TECHNISCHE UNIVERSITÄT WIEN

HEPHY INSTITUT FÜR HOCHENERGIEPHYSIK

DIPLOMARBEIT

Likelihood free inference with advanced machine learning techniques

zur Erlangung des akademischen Grades

Diplom-Ingenieurin

im Rahmen des Studiums

Technische Physik

eingereicht von

Lena Wild, BSc MA MA
Matrikelnummer: 11824551

ausgeführt am Institut für Hochenergiephysik
der Österreichischen Akademie der Wissenschaften

Betreuung
Betreuer: Privatdoz. DI Dr. Robert Schöfbeck

Wien, 12.06.2023

Unterschrift Verfasserin

Unterschrift Betreuer

Abstract

Although many theories locate hypothetical phenomena beyond the Standard Model at energy scales far above the current experimental reach, tiny deviations from the Standard Model prediction are expected to show in the tails of data collected with the Compact Muon Solenoid experiment at the Large Hadron Collider at CERN. These deviations can be described through the insertion of effective operators of mass dimension higher than four, treating the Standard Model as a low-energy approximation of the hypothetical high energy theory.

In this thesis, we adopt novel machine learning techniques based on simulation (simulation-based inference) to teach the machine the optimal test statistic according to the Neyman-Pearson lemma for four-fermion operator insertions in four-top production and production of two top and two bottom quarks. The centerpiece of this approach consists in exploiting the polynomial structure of the effective field theory prediction as a function of the coefficients of the operators, i.e., the Wilson coefficients. Hence, learning only a small number of coefficient functions allows to parametrize an optimal classifier in the full parameter space. With the learned coefficient functions at hand, we then set nuisance-free limits in an unbinned likelihood ratio test up to quadratic order in the polynomial expansion. In this way, we investigate the neural network's performance in learning the yield- and shape-related modifications, thus probing new forces between four heavy quarks.

On the machine learning side, we combine Deep Neural Networks with Long Short Time Memory layers to extract information not only from scaler observables, but also from the variable length jet system in analogy to speech recognition. By also probing this Multivariate Analysis setup in the simpler, more robust setting of multi-classification, we test, optimize and cross-validate the configuration of the network.

In instantiating a complete workflow of sample generation, training with simulation-based inference, and the limit setting procedure, we obtain projected limits on the Wilson coefficients of four-fermion operators in $\mathrm{t\bar{t}t\bar{t}}$ and $\mathrm{t\bar{t}b\bar{b}}$ with and without $\mathrm{t\bar{t}}$ background. Thus, we demonstrate the potential of these novel neural network architectures and machine learning techniques for future analyses.

Contents

Chapter 1

Introduction

In 2012, the discovery of the Higgs Boson at the Large Hadron Collider (LHC) [1] at CERN marked a climax in the fulminant triumph of the Standard Model of Particle Physics (SM) [2, 3]. However, phenomena like dark matter or the baryon asymmetry, not to mention the fundamental force of gravity, are not part of the SM explanation, just to name a few. Hence, the search for a physics beyond the Standard Model (BSM) has long begun, both on the experimental and the theoretical side.

As no traces of BSM physics are being found at the currently accessible energy scale of LHC, many BSM theories locate the new physics at energy scales Λ, considerably higher than the current LHC reach. Without investigating any specific of these vividly discussed approaches, the Standard Model Effective Field Theory (SMEFT) [4–9] provides a framework for parametrizing and re-evaluating LHC data on the look-out for tiny BSM signatures. In treating the SM as an effective limit of the BSM theory at low energy scale, SMEFT predicts tiny deviations in the observed quantities' tails due to BSM physics at Λ. At the heart of this framework are so-called effective field operators that modify the cross sections for SM processes while keeping the SM symmetries and its particle content intact [6].

In this thesis, we exploit the simple polynomial structure that the SMEFT predicts for BSM cross sections in the LHC energy range with a new, powerful Machine Learning approach. On the side of the technical implementation, we use a novel combination of Deep Neural Networks (DNNs) [10] and pair them with Long Short Term Memory (LSTM) [11–14] layers to optimally extract the BSM nuisances from scalar event-level observables and the jet system. On the side of the machine learning algorithm, we adapt the findings of Ref. [15–29] on simulation-based inference techniques to make the machine learn the optimal test statistic – an intractable quantity – from a tractable target. With four top quark production and production of two top and two bottom quarks as our physics cases, we then investigate all effective field operators that affect interactions between heavy quarks [9, 30, 31].

In chapter 2, we start by describing the LHC at CERN with a particular focus upon the technical details of the Compact Muon Solenoid (CMS) - the detector at the center of our analysis [32]. Subsequently, in chapter 3, we introduce our physics cases in the SMEFT notation and characterize the relevant four-fermion operators. In chapter 4, we derive the loss function in simulation-based inference to teach the optimal test statistic to the machine, whereas in chapter 5 we describe the technical details of our network. The optimization of the hyperparameters is subject of chapter 6, together with the sample generation. Additional to simulation-based inference, we will also perform multi-classification in $t\bar{t}t\bar{t}$ signal and $t\bar{t}$ background with our network configuration, to probe the implementation in a more sturdy setting and use it as a proxy. In the final chapter 7, we will evaluate the network's performance and compute the log likelihood ratio for single operator insertion.

Chapter 2

CERN, LHC and CMS

2.1 CERN - The European Council for Nuclear Research

In 1954, the European Council for Nuclear Research (*Conseil européen pour la recherche nucléaire*, CERN) was founded with the aim of performing world-class research in particle physics. Located at the Franco-Swiss border near Geneva, it provides scientists from all over the world with the tools to push the boarders of science and technology for the benefit of all [33]. Since then, the numerous collaborations at CERN reported a number of significant discoveries in the physics of fundamental particles – from which the discovery of the Higgs Boson in 2012 at Large Hadron Collider (LHC) [1] might well be the most prominent [2, 3]. Other milestones include, e.g., the discoveries of the W and Z bosons [34, 35] and CP-violation in the decays of neutral kaons [36]. Additionally, as experiments of this scale require an efficient and automated tool for sharing information with collaborators all around the world, the invention of the World Wide Web marks yet another groundbreaking developement at CERN [37].

Over the years, new accelerators have constantly been added to the accelerator complex at CERN up to the current status shown in Fig. 2.1, increasing the energy of the particle beams to a maximum of 6.5 TeV each [38]. Starting from the least energetic, the chain of subsequent energy boosts begins with the linear accelerator 4 (Linac4) that accelerates negatively charged hydrogen ions to 160 MeV. During injection into the subsequent machine, the Proton Synchrotron Booster (PSB), two electrons are stripped from the hydrogen ions, leaving behind only the protons. After reaching 2 GeV in the PSB, the proton beam is boosted to 26 GeV in the Proton Synchrotron (PS). Finally, the Super Proton Synchrotron reaches 450 GeV, leaving the proton beams ready to be injected in heart of CERN's accelerator chain, the LHC [1, 38]. There, protons reach 6.5 TeV each, as they are injected in two pipes and accelerated in opposite directions. Lastly, the two beams are brought to collision at a center of mass energy of 13 GeV, making the LHC the world's largest and most powerful particle accelerator [39].

Other facilities at CERN comprise, e.g., the Low Energy Ion Ring (LEIR), where the first creation of antihydrogen succeeded in 1996 [40], the time-of-flight detector for neutrons (n_TOF) and various other accelerators and decelerators.

2.2 LHC - The Large Hadron Collider

The heart of CERN's accelerator complex is the Large Hadron Collider, a ring of superconducting magnets and accelerating segments in a tunnel of 27 kilometer length. In ultra-high vacuum at temperatures around 1.9 K, protons travel in two separate beams in opposite directions to reach 6.5 TeV each, before they are directed into collisions in the experiments of the LHC [41].

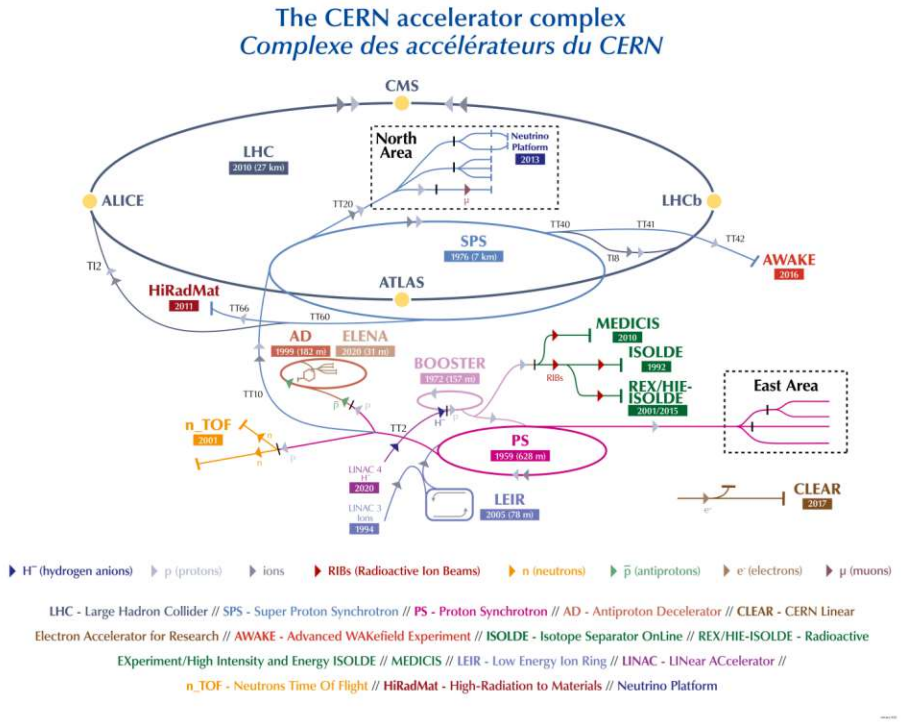


Figure 2.1: The accelerator complex at CERN. The chain of LINAC4, PSB, PS, SPS and finally LHC boosts protons to collision energies of 13 TeV. Image taken from Ref. [38].

The five main goals of the LHC are [41]:

1. **Confirming the origin of mass as explained by Robert Brout, Francois Englert and Peter Higgs in 1964.** When CERN announced the discovery of the Higgs boson in July 2012 [2, 3], this goal was partly achieved, whereas the detailled investigation of the Higgs is still ongoing. An example of current efforts is found in Ref. [42].

2. **Searching for a unified description of all four fundamental forces including gravity.** In this context, the theory of Supersymmetry (SUSY) is at the center of attention, as well as other theories that go beyond the Standard Model (SM) explanations. An overview over recent SUSY efforts in CMS and ATLAS is given in Ref. [43], whereas Ref. [44] contains the LHC Working Group Report on Effective Field Theories (EFT).

3. **Searching particles or phenomena responsible for dark matter and dark energy.** These searches are closely interconnected with the previously mentioned goals, as theories BSM are investigated. For example, recent investigations are described in the Dark Matter LHC Working Group Report [45].

4. **Investigating the questions related to matter and antimatter**, e.g., the proton anti-proton production reported by LHCb in 2017 [46, 47]. The detector is described in more detail below.

5. **Studying the physics of heavy-ion collisions.** The physics of strongly interacting matter at very high energy densities, the so-called quark-gluon plasma, is at the core of investigations at ALICE, one of the four main experiments at the LHC [48].

In addition to several fixed-target and antimatter experiments as well as other experimental setups that use the LHC injector chain, nine experiments are located directly at the LHC. Among these, the four main experiments are:

ATLAS (**A** **T**oroidal **L**HC **A**pparatu**s**) is the largest general-purpose detector at LHC, designed to cover a broad range of physical questions. A collaboration of more than 5500 scientists from 245 institutes in 42 countries (March 2022) investigates topics reaching from the discovery and subsequent study of the Higgs boson [2] to the search for extra dimensions, BSM physics and hints for dark matter. With its length of 46 m, height of 25 m and width of 25 m, the 7000 tons of ATLAS are the largest volume of all particle detectors ever constructed [49–51].

CMS The **C**ompact **M**uon **S**olenoid is the second general-purpose detector at the LHC. As the name suggests, a key task of the detector system is the reconstruction of muon tracks. With a height of 15 m and a length of only 21 m, it is rather compact compared to e.g., ATLAS. The core of CMS consists of the largest solenoid magnet ever made, generating a field of 3.8 T. Although the research questions of CMS and ATLAS are similar (SM and Higgs, search for BSM-physics, extra dimensions and dark matter), the technology used is different, allowing results to be cross-validated among the two collaborations. In May 2022, 5500 particle physicists, engineers, technicians, students and support staff from 241 institutes in 54 countries were contributing to the experiment [32, 52, 53].

ALICE (**A** **L**arge **I**on **C**ollider **E**xperiment) is one of the two smaller experiments at LHC, focussing on the specific field of strongly interacting matter at extreme energy densities. The detector is 26 m long, 16 m high and 16 m wide and specifically designed to study the so-called phase of quark-gluon plasma. In April 2022, almost 2000 scientists from 174 institutes in 40 countries were involved in the ALICE experiment [48, 54, 55].

LHCb The **L**arge **H**adron **C**ollider **b**eauty tackles another specific physical question: the investigation of the b-quark. The goal hereby is to investigate the tiny differences between matter and antimatter. Contrarily to the other detectors, the LHCb is not built around the collision point of the two beams, but consists of a row of subsequent components with an overall length of 21 m. In this configuration, it mainly detects decay products of the collision that travel in forward direction. 1565 scientists, engineers and technicians from 20 countries were part of the LHCb collaboration in March 2022 [46, 56, 57].

The location of the four largest experiments is schematically shown in Fig. 2.2.

To cover all five goals of LHC mentioned above, there are five smaller experiments at LHC. These experiments are listed below with references for detailled explanation:

- **TOTEM** (**Tot**al, **e**lastic and diffractive cross-section **m**easurement) [58, 59],
- **LHCf** (**L**arge **H**adron **C**ollider **f**orward) [60, 61],
- **MoEDAL** (**Mo**nopole and **E**xotics **D**etector **a**t the **L**HC) and its upgrade **MAPP** (**M**oEDAL **A**pparatus for **P**enetrating **P**articles) [62, 63], as well as the two newest facilities
- **FASER** (**F**orw**a**rd **S**earch **E**xperiment) [64, 65], and
- **SND@LHC** (**S**cattering and **N**eutrino **D**etector at the LHC) [66, 67].

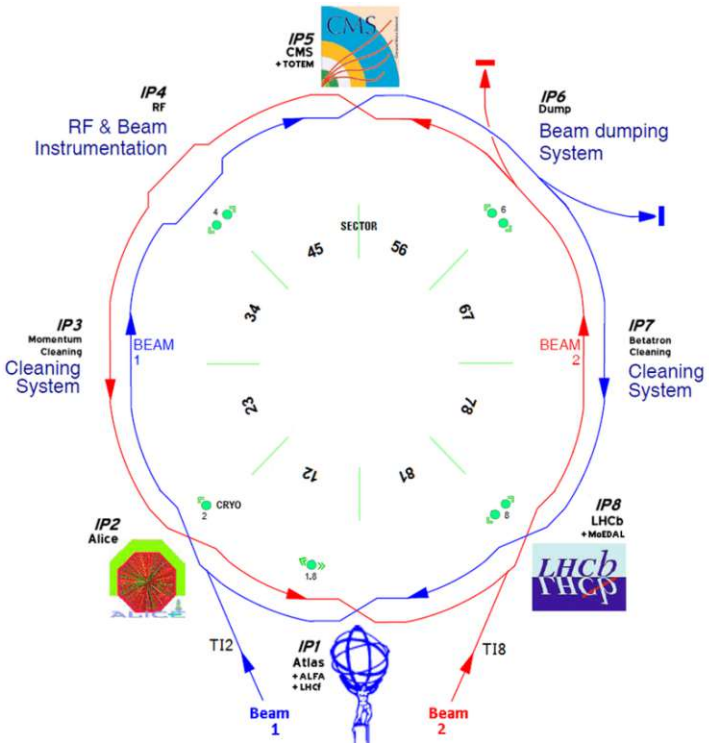


Figure 2.2: Sketch of LHC with the crossing points at IP1 (ATLAS), IP2 (ALICE), IP5 (CMS) and IP8 (LHCb). The beams switch position between inside and outside to assure equal length of the paths. Image taken from Ref. [68].

2.3 CMS - The Compact Muon Solenoid

This subsection is intended as a brief overview on the CMS detector and experiment. Further details regarding the detector's configuration and its components can be found in Ref. [32, 69–71].

Located at Point 5 – IP5 in Fig. 2.2 –, the Compact Muon Solenoid (CMS) detector is used as one of the two multi-purpose detectors at the LHC. Unlike other detectors, CMS was not build in-situ, but constructed in fifteen single "slices" that were then lowered into the cavern [32]. This allows the detector to be opened for maintenance and updates, for example during the winter shot-down in December 2022, shown in Fig. 2.3.
The experiment was built and designed to study proton-proton and lead-lead collisions at a centre-of-mass energy of up to 14 TeV and luminosities up to 10^{34} cm^{-2} s^{-1}. In particle physics, the luminosity $\mathcal{L}$ is given by

$$\mathcal{L}\sigma = \frac{dN}{dt} \tag{2.1}$$

and describes the number of recorded events in a certain time period given the cross section σ of the process.

Figure 2.3: CMS-detector opened for maintenance in December 2022.

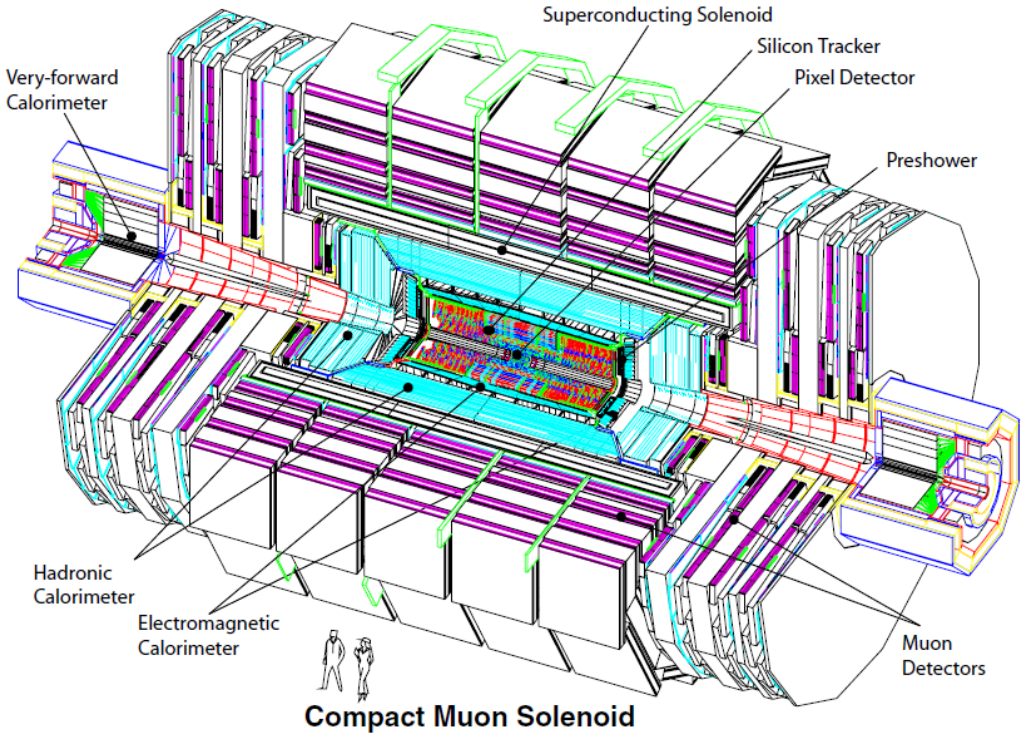


Figure 2.4: Sketch of the CMS detector. The superconducting solenoid contains three sub-detector systems: tracker, ECAL and HCAL. Outside the solenoid, the muon system is installed. Image taken from Ref. [72].

2.3.1 CMS detection systems

To fulfil the various needs of different analyses, a multitude of sub-systems is located in concentric cylinders around the collision point. The strongest magnet ever installed, a superconducting solenoid, bends tracks of charged particles coming from the proton-proton collisions by producing a field of 3.8 T. As the path of charged particles changes its curvature in the magnetic field $\mathbf{B}$ due to the Lorentz force $\mathbf{F} = q \cdot (\mathbf{v} \times \mathbf{B})$, measuring the particle's momentum is possible. For superconductivity to be stable, the helium-cooled coil of 6 m diameter and 12.5 m length with a stored energy of 2.6 GJ is operated at 4.5 K [32].

In order to detect the particles, trace back the path to possible secondary vertices and determine their momentum, three detector subsystems are hosted inside the compact solenoid:

Inner tracking system: A system of silicon pixels and silicon microstrips is used to reconstruct the tracks of charged particles. To gather accurate spatial information on primary and displaced vertices, the most sensitive tracking layers allow resolution up to 23 µm for single points, with the resolution decreasing to 100 µm for lower momenta. For muons, the resolution of the momentum is within some percentages. The 200 m^2 of active silicon area make the CMS tracker the largest silicon tracker ever build [32, 73].

Electromagnetic Calorimenter (ECAL): In this detector component, more than 65000 niobium doped lead tungstate ($PbWO_4$) crystals are used as scintillators to detect electrons and photons while measuring their energy. As scintillators emit light when electrons and photons pass through, an electrical current is generated in avalanche photodiodes and

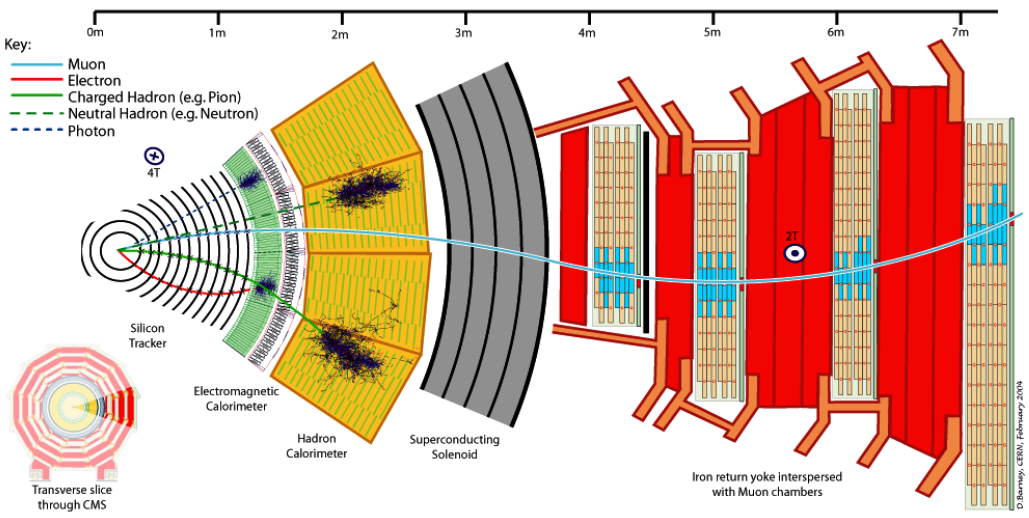


Figure 2.5: Slice of the CMS experiment in the barrel region with some exemplar particles and their path until detection. Image taken from Ref. [72].

vacuum phototriodes and passed on to the electronic system for readout. With the main challenges being stochastic effects, noise and non-uniformities, intercalibration errors and energy leakage, the energy resolution for, e.g., an electron of 120 GeV is 0.4% [32, 74].

Hadronic Calorimeter (HCAL) As the name suggests, the Hadronic Calorimeter is specially designed to detect strongly interacting particles, e.g., hadronic jets. Furthermore, also neutrinos or exotic particles can be indirectly detected via their missing transverse energy [69]. The calorimeter's multiple layers of brass absorbers and plastic scintillators lead to high rates of interaction with subsequent electronic readout of the particle shower information. Again, the information is then transferred to the electronic readout by means of quartz fibers [32, 75].

On the outside of the solenoid, the fourth detector system is installed. The **Muon System** is a distinct feature of the CMS detector as the latter has been specifically designed to detect muons over the full kinematic range of the LHC [32, 69]. In fact, due to muon having a significantly higher mass than electrons, they penetrate matter much deeper, leaving only tiny signals in the calorimeters. Hence, fairly all other particles of interest are already stopped before leaving the solenoid.

Due to the large area that needs to be covered for efficient muon detection, different technologies are used to fulfill the systems's key tasks of muon identification, momentum measurement, and triggering [32, 76]:

- **Drift Tubes (DTs)** are used in the so-called "barrel region", i.e., the "cylindrical" walls of CMS. When a muon passes the DT, the gas inside is ionized, with a subsequent electric impulse being created at a charged wire in the DT's center.
- **Cathode Strip Chambers (CSCs)** meet the demands of the endcap regions, i.e., the front and back end of the CMS detector, as the magnetic field is no longer uniform. The advantages of CSCs are their fast response time paired with fine segmentation, as well as resistance to radiation in the regions of elevated muon flux.

- Finally, **Resistive Plate Chambers (RPCs)** are used to complement DTs and CSCs in both barrel and endcap regions. They consist of double-gap chambers and are operated in avalanche mode. Hence, RPCs' extremely fast response time allows quick readout, making them an efficient trigger for determining the exact bunch crossing that at LHC take place every 25 ns. However, their positional resolution is inferior to DTs and CSCs [32, 76, 77].

2.3.2 Trigger and Data Acquisition at CMS

As interaction rates in proton-proton collisions are high at LHC, the interval between two subsequent beam crossings is only 25 ns long (an equivalent of 40 MHz). At the nominal design luminosity of 10^{34} cm^{-2} s^{-1}, approximately 20 proton-proton collisions take place within one bunch crossing. This entails the need for an efficient reduction of the recorded events, as such amounts of data would be impossible to process and store [32].

To accomplish this task, CMS has a two staged **trigger systems**. First, Level-1 Trigger (L1) of custom-designed, programmable electronics reduces the output rate from 40 MHz to a maximum of 100 kHz by using coarsely segmented data from the calorimeters and the muon system. The high-resolution data is held back in pipelined memories in the front-end electronics [32]. In this process, the L1 trigger has only a very short latent time of 3.2 µs. Then, the software-implemented High-Level Trigger (HLT) is hosted in a computer farm of around 1000 processing units and has access to the full data [32].

Based on this two-level trigger system, the **CMS Data Acquisition** (DAQ) system can cope with maximum input rates of 100 kHz from the L1 trigger which result in approximately 100 GB/s. Additionally, it hosts enough computational power for the HLT's software filter system [32, 78, 79]. The sketch of the triggers and the complete DAQ system of CMS are shown in Fig. 2.6.

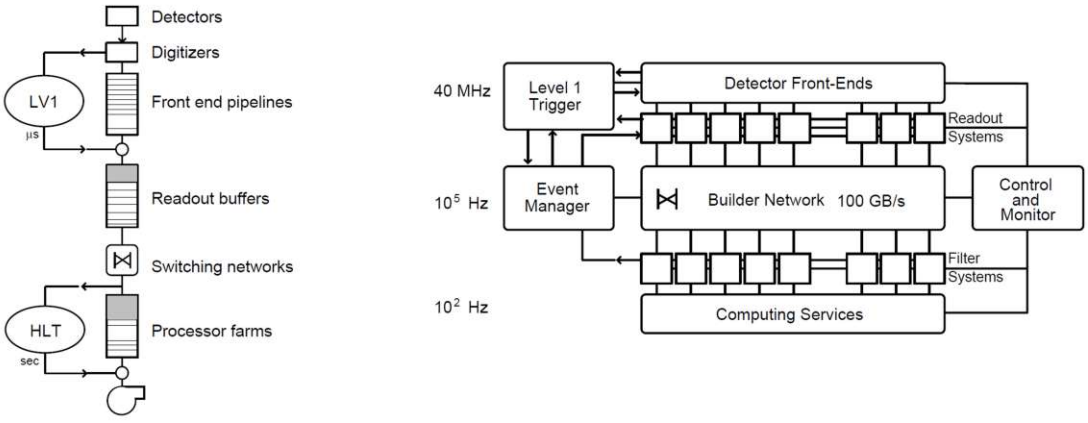


Figure 2.6: Left: Sketch of Level-1 trigger and High Level Trigger at CMS. Right: Complex of trigger and DAQ with respective working point frequencies. Images taken from Ref. [79].

Chapter 3

The Standard Model and Standard Model Effective Field Theory

3.1 The Standard Model and beyond

In the relatively short time span since its development, the SM of Particle Physics has proven to be the most precise theory of particles and their respective forces that we have. With its precise theoretical predictions at hand, subsequent discovery of a considerable number of new particles and processes in accelerator facilities was just a matter of time. Finally, in 2012, the discovery of the Higgs Boson at LHC marked the climax in a long series of spectacular discoveries [2, 3].

This incredible success can and will, however, not be the end of the line. In fact, significant problems in particle physics are still up to be solved even with the SM at hand. This is because there are already significant hints for BSM physics, i.e., for phenomena that outrun the explanation and theoretical description provided by the SM. Among these are for example the existence of dark matter (see e.g., Ref. [80, 81]), the excess of matter over antimatter – the so-called baryon asymmetry (e.g., Ref. [82]) – or the lightness of the electroweak scale (e.g., Ref. [83]).

In more general terms, the phrase "BSM physics" is used in three slightly different ways [84]:

1. "BSM physics" refers to phenomena with existing experimental evidence, that are however not accomodated by the SM explanation. Examples are dark matter and neutrino oscillations, as well as all phenomena concerning gravity. In fact, gravity – one of the four fundamental forces – is not part of the SM explanation.

2. "BSM physics" can be used for phenomena that can be accomodated by the SM, but only when some sort of ad hoc parametrization is introduced. Examples are the Yukawa coupling and the strong CP angle.

3. In a third sense, "BSM physics" can be used to generally refer to any extension of the SM that might or might not solve any of the puzzles mentioned above.

With this broad field of BSM physics at hand, it is no wonder that many novel theoretical descriptions, models and frameworks are currently discussed. Among these are, e.g., supersymmetry (SUSY) [85], composite models [86], two-Higgs doublet models [87] and models with extra spatial dimensions [88], just to mention a few. Additionally, current experimental conditions limit the accessible parameter space for probing new theories, which constrains the number of BSM models competing in explaining the same result [89]. Nevertheless, a broad variety of BSM models remains, and the search for the "new" physics – both experimentally and in terms of theoretical framework – has long begun. However, no evidence for BSM physics has been reported in terms of new particles or processes so far.

3.2 Standard Model Effective Field Theory

Contrarily to what has just been said, Standard Model Effective Field Theory (SMEFT) [4–9] does not directly probe any specific scenario of the BSM models mentioned before. Instead, the aim of the SMEFT approach is to re-evaluate LHC data for tiny signatures and indirect effects of BSM physics. In fact, the working hypothesis of SMEFT is that the BSM energy scale is much higher than the reach of the LHC. Hence, the discovery of BSM particles or processes is not to be expected within the LHC energy of 13 TeV. Nevertheless, the spectra of kinematic observables recorded in LHC collisions may still display subtle traces of the high-energy BSM phenomenon [4–9].

The SMEFT – the leading model in first and ongoing attempts to re-interpret collider results such as, e.g., Ref. [90] – parameterizes these subtle traces and deviations by keeping the SM symmetries and particle content intact. The SM, on the other hand, is treated merely as an effective theory with a range of applicability up to a certain energy scale Λ. The field theory with validity above Λ must then satisfy the following requirements [9]

1. The gauge group of $\mathrm{SU(3)_C \times SU(2)_L \times U(1)_Y}$ of the SM must be part of the theory.
2. All SM degrees of freedom must be incorporated as fundamental or composite fields.
3. At energies below Λ, the field theory must reduce to the SM.

3.2.1 The SMEFT Lagrangian at mass dimension 6

Especially in fullfilling requirement 3, most approaches have used the decoupling of heavy particles with masses of order Λ or above to reduce to the SM at energies $< \Lambda$. This leads to an extension of the SM Lagrangian at mass dimension 4 by operators of higher dimensions that are suppressed by powers of Λ [9]. The SMEFT Lagrangian therefore reads (up to mass dimension 6):

$$\mathcal{L}_{\mathrm{SM}} = \mathcal{L}^{(4)}_{\mathrm{SM}} + \frac{1}{\Lambda}\sum_k C^{(5)}_k \mathcal{O}^{(5)}_k + \frac{1}{\Lambda^2}\sum_k C^{(6)}_k \mathcal{O}^{(6)}_k + O\left(\frac{1}{\Lambda^3}\right) \tag{3.1}$$

Here, $\mathcal{O}^{(i)}_k$ are operators and $C^{(i)}_k$ are the so-called Wilson Coefficients, i.e., dimensionless coupling constants at dimension i. The number of operators can be very large – at mass dimension dimension 6, e.g., there are as many as 59 [9]. Naturally, not all operators affect all processes. Hence, we now look at our physics case first, before coming back to the relevant operators.

3.3 The physics cases: $\mathrm{t\bar{t}t\bar{t}}$ and $\mathrm{t\bar{t}b\bar{b}}$

In this thesis, we pay particular attention to the two physics cases of four-top production and the simultaneous production of two top and two bottom quarks. In the SM, these processes can be characterized as follows:

Four-top production The SM predicts the rare production of four top quarks ($\mathrm{t\bar{t}t\bar{t}}$) with a small cross section of $12.0^{+2.2}_{-2.5}$ fb at next leading perturbative order (NLO) in quantum chromodynamics (QCD) with electroweak corrections [93]. Recently, discovery has been reported by CMS and ATLAS [91, 94]. Previous efforts that ultimately led to the discovery can be found in Ref. [95–100]. After their, e.g., gluon induced production or production through a Higgs boson, top quarks predominantly decay to a bottom quark and a W boson. Then, the W boson decays either to leptons or to quarks, leading to a large number of different final states [100]. Hence, the $\mathrm{t\bar{t}t\bar{t}}$ process is characterized by tiny rates, but distinctive signatures in a wealthy and energetic final state.

