Study on observables carrying spin information in the $t\bar{t}H \rightarrow b\bar{b}$ process with the CMS detector

Master Thesis

von

Victor Serrano Herreros

vorgelegt der

Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen

im März 2025

erstellt im

I. Physikalischen Institut B

bei

Prüfer: Prof. Dr. Lutz Feld
Zweitprüfer: Prof. Dr. Johannes Erdmann

Contents

1 Introduction **1**

2 The Standard Model **2**
2.1 The top quark . . . 4
2.2 The Higgs boson . . . 5
2.2.1 Higgs production in association with a top quark pair . . . 6

3 The Large Hadron Collider **8**
3.1 Compact Muon Solenoid . . . 9
3.1.1 Coordinate system . . . 11

4 Spin and angular variables of $t\bar{t}H$ system **13**
4.1 Top quark pair dynamics . . . 14
4.1.1 Spin and polarization variables . . . 15
4.2 Boson decay variable . . . 17
4.3 $t\bar{t}+X$ system . . . 17

5 Simulated Samples **19**
5.1 Monte-Carlo simulation . . . 19
5.2 Particle Flow . . . 20
5.3 Object reconstruction . . . 21
5.3.1 Electrons . . . 21
5.3.2 Muons . . . 22
5.3.3 Jets . . . 22
5.3.4 Missing transverse momentum . . . 23

6 Event selection **24**
6.1 Polarization and spin observables in dileptonic events . . . 25
6.2 Quantifying the statistical separation power of observables . . . 26
6.3 Distributions of observables for the $t\bar{t}H$, $t\bar{t}Z$ and $t\bar{t}+b\bar{b}$ processes . . . 27

7 Event reconstruction **35**
7.1 Jet assignment and neutrino regression with the SPANet model . . . 35
7.1.1 Network training . . . 37
7.1.2 Jet assignment and neutrino regression performance . . . 39
7.1.3 Particle reconstruction efficiency . . . 41
7.2 Distributions of observables after particle reconstruction with SPANet . . . 44

8 Results and discussion **47**

References **49**

A Events used in SPANet training **56**

B Data paths of simulated samples **56**

C Separation power for the $t\bar{t}Z$ process **57**

D Anti-neutrino regression with SPANet **58**

E Two-dimensional plots for (anti-)neutrino regression **59**

F Top anti-quark reconstruction with SPANet **60**

G Observable reconstruction **61**
G.1 Additional reconstructed observables with SPANet 61
G.2 Ideal reconstruction of observables 64

Abstract

The associated production of a top-antitop quark pair with a Higgs boson ($t\bar{t}H$) provides a direct measurement of the top Yukawa coupling, which is crucial for precision tests of the Standard Model (SM). However, distinguishing the $t\bar{t}H(\rightarrow b\bar{b})$ process from the dominant $t\bar{t}+b\bar{b}$ background remains challenging due to their kinematic similarities. This work explores the novel use of spin and angular observables of the $t\bar{t}H$ system as discriminators between signal and background. A novel machine learning technique based on symmetry-preserving attention networks is employed on parton-jet assignment and neutrino regression with special focus on the signal process, enabling the reconstruction of the intermediate particles and, consequently, the relevant observables. This is the first time the above architecture is applied to the dileptonic channel of the signal process. Results are presented on the statistical separation power of these observables and their potential to contribute to the measurement of $t\bar{t}H$ during LHC Run 3.

1 Introduction

The study of top quark properties and their interactions with the Higgs boson provides a unique window into the Standard Model (SM) and potential physics beyond it [1, 2]. The associated production of a top-antitop quark pair with a Higgs boson ($t\bar{t}H$) is particularly significant, as it directly probes the Yukawa coupling of the top quark to the Higgs field. This interaction is of special interest due to the top quark's exceptional mass and highlights its unique role among elementary particles [3].

This thesis investigates the $t\bar{t}H(\rightarrow b\bar{b})$ process in the context of proton-proton collisions at the Large Hadron Collider (LHC) with the top quarks decaying leptonically. The decay channel where the Higgs boson produces a $b\bar{b}$ pair is chosen due to its relatively high branching ratio and its relevance in testing the SM predictions for heavy flavor physics [4]. However, the analysis of this process faces significant challenges, primarily due to the overwhelming associated top pair QCD background, particularly from $t\bar{t} + b\bar{b}$ production. The kinematic similarity between the signal and background makes their separation non-trivial, necessitating advanced methodologies to enhance the discrimination.

Previous studies have largely focused on improving signal extraction through multivariate analyses [5–7], but the potential of spin-based observables remains unexplored in this specific context. A promising approach explored in this work involves studying said observables carrying spin information of the $t\bar{t}$ system. The top quark, due to its short lifetime, decays before hadronizing, preserving its spin information in the angular distributions of its decay products. This characteristic allows the extraction of spin correlation and polarization observables, which can be used.

The core objective of this thesis is to exploit these spin-sensitive observables, in addition with some angular observables as well, that could be used as powerful discriminators between the signal and background processes. The analysis strategy involves defining a fiducial phase space through rigorous event selection criteria to enhance the identification of $t\bar{t}H$ events while suppressing backgrounds. Spin correlation and polarization observables are calculated at the particle level and adapted to include additional variables sensitive to the Higgs in the $t\bar{t}H$ process. These are reconstructed from objects such as leptons, jets and missing transverse energy using SPANet, a state-of-the-art machine learning framework that addresses key challenges in jet-parton assignment and neutrino reconstruction. Finally, the separation power of the observables is quantified, and the implications for the $t\bar{t}H$ measurement are discussed, providing valuable insights for studies of the Higgs sector.

This work represents a novel exploration of spin observables in the $t\bar{t}H(\rightarrow b\bar{b})$ channel, integrating theoretical insights with advanced computational tools to address the challenging signal extraction. The findings aim to contribute to the broader effort of precision testing the SM, in concrete on the measurement of $t\bar{t}H$ during Run 3.

2 The Standard Model

The *Standard Model* is one of the most remarkable achievements in modern physics, providing a unified framework to describe three of the four known fundamental forces and the particles that mediate them. It was not until the late 20^{th} century that the SM was finally established, when all the necessary concepts were developed and combined into a single theoretical framework; such as the gauge theory formalism of Yang and Mills [8], the discovery of asymptotic freedom by Gross, Wilczek, and Politzer [9, 10], the renormalization of gauge theories by ?t Hooft and Veltman [11], and the systematic classification of elementary particles into quarks, leptons, and bosons, among otherss. All these, alongside the unification of quantum mechanics and relativity into Quantum Field Theory, laid the foundation for the realization that matter at the smallest scales behaves very differently from what was expected in classical physics.
The earliest steps towards the SM were taken in the 1920s and 1930s when quantum mechanics was developed to explain the behaviour of particles at the atomic and subatomic levels. In 1928, Paul Dirac introduced the Dirac equation, which combined quantum mechanics with special relativity and predicted the existence of the positron -the first example of antimatter. As quantum theory advanced, it became clear that particles could be treated as excitations of fields that permeate space; and allowed to describe interactions between particles through the exchange of the "force-carrying" particles, giving birht to Quantum Field Theory.

One of the first successes in this new understanding was Quantum Electrodynamics (QED), which described the electromagnetic force. Developed by Richard Feynman, Julian Schwinger, and Sin-Itiro Tomonaga (among others) in the 1940s, QED provided a highly accurate theory of how electrons and photons interact. This was the starting building block towards the Standard Model. The strong nuclear force and weak nuclear force were much more challenging to understand. The strong force binds protons and neutrons together in the atomic nucleus, while the weak force is responsible for processes like radioactive decay. In the 1960s, Murray Gell-Mann and George Zweig proposed the existence of quarks that combine to form protons, neutrons and other hadrons, through strong force mediation. This was explained within the framework of QCD, which describes the interactions between quarks via the exchange of gluons; QCD also explains why quarks are never observed in isolation but are always bound together in hadrons, a phenomenon known as confinement. Around the same time, Sheldon Glashow, Abdus Salam, and Steven Weinberg worked on unifying the electromagnetic and weak force into what is known as the electroweak theory (EW).

By the early 1970s, the last pieces of the theory were put together. The EW theory and QCD were combined into a single theoretical framework under the group $SU(3)_C \times SU(2)_L \times U(1)_Y$, that describes the interactions of all known elementary particles: quarks, leptons, and bosons. Nevertheless, a crucial component remained absent, namely the explanation of mass. In gauge theories, the local symmetries prevents the presence of a mass term in the equations. Indeed, gauge vector bosons such as the W and Z —mediators of the weak force– are substantially massive [12, 13]. A solution to this problem came in the 1960s with a theoretical proposal put forward by Peter Higgs, François Englert, Robert Brout, and others [14, 15], introducing the concept of the Higgs field. The idea under this mechanism, also known as *Brout-Englert-Higgs (BEH) mechanism*, is to maintain the symmetries but rather to change the vacuum state.

The BEH mechanism relies on the concept of *spontaneous symmetry breaking (SSB)* to solve the issue of mass generation while preserving the gauge invariance of the theory [16, 17]. SSB occurs when the vacuum state of the system does not respect the symmetry of the equations. In the Standard Model, the Higgs field is a scalar field with a potential of the form

$$V(\phi) = -\mu^2|\phi|^2 + \lambda|\phi|^4, \tag{1}$$

where $\mu^2 > 0$ and $\lambda > 0$. This potential exhibits a "mexican hat" shape, leading to a non-zero vacuum expectation value (VEV) for the field, $v = \sqrt{\mu^2/\lambda}$. This VEV breaks the $SU(2)_L \times U(1)_Y$ symmetry of the electroweak theory down to the $U(1)_{EM}$ symmetry of electromagnetism.

The SSB mechanism is responsible of giving the particles mass in different ways. In the case of fermions, their masses arise through Yukawa interactions, which couple the fermion fields to the Higgs field. These interactions take the form $y_f \phi \bar{\psi}_L \psi_R$, where y_f is the Yukawa coupling constant. Because of SSB, the Higgs field acquires its VEV, which is responsible of generating mass terms of the form $m_f = \frac{y_f v}{\sqrt{2}}$. The specific value of y_f determines the mass of each fermion, explaining the observed mass hierarchy in the particle spectrum [16].

For gauge bosons, the W and Z bosons acquire mass by "absorbing" the degrees of freedom associated with the would-be Goldstone bosons resulting from SSB. The Higgs field, initially described by four degrees of freedom, leaves one as a physical scalar particle (the Higgs boson) while the remaining three provide the longitudinal components of the $W^{\pm}$ and Z bosons. The masses of these gauge bosons are given by $M_W = \frac{gv}{2}$ and $M_Z = \frac{M_W}{\cos\theta_W}$, where g is the gauge coupling constant and θ_W is the Weinberg angle. The photon, associated with the $U(1)_{EM}$ symmetry, remains massless, consistent with experimental observations of the electromagnetic force.

Over time, experimental evidence gradually confirmed the predictions of the Standard Model and played a fundamental role in measuring the free parameters of the model. The discovery of the W and Z bosons at CERN in 1983 [12, 13] was a major triumph for electroweak theory. Meanwhile, experiments confirmed the existence of all six quarks [19–22]: up, down, strange, charm, bottom and top, as well as the various types of leptons [23–25] (such as muons, taus and neutrinos). In 2012 the final puzzle piece fell into place, when the discovery of the Higgs boson was announced [26], validating the existence of the Higgs field and confirming the BEH mechanism as the process through which the particles acquire mass.

While the Standard Model is incredibly successful, it is not a complete theory, yet many unanswered question remain to be unravelled. It does not incorporate gravity, which is described by Einstein's General Theory of Relativity, nor does it explain dark matter, dark energy, the neutrino mass origin or the observed imbalance between matter and antimatter in the universe. Nevertheless, many active work trying to extend the SM has been done in the last decades, in order to develop a more comprehensive framework that addresses all these unresolved questions.

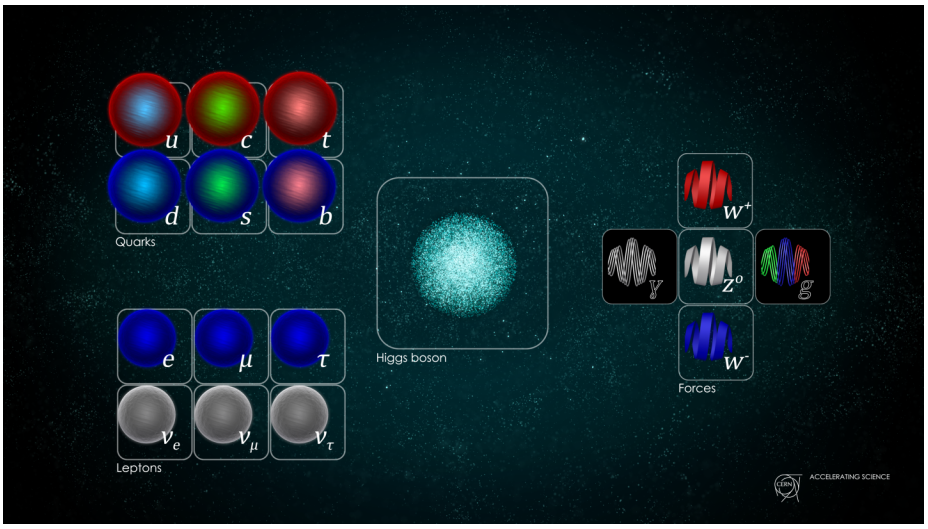

Figure 1: Illustration of the particles of the Standard Model. On the left are placed the fermions: the six quarks label with u, d, c, s, t and b, with the up-type quarks in red shell and the down-type in a blue shell, the three quark generations are illustrated with the blue, green and red inner colours. Also the six leptons in the bottom left, are drawn in blue for the charged one while the neutrinos are in grey. The scalar Higgs boson is placed in the center of the image. Gauge boson are on the right, with the weak force mediators in tree colours (in accordance with their charge), and the photon and the gluon with a dark background (massless particles). The gluons are also illustrated as containing colour charge. From [18]

2.1 The top quark

In the SM, the top quark holds a special place due to its unique properties, especially its large mass. The left-handed top quark is part of a weak-isospin doublet of the $SU(2)_L$ group -which describes weak interactions-, along with the bottom quark; it has an electric charge of $Q = +\frac{2}{3}$ and an isospin component $T_3 = +\frac{1}{2}$ [2]. Conversely, the right-handed top quark is a singlet state. The mass is a parameter from the theory that depends on the Yukawa coupling and the vacuum expectation value. For the top, such given mass is $m_t \approx \frac{\nu}{\sqrt{2}}$, with $\nu = 246\,\text{GeV}$ the VEV of the Higgs field; and therefore, a Yukawa coupling to the Higgs of the order of unity, reflecting explicitly the strong relation between the top quark and the Higgs boson.

What makes the top remarkable with respect to other quarks is that it is the only quark that is heavier than the W boson. As a result, it can decay weakly into a on-shell W boson and a bottom quark. Its decay width is amply dominated by this channel, $t \to Wb$, with $|V_{tb}| >> |V_{td}|, |V_{ts}|$. [1] With a short lifetime of about $\sim 5 \times 10^{-25}$ seconds, it is expected to decay before strong interactions have time to form bound states, behaving almost as a free particle during its brief existence. [2]

[1] V_{ij} refers to the elements of the Cabibbo-Kobayashi-Maskawa matrix [16], which gives the strength of the i-flavour quark decaying to a j-flavour quark via weak interaction. These flavour-changing currents are exclusively between up-type and down-type quarks.

[2] Some studies [27] propose that toponium ($t\bar{t}$-quarkonium-states) can be form with a $t\bar{t}$ binding time closer to the top lifetime, resulting in a peak in e^-e^+ scattering at the $t\bar{t}$ production threshold.

All these features make the top quark an excellent candidate to probe the theory. The large mass and Yukawa coupling to Higgs, give important loop contribution to precise measurements, such as the W boson mass which is correlated with the top and Higgs mass values and make substantial predictions to the vacuum stability. Additionally, it can be used to test the SM at the EW symmetry breaking scale and beyond. Precise measurements of its properties —mass, couplings, production cross sections, decay branching ratios, etc.) [2]—, can be used to search for deviations that may hint at new physics, and improve constraints on theoretical parameters. Although achieving precise measurements requires a high production rate of top quarks, hadron colliders such as the LHC, with their high energy and luminosity, enable a large production of top quark pairs necessary for these studies.

2.2 The Higgs boson

The Higgs boson, is a scalar particle with spin 0 and CP-even symmetry [1]. Its mass is a free parameter of the Standard Model (SM) and is given by $m_H \approx \sqrt{2\lambda}\nu$, where λ represents the self-coupling parameter in the SM scalar potential (Eq. 1). In 2012, the discovery of a particle consistent with the Higgs boson was announced, with a measured mass of $m_H \approx 125$ GeV [26].

The strength of the interactions between the Higgs boson and other particles is governed by Yukawa couplings, which are proportional to the mass of each particle. For gauge bosons, this relationship is quadratic in their mass, with a coupling of $g_{HVV} = \frac{2m_V^2}{\nu}$, while for fermions, the coupling is linearly proportional to the mass, given by $g_{Hf\bar{f}} = \frac{m_f}{\nu}$. Consequently, particles like the W, Z, and the top quark exhibit stronger couplings to the Higgs field compared to lighter quarks and leptons.

The Higgs boson is primarily produced in high-energy proton collisions at the Tevatron and LHC colliders, with the main production mechanisms[3] being gluon-gluon fusion (ggF), vector-boson fusion (VBF), production with an associated weak gauge boson (VH), production with a top-quark pair (ttH), and single-top associated production (tH). The center-of-mass energy $\sqrt{s}$ of the collision influences the cross sections for each production channel[1], with ggF being the dominant mechanism at the Tevatron and LHC due to the high density of gluons in protons at these energies.

Given that the branching ratios of the Higgs boson, i.e. the probabilities of decay into specific particles, depends on the mass of the Higgs itself (approximately 125 GeV), its decay channels can be predicted for this value. Ranking from the highest to the lowest as follows[4]: $H \to b\bar{b}$ and $H \to WW^*$, followed by $H \to gg$, $H \to \tau^+\tau^-$, $H \to c\bar{c}$, and $H \to ZZ^*$[3]. Decays with smaller branching ratios, yet accessible in experiments, include $H \to \gamma\gamma$, $H \to \gamma Z$, and $H \to \mu^+\mu^-$. In experimental analyses, a combination of production and decay channels is measured, providing critical insights into the Higgs boson couplings with SM heavy particles. This information is contained in the *signal strength* (μ), which quantifies the agreement between

[3]The "*" accounts for off-shell particles

observed rates and SM predictions, thereby serving as a benchmark for potential deviations from the SM.

2.2.1 Higgs production in association with a top quark pair

One of the most important processes for probing the interaction between the Higgs boson and the top quark is the associated production of the Higgs boson with a top-antitop quark pair. This process, $pp \rightarrow t\bar{t}H$, provides a direct probe of the top-Higgs Yukawa coupling. This coupling is of great interest because the top quark, being the heaviest particle in the Standard Model, has the strongest interaction with the Higgs field, and its large mass suggests a special role in electroweak symmetry breaking and in the dynamics of the Higgs mechanism. Additionally, the cross section has been extensively studied, with next-to-leading order (NLO) QCD and NLO electroweak corrections contributing to a more precise measurement of the top-Higgs interaction.

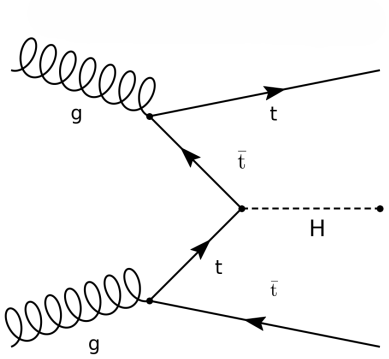

Figure 2: Feynman diagram at tree-level of the associated production of a Higgs with a top quark pair, $gg \rightarrow t\bar{t}H$.

The decay $t\bar{t}H(\rightarrow b\bar{b})$ is particularly significant, as it allows a direct measurement of the $t\bar{t}H$ production cross section. The inclusive NNLO in QCD cross section[28] is $\sigma_{t\bar{t}H} = 0.5070^{+0.9\%}_{3.0\%}$ [pb] at $\sqrt{s} = 13$ TeV, and a branching fraction of the Higgs decaying to bottom quark pair of approximately $\mathcal{BR} = 58\%$ [29]. These measurements are challenging due to the large QCD background, particularly the $t\bar{t}+b\bar{b}$ process, which closely matches the signal due to its similar final state configuration. The background is especially problematic because it is enhanced by higher-order QCD effects[3].

To mitigate these background contributions, advanced techniques are employed in several studies[5, 30], including the use of multivariate analysis (MVA) and machine learning algorithms. These kind of methods enhance the signal-to-background discrimination by identifying subtle differences in kinematic distributions between the signal and background processes. Another approach to improve the signal extraction involves studying additional Higgs decay channels[6], such as multilepton final states from $H \rightarrow WW^*, ZZ^*$, or $\tau^+\tau^-$, which have lower branching ratios but are less affected by the $t\bar{t}b\bar{b}$ background. Combining these channels with

$t\bar{t}H(\rightarrow b\bar{b})$ in a global fit improves the precision of the $t\bar{t}H$ coupling measurement and reduces the systematic uncertainties.

Lastly, the $t\bar{t}H(\rightarrow b\bar{b})$ process is not only crucial for measuring the top-Higgs Yukawa coupling, but also plays a key role in understanding the top quark singular role in the Standard Model. Deviations in the measured $t\bar{t}H$ cross section or the $H \rightarrow b\bar{b}$ branching ratio could indicate physics beyond, such as modifications to the Higgs sector or new interactions with the top quark.

3 The Large Hadron Collider

The Large Hadron Collider, located at CERN in Geneva, is the largest and most advanced particle accelerator in the world [31]. It lies 100 meters underground in a 27-kilometer circular tunnel, where protons and heavy ions are accelerated to nearly the speed of light [32]. The accelerator uses 1232 superconducting dipole magnets to steer particle beams along the circular trajectory and 392 quadrupole magnets to focus the beams tightly to a nominal trajectory [33]. These magnets are cooled to $-271.3°$C using liquid helium, achieving a temperature of 1.9 Kelvin, which allows for superconducting operation with zero electrical resistance [34]. This enables the generation of magnetic fields up to 8.33 T, essential for steering particles at energies nearly 7 TeV per beam [31, 33].

The LHC's design incorporates a hierarchy of pre-accelerators, which increase particle energies before injection into the main ring. These include the linear accelerator (LINAC2 for protons and LINAC3 for heavy ions), the Proton Synchrotron Booster (PSB), the Proton Synchrotron (PS), and the Super Proton Synchrotron (SPS), see Figure 3. Each stage is vital for optimizing beam quality and energy [35].

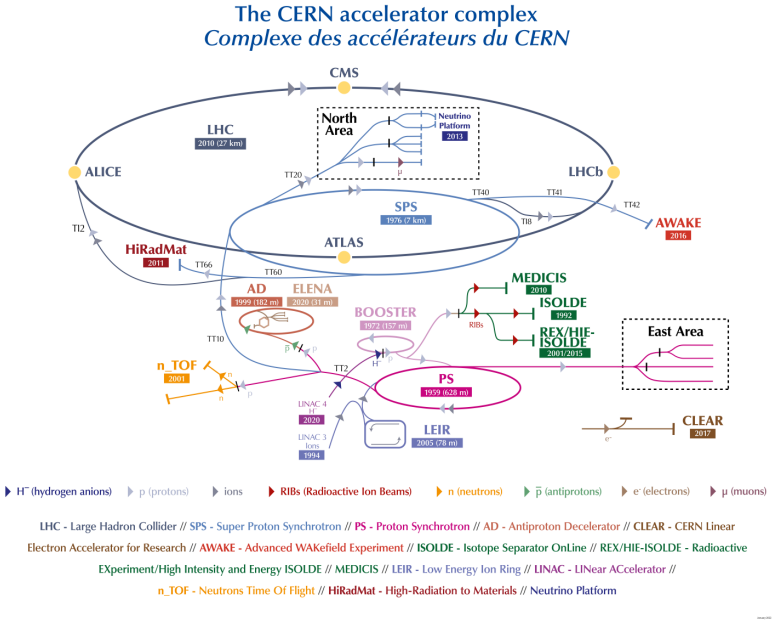

Figure 3: Schematic representation of the LHC (blue ring) with detectors indicated by yellow dots. Pre-acceleration stages include the SPS and PS rings. Image source: CERN.

The LHC operates with particle bunches containing approximately 1.15×10^{11} protons, spaced by 25 ns, resulting in a bunch crossing rate of 40 MHz at the interaction points [32]. The beams collide at four main detectors –ATLAS, CMS, LHCb, and ALICE–designed to investigate diverse phenomena ranging from Standard Model physics to potential discoveries beyond it, such as supersymmetry, dark matter candidates, and extra dimensions [31, 36].

To reach such high energies, the beam pipes are maintained at an ultrahigh vacuum of 10^{-13} atm, to prevent interactions between particles and residual gas molecules [37]. Managing the large data from the collisions, where up to 10^9 interactions per second can occur, requires advanced triggering systems and a specific designed computing infrastructure, to analyze petabytes of data annually [38]. This system is essential for handling the enormous data flow, allowing to focus on the most relevant events for further analysis.

Planned upgrades, including the High-Luminosity LHC (HL-LHC), aim to increase its luminosity by a factor of 10, enabling more precise measurements. For example, it is expected to produce 15 million of Higgs per year, compared to the three million from 2017 [39].

3.1 Compact Muon Solenoid

The CMS detector is built around a large superconducting solenoid magnet, which provides a magnetic field of $3.8\,\mathrm{T}$, enabling the precise measurement of the momentum of charged particles. The solenoid is covered by a steel flux-return yoke, which contributes to the total weight of $14,000\,\mathrm{t}$. The detector is compact, compared to others such as ATLAS [40], measuring $21\,\mathrm{m}$ in length and $15\,\mathrm{m}$ in height and width [41]. It has a modular design including and arrangement of specialized subsystems, each of one designed for the detection of different particle types and kinematic properties. A radial view of the detector layout is shown in Figure 4.

