GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

Fakultät für Physik

Bachelor's Thesis

Einsatz von Multivariaten Analysemethoden zur Untersuchung von Symmetrien der top-Higgs-Yukawa-Wechselwirkung in tHjb Ereignissen mittels pp-Kollisionsdaten des ATLAS-Detektors

Employing multivariate analysis algorithms to study symmetry properties of the top-Higgs-Yukawa interaction in tHjb events using pp collision data recorded with the ATLAS detector

prepared by

Luise Köhler

from Jena

at the II. Physikalisches Institut

Thesis number: II.Physik-UniGö-BSc-2026/05

Thesis period: 30th March 2026 until 5th July 2026

First referee: Prof. Dr. Arnulf Quadt

Second referee: Prof. Dr. Joern Grosse-Knetter

Contents

Abstract

Die CP-Symmetrie der Yukawa-Kopplung des Higgs-Bosons an Fermionen ist bislang experimentell nur schwach beschränkt. Aufgrund seiner hohen Masse ist das Top-Quark besonders geeignet, um diese Kopplung am Large Hadron Collider zu untersuchen. Diese Arbeit analysiert $t\bar{t}H$-, $tHjb$- und tWH-Produktion im voll-hadronischen Zerfallskanal $H \to \tau\tau$. Zur Klassifikation der Ereignisse in Signal- und dominante Untergrundprozesse wird ein Hypergraph-Neuronales Netz (HGNN) eingesetzt. Die Sensitivität CP-empfindlicher Observablen in den Signalprozessen wird untersucht. Auf Basis dieser Variablen werden für die Prozesse $tHjb$ und $t\bar{t}H$ jeweils binäre Klassifikatoren auf der Grundlage von Gradient Tree Boosting trainiert, um das Standardmodell-Szenario einer CP-symmetrischen Kopplung von CP-gemischten Szenarien zu unterscheiden, die durch das Higgs Characterisation Model beschrieben werden. Ihr Ergebnis wird in den zuvor durch das HGNN definierten $tHjb$- und $t\bar{t}H$-Signalregionen für verschiedene CP-Mischungswinkel analysiert. Zuletzt wird die Klassifikationsleistung des HGNN in zusätzlichen, durch die CP-Klassifikation definierten Signalregionen untersucht.

Stichwörter: Higgs-Boson, Top-Quark, CP-Symmetrie, Yukawa-Kopplung, Hypergraph Neuronales Netz, Gradient Tree Boosting, Large Hadron Collider

Abstract

The CP symmetry properties of the Yukawa coupling of the Higgs boson to fermions are not yet strongly constrained. Due to its large mass, the top quark provides an ideal probe of this coupling at the Large Hadron Collider. This analysis investigates $t\bar{t}H$, $tHjb$ and tWH production in the full-hadronic $H \to \tau\tau$ decay channel. A multi-class Hypergraph Neural Network (HGNN) is employed to classify events into the main signal and background processes. The performance of CP-sensitive observables is studied for each signal process. Using these observables as input features, binary Gradient Tree Boosting classifiers are trained separately for the $t\bar{t}H$ and $tHjb$ signal processes to distinguish the Standard Model scenario with a CP-symmetric coupling from CP-mixed scenarios described by the Higgs Characterisation model. The classifier outputs are evaluated in the HGNN-defined $tHjb$ and $t\bar{t}H$ signal regions for different CP-mixing scenarios. Lastly, the performance of the HGNN is assessed in additional signal regions defined by the CP-classification output.

Keywords: Higgs boson, top quark, CP symmetry, Yukawa coupling, Hypergraph Neural Network, Gradient Tree Boosting, Large Hadron Collider

1. Introduction

At the European Organization of Nuclear Research, known as CERN, research in fundamental particle physics seeks to advance our understanding of the constituents of matter and the interactions governing their behaviour. Experiments conducted at particle accelerators such as the Large Hadron Collider (LHC), currently the world's largest and highest-energy particle accelerator, are employed to test predictions of the Standard Model (SM) of Particle Physics, a quantum field theory generally considered the most successful theory to describe particles and their interactions to date. A more detailed description of the SM is provided in Chapter 2.1.

The ATLAS (A Torroidal LHC ApparatuS) experiment is one of four major detectors built to study proton-proton collisions at the LHC. In 2012, results from the ATLAS experiment and the CMS (Compact Muon Solenoid) experiment independently discovered the existence of a previously unobserved elementary particle consistent with the Higgs boson predicted in 1964 [1–6]. Subsequent precision measurements have further demonstrated this consistency and the discovered particle is now widely accepted as the Higgs boson within the SM framework. In the SM, the Higgs boson interacts with fermions via the Yukawa coupling, with a coupling strength proportional to the fermion mass. The top quark is the heaviest particle in the SM and therefore provides a particularly sensitive probe of the Yukawa coupling. Among the various Higgs boson production mechanisms at the LHC, the thesis will focus on production in association with a single top quark, specifically the $tHjb$ process in which a top quark and a Higgs boson are produced alongside a bottom quark and additional jets.

The SM does not contain sufficient charge-parity (CP) symmetry violation to explain the observed baryon asymmetry of the universe [7, 8]. Searches for new CP-violating interactions, e.g. in the couplings of the Higgs boson, are therefore important research targets at the LHC. While the coupling of the Higgs boson to heavy vector bosons (W, Z) has already been highly constrained in previous analyses, there are not yet many tests of the coupling to fermions. In the SM, the coupling of the Higgs boson to the top quark is predicted to be CP-even (i.e. symmetric under charge-parity transformation). Deviations from that prediction would cause a discrepancy between the predicted and measured cross sections for processes involving such a coupling. The production of a single top quark along with a Higgs boson (tH production) has not yet been observed at the LHC, primarily due to its small predicted cross section. However, with the increasing data volume from the recent data-taking period (extending to June 2026) and the high-luminosity upgrade scheduled to commence in 2030, observations of this process appear plausible in the near future. Therefore, its analysis is of great scientific interest.

This thesis investigates the performance of CP-sensitive observables constructed from kinematic quantities of objects measured with the detector and trains machine learning models that are based on these CP-sensitive observables and employ Gradient Tree Boosting to separate a CP-symmetric from a CP-symmetry-violating top-quark Yukawa coupling to the Higgs boson. The models are integrated into existing graph neural network architectures used for event classification, aiming to detect possible deviations from the SM prediction of a CP-even Yukawa coupling between the top quark and the Higgs boson in future analyses of tH production, particularly in $tHjb$ processes. As additional signal processes, $t\bar{t}H$ (top quark pair production in association with a Higgs boson) and tWH (tH production with an associated weak vector-boson) are studied.

2. Theoretical Background

In this Chapter, the theoretical framework underlying the topic of the thesis is introduced, starting with an overview of the Standard Model of Particle Physics in Section 2.1, the currently best-fitting theoretical description for elementary particles and their interactions. The Brout-Englert-Higgs mechanism and the Yukawa coupling of the top quark and the Higgs boson, as well as the relevant Higgs boson production and decay channels, are explained in detail in Sections 2.2 and 2.3.

2.1. The Standard Model of Particle Physics

The Standard Model (SM) of Particle Physics [9–21] is a theoretical model containing all known elementary particles and a description of three of the four fundamental forces as interactions between the particles. Mathematically, it is a renormalisable quantum field theory based on a set of fields (and corresponding quantum numbers) and the gauge symmetries $SU(3)_C \times SU(2)_L \times U(1)_Y$. The subscripts denote the corresponding charges: C refers to the colour charge governing the strong interaction, L specifies the coupling of the left-chiral component of fermions to the weak interaction, and Y denotes the weak hypercharge associated with the electroweak (EW) sector (see Section 2.2.1). The Lagrangian density is then fixed by requiring it to be locally gauge-invariant and renormalisable. It can be written as a sum of four terms [22],

$$\mathcal{L}_{\text{SM}} = \mathcal{L}_{\text{Gauge}} + \mathcal{L}_{\text{Matter}} + \mathcal{L}_{\text{Yukawa}} + \mathcal{L}_{\text{Higgs}} \quad . \tag{2.1}$$

The first component $\mathcal{L}_{\text{Gauge}}$ is a pure gauge Lagrangian corresponding to the kinetic energy and self-interaction of the gauge fields associated with the gauge bosons (also known as vector bosons) of spin 1. They mediate three of the four fundamental forces. The eight gluons are massless and carry a colour and an anti-colour charge, the quantum number mediating the strong force as described by the theory of Quantum Chromodynamics (QCD) [23, 24]. Noether's first theorem yields eight possible linearly independent linear combinations of colour charges, resulting in eight gluons. The photon is the massless and neutral gauge boson that mediates the electromagnetic interaction, interacting only with charged particles [25–32]. The weak force is mediated by the charged $W^{\pm}$ bosons and the neutral Z^0 boson [11–13].

The second term $\mathcal{L}_{\text{Matter}}$ contains the kinetic energy of the spin-1/2 particles of the SM, the fermions, and their interactions with the gauge fields. Fermions can be separated into six flavours (types) of leptons, which do not undergo strong interactions, and six flavours of quarks, which do. Both groups are divided into three generations. Each generation of leptons contains a charged lepton (electron e^-, muon μ^- and tau lepton τ^-) with third weak isospin component $I_3 = -\frac{1}{2}$ and a neutral and massless lepton, their respective neutrino $(\nu_e, \nu_\mu, \nu_\tau)$ with $I_3 = \frac{1}{2}$. Together, the charged lepton and its neutrino form a weak isospin doublet. Importantly, only the left-chiral components of the fermion fields transform as weak isospin doublets under $SU(2)_L$, whereas the right-chiral components are singlets under the symmetry and therefore do not interact weakly. Each quark generation consists of a quark of charge $+\frac{2}{3}$ (up-type: up, charm and top quark) and $I_3 = \frac{1}{2}$ and a quark of charge $-\frac{1}{3}$ (down-type: down, strange and bottom quark) and $I_3 = -\frac{1}{2}$, again forming weak isospin doublets.

The next term $\mathcal{L}_{\text{Yukawa}}$ describes the Yukawa interaction of the Higgs field [1–6], corresponding to the spin-0 Higgs boson, with fermions and will be described in Section 2.2.2. The last term $\mathcal{L}_{\text{Higgs}}$ contains the kinetic energy of the Higgs field and a potential term. It describes the mass and interaction terms for the gauge bosons and the Higgs boson and is further explained in Section 2.2.1.

Associated with each of the particles described previously, there is an antiparticle carrying the reversed sign on all additive quantum numbers, such as electric charge, and the same properties otherwise. Consequently, there exist particles that are their own antiparticles, such as the Higgs boson. A summary of the SM is shown in Figure 2.1.

To date, the SM is the most successful attempt at describing our universe at the smallest scales. Nonetheless, it cannot explain gravity, dark energy or dark matter. Dark matter is a postulated type of matter that would explain observed rotation velocity curves of galaxies, which cannot be explained by the visible matter and the particle content of the SM. A lot of theories predict dark matter particles light enough to be created at the LHC. Even though such a particle cannot be detected directly, if it were created, measurements of missing energy and momentum could provide a hint at its properties. One of the main objectives of the LHC and future colliders will be further tests of SM predictions.

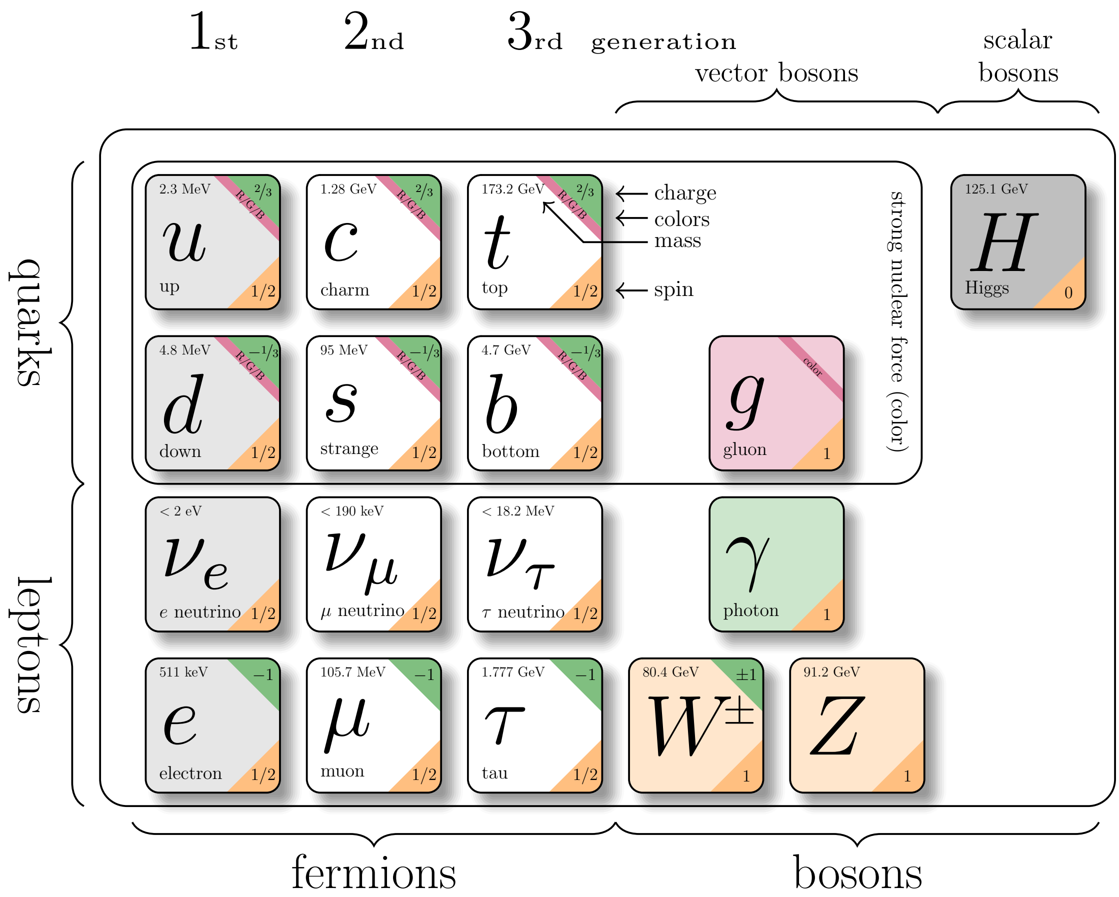


Figure 2.1.: The elementary particles predicted by the SM. Data from the Particle Data Group [33], figure reproduced from [34].

2.2. Higgs Mechanism and Yukawa Coupling

This Section provides an overview of the Brout-Englert-Higgs mechanism, which explains how mass terms for the $W^{\pm}$ and Z^0 gauge bosons arise through electroweak symmetry breaking. Furthermore, the Yukawa interaction which gives rise to fermionic mass terms and its CP symmetry properties as predicted by the SM will be introduced [10, 35–37].

2.2.1. The Brout-Englert-Higgs Mechanism

Experiments have confirmed both the W and Z boson to be massive particles. However, the gauge Lagrangian itself does not contain a mass term, as directly adding one would break the required invariance under $SU(2)_L \times U(1)_Y$. One solves this problem by defining a complex spin-0 four-component scalar field $\phi(T_w = \frac{1}{2}, Y_W = 1)$, where T_W denotes the weak isospin, the generator of $SU(2)_L$, and Y_W the weak hypercharge defined as

$$Y_W = 2(Q - T_w^3) \quad \text{and} \quad T_W^i = \sigma^i/2\,, \tag{2.2}$$

where σ^i are the Pauli matrices. The field ϕ defined as

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2}}(\phi_3 + i\phi_4) \\ \frac{1}{\sqrt{2}}(\phi_1 + i\phi_2) \end{pmatrix} \tag{2.3}$$

forms a doublet of $SU(2)_L$ and is known as the Higgs field. Here, the ϕ_i are real scalar fields. The Higgs Lagrangian $\mathcal{L}_{\text{Higgs}}$ is then defined as

$$\mathcal{L}_{\text{Higgs}} = (D_\mu\phi)^\dagger D_\mu\phi + \mu^2\phi^\dagger\phi - \lambda(\phi^\dagger\phi)^2, \quad \text{where} \quad \lambda > 0, \mu^2 > 0\,. \tag{2.4}$$

Here D_μ is the covariant derivative and μ and λ are real numbers. As a result from the choice of sign for the mass term, the potential part of $\mathcal{L}_{\text{Higgs}}$ given by $V(\phi) = -\mu^2\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2$ no longer has a minimum at $\phi = 0$, but at any non-zero ϕ fulfilling $\phi^\dagger\phi = -\mu^2/2\lambda$ as seen in Figure 2.2. After spontaneous symmetry breaking and choosing the unitary gauge, the Higgs field can be written as

$$\phi = \begin{pmatrix} 0 \\ \frac{1}{\sqrt{2}}(v + h(x)) \end{pmatrix}\,, \tag{2.5}$$

where $\langle\phi^0\rangle = \mu/\sqrt{2\lambda} = v/\sqrt{2}$ is the non-zero vacuum expectation value (VEV) of the Higgs field and h is the physical Higgs boson field describing quantum fluctuations around the VEV. Substituting into Equation (2.4) yields the mass and interaction terms given by

$$\begin{aligned}\mathcal{L}_{\text{Higgs}} =& \frac{1}{2}\partial_\mu h\partial^\mu h - \frac{1}{2}(\underbrace{2\lambda v^2}_{m_H^2})h^2 + \frac{1}{2}\underbrace{\frac{g_1^2v^2}{2}}_{m_W^2}W_\mu^- W^{+\mu} + \frac{1}{2}\underbrace{\frac{v^2}{4(g_1^2+g_2^2)}}_{m_Z^2}Z_\mu Z^\mu \\ & \underbrace{-\frac{1}{4}\lambda h^4 - \lambda v h^3 + (vh + \frac{h^2}{2})(\frac{g_1^2}{2}W_\mu^- W^{+\mu} + \frac{Z_\mu Z^\mu}{4(g_1^2+g_2^2)})}_{\text{interaction terms of the Higgs boson with the } W, Z \text{ bosons and itself)}}\,.\end{aligned} \tag{2.6}$$

Here $W_\mu^\pm$ and Z_μ denote the gauge fields corresponding to the $W^\pm$ and Z bosons, respectively. The constants g_1 and g_2 are the gauge coupling constants of the $SU(2)_L$ and $U(1)_Y$ gauge groups.

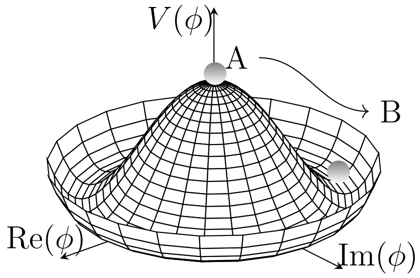


Figure 2.2.: Spontaneous symmetry breaking in the Higgs potential. Figure created using TikZ code adapted from J. Riebesell (MIT License).

2.2.2. Top-Higgs-Yukawa Coupling and CP Symmetry

The Yukawa interaction term $\mathcal{L}_{\text{Yukawa}}$ completes the $SU(2)_L \times U(1)_Y$ part of the Lagrangian, giving rise to fermionic mass terms and interaction terms of fermions with the Higgs field. The Lagrangian is given by

$$\mathcal{L}_{\text{Yukawa,f}} = -\frac{vg_f}{\sqrt{2}}(\bar{f}_L f_R + \bar{f}_R f_L) - \frac{g_f}{\sqrt{2}}(\bar{f}_L f_R + \bar{f}_R f_L)h \quad , \tag{2.7}$$

where f_L and f_R denote the left- and right-handed chiral components of the fermion field f (or anti-fermion field $\bar{f} = f^\dagger \gamma^0$ for $\bar{f}_L$ and $\bar{f}_R$). The strength of the Yukawa coupling is described by g_f and $m_f = \frac{vg_f}{\sqrt{2}}$ is the mass of the fermion. The resulting Lagrangian for the Yukawa coupling between the top quark and the Higgs boson in the SM is

$$\mathcal{L}_{tH-\text{Yukawa}} = -\frac{m_t}{v} t\bar{t}h \quad , \tag{2.8}$$

with $t\bar{t}$ representing the respective fermion fields. As such, the interaction is predicted by the SM to be symmetric under a combined charge conjugation parity (CP) transformation (CP-even). A modified Lagrangian for the coupling allowing CP-mixing (adding an antisymmetric or CP-odd component) is provided by the Higgs Characterisation (HC) model [38],

$$\mathcal{L}_{tH-\text{Yukawa,CP-mixing}} = -\frac{m_t}{v} \kappa_t \bar{t}(\cos(\alpha) + i\sin(\alpha)\gamma^5)th\,. \tag{2.9}$$

Here $\alpha \in (-\pi, \pi)$ is the mixing angle of the CP-even and CP-odd states and $\kappa_t \geq 0$ is a scaling parameter for the coupling strength [39]. For $\alpha = 0$ and $\kappa_t = 1$, Equation (2.9) reduces to the SM Lagrangian as written in Equation (2.8). In the case of a CP-mixed top-Higgs-Yukawa coupling, different cross sections for various processes involving the top-Higgs-Yukawa coupling would be obtained as shown in Table 2.1.

Table 2.1.: Next-to-Leading-Order (NLO) QCD predictions for the cross section of $t\bar{t}H$, $tHjb$ and tWH processes in pp collisions for different CP scenarios at $\sqrt{s} = 13.6\,\text{TeV}$ simulated with the HC model [40].

κ_t	α	Cross sections (pb)		
		ttH	$tHjb$	tWH
1	0 (SM, CP-even)	0.516	0.077	0.024
1	15	0.497	0.081	0.025
1	30	0.444	0.096	0.030
1	45 (CP max mixing)	0.371	0.125	0.039
1	60	0.299	0.172	0.051
1	75	0.247	0.242	0.066
1	90 (CP-odd)	0.227	0.333	0.085
-1	0	–	0.858	0.175
0.5	0	–	0.122	0.018
2	0	–	0.275	0.120
2	45	–	0.245	0.137

2.3. Higgs Boson Production and Decay

The production of a Higgs boson alongside other particles in proton-proton collisions at the LHC occurs through various processes. The cross sections predicted by the SM for processes involving the production of a Higgs boson as a function of the centre-of-mass energy $\sqrt{s}$ are shown in Figure 2.3.

At high-energy hadron colliders such as the LHC, the SM Higgs boson is prevalently produced through gluon-gluon fusion (ggF) mediated by a virtual quark loop with $\sigma_{\text{ggF}} = 51.66^{+4.7\%}_{-7.5\%}\,\text{pb}$ (N3LO QCD + NLO EW, NNLO PDF4LHC), vector boson fusion (VBF) with $\sigma_{\text{VBF}} = 4.109^{+2.5\%}_{-2.4\%}\,\text{pb}$ (N3LO QCD + NLO EW) and vector-boson-associated Higgs production with $\sigma_{\text{WH}} = 1.464^{+1.9\%}_{-2.0\%}\,\text{pb}$ and $\sigma_{\text{ZH}} = 0.947^{+3.3\%}_{-3.1\%}\,\text{pb}$ (both at NNLO QCD + NLO EW) as shown in Figure 2.4 [42–48]. In contrast to that, due to destructive interference between different amplitudes, the cross section for t-channel Higgs boson production along with a single top quark, is at a much lower level of only $\sigma_{tH/t-\text{ch.}} = 0.0853^{+6.7\%}_{-15.6\%}\,\text{pb}$ (NLO QCD) [42–44, 49, 50]. The corresponding $tHjb$ events, which are the focus of this analysis and have a final state consisting of a top quark, a Higgs boson, a bottom quark and at least one additional jet, have therefore not been observed yet. Furthermore, tWH (single-top Higgs production with a W boson) with $\sigma_{tWH} = 0.0174^{+6.0\%}_{-7.4\%}\,\text{pb}$ and $t\bar{t}H$ ($t\bar{t}$ pair production with a Higgs boson) with $\sigma_{t\bar{t}H} = 0.592^{+3.3\%}_{-3.7\%}\,\text{pb}$ are considered as signal processes. All values were computed at $\sqrt{s} = 13.6\,\text{TeV}$ and the Higgs mass $m_H = 125.09\,\text{GeV}$. Feynman diagrams of the three signal processes and the remaining

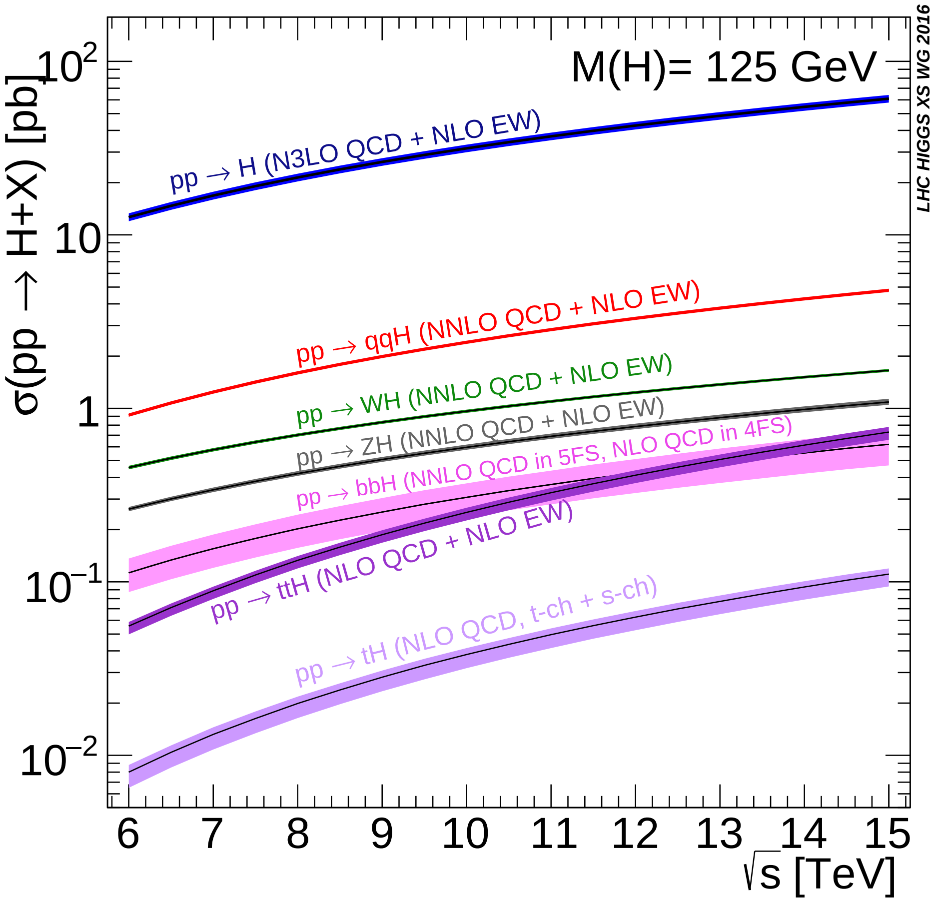


Figure 2.3.: The cross sections for SM processes involving the production of a Higgs boson in pp collisions as a function of $\sqrt{s}$ [41].

s-channel tH production are shown in Figures 2.5 and 2.6.

