

Search for heavy Higgs Bosons and Axion-Like Particles with the CMS Experiment via Deep Neural Networks

**Dissertation
zur Erlangung des Doktorgrades
an der Fakultät für Mathematik, Informatik und Naturwissenschaften
Fachbereich Physik
der Universität Hamburg**

vorgelegt von

Jörn Bach

Hamburg

2026

Gutachter/innen der Dissertation:

Prof. Dr. Christian Schwanenberger
Prof. Dr. Peer Stelldinger

Zusammensetzung der Prüfungskommission:

Prof. Dr. Jochen Liske
Prof. Dr. Gudrid Moortgat-Pick
Prof. Dr. Christian Schwanenberger
Prof. Dr. Peer Stelldinger
Dr. Alexander Grohsjean

Vorsitzende/r der Prüfungskommission:

Prof. Dr. Jochen Liske

Datum der Disputation:

07.05.2026

Vorsitzender des Fach-Promotionsausschusses PHYSIK:

Prof. Dr. Johannes Haller

Leiter des Fachbereichs PHYSIK:

Prof. Dr. Markus Drescher

Dekan der Fakultät MIN:

Prof. Dr.-Ing. Norbert Ritter

ABSTRACT

This interdisciplinary dissertation investigates several key physical and computational challenges inherent in high-energy physics analyses of top quark pair ($t\bar{t}$) production. As the heaviest known elementary particle, the top quark occupies a unique position within the Standard Model (SM), serving as a critical probe for new physics. With the advent of the Large Hadron Collider (LHC), physicists are now able to study precision effects in $t\bar{t}$ production with unprecedented detail.

The foundation of this research is a CMS (Compact Muon Solenoid) experiment analysis searching for heavy scalar or pseudoscalar particles that couple to top quark pairs via Yukawa interactions. A central result of this search is the observation of a significant excess at the $m_{t\bar{t}}$ production threshold. This work explores several interpretations of this excess, among them a reinterpretation of CMS results within the framework of Axion-Like Particles (ALPs) which could serve as part of an explanation for dark matter. Furthermore, the thesis evaluates the potential for interpreting such excesses in models incorporating CP-violation, supported by a sensitivity study conducted via Simulation-Based Inference (SBI).

To improve the sensitivity of future $t\bar{t}$ analyses, this work addresses the requirement for higher $m_{t\bar{t}}$ resolution and more accurate simulations. The former is discussed with the implementation of a Lorentz-algebra Transformer architecture designed to optimize the reconstruction of the top quark system. The latter is investigated by integration and validation of the `ttJ_MiNNLO` code within the CMS software framework to study the impact of next-to-next-to-leading-order (NNLO) corrections.

The dissertation further addresses the computational and statistical complexities of high-dimensional inference. In scenarios where the assumptions of Wilks' Theorem are not met, this work explores a novel approximation method for Feldman-Cousins inference.

Finally, the thesis tackles a critical bottleneck in applying machine learning to high-energy physics: the instability caused by simulated collision events with a negative weight through interference effects in Monte Carlo generated training datasets. To resolve this, a new algorithm (Pull by Partial Labels (PPL)) is introduced. Utilizing disambiguation-free partial label learning, the PPL algorithm is shown to stabilize training across controlled and experimental scenarios. Remarkably, the utility of the PPL method extends beyond physics; its interdisciplinary application is demonstrated in the field of "Machine Unlearning" (specifically class forgetting), where it performs competitively against established unlearning techniques.

Diese interdisziplinäre Dissertationsschrift behandelt einige der physikalischen und numerischen Schlüsselherausforderungen von Hochenergiephysikanalysen der Top Quark Paar Produktion.

Top Quarks stellen ein relevantes Forschungsfeld dar, weil sie als schwerstes bekanntes Elementarteilchen eine herausragende Stellung im Standard Modell der Teilchenphysik einnehmen. Mit dem LHC Collider sind Physiker erstmals in der Lage Präzisionseffekte in Top Quark Paaren zu studieren.

Diese Arbeit führt zuerst die theoretischen Zusammenhänge und das CMS (Compact Muon Solenoid) Experiment ein. Der Stand der Forschung im maschinellen Lernen in diesem Forschungsbereich wird ebenfalls erläutert. Die Grundlage für das Training vieler selbstlernender Algorithmen sind die Monte Carlo Simulationen. Auch diese werden eingeführt.

Die Grundlage dieser Arbeit ist eine CMS Analyse nach schweren skalaren oder pseudoskalaren Teilchen die mit einer Yukawa Interaktion an Top Quark Paare koppeln. Diese Analyse wird eingeführt und das Ergebnis wird erläutert. Ein signifikanter Exzess an der $m_{t\bar{t}}$ Top Paarproduktionsschwelle ist das zentrale Ergebnis dieser Suche.

Der Exzess kann in unterschiedlichen Interpretationen gedeutet werden. Diese Arbeit beleuchtet die Möglichkeit, dass es sich um ein Axion Like Particle handeln könnte und liefert eine Reinterpretation der CMS Ergebnisse in diesem Modell. Ebenso wird die Möglichkeit studiert, den Exzess künftig in einem theoretischen Modell, dass CP Verletzung erlaubt, zu interpretieren. Dafür wird eine Sensitivitätsstudie mit einer Simulation Based Inference (SBI) Methodik angefertigt.

Eine zentrale Rolle für die Erklärung der Beobachtung spielt die Sensitivität der Analyse. Künftige Analysen werden eine höhere $m_{t\bar{t}}$ Auflösung benötigen um mögliche Interpretationsszenarien zu unterscheiden. Eine Möglichkeit für einen Sensitivitätsgewinn ist der Einsatz von Methoden des maschinellen Lernens in der Rekonstruktion des Top Quark Systems. Hierfür wird eine Studie vorgestellt, die auf der Implementation einer Lorentz-Algebra Transformer Architektoren beruht.

Einen weiteren Sensitivitätsgewinn können genauere Monte Carlo Simulationen liefern. Der Effekt von Korrekturen in nächst-nächst höherer Ordnung wird anhand des `ttJ_MiNNLO` Codes studiert. Dafür wird dieser in CMS Software eingebaut und validiert.

Ebenfalls eine relevante Rolle spielt die Rechenzeiteffizienz in höherdimensionaler statistischer Inferenz. Die behandelte Analyse kann nicht vollumfänglich die Voraussetzungen des Theorems von Wilks erfüllen. Bereits die Simulation Based Inference kann in diesem Szenario helfen, es wird aber ebenso eine neue Näherungsmethode der Feldman-Cousins Inferenz studiert.

Um die Suche nach schweren pseudoskalaren Teilchen in Top Quark Paaren mithilfe von Klassifizierungsmodellen des maschinellen Lernens zu studieren muss das Problem negativer Gewichte in Trainingsdatensätzen gelöst werden. Dafür wird das Problem im Detail beschrieben und analysiert. Als Lösung wird ein neuer Algorithmus (PPL) vorgestellt, der mithilfe von disambiguierungsfreiem partiellen Label Lernens das Training stabilisieren kann. Die Überlegenheit des Algorithmus wird in kontrollierten, konstruierten und experimentnahen Szenarien bewiesen.

Eine interdisziplinäre Anwendung der PPL Methode kann im Klassenvergessen des Teilbereiches "Unlearning" des maschinellen Lernens gefunden werden. Diese bemerkenswert domänenübergreifende Anwendung wird anhand von Experimenten mit anderen klassischen "Unlearning" Methoden verglichen.

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen benutzt zu haben.

Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.

17.03.2026



Datum

Unterschrift der Doktorandin / des Doktoranden

CONTENTS

1	INTRODUCTION	1
2	LHC PHYSICS	3
2.1	Quantum Field Theories	3
2.2	The SM Lagrangian and its components	4
2.3	Open Questions of the Standard Model	10
2.4	Extensions of the Standard Model	11
3	THE EXPERIMENTAL SETUP	15
3.1	The Large Hadron Collider	15
3.2	The Compact Muon Solenoid Detector	15
3.3	Statistical Inference on experimental Data	18
4	MACHINE LEARNING IN HIGH ENERGY PHYSICS	21
4.1	A short introduction to Machine Learning	21
4.2	Machine Learning at the LHC	25
5	MODELLING OF PARTICLE COLLISIONS - MONTE CARLO EVENT GENERATION FOR EXPERIMENTS	29
5.1	Event Simulation at Hadron Colliders	29
5.2	MC Event Generators	29
5.3	Implementing the ttJ_MiNNLO Algorithm into CMS Software	31
6	SEARCH FOR HEAVY HIGGS BOSONS IN TOP-ANTITOP QUARK PAIR PRODUCTION	39
6.1	Data Analysis	39
6.2	Limit Setting beyond Wilks Theorem	47
6.3	CP Violation in Heavy Higgs to $t\bar{t}$ Decays	52
6.4	Towards Next Generation Analyses	61
7	SEARCH FOR AXION LIKE PARTICLES WITH THE CMS EXPERIMENT	67
7.1	Creating Templates via Reweighting	67
7.2	Deriving ALP Limits	70
7.3	ALP Limit Results	70
8	NEGATIVE WEIGHTS IN MACHINE LEARNING APPLICATIONS FOR THE LHC	77
8.1	Negative Event Weights	77
8.2	Assessing the Effect of negatively weighted Events on ML Classification Tasks	78
8.3	Mitigating Negative Event Weights	81
8.4	Proof of Concept for PPL: Revisiting MNIST	83
8.5	Evaluating Strategies for Dealing with negative Weights in Interference Scenarios: The Doubleslit Experiment	86
8.6	Analyzing Negative Weight Mitigation Method in a Mock Analysis Setup	91
9	UNLEARNING WITH PARTIAL LABEL LEARNING	99
9.1	Machine Unlearning	99
9.2	Appropriating the Pull by Partial Labels Approach for Unlearning	100
9.3	Discussion of Unlearning vs. negative Weights in Physics	103
10	SUMMARY	105
A	APPENDIX	107
A.1	Implementation of ttJ_MiNNLO into CMS Software	107
A.2	Additional Limit Results for CP violating Heavy Higgs Scenarios with MadMiners Morphing	110
A.3	More Distributions for Dileptonic Top-Antitop Pair Reconstruction with LGATR	114
A.4	ALP Reweighting Validation & Signal Templates	116
A.5	Negative Weights in a Doubleslit Toymodel - Less descriptive second Feature	121
A.6	Heavy Higgs Mock Analysis with different negative Weight Mitigation Schemes - Plots of Features	123
A.7	Using PPL and the Likelihood Ratio Trick to Reweight a Toy Distribution	124
	BIBLIOGRAPHY	127

NOTATION

FREQUENTLY USED SYMBOLS

Particle Physics

σ	Cross section
$\mathcal{M}_{fi}$	Transition matrix element between initial state i and final state f
$\mathcal{A}_{fi}$	Scattering amplitude
$\mathcal{L}$	Lagrangian density
p_T	Transverse momentum
η	Pseudorapidity, $\eta = -\ln \tan(\theta/2)$
ΔR	Angular separation, $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$
$\sqrt{s}$	Centre-of-mass energy
$m_{t\bar{t}}$	Invariant mass of the top quark pair system
m_t	Top quark pole mass ($m_t \approx 172.5 \text{ GeV}$)
Γ_t	Top quark decay width
p^μ	Four-momentum
α_s	Strong coupling constant
α_{em}	Electromagnetic coupling constant
f_a	ALP scale
μ_R, μ_F	Renormalisation and factorisation scales
H_T	Scalar sum of transverse momenta of all final-state objects
$E_T^{\text{miss}}, p_T^{\text{miss}}$	Missing transverse energy / momentum
Φ	Generic BSM spin-0 mediator (scalar or pseudoscalar)
H, A	Additional CP-even scalar and CP-odd pseudoscalar Higgs bosons
η_t	simplified model of $^1S_0^{[1]}$ pseudoscalar quasi-bound state of $t\bar{t}$
m_Φ, Γ_Φ	Mass and total decay width of Φ
$c_{t,\Phi}$	CP-even (scalar) top- Φ Yukawa coupling modifier
$\tilde{c}_{t,\Phi}$	CP-odd (pseudoscalar) top- Φ Yukawa coupling modifier
$g_{At\bar{t}}, g_{Ht\bar{t}}$	Coupling modifiers for pseudoscalar A and scalar H to top quarks
a	Axion-like particle (ALP) field
m_a	ALP mass
c_t	ALP-top quark coupling coefficient
$c_{\tilde{G}}$	ALP-gluon coupling coefficient
S_1, S_2	Physical mass eigenstates after loop-level mixing in the CP-violating model
c_{hel}	Spin correlation observable, $c_{\text{hel}} = c_{kk} + c_{rr} + c_{nn}$
c_{han}	Spin correlation observable, $c_{\text{han}} = -c_{kk} + c_{rr} + c_{nn}$

Machine Learning

$\hat{p}(x)$	Model-predicted class probability for input x
$\mathcal{L}_{\text{BCE}}$	Binary cross-entropy loss
$\mathcal{L}_{\text{CCE}}$	Categorical cross-entropy loss
$\mathcal{L}_{\text{MSE}}$	Mean squared error loss
$\mathcal{L}_{\text{CCL}}$	Classifier consistent loss
D_{retain}	Retain set (data to be preserved after unlearning)
D_{forget}	Forget set (data to be erased from the model)

Abbreviations

ALP	Axion-like particle
BSM	Beyond the Standard Model
CMS	Compact Muon Solenoid
DNN	Deep neural network
DKW	Dvoretzky-Kiefer-Wolfowitz (inequality)
FC	Feldman-Cousins
HL-LHC	High-luminosity Large Hadron Collider
LGATR	Lorentz Equivariant Geometric Algebra Transformer
LHC	Large Hadron Collider
MC	Monte Carlo
ME	Matrix element
MET	Missing transverse energy
MIA	Membership inference attack
ML	Machine learning
MSE	Mean squared error
NP	Nuisance parameter
PDF	Parton distribution function
PPL	Pull by Partial Labels
QCD	Quantum chromodynamics
QED	Quantum electrodynamics
QFT	Quantum field theory
ROC	Receiver operating characteristic
SBI	Simulation-based inference
SM	Standard Model of particle physics
2HDM	Two-Higgs-Doublet Model
VEV	Vacuum expectation value

This page is intentionally left blank.

INTRODUCTION

Particle physics is humanities daring endeavour to explain nature's smallest scales and understand its most fundamental interactions. An important contribution to this quest has come from the Large Hadron Collider and its experiments which have discovered the Higgs boson in 2012 [1, 2] and have thus completed our current best model, the Standard Model (SM). Since, the SM has been tested to unprecedented precision and many of its details have been understood in depth. Still, fundamental questions have not yet been answered and the search for beyond the SM (BSM) effects is central to the future.

Particularly the field of top quark pair physics provides a rich phenomenology of possible BSM contributions. Since the top quark is the heaviest known elementary particle, it couples most strongly to BSM particles which exhibit Yukawa coupling mechanisms. A popular theory in this regard is the idea of extended Higgs sectors. While the SM predicts a minimal Higgs sector with only the discovered h_{125} scalar at $m_h = 125$ GeV, there is no fundamental argument that excludes further Higgs particles. A simple addition to the SM could thus be another Higgs Doublet, giving rise to (among others) a new pseudoscalar and a scalar particle which in turn can decay to top quark pairs. Subsequently, analyses of the top decay products can give insights into possible intermediate scalar particles. The cleanest channel to search for such additional heavy Higgs bosons is the so-called dileptonic decay channel. It is characterized by both W bosons from the top quark decay decaying into a charged lepton and a neutrino. Here, spin correlations can be analyzed in addition to the invariant mass spectrum giving a handle on the symmetry structure of intermediate particles such as heavy Higgs bosons.

Chapter 2 and 3 give an introduction to the relevant theory for these analyses and explain the experimental setup of the CMS multi-purpose detector at the LHC.

In parallel to the experimental success of the LHC, new data analysis methods have developed and greatly enhanced the scientific impact the LHC program can have. The evolution of machine learning (ML) has fundamentally altered the way particle physics research is conducted. Furthermore, modern particle physics has proven to be a fruitful ground for ML innovation, with well-understood simulation data readily available and vast, complex data taken by the experiments. Chapter 4 therefore explains ML in the context of high energy physics.

This thesis explores exactly this intersection of particle physics and ML in the field of top quark pair production in the CMS experiment at the LHC.

First, Monte-Carlo simulations are investigated. Chapter 5 deals with the implementation and validation of the `ttJ_MiNNLO` generator for $t\bar{t}$ pairs which enables the CMS experiment to benefit from unprecedented accuracy in $t\bar{t}$ production simulations.

Chapter 6 explains the analysis setup for a search for heavy Higgs bosons. The statistical limit setting in cases where common approximations do not hold is evaluated and a proposal for a new, more efficient method is evaluated. A physically more challenging heavy Higgs scenario which includes CP violation is considered and a phenomenological study based on ML methods for direct likelihood estimation is conducted. To advance the sensitivity for all these heavy Higgs searches, a new physics-inspired ML based reconstruction method is implemented and evaluated.

Chapter 7 explores the signature of axion like particles in the CMS analysis presented before. These particles share their production mechanism through top quark coupling as well as the pseudoscalar structure. Additionally, they feature a direct gluon coupling. As they can be a promising candidate for dark matter, the full CMS search is interpreted with this signal model and exclusion limits are derived.

The results from the preceding chapters demonstrate the need for increased sensitivity in these searches. A possible solution can be the replacement of classical histogram based methods with ML classifiers. However, the particular peak-dip structure of (pseudo)scalar BSM contributions introduces a challenge for classical ML methods because of the presence of negatively weighted events. This challenge is analyzed in chapter 8. A solution based on partial label learning is proposed and tested across various scenarios.

A parallel to this problem can be found in the field of privacy for ML models. Unlearning data from a pretrained classifier is a challenging task that can be formalized in a similar fashion as the negative weights of the previous chapter. Logically, the same method from chapter 8 is proven in this scenario and compared to other unlearning methodologies.

Ultimately, chapter 10 summarizes the results and gives an outlook.

A concept is a brick. It can be used to build a courthouse of reason. Or it can be thrown through the window.

— Gilles Deleuze

The standard model of particle physics (SM) is the quantum field theory that describes all known fundamental particles and their interactions except for gravity.

2.1 QUANTUM FIELD THEORIES

Field theories emerged as the consequence of generalizing classical Lagrangian mechanics. They allowed to describe the interactions of continuous forces. The core object is the action functional:

$$S = \int d^4x \mathcal{L}(\phi_i, \partial\phi_i) \quad (2.1)$$

where $\mathcal{L}$ is the Lagrangian density which is a function of the fields ϕ_i . The principle of stationary action allows to extract the equations of motions. A systems full information is thus encoded in the Lagrangian density and the necessary boundary or initial conditions for solving these equations of motion.

2.1.1 Symmetries and Interactions

Noether's theorem associates symmetries with conserved quantities [3] marking an important shift in thought towards symmetry groups being fundamental to physical theories. Spacetime symmetries are thus directly connected to fundamental laws of nature such as the conservation of momentum which arises from translational symmetry. Demanding that physical theories should be invariant under inner (gauge) symmetries is a natural extension of this fact. Physics should stay invariant under transformations such as rotations.

Invariance under global gauge transformations of the form $\Psi \rightarrow \Psi' = e^{i\alpha} \Psi$ is trivial, since the free Dirac Lagrangian is invariant. For local symmetry transformations where the phase $\alpha = \alpha(x)$ is dependent on the spacetime coordinate, an extra term has to be added for the Lagrangian to be invariant. This gauge term introduces further particles into the theory, the gauge bosons or "force carriers". The SM is built on three fundamental gauge symmetries:

1. $U(1)_Y$ symmetry: unitary group of dimension 1 $\rightarrow$ electromagnetic force with hypercharge Y
2. $SU(2)_L$ symmetry: special unitary group of dimension 2 $\rightarrow$ weak force with weak isospin I
3. $SU(3)_C$ symmetry: special unitary group of dimension 3 $\rightarrow$ strong force with color charge C

or combined

$$SU(3)_C \times SU(2)_L \times U(1)_Y \quad (2.2)$$

In itself, these symmetries do not allow for explicit mass terms for the gauge bosons. The process of electroweak symmetry breaking (BEH mechanism) can introduce such terms and is further detailed in section 2.2.2.

In addition to these continuous symmetries, discrete symmetries can be observed. Unlike continuous symmetries, discrete symmetries form finite groups and do not lead to conserved currents through Noether's theorem. However, they impose important constraints on physical processes and selection rules for particle interactions. The most important symmetries are

- Parity Inversion : $\hat{P} \Psi(x, t) \rightarrow \Psi(-x, t)$
- Charge Conjugation: $\hat{C} \Psi(x, t) \rightarrow \bar{\Psi}(x, t)$
- Time Inversion: $\hat{T} \Psi(x, t) \rightarrow \Psi(x, -t)$

The combined CPT symmetry is proven to be exact for any local, Lorentz-invariant and unitary quantum field theory [4, 5] but the P and CP symmetries are violated by the weak force, as observed by experiments (Parity violation in the Wu experiment [6] and CP violation in K_L^0 decays [7]). This is a consequence of the chiral structure, the weak gauge bosons only couple to left-handed fermions and right handed antifermions. These findings carry important context for early universe thermodynamics and are closely linked to the matter-antimatter asymmetry, which is further detailed in 2.3.

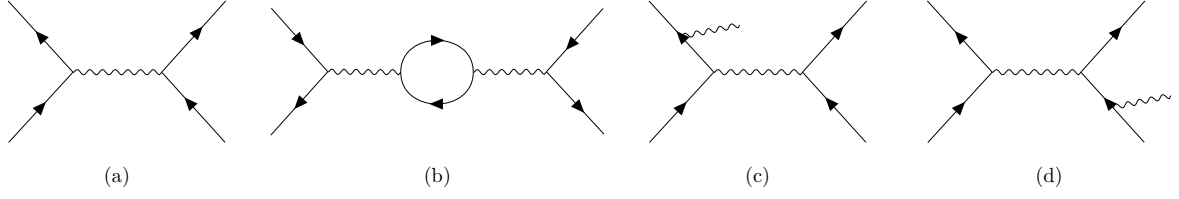


FIGURE 2.1: The first orders of perturbation theory illustrated as Feynman diagrams of QED processes. Here, the solid lines indicate fermions and the wavy lines indicate photons. (a) shows a leading-order tree-level process. (b) shows the first loop correction, which is next-to-leading order. (c) and (d) show initial and final state emissions, also called real corrections.

2.1.2 Matrix Elements and Perturbation Theory

The fundamental observable in particle physics is the cross section σ of a process, it gives the relative rate of a process occurring and thus encodes the physical probability. Fermis golden rule connects the cross section to the matrix element $\mathcal{M}$.

$$d\sigma = \frac{1}{\text{flux}} |\mathcal{M}_{fi}|^2 d\Phi_n \quad (2.3)$$

with the n -body Lorentz-invariant phase space element. The matrix element $\mathcal{M}_{fi}$ between the initial state i and the final state f can be computed from the sum of the amplitudes of all processes with indistinguishable initial and final states $\mathcal{M}_{fi} = \sum_i \mathcal{A}_{fi}$. The amplitudes can be expressed in a graphical form in Feynman diagrams which encode the amplitudes factors as vertices and legs. This sum is generally not analytically computable as there generally infinitely many intermediate states possible. In QFT, they can often be handled perturbatively as long as the coupling constant is small. Then the sum will converge and higher orders in the coupling constant can be treated as a correction to the lower orders. Figure 2.1(a) shows a tree level or leading order (LO) process for two-to-two scattering in quantum electrodynamics, which is of the lowest order in the coupling. It enters once for every vertex. Figure 2.1(b) shows a next-to-leading order (NLO) diagram that contributes to the same process. The fermionic loop introduces two new powers of the coupling into the equation and to correctly calculate the impact of this process, one has to integrate out all possible momenta for fermions in the loop. This commonly introduces divergences into the process. The diagrams Figure 2.1(c) and (d) show initial and final state emission. These processes can also introduce divergences for very soft or collinear emissions (Infrared or IR divergence). Generally, physical QFTs are considered to be non-divergent. The Kinoshita-Lee-Nauenberg theorem [8, 9] states, that the divergences from soft initial and final state emission cancel with the divergences from soft loops. Divergences introduced by very high momentum loops (Ultraviolet or UV divergences) have to be renormalized to allow for finite amplitudes. The dependence of the cross section on the absolute value squared introduces interference terms between different processes with indistinguishable initial and final states. Assuming two different processes $\mathcal{A}_1$ and $\mathcal{A}_2$ with the same initial and final states, the cross section is proportional to

$$\sigma \propto |\mathcal{M}|^2 = |\mathcal{A}_1 + \mathcal{A}_2|^2 = |\mathcal{A}_1|^2 + 2 \text{Re}(\mathcal{A}_1 \mathcal{A}_2^*) + |\mathcal{A}_2|^2 \quad (2.4)$$

While the squared absolute values are positive definite, the interference term can be negative. The overall cross section is positive definite to be physical. Negative interference contributions are called destructive interference and can also be observed in other areas of physics, such as optics.

2.2 THE SM LAGRANGIAN AND ITS COMPONENTS

The Standard Model Lagrangian reads

$$\mathcal{L}_{\text{SM}} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \bar{\psi}_i (i\gamma^\mu D_\mu - m_i) \psi_i + |D_\mu \phi|^2 - V(\phi) + \bar{\psi}_i y_{ij} \psi_j \phi + \text{h.c.} \quad (2.5)$$

Which encodes all relevant subatomic interactions and the particle content of the theory. The theories particle content can be split into four sectors that interact with each other through three fundamental forces. The structure of the SM is sketched in figure 2.2.

2.2.1 The Gauge Bosons

The interactions are mediated by gauge bosons that are constructed through the local gauge symmetries described above. Each symmetry gives rise to a specific and well defined set of gauge bosons.

The gauge boson of the $U(1)$ symmetry and thus the electromagnetic force is the photon (γ) and the associated

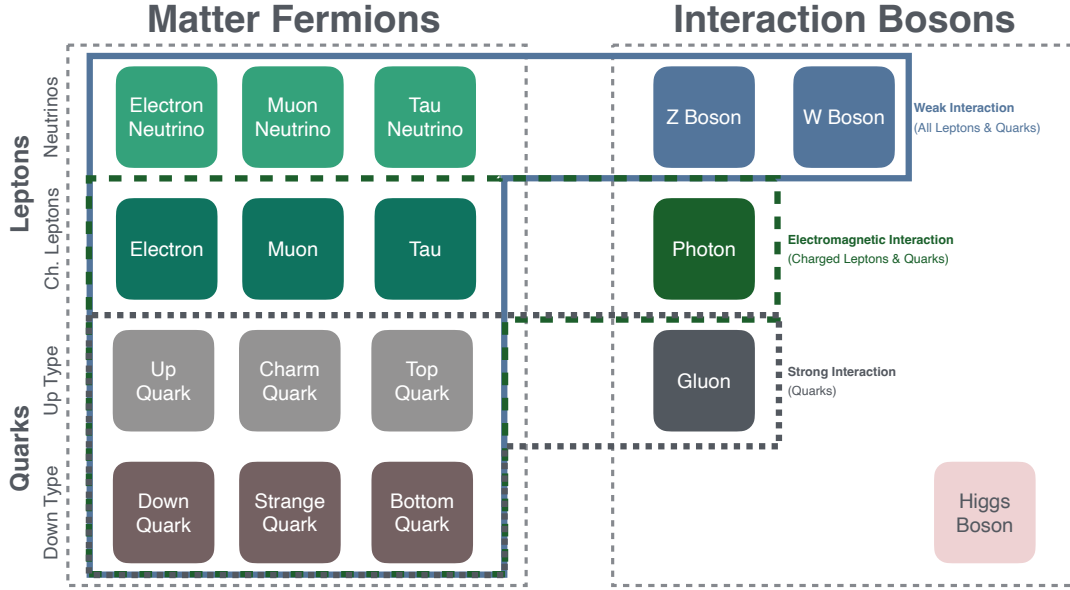


FIGURE 2.2: Particle Content of the Standard Model of particle physics.

field theory is Quantum Electrodynamics (QED). The photon is the quantization of the electromagnetic field, that has already been described in classical electrodynamics by the Maxwell equations. The conserved Noether charge associated to the $U(1)$ symmetry is the electrical charge. The coupling constant of QED is

$$\alpha_{\text{em}} = \frac{e^2}{4\pi}. \quad (2.6)$$

With e being the elementary charge. Quarks carry multiples of $\frac{1}{3}$ of the elementary charge and leptons carry either 1 e or -1 e . QFT considerations of renormalization result in an energy dependence of the coupling constant. α_{em} is close to $\frac{1}{137}$ for low energies but grows with larger energies. This phenomenon is known as running coupling. The $SU(2)$ symmetry gives rise to the gauge bosons of the weak interaction. The $SU(2)$ acts on a two dimensional vector space of spinors, which are called $SU(2)$ doublets. The doublets of the SM are formed by up- and down-type quarks or leptons and their respective neutrinos. The gauge bosons of this theory are three W bosons W_1 , W_2 and W_3 . They are not physically observable. To recover the physical $W^\pm$ and Z^0 bosons, electroweak (EW) unification and spontaneous symmetry breaking are required. For this, electromagnetic and weak gauge symmetries are combined: $U(1)_Y \times SU(2)$ with 3 W bosons and one B boson for $U(1)$. The Noether charges are the weak Isospin T and the weak hypercharge Y . This symmetry is then spontaneously broken and the electromagnetic $U(1)_{\text{em}}$ is recovered. The B and W_3 boson mix to form the electrically neutral γ and Z^0 bosons. Their mixing is determined by the weak mixing angle θ_W which can be measured experimentally. The other two W bosons form the physical charged W bosons W^+ and W^- . This process of electroweak symmetry breaking and the resulting particle is further discussed in chapter 2.2.2.

The $SU(3)$ symmetry of quantum chromodynamics (QCD) acts on three component vectors of spinors. These are labelled with the so called colorcharge of QCD, red, green and blue (or antired, antigreen and antiblue for antiparticles) which is also the Noether charge for this symmetry. The rather odd naming convention is chosen because either a combination of color and anticolor or a combination of all (anti-)colors form a neutral system in QCD analogous to how additive color mixing would give white from the primary colors [10]. The gauge bosons of this theory are the gluons. There are 8 of them and they carry a colorcharge and a different anticolorcharge. In contrast to QED, the QCD running coupling is inverted. α_s is small for big energies and grows (divergently) towards small energies. This gives rise to the phenomenon called color confinement. It is not possible to observe a bare color charge, the increasing coupling at low energies leads to hadronization below an energy threshold. On the opposite site, at high energies (and thus low distance regimes), the QCD interaction is weaker and the phenomenon of asymptotic freedom can be observed. It also has to be noted that for low energies, α_s grows to non-small values such that perturbation theory is not applicable anymore as terms of higher order will increase in impact. This regime must be treated in non-perturbative QCD (npQCD) in contrast to perturbative QCD (pQCD) calculations.

2.2.2 The Higgs Sector

Because explicit mass terms for the EW bosons would break gauge invariance [11], a different mechanism is needed to explain the short range nature (and thus observed mass of the W and Z bosons) of the EW interaction. The

Brout-Englert-Higgs mechanism (BEH) offers an explanation [12–14] based on spontaneous symmetry breaking. For this, an additional SU(2) doublet ϕ is added, introducing 4 degrees of freedom into the theory. It has a kinetic term $(D_\mu\phi)^\dagger D^\mu\phi$ which allows for couplings to the EW gauge bosons and a potential $V(\phi)$ is given by

$$V(\phi) = \mu^2\phi^\dagger\phi + \lambda(\phi^\dagger\phi)^2. \quad (2.7)$$

The parameters of the field μ^2 and λ are free parameters and assumed to adhere to $\mu^2 < 0$ and $\lambda > 0$ to form the "mexican hat" shape of the potential. In that case, there is a non-zero minimum at

$$|\phi| = v = \sqrt{\frac{-\mu^2}{\lambda}}. \quad (2.8)$$

which is also called the vacuum expectation value (VEV). The origin at $\phi = 0$ (which adheres to SU(2)×U(1) symmetry) is thus only metastable and the symmetry is simultaneously broken when the theory transitions from the high energy original state into the new minimum. If one considers the new field ϕ itself a fundamental field like the other SM fields, we can construct physical particles as fluctuations around this minimum. It can be written (by using SU(2) for gauge fixing) as

$$\phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix}. \quad (2.9)$$

Developing the Lagrangian in this expansion yields three massive bosons ($W^\pm$ and Z^0) and one massless boson (γ). Effectively, the four new degrees of freedom of the introduced field ϕ see three degrees of freedom absorbed into the electroweak bosons as masses and one degree of freedom expressed as a new particle. This massive scalar particle, often called Higgs-boson, can be identified as the h_{125} discovered by the ATLAS and CMS collaborations in 2012 [1, 2]. The masses follow from the factors in the expanded Lagrangian as

$$m_W = \frac{1}{2} \frac{e}{\theta_W} v, \quad m_Z = \frac{m_W}{\cos\theta_W} \quad \text{and} \quad m_H = \sqrt{-\mu^2}. \quad (2.10)$$

This also leads to the conclusion of $v = 246$ GeV. This mechanism can be extended for further SU(2) doublets, which is further discussed in 2.4.1. In addition to the gauge boson masses, the Yukawa interaction between the Higgs boson and the fermions allows for mass terms of the fermions. This mechanism thus also gives masses to the quarks and leptons in the SM. The value of the particles mass is then given by the Yukawa interaction strength between the Higgs and the particle. The term for the top mass (or analogously for any massive SM fermion except the neutrinos) is

$$g_{ht} = \sqrt{2} \frac{m_t}{v} \approx 1. \quad (2.11)$$

The masses can thus be verified by a measurement of the coupling or vice versa. This is an important test of the SM, which holds very well for the top quark [15].

