ETP-KA/2023-02

Implementation of an Optimal Statistical Inference to Reduce Systematic Uncertainties in the Higgs to Tau Tau Analysis at the CMS Experiment

For obtaining the academic degree

MASTER OF SCIENCE

at the Department of Physics
of the Karlsruhe Institute of Technology
accepted

MASTER THESIS

of

B. Sc. Artur, Artemij Monsch

at Institute of Experimental Particle Physics

Reviewer:	Prof. Dr. Markus Klute
Second Reviewer:	Priv. Doz. Dr. Roger Wolf
Advisor:	M. Sc. Lars Sowa

Karlsruhe, Monday 6th February, 2023

Erklärung zur Selbstständigkeit

Ich versichere, dass ich diese Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht und die Satzung des KIT zur Sicherung guter wissenschaftlicher Praxis in der gültigen Fassung vom 24.05.2018 beachtet habe.

Karlsruhe, den 06.02.2023, _____

B. Sc. Artur, Artemij Monsch

Als Prüfungsexemplar genehmigt von

Karlsruhe, den 06.02.2023, _____

Prof. Dr. Markus Klute

Contents

1. Introduction

Data analysis in scientific research plays an essential role in the extraction of valuable insights from data sets that are continuing to increase in amount and complexity. The field of High-energy physics (HEP) has seen substantial progress in its methods of data analysis, advancing with the increased possibilities of machine learning (ML) techniques and improving the accuracy of the obtained physics results. Neural Networks (NN) have become an integral part of classification tasks, discriminating between processes, and are frequently used for the extraction of ML-derived variables, utilizing them for statistical inference.
The typical analysis objective in HEP is the retrieval of the parameter of interest (POI) and its uncertainty by applying statistical inference, where the POI usually corresponds to the signal strength that is defined as a fraction of the measured signal process cross-section compared to its prediction. The thereby obtained uncertainty of the POI consists of a statistical and systematic part and is under constant scrutiny of ongoing reduction through the improvement of analysis techniques and the increase of the amount of analyzed data. HEP experiments are continuing their measurements, steadily increasing the amount of available data. The Compact Muon Solenoid (CMS) at the Large Hadron Collider (LHC) of the European Organization for Nuclear Research (CERN) as a prominent example will double the number of measured collisions after the third data-taking period (Run3) of the LHC [1] and further multiply the amount of available data after the upgrade to the High-Luminosity LHC [2]. The thereby achieved reduction of statistical uncertainties in many analyses will emphasize the importance of adequately addressing of the then dominant systematic uncertainties present in the analyses. Their mitigation is paramount in the next step on the improvement of the analysis objective uncertainty and requires the development of more sophisticated ML methods that are able to address them.
This work presents a strategy to minimize statistical and systematic uncertainties through the usage of NN, improving upon the previously proposed implementation [3]. The necessary HEP fundament for this work, with an emphasis on the CMS experiment, is given in chapter 2. Chapter 3 describes the POI estimation provided from statistical inference, where ML-derived variables are utilized. The same chapter also discusses the general usage of NNs and a method of NN optimization (training) based on the analysis objective. In chapter 4, a new training method is introduced, and the previously proposed implementation is examined upon its stability during the training. Based on the observation of the resulting training process, a modification of the existing method is presented that improves the stability of the training procedure and thereby provides the basis which allows for an application of the method in highly complex analyses. Further, the improved method is applied to a pseudo experiment consisting of one signal and one background process comparing the results to a conventional NN optimization as a benchmark. Chapter 5 introduces the extension of the new method to problems with multiple signal and background processes, thereby enabling the application for experiments that perform differential measurements. The results of the application are then discussed, demonstrating the conceptual differences to the usual NN optimization in the case of multiple processes.

In chapter 6 the developed technique is then applied on a subset of the data set that is used for the Standard Model H$\rightarrow\tau\tau$ analysis of the CMS experiment [4] containing multiple signal processes and several experimental and theoretical uncertainty sources that are utilized in the improved training method, achieving a reduction of the analysis objective that could not be addressed by the conventionally used NN optimization.

2. Analysis of H→ττ events in eτ final state at the CMS experiment

This chapter gives an overview the physics background necessary for high-energy physics experiments and discusses the process of data acquisition from detection to reconstruction exemplified by the CMS experiment at CERN. Furthermore, an outline of the CMS Standard Model H→ττ Analysis is presented, discussing the relevant physics processes for the analysis and the uncertainty sources.

2.1. The Standard Model and the Higgs Boson

The most accurate explanation of our understanding of elementary particles and their interaction is currently described by the Standard Model (SM) of particle physics [5–8], which was developed in the early second half of the 20th century. Classification of elementary particles within the Standard Model divides the known particles into fermions, which possess half-integer spin, and bosons with integer spin. These are summarized in figure 2.1. Fermions, which comprise matter particles, are further divided into quarks and leptons of three generations each. Quark types, also referred to as flavors, can be differentiated within their generation by their electric charge, with the up (u), charm (c), and top quark (t) possessing a positive electric charge of 2/3, and the down (d), strange (s), and bottom quark (b) having a negative electric charge of $-1/3$. Leptons within a generation can also be distinguished by their charge, with the negatively charged electron (e) in the first, muon (μ) in the second, and tau lepton (τ) in the third generation, paired with their corresponding neutrinos (ν_e, ν_μ, ν_τ) that have a neutral electric charge.
Bosons, which act as mediators of the three fundamental forces (excluding gravity), couple to particles with a corresponding charge. Photons (γ), as the mediator of the electromagnetic force, described by quantum electrodynamics (QED) couple only to particles with an electric charge. Similarly, gluons (g), which are the mediators of the strong nuclear force that is described through quantum chromodynamics (QCD), couple only to particles with a so-called color charge, specifically quarks and gluons themselves. The weak nuclear force is described by the exchange of Z and $W^\pm$ bosons, which couple to quarks and leptons.
The Standard Model is constructed on the principle of gauge invariance, where the underlying symmetries of the strong nuclear force ($SU(3)$), weak nuclear force ($SU(2)$), and electromagnetic force ($U(1)$) are maintained leading to the necessity of the introduced gauge fields, that are representing the bosons above, to be massless. The unification of the electromagnetic and weak nuclear force into the electroweak force, by Sheldon Glashow, Abdus Salam, and Steven Weinberg [9] however necessitated an explanation for symmetry breaking resulting in the emergence of a massive neutral Z and two charged $W^\pm$ bosons whose existence was confirmed by their discovery at CERN in 1983 [10–13].
A solution to this arising problem of massive gauge bosons of the weak nuclear force and the general problem of a mass-acquiring mechanism of the particles was resolved by the Brout-Englert-Higgs (BEH) mechanism [14–19] which proposed the introduction of a

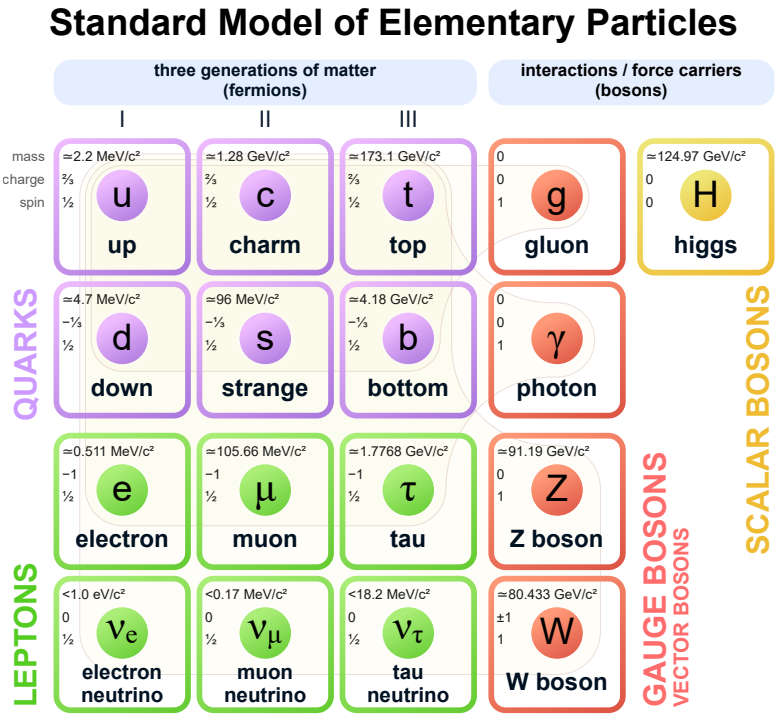


Figure 2.1.: A categorization of elementary particles of the Standard Model divided by colors corresponding to the particle types. Quarks (purple) and leptons (green) are separated into three generations. Bosons (orange) are shown on the right. The interaction between particles is described by the bosons in the enclosed groups shown by the thin lines. The Standard Model Higgs boson (yellow) on the right is shown without an enclosed group interacting with all particles except the gluon and photon. The mass (estimation), electric charge, and spin is provided for each particle in the upper left corner. Taken from [22]

spontaneous symmetry-breaking of the $SU(2) \times U(1)$ electroweak symmetry providing the mass to the Z and $W^{\pm}$ bosons and predicting an additional scalar boson. Furthermore, the BEH mechanism was also able to provide an explanation for the origin of the masses of the fermions through the Yukawa coupling of the Higgs boson to the fermions. This last piece of the Standard Model was confirmed by the discovery of the predicted Standard Model Higgs boson with a measured mass around $125\,\mathrm{GeV}/c^2$ by the CMS [20] and ATLAS [21] Collaborations at CERN in 2012.

The SM Higgs boson, an electrically neutral scalar boson, couples to all particles, with the exception of gluons and photons, and provides the mechanism through which the particles obtain their mass. The coupling to the Higgs boson is quadratic to the mass of the vector bosons and linear to the mass in the case of fermions as described by the Yukawa mechanism for the latter and can be depicted in figure 2.2. The ongoing goal at the Large Hadron Collider (LHC) is the study of the coupling structure of the Higgs boson which includes the studies of its production and decay rates. The main production mechanisms in descending occurrence given from the measured proton-proton collisions at the LHC are the gluon fusion (ggh), vector boson fusion (qqh), and the production in association with a vector boson (VH). The corresponding Feynman diagrams and measured cross sections are given in figure 2.3 and table 2.1. The branching ratios for the main decay modes are listed in table 2.2. The decay into two photons ($\mathrm{H} \to \gamma\gamma$) and four leptons via two Z bosons ($\mathrm{H} \to \mathrm{ZZ} \to 4\ell$) had a significant contribution to the discovery in 2012 despite their low branching fractions, as those decay products can be accurately measured by the detector. Other decay channels of the Higgs boson, pose several challenges e.g. a

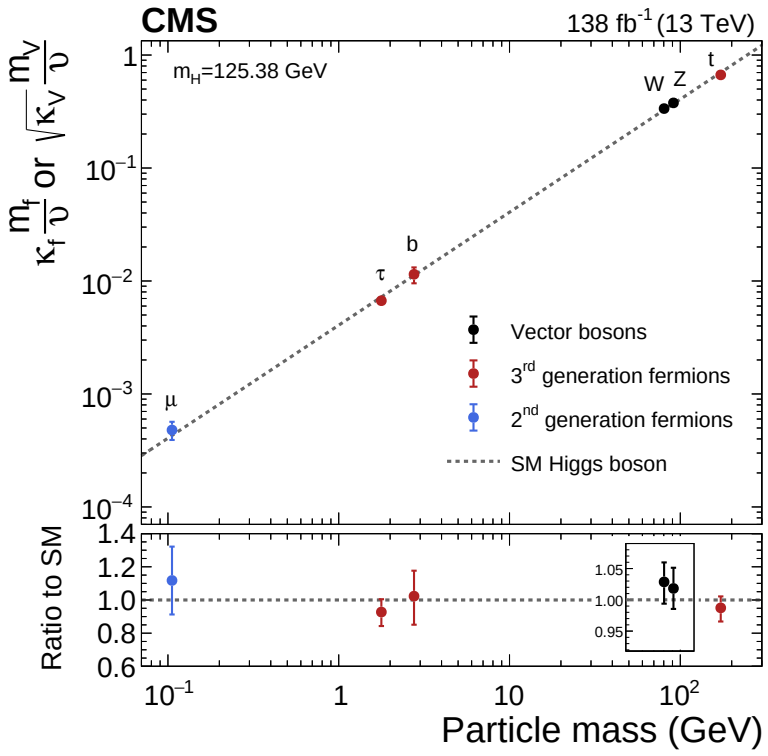
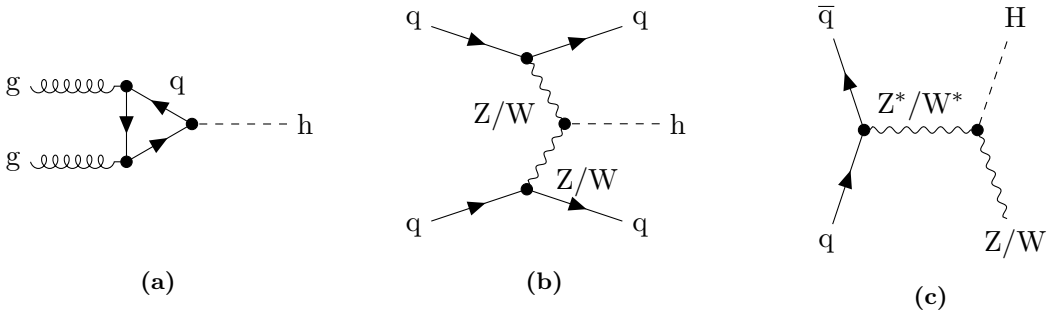


Figure 2.2.: Shown are the measured couplings of the Higgs boson to fermions and vector bosons as a function of the Higgs boson mass that was measured by the CMS collaboration to $m_{\mathrm{H}} = 125.38\,\mathrm{GeV}/c^2$ on the full data set of the LHC second data taking period. The dotted line corresponds to the Standard Model expectation with the square root of the coupling in the case of vector boson interaction with the Higgs boson and a linear coupling in the case of fermions, divided by the vacuum expectation value v. Taken from [23]

Table 2.1.: Shown are the cross-sections of the three main production modes at the LHC from the second data-taking period of LHC at a center of mass energy of $13\,\mathrm{TeV}/c^2$ with the largest contribution from the gluon fusion (ggh) process. The Higgs boson production in association with a vector boson (VH) is the sum of ZH and WH processes and the vector boson fusion (qqh) production mode lists the fiducial cross-section with QCD and electroweak corrections considering only on shell Higgs boson without off-shell effects as stated in [24].

Production process	ggh	qqh	VH
Cross section	48.31 pb	1.97 pb	2.24 pb



(a) **(b)** **(c)**

Figure 2.3.: Feynman diagrams of the three main Higgs production processes at the LHC with gluon fusion in (a), vector boson fusion in (b) and Higgs boson production associated with a vector boson in (c).

Table 2.2.: Branching fractions of the Higgs decay for a Higgs boson mass of $m_{\mathrm{H}} = 125.38\,\mathrm{GeV}/c^2$ as measured by CMS at a center of mass energy of $13\,\mathrm{TeV}/c^2$ as stated in [24].

Decay mode	Fraction
bb	0.58
WW	0.22
gg	0.082
ττ	0.062
cc	0.029
ZZ	0.028
γγ	0.0023
Zγ	0.0016
μμ	0.0002

large number of indistinguishable background processes in the decay into W bosons and b quarks, a weak coupling due to low mass (e.g. H→μμ) or difficulty in reconstructing final states that contain neutrinos.
The direct coupling of the leptons to the Higgs boson allows the measurement of the Yukawa coupling where the H→ττ decay channel poses the advantage of manageable background rates in comparison to the bb decay channel. A differential study of the Higgs production modes is another crucial step in improving the understanding of the couplings of the Higgs boson and is a main target for searches of deviations from Standard Model expectations.
The Standard Model, despite its great success, is limited in its description of more fundamental questions such as the reason for the observed wide mass range of the known particles or the explanation of dark matter which can be indirectly observed today [25]. A search for a deviation from the Standard Model might therefore be a first hint at new physics and coincides with the goal of this work towards a method development resulting in more precise measurements.

2.2. Compact Muon Solenoid experiment

The Compact Muon Solenoid (CMS) experiment at the European Organization for Nuclear Research (CERN) is one of the primary experiments operating along the 27 km Large Hadron Collider (LHC) used for proton-proton acceleration (figure 2.4). During the second data-taking period of the LHC (Run2) between 2016 and 2018, which operated at a center of mass energy of 13 TeV/c^2, a total of $162.85\,\mathrm{fb}^{-1}$ of collision data [1] was produced, of which $150.25\,\mathrm{fb}^{-1}$ was recorded by the CMS experiment [26].
The CMS detector is a classic 4π detector, that surrounds the LHC beam pipe at the collision point, arranging the four main detector elements, and the superconducting solenoid with a magnetic field of 3.8 T concentrically around the beam pipe, with only the muon system not enclosed by the solenoid, as illustrated in figure 2.5.
The determination of the trajectory of charged particles is performed using the tracker system which is the innermost layer of the detector. The curvature of the said trajectory is also used for the determination of the transverse momentum, which is utilized in the reconstruction of particle candidates that uses information from multiple detector parts. The tracking system is composed of silicon detector elements, specifically pixel detectors arranged around the beam pipe followed by silicon strips in two layers, with larger strips in the outermost layer. This configuration allows for a resolution of a few micrometers,

[1] The collision data is represented by the integrated luminosity $\int_t L$ where the Luminosity $L = N^2 A^{-1} f$ describes the number of collisions given N particles inside two bunches over their crossing collision area A and their collision frequency f. The conventionally used unit of barn correspond to $1\,\mathrm{b} = 10^{-28}\,\mathrm{cm}^2$.

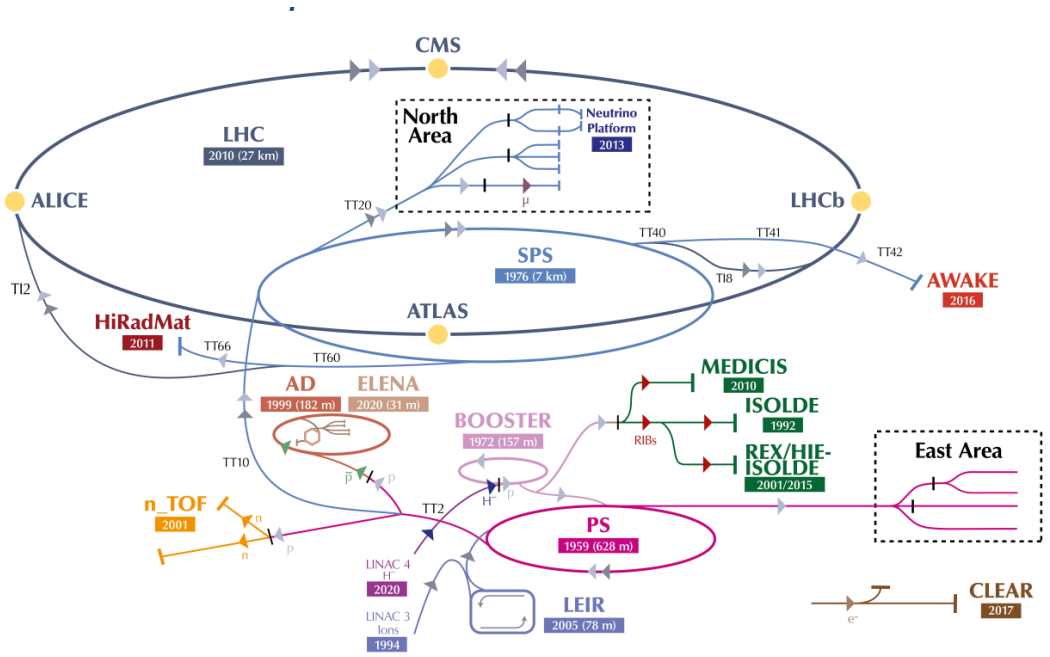


Figure 2.4.: An illustration of the accelerator complex with ongoing experiments. The CMS experiment is located at the LHC accelerator and can be seen in the upper part of the schematic. The coloring of the arrows corresponds to the particle types that are accelerated (decelerated) [27].

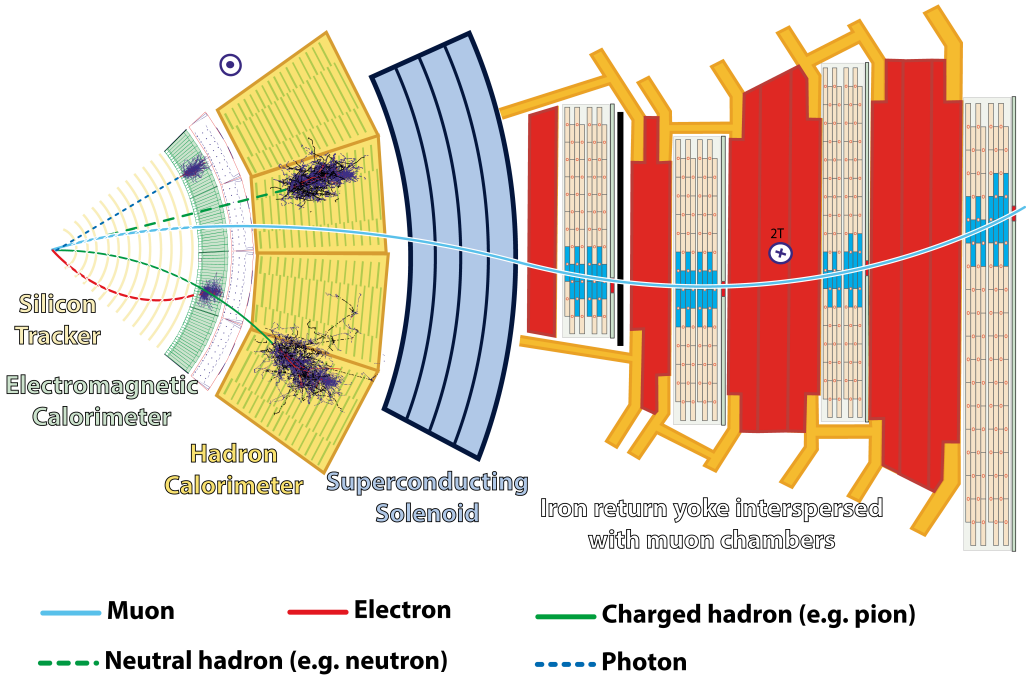


Figure 2.5.: A partial slice of the CMS detector is depicted highlighting the detector components used for particle detection. The signatures, seen in the detector, are distinguished between particles with only a track through all detector components for the muon, electromagnetic shower in ECAL with an additional track in the case for electrons, and no visible track in the tracker system for photons as well as a hadronic shower in HCAL with the distinguishing between neutral and charged hadrons upon the visible track in the tracker system. [28]

enabling a precise extrapolation to the interaction points of the decaying mother particles. The electromagnetic calorimeter (ECAL) is a scintillator detector and comprises the second layer. The energy measurement is conducted from the created electromagnetic showers by the electromagnetic interaction of particles such as electrons and photons. The ECAL is built of $PbWO_4$ crystals, which have a short radiation length of $X_0 = 0.89\,\text{cm}$, allowing for a compact design with around 25 radiation lengths and enabling the placement of the ECAL within the solenoid.
Particles that interact via the strong force propagate through the ECAL unaffected. Their energy deposit is measured in the hadronic calorimeter (HCAL), which is the last detector system placed within the solenoid. In contrast to the ECAL, which utilizes a homogenous design, layers of brass and plastic are used in the HCAL, with brass serving as an absorption material for the creation of hadronic showers, while plastic is used as the scintillation material for the energy measurement.
The muon system of the CMS detector serves as the final layer for measuring propagating muons, as these particles remain unaffected by prior detector systems. Drift tubes within the barrel region and cathode strip chambers in the endcaps, both integrated into the return yoke of the magnet, are used for measuring muon momentum through their curvature as they propagate through the last part of the detector. The information obtained from these detectors is then combined with that of the inner tracker system to achieve a better consistency of muon identification that originated from the collision point.
A crucial step after the initial measurement of the propagated decay products by the individual detector components is the trigger system which consists of two parts. The first-level trigger is implemented using field-programmable gate arrays (FPGAs) [29] utilizing the information from the ECAL, HCAL, and the muon system. The initial rate of the measured events, which corresponds to the bunch crossing rate of 40 MHz is reduced down to 100 kHz [30] by the utilization of the first level trigger. A further reduction down to manageable 1 kHz [31] is achieved by the final application of the high-level trigger (HLT) incorporating tracker information for the initial event reconstruction.
The trigger system utilizes kinematic variables such as the transverse momentum p_T or the direction of the reconstructed particle and its separation from nearby particles ΔR. Due to the invariance of the detected particles within the detector relative to the beam pipe, the cylindrical coordinate system is used for the description of particle trajectories. To address the direction of the particles the combination of the angle ϕ and pseudorapidity η is used, replacing the θ angle by the definition given by equation 2.1.

$$\eta = -\ln\left(\tan\left(\frac{\theta}{2}\right)\right) . \tag{2.1}$$

Hence the convenient representation of the distance between two objects ΔR is expressed by the differences between those two quantities (equation 2.2) and is used mainly for particle selection as discussed in the following section.

$$\Delta R = \sqrt{(\Delta\phi)^2 - (\Delta\eta)^2}\,. \tag{2.2}$$

2.3. Reconstruction and event selection of eτ final state from the H→ττ decay

The proton-proton collision products are tracked within the detector parts in form of multiple electric signals. The reconstruction of those signals into actual physics objects is performed utilizing the Particle Flow (PF) algorithm [32]. A combination of information from multiple detector components, as the particles or their decay products, propagate through the detector, is used by the PF for the reconstruction and assignment of physics objects to their observed tracks and originating points, the vertices. The primary vertex

is selected from cadidates of clustered reconstructed tracks that are closest to the z-axis and are compatible with the beam line. Collisions that are not originating from the primary vertex are referred to as pileup. The reconstructed objects consist e.g. of directly reconstructed electrons or muons and objects that require additional reconstruction steps e.g. jets or tau leptons. The following discussion briefly presents the used reconstruction and selection strategies for these physics objects.

- Electron selection is conducted upon electron candidates reconstructed by the PF algorithm, which utilizes the information from energy depositions in the ECAL and track information. The selection of these candidates is performed by the usage of a boosted decision tree (BDT) [33] incorporating additional information about the shape and specific energy depositions in the ECAL as well as further information about the track and optionally an isolation criterion with respect to other particles that is applied separately if not incorporated in the BDT.
- Muons are also derived from the PF reconstruction, incorporating information from the tracking system and the muon system. The selection applies constraints on the distance of the muons from the primary vertex (impact parameter) and the track quality to ensure the primary vertex is the muon origin point. Additionally, an isolation criterion, to photons, neutral, and charged particles from the particle flow reconstruction, is applied in order to suppress fake muon candidates originating from particle decays inside of jets.
- The identification of jets is initially conducted through clustering of PF reconstructed objects utilizing the anti-k_t algorithm [34] and only using objects originating from the primary vertex. Subsequently, the reconstructed jets are classified by the DeepJet algorithm [35], a neural network-based approach for the identification of jets originating from b quark decays. Selected jets are required to have a separation of $\Delta R > 0.5$ from the selected tau lepton candidates and have a transverse momentum of $p_T > 30\,\mathrm{GeV}/c$ within $|\eta| < 4.7$.
- Reconstruction of tau leptons is conducted in two steps by the application of the Hadron-Plus-Strip (HPS) algorithm [36] using the firstly reconstructed charged and neutral particles from PF as input. A grouping of hadronic tau decay is performed by the algorithm based on the quantity of charged hadrons, with a maximum of up to three, and the number of strips representing the π^0 of the tau lepton decay, with a maximum number of up to two. The corresponding fractions of thereby considered decays are summarized in table 2.3. Further identification and selection of the resulting reconstructed tau lepton candidates are conducted using the DeepTau algorithm [37], which utilizes a deep convolutional neural network. This final step completes the selection step, reducing the occurrence of falsely reconstructed tau leptons from jets, muons, and electrons.

The CMS SM H→ττ Analysis [4] uses for the differential measurement of the Higgs boson production modes the final states of the ττ decays listed in table 2.4, excluding ee and μμ final states due to their low branching fractions. For the application of the presented method, only a subset of the Run2 data is selected, namely the eτ final state of the 2017 data-taking year. The selection of those chosen eτ events that are used in the analysis is divided into three parts, each involving the application of various requirements and thresholds on the electron and tau lepton in order to reduce the misidentification rate and enrich the remaining data set with analysis-relevant processes.
The first step of the selection is represented by the mentioned trigger system with the level-one triggers and subsequent filters that need to be passed by an event in order to be recorded. An eτ event is required to have an electron in $|\eta| < 2.1$ with $p_T > 24\,\mathrm{GeV}/c$ as well as a hadronic decayed tau lepton with $p_T > 30\,\mathrm{GeV}/c$ and the same pseudorapidity

Table 2.3.: Leptonic and hadronic decay fractions of the tau lepton [38]. Decays with charged hadrons ($h^{\pm}$) summarize the decay fractions including $\pi^{\pm}$ and $K^{\pm}$ mesons.

Decay mode	Fraction
Leptonic	35.2
$\tau^- \to e^- \overline{\nu}_e \nu_\tau$	17.8
$\tau^- \to \mu^- \overline{\nu}_\mu \nu_\tau$	17.4
Hadronic	64.8
$\tau^- \to h^- \pi^0 \nu_\tau$	25.9
$\tau^- \to h^- \nu_\tau$	11.5
$\tau^- \to 2h^- h^+ \nu_\tau$	9.8
$\tau^- \to h^- 2\pi^0 \nu_\tau$	9.5
$\tau^- \to 2h^- h^+ \pi^0 \nu_\tau$	4.8
Other	3.3

Table 2.4.: Summary of the final state fractions of the ττ decay where τ_h denotes a hadronic tau decay as provided in table 2.3 and [38]. The ee and μμ final states are not used by the standard model analysis.

Final state	$\tau_h\tau_h$	$e\tau_h$	$\mu\tau_h$	$e\mu$	$\mu\mu$	ee
Fraction	0.42	0.23	0.23	0.06	0.03	0.03

region. Both objects are required to activate two preceding triggers, referred to as cross triggers, with a lower transversal momentum but an additional requirement of spacial separation of $\Delta R > 0.3$ between both objects. All events that pass those selections of the trigger and initial filter system (online analysis) are stored for later usage (offline analyses). The electrons that are present in the offline analysis are further required to fulfill either $p_T > 28\,\mathrm{GeV}/c$ and pass the single-electron trigger or have a $25 < p_T \leq 28\,\mathrm{GeV}/c$ and passing the electron-tau cross trigger with $|\eta| < 2.1$. Furthermore, the score generated by the BDT from the electron selection step must surpass the specified working point of 90 % efficiency. Additionally, the track of the electron has to match the primary vertex with a transverse (longitudinal) distance smaller than 0.045 cm (0.2 cm) and fulfill a p_T dependent isolation criterion if it was not used in combination with the BDT.
The recorded hadronic decayed tau lepton is required to fulfill $p_T > 30\,\mathrm{GeV}/c$ and $|\eta| < 2.3$, as well as passing a channel-specific criterion of the HPS algorithm. The displacement of the leading charged track from the reconstructed tau lepton must be smaller than 0.2 cm in the longitudinal direction. Furthermore, the discrimination value obtained from the DeepTau algorithm must exceed the threshold for the tight working point, and the discriminator against a false classification of electrons and muons as tau must surpass the threshold for the very loose working point, as defined in [37].
The distance of the electron and hadronic tau is required to fulfill $\Delta R > 0.5$ and both leptons must pass a missing energy filter based on the detection region within the CMS detector. Additionally, all considered electron-tau pairs must possess opposite charges and have a combined mass from electron and missing transverse energy mass smaller than $70\,\mathrm{GeV}/c^2$ to avoid an event overlap with the F_F method, which is discussed in the following section. The selection of the optimal electron-tau pair from the remaining candidates within an event is based on the transverse momentum and relative isolation, with preference given to the tau candidate with a DeepTau value closest to one and higher transverse momentum and better isolation in case of an unambiguous choice.
After the selection, an event contains either a background or a signal process that is known in the case of simulated events and is indicated by a corresponding label used for the machine learning application that will be discussed further in chapter 3. The Standard

Model analysis performs a differential measurement of the Higgs production modes given by the stage 1 Simplified Template Cross Section scheme (STXS) as described in [39], which provides a standardized binning approach within high energy physics. This binning introduces 14 signal classes, that are separated by the Higgs production mode, a number of jets, and kinematic variables, such as the transverse momentum of the reconstructed Higgs system or the invariant mass of two leading jets within an event. Within the scope of this work, this number is reduced to two signals that are derived from the Monte Carlo (MC) simulated signal events including the gluon fusion and vector boson fusion production modes which are described by the stage 0 STXS binning. Events containing background processes are further split into five classes and are described in the following section.

2.4. Background estimation

An accurate determination of the cross-section of signal processes requires a comprehensive understanding of the existing background processes, which cannot be further reduced due to the similarity in the final states. One approach to the background process estimation is the application of the MC simulation technique, conducting the event simulation and retrieving the full detector response of these events. This method is used for the signal processes and is applied for the following background processes, that are considered in the analysis.

- $\mathbf{Z \to \ell\ell}$ processes contain the decays of the Z boson into electrons (muons) in case of the eτ (eμ) final state, having the same branching ratio as tau leptons due to lepton universality. The inclusion of this process is particularly crucial when muons or electrons are incorrectly identified as tau leptons in areas with additional jets resulting from pileup.
- $\mathbf{t\bar{t}}$ processes mainly refers to the common decay mode of top quarks into bottom quarks and a W boson. Misidentification can occur in cases of leptonic and hadronic decay of the W bosons, thus creating decay products that can similarly be found in the signal process of eτ final state. This misidentification can be mitigated through the identification of jets originating from bottom quarks and a part of the estimation is covered by the F_F method described below.

An alternative method for background estimation for processes involving genuine tau leptons, not resulting from misidentification during reconstruction, is the τ-embedding method [40] that utilizes the principle of lepton universality. This approach involves the selection of an event containing a decay of a Z boson into two muons and the removal of those from the event. The two removed muons are then replaced with two tau leptons that are derived from the simulation accounting for the muon kinematics observed in the event. The benefit of this estimation method is the inclusion of additional effects such as pileup and detector effects directly from the measurement since they are more complex to simulate and present a source of uncertainty. Another benefit is the resulting high number of events that can be used for background estimation due to the high number of present Z→μμ events. The events containing the background process, estimated by this method, are assigned to a designated class for the machine learning application.

- $\mathbf{Z \to \tau\tau}$ process has its primarily contribution from the Drell-Yan process [41]. This background process can only be distinguished from signal processes by the invariant mass of the two tau leptons. As the decay of tau leptons includes neutrinos, which cannot be directly detected by the CMS detector, the reconstruction of the correct invariant mass poses an additional challenge of distinguishing from the signal processes.

Another technique used for background estimation directly from the data is the F_F method [42, 43]. It is used to estimate the below discussed QCD, W+jets, and $t\bar{t}$ processes, in

which jets from quarks or gluons can be misidentified as hadronic taus. For the machine learning application events that are derived by this method are assigned to a single class. The method measures the number of events not contained in the signal region in which hadronically decaying tau leptons would be identified similar to the signal region and the number of events in which the criteria for a hadronically decaying tau lepton are present but the reconstruction of it fails. The ratio of these two numbers defines the F_F value, which is determined separately in the QCD, W+jets, and $\mathrm{t\bar{t}}$-enriched process regions, which are orthogonal to the signal region. The F_F factors obtained in this way are combined and applied as weights to events in the application region that correspond to the signal region but where the criterion for a reconstructed hadronically decayed tau lepton has not been met. The resulting number of events is then the estimate of the number of events in the signal region in which jets are misidentified as hadronic tau leptons.

- **W+jets** process results in the potential misidentification of a jet as a hadronically decayed tau lepton in combination with a leptonic decayed W boson, or the misidentified jet as an electron and a hadronically decayed tau lepton from the W boson decay in the $e\tau$ final state.
- **QCD** process includes the remaining processes that contain multiple jets and also lead to the potential misidentification of jets as hadronic taus.

Minor background processes, such as the **EWKZ** and **di-boson** process contribute to final states similarly as the other background processes by introducing vector bosons which can produce similar final states as the signal processes due to misidentification with additional jets from pileup that are always present in pp collisions. For the machine learning application, both processes are assigned to a common class.

