies, and misidentification rates.

3.3 Calculation of the dependence of the cross section on Wilson coefficients

A first consistency check of the event generation is the validation of Equation (2.4). That relation can only be validated through a fit to a number of cross sections with varied Wilson coefficients because the interference and pure EFT cross sections σ_{int} and σ_{BSM} are not known. The cross sections used for this check can be found in the Appendix in Tables A.2 and A.3. These were calculated with 2000 events and a random seed of 0 each.

The fit function $\sigma(C) = \sigma_{\mathrm{SM}} + \sigma_{\mathrm{int}} C + \sigma_{\mathrm{BSM}} C^2$ is taken to be a quadratic polynomial. A low relative uncertainty of the fit parameters is judged as a sign of consistency. The parameters are found through a non-linear least squares fit. The statistical uncertainties of the cross sections calculated by MG5 are treated as standard deviations and one fit is performed for each coefficient (C_{tB} and C_{tW}). The calculated parameters are summarized in Table 3.1, while the cross sections and the fitted parabola can be seen in Figure 3.1. The standard deviation of the fit parameters is denoted by δ.

The goodness of the fit is assessed through the χ^2 value divided by the degrees of freedom. The calculated values are $5.8 \cdot 10^{-5}$ for C_{tW} and $1.1 \cdot 10^{-4}$ for C_{tB}. A value close to 1 indicates that the fit is consistent, while a much smaller value usually hints at overfitting, meaning that there are too many fitted parameters and the fit matches the observed values too closely. However, this is not a problem in this case as this is a fit to a theoretical prediction and a nearly perfect agreement with it is expected for an accurate simulation framework. The relative uncertainty of σ_{int} for C_{tB} is comparatively high with 33.84%, although this is explainable by the deviations of the points from a parabola around the origin as σ_{int} is the coefficient for the linear term responsible for the behaviour of the function around zero.

Table 3.1: Fit parameters for the dependence of the cross section on the Wilson coefficients C_{tW} and C_{tW} in the parameterization given by Equation (2.4). All cross sections are given in pb.

	C_{tW}	C_{tB}
σ_{SM}	1.3890 ± 0.0031	1.3954 ± 0.0028
$\sigma_{\mathrm{SM}}/\delta_{\sigma_{\mathrm{SM}}}$	0.22%	0.20%
σ_{int}	0.2683 ± 0.0012	0.00107 ± 0.00033
$\sigma_{\mathrm{int}}/\delta_{\sigma_{\mathrm{int}}}$	0.45%	33.84%
σ_{BSM}	0.0569 ± 0.0005	0.00325 ± 0.00006
$\sigma_{\mathrm{BSM}}/\delta_{\sigma_{\mathrm{BSM}}}$	0.88%	1.85%

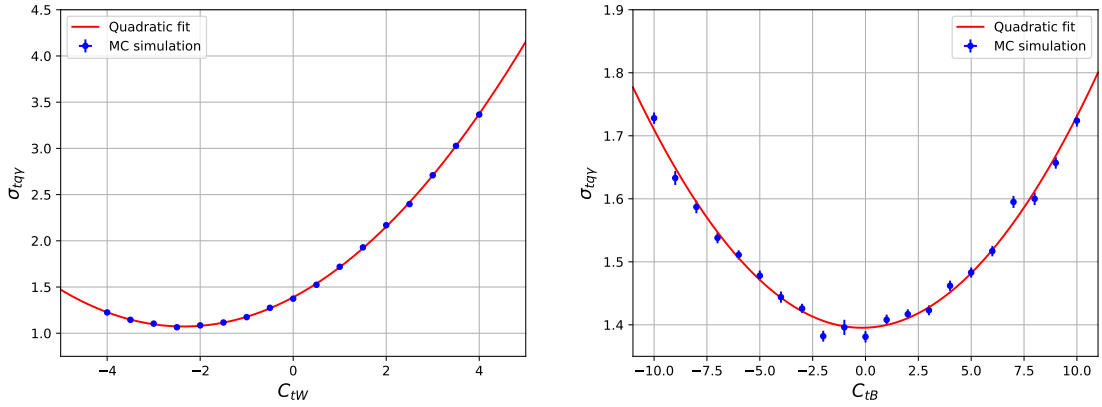


Figure 3.1: Dependence of the cross section on the Wilson coefficients C and the quadratic fits. The left plot shows C_{tW}, while the right one shows C_{tB}.

3.4 Investigation of control plots at truth level

The contributions to the event signature and cross section are expected to be different for the two considered operators. The cross section is assumed to be more sensitive to $\mathcal{O}_{tW}$ than to $\mathcal{O}_{tB}$. This is due to the Wtb vertex in the production of the top quark and the fact that the former operator affects both production vertices, leading to more possible insertions of it. A higher production rate of the top-quark because of a non-zero $\mathcal{O}_{tW}$ will influence the cross section more than an anomalous coupling to the photon. Therefore, it is expected that $\mathcal{O}_{tW}$ will influence the normalisations of the distributions of the object's kinematic variables more than their shapes, which are expected to be altered more by $\mathcal{O}_{tB}$. More specifically, it is reasonable to assume that the kinematic variables involving the photon will be influenced by these operators in particular. These could include the transverse momentum distribution of the photon $\gamma_{p_{\mathrm{T}}}$ and angular distances between the decay products of the top-quark and the photon.

A second consistency check entails the investigation of kinematic variable distributions at truth level based on these assumptions. These are termed control plots. For this purpose, six samples with 20000 events for both $tq\gamma$ and $t\bar{t}\gamma$ as well as for $C_{tB} = \pm 5$ and $C_{tW} = \pm 2$ are generated.

Four control plots for the transverse momentum of the photon $\gamma_{p_{\mathrm{T}}}$ and the invariant mass of the first top-quark candidate $m_{\ell\nu b}$ in the event are shown in Figure 3.2. A top-quark candidate is defined as a system of ℓ^+, b and ν from a top decay or ℓ^-, $\bar{b}$ and $\bar{\nu}$ from an antitop decay. The $\gamma_{p_{\mathrm{T}}}$ distributions are of particular importance to the whole analysis as it is expected that the EFT operators influence this variable because of the changed top-photon coupling. This is shown to be the case for all four variations of the considered operators and can be seen best in Figure 3.2b, where the distributions are normalised to unit area. The presumption that $\mathcal{O}_{tW}$ will primarily influence the cross section while $\mathcal{O}_{tB}$ will influence the shape of the kinematic variable distributions does not seem to apply for the distribution $\gamma_{p_{\mathrm{T}}}$. This conclusion is drawn because the shape differences in the upper part of the spectrum differ significantly for the two values of the same Wilson coefficients. The blue distribution ($C_{tW} = -2$) differs greatly from both SM backgrounds while the distribution of the $C_{tW} = -2$ sample (light orange) is close to both SM distributions.

The peak at about $m_t = 172.5\,\text{GeV}$ in the $m_{\ell\nu b}$ distribution is expected. It belongs to on-shell top quark candidates, where it is also possible that the top quark is first produced off-shell and then brought onto the mass shell by the radiation of the photon. There is also a washed-out, flatter peak at about 160 GeV which belongs to top-quark candidates that are already on-shell at production and then emit a photon. This case does not occur as often as photon radiation from other particles since there are a number of other electrically charged particles in the process that can emit the photon, so the 160 GeV peak is much flatter and wider. The peak is washed out because the photon can carry a spectrum of different transverse momenta.

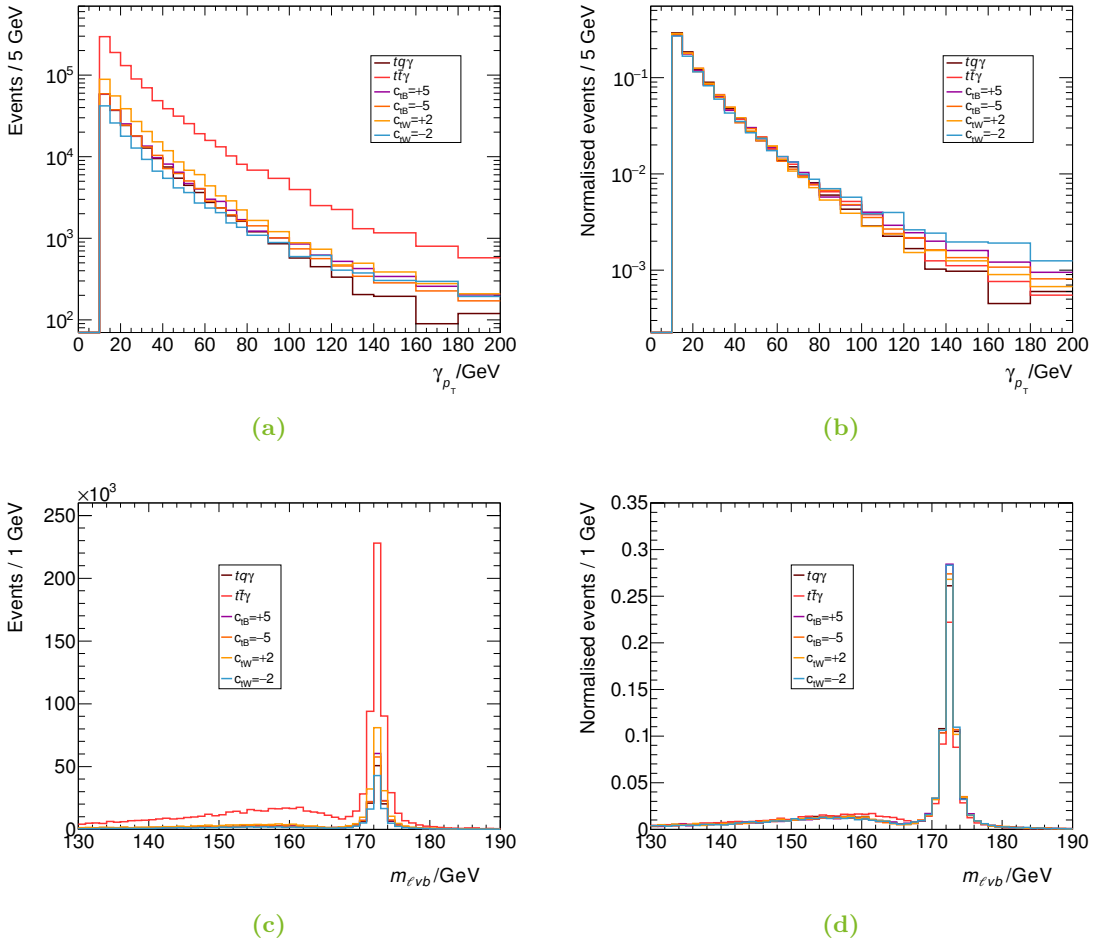


Figure 3.2: Unnormalised (a) and normalised (b) distributions of the transverse momentum p_{T} of the photon. The bin contents between 80 GeV and 140 GeV are averaged over 10 GeV and the bin contents between 140 GeV and 200 GeV are averaged over 20 GeV to smooth out statistical fluctuations. Unnormalised (c) and normalised (b) distributions of the invariant mass of the first top-quark candidate $m_{\ell\nu b}$. The histograms are normalised to unit area.