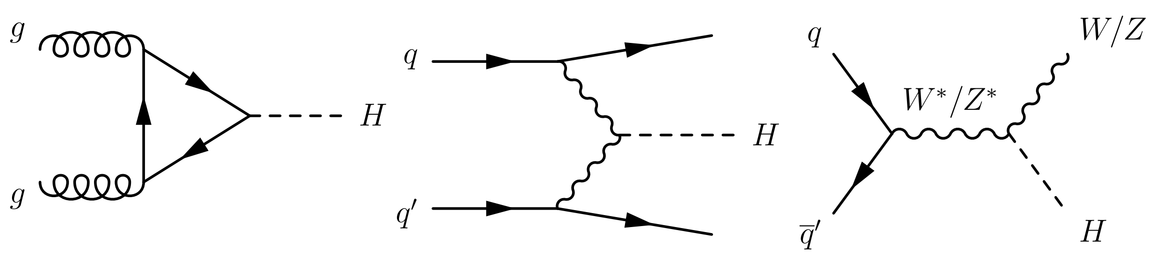
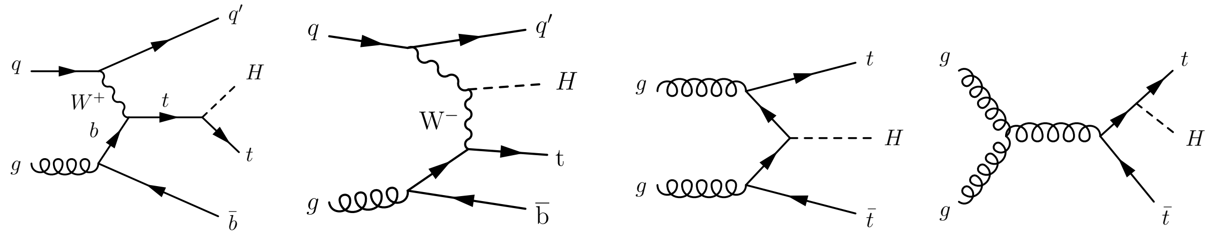
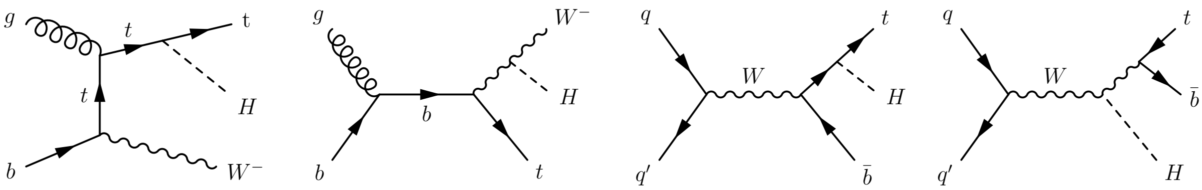


Figure 2.4.: Representative Feynman diagrams for the three dominant Higgs boson production processes: gluon fusion (ggF), weak vector-boson fusion (VBF) and weak vector-boson-associated Higgs production (VH).



(a) t-channel tH / $tHqb$ (b) $t\bar{t}H$

Figure 2.5.: Representative leading-order (LO) Feynman diagrams for (a) $tHqb$ production corresponding to the $tHjb$ signal considered in this analysis, and (b) $t\bar{t}H$ production.



(a) tWH (b) s-channel tH

Figure 2.6.: Representative leading-order Feynman diagrams for (a) tWH and (b) s-channel tH production.

A SM Higgs boson with a mass of 125 GeV has a total decay width of $\Gamma_H \simeq 4.1\,\text{MeV}$ which corresponds to a lifetime of about 1.6×10^{-22} s [51]. The most dominant decay channel is $H \to b\bar{b}$ with a branching ratio of $(5.82 \pm 0.10) \times 10^{-1}$. In the thesis, the $H \to \tau^+\tau^-$ decay with a branching ratio of $(6.27 \pm 0.15) \times 10^{-2}$ will be considered [51]. The tau leptons can each decay hadronically (τ_{had}) or leptonically (τ_{lep}) as shown in Figure 2.7, leading to three possible combinations in the final state of the Higgs boson. This analysis will focus on the full-hadronic decay $H \to \tau_{\text{had}}\tau_{\text{had}}$ which occurs in approximately 42% of all $\tau\tau$ decays [51].

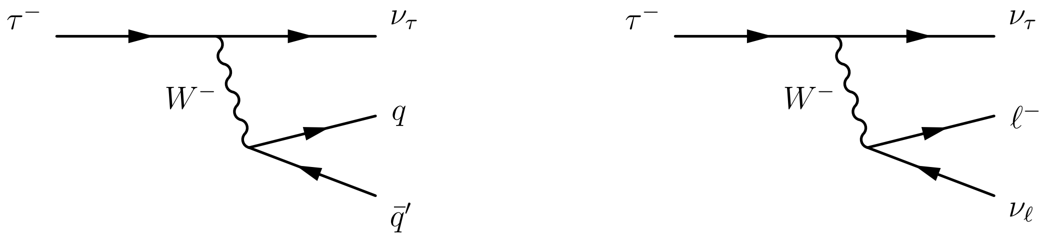


Figure 2.7.: Feynman diagrams for the leptonic (left) and hadronic (right) decay of a tau lepton.

3. Machine Learning

To address the challenge of distinguishing small signals from a large background, recent analysis routines employ supervised machine learning (ML) in training neural networks (NNs). Furthermore, the analysis employs boosted decision trees (BDTs) to distinguish different scenarios for the CP-symmetry of the top-Higgs-Yukawa interaction. Therefore, the basic concepts of these ML architectures are introduced in this Chapter.

3.1. Structure of a Neural Network

A single-layer neural network, the perceptron, was first designed by Rosenblatt in 1957 [52, 53]. In the perceptron, a set of inputs x_i is assigned weights w_i and then mapped to an output y by using an activation function σ,

$$y = \sigma(b + \sum_i w_i x_i)\,, \tag{3.1}$$

where the bias b assigned to the perceptron is functioning as a constant offset. The activation function introduces non-linearity to the neural network. Typical activation functions are, for instance, the sigmoid function

$$\sigma_s(x) = \frac{1}{1 + e^{-x}} \tag{3.2}$$

or the Rectified Linear Unit (ReLU) function $\mathrm{ReLU}(x) = \max\{0, x\}$[54]. The output layer of the perceptron is the only computational layer, so the computation is visible to the user. Multi-layer neural networks, on the other hand, contain multiple computational layers of neurons between the input and the output layer which are referred to as hidden layers. Successive layers feed into one another in the forward direction, meaning that the outputs x_i of the neurons in one layer are the inputs of the next layer, and the weights

w_{ij} describe the weight of input x_j to the output y_i of node i (with bias b_i), leading to

$$\begin{bmatrix} y_1 \\ ... \\ y_n \end{bmatrix} = \sigma \left(\begin{bmatrix} w_{11} & ... & w_{1m} \\ ... & & ... \\ w_{n1} & ... & w_{nm} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ ... \\ x_m \end{bmatrix} + \begin{bmatrix} b_1 \\ ... \\ b_n \end{bmatrix} \right) . \tag{3.3}$$

The output layer can also consist of multiple nodes.

3.2. Training and Validation

The training process of a neural network minimises the loss function L by optimising a set of learnable parameters θ (usually the weights and biases of the neurons). The function provides a mathematical description of the deviation between a model's prediction and the known true values for the training data. There are many types of loss functions suited for different tasks. Existing analysis routines on this topic often use the Binary Cross-Entropy (BCE) function defined as

$$L = \sum_n -w_n[y_n \cdot \ln(p_n) + (1 - y_n) \cdot \ln(1 - p_n)] \quad . \tag{3.4}$$

Here, w_n is the weight of event n, y_n is the true binary (0 or 1) class of the event also known as the target output, and p_n is the prediction of the network mapped to the range (0, 1) through an activation function. In the forward phase of the training, the model is given training data as input and calculates the output, out of which the loss is calculated by comparing to the known results. In the backward phase, the gradient $\vec{\nabla}_\theta$ of the loss with respect to the set of learnable parameters θ is calculated in the backward direction (from output to input layer) applying the chain rule of differential calculus. The result provides a local optimisation of the parameters, which is used to update the learnable parameters, which is, in the simplest case, done according to

$$\theta_{k+1} = \theta_k - \eta \cdot \partial_\theta L \quad , \tag{3.5}$$

where η is the learning rate, a hyperparameter of the model. A modified version for updating the parameters is used by the Adam optimiser [55] and given by

$$\theta_{k+1} = \theta_k - \eta \left(\frac{\sqrt{1 - \beta_2^k}}{1 - \beta_1^k} \cdot \frac{m_k}{\sqrt{v_k}} \right) . \tag{3.6}$$

Here β_1 and β_2 are hyperparameters known as decay rates and m_k,v_k are defined as

$$m_k = \beta_1 m_{k-1} + (1 - \beta_1)\partial_\theta L \qquad \text{and} \qquad v_k = \beta_2 v_{k-1} + (1 - \beta_2)(\partial_\theta L)^2. \tag{3.7}$$

This strategy is known as backpropagation or gradient descent [56]. With the updated parameters, the procedure is then repeated several times until the iterations converge towards a minimum in L.

The training set is usually divided into several batches, on which the model is consecutively trained. One training epoch is defined as a pass through the entire training data. Training a model for too many epochs can lead to overfitting: it can cause the model to become overly specialised to the training data, leading to poorer performance on unseen data. To mitigate the issue, the original data set is usually split into three samples: training, validation and test data. Throughout the training process, the model is fitted exclusively to the training data. After each epoch, the loss function is evaluated on the training and validation sets. As training progresses, both losses generally decrease, however, at a certain point, the loss on the validation data begins to increase again while the loss on training data continues to decrease. This divergence indicates the onset of overfitting, and training should be stopped at this stage to preserve the model's generalisation performance. The final model's performance is then evaluated on the test data [54].

To mitigate the issue of low statistics, k-fold cross validation, allowing for the usage of all events in training, is applied in this analysis. The dataset is split into $k = 4$ statistically independent subsets of equal size (*blocks* B0-B3) by generating an integer in the range $[0, k - 1]$ with the `TRandomRanluxpp` generator from the ROOT framework, initialised with a seed given by the sum of the event number and a fixed integer offset. The offset avoids potential issues with variables based on algorithms that use the event number as a seed for the training of underlying models, such as the b-tagger. Of the k blocks, one is retained as the test set, one is used as the validation set and the remaining two form the training set. Using each block as the test set exactly once, k separate *folds* of the model are trained. Each fold is then evaluated on its respective test set. A schematic overview of the k-fold cross validation for $k = 4$ is provided in Figure 3.1. When applying the model, each event is assigned to fold m using the same procedure that was employed for the assignment to blocks described above.

	Fold 0	Fold 1	Fold 2	Fold 3
B0	Test	Validation	Training	Training
B1	Training	Test	Validation	Training
B2	Training	Training	Test	Validation
B3	Validation	Training	Training	Test

Figure 3.1.: Schematic overview of the employed k-fold cross validation procedure with $k = 4$. Once all four folds are trained, the complete dataset is obtained by combining all four testing sets.

3.3. Dropout and Layer Normalisation

In the training of an NN, the use of regularisation techniques is needed to prevent overfitting. This can be done using dropout in certain layers, which are then called dropout layers. In each training step, a probability that randomly selected nodes are dropped, which means they do not get any inputs and thereby do not produce any outputs, is introduced. This probability is independent of previous dropouts, so the dropout can also affect multiple layers back-to-back. This stops the network from focusing on only some aspects, as it is forced to use all of its nodes to keep the loss low even if important nodes drop out. The dropout is only used in training as the network should be able to use its entire architecture in evaluation, testing and application.

Layer normalisation is a strategy that stabilises and accelerates the training, as it ensures the activations of each layer stay within a stable range, thereby smoothing gradients. It normalises all input features in a layer for each node according to

$$x' = \frac{x - \hat{x}}{\mathrm{var}(x)}\gamma + \beta \quad , \tag{3.8}$$

where the gain γ and bias β are learnable parameters, x is the input feature with the mean value $\hat{x}$ and variance $\mathrm{var}(x)$.

3.4. Hypergraph Neural Networks

In contrast to traditional neural networks described previously, graph-based convolutional neural networks (GNNs) can encode graph-structured data and show superiority on representation learning due to their ability to employ pairwise connections between data [57, 58]. A graph is defined as a pair $G = (V, E)$ with V being the set of nodes and E the set of unordered pairs $\{v_1, v_2\}$ of $v_1, v_2 \in V$ whose elements are called edges. While the degree (number of connected nodes) for all edges is fixed to be two in a graph, a hypergraph can encode higher-order data correlations through degree-free hyperedges (that can connect more than two nodes). A hypergraph NN (HGNN) handles these higher-order correlations in the hypergraph-structured input data through hyperedge convolution operations as described in Reference [59]. The analysis employs a trained HGNN for event classification as described in Section 6.3 that was developed as part of an internship [60] following up on a BSc thesis [61], further explanations are found in the given references.

3.5. Gradient Tree Boosting

A tree ensemble model uses the idea that many weaker models (in this case, binary decision trees) can form a stronger predictor when combined. Given a data set $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}$ ($|\mathcal{D}| = n$ and $y_i \in \mathbb{R}$) with n examples $\boldsymbol{x}_i \in \mathbb{R}^m$ and m features, the tree ensemble model predicts the output from K additive functions:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(\boldsymbol{x}_i), \qquad f_k \in \mathcal{F} \tag{3.9}$$

where $\mathcal{F} = \{f(\boldsymbol{x}) = w_{q(\boldsymbol{x})}\}(q : \mathbb{R}^m \to \{1, ...T\}, w \in \mathbb{R}^T)$ denotes the space of Classification and Regression Trees (CART). Each f_k is an independent tree structure q with T leaves, which are the terminal output nodes of the tree structure, and weight w_i representing the score on the i-th leaf. Gradient boosting builds the ensemble of trees iteratively. Let $\hat{y}_i^{(t)}$ be the prediction of the i-th instance at the t-th iteration. At boosting iteration t, a new tree f_t is added in order to minimise the regularised loss function

$$\mathcal{L}^{(t)} = \sum_i l(y_i, \hat{y}_i^{(t-1)}) + f_t(\boldsymbol{x}_i) + \Omega(f_t) \qquad \text{where} \qquad \Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2 . \tag{3.10}$$

Here l is a differentiable convex loss function that evaluates the accuracy of the prediction and $\Omega(f_t)$ penalises the complexity (many leaves or large leaf weights) of f_t [62]. Second-order approximation of the loss function around the current prediction $\hat{y}_i^{(t-1)}$ yields

$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n} [l(y_i, \hat{y}^{(t-1)}) + g_i f_t(\boldsymbol{x}_i) + \frac{1}{2} h_i f_t^2(\boldsymbol{x}_i)] + \Omega(f_t) \tag{3.11}$$

where $g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ and $h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$ are first- and second-order gradients of l w.r.t. the current prediction. Let $I_j = \{i | q(\boldsymbol{x}_i) = j\}$ be the set of training instances of leaf j. Removing the constant terms, substituting $f_t(\boldsymbol{x}_i) = w_j$ for all $i \in I_j$

and expanding $\Omega = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$, a simplified objective at step t

$$\tilde{\mathcal{L}}^{(t)} = \sum_{j=1}^{T} [(\sum_{i \in I_j} g_i) w_j + \frac{1}{2}(\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T \tag{3.12}$$

is obtained. Substituting $f_t(\boldsymbol{x}_i) = w_j$ for all $i \in I_j$ yields a quadratic function of the leaf weights w_j for the objective. Minimising the expression w.r.t. w_j gives the optimal weight w_j^* of leaf j

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \,. \tag{3.13}$$

Substituting these optimal weights back into the objective gives a scoring function $\tilde{\mathcal{L}}^{(t)}(q)$ to measure the quality of a tree structure q by

$$\tilde{\mathcal{L}}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{(\sum_{i \in I_j} g_i)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \,. \tag{3.14}$$

As it is usually not feasible to consider all possible tree structures q, the tree is grown from a single leaf using an iterative greedy algorithm. At each iteration, the algorithm considers candidate splits, i.e. binary partitions dividing the instances I of a node into left and right child nodes with instance sets I_L and I_R, where $I = I_L \cup I_R$. The loss reduction obtained by introducing the split is given by

$$\mathcal{L}_{\text{split}} = \frac{1}{2} \left[\frac{(\sum_{i \in I_L} g_i)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{(\sum_{i \in I_R} g_i)^2}{\sum_{i \in I_R} h_i + \lambda} - \frac{(\sum_{i \in I} g_i)^2}{\sum_{i \in I} h_i + \lambda} \right] - \gamma \,. \tag{3.15}$$

Among all possible split candidates, the split with the largest loss reduction is selected and the procedure is repeated recursively until a stopping criterion is reached [63].

4. Experimental Setup

Understanding the process of data acquisition is a necessary prerequisite to performing data analysis in any field, and in particular for complex processes such as those that occur at particle collider experiments, where a great number of particles produced during a collision have to be identified, and their properties measured as accurately as possible. Therefore, an overview of the Large Hadron Collider (LHC) as well as a description of the ATLAS detector and in particular the components most relevant to the thesis' subject is provided in this Chapter.

4.1. The Large Hadron Collider

The Large Hadron Collider (LHC) at CERN is the world's largest and highest-energy proton-proton collider with a circumference of about 27 km and a centre-of-mass energy of 13.6 TeV during the latest data taking period from 2022 to 2026 (Run 3). Through a system of dipole, quadrupole and higher multipole magnets and radiofrequency (RF) cavities, two bunched proton beams are focused and kept on a bent trajectory (travelling in opposite directions). At four primary collision points, the two beams are crossed at a nominal rate of 40 MHz corresponding to a bunch crossing every 25 ns [64]. During the data-taking period from 2022 to 2025 (Run 3 until the present point), every crossing produced on average 56 independent pp collisions, depending on the exact data-taking conditions [65]. Tight proton bunching allows for high luminosities. The production rate of a process is given by the product of its cross section σ_p and the instantaneous luminosity $\mathcal{L}$ as

$$\frac{\mathrm{d}N_p}{\mathrm{d}t} = \sigma_p \mathcal{L} \quad , \tag{4.1}$$

where the time-integrated luminosity $L = \int \mathcal{L} dt$ similarly results in the total number of events produced for a given process as $N_p = \sigma_p L$. The four collision points correspond to the four major experiments at LHC: ATLAS (A Large Toroidal LHC ApparatuS) and CMS (Compact Muon Solenoid) as general-purpose detectors, ALICE (A Large Ion Collider Experiment) focusing on heavy-ion physics and LHCb (Large Hadron Collider beauty) investigating hadrons containing bottom quarks.

4.2. The ATLAS Detector System

The multi-purpose ATLAS detector is a combination of several concentric cylindrical layers that operate on different detection mechanisms to allow for track reconstruction and precise energy and momentum measurements of most particles and their decay products. Certain particles, most notably neutrinos, cannot be detected and are instead inferred from calculations on missing transverse momentum. It is nominally forward-backward symmetric with respect to the interaction point with a diameter of about 25 m and a length of 44 m [66]. A schematic view of the detector is shown in Figure 4.1.

A right-handed coordinate system with its origin defined at the nominal interaction point is used in ATLAS. The z-axis is defined along the beam direction, the x-axis points to the centre of the accelerator ring and the y-axis points upwards. The transverse momentum p_{T}, transverse energy E_{T} and missing transverse energy (MET) are defined in the x-y plane. Spherical coordinates are commonly used, where ϕ denotes the azimuthal angle around the beam axis and θ is the polar angle to the z axis which is commonly expressed in terms of the pseudorapidity η defined as

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right) = \frac{1}{2}\ln\left(\frac{\|\boldsymbol{p}\| + p_z}{\|\boldsymbol{p}\| - p_z}\right) \quad . \tag{4.2}$$

Here $\boldsymbol{p}$ is the particle's momentum with the component p_z along the beam axis. At high energies, the pseudorapidity is approximately equal to the rapidity y given by

$$y = \frac{1}{2}\ln\left(\frac{E + p_z}{E - p_z}\right) \quad . \tag{4.3}$$

The difference between rapidities is invariant under Lorentz boosts along the beam axis.

As evident from its name, the ATLAS detector's magnetic system consists of one barrel and two end-cap large superconducting toroid magnets arranged in an eight-fold azimuthal symmetry and providing a magnetic field of 3.5 T field strength, used to deflect muons in the outer part of the detector, the muon chambers. Furthermore, a system of four thin superconducting solenoids provides a nearly homogeneous magnetic field in the z-direction of up to 2 T in the Inner Detector [66, 68]. The field causes charged particles to follow curved trajectories through the influence of the Lorentz force, directly relating the particle's momentum with its charge and the radius of its trajectory [69]. As a highly energetic incoming particle creates electron-hole pairs in a semiconductor by ionisation, or

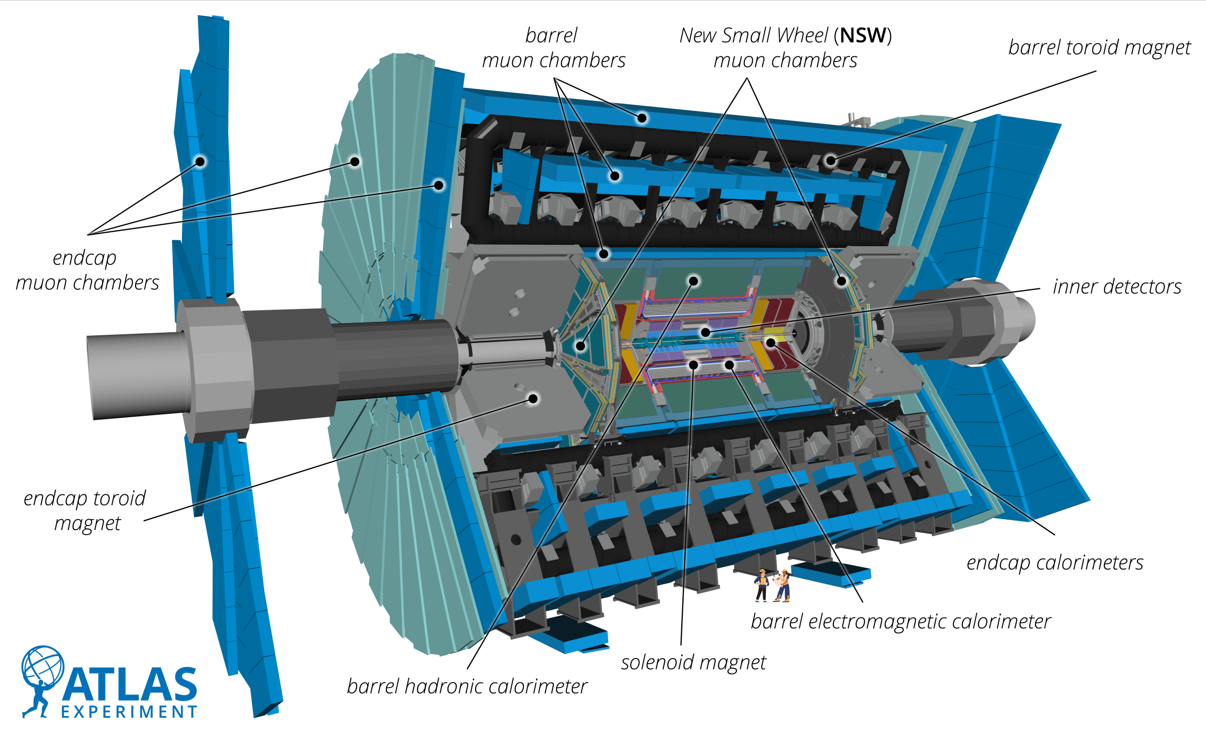


Figure 4.1.: A schematic illustration of the Atlas detector system, with human figures to scale [67].

electron-ion pairs in a gas, respectively, the resulting charge is collected through electrodes and measured as a signal providing information about the particle's trajectory (track).

Around the Inner Detector follows the calorimeter system designed to measure the energy of particles. The electromagnetic calorimeter is sensitive to photons, electrons and positrons. The incoming highly energetic particle creates a cascade of particles through Bremsstrahlung (emission of a photon by an $e^{\pm}$) and pair creation ($\gamma \rightarrow e^{\pm}$ in the field of a nucleus), called an electromagnetic (EM) shower. The number of particles grows until the average energy of the particles reaches the critical energy, below which ionisation becomes more likely. The induced signal is proportional to the number of particles in the shower maximum, which in turn directly correlates with the energy of the incoming particle. Hadronic showers follow the same principle: the hadrons interact through the strong force and new hadrons are created. They are more complex since the hadrons also induce EM showers overlapping with the hadronic part.

Muons are 200 times heavier than electrons, and therefore emit much lower levels of Bremsstrahlung. They are therefore, along with possible leakage from the calorimeters, the only particles detected in the outermost part of the detector, the muon chambers. Due to the toroidal field geometry of the outer magnetic field, combining the tracks in the muon chambers with the backwards-reconstructed tracks of muons in the Inner Detector can also function as a method to correct errors in the track reconstruction that are caused by imperfections in the cylindrical detector geometry [70].

4.2.1. The Inner Detector

The Inner Detector (ID) consists of three subsystems and is sensitive to particles with $|\eta| < 2.5$ [66, 71–73]. Directly around the beampipe are layers of finely segmented Silicon (Si) Pixel Detectors, achieving a spacepoint resolution of 10 µm and thereby making high-precision momentum measurements possible. Surrounding the Pixel Detector is the Semiconductor Tracker (SCT) which consists of Si microstrip modules and further tracks charged particles with a spacepoint resolution of 17 µm. The Transition Radiation Tracker (TRT), the ID's outermost part, consists of drift tubes filled with a Xenon-based gas. There, a charged particle ionises the gas atoms by transferring enough energy for a shell electron to escape the atomic potential and ionise other gas atoms, a process called charge avalanche, creating a signal collected at the spatially separated readout electrodes. The interspace between drift tubes consists of polyethylene. If a relativistic particle above a specific threshold for $\gamma = E/m$ crosses the interface, transition radiation is emitted, providing a way to distinguish particles that would leave identical traces in other detector parts, such as electrons and charged pions for energies in the range of $1 - 100$ GeV [74, 75].

4.2.2. The Calorimeter System

Enclosing the ID, the calorimeter system measures the energy of particles within the range of $|\eta| < 4.9$. It consists of two nested calorimeters: the inner electromagnetic calorimeter (ECAL) system based on Liquid Argon (LAr) and the outer hadronic calorimeter (HCAL) system, which uses scintillation tiles in addition to LAr. Both are designed as sampling calorimeters, a type of calorimeter with alternating layers of distinct active and passive media. The active layers measure the deposited energy through ionisation (LAr) or scintillation (tiles), whereas the shower is developed further in each layer of the passive material [76, 77].

Electromagnetic Calorimeter System

The ECAL system consists of a barrel region covering $|\eta| < 1.475$, two endcap regions on each side ($1.375 < |\eta| < 3.2$) and the Copper/LAr Forward Calorimeter ($3.1 < |\eta| < 4.9$).

A LAr presampler ($|\eta| < 1.8$) measures the energy loss before the ECAL, which is relevant for calibration. An accordion geometry of the absorber material and the LAr provides full 2π coverage in the azimuthal angle ϕ. Lead is used as the absorber material in all parts [66]. The ECAL measures energies of photons and electrons through their particle showers.

Hadronic Calorimeter System

The central part of the HCAL system detector is the barrel region divided into a central barrel part ($|\eta| < 1.0$) and two extended barrel parts ($0.8 < |\eta| < 1.7$). It uses scintillating tiles to detect hadronic showers and absorbent steel as the passive medium. Furthermore, there are two Copper-Tungsten/LAr endcap calorimeters ($1.5 < |\eta| < 3.2$) and the previously described Copper/LAr Forward Calorimeter ($3.1 < |\eta| < 4.9$) [66].

4.2.3. The Muon Spectrometer

Muons are minimum-ionising particles and therefore only deposit a small fraction of their energy in the calorimeters. The Muon Spectrometer (MS) is the outermost layer of the ATLAS detector and is responsible for high-precision measurements of the trajectory (track) of muons, and thus their momentum and charge, in about 4,000 individual muon chambers. The magnetic field caused by the toroidal magnets surrounding the barrel region [78] and two toroidal end-cap systems [79] deflects the muons according to the Lorentz force. This enables track measurements in a region of $|\eta| < 2.7$, with Monitored Drift Tubes (MDT) performing momentum measurements for $|\eta| < 2.0$ and Cathode Strip Chambers (CSC) covering the remaining range. The New Small Wheel (NSW) system installed after the end of Run 1 of the LHC contributes to the muon tracking in the range of $1.3 < |\eta| < 2.7$ and reduces the acceptance of fakes [80]. Other subsystems of the MS provide further information for L1 trigger decisions.

4.2.4. Trigger and Data Acquisition

The nominal bunch crossing rate of 40 MHz at the LHC leads to data production at a scale that vastly exceeds available storage capacity. However, since the vast majority of processes have been studied in great detail and are not of particular scientific interest, the amount of data that needs to be saved for further analysis can be reduced greatly through a two-level trigger system. The Level 1 (L1) trigger searches for relevant event signatures such as hadronically decaying tau leptons, using information from a subset of detectors and reducing the event rate to approximately 100 kHz. The software-based High Level

Trigger (HLT) is then seeded by Regions-of-Interest, as provided by the L1 trigger, and applies an event filter, selecting event data at a rate of approximately 3 kHz to be stored for further analysis [81].