2.5. Systematic uncertainties considered for the $e\tau$ final state

An integral part of any analysis is the appropriate addressing of systematic uncertainties as those contain effects that do not result from statistical fluctuations and the result of this address leads to an understanding of the impacts on the analysis objective. Systematic uncertainties are usually applied on histogram level resulting in a bin-wise upward and downward variation of bin contents and thus can be referred to as systematic variations. The application of such systematic variations can be achieved by two methods: the introduction of correction weights modifying the bin content without changing the bin assignment of an event, or through systematic shifts of specific quantities, e.g energy or momentum, and the subsequent propagation of these changes resulting in a different bin assignment in the analyzed distribution.
Systematic variations can affect the distribution to which they are applied in two ways: affecting the shape of the distribution of interest while preserving the yield or changing the yield without affecting the shape. In cases where both variations occur due to the introduced systematic uncertainty, both variations are treated as fully correlated during the statistical inference. In the following the main uncertainty sources of the $e\tau$ final state of the 2017 data set which is considered in the Standard Model analysis are discussed.

Efficiency uncertainties comprises lepton trigger efficiencies including single lepton triggers and cross triggers as well as the identification efficiency and tracking efficiencies of tau leptons in the embedding samples. Electron identification efficiency and lepton trigger efficiency introduce normalization uncertainties and are applied to yield estimations derived from the MC and embedded samples. Tau lepton efficiency uncertainties comprising identification, trigger, and tracking efficiency in embedding samples introduce normalization-changing variation in specific transversal momenta regions or certain decay modes where the combination of those uncertainties leads to shape and normalization-changing effects.

***F_F* method uncertainties** introduce shape-altering effects that are arising from the usage of this method for background estimation and are split depending on the correction from QCD, W/Z/$t\bar{t}$ (+jets) events. Those uncertainties are divided into statistical uncertainties that are derived from fit uncertainties and systematic uncertainties representing non-closure corrections. A detailed derivation of individual uncertainties related to the F_F method can be found in [44].

Energy scale uncertainties introduce a shape-altering effect that is applied in the MC data sets containing electrons, jets, tau leptons, and for missing transverse energy (MET). In the case of events created with the embedding method, a mixing of energy depositions in the calorimeters between tau leptons and previously cleaned muons might occur. Due to this reason, the energy scale uncertainties for tau leptons are additionally correlated with the MC samples per decay mode.

Bin-by-bin uncertainties introduce a shape-variation effect that originates from the limited number of used Monte-Carlo generated events for background process estimation. Those uncertainties are addressed as statistical uncertainties during the fit by the Barlow-Beeston approach [45] and its further adaption [46] (Barlow-Beeston lite) producing alternative shapes with the use of bin-associated errors simulating the statistical uncertainty in each bin.

Re-weighting uncertainties account for the correction of missing higher-order effects of matrix element calculation in Monte-Carlo simulated $t\bar{t}$ and $Z \to \ell\ell$ processes. A re-weighting is applied for the transverse momentum for both processes. In the case of $Z \to \ell\ell$ process an additional re-weighting correction is applied on the di-lepton mass.

Luminosity uncertainty affects the yield introducing a normalization variation of 2.6 % for all MC generated events not affecting embedding events and events from the F_F method as they are derived directly from the measurement data.

Signal theory uncertainties comprises cross-section and branching ratio uncertainties and introduces a combination of shape and normalization-changing effects that are applied on single signal bins split into gluon fusion and vector boson fusion as defined by the STXS stage 0 binning in [39]. Further splits are applied e.g. along the Higgs boson transverse momentum scale, the number of jets, or the mass of two leading jets. Additional uncertainties account for the migration between the distinct bins and addresses for missing higher-order corrections from the MC simulations. Normalization uncertainties resulting from the parton density functions are addressed separately for each of the chosen Higgs boson production modes.

Lepton to τ_h fake rate uncertainty accounts for the misclassifications of leptons as tau leptons and differentiates between the endcap and barrel region of the detector by introducing a corresponding uncertainty each.

Prefiring uncertainty mitigates the issue of the CMS detector in years 2016 and 2017 where a timing drift in the forward calorimeter caused the objects from the level one trigger to be assigned to previous events. This resulted in a lower efficiency as only one of three consecutive events are recorded due to trigger configurations.

Further uncertainties account for background normalization uncertainties of e.g. Z+jet events affecting the $Z \to \ell\ell$ process or shape-affecting variation in case of contamination of $t\bar{t}$ events with embedded samples with two genuine tau leptons.

For the application on the selected data set in chapter 6 only the uncertainties applied by the changing weights are considered since they are more advantageous for the calculation. All systematic uncertainties that are used for the application are given in appendix C. The collection does not include i.e. energy scale uncertainties, bin-by-bin uncertainties, or other

uncertainties that cannot be applied according to the above procedure of re-weighting. The following paragraph will shortly describe the abbreviation pattern of the used uncertainties that will be utilized and partly shown in the application step in chapter 6.
The selected uncertainties are distinguished by the corresponding in-, pre-, or suffixes, e.g. the uncertainties derived from the F_F method which are identified by the `_ff_` infix. Theory uncertainties are distinguished by their Higgs production mode, including gluon fusion (`_ggH_`) and vector boson fusion (`_qqH_`, `_vbf_`), separating between correlated uncertainties between experiments (`THU_`) and non-correlated uncertainties for the CMS experiment (`CMS_`). Uncertainties on the parton density function are introduced with the `pdf_` prefix each. Systematic uncertainties accounting for trigger or cross trigger (`_trigger_`, `_xtrigger_`) efficiencies (`_eff_`) are further separated between electron (`_l_`) and tau leptons (`_t_`) of the $e\tau$ final state (`_et_`). The latter are also differentiated according to their decay mode derived from the DeepTau algorithm. Efficiency uncertainties on electron identifications are summarized as one uncertainty with the `_e` suffix, while tau lepton identification is additionally differentiated by p_T regions, including $[30, 35, 40, 500, 1000, \text{inf}]\,\text{GeV}/c$ bins, and are similarly introduced for embedding events (`_emb_`). Efficiency uncertainties on the reconstruction from the HPS algorithm are denoted by the corresponding decay mode, as indicated by the `Prong` infix including the number of charged hadrons in the beginning and the presence of π^0 at the end. Both introduced process re-weighting uncertainties are distinguished by the corresponding suffixes: `_dyShape` for the Drell-Yan and `_ttbarShape` for the $t\bar{t}$ process. Luminosity and prefiring uncertainties are combined into one uncertainty each, indicated by the `lumi_` prefix and `_prefiring` suffix respectively. Cross-section uncertainties are denoted by `xsec`, for example, the cross-section uncertainty of the Z+jet process (`zjXsec`). Uncertainties accounting for the fake rates of electron misidentification as tau leptons (`_fake_`) are separated into the barrel (`_BA_`) and endcap (`_EC_`) region uncertainty.

3. Statistical inference in high energy physics

This chapter presents the basic methodology for statistical inference used for the retrieval of parameters and their associated uncertainties from a statistical model. Additionally, the utilization of quantities obtained from a machine learning approach is discussed together with its usage for statistical inference. Further, the structure of neural networks that are used in this work is given and an optimization strategy is outlined with a presentation of a method for the computation of an optimal training objective.

3.1. Statistical model and parameter estimation

The data obtained from the measurement, as described in chapter 2, is used to determine certain physics processes chosen depending on the analysis objective. Starting from a counting experiment, the measurement can be represented in form of cross-sections σ_X when accounting for the given Luminosity and compared against the standard model expectation σ_{SM}. The ratio of these cross-sections (equation 3.1) is defining the signal strength μ. The uncertainty on μ is given by σ_μ and is described by the confidence interval including statistical uncertainties or a combination of statistical and systematic uncertainties if those are included in the consideration.

$$\mu = \frac{\sigma_X}{\sigma_{\text{SM}}} \,. \tag{3.1}$$

The estimation of the signal strength is conducted by maximizing the likelihood function which represents the probability of measuring $\{x_1, x_2, ...x_N\}$ events given a probability density function (pdf) p that depends on $\{\theta_j\}$ parameters, with $\mu \in \{\theta_j\}$. The difficulty of this method is that the pdf $p(x_i|\{\theta_j\})$ is intractable in high energy physics and therefore a priori not known. This problem is mitigated by the usage of the Monte Carlo method, by simulating particle collisions for the experiment in order to retrieve an estimation of p. The simulated events can be summarized in histograms providing the physics expectation for chosen quantities such as the invariant mass. In combination with the conducted measurement, the statistical inference step can be applied, retrieving $\mu \pm \sigma_\mu$. For this approach, the content of every bin i of a histogram is viewed as an independent counting experiment that follows the Poisson distribution (equation 3.2), described by the number of observed events n_i given the expectation λ_i per bin i. The expectation in each bin is composed of the number of expected background and signal events from the simulation with the signal events being weighted by μ. Thus the Likelihood function to observe a set of specific bin configurations can be constructed as a product of the Poisson distributions

of all bins as done in equation 3.3. The resulting Likelihood is then only dependent on μ and the total number of observed events N.

$$\mathcal{P}\left(n_i|\lambda_i\right) = \frac{\lambda_i^{n_i}}{n_i!}\mathrm{e}^{-\lambda_i}\,, \tag{3.2}$$

$$\mathcal{L}\left(N,\mu\right) = \prod_{i=1}^{N_{\mathrm{bins}}} \mathcal{P}\left(n_i|\mu s_i + b_i\right)\,. \tag{3.3}$$

The parameter of interest (POI) μ and its uncertainty σ_μ is determined by minimizing the negative logarithm of the likelihood function and determining the 68 % confidence interval. This approach tends to be more stable than the maximization of the direct likelihood consisting of probabilities that usually have small numbers. This then addresses only the statistical uncertainty of μ that is accounted for by the Poisson distribution. The inclusion of systematic uncertainties is accomplished by the introduction of additional nuisance parameters (NP), that are denoted as $\{\theta_j\}$, which affect the signal and background expectations. Equation 3.3 is thereby extended to equation 3.4, accounting for the effects of the NP on the processes and multiplying the pdf $\{\mathcal{C}_j\}$ that is chosen for each nuisance parameter individually e.g. as the log-normal distribution in case of normalization uncertainties.

$$\mathcal{L}\left(N,\mu,\{\theta_j\}\right) = \prod_{i=1}^{N_{\mathrm{bins}}} \mathcal{P}\left(n_i\Big|\mu s_i\left(\{\theta_j\}\right) + b_i\left(\{\theta_j\}\right)\right) \prod_{j=1}^{M} \mathcal{C}_j\left(\theta_j\Big|\mu_{\theta_j},\sigma_{\theta_j}\right)\,. \tag{3.4}$$

The confidence interval σ_μ can then be derived from the compatibility of the best-fit value $\hat{\mu}$ with other possible μ values, which corresponds to an inverse hypothesis test. According to Neyman and Pearson [47], the most powerful test statistic is given by the ratio of the likelihoods of two hypotheses as sown in equation 3.5, where $\{\hat{\theta}_j\}_\mu$ corresponds to the best-fit values of the nuisance parameters at a fixed μ.

$$\lambda = \frac{\mathcal{L}\left(N,\mu,\{\hat{\theta}_j\}_\mu\right)}{\mathcal{L}\left(N,\hat{\mu},\{\hat{\theta}_j\}\right)}\,. \tag{3.5}$$

According to Wald [48] and Wilks [49], in the asymptotic case of large sample sizes and negligible nuisance parameter effects, λ can be approximated by a χ^2_k distribution, where k is defined as the difference of the number of degrees of freedom between two hypotheses and in case of only one parameter of interest, k would correspond to one. For the non-asymptotic case, the best-fit value $\hat{\mu}$ and the corresponding 68 % confidence interval σ_μ can be determined numerically by profiling the negative logarithm of equation 3.5.
The procedure of profiling is illustrated in figure 3.1 and is exemplified for one nuisance parameter θ. The negative log-likelihood (NLL) $-\ln\mathcal{L}(N,\mu,\theta)$ is shown as a surface in gray with the μ dependence in the x dimension and θ dependence in the y dimension. The blue dashed line corresponds to the NLL, with a similar form as equation 3.3, having a dependence on θ but without its consideration during the variation. In this example, this θ dependence is fixed to a value resulting in a slice of the likelihood surface that contains the global minimum in order to focus on the effects on σ_μ that are introduced by the consideration of θ. Since θ is usually not known when not explicitly considered in the likelihood, the overlap of the global minima and the minima given by the slice is not necessarily guaranteed. The dashed red line in the figure demonstrates the profiling of θ in the $\mathcal{L}(N,\mu,\theta)$ surface at a given μ within the shown interval. At each point of μ, a minimum in the θ plane is determined. Fulfilling the condition $\inf_\theta\left(-\ln\mathcal{L}(N,\mu,\theta)\right)$ for all μ within the shown interval then leads to a slice in $-\ln\mathcal{L}(N,\mu,\theta)$ that deviates from the blue dashed line, resulting in a lower NLL value at the same μ when not evaluated at the global minimum. This observation is visualized by the projection to one dimension,

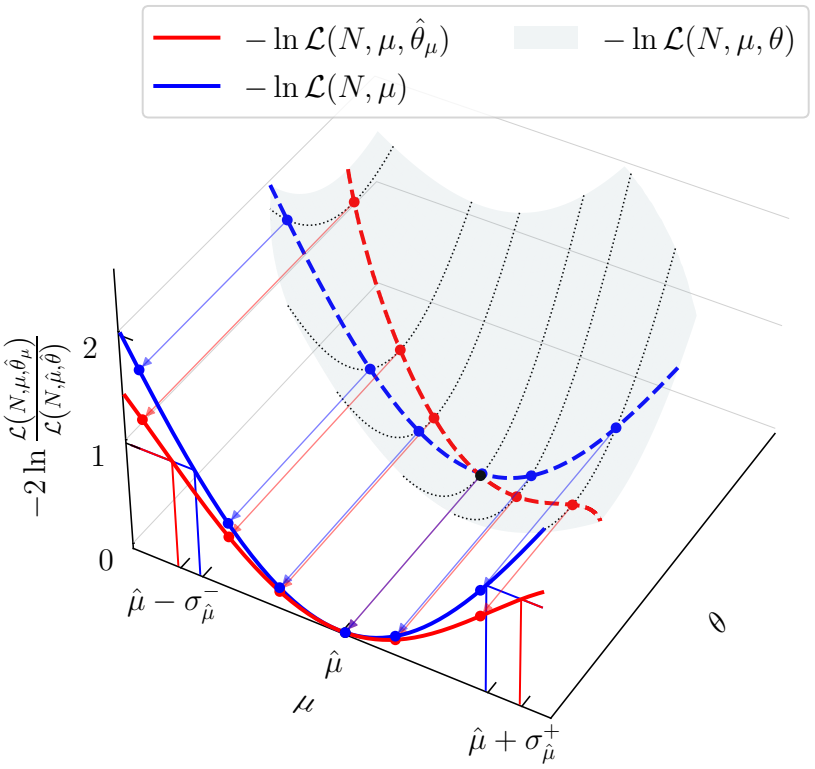


Figure 3.1.: Illustration of an exemplary negative logarithm of a Likelihood function with incorporated signal strength μ and a nuisance parameter θ dependence shown as a gray surface with larger values at the edges and a global minimum that is indicated by the black point. The blue dashed line shows the minimization in one dimension which corresponds to the case of no present nuisance parameters, thus the interval (horizontal lines) of the projected profile (solid blue) line represents the statistical uncertainty. In red the effects of the present nuisance, the parameter is profiled out for each μ which results in a widening of the projected profile and increased uncertainty, containing statistical and systematic uncertainty. The uncertainty is indicated by the $\hat{\mu} \pm \sigma^{\pm}_{\hat{\mu}}$ ticks on the $\hat{\mu}$ axis that projects to the height of one of the projected profiles on the z axis ($-2 \ln \lambda$).

shown by the solid blue and red non-straight lines, with the same minimum, leading to the same $\hat{\mu}$ in this example. The confidence interval of $\hat{\mu}$ is indicated by the blue and red straight lines, which are determined by the values of the NLL scans at a height of one, leading to a lower (higher) σ^-_μ (σ^+_μ) uncertainty on $\hat{\mu}$. As can be depicted from these scans the uncertainty on $\hat{\mu}$ increases upon the consideration of θ describing the statistical and systematic uncertainty on $\hat{\mu}$ whereas in comparison the blue scan describes the statistical only uncertainty on $\hat{\mu}$.
The study of the effects of the nuisance parameters on the uncertainty of the estimated signal strength and potential optimization of the analysis is performed blind, without knowledge of the measured data utilizing an Asimov data set [50, 51] as a replacement. The Asimov data represents the expectation value, thereby guaranteeing the independence from statistical fluctuations for the statistical inference, resulting in $\hat{\mu} = 1$, and is also used for the machine learning application which is discussed further in the next sections.

3.2. Feed forward neural networks

The application of the statistical inference using the statistical model described by equation 3.4 requires provided values of process expectations and the performed measurement. Those values can be extracted from physical quantities that can discriminate well between the processes, e.g. the invariant mass. Including additional quantities may enhance this discrimination and further improve the result of statistical inference. The limitation of this approach, however, is given by the finite number of simulated and performed measurements leading to low-populated bins e.g. in the case of multidimensional histograms.
A possible solution to the problem of high dimensionality can be a function f that maps a set of dimension $\mathbb{R}^{N\times M}$ to a set with the dimension $\mathbb{R}^{N\times C}$, where N represents the number of events present in a given data set and M the number of considered physics variables retrieved from those events. The best approach to finding the appropriate function for the mapping between two finite-dimensional spaces is given by utilizing a neural network (NN) as it can approximate any function if sufficiently many free parameters are provided for the representation [52]. The NN is then capable to reduce those M numbers of given variables to C ML-derived variables, with $C < M$ or a single number ($C = 1$), while ideally maintaining correlations between the provided M variables. In the case of simulated events, the physics process leading to the specific ML-derived variable is known and is associated by a corresponding label in the data set. Those variables and labels can subsequently be used in the statistical model e.g. equation 3.4 for the inference step, replacing physical quantities with the condensed information contained in a single or a handful of ML-derived variables which are summarized into one or two-dimensional histograms, circumventing the problem of high dimensionality.
The type of NN that is utilized by the CMS Standard Model H$\rightarrow\tau\tau$ analysis and used throughout this work is a Feed Forward NN, which comprises several artificial neurons [53] that are organized into l layers. All neurons in a layer are connected to each neuron in the subsequent layer, and the information from M variables (NN input) is thereby propagated through the complete NN, ending with the NN output as the output of the neurons from the last layer. Those neurons are specifically referred to as NN output nodes to distinguish them from the neurons used in the hidden layers, which are defined as the layers between the NN input and the NN output nodes. The information propagation of a single neuron in layer l is described by equation 3.6 where the outputs of the neurons from the previous layer $\{x_i^{(l-1)}\}$ are multiplied by the weights $\{w_i^{(l)}\}$ of the current layer l. Additionally, a bias value $b_i^{(l)}$ of the current layer is added to the sum of the O_i products. The resulting value is then transformed by a non-linear activation function φ and is used as input for the neurons in the subsequent layer:

$$x_i^{(l)} = \varphi\left(\sum_{j=1}^{O_i} w_{ij}^{(l)} x_i^{(l-1)} + b_i^{(l)}\right) . \tag{3.6}$$

The choice of non-linear activation functions φ is made for each layer of the neural network. The rectified linear unit (ReLU) [54, 55] as depicted in equation 3.7 is preferably used as the activation function for hidden layers as it tends to show better performance in comparison to symmetric activation functions e.g. tangent hyperbolic function [56]. For the activation function of the NN output layer, the Sigmoid or Softmax function is typically applied, depending on the number of present neurons as both have a codomain of [0, 1]. The Sigmoid function (equation 3.8) can be applied for binary classification tasks, differentiating between two processes. The statistical inference is then performed on a one-dimensional histogram of the NN output.
In contrast, the Softmax activation function (equation 3.9) is used in presence of multiple NN output nodes, representing the number of classes (C_{class}) that are targeted by an

analysis. The softmax function further applies a normalization between all classes for each event propagated through the NN, allowing for an interpretable probabilistic class association of the NN output. The Softmax codomain thereby consists of C_{class} numbers in the [0, 1] interval for each class.

$$\text{ReLU}(x) = \begin{cases} x & \text{if } x > 0\,, \\ 0 & \text{else} \end{cases} \quad , \tag{3.7}$$

$$\text{Sigmoid}(x) = \frac{1}{1 + \mathrm{e}^{-x}}\,, \tag{3.8}$$

$$\text{Softmax}(x_c) = \frac{e^{x_c}}{\sum_i^{C_{\text{class}}} e^{x_c}} \quad \text{for } c \in \{1, 2, ...C_{\text{class}}\}\,. \tag{3.9}$$

As the H→ττ analysis performs a differential measurement of the Higgs production assigning the production modes to unique classes on top of the background classes, the Softmax activation function is used for the present multiple output nodes. After the training, which optimizes the weights of the NN, and is discussed further in the following section, N events are propagated through the NN resulting in C_{class} ML-derived variables given by the NN output nodes. To avoid double counting only the nodes with the maximal output values are used for the statistical inference process. The NN output node with the maximal numerical value is used as the assignment of the propagated N events to one of the C_{class} classes. The position within the resulting class histograms is then determined by the numerical values. The resulting C_{class} histograms after the summarization of N events are then used in the statistical inference to retrieve $\mu \pm \sigma_\mu$ as discussed in section 3.1.

3.3. Neural network optimization

The optimization of the NN is conducted through an iterative process referred to as training. The weights $\{w_{ij}\}$ and biases $\{b_i\}$ of the NN are initialized randomly according to the Glorot normal method [57]. The iterative optimization of the weights and biases utilizes the method of gradient descent [58], by calculating the gradient of a loss function with respect to the NN weights and biases. The loss function is selected as the training objective and reduces the NN output to a scalar and serves as a metric for optimization with decreasing values for a successful training process, leading to a minimum in the parameter space of the NN weights and biases.
A common choice for the loss function in presence of multiple classes is the categorical cross-entropy, which can be depicted in equation 3.10 and will be referred to as CE. The CE function considers the number of classes for N events and calculates the score of the class assignment given by the data set corresponding to $y_i^{(c)}$ in an ideal case, and computes the logarithm of those NN prediction $\mathrm{f}(x_i)^{(c)}$. The class assignments $y_i^{(c)}$ thereby have the value of one for a specific class c and are zero otherwise. The weights for the individual classes, denoted as $w^{(c)}$, are given by the data set. They indicate the importance of the contribution of individual events to the total loss or can be chosen to mitigate an imbalance of events from different classes used for the training. The training on CE in combination with the Softmax activation function can be interpreted as the maximization of the likelihood of finding an event x_i in the corresponding class c. The minimization of CE loss improves the certainty of the NN upon an event classification leading to an event separation in the introduced number of classes in an optimal training result.

$$L_{\text{CE}} = -\sum_{c=1}^{C_{\text{class}}} w^{(c)} \sum_{i=1}^{N} y_i^{(c)} \log\left(\mathrm{f}(x_i)^{(c)}\right)\,. \tag{3.10}$$

In the case of binary classification problems, the separation of two processes, such as one signal and one background process, can be achieved by maximizing the distance of

those processes in the codomain of the activation function, with the Sigmoid function as a commonly used option. The computation of the training objective can be accomplished by modifying the CE loss for the binary case consisting of two classes $C_{\text{class}} = \{0, 1\}$ as shown in equation 3.11. This special case will be referred to as binary cross-entropy (BCE). The logarithm of the NN output can thereby be interpreted as the probability of an event corresponding to one of the two classes, with increased probability for events that are classified close to the class values of one ($P(y_i = 1) = \mathrm{f}(x_i)$) or zero ($P(y_i = 0) = 1 - \mathrm{f}(x_i)$) that are given by the y_i labels, derived from the data set. The weights w_i of an event are corresponding to the two class weights and are chosen with the same reasoning as $w^{(c)}$ in CE.

$$L_{\text{BCE}} = L_{\text{CE}}\Big|_{C_{\text{class}}=\{0,1\}} = -\sum_{i=1}^{N} w_i \left[y_i \log\left(\mathrm{f}(x_i)\right) + (1 - y_i) \log\left(1 - \mathrm{f}\left(x_i\right)\right)\right] . \tag{3.11}$$

The choice of the loss thereby provides the measuring of the difference between the current NN output prediction and a target given by the label y_i. This difference is successively minimized by the optimization process by adjusting the NN weights and biases. The optimization step t involves the gradient computation and is carried out on the computed loss from all N events for each epoch, or multiple times on the loss derived from successively propagated subsets of N' events with $0 < N' < N$ of the full training data set. The first approach is referred to as full-batch training, whereas the latter, that utilizes a split of the training data set into subsets, is referred to as mini-batch training. During the event propagation through the NN and the computation of the loss (forward pass) all applied operations for each event are stored in a computational graph. In the following backpropagation, the numeric computation of the derivatives of the previously applied operations is performed in the backward direction through the created graph (backward pass) utilizing the automatic differentiation of PyTorch package [59]. This calculation can be exemplarily expressed as a chain rule application, shown for weights w_{ij} in equation 3.12, starting with the derivation of the loss L at the optimization step t ($L^{(t)}$) with respect to the NN output, denoted as $\mathrm{f}(x)$, followed by the derivative of the NN output with respect to the activation function of the last layer l (φ_l) and ending with the last derivative with respect to the selected weights w_{ij}. This resulting gradient is then multiplied by the learning rate η, which scales the performed gradient step. The product is then used as a correction for the weights at the current epoch $w_{ij}^{(t)}$ as shown in equation 3.13. An analogous procedure is applied for the biases b_i.

$$\frac{\partial L^{(t)}}{\partial w_{ij}} = \frac{\partial L^{(t)}}{\partial \mathrm{f}(x)} \frac{\partial \mathrm{f}(x)}{\partial \varphi_l} \frac{\partial \varphi_l}{\partial x_i^{(l-1)}} \cdots \frac{\partial \varphi_i}{\partial w_{ij}} , \tag{3.12}$$

$$w_{ij}^{(t+1)} = w_{ij}^{(t)} - \eta \cdot \frac{\partial L^{(t)}}{\partial w_{ij}} . \tag{3.13}$$

This optimization process can be further improved by utilizing enhanced optimization algorithms, such as Adam [60] or NAdam [61, 62]. These algorithms change the basic gradient descent method by replacing the gradients in the product with the learning rate by adaptive moments. Those are calculated with the usage of the computed gradients and lead to faster convergence of the optimization process. Specifically, the Adam optimizer utilizes the gradient of the current epoch in its calculation of moments, while NAdam also incorporates the gradient from a previous epoch. This combination of gradients from the current and previous optimization steps results in an adjusted momentum, known as Nesterov momentum, that can further improve the optimization process.

3.4. Neural network optimization on the analysis objective

Both training objectives discussed in section 3.3 are commonly used in physics analysis and aim toward process separation. The resulting distributions exhibit process-enriched bins when summarized into histograms, thereby inherently leading to a reduction in statistical uncertainty upon the application of statistical inference. The important point however is that this reduction of statistical uncertainty is only a byproduct of the chosen training objectives and is not explicitly targeted during the training. Also, neither of those training objectives addresses systematic uncertainties that are present in an analysis.
In this regard, selecting σ_μ as the new training objective is an optimal choice, as it aligns with the analysis objective, addressing the statistical and systematic uncertainties of μ, thereby making the training uncertainty-aware. This consistency between the training and analysis objective is thereby of uttermost importance as it enables the learning of an appropriate dimensionality reduction of the M variables by the NN.
The utilization of σ_μ as a training objective rises the necessity of an analytical calculation of σ_μ, since the presented numeric approach of profiling, as described in section 3.1, is not applicable for the backpropagation. An asymptotic estimation of σ_μ can be obtained from the calculation of the Fisher information [63] as depicted in equation 3.14, where the calculation corresponds to the expectation of the Hessian matrix of NLL by calculating the second order derivative of NLL with respect its parameters μ and $\{\theta_j\}$.

$$\mathcal{F}_{ij} = \mathbb{E}\left[\left(\frac{\partial^2}{\partial x_n \partial x_m}\Big[-\log\mathcal{L}\left(N,\mu,\{\theta_j\}\right)\Big]\right)_{x_n,x_m=\{\mu,\{\theta_j\}\}}\right] . \tag{3.14}$$

The denoted expectation is addressed through the utilization of the Asimov data set and the full-batch training, as this approach improves the asymptotic estimation with the increased sample size used for the calculation. With this ansatz, the covariance matrix can be obtained by inverting the calculated Hessian matrix as it corresponds to the Fisher Information, resulting in the estimations of the variances for μ and θ_j given by

$$V_{ii} = \left(\mathcal{F}_{ii}\right)^{-1} . \tag{3.15}$$

The utilization of the likelihood in this estimation is thereby asymptotically efficient, ensuring that the estimation approaches the Cramer-Rao bound [64, 65]. By construction of equation 3.14, the uncertainty σ_μ corresponds to the square root of the first diagonal element of the covariance matrix $\sqrt{V_{11}}$ and provides the confidence interval on μ.
The thereby conducted calculation steps leading to the estimation of σ_μ provide a tractable gradient that is utilized by the backpropagation for the weight optimization as described in section 3.3. The only step in this calculation that poses a problem is the histogram creation step since the gradient of a histogram is not well defined for optimization purposes. A solution to this problem can be the introduction of a custom function that replaces the corresponding inapplicable gradient of the histogram, represented by a derivative in equation 3.12 during the backpropagation. A further discussion upon the choice of the appropriate custom function as the replacement will is presented in section 4.2.

4. Studies on a pseudo experiment for binary classification

This chapter introduces the general procedure for the performed pseudo-experiment studies, that have been used for this, and the next chapter, by defining the used data set for the binary classification and the neural network architecture as well as the general training procedure. Subsequently, an issue associated with calculating the signal strength uncertainty on binned NN output for weight optimization is demonstrated. An additional modification to the existing approach is presented to mitigate this problem and enable the extension of the training to more complex classification tasks e.g. the reduced CMS Standard Model H$\rightarrow\tau\tau$ analysis that will be described in chapter 6. Given this modification, the uncertainty-aware training on σ_μ as discussed in section 3.4 is then compared to the classical training on BCE loss in a simplified example. In the last step, a Taylor Coefficient Analysis (TCA) of the NN output function is conducted to highlight the differences in the NN response upon a performed uncertainty-aware and BCE training.

4.1. Neural network and experimental setup

The pseudo experiments in this and the following chapter consist of multiple processes that are each synthetically generated with 10^5 events from two-dimensional Gaussian distributions with a covariance matrix represented by a two-dimensional identity matrix. The processes are distinguished by their predefined expectation values and their normalization. For binary classification, the expectation value of the signal process is set to $(x_1, x_2) = (0, 0)$, and the background process to $(x_1, x_2) = (1, 1)$ as shown in figure 4.1. The normalizations are chosen to represent the case of 50 signal and 1000 background events for the statistical inference and uncertainty-aware training. The introduced shape-altering systematic variation changes the position of the background distribution in the x_2 by ± 1. For the training, the generated data set is divided into two halves, one half is used for weight optimization, and the other half for validation. The training results are evaluated on an independently generated test data set that is not further divided and contains a total of $2 \cdot 10^5$ events.
For the classification of signal and background, a fully connected feed-forward NN with a single hidden layer of 100 nodes and an applied ReLU activation function is used. The NN weights and biases are randomly initialized using Glorot normal initialization [57]. The number of input nodes, output nodes, and the activation function of the output layer will vary depending on the data set and considered task. The variables x_1 and x_2 are chosen as the inputs to the NN. For the binary classification the number of NN output nodes is set to one, for which the Sigmoid function is selected as the final activation function.
A full batch training is applied, with the consequence of one performed gradient descent step per epoch. The use of full batch training, instead of the commonly used mini-batches, is necessary to achieve the best asymptotic estimation of the covariance matrix and the associated uncertainty on the signal strength, as discussed in chapter 3. The training

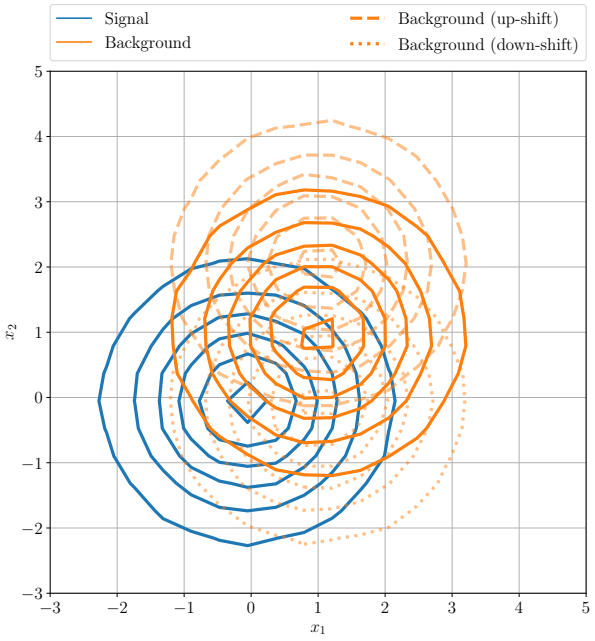


Figure 4.1.: A representation of the signal and background processes that are obtained from 10^5 events following a two-dimensional Gaussian distribution with a unity matrix used as covariance matrix. The mean values to model the distributions are $(x_1, x_2) = (0, 0)$ for signal and $(x_1, x_2) = (1, 1)$ for background. Systematic variations of the background process are shown as dashed and dotted lines for the ± 1 shift along the x_2 axis.

starts with a "warm-up" phase of 300 epochs, which uses BCE as the training objective and will be further motivated in the following sections. After the warm-up phase, the training objective is changed to σ_μ and stops after no improvement on the validation set is achieved for 1000 consecutive epochs. This stopping criterion will be referred to as patience with additional number of chosen epochs. The NN weights of the epoch with the best performance on the validation data set are saved and applied to the test sample after training. Further, NAdam is selected as the optimizer with a learning rate of 0.001 during the warm-up phase and 0.01 during the training on σ_μ. This setup will be used for the application steps and the comparison between the classic training approach and uncertainty-aware training unless stated otherwise.

