Karlsruher Institut für Technologie

Modern machine learning in the presence of systematic uncertainties for robust and optimized multivariate data analysis in high-energy particle physics

Zur Erlangung des akademischen Grades eines

DOKTORS DER NATURWISSENSCHAFTEN

von der KIT-Fakultät für Physik des
Karlsruher Instituts für Technologie (KIT)
angenommene

DISSERTATION

von

M.Sc. Stefan Wunsch

Tag der mündlichen Prüfung: 29. Januar 2021

Referent:	Prof. Dr. Günter Quast
Korreferent:	Priv. Doz. Dr. Roger Wolf
Betreuer am CERN:	Dr. Lorenzo Moneta

Karlsruhe Institute of Technology

Modern machine learning in the presence of systematic uncertainties for robust and optimized multivariate data analysis in high-energy particle physics

For obtaining the academic degree

DOKTOR DER NATURWISSENSCHAFTEN

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

DOCTORAL THESIS

of

M.Sc. Stefan Wunsch

Day of the oral exam: 29 January 2021

Reviewer:	Prof. Dr. Günter Quast
Second reviewer:	Priv. Doz. Dr. Roger Wolf
CERN supervisor:	Dr. Lorenzo Moneta

Abstract

In high energy particle physics, machine learning has already proven to be an indispensable technique to push data analysis to the limits. So far widely accepted and successfully applied in the event reconstruction at the LHC experiments, machine learning is today also increasingly often part of the final steps of an analysis and, for example, used to construct observables for the statistical inference of the physical parameters of interest. This thesis presents such a machine learning based analysis measuring the production of Standard Model Higgs bosons in the decay to two tau leptons at the CMS experiment and discusses the possibilities and challenges of machine learning at this stage of an analysis. To allow for a precise and reliable physics measurement, the application of the chosen machine learning model has to be well under control. Therefore, novel techniques are introduced to identify and control the dependence of the neural network function on features in the multidimensional input space. Further, possible improvements of machine learning based analysis strategies are studied. A novel solution is presented to maximize the expected sensitivity of the measurement to the physics of interest by incorporating information about known uncertainties in the optimization of the machine learning model, yielding an optimal statistical inference in the presence of systematic uncertainties.

Contents

CHAPTER 1

Introduction

Research in the field of high energy particle physics (HEP) is at the forefront of modern data analysis due to the complexity and the massive amount of data from the experiments. For example, the Compact Muon Solenoid (CMS) experiment integrated in the Large Hadron Collider (LHC) at the European Organization for Nuclear Research (CERN) has recorded in 2016 to 2018 data with an integrated luminosity of $150\,\mathrm{fb}^{-1}$ [1]. The data contains records of proton-proton collisions consisting of quadrillions of individual particle interactions measured with millions of detector channels [2]. This translates into exabytes of data ready to be analyzed by thousands of physicists around the globe.

Traditionally, the data is analyzed by isolating the interactions of interest with selection requirements on physical observables, e.g., the invariant mass of a decay system, motivated by the underlying theory and knowledge about the detector. Milestones in the development of the Standard Model (SM) of particle physics [3–8] have been achieved with such analysis strategies, for example the W and Z boson discoveries in the 1980s [9, 10].

In the following twenty years, not only the hardware of HEP experiments got more sophisticated but also the data analysis techniques evolved significantly. The community increasingly incorporated multivariate methods in their data analysis, at that time considerably driven by the TMVA project [11]. These developments resulted in the discovery of the Higgs boson in 2012 [12, 13] powered by multivariate techniques, mostly boosted decision trees. However, these techniques were primarily deployed in the intermediate steps of the data processing such as the electron and tau identification [14–16] and less in the final steps of a measurement, e.g., as observable for the statistical inference.

In the meantime, multivariate analysis techniques, now typically referred to as machine learning (ML) techniques, experienced an explosively growing attention in industry and research. Spectacularly visible in the massive improvements in the Large Scale Visual Recognition Challenge [17] with convolutional neural network (NN) architectures like AlexNet [18], the flexibility and applicability of ML started to flourish by expanding the mindset of NNs to a framework of decomposable pieces resulting in an enormous collection of new ML models.

This thesis studies and expands the application of such modern ML techniques to data analysis in HEP. Nowadays, after a widespread acceptance of ML in the first

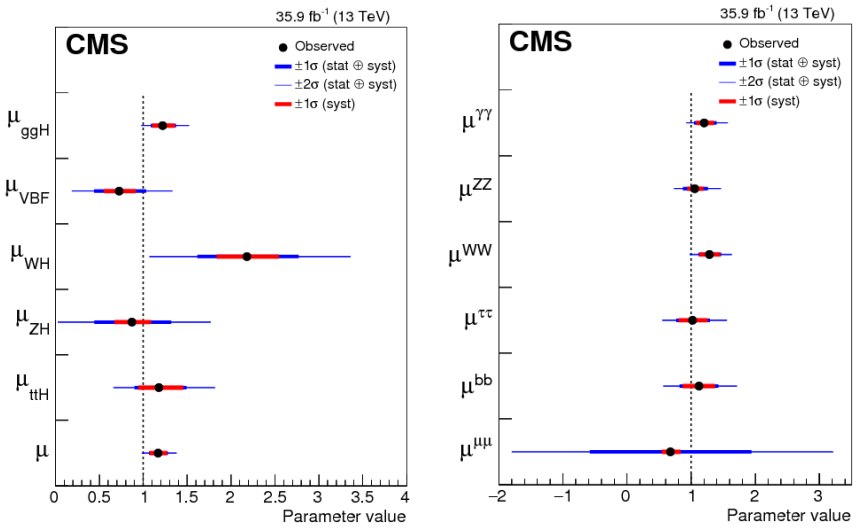

Figure 1.1: Summary of the measured signal strength modifiers with respect to the SM expectation per production mode (left) and per decay mode (right) of the Higgs boson [19].

steps of the data analysis toolchain like identification and reconstruction of particles, the community moves towards a more frequent application of ML also in the very-end steps of a measurement like the usage of fully ML based observables as direct input to the statistical inference. Since analysis in HEP is driven by precision and therefore the best possible knowledge of all contributing factors to the measurement are a crucial ingredient, heavily ML based analysis strategies pose new challenges to the analyzers. The full control of the applied ML techniques becomes a fundamental element of data analysis to be able to determine precisely the effects of statistical and systematic variations on the measurement.

This effort requires the understanding of the dependence of the ML model on features in the multidimensional input space. Chapter 3 discusses the state-of-the-art methods from the ML community and puts them in context with the requirements of such techniques for HEP analysis. A novel approach is presented to analyze the relevant dependencies of the NN function on characteristics in the multidimensional input space with special emphasis on the identification of higher-order features such as correlations between inputs.

Figure 1.1 shows the latest combined measurements from the CMS collaboration of the production and decay rates of the Higgs boson based on data taken in 2016 with an integrated luminosity of $36\,\mathrm{fb}^{-1}$. Already at this stage, the measurements are not anymore in the first place limited by the amount of data recorded, represented by the statistical uncertainty, but by the known unknowns of the analysis, the systematic uncertainties. The effect is even more pronounced in today's analyses with data from the full LHC Run 2 corresponding to $150\,\mathrm{fb}^{-1}$. This is demonstrated in chapter 2 for the differential measurement of the production of SM Higgs bosons in the decay to two tau leptons. Analyses in the future with data from LHC Run 3 and 4 with an anticipated delivered integrated luminosity of $300\,\mathrm{fb}^{-1}$ and $3000\,\mathrm{fb}^{-1}$ are expected to intensify this

development [20]. The projection raises the question how the existing analysis strategies in HEP could be adapted to achieve in such scenarios the best possible physics measurement. A possible approach to include the knowledge about systematic uncertainties in the design of the analysis strategy is the usage of ML techniques, which explicitly allow to control the dependence of the ML model on systematic variations. Chapter 4 reviews existing methods and presents a novel technique to reduce the dependence of the NN function to systematic variations in the multidimensional input space. A more direct approach is optimizing the analysis strategy based on the expected sensitivity to the physics target of interest. Chapter 5 studies techniques and their challenges to optimize with ML methods the analysis strategy directly on the objective of the statistical inference, e.g., the variance of a parameter of interest. A novel solution is presented to use information from binned Poisson likelihoods with nuisance parameters to maximize the expected sensitivity of the analysis resulting in an optimal statistical inference in the presence of systematic uncertainties.

CHAPTER 2

Machine learning based analysis of the production of Standard Model Higgs bosons in the decay to two tau leptons

This chapter contains the physics background and the details about a ML based data analysis strategy to perform a cross section measurement of the production of SM Higgs bosons in the decay to two tau leptons at the CMS experiment. Special emphasis is put on the event categorization and observables based on NNs, and the implications of the usage of ML on the statistical inference, for example the treatment of systematic uncertainties. The analysis is used as an example to discuss the challenges and their solutions to enable a robust data analysis while taking advantage of modern ML methods.

2.1 The Compact Muon Solenoid experiment at the Large Hadron Collider

The CMS experiment is located at CERN near Geneva and the Swiss-French border. The particle detector is placed in one of the four interaction points of the LHC where bunches of approximately 10^{11} protons collide at a rate of 40 MHz with a center of mass energy of 13 TeV. The protons reach nearly the speed of light which is made possible by the accelerator complex at CERN because previous state of the art accelerators such as the Proton Synchrotron Booster (PSB), Proton Synchrotron (PS) and Super Proton Synchrotron (SPS) can be used to preaccelerate the particles already at high energies before injecting them into the LHC. Figure 2.1 shows an overview over the CERN accelerator complex.

The CMS detector is built in layers around the interaction point, see figure 2.2 for a transverse slice of the detector. The innermost part is the tracking system consisting of layers of silicon pixel and strip detectors. Because each bunch crossing results on average for LHC Run 2 in about 30 collisions with many individual particles per interaction [1], detailed information about the path of the particles is crucial to reconstruct the events. The tracker is capable to record the path of the charged particles with a spatial resolution of a few µm. Due to the magnetic field of 3.8 T parallel to the beam pipes, the curvature

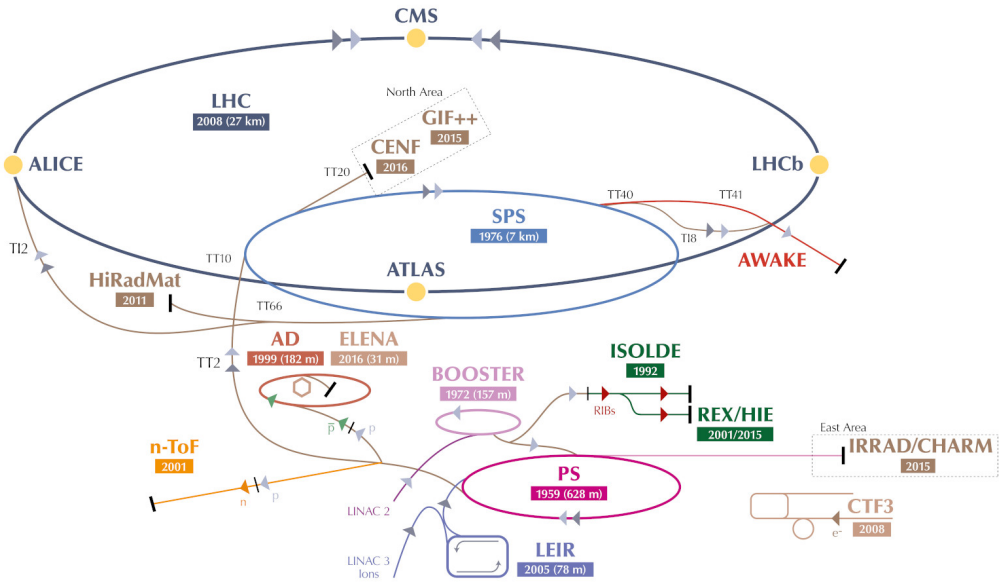

Figure 2.1: Overview over the CERN accelerator complex including the CMS experiment at the LHC [21].

of the tracks can be used to measure precisely the transverse momentum of the charged particles.

To reconstruct the full kinematic properties of the particles, also the energy has to be measured. For this purpose, first the electromagnetic calorimeter (ECAL) and then the hadronic calorimeter (HCAL) enclose the tracker, capturing electrons and photons in the first and hadrons in the second subdetector. Both are scintillator detectors, the ECAL is made from lead tungstate and the HCAL is made from plastic interleaved with brass.

From the known charged elementary particles, only muons can escape the first three subdetectors because of their low interaction rate with matter. Therefore, the outermost layer of the CMS detector, integrated into the return yoke of the magnet, is dedicated to the detection of muons. Built from different detector types, drift tubes in the barrel region and cathode strip chambers at the endcaps, the system provides additional information about the muon momentum. Because muons are often a hint for interesting physics, resistive plate chambers are used in addition to signal such events quickly to the trigger system.

Because the CMS detector is subject to a bunch crossing at a rate of 40 MHz and a full event has an information content after zero suppression of approximately 1 MB [2], recording every event for further analysis would result in an unmanageable data rate of about $40\,\mathrm{TB\,s^{-1}}$. To keep the interesting physics events, a two-stage trigger system is deployed. The first stage is implemented using field programmable gate arrays (FPGAs) based on information from the ECAL, HCAL and the muon system, deciding rapidly

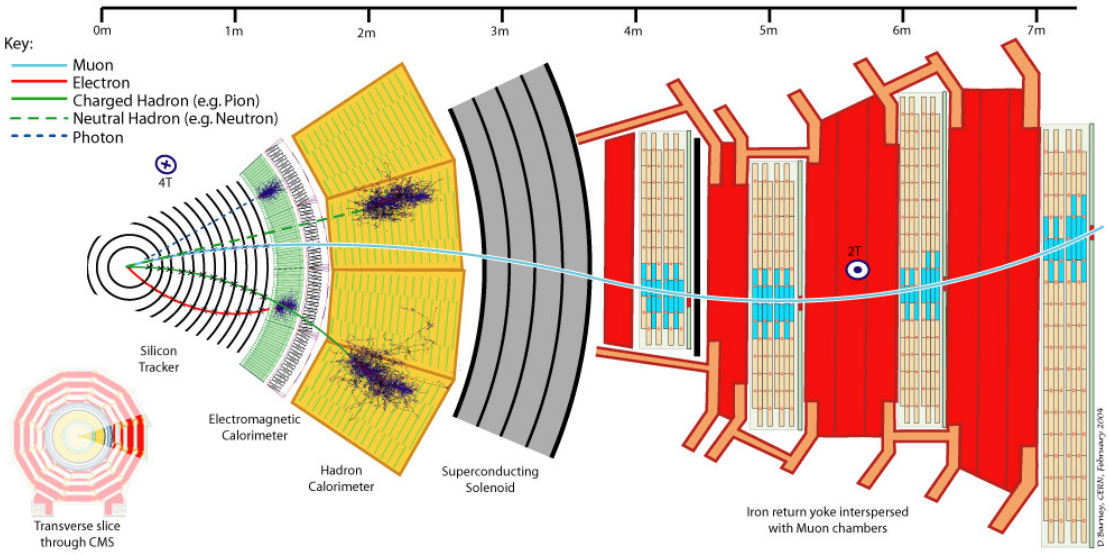

Figure 2.2: Transverse slice of the CMS detector with traces of muons, electrons, photons and hadrons traversing the subdetectors [22].

whether to pass the event to the second stage, the high level trigger (HLT), with a rate of about 100 kHz [23]. The HLT is a computing farm performing an initial reconstruction of the event including information from the tracking system, which allows to reject events based on detailed requirements for the observed interactions, finally reducing the rate of events written to the storage system to approximately 1 kHz [24].

Because of the shape of the CMS detector and the invariance of the decays in the transverse plane to the beams, the preferred coordinate system used to analyze the particle decays is cylindrical with the azimuth ϕ in the transverse plane to the colliding bunches and the z direction along the beam pipe. To specify a polar angle θ, typically the transformation called pseudorapidity is used, defined as

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

Further, commonly used is the metric ΔR as a measure of the distance between two objects, which is given by

$$\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}. \tag{2.2}$$

2.2 Production and decay of the Standard Model Higgs boson

The SM Higgs boson is produced in three main production processes. Feynman graphs of these processes are shown in figure 2.3. The cross section of the processes including

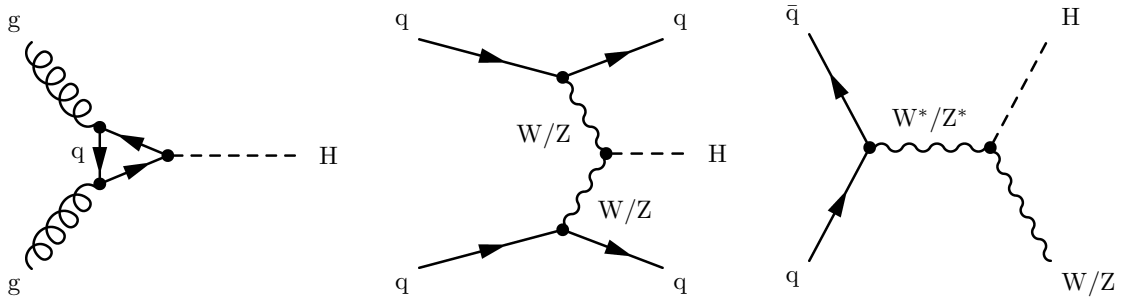

Figure 2.3: Feynman diagrams in leading order perturbation theory shown for the three main production processes of the SM Higgs boson, from left to right and with decreasing cross section: Higgs boson production from gluon fusion (ggH), Higgs boson production from vector boson fusion (qqH) and Higgs boson production associated with vector bosons (VH) [25].

a Higgs boson can be compared to well established SM processes such as the Z boson production in figure 2.4.

To understand the physical units of cross section and luminosity in context of data analysis and ML, the numbers have to be translated to counts. For bunches containing N particles each, an effective area A of the crossing and the frequency of the crossings f, the luminosity is in first approximation given by $L = fN^2A^{-1}$. It should be noted that the actual measurement and calibration of the luminosity at the CMS experiment is far more complex and precise, e.g., taking into account the exact beam shape measured with van der Meer scans [27]. The luminosity L can be interpreted as collisions per area and time with the unit $\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ and the integrated luminosity is given by $\int_t L dt$ with the unit cm^{-2}. In HEP, the integrated luminosity is often specified with the unit barn, which converts $10^{-28}\,\mathrm{cm}^{-2}$ into 1 b. The LHC reached during the Run 2 period a peak luminosity of about $2 \times 10^{-34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ and delivered an integrated luminosity of $162.85\,\mathrm{fb}^{-1}$ to the CMS experiment, which recorded $150.26\,\mathrm{fb}^{-1}$ for physics analysis (see figure 2.5).

Finally, to make the task to measure the cross section of the SM Higgs boson easily understandable in terms of data analysis, table 2.1 shows the number of recorded events available for analysis from the CMS experiment inclusively, for Z boson production and the three main Higgs boson production processes shown in figure 2.3. The rates of the processes are given as cross sections, typically with the unit barn, so that the integrated luminosity multiplied with the cross section of the physical process of interest results in the expected number of events in the dataset.

Because the analysis in this thesis studies the decay of the Higgs boson into two tau leptons, the numbers in table 2.1 have to be multiplied with the fraction to this decay mode to get an estimate for the expected Higgs events in the dataset. Table 2.2 shows the fractions of the decay modes of the SM Higgs boson, which are so far experimentally accessible at the LHC.

Finally, the decay products of the Higgs boson specified by the decay modes in table 2.2 decay in their final states eventually recorded by the CMS detector. For the $\tau\tau$ decay mode being subject of this thesis, the fractions of the final states are shown in table 2.3.

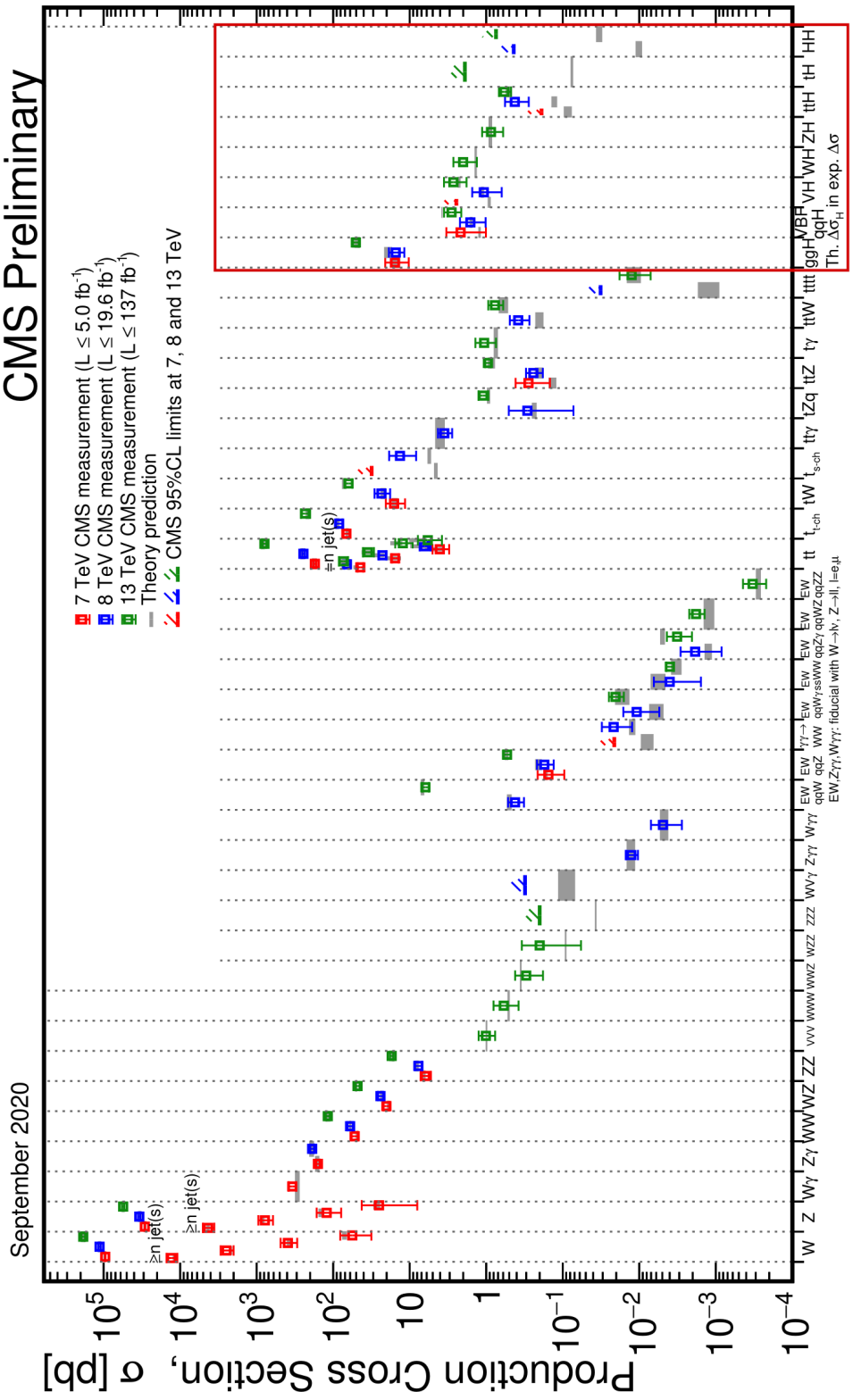

Figure 2.4: Summary of cross sections measured with the CMS detector at 7 TeV, 8 TeV and 13 TeV. The Higgs boson production processes are highlighted with the red box [26].

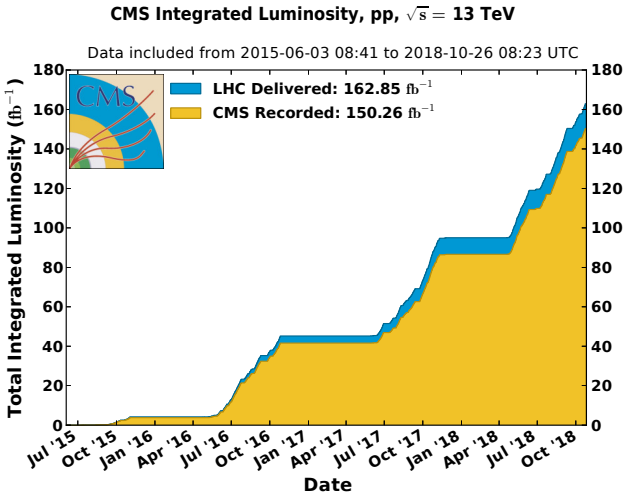

Figure 2.5: Integrated luminosity of proton proton collisions delivered by the LHC and recorded by the CMS experiment during the Run 2 period from end of 2015 to end of 2018 [1].

Table 2.1: Shown are the cross sections and expected number of events inclusively [28], for inclusive Z boson production [29], and the three main Higgs boson production processes ggH, qqH and VH [30]. The values are rounded to the significant digits and given for the full Run 2 dataset of the CMS experiment with 150 fb^{-1} at a center of mass energy of 13 TeV such as shown in figure 2.5.

Process	Total	Z	ggH	qqH	VH
Cross section	69 mb	19 nb	49 pb	2 pb	2 pb
Expected number of events for LHC Run 2	1×10^{16}	3×10^{9}	7×10^{6}	3×10^{5}	3×10^{5}

Table 2.2: Fractions of the decay modes for the SM Higgs boson, which are so far accessible at the LHC. The values are rounded to the significant digits [30].

Decay mode	bb	WW	$\tau\tau$	ZZ	$\gamma\gamma$
Fraction	0.58	0.22	0.063	0.026	0.002

Table 2.3: Fractions of the final states for the ditau decay. The symbol τ_h denotes the hadronic decay of the tau lepton. The values are rounded to the significant digits [31].

Final state	$\tau_\mathrm{h}\tau_\mathrm{h}$	$e\tau_\mathrm{h}$	$\mu\tau_\mathrm{h}$	$e\mu$	ee and $\mu\mu$
Fraction	0.42	0.23	0.23	0.06	0.06

2.3 Background processes and estimation methods

To measure the cross sections of the physical processes of interest, typically referred to as the signal processes, data analysis in HEP is driven by the precise estimation of all processes contributing to the data. This section describes the considered physical processes with similar final states than the signal processes, and referred to as the background processes, which are taken into account for the analysis described in this thesis. Figure 2.6 gives an example showing the invariant mass of the visible decay products of the ditau system as showcase for the physical processes and estimation methods described in the following.

$Z \rightarrow \tau\tau$ The major process for the production of lepton pairs at the LHC is the Drell Yan (DY) process [35]. Dominantly through the decay of a Z boson, a tau pair can be produced with the main difference to a $H \rightarrow \tau\tau$ decay being the invariant mass of the ditau system. Since tau leptons decay always in association with at least one neutrino, the possibilities to reconstruct the invariant mass of the ditau system are limited [36], leaving the $Z \rightarrow \tau\tau$ process as the most dominant and hardly reducible background process.

The process can be fully estimated from simulation taking into account up to four additional jets in the hard interaction. Also the tau embedding technique [32] can be used to estimate the events partially from data improving in particular the description of the jet related properties.

$Z \rightarrow ll$ Such as Z bosons can decay into tau pairs, also electron and muon pairs are the result from the DY process. Although the final state is not the same as for the $H \rightarrow \tau\tau$ decay, misidentifications of quark or gluon induced jets, electrons, or muons as hadronic tau decays can lead to such events ending up in the analysis selection.

The process can be simulated as part of the DY process including the Z decay to genuine taus or partially estimated with data driven techniques. The contribution of misidentified hadronic tau decays from jets can be derived from data using the FF method [33, 34].

$W + \mathbf{jets}$ W bosons are frequently produced in proton proton collision at the LHC. In association with an additional jet, the lepton from the W boson and the jet misidentified as hadronic tau decay build the same signature than a decay of a genuine tau pair.

A possible estimation technique is the simulation with up to four additional jets or the contribution can be fully described by the FF method.

$t\bar{t}$ Produced via gluon fusion and quark antiquark annihilation, top antitop pairs appear commonly in proton proton collisions at the LHC. Because the top quark decays almost exclusively via a W boson into a bottom quark, the ditop pair results in two W bosons, which have a variety of decay modes to produce the signature of a ditau pair. The two W bosons can either decay into a genuine tau pair or in a muon or electron and the

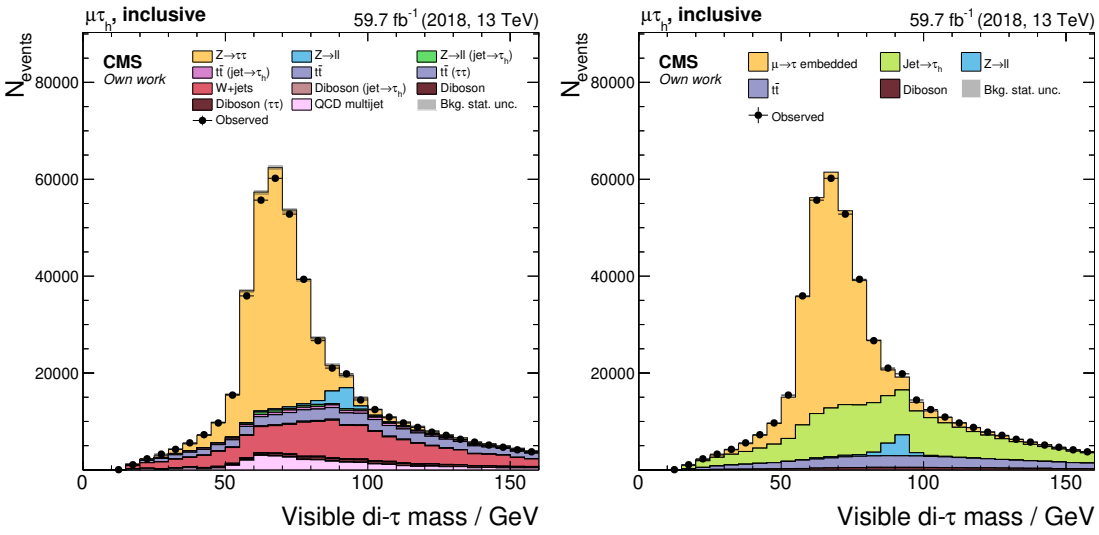

Figure 2.6: Invariant mass of the visible decay products of the ditau system in the $\mu\tau_\mathrm{h}$ channel with data taken in 2018 using background estimations from simulation except for the QCD multijet process (left) and with the data driven estimation techniques tau embedding [32] and the fake factor (FF) method [33, 34] (right).

same mechanisms than for $W+\text{jets}$ and $Z \to ll$ can result in misidentified hadronic tau decays.

Besides the full simulation of the process, the embedding and FF techniques can be used to estimate from data the genuine tau decays and the misidentifications from jets.

QCD multijet The QCD multijet background summarizes remaining processes with multiple jets in the final state leading to a misidentification of a hadronic tau decay. Because the simulation is challenging, a precise data analysis has to estimate the contribution from data. The FF method is suitable because the contribution originates from misidentified jets. Another approach is the estimation of the QCD multijet process from same signed ditau pairs in data and the subtraction of all known other processes, which is possible because the QCD multijet process is in first order independent from the charge of the reconstructed taus.

Minor background processes Minor background processes are diboson production and the production of single top quarks, both potentially decaying with similar signatures as the $H \to \tau\tau$ process. As explained above, the misidentfication of jets and the contribution of genuine taus can be absorbed into the embedding and the FF method. Otherwise, the contributions are estimated from simulation.

2.4 Event reconstruction and selection

An event at the CMS experiment is defined as the data of a single crossing of two bunches of protons, which contain for LHC Run 2 on average about 30 collisions. This section describes the necessary procedures to reconstruct from the detector signals the physical objects, e.g., electrons, taus or jets, which are studied to draw conclusions about the underlying physical processes. In a second step, the event selection reduces the reconstructed data to the relevant events for the analysis of the physical process of interest, in this analysis the decay of the SM Higgs boson.

2.4.1 Reconstruction

At CMS, the reconstruction of events is based on the particle flow (PF) algorithm [37] combining information from all subdetectors in a holistic approach to a consistent set of physical objects, namely electrons, photons, muons and neutral and charged hadrons. These objects are further processed by dedicated algorithms to reconstruct composed objects such as taus, jets and missing energy. The vertices of the collisions are determined from the reconstructed tracks of the particles and the primary vertex is defined as the vertex with the largest sum of the transverse momentum of all contributing physical objects. All other collisions are called pileup.

Jets Jets are clustered with the anti-k_t algorithm from the PF objects [38]. Further, the jets are classified with the DeepJet approach based on deep convolutional and recurrent NN architectures to determine the originating particle, which can be a light quark (up, down and strange), a heavy quark (charm and bottom) or a gluon [39]. Jet tagging specialized on discriminating jets from bottom quarks is in particular useful for this analysis to identify events from the $t\bar{t}$ process.

Electrons The PF algorithm builds electrons with information retrieved from the ECAL and the tracking system. To suppress misidentified electrons, a ML based identification algorithm is applied using boosted decision trees [14]. The approach uses information from the reconstruction such as the track quality or the structure of the energy deposits in the ECAL to reject particles such as photons and charged hadrons, which are falsely reconstructed as electrons.

Muons Muons are reconstructed similar to electrons with a combination of information from the tracking system and the muon subdetector. Because only muons pass the ECAL and HCAL, the identification of genuine muons is very high with a rate above 99 % [40].

Taus Tau leptons require similar to jets an additional reconstruction on top of final state objects from the PF algorithm. At the CMS experiment, the reconstruction algorithm is called hadron plus strips (HPS) [41] where the strips refer to the clustered photons and secondary electrons from π^0 decays. The strips are combined with the reconstructed charged hadrons to identify the decay mode of the hadronically decayed tau. The leptonic

Table 2.4: Fractions of the leptonic and hadronic decay modes of the tau lepton rounded to the significant digits. The symbol $h^{\pm}$ denotes a charged hadron [31, 41].

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

and hadronic decay modes and the respective fractions are summarized in table 2.4. Taus are naturally difficult to distinguish from quark or gluon induced jets but also misclassified from electrons and muons. Therefore a ML based multiclassification algorithm is designed using a deep convolutional NN architecture [42].

2.4.2 Selection

Such as visible from table 2.1, the dataset of the CMS experiment has an enormous size, also because the experiment serves many analyses and not only Higgs physics. For these reasons, a preselection of relevant events is performed, enriching the dataset with the processes of interest and reducing the dataset size significantly. Because the description of the full analysis carried out on the three data taking periods 2016, 2017 and 2018 of LHC Run 2 in the considered final states $e\tau_\mathrm{h}$, $\mu\tau_\mathrm{h}$, $\tau_\mathrm{h}\tau_\mathrm{h}$ and $e\mu$ does not serve the focus of this thesis, the following section is reduced compared to the full documentation given in [43, 44]. An overview is given for the event selection of the data taking period of 2018 and the $\mu\tau_\mathrm{h}$ final state, demonstrating the necessary procedures and the complexity for such a measurement, even before performing any statistical analysis with the data.

The selection is performed on multiple levels. Events are globally selected by requiring specific triggers fired at the time of data taking. Further, the event content is filtered with selections on the reconstructed objects, like muons, taus and jets, before the objects are identified which most likely result from the process of interest, here the decay of the Higgs boson. These objects are used to reconstruct the full decay system and a final selection is performed before the events enter the statistical inference, defining the dataset for the specific decay channel of the Higgs boson.

Trigger Triggers are selected from the available set of triggers defined by the CMS collaboration. The selection is performed based on the trigger path, which defines a specific setting of fired triggers on the first trigger stage, additional filters and a valid

object in the HLT reconstruction with additional requirements. In the $\mu\tau_h$ final state of this analysis, the selections correspond, in first order, to events triggered by an isolated muon with $p_T > 24\,\text{GeV}$ or events with an isolated muon with $p_T > 20\,\text{GeV}$ and a hadronically decayed tau with $p_T > 27\,\text{GeV}$ in $|\eta| < 2.1$. In addition, the objects that caused the event to be recorded are required to match the respective objects of the offline reconstructed Higgs boson decay in $\Delta R < 0.5$.

Muons The muons available for analysis are cleaned by removing muons based on a set of filters defined by the medium working point of the muon identification algorithm [45]. The requirements are based on the compatibility of the information from the tracking system to the activated segments in the muon detector, the number of valid hits in the tracker and the fit quality of the reconstructed track. Also the track of the muon has to match the primary vertex within 4.5 mm in the transverse direction and 20 mm in longitudinal direction. Finally, the muon is required to be isolated from photons and neutral and charged hadrons in a cone of $\Delta R < 0.4$, which is enforced by a threshold on the sum of the additional transverse energy by these particles relative to the muon p_T.

Taus The available taus reconstructed by the HPS algorithm are primarily cleaned by the multivariate tau identification. The analysis uses for the classifier in the $\mu\tau_h$ channel a tight working point for the discrimination against jets with about 40 % acceptance rate and close to 0.1 % misidentification rate. The working points chosen for the separation against electrons and muons perform with an acceptance rate of close to 99 % at a misidentification rate of about 1 % and 0.01 %, respectively [42]. In addition, the tau is required to have $p_T > 30\,\text{GeV}$ and a distance of the track to the primary vertex below 20 mm in longitudinal direction.

Jets Jets are selected for the analysis if they fullfill the requirements $p_T > 30\,\text{GeV}$ and $|\eta| < 4.7$. In case the jet is identified by the multivariate jet tagger being the result of a bottom quark, the kinematic requirements are set to $p_T > 20\,\text{GeV}$ and $|\eta| < 2.4$. Jets used in the analysis have to be separated from the muon and tau of the selected pair by $\Delta R > 0.5$.

Pair building and final selection A valid pair of a muon and a hadronically decayed tau is selected whereas in case of ambiguities pairs are preferred with better isolation and higher transverse momentum. The muon and tau of the pair are required to have opposite charge and being separated by $\Delta R > 0.5$. Further, the transverse mass of the muon plus the missing transverse energy have to be below 70 GeV to be independent of the data driven FF method. Finally, the best pair is selected for further analysis if no additional unmatched isolated electrons or muons are present in the event.