3.5 Showering, hadronisation and detector simulation software

The event generation with MG5 provides truth level information about the particles in the final state and their four-momenta. The partons may additionally emit gluons before and after the hard-scattering process. These are able to emit further radiation, leading to so-called

parton showers. The gluons and the final state quarks hadronise to form jets. Collisions between partons not contributing to the hard-scattering process also take place, which result in extra hadron production. This is termed underlying event. These phenomena are simulated with the program PYTHIA8 [21] interfaced with MG5. The result of this simulation is a collection of stable[3] particles.

These stable particles would pass through the experimental apparatus and generate a detector response in a collider experiment. Reconstruction and identification algorithms are employed to analyse the detector output and to translate the particles into physical objects with kinematic variables at the so-called detector level. These include photons, leptons, and jets. The algorithms have identification efficiencies and misidentification rates that influence the systematic uncertainties of the analysis results. In addition to that, a missing transverse momentum vector $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}} = -\sum_i \vec{p}_{\mathrm{T}}^{\,i}$ with magnitude $E_{\mathrm{T}}^{\mathrm{miss}}$ is constructed where the sum runs over the transverse momenta $\vec{p}_{\mathrm{T}}^{\,i}$ of all reconstructed objects, indexed by i. This formula is based on the fact that the sum of the transverse momenta of all particles should be equal to zero since the transverse plane is perpendicular to the collision axis. A considerable amount of missing transverse energy usually indicates the presence of neutrinos since they escape the detectors of typical collider experiments at the LHC. Neutrino reconstruction is not possible since the energies and particles in the initial state are not known with certainty in hadron collisions. A technique called b-tagging is of particular importance to analyses of processes with top-quarks. It is developed to identify jets originating from a bottom-quark (called "b-tagged") or other partons. The idea is to select events with top-quarks by requiring b-tags because top-quarks almost always decay into bottom-quarks. The b-tagging algorithms of ATLAS [18] and CMS are only applied to jets with $|\eta| < 2.5$.

The efficiencies of the algorithms and the reconstruction of particles have a considerable impact on the analysis and need to be simulated since the analysis strategy needs to account for the detector response. The program DELPHES [23] is employed for the detector simulation in this thesis. It comprises MC sampling of efficiencies and misidentification rates that may be dependent on kinematic variables of the particles. More sophisticated programs such as GEANT4 [24], which is often used in the analyses of ATLAS and CMS, allow for the explicit propagation of particles through the detector along with implementations of specific geometries of the experiment. However, these approaches, while improving the accuracy, are usually computationally expensive, which would be outside the scope of this thesis.

A parameterisation of the CMS detector response is used for the detector simulation in this thesis because its implementation in DELPHES is actively maintained by the CMS Collaboration. The jet clustering procedure in DELPHES is performed via the FastJet package [25] and the jet clustering algorithm used is the anti-k_t algorithm [26] with a radius parameter of $\Delta R = 0.5$. The b-tagging efficiency and the misidentification rate of charm-jets as b-tagged jets are only dependent on the transverse momenta of the jets and are shown in Fig. 3.3. The relations are

$$b_{\mathrm{eff}}(p_{\mathrm{T}}) = 0.85 \tanh(0.0025 \cdot p_{\mathrm{T}}/\mathrm{GeV}) \cdot \frac{25}{1 + 0.063 \cdot p_{\mathrm{T}}/\mathrm{GeV}}\,, \tag{3.3}$$

$$c_{\mathrm{misid}}(p_{\mathrm{T}}) = 0.25 \tanh(0.018 \cdot p_{\mathrm{T}}/\mathrm{GeV}) \cdot \frac{25}{1 + 0.013 \cdot p_{\mathrm{T}}/\mathrm{GeV}}\,. \tag{3.4}$$

[3] A particle is defined as stable if its mean lifetime τ exceeds $1\,\mathrm{cm} \cdot c \approx 3.34 \cdot 10^{-11}\,\mathrm{s}$ [22].

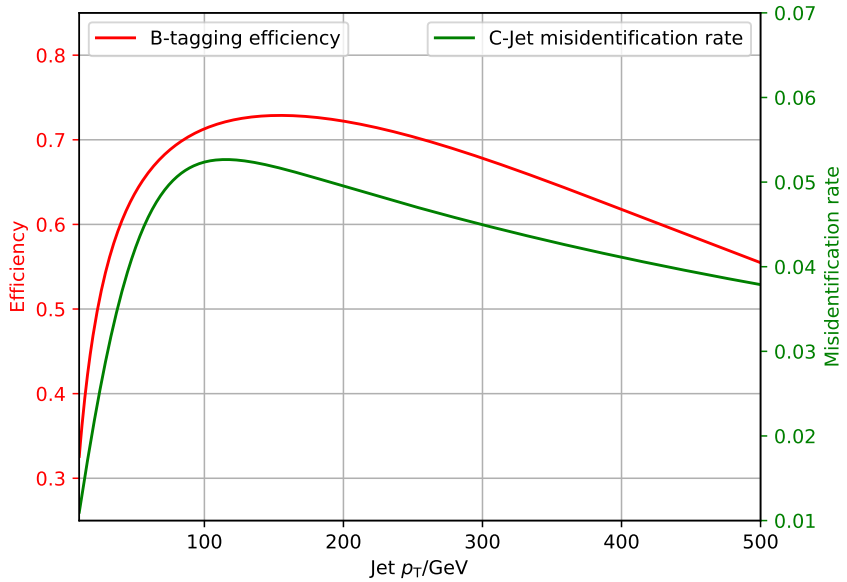


Figure 3.3: Plot of the b-tagging efficiency (left axis, red) and the misidentification rate of charm-jets as b-jets (right axis, green) as a function of the transverse momentum of the reconstructed jet in the employed detector simulation software DELPHES.

3.6 Discussion of the event selection

The event selection precedes the concrete analysis and comprises a set of requirements for the events that aim to select $tq\gamma$ events while rejecting $t\bar{t}\gamma$ events. The random seeds 0, 1, 2 and 3 are used to generate 50 000 events each for the six previous sample definitions. These are then merged to produce six samples of 200 000 events each. The cross sections of these samples are given in Table 3.2. The statistical uncertainties are neglected because they are of the order of 0.1%. The nominal SM cross section of $tq\gamma$ is not compatible with the fit parameters σ_{SM}, shown in Table 3.1, within one σ. This inconsistency is not investigated further since only the cross sections in Table 3.2 are used for the weighting of the events for the MVA analysi, although further studies should be conducted to resolve this disagreement.

Table 3.2: Nominal cross sections calculated with MG5. All cross sections are given in picobarn.

SM $tq\gamma$	$t\bar{t}\gamma$	$C_{tB} = 5$	$C_{tB} = -5$	$C_{tW} = 2$	$C_{tW} = -2$
1.424	7.489	1.52	1.504	2.209	1.107

In the following, two event selections with the same baseline selection are discussed. They are based on the event signature of $tq\gamma$ at detector level and a physical motivation is given for each selection requirement.

Exactly one photon with $p_{\text{T}} > 15\,\text{GeV}$ needs to be present in the event. This cut is set to a lower value than in the analysis of the CMS Collaboration (25 GeV [4]) in order to retain a higher number of events for the MVA studies. Further investigation would be required to find the value which would maximise the signal purity after the event selection. Exactly one lepton

with $p_{\mathrm{T}} > 27\,\mathrm{GeV}$ is required to select events with leptonic W boson decay. This value is typical in studies of single-top-quark production either as a cut or as a trigger threshold [4, 27]. Missing transverse energy of 30 GeV or more is required. This cut utilises the presence of neutrinos in the signal process and aims to cut away events without neutrinos, e.g. W bosons decaying hadronically in the signal and background processes or background processes with a Z boson decaying to charged leptons. Lastly, exactly one b-tagged jet with $p_{\mathrm{T}} > 25\,\mathrm{GeV}$ is required, targetting the b-jet from the top-quark decay as the jet from the hadronisation of the initial $\bar{b}$ quark seen in Figure 2.3 is mostly out of acceptance region for b-tagging [27]. This requirement cuts away a $t\bar{t}\gamma$ events with two b-tags, although a considerable amount of $t\bar{t}\gamma$ events remain after this cut as the b-tagging algorithms only have a limited identification efficiency.

The selections differ in the treatment of the characteristic forward jet of the $tq\gamma$ process formed by the hadronisation of q_{f} in Figure 2.3. From now on, a jet is defined as a forward jet if it is reconstructed with $p_{\mathrm{T}} > 25\,\mathrm{GeV}$ and $|\eta| > 2.5$. This high pseudorapidity results from the scattering of the initial up-type quark against the heavier bottom-quark. With this definition the forward jet is not considered for b-tagging algorithms as $|\eta| > 2.5$ is outside the b-tagging range for the ATLAS and CMS experiments. The first selection looks at the jet with the highest transverse momentum, termed the leading jet, that is not b-tagged and vetoes the event in case this jet is not a forward jet. Otherwise, the other requirements explained above are checked. This procedure is therefore termed leading untagged selection. The second selection demands exactly one forward jet and requires exactly two jets being present in the event. This selection aims to suppress $t\bar{t}\gamma$ events with at least one hadronically decaying W boson and is expected to suppress multi-jet background in a more complete analysis.

