Bayesian Categorization of Higgs Boson Events in the Four-Lepton Final State with the CMS Experiment at the LHC

Master's Thesis

submitted by

Philipp Windischhofer

ETH Zürich
Department of Physics
Otto-Stern-Weg 1
8093 Zürich
Switzerland

and

École Polytechnique
Laboratoire Leprince-Ringuet
91128 Palaiseau Cedex
France

Thesis Advisor:

Roberto Salerno

June 2018

Abstract

The highlight of Run 1 of the Large Hadron Collider (LHC) was the discovery of a Higgs boson, compatible in its properties with the Standard Model of Particle Physics. With Run 2 of the LHC in full swing and, so far, no additional degrees of freedom uncovered, the focus of experimental Higgs physics is now shifting to producing precision measurements of the boson's properties.

Its decay into four leptons via two intermediate Z bosons, $H \to ZZ^* \to 4\ell$, provides a clean, fully reconstructible final state with a high signal-to-background ratio. Accessing the couplings of the Higgs boson through the study of its various production mechanisms is one application of this channel. In order to maximize the statistical power of such an analysis, it is important to be able to separate events coming from different production modes. That is, the available dataset must be partitioned into mutually orthogonal event categories with a high purity in their targeted production channels.

This thesis presents a Bayesian approach to the event categorization problem. In this framework, the central role is played by the the degree of belief that a given event belongs to a certain production mode, encoded by Bayesian posterior probabilities. Events can then be classified by assigning them to the category with the highest posterior. This category is determined through iterative, pairwise comparisons of ratios of posterior probabilities. Exploiting the invariance of likelihood ratios under a certain class of dimensionality-reducing maps, the posterior ratios can be approximately computed via Bayes' theorem.

The performance of this approach is benchmarked in the context of an analysis carried out by the CMS Experiment at the LHC, targeting the four-lepton final state. This analysis is based on $41.5\,\mathrm{fb}^{-1}$ of data collected at $\sqrt{s} = 13\,\mathrm{TeV}$ during the 2017 pp run. The proposed Bayesian method is used to replace the event classification algorithm that is employed by the published analysis. This leads to a significant increase in the sensitivity to subleading Higgs boson production modes, as quantified by the expected uncertainties in their corresponding signal strength modifiers μ. The signal strengths are defined as the ratios of the observed to the expected Higgs boson yields.

The expected uncertainty in μ_{VBF} for production through vector boson fusion is reduced by about 10%. Similar improvements are observed for other production channels, ranging from about 6% in $\mu_{t\bar{t}H,tH}$ for associated production with a top quark pair or single top, up to 15% in $\mu_{VH\text{-hadr.}}$ for associated production with a hadronically decaying W or Z boson.

Contents

1 Introduction: The World at our Fingertips **1**
1.1 A Reductionist's Guide to the Higgs . . . 1
1.2 An Elementary Scalar Particle . . . 3
1.3 Exploiting the Four-Lepton Final State with the CMS Experiment . . . 4
1.3.1 Signal Event Selection . . . 4
1.3.2 Kinematic Discriminants . . . 5
1.3.3 Event Categorization . . . 6
1.3.4 Background Estimation . . . 7
1.3.5 Statistical Analysis . . . 7
1.3.6 Systematic Uncertainties . . . 8
1.4 Into the Future . . . 9

2 A Bayesian Approach to Event Categorization **10**
2.1 Theoretical Foundations . . . 10
2.2 Practical Implementation . . . 13
2.2.1 Definition of Categories . . . 13
2.2.2 Choice of Parameterization . . . 14
2.2.3 Determining the Parameters . . . 15
2.2.3.1 Training Datasets . . . 16
2.2.3.2 Constructing the Feature Vector . . . 16
2.2.3.3 Preprocessing . . . 18
2.2.3.4 Training and Regularization . . . 18
2.2.3.5 Hyperparameters . . . 20
2.2.4 Calibration . . . 20
2.2.5 Finding the Priors . . . 21
2.2.5.1 Punzi's Purity Measure . . . 22
2.2.5.2 Bayesian Optimization . . . 24
2.2.6 Intransitive Games . . . 26
2.3 Summary . . . 27

3 Results **28**
3.1 Updating the Signal Model . . . 28
3.2 Evaluating Systematic Uncertainties . . . 28
3.3 Results . . . 30

4 Summary and Outlook: Into the High Luminosity Future **32**

Bibliography **33**

A Event Information **36**

B Reverse-Engineering Neural Networks **37**

Chapter 1
The World at our Fingertips

Modern particle physics has already come a long way in explaining the microscopic origins of our macroscopic world. The observation of a Higgs boson in 2012 represents another big leap forward, and its continued study promises ample opportunity for further insights. The conceptual origins and experimental signatures of this unique particle are introduced in Sections 1.1 and 1.2. Section 1.3 completes this introductory chapter by outlining a specific experimental approach to probe it.

1.1 A Reductionist's Guide to the Higgs[1]

The formulation of relativity and quantum mechanics was undoubtedly the central achievement of physics in the last century. Individually, these two theories have revolutionized our view of nature in very different ways, and their union forms the foundation of our present understanding of the smallest accessible length scales.

Indeed, the principles of relativity and quantum mechanics, once taken together, are extremely strong and severely constrain the phenomena that nature can produce. At its most fundamental level, the Poincaré group encodes all symmetries of a matterless, flat spacetime: rotations, translations and Lorentz boosts. These symmetries also lie at the heart of any practical notion of an elementary particle, the quantum of matter. Experimentally, particles have certain intrinsic properties on which observers in different reference frames will always agree, i.e. features that are *invariant* under Poincaré transformations. On the other hand, they can be observed in different states of motion, i.e. with different (angular) momenta, but still be considered the *same* particle. The nature of a particle, therefore, is captured by the collection of all states that can be produced by first taking a particle at rest and then transforming into all conceivable reference frames. Mathematically, a particle transforms under an irreducible representation of the Poincaré group, with its invariant properties – mass and spin – taking the role of Casimir invariants [1].

Quantum mechanics now commands the Poincaré group to act on its representations in a *unitary* way – and the only unitary representations admitted by its non-compactness are of infinite dimension. Relativistic quantum *fields* are thus singled out as a fitting description of elementary particles.

It is remarkable just *how* predictive this framework of quantum fields really is. The connection between spin and statistics, completely mysterious in nonrelativistic quantum mechanics, was found to be a consequence of microcausality and the positivity of energies. While the spin of a *free* particle is unbounded by symmetry, any *interacting* particle at low energies must have a spin of not more than two – completely independent of its behaviour at inaccessibly high scales (see Section 13.1 of [2]). Even more, there can exist only a *single* species of interacting massless spin-two particles, and they are necessarily associated with a long-ranged force that couples universally to energy, that is, gravitation[2]. While the existence of gravity as a massless spin-two field is thus intimately linked to general covariance, the presence of massless vector bosons is associated with an internal gauge redundancy of the field theory [3].

The framework of gauge theories proves to be an incredibly efficient way to encode information, where the structure of the gauge symmetry group has immediate physical consequences for the particle content of the theory. In the same spirit, the presence of a special, additional internal symmetry [4] guarantees the *unitarity* of physically relevant scattering processes, as described by the quantum theory.

Wilson explained that, in the low-energy limit, the *structure* of any observable interaction is heavily constrained as well. Whatever the correct theory prescribes at very high energy scales; at low energies, the *relevant* operators are always going to be *renormalizable* [5].

A major conceptual challenge was posed by the discovery of the mediators of the *weak interaction*, the W and Z bosons. While their self-interactions can easily be explained by a non-abelian gauge group, the fact that these bosons are *massive* is more difficult to reconcile with the consistency of a gauge theory. Especially the *longitudinal* spin polarizations of the new gauge bosons seem, at first sight, to be incompatible with perturbative unitarity and renormalizability.

Figures 1.1a and 1.1b show two possible contributions to the production of a top quark pair through the fusion of two longitudinally polarized W bosons. The cross section resulting from these two tree diagrams however surpasses the unitarity bound for centre-of-mass energies of the two W bosons in excess of about

[1] This section draws inspiration from a talk given shortly after the observation of a Higgs boson in 2012 by Nima Arkani-Hamed at the Institute for Advanced Study in Princeton. It is titled "The Inevitability of Physical Laws: Why the Higgs Has to Exist".

[2] From this perspective, the *equivalence principle*, the starting point for General Relativity, it not so much an axiom, but rather a *consequence* of quantum mechanics and Lorentz invariance.

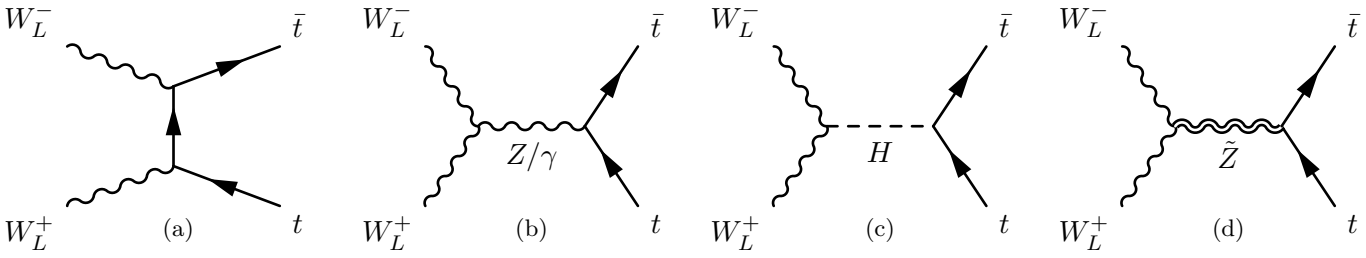


Figure 1.1: Unitarity-violating contributions to top quark pair production in (a) and (b). Possible solutions to the problem of unitarity in (c) and (d): a new vector or scalar boson can subtract the problematic contributions.

1 TeV. This problem can be made to disappear if the theory includes yet another degree of freedom. From the structure of the above diagrams, Lorentz invariance and quantum mechanics, *the only possibility* for this additional state is to have a spin of either zero or one.

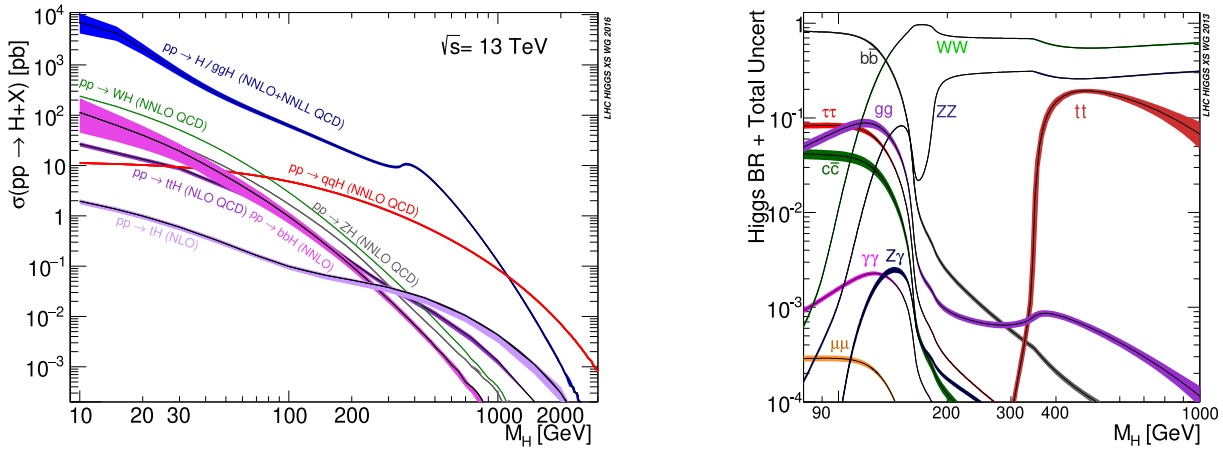
The first option is realized through the *Higgs mechanism*[3]. The new elementary scalar H, the Higgs boson, now leads to contributions such as in Figure 1.1c. By adjusting its couplings to the other states in the model, all unitarity violations can be eliminated. The calculation shows that, for this to occur, the Higgs boson coupling strengths must be *proportional to the masses* of the corresponding particles. Now, the theory in this manifestly unitary form is still nonrenormalizable. However, its S-matrix is *identical* to the physical part of the S-matrix of an enlarged, renormalizable gauge theory that also contains additional Goldstone modes, attributed to a *spontaneous reduction* of the full gauge symmetry. There exists a family of gauges that smoothly interpolate between these two models – that are therefore *both* unitary and renormalizable.

A consistent gauge theory of the weak force therefore *does* exist. Even more, its formulation makes apparent the common origins of the weak and electromagnetic interactions, indistinguishable at temperatures $T >> 100\,\text{GeV}$. In this regime, found in the very young universe, the model is governed by the gauge group $SU(3)_C \times SU(2)_L \times U(1)_Y$. As the temperature drops, a phase transition occurs, provoked by the shape of the Higgs potential. This spontaneously reduces the gauge symmetry to $SU(3)_C \times U(1)_Q$ and lifts the degeneracy between the weak and electromagnetic forces. In this theory, the Standard Model of Particle Physics (SM), the breaking leads to a *minimal* phenomenological Higgs sector, introduces a scalar Higgs boson as its single physical state and fully fixes its couplings in exactly the way required by unitarity. The only parameter that is left undetermined is its mass.

As it turns out, the Higgs boson mass m_H strongly impacts the channels through which the particle can be produced or decay, and therefore sets its phenomenological signature. Figure 1.2a illustrates the cross sections for the production of a SM Higgs boson in proton-proton (pp) collisions, $\sigma(pp \to H + X)$, and shows their dependence on the Higgs boson mass. Apart from very high m_H in excess of 1 TeV, the dominant production channel proceeds through the fusion of two gluons, ggH. Since there exists no direct coupling of the Higgs boson to gluons, this process must involve at least one quark loop, see Figure 1.3a. Owing to its high mass, the top quark provides the largest contribution. In the intermediate Higgs boson mass range from about 100 to 500 GeV, vector boson fusion (VBF) forms the next-to-leading production mode, already suppressed by one order of magnitude. Here, the Higgs boson is produced from two weak bosons (W or Z), radiated off the incoming quarks or antiquarks. These tend to produce forward-pointing quark jets in the final state, $pp \to qqH$. Vector boson fusion is illustrated in Figure 1.3b. The SM Higgs boson can also be produced in association with a $V = W, Z$ boson, see Figure 1.3c. The vector boson in the final state can decay leptonically (VH-leptonic) or hadronically (VH-hadronic), leading to different event signatures. Finally, associated production with a pair of top or bottom quarks ($t\bar{t}H$ or $b\bar{b}H$, see Figure 1.3d) or with a *single* top quark (tH) exists.

Figure 1.2b highlights the dominant decay processes of the SM Higgs boson in terms of their branching fractions. For $m_H > 160\,\text{GeV}$, the boson decays almost exclusively into two W or Z bosons (WW or ZZ) or, if kinematically allowed, into a pair of top quarks ($t\bar{t}$). In the low mass range below 90 GeV, the mass-proportional couplings to the b and c quarks as well as τ leptons determine the hierarchy of the most important decay modes, while the decay into weak bosons is suppressed. The transition between these two regimes occurs in the mass region around 120 to 140 GeV, where also the decays $H \to \gamma\gamma$ and $H \to Z\gamma$ are enhanced. This makes this part of phase space particularly rich, phenomenologically.

[3]The second possibility is, in effect, a *geometrical* version of the Higgs mechanism. A new vector boson $\tilde{Z}$ can help to achieve unitarity in $W_L W_L \to t\bar{t}$ as in Figure 1.1d, and also in $W_L W_L \to W_L W_L$ scattering. However, this new particle itself would suffer from similar problems, requiring yet another state $\tilde{\tilde{Z}}$ to regularize $\tilde{Z}$ scattering, and so on. Indeed, this infinite tower of new particles arises in theories with a compactified additional spacetime dimension, and these models are known to be perfectly unitary [6].



(a) (b)

Figure 1.2: SM Higgs boson production cross sections $\sigma(pp \to H + X)$ for different mass hypotheses at $\sqrt{s} = 13\,\text{TeV}$ are shown in (a). Figure (b) illustrates the dominant decay channels of the SM Higgs boson and their respective branching ratios as a function of m_H. Figures taken from [7, 8] (modified).

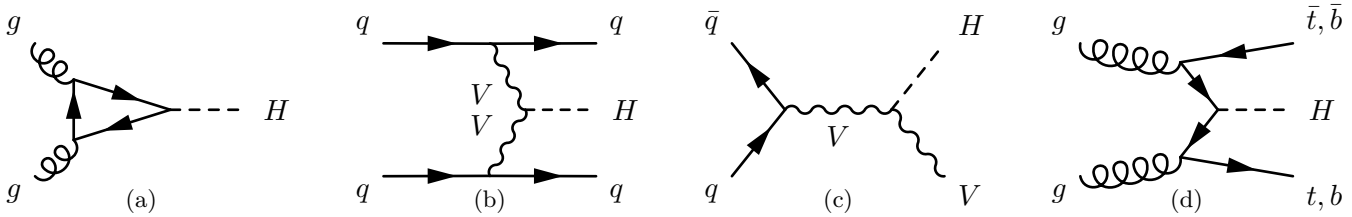


Figure 1.3: Most important SM Higgs boson production modes at hadron colliders. Figure (a) shows the leading contribution to gluon fusion, vector boson fusion is shown in (b). Associated production with a vector boson (ZH or WH) is shown in (c), while (d) illustrates Higgs boson production in association with a top or bottom quark pair, starting from a gluon initial state. The cross sections for these processes at $\sqrt{s} = 13\,\text{TeV}$ and $m_H = 125.09\,\text{GeV}$ are $\sigma_{ggH} = 48.5\,\text{pb}$, $\sigma_{\text{VBF}} = 3.8\,\text{pb}$, $\sigma_{ZH} = 0.88\,\text{pb}$, $\sigma_{WH} = 1.4\,\text{pb}$ and $\sigma_{t\bar{t}H} = 0.51\,\text{pb}$.

1.2 An Elementary Scalar Particle

In the search for the SM Higgs boson, the two high-luminosity experiments at the Large Hadron Collider (LHC), ATLAS and CMS, announced the observation of an electrically neutral, scalar particle H with a mass of about 125 GeV in July 2012 [9, 10]. This discovery was based on the analysis of approximately $5\,\text{fb}^{-1}$ of pp collision data collected at $\sqrt{s} = 7\,\text{TeV}$ and $5\,\text{fb}^{-1}$ at 8 TeV. Both experiments targeted several final states of the new boson, with $H \to ZZ$ and $H \to \gamma\gamma$ being the most sensitive. At the mass of the observed particle, these are expected to be comparably rare channels, with branching fractions of $\text{Br}(H \to ZZ) = 2.64 \cdot 10^{-2}$ and $\text{Br}(H \to \gamma\gamma) = 2.27 \cdot 10^{-3}$ respectively. However, their clean experimental signatures make them preferable to dominating decays such as $H \to b\bar{b}$. The latter in particular is very hard to observe at a hadron collider, owing to the severe QCD background.

A larger dataset became available by the end of Run 1 of the LHC in 2013, amounting to about $5\,\text{fb}^{-1}$ at $\sqrt{s} = 7\,\text{TeV}$ and $20\,\text{fb}^{-1}$ at 8 TeV. These data allowed a more thorough assessment of the boson's properties, which were found to be consistent with a spin-parity quantum number assignment of $J^{PC} = 0^{++}$ [11]. Measurements of its production and decay rates were undertaken and its couplings to other SM particles were determined with an accuracy of about 20% [12]. Also the mass measurement was refined, yielding $m_H = 125.09 \pm 0.24\,\text{GeV}$ for the full dataset [13]. The central result of Run 1, therefore, was to firmly establish the existence of a new particle, and to confirm that its properties are indeed consistent with those expected for the SM Higgs boson.

Run 2 of the LHC commenced in 2015 at an increased centre-of-mass energy of $\sqrt{s} = 13\,\text{TeV}$. By June 2018, more than $100\,\text{fb}^{-1}$ at this energy have already been delivered to ATLAS and CMS. This allows to continue the exploration of the properties of the discovered Higgs boson and to constrain production and decay modes that were not previously accessible.

More concretely, the availability of a larger dataset will allow to focus on very clean, but comparably rare

channels to aid these measurements. The process $H \to ZZ^* \to 4\ell$[4] is especially convenient. Its final state will contain four leptons originating from the primary vertex, that is, either $4e$, 4μ or $2e2\mu$. Experimentally, these *prompt* leptons can be fully and accurately reconstructed. Thus, the complete kinematic configuration of the final state is easily accessible and can be used to extract additional information about the Higgs boson couplings and potential anomalous contributions. The Higgs boson mass can be inferred with good precision from the invariant mass of the four leptons, $m_{4\ell}$. Even though the branching ratio is very small, $\mathrm{Br}(H \to ZZ^* \to 4\ell) = 1.2 \cdot 10^{-4}$, this channel features a very good signal-to-background ratio.

The irreducible backgrounds for this process arise from a ZZ or $Z\gamma^*$ pair produced from a $q\bar{q}$ initial state, $q\bar{q} \to ZZ$, or through the fusion of two gluons, $gg \to ZZ$. Also the production of a single Z boson, followed by its decay into four leptons, $Z \to 4\ell$, contributes. Reducible backgrounds involve processes where one or more prompt leptons are faked by decaying heavy-flavour hadrons or mesons within jets. These are Z + jets, $t\bar{t}$ + jets, $Z\gamma$ + jets, WZ + jets and WW + jets. This class of backgrounds is collectively denoted as $Z + X$. In the $m_{4\ell}$ signal region close to the Higgs boson peak, the irreducible $q\bar{q} \to ZZ$ and the reducible $Z + X$ backgrounds dominate. Owing to its peaking structure, the latter is particularly inconvenient and needs to be well under control.

1.3 Exploiting the Four-Lepton Final State with the CMS Experiment

The CMS Experiment is a general-purpose detector that was designed to cover a diverse physics programme. An important feature of CMS is its superconducting solenoid with an internal diameter of 6 m, providing a magnetic field of 3.8 T. A large semiconductor tracker system, a lead-tungstate crystal electromagnetic calorimeter as well as a brass/scintillator sampling calorimeter are located in the interior of the magnet. An extensive muon system based on gas-ionization detectors is integrated into the steel return yoke outside the solenoid. In-depth information about the detector can be found in [14]. The coordinate system used in the following is explained in Appendix A.1.

The work presented in this thesis is carried out in the context of a $H \to ZZ^* \to 4\ell$ analysis performed by CMS [15]. This analysis is based on $41.5\,\mathrm{fb}^{-1}$ of data, collected at $\sqrt{s} = 13\,\mathrm{TeV}$ during the LHC's 2017 pp run. The analysis also performs a combination with the 2016 dataset comprising $35.9\,\mathrm{fb}^{-1}$. The corresponding paper based on 2016 data alone can be found in [16].

The main result of this analysis is the measurement of the Higgs boson *signal strength modifiers* $\mu = \sigma^{\mathrm{obs}}/\sigma^{\mathrm{SM}}$, i.e. the scale factors for the observed signal yields relative to the yields prescribed by the SM. Deviations from unity would indicate modified couplings of the Higgs boson in its production or decay into four leptons. First, one can measure an inclusive signal strength μ_{global} that is sensitive to modifications of the total event yield, independent of the Higgs boson production mode. More information about individual couplings can be extracted by allowing the signal strengths μ_p to differ for each production process p. The analysis attempts to measure μ_p for five different production modes: ggH, VBF, $t\bar{t}H$ as well as for VH-leptonic and VH-hadronic[5]. Finally, one can also assert common signal strength modifiers for the coupling to bosons, relevant only for the VBF and VH production modes, and fermions. This leads to two free signal strength parameters only, $\mu_{\mathrm{VBF},VH}$ and $\mu_{ggH,t\bar{t}H,b\bar{b}H,tH}$.