2.5 Statistical inference

The statistical inference in this analysis compares the observation of the CMS experiment with the expectation from the SM through selected observables resulting in the measurement of the parameters of interest (POIs). The observables can be physical quantities such as the invariant mass of the reconstructed Higgs boson but also abstract variables like the output from a NN function. Because the procedures for the statistical inference are independent from the chosen observables, the discussion about the selection of the observables is continued in section 2.6. This section describes the statistical model and methods used to determine the cross section of the Higgs boson from the data of the CMS experiment in consideration of all known uncertainties.

2.5.1 Analysis objective

From the physics point of view the objective of the analysis is the cross section measurement of the Higgs boson decaying into two tau leptons. The concept of a cross section is translated to the statistical framework as the parameter μ called the signal strength modifier, which scales the expected events for the SM $H \to \tau\tau$ process linearly. In consequence, the subject of the statistical inference is the parameter estimation of the POI μ given as the bestfit value $\hat{\mu}$ and the according uncertainty σ_μ in units of standard deviations. Previous measurements of the signal strength were $\mu = 0.78 \pm 0.27$ for the evidence of $H \to \tau\tau$ at the CMS experiment in 2014 [46] and $\mu = 1.09 \pm 0.27$ for the discovery in 2018 [47].

For current analyses and measurements in the future, differential measurements become experimentally accessible to probe the SM with higher granularity. The CMS collaboration follows a framework recommended by the simplified template cross section (STXS) working group [48, 49] in collaboration with the ATLAS experiment and theorists. The collaboration has the purpose to specify phase spaces of the SM Higgs boson kinematic properties in which a cross section measurement is testing the SM with minimal uncertainties on the theory to maximize the sensitivity to the underlying physics. A further advantage is the coordination of different analyses supporting a global combination of differential measurements across experiments and final states.

In stage 0 of the STXS framework, the cross section measurement is split by the production processes of the Higgs boson of which this analysis is sensitive to ggH, qqH and VH. It should be noted that the measurement qqH also contains the events from VH in which the vector boson decays hadronically. The statistical framework evolves accordingly by replacing the inclusive POI μ with three dedicated parameters and the statistical inference is performed with a multidimensional fit estimating all POIs concurrently. In the next stage of the framework, STXS stage 1, the measurements are split by jet multiplicity and the kinematic properties of the Higgs boson. Figure 2.7 gives an overview over the subcategories for ggH and qqH. If the analyses are not sensitive to all categories in STXS stage 1, the respective POIs can be merged by measuring the parameters combined into a single POI.

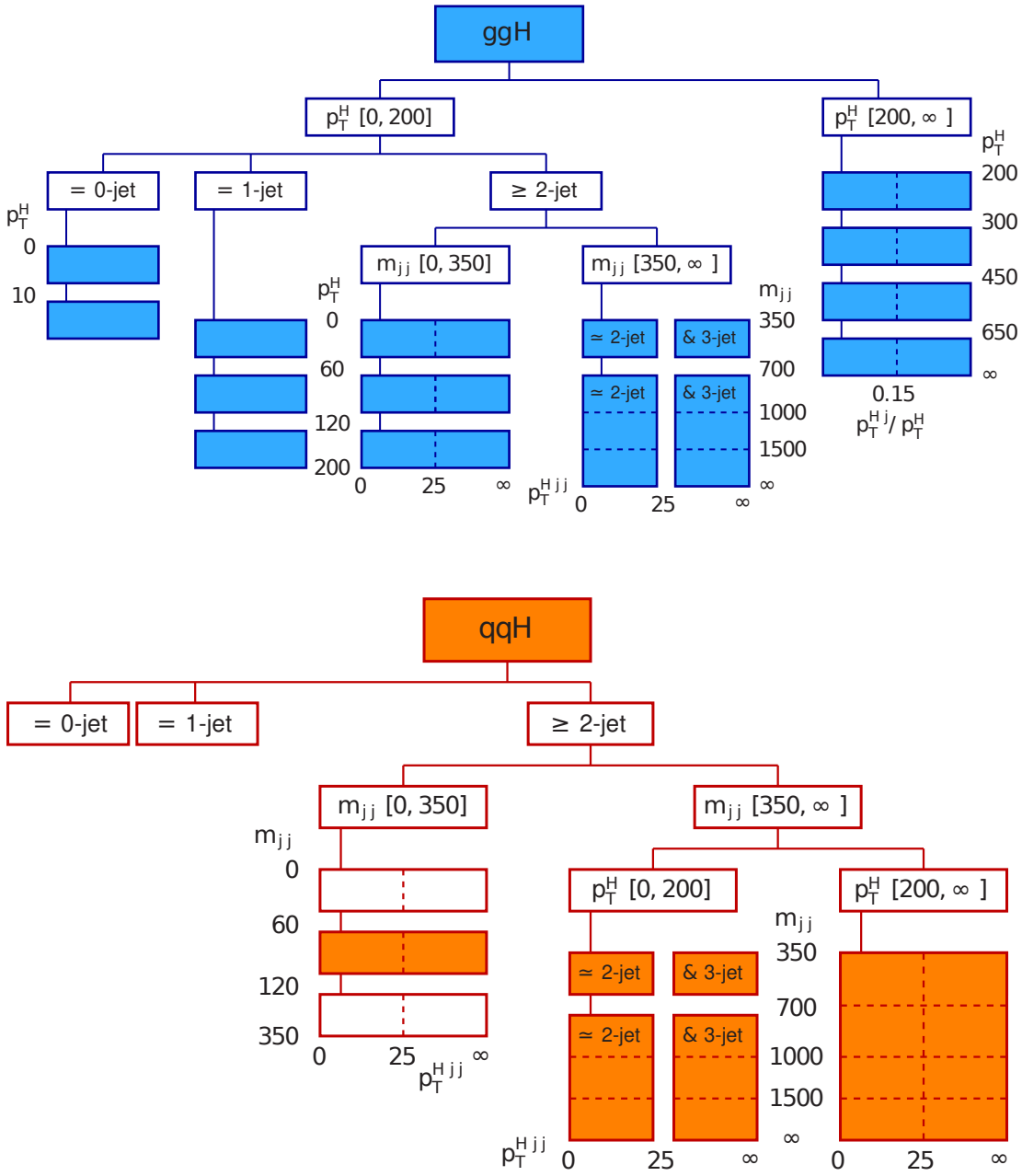

Figure 2.7: Stage 1 of the STXS framework for ggH and qqH [48, 49]

2.5.2 Statistical model

Most analyses published by the CMS collaboration are based on the concept of counting experiments. The reason is that the probability density $p(\boldsymbol{x}|\boldsymbol{\theta})$ of the observables $\boldsymbol{x}$ given the parameters of the statistical model $\boldsymbol{\theta}$ is not analytically known. Analysis in HEP typically circumvents this problem with a full simulation of the experiment, from the particle collision to the detector response, eventually sampling the distribution $p(\boldsymbol{x}|\boldsymbol{\theta})$. Because the simulation is computationally expensive and the number of simulated events required to sample the distribution in a high dimensional space spanned by $\boldsymbol{x}$ grows exponentially, the statistical model is built on a summary of $\boldsymbol{x}$ to reduce the dimensionality of the problem. A suitable summary is a count with the observation k, the expectation λ and the Poisson distribution

$$\mathcal{P}(k|\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}. \tag{2.3}$$

In this space, the probability density is accessible because the expectation for the counts is given by the simulation. For a single signal s and background b the expectation is formulated as $\mu s + b$ introducing the POI in the statistical model. Using Poisson statistics, the likelihood function for a histogram with h bins is

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{h} \mathcal{P}(d_i|\mu s_i + b_i) \tag{2.4}$$

with the parameters $\boldsymbol{\theta} = \{\mu\}$, the observed counts $\boldsymbol{d}$ and the expected counts of the signal and background process $\boldsymbol{s}$ and $\boldsymbol{b}$, respectively. The statistical model expands naturally to multiple signal and background processes by adding additional terms and POIs to the expectation of the Poisson distribution.

2.5.3 Systematic uncertainties

Finding the bestfit values for the POIs is not the actual complication for the statistical inference in HEP but the challenge is the precise determination of the according uncertainties. Uncertainties are split into two groups, statistical and systematic uncertainties. The statistical uncertainties are naturally included in the likelihood function via the Poisson distribution while systematic uncertainties require an extension of the statistical model.

The statistical approach to include systematic uncertainties in the estimation of the POIs is including additional parameters to $\boldsymbol{\theta}$, namely the nuisance parameters (NPs). The concept is that the NPs can compensate the expectation for the signals, resulting in an estimation of the POIs with an increased variance. Consequently, the physics measurement has an increased uncertainty. A possible implementation of a systematic uncertainty on the expectation of $\boldsymbol{b}$ is

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{h} \mathcal{P}(d_i|\mu s_i + \eta b_i) \cdot \mathcal{C}(\eta) \tag{2.5}$$

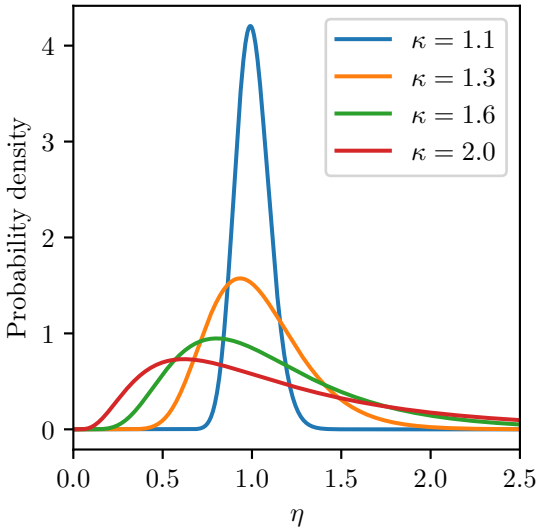

Figure 2.8: Lognormal distribution for various values of κ

with a given NP η and the constraint term $\mathcal{C}$. In contrast to POIs, NPs are typically constraint parameters of the statistical model and the apriori knowledge about the probability that the NP takes a specific value in the bestfit configuration of $\boldsymbol{\theta} = \{\mu, \eta\}$ is specified by the probability density $\mathcal{C}(\eta)$. Because the statistical inference of a typical analysis in HEP is carried out with multiple hundreds of NPs, the probability density is not specifically implemented for each NP but two groups of systematic uncertainties are defined.

Similar to the example above, normalization uncertainties scale a process with a factor η following the lognormal distribution

$$\mathcal{C}_{\ln\mathrm{N}}(\eta, \kappa) = \frac{1}{\sqrt{2\pi}\eta \log \kappa} \exp\left(-\frac{1}{2}\left(\frac{\log \eta}{\log \kappa}\right)^2\right) \tag{2.6}$$

parametrized by κ representing the uncertainty on the yield. Figure 2.8 shows the distribution for various values of κ. This constraint is suitable for physics analysis because the lognormal distribution is bound to positive values and therefore prevents unphysical values for the NP which would result in negative counts [50].

More complex systematic uncertainties that scale not only the yield of a process inclusively but may have an individual impact on each bin of the likelihood are called shape uncertainties. The distribution of the NP is parametrized given three values per bin, the expectations of the process at the nominal value I_0 and expectations of the upshifted and downshifted values $I_\pm$ due to the systematic uncertainty. The shifted values represent the instantiations of the systematic uncertainty with the probabilities of $\pm 1\sigma$ Gaussian standard deviations and all other values are interpolated between these and the nominal value. This concept is implemented in the likelihood using a normal distribution for the constraint term $\mathcal{C}(\eta)$ and a transformation of the NP for the modification of the

expectation, for example for the background process given by

$$b_i \to I(I_0, I_+, I_-, \eta) \cdot b_i. \tag{2.7}$$

The interpolation can be implemented to be linear, however, a piecewise linear function has a discontinuous first derivative causing difficulties with the minimization of the likelihood based on gradient descent. Taking the requirements of a continuous first and also second derivative into account the interpolation rule

$$I(I_0, I_+, I_-, \eta) = \begin{cases} (I_+/I_0)^\eta & \eta > 1 \\ 1 + \sum_{i=1}^{6} a_i \eta^i & |\eta| < 1 \\ (I_-/I_0)^\eta & \eta < -1 \end{cases} \tag{2.8}$$

has proven being suitable [51]. It should be noted that for a normal distributed NP η the extrapolations with $|\eta| > 1$ are again lognormal distributed because of $(I_\pm/I_0)^\eta = e^{\eta \log(I_\pm/I_0)}$. Figure 2.9 shows the interpolation for different configurations of nominal and shifted values.

2.5.4 Parameter estimation

The previous sections discussed how in such an analysis the statistical model is built. Consequently, this section covers the procedure to infer from the model estimates for the POIs including an uncertainty, or using a terminology more common to statistics, a confidence interval.

The parameter estimation finds the set of best fitting values for all parameters of the statistical model $\hat{\boldsymbol{\theta}}$ with respect to the data in a global minimization of the negative logarithmic likelihood (NLL) $-\log \mathcal{L}(\boldsymbol{\theta})$. Typically the NLL is minimized because many multiplications of probabilities are numerically unstable and since the logarithm is a monotonic function the minimization of the NLL results in the same values for $\hat{\boldsymbol{\theta}}$. Using the likelihood as objective function is sensible because the estimator is consistent and efficient, which means that the estimate $\hat{\boldsymbol{\theta}}$ converges with increasing sample size to the truth values with a minimal variance.

More complex than measuring the best fit values of the POIs is the construction of an according confidence interval. In the Neyman construction [52], the confidence interval is determined from the inversion of an hypothesis test by evaluating the compatibility of the best fit value of the POI, for example $\hat{\mu}$, to other possible values μ. Such as known from Neyman and Pearson [53], the most powerful statistic for such an hypothesis test is the likelihood ratio

$$\lambda = \frac{\mathcal{L}(\mu, \hat{\boldsymbol{\theta}}_\mu)}{\mathcal{L}(\hat{\boldsymbol{\theta}})} \tag{2.9}$$

with $\hat{\boldsymbol{\theta}}_\mu$ being the best fit values of the parameters given a fix value for μ. The upper and lower bounds μ_{up} and μ_{down} correspond to the values of μ for which the distribution of λ has the respective bound at $\lambda(\hat{\mu})$ for the interval of interest, for example a central

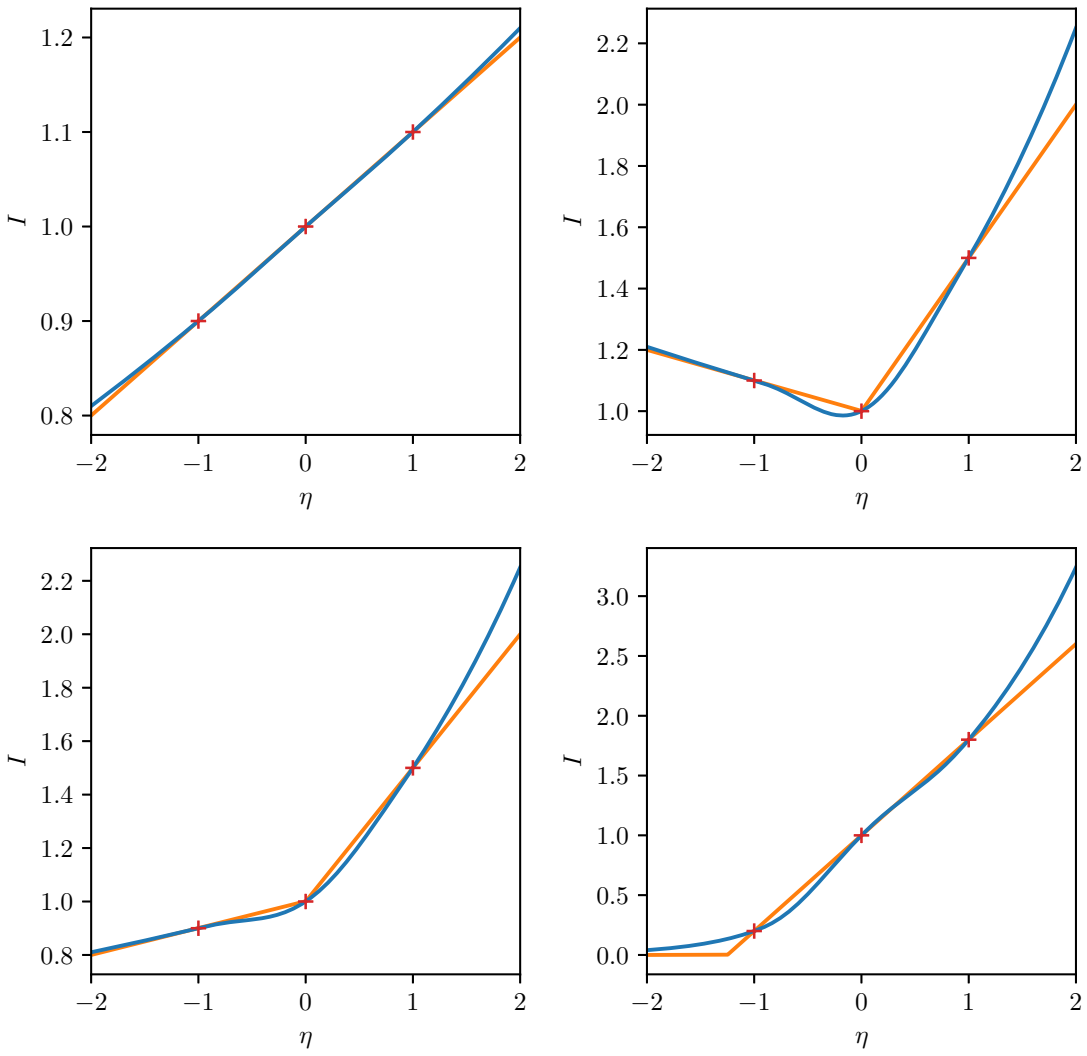

Figure 2.9: Examples for the interpolation between nominal, upshifted and downshifted values (red crosses) used for the modeling of shape uncertainties. Shown is a comparison of the polynomial interpolation (blue lines) from equation 2.8 compared to a piecewise linear extrapolation (orange lines).

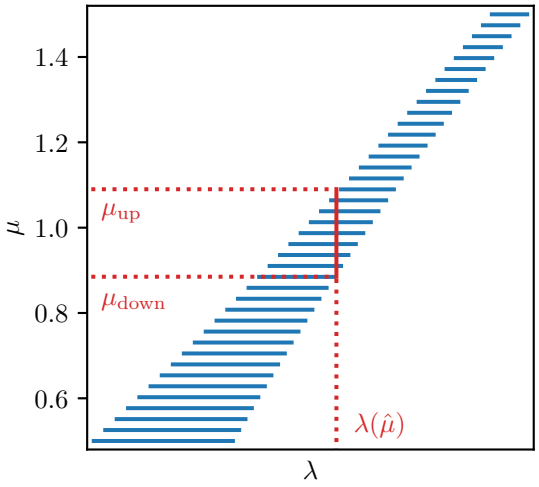

Figure 2.10: An exemplary Neyman construction such as used to find the corresponding confidence interval for the best fit value $\hat{\mu}$. The blue solid lines represent the interval of interest for the distribution of the likelihood ratio λ with a fix value for μ. The confidence interval is determined from the values for μ at which the measurement $\lambda(\hat{\mu})$ is at the upper or lower bound of the interval, respectively.

interval covering a range equal to $1\,\sigma$ in standard deviations. Figure 2.10 visualizes the Neyman construction with an example.

Because the Neyman construction is computationally expensive, the confidence intervals are typically approximated using the theorems of Wilks [54] and Wald [55], which proof that the distribution of the test statistic $-2\log\lambda$ approaches with a suitably large sample size and under the null hypothesis a χ^2_k distribution with k degrees of freedom. k is given by the difference in number of free parameters of the null hypothesis and the alternative hypothesis, and hence for the example with the single POI μ the test statistic follows a χ^2_1 distribution, which is equal to a squared normal distributed random variable. Consequently, the central confidence interval covering a range equal to $n\,\sigma$ in standard deviations is given by the two solutions for μ of the equation $-2\log\lambda(\mu) = n^2$. It should be noted that if the requirements for the asymptotic case are fullfilled, the confidence interval is symmetric because the profile of the test statistic $-2\log\lambda$ is a parabola. Also in the case of non parabolic NLL functions, such as often present in HEP analyses, it can be shown that because of invariance properties of the likelihood ratio λ the approach is still valid [56, 57].

To prevent a bias of the analysis, for example by reoptimizing bin edges or the choice of the observables, the development is done blind to the data. This is possible due to the usage of an Asimov dataset [58], which replaces the data with the nominal expectation. In the case with the single POI μ shown in equation 2.5, the Asimov dataset is $\boldsymbol{d} = \boldsymbol{s} + \boldsymbol{b}$ with $\mu = 1$ representing the expectation from the SM.

2.6 Observables and event categorization based on multiclass neural networks

The analysis leverages multivariate analysis methods with the usage of multiclass classifiers implemented with NNs. This section highlights the technical details and peculiarities of the analysis strategy in the context of ML in contrast to [44], which focuses on the physics results. An all inclusive technical documentation of the analysis is provided by [43].

2.6.1 Overview

The NNs are set up as multiclass classifiers given as inputs information about the event, the reconstructed Higgs boson system and additional jets. The input variables are summarized in table 2.5. The outputs of the NNs are trained to predict the probability that the event belongs to a specific class, which are designed to support the separation between the dominant background processes and the signal processes of interest. With the target to measure the differential cross sections in the STXS stage 1 framework, a single class for each signal is the baseline. Because not all of these signals are experimentally accessible, related and non distinguishable signals are merged for the NN training to reduce the complexity of the task. In addition, each dominant background process is assigned a dedicated class and eventually all remaining background processes are merged into a miscellaneous class. A summary of the event classes is shown in table 2.6.

Assuming a successful training of the multiclass NN, the predictions cannot be put directly into the statistical model such as described in section 2.5. The first issue is that the output of the NN is a multidimensional observable and since high dimensional spaces are only sparsely populated, a counting experiment is not possible. Second, the statistical model allows to use each event only once, which would otherwise result in double counting of observations and consequently falsifying the measurement. Taking these restrictions into consideration, a suitable summary of the NN output is the usage of the largest probability per event as observable. This transformation categorizes the events and reduces the dimensionality $\mathbb{R}^N$ of the NN output to N scalar observables, each dedicated to a separate event category. Double counting is naturally prevented because the largest probability is defined only once per event. Figure 2.11 provides a graphical overview over the analysis strategy from the input variables of the NN to the final observables entering the statistical inference.

2.6.2 Neural network architecture and training procedure

The used NN architecture is a fully connected feed forward network with two hidden layers and 200 nodes each [60]. The activation of the hidden layers is a hyperbolic tangent and a softmax function is used for the output layer. The weights of the layers are initialized with the Glorot algorithm [61]. For regularization during the training, dropout [62] with a probability of 0.3 and an L2 regularization term [63] with a factor of 10^{-5} are applied.

Table 2.5: Summary of the event properties used as input for the multiclass NNs. All variables are used in all four decay channels unless stated otherwise.

Identifier	Description
`pt_1`	p_T of the first object in the pair, e.g., μ in $\mu\tau_\mathrm{h}$
`pt_2`	p_T of the second object in the pair, e.g., τ_h in $\mu\tau_\mathrm{h}$
`mTdileptonMET`	Transverse mass of the dilepton pair including missing energy, only used in $e\mu$
`jpt_1`	p_T of the leading jet with the highest p_T
`jpt_2`	p_T of the subleading jet with the second highest p_T
`njets`	Number of jets
`nbtag`	Number of jets originating from a bottom quark, not used in $e\mu$
`mjj`	Invariant mass of the leading and subleading jet
`jdeta`	Difference in η between the leading and subleading jet
`dijetpt`	p_T of the dijet system built from leading and subleading jet
`m_sv`	Invariant mass of the Higgs boson system reconstructed with the SVFIT algorithm [36]
`m_vis`	Visible mass of the Higgs boson system
`pt_vis`	Visible p_T of the Higgs boson system
`DiTauDeltaR`	ΔR between the tau leptons of the Higgs boson system
`ME_q2v1`	MELA variable [59]
`ME_q2v2`	MELA variable

The loss function, in detail discussed in section 2.6.4, is optimized with respect to the trainable parameters using the Adam optimizer [64] with a learning rate of 10^{-4}.

Because the analysis has the target to maximize the sensitivity of the measurement, all events should be used for the statistical inference, which is ensured by a two fold approach. The overall dataset, including data and simulated events, is separated in two equal halves. On each half a separate training and validation is performed using 75 % of the events for the optimization of the model and the rest for the monitoring of the training progress. The training is stopped if the loss does not improve on the validation split for 50×10^3 consecutive gradient steps eventually selecting the model with the lowest validation loss for the testing and application on the other half of the dataset. The applied procedure to calculate efficiently the gradients from the loss function is discussed in section 2.6.4. Because each of the three data taking periods comes with a dedicated set of simulated events due to changing detector conditions and each decay channel requires a separate training due to the different event topologies of the final states, the two fold approach results in 24 separate NNs and trainings for the full analysis. Section 2.6.5 presents a solution for an unified NN training using conditional input variables to reduce the complexity of the analysis.

Table 2.6: Summary of the event classes used for the training of the NNs. The symbol $p_{\mathrm{T}}^{\mathrm{H}}$ denotes the transverse momentum of the Higgs boson, N_{jets} the number of jets and m_{jj} the invariant mass of the dijet system built from the leading and subleading jets. The signal classes are the same in all decay channels and the background classes are adapted to the dominant background processes.

Signal classes			
Process	N_{jets}	$p_{\mathrm{T}}^{\mathrm{H}}$	m_{jj}
ggH	≥ 0	$[200\,\mathrm{GeV}, 300\,\mathrm{GeV}]$	-
ggH	≥ 0	$> 200\,\mathrm{GeV}$	-
ggH	0	$< 10\,\mathrm{GeV}$	-
ggH	0	$[10\,\mathrm{GeV}, 200\,\mathrm{GeV}]$	-
ggH	1	$< 60\,\mathrm{GeV}$	-
ggH	1	$[60\,\mathrm{GeV}, 120\,\mathrm{GeV}]$	-
ggH	1	$[120\,\mathrm{GeV}, 200\,\mathrm{GeV}]$	-
ggH	≥ 2	$< 60\,\mathrm{GeV}$	$< 350\,\mathrm{GeV}$
ggH	≥ 2	$[60\,\mathrm{GeV}, 120\,\mathrm{GeV}]$	$< 350\,\mathrm{GeV}$
ggH	≥ 2	$[120\,\mathrm{GeV}, 200\,\mathrm{GeV}]$	$< 350\,\mathrm{GeV}$
ggH	≥ 2	$< 200\,\mathrm{GeV}$	$> 350\,\mathrm{GeV}$
qqH	≥ 2	$< 200\,\mathrm{GeV}$	$[350\,\mathrm{GeV}, 700\,\mathrm{GeV}]$
qqH	≥ 2	$< 200\,\mathrm{GeV}$	$> 700\,\mathrm{GeV}$
qqH	≥ 2	-	$< 350\,\mathrm{GeV}$
qqH	≥ 2	$> 200\,\mathrm{GeV}$	$> 350\,\mathrm{GeV}$

Background classes	
Process	Decay channels
$Z \to \tau\tau$	$e\mu$, $e\tau_{\mathrm{h}}$, $\mu\tau_{\mathrm{h}}$, $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$
$Z \to ll$	$e\tau_{\mathrm{h}}$, $\mu\tau_{\mathrm{h}}$
$t\bar{t}$	$e\mu$, $e\tau_{\mathrm{h}}$, $\mu\tau_{\mathrm{h}}$
Misidentifications from jet $\to \tau_{\mathrm{h}}$	$e\mu$, $e\tau_{\mathrm{h}}$, $\mu\tau_{\mathrm{h}}$, $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$
Diboson	$e\mu$
Miscellaneous	$e\mu$, $e\tau_{\mathrm{h}}$, $\mu\tau_{\mathrm{h}}$, $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$

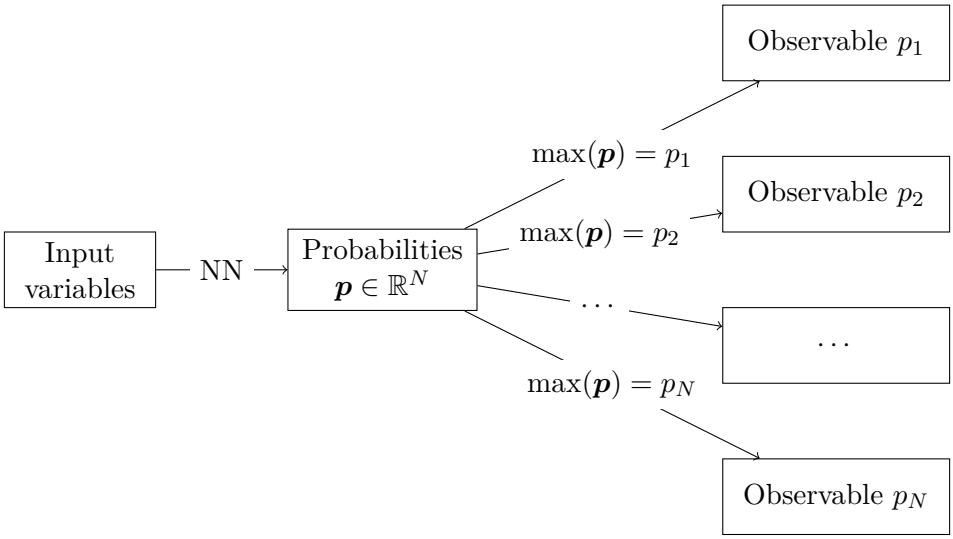

Figure 2.11: Overview over the transformation of the input variables via the multiclass NN with N classes to the final categories and observables entering the statistical inference

2.6.3 Performance evaluation

Before the trained NN is applied to the data, the success of the training is evaluated on the respective second half of the training dataset. It should be noted that a single metric to quantify the quality of the training is not strictly defined for a multiclass classification task. Further, the analysis task, namely the differential measurement of the $H \to \tau\tau$ cross section, is not fully congruent with the training objective and is also a multidimensional construct without a well defined scalar metric to measure the NN performance. The congruency between the training objective and the analysis objective is discussed in chapter 5 where a novel solution for a holistic ML based analysis strategy is presented.

The chosen metric to validate the success of the NN training is the confusion matrix. The matrix compares for each event class the fraction of the events being identified or misidentfied by the NN. Summarized are the event weights to adjust for the importance of simulated events and eventually the sum of all event weights are normalized per class reflecting the training objective, which treats each event class with equal importance. Such a confusion matrix can be normalized again for each predicted or true event class, yielding the purity and efficiency of the respective classes. An examplary efficiency representation of the confusion matrix is shown in figure 2.12.

2.6.4 Model optimization with event weights and imbalanced datasets

The training uses the categorical cross entropy (CE) as loss function

$$\mathrm{CE} = -\sum_{i=1}^{B} w^{(i)} \sum_{j=1}^{N} y_j^{(i)} \log(p_j^{(i)}) \tag{2.10}$$

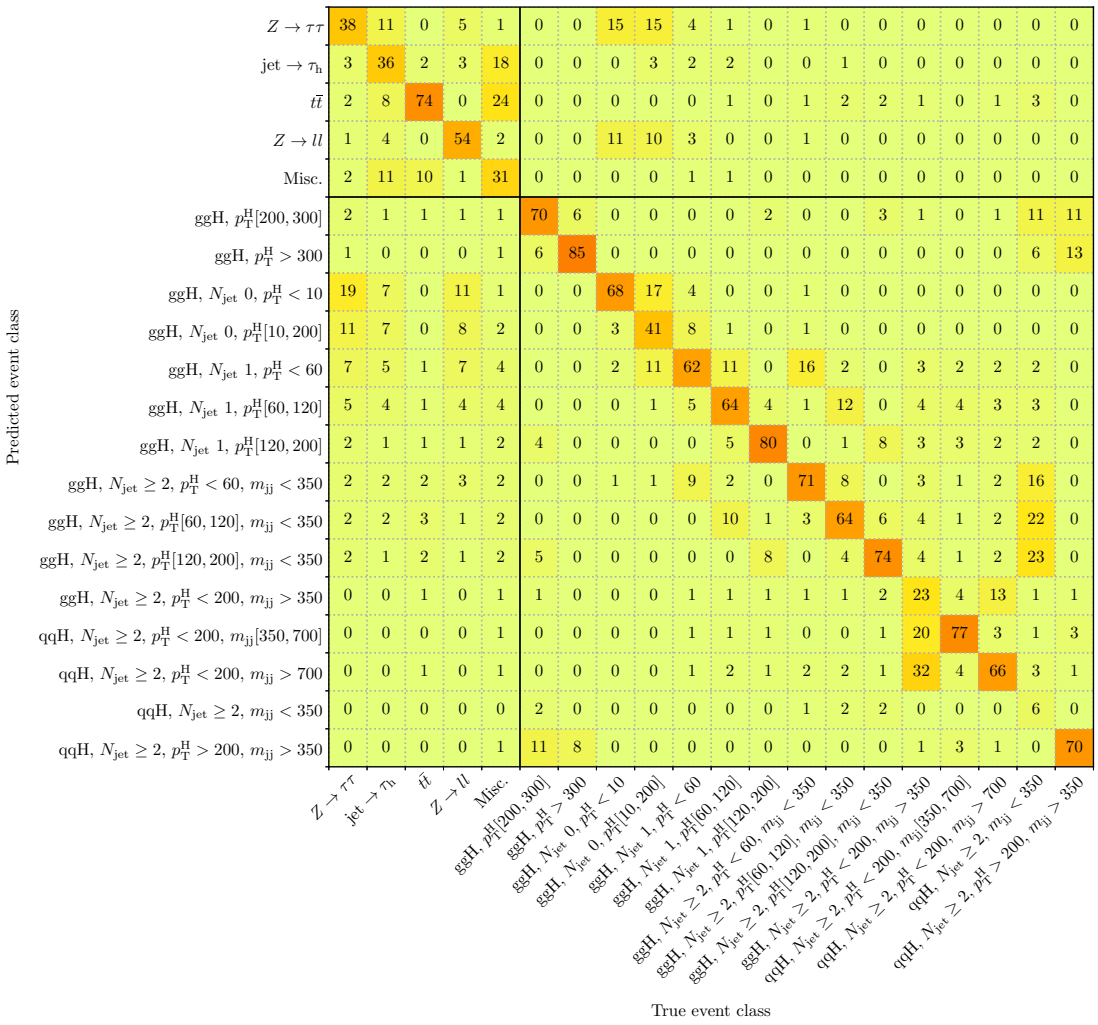

Figure 2.12: The confusion matrix is normalized for each true event class yielding the efficiency representation, which shows the respective values in percent. The matrix is evaluated for the $\mu\tau_h$ channel and the 2016 data taking period. The solid black line separates the signal and background classes.

with the number of events B, the number of classes N, the predictions of the NN $\boldsymbol{p}$ and the vector $\boldsymbol{y}$ with a one for the element corresponding to the true class and zeros otherwise. The term w is used to encode the importance for the classification task for each event contributing to the loss. The optimization of the CE can be understood as the minimization of the NLL $-\log\left(\prod_i p(\boldsymbol{x}_i|y_i)\right)$ with $\boldsymbol{x}$ being the NN inputs yielding a strong motivation for this training objective in classification tasks. Chapter 5 discusses in detail the suitability of the CE loss with respect to the analysis objective and alternative solutions are proposed.

Traditionally, each gradient step is computed from a batch of B events randomly drawn from the training dataset. Neglecting the weight term w, the importance of each event class is then given by the frequency of respective events in the trainig dataset. Because in HEP the simulated events are typically oversampled compared to the expectation in data, for example the simulation of the signals, the frequency of a process in the training dataset does not reflect the importance of the class. Therefore, the frequency of simulated events has to be scaled down by a statistical weight to match the distribution in the data. Further, the simulation is subject to corrections of the variable distributions being implemented with statistical weights, for example in this analysis the reweighting of the transverse momentum of the Z boson. These event weights have to be taken into account in the NN training to model the classification task correctly. An exemplary distribution of the event weights is presented in figure 2.13 showing that for a random sampling of the batch the highly imbalanced training dataset with respect to the number of events per class would result in an imprecise gradient for the less frequent processes, especially the signal classes. Therefore, the batch is built by selecting from each class randomly the same number of events providing a balanced precision for the optimization across all classes. For the training of the NNs used in this analysis, 30 events per class are chosen. In addition, the weights are normalized per class maintaining the relation between the events but encoding an equal importance for each class in the NN training. The impact of a balanced selection of events for the gradient computation is discussed in detail in [65].

2.6.5 Transfer learning with data driven training and conditional inputs

Analysis in HEP deals with information from a variety of datasets which are not strictly following the same underlying distribution. The two major domains are the data from the experiment and the simulated events. But at the same time the detector conditions change over time resulting in a varying ground truth of the data distribution, which is covered by dedicated simulations for the different data taking periods. This section presents the applied strategies to incorporate this information into the NN training.