5. Data and Simulated Event Samples

Simulated Monte Carlo (MC) event samples are used to model the $tHjb$, tWH and $t\bar{t}H$ ($H \to \tau_{\text{had.}}\tau_{\text{had.}}$) signal processes and the SM background contributions. Additional signal samples are modelled with different CP-mixing scenarios of the top-Higgs coupling as described by the HC model in Equation (2.9). The simulations correspond to the ATLAS detector data taking period from 2022 to 2024 in pp collisions at an integrated luminosity of approximately 161 fb^{-1} with 1.9% uncertainty [82] and 13.6 TeV centre-of-mass energy [83]. The generator configurations are summarised in Table 5.1.

Table 5.1.: Generator configurations used in the production of signal and background samples with the mass of the top quark and SM Higgs boson set to 172.5 GeV and 125 GeV, respectively. Matrix element (ME) order refers to the perturbative order in the strong coupling constant α_s. Tune denotes the parton shower tune, either A14 with the NNPDF2.3LO set for PYTHIA 8 or the default tune for SHERPA. Leading-logarithm photon emission is included in all samples and either modelled by the parton shower generator or by PHOTOS [84].

Physics process	**Event generator**	**ME order**	**Parton shower**	**PDF (ME)**	**Tune**
$t\bar{t}H$	POWHEG BOX v06-03	NLO	PYTHIA 8	NNPDF3.0NLO	A14 NNPDF2.3LO
$tHjb$	MADGRAPH5_AMC@NLO	NLO	PYTHIA 8	NNPDF3.0NNLO	A14 NNPDF2.3LO
tWH	MADGRAPH5_AMC@NLO	NLO	PYTHIA 8	NNPDF3.0NNLO	A14 NNPDF2.3LO
CP-mixed samples	MADGRAPH5_AMC@NLO	NLO	PYTHIA 8	NNPDF3.0NLO	A14 NNPDF2.3LO
$t\bar{t}$	POWHEG BOX v2	NLO	PYTHIA 8	NNPDF3.0NLO	A14 NNPDF2.3LO
tW	POWHEG BOX v2	NLO	PYTHIA 8	NNPDF3.0NLO	A14 NNPDF2.3LO
Single top (s-, t-ch.)	POWHEG BOX v2	NLO	PYTHIA 8	NNPDF3.0NLO	A14 NNPDF2.3LO
$t\bar{t}Z/\gamma^*$	MADGRAPH5_AMC@NLO	NLO	PYTHIA 8	NNPDF3.0NLO	A14 NNPDF2.3LO
$t\bar{t}W$	SHERPA 2.2.14	MEPS@NLO	SHERPA	NNPDF3.0NNLO	SHERPA
ggF H production	POWHEG BOX v06-03	NNLOPS	PYTHIA 8	NNPDF3.0NLO	A14 NNPDF2.3LO
VBF H production	POWHEG BOX v06-03	NLO	PYTHIA 8	NNPDF3.0NLO	A14 NNPDF2.3LO
$qq \to WH/ZH$, $gg \to ZH$	POWHEG BOX v06-03	NLO	PYTHIA 8	NNPDF3.0NLO	A14 NNPDF2.3LO
V + jets	SHERPA 2.2.14	MEPS@NLO	SHERPA	NNPDF3.0NNLO	SHERPA
Diboson (VV)	SHERPA 2.2.14	MEPS@NLO	SHERPA	NNPDF3.0NNLO	SHERPA
$gg \to ZZ \to 4\ell$	POWHEG BOX	NLO	PYTHIA 8	NNPDF3.0NLO	A14 NNPDF2.3LO

The SM $t\bar{t}H$ samples are modelled at NLO QCD in the five-flavour scheme with POWHEG BOX v.06-03 [85–89] with the NNPDF3.0NLO parton density function (PDF) set [90]. The renormalisation and factorisation scales are set to $\sqrt[3]{m_{\mathrm{T}}(t) \cdot m_{\mathrm{T}}(\bar{t}) \cdot m_{\mathrm{T}}(H)}$. The events are interfaced to Pythia 8.310 [91] for parton showering, hadronisation and the underlying event with the A14 tune [92] and NNPDF2.3LO PDF set [90]. The SM tWH and $tHjb$ samples are produced by Madgraph5_aMC@NLO with the CT10 PDF set [93] and interfaced to Pythia 8.309 with the A14 tune [92]. Signal samples for the CP-mixed hypotheses are generated with Madgraph5_aMC@NLO and interfaced to Pythia 8.308.

Jet-associated Vector boson production (V + jets) including the dominant $Z \to \tau\tau$ background is modelled using the Sherpa 2.2.14 [94] generator with the MEPS@NLO prescription [95–98]. LO MEs for up to four partons and NLO MEs for up to two partons are calculated with the Comix [99] and OpenLoops 2 [100–102] libraries. The NNPDF3.0NNLO PDF set [90] is used and the samples are normalised to the NNLO prediction [103]. Top quark backgrounds are modelled at NLO QCD in the five-flavour scheme with the damping parameter $h_{\text{damp}} = 1.5\ m_t$ [104] with Powheg Box v2 [85–88] and interfaced to Pythia 8.308 [91]. Additional backgrounds from $t\bar{t}V$, diboson production and other Higgs boson production modes (ggF/VBF) are generated as listed in Table 5.1.

All MC samples produced with Powheg Box and Madgraph5_aMC@NLO are interfaced to EvtGen v2 [105] for the decays of bottom and charm hadrons. Additional pp collisions (pile-up) are generated with Pythia 8.186 [106] (A3 tune [107], MSTW2008LO PDF set [108]) and overlaid onto the hard scattering. All samples are processed through the standard Atlas detector simulation [109] based on Geant4 [110], and a subset further employs the Atlas fast calorimeter simulation. A unique Dataset Identifier (DSID) is assigned to each sample. The samples are split into the MC23a, MC23d and MC23e production campaigns, simulated with the detector and pile-up conditions present at Atlas data-taking in 2022, 2023 and 2024. The event weights are defined as

$$w_{\text{event}} = \frac{w_{\text{mc}} \cdot w_{\text{pileup}} \cdot S_{\text{JVT}} \cdot S_{\text{flavour-tag}} S_{\text{ID}(\tau_1)} \cdot S_{\text{ID}(\tau_2)} \cdot L \cdot \sigma}{\sum_i w_{\text{mc},i}} \ , \tag{5.1}$$

where w_{mc} is the generator-level MC statistics weight, w_{pileup} accounts for pile-up effects, S_{JVT} and $S_{\text{flavour-tag}}$ are scale factors for the jet vertex tagger and flavour tagger, $S_{\text{ID}(\tau_k)}$ are scale factors derived for the τ lepton identification, L is the integrated luminosity of the simulated data-taking period and σ the cross section of the simulated process. Truth-level information from TopCPToolkit [111] is used for the $t\bar{t}H$ samples. At the time of this analysis, the $tHjb$ and tWH samples with $\alpha = 45°$ are only available for the MC23a and MC23d campaigns. The DSIDs of all analysed samples are listed in Appendix A.

6. Object Definition and Event Selection

To reconstruct an event from the signals measured by the ATLAS detector, object definitions that specify how the analysis attributes signals to physics objects are given in Section 6.1. To focus on events with signatures matching those expected of the primary signal process ($tHjb$) and corresponding decay channels of interest, preselection criteria are applied as defined in Section 6.2. The categorisation of events into two signal ($tHjb$ and $t\bar{t}H$) and two main background ($Z \to \tau\tau$ and dileptonic $t\bar{t}$ or tW) processes using a HGNN is described in Section 6.3.

6.1. Object Definitions

Dedicated algorithms perform the reconstruction of physics objects such as electrons or jets from the measured energy deposits in the calorimeters and individual hits in the tracking detectors and the muon spectrometer. The resulting information for each object is a reconstructed four-momentum (requiring dedicated calibrations) and additional properties, e.g. the charge. In the following text, the terms *leading* and *sub-leading* refer to the object with the highest and second-highest transverse momentum.

The flight trajectories of charged particles traversing the Inner Detector are reconstructed as *tracks* from individual hits in the silicon pixel and strip detectors, as well as drift circles in the Transition Radiation Detector. To identify the location of the hard-scattering interaction, the primary vertex is reconstructed from at least two tracks and defined as the vertex with the highest sum of p_{T}^2 of the associated tracks.

Due to the colour confinement imposed by QCD, described by the large value of α_s at low energy scales, coloured particles such as quarks and gluons hadronise and produce collimated streams of colour-neutral hadrons. These *jets* are reconstructed by applying the anti-k_t algorithm [112, 113] with a jet radius $R = 0.4$ to topological energy clusters in the calorimeter and tracks processed with a particle-flow algorithm [114]. The jet energy is corrected for noise threshold effects, non-compensating calorimeter response, pile-up contamination and energy loss from the inactive material due to the design as a sampling calorimeter [115]. Only jets with transverse momentum $p_{\mathrm{T}} \geq 20\,\mathrm{GeV}$ and pseudorapidity $|\eta| < 4.5$ are considered. Jet vertex tagger (JVT) algorithms are employed to remove jets with $p_{\mathrm{T}} < 60\,\mathrm{GeV}$ identified as not associated with the primary vertex [116, 117].

Flavour tagging separates other hadronic jets (such as light-jets or hadronic τ-lepton decays) from b-jets containing b-hadrons and is of special importance for top quark analyses as top quarks predominantly decay into a W boson and a b-quark. The tagging of b-jets relies on the high mass, high decay multiplicity and long lifetime ($\tau \sim 10^{-12}\,\mathrm{s}$) of b-hadrons, often causing a displaced secondary vertex. The transformer-based GN2v01 flavour tagger [118] is employed with a working point (WP) of 85%, where b-jets are positively identified as such with 85% efficiency as measured in $t\bar{t}$ events. The *bachelor jet* is defined as the leading non-b-tagged jet that is not matched to the best top quark candidate as determined by the HGNN employed in this analysis for multiclass event categorisation into signal and background processes. Further details are found in Section 6.3.

Hadronically decaying τ-leptons produce a neutrino and visible decay products ($\tau_{\mathrm{had-vis}}$). Their reconstruction is seeded by jets reconstructed from topological clusters in the calorimeter deposits employing the anti-k_t algorithm with $R = 0.4$ [112, 113]. A recursive NN (RNN) algorithm is used to identify tracks originating from a $\tau_{\mathrm{had-vis}}$. To reject $\tau_{\mathrm{had-vis}}$ candidates composed of quark- or gluon-induced jets, the transformer-based GNN identification algorithm GNTau that combines track, cluster and jet information with global jet properties is employed with a medium WP [119]. Additionally, the number of τ-leptons that also pass the tight WP is stored. Candidates $\tau_{\mathrm{had-vis}}$ are required to fulfil $p_{\mathrm{T}} \geq 25\,\mathrm{GeV}$ and $|\eta| < 2.47$.

Electrons are reconstructed from matching ID tracks and ECAL energy deposits [120] and identified with a medium WP employing a likelihood criterion based on track properties and calorimeter shower shapes [120, 121]. Electrons with $p_{\mathrm{T}} \geq 15\,\mathrm{GeV}$ and $|\eta_{\mathrm{cluster}}| < 2.47$ are retained and a loose isolation criterion is required. Muons are reconstructed through a combined fit of ID and MS tracks. They are required to satisfy $p_{\mathrm{T}} \geq 10\,\mathrm{GeV}$, $|\eta| < 2.47$, pass a loose identification and tight isolation with a variable isolation radius [122].

To resolve the ambiguity of reconstructed objects sharing some of their constituent tracks or calorimeter clusters, an overlap removal procedure is applied to the objects defined above using the AssociationUtils package with the standard WP as described in Reference [123].

The total transverse momentum is conserved and approximately zero before the central pp collision. The missing transverse momentum $\boldsymbol{E}_{\mathrm{T}}^{\mathrm{miss}}$ contains information about neutrinos and other non-detectable particles and is defined as the negative vector sum of the transverse momenta of leptons, jets, $\tau_{\mathrm{had-vis}}$ and a soft term summing all $\boldsymbol{p}_{\mathrm{T}}$ of tracks matched to the primary vertex but not to any of the previous objects [124]. Its magnitude is referred to as the missing transverse energy $E_{\mathrm{T}}^{\mathrm{miss}}$, for $E_{\mathrm{T}}^{\mathrm{miss}} \neq 0$ the azimuthal angle ϕ_{MET} is defined as shown in Equation (6.1).

$$\phi_{\mathrm{MET}} = \begin{cases} \arccos\left(\frac{p_{x,\mathrm{miss}}}{E_{\mathrm{T}}^{\mathrm{miss}}}\right) & \text{if } p_{y,\mathrm{miss}} \geq 0 \\ -\arccos\left(\frac{p_{x,\mathrm{miss}}}{E_{\mathrm{T}}^{\mathrm{miss}}}\right) & \text{if } p_{y,\mathrm{miss}} < 0 \end{cases} \tag{6.1}$$

A summary of the applied object definitions is shown in Table 6.1.

Table 6.1.: Summary of the object definitions adapted in the analysis.

object	**definition**
jets	anti-k_t algorithm ($R = 0.4$) + particle-flow algorithm
	$p_{\mathrm{T}} \geq 20\,\mathrm{GeV}$ and $\lvert\eta\rvert < 4.5$
	pass JVT
b-tagged jets	all requirements for jets
	pass GN2v01 flavour tagger at 85% WP
bachelor jet	leading non-b-tagged jet
	that is not matched to the best top quark candidate
$\tau_{\mathrm{had-vis}}$	anti-k_t algorithm ($R = 0.4$)
	RNN for track identification
	identification with GNTau at medium WP
	$p_{\mathrm{T}} \geq 15\,\mathrm{GeV}$ and $\lvert\eta\rvert < 2.47$
electron	medium identification, loose isolation
	$p_{\mathrm{T}} \geq 15\,\mathrm{GeV}$ and $\lvert\eta_{\mathrm{cluster}}\rvert < 2.47$
muon	loose identification, tight isolation
	$p_{\mathrm{T}} \geq 10\,\mathrm{GeV}$ and $\lvert\eta\rvert < 2.47$

6.2. Event Selection

Preselected events must have a reconstructed primary vertex with at least two associated tracks with transverse momenta $p_{\mathrm{T}} \geq 500\,\mathrm{MeV}$, ensuring a high likelihood for a hard parton interaction. The ATLAS trigger system for LHC Run 3 is employed [81]. Preselected data events need to be part of the ATLAS good run list (GRL) files that record luminosity blocks (of one minute length) where all parts of the ATLAS detector were operational and the collected data is deemed usable for analyses.

To isolate kinematic regions with enhanced signal and simultaneously reject background events, the preselection further imposes topological and kinematic cuts on the reconstructed objects. As the analysis focuses on tH and $t\bar{t}H$ events with $H \to \tau\tau$ in the full-hadronic decay channel, visible decay products $\tau_{\mathrm{had-vis}}$ of two τ-leptons of opposite charge and no additional leptons are required. The Higgs boson candidate is formed by the vector momentum sum of the visible decay products $\tau_{\mathrm{had-vis}}$ and $\boldsymbol{E}_{\mathrm{T}}^{\mathrm{miss}}$. Its four-momentum is calculated using the likelihood-based Missing Mass Calculator (MMC) technique [125].

Both reconstructed visible τ-lepton candidates $\tau_{\mathrm{had-vis}}$ need to fulfil $p_{\mathrm{T}} \geq 30\,\mathrm{GeV}$, with the leading $\tau_{\mathrm{had-vis}}$ passing a stricter cut of $p_{\mathrm{T}} \geq 40\,\mathrm{GeV}$. The azimuthal distance is required to fulfil $0.6 \leq \Delta R_{\tau_{\mathrm{had-vis}}\tau_{\mathrm{had-vis}}} \leq 3.2$. At least one b-tagged jet and in total at least 3 jets (from the hadronically decaying top quark) are required. All jets need to satisfy $p_{\mathrm{T}} \geq 25\,\mathrm{GeV}$. Furthermore, the existence of a jet with $|\eta| < 3.2$ and $p_{\mathrm{T}} \geq 70\,\mathrm{GeV}$ is required due to a corresponding L1 requirement of the di-τ trigger employed in the analysis [81]. The fraction of the τ-lepton's momentum carried by its visible decay products, computed with the $\boldsymbol{E}_{\mathrm{T}}^{\mathrm{miss}}$ components in the collinear approximation and defined as x_1 and x_2 for the leading and sub-leading $\tau_{\mathrm{had-vis}}$, is required to satisfy $0.1 \leq x_{1,2} \leq 1.4$. A summary of the preselection criteria is provided in Table 6.2. An additional *analysis selection* is defined for usage in Section 6.3 and 8.5 by requiring that an event passes all preselection criteria, and imposing an additional cut that at least one τ-lepton needs to pass the tight WP of the GNTau algorithm [119].

6.3. Event Categorisation

To identify the $tHjb$ and $t\bar{t}H$ signal processes and improve background rejection, a trained HGNN developed during an internship [60], following a BSc thesis [61], is employed. Top-

Table 6.2.: Summary of the preselection criteria applied in the analysis.

object	**preselection requirements**
electron	$n_e = 0$
muon	$n_\mu = 0$
$\tau_{\mathrm{had-vis}}$	$n_\tau = 2$
	opposite charge
	$p_\mathrm{T} \geq 30\,\mathrm{GeV}$
	leading $\tau_{\mathrm{had-vis}}$: $p_\mathrm{T} \geq 40\,\mathrm{GeV}$
	$0.6 \leq \Delta R_{\tau_{\mathrm{had-vis}}\tau_{\mathrm{had-vis}}} \leq 3.2$
x_i	$0.1 \leq x_{1,2} \leq 1.4$
jets	$p_\mathrm{T} \geq 25\,\mathrm{GeV}$
	$n_j \geq 3$
	$n_b \geq 1$
	$\exists$ jet with $\lvert\eta\rvert \leq 3.2$ and $p_\mathrm{T} \geq 70\,\mathrm{GeV}$

quark candidates are reconstructed using hyperedges, with the highest-scoring candidate selected first and the second candidate identified after removing the three jets assigned to the best candidate. More details on the HGNN architecture, the input features and the truth matching procedure that was used can be found in the given references. The objects of interest for this analysis are the leading and sub-leading top quark, τ-lepton (τ_1 and τ_2), b-jet and light (u, d, c, s- or gluon) jet, as well as the bachelor jet (see Section 6.1). The distributions of kinematic variables p_T and η of all objects of interest with the analysis selection applied as described in Section 6.2 are given in Figures 6.1 and 6.2.

Events are categorised into four separate classes according to their highest prediction score among the four classes: two for the signals ($t\bar{t}H$ and $tHjb$) and two for the main backgrounds, $Z(\tau\tau)$ and dileptonic top background ($t\bar{t}$ and tW). Events are assigned to signal regions (SRs) or control regions (CRs). The SRs are designed to maximise the sensitivity to the signal processes, while the CRs are enriched in the dominant background processes and used to constrain their normalisation and reduce the associated systematic uncertainties in the statistical fit. For all SRs and CRs, the analysis selection needs to be fulfilled. For the $t\bar{t}H$ and $tHjb$ SR, events are required to be categorised into the $t\bar{t}H$ or $tHjb$ class by the HGNN, respectively. Similarly, for the $Z(\tau\tau)$ and dileptonic $t\bar{t}/tW$ CR, events are required to be categorised into the $Z(\tau\tau)$ or dileptonic $t\bar{t}/tW$ background class, respectively. The distributions of kinematic variables p_T and η of all objects of interest in all SRs and CRs are given in Figures B.1–B.8 in Appendix B.

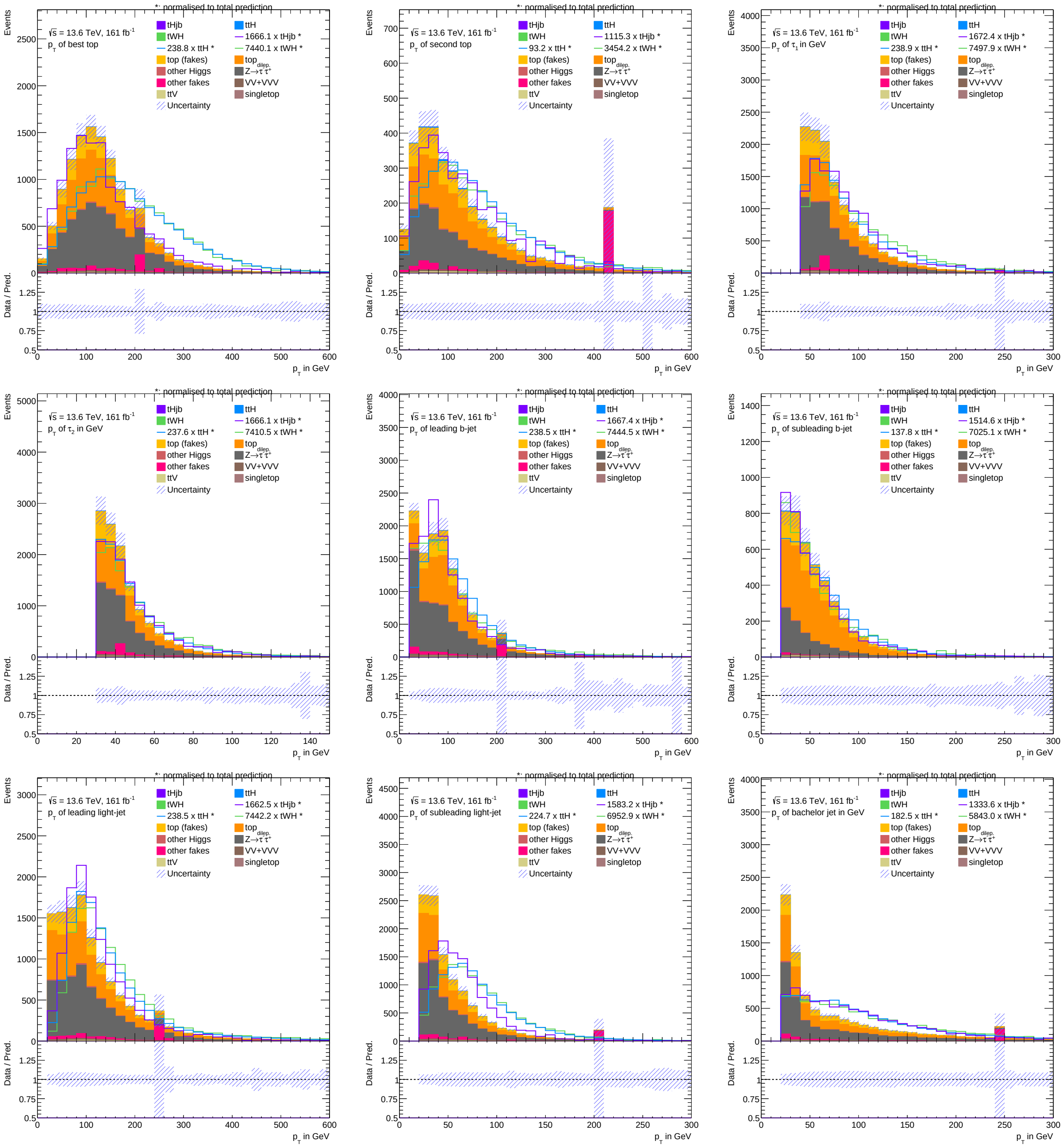


Figure 6.1.: Transverse momentum p_{T} of the best and second-best top quark candidate determined by the HGNN, the leading (τ_1) and sub-leading (τ_2) τ lepton, b-jet and light-jet, and the bachelor jet with the analysis selection applied as described in Section 6.2.

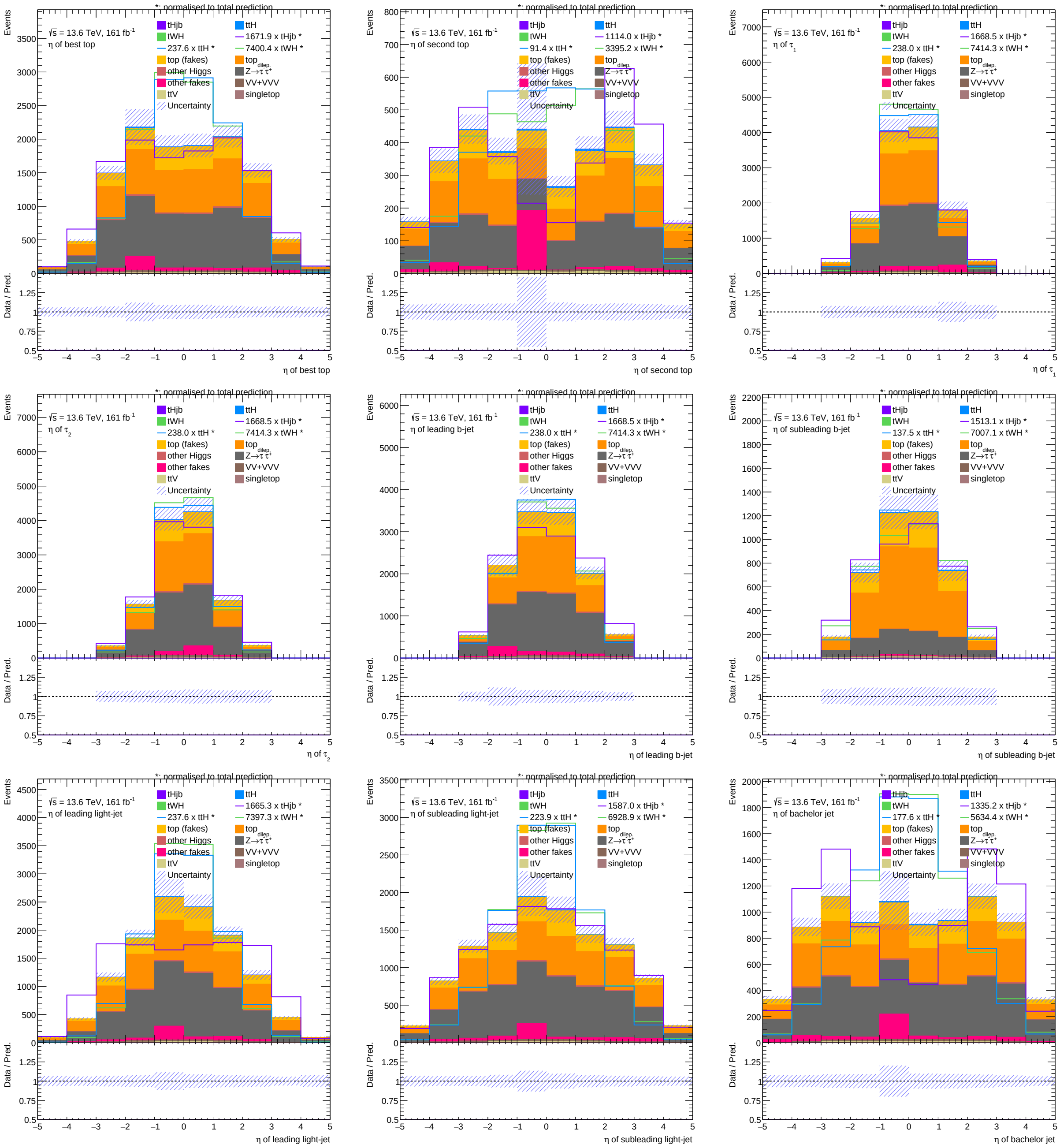


Figure 6.2.: Pseudorapidity η of the best and second-best top quark candidate determined by the HGNN, the leading (τ_1) and sub-leading (τ_2) τ lepton, b-jet and light-jet, and the bachelor jet with the analysis selection applied as described in Section 6.2.

7. CP-sensitive Observables

Top-associated Higgs boson production can constrain CP properties of the top-Higgs Yukawa coupling through rate measurements, kinematic variable distributions and CP-odd observables. While CP-odd observables would in principle provide an unambiguous method to probe CP symmetry, their reconstruction from measurements would require knowledge of the polarisation of the top quark and a method for differentiating between up- and down-type quarks in the hadronic decay channel of interest. For the current Run 3 dataset, this is not feasible. This analysis therefore focuses on specifically constructed kinematic CP-sensitive observables defined in Section 7.1. Their distributions are provided in Section 7.2.

7.1. Definitions

The CP-sensitive observables used in this analysis are defined in the right-handed coordinate system described in Section 4.2 as employed in ATLAS, writing three-momenta as $\boldsymbol{p} = (p^x, p^y, p^z)$. Subscripts indicate the particle that is referred to. The observables are taken from existing proposals in literature [126–142] and were originally developed for the $t\bar{t}H$ signal process where they are computed between the top and anti-top quark. None of the variables evaluated in their respective rest frames require a distinction between the top and anti-top quark, since such a distinction would be difficult to realise in experiments.