4.2. Discretization of the neural network output

The discretization of the NN output for the calculation of σ_μ poses a challenge for weight optimization since the gradient of a histogram bin is either zero or infinity. Therefore the backpropagation, as described in chapter 3 cannot be based on a naive gradient calculation. To address the issue of the inapplicable backpropagation, two approaches can be taken. The first option, for approximating the histogram bin-wise with a differentiable function [66] has the disadvantage of training only on an approximation of the analysis objective. An alternative approach, which is adapted in this work, involves the replacement of the histogram gradient in the backpropagation step with a custom function [3] and not altering the calculation of the analysis objective in the forward path.
One option for the custom function is to use a constant non-zero gradient across the entire histogram, effectively skipping this step in the backward pass and introducing only a scaling factor for the final gradient. This approach has the caveat that during the training the used optimizer cannot derive information from the change of the NN output for events inside of bins unless a given event undergoes a change in the bin assignment.
Therefore, a more sophisticated modification to the backward pass is applied, by replacing

the histogram gradient with a custom function as proposed in [3]. To demonstrate the effect of this proposal and its future modification, the setup as described in [3] is chosen for a better comparison of the influence on the training upon the choice of the custom function. The setup further utilizes the Adam optimizer with a learning rate of 0.001. The replacement function for the histogram gradient is built up from the gradients of Gaussian distributions with the standard deviation equal to half of the bin width and the bin center as the mean. The substitution is performed bin-wise, and the resulting bin gradients are combined into a total histogram gradient.
This procedure is illustrated for 20 individual samples in figure 4.2a, where the first line of the figure shows the NN output as vertical lines, indicating the predicted numerical value of the NN, lying between zero and one. The figure below summarizes those 20 samples in form of a histogram, that is further used for the Likelihood calculation in the forward pass. The following blue lines below the histogram correspond to the custom functions used for the replacement of the bin gradients in the backward pass. Each bin gradient thereby exhibits maxima at the bin edges with a sign flip within each bin and are approaching zero outside the corresponding bin. The total histogram gradient, as a combination of those gradients, is shown at the bottom in form of an orange line and is used in the backpropagation as a replacement for the histogram gradient. The calculation of the histogram gradient is conducted by performing a summation of all bin gradients. From the overlapping extrema of each bin gradient, the sum exhibits low amplitudes in the central bins of the histogram and maxima at the edges.
To illustrate the effect of this custom function in the backpropagation, the evolution of the NN output f(x) for 50 randomly picked signal and background events is tracked during the training, as shown in figure 4.2b. The figure is vertically divided by the dashed line at epoch 300, indicating the training on BCE loss in the warm-up phase, followed by the training on σ_μ and the black horizontal lines correspond to the bin edges of the histogram used to summarize the the NN output. The evolution of the NN output shows a signal and background process separation during the warm-up phase, populating all bins of the histogram. After the warm-up phase, a steep aggregation of events into a few bins is observed, followed by oscillations that spread the NN classification. This oscillation, however, does not lead to a redistribution of the events into more bins, leaving the remaining bins of the histogram unpopulated for the remaining training.
Figure 4.2c shows the loss evolution for the corresponding training procedure and is divided into two parts: the upper part of the figure shows the evolution of the BCE loss, whereas the lower part shows the evolution of σ_μ loss. The training is performed on BCE during the warm-up phase and on σ_μ afterward. The solid blue line shows the training loss and indicates the switch from BCE to σ_μ by the dashed vertical black line at epoch 300. The validation loss, which is calculated for both losses at each epoch, is shown by the orange lines and is further divided into two parts, indicating its usage for the validation step as a solid line and a calculation only for the illustrative purpose in form of a dotted line. As observed, the evolution of the BCE loss decreases during its usage in the warm-up phase and is followed by a steep increase after the switch leading to a plateau after epoch 400. The σ_μ loss, on the other hand, exhibits a short-term fluctuation before settling into a plateau that lasts until the end of the warm-up phase. Afterwards σ_μ decreases briefly and shows erratic behavior for the remaining training.
The observed separation between signal and background processes during the warm-up phase, as the result the training on BCE, also reduces the σ_μ as indicated by 4.2c. A good separation of signal and background events, which leads to the enrichment of events in the lowest and highest bins of the histogramming NN output, is maximized after a fraction of the warm-up phase, as indicated by the emerging plateau in the σ_μ evolution. Further adjustments to the NN weights do not affect σ_μ as the focus of the NN optimization shifts to the adjustment of event position inside the central bins, which poses only a minor

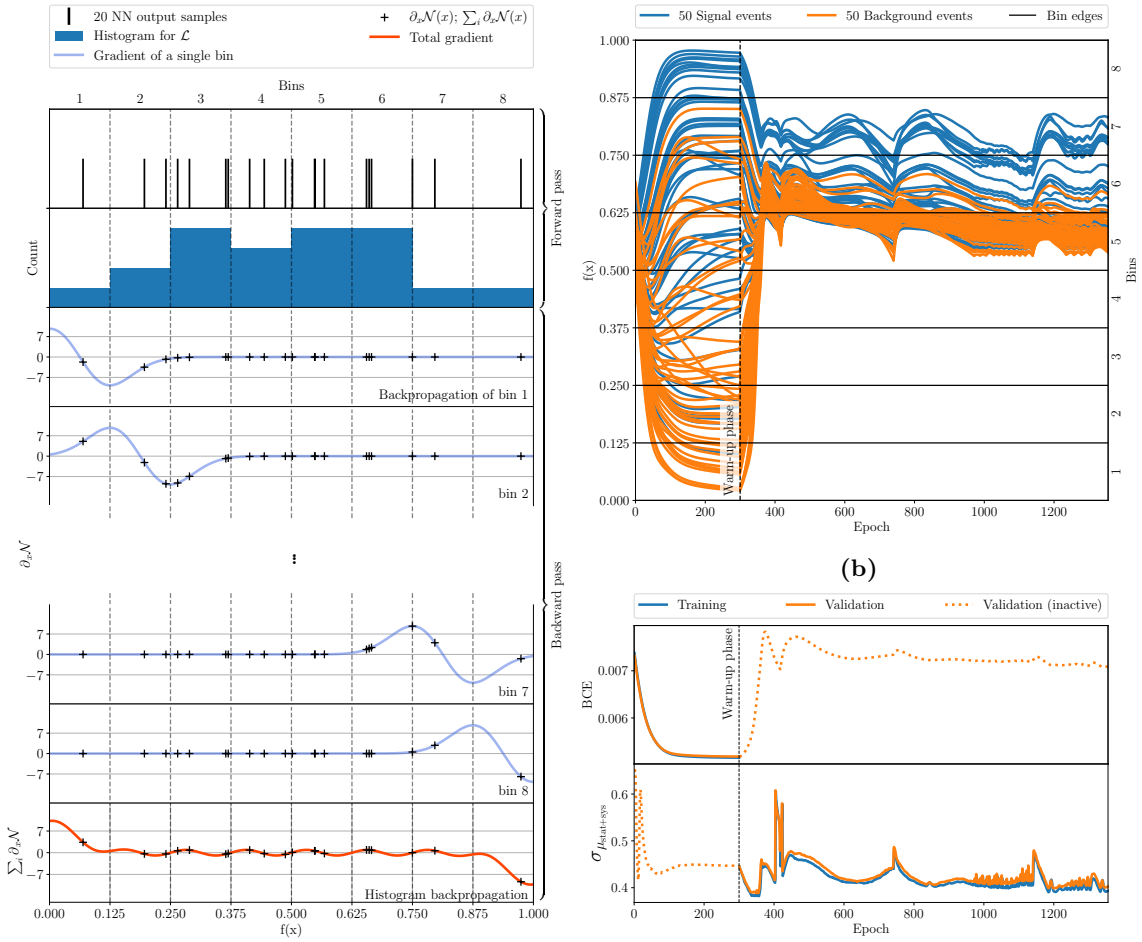


Figure 4.2.: The bin-wise calculation of the histogram gradient as proposed in [3] is shown in (a) for 20 individual random samples. The evolution of the NN classification of 50 randomly chosen signal and background event each is displayed in (b) for the complete full training. The end of the BCE warm-up phase is indicated by the dashed line at epoch 300. The loss evolution of the training (c) also indicates this switch. The loss evolution is further split into the BCE loss (top) and σ_μ loss (bottom). The dotted line indicates the computed validation loss that is not used for the validation step and switches from σ_μ to BCE after the warm-up phase. Further information is provided in the text.

contribution to σ_μ.
After the completion of the warm-up phase, a collapse of events into fewer bins occurs during the next 50 epochs where the minimum of σ_μ is found. Since this collapse mainly moves the background events up to the signal events, the BCE value increases, indicating a misclassification from the perspective of BCE that is not further affecting the training. The distribution after this collapse however does not necessarily correspond to an unambiguous absolute minimum of the training process, as demonstrated by the subsequent jumps in σ_μ loss and the attempts to redistribute events into neighboring bins as shown in figure 4.2b. The phenomenon of the NN output function f(x) collapsing into very few bins is thereby independent of the exact configuration of the warm-up phase. Training without a warm-up phase does not show a prominent collapse since the events are located inside the central bins of the histogram at the start of the training due to the NN weights initialization and the usage of Sigmoid as the final activation function. Their movement throughout the training however remains grouped within a few bins showing a similar erratic loss evolution but at a slower pace. The exact number of populated bins and their position is task-specific

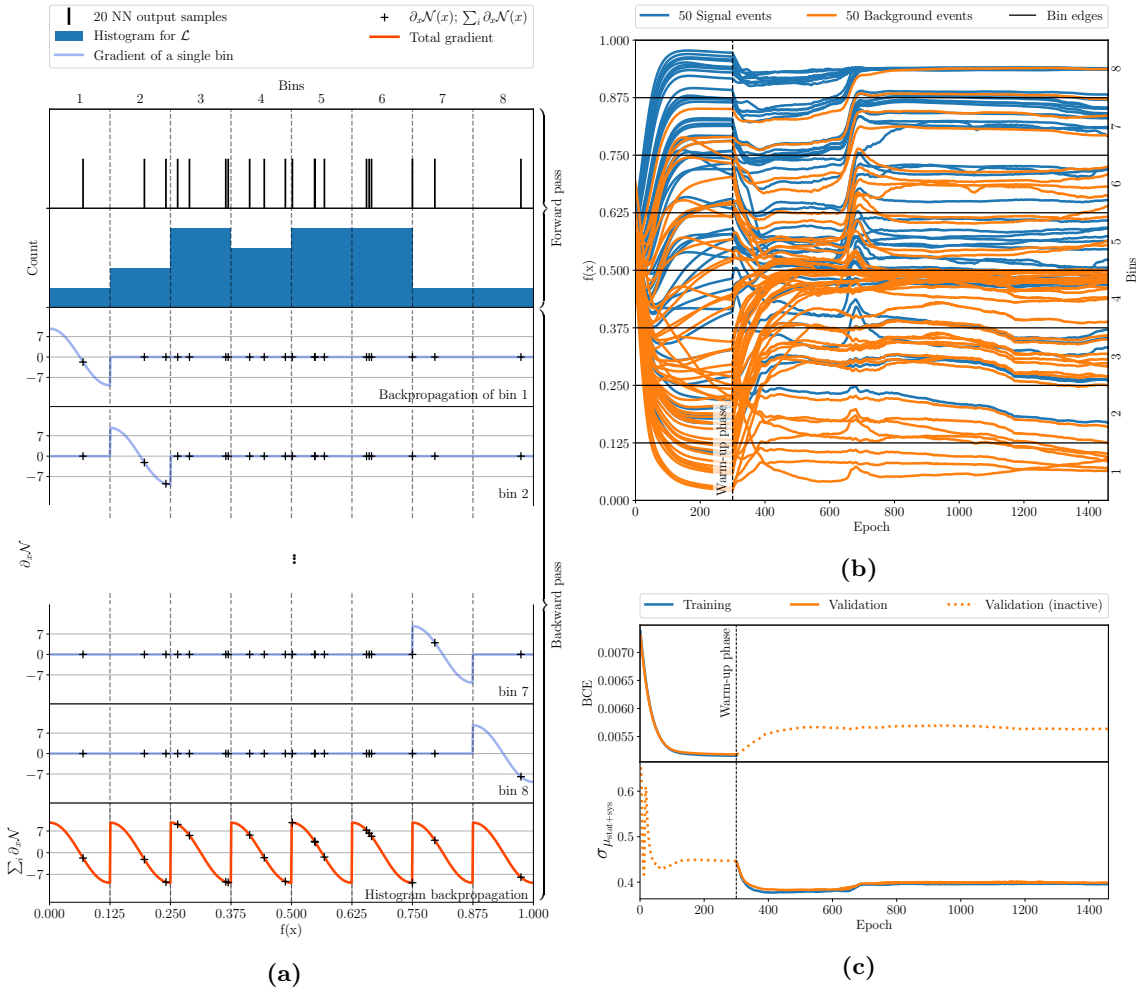


Figure 4.3.: Shown are the same quantities as in figure 4.2 differ in the bin-wise calculation of the histogram gradient, as shown in (a) by the introduction of the maximum affecting range of each bin gradient. This modification further affects the NN classification that does not result in the collapse into a few bins after the warm-up phase (dashed black line) as shown in (b) for the same 50 randomly selected signal and background events as in figure 4.2b. The corresponding loss evolution, shown in (c) indicates a stable training in comparison to figure 4.2c with a minimum at epoch 459.

and further depends on the weight initialization.
With such a collapse or the absence of a spread of events, the optimal minimum may not be found without a warm-up phase, as the performed search requires a persistent spread across all bins. In the following, a modification to the custom function is proposed that replaces the gradient during backpropagation with the goal to mitigate the collapse of the NN output function into a few histogram bins. This modification still relies on the gradient of a Gaussian distribution, but restricts its range to each corresponding bin boundary, thereby only locally affecting events within each bin. The impact of this modification on the total histogram gradient is shown in figure 4.3a. In contrast to figure 4.2a no long-range effects occur that affect the amplitude of the total gradient that shows the same maximal amplitude for bins in the center of the histogram as well as the outermost bins.
The backpropagation that contains this modification leads to a better distribution of events, reducing the tendency of NN outputs aggregation in only a small number of bins, as shown in figure 4.3b. This allows for a more stable search for an optimal result. The found loss of 0.382 in this course (figure 4.3c) is numerically comparable to the minimal loss of 0.388 found with the previously proposed modification of the backpropagation. The main

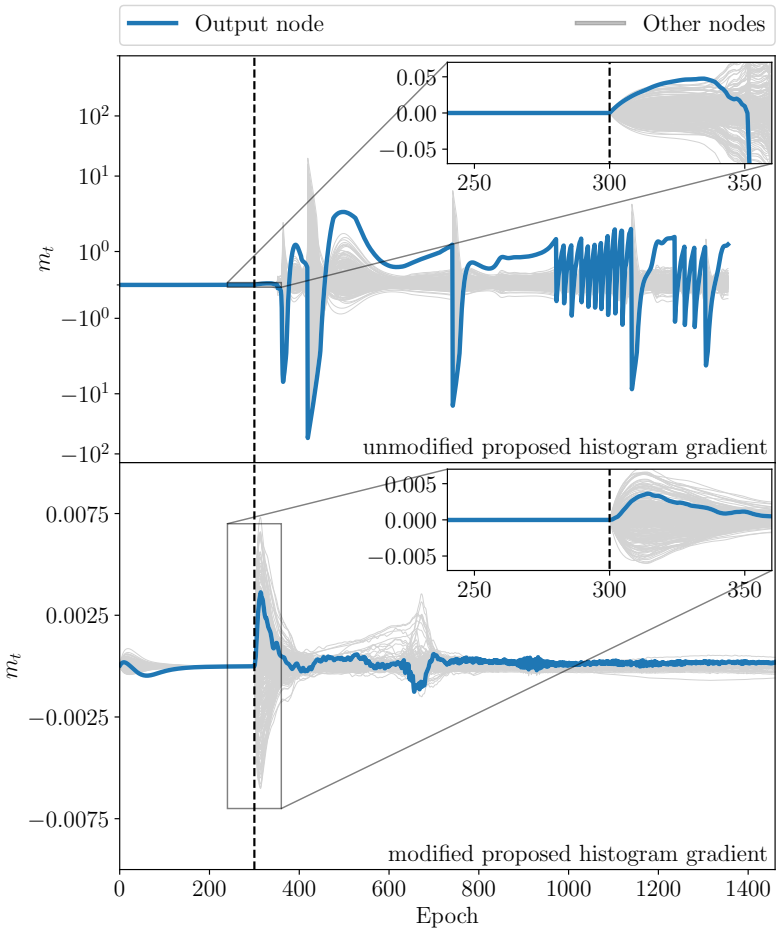


Figure 4.4.: Evolution of the momentum of the NN output node (blue) and all other nodes (gray) of the used NN. The in [3] proposed histogram gradient used during the backpropagation is shown in the top part and its modification as discussed in the text and illustrated in figure 4.3a is shown at the bottom. A window of $[240, 360]$ epochs is zoomed out to indicate the major difference of both gradients after the switch from BCE to σ_μ loss during the training. The shown range on the y-axis varies between the figures as the amplitude differs between both histogram gradients.

differences between the two approaches are the event distributions and the fact that with the previously proposed modification the optimal solution is only found during the collapse phase. With more complex data sets or additional uncertainties, the search for an optimal solution might take much longer and turn the search during the collapse phase unfeasible. The training procedure, as stated above, uses the Adam optimizer where the weights are updated by utilizing the adaptive momentum as shown in equation 4.1. Therefore, the collapse can also be examined by observing the momentum during the optimization process. The estimation of this momentum m_t is calculated for each weight in the NN after an epoch t for the full batch training by performing a sum of the momentum from the previous epoch m_{t-1} ($m_0 = 0$) multiplied by a constant decay rate β and the gradient of the current epoch g_t weighted with $(1 - \beta)$. The decay rate is chosen to be 0.9.

$$m_t = \beta m_{t-1} + (1 - \beta) g_t \,. \tag{4.1}$$

In figure 4.4, the evolution of the momentum of the NN output node is shown to indicate the differences between both custom functions used for the histogram gradient replacement. The momentum of the NN output node is indicated by the blue line, whereas all other neurons are shown in gray to indicate the general trend of the evolution. Both variants of

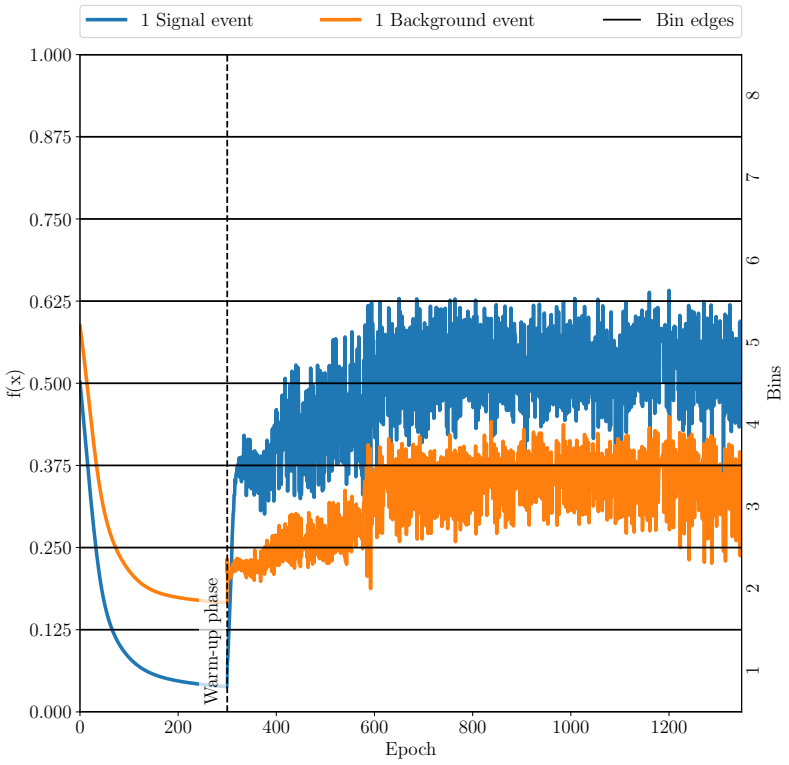


Figure 4.5.: Evolution of the NN classification of one selected signal and background event during the course of the training. For the training, the Adam optimizer was replaced by the NAdam and the learning rate after the warm-up phase was increased from 0.001 to 0.01.

the used histogram gradients show an increase in momentum shortly after the warm-up phase. The increase for the unmodified custom function (top) is about a magnitude larger than the momentum of the modified custom function (bottom) that is used for the gradient. In both cases, the momentum is decreasing shortly afterwards, but while the momentum of the modified gradient returns to the previously observed magnitude, the decrease in the unmodified gradient overshoots, followed by oscillations with amplitudes that exceed magnitude values up to 20. In comparison, the magnitude of the momenta of the modified custom function does not exceed $3 \cdot 10^{-3}$ after the initial increase of up to $7 \cdot 10^{-3}$. This oscillatory behavior can also be observed for the momenta of the other weights. In the case of the modified custom function that is used for the gradient replacement, a more independent evolution of the remaining momenta is observed after a decrease around epoch 400. From this observation, it can be concluded, that a reduction of the collapsing effect, caused by the introduction of the unmodified custom function for the histogram gradient, can be achieved and stabilize the training procedure as indicated by the evolution of the momenta. This enables further weight optimization after the warm-up phase, achieving an optimal result, as can be observed for the lower figure after epoch 400, where the optimizer further adapts several weights to the new training objective.
As a final step, in addition to the modification of the proposed function for the histogram gradient, an increase in the learning rate and the change of the optimizer from Adam to NAdam is applied. Thereby enhanced event movement improves the search for the optimal minimum by increasing the exploration of more bin combinations in order to find an optimal distribution of signal-enriched bins and a minimal effect from the systematic variations. NAdam incorporates the gradient from the previous optimization step g_{t-1} in addition to the current gradient g_t in the calculation of the momentum m_t. This extension of the calculation of the momentum provides the possibility to incorporate information

about different bin configurations in case of an undergone bin change of specific events between t and $t-1$. Since the numeric change in the σ_μ loss only occurs by a performed bin change of at least one event the introduction of an increased learning rate accelerates the rate of performed bin changes. The effect of those changes is illustrated for one signal and one background event in figure 4.5.
As this problem of the discretization of the NN output is fundamentally caused by the choice of the function that is used for the replacement of the histogram gradient it is not limited to the specific data set that is chosen in this chapter and can also be observed e.g. on the ATLAS Higgs Boson Machine Learning Challenge data set [67] as shown in appendix A.

4.3. Effect of the systematic-aware training on the neural network output function

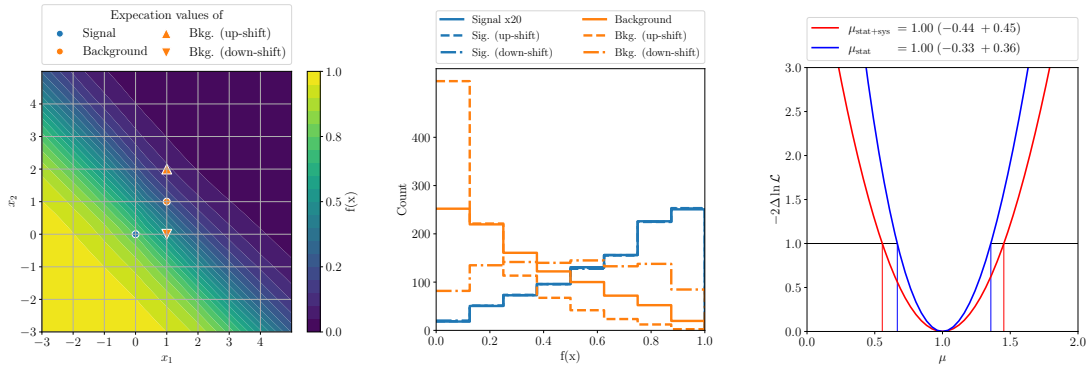
To evaluate the effects of the uncertainty-aware training, a comparison with a classical training method is conducted by performing training using the BCE loss, following the same procedure as described in section 4.1. The output plane of the best performing NN from the BCE training on the validation data set is shown in figure 4.6a, along with the mean values of the background, signal, and the up and down background shifts induced by the systematic variation. As expected, the BCE training is agnostic to the presence of systematic variations, as those are not incorporated into the training process. The training results in a decision plane that indicates the optimal spatial separation between signal and background. The nominal NN output of the best-performing NN, together with the propagated up and down variations for signal and background, are shown in figure 4.6b and used as inputs for the statistical inference to retrieve the signal strength with corresponding uncertainty. This result is presented in the form of a likelihood scan, shown in figure 4.6c. The blue scans represents the case of only statistical uncertainties ($\mu_{\text{stat}} = 1.00^{+0.36}_{-0.33}$), while in red both statistical and systematic uncertainties are considered leading to the benchmark value of

$$\mu_{\text{stat+sys}} = 1.00^{+0.45}_{-0.44}\,.$$

Performing the uncertainty-aware training as described in section 4.1 leads to an identical result of $\mu_{\text{stat}} = 1.00^{+0.36}_{-0.33}$ when considering only statistical uncertainties. Taking into account both statistical and systematic uncertainties leads then to an improved result of

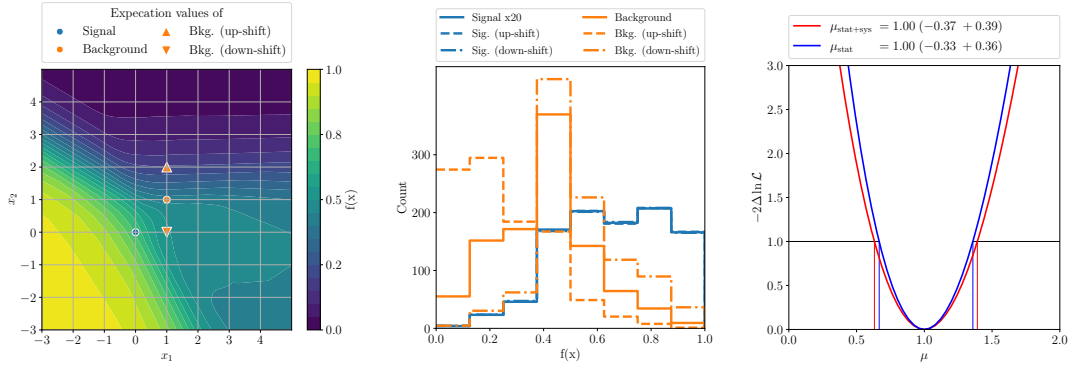
$$\mu_{\text{stat+sys}} = 1.00^{+0.39}_{-0.37}\,,$$

as can be seen in figure 4.7c. The shown values in the decision plane presented in figure 4.7b can not be interpreted anymore as the probability of an event being associated with signal or background, as for the BCE training. The order of bins is not relevant for the statistical inference nor for the calculation of σ_μ, the observed enhancement of signal events on the right side of the histogram does not indicate any inherent significance. Rather, this distribution is a result of the initial classification of signal events being on the right side after the warm-up phase. Similarly, the values around 0.5 in the decision plane (figure 4.7a) that predominantly include the regions affected by the systematic variation in the used data set should not be interpreted as an indication of the NN being indecisive about the association of the events to be signal or background. It should be noted that the training objective is to move background and signal events to create signal-enriched bins and minimize the effect of the introduced systematic variation. Moving the background events one bin away from the three signal-enriched bins seems to be sufficient separation, resulting in a larger concentration of background events to the left of a value of approximately 0.5. The movement of background events from the left of the histogram towards the central bin can mostly be attributed to the not completely eliminated effect of event aggregation into



(a) **(b)** **(c)**

Figure 4.6.: Results of the training on the BCE loss showing the decision-plane of the best performing NN on the validation data set in (a), the histogram of the NN output of the nominal and shifted signal and background processes in (b) and the likelihood scans for the signal strength estimation in (c). The scans in blue shows the statistical-only uncertainty, and in red the combined statistical and systematic uncertainty.

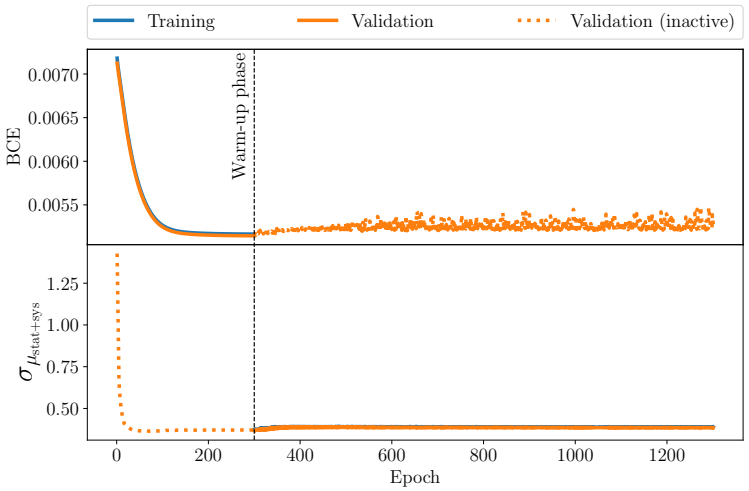


(a) **(b)** **(c)**

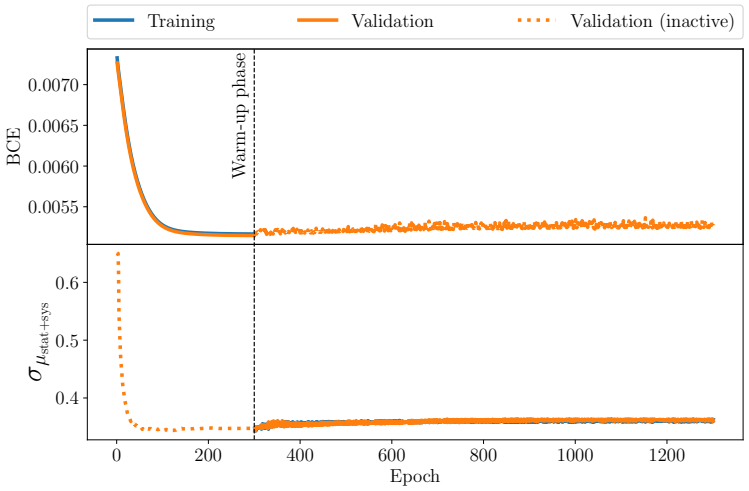
Figure 4.7.: Results of the uncertainty-aware training. The same quantities as in figure 4.6 are shown.

fewer bins as can be seen in figure 4.3b in the first 100 epochs after the warm-up phase. Subsequently, two additional types of uncertainty-aware training can be performed: considering systematic variations that are induced by normalization uncertainties or the absence of any systematic variations. An example for a normalization uncertainty can be given in form of a 10 % up and down variation of the background process. This corresponds to a weight correction factor of 1.1 (0.9) in case of upward (downward) variation for all background events. The absence of systematic variations is achieved by calculating σ_μ without the addition of any systematic variations laying the focus on the minimization of the statistical uncertainty.

The loss evolution for the normalization-induced systematic variation is shown in figure 4.8a and indicates, that the uncertainty aware-training is incapable of addressing this type of systematic variation as it shows no further improvement after the warm-up phase. A similar observation is seen for the case of an absent systematic variation in figure 4.8b, showing that the best possible result of minimizing the statistical uncertainty of σ_μ is achieved by BCE. The strategy of moving background events that create a systematic variation out of signal-enriched bins is not applicable in case of normalization uncertainties since every background event contributes to the normalization uncertainty. A shift of any background event would also affect nearby signal events thus resulting in a worsening of statistical uncertainty upon movement. In contrast, the presence of any sufficiently large



(a)



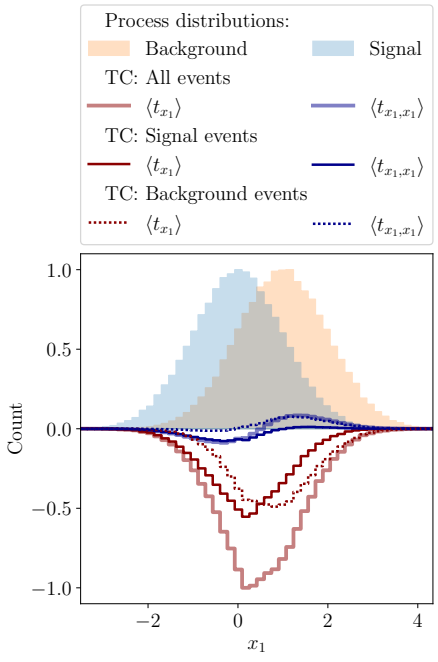
(b)

Figure 4.8.: Loss evolution of the uncertainty-aware training in case of (a) 10 % uncertainty in the normalization of the background process and (b) only statistical uncertainties. The vertical dashed lines indicate the end of the warm-up phase. Evaluated dotted (inactive) validation losses are not used for the validation step as discussed in section 4.2.

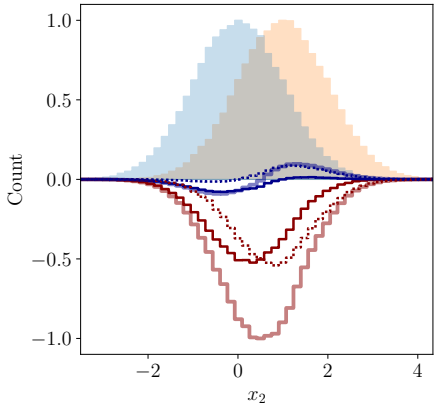
shape-altering systematic variation does not affect all events equally and forces the NN to find a trade-off between an increase of the reduced statistical uncertainty as obtained during the warm-up phase and the reduction of the uncertainty resulting from the systematic variation. The determination of a sufficiently large shape-altering systematic variation in the given example and the effects of the NN decision are investigated in the following section.

4.4. Neural network decision taking in presence of systematic uncertainties

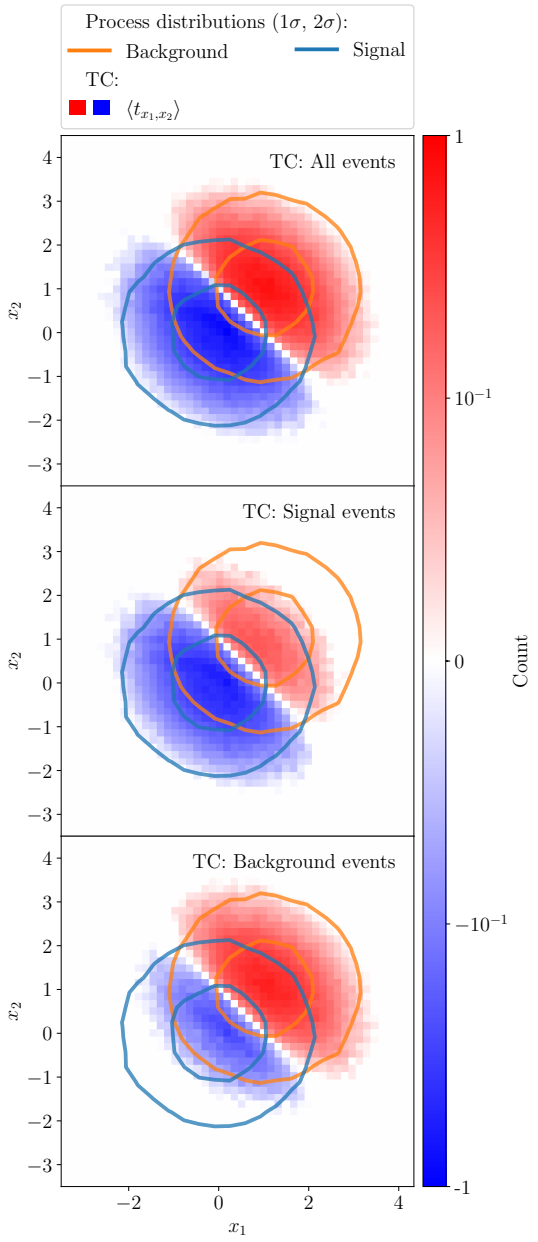
To identify the most influential input variables on the NN output function, either individually or in combination with other variables, a TCA, as described in [68], can be performed. For this analysis, only Taylor coefficients (TC) of the first $\langle t_{x_i} \rangle$ and second order $\langle t_{x_i,x_j} \rangle$ are considered. The approach followed here deviates from [68] by not summarizing the TCs, that are obtained from the test data set as the aim is set to the identification of specific regions in the parameter space that are important for the NN decision.
The derived TCs from the test data set are used as weights for one or two-dimensional histograms that are created from the NN input variables of the test data set. Thereby



(a) TCs of: $\langle t_{x_1} \rangle$, $\langle t_{x_1,x_1} \rangle$



(b) TCs of: $\langle t_{x_2} \rangle$, $\langle t_{x_2,x_2} \rangle$



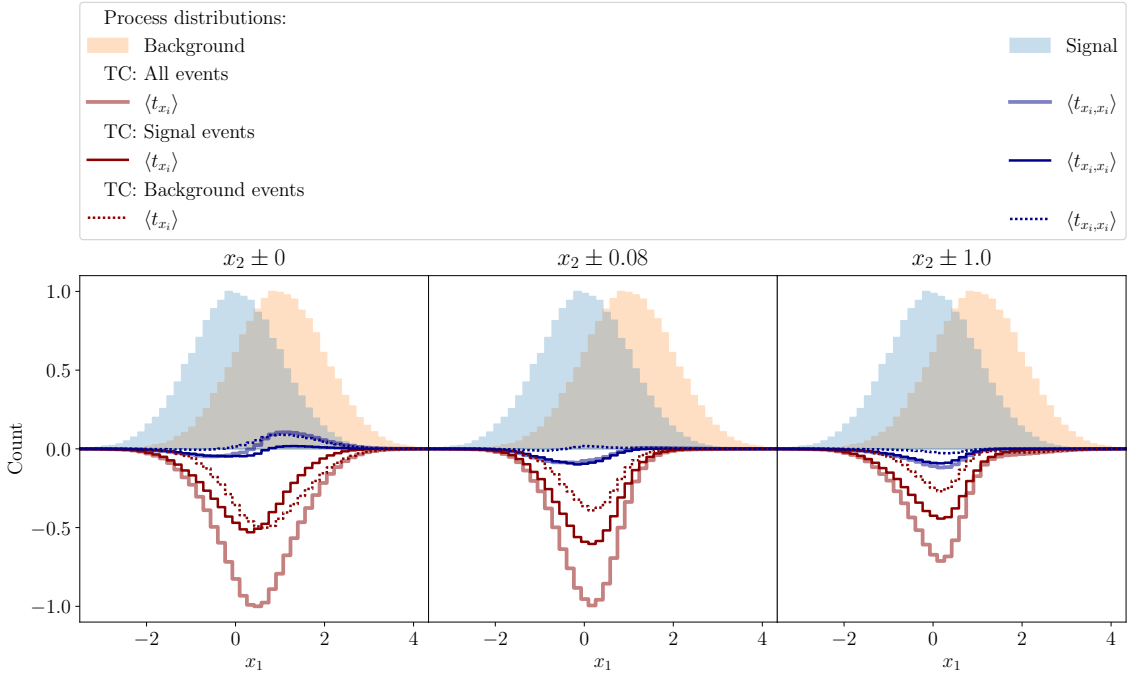
(c) TC of: $\langle t_{x_1,x_2} \rangle$

Figure 4.9.: Distributions of event coordinates weighted by their corresponding TCs with (a) for $\langle t_{x_1} \rangle$ and $\langle t_{x_1,x_1} \rangle$, (b) for $\langle t_{x_2} \rangle$ and $\langle t_{x_2,x_2} \rangle$ and (c) for $\langle t_{x_1,x_2} \rangle$. Filled histograms in (a) and (b) as well as 1σ and 2σ contours in (c) visualize the signal and background process. The unfilled histograms in (a) and (b) show the first (second) order TC in red (blue). Each TC is further divided upon the used events for the TC calculation by unfilled histograms with solid thick lines for all events, solid thin lines for signal events, and thin dotted lines for background events. This separation in (c) is shown by three different subfigures with a corresponding annotation.