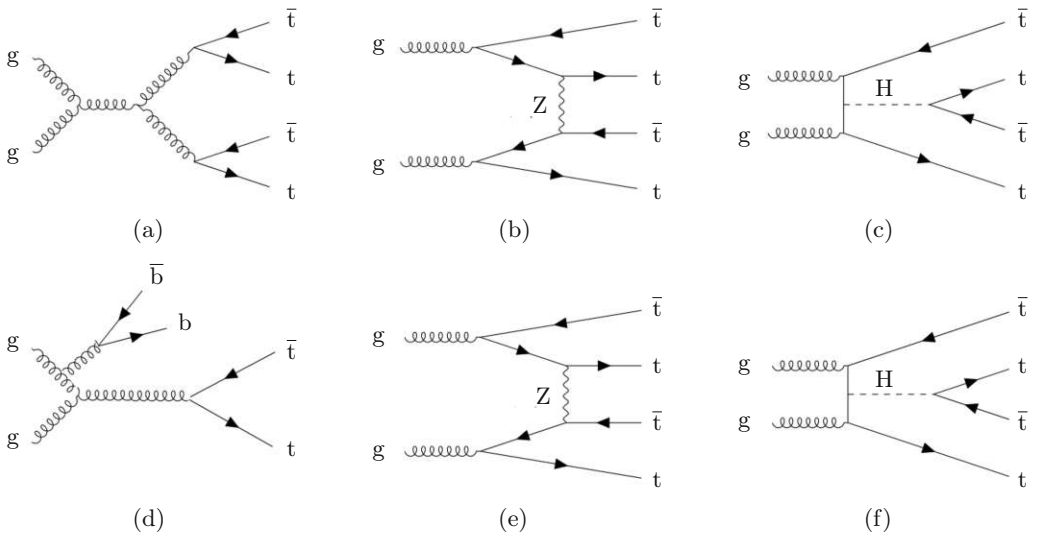


Figure 3.1: Production of $t\bar{t}t\bar{t}$ through (a) gluon induced strong interaction, (b) the exchange of a Z boson or (c) a Higgs boson. Production of $t\bar{t}b\bar{b}$ through (d) gluon induced strong interaction, (e) the exchange of a Z boson or (f) a Higgs boson. Figure partially taken from [91] and [92] with personal adaptions.

Production of 2 top and 2 bottom quarks Since Run 1, the production of two top and two bottom quarks $t\bar{t}b\bar{b}$ has been studied mostly for "generator tuning", as the extra $b\bar{b}$ is challenging to model and systematically limited. An overview over these efforts is given in Ref. [101]. In the last years, CMS and ATLAS measured various cross sections for different N_ℓ and $N_{\text{b-jets}}$ in the hadronic/$1\ell/2\ell$ and $1\ell + e\mu$ selection, respectively. These efforts are found in Ref. [92, 102–104].

Feynman diagrams of $t\bar{t}t\bar{t}$ and $t\bar{t}b\bar{b}$ for gluon induced production, production through a Z boson or through the Higgs are shown in Fig. 3.1.

3.4 Multi-classification in $t\bar{t}t\bar{t}$ and backgrounds

As stated in the introduction, we will use our MVA architecture (to be described in detail in the following chapters) not only for simulation-based inference, but also for multi-classification in the $t\bar{t}t\bar{t}$ signal plus $t\bar{t}b\bar{b}$, $t\bar{t}c\bar{c}$, $t\bar{t}$+light jets backgrounds setting. This is mainly because machine learning algorithms in this respect are well-known and provide assured results. Hence, we can use this classification task as a proxy for our network's configuration while relying only on sturdy machine learning algorithms, especially a simple loss function. Therefore, we only briefly describe the physical background of multi-classification for the sake of completeness before moving on to the SMEFT effects and the BSM physics – the key interest of this thesis.

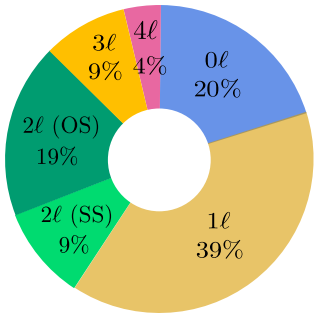


Figure 3.2: Branching ratios for $t\bar{t}t\bar{t}$ decays into leptonic final states. Figure taken from Ref. [98].

In this work, a investigate specific final state of the $t\bar{t}t\bar{t}$ decay: The two lepton opposite sign final state (2ℓOS). The branching ratio for 2ℓOS is 19%, as shown in Fig. 3.2. In this channel, large production rates of $t\bar{t}$ with a pair of heavy-flavour jets ($t\bar{t}b\bar{b}$, $t\bar{t}c\bar{c}$) require an efficient signal-to-background separation for an adequate sensitivity for the $t\bar{t}t\bar{t}$ discrimination. Other backgrounds such as $t\bar{t}Z$ and $t\bar{t}H$ will not be considered in this thesis, as we are not interested in the classification itself, but more in its behaviour in the machine learning context. In this respect, the high jet multiplicities of up to ten jets in the 2ℓOS channel make it a perfect set up to test LSTMs, that we will introduce in chapter 5.

3.5 SMEFT in $t\bar{t}t\bar{t}$ / $t\bar{t}b\bar{b}$

Coming back to the SMEFT, it follows naturally from our physics cases that the main interest of this thesis concerns new BSM forces among top quarks ($t\bar{t}t\bar{t}$) and, in a more general sense, heavy quarks ($t\bar{t}b\bar{b}$). In fact, EFT effects are almost complementary in $t\bar{t}t\bar{t}$ and $t\bar{t}b\bar{b}$, as shown in the Feynman diagrams of the effective interaction in Fig. 3.3, (b) and (d), based on Ref. [30] and [31].

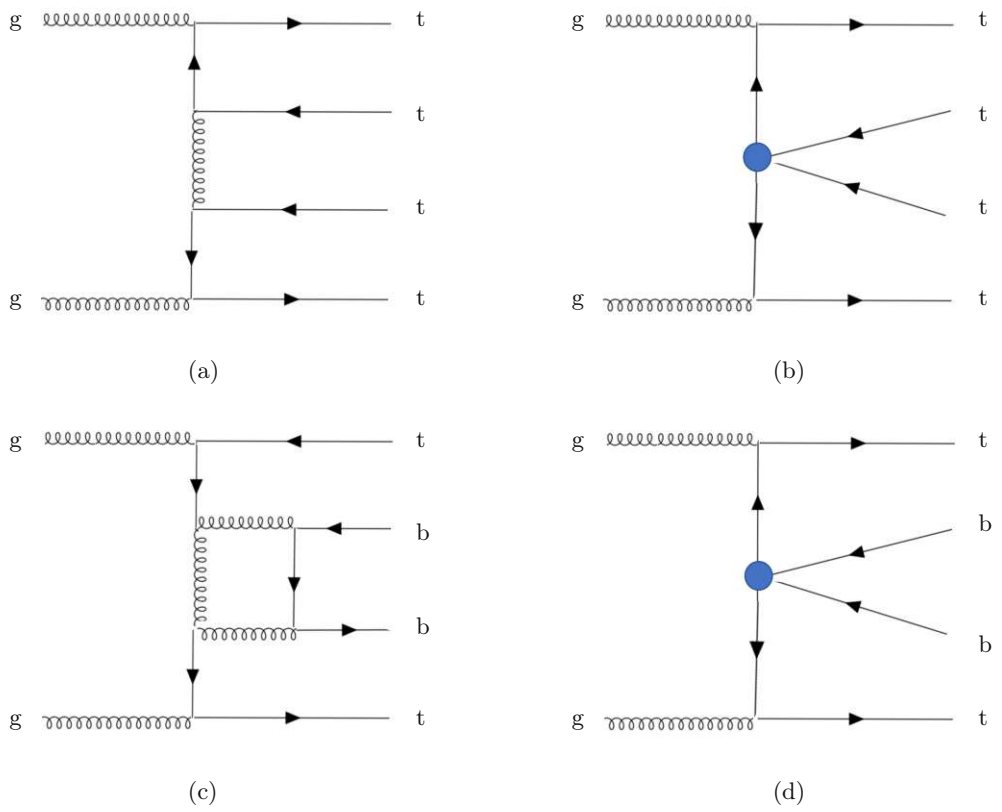


Figure 3.3: For $t\bar{t}t\bar{t}$ production, the SM interaction in (a) is replaced by an effective interaction in (b) in SMEFT. An equivalent Feynman diagram of the effective interaction can be drawn for $t\bar{t}b\bar{b}$ (d). The blue circle represents the insertion of one dimension-six operator which corresponds to a short-distance interaction between four heavy quarks at an energy scale Λ. The production for both $t\bar{t}t\bar{t}$ and $t\bar{t}b\bar{b}$ here is gluon induced.

As operators at mass dimension five only generate couplings that violate the baryon or lepton number, the first expansion of the SM-Lagrangian in terms of SMEFT at mass dimension 6 is [31]

$$\mathcal{L}_{\mathrm{SM}} = \mathcal{L}^{(4)}_{\mathrm{SM}} + \frac{1}{\Lambda^2}\sum_k C^{(6)}_k \mathcal{O}^{(6)}_k + O\left(\frac{1}{\Lambda^3}\right). \tag{3.2}$$

Following Ref. [31], we impose a $\mathrm{U}(2)_\mathrm{q} \times \mathrm{U}(2)_\mathrm{u} \times \mathrm{U}(2)_\mathrm{d}$ flavor symmetry in the light quark sector. Now, we consider all four-fermion operators that describe short-distance interactions between third generation quarks at the energy scale Λ. The operators are mainly taken from Ref. [31], with additional $t\bar{t}t\bar{t}$-specific operators from previous SMEFT studies on this process in Ref. [90, 105–107] and especially Ref. [30]. Since we train $\mathrm{Re}(\mathcal{O})$ and $\mathrm{Im}(\mathcal{O})$ for some operators, these are listed separately in the table. Additionally, we indicate with a tick if the operator affects $t\bar{t}t\bar{t}$, $t\bar{t}b\bar{b}$ or both.

No.	Operator	$t\bar{t}t\bar{t}$	$t\bar{t}b\bar{b}$
1	$\mathcal{O}^1_{QQ} = \frac{1}{2}\,(\bar{Q}\,\gamma_\mu\,Q)\,(\bar{Q}\,\gamma^\mu\,Q)$,	✓	✓
2	$\mathcal{O}^8_{QQ} = \frac{1}{2}\,(\bar{Q}\,\gamma_\mu\,T^A\,Q)\,(\bar{Q}\,\gamma^\mu\,T^A\,Q)$,	✓	✓
3	$\mathcal{O}^1_{tb} = (\bar{t}\,\gamma_\mu\,t)\,(\bar{b}\,\gamma_\mu\,b)$,		✓
4	$\mathcal{O}^8_{tb} = (\bar{t}\,\gamma_\mu\,T^A\,t)\,(\bar{b}\,\gamma^\mu\,T^A\,b)$,		✓
5	$\mathcal{O}^1_{tt} = (\bar{t}\,\gamma_\mu\,t)\,(\bar{t}\,\gamma_\mu\,t)$,	✓	
6	$\mathcal{O}^1_{bb} = (\bar{b}\,\gamma_\mu\,b)\,(\bar{b}\,\gamma_\mu\,b)$,		
7	$\mathcal{O}^1_{Qb} = (\bar{Q}\,\gamma_\mu\,Q)\,(\bar{b}\,\gamma_\mu\,b)$,		✓
8	$\mathcal{O}^8_{Qb} = (\bar{Q}\,\gamma_\mu\,T^A\,Q)\,(\bar{b}\,\gamma^\mu\,T^A\,b)$,		✓
9	$\mathcal{O}^1_{Qt} = (\bar{Q}\,\gamma_\mu\,Q)\,(\bar{t}\,\gamma_\mu\,t)$,	✓	✓
10	$\mathcal{O}^8_{Qt} = (\bar{Q}\,\gamma_\mu\,T^A\,Q)\,(\bar{t}\,\gamma^\mu\,T^A\,t)$,	✓	✓
11	$\mathrm{Re}(\mathcal{O}^1_{QtQb}) = \mathrm{Re}((\bar{Q}\,t)\,\epsilon\,(\bar{Q}\,b))$,		✓
12	$\mathrm{Im}(\mathcal{O}^1_{QtQb}) = \mathrm{Im}((\bar{Q}\,t)\,\epsilon\,(\bar{Q}\,b))$,		✓
13	$\mathrm{Re}(\mathcal{O}^8_{QtQb}) = \mathrm{Re}((\bar{Q}\,T^A\,t)\,\epsilon\,(\bar{Q}\,T^A\,b))$,		✓
14	$\mathrm{Im}(\mathcal{O}^8_{QtQb}) = \mathrm{Im}((\bar{Q}\,T^A\,t)\,\epsilon\,(\bar{Q}\,T^A\,b))$,		✓
15	$\mathrm{Re}(\mathcal{O}_{tH}) = \mathrm{Re}((H^\dagger H)\,(\bar{Q}\,\tilde{H}\,t))$	✓	✓
16	$\mathrm{Im}(\mathcal{O}_{tH}) = \mathrm{Im}((H^\dagger H)\,(\bar{Q}\,\tilde{H}\,t))$	✓	✓

Table 3.1: Collection of all EFT-operators in $t\bar{t}t\bar{t}$ and $t\bar{t}b\bar{b}$ probed in this thesis.

In the table above, Q denotes the left-handed SU(2) doublet of top and bottom quarks, t/b represents the right-handed top/bottom quark. T^A is the generator of SU(3) and denotes the totally antisymmetric Levi-Civitas tensor in SU(2) space [31]. We note that the operators $\mathcal{O}^1_{Qt}, \mathcal{O}^8_{Qt}, \mathcal{O}^1_{QQ}$ and $\mathcal{O}^8_{QQ}$ affect both $t\bar{t}t\bar{t}$ and $t\bar{t}b\bar{b}$.

When it comes to how the individual operators affect the cross section, this effect can be expressed by means of the corresponding Wilson coefficients C_i. Again following Ref. [31], we absorb the factor $1/\Lambda^2$ into C_i, and write the SMEFT cross section as

$$\sigma_{t\bar{t}b\bar{b}(t\bar{t}t\bar{t})} = \sigma^{\mathrm{SM}}_{t\bar{t}b\bar{b}(t\bar{t}t\bar{t})}\left(1 + \sum_i p^i_1 C_i + \sum_j \sum_{i\leq j} p^{ij}_2 C_i C_j.\right) \tag{3.3}$$

The coefficients of the fit to the cross section from the theory paper Ref. [31] can be found in Fig. 3.4. There, the EFT operators have been turned on one after the other:

$$\sigma_{\mathcal{O}_i} = \sigma^{\mathrm{SM}}\left(1 + p_1 C_i + p_2 C_i^2\right). \tag{3.4}$$

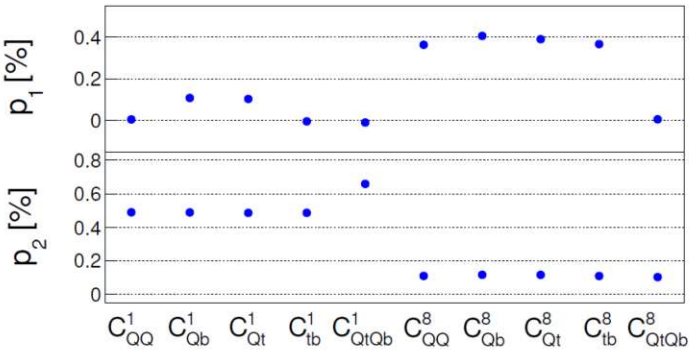


Figure 3.4: Coefficients of the fit to the cross section $\sigma = \sigma^{\mathrm{SM}}(1 + p_1 C_i + p_2 C_i^2)$ for the EFT operators affecting $t\bar{t}b\bar{b}$ turned on one by one. Figure taken from Ref. [31].

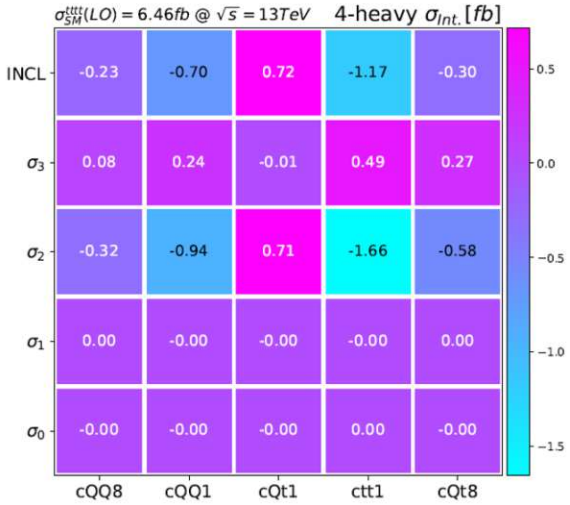


Figure 3.5: Interference strength of the EFT operators affecting $t\bar{t}t\bar{t}$. The top row shows the effect in the total inclusive prediction with all contributions from QCD and electro-weak force. The subsequent rows show the effects at different powers of α_s. Figure taken from Ref. [30].

Chapter 4

Learning the SMEFT with Simulation-based Inference

4.1 Simulation based inference

4.1.1 The optimal test statistic

With what has been discussed on SM and BSM physics so far, it is clear that the target of our analysis is to decide between two hypotheses $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ – the first being a BSM model, the latter the SM. From a statistical viewpoint, this can be done by computing the test statistic. At the heart of simulation-based (or likelihood-free) inference is the Neyman-Pearson lemma, that gives us the optimal test statistic for discriminating between two hypotheses $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$ [108, 109]. It states that the negative log-likelihood ratio

$$q_{\boldsymbol{\theta}}(\mathcal{D}) = -\log \frac{L(\mathcal{D}|\boldsymbol{\theta})}{L(\mathcal{D}|\boldsymbol{\theta}_0)} \tag{4.1}$$

is the most powerful test statistic available. Here, $\mathcal{D}$ denotes a data set of N events, each with a feature vector $\boldsymbol{x}_i$,

$$\mathcal{D} = \{\boldsymbol{x}_i\}_{i=1}^{N}. \tag{4.2}$$

The likelihood function $L(\mathcal{D}|\boldsymbol{\theta})$ therefore gives the probability to observe $\mathcal{D}$ under the hypothesis indicated by $\boldsymbol{\theta}$.

With the Neyman-Pearson lemma as a starting point, many efforts have been made to teach the BSM effects to the machine. Among these are, e.g., Ref. [15–29]. The following derivations use the just mentioned references. Especially, our general nomenclature and the limit setting procedure follow Ref. [24], while the derivation of the loss function is explicitly based on Ref. [15, 22, 27–29].

In our context of particle physics, it is possible to write the likelihood function as a product of the Poisson contribution

$$\mathrm{P}_{\mathcal{L}\sigma(\boldsymbol{\theta})}(N) = p(N|\boldsymbol{\theta}) \sim N, \tag{4.3}$$

i.e., the observation of N events, and a second contribution of the normalized probability density function, that event-wise reads [110]

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{1}{\sigma(\boldsymbol{\theta})} \frac{\mathrm{d}\sigma_{\boldsymbol{\theta}}(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}}. \tag{4.4}$$

In the two equations above, $\sigma(\boldsymbol{\theta})$ is the inclusive cross section and $\mathrm{d}\sigma_{\boldsymbol{\theta}}(\boldsymbol{x})/\mathrm{d}\boldsymbol{x}$ is the detector-level differential cross section. $\mathcal{L}$ denotes the integrated luminosity. Hence, we can write the likelihood

function $L(\mathcal{D}|\boldsymbol{\theta})$ as

$$L(\mathcal{D}|\boldsymbol{\theta}) = \mathrm{P}_{\mathcal{L}\sigma(\boldsymbol{\theta})}(N) \times \prod_{i=1}^{N} p(\boldsymbol{x}_i|\boldsymbol{\theta}) = \frac{e^{-\mathcal{L}\sigma(\boldsymbol{\theta})}}{N!} \times \prod_{i=1}^{N} \mathcal{L}\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}_i|\boldsymbol{\theta}) \tag{4.5}$$

We then inject this expression for the likelihood function in equation Eq. 4.1

$$\begin{aligned} q_{\boldsymbol{\theta}}(\mathcal{D}) &= -\log\left(\frac{e^{-\mathcal{L}\sigma(\boldsymbol{\theta})} \times \prod_{i=1}^{N} \mathcal{L}\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}_i|\boldsymbol{\theta})}{e^{-\mathcal{L}\sigma(\boldsymbol{\theta}_0)} \times \prod_{i=1}^{N} \mathcal{L}\sigma(\boldsymbol{\theta}_0) p(\boldsymbol{x}_i|\boldsymbol{\theta})}\right) \\ &= \mathcal{L}(\sigma(\boldsymbol{\theta}) - \sigma(\boldsymbol{\theta}_0)) - \sum_{i=1}^{N} \log R(\boldsymbol{x}_i|\boldsymbol{\theta}, \boldsymbol{\theta}_0) \end{aligned} \tag{4.6}$$

with

$$R(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \frac{\mathrm{d}\sigma_{\boldsymbol{\theta}}(\boldsymbol{x})/\mathrm{d}\boldsymbol{x}}{\mathrm{d}\sigma_{\boldsymbol{\theta}_0}(\boldsymbol{x})/\mathrm{d}\boldsymbol{x}} = \frac{\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta})}{\sigma(\boldsymbol{\theta}_0) p(\boldsymbol{x}|\boldsymbol{\theta}_0)}, \tag{4.7}$$

i.e., the differential cross section ration for the two hypotheses $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$.

As the first term in Eq. 4.6 – $\mathcal{L}(\sigma(\boldsymbol{\theta}) - \sigma(\boldsymbol{\theta}_0))$ – is independent from the feature vector $\boldsymbol{x}$, it can be retrieved from simulation or analytical calculation. What is therefore left to compute is the (negative logarithm) of the differential cross section ratio based on the feature vector $\boldsymbol{x}$. As the inclusive cross section is a known value for both SM and BSM processes, $R(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ is equivalent to the detector level likelihood ratio multiplied by a known factor

$$R(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\theta}_0) \propto \frac{p(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x}|\boldsymbol{\theta}_0)} = r(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\theta}_0). \tag{4.8}$$

Consequently, we only need to know the detector level likelihood ratio to have the most powerful test statistic at hand.

4.1.2 Learning EFT effects from simulation

With the SM-EFT extension of the SM Lagrangian $\mathcal{L}_{\mathrm{SM}}$ at hand [9, 31],

$$\mathcal{L}_{\mathrm{SM}} = \mathcal{L}_{\mathrm{SM}}^{(4)} + \frac{1}{\Lambda^2} \sum_k C_k^{(6)} \mathcal{O}_k^{(6)} + O\left(\frac{1}{\Lambda^3}\right), \tag{4.9}$$

we know that the insertion of a single EFT operator $\mathcal{O}_k$ influences the differential cross section at parton level, which we write by adopting the nomenclature from Ref. [24] as

$$\mathrm{d}\sigma(\boldsymbol{\theta}) \propto |\mathcal{M}_{\mathrm{SM}}(\boldsymbol{z}) + \theta_k \mathcal{M}_{\mathrm{BSM}}^k(\boldsymbol{z})|^2 \mathrm{d}\boldsymbol{z} \tag{4.10}$$

where $\boldsymbol{\theta}$ denotes the respective insertion of the Wilson Coefficient. Hence, we see from Eq. 4.10 that the two leading EFT contributions enter in the form of a second order polynomial, with the leading BSM effect being the interference term between SM and BSM amplitudes,

$$2\theta_k \mathrm{Re}(\mathcal{M}_{\mathrm{SM}}^*(\boldsymbol{z}) \mathcal{M}_{\mathrm{BSM}}^k(\boldsymbol{z})), \tag{4.11}$$

and the quadratic term containing the BSM effects only,

$$\theta_k^2 |\mathcal{M}_{\mathrm{BSM}}^k(\boldsymbol{z})|. \tag{4.12}$$

The polynomial dependence of both $\sigma(\boldsymbol{\theta})$ and $\mathrm{d}\sigma(\boldsymbol{\theta})/\mathrm{d}\boldsymbol{z}$ also enters in the parton level likelihood, that in analogy to Eq. 4.4 reads

$$p(\boldsymbol{z}_i|\boldsymbol{\theta}) = \frac{1}{\sigma(\boldsymbol{\theta})}\frac{\mathrm{d}\sigma_{\boldsymbol{\theta}}(\boldsymbol{z}_i)}{\mathrm{d}\boldsymbol{z}}. \tag{4.13}$$

When we then simulate at parton-level in perturbation theory, what we get is a sample of events with parton-level configurations $\boldsymbol{z}_i$ distributed according to $p(\boldsymbol{z}_i|\boldsymbol{\theta}_{\mathrm{ref}})$, where $\boldsymbol{\theta}_{\mathrm{ref}}$ might or might not be the SM point in the $\boldsymbol{\theta}$-space.

Given the polynomical structure of $\sigma(\boldsymbol{\theta})$ and $\mathrm{d}\sigma(\boldsymbol{\theta})/\mathrm{d}\boldsymbol{z}$, once the $\boldsymbol{\theta}$-independent terms $\mathcal{M}_{\mathrm{SM}}(\boldsymbol{z})$ and $\mathcal{M}^k_{\mathrm{BSM}}(\boldsymbol{z})$ are known for a specific set of $\boldsymbol{z}_i$, we know the $\mathrm{d}\sigma(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$. In fact, the polynomial structure allows to infer the information on the full $\boldsymbol{\theta}$-dependence of every $\boldsymbol{z}$ when a sufficient number of linearly independent $\boldsymbol{\theta}$ values is at hand. With this procedure, we can obtain analytic expressions for $w_i(\boldsymbol{\theta})$, i.e., the per-event weight function describing the $\boldsymbol{\theta}$-dependence over the whole range of $\boldsymbol{\theta}$ values [27].

Next, we scale the sample to the expected luminosity $\mathcal{L}$ and approximate the differential cross section at parton-level in a small section $\Delta\boldsymbol{z}$ of the phase space

$$\int_{\Delta\boldsymbol{z}} \frac{\mathrm{d}\sigma_{\boldsymbol{\theta}}(\boldsymbol{z})}{\mathrm{d}\boldsymbol{z}}\mathrm{d}\boldsymbol{z} \approx \frac{\mathrm{d}\sigma_{\boldsymbol{\theta}}(\boldsymbol{z})}{\mathrm{d}\boldsymbol{z}}\Delta\boldsymbol{z} \approx \frac{1}{\mathcal{L}}\sum_{\boldsymbol{z}_i\in\Delta\boldsymbol{z}} w_i(\boldsymbol{\theta}), \tag{4.14}$$

which approximately holds if the differential cross section in $\Delta\boldsymbol{z}$ does not vary excessively and we consider a large number of events [24]. Integrating over all $\Delta\boldsymbol{z}$ allows us to tie the weights to the generator-level likelihood $p(\boldsymbol{z}_i|\boldsymbol{\theta})$:

$$\sum_{i=1}^{N_{\mathrm{events}}} w_i = \mathcal{L}\sigma(\boldsymbol{\theta}). \tag{4.15}$$

How can we exploit this relations when it comes to our simulated samples? First, we consider the sampling of events with probability $p(\boldsymbol{x}|\boldsymbol{\theta})$. We can factorise $p(\boldsymbol{x}|\boldsymbol{\theta})$ as $p(\boldsymbol{x}, \boldsymbol{z}_{\mathrm{d}}, \boldsymbol{z}_{\mathrm{s}}, \boldsymbol{z}_{\mathrm{p}}|\boldsymbol{\theta})$ and integrate out the latent space in Eq. 4.16 below. Here, $\boldsymbol{x}$ denotes the per-event features after full simulation in forward mode, $\boldsymbol{z}_{\mathrm{d}}$ denotes the feature vector at detector level, $\boldsymbol{z}_{\mathrm{s}}$ at shower level and $\boldsymbol{z}_{\mathrm{p}}$ at parton level. The true likelihood at detector, shower and parton level is intractable

$$\begin{aligned} p(\boldsymbol{x}|\boldsymbol{\theta}) &= \int \mathrm{d}\boldsymbol{z}_{\mathrm{d}}\mathrm{d}\boldsymbol{z}_{\mathrm{s}}\mathrm{d}\boldsymbol{z}_{\mathrm{p}} p(\boldsymbol{x}, \boldsymbol{z}_{\mathrm{d}}, \boldsymbol{z}_{\mathrm{s}}, \boldsymbol{z}_{\mathrm{p}}|\boldsymbol{\theta}) \\ &= \int \mathrm{d}\boldsymbol{z}_{\mathrm{d}}\mathrm{d}\boldsymbol{z}_{\mathrm{s}}\mathrm{d}\boldsymbol{z}_{\mathrm{p}} p(\boldsymbol{x}|\boldsymbol{z}_{\mathrm{d}})p(\boldsymbol{z}_{\mathrm{d}}|\boldsymbol{z}_{\mathrm{s}})p(\boldsymbol{z}_{\mathrm{s}}|\boldsymbol{z}_{\mathrm{p}})p(\boldsymbol{z}_{\mathrm{p}}|\boldsymbol{\theta}). \end{aligned} \tag{4.16}$$

The huge latent space spanned by the integrals over $\mathrm{d}\boldsymbol{z}_d$, $\mathrm{d}\boldsymbol{z}_p$ and $\mathrm{d}\boldsymbol{z}_s$ is the true crux of the EFT measurement problem, as it can involve millions of random numbers. As simulation is always run in forward mode, this immense latent space makes it impossible to trace back the likelihood function from a set of observables $\boldsymbol{x}$ to their parton-level configuration $\boldsymbol{z}$. An evaluation of the theory parameters given a certain observation is therefore impossible [15, 22, 27–29].

However, in the differential cross section ratio, we can exploit the simplicity of the joint space, as all intractable factors cancel[1]:

$$r(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \frac{p(\boldsymbol{x}, \boldsymbol{z}_{\mathrm{d}}, \boldsymbol{z}_{\mathrm{s}}, \boldsymbol{z}_{\mathrm{p}}|\boldsymbol{\theta})}{p(\boldsymbol{x}, \boldsymbol{z}_{\mathrm{d}}, \boldsymbol{z}_{\mathrm{s}}, \boldsymbol{z}_{\mathrm{p}}|\boldsymbol{\theta}_0)} = \frac{p(\boldsymbol{x}|\boldsymbol{z}_{\mathrm{d}})p(\boldsymbol{z}_{\mathrm{d}}|\boldsymbol{z}_{\mathrm{s}})p(\boldsymbol{z}_{\mathrm{s}}|\boldsymbol{z}_{\mathrm{p}})p(\boldsymbol{z}_{\mathrm{p}}|\boldsymbol{\theta})}{p(\boldsymbol{x}|\boldsymbol{z}_{\mathrm{d}})p(\boldsymbol{z}_{\mathrm{d}}|\boldsymbol{z}_{\mathrm{s}})p(\boldsymbol{z}_{\mathrm{s}}|\boldsymbol{z}_{\mathrm{p}})p(\boldsymbol{z}_{\mathrm{p}}|\boldsymbol{\theta}_0)} = \frac{p(\boldsymbol{z}_{\mathrm{p}}|\boldsymbol{\theta})}{p(\boldsymbol{z}_{\mathrm{p}}|\boldsymbol{\theta}_0)} \tag{4.17}$$

We can relate this expression to the per-event weights $w_i(\boldsymbol{\theta})$ by combining Eq. 4.13 and 4.14 to

$$\frac{w_i(\boldsymbol{\theta})}{\mathcal{L}\sigma(\boldsymbol{\theta})} = \frac{1}{\sigma(\boldsymbol{\theta})}\frac{\mathrm{d}\sigma_{\boldsymbol{\theta}}(\boldsymbol{z})}{\mathrm{d}\boldsymbol{z}}\bigg|_{\boldsymbol{z}=\boldsymbol{z}_i} = p(\boldsymbol{z}_i|\boldsymbol{\theta}) \tag{4.18}$$

[1] This is not true for all cases. Some EFT insertions also affect the parton shower [24].

and injecting this into Eq. 4.17

$$r(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \frac{p(\boldsymbol{z}_\mathrm{i}|\boldsymbol{\theta})}{p(\boldsymbol{z}_\mathrm{i}|\boldsymbol{\theta}_0)} = \frac{\sigma(\boldsymbol{\theta}_0)}{\sigma(\boldsymbol{\theta})} \frac{w_i(\boldsymbol{\theta})}{w_i(\boldsymbol{\theta}_0)}. \tag{4.19}$$

It must be emphasized that this does not in any way allow to evaluate the likelihood $p(\boldsymbol{x}|\boldsymbol{\theta})$ in the cross section ratio $R(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\theta}_0)$, as the integral over the latent space is still intractable. A collection of all relevant quantities is given in Tab. 4.1, following the color code of Ref. [27].