For the $tHjb$ signal, the CP-sensitive variables are constructed using the best top candidate and the bachelor jet, which is expected to originate from the light spectator quark and denoted by a B subscript. The Higgs boson is referred to by the subscript H. A normal vector $\boldsymbol{n} = (0, 0, 1)$ is defined. An overview of the CP-sensitive observables used in the analysis and the respective rest frames in which they are analysed is provided in Table 7.1.

To transform into the rest frame of system S defined by $\boldsymbol{p}_S = 0$ (where S is replaced by tB or $t\bar{t}$ for the corresponding signal processes) starting from the lab frame, first a rotation around the z axis is applied such that $\boldsymbol{p}_{\mathrm{T},S}$ is parallel to the x-axis. A boost along the

Table 7.1.: Overview of the CP-sensitive observables considered in this analysis, including their definition and the rest frames in which they are analysed. The placeholder X should be read as $\bar{t}$ for the $t\bar{t}H$ signal process and as B for the $tHjb$ and tWH signal processes.

observable	definition	reference frame
$p_{T,H}$	directly available from MMC	lab frame
$\lvert\Delta\eta_{tX}\rvert$	$\lvert\eta_t - \eta_X\rvert$	lab frame
$\lvert\Delta\phi_{tX}\rvert$	$\lvert\phi_t - \phi_X\rvert$	lab frame
$b_{2/tX}$	$\frac{(\boldsymbol{p}_t \times \boldsymbol{n})(\boldsymbol{p}_X \times \boldsymbol{n})}{\lvert\boldsymbol{p}_t\rvert \cdot \lvert\boldsymbol{p}_X\rvert}$	lab frame
$b_{4/tX}$	$\frac{p_t^z p_X^z}{\lvert\boldsymbol{p}_t\rvert \cdot \lvert\boldsymbol{p}_X\rvert}$	lab frame
$\lvert\cos\theta^*\rvert_{tX}$	$\frac{\lvert\boldsymbol{p}_t \cdot \boldsymbol{n}\rvert}{\lvert\boldsymbol{p}_t\rvert \cdot \lvert\boldsymbol{n}\rvert}$	tX rest frame

x-axis is performed so that $\boldsymbol{p}_{\mathrm{T},S} = 0$. Then, a boost along the z-direction ensures $\boldsymbol{p}_S = 0$. The procedure is illustrated in Figure 7.1.

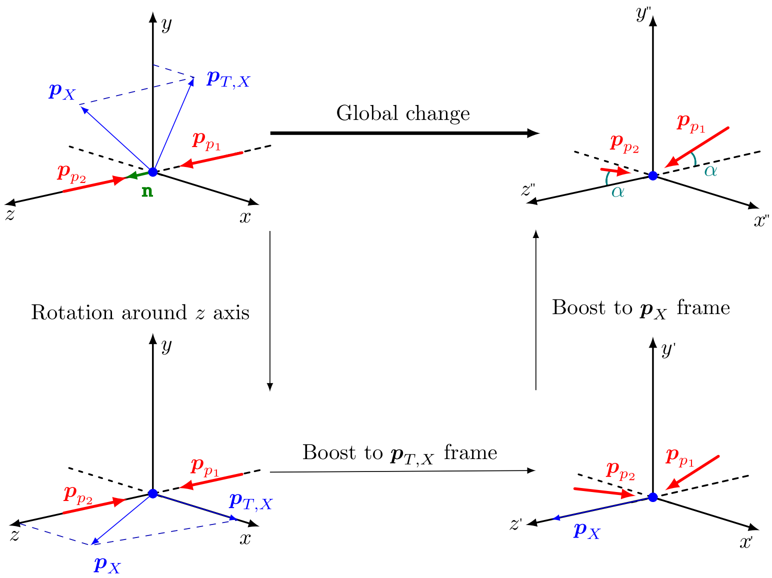


Figure 7.1.: The transformation from the lab frame to the rest frame of system S defined by $\boldsymbol{p}_S = 0$ as adapted in this work, where S corresponds to tB for the $tHjb$ and tWH signal and to $t\bar{t}$ for the $t\bar{t}H$ signal [143].

Because the bachelor jet is expected to originate from the light spectator quark as seen in Figure 2.5a and the corresponding truth-level information is not stored directly in the analysis framework that was used [111], CP-sensitive observables for the tHjb signal are computed only from reconstructed objects. The impact of the reconstruction step is instead evaluated using the $t\bar{t}H$ process, where the CP-sensitive observables are constructed between the $t\bar{t}$ pair on which truth-level information is available.

7.2. Distributions

The distributions of the CP variables presented in Table 7.1 for the $tHjb$ process computed from reconstructed objects for MC samples generated with different CP mixing angles α (see Equation (2.9)) are shown in Figure 7.2. For the $t\bar{t}H$ signal, the variables are computed from truth-level MC information and reconstructed objects as shown in Figure 7.3. The separation power of the observables is observed to be significantly reduced for the reconstructed objects compared to truth-level information. The effect is particularly prominent in the case of $b_{4/t\bar{t}}$ and $|\Delta\phi_{t\bar{t}}|$. Migration matrices indicating the bin-by-bin migration from a truth bin to a reconstructed bin for each variable are comprised in Appendix C.

For the tWH signal process, the CP-sensitive quantities are not constructed as the bachelor jet does not correspond to a physical object of interest in the study of CP properties.

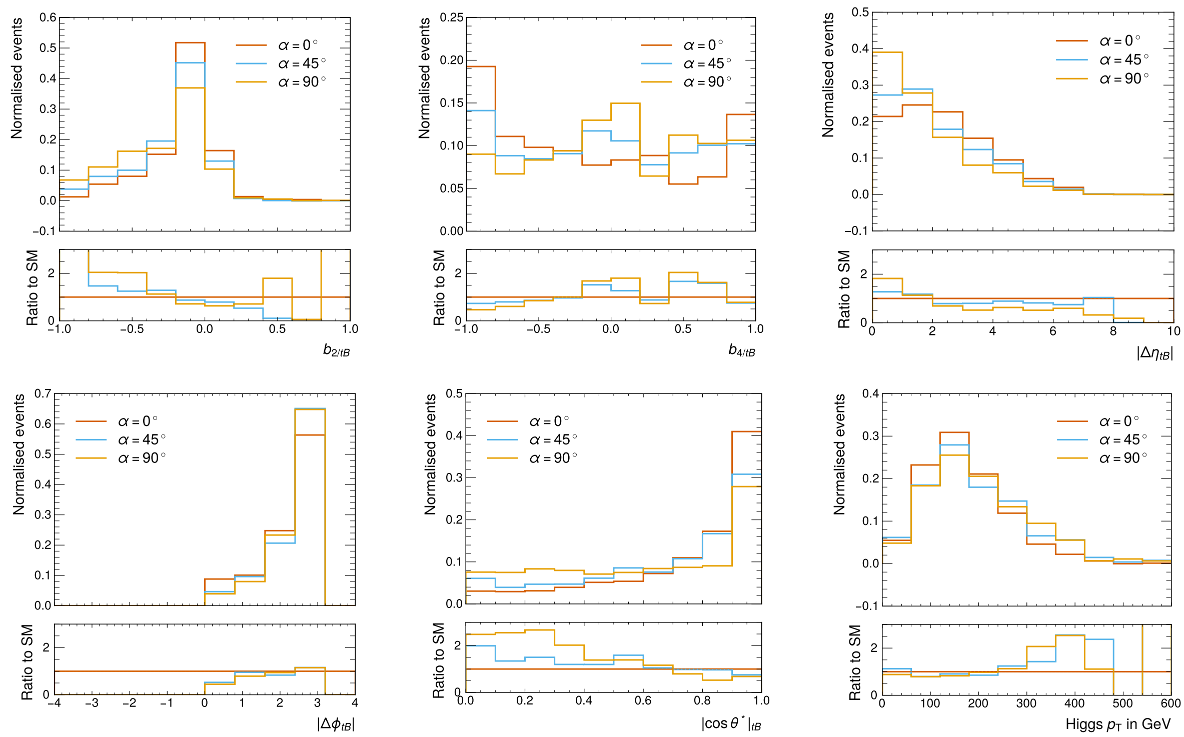


Figure 7.2.: Distributions of the CP-sensitive observables from Table 7.1 computed from the reconstructed top quark and bachelor jet for MC samples of the $tHjb$ process simulated with $\alpha = 0°, 45°$ and $90°$. The analysis selection is applied. The bottom panel shows the ratio to the SM prediction ($\alpha = 0°$).

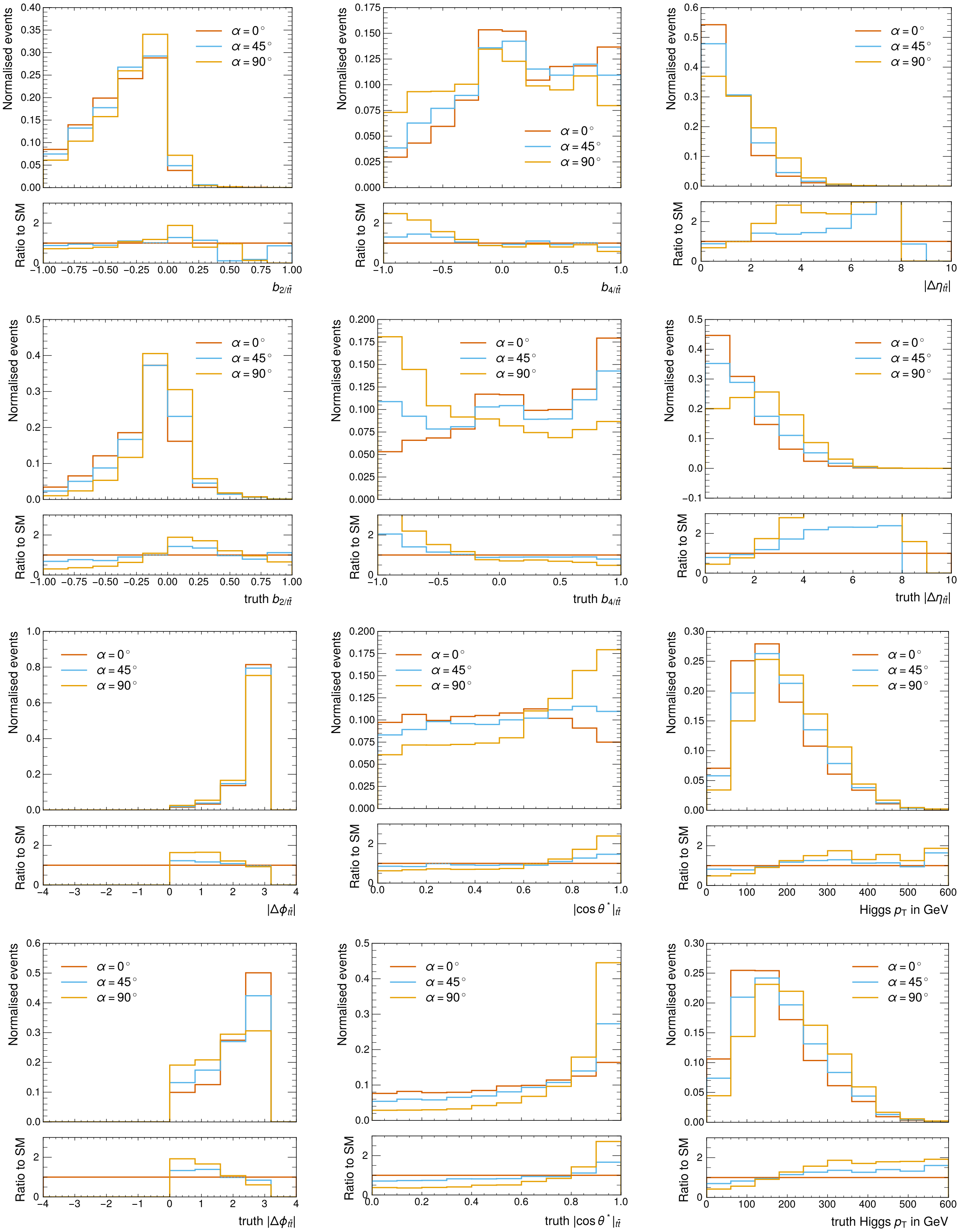


Figure 7.3.: Comparison of the CP-sensitive observables from Table 7.1 computed from reconstructed objects to a computation from MC truth information in the $t\bar{t}H$ signal process, using samples simulated with a mixing angle $\alpha = 0°, 45°$ and $90°$. The analysis selection is applied. The bottom panel shows the ratio to the SM prediction ($\alpha = 0°$).

8. Separating CP-Hypotheses

To discriminate between CP-even and CP-mixed top-Higgs Yukawa couplings in the $tHjb$ and $t\bar{t}H$ signal processes, binary classifiers employing Gradient Tree Boosting as specified in Section 8.1 are trained and evaluated according to Section 8.2. The CP-sensitive observables are computed between the relevant objects as introduced in Chapter 7. Separate classifiers are trained for the two processes as described in Section 8.3 and 8.4 and applied in their corresponding signal regions. The performance of the HGNN is evaluated in new regions defined according to the CP classification. The results are shown in Section 8.5.

8.1. Binary Classifier Specifications

MC-generated samples for different CP mixing angles α (see Equation (2.9)) from 0° (CP-even) to 90° (CP-odd) in steps of 15° are processed [111, 144, 145]. The CP-sensitive quantities $b_{2/tX}, b_{4/tX}$, $|\cos\theta^{\star}|_{tX}$, $|\Delta\eta_{tX}|$, $|\Delta\phi_{tX}|$ and $p_{T,H}$ (where tX corresponds to $t\bar{t}$ or tB as described previously) are computed and used as an input to the classifier. Another kinematic variable of the Higgs boson, η_H, was not yet investigated as a CP-sensitive observable in existing proposals, but it was found to improve the classifier's performance in some cases and therefore added as an input variable.

The binary classification is based on XGBoost (Extreme Gradient Boosting), a scalable ML system employing Gradient Tree Boosting as explained in Section 3.5 [63, 146]. Class weights are balanced by taking the sum of the weights for each class, selecting the maximum and scaling the weights for the other class by a factor calculated as the maximum sum divided by the sum of weights of that class. Events where one of the top quarks or the bachelor jet is not reconstructed contain placeholder values set to -999.0 for the CP observables and are omitted in training. Negative MC weights are part of the complete SM prediction, however, including them in training can lead to a negative weighted average loss per batch which causes the minimisation procedure of the training to break down as the loss diverges towards negative infinity. To remove events with negative MC weights while preserving the physical meaning, the positive weights are re-weighted by a sample-wise computed scale factor $\sum_i w_i / \sum_{i,w_i>0} w_i$ to account for the normalisation impact of the negative weights. A summary table of the MC statistics available for training the binary classifiers after removing events with placeholder values or negative MC weights is provided in Table D.1 in Appendix D.1.

The weighted binary cross-entropy loss function

$$L = \sum -w_n [y_n \cdot \ln p_n + (1 - y_n) \ln(1 - p_n)] \,, \tag{8.1}$$

is used for training, where w_n is the weight of event n, y_n is the true label (0 for CP-even, 1 for CP-mixed) and p_n is the prediction, a continuous output score from 0 to 1 with 0 corresponding to predicting the CP-even and 1 the CP-odd scenario.

The model is specified through various parameters that are described in the documentation of XGBoost [147, 148] and set according to Table D.2 provided in Appendix D.2.

8.2. Evaluation Metrics

This section describes the evaluation metrics implemented to judge the model's performance. The model is trained and evaluated using four-fold cross validation as described in Section 3.2. Four independent models (folds) are trained and each event is evaluated only by the fold for which it belongs to the corresponding test set. This ensures every event remains statistically independent of the model that evaluates it and contributes exactly once to the final estimates of the model's performance.

For each fold, the false positive rate (FPR) and true positive rate (TPR) for a set of events $\mathcal{O} \subseteq \mathcal{E}$ where $\mathcal{E}$ denotes the test set of the fold under consideration, are defined as

$$\mathrm{FPR}_{\mathcal{O}} = \frac{\sum_{i\in\mathcal{O}} w_i^{\mathrm{CP-even}}}{\sum_{j\in\mathcal{E}} w_j^{\mathrm{CP-even}}} \qquad \text{and} \qquad \mathrm{TPR}_{\mathcal{O}} = \frac{\sum_{i\in\mathcal{O}} w_i^{\mathrm{CP-odd}}}{\sum_{j\in\mathcal{E}} w_j^{\mathrm{CP-odd}}} \tag{8.2}$$

where w are the corresponding event weights. The set $\mathcal{O}$ is parametrised by the treshold p_{cut} set on the output score p_n of the ML model for event n as $\mathcal{O}(p) = \{\text{event } n \mid p_{\mathrm{cut}} \leq p_n\}$. The receiver operating characteristic (ROC) curve illustrates the performance of the binary classifier at different threshold values applied to the output score, above which an event is classified as CP-odd in this analysis. A separate ROC curve and corresponding area under the ROC curve (AUC) are computed for each fold. The AUC is calculated by approximating the integral with the trapezoid rule,

$$\mathrm{AUC} \simeq \sum_{i=0}^{n_{\mathrm{bins}}^{\mathrm{AUC}}-1} \frac{\mathrm{TPR}(\mathrm{FPR}_{i+1}) + \mathrm{TPR}(\mathrm{FPR}_{i})}{2} (\mathrm{FPR}_{i+1} - \mathrm{FPR}_{i}) \,, \tag{8.3}$$

with $n_{\mathrm{bins}}^{\mathrm{AUC}}$ being the number of bins for the numerical integration. A higher AUC implies a better-performing model.

To visualise the classifier's performance, a confusion matrix using a cut of 0.5 on the model output score to discriminate between CP mixing angle 0° and 90° (or another mixing angle $\neq 0°$) is used. The confusion matrix is computed from the combined set of out-of-fold predictions obtained from all four test folds and therefore corresponds to the performance of the classifier on the full dataset while maintaining statistical independence between training and evaluation samples.

To study the importance of individual input features for the model's performance, the technique of permutation importance is used. For each fold and each input variable, the entries of that input variable are randomly permuted between events of the corresponding test set while all other variables are left unchanged. The AUC obtained with the shuffled set is compared to the AUC of the nominal (unshuffled) set, $\mathrm{AUC}_{\mathrm{nom.}}$ The resulting relative degradation in the model's performance, the *permutation importance*, is measured as

$$\Delta\mathrm{AUC} = \frac{\mathrm{AUC}_{\mathrm{nom.}} - \mathrm{AUC}}{\mathrm{AUC}_{\mathrm{nom.}}} \tag{8.4}$$

If the input feature carries significant information on the classification task (in this case, the CP mixing angle), a high permutation importance of that feature is expected. The permutation importance is evaluated independently for each fold and the average over the four folds is chosen as the final permutation importance of the input feature. The variation between folds is used to estimate the stability of the feature's importance.

Additionally, the separation power of the individual input features and the output of the trained model is studied. Taking the number of bins to be $n_{\mathrm{bin}}^{\mathrm{sep.}} = 10$, the separation $\mathcal{S}$ provided by a variable is given by

$$\mathcal{S} = \frac{1}{2} \sum_{i}^{n_{\mathrm{bin}}^{\mathrm{sep.}}} \frac{(S_i - B_i)^2}{(S_i + B_i)} \tag{8.5}$$

where S_i and B_i are the normalised number of CP-odd (signal) and CP-even (background) events in the i-th bin of the variable's distribution. Therefore, a comparison of the separation powers provided by different variables is only meaningful if the same number of bins is used in all histograms. For the classifier output, the separation is computed using the combined set of out-of-fold predictions from all four test folds.

8.3. Binary Classification for the $tHjb$ Signal

A binary classifier based on the XGBoost package [146] with its parameters set as described in Section 8.1 is trained by employing a four-fold cross validation procedure. Each fully trained fold is then evaluated on its respective test set according to the metrics set in Section 8.2.

8.3.1. Distinguishing the CP-even and CP-odd Scenario

The logarithm $\ln(L)$ of the loss function given by Equation (8.1), computed on the training and validation set for each fold of the binary classifier trained on MC-generated samples of the $tHjb$ process for a completely CP-even vs CP-odd coupling is shown on the left in Figure 8.1, whereas the ROC curve for each fold of the classifier evaluated on its respective test set is shown on the right. While large differences in the number of trees between the folds as well as slightly different limits for $\ln(L)$ can be observed, all folds reach very similar values of about 0.65 for the AUC defined according to Equation (8.3).

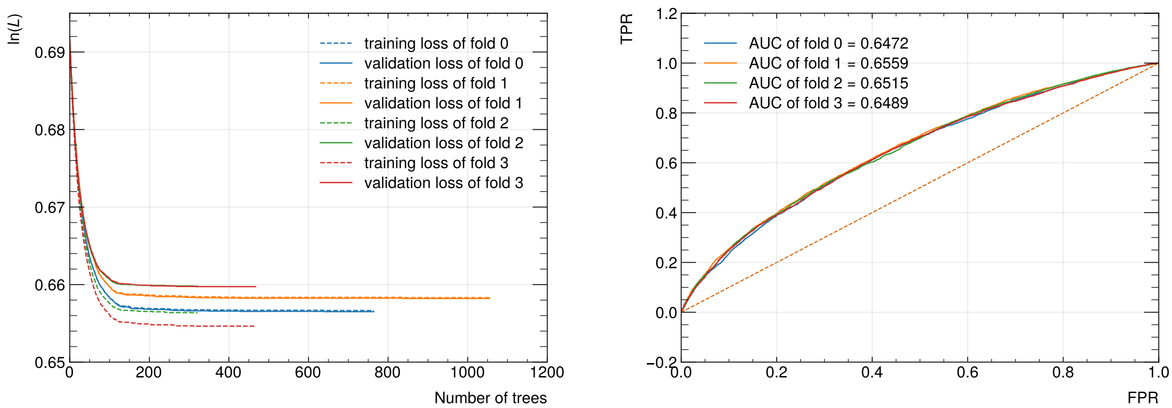


Figure 8.1.: Logarithm $\ln(L)$ of the training and validation loss plotted against the number of trees (left) and ROC curve (right) for each fold of the binary classifier trained on $tHjb$ samples for CP-even (mixing angle 0°) and CP-odd (90°) top-Higgs-Yukawa coupling.

The confusion matrix of the classifier, using a cut of 0.5 on the model output score to discriminate between CP mixing angle 0° and 90°, is shown on the left in Figure 8.2. The matrix is normalised per row. The ranked permutation importance for all input variables of the CP-even vs CP-odd classifier for the $tHjb$ signal process is shown in Figure 8.2 on the right.

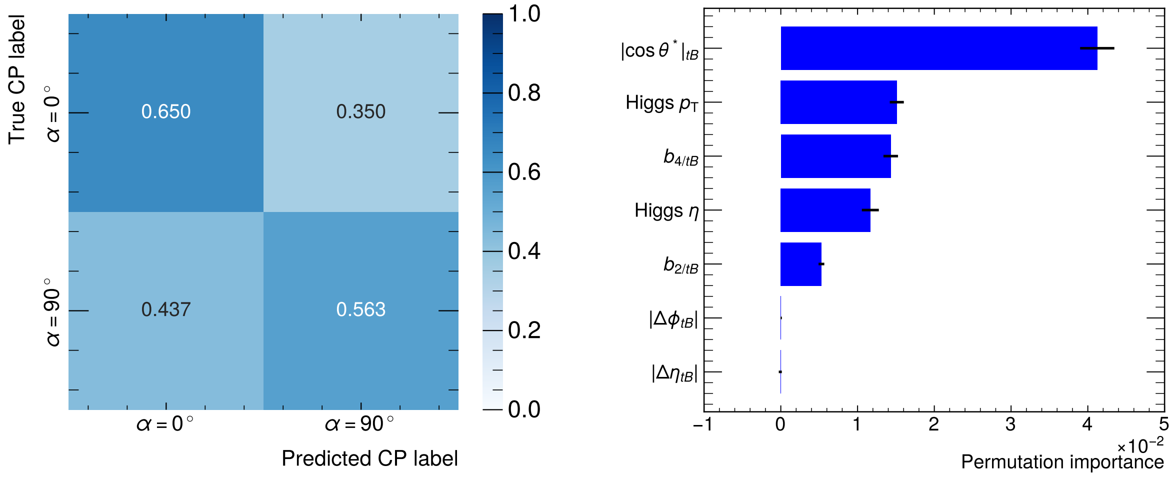
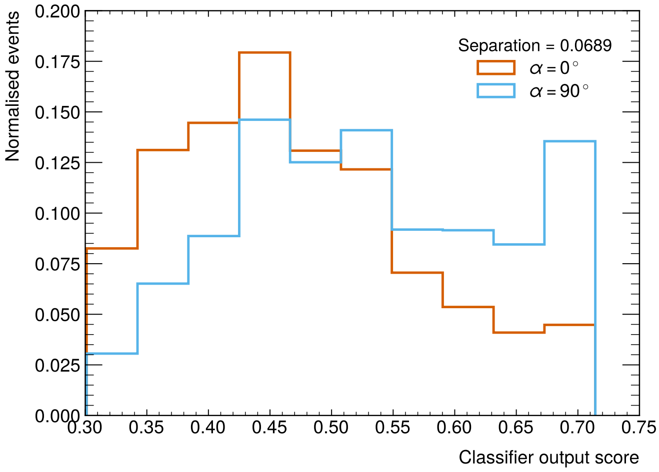


Figure 8.2.: Confusion matrix normalised per row, employing a cut of 0.5 (left) on the output of the binary CP-even (mixing angle 0°) vs CP-odd (90°) classifier for the $tHjb$ signal, and ranked permutation importance of its input features (right).

The separation power of the individual input features and the trained model is calculated according to Equation (8.5) with the number of bins chosen to be $n_{\text{bin}}^{\text{sep.}} = 10$. The normalised number of events as a function of the model output score and the calculated separation power of the score and each input variable for the CP-even vs CP-odd classifier for the $tHjb$ signal are shown in Figure 8.3. It can be seen that the ML model provides an increased separation power compared to the individual variables, however due to the insufficient separation contained in the input variables the separation power of the final model is also limited. Furthermore, a discrepancy in the rankings of input features according to separation power and permutation importance is noticed when comparing the Table in Figure 8.3 to the permutation importance plot on the right in Figure 8.2.



Variable	Separation
output score	6.895%
$\lvert\cos\theta^*\rvert_{tB}$	4.436%
b_4	3.722%
b_2	3.665%
$\lvert\Delta\eta_{tB}\rvert$	3.500%
η_H	2.305%
$p_{T,H}$	2.144%
$\lvert\Delta\phi_{tB}\rvert$	0.545%

Figure 8.3.: Normalised number of events plotted against the output score of the CP-even (mixing angle 0°) and CP-odd (90°) classifier (left) and ranked separation power of the CP-sensitive observables listed in Table 7.1 and the output score (right) for the $tHjb$ signal.

8.3.2. Separation Power for Comparing the CP-even Scenario to Different CP-mixed Scenarios

The training and evaluation process outlined in the previous section for distinguishing the CP-even and CP-odd scenario can be repeated for five separate classifiers with an identical architecture that distinguish the CP-even scenario from the available CP-mixed scenarios $\alpha = 15°, 30°, 45°, 60°$ and $75°$, respectively. For each of the classifiers, the separation power of each input variable and the output score of the trained model is computed as described in the previous section. A summary of the computed separation powers for the individual input features and output score for all classifiers is provided in Table 8.1. As the current exclusion limit at 95% confidence level for the CP mixing angle is at $\alpha_{\text{excl.}} = 43°$ [149], the performance of the classifier distinguishing the CP-even against the maximally CP-mixed scenario ($\alpha = 45°$) is of special importance. Therefore, all plots provided in Section 8.3.1 are reproduced for this classifier and shown in Figures E.1–E.3 in Appendix E.

8.4. Binary Classification for the $t\bar{t}H$ Signal

A binary classifier with the same architecture as the classifier for the $tHjb$ signal processes is implemented to study the $t\bar{t}H$ signal. In this case, the CP-sensitive quantities given as input features are constructed between the top and anti-top quark. All evaluation metrics are applied as described for the $tHjb$ process in Section 8.3.

Table 8.1.: Separation power in % for the CP-sensitive observables listed in Table 7.1, the additional input variable η_H and the classifier output score for distinguishing different scenarios for the CP mixing angle α from Equation (2.9) in $tHjb$ samples. Darker colour corresponds to larger separation.