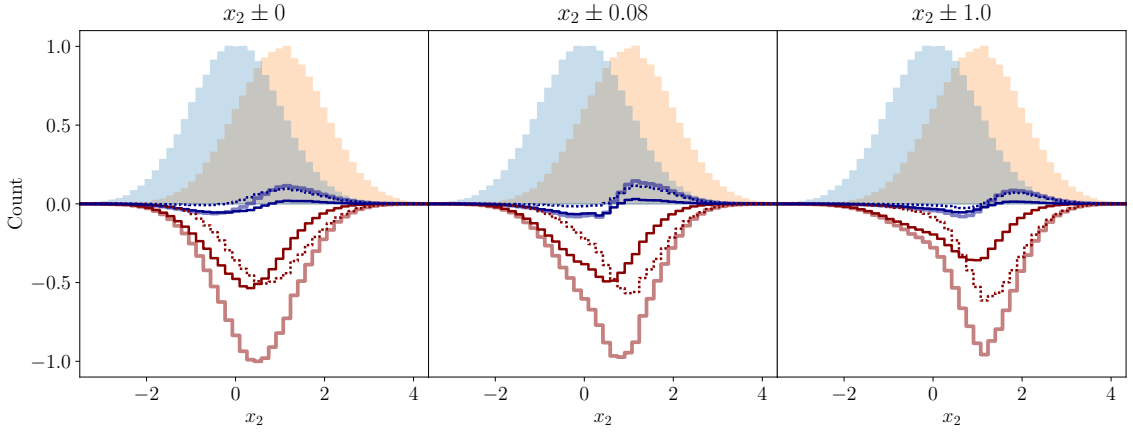
resulting histograms are exemplarily created on the performed benchmark from section 4.3 and are shown in figure 4.9. The one dimensional histograms of $\langle t_{x_1} \rangle$ and $\langle t_{x_1,x_1} \rangle$ TCs are summarized in figure 4.9a, whereas the $\langle t_{x_2} \rangle$ and $\langle t_{x_2,x_2} \rangle$ TC are summarized in figure 4.9b. The $\langle t_{x_1,x_2} \rangle$ TC is shown in form of a two-dimensional histogram in figure 4.9c. The filled histograms in figures 4.9a and 4.9b correspond to the signal and background distributions. These have been added for illustrative purposes to provide a context of the associated TCs with respect to the events in the existing data set. For $\langle t_{x_1,x_2} \rangle$ in figure 4.9c, the signal,

and background processes are shown as 1σ and 2σ contours. Further, all histograms of a figure are scaled by an arbitrary factor to display the relative difference between different TCs of the same input variables.
The unfilled histograms in figures 4.9a and 4.9b show the first and second order TCs of the x_1 and x_2 coordinates respectively with the red colored histograms indicating the first order and in blue the second order TCs. The thick light red and blue unfilled histograms represent the corresponding TCs that are created using all events of the test data set. TCs that are created only using signal (background) processes are indicated by the unfilled histograms with thin red and blue solid (dotted) lines. The filled histograms in figure 4.9c are equivalent to the TC histograms from the one-dimensional distributions of figures 4.9a and 4.9b. The separation of the TCs in figure 4.9c are given by three subfigures that show TCs calcuated on all events in the upper subfigure, followed by the TCs that are calculated using only signal events in the middle subfigure and TCs calculated on background events in the subfigure at the bottom.
The high amplitude of the first-order TC relative to the second-order TC is partially induced by the use of the ReLU activation function in the hidden layer. Due to this choice, the second derivatives yield non-zero values only for the derivatives of the NN output node. Based on figures 4.9a and 4.9b, the first derivative with respect to the given coordinates appears to be most significant for the decision of the NN in the overlap region between the signal and background process. Further, a distinction between signal and background events of the first derivative can be seen with the maximal values being located closer to the corresponding processes. Second-order TC display a definite process contribution, as indicated by the sign flip in their distribution, where the sum of second-order TCs is close to zero except in the vicinity of the mean of the corresponding processes, which exhibits negative (positive) TC values for TCs from signal (background) events.
To investigate the effect of the uncertainty-aware training on NN output function the described procedure for the creation of the TC histograms is applied on the systematic variation, as introduced in section 4.1 and varied from 0 to ± 1. This variation is performed in steps of 0.01 between 0 and 0.2 and in steps of 0.1 onwards. In order to account for any fluctuations resulting from the random weight initialization 100 trainings for each changing systematic variation are performed and the mean of thus resulting Taylor coefficients is used as a weight for the described Taylor coefficient histograms. The histograms of the Taylor coefficient that results from the systematic variations of 0, 0.08, and 1.0 are shown in figures 4.10a, 4.10b and 4.10c. The effects between 0.08 and 1.0 show a gradual change in the TCs with increasing systematic variations and no change between 0 and 0.08 is observed.
For an absent systematic variation, the obtained Taylor coefficients are comparable with the BCE training, which is assuring in view of the discussion in section 4.3 and can be depicted from the left figures in figure 4.10 where in case of $\langle t_{x_1,x_2} \rangle$ the left three figures show a separation similar to figure 4.9c. A deviation from this state can be seen after the systematic variation exceeds the value of 0.08 which is data set and task-specific.
The distinction between signal and background events from $\langle t_{x_1} \rangle$ vanishes, as the TCs originating from signal and background events, overlap and move towards the mean of the signal process as can be seen in the central subfigure in figure 4.10a. Further, the overall importance of $\langle t_{x_1} \rangle$ and the contribution from the background events decreases with an increasing systematic variation, visible by comparing the TCs of $x_2 \pm 0.08$ and $x_2 \pm 1.0$ whereas the $\langle t_{x_1,x_1} \rangle$ maintains its importance, for the signal events. In contrast, $\langle t_{x_2} \rangle$ and $\langle t_{x_2,x_2} \rangle$ preserve their importance with increasing systematic variations but also perform a shift, in this case towards the background process. The $\langle t_{x_2} \rangle$ TC maintains the long tail in the region of signal process but decreases in importance upon the increase of the systematic variation as can be seen in the middle and right subfigure of figure 4.10c.
With increasing systematic variations the importance of $\langle t_{x_1,x_2} \rangle$ (figure 4.10c) decreases

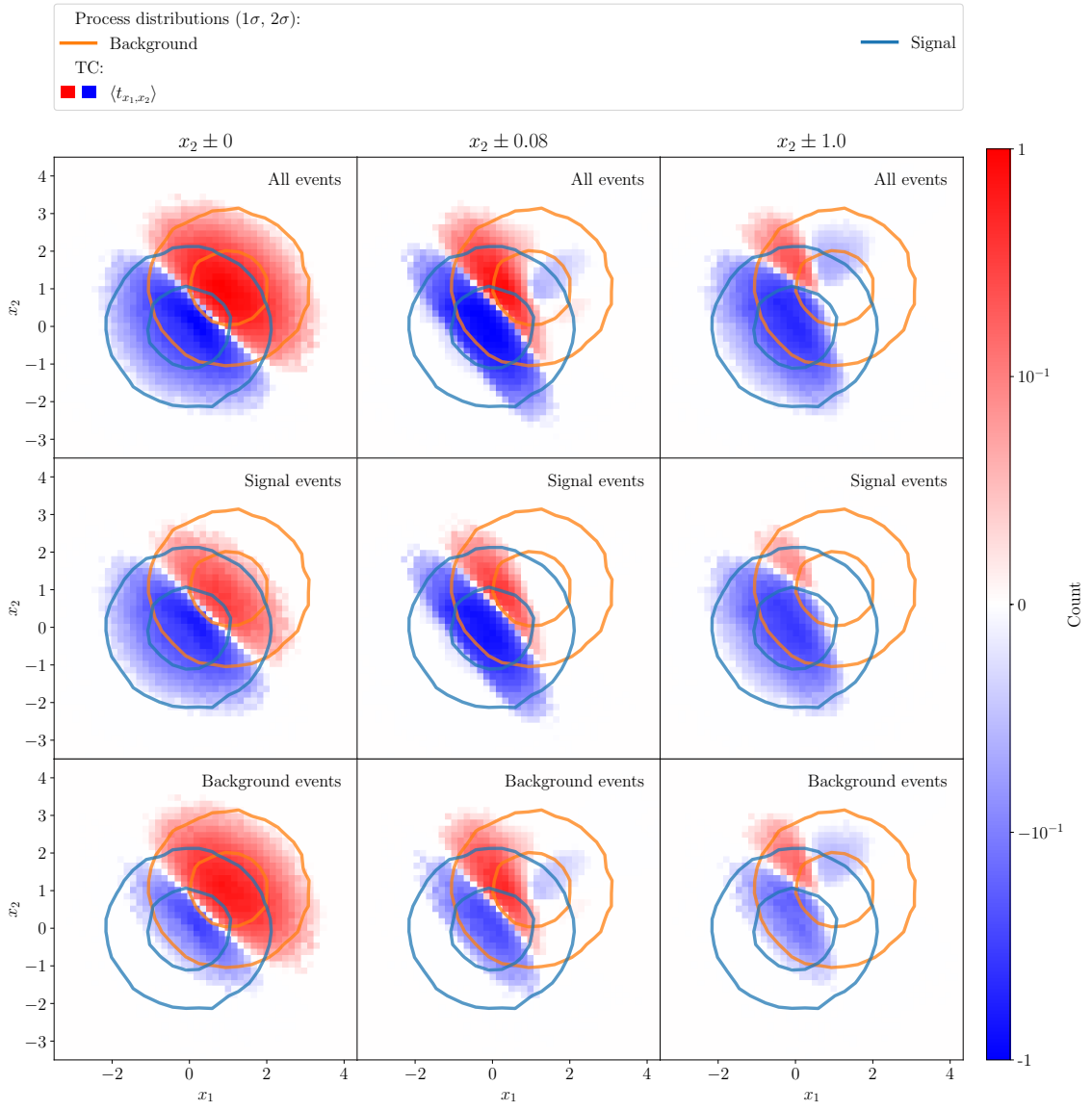
similarly to $\langle t_{x_1} \rangle$ and $\langle t_{x_1,x_1} \rangle$ in comparison to the BCE training. Especially the background and signal process regions that are affected by the systematic variations lose importance for the decision of the NN since the possibility of the selection of signal processes decreases with increasing systematic variation. On the other hand, the signal region that is unaffected by the systematic variations maintains its importance with increasing systematic variation as it remains the only region of signal processes that are less affected by the systematic variations and has the largest contribution to the signal enriched bins of figure 4.7b.
Hence the overall observation indicates a decorrelating effect of most of the second-order TCs upon the application of uncertainty-aware training in comparison to the BCE training, no longer exerting a separating power between signal and background processes. The focus is shifted towards the identification of signal-enriched regions in case of $\langle t_{x_1} \rangle$, $\langle t_{x_1,x_1} \rangle$ and $\langle t_{x_1,x_2} \rangle$. The $\langle t_{x_2} \rangle$ and $\langle t_{x_2,x_2} \rangle$ TCs shows a shift towards the background process but otherwise remains comparable to the results from the BCE training.



(a) TCs of: $\langle t_{x_1} \rangle$, $\langle t_{x_1,x_1} \rangle$



(b) TCs of: $\langle t_{x_2} \rangle$, $\langle t_{x_2,x_2} \rangle$



(c) TC of: $\langle t_{x_1,x_2} \rangle$

Figure 4.10.: The same quantities as in figure 4.9 are shown. The figures are further divided upon the addressed systematic variation during the uncertainty-aware training of $x_2 \pm 0$ in the left, $x_2 \pm 0.08$ in the middle, and $x_2 \pm 1$ in the right subfigure.

5. Extension of the pseudo experiment to multiple classes

In this chapter, an extension to the application of the uncertainty-aware training is performed in order to solve classification tasks with multiple classes. Two approaches are presented in this chapter. A comparison of the effectiveness of the introduced modification and the unmodified approach of uncertainty-aware training in presence of multiple classes is conducted afterwards with regard to an extension to more complex problems.

5.1. Extension of experiment configuration

For an application of uncertainty-aware training to a task that introduces multiple classes modifications to the existing setup are required. Two additional processes that are modeled similarly to the signal and background process in the binary case are added to the data set with an additional signal process at $(x_1, x_2) = (0.5, 3)$ and a second background process at $(x_1, x_2) = (-1, 2)$ as shown in figure 5.1. This extension increases the total number of events of the training, validation, and test data set by a factor of two. In addition to the existing systematic variation of the first background process, a second systematic variation affecting the second background process is introduced, with a ± 1 variation along the x_1 axis. Additionally, the signal processes are reweighted to yield 100 signal events per signal process instead of the previously chosen 50 events for the binary classification. The yield of 1000 events for each background process remains unchanged. An increase in the yields of the signal processes reduces the resulting dominating statistical uncertainty of the signal strengths that inevitably increases due to the introduction of additional background processes shifting the importance back to addressing the reduction of the systematic uncertainties introduced by the systematic variations.
To align the network architecture with the extended data set, the number of output nodes is raised to four, assigning each process uniquely to an output node. The Sigmoid activation function of the output node is replaced with the Softmax activation function and the loss function is changed to CE. With this regard, the loss of the warm-up phase of the uncertainty-aware training is also changed from BCE to CE in order to align with the binary application by addressing the minimization of statistical uncertainty during the warm-up phase.
The class assignment of propagated events for the uncertainty-aware training is performed analogously to the class assignment that is performed for the statistical inference step as described in section 3.2. The likelihood used for the computation of σ_μ and the statistical inference in the binary case as defined in equation 3.4 is extended to multiple classes with C_{class} being the total number of classes and P (P') distinguishable signal s_k (background $b_{k'}$) processes as shown in equation 5.1. With this extension, the introduced nuisance parameters $\{\theta_j\}$ that represent the systematic uncertainties affect specific signal or background processes or a combination of those. In the presence of multiple signal

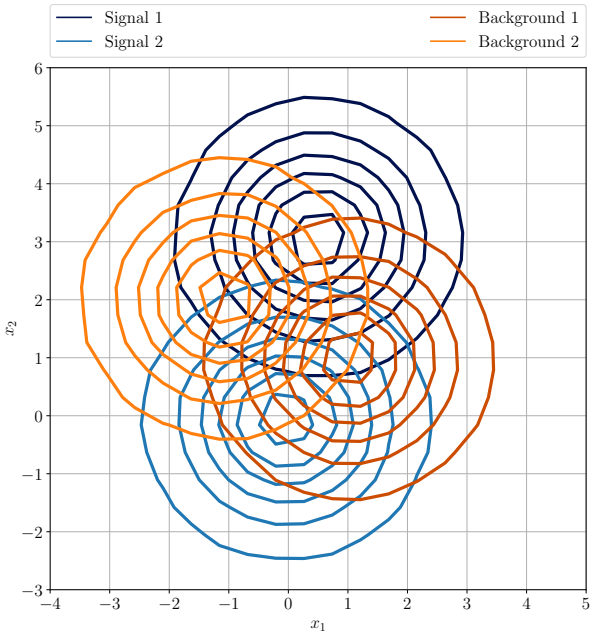


Figure 5.1.: A representation of the signal and background processes where each process is modeled with 10^5 events by a two-dimensional Gaussian distribution each with a two-dimensional unity matrix used for the covariance matrix for all processes. The mean value used to model the distributions is $(x_1, x_2) = (0, 0)$ for Signal 1 and $(x_1, x_2) = (0.5, 3)$ for Signal 2 process. The mean value of the Background 1 process is set to $(x_1, x_2) = (1, 1)$ whereas the mean of the Background 2 is chosen to be at $(x_1, x_2) = (-1, 2)$.

processes the training objective of the uncertainty-aware training is extended to a sum of signal strength uncertainties, denoted as $\sum_i \sigma_{\mu_i}$.

$$\mathcal{L}\left(N, \mu, \{\theta_j\}\right) = \prod_{c=1}^{C_{\text{class}}} \prod_{i=1}^{N_{\text{bins}}} \mathcal{P}\left(n_i \middle| \sum_{k=1}^{P} \mu_k s_{i,k}\left(\{\theta_j\}\right) + \sum_{k'=1}^{P'} b_{i,k'}\left(\{\theta_j\}\right)\right) \times \prod_{j=1}^{M} \mathcal{C}\left(\theta_j \middle| \mu_{\theta_j}, \sigma_{\theta_j}\right) . \tag{5.1}$$

5.2. Application in presence of two systematic uncertainties and multiple classes

To indicate the improvement in the signal strength uncertainties by applying the uncertainty-aware training in presence of multiple classes a benchmark is conducted, by performing a CE training. To avoid problems from overtraining a lower patience threshold of only one epoch is chosen since the combination of the simplified data set and the chosen full-batch training strategy allows for fast learning of the correct process assignment to the appropriate classes. The resulting classification of the CE training is shown in figure 5.2a, where bins that contained fewer than 10 events have been merged with the neighboring bins for the statistical inference. As the occurrence of low-populated or empty bins is observed in the histogram edges the merging is performed adding those bins with their neighboring bins towards the center of the histogram until the condition of 10 events in the combined bin is fulfilled. Additionally to the nominal processes the effect of the systematic variation of the Background 2 in the x_1 plane by ± 1 is shown in gray, where all up and downward variations of individual background processes resulting from this shift are added up in each bin. The resulting variations on signal processes are shown individually. The evolution of the loss, displayed similarly to the loss evolution in chapter 4, is shown in figure 5.2b a

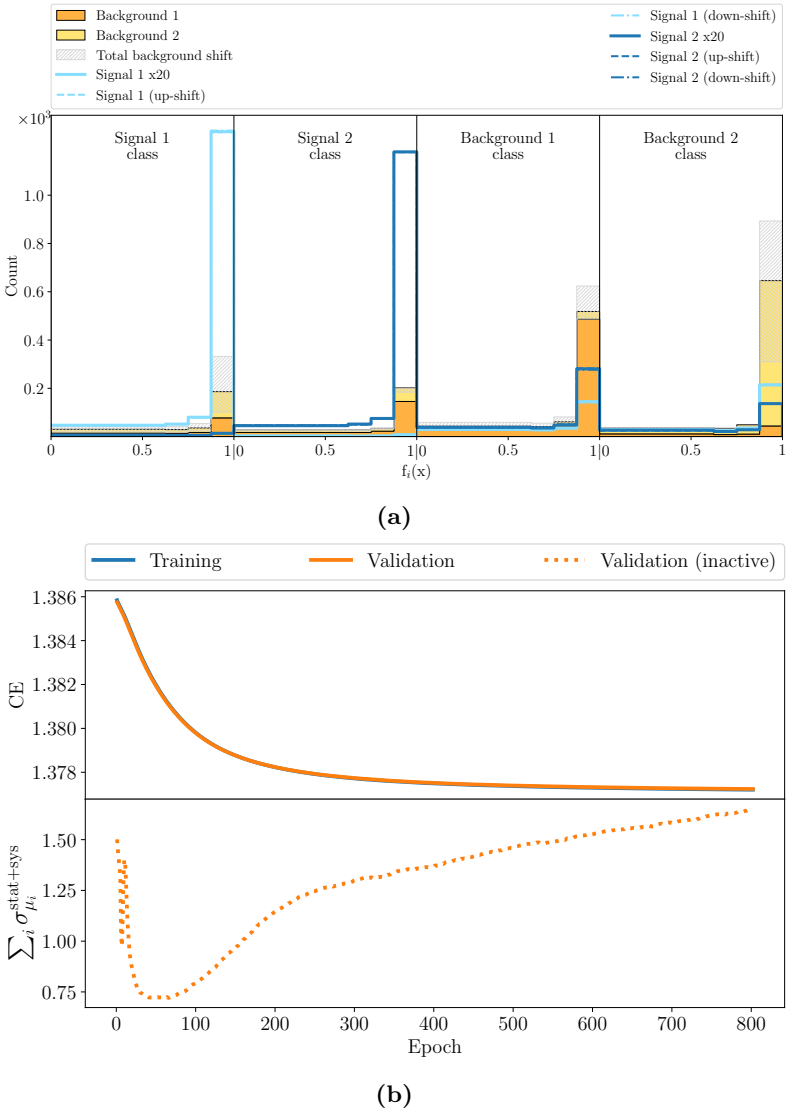
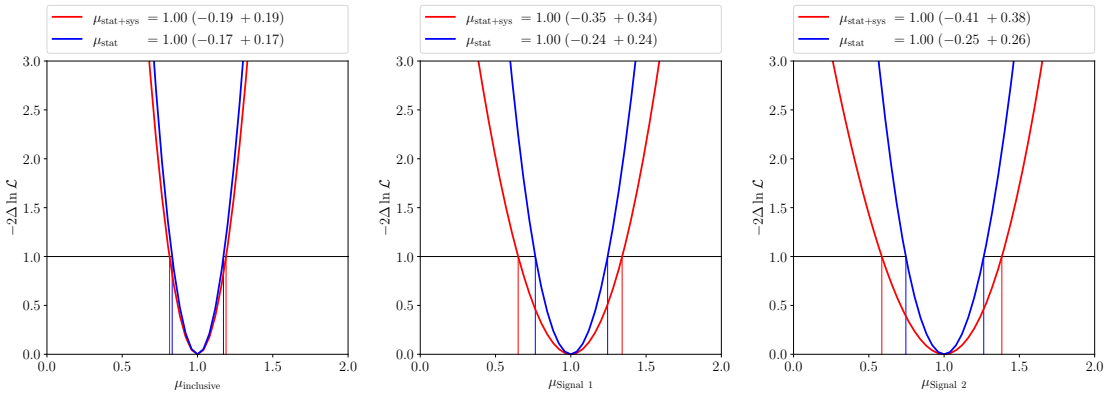


Figure 5.2.: NN output on the test data set after the CE training is shown in (a) displaying nominal signal and background processes. The total effect of the systematic variation of Background 2 by ± 1 along x_1 is summed up for all background processes in each bin and is shown in gray. The effect of systematic variation on signal processes is displayed separately for each signal process. Further information about binning is given in the text. The loss evolution, shown in (b), is displayed similarly to chapter 4 and shows the evolution of the CE loss in the upper part and the illustrative $\sum_i \sigma_{\mu_i}$ loss in the lower part.

displays a continuous decrease of the CE loss in the upper part of the figure. On the other hand, the illustrative displayed $\sum_i \sigma_{\mu_i}$ loss increases after a short decrease to a minimum around 0.75 in the first 100 epochs. It indicates, that the optimization based on the CE does not automatically imply a minimal result also for $\sum_i \sigma_{\mu_i}$, particularly in presence of systematic variations. This can be explained by the aggregation of the events into a few bins leading to an increase in $\sum_i \sigma_{\mu_i}$ as discussed in chapter 4. Resulting likelihood scans extracted from the signal strength estimations after the statistical inference are presented in figure 5.3. Figure 5.3a shows the inclusive signal strength of both signal processes. Figures 5.3b and 5.3c show the results for one of the signal strengths being selected as the



(a) **(b)** **(c)**

Figure 5.3.: Inclusive signal strength estimation from the CE training are shown in (a). Further, (b) shows the differential signal strength estimation of Signal 1 and (c) of Signal 2. The blue likelihood scans show statistical-only uncertainties whereas the red scans display statistical and systematic uncertainties.

parameter of interest. The signal strength uncertainties from the performed benchmark, taking into account statistical and systematic uncertainties, results in:

$$\mu_{\text{inclusive}} = 1.00^{+0.19}_{-0.19}\,,$$
$$\mu_{\text{Signal 1}} = 1.00^{+0.34}_{-0.35}\,,$$
$$\mu_{\text{Signal 2}} = 1.00^{+0.38}_{-0.41}\,.$$

In the next step, uncertainty-aware training is applied with the in section 5.1 discussed changes. The application is expected to result in a slight deviation from the results of the CE training, as obtained after the warm-up phase, accounting for the presence of systematic variations in the final distribution. Figure 5.4a shows an example training that follows this expectation. The distributions are broadened with respect to the CE training and the event associated to their assigned event classes are preserved. From the loss evolution in figure 5.4b the observed increase in $\sum_i \sigma_{\mu_i}$ during the warm-up phase is successfully minimized after the warm-up phase. The CE loss shows a steep increase remaining at a plateau of around 1.380 for the rest of the training. This increase can be mainly attributed to the observed movement of events from the right, broadening the distribution. The signal strength estimations, given this result of the uncertainty-aware training, are shown in figure 5.5 and are split similarly to the benchmark, to:

$$\mu_{\text{inclusive}} = 1.00^{+0.17}_{-0.16}\,,$$
$$\mu_{\text{Signal 1}} = 1.00^{+0.25}_{-0.24}\,,$$
$$\mu_{\text{Signal 2}} = 1.00^{+0.29}_{-0.28}\,.$$

A central issue of the uncertainty-aware training in this extension to multiple classes is that prior assumptions about the event classification induced by the CE warm-up phase are not taken into account anymore during the training on $\sum_i \sigma_{\mu_i}$. The introduction of additional classes as stated in equation 5.1 leads only to an introduction of additional bins, but the order of the bins or the information about the allocated classes is not used beyond the CE warm-up phase. This problem becomes more pronounced when introducing additional processes, additional systematic variations, or a more complex task than the one outlined in this chapter. This is demonstrated by the introduction of an additional systematic variation increasing the number of total systematic variations up to three. The third introduced systematic variation affects the second background, introducing its variation

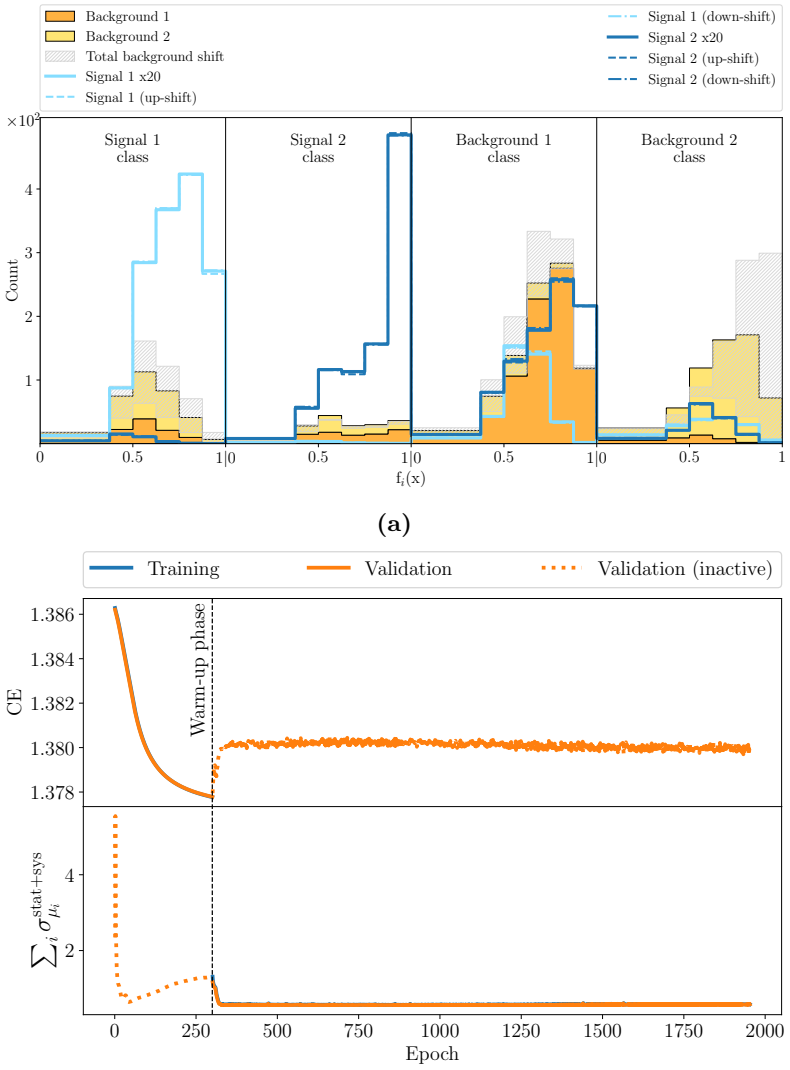


Figure 5.4.: Resulting NN output (a) and the loss evolution (b) of an uncertainty-aware training in presence of multiple classes that preserves the event assignments to classes from the warm-up phase on CE loss. The same quantities as in figure 5.2 are shown.

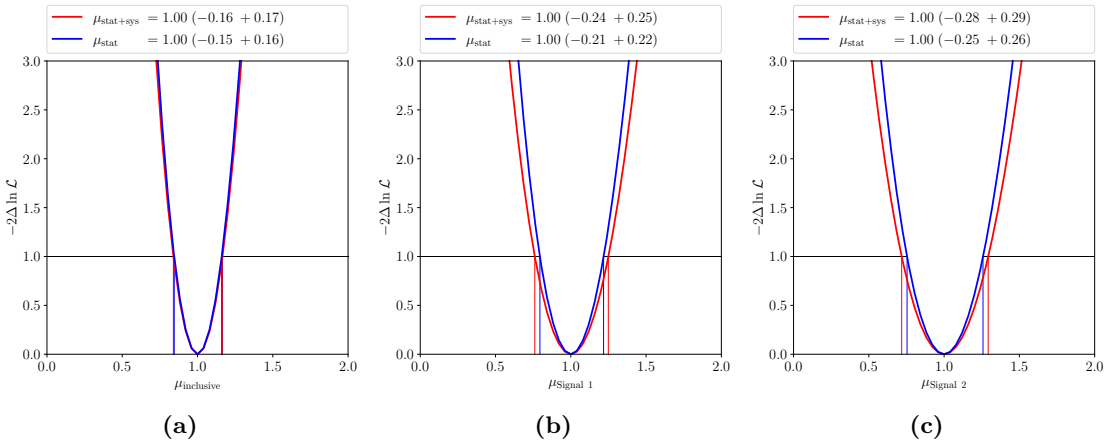


Figure 5.5.: Results of the statistical inference after the uncertainty-aware training of figure 5.4 with similar quantities as shown in figure 5.3, displaying (a) the inclusive signal strength estimation and the differential signal strength estimations of (b) Signal 1 and (c) Signal 2.

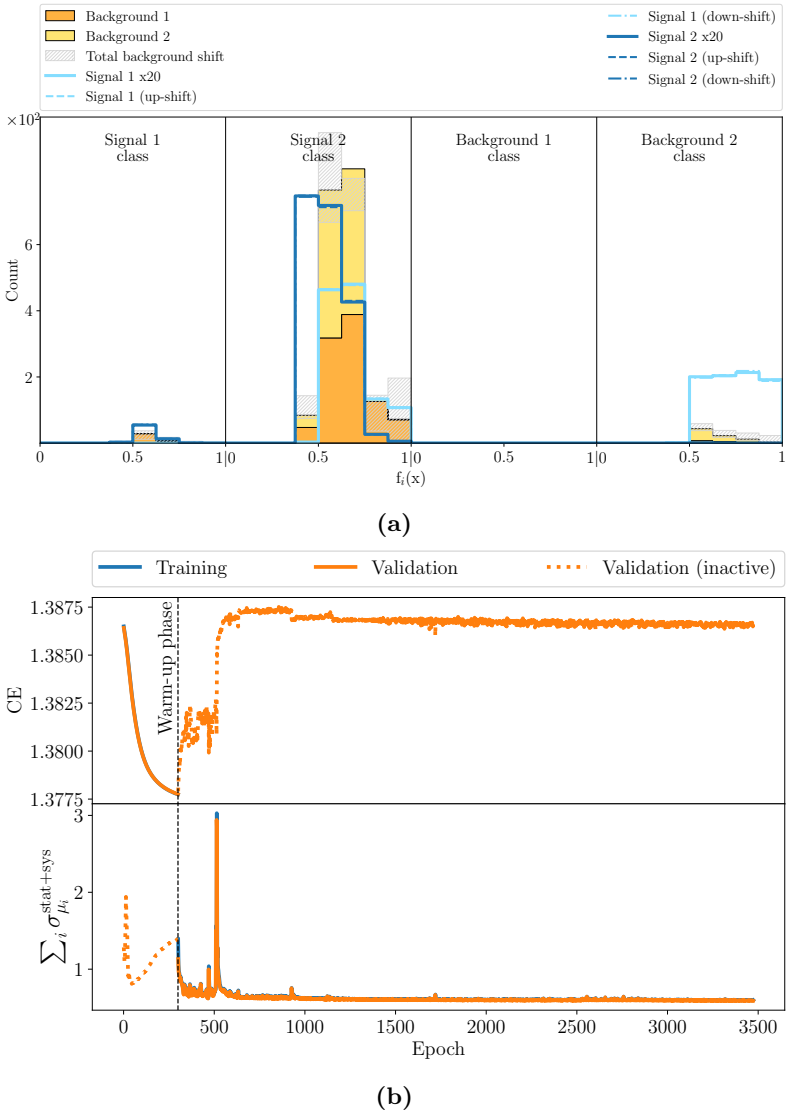


Figure 5.6.: Resulting NN output (a) and the loss evolution (b) of an uncertainty-aware training in presence of multiple classes that shows the problem of empty classes and misclassified events. Shown are the same quantities as in figure 6.6 but not merged low-populated bins inside of not empty classes.

along the x_2 plane by ± 1 analogous to the systematic variation on the first background process. The NN output of the best-performing epoch on the validation data set is shown in figure 5.6a. The corresponding loss evolution is shown in figure 5.6b and indicates a similar behavior of the NN in the first 200 epochs after the warm-up phase, where the broadening of the distributions in each class leads to a steep increase in the CE loss that then stays at a plateau with a slightly decreasing value. This training however shows a second steep increase after epoch 500 to a new plateau, indicating a further change in the event classification. The short increase in $\sum_i \sigma_{\mu_i}$ during the same epochs indicates a grouping and redistribution of events. The resulting distribution of the best-performing NN of this training, indicates a deviation in the event assignment from the classification obtained by the CE training and the appearance of empty classes. The problem of misclassification is mainly illustrated by the Signal 1 events being classified and further enriched in the Background 2 class or the increased number of background events in the Signal 2 class. Further, the Background 1 class remains empty after the redistribution. This result is not unexpected as the training process does not have to reproduce the previous class

assigned of events or populate the introduced classes after their changes when training on $\sum_i \sigma_{\mu_i}$. In terms of the training process the performed redistribution decreased $\sum_i \sigma_{\mu_i}$ and the loss evolution shows no further irregularities in the training process, finding the best result around epoch 2400. The goal of the uncertainty-aware training is only the minimization of $\sum_i \sigma_{\mu_i}$, which is not penalized in any way, upon the removal of events from their associated classes. Based on the NN output, two main issues can be stated. The first issue is the lack of a classification of events during training into predefined classes, leading to misclassification in the standard interpretation. Since the minimization of $\sum_i \sigma_{\mu_i}$ as the training objective does not correspond to a standard classification task, the weight optimization can lead to phase space regions where such a form of misclassification may occur. The second related issue is the potential occurence of empty classes. In general, the presence of very low-populated or even empty bins may lead to technical problems during the statistical inference and needs to be addressed, e.g. by merging of low-populated bins. Aside from its necessity, the application of bin merging implies a deviation from the actual training result, as the smaller number of bins was not used during training. This should be avoided as the bin configuration where the systematic variations of neighboring bins lead to a cancelation of systematic variations may be compromised. The problem of empty classes, however, cannot be solved by bin merging and can lead to difficulties during the statistical inference if the application to data might results in a classification of events inside an empty class. Two approaches are presented in the following two sections to address the issue of empty classes and potentially maintain the desired classification as implied by the warm-up phase.