The CMS detector is composed of several key components, each designed to measure different types of particles and properties [41]. [These include the *tracker*, which precisely records the trajectories of charged particles; the *electromagnetic calorimeter* (ECAL), which measures the energy of photons and electrons; the *hadronic calorimeter* (HCAL), for detecting hadrons like protons and neutrons; and the *muon system*, which surrounds the other layers and identifies muons as they penetrate through the entire detector.] Together, these components are arranged around the solenoid magnet, enabling CMS to reconstruct complex collision events with high precision and cover a wide range of interactions. In Figure 4 a radial section of the CMS detector is shown with all the corresponding modules.

The CMS detector consists of a series of specialized subsystems, each optimized to detect specific particle types and measure their properties [41]. These include (from the innermost to the outermost layer in the detector) the *tracker*, the *electromagnetic calorimeter (ECAL)*, the *hadronic calorimeter (HCAL)*, and the *muon system*. Together, these subsystems are arranged around the solenoid magnet –except the muon system which lies outside–, in a structure by layers that enables the CMS detector to reconstruct the complex events produced in the collisions. This modular design allows a coverage of a wide range of particle interactions.

Inner tracker

The inner tracker of the CMS detector, located closest to the beamline, is a system designed to track the paths of charged particles originated from the collision points. It consists of two main sub-detectors: the silicon pixel detector (SPD) and the silicon strip tracker (SST).

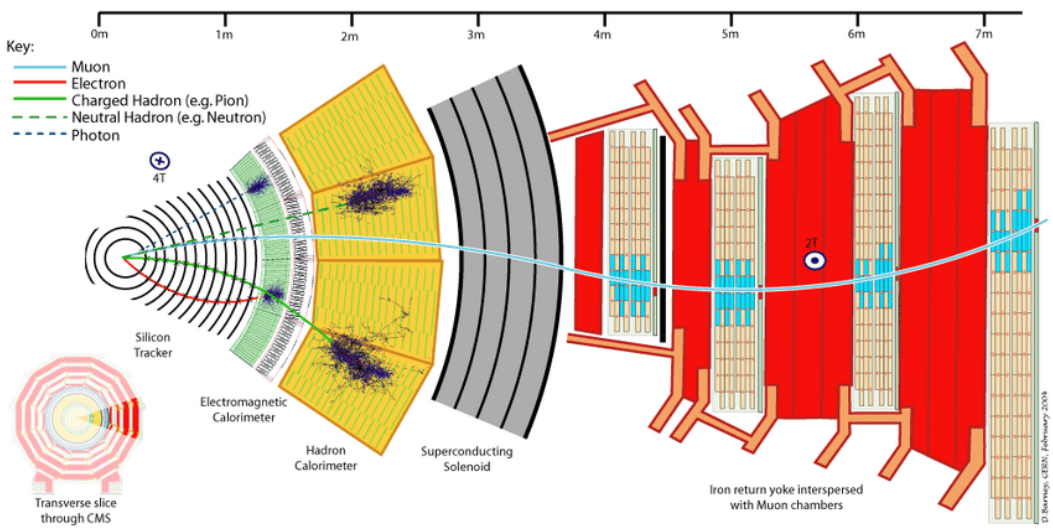

Figure 4: Radial section of the CMS detector, with the main subsystems and their arrangements. From *CERN*, for the bene?t of the CMS Collaboration.

- The *silicon pixel detector*, the innermost layer of the tracker, standouts in vertex reconstruction, providing precise spatial measurements essential for differentiating collision vertices and identifying secondary vertices from decays such as b-quarks. It is composed of modular silicon sensor arrays, with each module containing 160×416 pixels. The individual pixel size is $100 \times 150\,\mu\text{m}^2$, which allows for an exceptional spatial resolution of approximately $10\,\mu$m [41]. The SPD architecture is divided into a barrel section, which consist on four concentric cylindrical layers, and forward disks, three on each side, with a pseudorapidity of $|\eta| < 3.0$. The barrel's innermost layer is situated at $29\,$mm from the beamline, making it highly sensitive to particles produced immediately after a collision. The pixel detector provides precise information on the transverse impact parameter, which plays a big role for tagging long-lived particles such as B-mesons [42].

- Covering the pixel detector lies the *silicon strip tracker*, responsible for measuring the momentum and paths of charged particles. It covers an acceptance range of $|\eta| < 2.5$ and has a total active area of $198\,\text{m}^2$ [41]. The SST is designed with a barrel-endcap geometry, comprising 10 barrel layers in total, divided into two groups: the inner barrel with four layers and the outer barrel with six. In addition, the endcap regions has three inner disks and two outer disks on each side of the barrel. Each silicon strip module includes series of micro-strips oriented at precise angles for an optimal spatial resolution in both the radial and longitudinal directions, which is critical for reconstructing high-momentum tracks in complex events [41]. The SST is crucial for maintaining tracking efficiency in regions of the detector where the pixel resolution diminishes [42].

The combination of the pixel and strip detectors enables the inner tracker to perform in a high-luminosity environment, ensuring efficient track reconstruction in events with high particle multiplicities and significant pile-up [42]. The entire tracker is housed within a hermetic structure that minimizes material interactions, reducing effects such as bremsstrahlung radiation and secondary scattering [41].

Electromagnetic calorimeter

The ECAL is the next module outwards from the tracker. It is constructed from lead tungstate ($PbWO_4$) scintillating crystals, chosen for their high density and excellent optical properties [41]. The ECAL is divided into a barrel section and two endcaps, collectively covering a pseudorapidity range of $|\eta| < 3.0$. Photons and electrons deposit their energy within these crystals, producing scintillation light that is converted into an electrical signal by silicon photodiodes (in the barrel) and vacuum phototriodes (in the endcaps). The ECAL has an excellent energy resolution, which is a key feature for identifying and reconstructing photons and electrons with high precision [42].

Hadronic calorimeter

Surrounding the ECAL is placed the HCAL, responsible of measuring the energy of hadrons such as protons, neutrons, and pions. The HCAL is designed with alternating layers of brass absorber and plastic scintillator tiles. The light produced in the scintillators is guided through wavelength-shifting fibers to hybrid photodetectors, enabling energy measurements with good resolution [42]. The HCAL consists of a barrel section, endcaps, and additional forward calorimeters that extend the coverage to $3.0 < |\eta| < 5.2$. These forward calorimeters play an important role in the precision measurement of the transverse energy, particularly in high-energy jet events. Additionally, an outer hadron calorimeter, located outside the solenoid, helps capture the shower tails of energetic jets and helps improving the rejection of certain (hadronic) backgrounds in events involving muons [41].

Muon system

The muon system consist on the outermost layer of the CMS detector, designed to identify and measure muons, which penetrate the calorimeters and the solenoid. This subsystem is embedded within the steel flux-return yoke, which also confines the magnetic field of the solenoid. The muon detection system is built of three types of detectors: drift tubes (DTs) in the barrel region, cathode strip chambers (CSCs) in the endcaps, and resistive plate chambers (RPCs) which can be found both regions, the barrel and the endcap. The DTs and CSCs provide spatial resolution, which enables to reconstruct the muon trajectories, while the RPCs provides fast timing information vital for triggering. Together, all these systems ensure high efficiency and excellent momentum resolution for muons across the detector's acceptance range of $|\eta| < 2.4$ [42]. Recent upgrades, including the addition of gas electron multiplier (GEM) chambers, have further improved performance in the forward regions, where higher particle fluxes implies additional challenges [41].

3.1.1 Coordinate system

The CMS coordinate system is chosen taking the origin as the nominal collision point inside the CMS detector. The x-axis is pointing to the center of the LHC circumference, the y-axis pointing upwards, and the z-axis in the direction of the beam according to $\hat{z} = \hat{x} \times \hat{y}$. The spherical coordinates are standard: r is the radial coordinate, the azimuthal angle, ϕ, is measured from the x-axis and confined in the x-y plane, and the polar angle θ starts from the z-axis

in the z-y plane. A more convenient coordinate, instead of the polar angle, is defined as the pseudorapidity, $\eta \equiv \ln\tan(\theta/2)$. Transverse quantities to the beam direction as the energy or momentum are denoted as E_T and p_T, respectively; inequality in the measured energy in such transverse plane is named E_T^{miss} or MET (missing transverse energy). A common variable used to defined angular distances between two objects is $\Delta R \equiv \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$.

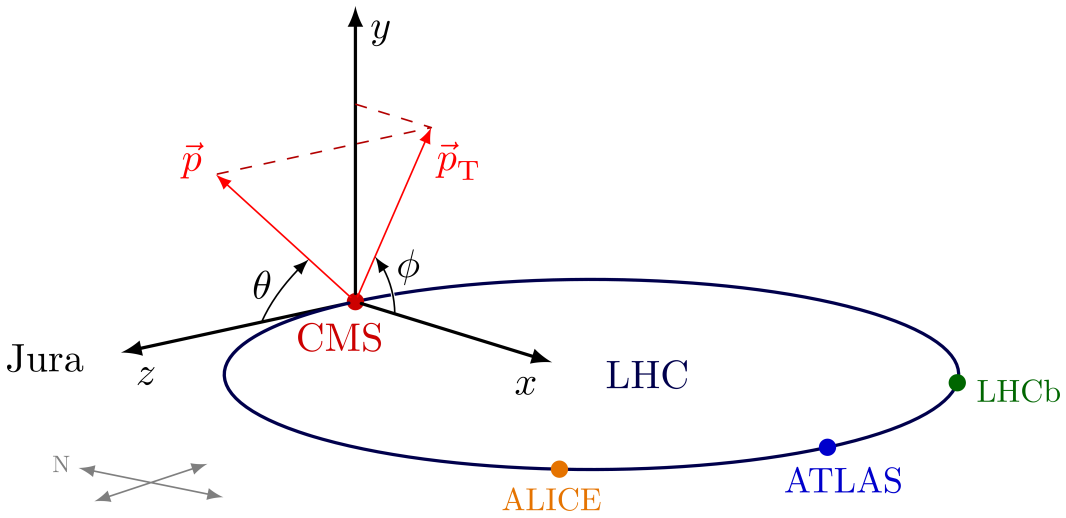

Figure 5: CMS coordinates system taking as origin the nominal collision point at CMS detector. x-axis pointing to the center of the LHC circumference, y-axis pointing upwards, and the z-axis in the direction of the beam given $\hat{z} = \hat{x} \times \hat{y}$ components. Transverse momentum ($\vec{p}_T$) is confined to the x-y plane. From [43]

4 Spin and angular variables of $t\bar{t}H$ system

Spin, as the intrinsic quantum angular momentum of particles, is not only a fundamental characteristic of matter but also a direct manifestation of the underlying symmetries of the quantum field theories describing the SM [44]. Its study provides insights into the structure of interactions, linking particle dynamics to conserved quantities such as total angular momentum and parity. Furthermore, spin observables are highly sensitive to the helicity structure of the interactions, making them a powerful probe for testing the chiral nature of electroweak couplings [45].

In particular, spin correlations between particles are sensitive to the production and decay of such. They encode valuable information about the mechanisms acting during these processes, which may provide singular footprints characteristic of each [46]. Moreover, the spin has a major role when transferring information to decay products, especially for unstable particles such as top quarks. Due to the short lifetime of the top, its spin state is preserved until its decay, transferring the information to the angular distributions of its decay products. This unique feature allows the reconstruction of the initial spin state from experimental data, enhancing the capability of using the spin to characterize complex processes [47].

The study of spin correlation in systems involving top quark pairs has been deeply researched in the literature [48, 49]. In particular, the associated production of a boson with a top quark and antiquark can be significantly different considering the spin of the boson itself. Thus, studying the spin information of a system could highlight possible differences between processes with similar topologies but involving distinct intermediate particles.

In this work the associated production of a boson with a top quark pair ($t\bar{t}X$ —where X can be any type of boson) is decoupled as two independent systems consisting on the two top quarks ($t\bar{t}$) and the boson itself decaying to two bottom quarks ($X \to b\bar{b}$). Within this framework, relevant variables are defined sensitive to each sub-system, featuring the information contained in each type of particle. A complete study of the spin correlation involving the production and decay of the $gg \to t\bar{t}Z$ system (considered as one unified system) can be found in [??], however the complexity of the analysis is significantly increased.

Concerning the process $t\bar{t}H(\to b\bar{b})$, the large background of $t\bar{t}$ with heavy-flavour jets [50] obscures the signal extraction due to the similar final state, i.e. heavy-flavour jets and leptons. In the work of [5], this is already treated using a boosted decision tree to select the more prominent variables with highest background reduction. In this work, a new set of variables is proposed focusing on spin-related and angular observables, where differences between the production mechanisms of the resonances ($t\bar{t}H$ and $t\bar{t}Z$) and the former background could manifest through the reconstructed final state particles.

4.1 Top quark pair dynamics

Top quark properties have been intensively studied at the LHC collider, which produces a large number of these particles [22]. The top quark-antiquark pair spin correlation provides a unique test for the Standard Model and opportunity to search for potential deviations. This correlation can be described through spin-information variables (see Eq. 5), which can be used to probe the interactions within the $t\bar{t}$ pair. Moreover, recent studies in ATLAS [51] and CMS [52] have reported the observation of quantum entanglement in top quark-antiquark pair production, using observables derived from the spin correlation matrix, thus highlighting the importance of spin-based measurements in understanding quantum properties and testing fundamental theories of particle interactions.

The brief lifetime of the top quark ($\sim 10^{-25}$ s), which is shorter than both the time-scale for QCD hadronization, $^1/_{\Lambda_{QCD}} \sim 10^{-24}$ s, and the spin decorrelation time, $^{m_t}/_{\Lambda^2_{QCD}} \sim 10^{-21}$ s, allows for the direct measurement of the top quark's spin properties from its decay products [47]. The top quark predominantly decays via electroweak interactions into a W^+ boson and a b quark (or a W^- boson and a $\bar{b}$ quark for the top antiquark), with the W^+ (W^-) decaying either hadronically into a $q\bar{q}'$ pair or leptonically into $\ell^+\nu$ ($\ell^-\bar{\nu}$), where ℓ represents electrons, muons, or taus. The channel with both W boson decaying to leptons, *dileptonic*, is of special interest because for the leptons the spin analyzing power [53], which amounts for the top spin information preserved in its decays, is $\kappa_{\ell^+} \approx 1$ (with opposite sigh for the antiparticles), providing the most sensitive channel to top spin effects. Allowing the study of top spin correlation properties through the angular distributions of the leptons.

At the LHC, top quark pairs are mainly produced via gluon-gluon fusion ($gg \to t\bar{t}$), with a smaller contribution from quark-antiquark annihilation ($q\bar{q} \to t\bar{t}$) [2]. In the SM, top quarks are unpolarized at leading order (LO) due to the parity-conserving nature of QCD interactions, which implies no preferred spin direction (longitudinal polarization), and the approximate time invariance of these interactions, which suppresses transverse polarization. Nonetheless, electroweak corrections and absorptive terms at one loop introduce minor contributions to both longitudinal and transverse polarizations, each below 1% [48].

Although the top quarks are initially unpolarized—production via strong interaction—the spins of the $t\bar{t}$ pair are strongly correlated. The nature of these correlations depends on the invariant mass of the pair, $m_{t\bar{t}}$. Near the production threshold ($m_{t\bar{t}} \sim 2m_t$), gluon-gluon fusion predominantly produces $t\bar{t}$ pairs in maximally entangled spin-singlet states (Bell states), characterized by anti-parallel spins along any chosen axis. Conversely, quark-antiquark annihilation at threshold produces separable $t\bar{t}$ pairs with no definitive spin correlation. As a result, the physical state of $t\bar{t}$ pairs at threshold is a mixed state, with the degree of entanglement determined by the relative contributions of the gg and $q\bar{q}$ production channels. While the behavior of $t\bar{t}$ pairs near threshold has been extensively studied and is well understood [51, 52], the broader picture over the full phase space remains more intricate. At low transverse momentum ($p_T < m_t$), $t\bar{t}$ pairs are primarily produced via gg fusion in an s-wave state, resulting in maximally anti-correlated spins along any axis. At high p_T, however, the pairs are produced through chiral mechanisms, with their spins becoming aligned (parallel) along

the direction of the $t\bar{t}$ system. The interference effects of these distinct spin configurations are captured in the off-diagonal elements of the spin-correlation matrix. These features not only make the study of $t\bar{t}$ spin dynamics necessary, but also offer a sensitive testing ground for the associated top pair production with additional particles.

4.1.1 Spin and polarization variables

To quantify the polarization and spin correlation strength of the $t\bar{t}$ system, it is essential to establish a clear understanding of the production dynamics and define an appropriate reference basis for measuring the relevant observables. Given the narrow width of the top quark, the production and decay of the $t\bar{t}$ pair can be effectively treated as independent processes, enabling the factorization of their respective spin density matrices. Under this approximation, the squared matrix element for $t\bar{t}$ production (via gluon-gluon fusion and/or quark-antiquark annihilation) followed by decay into two leptons (after spin and color summation) can then be expressed as [47]:

$$|\mathcal{M}\{gg/q\bar{q} \to t\bar{t} \to (\ell^+\nu b)(\ell^-\bar{\nu}\bar{b})\}|^2 \approx \text{tr}\,[\rho R\bar{\rho}], \tag{2}$$

where ℓ denotes either electrons or muons. Here, R is the spin density matrix describing the $t\bar{t}$ pair production, encapsulating the spin correlations between the top quark and antiquark, whereas, ρ and $\bar{\rho}$ are the spin density matrices associated with the decay of the top quark and antiquark, respectively.

The production spin density matrix, R, can be further decomposed into a basis for the top quark and antiquark spin states. Using the Pauli matrices σ^i as a spin operator basis, this decomposition takes the following form:

$$R \propto \mathcal{A}\mathbb{I}\otimes\mathbb{I} + \sum_{i=1}^{3}\left(\mathcal{B}_i^+\sigma^i\otimes\mathbb{I} + \mathcal{B}_i^-\mathbb{I}\otimes\sigma^i\right) + \sum_{i,j=1}^{3}\mathcal{C}_{ij}\sigma^i\otimes\sigma^j. \tag{3}$$

Here, $\mathbb{I}$ is the 2×2 identity matrix, and σ^i are the Pauli matrices, which form a natural basis for spin states aligned with the polarization axes of the top quark. The first matrix in the tensor product acts on the top spin space, while the second operates in the top antiquark spin space. The coefficient $\mathcal{A}$ is a scalar function related to the spin-independent $t\bar{t}$ production rate, encoding contributions to the total cross section and top kinematics. $\mathcal{B}_i^{\pm}$ represent three-vectors of functions that quantify the degree of polarization of the top quark $(+)$ and antiquark $(-)$ along each axis of the chosen reference frame. Lastly, $\mathcal{C}_{ij}$ is a 3×3 tensor of functions that characterizes the correlation between the $t-\bar{t}$ spins.

The choice of spin basis is not unique, but it is common in the literature [48] to adopt a specific orthonormal basis that simplifies the decomposition of $\mathcal{A}$, $\boldsymbol{\mathcal{B}}^{\pm}$ and $\mathcal{C}$. The basis is defined in the zero-momentum frame (ZMF) of the $t\bar{t}$ pair, where all particles are boosted into this frame. The direction of the incoming parton, $\hat{\boldsymbol{p}}$, and the outgoing top quark in the ZMF, $\hat{\boldsymbol{k}}$, are used

to define the perpendicular direction to the scattering plane, $\hat{\boldsymbol{n}}$. To complete the set, the third element $\hat{\boldsymbol{r}}$ is constructed as the vector orthogonal to both $\hat{\boldsymbol{n}}$ and $\hat{\boldsymbol{k}}$; together, these three vectors form a right-handed orthonormal basis:

$$\{\hat{\boldsymbol{k}}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{r}}\}: \quad \hat{\boldsymbol{n}} = \frac{1}{r}(\hat{\boldsymbol{p}} \times \hat{\boldsymbol{k}}),\ \hat{\boldsymbol{r}} = \frac{1}{r}(\hat{\boldsymbol{p}} - y\hat{\boldsymbol{k}}) \quad \text{with} \quad y = \hat{\boldsymbol{k}} \cdot \hat{\boldsymbol{p}},\ r = \sqrt{1 - c^2} \tag{4}$$

The variables r and y correspond to $r = \sin\Theta$ and $y = \cos\Theta$, where Θ is the scattering angle of the top quark. Figure 6 illustrates this coordinate system, which is employed to define the spin coefficients.

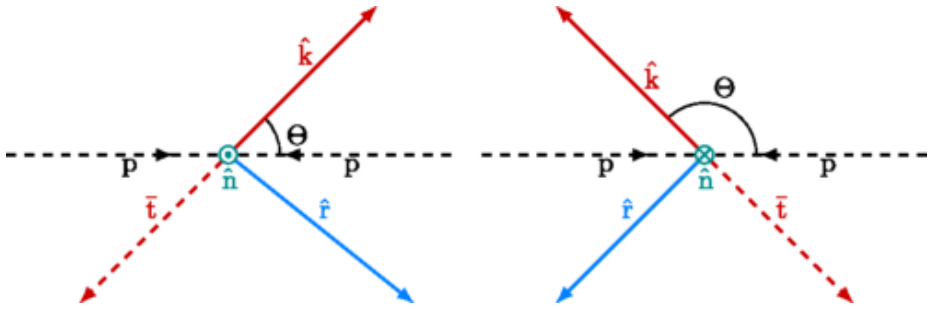

Figure 6: Reference system used to define the polarization and spin-correlation observables, from [47]. The direction of the incoming partons is (dashed black lines) $\boldsymbol{p}$, $\hat{\boldsymbol{k}}$ is chosen as the in the direction outgoing top quark (solid red line). $\hat{\boldsymbol{n}}$ (blue circle with a dot/cross) is build as the orthogonal direction to the scattering plane define by $\boldsymbol{p}$ and $\hat{\boldsymbol{k}}$. The $\hat{\boldsymbol{r}}$ vector (solid blue line) is the cross product of $\hat{\boldsymbol{k}}$ and $\hat{\boldsymbol{n}}$. The set $\hat{\boldsymbol{k}}, \hat{\boldsymbol{n}}, \hat{\boldsymbol{r}}$ form a right-handed orthonormal basis.

In this basis, the functions expand as follows:

$$\begin{aligned}
\mathcal{B}_i^{\pm} &= b_k^{\pm}\, \hat{k}_i + b_r^{\pm}\, \hat{r}_i + b_n^{\pm}\, \hat{n}_i, \\
\mathcal{C}_{ij} &= c_{kk}\, \hat{k}_i \hat{k}_j + c_{rr}\, \hat{r}_i \hat{r}_j + c_{nn}\, \hat{n}_i \hat{n}_j \\
&\quad + c_{rk}(\hat{r}_i \hat{k}_j + \hat{k}_i \hat{r}_j) + c_{nr}(\hat{n}_i \hat{r}_j + \hat{r}_i \hat{n}_j) + c_{kn}(\hat{k}_i \hat{n}_j + \hat{n}_i \hat{k}_j) \\
&\quad + c_n(\hat{r}_i \hat{k}_j - \hat{k}_i \hat{r}_j) + c_k(\hat{n}_i \hat{r}_j - \hat{r}_i \hat{n}_j) + c_r(\hat{k}_i \hat{n}_j - \hat{n}_i \hat{k}_j).
\end{aligned} \tag{5}$$

In total, the expansion yields 15 independent coefficients with symmetric and antisymmetric components in $\mathcal{C}$. These coefficients depend on the center-of-mass energy squared of the partons, $\hat{s}$, and the top quark scattering angle y.

As suggested in [47], the Bose-Einstein symmetry of the gg initial state imposes specific constraints on the definitions of the $\hat{\boldsymbol{r}}$ and $\hat{\boldsymbol{n}}$ axes, as these axes are odd under this symmetry. To account for these constraints and enable measurements of non-vanishing values for the corresponding coefficient functions, a redefinition of the axes is necessary. The modified basis is defined as follows:

$$(\hat{\boldsymbol{k}}, \hat{\boldsymbol{r}}, \hat{\boldsymbol{n}}) \to (\hat{\boldsymbol{k}},\ \mathrm{sign}(\cos\Theta)\hat{\boldsymbol{r}},\ \mathrm{sign}(\cos\Theta)\hat{\boldsymbol{n}}), \tag{6}$$

where the sign of $\cos\Theta$, which is odd under Bose-Einstein symmetry, establishes a consistent "forward" direction for each event. This redefinition ensures that the axes align correctly with the symmetries of the $t\bar{t}$ production process. In Figure 6, this "axis per event" adjustment is illustrated by the flipping of the signs of $\hat{\boldsymbol{r}}$ and $\hat{\boldsymbol{n}}$ at $\Theta = \pi/2$. It is important to account for this feature when constructing the observables for the coefficient functions introduced in Section 6.3.

In addition, inspection of additional variables —which have been shown to also be sensitive to $t\bar{t}$ spin correlations [54, 55]— can help identify significant variations in the distributions of processes involving a top-antitop pair. For instance, $|\Delta\eta_{\ell\bar{\ell}}|$ and $|\Delta\phi_{\ell\bar{\ell}}|$, measure the differences in pseudorapidity and azimuthal angle between the two charged leptons in dileptonic final states. Observations of $t\bar{t}$ spin correlation have been reported [49] using the azimuthal angle difference, $|\Delta\phi_{\ell\bar{\ell}}|$. The presence of a third body (X) can alter the angular distribution of the leptons due to the mass of the associated particle.[4]

Another set of variables concerning the $t\bar{t}$ system are $c_{\mathbf{hel}}$ and $c_{\mathbf{han}}$, although not completely independent of the former spin and polarization coefficients, can highlight differences between states with distinct parity properties when produced with an intermediate particle X. Here, c_{hel} is defined as the cosine of the angle between the two leptons in the $\{\hat{\boldsymbol{k}}, \hat{\boldsymbol{r}}, \hat{\boldsymbol{n}}\}$ basis[5]. Meanwhile, c_{han} is a linear combination of the spin correlation coefficients, designed to enhance sensitivity to parity-violating effects.

4.2 Boson decay variable

The sub-system involving a boson decaying to a bottom pair ($X \to b\bar{b}$) is targeted using the angular distance of the decay products, $|\cos\theta^*|$. This variable is defined as the magnitude of the cosine of the angle between the X particle and one decaying quark in the ZMF of X. With this, the intrinsic spin of the boson particle is targeted. Spin-0 particles will produce isotropic angular distributions, while spin-1 particles will exhibit preferred directions in their decay products. Since the bottom quarks are indistinguishable in the detector, it is necessary a measurement independent of whether the quark or anti-quark is chosen. For this reason the absolute value of the cosine is taken, $|\cos\theta^*| = \frac{|\cos\theta^*| + |\cos(\theta^* + \pi)|}{2}$. Figure 7 illustrates the definition of this variable.