In the following sections, the relevant aspects of the analysis will be described in some detail, more information can be found in [15].

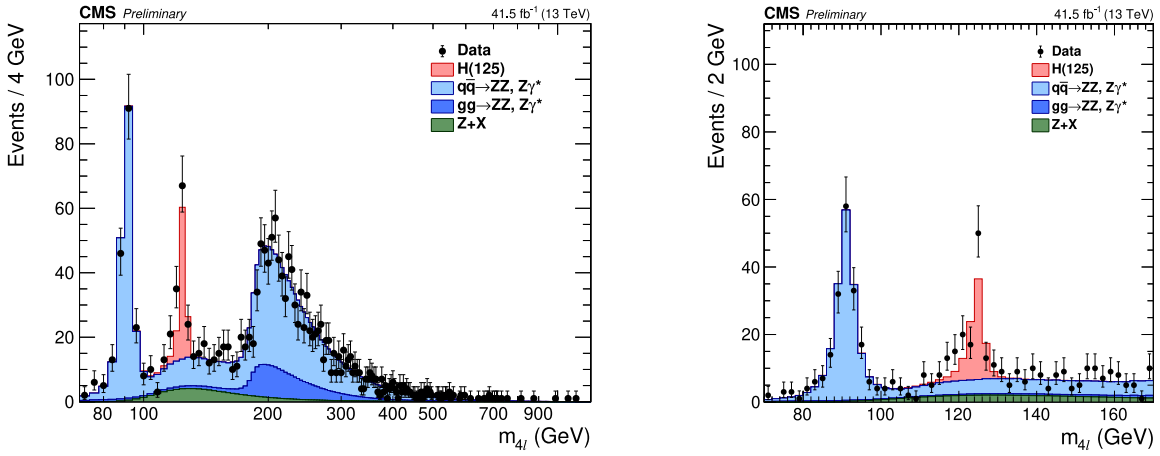
1.3.1 Signal Event Selection

The analysis builds on data that have been recorded by a series of triggers, requiring the presence of either one, two or three leptons in the event. At the level of the trigger, the leptons must satisfy a number of lax quality requirements. Higgs boson candidates are then built by starting from a subsample with at least *four* isolated and well-reconstructed leptons. To be selected, these leptons must pass a *tight* identification cut.

In a first step, Z candidates are built by combining pairs of same-flavour, opposite-sign leptons. In a second step, pairs of Z candidates are combined into ZZ candidates. For a ZZ candidate, the Z candidate whose invariant mass is closest to the nominal mass of the Z boson is labelled Z_1, the other one is labelled Z_2. The formed ZZ candidates are subject to several kinematic cuts that seek to protect against selecting an on-shell Z together with a low-mass dilepton resonance, or against incorporating non-prompt leptons. In particular, $m_{4\ell} > 70\,\mathrm{GeV}$ is required. In events that contain more than a single ZZ candidate, only the candidate with the highest associated value of $\mathcal{D}_{\mathrm{bkg}}$ (defined in Section 1.3.2) is kept. A full description of the

[4] As usual in a hadron collider context, a *lepton* will generally be an electron or a muon only.

[5] Both VH-hadronic and VH-leptonic probe the same HVV vertex. In the context of the measurement of Stage 0 Simplified Template Cross Sections (STXS) for these processes, VH-hadronic is grouped together with VBF, see the prescription in Section III.2.2.b of [7]. CMS decided to divert from this convention and to provide results for VH-hadronic and VH-leptonic separately.



(a) (b)

Figure 1.4: Shown is the observed spectrum of the four-lepton invariant mass $m_{4\ell}$ in the full experimental range in (a) and centred on the Higgs boson peak in (b). Clearly visible is the Z peak at $m_Z = 91.2\,\text{GeV}$ and the ZZ continuum starting at about $2\,m_Z$. The irreducible $gg \to ZZ$ and $q\bar{q} \to ZZ$ backgrounds as well as the reducible $Z + X$ background are also shown. Figure taken from [15].

event selection procedure can be found in [16]. The set of all selected ZZ candidates will then form the input to the subsequent analysis steps. Figure 1.4 shows the observed $m_{4\ell}$ spectrum and its composition.

1.3.2 Kinematic Discriminants

The analysis makes heavy use of the kinematics of the final state, taking into account both the decay products of the Higgs boson as well as additional particles associated with its production. Starting from this kinematic information, a number of discriminating variables are defined that seek to distinguish between Higgs signal and irreducible background, or between different Higgs boson production modes.

These discriminants are derived from first principles, i.e. they use matrix element calculations for the processes they intend to separate. That is to say, theoretical predictions for the differential production cross section or decay rate are employed for the process under study. These objects, evaluated for the kinematic configuration of an *observed* event, provide a measure of the compatibility with the underlying process.

For example, one can define a discriminant sensitive to the kinematic differences of the final state leptons in the $H \to 4\ell$ signal and $q\bar{q} \to ZZ \to 4\ell$ background as

$$\mathcal{D}_{\text{bkg}} = \left[1 + \frac{\mathcal{P}(\mathbf{\Omega}^{H\to 4\ell}|q\bar{q} \to ZZ)}{\mathcal{P}(\mathbf{\Omega}^{H\to 4\ell}|H \to 4\ell)}\right]^{-1} . \tag{1.1}$$

Here, $\mathbf{\Omega}^{H\to 4\ell}$ is the vector of kinematic observables that fully describe the decay $H \to 4\ell$. $\mathcal{P}(\mathbf{\Omega}^{H\to 4\ell}|q\bar{q} \to ZZ)$ and $\mathcal{P}(\mathbf{\Omega}^{H\to 4\ell}|H \to 4\ell)$ are the probabilities for an observed event with $\mathbf{\Omega}^{H\to 4\ell}$ to have originated from the irreducible $q\bar{q}$ background or the signal process $H \to 4\ell$. These probabilities are obtained within the MELA framework [17]. $\mathcal{P}(\mathbf{\Omega}^{H\to 4\ell}|H \to 4\ell)$ is computed from the fully differential decay rate $d\Gamma^{H\to 4\ell}/d\mathbf{\Omega}^{H\to 4\ell}$, which is known analytically. $\mathcal{P}(\mathbf{\Omega}^{H\to 4\ell}|q\bar{q} \to ZZ)$, in absence of an analytic result, is constructed directly from numerically evaluated Monte Carlo (MC) generator matrix elements for the $q\bar{q}$ background. By definition, $\mathcal{D}_{\text{bkg}}$ lies in the interval $[0, 1]$. Low values correspond to background-like kinematics.

Moreover, one can build discriminants that separate different Higgs boson production modes, drawing on information from production kinematics alone. A variable sensitive to the kinematics of VBF and ggH events, both with two observed jets, can be defined as

$$\mathcal{D}_{\text{VBF-2j},ggH} = \left[1 + \frac{\mathcal{P}(\mathbf{\Omega}^{H+JJ}|ggH)}{\mathcal{P}(\mathbf{\Omega}^{H+JJ}|\text{VBF})}\right]^{-1} . \tag{1.2}$$

Now, $\mathbf{\Omega}^{H+JJ}$ denotes the vector of variables that fully define the *production* of a Higgs boson in association with two jets. $\mathcal{P}(\mathbf{\Omega}^{H+JJ}|ggH)$ and $\mathcal{P}(\mathbf{\Omega}^{H+JJ}|\text{VBF})$ are again the probabilities to observe this kinematic configuration in a ggH or VBF event[6]. Further discriminating variables $\mathcal{D}_{p,p'}$ that separate between any two

[6]Even though ggH does not involve any jets at leading order, at higher orders, the incoming gluons can split and generate gluon jets.

production modes p and p' can easily be built by combining the probabilities $\mathcal{P}(\mathbf{\Omega}^{H+JJ}|p)$ and $\mathcal{P}(\mathbf{\Omega}^{H+JJ}|p')$ into expressions analogous to Equation 1.2. Beside the discriminant defined above, the analysis also uses $\mathcal{D}_{VH\text{-hadr.},ggH}$ and $\mathcal{D}_{\text{VBF-1j},ggH}$ that are of this type. The latter is sensitive to VBF events with only one reconstructed jet. It is obtained from $\mathcal{D}_{\text{VBF-2j},ggH}$ by integrating over the kinematics of the unobserved jet.

$\mathcal{D}_{\text{bkg}}$ as defined above is completely ignorant about the production of the Higgs boson. However, one can combine production and decay kinematics into the same discriminant and in this way specialize $\mathcal{D}_{\text{bkg}}$ to a specific production mode. This allows for a more efficient separation of the targeted Higgs boson production channel from other (perhaps more abundant) production processes as well as from reducible and irreducible background. Such *production-specific background discriminants* for the VBF and VH-hadronic production modes can be defined as

$$\begin{aligned}\mathcal{D}_{\text{VBF-2j,bkg}} &= \left[1 + c_{\text{VBF}}(m_{4\ell})\frac{\mathcal{P}(\mathbf{\Omega}|\text{VBS}) + \mathcal{P}(\mathbf{\Omega}|VVV) + \mathcal{P}(\mathbf{\Omega}|\text{QCD})}{\mathcal{P}(\mathbf{\Omega}|\text{VBF}) + \mathcal{P}(\mathbf{\Omega}|VH\text{-hadr.})}\right]^{-1}, \\ \mathcal{D}_{VH\text{-hadr.,bkg}} &= \left[1 + c_{VH\text{-hadr.}}(m_{4\ell})\frac{\mathcal{P}(\mathbf{\Omega}|\text{VBS}) + \mathcal{P}(\mathbf{\Omega}|VVV) + \mathcal{P}(\mathbf{\Omega}|\text{QCD})}{\mathcal{P}(\mathbf{\Omega}|\text{VBF}) + \mathcal{P}(\mathbf{\Omega}|VH\text{-hadr.})}\right]^{-1}.\end{aligned} \tag{1.3}$$

Beside four prompt leptons from the Higgs boson decay, production through the VBF and VH-hadronic processes leads in addition to two jets in the final state. Correspondingly, $\mathbf{\Omega}$ is now given by the union of $\mathbf{\Omega}^{H+JJ}$ and $\mathbf{\Omega}^{H\to 4\ell}$. Background processes that can generate four leptons and two jets include vector boson scattering (VBS) and tri-boson production (VVV). These are therefore taken into account in the above expressions in addition to QCD production. Note that the same signal probabilities $\mathcal{P}(\mathbf{\Omega}|\text{VBF})$ and $\mathcal{P}(\mathbf{\Omega}|VH\text{-hadr.})$ appear in both discriminants. However, this does not lead to a loss of power, since, in the context of the analysis, the discriminants will never be used to directly distinguish between events coming from VBF and VH-hadronic. In addition, the $m_{4\ell}$-dependent constants c_{VBF} and $c_{VH\text{-hadr.}}$ are introduced to optimize each discriminant separately.

The (production-specific) background discriminants in Equations 1.1 and 1.3 and the production discriminants of the type of Equation 1.2 are used for different purposes in the analysis. The latter enter into the definition of event categories, as will be explained in the next section, while the background discriminants are important for the final statistical analysis itself, see Section 1.3.5.

Appendix A.2 gives a more detailed description of the observables from which the vectors $\mathbf{\Omega}^{H+JJ}$ and $\mathbf{\Omega}^{H\to 4\ell}$ are composed.

1.3.3 Event Categorization

To improve the sensitivity of the analysis to different Higgs boson production modes, the events that pass the selection procedure in Section 1.3.1 are divided into a number of mutually exclusive categories. These categories each target a specific signal process. Currently, the analysis [15] defines seven such classes, six of which are *tagged* categories, that is, they seek to isolate a production mode where the Higgs boson is nominally accompanied by associated particles that can serve as the *tag*.

Two of the tagged categories (VBF-2jet and VBF-1jet) target the VBF production mode, with two or one reconstructed jet(s), respectively. The VH-hadronic and VH-leptonic categories target associated production with a W or Z boson which then decays hadronically or leptonically. Two further categories try to select associated production with a top quark pair, with no ($t\bar{t}H$-hadronic) or at least one ($t\bar{t}H$-leptonic) observed lepton from the subsequent $t \to Wb$ decays. The remaining *untagged* category groups all events that are not compatible with any of these tagged states.

The classification procedure that is currently employed by the analysis can be described best as a *greedy* selection, following a predefined sequence. This algorithm is illustrated in Figure 1.5. Each event is tested sequentially against the individual categories. At each step, a rectangular selection cut is applied. These cuts operate on the production discriminants $\mathcal{D}_{\text{VBF-1j},ggH}$, $\mathcal{D}_{\text{VBF-2j},ggH}$ and $\mathcal{D}_{VH-\text{hadr.},ggH}$ (as continuous variables), as well as on *primitive* event information such as the number of jets, b-tagged jets and leptons (as discrete variables). An event that passes the selection is assigned to the respective category. Failing events move on and are tested against the next category in the sequence. An event that is not selected by any of the tagged categories is automatically placed into the *untagged* bin.

Originally, the analysis employed in addition a VH-MET category. It was served by the same sequential classification procedure, with the corresponding cuts inserted downstream of those filling the $t\bar{t}H$-leptonic category in Figure 1.5. VH-MET aims to select VH events with pronounced missing transverse energy (MET). This signature arises for example for ZH events in which the Z then decays into two neutrinos. Due to problems with the modelling of the MET in 2017 data, this category had to be dropped for the final publication, as a sufficient agreement between data and MC could not be guaranteed.

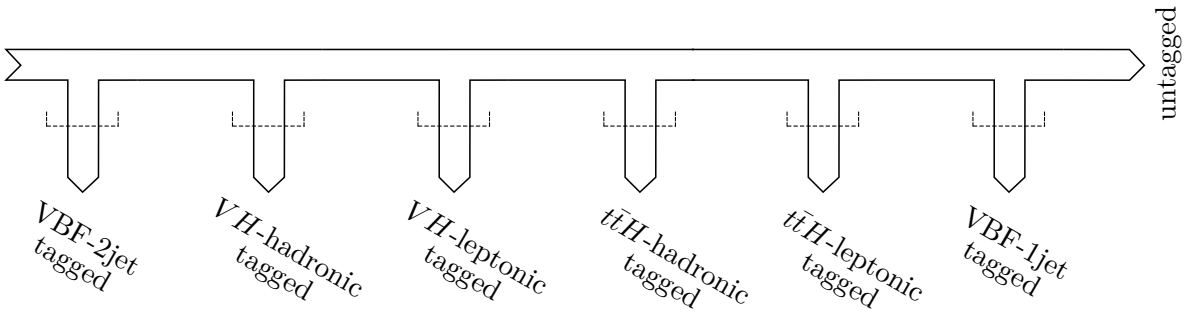


Figure 1.5: Illustration of the currently employed event categorization procedure. Events that pass the event selection enter from the left and are tested against the different categories in the sequence shown. Dashed lines indicate where selection cuts are imposed. The events are allocated to the *first* category in the sequence for which the selection criteria are satisfied.

The quality of the event categorization depends in a nontrivial way on the cuts that are used at different positions along the sequence. These cuts were originally determined through a process that resembles the construction of a decision tree. In a first step, a large number of very exclusive categories is defined, one for every possible combination of the discrete event variables listed above. The cuts on the continuous discriminating variables are chosen such as to achieve a signal efficiency of 80% (for $\mathcal{D}_{\text{VBF-2j},ggH}$ and $\mathcal{D}_{\text{VBF-1j},ggH}$) or 50% (for $\mathcal{D}_{VH\text{-hadr.},ggH}$), respectively. In a second step, these bins are manually merged into the seven final event categories in such a way as to achieve a reasonable purity in their targeted production modes.

1.3.4 Background Estimation

All measurements performed by the analysis make use of events from the *signal region*, populated by all selected events that lie in a $105 < m_{4\ell} < 140\,\text{GeV}$ mass window around the Higgs boson peak. To be able to meaningfully extract and quantify any potential signal, the characteristics of the contributing background processes must be well understood. Those of the irreducible backgrounds $gg \to ZZ$ and $q\bar{q} \to ZZ$ can be reliably estimated using simulated events from MC. On the other hand, owing to the many different physics processes that contribute to the reducible $Z + X$ background, an estimation solely from MC is not feasible. Their contribution to the signal region is thus estimated using two independent methods, based on separate data control regions.

The first method starts from a sample of events with a Z_1 candidate (defined as in Section 1.3.1) and two additional *same-flavour, opposite-sign (OS)* leptons. This dataset is then split into two control regions, both orthogonal to the signal region. These two samples contain events where exactly one or both of the additional leptons are *loose leptons*. That is, they must fail the *tight* identification cut prescribed in the signal event selection, but *pass* the identification with a relaxed working point. These two categories thus naturally target background processes that produce either two (Z + jets, $t\bar{t}$ + jets) or three prompt leptons (WZ + jets). The yield in the signal region is then estimated from the event count in these two samples and the corresponding misidentification rates, that is, the fraction of non-signal leptons that nevertheless manage to pass the signal selection cuts. These fake rates generally depend on the transverse momentum p_T and the pseudorapidity η of the lepton.

The second method uses events that have a Z_1 candidate and two additional *same-flavour, same-sign (SS)* leptons. By construction, this sample is again orthogonal to the signal region. The additional leptons are both required to be *loose*, but can also be *tight*. Then, again, the expected $Z + X$ yield in the signal region is obtained by using the corresponding *fake rates*, i.e. the fraction of *loose* leptons that are also *tight* and thereby satisfy the signal selection cuts. Since this control region – populated by SS leptons – is used to estimate the yield of fake OS leptons in the signal region, differences between events belonging to these two classes must be accounted for. This correction is captured by an additional factor of order unity, computed from MC.

The fake rates for electrons and muons are measured, separately for both methods, from event samples containing a Z_1 candidate and exactly one additional lepton that passes at least the *loose* identification cut. Both procedures result in independent estimates for the yield of the reducible background, compatible within their respective uncertainties. The final result for the $Z + X$ yield is then obtained as the weighted average of the individual estimates.

1.3.5 Statistical Analysis

The extraction of the signal strength modifiers requires the definition of a model of the analysis from a statistical point of view. The signal strengths then form the free parameters of this model. To maximize the

power of the obtained results, the model should make use of the entire information available about each event. For the present analysis, it ultimately relies on four observables per event: the event category c and final state $f \in \{4e, 4\mu, 2e2\mu\}$ assigned to it, as well as the four-lepton invariant mass $m_{4\ell}$ and a background discriminant $\mathcal{D}^c_{\text{bkg}}$. Only events from the $m_{4\ell}$ signal region are used.

To exploit the strong correlations between $m_{4\ell}$ and $\mathcal{D}^c_{\text{bkg}}$, two-dimensional probability distributions $\mathcal{L}^{\text{2D}}_{p,c,f}$ in these variables are defined for the expected signal produced by a SM Higgs boson. They are built as a product of two terms, separately for each studied production mode p, event category c and final state f,

$$\mathcal{L}^{\text{2D}}_{p,c,f}(\mathbf{s}) = \mathcal{L}_{p,c,f}(m_{4\ell}; m_H)\ \mathcal{L}_{p,c,f}(\mathcal{D}^c_{\text{bkg}}|m_{4\ell}). \tag{1.4}$$

Here, the vector $\mathbf{s}$ groups the two continuous observables, $\mathbf{s} = (m_{4\ell}, \mathcal{D}^c_{\text{bkg}})$. The background discriminant $\mathcal{D}^c_{\text{bkg}}$ is dependent on the event category. For the VBF-2jet and VH-hadronic categories, the combined discriminants introduced in Equation 1.3 are used. For all other categories, $\mathcal{D}^c_{\text{bkg}}$ is equal to $\mathcal{D}_{\text{bkg}}$ as defined in Equation 1.1. The $m_{4\ell}$ dimension in this density is unbinned due to the small number of events expected to lie in the Higgs boson peak. That is, the model for the peak shape, $\mathcal{L}_{p,c,f}(m_{4\ell}; m_H)$, uses an analytic parameterization of the $m_{4\ell}$ distribution obtained from MC. In addition, it is parameterized as a function of the Higgs boson mass m_H. This brings additional flexibility and allows to modify the mass hypothesis as new measurements become available. One can also treat m_H on the same footing as the signal strengths and *perform* a mass measurement. The conditional density $\mathcal{L}_{p,c,f}(\mathcal{D}^c_{\text{bkg}}|m_{4\ell})$ is implemented as a histogram and also taken from MC. Analogous templates are created for each background process.

Besides the parameterization of the *shape*, which is taken into account by the densities in Equation 1.4, also the signal *yields* that are expected for the SM Higgs boson are modelled. Again, these are specified for each production mode in each event category and final state, parameterized as a function of m_H. The expected background yields are also recorded for each final state and category but, of course, do not depend on m_H. The analysis model is then completely defined by the shape and yield parameterizations.

The signal strength modifiers μ_p and their uncertainties are finally determined through a maximum-likelihood fit to the observed data. Schematically, the likelihood function induced by the analysis model is

$$\mathcal{L}(\boldsymbol{\mu}) = \prod_{(c,f)} \text{Po}\left(n_{c,f,\text{obs}}; n_{s,c,f}(\boldsymbol{\mu}) + n_{b,c,f}\right) \prod_{\mathbf{s}\in(c,f)} \mathcal{L}^{\text{2D}}_{c,f}(\mathbf{s}; \boldsymbol{\mu}), \tag{1.5}$$

where the parametric dependence on m_H has been suppressed for clarity. The likelihood consists of two main contributions, sensitive to the signal yields and the signal shape, respectively.

In the first, Po denotes the density of the Poisson distribution and $\boldsymbol{\mu}$ is the vector of the μ_p. The signal yield expected for category c and final state f is denoted as $n_{s,c,f}(\boldsymbol{\mu})$. This number sums over all production modes, scaled by their respective μ_p. The expected background yield is denoted as $n_{b,c,f}$. The number of actually observed events in a bin (c, f) is written as $n_{c,f,\text{obs}}$.

In the second term, the density $\mathcal{L}^{\text{2D}}_{c,f}(\mathbf{s}; \boldsymbol{\mu})$ encodes the expected distribution of $\mathbf{s}$ in a given category and final state. It is given by the convex combination of the $\mathcal{L}^{\text{2D}}_{p,c,f}(\mathbf{s})$ over all considered signal production modes and backgrounds, weighted by their corresponding yields. The products in the likelihood run over all combinations of event categories and final states, and all events $\mathbf{s}$ in each such bin.

Note that the production mode signal strengths μ_p as defined by the analysis do not directly correspond to specific *signal processes*. For example, $\mu_{VH\text{-lept.}}$ acts as an inclusive scale factor for the rates of the processes $ZH, Z \to \nu\bar{\nu}$; $ZH, Z \to 2\ell$ and $WH, W \to \ell\nu$. One could also define exclusive signal strengths μ_s for the rate of each of these processes individually. However, they differ only in the decay of the associated Z boson and therefore all probe the same HVV coupling. Nevertheless, this concept will turn out to be useful and the μ_s will appear briefly in Section 2.2.5.

Measurements of more inclusive signal strengths such as μ_{global} or $\mu_{\text{VBF},VH}$ are made by identifying the corresponding μ_p, i.e. by varying them together during the fit. In a similar way, the subleading $b\bar{b}H$ and tH production modes are absorbed into ggH and $t\bar{t}H$, i.e. the analysis will produce scale factors $\mu_{ggH,b\bar{b}H}$ and $\mu_{t\bar{t}H,tH}$.