To minimize the difference between simulation and data for the $Z \to \tau\tau$ process, the training uses only partially simulated events. The embedding technique takes $Z \to \mu\mu$ events from data, removes the detector hits from the muons and inserts instead the simulation of two decaying taus [32]. This technique yields an improved description of the $Z \to \tau\tau$ process because complex objects in the event such as additional jets are incorporated directly from data, successfully minimizing the required transfer learning

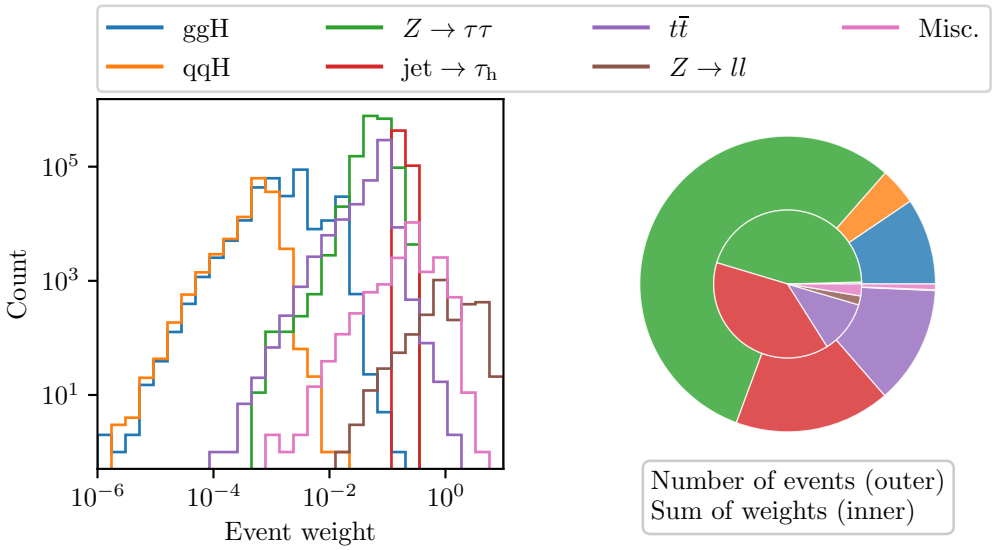

Figure 2.13: Distribution of the event weights in the training dataset for the $\mu\tau_h$ channel from the 2016 data taking period. $Z \to \tau\tau$ refers to the dataset from embedded events and jet $\to \tau_h$ refers to the dataset from the FF application region. See section 2.6.5 for details about the selection of the datasets. The inner and outer pie chart on the right show per class the number of events and the sum of event weights, respectively. The differential signal samples reflecting the STXS stage 1 framework are merged to the classes ggH and qqH.

between the source and target domain. An example for the improved event description due to embedded events is shown in figure 2.14.

A more complex scenario is given by the QCD multijet process entering the analysis with jet $\to \tau_h$ misidentifications. Because suitable simulated events are not available due to the complexity of the process, the only possibility to include this class of events in the training is extracting information from the data. Since the NN training is performed fully supervised, a subset of the data is used which is highly dominated by such events. The tau identification is the major discriminator against events from tau misidentifications and therefore a suitable dataset is built from events with a loose tau identification excluding events required for the analysis. Because the selection does not only target QCD events but all events from jet $\to \tau_h$ misidentifications, this event class also includes events from the $t\bar{t}$ and W + jets processes, which simulated events are consequently removed from the training dataset. The task to minimize the difference of this dataset to the analysis selection is part of the FF method [33, 34]. The FF method derives from additional determination regions in data and simulation the probabilities that in the application region, which is congruent with the selection described before, an event belongs to $t\bar{t}$, W + jets or QCD and corrects for events not originating from jet $\to \tau_h$ misidentifications. These probabilities are used as weights in the training to model the distribution of the jet $\to \tau_h$ dataset as close as possible to the target in the analysis. An example for the contribution of the processes to the described selection in data is shown in figure 2.15.

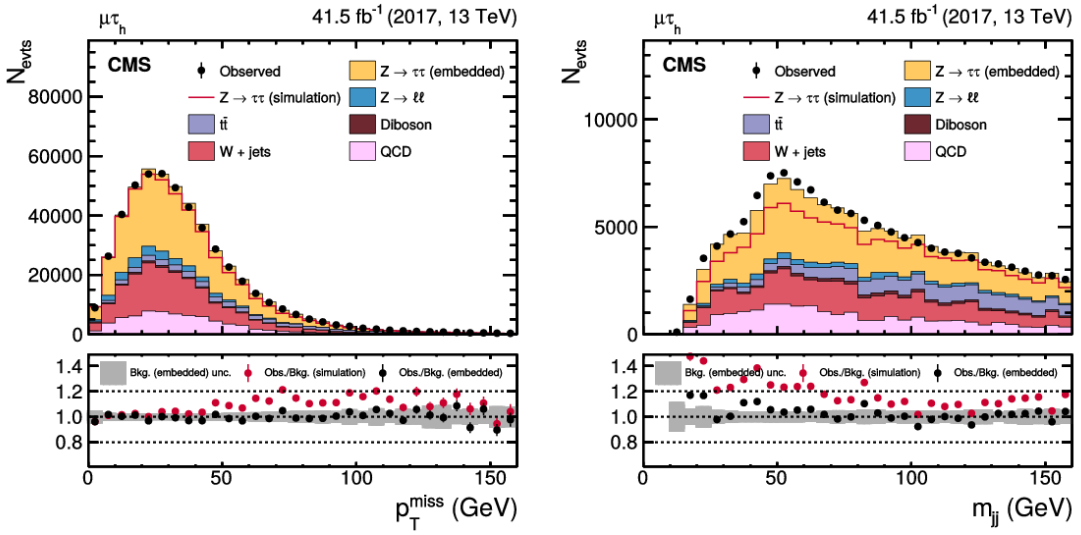

Figure 2.14: Comparison of embedded and fully simulated events exemplary shown for the missing transverse momentum and the invariant mass of the leading and subleading jet [32]

The previously described measures minimize the differences between the datasets used for the training of the NN and the data. But the knowledge about similarities also can be incorporated in the training procedure to enable transfer learning between datasets. Such an approach is followed in this analysis with an inclusive training of a single NN for all data taking periods given additional boolean input variables specifying the period. The usage of the condition information in the training allows to perform the analysis with 8 NNs instead of 24, which reduces the complexity significantly and improves the sensitivity up to 10 %. A detailed study is performed in [66].

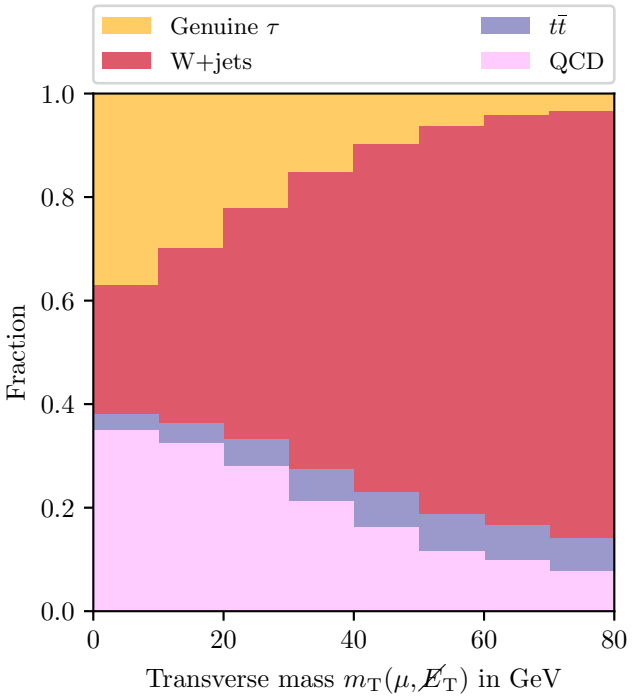

Figure 2.15: Composition of the data in the $\mu\tau_{\mathrm{h}}$ channel from the 2018 data taking period as a function of the transverse mass of the muon including the missing transverse energy. The shown selection has a loose tau identification but is excluding events entering the analysis and is used for the NN training to identify events with jet $\rightarrow \tau_{\mathrm{h}}$ misidentifications.

2.7 Robust multivariate analysis in the presence of systematic uncertainties

This section discusses the challenges resulting from the usage of ML methods for the observables and event categorization in the presence of systematic uncertainties and presents techniques to maintain a robust and reliable analysis.

2.7.1 Challenges and strategies

Measurements in HEP are driven by a detailed understanding of the analysis from the particle collisions to the statistical inference. Such as shown in section 2.5, all known uncertainties are incorporated in the statistical model, which enables data analysis in HEP to perform complex measurements with highest precision. Therefore, the major effort of such an analysis is the precise understanding of the data, simulated events and the according systematic uncertainties. At the same time, the major risk for an analysis to draw incorrect or imprecise conclusions from the data is the absence of uncertainties in the statistical model. Missing or underestimated uncertainties potentially result in the interpretation of uncovered mismodelings of the expectation as a feature

in the data caused by the physical process of interest. In the worst case an analysis may falsely discover a new particle or falsely announce a deviation from the SM. The traditional solution is a thorough study of the agreement between data and expectation for all relevant marginal distributions of the dataset, for example the variables listed in table 2.5, and for selections of the data with known compatibility in consideration of the known uncertainties. In summary, the strategy is the comprehensive description of the expectation for such control regions in the statistical model and the extrapolation of this knowledge in the regions with signal expectation, which enables a precise interpretation of the data.

This approach served well in numerous analyses with strategies that select the signal region based on information from the marginal distributions. However, the extensive usage of ML for the observables and event categorization amplifies the concerns about uncovered mismodelings. Because ML methods are explicitly designed to exploit information from the multidimensional input space and the additional gain in sensitivity compared to selections based on the marginal distributions is a result of the sensitivity of the ML method to higher order features in the data, the strategy to ensure a robust and reliable analysis must be revised.

It should be noted that ML methods do not carry any systematic uncertainty by themselves. After the training of the method, the free parameters of the model are frozen and constant for the analysis. Therefore, any ML model, for example the NNs in this analysis, can be treated as a function mapping the input space spanned by the input variables to the output space of the predictions, which is not different from constructing an invariant mass from the kinematic properties of particles. The differences to conventional functions are the dimensionality of the input space potentially being of a considerably higher dimension and the hidden relations between the input and output space. The strategy and developments presented in this thesis address these new challenges with an extended validation of the multidimensional input space and novel techniques to study and control the dependence of the ML model on higher order features in the dataset.

2.7.2 Input space and model validation

To be sure that the outputs of the ML model are well described by the statistical model, the description of the inputs has to be ensured for the N dimensional input space spanned by the N input variables. Because the statistical model, such as introduced in section 2.5, is based on a binned Poisson likelihood, a holistic approach to quantify the agreement of the expectation and data is not possible. Therefore, the N dimensional space is separated into subspaces, for example 1D subspaces equal to the marginal distributions or 2D subspaces incorporating also the correlations between pairs of variables. These subspaces can be summarized with counts such as required by the statistical model, allowing to use statistical tests to quantify the agreement between expectation and data including all uncertainties. A suitable statistical test is the saturated goodness of fit (GoF) test [67, 68].

The saturated GoF test is using the likelihood as test statistic normalized to the likelihood of the saturated model. The saturated model has an expectation exactly equal

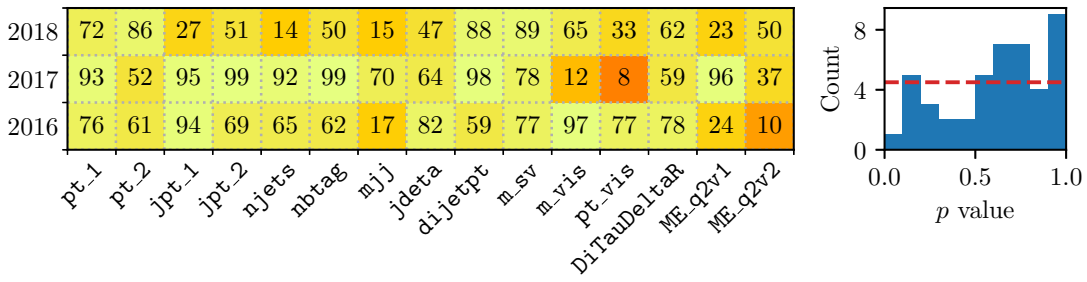

Figure 2.16: Saturated GoF tests for the $\mu\tau_h$ channel carried out for the 1D subspaces of the variables listed in table 2.5. The p values of the GoF test are shown in percent in the matrix on the left and the results are summarized in the histogram on the right. The agreement is tested for each variable with ten equally populated bins in data and the full statistical model including systematic uncertainties. Testing with a χ^2 test the null hypothesis that the p values of the GoF tests are equally distributed (red dashed line) results in a p value of 0.14 quantifying the good overall agreement.

to the data and therefore results in the maximum possible likelihood for the given data yielding a suitable foundation for the test statistic [67]. The distribution of the test statistic is sampled with pseudo experiments enabling the quantification of the compatibility with the statistical model including systematic uncertainties. The test is suited for the validation in higher dimensions because each bin of a multidimensional histogram is treated independently, which avoids ambiguities of the result due to ordering. The test results are summarized by setting a threshold on the p value, typically 5 %, and counting the tests, which fail the requirement. The expectation for the 5 % threshold is to observe at the maximum the same percentage of failing tests, indicating a good agreement between the expectation and data for the studied subspaces of the dataset. Alternatively the distribution of the p values from the GoF tests is expected to be equally distributed, which can be tested statistically such as carried out in figure 2.16 and figure 2.17.

In practice, the approach poses two challenges. First, the number of tests increases exponentially following the binomial coefficient $\frac{N!}{D!(N-D)!}$ with N the number of variables and D the dimension of the subspaces of interest. The number of tests in $\mathcal{O}(N^D)$ yield a computational challenge because the counts have to be computed from terabytes of data for each data taking period and decay channel. Second, the curse of dimensionality forces a coarse binning for higher dimensions or, in reverse, limits the dimensionality of the testable subspaces given a fix granularity for the histograms.

Figure 2.16 shows exemplary results of the 1D GoF tests in this analysis evaluated for the variables listed in table 2.5. According tests of the 2D subspaces are presented in figure 2.17, which test in addition the correlations of variable pairs. Figure 2.17 summarizes 105 GoF tests for the $\mu\tau_h$ channel and the 2016 data taking period being only $\frac{1}{12}$ of the full Run 2 analysis. A full summary of the GoF test results can be found in [43].

As obvious from the discussion, a brute force approach for the input space validation is

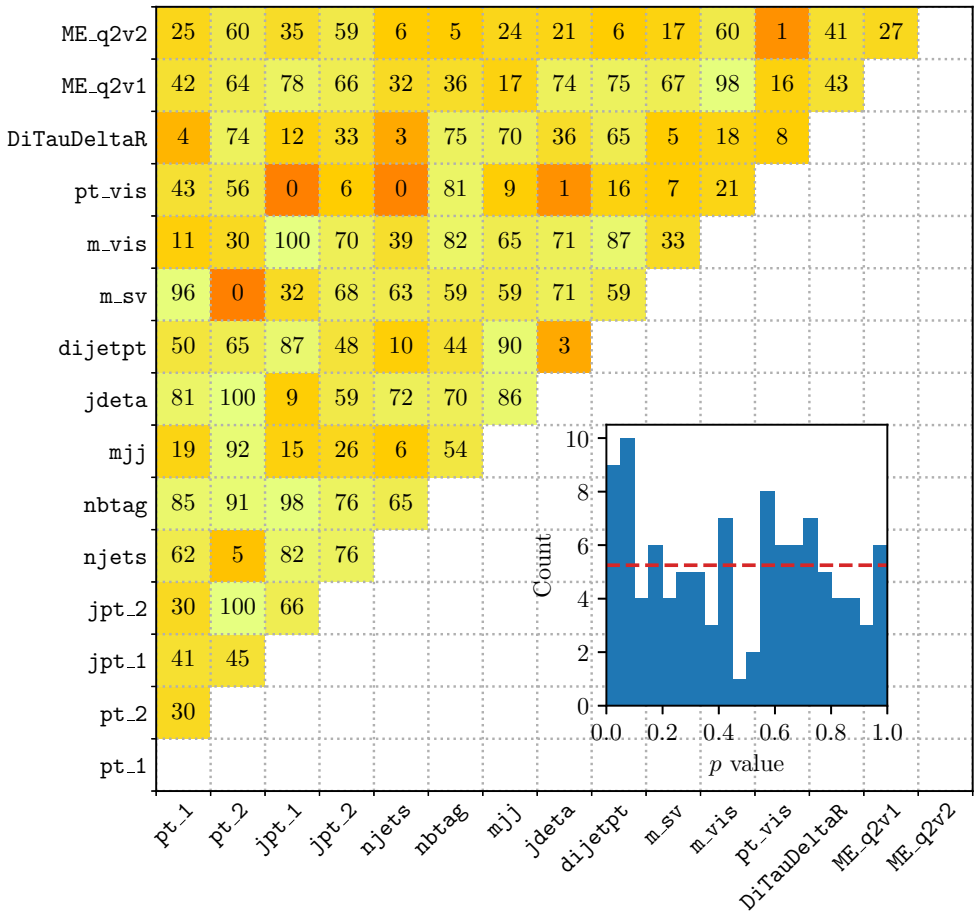

Figure 2.17: Saturated GoF tests for the $\mu\tau_h$ channel and the 2016 data taking period carried out for the 2D subspaces of the variables listed in table 2.5. The p values of the GoF test are shown in percent in the matrix and the results are summarized in the histogram. The 2D histogram as input for the statistical model is built with the same binning in each variable than used for the 1D tests in figure 2.16 summing up to 100 bins per histogram. Testing with a χ^2 test the null hypothesis that the p values of the GoF tests are equally distributed (red dashed line) results in a p value of 0.48 quantifying the good overall agreement.

challenging. Therefore, the detailed understanding of the relations between the ML output and the input space is an additional integral part of a practicable validation strategy for ML based analyses. Chapter 3 is dedicated to this topic and presents techniques to reveal the dependencies of the NN function on features in the input space introducing a novel technique to analyze the dependence on selected subspaces. Such techniques allow to focus the validation of the input space on the subspaces that contribute most to the response of the model, which improves significantly the practicability of the discussed strategy for a robust multivariate analysis.

In addition, a detailed analysis of the ML model is useful to verify the validity of the learned relations itself serving as an additional protection against undiscovered mismodelings. Due to the long history of data analysis in HEP and detailed predictions from theory, the dominant variables contributing to the separation of signal and background processes are known, which allows to identify unexpected features picked up by the trained model and can trigger further investigations.

The technique to analyse the NN function based on a Taylor expansion is discussed in detail in chapter 3. Figure 2.18 visualizes the dependence of the NN function on the input variables in first order, revealing the dominant variables used by the NN for the classification task. Table 2.7 summarizes the most influential features up to the second order, which enables a more detailed insight into the NN function, e.g., the impact of information hidden in the relations of variables. The Taylor expansion of NN functions and the interpretation of the coefficients is extensively discussed in chapter 3.

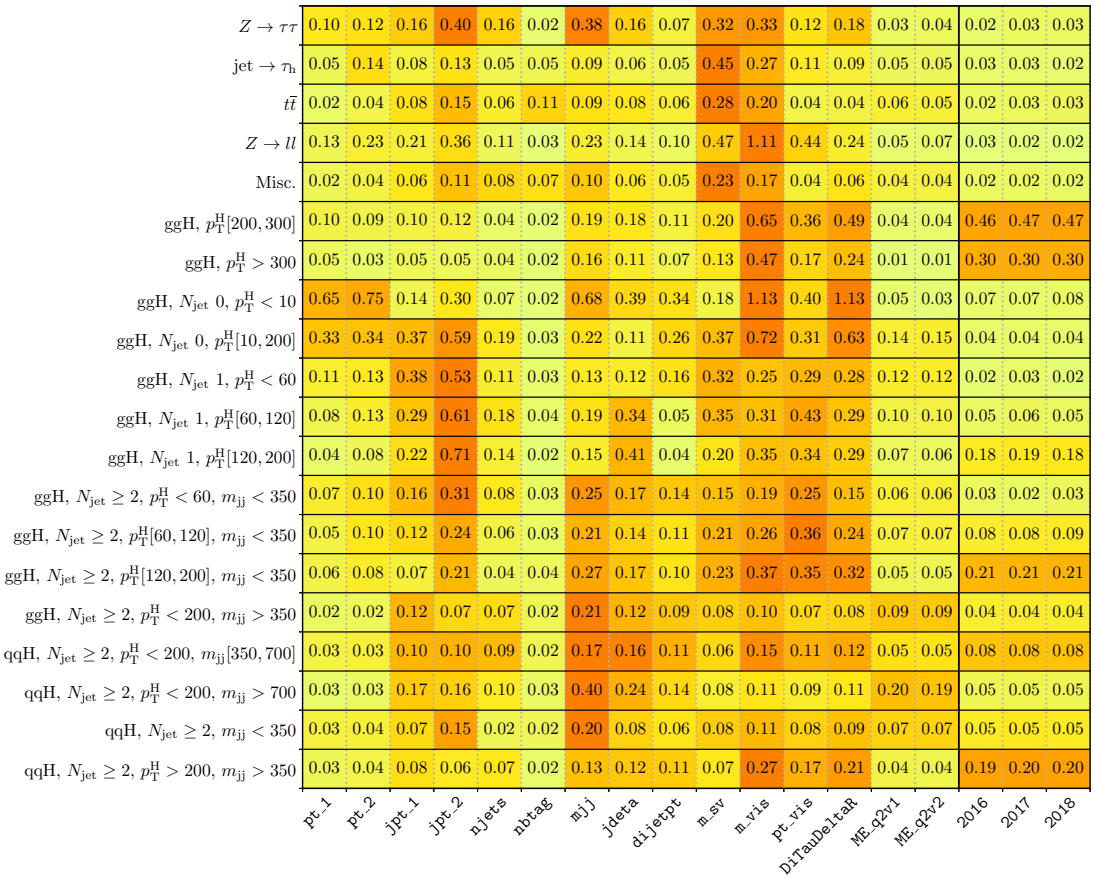

Figure 2.18: The sensitivity scores derived from the first order Taylor coefficients of a NN trained for event categorization in the $\mu\tau_h$ channel. The matrix shows the values of the scores, independently highlighted per row with the coloring. The solid vertical line separates the conditional features providing information about the data taking period, see section 2.6.5. The overview shows the importance of the (invariant) mass of the ditau system and the dijet system such as made use of in previous $H \to \tau\tau$ analyses [46, 47]. It should be noted that the variables are not independent but carry similar information. For example, the information about two or more jets being present in the events is not only carried by `njets` but also implicitly by `mjj`, which is set to a default value for less than two jets. Similarly, the variables describing the mass of the ditau system, `m_vis` and `m_sv`, carry to a large extent the same information, giving the NN the choice during the training.

Table 2.7: The sensitivity scores derived from the Taylor coefficients up to the second order of a NN trained for event categorization in the $\mu\tau_h$ channel. Shown are the ten features with the largest scores representing the subspaces up to second order with most influence on the respective NN output. The appearance of the conditional variables `2016`, `2017` and `2018` (see section 2.6.5) indicates that the NN has learned information dependent on the data taking period. Second order coefficients with the same variable indicate the discrimination by a peaking distribution, for example frequently appearing for the mass variables `m_vis`, `m_sv` and `mjj`. The interpretation of the coefficients is extensively discussed in chapter 3.

$Z \to \tau\tau$			jet $\to \tau_h$			$t\bar{t}$			$Z \to ll$			Misc.		
m_vis	DiTauDeltaR	1.83	m_sv	m_sv	3.18	m_sv	m_sv	1.81	m_vis	m_vis	12.94	m_sv	m_sv	1.69
m_sv	m_sv	1.75	m_vis	m_sv	2.93	m_vis	m_sv	1.51	m_vis	DiTauDeltaR	2.50	m_vis	m_sv	1.50
jdeta	mjj	1.43	m_vis	m_vis	1.10	m_vis	m_vis	0.45	m_vis	m_sv	2.29	m_vis	m_vis	0.48
mjj	mjj	1.37	mjj	mjj	0.91	m_sv		0.28	m_sv	m_sv	1.95	mjj	mjj	0.47
m_vis	m_sv	1.30	jdeta	mjj	0.82	m_sv	2017	0.23	jdeta	mjj	1.80	jdeta	mjj	0.42
m_vis	mjj	1.20	m_sv		0.45	m_sv	2016	0.22	m_vis	pt_vis	1.69	m_sv		0.23
mjj	DiTauDeltaR	1.15	m_vis	DiTauDeltaR	0.44	mjj	mjj	0.20	pt_2	m_vis	1.67	m_sv	jpt_2	0.21
m_vis	m_vis	1.10	pt_2	m_vis	0.36	m_sv	2018	0.20	m_vis	mjj	1.63	m_sv	2017	0.21
pt_1	DiTauDeltaR	1.07	m_sv	2016	0.35	m_vis		0.20	mjj	mjj	1.57	m_sv	2016	0.20
pt_2	DiTauDeltaR	1.04	m_sv	2017	0.35	jdeta	mjj	0.17	pt_1	m_vis	1.43	mjj	jpt_2	0.18

ggH, $p_T^H[200,300]$			ggH, $p_T^H > 300$			ggH, N_{jet} 0, $p_T^H < 10$			ggH, N_{jet} 0, $p_T^H[10,200]$			ggH, N_{jet} 1, $p_T^H < 60$		
DiTauDeltaR	2018	2.64	m_vis	2017	1.40	m_vis	DiTauDeltaR	9.72	m_vis	DiTauDeltaR	5.36	mjj	mjj	3.70
DiTauDeltaR	2016	2.63	m_vis	2016	1.40	m_vis	m_vis	7.76	m_vis	m_vis	4.52	jdeta	mjj	2.73
DiTauDeltaR	2017	2.62	m_vis	2018	1.39	pt_1	DiTauDeltaR	5.71	jdeta	mjj	3.35	m_vis	m_vis	1.36
m_vis	2018	2.26	m_vis	m_vis	1.33	mjj	mjj	5.61	mjj	mjj	3.20	jpt_1	jpt_1	1.25
m_vis	2017	2.26	2016	2017	1.21	jdeta	mjj	5.52	pt_1	DiTauDeltaR	3.14	m_sv	m_sv	1.22
m_vis	2016	2.24	2016	2018	1.20	pt_1	m_vis	5.43	pt_1	m_vis	2.97	m_vis	m_sv	1.12
2017	2018	2.23	2017	2018	1.20	pt_2	DiTauDeltaR	5.36	pt_2	DiTauDeltaR	2.95	jpt_1	jpt_2	1.06
2016	2018	2.22	DiTauDeltaR	2016	1.10	DiTauDeltaR	DiTauDeltaR	5.27	pt_2	m_vis	2.89	pt_vis	DiTauDeltaR	0.94
m_vis	DiTauDeltaR	2.22	DiTauDeltaR	2017	1.10	pt_2	m_vis	5.20	DiTauDeltaR	DiTauDeltaR	2.89	jpt_2	jpt_2	0.87
2016	2017	2.22	DiTauDeltaR	2018	1.09	dijetpt	DiTauDeltaR	4.45	dijetpt	DiTauDeltaR	2.25	mjj	jpt_2	0.81

ggH, N_{jet} 1, $p_T^H[60,120]$			ggH, N_{jet} 1, $p_T^H[120,200]$			ggH, $N_{jet} \geq 2$, $p_T^H < 60$, $m_{jj} < 350$			ggH, $N_{jet} \geq 2$, $p_T^H[60,120]$, $m_{jj} < 350$			ggH, $N_{jet} \geq 2$, $p_T^H[120,200]$, $m_{jj} < 350$		
jdeta	mjj	2.08	mjj	mjj	1.95	mjj	mjj	1.06	pt_vis	DiTauDeltaR	1.24	pt_vis	DiTauDeltaR	1.57
mjj	mjj	2.07	jdeta	mjj	1.82	m_sv	m_sv	0.96	m_sv	m_sv	1.23	m_sv	m_sv	1.28
m_sv	m_sv	1.93	pt_vis	DiTauDeltaR	1.55	m_vis	m_sv	0.86	m_vis	m_sv	1.00	m_vis	pt_vis	1.17
pt_vis	DiTauDeltaR	1.60	m_vis	pt_vis	1.33	m_vis	m_vis	0.72	m_vis	pt_vis	0.82	mjj	mjj	1.12
m_vis	m_sv	1.48	jpt_2	jpt_2	1.22	jdeta	jpt_2	0.69	mjj	mjj	0.79	m_vis	DiTauDeltaR	1.07
m_vis	m_vis	1.09	jdeta	jpt_2	1.18	jdeta	mjj	0.68	pt_vis	pt_vis	0.76	DiTauDeltaR	2017	1.03
pt_vis	pt_vis	0.98	m_sv	m_sv	1.14	pt_vis	DiTauDeltaR	0.64	m_vis	m_vis	0.69	DiTauDeltaR	2018	1.03
m_vis	pt_vis	0.92	m_vis	DiTauDeltaR	1.11	jpt_2	jpt_2	0.64	DiTauDeltaR	DiTauDeltaR	0.69	DiTauDeltaR	2016	1.02
DiTauDeltaR	DiTauDeltaR	0.91	DiTauDeltaR	2018	1.02	mjj	jpt_2	0.62	m_vis	DiTauDeltaR	0.54	DiTauDeltaR	DiTauDeltaR	1.01
jpt_2	jpt_2	0.84	DiTauDeltaR	2016	1.01	jpt_1	jpt_2	0.51	jdeta	mjj	0.50	m_vis	m_vis	0.87

ggH, $N_{jet} \geq 2$, $p_T^H < 200$, $m_{jj} > 350$			qqH, $N_{jet} \geq 2$, $p_T^H < 200$, $m_{jj}[350,700]$			qqH, $N_{jet} \geq 2$, $p_T^H < 200$, $m_{jj} > 700$			qqH, $N_{jet} \geq 2$, $m_{jj} < 350$			qqH, $N_{jet} \geq 2$, $p_T^H > 200$, $m_{jj} > 350$		
mjj	mjj	0.81	m_vis	DiTauDeltaR	0.45	jdeta	mjj	1.19	mjj	mjj	1.12	DiTauDeltaR	2018	1.18
jdeta	mjj	0.80	pt_vis	DiTauDeltaR	0.41	mjj	mjj	1.05	jdeta	mjj	0.55	DiTauDeltaR	2017	1.17
m_sv	m_sv	0.43	DiTauDeltaR	2018	0.41	jdeta	jdeta	0.41	m_sv	m_sv	0.40	DiTauDeltaR	2016	1.17
m_vis	m_sv	0.34	m_vis	pt_vis	0.41	mjj		0.40	pt_vis	DiTauDeltaR	0.39	m_vis	DiTauDeltaR	1.11
m_vis	m_vis	0.26	DiTauDeltaR	2016	0.41	m_sv	m_sv	0.36	m_vis	DiTauDeltaR	0.35	m_vis	2018	1.02
pt_vis	DiTauDeltaR	0.26	DiTauDeltaR	2017	0.41	mjj	dijetpt	0.35	m_vis	pt_vis	0.32	m_vis	2017	1.01
jpt_1	dijetpt	0.26	jdeta	mjj	0.40	mjj	jpt_2	0.34	m_vis	m_vis	0.31	m_vis	2016	1.00
m_vis	DiTauDeltaR	0.26	m_vis	m_vis	0.39	ME_q2v1	ME_q2v2	0.32	m_vis	m_sv	0.29	2017	2018	0.98
m_vis	pt_vis	0.25	m_vis	2018	0.37	dijetpt	DiTauDeltaR	0.30	DiTauDeltaR	2018	0.28	2016	2018	0.97
jdeta	jdeta	0.22	m_vis	2017	0.37	m_vis	m_sv	0.30	DiTauDeltaR	2017	0.28	2016	2017	0.97

2.8 Differential measurement of the Standard Model $H \to \tau\tau$ cross section

This section presents the results of the analysis with focus on the differential measurement of the SM Higgs boson cross section in the STXS framework and discusses potential improvements of the analysis strategy for future measurements.

Figure 2.19 shows the measurement of the signal strengths parameters associated with the signals defined in the STXS framework and figure 2.20 shows the correlation of these POIs. In addition, figure 2.21 presents the signal strength parameters scaled to the actual cross sections. Figures visualizing the data and expectation being input to the statistical inference in the $\mu\tau_{\mathrm{h}}$ channel for the data taking period of 2018 are presented in appendix A. The results show a tension with the expectation from the SM of $3\,\sigma$ in Gaussian standard deviations for ggH events with no jets. This is an interesting finding and future measurements and combinations have to show whether the deviation is of a statistical nature or due to an actual physics effect, which is revealed in this analysis by the unprecedented fine granularity of the measurement. The presented results confirm the importance of the STXS framework, which fosters the combination of such measurements with other analyses and enables to gain a maximum of sensitivity to new physics.

The measurements in the STXS framework facilitate the organized search for a physics model, which supersedes the SM. The new model may be able to predict the existence of further fundamental particles, for example such as part of the Minimal Supersymmetric Standard Model (MSSM) [69–71], and could explain unsolved questions in physics, e.g., the existence of dark matter. Projects such as HiggsBounds and HiggsSignals [72–76] take the measurements to perform an interpretation of the experimental results in terms of the compatibility with other physics models, which is an important contribution to guide the experiments to interesting analyses and measurements. These studies are also the reason why the correlation of the POIs shown in figure 2.20 is highly important, because this information enables in such applications the statistically sound interpretation and combination of the measurements.

Figure 2.19 contains also valuable information with respect to the applied analysis strategy. Even though the analysis measures concurrently 12 signals with cross sections ranging from about 10 fb to 1000 fb, the results are not heavily dominated by the statistical uncertainty. Stronger pronounced for the more frequent ggH process, the systematic and theory uncertainties have a significant impact on the absolute uncertainty. This finding means for future analyses that a substantial reduction of the absolute uncertainty following $1/\sqrt{N}$ with an increasing number of events N can not be expected. Another important ingredient is the discussion carried out in chapter 5, which presents the evidence that the used ML based analysis strategy with the NNs trained on the CE loss is a powerful approach to optimize the estimate of the POIs with respect the statistical uncertainty. These points indicate that future analysis may not be able to improve dramatically with respect the statistical component of the uncertainty but have to take into account all contributing uncertainties to achieve further significant improvements.