2.2.3 The Leptonic Sector

The fermionic particles that do not interact via the strong force are grouped into the leptonic sector. There are three charged leptons (e , μ and τ) and their non-charged neutrino (ν_e , ν_μ and ν_τ) counterparts which form SU(2) doublets. The electron being the least massive is the only stable charged lepton. Muons, having ≈ 200 the mass of an electron, decay almost exclusively via the weak interaction into electrons. Their average free travel length is large compared to that of other elementary particles thus necessitating specialized detector components for reconstruction. The heaviest charged lepton, τ , has a mass of $m_\tau \approx 1.7$ GeV. It can thus also decay hadronically, making it very challenging to reconstruct.

Neutrinos only interact via the weak interaction so they are infeasible to detect at collider experiments. Information about them can be inferred from missing transverse momentum albeit with a very limited precision. There is no proven mechanism to give neutrinos masses as their right handed fields would not interact and they are considered massless in the SM. However, the existence of neutrino oscillations implies a mass basis [16], rendering them massive particles. Cosmological observations place strict bounds on the neutrino masses [17]. Their low masses also rule neutrinos out as dark matter candidates. They would need to be considered as hot dark matter which would wash out cosmic structure formation, which is not observed.

2.2.4 The Quark Sector

The quarks are fermionic particles that interact through the strong nuclear force and carry its color-charge. They are further divided into up and down type quarks. Up type quarks carry $2/3$ of an electrical elementary charge e , while down type quarks are negatively charged with $-1/3$ e . The lightest quarks of their respective category are the up-quark and down-quark which form protons and neutrons and thus all of the regular nucleons from the periodic table of elements. Their heavier siblings of the second generation are the charm and strange quark respectively. The

heaviest quark generation is comprised of the down type bottom quark and the up type top quark. The top quark has some special features due to its high mass of ≈ 172.5 GeV [18] and is the focus of the analysis in this thesis. Quarks interact, in addition to the strong force, also through the electroweak force. The electroweak interacting quarks are however not the same as the quarks in the mass basis. The rotation between these bases is described by the Cabibbo–Kobayashi–Maskawa matrix (CKM). The relationship is

$$\begin{bmatrix} d' \\ s' \\ b' \end{bmatrix} = \begin{bmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{bmatrix} \begin{bmatrix} d \\ s \\ b \end{bmatrix}. \quad (2.12)$$

The matrix elements of the CKM matrix are often parametrized in three mixing angles and one CP violating phase. They are experimentally measured and the diagonal elements of the matrix dominate. Off-diagonal elements are small which means that mixing between quark generations is small. This can be observed in the branching ratio of tops decaying into b quarks. Decays into other quark generations are called Cabbibo-suppressed or CKM-suppressed.

Due to the quark confining nature of the strong interaction, it is impossible to observe bare quarks. They either decay (in the case of the top quark) or form bound states. Quark bound states are called hadrons. They are further divided into mesons (integer spin) and baryons (spin $\frac{1}{2}$). Mesons are quark-antiquark particles, like the pions ($u\bar{d}$). Baryons are three quark combinations like the proton (uud) and the neutron (udd). At higher energies, it can be observed that bound states of hadrons have a rich inner structure due to virtual gluons and virtual quarks. This is especially important at hadron colliders and is further discussed in chapter 5.1.

2.2.5 The Top Quark

The top quark is the heaviest fundamental particle described in the SM. Its mass is over $40\times$ that of its quark generation partner. The top quarks decay to W bosons and b quarks. Due to Cabbibo suppression, the branching ratio (BR) is almost 1. Throughout this thesis, this assumption is made. Subsequently, one classifies the top quark decays by the W boson decays. These can decay leptonically into a lepton l and a neutrino ν (BR of 0.332) or hadronically into a quark-antiquark pair (BR of 0.665). Top quarks are predominantly produced in pairs in proton-proton collisions (as in the LHC at 13 TeV, see chapter 3), where the cross section is $\sigma_{t\bar{t}} \approx 830$ pb vs $\sigma_{\text{single } t} \approx 215$ pb). The decays of the $t\bar{t}$ system are thus classified into three classes.

- fully hadronic: $t\bar{t} \rightarrow W^- W^+ b \bar{b} \rightarrow b \bar{b} q \bar{q}' \bar{q}'$
- semi-leptonic (or ℓj): $t\bar{t} \rightarrow W^- W^+ b \bar{b} \rightarrow b \bar{b} q \bar{q}' l \bar{\nu}_l$
- dileptonic (or $\ell\ell$): $t\bar{t} \rightarrow W^- W^+ b \bar{b} \rightarrow b \bar{b} l \bar{l} \bar{\nu}_l \nu_l$

In this thesis, leptonic decays are not accounting for τ final states ¹ as these are hard to analyze because of the aforementioned complex decays of the τ . With a decay width of $\Gamma = 1.36$ GeV it decays quicker than it can hadronize ($\Lambda_{\text{QCD}} = 250$ MeV). Conventionally, it is assumed that the top quark thus only exists as a bare particle and does not form hadrons. This assumption is further discussed in 6.1.2. Because of this behavior, one can study the top properties in a unique setting that no other quark allows for. In particular, spin polarization and correlation observables for $t\bar{t}$ systems are accessible in the decay products [19]. These have been studied at the Tevatron [20, 21] and the LHC [22, 23]. As the information in these observables is central to the search conducted, a dedicated chapter on $t\bar{t}$ events and the spin information in $t\bar{t}$ events is presented here.

2.2.5.1 Top Quark Pair Production and Spin Information

The top quark pair production process has three relevant leading-order diagrams at a hadron collider as shown in figure 2.3. These can be summarized into quark-antiquark annihilation and gluon fusion. Since the LHC is a hadron hadron collider, antiquarks are only accessible as sea quarks in the protons and due to the high energy the low x tail of the gluon pdf is dominant, where gluons are highly suppressing the sea quark contributions. The gluon fusion process is thus dominant. NLO contributions add more diagrams while also modifying the shown processes through initial and final state radiation. There are normalization and kinematic effects from taking NLO corrections into account. Furthermore, NNLO effects have been studied theoretically [24]. It has been demonstrated that, in the context of unfolded measurements of $p_{T,t}$ and double-differential distributions, only a framework incorporating NNLO corrections facilitates a precise description. A more detailed discussion of this topic can be found in chapter 5.2.1.2.

A top quark pair system consists of two top quarks with spin $s = \frac{1}{2}$. The systems total state can thus be $S = 0$ (scalar) or $S = 1$ (vector). Additionally, an orbital angular momentum L of non-negative integer value has to be considered. A commonly used method to describe the $t\bar{t}$ spin state is thus - akin to atomic physics term symbol -

$$^{2S+1}L_J \quad (2.13)$$

¹ except for $\tau \rightarrow e^\mp \mu^\pm$

where the total angular momentum is $\vec{J} = \vec{S} + \vec{L}$. L is often given in the spectroscopic notation of letters (S for $L = 0$, P for $L = 1$, D for $L = 2$, F for $L = 3$, ...). In hadron collisions, a mixture of these states is produced. The initial state and the possible presence of resonant intermediate particles play a role in calculating the exact contribution of each spin state.

To understand the spin analyzing power of the decay leptons, one has to remember that the electroweak force that is facilitating the decay is maximally parity violating and that the b quark, leptons and neutrinos are almost massless compared to the top quark. Considerations of conservation of angular momentum lead to the situation that the charged antilepton is preferably emitted into the top quark spin direction. The charged lepton in the antitop decay is emitted into the opposite direction of the antitop quark spin. For final states including hadronization (jets) as well as the leptonic decay of the other quarks, this statement does not hold as the strong force that effects these processes is not related to the particles helicity. The spin information is only conserved in the parity violating decay of the W boson into a lepton and a neutrino.

As discussed before, the top quark decay products allow to study the properties of the bare quark due to its fast decay. The following introduces the spin density matrix as a tool to quantize the spin polarisation and correlation effects. The spin density matrix $\mathbf{R}$ - as defined in Ref. [19, 25, 26] - decomposes the differential production cross section of $t\bar{t}$ pairs into a kinematic component, a polarisation component and a spin correlation component. It is dependent on the center of mass energy $\hat{s}$ and can be expressed as

$$\mathbf{R} = A\mathbb{1} \otimes \mathbb{1} + B_i^1 \sigma_i \otimes \mathbb{1} + B_i^2 \mathbb{1} \otimes \sigma_i + C_{ij} \sigma_i \otimes \sigma_j \quad (2.14)$$

$\mathbb{1}$ is the 2D unity matrix, σ_i are the Pauli matrices. A determines the overall cross section of the $t\bar{t}$ production differential in kinematic properties such as $m_{t\bar{t}}$. The $\vec{B}^{1/2}$ vectors describe the spin polarizations of the top and antitop quark. These will not be further discussed as they are approximately zero in the SM.

The spin correlation matrix C is a 3×3 matrix that encodes the correlation information. The i,j th element is indicative of the correlation between the i th component of the top quark spin and the j th component of the antitop quark spin ($-1 \leq C_{ij} \leq 1$). The spin sensitive $\vec{B}$ and C are dependent on the top scattering angle θ_t^* between the top quark and the beam axis. The evaluation of predictions needs a basis, here the basis $\{\hat{k}, \hat{r}, \hat{n}\}$ as given in Ref. [19] is chosen. $\hat{k}$ is the direction of the top quark. Subsequently, $\hat{r}$ and $\hat{n}$ are given by

$$\{\hat{r}, \hat{k}, \hat{n}\}: \quad \hat{r} = \frac{1}{r}(\hat{p} - y\hat{k}), \quad \hat{n} = \frac{1}{r}(\hat{p} \times \hat{k}), \quad y = \hat{k} \cdot \hat{p} = \cos \theta_t^*, \quad r = \sqrt{1 - y^2}. \quad (2.15)$$

with $\hat{p} = (0, 0, 1)$ being the beam axis as a proxy for the incoming parton direction. This right handed basis system is visualized in figure 2.4.

Following the motivation given above, the lepton unit momenta in their respective (anti-)top quark rest system (here written as $\hat{\ell}_t^+$ for the top quarks antilepton and $\hat{\ell}_{\bar{t}}^-$ for the antitop quarks lepton) encodes the spin information, the differential normalized cross section is given by

$$\frac{1}{\sigma} \frac{d\sigma}{d\Omega_+ d\Omega_-} = \frac{1}{(4\pi)^2} \left(1 + \vec{B}^1 \cdot \hat{\ell}_t^+ + \vec{B}^2 \cdot \hat{\ell}_{\bar{t}}^- + (\hat{\ell}_t^+)^T \mathbf{C} \hat{\ell}_{\bar{t}}^- \right). \quad (2.16)$$

The basis $\{\hat{k}, \hat{r}, \hat{n}\}$ has the advantage, that this can be decomposed into components, which in turn can be used to define observables which are sensitive to a wide range of symmetries in the $t\bar{t}$ productions. The diagonal components of C are induced by interactions conserving C and CP symmetry, so they can be large for QCD $t\bar{t}$ production. The off-diagonal elements are induced by C and CP violating interactions, so one can explicitly test for different types of C and CP violating $t\bar{t}$ production using their sums and differences. Integrating out the azimuthal

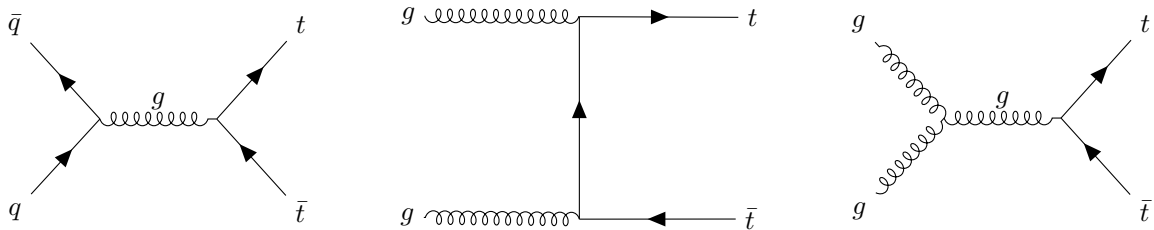


FIGURE 2.3: The most important leading order processes of top quark pair production at a hadron collider.

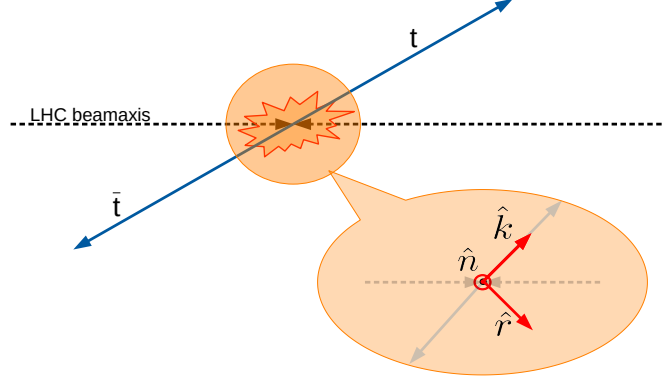


FIGURE 2.4: Schematic of the top quark system with the set of basis unit vectors $\{\hat{k}, \hat{r}, \hat{n}\}$ that are used for the spin correlation matrix.

angles from equation 2.16, one can derive single and double differential distributions for the scattering angle of the top (θ_1^i) and antitop (θ_2^j) with respect to the basis vectors. The single differential distribution are given by

$$\frac{1}{\sigma} \frac{d\sigma}{d \cos \theta_a^i} = \frac{1}{2} \left(1 + B_i^a \cos \theta_a^i \right), \quad (2.17)$$

and the double differential distributions for the components of the spin correlation matrix

$$\frac{1}{\sigma} \frac{d\sigma}{d \cos \theta_1^i} \frac{d\sigma}{d \cos \theta_2^j} = \frac{1}{2} \left(1 - C_{ij} \cos \theta_1^i \cos \theta_2^j \right) \ln \left(\frac{1}{|\cos \theta_1^i \cos \theta_2^j|} \right). \quad (2.18)$$

The sum and differences of C_{ij} , which are sensitive to discrete symmetries, follow the equation

$$\frac{1}{\sigma} \frac{d\sigma}{dx_{\pm}} = \frac{1}{2} \left(1 - \frac{C_{ij} \pm C_{ji}}{2} x \right) \cos^{-1} |x_{\pm}|, \quad \text{with } x_{\pm} = \cos \theta_1^i \cos \theta_2^j \pm \cos \theta_1^j \cos \theta_2^i \quad (2.19)$$

for $i \neq j$. These equations describe the analytical relation between the observable differential distributions and the coefficients to be extracted from this observation. Following Ref. [27], we denote the observables connected to the coefficients C_{ij} in lowercase letters, so $c_{ij} \equiv \cos \theta_1^i \cos \theta_2^j$. In addition to the 6 $c_{ij} \pm c_{ji}$ observables which are sensitive to the CP and P properties of the $t\bar{t}$ system, we construct two observables which are sensitive to the $t\bar{t}$ systems spin state, c_{hel} and c_{han} . Both are linear combinations of the diagonal elements of the spin density matrix.

$$c_{\text{hel}} = +c_{kk} + c_{rr} + c_{nn} \quad (2.20)$$

and

$$c_{\text{han}} = -c_{kk} + c_{rr} + c_{nn} \quad (2.21)$$

It is evident, that $c_{\text{hel}} = \hat{\ell}_t^+ \cdot \hat{\ell}_{\bar{t}}^-$ and c_{han} is the dot product of the lepton unit momenta with the top direction component ($\hat{k}$) sign-flipped in one of the vectors. There is also an analytical relation for these variables.

$$\frac{1}{\sigma} \frac{d\sigma}{dc_{\text{hel}}} = \frac{1}{2} \left(1 - D^{(1)} c_{\text{hel}} \right) \quad (2.22)$$

and

$$\frac{1}{\sigma} \frac{d\sigma}{dc_{\text{han}}} = \frac{1}{2} \left(1 - D^{(k)} c_{\text{han}} \right) \quad (2.23)$$

with the coefficients $D^{(1)} = \frac{1}{3} \sum_{i \in \{k,r,n\}} C_{ii}$ and $D^{(k)} = \frac{1}{3} (-C_{kk} + C_{rr} + C_{nn})$. As discussed in Ref. [25], the D coefficients can be used to test for quantum entanglement as done in the observation papers of ATLAS and CMS [28, 29]. $D^{(1)}$ is maximally negative for pure singlet states, so that the resulting slope in the c_{hel} observable is positive. For triplet states, $C^{(k)}$ is maximally positive, so the c_{han} observable has the maximally negative slope. The combination of c_{hel} and c_{han} is a suitable set of variables to distinguish singlet and triplet states. An overview over all of the discussed variables which are used in this work is given in table 2.1. In general, it should be noted that since all the observables are computed by boosting the laboratory frame leptons into the $t\bar{t}$ ZMF and then boosting the leptons into their respective top parent ZMFs, a dependence on the kinematics of the final state is largely eliminated. While observables in the laboratory frame, like the azimuthal angle difference between the leptons $\Delta\phi_{\ell\ell}$, can be measured with more precision, they are also impacted by the $t\bar{t}$ kinematics and as such they are less sensitive to the aforementioned effects.

Observable	Sensitivity to
$c_{rk} + c_{kr}$	P-, CP even
$c_{rk} - c_{kr}$	P even, CP odd
$c_{nr} + c_{rn}$	P odd, CP even
$c_{nr} - c_{rn}$	P odd, CP odd
$c_{nk} + c_{kn}$	P odd, CP even
$c_{nk} - c_{kn}$	P odd, CP odd
$c_{kk} + c_{rr} + c_{nn}$	Spin Singlet vs Triplet
$-c_{kk} + c_{rr} + c_{nn}$	Spin Singlet vs Triplet

TABLE 2.1: Overview over the spin correlation variables derived from the spin density matrix.

2.3 OPEN QUESTIONS OF THE STANDARD MODEL

The standard model explains all observed particles at current colliders and describes the observed interactions with great accuracy, bar the gravitational interaction which is not described. However, there is a set of open problems which can be classified into aesthetic problems and incompleteness.

The aesthetic problems include the Higgs mass hierarchy problem, the strong CP problem and critique of the number of free parameters. The Higgs mass renormalization includes terms with corrections of very high energy up to the Planck scale, yet the measured Higgs mass is 125 GeV. A 'fine tuning' explanation, which requires the corrections to cancel very accurately, is possible but explanations from first principles are actively searched for.

Another aesthetic problem is posed by the free parameter of the strong CP phase. Due to CP violating effects in non perturbative QCD, one can write an effective QCD Lagrangian that contains a term

$$\mathcal{L}_{\text{QCD}} \supset \theta \frac{g_s^2}{32\pi^2} G\tilde{G} \quad (2.24)$$

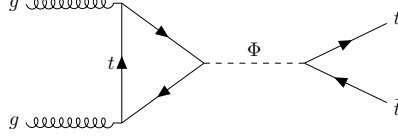
with the gluon field strength G and its dual $\tilde{G}$ [30] where the color and Lorentz indices are omitted for brevity. This term is CP violating and would thus allow for CP violating processes in QCD. The angle $\theta \in [0, 2\pi)$ can theoretically take any value, yet measurements of the neutron electric dipole moment indicate that it is extremely close to 0 [31] without an explanation from first principles. The QCD axion theory could solve this issue and is further discussed in 2.4.2.

Perhaps the most famous incompleteness of the SM is the observation of dark matter (DM) on almost all length scales except the quantum level. Observations in astronomy have indicated that there is a substantial amount of matter in the universe that cannot be explained by the SM. On galactic scales, missing mass can be observed by measuring rotational curves of galaxies. The profile of visible matter suggest that the radial velocity should drop off with larger radius r , but observations provide evidence for galaxy rotation to be flat to much higher radii [32]. Using Virial's theorem, first estimates on galaxy clusters indicated that the visible matter only accounts for a small percentage of the gravitationally bound mass [33]. Another strong evidence for dark matter on galaxy cluster scales come from gravitational lensing measurements [34]. The curvature of spacetime induced by the mass of the galaxy clusters is several times higher than visible matter would indicate, which leads to a stronger lensing effect on light which then can be measured. On cosmological scales, there is evidence for excess invisible matter from baryonic acoustic oscillations and their imprint on the cosmic microwave background [17]. These detailed measurements of the large scale structure of the universe also allow to derive constraints on the type of DM. The current astronomical observations require DM to be

1. electrically neutral and colorless, so neither strongly or electrodynamically interacting
2. non-baryonic
3. cold, so slowly moving compared to c
4. long lived.

There are multiple approaches to solve this problem, most of them focus on new particles beyond the SM. Some of the theories are discussed in this thesis and are expanded upon in the following chapter.

Another outstanding problem in the context of the SM is the matter-antimatter asymmetry in the universe. The Standard Model of cosmology, when combined with the Standard Model of particle physics, predicts that matter and antimatter should have been created in about equal quantities during the Big Bang. If this symmetry had been perfect, matter and antimatter would have annihilated into photons, leaving behind a Universe with no baryonic matter. However, observations give a baryon-to-photon ratio of $\eta_B = \frac{n_B - n_{\bar{B}}}{n_\gamma} = 6.1 \times 10^{-10}$ [17]. In the context of Baryogenesis, the theory that explains the baryon content of the universe, there are three conditions that need to be met to explain that number [35]. The so called Sakharov conditions are

FIGURE 2.5: The Feynman graph of the considered process of $gg \rightarrow \Phi \rightarrow t\bar{t}$.

1. Baryon number violation - processes must exist that change the net baryon number
2. C and CP violation at a large enough rate in the early universe
3. Departure from thermal equilibrium - to prevent CPT invariance from enforcing equal baryon and antibaryon production rates

The SM can satisfy all conditions in principle, but the amplitude of CP violation of the EW interaction is not large enough to explain the observed matter-antimatter asymmetry [36]. Consequently, BSM physics is searched for that allows for new CP violating processes to occur at a large enough rate. At last, it needs mentioning that the minimal SM cannot explain neutrino masses.

2.4 EXTENSIONS OF THE STANDARD MODEL

The aforementioned problems have prompted a multitude of theories to solve these. Many of these theories replicate the SM in sufficiently low energetic regimes to confirm to the agreement between data and SM predictions, so they are often called beyond the standard model (BSM) theories. Discussing them all in depth is impossible, so the following chapter will focus on two theories which have been experimentally confronted in the context of this thesis. Both predict spin-0 states that would be observable in $t\bar{t}$ decays which ties them together experimentally but fundamentally they are different theories born from different motivations.

2.4.1 Extended Higgs Sectors

The BEH mechanism described in 2.2.2 has no inherent limitation to a single added $SU(2)$. Adding more doublets produces a richer structure of the Higgs sector after EW symmetry breaking. The simplest case is the Two-Higgs-Doublet Model (2HDM) [37]. Here, one additional complex $SU(2)$ doublet is added to the BEH mechanism. The four new degrees of freedom manifest in new particles such that there is 5 physical Higgs fields after symmetry breaking (two electrically neutral scalar bosons h and H , one electrically neutral pseudoscalar boson A and two electrically charged bosons $H^\pm$). These 2HDM models arise in the context of supersymmetric [38] and axion extensions of the SM [39]. One can identify the h with the observed Higgs boson at 125 GeV (h_{125}) and then set out to search for the additional (charged) Higgs bosons. The model space has been constrained in global fits [40] and in direct searches [41, 42] but there are regions of phase space left in reach of the LHC experiments.

One way to provide widely interpretable limits, is the usage of simplified models. They introduce the new Higgs particles with free masses and widths as well as free coupling modifiers to the SM particles involved in the decay and production. This way, exclusion limits are not constrained to a particular realization of the 2HDM model and can be generally used by theoretical efforts to constrain particular models. A simplified model that is of particular interest in this thesis considers only the pseudoscalar A and scalar H to be within reach, so the charged Higgses contribution is neglected. The Lagrangian of an additional BSM spin-0 scalar Φ coupling only to $t\bar{t}$ via Yukawa interactions can be written as

$$\mathcal{L}_\Phi = \frac{1}{2} (\partial_\mu \Phi) (\partial^\mu \Phi) + \frac{m_\Phi^2}{2} \Phi^2 + \frac{m_t}{v} \bar{t} \Phi (c_{t,\Phi} + i\gamma_5 \tilde{c}_{t,\Phi}) t \quad (2.25)$$

with the mass m_Φ and the CP-even $c_{t,\Phi}$ and CP-odd $\tilde{c}_{t,\Phi}$ coupling modifiers. Note, that Φ is a generic mass eigenstate of a mixture of the CP-even H boson and the CP-odd A boson. In the case of no CP violation in the extended Higgs sector, the A and H don't mix and $\Phi \in \{A, H\}$. The couplings are mutually exclusive in the two cases

- $\Phi = A \rightarrow \tilde{c}_{t,\Phi} = g_{A t \bar{t}}$ and $c_{t,\Phi} = 0$
- $\Phi = H \rightarrow c_{t,\Phi} = g_{H t \bar{t}}$ and $\tilde{c}_{t,\Phi} = 0$.

These particles, like the SM h_{125} counterpart, are predominantly produced in gluon fusion through a $t\bar{t}$ loop at the LHC. This thesis considers the $t\bar{t}$ decays, as these are dominant in the mass phase space beyond the energy threshold that enables $t\bar{t}$ production (for $m_\Phi > 2m_t \approx 350$ GeV). The Feynman graph of this process is shown in figure 2.5. Since the final state is identical to the SM $t\bar{t}$ production, interference terms have to be considered.

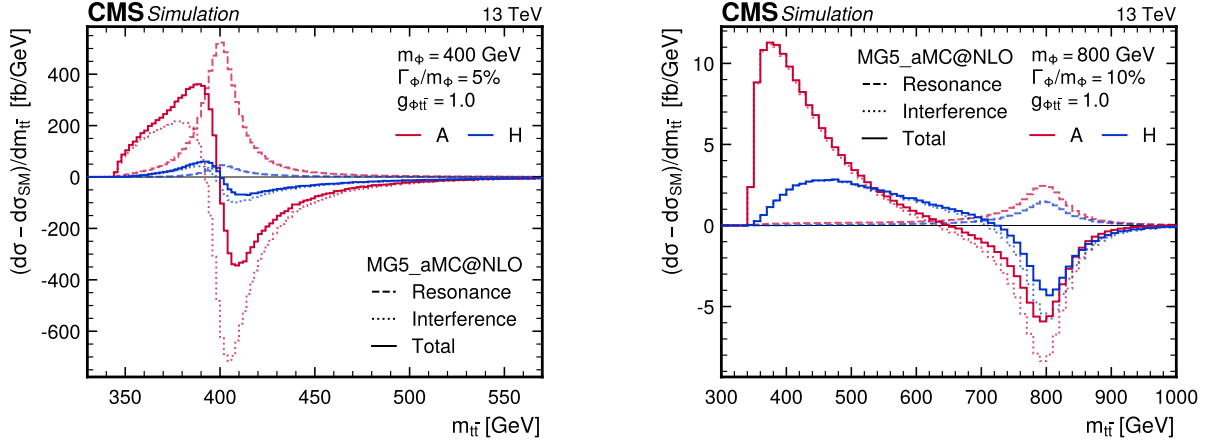


FIGURE 2.6: Differential cross sections for $A/H \rightarrow t\bar{t}$ as a function of $m_{t\bar{t}}$ with the SM $t\bar{t}$ production subtracted. Two cases for both A and H are shown, $m_\Phi = 400$ GeV with a relative width of $\Gamma_\Phi/m_\Phi = 0.05$ in the left plot and $m_\Phi = 800$ GeV with a relative width of $\Gamma_\Phi/m_\Phi = 0.1$ in the right plot. The interference and resonant contributions are given individually (dotted and dashed line) and summed (solid line). The plot is taken from Ref. [43].

Calculations show, that the interference term has a big impact on the shape in the $m_{t\bar{t}}$ spectrum, as it introduces a peak-dip structure at the value of the mediator mass [44]. The height and width of these features is determined by the masses and widths. It is also significantly different for pseudoscalar and scalar mediators. Two cases are plotted in figure 2.6. One can observe an additional broad peak at low $m_{t\bar{t}}$ values. This is induced by the gluon PDF which peaks at low $m_{t\bar{t}}$ and thus gives rise to an increased constructive interference effect. For statistical analysis, it is also important to note the scaling behavior of the differential cross section. The interference term scales with the Φ to $t\bar{t}$ coupling squared while the resonant term scales to the power of 4.

$$d\sigma = d\sigma^{\text{SM}} + g_{A/Ht\bar{t}}^2 d\sigma^{\text{int}} + g_{A/Ht\bar{t}}^4 d\sigma^{\text{res}} \quad (2.26)$$

with the differential cross sections $d\sigma^{\text{SM}}$ for the SM term, $d\sigma^{\text{int}}$ for the SM-A/H interference term and the $d\sigma^{\text{res}}$ for the resonant A/H term.

2.4.1.1 CP Violating Extended Higgs Sectors

The CP-violating case warrants special attention, as the mixing between A and H introduces further interference terms between the two BSM scalars. In this general case, Φ_1 and Φ_2 are taken to be the tree level mass eigenstates which are mixed realizations of the CP eigenstates A and H. The paper [45] highlights the necessity to correct for mixing contributions that arise from higher order EW corrections. This is described in the Z-factor formalism [46, 47] and introduces a further complex 2×2 mixing matrix between the tree level states such that

$$c_{t,1} \rightarrow Z_{\Phi_1, \text{tree}}^{\Phi_1} c_{t,1} + Z_{\Phi_2, \text{tree}}^{\Phi_1} c_{t,2}, \quad (2.27)$$

$$\tilde{c}_{t,1} \rightarrow Z_{\Phi_1, \text{tree}}^{\Phi_1} \tilde{c}_{t,1} + Z_{\Phi_2, \text{tree}}^{\Phi_1} \tilde{c}_{t,2}, \quad (2.28)$$

$$c_{t,2} \rightarrow Z_{\Phi_2, \text{tree}}^{\Phi_2} c_{t,2} + Z_{\Phi_1, \text{tree}}^{\Phi_2} c_{t,1}, \quad (2.29)$$

$$\tilde{c}_{t,2} \rightarrow Z_{\Phi_2, \text{tree}}^{\Phi_2} \tilde{c}_{t,2} + Z_{\Phi_1, \text{tree}}^{\Phi_2} \tilde{c}_{t,1}. \quad (2.30)$$

This mixing is dependent on the mass degeneracy of the tree-level scalars. It is found to be small (with a diagonal Z matrix) for a large mass separation, while the mixing from the EW correction is important in the near mass degenerate case. Since the Z-matrix is complex, further imaginary terms can enter the amplitude and lead to a richer shape variety in the $m_{t\bar{t}}$ spectrum. The differential cross section as computed in Ref. [45] scales with

$$\begin{aligned} d\sigma = & d\sigma^{\text{SM}} + (c_{t,1} + i\gamma_5 \tilde{c}_{t,1})^4 d\sigma^{\text{res}, S_1} + (c_{t,2} + i\gamma_5 \tilde{c}_{t,2})^4 d\sigma^{\text{res}, S_2} + \\ & + (c_{t,1} + i\gamma_5 \tilde{c}_{t,1})^2 d\sigma^{\text{int}, S_1 \text{ to SM}} + (c_{t,2} + i\gamma_5 \tilde{c}_{t,2})^2 d\sigma^{\text{int}, S_2 \text{ to SM}} + \\ & + (c_{t,1} + i\gamma_5 \tilde{c}_{t,1})^2 \times (c_{t,2} + i\gamma_5 \tilde{c}_{t,2})^2 d\sigma^{\text{int}, S_1 \text{ to } S_2}. \end{aligned} \quad (2.31)$$

Here, the physical states after loop-level mixing are denoted as S_1 and S_2 and the coupling modifiers are taken after the loop-level replacement discussed above. There are two interference terms for $S_{1/2}$ to SM $t\bar{t}$ background interference and one interference term for the signal-signal interference of S_1 to S_2 . It also has to be noted that this simplified CP violating model can be mapped upon the 2HDM models discussed earlier.

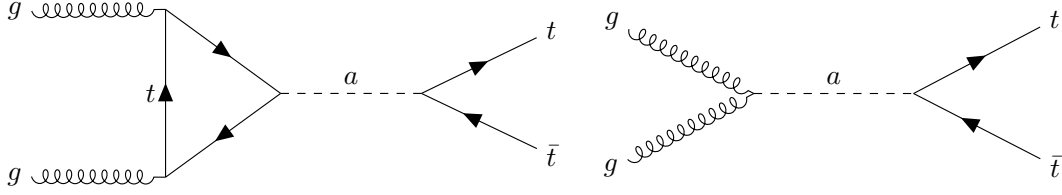


FIGURE 2.7: The considered production graphs for the $\text{ALP} \rightarrow t\bar{t}$ process. The left graph shows the case of the top-ALP coupling, while the right graph accounts for the direct ALP-gluon coupling.

2.4.2 Axion-Like Particles

Axions were devised as a solution to the strong CP which is described above in 2.3 [48, 49]. They are constructed based on the Peccei-Quinn mechanism [50, 51]. For this, a new symmetry $U(1)_{\text{PQ}}$ is introduced which is spontaneously broken at a scale f_a . A new pseudoscalar particle results from this, called axion (a) with a shift symmetry $a \rightarrow a + \kappa f_a$, where κ is an arbitrary parameter. The new particle has a Lagrangian that reads

$$\mathcal{L}_a = \frac{1}{2}(\partial_\mu a)(\partial^\mu a) + \frac{\alpha_S}{8\pi} \frac{a}{f_a} G_{\mu\nu} \tilde{G}^{\mu\nu} + \frac{m_a^2}{2} a^2 + \text{interaction terms} \quad (2.32)$$

with the color indices of the gluon field G omitted for readability. The $G\tilde{G}$ term indicates that a sum of $\mathcal{L}_a$ and $\mathcal{L}_{\text{QCD}}$ can be constructed that obeys

$$\frac{a}{f_a} + \theta = 0. \quad (2.33)$$

This would explain the missing CP violation in the QCD observations. At the same time, $\mathcal{L}_a$ is non-renormalizable, so it needs to be understood as an effective low-energy theory below the scale f_a . There are many suggestions for UV complete theories that have this effect [52]. These completions can also introduce the interaction terms indicated in equation 2.32. Using this mechanism to construct pseudoscalar particles that do not necessarily solve the strong CP problem is possible and the resulting models might still be physically well motivated, for example as a DM candidate. The particles in these more general models are called axion-like particles (ALPs).