1.3.6 Systematic Uncertainties

The analysis is affected by a number of systematic uncertainties. The leading *experimental* sources stem from the integrated luminosity (2.3%) and the selection and reconstruction efficiencies for the signal leptons (3 - 12.5%, depending on the lepton flavour). The latter are estimated from $Z \to ee$, $Z \to \mu\mu$ and $J/\Psi \to \mu\mu$ events. This leads to larger uncertainties in the efficiencies for low-p_T electrons, which are not commonly present in these samples. However, the off-shell Z in the signal channel $H \to ZZ^*$ can produce precisely such

low-p_T electrons, which explains their big impact on the analysis. Additional experimental uncertainties arise from the estimation of the reducible background. The differences in the composition of the $Z_1 + \ell$ sample and the control regions that are used for the determination and the application of the lepton fake rates account for an uncertainty of about 35% in the $Z + X$ yields. An additional shape uncertainty in the estimated $m_{4\ell}$ distribution of this background contributes as well. These systematics also cover the statistical uncertainty arising from the limited number of events in the control regions.

The most important *theoretical* sources include uncertainties in the renormalization and factorization scales, the used parton distribution functions (PDFs) as well as the modelling of hadronization and the underlying event. The theoretical uncertainties in the prediction of the ggH cross section are handled by the prescription outlined in Section I.4.2.a of [7]. The effect of the uncertainty in the PDF set is estimated through the variations observed when using different replicas of the default PDFs, fluctuating around their nominal parameterization. These theoretical sources can lead to an uncertainty in the event yield of up to 40% for the main production modes, depending on the category. Together with the experimental uncertainty in the jet energy scale, they also represent the leading sources of uncertainty in the event categorization, i.e. they account for the *migration* of events between different categories.

All systematic uncertainties are taken into account in the signal strength fits by introducing a separate nuisance parameter for each source, leading to additional terms in the likelihood in Equation 1.5. Further practical details on the actual evaluation of the relevant systematics can be found in Section 3.2.

1.4 Into the Future

In the six years that have already passed since its discovery, the Higgs boson has been carefully scrutinized. So far, all of the studied properties, within their respective uncertainties, are fully compatible with the hypothesis of a SM Higgs boson. On the other hand, no direct evidence for the existence of additional degrees of freedom beyond the framework of the SM (BSM) could yet be obtained. These new particles are required by a range of theories that aim to extend the SM to solve some of its shortcomings, such as models of low-scale supersymmetry [18] or compact extra dimensions.

In this situation, highly precise measurements of critical observables need to be made, for they can provide *indirect* information about BSM physics at high scales. The Higgs sector itself and thus the couplings of the Higgs boson to other particles will play a central role in these efforts. In the long term, the measurement of the Higgs boson self couplings (or, equivalently, the reconstruction of the Higgs potential) will allow to assess whether the minimal symmetry breaking mechanism in the SM is indeed realized in nature, or if more complicated dynamics are involved [19–21].

The data sample of about $150\,\mathrm{fb}^{-1}$ that will have been collected by the end of Run 2 of the LHC is a first step and a necessary prerequisite to embark on this journey. On the other hand, it is the responsibility of the experiments to take all necessary steps in order to use the information contained in the recorded data to the fullest extent and, given the huge scale of particle physics experiments, the available resources in the most economic way possible.

The present thesis seeks to take a small step in this direction. Indeed, the four-lepton final state with its wealth of available kinematic information provides an ideal starting point and testbed. This includes the kinematic discriminants that manage to condense much of the raw kinematic information into a small number of powerful observables, but also concerns this primitive, low-level event information itself.

At the same time, owing to the variety of production modes that are accessible through the four-lepton channel, its observables will span a large dynamic range of characteristics. There are many places where a global, unified view of the entire information content would be beneficial to enhance the sensitivity of an analysis. Naturally, such a paradigm is most powerful and appealing when one tries to exploit small, systematic differences between events, that is, during the phase of *event categorization*.

Chapter 2 will present a classification algorithm, based on Bayesian decision theory, that is capable of combining all of the information in an event in a flexible and transparent manner. In Chapter 3, this method will replace the event categorization that is currently used in the $H \to 4l$ analysis outlined above. Improvements in the expected sensitivity of the analysis will be evaluated and quantified. Chapter 4 finally summarizes the results and gives an outlook to future developments.

Chapter 2

A Bayesian Approach to Event Categorization

Statistical decision theory [22] provides a rigorous and theoretically well-founded framework to decision-finding based on statistical observations, i.e. in incompletely known situations. In this chapter, we will adopt the attitude of a Bayesian statistician and formulate the event categorization problem as a question in Bayesian decision theory [23]. In Section 2.1, this will allow us to derive a classification algorithm that minimizes the total probability that an event is misclassified. Most steps will be presented with the aforementioned $H \to 4\ell$ analysis in mind, but easily generalize to other situations. Section 2.2 then shows how the necessary computations can be organized and carried out in practice. Section 2.3 will summarize the central points.

2.1 Theoretical Foundations

Given an event, we can associate with it a *feature vector* $\mathbf{e}$ which contains all observables that are deemed necessary to approach the categorization problem. Individual events will generally have fluctuating feature vectors, even if they come from the same physical process; each observed instance of $\mathbf{e}$ therefore corresponds to a realization of a random vector $\mathbf{E}$. Given such a realization $\mathbf{e}$, the goal is then to find the event category $\bar{c}$ that provides the highest degree of compatibility. In principle, the set of categories – the *hypothesis space* – can be continuous. However, for the application discussed here, we can directly specialize to finitely many event categories.

To make this idea more precise, we need to define a *cost function* that allows to rate the quality of any given categorization rule. Note that the feature vector $\mathbf{e}$ need not be unique, i.e. any given $\mathbf{e}$ may well be compatible with more than one event category. Also, more than one physical process may produce events with (nearly) the same $\mathbf{e}$. The only notion that is well-defined, therefore, is the concept of an *expected* cost.

Let us denote by $l(c, c')$ the cost (or loss) generated if an event targeted by category c' is (perhaps wrongly) assigned to category c. Given an event $\mathbf{e}$ and a *categorization rule* that places $\mathbf{e}$ into category $c = c(\mathbf{e})$, the *expected* cost for this event will take the form

$$L(\mathbf{e}, c) = \sum_{c' \in \mathcal{C}} l(c, c')\, p(c'|\mathbf{e}), \tag{2.1}$$

and hence serves as a measure for the accuracy of the categorization rule. Here, $\mathcal{C}$ is the set of all defined event categories and $p(c'|\mathbf{e})$ is the probability[1] that an event with feature vector $\mathbf{e}$ actually belongs to category $c' \in \mathcal{C}$. In practice, the set $\mathcal{C}$ of event categories should be inclusive enough such that it can fully accommodate any event. The $p(c'|\mathbf{e})$ will then satisfy a completeness relation, $\sum_{c' \in \mathcal{C}} p(c'|\mathbf{e}) = 1\ \forall \mathbf{e}$.

Now, any *misclassified* event will not contribute to the *signal* of its own category c', but will show up as *background* to category c instead[2]. Therefore, it is natural to choose $l(c, c')$ to be of the form

$$l(c, c') = 1 - \delta_{cc'}, \tag{2.2}$$

that is, *any* misclassified event will result in the same contribution to the overall loss, independent of which categories are actually involved.

With this choice of cost function, the optimal categorization rule $\bar{c}(\mathbf{e})$ should now minimize the cost in Equation 2.1. Thus,

$$\bar{c}(\mathbf{e}) = \arg\min_{c \in \mathcal{C}} L(\mathbf{e}, c) = \arg\max_{c \in \mathcal{C}} p(c|\mathbf{e}), \tag{2.3}$$

i.e. the optimal classifier should always choose the category with the highest probability $p(c|\mathbf{e})$.

We can now use Bayes' theorem to relate the *posterior* $p(c|\mathbf{e})$ to the *likelihood* for category c, $p_{\mathbf{E}}(\mathbf{e}|c)$, and the *prior probability* $p(c)$. Therefore, we can also write Equation 2.3 as

[1] Of course, this (Bayesian) probability depends heavily on the definition of the event categories, i.e. the characteristics of the events targeted by them. See Section 2.2.1 for more details.

[2] At this abstract level, "background" refers to all events that are placed into an undesired category, regardless of whether they belong to the $H \to 4\ell$ signal or the reducible and irreducible backgrounds. This subtlety will show up again several times in what follows.

$$\bar{c}(\mathbf{e}) = \arg\max_{c \in \mathcal{C}} p_{\mathbf{E}}(\mathbf{e}|c)\, p(c). \tag{2.4}$$

The *evidence* $p(\mathbf{e}) = \sum_{c'} p_{\mathbf{E}}(\mathbf{e}|c')\, p(c')$, which appears in the denominator of Bayes' theorem, is irrelevant for our purposes. It only guarantees the correct overall normalization of the posterior as a probability, but will not modify *which* category actually exhibits the highest posterior in absolute terms. Indeed, any such global rescaling of the posteriors must be irrelevant for the task of finding $\bar{c}$, so the true discriminative power does not lie in the numerical value of $p(c|\mathbf{e}) \propto p_{\mathbf{E}}(\mathbf{e}|c)\, p(c)$ itself, but rather in the ratios

$$R_{c,c'}(\mathbf{e}) = \frac{p(c|\mathbf{e})}{p(c'|\mathbf{e})} = \frac{p_{\mathbf{E}}(\mathbf{e}|c)}{p_{\mathbf{E}}(\mathbf{e}|c')} \frac{p(c)}{p(c')} = r_{c,c'}(\mathbf{e}) \frac{p(c)}{p(c')} \tag{2.5}$$

for $c \neq c'$. Here, $r_{c,c'}$ is the ratio of the likelihoods $p_{\mathbf{E}}(\mathbf{e}|c)$ and $p_{\mathbf{E}}(\mathbf{e}|c')$.

In particular, knowing all the ratios specified in Equation 2.5 is equivalent to knowing the solution to Equation 2.4. Operatively, one could compute $\bar{c}$ by employing a "voting" procedure: for every ratio $R_{c,c'}$, the category c will receive a vote if $R_{c,c'} > 1$ and c' will receive a vote if instead $R_{c,c'} < 1$. At the end, the category with the highest posterior will have received the largest number of votes[3][4].

Unfortunately, in particle physics, the likelihoods $p_{\mathbf{E}}(\mathbf{e}|c)$ (and therefore also likelihood ratios, such as the ones appearing in Equation 2.5) are virtually never known directly, i.e. available in a way that would permit a straightforward numerical evaluation. Rather, likelihoods are used as generative models from which simulated events $\mathbf{e}$ are drawn following a MC procedure.

One might envisage to reconstruct the full likelihood (and therefore the posterior) by applying a nonparametric density estimator to a set of simulated events. However, the feature vector $\mathbf{e}$ usually is of very high dimension, and so a prohibitively large number of events would be needed in order to ensure a sufficient quality of the reconstructed probability distribution.

However, for our purposes, not the likelihoods $p_{\mathbf{E}}(\mathbf{e}|c)$ themselves are needed, but only their *ratios*. Those turn out to be invariant under a certain class of dimensionality-reducing maps. The main result of the paper [24], as far as the present application is concerned, lies in the following

Theorem 1. *Let $\mathbf{E}$ be a random vector with realizations $\mathbf{e}$ in $U \subseteq \mathbb{R}^n$ and probability density $p_{\mathbf{E}}(\mathbf{e}|c)$ and let $s_{c,c'} : \mathbb{R}^n \to \mathbb{R}$ be a function monotonic with the density ratio $r_{c,c'} = p_{\mathbf{E}}(\mathbf{e}|c)/p_{\mathbf{E}}(\mathbf{e}|c')$, for given categories c and c'. Then, the density ratio is invariant under the transformation induced by $s_{c,c'}$,*

$$r_{c,c'}(\mathbf{e}) = \frac{p_{\mathbf{E}}(\mathbf{e}|c)}{p_{\mathbf{E}}(\mathbf{e}|c')} = \frac{p_{\mathrm{U}}(u = s_{c,c'}(\mathbf{e})|c)}{p_{\mathrm{U}}(u = s_{c,c'}(\mathbf{e})|c')}, \tag{2.6}$$

where $p_{\mathrm{U}}(u = s_{c,c'}(\mathbf{e}))$ is the probability density of $\mathrm{U} = s_{c,c'}(\mathbf{E})$.

A proof of this theorem is presented in [24]. Below, we give an alternative, simplified proof, somewhat similar in spirit.

Proof. We first show that the theorem holds in the special case where $s_{c,c'}(\mathbf{e}) = r_{c,c'}(\mathbf{e})$, i.e. that

$$\frac{p_{\mathrm{U}}(u = r_{c,c'}(\mathbf{e})|c)}{p_{\mathrm{U}}(u = r_{c,c'}(\mathbf{e})|c')} = r_{c,c'}(\mathbf{e}).$$

Starting from the definition of the induced density $p_{\mathrm{U}}(u = r_{c,c'}(\mathbf{e})|c)$, we indeed find

$$\begin{aligned} p_{\mathrm{U}}(u = r_{c,c'}(\mathbf{e})|c) &= \frac{d}{du'}\bigg|_u \int_{r_{c,c'}(\mathbf{e}')<u'} d\mathbf{e}'\, p_{\mathbf{E}}(\mathbf{e}'|c) = \frac{d}{du'}\bigg|_u \int d\mathbf{e}'\, \theta(u' - r_{c,c'}(\mathbf{e}'))\, p_{\mathbf{E}}(\mathbf{e}'|c) = \\ &= \int d\mathbf{e}'\, \delta(u - r_{c,c'}(\mathbf{e}'))\, p_{\mathbf{E}}(\mathbf{e}'|c) = \int d\mathbf{e}'\, \delta(u - r_{c,c'}(\mathbf{e}'))\, r_{c,c'}(\mathbf{e}')\, p_{\mathbf{E}}(\mathbf{e}'|c') = \\ &= r_{c,c'}(\mathbf{e}) \int d\mathbf{e}'\, \delta(u - r_{c,c'}(\mathbf{e}'))\, p_{\mathbf{E}}(\mathbf{e}'|c') = r_{c,c'}(\mathbf{e}) \cdot p_{\mathrm{U}}(u = r_{c,c'}(\mathbf{e})|c'), \end{aligned} \tag{2.7}$$

where the fourth equality holds due to the definition of $r_{c,c'}$.

[3] This is simply the fact that in a finite set of real numbers, *only* the largest number will have the property that it is larger than *any* of the other numbers.

[4] See Section 2.2.6 for what happens if two categories receive the same number of votes.

The full theorem now follows from a change of variables. Since, by assumption, $s_{c,c'}(\mathbf{e})$ and $r_{c,c'}(\mathbf{e})$ are related by a monotonous transformation $s_{c,c'}(\mathbf{e}) = s_{c,c'}(r_{c,c'}(\mathbf{e}))$, their respective densities are linked only by a rescaling

$$p_{\mathrm{U}}(u = s_{c,c'}(\mathbf{e})|c) = p_{\mathrm{U}}(u = r_{c,c'}(\mathbf{e})|c) \cdot \left| \frac{dr_{c,c'}}{ds_{c,c'}} \right|, \tag{2.8}$$

and therefore

$$r_{c,c'}(\mathbf{e}) = \frac{p_{\mathrm{U}}(u = r_{c,c'}(\mathbf{e})|c)}{p_{\mathrm{U}}(u = r_{c,c'}(\mathbf{e})|c')} = \frac{p_{\mathrm{U}}(u = s_{c,c'}(\mathbf{e})|c)}{p_{\mathrm{U}}(u = s_{c,c'}(\mathbf{e})|c')}. \tag{2.9}$$

□

With the aid of this result, once such a map $s_{c,c'}(\mathbf{e})$ is available, the likelihood ratio $r_{c,c'}$ can be determined by estimating two *one-dimensional* distributions p_{U} and computing their ratio. This represents a substantial simplification of the original problem, also in terms of the size required for the dataset on which the density estimation is performed.

Furthermore, as the authors of [24] continue to point out, the problem of constructing an appropriate function $s_{c,c'}(\mathbf{e})$ is equivalent to finding the solution to a functional minimization problem, as stated by

Theorem 2. *Let* $\mathbf{E}$ *be a random vector with realizations* $\mathbf{e}$ *in* $U \subseteq \mathbb{R}^n$ *and probability density* $p_{\mathbf{E}}(\mathbf{e}|c)$*. For any fixed pair of categories* (c, c') *with* $c \neq c'$*, let* $\mathcal{E}_c$*,* $\mathcal{E}_{c'}$ *be collections of realizations of* $\mathbf{E}$ *drawn from the distributions* $p_{\mathbf{E}}(\mathbf{e}|c)$*,* $p_{\mathbf{E}}(\mathbf{e}|c')$ *and let their cardinalities be* $|\mathcal{E}_c| = N_c$ *and* $|\mathcal{E}_{c'}| = N_{c'}$*. For any* $\mathbf{e} \in \mathcal{E}_c \cup \mathcal{E}_{c'}$*, define the indicator function* $y(\mathbf{e}) = 1$ *if* $\mathbf{e} \in \mathcal{E}_c$ *and* $y(\mathbf{e}) = 0$ *if* $\mathbf{e} \in \mathcal{E}_{c'}$*. Furthermore, define the weights* $w(\mathbf{e}) = w_c \in \mathbb{R}$ *if* $\mathbf{e} \in \mathcal{E}_c$ *and* $w(\mathbf{e}) = w_{c'} \in \mathbb{R}$ *if* $\mathbf{e} \in \mathcal{E}_{c'}$*.*

Then, in the limit $N = N_c + N_{c'} \to \infty$ *with finite* $\frac{N_c}{N} \to \mu_c$ *and* $\frac{N_{c'}}{N} \to \mu_{c'}$ *the function* $s_{c,c'} : \mathbb{R}^n \to \mathbb{R}$ *with*

$$s_{c,c'} = \arg\min_{\mathfrak{s}} L[\mathfrak{s}] = \arg\min_{\mathfrak{s}} \frac{1}{N} \sum_{\mathbf{e} \in \mathcal{E}_c \cup \mathcal{E}_{c'}} w(\mathbf{e}) \cdot [y(\mathbf{e}) - \mathfrak{s}(\mathbf{e})]^2 \tag{2.10}$$

is a monotonous function of the density ratio $r_{c,c'}(\mathbf{e}) = \frac{p_{\mathbf{E}}(\mathbf{e}|c)}{p_{\mathbf{E}}(\mathbf{e}|c')}$*.*

Proof. With the definitions above, the sum in Equation 2.10 separates,

$$L[\mathfrak{s}] = w_c \frac{N_c}{N} \frac{1}{N_c} \sum_{\mathbf{e} \in \mathcal{E}_c} [1 - \mathfrak{s}(\mathbf{e})]^2 + w_{c'} \frac{N_{c'}}{N} \frac{1}{N_{c'}} \sum_{\mathbf{e} \in \mathcal{E}_{c'}} \mathfrak{s}(\mathbf{e})^2. \tag{2.11}$$

In the limit indicated, this becomes

$$\lim_{N \to \infty} L[\mathfrak{s}] = \int d\mathbf{e} \left[w_c \mu_c \, p_{\mathbf{E}}(\mathbf{e}|c) \, [1 - \mathfrak{s}(\mathbf{e})]^2 + w_{c'} \mu_{c'} \, p_{\mathbf{E}}(\mathbf{e}|c') \, \mathfrak{s}(\mathbf{e})^2 \right]. \tag{2.12}$$

An extremum can be found in the usual way by demanding stationary w.r.t. arbitrary variations $\delta\mathfrak{s}$,

$$0 = \int d\mathbf{e} \; \delta s_{c,c'}(\mathbf{e}) \left[-w_c \mu_c \, p_{\mathbf{E}}(\mathbf{e}|c) \, [1 - s_{c,c'}(\mathbf{e})] + w_{c'} \mu_{c'} \, p_{\mathbf{E}}(\mathbf{e}|c') \, s_{c,c'}(\mathbf{e}) \right], \tag{2.13}$$

and so

$$s_{c,c'}(\mathbf{e}) = \frac{1}{1 + \frac{w_{c'}}{w_c} \frac{\mu_{c'}}{\mu_c} \frac{p_{\mathbf{E}}(\mathbf{e}|c')}{p_{\mathbf{E}}(\mathbf{e}|c)}}, \tag{2.14}$$

which is indeed a monotonous function of the likelihood ratio $r_{c,c'}$. That this extremum is indeed a *minimum* is easily seen from the structure of the functional, or by taking second functional derivatives. Note that the choice $w_c = \frac{1}{\mu_c}$ and $w_{c'} = \frac{1}{\mu_{c'}}$ for the weights allows to further simplify this expression and yields a form completely analogous to the discriminants in Section 1.3.2.

Note also that the functional $L[\mathfrak{s}]$ in Equation 2.10 is nothing but the *mean squared error* (MSE) induced by a variable $\mathfrak{s}(\mathbf{e})$ that seeks to reproduce the indicator function $y(\mathbf{e})$, that is, to split $\mathcal{E}_c \cup \mathcal{E}_{c'}$ back into $\mathcal{E}_c$ and $\mathcal{E}_{c'}$. □

Satisfying the minimality condition in Equation 2.10 is a much more concrete goal – and easier to achieve approximately – than to find a function $s_{c,c'}(\mathbf{e})$ with the abstract property of being monotonous with the ratio $r_{c,c'}$. Exploiting the connection between these two problems, a strategy to solve the original classification problem in Equation 2.4 has become apparent:

- Given a set $\mathcal{C}$ consisting of $|\mathcal{C}|$ event categories, form all $\binom{|\mathcal{C}|}{2}$ possible pairs of categories $(c, c') \in \mathcal{C} \times \mathcal{C}$ with $c \neq c'$.
- For each pair (c, c'), choose a sufficiently general parameterization of a function $s_{c,c'} : \mathbb{R}^n \to \mathbb{R}$. Determine the parameters by requiring the functional in Equation 2.10 to be minimized. According to Theorem 2, this will automatically result in the fitted $s_{c,c'}(\mathbf{e})$ being monotonous with $r_{c,c'}$. The collections $\mathcal{E}_c$ and $\mathcal{E}_{c'}$ that are used in this step form part of the definition of the algorithm and must be specified externally.
- For all functions $s_{c,c'}$, estimate the one-dimensional distributions $p_{\mathrm{U}}(u = s_{c,c'}(\mathbf{e})|c)$ and $p_{\mathrm{U}}(u = s_{c,c'}(\mathbf{e})|c')$. Make use of Theorem 1 to (approximately) compute $r_{c,c'}$, and, given the priors $p(c)$ and $p(c')$, also find $R_{c,c'}$. As yet, the priors are not specified and form the free parameters of the classification algorithm.
- Use $R_{c,c'}$ computed for all pairs (c, c') to pick the category with maximum posterior $p_{\mathbf{E}}(c|\mathbf{e})\, p(c)$ and thus determine the solution to Equation 2.4.

2.2 Practical Implementation

With the conceptual structure and the computational steps of the algorithm now specified, the following sections will describe their practical realization. First, Section 2.2.1 will explain the definition of the event categories $c \in \mathcal{C}$ in the Bayesian framework. Section 2.2.2 will fix a convenient parameterization of the functions $s_{c,c'}$. Sections 2.2.3 and 2.2.4 explain how these functions are computed in practice and how the likelihood ratios $r_{c,c'}$ are estimated. Section 2.2.5 will fix the category priors, allowing Section 2.2.6 to finally determine the optimal category for each event. More implementation-specific details, a tutorial of how to use the developed tools as well as the complete source code of the project can be found in [25].