Quantity		tractable?
$p(\boldsymbol{x}\|\boldsymbol{\theta})$	$= \int \mathrm{d}\boldsymbol{z}\ p(\boldsymbol{x}\|\boldsymbol{z})p(\boldsymbol{z}\|\boldsymbol{\theta})$	No
$p(\boldsymbol{x}_i\|\boldsymbol{\theta})$	$= \frac{1}{\sigma(\boldsymbol{\theta})} \frac{\mathrm{d}\sigma_{\boldsymbol{\theta}}(\boldsymbol{x}_i)}{\mathrm{d}\boldsymbol{x}}$	No
$R(\boldsymbol{x}\|\boldsymbol{\theta}, \boldsymbol{\theta}_0)$	$= \frac{\sigma(\boldsymbol{\theta})p(\boldsymbol{x}\|\boldsymbol{\theta})}{\sigma(\boldsymbol{\theta}_0)p(\boldsymbol{x}\|\boldsymbol{\theta}_0)}$	No
$r(\boldsymbol{x}\|\boldsymbol{\theta}, \boldsymbol{\theta}_0)$	$= \frac{p(\boldsymbol{x}\|\boldsymbol{\theta})}{p(\boldsymbol{x}\|\boldsymbol{\theta}_0)}$	No
$p(\boldsymbol{z}_i\|\boldsymbol{\theta})$	$= \frac{1}{\sigma(\boldsymbol{\theta})} \frac{\mathrm{d}\sigma_{\boldsymbol{\theta}}(\boldsymbol{z}_i)}{\mathrm{d}\boldsymbol{z}}$	Yes
$R(\boldsymbol{x}, \boldsymbol{z}\|\boldsymbol{\theta}, \boldsymbol{\theta}_0)$	$= \frac{\sigma(\boldsymbol{\theta})p(\boldsymbol{z}\|\boldsymbol{\theta})}{\sigma(\boldsymbol{\theta}_0)p(\boldsymbol{z}\|\boldsymbol{\theta}_0)}$	Yes
$r(\boldsymbol{x}, \boldsymbol{z}\|\boldsymbol{\theta}, \boldsymbol{\theta}_0)$	$= \frac{p(\boldsymbol{z}\|\boldsymbol{\theta})}{p(\boldsymbol{z}\|\boldsymbol{\theta}_0)}$	Yes
$r(\boldsymbol{x}_i, \boldsymbol{z}_i\|\boldsymbol{\theta}, \boldsymbol{\theta}_0)$	$= \frac{p(\boldsymbol{z}_i\|\boldsymbol{\theta})}{p(\boldsymbol{z}_i\|\boldsymbol{\theta}_0)} = \frac{\sigma(\boldsymbol{\theta}_0)}{\sigma(\boldsymbol{\theta})} \frac{w_i(\boldsymbol{\theta})}{w_i(\boldsymbol{\theta}_0)}$	Yes

Table 4.1: Tractable and intractable quantities. Following Ref. [27], we mark intactable quantities red. Consequently, tractable quantities are marked blue.

How can we now make our network learn the optimal test statistic $q_{\boldsymbol{\theta}}(\mathcal{D})$ according to the Neyman-Pearson lemma Eq. 4.1, when the quantity $p(\boldsymbol{x}|\boldsymbol{\theta})$ and therefore $R(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ are intractable? According to Ref. [15, 22, 27–29], the trick now consists in using the L^2 squared loss functional for functions $\hat{g}(\boldsymbol{x})$ that only depend on $\boldsymbol{x}$ but try to approximate the target $g(\boldsymbol{x}, \boldsymbol{z})$,

$$L[\hat{g}(\boldsymbol{x})] = \int \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z}\ p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})\ |g(\boldsymbol{x}, \boldsymbol{z}) - \hat{g}(\boldsymbol{x})|^2 \tag{4.20}$$

For variation calculus for $\hat{g}(\boldsymbol{x})$ based on the derivation in Ref. [27], we write the integral as

$$\begin{aligned} &\int \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{z}\ p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})\ |g(\boldsymbol{x}, \boldsymbol{z}) - \hat{g}(\boldsymbol{x})|^2 = \\ &\int \mathrm{d}\boldsymbol{x} \underbrace{\left(\hat{g}^2(\boldsymbol{x}) \int \mathrm{d}\boldsymbol{z}\ p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) - 2\hat{g}(\boldsymbol{x}) \int \mathrm{d}\boldsymbol{z}\ p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})\ g(\boldsymbol{x}, \boldsymbol{z}) + \int \mathrm{d}\boldsymbol{z}\ p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) g^2(\boldsymbol{x}, \boldsymbol{z}) \right)}_{:= F(\boldsymbol{x})} \end{aligned} \tag{4.21}$$

and search for the function $g^*(\boldsymbol{x})$ so that $L[\hat{g}(\boldsymbol{x})]$ is extreme

$$\left. \frac{\delta F}{\delta \hat{g}} \right|_{\hat{g}=g^*} = 0. \tag{4.22}$$

This condition yields

$$0 = \left(-2\hat{g} \int \mathrm{d}\boldsymbol{z}\ p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) + 2 \int \mathrm{d}\boldsymbol{z}\ p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) g(\boldsymbol{x}, \boldsymbol{z}) \right) \Bigg|_{\hat{g}=g^*}. \tag{4.23}$$

After calculating the first integral (which evaluates to $p(\boldsymbol{x}|\boldsymbol{\theta})$) we solve for g^*

$$g^* = \frac{1}{p(\boldsymbol{x}|\boldsymbol{\theta})} \int \mathrm{d}\boldsymbol{z}\ p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) g(\boldsymbol{x}, \boldsymbol{z}). \tag{4.24}$$

We can now substitute the target with the tractable (!) function $r(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}_0, \boldsymbol{\theta})$, which is the coefficient of two intractable functions $p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}_0)$ and $p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})$

$$\begin{aligned} g^* &= \frac{1}{p(\boldsymbol{x}|\boldsymbol{\theta})} \int \mathrm{d}\boldsymbol{z}\; p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) r(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}_0, \boldsymbol{\theta}) \\ &= \frac{1}{p(\boldsymbol{x}|\boldsymbol{\theta})} \int \mathrm{d}\boldsymbol{z}\; \cancel{p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})} \frac{p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}_0)}{\cancel{p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})}} \\ &= \frac{p(\boldsymbol{x}|\boldsymbol{\theta}_0)}{p(\boldsymbol{x}|\boldsymbol{\theta})} = r(\boldsymbol{x}|\boldsymbol{\theta}_0, \boldsymbol{\theta}). \end{aligned} \tag{4.25}$$

Hence, we conclude that by minimizing the squared loss of $r(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}_0, \boldsymbol{\theta})$,

$$L[\hat{r}(\boldsymbol{x}|\boldsymbol{\theta}_0, \boldsymbol{\theta})] = \sum_{(\boldsymbol{x}_i, \boldsymbol{z}_i) \sim p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})}^{N_{\text{events}}} p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) |r(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}_0, \boldsymbol{\theta}) - \hat{r}(\boldsymbol{x}|\boldsymbol{\theta}_0, \boldsymbol{\theta})|^2, \tag{4.26}$$

we can regress on the true likelihood ratio!

4.1.3 Learning the polynomial dependence

Next, we exploit that Eq. 4.10 implies a polynomial dependence of the likelihood ratio at detector level $R(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ because of the polynomial dependence of $\mathrm{d}\sigma(\boldsymbol{\theta})$, even if the detector-level likelihood itself is intractable. By slightly adapting what has been done in Ref. [24] to our use case, we write

$$\begin{aligned} R(\boldsymbol{x}|\boldsymbol{\theta}, \boldsymbol{\theta}_0) &= \frac{\mathrm{d}\sigma_{\boldsymbol{\theta}}(\boldsymbol{x})/\mathrm{d}\boldsymbol{x}}{\mathrm{d}\sigma_{\boldsymbol{\theta}_0}(\boldsymbol{x})/\mathrm{d}\boldsymbol{x}} = \frac{\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta})}{\sigma(\boldsymbol{\theta}_0) p(\boldsymbol{x}|\boldsymbol{\theta}_0)} \\ &= 1 + (\theta - \theta_0) \frac{\partial(\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta}))}{\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta})} + (\theta - \theta_0)^2 \frac{\partial^2(\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta}))}{\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta})} \\ &= 1 + (\theta - \theta_0) R_{\text{lin}}(\boldsymbol{x}) + (\theta - \theta_0)^2 R_{\text{quad}}(\boldsymbol{x}). \end{aligned} \tag{4.27}$$

When choosing an equivalent ansatz for the tractable joint likelihood ratio $R(\boldsymbol{x}, \boldsymbol{z}, |\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ as the ansatz for our training target,

$$R(\boldsymbol{x}, \boldsymbol{z}, |\boldsymbol{\theta}, \boldsymbol{\theta}_0) = 1 + (\theta - \theta_0) \frac{\partial(\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}))}{\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})} + (\theta - \theta_0)^2 \frac{\partial^2(\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}))}{\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})} \tag{4.28}$$

we find with Eq. 4.25[2]

$$\begin{aligned} g^* &= \frac{1}{p(\boldsymbol{x}|\boldsymbol{\theta})} \int \mathrm{d}\boldsymbol{z}\; p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) R(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}_0, \boldsymbol{\theta}) \\ &= \frac{1}{p(\boldsymbol{x}|\boldsymbol{\theta})} \int \mathrm{d}\boldsymbol{z}\; p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}) \left(1 + (\theta - \theta_0) \frac{\partial(\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}))}{\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})} + (\theta - \theta_0)^2 \frac{\partial^2(\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta}))}{\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{\theta})}\right) \\ &= \frac{1}{p(\boldsymbol{x}|\boldsymbol{\theta})} \left(p(\boldsymbol{x}|\boldsymbol{\theta}) + (\theta - \theta_0) \frac{\partial(\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta}))}{\sigma(\boldsymbol{\theta})} + (\theta - \theta_0)^2 \frac{\partial^2(\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta}))}{\sigma(\boldsymbol{\theta})}\right) \\ &= 1 + (\theta - \theta_0) \underbrace{\frac{\partial(\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta}))}{\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta})}}_{R_{\text{lin}}(\boldsymbol{x})} + (\theta - \theta_0)^2 \underbrace{\frac{\partial^2(\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta}))}{\sigma(\boldsymbol{\theta}) p(\boldsymbol{x}|\boldsymbol{\theta})}}_{R_{\text{quad}}(\boldsymbol{x})} \end{aligned} \tag{4.29}$$

[2] As $R(\boldsymbol{x}, \boldsymbol{z}, |\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ and $r(\boldsymbol{x}, \boldsymbol{z}, |\boldsymbol{\theta}, \boldsymbol{\theta}_0)$ are proportional, they can be used interchangably.

Hence, we regress exactly in the coefficient functions of the intractable likelihood ratio $R_{\text{lin}}(\boldsymbol{x})$ and $R_{\text{quad}}(\boldsymbol{x})$.

In the last step, we identify the regression targets in Eq. 4.29 with our weight functions. The weights around the value $\boldsymbol{\theta}_0$ can be written event-wise as

$$w_i(\boldsymbol{\theta}) = w_{i,0} + (\theta - \theta_0) w_{i,1} + (\theta - \theta_0)^2 w_{i,2}, \tag{4.30}$$

where the coefficient $w_{i,0}$ corresponds to $w(\boldsymbol{\theta}_0)$, i.e., in our case the SM weight. The coefficients $w_{i,1}$ and $w_{i,2}$ are the BSM-operator specific weights and known from the event-simulation (see chapter 6). Hence, we set the $\boldsymbol{\theta}$-value at the SM point to zero and reformulate the regression target

$$R(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta}) = \frac{\sigma(\boldsymbol{\theta})}{\sigma(SM)} r(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta}, \boldsymbol{\theta}_0) = \frac{w_i(\boldsymbol{\theta})}{w_i(\boldsymbol{\theta}_0)} = 1 + \theta \frac{w_{i,1}}{w_{i,0}} + \theta^2 \frac{w_{i,2}}{w_{i,0}}. \tag{4.31}$$

Finally, we inject the polynomial ansatz of Eq. 4.27 in the loss function of Eq. 4.26, which then reads

$$\begin{aligned} L &= \sum_{\boldsymbol{\theta} \in \mathcal{B}} \sum_i w_{i,0} (R(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta}) - \hat{R}(\boldsymbol{x}|\boldsymbol{\theta}))^2 \\ &= \sum_{\boldsymbol{\theta} \in \mathcal{B}} \sum_i w_{i,0} (\cancel{1} + \theta \frac{w_{i,1}}{w_{i,0}} + \theta^2 \frac{w_{i,2}}{w_{i,0}} - \cancel{1} - \theta R_{\text{lin}}(\boldsymbol{x}_i) - \theta^2 R_{\text{quad}}(\boldsymbol{x}_i))^2 \\ &= \sum_{\boldsymbol{\theta} \in \mathcal{B}} \sum_i w_{i,0} \left[\theta \left(\frac{w_{i,1}}{w_{i,0}} - R_{\text{lin}}(\boldsymbol{x}_i) \right) + \theta^2 \left(\frac{w_{i,2}}{w_{i,0}} - R_{\text{quad}}(\boldsymbol{x}_i) \right) \right]^2 \end{aligned} \tag{4.32}$$

Here, $\mathcal{B}$ contains two different, arbitrarily chosen base points for $\boldsymbol{\theta} \neq \boldsymbol{\theta_0}$. As we fit a quadratic polynomial, for the loss function to be expressive enough $\mathcal{B}$ includes two values for $\boldsymbol{\theta} \neq 0$ in addition to the SM value $\boldsymbol{\theta}_0 \neq 0$.

4.2 Limit setting in 1D and 2D

For evaluating the network's performance once the training has been successfully concluded, we first check the convergence in bins of the scalar event-level observables. The simple condition

$$\begin{aligned} \sum_{\boldsymbol{x}_i \in \text{bin}} w_{0,i} R_{\text{lin}}(\boldsymbol{x}_i) &\stackrel{?}{=} \sum_{\boldsymbol{x}_i \in \text{bin}} w_{1,i} \\ \sum_{\boldsymbol{x}_i \in \text{bin}} w_{0,i} R_{\text{quad}}(\boldsymbol{x}_i) &\stackrel{?}{=} \sum_{\boldsymbol{x}_i \in \text{bin}} w_{2,i} \end{aligned} \tag{4.33}$$

holds for a converged training. In fact, when minimizing L in Eq. 4.32, the expression in the brackets vanishes for

$$\begin{aligned} R_{\text{lin}}(\boldsymbol{x}_i) &\to \frac{w_{i,1}}{w_{i,0}} \\ R_{\text{quad}}(\boldsymbol{x}_i) &\to \frac{w_{i,2}}{w_{i,0}}. \end{aligned} \tag{4.34}$$

Hence, we can see the convergence in bins of the scalar event-level features we feed the network as $\boldsymbol{x}$-values.

However, this is only a first check in the training process. To effectively set limits, we need to perform hypothesis tests [110]. From the Neyman-Pearson lemma (Eq. 4.1) we know that in the absence of nuisances and for an unbinned likelihood ratio test, the optimal test statistic is

$$q_{\boldsymbol{\theta}}(\mathcal{D}) = \underbrace{\mathcal{L}(\sigma_{\boldsymbol{\theta}} - \sigma_{\text{SM}})}_{=\text{const}} - \sum_{i=1}^{N} \log R(\boldsymbol{x}_i|\boldsymbol{\theta}, \text{SM}) \tag{4.35}$$

where $R(\boldsymbol{x}_i|\boldsymbol{\theta}, \text{SM})$ is the differential cross section ratio at detector-level. With simulation-based inference, we estimated $R(\boldsymbol{x}_i|\boldsymbol{\theta}, \text{SM})$ as

$$R(\boldsymbol{x}_i|\boldsymbol{\theta}, \text{SM}) \approx \hat{R}(\boldsymbol{x}_i|\boldsymbol{\theta}, \text{SM}) = 1 + \theta R_{\text{lin}}(\boldsymbol{x}_i) + \theta^2 R_{\text{quad}}(\boldsymbol{x}_i), \tag{4.36}$$

with $R_{\text{lin}}(\boldsymbol{x}_i)$ and $R_{\text{quad}}(\boldsymbol{x}_i)$ being the output of our neural network. For every hypothesis test, we then define the test statistic $t_{\boldsymbol{\theta}}(\boldsymbol{x})$ as

$$t_{\boldsymbol{\theta}}(\boldsymbol{x}_i) \equiv \hat{R}(\boldsymbol{x}_i|\boldsymbol{\theta}, \text{SM}). \tag{4.37}$$

Next, we want to make the analysis a binned analysis, as has been done similarly in Ref. [24]. Here, the problem is that the range of values of $t_{\boldsymbol{\theta}}(\boldsymbol{x})$ depends parametrically on $\boldsymbol{\theta}$. In other words, we cannot define a uniform binning over the whole range of $\boldsymbol{\theta}$-values because there will always be a $\boldsymbol{\theta}$ either large or small where all events are concentrated in the first or the last bin. Hence, we need to dynamically adjust the binning choice for each individual value of $\boldsymbol{\theta}$.

First, we make a finely binned histogram of $p(t_{\boldsymbol{\theta}}(\boldsymbol{x})|\text{SM})$, i.e., we obtain the probability density function of the test statistic for a value $\boldsymbol{\theta}$ under the SM hypothesis. In practical terms, this means filling the histogram of $t_{\boldsymbol{\theta}}(\boldsymbol{x})$ with the SM weights. Next, we make this distribution a binned, flat distribution under the SM hypothesis by computing the weighted quantiles of $p(t_{\boldsymbol{\theta}}(\boldsymbol{x})|\text{SM})$ in steps of 10%. With this new binning choice, $p(t_{\boldsymbol{\theta}}(\boldsymbol{x})|\text{SM})$ becomes flat by construction with the yield in each bin, N^i_{events}, being

$$\sum_{i=1}^{n_{\text{bins}}} N^i_{\text{events}} = \sum_{i=1}^{n_{\text{bins}}} \frac{\mathcal{L}\sigma_{\text{SM}}}{n_{\text{bins}}} = \mathcal{L}\sigma_{\text{SM}}, \qquad n_{\text{bins}} = 10 \tag{4.38}$$

Then, we fill the bins with the test statistic evaluated at the BSM-point $\boldsymbol{\theta}$, i.e., $p(t_{\boldsymbol{\theta}}(\boldsymbol{x})|\boldsymbol{\theta})$. Now, the last bin contains those 10% of events that changes the most with respect to $\boldsymbol{\theta}$ – which is exactly what we wanted to achieve in first place.

At this point, it is important to distinguish between two "modes" of evaluation: Keep the full information learned in the training, or rule out the BSM-related changes that alter the overall yield. The first consists in weighting the test statistic with the full BSM weights and normalize the SM weights (and the SM term in the BSM weights) so that the overall yield corresponds to the integrated luminosity

$$\begin{aligned} w_{i,\text{SM}} &= \frac{\mathcal{L}\sigma_{\text{SM}} w_{i,0}}{\sum_j w_{j,0}} \\ w_{i,\text{BSM}} &= \frac{\mathcal{L}\sigma_{\text{SM}} \left(w_{i,0} + \theta w_{i,1} + \theta^2 w_{i,2}\right)}{\sum_j w_{j,0}} \end{aligned} \tag{4.39}$$

This is referred to as "full information" in the following parts of the thesis. The second "mode" of evaluation is referred to as "shape effects only" and consists in normalizing both SM and BSM yields to the same expected number of events according to the integrated luminosity.

$$\begin{aligned} w_{i,\text{SM}} &= \frac{\mathcal{L}\sigma_{\text{SM}} w_{i,0}}{\sum_j w_{j,0}} \\ w_{i,\text{BSM}} &= \frac{\mathcal{L}\sigma_{\text{SM}} \left(w_{i,0} + \theta w_{i,1} + \theta^2 w_{i,2}\right)}{\sum_j \left(w_{j,0} + \theta w_{j,1} + \theta^2 w_{j,2}\right)} \end{aligned} \tag{4.40}$$

With this weighting choice, we are sensitive only to those effects that the network is able to extract from the BSM shape information.

Finally, we compute the expected likelihood ratio by summing over the Poisson bins for SM and BSM pdf [110]:

$$q_{\boldsymbol{\theta},\text{binned}}(\mathcal{D}) = \sum_{i=1}^{n_\text{bins}} \left(\lambda_i(\boldsymbol{\theta}) - \lambda_i(\text{SM}) - n_i \log \frac{\lambda_i(\boldsymbol{\theta})}{\lambda_i(\text{SM})} \right) \tag{4.41}$$

where λ_i is the prediction in the respective bin and n_i corresponds to the observation which we replace with the SM prediction $\lambda_i(\text{SM})$. When we then multiply the expression in Eq. 4.41 with -2, we can evaluate where 1σ and 2σ with respect to the value of $\boldsymbol{\theta}$ according to Tab. 4.2 below.

$(1-\alpha)(\%)$	$m=1$	$m=2$
68.27	1.00	2.30
90.00	2.71	4.61
95.00	3.84	5.99
95.45	4.00	6.18
99.00	6.63	9.21
99.73	9.00	11.83

Table 4.2: Values of $2\Delta\ln L$ corresponding to the coverage probability $1-\alpha$ in % in the limit of large sample data. The value of m indicates the number of simultaneously estimated parameters. Values taken from chapter 40: *Statistics* in Ref. [110].

Lastly, we want to constrain two operators $\mathcal{O}_1$ and $\mathcal{O}_2$ simultaneously based on our 1D training. Hence, we write the combined test statistic $t^{2\text{D}}_{\boldsymbol{\theta},\boldsymbol{\theta}'}$ for $\mathcal{B}$ containing the values $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ for the two trained coefficients as

$$t^{2\text{D}}_{\boldsymbol{\theta},\boldsymbol{\theta}'}(\boldsymbol{x}) = R^{2\text{D}}_{\mathcal{O}_1,\mathcal{O}_2}(\boldsymbol{x}|\boldsymbol{\theta},\boldsymbol{\theta}',\text{SM}) = 1 + \sum_{i\in\mathcal{B}} \theta_i R^{(i)}_\text{lin}(\boldsymbol{x}) + \sum_{j\in\mathcal{B}} \sum_{i\leq j} \theta_i\theta_j R^{(i)}_\text{quad}(\boldsymbol{x}) R^{(j)}_\text{quad}(\boldsymbol{x}) \tag{4.42}$$

which we approximate with our training results for one operator $R_{\mathcal{O}_1}(\boldsymbol{x}_i|\boldsymbol{\theta},\text{SM})$ and $R_{\mathcal{O}_2}(\boldsymbol{x}_i|\boldsymbol{\theta}',\text{SM})$ as

$$t^{2\text{D}}_{\boldsymbol{\theta},\boldsymbol{\theta}'}(\boldsymbol{x}) \approx 1 + \theta R^{(\mathcal{O}_1)}_\text{lin} + \theta' R^{(\mathcal{O}_2)}_\text{lin} + \theta^2 R^{(\mathcal{O}_1)}_\text{quad} + \theta'^2 R^{(\mathcal{O}_2)}_\text{quad}. \tag{4.43}$$

In the 2D evaluation, the BSM weights need to be changed according to

$$w_{i,\text{BSM}} = w_{i,0} + \theta w^{(\mathcal{O}_1)}_{i,1} + \theta' w^{(\mathcal{O}_2)}_{i,1} + \theta^2 w^{(\mathcal{O}_1)}_{i,2} + \theta'^2 w^{(\mathcal{O}_2)}_{i,2} + 2\theta\theta' w^{(\mathcal{O}_1,\mathcal{O}_2)}_{i,12}, \tag{4.44}$$

before normalizing again to the same integrated luminosity as the SM. When setting the limits, we need to choose $m=2$ in Tab. 4.2. With the theoretical background at hand, we now move to the network architecture and the details of technical implementation.

Chapter 5

Multivariate Analysis

5.1 Basic Priniciples of MVA

As the BSM physic causes various effects in the observed quantities, Multivariate Analysis (MVA) provides us with a powerful tool to simultaneously investigate these effects while tracing them back to the two desired numbers, i.e., the linear and the quadratic dependence of the prediction. In this respect, a broad variety of existing algorithms to optimally extract the information from the input features is at hand. Hence, we first describe the components within our neural network's architecture. As it is almost identical for simulation-based inference and multi-classification, we will discuss both training setups simultaneously if not explicitly stated ad locum.

5.1.1 Deep Neural Networks (DNNs)

The first component of choice – a Deep Neural Network of multiple stacked dense layers – follows naturally from the structure of our data [10]. In fact, for every simulated event we can retrieve a large number of scalar event-level observables, i.e., our feature vector $\boldsymbol{x}$. Hence, we start by setting up a DNN of dense layers that takes $\boldsymbol{x}$ as input and passes it through a number of subsequent hidden layers. The working principle of a **dense − or fully-connected − layer** is encoded in the linear equation

$$\boldsymbol{y} = \mathbf{W}\boldsymbol{x} + \boldsymbol{b}, \tag{5.1}$$

where the feature vector $\boldsymbol{x}$ is multiplied with the layer's weights $\mathbf{W}$ to get the layer's output $\boldsymbol{y}$. $\boldsymbol{b}$ denotes an eventual bias applied in the training [111]. For every linear (dense, fully connected) layer, the input shape is determined by the length of the input feature vector $\boldsymbol{x}$, whereas the shape of the output is a free hyperparameter and therefore subject to optimization. To give the neural network suitable degrees of freedom, our architecture comprises a sequence of multiple fully connected layers. As depicted in Fig. 5.1, every neuron on the input side is connected to every neuron at the output side.

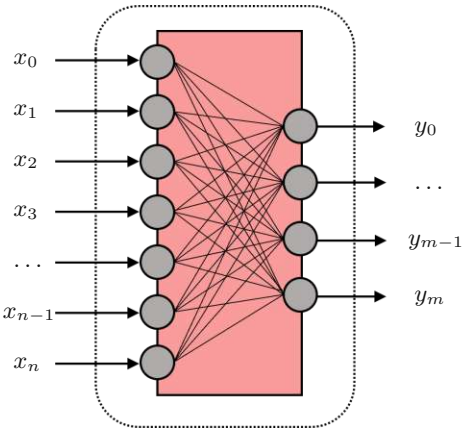


Figure 5.1: Fully connected linear layer: Every input is connected to every output.

This fully connected setup is not the desired configuration, as we want to bring in non-linearity in the training for further degrees of freedom. In this respect, the tools of choice are activation functions that determine whether a neuron fires or not. In our case, we choose the **Rectified**

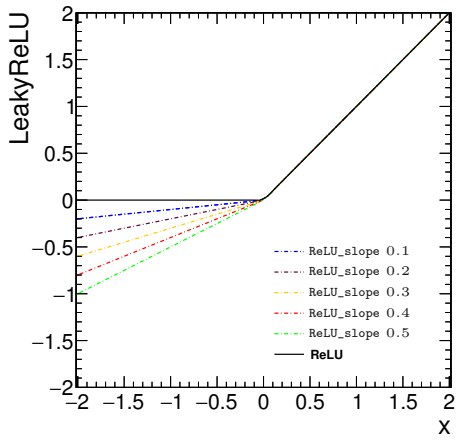


Figure 5.2: Comparison of $\mathrm{ReLU}(x_i)$ and $\mathrm{LeakyReLU}(x_i)$.

Linear Unit function (ReLU) [112] applied to every input feature x_i,

$$\mathrm{ReLU}(x_i) = (x_i)^+ = \max(0, x_i). \tag{5.2}$$

However, we find that cutting off all inputs smaller than 0 might be too strong of a restriction. Hence, we adapt the ReLU to a so-called Leaky Rectified Linear Unit function (LeakyReLU) [113] with `ReLU_slope` being the negative slope of a linear function and another tunable hyperparameter, as shown in Fig. 5.2:

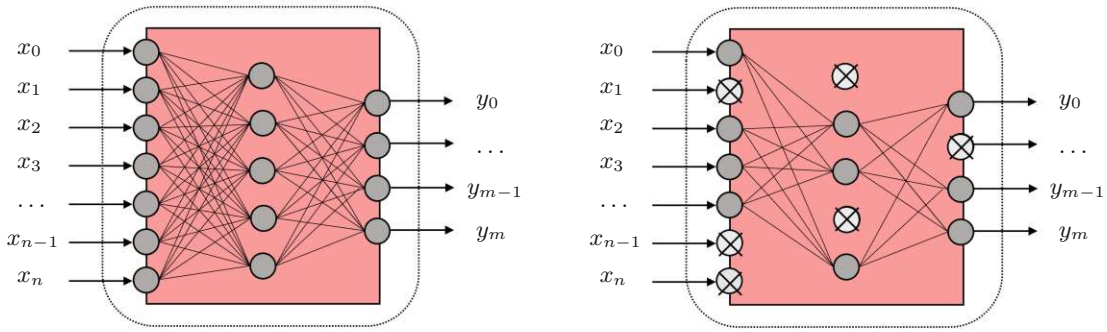
$$\mathrm{LeakyReLU}(x_i) = \begin{cases} x_i, & \text{if } x \geq 0 \\ \texttt{ReLU\_slope} \times x_i, & \text{otherwise} \end{cases}$$

For the multi-classification network, we apply the Softmax function [114] in the last step, so that all output is mapped to the range of [0,1]:

$$\mathrm{Softmax}(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)} \tag{5.3}$$

Activation functions are of great use to tackle the risk of overtraining, i.e., loosing the network's adaptability to new training data by just learning the training data sample. An additional helper are **Dropout layers** [115] that help prevent the co-adaption of neurons by randomly zeroing elements of the input vector with probability p. This means that while training, Dropout layer draw samples from an exponential number of "thinned" networks. When evaluating the network, all predictions of the thinned configurations are averaged by a single un-thinned network with smaller weights [116]. In our case, samples of thinned networks are randomly drown from a Bernoulli distribution [117]. A schematic depiction of the working principle of dropout layers can be found in Fig. 5.3. In terms of hyperparameters for optimization, we gain an additional handle in the dropout probability p stored in the `dropout` variable when tuning the network.

To sum things up, we now have several hyperparameters that need to be optimised in the training for the DNN: the number of entries in the output vector of the first layer, ergo, the hidden size of the first dense layer (`hs1`) and of the second dense layer (`hs2`). Additionally, we need to choose the slope of the LeakyReLU `ReLU_slope` and the probability p for a single neuron to be deactivated in the training, i.e., the `dropout` parameter. The input size of the first layer and the output size of the final layer are determined by the number of scalar event-level input features n and the number of targets. In our case this number is 2 in simulation-based inference and 4 in multi-classification. The final configuration of the DNN is shown in Fig. 5.4.



(a) fully connected layers

(b) dense layers + dropout

Figure 5.3: (a) Two stacked fully connected layers. Every input neuron is connected to every output neuron. (b) Two stacked fully connected layers with dropout. Some neurons are randomly deactivated.

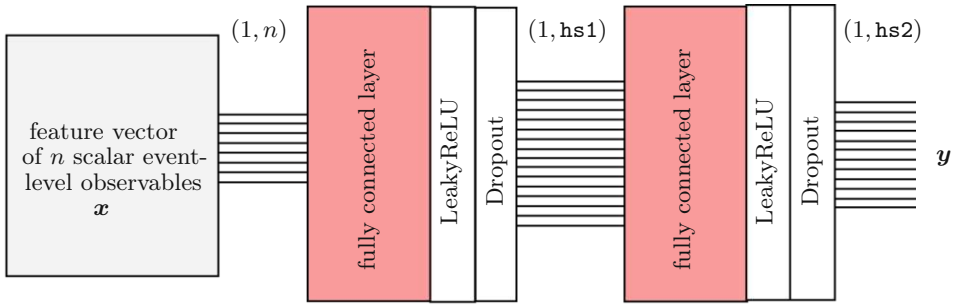


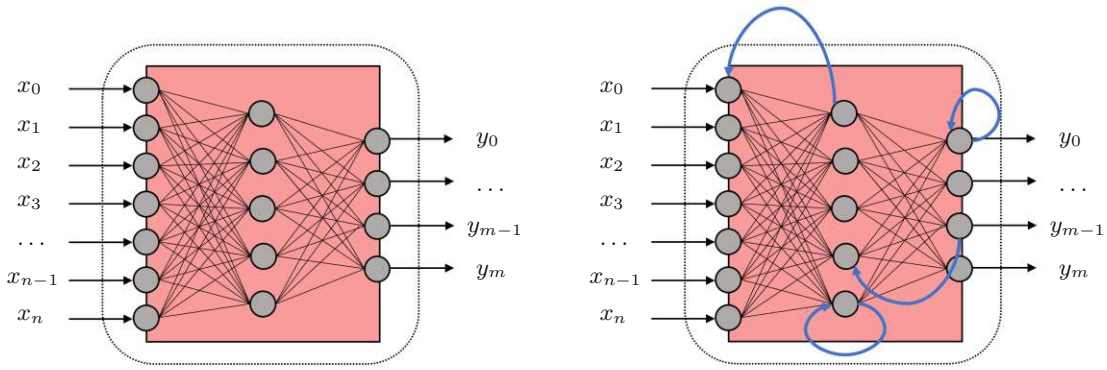
Figure 5.4: The DNN consists of two stacked dense layers of hidden sizes `hs1` and `hs2` with LeakyReLUs and Dropout layers in between. The shape of the input vector is $(1, n)$, the shape after each hidden layer i is $(1, \texttt{hs}_i)$.

5.1.2 Recurrent Neural Networks (RNNs)

To extract even more information from our data, we choose another class of algorithms to pair with the DNN layers: Recurrent Neural Networks (RNNs) [118] and, in particular, Long Short Term Memory (LSTMs) [11–14]. In fact, not all BSM effects might be detectable when only looking at scalar event-level based observables. Hence, we want our network to search for correlations between different kinematic variables in the multiple jets of one event.