5.2.1. One-class modification

The following ansatz addresses two issues of the unmodified multi-class uncertainty-aware training. One of the issues is the emerging empty classes, while the second issue poses a more fundamental problem of losing NN output information when performing the training. During the training on $\sum_i \sigma_{\mu_i}$ in case of multiple signal classes or on σ_μ for one signal class, all non-maximal NN output nodes are discarded for every event, leaving only the output node with the maximum value, which is then used for uncertainty-aware training. This discard of NN output reduces the amount of information that can be used for weight optimization during the training.
Both issues can be addressed by using only one class, reducing the number of NN output nodes to one, and switching back to the Sigmoid activation function. During the warm-up phase, all signal and background processes are combined into one signal and one background class correspondingly for the training on the BCE, resulting in a binary classification task, as described in chapter 4. When switching to the training based on $\sum_i \sigma_{\mu_i}$, all individual processes are distinguished again considering all signal and background processes for the loss minimization. To account for the fact that additional bins are introduced when considering multiple classes, as discussed in section 5.1, the number of used bins for the training is increased from previously used 8 to 16 bins.
A typical result from a performed training is shown in figure 5.7. The NN output exhibits signal-enriched regions near histogram edges with the remaining background events being classified towards the central bins. The interpretation is comparable to the result obtained for the binary case in section 4.3 with the main difference of multiple present signal processes that leaves the position of signal-enriched bins within the histogram arbitrarily up to the point of least overlap between the signal processes and the remaining backgrounds.
The resulting signal strength estimations are presented in form of a performed ensemble test in section 5.3 and are compared to the naive uncertainty-aware training of the previous section and the second introduced modification that is presented in the following section.

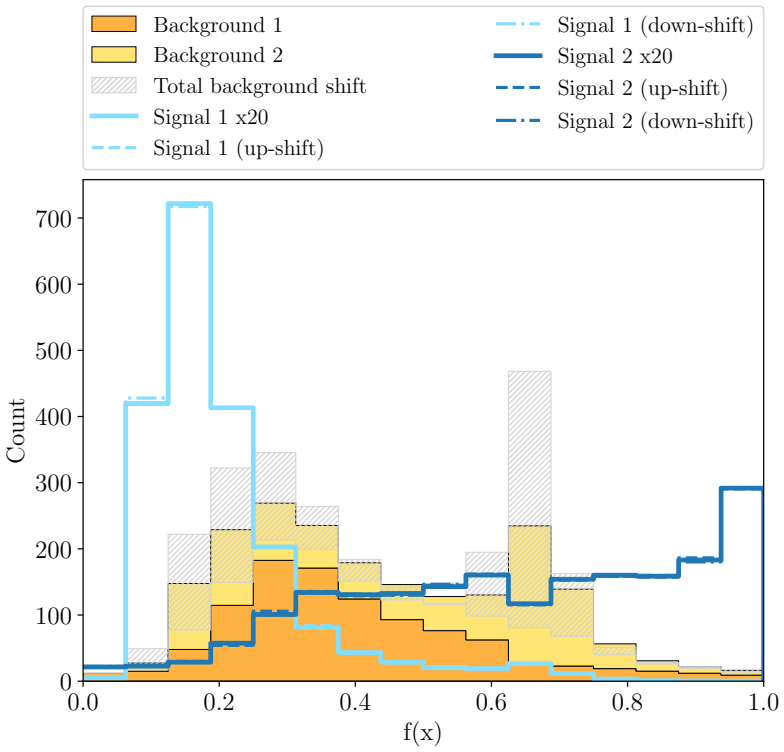


Figure 5.7.: Summarized NN output of the one-class modification of uncertainty-aware training in presence of multiple classes. Displayed are the nominal signal and background processes. The total effect of the systematic variation of Background 2 by ± 1 along x_1 is summed up for all background processes in each bin and shown in gray. Resulting systematic variations on signal processes are shown separately for each signal process.

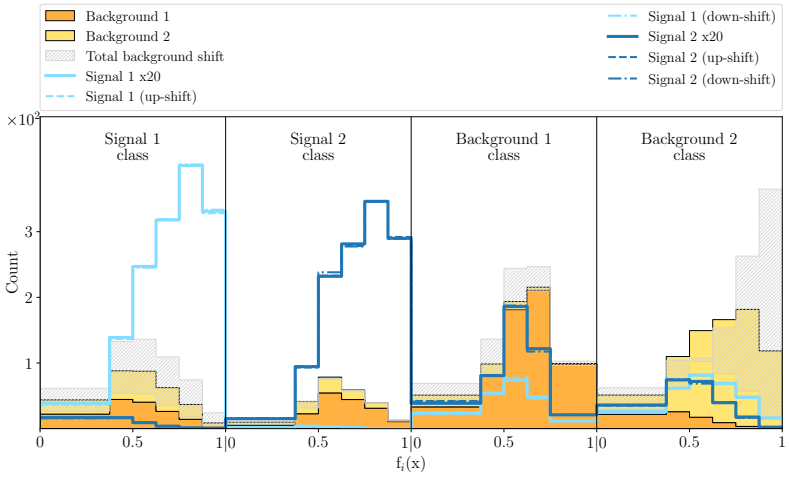
5.2.2. Constrained uncertainty aware training

The problem of empty classes can also be addressed by enforcing the preservation of the classification after the warm-up phase maintaining the multiple output nodes of the NN for the unique assignment of events to selected classes. The realization is built on the Modified differential Method of Multipliers as presented in [69] to solve constrained differential optimization problems. An extension of the present loss is performed by introducing a penalty term of form $\lambda g(\cdot)$ as noted in equation 5.2 where $g(\cdot)$ represents an additional function that is calculated during the training and multiplied by a learnable parameter λ that is set to zero if $g(\cdot) < 0$, evaluated for each epoch. This function is constructed as the difference between the loss function L that is used during the warm-up phase and an arbitrarily selected constant L' as shown in equation 5.3. The goal of this approach is the preservation of the event classification that is obtained at the end of the warm-up phase of the uncertainty-aware training using L. Therefore the constant is set to the value of the warm-up loss at the end of the warm-up phase and is denoted as L'.

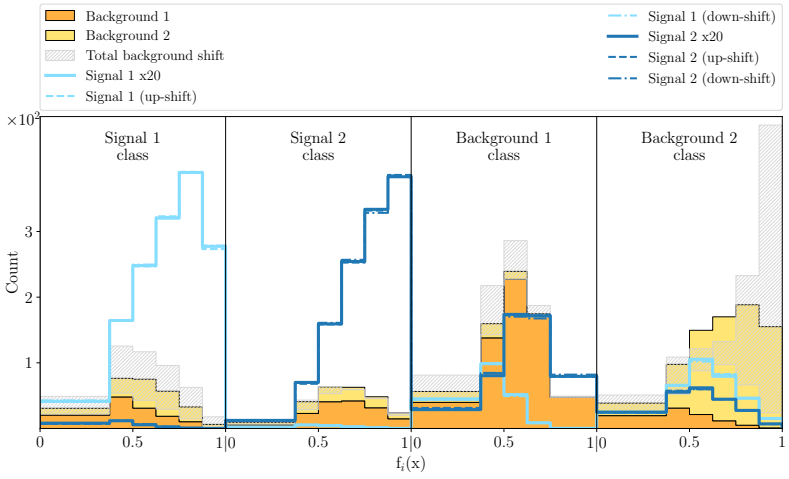
$$\text{Loss} = \sum_i \sigma_{\mu_i} + \lambda g(\cdot)\,, \tag{5.2}$$

$$g(\cdot) = L - L'\,. \tag{5.3}$$

Further, the difference from the unmodified multi-class approach is the change in the warm-up loss and the final activation function of the NN. The latter is changed from Softmax to Sigmoid and the warm-up loss is set to a modified BCE ($L_{\text{BCE}'}$), as defined in equation 5.4, to address multiple classes. The constant value L' is chosen as the value of $L_{\text{BCE}'}$ at the end of the warm-up phase ($L'_{\text{BCE}'}$). Similar to BCE, as defined in equation 3.11, this



(a)



(b)

Figure 5.8.: NN outputs of the constrained-loss modification of uncertainty-aware training in presence of multiple classes on the test data set (a) at the end of the warm-up phase and (b) for the best-performing NN on the validation data set. Shown are the nominal processes, the total background shift, and individual signal variations that are resulting from the same systematic variations and are further described in figure 5.2a.

modification accumulates processes in the corresponding classes as defined by the label $y_i^{(c)}$ in the data set ($y_i^{(c)} = 1$). The prediction given by the NN output can be interpreted similarly as the probability that a given process corresponds to a specific class. All processes that do not correspond to a certain class ($y_i^{(c)} = 0$) are predicted with a low value. Since only the maximum valued node of the NN output of each event enters the $\sum_i \sigma_{\mu_i}$ calculation the values of the remaining classes are discarded and an additional normalization of the NN output in form of an introduced Softmax activation function is an unnecessary constraint.

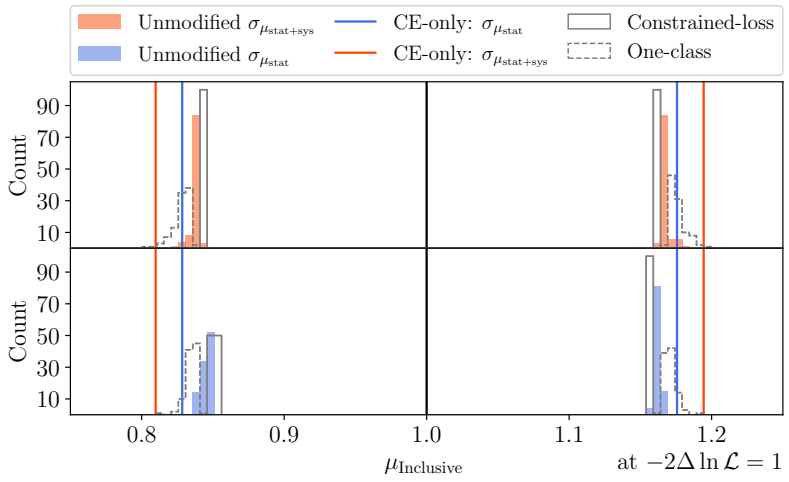
$$L_{\mathrm{BCE}'} = -\sum_{i=1}^{N} w_i \sum_{c=1}^{C_{\mathrm{class}}} \left[y_i^{(c)} \log\left(\mathrm{f}(x_i)^{(c)}\right) + \left(1 - y_i^{(c)}\right) \log\left(1 - \mathrm{f}(x_i)^{(c)}\right) \right] . \tag{5.4}$$

The NN output, evaluated on the test data set from a NN at the end of the warm-up phase and the best-performing NN on a validation data set is exemplarily shown in figures 5.8a and 5.8b respectively and displays a broad event distribution inside the classes. This broader event distribution, especially at the end of the warm-up phase, is the reason for not using the CE in favor of the modified BCE. Training on CE results in a less broad distribution with the most populated bins at the right of each corresponding class histogram.

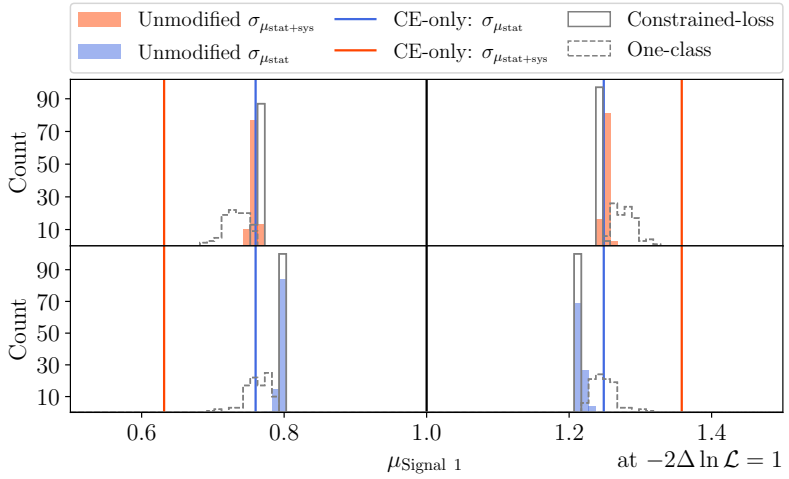
The use of CE during the warm-up phase therefore requires a broadening of the distribution afterwards, as indicated by the steep increase observed in figure 5.6b that is followed by a plateau. Given this observation, the construction of $g(\cdot)$ can also be accomplished with the CE and this known plateau value as the set constant representing the ideal result of the uncertainty-aware training that maintains the class assignment. Unfortunately, this plateau value is apriori not known for a specific task. It might vary depending on the randomly initialized weights, or could not be identified if the movement of events into different classes and fewer bins after the warm-up phase.
The constraint placed upon λ is conditionally based on the value of $g(\cdot)$. In cases where the uncertainty-aware training reduces $\sum_i \sigma_{\mu_i}$ and improves the class assignment of events, compared to the end of the warm-up phase, λ is set to zero. According to this condition, λ affects the loss only in cases where the class assignment gets worse in comparison to the retrieved classification at the end of the warm-up phase forcing a recreation of a similar class assignment. This procedure allows the NN to temporarily perform a redistribution of events, thereby worsening the $L_{\mathrm{BCE}'}$ and performing the optimization in a different phase space after the recreation of a similar class assignment that is also evaluated to $L'_{\mathrm{BCE}'}$. A continuous update of $L'_{\mathrm{BCE}'}$ would result in a continuous constraint, limiting the phase space exploration of the optimizer by restricting the possible event redistributions.

5.3. Comparison of the modifications of the uncertainty-aware training in presence of multiple classes

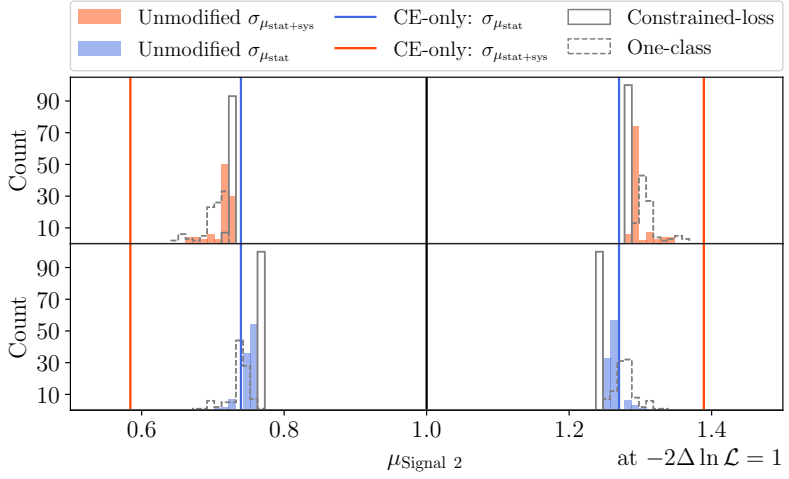
To compare the effectiveness of the two proposed modifications of the uncertainty-aware training to the unmodified version, the results are compared based on ensembles of 100 trainings conducted for each approach by randomly initializing the NN weights using varying seeds. The comparison is performed on the results of the statistical inference summarizing the extracted signal strengths uncertainties similarly to figure 5.3.
The results of the ensembles, combined with the previously created benchmark of the training on the CE are displayed in figure 5.9. The figure is divided into an inclusive signal strength estimation, as shown in figure 5.9a and differential signal strength estimation shown in figure 5.9b for Signal 1 and in figure 5.9c for Signal 2. The benchmark result of the CE training is indicated by the solid vertical lines, which show the uncertainties from the likelihood scan for the case of considering only statistical uncertainties (blue) and both statistical and systematic uncertainties (red). The black line marks the best-fit value, which, as expected, does not differ from one due to the use of an Asimov data set. The resulting uncertainties of the likelihood scans from the 100 trainings are indicated in form of histograms. The lower part of the figures thereby summarizes the results of statistical uncertainties whereas the upper histograms summarize the results of the statistical and systematic uncertainties. The filled histograms show the results of the unmodified multi-class uncertainty-aware training as described in section 5.2, while the unfilled histograms show the results of the two introduced modifications as a comparison. The unfilled histograms with a dashed gray line summarize the one-class approach as discussed in subsection 5.2.1 and the unfilled histograms with a solid gray line the results of the in subsection 5.2.2 discussed constrained-loss approach. All lines and histograms on the left (right) side of the best-fit value corresponds to the lower (upper) value of the signal strength uncertainties. For this comparison, all trainings of the unmodified multi-class uncertainty-aware training, which showed the issue of empty classes have been re-initialized. Cases without empty classes were considered regardless of the degree of misclassification. For the benchmark values, only one training result is shown.
The successful reduction of the overall uncertainty of all three variants of the multi-class uncertainty-aware training can be observed in the upper parts of the figures, where the distributions are not exceeding the overall uncertainty given by the CE training. Another important point is an improvement in cases where only the statistical uncertainty is



(a)



(b)



(c)

Figure 5.9.: Results of the statistical inference comparing the unmodified uncertainty-aware training to the two introduced ansatzes in form of an ensemble test with a sample size of 100. Inclusive signal strength estimations are shown in (a). Differential signal strength estimations are shown in (b) for Signal 1 and in (c) for Signal 2. The vertical black lines indicate the estimated signal strength. Statistical (statistical and systematic) uncertainties from the likelihood scans of the conducted benchmark are shown as blue (red) vertical lines. Filled histograms summarize the result of the unmodified uncertainty-aware training in a similar way. Unfilled histograms with a solid (dashed) line show the results for the constrained-loss (one-class) approach. Further information on the figure is given in the text.

considered during the statistical inference. This can be seen independent of the applied training where regions of the distributions in the lower parts of the figures exceed the set benchmark value towards the estimated value, indicating a lower statistical uncertainty. This can exemplarily be described by the comparison of the constrained-loss approach or the unmodified uncertainty-aware training with the CE training.
All these three trainings utilize multiple classes for event separation. During the CE training, an attempt is made to achieve the best possible assignment of events to the appropriate classes. A misclassification to some extent is inevitable, especially in particle physics problems where events with the same topology arise from different processes. Optimization on CE corrects the occurring misclassification across all processes performing a trade-off where the purity of e.g. a background class is improved at the expense of increasing the contamination by misclassified events in other classes including the signal classes and vice versa. This is a common occurrence since the CE loss does not differentiate between the individual classes that have been defined for the training and the event weights only introduce a correction upon the observed number of process events.
The goal of minimizing $\sum_i \sigma_{\mu_i}$ on the other hand is achieved by enriching the signal classes with signal events and is conducted during uncertainty-aware training independent of the number of signal processes. A trade-off, where the purity of a background class is increased at the cost of an increased contamination of or by signal events can ideally only occur for one scenario: The reduction of the systematic variation outweighs the increase in statistical uncertainty. Since the warm-up phase reduces the statistical uncertainty this trade-off leads to a decrease in the overall uncertainty. The purity of the background classes is no longer a relevant factor, as they contribute little to the calculation of $\sum_i \sigma_{\mu_i}$, due to the significantly lower number of signal events in these classes compared to the signal classes with enriched signal events.
Overall, the one-class approach shows worse results when compared to the other two approaches. Individual trainings even show an inclusive overall uncertainty that is not minimized despite the minimization of the uncertainties of the differential signal strengths. This issue can be attributed in the first place to the increased instability in training that becomes more dependent on the initialized NN weights, which is reflected in the larger spread of the distributions of the one-class approach in comparison to the other approaches with the most probable value being consistently worse.
On the other hand, the constrained-loss approach significantly reduces the spread of the distributions and shows a better most probable value than the other approaches. This result leaves the one-class ansatz with the only advantage of having a performed training purely on the analysis objective whereas the constrained-loss approach depends on the introduction of two additional hyperparameters, namely the function $g(\cdot)$ and the constant value $L'_{\mathrm{BCE}'}$.
The second issue of the one-class approach is the limited space when considering multiple signal processes in addition to the considered background processes in a single histogram. With an increasing number of signal processes the challenge to achieve a good signal separation between the individual signal processes and the background processes leads to an overall worse result.
The introduction of additional histograms is more advantageous as it introduces an additional dimension for the event separation. An increase in the number of used bins for the one-class approach does not resolve this issue, as shown in appendix B, since the movement of the events from the NN output remains grouped in the event overlapping regions where signal and background processes cannot be distinguished. Therefore, the one-class approach might be particularly suitable for tasks that introduce only one signal process in addition so several background processes whereas the constrained approach is more suitable for tasks comprising multiple signal processes in addition to the present background processes.

6. Application on reduced standard model H→ττ data set

This chapter presents the application of the uncertainty-aware training on a more complex problem, namely on the reduced CMS data that is used in [4]. First, the modified setup resulting from the changed data is described, followed by an application of binary classification with an additional outline of the necessity of including the most important systematic uncertainties during the training. The chapter concludes with the application of the uncertainty-aware training on this data set considering multi-class classification by using the modified versions of uncertainty-aware training that were introduced in chapter 5.

6.1. Setup, analysis procedure, and CMS data set

The application of the uncertainty-aware training on a realistic example is performed utilizing the data set that is used for the CMS Standard Model H $\rightarrow \tau\tau$ analysis [4]. To set the focus on the demonstration three reductions to the analysis and the considered data set are applied. From the three years of the second data-taking period of LHC and four final states, only the data of 2017 and the eτ final state are considered with the aim of data reduction to reduce the training time. The selection of the eτ final state described in chapter 2 is sufficient, as an analogous approach for the training of NN with the same architecture is applied for the other final states and the outputs of all NNs are combined using the nominal and shifted test data sets for the statistical inference. The choice on eτ final state over μτ is done from the consideration of resulting higher uncertainty sources from the combination of electrons and hadronic taus in comparison to muons and the introduction of electron uncertainty sources in comparison to the ττ final state. Further, as described in section 2.5, only systematic uncertainties that are provided in form of weight corrections are used in this chapter. The 86 thereby considered systematic uncertainties are summarized in appendix C in descending order by their impact on the estimated signal strength from the statistical inference performed on the BCE training for binary classification that will be discussed in detail in the following section. This restriction also contributes to the reduction of training time, as these uncertainties do not require an additional propagation of shifted data sets to quantify the effects of the systematic variations on the NN output. Instead, a direct application of the weights to the NN output is performed.
The data set used for this application is divided into two halves, referred to as the first and second fold, following the proposed approach from the analysis to make use of the full data set for the NN training and the statistical inference. The following split of the resulting data set from each fold that is used for the training (validation) step is changed from previously used 75 % (25 %), respectively, to 50 % with the goal to improve the validation result due to the applied full-batch training. Testing and thus the extraction of the NN outputs for the statistical inference is performed on the correspondingly other folds and combined afterwards, allowing for the benefit of not losing data due to the performed

training and thereby improving the statistical uncertainty of the final measurement. The NN architectures are identical for both folds, as described in section 4.1 with the number of output nodes varying according to the specific task being addressed. The number of input nodes is set to 15, based on the input variables proposed by the analysis, with an importance-based selection determined by the impact on the NN output obtained from the method described in [3]. The variables are scaled by their standard deviation and shifted by their mean before the training.
These variables include the visible di-τ mass m_{vis} and the fully reconstructed di-τ mass $m_{\text{sv}}^{\text{Puppi}}$ using the pile-up per particle (PUPPI) algorithm [70]. The variable m_{vis} mainly discriminates signal events against $Z \to \tau\tau$ background by considering only the visible decay products, whereas the latter variable includes the undetectable neutrinos in the reconstruction. Further di-τ properties are considered in addition to the reconstructed masses such as the transverse momentum of the first (second) tau lepton $p_T(\tau_1)$ ($p_T(\tau_2)$) as well as their combined visible transverse momentum p_T^{vis} and their angular distance $\Delta R(\tau_1, \tau_2)$. For the characterization of the different jet topologies additional kinematic quantities are added including the transverse momentum of the leading (subleading) jet $p_T(j_1)$ ($p_T(j_2)$), their difference in pseudorapidity $\Delta\eta_{jj}$, as well as the combined mass of those jets m_{jj} and the transverse momentum $p_T(jj)$ resulting from this combination. The number of jets N_{Jet} alongside the number of jets that are classified to be originating from a bottom quark decay N_{Btag} are added mainly to identify events from the $t\bar{t}$ background. Since the number of jets can increase due to the presence of pile-up resulting from additional pp collisions or a misidentification of additional jets the combination of those two variables can be used for the identification of a signal event from the selected eτ final state as it is expected to observe at least one jet that is not originating from a bottom quark decay. To synchronize the NN input across final states two additional quantities using the matrix element likelihood approach (MELA), `MELA_q2v1` and `MELA_q2v2`, as described in [71, 72], are added to the input variables. These discriminating variables estimate the momentum transfer for the first and second exchanged vector bosons, differentiating between the vector boson fusion originating signal process and Drell-Yan process topologies by calculating their likelihood ratio, thus reflecting the probability of the observation of their occurrence. For the application of multi-class uncertainty-aware training, the separation of signal and background processes is performed as proposed in the original analysis. Background events that contain genuine $\tau\tau$ events, which are obtained through the τ-embedding method, as described in section 2.4, are assigned to the `emb` class analogous to the `ff` class, which consists of events that contain jets that are misidentified as tau leptons and are estimated using the F_F method. All remaining classes contain events that are purely derived from simulation. For the signal processes the STXS stage 0 binning [39] is applied, which splits the signal process into the qqh and ggh production modes which are assigned to two unique classes. The estimated events resulting from the $t\bar{t}$ background, are assigned to the `tt` class, while the `zll` class contains events from the $Z \to \ell\ell$ process. Events resulting from the electro-weak production of Z bosons and di-boson production, as mentioned in section 2.4, have a minor contribution to the overall background due to the choice of the final state and are therefore subsumed into the additional `misc` class preventing the misidentification of those events with the signal processes.

6.2. Application of uncertainty-aware training for binary classification

For an application of binary classification, all signal and background processes that have been discussed in the previous section are combined into one single signal and background process correspondingly. This changes the influence of the normalization uncertainties that are described in section 2.5. Due to the aggregation of the background and signal processes systematic variations introduced by normalization uncertainties affect only a part

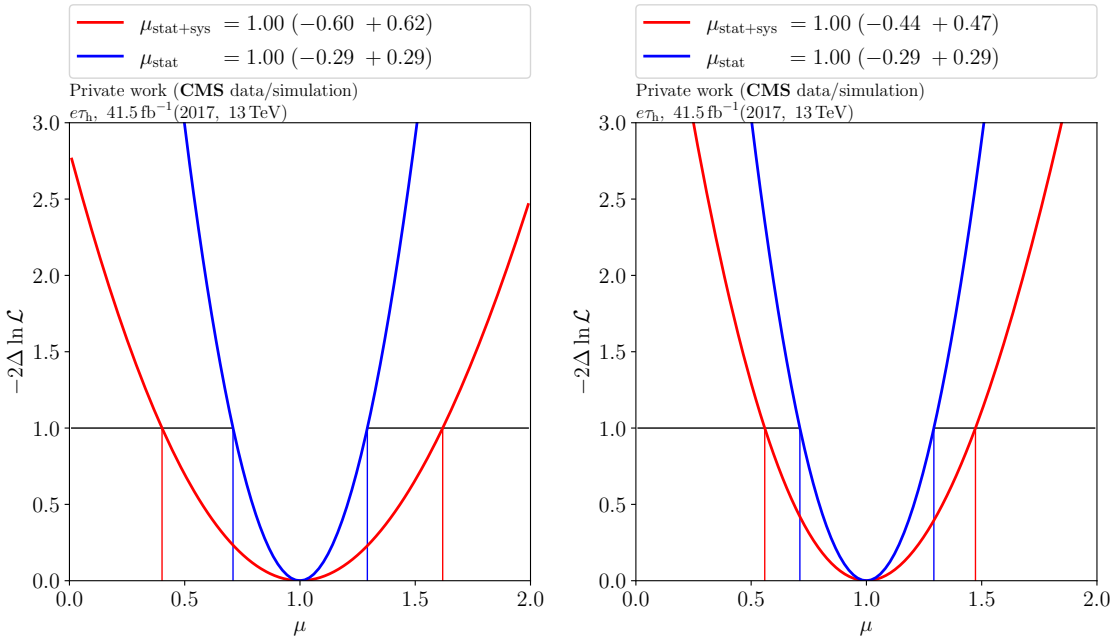


Figure 6.1.: Likelihood scans of the estimation of the signal strength (a) for the BCE training and (b) the uncertainty-aware training, considering all 86 uncertainties. The estimation of only statistical uncertainty is shown in blue, whereas the estimation of statistical and systematic uncertainties is shown in red.

of the background or signal process or their combination, thus introducing a shape-altering effect and are addressed correspondingly in the statistical inference. The benchmark for the application of binary classification is conducted using the BCE loss, followed by the statistical inference applying all 86 systematic uncertainties and using the NN output of the simulated data. This yields an estimation of the signal strength, as shown in figure 6.1a of

$$\mu_{\text{stat+sys}} = 1.00^{+0.62}_{-0.60}\,.$$

By performing uncertainty-aware training and considering all systematic uncertainties during the training step, an improvement in the signal strength uncertainty to

$$\mu_{\text{stat+sys}} = 1.00^{+0.47}_{-0.44}\,,$$

can be achieved, as can be depicted from the scan of the likelihood shown in figure 6.1b.

In addition to the estimation of the signal strength and its uncertainty through the statistical inference, the effects of the introduced systematic uncertainties can be examined to estimate their impact on the retrieved signal strength estimation. This impact estimation is acquired by the measurement of the strength of the dependence of the resulting shift on the estimated signal strength upon a variation of a nuisance parameter by $\pm 1\sigma$. This can be performed for each introduced systematic uncertainty, represented by their corresponding nuisance parameter and thus creating a ranking of the impact on the estimated signal strength. The 20 uncertainties with the highest impact on the signal strength, according to the benchmark result, are shown in figure 6.2 as gray lines. The change in the impact after the application of uncertainty-aware training is depicted as colored bars corresponding to the $\pm 1\sigma$ shift of the individual nuisance parameters. Notable is the reduction of the impact on the estimated signal strength from the 10 leading uncertainties. The observed trend, of performed reduction of uncertainties with a higher impact at the cost of an impact increase of less important uncertainties, can be further seen in appendix C for

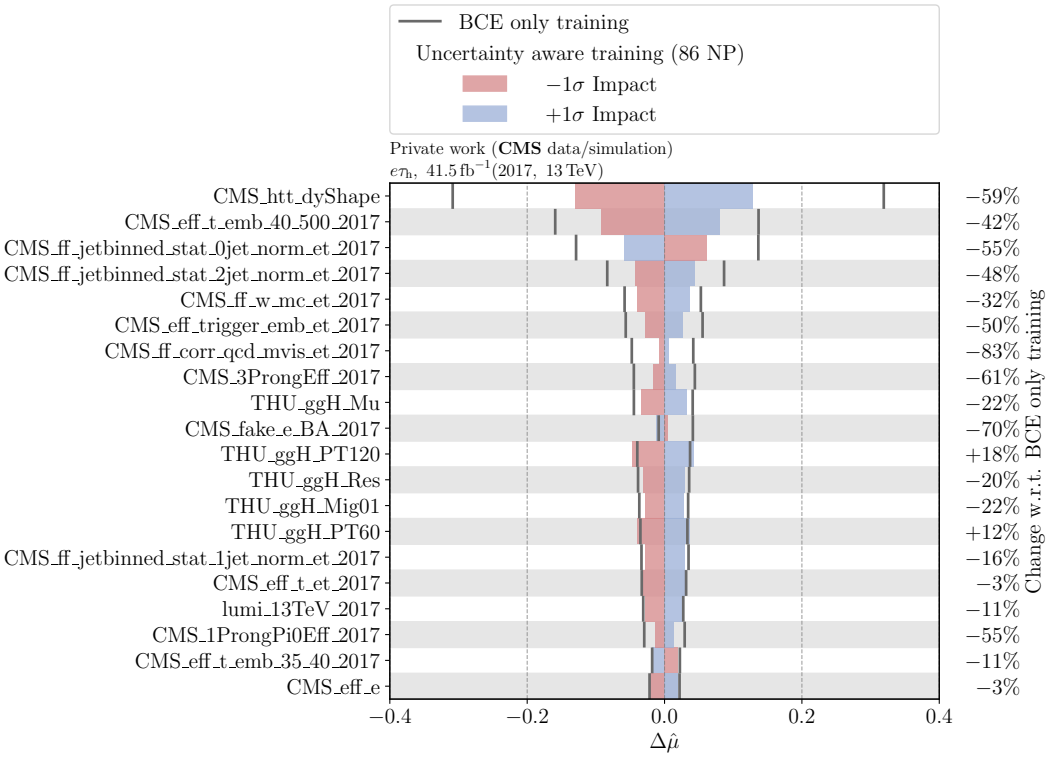


Figure 6.2.: Summary of 20 uncertainties with the highest impact on the estimated signal strength from the training on the BCE loss (gray) in comparison to the applied uncertainty-aware training where the colored bars indicate the impact of a performed $\pm 1\sigma$ shift of the nuisance parameter that corresponds to the systematic uncertainty. The full comparison is shown in appendix D.

the remaining uncertainties. The first two examples of an increase are the differential theory uncertainties on the ggh production mode in the transverse momentum bins of $p_T = 120\,\text{GeV}/c$ and $p_T = 60\,\text{GeV}/c$. Their overall lower importance compared to the results of the original analysis is due to the restriction on a subset of the data, as these uncertainties arise throughout all final states, while most of the uncertainties considered in this study only affect this specifically chosen subset of the data.