The model that is considered in this thesis follows the derivation in Ref. [53]. They rewrite the interaction terms of the ALP to fermions as Yukawa-like interactions and only consider the top-ALP interaction and the explicit gluon term such that the Lagrangian follows as

$$\mathcal{L} = \mathcal{L}_{\text{SM}} + (\partial_\mu a)(\partial^\mu a) + \frac{m_a^2}{2} a^2 - c_{\tilde{G}} \frac{a}{f_a} G_{\mu\nu} \tilde{G}^{\mu\nu} + i c_t m_t \frac{a}{f_a} \bar{t} \gamma^5 t. \quad (2.34)$$

Note, that the resulting Lagrangian is similar to the pseudoscalar case $\Phi = A$ in 2.25 for $c_{\tilde{G}=0}$. For $c_{\tilde{G} \neq 0}$, an additional production case via direct coupling to gluons has to be considered. The $t\bar{t}$ production processes are shown in figure 2.7. The phenomenological study of this model [53] suggests that for some scenarios, this added direct gluon coupling could allow to distinguish the extended Higgs sector pseudoscalars from the ALP. It is thus an interesting process to study with experimental data. The results of this study are detailed in chapter 7.

THE EXPERIMENTAL SETUP

Matter signals to its lost voyagers, telling them that their quest is vain, and that their homeland already lies in ashes behind them.

— Nick Land

To access fundamental physics processes at the highest possible energies, collisions of protons with protons and heavier nuclei with heavy nuclei are generated by the Large Hadron Collider at four interaction points. This thesis works with data from the all-purpose "Compact Muon Solenoid" (CMS) detector. The following chapter gives an overview over the experimental setup of the LHC and the CMS detector, its subdetectors and basic analysis chain.

3.1 THE LARGE HADRON COLLIDER

The LHC is a hadron collider designed to collide protons and ions at the highest possible energies [54]. It is located near Geneva in Switzerland and France and is built into the 26.7 km long former LEP tunnel, between 45m and 170m underground. It makes use of superconducting magnets to keep protons up to 7 TeV and heavy ions such as lead and oxygen up to 2.7 TeV on the circular orbit. This results in a maximal center of mass energy of $\sqrt{s} = 14$ TeV. The electromagnet radio-frequency cavities that achieve these energies are also superconducting. The design capacity allows for 2800 proton bunches per beam with overall 3×10^{14} protons in the ring with a bunch crossing rate of 40 MHz. The design luminosity is $L = 10^{34} \text{cm}^{-2}\text{s}^{-1}$. The high-luminosity upgrade (HL-LHC), planned to start data-taking in 2028, will increase this peak instantaneous luminosity by a factor of 10. LHC delivers colliding beams to four experiments: ATLAS [55], CMS [56], LHCb [57] and ALICE [58]. While LHCb focuses on flavor physics with B-mesons and ALICE on heavy ion physics, ATLAS and CMS were built with general purpose detection principles in mind to test the SM in general and especially at the highest energies. A design target was the detection of the Higgs boson, which was achieved by ATLAS and CMS simultaneously in 2012 [1, 2]. Another design target was the detection of BSM physics in the TeV mass range, which still lies in the realm of speculation. As of 2025, LHC has gone through three runs of data-taking. Run 1 was conducted between 2010 and 2012 with $\sqrt{s} = 7$ TeV and $\sqrt{s} = 8$ TeV and yielded 29 fb^{-1} of data for ATLAS and CMS. Run 2 was ongoing between 2016 and 2018 at $\sqrt{s} = 13$ TeV and yielded 140 fb^{-1} . As of 2025, Run 3 is still ongoing. It was planned to take 300 fb^{-1} at $\sqrt{s} = 13.6$ TeV. Run 4 of the aforementioned HL-LHC era is projected to achieve an overall integrated luminosity of 3 ab^{-1} [59].

3.2 THE COMPACT MUON SOLENOID DETECTOR

The CMS detector [56] is designed around the superconducting solenoid magnet lending its name to the experiment. It generates a uniform 3.8 T magnetic field in the tracker and calorimeter components inside the magnet construction as well as in the muon system situated outside the magnet. The detector is designed as an hermetic experiment, with barrel detector components covering $|\eta| \leq 1.5$ and endcap subdetectors covering the very forward and very backward $|\eta|$ range. CMS consists makes use of an onion-shell design to characterize and measure different particle types in different subdetectors. The detector schematic is shown in figure 3.1.

3.2.1 Subdetectors

The following will detail the subdetectors from the inside to the outside of CMS.

Tracker

The innermost subdetector is the tracker. It is divided into a pixel tracker for the region closest to the beam pipe and a silicon strip tracker outside this region. Tracker design has to accomplish two-fold goals: Deliver accurate track and vertex finding while using a low material budget as to not attenuate the particles energy. The pixel tracker consists of three layers in the barrel and two endcaps. It is mainly designed to accurately reconstruct vertices which are used in particle identification of heavy flavor quarks and to discriminate against pile-up events. The silicon strip tracker is divided into four parts: the tracker inner barrel (TIB), tracker outer barrel (TOB), tracker inner discs (TID) and tracker endcap (TEC). Its main tasks is the reconstruction of particle tracks for momenta measurement. The tracker is almost hermetic, with only the very high $|\eta| > 2.4$ range being left out due to the very highly radiative environment there.

Electromagnetic Calorimeter

The electromagnetic calorimeter (ECAL) is equipped with scintillating lead tungstate crystals (PbWO_4) which allows to reconstruct the energy of electromagnetic cascades of photons and electron-positron pairs (showers). The ECAL is divided into a barrel section and an endcap ECAL which has an additional preshower module in front of

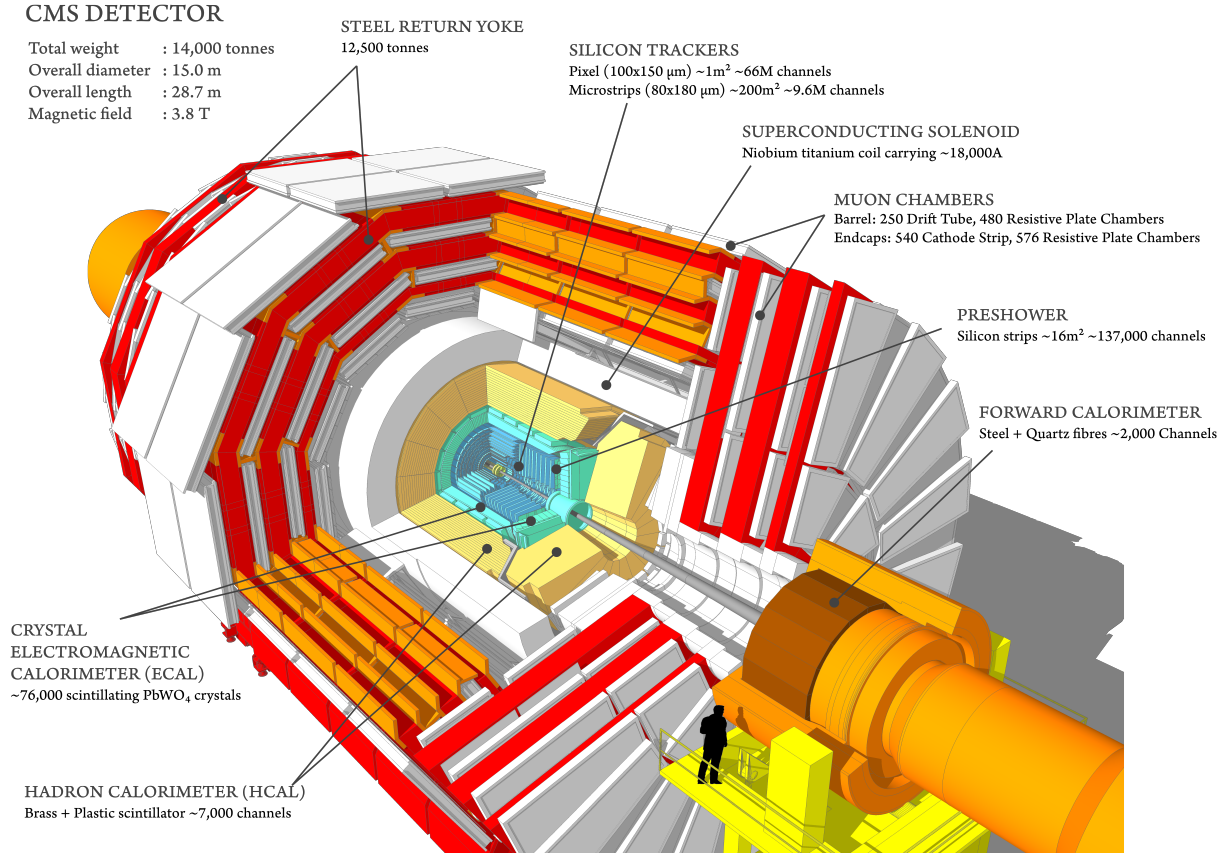


FIGURE 3.1: A schematic of the CMS detector. Taken from [60].

it. The preshower aids the detection of neutral pions which decay into two photons and would otherwise often be misreconstructed. The crystals are read out through photodetectors attached to their back. The readout logic allows to keep the energetic contributions of different bunch crossings separate by fitting lineshapes derived by calibration measurements. The short radiation length of the lead tungstate ($X_0 \approx 0.89 \text{ cm}$) for electrons and photons allows to fully contain most EW showers.

Hadronic Calorimeter

With the vast multiplicity of hadronic particles in proton-proton collisions, the hadronic calorimeter (HCAL) needs to deliver a fast and accurate energy measurement to enable full event reconstruction. This is achieved by a sampling calorimeter design. The HCAL consists of layers of steel and brass absorber plates with plastic scintillator tiles in between which are read out through photodetectors. The high material budget is necessary as hadrons are heavier than photons and electrons and have much more penetrating power. The CMS HCAL is divided into towers. These span the barrel (tower 1-16) and the forward region (tower 17-29). In the area $|\eta| < 1.26$ a further layer of scintillating material is placed outside the solenoid magnet to capture highly energetic jets. Additionally, a very forward ($3.0 < |\eta| < 5.0$) calorimeter is employed to cover the high jet activity that is expected here for proton-proton collisions.

Muon System

Outside of the solenoid magnet is the muon system. As muons penetrate even the HCAL without much energy loss, they are reconstructed by their tracks in the iron return yoke of the solenoid. Three different detection principles are used for the muon system. The drift tubes (DTs) in the barrel region ($|\eta| < 1.2$) provide the highest positional and directional sensitivity. In the forward region ($1.2 < |\eta| < 2.5$), cathode strip chambers (CSCs) are employed. These provide slightly less resolution but a higher readout rate to counter higher muon rates in the forward region. The full muon system is additionally equipped with fast-response resistive plate chambers (RPCs) which can be used to trigger on muon information. The muon system was designed with the "golden channel" of Higgs detection in mind ($H \rightarrow ZZ \rightarrow 4\mu$) such that CMS delivers a 100% trigger efficiency on these events [56].

3.2.2 Event Reconstruction

3.2.2.1 Trigger

The 40 MHz bunch crossing rate of the LHC machine is too large to process all events in real time. CMS employs a multi-level trigger system to only process a particularly interesting subset of events that keeps the data rate at $\mathcal{O}(1 \text{ GB/s})$ which equates to a 1 kHz event rate. The first trigger tier is the level 1 (L1) trigger which relies on trigger

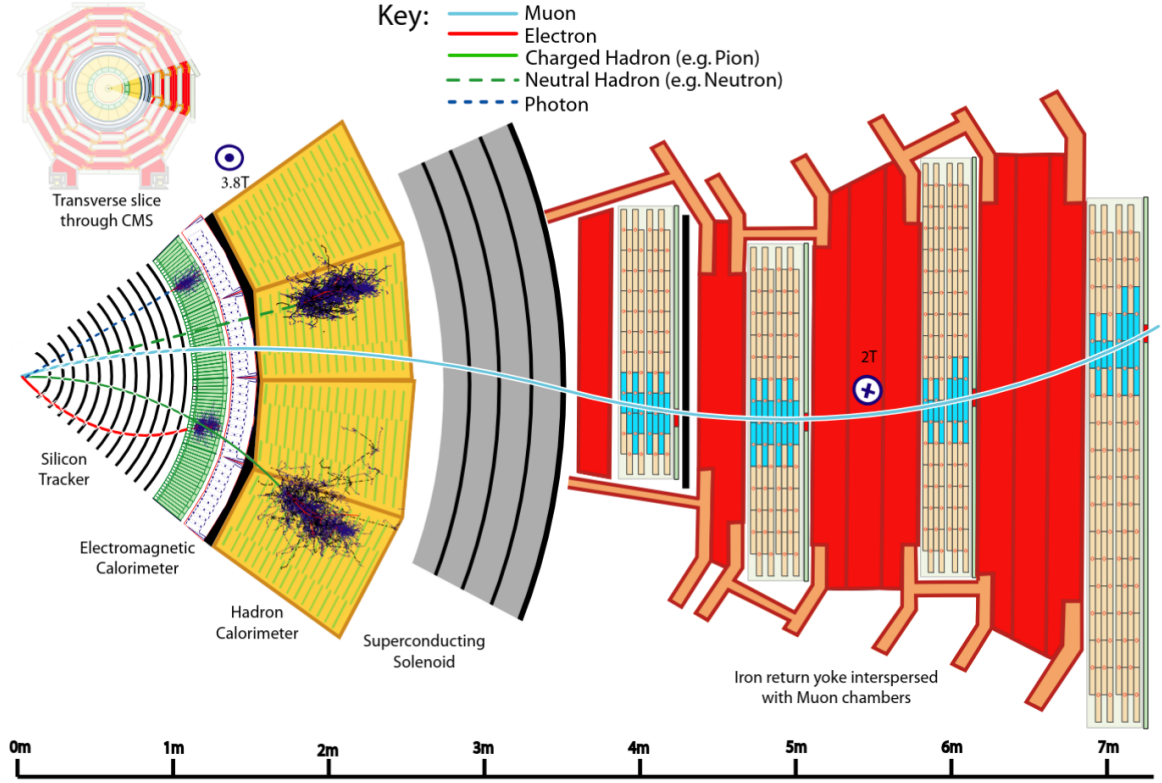


FIGURE 3.2: A slice of the CMS detector with a sketch of particle interactions in the different subdetector systems. Figure is taken from [61].

primitives sent from fast detector components. These are ECAL, HCAL and muon RPCs. The data is processed on field programmable gate arrays (FPGAs) and decisions are made based on simple logic, such as p_T thresholds for certain objects. The L1 triggered data already only has a rate of 100kHz. With an accept decision of the L1 trigger, the full detector is read out for the event. A CPU and GPU based high level trigger (HLT) system combines the full information into a decision logic for events that are particularly interesting for analysis. The decision process is based on a "trigger menu" with multiple possible selections that can apply to an event to accommodate for different analysis requirements. The 1kHz output stream of the HLT is saved on disk and is handed over for full reconstruction and subsequent offline analysis.

3.2.2.2 Particle Flow Event Reconstruction

The full CMS reconstruction based on the particle flow (PF) algorithm [61]. Figure 3.2 details common particle interactions in a slice of the CMS detector. Subdetector information is combined to identify particles and based on the identification, the particles energy and momentum are reconstructed. In detail, tracks are extrapolated into the calorimeters from the tracker. Muons are identified through their signature in the muon system and the tracker. Calorimeter clusters are assigned to tracks if they fall into the same region. A combined track and clusters are identified as a charged hadron if it is not a muon and not considered further in the reconstruction of the rest of the event. Electrons are corrected for Bremsstrahlung photon emission and their ECAL deposits are added up. Remaining calorimeter clusters are identified as photons (in the case of an ECAL energy deposit) and as neutral hadrons (for HCAL clusters).

The resulting physics objects are used to calculate particle momenta and energy. Jets are clustered from the physics objects reconstructed prior. An iterative anti- k_T algorithm is used to construct them [62]. The algorithm is infrared safe and the jet shape does not vary with soft radiative emission while being computationally efficient. Anti- k_T is instantiated with a jet radius of 0.4 in the $\eta - \phi$ plane to capture most of the partonic jet constituents while also being fine-grained enough to tell different origin jets apart.

In a hadron collider, the partons taking part in the original hard scattering cannot be predicted with certainty. As a result, the overall momentum conservation cannot be reconstructed. Since the transverse momentum vectors of any event should sum to 0 regardless, the p_T is used to approximately reconstruct particles which escape the detector (neutrinos). The missing transverse momentum (MET) is thus the negative vector of the transverse momentum vector sum of all PF candidates in an event. MET relies on the uncertainty of all other momentum measurements and is thus the most difficult quantity to reconstruct.

3.2.3 CMSSW Software Environment

The CMS SoftWare (CMSSW) [63] is the online- and offline software suite used by the CMS collaboration for many tasks. It is based on a wide range of software written in many different languages such as C++ steered by python configuration files.

The central experimental data is organized in the Event Data Model (EDM). The data is structured in tiers of ascending abstraction. The lowest level data tier is the RAW event format which includes detector information such as the L1 and HLT decision data. The AOD (Analysis Object Data) contains only high level objects intended for physics analysis. MINIAOD and NANOAOD contain even more streamlined, higher level information for efficient data analysis. The work in this thesis was mostly carried out on NANOAOD level data. A special case is the NANOGEN dataformat as it is structured just like the normal NANOAOD format but is only filled with information that is available from MC simulation excluding the detector simulation. FULLSIM simulated data on the other hand includes all the detector response including RAW information.

3.3 STATISTICAL INFERENCE ON EXPERIMENTAL DATA

The ultimate goal of any particle physics analysis is a statement about an underlying parameter in a physical model. This can be a measurement of a well established parameter of the SM, an exclusion statement on a BSM model parameter or an observation of an undiscovered process. This task is not trivial due to the fact that the experimentally measured quantities are random in nature from statistical uncertainties and from the inherent random nature of quantum physics. A statistically sound method of parameter extraction is thus needed. Standard procedure for the LHC experiments is outlined in Ref. [64]. This work deals mostly with binned profile maximum likelihood fits which are interpreted in a frequentist framework, where statistical inference is based on the distribution of data over repeated hypothetical experiments, with model parameters considered fixed but unknown.

3.3.1 Profile Likelihood Fits

To get an estimate for the parameters of interest (POIs) a profile likelihood fit is performed. To account for systematic uncertainties, nuisance parameters (NPs) are usually included in this fit. The fit can happen on binned information in histogram templates or on a continuous distribution. To perform such a fit, a likelihood model is needed. This likelihood function maps the parameters θ (NP and POIs) and a (simulated) observation X to the likelihood to obtain this result. It forms the physical language that the statistical analysis builds upon, all information about uncertainties, auxiliary measurements and the details of the (expected) signal are encoded into this model. The likelihood is a statistical measure that differs from probability density functions in the sense that they are not normalized to unity. For the likelihood one writes

$$\mathcal{L}(\theta) = P(X|\theta) \quad (3.1)$$

where $P(X|\theta)$ is the probability to obtain an observation X under the assumption of the true parameters θ . The idea of maximum-likelihood parameter estimation is that maximizing $\mathcal{L}(\theta)$ yields the most probable parameter set θ . Most implementations, in particular in the ROOT framework [65], achieve this by minimizing the negative log-likelihood

$$\text{NLL} = -2 \ln \mathcal{L}. \quad (3.2)$$

Minimization is done simultaneously for NPs and POIs. Constructing a likelihood is an arbitrarily complex problem. In the case of a binned likelihood, it is assumed that each bin is orthogonal to each other and can be treated as an independent counting experiment. The poisson distribution is used to estimate the likelihood of bin i to observed N_i events in the given bin

$$L_i(\theta) = \text{Poisson}(N_i | n_i(\theta)) \quad (3.3)$$

with the mean expected number of events $n_i(\theta)$. This mean expectation is given from the input template. The overall likelihood is given by the product of the per-bin likelihoods and the constraint term for the NPs (θ_{NP}). Here, we introduce the different treatment of POIs (θ_{POI}) and the NPs

$$L(\theta_{\text{POI}}, \theta_{\text{NP}}) = \prod_i L_i(\theta_{\text{POI}}, \theta_{\text{NP}}) G(\theta_{\text{NP}}). \quad (3.4)$$

The function G encodes prior knowledge on the NPs, for example from auxiliary measurements. A common constraint term could be coming from an experimental uncertainty u_1 that was measured to have a mean value $\hat{u}_1$ and a standard deviation σ_{u_1} . The function G is then proportional to $\mathcal{N}(\hat{u}_1, \sigma_{u_1})$. It is common to normalize the corresponding NPs to the standard deviation

$$\theta_{u_1} = (u_1 - \hat{u}_1) / \sigma_{u_1}, \quad (3.5)$$

so a one standard deviation effect is given by $\theta_{u_1} = \pm 1$. This also gives a handle of checking fit results for consistency. If a fit fixes a NP to a value that is in disagreement with an auxiliary measurement it shows in the so called pull plot. A pull is defined as the fitted shift from its nominal value, normalized by its prior uncertainty,

$$\text{pull} = \frac{\hat{\theta}_{\text{NP}} - \theta_{\text{NP}}}{\Delta \theta_{\text{NP}}}. \quad (3.6)$$

Pulls close to zero indicate consistency with the prior, while large pulls may signal tension between the model and the data or underestimated uncertainties. Another hint at a wrongly assigned uncertainty is a NP exhibiting less or more uncertainty than the prefit value. A severely constrained NP (showing a smaller uncertainty after the fit) might hint at the prior being chosen wrongly.

3.3.2 Statistical Hypothesis Testing

The goal of high energy physics analyses is often a statistical statement on the probability of obtaining a certain hypothesis. For this, tools from the statistical field of hypothesis testing are employed. For hypothesis testing, one needs to define at least two hypotheses. A null hypothesis, which is assumed to be true and an alternative hypothesis. For particle physics searches, the null hypothesis is usually also called the background-only hypothesis, in the special cases of BSM-searches, this coincides with the SM hypothesis.

To discriminate between the two hypotheses, a test statistic (t) is defined. This is a real valued variable that has different, known distributions for the different hypotheses. The test statistic can be chosen freely but the most powerful choice [66] is the (negative log-)likelihood ratio (LR) λ

$$\lambda(\theta_{\text{POI}}) = -2 \ln \frac{L(\theta_{\text{POI}}, \hat{\theta}_{\text{NP}})}{L(\hat{\theta}_{\text{POI}}, \hat{\theta}_{\text{NP}})}. \quad (3.7)$$

λ is a function only of the POIs as the nuisances in the numerator are profiled over to maximize the likelihood for the given POI value. The denominator is fixed to the overall maximum likelihood estimate. This defines the agreement of data to the POI value θ_{POI} . A low value of λ translates to good agreement and high values point at disagreement. The numerical values of λ follow a probability density $f(\lambda|\theta_{\text{POI}})$ which itself is a function of θ_{POI} . Actual physical data will determine an observed value λ^{obs} which can be used to exclude a set of POIs with a confidence level (CL) α

$$P(\lambda(\theta_{\text{POI}}) < \lambda^{\text{obs}}) = \int_0^{\lambda^{\text{obs}}} f(\lambda|\theta_{\text{POI}}) d\lambda \stackrel{!}{>} \alpha. \quad (3.8)$$

One can thus find a critical value $c_\alpha(\theta_{\text{POI}})$ with $P(\lambda \leq c_\alpha|\theta_{\text{POI}}) = \alpha$ and exclude all POI for which $\lambda^{\text{obs}} > c_\alpha$. This can be reformulated as a p-value

$$P(\lambda(\theta_{\text{POI}}) > \lambda^{\text{obs}}) = \underbrace{\int_{\lambda^{\text{obs}}}^{\infty} f(\lambda|\theta_{\text{POI}}) d\lambda}_{=p_{\theta_{\text{POI}}}} \stackrel{!}{<} 1 - \alpha. \quad (3.9)$$

A p-value encodes the probability to have an equal or more extreme outcome given the hypothesis. A small p-value is thus evidence against the hypothesis.

So called "toy-based" approaches evaluate the integral with a MC method. A set of toy distributions for a given hypothesis is sampled and compared to the observation. This is a method that is guaranteed to give correct coverage but it is generally computationally expensive.

Asymptotic methods rely on the observation of Wilk's theorem [67]. For large sample sizes approaching ∞ the $-\ln \lambda$ distribution will approach a chi-squared (χ^2) distribution under the null hypothesis. The χ^2 distribution will have degrees of freedom of the dimensionality of the signal model. Exclusion regions can thus be constructed from the observed test statistic alone and no further MC calculations are necessary. However, this asymptotic result relies on several regularity conditions: the null hypothesis must be nested within the alternative, the true parameter values must lie in the interior of the parameter space, the likelihood must be sufficiently smooth and differentiable, and the maximum likelihood estimators must be consistent and asymptotically normal. This can happen with non-linear signal models as quadratic coupling coefficients θ_{POI}^2 are always positive [68] or with NPs which are physically bounded.

3.3.2.1 Upper- and Two-sided Limits

To exclude regions of parameter space, where little or no signal is observed, upper limits on a parameter θ_{POI} are sought for which all $\theta_{\text{POI}} > \theta_{\text{POI}}^{\text{up}}$ are excluded. This is done with the CL_S criterion [69]. To exclude cases where lower θ_{POI} values are excluded because higher θ_{POI} fit the data better, the test statistic is redefined to have a lower bound at the best fit $\hat{\theta}_{\text{POI}}$

$$q(\theta_{\text{POI}}) = \begin{cases} \lambda(\theta_{\text{POI}}) & \text{for } \hat{\theta}_{\text{POI}} \leq \theta_{\text{POI}} \\ 0 & \text{for } \hat{\theta}_{\text{POI}} > \theta_{\text{POI}}. \end{cases} \quad (3.10)$$

based on q the CL_S value is defined as the ratio of signal+background over the background.

$$\text{CL}_S(\theta_{\text{POI}}) = \frac{p_{s+b}}{1 - p_b} \quad (3.11)$$

with the p-values for signal+background and background-only hypothesis. This p-value is defined over the space of the q-test statistic

$$p_{s+b} = \int_0^{q^{\text{obs}}} f(q|\theta_{\text{POI}})dq \quad (3.12)$$

and according for the background only p-value at $\theta_{\text{POI}} = 0$. Using the ratio of signal+background over background-only prevents small signal values from being excluded when the background-only hypothesis also badly fits the data. An exclusion of confidence α is claimed at $\text{CL}_S < \alpha$. This method ensures that exclusion regions do not overcover.

On the other hand, if establishing an observation, the background-only hypothesis is commonly excluded and the p-value is quoted. The null hypothesis is tested for and the p-value is compared to the standard normal distribution. A p-value of 0.05 is thus translated to $\approx 95\%$ under the normal distribution which is translated into 2 standard deviations. It is generally accepted in HEP to claim an observation at p-values that translate to five standard deviations.

This method of choosing between upper limits and observation carries the risk that the ad-hoc assumption can change the derived result drastically. A method to ensure the correct limits are drawn in any scenario is given by the Feldman-Cousins (FC) technique [70]. FC limits introduce an objective ordering criterion. For each parameter θ_{POI} , the acceptance region is constructed by including outcomes in order of decreasing likelihood-ratio (so increasing λ) until the desired confidence level is reached. While FC intervals are considered more rigorous from a statistical standpoint, they are computationally demanding, as they require ensemble tests for each parameter value.

The analysis in this thesis does not use the asymptotic method based on Wilks theorem for 2D limits. The impact of this and strategies to alleviate the computational burden of the FC method in this case are discussed in section 6.2.

Design no longer leads back towards a divine origin (...) Planning is the creationist symptom of underdesigned software circuits.

— Nick Land

Since particle physics experiments have seen digital data acquisition historically early, modern High Energy Physics is a data driven science from experimental side as well as from theoretical simulations. This fact leads to the availability of comparatively big datasets which are accompanied by similarly big simulated (and thus understood) datasets. Subsequently, HEP offers a wide variety of possibilities to deploy machine learning. Historically, HEP has seen ML algorithms such as Boosted Decision Trees (BDTs), Multi-Layer Perceptrons (MLPs) and recently has been at the edge of the general Deep Learning trend of the 2010s.

4.1 A SHORT INTRODUCTION TO MACHINE LEARNING

Machine Learning comprises all disciplines that make use of algorithms that optimize another algorithm ('model') for a certain task. This is commonly referred to as 'training' or 'fitting' a model. While culturally this is often synonymous with a notion of intelligence, one actually has to precisely distinguish between 'artificial intelligence' and ML.

Generally, one may classify ML algorithms into three separate categories:

1. supervised algorithms: Rely on labelled data and learn from feedback to any output the model predicts to a data point.
2. unsupervised algorithms: Deal with unlabelled data by learning similarities on the feature space.
3. reinforcement learning: Each training step, the model is either reinforced or punished with respect to a goal function.

Some algorithms fall between categories, for example semi-supervised algorithms.

4.1.1 Deep Neural Networks

With the rise in GPU accelerated computing power in the 2010s, Deep Neural Networks (DNNs) have manifested as a commonly used tool in ML applications [71]. Since they also play a big role in particle physics and especially in this thesis, the next chapter offers a more in-depth description of their structure and use.

For most purposes in HEP, it is helpful to think about neural networks as a numerically approximated fit function of the inputs and (many) model parameters θ . θ is found by minimizing a loss in the training process. The most important building block of DNNs is the artificial neuron (also called node). It is, mathematically speaking, a linear function with parameters that get tuned during training. A nonlinear activation function is linked to introduce the capability to model non-linear functions such that we may write

$$f(\vec{X}) = a\left(\sum_{j=1}^n w_j x_j + b\right) \quad (4.1)$$

where w_i are the trainable components of the weight matrix, b is a trainable bias and $a(x)$ is the aforementioned non-linear, differentiable activation function. Popular examples for activation functions are rectified linear unit functions (ReLU) [72], sigmoid and tangens hyperbolicus. These nodes are used in layers of different broadness (more nodes in parallel) and varying amounts of these layers can be stacked upon each other ('depth') to form a 'multilayer perceptron'. One does refer to the first layer as an input layer, the last one as an output layer and intermediate layers are 'hidden layers'. Typically, a artificial neural network is considered 'deep' when there are multiple hidden layers. One key element of the DNNs success is the Universal Approximation Theorem. It states that one may approximate any function, that one can sample from, by a sufficiently capable neural network [73]. This is not a promise of actually finding the DNN configuration that solves the problem but gives a strong motivation to use a DNN in many cases, explaining their widespread usage.

4.1.2 Loss Functions

To actually train a DNN algorithm, backpropagation is used. Since equation 4.1 is differentiable, one may use gradient based optimization methods with DNNs. A differentiable goal function (loss function) is formulated

suitable for the task at hand. Here, training a ML algorithm can differ from classical optimization: The goal of the model is to perform well on a test set with respect to a metric that might be intractable. During training, one minimizes a surrogate target function - the loss - on the training set. It is thus important to choose a loss function that instills the correct behavior into the model.

The following will give two examples for the most popular choices and their context in DNN training. In addition to the aforementioned supervised or unsupervised targets, it is also necessary to distinguish classification and regression tasks. Classification tasks have a discrete target space, the goal is to group each sample to a class. A popular loss function for this task regime is the (binary/categorical) Cross Entropy $\mathcal{L}_{CE}$ (BCE for binary classification, CCE for multiclass classification). It is given by the sum over the K classes of the product of target value $y(\mathbf{x}_i)$ and the logarithm of the output probability $p(\mathbf{x}_i)$,

$$\mathcal{L}_{CE} = - \sum_{i \in \mathcal{X}} \sum_{k \in K} y_k(\mathbf{x}_i) \log(p_k(\mathbf{x}_i)). \quad (4.2)$$

The second summation runs over the samples in the set $\mathcal{X}$ that is used for loss estimation. In the discrete case, this is equivalent to the sum of the Shannon Entropy and the Kullback-Leibler divergence. Minimizing the Cross Entropy simultaneously maximizes the logarithmic likelihood for an estimator that outputs $p_k(\mathbf{x}_i)$ [71].

For regression tasks, where one optimizes the model output with respect to a continuous target space, the mean-squared error is a popular choice of loss function:

$$\mathcal{L}_{MSE} = \frac{1}{N} \sum_{i \in \mathcal{X}} (y(\mathbf{x}_i) - p(\mathbf{x}_i))^2 \quad (4.3)$$

where y are the targets and p are the model outputs. This generalizes to vector targets $\mathbf{Y}$ and vector outputs $\mathbf{P}$. The MSE is a distance based loss, while it is convenient to implement and often yields good optimization performance, its heavy weighting of outliers can be a problem that leads to suboptimal performance. The shown loss function examples, while popular, are not a general solution to these problems and choosing a best-performing loss function is an important factor in successful DNN training.

Losses are often computed on mini-batches to aid with computational performance and to regularize by having a noisy subsample for each DNN update. For this, a sampled subset of the training data is passed through the network and the loss derivative is computed. Then, the network weights are updated and the procedure is repeated for the next batch.