Variable	0° vs 15°	0° vs 30°	0° vs 45°	0° vs 60°	0° vs 75°	0° vs 90°
output score	0.001	0.412	2.616	5.082	5.875	6.895
$\lvert\cos\theta^*\rvert$	0.044	0.638	1.634	3.196	3.626	4.436
b_4	0.056	0.484	1.235	2.498	2.975	3.722
b_2	0.088	0.720	1.506	2.768	3.040	3.665
$\lvert\Delta\eta_{tB}\rvert$	0.057	0.472	1.126	2.444	2.827	3.500
η_H	0.070	0.649	1.065	1.834	2.112	2.305
$p_{\mathrm{T},H}$	0.105	1.079	1.907	2.362	2.271	2.144
$\lvert\Delta\phi_{tB}\rvert$	0.049	0.039	0.243	0.312	0.390	0.545

8.4.1. Distinguishing the CP-even and CP-odd Scenario

A binary classifier is trained on MC-generated $t\bar{t}H$ samples with a completely CP-even vs CP-odd coupling as described for the $tHjb$ process in Section 8.3.1. The logarithm $\ln(L)$ of the loss function given by Equation (8.1), computed on the training and validation set for each fold, is shown on the left in Figure 8.1, whereas the ROC curve for each fold is shown on the right. Similarly to the classifiers trained on the $tHjb$ samples described in Section 8.3.1, large differences in the number of trees between the folds as well as slightly different limits for $\ln(L)$ can be observed. However, all folds again reach very similar values of about 0.656 for the AUC defined according to Equation (8.3).

The confusion matrix of the classifier for the $t\bar{t}H$ signal is shown on the left in Figure 8.5, and the ranked permutation importance for all input variables is given on the right. All metrics are calculated following the description provided for the $tHjb$ process in Section 8.3.1. Compared to the permutation importance observed for the classifier employed in the $tHjb$ signal process (see Figure 8.2 on the right), the input variables show very different permutation importances.

The normalised number of events as a function of the model output score and the calculated separation power of the score and each input variable are shown in Figure 8.6. In this case, no deviations between rankings of the input variables according to separation power and permutation importance are observed. Compared to the model trained for the $tHjb$ signal, fewer variables contain information that improves the classifier.

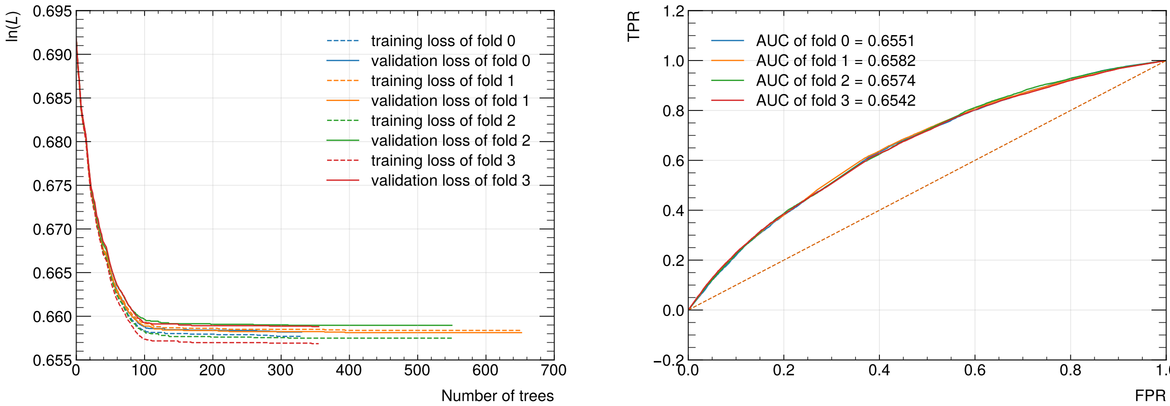


Figure 8.4.: Training and validation loss plotted against the number of trees (left) and ROC curve (right) of the binary classifier trained on $t\bar{t}H$ samples for CP-even (mixing angle 0°) and CP-odd (90°) top-Higgs-Yukawa coupling.

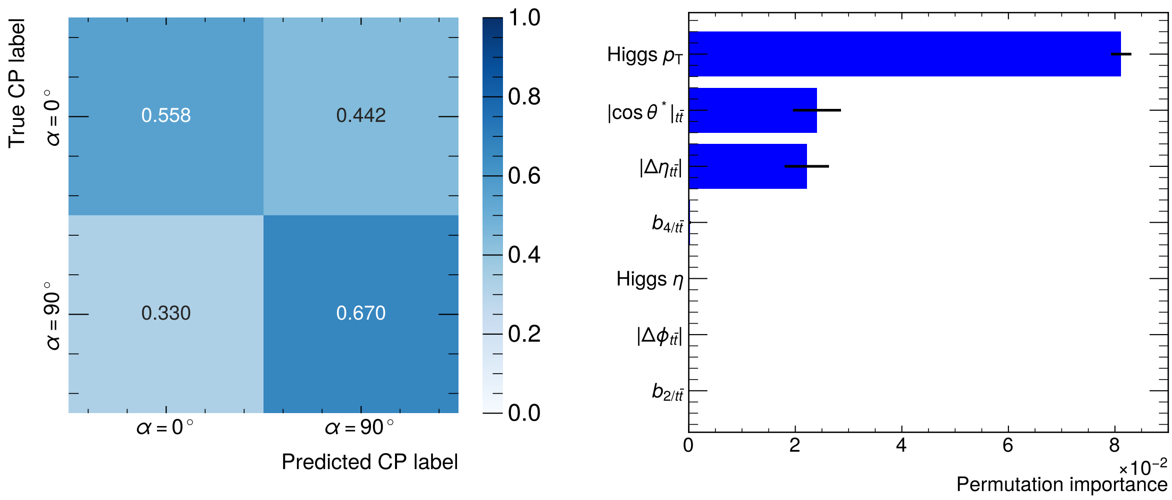
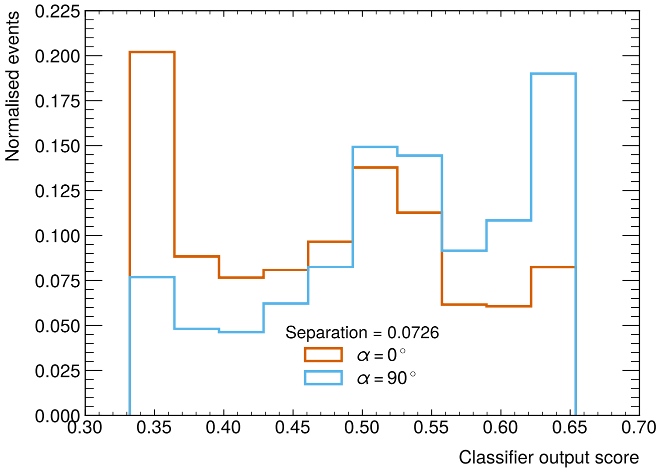


Figure 8.5.: Confusion matrix normalised per row and employing a cut of 0.5 (left) of the binary CP-even (mixing angle 0°) vs CP-odd (90°) classifier for the $t\bar{t}H$ signal processes, and ranked permutation importance of its importance (right).

8.4.2. Separation Power for Comparing the CP-even Scenario to Different CP-mixed Scenarios

The procedure described in Section 8.3.2 is repeated to train five classifiers that distinguish the CP-even scenario from $\alpha = 15°, 30°, 45°, 60°$ and $75°$ in the $t\bar{t}H$ process. A summary of the separation powers for the individual input features and output score for all classifiers is provided in Table 8.2. All plots provided in Section 8.4.1 are reproduced for a classifier distinguishing $\alpha = 0°$ and $\alpha = 45°$ and shown in Figures E.4–E.6 in Appendix E.



Variable	Separation
output score	7.257%
$p_{T,H}$	3.306%
$\|\cos\theta^*\|_{t\bar{t}}$	3.257%
$\|\Delta\eta_{t\bar{t}}\|$	3.250%
b_4	2.258%
b_2	0.942%
η_H	0.921%
$\|\Delta\phi_{t\bar{t}}\|$	0.277%

Figure 8.6.: Normalised number of events plotted against the output score of the CP-even (mixing angle 0°) and CP-odd (90°) classifier (left) and ranked separation power of the CP-sensitive observables listed in Table 7.1 and the output score (right) for the $t\bar{t}H$ signal.

Table 8.2.: Separation power in % for the observables listed in Table 7.1, the additional input variable η_H and the output score for distinguishing different CP-mixed scenarios in $t\bar{t}H$ samples. Darker colour corresponds to larger separation.

Variable	0° vs 15°	0° vs 30°	0° vs 45°	0° vs 60°	0° vs 75°	0° vs 90°
output score	0.001	0.001	1.099	3.371	5.966	7.257
$p_{T,H}$	0.020	0.101	0.513	1.435	2.613	3.306
$\|\cos\theta^*\|$	0.007	0.146	0.625	1.656	2.788	3.257
$\|\Delta\eta_{t\bar{t}}\|$	0.009	0.154	0.658	1.678	2.772	3.250
b_4	0.016	0.138	0.465	1.224	1.909	2.258
b_2	0.004	0.029	0.175	0.498	0.804	0.942
η_H	0.010	0.034	0.117	0.346	0.641	0.921
$\|\Delta\phi_{t\bar{t}}\|$	0.011	0.024	0.102	0.252	0.396	0.490

8.5. Integration into the HGNN used for Event Classification

The binary classifiers described in the previous sections discriminate between different scenarios for the CP mixing angle α (see Equation (2.9)) only by comparing the shape of the distributions of the CP-sensitive variables introduced in Chapter 7. A CP-symmetry-violating top-Higgs-Yukawa coupling would also cause a significant change in cross sections as listed in Table 2.1, yielding more signal events for the $tHjb$ and tWH process and less for the $t\bar{t}H$ process. A change in the cross sections could provide more information than utilising only the differences in the shape of distributions for the objective of distinguishing different scenarios for the CP-mixing. An increase above the expected cross section could however also be caused by different BSM effects, whereas the CP variables are designed to be relatively insensitive to other effects except for a CP-mixed scenario of the top-Higgs-Yukawa coupling. The impact of the shape distributions of CP-sensitive variables in comparison to the expected changes in cross section is analysed in this Section.

A HGNN is used as described in Section 6.3 to reconstruct the best top/anti-top quark candidates and categorise events into four separate classes according to their highest prediction score among the four classes: $t\bar{t}H$, $tHjb$, $Z(\tau\tau)$ and dileptonic top background ($t\bar{t}$ and tW) [60, 61]. The $t\bar{t}H$ and $tHjb$ SRs are defined as described in Section 6.3 according to the respective predicted output class by the HGNN, along with fulfilling the analysis selection defined in Section 6.2.

In these signal regions, the respective binary classifiers that were trained on the $tHjb$ or $t\bar{t}H$ samples as described in the previous sections are applied. In Figure 8.7, the output of the binary classifier trained to distinguish CP-even and CP-odd $tHjb$ samples is plotted in the $tHjb$ SR, using the CP-even (on the left) or CP-odd (on the right) signal samples as input, along with CP-even background samples (if the coupling is involved). Similarly, in Figure 8.8, the output of the classifier trained to distinguish CP-even and CP-odd $t\bar{t}H$ samples is plotted in the $t\bar{t}H$ SR for the same input samples. No significant change caused by possible selection-specific acceptance effects of the expected increase in cross section for the $tHjb$ process and decrease for the $t\bar{t}H$ process is observed in the MC samples. When comparing the shape distributions of the classification for the $t\bar{t}H$ signal in Figure 8.8 represented by the solid lines, a nearly flat distribution is observed for the CP-even signal samples. However, for the CP-odd samples, a higher content with a factor of approximately two compared to the CP-even (score ≤ 0.5) bin is observed in the CP-odd bin (score > 0.5) for all three signal samples. For the classifier trained on the $tHjb$, however, the content in the CP-even bin is much higher for both the CP-even and CP-odd signal samples, showing a lower performance level compared to the $t\bar{t}H$ classifier. Furthermore, as seen in Figure 8.8, the $t\bar{t}H$ SR contains many $tHjb$ events in the CP-odd case, whereas in the CP-even case the $tHjb$ is almost completely filtered out by the HGNN. This shows the performance of the HGNN used in event classification is likely impacted in the case of CP-mixing, as it was solely trained on CP-even input samples.

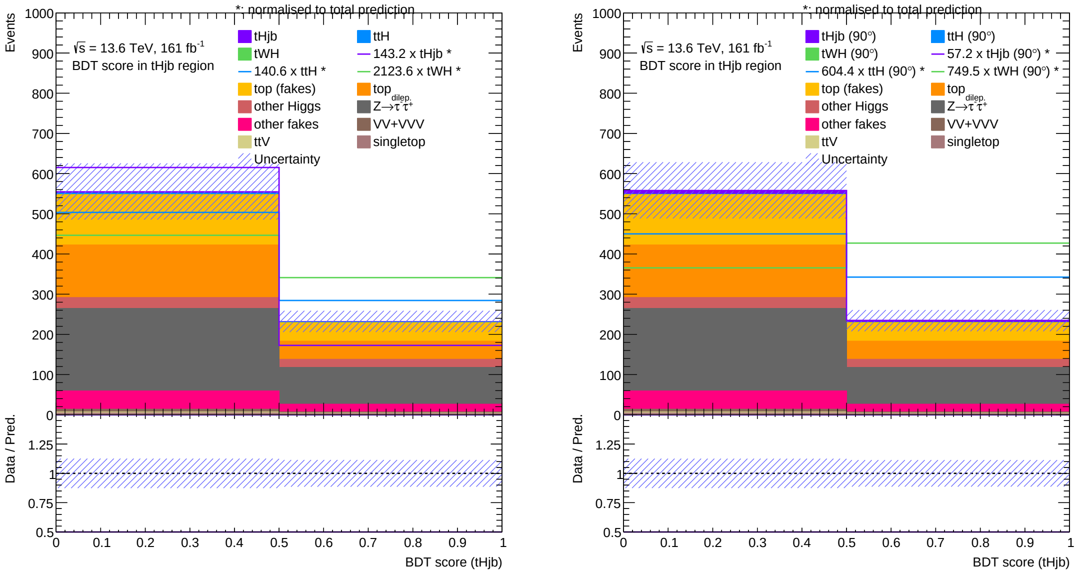


Figure 8.7.: Output score of the binary CP angle classifier trained to analyse the $tHjb$ signal in the $tHjb$ SR defined by the HGNN. The top-Higgs-Yukawa coupling for all three signal samples is simulated with $\alpha = 0°$ on the left and $\alpha = 90°$ on the right using the HC model shown in Equation (2.9). The background samples are modelled with $\alpha = 0°$.

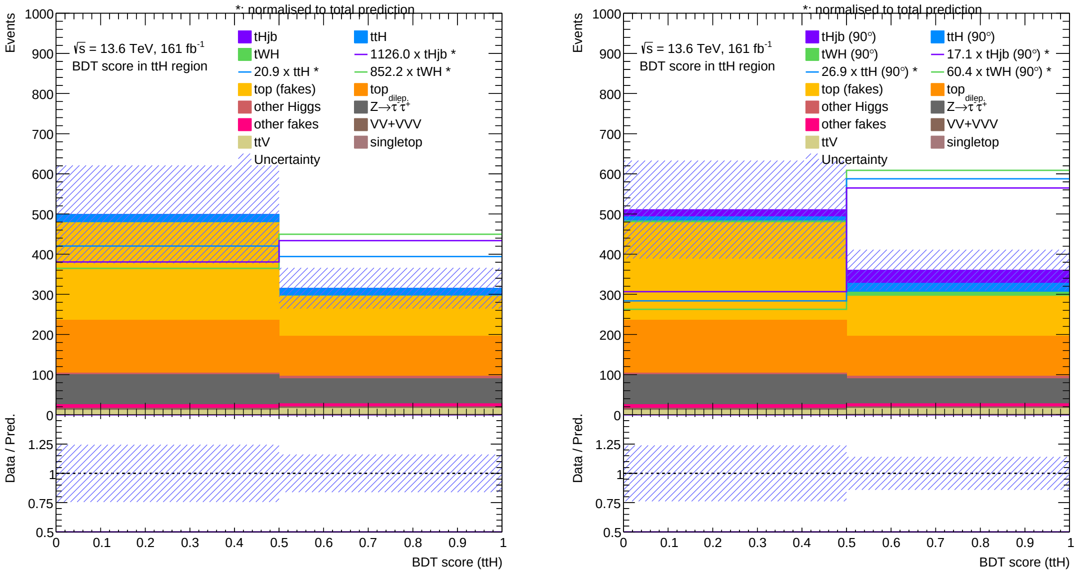


Figure 8.8.: Output score of the binary CP angle classifier trained to analyse the $t\bar{t}H$ signal in the $t\bar{t}H$ SR defined by the HGNN. The top-Higgs-Yukawa coupling for all three signal samples is simulated with $\alpha = 0°$ on the left and $\alpha = 90°$ on the right using the HC model shown in Equation (2.9). The background samples are modelled with $\alpha = 0°$.

In the $tHjb$ SR defined by the output of the HGNN, two separate SRs are defined: the $tHjb_{\alpha=0^\circ}$ SR is defined by the same criteria as the $tHjb$ SR along with imposing an additional cut of `score<0.5` on the output score of the binary CP-even ($\alpha = 0^\circ$) vs CP-odd ($\alpha = 90^\circ$) classifier for the $tHjb$ signal. The $tHjb_{\alpha=90^\circ}$ instead imposes an additional cut of `score<0.5` on the output score. The distribution of the output score for the $tHjb$ class of the HGNN in the $tHjb_{\alpha=0^\circ}$ and the $tHjb_{\alpha=90^\circ}$ SR is shown in Figure 8.9. The $tHjb$, tWH and $t\bar{t}H$ signal samples are simulated using the HC model with a mixing angle $\alpha = 0^\circ$ in the top row, $\alpha = 45^\circ$ in the middle row and $\alpha = 90^\circ$ in the bottom row for the top-Higgs-Yukawa coupling. The background samples are modelled with $\alpha = 0^\circ$. A severe degradation in the event classification performance of the HGNN for the $tHjb$ signal is observed in the shape distributions of the $tHjb$-score for the CP-mixed $tHjb$ samples in both the CP-even and CP-odd region. The $tHjb$ output score of the multi-class HGNN does not provide a good separation between the background samples (or other signal samples) and the CP-mixed $tHjb$ sample.

An analogous procedure is applied in the $t\bar{t}H$ SR defined by the output of the HGNN, where again two separate SRs are defined: the $t\bar{t}H_{\alpha=0^\circ}$ SR is defined by the same criteria as the $t\bar{t}H$ SR along with imposing an additional cut of `score<0.5` on the output score of the binary CP-even ($\alpha = 0^\circ$) vs CP-odd ($\alpha = 90^\circ$) classifier for the $t\bar{t}H$ signal. The $t\bar{t}H_{\alpha=90^\circ}$ instead imposes an additional cut of score ≥ 0.5 on the output score. The distribution of the output score for the $t\bar{t}H$ class of the HGNN in the $t\bar{t}H_{\alpha=0^\circ}$ and the $t\bar{t}H_{\alpha=90^\circ}$ SR is shown in Figure 8.10, following the same three cases for the modelling of the signal and background samples described above. In contrast to the previous study, the performance of the HGNN in the classification of the $t\bar{t}H$ signal does not show any significant decrease. However, the $tHjb$ and tWH signal samples show a shift to higher $t\bar{t}H$ scores in the CP-mixed samples, likely leading to a reduced ability of the HGNN to distinguish between different signal samples.

8.6. Systematic Uncertainties

Due to the time frame of this thesis, systematic uncertainties are not studied in the analysis of the binary classifiers employed to distinguish different scenarios for the CP mixing angle α that were described in Section 8.3 and 8.4. However, the analysis of the HGNN that was used for event categorisation into the two signal processes and two main backgrounds estimates systematic uncertainties as explained in [60, 61].

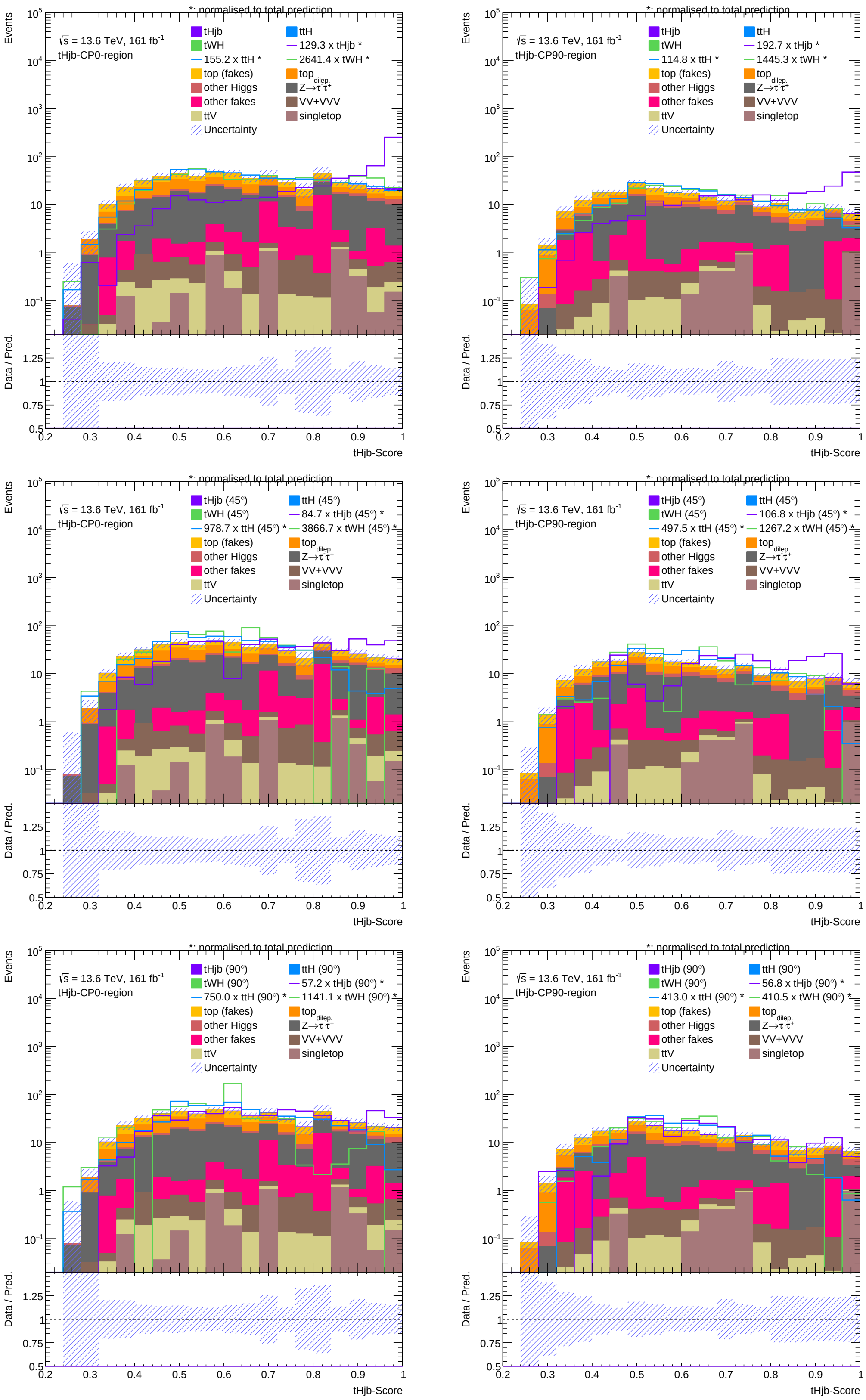


Figure 8.9.: Distributions of the output score for the $tHjb$ class of the HGNN in the $tHjb_{\alpha=0^\circ}$ (left) and the $tHjb_{\alpha=90^\circ}$ (right) SR. The top-Higgs-Yukawa coupling for all signal samples is simulated using the HC model with $\alpha = 0^\circ$ in the top row, $\alpha = 45^\circ$ in the middle row and $\alpha = 90^\circ$ in the bottom row. The background samples are modelled with $\alpha = 0^\circ$.

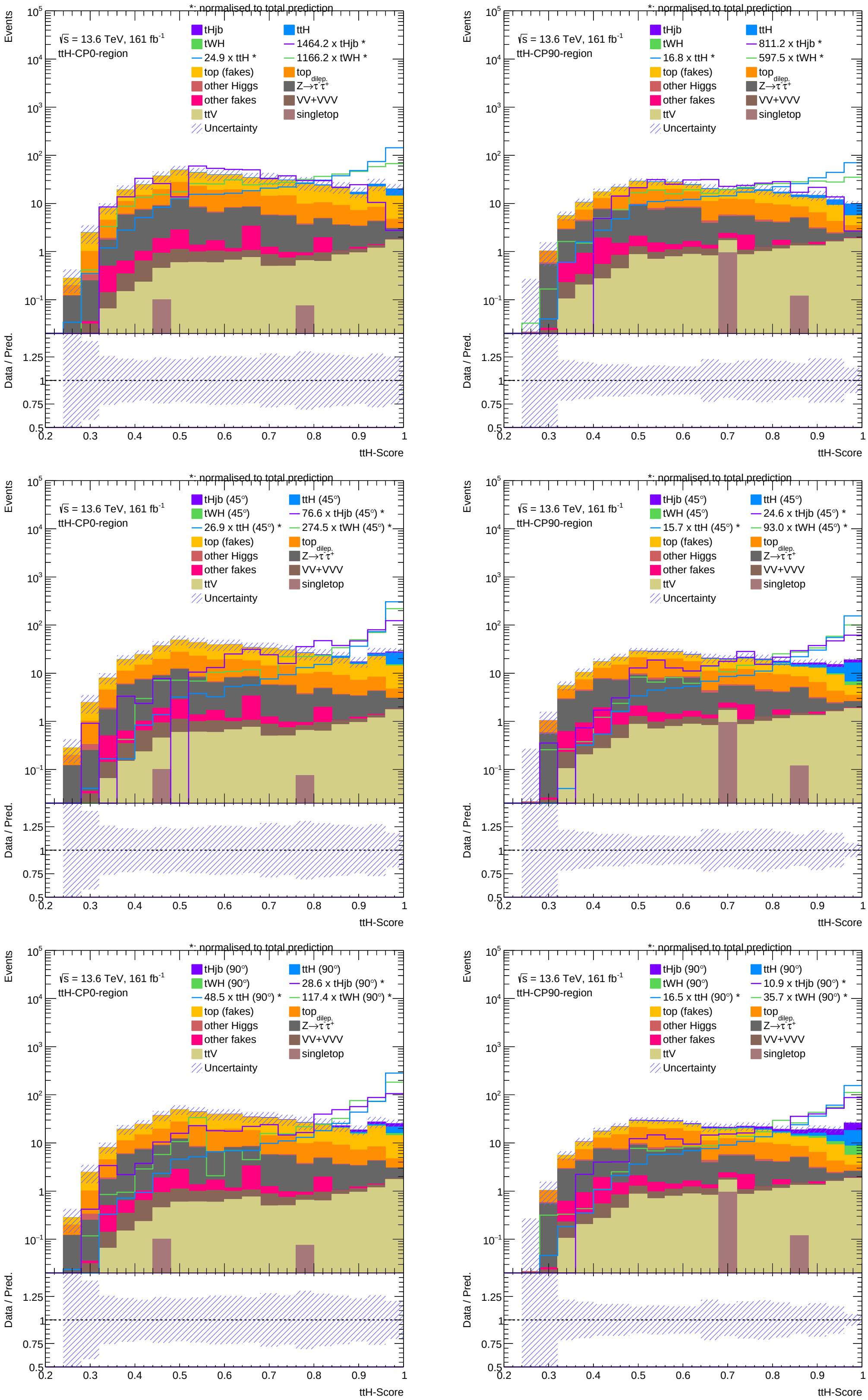


Figure 8.10.: Distributions of the output score for the $t\bar{t}H$ class of the HGNN in the $t\bar{t}H_{\alpha=0^\circ}$ and the $t\bar{t}H_{\alpha=90^\circ}$ SR. The top-Higgs-Yukawa coupling for all signal samples is simulated using the HC model with $\alpha = 0^\circ$ in the top row, $\alpha = 45^\circ$ in the middle row and $\alpha = 90^\circ$ in the bottom row. The background is modelled with $\alpha = 0^\circ$.