Both selections are compared based on the expected significances $S_i/\sqrt{B}$ they achieve for each of the four signal samples i, where S_i denotes the sum of selected signal weights of that sample and B the sum of selected background weights. Another approach would be to compare the expected significances for EFT enriched and SM $tq\gamma$ as signal and only $t\bar{t}\gamma$ as background because it is assumed to be challenging to select EFT enriched $tq\gamma$ events with a linear event selection. This approach is not chosen, however, to keep the definition of signal and background consistent throughout this thesis. The results of the two event selections are given in Table A.4 in the Appendix. The selection requiring exactly one forward jet vetoes considerably more events than the leading untagged selection. The latter selection achieves considerably higher expected significances than the former. These higher significances warrant the use of the leading untagged selection to select the events before further analysis in Chapter 4. It would be possible to implement further cuts to suppress other neglected background processes. These are not implemented to achieve as high statistics as possible to reduce the statistical uncertainties of the results of the MVA studies. For example, one of these other requirements would be to veto events with $70\,\mathrm{GeV} < m_{\ell\gamma} < 110\,\mathrm{GeV}$ in order to suppress fake photon background from Z + jets events.

4 Analysis of the sensitivity of kinematic variables to Wilson coefficients

The goal of the MVA studies is to achieve maximum discrimination between the EFT enriched and the SM background samples. The two methods used in this thesis to accomplish this are presented along with their classification accuracies. The method that performs better is used for further analysis and the bin-wise expected significances for the output of the classifiers and for the most discriminating variables are compared. Finally, the limitations of this thesis, resulting from several necessary assumptions and approximations, are discussed.

4.1 Introducing the employed multivariate analysis methods

The application of machine learning and MVA in HEP leads to more powerful ways of separating signal and background by combining many variables and their nonlinear correlations than with simple linear cuts on kinematic variables. The concrete goal in this analysis is to train a classifier with labelled data so that it can distinguish between signal (EFT enriched $tq\gamma$) and SM background, which are called classes. This is assessed by the accuracy, defined as the fraction of correct predictions and the total number of events. An accuracy higher than 50% would already be a success because no studies have been performed on the discrimination of EFT enriched samples and SM background yet and this would prove the possibility to discriminate between the two classes in principle. The classification is performed on the basis of n features. The classifier translates these features into an output, e.g. between -1 and 1. A discrimination threshold, also termed cut, can be applied to the output, meaning that all events that are assigned an output value greater than a certain predefined cut value are classified as signal, whereas all other events are classified as background.

The idea of linear discriminant analysis (LDA) is to create a linear combination of features on which a cut allows a separation of the two classes. A so-called decision surface is defined, which is an n-dimensional hyperplane. A projection of an event with its features on this plane results in the output of this classifier. This simple method represents a benchmark classifier that can be used if more sophisticated methods do not perform significantly better. The BDT method is based on the central concept of boosting and makes use of many estimators, also called weak learners. The idea of boosting is that individual learners are trained sequentially so that the individual learners can benefit from the weaknesses of others during the training step. The individual training events receive larger weights for training if the previous learners disagree about the classification or if they predict the wrong label. The individual weak learners are called weak because they perform only slightly better than random guess. Only their combination can result in a skillful classifier. It is possible to access the so-called feature importances of the BDT, which sum to one and rank the given features in terms of their discrimination power. This is used for the physical interpretation as these importances might represent the EFT induced differences in the kinematic variable distributions compared to the SM predictions.

The BDT and the LDA are implemented within the Python library scikit-learn (version 0.18.1) [28], referred to as SKLEARN. The algorithm AdaBoost is used to construct the BDT. This general introduction should suffice since the concrete mechanisms of the methods is not important for this thesis. More information can be found this publication about AdaBoost [29].

4.2 Comparing boosted decision trees and linear discrimination analysis to classify EFT enriched data

Both models are evaluated in three scenarios due to the different possible combinations of background processes: Firstly, the models are evaluated against both backgrounds, then against $tq\gamma$ and lastly against $t\bar{t}\gamma$ in order to estimate the discrimination power of the different features and assess the skill of the models in the classification against SM background. The data is split into a training and a test dataset with an 80 to 20 ratio.

The splitting is performed in a stratified fashion, meaning that the class percentages are preserved for the training and test sets. The model is trained with the training set and the final evaluation score is calculated with the test set to detect the phenomenon of overfitting, where the model focuses on statistical fluctuations in the training data and fits it too closely. A significantly higher accuracy on the training set than on the test set is a sign of overfitting.

The input features are realised as kinematic variables in this thesis. Only the values of the leading physical objects are considered. The adopted variables are the transverse momenta and the pseudorapidities of the jet, b-tagged jet, forward jet, photon, and lepton; the missing transverse energy; the sum of the transverse momenta of the photon, lepton, b-tagged jet, forward jet and the missing transverse energy (commonly referred to as H_{T}); the invariant mass of the lepton, photon and jet in all 2-Permutations and the angular distances of photon to lepton, b-tagged jet and forward jet, respectively. A number of chosen variables involve the leading photon as the EFT operators are expected to influence these variables the most. This assumption has been validated qualitatively with the investigation of the photon transverse momentum distributions in Chapter 3.4. The other variables are chosen as they involve other process-specific physical objects, namely the jets as well as the lepton and the neutrino (as missing transverse energy) from leptonic W boson decay.

While LDA is not dependent on any parameters, a BDT has two parameters: The number of weak learners, termed n_estimators by SKLEARN, and the learning rate, which describes how much the model changes after each iteration of training. These parameters are called hyperparameters and are tuned before the training of the final model for each scenario. This tuning comprises a grid search over possible values for the two parameters through a technique called k-fold cross validation (CV). The data is split into k stratified parts and k steps of training and evaluation are performed. In the i-th step the i-th part is used as the test set and the remaining $k-1$ data sets are combined to form the training set. The CV score is the average of the accuracy of all steps. The grid search is performed with 5-fold CV on the training set over the following values:

$$\begin{aligned} \text{n\_estimators} &: 50,\ 80,\ 90,\ 100,\ 110,\ 120,\ 150, \\ \text{learning rate} &: 0.01,\ 0.05,\ 0.075,\ 0.1,\ 0.125,\ 0.15,\ 0.3,\ 0.5,\ 1. \end{aligned}$$

preliminary studies with the samples at hand have shown that these values enclose the most promising hyperparameters. The pair of parameters offering the highest CV scores is chosen for each final model. The optimal hyperparameters found through this procedure are given in Table A.5 in the Appendix.

The weights for the training are adjusted in the following way only for the training of the BDTs: The sum of weights for the signal events S and for the background events B are calculated. All signal weights are multiplied by B/S, which has the effect that the new sums of weights are equal. This procedure is adopted because, otherwise, the BDT would concentrate more on the correct classification of $t\bar{t}\gamma$ as background due to the higher cross section since the weak learners concentrate on events with larger weights at the start of the training. This

is not desired as the analysis concentrates on the characteristics of the EFT enriched samples. Preparatory studies using the samples at hand have validated that the average accuracy is lower when this reweighting procedure is not used.

SKLEARN's LDA implementation fits a Gaussian density to each class and training with weighted events is not possible in that case. Accuracies are still calculated with weighted events to emulate empirical data. The fit is performed with the shrinkage tool of the SKLEARN implementation. This means that the off-diagonal elements of the sample covariance matrix of the input variables are shrunk to improve its estimation. Usually, this is only necessary when the number of samples is small compared to the number of features but the tool is also employed in this analysis because it improves the performance of the LDAs. The optimal shrinkage parameter is determined automatically by the tool.

The accuracies of the tuned models are listed in Table 4.1. The nominal training accuracies are calculated as the sample mean of the accuracy scores in 5-fold CV. 95% confidence intervals $[\overline{x} - \frac{s_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}, \overline{x} + \frac{s_n}{\sqrt{n}} t_{n-1,1-\frac{\alpha}{2}}]$ are constructed for each training accuracy, where $\overline{x}$ is the arithmetic mean of the $n = 5$ scores, s_n is the corrected sample standard deviation and $t_{n-1,1-\frac{\alpha}{2}}$ is the $(1-\alpha)$-quantile of Student's t_{n-1} distribution with the significance level α. This assumes that the training accuracies follow a normal distribution.

Table 4.1: Accuracies on the training and test sets for the final models. The leftmost column describes the background composition used to train (T) and evaluate (E) the model. The accuracies on the test set consisting of both backgrounds with the model trained using $t\bar{t}\gamma$ are presented in the last row (E: both). The accuracies on the training dataset are obtained through a 5-fold CV, they denote 95% confidence intervals. All accuracies are given in percent.

			$C_{tB} = +5$	$C_{tB} = -5$	$C_{tW} = -2$	$C_{tW} = -2$
T/E: both	BDT	Training	62.5 ± 3.5	62.5 ± 1.7	58.8 ± 2.2	69.6 ± 2.4
		Test	60.7	63.7	59.3	67.1
	LDA	Training	71.4 ± 2.5	73.0 ± 2.5	64.6 ± 2.2	69.3 ± 1.8
		Test	72.1	72.1	63.4	70.8
T/E: $tq\gamma$	BDT	Training	51.6 ± 1.8	52.7 ± 1.4	39.9 ± 1.1	52.7 ± 1.9
		Test	54.5	53.3	42.5	53.5
	LDA	Training	53.8 ± 2.5	54.2 ± 1.9	50.7 ± 2.7	53.9 ± 0.8
		Test	52.7	56.1	49.5	56.8
T/E: $t\bar{t}\gamma$	BDT	Training	63.5 ± 2.8	64.1 ± 4.1	63.2 ± 4.7	62.6 ± 3.9
		Test	66.3	64.0	62.8	62.4
	LDA	Training	31.9 ± 1.6	34.3 ± 1.6	46.5 ± 1.3	31.8 ± 0.8
		Test	33.6	32.9	46.9	32.2
T: $t\bar{t}\gamma$ E: both	BDT	Test	61.3	64.4	61.7	62.2
	LDA	Test	45.3	28.9	43.5	46.0

The performance of the BDT with classification against both backgrounds is above 50% for all coefficients. The nominal LDA accuracies are significantly higher than the ones achieved by the BDTs. This is unexpected because there is considerable overlap between signal and background since both contain SM $tq\gamma$ contributions. A reason for this behaviour has not been found and further studies on this phenomenon are needed. The performance of the BDT classifying the $C_{tW} = -2$ is significantly better than the performances of the other three

models. This may result from more pronounced shape differences due to the effects of the EFT operators. The control plot 3.2b does indicate this fact for the γ_{p_T} distribution, which is the feature with the highest feature importance for this model by far. The accuracy on the test set is not contained in the 95% confidence interval for the training set accuracy for this sample, indicating overfitting.