4.1.3 Weighted Training Samples

Introducing sample weights can be beneficial for DNN performance to overcome unbalanced data or to carry over a (physical) information that is stored in the sample weight of the training set. Unbalanced data will be detrimental to the training performance on the minority class, as the training loss minimization will be disfavored against the majority class. The minority class in particle physics is often the analysis target as rare processes are the most interesting to study. A maximal performance in classifying the minority class is crucial. These problems are commonly referred to as ‘needle in the haystack’ problems. Balancing the weight between minority and majority class will lead to a balanced loss that minimizes the error on both classes equally. Commonly, these weights are applied by multiplying the batch-loss with the average sample weight of the batch. Since the gradient operator is linear, this will result in an importance weighting of the gradient, resulting in the desired weighted DNN parameter update.

4.1.4 Partial Label Learning

Chapters 9 and 8 are employing a DNN method that is based on Partial Labels. Partial Label Learning (PLL) is defined as the task of training a ML algorithm on data points that each have labels from potentially multiple classes, only one of which is correct [74]. This places it in the group of methods based on multilabelled data. Multilabelled data also comprises data which naturally has multiple correct labels and partially multilabelled data, where a subset of the labels for each instance in training data is correct [75]. Figure 4.1 gives a graphical overview over the different types of multilabels, also contrasting it to the uniquely labelled data that has been discussed as the standard case. Multilabels are often easier to come by in cases where labelling is an expensive task. PLL especially has a usecase in image recognition scenarios where multiple objects are in the image. This extends to bioinformatics, web mining and facial recognition [76]. A common strategy for learning from partial labels is to disambiguate the set of candidate labels either by using averaged modeling outputs of the different label set entries (average-based approaches) [74] or by regarding the ground truth label as an intermediate variable which needs to be identified via an iterative algorithm (identification-based approaches) [77]. Thus, both types of approaches force the output distribution into a certain prediction, either to a uniform distribution of probabilities for the possible labels or to a concentration in one single label. However, for the work regarding negative weighted data points, it is interesting to consider third type of PLL-approaches, which is called disambiguation-free [76, 78, 79] as they don’t explicitly force the output into a certain distribution as long as the probabilities concentrate on the allowed labels.

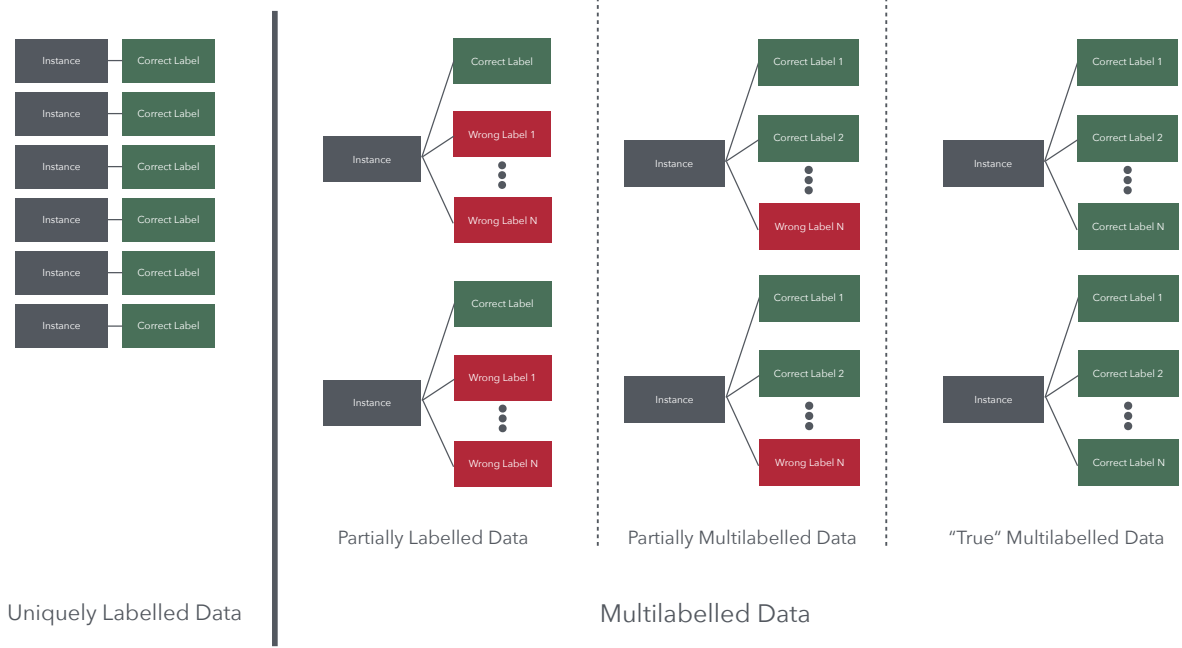


FIGURE 4.1: Schematic overview over partially labelled data in contrast to other multilabelled training data and uniquely labelled data.

A particular focus in this work is placed on the risk estimator proposed by Feng et al. [76], due to its simplicity and applicability to arbitrary classification networks. The risk estimator is defined as

$$\hat{R}_{CC} = -\frac{1}{n} \sum_i \log \left(\frac{1}{2^{2k-1} - 1} \sum_{y \in Y_i} \frac{\exp(f_y(\mathbf{x}_i))}{\sum_j \exp(f_j(\mathbf{x}_i))} \right). \quad (4.4)$$

Here, n denotes the dataset length (or batch length in batched training), k is the number of classes, Y_i is the candidate label set for the data point i and $f_y(\mathbf{x}_i)$ is the logit output of the training model, e.g. a DNN. The last ratio of exponentials takes the function of the Softmax function. The sum in the denominator runs over $j \in Y_i$ as a normalization. This risk estimator can be adapted into a loss function, which shall be referred to as classifier consistent loss (CCL). The loss function is simplified by leaving out the first factor in the logarithm because it does not contribute in differentiation. The risk estimator (and subsequently the loss) is proven to be classifier consistent [76], which means that the classifier optimized by this method converges to the argmin of the risk.

4.1.5 DNN Architectures

The simple multilayer perceptron architecture that has been described in 4.1.1 is commonly referred to as "fully connected" or "densely connected". With the same building blocks of artificial neurons one can realize other architectures with sparser or specialized connection layouts. Optimizing the network architecture to suit a problem is a mighty tool to overcome complex problems.

Convolutional Neural Networks (CNNs) use a DNN architecture that relies on the convolution operation instead of the matrix multiplication of dense layers [80]. Data that is known to be grid-like can be easier captured by convolutions. For a convolution in the ML sense a set of learned kernel matrices (or "filters") is moved across the input vector and the dot product is calculated at each step. This has the advantage, that significantly fewer parameters need to be learned, as they are shared across the whole input vector. In addition, spacial features that span across a few datapoints inside the input vector are easier learned and when one builds a model of subsequent convolutional layers, one can benefit from hierarchical feature extraction. Convolutional layers are often paired with pooling layers, which reduce the dimensionality. Popular forms are max-pooling and average-pooling [81]. They serve as a form of regularization, keep translation equivariance and reduce computational cost.

Residual Networks (ResNet) [82] employ another popular DNN architecture originally perceived for learning image data. They were introduced to address the degradation problem in very deep architectures, where increasing the network depth leads to higher training error. Instead of requiring a stack of layers to directly learn a desired underlying mapping $\mathcal{H}(x)$, ResNets are designed to learn a residual mapping $\mathcal{F}(x) := \mathcal{H}(x) - x$. The output of a residual block is then computed as:

$$y = \sigma(\mathcal{F}(x, \{W_i\})) + x \quad (4.5)$$

where σ is an activation function. This architecture utilizes "skip connections" (or shortcut connections) that allow the gradient to flow through more effectively during backpropagation, mitigating the vanishing gradient problem.

Graph Neural Networks (GNNs) are another form of DNN architecture that relies on structure in the data to leverage sparsity. GNNs operate on graph structured data. Graphs are built from nodes and edges and allow for irregular shapes. The idea of GNNs is that the output of a layer is computed from the data on the concerned node itself and from the aggregation of information of neighbouring edges and nodes. Thus, information of local and global contexts in the graph can be captured without having to compute a sparse matrix representation of the data. **Transformer Networks** or short transformers are an architecture based on the self-attention mechanism that has become state-of-the-art for tasks with sequential and contextual data [83]. The output of an attention layer is computed by the matrix product of a query (Q), key (K) and value (V) vector following

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (4.6)$$

query, key and value are all linear embeddings of the input to the attention layer x_i . Here, the softmax, a function to convert a set of numbers into probabilities that sum to 1, of query and key scaled by the dimensionality d_k gives the attention weight that is then applied to the value. The result is scalable and captures importance and interaction of features. A transformer network is build from a embedding layer that is then followed by (multiple) attention and feedforward layers. Commonly these are used in an encoder-decoder style. A final un-embedding layer yields the output result in the target space. Common encodings are tokenized words (for language processing).

Physics Informed Network Design Since the data collected in collider experiments has many underlying theoretical constraints, it is often advantageous to construct a network with this structure in mind. This also benefits the explainability because the physics assumptions can be tested based on the performance of the network [84]. A core concept of physics structure is symmetry and thus equivariant model architectures are often used in physics informed ML applications. Equivariance is the notion that the output of a model $f(x)$ should remain the same if one acts with a symmetry transformation g on the inputs or on the outputs of the model

$$f(g \cdot x) = g \cdot f(x). \quad (4.7)$$

Among many applications, two such network designs shall be highlighted here, as they have been tested in the scope of this work.

Particle Flow Networks (PFN) [85] are based on Deep Sets [86] which allow to learn from point clouds datastructures. The idea is to approximate observables from a function of the sum of mappings from the particle features p_i ,

$$PFN = F\left(\sum_{i=1}^M \Phi(p_i)\right). \quad (4.8)$$

where F is a global DNN acting on the sum of the outputs of $\Phi(p_i)$ DNNs computed from the particle features. This imposes equivariance on the order of the inputs. It does not require any other symmetries and can thus flexibly be used for event classification, jet tagging or other tasks.

A more rigid idea of physics symmetry informed DNNs is implemented in the Lorentz Equivariant Geometric Algebra Transformer (LGATR) [87, 88]. It is based on the Geometric Algebra Transformer [89] and specifies the implementation of any geometric algebra to the underlying geometric algebra of particle physics - the Lorentz algebra. In practice, this means that the computations of the transformer and the input and hidden objects are replaced with Lorentz geometric objects. The elements of the algebra are multivectors that can be embedded into the $\mathbb{R}^{16}$ space. These multivectors are constructed from sums of four real vectors $\gamma^\mu \in \mathbb{R}^4$ and their antisymmetric products of multiple of these vectors. The degrees of freedom of a multivector are thus a scalar x^S , a pseudoscalar x^P , a vector x_μ^V , an axialvector x_μ^A and a bivector $x_{\mu\nu}^B$

$$x = \begin{pmatrix} x^S \\ x_\mu^V \\ x_{\mu\nu}^B \\ x_\mu^A \\ x^P \end{pmatrix}. \quad (4.9)$$

A particle can be represented as a multivector with a scalar component (x^S = particle ID) and a momentum fourvector $x_\mu^V = p_\mu$ and all other components set to 0. To construct a transformer that is Lorentz-equivariant and operates on these objects, one has to replace the usual layer operations with equivariant counterparts. This is detailed in [87]. The final LGATR network is then constructed similar to a transformer encoder, with multiple blocks of LGATR-attention layers stacked, interleaved with LGATR-dense layers. It can be used for any classification or regression task where the symmetric structure is advantageous. The authors provide performance proofs for regression of QFT amplitudes, jet tagging and event generation tasks [87].

4.2 MACHINE LEARNING AT THE LHC

The LHC environment enables unprecedented data set size and its start of operation coincided with the rise of modern ML applications. Naturally, many tasks have been optimized with ML early on. This chapter will focus on the experimental usage of ML techniques that are explored in the scope of this thesis. Further literature also covers accelerator operational aspects [90], theory developments and experimental usages [91]. The work in chapter 8 focuses on event classification with ML discriminators. Other ML work that is explored in this project is associated to Simulation Based Inference (SBI) and event reconstruction.

4.2.1 Event Classification

The advantage of using ML models in event classification is straight forward: Sensitivity to desired processes can be enhanced by combining the discrimination power of multiple high-level observables in a ML model. Other research even suggests that forming images of the energy deposits can yield an increase in performance [92]. The basic use case of event classification for LHC applications is supervised training on MC data. With MC simulations for most physics processes large datasets of labeled data are obtainable. Apart from boosted decision trees (BDTs) that have been used in these applications for decades [93], DNNs are successful in these tasks. As described above, the binary Cross Entropy is the usual choice of loss function as it optimizes for the logarithmic likelihood of correctly categorizing a sample. This fact has further relevance for particle physics as using a Sigmoid (for binary classification) or SoftMax (for multiclass classification) activation function in the output layer allows the DNN outputs to be interpreted as probabilities. A sufficiently trained DNN thus approximates the probability distribution for signal in MC data [91]:

$$p_{\text{MC data},S} = p_{\text{model},S} \quad (4.10)$$

This is a powerful finding as it allows to interpret the network output as the expectation value for signal and background probability of a given datapoint.

The main challenge in LHC application is the training data. While MC simulation data is labeled, abundantly available and physically well motivated, precision and biases in MC over real data remain a concern. There are weakly supervised and unsupervised methods that alleviate this problem by allowing to train on experimental data directly. Training without labeled data is a structurally more difficult task, so these models commonly perform worse compared to supervised classification. These methods are commonly used in anomaly detection where the signal is unknown [94–96].

4.2.1.1 The Likelihood Ratio Trick

Using the BCE loss (as in equation 4.2) for classification has a further benefit. If one assumes a global minimum, analytical calculation yields:

$$\frac{\partial \mathcal{L}}{\partial \theta} \stackrel{!}{=} 0 \quad (4.11)$$

$$0 \stackrel{!}{=} y(\mathbf{x}_i) \frac{\partial \mathcal{L}}{\partial \theta} \log(p(\mathbf{x}_i)) + (1 - y(\mathbf{x}_i)) \frac{\partial \mathcal{L}}{\partial \theta} \log(1 - p(\mathbf{x}_i)) \quad (4.12)$$

which, after application of the chain rule and some algebraic simplification becomes

$$\frac{p(\mathbf{x}_i)}{1 - p(\mathbf{x}_i)} = \frac{y(\mathbf{x}_i)}{1 - y(\mathbf{x}_i)} \quad (4.13)$$

Identifying $y(\mathbf{x}_i)$ as the likelihood of a data point belonging to the class 1 and $1 - y(\mathbf{x}_i)$ being the likelihood of a data point belonging to class 0, we see that a simple function of the classifier output $p(\mathbf{x}_i)$ will give the likelihood ratio between the two classes. This is a reformulated finding of the fact, that minimization with respect to the Cross Entropy is equivalent to the maximization of the log-likelihood. In practice, it is often enough to have a classifier that approaches the global minimum well enough. Likelihood ratios can then be used directly as a test statistic for parameter inference or in other applications such as reweighting of datasets [97].

4.2.2 Event Reconstruction

Event reconstruction describes the task to infer the kinematic information of the constituents of the hard scattering event from the decay products that the detector measures. Depending on the variables and the process that one reconstructs, this is posed as a one or multidimensional regression tasks. The differing final states pose different challenges. If there is a large multiplicity of jets or ambiguity in the objects origin, the task of matching the objects to the original particles is the biggest challenge. It can thus be seen as a permutation task. On the other hand, final states with many neutrinos will have a lot of missing information. Here, it is important to predict the distribution of the initial particles well despite the ambiguous MET. Detector effect are also always a concern.

This thesis deals with variables derived from top quark pair production events, so reconstructing the top system is important. There are examples of exploratory ML studies on top-antitop reconstruction in literature [98–100], but

due to the need of an exact understanding of induced biases and out of distribution behavior, the analysis in chapter 7 only considers an analytical approach, which is detailed in section 6.1.1.5. Chapter 6.4 provides a preliminary study on the feasibility of using the LGATR architecture for the reconstruction of dileptonically decayed top quarks.

4.2.3 Simulation Based Inference

Chapter 5 illustrates how mechanistic feedforward models are used to generate predictions for well motivated but analytically intractable theories such as QCD. To infer fundamental parameters of that theory from observations, the inverse problem must be solved: An analytically intractable likelihood of the underlying parameters is to be inferred. A ML driven approach to this problem is given by the field of simulation based inference (SBI). Here, the classical construction of summary statistics from (binned) observables is omitted for direct inference on a learned likelihood model that was trained on the simulation output. This method has been adopted in many areas of science and engineering as high fidelity simulations have become commonplace. A detailed overview of the field is given in Ref. [101].

The work here focuses on inferring QFT parameters from Monte Carlo simulations with the MadMiner framework [102]. The statistical analysis chain for standard HEP analysis is detailed in section 3.3. Here, a likelihood is constructed based on the poisson distribution using a limited number of binned high level search variables. The likelihood is then used for inference. SBI keeps the inference procedure intact but proposes a direct way to constructing the likelihood from the MC output. MadMiner adds the idea of augmented information [103] for improved learning of the LR. This is based on the observation, that the event probability density for an outcome in a variable x under the assumption of theory parameter θ factorizes into (hidden) intermediate particle momenta z_{hidden} along the MC chain

$$p(x|\theta) = \int dz_{\text{detector}} \int dz_{\text{shower}} \int dz p(x|z_{\text{detector}}) p(z_{\text{detector}}|z_{\text{shower}}) p(z|\theta). \quad (4.14)$$

The lowest hidden state is given by the ME level parton momenta z . Two useful variables are the joint likelihood ratio $r(x, z|\theta_0, \theta_1)$ and the joint score $t(x, z|\theta_0)$. The joint likelihood ratio is defined as the per event likelihood ratio for ME level momenta between a theory parameter θ_0 and a reference theory parameter θ_1

$$r(x, z|\theta_0, \theta_1) = \frac{p(z|\theta_0)}{p(z|\theta_1)}. \quad (4.15)$$

The joint score is the per event relative gradient of the likelihood with respect to the theory parameter

$$t(x, z|\theta_0) = \frac{\nabla_{\theta} p(z, \theta)}{p(z, \theta)} \Big|_{\theta_0}. \quad (4.16)$$

The joint score as well as the joint likelihood ratio can be extracted at ME level from a ME generator for each event. MadMiner delivers an implementation for MG5_aMC@NLO but the method is in theory also applicable to other generator codes such as Powheg. Following Ref. [103], regression on the joint ME level quantities allows to train an estimator for the intractable observable-level likelihood ratio and score.

There are multiple ways to construct such a regression.

The *SCore-Augmented Neural Density Approximates Likelihood* (SCANDAL) method constructs a parametrized neural density estimator for the likelihood $\hat{p}(x|\theta)$. The neural model can be freely chosen as long as it is differentiable with respect to θ . It regresses on the loss

$$\mathcal{L}_{\text{SCANDAL}} = \mathcal{L}_{\text{MLE}} + \alpha \frac{1}{N} \sum_i |t(x_i, z_i) - \nabla_{\theta} \log \hat{p}(x)|^2. \quad (4.17)$$

Here, $\mathcal{L}_{\text{MLE}}$ denotes any maximum likelihood loss between $\hat{p}$ and the joint likelihood $p(x_i|\theta)$. α is a hyperparameter steering the impact of the additional loss term and N denotes the number of events in the batch. SCANDAL gives a direct estimate of the likelihood and not its ratio.

The *Approximate Likelihood with Improved Cross-Entropy estimator and Score* (ALICES) method [104] is based on the likelihood ratio trick for DNN discriminators introduced in section 4.2.1.1. Instead of discriminating between integer labels $y_i \in \{0, 1\}$ the joint likelihood ratio of signal θ_1 over background θ_0 + signal θ_1 is used as an augmented information. The target function is

$$\mathcal{L}_{\text{ALICES}} = -\frac{1}{N} \sum_i [s(x_i, z_i|\theta_0, \theta_1) \log \hat{s}(x_i) + (1 - s(x_i, z_i|\theta_0, \theta_1)) \log(1 - \hat{s}(x_i))] + \mathcal{L}_{\text{score}} \quad (4.18)$$

with the function $s(x_i, z_i|\theta_0, \theta_1)$

$$s(x_i, z_i|\theta_0, \theta_1) = \frac{p(x_i, z_i|\theta_1)}{p(x_i, z_i|\theta_1) + p(x_i, z_i|\theta_0)}. \quad (4.19)$$

This quantity is extractable at ME-level and improves the performance of the ALICES method with respect to standard BCE based discrimination. The added score loss term is defined to be

$$\mathcal{L}_{\text{score}} = \alpha(1 - y_i) \left| t(x_i, z_i | \theta_0, \theta_1) - \nabla_{\theta} \log \left(\frac{1 - \hat{s}(x_i | \theta, \theta_1)}{\hat{s}(x_i | \theta, \theta_1)} \right) \right|_{\theta_0}^2. \quad (4.20)$$

This can be seen as a regulatory term that ensures the convergence of the ALICES estimator towards correct local tangential behavior of the likelihood function. Again, any DNN which is smooth with respect to θ parametrization is a valid choice for an ALICES estimator.

The *Score Approximates Likelihood Locally* (SALLY) method uses the fact that a likelihood can be approximated to linear order around a parameter point θ_{ref} as

$$p_{\text{local}}(x|\theta) = \frac{1}{Z(\theta)} p(t(x|\theta_{\text{ref}})|\theta_{\text{ref}}) \exp[t(x|\theta_{\text{ref}}) \cdot (\theta - \theta_{\text{ref}})]. \quad (4.21)$$

For SALLY, the score $t(x|\theta_{\text{ref}})$ is estimated by minimizing the mean squared error between the joint score and the DNN estimator $\hat{t}(x|\theta_{\text{ref}})$. To accommodate for the effect of the normalization of the constant $Z(\theta)$, any multivariate density estimation technique can be used. Due to the local nature of this approach, its scaling to high-dimensional parameter spaces is superior to other methods. However, the predicted limits will be worse, the further away from the reference point the SALLY method is used. *Score Approximates Likelihood Locally IN One dimension* (SALLINO) is a method building on this idea but absorbing the $\theta_0 - \theta_1$ dependence into one dimension to offer an even better scaling with the parameter space dimensionality.

In addition to the inference of limits, the DNNs trained here share the feature of differentiability. This allows to extract the relative information content of high-level observables x via the Fisher information matrix [105]

$$I_{ij} = \mathbb{E} \left[\frac{\partial \log p_{\text{full}}(\{x\}|\theta)}{\partial \theta_i} \frac{\partial \log p_{\text{full}}(\{x\}|\theta)}{\partial \theta_j} \Big|_{\theta} \right]. \quad (4.22)$$

The minimal covariance of an estimator $\hat{\theta}$ is related to the inverse Fisher matrix according to the Cramér-Rao bound and thus the maximal precision level on a parameter can be inferred directly [106, 107]. For example, the minimal error on a parameter in a one-dimensional case is $\Delta\theta \geq 1/\sqrt{\mathcal{I}}$. The Fisher information matrix can be approximated by a score estimator and thus we can use the MadMiner framework to gain insight into the usefulness of a particular variable in a future analysis.

The MadMiner publications present their work in the context of effective field theories, where highly dimensional theory parameter spaces and limited MC sample size are the constraining issues for inference of physical limits. All of these problems are also encountered in the case of CP violating heavy Higgs bosons in chapter 6.3.1 where the MadMiner framework sees use for a phenomenological study of the parameter space.

MODELLING OF PARTICLE COLLISIONS - MONTE CARLO EVENT GENERATION FOR EXPERIMENTS

Not to withdraw from the process, but to go further, to "accelerate the process," as Nietzsche put it: in this matter, the truth is that we haven't seen anything yet.

— Gilles Deleuze

In modern particle physics, experimental results are often derived from comparisons of distributions of events rather than single events to predictions. The underlying theory is described in Chapter 2. This approach enables the estimation of probabilities for the occurrence of processes but the framework of QFT does not allow for analytical prediction of events without accounting for the statistical nature of the quantum world. To generate simulated events suitable for comparison with experimental data, dedicated algorithms are required that can handle the highly complex and multi-scale nature of high-energy collisions. A central challenge arises from the vast range of energy and length scales involved, spanning the initial short-distance quantum interaction to the macroscopic detector response across the full volume of, for example, the CMS detector. In modern high-energy physics analyses, Monte Carlo (MC) event generators are used to model the hard scattering process, parton shower algorithms simulate additional QCD and QED radiation and hadronization effects, and detailed detector simulations reproduce the detector response. The complete simulation chain produces fully reconstructed events that can be directly compared to experimental data. The Monte-Carlo work presented in this chapter was mostly conducted as service work of the author to gain authorship status with the CMS experiment.

5.1 EVENT SIMULATION AT HADRON COLLIDERS

The LHC collides hadrons at energy scales of multiple TeV. Hadrons have a rich inner structure due to the confining nature of QCD which is analytically intractable. For hard scattering processes of two protons into a final state X , where the transverse momentum scale of X is large compared to the QCD scale, the QCD factorization theorem is valid [108]. It allows to factorize the cross section into a hard term $\hat{\sigma}$ and into the parton density function (PDF) $f_{\text{part}}(x)$. The total cross section is given by

$$\sigma(pp \rightarrow X) = \int_0^1 dx_1 \int_0^1 dx_2 \sum_{\text{part}_1, \text{part}_2} f_{\text{part}_1}(x_1) f_{\text{part}_2}(x_2) \hat{\sigma}(\text{part}_1(x_1 p_{\text{proton}}) + \text{part}_2(x_2 p_{\text{proton}}) \rightarrow X). \quad (5.1)$$

Here, the sums runs over all possible combinations of partons part_1 and part_2 . Their momentum fractions x_1 and x_2 are integrated over and the proton momentum p_{proton} is given by the accelerator.

The theorem allows the hard process to be described in perturbation theory and the inner structure of the hadrons is modelled statistically via the PDF. In practice, PDFs are fitted to data as for the QCD bound partons no analytical calculation is available. It also has to be noted that the PDFs are dependent on the chosen factorization scale μ_F where the hard, perturbative and the soft, non-perturbative part are separated. This choice is not unique and it is often handled as an uncertainty in calculations. For top quark pair production the top mass can be an adequate choice $\mu_F = m_t$ or dynamical per event choices are taken into consideration such as $\mu_F = \frac{H_T}{2}$ where H_T is defined as $H_T = \sum_i \sqrt{m_i^2 + p_{T,i}^2}$ with the index i running over all the events constituents.

5.2 MC EVENT GENERATORS

The "hard scattering" in a collider event accounts for the most energetic process between partons in a collision. Its matrix element (ME) can therefore be computed in perturbative QCD (pQCD) ¹, as the interaction is very short range. In MC simulations, hard scattering is handled separately from the underlying event, which includes beam remnants and softer parton-parton interactions. MC event generators compute the ME of the process up to a given order of pQCD and sample from the resulting phase space. It is generally infeasible to calculate the ME densely in the full phase space, so generator codes typically divide the phase space into a grid. A refinement algorithm usually allows for computationally efficient calculation across these grids.

While fixed-order pQCD calculations provide high accuracy for the hard process, they fail to describe the complex structure of the hadronic final states observed in detectors. Conversely, PS MC codes provide a realistic description of high-multiplicity final states (e.g. from NLO corrections or the emission of additional partons at the same order) via soft and collinear approximations but need a ME generator to cover the hard process generation. To bridge this gap, "matching" algorithms were developed to combine ME generators and PS. The primary challenge in this combination is double counting: both the ME real emissions (in NLO or NNLO pQCD ME generators) corrections and the PS radiating the first gluon can cover the same phase space region.

¹ There are also electroweak processes with an electroweak perturbative ME. The text continues to discuss the QCD case.

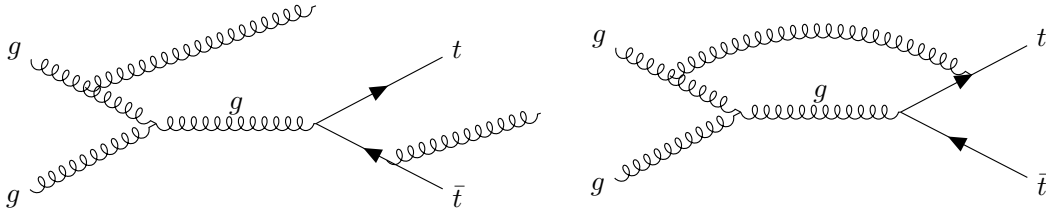


FIGURE 5.1: Two example Feynman graphs of second order corrections to $t\bar{t}$ production. The left graph show initial and final state radiation. The right graph exemplifies the scenario of a gluon exchange between initial and final state.

5.2.1 The POWHEG BOX

The POWHEG BOX [109] is a collection of algorithms to implement (N)NLO ME calculations into showering building on the POSITIVE WEIGHT HARDEST EMISSION GENERATOR (POWHEG) algorithm. The core idea of this method is to generate the hardest radiation emission first, using the exact NLO matrix element, and then interface this with a parton shower that is vetoed to prevent it from generating radiation harder than the one already generated. This is one method to enable NLO accuracy in ME+PS calculation. At NLO, POWHEG generates no negatively weighted events, which differentiates it from other methods for NLO ME+PS generation.

5.2.1.1 The hvq Generator

The hvq (heavy quark) package [110], implemented within the POWHEG BOX v2 framework, represents the standard tool for simulating top quark pair production at NLO in QCD. hvq relies on the narrow width approximation. In this approximation, the top quark is treated as a stable particle with a fixed mass and zero width during the computation of the hard scattering matrix element. Consequently, the process is generated effectively as a $2 \rightarrow 2$ scattering process ($pp \rightarrow t\bar{t}$). hvq generates the hardest radiation emission either from the initial state or from the final state using the exact NLO matrix element according to the POWHEG method. However, because the top quarks are treated as stable in the production step, FSR is restricted to emissions off the top quark prior to its decay. Radiation from the decay products (e.g., the b quark) is typically handled by the PS.

Since the top quark is unstable in the Standard Model, its decay is handled in a subsequent step after the production of the $t\bar{t}$ pair. This is achieved by uniformly sampling the kinematically allowed decay phase space for the top and antitop quarks in their respective rest frames. The events are kept with a probability according to the full matrix elements $pp \rightarrow t\bar{t} \rightarrow \{b\bar{b}l\bar{l}\nu_l\bar{\nu}_l, b\bar{b}q\bar{q}l\nu_l, b\bar{b}q\bar{q}qq\}$ evaluated at LO in QCD. This procedure ensures that spin correlations between the top and antitop quarks are preserved at LO accuracy. This algorithm is also known as the MADSPIN algorithm [111].

While the matrix element calculation assumes stable top quarks (Narrow Width Approximation (NWA)), finite width effects are approximated in hvq by reshuffling the momenta of the decay products such that the invariant mass of the top quark candidates follows a Breit-Wigner distribution centered at the top quark mass. However, this is a post-processing step and does not account for off-shell effects at the matrix-element level or interference with non-resonant backgrounds.

5.2.1.2 The $t\bar{t}J$ -MiNNLO Generator

$t\bar{t}J$ -MiNNLO is a POWHEG BOX v2 algorithm that enables NNLO accuracy for $t\bar{t}$ production. It uses the MiNNLO_{PS} (Multiscale improved NNLO) method [112] which is an expansion of the MiNLO method [113] and MiNLO' method [114].

MiNLO improves NLO multijet calculations by implementing a dynamical, process-dependent choice of renormalization and factorization scales and by incorporating modified Sudakov form factors. These Sudakov factors resum logarithmically enhanced contributions and describe the probability that a parton does not emit additional soft or collinear radiation above a given scale. The improved MiNLO' procedure [114] is used to regularize standard NLO+PS POWHEG calculations for processes of the type $F + \text{jet}$, such as Higgs+jet or W/Z +jet production. In fixed-order NLO calculations, these processes exhibit divergences when the transverse momentum of the additional jet becomes small or collinear. By consistently resumming these logarithmic contributions, MiNLO' renders the cross section finite in this region and achieves NLO accuracy also for observables inclusive in the additional jet. MiNNLO_{PS} builds on MiNLO' and includes higher order corrections derived from momentum-space resummation of the transverse-momentum spectrum. It can retain the leading log accuracy of the shower program and does not rely on a merging scale. This way, consistent NNLO accuracy of the inclusive production can be achieved.

Practically, the code generates F and 2 jets in the ME generator stage and subsequent emission is handled by the PS code. This method was originally derived for color singlet production but was later extended to heavy quark pair production which features color charge in the final state. For $t\bar{t}$ production this is the `ttJ_MiNNLO` generator used in this work [115, 116]. The color charge enables gluon exchange between initial and final states which leads to non-trivial color interference effects. An exemplary Feynman graph is shown in figure 5.1 (right).

`ttJ_MiNNLO` inherits the treatment of top quark decays from the same `MADSPIN` algorithm implemented into `hvg` which yields LO accuracy. Off-shell effects are also implemented following the `hvg` algorithm employing a narrow width approximation.

5.3 IMPLEMENTING THE `ttJ_MiNNLO` ALGORITHM INTO CMS SOFTWARE

The experimental collaborations employ MC generators in their own respective frameworks. To make NNLO $t\bar{t}$ production available to the experiment, `ttJ_MiNNLO` has to be implemented into the `CMSSW` software and subsequently validated. `CMSSW` interacts with `POWHEG` with a suite of scripts in the "genproductions" environment. These scripts enable compilation of the process and handle the generation of MC grids. These grids are subsequently stored in "gridpacks" which allow for portable, efficient and reproduceable event generation.

To build `ttJ_MiNNLO` gridpacks, the code itself and all dependencies have to be compiled with relative paths. This is achieved by a compiler patch. Some extra care had to be taken for the needed external `OpenLoops` library [117] which handles calculation of tree and one-loop scattering amplitudes. It continuously failed to find the installed `OpenLoops` libraries when the paths to the installation files were set at runtime. The issue was ultimately circumvented by a recompilation of the needed `OpenLoops` libraries at the time of execution of the gridpack. This provided stable usage of this external dependency.