Chapter 4 discusses in detail the usage of modern ML techniques to implement the

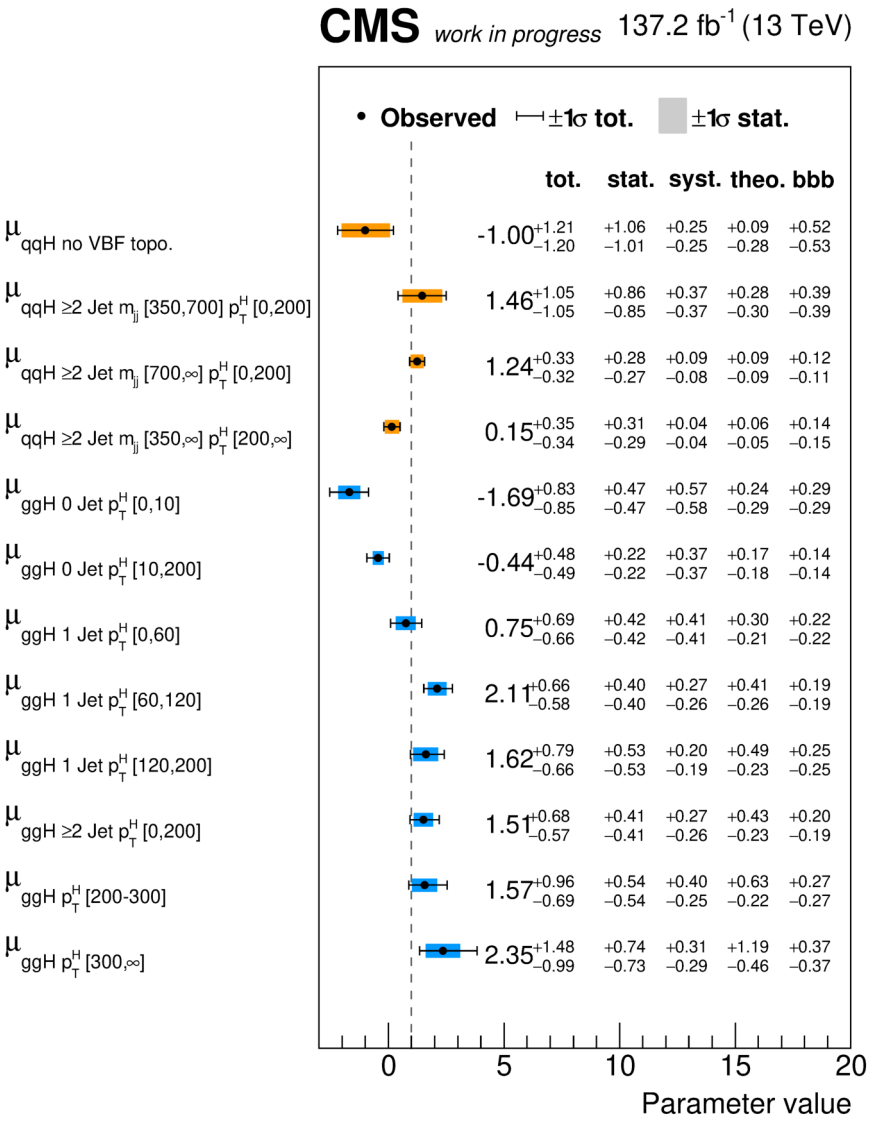

Figure 2.19: Differential measurement of signal strength for the SM Higgs boson in the decay to two tau leptons. The columns indicate the contribution of different uncertainty categories to the total uncertainty. The categories are split into the statistical component, the theory uncertainties, the bin-by-bin uncertainties and all other systematic uncertainties. The correlation matrix of the POIs is shown in figure 2.20. [44]

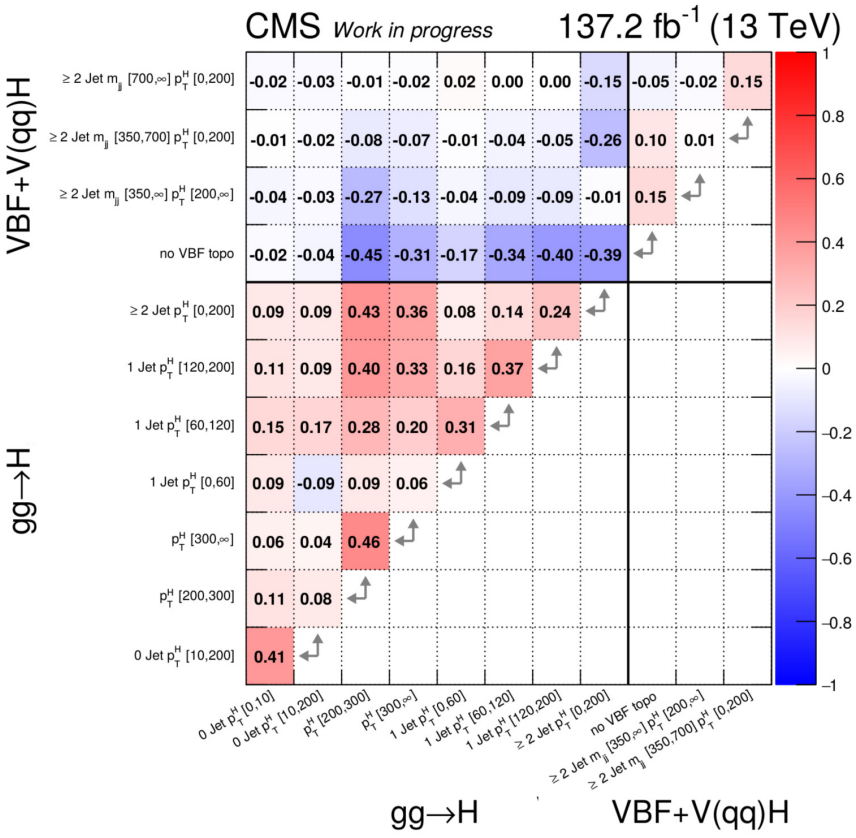

Figure 2.20: Correlation of the measured signal strength parameters presented in figure 2.19 [44]

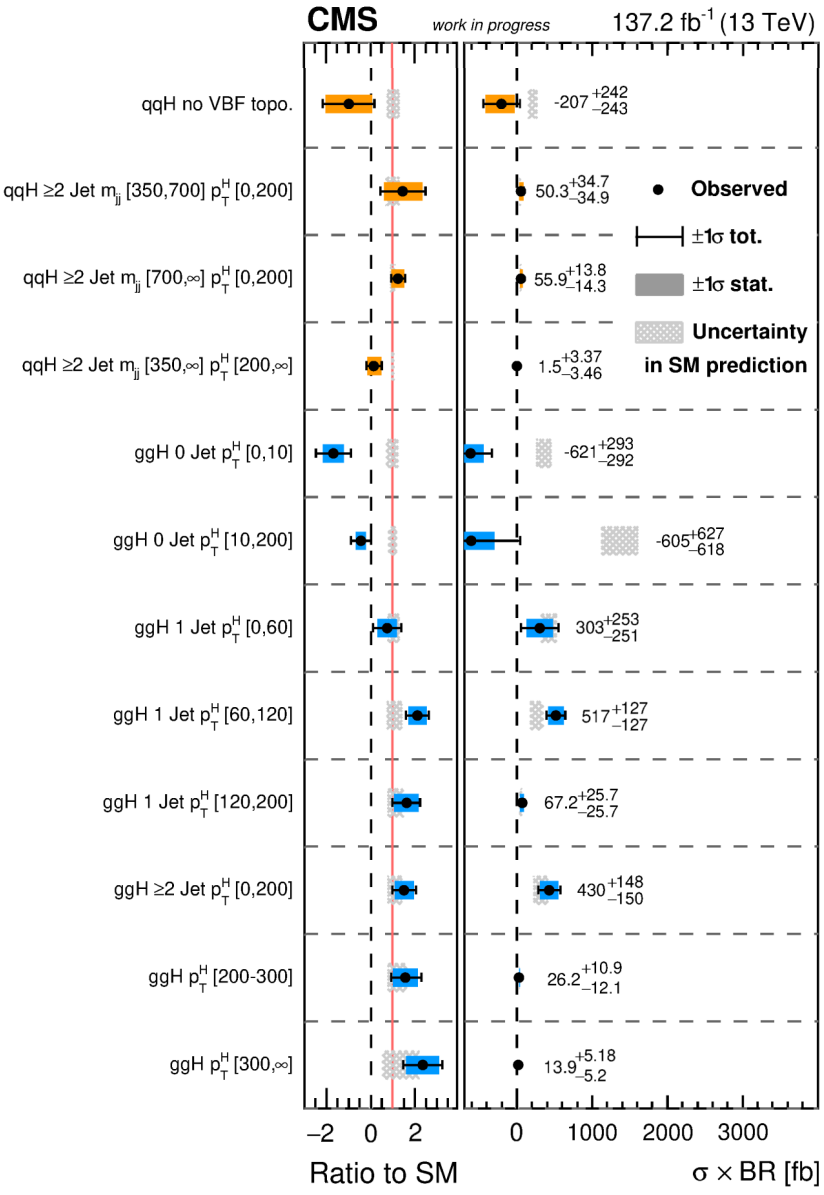

Figure 2.21: Differential cross section measurement of the SM Higgs boson in the decay to two tau leptons. Adapted from [44].

decorrelation of the NN function to any variation in the input space. These techniques allow to easily remove the propagation of systematic uncertainties on the result of the measurement. However, the examples in chapter 4 show that improving the actual analysis objective, e.g., the absolute uncertainty of a POI, is highly non trivial. The optimization is coupled between the statistical and systematic part of the uncertainty and manually controlled by a hyperparameter. The examples show that the existence of a systematic uncertainty does not guarantee that a decorrelation improves the analysis but potentially even worsens the absolute uncertainty of the POI because the suppressed information may be crucial for the separation of the signal and background processes. In summary, the optimization problem is a complex tradeoff between all contributing uncertainties.

A possible solution to this optimization problem is discussed in chapter 5, which proposes a computational efficient method to find an optimal tradeoff between all contributing uncertainties. The differential formulation of the analysis objective, e.g., the absolute uncertainty of the estimate of a POI, allows to converge to an optimal solution with a balance between the statistical and systematic uncertainty. Shown in examples in chapter 5, the optimization can result in a reduced absolute uncertainty compared to the optimization purely based on the statistical component such as in typical ML based analysis strategies. These novel strategies for analysis in HEP have the potential to offer significant improvements for future measurements in the upcoming precision era of Higgs physics at the LHC.

CHAPTER 3

Understanding the dependence of the machine learning model on features in the input space

This chapter discusses the requirement for multivariate data analysis in HEP to be able to reveal the dependence of the ML model on the features in the input space. Existing solutions from the ML community are analyzed with respect to their suitability for the usage in physics data analysis to ensure robust and reliable measurements. Finally, a novel technique is presented, which allows to study the dependence of the NN function on selected subspaces of the input space, tailored to the challenges posed by multivariate data analysis in HEP.

3.1 About the special requirements for multivariate data analysis in high-energy particle physics

Chapter 2 discusses in detail the required procedures to perform precise measurements in HEP. A crucial aspect of the data analysis is the detailed understanding of all contributing parts so that all uncertainties of the measurement can be reflected in the statistical model enabling a reliable statement about the uncertainty of the result. Such as discussed in more detail in section 2.7, a major concern regarding multivariate analysis in HEP is the misinterpretation of data due to uncovered mismodelings in the expectation of the statistical model. ML in the analysis strategy amplifies these concerns since the analysis may be highly sensitive to mismodeled higher order features in the dataset because methods such as NNs draw their performance explicitly from such features. The solution is the thorough validation of the input space to validate the expectation of the statistical model including all statistical and systematic uncertainties. A decisive role for such a solution plays the ability to understand the relations between the multidimensional input space and the outputs of the ML model to be able to guide the validation of the input space. In addition, the sensitivity of the ML model to the features in the dataset can be compared to the expectation from the knowledge about the detector and the underlying theory, providing an additional layer of scrutiny.

The requirements set out above are not fully congruent with the targets of related techniques and the literature in the ML community. Mainly driven by computer vision,

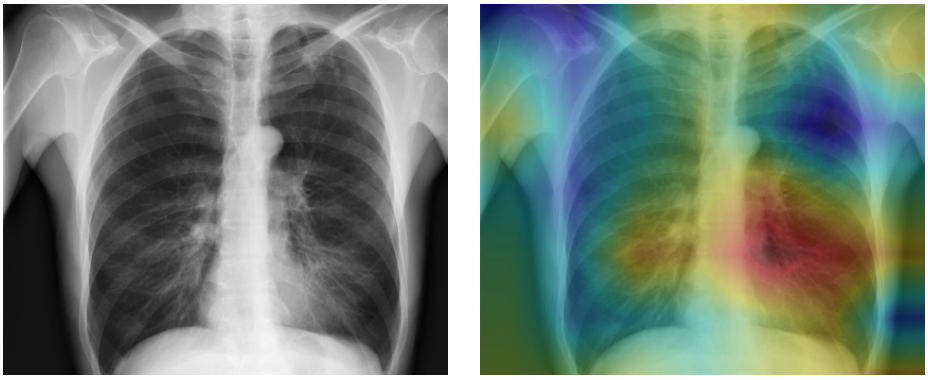

Figure 3.1: An example usecase of techniques to understand the dependence of a ML model on the inputs outside of HEP. The CheXNet model uses an input X-ray image (left) to classify the pathology of the patient, e.g., pneumonia. Class activation maps [77] are used to attribute the classification to the respective parts of the input image (right), which enables a validation of the model response. [78]

the ML community focuses on the attribution of the model response to the single inputs. An example is shown in figure 3.1, which attributes the classification of the X-ray image as pneumonia positive to the respective parts of the input image adding valuable information to validate the response of the ML model. The methods developed with such usecases in mind focus on the full redistribution of the model response on the inputs and do not foresee to analyze the impact of the correlation between inputs on the output, missing a crucial piece of information for a differentiated validation strategy of multivariate analysis in HEP.

The second difference is that most techniques from the ML community analyze a single example of a population. For example in figure 3.1 the question is which part of this specific image contributes most to the classification of the patient's pathology, not what makes an X-ray image revealing pneumonia in general. However, the data analysis in this thesis operates on the full dataset, which sets the focus of such techniques to the latter question and requires novel solutions or a suitable aggregation strategy for the sensitivity analysis of single predictions.

Section 3.2 gives an overview over existing approaches from literature and provides the context for the novel approach introduced in section 3.3 tailored to the challenges of data analysis in HEP.

3.2 Overview over existing approaches

The interest in understanding precisely the mechanics of multivariate analysis techniques is not a novel field and hence the literature provides many different solutions accumulated over time. However, this section points out the most prominent approaches and discusses why the capabilities of the methods are not fully covering the requirements for analysis in HEP.

One of the most prominent classic solutions to identify the most influential features in a multivariate analysis is the principal component analysis (PCA) [79]. The PCA approach transforms the coordinate system of the input space into a linearly uncorrelated one. Next, the eigenvalues of the transformation matrix can be interpreted as the information of the respective axis, revealing the most prominent features of the dataset. The issue to be pointed out for this approach is that there is no guarantee that the ML method will pick up any of the features during training. Following, PCA is not able to provide a measure for the strength of the dependence between the input and output space of the ML model, being a crucial input to rank the most influential features contributing to the measurement.

Another frequently applied method to identify influential inputs of the ML model is the leave-one-out approach. The technique chooses a metric, e.g., the area under the curve (AUC) of a receiver operating characteristic (ROC), measures the influence of the inputs by removing them one by one and retrains the ML model each time. Then, the least influential input is dropped and the procedure is repeated until a single input is left. This approach is picked as an example to point out that methods which modify the trainable parameters of the model cannot guarantee that the model applied in the analysis has the same dependencies such as identified by the respective method. In particular, the issue is amplified by redundant information in the dataset, which offers multiple equally efficient solutions for a task. Such a scenario is typical for HEP datasets, for example due to shared information by the kinematics of constituents and the invariant mass of the decay system.

The largest group of methods is based on the analysis of the gradient from the NN output to the input variables. The underlying idea is that a large gradient tells that a small change of the input has a large influence on the respective output, hence revealing the influential features picked up by the ML model during training. The practicability of such an approach is supported by the modern computing infrastructure used for the implementation of NNs, which is based on computational graph libraries [80–82] natively supporting the calculation of analytic gradients due to automatic differentiation. The usage of information from gradients to explain decisions has shown to be applicable to simple classifications tasks [83] early on and also has been adopted in computer vision applications to identify the significant parts of an image with saliency maps [84]. A similar approach is the identification of relevant inputs with redistribution rules propagating the output of the NN back to the inputs [85], which is closely related to using the gradient times the input as a measure of importance [86]. The method introduced in section 3.3 is also based on a gradient analysis, although the literature focuses on the explanation of the NN response solely in terms of the input variables and does not take into account the decomposition of the input space such as it is important for the application in this analysis.

The interpretation of the decision of a NN model beyond the attribution to the plain input variables is also performed by [87]. The technique is based on learning the linear dependence between the activation of the nodes and a concept defined by a set of labeled examples, enabling an explanation of the NN response in terms of any desired high level feature. Even though not trivially applicable to the scenario in this analysis, the approach

of the method is promising to achieve an easily understandable interpretation of decisions taken by NNs.

3.3 Identifying the relevant dependencies of the neural network function on characteristics of the multidimensional input space

The following sections introduce a novel method to analyze NNs by decomposing the NN function with respect to the subspaces of the multidimensional input space, which enables a fine granular sensitivity analysis tailored for the use case in HEP. A dedicated paper about this method was published in [88].

3.3.1 Method

The new method is centered around the idea to decompose the NN function $f(\boldsymbol{x})$ with a Taylor expansion and associate the coefficients of the decomposition with subspaces of the input space. The decomposition is performed at each element of the dataset $D = \{\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N\}$ with size N, which represents the points in the input space being of interest for the understanding of the NN. D is typically the test split of the available dataset, although also a subset can be of interest, e.g., the analysis of the NN response for a specific class of events. The choice of the examples in D defines the question asked about the NN function because, for example, the dependence of the NN to the input space for the signal class may be considerably different from the inclusive dependence.

The Taylor coefficients carry information about the dependence of the NN function in the respective subspace of $\boldsymbol{x}$. For example in the expansion up to the second order for $\boldsymbol{x} = (x_1, x_2)$ at the point $\boldsymbol{a}$ given by

$$\begin{aligned} T(\boldsymbol{x}) &= f(\boldsymbol{a}) + (x_1 - a_1)\partial_{x_1} f(\boldsymbol{a}) + (x_2 - a_2)\partial_{x_2} f(\boldsymbol{a}) \\ &\quad + \frac{1}{2!}\Big((x_1 - a_1)^2 \partial_{x_1 x_1} f(\boldsymbol{a}) + 2(x_1 - a_1)(x_2 - a_2)\partial_{x_1 x_2} f(\boldsymbol{a}) + (x_2 - a_2)^2 \partial_{x_2 x_2} f(\boldsymbol{a})\Big) \\ &= f(\boldsymbol{a}) + (x_1 - a_1) t_{x_1} + (x_2 - a_2) t_{x_2} \\ &\quad + (x_1 - a_1)^2 t_{x_1 x_1} + (x_1 - a_1)(x_2 - a_2) t_{x_1 x_2} + (x_2 - a_2)^2 t_{x_2 x_2} \end{aligned} \tag{3.1}$$

the Taylor coefficient t_i signals the sensitivity of the NN function f to the respective input space. The symbols t_{x_1} and t_{x_2} represent the dependence of f to the one dimensional input spaces and $t_{x_1 x_2}$ incorporates in addition information from the correlation between x_1 and x_2 in the two dimensional subspace. The terms $t_{x_1 x_1}$ and $t_{x_2 x_2}$ are associated with long range dependencies, explained with examples in the following sections.

It should be noted that due to $\partial_x f(cx) = c\partial_x f(x)$ the coefficients t_i are only comparable if the inputs are scaled to the same mean and variance. The standardization of the inputs is a typical preprocessing technique in ML using the transformation $x \to (x - \mu)/\sigma$ with μ and σ being the mean and the standard deviation of x. Further, the higher order coefficients require also a higher order differentiability of the NN function f, which may

not be given since the optimization in ML is based on first order derivatives. A typical example for an operation without a suitable second order derivative is the rectified linear unit (ReLU) [89] $\max(x, 0)$ being suitable and computationally efficient for typical ML applications but having an overall gradient of zero at second order. Therefore, an alternative choice for the activation function is the hyperbolic tangent function.

Eventually, the coefficients t_i are summarized over the examples in D to condense the information to a set of scores indicating the dependence of the NN function f to the respective subspace of the input space. A suitable aggregation is

$$\langle t_i \rangle = \frac{\sum_{j=1}^{N} w_j \left| t_i \left(\boldsymbol{a}_j \right) \right|}{\sum_{j=1}^{N} w_j} \tag{3.2}$$

with $\boldsymbol{w}$ being the weights of the events in D. The sign of the coefficient t_i carries the information whether the respective subspace contributes to the increase or decrease of the NN output, however, since a measure of importance is desired, the absolute value is chosen. Further, the weights $\boldsymbol{w}$ are potentially renormalized to reflect the task of the training, e.g., the classification of the classes with equal importance such as described in section 2.6. To reflect the sensitivity of the NN with respect to the training objective, the renormalized weights must be used for the computation of the aggregation.

3.3.2 Application on simple examples based on pseudo experiments

The suitability of the Taylor coefficients to perform a detailed analysis of the dependence of the NN function on the input space is demonstrated with simple examples based on pseudo experiments. The observations for the two classes signal and background are drawn from multivariate normal distributions in two dimensions to define binary classification tasks. The parameters of the multivariate normal distributions for each task are shown in table 3.1. Simulated are 10^5 events, which are split in half for the optimization of the NN and the monitoring of the training. The chosen NN architecture consists of a single hidden layer with 100 nodes and hyperbolic tangent activations. The gradients are computed on the first half of the training dataset and applied to the trainable parameters using the Adam optimizer [64]. The loss function is the CE and the training is stopped if the loss did not improve three times in a row on the second half of the training dataset. The presented results from the analysis of the Taylor coefficients are computed on a statistically independent dataset with 10^5 examples.

Figure 3.2 shows the distributions of the classes for the classification tasks and the according sensitivities from the Taylor coefficients. In addition, figure 3.3 shows the distribution of the derivatives in the input space up to the second order, revealing the underlying information aggregated by the sensitivity scores $\langle t_i \rangle$.

The task A is the most simple classification task, defined by two uncorrelated multivariate normal distributed classes with different means. The expectation is to find most sensitivity for the marginal distributions of x_1 and x_2 with the same importance due to the same distance on each axis, which is confirmed by the scores $\langle t_{x_1} \rangle$ and $\langle t_{x_2} \rangle$. Task B reduces the distance of the means on the x_2 axis by half, which halves accordingly the

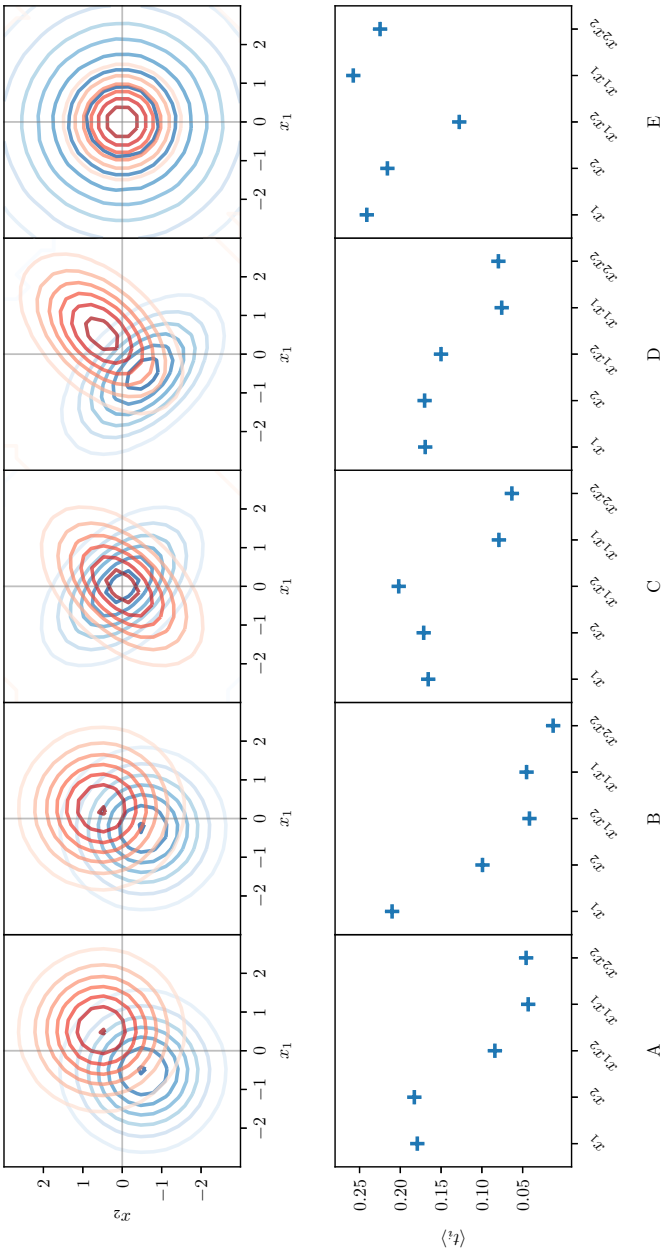

Figure 3.2: The upper row shows the distribution of the signal (red) and background (blue) classes following multivariate normal distributions with the parameters given in table 3.1. The lower row shows the resulting sensitivity scores $\langle t_i \rangle$ for the respective task in each column. In addition, the according distributions of the derivatives in the input space are visualized in figure 3.3.

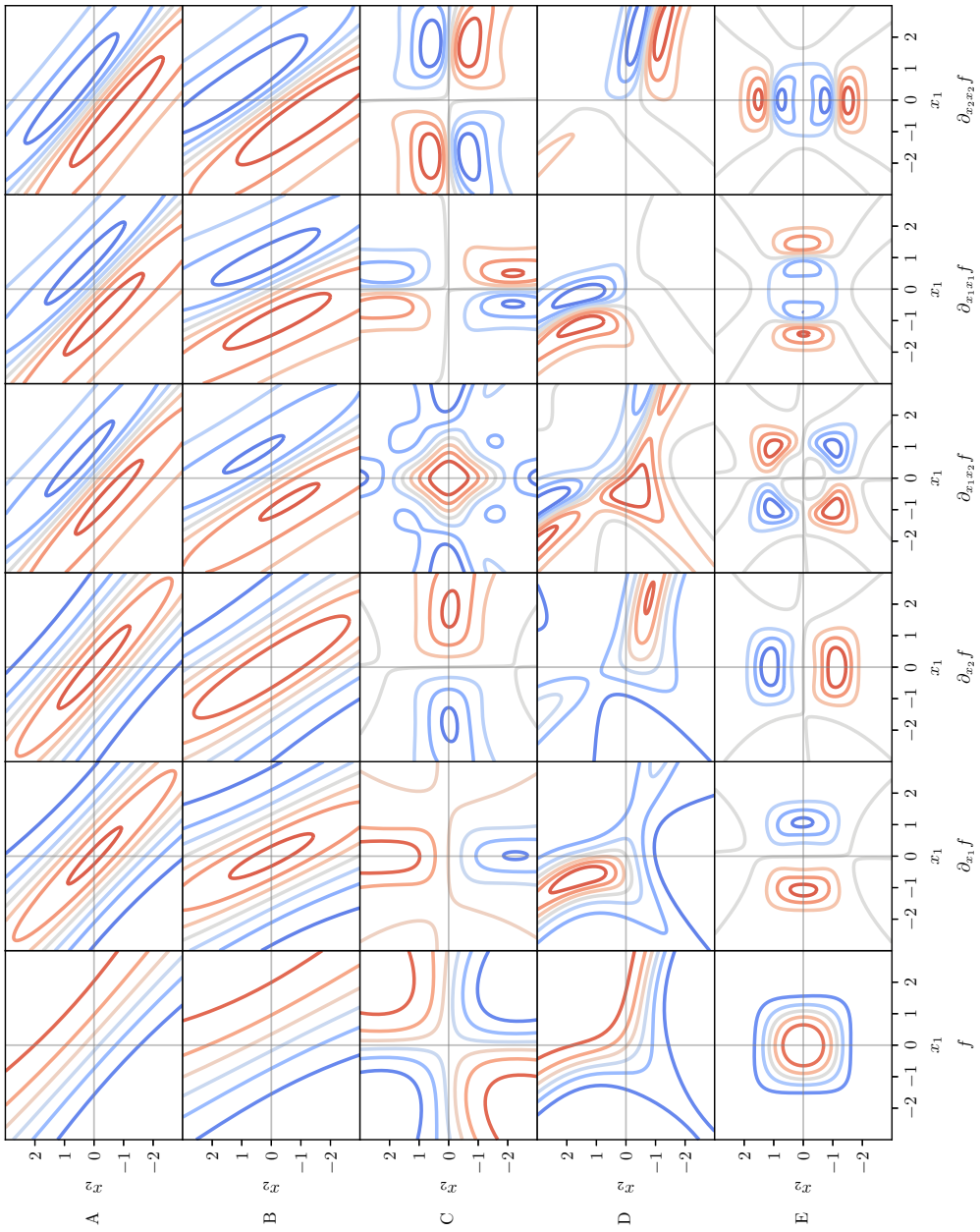

Figure 3.3: Each row relates to one of the simple tasks specified by the multivariate normal distributions for signal and background given in table 3.1 and figure 3.2. The first column visualizes the NN function separating signal from background and the following columns show the derivatives of the NN function up to the second order. It should be noted that the colormaps are normalized per subfigure indicating small values in red and large values in blue and therefore reveal only the relative distribution of the values. The overall scale is visible from the aggregation in figure 3.2.

Table 3.1: Parameters of the multivariate normal distributions defining the binary classification tasks of the simple examples demonstrating the analysis of the Taylor coefficients in different scenarios

Task	Mean				Covariance	
	Signal (x_1, x_2)		Background (x_1, x_2)		Signal	Background
A	0.5	0.5	−0.5	−0.5	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
B	0.5	0.25	−0.5	−0.25	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
C	0	0	0	0	$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$
D	0.5	0.5	−0.5	−0.5	$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$
E	0	0	0	0	$\begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$	$\begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}$

importance of the marginal distribution of x_2 but keeps $\langle t_{x_1} \rangle$ unchanged. To maximize for example the importance of the two dimensional subspace related to the score $\langle t_{x_1 x_2} \rangle$, task C chooses for both classes the same mean but opposite correlations. Consequently, the score related to the $x_1 x_2$ subspace is most prominent. However, $\langle t_{x_1} \rangle$ and $\langle t_{x_2} \rangle$ remain relevant, also in the case of the same mean of both distributions. The reasons are the regions around $(\pm 2, 0)$ and $(0, \pm 2)$, where signal events can be separated from the background class solely with information from the marginal distribution of the respective axis. The effect is clearly visible for task C in the first order derivatives $\partial_{x_1} f$ and $\partial_{x_2} f$ shown in figure 3.3, which mark the regions with a high sensitivity to x_1 or x_2. In comparison, the derivative $\partial_{x_1 x_2} f$ is most prominent around $(0, 0)$ at the bulk of the distribution, resulting in the largest score for $\langle t_{x_1 x_2} \rangle$. Task C is a prime example for the visualization of the potentially completely different sensitivity of the NN function to features in the input space based on the region in the multivariate input space. The sensitivity scores computed from the Taylor coefficients dissect successfully the dependencies of the NN response to the input space, providing a detailed summary of the learned relations. Task D combines the separation from the means in task A and the correlation in task C, whose sensitivity analysis shows a similar picture than for task A but adds the expected additional importance to $\langle t_{x_1 x_2} \rangle$ due to the newly introduced correlations. Finally, task E uses the same mean for both classes but chooses a different variance for signal and background. The task is designed to visualize the meaning of the second order scores $\langle t_{x_1 x_1} \rangle$ and $\langle t_{x_2 x_2} \rangle$, which represent the sensitivity to more complex structures in the marginal distributions. The difference with respect to the related first order scores is the separability of the classes not by a single threshold but by the variance of the variable. Figure 3.4 presents the development of $\langle t_{x_1 x_1} \rangle$ for NNs trained on variations of task A, which shows that this second order feature can be detected independently and scales as

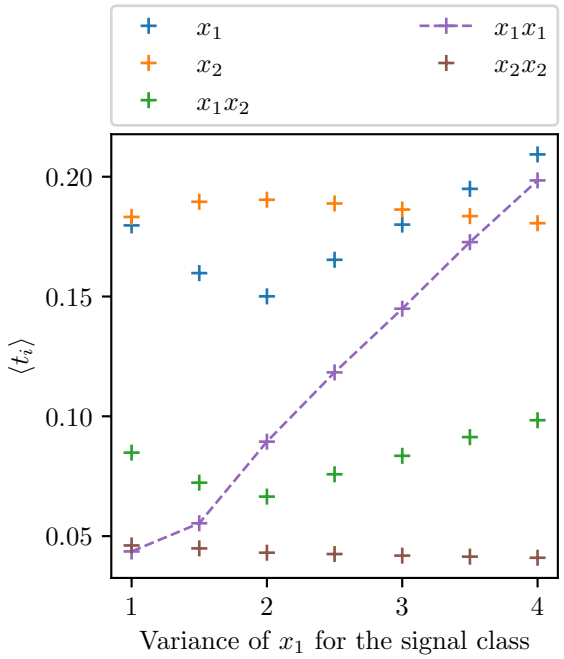

Figure 3.4: Variations of task A with an increasing variance of x_1 for the signal class result in the expected increased sensitivity of the NN function to the second order feature $\langle t_{x_1x_1} \rangle$, whereas a similar sensitivity to all other features is maintained.

expected with the variance of x_1.

3.3.3 Analysis of the learning progress

The sensitivity of the NN function is not only interesting in the case of a fully converged NN but enables also an in depth analysis of the learning progress. Figure 3.5 shows the sensitivity analysis carried out during the training on task D after each gradient step, revealing interesting details about the training progress visible due to the rise in sensitivity to the first and second order features. The figure uses as metric the AUC of the ROC, which allows to estimate the success of solving the task at each step and enables to visualize when the training stops to extract additional information from the data to separate signal and background. The progress shows that the first order features are learned quickly in only a few gradient steps leading already to most of the separation between signal and background. A second rise in separation power is accompanied by an increased sensitivity to $\langle t_{x_1x_2} \rangle$, indicating the successful usage of the additional information hidden in the different correlations of the signal and background events. At gradient step 339, the early stopping rule triggers since the validation loss has not improved three times in a row, marking the convergence of the training. However, longer training increases the sensitivity to the second order features, although without a notable impact on the AUC. This finding allows to validate the expectation that

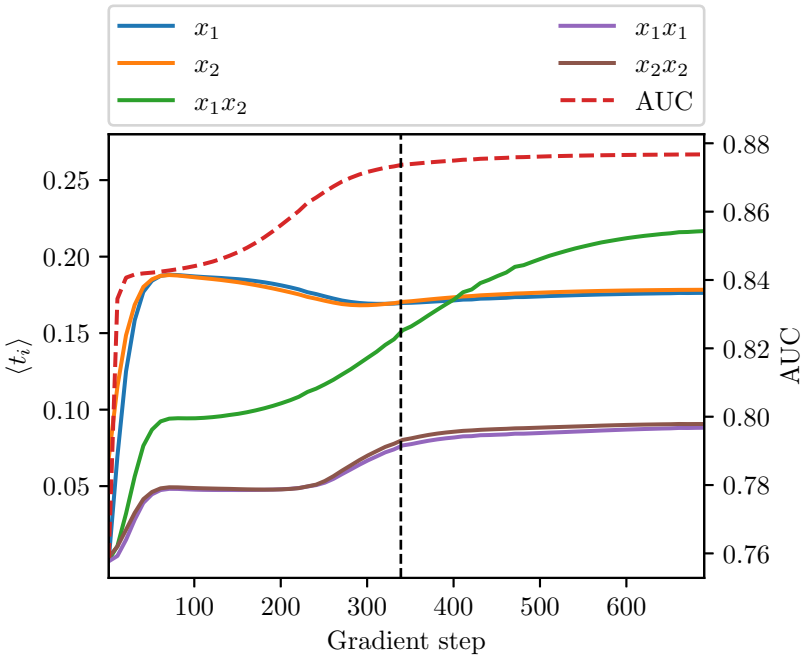

Figure 3.5: The training is based on task D in figure 3.2 with the sensitivity analysis being carried out after each gradient step. The AUC of the ROC is shown on the second axis to the right measuring the success in solving the task. The vertical dashed line indicates the trigger of the early stopping rule at gradient step 339 which marks that the validation loss has not improved three times in a row.

simpler features like the separation in marginal distributions are learned first and more complicated relations in higher order spaces such as differences in correlations require more steps to be picked up by the NN. Also the analysis shows that it is favorable to use early stopping to keep the dependence of the NN function on the input space as simple as possible.

3.3.4 Application on an example from high energy particle physics

This section applies the sensitivity analysis leveraging the information hidden in the Taylor expansion of the NN function on a more complex example from HEP. The example is taken from the Higgs ML challenge [90], which represents a simplified search for the SM Higgs boson in the final state of an electron or muon and a hadronically decayed tau in events at the ATLAS detector, being closely related to the analysis discussed in chapter 2. The dataset consists of 2.5×10^5 events for training and 5.5×10^5 events for testing including labels for each event indicating the origin from a Higgs boson or a background process. The background events are a mixture of the contributing processes but not treated separately during training, which is set up as a binary classification task. A third of the simulated events are signal events, which rate is scaled with event weights

to the expectation of the SM. The chosen NN architecture and training procedure are the same as for the simple example in section 3.3.2 but is using in addition the standardization of the inputs with the $(x - \mu)/\sigma$ transformation.

The inputs are 30 variables ranging from the reconstructed invariant mass of the Higgs boson `DER_mass_MMC` to the primary kinematic properties of the constituents in the final state. From the 30 variables, 17 are primary variables and 13 are derived quantities, marked in the variable names with `PRI` and `DER`, respectively. The list of variables with the exact definitions and their physical meaning is given in [90] but can also be easily inferred from the variable names, in which `lep` refers to the electron or muon and `tau` to the hadronically decayed tau.

The results of the sensitivity analysis are shown in table 3.2, which lists the first and last 20 ranks ordered by sensitivity and the respective contributing variables. The 30 input variables result in 495 sensitivity scores $\langle t_i \rangle$ derived from the first and second order Taylor coefficients. The distribution of the first and second order scores is visible on the left in figure 3.6, which shows a steep fall. The slope indicates that the NN uses only a few features to achieve most of the separation power to solve the binary classification task. To prove this assumption, the NN is retrained once with only the variables contributing to ranks above a threshold on the sensitivity score and again with all other variables excluding the previous selection. The threshold is chosen as the middle point between the largest and smallest sensitivity score at 1.523×10^{-3} and selects the first 11 ranks in table 3.2. The right hand side of figure 3.6 shows the AUC of the ROC being identical for the training on all variables and the subset of variables identified as the sensitive inputs, which validates that the method identified the important features contributing to the NN output. As expected, removing the important variables results in a degraded AUC score. Taking the discussions in chapter 2 into consideration, the input space could be reduced from 30 to 6 dimensions without a loss of performance but with a massively reduced complexity of the input space validation. The sensitivity analysis reveals that the NN function has a significant dependency on only 2 % of the 495 subspaces up to the second order, which allows to perform a precisely targeted verification of the statistical model enabling a robust analysis strategy. The first variable related to the azimuth appears at rank 82 with a score of 12 % with respect to the first rank.

Figure 3.7 visualizes the placement of a subset of variables in the ranking. The figure shows the cumulated count of a variable contributing to a sensitivity score from the first to the last rank. Following, a rise early on indicates that the variable is often present in the first ranks and therefore the uppermost line refers to the variable with most impact. A normalized AUC is computed and shown in the legend, performing an efficient aggregation of the information in the ranking with respect to the variable importance. The AUC scores show that well known quantities contribute most to the separation of the signal of the Higgs boson from the background processes, for example, the reconstructed mass of the Higgs boson (`DER_mass_vis` and `DER_mass_MCC`) or the invariant mass of the dijet system. On the contrary, variables associated with the azimuth contribute very little to the separation power of the NN because collider events are invariant in the transverse plane.

To provide additional detail, figure 3.8 shows the marginal distributions of the variables,

Table 3.2: The results of the sensitivity analysis up to the second order with the first and last 20 ranks ordered by sensitivity and the corresponding contributing variables. It should be noted that a single variable in a row indicates a sensitivity score from a first order Taylor coefficient whereas all others are derived from second order coefficients.