The accuracies for the classification of the EFT enriched samples and $tq\gamma$ are only slightly above 50% and 50% is contained in the confidence interval for $C_{tB} = +5$. The accuracies on the test set are above the nominal training accuracies, indicating a small, but consistent discrimination power. This is expected because the separation of EFT enriched samples and SM samples is a challenging task since the EFT enriched samples do contain SM contributions as well. The BDTs and the LDAs achieve similar accuracies. The sample with $C_{tW} = +2$ is an exception, the BDT does not learn to classify the events and performs significantly worse than random guess and the LDA. A Wilson coefficient of $C_{tW} = +2$ seems to affect the normalisations of the distributions more than the shapes, which is seen in the control plot 3.2b of the γ_{p_T} distribution. There, the light orange curve belonging to $C_{tW} = +2$ is more similar to the SM $tq\gamma$ curve (brown) than the curves belonging to the other EFT enriched samples. It is not possible for the BDT, and to a smaller extent the LDA, to learn these normalisation differences, as mainly the weights are changed. This model is not considered for the analysis of feature importances due to its apparent lack of skill.

The most discriminating feature for the $C_{tB} = \pm 5$ samples against the SM $tq\gamma$ background is the transverse momentum of the photon with feature importances of 0.191 and 0.125, while the pseudorapidity of the forward jet (0.109) and the invariant mass of lepton and photon (0.109) also have significant discrimination power for the $C_{tB} = +5$ sample. The second and third most discriminating features for $C_{tB} = -5$ are the transverse momentum of the lepton with a feature importance of 0.117 and the invariant mass of the jet and the lepton (0.1). There are only three discriminating features for the $C_{tW} = -2$ sample: The transverse momentum of the photon, the lepton and the invariant mass of the lepton and the photon with feature importances of 0.422, 0.389 and 0.189, respectively. It is not advisable to interpret these importances directly as effects of the EFT operators since all feature importances are compatible with zero due to the low statistics. The feature importances should only be interpreted as indications of possible shape differences due to the EFT operators. Further studies with more MC statistics would lead to more precise estimations of these differences.

The BDTs perform well when trained and evaluated using only $t\bar{t}\gamma$ as background. The combination of multiple kinematic variables leads to accuracies above 60% for all four EFT enriched samples. It has been validated that this is not due to the EFT contributions in the signal as a binary BDT classification of SM $tq\gamma$ and SM $t\bar{t}\gamma$ yields similar results that are compatible within the uncertainties. This is an important result because it shows that the application of MVA for the classification of signal and SM background does not depend on possible EFT contributions in the signal. The LDAs fail to classify EFT enriched $tq\gamma$ against $t\bar{t}\gamma$. It is likely that the two processes are not linearly separable in the feature space. This argument is supported by the fact that the five to seven most discriminant features of the BDTs have similar feature importances, indicating that the shape differences are spread evenly in several dimensions.

It is also investigated if it is possible to achieve a satisfactory accuracy when a model is trained with $t\bar{t}\gamma$ background and evaluated against both backgrounds. The BDTs manage to achieve accuracies over 60% with this procedure. It seems that the classifiers learn the effects of the EFT operators during training against another process and can apply this to data that are composed in a different way. The LDAs perform significantly worse than random

guess which is not surprising as a linear separation through a n-dimensional hyperplane is not adaptive to a redefinition of the signal and background classes.

All in all, it is shown that the separation of EFT enriched $tq\gamma$ from all three combinations of the two underground processes is possible in principle. The LDA is superior to the BDT in the discrimination against both processes. When separating against $tq\gamma$, both methods perform similarly well and only the BDTs allow the separation of the signal and $t\bar{t}\gamma$. It is expected that the performance of the LDA would decrease considerably with the inclusion of more background processes as it becomes more and more difficult to separate EFT enriched $tq\gamma$ and SM background linearly. Only the sample with $C_{tW} = -2$ shows signs of overfitting when trained against both backgrounds. Other than that, no clear signs of overfitting can be inferred from the accuracies since all other test accuracies are below the lower interval limit of the confidence intervals. Based on this, further studies are performed using BDTs as classifiers.

The sample with $C_{tB} = 5$ serves as a representative sample for the subsequent analysis as no outliers are present in the accuracies achieved for this sample. Additional material for the other three samples can be found in the Appendix A.2 and the results presented there are included in the Summary.

Another possibility to investigate the performance of a model is the so-called receiver operating characteristic curve (ROC curve), which highlights the skill of the model with respect to possible cuts on the output, together with the area under the ROC curve, referred to as AUC. The true positive rate (TPR), which is also known as sensitivity and which is defined as the number of correctly classified signal events divided by the total number of events N, is plotted against the false positive rate (FPR), the ratio of background events falsely classified as signal events to N. A random guess model would have a bisecting line characterised by TPR = FPR with an AUC of 0.5. Therefore, an AUC considerably higher than 0.5 is an indication of good performance. The ROC curves for the three possible background combinations are shown in Figure 4.1, a-c. These are calculated using the whole dataset and reflect the accuracies consistently since the AUCs can be associated with them.

The classification of EFT enriched $tq\gamma$ and both backgrounds occurs with high variance as the 1σ-band in Figure 4.1a touches the random guess curve. The two backgrounds are artificially defined as a single class even though they are physically different processes. Because of that, the background events have a higher variance in the feature space than the physically uniform backgrounds for the other scenarios and different splits might result in a considerably uneven makeup of the background. The stratified sampling does not address this problem as the stratification is only performed according to the class definitions.

The model trained with SM $tq\gamma$ background classifies with lower variance as can be seen in Figure 4.1b. It is expected that the performance of the model is only slightly better than random guessing as the signal is only EFT enriched and not pure. This is confirmed by the ROC curve and the achieved accuracies. In contrast, the ROC curve for training against $t\bar{t}\gamma$ with an AUC=70 ± 2 is clearly better than random guess and the classification is performed with low variance. This supports the prior assumption of a well-performing model due to its high accuracy.

The subsequent analysis is carried out on the BDT with training and evaluation using both backgrounds because this is a realistic setting for an analysis with empirical data and the BDT performs reasonably well in that situation. It is common to use both training and test data for the evaluation of the final model, as it is usually not possible to afford to lose more than half of the MC simulation data in the form of training data for the evaluation, as the statistical uncertainties would then increase significantly. The decision to include the training

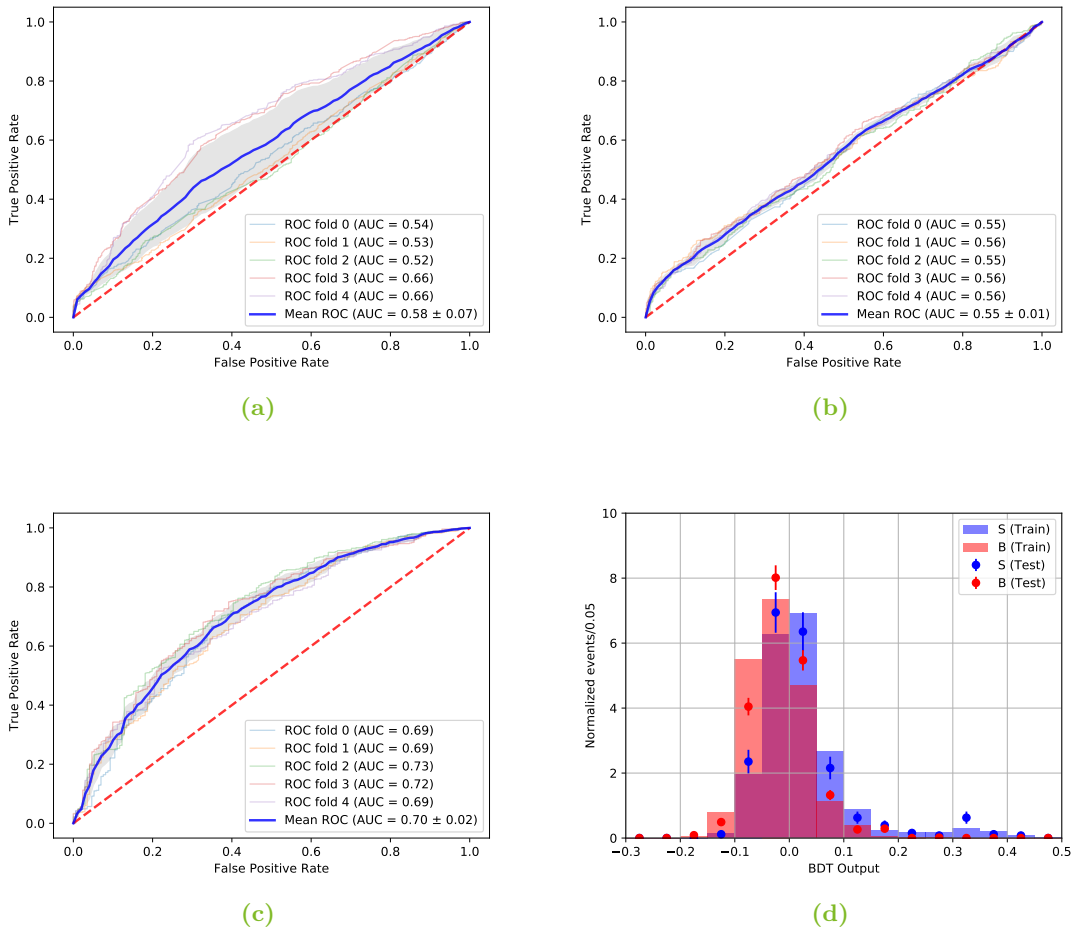


Figure 4.1: ROC curves for the BDTs with training and evaluation using both backgrounds (a), using $tq\gamma$ (b) and using $t\bar{t}\gamma$ (c), with the $C_{tB} = +5$ sample as the signal. The dashed red line represents the ROC of a random guess classifier while the shaded grey region is the 1σ-band of the mean ROC curve calculated with the corrected sample standard deviation from the individual CV folds. The BDT output for training and evaluation using both backgrounds is shown separately for training and test data in (d). The signal is shown in blue, while the background events are shown in red. The histograms are normalised to unit area and the error bars are Poisson uncertainties.

data in the final evaluation must be based on thorough previous checks for overfitting because it is a basic principle of the MVA to evaluate the tuned and trained model on previously unseen data.