The compiled `POWHEG` code is then used to prepare MC grids [118]. The physical and technical parameters used in the `powheg.input` file are attached in the appendix A.1.

NNLO MC grid generation is a stability challenge due to the large final state phase space that needs to be covered. A few measures had to be taken to minimize stability concerns.

To achieve the necessary statistical precision for the phase space, the grid computation is highly parallelized. MC grids are computed using 400 parallel jobs for `POWHEG` stages 1 through 3. Crucially, the parameters `ncall1` (set to 50,000) and `ncall2` (set to 100,000) define the sampling density per job. In stage 1, the `ncall1` iterations are used for grid adaptation, optimizing the integration grid to map the peaks of the differential cross-section (importance sampling). In stage 2, the `ncall2` iterations perform the actual cross-section integration. The high degree of parallelization results in a total effective statistical sample of 4×10^7 calls for the integration step, which is essential to resolve the NNLO final state. Because the grid computation is very time expensive (5-15 days on CERN LXPLUS computation infrastructure), some jobs fail during the computations due to hardware disconnects or other instabilities. It was found that the high number of parallel jobs can mitigate this issue as the results will be stable even if a few points in the sample are missing.

During grid computation and event generation (`POWHEG` step 4), the `POWHEG` folding mechanism is used with 5 folds in ξ, y and ϕ direction. In `POWHEG`, ξ describes the energy fraction of radiated partons, y describes the polar angle of radiation and ϕ describes the azimuthal angle and the folding describes the number of points that computations are averaged over in these parameters. These folding parameters differ from the CMS settings of `hvg` which sets these to 1. `ttJ_MiNNLO` requires large folding parameters as its additional real emission makes the radiation phase space more challenging to control. It was observed that the folding keeps the event generation from failing at a small cost of computational performance. Except for the folding, all other technical MC parameters were chosen from the suggested `powheg.input` of the authors which is shipped with the code. The authors suggestion for folding was set to 1.

The folding parameter also controls the negative weight fraction (see also section 8.1). Higher folding values minimize the need for negatively weighted events. Increasing the fold parameter in `ttJ_MiNNLO` from 1 to 5 decreases the negative weight fraction from 20% to 5%.

5.3.1 Validating the `ttJ_MiNNLO` Implementation

To check the validity of the generated events, the phenomenological analysis in the paper [116], which compares the `MiNNLOPS` to NLO+PS results, is taken as guidance to compare distributions to. The paper uses different physical parameters for the top mass (assuming $m_t = 173.7$ GeV) and widths of $\Gamma_t = \Gamma_W = 0$ GeV instead of the CMS parameters shown in table 5.1. Also, the comparison generator is based on the `MiNNLO'` method to achieve NLO+PS accuracy. In this work, `hvg` was chosen as a comparison NLO+PS generator as it is the standard CMS $t\bar{t}$ generator. It however does not rely on `MiNNLO'`. It achieves NLO+PS accuracy natively.

General qualitative statements about the distributions are comparable, because the comparisons in this work are conducted with respect to `hvg` using consistent physical parameters. Only normalized distributions are shown in the following as in practice cross section information can be inferred more accurately from inclusive calculations such as `TOP++` [122] which are NNLO and NNLL (Next-to-next-to leading log accuracy in resummations) accurate. This is also done in the later chapters discussing the search for heavy Higgs bosons.

Parameter	Value
<i>Physical Parameters</i>	
Top mass m_{top}	172.5 GeV
Top width Γ_{top}	1.31 GeV
W boson mass m_W	80.4 GeV
W boson width Γ_W	2.141 GeV
<i>Beam Parameters</i>	
$\sqrt{s}$	13 TeV
nominal PDF	NNPDF31_nnlo_hessian_pdfas [119]
<i>Showering Parameters</i>	
Shower Code	PYTHIA 8.309 [120]
Tune	CP5 [121]
POWHEG:veto	on
POWHEG:pThard	0
POWHEG:pTemt	0
POWHEG:pTdef	1
POWHEG:emitted	0
SpaceShower:pTmaxMatch	2
TimeShower:pTmaxMatch	2

TABLE 5.1: Physical parameters for the $t\bar{t}J$ -MiNNLO event generation experiments.

5.3.1.1 Semileptonic Decays

The semileptonic decays are compared using the RIVET [123] routine CMS_2018_I1663958 which is based on a CMS unfolded measurement [124]. The event selection necessitates a semi-leptonic topology characterized by exactly one signal lepton, zero veto leptons, at least two b -tagged jets, and at least two light-flavor jets. Particle-level objects are defined using the visible final state $|\eta| < 6$. Leptons (electrons and muons) are "dressed" with photons within a cone of $\Delta R = 0.1$ and are categorized as signal candidates if they satisfy $p_T > 30$ GeV, $|\eta| < 2.4$, and a scalar relative isolation requirement of $I_{\text{rel}} < 0.35$ within a cone of $\Delta R = 0.4$. Events containing veto leptons ($15 < p_T < 30$ GeV) are rejected to ensure orthogonality. Jets are clustered using the anti- k_t algorithm with $R = 0.4$ and must satisfy $p_T > 25$ GeV and $|\eta| < 2.4$. An overlap removal procedure is applied to discard any jet found within $\Delta R < 0.4$ of a signal lepton, veto lepton, or isolated photon.

The $t\bar{t}$ system is then reconstructed via a kinematic fit that minimizes the squared deviation from the W boson and top quark pole masses (using standard CMS parameters $m_W = 80.4$ GeV and $m_t = 172.5$ GeV) across all permutations of the selected jets. Finally, additional jets are identified as those with $p_T > 30$ GeV that are not identical ($\Delta R > 10^{-5}$) to the constituents of the reconstructed top decay products.

The results of this analysis are presented in figure 5.2 for the $t\bar{t}J$ -MiNNLO and hvq generators. They are compared to 35.8 fb^{-1} of unfolded CMS data taken at $\sqrt{s} = 13$ TeV. The error bars account for the MC statistical uncertainty. The first row shows kinematic variables for the $t\bar{t}$ pair. The qualitative behavior is similar to the behavior reported in figure 4 of Ref. [116]. For high values of $|\eta_{t\bar{t}}|$ some fluctuations can be observed. There is a large discrepancy in the lowest $m_{t\bar{t}}$ bin, exceeding a factor of 1.2 of data to MC for normalized distributions. This discrepancy is evident in NLO+PS and MiNNLO_{PS} predictions. For higher values of $m_{t\bar{t}}$ there is an upwards trend in data with respect to $t\bar{t}J$ -MiNNLO. Again, MiNNLO_{PS} does not offer a large gain in data agreement over NLO+PS. This changes for $p_{t\bar{t}}$. Events with $p_{t\bar{t}} > 500$ GeV are well described by $t\bar{t}J$ -MiNNLO while hvq overshoots the data by $>10 - 20\%$. This is consistent with prior findings that the $t\bar{t}$ p_t distribution can only adequately be described in NNLO accuracy for $t\bar{t}$ production [125].

The pseudorapidity distributions show no disagreements between hvq and $t\bar{t}J$ -MiNNLO which mirrors the findings in Ref. [116]. In high $|\eta|$ regions, some data to MC disagreement is still evident for the jet associated to the first W boson, the b -jet associated to the hadronically decaying W boson and the b -jet associated with the leptonically decaying W boson.

The p_T distributions of the same particles show an interesting behavior. hvq and $t\bar{t}J$ -MiNNLO disagree by up to 10% in shape for the p_T of the jet associated to the first W boson. In high p_T regions, hvq provides the better description of the data shape. This is not visible in figure 4 of Ref. [116] as the normalization effect hides the shape effect. The relative ratio of NLO to NNLO agrees between the authors' findings and this implementation. For the p_T spectra of the b quarks, $t\bar{t}J$ -MiNNLO offers a good description of data, as hvq does. In case of high p_T b -quarks associated to

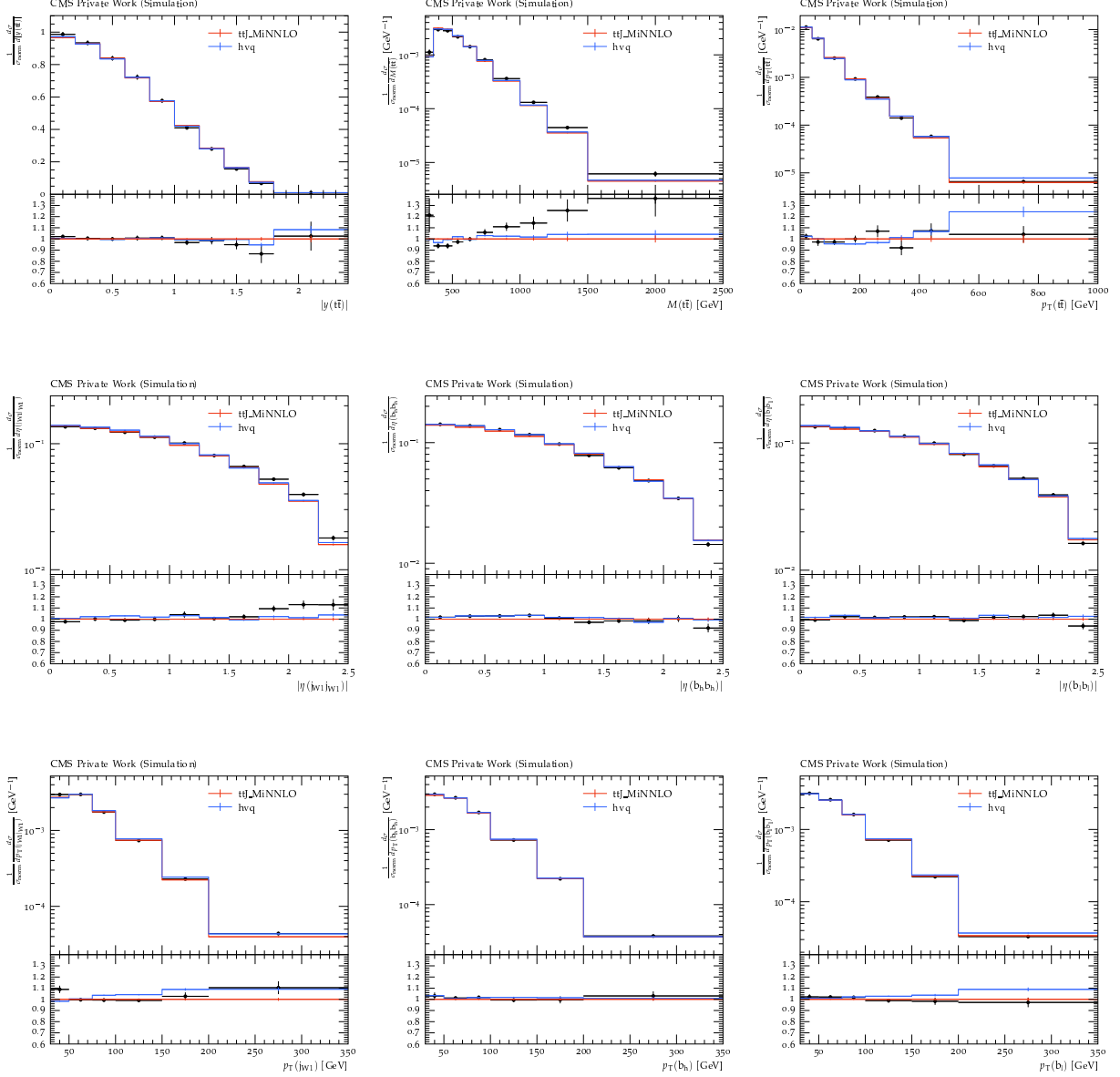


FIGURE 5.2: Comparison of `ttj_minnlo` (red), `hvq` (blue) predictions with CMS data [124] for semileptonic ME-level decays. Displayed are the normalised differential cross-sections of the absolute rapidity, invariant mass and transverse momentum of the $t\bar{t}$ system (upper row, from left to right), the pseudorapidity of the jet from the first W boson, the b quark of the hadronically decaying top and the b quark of the leptonically decaying top (middle row, from left to right), and the transverse momentum of the jet from the first W boson, the b quark of the hadronically decaying top and the b quark of the leptonically decaying top (lower row, from left to right). Each panel has a lower part for a ratio plot where the ratio to the `ttj_minnlo` prediction is shown.

the leptonically decaying W boson, the description of ttJ_MiNNLO is superior to hvq. The authors of Ref. [116] of Ref. [116] found a significant 20% overshoot of NLO and NNLO predictions compared to data in $p_T > 200$ GeV. In this implementation, the NNLO calculation agrees with the data within 5%. Qualitatively, the behavior agrees between the two implementations. MC underestimates data in the low p_T region and overestimates it in the high p_T . One part of the quantitative disagreement between Ref. [116] and this work can be attributed to the normalization, the different tune (Monash 2013 [126] vs. CP5 [121]). The presented results are compatible with the original ones of the authors within their quoted uncertainty.

Overall, the implementation of ttJ_MiNNLO shows sensible results for semileptonic $t\bar{t}$ decays while offering significantly improved description of data for the p_T spectrum of the $t\bar{t}$ pair.

5.3.1.2 Dileptonic Decays

For the dileptonic decay channel, the $t\bar{t}$ generators are compared to 36.1 fb^{-1} of data taken at $\sqrt{s} = 13 \text{ TeV}$ from the ATLAS experiment [127] implemented into RIVET as ATLAS_2019_I1759875. The particle-level selection utilizes prompt electrons and muons (excluding those from hadron decays). These leptons are, again, "dressed" with prompt photons found within a cone of size $\Delta R = 0.1$ to recover final-state radiation. Both the dressed electrons and muons are required to satisfy kinematic cuts of $p_T > 20 \text{ GeV}$ and $|\eta| < 2.5$. The event selection isolates the $e^\pm \mu^\mp$ topology by requiring at least one prompt electron and one prompt muon with opposite electric charges. The results are shown in figure 5.3, the error bars represent again the statistical uncertainty. The first two rows show kinematic features of the decay leptons. In the paper [116], these are shown to be flat within fluctuation between NLO+PS and MiNNLO_{PS} predictions and data. The visible effect of NNLO accuracy that the authors quote is the higher cross section. In this work, a flat shape in the ratio between hvq and ttJ_MiNNLO is thus expected. This is validated by observing the normalized distributions in figure 5.3, within statistical fluctuations hvq and ttJ_MiNNLO agree across the lepton p_T , $m_{e\mu}$, lepton pseudorapidity $|\eta_\ell|$ and the angle $\Delta\phi_{e\mu}$ between the leptons. In the low $m_{e\mu}$ region, an underestimate is also observed, but the ratio between ttJ_MiNNLO and data is closer to 1, than the original authors quote, as the normalization shifts the whole data distribution down. In a normalized ratio, the intermediate $m_{e\mu}$ region between 200 and 400 GeV is thus overestimated by the ttJ_MiNNLO prediction. In the final $m_{e\mu}$ bin with $m_{e\mu} > 400 \text{ GeV}$, the ttJ_MiNNLO prediction shows an improvement over the hvq prediction. The differential distributions in row 3 and row 4 also show no significant disagreements apart from statistical fluctuations. Notably, also the shape change from negative slope below $m_{e\mu} < 80 \text{ GeV}$ to positive for higher $m_{e\mu}$ for the $\Delta\phi_{e\mu}$ distribution is consistently modelled between this ttJ_MiNNLO implementation and the original authors distributions. In the dileptonic $t\bar{t}$ decay channel the implementation is thus also ultimately consistent with the published ttJ_MiNNLO findings. This raises confidence in the correct implementation.

5.3.2 Benefits of NNLO $t\bar{t}$ Generation

As previously mentioned, the experimental collaborations have established that the mismodelling of the $t\bar{t}$ systems top p_T can only be resolved by incorporating NNLO corrections into the simulation of $t\bar{t}$ production [125]. Observed top p_T distributions are softer than NLO predictions. The additional real radiation can carry away some part of the p_T and thus correct for this issue. A driving motivation to use the vastly more complex ttJ_MiNNLO generator over hvq is to solve this issue without a need for further data driven corrections such as the top p_T reweighting in analysis where a well modelled top p_T distribution is crucial.

To validate that the issue is indeed solved, a sample of $t\bar{t}$ pairs was produced with ttJ_MiNNLO without inclusion of the internal decay. Leaving the decays out of the ME calculation allows the calculated quantities at ME level to be guaranteed NNLO accurate as the ME-level decays are only LO accurate in ttJ_MiNNLO as well as hvq. Figure 5.4 shows the effect in the left panel. The p_T of the top quark is shown before parton shower. The predicted p_T distribution of ttJ_MiNNLO is indeed softer, predicting more top quarks with $p_T < 150 \text{ GeV}$ than hvq. Subsequently, it predicts less events in the high p_T tail. Statistical uncertainty is higher here but the trend follows the findings in Ref. [125] quantitatively.

The right panel of figure 5.4 shows another advantage of direct NNLO top quark pair MC. The distributions of additional jets (not counting the jets originating from the $t\bar{t}$ decay) is shown after parton showering. It can be seen that NLO and NNLO predictions agree for lower multiplicities within a few percent. For events with more than 5 additional jets, the NNLO prediction is $\sim 10\%$ higher than NLO. This effect is induced by the higher number of ME emission in addition to the matching to the parton shower. While ttJ_MiNNLO generates one more real emission than hvq the large differences between the generators start to show only in the regime of extra emissions (≥ 5) that are handled by the parton shower. Because the POWHEG veto is set at the p_T scale of the additional emission for ttJ_MiNNLO, the subsequent lower p_T emissions are better modelled by the PS code that generally performs best in the soft regime.

5.3.3 Outstanding Validation Topics

During validation it was found that the m_t modelling of hvq and ttJ_MiNNLO are incompatible. Both generators calculate the ME for stable tops on-shell in the NWA, which would lead to a hard peak at the chosen top mass

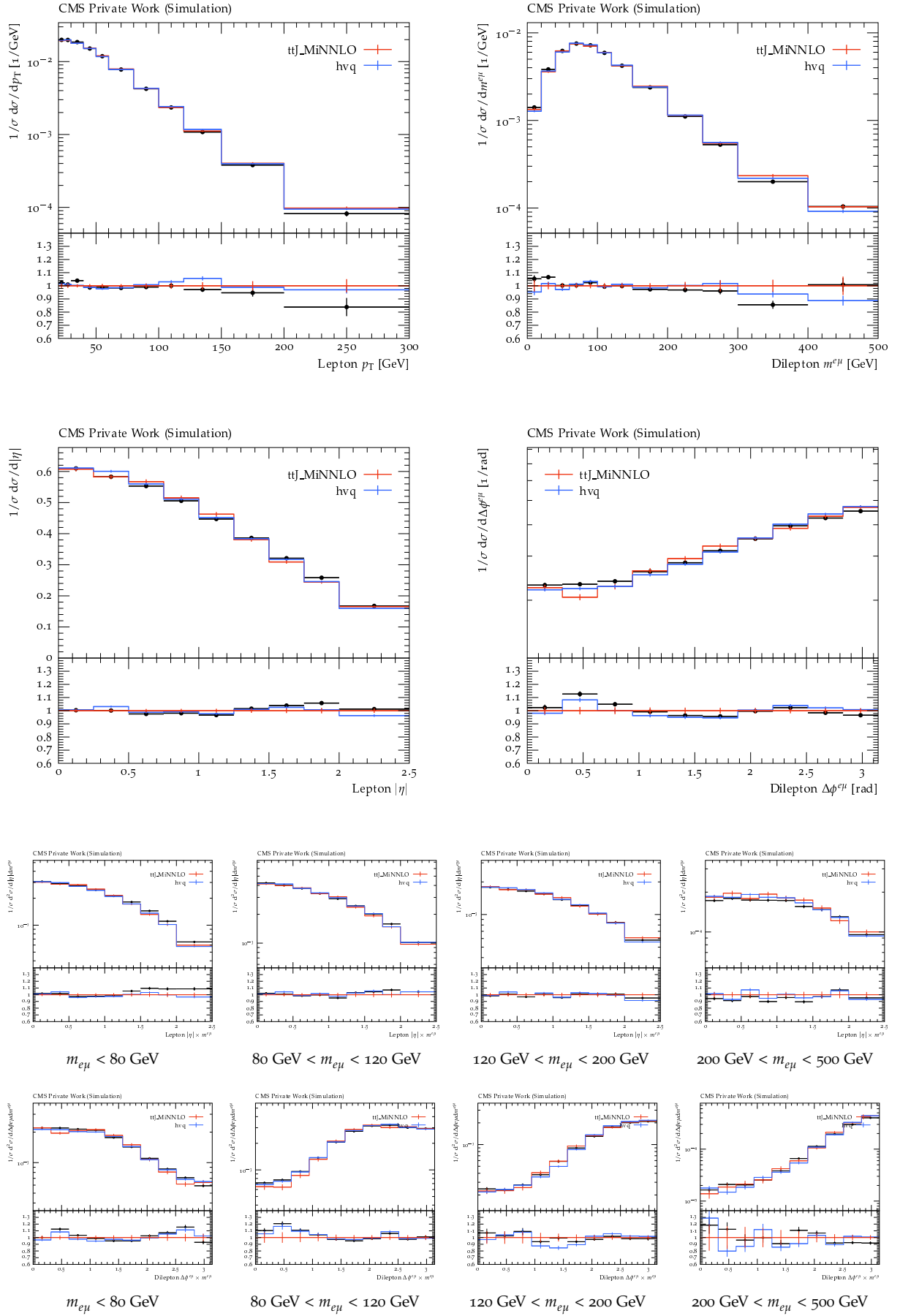


FIGURE 5.3: Comparison of ttj_minNLO (red), hvq (blue) predictions with ATLAS data [127] for dileptonic ME-level decays. Shown are normalized distributions for the lepton p_T (upper left), dilepton $m_{e\mu}$ (upper right), lepton $|\eta|$ (second row left), the opening angle between the leptons $\Delta\phi^{e\mu}$ (second row right), the lepton $|\eta|$ in four bins of the $m_{e\mu}$ ($m_{e\mu} < 80$ GeV, $80 < m_{e\mu} < 120$ GeV, $120 < m_{e\mu} < 200$ GeV, $200 < m_{e\mu} < 500$ GeV, third row left to right) and the dilepton $\Delta\phi^{e\mu}$ in the same $m_{e\mu}$ bins (fourth row, left to right). Each panel has a lower part for a ratio plot where the ratio to the ttj_minNLO prediction is shown.

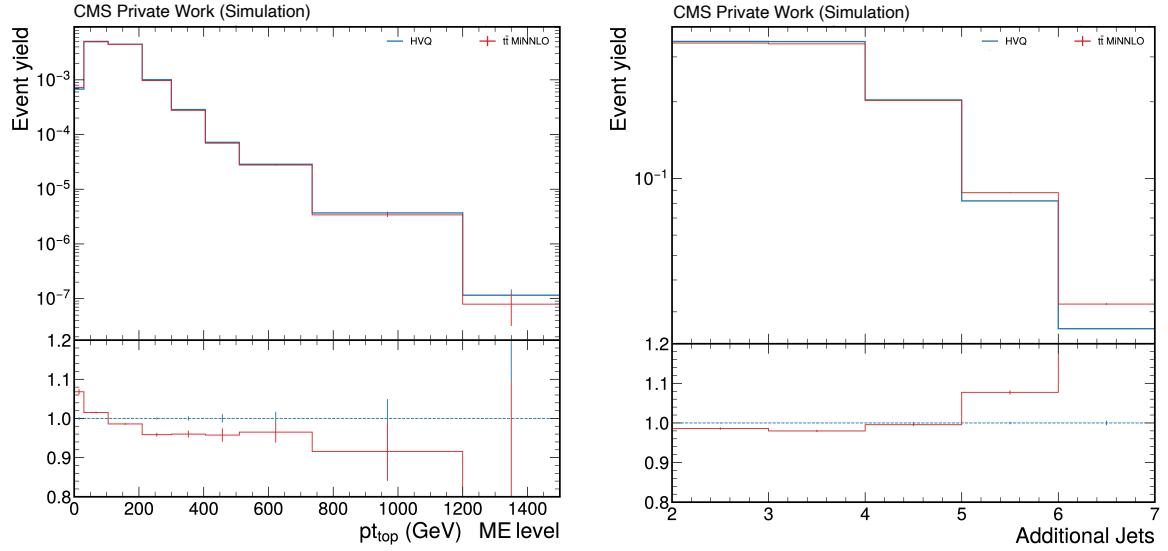


FIGURE 5.4: Comparison of hvq and ttJ_MiNNLO in additional jet modelling and top quark p_T modelling. The left plot shows the normalized histogram of top quark p_T at ME level, before any parton shower. The right plot shows the normalized event yield in different bins accounting for the additional jets, not counting the jets arising from the top quarks decay, found in that event after parton showering. The ratio plots show the ratio of ttJ_MiNNLO to hvq. Note that this is the inverted ratio compared to other plots in this chapter. This was chosen deliberately for the sake of comparability to Ref. [125].

of 172.5 GeV. Enabling the decay at ME level also enables the smearing algorithm described in section 5.2.1.1. For computational reasons, the smearing around the top mass is truncated at a set value given in multiples of the top width for ttJ_MiNNLO and hvq only allows values of m_t between the mass of the W boson and $\sqrt{2} \times m_t$. As the $\pm 10 \times \Gamma_t$ window around m_t for ttJ_MiNNLO is more constrained than the hvq handling, the distributions show a large disagreement starting at ± 10 GeV from the chosen top mass parameter. This effect can be observed at the ME level in the right panel of figure 5.5.

These modelling choices in the approximate description propagate to observable-level predictions and can significantly distort kinematic distributions even after parton showering. The left panel of figure 5.5 illustrates this for the m_t distribution, where deviations of up to 20% are observed with the chosen binning. Such effects are critical for analyses sensitive to off-shell and virtual top-quark contributions. A fully consistent treatment of off-shell effects requires the use of a dedicated generator such as b_bbar_4l [128].

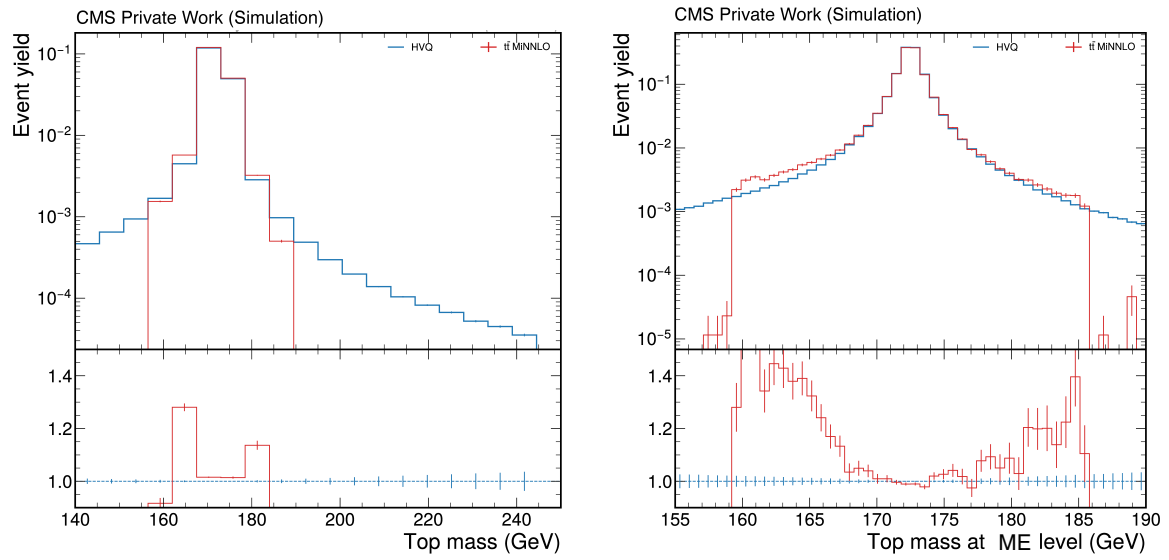


FIGURE 5.5: Differences in top mass modelling between `hvq` and `ttj_minnlo`. The left plot shows the top mass distribution after parton shower and the right plot shows the top mass distribution on ME level. The data was produced with dileptonic decays enabled on ME level for both simulation codes.

In summary, the `ttJ_MiNNLO` NNLO MC generator for $t\bar{t}$ production has been implemented into CMS software and its basic usage has been validated. Especially the top p_T distribution has shown to be better modelled compared to `hVq`. While some aspects, such as the m_t distribution, need further investigation, the baseline for a new level of accuracy in $t\bar{t}$ analyses has been established.

SEARCH FOR HEAVY HIGGS BOSONS IN TOP-ANTITOP QUARK PAIR PRODUCTION

Couplings between BSM spin-0 particles and $t\bar{t}$ are predicted in many extensions of the SM and can potentially help to explain DM, contribute to CP violation or more. Modelling of such states and their impact on $t\bar{t}$ final states has been discussed in chapter 2.4.1. In addition to the extended Higgs sectors discussed there, an important pseudoscalar contribution at the top quark pair production threshold ($E \approx 2m_t$) can come from quasi bound states of top quarks.

6.1 DATA ANALYSIS

This chapter summarizes the search for heavy pseudoscalar and scalar particles in $t\bar{t}$ final states. As primary author of Ref. [43] and Ref. [129], the author contributed to the 2D exclusion contours. In addition, the analysis setup of the CMS analysis, which was worked out in Ref. [130] and Ref. [98], is used to search for ALPs. This search is detailed in a dedicated chapter 7.

The computation of the 2D exclusion contours uncovered computational challenges which are discussed in section 6.2.

As pioneering work for future extensions of this analysis, the author provides an outlook on a more complex realization of (pseudo)scalar additions to the SM given in 6.3.1 and the challenges of higher dimensional theory parameter spaces are addressed with a pilot study utilizing SBI techniques.

Additionally, future iterations of this analysis will rely on improved $t\bar{t}$ reconstruction. Building on Ref. [98], ML techniques are introduced and work on a state of the art physics-inspired transformer architecture for top reconstruction is discussed in 6.4.

6.1.1 Analysis Setup

The analysis is conducted in two separate channels, the single-lepton channel comprised of final states with at least 3 jets (ℓj channel) and the dileptonic channel with exactly two opposite sign leptons ($\ell\ell$). They are statistically combined. The data for this analysis was taken in LHC Run 2 from 2016 to 2018 and amounts to 138 fb^{-1} . It is divided into four eras, two for 2016¹ and one for 2017 and 2018. The analyzed data has been re-reconstructed and recalibrated in the so called CMS Run 2 "Ultralegacy" (UL) effort. A combination of single lepton and dilepton triggers was used with slightly varying p_T thresholds between eras.

As the $\ell\ell$ channel setup is utilized for the search for ALPs in chapter 7 it will be described in more detail. The ℓj channel setup will be briefly explained as it impacts the results of section 6.2, a full description is provided in Ref. [43].

6.1.1.1 Selection of the $\ell\ell$ channel

In the $\ell\ell$ channel, events are considered which feature exactly one pair of opposite charge leptons, either electrons or muons. This results in three channels (ee , $\mu\mu$ and $e\mu$). Additionally, two jets are required with $p_T > 30 \text{ GeV}$ with one being b tagged by the DeepJet algorithm at a CMS defined medium working point [132, 133].

Jets are clustered by the anti- k_T algorithm [134] with a distance parameter of $\Delta R = 0.4$ and the jet momentum is measured from the vectorial sum of all constituent particles. The measured momentum resolution is 5 - 10% in all p_T ranges within the detector acceptance. Jet energy corrections are derived from simulation, and additional in situ measurements of the momentum imbalance in dijet, photon+jet and Z+jet events are used to calibrate the jet energy scale. The jet energy resolution is 15-20% for low energies at 30 GeV and goes down to 5% at 1000 GeV. Jets need to have $|\eta| < 2.4$ and have $p_T > 30 \text{ GeV}$ to be considered. They are also required to have a distance of $\Delta R > 0.4$ from any selected lepton.

Leptons are selected if $|\eta| < 2.4$ and $p_T > 20 \text{ GeV}$. A section between the barrel ECAL and the endcap ECAL is removed. Events are thus not considered if they fall into $1.44 < |\eta| < 1.57$. Additionally, electrons need to fulfill an MVA based ID criterium [135] which facilitates 90% background rejection at $\sim 90\%$ signal efficiency and includes an isolation requirement. Muons need to pass an ID requirement [136] and an isolation requirement on the relative isolation variable I_{rel} .

The leading lepton needs to pass $p_T > 25 \text{ GeV}$ and the subleading lepton $p_T > 20 \text{ GeV}$. To cut away low mass resonances, a cut on the invariant dilepton mass $m_{\ell\ell} > 20 \text{ GeV}$ is imposed. For the same-flavor lepton channels, a Z

¹ The data of the year 2016 is divided into the "2016pre" and "2016post" eras for the periods before and after the modification of the APV readout chip which affects track hit reconstruction [131].

Process	MC Simulation QCD order	ME Generator	Cross section [pb]	Cross section Source
$t\bar{t}$	NLO	POWHEG v2 (hvq [143])	833.9	Top++ [122]
tW	NLO	POWHEG (ST_wtch [144])	v2 71.7	Ref. [145]
$Z\gamma$ +jets	NNLO	POWHEG v2 (Zj MiNNLO [146, 147])	24.7×10^3	FEWZ [148, 149]
t -channel single top	NLO	POWHEG v2 (ST_tch [150]) + MADSPIN [151]	217.0	HATHOR [139, 152]
s -channel single top	NLO	MG5_AMC@NLO	10.3	HATHOR [139, 152]
$t\bar{t}W$	NLO	MG5_AMC@NLO	0.64	MG5_AMC@NLO
$t\bar{t}Z$	NLO	MG5_AMC@NLO	0.75	MG5_AMC@NLO
WW	LO	PYTHIA 8.2	118.7	[153]
WZ	LO	PYTHIA 8.2	47.1	MCFM [154]
ZZ	LO	PYTHIA 8.2	16.5	MCFM [154]

TABLE 6.1: An overview of the different background processes considered in the search for heavy (pseudo-)scalars. They are listed with the ME generator used to simulate them, the QCD order of that generator as well as the used cross section normalization and the source for that number. For all samples, PYTHIA 8.2 is used for showering and hadronization.

window is cut out $|m_{\ell\ell} - m_Z| < 15$ GeV. To account for the expected neutrinos from the W boson decays, $p_T^{\text{miss}} > 40$ GeV is required.