The systematic uncertainty with the highest impact on the estimated signal strength in both the benchmark analysis and the analysis based on the uncertainty-aware training is the shape-altering uncertainty from the Drell-Yan (DY) process. Its effects on the NN output are shown in figures 6.3a and 6.3b for the BCE and uncertainty-aware training, where the upper plots correspond to the nominal NN outputs, supplemented by the relative up and down variations of the signal (blue ratio) and background processes (orange ratio) for DY as the selected systematic variation with respect to the corresponding nominal process in each bin. As this uncertainty only applies to events originating from the $\text{Z} \to \ell\ell$ process, the signal process remains unaffected. For the BCE training, the largest effect of this systematic variation is observed for the background process in the bins with enriched signal, thereby explaining the observed large impact on the estimated signal strength. The effect of the same systematic variation after the application of uncertainty-aware training changes to a flatter distribution with comparable high relative amplitude, as shown in figure 6.3b. This reduction in the shape-altering effect is achieved through the redistribution of affected events across more bins, creating a systematic variation similar to one that can be introduced by a normalization uncertainty. The deviation of the first bin, which exhibits a lower relative variation in comparison to the remaining bins can be attributed to the

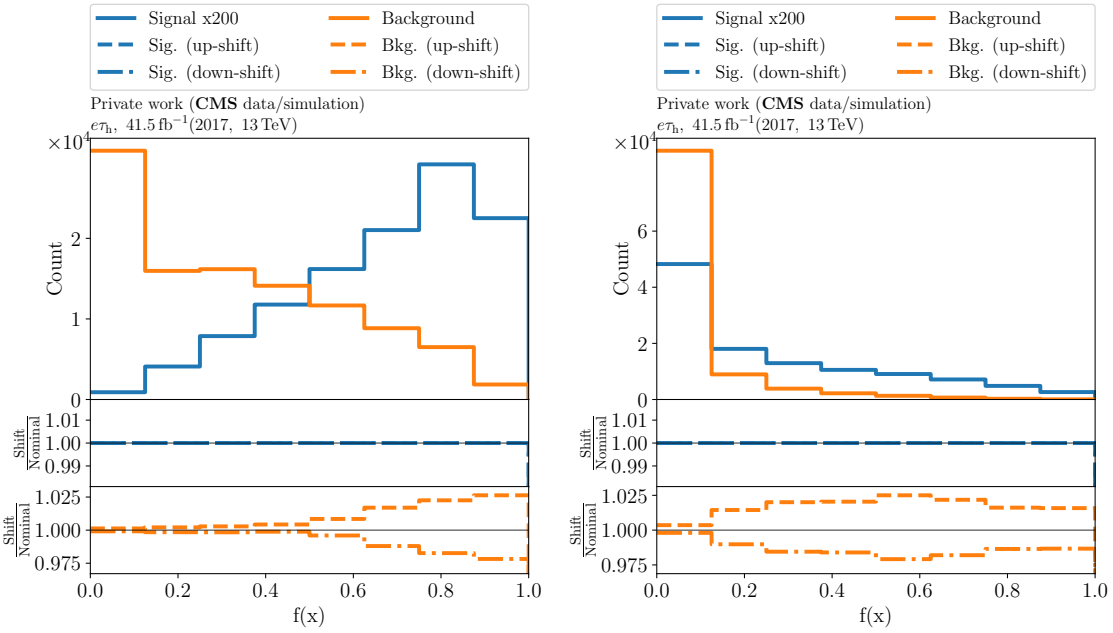


Figure 6.3.: NN output of (a) the BCE training and (b) uncertainty-aware training. Shown are the nominal signal and background processes (top rows) supplemented by the relative up and down variation of the signal (blue ratio) and background processes (orange ratio) with respect to the corresponding nominal processes in each bin. Shown is the systematic variation resulting from DY uncertainty affecting the $Z \to \ell\ell$ process in the combined background.

large number of overall background events contained in this bin, thus resulting in a lower ratio despite having a higher absolute difference between the up and downward variation. Thus changed effects of the systematic variation are better constrained by the performed fit, leading to a reduced impact in comparison to the result from the CE training.
To indicate the effects of both training methods on the NN output function, the change of the importance of the input variables is investigated using the TCA method as described in [68], considering Taylor Coefficients (TC) up to the second order. The computation of the mean of the absolute values of the TCs is performed by combining the TCs retrieved from the NNs of both folds. Figure 6.4 shows the resulting importance ranking displaying the 20 most important input variables and their correlations. The ordering is based on the results from the BCE training that are shown in gray with the results of the uncertainty-aware training shown in black for comparison.
After the application of the uncertainty-aware training, an overall decrease in the importance of variables is observed that previously showed high importance during the BCE training. This reduction is similar to the observation seen in section 4.4 and is e.g. displayed by the first order TC of the m_{vis} input variable. In the case of the BCE training, this variable shows the second highest importance, thus indicating high separation power upon the distinction between signal and background process. However, as m_{vis} does not incorporate the information of present neutrinos in comparison to $m_{\text{sv}}^{\text{Puppi}}$ a misclassification due to a wrong mass reconstruction is more likely to occur. In the case of wrong-reconstructed masses, the resulting discrimination using m_{vis} can lead to an undesired increased aggregation of misidentified background events in signal-enriched regions. Thus the introduced systematic variations in signal-enriched bins, with the DY uncertainty as one of the prominent examples, are leading to an increase in the uncertainty on the estimated signal strength. The conducted application of uncertainty-aware training indicates the consideration of the

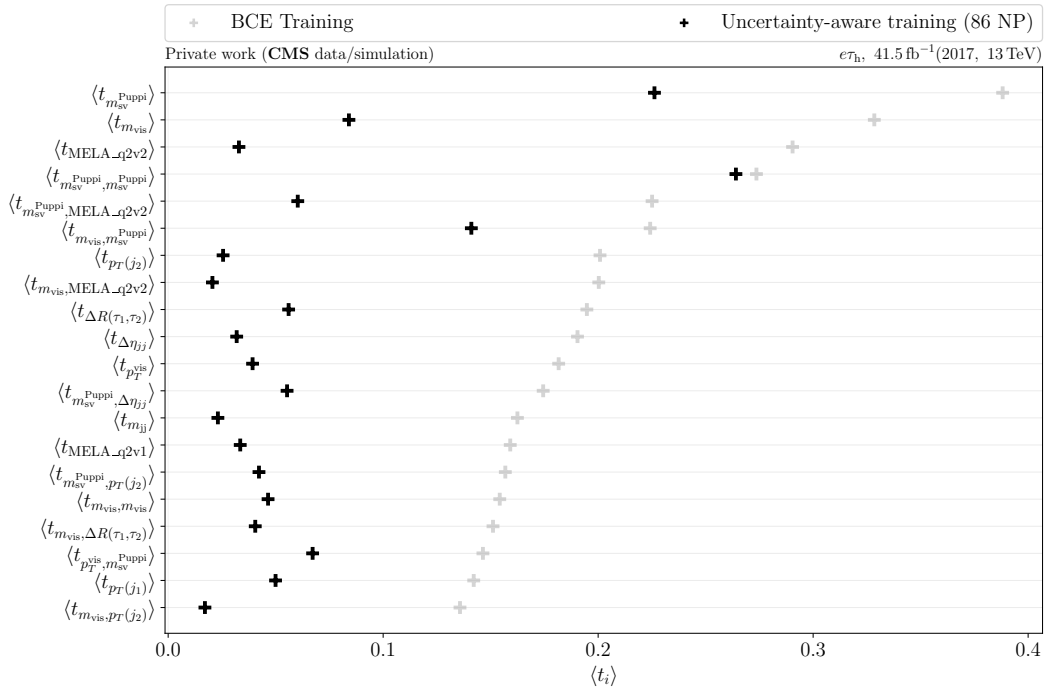
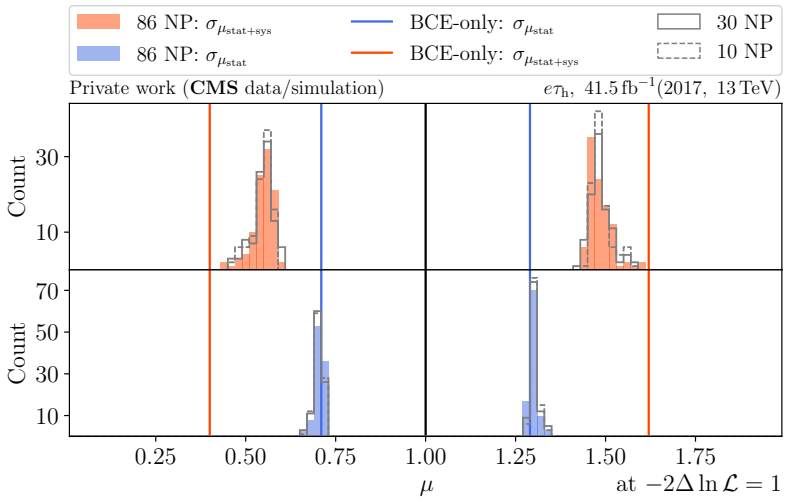
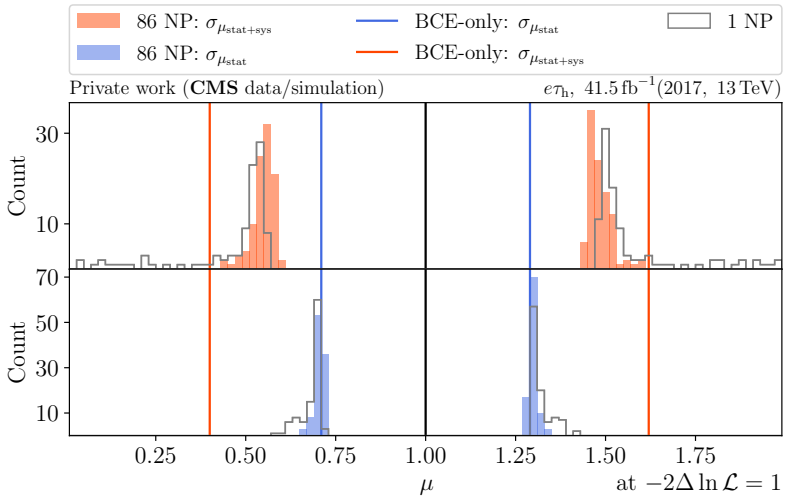


Figure 6.4.: The change of the 20 highest Taylor coefficients (TC) from the training on the BCE loss (gray) in comparison to the uncertainty-aware training (black) is shown indicating the most important input variables or the correlation between input variables for the NN decision. The shown Taylor coefficients are summarized by performing a mean of the absolute values of the TC from each event as proposed in [68]. The remaining TC can be depicted in appendix E.

present uncertainties by the NN leading to a reduction of $\langle t_{m_{\text{vis}}} \rangle$ and the decrease of the impact of the DY uncertainty in figure 6.2. A summary of all following first and second-order TC of this comparison can be further depicted in appendix E. Similar to the conducted ensemble test, which is discussed in section 5.3 an additional study regarding the stability of the training is conducted due to the increased complexity resulting from the usage of a higher number of systematic variations. The results of the statistical inference from the performed ensemble tests are summarized similarly to figure 5.9 in figure 6.5a. The filled histograms show the results of 100 performed statistical inferences of uncertainty-aware trainings, which utilized all 86 systematic variations. The shown distribution indicates the ability of the uncertainty-aware training to minimize the effects of those systematic variations in most of the training runs. Only a few exceptions are observed where no improvement compared to the benchmark analysis is seen, as indicated by the end of the tails of the upper distributions. This can partly be attributed to the discretization of the NN output discussed in section 4.2 resulting in the inability to find an optimal bin combination. An unfavorable initialization of the NN weights that compromise a successful optimization is another possibility.
An additional examination is performed as to whether a subset of the considered systematic variations used during the training might be sufficient for a comparable minimization of the overall uncertainty compared to an evaluation using the entire set of provided uncertainties. To address this, the 30 and 10 systematic variations with the largest impact, as identified for the benchmark analysis, are selected and used for 100 training runs each. The performed statistical inference in both cases accounts for all 86 uncertainties. The results are shown in figure 6.5a by the unfilled histograms with solid (dashed) lines indicating a training on 30 (10) systematic variations. The observation shows that the consideration of only a subset



(a)



(b)

Figure 6.5.: Summarized results of the application of statistical inference upon 100 performed uncertainty-aware trainings similar to figure 5.9. The filled histograms show the results of trainings that uses all 86 uncertainties. The trainings using 30 (10) most important systematic uncertainties are shown as unfilled histograms with a solid (dashed) line in (a). The results of the utilization of only the most important uncertainty is shown as an unfilled histogram in (b). The vertical blue (red) lines correspond to the statistical (statistical and systematic) uncertainties that are retrieved from the BCE training. Further information is given in the text.

of the systematic variations, in particular, the uncertainties with the highest impact is sufficient for a successful minimization of the overall uncertainty. This results in a decrease of the task complexity and the training time of an epoch by approximately a factor of 5.6 (7.5) comparing 86 to 30 (10) systematic variations[1]. This reduction originates mainly from the reduction of complexity of the calculation of the inverse of the Hessian matrix for each epoch[2]. However, problems arise when too few systematic variations are included during the uncertainty-aware training, making the application of a specifically targeted

[1]The calculated factors refer to the required time per epoch, whereby the average value of 1000 epochs from a single training is used. The mean time that is required for an epoch for this specific data set training with 86, 30, and 10 uncertainties are (5.77 ± 0.07) s, (1.03 ± 0.02) s and (0.77 ± 0.01) s.

[2]The approximate fraction of the total time for an epoch that is used for this calculation in the case of 86 (30, 10) systematic variations is 85 % (74 %, 63 %).

systematic variation reduction less applicable. An extreme example is shown in figure 6.5b, where only the systematic variation caused by the DY uncertainty is considered during the training. The summarized results are shown by the unfilled histogram, displaying prominent tails. This indicates a worse result than the training on the BCE loss showing the lack of flexibility to perform the described trade-off between the increase of impact of less important systematic variations and a decrease of systematic variations with a greater impact on the overall uncertainty. This can lead to an increase in the effect of systematic variations raising the impact of the next most important uncertainties that are not considered during the training, eventually worsening the overall uncertainty. Therefore, in the case of an application, where the number of considered systematic variations differs between the training and the statistical inference, the uncertainties that have the highest impact on the estimated signal strength must be identified beforehand e.g. given a performed benchmark.

6.3. Application of uncertainty-aware training in presence of multiple classes

In the following, an extension of the uncertainty-aware training to multiple classes that are discussed in section 6.1 is performed. The usage of seven classes, without the consideration of the additionally introduced complexity by systematic variations, fulfills the condition of a complex problem as mentioned in section 5.2. An application of the unmodified approach of the uncertainty-aware training is therefore impractical and will not be considered in the comparison due to the issue of potentially emerging empty classes during the training. With this regard, only the two introduced modifications to the uncertainty-aware training are compared against each other, and against the performed benchmark analysis.
The benchmark analysis is conducted with a lower patience for the training than the uncertainty-aware training, similar to the discussion in section 5.2. However, due to the complexity of the current task, the patience is set to 100 epochs. To ensure better comparability between the benchmark and the uncertainty-aware trainings, the NN architecture and binning described earlier are used for this task instead of the optimized NN architecture and binning of the original analysis.
The resulting NN output of the CE training is shown in figure 6.6a. The upper part shows the nominal background and signal processes on a log scale and indicates a high level of contamination of ggh events in the qqh class, which could only be marginally improved with a larger NN architecture as this feature is also observed in the original analysis. All low-populated bins are merged by the procedure described in section 5.2. The lower part of the figure shows the ratio between the summed yields of all background processes and the corresponding signal process divided by the summed yields of all background processes for each bin, indicating signal-enriched regions. For this purpose, the scaling of signal processes as indicated in the upper part of the figure is not applied. The systematic variation of all background processes that are affected by the selected uncertainty is shown similarly in gray in the ratio plot. The chosen systematic variation represents the effects of the DY uncertainty as its impact on the estimation of inclusive signal strength remains high as can be depicted in appendix F.
The results of the statistical inference are shown in figures 6.6b, 6.6c and 6.6d . They are divided into an inclusive measurement and the differential measurements of individual Higgs boson production modes. As indicated by the NN output, the overall uncertainty on qqh production is dominated by statistical uncertainties. This is supported by the inability of the NN to distinguish between the qqh and ggh processes, which results in

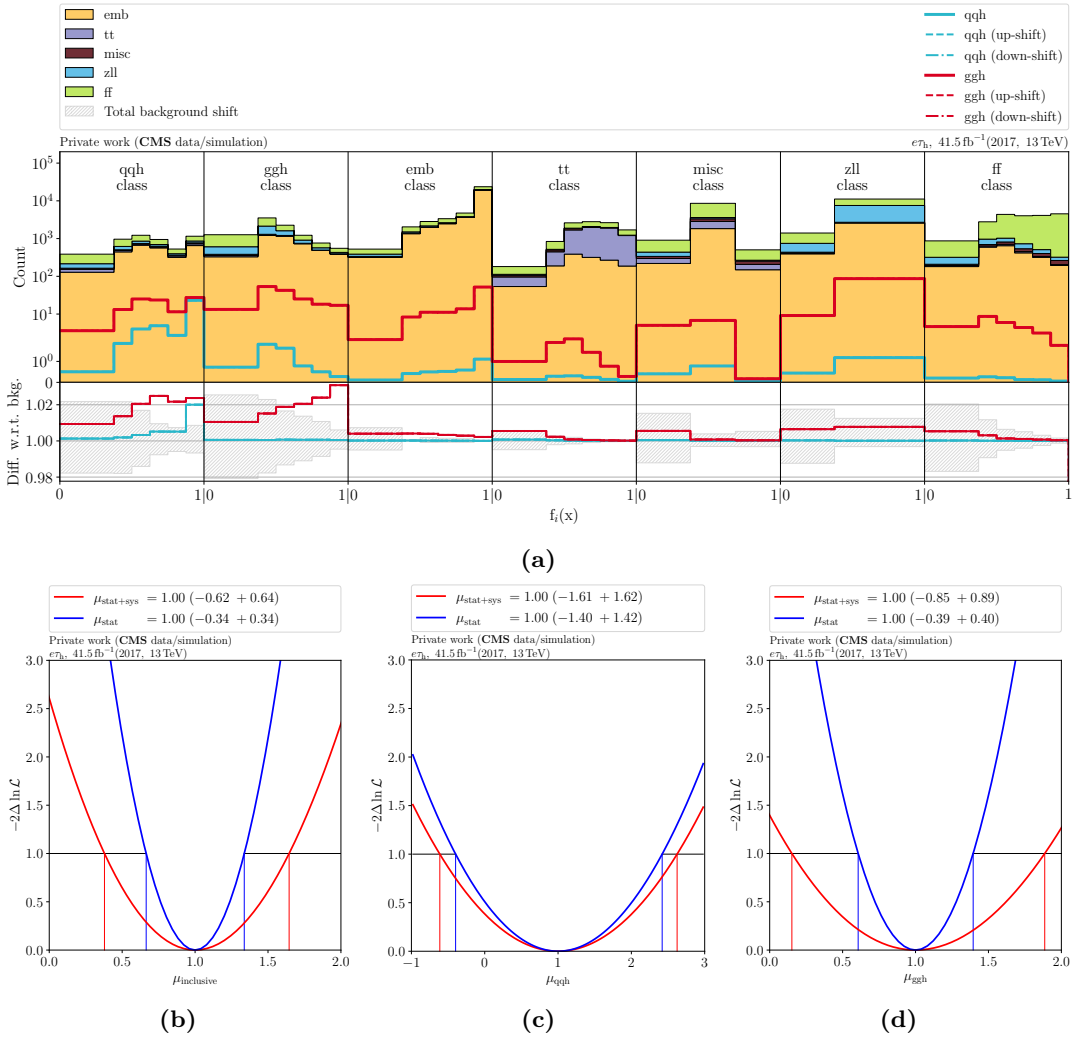
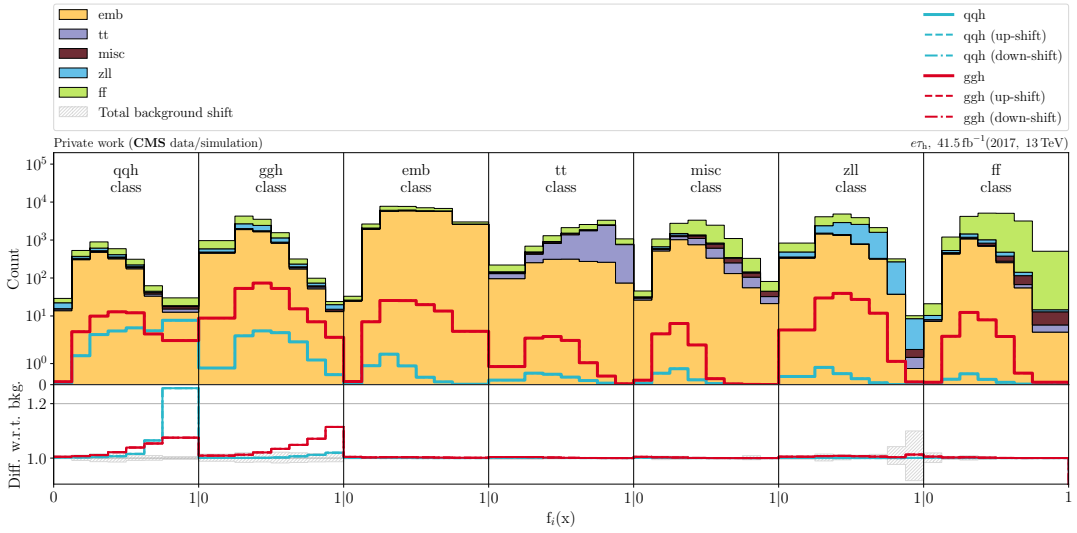


Figure 6.6.: The NN output for the CE training is shown in (a), displaying the nominal processes on a log scale in the upper half. Low-populated or empty bins are merged with the neighboring bins as exemplarily shown in the `zll` or `misc` classes. The lower part shows the ratio between the summed yields of all background processes and the corresponding signal process, divided by the summed yields of all background processes in each bin, indicating signal-enriched regions. The systematic variation is shown similarly in gray and displays the effect of the DY uncertainty. Used classes and processes are described in the text. The inclusive (b) and differential signal strength estimation of the (c) qqh and (d) ggh process of preformed statistical inference of CE training are shown, considering all 86 uncertainties. Results that only consider statistical uncertainties are indicated by the blue scans and the consideration of statistical and systematic uncertainties is shown in red.

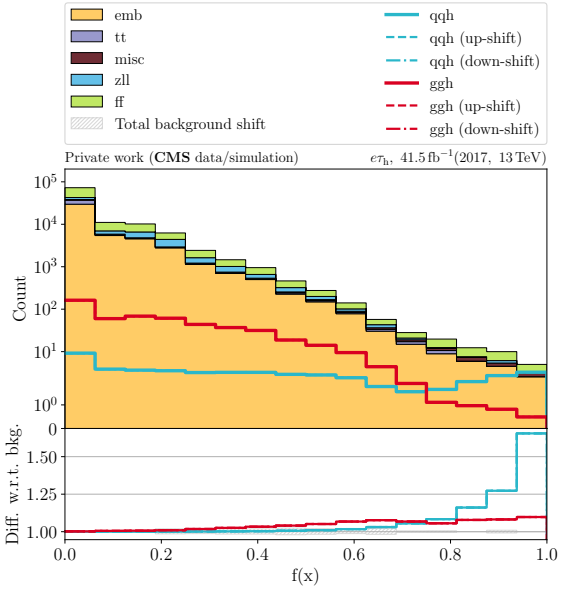
only one highly enriched bin with qqh events. The resulting values of the signal strength uncertainties that are used as benchmark for the uncertainty-aware training are:

$$\mu_{\text{inclusive}} = 1.00^{+0.64}_{-0.62}\,,$$
$$\mu_{\text{qqh}} = 1.00^{+1.62}_{-1.61}\,,$$
$$\mu_{\text{ggh}} = 1.00^{+0.89}_{-0.85}\,.$$

For the application of the uncertainty-aware training using the constrained-loss modification an adaption to the described procedure in chapter 5 is made. An increase in the learning rate from 0.001 to 0.01 and a reduction of the warm-up phase from 300 epochs to 150 epochs are applied. The usage of the lower learning rate and a warm-up phase of 300



(a)

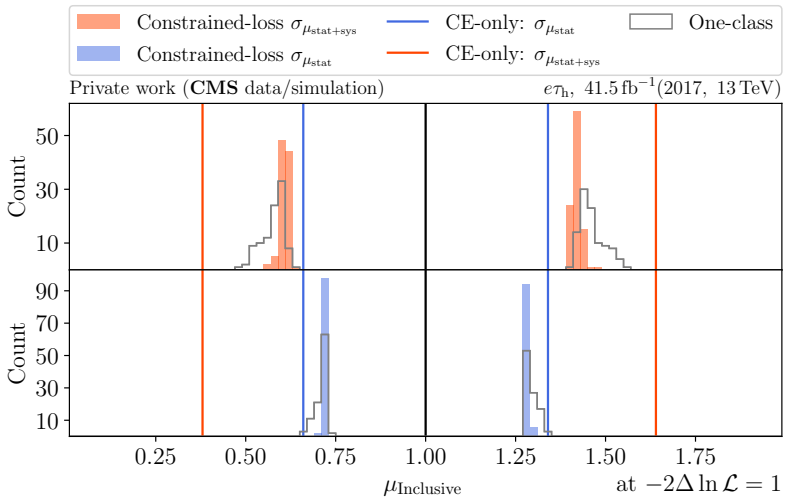


(b)

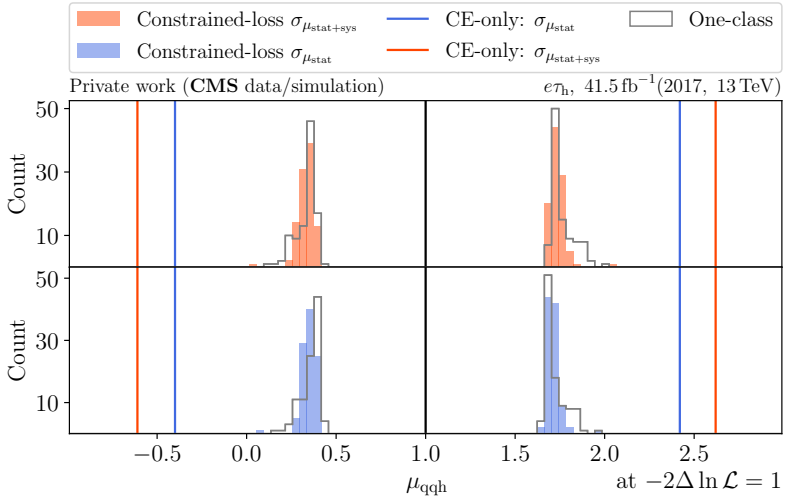
Figure 6.7.: The NN output of the (a) constrained-loss approach and for the (b) one class approach of the uncertainty aware training, shown similarly to figure 6.6a.

epochs results in a worse classification of events into the corresponding classes and requires a longer warm-up phase that is mediated by the increase of the learning rate to 0.01. with this increase, an optimal distribution after around 200 epochs is acquired leading to a plateau in the loss evolution. To ensure the possibility of λ being able to be set to zero or become sufficiently small after the warm-up phase, the duration of the warm-up phase is reduced to 150 epochs in order to allow for enough flexibility in the event redistribution before the penalty comes into action. On the other hand, the training procedure of the one-class approach remains unchanged with respect to the study described in section 5.2.1. Two exemplary NN outputs of those modifications are illustrated in figure 6.7 and indicate an improved signal separation in comparison to the CE training. Both approaches leads to bins with enriched qqh events, thereby significantly reducing the statistical uncertainty on the qqh process, which is inherently present in any trainings due to the imbalance of the signal yields. The nominal yield of the qqh process exhibits only around 51 events, compared to 516 events from the ggh process.

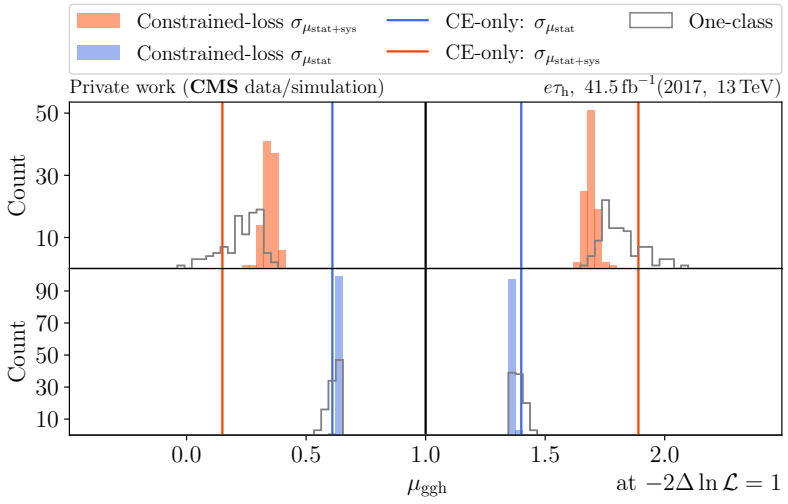
The results of the statistical inference are shown in figure 6.8 in form of conducted ensemble test with a sample size of 100, similar to the discussion in section 5.3. The filled histograms summarize the result given by the constrained approach, and the unfilled histograms are the results of the one-class approach. The benchmark values, indicating the statistical and systematic uncertainties of the conducted CE training, correspond to the values given in figure 6.6. Since both approaches use the sum of the signal strength uncertainties as the loss, a successful minimization of the inclusive signal strength can generally be achieved, as indicated in figure 6.8a. For the given example the one-class approach primarily minimizes the overall uncertainty by minimizing the uncertainty on qqh, separating the qqh events to the right edge of the histogram, while the constrained-loss approach mainly achieves this by sorting out non-qqh events from the `qqh` class. The additional class histograms of the constrained-loss approach show more benefit for the uncertainty minimization of the ggh Higgs production mode, as both methods exhibit significant confusion between the ggh events and the background events. As indicated by figure 6.8c, the minimization of statistical and systematic uncertainty of the differential ggh signal strength does not always improve with respect to the benchmark analysis in case of the one-class approach, as shown by the distributions lying partly outside of the benchmark boundary of the CE training. A similar behavior is observed in the study in section 5.3, with the main difference of the large imbalance between the different signal processes in terms of their expected yields. Therefore, the minimization of the uncertainty of the qqh signal process is favored at the expense of an increase in the uncertainty of the ggh process, under the condition of minimizing the sum of the signal strength uncertainties of both processes. In this regard, the constrained-loss modification leads to a better result.
An extension to a more granular binning of the Higgs boson production modes that are introduced in the STXS stage 1 binning [24] would lead to 14 individual signal classes that have to be considered during the training and statistical inference in addition to the five present background classes. With the results shown in this chapter and in light of the application on a task that comprises multiple classes, the constrained-loss approach shows an overall more promising performance by consistently minimizing the signal strengths of the considered signal classes individually, irrespective of observed imbalances in the expected signal yields. Further, the constrained-loss approach provides more deterministic results as indicated by the lower spread upon the training repetition in comparison to the one-class approach, making it less error-prone upon different weight initializations.



(a)



(b)



(c)

Figure 6.8.: Summary of the statistical inference results from 100 trainings comparing the one-class and constrained-loss approach of the uncertainty-aware training, utilizing 86 systematic uncertainties. The results in the case of an estimation of the inclusive signal strength are shown in (a) whereas the differential signal strength estimations are depicted in (b)for qqh and in (c) for ggh process. The structure is similar to figure 5.9 and is in detail explained in section 5.3.

7. Summary and outlook

The application of several machine learning (ML) techniques in high energy physics (HEP) has proven to be valuable in many areas of conducted analyses. The contribution to object identification and reconstruction in the detector, process classification or the extraction of ML-derived variables, that are used for statistical inference has led to a steady improvement of the measurement of the parameter of interests (POI) and its confidence interval. The differential measurement of the Higgs production modes that was performed within the Standard Model $H\to\tau\tau$ analysis [4] with provided data from the Compact Muon Solenoid (CMS) experiment utilized several of those ML techniques and improved the analysis results in comparison to the traditional analysis strategy [73].
In chapter 3 an outline of the classic utilization of neural networks (NN) was presented describing the training objective of process separation given a data set using the cross-entropy (CE) as the training objective for tasks with multiple classes or the binary cross-entropy (BCE) in case of two classes, that are referred to as signal or background processes. In contrast to that, a novel method for NN training was presented introducing an optimization directly on the analysis objective: the uncertainty of the POI.
A demonstration of this novel technique was presented in chapter 4 and the proposed implementation was examined with regard to the stability of the training procedure. Upon this examination, the occurring problem of a collapsing NN output during the training was identified and a solution was introduced by modifying the existing training procedure, thereby mitigating the collapse, allowing for more explorative training. This method was then first demonstrated in a pseudo experiment with one signal and one background process, where the direct training on the analysis objective achieved an uncertainty reduction of approximately 15 % compared to conventional training on BCE. On top of that, a Taylor Coefficient Analysis was conducted, investigating the differences in the NN decision between the two methods. It was found that in the case of the training on the analysis objective, the focus of the NN shifts to the signal process identification with a consideration of the additionally introduced uncertainty effects on it.
Chapter 5 extended this novel NN optimization method further to be applicable for tasks that contain multiple processes, by presenting two variants in order to maintain the interpretability that is given by the conventional utilization of CE. In addition to the successful minimization of the systematic uncertainties, that were added to the problem, the new training approach was also able to provide an optimal minimization of the statistical part of the POI uncertainty which outperformed the conventional CE training.
In chapter 6 this method was applied to a subset of CMS data set that was used for the $H\to\tau\tau$ analysis evaluating the improvement on the POI uncertainty in the binary case and upon the introduction of multiple processes using several realistic detector and theory uncertainties whose effects were applied as weight corrections on histogram level. The conducted application of this novel approach has achieved a significant reduction in the uncertainty of around 25 % one single POI and 35 %, 57 % and 25 % for multiple POIs and provided a better signal separation in comparison to CE for signal processes whose uncertainties are dominated by statistics.

This study showed an optimized inference based method on a reduced data set of the CMS experiment considering a subset of uncertainty sources. A possible further step would be the inclusion of the remaining uncertainties and the extension of the training on the complete data set that is provided by the second data-taking period of the Large Hadron Collider. Another direction would be the building of a statistical model with incorporated uncertainties without the necessity of a histogram creation and an incorporation of an appropriate penalty for the process misclassification. This approach would thereby improve the training procedure as it would not be hindered by the discretization of the NN output and not necessitate a modification for an appropriate class assignment for problems that introduce multiple processes.

Appendix

A. Neural network output collapse on ATLAS Higgs Machine Learning data set

In the following, the problem of the collapsing NN output function due to the discretization of the NN output is displayed on the ATLAS Higgs Boson Machine Learning Challenge data set [67]. The collapse and the proposed modifications are visualized in form of event movement, similar to section 4.2 and the same NN architecture is used as described in section 4.2.
Utilized the proposed setup of [3], where the following variables are used for the NN input:`DER_mass_vis`, `DER_pt_h`, and `DER_deltaeta_jet_jet`. The systematic variation is introduced by a 10 % variation of `PRI_met` variable. The correction weights retrieved from this variation are applied as weight corrections to the NN output.

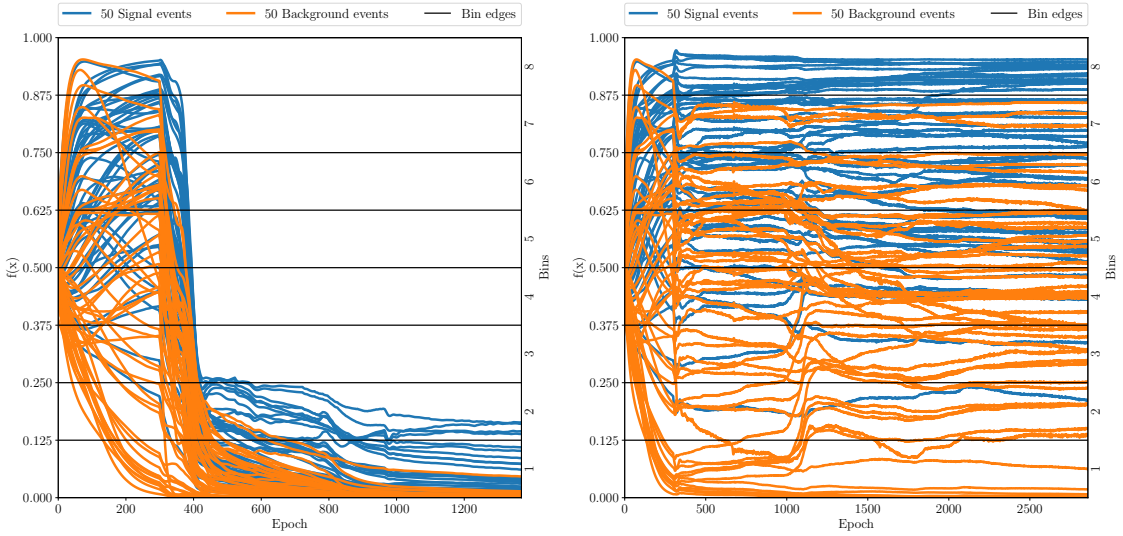
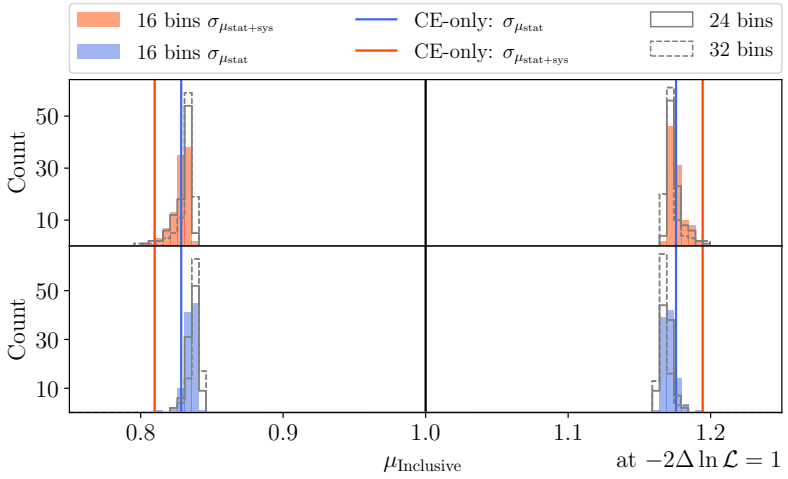


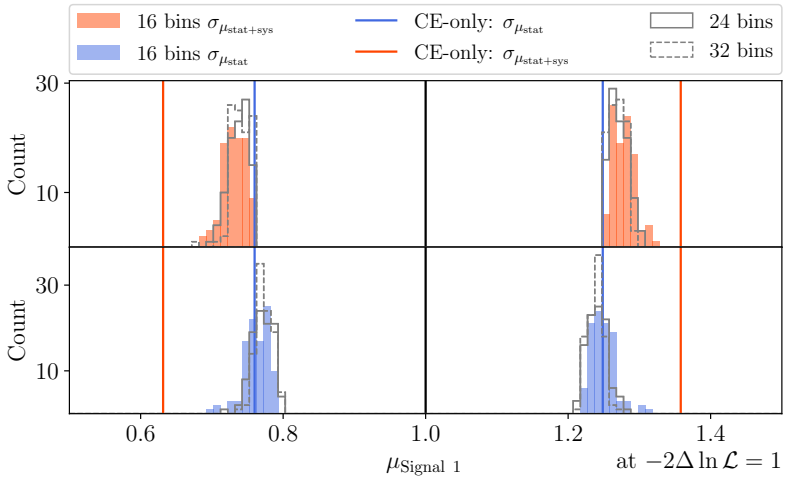
Figure A.1.: Shown are the event movements of 50 randomly chosen signal and background events of the ATLAS Higgs Boson Machine Learning Challenge data set. A warm-up phase of 300 epochs on BCE is applied in both cases. The resulting event movements of the proposed custom function for the histogram gradient [3] are shown in (a), whereas the results of the modification of the custom function as proposed in section 4.2 are shown in (b). Further information is given in section 4.2.