A first check based on the comparison of accuracies on the training and test sets has already been performed for all models. Another possible check may be the examination of the output distribution for training and test data by eye, shown in Figure 4.1d. Slight overfitting is estimated to be present with this technique because the background events in the test data (red dots) are slightly above the background distribution in the training data (red bars) to the right of the typical discrimination threshold 0, indicating that more background events are classified as signal for the test data than for the training data, which is a sign of worse performance. This is significant as the deviations are not compatible within one standard deviation. It is also possible to test the similarity of the output distributions of the training data and the test data for signal and background, respectively, using a Kolmogorov-Smirnov (KS) test. The null hypothesis H_0 is that the histograms of the BDT output for the training data and test data come from the same probability distribution. This is essentially a check whether the two histograms are similar enough that overfitting may be presumed to be absent. H_0 cannot be rejected at a significance level of 5%[1] for both training and test sets of all samples [2]. However, the KS test has low statistical power and is therefore considered a conservative test. Rejection of H_0 would be a strong sign for overfitting but the result that H_0 cannot be rejected may not be interpreted as an absence of overfitting.

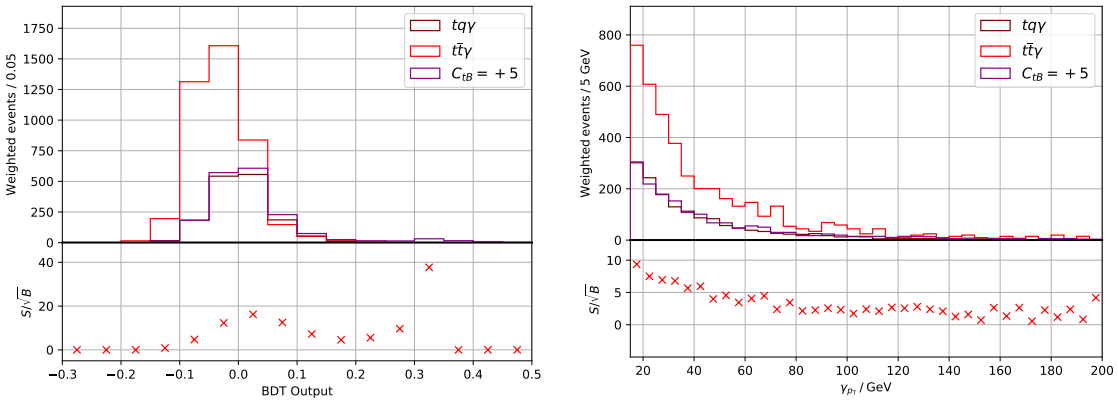
Overall no definite signs of overfitting are present for the $C_{tB} = 5$ sample and the whole data are used for the final step of the analysis. The BDT output is compared to the distribution of the most discriminating feature, which is the transverse momentum of the photon with a feature importance of 0.19 with the following feature importances being on the order of 0.1, with regard to the bin-wise expected significances $S_i/\sqrt{B_i}$, where S_i and B_i are the weighted signal and background events in the i-th bin, respectively. These expected significances depend on the choice of binning, which is selected here by eye. No particular emphasis is placed on the selection of the optimal binning since the discriminant of the BDT and the individual kinematic variables are only roughly compared according to the expected significances. This study is important in the preparation for an analysis with empirical data, as it can be used to assess the quality of the chosen MVA method.

The two histograms are shown in Figure 4.2. It can be seen that the output distribution in 4.2a is mostly concentrated around zero with a small tail of mostly signal events to the right of zero, which is also apparent in Figure 4.1d. These signal events could be interpreted as events with EFT contribution, which would mean that the BDT is able to separate pure EFT $tq\gamma$ and SM $tq\gamma$ to some degree even in classification with multiple backgrounds. It is possible to achieve a bin-wise expected significance as high as 37.7 with approximately 31.2 signal events and 0.7 background events when considering the bin $[0.3, 0.35]$ in the long tail of the BDT output. Therefore, these simplified studies suggest a discrimination threshold of 0.3 for an analysis to extract a $C_{tB} = +5$ signal yield. It is assumed that this expected significance will be considerably lower with the inclusion of the background processes neglected in this thesis. Nevertheless, the BDT can adapt a little by taking into account the shape differences of the distributions of the other background processes during training and learning from them.

This kind of learning process is not possible when only considering single kinematic variables. The achievable significances for the γ_{p_T} spectrum in Figure 4.2b range from 1.7 to 4.5 in the

[1] A significance level of 5% is commonly used and means that if H_0 is correct, the probability that it will be falsely rejected must not be more than 5%.

[2] The plots of the BDT output distribution for the other samples can be found in Chapter A.2 of the Appendix.



(a) (b)

Figure 4.2: BDT output distribution for the $C_{tB} = 5$ sample with training and evaluation using both background processes (a) and the transverse momentum of the leading photon for the three samples (b). The lower plots show the expected significance per bin, which is set to zero if no background events are contained in the bin.

middle part of the spectrum from 60 to 140 GeV. Only this region is considered since the comparatively higher expected significances in the spectrum's lower part are only due to the normalisation in the bins and the upper part is prone to fluctuations that do not represent a physical excess of signal events over background events. More specifically, the γ_{p_T} spectra of the neglected background processes can be assumed to be similarly soft and uniformly decreasing which would lead to pronounced and approximately homogenous contamination of the combined spectrum with background processes.

4.3 Discussion of approximations, assumptions and resulting limitations

A number of approximations are made in this thesis which need to be discussed in order to assess the quality and significance of the presented results. To begin with, most of the approximations stem from the limited scope of the thesis. Only two background processes, $tq\gamma$ and $t\bar{t}\gamma$, are simulated. The restrictive event selection cuts are necessary to prepare the data for further analysis. However, they cut away a high number of events so that the MVA studies are hampered by the low statistics. Samples with a higher number of events would reduce the statistical uncertainties on the accuracies and the ROC curve and improve the meaningfulness of these studies overall.

The estimation of sources of uncertainties is also largely neglected. A dedicated estimation of systematic uncertainties concerning PDF variations, different values for coupling constants, particle masses and scales would need to be performed in order to arrive at results with credible uncertainties. An analysis with empirical data using these kinds of EFT studies as a basis for a model-independent interpretation would also need to connect experimental uncertainties, such as uncertainties concerning the object reconstruction, with the MC simulations.

The whole analysis is significantly hampered by the fact that there is no way to generate a consistent sample without the SM contributions in MG5 with the DIM6TOP model. If such a sample could be obtained, the interpretation of the excess signal events at high BDT outputs as events with an EFT contribution could be tested. These validations are vital in the preparation of a full analysis to decide on a consistent analysis strategy.

5 Summary

The process of t-channel single-top-quark production in association with a photon has been investigated in MC simulations with an EFT approach. EFT enriched samples with the operators $\mathcal{O}_{tW}$ and $\mathcal{O}_{tB}$ were generated and treated as signal, while the SM background consists of $tq\gamma$ and $t\bar{t}\gamma$. The consistency of the event generation has been validated through control plots at truth level and through a parameterisation of the Wilson coefficient dependence of the cross section.

Multivariate analysis methods have been applied to classify the EFT enriched samples and SM backgrounds. Linear discriminant analysis (LDA) performs better than boosted decision trees (BDT) when discriminating against both backgrounds. The two methods yield comparable results when classifying EFT enriched $tq\gamma$ and SM $tq\gamma$, which corresponds to an identification of the effects of the EFT operators. The most discriminating variable is the transverse momentum of the leading photon for all samples, assessed by feature importances of the BDTs. The classification of the signal and $t\bar{t}\gamma$ background through BDTs is successful with accuracies above 60% and it was shown that LDA fails to separate these two classes.

The highest achievable expected significances $S/\sqrt{B}$ per bin using the histogrammed BDT outputs have been demonstrated to lie between 17.6 for the sample with a Wilson coefficient of $C_{tB} = -5$ and 37.7 for the $C_{tB} = +5$ sample. These are considerably higher than the expected significances per bin for the distributions of the most discrimating kinematic variable per sample, which range from 4.5 to 11.4. The latter are expected to be significantly lower when including other background processes. It is expected that the BDTs are more robust with respect to the inclusion of further background processes, as they are presumed to be able to learn from the shape differences of the distributions of the other processes.

All in all, the goal of this thesis has been achieved because it was shown that the MVA based classification of EFT enriched $tq\gamma$ and SM backgrounds is possible using multivariate analysis methods. Sensitivity of the classification to EFT contributions is presumed. Substantial approximations have been made due to the limited scope of this thesis. For example, important background processes like $t\bar{t}$ with fake photons were omitted, the samples used for the MVA methods had low statistics and the estimation of uncertainties was largely neglected. These approximations lead to the fact that the concrete numerical results are not very precise and should only be interpreted as rough estimates. It is suggested that further studies be based on the analysis strategy presented in this thesis and focus on a gradual reduction of the approximations made here in order to obtain more meaningful results.

A Appendix

A.1 Additional Tables

Table A.1: Event yields after the event selection in data and for each SM contribution. The expected yields are presented with their statistical and systematic uncertainties combined. This Table is literally quoted from [4] (Table 1). The signal is defined as single-top-quark production.

Process	Event yield
$t\bar{t} + \gamma$	1401 ± 131
$W\gamma +$ jets	329 ± 78
$Z\gamma + jets$	323 ± 55
Misidentified photon	374 ± 74
$t\gamma$ (s- and tW-channel)	57 ± 8
$VV\gamma$	8 ± 3
Total background	2401 ± 178
Expected signal	154 ± 24
Total SM prediction	2555 ± 180
Data	2535

Table A.2: Cross sections calculated by MG5 and their dependence on the Wilson Coefficient C_{tW}. All cross sections are given in pb.

C_{tW}	σ
-4	1.225 ± 0.0072
-3.5	1.145 ± 0.0065
-3	1.104 ± 0.0055
-2.5	1.065 ± 0.0065
-2	1.085 ± 0.0059
-1.5	1.116 ± 0.0064
-1	1.174 ± 0.0068
-0.5	1.275 ± 0.0081
0	1.374 ± 0.0089
0.5	1.524 ± 0.013
1	1.719 ± 0.010
1.5	1.929 ± 0.010
2	2.169 ± 0.011
2.5	2.397 ± 0.013
3	2.709 ± 0.015
3.5	3.027 ± 0.018
4	3.366 ± 0.021

Table A.3: Cross sections calculated by MG5 and their dependence on the Wilson Coefficient C_{tB}. All cross sections are given in pb.