4.3 $t\bar{t} + X$ system

Although the formalism described in Sect. 4.1 was derived for $t\bar{t}$ production, it can be extended to $ij \to t\bar{t}X$ topologies ($2 \to 3$ processes), maintaining the same structure for the variables defined above. This approximation has proven to be a valuable framework for various analyses in dileptonic and lepton-plus-jets channels [47, 48, 54]. In the context of associated boson production ($pp \to t\bar{t}X$), the formalism applies directly to the $t\bar{t}$ pair. However, additional observables sensitive to the boson can further be included in the study.

[4] These calculations are performed in the laboratory frame, making them independent of boosts to the $t\bar{t}$ ZMF.
[5] It can be shown that c_{hel} is equivalent to the trace of the spin correlation matrix.

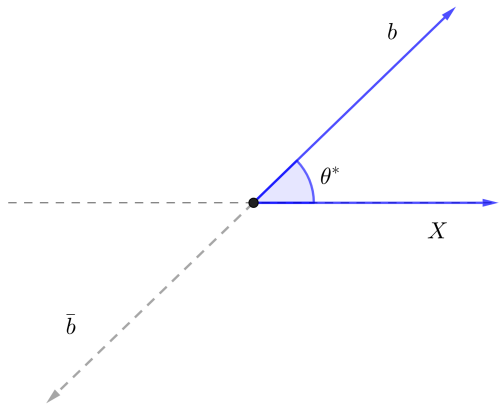

Figure 7: Illustration of θ^* as the angle between the X particle and one of its decay b-quarks in the rest frame of X. The black dot represents the X particle at rest. The two quarks have opposite directions in this frame.

In particular, the angular distribution between the three relevant particles is targeted in this work with the $\min(\Delta R_{\{t,\bar{t},X\}})$ variable. It consist on the minimum ΔR value among all three combinations of particles $\{t, \bar{t}, X\}$. Although this variable is correlated with the angular differences of the leptons, its inclusion is motivated as it provides a direct probe of the properties of the X particle.

This variable, sensitive to the whole $t\bar{t}X$ system, has the same caveat as $|\cos\theta^*|$, which is reconstructing the associated X particle . A detailed argument about how this is performed is presented in Sect. 7, however the difficulties it presents justify the caution when choosing variables that are dependent of the method used for the reconstruction of particles.

5 Simulated Samples

For this work simulated samples of proton-proton collision at LHC with a center-of-mass energy $\sqrt{s} = 13.6$ TeV, are used, corresponding to Run 3 conditions. Three distinct processes are analyzed: $t\bar{t}H(\rightarrow b\bar{b})$, which is the signal process of interest, where a Higgs boson is produced in association with a top-antitop quark pair, with the subsequently decay of the Higgs to a pair of b-quarks. Also, the pair production of top quarks with in association with jets, $t\bar{t}$+jets, as primary background process which closely resembles the topology of the signal. One last sample, is the Z boson production is association with a top quark pair $t\bar{t}Z(\rightarrow Q\bar{Q})$—with the Z boson decaying to any pair of quarks—, primary used as a control region process, that can be used to validate background model.

The datasets are simulated under the data-taking conditions of the years 2022 and 2023, corresponding to the initial phase of Run 3. Each year is further subdivided into two periods due to significant detector issues and changes in its operational configuration. These divisions ensure that variations in detector performance are properly accounted for during the analysis.

The 2022 datasets are categorized as *preEE* and *postEE*, reflecting a critical issue with the ECAL Endcap (EE) in September 2022 [56].Additionally, the 2023 datasets are also divided into *preBPix* and *postBPix*, due to a problem affecting the Barrel Pixel Detector (BPix) after the first technical stop in June 2023 [57].

5.1 Monte-Carlo simulation

The simulation of the signal and background processes in this analysis is performed using general-purpose MonteCarlo (MC) generators [58].

The event generation is performed at next-to-leading order (NLO) in QCD using the event generator POWHEG[59] for the signal $t\bar{t}H$ and background $t\bar{t}$ samples, and MADGRAPH5_AMCATNLO [60] for $t\bar{t}Z$. These event generators calculate the matrix element (ME) of the hard scattering and the relevant parton distribution functions (PDFs) are chosen accordingly. For all the samples the PDFs from the NNPDF3.1 set is employed [61].

The former simulation of the hard scattering is then interfaced with the hadronization process. For this task, PYTHIA8 generator is used [62], which simulates the fragmentation of quarks and gluons into hadrons, including both initial (IFR) and final state radiation (FSR). The hadronization process is modeled according to the QCD theory, ensuring that the generated particles match those observed in experiments.

Finally, the response of the detector is simulated using the GEANT4 framework [63], which models the interactions of the simulated particles with the detector material. This step produces simulated detector data, allowing the same reconstruction algorithms applied to real data to be used on the MC events. The detailed detector geometry and material properties are incorporated in this step to achieve a higher level of accuracy.

The MC samples are further adjusted to match the pileup conditions observed in the data. Since the pileup distributions in the data and simulation may differ, an additional weight is applied to the simulated events to match the pileup distributions. A list of all the simulated signal and background processes, along with their respective cross sections and relevant event generators, can be found in the Appendix B.

5.2 Particle Flow

The Particle Flow (PF) algorithm is designed to combine information from various detector subsystems, such as the tracking system, the calorimeters (ECAL and HCAL), and the muon system, to identify and reconstruct particles produced in collisions [64]. This multi-detector approach enables more accurate and efficient particle reconstruction by leveraging the available data from all subsystems. The algorithm follows a sequential procedure, consisting of the following steps:

1. **Tracking and clustering**: Charged particle tracks are reconstructed in the tracking system using a combinatorial approach [65]. This algorithm builds trajectories by iterating through the tracker layers, generating "seeds" from compatible hits, and performing a fit to determine the momentum and direction of each track. In parallel, energy deposits in the calorimeters are grouped into clusters using a topological clustering algorithm. This process begins by identifying energy peaks (seeds) and combining adjacent cells that exceed noise thresholds. The clusters are then fitted under a Gaussian mixture model to estimate their total energy and position.

2. **Information matching**: The PF algorithm correlates tracks with energy clusters and muon system hits to identify particles. Tracks from the inner tracker are matched to energy deposits in the ECAL and HCAL, as well as to signals in the muon chambers. This matching reduces redundancies and guarantees an optimal performance of the subdetectors.

3. **Particle identification**: Here, the particle identification follows a hierarchical approach based on the discriminability of particle signatures. Muons are firstly identified, as their distinct signals in the muon system make them the most straightforward to classify. Then, electrons are identified next, using the association between ECAL clusters and tracks. The reconstruction uses bremsstrahlung effects by combining nearby clusters into superclusters. Charged hadrons are assigned by matching the remaining tracks to energy clusters in the calorimeters. Finally, neutral particles such as photons and neutral hadrons are identified using residual calorimetric energy clusters not associated with tracks.

4. **Removal of used information**: After each particle is identified, the corresponding information from the subdetectors is removed from further consideration. This ensures that subsequent steps operate on unutilized data, preventing duplication or misclassification.

5.3 Object reconstruction

The reconstruction and identification of physical objects, such as particles, jets, and missing transverse momentum, is essential for analyzing the data collected by the CMS detector. The information from various subdetectors is processed to reconstruct the kinematic properties and identities of the particles produced in each event.

To achieve this, the PF algorithm plays a central role in the reconstruction process. Once PF blocks are formed, they are further analyzed to identify and reconstruct physical objects, which includes matching the reconstructed particles to specific signatures required for the analysis. In this step, criteria are optimized for each object type to enhance the precision of the reconstruction process while minimizing misidentifications. The Particle Flow framework helps streamline the identification and reconstruction of objects like electrons, muons, and jets, and it provides a detailed event description that mirrors the physical processes in proton-proton collisions.

In this subsection, key components and criteria involved in object reconstruction are introduced, setting the stage for the detailed definitions and reconstructions necessary later in the analysis.

5.3.1 Electrons

Electrons are reconstructed in the CMS detector by combining information from the tracking system and the ECAL. Due to their electric charge, electrons leave a distinct track in the inner tracking system, which is associated with a corresponding energy deposit in the ECAL, resulting from the electromagnetic shower they produce. This shower is influenced by Bremsstrahlung radiation, causing energy loss, which is taken into account during reconstruction. To accurately identify electrons, several criteria are applied. The reconstructed track must match the energy deposit in the ECAL, forming a supercluster (SC), and the shower shape in the ECAL is used as a key discriminator, with electrons typically having a narrower, more compact shower than hadrons. Isolation and identification cuts further reduce misidentifications, particularly from jets and photons, by requiring smaller energy deposits in the HCAL relative to the ECAL and applying isolation requirements to minimize contamination from nearby particles. Due to inefficiencies in the transition region between the barrel and endcap sections of the ECAL, electrons with pseudorapidity in the range $1.4442 < |\eta| < 1.5660$ are excluded.

For the data taking period of Run 3, two main approaches for electron identification are possible in the CMS detector: a cut-based identification—used in previous years— and a novel multivariate analysis (MVA) technique. In this work we use the MVA identification algorithm [66]. This method uses a combination of multiple variables to assign a probability to each electron candidate being a true electron [67]. The MVA approach incorporates more information and can provide better identification efficiency, particularly in cases where the separation between electron and non-electron particles is less clear. For the analysis, electron identification

criteria including isolation requirements and a working point with an efficiency of 80% are used [68].

5.3.2 Muons

Muons are primarily identified in the CMS detector using the inner tracking system and the muon chambers, as they deposit minimal energy in the calorimeters. Tracks are independently reconstructed in both subsystems and are considered "global muons" if their trajectories are compatible. To ensure precise identification, global muons must meet several quality criteria, including cuts on the fit to track measuring the goodness (χ^2) and a requirement for the reconstructed muon trajectory to intersect with multiple detection layers in the muon chambers, minimizing misidentifications from hadrons that penetrate into the muon system. Additionally, cuts on matched tracker and muon chamber hits help reduce muons from "in flight" decays. To further improve accuracy, the position of the reconstructed track relative to the primary vertex is restricted, with transverse (d_{xy}) and longitudinal (d_z) distances (impact parameter) requirements to minimize pileup effects. Isolation conditions are also applied, ensuring that the sum of the p_T of hadrons and photons within a cone of $\Delta R < 0.4$ around the muon, divided by the muon p_T, stays below a specified threshold. Corrections for pileup contributions are included in this calculation.

ID requirements are applied to candidates within the range $15 < p_T < 200$ GeV, and a reconstruction acceptance of $|\eta| < 2.4$ [69]. For the "tight" identification, additional conditions are imposed, such as a normalized $x^2/n_{dof} < 10$ for the global track fit, more than one matched station in the muon chambers, and more than five hits in the tracker. Isolation, vertex alignment, and other identification criteria are included when defining the working point. Muon reconstruction efficiency for the tight ID and tight PF isolation (corresponding to a Particle Flow isolation $\Delta R < 0.15$) is about $< 96\%$ [70] for the 2022 and 2023 period at $\sqrt{s} = 13.6$ TeV.

5.3.3 Jets

The reconstruction of jets aims to identify quarks and gluons by clustering their associated hadronic showers. In this thesis, the anti-k_t algorithm [71] is used with a distance parameter of $\Delta R = 0.4$ to match particle candidates into jets. To account the high number of particles per event during Run 3, the *Pileup Per Particle Identification* (PUPPI) algorithm is employed to moderate the impact of pileup on jet reconstruction [72].

Only jets with a $p_T > 30$ GeV and a pseudorapidity $|\eta| < 2.4$ are considered. Jets within $\Delta R < 0.4$ of any selected lepton are vetoed to prevent overlap between leptons and jets. Additionally, jet energy corrections (JEC) are applied. These corrections account for contributions from pileup, detector response, and residual differences between simulation and data.

Jet ID requirements are applied to restrict the energy fractions and composition of jet constituents, improving the purity of the reconstructed jets and reducing misidentifications. Moreover, events with jets in regions with irregular detector responses are excluded using veto maps.

b-tagging

Identifying jets originating from bottom quarks (*b-tagging*) is a fundamental task in analyses involving top quark decays. This process relies on advanced multivariate algorithms that exploit a wide range of jet features, such as the position of secondary vertices and the particle composition within the jet.

For Run 2, the state-of-the-art tagger was DeepJet [73], a deep neural network designed for multi-classification tasks. It utilizes low-level properties of charged and neutral particle-flow jet constituents, complemented by secondary vertex information, achieving excellent heavy flavor tagging performance.

For Run 3, new-generation algorithms have been introduced to further enhance b-tagging performance. The ParticleNet algorithm [74], represents jets as unordered "particle clouds" rather than ordered collections of constituents, preserving permutation invariance. It uses a dynamic graph convolutional neural network to effectively incorporate low-level jet information, excelling in heavy flavor tagging and hadronic tau identification. Additionally, the RobustParTAK4 model employs a modified ParticleTransformer [75] architecture, which introduces pairwise interaction features between jet constituents and secondary vertices. It also incorporates Adversarial Training to improve robustness against mismodeling in Monte Carlo simulations. This feature distorts input features during training to help the model classify jets more reliably across varied input distributions.

The last two algorithms represent significant advancements in b-tagging, leveraging state-of-the-art neural network architectures to enhance sensitivity and robustness in identifying bottom quark jets. Their complementary designs enable optimized performance across a range of jet topologies ensuring precise and efficient classification.

5.3.4 Missing transverse momentum

The missing transverse momentum ($\boldsymbol{p_T^{\text{miss}}}$), also called missing transverse energy or MET, is a crucial observable in this thesis which dileptonic decays from W bosons. It is defined as the negative vectorial sum of the transverse momenta of all detected final state particles:

$$\boldsymbol{p_T^{\text{miss}}} = -\sum_i \boldsymbol{p_{T,i}}.$$

Due to energy-momentum conservation and the very small transverse momentum of the colliding protons before the interaction, $\boldsymbol{p_T^{\text{miss}}}$ corresponds to the transverse momentum carried by undetected particles. It is common, in processes such as top quark pair production, that $\boldsymbol{p_T^{\text{miss}}}$ primarily arises from neutrinos escaping the detector.

However, mismeasurements of the p_T of reconstructed particles can contribute substantially to $\boldsymbol{p_T^{\text{miss}}}$, which requires to be taken in consideration when calculating the MET and its uncertainties. The recommendation from CMS is to use the PUPPI algorithm for the MET reconstruction.

6 Event selection

A baseline selection to target events containing the key particles needed for the analysis, is build with a number of cuts applied to the data. These cuts define a specific phase space carefully studied to optimize the object selection, as suggested in the work of K. Krasenbrink. With this, non relevant events are filter-out providing a relatively clearer signal sample for $t\bar{t}H(\rightarrow b\bar{b})$ dileptonic. In particular, the selection targets the decay products of the top quark, the top antiquark, and a third particle, which may be a Higgs boson, a gluon, or a Z boson, depending on the process under consideration.

1. At least four jets $N_{\text{jets}} \geq 4$, with minimum jet transverse momentum $p_{T,\text{jet}} > 30$ GeV, and maximum jet pseudorapidity $|\eta_{\text{jet}}| < 2.4$.
2. Exactly two oppositely charged leptons (any combination of electrons and muons is considered).
3. Muons within the range $|\eta| < 2.4$.
4. Electrons within the range $|\eta| < 2.5$.
5. A leading(subleading) lepton with $p_T > 25(15)$ GeV.
6. Invariant mass of e^+e^- and $\mu^+\mu^-$ pairs: $m_{ee/\mu\mu} < 86$ GeV or $m_{ee/\mu\mu} > 96$ GeV.
7. Minimum invariant mass of same-flavor lepton pairs: $m_{ee/\mu\mu} > 20$ GeV.

The first cut is the minimal requirement to select events with jets originating from the decay of top quarks or boson —this plays a crucial role for the event reconstruction presented in Sect. 7. While for a high signal purity it would be necessary to include at least three or four b-tagged jets, statistical limitations for reconstruction—explained also in Section 7— led to a more relaxed constraint, with no minimum number of b-tagged jets required. The requirement for oppositely charged leptons, targets events where both top quarks decay through the dileptonic channel. For same-flavor leptons, additional criteria on their invariant mass are imposed to exclude the Drell-Yan window of Z boson production. Additional cuts are based on the object reconstruction specifications for each type of particle following the IDs criteria.

On the following, the concept *detector-level* information will be used numerous times. It refers to any process or object reconstruction with the information provided from the detector, i.e., objects presented in Sect. 5.3. For example, jets are the detector-level information of quarks and gluons after parton showering. In the case of leptons, these correspond to the same object before and after detector-level, however, additional challenges can arise, such as non-prompt leptons which will increase the complexity of the identification process.

On the other hand, the *particle-level* information, describes the objects independent of the detector effects; such as bare quarks, leptons and neutrinos among others. Accessing this kind of information is one strength of using MC simulations, which allow the study of processes at a elementary-particle stage. Whose results can help the understanding of observations made

with experimental data and also provide useful information of detector effects. However, particle-level analyses must be carried and interpreted with caution, since MC samples are typically based on the SM and possible BSM effects will not be visible. Nevertheless, this type of limitation is not expected to impact in the work.

Before continuing is important to clarify some aspects. Although the following analysis focuses exclusively at particle-level it is necessary to reconstruct the intermediate particles in order to build the observables. For such task, some requirements are applied to the particles which is a sort of "event reconstruction" at particle-level stage which is not common in the majority of analysis. The idea of introducing the above event selection, which will be use later in the following section with detector-level particles, is to make the reader familiar with applying a criteria to filter-out events that are not important for the signal extraction.

It would be possible to construct the observables directly from the particle-level information of the top quarks and the associated gluon (depending on the process). However, to ensure a meaningful comparison between particle- and detector-level results, the observables must be built following the same reconstruction steps. Therefore, in the next sub-sections, the derivation will start from the decay products—quarks, gluons, and leptons—ensuring consistency throughout the analysis.

6.1 Polarization and spin observables in dileptonic events

The treatment given in Section 4.1, where the coefficients function were defined and a specific basis was presented, was mainly theoretical and using exclusively top spin states. From an experimental point of view, to build the observables, a slightly different basis must be chosen along with the particles used to reconstruct the top quarks.

As mentioned above, the focus relies on LHC $t\bar{t}$ production with top quarks decaying in the dileptonic channel:

$$pp \to t\bar{t}X \to \ell^+\ell^- + \text{jets} + \text{MET}, \tag{7}$$

with $\ell = e, \mu$. The final products are described as detector-level objects, with the jets as the result of hadronization of the bottom quarks and the MET as the visible neutrinos' signal.

Since the polarization and spin observables cannot be measured directly using the top quarks, the lepton directions are used instead due to its high spin analysing power ($\kappa_{\ell^+} \approx 1$). Additionally, for proton-proton collisions, a basis based on Eq. (4) is built, choosing $\hat{\boldsymbol{p}}_{\boldsymbol{p}} = (0, 0, 1)$ as the direction of the proton beam in the $+\hat{\boldsymbol{z}}$ direction in the CMS laboratory reference frame (see Sect. 3.1.1), and constructing $\hat{\boldsymbol{n}}_{\boldsymbol{p}}$ and $\hat{\boldsymbol{r}}_{\boldsymbol{p}}$ accordingly.

With all these considerations, the four-fold angular distribution for the leptons can be derived and obtaining the concerning observables [47]. The observables with their corresponding coefficient function are summarized in Table 1. The notation used here is lower-index of 1 and 2 for the positive and negative lepton, respectively. While θ_1^i and θ_2^j are the angles of such

Table 1: Observables and their corresponding polarization and spin coefficient functions.

Observable	Coefficient function
$\cos\theta_1^k$	b_k^+
$\cos\theta_2^k$	b_k
$\cos\theta_1^r$	b_r^+
$\cos\theta_2^r$	b_r
$\cos\theta_1^n$	b_n^+
$\cos\theta_2^n$	b_n
$\cos\theta_1^k \cos\theta_2^k$	c_{kk}
$\cos\theta_1^r \cos\theta_2^r$	c_{rr}
$\cos\theta_1^n \cos\theta_2^n$	c_{nn}
$\cos\theta_1^r \cos\theta_2^k + \cos\theta_1^k \cos\theta_2^r$	c_{rk}
$\cos\theta_1^r \cos\theta_2^k - \cos\theta_1^k \cos\theta_2^r$	c_n
$\cos\theta_1^n \cos\theta_2^r + \cos\theta_1^r \cos\theta_2^n$	c_{nr}
$\cos\theta_1^n \cos\theta_2^r - \cos\theta_1^r \cos\theta_2^n$	c_k
$\cos\theta_1^k \cos\theta_2^n + \cos\theta_1^n \cos\theta_2^k$	c_{kn}
$\cos\theta_1^k \cos\theta_2^n - \cos\theta_1^n \cos\theta_2^k$	c_r
c_{hel}	$-(c_{kk} + c_{rr} + c_{nn})/3$

leptons with respect to the direction of the i and j axis ($\hat{\boldsymbol{k}}_{\boldsymbol{p}}$, $\hat{\boldsymbol{r}}_{\boldsymbol{p}}$ and $\hat{\boldsymbol{n}}_{\boldsymbol{p}}$), in the rest frame of each parent top quark.

6.2 Quantifying the statistical separation power of observables

In order to quantitatively assess the difference in shape between the signal and background distributions, it is necessary to define a metric for their distinction. One widely used method in multivariate analysis (MVA) [67] is the *separation* of a variable x, denoted by S^2, which is defined as:

$$S^2 = \frac{1}{2} \int \frac{\left(\hat{f}_S(x) - \hat{f}_B(x)\right)^2}{\hat{f}_S(x) + \hat{f}_B(x)} \, dx \tag{8}$$

where $\hat{f}_S(x)$ and $\hat{f}_B(x)$ represent the probability density functions of the signal and background distributions of x, respectively. In this context, the separation S^2 is zero when the signal and background distributions are identical in shape, and it is one when the two distributions do not overlap.

Equation (8) can also be expressed in terms of the histogram binning of the distributions, allowing for a discrete calculation of S^2. Additionally, the uncertainty on S^2 can also be eval-

uated. An expression for the discrete sum is given by:

$$
\begin{aligned}
S^2 &= \frac{1}{2}\sum_i \frac{\left(f_{i,S} - f_{i,B}\right)^2}{f_{i,S} + f_{i,B}} \\
\sigma_{S^2} &= \frac{1}{2}\sum_i \frac{f_{i,S} - f_{i,B}}{\left(f_{i,S} + f_{i,B}\right)^2} \sqrt{\left(f_{i,S} + 3f_{i,B}\right)^2 \sigma_{i,S}^2 + \left(-3f_{i,S} + f_{i,B}\right)^2 \sigma_{i,B}^2},
\end{aligned}
\tag{9}
$$

where the index i runs over each bin of the histogram. Here, $f_{i,S}$ and $f_{i,B}$ denote the number of events in the i-th bin for the signal and background, respectively, and $\sigma_{i,S}$ and $\sigma_{i,B}$ are the associated uncertainties for each bin. The distributions are normalized following $\sum_i f_{i,S} = \sum_i f_{i,B} = 1$.

6.3 Distributions of observables for the $t\bar{t}H$, $t\bar{t}Z$ and $t\bar{t} + b\bar{b}$ processes

An initial examination of the observables is performed at particle level (i.e., without applying the event selection criteria from Sect. 6). This approach provides valuable insights to optimize the analysis described in Sect. 7.2. Specifically, the spin correlation observables presented in section 4.1 are evaluated alongside additional variables introduced in Sects. 4.2 and 4.3. By analyzing these distributions, it is possible to discard variables that exhibit poor separation between signal and background, allowing the subsequent study at detector level focusing exclusively on the most discriminative observables.

The particle-level distributions, evaluated in the inclusive phase space, are presented in Figures 8, 9, and 10. Additionally, the transverse momentum (p_T) of the bottom quark pair closest in ΔR is included in the analysis for comparison purposes (Figure 11). This variable ranks among the most relevant in the BDT employed for signal-to-background enhancement [5]. Its inclusion here serves as a "standard candle," providing a benchmark for comparing the separation performance of the variables from this work. The distributions consider the three samples, with the signal process $t\bar{t}H$, where the Higgs boson decays to a $b\bar{b}$ pair. The only requirement in this inclusive phase space the presence of dileptonic events, characterized by two oppositely charged leptons of any flavour ($e^{\pm}$ or $\mu^{\pm}$), two neutrinos ($\nu\bar{\nu}$), and four bottom quarks. This setup allows the reconstruction of the necessary particles for calculating the observables.

Figures 8 and 9 display the polarization and spin correlation observables, respectively. These observables are calculated by projecting the momentum of the lepton ($\hat{\boldsymbol{l}}_1$) or antilepton ($\hat{\boldsymbol{l}}_2$) onto the coordinate axes $\{\hat{\boldsymbol{k}}, \hat{\boldsymbol{r}}, \hat{\boldsymbol{n}}\}$. Under the assumption that the top and antitop quarks share the same polarization, the observables are expressed as sums and differences of the two particles' contributions. Additional angular observables are shown in Figure 10, including the azimuthal ($\Delta\phi$) and pseudorapidity ($\Delta\eta$) differences of the lepton-antilepton pair, the variable $|\cos\theta^*|$ (see Figure 7), and the minimum ΔR among any combination of the Higgs boson, top quark, and antitop quark.

Distributions are normalized across the three samples, since the focus remain on the shape comparison between signal and background, providing useful insights for the signal signif-

icance. Such normalization is essential for evaluating the separation power of the variables and improves the visualization of the distributions, given the substantial differences in cross sections between the processes.