2.2.1 Definition of Categories

In Bayesian classification, the defining properties of the individual categories, i.e. the type of events they target, are encoded in the posteriors $p(c|\mathbf{e})$. Ultimately, these probabilities are defined through the functions $s_{c,c'}$, which in turn encapsulate the properties of the collections $\mathcal{E}_c$ and $\mathcal{E}_{c'}$ from which they are determined. In this way, we are led to define the individual event categories *implicitly* based on the samples in the collections $\mathcal{E}_c$. These contain events drawn from the distributions $p_{\mathbf{E}}(\mathbf{e}|c)$, which specify the intended content of the categories. This approach is typical of *data-driven* methods that build directly on the expected event characteristics rather than any a-priori (physics) knowledge.

The set $\mathcal{C}$ of categories and their associated $\mathcal{E}_c$ must therefore be specified. This must be done in a way that recognizes some particularities of the Bayesian method on the one hand and maintains compatibility with the event categories defined for the analysis on the other hand. As we will see, the event categories on which the Bayesian classification algorithm operates will turn out to be *different* from those appearing in the signal model of Section 1.3.5. A mapping is therefore defined to relate the two sets.

Some of the categories used by the analysis will consist of a mixture of different physics processes. For example, the VH-leptonic category targets both ZH, $Z \to 2\ell$ and WH, $W \to \ell\nu$. Its likelihood $p_{\mathbf{E}}(\mathbf{e}|VH\text{-lept.})$ therefore is some superposition of $p_{\mathbf{E}}(\mathbf{e}|ZH\text{-lept.})$ and $p_{\mathbf{E}}(\mathbf{e}|WH\text{-lept.})$, each of which is expected to have different characteristics. To respect these differences, it was decided to clearly separate the two processes for the purposes of the classification. That is, independent ZH-leptonic and WH-leptonic categories are kept and then merged at the end. The situation for VH-hadronic is handled in the same way. Conversely, the event categories VBF-1jet and VBF-2jet target the same process and thus contain events with similar characteristics (e.g. very forward jets). Therefore, the classification algorithm manages only a *single* VBF category. Events classified as such are then split into VBF-1jet and VBF-2jet depending on the number of jets in the event. Events with no reconstructed jets are routed into *untagged* instead[5].

Furthermore, new categories are introduced for the two background processes that are most dominant in the signal region, $Z + X$ and $q\bar{q} \to ZZ$. Events coming from one of these are mapped to the *untagged* category of the analysis[6]. No dedicated categories are defined for the remaining $gg \to ZZ$ background and the tH and $b\bar{b}H$ signal processes. This means that the completeness property of $p(c|\mathbf{e})$ mentioned below Equation 2.1 will not be strictly fulfilled. However, the yields of these processes are negligible in the $m_{4\ell}$ domain close to the Higgs boson peak. Completeness thus holds to a very good approximation for the relevant events in the signal

[5] These events turn out to be about 5% pure in VBF, similar to the *untagged* category.

[6] In principle, one should keep these additional background-enriched categories and also include them in the statistical model of the analysis. Here, we strive to disentangle the event classification and the signal strength fit, and therefore deliver the result of the event classification in terms of the originally defined categories. This entitles abusing the *untagged* category to handle the background.

classification category $c \in \mathcal{C}$	targeted signal process	analysis category $c \in \mathcal{C}_A$
VH-MET	ZH, $Z \to \nu\bar{\nu}$	VH-MET
$t\bar{t}H$-hadronic	$t\bar{t}H$, $t\bar{t} \to 0\ell + X$	$t\bar{t}H$-hadronic
$t\bar{t}H$-leptonic	$t\bar{t}H$, $t\bar{t} \to 1, 2\ell + X$	$t\bar{t}H$-leptonic
ZH-leptonic WH-leptonic	ZH, $Z \to 2\ell$ WH, $W \to \ell\nu$	VH-leptonic
ZH-hadronic WH-hadronic	ZH, $Z \to 0\ell\, 0\nu + X$ WH, $W \to 0\ell + X$	VH-hadronic
VBF	VBF	VBF-1jet VBF-2jet
ggH $Z + X$ $q\bar{q} \to ZZ$	ggH $Z + X$ $q\bar{q} \to ZZ$	untagged

Table 2.1: Shown is the set of 11 categories used internally by the event classification algorithm and their targeted physics processes that determine $\mathcal{E}_c$. Here, X denotes anything other than an explicitly mentioned lepton or neutrino. The third column indicates the mapping to the eight event categories originally employed by the analysis.

region. To see why completeness is important, consider the case of trying to classify an event that is targeted by a hypothetical category c_h not defined in $\mathcal{C}$. Then, the posteriors $p(c|\mathbf{e})$ for $c \in \mathcal{C}$ may or may not express any clear preference. In any case, the category eventually assigned to this event will be incorrect. Even worse, since the deciding posterior $p(c_h|\mathbf{e})$ is never even evaluated, this decision will be made based only on the noise in the $p(c|\mathbf{e})$ for $c \in \mathcal{C}$.

The analysis suffered from modelling issues for the missing transverse energy (MET). These could not be resolved in time for publication and led to the elimination of the VH-MET category. However, the unsatisfactory agreement of the MET in data and MC is irrelevant for our purposes, as a full unblinding of the analysis is not required. The VH-MET category is therefore re-introduced and a set of eight event categories is used in the signal model, as originally foreseen.

Note that, conceptually, there is a difference between the set $\mathcal{C}$ of categories used internally by the event classification and those that enter into the statistical model of the analysis. We will thus denote the collection of the eight analysis categories as $\mathcal{C}_A$ instead. Also, a category index that can take values in $\mathcal{C}$ will be labelled by a calligraphic c, while a straight c will index $\mathcal{C}_A$, as before. Table 2.1 shows a summary of the eleven event categories in $\mathcal{C}$, the signal processes they target, and how they relate to those in $\mathcal{C}_A$.

2.2.2 Choice of Parameterization

For the purpose of choosing a suitable parameterization of the dimensionality-reducing functions $s_{c,c'}$, consider again Theorem 2. Its proof shows that, by placing a cut on $s_{c,c'}$, we obtain a *binary* categorization rule that can separate between events from categories c and c'. Conversely, any algorithm that can build such a binary classifier with continuous output will result in some function[7] $\tilde{s}_{c,c'}(\mathbf{e})$. This candidate will then be a suitable function $s_{c,c'}$ if it also minimizes the MSE in Equation 2.10.

A binary classification algorithm is said to be *universally consistent* if it is able to achieve (probabilistic) convergence to the MSE minimum independent of the characteristics of the events in $\mathcal{E}_c \cup \mathcal{E}_{c'}$. Many algorithms are known to have this property and can therefore be used to parameterize $s_{c,c'}$, among them kernel-based methods, AdaBoost as a classification meta-algorithm [26] and classifiers based on artificial neural networks [27]. We expect $s_{c,c'}$ to be a reasonably smooth function of the feature vector $\mathbf{e}$. Therefore, we should prefer models that manifestly lead to smooth parameterizations. In this spirit, multilayer perceptrons, a simple class of feedforward artificial neural networks, provide an ideal candidate[8].

Multilayer perceptrons (MLPs) aim to implement a highly idealized and simplified model of a network of biological neurons. In such systems, many individual neurons are densely connected with one another through a network of synapses. The central feature of biological neurons is their nonlinearity: a neuron becomes active and "fires" (i.e. produces an electrical signal at its output) only if the sum of the signals received at its inputs

[7] Since these algorithms explicitly use $\mathcal{E}_c \cup \mathcal{E}_{c'}$ to find $\tilde{s}_{c,c'}$, one refers to this event sample also as the *training dataset*. The MSE then compares the truth about an event's identity – as encoded by the indicator function – with the actual numerical output of the classifier. Since the truth is available for every event in the training dataset, the process of building a classifier in this way is known as *supervised training*.

[8] Beside the requirement that the MSE be minimized, the implemented model for $s_{c,c'}$ must *generalize* well, i.e. allow extrapolation to events not found in the training dataset. For neural networks, this point is touched upon again in Section 2.2.3.4.

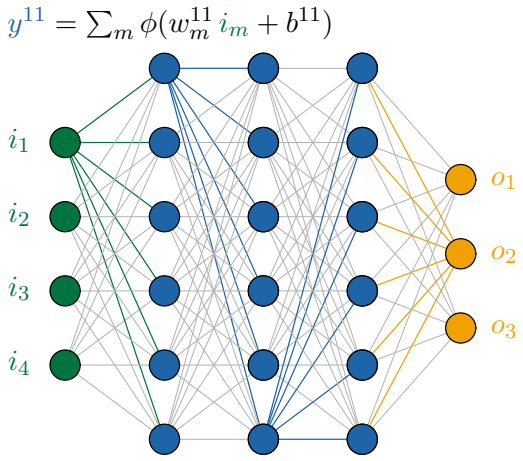


Figure 2.1: Illustration of the layout of a simple multilayer perceptron. The network consists of *input neurons* (green), *hidden neurons* (blue) and *output neurons* (orange). In this case, the hidden neurons are arranged in three hidden layers with six neurons each. Each neuron computes a function of the type of Equation 2.15. Shown is the corresponding equation for y^{11}, the output produced by the top neuron in the first hidden layer. Lines connecting two neurons represent the output of one neuron being used as the input of the next. The entire network in this example will then implement a function $\mathbb{R}^4 \to \mathbb{R}^3$.

exceeds a certain threshold. From an idealized mathematical point of view, each neuron therefore implements a function of the form

$$y = \sum_m \phi(w_m x_m + b), \tag{2.15}$$

where $y \in \mathbb{R}$ is the neuron's output and the $x_m \in \mathbb{R}$ form the inputs. The $w_m \in \mathbb{R}$ are the components of a *weight vector* associated with this neuron, and $b \in \mathbb{R}$ is a *bias* term. The nonlinear function $\phi : \mathbb{R} \to \mathbb{R}$ is called the *activation function* and is usually chosen to be of sigmoidal shape.

In an MLP, many of these elementary neurons are connected in a grid-like arrangement, see Figure 2.1 for an illustration. The weights and biases of each neuron form the adjustable parameters of the network. In effect, the entire arrangement implements a complicated composition of functions of the form of Equation 2.15. It can be shown that, under very general assumptions on the form of the activation function, MLPs act as universal approximators. That is, they can, for a proper choice of their parameters, indeed approximate any smooth function arbitrarily precisely[9] [28].

Apart from the choice of using MLPs to parameterize $s_{c,c'}$, there exists additional leeway that is worth exploiting. On the one hand, one might choose to implement $s_{c,c'}$ through a single, "inclusive" MLP, based on the entire training dataset $\mathcal{E}_c \cup \mathcal{E}_{c'}$ associated to these categories. Alternatively, one can regard $s_{c,c'}$ as a piecewise function, with each component being handled by a separate, more specialized MLP. One could thus better control performance in different corners of event space, but one would on the other hand suffer from smaller training datasets for each of the components.

This latter approach was chosen for $s_{\mathrm{VBF},ggH}$, which is implemented by three MLPs. Each component handles events with a different number of jets (≥ 2, 1 or 0 jets). In this specific case, this was observed to lead to an improved classification performance, since ggH forms the dominant background to the VBF process and jet kinematics provide a powerful discriminant between the two. In addition, the available training datasets for these two processes contain enough events to allow for this splitting. All other $s_{c,c'}$ are implemented by a single MLP each. With this setup, the 11 event categories defined in Table 2.1 require 55 functions $s_{c,c'}$, parameterized by a total of 57 MLPs.

2.2.3 Determining the Parameters

The degree to which the $s_{c,c'}$ fail to be monotonic functions of the $r_{c,c'}$ is directly related to their failure in minimizing the MSE in Equation 2.10. In theory, MLPs are powerful enough to find this global minimum, although care must be taken to achieve good performance in practice. The following sections describe potential difficulties in finding the correct set of MLP parameters and highlight the implemented countermeasures.

[9] Intuitively, this is a somewhat stronger form of the Kolmogorov superposition theorem which states that any continuous function $f : \mathbb{R}^n \to \mathbb{R}$ can be written as a composition of continuous functions h_i, $g_{i,j} : \mathbb{R} \to \mathbb{R}$ in the form $f(x_1, \ldots, x_n) = \sum_{i=0}^{2n} h_i \left(\sum_{j=0}^{n} g_{i,j}(x_j) \right)$. A-priori, the h_i and $g_{i,j}$ can all be different.

2.2.3.1 Training Datasets

For each category $c \in \mathcal{C}$, a sample of events drawn from the corresponding intended event composition $p_{\mathbf{E}}(\mathbf{e}|c)$ is available. These consist either of events from the control regions introduced in Section 1.3.4 (for $c = Z + X$) or of simulated events (for all other categories).

The MC samples for the ggH, VBF, VH and $t\bar{t}H$ signal processes are generated with POWHEG [29, 30] at next-to-leading order (NLO) in perturbative QCD. The tH and $b\bar{b}H$ processes are generated using JHUGEN [31], which also models the decay $H \to 4\ell$ for all samples. Showering is performed using PYTHIA 8 [32]. The irreducible $q\bar{q} \to ZZ$ background is also generated at NLO using POWHEG and PYTHIA, with the same settings as for the signal. The $gg \to ZZ$ background is simulated at leading order (LO) with MCFM [33]. All samples are generated with the NNPDF 3.1 NLO parton distribution functions and the signal samples use $m_H = 125\,\text{GeV}$.

These samples are then partitioned randomly into a training, a validation and a test dataset. The *training* dataset will be used to fit the MLP parameters in the sense of Theorem 2 and therefore construct $s_{c,c'}$. The *validation* dataset serves as a regularizer during the training and is also used to select proper values for the category priors $p(c)$. Also, any other free hyperparameters are determined by using this dataset. The *test* dataset is not needed in any way to define the categorization algorithm. It is solely used to perform an unbiased evaluation of the overall analysis performance at the end.

For all signal processes, the training dataset consists of 50% of all available simulated events, and the validation and test datasets hold 25% each. For the $q\bar{q} \to ZZ$ background, a smaller training sample of 33% of all events was used, with a validation dataset of 17% of all events. For $Z + X$, the training and validation datasets were reduced further to 13% and 7%, respectively. The remaining 80% of the events are then available for the estimation of the $Z + X$ yield in the signal region as outlined in Section 1.3.4.

All events contained in the training datasets that pass the analysis event selection are also used for the training. Note that this includes in particular events *outside* the $105 < m_{4\ell} < 140\,\text{GeV}$ signal region that is used for the measurement of the signal strength modifiers. This was found to be beneficial for the identification of ZH-leptonic events. They feature a distinctive *shoulder* in the $m_{4\ell}$ spectrum to the right of the Higgs boson peak. This is caused by associated leptons from the Z boson decay that are misidentified as coming from the decay of the Higgs boson. By using the *entire* shoulder, and not merely the part that lies within the signal region, the training dataset grows significantly in size. Building on these extra events to reinforce their characteristic features was found to improve the performance not only globally, but also in the relevant $m_{4\ell}$ signal region. A similar fact was observed for the $q\bar{q} \to ZZ$ background, for which the majority of events are expected to lie in the Z peak and the ZZ continuum, i.e. outside the signal region. However, no adverse effects have been observed by including also these events in the training datasets.

2.2.3.2 Constructing the Feature Vector

The functions $s_{c,c'}$ operate on the feature vector $\mathbf{e}$. From which observables $\mathbf{e}$ gets constructed is, however, not predetermined. On the one hand, the feature vector should contain those variables with the highest discriminative power to separate c from c'. On the other hand, the number of input features should be restricted to a reasonable minimum. In addition, the event classification and its performance should not be dependent on m_H, the free parameter of the analysis signal model. The feature vector must therefore be designed in such a way that m_H can not be reconstructed from it.

It is far from clear a-priori which combination of input variables will lead to the best discriminative performance. Moreover, this choice will depend on which categories are concerned, i.e. can be different for each $s_{c,c'}$. An exhaustive search in the space of input variables is infeasible due to the large number of possible combinations and the huge computational effort required. Therefore, the following staged approach was chosen to arrive at a reasonable guess for a well-performing combination of input features, individually for each $s_{c,c'}$.

First, a manual preselection of variables is performed, starting from low-level kinematic properties of the final state particles and the kinematic discriminants based on them. This step aims to remove features that are either not applicable or otherwise irrelevant for individual $s_{c,c'}$. For example, for the piecewise function $s_{\text{VBF},ggH}$, some restrictions must be obeyed concerning the production discriminants $\mathcal{D}_{p,p'}$. Most of the $\mathcal{D}_{p,p'}$ are defined only for events with at least two jets. Correspondingly, only $\mathcal{D}_{\text{VBF-1j},ggH}$ is available for its 1-jet component, while its 0-jet component can not access any production discriminant at all. A-priori, *all* kinematic discriminants are usable for the other $s_{c,c'}$. In addition, the primitive event variables contained in $\mathbf{\Omega}^{H\to 4\ell}$ are made accessible only to those $s_{c,c'}$ that try to separate between a signal category and a background component[10]. In this step, also the leading information on m_H is removed by masking the four-lepton invariant

[10] As a temporary fix to avert the current modelling issues for the MET, this variable is blocked for any $s_{Z+X,c'}$, i.e. any MLP

mass $m_{4\ell}$ in the region below 150 GeV, i.e. for the entire signal region. Furthermore, the kinematic properties (p_T, η and the azimuth ϕ) of only the two hardest jets in each event are forwarded[11].

Second, a boosted decision tree (BDT) is trained to discriminate between events from any two categories c and c'. This uses the exact same training dataset that is going to be used later for the fitting of the MLP, starting from the input features selected for each $s_{c,c'}$ in the first stage. The training of a BDT is relatively cheap, computationally, and allows this step to be completed in a reasonable time frame. In essence, the BDT is used here as a surrogate for the corresponding MLP, expecting it to capture correlations between variables that can then be fully exploited by the latter. Indeed, a trained BDT naturally provides an *importance score* for each feature, namely the number of times that a cut is placed on it by the decision tree ensemble. Based on this metric, any variable used by the BDT for more than a certain fraction of all imposed cuts, here 1%, is finally included in the feature vector $\mathbf{e}$. Starting from 42 available input variables, this selection process allocates between 10 to 31 observables to the feature vectors, depending on $s_{c,c'}$. Operatively, the training of the BDTs is done using *gradient boosting* [34] as implemented in the `XGBoost` package [35].

Note that, in particular, *no* manual post-selection of input features is made after this step. In line with the data-driven nature of Bayesian classification adopted here, variables are selected based only on their observed characteristics and correlations. No physics knowledge, other than the minimum amount required for the first selection step, is assumed.

Figure 2.2 summarizes the resulting collection of input variables obtained in this way, together with the importance weight placed on each feature. Horizontally, this plot lists the 55 functions $s_{c,c'}$. The three components of $s_{\text{VBF},ggH}$ are shown separately. Input variables that pass the selection for at least one MLP are shown vertically. From top to bottom, these include the kinematic production and background discriminants $\mathcal{D}_{p,p'}$ and $\mathcal{D}_{\text{bkg}}$ as well as kinematic properties (p_T, η, ϕ) of associated leptons $e\ell$ and jets j. A subscripted (i) refers to the i-th hardest object, ranked by p_T. Also listed are masses m, decay flavours $\mathfrak{f} \in \{2e, 2\mu\}$ and transverse momenta p_T of the Z candidates formed during the event selection. Properties of the selected ZZ candidate are used as well. Here, $m_{4\ell}$ refers to the *masked* four-lepton invariant mass and $\sigma(m_{4\ell})$ is its estimated resolution[12]. Also available are the numbers of jets $n(j)$, b-tagged jets $n(j_b)$, associated leptons $n(e\ell)$ and additional Z candidates $n(eZ)$. Finally shown are the angles contained in $\mathbf{\Omega}^{H\to 4\ell}$.

The interpretation of Figure 2.2 requires some care. On the one hand, variables may pass the selection not because of their intrinsic power, but because correlating them with *other* features is beneficial. On the other hand, the highest-ranked variables (i.e. those that are used most often by the trained decision tree ensemble) are expected to carry significant discriminative power on their own. As we will see, these are the variables that also a *physics-driven* selection would include. For example, information about associated leptons in the event is preferentially added to the feature vector as soon as the VH-leptonic or $t\bar{t}H$-leptonic categories are involved. Indeed, these contain additional leptons from the decays of the V bosons. The MET receives high weights for $s_{c,c'}$ that are concerned with events from VH-MET (by definition of its targeted signal process) as well as the WH-leptonic and $t\bar{t}H$-leptonic categories (due to the characteristic presence of neutrinos from $W \to \ell\nu$ decays).

To separate VBF from ggH events, each with one reconstructed jet, the most important variables are expected to concern the kinematics of the single jet. Indeed, this is reflected in the selection made by the BDT: $\eta(j_{(1)})$, $p_T(j_{(1)})$ and $\mathcal{D}_{\text{VBF-1}j,ggH}$ are the highest-ranked features for the 1-jet component of $s_{\text{VBF},ggH}$. In the case of two observed jets, emphasis is given instead to the kinematic discriminant $\mathcal{D}_{\text{VBF-2j},ggH}$. This variable seems to *shadow* the raw kinematic information of the jets themselves, which tendentially receive lower scores. Using the technique of *activation maximization* introduced in Section 2.2.3.4 and Appendix B, this effect was found to persist also on the level of the MLP that implements the 2-jet component of $s_{\text{VBF},ggH}$. Only if the kinematic discriminants are blocked as inputs does the low-level jet kinematics get exploited fully.

The distinction between physics- and data-driven feature selection becomes blurred for $\mathcal{D}_{\text{bkg}}$. As expected, this variable ranks high for those $s_{c,c'}$ that discriminate between signal and reducible or irreducible background. Interestingly, it also receives a high score for $s_{WH\text{-lept.},ZH\text{-lept.}}$, which only concerns signal events. This is due to ZH-leptonic events from the shoulder in $m_{4\ell}$. Having misidentified leptons, these events feature less signal-

that would be exposed to events coming both from data (albeit only from the control regions) and MC during the training process.

[11] This has to do with the fact that the used MC samples are produced at NLO. For ggH, for example, only the leading jet is modelled at the matrix element level, and fully correct. The next-to-leading jet is already showered off the first jet or the incoming gluons, but still follows reasonably well the distributions observed in data. For any additional jet in the event this agreement is already much worse. Furthermore, only jets with $p_T > 30$ GeV are used, again to ensure good correspondence between data and MC.

[12] Even though the resolution $\sigma(m_{4\ell})$ is generally correlated with $m_{4\ell}$, for events in the $m_{4\ell}$ signal region, the correlation is negligible. This variable is therefore not masked like $m_{4\ell}$ itself; moreover, it was confirmed explicitly that this does not introduce any undesired m_H-dependence of the classification performance.

like 4ℓ kinematics and consequently are assigned lower $\mathcal{D}_{\text{bkg}}$ values. Inadvertently, a *background* discriminant thereby becomes sensitive to differences between various *signal* processes.

Beside $\mathcal{D}_{\text{bkg}}$, the masses of the Z_1 and Z_2 candidates are considered important to isolate the $q\bar{q} \to ZZ$ background. This is a result of the fact that no $m_{4\ell}$ cut was placed on the training datasets. Then, the masses $m(Z_{1,2})$ allow to separate the events in the ZZ continuum from the Higgs boson signal where Z_2 is off-shell. If the training dataset is restricted to the $m_{4\ell}$ signal region, both masses behave similar for signal and background and also their assigned importance scores diminish.