B. One-class ansatz on toy data set with an increased number of bins

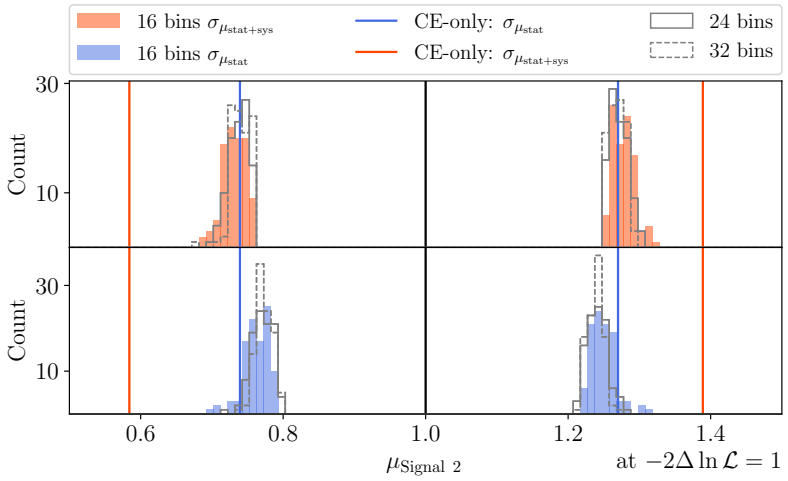
Shown are the summarized results of a conducted study on the used number of bins for the one-class approach on the pseudo experiment as discussed in section 5.1. Figure B.2 summarizes the results considering 24 and 32 bins in comparison to the 16 bins that are shown in section 5.2. A lower number of bins is not used due to separation issues of the processes.
The displayed figure is split into multiple figures, that are shown across multiple pages in order to improve readability.



(a)



(b)

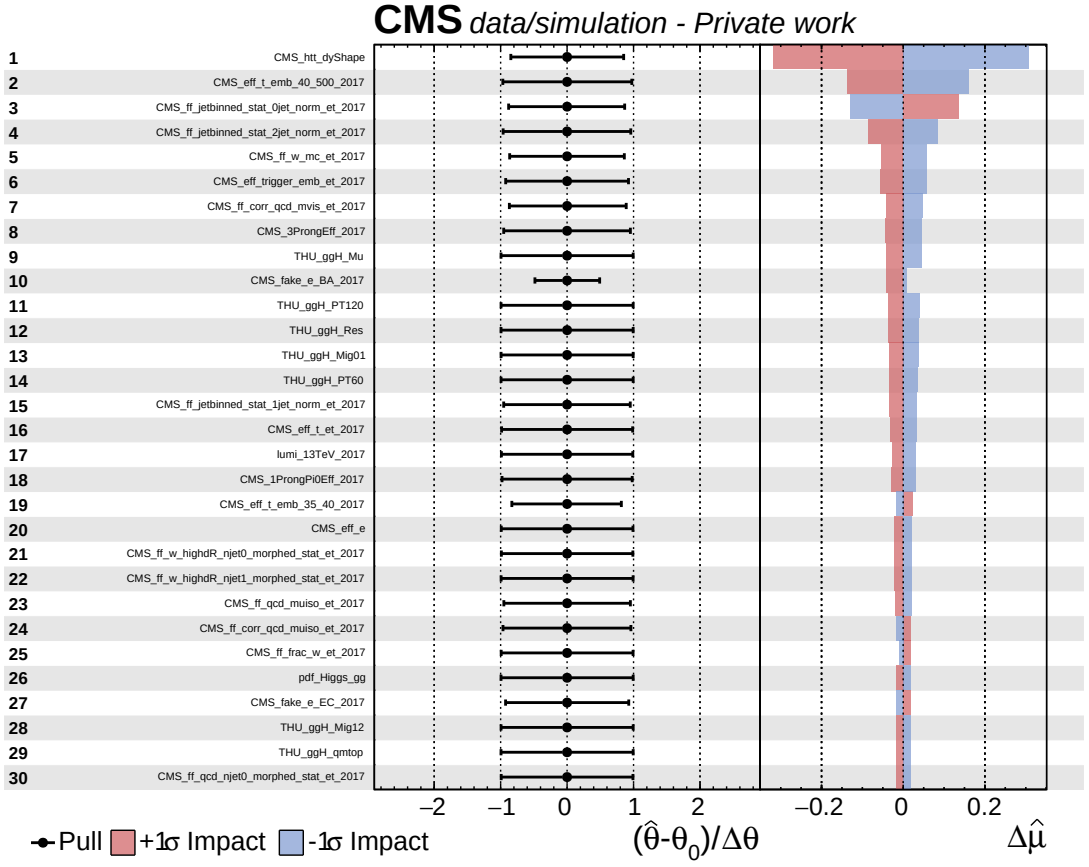


(c)

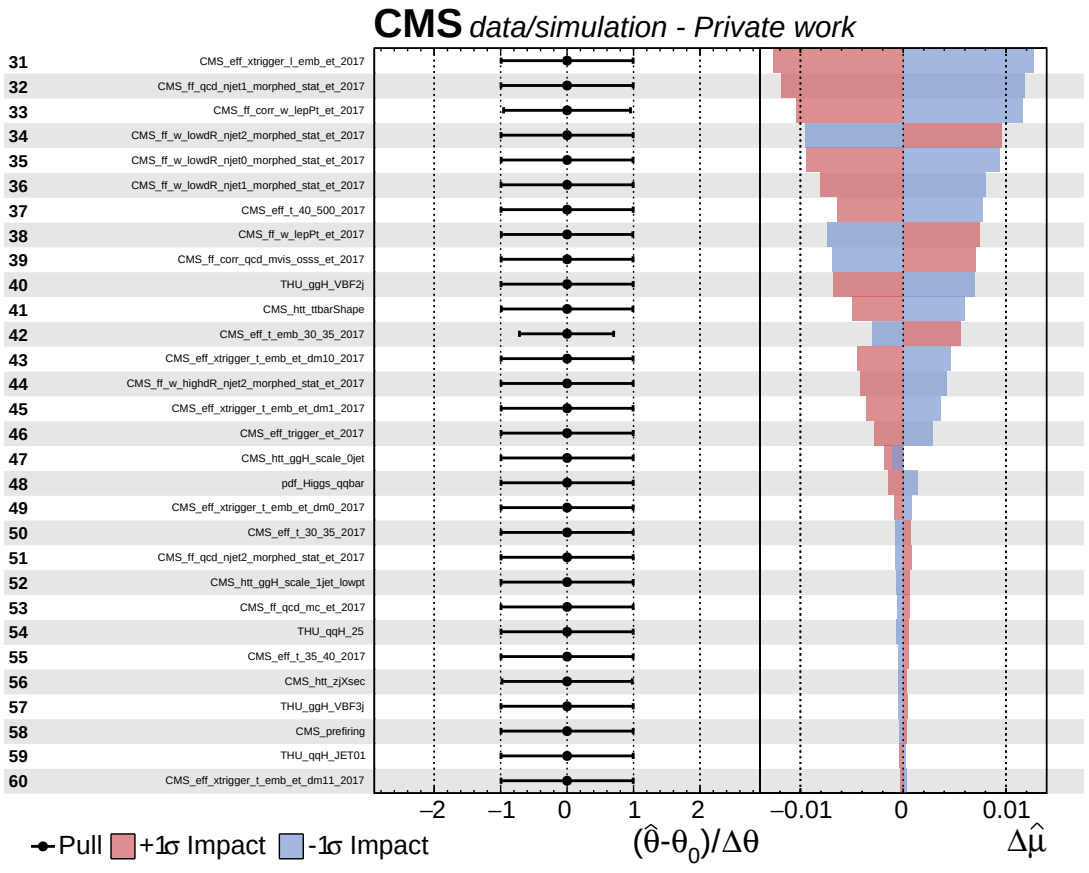
Figure B.2.: Results of the statistical inference comparing different numbers of used bins in the one-class approach on the pseudo experiment as discussed in subsection 5.2.1 as an ensemble test with a sample size of 100. Inclusive signal strength estimations are shown in (a). Differential signal strength estimations are shown in (b) for Signal 1 and in (c) for Signal 2. The vertical black line in each figure indicates the estimated signal strength. Statistical (statistical and systematic) uncertainties from the likelihood scans of the conducted benchmark using the CE loss are shown as blue (red) vertical lines. Filled histograms summarize the result of the one-class approach training using 16 bins as shown in section 5.2. Unfilled histograms with a solid (dashed) line show the results for the one-class approach using 24 (32) bins.

C. Uncertainty impacts in case of binary classification for the reduced CMS SM H→ττ data set

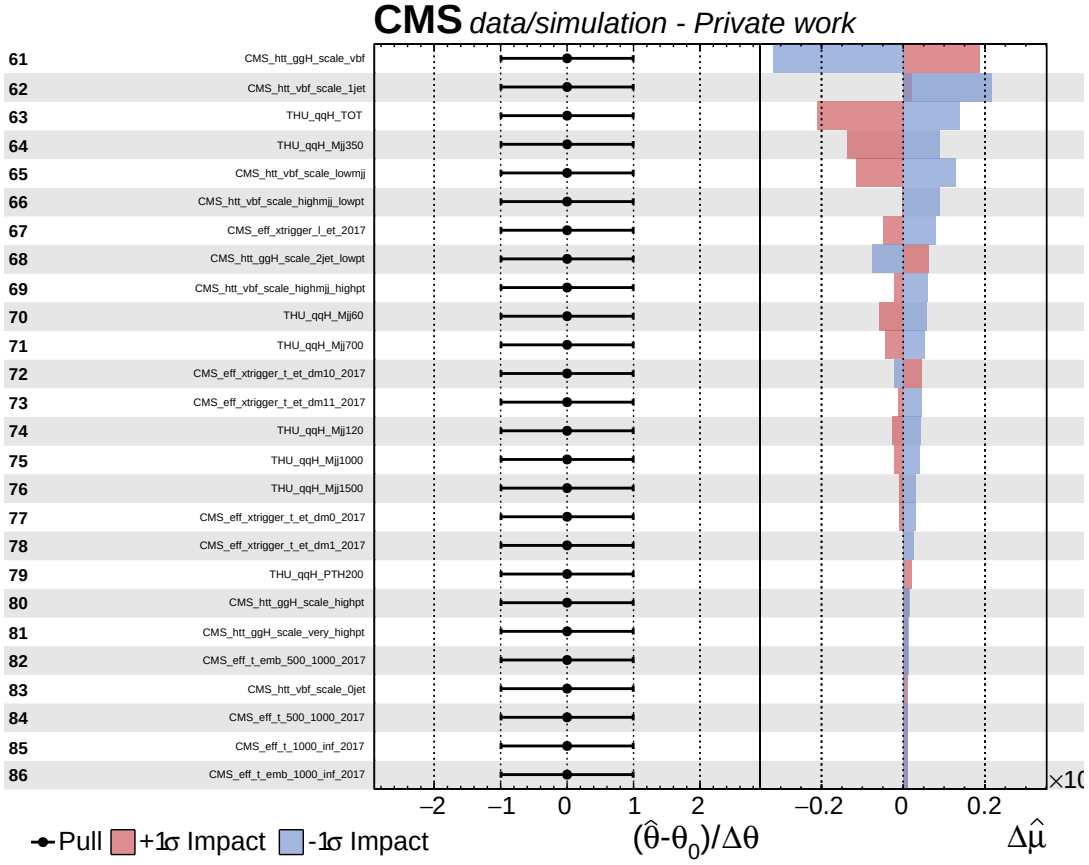
Listed are all uncertainties, that are considered in the application of the uncertainty-aware training on the reduced CMS data set that is used for the CMS Standard Model analysis [4]. The following figures list the uncertainties in descending order, depending on their impacts on the estimated signal strength uncertainty (right) that are retrieved from the statistical inference of a BCE training as discussed in section 6.2. The showed pulls are not deviating from zero due to the use of the Asimov data set and are omitted in following plots. A description of the used pre-, in-, and suffixes for the names are given in section 2.5. The shown list is split into multiple parts, that are shown across multiple pages in order to improve readability and show the change between different scales of the impacts.



(a)



(b)

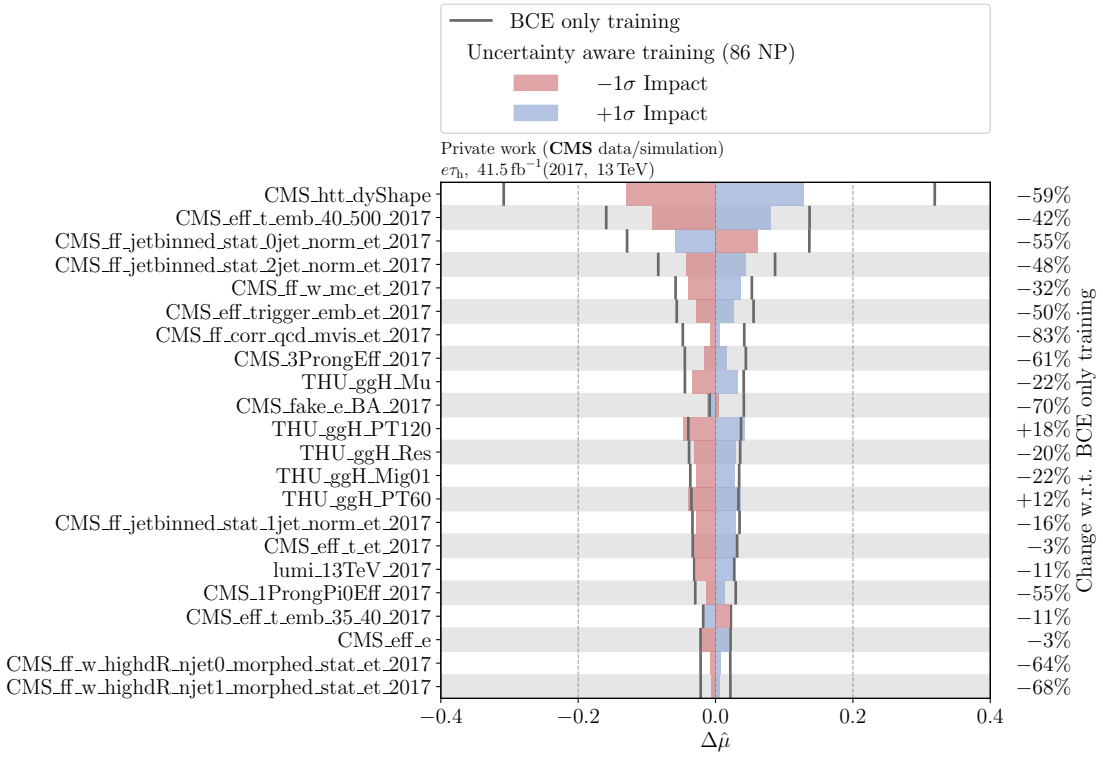


(c)

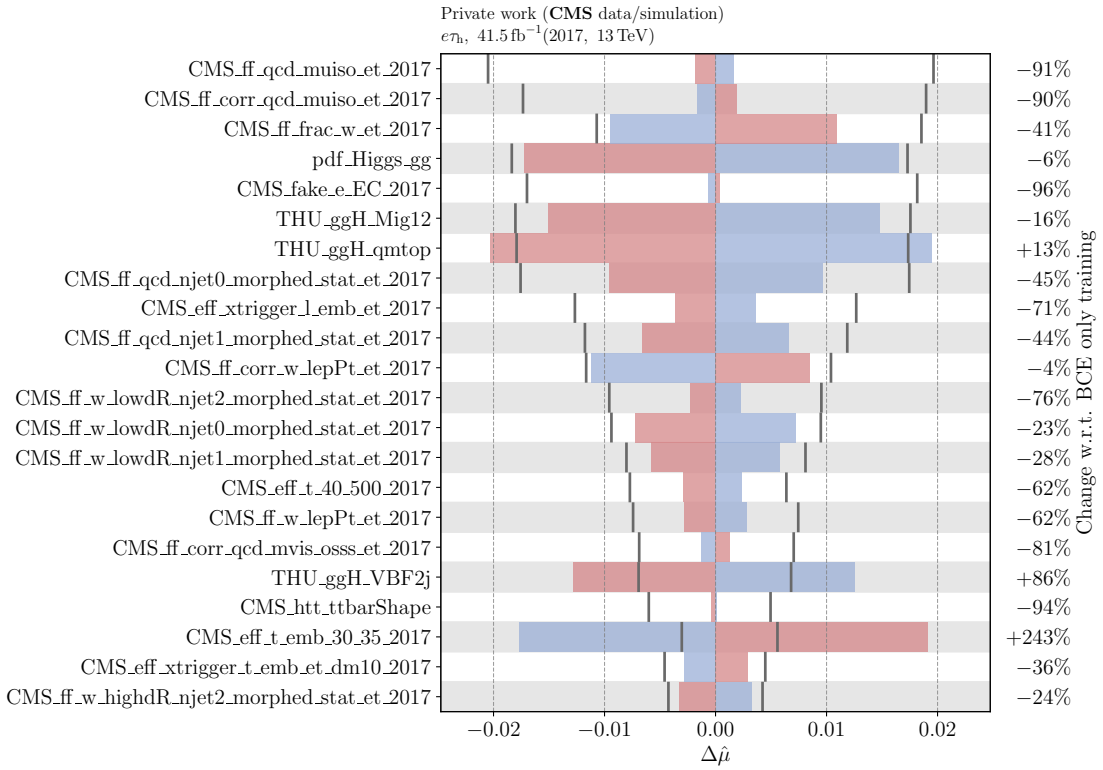
Figure C.3.: Impacts of the used uncertainties from the BCE training as conducted in section 6.2.

D. Impact changes of uncertainties between BCE and uncertainty-aware training in case of binary classification for the reduced CMS SM H→ττ data set

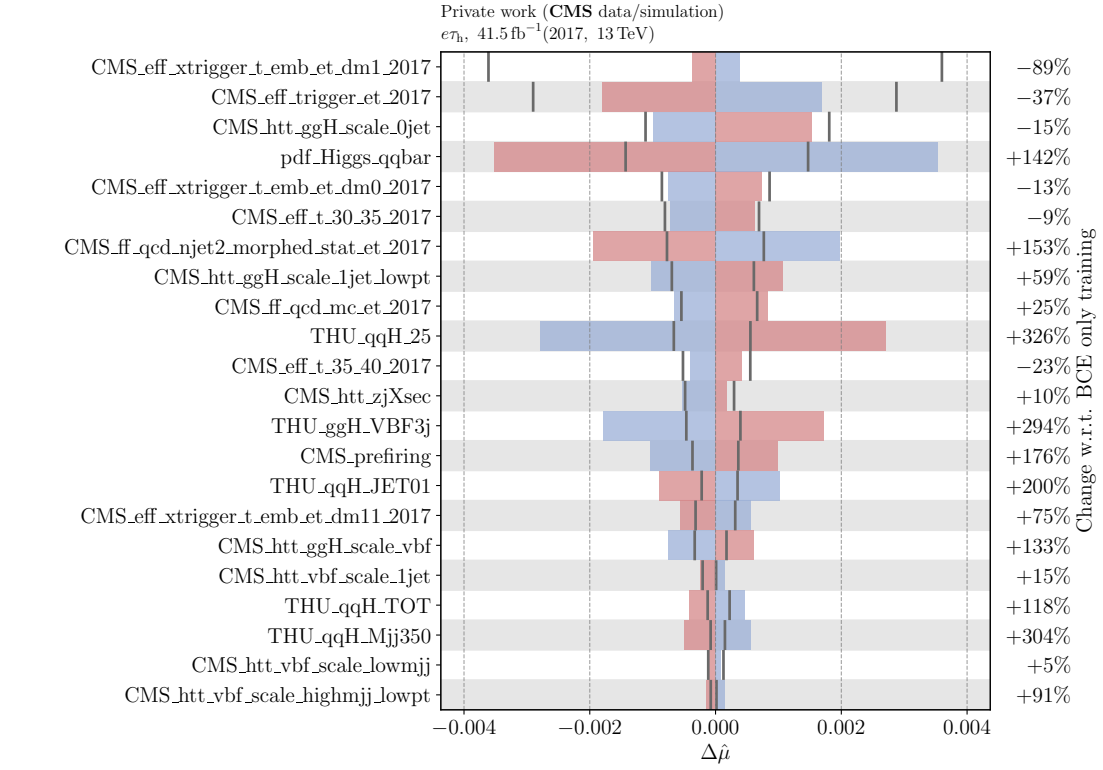
Changes of the impacts of all uncertainties between BCE and uncertainty-aware training on the reduced CMS data set that is used for the CMS Standard Model analysis [4]. A description of the used pre-, in-, and suffixes are given in section 2.5. The following figures list the uncertainties in descending order, depending on the impacts retrieved by the statistical inference from the results of BCE training as discussed in section 6.2.
The displayed list is split into multiple parts, that are shown across multiple pages in order to improve readability and show the change between different scales of the impacts.



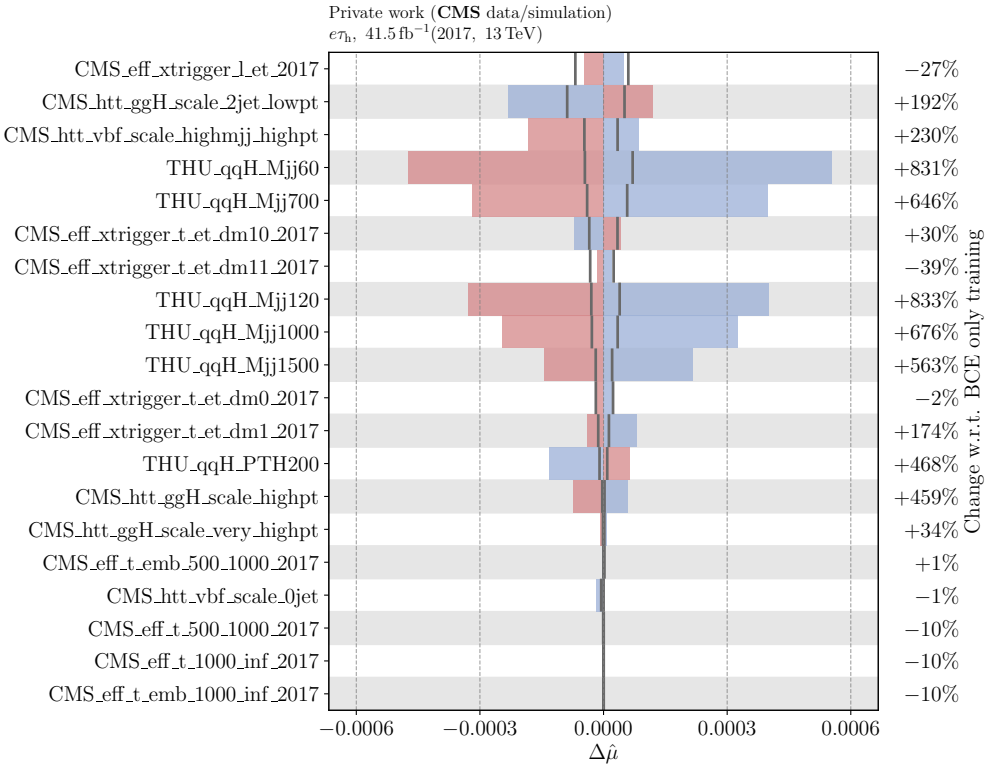
(a)


Private work (CMS data/simulation)
$e\tau_h$, 41.5 fb^{-1}(2017, 13 TeV)
CMS_ff_qcd_muiso_et_2017 −91%
CMS_ff_corr_qcd_muiso_et_2017 −90%
CMS_ff_frac_w_et_2017 −41%
pdf_Higgs_gg −6%
CMS_fake_e_EC_2017 −96%
THU_ggH_Mig12 −16%
THU_ggH_qmtop +13%
CMS_ff_qcd_njet0_morphed_stat_et_2017 −45%
CMS_eff_xtrigger_l_emb_et_2017 −71%
CMS_ff_qcd_njet1_morphed_stat_et_2017 −44%
CMS_ff_corr_w_lepPt_et_2017 −4%
CMS_ff_w_lowdR_njet2_morphed_stat_et_2017 −76%
CMS_ff_w_lowdR_njet0_morphed_stat_et_2017 −23%
CMS_ff_w_lowdR_njet1_morphed_stat_et_2017 −28%
CMS_eff_t_40_500_2017 −62%
CMS_ff_w_lepPt_et_2017 −62%
CMS_ff_corr_qcd_mvis_osss_et_2017 −81%
THU_ggH_VBF2j +86%
CMS_htt_ttbarShape −94%
CMS_eff_t_emb_30_35_2017 +243%
CMS_eff_xtrigger_t_emb_et_dm10_2017 −36%
CMS_ff_w_highdR_njet2_morphed_stat_et_2017 −24%
−0.02 −0.01 0.00 0.01 0.02
$\Delta\hat{\mu}$
Change w.r.t. BCE only training

(b)


Private work (CMS data/simulation)
$e\tau_h$, 41.5 fb^{-1}(2017, 13 TeV)
CMS_eff_xtrigger_t_emb_et_dm1_2017 −89%
CMS_eff_trigger_et_2017 −37%
CMS_htt_ggH_scale_0jet −15%
pdf_Higgs_qqbar +142%
CMS_eff_xtrigger_t_emb_et_dm0_2017 −13%
CMS_eff_t_30_35_2017 −9%
CMS_ff_qcd_njet2_morphed_stat_et_2017 +153%
CMS_htt_ggH_scale_1jet_lowpt +59%
CMS_ff_qcd_mc_et_2017 +25%
THU_qqH_25 +326%
CMS_eff_t_35_40_2017 −23%
CMS_htt_zjXsec +10%
THU_ggH_VBF3j +294%
CMS_prefiring +176%
THU_qqH_JET01 +200%
CMS_eff_xtrigger_t_emb_et_dm11_2017 +75%
CMS_htt_ggH_scale_vbf +133%
CMS_htt_vbf_scale_1jet +15%
THU_qqH_TOT +118%
THU_qqH_Mjj350 +304%
CMS_htt_vbf_scale_lowmjj +5%
CMS_htt_vbf_scale_highmjj_lowpt +91%
−0.004 −0.002 0.000 0.002 0.004
$\Delta\hat{\mu}$
Change w.r.t. BCE only training

(c)

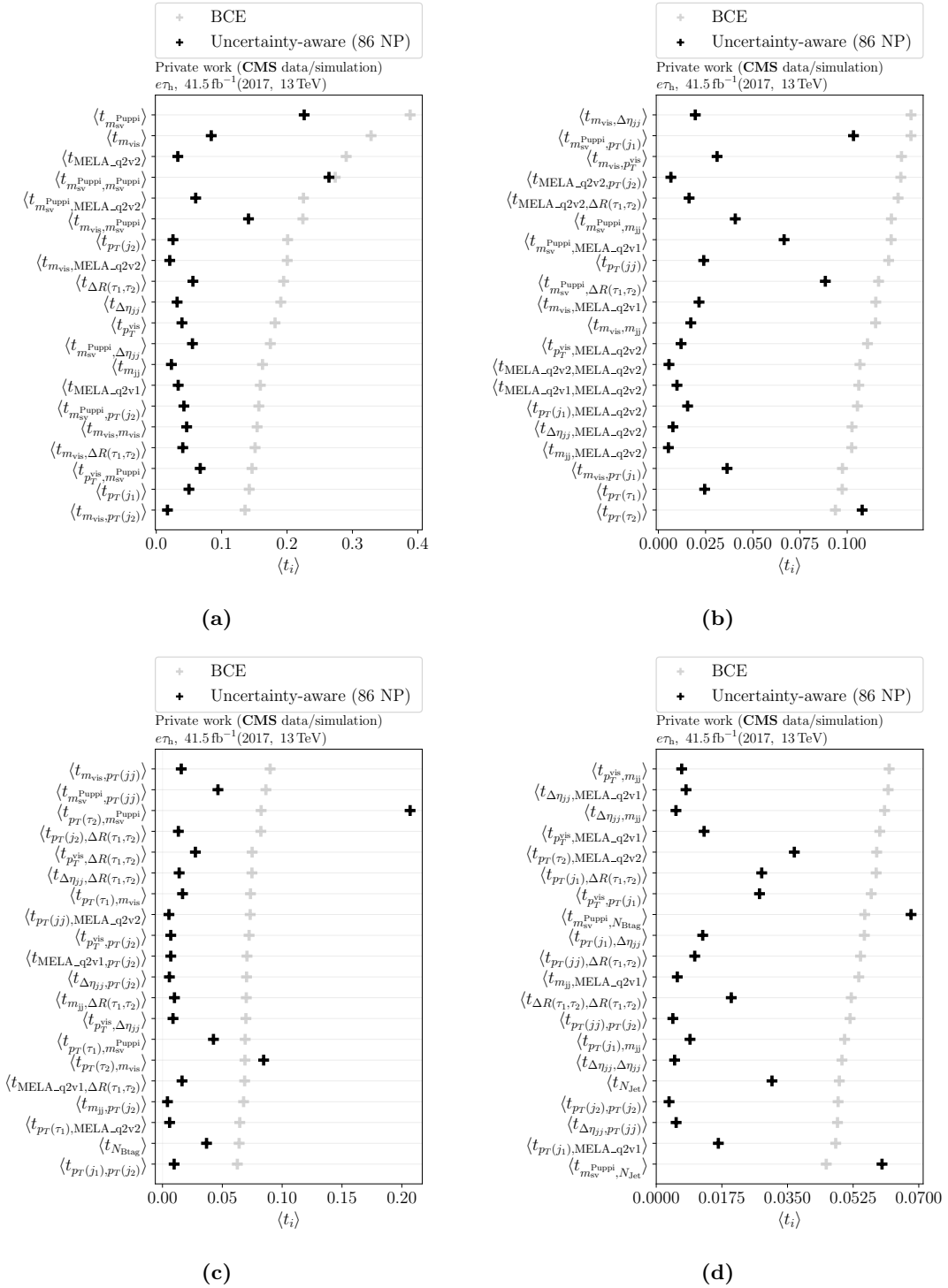


(d)

Figure D.4.: Impact changes on the estimated signal strength from the training on the BCE loss (grey) in comparison to the applied uncertainty-aware training as described in section 6.2. Colored bars indicate the impact of a performed $\pm 1\sigma$ shift of the nuisance parameter that corresponds to the systematic uncertainty. The impacts are split from (a) to (d) in descending order, showing different ranges. Further information is given at the beginning of appendix D and in section 6.2.

E. Comparison in the change of the importance of the input variables in the binary case of the reduced CMS Standard Model H→ττ data set

Full list of the Taylor Coefficient Analysis [68] conducted on the reduced CMS data set that is used for the CMS Standard Model H→ττ analysis in case of binary classification as discussed in section 6.2.



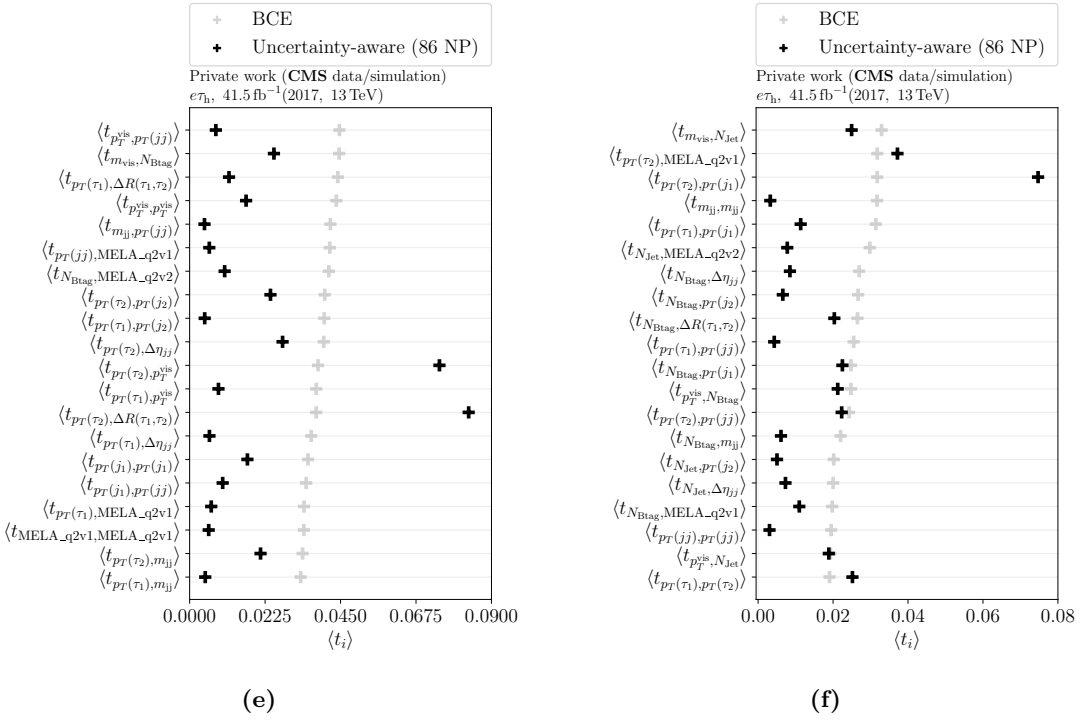
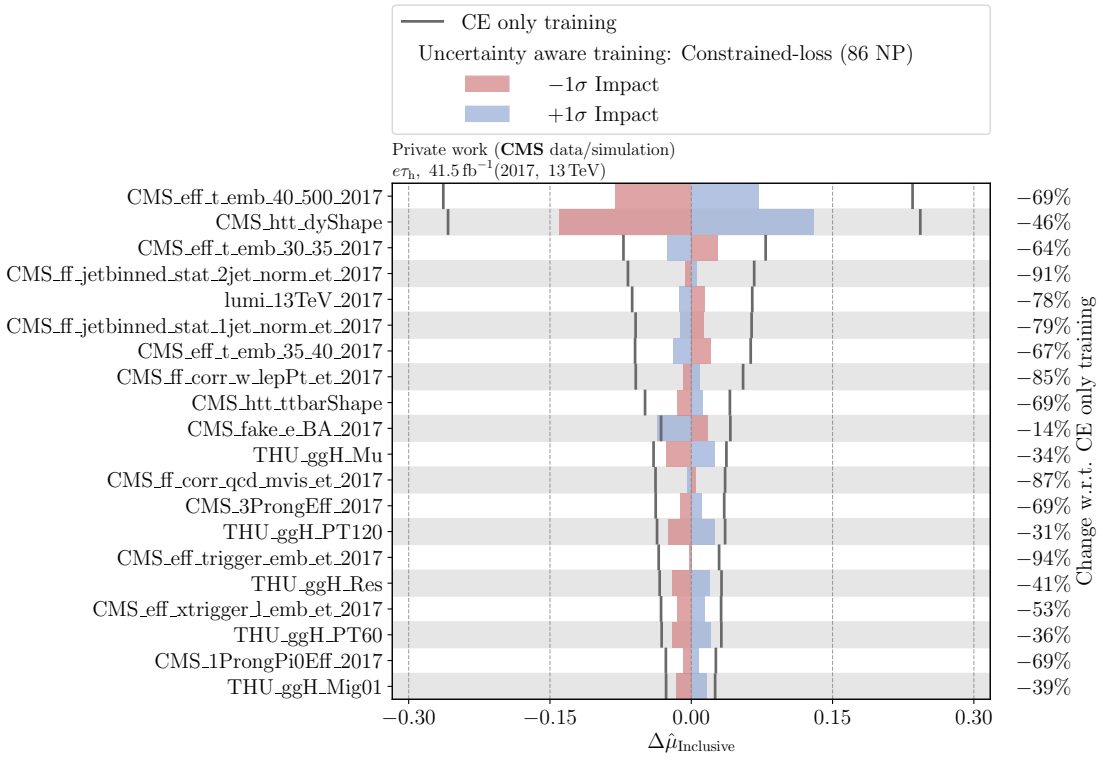


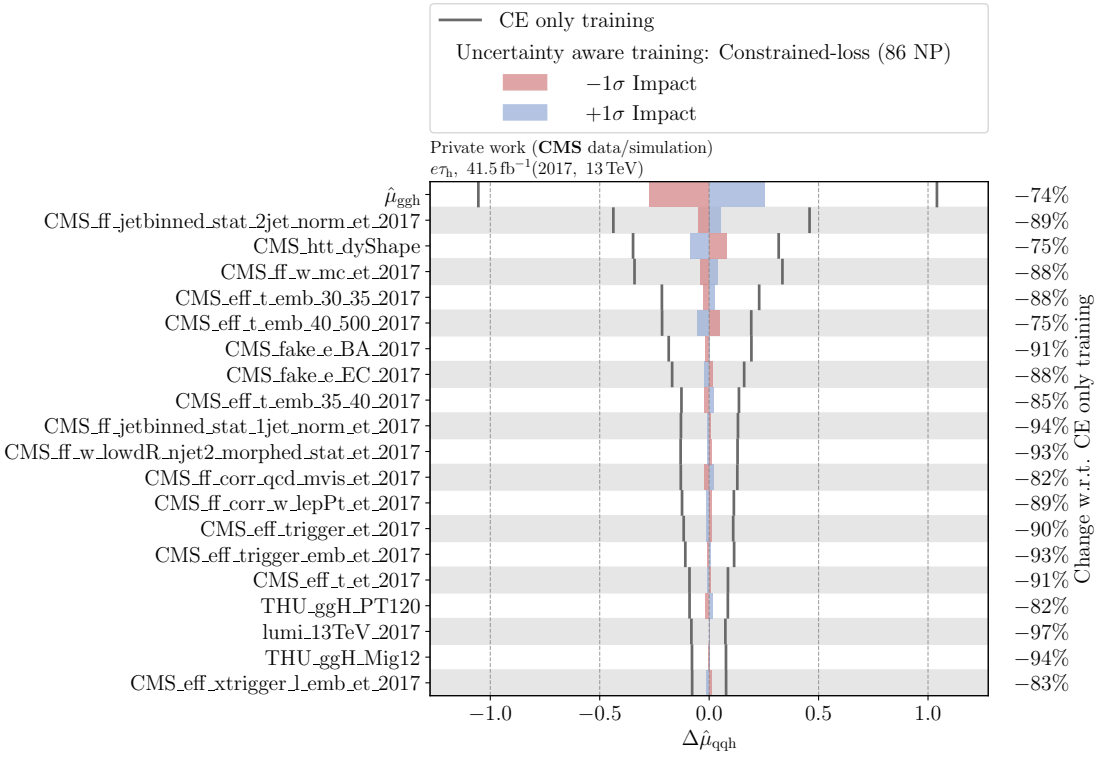
Figure E.5.: Shown are the results of a conducted Taylor Coefficient Analysis [68] on a reduced CMS data set used for the CMS Standard Model H→ττ Analysis [4] in case of a binary classification as discussed in section 6.2. The Taylor coefficients are retrieved from both folds of used NNs as proposed by the analysis. The results of the BCE training are shown in gray, whereas the results of the uncertainty-aware training utilizing all 86 introduced systematic uncertainties are shown in black. The Taylor coefficients are shown descending in their importance based on the BCE loss ranging across multiple figures form (a) to (g).