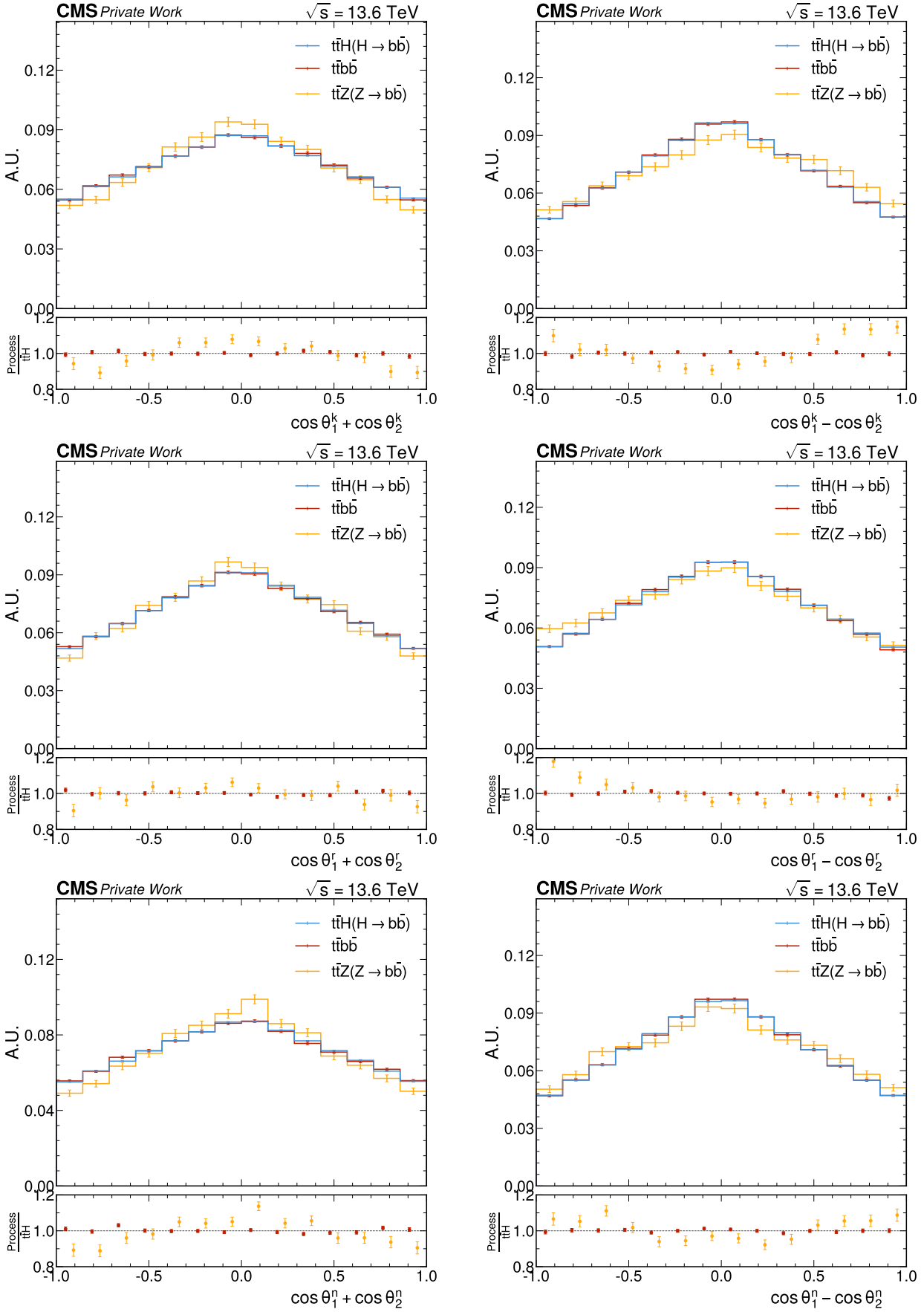

Figure 8: Sum (left) and difference (right) of the polarization observables $\cos\theta_1^i \pm \cos\theta_2^i$ shown in Table 1, projected on different axis. Three processes are included: the $t\bar{t}H(\rightarrow b\bar{b})$ signal (blue line), $t\bar{t}+b\bar{b}$ background (red line) and $t\bar{t}Z(\rightarrow b\bar{b})$ (yellow line). Uncertainties are purely statistical. Ratio plots (below) are calculated for the background and $t\bar{t}Z$ over the signal.

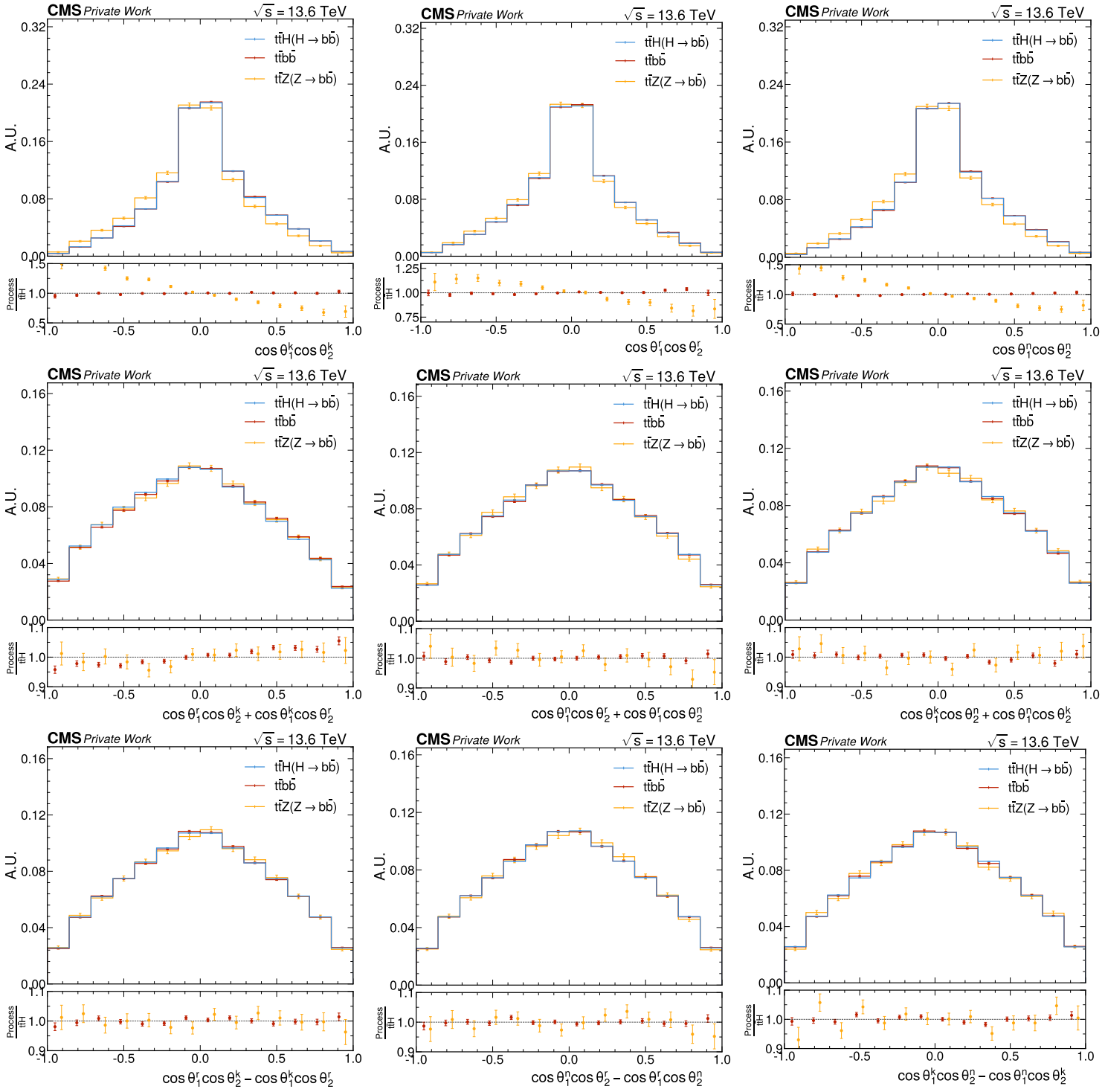

Figure 9: Spin correlation observables from Table 1 projected on three basis axes. Three processes are included: the $t\bar{t}H(\rightarrow b\bar{b})$ signal (blue line), $t\bar{t}+b\bar{b}$ background (red line) and $t\bar{t}Z(\rightarrow b\bar{b})$ (yellow line). Uncertainties are purely statistical. Ratio plots (below) are calculated for the background and $t\bar{t}Z$ over the signal..

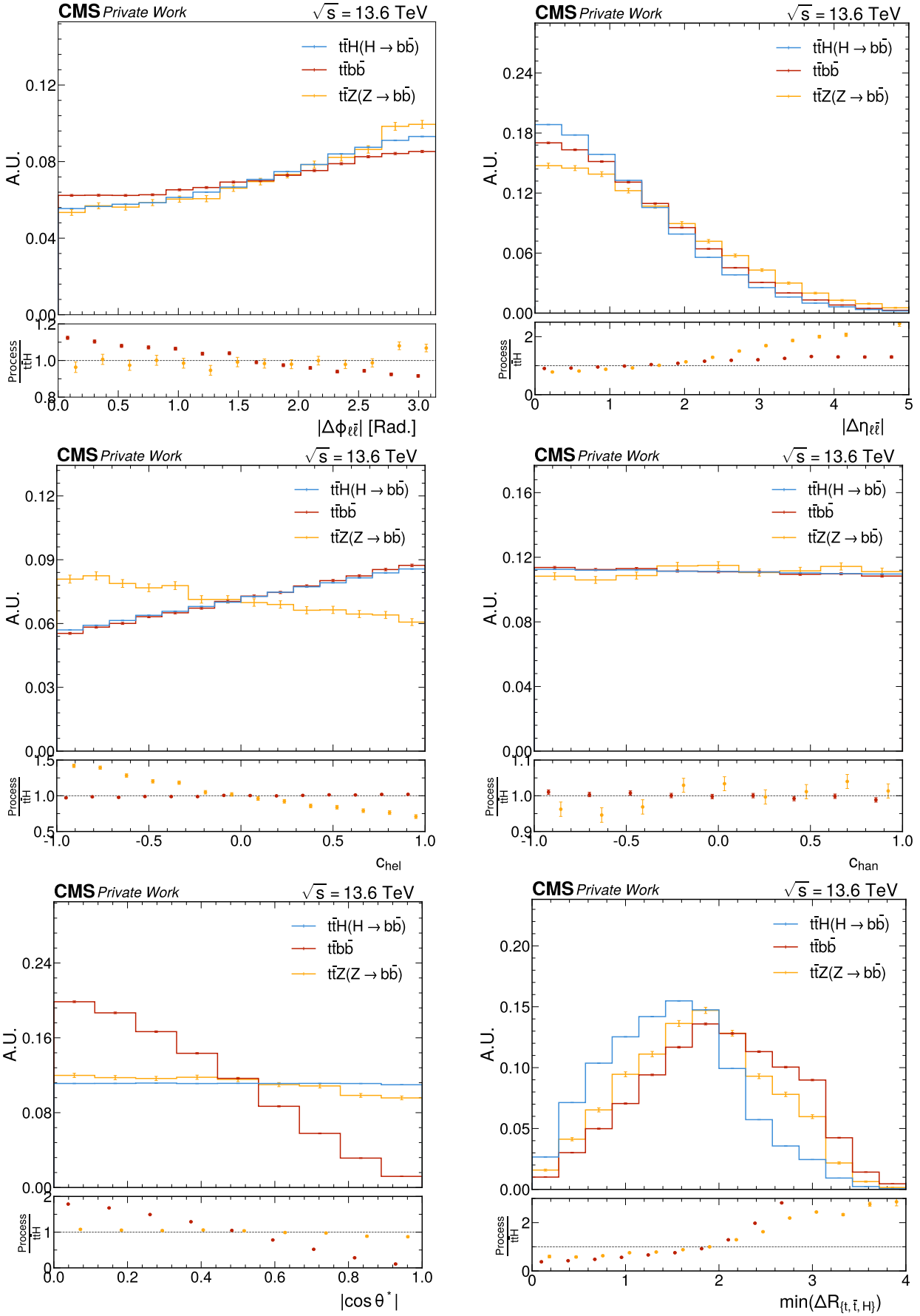

Figure 10: Distributions of absolute differences in ϕ and η of the lepton-antilepton pair (top row). c_{hel} and c_{han} spin information variables (middle row). The $|\cos\theta^*|$ (bottom left), is the angle between the bottom quark in the rest frame of a gluon particle, and the direction of the gluon. The minimun ΔR (bottom right) of any combination from the top, antitop and Higgs. Three processes are included: the $t\bar{t}H(\rightarrow b\bar{b})$ signal (blue line), $t\bar{t}+b\bar{b}$ background (red line) and $t\bar{t}Z(\rightarrow b\bar{b})$ (yellow line). Uncertainties are purely statistical. Ratio plots (below) are calculated for the background and $t\bar{t}Z$ over the signal.

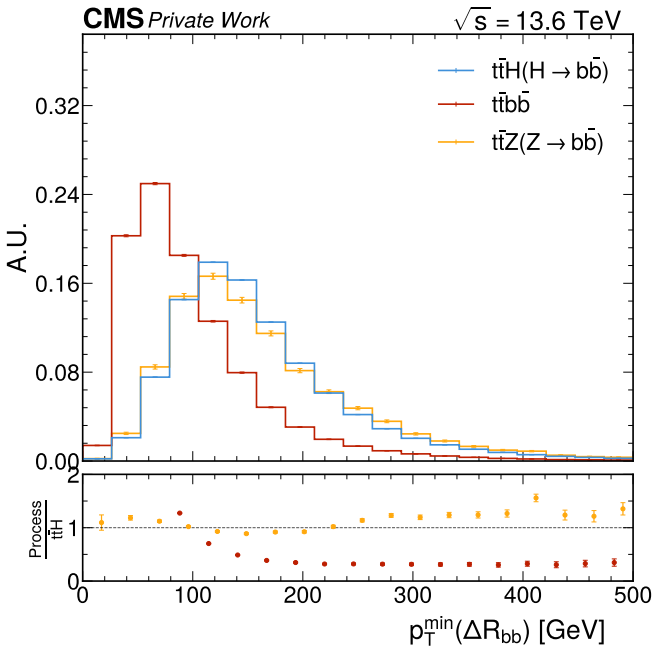

Figure 11: Transversal momentum of the two closest, in ΔR, bottom quarks. Three processes are included: the $t\bar{t}H(\rightarrow b\bar{b})$ signal (blue line), $t\bar{t}+b\bar{b}$ background (red line) and $t\bar{t}Z(\rightarrow b\bar{b})$ (yellow line). Uncertainties are purely statistical. Ratio plots (below) are calculated for the background and $t\bar{t}Z$ over the signal.

Only statistical uncertainties are considered in this analysis, represented by the error bars in the distributions. These uncertainties are calculated assuming Poisson statistics, with the standard deviation given by $\sigma = \sqrt{N}$, where N represents the number of counts in each bin. Additionally, ratio plots are included below each distribution, illustrating the fraction of events in the background samples relative to the $t\bar{t}H$ signal process. These ratio plots provide valuable insights into the discriminating power of the variables under study.

Inspecting Figures 8 and 9, the polarization and spin correlation observables of the $t\bar{t}$ system exhibit nearly identical shapes for the signal and background distributions, with peaks centered around zero, as predicted by theory [48]. Minor discrepancies are observed in the $t\bar{t}Z$ sample for the polarization observables and the diagonal spin correlation coefficients (C_{ii}), although these differences present a relative low separation power. However, these observables offer limited separation power for distinguishing signal from background and will therefore be excluded from further analyses in the subsequent section. Diagonal elements = chel

Regarding the angular variables in Figure 10, several distributions show greater discrimination. The angular differences of the lepton-antilepton pair ($|\phi_{\ell\bar{\ell}}|$ and $|\eta_{\ell\bar{\ell}}|$, top row) display similar shapes across all samples, with minor variations at extreme values. The observables c_{hel} and c_{han} exhibit similar trends for the $t\bar{t}H$ signal and the $t\bar{t}+b\bar{b}$ background (elaborar diciendo que tienen mas predominancia hacia uno lo que refleja una mayor correlacion de los top quarks), whereas the $t\bar{t}Z$ sample shows an opposite behaviour for c_{hel}.

The variable $|\cos\theta^*|$ stands out as the most distinctive when comparing shapes between the $t\bar{t}H$ signal and the main $t\bar{t}+b\bar{b}$ background. This result aligns with theoretical expectations; the Higgs boson, being a scalar particle with spin-0, decays isotropically into a $b\bar{b}$ pair, resulting in a flat angular distribution. In contrast, the background samples ($t\bar{t}Z$ and extra jets from $t\bar{t}$+jets production) involve vector bosons with spin-1, leading to characteristic angular distributions for the $b\bar{b}$ pair [16]. These differences are clearly observed in the simulated data.

Concerning the last observable in Figure 10 (bottom right), representing the minimum ΔR among any combination of the Higgs, top, and antitop quarks, also shows significant separation between signal and background. For the $t\bar{t}H$ sample, events with lower ΔR values are more frequent, indicating that the Higgs boson is often more collinear with the top quark compared to other bosons such as the Z or jets from $t\bar{t}$+jets production.

The reference variable used in the analysis [5] is presented in Figure 11. The distribution for the $t\bar{t}+b\bar{b}$ background peaks at lower p_T values compared to the $t\bar{t}H$ and $t\bar{t}Z$ samples. It is important to note that this variable is not entirely independent of $\min(\Delta R_{\{t,\bar{t},H\}})$, as both are influenced by the kinematics of bottom quarks. However, $p_T^{\min}(\Delta R_{bb})$ is closely tied to the transverse momentum of the bosonic particles, whereas the new variable introduced in this work focuses on the angular distance between the relevant boson and the top quarks.

Table 2 summarizes the calculated separation values for each variable, along with their corresponding statistical uncertainties. These values provide quantitative metrics to assess the discriminating power of each observable in separating the $t\bar{t}H$ signal from the background processes.

Based on these results, five variables are selected for further study at particle-level, ranked by their separation power (from highest to lowest): $|\cos\theta^*|$, $\min(\Delta R_{t,\bar{t},H})$, $|\Delta\eta_{\ell\bar{\ell}}|$, $|\Delta\phi_{\ell\bar{\ell}}|$, and c_{hel}. There is a difference of two orders of magnitude in separation power between the first two variables and the lepton-pair observables. Nevertheless, the latter are included in the analysis due to their independence from the reconstruction of the top quarks and the boson. Additionally, although c_{hel} has a relatively low separation power, it offers valuable information for distinguishing the $t\bar{t}Z$ sample from the signal.

The motivation for including the $t\bar{t}Z$ process in the analysis was to use it as a control region, orthogonal to the signal, to model the challenging $t\bar{t}+b\bar{b}$ background. However, it has been shown to provide relatively good separation power for the c_{hel} variable. Therefore, in future analyses aiming to compare $t\bar{t}Z$ and $t\bar{t}H$, incorporating this process could be beneficial. The separation power of $t\bar{t}Z$ and the signal can be found in appendix C.

Table 2: Summary of S^2 values with the statistical uncertainty for the $t\bar{t} + b\bar{b}$ background and $t\bar{t}Hb\bar{b}$ signal samples for each variable, in descending order. The last row contains one of the best variables for the BDT used in the [5] analysis.

Variable	$S^2 \pm$ (stat.)
$\lvert\cos\theta^*\rvert$	$(1.0078 \pm 0.0055) \times 10^{-1}$
$\min(\Delta R_{t,\bar{t},H})$	$(9.840 \pm 0.045) \times 10^{-2}$
$\lvert\Delta\eta_{\ell\bar{\ell}}\rvert$	$(2.554 \pm 0.079) \times 10^{-3}$
$\lvert\Delta\phi_{\ell\bar{\ell}}\rvert$	$(1.120 \pm 0.055) \times 10^{-3}$
$\cos\theta_1^r \cos\theta_2^k + \cos\theta_1^k \cos\theta_1^r$	$(1.22 \pm 0.18) \times 10^{-4}$
c_{hel}	$(5.1 \pm 1.2) \times 10^{-5}$
$\cos\theta_1^n + \cos\theta_2^n$	$(3.4 \pm 1.1) \times 10^{-5}$
$\cos\theta_1^r \cos\theta_2^r$	$(2.91 \pm 0.88) \times 10^{-5}$
$\cos\theta_1^n \cos\theta_2^n$	$(2.57 \pm 0.84) \times 10^{-5}$
$\cos\theta_1^r + \cos\theta_2^r$	$(2.24 \pm 0.90) \times 10^{-5}$
$\cos\theta_1^r - \cos\theta_2^r$	$(2.20 \pm 0.89) \times 10^{-5}$
$\cos\theta_1^k \cos\theta_2^n - \cos\theta_1^n \cos\theta_2^k$	$(2.18 \pm 0.77) \times 10^{-5}$
$\cos\theta_1^k \cos\theta_2^k$	$(2.07 \pm 0.75) \times 10^{-5}$
$\cos\theta_1^k \cos\theta_2^n + \cos\theta_1^n \cos\theta_2^k$	$(1.97 \pm 0.73) \times 10^{-5}$
$\cos\theta_1^k + \cos\theta_2^k$	$(1.80 \pm 0.83) \times 10^{-5}$
$\cos\theta_1^r \cos\theta_2^k - \cos\theta_1^k \cos\theta_2^r$	$(1.70 \pm 0.68) \times 10^{-5}$
c_{han}	$(1.60 \pm 0.66) \times 10^{-5}$
$\cos\theta_1^n - \cos\theta_2^n$	$(1.23 \pm 0.65) \times 10^{-5}$
$\cos\theta_1^n \cos\theta_2^r + \cos\theta_1^r \cos\theta_2^n$	$(1.22 \pm 0.58) \times 10^{-5}$
$\cos\theta_1^k - \cos\theta_2^k$	$(1.13 \pm 0.63) \times 10^{-5}$
$\cos\theta_1^n \cos\theta_2^r - \cos\theta_1^r \cos\theta_2^n$	$(1.01 \pm 0.52) \times 10^{-5}$
$p_T[\min(\Delta R_{bb})]$	$(2.1283 \pm 0.0074) \times 10^{-1}$

7 Event reconstruction

Until this point, all observables have been calculated at particle-level. While these results are important from a theoretical perspective, they cannot be directly compared to experimental data. Instead, the analysis must be performed at detector-level, where physical observables are derived from the objects identified by the detector (Sect. 5.3). This section details the process of reconstructing all relevant particles from detector-level information and generating the corresponding observables.

Reconstructing events from detector-level information poses significant challenges. In this work, the focus is on the $t\bar{t}H$ signal, with two main problems of interest. The first concerns the dileptonic decay channel, where the reconstruction of both neutrinos is required. At detector level, the only measurable quantity related to neutrinos is the missing transverse momentum ($\not{p}_T$). Typically, analytic approaches are used for this reconstruction [76], involving six parameters in total (three momentum components for each neutrino, assuming they are massless). These methods use the known masses of the W-boson and the top quark to constrain the free parameters and recover the neutrino kinematics.

The second major challenge is jet assignment. At detector-level, quarks and gluons cannot be detected at its elementary state, leaving a trace of particles in a certain direction (jets). Commonly, designated scores like b-tagging, c-tagging, or light-flavor tagging are employed [77] to identify the jets provenance, i.e. how likely they are originated from a specific quark. These scores estimate the likelihood that a jet originates from a specific type of quark, using indirect methods based on secondary vertex reconstruction, track information, and impact parameter algorithms [78]. However, these approaches rely on limited information. Recent advancements in machine learning have shown promising results in jet assignment tasks, leveraging state-of-the-art techniques [79], [80].

This work employs SPANet, a deep learning architecture, to address both challenges in particle reconstruction. With the network is possible to assign jets originating from b-quarks to their respective parent particles (top quark, antitop quark, and Higgs boson), enabling the calculation of observables at detector level. The following sections describe the architecture of SPANet and its performance in the dileptonic channel.

7.1 Jet assignment and neutrino regression with the SPANet model

Reconstructing heavy particles such as the top quark and Higgs boson presents significant challenges, many of which can be addressed using machine learning techniques. SPANet [81] is a specialized neural network architecture which target events involving top quarks, offering a novel approach to event reconstruction through symmetry-preserving attention mechanisms. These mechanisms explicitly incorporate the intrinsic symmetries of the problem, providing an efficient framework for solving jet-parton assignment tasks.

Traditionally, jet assignment has been addressed using analytic methods such as χ^2 fits [82] or

kinematic likelihood methods [83]. However, these approaches face some limitations such as the combinatorial explosion caused by the need to evaluate all possible jet permutations. In high-energy physics experiments, such as those conducted at the LHC, processes with high jet multiplicity and backgrounds with similar topology as signal, further degrade the efficiency and accuracy of these standard methods.

SPANet overcomes these challenges by leveraging symmetries to significantly reduce the number of jet permutations required. Initially developed for the reconstruction of fully hadronic $t\bar{t}$ (i.e. with the W from the top decaying exclusively to a pair of quarks, which produces six jets in the final state), SPANet has been adapted to handle an arbitrary number of objects. Moreover, it incorporates auxiliary features such as regression for neutrino reconstruction in semi- and dileptonic events, as well as signal-vs-background classification. By efficiently assigning jets to their parent partons, SPANet not only enhances reconstruction accuracy but also reduces computational complexity, making it a robust tool for reconstructing heavy particle decays in high-energy physics.

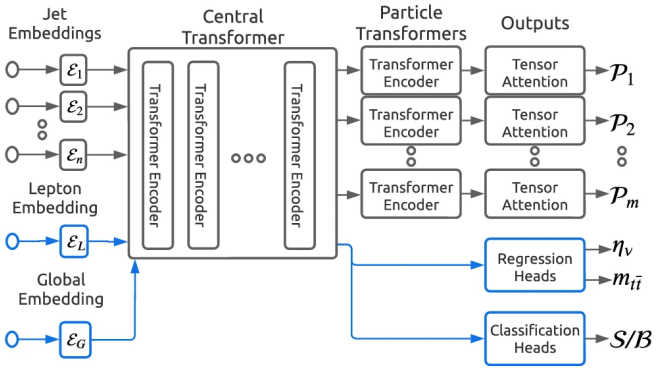

Figure 12: Diagram of the SPANet architecture. Data flows from left to right: inputs are denoted by ε_i, assignment outputs by P_j, regression results such as η_ν and $m_{t\bar{t}}$, and a classification output S/B. Components from previous implementations are shown in black, while newly introduced elements are highlighted in blue. Taken from [84].