6.1.1.2 Experimental Corrections

To ensure good agreement between data and simulation in the various distributions, several corrections are applied to the simulated samples. Most of these corrections are derived centrally by the CMS collaboration and include trigger scale factors, lepton scale factors, pileup reweighting, jet energy corrections, MET corrections, b-tagging scale factors, and ECAL L1 prefire corrections [43].

6.1.1.3 Background Modelling

The selection of the $\ell\ell$ channel achieves $\sim 90\%$ purity in $t\bar{t}$. Remaining non- $t\bar{t}$ backgrounds are single top production, Z+jet production, diboson and $t\bar{t}$ V production. QCD multijet production with non-prompt or misidentified leptons contributes only a miniscule effect and was not further considered. The background contributions are modelled in MC simulation, an overview can be obtained from table 6.1. In the case of Z+jets production a data driven normalization was applied. The normalization is taken from the Z-window control region ($|m_{\ell\ell} - m_Z| > 15$ GeV) which was cut out from the remainder of the analysis [137].

Particular care has to be taken with the non-reducible $t\bar{t}$ background. It was simulated with POWHEG hvq at fixed order perturbative QCD NLO accuracy. Then, a 2D reweighting to NNLO accuracy binned in $m_{t\bar{t}}$ and the scattering angle between the top and the beam in the $t\bar{t}$ restframe $\cos\theta^*$ was applied. The NNLO prediction was computed with the MATRIX program [138] and shows corrections of 1-2%. Additionally, NLO EW corrections are applied in the same reweighting procedure. These are computed in HATHOR [139] [140–142]. Notable contributions come from virtual Higgs particle exchanges between the top and antitop. They can influence the spin state of the $t\bar{t}$ system. HATHOR is only accurate up to LO in α_S and thus there is ambiguity in how to apply the correction to the NLO simulation (with NNLO corrections applied). An additive and multiplicative scheme can be employed [43]. The multiplicative scheme leads to 2-4% corrections and the additive approach to 1-2%. The multiplicative scheme was chosen and a shape uncertainty was assigned to the difference to the additive scheme. Additionally, an uncertainty was added for varied y_t as this impacts the shape and the magnitude of the EW correction.

6.1.1.4 Systematic Uncertainties

A set of theory and experimental uncertainties is implemented to cover systematic effects. The theory uncertainties are listed in table 6.2 and the experimental uncertainties are listed in table 6.3.

Uncertainties estimated from separate MC samples are smoothed to minimize the impact of the statistical fluctuations in these samples. The template smoothing is based on the LOWESS algorithm [168, 169].

Source	Type	Description	Affected Processes
<i>Modeling and Scales</i>			
Top quark Yukawa coupling	Shape	Varied within $1.00^{+0.11}_{-0.11}$ to cover NLO EW corrections [155] .	$t\bar{t}$
EW correction scheme	Shape	Difference between additive and multiplicative combination of QCD and EW corrections.	$t\bar{t}$
Top quark mass (m_t)	Shape	Simulated ± 3 GeV shifts rescaled to emulate ± 1 GeV uncertainty. Considered to be fully correlated between signals and background	$t\bar{t}$, A/H, η_t
ME scales (μ_R, μ_F)	Shape	Renormalization and factorization scales varied independently by factors of 0.5 and 2.0 to cover missing higher orders.	$t\bar{t}$, A/H, η_t , single t , Z+jets
PS ISR/FSR	Shape	Scales for α_S in parton shower simulation varied by factors of 0.5 and 2.0.	$t\bar{t}$, A/H, η_t , single t , Z+jets
Color reconnection	Shape	Comparison between QCD-inspired and early resonance decay models in PYTHIA.	$t\bar{t}$
ME/PS matching (h_{damp})	Shape	POWHEG damping parameter varied.	$t\bar{t}$
PDF	Shape	Variations based on 100 NNPDF3.1 replicas. A principal component analysis on the search template is computed and the leading component is chosen. Additionally, α_S uncertainty in PDF is added.	$t\bar{t}$
<i>Background Normalizations</i>			
$t\bar{t}$ norm.	Norm.	Shifts in NNLO + NNLL $t\bar{t}$ cross section as effects of ME scale variation and m_t are computed with Top++ [122]. They are treated fully correlated to the shape systematics, no new NPs are introduced	$t\bar{t}$
Single top (and tW) norm.	Norm.	Constrained with a normalization uncertainty of 15% [156, 157].	Single t
$t\bar{t}V$ norm.	Norm.	Constrained with a normalization uncertainty of 30% [158, 159].	$t\bar{t}V$
Z+jets norm.	Norm.	Constrained with a normalization uncertainty of 5% [160].	Z+jets
Diboson norm.	Norm.	Constrained with a normalization uncertainty of 30% .	Diboson
η_t norm.	Norm.	Treated as an unconstrained free parameter in the relevant fits.	η_t

TABLE 6.2: Theoretical systematic uncertainties considered in the dileptonic ($\ell\ell$) channel analysis, with descriptions of the variations applied.

Source	Type	Description	Affected Processes
<i>Luminosity and Pileup</i>			
Integrated luminosity	Norm.	Overall 1.6% uncertainty split into individual NPs with varying correlations across eras [161–163] .	All
Pileup	Shape	Effective inelastic proton-proton cross section varied by 4.6% [164].	All
<i>Jet and MET Reconstruction</i>			
Jet energy scale (JES)	Shape	NPs covering absolute/relative scales, flavor response, and pileup offsets [165].	All
Jet energy resolution (JER)	Shape	Variation of resolution scale factors applied to smearing of simulated jets [165].	All
Unclustered p_T^{miss}	Shape	Energy of unclustered PF candidates varied by their respective resolutions [166].	All
<i>Flavor Tagging</i>			
b tagging (heavy flavor)	Shape	Effects of parton shower modeling, leptons in jets, and JES impact on b/c jets [133].	All
b tagging (light flavor)	Shape	Accounts for mistagging rates of light-flavor jets [133].	All
<i>Leptons and Triggers</i>			
Dilepton triggers	Shape	Uncertainty on scale factors correcting trigger efficiency differences (ee, e μ , $\mu\mu$) [135, 136].	All
Electron identification	Shape	Uncertainty on scale factors correcting ID efficiency [135].	All
Muon identification	Shape	Uncertainty on scale factors correcting ID efficiency [136].	All
ECAL L1 trigger	Shape	Correction for L1 timing shift inefficiency in 2016/2017 eras ($ \eta > 2.4$) [167].	All

TABLE 6.3: Experimental systematic uncertainties considered in the dileptonic ($\ell\ell$) channel analysis, with descriptions of the variations applied.

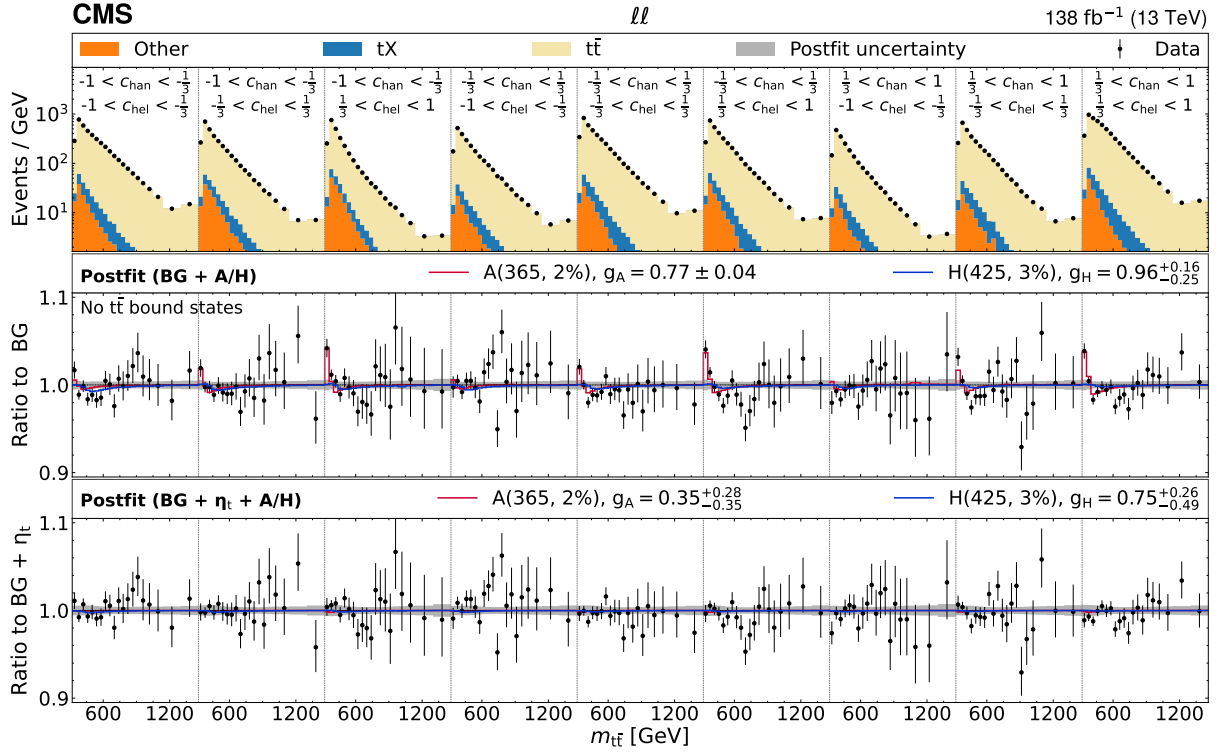


FIGURE 6.1: Postfit data distribution for A with $m_a = 365$ GeV, $w_a = 2\%$ and H with $m_H = 425$ GeV, $w_H = 3\%$ in the $m_{t\bar{t}} \times c_{\text{hel}} \times c_{\text{han}}$ (20 bins $\times$ 3 bins $\times$ 3 bins) search template for the $\ell\ell$ channel. The ratio plots are shown in the lower panel for both fit scenarios including a $t\bar{t}$ bound state contribution η_t (lower panel) and without η_t (middle panel). Figure is taken from [43].

6.1.1.5 Analytic Top Reconstruction in the $\ell\ell$ Channel

To reconstruct the kinematic properties of the dileptonically decayed $t\bar{t}$ system, an analytic approach is used [170]. First, the b quarks and charged leptons are matched in a maximum likelihood fit choosing the m_{ℓ^+b} and $m_{\ell^- \bar{b}}$ pair which maximizes the product of the likelihoods with respect to MC simulation data.

Two top quark fourvectors translate into eight unknown quantities. Since ν escape the detector unmeasured, only their transverse momentum can be constrained from the missing transversal moment p_T^{miss} . This constraint is implemented as the assumption that all p_T^{miss} comes from the massless ($p_\nu^2 = 0$) neutrinos

$$y p_{\nu_{\ell}, x/y} + p_{\bar{\nu}_{\ell}, x/y} = p_{x/y}^{\text{miss}}. \quad (6.1)$$

Also, all top quarks and W bosons are required to be on-shell $p_t^2 = m_t^2$. With these constraints included, one is left with a four fold ambiguity in the results.

Due to detector resolution effects and violation of the assumptions, events have either 4, 2 or 0 solutions. The lepton and b jet inputs to this algorithm are smeared to the standard deviation of the detector resolution and the m_W is sampled from a Breit-Wigner distribution with the width Γ_W . The resulting tops are computed 100x and weighted according to the product of the likelihood for the $\{m_{\ell^- \bar{b}}, m_{\ell^+ b}\}$ pair. A weighted average of the top quarks is then taken as the final result.

6.1.1.6 Search Variables

From the reconstructed tops, the $m_{t\bar{t}}$, c_{hel} and c_{han} values are computed. $m_{t\bar{t}}$ offers discrimination between different mass and width hypotheses.

The spin correlation observables c_{hel} (see equation 2.20) and c_{han} (see equation 2.21) have been introduced in detail in section 2.2.5.1. c_{hel} allows to discriminate pseudoscalar and scalar mediators from each other and SM $t\bar{t}$ systems. A strong positive linear slope is expected for A in the c_{hel} distribution, a negative slope for H and the SM contribution has a smaller positive slope.

c_{han} has a discriminative power towards scalar mediators from both pseudoscalars and SM $t\bar{t}$ contributions. A negative linear slope is visible in the c_{han} spectrum for H hypotheses. The positive slopes of A is smaller than in c_{hel} but still larger than the SM contributions positive slope.

The variables are combined into a 3D template of $20 \times 3 \times 3$ bins in $m_{t\bar{t}} \times c_{\text{hel}} \times c_{\text{han}}$. Subsequently, pseudoscalar hypotheses have the most signal-enriched bins in the highest c_{hel} and c_{han} bin and scalar hypotheses are most prominent in the lowest c_{hel} and c_{han} bin.

6.1.1.7 Setup of the ℓj channel

For the ℓj channel, events with at least three jets and exactly one lepton are selected. At least two of the jets have to be b tagged by the DEEPJET algorithm [132]. The selected lepton (electron or muon) is required to have a transverse momentum of $p_T > 30$ GeV and $|\eta| < 2.4$, with a higher threshold of 34 GeV applied to electrons in the 2017 and 2018 data-taking periods to account for trigger thresholds. To ensure orthogonality with the dilepton channel, events containing additional loosely identified leptons with $p_T > 20$ GeV are vetoed.

The kinematics of the top quark pair system are reconstructed using the method described in [171]. The longitudinal momentum of the neutrino is estimated by imposing a W boson mass constraint on the system formed by the lepton and the missing transverse momentum. The assignment of jets to the parton-level topology is determined by maximizing a likelihood function based on the expected invariant masses of the hadronically decaying top quark and W boson. The method differs for four and three jet scenarios. For events with exactly three jets, a correction derived from simulation is applied to the reconstructed hadronic top quark momentum to account for the missing jet energy [172].

Signal extraction is performed using two-dimensional templates of the invariant mass of the $t\bar{t}$ system, $m_{t\bar{t}}$, and the cosine of the polarization angle, $|\cos \theta_{t\bar{t}}^*|$. The latter is defined as the angle between the leptonically decaying top quark in the zero-momentum frame and the direction of the $t\bar{t}$ system in the laboratory frame. This angular variable discriminates between the signal and the standard model background: scalar and pseudoscalar resonances result in isotropic top quark production (flat distribution), while the standard model background peaks at high values of $|\cos \theta_{t\bar{t}}^*|$.

Similar to the $\ell\ell$ channel, $t\bar{t}$, η_t and single top background processes are estimated with MC simulations. Backgrounds from QCD multijet and electroweak processes are estimated using a data-driven control region defined by events with zero b -tagged jets. A detailed discussion of the uncertainties is available in the publication [43].

6.1.2 Contributions of $t\bar{t}$ bound states at the threshold

At the production threshold of top quark pair production (at $m_{t\bar{t}} \approx 2m_t$) calculations [173–175] predict a signature competing with pseudoscalar contributions. It arises from top quark bound state effects which can be calculated in non-relativistic QCD and are generally not included in perturbative NLO $t\bar{t}$ simulations. Classically, top quarks do not form long lived hadrons due to their short lifetimes but the NRQCD predictions show a cross section enhancement of a few picobarns. The only attractive "toponium" states at the LHC are formed from color singlet contributions. The dominant contribution comes from the $^1S_0^{[1]}$ quasi-bound state η_t .

The CMS publication [129] discusses a search for such a contribution with the $\ell\ell$ setup described above. The bound state is modelled by a simplified η_t model which corresponds to a pseudoscalar with mass $m_{\eta_t} = 343$ GeV and width $\Gamma_{\eta_t} = 2.8$ GeV as given in equation 2.25. This simplified model acts as a proxy and allows to extract the cross section σ_{η_t} of such effects. Additionally, by testing for a similarly constructed scalar contribution χ_t , the CP state of the threshold effects can be probed.

The observed cross section for the η_t model is

$$\sigma(\eta_t) = 8.8 \pm 0.5(\text{stat})_{-1.3}^{+1.1}(\text{syst}) \text{ pb} = 8.8_{-1.4}^{+1.2} \text{ pb}. \quad (6.2)$$

In a simultaneous fit of χ_t and η_t , it is confirmed that the data prefers a pseudoscalar hypothesis. The extracted $\sigma(\chi_t)$ is compatible with 0 within one standard deviation, while the η_t excess remains significant. Similar findings are discussed for the A/H interpretation in section 6.1.4.

In a follow up study of the ATLAS collaboration the excess was confirmed [176]. They extract a cross section of

$$\sigma(t\bar{t}_{\text{NRQCD}}) = 9.0 \pm 1.3 \text{ pb}. \quad (6.3)$$

Here, an approximate NRQCD model is used that uses a reweighting approach derived from the NRQCD LO Green's function of toponium [177]. For the same η_t model that CMS uses they find

$$\sigma(\eta_t) = 13.4 \pm 1.9 \text{ pb}. \quad (6.4)$$

This result exceeds the comparable CMS result by ~ 2 standard deviations.

6.1.3 Statistical Analysis

The templates are fitted with the maximum likelihood fit method as described in section 3.3. A poisson likelihood is defined as

$$L(p_\Phi, \mu(\eta_t), \nu) = \left(\prod_i \frac{\lambda_i(p_\Phi, \mu(\eta_t), \nu)^{n_i}}{n_i!} e^{-\lambda_i(p_\Phi, \mu(\eta_t), \nu)} \right) G(\nu), \quad (6.5)$$

with the yield given by

$$\lambda_i(p_\Phi, \mu(\eta_t), \nu) = S_i^\Phi(p_\Phi, \nu) + S_i^{\eta_t}(\mu(\eta_t), \nu) + B_i(\nu). \quad (6.6)$$

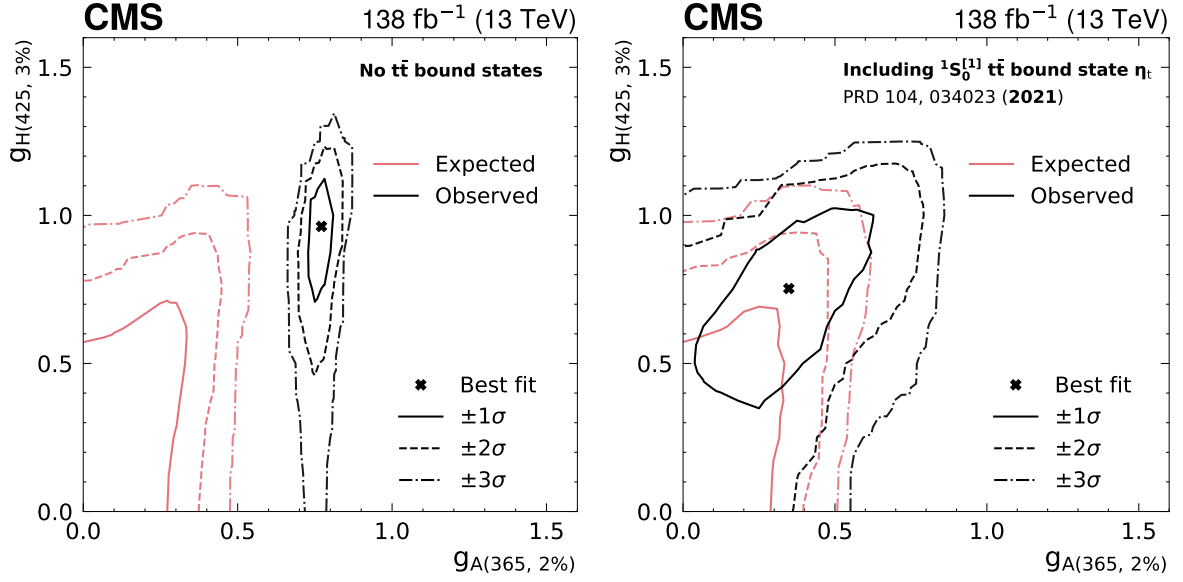


FIGURE 6.2: Frequentist FC exclusion contours for the scenario $m_A = 365$ GeV, $\Gamma_A = 2\%$ and $m_H = 425$ GeV, $\Gamma_H = 3\%$ including η_t in the background model (right panel) and not including the η_t model as a background (left panel). The red lines describe the expectation and the black line describe the observed exclusion for different confidence levels. The plots are taken from [43].

Here, S_i^Φ describes the yield of the pseudoscalar and scalar hypotheses Φ in bin i , $S_i^{\eta_t}$ describes the yield of the η_t simplified model for a $t\bar{t}$ bound state in that bin and B_i describes the full background including the perturbative QCD $t\bar{t}$ production in the respective bin. S_i^Φ is dependent on the signal model parameters p_Φ , $S_i^{\eta_t}$ is dependent on the η_t signal strength $\mu(\eta_t)$ and all contributions are dependent on the vector of NPs v . $G(v)$ is the product of probability density functions signaling the external constraints on the NPs.

The Φ yield is described by

$$S_i(p_\Phi, v) = \mu \sum_{\Phi=A,H} \left(g_{\Phi t\bar{t}}^4 s_{R,i}^\Phi(m_\Phi, \Gamma_\Phi, v) + g_{\Phi t\bar{t}}^2 s_{I,i}^\Phi(m_\Phi, \Gamma_\Phi, v) \right), \quad (6.7)$$

which encodes the resonant s_R^Φ and interference parts s_I^Φ . Here, the Φ parameters p_Φ are explicitly written out as the Φ -boson mass m_Φ , the width Γ_Φ and the coupling modifier $g_{\Phi t\bar{t}}$.

The background-only model is constructed by setting $g_{\Phi t\bar{t}}$ to 0. Results are computed with and without η_t in the background, for no η_t contributions the η_t signal strength is set to 0.

A modified LHC profile likelihood test statistic $\tilde{q}_\mu$ is used [178]. For testing Φ hypotheses an auxiliary signal strength modifier μ is introduced because the Φ signal strength scales nonlinearly with $g_{\Phi t\bar{t}}$. For $\mu = 0$ the SM case is probed and $\mu = 1$ yields the Φ hypothesis. The test statistic then follows the true $g_{\Phi t\bar{t}}$ test statistic for a given value of $g_{\Phi t\bar{t}}$. The asymptotic approximation is used to derive results for the CL_s criterion. The space of $g_{\Phi t\bar{t}}$ can then be scanned to determine which coupling modifier values are excluded. This methodology was tested in [179] and found to yield correct exclusion limits. One dimensional limits in either A or H are calculated by fixing the other contribution to 0.

In the case of two dimensional interpretation, the asymptotic approximation is dropped and the Feldman-Cousins method is used.

6.1.4 Results and Nature of the Excess

The extracted η_t cross section from CMS [129] is able to explain the excess at the $t\bar{t}$ threshold, no further significant excess is observed. This is exemplified in figure 6.2, where the results for H with $m_H = 425$ GeV and $\Gamma_H/m_H = 3\%$ and A with $m_A = 365$ GeV and $\Gamma_A/m_A = 2\%$ are presented. This parameter point accounts for the biggest observed excess.

The left plot shows the exclusion contour without the η_t in the background model. A clear difference between the expected result (red) and the observed (black) result can be seen. Along the pseudoscalar axis g_A the excess is greater than five standard deviations, three standard deviations are shown as a contour in the plot. The three SD contour reaches $g_H = 0$ along the scalar coupling modifier direction. The pseudoscalar nature of the excess consistent with the findings in [129] is recovered.

The right panel shows the scenario with η_t in the background. The observed contour is compatible with the zero hypothesis within two standard deviations in both coupling modifiers. A similar finding is obtained by analyzing the postfit data distributions, shown in figure 6.1. These findings allow for three interpretations of the excess:

1. The excess can be explained by a quasibound state eta without the need of assuming a BSM contribution through an elementary pseudoscalar.
2. The excess can hint to a BSM elementary pseudoscalar in case $t\bar{t}$ bound state effects are small.
3. The excess could be a mixture of $t\bar{t}$ bound state effects and a BSM pseudoscalar contribution.

None of the scenarios is a priori preferred by data. To disentangle a BSM pseudoscalar from a SM $t\bar{t}$ quasibound state, an improved $m_{t\bar{t}}$ resolution is necessary. The implications of this need for better resolution will be further discussed in section 6.4.

With the presented analysis setup one can derive constraints on the parameter space of A and H scenarios. This allows for reinterpretation in specific (for example 2HDM) models with a set relation between the parameters of A and H, because the analysis was conducted with a generic scalar and pseudoscalar hypothesis, covering the full set of A and H parameter combinations. A set of 2D contours is derived with an η_t contribution included as a background with a free floating normalization.

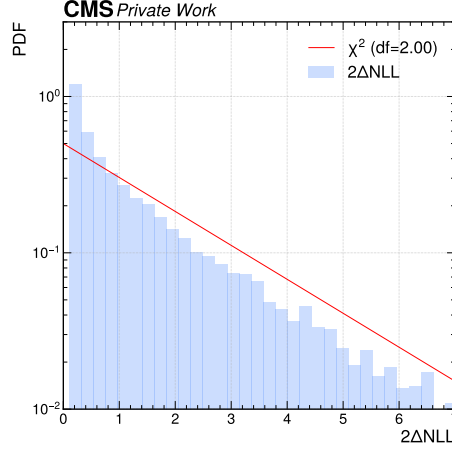


FIGURE 6.3: $2\Delta\text{NLL}$ probability distribution function for toys drawn from $g_{A\bar{t}\bar{t}} = g_{H\bar{t}\bar{t}} = 0$ with η_t in the background, plotted against the asymptotic assumption of χ^2 with two degrees of freedom.

6.2 LIMIT SETTING BEYOND WILKS THEOREM

Wilks theorem can be violated by a number of reasons during the analysis. Wilks theorem presupposes that the underlying true value that is tested for lies inside the parameter space and not on the boundary. POIs that are limited by a boundary and can physically plausibly take the value on the limits break this assumption. Such a problem can arise from signal strengths that can be 0 but generally not negative. In this analysis, there are quadratic and quartic couplings in the signal model, as shown in equation 6.7 for correctly scaling the interference and resonant contribution. Subsequently, negative values are ruled out while 0 is a physically meaningful result (SM only hypothesis). Additionally, NPs can break the assumptions of Wilks theorem too.

Generally, Wilks theorem predicts the asymptotic distribution of the test statistic $\lambda(\theta_{\text{POI}})$ ($2\Delta\text{NLL}$) from equation 3.7 to follow a χ^2 distribution with degrees of freedom equal to the number of POIs. In the scenario of A and H, the $\lambda(\theta_{\text{POI}})$ distribution should thus follow a χ^2 distribution with two degrees of freedom. Figure 6.3 shows that this assumption is violated. Here, one can see the empirical distribution of $2\Delta\text{NLL}$ values of toys in blue. These toys are based on pseudo-datasets drawn from the likelihood model of the analysis for the hypothesis $g_{A\bar{t}\bar{t}} = g_{H\bar{t}\bar{t}} = 0$. If Wilks theorem would hold, the toy distribution would match the prediction of the χ^2 (red) line. One could then immediately read off the p-values from the $2\Delta\text{NLL}$ values of the χ^2 distribution and could circumvent the creating of pseudo-datasets. However, here the χ^2 distribution falls off slower and would thus overestimate the drawn limit in this case.

The statistically sound solution is the usage of the Feldman Cousins technique [70]. Toys are computed for all parameter points on the $g_{A\bar{t}\bar{t}} - g_{H\bar{t}\bar{t}}$ -grid that is being tested, the method is thus fully empirical and relies on no asymptotic assumptions. For each $g_{A\bar{t}\bar{t}} - g_{H\bar{t}\bar{t}}$ hypothesis a maximum likelihood fit on the observed data is carried out to obtain a "best fit" result. Toys that pass the $\lambda(\theta_{\text{POI}})$ of the best fit result at that parameter point are counted and exclusion limits are drawn based on the passing toy distribution on the grid. This process needs a sufficient number of toys for each point to converge, which is extremely computationally expensive. On average each grid point requires ~ 1100 toys. There are ~ 175 grid points per exclusion contour on the $g_{A\bar{t}\bar{t}} - g_{H\bar{t}\bar{t}}$ -grid. Accordingly, this averages to $\sim 200,000$ toys per contour. Two contours (expected and observed) are needed per parameter point. This implementation leverages an iterative algorithm that starts of with a coarse grid in the coupling modifiers $g_{A\bar{t}\bar{t}}$ and $g_{H\bar{t}\bar{t}}$ for a given mass-width scenario for A and H. Around the contours of the first iteration, a finer gridding is chosen in a following step, such that a more granular resolution for the contour is possible. These iterations are repeated two or three times until a visually satisfying result is achieved which does not feature an overly blocky appearance of the contour.

Since the generated toys have to be fitted to obtain a ΔNLL result, instabilities in the fits are propagated into granular areas in the coupling modifier space and a residual "jankyness" is often observed in contours. An example of this is shown in figure 6.4. The FC contour on the left shows an unstable 1SD (68% CL) contour between $0.5 \leq g_{H\bar{t}\bar{t}} \leq 0.9$ and $0.1 \leq g_{A\bar{t}\bar{t}} \leq 0.8$. The contour based on the NLL is smooth, but there are some non-excluded areas inside the excluded space that show the same instabilities in the fit. Computing vastly more toys is a solution but prohibitively expensive for a full grid of masses and widths.

A middle ground between purely asymptotic contours and fully toy based FC contours can be found in a method with shared toys at the background-only hypothesis henceforth referred to as the shared toys method.

For $g_{A\bar{t}\bar{t}} = g_{H\bar{t}\bar{t}} = 0$, the mass and width of the tested particles does not matter. Only the SM will influence this hypothesis which is shared for all signal hypotheses. One can thus once generate a large set of toys for this point and check the observed best fits on the $g_{A\bar{t}\bar{t}} - g_{H\bar{t}\bar{t}}$ grid against the distribution in the $\{0,0\}$ point instead of the toys

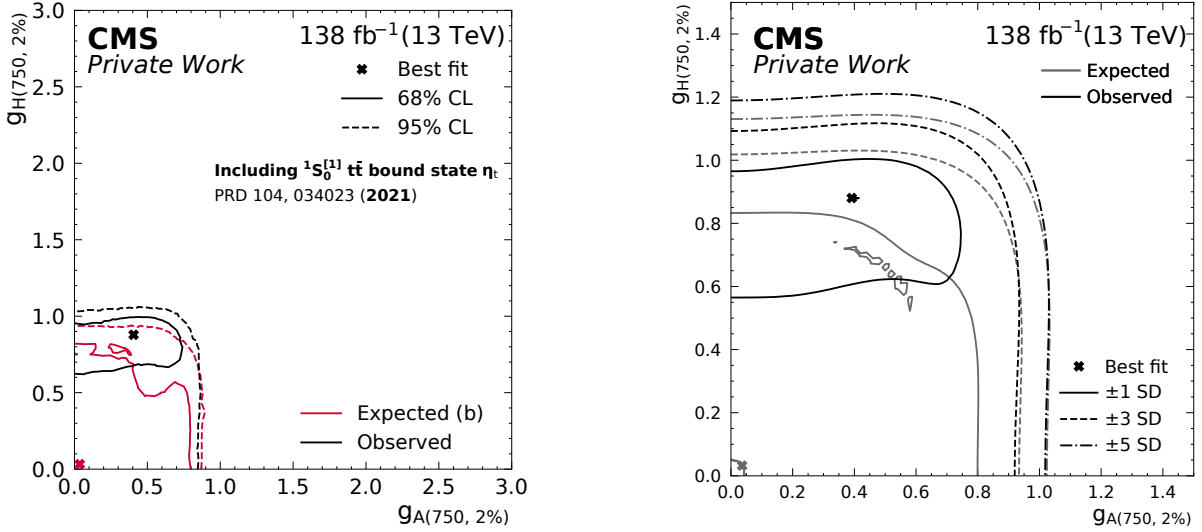


FIGURE 6.4: FC contour exclusion contour (left) and NLL based exclusion contour (right) for the scenario with A (750 GeV mass and 2% width) and H (750 GeV mass and 2% width). The plots follow the scheme of figure 6.2.

at the specific grid point. This rephrases the statistical question from a goodness of fit at each parameter point to a compatibility test to the background-only hypothesis. This method conceptionally yields very similar outcomes for exclusion contours and it is not reliant on asymptotic approximations.

6.2.1 Estimating the Uncertainty of Toy Sampling

For additional confidence in the toy sampling, one can estimate the statistical error of the toy distribution. The Dvoretzky–Kiefer–Wolfowitz (DKW) inequality [180] gives a bound on the difference between the sampled cumulative distribution function (CDF) $F_n(x)$ and the unknown true distribution function $F(x)$

$$\Pr\left(\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| > \epsilon\right) \leq Ce^{-2n\epsilon^2} \quad \text{for every } \epsilon. \quad (6.8)$$

Here, n denotes the sample size and C is a constant which is found to be 2 [181]. The DKW inequality allows to construct confidence band around a sampled CDF

$$F_n(x) - \delta \leq F(x) \leq F_n(x) + \delta \quad \text{with } \delta = \sqrt{\frac{\ln(2/\alpha)}{2n}} \quad (6.9)$$

with α being the confidence level. With this formula it is possible to correct the toy distribution for a sampling error to a given uncertainty. In this work, 95% CL was adopted if not indicated otherwise. Figure 6.5 visualizes the procedure. The ΔNLL distribution of the toys allows to compute the CDF for the toys. Then, a confidence band around this sampled distribution can be computed with equation 6.9. The plot visualizes the 68, 95 and 99.73 percentile values of the toy distribution. The confidence intervals on these are inferred for the points where the $\text{CDF} - \delta$ and $\text{CDF} + \delta$ reach these percentiles respectively. The intervals are reasonably small for the 68 and 95 percentile but the tail of the distribution is much less populated leading to a large uncertainty on the 99.73 percentile estimate. The shown result is based on ~ 130000 toys at $g_{A\bar{t}t} = g_{H\bar{t}t} = 0$.