C_{tB}	σ
−10	1.728 ± 0.009
−9	1.633 ± 0.011
−8	1.587 ± 0.01
−7	1.538 ± 0.0086
−6	1.511 ± 0.0073
−5	1.478 ± 0.0081
−4	1.444 ± 0.0089
−3	1.426 ± 0.0074
−2	1.382 ± 0.0086
−1	1.396 ± 0.012
0	1.381 ± 0.009
1	1.408 ± 0.0083
2	1.417 ± 0.0076
3	1.423 ± 0.0078
4	1.462 ± 0.0081
5	1.483 ± 0.0081
6	1.517 ± 0.0083
7	1.595 ± 0.0095
8	1.6 ± 0.01
9	1.657 ± 0.0092
10	1.724 ± 0.0095

Table A.4: Event yields, weighted event yields and expected significances $S/\sqrt{B}$ after applying the leading untagged selection (LUS) and the selection requiring exactly one forward jet and exactly two jets (1FJ). The accepted (acc.) yields are rounded to integers and the expected significances are rounded to one decimal place.

		$tq\gamma$	$t\bar{t}\gamma$	$C_{tB} = 5$	$C_{tB} = -5$	$C_{tW} = 2$	$C_{tW} = -2$
LUS	acc. (weighted)	1549	4195	1786	1723	3094	1843
	acc. (events)	2218	856	2551	2462	2213	2632
	$S_i/\sqrt{B}$			23.6	22.7	40.8	24.3
1FJ	acc. (weighted)	446	715	500	502	850	485
	acc. (events)	640	146	715	718	606	692
	$S_i/\sqrt{B}$			14.7	14.7	24.9	14.2

Table A.5: Hyperparameters found through grid search and 5-fold cross validation on the training sets.

		$C_{tB} = 5$	$C_{tB} = -5$	$C_{tW} = 2$	$C_{tW} = -2$
Against both backgrounds	n_estimators	110	80	80	90
	learning rate	0.05	0.125	0.125	0.01
Against $tq\gamma$	n_estimators	110	120	80	90
	learning rate	0.1	0.125	0.01	0.15
Against $t\bar{t}\gamma$	n_estimators	110	120	100	120
	learning rate	0.15	0.15	0.15	0.3

A.2 Additional material for the multivariate analysis

The ROC curves, the output distribution and expected significances for the three remaining samples with $C_{tB} = -5$ and $C_{tW} = \pm 2$ are presented and briefly discussed in this chapter. The output distribution and the expected significances are calculated with training and evaluation using both SM $tq\gamma$ and $t\bar{t}\gamma$ as background processes.

The ROC curves in Figure A.1, a-c, show that the classification of $tq\gamma$ with $C_{tB} = -5$ and SM background is possible for all three background combinations. The evaluation against both backgrounds occurs with higher variance and the 1σ-band penetrates the random guess curve slightly. The classification of EFT enriched $tq\gamma$ and $t\bar{t}\gamma$ is the most successful with a nominal AUC of 0.70. The output distribution in Figure A.1d only shows slight signs of overfitting as the nominal number of background events to the right of zero is slightly higher for the test set than for the training set. An excess of signal events is present in the two bins from 0.7 to 0.8 similarly to the $C_{tB} = +5$ sample. Further validation would be necessary to assess the interpretation that these events have EFT contribution.

The highest achievable expected significance per bin is 17.6 for the BDT output in Figure A.2a. No background events are observed for output values larger than 0.35 so that the expected significance cannot be calculated for output values larger than 0.35. This suggests a discrimination threshold around this value which would lead to approximately 45 signal events with no background contamination. The transverse momentum of the leading photon γ_{p_T} is the most discriminating variable with a feature importance of 0.125. Only the middle part of the γ_{p_T} spectrum in Figure A.2b is considered due to increasing background contamination with the inclusion of the neglected background processes in the lower part of the spectrum and considerable fluctuations in the upper part. The expected significances in the region from 60 GeV to 140 GeV are below 5.

The classification of $tq\gamma$ with $C_{tW} = +2$ and SM background is highlighted with ROC curves in Figure A.3, a-c. These show that discrimination is possible for all three background combinations, although the classification against both backgrounds does not seem to be convincing as three of the five ROC curves for the individual CV folds cross the random guess line. The ROC curves with classification of EFT enriched $tq\gamma$ and SM $tq\gamma$ are very narrow and only slightly above random guess. Again, discrimination against $t\bar{t}\gamma$ is the most successful. The output distribution in Figure A.3d shows considerable signs of overfitting as the nominal number of background events to the right of zero is slightly higher for the test set than for the training set while the signal events of the two sets are compatible within the statistical uncertainty in that range. No excess of signal events is observed for values larger than 0.3 and the output is more narrow than for the two C_{tB} samples. This indicates that a value of $C_{tW} = +2$ influences the cross section more than the shapes of the kinematic variable distributions. Nevertheless, the output distribution appears to have a Gaussian shape and the BDT seems to classify the events consistently, although with moderate skill.

It is possible to achieve an expected significance of 35.6 in the bin $[0, 0.05]$ for the BDT output shown in Figure A.4a. However, this high value is due to the considerably higher cross section for the $C_{tW} = +2$ sample compared to the other EFT enriched samples. Additionally, this value is likely to decrease greatly with the inclusion of neglected background processes as this bin is close to the typical cut on the BDT output of zero. The transverse momentum of the leading lepton is the most discriminating variable for $C_{tW} = +2$ with a feature importance of 0.175 and its spectrum is shown Figure A.4b. It is difficult to compare the bin-wise expected significances that are as high as 9.1 in the middle part of the spectrum as the cross section for this sample is a lot higher than the cross sections for the other considered Wilson coefficient

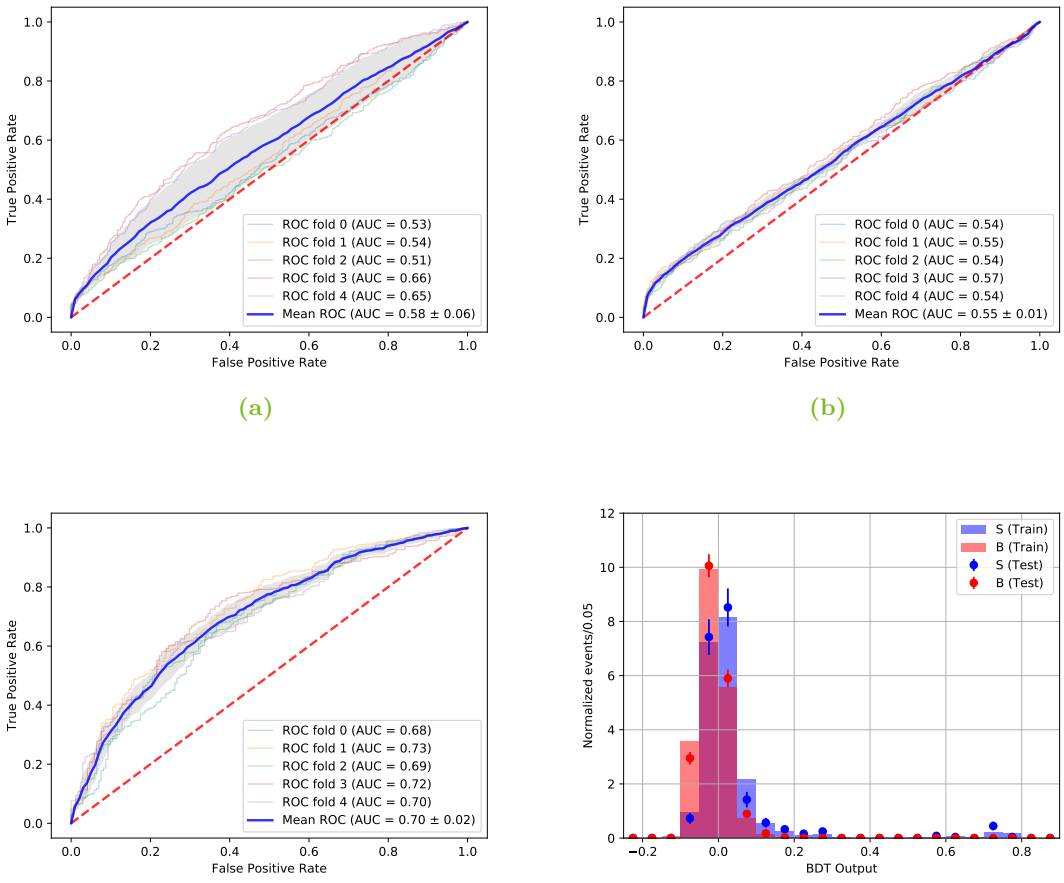
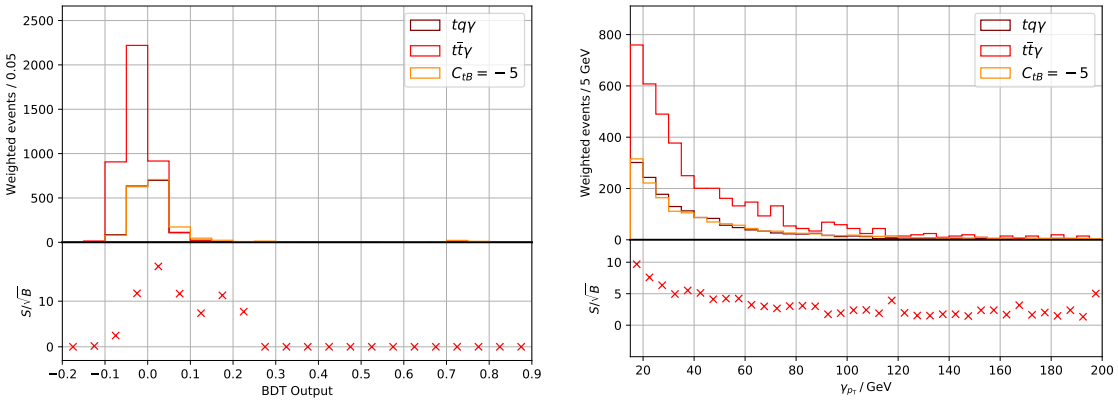


Figure A.1: ROC curves for the BDTs with training and evaluation using both backgrounds (a), using $tq\gamma$ (b) and using $t\bar{t}\gamma$ (c), with the $C_{tB} = -5$ sample as the signal. The dashed red line represents the ROC of a random guess classifier while the shaded grey region is the 1σ-band of the mean ROC curve calculated with the corrected sample standard deviation from the individual CV folds. The BDT output for training and evaluation using both backgrounds is shown separately for training and test data in (d). The signal is shown in blue, while the background events are shown in red. The histograms are normalised to unit area and the error bars are Poisson uncertainties.