The search for correlations in the jet system resembles the problem of spoken language recognition. In fact, when the machine tries to extract information from a sentence, it is a priori unclear where the most relevant piece might be located [12]. Similarly, we can teach the machine our "sentences", i.e., the collection of features of single jets within our complete jet system, just as one would do in spoken language processing. To then scan the sentence for the most relevant piece, the network must display time-dependent behaviour. Such time-dependent structures are encoded in algorithms of Recurrent Neural Networks (RNN) [118] and Long Short Term Memory (LSTM) [11–14], which we will describe in the following.

Going back to our Deep Neural Networks, where does the main difference to RNNs lie? It helps to take into consideration that our configuration of fully connected layers with dropout and activation functions from Fig. 5.4 is sometimes referred to as Feedforward Neural Networks (FNN).



(a) DNN (b) RNN

Figure 5.5: (a) Two stacked fully connected layers. This DNN configuration allows information to be passed from one node to the other in forward direction only. (b) The RNN allows information to be passed forward and backward, mimicing temporal dynamic behaviour.

In fact, information is passed through the different components in only one direction, i.e., forward. Contrarily, RNNs have additional so-called feedback connections. Opposed to DNNs, RNN therefore allow input from one node to affect the same node's subsequent input, as shown in Fig. 5.5, (b). With this intrinsic "backward" connections, RNNs mirror temporal dynamic behaviour.

5.1.3 Long Short Term Memory (LSTM)

As RNNs are likely to run into vanishing and exploding gradients, LSTMs as a subcategory of RNNs are specifically designed to overcome the first of this two malfunctions in RNNs [11]. Both of them are a consequence of the updating timescales for weights and biases in RNNs:

Long Term Memory: The so-called long term memory is connected to the updating of weights and biases *between different subsequent layers*. This is denoted with "long", as the updating is performed only after one complete training epoch, i.e., when all available input has been passed through the complete network once.

Short Term Memory: The so-called short term memory is updated more frequently, namely once per time step. Therefore, it is connected to the activation patterns of the nodes and the update of the RNN's *feedback connections*.

LSTMs help in tackling the vanishing gradient problem related to the long term memory. When performing the backpropagation operation, the gradient might vanish because of limitations in finite-precision numbers used. Here, the LSTM's workaround consists in their long short term memory, which allows gradients to stay unchanged [11–13].

Similarly to several words in a sentence in language recognition, a large number of jets originating from the $t\bar{t}t\bar{t}$ or $t\bar{t}b\bar{b}$ parents gives us the constituents of our input. Hence, the latter is no longer a simple feature vector $\boldsymbol{x}$ of shape $(1, n)$ as with the DNN, but a jet array $\mathbf{z}$ of variable length up to a defined maximum number of `N_jet_max` jets, each with m kinematic features. The working principle of an LSTM cell is given in the following equations, all performed once per element

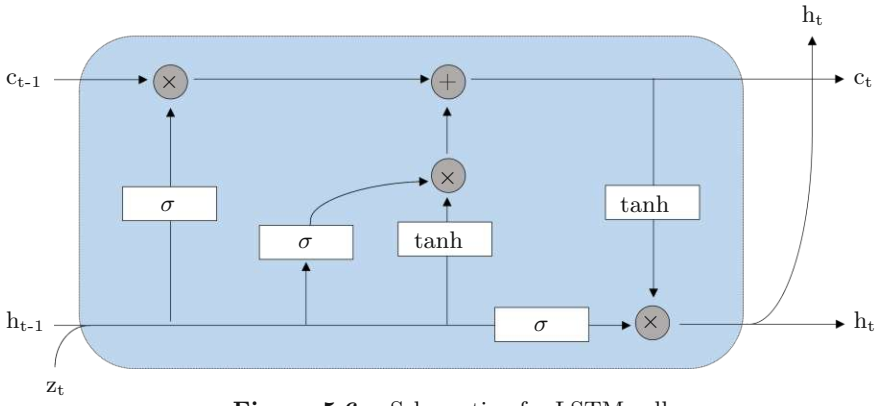


Figure 5.6: Schematic of a LSTM cell.

$\boldsymbol{z_t}$ with $t \in \{0, 1, ..., \texttt{N\_jet\_max}\}$ for the sequential input $\mathbf{z}$ of shape $(1, m, \texttt{N\_jet\_max})$ [119]:

$$\begin{aligned} i_t &= \sigma(W_{ii} z_t + b_{ii} + W_{hi} h_{t-1} + b_{hi}) \\ f_t &= \sigma(W_{if} z_t + b_{if} + W_{hf} h_{t-1} + b_{hf}) \\ g_t &= \tanh(W_{ig} z_t + b_{ig} + W_{hg} h_{t-1} + b_{hg}) \\ o_t &= \sigma(W_{io} z_t + b_{io} + W_{ho} ht - 1 + b_{ho}) \\ c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \tag{5.4}$$

In Eq. 5.4, i_t, f_t, g_t and o_t are the input, forget, cell and output gates. The LSTM's cell and hidden states are denoted by c_t and h_t, respectively. While training, the network learns the weight matrices in $\mathbf{W}$ and an eventual bias vector $\boldsymbol{b}$. $\odot$ stands for the Hadamard product. σ is the cell's activation function, i.e., the sigmoid function given by $\sigma(x_i) = (1 + \exp(-x_i))^{-1}$.

With adding LSTMs to our network architecture, we have two additional hyperparameters to optimize. The first one is the number of stacked LSTM layers (`num_layers`), the second is the output size (hidden size) of the LSTM (`hs_lstm`). Additionally, we need to decide a suitable length for `N_jet_max`.

5.2 The MVA architecture

5.2.1 The DNN+LSTM Confguration

For the final architecture, we combine both DNN and RNN components. For the scalar event-level input features stored in the input vector $\boldsymbol{x}$, we keep the configuration from Fig. 5.4. In parallel, we feed jet-array based input stored in $\mathbf{z}$ of variable length up to `N_jet_max` into one or multiple stacked LSTM layer(s). The output of the DNN is a vector $\boldsymbol{y'}$ of shape $(1, \texttt{hs2})$, whereas the output of the LSTM is a vector $\boldsymbol{y''}$ of shape $(1, \texttt{hs\_LSTM})$.

At this point, we need to concatenate the output $\boldsymbol{y'}$ and $\boldsymbol{y''}$. To combine the information gathered from DNN and LSTM, we pass the concatenated output through another DNN layer of hidden size `hs_comb`. Finally, the last dense layer gives us the two/four desired output values. A last detail is worth noting: we add LeakyReLUs and dropout layers not only after every dense layer, but also on the LSTM output.

The MVA architecture used for training is sketched in Fig. 5.7. The number of dense layers with the respective activation function have proven to be appropriate for our task. For the optimization of all aforementioned hyperparameters see chapter 6.

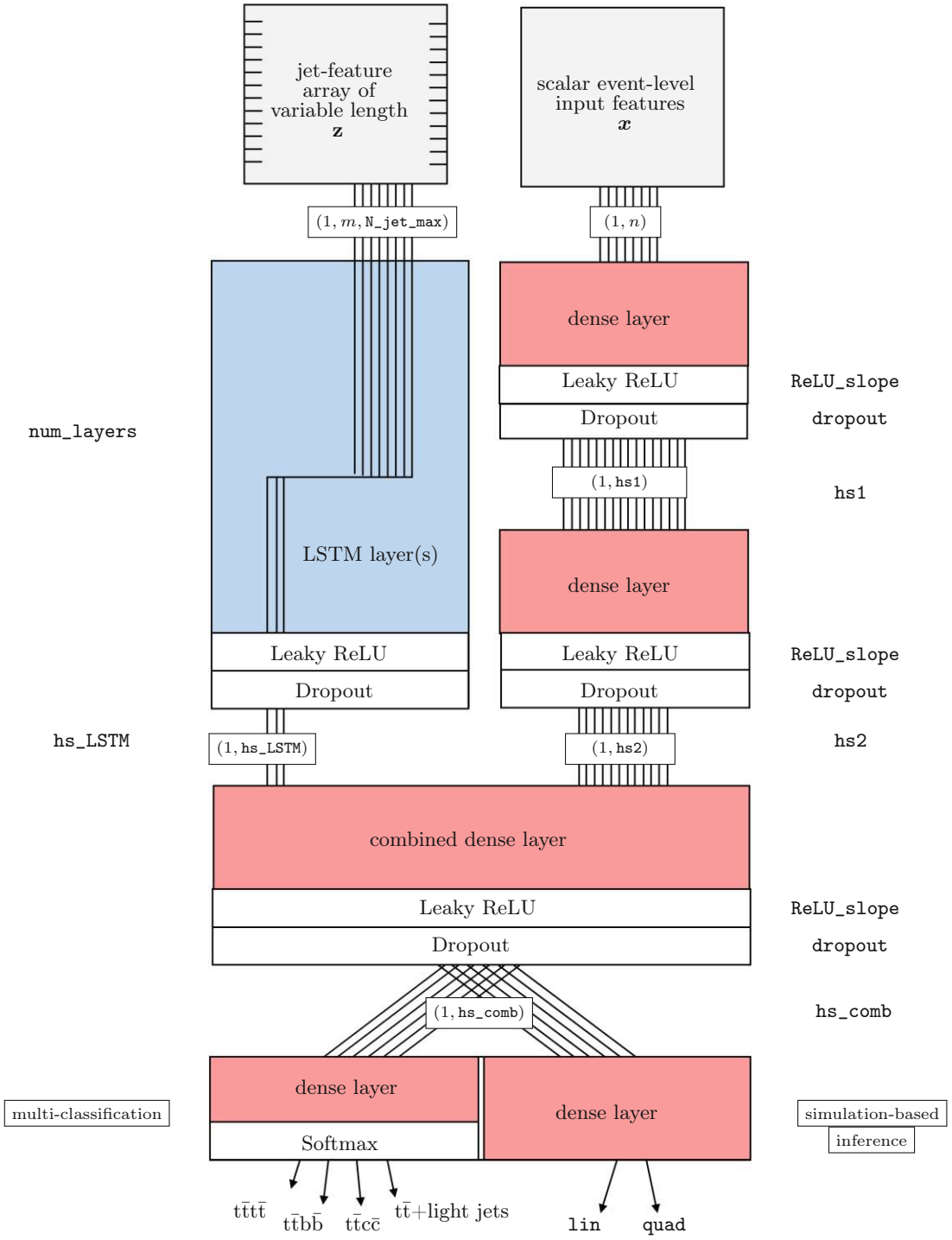


Figure 5.7: MVA Neural Network Architecture with shapes and hyperparameters. The general setup is used for simulation-based inference and multi-classification likewise, with the only difference being the configuration of the last dense layer, as indicated above.

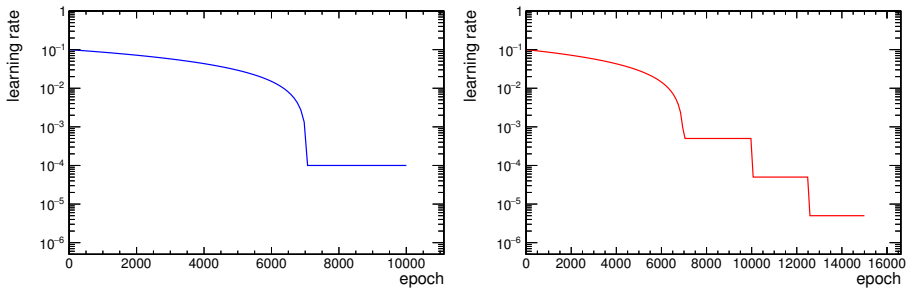
5.2.2 Optimizer and Scheduler

Lastly, we must define our model's optimizer and scheduler. In this context, the optimizer's task is to change weights of the network in order to minimize the loss. Contrarily, the scheduler models only the learning rate for an optimal adaption to the learning process [120, 121]. As an optimizer, we use the adam algorithm – an algorithm for first-order gradient-based optimization of stochastic objective functions, i.e., our target weight functions. The key feature of this approach is found in its estimation of lower-order moments that are then used for optimally updating the training weights. Further detail is found in the original paper in Ref. [122].

The learning rate scheduler demands more in-depth tuning, as the BSM-effects result only in minimal nuisances in the training data with respect to the SM. In fact, starting with a relatively high learning rate of 0.1 is necessary to not get stuck in a local minimum of the parameter space. Then, we decay the learning rate linearly until reaching 7/10 of the total epochs. At this point, we continue training until the total number of 10000 epochs at a relatively low learning rate of 0.0001. The slope of the linear decay is referred to as `s_factor` and takes the value of 0.001 for training with signal only. Hence, we implement a scheduler that adapts the learning rate according to the sketch in Fig. 5.8, (a).

When training with signal and background, an adequate tuning of the learning rate is even more crucial for an optimal training process. Here, it turns out that the best performance is reached if we start with an initial learning rate of 0.1 as before, but turn it down to slightly higher value of 0.005 after 7/10 of the total epochs. When reaching the total number of 10000 epochs, we assess the model's performance. In most cases, it turns out useful to continue training with a learning rate of 5e-05 for some 2000 epochs and then, if necessary, go to a very tiny learning rate of only 5e-06. This assures that the small BSM nuisances are adequately learned even when there is background present. The decay of the learning rate is shown in Fig. 5.8, (b).

With respect to our hyperparameter set, we now add the parameter `w_decay` that represents the weight decay in the adam optimizer (as we set it to zero this has no impact on the training). Next, the initial learning rate `start_lr` that is 0.1 for both signal and signal+background trainings. Lastly, we have the target learning rate at 7/10 of the total number of epochs `target_lr` or the `s_factor`, respectively. The combination of these two parameter is (1e-04 / 0.001) when training on signal data only, and (5e-04 / 0.005) for the first learning rate plateau with signal and background. If needed, two plateaus of variable length at a `target_lr` of 5e-05 or 5e-06 are added. A complete collection of all hyperparameter can be found in the next chapter 6, Tab. 6.1.



(a) training with signal **(b)** training with signal+background

Figure 5.8: The optimal learning rate evolution for training with (a) signal only and (b) signal+background. In the latter case, scheduling the learning rate down after the first training part increases the probability that small nuisances of the EFT-weights in the signal are learned properly by the network.

5.2.3 The Loss Function

Coming back to the network's two tasks – multi-classification and simulation-based inference in SMEFT – the loss function is where the two MVA setups fundamentally differ. The loss function for simulation-based inference has been discussed extensively in the dedicated chapter 4. Here, we just repeat the final expression, which is given by

$$L = \sum_{\theta \in \mathcal{B}} \sum_i w_{i,0} \left[\theta \left(\frac{w_{i,1}}{w_{i,0}} - R_{\text{lin}}(x_i) \right) + \theta^2 \left(\frac{w_{i,2}}{w_{i,0}} - R_{\text{quad}}(x_i) \right) \right]^2 , \tag{5.5}$$

where $\mathcal{B}$ contains two different, arbitrarily chosen base points $\theta \neq 0$, $w_{i,j}$ are the per-event weights from simulation, and R_{lin} and R_{quad} are the regression targets in the polynomial dependence of the detector-level likelihood ratio.

For multi-classification, we use the PyTorch build-in mean square error loss function between input $\boldsymbol{x}$ and target $\boldsymbol{y}$ [123],

$$L = \frac{1}{N_{\text{events}}} \sum_{i=1}^{N_{\text{events}}} (\boldsymbol{x}_i - \boldsymbol{y}_i)^2. \tag{5.6}$$

As this loss function is more sturdy and robust than the complex loss function for simulation-based inference, this makes the multi-classification an ideal proxy for testing the network's configuration in an easy to control setup.

5.3 The MVA input features

5.3.1 Training variable selection

For training, we use 40 scalar event-level input features that are fed into the dense layers in Fig. 5.7. The input features comprise variables regarding jets, leptons, b-tagged jets and missing energy and are characterized in Tab. 5.1 below.

From the kinematic variables of transverse momentum p_{T}, pseudorapidity η and azimuthal angle ϕ we can derive the angular distance ΔR with the definition of the coordinate system of CMS from Fig. 5.9

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}. \tag{5.7}$$

Furthermore, we calculate the invariant mass of two decay products as

$$m^2 = p^{\mu_1} p_{\mu_2}, \tag{5.8}$$

which for a 4-momentum p^{μ_i} in the limit of massless particles

$$p^{\mu_i} = |\boldsymbol{p_i}| \begin{pmatrix} \sqrt{1 + \frac{m_i^2}{|\boldsymbol{p_i}|^2}} \\ \cos(\phi_i)\sin(\theta_i) \\ \sin(\phi_i)\sin(\theta_i) \\ \cos(\theta_i) \end{pmatrix} = |\boldsymbol{p}_{\text{T}_i}| \begin{pmatrix} \sqrt{\cosh^2(\eta_i) + \frac{m_i^2}{|\boldsymbol{p}_{\text{T}_i}|^2}} \\ \cos(\phi_i) \\ \sin(\phi_i) \\ \sinh(\eta_i) \end{pmatrix} \overset{\text{m=0}}{=} |\boldsymbol{p}_{\text{T}_i}| \begin{pmatrix} \cosh(\eta_i) \\ \cos(\phi_i) \\ \sin(\phi_i) \\ \sinh(\eta_i) \end{pmatrix} \tag{5.9}$$

evaluates to

$$m_{\text{T}}{}^2 = 2 p_{\text{T}_1} p_{\text{T}_2} (\cosh(\eta_1 - \eta_2) - \cos(\phi_1 - \phi_2)). \tag{5.10}$$

Finally, we use the m_{T2} constructed from the missing transverse momentum $\not{p}_{\text{T}}$ and the particles specified in the entries 38 - 40 in Tab. 5.1 below according to

$$M_{\text{T2}}^2 \equiv \min_{\not{p}_{\text{T}_1} + \not{p}_{\text{T}_2} = \not{p}_{\text{T}}} [\max\{m_{\text{T}}^2(p_1, \not{p}_{\text{T}_1}), m_{\text{T}}^2(p_1, \not{p}_{\text{T}_2})\}] \tag{5.11}$$

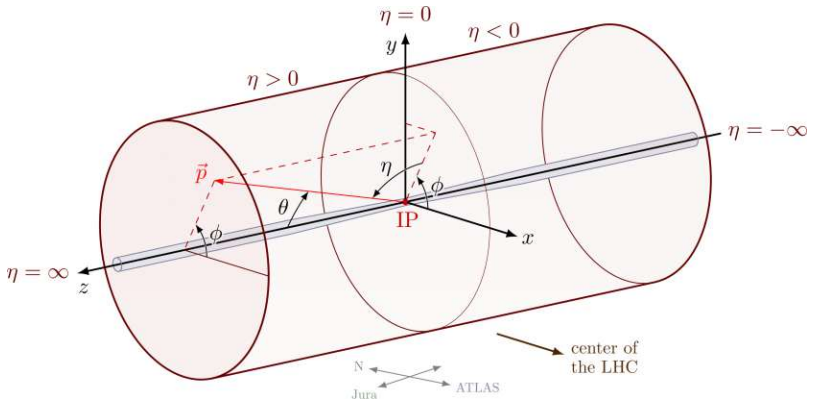


Figure 5.9: The coordinate system of CMS. Image taken from Ref. [124].

with unknown fractions of missing momentum $\not{p}_{\mathrm{T}_i}$ and the observed missing momentum $\not{p}_{\mathrm{T}}$ [125, 126]. The following table enumerates all variables used as DNN input:

No.	Symbol	Definition
1	N_{recoJet}	jet multiplicity
2	H_{T}	scalar sum of the transverse momentum of all jets in an event
3	$H_{\mathrm{T},b}$	scalar sum of the transverse momentum of b-tagged jets
4	$H_{\mathrm{T,ratio}}$	scalar sum of p_{T} of the first four jets divided by H_{T}
5	$p_{\mathrm{T}}(\mathrm{j}_0)$	transverse momentum of the leading jet
6	$\eta(\mathrm{j}_0)$	pseudorapidity of the leading jet
7	$p_{\mathrm{T}}(\mathrm{j}_1)$	transverse momentum of the subleading jet
8	$\eta(\mathrm{j}_1)$	pseudorapidity of the subleading jet
9	$p_{\mathrm{T}}(\mathrm{j}_2)$	transverse momentum of the third leading jet
10	$\eta(\mathrm{j}_2)$	pseudorapidity of the third leading jet
11	$p_{\mathrm{T}}(\mathrm{j}_3)$	transverse momentum of the forth leading jet
12	$p_{\mathrm{T}}(\mathrm{j}_4)$	transverse momentum of the fifth leading jet
13	$p_{\mathrm{T}}(\mathrm{j}_5)$	transverse momentum of the sixth leading jet
14	$p_{\mathrm{T}}(\mathrm{j}_6)$	transverse momentum of the seventh leading jet
15	$p_{\mathrm{T}}(\mathrm{j}_7)$	transverse momentum of the eighth leading jet
16	$p_{\mathrm{T}}(\mathrm{b}_0)$	transverse momentum of the leading b-tagged jet
17	$p_{\mathrm{T}}(\mathrm{b}_1)$	transverse momentum of the subleading b-tagged jet
18	$p_{\mathrm{T}}(\ell_0)$	transverse momentum of the leading lepton
19	$\eta(\ell_0)$	pseudorapidity of the leading lepton
20	$p_{\mathrm{T}}(\ell_1)$	transverse momentum of the subleading lepton
21	$\eta_(\ell_1)$	pseudorapidity of the subleading lepton
22	$m_{\mathrm{T}}(\ell_0)$	transverse mass of the leading lepton
23	$m_{\mathrm{T}}(\ell_1)$	transverse mass of the subleading lepton
24	$m_{\mathrm{T}}(\ell_0, \ell_1)$	invariant mass of the lepton system
25	$m_{\mathrm{T}}(\mathrm{j}_0)$	transverse mass of the leading jet and the leading lepton
26	$m_{\mathrm{T}}(\mathrm{j}_1)$	transverse mass of the subleading jet and the leading lepton
27	$m_{\mathrm{T}}(\mathrm{j}_0, \mathrm{j}_1)$	invariant mass of the leading / subleading jet system
28	$\not{E}_{\mathrm{T}}$	missing transverse energy
29	$\Delta\phi(\ell_0, \ell_1)$	difference in azimuthal angle of the leading and subleading lepton
30	$\Delta\phi(\mathrm{j}_0, \mathrm{j}_1)$	difference in azimuthal angle of the leading and subleading jet
31	$\Delta\eta(\ell_0, \ell_1)$	difference in pseudorapidity of the leading and subleading lepton

32	$\Delta\eta(\mathrm{j}_0, \mathrm{j}_1)$	difference in pseudorapidity of the leading and subleading jet
33[1]	m_{4b}	mass of the system of 4 b-jets if there are as many else 0
34[1]	$\Delta R_0(\ell, \mathrm{b})$	minimum angular distance in the lepton/b-jet system
35[1]	$\Delta R_1(\ell, \mathrm{b})$	second smallest angular distance in the lepton/b-jet system
36[1]	$\Delta R(\mathrm{b}, \mathrm{b})$	minimum angular distance in the b-jet system
37[1]	$\Delta R(\ell, \ell)$	minimum angular distance between leading/subleading lepton
38[1]	$m_{\mathrm{T2}}(\ell, \ell)$	m_{T2} variable from leading and subleading leptons [125, 126]
39[1]	$m_{\mathrm{T2}}(\mathrm{b}, \mathrm{b})$	m_{T2} variable from leading and subleading b-jets [125, 126]
40[1]	$m_{\mathrm{T2}}(\mathrm{b} + \ell)$	m_{T2} variable from two leading leptons and b-jets [125, 126]
41[2]	$\mathrm{BTag}_{\mathrm{j}_0}$	b-tagging score of the leading jet
42[2]	$\mathrm{BTag}_{\mathrm{j}_1}$	b-tagging score of subleading jet
43[2]	$\mathrm{BTag}_{\mathrm{j}_2}$	b-tagging score of third leading jet

Table 5.1: Scalar event-level input features for dense layers (DNNs)

For the LSTM-layers in simulation-based inference, we have only four features as we are limited by the DELPHES parametrisation of the reconstructed jets [127, 128]. The features are listed in Tab. 5.2 below, whereas details on the sample generation will be given in the next chapter 6. In addition to the kinematic variables p_{T}, η and ϕ the binary b-tagging criterion is 0 if the jet is not b-tagged and 1 if the contrary is true.

No.	Symbol	Definition
1	p_{T}	transverse momentum of the jet
2	η	pseudorapidity of the jet
3	ϕ	azimuthal angle of the jet
4	bTag	binary b-tagging criterion

Table 5.2: Jet array features for LSTM layer(s) in simulation-based inference.

Contrarily, in multi-classification we have the possibility to use more features as the training data is not processed with the limited detector simulation in DELPHES. In addition to the kinematic quantities p_{T}, η and ϕ of each jet, we can also feed our LSTMs discriminator quantities between b-quarks, b-quarks and leptons as well as between light quarks and gluons. Additional features comprise c-quark vs b-quark discriminators and the pile-up probability. The complete list can be found in Tab. 5.3 below.

No.	Symbol	Definition
1	p_{T}	transverse momentum of the jet
2	η	pseudorapidity of the jet
3	ϕ	azimuthal angle of the jet
4	BTagB	b+bb+lepb tag discriminator
5	BTagCvB	c vs b+bb+lepb discriminator
6	BTagQG	gluon vs light quark discriminator
7	puId	pile-up jet probability
8	qgl	quark vs gluon likelihood discriminator

Table 5.3: Jet array features for LSTM layer(s) in multi-classification.

[1] Feature not used in multi-classification.
[2] Feature not used in simulation-based inference.

Chapter 6

Data Generation and Training

6.1 The input data

6.1.1 MadGraph5_aMC@NLO and DELPHES

For simulating the $t\bar{t}t\bar{t}$ and $t\bar{t}b\bar{b}$ signals for simulation-based inference, we use the framework MadGraph5_aMC@NLO. MadGraph5_aMC@NLO is a framework that creates code to compute tree-level and next-to-leading order cross sections, matches the output to parton shower simulations and merges samples that differ by light-parton multiplicities only [129]. The user's input in all this is limited to the input of the physical quantities:

```
    import model SMEFTsim_topU3l_MwScheme_UFO-massless_4t
    define p = g u c d s u~ c~ d~ s~
    define j = g u c d s u~ c~ d~ s~
    define l+ = e+ mu+ ta+
    define l- = e- mu- ta-
    define vl = ve vm vt
    define vl~ = ve~ vm~ vt~
    define p = p b b~
    define j = j b b~
⟶  generate p p > t t~ t t~ SMHLOOP=0 NPprop=0 NP=1 @0
    output TTTT_MS -nojpeg
```

for $t\bar{t}t\bar{t}$ and

```
    ...
⟶  generate p p > t t  b b  SMHLOOP=0 NPprop=0 NP=1 @0
    output TTbb_MS -nojpeg
```

for $t\bar{t}b\bar{b}$. This means, we simulate the decays with no extra jet for our signal processes.

The use of SMEFTsim_3 allows us to include new physics (`NP=1`), as this package is specifically designed for automated computations in SMEFT [130]. The `topU3l` extension matches our requirements that we choose in accordance with the interpretation of top-quark LHC measurements in SMEFT, again follwing Ref. [31] based on Ref. [90]. In this process, we use MadSpin implemented in MadGraph5_aMC@NLO to decay narrow resonances while preserving spin correlation [131]. Exemplar histograms of the kinematic variable H_T for our simulated events at generator level are shown in Fig. 6.1. In Fig. 6.2, histograms of the EFT weights for all operators in Tab. 3.1, chapter 3 are shown.

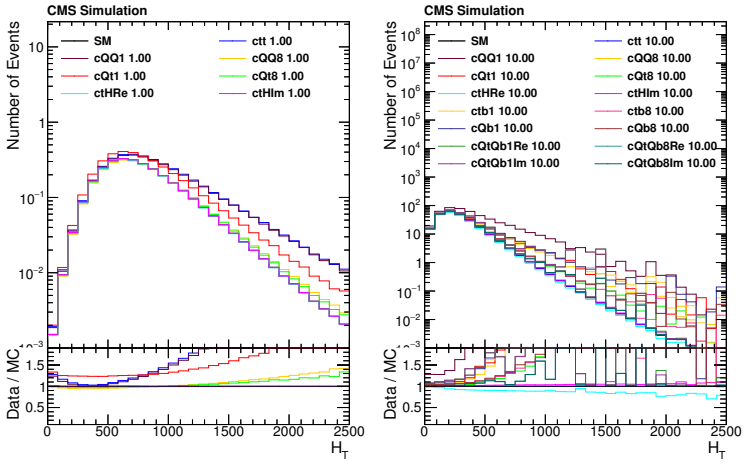
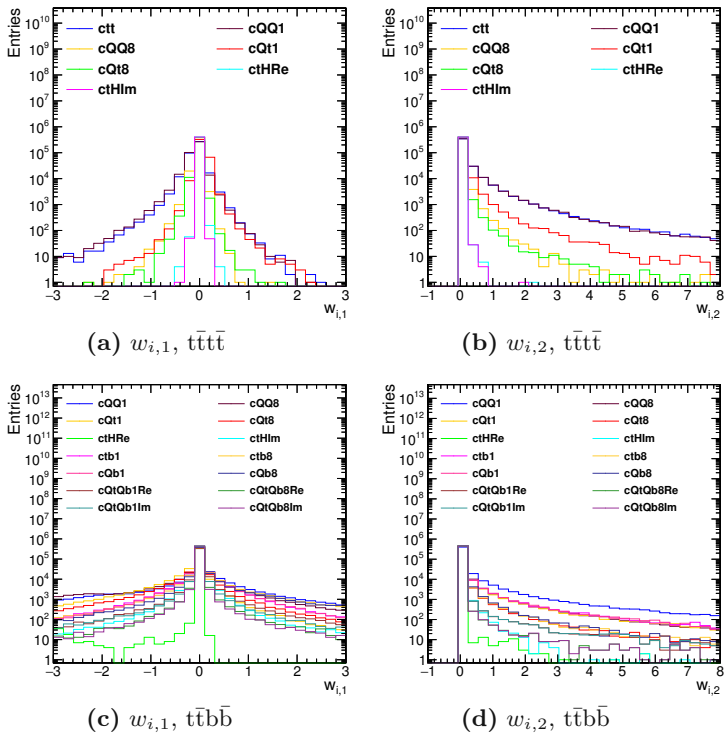


Figure 6.1: Histogram of (a) $t\bar{t}t\bar{t}$ and (b) $t\bar{t}b\bar{b}$, no event selection applied. In $t\bar{t}t\bar{t}$, EFT operators are weighted with $\theta = 1$, in $t\bar{t}b\bar{b}$ with 10.



(a) $w_{i,1}$, $t\bar{t}t\bar{t}$ **(b)** $w_{i,2}$, $t\bar{t}t\bar{t}$

(c) $w_{i,1}$, $t\bar{t}b\bar{b}$ **(d)** $w_{i,2}$, $t\bar{t}b\bar{b}$

Figure 6.2: Histogram of the linear and quadratic coefficients $w_{i,1}$ and $w_{i,2}$ of the EFT weights $w_i = w_{i,0} + \theta w_{i,1} + \theta^2 w_{i,2}$ at generator level in $t\bar{t}t\bar{t}$ (a)-(b) and $t\bar{t}b\bar{b}$ (c)-(d), no event selection applied.

Then, we process the events with the model of the CMS detector in DELPHES, a framework that provides fast simulation of the CMS detector [127, 128]. The generator level samples are used to subsequently simulate the detector and retrieve output observables as, e.g., isolated leptons, missing transverse energy and collection of jets in an event. In this process, DELPHES takes into account subdetector resolutions while smearing the kinematics of the particles in the final states. Additionally, DELPHES includes the following six features [127, 128]:

1. geometrical implementation of the detector,
2. the magnetic field and its effect on charged particles' tracks,
3. reconstruction of photons, leptons, jets, b-jets, τ-jets, missing transverse energy,
4. lepton isolation,
5. trigger emulation and
6. an event display.

However, being only a fast detector simulation, DELPHES comes with limitations in comparison to the full detector simulation GEANT4 [132]: an idealised geometry (uniform, symmetric around the beam axis, no cracks, no material), no secondary interactions, no multiple scatterings, neglected photon conversion and bremsstrahlung [127, 128].

For training with signal and background in simulation-based inference, we use the $t\bar{t}$ sample produced with POWHEG [133] at next leading order. The shower has been simulated with PYTHIA [134] and the detector simulation with GEANT4.

For multi-classification, we use $t\bar{t}t\bar{t}$ produced with MADGRAPH5_aMC@NLO as signal sample. The shower has been simulated with PYTHIA and the detector response with GEANT4. The three background categories result from the $t\bar{t}$ background, which we divide in $t\bar{t}b\bar{b}$, $t\bar{t}c\bar{c}$ and $t\bar{t}$+light jets components.

6.1.2 Event selection

Next, we follow Ref. [31] mimicking the analysis in Ref. [102] and apply the following requirements for our selection in $t\bar{t}t\bar{t}$ and $t\bar{t}b\bar{b}$ in simulation-based inference: Each event must have two reconstructed leptons with transverse momentum $p_T > 20\,\text{GeV}$ and pseudorapidity $|\eta| < 2.4$, missing transverse energy must be $\not{E}_T > 30\,\text{GeV}$. Additionally, we require every event to have at least four jets with $p_T > 30\,\text{GeV}$ and $|\eta| < 2.5$, with at least two of them being b-tagged jets.