A subsequent proposal of this work is the adaption of the explained error approximation for the FC results. One can easily and statistically rigorously estimate the error on FC exclusion contours by using the DKW uncertainty bands. However, the FC contours in this analysis feature differing number of toys for each point on the coupling modifier grid due to the iterative algorithm. Thus, the error estimation is not comparable between neighbouring points. The uncertainty band is not smooth and does not offer interpretability in areas of the parameter grid but only at the exact point it was derived at. DKW error bands are thus omitted for FC results in the following. The uncertainty estimates are shown for the shared toys method and can thus be evaluated in the next section.

6.2.2 Comparing the Shared Toys Method to the Feldman Cousins Method

A set of exclusion contours was produced to compare the FC results to the results derived with the shared toy method. Figure 6.6 shows 3 mass-width scenarios. The left column shows the exclusion contours derived with the FC method and the right column is based on results using the shared toys method. The exclusion contours generally agree well in shape and covered area.

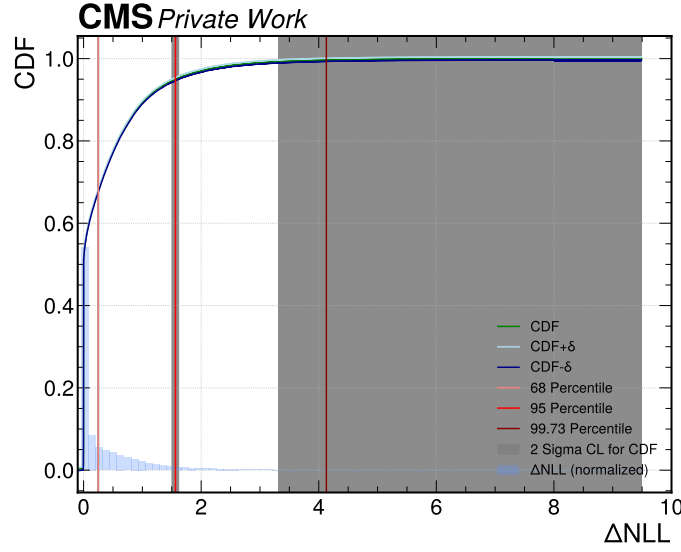


FIGURE 6.5: ΔNLL distribution with the according cumulative distribution. The $\pm\delta$ CL band around the CDF is drawn with the DKW inequality of equation 6.9. The 68, 95 and 99.73 percentile are shown with grey shaded 2SD bands derived from the uncertainty on the CDF.

The shape of the $m_A = 1000$ GeV $w_A = 2.0\%$, $m_H = 750$ GeV $w_H = 2.0\%$ scenario notably shows how the contour derived from shared $\{0,0\}$ toys also accommodates an observed best fit which is incompatible with 0 at 1 SD but compatible with 0 at 2SD. Here, also one can observe the effect of the DKW error band. It covers a larger area where the FC result shows characteristic instabilities. The effect of toy statistics impacts this region more, explaining why the FC contour is not smooth. Overall, the $\{0,0\}$ contours are smoother as a finer grid can be sampled without a large computational cost. The shared toy method however overestimates the excluded regions in the shown cases. It is particularly apparent in the 68% CL lines of the shown signal hypotheses. The excluded coupling values of $g_{A\bar{t}t}$ and $g_{H\bar{t}t}$ differ by 10-20% along the axes of the plots. The observed deviations can be pinned to the underlying toy distribution. The background only hypothesis changes for a given signal fit, as the NPs will be fixed to different values. This is quantified in the pulls and constraints and the impact on the toy distribution is visualized in figure 6.7. The left plot shows NP distributions sampled from toys drawn at $\{0,0\}$ compared to toys sampled from a signal point where $g_{A\bar{t}t} = g_{H\bar{t}t} = 1$. Most NPs will vary only slightly with the signal hypothesis but as evident from the electroweak Yukawa NP, some distributions can be shifted. This effect is then propagated into the $2\Delta\text{NLL}$ distribution which is shown in the right plot. The toy distribution at $\{0,0\}$ peaks at 0 while the respective distribution sampled at $\{1,1\}$ is shifted to higher values of $2\Delta\text{NLL}$. A limit based on this distribution will thus be weaker. This effect drives the more conservative contours of the FC approach in figure 6.6. It is less pronounced at higher percentiles of the $2\Delta\text{NLL}$ distribution as there the tails of the toy distributions are more similar.

Redoing the fit of the shared toys with the signal hypothesis does not significantly change the $2\Delta\text{NLL}$ distribution as apparent in the plot. Refitting only has a miniscule effect on the contour because the NPs are determined during toy sampling. It does not remedy the differences in the contours.

Setting limits based on the background-only hypothesis should be viewed as a different statistical hypothesis test. It tests the compatibility with the background model at the $\{0,0\}$ point and can thus yield different results compared to FC hypothesis testing. However, in many cases, especially with small pulls and constraints on NPs, it can be used as an approximation of the FC procedure. All results derived here are based on 130000 $\{0,0\}$ toys and the respective profile likelihood fits. This is fewer toys than a single FC contour. Computationally, the procedure is thus comparable to the asymptotic result while retaining the advantage to not rely on the assumptions of Wilks theorem.

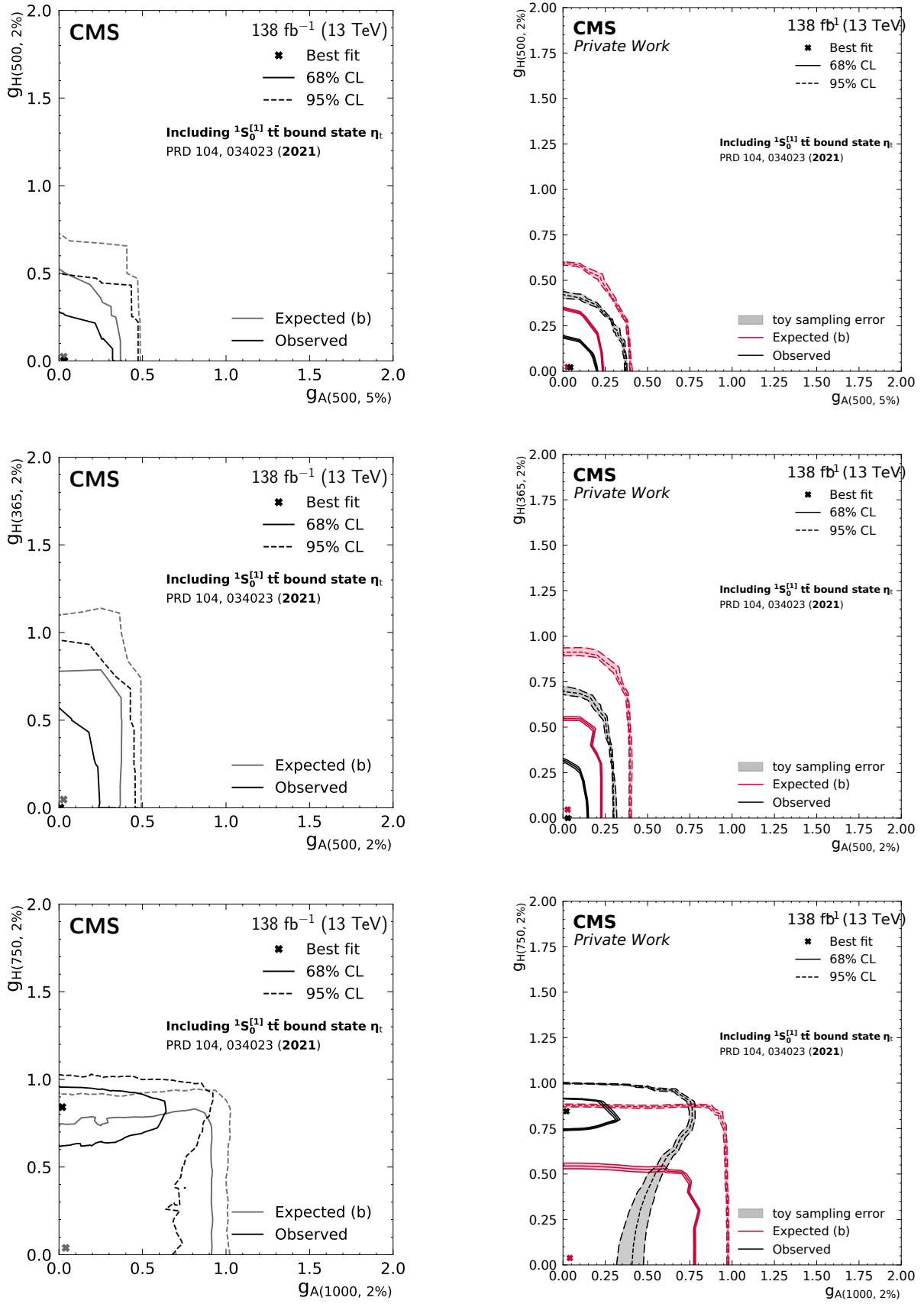


FIGURE 6.6: Contours derived with the FC procedure (left column) and the shared toy procedure outlined in the text (right column).

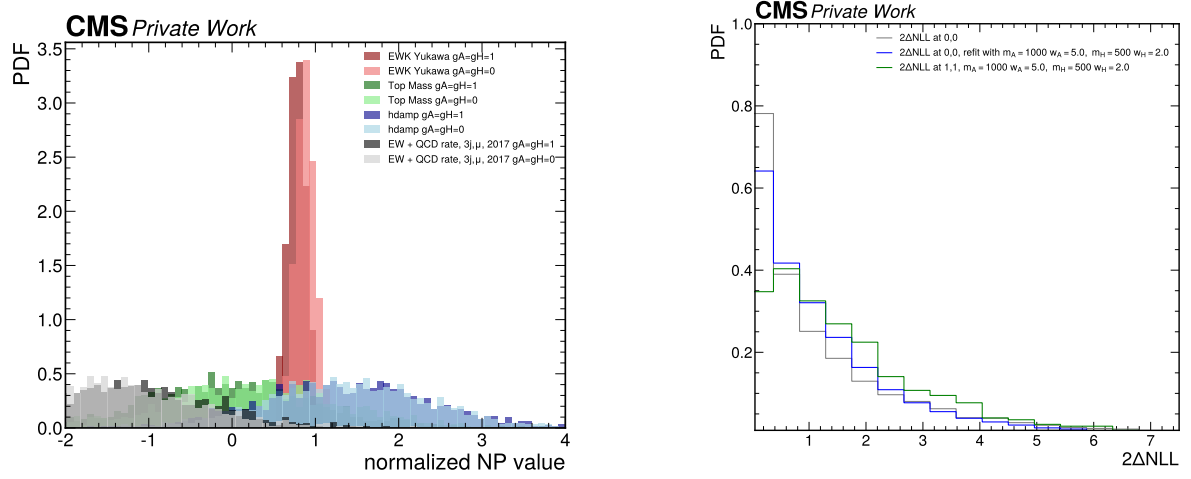


FIGURE 6.7: The left panel shows the probability distribution function of the normalized NP values for different systematic uncertainties at the $\{0,0\}$ and the $\{1,1\}$ point. Most systematics are not dependent on the value of $g_{A\bar{t}t}/g_{H\bar{t}t}$ at which they are sampled but the electroweak Yukawa nuisance shows a different peak based on the signal hypothesis value. The right panel shows the $2\Delta\text{NLL}$ distribution of a set of toys drawn at the background-only hypothesis (grey), at the signal hypothesis $g_{A\bar{t}t} = g_{H\bar{t}t} = 1$ for $m_A = 1000$ GeV, $\Gamma_A/m_A = 5\%$, $m_H = 500$ GeV, $\Gamma_H = 2\%$ (green) and drawn from the background-only hypothesis but refit with the same signal hypothesis (blue). The refit toy distribution lays between the other two distributions but is closer to the background only hypothesis.

6.3 CP VIOLATION IN HEAVY HIGGS TO $t\bar{t}$ DECAYS

In the follow section the search described in section 6.1 will be generalized to a scenario where the CP eigenstates A and H can mix. The theory of this model is described in section 2.4.1.1. Because of the unique challenges in this scenario, a new analysis method is introduced in form of SBI which is described in theoretical detail in section 4.2.3.

6.3.1 Phenomenological Study of CP Violation in $t\bar{t}$ Events with Simulation Based Inference Tools

The simplified model for CP violating additional heavy scalars has a highly multidimensional parameter space. It is thus infeasible to study with common analysis tools as the curse of dimensionality introduces high computational costs. With two masses, two widths, four coupling modifiers and Z-matrix mixing with an additional 4 complex parameters, the theory space is 16-dimensional. The computational cost of a grid of MC samples across this space scales exponentially with the number of parameters and it is thus prohibitive to generate a covering set of simulations.

Still, one can deduce interesting areas of parameter space and learn about the discrimination power of variables by conducting a study based on SBI. The SBI method and the MadMiner framework are described in detail in 4.2.3.

6.3.1.1 Simulation and Inference Setup

For this study, the CP violating simplified heavy scalar model from Ref. [45] is taken in the "Universal FeynRules Output" (UFO) model format [182]. It includes the desired processes in leading order in α_s .

MadMiner handles the computational cost of MC sample production by "generator morphing" which was inspired by effective field theory (EFT) considerations. If a likelihood can be factorized into a factor only dependent on the POI and factors dependent on the momenta z it can be computed from a few basis points θ_{basis} and morphed into a full grid of predictions with the factorization equation

$$p(x|\theta) = \sum_c \tilde{w}_c(\theta) f_c(z). \quad (6.10)$$

Here, the sum runs over the amplitude components c and the POI dependence is absorbed into $w_c(\theta)$ while the momentum dependence is absorbed into $f_c(z)$. A factorization of this form for the coupling modifier of the CP violating theory is presented in equation 2.31. The dependence on the coupling modifiers $c_{t,1}, \bar{c}_{t,1}, c_{t,2}, \bar{c}_{t,2}$ is factored out from the differential cross section $d\sigma$. The theory does not factorize equally trivially in the mass, width and Z-mixing parameters. For this study, a benchmark point in these parameters was thus chosen, given in table 6.4. Thus, the POIs here are the CP-even and CP-odd coupling modifiers between the additional heavy scalars S_1

Theory Parameter	MadGraph Name	Benchmark Value
Mass of Scalar 1	MASS 35	500 GeV
Mass of Scalar 2	MASS 36	600 GeV
Width of Scalar 1	WIDTH 35	50 GeV
Width of Scalar 2	WIDTH 36	60 GeV
$Z_{\Phi_1, \text{tree}}^{\Phi_1}$	FRBLOCK 9	1.0
$Z_{\Phi_2, \text{tree}}^{\Phi_1}$	FRBLOCK 10	0.0
$Z_{\Phi_1, \text{tree}}^{\Phi_2}$	FRBLOCK 11	0.0
$Z_{\Phi_2, \text{tree}}^{\Phi_2}$	FRBLOCK 12	1.0

TABLE 6.4: Parameter values for the benchmark point considered in the MadMiner study of CP violating heavy scalars.

and S_2 and $t\bar{t}$: $\{c_{t,1}, c_{t,2}, \bar{c}_{t,1}, \bar{c}_{t,2}\}$ which were limited to the range $[-1,1]$. Building a morphing setup with these POIs and allowing for interference terms between the signals, signal-SM interference and resonant contributions from S_1, S_2 and SM $t\bar{t}$ production, yields a morphing basis of 70 points with MadMiners automation ². A method to verify that the chosen morphing basis is optimal, is to calculate the morphing weights across the parameter space. The statistical error of MC samples scales with the inverse squareroot of the sample size, but the individual uncertainty at a parameter point gets multiplied by the morphing weight. Subsequently, the relative statistical error of a histogram bin σ_i is dependent on the morphing weight $w_c(\theta)$ and the number of events in that respective bin n_i via

$$\frac{\sigma_i}{n_i} = \frac{w_c(\theta)}{\sqrt{n_i}}. \quad (6.11)$$

² To reconstruct the theory prediction for arbitrary values of the 4 parameters considered here, the morphing basis must span all independent terms appearing in the squared matrix element up to fourth order. The number of such terms corresponds to the number of monomials of total degree ≤ 4 in 4 variables, given by $\sum_{d=0}^4 \binom{4+d-1}{d} = 1 + 4 + 10 + 20 + 35 = 70$.

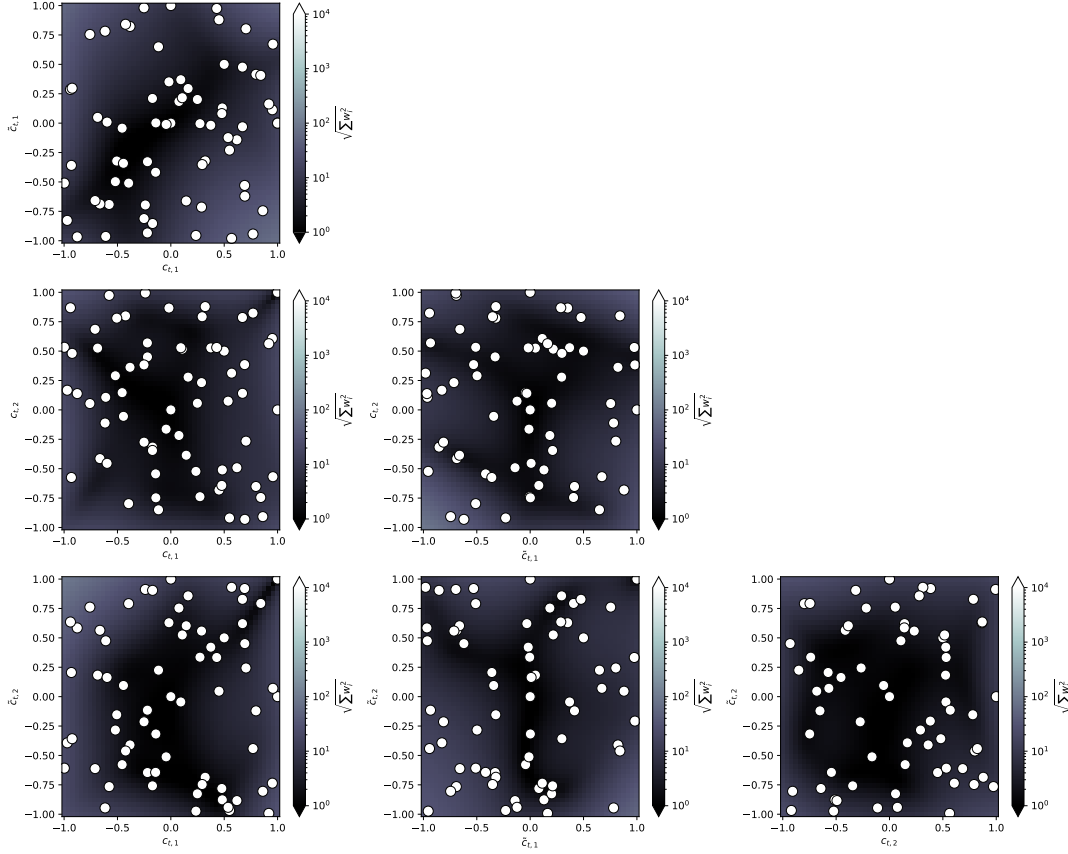


FIGURE 6.8: Morphing basis of the CP violating heavy scalar model MadMiner study. The basis points in the 4D POI parameter space were chosen by the MadMiner morphing algorithm and are shown as white circles. The grey-scale colors show the estimation of the statistical errors by showing the morphing weight values $\sqrt{\sum w_i^2}$.

Thus, large weights will translate into a large statistical error. Figure 6.8 shows slices through the parameter space with the automatically chosen benchmark points marked. The colormap visualizes the weights, which reach $\mathcal{O}(100)$. The average weight is generally smallest around the SM benchmark point $c_{t,1} = c_{t,2} = \tilde{c}_{t,1} = \tilde{c}_{t,2} = 0$. The derived results will thus be most dependable for small deviations around the SM. This is in line with observations in EFT scenarios [183].

With a basis for the cross section calculation set, events need to be generated. MadGraph5 V2_9_14 [184] was used for generating 22,000 events for each benchmark point. This results in ~ 1.5 million events across the whole parameter space. The generation command `gg > tt (t > bw+, w+ > l+vl), (t > bw-, w- > l-vl)` was used. It includes all signal and background diagrams contributing to $t\bar{t}$ production with a dileptonic decay. The matrix element is evaluated at 13 TeV with the NNPDF2.3 pdf set including QED corrections [185]. The parton showering is subsequently executed by Pythia8.306 [186].

An analysis on parton level is implemented with the MadMiner ‘analysis’ subpackage. It involves 10% smearing for generator level quantities (energy, p_T , η , ϕ for all particles) to account for detector resolution and efficiency. Further detector effects were foregone in the rest of the work. A set of variables is later used for training the networks, together with the parton level cross sections. In this work, the spin sensitive variables from table 2.1 are used as well as some kinematic variables ($m_{t\bar{t}}$, the leading jet p_T and p_T^{miss}) are included. Some derived distributions are shown for some benchmark points in figure 6.9. Note, that the shown distributions are normalized to better visualize shape effects. Cross section effects are of the magnitude $\mathcal{O}(1\%)$. Since the shown distributions are sampled from the benchmark points, they do not have an additional morphing weight uncertainty. The produced 22,000 events in 10 bins translate into a per bin statistical uncertainty of $\sim 0.05\%$. Shown are the SM benchmark, a scenario with two pseudoscalar particles, the 2HDM case with one scalar and one pseudoscalar particle and a mixed scenario with two heavy particles with mixed couplings ($c_{t,1} = c_{t,2} = \tilde{c}_{t,1} = \tilde{c}_{t,2} = 0.5$). The $m_{t\bar{t}}$ variable offers the most sizeable shape effect between the scenarios but it is susceptible to null-signatures for scenarios across the theory space as discussed in [45]. c_{hel} is also well established from experimental study and known to be suitable to distinguish (pseudo)scalar contributions from the SM. By construction, it is not suitable to detect CP violating contributions. The plot shows no slope for the mixed scenario as the pseudoscalar and scalar contribution to c_{hel} cancel. The sums of quadratic spin correlation matrix components can fulfill this task, but the parton level distributions already show very minor effects, the most promising feature can be seen in the $c_{nr} - c_{rn}$ variable, where a negative slope is visible for the mixed scenario. $c_{nk} - c_{kn}$ is also expected to exhibit sensitivity, as it is also P odd and CP odd. The analytical

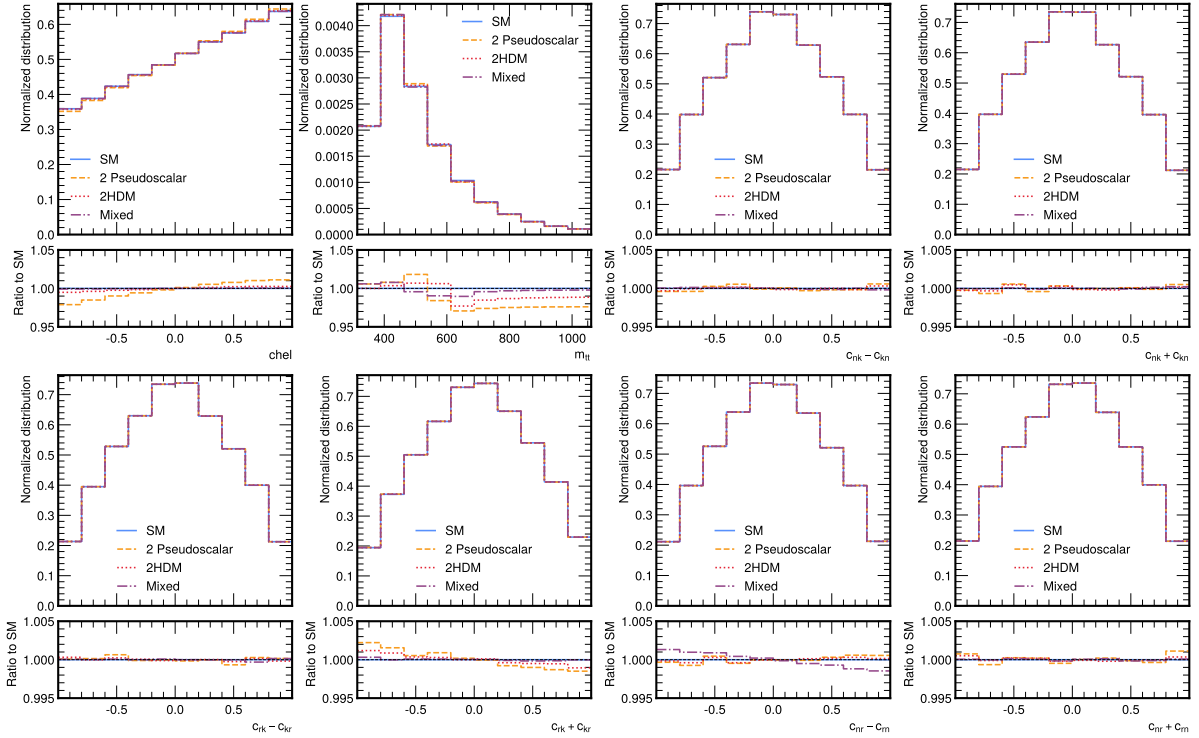


FIGURE 6.9: Distributions of the spin sensitive variables and $m_{t\bar{t}}$. Depicted are parton level quantities for the SM ($c_{t,1} = c_{t,2} = \tilde{c}_{t,1} = \tilde{c}_{t,2} = 0$), a benchmark point with two pseudoscalars ($c_{t,1} = c_{t,2} = 0, \tilde{c}_{t,1} = \tilde{c}_{t,2} = 1$), a benchmark point with the 2HDM configuration of one pseudoscalar and one scalar ($c_{t,1} = 0, c_{t,2} = 1, \tilde{c}_{t,1} = 1, \tilde{c}_{t,2} = 0$) and a mixed benchmark point ($c_{t,1} = c_{t,2} = \tilde{c}_{t,1} = \tilde{c}_{t,2} = 0.5$). The ratio plots are computed with respect to the SM histogram. Note, that the axis ranges for the ratio plots are smaller for the spin variables.

calculation in Ref. [19] predicts a slope which is roughly one quarter less prominent at 13 TeV than $c_{nr} - c_{rn}$. This explains why no effect can be observed for $c_{nk} - c_{kn}$ in this sample.

The training data includes all calculated observables and the added parton level cross section information used for data augmentation as described in 4.2.3. Adding observables with no apparent separation power in their histograms can be beneficial to network training due to correlations. A training set is constructed from the benchmark points. An ALICES, a SALLY and a SALLINO model [102] with a fully connected DNN as a backbone are trained

Hyperparameter	Value for ALICES	Value for SALLY & SALLINO
Learning Rate	1×10^{-5}	1×10^{-5}
Optimizer	ADAM	ADAM
Batch Size	128	128
Number of Layers	2	2
Nodes per Layer	(32, 64)	(128, 128)
Dropout	0.05	0.0
Training Epochs	50	20

TABLE 6.5: Training Setups for the ALICES and SALLY/SALLINO methods.

with hyperparameters detailed in table 6.5. A range of different hyperparameter setups was tried as overfitting was a problem leading to biased best-fit estimations and overconfident limits. The convergence with the given hyperparameters is deemed to be sufficient. The loss curves are visualized in figure 6.10 and show that a stable minimum has been reached. The trained models are then used to infer limits for a 139 fb^{-1} , 300 fb^{-1} and a 3000 fb^{-1} scenario. This is done to gauge the maximal sensitivity that can be reached with the Run 2, Run 3 and HL-LHC data respectively. 95% CL exclusion bands are drawn. For comparison, a histogram based likelihood, based on the c_{hel} and $m_{t\bar{t}}$ histograms is evaluated as well. This one uses the asymptotic approximation and is, as examined in section 6.2, flawed. It is still included for comparison.

6.3.1.2 Exclusion Limit Results at the SM benchmark Point

The 4D likelihood is evaluated on a grid with 10 bins in each POI direction for values between -1 and 1 and expected p-values are computed from that result as explained in section 3.3. At first, the assumed truth value was placed at

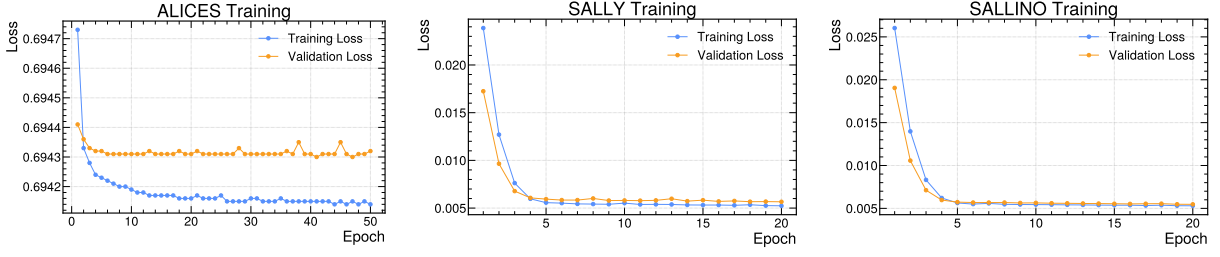


FIGURE 6.10: Loss Curves for the ALICES (left), SALLY (middle) and SALLINO (right) networks trained with the hyperparameter setup detailed in table 6.5.

the SM benchmark with $c_{t,1} = c_{t,2} = \tilde{c}_{t,1} = \tilde{c}_{t,2} = 0$.

As the result is four dimensional, it needs to either projected or profiled to show 2D contours. A profiling approach that maximizes the likelihood along the hidden axes was adopted to present the results in figure 6.11. The points marked with stars are the best fit values for the likelihood estimator encoded into the color. The resulting contour shows the excluded regions at 95% CL if no BSM effects are present, giving a measure of sensitivity to such processes in the respective datasets.

It can be seen, that the sensitivity is largest for the pseudoscalar coupling $\tilde{c}_{t,1/2}$ for any method except SALLINO. Variations in this coupling have a larger impact on the cross section. For lower masses, a higher cross section is expected so the sensitivity is larger for the couplings of the lighter S_1 than the corresponding coupling of the heavier S_2 . In fact, the aforementioned mass effect and the scalar to pseudoscalar effect are roughly equal in magnitude such that the exclusion contour is round in the corresponding plane of $c_{t,1} - \tilde{c}_{t,2}$.

For SALLY and ALICES, the exclusion power for the scalar coupling of the heavier particle S_2 is not sufficient at $\mathcal{L} = 300 \text{ fb}^{-1}$ to be contained in the tested parameter space up to $c_{t,2} = 1$. Values of $|\tilde{c}_{t,1}| > \sim 0.5(0.25)$, $|\tilde{c}_{t,2}| > \sim 0.75(0.35)$ and $|c_{t,1}| > \sim 0.6(0.25)$ are expected to be excluded for $\mathcal{L} = 300 \text{ fb}^{-1}$ ($\mathcal{L} = 3000 \text{ fb}^{-1}$) with the SALLY method. The exclusion contours are all following elliptic shapes with their half axes being aligned with the coupling axes. This shows that no cancellations of signatures happen into the CP violating directions for the trivial Z matrix mixings considered here.

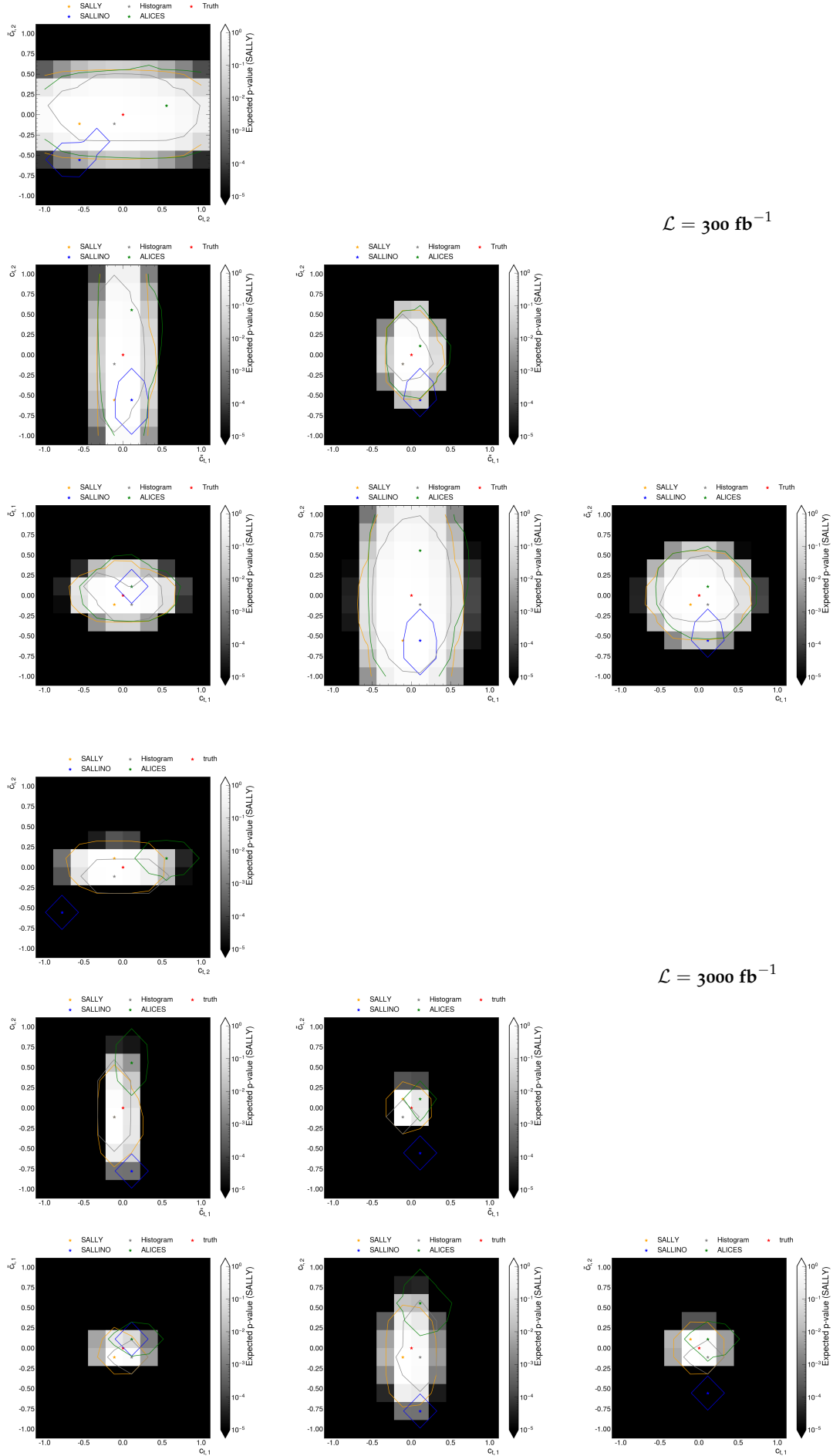
Qualitatively, this discussion holds for the histogram, SALLY and ALICES method for $\mathcal{L} = 300 \text{ fb}^{-1}$.