F. Comparison of the uncertainty impacts in the multi-class case on the reduced CMS Standard Model H→ττ data set

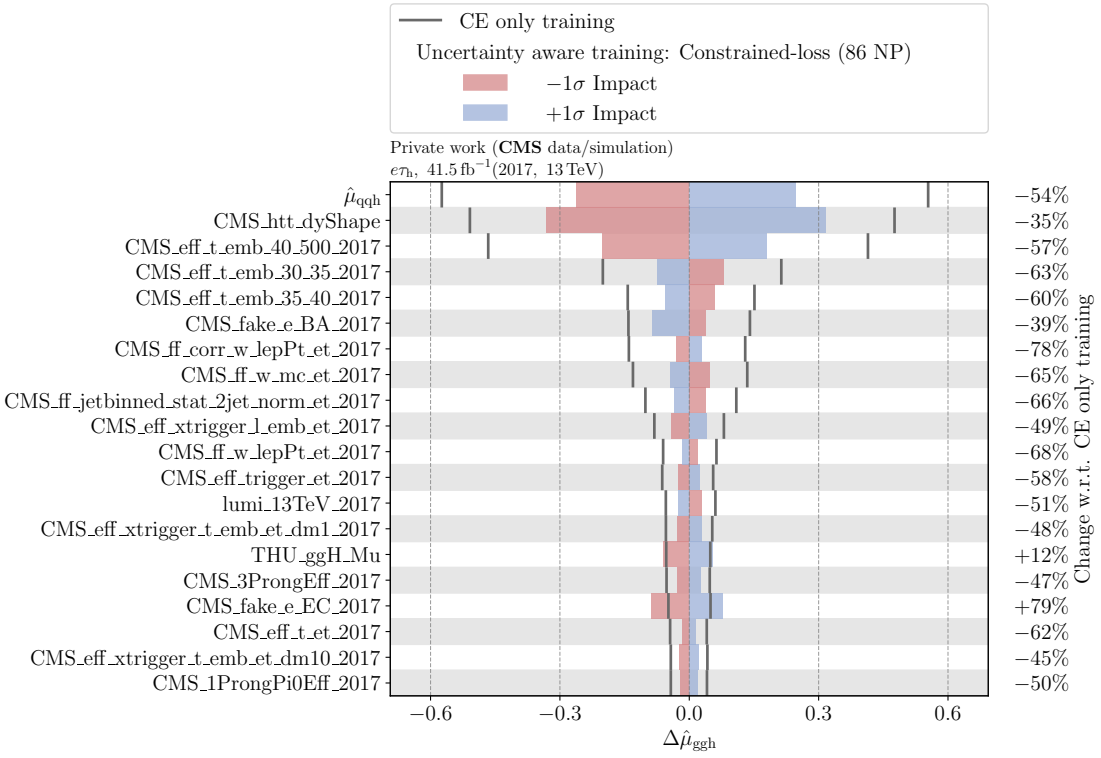
Changes of the impacts of the 20 most impactful uncertainties of the conducted training on CE and the two introduced modifications to the uncertainty-aware training in presence of multiple classes that are discussed in section 5.2. The following results are retrieved from the reduced CMS data set that is used for the CMS Standard Model analysis [4]. A description of the used pre-, in-, and suffixes are given in section 2.5 and further information about the used data is provided in chapter 6.
The shown list is split into multiple parts, that are shown across multiple pages in order to improve readability and show the change between different scales of the impacts.



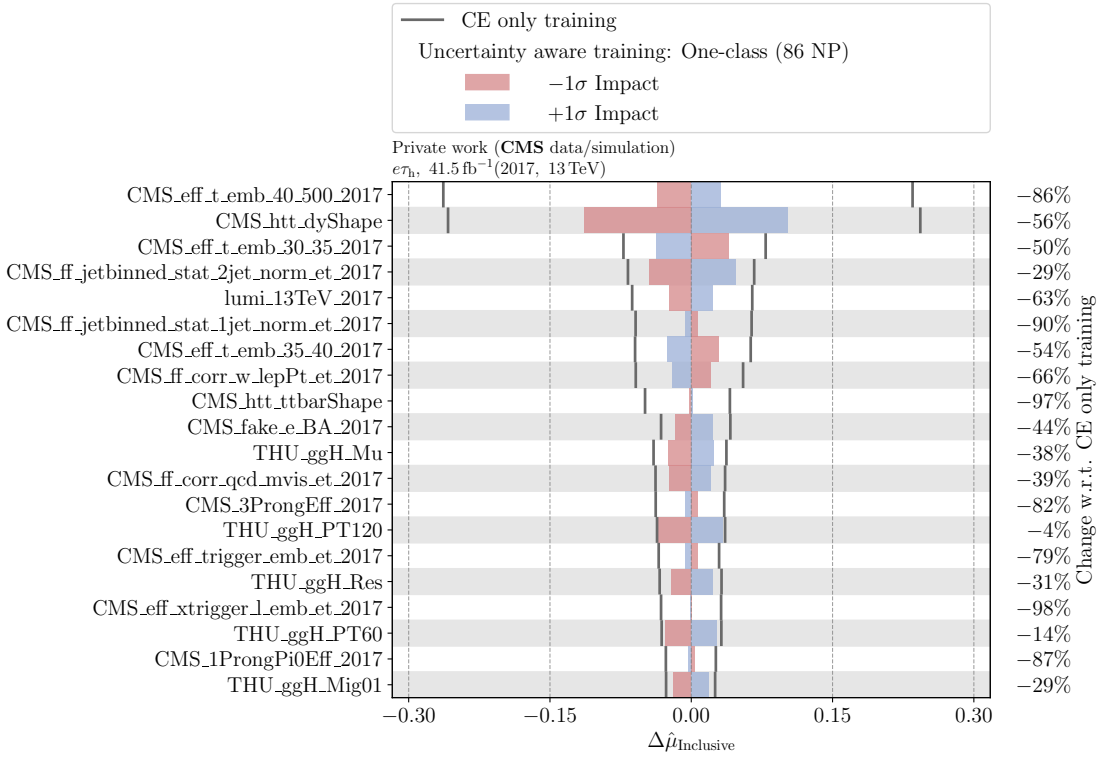
(a)


CE only training
Uncertainty aware training: Constrained-loss (86 NP)
-1σ Impact
$+1\sigma$ Impact
Private work (**CMS** data/simulation)
$e\tau_h$, 41.5 fb^{-1}(2017, 13 TeV)
$\hat{\mu}_{ggh}$ −74%
CMS_ff_jetbinned_stat_2jet_norm_et_2017 −89%
CMS_htt_dyShape −75%
CMS_ff_w_mc_et_2017 −88%
CMS_eff_t_emb_30_35_2017 −88%
CMS_eff_t_emb_40_500_2017 −75%
CMS_fake_e_BA_2017 −91%
CMS_fake_e_EC_2017 −88%
CMS_eff_t_emb_35_40_2017 −85%
CMS_ff_jetbinned_stat_1jet_norm_et_2017 −94%
CMS_ff_w_lowdR_njet2_morphed_stat_et_2017 −93%
CMS_ff_corr_qcd_mvis_et_2017 −82%
CMS_ff_corr_w_lepPt_et_2017 −89%
CMS_eff_trigger_et_2017 −90%
CMS_eff_trigger_emb_et_2017 −93%
CMS_eff_t_et_2017 −91%
THU_ggH_PT120 −82%
lumi_13TeV_2017 −97%
THU_ggH_Mig12 −94%
CMS_eff_xtrigger_l_emb_et_2017 −83%
Change w.r.t. CE only training
−1.0 −0.5 0.0 0.5 1.0
$\Delta\hat{\mu}_{qqh}$

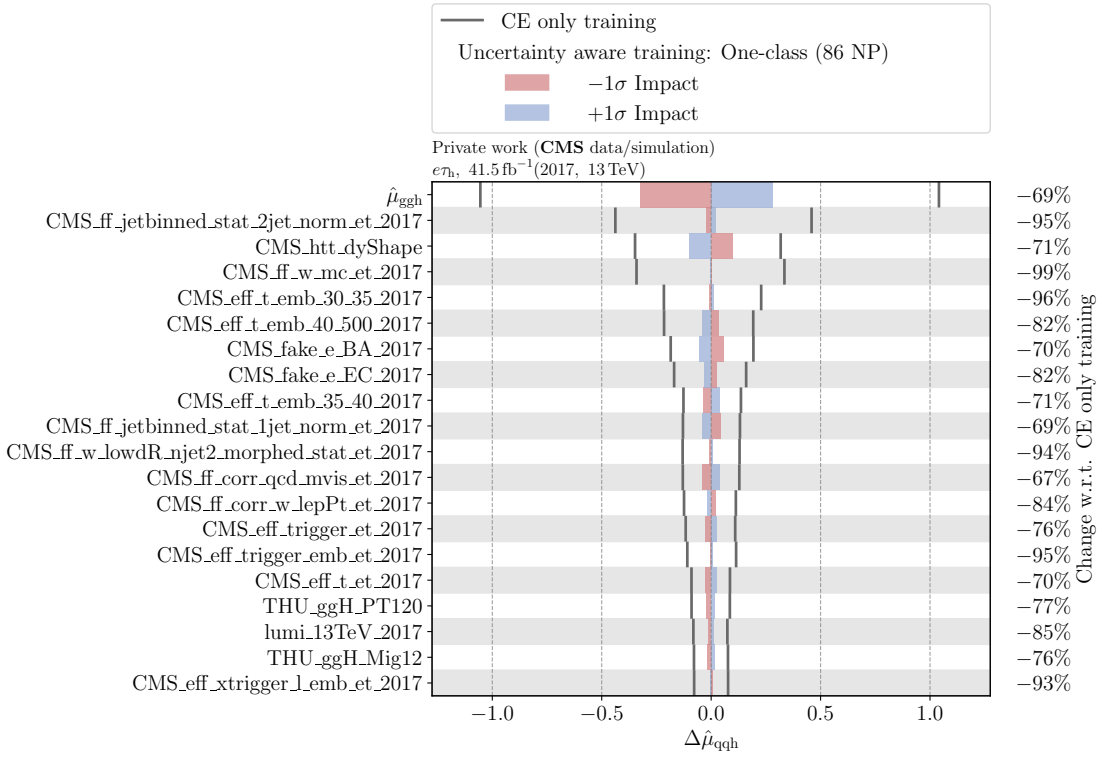
(b)


CE only training
Uncertainty aware training: Constrained-loss (86 NP)
-1σ Impact
$+1\sigma$ Impact
Private work (**CMS** data/simulation)
$e\tau_h$, 41.5 fb^{-1}(2017, 13 TeV)
$\hat{\mu}_{qqh}$ −54%
CMS_htt_dyShape −35%
CMS_eff_t_emb_40_500_2017 −57%
CMS_eff_t_emb_30_35_2017 −63%
CMS_eff_t_emb_35_40_2017 −60%
CMS_fake_e_BA_2017 −39%
CMS_ff_corr_w_lepPt_et_2017 −78%
CMS_ff_w_mc_et_2017 −65%
CMS_ff_jetbinned_stat_2jet_norm_et_2017 −66%
CMS_eff_xtrigger_l_emb_et_2017 −49%
CMS_ff_w_lepPt_et_2017 −68%
CMS_eff_trigger_et_2017 −58%
lumi_13TeV_2017 −51%
CMS_eff_xtrigger_t_emb_et_dm1_2017 −48%
THU_ggH_Mu +12%
CMS_3ProngEff_2017 −47%
CMS_fake_e_EC_2017 +79%
CMS_eff_t_et_2017 −62%
CMS_eff_xtrigger_t_emb_et_dm10_2017 −45%
CMS_1ProngPi0Eff_2017 −50%
Change w.r.t. CE only training
−0.6 −0.3 0.0 0.3 0.6
$\Delta\hat{\mu}_{ggh}$

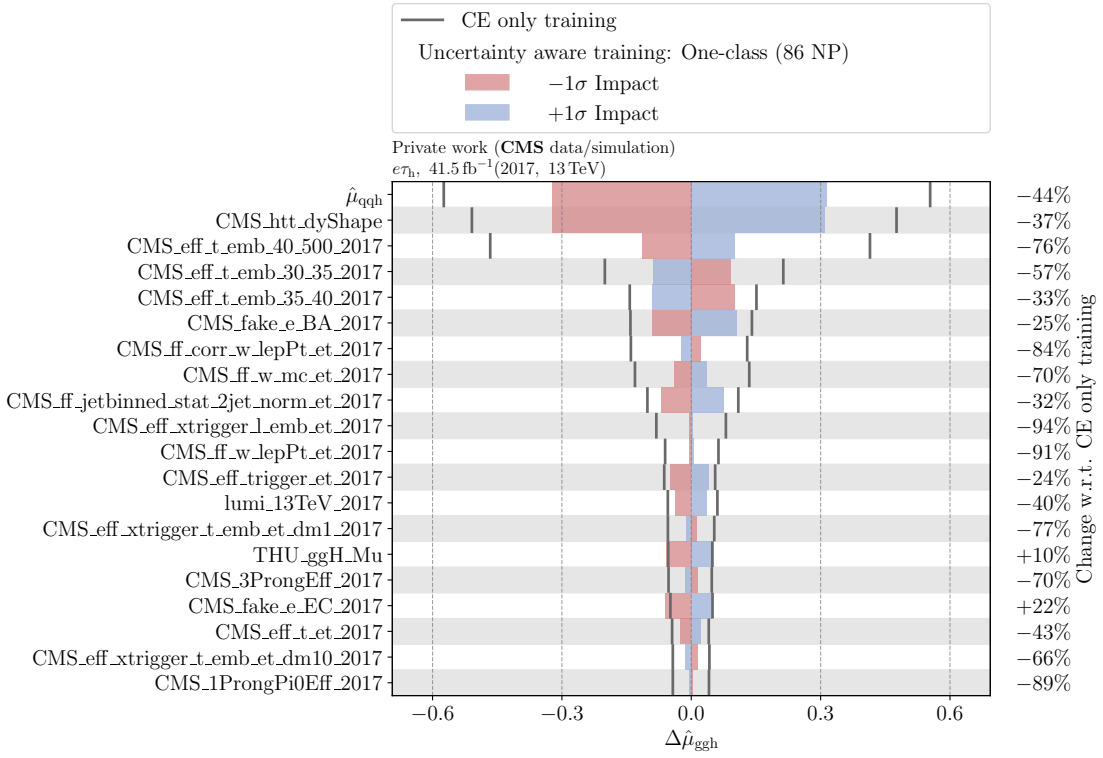
(c)


CE only training
Uncertainty aware training: One-class (86 NP)
-1σ Impact
$+1\sigma$ Impact
Private work (**CMS** data/simulation)
$e\tau_h$, 41.5 fb^{-1}(2017, 13 TeV)
CMS_eff_t_emb_40_500_2017 −86%
CMS_htt_dyShape −56%
CMS_eff_t_emb_30_35_2017 −50%
CMS_ff_jetbinned_stat_2jet_norm_et_2017 −29%
lumi_13TeV_2017 −63%
CMS_ff_jetbinned_stat_1jet_norm_et_2017 −90%
CMS_eff_t_emb_35_40_2017 −54%
CMS_ff_corr_w_lepPt_et_2017 −66%
CMS_htt_ttbarShape −97%
CMS_fake_e_BA_2017 −44%
THU_ggH_Mu −38%
CMS_ff_corr_qcd_mvis_et_2017 −39%
CMS_3ProngEff_2017 −82%
THU_ggH_PT120 −4%
CMS_eff_trigger_emb_et_2017 −79%
THU_ggH_Res −31%
CMS_eff_xtrigger_l_emb_et_2017 −98%
THU_ggH_PT60 −14%
CMS_1ProngPi0Eff_2017 −87%
THU_ggH_Mig01 −29%
−0.30
−0.15
0.00
0.15
0.30
$\Delta\hat{\mu}_{\text{Inclusive}}$
Change w.r.t. CE only training

(d)


CE only training
Uncertainty aware training: One-class (86 NP)
-1σ Impact
$+1\sigma$ Impact
Private work (**CMS** data/simulation)
$e\tau_h$, 41.5 fb^{-1}(2017, 13 TeV)
$\hat{\mu}_{\text{ggh}}$ −69%
CMS_ff_jetbinned_stat_2jet_norm_et_2017 −95%
CMS_htt_dyShape −71%
CMS_ff_w_mc_et_2017 −99%
CMS_eff_t_emb_30_35_2017 −96%
CMS_eff_t_emb_40_500_2017 −82%
CMS_fake_e_BA_2017 −70%
CMS_fake_e_EC_2017 −82%
CMS_eff_t_emb_35_40_2017 −71%
CMS_ff_jetbinned_stat_1jet_norm_et_2017 −69%
CMS_ff_w_lowdR_njet2_morphed_stat_et_2017 −94%
CMS_ff_corr_qcd_mvis_et_2017 −67%
CMS_ff_corr_w_lepPt_et_2017 −84%
CMS_eff_trigger_et_2017 −76%
CMS_eff_trigger_emb_et_2017 −95%
CMS_eff_t_et_2017 −70%
THU_ggH_PT120 −77%
lumi_13TeV_2017 −85%
THU_ggH_Mig12 −76%
CMS_eff_xtrigger_l_emb_et_2017 −93%
−1.0
−0.5
0.0
0.5
1.0
$\Delta\hat{\mu}_{\text{qqh}}$
Change w.r.t. CE only training

(e)



(f)

Figure F.6.: Shown are the 20 most impactful uncertainties on the estimated signal strength, ordered by the results of the training on CE loss. Utilized is the reduced CMS data set used for the CMS Standard Model H→ττ analysis [4] using multiple classes as described in section 6.1. The change is shown between the CE loss (gray) and one of the introduced approaches for the uncertainty-aware training that are described in section 5.2. Colored bars indicate the impact of a performed $\pm 1\sigma$ shift of the nuisance parameter that corresponds to the systematic uncertainty. The impacts are split into two parts. The first half displays the results of the estimated (a) inclusive signal strength, (b) differential signal strength of the qqh process, and (c) differential signal strength of the qqh process of the constrained-loss approach, whereas the second half, containing (d), (e), and (f) displays the results of the one-class approach accordingly. Further information is given at the beginning of appendix D.

References

[1] Stephane Fartoukh et al. *LHC Configuration and Operational Scenario for Run 3*. Tech. rep. Geneva: CERN, 2021. URL: https://cds.cern.ch/record/2790409.

[2] I. Zurbano Fernandez et al. "High-Luminosity Large Hadron Collider (HL-LHC): Technical design report". In: 10/2020 (Dec. 2020). Ed. by I. Béjar Alonso et al. DOI: 10.23731/CYRM-2020-0010.

[3] Stefan Wunsch et al. "Optimal statistical inference in the presence of systematic uncertainties using neural network optimization based on binned Poisson likelihoods with nuisance parameters". en. In: *Comput. Softw. Big Sci.* 5.1 (Dec. 2021).

[4] CMS Collaboration. *Measurements of Higgs boson production in the decay channel with a pair of leptons in proton-proton collisions at $\sqrt{s}$ = 13 TeV*. 2022. DOI: 10.48550/ARXIV.2204.12957.

[5] Sheldon L Glashow. "Partial-symmetries of weak interactions". en. In: *Nucl. Phys.* 22.4 (Feb. 1961), pp. 579–588.

[6] Steven Weinberg. "A model of leptons". In: *Phys. Rev. Lett.* 19.21 (Nov. 1967), pp. 1264–1266.

[7] Abdus Salam. "Weak and electromagnetic interactions". In: *Selected Papers of Abdus Salam*. WORLD SCIENTIFIC, May 1994, pp. 244–254.

[8] Mary K. Gaillard, Paul D. Grannis, and Frank J. Sciulli. "The standard model of particle physics". In: *Reviews of Modern Physics* 71.2 (Mar. 1999), S96–S111. DOI: 10.1103/revmodphys.71.s96.

[9] Sheldon L. Glashow. "Partial-symmetries of weak interactions". In: *Nuclear Physics* 22.4 (1961), pp. 579–588. ISSN: 0029-5582. DOI: https://doi.org/10.1016/0029-5582(61)90469-2.

[10] G. Arnison et al. "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.

[11] G. Arnison et al. "Experimental observation of lepton pairs of invariant mass around 95 GeV/c2 at the CERN SPS collider". In: *Physics Letters B* 126.5 (1983), pp. 398–410. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(83)90188-0.

[12] P. Bagnaia et al. "Evidence for Z0e+e at the CERN pp collider". In: *Physics Letters B* 129.1 (1983), pp. 130–140. ISSN: 0370-2693. DOI: https://doi.org/10.1016/0370-2693(83)90744-X.

[13] M. Banner et al. "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.

[14] F Englert and R Brout. “Broken symmetry and the mass of gauge vector mesons”. In: *Phys. Rev. Lett.* 13.9 (Aug. 1964), pp. 321–323.

[15] P W Higgs. “Broken symmetries, massless particles and gauge fields”. en. In: *Phys. Lett.* 12.2 (Sept. 1964), pp. 132–133.

[16] Peter W Higgs. “Broken symmetries and the masses of gauge bosons”. In: *Phys. Rev. Lett.* 13.16 (Oct. 1964), pp. 508–509.

[17] G S Guralnik, C R Hagen, and T W B Kibble. “Global Conservation Laws and Massless Particles”. In: *Phys. Rev. Lett.* 13.20 (Nov. 1964), pp. 585–587.

[18] Peter W Higgs. “Spontaneous symmetry breakdown without massless bosons”. en. In: *Phys. Rev.* 145.4 (May 1966), pp. 1156–1163.

[19] T W B Kibble. “Symmetry breaking in non-abelian gauge theories”. In: *Phys. Rev.* 155.5 (Mar. 1967), pp. 1554–1561.

[20] S. Chatrchyan et al. “Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC”. In: *Physics Letters B* 716.1 (Sept. 2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021.

[21] G. Aad et al. “Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC”. In: *Physics Letters B* 716.1 (Sept. 2012), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020.

[22] Cush MissMJ. *Standard Model of Elementary Particles*. [Online; accessed 15-January-2023]. 2022. URL: https://en.wikipedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg#file.

[23] “A portrait of the Higgs boson by the CMS experiment ten years after the discovery”. In: *Nature* 607.7917 (July 2022), pp. 60–68. DOI: 10.1038/s41586-022-04892-x.

[24] CERN. *CERN Yellow Reports: Monographs, Vol 2 (2017): Handbook of LHC Higgs cross sections: 4. Deciphering the nature of the Higgs sector*. en. 2017. DOI: 10.23731/CYRM-2017-002.

[25] Douglas Clowe et al. “A Direct Empirical Proof of the Existence of Dark Matter*”. In: *The Astrophysical Journal* 648.2 (Aug. 2006), p. L109. DOI: 10.1086/508162.

[26] The CMS Collaboration. *Recorded integrated luminosity at the CMS experiment*. [Online; accessed 03-January-2023]. 2019. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

[27] Ewa Lopienska. “The CERN accelerator complex, layout in 2022. Complexe des accélérateurs du CERN en janvier 2022”. In: (2022). General Photo. URL: https://cds.cern.ch/record/2800984.

[28] David Barney. “CMS Detector Slice”. CMS Collection. 2016. URL: https://cds.cern.ch/record/2120661.

[29] J. Rose, A. El Gamal, and A. Sangiovanni-Vincentelli. “Architecture of field-programmable gate arrays”. In: *Proceedings of the IEEE* 81.7 (1993), pp. 1013–1029. DOI: 10.1109/5.231340.

[30] CMScollaboration, A Tapper and Darin Acosta. *CMS Technical Design Report for the Level-1 Trigger Upgrade*. Tech. rep. Additional contacts: Jeffrey Spalding, Fermilab, Jeffrey.Spalding@cern.ch Didier Contardo, Universite Claude Bernard-Lyon I, didier.claude.contardo@cern.ch. 2013. URL: https://cds.cern.ch/record/1556311.

[31] Tomasz Bawej et al. “The New CMS DAQ System for Run-2 of the LHC”. In: *IEEE Transactions on Nuclear Science* 62.3 (June 2015). DOI: 10.1109/TNS.2015.2426216.

[32] A.M. Sirunyan et al. "Particle-flow reconstruction and global event description with the CMS detector". In: *Journal of Instrumentation* 12.10 (Oct. 2017), P10003. DOI: 10.1088/1748-0221/12/10/P10003.

[33] "Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s}$= 8 TeV". In: *J. Instrum.* 10.06 (June 2015), P06005–P06005.

[34] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. "The anti-k_t jet clustering algorithm". In: (Feb. 2008). arXiv: 0802.1189 [hep-ph].

[35] Markus Stoye and CMS collaboration. "Deep learning in jet reconstruction at CMS". In: *J. Phys. Conf. Ser.* 1085 (Sept. 2018), p. 042029.

[36] CMS Collaboration. "Performance of reconstruction and identification of τ leptons decaying to hadrons and ν_τ in pp collisions at $\sqrt{s} = 13$ TeV". In: (Sept. 2018). arXiv: 1809.02816 [hep-ex].

[37] "Performance of the DeepTau algorithm for the discrimination of taus against jets, electron, and muons". In: (2019). URL: https://cds.cern.ch/record/2694158.

[38] Particle Data Group, R. L. Workman et al. "Review of Particle Physics". In: *PTEP* 2022 (2022), p. 083C01. DOI: 10.1093/ptep/ptac097.

[39] CERN. *CERN Yellow Reports: Monographs, Vol 2 (2017): Handbook of LHC Higgs cross sections: 4. Deciphering the nature of the Higgs sector*. en. 2017. DOI: 10.23731/CYRM-2017-002.

[40] A.M. Sirunyan et al. "An embedding technique to determine $\tau\tau$ backgrounds in proton-proton collision data". In: *Journal of Instrumentation* 14.06 (June 2019), P06032–P06032. DOI: 10.1088/1748-0221/14/06/p06032.

[41] Sidney D. Drell and Tung-Mow Yan. "Massive Lepton-Pair Production in Hadron-Hadron Collisions at High Energies". In: *Phys. Rev. Lett.* 25 (5 Aug. 1970), pp. 316–320. DOI: 10.1103/PhysRevLett.25.316.

[42] A. M. et al. Sirunyan. "Measurement of the $Z/\gamma^* \to \tau\tau$ cross section in pp collisions at $\sqrt{s} = 13\,\text{TeV}$ and validation of τ lepton analysis techniques". In: *The European Physical Journal C* 78.9 (Sept. 2018), p. 708. ISSN: 1434-6052. DOI: 10.1140/epjc/s10052-018-6146-9.

[43] J. Andrejkovic et al. *Data-driven background estimation of fake-tau backgrounds in di-tau final states with 2016 and 2017 data*. en. Analysis Note CMS AN-2018/257. 2019. URL: https://cms.cern.ch/iCMS/jsp/db_notes/noteInfo.jsp?cmsnoteid=CMS%5C%20AN-2018/257.

[44] J. Andrejkovic et al. *Measurement of Higgs(125) boson properties in decays to a pair of tau leptons with full Run II data using Machine-Learning techniques*. en. Analysis Note CMS AN-2019/177. 2019. URL: https://cms.cern.ch/iCMS/jsp/db_notes/noteInfo.jsp?cmsnoteid=CMS%5C%20AN-2019/177.

[45] Roger Barlow and Christine Beeston. "Fitting using finite Monte Carlo samples". In: *Computer Physics Communications* 77.2 (1993), pp. 219–228. ISSN: 0010-4655. DOI: https://doi.org/10.1016/0010-4655(93)90005-W.

[46] J. S. Conway. *Incorporating Nuisance Parameters in Likelihoods for Multisource Spectra*. 2011. DOI: 10.48550/ARXIV.1103.0354.

[47] Jerzy Neyman and Egon S. Pearson. "On the Problem of the Most Efficient Tests of Statistical Hypotheses". In: *Philosophical Transactions of the Royal Society A* 231 (1933), pp. 289–337.

[48] Abraham Wald. “Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations is Large”. In: *Transactions of the American Mathematical Society* 54.3 (1943), pp. 426–482. ISSN: 00029947. URL: http://www.jstor.org/stable/1990256 (visited on Jan. 27, 2023).

[49] S S Wilks. “The large-sample distribution of the likelihood ratio for testing composite hypotheses”. In: *Ann. Math. Stat.* 9.1 (Mar. 1938), pp. 60–62.

[50] Isaac Asimov. *Franchise, in Isaac Asimov: The Complete Stories, Vol. 1*. Broadway Books, 1990.

[51] Glen Cowan et al. “Asymptotic formulae for likelihood-based tests of new physics”. In: *The European Physical Journal C* 71.2 (Feb. 2011). DOI: 10.1140/epjc/s10052-011-1554-0.

[52] Kurt Hornik, Maxwell B. Stinchcombe, and Halbert L. White. “Multilayer feedforward networks are universal approximators”. In: *Neural Networks* 2 (1989), pp. 359–366.

[53] F Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain”. en. In: *Psychol. Rev.* 65.6 (Nov. 1958), pp. 386–408.

[54] Kunihiko Fukushima. “Cognitron: A self-organizing multilayered neural network”. In: *Biological Cybernetics* 20.3 (Sept. 1975), pp. 121–136. ISSN: 1432-0770. DOI: 10.1007/BF00342633.

[55] Vinod Nair and Geoffrey E. Hinton. “Rectified Linear Units Improve Restricted Boltzmann Machines”. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML’10. Haifa, Israel: Omnipress, 2010, pp. 807–814. ISBN: 9781605589077.

[56] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. “Deep Sparse Rectifier Neural Networks”. In: *International Conference on Artificial Intelligence and Statistics*. 2011.

[57] Xavier Glorot and Yoshua Bengio. “Understanding the difficulty of training deep feedforward neural networks”. In: *International Conference on Artificial Intelligence and Statistics*. 2010.

[58] A. Cauchy. *Méthode générale pour la résolution des systèmes déquations simultanées*. Paris: C. R. Acad. Sci. Paris, 1847, 25:536–538.

[59] Adam Paszke et al. “PyTorch: An Imperative Style, High-Performance Deep Learning Library”. In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[60] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *CoRR* abs/1412.6980 (2014).

[61] Y. NESTEROV. “A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$”. In: *Doklady AN USSR* 269 (1983), pp. 543–547. URL: https://cir.nii.ac.jp/crid/1570572699326076416.

[62] Timothy Dozat. “Incorporating Nesterov Momentum into Adam”. In: 2016.

[63] Rory A. Fisher. “Theory of Statistical Estimation”. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 22 (1925), pp. 700–725.

[64] C. Radhakrishna Rao. “Information and the Accuracy Attainable in the Estimation of Statistical Parameters”. In: *Breakthroughs in Statistics: Foundations and Basic Theory*. Ed. by Samuel Kotz and Norman L. Johnson. New York, NY: Springer New York, 1992, pp. 235–247. ISBN: 978-1-4612-0919-5. DOI: 10.1007/978-1-4612-0919-5_16.

[65] Harald Cramér. *Mathematical Methods of Statistics* -. Kassel: Princeton University Press, 1999. ISBN: 069-1-005-478-.

[66] Pablo de Castro and Tommaso Dorigo. "INFERNO: Inference-Aware Neural Optimisation". In: *Computer Physics Communications* 244 (2019), pp. 170–179. ISSN: 0010-4655. DOI: https://doi.org/10.1016/j.cpc.2019.06.007.

[67] ATLAS Collaboration. *Dataset from the ATLAS Higgs boson machine learning Challenge 2014*. 2014.

[68] Stefan Wunsch et al. "Identifying the relevant dependencies of the neural network response on characteristics of the input space". en. In: *Comput. Softw. Big Sci.* 2.1 (Nov. 2018).

[69] John Platt and Alan Barr. "Constrained Differential Optimization". In: *Neural Information Processing Systems*. Ed. by D. Anderson. Vol. 0. American Institute of Physics, 1987. URL: https://proceedings.neurips.cc/paper/1987/file/a87ff679a2f3e71d9181a67b7542122c-Paper.pdf.

[70] Daniele Bertolini et al. "Pileup per particle identification". In: *Journal of High Energy Physics* 2014.10 (Oct. 2014), p. 59. ISSN: 1029-8479. DOI: 10.1007/JHEP10(2014)059.

[71] Andrei V. Gritsan et al. "Constraining anomalous Higgs boson couplings to the heavy-flavor fermions using matrix element techniques". In: *Phys. Rev. D* 94 (5 Sept. 2016), p. 055023. DOI: 10.1103/PhysRevD.94.055023.

[72] Sara Bolognesi et al. "Spin and parity of a single-produced resonance at the LHC". In: *Phys. Rev. D* 86 (9 Nov. 2012), p. 095031. DOI: 10.1103/PhysRevD.86.095031.

[73] CMS, *Measurement of Higgs boson production in the decay channel with a pair of τ leptons*. Tech. rep. Geneva: CERN, 2020. URL: https://cds.cern.ch/record/2725590.