(a) (b)

Figure A.2: BDT output distribution for the $C_{tB} = -5$ sample with training and evaluation using both background processes (a) and the transverse momentum of the leading photon, the most discriminating variable for this sample (b). The lower plots show the expected significance per bin, which is set to zero if the bin does not contain background events.

values. Because of that, it is a consistent result that the expected significances are higher than the ones achievable for the other samples.

The ROC curves for the classification of $tq\gamma$ with $C_{tW} = -2$ and SM background can be seen in Figure A.5, a-c. The evaluation against both backgrounds is considered unsuccessful as the AUC of 0.53 ± 0.04 is compatible with a random guess. The discrimination against SM $tq\gamma$ occurs with low variance as the AUC is 0.56 ± 0.01, indicating the presence of a few pronounced differences induced by the EFT operators. This interpretation is supported by the fact that there are only three discriminating features for the $C_{tW} = -2$ sample: The transverse momentum of the leading photon, the leading lepton and the invariant mass of the leading lepton and the leading photon. The classification of EFT enriched $tq\gamma$ and $t\bar{t}\gamma$ is successful with a nominal AUC of 0.68. The output distribution in Figure A.5d looks peculiar: A big peak to the left of zero is followed by an empty bin and a flatter, washed-out peak. An excess of signal over background events is present as a tail at high BDT outputs. These odd features are difficult to interpret consistently and might be an artifact of the limited amount of data available in this thesis.

The achievable expected significances per bin should be treated with caution due to the peculiarity of the output. A value of 31.2 is achievable in the $[0.55, 0.6]$ bin of the BDT output which can be seen in A.6a. The angular separation of the leading photon and the leading b-tagged jet is the most discriminating feature with a feature importance of 0.389 and its distribution is shown in A.6b. It is possible to achieve a bin-wise expected significance of 11.4 with this kinematic variable. Although the shape differences between $tq\gamma$ and $t\bar{t}\gamma$ are visible in the plot (the distribution for $tq\gamma$ does not fall of as sharply as the $t\bar{t}\gamma$ distribution), they are not pronounced enough to achieve a satisfactory significance as the inclusion of neglected background processes will contaminate the spectrum considerably.

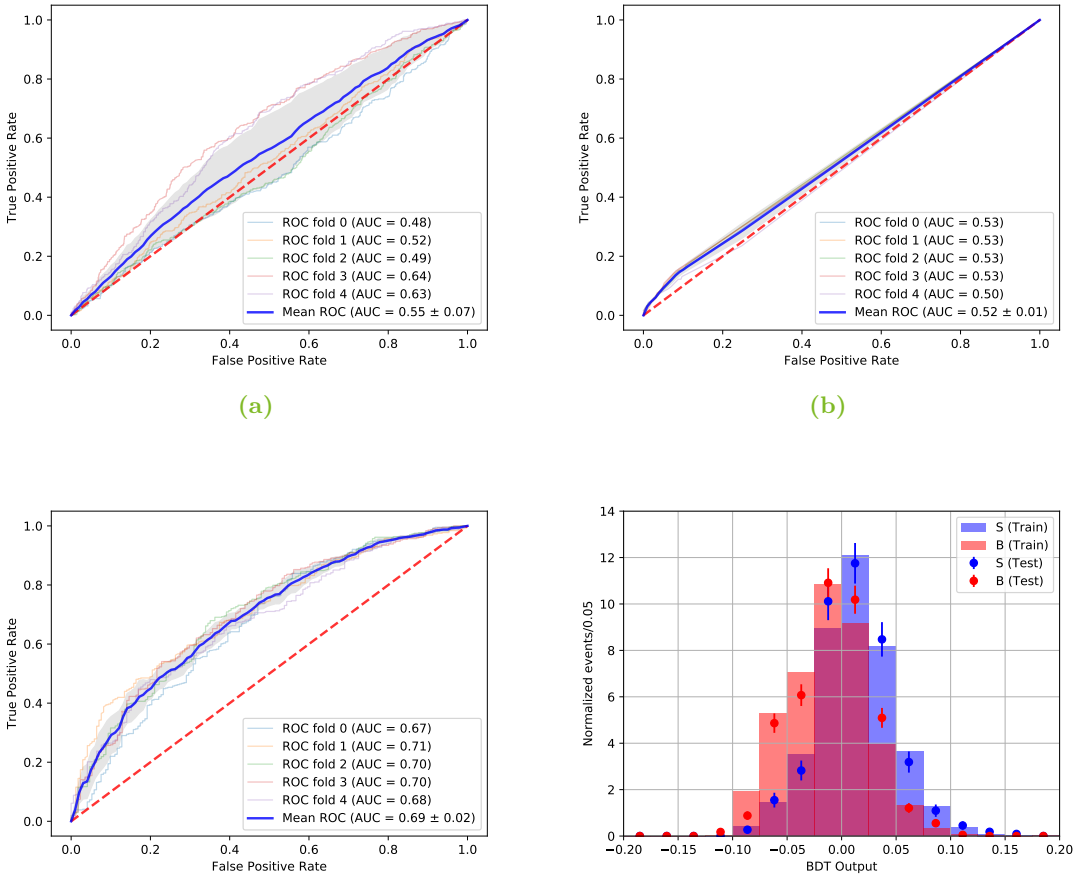
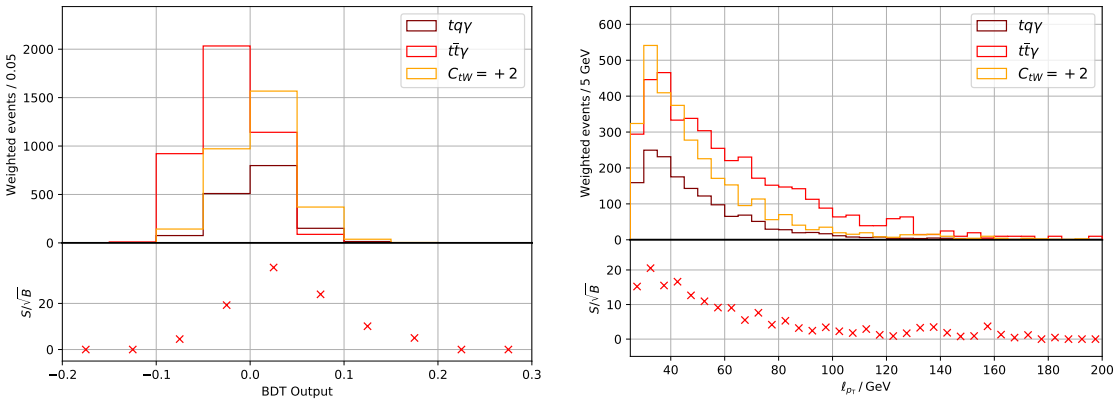


Figure A.3: ROC curves for the BDTs with training and evaluation using both backgrounds (a), using $tq\gamma$ (b) and using $t\bar{t}\gamma$ (c), with the $C_{tW} = +2$ sample as the signal. The dashed red line represents the ROC of a random guess classifier while the shaded grey region is the 1σ-band of the mean ROC curve calculated with the corrected sample standard deviation from the individual CV folds. The BDT output for training and evaluation using both backgrounds is shown separately for training and test data in (d). The signal is shown in blue, while the background events are shown in red. The histograms are normalised to unit area and the error bars are Poisson uncertainties.



(a) **(b)**

Figure A.4: BDT output distribution for the $C_{tW} = +2$ sample with training and evaluation using both background processes (a) and the transverse momentum of the leading lepton, the most discriminating variable for this sample (b). The lower plots show the expected significance per bin, which is set to zero if the bin does not contain background events.

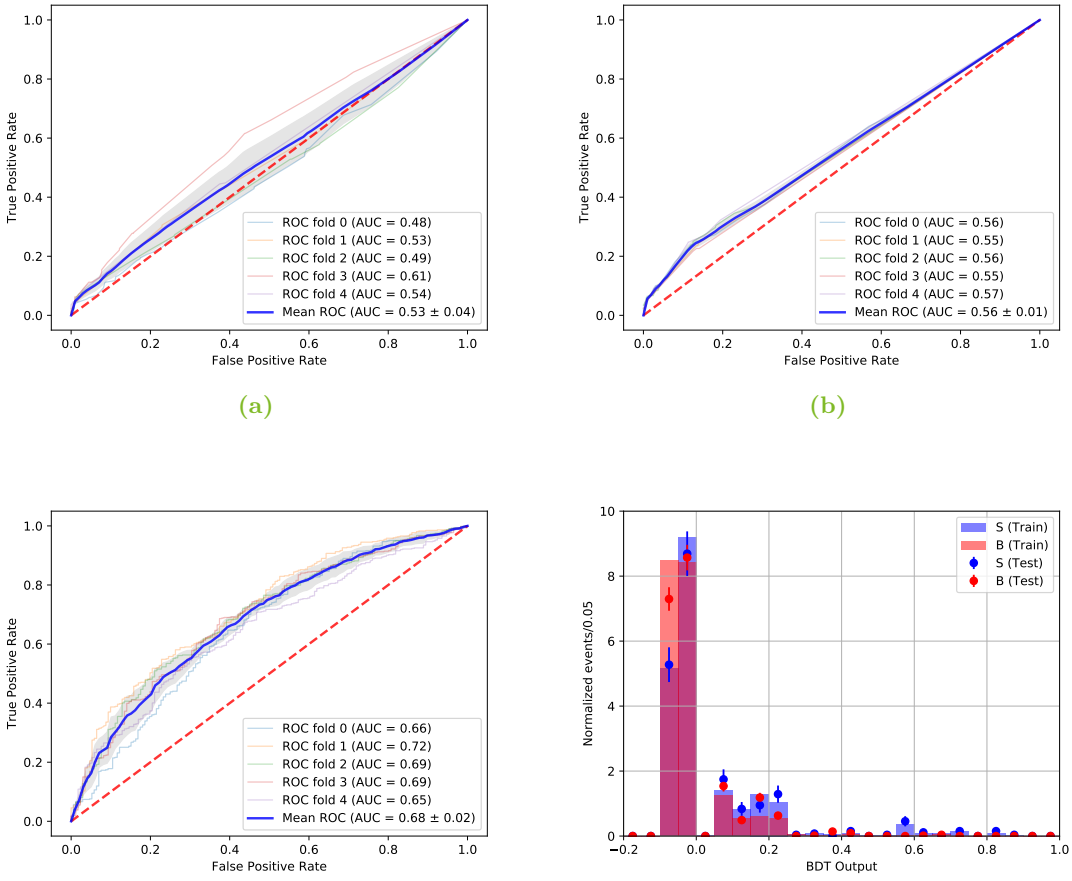
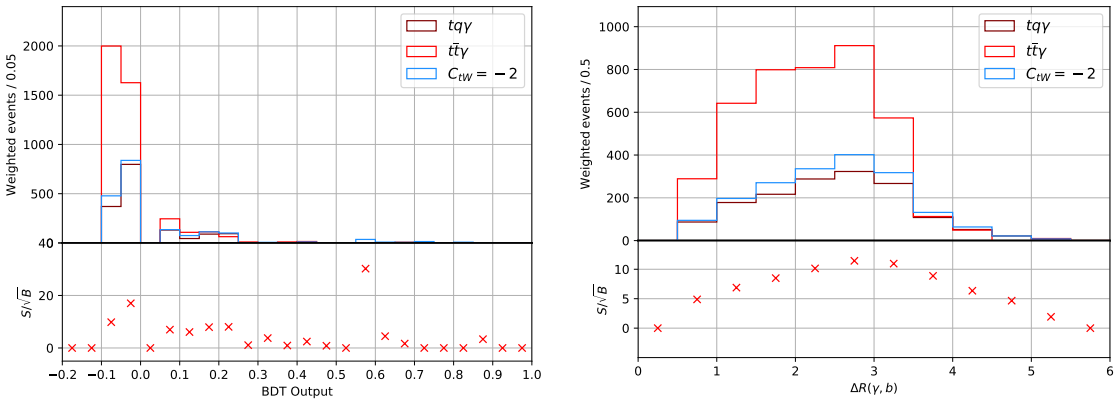