When training with backgrounds, the selection changes as follows: In $t\bar{t}t\bar{t}$+background, we require two or more leptons with $p_T > 20\,\text{GeV}$ and $|\eta| < 2.4$, six or more jets with $p_T > 30\,\text{GeV}$ and $|\eta| < 2.5$, one or more of which must be a b-jet. Additionally, we apply a cut on $H_T = 500\,\text{GeV}$. In $t\bar{t}b\bar{b}$+background, we require two or more leptons with $p_T > 20\,\text{GeV}$ and $|\eta| < 2.4$, four or more jets with $p_T > 30\,\text{GeV}$ and $|\eta| < 2.5$, one or more of which must be a b-jet to consistently remove all extra double b-jet activity. For the background, we assume that no EFT weights are present, i.e., that the EFT operators just act upon the signal whereas the background is described by the SM only. For all processes, we normalize to the same integrated luminosity of $\mathcal{L} = 300\,\text{fb}^{-1}$. Histograms of selected input features can be found in Fig. 6.3 - Fig. 6.6 on the next pages.

For multi-classification, we require a minimum number of 4 jets, of which at least three are b-tagged. Additionally, we require $H_T > 500\,\text{GeV}$ and use the two lepton opposite sign channel. Histograms of selected input features can be found in Fig. 6.7.

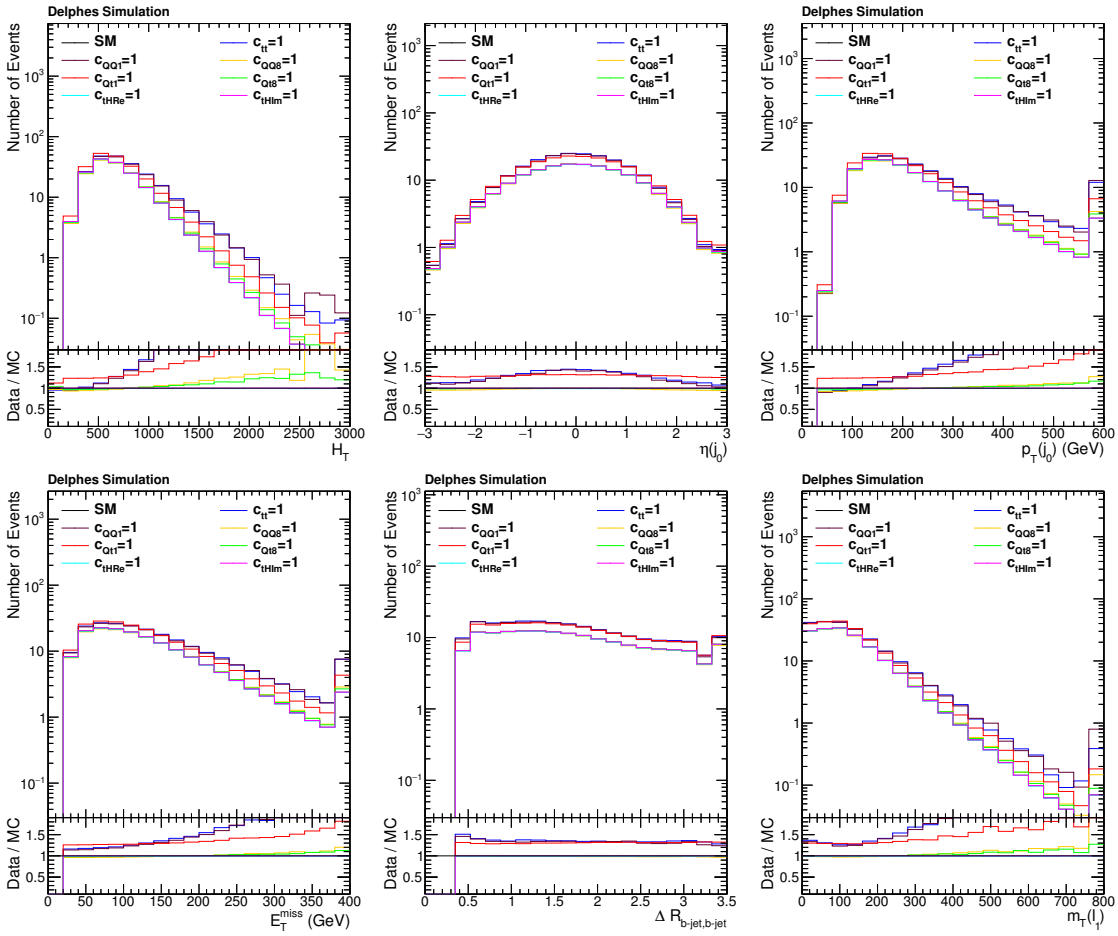


Figure 6.3: Histogram of selected input features for $t\bar{t}t\bar{t}$.

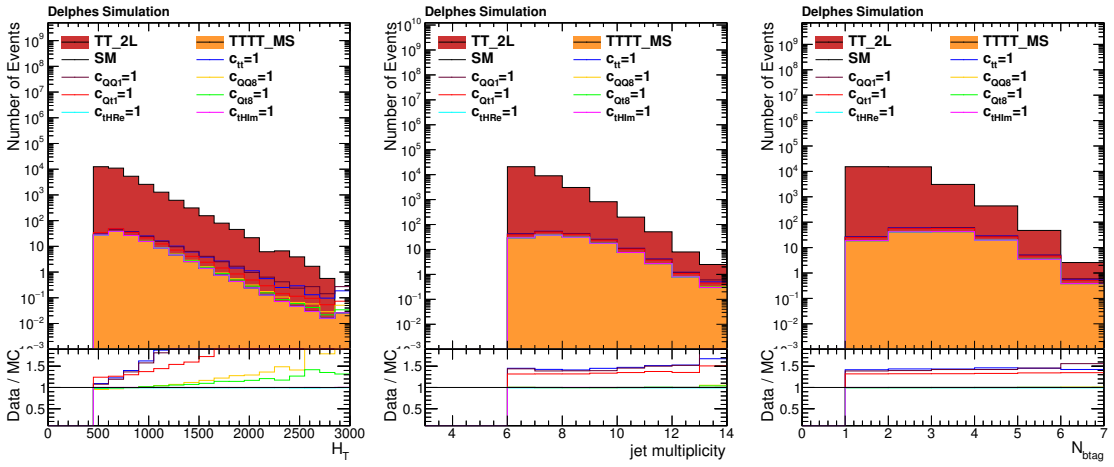


Figure 6.4: Histogram of selected input features for $t\bar{t}t\bar{t}$ with $t\bar{t}$ background

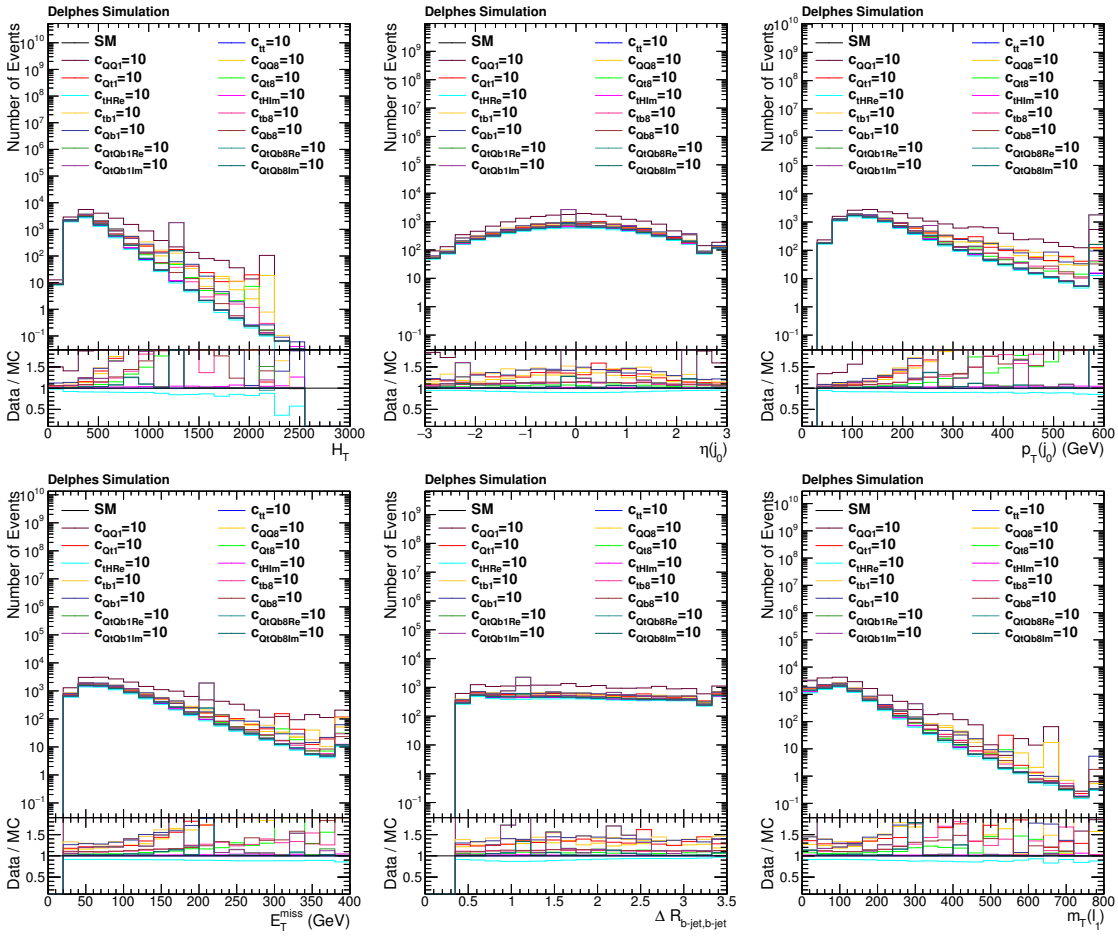


Figure 6.5: Histogram of selected input features for $t\bar{t}b\bar{b}$.

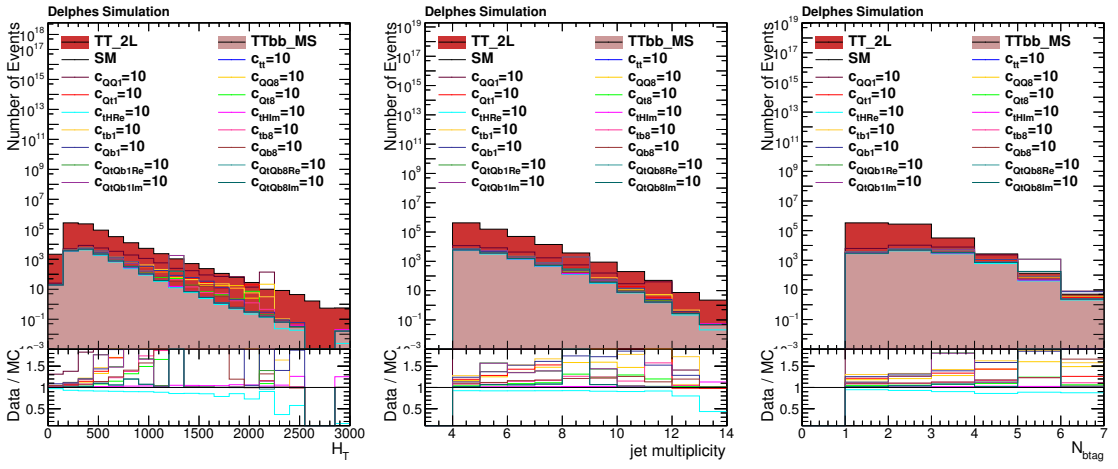


Figure 6.6: Histogram of selected input features for $t\bar{t}b\bar{b}$ with $t\bar{t}$ background

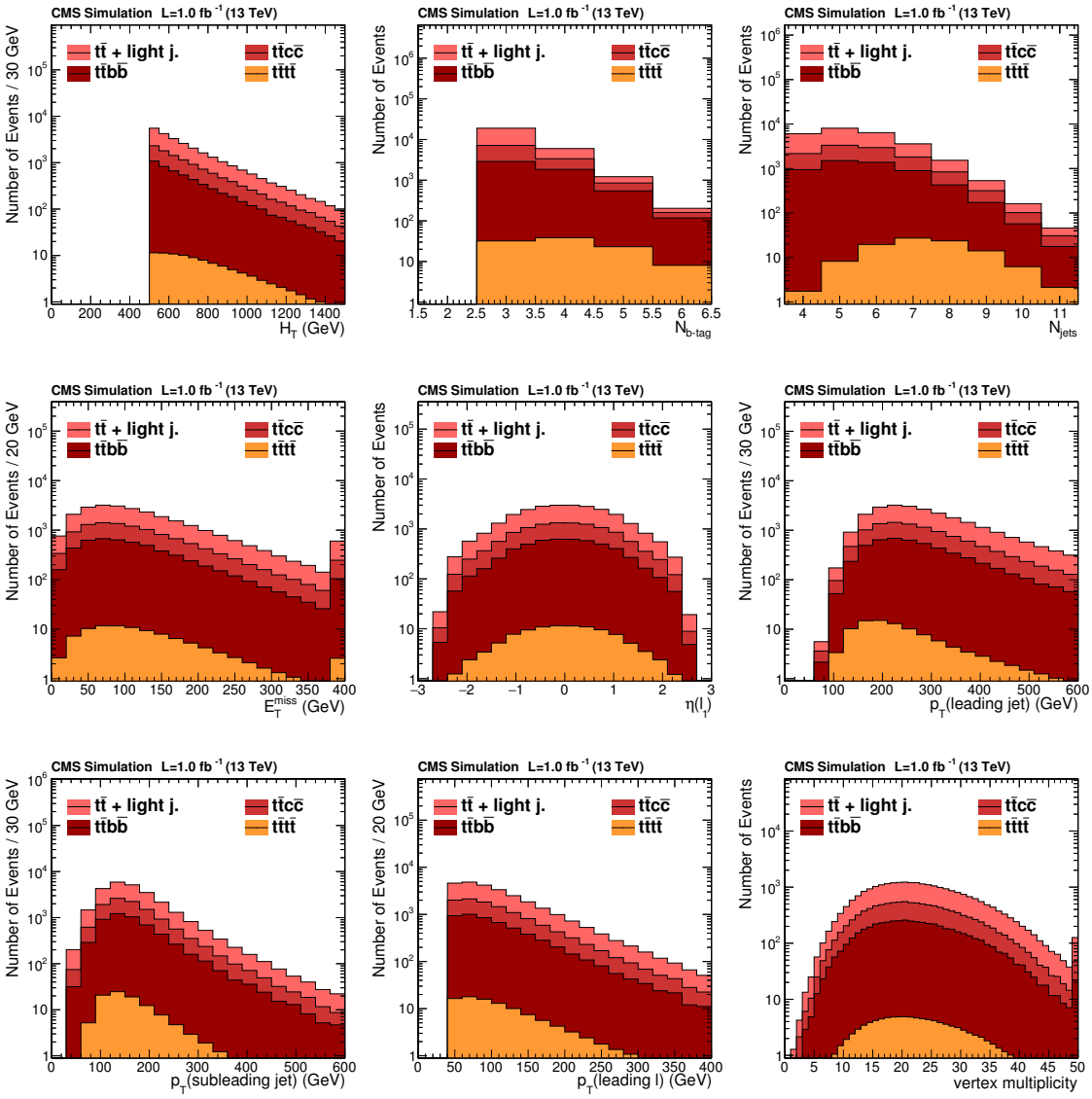


Figure 6.7: Histogram of selected input features for multi-classification of $t\bar{t}t\bar{t}$, $t\bar{t}b\bar{b}$, $t\bar{t}c\bar{c}$ and $t\bar{t}$+light jets.

6.2 Hyperparameter optimization

With the MVA architecture of chapter 5 and the simulated samples at hand, we now optimize the tunable hyperparameter. In Tab. 6.1, a summary of all hyperparameters related to the DNN or LSTM part of the network are listed, together with the general training parameters.

No.	Variable	Definition	default
DNN			
1	`hs1`	hidden size, first dense layer	80
2	`hs2`	hidden size, second dense layer	45
3	`hs_comb`	hidden size of the combined dense layer if `LSTM == True`, hidden size of third dense layer if `LSTM == False`	5
4	`ReLU_slope`	negative slope of LeakyReLU(x) for $x < 0$	0.5
5	`dropout`	dropout probability in dropout layers	0.5
LSTM			
6	`num_layers`	number of stacked LSTM layers	2
7	`hs_lstm`	output size, LSTM	4
8	`ReLU_slope`	negative slope of LeakyReLU(x) for $x < 0$ after LSTM	0.5
9	`dropout`	dropout probability in dropout layer after LSTM	0.5
General setup			
10a	`n_epochs` (sig)	number of training epochs for signal only	10000
10b	`n_epochs` (sig+bkg)	number of training epochs for signal+background	15000
11	`batches`	number of batches	1
12	`optimizer`	optimizer	adam [122]
13	`w_decay`	rate of decaying weights	0
14	`scheduler`	scheduler	linear+ flat
15a	`s_factor`(sig)	scheduler decay factor	0.001
15b	`s_factor` (sig+bkg)	scheduler decay factor	0.005
16	`LSTM`	add LSTM layer(s)	`False`
17	`start_lr`	start learning rate	0.1
18a	`target_lr` (sig)	target learning rate (signal)	1e-04
18b	`target_lr` (sig+bkg)	target learning rate (signal+background)	5e-04 5e-05 5e-06

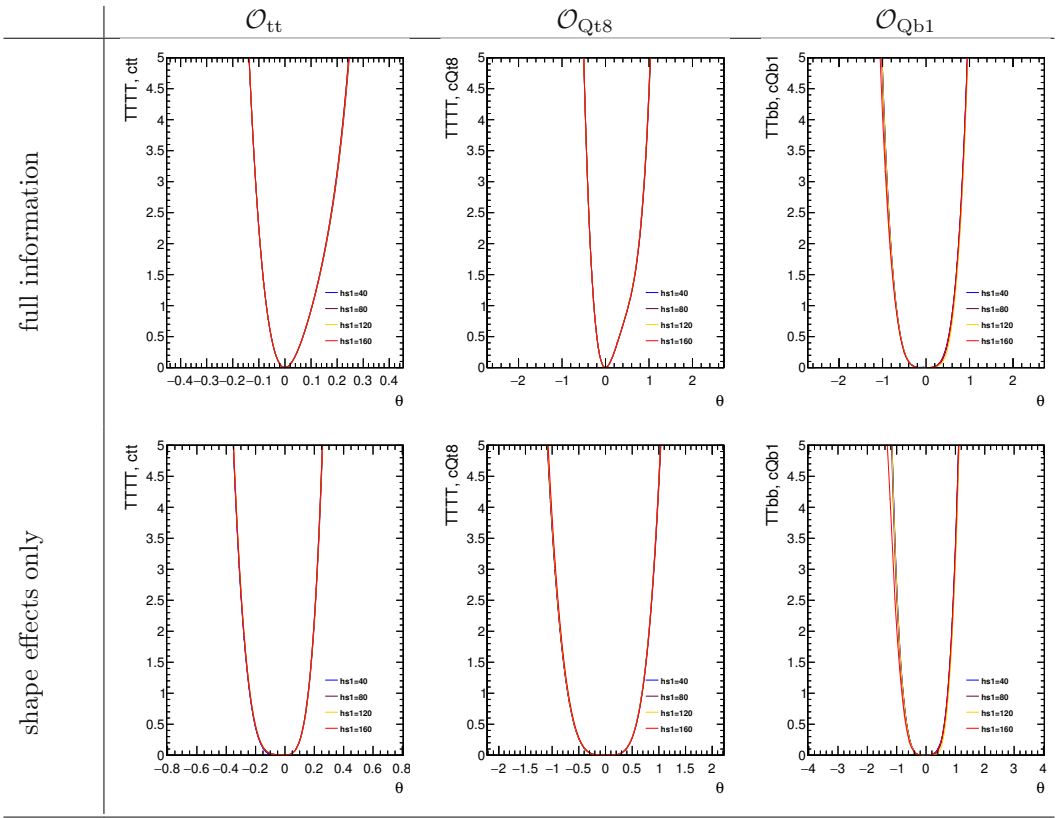
Table 6.1: Collection of all tunable hyperparameter and general components of the MVA architecture with their respective default value before hyperparameter optimization.

We perform the hyperparameter optimization in the simulation-based inference setup. This is not only because learning the BSM signatures is the central aim of this thesis, but also because the task is far more complicated than the relatively simple multi-classification. Hence, optimizing in the more delicate setup is a sensible choice to prevent pseudo-optimization in an already robust setting.

6.2.1 Hyperparameter optimization DNN

As different operators have different effects on the shape information and given their various interference strength, we choose three operators for the hyperparameter optimization of the DNN. The first one is $\mathcal{O}_{\mathrm{tt}}$ in $\mathrm{t\bar{t}t\bar{t}}$, as it is among those operators with the strongest effects. Next, $\mathcal{O}_{\mathrm{Qt8}}$ in $\mathrm{t\bar{t}t\bar{t}}$, with tiny yield variations and mostly BSM information in the shape. Lastly, we probe $\mathcal{O}_{\mathrm{Qb1}}$ in $\mathrm{t\bar{t}b\bar{b}}$, to assure hyperparameter optimization is suitable for both signal samples $\mathrm{t\bar{t}t\bar{t}}$ and $\mathrm{t\bar{t}b\bar{b}}$. In the training setups 1-4 described in Tab. 6.2, we vary the size of the first hidden layer, `hs1`, between 40 and 160, where 40 is the total number of scalar event-level input features. The size of the subsequent layers is equivalent to our default configuration. For performance evaluation, we compute the LLR as described in section 4.2 in chapter 4. Furthermore, we probe the configuration not only on the complete learning output denoted by "full information" in the following tables, but we also probe the shape-sensitive part. As described in chapter 4, we achieve this second evaluation mode by normalizing both SM and BSM yield to the same expected number of events, i.e., $\mathcal{L}\sigma_{\mathrm{SM}}$. This second choice of normalization is referred to as "shape effects only".

The results are displayed in Tab. 6.2 below. We see only minor differences in performance with varying hidden size of the first dense layer in $\mathcal{O}_{\mathrm{Qb1}}$ for an overly large `hs1` and no differences at all for the operators in $\mathrm{t\bar{t}t\bar{t}}$. Hence, the configuration of a first hidden size of 80 – two times the number of scalar event-level input features – is adequate.



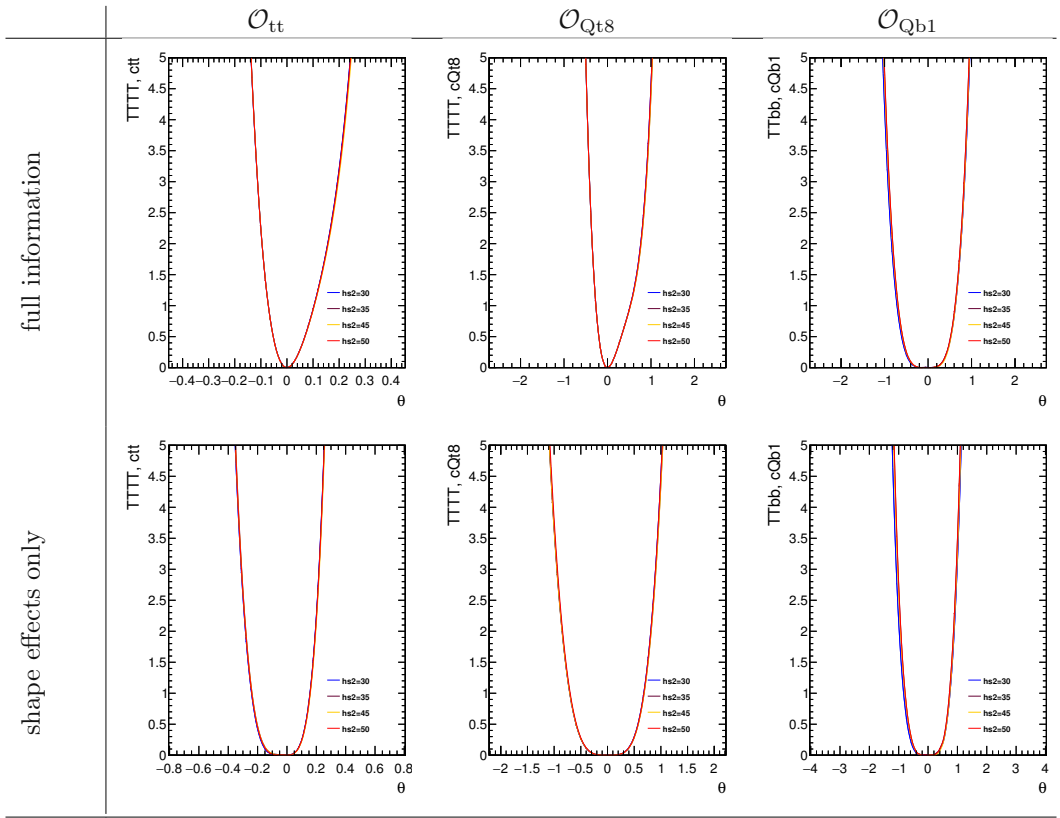
No.	hs1	hs2	hs_comb	LSTM	num_layers	hs_LSTM
1-4	40 / 80 / 120 / 160	45	5	-	-	-

Table 6.2: Hyperparameter optimization: hidden size of the first dense layer (`hs1`)

In the training setups 5-8 in Tab. 6.3, we probe the optimal size of the second dense layer. As before with `hs1`, performance is decreasing very slightly for $\mathcal{O}_{\mathrm{Qb1}}$ in case `hs2` is chosen to be excessively small. In $t\bar{t}t\bar{t}$, which generally has more pronounced EFT effects visible in the shapes, the performance is unaltered. Hence, the configuration of `hs2 = 45`, corresponding to the number of scalar event-level based input features plus 5, meets the requirements.

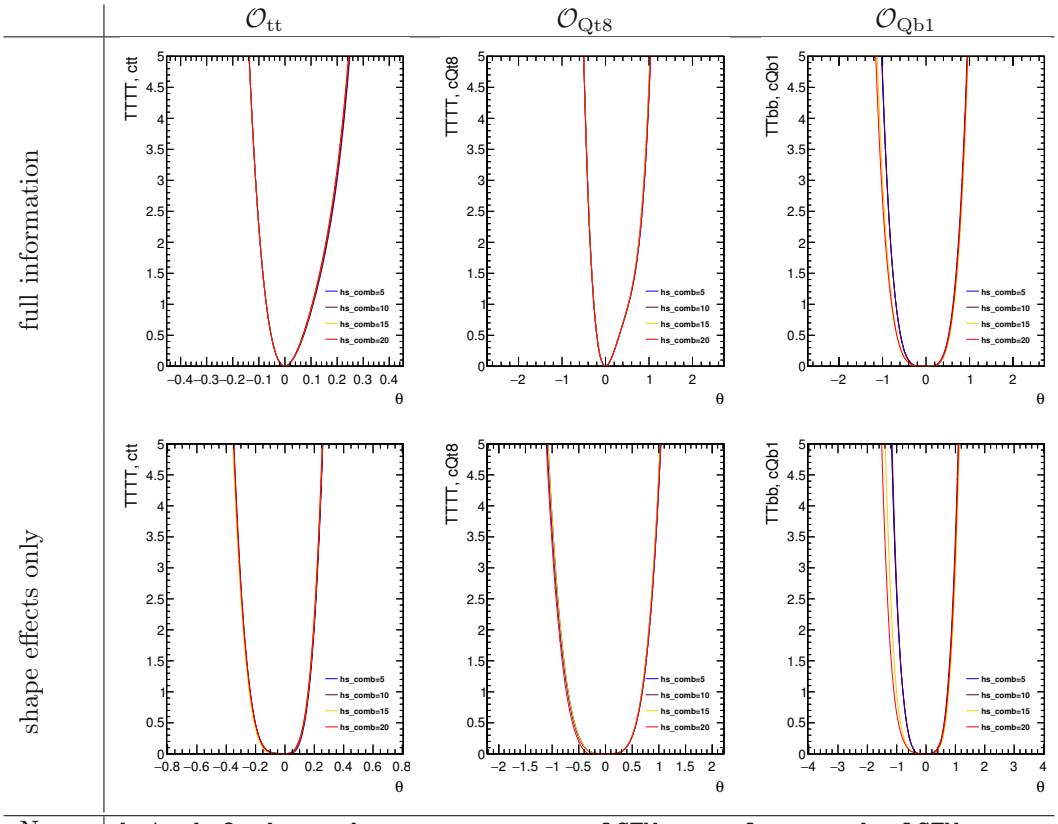
Last for the DNN, we examine the optimal size of the third dense layer in case no LSTMs are added, or, if LSTMs are added in a later step, the hidden size of the combined layer for the concatenated DNN and LSTM output, `hs_comb`. This corresponds to the training setups 9-12 in Tab. 6.4. For $\mathcal{O}_{\mathrm{tt}}$ and $\mathcal{O}_{\mathrm{Qt8}}$ no effects connected to variations in `hs_comb` are visible. In $\mathcal{O}_{\mathrm{Qb1}}$, however, performance decreases significantly with `hs_comb` being too large, especially when reduced to the shape effects only. Hence, the optimal configuration requires the last hidden layer to be relatively small, and just slightly larger than the training target length.

Finally, the output size of the last layer is not subject to opimization, as it is fixed by the target length, i.e. 2 for simulation-based inference – $R_{\mathrm{lin}}, R_{\mathrm{quad}}$ – and 4 for multi-classification – the probability for $t\bar{t}t\bar{t}$, $t\bar{t}b\bar{b}$, $t\bar{t}c\bar{c}$ or $t\bar{t}$+light jets. With respect to our DNN layers, it should be noted that the network is quite prone to overtraining. To prevent this, dropout layers and activation functions are used as described in detail in the previous chapter 5. In this respect, a dropout probability of 0.5 paired with a negative slope of 0.5 in the LeakyReLU has shown to be adequate.



No.	hs1	hs2	hs_comb	LSTM	num_layers	hs_LSTM
5-8	80	30 / 35 / 45 / 50	5	-	-	-

Table 6.3: Hyperparameter optimization: hidden size of the second dense layer (`hs2`)



No.	hs1	hs2	hs_comb	LSTM	num_layers	hs_LSTM
9-12	80	45	5 / 10 / 15 / 20	-	-	-

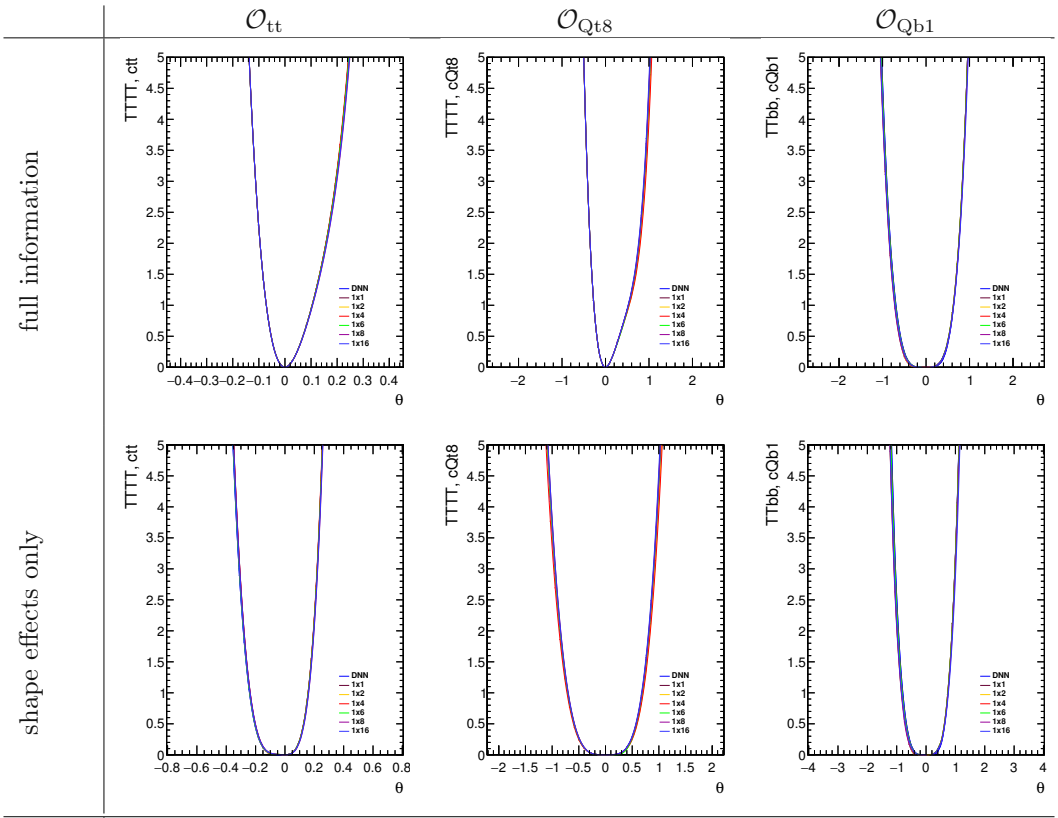
Table 6.4: Hyperparameter optimization: hidden size of the combined layer (`hs_comb`)

The hyperparameter optimization has been performed with the BSM effects and simulation-based inference. However, the results are assumed to be valid also for multi-classification, because of the almost identical network structure. As the network for multi-classification learns much faster and converges quickly, the qualitative learning rate adjustment in Fig. 5.8 of chapter 5 is maintained, but the number of epochs can be reduced to 1/10 of the total epochs stated there, i.e., `n_epochs`=1000.