Several collections of input features have been investigated, differing in the cumulative number of selected variables, summed over all $s_{c,c'}$. With the BDTs fully trained, these variations are easily obtained by changing the acceptance threshold on the importance score. The selection presented here ranks high in terms of the number of assigned features and the achieved performance. However, tests indicate that the total number of variables (as measured by the cumulative sum) can be reduced by about 25% without significantly impairing the overall performance of the classification. If *no* selection of input features is made and all available input features are used for every $s_{c,c'}$, classification performance decreases by about 2-4%.

2.2.3.3 Preprocessing

The feature vector as composed by the above method will generally consist of variables with very different characteristics, e.g. wildly contrasting numerical ranges and distributions. This can have adverse effects when such data are directly used as inputs to an MLP. For example, the activation functions of individual neurons may easily get *saturated* and cause problems during the training procedure[13]. Further preprocessing steps are therefore needed.

Most fundamentally, a variable may be *periodic* (i.e. the two boundary points of its allowed range are identified, as is the case for angles) or *nonperiodic*. However, correlating multiple such variables is cumbersome if the chosen representation does not explicitly respect the periodicity. To this end, any angle ϕ that is selected as input feature gets *encoded* in the feature vector as the tuple $(\sin\phi, \cos\phi)$, thereby manifestly preserving the identification $0 \sim 2\pi$.

Some variables may not be defined for individual events, such as the pseudorapidity of the second-hardest jet in an event containing only one jet. By default, such a non-existing observable is set to a predefined value during the preprocessing step. Note that this is certainly not an optimal solution, as it may not always be possible to choose a default value that lies outside the physically allowed regime. Thus, an alternative solution has been investigated, using a different neural network architecture[14]. However, no performance gain was visible at the expense of a significantly increased model complexity and prolonged training times.

At this point, the feature vector may also contain observables that are strongly correlated with one another. In fact, these variables may pass the BDT-based selection precisely *because* of their correlation: a cut placed on either one will have identical effects. It is therefore desirable to remove this correlation, as follows. The feature vectors of the events in the training dataset $\mathcal{E}_c \cup \mathcal{E}_{c'}$ for $s_{c,c'}$ are realizations of a random vector $\mathbf{E}$. These realizations are sampled from the distribution $\mu_c\, p_{\mathbf{E}}(\mathbf{e}|c) + \mu_{c'}\, p_{\mathbf{E}}(\mathbf{e}|c')$. Through a transformation $\mathbf{e} \to R\mathbf{e}$, where R is an orthogonal matrix, the covariance matrix of $R\mathbf{E}$ can then be made diagonal. All correlations between the different components of the transformed feature vector thus vanish. In addition, each transformed component can be rescaled to have a mean of zero and a variance of one. This procedure, known as *PCA whitening*, will then generate the final inputs that are presented to the MLPs.

2.2.3.4 Training and Regularization

During the training procedure, the parameters of the MLP are iteratively adjusted in order to minimize the MSE in Equation 2.10 and to implement the desired $s_{c,c'}$. The weights w_c and $w_{c'}$ that were left unspecified in Theorem 2 are now also fixed to $w_c = \frac{1}{\mu_c} = \frac{N}{N_c}$ and $w_{c'} = \frac{1}{\mu_{c'}} = \frac{N}{N_{c'}}$. As the proof of this theorem shows, the specific choice of these weights should not be of any relevance to the monotonicity of the resulting $s_{c,c'}$. In practice, however, some training datasets turn out to be very *unbalanced*, i.e. $|\mathcal{E}_c|$ is very different from $|\mathcal{E}_{c'}|$. In such a case, a nonzero, but trivial, categorization performance can be achieved by simply assigning all events to the majority class, i.e. by a constant $s_{c,c'}(\mathbf{e})$. For our application, such a result is clearly undesired. Intuitively, the choice of the weights made above then puts a stronger emphasis on the events from the minority class and penalizes this behavior.

[13] In the training phase, the parameters of the network are optimized through a gradient descent to minimize the loss functional. A saturating activation function would cause the gradient of the loss function to vanish for individual weights. Those parameters would then no longer be updated.

[14] Instead of using an MLP alone to compute $s_{c,c'}$, it was combined with a recurrent neural network (in practice, the more robust LSTMs [36] have been used). Recurrent neural networks naturally allow for the processing of *sequences*. This class of models can therefore accept kinematic input of variable length, e.g. lists of jet or lepton kinematics.

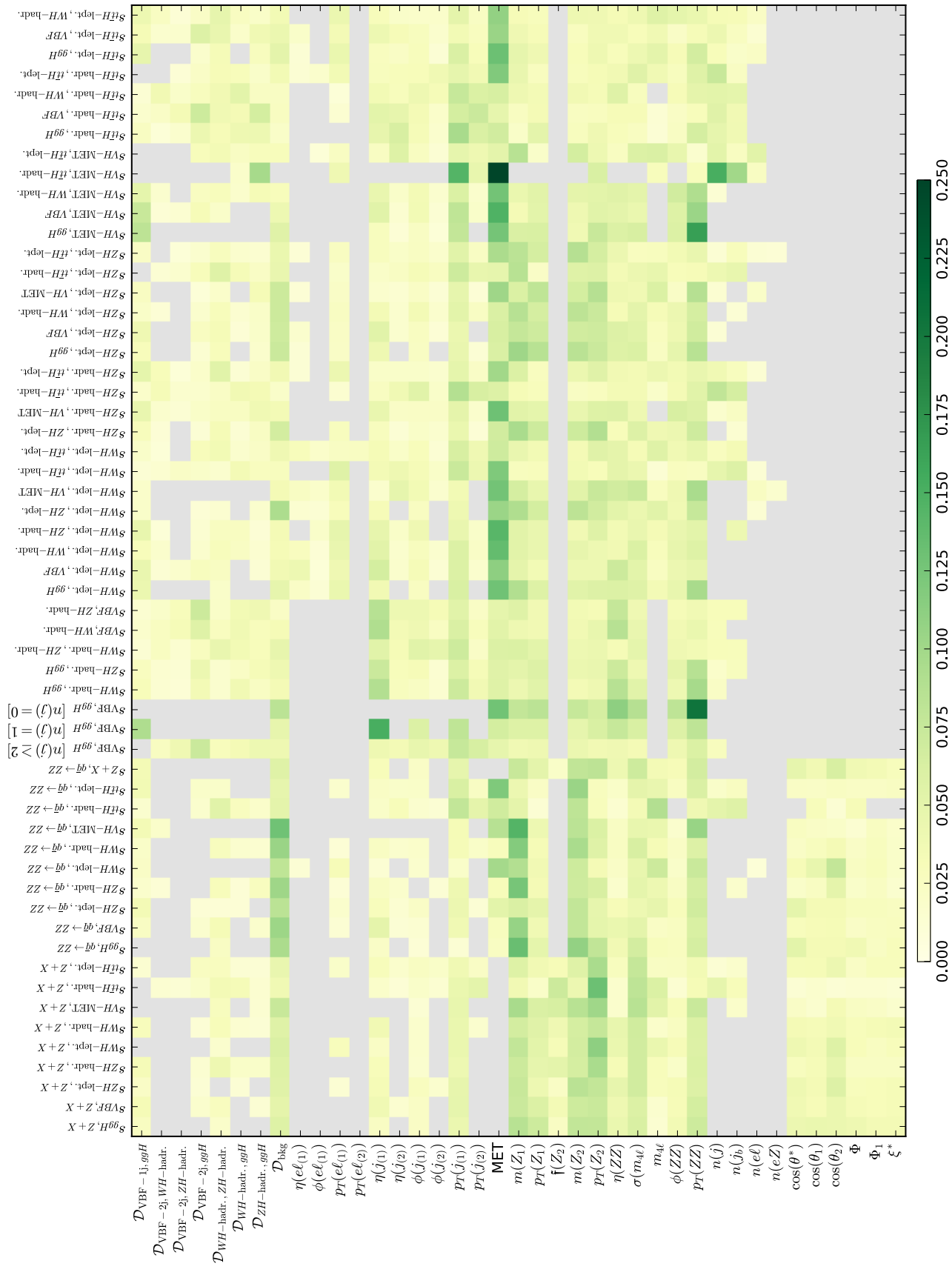


Figure 2.2: Observables selected as input for a particular function $s_{c,c'}$, listed horizontally. The available input features are shown vertically. The coloring indicates the importance score attached to a variable by the BDT, normalized to one for each $s_{c,c'}$. Variables that are not selected are shown in gray.

Operatively, the training of the MLPs is performed through a gradient descent in the space of network parameters. Two different variants of the classical gradient descent algorithm have been evaluated: *stochastic gradient descent* (SGD) [37] and the adaptive *Adam* algorithm [38]. The final performance achieved by these two methods was found to be very similar, although SGD required more training iterations to reach the same level of performance. All results presented in the following thus use Adam. The training of the MLPs is carried out using the `Keras` package [39] together with `TensorFlow` [40] as backend.

The training procedure, owing to its iterative nature, is computationally expensive. On the other hand, the evaluation of an already trained network consists in a series of matrix multiplications and can be handled very efficiently. The computation of the corresponding $s_{c,c'}$ for a given event is therefore very economical in terms of the needed resources.

The low-level details of this calculation are represented by the states of the internal neurons. Unfortunately, these do not relate readily to more classical concepts, such as the ideas that go into the computation of matrix element discriminants. Graphical visualizations of a trained MLP can help to build up intuition about how individual input variables get used and correlated. Somewhat outside the main line of development of this thesis, Appendix B highlights such a visualization method and gives a few examples. It also explains why these can be interesting topics to study in their own right.

With all its parameters fixed after the training, we expect the resulting function $s_{c,c'}$ to behave identical, regardless of whether it is applied to events coming from the training dataset itself, or from any other independent set of samples. This is, however, not true if the MLP is *overfitting* the training dataset, i.e. incorporating statistical fluctuations instead of actually present characteristics of the data. Several techniques are employed during the training process to limit this undesired phenomenon. First, *dropout regularization* is used, whereby a fixed fraction of the network's neurons are chosen randomly and temporarily excluded from the gradient descent. Second, *early stopping regularization* halts the training process as soon as the network's performance, as measured by the validation loss (i.e. the MSE computed for the validation dataset) ceases to improve. These two methods are indeed sufficient to avoid overtraining, see Figure 2.4.

2.2.3.5 Hyperparameters

The internal structure of an MLP in terms of its neurons is already partially fixed by the properties of the scalar function $s_{c,c'}$ that it implements. Indeed, by construction, each network will have one output neuron only, while the number of input neurons is fixed by the dimensionality of the chosen feature vector. However, the number of hidden layers in the network and the number of neurons that make up each hidden layer are undetermined. They are part of the adjustable *hyperparameters* of the network.

In principle, one could perform an optimization of the hyperparameters in the same sense as the neuron weights were updated during the training. One would pick a set of hyperparameters, train the corresponding MLP and take the validation loss of the *trained* network as a measure of its performance. Since this cost function is enormously expensive to evaluate (with training times typically of several hours), Bayesian optimization methods are used to limit the number of cost function calls, i.e. the number of evaluated hyperparameter combinations (see also Section 2.2.5 for a different application of this method).

Since the MLPs used for the present application will contain comparably few neurons, a simplified, but significantly faster approach was chosen instead. Rather than allowing each hidden layer to have a different number of neurons, the same value was used for all hidden layers. Together with the number of hidden layers, this leaves two hyperparameters that are most relevant to the performance of the MLP[15]. Then, networks are trained in parallel for several different values of these two parameters. Finally, the combination leading to the best performance on the validation dataset is retained.

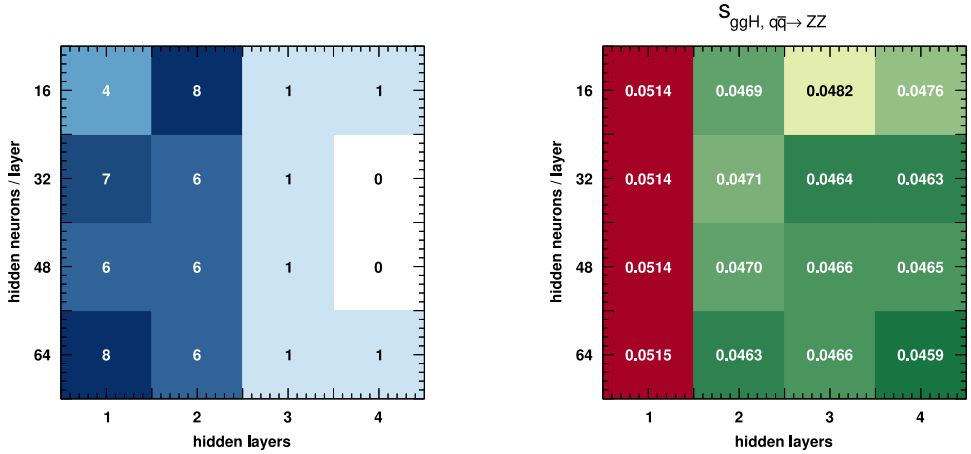
Figure 2.3 shows that, in most cases, the best-performing networks are very small, with only one or two hidden layers and around fifty neurons per hidden layer. As a rule of thumb, the number of weights that parameterize the network should be of the same order as the size of the training dataset. This is indeed the case here, with both numbers being $\mathcal{O}(10^4)$.

2.2.4 Calibration

For the computation of the likelihood ratios $r_{c,c'}$, the one-dimensional densities $p_{\mathrm{U}}(u = s_{c,c'}(\mathbf{e})|c)$ and $p_{\mathrm{U}}(u = s_{c,c'}(\mathbf{e})|c')$ must be estimated. This step is also called *calibration*, for it is the ratio of these probabilities that allows one to compute the numerically correct value of $r_{c,c'}$, given an *uncalibrated* $s_{c,c'}$.

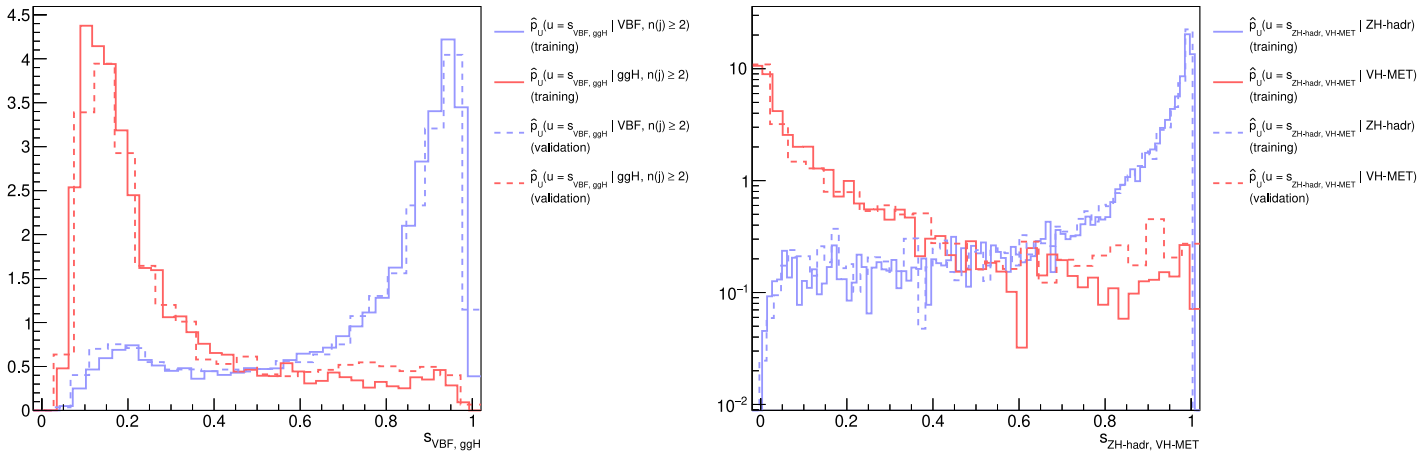
Histograms form very convenient nonparametric density estimators. Their only adjustable hyperparameter is the number of bins, or, equivalently, the bin width. This parameter should be chosen in such a way that

[15] In principle, also the activation functions ϕ are part of the hyperparameters. In our implementation, fixed choices have been made, using *rectified linear* activation functions [41] for all hidden neurons, and a sigmoid for the output neuron.



(a) (b)

Figure 2.3: Result of a sweep of network hyperparameters. Figure (a) shows a histogram of the optimized hyperparameters of the 57 MLPs needed for the classification algorithm. Only very few networks require more than two hidden layers, and if they do, the improvement w.r.t. a model with fewer layers is often below 1%. This is illustrated in Figure (b). It shows the the achieved MSE on the validation dataset as a function of the hyperparameters of the MLP that parameterizes $s_{ggH,q\bar{q}\to ZZ}$. It performs optimally with four hidden layers.



(a) (b)

Figure 2.4: Estimated probability distributions $\hat{p}_U$ of $s_{\mathrm{VBF},ggH}(\mathbf{e})$ for events with at least two jets in (a) and for $s_{ZH\text{-hadr.},VH\text{-MET}}(\mathbf{e})$ in (b). No significant overtraining is present, as the distributions are very similar regardless of whether the estimation is performed on the training or on the validation dataset. Note that, since these two event samples are of different sizes, the bin widths differ as well, as dictated by the method of Diaconis and Freedman.

the *estimated* distribution $\hat{p}_U(u)$ follows as closely as possible the *true*, but ultimately unknown, distribution $p_U(u)$. If one demands to minimize the *mean squared error* of the density estimation, i.e. one asks that

$$\int du\ (p_U(u) - \hat{p}_U(u))^2 \to \text{min.},$$

then, as Freedman and Diaconis show [42], the ideal bin width h is given by

$$h = 2\frac{\mathrm{IQR}}{n^{1/3}}. \tag{2.16}$$

Here, IQR is the interquartile range of the data and n is the number of samples used to fill the histogram. Figure 2.4 illustrates two practical examples of distributions p_U that were estimated in this way.

2.2.5 Finding the Priors

In Bayes' theorem, priors have the important role of converting a likelihood into a quantity proportional to the posterior probability. Only once the priors are specified becomes the concept of a posterior even meaningfully *defined*.

In the context of Bayesian classification, the priors $p(c)$ allow to control how *acceptive* any given category should be: if it is assigned a large prior, also its posterior probability will be large. An event is therefore likely to be assigned to this class unless there is strong evidence against it, i.e. the likelihood is very small. According to Equation 2.5, not the priors themselves, but only their *ratios* are important. That is, instead of having $|\mathcal{C}| = 11$ priors in our case, one for each category, there are only $|\mathcal{C}| - 1 = 10$ relevant free parameters in total. In practice, this extra degree of freedom can be eliminated simply by choosing an arbitrary category prior and setting it to a fixed default value.

We are going to take the pragmatic attitude that the "correct" priors $p(c)$ are those that lead to a high categorization performance. Indeed, by tuning the priors, one can optimize signal efficiency and background rejection for any given category. Indeed, both a good signal *purity* and a sufficient signal *acceptance* will be required in the subsequent statistical analysis to extract the Higgs boson signal strength modifiers. Up to now, we have not specified in which way the compromise between the two should be resolved.

2.2.5.1 Punzi's Purity Measure

Clearly, the event categorization should be tuned in such a way as to maximize the discovery potential of the $H \to 4\ell$ analysis. Indeed, only the dominant ggH production channel has so far been "discovered" in the sense of a 5σ incompatibility with the hypothesis $\mu_{ggH} = 0$. All others have not yet been seen individually in the four-lepton final state and thus it is for those subleading production processes that the sensitivity of the analysis should be optimized.

As was laid out in Section 1.3.5, the signal strength modifiers μ_p of the various Higgs boson production modes are determined in a fit of the analyis model to the observed events, involving all eight event categories c and all final states f. Therefore, in general, *all categories* will contribute to the extraction of any given μ_p. However, those that specifically target this process will have the highest purity and event yield, and therefore also the greatest impact on the fit.

The relationship between the composition of the individual categories and the sensitivity of the analysis is thus somewhat obscured by the details of the fitting procedure. Thus, it is not immediately clear how the former should look like in order to optimize the latter. However, we can ponder performing a simplified analysis in which the μ_p are not extracted directly, but each event category c is used *individually* to constrain the strength of its targeted *signal process* according to Table 2.1. This gives rise to signal strength modifiers $\mu_s = \mu_{s(c)}$ that are, generally, more exclusive than the μ_p. In line with the above, only the *tagged* categories and their related μ_s are considered at this stage. For each category, this corresponds to a simple Poisson counting experiment in the presence of background. We now choose the priors such that the sensitivity to each μ_s in this simplified setting is optimized. Since each of the production mode signal strengths μ_p correlates the information contained in one or several μ_s, we can then also expect the performance of the original, more complicated fit to be reasonably well tuned.

Against this backdrop, it remains to specify a performance measure for each category c individually, sensitive to the discovery potential for its targeted signal process. The concept of purity introduced by Punzi in [43] provides a very convenient metric. It is related to the size of the region $\mathcal{S}_c$ in parameter space for which the counting experiment in category c is *conclusive* in the sense explained below. In our case, the parameter space is the positive real axis and thus one-dimensional, as every category is used to assess the strength of a single signal process only. To apply Punzi's sensitivity measure, we also need to specify a statistical test that is used to draw conclusions from an excess seen in the data, i.e. determines whether the null hypothesis $\mu_s = 0$ must be given up. In our simplified setting, the only test statistic that can be used for this purpose is the number of events in category c, n_c. The critical region of a test sensitive to a signal contribution on top of the background will then be of the form $\Omega_c = \{n_c : n_c > n_{c,\min}\}$.

The *sensitivity region* of category c, $\mathcal{S}_c$, is then defined as the set

$$\mathcal{S}_c = \{\mu_s : 1 - \beta_\alpha(\mu_s) \geq \mathrm{CL}\}, \tag{2.17}$$

where α is the chosen *significance level* of the test and $1-\beta$ denotes its *power*, CL is a user-defined probability explained below. For the one-sided test prescribed above, the sensitivity region will always be of the form $\mathcal{S}_c = \{\mu_s : \mu_s > \mu_{s(c),\min}\}$.

Now, $\mathcal{S}_c$ is precisely the region in parameter space in which μ_s will *either* be discovered with a certain probability *or* can be excluded at a certain confidence level:

- In case the experiment observes a number of events $n_{c,\mathrm{obs}}$ that lies in the critical region of the test, $n_{c,\mathrm{obs}} \in \Omega_c$, then the hypothesis $\mu_s = 0$ must be rejected at the chosen significance level α. If the true (but unknown) value of the signal strength μ_s lies within $\mathcal{S}_c$, by definition, this happens with a probability of *at least* CL.

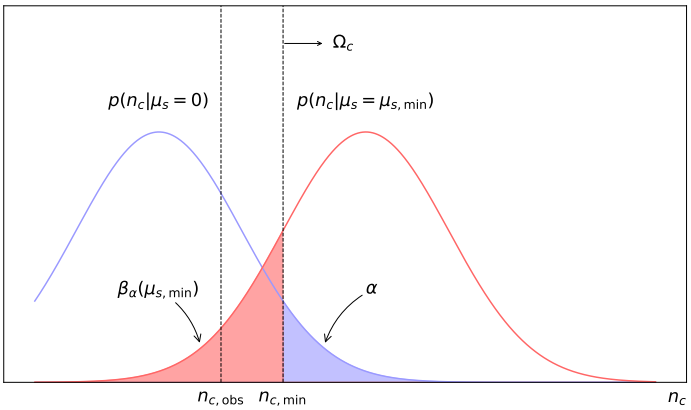


Figure 2.5: Illustration of Punzi's purity measure for a simple counting experiment in category c. Shown are the distributions of the test statistic n_c for the null hypothesis $\mu_s = 0$ (blue) and for the hypothesis $\mu_s = \mu_{s,\min}$ (red). This saturates the inequality in Equation 2.17, such that the red-colored area amounts to exactly $1-\text{CL}$. For any $n_{c,\text{obs}}$ outside the critical region, i.e. $n_{c,\text{obs}} < n_{c,\min}$, the standard Neyman construction will therefore place $\mu_{s,\min}$ *outside* the one-sided confidence interval for μ_s at confidence level CL. The same is true for all $\mu_s > \mu_{s,\min}$, so all $\mu_s \in \mathcal{S}_c$ get excluded.