Figure 12 illustrates the SPANet architecture, with new added features (highlighted in blue) enabling the reconstruction of dileptonic events. The input consists of vectors representing the kinematic properties of detector-level objects, such as jets, leptons, and the missing transverse momentum ($\boldsymbol{p}_{\boldsymbol{T}}^{\text{miss}}$). These vectors are embedded into a high-dimensional latent space and processed by a central transformer encoder [84]. This encoder integrates relative information between all jets, capturing angular correlations, energy distributions, and global event patterns. To ensure permutation invariance and avoid introducing an artificial order among the inputs, positional embeddings are omitted.

Following the central encoder, SPANet employs individual transformers customized to each resonance particle. These transformers extract relevant features for reconstructing specific particles, utilizing learned symmetries to optimize the jet assignment process. A key innovation of SPANet is its ability to compute the joint probability of each jet being assigned to a given res-

onance particle. This is accomplished through a symmetric tensor attention layer [85], which embeds symmetry constraints directly into the attention mechanism. The resulting output is a joint probability distribution over jets, reflecting their likelihood of being associated with particular resonance particles.

To address the factorial growth of computational complexity associated with evaluating all jet permutations, SPANet divides the network into separate branches for each resonance particle. This design reduces the runtime complexity from $O(N!)$ to $O(N^{k_p})$, where k_p represents the number of decay products of the resonance particle. This efficiency enables SPANet to perform robust event reconstruction even in scenarios with high jet multiplicity, making it particularly well-suited for tasks such as dileptonic $t\bar{t}H$ reconstruction.

7.1.1 Network training

The training of SPANet is performed using the $t\bar{t}H$ sample, applying the event selection criteria outlined in Sect. 6. The first step involves generating the true assignment labels, $\mathcal{T}$, which encode the particle-level information, specifically the provenance of b-quarks (i.e., whether they originate from the top quark or the Higgs boson). This assignment process matches each quark one-to-one with a corresponding detector-level jet, provided that their angular separation satisfies $\Delta R < 0.4$. This process enables the network to correctly associate jets with their respective resonance particles, allowing it to learn essential features from the data, and improving the performance to distinguish between jets originating from different resonance particles.

Additionally, for leptonic events, neutrino information is crucial. SPANet is capable of reconstructing both neutrinos (in dileptonic events) through a regression module introduced in the latest version of the model [81]. The kinematics of the neutrino and antineutrino are derived at the particle level and subsequently introduced into the regression part of the network configuration (see Fig. 13). The optimal set of variables for the regression are the three-momentum components of the neutrino and antineutrino: p_x, p_y and p_z.

The following information is used as input for the training process:

- **Jets:** p_T, η, ϕ, M, b-tagging, $\Delta R_{j,\ell}$ and $m_{j,\ell}$.
- **Leptons:** p_T, η, ϕ, M and a lepton-flavour key.
- **MET:** p_T and ϕ.

Here, p_T, η, ϕ, and M (mass) represent the kinematic information of the objects. Additionally, the b-tagging score (ranging from 0 to 1) is used to identify jets that are more likely to originate from a bottom quark. Three b-tagging algorithms, as described in Sect. 5.3.3, are included as inputs. Although these algorithms are correlated, they provide complementary information that aids in identifying jets from resonance particles with high b-tagging scores. Additionally, the ΔR of the jets with the leptons as well their invariant mass (jet and lepton), are included in the network. This is behind the idea that jets coming from the top quark will have certain

INPUTS		**EVENT**	**REGRESSIONS**
– Sequential:	**– Global:**	**– t**	**– Event:**
— Jets:	**— Leptons:**	— b	— ν, p_x
—— p_T	—— p_T	**– $\bar{t}$**	— ν, p_y
—— η	—— η	— b	— ν, p_z
—— ϕ	—— ϕ	**– H**	— $\bar{\nu}, p_x$
—— M	—— M	— b_1	— $\bar{\nu}, p_y$
—— *b-tagging*	—— *flavour*	— b_2	— $\bar{\nu}, p_z$
—— $\Delta R_{j,\ell^+}$			
—— $\Delta R_{j,\ell^-}$	**— MET**	**PERMUTATIONS**	**CLASSIFICATIONS**
—— m_{j,ℓ^+}	—— p_T	**– H**	
—— m_{j,ℓ^-}	——ϕ	— $[b_1, b_2]$	

Figure 13: Configuration of SPANet for training: The INPUTS include the kinematics of leptons and jets, as well as the missing transverse momentum. Neutrino information is incorporated into the REGRESSIONS module. The event topology, describing the resonance particles and their decay products, is specified in the EVENT section. Symmetries are integrated through the PERMUTATIONS module. A signal-to-background CLASSIFICATION task is also supported, though not used in this work.

correlation to the leptons as it would be a "pseudo-top" reconstruction despite not including the neutrino. While jets not coming from the top will have zero correlation with the leptons. Thus introducing this information could help SPANet identifying correctly the provenance of jets. For leptons, a boolean input is used to distinguish whether the lepton is an electron or a muon, accounting for differences in reconstruction efficiencies and misidentification rates [68, 70]. This information allows SPANet to adjust its training for events with muons or electrons (with reconstruction efficiencies of 99% and 80%, respectively), improving overall reliability.

The sample is divided into three subsets for the training process: a training set, a validation set, and a test set. The training set is used to train the network's parameter, while the validation set helps monitor and avoid overfitting. The test set is kept independent from the training process and is used to evaluate the performance on unseen data. The number of events and the corresponding fractions for each of these subsets are shown in Appendix A.

The hyperparameter tuning proposed by [85] was specifically optimized for the $t\bar{t}H$ (full-hadronic) process and is applied in this work. These optimized values are provided in the Appendix.

This analysis represents the first application of SPANet to a dileptonic sample, testing its capabilities in a more complex environment. The results obtained are novel and provide valuable insights into SPANet's performance in reconstructing events with leptonic final states.

7.1.2 Jet assignment and neutrino regression performance

The performance of SPANet is evaluated using the test sub-sample, focusing on *fully-reconstructable* events. These are defined as events where all resonance particles can be fully reconstructed because their decay products are identified. This distinction is crucial, as certain events may contain four jets, but not all originate from the relevant particles; some jets may result from extra processes, such as parton showers. By selecting fully reconstructable events, it is possible to directly compare SPANet's predictions with the complete particle-level information. This allows an assessment of the network's ability to correctly assign jets to their respective parent particles. The efficiency (ε) of the jet assignment task is defined as the ratio of the number of events with correct assignments to the total number of fully reconstructable events.

Jet assignment efficiencies are calculated separately for each resonance particle in the $t\bar{t}H$ sample. For the top and antitop quarks, the task involves assigning one jet originating from the b- and $\bar{b}$-quarks, respectively, since the W-bosons decay leptonically. For the Higgs boson, both jets from the $b\bar{b}$ pair must be correctly identified. The reconstruction efficiencies for each particle are presented in Table 3.

Table 3: Reconstruction efficiencies for the Higgs, top and antitop quarks using SPANet.

	Efficiency (%)
Higgs (H)	38.4
Top (t)	65.5
Antitop ($\bar{t}$)	65.2

SPANet achieves notable reconstruction efficiencies, particularly for the Higgs boson. The efficiency for $H \to b\bar{b}$ outperforms previous works using analytic methods for this production and decay channel [86]. To further investigate the network's performance, the Higgs invariant mass and transverse momentum (p_T) are reconstructed and shown in Figure 14. The events are categorized based on the number of jets correctly assigned: *2-correct* (both jets assigned correctly), *1-correct* (only one jet assigned correctly), and *0-correct* (neither jet assigned correctly).

The *1-correct* category constitutes the largest contribution, accounting for 52.8% of the events, followed by the *2-correct* category with 38.4% and the *0-correct* category with 8.8%. These results highlight the potential of SPANet in achieving reliable Higgs reconstruction.

A key characteristic of SPANet is its neutrino regression capability. The network predicts the momentum components of the neutrinos using the input variables provided during training (see Sect. 7.1.1). These predictions are compared against the true particle-level neutrino kinematics to evaluate the network's performance. Figure 15 shows the distributions for the neutrino's three-momentum components (p_x, p_y, and p_z) and energy (E). The term *ideal regression* refers to the scenario where the network's predictions exactly match the particle-level neutrino values. Similar results for the antineutrino are included in Appendix D, displaying identical trends in their distributions.

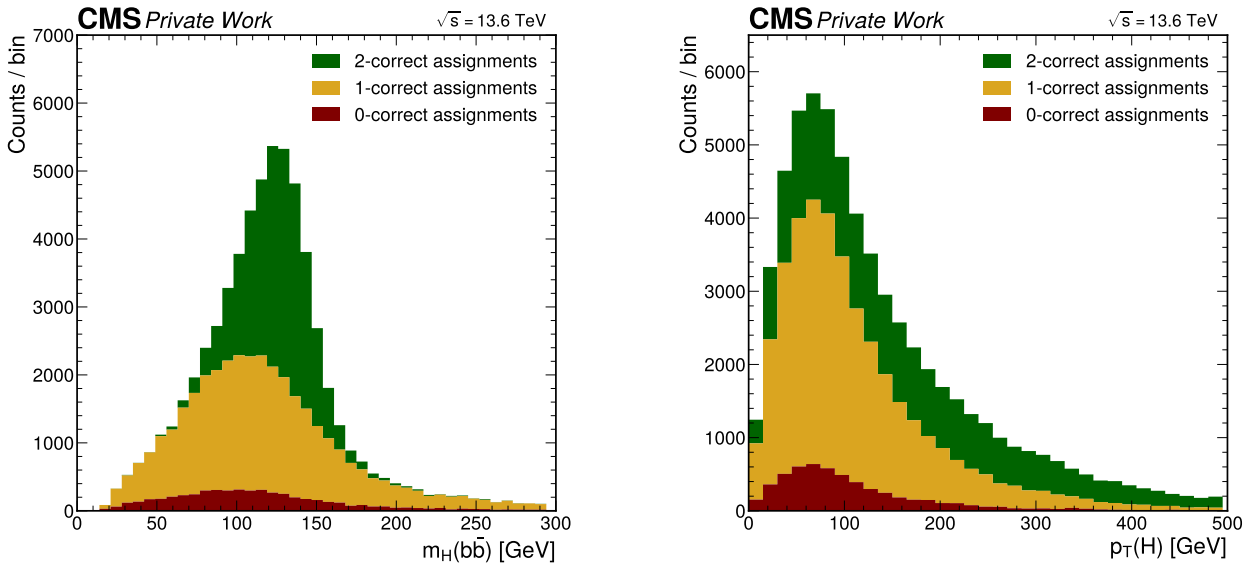

Figure 14: Invariant mass and transverse momentum of the Higgs boson reconstructed from two jets using SPANet. The distributions are categorized by events where both jets are correctly assigned (green), only one jet is correctly assigned (orange), and no jets are correctly assigned (red).

From Figure 15, it can be observed that the transverse momentum components (p_x and p_y) exhibit narrower distributions centred around zero, with an excess of events with transversal momentum closer to this value. It may be the caused for SPANet over-fitting events to the maximum of distributions as also happens with the energy near the peak around 50 GeV. Moreover, the regression for the longitudinal momentum p_z would be the direct result of SPANet predicting more events with low p_T and therefore higher p_z. However, the excess of events for p_z is not accumulated in the tails, instead, they are in a intermediate place between the peak and said tails. This is in agreement with the network predicting, events with less energy as observed in the bottom right of Figure 15. In general, SPANet seems to be biased towards "more frequent" events, which gives robustness in the presence of outliers, but can aslo negatively affect to the performance of the regression. The regression for the anti-neutrino produces the same results and can be found in Appx. D.

Additionally, the two-dimensional distributions of SPANet prediction versus the true particle-level for some neutrino and anti-neutrino variables are provided in Appendix E. The diagonal red line corresponds to perfect agreement between predictions and true values, which can be used as a reference for good performance. For most of the variables the agreement is good at low values, however, the prediction is notably worse for high values. What is expected, given that the network cannot learn many features (at high values) with a small number of events, compare to the low values where most of the events accumulate.

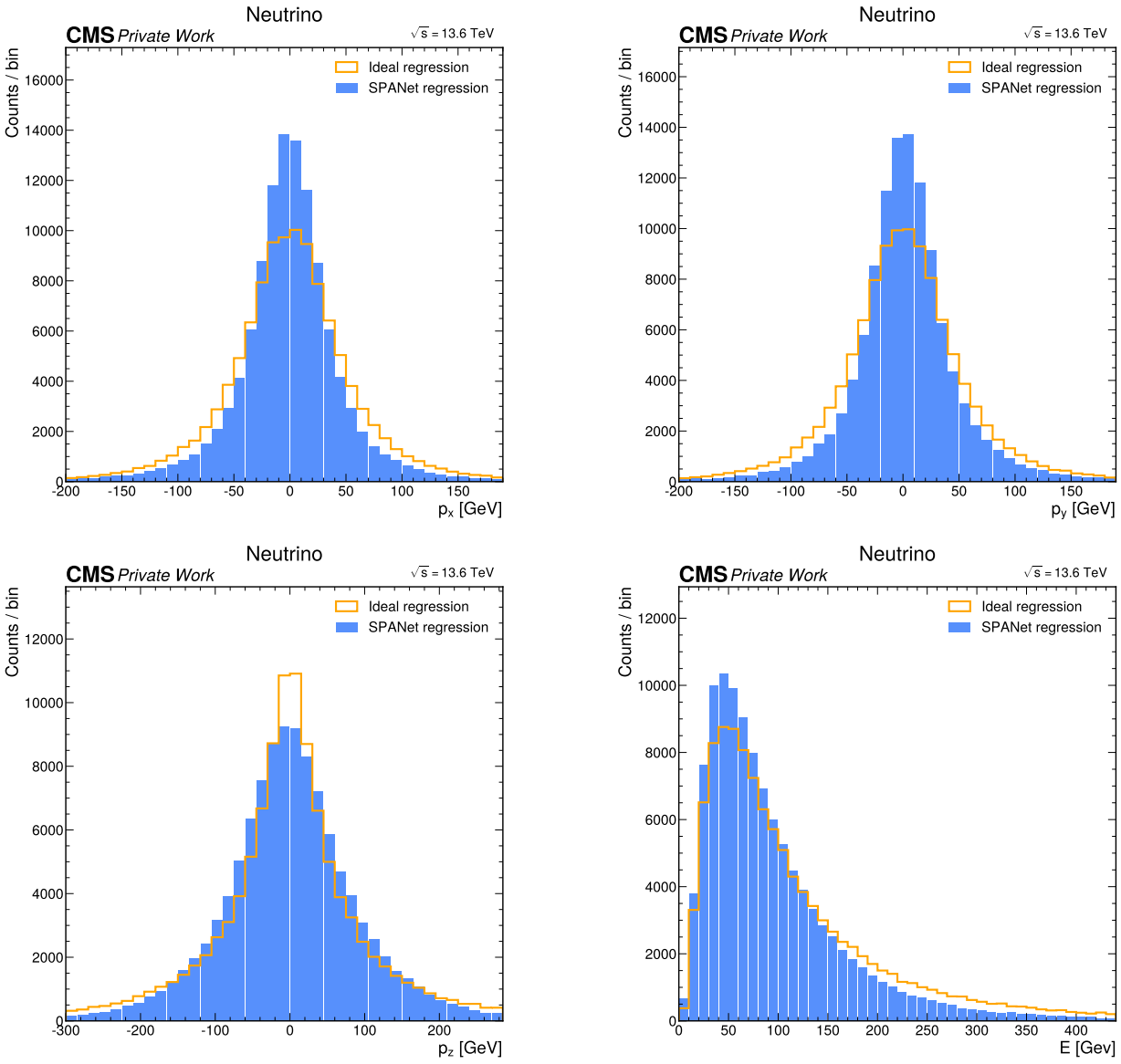

Figure 15: Regression of the neutrino momentum components (p_x, p_y and p_z) and energy (E) with SPANet (blue histogram). Ideal regression is the distribution assuming a 100% efficiency of the network (built with the particle-level neutrino information and without considering detector effects).

7.1.3 Particle reconstruction efficiency

Once the jets are assigned and the neutrino momentum are obtained from the regression, the top quarks and Higgs boson can be reconstructed. For the Higgs boson, only jets are required, while the reconstruction of the (anti)top quark requires a jet, a positively(negatively) charged lepton, and the (anti)neutrino.

Unlike in the previous section, the reconstruction here includes not only fully reconstructable events but also events where one or more jets from the resonance particles are missing. This approach better reflects the conditions encountered when analyzing real data. The distributions for the Higgs boson and the top quark are shown in Figures 16 and 17, respectively. An analogous distribution for the antitop quark is provided in Appendix F. These distributions are categorized into three histograms:

1. *Particle-level distribution:* The true kinematic properties of the particles, derived without detector effects.

2. *Ideal reconstruction:* Detector-level objects are used for jets and leptons, assuming 100% jet assignment efficiency and perfect neutrino regression. This includes only fully reconstructable events.

3. *SPANet reconstruction:* Jets and neutrinos are predicted by SPANet, while the lepton kinematics remain independent of the network.

To improve the reconstructed particle properties, jet energy regressions are applied to the jet momenta. This ensures that the invariant mass distributions are centered around their respective pole masses. All histograms are normalized to unity for easier comparison.

Figure 16 shows the reconstructed distributions for the Higgs boson from two jets (coming from the hadronization of of the $b\bar{b}$ pair). The invariant mass distribution (top left) exhibits similar mean values across all histograms, though the root-mean-square (RMS) values are larger for the reconstructed cases. Due to the high resolution of the particle-level distribution, a subplot focusing on the two reconstructions is included for clarity. Here, the SPANet reconstruction displays a leftward tail – with a smooth extended tail as well on the right side–, mainly due to events in the *1-correct* category (as shown in Figure 14). This broader distribution is entirely attributed to jet assignment inefficiencies.

For the transverse momentum (top right), the difference between the particle-level and ideal reconstructions arises from detector resolution effects, though these are less pronounced than for the mass. The SPANet prediction, influenced by the *1-correct* category, also shows a slight shift to lower values. Note that the 2-correct contribution from Figure 14 is equivalent to the ideal reconstruction. The η and ϕ distributions (bottom row) are consistent with the particle-level results within uncertainties. Differences in the η shape for the reconstructions are attributed to detector smearing effects.

Figure 17 illustrates the reconstructed distributions for the top quark. Unlike the Higgs, the reconstruction involves a jet, lepton, and neutrino, making the regression a critical factor in the shape of the distributions. The invariant mass (top left) from SPANet exhibits a smaller mean value compared to the ideal or particle-level cases, likely due to systematic biases in the neutrino regression. The RMS is found to be narrower than the Higgs reconstruction, mainly attributed to the higher jet assignment efficiency due to requiring only one jet.

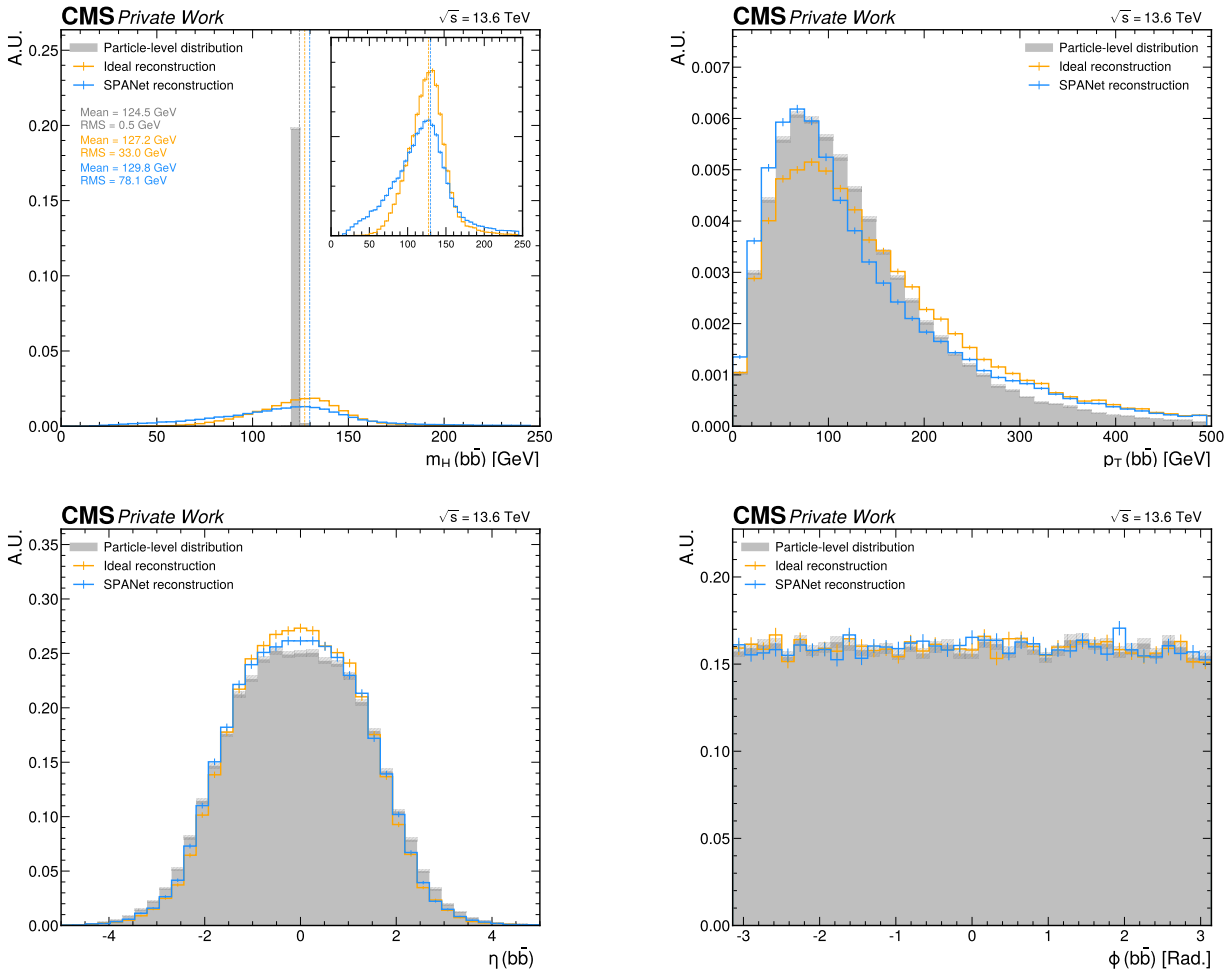

Figure 16: Reconstruction of the Higgs boson using SPANet. Distributions include invariant mass (top left), transverse momentum (top right), pseudorapidity (bottom left), and azimuthal angle (bottom right).

The transverse momentum distribution (top right) for SPANet shows a rightward shift compared to the ideal case. This behavior cannot be explained by jet assignment, as only one jet is used, but is instead linked to the neutrino regression. As observed in Figure 15, SPANet tends to underestimate the transverse momentum components (p_x and p_y). This imbalance causes an overestimation of the reconstructed p_T for the top quark. In other words, when a top quark decay in the top rest frame, the bottom quark and a W boson have the same transversal momentum in opposite directions. Also, the neutrino and lepton momentums are boosted in the direction of the W boson. Hence, if the predicted neutrino p_T is lower, the W boson will also have a lower momentum and consequently the reconstructed top quark will have a not zero momentum in its rest frame.

The η and ϕ distributions (bottom row) align closely with the particle-level results, with a

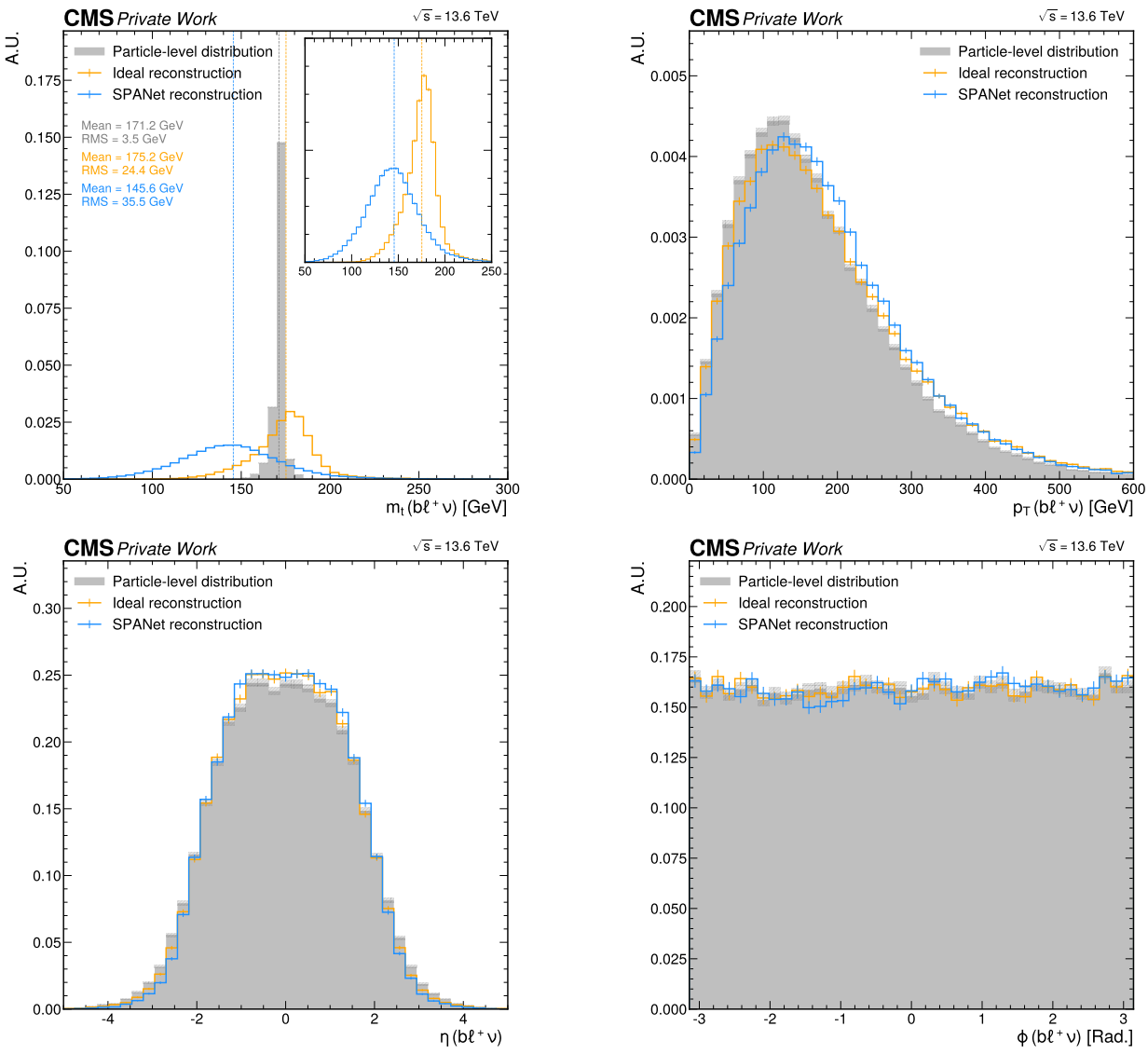

Figure 17: Reconstruction of the top quark using SPANet. Distributions include invariant mass (top left), transverse momentum (top right), pseudorapidity (bottom left), and azimuthal angle (bottom right).

notable plateau in η. This feature, present in both the ideal and SPANet reconstructions, is not only consequence of the neutrino but also from the resolution of the leptons.