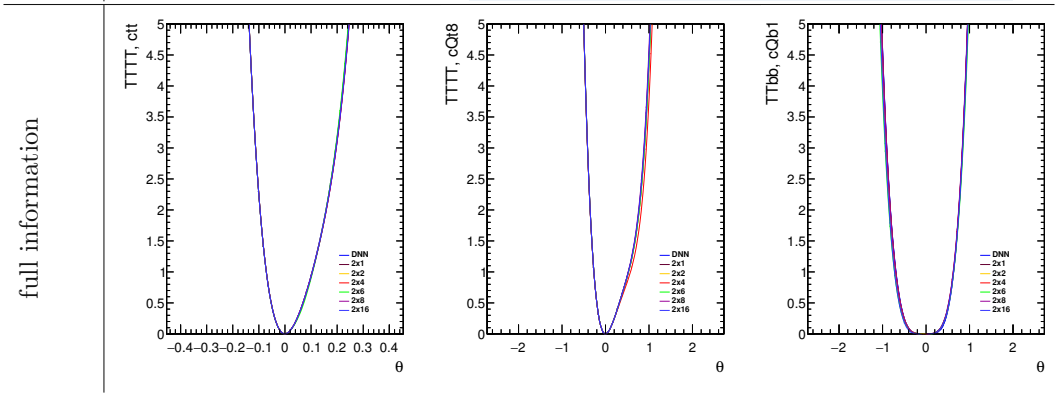
6.2.2 Hyperparameter optimization LSTM

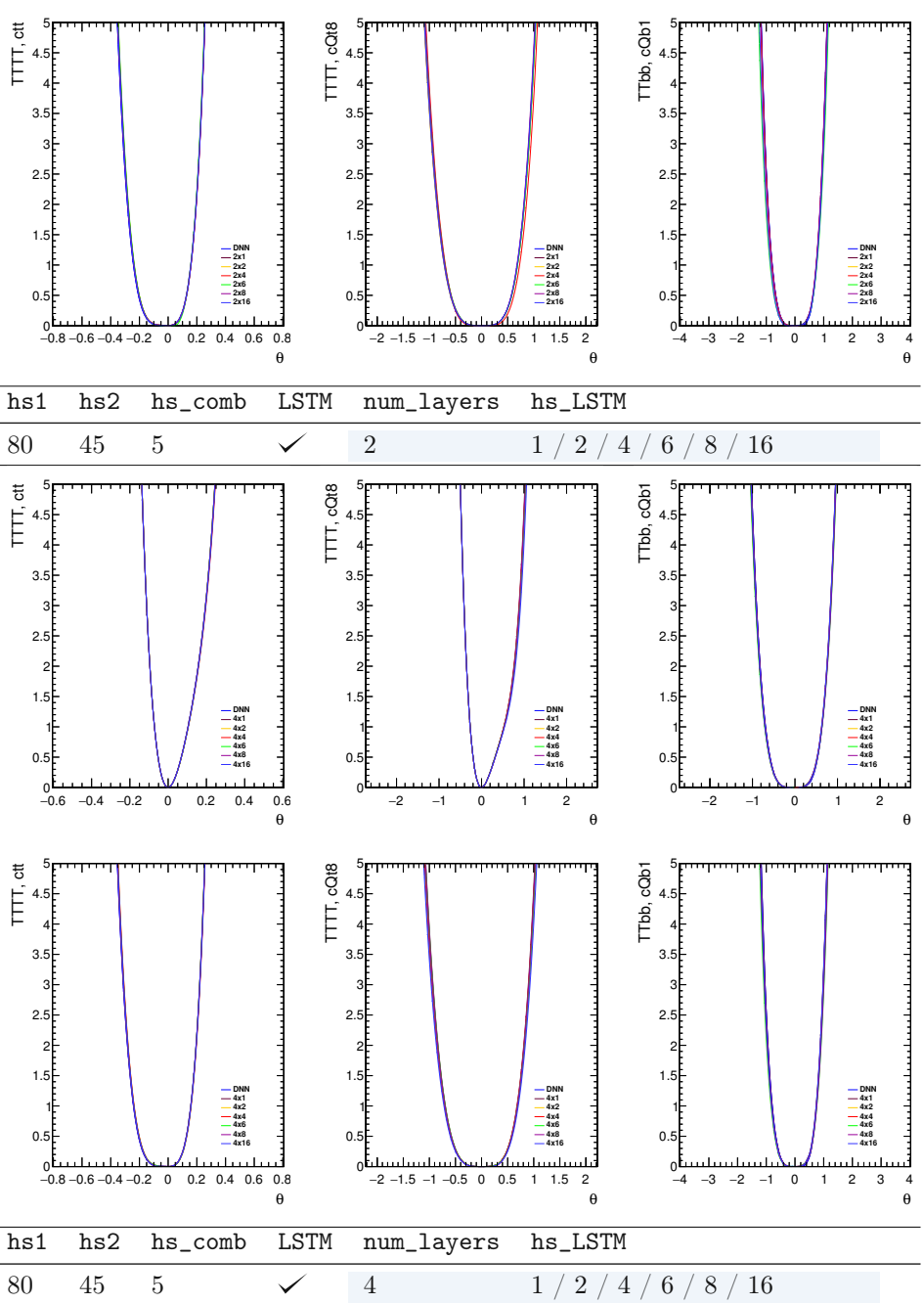
For the LSTMs, we try to optimize performance by training with different configurations of the two main tunable parameters: the number of layers and the hidden size of the LSTM. We identify the combinations of these two parameters with the syntax `num_layers` × `hs_LSTM` in the plots on the following pages.

Contrarily to our assumption, the LSTM configuration does not add any gain in performance compared to the DNN only setup. This is confirmed by the LLR on the follwing two pages in Tab. 6.5 for different number of layers and different hidden sizes of the LSTM. We test a broad variety of configurations to rule out that the LSTM's additional value might be limited to dedicated configurations only. Additionally, we always consider the LLR with full trained information and the LLR with shape effects only, in case the gains of an LSTM might be visible only in one of the settings.



No.	hs1	hs2	hs_comb	LSTM	num_layers	hs_LSTM
13-18	80	45	5	✓	1	1 / 2 / 4 / 6 / 8 / 16





No.	hs1	hs2	hs_comb	LSTM	num_layers	hs_LSTM
19-24	80	45	5	✓	2	1 / 2 / 4 / 6 / 8 / 16
25-30	80	45	5	✓	4	1 / 2 / 4 / 6 / 8 / 16

Table 6.5: Hyperparameter optimization for the LSTM (`num_layers`×`hs_LSTM`)

The reason for this behaviour is to be investigated. Possible explanations are listed below:

1. When training with signal only, the DNN fully extracts all available information. Hence, no improvement is seen for additional LSTMs. If this were the case, training with backgrounds must behave differently. No systematic parameter study has been conducted in this respect. However, training in different operators with different LSTM configuration for all two signal processes with additional $t\bar{t}$ background has not shown any hints for this hypothesis to be true. As an example, the LLRs for three trainings with background are shown in Fig. 6.8.
2. The additional LSTMs are not configured properly. If this were the case, the additional LSTMs will not enhance performance in multi-classification either. This hypothesis is to be investigated in the following subsection.
3. The LSTMs are configured properly but the input data is not sufficient. The limitations of the DELPHES parametrization of the detector limit the expressivity of the data. If this were the case, LSTMs in multi-classification should significantly enhance the network's performance, as they get a considerably larger set of input features.

In the following subsection, we use our multi-classification setup with an identical neural network to reject the second and back the third hypothesis.

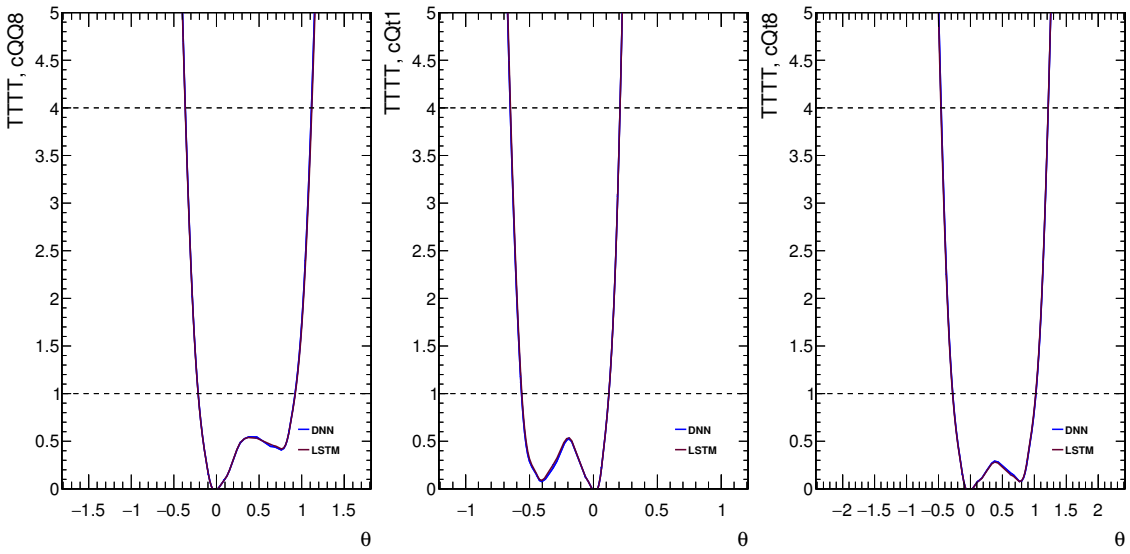


Figure 6.8: LSTMs do not improve the performance with respect to the DNN only configuration in signal+background training. Here, we used 1 LSTM layer of hidden size 20 for the operators $\mathcal{O}_{\mathrm{Qt8}}$, $\mathcal{O}_{\mathrm{Qt1}}$ and $\mathcal{O}_{\mathrm{QQ8}}$ in $t\bar{t}t\bar{t}$.

6.2.3 LSTM and multi-classification

We start by comparing LSTM configurations for different number of layers and hidden sizes. To evaluate the performance, we first make a histogram of the output probabilities for the four different categories when samples of a single category are fed into the trained network. An example of such a histogram can be seen in Fig. 6.9, where the trained network has been fed with all four samples and the probability for $t\bar{t}t\bar{t}$ in each individual sample category has been retrieved. From the histograms for signal and background, we then compute the ROC-curve with the training output.

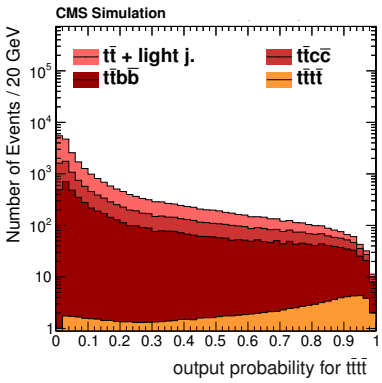
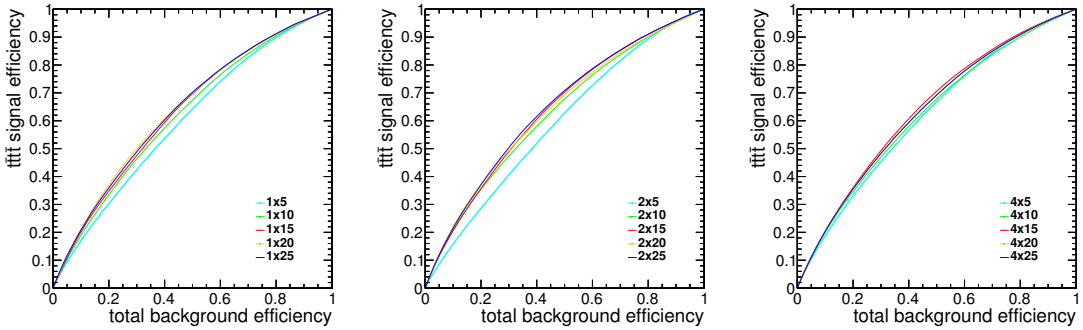


Figure 6.9: Histogram of the neural network's output for the $t\bar{t}t\bar{t}$ probability.



hs1	hs2	hs_comb	LSTM	num_layers	hs_LSTM
70	40	5	✓	1 / 2 / 4	5 / 10 / 15 / 20 / 25

Table 6.6: Hyperparameter optimization: LSTM configuration in multi-classification

From the shapes we deduce that the hidden size of the LSTM paired with the DNN default configuration[1] ought to be 20 or 25. Comparing configurations of hidden size 20 shows that stacking multiple layers does not improve the network's performance, as all curves in Fig. 6.10 perfectly overlap. The optimal LSTM configuration for multi-classification is therefore `num_layers`=1 and `hs_LSTM`=20 or 25.

What is left to investigate is the performance of a network with an additional LSTM layer compared to a DNN-only set up. With the optimal DNN configuration from the previous hyperparameter optimization, the enhancement of the network due to the LSTM is significant. This can be seen by comparing the LSTM's red ROC-curve in Fig. 6.11 to the DNN's blue shape. Hence, we deduce that the LSTMs are implemented correctly and that they are able to extract information beyond the DNN.

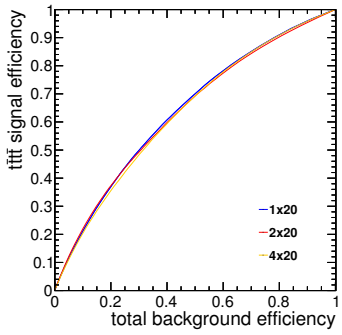
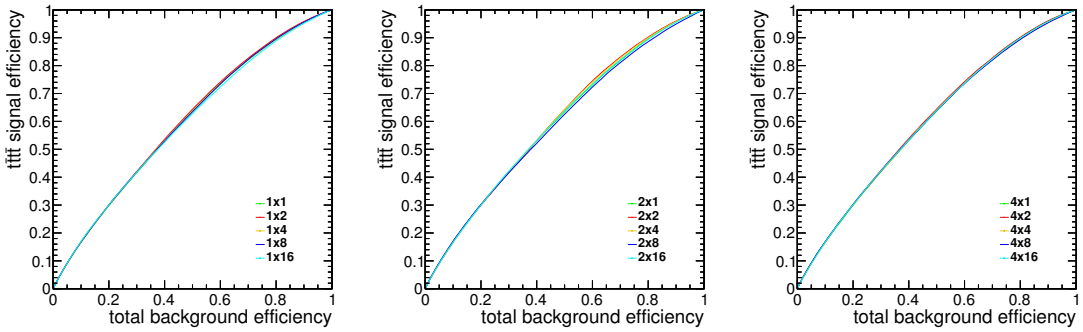


Figure 6.10: Roc curves for LSTMs with different number of layers (1/2/4) and hidden size 20.

[1]The default values for `hs1` is 80 in simulation-based inference, which corresponds to 2× the number of scalar event-level input features (=40). In multi-classification, the number of scalar event-level input features is 35, hence `hs1` $= 2 \times 35 = 70$. Similarly, the optimal `hs2` equals the number of DNN input features +5.

However, this is not what we have found for simulation-based inference. To understand what the reason behind this might be, we compare the inputs of the LSTMs for inference and multi-classification, respectively. On the one hand, due to the limited DELPHES parametrization of the detector we have only four kinematic observables to feed the LSTM jets in inference. The number of inputs for classification, on the other hand, is significantly higher and comprises observables that are not (already partly) present in the DNN variables. Hence, it is reasonable to assume that we simply do not feed our LSTMs enough information beyond what the DNN already knows.

To back this claim, we deliberately restrict the multi-classification input to the kinematic variables p_T, η and ϕ and train the LSTM again. As we cannot deduce the optimal LSTM output size for this case – the number of LSTM inputs drastically changed – we again try different hidden sizes. The configurations and results are summarized in Tab. 6.7.



hs1	hs2	hs_comb	LSTM	num_layers	hs_LSTM
70	40	5	✓	1 / 2 / 4	1 / 2 / 4 / 8 / 16

Table 6.7: Hyperparameter optimization: Reduced LSTM configuration in multi-classification

The results are clear: The different configurations all yield the same output, and what is more, do not further improve the network's performance. In fact, in Fig. 6.11 below, the roc curves for DNN only and DNN + "reduced" LSTM perfectly overlap. Hence, we conclude that our third hypothesis is correct.

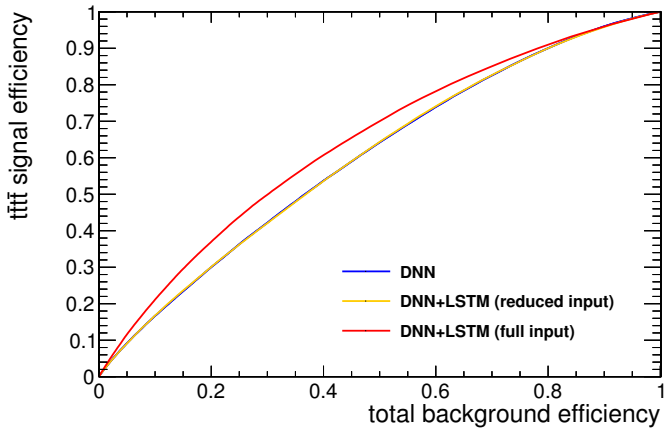


Figure 6.11: Comparison of DNN, DNN+LSTM (reduced features) and DNN+LSTM (full features). The performance of the latter is significantly better, whereas the first two are identical.

Chapter 7

Results

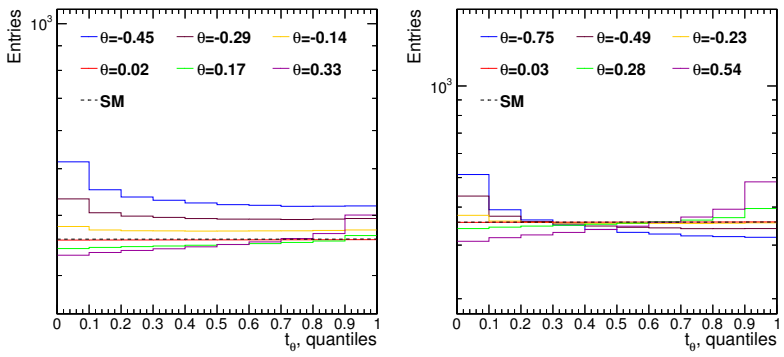
7.1 Training for signal only

We start by computing the LLR for 1D training with single operator insertions without additional $t\bar{t}$ background. As described in chapter 4, a first hint for successful training is the condition

$$\sum_{\mathbf{x_i}\in\text{bin}} w_{0,i} R_{\text{lin(quad)}}(\mathbf{x_i}) \stackrel{?}{=} \sum_{\mathbf{x_i}\in\text{bin}} w_{1(2),i}, \tag{7.1}$$

i.e., convergence in the bins of scalar event-level observables. Training the network as described in the chapters 5 and 6 adequately meets this condition. An exemplar collection of such histograms of $R_{\text{lin(quad)}}(\mathbf{x})$ in bins of the DNN input features is shown in Fig. 7.2 on the next page for $\mathcal{O}_{\text{QQ1}}$ in $t\bar{t}t\bar{t}$.

To set limits, we compute the 1D and 2D log-likelihood ratios using (a) full information from training or (b) only the information regarding the shape. The two modes of evaluation differ in their normalization, which can be seen when we make histograms of the test statistic $t_{\boldsymbol{\theta}}$ for $\theta \neq 0$, i.e., not at the SM point where the test statistic is flat under the respective binning choice. Again for $\mathcal{O}_{\text{QQ1}}$ in $t\bar{t}t\bar{t}$ as an example, histograms for the two modes are shown in Fig. 7.1. For (a), the total number of events is modified because of the operator insertion, for (b), the operator influences only the shape with respect to the SM, keeping the sum of all per-bin yields constant.



(a) full information **(b)** shape effects only

Figure 7.1: The transformed test statistic for $\mathcal{O}_{\text{QQ1}}$ in $t\bar{t}t\bar{t}$ according to the procedure described in Chapter 4, section 4.2. The bins are then used in expression Eq. 4.41. The values of $\boldsymbol{\theta}$ for which the test statistic has been calculated differ between the left and right figure. This is a deliberate choice to better adapt to the most sensitive $\boldsymbol{\theta}$-range according to the LLR shapes.

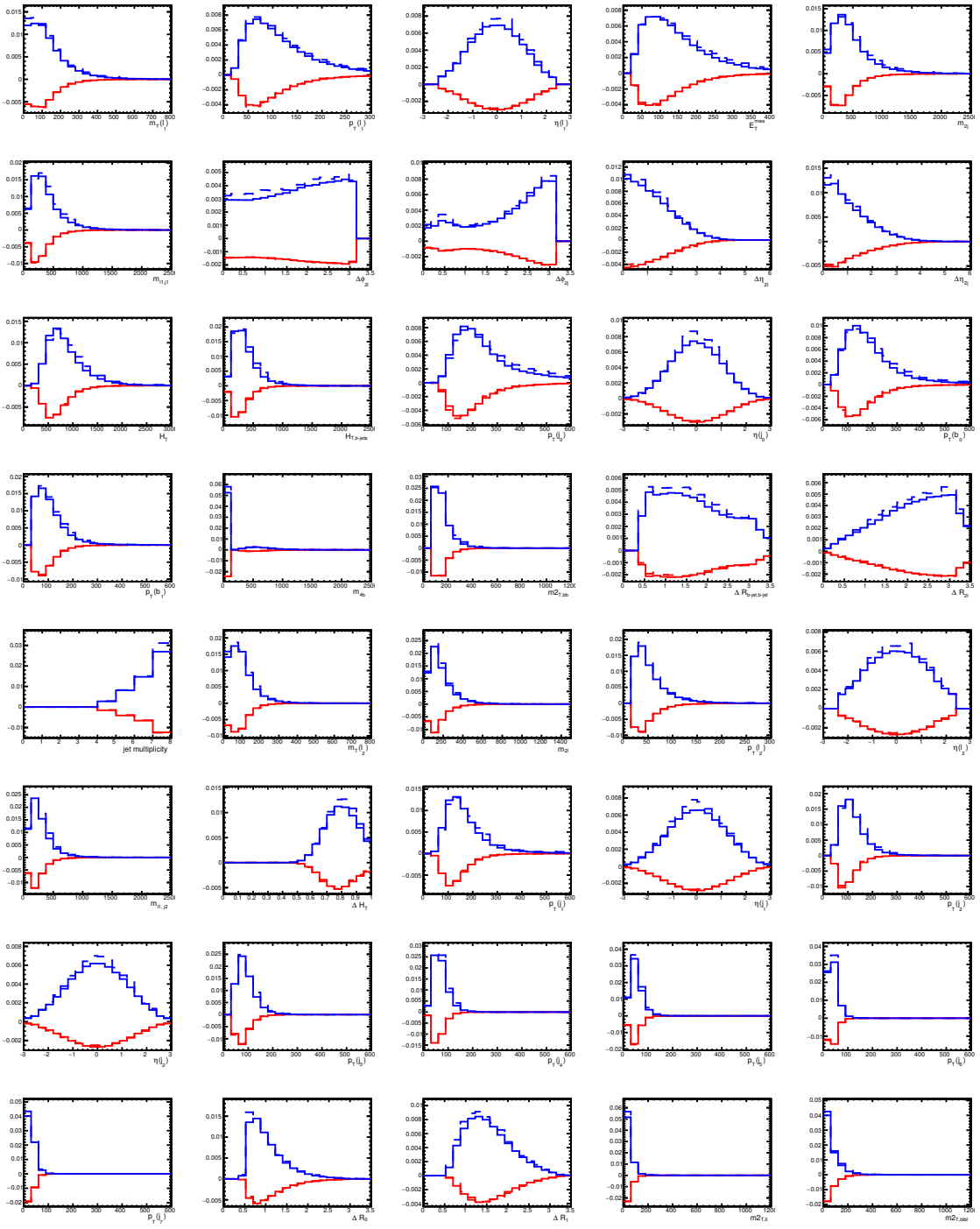


Figure 7.2: Convergence in bins of all scalar event-level observables for $\mathcal{O}_{\mathrm{QQ1}}$ in $\mathrm{t\bar{t}t\bar{t}}$. The solid red line denotes the left side of Eq. 7.1 for R_{lin}, i.e., the training output, whereas the dashed red line marks the right side, i.e., the truth. Blue lines are the equivalent for R_{quad}.

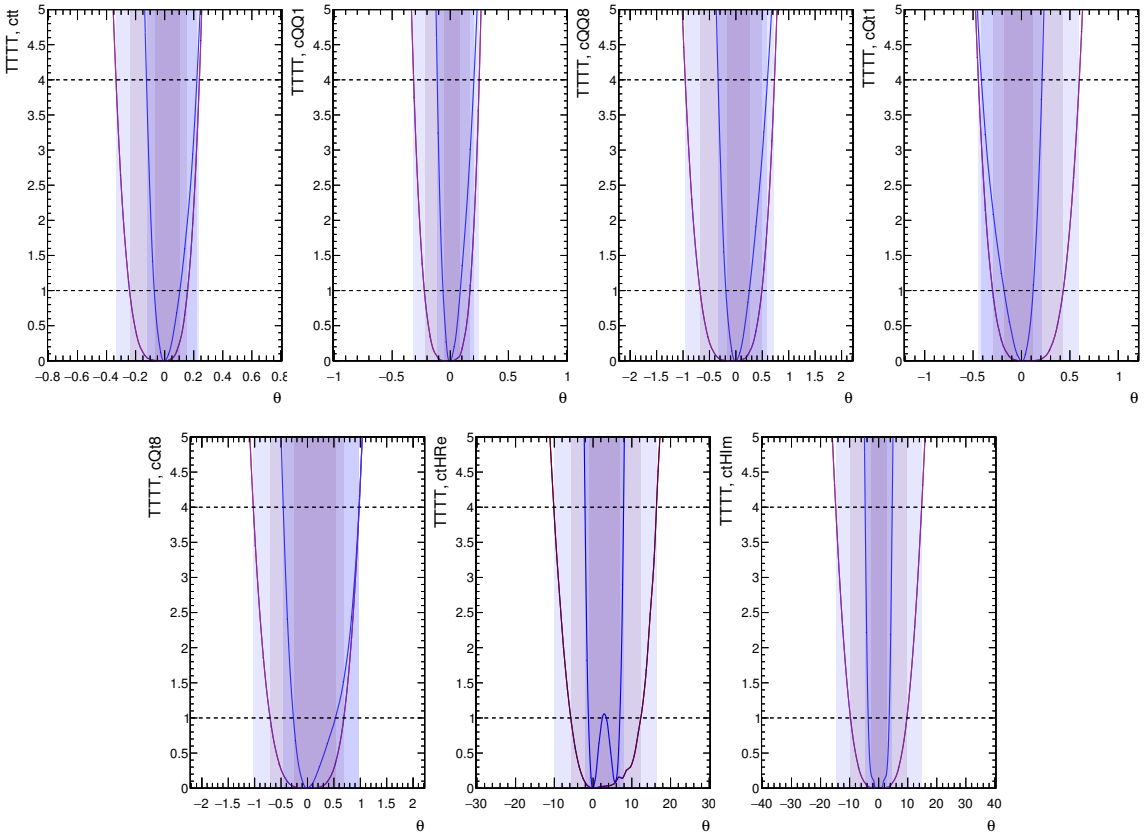


Figure 7.3: 1D LLR for all operators in $t\bar{t}t\bar{t}$, signal only. The shapes have been calculated using the **full information** or **only the shape information** about the EFT effects learned in the training. These effects affect both variations in the yield per bin and changes in the shape of the pdf $p(\boldsymbol{x}|\boldsymbol{\theta})$ with respect to the SM pdf $p(\boldsymbol{x}|SM)$. The blue and purple boxes indicate 2σ and 1σ according to Tab. 4.2, respectively. The network has been trained with **signal only**.

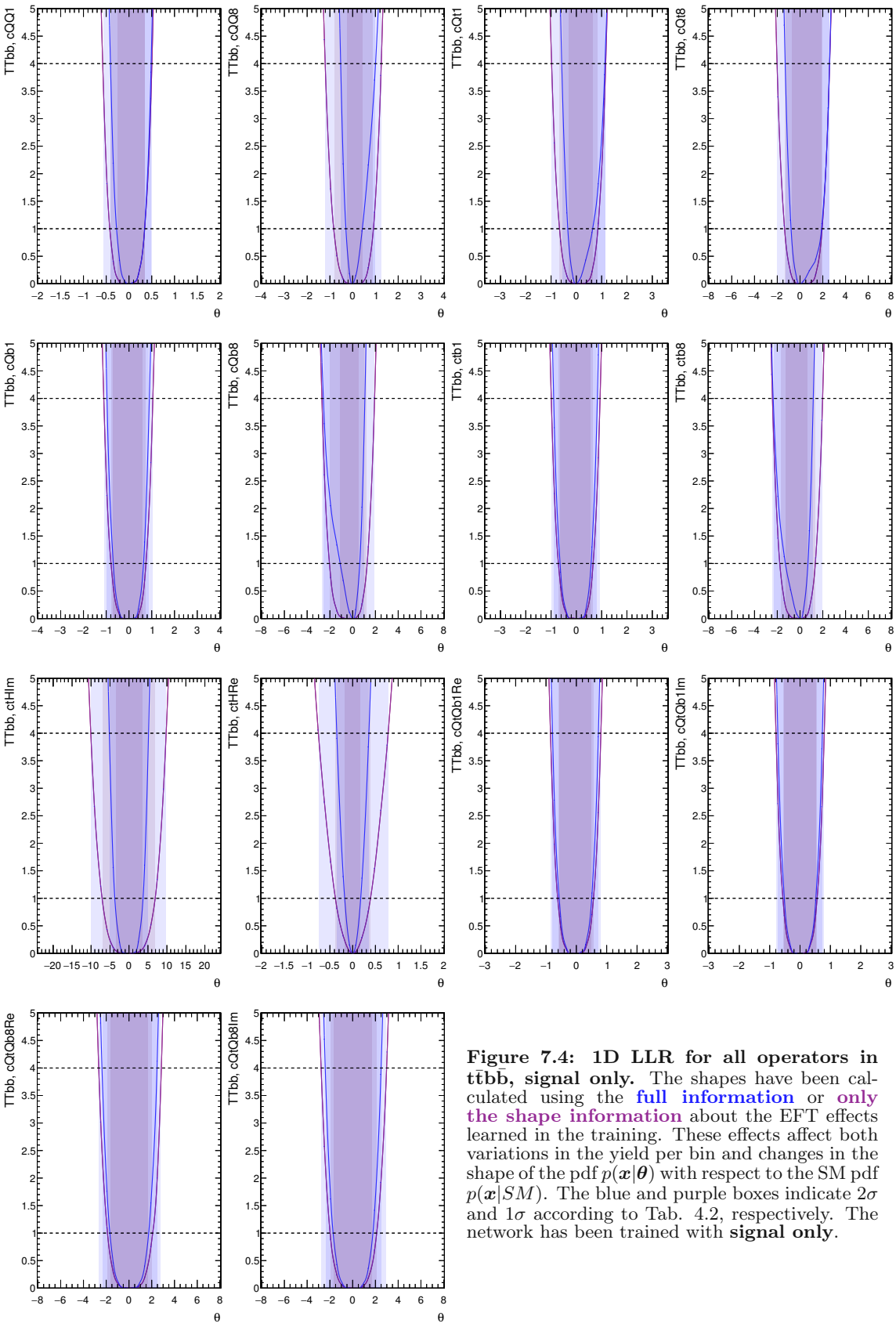


Figure 7.4: 1D LLR for all operators in $t\bar{t}b\bar{b}$, signal only. The shapes have been calculated using the **full information** or **only the shape information** about the EFT effects learned in the training. These effects affect both variations in the yield per bin and changes in the shape of the pdf $p(\boldsymbol{x}|\boldsymbol{\theta})$ with respect to the SM pdf $p(\boldsymbol{x}|SM)$. The blue and purple boxes indicate 2σ and 1σ according to Tab. 4.2, respectively. The network has been trained with **signal only**.

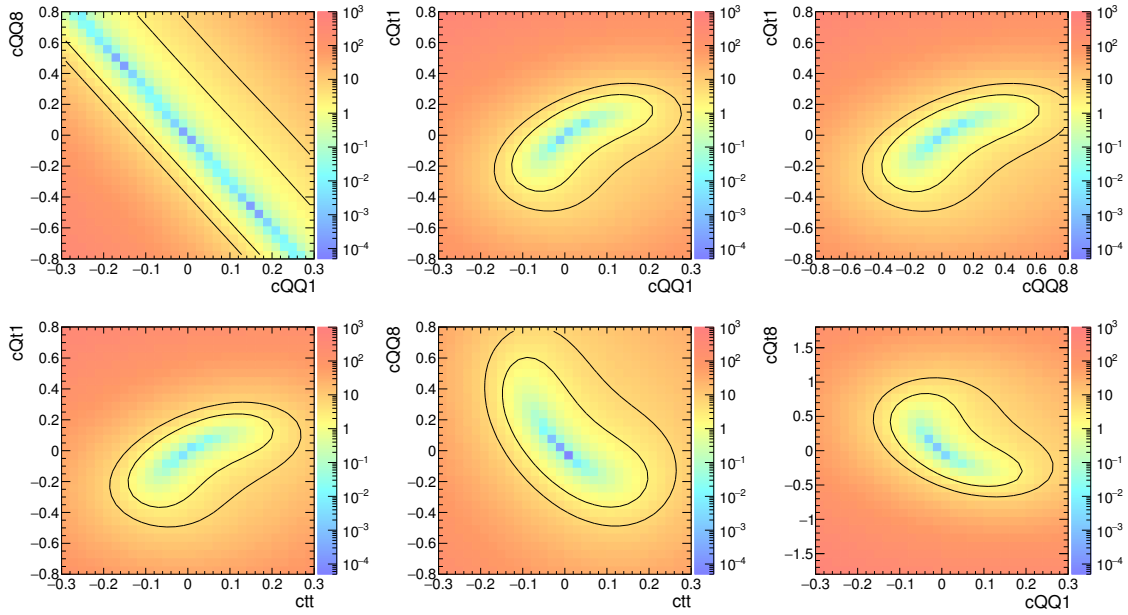


Figure 7.5: 2D LLR (full information) for selected operators in $t\bar{t}t\bar{t}$, signal only. The 2D limits above have been calculated using the **full information** about the EFT effects learned in the training for selected operator combinations in $t\bar{t}t\bar{t}$.