Figure A.5: ROC curves for the BDTs with training and evaluation using both backgrounds (a), using $tq\gamma$ (b) and using $t\bar{t}\gamma$ (c), with the $C_{tW} = -2$ sample as the signal. The dashed red line represents the ROC of a random guess classifier while the shaded grey region is the 1σ-band of the mean ROC curve calculated with the corrected sample standard deviation from the individual CV folds. The BDT output for training and evaluation using both backgrounds is shown separately for training and test data in (d). The signal is shown in blue, while the background events are shown in red. The histograms are normalised to unit area and the error bars are Poisson uncertainties.



(a) (b)

Figure A.6: BDT output distribution for the $C_{tW} = -2$ sample with training and evaluation using both background processes (a) and the angular separation ΔR of the leading photon and the leading b-tagged jet, the most discriminating variable for this sample (b). The lower plots show the expected significance per bin, which is set to zero if the bin does not contain background events.

Bibliography

[1] D0 Collaboration, *Observation of Single Top Quark Production*, Phys. Rev. Lett. **103** (2009) 092001, arXiv: 0903.0850 [hep-ex].

[2] CDF Collaboration, *First Observation of Electroweak Single Top Quark Production*, Phys. Rev. Lett. **103** (2009) 092002, arXiv: 0903.0885 [hep-ex].

[3] Tevatron Electroweak Working Group, *Combination of CDF and D0 Measurements of the Single Top Production Cross Section*, (2009), arXiv: 0908.2171 [hep-ex].

[4] CMS Collaboration, *Evidence for the associated production of a single top quark and a photon in proton-proton collisions at* $\sqrt{s} = 13$ *TeV*, Phys. Rev. Lett. **121** (2018) 221802, arXiv: 1808.02913 [hep-ex].

[5] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, JHEP **07** (2014) 079, arXiv: 1405.0301 [hep-ph].

[6] D. Galbraith and C. Burgard, *Example: Standard model of physics*, http://www.texample.net/tikz/examples/model-physics/.

[7] M. et al. (Particle Data Group) Tanabashi, Phys. Rev. D **98** (2018) 030001, eprint: 1802.07237.

[8] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Phys. Lett. B **716** (2012) 1, arXiv: 1207.7214 [hep-ex].

[9] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. B **716** (2012) 30, arXiv: 1207.7235 [hep-ex].

[10] F. Zwicky, *Die Rotverschiebung von extragalaktischen Nebeln*, Helv. Phys. Acta **6** (1933) 110.

[11] L. Canetti, M. Drewes, and M. Shaposhnikov, *Matter and antimatter in the universe*, New Journal of Physics **14** (2012) 095012.

[12] E. Fermi, *Versuch einer Theorie der β-Strahlen. I*, Z. Phys. **88** (1934) 161.

[13] W. Buchmüller and D. Wyler, *Effective Lagrangian Analysis of New Interactions and Flavor Conservation*, Nucl. Phys. B **268** (1986) 621.

[14] D. Barducci et al., *Interpreting top-quark LHC measurements in the standard-model effective field theory*, (2018), arXiv: 1802.07237 [hep-ph].

[15] ATLAS Collaboration, *Observation of top-quark pair production in association with a photon and measurement of the* $t\bar{t}\gamma$ *production cross section in pp collisions at* $\sqrt{s} = 7$ *TeV using the ATLAS detector*, Phys. Rev. D **91** (2015) 072007, arXiv: 1502.00586 [hep-ex].

[16] M. Ghneimat, *Probing the top-quark coupling to the photon through the cross-section measurement of* $t\bar{t}\gamma$ *production in pp collisions with the ATLAS detector*, (2018), URN: urn:nbn:de:hbz:5n-50874.

[17] CMS Collaboration, *The CMS Experiment at the CERN LHC*, JINST **3** (2008) S08004.

[18] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, JINST **3** (2008) S08003.

[19] A. Buckley et al., *LHAPDF6: parton density access in the LHC precision era*, Eur. Phys. J. C **75** (2015) 132, arXiv: 1412.7420 [hep-ph].

[20] ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC*, (2019).

[21] T. Sjöstrand et al., *An Introduction to PYTHIA 8.2*, Comput. Phys. Commun. **191** (2015) 159, arXiv: 1410.3012 [hep-ph].

[22] ATLAS Collaboration, *Measurement of the $t\bar{t}Z$ and $t\bar{t}W$ production cross sections in multilepton final states using 3.2 fb^{-1} of pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Eur. Phys. J. C **77** (2017) 40, arXiv: 1609.01599 [hep-ex].

[23] J. de Favereau et al., *DELPHES 3, A modular framework for fast simulation of a generic collider experiment*, JHEP **02** (2014) 057, arXiv: 1307.6346 [hep-ex].

[24] S. Agostinelli et al., *GEANT4: A Simulation toolkit*, Nucl. Instrum. Meth. A **506** (2003) 250.

[25] M. Cacciari, G. P. Salam, and G. Soyez, *FastJet User Manual*, Eur. Phys. J. C **72** (2012) 1896, arXiv: 1111.6097 [hep-ph].

[26] M. Cacciari, G. P. Salam, and G. Soyez, *The anti-k_t jet clustering algorithm*, JHEP **04** (2008) 063, arXiv: 0802.1189 [hep-ph].

[27] CMS Collaboration, *Measurement of the Single-Top-Quark t-Channel Cross Section in pp Collisions at $\sqrt{s} = 7$ TeV*, JHEP **12** (2012) 035, arXiv: 1209.4533 [hep-ex].

[28] F. Pedregosa et al., *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research **12** (2011) 2825.

[29] Y. Freund, R. Schapire, and N. Abe, *A short introduction to boosting*, Journal-Japanese Society For Artificial Intelligence **14** (1999) 1612.

Danksagung

Zuallererst bedanke ich mich bei Herrn Prof. Dr. Kröninger, es mir ermöglicht zu haben, die vorliegende Bachelorarbeit beim Lehrstuhl für Experimentelle Physik IV zu schreiben, sowie für die gute Betreuung. Außerdem danke ich Herrn Prof. Dr. Dr. Rhode für die Zweitkorrektur meiner Arbeit und Herrn Priv.-Doz. Dr. Johannes Erdmann für seine Betreuung, welche aus vielen Gesprächen über die Analysetechniken und die zugrundeliegende Physik und dem Korrekturlesen zweier Kapitel bestand. Ich danke Björn Wendland für die zahlreichen Gespräche über und Anregungen für sämtliche Aspekte der Arbeit und das Korrekturlesen der Arbeit. Für das Korrekturlesen des Theoriekapitels danke ich Cornelius Grunwald.

Ich möchte meinen Dank an den gesamten Lehrstuhl ausweiten. Ein professioneller, aber stets freundlicher Umgang trägt zu einer produktiven Atmosphäre bei, in der Fragen aller Art bereitwillig beantwortet werden. Auch danke ich den anderen Bachelorstudenten am Lehrstuhl. Die Diskussionen mit Ihnen haben interessante Impulse für Aspekte der Arbeit geliefert und ich danke insbesondere meinen Bürokollegen für die freundliche und konstruktive Atmosphäre.

Zuletzt möchte ich meinen Freunden und meiner Familie für die Unterstützung während dieser Arbeit und während des gesamten Studiums danken.

Eidesstattliche Versicherung

Ich versichere hiermit an Eides statt, dass ich die vorliegende Abschlussarbeit mit dem Titel "Studies for the sensitivity to dimension six operators in the context of effective field theories in single-top-quark production with a photon" selbstständig und ohne unzulässige fremde Hilfe erbracht habe. Ich habe keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, sowie wörtliche und sinngemäße Zitate kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Ort, Datum	Unterschrift

Belehrung

Wer vorsätzlich gegen eine die Täuschung über Prüfungsleistungen betreffende Regelung einer Hochschulprüfungsordnung verstößt, handelt ordnungswidrig. Die Ordnungswidrigkeit kann mit einer Geldbuße von bis zu 50 000 € geahndet werden. Zuständige Verwaltungsbehörde für die Verfolgung und Ahndung von Ordnungswidrigkeiten ist der Kanzler/die Kanzlerin der Technischen Universität Dortmund. Im Falle eines mehrfachen oder sonstigen schwerwiegenden Täuschungsversuches kann der Prüfling zudem exmatrikuliert werden (§ 63 Abs. 5 Hochschulgesetz –HG–).

Die Abgabe einer falschen Versicherung an Eides statt wird mit Freiheitsstrafe bis zu 3 Jahren oder mit Geldstrafe bestraft.

Die Technische Universität Dortmund wird ggf. elektronische Vergleichswerkzeuge (wie z. B. die Software "turnitin") zur Überprüfung von Ordnungswidrigkeiten in Prüfungsverfahren nutzen.

Die oben stehende Belehrung habe ich zur Kenntnis genommen.

Ort, Datum	Unterschrift