7.2 Distributions of observables after particle reconstruction with SPANet

Once the resonance particles are reconstructed, the selected observables can be calculated at detector-level. Following the preselection derived in Sect. 6.3, only variables with the highest separation power are included in this analysis. The shapes of the reconstructed distributions are expected to differ from their particle-level counterparts due to three main factors: (1) the

use of detector-level objects (jets instead of quarks), (2) the performance of SPANet in jet assignment and neutrino regression, and (3) the fiducial phase space defined by the event selection criteria. The reconstructed observables are shown in Figure 18, alongside their particle-level equivalents for comparison (dashed lines).[6]

From Figure 18, several interesting trends can be observed. The angular differences of the leptons ($|\Delta\phi_{\ell\bar{\ell}}|$ and $|\Delta\eta_{\ell\bar{\ell}}|$, top row) remain unaffected by SPANet, as their reconstruction is independent of jet assignment or neutrino regression. The minor deviations from the particle-level shapes are attributed to lepton reconstruction efficiencies and the p_T threshold applied in the detector trigger.

More significant changes are observed in the c_{hel} distribution (middle left). For the $t\bar{t}H$ signal and $t\bar{t}+b\bar{b}$ background, the distribution shows increased slopes in all cases, while for the $t\bar{t}Z$ background, the trend reverses compared to the particle-level shape. This behavior reduces the discriminating power of c_{hel} between $t\bar{t}H$ and $t\bar{t}Z$. Similarly, the $|\cos\theta^*|$ distribution (middle right) exhibits a significant flattening for the $t\bar{t}+b\bar{b}$ sample, losing the characteristic shape associated with vector boson decays. For large $|\cos\theta^*|$ values, a notable reduction in event counts is observed, particularly for the $t\bar{t}H$ signal, where an isotropic decay is expected. This behavior, already present at particle level, appears to be exacerbated by the reconstruction process in the exclusive phase space.

Conversely, the minimum ΔR between any combination of particles (bottom left) remains consistent after reconstruction. While the difference between the signal and background distributions has decreased slightly, the $t\bar{t}H$ signal still dominates at lower ΔR values, maintaining its discriminating power. Lastly, the variable used for comparison with the analysis in **[??]** (bottom right) shows a significant loss of separation power at detector level. The distinctive peak of the $t\bar{t}+b\bar{b}$ background at low p_T is reduced, likely due to the p_T thresholds applied to jets. Additionally, the $t\bar{t}H$ distribution has lost its distinctive peak near the Higgs mass, instead displaying a broader shape, which can be attributed to the limited resolution of jets compared to quarks.

The separation power of these observables at detector-level is quantified in Sect. 8 (Table 4), where the implications of these results are discussed along with potential avenues for improving the analysis.

[6]The particle-level distributions shown here differ from those in Figures 8, 9, 10, and 11 due to the phase space selection. While the previous particle-level distributions correspond to the inclusive phase space, the ones in this section are calculated within the fiducial phase space defined by the event selection criteria.

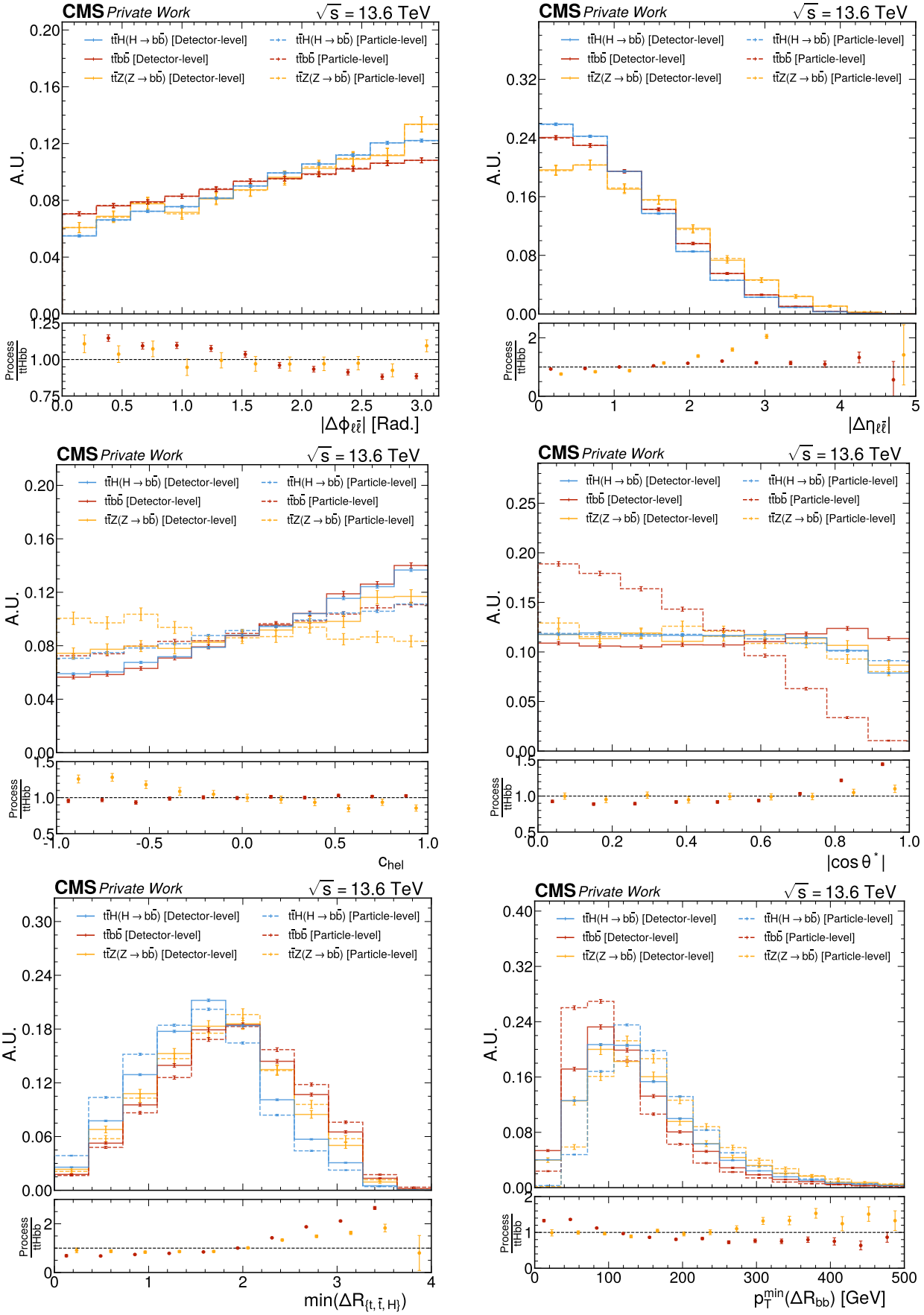

Figure 18: Observables reconstructed at detector level using SPANet (solid lines) and at particle level (dashed lines) in the fiducial phase space defined in Sect. 6. Three processes are included: the $t\bar{t}H(\rightarrow b\bar{b})$ signal (blue line), $t\bar{t}+b\bar{b}$ background (red line) and $t\bar{t}Z(\rightarrow b\bar{b})$ (yellow line). Uncertainties are purely statistical. Ratio plots (below) are calculated for the background and $t\bar{t}Z$ over the signal.

8 Results and discussion

Following the reconstruction of observables in the previous section, the separation power between the signal and the main $t\bar{t} + b\bar{b}$ background was evaluated to identify variables that could provide improvements $t\bar{t}H$-signal analyses. The results are presented in Table 4, where variables are sorted from highest to lowest separation power (the last row corresponds to a variable used in the work of [5]).

Table 4: Summary of separation $S^2 \pm$ (stat.) for the optimized selection.

Variable	$S^2 \pm$ (stat.)
$\min(\Delta R_{\{t,\bar{t},X\}})$	$(2.906 \pm 0.086) \times 10^{-2}$
$\lvert\cos\theta^*\rvert$	$(5.69 \pm 0.40) \times 10^{-3}$
$\lvert\Delta\phi_{\ell\bar{\ell}}\rvert$	$(3.00 \pm 0.30) \times 10^{-3}$
$\lvert\Delta\eta_{\ell\bar{\ell}}\rvert$	$(1.50 \pm 0.21) \times 10^{-3}$
c_{hel}	$(1.82 \pm 0.77) \times 10^{-4}$
$p_T[\min(\Delta R_{bb})]$	$(9.51 \pm 0.63) \times 10^{-3}$

The highest separation power is achieved by $\min(\Delta R_{\{t,\bar{t},X\}})$, with only minor changes in its shape from particle- to detector-level. However, its separation value is reduced by approximately a factor of three. For $|\cos\theta^*|$, the reduction is even more pronounced, decreasing by two orders of magnitude. In contrast, the separation for $|\Delta\phi_{\ell\bar{\ell}}|$ and $|\Delta\eta_{\ell\bar{\ell}}|$ remains largely unaffected. These larger uncertainties stem primary from the smaller event sample used after applying the event selection criteria. Regarding $p_T[\min(\Delta R_{bb})]$, its separation power is significantly diminished at detector level, decreasing by two orders of magnitude. Notably, this variable is independent of the SPANet jet assignment process, as it only involves b-tagged jets, making its degradation purely attributable to detector effects.

The influence of SPANet on the shapes of the reconstructed observables (Figure 18) and the separation values (Table 4) is a key aspect to consider. To disentangle the contributions from SPANet efficiency and detector effects, the same distributions were replotted in Appendix **??**, replacing particle-level shapes with ideal reconstructions (independent of SPANet). These plots reveal that significant differences in shape are primarily due to SPANet efficiency. For example, the $|\cos\theta^*|$ distribution for the $t\bar{t} + b\bar{b}$ background keeps its particle-level shape at detector level. However, the jet assignment by SPANet flattens this distribution, reducing by a large amount its discrimination power.

Regarding the potential bias of the observables toward $t\bar{t}H$-like shapes, the separation power with respect to the $t\bar{t}Z$ process has not been deemed a critical factor for this study and is therefore their values are advisedly not presented.

Despite these limitations, SPANet has proven an effective tool for reconstructing the resonance particles necessary for the analysis. It successfully addresses the challenging task of jet-parton matching and provides the necessary neutrino reconstruction for leptonic analyses. In fact, the

Higgs reconstruction efficiency of 38% surpasses previous studies in this channel, achieving this without compromising the reconstruction of top quarks. However, certain biases introduced by SPANet are evident in Figure 18. For instance, the c_{hel} and $|\cos\theta^*|$ distributions for $t\bar{t}Z$ and $t\bar{t}+b\bar{b}$, respectively, once noticeably different at particle-level, are reconstructed with a shape more alike to the $t\bar{t}H$ signal. In the case ofc_{hel}, it changes the slope from negative to positive, whereas for $|\cos\theta^*|$ is predicted as a flat distribution. Hence, although using the signal sample $t\bar{t}H$ for the training of SPANet to enhance the reconstruction of the Higgs, it has this counterpart of introducing a leaning to the signal shape. These changes reduce the ability of these variables to effectively separate signal from background, a direct consequence of SPANet signal-focused training.

While this study demonstrates promising results in identifying new variables, some considerations should be noted. First, no requirement was imposed on the number of b-tagged jets. This would be relevant in the case of extending the analysis on real data where constraining the number of b-jets could enhance the separation between signal and backgrounds by rejecting events that lack relevant information. This was not feasible here due to statistical limitations, although SPANet training incorporated b-tagging scores to provide information relative to the b-jets. Additionally, training SPANet on both $t\bar{t}H$ and $t\bar{t}+b\bar{b}$ samples simultaneously could mitigate biases and improve reconstruction consistency across datasets. Such an approach could restore the discrimination potential of variables like $|\cos\theta^*|$ for the main background.

Within this work, it has been explored the application of SPANet for the reconstruction of $t\bar{t}H$ events, focusing on the identification and evaluation of spin-sensitive observables. The results demonstrate the utility of SPANet in addressing key challenges in jet-parton assignment and neutrino reconstruction, achieving competitive performance compared to traditional approaches. While the efficiency in reconstructing the Higgs boson is noteworthy, certain biases in the network's predictions highlight areas for future improvement. Specifically, refining training strategies to incorporate background samples and exploring alternative methods to reduce bias will enhance the robustness of the reconstructed observables. These findings contribute to advancing the precision analysis of $t\bar{t}H$ processes and highlight the potential for SPANet to play a pivotal role in future high-energy physics experiments, particularly in probing the Higgs mechanism and its couplings with unprecedented accuracy.

References

[1] J. Ellis. "Higgs Physics". In: *Proceedings of the 2013 European School of High-Energy Physics*. Ed. by M. Mulders and G. Perez. CERN-2015-004. King's College London and CERN. Geneva: CERN, 2015.

[2] T.M. Liss, F. Maltoni, and A. Quadt. *The Top Quark*. Updated September 2013. Review, University of Illinois, University of Catholique de Louvain, University of Göttingen. 2013. URL: https://pdg.lbl.gov/2013/reviews/rpp2013-rev-top-quark.pdf.

[3] M. Carena et al. *Status of Higgs Boson Physics*. Revised August 2023. Review, FNAL, University of Chicago, Kavli Institute, DESY Hamburg, Humboldt University, MPI Munich, UC San Diego. 2023. URL: https://pdg.lbl.gov/2023/reviews/rpp2023-rev-higgs-boson.pdf.

[4] R.L. Workman and et al. (Particle Data Group). "Gauge and Higgs Bosons". In: *Progress of Theoretical and Experimental Physics* 2022 (2022), p. 083C01.

[5] CMS Collaboration. "Measurement of the $t\bar{t}H$ and tH production rates in the $H \to b\bar{b}$ decay channel using proton-proton collision data at $\sqrt{s} = 13$ TeV". In: *Submitted to JHEP* (2024). DOI: 10.48550/arXiv.2407.10896. arXiv: 2407.10896 [hep-ex].

[6] Mohammad Mobassir Ameen. *Measurement of the ttH production cross-section in multi-leptonic final states in pp collisions at a centre-of-mass energy of 13 TeV with the CMS detector*. Tech. rep. Geneva: CERN, 2024. URL: https://cds.cern.ch/record/2917567.

[7] Niccolo Moretti et al. "Measuring the signal strength in $t\bar{t}H$ with $H \to b\bar{b}$". In: *Physical Review D* 93.1 (Jan. 2016). ISSN: 2470-0029. DOI: 10.1103/physrevd.93.014019. URL: http://dx.doi.org/10.1103/PhysRevD.93.014019.

[8] C. N. Yang and R. L. Mills. "Conservation of Isotopic Spin and Isotopic Gauge Invariance". In: *Physical Review* 96.1 (1954), p. 191. DOI: 10.1103/PhysRev.96.191.

[9] D. J. Gross and F. Wilczek. "Ultraviolet Behavior of Non-Abelian Gauge Theories". In: *Physical Review Letters* 30.26 (1973), p. 1343. DOI: 10.1103/PhysRevLett.30.1343.

[10] H. D. Politzer. "Reliable Perturbative Results for Strong Interactions". In: *Physical Review Letters* 30.26 (1973), p. 1346. DOI: 10.1103/PhysRevLett.30.1346.

[11] G. 't Hooft and M. Veltman. "Regularization and Renormalization of Gauge Fields". In: *Nuclear Physics B* 44.1 (1972), pp. 189–213. DOI: 10.1016/0550-3213(72)90279-9.

[12] G. Arnison and others [UA1 Collaboration]. "Experimental Observation of Isolated Large Transverse Energy Electrons with Associated Missing Energy at $\sqrt{s} = 540$ GeV and Their Interpretation as Evidence for W Boson Production". In: *Physics Letters B* 122.1-2 (1983), pp. 103–116. DOI: 10.1016/0370-2693(83)91177-2.

[13] G. Arnison and others [UA1 Collaboration]. "Experimental Observation of Lepton Pairs of Invariant Mass around 95 GeV/c^2 at the CERN $\bar{p}p$ Collider". In: *Physics Letters B* 126.5-6 (1983), pp. 398–410. DOI: 10.1016/0370-2693(83)90188-0.

[14] F. Englert and R. Brout. "Broken Symmetry and the Mass of Gauge Vector Mesons". In: *Physical Review Letters* 13.9 (1964), pp. 321–323. DOI: 10.1103/PhysRevLett.13.321.

[15] P. W. Higgs. "Broken Symmetries and the Masses of Gauge Bosons". In: *Physical Review Letters* 13.16 (1964), pp. 508–509. DOI: 10.1103/PhysRevLett.13.508.

[16] Michael E. Peskin and Daniel V. Schroeder. *An Introduction to Quantum Field Theory*. Reading, USA: Addison-Wesley (1995) 842 p. Westview Press, 1995.

[17] Francis Halzen and Alan D. Martin. *Quarks and Leptons: An Introductory Course in Modern Particle Physics*. Wiley, 1984.

[18] CERN. *The Standard Model*. Accessed: 2024-10-11. 2024. URL: https://home.cern/science/physics/standard-model.

[19] A. Alavi-Harati et al. "Observation of Direct CP Violation in $K \to \pi\pi$ Decays". In: *Physical Review Letters* 83.22 (1998), p. 922.

[20] J. J. Aubert et al. "Experimental Observation of a Heavy Particle J/ψ". In: *Physical Review Letters* 33.23 (1974), p. 1404. DOI: 10.1103/PhysRevLett.33.1404.

[21] S. W. Herb et al. "Observation of a Dimuon Resonance at 9.5 GeV in 400 GeV Proton-Nucleus Collisions". In: *Physical Review Letters* 39.5 (1977), p. 252. DOI: 10.1103/PhysRevLett.39.252.

[22] F. Abe and others [CDF Collaboration]. "Observation of Top Quark Production in $p\bar{p}$ Collisions". In: *Physical Review Letters* 74.14 (1995), p. 2626. DOI: 10.1103/PhysRevLett.74.2626.

[23] M. L. Perl and others [SLAC-LBL Collaboration]. "Evidence for Anomalous Lepton Production in e^+e^- Annihilation". In: *Physical Review Letters* 35.22 (1975), p. 1489. DOI: 10.1103/PhysRevLett.35.1489.

[24] C. L. Cowan and F. Reines. "Detection of the Free Neutrino: A Confirmation". In: *Science* 124.3212 (1956), pp. 103–104. DOI: 10.1126/science.124.3212.103.

[25] G. Danby et al. "Observation of High-Energy Neutrino Reactions and the Existence of Two Kinds of Neutrinos". In: *Physical Review Letters* 9.1 (1962), p. 36. DOI: 10.1103/PhysRevLett.9.36.

[26] ATLAS Collaboration and CMS Collaboration. "Observation of a new particle at a mass of 125 GeV with the ATLAS and CMS experiments at the LHC". In: *Physics Letters B* 716.1 (2012), pp. 1–29. DOI: 10.1016/j.physletb.2012.08.020.

[27] J. A. Aguilar-Saavedra. "Toponium Hunter's Guide". In: *arXiv* 2407.20330v2 (2024). IFT-UAM/CSIC-24-112, CERN-TH-2024-123. URL: https://arxiv.org/abs/2407.20330.

[28] Stefano Catani et al. "Higgs boson production in association with a top-antitop quark pair in next-to-next-to leading order QCD". Version 2. In: (2022). Submitted on 14 Oct 2022, last revised 8 Mar 2023. DOI: https://doi.org/10.48550/arXiv.2210.07846. URL: https://doi.org/10.48550/arXiv.2210.07846.

[29] D. de Florian et al. *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*. arXiv:1610.07922 [hep-ph]. 2017. DOI: https://doi.org/10.48550/arXiv.1610.07922.

[30] Roberto Santos et al. "Machine learning techniques in searches for $t\bar{t}h$ in the $h \to b\bar{b}$ decay channel". In: *arXiv* 1610.03088 (2016). Submitted on 10 Oct 2016, last revised 30 Nov 2016. DOI: https://doi.org/10.48550/arXiv.1610.03088.

[31] L. Evans and P. Bryant. "LHC Machine". In: *JINST* 3 (2008), S08001. DOI: 10.1088/1748-0221/3/08/S08001.

[32] *The Large Hadron Collider: A marvel of modern engineering*. https://home.cern/science/accelerators/large-hadron-collider. 2023.

[33] *Pulling together: Superconducting electromagnets*. https://home.cern/science/engineering/pulling-together-superconducting-electromagnets. 2023.

[34] *Cryogenics: Low temperatures, high performance*. https://home.cern/science/engineering/cryogenics-low-temperatures-high-performance. 2023.

[35] *The CERN accelerator complex*. https://home.cern/science/accelerators/accelerator-complex. 2023.

[36] S. et al. (CMS Collaboration) Chatrchyan. "The CMS experiment at the CERN LHC". In: *JINST* 3 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004.

[37] *A vacuum as empty as interstellar space*. https://home.cern/science/engineering/vacuum-empty-interstellar-space. 2023.

[38] *The Worldwide LHC Computing Grid (WLCG)*. https://home.cern/science/computing/grid. 2023.

[39] *High-Luminosity LHC*. https://home.cern/science/accelerators/high-luminosity-lhc. 2023.

[40] ATLAS Collaboration. "The ATLAS Experiment at the CERN Large Hadron Collider: A Description of the Detector Configuration for Run 3". In: *JINST* 19 (2024), P05063. DOI: 10.1088/1748-0221/19/05/P05063. arXiv: 2305.16623 [physics.ins-det].

[41] CMS Collaboration. "Development of the CMS detector for the CERN LHC Run 3". In: *Journal of Instrumentation (JINST)* 19 (2024), P05064. DOI: 10.1088/1748-0221/19/05/P05064. arXiv: 2309.05466 [physics.ins-det].

[42] CMS Collaboration. *Overview of CMS Physics Goals and Detector*. https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookCMSExperiment. Accessed: 03.11.2024.

[43] Izaak Neutelings. *CMS coordinate system*. Accessed: 2024-11-06. 2024. URL: https://tikz.net/axis3d_cms/.

[44] Lian-Tao Wang and Itay Yavin. "A Review of Spin Determination at the LHC". In: *International Journal of Modern Physics A* 23 (2008), pp. 4647–4668. arXiv: 0802.2726 [hep-ph].

[45] M.G. Echevarria et al. "Spin Physics with a Fixed-Target Experiment at the LHC". In: *Proceedings of the 23rd International Spin Physics Symposium (SPIN 2018)*. Vol. 2019. SISSA, 2019, p. 063. DOI: 10.22323/1.346.0063. arXiv: 1903.03379 [hep-ex].

[46] J. Bernabeu and A. Segarra. "The W and Z boson spin observables as messengers of new physics at LHC". In: (2017). arXiv: 1711.04250 [hep-ph].

[47] CMS Collaboration. "Measurement of the top quark polarization and $t\bar{t}$ spin correlations using dilepton final states in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Physical Review D* 100.7 (2019), p. 072002. DOI: 10.1103/PhysRevD.100.072002. arXiv: 1907.03729 [hep-ex].

[48] Werner Bernreuther, Dennis Heisler, and Zong-Guo Si. "A set of top quark spin correlation and polarization observables for the LHC: Standard Model predictions and new physics contributions". In: *Journal of High Energy Physics* 12 (2015), p. 026. DOI: 10.1007/JHEP12(2015)026. arXiv: 1508.05271 [hep-ph].

[49] ATLAS Collaboration. "Observation of spin correlation in $t\bar{t}$ events from pp collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector". In: *Phys. Rev. Lett.* 108 (2012). DOI: 10.48550/arXiv.1203.4081. arXiv: 1212.4888 [hep-ph].

[50] Lars Ferencz et al. "Study of ttbb and ttW background modelling for ttH analyses". In: (2023). LHC Higgs WG report. arXiv: 2301.11670 [hep-ex].

[51] ATLAS Collaboration. "Observation of quantum entanglement in top-quark pairs using the ATLAS detector". In: *Nature* 633 (2024), p. 542. DOI: 10.48550/arXiv.2311.07288. arXiv: 2311.07288 [hep-ex].

[52] CMS Collaboration. "Observation of quantum entanglement in top quark pair production in proton-proton collisions at $\sqrt{s} = 13$ TeV". In: *Reports on Progress in Physics* 87 (2024), p. 117801. DOI: 10.1088/1361-6633/ad7e4d. arXiv: 2406.03976 [hep-ex].

[53] A. Brandenburg, Z.G. Si, and P. Uwer. "QCD-corrected spin analysing power of jets in decays of polarized top quarks". In: *Physics Letters B* 539 (2002), pp. 235–241. DOI: 10.1016/S0370-2693(02)02098-1. arXiv: hep-ph/0205023 [hep-ph].

[54] Matthew Baumgart and Brock Tweedie. "A New Twist on Top Quark Spin Correlations". In: *Journal of High Energy Physics* 03 (2013), p. 117. DOI: 10.1007/JHEP03(2013)117. arXiv: 1212.4888 [hep-ph].

[55] A. A. Anuar. "Top Quark Spin and Polarization Properties in Searches for New Phenomena with the CMS Detector at the LHC". DESY-THESIS-2020-001, Dissertation, Universität Hamburg, 2019. PhD thesis. Hamburg: Universität Hamburg, 2019. DOI: 10.3204/PUBDB-2020-00203.

[56] Stefano Argiro. *Problems and solutions: the ECAL leak story*. Accessed: 2024-11-25. 2024. URL: https://cms.cern/news/problems-and-solutions-ecal-leak-story.

[57] CMS Collaboration. *Pixel Detector Performance in Run 3*. Tech. rep. CMS-DP-2022-067; CERN-CMS-DP-2022-067. Accessed: 2024-11-19. Geneva, Switzerland: CERN, 2022. URL: https://cds.cern.ch/record/2856417.