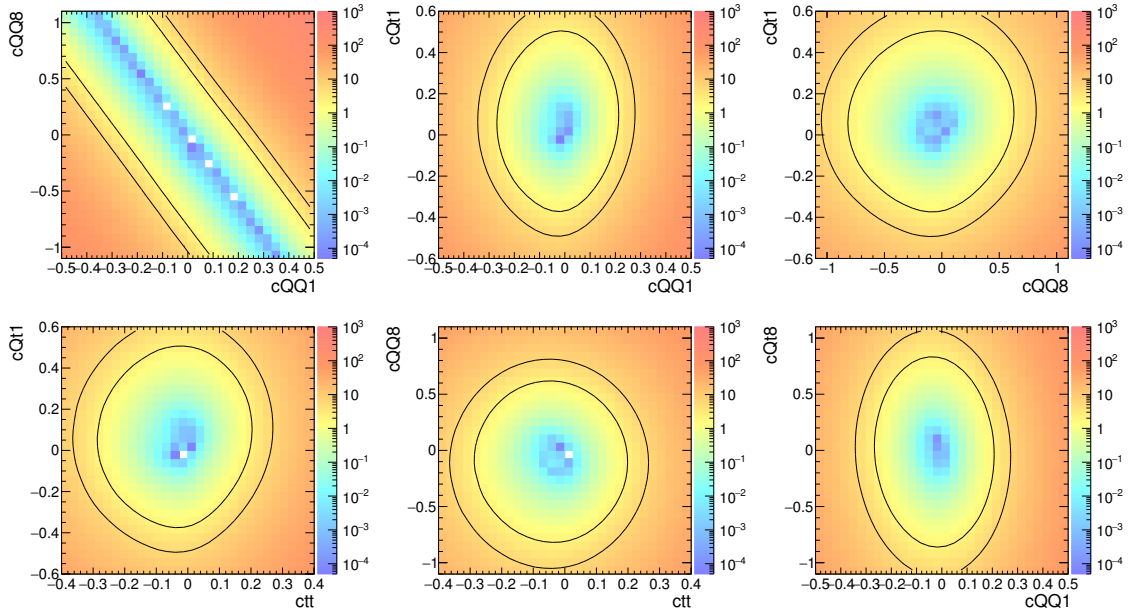


Figure 7.6: 2D LLR (shape effects only) for selected operators in $t\bar{t}t\bar{t}$, signal only. The 2D limits below have been calculated using **only the shape information** from the EFT effects learned in the training for selected operator combinations in $t\bar{t}t\bar{t}$.

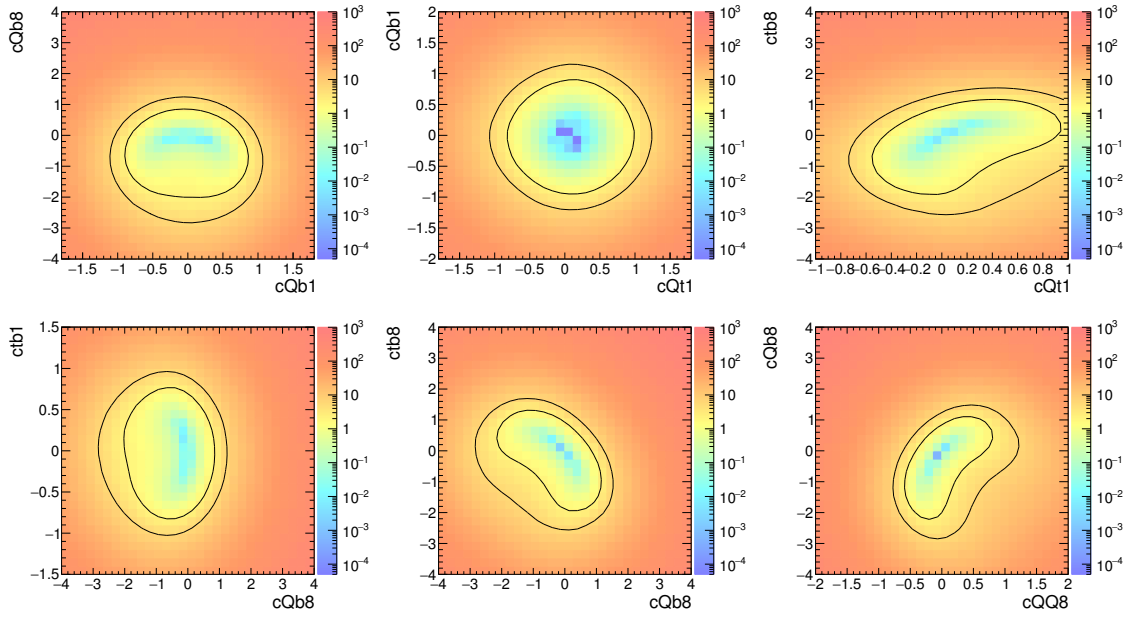


Figure 7.7: 2D LLR (full information) for selected operators in $t\bar{t}b\bar{b}$, signal only. The 2D limits above have been calculated using the **full information** about the EFT effects learned in the training for selected operator combinations in $t\bar{t}b\bar{b}$.

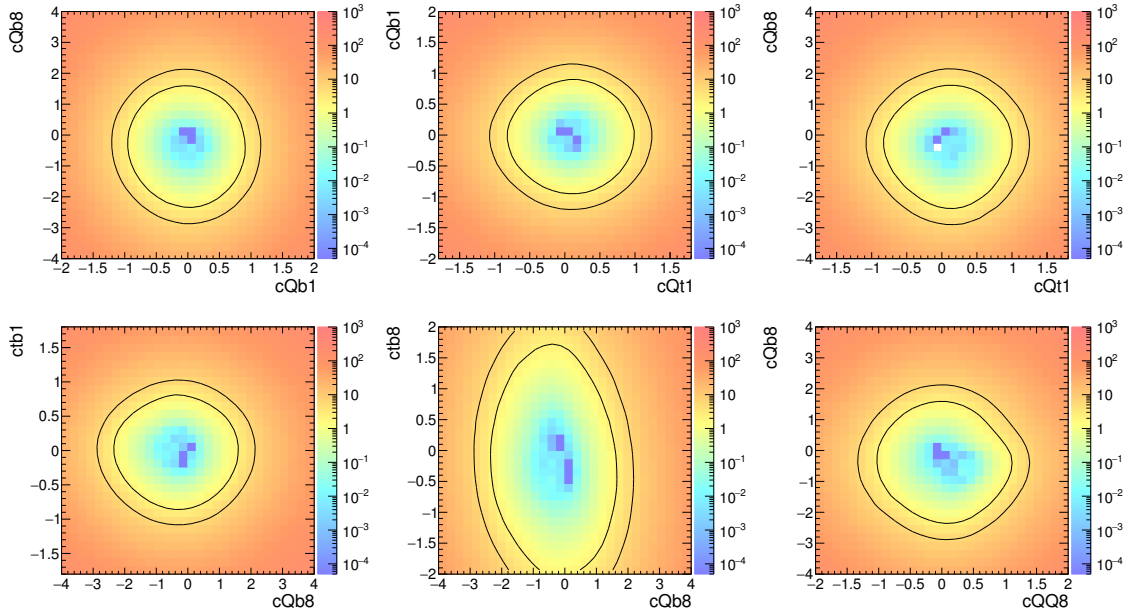


Figure 7.8: 2D LLR (shape effects only) for selected operators in $t\bar{t}b\bar{b}$, signal only. The 2D limits below have been calculated using **only the shape information** from the EFT effects learned in the training for selected operator combinations in $t\bar{t}b\bar{b}$.

7.2 Training with signal and background

Finally, we train selected operators using a combination of signal and background samples, i.e., the ones with the greatest effects on the shapes. The background consists of $t\bar{t}$ events, whereas the signals are again $t\bar{t}t\bar{t}$ and $t\bar{t}b\bar{b}$, respectively. The event selections are performed as described in section 6.1.2 in chapter 6.

It is important to note that we do not tell the machine about the signal and background events. Instead, we let it treat every event exactly the same, except from the SM reference weight $w_{i,0}$ we assign to every event in the loss function. This weight is related to the cross section of the process, and we normalize to the integrated luminosity of $\mathcal{L} = 300$ fb^{-1},

$$w_{i,\mathrm{SM}}^{\mathrm{sig}} = \frac{\mathcal{L}\sigma_{\mathrm{SM}}^{\mathrm{sig}} w_{i,0}^{\mathrm{sig}}}{\sum_j w_{j,0}^{\mathrm{sig}}} \tag{7.2}$$

for our DELPHES signal samples $t\bar{t}t\bar{t}$ and $t\bar{t}b\bar{b}$. For the $t\bar{t}$ background simulated with GEANT, the events are also assigned per-event weights that model interference effects from next-leading order (NLO) calculations and can also be negative.

The histograms in bins of scalar event-level observables in Fig. 7.10 on the next page and the computation of the 1D and 2D LLRs are then just identical to the training setting without background.

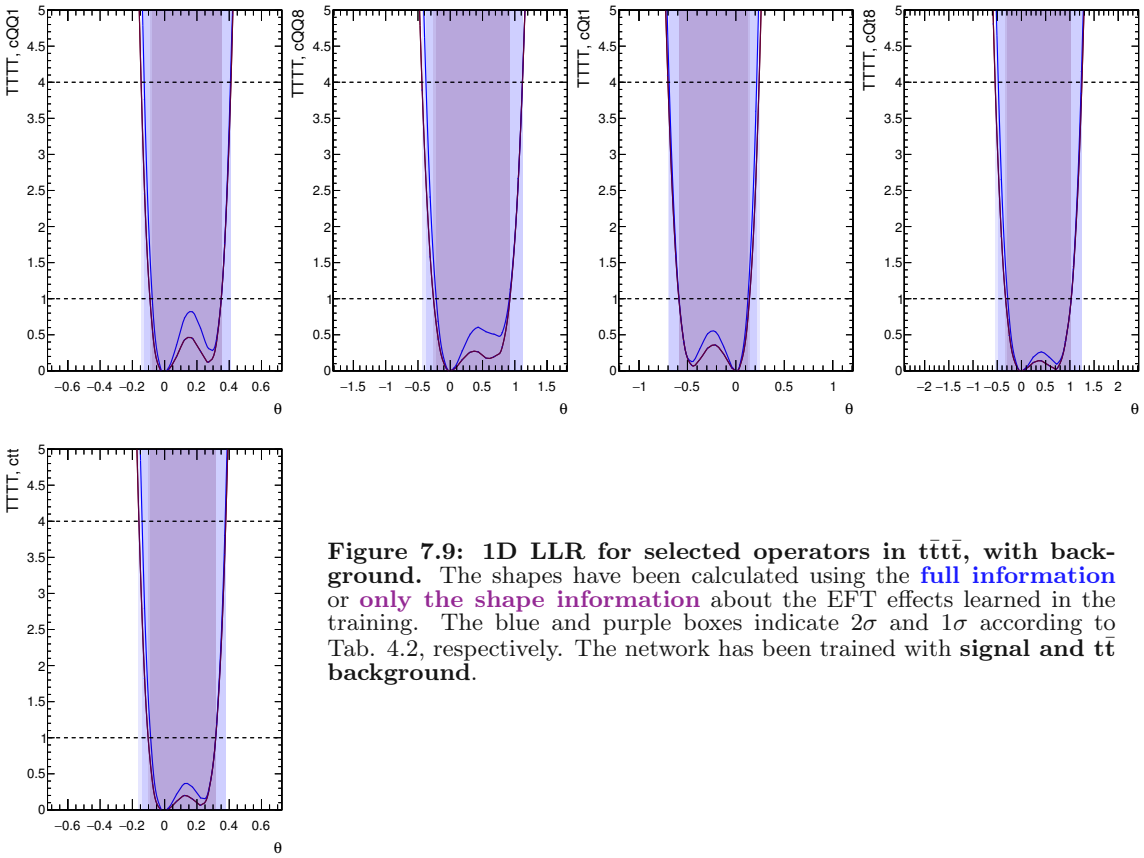


Figure 7.9: 1D LLR for selected operators in $t\bar{t}t\bar{t}$, with background. The shapes have been calculated using the **full information** or **only the shape information** about the EFT effects learned in the training. The blue and purple boxes indicate 2σ and 1σ according to Tab. 4.2, respectively. The network has been trained with **signal and $t\bar{t}$ background**.

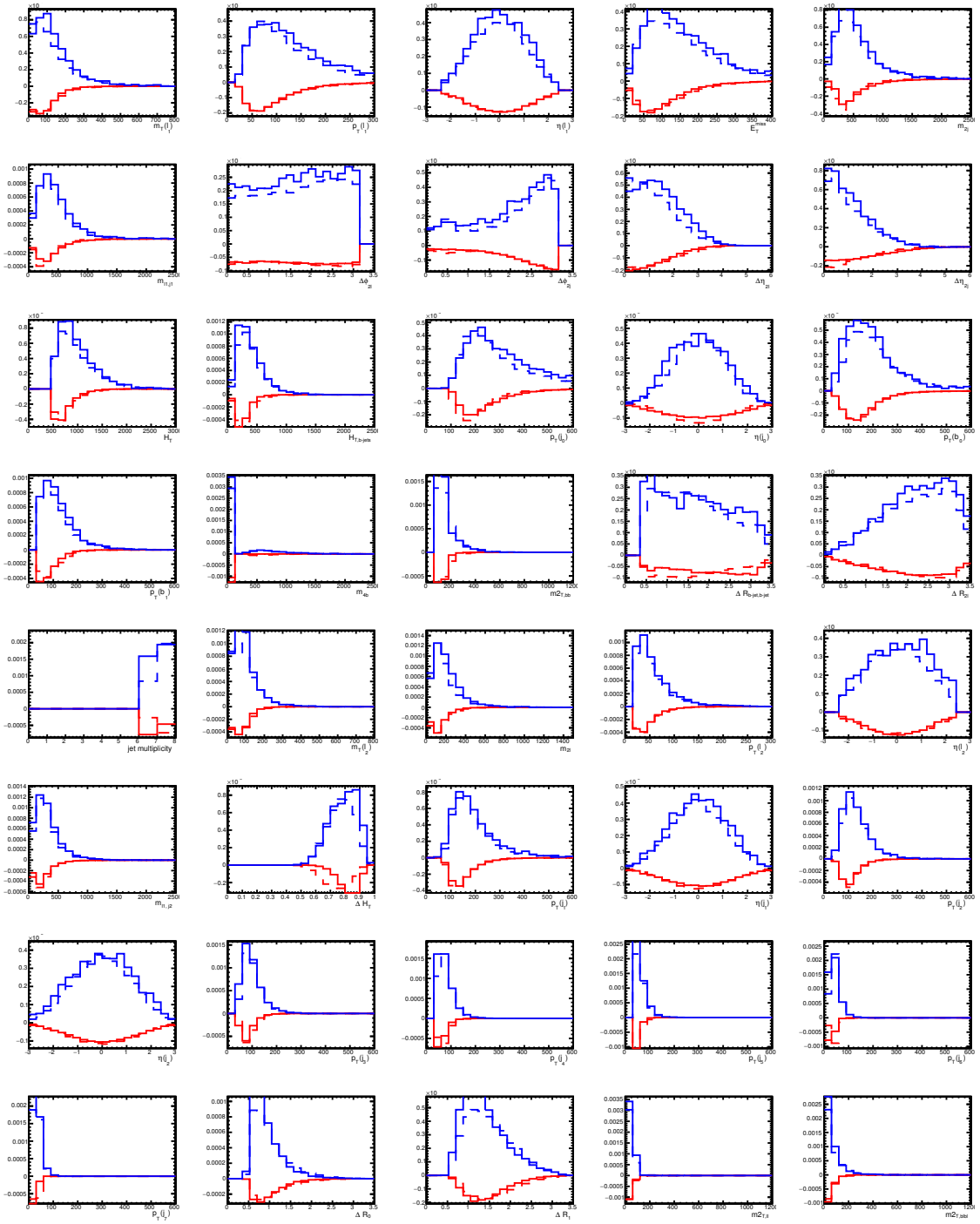


Figure 7.10: Convergence in bins of all scalar event-level observables for $\mathcal{O}_{\mathrm{QQ1}}$ in $\mathrm{t\bar{t}t\bar{t}}$ with $\mathrm{t\bar{t}}$ background. The solid red line denotes the left side of Eq. 7.1 for R_{lin}, i.e., the training output, whereas the dashed red line marks the right side, i.e., the truth. Blue lines are the equivalent for R_{quad}.

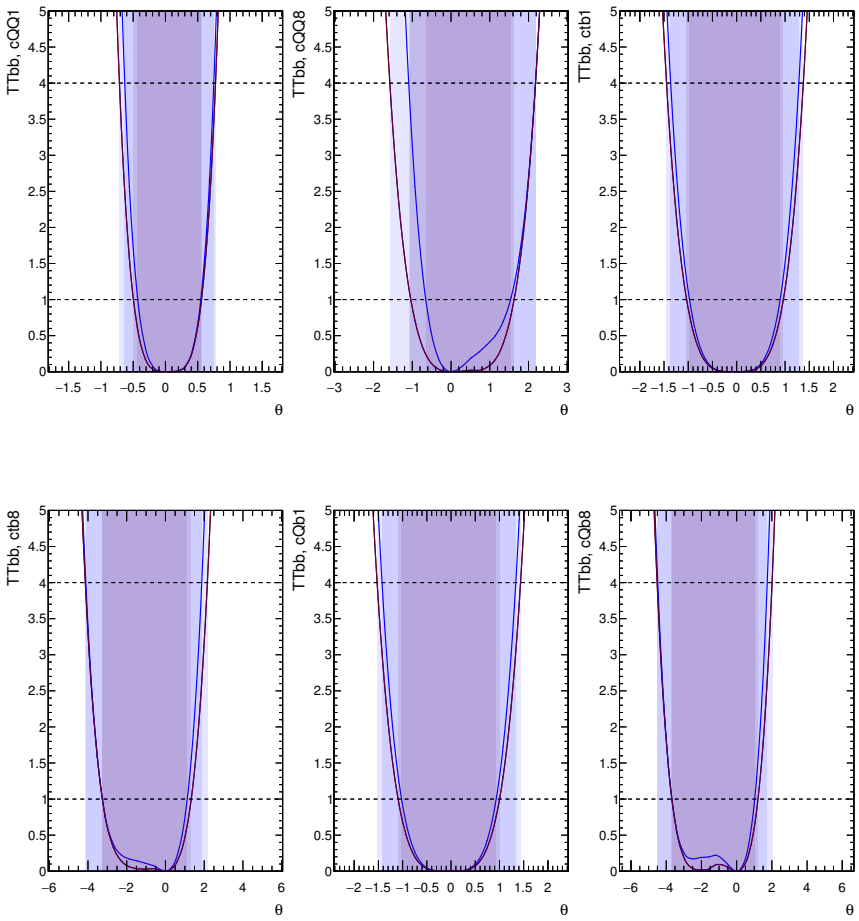


Figure 7.11: 1D LLR for selected operators in $t\bar{t}b\bar{b}$, with background. The shapes have been calculated using the **full information** or **only the shape information** about the EFT effects learned in the training. The blue and purple boxes indicate 2σ and 1σ according to Tab. 4.2, respectively. The network has been trained with **signal and $t\bar{t}$ background**.

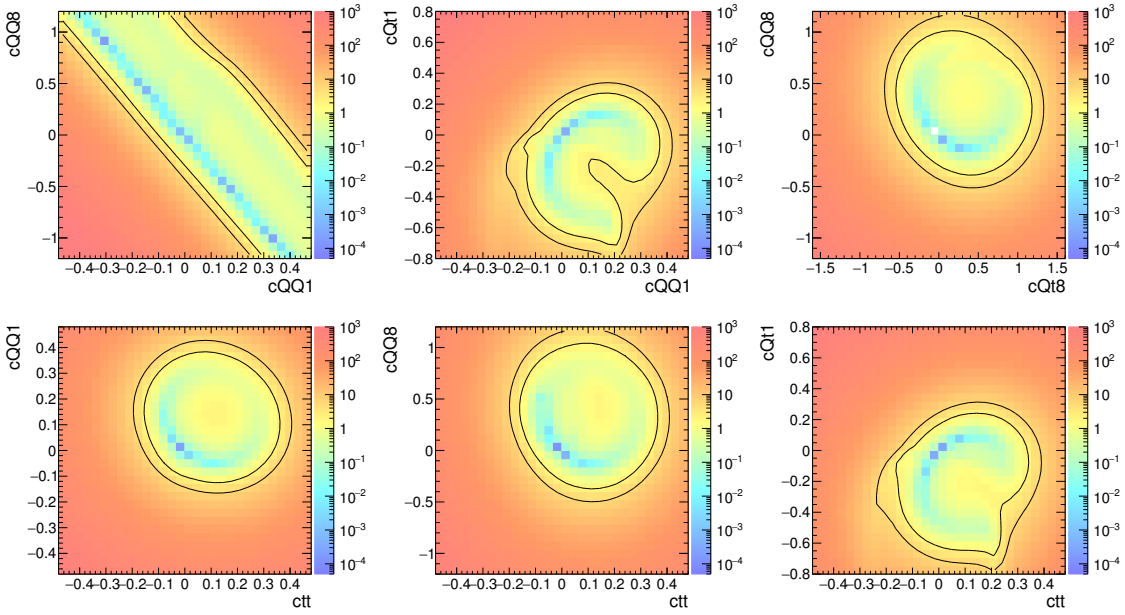


Figure 7.12: 2D LLR (full information) for selected operators in $t\bar{t}t\bar{t}$ with $t\bar{t}$ background. The 2D limits above have been calculated using the **full information** about the EFT effects learned in the training for selected operator combinations in $t\bar{t}t\bar{t}$

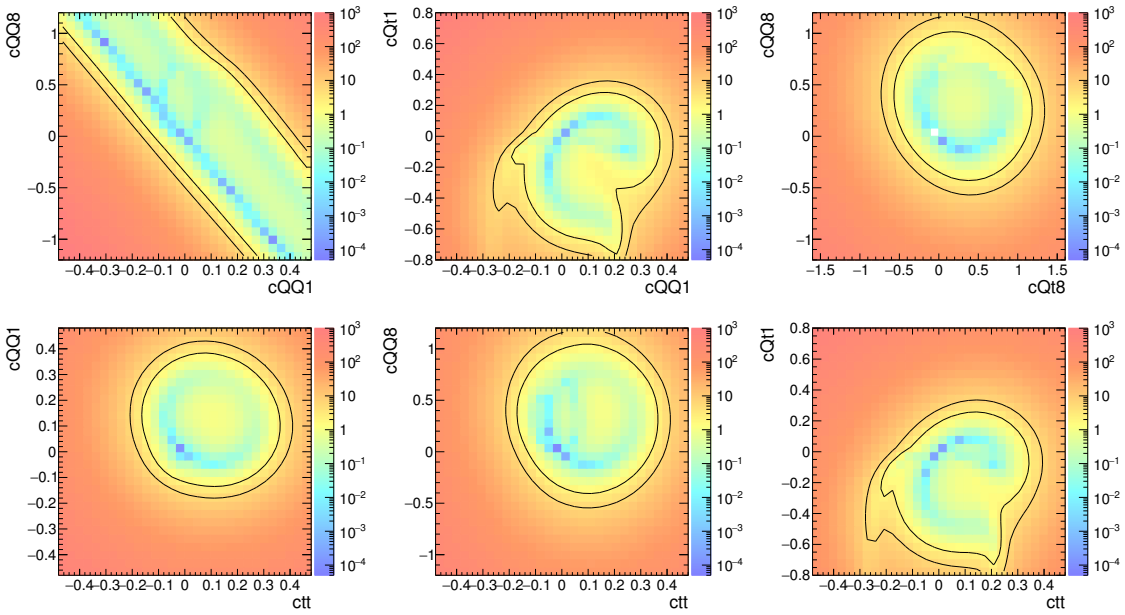


Figure 7.13: 2D LLR (shape effects only) for selected operators in $t\bar{t}t\bar{t}$ with $t\bar{t}$ background. The 2D limits below have been calculated using **only the shape information** from the EFT effects learned in the training for selected operator combinations in $t\bar{t}t\bar{t}$

.

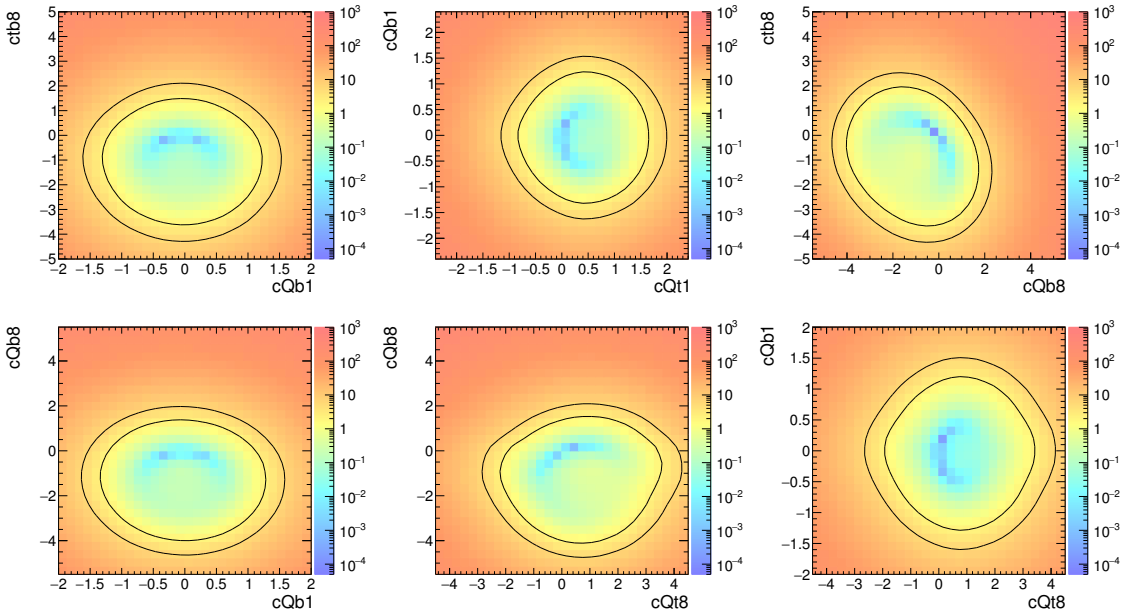


Figure 7.14: 2D LLR (full information) for selected operators in $t\bar{t}b\bar{b}$ with $t\bar{t}$ background. The 2D limits above have been calculated using the **full information** about the EFT effects learned in the training for selected operator combinations in $t\bar{t}b\bar{b}$.

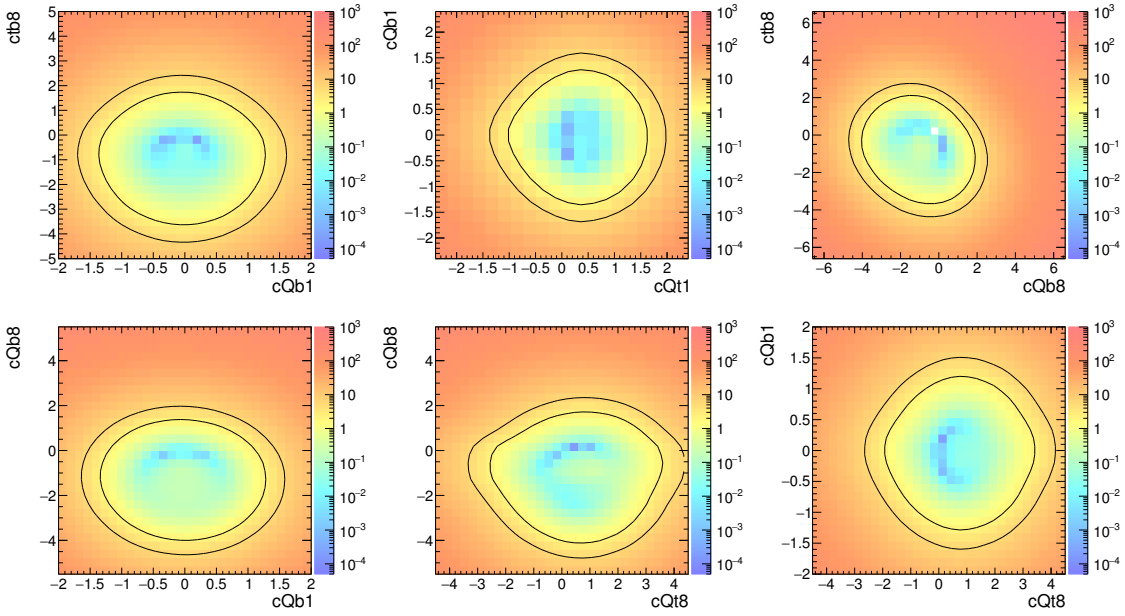


Figure 7.15: 2D LLR (shape effects only) for selected operators in $t\bar{t}b\bar{b}$ with $t\bar{t}$ background. The 2D limits below have been calculated using **only the shape information** from the EFT effects learned in the training for selected operator combinations in $t\bar{t}b\bar{b}$.

7.3 Summary of the limit setting

Looking at the 1D and 2D limits on the Wilson coefficients on the previous pages, we see that we can effectively constrain a large amount of four-fermion operators in $t\bar{t}t\bar{t}$ and $t\bar{t}b\bar{b}$ with and without backgrounds. Hereinafter, we will briefly discuss peculiarities and characteristics of the LLRs just computed:

- Regarding the **general shape of the LLR contures in 1D**, we note that it follows a quartic polynomial form. This is to be expected, as the yield itself is a quadratic function in the EFT parameters and hence a quadratic polynomial in $\boldsymbol{\theta}$. In a Gaussian approximation of the 1D limits on the Wilson coefficients, one can then interpret the LLR shapes as one dimensional Gaussian confidence intervals. The deviation from the Gaussian parabola is due to the quadratic term in the EFT, that leads to the quartic shape.

- When **training and evaluating signal only in $t\bar{t}t\bar{t}$**, the yield information considerably improves the discriminative power of our parametrized classifier. When removing the yield variations in evaluation, performance decreases and the shapes display a more quartic behaviour. In **training and evaluating signal only in $t\bar{t}b\bar{b}$**, the effect is a little less pronounced for operators like, e.g., $\mathcal{O}_{\mathrm{tb1}}$, $\mathcal{O}_{\mathrm{Qb1}}$, $\mathcal{O}_{\mathrm{QtQb1}}$ and $\mathcal{O}_{\mathrm{QtQb8}}$.

- When **evaluating the 2D LLRs in $t\bar{t}t\bar{t}$ for the combination of the operators $\mathcal{O}_{\mathbf{QQ1}}$ and $\mathcal{O}_{\mathbf{QQ8}}$**, the shape displays a peculiar form. This is in accordance with the underlying theory structure of SMEFT for operators involving only left-handed heavy quarks. In fact, Ref. [90] gives the following relation between those operators and the color singlet and octets that would or would not interfere with the QCD amplitudes:

$$\begin{pmatrix} \mathcal{O}_{\mathrm{qq}}^{1(3333)} \\ \mathcal{O}_{\mathrm{qq}}^{3(3333)} \end{pmatrix} = \begin{pmatrix} 1 & -1/3 \\ 0 & 4 \end{pmatrix}^T \begin{pmatrix} (\bar{Q}\,\gamma_\mu\,Q)\,(\bar{Q}\,\gamma^\mu\,Q) \\ (\bar{Q}\,\gamma_\mu\,T^A\,Q)\,(\bar{Q}\,\gamma^\mu\,T^A\,Q) \end{pmatrix} \tag{7.3}$$

  When looking at our operators in Tab. 3.1, we can identify $(\bar{Q}\,\gamma_\mu\,Q)\,(\bar{Q}\,\gamma^\mu\,Q)$ with $2\mathcal{O}_{\mathrm{QQ1}}$ and $(\bar{Q}\,\gamma_\mu\,T^A\,Q)\,(\bar{Q}\,\gamma^\mu\,T^A\,Q)$ with $2\mathcal{O}_{\mathrm{QQ8}}$. Hence, in the 2D limit, we effectively constrain $\mathcal{O}_{\mathrm{qq}}^{1(3333)}$ in the rotated basis with a ratio of $1/3$ between $\mathcal{O}_{\mathrm{QQ1}}$ and $\mathcal{O}_{\mathrm{QQ8}}$, as seen in the previous plots.