Rank	Variables		$\langle t_i \rangle \times 10^{-3}$
1	DER_mass_vis	DER_pt_ratio_lep_tau	3.061
2	DER_deltar_tau_lep	DER_mass_vis	2.852
3	DER_mass_vis	PRI_lep_pt	2.722
4	DER_deltar_tau_lep	DER_pt_ratio_lep_tau	2.318
5	DER_pt_ratio_lep_tau	PRI_lep_pt	2.182
6	DER_mass_vis	DER_mass_vis	2.144
7	DER_mass_MMC	DER_mass_vis	2.056
8	DER_deltar_tau_lep	PRI_lep_pt	2.023
9	DER_mass_jet_jet	DER_mass_vis	1.837
10	DER_mass_vis		1.806
11	DER_mass_MMC	DER_pt_ratio_lep_tau	1.539
12	DER_mass_transverse_met_lep	DER_mass_vis	1.478
13	DER_mass_jet_jet	DER_pt_ratio_lep_tau	1.447
14	DER_deltar_tau_lep	DER_mass_MMC	1.446
15	DER_pt_ratio_lep_tau		1.443
16	DER_mass_MMC	PRI_lep_pt	1.438
17	DER_deltar_tau_lep	DER_mass_jet_jet	1.366
18	DER_deltar_tau_lep		1.355
19	DER_mass_jet_jet	PRI_lep_pt	1.337
20	DER_mass_MMC	DER_mass_MMC	1.312
	...		
476	PRI_tau_eta	PRI_tau_phi	0.020
477	PRI_jet_leading_pt	PRI_met_phi	0.019
478	PRI_jet_leading_eta	PRI_jet_subleading_eta	0.019
479	PRI_jet_leading_eta	PRI_lep_phi	0.019
480	PRI_jet_subleading_phi	PRI_lep_phi	0.019
481	DER_sum_pt	PRI_tau_phi	0.019
482	DER_sum_pt	PRI_met_phi	0.019
483	PRI_jet_num	PRI_met_phi	0.018
484	DER_prodeta_jet_jet	PRI_met_phi	0.018
485	PRI_lep_eta	PRI_met_phi	0.018
486	DER_pt_tot	PRI_met_phi	0.018
487	PRI_jet_subleading_phi	PRI_met_phi	0.018
488	DER_sum_pt	PRI_tau_eta	0.018
489	PRI_lep_eta	PRI_tau_phi	0.018
490	PRI_jet_num	PRI_lep_phi	0.017
491	PRI_jet_leading_eta	PRI_lep_eta	0.017
492	PRI_jet_leading_eta	PRI_met_phi	0.016
493	PRI_met_phi	PRI_tau_eta	0.015
494	PRI_lep_phi	PRI_tau_eta	0.015
495	PRI_jet_leading_eta	PRI_tau_phi	0.014

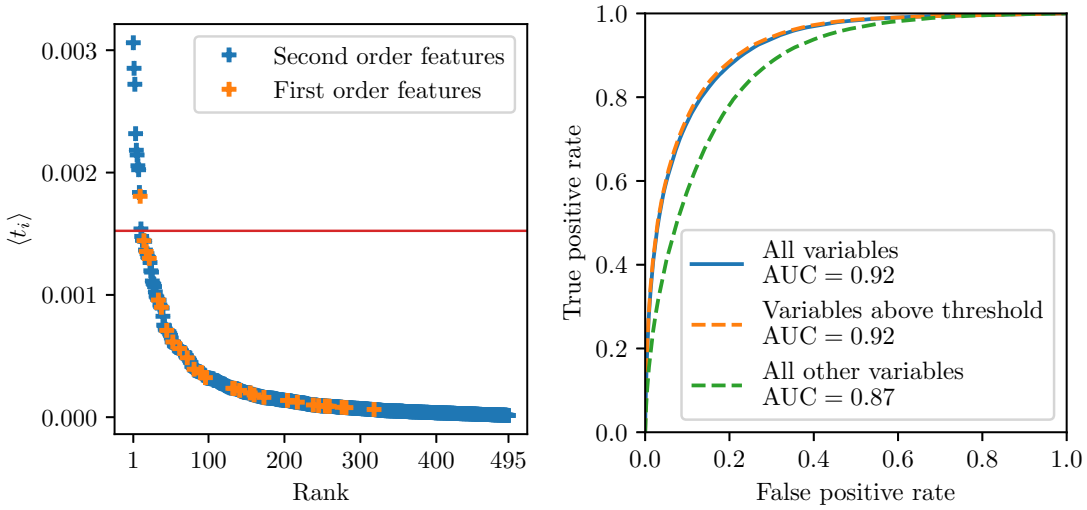

Figure 3.6: The left figure visualizes the distribution of the ranked first and second order features with a steep fall of the respective sensitivity scores. The threshold at 1.523×10^{-3} is put in the middle of the largest and smallest score and used to select the most influential variables. The right hand side shows the ROCs for the training with all variables, only the variables contributing to the ranks above the threshold and all other variables, successfully validating that the sensitivity analysis has identified the most important features.

which contribute to the first 11 ranks in table 3.2 and the according two dimensional distributions of their combinations. Each subfigure is labeled with the respective placement in the ranking, which allows to relate the sensitivity of the NN function to features in these spaces. It should be noted that the figure shows the input variables in the transformed space after application of the $x \to (x - \mu)/\sigma$ transformation, such as the inputs are given to the NN function. Especially in more complex learning tasks, the measured sensitivity of the NN output does not have to map strictly to a striking feature in the input space because there is no guarantee that the NN has learned the perfect classifier which reflects exactly the likelihood for finding signal or background. Therefore, analyzing the relations between table 3.2 and figure 3.8 must be carried out with caution. Nevertheless, all distributions of signal and background related to the first ranks show a clear separation of the classes, which verifies that the NN has learned useful relations to solve the classification objective. Examples which reveal prominent features that are only accessible in the two dimensional subspace can be found in the subfigures related to rank 1, 2 and 5, whereas for the presented two dimensional distributions with the lowest sensitivity scores, see 17, 19 and 32, such features are not visible. Further, the subfigure related to rank 6 is an example for the second order coefficient highlighted in the study in figure 3.4, which is sensitive to information hidden in the variance of overlain distributions.

To demonstrate the usefulness of the sensitivity analysis for the validation of the NN model, the variables `DER_mass_vis` and `DER_mass_MMC` are replaced by normal distribu-

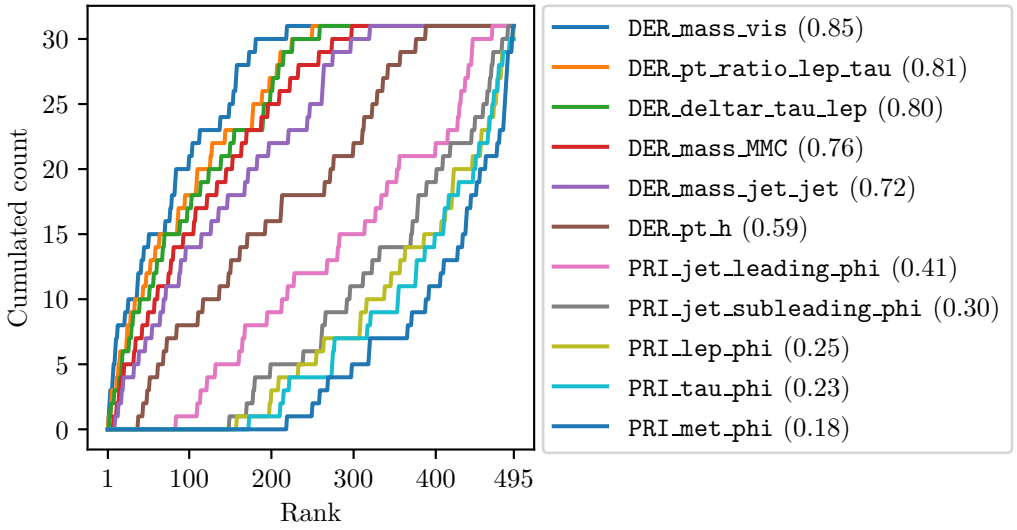

Figure 3.7: Computed is the cumulated count of appearances of the respective variable in the ranking in table 3.2 from the first to the last rank. The score in the legend represents the AUC of the cumulated count normalized to the area of the figure given by 31×494.

tions. For the background events, the values are uncorrelated, however, the signal events are fully correlated in these two variables. Such an error may happen by a mistake in the variable assignment, but is also possible due to complex issues in the simulation. An exemplary mistake is the usage of the same random seed to simulate the properties of objects, which may introduce artificial correlations. It should be noted that this artificial feature is not visible by checking the marginal distributions, only by an analysis of the two dimensional subspace. The training on this scenario gives an AUC of 0.96, outperforming the training on the unaltered data significantly but the performance is massively degraded on data without the correlation of the signal events, resulting in an AUC of 0.84. However, the sensitivity analysis for this NN shows on the first rank the two dimensional subspace of `DER_mass_vis` and `DER_mass_MMC` with a score of 3.1×10^{-2}, successfully identifying the cause of the apparent superior performance. Also, the score is one magnitude larger than others, for example see table 3.2 as reference, clearly indicating a striking feature in that subspace. The validation of the input space guided by the sensitivity analysis of the NN function is able to identify the mismodeling in the simulated events right away whereas in a generic validation strategy the subspace is only one of many.

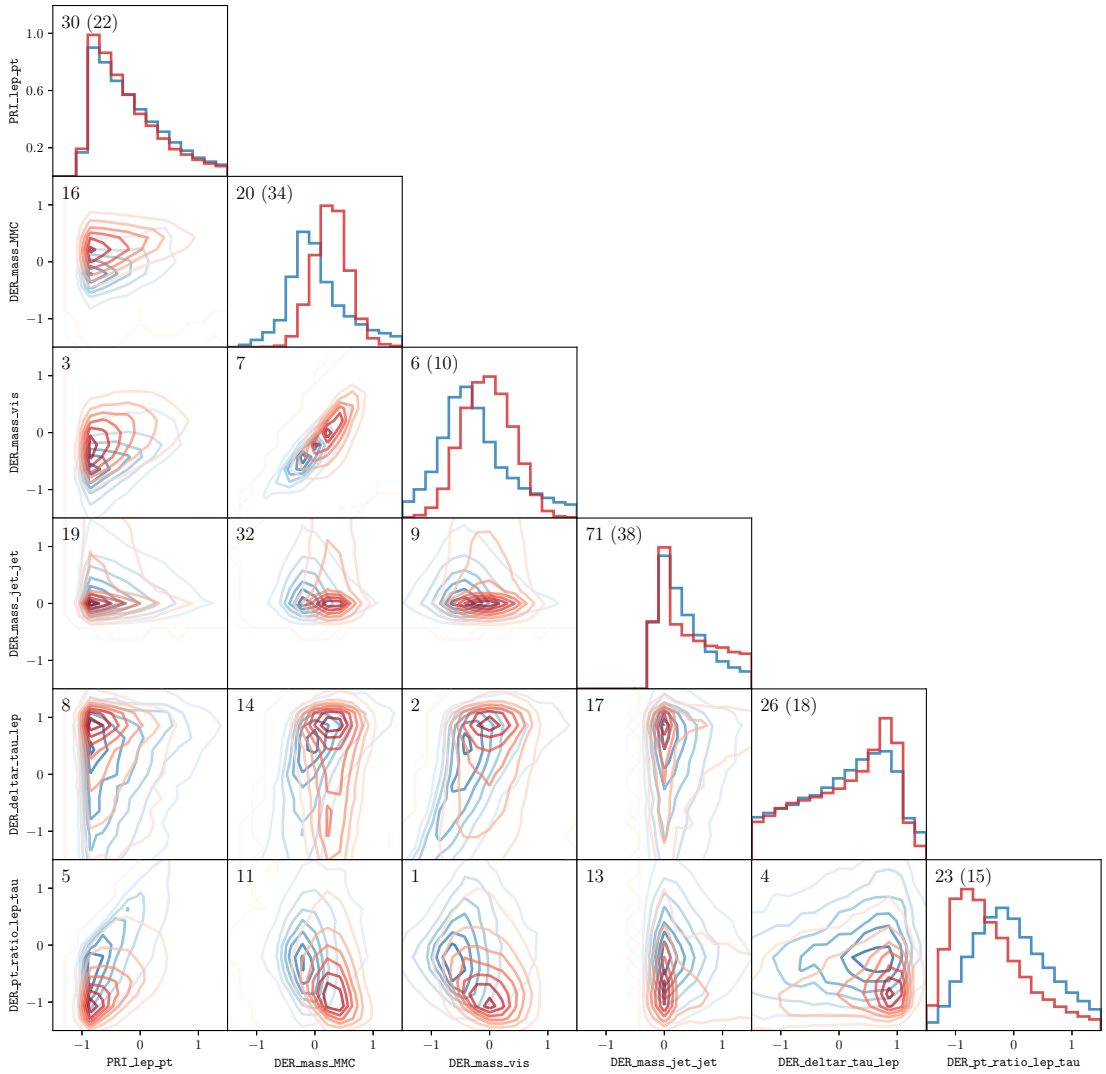

Figure 3.8: Distribution of the preprocessed input variables, which contribute to the first 11 ranks in table 3.2 shown separately for signal (red) and background (blue). The main diagonal elements of the grid show the marginal distributions and all other subfigures depict the two dimensional distribution of the respective variable pairs. The numbers indicate the placement in the ranking, whereas on the main diagonal in addition the ranking with respect to the first order coefficient is given in the brackets.

CHAPTER 4

Controlling the dependence of the machine learning model on systematic variations

This chapter discusses techniques which allow to gain a fine granular control over the dependence of the ML model on the input space being important for data analysis in HEP due to the presence of systematic uncertainties. Existing solutions are discussed and a novel approach is presented, which allows the inclusion of information about variations in the inputs space in the NN optimization, enabling robust and reliable measurements with respect to systematic uncertainties.

4.1 About the necessity of full control over the machine learning model for data analysis in high-energy particle physics

In HEP, many use cases exist for applications, which require full control over the ML model. An example from object reconstruction is jet tagging, which classifies reconstructed jets for further analysis, for example with respect to the originating particle [39] or advanced features like the jet substructure [91]. Such taggers are often required to be invariant in specific properties of the object, e.g., the tagger for the jet substructure is desired to be insensitive to the jet mass to reduce the impact of systematic uncertainties in the background modeling [92–94]. However, the classification efficiency depends on the jet mass, which is why the ML model develops the undesired dependency during the training. Without special techniques to take control over the dependencies of the ML model in a fine granular way, the only solution is the complete removal of the corresponding information from the training. This simple solution raises in turn the issue that the performance of the tagger might be heavily degraded. But more problematic is that the complete removal of the information about the jet mass is complicated by its correlations to other inputs such as the kinematic properties of the constituents. Most of the methods discussed in section 4.2 and the novel approach in section 4.3 are specialized on NNs and allow to control the dependence of the NN function on specific features in the input space, which enables an optimization of the model that ignores given information explicitly. Another use case being more closely related to the analysis

presented in chapter 2 is the application of such techniques for the construction of ML based observables being input to the statistical inference. Once the input space and the statistical model are scrutinized with the validation strategy discussed in chapter 2 and the techniques presented in chapter 3, potentially discovered mismodelings have to be treated accordingly. A thinkable scenario is that a mismodeling is successfully identified but a precise determination of the related uncertainties is not accessible, which forces the analyst to remove the respective information from the analysis in favor of a more robust measurement. A related scenario is the precise knowledge about a systematic uncertainty, which impairs the sensitivity of the measurement significantly and in consequence an improvement is expected from a less affected ML based observable, which reduces the propagation of the respective uncertainty into the measurement. Chapter 5 discusses this scenario extensively.

In summary, the availability of methods to control the dependency of the ML model on the input space is a necessity for multivariate analysis in HEP to tackle challenges posed by systematic uncertainties in the measurement. Techniques are required, which are able to manipulate the dependency of the ML model to arbitrary variations in the multidimensional input space, offering capabilities beyond the decorrelation against a single input variable. Section 4.2 gives an overview over existing solutions and discusses the challenges, which are addressed by the novel method introduced in section 4.3.

4.2 Overview over existing approaches

Frequently used for ML tasks in HEP are boosted decision trees (BDTs), which consist of decision trees whose inputs are weighted in each training iteration with respect to previous misidentifications to optimize the training objective. The reweighting during the training can be modified to take into consideration besides the misidentification rate also the uniformity of the classification with respect to a subspace of the input space [95, 96]. The technique can be used in HEP analysis to reduce the dependence of the BDT on the observables of interest, successfully demonstrated in [97] with a Dalitz plot analysis. A technical challenge is the application of the method because a custom implementation of BDTs is required whereas the following NN based techniques can be implemented with existing software validated by a large user base.

Based on NNs, the concept of adversarial neural networks (ANNs) [98] is suitable to implement the penalty to variations in the input space. The idea is to train a second NN with the objective to categorize the output of the actual classifier with respect to the variation [99]. The capability of the adversary to perform the categorization is introduced in the training objective as a minimax problem, forcing the actual classifier to be invariant with respect to the variation. The NNs are trained in an alternating manner with the trainable parameters fixed for the respective other NN. The approach has been proven to be successful in HEP applications, for example to mitigate systematic uncertainties in the search for new physics [100, 101]. However, the application of the technique is challenging due to the instability of the minimax optimization [98, 102], the large number of hyper parameters introduced by the second NN and the choices for

the alternating training procedure. Further, the variations in the input space must be available as modified datasets and cannot be given in the form of statistical weights, which restricts the use cases in HEP substantially. Encoding the information about the variations with statistical weights reduces the complexity of an analysis significantly because an analysis in HEP typically processes a large amount of data, which makes the reprocessing of modified datasets computationally expensive.

For these reasons, methods are desired with a smaller complexity than an approach with ANNs, which also support the description of variations in the input space with statistical weights. A solution satisfying these requirements is proposed in [103]. It solves the task by introducing a penalty term in the training objective, which reflects a measure of correlation between the NN output and the variation of interest. The method is shown to achieve similar performance to an ANNs at the cost of a single additional hyperparameter and improved convergence properties. Further, section 4.3 presents a novel technique with similar properties based on the approximation of counts and histograms, which are used to penalize the dependence of the NN response on variations in the input space.

Another group of methods implement the decorrelation to systematic uncertainties implicitly by optimizing the analysis objective instead of including a penalty on a specific variation in the input space. The analysis objective can be implemented with an approximation of the signal significance [104, 105] or with the variance of the signal strength based on the likelihood [106, 107]. Including systematic uncertainties in the objective results in an optimization, which reduces the dependence on the systematic variations in the input space if the objective improves with the mitigation. It should be noted that the usefulness of an uncertainty mitigation with respect to the analysis objective is a priori not known. However, the techniques based on the optimization of the analysis objective include this tradeoff naturally in the training whereas methods with an explicit penalty on the systematic variation must introduce a new hyperparameter. The number of introduced hyperparameters are not manageable for complex analyses with hundreds of systematic uncertainties, restricting the use case of methods, which penalize systematic variations explicitly. Chapter 5 is dedicated to the optimization of the ML model on the analysis objective and discusses the challenges and solutions.

4.3 Controlling the dependence of the neural network function on systematic variations in the multidimensional input space

This section introduces a novel method to control the dependence of the NN function to variations in the input space, for example being useful for analysis in HEP to decorrelate against physical observables or to reduce the impact of systematic uncertainties in the measurement. A dedicated paper about this method was published in [108].

4.3.1 Method

The goal of the new method described in this section is the construction of a loss function, which penalizes the dependence of the NN function $f(\boldsymbol{x})$ on variations $\boldsymbol{x} + \boldsymbol{\Delta}$ from the nominal input space $\boldsymbol{x}$. The varied inputs $\boldsymbol{x} + \boldsymbol{\Delta}_{\text{up,down}}$ can represent systematic shifts, which define the systematic uncertainties in the statistical model in order to decorrelate against the respective uncertainty, see section 2.5 for details. Another possible scenario in HEP analysis is using different mass hypothesis for the definition of the variation $\boldsymbol{\Delta}_{\text{mass}} = \boldsymbol{x}_{\text{mass}_1} - \boldsymbol{x}_{\text{mass}_2}$ to suppress any kinematic bias in the measurement due to the sensitivity of the NN to the observable.

Since one of the requirements for the method is the support of variations described by statistical weights, the typical procedure to compare $f(\boldsymbol{x})$ to $f(\boldsymbol{x} + \boldsymbol{\Delta})$ are counts and histograms. However, the mathematical operation of a count is not differentiable on the bin edges and zero otherwise, which prohibits the computation of the analytic gradient with automatic differentiation. Therefore, this method uses an approximation for the count operation with a valid gradient. A suitable approximation is a Gaussian function $\mathcal{G}$ with the standard deviation set to the half width of the bin and normalized to $\max(\mathcal{G}) = 1$. Figure 4.1 visualizes the setup in one dimension, but a similar setup is possible with a multivariate Gaussian for multidimensional histograms. The gradient of $\mathcal{G}$ is well defined, peaking at the bin boundaries and zero at the bin center, which encodes for the gradient with respect to the trainable parameters a large change for events on the boundaries with a decreasing strength towards the center and in a distance to the boundaries. It should be noted that the overall scale of the gradient is not of importance since the gradient is scaled by the learning rate, which is a hyperparameter of the training and typically dynamically adapted by the optimizer algorithm [64].

For N examples in the dataset $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, the count operation translates to $\mathcal{N}_k = \sum_{i=1}^{N} \mathcal{G}_k\left(f\left(\boldsymbol{x}_i\right)\right)$, given the bins $\boldsymbol{k}$ in the histogram of the NN output $f(\boldsymbol{x})$. The blurred representation of the histogram can now be used to construct a mathematical expression, which penalizes the variation of the NN with respect to the variations $\boldsymbol{\Delta}$. Chosen for this method is the squared relative change of $\mathcal{N}_k$ with respect to the variation $\boldsymbol{\Delta}$ given by

$$\Lambda\left(\boldsymbol{x}, \boldsymbol{\Delta}\right) = \frac{1}{n_k} \sum_{k} \left(\frac{\mathcal{N}_k\left(f\left(\boldsymbol{x}\right)\right) - \mathcal{N}_k\left(f\left(\boldsymbol{x} + \boldsymbol{\Delta}\right)\right)}{\mathcal{N}_k\left(f\left(\boldsymbol{x}\right)\right)} \right)^2 \tag{4.1}$$

with the number of bins n_k.

This formulation for the similarity of $f(\boldsymbol{x})$ and $f(\boldsymbol{x} + \boldsymbol{\Delta})$ can be appended to the actual training objective, for example a classification task using the CE function, which results in the loss function

$$L_\Lambda = L_{\text{CE}} + \lambda\Lambda \tag{4.2}$$

with the hyperparamter λ controlling the strength of the penalty.

4.3.2 Application on a simple example based on pseudo experiments

This section presents the proposed method using a simple example based on pseudo experiments. The input space of the example is shown in figure 4.2, spanned by the

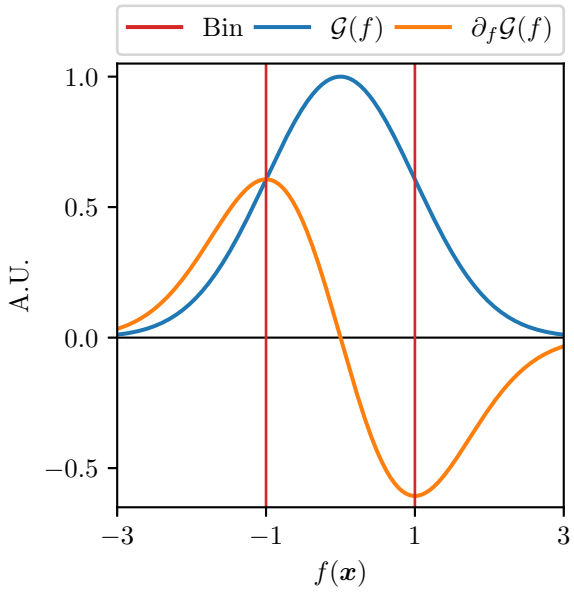

Figure 4.1: Approximation of the bin function using a Gaussian function $\mathcal{G}$ with the standard deviation set to the half width of the bin and normalized to $\max(\mathcal{G}) = 1$.

variables x_1 and x_2, and is populated by the two processes signal and background. In addition, variations from the nominal distribution are introduced for the background process, given by two discrete shifts $x_2 \pm 1$ of the mean. These variations can be interpreted as the $\pm 1\sigma$ variations of a systematic uncertainty in context of a statistical model such as described in section 2.5. It should be noted that the scenario is deliberately the same as discussed in [99], which proposes the usage of ANNs to solve the same task.

The training task without consideration of the variations is the classification of the signal and background process optimized with the CE function. The NN consists of two hidden layers with 200 nodes each and ReLU activations [89]. The output layer has a single node and the sigmoid activation function. The trainable parameters are optimized on 5×10^4 examples, a batch size of 10^3 per gradient step and the Adam algorithm [64]. Once the training has not improved for five epochs in a row on a statistically independent dataset of the same size, the training is stopped. Eventually, all results are evaluated on a separate dataset with 10^5 examples. For the construction of the penalty term Λ, ten equidistant bins are used in the range $[0, 1]$ of the NN output and the hyperparameter λ is set to 20. The variations are implemented with modified versions of the nominal dataset, but the usage of statistical weights for the same purpose is demonstrated in section 4.3.3.

The impact of the additional term Λ in the loss function is visualized in figures 4.3 and 4.4. Figure 4.3 shows the distribution of the NN output f for a training with only the CE loss and the same distribution if the penalty term Λ is used in addition. The change of the NN output with respect to these variations is minimized for the NN function f_{L_Λ},

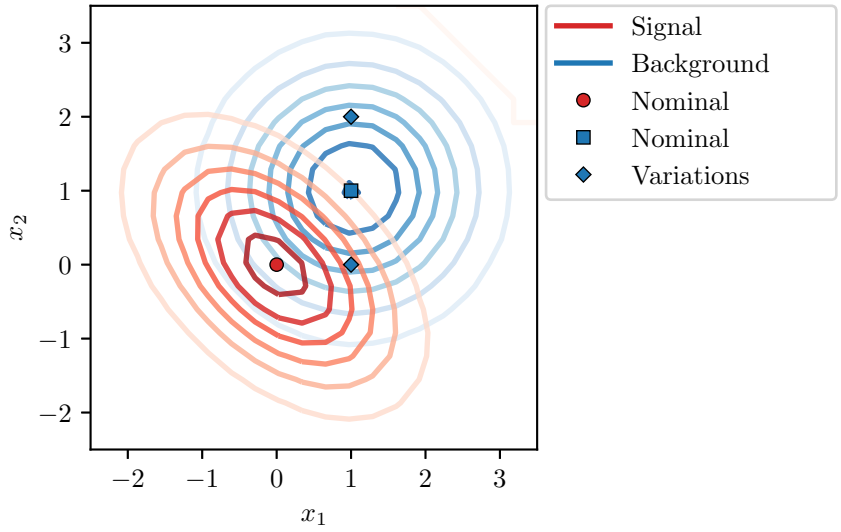

Figure 4.2: Simple example based on pseudo experiments with the input space spanned by the variables x_1 and x_2. The training task is the classification of the classes signal and background with the means $(0\,0)$ and $(1\,1)$, and the covariance matrices $\begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, respectively. The proposed method to control the dependence of the NN function on the input space is studied using variations of the mean of the background distribution with $x_2 \pm 1$.

whereas $f_{L_{\text{CE}}}$ is strongly influenced by the variation of x_2. The successful decorrelation against the variation is visible in figure 4.4, which shows the distribution of the NN output in the input space. For the training including the additional term Λ in the optimization, the surface of the NN function is aligned with the direction of the shifts, which results in a robust classifier with respect to the variations. Translated to the physics use case, the NN function avoids the propagation of the systematic uncertainty to the output, which provides a robust observable for a measurement. The figures can be compared to the distributions in [99] using an ANN, which show a similar outcome.

A suitable metric to measure the impact of the decorrelation with respect to the classification task is the AUC of the ROC, which is evaluated on the nominal dataset and the instances with the variations. The ROCs on the dedicated datasets measure the dependence of the classification on the actual instantiation of the variation in data. Figure 4.5 shows large fluctuations in the AUC for the NN solely trained on the CE loss, whereas the NN trained with the additional penalty term Λ has the same performance in all cases but at the cost of a reduced separation power.

The impact of the hyperparamter λ is studied in figure 4.6. The figure shows for different λ parameters the resulting classification boundary at $f(\boldsymbol{x}) = 0.5$. Setting $\lambda = 20$, such as used in the discussion before, fully decorrelates against the variations, but the method also allows to control the dependence in a fine granular way. For example $\lambda = 0.1$ shows just a slight rotation of the classification boundary. The NN function with $\lambda = 0.5$ rotates the classification boundary further towards the fully decorrelated case, though the boundary

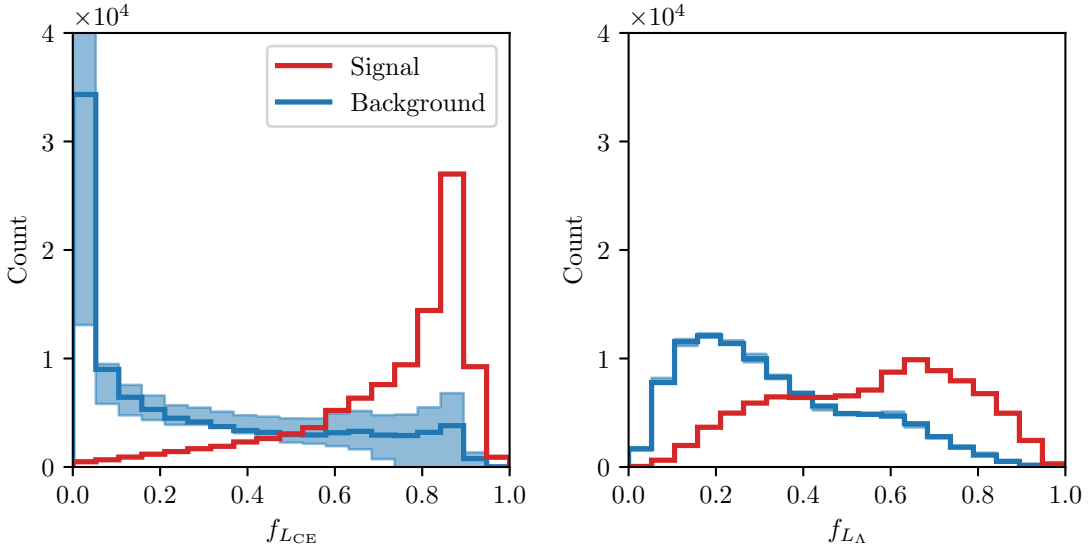

Figure 4.3: Output distributions of the NN function f optimized with only the CE loss (left) and the penalty terms Λ in addition (right). The band around the nominal distribution of the background process shows the change of the NN response with respect to the introduced variations.

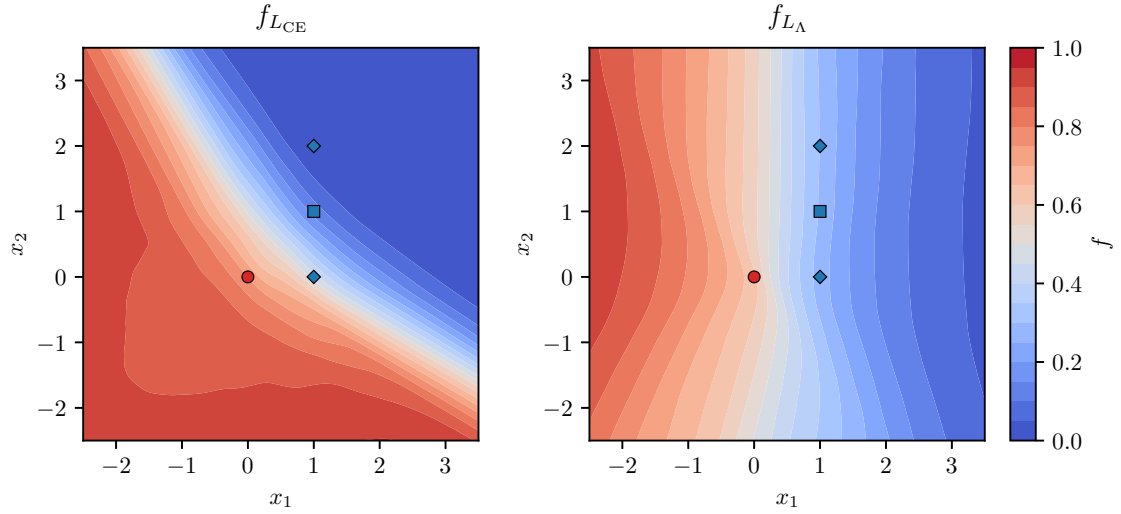

Figure 4.4: Distribution of the NN function in the input space for the training using only the CE loss (left) and the penalty term Λ in addition (right). The markers show the means of the nominal distributions and variations, such as labeled in figure 4.2.

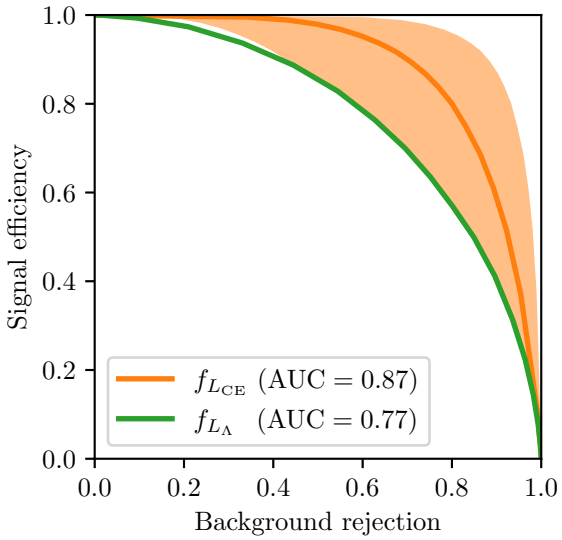

Figure 4.5: The ROCs are evaluated for the nominal dataset and separate datasets with the variations, shown as the colored bands, with the NN only trained on the CE loss compared to a training with the penalty term Λ applied in addition. The NN trained with the penalty term has the same performance on all datasets.

has a kink at $(-0.5, 2)$. This is explained by the mathematical constraints implemented by the loss function with no strong penalty in phase spaces with little population of any process. Therefore, the NN function shows in these regions effects from the weight initialization and the statistical component of the simulated dataset. $\lambda = 5$ shows again an increase in the decorrelation and is explicitly included to discuss another feature due to the mathematical formulation of the loss function. Since the additional term Λ consists of two components for the two variations and the fact that the optimization of a NN rarely converges to a global minimum, only a weak penalty exists to prevent an unequal decorrelation of multiple variations. The equal optimization of the two variations could be strongly enforced by adding an additional term in the loss function, for example $(\Lambda(\Delta_{\mathrm{up}}) - \Lambda(\Delta_{\mathrm{down}}))^2$, but is also strengthened naturally with a large value for λ towards the full decorrelation.

4.3.3 Application on an example from high energy particle physics

This section discusses the proposed method in context of a more complex example from HEP. Similar to section 3.3.4, the dataset of the Higgs ML challenge is used. Details about the dataset can be found in the referenced section and in [90]. In summary, the dataset consists of simulated events described by 30 variables, which are used to classify the events either as signal originating from a SM Higgs boson decaying into a lepton and a hadronic tau or as background from a mixture of processes with a similar signature. In addition to the original dataset, a systematic uncertainty is introduced resembling the finite accuracy

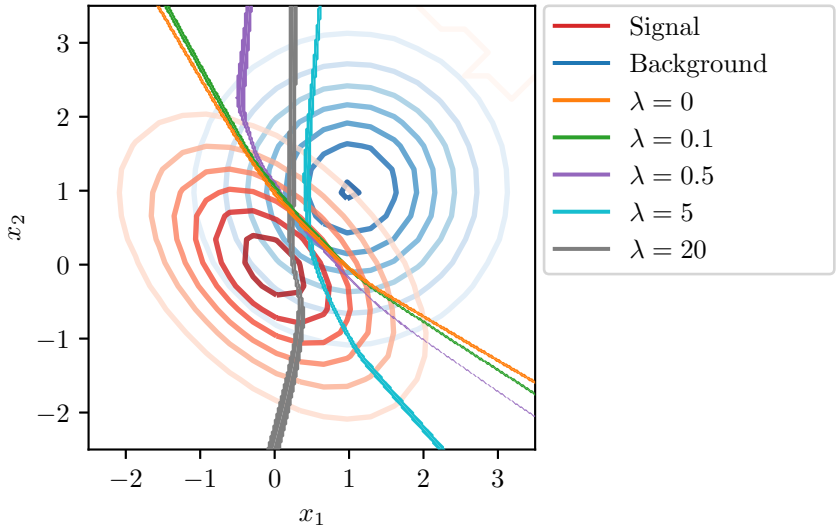

Figure 4.6: The classification boundary of $f(\boldsymbol{x})$ in the range $[0.48, 0.52]$ is visualized in the input space for various values of the hyperparameter λ. The boundaries overlay the distributions of the signal and background classes in the nominal case. Figure 4.2 shows the input space with the variations.

for the reconstruction of tau leptons in the detector. The example assumes a typical 3 % uncertainty on the tau momentum p_{T}^{τ} [109], which is named `PRI_tau_pt` in the dataset. The $\pm 1\sigma$ shifts are implemented with the transformation $\texttt{PRI\_tau\_pt} \cdot (1 \pm 0.03)$ using statistical weights. The propagation of the systematic uncertainty on correlated variables such as the invariant ditau mass $m_{\tau\tau}$ (`DER_mass_MMC`) and the missing transverse energy $\not{E}_{\mathrm{T}}$ (`PRI_met`) is visible in figure 4.7. To model correctly the migration effects in and out of the event selection, only events with $p_{\mathrm{T}}^{\tau} > 25\,\mathrm{GeV}$ are used for the NN training and the results, which is well above the 3 % variation of the minimal p_{T}^{τ} of 20 GeV in the original dataset.

The NN architecture is the same as for the simple example in section 4.3.2. The training uses 75 % of the training dataset for the optimization of the trainable parameters with the Adam algorithm [64] and a batch size of 10^3. If the loss does not decrease for ten epochs in a row on the remaining dataset, the training is stopped. The loss function consists of the CE function and the additional penalty term Λ with the hyperparameters $\lambda = 20$ and 20 equidistant bins in the range $[0, 1]$ of the NN output.