- In case the experiment does *not* lead to the discovery of $\mu_s \neq 0$, i.e. $n_{c,\text{obs}}$ lies outside the critical region, then the confidence interval for μ_s at confidence level CL will exclude (at least) *all* values of $\mu_s \in \mathcal{S}_c$. Figure 2.5 tries to make this fact plausible.

Clearly, if one choice of priors leads to a larger sensitivity region than another, the former can well be called more *sensitive* to the signal process μ_s. Thus, we can define the Punzi purity $\mathcal{P}_c$ of category c as a measure of the size of $\mathcal{S}_c$ as

$$\mathcal{P}_c = \frac{1}{\mu_{s(c),\min}}.$$

To derive an explicit formula for $\mathcal{P}_c$, consider again Figure 2.5. For our counting experiment, the n_c will be Poisson distributed, $n_c \sim \text{Po}(n_{s,c}(\mu_s) + n_{b,c})$. Here, $n_{s,c}(\mu_s)$ and $n_{b,c}$ are the expected number of signal and background events in category c, respectively (we do not distinguish between different final states here, hence the absence of the index f). Owing to the discrete nature of the Poisson distribution, discretization effects will also be visible in the resulting expression for $\mathcal{P}_c$. For the purpose of optimizing the performance of the categorization, however, an analytic, continuous parameterization of the exact result will be sufficient. A simple way to obtain one is to use a Gaussian as an approximation for the density of the Poisson distribution, $n_c \sim \mathcal{N}(\mu, \sigma^2) \sim \mathcal{N}(n_{s,c}(\mu_s) + n_{b,c}, n_{s,c}(\mu_s) + n_{b,c})$. For $\mu_s = \mu_{s,\min}$, the inequality in Equation 2.17 saturates and the mean values of the Gaussians for $\mu_s = 0$ and $\mu_s = \mu_{s,\min}$ become related in the following way,

$$n_{s,c}(0) + a\sqrt{n_{s,c}(0) + n_{b,c}} + b\sqrt{n_{s,c}(\mu_{s,\min}) + n_{b,c}} = n_{s,c}(\mu_{s,\min}). \tag{2.18}$$

Here, $a = \Phi^{-1}(1-\alpha)$ and $b = \Phi^{-1}(\text{CL})$, where Φ is the cumulative distribution function (CDF) of the standard normal distribution. The expected number of signal events $n_{s,c}$ can be related to the signal efficiency ϵ_c achieved by category c, the integrated luminosity L_{int} and the SM cross section for the targeted signal process σ_s^{SM},

$$n_{s,c}(\mu_s) = \epsilon_c \cdot L_{\text{int}} \cdot \sigma_s^{\text{SM}} \cdot \mu_s.$$

Making use of this relation to solve for $\mu_{\min}$ in Equation 2.18, one obtains

$$\mathcal{P}_c = \frac{1}{\mu_{s(c),\min}} \sim \frac{\epsilon_c}{b^2 + 2a\sqrt{n_{b,c}} + b\sqrt{b^2 + 4a\sqrt{n_{b,c}} + 4n_{b,c}}}, \tag{2.19}$$

where we have dropped irrelevant overall constant factors involving σ_s^{SM} and L_{int}.

The Gaussian approximation becomes problematic for low values of $n_{b,c}$ and large values of a, b, i.e. in situations where the difference between Gaussian and Poissonian tail integrals becomes noticeable. One can improve the approximation made above by taking into account these differences at the next order in a and b. This leads to a modification of the denominator in Equation 2.19. The explicit expression is given in [43].

A word is in order about the exact meaning of *background* in this context. Clearly, from the above considerations of a simple counting experiment, the term should include both reducible and irreducible backgrounds,

as well as misclassified Higgs *signal* events. However, the full analysis determines the signal strengths μ_p using two-dimensional templates that involve $\mathcal{D}^c_{\text{bkg}}$. That is to say that events from the reducible and irreducible backgrounds get separated from the $H \to 4\ell$ signal along the direction of $\mathcal{D}^c_{\text{bkg}}$ during the fitting procedure. Conversely, the presence of a certain number of these background events is much less severe than the presence of misclassified signal events. Therefore, the numbers $n_{b,c}$ in Equation 2.19 will *only* include this latter type of background. In this sense, we perform the counting experiment only for events with signal-like $\mathcal{D}^c_{\text{bkg}}$ values.

2.2.5.2 Bayesian Optimization

The priors $p(c)$ now need to be adjusted in such a way as to maximize the Punzi purity $\mathcal{P}_c$ for every category. To do so, the signal efficiency ϵ_c as well as the number of expected background events $n_{b,c}$ in each category must be computed. To determine these values, it is necessary to apply the categorization procedure with the proposed set of priors to (some fraction of) the validation dataset. Therefore, this is a very expensive problem, and the number of prior combinations that are evaluated during the optimization must be reduced to a minimum. In addition, the purities of the individual categories are not independent, i.e. the Punzi values of all categories will generally be affected by a change in the priors. Furthermore, any performance gain compared to the current event categorization should ideally be visible across *all* categories, rather than just a single one.

To incorporate the above requirements, we first need to define a scalar-valued utility function on the space of priors. Higher values of the utility function then correspond to preferable categorization outcomes: equally enhanced Punzi purities for all categories. Moreover, the utility function should not depend on the absolute scale of the $\mathcal{P}_c$, but only on their relative variations. To this end, we also define a reference Punzi value for each category, $\mathcal{P}_c^{\text{ref}}$. In practice, these will be the Punzi purities achieved by the currently employed categorization. The chosen utility function will then quantify the categorization performance based on the *relative* purity improvement in category c,

$$\Delta\mathcal{P}_c = \frac{\mathcal{P}_c - \mathcal{P}_c^{\text{ref}}}{\mathcal{P}_c^{\text{ref}}}. \tag{2.20}$$

A family of utility functions that was found to work well in practice is

$$u(\boldsymbol{\Delta}\mathcal{P}) = \frac{1}{\sqrt{|\mathcal{C}_A|}} \sum_{c\in\mathcal{C}_A} \left[\left(\frac{1}{\gamma}\right)^m - \left(W(\Delta\mathcal{P}_c) - \frac{1}{\gamma}\right)^m\right], \tag{2.21}$$

where $\gamma \in \mathbb{R}$ and $m \in 2\mathbb{Z}$ are free parameters, and $\boldsymbol{\Delta}\mathcal{P}$ is the vector of the $\Delta\mathcal{P}_c$. The *weighted improvement* $W(\Delta\mathcal{P}_c)$ is

$$W(\Delta\mathcal{P}_c) = \begin{cases} \frac{1}{\gamma}\tanh(\gamma\,\Delta\mathcal{P}_c) & \Delta\mathcal{P}_c > 0 \\ \Delta\mathcal{P}_c & \Delta\mathcal{P}_c \leq 0 \end{cases}. \tag{2.22}$$

This utility function has a number of nice properties. The choice of a saturating $W(\Delta\mathcal{P}_c)$ ensures that extremely large performance gains in one category are discounted if at the same time other categories perform poorly. The parameter γ is related to the onset of saturation at about $\Delta\mathcal{P}_c \sim \frac{1}{\gamma}$. In addition, we have that

$$\frac{du}{d\Delta\mathcal{P}_c} - \frac{du}{d\Delta\mathcal{P}_{c'}} < 0,$$

if $\Delta\mathcal{P}_c > \Delta\mathcal{P}_{c'}$ and both are positive. This has a balancing effect that tries to achieve uniform improvement in all categories. Moreover, $u = 0$ for a categorization performance that exactly matches that of the reference.

The optimal choice of priors is now the one that maximizes the utility function in Equation 2.21. As already alluded to above, this function internally iterates over a large dataset and is therefore very expensive to compute, with evaluation times typically of the order of several minutes. In addition, random numbers play a role in the process of finding the optimal category for an event, as will be explained in Section 2.2.6. This means that the computed Punzi purities will fluctuate slightly from evaluation to evaluation, i.e. the utility function is *noisy*. In addition, there exists a compromise between the evaluation time and the amount of noise that is present: using a larger part of the validation dataset is costly, but increases the statistics and hence the precision of the computed utility value.

Bayesian optimization methods [44] provide a handle to the maximization of such a noisy, expensive utility function. This framework models the utility u as a field of random variables, that is, a *stochastic process*. Here, the argument vector $\boldsymbol{\Delta}\mathcal{P}$ acts as a continuous index for the individual random variables $u(\boldsymbol{\Delta}\mathcal{P})$. Gaussian

processes are most commonly used[16]. A characteristic feature of a Gaussian process is that any finite set of i random variables $\{u(\mathbf{\Delta}\mathcal{P}_{1:i})\}$ follows a multivariate Gaussian distribution. A Gaussian process can therefore also be seen as a probability distribution on the space of functions $u(\mathbf{\Delta}\mathcal{P})$, which, in analogy to the finite dimensional case, is characterized by a *mean function* $\mu(\mathbf{\Delta}\mathcal{P})$ and a *covariance function* $\sigma^2(\mathbf{\Delta}\mathcal{P}, \mathbf{\Delta}\mathcal{P}')$.

Stochastic processes provide a rigorous way to use the entire information available about the utility function. In this context, the Gaussian process above plays the role of a prior, i.e. it encodes properties that the utility function is known to have a-priori. In addition, a certain number of "observations", i.e. evaluations of the utility function at certain points, are available. These observations then induce a *posterior* distribution on the space of functions. This updated model of the utility function is used to determine the position of the *next* observation in an optimal way, made more precise further below. In this sense, Bayesian optimization trades the original, very expensive maximization problem defined on the utility function for a more manageable optimization problem operating on the posterior.

More concretely, let $\mathcal{D}_{1:i} = \{u(\mathbf{\Delta}\mathcal{P}_{1:i})\} = \{u_{1:i}\}$ denote i observations of the utility function, e.g. taken from previous iterations of the algorithm. In our application, the Gaussian process prior will be characterized by a mean function that is identically zero, i.e. always identifies categorization performance to be equal to the reference, independent of the category priors $p(c)$. Then, the already observed utility values $u_{1:i}$ and the new, yet to be determined value u_{i+1} are jointly Gaussian with a mean of zero,

$$\begin{bmatrix} u_{1:i} \\ u_{i+1} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_{kl}^2 & \sigma_k^2 \\ \sigma_l^2 & \sigma^2(\mathbf{\Delta}\mathcal{P}_{i+1}, \mathbf{\Delta}\mathcal{P}_{i+1}), \end{bmatrix}\right) \tag{2.23}$$

where $\sigma_{kl}^2 = \sigma^2(\mathbf{\Delta}\mathcal{P}_k, \mathbf{\Delta}\mathcal{P}_l)$ and $\sigma_k^2 = \sigma^2(\mathbf{\Delta}\mathcal{P}_{i+1}, \mathbf{\Delta}\mathcal{P}_k)$. Using the Sherman-Morrison formula to compute the inverse of the covariance matrix in Equation 2.23, this density can be refactored into the posterior

$$p(u_{i+1}|\mathcal{D}_{1:i}, \mathbf{\Delta}\mathcal{P}_{i+1}) \sim \mathcal{N}(\mu_{i+1}(\mathbf{\Delta}\mathcal{P}_{i+1}), \sigma_{i+1}^2(\mathbf{\Delta}\mathcal{P}_{i+1})), \tag{2.24}$$

with

$$\mu_{i+1}(\mathbf{\Delta}\mathcal{P}_{i+1}) = \sigma_k^2\, \sigma_{kl}^{-2}\, u_l, \tag{2.25}$$

$$\sigma_{i+1}^2(\mathbf{\Delta}\mathcal{P}_{i+1}) = \sigma^2(\mathbf{\Delta}\mathcal{P}_{i+1}, \mathbf{\Delta}\mathcal{P}_{i+1}) - \sigma_k^2\, \sigma_{kl}^{-2}\, \sigma_l^2. \tag{2.26}$$

Here, σ_{kl}^{-2} is the inverse of the matrix σ_{kl}^2 in Equation 2.23. The posterior is sometimes also called the *acquisition function*, for reasons that will be clear shortly.

Given this posterior, the goal is now to determine the position $\mathbf{\Delta}\mathcal{P}_{i+1}$ where the next observation should be made. Naturally, this choice should build on the knowledge encoded in the posterior and preferentially pick points where the original utility function u is likely to have a maximum. A simple method is to choose the point where the *upper confidence bound* of the posterior has its maximum,

$$\mathbf{\Delta}\mathcal{P}_{i+1} = \arg\max_{\mathbf{\Delta}\mathcal{P}} \left[\mu_{i+1}(\mathbf{\Delta}\mathcal{P}) + \kappa\, \sigma_{i+1}(\mathbf{\Delta}\mathcal{P})\right]. \tag{2.27}$$

In this expression, κ is an adjustable parameter that controls the tradeoff between *exploitation* of known candidates for global maxima, and the *exploration*, i.e. the search for new such candidates.

Another option, which was found to work better in practice, is to select $\mathbf{\Delta}\mathcal{P}_{i+1}$ such that the *probability of improvement* of the utility function gets maximized at this point. That is, one sets

$$\mathbf{\Delta}\mathcal{P}_{i+1} = \arg\max_{\mathbf{\Delta}\mathcal{P}} p(u(\mathbf{\Delta}\mathcal{P}) > u_{\max} + \xi) = \arg\max_{\mathbf{\Delta}\mathcal{P}} \Phi\left(\frac{\mu_{i+1}(\mathbf{\Delta}\mathcal{P}) - u_{\max} - \xi}{\sigma_{i+1}(\mathbf{\Delta}\mathcal{P})}\right), \tag{2.28}$$

where $u_{\max}$ is the highest utility value observed so far, $u_{\max} = \max_{\mathbf{\Delta}\mathcal{P}\in\mathbf{\Delta}\mathcal{P}_{1:i}} u(\mathbf{\Delta}\mathcal{P})$. As before, Φ is the CDF of the standard normal distribution. High values of the parameter ξ encourage exploration, while low values lead to exploitation instead. Implementing a *cooling* schedule for ξ has been found to be very beneficial: high values at the start of the optimization lead to a good exploration of the parameter space, while low values at the end of the process will fine-tune the best candidate for the global maximum.

Nothing has been said so far about the choice of the covariance function $\sigma^2(\mathbf{\Delta}\mathcal{P}, \mathbf{\Delta}\mathcal{P}')$. Indeed, this *kernel* has a large impact on the performance of the algorithm, since it controls very directly the characteristics of the posterior and therefore the positions of the requested new utility observations. The standard choice

$$\sigma^2(\mathbf{\Delta}\mathcal{P}, \mathbf{\Delta}\mathcal{P}') = \exp\left(-\frac{1}{2}\|\mathbf{\Delta}\mathcal{P} - \mathbf{\Delta}\mathcal{P}'\|^2\right)$$

[16]This is mainly due to the convenient marginalization properties of the Gaussian distribution. Recently, the Student-t process was put forward as a promising alternative [45].

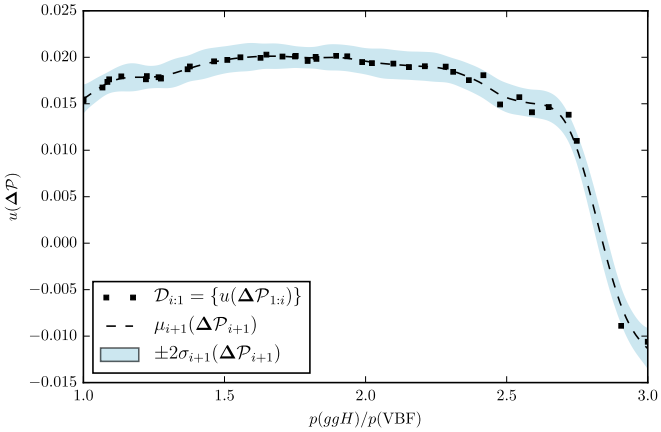


Figure 2.6: Optimization of the utility function $u(\boldsymbol{\Delta}\mathcal{P})$ restricted to the two VBF categories, allowing to fix $p(\mathrm{VBF})$ in relation to $p(ggH)$. Shown are observations $\mathcal{D}_{1:i} = \{u(\boldsymbol{\Delta}\mathcal{P}_{1:i})\}$ of the utility function as well as the posterior mean and standard deviation, according to Equations 2.25 and 2.26.

has been found to be too smooth for our purposes. Good results have been obtained when using the Matérn kernel instead [44]. Note that an additional diagonal contribution $\sigma^2(\boldsymbol{\Delta}\mathcal{P}, \boldsymbol{\Delta}\mathcal{P}') \sim \delta_{\boldsymbol{\Delta}\mathcal{P}\boldsymbol{\Delta}\mathcal{P}'}$ naturally allows to take into account the noise level present in the individual observations.

However, even equipped with this powerful tool of Bayesian optimization, a *direct* maximization of the full utility function in all priors simultaneously is infeasible. While the above statement is true that the variation of any prior will generally affect *all* categories, in practice, the individual categories (and priors) *do* decouple partially. For example, as can be seen from Figure 3.2a below, both VBF categories consist almost exclusively of events from the ggH and VBF processes. Therefore, the prior $p(\mathrm{VBF})$ can be found by *restricting* the sum in Equation 2.21 to these two categories[17] as a subset of $\mathcal{C}_A$ and varying only $p(ggH)/p(\mathrm{VBF})$. Figure 2.6 illustrates this step. With $p(\mathrm{VBF})$ determined in terms of $p(ggH)$, the $t\bar{t}H$ subspace also decouples almost perfectly. The priors $p(t\bar{t}H\text{-leptonic})$ and $p(t\bar{t}H\text{-hadronic})$ can thus be found by restricting the utility function to the two $t\bar{t}H$ event categories and keeping the previously determined priors fixed. In a similar fashion, this scheme can be continued to find the remaining category priors, trading the original ten-dimensional problem for a series of manageable one- and two-dimensional optimization problems.

At the end of this sequential optimization procedure, which can also be regarded as a type of regularization, a good candidate for an optimal set of priors is available. As a last step, all priors are finally varied and optimized simultaneously, restricted to a small hypercube centred on the previously found candidate optimum. The practical implementation of these steps makes use of the software package in [44].

The impact of well-chosen priors on the performance of the categorization is tremendous. As measured by the relative Punzi score defined in Equation 2.20, all event categories manage to attain values in a range between $\Delta\mathcal{P}_{t\bar{t}H\text{-hadr.}} = 0.02$ and $\Delta\mathcal{P}_{\text{VBF-1jet}} = 0.16$ for priors that maximize the utility function in Equation 2.21. This is to be contrasted with values as low as $\Delta\mathcal{P}_{VH\text{-MET}} = -0.42$ for the default choice of flat category priors.

2.2.6 Intransitive Games

With the category priors now determined, all posterior ratios $R_{c,c'}$ are available. As mentioned at the outset of Section 2.1, the category with maximum posterior probability, and hence the result of the classification algorithm, can then be found by a simple voting scheme. This is certainly true if the computed $r_{c,c'}$ were indeed the *correct* ones. In practice, however, they can only be regarded as reasonable approximations. In particular, it is not even guaranteed that the constructed $R_{c,c'}$ can be used to compare categories in a transitive way: a situation with 3 categories c, c', c'' and $R_{c,c'}$, $R_{c',c''}$, $R_{c'',c}$ all greater than 1 is very well possible. Figure 2.7 shows a practical illustration of this fact, produced from a simulated ggH event.

Owing to this intransitivity, situations exist where two or more "winning" categories are found, i.e. several categories receive the same number of votes. In the implemented voting procedure, such a situation is arbitrated as follows: first, a reduced voting is carried out, taking into account only the competing winners. This step is repeated for as long as the number of winners is *smaller* than the number of categories that participate in the voting. This procedure will terminate either when a single winner is found, or, in the presence of a directed cycle such as in Figure 2.7, the categories belonging to this loop are isolated. At this stage, no winner

[17] The normalization of the utility function has been chosen such that the variance of $u(\boldsymbol{\Delta}\mathcal{P})$, as a random variable, is approximately independent of the number of categories included in the sum.

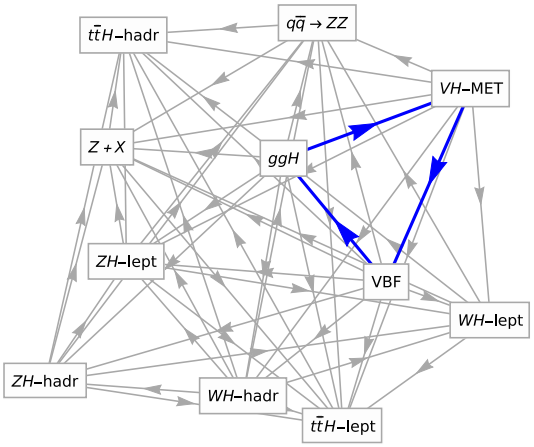


Figure 2.7: Illustration of the relations among the event categories induced by the pairwise posterior ratios. An arrow is drawn pointing from category c to category c' if $R_{c,c'} > 1$. Note the directed cycle formed by the three categories VH-MET, VBF and ggH, an indication of the intransitivity one encounters when comparing categories using approximated likelihood ratios.

can be determined based solely on information from the posterior ratios $R_{c,c'}$. Thus, it is investigated whether the priors alone single out a winner, i.e. the above algorithm is repeated using only the prior ratios $p_{c,c'} = \frac{p(c)}{p(c')}$ instead of $R_{c,c'}$. If this fails as well, the winner is chosen randomly from the leftover categories and returned as the result of the categorization procedure.

It is interesting to generalize the problem of finding the winning category, i.e. determining the solution to Equation 2.4. Indeed, intransitivity is a very common occurrence in the world of sports[18]. We can identify event categories with players in a tournament and declare that a category c will win against c' in a direct match iff $R_{c,c'} > 1$. In this context, the above voting prescription for determining the winning category is known under the term of a *single round-robin tournament*, which is commonly found in chess or football[19]. Another very common scheme is that of a *single-elimination tournament*, typically used in tennis. In the first round of the tournament, pairs of players are matched randomly and play each other. The winners then recursively repeat the same process among themselves while the losers drop out of the competition. This scheme has been implemented as an alternative method, with the following small modification. Contrary to the round-robin case, the final winning category will now depend on the allocation of pairs in the first round, called *seeding*. To obtain a robust indicator of the actual "strength" of a category independent of this assignment, a fixed number of single-elimination tournaments are played, each time with a randomly chosen seeding. If a single category manages to win more than half of the played tournaments, it is returned as the result of the categorization. Otherwise, additional tournaments are played for a fixed, but much larger number of times. If a category can now be identified as the winner, it is returned, otherwise, the same arbitration process is followed as in the round-robin case.

Note that in the absence of any intransitivity, both methods give the same results. In comparison, the single-elimination tournament tends to perform marginally better than the round-robin prescription, with performance improvements ranging from 1% to 3%.

2.3 Summary

We have seen how Bayesian decision theory can be used to define a well-justified approach to the problem of partitioning $H \to 4\ell$ events according to their production modes. The central objects in this procedure are formed by the Bayesian posterior probabilities that an event belongs to a certain event category. Events are then classified by placing them in the category with the highest posterior. This class is found through pairwise comparisons of ratios of posterior probabilities, resembling a tournament.