[58] Andy Buckley et al. "General-purpose event generators for LHC physics". In: *Physics Reports* 504 (2011), pp. 145–233. DOI: 10.1016/j.physrep.2011.03.005. arXiv: 1101.2599 [hep-ph].

[59] Stefano Frixione, Paolo Nason, and Carlo Oleari. "Matching NLO QCD computations with parton shower simulations: the POWHEG method". In: *Journal of High Energy Physics* 2007.11 (Nov. 2007), 070?070. ISSN: 1029-8479. DOI: 10.1088/1126-6708/2007/11/070. URL: http://dx.doi.org/10.1088/1126-6708/2007/11/070.

[60] J. Alwall et al. "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations". In: *Journal of High Energy Physics* 07 (2014), p. 079. DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301 [hep-ph].

[61] The NNPDF Collaboration: Richard D. Ball et al. "Parton distributions from high-precision collider data". In: *European Physical Journal C* 77 (2017), p. 663. DOI: 10.1140/epjc/s10052-017-5199-5. arXiv: 1706.00428 [hep-ph].

[62] Christian Bierlich et al. "A comprehensive guide to the physics and usage of PYTHIA 8.3". In: *arXiv* 2203.11601 (2022). Accessed: 2024-11-19. URL: https://doi.org/10.48550/arXiv.2203.11601.

[63] Geant4 Collaboration. *Geant4. A simulation toolkit - Physics Reference Manual Release 11.2*. Version Rev8.0. Accessed: 2024-11-25. Dec. 2023. URL: https://geant4.web.cern.ch/.

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

[65] Wolfgang Adam et al. *Track Reconstruction in the CMS tracker*. CMS Note CMS-NOTE-2006-041. CERN, Dec. 2006.

[66] CMS Collaboration. *MVA Based Electron ID for Run 3*. https://twiki.cern.ch/twiki/bin/view/CMS/MultivariateElectronIdentificationRun3. Accessed: 12.11.2024.

[67] A. Hoecker et al. *TMVA - Toolkit for Multivariate Data Analysis*. Preprint available at arXiv:physics/0703039 [physics.data-an]. 2007. URL: https://doi.org/10.48550/arXiv.physics/0703039.

[68] CMS Collaboration. *Performance of electron reconstruction at High Level Trigger using data collected at the CMS experiment at CERN in 2023*. Tech. Report CMSDP-2024/041. CMS Performance Note, July 2024. URL: https://cms-docdb.cern.ch.

[69] CMS Collaboration. *Muon recommendations for 2022 data and Monte Carlo*. https://twiki.cern.ch/twiki/bin/view/CMS/MuonRun32022#ID_efficiencies_AN1. Accessed: 12.11.2024.

[70] CMS Collaboration. *Muon ID and Isolation efficiencies with muons in proton-proton collisions at* $\sqrt{s} = 13.6$ *TeV*. Tech. Report CMSDP-2024/067. Version v2. CMS Performance Note, July 2024. URL: https://cms-docdb.cern.ch.

[71] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The anti-k_t jet clustering algorithm". In: *JHEP* 04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189 [hep-ph].

[72] CMS Collaboration. "Pileup mitigation at CMS in 13 TeV data". In: *Journal of Instrumentation* 15 (2020), P09018. DOI: 10.1088/1748-0221/15/09/P09018. arXiv: 2003.00503 [hep-ex].

[73] Emil Bols et al. "Jet Flavour Classification Using DeepJet". In: *Journal of Instrumentation (JINST)* 15 (2020), P12012. DOI: 10.1088/1748-0221/15/12/P12012. arXiv: 2008.10519 [hep-ex].

[74] Huilin Qu and Loukas Gouskos. "ParticleNet: Jet Tagging via Particle Clouds". In: *Physical Review D* 101 (2020), p. 056019. DOI: 10.1103/PhysRevD.101.056019. arXiv: 1902.08570 [hep-ph].

[75] Huilin Qu, Congqiao Li, and Sitian Qian. "Particle Transformer for Jet Tagging". In: *Proceedings of the 39th International Conference on Machine Learning (ICML)* (2022). DOI: 10.48550/arXiv.2202.03772. arXiv: 2202.03772 [hep-ph].

[76] CMS Collaboration. "Measurement of the differential cross section for top quark pair production in pp collisions at $\sqrt{s} = 8$ TeV". In: *Eur. Phys. J. C* 75 (2015), p. 542. DOI: 10.1140/epjc/s10052-015-3709-x. arXiv: 1505.04480 [hep-ex].

[77] Uttiya Sarkar and CMS Collaboration. *Run 3 performance and advances in heavy-flavor jet tagging in CMS*. Tech. rep. CMS-CR-2024-247. Presented at the 42nd International Conference on High Energy Physics (ICHEP 2024), Prague, Czech Republic, 18–24 July 2024. CMS Note, Oct. 2024, p. 14.

[78] CMS Collaboration. "Identification of b-quark jets with the CMS experiment". In: *JINST* 8 (2013), P04013. DOI: 10.1088/1748-0221/8/04/P04013. arXiv: 1211.4462 [hep-ex].

[79] Martin Erdmann, Benjamin Fischer, and Marcel Rieger. "Jet-Parton Assignment in ttH Events using Deep Learning". In: *JINST* 12 (2017), P08020. DOI: 10.1088/1748-0221/12/08/P08020. arXiv: 1706.01117 [hep-ex].

[80] Jason Sang Hun Lee et al. "Zero-Permutation Jet-Parton Assignment using a Self-Attention Network". In: *J. Korean Phys. Soc.* 84 (2024), pp. 427–438. DOI: 10.1007/s40042-024-01037-3. arXiv: 2012.03542 [hep-ex].

[81] Michael James Fenton et al. "Reconstruction of Unstable Heavy Particles Using Deep Symmetry-Preserving Attention Networks". In: *Communications Physics* 7 (2024), p. 139. DOI: 10.1038/s42005-024-01627-4. arXiv: 2309.01886 [hep-ex].

[82] Scott Stuart Snyder. "Measurement of the Top Quark Mass at D0". FERMILAB-THESIS-1995-27. PhD Thesis. State University of New York, Stony Brook, 1995. DOI: 10.2172/1422822.

[83] Johannes Erdmann et al. "A likelihood-based reconstruction algorithm for top-quark pairs and the KLFitter framework". In: *Nuclear Instruments and Methods in Physics Research Section A* 748 (2014), pp. 18–25. DOI: 10.1016/j.nima.2014.02.029. arXiv: 1312.5595 [hep-ex].

[84] Ashish Vaswani et al. "Attention Is All You Need". In: (2017). DOI: 10.48550/arXiv.1706.03762.

[85] Alexander Shmakov et al. "SPANet: Generalized Permutationless Set Assignment for Particle Physics using Symmetry Preserving Attention". In: *SciPost Physics* 12 (2022), p. 178. DOI: 10.21468/SciPostPhys.12.5.178. arXiv: 2106.03898 [hep-ex].

[86] Charis Koraka. "Study of the Higgs boson production in association with two top quarks and its subsequent decay to a $b\bar{b}$ pair with the CMS detector at the CERN LHC". Published: Nov 29, 2021, Experiments: CERN-LHC-CMS. PhD Thesis. Athens National and Kapodistrian University, 2021. DOI: 10.12681/eadd/50506. URL: https://hdl.handle.net/10442/hedi/50506.

A Events used in SPANet training

Table 5: Number of events from the $t\bar{t}H$ sample for training, validation, and testing of SPANet. The percentage represents the fraction of the total number of events in the sample.

	Events
Training (75%)	468,000
Validation (5%)	25,000
Test (20%)	111,000

B Data paths of simulated samples

Table 6: Data paths of simulated samples for $t\bar{t}H$, $t\bar{t}$, and $t\bar{t}Z$. Each sample corresponds to a specific period and is identified by the name given in the data aggregation system employed by CMS.

Sample	Period	Sample Path
$t\bar{t}H$	2022 preEE	/TTH_Hto2B_M-125_TuneCP5_13p6TeV_powheg-pythia8/Run3Summer22NanoAODv12-130X_mcRun3_2022_realistic_v5-v3/NANOAODSIM
	2022 postEE	/TTH_Hto2B_M-125_TuneCP5_13p6TeV_powheg-pythia8/Run3Summer22EENanoAODv12-130X_mcRun3_2022_realistic_postEE_v6-v3/NANOAODSIM
	2023 preBPix	/TTH_Hto2B_M-125_TuneCP5_13p6TeV_powheg-pythia8/Run3Summer23NanoAODv12-130X_mcRun3_2023_realistic_v14-v3/NANOAODSIM
	2023 postBPix	/TTH_Hto2B_M-125_TuneCP5_13p6TeV_powheg-pythia8/Run3Summer23BPixNanoAODv12-130X_mcRun3 _2023_realistic_postBPix_v2-v3/NANOAODSIM
$t\bar{t}$	2022 preEE	/TTto2L2Nu_TuneCP5_13p6TeV_powheg-pythia8/Run3Summer22NanoAODv12-130X_mcRun3_2022_realistic_v5-v2/NANOAODSIM
	2022 postEE	/TTto2L2Nu_TuneCP5_13p6TeV_powheg-pythia8/Run3Summer22EENanoAODv12-130X_mcRun3_2022_realistic_postEE_v6-v2/NANOAODSIM
	2023 preBPix	/TTto2L2Nu_TuneCP5_13p6TeV_powheg-pythia8/Run3Summer23NanoAODv12-130X_mcRun3_2023_realistic_v14-v2/NANOAODSIM
	2023 postBPix	/TTto2L2Nu_TuneCP5_13p6TeV_powheg-pythia8/Run3Summer23BPixNanoAODv12-130X_mcRun3_2023_realistic_postBPix_v2-v3/NANOAODSIM
$t\bar{t}Z$	2022 preEE	/TTZ-ZtoQQ-1Jets_TuneCP5_13p6TeV_amcatnloFXFX-pythia8/Run3Summer22NanoAODv12-130X_mcRun3_2022_realistic_v5-v2/NANOAODSIM
	2022 postEE	/TTZ-ZtoQQ-1Jets_TuneCP5_13p6TeV_amcatnloFXFX-pythia8/Run3Summer22EENanoAODv12-130X_mcRun3_2022_realistic_postEE_v6-v2/NANOAODSIM
	2023 postBPix	/TTZ-ZtoQQ-1Jets_TuneCP5_13p6TeV_amcatnloFXFX-pythia8/Run3Summer23BPixNanoAODv12-130X_mcRun3_2023_realistic_postBPix_v6-v2/NANOAODSIM

C Separation power for the $t\bar{t}Z$ process

Table 7: Summary of S^2 values with the statistical uncertainty for different variables for the samples $t\bar{t}Z$ and $t\bar{t}H$.

Variable	$S^2 \pm$ (stat.)
$\min(\Delta R_{t,\bar{t},H})$	$(3.940 \pm 0.008) \times 10^{-2}$
$\lvert\cos\theta^*\rvert$	$(1.29 \pm 0.17) \times 10^{-3}$
c_{hel}	$(1.178 \pm 0.005) \times 10^{-2}$
c_{han}	$(3.082 \pm 0.008) \times 10^{-4}$
$\lvert\Delta\eta_{\ell\bar{\ell}}\rvert$	$(1.749 \pm 0.005) \times 10^{-3}$
$\lvert\Delta\phi_{\ell\bar{\ell}}\rvert$	$(3.527 \pm 0.009) \times 10^{-4}$
$\cos\theta_1^k - \cos\theta_2^k$	$(1.616 \pm 0.022) \times 10^{-3}$
$\cos\theta_1^k + \cos\theta_2^k$	$(1.016 \pm 0.018) \times 10^{-3}$
$\cos\theta_1^r - \cos\theta_2^r$	$(0.734 \pm 0.014) \times 10^{-3}$
$\cos\theta_1^r + \cos\theta_2^r$	$(0.466 \pm 0.012) \times 10^{-3}$
$\cos\theta_1^n - \cos\theta_2^n$	$(0.861 \pm 0.016) \times 10^{-3}$
$\cos\theta_1^n + \cos\theta_2^n$	$(1.263 \pm 0.019) \times 10^{-3}$
$\cos\theta_1^k \cos\theta_2^k$	$(6.898 \pm 0.038) \times 10^{-3}$
$\cos\theta_1^r \cos\theta_2^r$	$(1.471 \pm 0.019) \times 10^{-3}$
$\cos\theta_1^n \cos\theta_2^n$	$(4.447 \pm 0.031) \times 10^{-3}$
$\cos\theta_1^k \cos\theta_2^n - \cos\theta_1^n \cos\theta_2^k$	$(0.224 \pm 0.071) \times 10^{-3}$
$\cos\theta_1^k \cos\theta_2^n + \cos\theta_1^n \cos\theta_2^k$	$(0.155 \pm 0.059) \times 10^{-3}$
$\cos\theta_1^n \cos\theta_2^r - \cos\theta_1^r \cos\theta_2^n$	$(0.122 \pm 0.052) \times 10^{-3}$
$\cos\theta_1^n \cos\theta_2^r + \cos\theta_1^r \cos\theta_2^n$	$(0.172 \pm 0.063) \times 10^{-3}$
$\cos\theta_1^r \cos\theta_2^k - \cos\theta_1^k \cos\theta_2^r$	$(0.078 \pm 0.042) \times 10^{-3}$
$\cos\theta_1^r \cos\theta_2^k + \cos\theta_1^k \cos\theta_2^r$	$(0.119 \pm 0.052) \times 10^{-3}$
$p_T[\min(\Delta R_{bb})]$	$(3.216 \pm 0.002) \times 10^{-3}$

D Anti-neutrino regression with SPANet

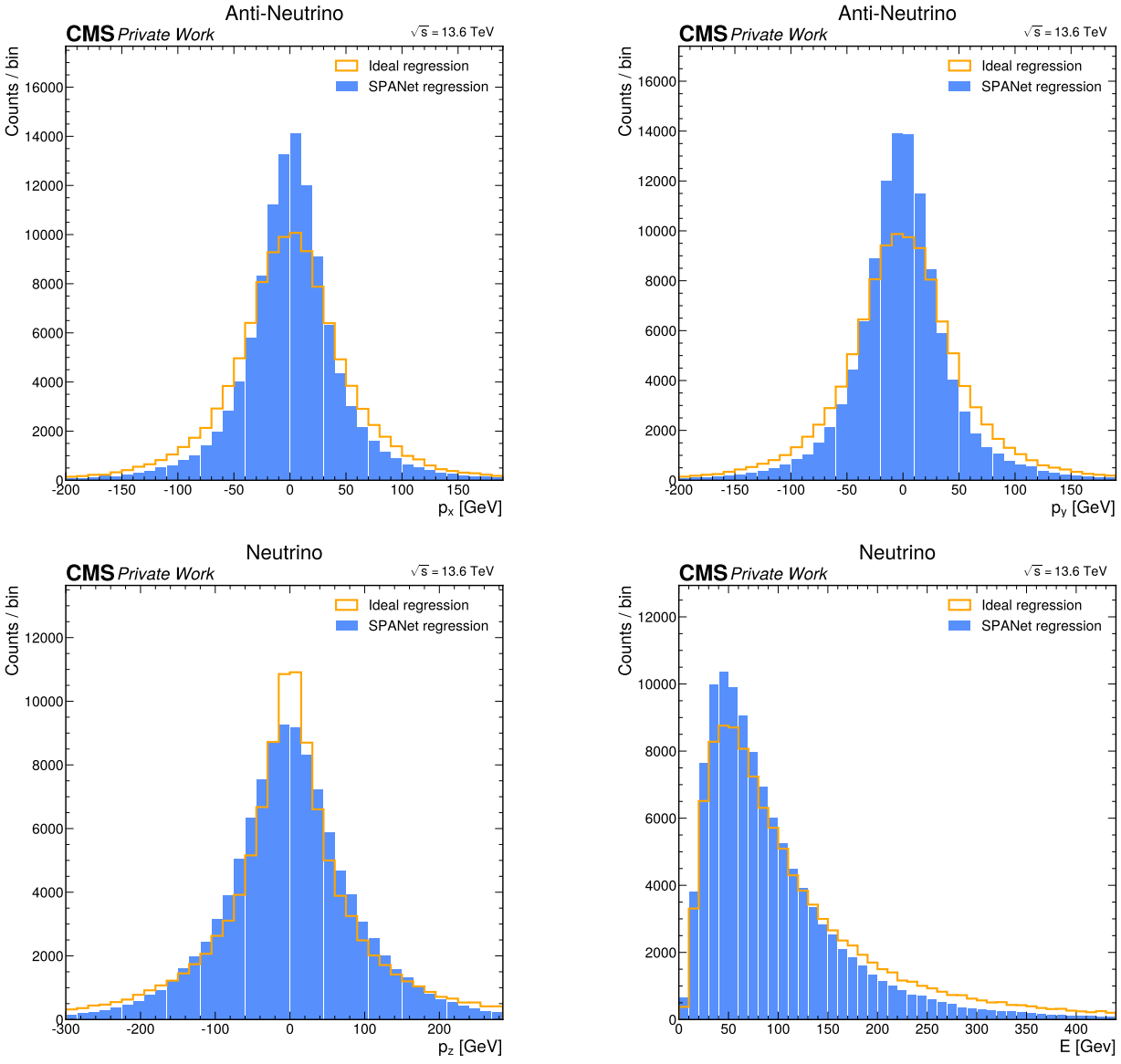

Figure 19: Regression of the anti-neutrino momentum components (p_x, p_y and p_z) and energy (E) with SPANet (blue histogram). Ideal regression is the distribution assuming a 100% efficiency of the network (built with the particle-level neutrino information and without considering detector effects).

E Two-dimensional plots for (anti-)neutrino regression

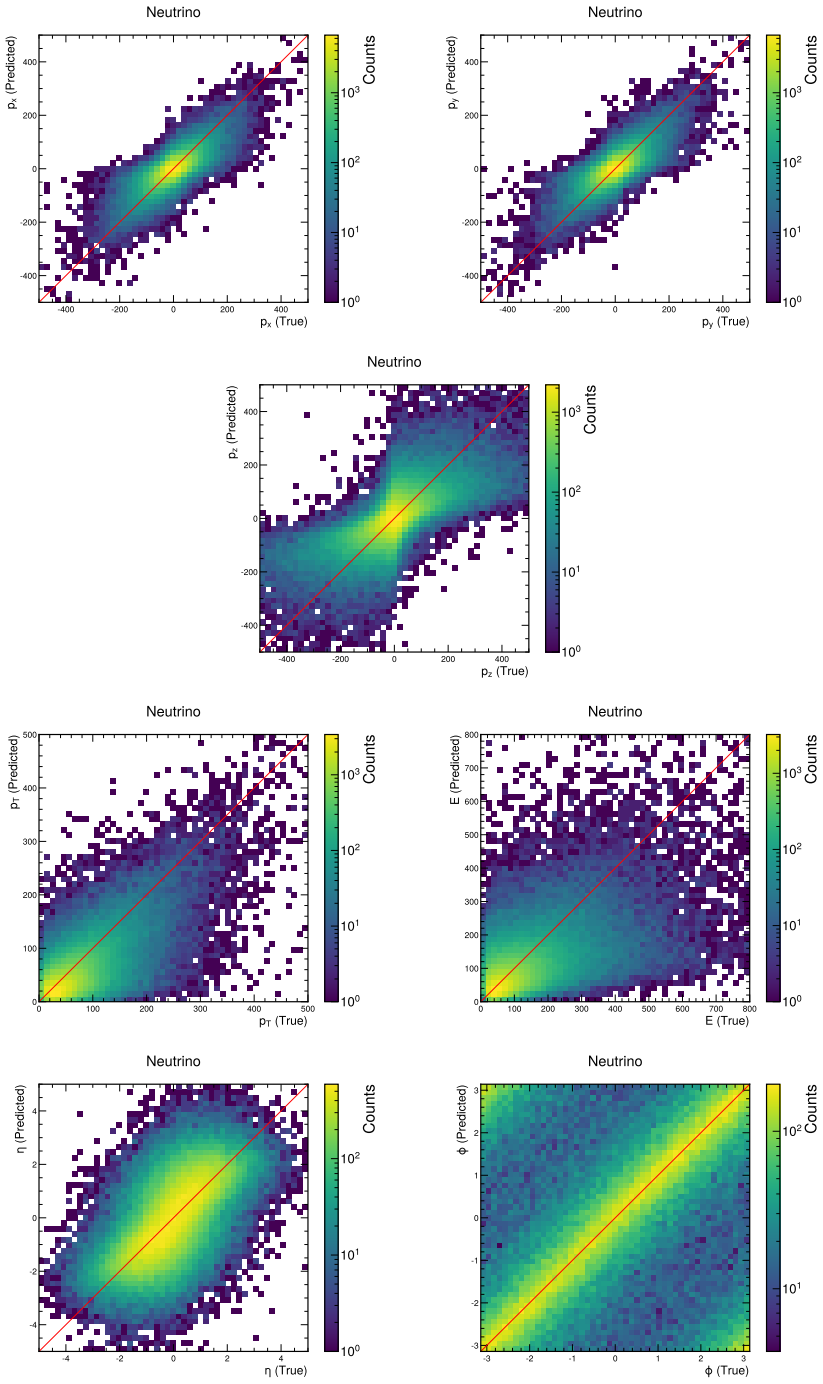

Figure 20: Two-dimensional plots for several kinematic variables of the neutrino. The x-axis correspond with the true value of the neutrino (at particle-level) and the y-axis is the SPANet prediction. The red diagonal line is the case when the prediction matches the true value.

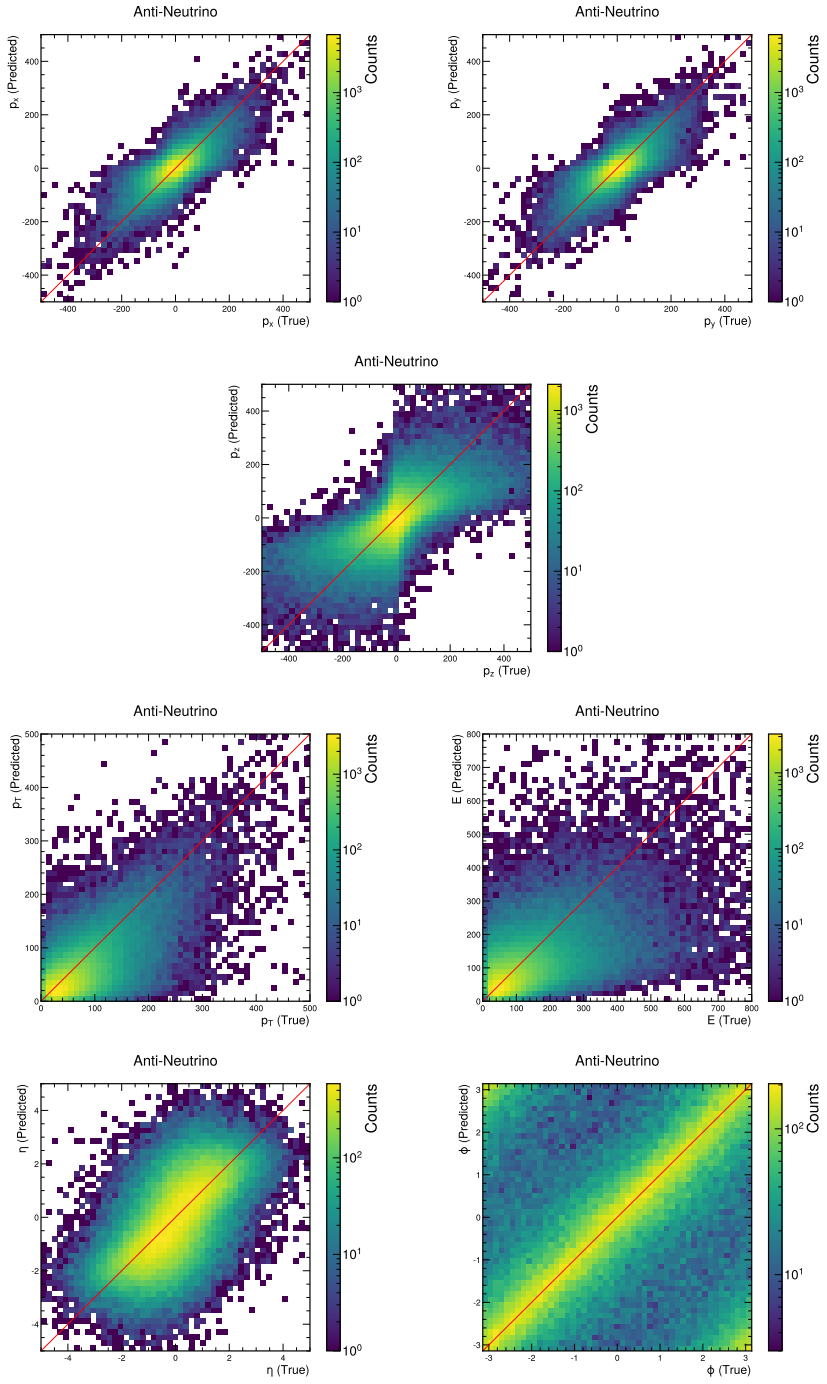

Figure 21: Two-dimensional plots for several kinematic variables of the anti-neutrino. The x-axis correspond with the true value of the neutrino (at particle-level) and the y-axis is the SPANet prediction. The red diagonal line is the case when the prediction matches the true value.