Figure 4.8 shows the distribution of the NN output for the training solely on the CE function and the penalty term Λ in addition. The bands around the nominal distributions visualize the uncertainty of the NN output caused by the introduced systematic uncertainty on p_{T}^{τ}. A decorrelation of the NN response with respect to the systematic uncertainty can be observed for the signal process. However, the distribution of the background events has an unchanged uncertainty band, which is explained by the dominant normalization uncertainty for this process. Easily visible in figure 4.7, the background

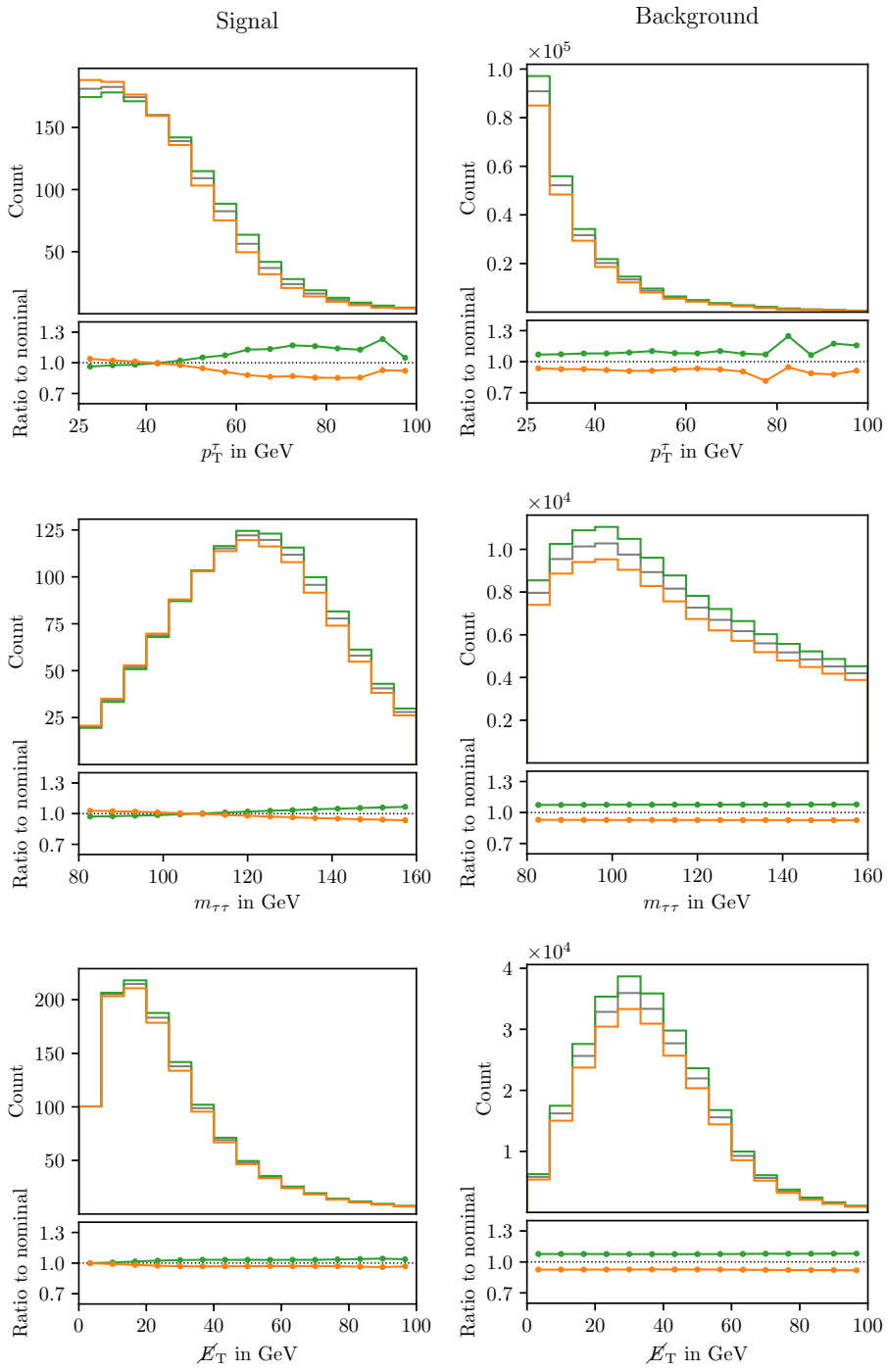

Figure 4.7: Example distributions of variables in the Higgs ML dataset for the signal (left) and background (right) classes. The ratio below each figure compares the nominal distribution (grey) to the $\pm 1\sigma$ up (green) and down (orange) variations describing the systematic uncertainty on p_{T}^{τ}.

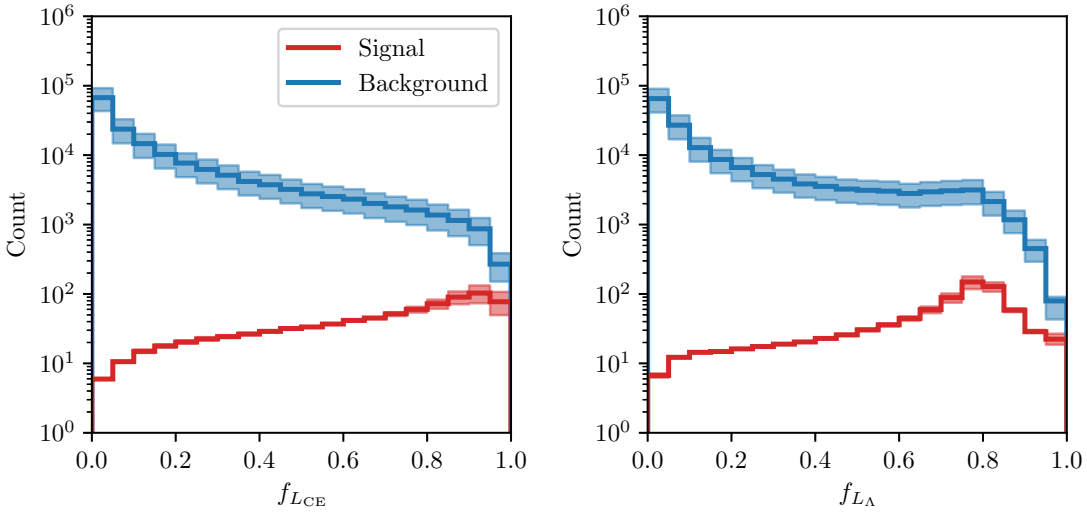

Figure 4.8: Distribution of the NN output for the NN function trained solely on the CE function (left) and the penalty term Λ in addition (right). The band around the nominal distributions are enlarged by a factor of 5 for improved visibility and show the change of the NN response due to the introduced systematic uncertainty on p_{T}^{τ}.

class has a steeply falling distribution directly at the boundary of the analysis selection at $p_{\mathrm{T}}^{\tau} = 25\,\mathrm{GeV}$, which causes a large number of events to migrate in and out of the selection. On the contrary, the signal events peak with a distance to the boundary, which results in a more pronounced shape altering effect of the systematic uncertainty. A crucial detail about such decorrelation techniques is the fact that it is impossible to decorrelate against normalization uncertainties since such variations cannot be mitigated by construction. Therefore, the effect of the penalty term Λ in this example is expected to be mainly visible for the signal process, which is the case in figure 4.8.

The effect of the decorrelation with the penalty term Λ can be more precisely studied with a fit of the statistical model, which uses the NN output as observable. The construction of the model follows the description in section 2.5 with an Asimov dataset, but introduces two nuisance parameters η_{norm} and η_{shape} for the systematic uncertainty on p_{T}^{τ}, which factorize the effects on the normalization and the shape. The normalization effect is found to be 2.2 % and 7.6 % for the signal and background process, respectively. It should be noted that this setup with two independent parameters is not fully correct with respect the originally introduced systematic uncertainty, but allows to estimate precisely the effect of Λ on the normalization and shape by studying the correlations of the parameters η_{norm} and η_{shape} with the signal strength μ. Both statistical models, using $f_{L_{\mathrm{CE}}}$ and f_{L_Λ} as observable, have a correlation of η_{norm} with μ of 35 %. However, the parameter η_{shape} has a correlation of 55 % with μ in the case of $f_{L_{\mathrm{CE}}}$ but just 5 % for f_{L_Λ}, which validates the successful decorrelation of the NN function using the penalty term Λ.

The resulting significance of the signal process is shown in figure 4.9 for various values of

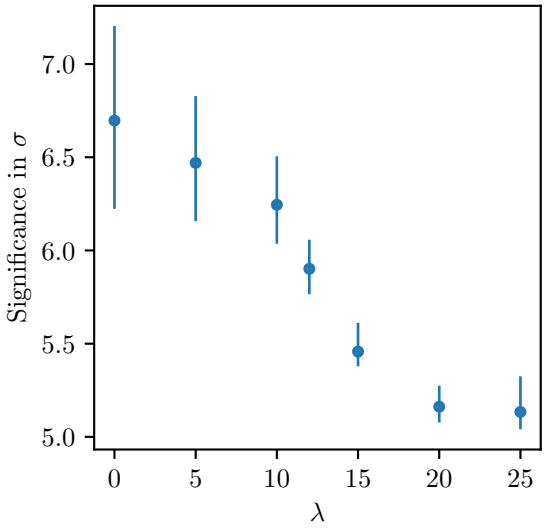

Figure 4.9: Significance of the signal process with respect to various values of λ. The points with $\lambda = 0$ and $\lambda = 20$ correspond to the two cases shown in figure 4.8. The bars are enlarged with a factor of five for improved visibility.

λ. In contrast to the previous discussion, these results are retrieved with a single nuisance parameter representing the systematic uncertainty. The bars visualize the outcome for the significance if the statistical inference is applied on an Asimov dataset with the $\pm 1\sigma$ variation in the data compared to the nominal expectation indicated by the marker. The statistical model itself stays the same for all experiments. A tradeoff is clearly visible between the nominal significance and the variation thereof with the $\pm 1\sigma$ instantiations of the systematic variation in data. The highest significance of $6.7\,\sigma$ is continuously reduced with increasing values of λ due to the removal of information from p_{T}^{τ}, falling to $5.2\,\sigma$ at $\lambda = 20$. But the reduced nominal significance comes with an increase in the robustness of the measurement. The variation of the significance falls from $7.5\,\%$ at $\lambda = 0$ to $1.8\,\%$ at $\lambda = 20$, successfully improving the reliability of the result with respect to the systematic uncertainty.

To visualize the event selection performed by the NN, the upper panel in figure 4.10 shows the distribution of p_{T}^{τ} inclusively and in a signal enriched region. The signal enriched subsets are obtained from events with a NN output larger than 0.7, see figure 4.8 for reference. The distributions show for $f_{L_{\mathrm{CE}}}$ large sculpting effects, especially for the background process, which follows a signal like distribution. However, the distributions for $f_{L_{\Lambda}}$ are close to the inclusive distributions, which validates that the NN has neglected the information from the variable p_{T}^{τ} to separate signal and background. In the lower panel in figure 4.10, the procedure is repeated for the reconstructed invariant mass of the Higgs boson $m_{\tau\tau}$. For this variable a strong dependence of the NN output with respect to $m_{\tau\tau}$ is visible for both trainings. The distributions for signal and background related to $f_{L_{\mathrm{CE}}}$ are very similar in the signal enriched region, which shows that no residual information

for the separation of the classes is left. But the same distributions in the signal enriched region for f_{L_Λ} can still be used to separate signal and background, although also a clear separation to the inclusive background distribution is visible. This is explained by the shared information between p_{T}^{τ} and $m_{\tau\tau}$ being removed from the training by the penalty term Λ, which enables the optimization to rely only on partial information of $m_{\tau\tau}$.

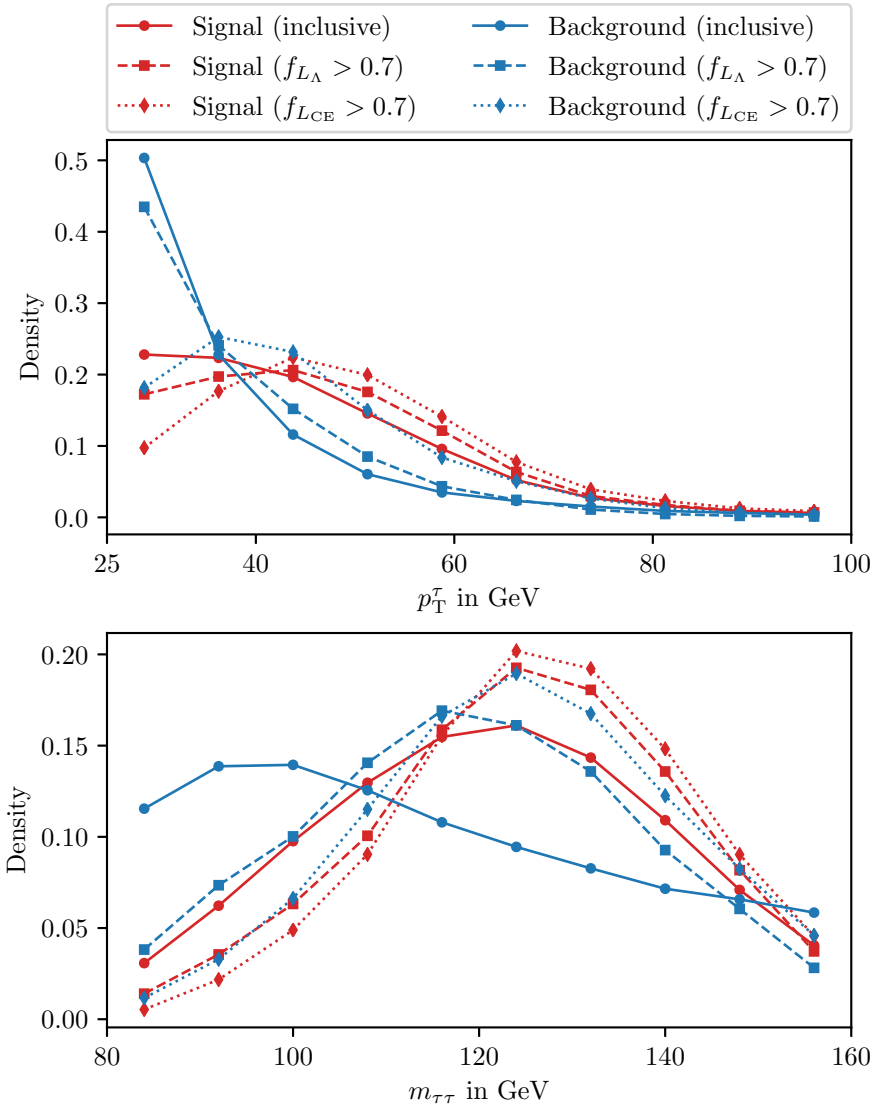

Figure 4.10: Distributions of p_T^τ and $m_{\tau\tau}$ for the signal and background class, shown inclusively and in a signal enriched region defined by the NN output being greater than 0.7.

CHAPTER 5

Optimal statistical inference with model optimization based on likelihood information

This chapter studies novel strategies for data analysis in HEP. Typical approaches used in current HEP analyses are analyzed in order to identify the insufficiencies, which lead to suboptimal results. Finally, a novel approach is introduced, which is based on the optimization of a ML model using likelihood information and allows to find an optimal observable in consideration of all statistical and systematic uncertainties. The capability of the new method to perform data analysis with the statistically optimal sensitivity is demonstrated with a simple example based on pseudo experiments and the suitability of the approach in practise is shown with a more complex example in the context of HEP.

5.1 About the efficiency of data analysis in high energy particle physics

A typical task of data analysis in HEP is the inference of the presence or absence of a physical process in data by performing a search which may result in a discovery, or the measurement of the cross section once the process is established. Most analyses, such as those performed by the CMS and ATLAS collaborations at the LHC, solve these analysis tasks by following the cut and count approach, which compares in selected spaces of the dataset the compatibility of observed counts with the signal and background hypotheses. The simple cut and count strategy has improved over the decades by using histograms instead of single counts and a superior performance has been achieved with the usage of multivariate analysis techniques such as BDTs and NNs for the event selection. This strategy for data analysis in HEP is established since the early 2000s and led to the discovery of the Higgs boson in 2012 [12, 13]. Further improvements have been observed by using the output of such ML methods as the observable in the statistical inference instead of using physically motivated quantities [44, 110–112]. The analysis described in chapter 2 is one of this kind and shows an improved sensitivity compared to a cut based strategy [113].

While analyzing the development of data analysis in HEP in the past, the question rises why improved strategies were found repeatedly over years and the naturally following

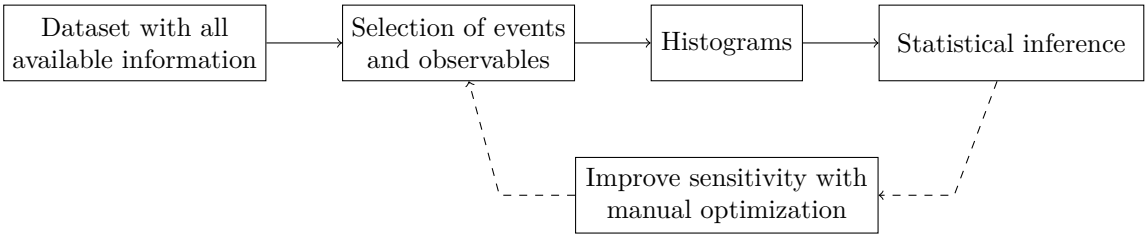

Figure 5.1: Overview over the typical strategy for data analysis in HEP, for example followed by most analyses of the Higgs boson carried out by the CMS and ATLAS collaborations at the LHC.

question is whether there is at the end an optimal strategy for data analysis in HEP. These questions can be addressed by analyzing the typical data analysis strategy, visualized in figure 5.1. Assuming that the data analysis is based on an initial dataset containing all available information, the statistically optimal result could be achieved by performing the statistical inference directly on this high dimensional dataset. Powerful statistical tools are available with the Neyman Pearson lemma [53] and the theorem by Wilks and Wald [54, 55], which promise an asymptotically optimal result. However, the key of these statistical methods is the likelihood function, which is not available in this high dimensional input space for analysis in HEP. Because the expectation in the statistical model is derived with simulation, the complexity of today's experiments and the curse of dimensionality makes the derivation of probability distributions unfeasible in high dimensional spaces. Therefore, the typical analysis strategy is heavily based on the reduction of the dimensionality so that the statistical methods can be applied on a low dimensional dataset. After the event selection, which picks the events to be considered in the statistical inference, the first step of the dimensionality reduction is the selection or crafting of sensitive observables. Driven by years of experience and detailed knowledge about the detector and the underlying theory, the observables are designed to contain as much information as possible about the process of interest. Besides the usage of high level observables, the complexity of the problem is reduced further by summarizing the information with histograms. Histograms are especially suitable for the dimensionality reduction because the probability distribution of counts is well known and makes the likelihood function easily accessible. At the end, the complexity of the initial dataset containing approximately $\mathcal{O}(10^7)$ events described by $\mathcal{O}(100)$ variables is reduced to $\mathcal{O}(10)$ histograms representing the probability density of $\mathcal{O}(1)$ high level observables. The optimization of the analysis is typically carried out manually by studying the impact of the free parameters in the analysis procedure, e.g., the choice of the observable or the number of bins, on the result of the statistical inference. Since analyses in HEP are build up blindly to the actual data using an artificial Asimov dataset [58], the iterative optimization of the free parameters does not bias the result.

All inefficiencies of the previously described analysis strategy stem from the dimensionality reduction. The selected high level observables may not carry all information available in the initial dataset and also summarizing the information with histograms

is subject to information loss. These inefficiencies can be addressed with more powerful high level observables, for example the output of a NN, and a more fine granular binning for the histograms. However, it should be noted that the optimal sensitivity of an analysis is a priori not known and therefore no one can tell whether the current best analysis strategy could still be significantly improved.

In addition to the challenges posed by the dimensionality reduction, data analysis in HEP is subject to systematic uncertainties, which has direct implications on the search for the optimal observables. An observable could be optimal with respect to the properties predicted by theory but the experimental uncertainties may degrade the sensitivity considerably. The number of systematic uncertainties, which contribute to a typical HEP analysis, is in $\mathcal{O}(100)$, turning the task to find the best possible observable highly challenging.

The previously discussed reasons contribute to the fact that no definite solution for data analysis in HEP is established. Data analysis in HEP is very complex and offers a large parameter space to optimize the specific analysis objective. The manual optimization of these parameters increases the required time to develop such an analysis substantially and does not promise an optimal solution. To tackle these conceptual issues, the following sections discuss modern approaches for data analysis in HEP, which promise asymptotically optimal results. A novel analysis strategy is proposed, which uses modern ML techniques to optimize the free parameters of the system directly on the result of the statistical inference, while taking into account all statistical and systematic uncertainties. The proposed approach is not subject to excessive manual optimization, yielding better and faster results from analyses in HEP. Section 5.2 gives an overview over the current literature and existing solutions, which is followed by section 5.3 introducing a novel approach to achieve the previously discussed goals. The novel method is studied in section 5.3.2 using a simple example with pseudo experiments and in section 5.3.3 with a more complex example from HEP.

5.2 Overview over existing approaches

Various methods have been developed in the HEP community to improve the data analysis procedure compared to the strategy discussed in section 5.1.

For the discovery of the Higgs boson, an innovative strategy was followed in the analysis of the decay $H \to ZZ \to 4l$ [12, 114]. The analysis strategy uses the matrix element method [59, 115, 116], also known as MELA, to craft a powerful observable given by the likelihood ratio, which reflects the probability that the observed kinematic configuration of the two Z bosons originates from the signal process or the dominant background $\mathrm{ZZ}/\mathrm{Z}\gamma^*$. As known from Neyman Pearson [53], the likelihood ratio is the most powerful discriminator and therefore the dimensionality reduction from the kinematic properties of the reconstructed Z bosons to the scalar response of the method is in theory optimal. The observable is only optimal in theory because the detector effects have to be taken into account, which can be done either with analytic approximations or by simulation. The first solution affects the performance of the observable whereas the second solution

requires significant computing resources. Another challenge is the computation of the likelihood ratio, which is not trivial and therefore not available for all physical processes. Further, the observable is only optimal with respect to the kinematic properties of the decay products but does not include additional information such as the discriminators for the object identification or other algorithms being part of the reconstruction. Although the MELA approach is for these reasons in practise not optimal, the method provides a powerful discriminator, which also has been exploited in HEP analyses as input to subsequent ML methods, for example in [111] and in the analysis presented in chapter 2.

While the matrix element method computes the likelihood ratio directly using the underlying theoretical model, another group of methods performs this task with information from simulation. The advantage is the correct modeling of the detector response and the possibility to include all available information, which promises a more accurate and hence more powerful estimate of the likelihood ratio. In HEP, the likelihood in the input space can not be derived analytically, but the simulation is able to generate events from the probability distribution of the respective physics model. ML methods such as NNs can be used to learn the likelihood function, which then serves as an approximation of the intractable analytic form. First discussed in [117], this approach has been adopted by a group of methods [118–121], which provide an optimal projection of the information in the input space to a single observable. A challenge is the technical infrastructure to carry out the training, which is addressed by [122] but optimized for phenomenological studies. Also, the methods require a significant amount of simulated events, which poses a computational challenge for complex simulations like the CMS detector. Further, the methods are not explicitly designed to include systematic uncertainties, although additional parameters could be added to the simulation to reflect the systematic variations in the input space. The applicability of these methods for analyses such as described in chapter 2 is still limited because typically not all systematic variations are described by explicit shifts in the input space but by statistical weights.

Another group of techniques puts the focus on the inclusion of systematic uncertainties in the optimization of the dimensionality reduction. First proposed in [106, 107], NNs can be used to learn a dimensionality reduction, which is optimal with respect to the measurement. To achieve this goal, the training objective is aligned with the analysis objective by using the performance of the statistical inference, for example the variance of the POI, as the loss function. The challenge for likelihoods based on Poisson statistics is the count operation, which has no suitable derivative to use automatic differentiation for the optimization and therefore prohibits the convenient usage of modern NN frameworks based on computational graphs. The method proposed in [106] solves the issue by using the softmax function with a small temperature parameter as a replacement for the count operation, which successfully enables again the computation of a valid gradient but uses an approximation of a count as input to the Poisson statistics, which may affect the validity of the statistical inference. In contrast to this solution, the novel method introduced in section 5.3 keeps the count operation unchanged and makes an approximation solely for the gradient.

In principle, any ML technique, which mitigates systematic uncertainties can be used to construct a dimensionality reduction with improved properties compared to a train-

ing only on the nominal dataset. Chapter 4 is dedicated to such techniques with an overview over existing methods in section 4.2. Using such methods with many systematic uncertainties poses a practical challenge since every uncertainty introduces a new hyperparameter, which must be tuned manually. The advantage of the novel technique in section 5.3 is the implicit optimization of these parameters with respect to the analysis objective, which significantly simplifies the usage for data analysis in HEP with $\mathcal{O}(100)$ systematic uncertainties.

5.3 Optimal statistical inference in the presence of systematic uncertainties using neural network optimization based on binned Poisson likelihoods with nuisance parameters

This section introduces a modern approach for analysis in HEP using NNs to find the optimal dimensionality reduction, which is enabled by a novel solution to train on information from the binned Poisson likelihood including nuisance parameters. The approximated gradient for the count operation allows to keep the exact formulation of the likelihood for the statistical inference and at the same time enables the optimization of the NN function with respect to the analysis objective in consideration of all statistical and systematic uncertainties. A dedicated paper about this method was published in [123].

5.3.1 Method

The analysis strategy powered by the techniques discussed in this section is based on using the analysis objective, for example the variance of the POI, as loss function of the NN optimization, which is then supposed to learn an optimal dimensionality reduction. To enable the optimization with modern ML tools using automatic differentiation, the likelihood $\mathcal{L}(\boldsymbol{\theta})$ with the POIs and NPs $\boldsymbol{\theta}$ is required to be analytically differentiable. Although the statistical framework used in most HEP analyses at the LHC is based on binned Poisson likelihoods, this property is not a requirement since the typical statistical tools are based on numerical differentiation [51, 56, 124]. However, numerical differentiation is not suitable for the optimization of a NN since the derivative for each parameter requires at least two evaluations of the likelihood, which is computationally unfeasible for the typical number of trainable parameters in NN functions. To solve this problem, it has to be taken care that the established statistical methods are still usable with the novel analysis strategy, since the validation of new methods for the statistical inference requires exhaustive studies by the HEP community to ensure reliable results. The likelihood function

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{h} \mathcal{P}(d_i | \mu s_i + b_i) \tag{5.1}$$

with the data $\boldsymbol{d}$, the signal and background expectations $\boldsymbol{s}$ and $\boldsymbol{b}$ with the signal strength modifier μ, and the Poisson function $\mathcal{P}$ has only a single mathematical operation that

is not analytically differentiable, which is the count operation used to fill the h bins of the histogram. Although the summary statistic has to be a count to be usable with the established methods for statistical inference using Poisson statistics, the computation of the gradient itself for the training can be modified. Chosen is the gradient of the count approximation already used in chapter 4, which is based on the Gaussian function $\mathcal{G}$ with the standard deviation set to the half width of the bin and normalized to $\max(\mathcal{G}) = 1$. Figure 4.1 visualizes the approximation. It should be noted that compared to the method in chapter 4 only the gradient is replaced to enable the optimization of the NN parameters, not the computation of the count itself in general.

All mathematical operations used to add NPs to the statistical part of the likelihood in equation 5.1 are analytically differentiable, see section 2.5 for a detailed description of the mathematics. To simplify the following introduction of the method, the parameters in the likelihood $\boldsymbol{\theta} = (\mu, \eta)$ are reduced to the signal modifier μ as POI and the NP η. The likelihood including the NP is given by

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{h} \mathcal{P}(d_i | \mu s_i + b_i + \eta \Delta_i) \cdot \mathcal{N}(\eta) \tag{5.2}$$

with the systematic variations $\boldsymbol{\Delta}$ and the standard normal distribution $\mathcal{N}$ as constraint term for the NP η. In case the systematic variation is not symmetric, the impact of the NP on the expectation in the likelihood can be written as $\max(\eta, 0)\, \boldsymbol{\Delta}_{\text{up}} + \min(\eta, 0)\, \boldsymbol{\Delta}_{\text{down}}$.

The summary of the information in the dataset $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ with n events described by d variables is implemented using the NN function $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\omega})$ with the trainable parameters $\boldsymbol{\omega}$. Although the examples in the following are restricted to a one dimensional histogram, the gradient approximation is also usable with a multivariate Gaussian function and k output nodes for the NN $\boldsymbol{f}$.

Since the performance of an analysis is typically measured in terms of the variance of the estimate for the POIs, the optimization of the trainable parameters $\boldsymbol{\omega}$ is performed with respect to this metric to align the analysis objective and the objective of the NN training. To do so, an analytically differentiable formulation of the variance of the POI is required. In the statistical inference of a HEP analysis, the variance, or uncertainty, of the estimate for the POI is retrieved by profiling the likelihood ratio, section 2.5 gives additional details. Since the procedure of profiling the POI, which means minimizing at each point of the likelihood scan all NPs, is not a closed analytic formula, an asymptotic estimate for the variance is used given by the Fisher information [125]. The Fisher information of the likelihood

$$F_{ij} = \mathrm{E}\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j}\left(-\log \mathcal{L}(\boldsymbol{\theta})\right)\right] \tag{5.3}$$

can be used to estimate the variance of the estimate for any parameter in $\boldsymbol{\theta}$. To include this approximation successfully in the method, the expectation denoted by $\mathrm{E}[\ldots]$ in the formula is removed by using as data the Asimov dataset [58], which reflects the median expected outcome of the measurement. Then, the variance of the estimate for the parameter θ_i is given by

$$V_{ii} = F_{ii}^{-1}. \tag{5.4}$$

This holds true because the estimation of the parameters $\boldsymbol{\theta}$ using the likelihood is asymptotically efficient, which implies that the variance is close to the Cramer Rao bound [126, 127]. Assuming that the first parameter in $\boldsymbol{\theta}$ is the POI μ, the loss function used to minimize the trainable parameters $\boldsymbol{\omega}$ of the NN function $\boldsymbol{f}$ is V_{11}.

It should be noted that the computation of the asymptotic variance does not require the likelihood ratio but only the likelihood itself. The denominator in the general likelihood ratio test is fixed to the global best fit, so that the logarithm and the derivative in the Fisher information remove the constant completely, simplifying the implementation compared to the profile of the likelihood ratio. Also not trivially visible is the mechanism, which combines the information about the parameters in $\boldsymbol{\theta}$, for example for the element V_{11} related to the POI μ. The matrix inversion of F_{ij} is the key element, which mathematically combines the parameters and mixes the NPs with the POIs. Related to the example in equation 5.2, this allows that an improved constraint of the NP η can reduce the variance of the POI μ.

The question whether the optimization should be capable to find an optimal dimensionality reduction can be addressed by analyzing the potential inefficiencies. Assuming that the optimization converges, two operations may reduce the efficiency of the summary statistic. A possible degradation could happen if the NN with the chosen number of outputs k and the trainable parameters $\boldsymbol{\omega}$ is not capable to learn a sufficient statistic, meaning that the dimensionality reduction from d to k is not lossless. Trivially visible is that in the case of $k \to d$, the NN function becomes a sufficient statistic, but most interesting is the case with $k = 1$, which is the desired case to simplify the usage of the method in practise. In general, a strong statement about the efficiency of the dimensionality reduction given by the NN function $\boldsymbol{f}$ is not possible, but a very typical analysis case in HEP can be discussed. Assuming an analysis, which has as objective to measure the cross section of a signal process via the signal strength modifier μ obstructed by several background processes, it can be shown that the NN function trained on the CE function is a sufficient statistic if no systematic uncertainties are considered [66, 106]. The CE function minimizes the statistical part of the likelihood, such as shown in equation 5.1, with the setting of a binary classification task separating the signal from the sum of the backgrounds. Therefore, the optimization of V_{11} without NPs in the likelihood is expected to result in the same variance of μ than a training on the CE loss, which also can be observed in the examples discussed in sections 5.3.2 and 5.3.3. Following, the inclusion of NPs in the likelihood, see equation 5.2, implements just a modified version of a case in which the NN function is already a sufficient statistic. The simple example in section 5.3.2 has a tractable likelihood in the input space, which is used to show that the proposed method reaches the optimal performance. Also well known is that a NN with a sufficient complexity given by the number of trainable parameters $\boldsymbol{\omega}$ can learn any function. It should be noted that this discussion implies that typical ML based analysis strategies based on classification objectives are already using an optimal approach with respect to the statistical uncertainty of the analysis and all gain from this method stems from the inclusion of the systematic uncertainties in the training objective. But with increasing data statistics from the experiments delivered by Run 2 and 3 of the LHC, analyses are expected to be increasingly dominated by systematic uncertainties, which

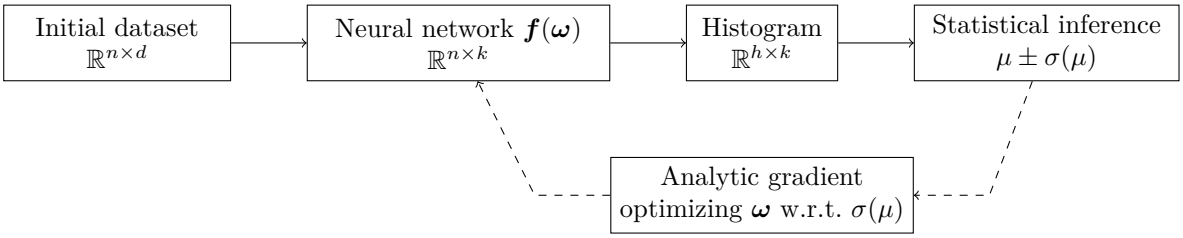

Figure 5.2: Overview over the novel method introduced in section 5.3. The dimensionality of the initial dataset with n events and d variables is reduced by a NN $\boldsymbol{f}$ to k observables, which enter the binned Poisson likelihood with an histogram built by h bins. The approximation of the gradient for the count operation allows to use automatic differentiation to optimize the trainable parameters $\boldsymbol{\omega}$ of the NN function $\boldsymbol{f}$ with respect to the variance of the POI μ. The optimization takes into account the information about systematic uncertainties, which are integrated in the likelihood with NPs.

makes the efficient inclusion of the information about systematic uncertainties in the optimization of the analysis strategy an influential improvement. Another potential lossy operation is the histogram. With an increasing number of bins h, the histogram converges to the continuous description of the probability density of $\boldsymbol{f}$, which is by construction optimal. The examples in the following sections show that already very few bins can be sufficient to reach an optimal result.

An overview over the method is given in figure 5.2, which can be compared to the typical analysis strategy in figure 5.1. The novel method described in this section allows to automatize the optimization of the free parameters in the analysis strategy with modern ML methods, which results in better and faster results of HEP analyses. Additional details about the application of the technique are presented in section 5.3.2 with a simple example based on pseudo experiments and section 5.3.3 with a more complex example from HEP.

5.3.2 Application on a simple example based on pseudo experiments

This section presents the capabilities of the proposed method with a simple example based on pseudo experiments. The example uses a two dimensional input space spanned by the variables x_1 and x_2, which is populated by the two classes signal and background. Figure 5.3 shows the distribution of the classes in the input space, which are uncorrelated multinominal Gaussian distributions with the means $(\,0\;0\,)$ and $(\,1\;1\,)$, respectively. The example is enhanced by a systematic uncertainty on the mean of the background process parametrized by the shifts $x_2 \pm 1$, which represent the 1σ variations of the uncertainty. The variations are implemented with duplicates of the dataset containing the applied shifts, but a reweighting of the nominal dataset is also possible on the level of the histogram. Further, the expectations of the processes are normalized with statistical weights to 50 signal and 10^3 background events in the dataset, which represents a typical scenario in a HEP analysis of a rare physical process.

The applied NN consists of a single hidden layer with 100 nodes and ReLU activa-

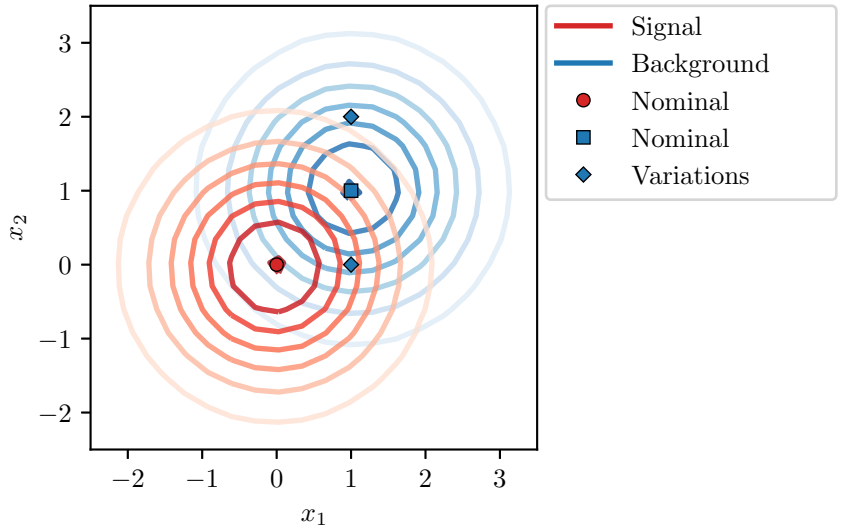

Figure 5.3: Distributions of the signal and background classes in the input space with a systematic uncertainty on the mean of the background process parametrized by the shifts $x_2 \pm 1$, which represent the 1σ variations of the uncertainty.

tions [89]. The output layer has a single node with a sigmoid activation, which limits the range of values to $(0, 1)$. The trainable parameters are initialized with the Glorot algorithm [61] and optimized with the Adam optimizer [64]. The gradients for the optimization are computed on a dataset with 10^5 samples and the training is stopped if the loss does not improve on an independent validation dataset of the same size for 100 gradient steps in a row. The model with the smallest validation loss is used for all following results, which are computed based on another independent dataset consisting of 10^5 samples. An improved convergence has been observed if the model is trained first only on the statistical part of the likelihood, which is done in this example for 30 gradient steps.