- When **training and evaluating signal and background in $t\bar{t}t\bar{t}$ and $t\bar{t}b\bar{b}$**, the differences in the 1D shapes between the two "modes" of evaluation almost vanish. Hence, we conclude that the most efficient handle to tease out the EFT nuisances from a large $t\bar{t}$ background not affected by the respective operators is the variation in the shape. The increased discriminative power of the yield changes is still present at values of $\boldsymbol{\theta}$ very close to the SM point. However, this relevant $\boldsymbol{\theta}$-range, where the yield variations would allow tighter limits than just the shape information, is well below the 1σ. When looking at the 2D ratios (especially in $t\bar{t}b\bar{b}$), there is almost no difference.

- Furthermore, when **training and evaluating signal and background**, one notices that the LLR shapes are generally less "smooth" than in the signal only case. This is especially the case for the 2D limits in $t\bar{t}t\bar{t}$, e.g., for the operator combinations of $\mathcal{O}_{\mathrm{QQ1}} + \mathcal{O}_{\mathrm{Qt1}}$ or $\mathcal{O}_{\mathrm{tt}} + \mathcal{O}_{\mathrm{Qt1}}$. This behaviour is mirrored in the histograms of the input features for certain values of $\boldsymbol{\theta}$, where the EFT effects are either present only in form of very flat changes that are almost equal in each bin or almost no variations in yield.

Chapter 8

Conclusion and Outlook

In this thesis, we instantiated a complete work-flow from sample generation to training a neural network with simulation-based inference and setting first nuisance-free limits for single operator insertions in the Standard Model Effective Field Theory formalism.

At the heart of our machine learning approach, we exploited the fact that the beyond the Standard Model effects of effective field operator insertions enter quadratically in the theory prediction for the cross section at a given value of the Wilson coefficients with respect to the Standard Model. This polynomial dependence allowed us to construct a loss function so that we could regress in the intractable quantity of the joint likelihood ratio at detector level while using the tractable event weights from simulation as a training target. By choosing a posteriori the value of $\boldsymbol{\theta}$ we want to be optimal to, we thus circumnavigated the necessity to train separate networks for distinct values in our parameter space.

As training data, we produced signal samples for our physics cases of four top quark production and simultaneous production of two top and two bottom quarks. From the data, we retrieved scalar event-level and jet-array system based observables to feed into our Multivariate Analysis architecture. In terms of machine learning, we used a combination of Deep Neural Networks (DNNs) and Long Short Term Memory (LSTM) layers to optimally extract the beyond the Standard Model information present in the input features. After optimizing the hyperparameters related to the DNN components, however, we found that the LSTMs could not further contribute to the network's performance. An investigation of the causes for this unexpected behaviour took advantage of the fact that our Multivariate Analysis architecture could also be used for multi-classification with minimal modifications.

When using the training setup with multi-classification as a proxy for LSTM testing, we saw a considerable improvement of performance when adding an optimized LSTM configuration to the DNN. After artificially restraining the LSTM input to the features we had used in simulation-based inference before, the performance degraded again, which is a strong hint that the limitations of our detector parametrization in DELPHES are responsible for the weak performance of additional LSTMs.

Finally, we set limits for single operator insertions, both for training with signal samples only and with an additional $t\bar{t}$ background. In this respect, we investigated the yield related improvements in the limits compared to the log-likelihood ratio shapes with variations only in the shapes. Herein, we found that the signal-only performance of the training profits considerably from the yield-related changes, whereas the changes in shape provide the most sensitive handle when background is added. In 2D, we saw correlations between some operator combinations, which we could trace back to the underlying theoretical background.

As an outlook, this thesis opens up interesting perspectives on attempts to learn the EFT effects through simulation-based inference, not least because of intrinsic technical limitations of this work. In fact, making use of samples with a more accurate detector parametrization could fully exploit the LSTM layers' power. If this is comparable to what has shown in multi-classification, the expected improvement could be significant. Additionally, making use of Graph Neural Networks could also lead to an increased performance of the neural network.

Lastly, another point should not be left unmentioned: The limits on Wilson coefficients have been computed in an ideal setting without any addition of systematical uncertainties or nuisances. It remains to investigate in the future if the discriminative power is still present thereafter, especially in the systematically limited, difficult to model $t\bar{t}b\bar{b}$ production.

Acknowledgements

This thesis would not have been possible without the help of many people. First and foremost, I have to thank my supervisor, Privatdoz. DI Dr. Robert Schöfbeck, for his availability whenever I ran into unforeseen problems or just had any sort of question regarding the thesis and more. His expertise and encouragement were the key to complete this analysis.

I am also grateful to the whole CMS analysis group at PSK for always being eager to share physical and technical expertise.

Furthermore, I would be remiss in not mentioning my family and Ben Heinrich, who always supported me in every problem I could possibly imagine.

List of Figures

Bibliography

[1] L. Evans and P. Bryant. "LHC Machine". In: *Journal of Instrumentation* 3.8 (2008), S08001. DOI: 10.1088/1748-0221/3/08/S08001. URL: https://dx.doi.org/10.1088/1748-0221/3/08/S08001.

[2] The ATLAS Collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". In: *Physical Letters* B716 (2012), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020. eprint: 1207.7214. URL: https://doi.org/10.1016%2Fj.physletb.2012.08.020.

[3] The CMS Collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". In: *Physical Letters* B716 (2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021. URL: https://doi.org/10.1016%2Fj.physletb.2012.08.021.

[4] C.P. Burgess. "An Introduction to Effective Field Theory". In: *Annual Review of Nuclear and Particle Science* 57.1 (2007), pp. 329–362. DOI: 10.1146/annurev.nucl.56.080805.140508.

[5] C. J. C. Burges and H. J. Schnitzer. "Virtual Effects of Excited Quarks as Probes of a Possible New Hadronic Mass Scale". In: *Nuclear Physics B* 228 (1983), pp. 464–500. DOI: 10.1016/0550-3213(83)90555-2.

[6] C. N. Leung, S. T. Love, and S. Rao. "Low-Energy Manifestations of a New Interaction Scale: Operator Analysis". In: *Zeitschrift für Physik C* 31 (1986), pp. 433–437. DOI: 10.1007/BF01588041.

[7] W. Buchmuller and D. Wyler. "Effective Lagrangian Analysis of New Interactions and Flavor Conservation". In: *Nuclear Physics B* 268 (1986), pp. 621–653. DOI: 10.1016/0550-3213(86)90262-2.

[8] I. Brivio and M. Trott. "The standard model as an effective field theory". In: *Physics Reports* 793 (2019), pp. 1–98. DOI: 10.1016/j.physrep.2018.11.002.

[9] B. Grzadkowski et al. "Dimension-six terms in the Standard Model Lagrangian". In: *Journal of High Energy Physics* 2010.10 (2010). DOI: 10.1007/jhep10(2010)085. URL: https://doi.org/10.1007%2Fjhep10%282010%29085.

[10] J. Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural Networks* 61 (2015), pp. 85–117. DOI: 10.1016/j.neunet.2014.09.003.

[11] Y. Hu et al. *Overcoming the vanishing gradient problem in plain recurrent networks*. 2018. DOI: 10.48550/ARXIV.1801.06105. URL: https://arxiv.org/abs/1801.06105.

[12] X. Tian et al. *Deep LSTM for Large Vocabulary Continuous Speech Recognition*. 2017. DOI: 10.48550/ARXIV.1703.07090. URL: https://arxiv.org/abs/1703.07090.

[13] S. Hochreiter and J. Schmidhuber. "Long Short-term Memory". In: *Neural computation* 9 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[14] F. Gers, J. Schmidhuber, and F. Cummins. "Learning to Forget: Continual Prediction with LSTM". In: *Neural computation* 12 (2000), pp. 2451–71. DOI: 10.1162/089976600300015015.

[15] K. Cranmer, J. Pavez, and G. Louppe. *Approximating Likelihood Ratios with Calibrated Discriminative Classifiers*. 2016. arXiv: 1506.02169 [stat.AP].

[16] J. Brehmer et al. "Benchmarking simplified template cross sections in W H production". In: *Journal of High Energy Physics* 2019.11 (2019). DOI: 10.1007/jhep11(2019)034.

[17] A. Butter et al. *Back to the Formula – LHC Edition*. 2023. arXiv: 2109.10414 [hep-ph].

[18] A. Valassi. "Optimising HEP parameter fits via Monte Carlo weight derivative regression". In: *EPJ Web of Conferences* 245 (2020). Ed. by C. Doglioni et al., p. 06038. DOI: 10.1051/epjconf/202024506038.

[19] R. Gomez Ambrosio et al. "Unbinned multivariate observables for global SMEFT analyses from machine learning". In: *Journal of High Energy Physics* 2023.3 (2023). DOI: 10.1007/jhep03(2023)033.

[20] J. Hollingsworth and D. Whiteson. *Resonance Searches with Machine Learned Likelihood Ratios*. 2020. arXiv: 2002.04699 [hep-ph].

[21] J. Brehmer and K. Cranmer. "Simulation-based inference methods for particle physics". In: *Nature Reviews Physics* 3.305 (2021). DOI: https://doi.org/10.1038/s42254-021-00305-6.

[22] J. Brehmer et al. "Constraining Effective Field Theories with Machine Learning". In: *Physical Review Letters* 121.11 (2018). DOI: 10.1103/physrevlett.121.111801.

[23] S. Chatterjee et al. *Learning the EFT likelihood with tree boosting*. 2022. arXiv: 2205.12976 [hep-ph].

[24] S. Chatterjee et al. "Tree boosting for learning EFT parameters". In: *Computer Physics Communications* 277 (2022), p. 108385. DOI: 10.1016/j.cpc.2022.108385. URL: https://doi.org/10.1016%2Fj.cpc.2022.108385.

[25] L. Pacchiardi and R. Dutta. *Likelihood-Free Inference with Generative Neural Networks via Scoring Rule Minimization*. 2022. arXiv: 2205.15784 [stat.CO].

[26] S. Chen et al. "Parametrized classifiers for optimal EFT sensitivity". In: *Journal of High Energy Physics* 2021.5 (2021). DOI: 10.1007/jhep05(2021)247.

[27] J. Brehmer et al. "A guide to constraining effective field theories with machine learning". In: *Physical Review D* 98 (5 2018), p. 052004. DOI: 10.1103/PhysRevD.98.052004. URL: https://link.aps.org/doi/10.1103/PhysRevD.98.052004.

[28] J. Brehmer et al. "Mining gold from implicit models to improve likelihood-free inference". In: *Proceedings of the National Academy of Sciences* 117.10 (2020), pp. 5242–5249. DOI: 10.1073/pnas.1915980117.

[29] J. Brehmer et al. "MadMiner: Machine learning-based inference for particle physics". In: *Comput. Softw. Big Sci.* 4.1 (2020), p. 3. DOI: 10.1007/s41781-020-0035-2. arXiv: 1907.10621 [hep-ph].

[30] R. Aoude et al. "Complete SMEFT predictions for four top quark production at hadron colliders". In: *Journal of High Energy Physics* 2022.10 (2022). DOI: 10.1007/jhep10(2022)163. URL: https://doi.org/10.1007%2Fjhep10%282022%29163.

[31] J. D'Hondt et al. "Learning to pinpoint effective operators at the LHC: a study of the $t\bar{t}b\bar{b}$ signature". In: *Journal of High Energy Physics* 2018.11 (2018). DOI: 10.1007/jhep11(2018)131. URL: https://doi.org/10.1007%2Fjhep11%282018%29131.

[32] The CMS Collaboration. "The CMS experiment at the CERN LHC". In: *Journal of Instrumentation* 3.08 (2008). DOI: 10.1088/1748-0221/3/08/S08004. URL: https://dx.doi.org/10.1088/1748-0221/3/08/S08004.

[33] CERN. *What is CERN's mission?* 2023. URL: https://home.cern/about/who-we-are/our-mission (visited on 05/11/2023).

[34] The UA1 Collaboration. "Experimental observation of isolated large transverse energy electrons with associated missing energy at s=540 GeV". In: *Physics Letters B* 122.1 (1983), pp. 103–116. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(83)91177-2. URL: https://www.sciencedirect.com/science/article/pii/0370269383911772.

[35] The UA2 Collaboration. "Observation of single isolated electrons of high transverse momentum in events with missing transverse energy at the CERN pp collider". In: *Physics Letters B* 122.5 (1983), pp. 476–485. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(83)91605-2. URL: https://www.sciencedirect.com/science/article/pii/0370269383916052.

[36] The NA48 Collaboration. "A new measurement of direct CP violation in two pion decays of the neutral kaon". In: *Physics Letters B* 465.1 (1999), pp. 335–348. ISSN: 0370-2693. DOI: https://doi.org/10.1016/S0370-2693(99)01030-8. URL: https://www.sciencedirect.com/science/article/pii/S0370269399010308.

[37] CERN. *The birth of the Web*. 2023. URL: https://home.web.cern.ch/science/computing/birth-web (visited on 05/11/2023).

[38] CERN. *CERN's accelerator complex*. 2023. URL: https://www.home.cern/science/accelerators/accelerator-complex (visited on 05/11/2023).

[39] CERN. *The Large Hadron Collider*. 2023. URL: https://home.cern/science/accelerators/large-hadron-collider (visited on 05/11/2023).

[40] G. Baur et al. "Production of antihydrogen". In: *Physics Letters B* 368.3 (1996), pp. 251–258. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(96)00005-6. URL: https://www.sciencedirect.com/science/article/pii/0370269396000056.

[41] CERN (Education, Communications and Outreach Group). "LHC. The guide. FAQ". In: *CERN-Brochure-2021-004-Eng* (2021).

[42] The CMS Collaboration. "A portrait of the Higgs boson by the CMS experiment ten years after the discovery". In: *Nature* 607.7917 (2022), pp. 60–68. DOI: 10.1038/s41586-022-04892-x. URL: https://doi.org/10.1038%2Fs41586-022-04892-x.

[43] ATLAS and CMS Collaborations, Vasiliki A. Mitsou for the collaborations. "SUSY searches in ATLAS and CMS". In: *Proceedings of Science* CORFU 2019 (2020), p. 050. DOI: 10.22323/1.376.0050.

[44] N. Castro et al. *LHC EFT WG Report: Experimental Measurements and Observables*. 2022. arXiv: 2211.08353 [hep-ph].

[45] T. Abe et al. *LHC Dark Matter Working Group: Next-generation spin-0 dark matter models*. 2018. arXiv: 1810.09420 [hep-ex].

[46] The LHCb Collaboration. "The LHCb Detector at the LHC". In: *Journal of Instrumentation* 3 (2008), S08005. DOI: 10.1088/1748-0221/3/08/S08005.

[47] The LHCb Collaboration. "Measurement of Antiproton Production in pHe Collisions at $\sqrt{s_{NN}} = 110$ GeV". In: *Physical Review Letters* 121.22 (2018), p. 222001. DOI: 10.1103/PhysRevLett.121.222001. arXiv: 1808.06127 [hep-ex].

[48] The ALICE Collaboration. "The ALICE experiment at the CERN LHC". In: *Journal of Instrumentation* 3 (2008). DOI: 10.1088/1748-0221/3/08/S08002.

[49] The ATLAS Collaboration. "The ATLAS Experiment at the CERN Large Hadron Collider". In: *Journal of Instrumentation* 3 (2008). Also published by CERN Geneva in 2010, S08003. DOI: 10.1088/1748-0221/3/08/S08003. URL: https://cds.cern.ch/record/1129811.

[50] The ATLAS Collaboration. *Detector Technology. The ATLAS Detector*. 2023. URL: https://atlas.cern/Discover/Detector (visited on 05/12/2023).

[51] CERN. *ATLAS*. 2023. URL: https://home.cern/science/experiments/atlas (visited on 05/12/2023).

[52] The CMS Collaboration. *Detector. About CMS*. 2023. URL: https://cms.cern/detector (visited on 05/12/2023).

[53] CERN. *CMS*. 2023. URL: https://home.cern/science/experiments/cms (visited on 05/12/2023).

[54] The ALICE Collaboration. *The experiment*. 2023. URL: https://alice.cern/experiment (visited on 05/12/2023).

[55] CERN. *ALICE*. 2023. URL: https://home.cern/science/experiments/alice (visited on 05/12/2023).

[56] The LHCb Collaboration. *Large Hadron Collider beauty experiment*. 2023. URL: https://lhcb-outreach.web.cern.ch/ (visited on 05/12/2023).

[57] CERN. *LHCb*. 2023. URL: https://home.cern/science/experiments/lhcb (visited on 05/12/2023).

[58] The TOTEM Collaboration. "The TOTEM experiment at the CERN Large Hadron Collider". In: *Journal of Instrumentation* 3 (2008), S08007. DOI: 10.1088/1748-0221/3/08/S08007.

[59] The TOTEM Collaboration. *The TOTEM experiment*. 2023. URL: https://totem-experiment.web.cern.ch/ (visited on 05/12/2023).

[60] The LHCf Collaboration. "The LHCf detector at the CERN Large Hadron Collider". In: *Journal of Instrumentation* 3 (2008), S08006. DOI: 10.1088/1748-0221/3/08/S08006.

[61] CERN. *LHCf*. 2023. URL: https://home.cern/science/experiments/lhcf (visited on 05/12/2023).

[62] J. L. Pinfold. "The MoEDAL Experiment at the LHC – a New Light on the Terascale Frontier". In: *Journal of Physics: Conference Series* 631.1 (2015), p. 012014. DOI: 10.1088/1742-6596/631/1/012014. URL: https://dx.doi.org/10.1088/1742-6596/631/1/012014.

[63] The MoEDAL-MAPP Collaboration. *MoEDAL-MAPP Experiment*. 2023. URL: https://moedal.web.cern.ch/ (visited on 05/12/2023).

[64] The FASER Collaboration. *The FASER Detector*. 2022. arXiv: 2207.11427 [physics.ins-det].

[65] The FASER Collaboration. *FASER: About the experiment*. 2023. URL: https://faser.web.cern.ch/index.php/about-the-experiment (visited on 05/12/2023).

[66] The SND@LHC Collaboration. *SND@LHC: The Scattering and Neutrino Detector at the LHC*. 2023. arXiv: 2210.02784 [hep-ex].

[67] The SND@LHC Collaboration. *SND@LHC: Scattering and Neutrino Detector at the LHC*. 2023. URL: https://snd-lhc.web.cern.ch/ (visited on 05/12/2023).

[68] J. Wenninger. "Machine Protection and Operation for LHC". en. In: *CERN Yellow Reports* (2016), Vol 2 (2016): Proceedings of the 2014 Joint International Accelerator School: Beam Loss and Accelerator Protection. DOI: 10.5170/CERN-2016-002.377. URL: https://e-publishing.cern.ch/index.php/CYR/article/view/242.

[69] The CMS Collaboration. *CMS, the Compact Muon Solenoid: Technical proposal*. LHC technical proposal. Geneva: CERN, 1994. URL: http://cds.cern.ch/record/290969.

[70] The CMS Collaboration. *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*. Technical design report. CMS. Geneva: CERN, 2006. URL: https://cds.cern.ch/record/922757.

[71] The CMS Collaboration. "CMS technical design report, volume II: Physics performance". In: *Journal of Physics G* 34.6 (2007), pp. 995–1579. DOI: 10.1088/0954-3899/34/6/S01.

[72] The CMS Collaboration. "CMS Physics Technical Design Report: Addendum on High Density QCD with Heavy Ions". In: *Journal of Physics G: Nuclear and Particle Physics* 34.11 (2007), p. 2307. DOI: 10.1088/0954-3899/34/11/008. URL: https://dx.doi.org/10.1088/0954-3899/34/11/008.

[73] The CMS Collaboration. *The CMS tracker system project: Technical Design Report*. Technical design report. CMS. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/368412.

[74] The CMS Collaboration. *The CMS electromagnetic calorimeter project: Technical Design Report*. Technical design report. CMS. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/349375.

[75] The CMS Collaboration. *The CMS hadron calorimeter project: Technical Design Report*. Technical design report. CMS. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/357153.

[76] The CMS Collaboration. *The CMS muon project: Technical Design Report*. Technical design report. CMS. Geneva: CERN, 1997. URL: https://cds.cern.ch/record/343814.

[77] M. De Giorgi et al. *Design and simulations of the trigger electronics for the CMS muon barrel chambers*. 2005. URL: http://cds.cern.ch/record/1062706.

[78] The CMS Collaboration. *CMS. The TriDAS project. Technical design report, vol. 1: The trigger systems*. Technical design report. CMS. Geneva: CERN, 2000. URL: https://cds.cern.ch/record/706847.

[79] The CMS Collaboration. *CMS The TriDAS Project: Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger. CMS trigger and data-acquisition project*. Technical design report. CMS. Geneva: CERN, 2002. URL: http://cds.cern.ch/record/578006.

[80] E. Corbelli and P. Salucci. "The extended rotation curve and the dark matter halo of M33". In: *Monthly Notices of the Royal Astronomical Society* 311.2 (2000), pp. 441–447. DOI: 10.1046/j.1365-8711.2000.03075.x. URL: https://doi.org/10.1046%2Fj.1365-8711.2000.03075.x.

[81] The Planck Collaboration. "Planck 2018 results". In: *Astronomy & Astrophysics* 641 (2020), A6. DOI: 10.1051/0004-6361/201833910.

[82] J.R. Espinosa et al. "Electroweak baryogenesis in non-minimal composite Higgs models". In: *Journal of Cosmology and Astroparticle Physics* 2012.1 (2012), p. 012. DOI: 10.1088/1475-7516/2012/01/012.

[83] E. Witten. "Dynamical breaking of supersymmetry". In: *Nuclear Physics B* 188.3 (1981), pp. 513–554. ISSN: 0550-3213. DOI: https://doi.org/10.1016/0550-3213(81)90006-7.

[84] J.D. Lykken. *Beyond the Standard Model*. 2011. arXiv: 1005.1676 [hep-ph].

[85] S. P. Martin. "A Supersymmetry Primer". In: *Perspectives on Supersymmetry*. World Scientific, 1998, pp. 1–98. DOI: 10.1142/9789812839657_0001.

[86] G. Cacciapaglia et al. "Composite scalars at the LHC: the Higgs, the Sextet and the Octet". In: *Journal of High Energy Physics* 2015.11 (2015). DOI: 10.1007/jhep11(2015)201.

[87] D. Dicus, A. Stange, and S. Willenbrock. "Higgs decay to top quarks at hadron colliders". In: *Physics Letters B* 333.1-2 (1994), pp. 126–131. DOI: 10.1016/0370-2693(94)91017-0.

[88] G. Cacciapaglia, A. Deandrea, and J. Llodra-Perez. "A dark matter candidate from Lorentz invariance in 6D". In: *Journal of High Energy Physics* 2010.3 (2010). DOI: 10.1007/jhep03(2010)083.

[89] P. Binétruy et al. "Relating incomplete data and incomplete theory". In: *Physical Review D* 70.9 (2004). DOI: 10.1103/physrevd.70.095006. URL: https://doi.org/10.1103%2Fphysrevd.70.095006.

[90] J. A. Aguilar Saavedra et al. *Interpreting top-quark LHC measurements in the standard-model effective field theory*. 2018. arXiv: 1802.07237 [hep-ph].

[91] The CMS Collaboration. *Observation of four top quark production in proton-proton collisions at* $\sqrt{s} = 13$ TeV. Tech. rep. Geneva: CERN, 2023. URL: https://cds.cern.ch/record/2853304.

[92] The ATLAS Collaboration. "Measurements of inclusive and differential fiducial cross-sections of $t\bar{t}$ production with additional heavy-flavour jets in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *Journal of High Energy Physics* 2019 (2019). DOI: 10.1007/jhep04(2019)046. URL: https://doi.org/10.1007%2Fjhep04%282019%29046.

[93] R. Frederix, D. Pagani, and M. Zaro. "Large NLO corrections $t\bar{t}W^{\pm}$ and $t\bar{t}t\bar{t}$ hadroproduction from supposedly subleading EW contributions". In: *Journal of High Energy Physics* 2 (2018). DOI: 10.1007/jhep02(2018)031.

[94] The ATLAS Collaboration. *Observation of four-top-quark production in the multilepton final state with the ATLAS detector*. 2023. arXiv: 2303.15061 [hep-ex].

[95] The ATLAS Collaboration. "Evidence for $t\bar{t}t\bar{t}$ production in the multilepton final state in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *European Physical Journal C* 80.11 (2020), p. 1085. DOI: 10.1140/epjc/s10052-020-08509-3. arXiv: 2007.14858 [hep-ex].

[96] The ATLAS Collaboration. *Search for* $t\bar{t}H/A \to t\bar{t}t\bar{t}$ *production in the multilepton final state in proton-proton collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*. Tech. rep. Geneva: CERN, 2022. URL: https://cds.cern.ch/record/2805212.

[97] The ATLAS-Collaboration. "Measurements of inclusive and differential fiducial cross-sections of $t\bar{t}t\bar{t}$ production with additional heavy-flavour jets in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". In: *Journal of High Energy Physics* 2019.4 (2019). DOI: 10.1007/jhep04(2019)046. URL: https://doi.org/10.1007%2Fjhep04%282019%29046.

[98] The CMS Collaboration. "Search for production of four top quarks in final states with same-sign or multiple leptons in proton–proton collisions at $\sqrt{13}$ TeV". In: *European Physical Journal C* 80.2 (2020). DOI: 10.1140/epjc/s10052-019-7593-7.

[99] The CMS Collaboration. "Measurement of the cross section for $t\bar{t}$ production with additional jets and b jets in pp collisions at $\sqrt{s} = 13$ TeV". In: *Journal of High Energy Physics* 2020.7 (2020). DOI: 10.1007/jhep07(2020)125. URL: https://doi.org/10.1007%2Fjhep07%282020%29125.

[100] The CMS Collaboration. *Evidence for the simultaneous production of four top quarks in proton-proton collisions at* $\sqrt{s}$ *=13 TeV*. Tech. rep. Geneva: CERN, 2022. URL: https://cds.cern.ch/record/2827591.

[101] M. Cristinziani and M. Mulders. "Top-quark physics at the Large Hadron Collider". In: *Journal of Physics G: Nuclear and Particle Physics* 44.6 (2017), p. 063001. DOI: 10.1088/1361-6471/44/6/063001. URL: https://doi.org/10.1088%2F1361-6471%2F44%2F6%2F063001.

[102] The CMS Collaboration. "Measurements of $t\bar{t}$ cross sections in association with b jets and inclusive jets and their ratio using dilepton final states in pp collisions at $\sqrt{s}$ = 13 TeV". In: *Physics Letters B* 776 (2018), pp. 355–378. DOI: 10.1016/j.physletb.2017.11.043. arXiv: 1705.10141 [hep-ex].

[103] The CMS Collaboration. "Measurement of the cross section for $t\bar{t}$ production with additional jets and b jets in pp collisions at $\sqrt{s}$ = 13 TeV". In: *Journal of High Energy Physics* 2020.7 (2020). DOI: 10.1007/jhep07(2020)125. URL: https://doi.org/10.1007%2Fjhep07%282020%29125.

[104] The CMS Collaboration. "Measurement of the $t\bar{t}b\bar{b}$ production cross section in the all-jet final state in pp collisions at $\sqrt{s}$=13 TeV". In: *Physics Letters B* 803 (2020), p. 135285. ISSN: 0370-2693. DOI: https://doi.org/10.1016/j.physletb.2020.135285. URL: https://www.sciencedirect.com/science/article/pii/S0370269320300897.

[105] O. Bessidskaia Bylund et al. "Probing top quark neutral couplings in the Standard Model Effective Field Theory at NLO in QCD". In: *Journal of High Energy Physics* 05 (2016), p. 052. DOI: 10.1007/JHEP05(2016)052. arXiv: 1601.08193 [hep-ph].

[106] C. Zhang and S. Willenbrock. "Effective-field-theory approach to top-quark production and decay". In: *Physical Review D* 83 (3 2011), p. 034006. DOI: 10.1103/PhysRevD.83.034006. URL: https://link.aps.org/doi/10.1103/PhysRevD.83.034006.

[107] C. Degrande et al. "Non-resonant new physics in top pair production at hadron colliders". In: *Journal of High Energy Physics* 2011.3 (2011). DOI: 10.1007/jhep03(2011)125. URL: https://doi.org/10.1007%2Fjhep03%282011%29125.

[108] K. Cranmer. "Practical Statistics for the LHC". In: *2011 European School of High-Energy Physics*. 2014, pp. 267–308. DOI: 10.5170/CERN-2014-003.267. arXiv: 1503.07622 [physics.data-an].

[109] C.K. Khosa, V. Sanz, and M. Soughton. "A simple guide from machine learning outputs to statistical criteria in particle physics". In: *SciPost Phys. Core* 5 (2022), p. 050. DOI: 10.21468/SciPostPhysCore.5.4.050. URL: https://scipost.org/10.21468/SciPostPhysCore.5.4.050.

[110] The Particle Data Group. "Review of Particle Physics". In: *Progress of Theoretical and Experimental Physics* 2022 (2022), p. 083C01. DOI: 10.1093/ptep/ptac097.

[111] PyTorch. *LINEAR*. 2023. URL: https://pytorch.org/docs/stable/generated/torch.nn.Linear.html (visited on 06/12/2023).

[112] PyTorch. *ReLU*. 2023. URL: https://pytorch.org/docs/stable/generated/torch.nn.ReLU.html (visited on 06/09/2023).

[113] PyTorch. *LeakyReLU*. 2023. URL: https://pytorch.org/docs/stable/generated/torch.nn.LeakyReLU.html (visited on 06/09/2023).

[114] PyTorch. *Softmax*. 2023. URL: https://pytorch.org/docs/stable/generated/torch.nn.Softmax.html (visited on 06/09/2023).

[115] PyTorch. *Dropout*. 2023. URL: https://pytorch.org/docs/stable/generated/torch.nn.Dropout.html (visited on 06/09/2023).

[116] G.E. Hinton et al. "Improving neural networks by preventing co-adaptation of feature detectors". In: *CoRR* abs/1207.0580 (2012). arXiv: 1207.0580. URL: http://arxiv.org/abs/1207.0580.

[117] A. Labach, H. Salehinejad, and S. Valaee. "Survey of Dropout Methods for Deep Neural Networks". In: *CoRR* abs/1904.13310 (2019). arXiv: 1904.13310. URL: http://arxiv.org/abs/1904.13310.

[118] R. M. Schmidt. *Recurrent Neural Networks (RNNs): A gentle Introduction and Overview*. 2019. DOI: 10.48550/ARXIV.1912.05911. URL: https://arxiv.org/abs/1912.05911.

[119] PyTorch. *LSTM*. 2023. URL: https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html (visited on 05/19/2023).

[120] J. Taylor et al. *Optimizing the optimizer for data driven deep neural networks and physics informed neural networks*. 2022. arXiv: 2205.07430 [cs.LG].

[121] A. Senior et al. "An empirical study of learning rates in deep neural networks for speech recognition". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. 2013, pp. 6724–6728. DOI: 10.1109/ICASSP.2013.6638963.

[122] D.P. Kingma and J. Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[123] PyTorch. *MSELOSS*. 2023. URL: https://pytorch.org/docs/stable/generated/torch.nn.MSELoss.html (visited on 06/09/2023).

[124] I. Neutelings. *CMS coordinate system*. 2023. URL: https://tikz.net/axis3d_cms/ (visited on 05/24/2023).

[125] C. G. Lester and D. J. Summers. "Measuring masses of semiinvisibly decaying particles pair produced at hadron colliders". In: *Physics Letters B* 463 (1999), pp. 99–103. DOI: 10.1016/S0370-2693(99)00945-4. arXiv: hep-ph/9906349.

[126] C. G. Lester. "The stransverse mass, MT2, in special cases". In: *Journal of High Energy Physics* 05 (2011), p. 076. DOI: 10.1007/JHEP05(2011)076. arXiv: 1103.5682 [hep-ph].

[127] S. Ovyn, X. Rouby, and V. Lemaitre. *DELPHES, a framework for fast simulation of a generic collider experiment*. 2009. arXiv: 0903.2225 [hep-ph].

[128] J. de Favereau et al. "DELPHES 3, A modular framework for fast simulation of a generic collider experiment". In: *Journal of High Energy Physics* 02 (2014), p. 057. DOI: 10.1007/JHEP02(2014)057. arXiv: 1307.6346 [hep-ex].

[129] 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". In: *Journal of High Energy Physics* 2014.7 (2014). DOI: 10.1007/jhep07(2014)079. URL: https://doi.org/10.1007%2Fjhep07%282014%29079.

[130] I. Brivio, Y. Jiang, and M. Trott. "The SMEFTsim package, theory and tools". In: *Journal of High Energy Physics* 2017.12 (2017). DOI: 10.1007/jhep12(2017)070. URL: https://doi.org/10.1007%2Fjhep12%282017%29070.

[131] P. Artoisenet et al. "Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations". In: *Journal of High Energy Physics* 2013.3 (2013). DOI: 10.1007/jhep03(2013)015. URL: https://doi.org/10.1007%2Fjhep03%282013%29015.

[132] The GEANT4 Collaboration. "GEANT4–a simulation toolkit". In: *Nucl. Instrum. Meth. A* 506 (2003), pp. 250–303. DOI: 10.1016/S0168-9002(03)01368-8.

[133] S. Alioli et al. "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX". In: *Journal of High Energy Physics* 2010.6 (2010). DOI: 10.1007/jhep06(2010)043. URL: https://doi.org/10.1007%2Fjhep06%282010%29043.

[134] T. Sjöstrand et al. "An introduction to PYTHIA 8.2". In: *Computer Physics Communications* 191 (2015), pp. 159–177. DOI: 10.1016/j.cpc.2015.01.024. URL: https://doi.org/10.1016%2Fj.cpc.2015.01.024.