In practice, the posterior ratios can be obtained from the ratios of the corresponding likelihoods and priors via Bayes' theorem. The former are approximately computed by exploiting their invariance under a certain type of dimensionality-reducing function, conveniently parameterized by artificial neural networks. The latter are finally found directly by demanding the optimality of the classification procedure w.r.t. the Punzi purity, a measure of the sensitivity of an analysis.

[18]For example, in the 2018 Chess World Champion Candidates Tournament, Karjakin beat Kramnik, who in turn won against Aronian. Aronian closed the loop by beating Karjakin.

[19]However, these leagues usually employ *double* round-robins, in which any two players meet twice as the tournament unfolds; playing both white and black in chess, or home and away in football. This subtlety is not relevant for our purposes.

Chapter 3
Results

While Chapter 2 defined a Bayesian method for event classification and described its practical implementation, it will be the content of the present Chapter to evaluate its performance in a realistic setting. To this end, the new method is employed to replace the event classification algorithm currently used by the $H \to 4\ell$ analysis introduced in Section 1.3. Apart from the categorization, the remaining analysis procedure will stay unmodified. Both the current and Bayesian algorithms are then used to obtain results for the signal strengths and their uncertainties *expected* for a SM Higgs boson and the 2017 dataset. Their comparison will provide a meaningful benchmark of the classification performance and the resulting sensitivity of the analysis.

To integrate the Bayesian classification algorithm into the analysis, its statistical model needs to be partially updated. This is done in Section 3.1. Section 3.2 will then assess sources of systematic uncertainties affecting the new classification procedure, and Section 3.3 will outline the final results.

3.1 Updating the Signal Model

The signal model defined in Section 1.3.5 consists of parameterizations of the event yields in every category as well as the event shape as measured by the variables $m_{4\ell}$ and $\mathcal{D}^{c}_{\text{bkg}}$. In principle, both will be affected by a change in the categorization algorithm. However, the *leading* modifications to the sensitivity induced by a new classification are going to be captured already by the resulting changes in the event yields. The modifications in the event shape will be subdominant[1]. Therefore, only the parameterization of the event yields as a function of m_H needs to be updated for Bayesian categorization. The two-dimensional templates created for the present event classification algorithm can be reused.

To perform the yield parameterization, the composition of the eight event categories is evaluated on MC for several Higgs boson mass points in the range from 120 to 130 GeV, separately for each final state. To estimate the expected yields at values of m_H for which no MC samples were available, polynomials are fitted to these datapoints. Depending on the absolute event yield in a category and the number of mass points for which signal MC samples were available, these polynomials are of degree zero, one or two. Figure 3.1 shows examples of parameterized event yields for these three cases. A clear positive trend in all final states is visible as one increases m_H. The yields of ggH events in the untagged category and that of $t\bar{t}H$-hadronic events in the $t\bar{t}H$-leptonic category increase by about a factor of two when going from $m_H = 120\,\text{GeV}$ to $m_H = 130\,\text{GeV}$. This is consistent with the expected enhancement in the product of Higgs boson production cross section and the branching ratio for these final states. Note that, across all categories and production modes, the $4e$ final state produces a lower yield than 4μ, owing to the lower reconstruction efficiency for electrons. From lepton universality, one would expect the two to be equal. The same fact causes the $2e2\mu$ yield to be lower than the expected value of twice the $4e$ or 4μ yield.

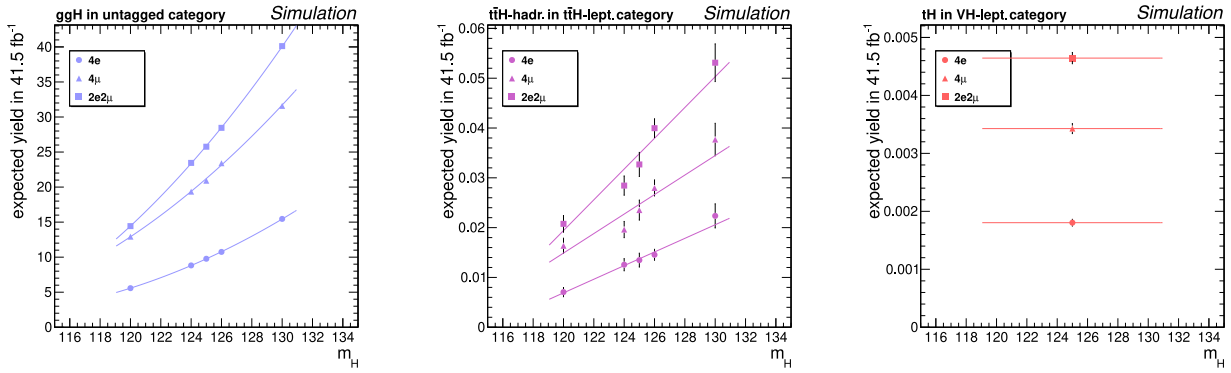
The determination of the expected background yields is also performed using MC for the $q\bar{q} \to ZZ$ and $gg \to ZZ$ backgrounds, and based on the data control regions for the $Z+X$ background, as outlined in Section 1.3.4. Of course, in this case, no parameterization in terms of m_H exists. No unblinding of the signal region is performed.

3.2 Evaluating Systematic Uncertainties

The measurement of the inclusive Higgs boson signal strength μ_{global} is starting to become dominated by systematic uncertainties for the combined 2016 + 2017 dataset comprising $77.4\,\text{fb}^{-1}$. Even in this case, the signal strengths μ_p for the tagged production modes remain dominated by statistical uncertainties. Although our focus so far lay on the μ_p, assessing relevant systematics is crucial for a reliable comparison – significantly increased systematic uncertainties in the Bayesian event classification could partially offset an improved categorization performance.

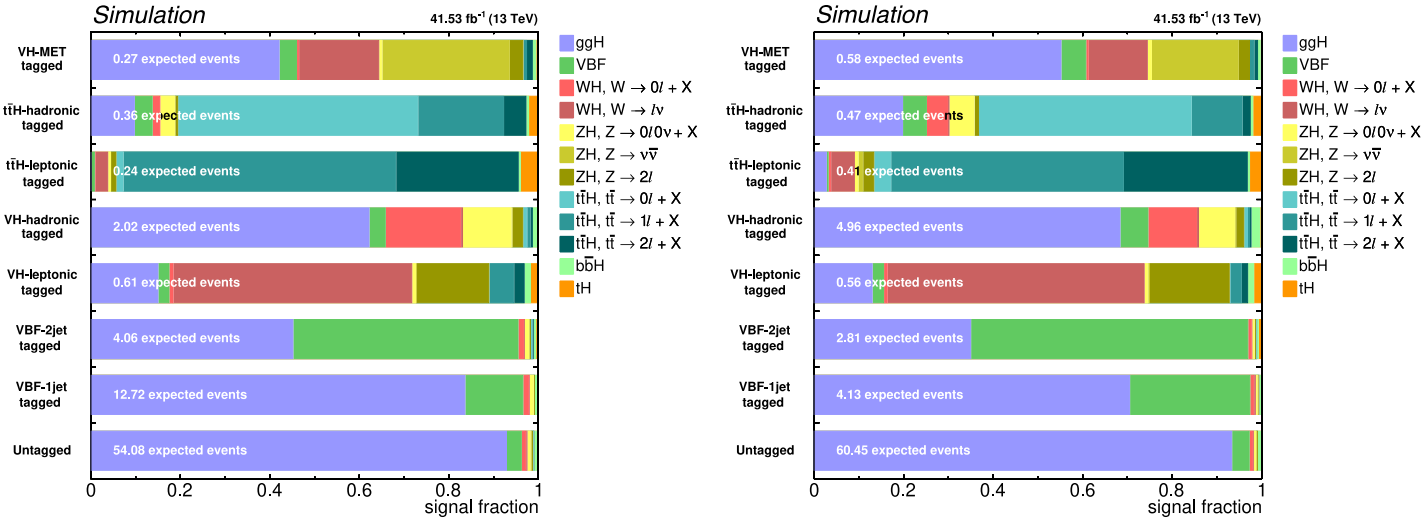
First, systematic uncertainties on the feature vectors $\mathbf{e}$ will directly affect the categorization and lead to the migration of events between individual categories. An essential contribution comes from the uncertainty in the jet energy scale (JES). To assess its impact, jet energies are varied within their corresponding $\pm 1\sigma$ calibration uncertainties and the effect on the signal parameterization is recorded. Uncertainties in the signal shape are

[1] In most categories, the lineshape $\mathcal{L}_{c,f}(m_{4\ell})$ will just contain a peak centred at m_H. The only exception is VH-leptonic, where an additional shoulder is visible due to associated leptons being misidentified as signal leptons from the Higgs boson decay. Only in the latter case, small effects due to a different classification procedure are expected.



(a) (b) (c)

Figure 3.1: Examples of yield parameterizations in the region $105 < m_{4\ell} < 140\,\text{GeV}$. Figure (a) shows the presence of ggH events in the *untagged* category, separate for each final state. Figures (b) and (c) are the analogous plots for $t\bar{t}H$-hadronic events in the $t\bar{t}H$-leptonic category and for tH events in the VH-leptonic category.



(a) current categorization (b) Bayesian categorization

Figure 3.2: Detailed view of the signal composition of the individual categories for $m_H = 125\,\text{GeV}$ in the mass window $105 < m_{4\ell} < 140\,\text{GeV}$. Figure (a) shows the result of the currently employed algorithm, Figure (b) is valid for the Bayesian method. For the signal processes, X denotes anything other than an explicitly stated lepton or neutrino.

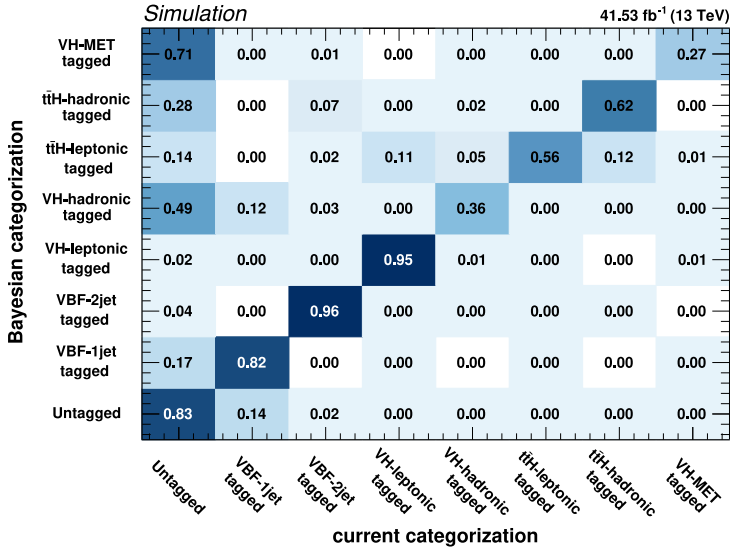


Figure 3.3: Shown is the signal composition of event categories defined by the Bayesian categorization algorithm in terms of categories as defined by the current categorization scheme, in the signal region $105 < m_{4\ell} < 140\,\text{GeV}$. The entries in each row sum to one and therefore directly correspond to percentages.

well-covered by those in the yields, and are therefore not separately taken into account. Event migration due to the JES uncertainty is found to be of a similar magnitude for both the current and Bayesian categorization algorithms. Values range from a 2% uncertainty in the hadronic $t\bar{t}H$ yield in its associated category, up to a 20% uncertainty in the ggH yield in the VBF-2jet category. The uncertainty in the lepton energy scale is propagated in a similar fashion. Its contribution to the uncertainty in the categorization is determined to be negligible[2], affecting the yields at levels far below 1%. The imprecise knowledge of the b-tagging efficiency generates uncertainties of up to 3% in the $t\bar{t}H$-leptonic yield in the VH-leptonic category. This source is assessed by varying the respective data-to-simulation scale factors within their assigned uncertainties. Uncertainties in the modelling of hadronization and the underlying event are evaluated from dedicated MC samples where the modelling parameters have been modified. In practice, this theoretical uncertainty gets combined with those coming from the variation of the factorization and renormalization scales. Again, they are of similar magnitude for both algorithms, ranging from 2% to about 30% for the main Higgs boson production modes in their corresponding categories.

Second, categorization-independent systematic uncertainties are evaluated and included as well, for sake of completeness and to provide an absolute measure of the expected analysis sensitivity. These include uncertainties in the integrated luminosity, the lepton selection efficiencies as well as the pileup.

A comment is in order about the use of events from the $Z + X$ control regions for the training (and validation) of several MLPs, as mentioned in Section 2.2.3.1. Effectively, this will reduce the number of events available for the estimation of the $Z + X$ yield in the signal region and therefore lead to an elevated statistical uncertainty. Thus, the assigned uncertainty in the $Z + X$ yields was increased from about 40% to 60% for all categories and final states. This includes the uncertainties in the fake rates and the estimated $m_{4\ell}$ distribution and also (over)covers the increased statistical uncertainty.

3.3 Results

Figure 3.2 compares the expected signal composition of the individual event categories for the currently used and Bayesian categorization schemes. Figure 3.3 relates them in terms of the expected migration of events between categories as defined by either algorithm.

In these plots, a strong hierarchy between the categories is visible, with *untagged* containing the vast majority of all events. This category is more than 90% pure in ggH events and achieves a similar composition for both algorithms. On the other hand, the VH-MET category more than doubles its expected yield when defined by the Bayesian method. It absorbs a visible fraction of events that were previously located in the untagged category. Although the contamination by ggH increases, it contains about 50% more genuine $ZH, Z \to \nu\bar{\nu}$ signal events than before. The situation is similar for the VH-hadronic category, which now holds about 65% more $WH, W \to 0\ell{+}X$ signal events, at the expense of a higher contamination by the ggH and VBF processes. Nevertheless, its assigned Punzi purity increases by about 55% compared to the current classification scheme. Also the $t\bar{t}H$-leptonic category manages to increase its expected yield significantly, mainly due to the attraction of signal events that were previously misclassified and located in the $t\bar{t}H$-hadronic, VH-leptonic and untagged categories. The Bayesian VBF-2jet category is composed almost exclusively of events that were classified as such also by the current algorithm. However, it trades a higher purity in VBF events against a lower total yield. A similar statement is true for the VBF-1jet category. No big changes occur for the VH-leptonic category, which features a similar composition and event yield for both classification algorithms.

Based on this expected composition of the event categories, Figure 3.4 shows the expected results for the signal strengths μ_p of the five main Higgs boson production modes. Figure 3.5 combines the equivalent plots for the fermionic and bosonic scale factors $\mu_{ggH,t\bar{t}H,b\bar{b}H,tH}$ and $\mu_{\mathrm{VBF},VH}$. Both plots are produced by performing a fit of the signal model to an Asimov dataset generated for the hypothesis of a SM Higgs boson. This dataset is constructed artificially in such a way that the fit returns the *true* signal strengths together with their *expected* uncertainties [46, 47], i.e. those found in the mean of a large ensemble of experiments.

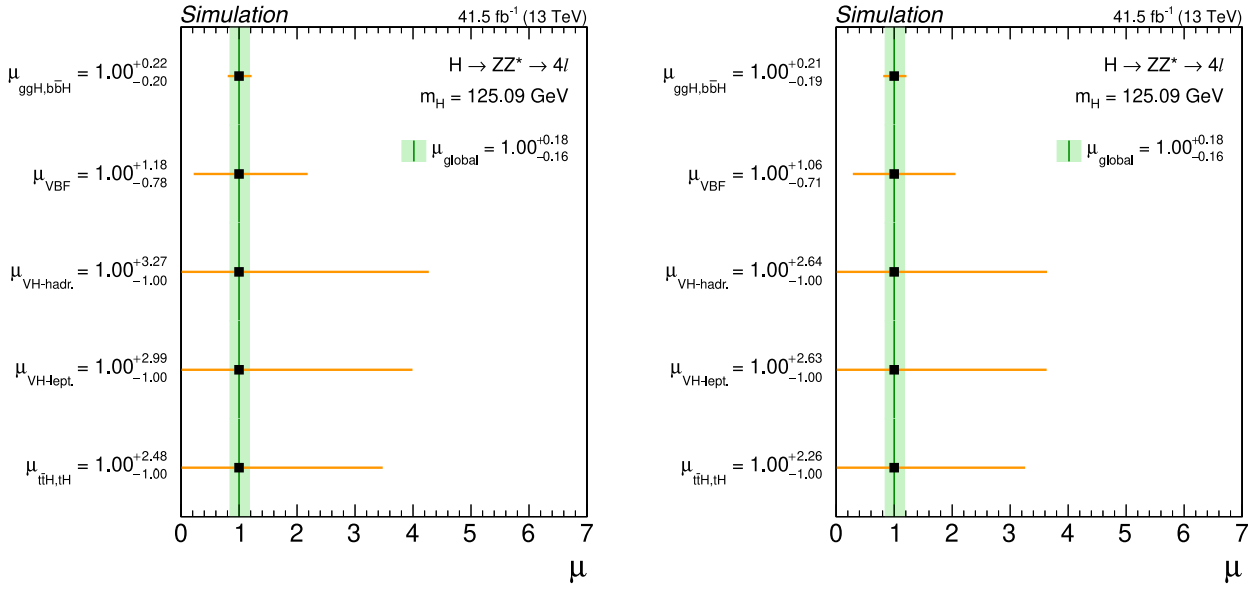
As is apparent from Figures 3.4a and 3.4b, a significant enhancement of the sensitivity of the analysis is expected for all tagged production modes when employing Bayesian event classification. The reduction of the expected uncertainties in μ_{VBF} and $\mu_{t\bar{t}H,tH}$ amount to about 10% and 6%, respectively. The VH production mode can also be constrained better now, with the expected uncertainties in $\mu_{VH\text{-hadr.}}$ and $\mu_{VH\text{-lept.}}$ reduced by about 15% and 10% respectively[3]. The sensitivity to $\mu_{ggH,b\bar{b}H}$ improved only marginally. The inclusive

[2]The situation is different if one attempts to measure m_H using the four-lepton final state. In this case, lepton energy scale uncertainties are an important source of systematics.

[3]Figure 3.4a directly corresponds to the first row of Table 3 of [15]. Compared to the values given there, the large difference in the expected uncertainty in $\mu_{VH\text{-lept.}}$ is caused by the VH-MET category, which is not present in the published analysis. The small remaining differences are caused by VH-MET events migrating to the remaining seven categories, thereby changing their

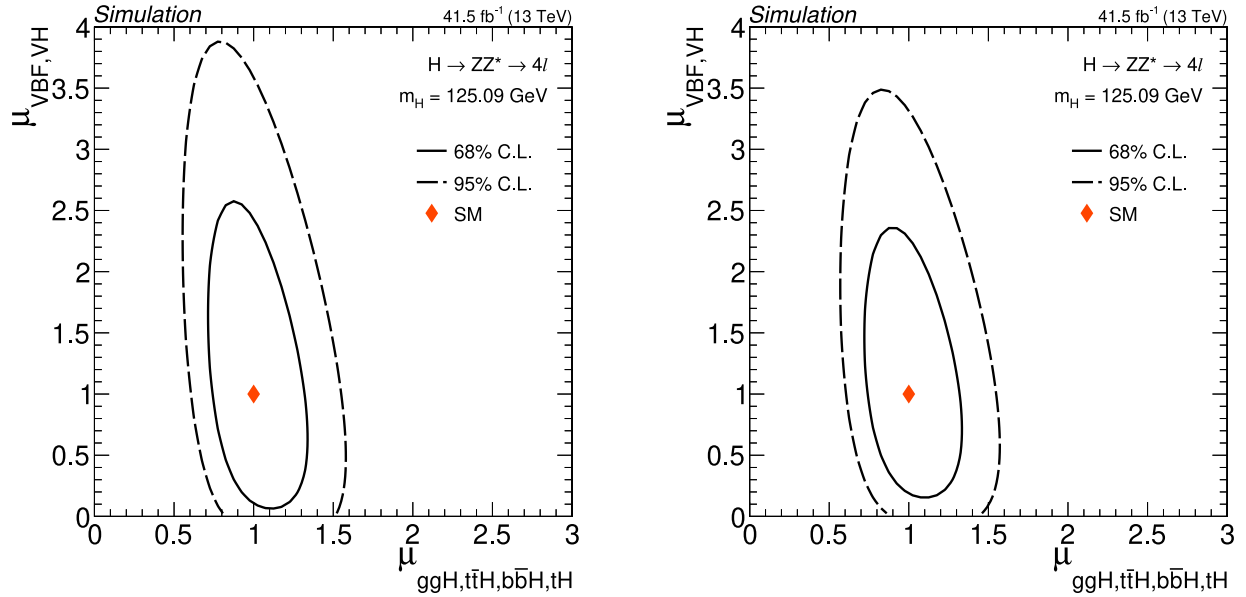
Higgs boson signal strength μ_{global} is independent of the classification and therefore did not change.

Improvements induced by the Bayesian method are also visible when comparing Figures 3.5a and 3.5b. This algorithm leads to expected signal strengths of $\mu_{\text{VBF},VH} = 1.00^{+0.83}_{-0.61}$ and $\mu_{ggH,t\bar{t}H,b\bar{b}H,tH} = 1.00^{+0.19}_{-0.21}$, while the current classification results in $\mu_{\text{VBF},VH} = 1.00^{+0.96}_{-0.69}$ and $\mu_{ggH,t\bar{t}H,b\bar{b}H,tH} = 1.00^{+0.22}_{-0.19}$. Especially the bosonic signal strength $\mu_{\text{VBF},VH}$ is much better constrained by the new method, mainly driven by the improvement in μ_{VBF}. The fermionic signal strength, dominated by the ggH component, remains almost unchanged.



(a) current categorization (b) Bayesian categorization

Figure 3.4: Expected signal strengths for the main SM Higgs boson production modes μ_p and the inclusive signal strength μ_{global}, for the currently used event categorization in (a) and for Bayesian classification in (b). The horizontal bars and the filled bands indicate expected $\pm 1\sigma$ uncertainties, including both statistical and systematic sources.



(a) current categorization (b) Bayesian categorization

Figure 3.5: Results of a 2D likelihood scan in the fermionic and bosonic signal strengths $\mu_{ggH,t\bar{t}H,b\bar{b}H,tH}$ and $\mu_{\text{VBF},VH}$, for the currently used event categorization in (a) and for Bayesian classification in (b). The expected best-fit value for a SM Higgs boson is marked by a red diamond, also shown are the expected 68% and 95% CL regions.

expected composition slightly.

Chapter 4
Summary and Outlook

Into the High Luminosity Future

The principles behind Relativity and Quantum Mechanics cause relativistic quantum fields to be a good low-energy description of nature, regardless of how its dynamics at very high energies really look like. On the one hand, this turns the Standard Model of Particle Physics into a powerful and effective description of phenomena observed at experimentally accessible energy scales. On the other hand, it makes correcting and extending this theory a daunting task.

Most recently, the discovery of a new scalar particle has once again confirmed the validity of this theoretical framework: all measurements can be fully explained by the hypothesis that this state is indeed the single Higgs boson H found in the Standard Model. The Higgs mechanism is central to the structure of this theory, triggering the breaking of the electroweak symmetry and thereby initiating the separation of the weak and electromagnetic interactions. Consequently, its detailed and precise exploration takes a central role also in the experimental programmes of present and future collider experiments.

The decay channel $H \to ZZ^* \to 4\ell$ provides a clean signature that lends itself very well to the study of different production mechanisms of the Higgs boson, and to the measurement of the corresponding couplings. The work presented in this thesis, in line with present efforts by the CMS Experiment at the Large Hadron Collider, provides a small step towards enhancing the experimental sensitivity to this channel.