Because this example has a tractable likelihood in the input space, the statistically best possible result for the measurement of the signal strength modifier μ is known. Figure 5.4 shows the profile of the likelihood ratio with the optimal constraint of the POI found as $\mu = 1.0^{+0.37}_{-0.35}$. The profile is always computed twice, once with only the statistical uncertainty and again with the systematic uncertainty in addition, which visualizes the impact of the systematic uncertainty on the result. It should be noted that the profile shows the median expected result enforced by the used Asimov dataset, which also explains that the best fit value of μ is always at 1.0.

In addition to the optimal result shown in figure 5.4, a NN is trained with the CE loss on a binary classification task. To encode the probability of signal and background correctly, the CE loss is computed with sample weights, which encode the expected rate of the processes. Eight bins are used for the histogram, which counts are the inputs for the estimate of the signal strength modifier μ. Such as discussed above, the expectation

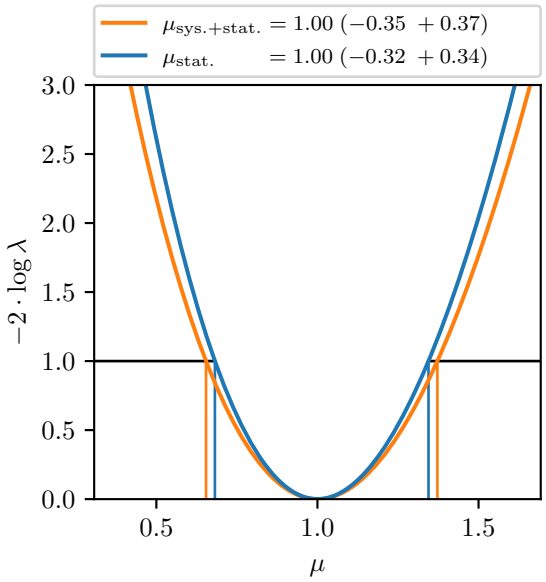

Figure 5.4: Profile of the likelihood defined in the input space with only the statistical uncertainty (blue) and the systematic uncertainty in addition (orange), which represents the statistically best possible estimate for the POI μ.

is that this analysis strategy is optimal with respect to the statistical uncertainty of the scenario. As known from the optimal solution in figure 5.4, the best possible estimate of the POI in consideration of only the statistical component of the likelihood is $\mu = 1.0^{+0.34}_{-0.32}$. The top row in figure 5.5 shows the distribution of the NN output, the NN function in the input space and the profile of the likelihood ratio. The profiles reveal that the CE approach achieves an almost optimal performance with respect to the statistical uncertainty given by $\mu = 1.0^{+0.35}_{-0.33}$ but is massively degraded by 32 % to $\mu = 1.0^{+0.45}_{-0.44}$ if the systematic uncertainty is considered in the statistical inference. The NN function in the input space shows the expected decision plane, which is aligned with the separation by the means of the signal and background distributions.

A similar performance is expected by the analysis strategy with the NLL based loss function but without including the systematic uncertainty in the training. In this scenario, the CE loss and the NLL loss optimize the same target, which is shown in the middle row of figure 5.5. The profile of the likelihood ratio gives results very similar to the CE approach, which confirms the expectation. However, the distribution of the NN function has a different shape, which shows that the training has found a different solution with a very similar outcome for the estimate of μ. Also the NN function in the input space is not perfectly aligned along the diagonal such as visible for the CE loss. These effects have no impact on the variance of the estimate for μ, because the regions close to the border of the shown input space are only sparsely populated.

If the systematic uncertainty is included in the NLL loss, the variance of the estimate for μ improves significantly. The bottom row in figure 5.5 shows the results. The constraint of

$\mu = 1.0^{+0.39}_{-0.36}$ is just 4 % worse than the best possible result in figure 5.4. This confirms that the strategy allows to achieve an asymptotically optimal result with full consideration of all statistical and systematic uncertainties. The NN function in the input space shows that the decision plane is slightly rotated towards the x_2 axis, which is a partial decorrelation against the systematic uncertainty on the mean of the background process in x_2. In contrast to the decorrelation techniques discussed in chapter 4, this method automatically optimizes the impact of the systematic uncertainty on the analysis objective without additional hyperparameters. The correlation of the POI μ to the NP η is reduced from 64 % to 13 % comparing the CE approach with the training on the NLL based loss including the systematic uncertainty. The effect is also visible in the distribution of the NN output in figure 5.5, which shows for the NLL approach small systematic uncertainties in the sensitive bins.

5.3.3 Application on an example from high energy particle physics

The example from HEP presented in this section makes use of the dataset from the Higgs ML challenge, which implements a search for the SM Higgs boson in the decay to a ditau pair with a lepton and a hadronic tau in the final state. The dataset is described in detail in section 3.3.4 and in [90]. In summary, the dataset consists of about 6×10^5 simulated events with 30 variables each. From the 30 variables the missing transverse energy $\not{E}_\mathrm{T}$ (`PRI_met`), the visible mass $m_\mathrm{vis}^{l\tau}$ (`DER_mass_vis`), the transverse momentum of the reconstructed Higgs boson p_T^H (`DER_pt_h`) and the distance in the pseudo rapidity of the leading and subleading jets $\Delta\eta_\mathrm{jet,jet}$ (`DER_deltaeta_jet_jet`) are selected. The distributions of the variables are shown in figure 5.6 for the signal and the background classes, which consist of a mixture of the contributing physical processes. The selected variable $\not{E}_\mathrm{T}$ is not an input to the NN but the three other variables span the input space. The variable $\not{E}_\mathrm{T}$ is used to implement a systematic uncertainty described by the shifts $\not{E}_\mathrm{T} \cdot (1.0 \pm 0.1)$, which represents a 10 % uncertainty on the scale of the missing transverse energy. The impact of the variations on the signal and background expectations are shown in figure 5.6. The variations are chosen as described because only the contribution of a systematic uncertainty, which does not affect the normalization of the process can be mitigated. The introduced systematic uncertainty does not have any normalization effect and therefore the proposed method is expected to have a strong impact on the measurement of the signal strength modifier μ. Further, the missing transverse energy is on purpose not an input to the NN to present a more complex example in which the source of the systematic uncertainty is not a direct input to the training but has to be inferred implicitly via the correlation to the other variables. Statistical weights are used to describe the systematic variations in the input space, which proves the capability of the method to integrate such information and presents a computationally more efficient solution than duplicates of the nominal dataset. Finally, the dataset is reduced to only those events, which have all of the selected variables defined. The signal expectation is scaled by a factor of two, which results in 244.0 and 35140.1 (106505 and 131480) weighted (unweighted) events for the signal and background process, respectively. Inclusively, the ratio of signal to background is 1 to 144, which represents a realistic scenario in HEP

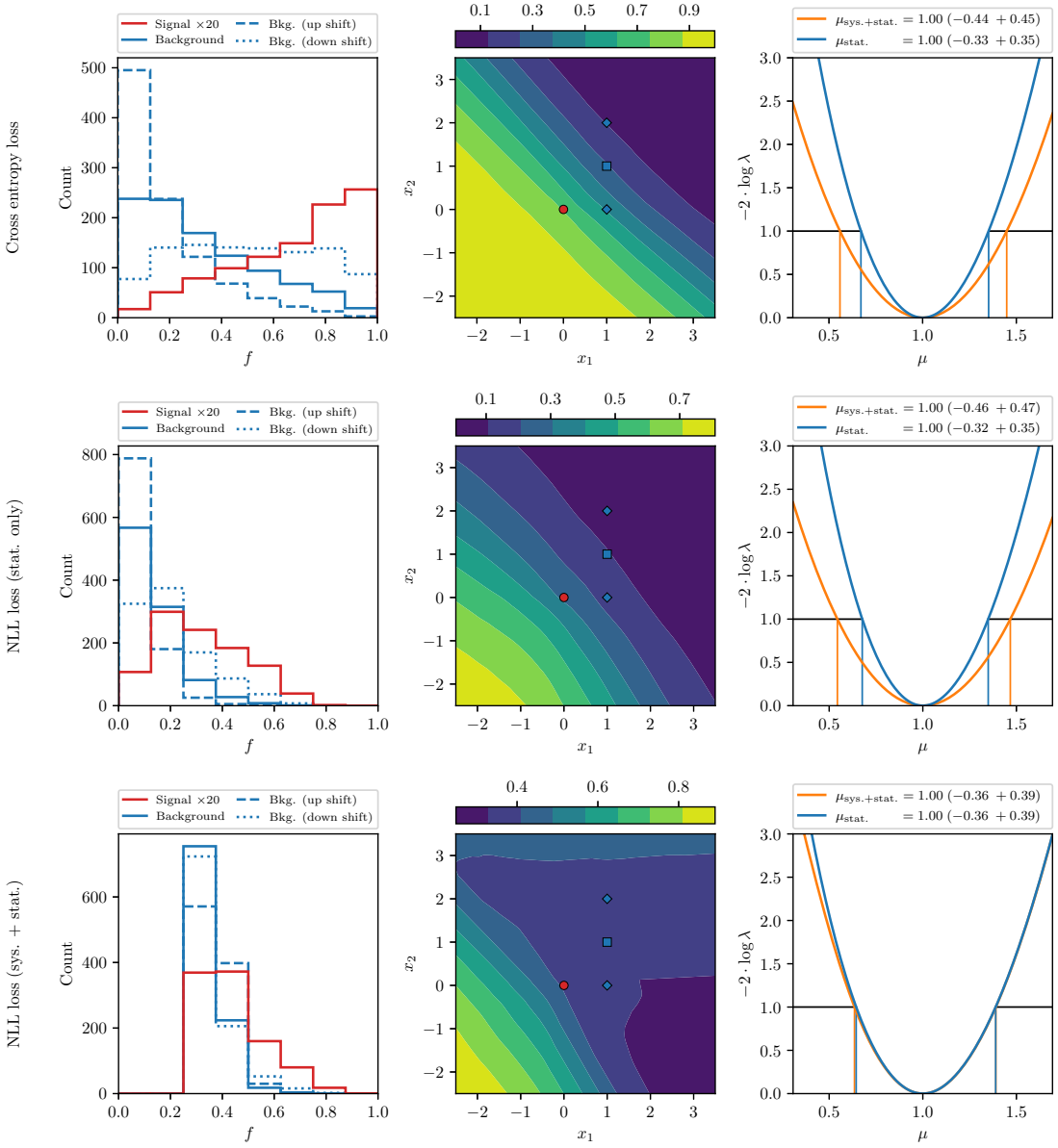

Figure 5.5: The simple example based on pseudo experiments is evaluated for a training on the CE loss (top row), the NLL loss without systematic uncertainty (middle row) and the NLL loss including the systematic uncertainty (bottom row). Each scenario is studied with the distribution of the NN output (left column), the NN function in the input space (middle column) and the profile of the likelihood ratio considering only the statistical component of the likelihood and the systematic uncertainty in addition (right column).

similar to the analysis described in chapter 2.

The study is carried out with the same NN architecture and training procedure than in section 5.3.2 using two thirds of the dataset for training and validation and the rest for the following results. The training and validation splits the respective subset of the data again in half and the expectation of the signal and background processes are scaled with statistical weights to the inclusive expectation to model the correct statistical uncertainties in the likelihood. For the inputs of the NN, the transformation $(x - \overline{x})/\sigma(x)$ with the mean $\overline{x}$ and the standard deviation $\sigma(x)$ is applied to standardize the value ranges of the variables.

Figure 5.7 shows the same scenarios than for the simple example in section 5.3.2, which use the CE loss, the NLL loss with only the statistical uncertainty in the training and the NLL loss including the systematic uncertainty. Because an optimal solution is not known due to the intractable likelihood in the input space, the CE loss represents with the estimate $\mu = 1.0^{+0.69}_{-0.69}$ the baseline for the comparison to the proposed method. However, to validate that the strategy based on the classification objective results as expected in an optimal result with respect to the statistical component of the likelihood, the profiles of the likelihood ratio can be compared to a training on the NLL loss without the systematic uncertainty. Figure 5.7 shows a close match between the two scenarios with the characteristic gap between the profiles with and without the systematic uncertainty in the statistical inference. The training on the NLL loss including the systematic uncertainty gives an estimate of $\mu = 1.0^{+0.61}_{-0.60}$, which improves the result from the CE approach by 13 %. The comparison of the profiles shows again that the proposed method successfully performs a trade off between an estimate of μ being only influenced by the statistical uncertainty and an improved result in consideration of the systematic uncertainty. Similar to the simple example, the correlation of the POI μ to the NP η drops from 69 % in the scenario with the CE loss to 4 % for the training on the NLL loss including systematic uncertainties.

To study the impact of the number of bins on the proposed method, figure 5.8 shows for different number of bins the correlation of the POI μ with the NP η and the constraint of the POI μ. Compared to the training on the CE loss, the NLL based training shows a reduced correlation in all configurations from two to 64 bins. An improvement in the correlation of μ with η is visible up to eight bins, which indicates that two and four bins do not offer enough degrees of freedom to achieve the same performance as eight bins or more. In terms of the analysis objective, the constraint of the POI μ, the novel method has a stable average improvement of 10 % compared to the CE approach, which indicates that the NLL loss makes best use of the available information in all configurations. Also notable is that the difference between the statistical inference with and without consideration of the systematic uncertainty has a constant average distance of 0.18 for the CE loss and about 0.01 for the proposed method. However, it should be noted that the statistical model in an actual HEP analysis also contains contributions that take into consideration the uncertainty on the expectation from simulation. These uncertainties are introduced by the finite statistics of the simulated events, which prevent that the constraint of μ scales with many bins such as visible in figure 5.8. Typically an analysis tries to use as little bins as possible to optimize the description of the contributing

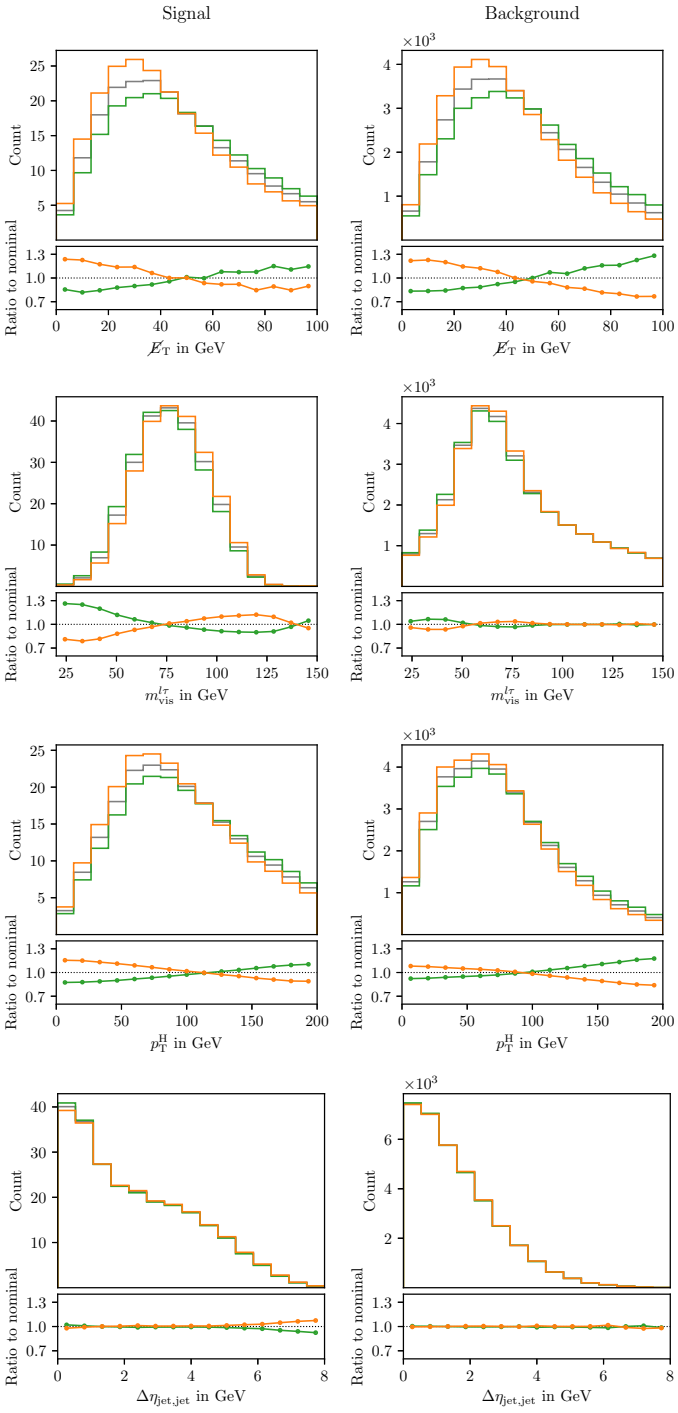

Figure 5.6: Selected variables of the Higgs ML dataset for the signal (left) and background (right) process. Each subplot shows the ratio of the systematic variation implemented by $\not{E}_T \cdot (1.0 \pm 0.1)$ to the nominal expectation.

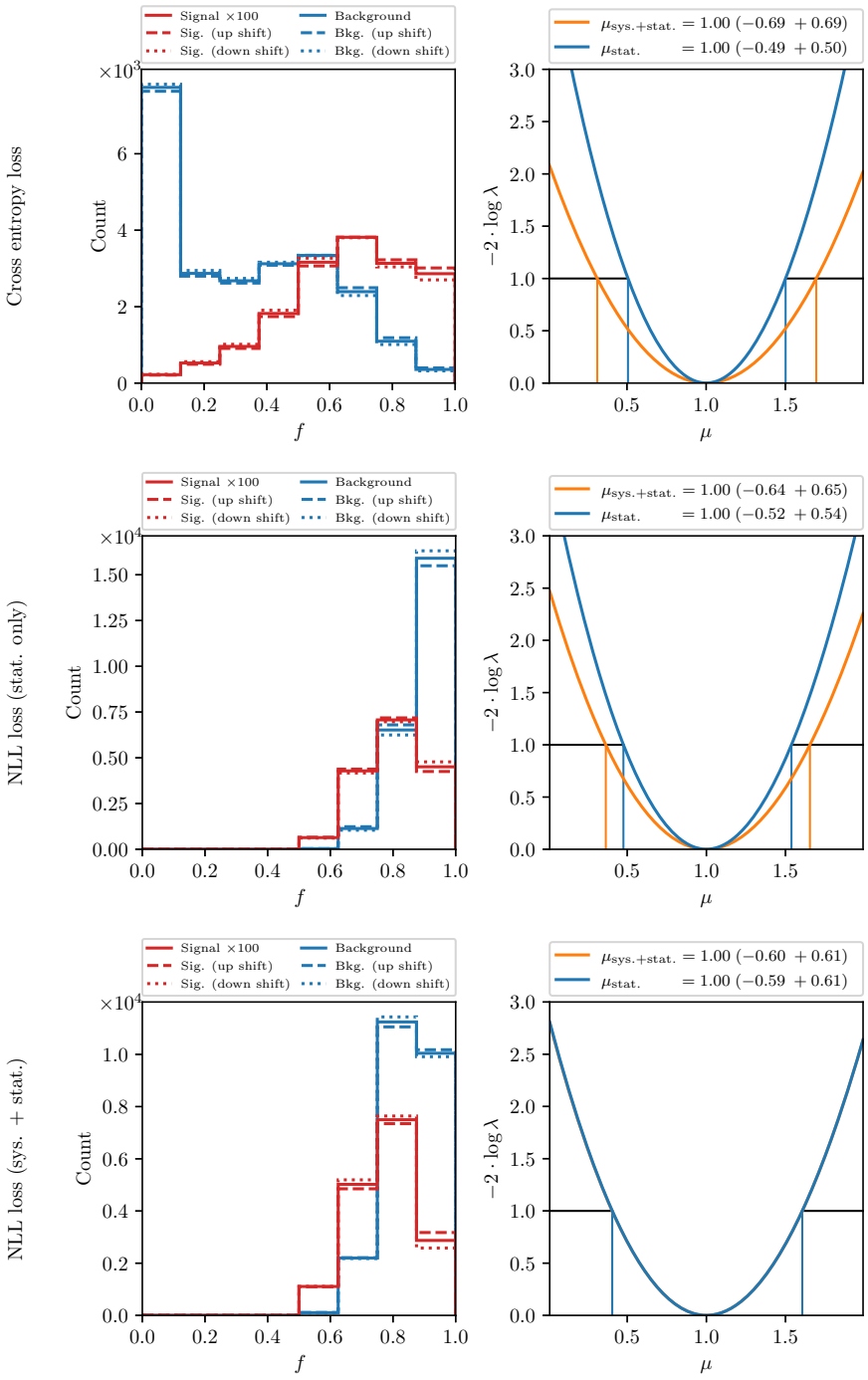

Figure 5.7: The example from HEP is evaluated for the CE loss (top row), the NLL loss without systematic uncertainty (middle row) and the NLL loss including the systematic uncertainty (bottom row). Each scenario is studied with the distribution of the NN output (left column) and the profile of the likelihood ratio considering only the statistical component of the likelihood and the systematic uncertainty in addition (right column).

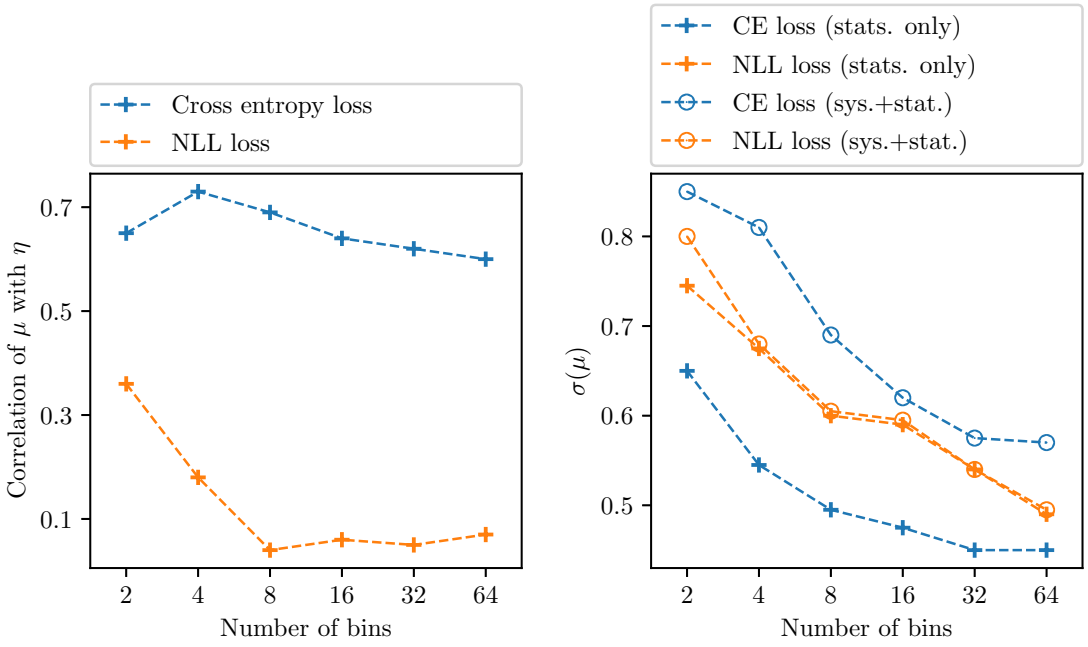

Figure 5.8: Impact of the number of bins on the result of the statistical inference shown by the correlation of the POI μ with η (left) and the constraint of the POI μ (right), evaluated for the CE approach and the training on the NLL based loss including the systematic uncertainty. It should be noted that the number of bins is the same in the training of the NLL loss and for the estimate of μ based on the resulting histogram.

processes, which improves the reliability of the analysis and the result of the statistical inference due to the reduced systematic uncertainties.

CHAPTER 6

Conclusion

This chapter presents in section 6.1 a summary of this work and discusses in section 6.2 possible future fields of research with respect to the topics discussed in the previous chapters.

6.1 Summary

Beyond any doubts, machine learning (ML) has become an indispensable tool in high energy particle physics (HEP) to achieve the best possible results in the analysis of experimental data. The success of ML in data analysis in HEP pushes the usage of modern multivariate techniques to the limits and finds application in almost every step of the analysis strategy. Especially if ML is used to perform the event selection or to create observables for the statistical inference, hidden dependencies of the ML model on the high dimensional input space pose novel challenges to ensure robust and reliable results for precise physics measurements. At the same time, modern ML powers novel strategies for data analysis in HEP promising highest sensitivity to the physics of interest.

Such an analysis using neural networks (NNs) for event selection and final observables is presented in chapter 2, which has the objective to study the decay of the Standard Model (SM) Higgs boson into two tau leptons in data from the Compact Muon Solenoid (CMS) experiment at the Large Hadron Collider (LHC). The procedures to realize a precise measurement of the inclusive and differential cross sections in the simplified template cross section (STXS) framework [48] is described with focus on the implications of the massive usage of ML techniques on the statistical inference in presence of systematic uncertainties. The ML based analysis strategy led to the most precise measurements of the SM Higgs boson cross sections in the decay to tau leptons [44] with a significantly improved sensitivity compared to traditional analysis strategies [113]. Figure 6.1 presents the results of the differential measurement, which is an invaluable resource to constrain the theoretical framework of the SM and beyond.

Chapter 3 puts focus on the challenge to validate the input space and the used ML model for such an analysis strategy. A novel method is introduced, which allows to understand the dependence of the NNs on the features in the multidimensional input space [88]. The capability of the technique to identify and quantify the importance of

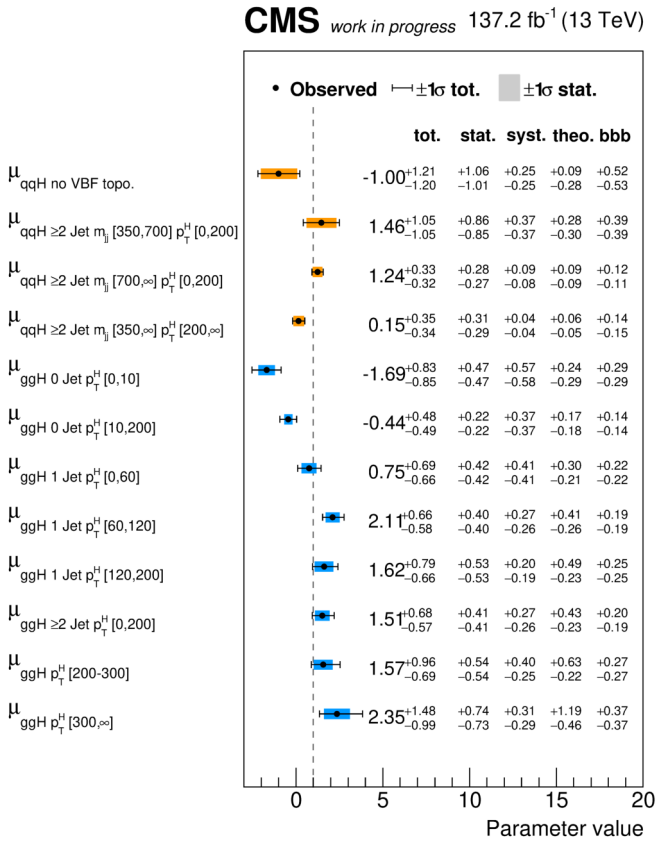

Figure 6.1: Differential measurement of the SM Higgs boson cross section in the STXS framework extracted from data of the CMS detector at the LHC [44]

higher order features, such as correlations between variables, on the NN output is crucial to find the subspaces of the input space with highest importance. This enables efficient validation strategies to quantify the agreement between expectation and data in the high dimensional input space, which ensures the detection and sufficient coverage of potential mismodelings. Such a validation strategy is applied in the analysis described in chapter 2 and contributes significantly to the robustness and reliability of the measurement.

Known features in the input space, which are not desired to contribute to the analysis result, are challenging to mitigate with multivariate techniques, which are in principle sensitive to any information in the presented dataset. In consequence, novel techniques are required, which allow to control the dependence of the ML model on features in the input space. Chapter 4 presents a novel technique for NNs to control the information, which contributes to the optimization of the model during the training and allows to suppress in a fine granular way any feature in the high dimensional input space [108]. The proposed method introduces only a minimal set of additional hyperparameters and

is simple to integrate with typical objectives in HEP analyses such as the separation of physical processes in data.

Going a step further, chapter 5 discusses the efficiency of today's data analysis strategies in HEP and proposes a novel approach to reach a statistically optimal result. Based on modern ML methods, the introduced analysis strategy combines the analysis objective with an automatic optimization of the dimensionality reduction required to build the statistical model for the inference of the parameters of interest (POIs) [123]. The key development is the formulation of an analytical derivative of the count operation, which allows to use automatic differentiation with modern computation frameworks and enables the usage of binned Poisson likelihoods in the training objective. An example with a known likelihood in the input space is used to demonstrate that the analysis strategy reaches asymptotically the optimal sensitivity. Further, a more complex example presents the application of the method in a typical data analysis scenario from HEP and finds in the presence of systematic uncertainties a superior sensitivity to the physics of interest.

6.2 Outlook

Chapter 2 presents a differential measurement of the SM Higgs boson in the decay to two tau leptons in the STXS framework, which has the purpose to unify comprehensively differential measurements and targets the combination of results across analyses and experiments. The analysis strategy discussed in chapter 2 is highly optimized on such STXS measurements and has set a reference for the next decade, in terms of the physics results but also with respect to the applied data analysis techniques.

Looking towards Run 3 and 4 of the LHC, future analyses of data from the experiments, such as CMS and ATLAS, will be heavily challenged to gain from the increased integrated luminosity of $300\,\mathrm{fb}^{-1}$ and $3000\,\mathrm{fb}^{-1}$, respectively [20]. The pileup of additional proton proton collisions is expected to increase from approximately 30 proton proton interactions today to over 130 interactions for a typical event during the high luminosity LHC phase in Run 4 [128]. This development increases substantially the systematic uncertainties of future measurements compared to the additional sensitivity gained by the growing number of recorded events. In context of the analysis of the SM Higgs boson, the data recorded from Run 3 and 4 are intended to be used to perform measurements at the highest precision to test the predictions of the SM with all available information. These upcoming analyses will be significantly stronger impaired by systematic uncertainties than the analysis discussed in chapter 2, which will require novel strategies for the analysis of the data to perform the best possible measurements.

Certainly, modern ML techniques will play an important role to achieve these goals. The novel techniques introduced in chapter 3 and 4 allow to identify and control the dependence of the ML model on features in the input space and therefore enable robust measurement in the presence of systematic uncertainties. Although such methods are crucial for ML based analyses with reliable results, these methods do not necessarily promise an analysis with an optimal sensitivity to the physics of interest. Chapter 5 discusses the efficiency of typical analyses in HEP and identifies the parts of today's ML

based analysis strategies, which allow to improve the results. The results from chapter 5 are highly promising and show the optimal use of the experimental data enabled by the integration of the knowledge about the systematic uncertainties of the experiment.

An important question is whether, or how, such novel data analysis techniques can be safely used for data analysis in HEP. Analyses in HEP are driven by precision and statistically sound results, which allow to guide the experimental field with multibillion dollar experiments in the right direction. The question to be answered is whether the inclusion of systematic uncertainties in an automatic optimization of the analysis strategy affects the reliability of the statistical inference. A possible scenario is a ML based observable entering the statistical inference, which is tuned to mitigate exactly the systematic variations such as modeled in the expectation of the statistical model but is in reality highly sensitive to the actual variations in data. This problem is very similar to the overfitting problem in ML, which describes the problem that an overtrained ML model is getting sensitive to the statistical fluctuations in the training dataset but eventually does not generalize well to new data. This could mean for novel techniques, which include the knowledge about systematic variations in the training, that two different estimates of the systematic uncertainty are required to guarantee the generalization of the learned relations to new data. Finding a safe strategy to apply such novel ML based data analysis techniques, which make use of the knowledge about systematic uncertainties is an important field of study, but currently not followed by many researchers around the globe. Because of the discussed points, HEP is for good reasons a conservative scientific field with respect to ML based analysis strategies and the adoption of novel methods may take years to be validated and accepted. However, to extract all possible information out of the precious data from past and upcoming runs of the LHC, the efforts to study such optimized analysis strategies cannot start soon enough and present a rich scientific field for the future.

APPENDIX A

Data and expectation in the $\mu\tau_{\mathrm{h}}$ channel for the data taking period of 2018

The following figures present the data and expectation in the $\mu\tau_{\mathrm{h}}$ channel for the data taking period of 2018, which are input to the measurement of the cross section in the STXS framework. The results of the measurement, which combines all available data of the LHC Run 2 are discussed in section 2.8.