F Top anti-quark reconstruction with SPANet

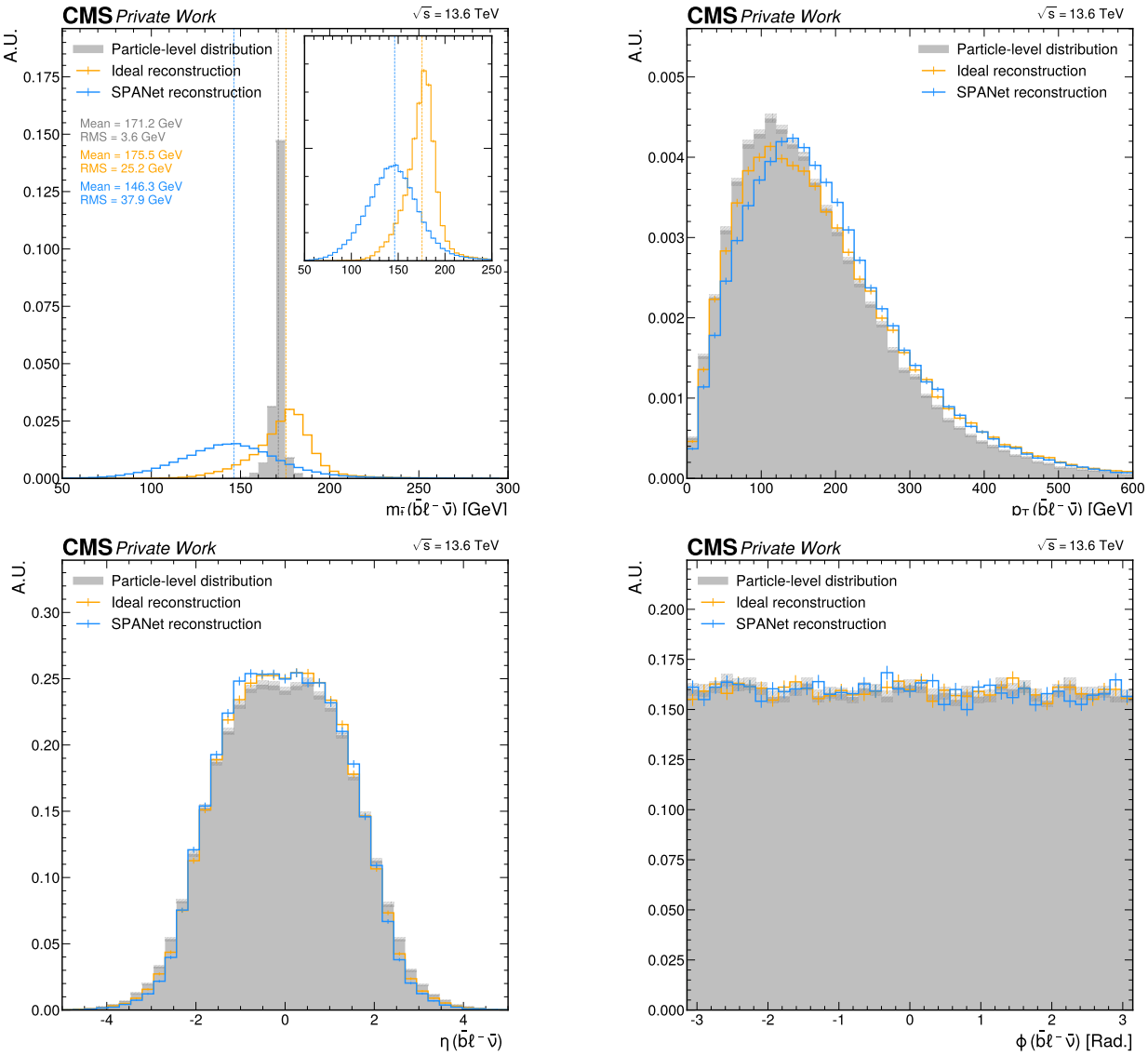

Figure 22: Top anti-quark reconstruction

G Observable reconstruction

G.1 Additional reconstructed observables with SPANet

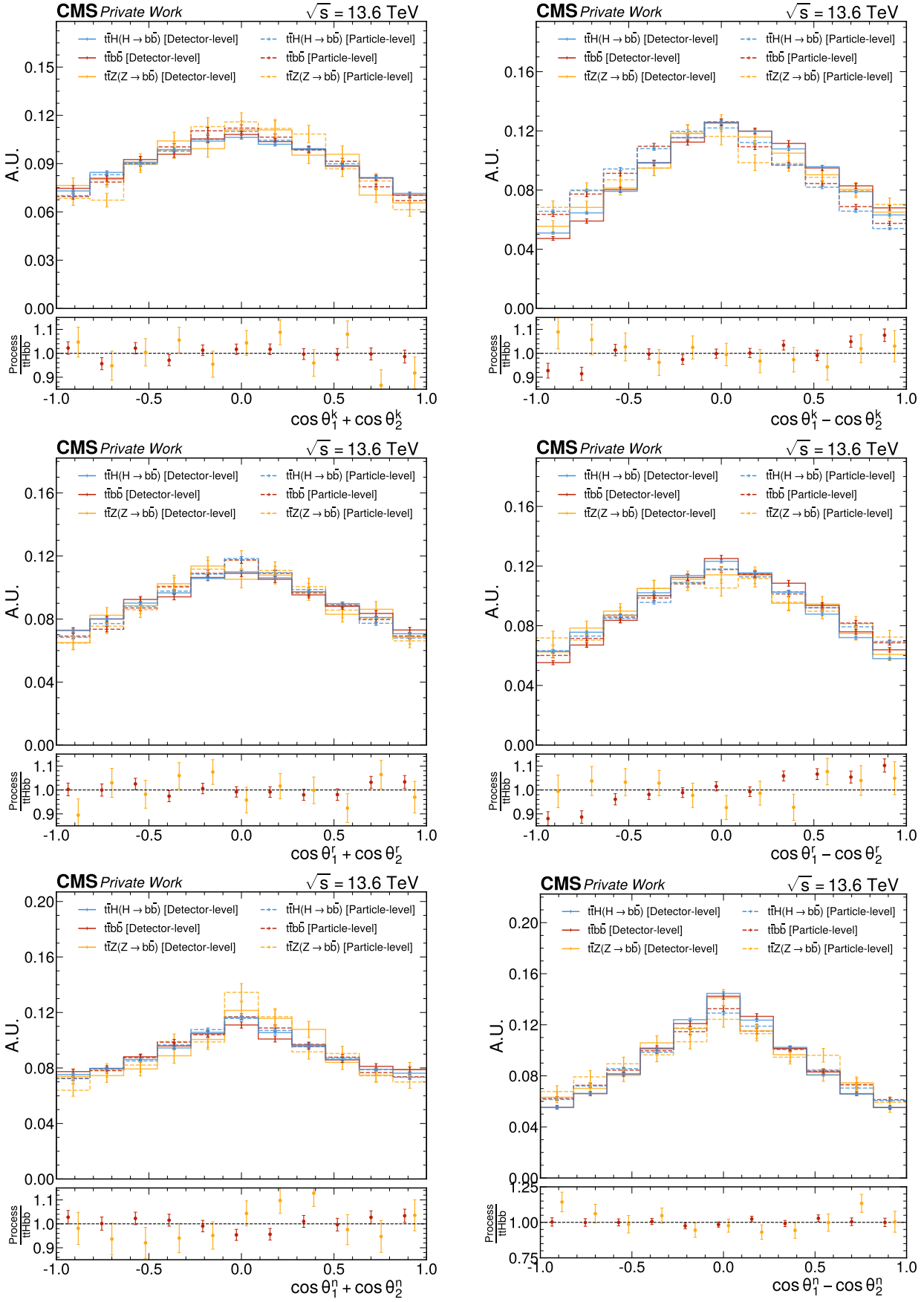

Figure 23: Polarization observables reconstructed at detector level using SPANet (solid lines) and at particle level (dashed lines) in the fiducial phase space defined in Sect. 6. Three processes are included: the $t\bar{t}H(\rightarrow b\bar{b})$ signal (blue line), $t\bar{t}+b\bar{b}$ background (red line) and $t\bar{t}Z(\rightarrow b\bar{b})$ (yellow line). Uncertainties are purely statistical. Ratio plots (below) are calculated for the background and $t\bar{t}Z$ over the signal.

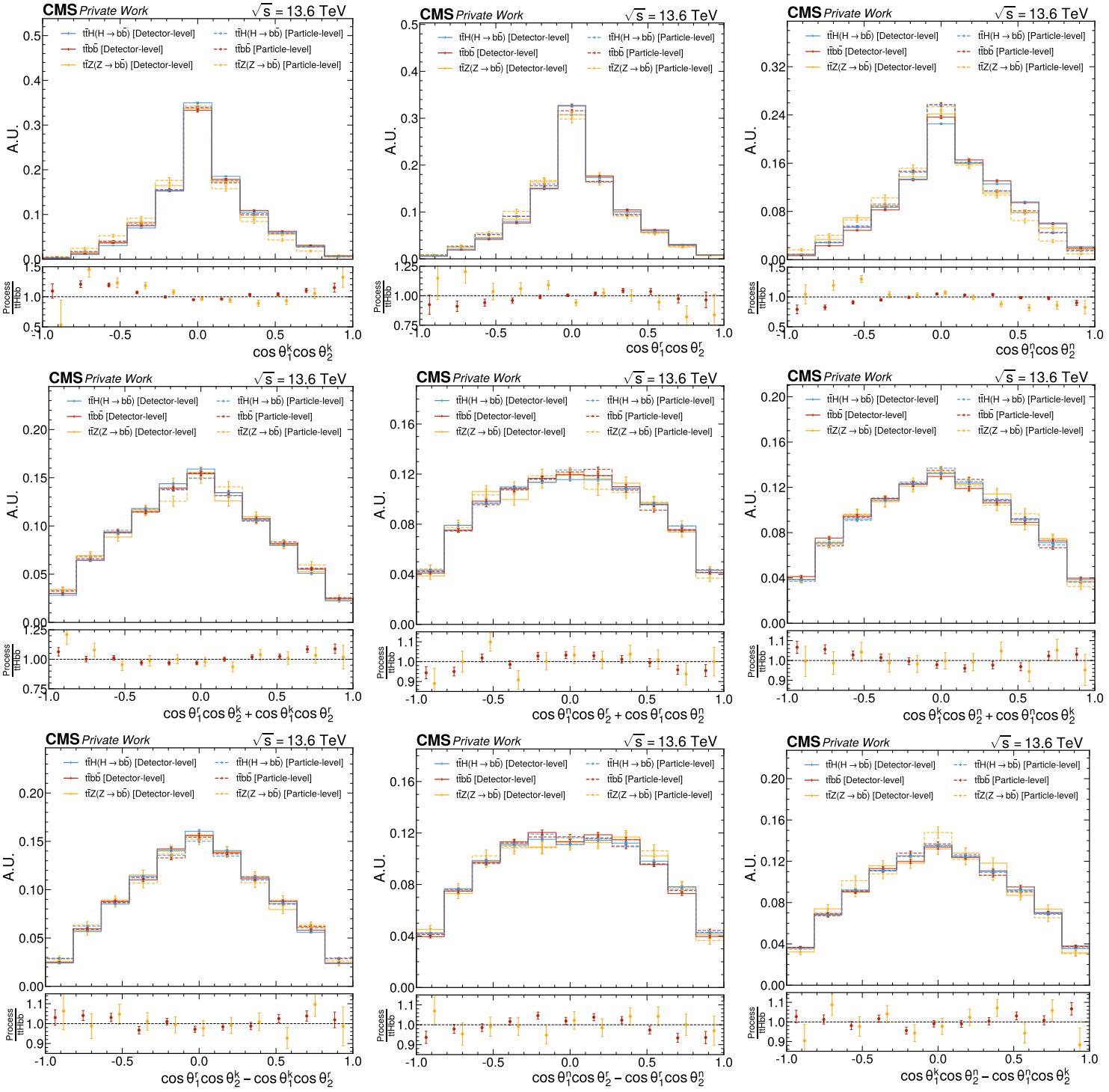

Figure 24: Spin correlation observables observables reconstructed at detector level using SPANet (solid lines) and at particle level (dashed lines) in the fiducial phase space defined in Sect. 6. Three processes are included: the $t\bar{t}H(\rightarrow b\bar{b})$ signal (blue line), $t\bar{t}+b\bar{b}$ background (red line) and $t\bar{t}Z(\rightarrow b\bar{b})$ (yellow line). Uncertainties are purely statistical. Ratio plots (below) are calculated for the background and $t\bar{t}Z$ over the signal.

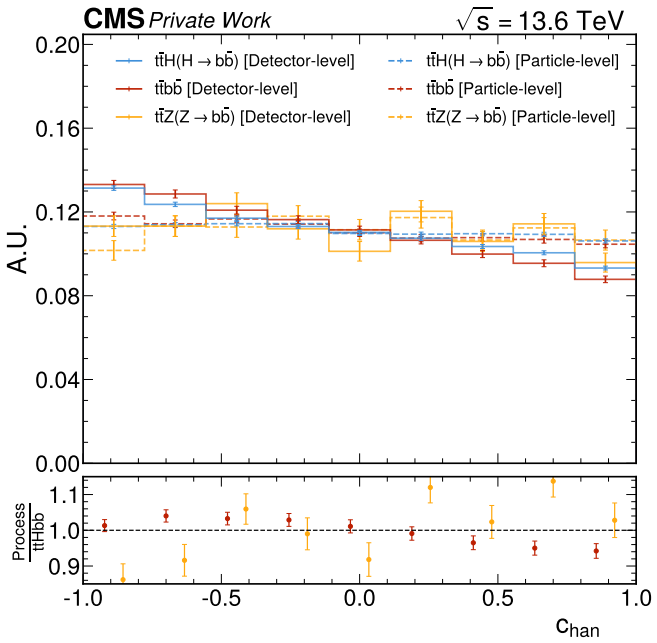

Figure 25: c_{han} spin information variable reconstructed at detector level using SPANet (solid lines) and at particle level (dashed lines) in the fiducial phase space defined in Sect. 6. Three processes are included: the $t\bar{t}H(\rightarrow b\bar{b})$ signal (blue line), $t\bar{t}+b\bar{b}$ background (red line) and $t\bar{t}Z(\rightarrow b\bar{b})$ (yellow line). Uncertainties are purely statistical. Ratio plots (below) are calculated for the background and $t\bar{t}Z$ over the signal.

G.2 Ideal reconstruction of observables

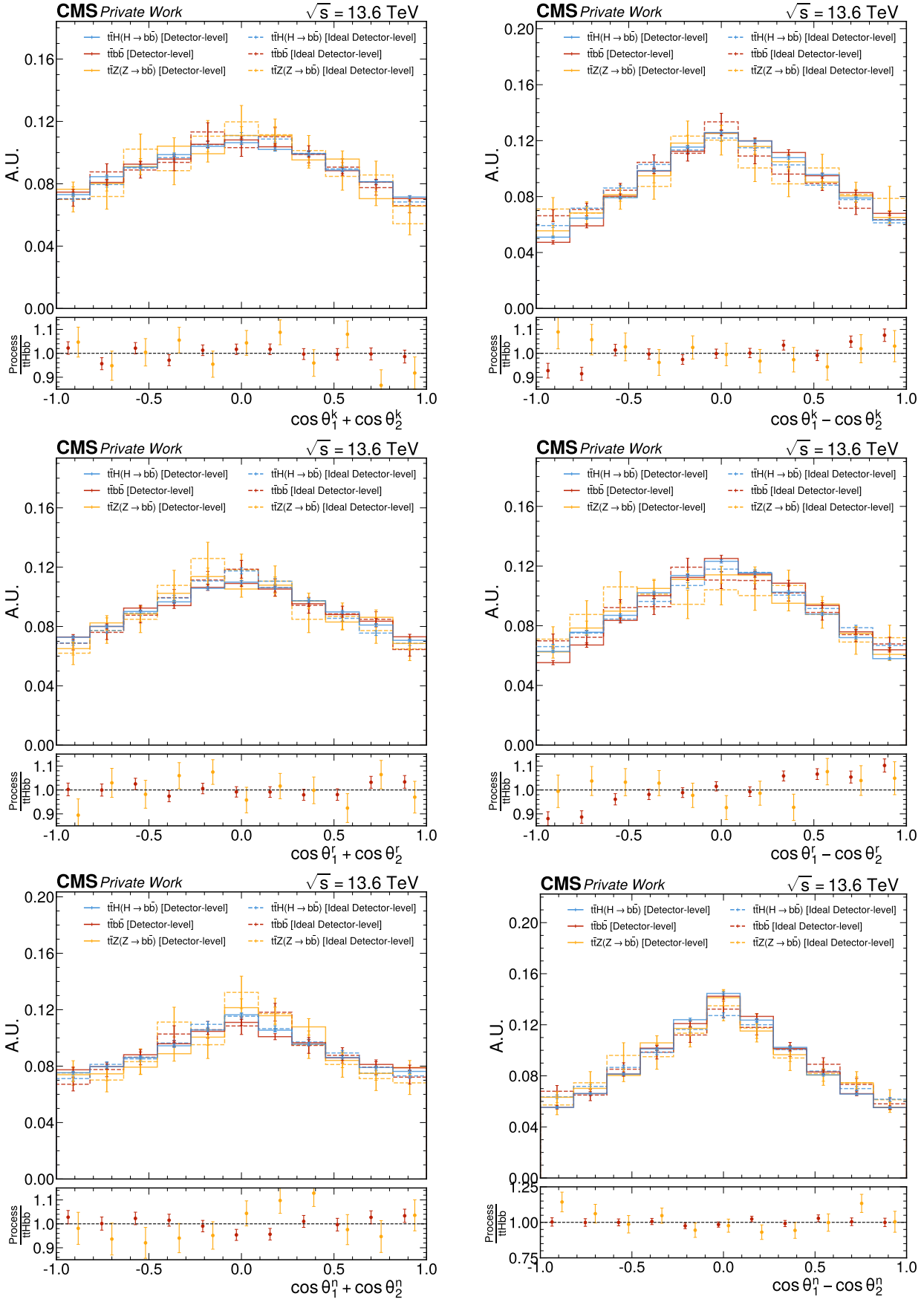

Figure 26: Polarization observables reconstructed at detector level using SPANet (solid lines) and assuming a perfect reconstruction with the network (dashed lines). Three processes are included: the $t\bar{t}H(\rightarrow b\bar{b})$ signal (blue line), $t\bar{t}+b\bar{b}$ background (red line) and $t\bar{t}Z(\rightarrow b\bar{b})$ (yellow line). Uncertainties are purely statistical. Ratio plots (below) are calculated for the background and $t\bar{t}Z$ over the signal.

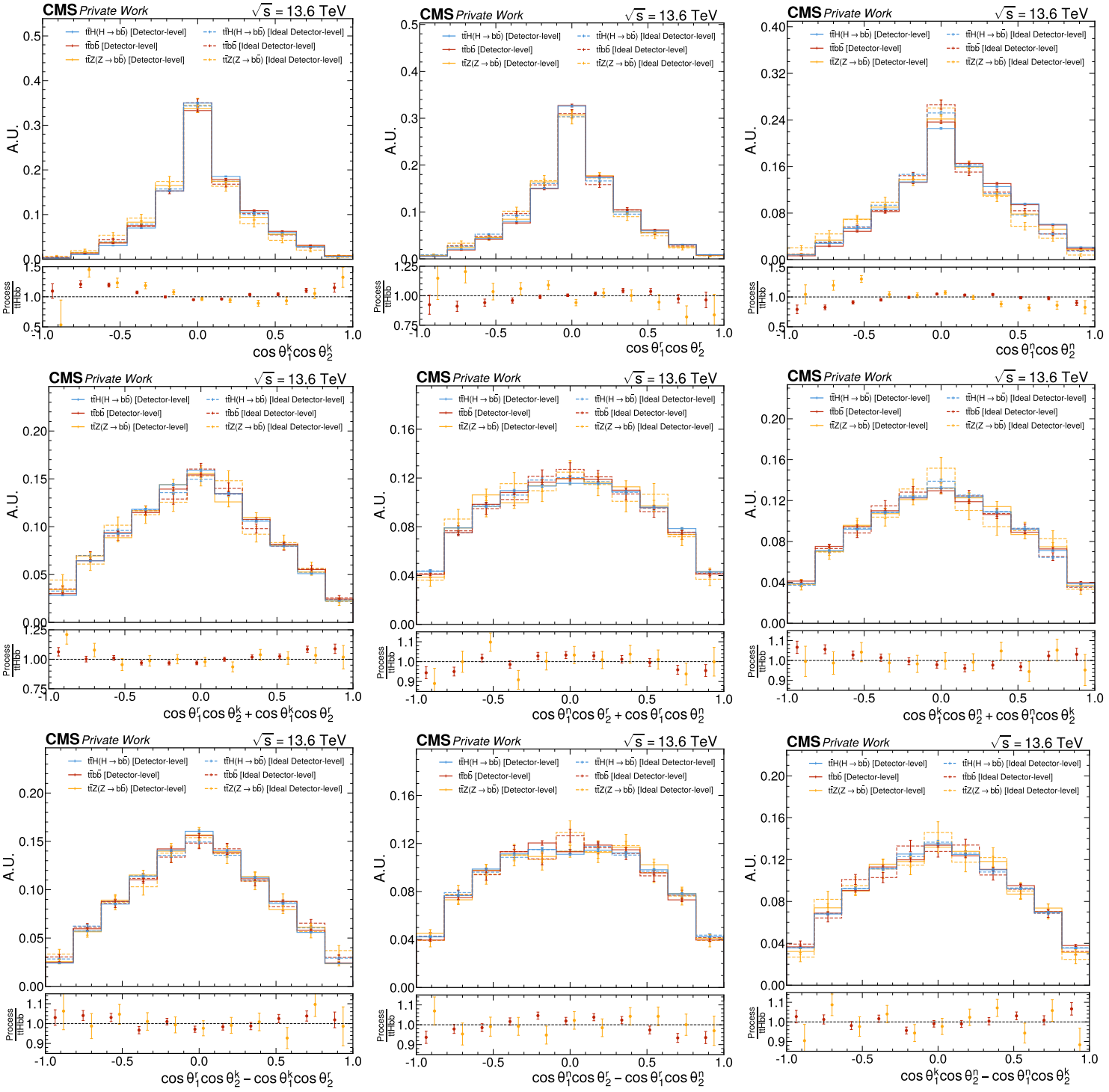

Figure 27: Spin correlation observables observables reconstructed at detector level using SPANet (solid lines) and assuming a perfect reconstruction with the network (dashed lines). Three processes are included: the $t\bar{t}H(\rightarrow b\bar{b})$ signal (blue line), $t\bar{t}+b\bar{b}$ background (red line) and $t\bar{t}Z(\rightarrow b\bar{b})$ (yellow line). Uncertainties are purely statistical. Ratio plots (below) are calculated for the background and $t\bar{t}Z$ over the signal.

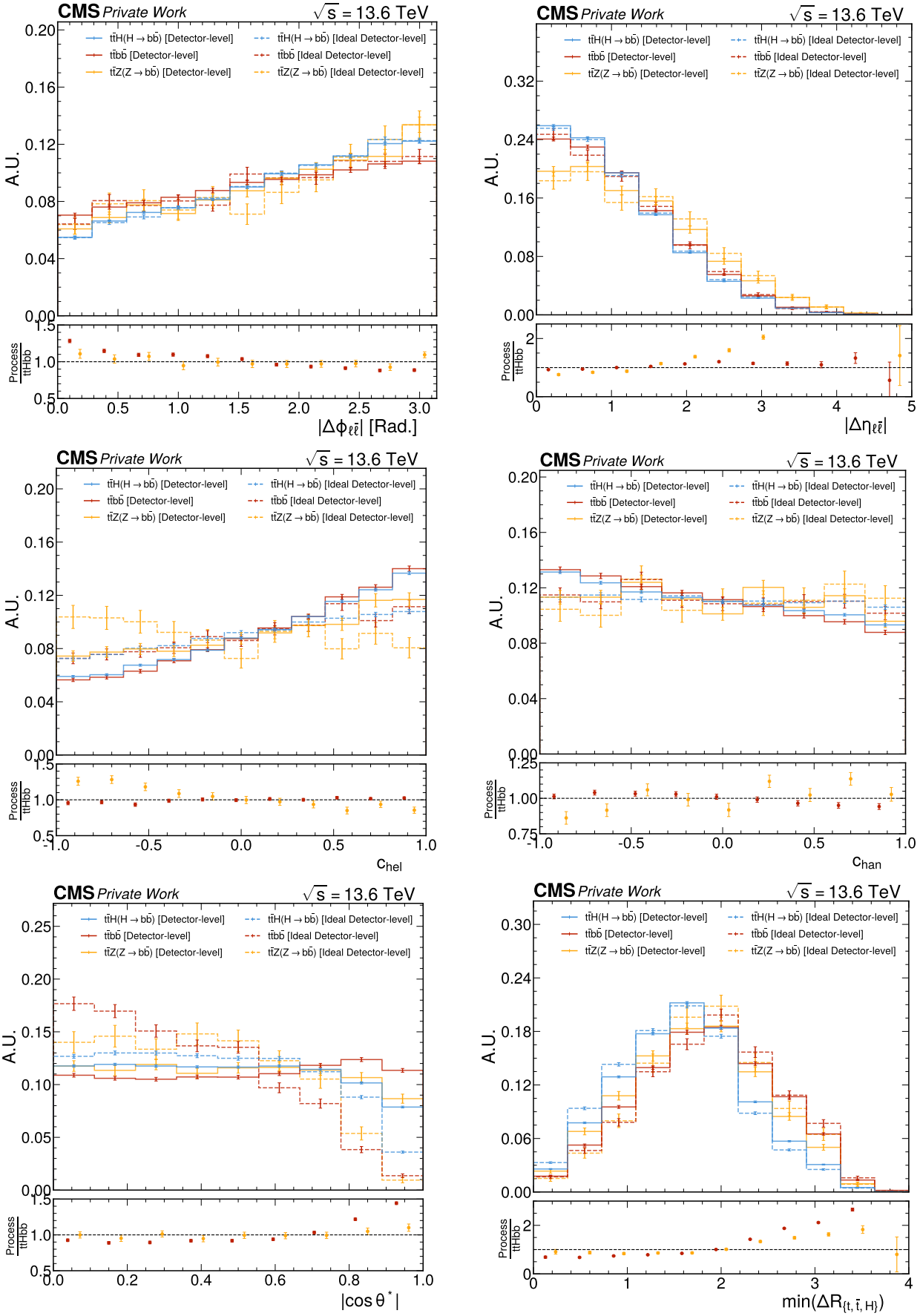

Figure 28: Spin-information and angular observables reconstructed at detector level using SPANet (solid lines) and assuming a perfect reconstruction with the network (dashed lines). Three processes are included: the $t\bar{t}H(\rightarrow b\bar{b})$ signal (blue line), $t\bar{t}+b\bar{b}$ background (red line) and $t\bar{t}Z(\rightarrow b\bar{b})$ (yellow line). Uncertainties are purely statistical. Ratio plots (below) are calculated for the background and $t\bar{t}Z$ over the signal.