Based on the formal methods of Bayesian decision theory, an updated event classification algorithm has been described, implemented and tested. Exploiting an ensemble of artificial neural networks for the computation of likelihood ratios, this method allows to draw on the collective information contained in a large number of observables. At the same time, these computations are carried out in a very structured and transparent manner and allow the impact and importance of each input variable to be easily assessed. In this scheme, the event categories themselves are specified *implicitly* through samples of simulated events, generated for their desired signal processes. This represents a departure from more physics-driven paradigms, which often define the categories explicitly through a series of manually engineered selection cuts placed on a handful of event variables.

This new procedure brings about significant improvements in the experimental sensitivity to the signal strength modifiers μ for the five main SM Higgs boson production modes. These were quantified by comparing to a CMS analysis targeting the four-lepton final state, based on $41.5\,\mathrm{fb}^{-1}$ of pp data recorded in 2017. In this context, the expected uncertainty in $\mu_{t\bar{t}H,tH}$ for associated production of a Higgs boson with a top quark pair or single top was reduced by 6%. For vector boson fusion and associated production with a W or Z boson, the gains are even more significant and lie in the range from 10% to 15%.

This success is very encouraging and motivates further, more ambitious efforts in preparing for future precision measurements. Indeed, the amount of recorded data will increase by more than one order of magnitude over the coming decades. From a physics point of view, this calls for an efficient organizing principle that can encode and compress this wealth of experimental measurements into a small number of critical parameters. Effective field theories provide such a universal parameterization in terms of Wilson coefficients, the coupling constants of effective, nonrenormalizable interactions that extend the Standard Model [48]. Future analysis chains will then need to cope with the complex problem of delivering simultaneous measurements – or exclusion limits – of a large number of these parameters. They will also have to be prepared to exploit the dataset to the fullest extent by making use of the entire information delivered by the detector at every processing step.

Current developments have shown [49, 50] how the concepts that were employed here for the classification of Higgs boson events can be extended further. Building extensively on the simulation of the involved physics processes and detectors, the global likelihood needed for the measurement or limit-setting procedure can be estimated *directly* from the raw event information. Moreover, these methods scale well to high-dimensional parameter spaces, making them a prime candidate for extending the physics reach of present and upcoming analyses.

Therefore, with many new and promising ideas looming at the horizon, the era of high luminosities and precision physics will be shaped just as much by advances in analysis strategies as it will be supported by progress in detector design. We can be confident that both taken together will, if not enable a discovery, at least point us to the best streetlight under which to look for it.

Bibliography

[1] E. Wigner, "On Unitary Representations of the Inhomogeneous Lorentz Group", Annals of Mathematics **40** (1939), doi:10.2307/1968551.

[2] S. Weinberg, "The Quantum Theory of Fields", Vol. 1, Cambridge University Press (1995).

[3] S. Weinberg, "Derivation of Gauge Invariance and the Equivalence Principle from Lorentz Invariance of the S-Matrix", Phys. Lett. **9** (1964), doi:10.1016/0031-9163(64)90396-8.

[4] C. Becchi, "Introduction to BRS Symmetry", (2009), arXiv:hep-th/9607181.

[5] K. G. Wilson and J. Kogut, "The Renormalization Group and the ϵ Expansion", Phys. Rep. **12** (1974), doi:10.1016/0370-1573(74)90023-4.

[6] R. S. Chivukula, D. A. Dicus, and H.-J. He, "Unitarity of Compactified Five-Dimensional Yang–Mills Theory", Phys. Lett. B **525** (2002), doi:10.1016/S0370-2693(01)01435-6, arXiv:hep-ph/0111016v3.

[7] LHC Higgs Cross Section Working Group, D. de Florian et al., "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector", CERN Yellow Reports: Monographs, Vol. 2/2017, doi:10.23731/CYRM-2017-002, arXiv:1610.07922 [hep-ph].

[8] LHC Higgs Cross Section Working Group, S. Heinemeier et al., "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties", CERN Yellow Reports: Monographs, Vol. 4/2013 , doi:10.5170/CERN-2013-004, arXiv:1307.1347 [hep-ph].

[9] The CMS Collaboration, "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC", Phys. Lett. B **716** (2012), doi:10.1016/j.physletb.2012.08.021, arXiv:1207.7235 [hep-ex].

[10] The ATLAS Collaboration, "Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC", Phys. Lett. B **716** (2012), doi:10.1016/j.physletb.2012.08.020, arXiv:1207.7214 [hep-ex].

[11] The CMS Collaboration, "Constraints on the Spin-Parity and Anomalous HVV Couplings of the Higgs Boson in Proton Collisions at 7 and 8 TeV", Phys. Rev. D **92**, 012004 (2015), doi:10.1103/PhysRevD.92.012004, arXiv:1411.3441 [hep-ex].

[12] The ATLAS and CMS Collaborations, "Measurements of the Higgs Boson Production and Decay Rates and Constraints on its Couplings from a Combined ATLAS and CMS Analysis of the LHC pp Collision Data at $\sqrt{s} = 7$ and 8 TeV.", JHEP **16**, 45 (2016), doi:10.1007/JHEP08(2016)045, arXiv:1606.02266 [hep-ex].

[13] The ATLAS and CMS Collaborations, "Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments", Phys. Rev. Lett. **114** (2015), doi:10.1103/PhysRevLett.114.191803, arXiv:1503.07589 [hep-ex].

[14] The CMS Collaboration, "The CMS Experiment at the CERN LHC", JINST **3**, S08004 (2008), doi:10.1088/1748-0221/3/08/S08004.

[15] The CMS Collaboration, "Measurements of Properties of the Higgs Boson in the Four-Lepton Final State at $\sqrt{s} = 13$ TeV", CMS-PAS-HIG-18-001 (2018), http://cds.cern.ch/record/2621419/files/HIG-18-001-pas.pdf (visited on 06/14/2018).

[16] The CMS Collaboration, "Measurements of Properties of the Higgs Boson Decaying into the Four-Lepton Final State in pp Collisions at $\sqrt{s} = 13$ TeV", JHEP **17**, 47 (2017), doi:10.1007/JHEP11(2017)047, arXiv:1706.09936 [hep-ex].

[17] I. Anderson et al., "Constraining Anomalous HVV Interactions at Proton and Lepton Colliders", Phys. Rev. D **89**, 035007 (2014), doi:10.1103/PhysRevD.89.035007, arXiv:1309.4819.

[18] The ATLAS Collaboration, "Supersymmetry Physics Results", https://twiki.cern.ch/twiki/bin/view/AtlasPublic/SupersymmetryPublicResults (visited on 06/14/2018).

[19] C. Englert et al., "Precision Measurements of Higgs Couplings: Implications for New Physics Scales", J. Phys. G **41**, 113001 (2014), doi:10.1088/0954-3899/41/11/113001, arXiv:1403.7191 [hep-ph].

[20] F. Goertz, A. Papaefstathiou, L. Yang, and J. Zurita, "Higgs Boson Self-Coupling Measurements using Ratios of Cross Sections", JHEP **13**, 16 (2013), doi:10.1007/JHEP06(2013)016, arXiv:1301.3492 [hep-ph].

[21] S. D. Vita, C. Grojean, G. Panico, M. Riembau, and T. Vantalon, "A Global View on the Higgs Self-Coupling", JHEP **17**, 69 (2017), doi:10.1007/JHEP09(2017)069, arXiv:1704.01953 [hep-ph].

[22] J. Pratt, H. Raiffa, and R. Schlaifer, "Introduction to Statistical Decision Theory", MIT Press (2008).

[23] M. A. T. Figueiredo, "Lecture Notes on Bayesian Estimation and Classification", Instituto de Telecomunicações, and Instituto Superior Técnico, Lisbon, (2004), http://www.lx.it.pt/~mtf/learning/Bayes_lecture_notes.pdf (visited on 06/14/2018).

[24] K. Cranmer, J. Pavez, and G. Louppe, "Approximating Likelihood Ratios with Calibrated Discriminative Classifiers", (2016), arXiv:1506.02169v2 [stat.AP].

[25] P. Windischhofer, "Bayes4Leptons", GitHub repository, https://github.com/philippwindischhofer/ZZAnalysis (visited on 06/14/2018).

[26] P. L. Bartlett and M. Traskin, "AdaBoost is Consistent", Journal of Machine Learning Research **8** (2007).

[27] A. Faragó and G. Lugosi, "Strong Universal Consistency of Neural Network Classifiers", IEEE Transactions on Information Theory **39** (1993), doi:10.1109/18.243433.

[28] K. Hornik, "Approximation Capabilities of Multilayer Feedforward Networks", Neural Networks **4** (1991), doi:10.1016/0893-6080(91)90009-T.

[29] E. Bagnaschi, G. Degrassi, P. Slavich, and A. Vicini, "Higgs Production via Gluon Fusion in the POWHEG Approach in the SM and in the MSSM", JHEP **12**, 88 (2012), doi:10.1007/JHEP02(2012)088, arXiv:1111.2854 [hep-ph].

[30] P. Nason and C. Oleari, "NLO Higgs Boson Production via Vector-Boson Fusion Matched with Shower in POWHEG", JHEP **10**, 37 (2010), doi:10.1007/JHEP02(2010)037, arXiv:0911.5299 [hep-ph].

[31] Y. Gao, A. Gritsan, Z. Guo, K. Melnikov, M. Schulze, and N. Tran, "Spin Determination of Single-Produced Resonances at Hadron Colliders", Phys. Rev. D **81**, 075022 (2010), doi:10.1103/PhysRevD.81.075022, arXiv:1001.3396 [hep-ph].

[32] T. Sjöstrand et al., "An Introduction to PYTHIA 8.2", Comput. Phys. Commun. **191** (2015), doi:10.1016/j.cpc.2015.01.024, arXiv:1410.3012 [hep-ph].

[33] J. M. Campbell, R. K. Ellis, and C. Williams, "Bounding the Higgs Width at the LHC Using Full Analytic Results for $gg \to 2e2\mu$", JHEP **14**, 60 (2014), doi:10.1007/JHEP04(2014)060, arXiv:1311.3589 [hep-ph].

[34] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", Annals of Statistics **29** (2001), doi:10.1214/aos/1013203451.

[35] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System", Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016), doi:10.1145/2939672.2939785, arXiv:1603.02754 [cs.LG].

[36] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory", Neural Comput. **9** (1997), doi:10.1162/neco.1997.9.8.1735.

[37] L. Bottou, "Large-Scale Machine Learning with Stochastic Gradient Descent", Proceedings of COMPSTAT'2010 (2010), doi:10.1007/978-3-7908-2604-3_16.

[38] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization", (2017), arXiv:1412.6980v9 [cs.LG].

[39] F. Chollet, "Keras", GitHub repository, https://github.com/fchollet/keras (visited on 06/14/2018).

[40] M. Abadi et al., "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems", (2016), arXiv:1603.04467 [cs.DC].

[41] R. K. Srivastava, J. Masci, F. Gomez, and J. Schmidhuber, "Understanding Locally Competitive Networks", (2015), arXiv:1410.1165 [cs.NE].

[42] D. Freedman and P. Diaconis, "On the Histogram as a Density Estimator: L_2 Theory", Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete **57** (1981), doi:10.1007/BF01025868.

[43] G. Punzi, "Sensitivity of Searches for New Signals and its Optimization", (2003), arXiv:physics/0308063 [physics.data-an].

[44] E. Brochu, V. Cora, and N. de Freitas, "A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modelling and Hierarchical Reinforcement Learning", (2010), arXiv:1012.2599 [cs.LG].

[45] A. Shah, A. G. Wilson, and Z. Ghahramani, "Student-t Processes as Alternatives to Gaussian Processes", (2014), arXiv:1402.4306 [stat.ML].

[46] The ATLAS and CMS Collaborations and the LHC Higgs Combination Group, "Procedure for the LHC Higgs Boson Search Combination in Summer 2011", ATL-PHYS-PUB-2011-11 / CMS NOTE-2011/005 (2011), http://cds.cern.ch/record/1379837/files/NOTE2011_005.pdf (visited on 06/14/2018).

[47] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic Formulae for Likelihood-Based Tests of New Physics", Eur. Phys. J. C **71** (2011), doi:10.1140/epjc/s10052-011-1554-0, arXiv:1007.1727 [physics.data-an].

[48] B. Henning, X. Lu, and H. Murayama, "How to use the Standard Model Effective Theory", (2015), arXiv:1412.1837 [hep-ph].

[49] J. Brehmer, K. Cranmer, G. Louppe, and J. Pavez, "A Guide to Constraining Effective Field Theories with Machine Learning", (2018), arXiv:1805.00020 [hep-ph].

[50] J. Brehmer, G. Louppe, J. Pavez, and K. Cranmer, "Mining Gold from Implicit Models to Improve Likelihood-Free Inference", (2018), arXiv:1805.12244 [stat.ML].

[51] S. Regnard, "Measurements of Higgs Boson Properties in the Four-Lepton Final State at $\sqrt{s} = 13\,\text{TeV}$ with the CMS Experiment at the LHC", PhD Thesis (École Polytechnique, 2016), https://cds.cern.ch/record/2255120/files/TS2017_002_2.pdf (visited on 06/14/2018).

[52] T. Roxlo and M. Reece, "Opening the Black Box of Neural Nets: Case Studies in Stop / Top Discrimination", (2018), arXiv:1804.09278 [hep-ph].

Appendix A
Event Information

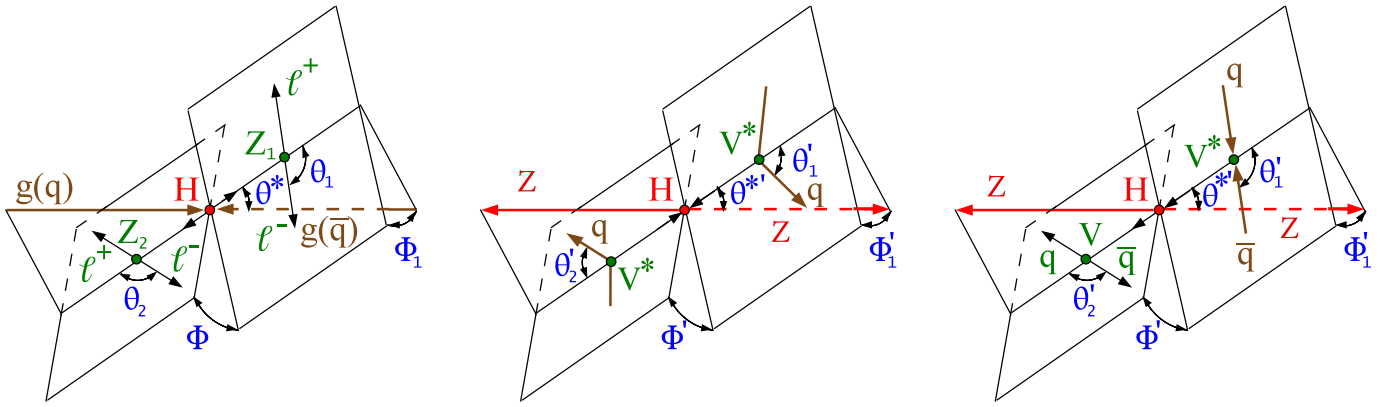
A.1 Coordinate System

The CMS Experiment employs a coordinate system whose x-axis points towards the centre of the LHC ring and the y axis points vertically upwards. As a right-handed system, the positive z axis lies in the direction of the counter-clockwise revolving proton beam. The azimuthal angle ϕ is defined in the x/y-plane and is measured from the x-axis. The radial distance from the interaction point in the centre of the detector is labelled as r. The polar angle θ lies in the z/r-plane. In many cases, the pseudorapidity $\eta = -\log\tan\frac{\theta}{2}$ is a more convenient quantity than the polar angle itself.

A.2 Event Kinematics

The kinematic discriminant $\mathcal{D}_{\text{bkg}}$, sensitive to the *decay* of the Higgs boson into four leptons, $H \to 4\ell$, is computed from $\mathbf{\Omega}^{H\to 4\ell}$. This vector contains a set of observables that fully specify the kinematics of this decay process. For the decay of a massive scalar boson into four on-shell leptons, in total eight degrees of freedom are needed. One possible choice is to use the four-lepton invariant mass $m_{4\ell}$, the masses $m(Z_{1,2})$ of the Z_1 and Z_2 candidates and in addition five angles, θ_1, θ_2, θ^*, Φ and Φ_1, all defined in Figure A.1a. Note that this still leaves arbitrary (but irrelevant) rotations around the beam axis, symbolized by the incoming quarks or gluons. This freedom is fixed by another angle ξ^*.

The *production* discriminants $\mathcal{D}_{\text{VBF-2j},ggH}$ and $\mathcal{D}_{VH\text{-hadr.},ggH}$ are sensitive to the kinematics of the VBF and VH-hadronic production modes for events with two reconstructed jets. Both make use of the kinematic vector $\mathbf{\Omega}^{H+JJ}$. This vector holds the information to completely specify the kinematics of a Higgs boson produced in association with two jets. Again, five angles are required to define the event at leading order. One possible choice is to use θ_1', θ_2', $\theta^{*\prime}$, Φ' and Φ_1'. As explained in Figures A.1b and A.1c, they are defined differently from the unprimed quantities above. Moreover, their definitions depend on the production mode under consideration.



(a) $H \to 4\ell$ decay (b) VBF production (c) VH-hadronic production

Figure A.1: Figure (a) illustrates the definition of the kinematic angles in $\mathbf{\Omega}^{H\to 4\ell}$ for the $H \to 4\ell$ decay. The angles θ_1 and θ_2 are defined in the rest frames of the corresponding Z bosons. The remaining angles Φ, Φ_1 and θ^* are defined in the rest frame of the Higgs boson. A sixth angle, ξ^*, is not shown. Also defined in the Higgs boson rest frame, it specifies the azimuth of Z_1 w.r.t. the x-axis of the detector frame. Figures (b) and (c) define the kinematic angles in $\mathbf{\Omega}^{H+JJ}$ in the case of the VBF and VH-hadronic production modes. In both situations, the Higgs boson is produced in association with two jets, but the definition of the individual angles is different. As before, Φ', Φ_1' and $\theta^{*\prime}$ are defined in the rest frame of the Higgs boson. For the VBF production mode, θ_1' and θ_2' are defined in the detector frame; for VH-hadronic, they are defined in the rest frames of V^* and V, respectively. Figures taken from [51] (modified).

Appendix B
Reverse-Engineering Neural Networks

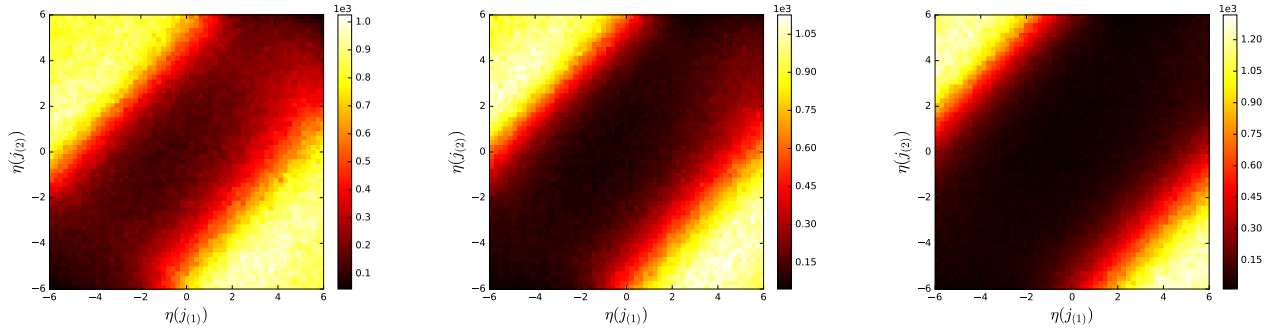
The functions $s_{c,c'}$ introduced in Section 2.1 are chosen such as to minimize the mean squared error in Equation 2.10 or, equivalently, to be monotonous in the likelihood ratios $r_{c,c'}$. Apart from this requirement, the details of the functional mapping are not predetermined. These details are encoded in the weights of the internal neurons of the multilayer perceptrons (MLPs) that compute the $s_{c,c'}$. By definition, the hidden neurons are not directly visible or accessible. Still, it is beneficial to gain some insight into exactly how individual components of the feature vector enter into $s_{c,c'}$. That is, we would like to identify the relevant, characteristic features of those events that lead to high values of $s_{c,c'}$, i.e. provide evidence for c over c'.

To this end, we can artificially create events $\mathbf{e}$ that strongly *activate*[1] the output neuron of the MLP for a specific $s_{c,c'}$. This is easily accomplished by starting from a randomly chosen $\mathbf{e}$ and then performing a gradient *ascent* in its components. This process is stopped once $s_{c,c'}(\mathbf{e})$ reaches a specific threshold. If the procedure is iterated many times, the characteristic features of the resulting collection of synthesized events $\mathbf{e}$ can well be called *relevant*, at least from the point of view of $s_{c,c'}$. Graphical visualizations can then easily be obtained by simply histogramming individual components of the generated feature vectors.

Note that the events produced in this way need not be physical, i.e. they may not (and will not, in general) conserve momentum or energy, or even contain an integer number of jets. Therefore, in principle, of all generated events, only the physical ones should be retained. However, the *dominant* input features in $\mathbf{e}$ are expected to produce high values for $s_{c,c'}$ *regardless* of whether or not the other variables in the event are physical. Therefore, in the examples presented here, the restriction to physical events is made only in the subspace of actually histogrammed variables, considerably speeding up the process.

Figure B.1 shows such an *activation maximization diagram*, obtained for the MLP implementing the ≥ 2-jet component of $s_{\mathrm{VBF},\,ggH}$. By histogramming only in the pseudorapidities of the two hardest jets, $\eta(j_{(1)})$ and $\eta(j_{(2)})$, all other event variables are effectively averaged out. As expected, $s_{\mathrm{VBF},\,ggH}$ produces high values if the two jets are separated by a large $\Delta\eta$, indeed a characteristic feature of VBF events. Two central jets, or two forward jets with low $\Delta\eta$ lead to the event being regarded as ggH-like. Also, the presence of a clear boundary separating these two regions confirms the conjecture made above: high $\Delta\eta$ is *necessary* for the network to deliver high output, *independent* of the values of the other observables. This boundary gradually becomes sharper as one demands higher values for $s_{\mathrm{VBF},\,ggH}$, i.e. restricts to events that are deemed *very* VBF-like.

Here, the technique of activation maximization managed to recover a (well-known) characteristic feature of the data. The authors of [52] use the same method to extract useful correlations between event variables that can aid in discriminating the decay of supersymmetric particles from Standard Model backgrounds. Also in this case, the authors confirmed that neural networks internally make use of concepts very similar to manually engineered discriminative variables. In some cases, the networks had even implemented simpler, but equally effective, variations.



(a) $s_{\mathrm{VBF},\,ggH} > 0.5$ (b) $s_{\mathrm{VBF},\,ggH} > 0.7$ (c) $s_{\mathrm{VBF},\,ggH} > 0.9$

Figure B.1: Activation maximization diagrams generated for the neural network that implements $s_{\mathrm{VBF},\,ggH}$. The number of jets in the event is fixed to two. In this specific example, matrix element discriminants are *not* used as inputs to the MLP. Shown is the projection of the generated feature vectors onto the plane spanned by the pseudorapidities of the two jets. Events with high $\Delta\eta$ between the jets lead to significant activation.

[1] This terminology is borrowed from biology. When a biological neuron becomes activated and fires, it typically produces a spike train at its output. In the context of artificial neural networks, the term simply means a high numerical output produced by a certain neuron.