9. Conclusion and Outlook

Based on cross section measurements, a CP-symmetry-violating coupling of the Higgs boson to the top quark as described in the HC model, characterised by the mixing angle α, is currently constrained to $|\alpha| < 43°$ at 95% confidence level [149]. However, the cross section of a process is not only influenced by a possible CP-mixing in the coupling, but also sensitive to various other effects. Therefore, several observables that are mostly insensitive to other parameters and provide an exclusive probe of the CP-symmetry properties are proposed in the literature [143] and were analysed in this thesis. It was observed that the performance of the observables is severely impacted by detector-level efficiency and reconstruction effects when comparing the results obtained from reconstructed objects to particle-level information from the MC generator in the $t\bar{t}H$ signal process. A similar degradation is expected for the $tHjb$ signal process, accessing the particle-level MC information for the light spectator quark that is expected to correspond to the bachelor jet used to compute the CP-sensitive observables would be necessary to quantify this statement and was not possible in the time frame of this thesis.

The limited separation achieved by the considered CP-sensitive observables is partly caused by the signature expected for tH and $t\bar{t}H$ production with $H \to \tau^+\tau^-$ in the full-hadronic final state. Due to the high hadronic activity, the probability that the constructed bachelor jet does not correspond to the light spectator quark, meaning that the wrong object is selected to compute the CP-sensitive observables, is much higher than the probability of selecting the wrong objects in "cleaner" final states such as $H \to \gamma\gamma$. Future studies could investigate the observables in other reference frames [143]. Furthermore, while the STXS v1.2 framework is primarily designed for differential cross section measurements rather than CP studies, adopting its binning for the transverse momentum of the Higgs boson, choosing the bin edges at $p_{\mathrm{T},H} = [0, 60, 120, 300, 450, +\infty]$ GeV, and a binning for the other variables that assigns less densely populated regions in the tails of the variable distributions to wider bins, could be employed in future analyses.

Employing multivariate analysis methods by training dedicated gradient boosted tree classifiers for the CP-even and CP-odd hypotheses in the $tHjb$ and $t\bar{t}H$ signal processes provides an increased separation compared to the individual input variables. Despite the substantially larger training statistics available for $t\bar{t}H$, both classifiers achieve a comparable performance with an AUC of approximately 0.65. The impact of individual input variables, ranked according to the permutation importance metric and separation powers, differs considerably between the two processes. As the CP-sensitive observables were originally proposed for the $t\bar{t}H$ process and simply transferred to the $tHjb$ process with an adapted computation targeted towards the light spectator quark instead of the second top quark from the $t\bar{t}H$ process, this result suggests that some of the CP-observables might not work equally well for the $tHjb$ process.

The binary CP classifier trained on the $tHjb$ samples shows only limited discrimination in the $tHjb$ SR, while the equivalent $t\bar{t}H$ classifier in the $t\bar{t}H$ SR provides moderate separation for all signal samples. The HGNN performs similarly in the CP-even and CP-odd SR for both the $t\bar{t}H$ and $tHjb$ processes and a significant degradation in event classification performance is observed for the CP-mixed samples. The separations of the $tHjb$ signal from background and between signal samples are severely reduced in the CP-mixed cases. Since the CP-sensitive observables that were analysed provide only limited discrimination, an alternative approach would be to train an HGNN directly on the four-momenta of reconstructed objects, optionally providing the CP-observables as additional input. Such a network could simultaneously perform the classification in different processes and distinguish between CP-even and CP-odd hypotheses by creating two classes for each signal process, corresponding to a CP-even or CP-mixed coupling.

For the tWH signal process, the bachelor jet does not correspond to a physical object of interest in studying the CP-symmetry properties of the top-Higgs-Yukawa coupling. However, the similarity between the W boson and the second top quark in leading-order Feynman diagrams in Figures 2.5b and 2.6a suggests that the hadronically decaying W boson ($W \to q\bar{q}'$) could be employed to compute the CP-sensitive variables for the tWH process in place of the bachelor jet that was used to study the $tHjb$ signal.

Overall, the thesis analysed CP-sensitive observables and their usage with ML methods for probing the top-Higgs Yukawa coupling in tH and $t\bar{t}H$ production with $H \to \tau\tau$ in the full-hadronic final state. While the challenging event topology limits the performance of the current observables, the results identify several promising directions for future developments and demonstrate the potential of this channel for future analyses.

Bibliography

[1] P. W. Higgs, *Broken Symmetries, Massless Particles and Gauge Fields*, Phys. Lett. **12**, 132 (1964)

[2] F. Englert, R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*, Phys. Rev. Lett. **13**, 321 (1964)

[3] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, Phys. Rev. Lett. **13**, 508 (1964)

[4] G. S. Guralnik, C. R. Hagen, T. W. B. Kibble, *Global Conservation Laws and Massless Particles*, Phys. Rev. Lett. **13**, 585 (1964)

[5] P. W. Higgs, *Spontaneous Symmetry Breakdown without Massless Bosons*, Phys. Rev. **145**, 1156 (1966)

[6] T. W. B. Kibble, *Symmetry Breaking in Non-Abelian Gauge Theories*, Phys. Rev. **155**, 1554 (1967)

[7] M. B. Gavela, et al., *Standard model CP violation and baryon asymmetry*, Mod. Phys. Lett. A **9**, 795 (1994)

[8] P. Huet, E. Sather, *Electroweak baryogenesis and standard model CP violation*, Phys. Rev. D **51**, 379 (1995)

[9] S. L. Glashow, *Partial Symmetries of Weak Interactions*, Nucl. Phys. **22**, 579 (1961)

[10] S. Weinberg, *A Model of Leptons*, Phys. Rev. Lett. **19**, 1264 (1967)

[11] A. Salam, *Weak and Electromagnetic Interactions*, Almqvist & Wiksell, Stockholm, nobel symposium 8 edition (1968)

[12] S. L. Glashow, J. Iliopoulos, L. Maiani, *Weak Interactions with Lepton-Hadron Symmetry*, Phys. Rev. D **2**, 1285 (1970)

[13] H. Georgi, S. L. Glashow, *Unified Weak and Electromagnetic Interactions without Neutral Currents*, Phys. Rev. Lett. **28**, 1494 (1972)

[14] H. D. Politzer, *Reliable Perturbative Results for Strong Interactions*, Phys. Rev. Lett. **30**, 1346 (1973)

[15] H. D. Politzer, *Asymptotic Freedom: An Approach to Strong Interactions*, Phys. Rept. **14**, 129 (1974)

[16] D. J. Gross, F. Wilczek, *Asymptotically Free Gauge Theories*, Phys. Rev. D **8**, 3633 (1973)

[17] S. Weinberg, *The Making of the Standard Model*, Eur. Phys. J. C **34**, 5 (2004)

[18] G. 'tHooft, *Renormalization of massless Yang-Mills fields*, Nucl. Phys. B **33(1)**, 173 (1971)

[19] G. 't Hooft, *Renormalizable Lagrangians For Massive Yang-Mills Fields*, Nucl. Phys. B **35**, 167 (1971)

[20] G. 't Hooft, M. Veltmann, *Regularization And Renormalization Of Gauge Fields*, Nucl. Phys. B **44**, 189 (1972)

[21] G. 't Hooft, M. Veltmann, *Combinatorics of gauge fields*, Nucl. Phys. B **50**, 318 (1972)

[22] A. Quadt, *Top quark physics at hadron colliders*, Eur. Phys. J. **C48**, 835 (2006)

[23] H. Fritzsch, M. Gell-Mann, H. Leutwyler, *Advantages of the color octet gluon picture*, Phys. Lett. B **47(4)**, 365 (1973)

[24] M. Gell-Mann, *A Schematic Model of Baryons and Mesons*, Phys. Lett. **8**, 214 (1964)

[25] S. Tomonaga, *On a Relativistically Invariant Formulation of the Quantum Theory of Wave Fields*, Prog. Theor. Phys. **1(2)**, 27 (1946)

[26] J. Schwinger, *On Quantum-Electrodynamics and the Magnetic Moment of the Electron*, Phys. Rev. **73**, 416 (1948)

[27] J. Schwinger, *Quantum Electrodynamics. I. A Covariant Formulation*, Phys. Rev. **74**, 1439 (1948)

[28] R. P. Feynman, *Space-Time Approach to Quantum Electrodynamics*, Phys. Rev. **76**, 769 (1949)

[29] R. P. Feynman, *The Theory of Positrons*, Phys. Rev. **76**, 749 (1949)

[30] R. P. Feynman, *Mathematical Formulation of the Quantum Theory of Electromagnetic Interaction*, Phys. Rev. **80**, 440 (1950)

[31] F. J. Dyson, *The Radiation Theories of Tomonaga, Schwinger, and Feynman*, Phys. Rev. **75**, 486 (1949)

[32] F. J. Dyson, *The S Matrix in Quantum Electrodynamics*, Phys. Rev. **75**, 1736 (1949)

[33] R. L. Workman, et al. (Particle Data Group), *Review of Particle Physics*, PTEP **2022**, 083C01 (2022)

[34] S. Aulich, *Accessing Quantum Entanglement in off-shell $H^* \to ZZ$ with Separation of Polarisation States*, B.Sc. Thesis, Georg-August-University of Göttingen (2023), II.Physik-UniGö-BSc-2023/05

[35] M. E. Peskin, D. V. Schroeder, *An Introduction to quantum field theory*, Addison-Wesley, Reading, USA (1995)

[36] G. Sterman, *An Introduction to Quantum Field Theory*, Cambridge University Press (1993)

[37] S. Dawson, *Introduction to electroweak symmetry breaking*, in *ICTP Summer School in High-Energy Physics and Cosmology*, pages 1–83 (1998)

[38] F. Demartin, F. Maltoni, K. Mawatari, B. Page, M. Zaro, *Higgs characterisation at NLO in QCD: CP properties of the top-quark Yukawa interaction*, Eur. Phys. J. C **74(9)**, 3065 (2014), 1407.5089

[39] G. Merz, *Precision Measurements of Higgs Boson Couplings in the Diphoton Decay Channel with Run-2 of the ATLAS Detector*, Ph.D. Thesis, University of Michigan (2021)

[40] B. Bullard, et al., *Measurement of CP structure in the top quark Yukawa coupling in $t\bar{t}H$ and tH production in the $H \to \gamma\gamma$ decay channel with partial Run 3 dataset*, Internal Note, CERN, Geneva (2025)

[41] D. de Florian, et al. (LHC Higgs Cross Section Working Group), *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*, CERN Yellow Rep. Monogr. **2**, 1 (2017)

[42] LHC Higgs WG1, *Predictions for Production Cross Sections of the Higgs Boson at the LHC and HL-LHC* (LHCHWG-2026-005)

[43] R. D. Ball, et al. (PDF4LHCWorkingGroup), *The PDF4LHC21 combination of global PDF fits for the LHC Run III*, J. Phys. G **49**, 080501 (2022)

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

[45] F. A. Dreyer, A. Karlberg, *Vector-Boson Fusion Higgs Production at Three Loops in QCD*, Phys. Rev. Lett. **117(7)**, 072001 (2016)

[46] A. Denner, et al., *HAWK 2.0: A Monte Carlo program for Higgs production in vector-boson fusion and Higgs strahlung at hadron colliders*, Comput. Phys. Commun. **195**, 161 (2015)

[47] T. Liu, K. Melnikov, A. A. Penin, *Nonfactorizable QCD Effects in Higgs Boson Production via Vector Boson Fusion*, Phys. Rev. Lett. **123(12)**, 122002 (2019)

[48] K. Asteriadis, C. Brønnum-Hansen, K. Melnikov, *Nonfactorizable corrections to Higgs boson production in weak boson fusion*, Phys. Rev. D **109(1)**, 014031 (2024)

[49] F. Demartin, et al., *Higgs production in association with a single top quark at the LHC*, Eur. Phys. J. C **75(6)**, 267 (2015)

[50] F. Demartin, et al., *tWH associated production at the LHC*, Eur. Phys. J. C **77(1)**, 34 (2017)

[51] F. Takahashi, et al. (Particle Data Group), *Review of Particle Physics*, Int. J. Mod. Phys. A **41**, 2630011 (2026)

[52] F. Rosenblatt, *The perceptron - A perceiving and recognizing automaton*, Technical Report 85-460-1, Cornell Aeronautical Laboratory, Ithaca, New York (1957)

[53] F. Rosenblatt, *The perceptron: A probabilistic model for information storage and organization in the brain*, Psychol. Rev. **65**, 386 (1958)

[54] C. C. Aggarwal, *Neural Networks and Deep Learning: A Textbook*, Mathematics and Statistics, Springer Cham, 2nd edition (2023)

[55] D. P. Kingma, J. Ba, *Adam: A Method for Stochastic Optimization*, in *International Conference on Learning Representations (ICLR)* (2015)

[56] D. E. Rumelhart, G. E. Hinton, R. J. Williams, *Learning representations by back-propagating errors*, Nature **323(6088)**, 533 (1986)

[57] T. N. Kipf, M. Welling, *Semi-Supervised Classification with Graph Convolutional Networks*, International Conference on Learning Representations (2016), 1609.02907

[58] M. Defferrard, X. Bresson, P. Vandergheynst, *Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering*, Annual Conference on Neural Information Processing Systems (2016), 1606.09375

[59] Y. Feng, et al., *Hypergraph Neural Networks*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3558–3565 (2019)

[60] M. Adler, *Exploring hypergraph neural networks to employ multi-particle correlations in single-top-quarks associated Higgs-boson production via an auxiliary task to identify the correct top-quark decay products*, https://indico.cern.ch/event/1669803/contributions/7019836/attachments/3263557/5828109/Max_Adler%20(3).pdf (2026), Internal Presentation, Accessed: 2026-06-30

[61] M. Adler, *Studies of the top-Higgs Yukawa interaction via Higgs boson production in association with a single top quark in a final state with two tau leptons and jets using proton-proton collision data recorded with the ATLAS detector*, B.Sc. Thesis, Georg-August-University of Göttingen (2025), II. Physik-UniGö-BSc-2025/04

[62] J. H. Friedman, *Greedy function approximation: A gradient boosting machine.*, Ann. Stat. **29(5)**, 1189 (2001)

[63] T. Chen, C. Guestrin, *XGBoost: A Scalable Tree Boosting System*, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, ACM, New York, NY, USA (2016)

[64] O. S. Brüning, et al., *LHC Design Report Vol. 1: The LHC Main Ring*, CERN Yellow Rep. Monogr., CERN, Geneva (2004)

[65] The ATLAS Collaboration, *Public ATLAS Online Luminosity Plots for Run-3 of the LHC*, https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun3 (2025), accessed: 2026-03-26

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

[67] R. M. Bianchi, ATLAS Collaboration, *ATLAS experiment schematic or layout illustration*, https://cds.cern.ch/record/2837191 (2022), General Photo, Accessed: 2026-03-26

[68] The ATLAS Collaboration, *ATLAS magnet system: Technical Design Report*, Technical Report CERN-LHCC-97-018, ATLAS-TDR-6, CERN, Geneva (1997)

[69] H. A. Lorentz, *La théorie électromagnétique de Maxwell et son application aux corps mouvants*, Archives Néerlandaises des Sciences Exactes et Naturelles **25** (1892)

[70] The ATLAS Collaboration, *ATLAS Muon Spectrometer: Technical Design Report*, Technical Report CERN-LHCC-97-022, CERN, Geneva (1997)

[71] The ATLAS Collaboration, *Technical Design Report for the ATLAS Inner Tracker Pixel Detector*, Technical Report CERN-LHCC-2017-021, ATLAS-TDR-030, CERN, Geneva (2017)

[72] The ATLAS Collaboration, *ATLAS Pixel Detector Electronics and Sensors*, JINST **3**, P07007 (2008)

[73] A. Ahmad, et al., *The silicon microstrip sensors of the ATLAS semiconductor tracker*, Nucl. Instrum. Meth. A **578**, 98 (2007)

[74] B. Mindur, ATLAS Collaboration, *ATLAS Transition Radiation Tracker (TRT): Straw tubes for tracking and particle identification at the Large Hadron Collider*, Nucl. Instrum. Meth. A **845**, 257 (2017)

[75] N. Belyaev, et al., *Development of Transition Radiation Detectors for hadron identification at TeV energy scale*, J. Phys. Conf. Ser. **1390(1)**, 012126 (2019)

[76] H. Kolanoski, N. Wermes, *Particle Detectors*, Oxford University Press (2020)

[77] The ATLAS Collaboration, *ATLAS Liquid Argon Calorimeter: Technical Design Report*, Technical report, CERN, Geneva (1996)

[78] The ATLAS Collaboration (ATLAS), *ATLAS Barrel Toroid: Technical Design Report*, Technical report, CERN, Geneva (1997)

[79] The ATLAS Collaboration (ATLAS), *ATLAS Endcap Toroids: Technical Design Report*, Technical report, CERN, Geneva (1997)

[80] T. Kawamoto, et al., *New Small Wheel Technical Design Report*, Technical report, CERN, Geneva (2013), CERN-LHCC-2013-006

[81] The ATLAS Collaboration, *The ATLAS trigger system for LHC Run 3 and trigger performance in 2022*, JINST **19(06)**, 6 (2024)

[82] The ATLAS Collaboration, *Preliminary analysis of the luminosity calibration for the ATLAS 13.6 TeV data recorded in 2023*, ATLAS Note, CERN, Geneva (2024)

[83] The ATLAS Collaboration, *Preliminary luminosity calibration of the ATLAS 13.6 TeV data recorded in 2024 and combination with the 2022 and 2023 measurements*, ATLAS Note, CERN, Geneva (2025)

[84] P. Golonka, Z. Was, *PHOTOS Monte Carlo: A Precision tool for QED corrections in Z and W decays*, Eur. Phys. J. C **45**, 97 (2006)

[85] S. Frixione, P. Nason, G. Ridolfi, *A Positive-weight next-to-leading-order Monte Carlo for heavy flavour hadroproduction*, JHEP **09**, 126 (2007)

[86] P. Nason, *A New method for combining NLO QCD with shower Monte Carlo algorithms*, JHEP **11**, 40 (2004)

[87] S. Frixione, P. Nason, C. Oleari, *Matching NLO QCD computations with Parton Shower simulations: the POWHEG method*, JHEP **11**, 70 (2007)

[88] S. Alioli, et al., *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, JHEP **06**, 43 (2010)

[89] H. B. Hartanto, et al., *Higgs boson production in association with top quarks in the POWHEG BOX*, Phys. Rev. D **91(9)**, 094003 (2015)

[90] R. D. Ball, et al. (NNPDF), *Parton distributions for the LHC Run II*, JHEP **04**, 40 (2015)

[91] C. Bierlich, et al., *A comprehensive guide to the physics and usage of PYTHIA 8.3*, SciPost Phys. Codeb. **2022**, 8 (2022)

[92] The ATLAS Collaboration, *ATLAS Pythia 8 tunes to 7 TeV data*, ATLAS Note, CERN, Geneva (2014), ATL-PHYS-PUB-2014-021

[93] H.-L. Lai, et al., *New parton distributions for collider physics*, Phys. Rev. D **82**, 074024 (2010)

[94] E. Bothmann, et al. (Sherpa), *Event Generation with Sherpa 2.2*, SciPost Phys. **7(3)**, 34 (2019)

[95] S. Hoeche, et al., *A critical appraisal of NLO+PS matching methods*, JHEP **09**, 49 (2012)

[96] S. Hoeche, et al., *QCD matrix elements + parton showers: The NLO case*, JHEP **04**, 27 (2013)

[97] S. Catani, et al., *QCD matrix elements + parton showers*, JHEP **11**, 63 (2001)

[98] S. Hoeche, et al., *QCD matrix elements and truncated showers*, JHEP **05**, 53 (2009)

[99] T. Gleisberg, S. Hoeche, *Comix, a new matrix element generator*, JHEP **12**, 39 (2008)

[100] F. Buccioni, et al., *OpenLoops 2*, Eur. Phys. J. C **79(10)**, 866 (2019)

[101] F. Cascioli, P. Maierhofer, S. Pozzorini, *Scattering Amplitudes with Open Loops*, Phys. Rev. Lett. **108**, 111601 (2012)

[102] A. Denner, S. Dittmaier, L. Hofer, *Collier: a fortran-based Complex One-Loop Library in Extended Regularizations*, Comput. Phys. Commun. **212**, 220 (2017)

[103] C. Anastasiou, et al., *High precision QCD at hadron colliders: Electroweak gauge boson rapidity distributions at NNLO*, Phys. Rev. D **69**, 094008 (2004)

[104] The ATLAS Collaboration, *Studies on top-quark Monte Carlo modelling for Top2016*, ATLAS Note, CERN, Geneva (2016)

[105] D. J. Lange, *The EvtGen particle decay simulation package*, Nucl. Instrum. Meth. A **462**, 152 (2001)

[106] T. Sjostrand, S. Mrenna, P. Z. Skands, *A Brief Introduction to PYTHIA 8.1*, Comput. Phys. Commun. **178**, 852 (2008)

[107] The ATLAS Collaboration, *The Pythia 8 A3 tune description of ATLAS minimum bias and inelastic measurements incorporating the Donnachie-Landshoff diffractive model*, ATLAS Note, CERN, Geneva (2016)

[108] A. D. Martin, et al., *Parton distributions for the LHC*, Eur. Phys. J. C **63**, 189 (2009)

[109] The ATLAS Collaboration, *The ATLAS Simulation Infrastructure*, Eur. Phys. J. C **70**, 823 (2010)

[110] S. Agostinelli, et al. (GEANT4), *GEANT4 - A Simulation Toolkit*, Nucl. Instrum. Meth. A **506**, 250 (2003)

[111] The ATLAS Collaboration, *TopCPToolkit*, v3.2.0, Accessed: 2026-07-01, URL https://doi.org/10.5281/zenodo.19683083

[112] M. Cacciari, G. P. Salam, G. Soyez, *The anti-k_t jet clustering algorithm*, JHEP **04**, 63 (2008)

[113] M. Cacciari, G. P. Salam, G. Soyez, *FastJet User Manual*, Eur. Phys. J. C **72**, 1896 (2012)

[114] The ATLAS Collaboration, *Jet reconstruction and performance using particle flow with the ATLAS Detector*, Eur. Phys. J. C **77(7)**, 466 (2017)

[115] The ATLAS Collaboration, *Jet energy scale and resolution measured in proton–proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Eur. Phys. J. C **81(8)**, 689 (2021)

[116] The ATLAS Collaboration, *Performance of pile-up mitigation techniques for jets in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector*, Eur. Phys. J. C **76(11)**, 581 (2016)

[117] The ATLAS Collaboration, *Identification and rejection of pile-up jets at high pseudorapidity with the ATLAS detector*, Eur. Phys. J. C **77(9)**, 580 (2017), [Erratum: Eur.Phys.J.C 77, 712 (2017)]

[118] The ATLAS Collaboration, *Transforming jet flavour tagging at ATLAS*, Nature Commun. **17(1)**, 541 (2026)

[119] A. Vaswani, et al., *Attention Is All You Need*, in *Advances in Neural Information Processing Systems*, volume 30, Curran Associates, Inc. (2017)

[120] The ATLAS Collaboration, *Electron and photon efficiencies in LHC Run 2 with the ATLAS experiment*, JHEP **05**, 162 (2024)

[121] The ATLAS Collaboration, *Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton-proton collision data at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. C **79(8)**, 639 (2019)

[122] The ATLAS Collaboration, *Muon reconstruction and identification efficiency in ATLAS using the full Run 2 pp collision data set at* $\sqrt{s} = 13$ *TeV*, Eur. Phys. J. C **81(7)**, 578 (2021)

[123] The ATLAS Collaboration, *Measurements of Higgs boson production cross-sections in the* $H \to \tau^+\tau^-$ *decay channel in pp collisions at* $\sqrt{s}$ *= 13 TeV with the ATLAS detector*, JHEP **08**, 175 (2022)

[124] The ATLAS Collaboration, *The performance of missing transverse momentum reconstruction and its significance with the ATLAS detector using 140 fb*$^{-1}$ *of* $\sqrt{s} = 13$ *TeV pp collisions*, Eur. Phys. J. C **85(6)**, 606 (2025)

[125] A. Elagin, et al., *A New Mass Reconstruction Technique for Resonances Decaying to di-tau*, Nucl. Instrum. Meth. A **654**, 481 (2011)

[126] J. Ellis, et al., *Disentangling Higgs-Top Couplings in Associated Production*, JHEP **04**, 4 (2014)

[127] X.-G. He, G.-N. Li, Y.-J. Zheng, *Probing Higgs boson CP Properties with* $t\bar{t}H$ *at the LHC and the 100 TeV pp collider*, Int. J. Mod. Phys. A **30(25)**, 1550156 (2015)

[128] F. Demartin, et al., *Higgs production in association with a single top quark at the LHC*, Eur. Phys. J. C **75(6)**, 267 (2015)

[129] F. Demartin, et al., *tWH associated production at the LHC*, Eur. Phys. J. C **77(1)**, 34 (2017)

[130] R. K. Barman, D. Gonçalves, F. Kling, *Machine learning the Higgs boson-top quark CP phase*, Phys. Rev. D **105(3)**, 035023 (2022)

[131] D. Azevedo, et al., *CP-violation, asymmetries and interferences in* $t\bar{t}\phi$, JHEP **09**, 246 (2022)

[132] J. Ackerschott, et al., *Returning CP-observables to the frames they belong*, SciPost Phys. **17(1)**, 1 (2024)

[133] H. Bahl, S. Brass, *Constraining* $\mathcal{CP}$*-violation in the Higgs-top-quark interaction using machine-learning-based inference*, JHEP **03**, 17 (2022)

[134] T. Martini, et al., *Probing the CP structure of the top quark Yukawa coupling: Loop sensitivity versus on-shell sensitivity*, Phys. Rev. D **104(5)**, 055045 (2021)

[135] A. V. Gritsan, et al., *Constraining anomalous Higgs boson couplings to the heavy flavor fermions using matrix element techniques*, Phys. Rev. D **94(5)**, 055023 (2016)

[136] D. Gonçalves, K. Kong, J. H. Kim, *Probing the top-Higgs Yukawa CP structure in dileptonic $t\bar{t}h$ with M_2-assisted reconstruction*, JHEP **06**, 79 (2018)

[137] J. F. Gunion, X.-G. He, *Determining the CP nature of a neutral Higgs boson at the LHC*, Phys. Rev. Lett. **76**, 4468 (1996)

[138] D. Azevedo, et al., *CP tests of Higgs couplings in $t\bar{t}h$ semileptonic events at the LHC*, Phys. Rev. D **98(3)**, 033004 (2018)

[139] Q.-H. Cao, et al., *A New Observable for Measuring CP Property of Top-Higgs Interaction*, Chin. Phys. C **45(2)**, 023117 (2021)

[140] J. Yue, *Enhanced thj signal at the LHC with $h \to \gamma\gamma$ decay and $\mathcal{CP}$-violating top-Higgs coupling*, Phys. Lett. B **744**, 131 (2015)

[141] F. Demartin, et al., *Higgs characterisation at NLO in QCD: CP properties of the top-quark Yukawa interaction*, Eur. Phys. J. C **74(9)**, 3065 (2014)

[142] M. R. Buckley, D. Goncalves, *Boosting the Direct CP Measurement of the Higgs-Top Coupling*, Phys. Rev. Lett. **116(9)**, 091801 (2016)

[143] H. Bahl, et al., *$\mathcal{CP}$-sensitive simplified template cross-sections for $t\bar{t}H$*, JHEP **10**, 214 (2024)

[144] The ATLAS Collaboration, *FastFrames*, v6.1.0, Accessed: 2026-07-01, URL `https://doi.org/10.5281/zenodo.14773274`

[145] The ATLAS Collaboration, *TRExFitter*, v1.9.0, Accessed: 2026-07-01, URL `https://doi.org/10.5281/zenodo.14845712`

[146] T. Chen, et al., *XGBoost: Extreme Gradient Boosting* (2026), PyPI package v2.1.4, Accessed: 2026-06-28, URL `https://github.com/dmlc/xgboost`

[147] XGBoost Developers, *XGBoost Parameters*, `https://xgboost.readthedocs.io/en/stable/parameter.html` (2026), accessed: 2026-06-28

[148] XGBoost Developers, *Python API Reference*, `https://xgboost.readthedocs.io/en/stable/python/python_api.html` (2026), accessed: 2026-06-28

[149] The ATLAS Collaboration, *CP Properties of Higgs Boson Interactions with Top Quarks in the $t\bar{t}H$ and tH Processes Using $H \to \gamma\gamma$ with the ATLAS Detector*, Phys. Rev. Lett. **125(6)**, 061802 (2020)

A. List of Dataset Identifiers

Table A.1.: Summary of all MC samples used in this analysis and their DSIDs grouped according to the categories used for plotting, with the mass of the top quark and SM Higgs boson set to 172.5 GeV and 125 GeV, respectively.