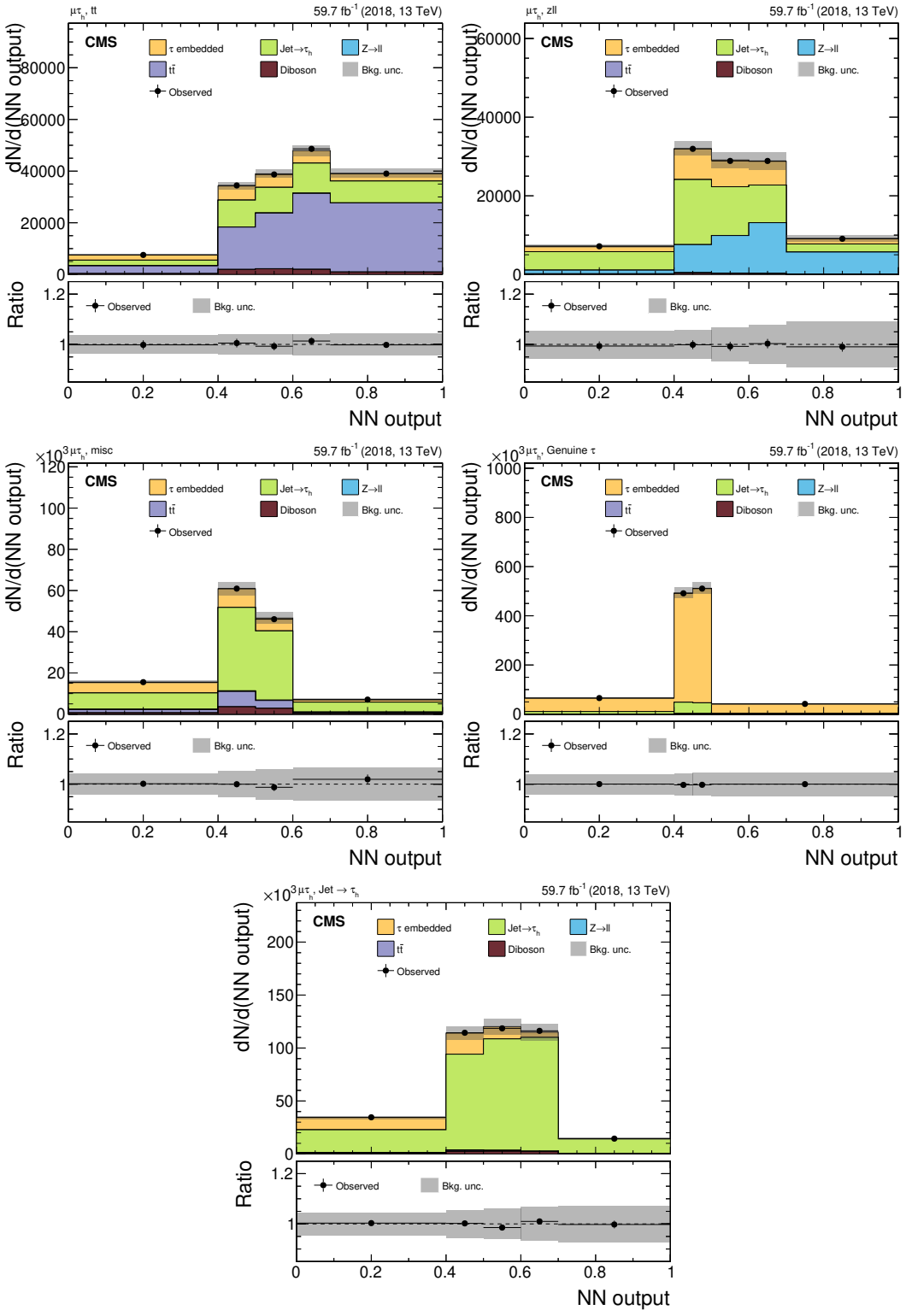

Figure A.1: Data and expectation in the $\mu\tau_{\mathrm{h}}$ channel for the data taking period of 2018 in the background categories

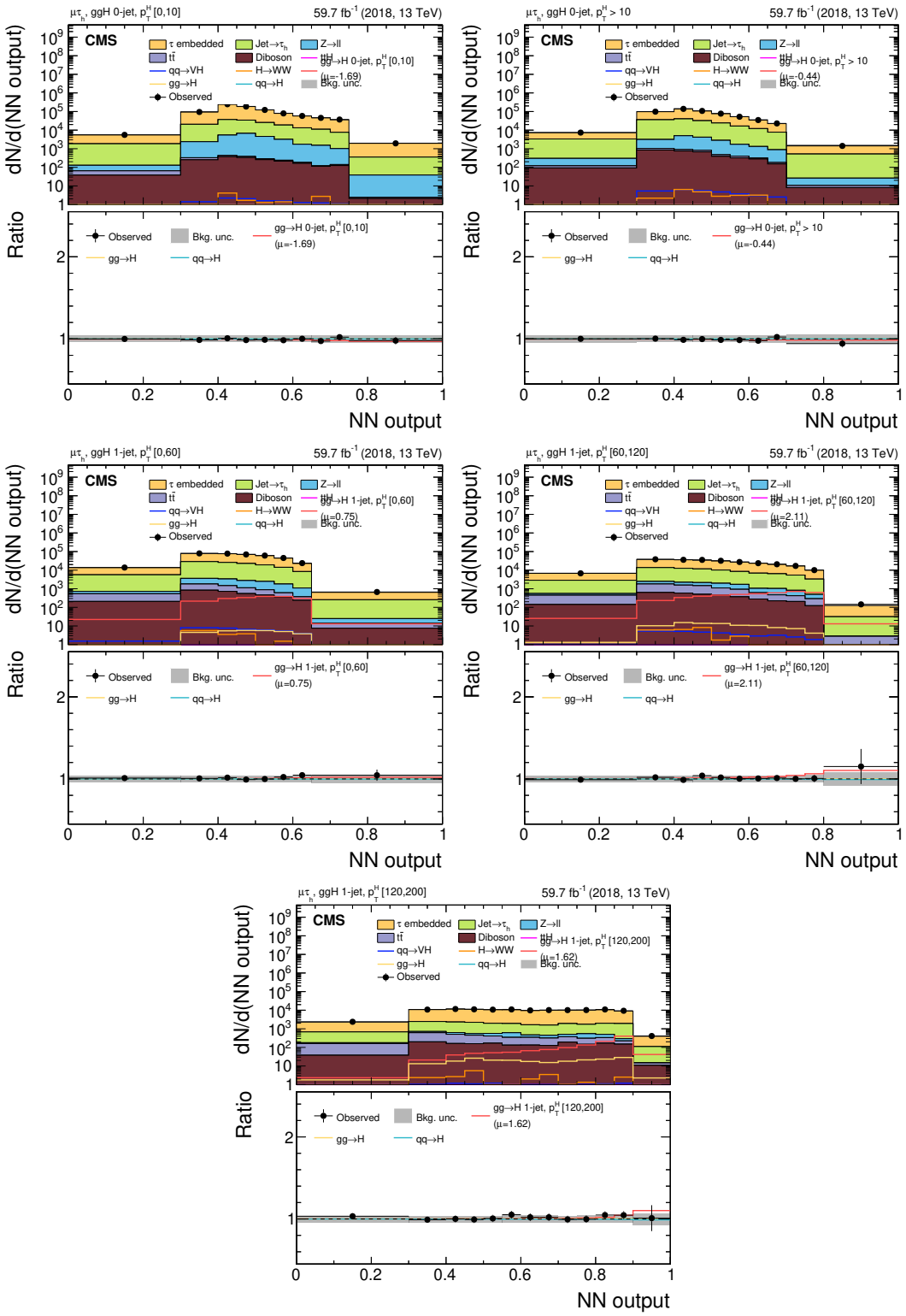

Figure A.2: Data and expectation in the $\mu\tau_h$ channel for the data taking period of 2018 in the Higgs boson production from gluon fusion (ggH) categories

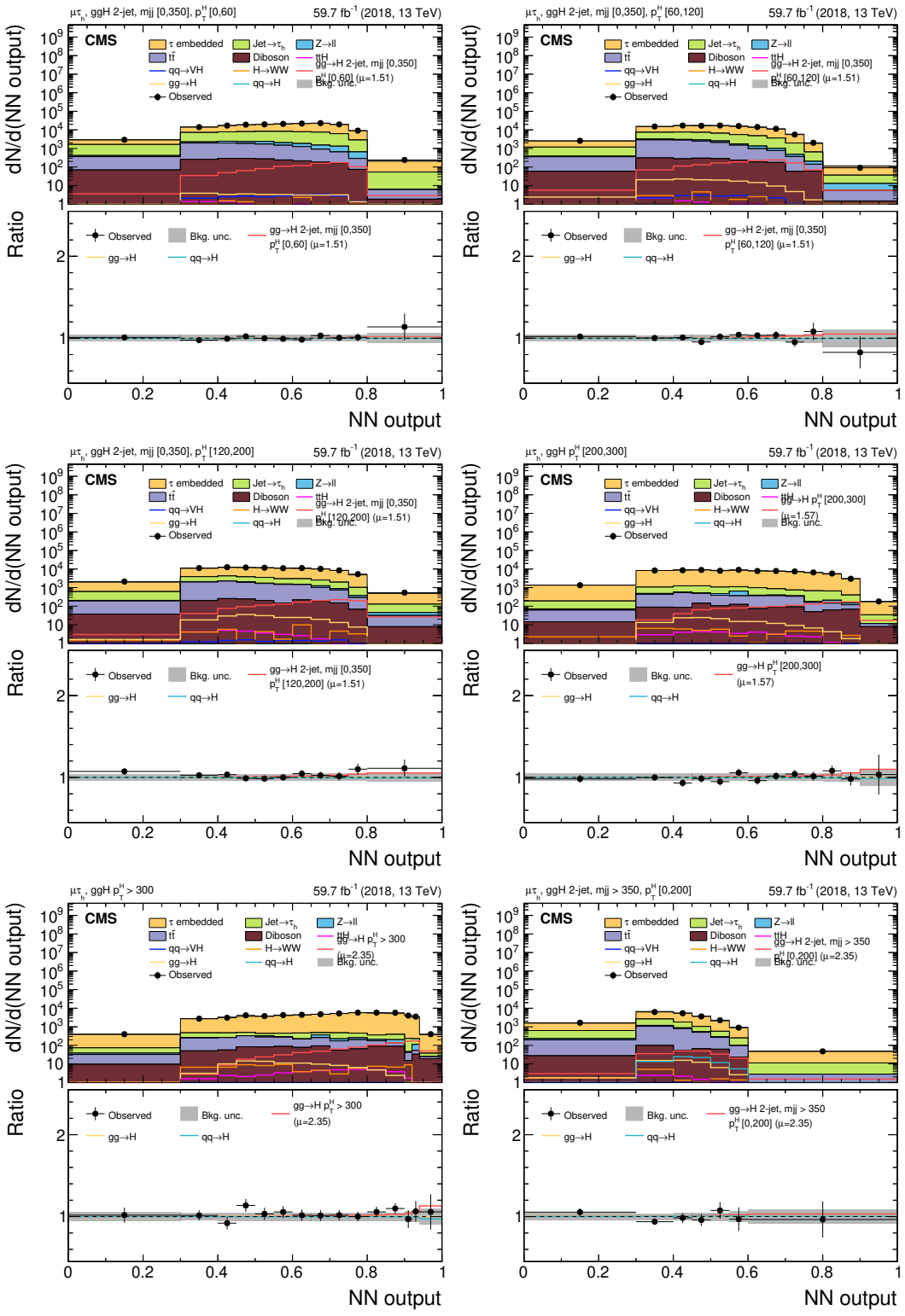

Figure A.3: Data and expectation in the $\mu\tau_h$ channel for the data taking period of 2018 in the ggH categories

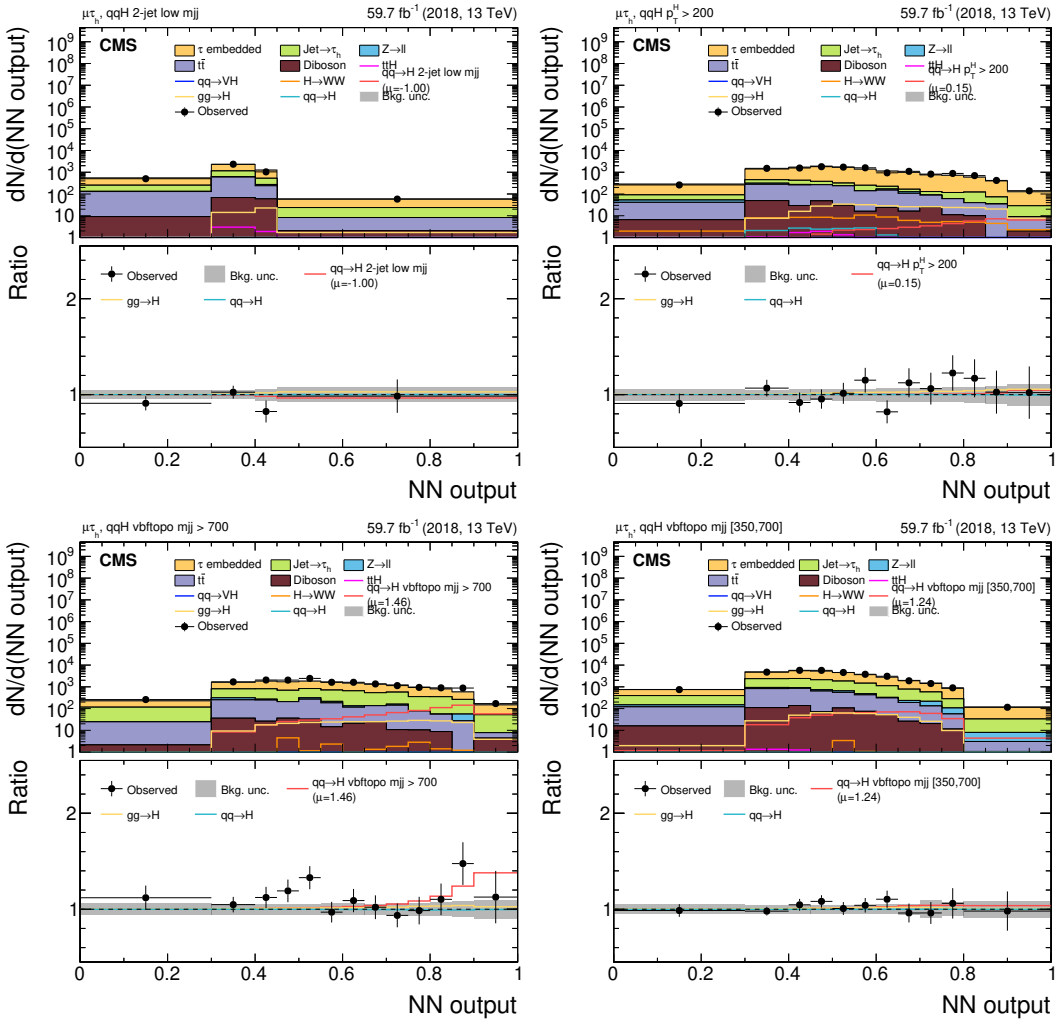

Figure A.4: Data and expectation in the $\mu\tau_{\mathrm{h}}$ channel for the data taking period of 2018 in the Higgs boson production from vector boson fusion (qqH) categories

APPENDIX B

Abbreviations

ANN adversarial neural network. 60, 61, 63, 64

AUC area under the curve. 46, 52–54, 57, 64

BDT boosted decision tree. 60, 73

CE cross entropy. 27, 29, 39, 48, 62–67, 69, 79, 81–85, 87, 88

CERN the European Organization for Nuclear Research. 3, 6

CMS Compact Muon Solenoid. 3, 4, 6–11, 14, 15, 17, 19, 73, 74, 76, 89–91

DY Drell Yan. 12

ECAL electromagnetic calorimeter. 7, 14

FF fake factor. 12, 13, 16, 30

FPGA field programmable gate array. 7

ggH Higgs boson production from gluon fusion. 9, 11, 17, 18, 26, 30, 39, 95, 96

GoF goodness of fit. 33–35

HCAL hadronic calorimeter. 7, 14

HEP high energy particle physics. 3–5, 9, 12, 19, 20, 23, 29, 32, 36, 43–47, 53, 59–62, 66, 73–80, 83, 85, 87, 89, 91, 92

HLT high level trigger. 8, 16

HPS hadron plus strips. 14, 16

LHC Large Hadron Collider. 3, 4, 6, 7, 9, 11, 12, 14, 15, 43, 73, 74, 77, 79, 89–93

ML machine learning. 3–6, 9, 14, 15, 24, 27, 32, 33, 36, 39, 43–48, 53, 59–61, 66, 68, 73, 75–77, 79, 80, 83, 86, 89–92

MSSM Minimal Supersymmetric Standard Model. 39

NLL negative logarithmic likelihood. 21, 23, 29, 82–85, 87, 88

NN neural network. 3–6, 14, 15, 17, 24–27, 29–33, 36–39, 43, 44, 46–48, 50–54, 56, 57, 59–67, 69, 70, 72, 73, 75–85, 87, 89, 90

NP nuisance parameter. 19–21, 77–80, 83, 85

PCA principal component analysis. 46

PF particle flow. 14

POI parameter of interest. 17, 19–21, 23, 39, 40, 43, 76–83, 85, 88, 91

PS Proton Synchrotron. 6

PSB Proton Synchrotron Booster. 6

qqH Higgs boson production from vector boson fusion. 9, 11, 17, 18, 26, 30, 97

ReLU rectified linear unit. 48, 63, 80

ROC receiver operating characteristic. 46, 52–54, 56, 64, 66

SM Standard Model. 3, 4, 6, 8, 9, 11, 14, 17, 23, 33, 39, 40, 42, 53, 54, 66, 83, 89–91

SPS Super Proton Synchrotron. 6

STXS simplified template cross section. 17, 18, 24, 30, 39, 89–91, 93

VH Higgs boson production associated with vector bosons. 9, 11, 17

APPENDIX C

Bibliography

[1] The CMS collaboration. "Recorded integrated luminosity at the CMS experiment". 2019. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults (visited on 05/02/2020).

[2] The CMS collaboration. "CMS Physics: Technical Design Report Volume 1: Detector Performance and Software". Technical Design Report CMS. Geneva: CERN, 2006.

[3] Sheldon L Glashow. "Partial-symmetries of weak interactions". *Nuclear physics* 22.4 (1961), pp. 579–588.

[4] Steven Weinberg. "A model of leptons". *Physical review letters* 19.21 (1967), p. 1264.

[5] Abdus Salam and John Ward. "Electromagnetic and weak interactions". *Physics Letters* 13.2 (1964), pp. 168–171.

[6] François Englert and Robert Brout. "Broken symmetry and the mass of gauge vector mesons". *Physical Review Letters* 13.9 (1964), p. 321.

[7] Peter Higgs. "Broken symmetries and the masses of gauge bosons". *Physical Review Letters* 13.16 (1964), p. 508.

[8] Gerald Guralnik, Carl Hagen, and Thomas Kibble. "Global conservation laws and massless particles". *Physical Review Letters* 13.20 (1964), p. 585.

[9] The UA1 collaboration. "Experimental observation of isolated large transverse energy electrons with associated missing energy at $\sqrt{s} = 540\,\text{GeV}$". *Phys. Lett. B* 122.CERN-EP-83-13 (Jan. 1983), 103–116. 31 p.
DOI: 10.5170/CERN-1983-004.123.

[10] The UA1 collaboration. "Experimental observation of lepton pairs of invariant mass around $95\,\text{GeV/c}^2$ at the CERN SPS collider". *Phys. Lett. B* 126.CERN-EP-83-073 (June 1985), 398–410. 17 p.
DOI: 10.1016/0370-2693(83)90188-0.

[11] Andreas Hoecker et al. "TMVA: Toolkit for Multivariate Data Analysis". *PoS* ACAT (2007), p. 040. arXiv: physics/0703039.

[12] The CMS collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". *Physics Letters B* 716.1 (Sept. 2012), pp. 30–61. ISSN: 0370-2693.
DOI: 10.1016/j.physletb.2012.08.021.

[13] The ATLAS collaboration. "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC". *Physics Letters B* 716.1 (Sept. 2012), pp. 1–29. ISSN: 0370-2693.
DOI: 10.1016/j.physletb.2012.08.020.

[14] The CMS collaboration. "Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$TeV". *Journal of Instrumentation* 10.06 (June 2015), P06005–P06005. ISSN: 1748-0221.
DOI: 10.1088/1748-0221/10/06/p06005.

[15] The CMS collaboration. "Reconstruction and identification of tau lepton decays to hadrons and tau neutrino at CMS". *Journal of Instrumentation* 11.01 (Jan. 2016), P01019–P01019. ISSN: 1748-0221.
DOI: 10.1088/1748-0221/11/01/p01019.

[16] The ATLAS collaboration. "Identification and energy calibration of hadronically decaying tau leptons with the ATLAS experiment in pp collisions at $\sqrt{s} = 8$ TeV". *The European Physical Journal C* 75.7 (July 2015). ISSN: 1434-6052.
DOI: 10.1140/epjc/s10052-015-3500-z.

[17] Olga Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252.
DOI: 10.1007/s11263-015-0816-y.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[19] The CMS collaboration. "Combined measurements of Higgs boson couplings in proton–proton collisions at $\sqrt{s} = 13\,\text{TeV}$". *The European Physical Journal C* 79.5 (May 2019). ISSN: 1434-6052.
DOI: 10.1140/epjc/s10052-019-6909-y.

[20] Apollinari G. et al. "High-Luminosity Large Hadron Collider (HL-LHC): Technical Design Report V. 0.1". CERN Yellow Reports: Monographs. Geneva: CERN, 2017.
DOI: 10.23731/CYRM-2017-004.

[21] CERN. "The CERN accelerator complex". 2016. URL: http://cds.cern.ch/record/2225847 (visited on 05/20/2020).

[22] The CMS collaboration. "Slice of the CMS detector". 2015. URL: https://cds.cern.ch/record/2628641 (visited on 05/20/2020).

[23] The CMS collaboration. "CMS Technical Design Report for the Level-1 Trigger Upgrade". Tech. rep. CERN-LHCC-2013-011. CMS-TDR-12. June 2013.

[24] Tomasz Bawej et al. "The new CMS DAQ system for run-2 of the LHC". *IEEE Transactions on Nuclear Science* 62.3 (May 2015).
DOI: 10.1109/TNS.2015.2426216.

[25] The CMS collaboration. "Evidence for the 125 GeV Higgs boson decaying to a pair of tau leptons". *Journal of High Energy Physics* 2014.5 (May 2014). ISSN: 1029-8479.
DOI: 10.1007/jhep05(2014)104.

[26] The CMS collaboration. "Summary of CMS cross section measurements". 2020. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsCombined (visited on 11/24/2020).

[27] The CMS collaboration. "CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV". Tech. rep. CMS-PAS-LUM-18-002. Geneva: CERN, 2019.

[28] The CMS collaboration. "Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 13\,\text{TeV}$". *Journal of High Energy Physics* 2018.7 (July 2018). ISSN: 1029-8479.
DOI: 10.1007/jhep07(2018)161.

[29] The CMS collaboration. "Measurement of inclusive W and Z boson production cross sections in pp collisions at $\sqrt{s} = 13\,\text{TeV}$". Tech. rep. CMS-PAS-SMP-15-004. Geneva: CERN, 2015.

[30] The LHC Higgs Cross Section Working Group collaboration. "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector". CERN Yellow Reports: Monographs. Oct. 2016.
DOI: 10.23731/CYRM-2017-002.

[31] Particle Data Group Collaboration. "Review of Particle Physics". *Phys. Rev. D* 98.3 (2018), p. 030001.
DOI: 10.1103/PhysRevD.98.030001.

[32] The CMS collaboration. "An embedding technique to determine $\tau\tau$ backgrounds in proton-proton collision data". *JINST* 14.arXiv:1903.01216. CMS-TAU-18-001-003. 06 (Mar. 2019), P06032. 57 p.
DOI: 10.1088/1748-0221/14/06/P06032.

[33] The CMS collaboration. "Measurement of the $\text{Z}\gamma^* \to \tau\tau$ cross section in pp collisions at $\sqrt{s} = 13$ TeV and validation of τ lepton analysis techniques". *Eur. Phys. J. C* 78.9 (2018), p. 708.
DOI: 10.1140/epjc/s10052-018-6146-9. arXiv: 1801.03535 [hep-ex].

[34] The CMS collaboration. "Search for additional neutral MSSM Higgs bosons in the di-tau final state in *pp* collisions at $\sqrt{s} = 13$ TeV". Tech. rep. CMS-PAS-HIG-17-020. Geneva: CERN, 2017.

[35] Sidney D. Drell and Tung-Mow Yan. "Massive Lepton-Pair Production in Hadron-Hadron Collisions at High Energies". *Phys. Rev. Lett.* 25 (13 Sept. 1970), pp. 902–902.
DOI: 10.1103/PhysRevLett.25.902.2.

[36] Lorenzo Bianchini et al. "Reconstruction of the Higgs mass in $H \to \tau\tau$ Events by Dynamical Likelihood techniques". *J. Phys. Conf. Ser.* 513 (2014). Ed. by D.L. Groep and D. Bonacorsi, p. 022035.
DOI: 10.1088/1742-6596/513/2/022035.

[37] The CMS collaboration. "Particle-flow reconstruction and global event description with the CMS detector". *Journal of Instrumentation* 12.10 (Oct. 2017), P10003–P10003. ISSN: 1748-0221.
DOI: 10.1088/1748-0221/12/10/p10003.

[38] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. "The anti-k_t jet clustering algorithm". *Journal of High Energy Physics* 2008.04 (Apr. 2008), pp. 063–063. ISSN: 1029-8479.
DOI: 10.1088/1126-6708/2008/04/063.

[39] The CMS collaboration. *Journal of Physics: Conference Series* 1085 (Sept. 2018), p. 042029.
DOI: 10.1088/1742-6596/1085/4/042029.

[40] The CMS collaboration. "Performance of CMS Muon Reconstruction in *pp* Collision Events at $\sqrt{s} = 7$ TeV". *JINST* 7 (2012), P10002.
DOI: 10.1088/1748-0221/7/10/P10002.

[41] The CMS collaboration. "Performance of reconstruction and identification of tau leptons decaying to hadrons and tau neutrinos in pp collisions at $\sqrt{s}$ =13 TeV". *Journal of Instrumentation* 13.10 (Oct. 2018), P10005–P10005. ISSN: 1748-0221.
DOI: 10.1088/1748-0221/13/10/p10005.

[42] The CMS collaboration. "Performance of the DeepTau Algorithm for the Discrimination of Taus against Jets, Electron, and Muons". Tech. rep. 2019.

[43] Andrejkovic Janik et al. "Measurement of Higgs boson properties in decays to a pair of tau leptons with full Run II data using Machine-Learning techniques" (2020). CMS analysis note, version 7.

[44] Sebastian Wozniewski. "Differential cross section measurements in the $H \to \tau\tau$ decay channel with CMS data of proton-proton collisions at the Large Hadron Collider at CERN" (2020). PhD thesis, Karlsruhe Institute of Technology.

[45] The CMS collaboration. "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV". *Journal of Instrumentation* 13.06 (June 2018), P06015–P06015.
DOI: 10.1088/1748-0221/13/06/p06015.

[46] The CMS collaboration. "Evidence for the 125 GeV Higgs boson decaying to a pair of tau leptons". *Journal of High Energy Physics* 2014.5 (May 2014). ISSN: 1029-8479.
DOI: 10.1007/jhep05(2014)104.

[47] The CMS collaboration. "Observation of the Higgs boson decay to a pair of tau leptons with the CMS detector". *Physics Letters B* 779 (Apr. 2018), pp. 283–316. ISSN: 0370-2693.
DOI: 10.1016/j.physletb.2018.02.004.

[48] The LHC Higgs cross section working group. "Simplified template cross sections working group". 2020. URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWGFiducialAndSTXS (visited on 05/28/2020).

[49] Nicolas Berger et al. "Simplified Template Cross Sections - Stage 1.1" (2019). arXiv: 1906.02754 [hep-ph].

[50] The ATLAS Collaboration Collaboration. "Procedure for the LHC Higgs boson search combination in Summer 2011". Tech. rep. CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11. Geneva: CERN, Aug. 2011.

[51] Kyle Cranmer et al. "HistFactory: A tool for creating statistical models for use with RooFit and RooStats". Tech. rep. CERN-OPEN-2012-016. Jan. 2012.

[52] Jerzy Neyman. "Outline of a theory of statistical estimation based on the classical theory of probability". *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 236.767 (1937), pp. 333–380.

[53] Jerzy Neyman and Egon Sharpe Pearson. "On the problem of the most efficient tests of statistical hypotheses". *Philosophical Transactions of the Royal Society of London. Series A* 231.694-706 (1933), pp. 289–337.

[54] Samuel S Wilks. "The large-sample distribution of the likelihood ratio for testing composite hypotheses". *The annals of mathematical statistics* 9.1 (1938), pp. 60–62.

[55] Abraham Wald. "Tests of statistical hypotheses concerning several parameters when the number of observations is large". *Transactions of the American Mathematical society* 54.3 (1943), pp. 426–482.

[56] KS Cranmer, A Lazzaro, et al. "The Roostats Project". *PoS ACAT2010* 57 (2010).

[57] Fred James. "Minuit: Function minimization and error analysis reference manual". Tech. rep. CERN, 1998.

[58] Glen Cowan et al. "Asymptotic formulae for likelihood-based tests of new physics". *The European Physical Journal C* 71.2 (2011), p. 1554.

[59] Yanyan Gao et al. "Spin determination of single-produced resonances at hadron colliders". *Physical Review D* 81.7 (Apr. 2010). ISSN: 1550-2368.
DOI: 10.1103/physrevd.81.075022.

[60] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. "Deep Learning". http://www.deeplearningbook.org. MIT Press, 2016.

[61] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 2010, pp. 249–256.

[62] Nitish Srivastava et al. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting". *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958.

[63] Andrei N Tikhonov. "Solution of incorrectly formulated problems and the regularization method" (1963).

[64] Diederik P. Kingma and Jimmy Ba. "Adam: A Method for Stochastic Optimization" (2014). arXiv: 1412.6980.

[65] Moritz Scham. "Standard Model $H \to \tau\tau$ Analysis with a Neural Network Trained on a Mix of Simulation and Data Samples". Master's thesis, Karlsruhe Institute of Technology, ETP-KA/2020-20. MA thesis. 2020.

[66] Simon Jörger. "Studies of the usage of neural networks in particle physics analyses" (2020). Master's thesis, Karlsruhe Institute of Technology, ETP-KA/2020-10.

[67] Steve Baker and Robert D Cousins. "Clarification of the use of chi-square and likelihood functions in fits to histograms". *Nuclear Instruments and Methods in Physics Research* 221.2 (1984), pp. 437–442.

[68] Robert D Cousins. "Generalization of chisquare goodness-of-fit test for binned data using saturated models, with application to histograms". *URL http://www. physics. ucla. edu/˜ cousins/stats/cousins_saturated. pdf* (2013).

[69] Howard Haber and Gordon Kane. "The search for supersymmetry: probing physics beyond the standard model". *Physics Reports* 117.2-4 (1985), pp. 75–263.

[70] Riccardo Barbieri. "Looking beyond the standard model: the supersymmetric option". *La Rivista del Nuovo Cimento (1978-1999)* 11.4 (1988), pp. 1–45.

[71] Sven Heinemeyer, Oscar Stål, and Georg Weiglein. "Interpreting the LHC Higgs search results in the MSSM". *Physics Letters B* 710.1 (2012), pp. 201–206.

[72] Philip Bechtle et al. "HiggsBounds: confronting arbitrary Higgs sectors with exclusion bounds from LEP and the Tevatron". *Computer Physics Communications* 181.1 (2010), pp. 138–167.

[73] Philip Bechtle et al. "HiggsSignals: Confronting arbitrary Higgs sectors with measurements at the Tevatron and the LHC". *The European Physical Journal C* 74.2 (2014), p. 2711.

[74] Philip Bechtle et al. "Probing the Standard Model with Higgs signal rates from the Tevatron, the LHC and a future ILC". *Journal of High Energy Physics* 2014.11 (2014), p. 39.

[75] Philip Bechtle et al. "Applying exclusion likelihoods from LHC searches to extended Higgs sectors". *The European Physical Journal C* 75.9 (2015), p. 421.

[76] Philip Bechtle et al. *HiggsBounds-5: Testing Higgs Sectors in the LHC 13 TeV Era*. 2020. arXiv: 2006.06007 [hep-ph].

[77] Bolei Zhou et al. “Learning Deep Features for Discriminative Localization” (2015). arXiv: 1512.04150 [cs.CV].

[78] Pranav Rajpurkar et al. “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning” (2017). arXiv: 1711.05225 [cs.CV].

[79] Svante Wold, Kim Esbensen, and Paul Geladi. “Principal component analysis”. *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), pp. 37–52.

[80] Martin Abadi et al. “Tensorflow: A system for large-scale machine learning”. *12th USENIX symposium on operating systems design and implementation*. 2016, pp. 265–283.

[81] Adam Paszke et al. “Automatic differentiation in pytorch” (2017).

[82] Tianqi Chen et al. “Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems”. *arXiv preprint arXiv:1512.01274* (2015).

[83] David Baehrens et al. “How to explain individual classification decisions”. *The Journal of Machine Learning Research* 11 (2010), pp. 1803–1831.

[84] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. “Deep inside convolutional networks: Visualising image classification models and saliency maps” (2014).

[85] Sebastian Bach et al. “On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation”. *PloS one* 10.7 (2015), e0130140.

[86] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. “Learning important features through propagating activation differences”. *arXiv preprint arXiv:1704.02685* (2017).

[87] Been Kim et al. “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)”. *arXiv preprint arXiv:1711.11279* (2017).

[88] Stefan Wunsch et al. “Identifying the Relevant Dependencies of the Neural Network Response on Characteristics of the Input Space”. *Computing and Software for Big Science* 2.1 (Sept. 2018). ISSN: 2510-2044. DOI: 10.1007/s41781-018-0012-1.

[89] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. “Deep sparse rectifier neural networks”. *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. 2011, pp. 315–323.

[90] Claire Adam-Bourdarios et al. “The Higgs machine learning challenge”. *J. Phys. Conf. Ser.* Vol. 664. 7. 2015.

[91] Jesse Thaler and Ken Van Tilburg. “Identifying boosted objects with N-subjettiness”. *Journal of High Energy Physics* 2011.3 (Mar. 2011). ISSN: 1029-8479. DOI: 10.1007/jhep03(2011)015.

[92] Chase Shimmin et al. “Decorrelated jet substructure tagging using adversarial neural networks”. *Physical Review D* 96.7 (Oct. 2017). ISSN: 2470-0029. DOI: 10.1103/physrevd.96.074034.

[93] Layne Bradshaw et al. "Mass agnostic jet taggers". *SciPost Phys.* 8.1 (2020), p. 011.

[94] ATLAS Collaboration Collaboration. "Performance of mass-decorrelated jet substructure observables for hadronic two-body decay tagging in ATLAS". Tech. rep. ATL-PHYS-PUB-2018-014. Geneva: CERN, July 2018.

[95] Justin Stevens and Mike Williams. "uBoost: A boosting method for producing uniform selection efficiencies from multivariate classifiers". *Journal of Instrumentation* 8.12 (2013), P12013.

[96] Alex Rogozhnikov et al. "New approaches for boosting to uniformity". *Journal of Instrumentation* 10.03 (2015), T03002.

[97] The LHCb collaboration. "Measurement of the CP-violating phase ϕ_S from $B_s^0 \to J/\psi\pi^+ + \pi^-$ decays in 13 TeV pp collisions". *Physics Letters B* 797 (2019), p. 134789. ISSN: 0370-2693.
DOI: https://doi.org/10.1016/j.physletb.2019.07.036.

[98] Ian Goodfellow et al. "Generative adversarial nets". *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[99] Gilles Louppe, Michael Kagan, and Kyle Cranmer. "Learning to pivot with adversarial networks". *Advances in neural information processing systems*. 2017, pp. 981–990.

[100] Christoph Englert et al. "Machine learning uncertainties with adversarial neural networks". *The European Physical Journal C* 79.1 (2019), pp. 1–10.

[101] Andrew Blance, Michael Spannowsky, and Philip Waite. "Adversarially-trained autoencoders for robust unsupervised new physics searches". *Journal of High Energy Physics* 2019.10 (Oct. 2019). ISSN: 1029-8479.
DOI: 10.1007/jhep10(2019)047.

[102] Constantinos Daskalakis and Ioannis Panageas. "The limit points of (optimistic) gradient descent in min-max optimization". *Advances in Neural Information Processing Systems*. 2018, pp. 9236–9246.

[103] Gregor Kasieczka and David Shih. "DisCo Fever: Robust Networks Through Distance Correlation" (2020). arXiv: 2001.05310 [hep-ph].

[104] Li-Gang Xia. "QBDT, a new boosting decision tree method with systematical uncertainties into training for High Energy Physics". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 930 (2019), pp. 15–26.

[105] Adam Elwood and Dirk Krücker. "Direct optimisation of the discovery significance when training neural networks to search for new physics in particle colliders" (2018). arXiv: 1806.00322 [hep-ex].

[106] Pablo De Castro and Tommaso Dorigo. "INFERNO: inference-aware neural optimisation". *Computer Physics Communications* 244 (2019), pp. 170–179.

[107] Tom Charnock, Guilhem Lavaux, and Benjamin D. Wandelt. "Automatic physical inference with information maximizing neural networks". *Physical Review D* 97.8 (Apr. 2018). ISSN: 2470-0029.
DOI: 10.1103/physrevd.97.083004.

[108] Stefan Wunsch et al. "Reducing the Dependence of the Neural Network Function to Systematic Uncertainties in the Input Space". *Computing and Software for Big Science* 4.1 (Feb. 2020). ISSN: 2510-2044.
DOI: 10.1007/s41781-020-00037-9.

[109] The ATLAS collaboration. "Cross-section measurements of the Higgs boson decaying into a pair of tau leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". *Physical Review D* 99.7 (Apr. 2019). ISSN: 2470-0029.
DOI: 10.1103/physrevd.99.072001.

[110] The CMS collaboration. "Measurement of Higgs boson production and decay to the $\tau\tau$ final state". Tech. rep. CMS-PAS-HIG-18-032. Geneva: CERN, 2019.

[111] The CMS collaboration. "Search for $\mathrm{t\bar{t}H}$ production in the $\mathrm{H} \to \mathrm{b\bar{b}}$ decay channel with leptonic $\mathrm{t\bar{t}}$ decays in proton-proton collisions at $\sqrt{s} = 13$ TeV". *Journal of High Energy Physics* 2019.3 (Mar. 2019). ISSN: 1029-8479.
DOI: 10.1007/jhep03(2019)026.

[112] The ATLAS collaboration. "Search for the Standard Model Higgs boson produced in association with top quarks and decaying into $b\bar{b}$ pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector". *Physical Review D* 97.7 (Apr. 2018). ISSN: 2470-0029.
DOI: 10.1103/physrevd.97.072016.

[113] The CMS collaboration. "Measurement of Higgs boson production in the decay channel with a pair of τ leptons". Tech. rep. CMS-PAS-HIG-19-010. Geneva: CERN, 2020.

[114] The CMS collaboration. "Study of the mass and spin-parity of the Higgs boson candidate via its decays to Z boson pairs". *Physical Review Letters* 110.8 (Feb. 2013). ISSN: 1079-7114.
DOI: 10.1103/physrevlett.110.081803.

[115] Andrei V. Gritsan et al. "Constraining anomalous Higgs boson couplings to the heavy-flavor fermions using matrix element techniques". *Physical Review D* 94.5 (Sept. 2016). ISSN: 2470-0029.
DOI: 10.1103/physrevd.94.055023.

[116] Andrei V. Gritsan et al. "New features in the JHU generator framework: constraining Higgs boson properties from on-shell and off-shell production" (2020). arXiv: 2002.09888 [hep-ph].

[117] Radford Neal et al. "Computing likelihood functions for high-energy physics experiments when distributions are defined by simulators with nuisance parameters" (2008).

[118] Markus Stoye et al. *Likelihood-free inference with an improved cross-entropy estimator*. 2018. arXiv: 1808.00973 [stat.ML].

[119] Johann Brehmer et al. "Constraining Effective Field Theories with Machine Learning". *Physical Review Letters* 121.11 (Sept. 2018). ISSN: 1079-7114. DOI: 10.1103/physrevlett.121.111801.

[120] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. *The frontier of simulation-based inference*. 2019. arXiv: 1911.01429 [stat.ML].

[121] Johann Brehmer et al. "Mining gold from implicit models to improve likelihood-free inference". *Proceedings of the National Academy of Sciences* 117.10 (Feb. 2020), pp. 5242–5249. ISSN: 1091-6490. DOI: 10.1073/pnas.1915980117.

[122] Johann Brehmer et al. "MadMiner: Machine learning-based inference for particle physics" (2019). arXiv: 1907.10621 [hep-ph].

[123] 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". *Computing and Software for Big Science* 5.4 (Jan. 2021). DOI: 10.1007/s41781-020-00049-5.

[124] Wouter Verkerke and David Kirkby. "The RooFit toolkit for data modeling". *Statistical Problems in Particle Physics, Astrophysics and Cosmology*. World Scientific, 2006, pp. 186–189.

[125] Ronald Aylmer Fisher. "Theory of statistical estimation". *Mathematical Proceedings of the Cambridge Philosophical Society*. Vol. 22. 5. Cambridge University Press. 1925, pp. 700–725.

[126] Harald Cramér. "Mathematical methods of statistics". Vol. 43. Princeton university press, 1999.

[127] C Radhakrishna Rao. "Information and the accuracy attainable in the estimation of statistical parameters". *Breakthroughs in statistics*. Springer, 1992, pp. 235–247.

[128] The ATLAS and CMS collaborations. "Expected pileup values at the HL-LHC". Tech. rep. Geneva: CERN, July 2013.

Acknowledgements

I thank my supervisors Günter Quast and Roger Wolf for their exceptional support since my first days in the field of particle physics in 2016. They were always available for technical discussions, but also for personal advice, and contributed significantly to the smooth progress of my journey.

Thanks to Lorenzo Moneta for his supervision, who supported me on my way at CERN from the very beginning. I am very grateful for the opportunity to work on many interesting topics and have pleasant memories of the exciting participation in cutting edge research, from Geneva over Tel Aviv to Adelaide.

I would like to thank the scientific communities in which I could follow and discuss my research. In particular many thanks to CERN and the Gentner program, which made my stay possible at such an exciting place. Thanks to the CMS collaboration, the department of experimental particle physics at KIT, the software development group of the experimental physics department at CERN and the graduate colleges GRK1964 and KSETA for being the foundation of my work.

I thank all of my colleagues at KIT and CERN for the exceptional working environment. Whether at CERN or mostly remote at KIT, I always had people around me with whom I enjoyed working very much. I will refrain from listing all the extraordinary people I have met and worked with over the last years, but be assured that I have greatly appreciated my daily life as a PhD thanks to you. Also thanks to all of you who visited me at CERN or spent time with me on various trips around the world, for professional reasons or for our own pleasure.

And a special thanks to Isabell, who made the brave decision to move with me abroad right after finishing her degree at KIT, laying ground for exciting years in France and Switzerland.