Category	Process	DSID	Generator
$t\bar{t}H$ (SM)	$t\bar{t}H$	6034(19–21)	Powheg Box v06-03 + Pythia 8.310 + Evtgen v2.2.1
$tHjb$ (SM)	$tHjb$	546585	Madgraph5_aMC@NLO + Pythia 8.309 + Evtgen v2.1.1
tWH (SM)	tWH	546586	Madgraph5_aMC@NLO + Pythia 8.309 + Evtgen v2.1.1
CP-mixed (HC)	$t\bar{t}H$	5746(35–41)	Madgraph5_aMC@NLO + Pythia 8.308 + Evtgen v2.2.1
	$tHjb$	5746(53–59)	Madgraph5_aMC@NLO + Pythia 8.308 + Evtgen v2.2.1
	tWH	5746(42–48	Madgraph5_aMC@NLO + Pythia 8.308 + Evtgen v2.2.1
top (dilep.)	$t\bar{t}$ (dilep.)	601230	Powheg Box v2 + Pythia 8.308 + Evtgen v2.1.1
	tW (dilep.)	6013(53–54)	Powheg Box v2 + Pythia 8.308 + Evtgen v2.1.1
top fakes	$t\bar{t}$ (allhad.)	601237	Powheg Box v2 + Pythia 8.308 + Evtgen v2.1.1
	$t\bar{t}$ (single lep.)	601229	Powheg Box v2 + Pythia 8.308 + Evtgen v2.1.1
single top	t-ch. (tb lep.)	6013(48–59)	Powheg Box v2 + Pythia 8.308 + Evtgen v2.1.1
	t-ch. (tqb lep.)	6013(50–51)	Powheg Box v2 + Pythia 8.308 + Evtgen v2.1.1
$t\bar{t}V$	$t\bar{t}Z/\gamma^*(\ell\ell)$	5220(24,28,32)	Madgraph5_aMC@NLO + Pythia 8.309 + Evtgen v2.1.1
	$t\bar{t}Z/\gamma^*(qq)$	522036	Madgraph5_aMC@NLO + Pythia 8.309 + Evtgen v2.1.1
	$t\bar{t}Z/\gamma^*(\nu\nu)$	522040	Madgraph5_aMC@NLO + Pythia 8.309 + Evtgen v2.1.1
	$t\bar{t}W$	700997	Sherpa 2.2.14
other Higgs	$qq \to V(\text{incl.})H(\tau\tau)$	6034(26–28)	Powheg Box v06-03 + Pythia 8.310 + Evtgen v2.2.1
	$gg \to Z(\text{incl.})H(\tau\tau)$	603418	Powheg Box v06-03 + Pythia 8.310 + Evtgen v2.2.1
	ggF $H(\tau\tau)$	603414	Powheg Box v06-03 + Pythia 8.310 + Evtgen v2.2.1
	VBF $H(\tau\tau)$	603422	Powheg Box v06-03 + Pythia 8.310 + Evtgen v2.2.1
$Z \to \tau^+\tau^-$	$Z/\gamma^*(\tau\tau)$	7007(92–94)	Sherpa 2.2.14
	$Z/\gamma^*(\tau\tau)$ with $10 < m_{Z/\gamma^*}[\text{GeV}] < 40$	7009(01–03)	Sherpa 2.2.14
other fakes	$Z/\gamma^*(ee/\mu\mu)$	7007(86-91)	Sherpa 2.2.14
	$Z/\gamma^*(ee/\mu\mu)$ with $10 < m_{Z/\gamma^*}[\text{GeV}] < 40$	700(895–900)	Sherpa 2.2.14
	$W^\pm(\ell\nu)$	7007(77-85)	Sherpa 2.2.14
	dijet prod.	6017(00,02,06)	Powheg Box v05-05 + Pythia 8.308 + Evtgen v2.1.1
	dijet prod.	6017(01,03,05,07)	Powheg Box v06-02 + Pythia 8.308 + Evtgen v2.1.1
$VV + VVV$	ZZ	701(085,090,095,100)	Sherpa 2.2.14
	WZ/WW	7011(05,10,15,20,25)	Sherpa 2.2.14
	$gg \to ZZ \to 4\ell$	603293	Powheg Box + Pythia 8.314 + Evtgen v2.2.1

B. Kinematic Variable Distributions in all SRs and CRs of the Event Categorisation

B.1. Distributions of p_T and η in the $tHjb$ SR

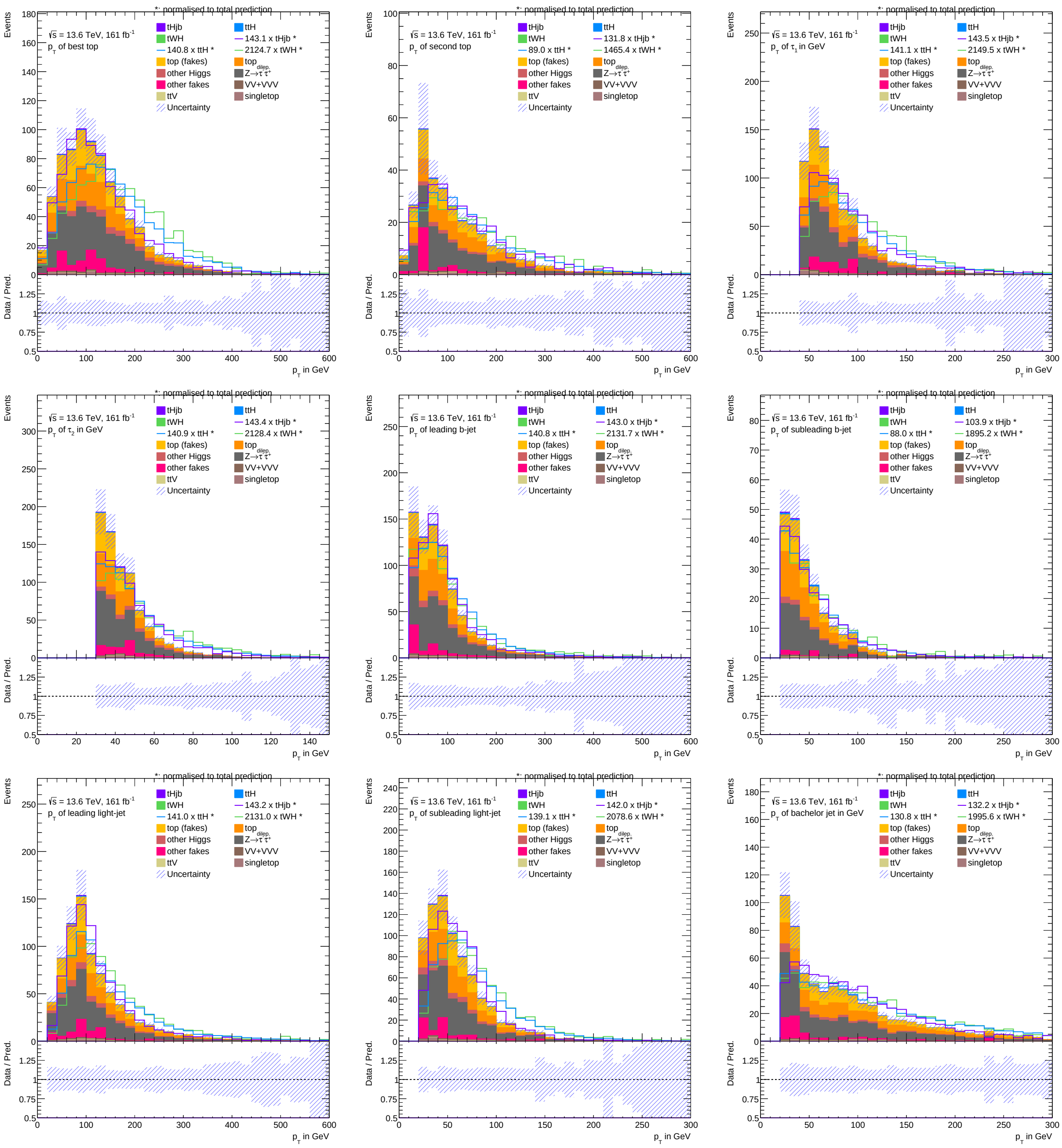


Figure B.1.: Transverse momentum p_T of the best and second-best top quark candidate determined by the HGNN, the leading (τ_1) and sub-leading (τ_2) τ lepton, b-jet and light-jet, and the bachelor jet in the $tHjb$ SR defined in Section 6.3. The analysis selection is applied.

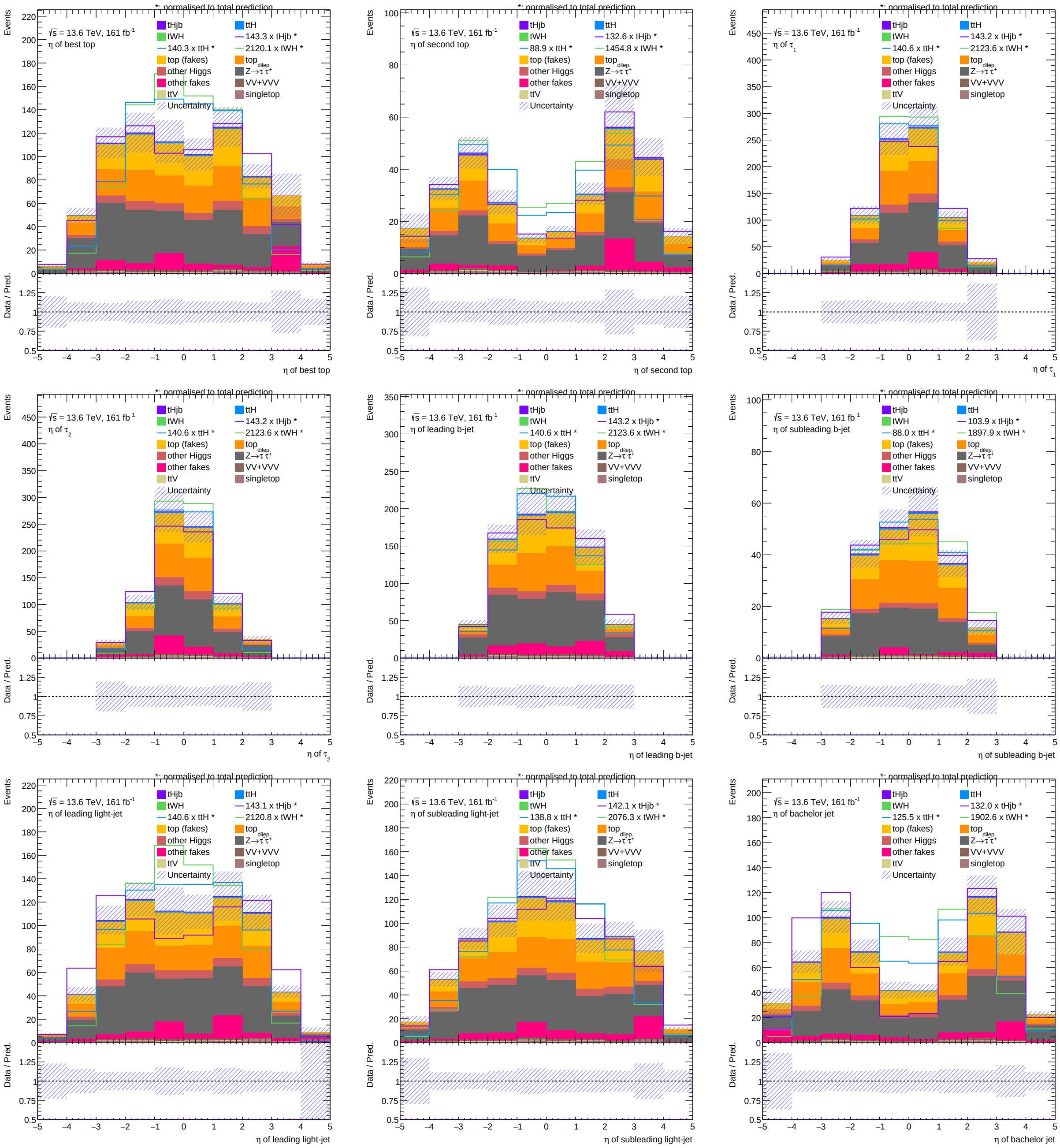


Figure B.2.: Pseudorapidity η of the best and second-best top quark candidate determined by the HGNN, the leading (τ_1) and sub-leading (τ_2) τ lepton, b-jet and light-jet, and the bachelor jet in the $tHjb$ SR defined in Section 6.3. The analysis selection is applied.

B.2. Distributions of p_{T} and η in the $t\bar{t}H$ SR

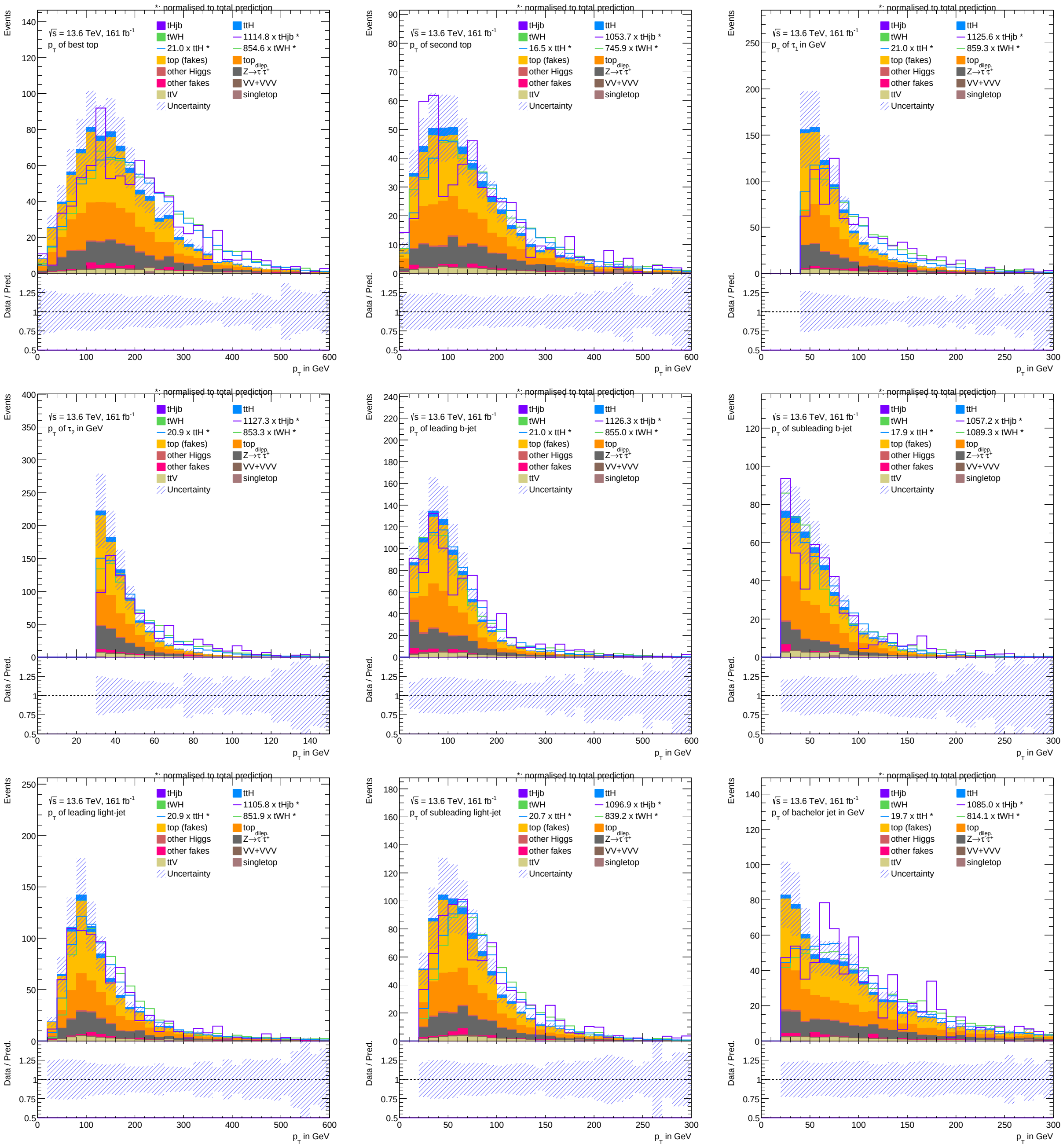


Figure B.3.: Transverse momentum p_{T} of the best and second-best top quark candidate determined by the HGNN, the leading (τ_1) and sub-leading (τ_2) τ lepton, b-jet and light-jet, and the bachelor jet in the $t\bar{t}H$ SR defined in Section 6.3. The analysis selection is applied.

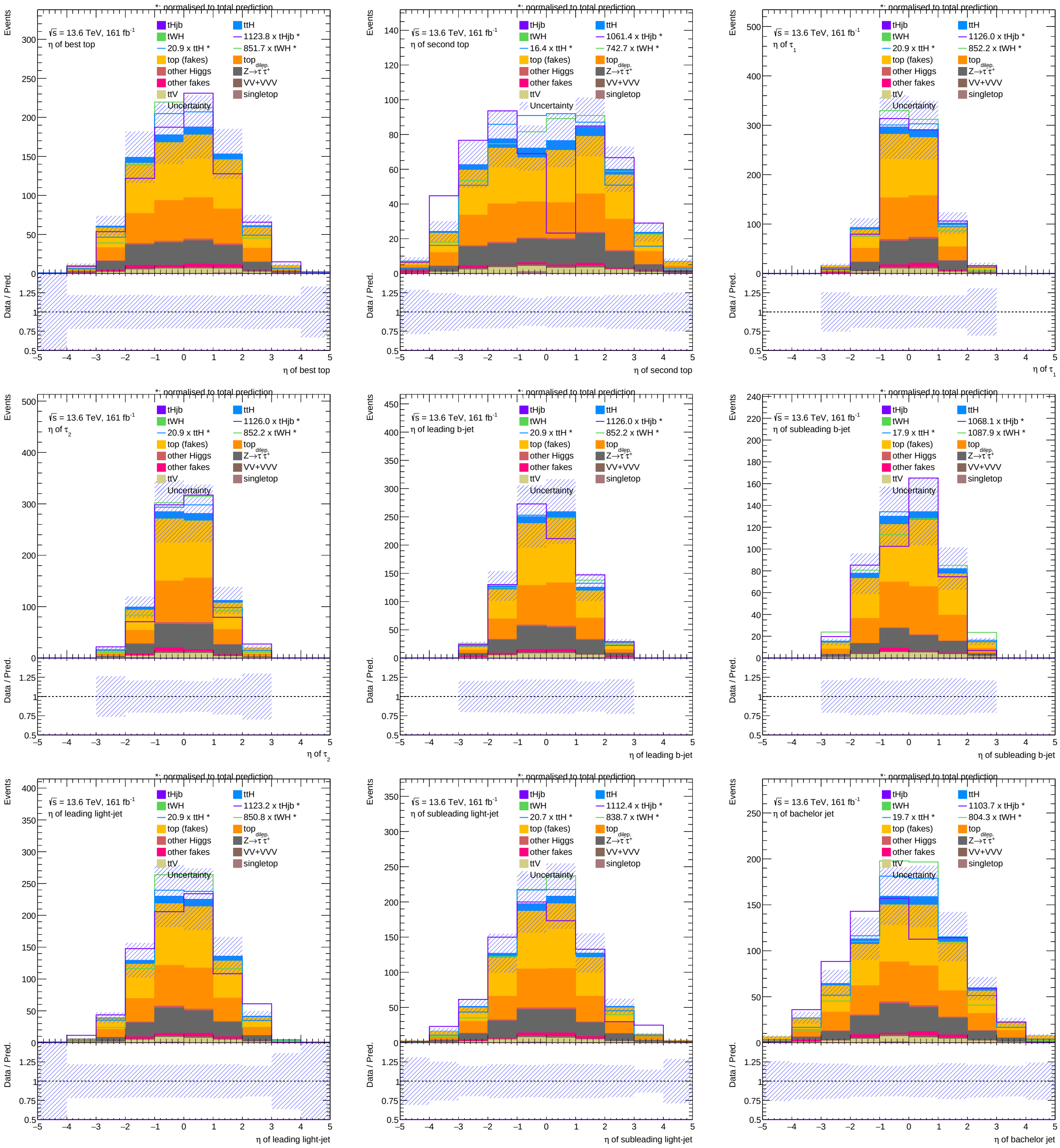


Figure B.4.: Pseudorapidity η of the best and second-best top quark candidate determined by the HGNN, the leading (τ_1) and sub-leading (τ_2) τ lepton, b-jet and light-jet, and the bachelor jet in the $t\bar{t}H$ SR defined in Section 6.3. The analysis selection is applied.

B.3. Distributions of p_{T} and η in the $Z(\tau\tau)$ CR

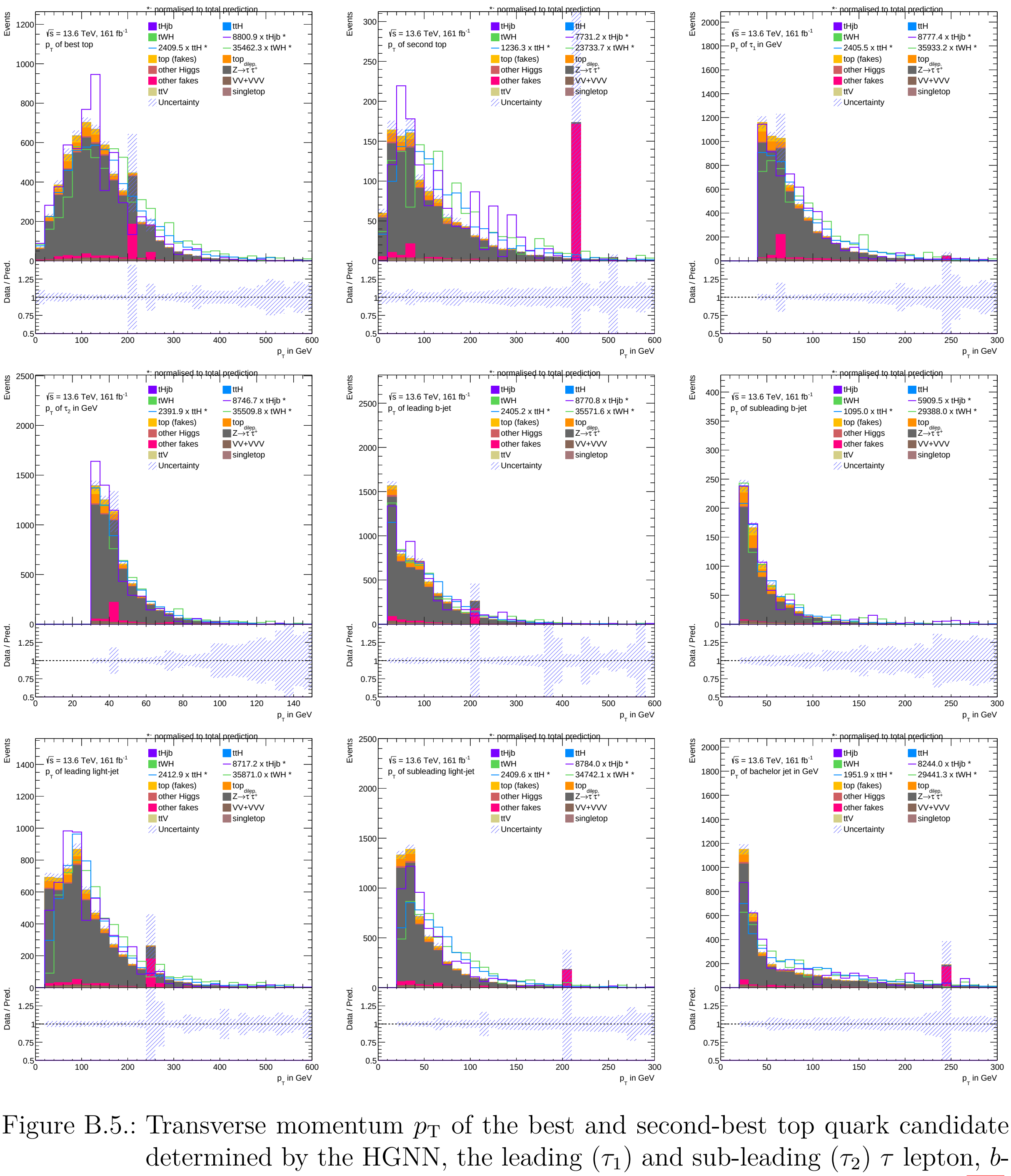


Figure B.5.: Transverse momentum p_{T} of the best and second-best top quark candidate determined by the HGNN, the leading (τ_1) and sub-leading (τ_2) τ lepton, b-jet and light-jet, and the bachelor jet in the $Z(\tau\tau)$ CR defined in Section 6.3. The analysis selection is applied.

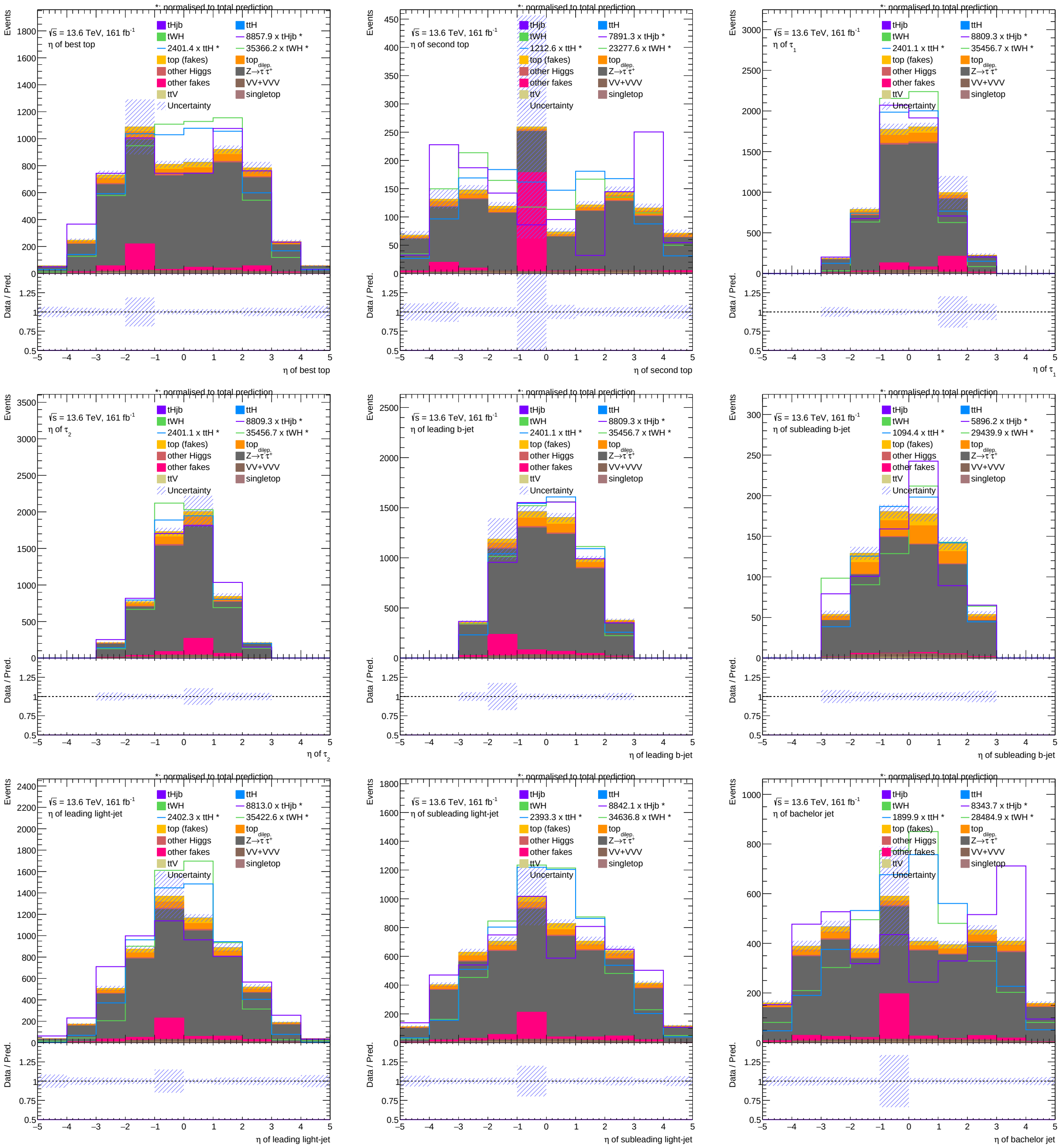


Figure B.6.: Pseudorapidity η of the best and second-best top quark candidate determined by the HGNN, the leading (τ_1) and sub-leading (τ_2) τ lepton, b-jet and light-jet, and the bachelor jet in the $Z(\tau\tau)$ CR defined in Section 6.3. The analysis selection is applied.

B.4. Distributions of p_T and η in the dileptonic top background ($t\bar{t}$ and tW) CR

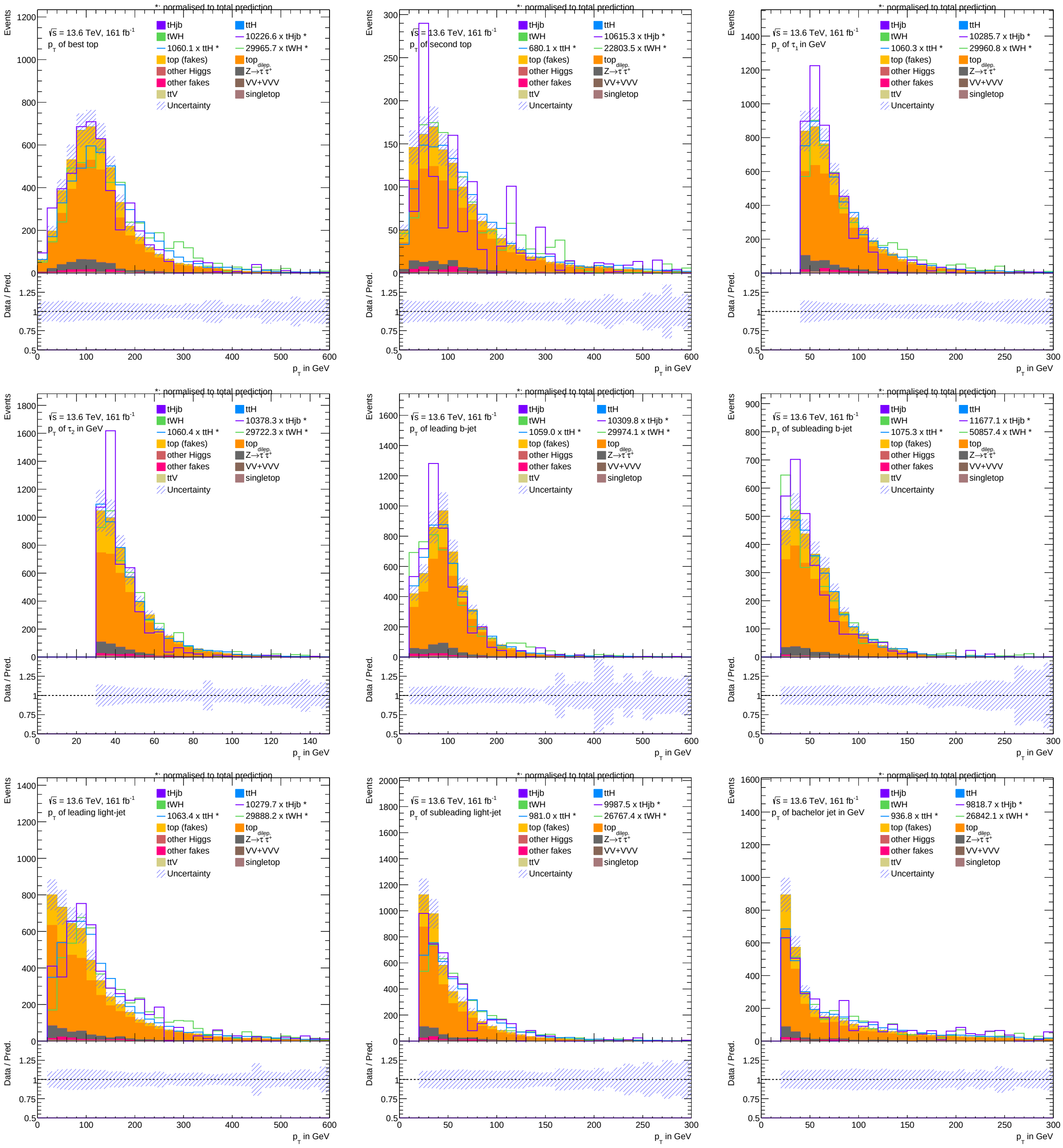


Figure B.7.: Transverse momentum p_T of the best and second-best top quark candidate determined by the HGNN, the leading (τ_1) and sub-leading (τ_2) τ lepton, b-jet and light-jet, and the bachelor jet in the dileptonic $t\bar{t}/tW$ CR defined in Section 6.3. The analysis selection is applied.

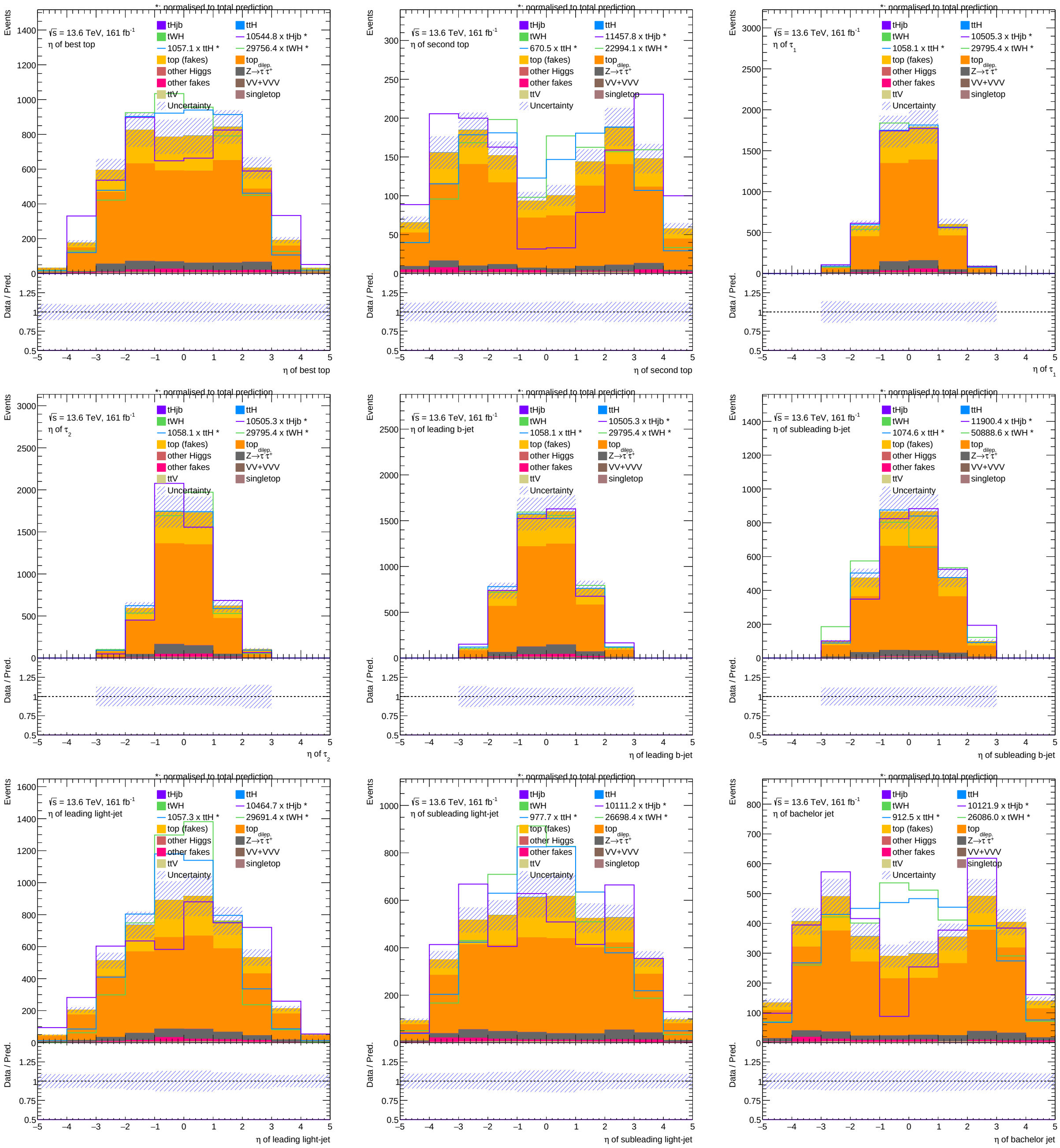


Figure B.8.: Pseudorapidity η of the best and second-best top quark candidate determined by the HGNN, the leading (τ_1) and sub-leading (τ_2) τ lepton, b-jet and light-jet, and the bachelor jet in the dileptonic $t\bar{t}/tW$ CR defined in Section 6.3. The analysis selection is applied.

C. Migration Matrices for CP-sensitive Variables in the $t\bar{t}H$ Signal Process

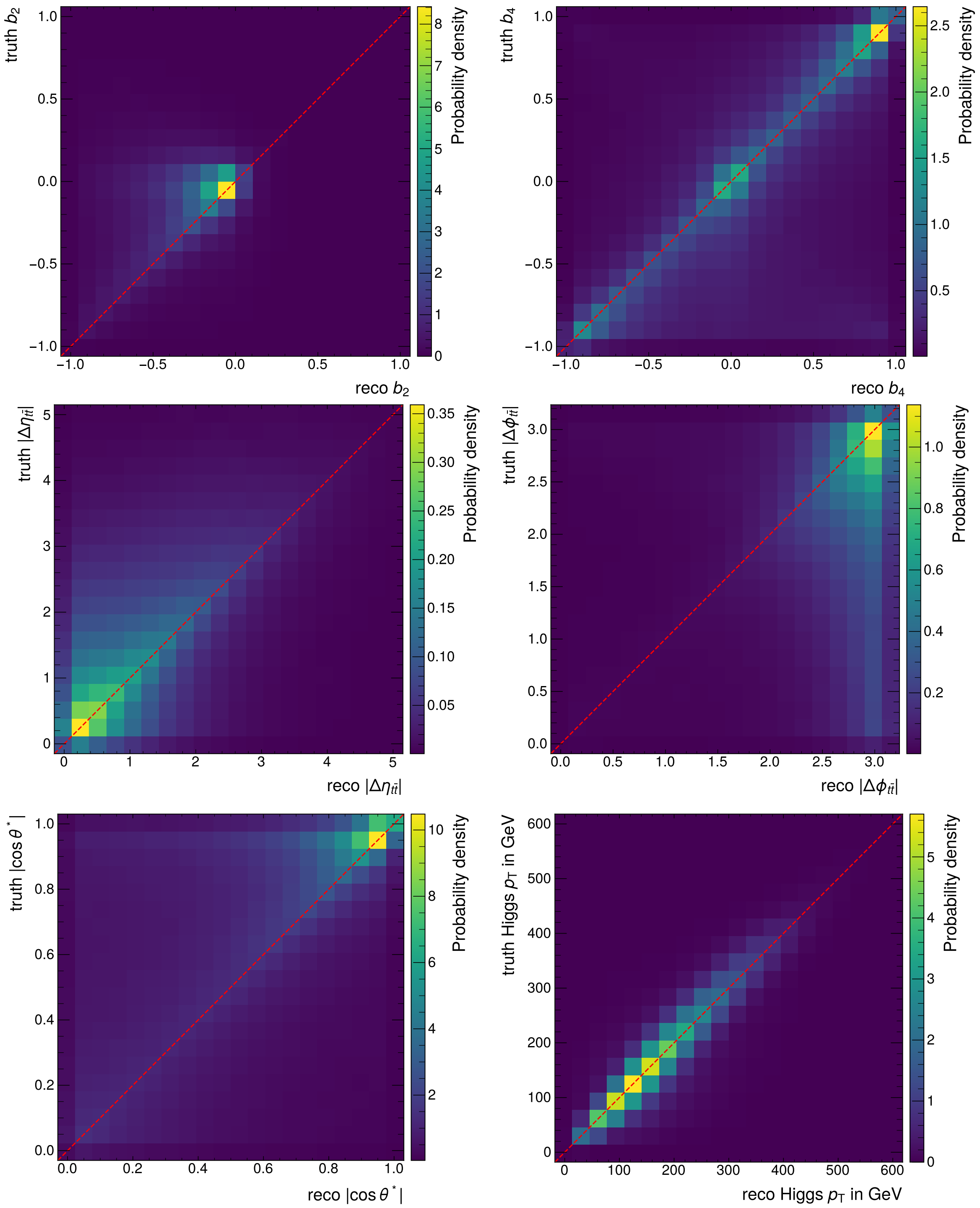


Figure C.1.: Migration matrices illustrating the bin-by-bin migration from truth to reconstructed bin of the CP-sensitive observables from Table 7.1 computed between the top and anti-top in the $t\bar{t}H$ signal process using MC-generated samples with different CP mixing angles between 0° and 90°.

D. Additional Information on the CP Binary Classification

D.1. Monte Carlo Statistics available for Training

Table D.1.: Number of events kept after the event selection described in Section 6.2 and the cleaning procedure (removing events with an unreconstructed bachelor jet or top quark, and events with negative MC weights. Also shown is the signed sum of the full event weights, including the cross section and branching ratio, corresponding to the expected number of events 2022-2024 LHC Run 3 ATLAS data. For the $tHjb$ sample with $\alpha = 45°$ where only the MC23a and MC23d campaigns (data-taking conditions in 2022 and 2023) are available, the corresponding value is scaled by the ratio of integrated luminosities to estimate the full 2022-2024 statistics and shown in brackets.

sample	**number of events**	**sum of signed weights**
$tHjb$, $\alpha = 0°$	39405	16.17
$tHjb$, $\alpha = 15°$	20495	17.27
$tHjb$, $\alpha = 30°$	8603	22.36
$tHjb$, $\alpha = 45°$	7788	10.43 (32.57)
$tHjb$, $\alpha = 60°$	28373	38.23
$tHjb$, $\alpha = 75°$	29031	53.73
$tHjb$, $\alpha = 90°$	28612	72.18
$t\bar{t}H$, $\alpha = 0°$	139987	44.72
$t\bar{t}H$, $\alpha = 15°$	83355	52.26
$t\bar{t}H$, $\alpha = 30°$	86122	48.18
$t\bar{t}H$, $\alpha = 45°$	28727	44.29
$t\bar{t}H$, $\alpha = 60°$	81141	33.97
$t\bar{t}H$, $\alpha = 75°$	112976	36.95
$t\bar{t}H$, $\alpha = 90°$	37820	12.00

D.2. Parameters employed in the Training of all CP Binary Classifiers

Table D.2.: Settings of the parameters used to describe the binary classifier based on the XGBoost package [146]. Explanations on the parameters and default values for all possible parameters are listed in the documentation [147, 148]. For parameters that are not listed in this table, the default values are used.

parameter	value	description
`booster`	`gbtree`	type of booster used for training (gradient-boosted decision trees)
`objective`	`binary:logistic`	objective function for binary classification with probabilistic output
`eta`	0.03	learning rate controlling the contribution of each tree
`gamma`	1	minimum loss reduction required to split a node
`max_depth`	3	maximum depth of individual trees
`min_child_weight`	5	minimum sum of instance weights required in a child node
`n_estimators`	5000	maximum number of trees
`subsample`	0.8	fraction of training events sampled for each tree
`colsample_bytree`	0.8	fraction of input features sampled for each tree
`reg_alpha`	1	L1 regularisation strength on leaf weights
`reg_lambda`	5	L2 regularisation strength on leaf weights
`tree_method`	hist	histogram-based tree construction algorithm
`random_state`	42	seed for reproducible random number generation.
`early_stopping_rounds`	200	stops training if the validation metric does not improve for 200 rounds.

E. Distinguishing the CP-even and maximally CP-mixed Scenario in the $tHjb$ and $t\bar{t}H$ Process

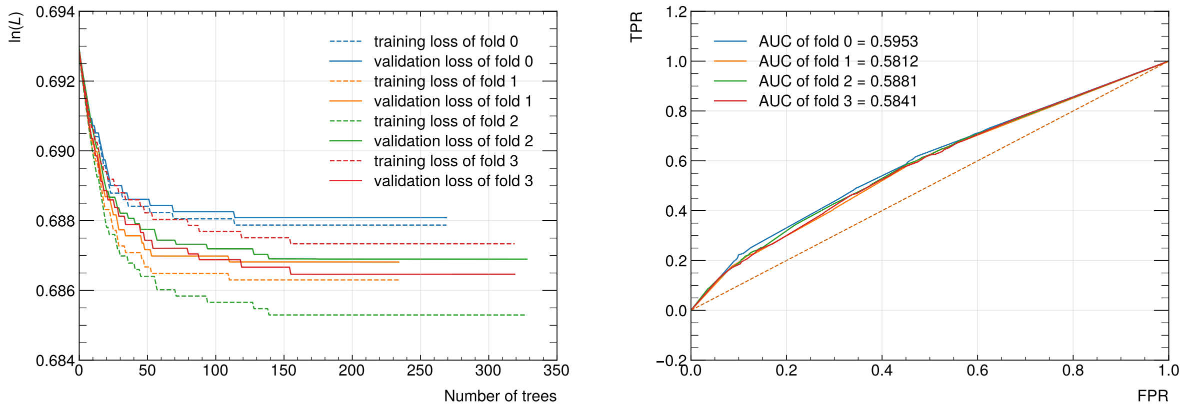


Figure E.1.: Logarithm $\ln(L)$ of the training and validation loss plotted against the number of trees (left) and ROC curve (right) for each fold of the binary classifier trained on $tHjb$ samples for CP-even (mixing angle 0°) and maximally CP-mixed (45°) top-Higgs-Yukawa coupling.

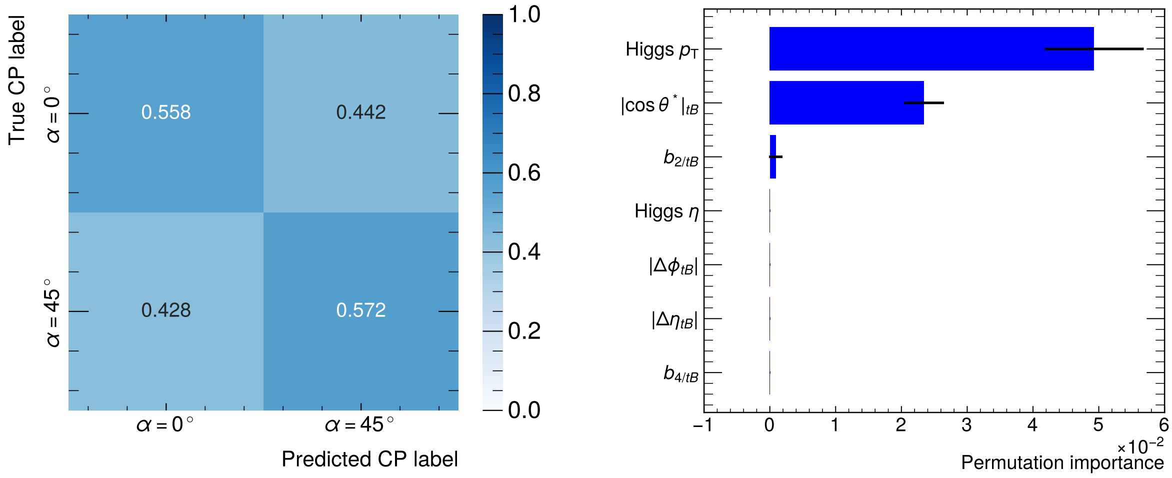
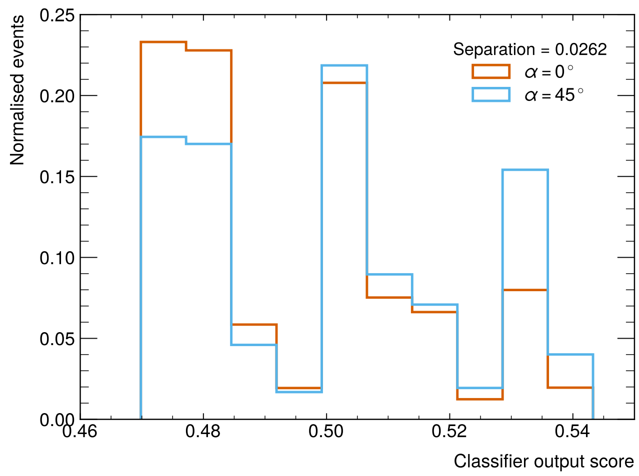


Figure E.2.: Confusion matrix normalised per row, employing a cut of 0.5 (left) on the output of the binary CP-even (mixing angle 0°) vs maximally CP-mixed (45°) classifier for the $tHjb$ signal, and ranked permutation importance of its input features (right).



Variable	Separation
output score	2.616%
$p_{T,H}$	1.907%
$\lvert\cos\theta^*\rvert$	1.634%
b_2	1.506%
b_4	1.235%
$\Delta\eta_{tB}$	1.126%
η_H	1.065%
$\Delta\phi_{tB}$	0.243%

Figure E.3.: Normalised number of events plotted against the output score of the CP-even (mixing angle 0°) vs maximally CP-mixed (45°) classifier (left) and ranked separation power of the CP-sensitive observables listed in Table 7.1 and the output score (right) for the $tHjb$ signal.

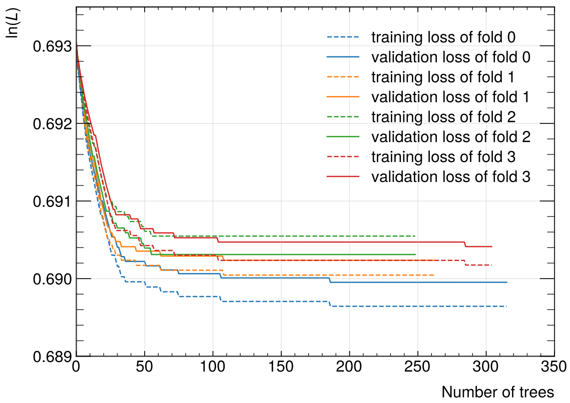


Figure E.4.: Logarithm $\ln(L)$ of the training and validation loss plotted against the number of trees (left) and ROC curve (right) for each fold of the binary classifier trained on $t\bar{t}H$ samples for CP-even (mixing angle 0°) and maximally CP-mixed (45°) top-Higgs-Yukawa coupling.

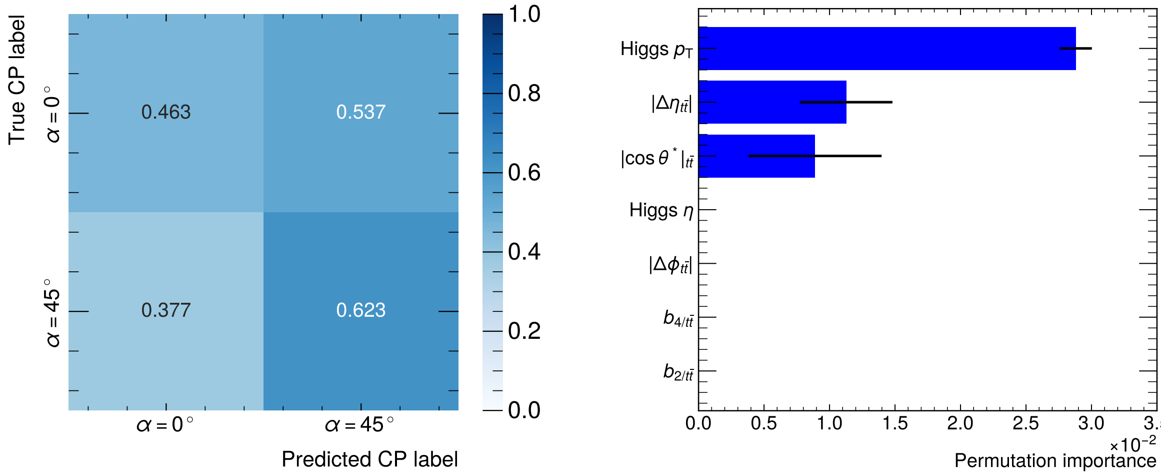
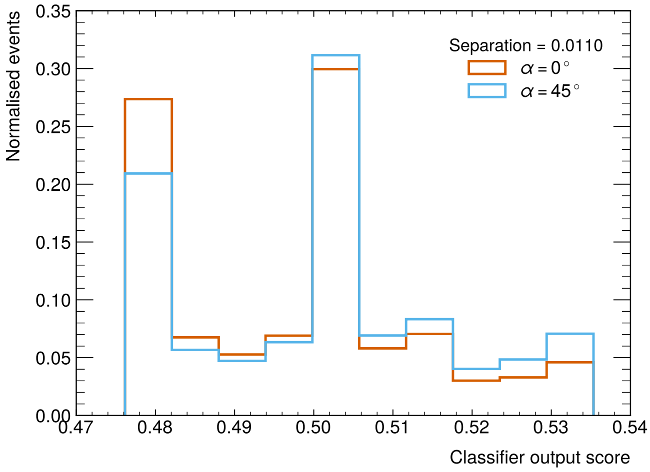


Figure E.5.: Confusion matrix normalised per row, employing a cut of 0.5 (left) on the output of the binary CP-even (mixing angle 0°) vs maximally CP-mixed (45°) classifier for the $t\bar{t}H$ signal, and ranked permutation importance of its input features (right).



Variable	Separation
output score	1.099%
$\Delta\eta_{tB}$	0.658%
$\lvert\cos\theta^*\rvert$	0.625%
$p_{T,H}$	0.513%
b_4	0.465%
b_2	0.175%
η_H	0.117%
$\Delta\phi_{tB}$	0.102%

Figure E.6.: Normalised number of events plotted against the output score of the CP-even (mixing angle 0°) vs maximally CP-mixed (45°) classifier (left) and ranked separation power of the CP-sensitive observables listed in Table 7.1 and the output score (right) for the $t\bar{t}H$ signal.

Acknowledgements

I would like to express my sincerest gratitude to all people without whose guidance and support this thesis would not have been possible. First and foremost, I want to thank my supervisor, Prof. Arnulf Quadt, for bringing his years of expertise and knowledge to this topic and providing me with an opportunity to contribute to analysis on the large scale of the ATLAS Collaboration, and all people without whose contributions the operation of this collaboration would not be possible. I also thank Prof. Jörn Große-Knetter for his help as the second referee.

I extend my sincere thanks to my advisor, PD. Dr. Karsten Köneke, for his continuous support and guidance that ensured the smooth process and successful completion of this thesis. I also want to thank Dr. Steffen Korn for many helpful insights and for always having an open ear in case of any problems when getting to know the complex analysis framework. Your help was much appreciated. I am also grateful to everyone who took part in proofreading this thesis.

Most importantly, I want to thank my family and friends for their unwavering love and support that carried me through my bad days. To my mom: I would not be where I am today without you.

Erklärung zur Nutzung von ChatGPT und vergleichbaren Werkzeugen im Rahmen von Prüfungen

In der hier vorliegenden Arbeit habe ich ChatGPT oder eine andere KI wie folgt genutzt:

- ☐ gar nicht
- ☐ bei der Ideenfindung
- ☐ bei der Erstellung der Gliederung
- ☐ zum Erstellen einzelner Passagen, insgesamt im Umfang von ______________% am gesamten Text
- ☐ zur Entwicklung von Software-Quelltexten
- ☒ zur Optimierung oder Umstrukturierung von Software-Quelltexten
- ☐ zum Korrekturlesen oder Optimieren
- ☒ Weiteres, nämlich: Spellchecking (without generative AI)

Ich versichere, alle Nutzungen vollständig angegeben zu haben. Fehlende oder fehlerhafte Angaben werden als Täuschungsversuch gewertet.

L. Köhler

Luise Köhler
5. Juli, 2026

Erklärung nach §13(9) der Prüfungsordnung für den Bachelor-Studiengang Physik und den Master-Studiengang Physik an der Universität Göttingen:

Hiermit erkläre ich, dass ich diese Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe und alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen wurden, als solche kenntlich gemacht habe.

Darüberhinaus erkläre ich, dass diese Abschlussarbeit nicht, auch nicht auszugsweise, im Rahmen einer nichtbestandenen Prüfung an dieser oder einer anderen Hochschule eingereicht wurde.

Göttingen, den July 5, 2026

(Luise Köhler)