The SALLINO method fails to predict a sensible exclusion contour in all scenarios. The improved scaling to multidimensional spaces leads to a more confident estimate which shows to be biased. While the predicted best fit always falls into the 95% CL contour of the comparable SALLY method, SALLINO's estimate shows overconfidence in the excluded region, ultimately excluding the truth value.

ALICES shows a comparable issue for the HL-LHC forecast. The bias is largest along the axis of least sensitivity $c_{t,2}$. The SALLY method shows the most stable performance, tracing the shape of the histogram method. Notably, the limits of the SALLY method are more conservative than the histogram. This is unexpected as the SALLY method has more information at its disposal. As noted before, the histogram method is relying on Wilks theorem which is violated in this scenario. As the overall shape seems plausible it is likely that this only has a quantitative effect and the histogram method is overestimating the exclusion region. Overcoming the asymptotic approximation with a FC approach is not viable as it is already prohibitively expensive in the lower dimensional parameter space of the A/H analysis presented earlier in section 6.2.

In addition to the data preparation and training method presented here, a training set was constructed by reweighting and usage of the morphing algorithm. For this, the MC events generated at the benchmark points were reweighted to a set of sampled points across the theory space. The morphing algorithm was then used to calculate the information used for data augmentation. It was found that sampling from the benchmark points as done for the nominal results presented above improves stability. Limit results for the morphing method are shown in the appendix A.2.

These are the first projections for sensitivity to these couplings.

FIGURE 6.11: Profiled limits for 300 fb^{-1} (top) and 3000 fb^{-1} (bottom) for the SM ($c_{t,1} = c_{t,2} = \tilde{c}_{t,1} = \tilde{c}_{t,2} = 0$) scenario.

6.3.1.3 Exclusion Limit Results in a Scenario with CP Violation

Another important test of the methods is the behavior of the networks when encountering a signal. The limits were recomputed for an assumed signal scenario with the "Mixed" benchmark point with $c_{t,1} = c_{t,2} = \tilde{c}_{t,1} = \tilde{c}_{t,2} = 0.5$. This scenario has a CP violating contribution, so it is especially interesting to see if a CP non-violating scenario can be excluded in this case. Also, the trained SALLY and SALLINO networks assumptions are less correct here. The expansion used for constructing the likelihood from the trained model is evaluated away from the origin.

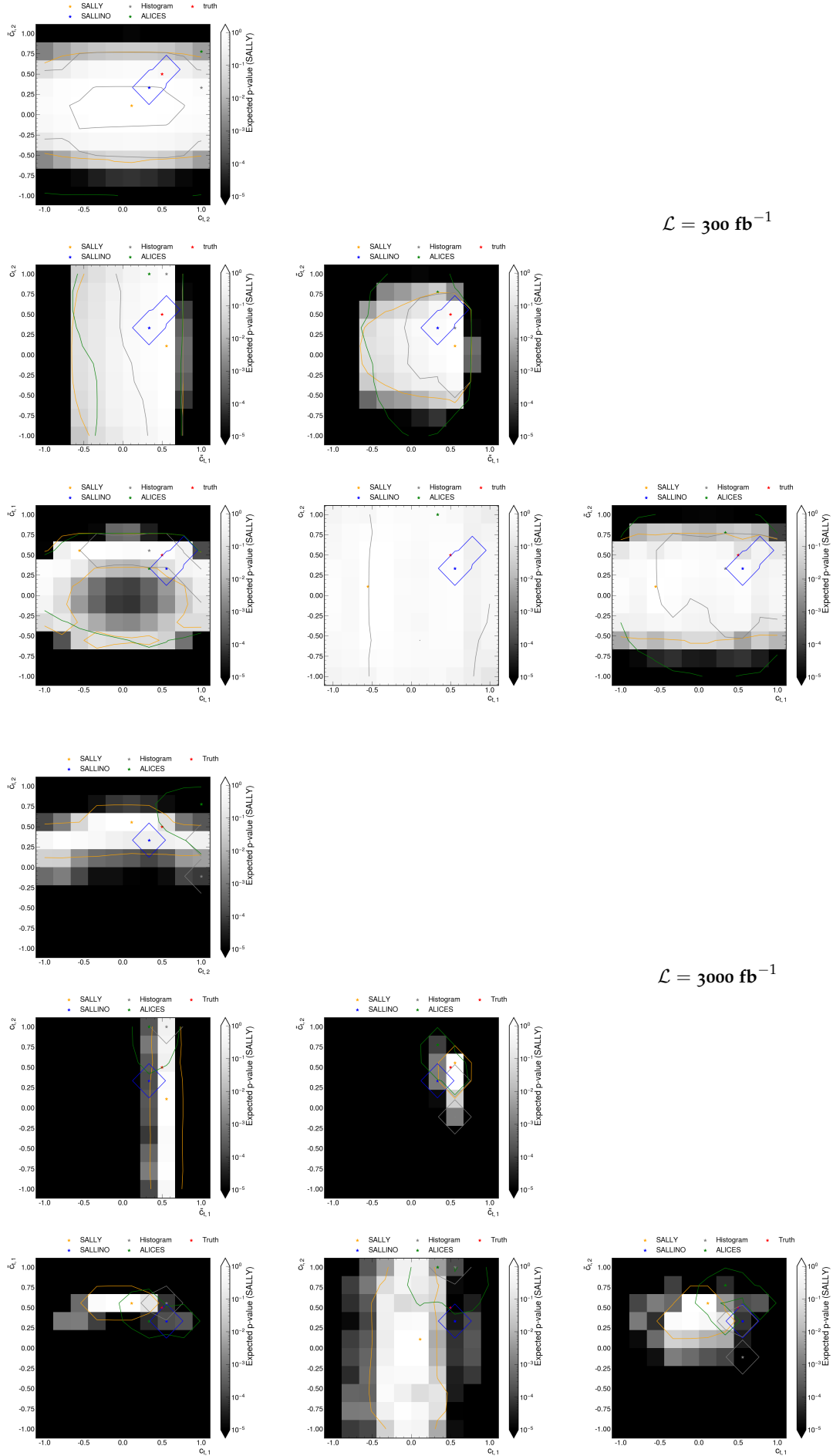
Two sets of profiled 2D contours are shown in figure 6.12. It is evident that the exclusion strength is smaller away from the SM point. For the scalar axes at 300 fb^{-1} luminosity no values on the tested grid can be excluded by SALLY or ALICES. Comparing to the histogram methods shape, once again SALLY is the most similar, while predicting less conservative contours. The exclusion contours seem to be generally symmetric around the SM point of $c_{t,1} = c_{t,2} = \tilde{c}_{t,1} = \tilde{c}_{t,2} = 0.0$. This can be explained by the fact that at 300 fb^{-1} the minor effects of the CP sensitive variables are not yet decisive for sensitivity and the cross section effects are stronger. With more integrated luminosity, the information from the additional variables can break this ambiguity. Indeed, the finding breaks down in the HL-LHC projection. The histogram method shows large biases away from the truth point, even excluding the injected signal. This is further evidence that the asymptotic approximation indeed is wrong. It has to be noted that while SALLINO is once again the most confident method, it is less biased in its predictions than for the SM scenario described in the previous section. For particles with mixed couplings it rarely excludes the injected signal. This seems to happen by chance as further tests with other scenarios away from the background-only case do not support this performance. While again ALICES and SALLY trace similar shapes in the 300 fb^{-1} plots, ALICES differs for higher luminosities. Especially in the scalar directions it is more confident than SALLY while introducing a one-sided exclusion, excluding smaller values in $c_{t,1/2}$. It is expected that SALLY, because of its local expansion, is not as powerful in directions of low information. Indeed, it shows symmetric contours around the origin in all directions except the pseudoscalar ones. The $\mathcal{L} = 3000 \text{ fb}^{-1}$ results show that there is sensitivity to exclude the SM case. This holds across methods and thus gives hope for further study of this process in the HL-LHC era. Overall, especially at the HL-LHC forecast, the results do not appear to be stable. Large biases and overconfidence in the excluded regions can be observed. A full analysis based on the current implementation can be ruled out as the instability is expected to be more impactful for a model with comparably complicated systematical uncertainties as the search for CP even or CP odd heavy scalars. Since the time of this study, the issue has been acknowledged by theorists and pinned to the morphing algorithm introducing unpredictable uncertainties in high dimensional parameter spaces. This can be caused by large local variations in weights that do not show up in the diagnostics plots that have been presented here. An improved morphing aware approach to simulation based inference as proposed in [187] can further enhance performance and most importantly overcome stability concerns when working in high dimensional spaces. The authors of that work also make use of the better scaling of physics informed LGATR transformers as DNN regressors. The results shown here are thus to be understood as a preliminary study gauging what current methods can do in theory models that still need to be deemed too complex to be explored without further assumptions. Further limiting the parameter space of the CP violating model explored here is not a priori motivated but necessary for a full experimental analysis.

6.3.1.4 Information Content of CP sensitive Variables

The following shall still give recommendation for further experimental exploration based on the information present in this study. The framework of the Fisher information matrix is introduced in section 4.2.3. The Fisher matrix is evaluated for the full information with the SALLY method and individually for 1D histograms of the spin variables to compare their discriminating power. The full shape and cross section information matrix evaluated at the SM point with the SALLY method is

$$\mathcal{I}_{ij}(\text{SM}) = \begin{pmatrix} 0.6710 & -0.1691 & -0.0266 & 0.0013 \\ -0.1691 & 1.0000 & 0.0526 & -0.1436 \\ -0.0266 & 0.0526 & 0.7092 & -0.1653 \\ 0.0013 & -0.1436 & -0.1654 & 0.8007 \end{pmatrix}. \quad (6.12)$$

For easier interpretation, these values have been normalized to the maximum. The first row relates to the $c_{t,1}$ parameter, the second row to $\tilde{c}_{t,1}$, the third row to $c_{t,2}$ and the fourth row to $\tilde{c}_{t,2}$. The higher the Fisher information, the more precise a parameter can be estimated. One can relate the limit results from figure 6.11 to the results of the matrix 6.12. As already evident in the profiled limits, the highest information is accessible for the less massive particles' S_1 pseudoscalar coupling $\tilde{c}_{t,1}$. There are large correlations to the scalar coupling of the same particle

FIGURE 6.12: Profiled limits for 300 fb^{-1} (top) and 3000 fb^{-1} (bottom) for the "Mixed" ($c_{t,1} = c_{t,2} = \tilde{c}_{t,1} = \tilde{c}_{t,2} = 0.5$) scenario.

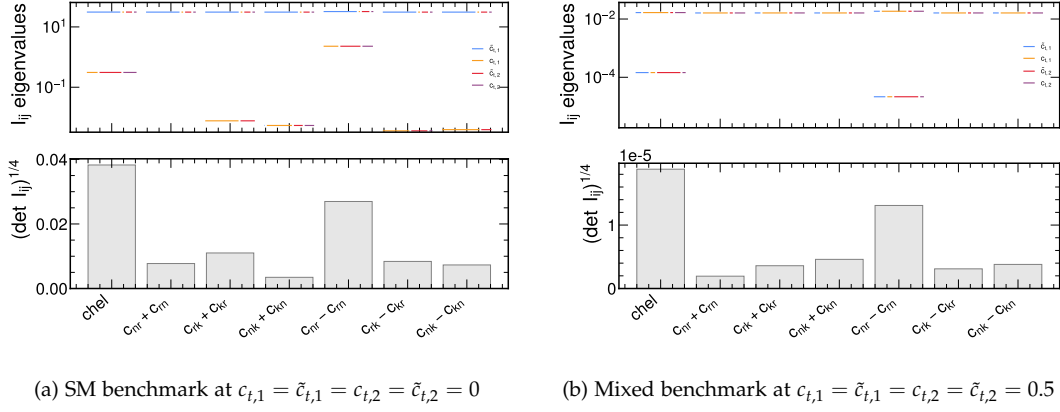


FIGURE 6.13: Fisher information decomposition. The lower plots show the determinant for each Fisher information matrix computed with the indicated variable. The upper panel shows the eigenvalue decomposition of the matrix with the color bars in each eigenvalue indicating the component magnitudes of the corresponding eigenvector. All matrices were normalized by the determinant of the full truth level information matrix before further processing. Only two or one of the four eigenvectors are visible for each of the 1D fisher matrices on the scale shown here.

$c_{t,1}$ and the pseudoscalar coupling of S_2 $\tilde{c}_{t,2}$. The eigenvalue decomposition of equation 6.12 with the respective eigenvectors v is

$$\lambda_1 = 1.1602, \quad v_1 = \begin{pmatrix} 0.2925 \\ -0.7996 \\ -0.2744 \\ 0.4469 \end{pmatrix} \quad (6.13)$$

$$\lambda_2 = 0.8457, \quad v_2 = \begin{pmatrix} -0.3252 \\ 0.4274 \\ -0.5483 \\ 0.6410 \end{pmatrix} \quad (6.14)$$

$$\lambda_3 = 0.5498, \quad v_3 = \begin{pmatrix} -0.5559 \\ -0.3168 \\ -0.5476 \\ -0.539 \end{pmatrix} \quad (6.15)$$

$$\lambda_4 = 0.6252, \quad v_4 = \begin{pmatrix} -0.7069 \\ -0.2784 \\ 0.5694 \\ 0.31402 \end{pmatrix} \quad (6.16)$$

The eigenvectors describe combinations of variables that can be estimated most precisely together and the eigenvalue gives a reference for the information content of that particular combination of variables.

To study the discriminative power of individual variables, the fisher information of one dimensional histograms are plotted in figure 6.13a for the SM benchmark and in figure 6.13b for the mixed parameter point. They are normalized to the determinant of the corresponding full information matrix. The determinants are shown as an indicator for the general sensitivity that this variable offers. The upper panels show an eigenvalue decomposition with the colors indicating the magnitude of the components of the corresponding eigenvector. This allows to read off sensitivities to combinations of POIs. Notably, c_{hel} and $c_{nr} - c_{rn}$ offer the best sensitivity. While c_{hel} has a good separation power between pseudoscalar BSM contributions and SM, it does not have sensitivity to the CP violation itself by construction. $c_{nr} - c_{rn}$ has separation power to CP violating scenarios, so a combination of the two variables can offer a good selection of variables for experimental study. $m_{t\bar{t}}$ is not pictured as it is several times more powerful and well studied.

It has to be noted that the 1D histograms are far less powerful for non-vanishing couplings compared to the SM case and the most sensitive parameter combinations change considerably. It is unexpected to see $c_{t,1}$ having a higher sensitivity than $\tilde{c}_{t,1}$ at the mixed benchmark. This is also not observed for the full SALLY information matrix (see appendix A.2), showing the need for higher dimensional inference in this area.

6.3.1.5 Computational Aspects

To conclude this study, the aspect of computational cost needs to be mentioned. As it has been the critical problem in the case discussed in section 6.2 it is natural that this higher dimensional parameter space comes at an additional expense. Because of that, statistically better suited FC contours have not been considered. The evaluation of the likelihood in 4 POIs with the asymptotic approximation is comparatively cheap and was executed on a DESY Maxwell CPU³ node in ~ 100 s. SALLY and SALLINO are more expensive to evaluate with ALICES being the most costly. The exact times are given in table 6.6. One can estimate the FC approach to multiply the histogram

	Histogram	SALLINO	SALLY	ALICES
Inference time [s]	126 ± 28	131 ± 0.5	215 ± 19	572 ± 100

TABLE 6.6: Inference times for likelihood calculation with 4 POIs, the average was taken over 5 computations.

inference time with the number of grid points. In the considered scenario with 10 bins in each POI, it would thus consume 100×10^4 s. For the ML methods one has to factor in training time. This computation time is an acceptable overhead, as only one training is needed for repeated inference in scenarios concerning different injected signals. The considered models in this study were all small enough to be trained in less than 10 minutes on DESY Maxwell GPU nodes equipped with NVIDIA Tesla P100 [188].

The major bottleneck for future analyses in high dimensional parameter spaces will be the training data. MC simulations can effectively be parallelized but the computational effort for high dimensional parameter grids still ultimately scales exponentially. Indeed, this study would have been infeasible if on top of the four coupling parameters also training data in the directions of Z-matrix mixing and mass and widths of the particles would have been needed.

³ Using a AMD EPYC 7402 24-Core Processor Node with 2.8 GHz and 8 GB of memory [188]

6.4 TOWARDS NEXT GENERATION ANALYSES

With more data from LHC run 3 and advancements in resolution from CMS detector upgrades, future analysis will enhance the sensitivity to SM and BSM effects in the dileptonic $t\bar{t}$ channel. To maximize these benefits and reach optimal performance, also the analyses will need to be optimized. A major concern in the search for heavy (pseudo)scalars is the mass resolution. It is severely affected by the choice of reconstruction algorithm. The analytic top reconstruction described in 6.1.1.5 delivers a well studied baseline but its performance is a bottleneck in the analysis. The reconstructed $m_{t\bar{t}}$ variable has a resolution of 15-25%, making it impossible to discriminate between η_t and BSM pseudoscalar contributions. A future approach to solve this issue could come from superior reconstruction algorithms in the form of DNNs. This chapter will provide an idea based on physics informed transformer networks and establish its potential.

To study this, relative bias and resolution will be considered as well as the predicted distributions. The relative bias $\mathcal{B}$ of an observable x is defined as

$$\mathcal{B} = \frac{1}{N} \sum_i \frac{x_{i,\text{predicted}} - x_{i,\text{true}}}{x_{i,\text{true}}}. \quad (6.17)$$

and gives a relative metric for the average difference between the predicted and true value. Here, N denotes the sample size over which the bias is computed and the sum i runs over the individual events with the prediction $x_{\text{predicted}}$ and the truth value x_{true} for the observable x . The resolution $\mathcal{R}$ is defined as

$$\mathcal{R} = \sqrt{\frac{1}{N} \sum_i \left(\frac{x_{i,\text{predicted}} - x_{i,\text{true}}}{x_{i,\text{true}}} \right)^2 - \mathcal{B}^2}. \quad (6.18)$$

These metrics are important in conjunction. Accurate distributional reconstruction is needed as large shape effects can influence the analysis result when fitting shapes. However, correct shapes with bad resolution and high bias can lead to washing out of key features also negatively impacting the sensitivity of the analysis. A comprehensive assessment of reconstruction quality therefore requires balancing global agreement in distributions with precise, unbiased event-level performance.

Using multivariate techniques to solve the problem of top reconstruction in the dileptonic channel is a powerful idea, as the problem of particle assignment and neutrino momentum can be solved simultaneously to maximize the overall performance. The analytic approach in section 6.1.1.5 also introduces artificial constraints which are necessary to find an analytic solution but can be violated and will then result in incorrect predictions. A DNN solves this issue and can combine the aforementioned benefits.

A study on a fully connected DNN architecture has already been presented in Ref. [98]. It features a two-fold design, with a prior DNN network which identifies and matches the b quark jets to their leptonic partners. With the output of this b jet classification network a second stage is fed that predicts the top and antitop momenta as an 8-dimensional vector. This simultaneous method was chosen as the per event observables, such as MET inherently correlate the output of the top and the antitop momentum. The networks architecture is based on fully connected layers: The b jet classifier is four layers deep with 200 nodes each and the momentum reconstruction network consists of 4 final layers with 800 nodes each. Here, the first two layers are used for the preprocessing of the leptonic and MET information and only the final 2 layers act on the output of the b quarks concatenated with the rest of the preprocessed event information. The full model has 3,353,706 parameters. It was found that this DNN can outperform the analytic method significantly, with the $m_{t\bar{t}}$ resolution at the $t\bar{t}$ threshold being improved from 25% to 10%.

Since these studies have been conducted, the class of transformer networks has seen an increase in usage in physics tasks. They scale well and the attention mechanism allows for information of short and long range correlations inside an event to be efficiently learned [88]. These advantages are amplified by physically motivated architecture choices of the LGATR network described in more detail in 4.1.5. Motivated by these considerations, a LGATR network was implemented for the task of top quark reconstruction in dileptonic $t\bar{t}$ events.

The training data used is similar to prior DNN studies in [98], containing 1 million dileptonic $t\bar{t}$ events generated with POWHEG v2 $h\nu q + \text{PYTHIA}$ and subsequent CMS detector simulation. They have passed through the analysis selection described in this chapter but cuts on lepton p_T , jet p_T , $m_{\ell\ell}$ and MET p_T have been omitted. The data was split into 700,000 training samples and 300,000 samples which were used for validation. The inputs to the network consists of fourmomenta of the leading 7 jets, the lepton and antilepton momenta as well as b and anti- b quark. The b quarks and leptons are thus implicitly already matched in the data with the analytic m_{lb} technique described. This was considered to be not disadvantageous as the network could learn to reassign this mapping if needed. MET was included as a fourvector with the 0th component being the p_T^{miss} , and the p_z component being left 0. Experiments have been carried out with the p_T^{miss} being passed as a scalar and the MET z component set to the corresponding η value. None of these techniques are inherently correct as MET breaks Lorentz symmetry.

Hyperparameter	Value for LGATR
Learning Rate	0.0001
Batch Size	256
Training Iterations	50000 $\approx$ 18 epochs
Optimizer	ScheduleFree [189]
Attention Blocks	12
Number of Heads per Block	14
Hidden Multivector Channels	12
Hidden Scalar Channels	10
Number of Parameters	409634

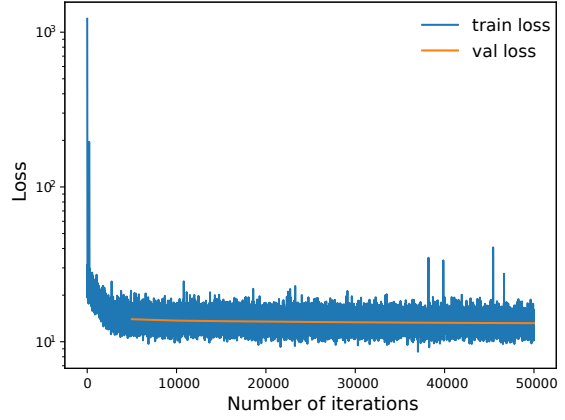
TABLE 6.7: Training Setup for the LGATR $t\bar{t}$ reconstruction network.

FIGURE 6.14: Convergence plot for the LGATR network showing the loss on the training sample (blue) and the validation sample (orange) as a function of the number of training iterations.

Setting the z component to 0 offers a trade-off between passing only correct information and retaining the leftover angular information present in MET x and y components. Inputs to the network are scaled by $1/4$ of the Minkowski norm of the fourvector. This is a Lorentz invariant method of prescaling inputs into $\mathcal{O}(1)$. It was done to overcome numeric issues with convergence as high target and input values introduced instabilities. The network outputs two multivector channels from which the first token is used. Then, the (anti-)top fourvector is extracted from this token. This is an arbitrary choice. By choosing one token during training, the activation of the other tokens will be 0 regardless as no gradient is generated.

Using two output channels ensures that each top can transform independently. Since they are connected until the final output layer, the transformation properties of the $t\bar{t}$ system are also retained. The network is trained with a MSE loss between the output quark fourvectors and the targets. The details of the network training can be taken from table 6.7. Compared to the fully connected DNN presented above, it is notably smaller, only using $\sim 10\%$ of the trainable parameters. With the multivector structure, the used GPU memory of ~ 16 GB during training is still at the limit that the DESY Maxwell GPU nodes offer.

The transformers convergence is smooth as depicted in figure 6.14. No indicators of overtraining are present, the validation loss falls into the stochastic noise of the training loss. The schedule free optimizer [189] was found to smoothen the convergence substantially.

The network is optimized on the single quark top and antitop kinematic quantities which are however not the targets of the analysis in this chapter. Thus, these distributions are shown in the appendix A.3 and they show quantitative and qualitative disagreements. As a more relevant metric with a higher degree of abstraction, the $t\bar{t}$ fourmomentum component distributions are shown in figure 6.15 combining the information of the top and antitop kinematics. There are sizeable shape effects visible. They are generally comparable to the DNN, except for the energy (0th component) which is modelled worse by LGATR. Shape disagreements in distributions are expected to a degree as the instance based MSE loss optimizes the per event quantity but not necessarily the distribution. At a high level of generalization, this should however be resolved. The x, y and z components are symmetric around zero and thus easier to model, the energy component has an asymmetric profile with a drastic fall-off close to the top pair production threshold. This is underestimated by the LGATR network while it overestimates the amount of events that fall into the energy range 500 - 700 GeV. It can be explained by the target MSE metric, which pulls events towards a mean and thus favors gaussian shapes over irregular, skewed distributions. The performance difference on the momentum components can be explained by their magnitude and thus their impact on the single events loss. In z direction, top quarks generally have more momentum. In fact, the average magnitude of the fourvector components in the evaluation dataset is $\sim 45\%$ 0th component, $\sim 31\%$ in z direction and $\sim 11\%$ in both x and y direction.

While the modelling of distributions is important, event reconstruction methods need to be judged on individual events as well. Figure 6.16 shows the standard deviation for the reconstructed spin correlation variables c_{hel} and c_{han} from their respective truth values as well as the resolution for the top quark p_T and $m_{t\bar{t}}$. The spin correlation results are given as the absolute standard deviation as their relative resolution results would be dominated by the $1/x$ normalization scaling close to 0. For the p_T and the $m_{t\bar{t}}$ only the ranges needed for analysis are considered and thus the resolution is shown. It is evident that for the top p_T and c_{hel} , the LGATR network performs better than the DNN and the analytical method. The c_{han} variable shows worse resolution by LGATR compared to the analytical approach. The DNN benchmark shows a steep gradient in resolution with c_{han} such that there is an area of c_{han} phase space that shows better resolution by LGATR. For $m_{t\bar{t}}$, LGATR outperforms the analytical method across the phase space. The DNN outperforms LGATR at low $m_{t\bar{t}}$ but their resolution is comparable for $m_{t\bar{t}} \geq 600$ GeV.

Figure 6.17 shows the bias results for the same variables. Generally, LGATR shows larger biases than the other methods. While the qualitative bias shapes agree for almost all variables, LGATR shows a particular effect in c_{hel} . The bias distribution has a positive peak for values between -0.75 and 0.0, indicating that events with these

c_{hel} values are systemically being overestimated. These events simultaneously feature a high standard deviation lowering confidence in the LGATR prediction in this region of phase space. As the relationship between the spin correlation variables and the fourmomenta is highly non-trivial, it is not an easy task to diagnose these effects. An explanation could however be that the shape effects in the energy component of the fourvectors and the unequal reconstruction performance on the p_T and the p_z part of the fourmomenta can lead to such behavior. The boosting into the $t\bar{t}$ and individual top and antitop restframes required for c_{hel} largely relies on the energy component. The angular relationship between the leptons in that restframe is sensitive to the inequality between the spatial components.

For the analysis discussed in this thesis, an unrolled search template in $m_{t\bar{t}}$, c_{hel} and c_{han} is used, which is shown

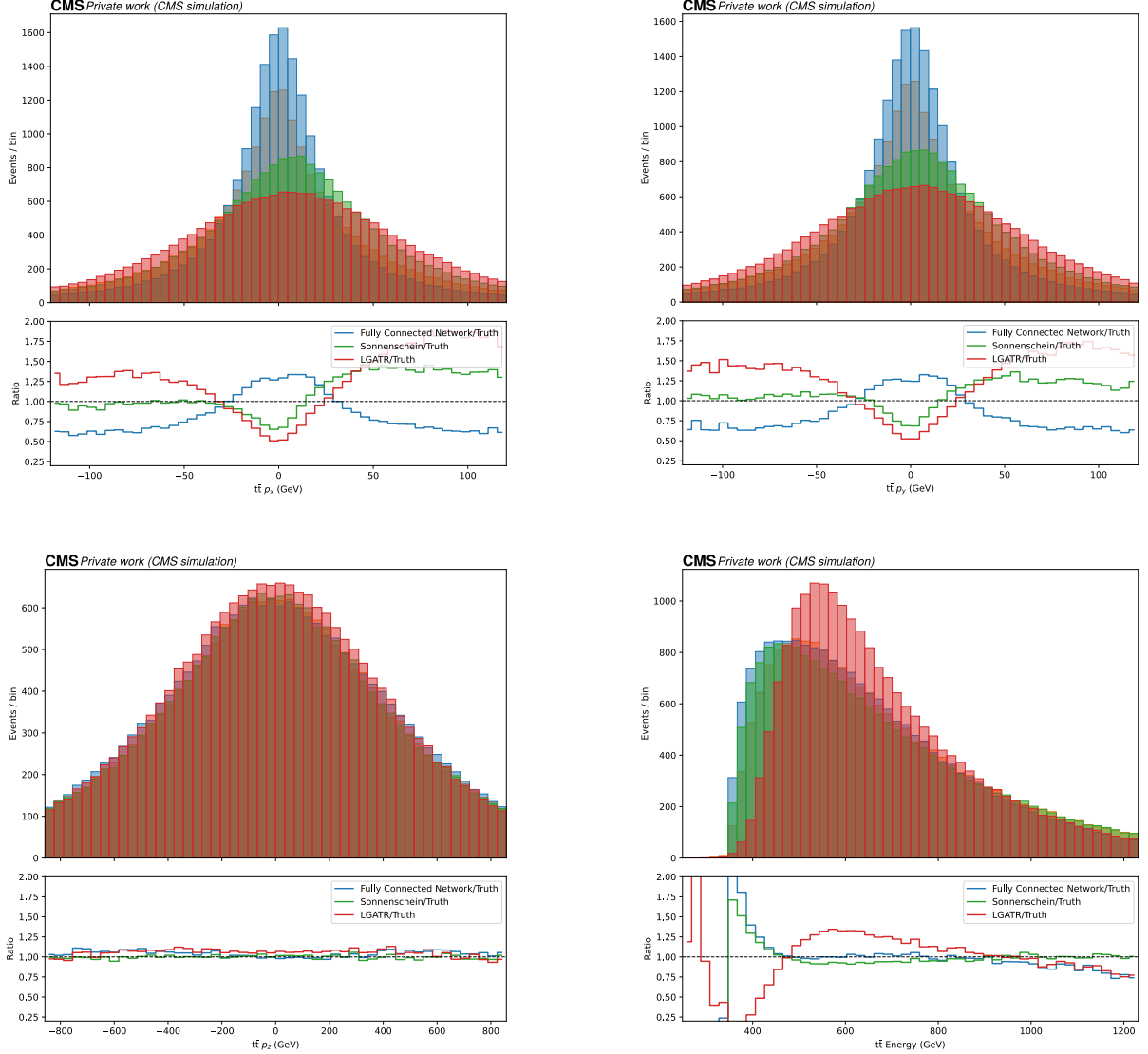


FIGURE 6.15: Prediction vs. truth of the momentum fourvector components for the $t\bar{t}$ system. The MC truth (orange) is shown against the analytic prediction [170] (green) and the DNN from [98] (blue) as well as the LGATR network trained in this work (red).

in figure 6.18. The resulting shapes show that the mismodelling of the spin correlations and the $m_{t\bar{t}}$ accumulates, the ratio to the truth is larger than in the one dimensional histograms. The $m_{t\bar{t}}$ shape of LGATR leads to a $m_{t\bar{t}}$ peak which is shifted with respect to the true distribution. This error is consistent in all spin correlation bins. The DNN on the other hand reconstructs the shape of the $m_{t\bar{t}}$ better but over- or undershoots the yield in the spin correlation bins. LGATR consistently models the height of the $m_{t\bar{t}}$ peak better across c_{hel} and c_{han} values. Both methods show large disagreements and thus motivate the usage of the analytic method in this analysis. The Sonnenschein method also shows disagreements but is most consistent overall, mostly only showing large disagreements with the truth for the sparsely populated high $m_{t\bar{t}}$ tails in the $c_{\text{hel}}-c_{\text{han}}$ bins.

To summarize, the LGATR study shows that more recent, powerful transformer architectures can improve reconstruction as already evident in some observables such as top quark p_T even with few trainable model parameters. As of now, this potential is hindered by major challenges that need to be overcome: The unequal performance on different fourvector components and the bias effects in predicted tops energy. Possible solutions to these problems

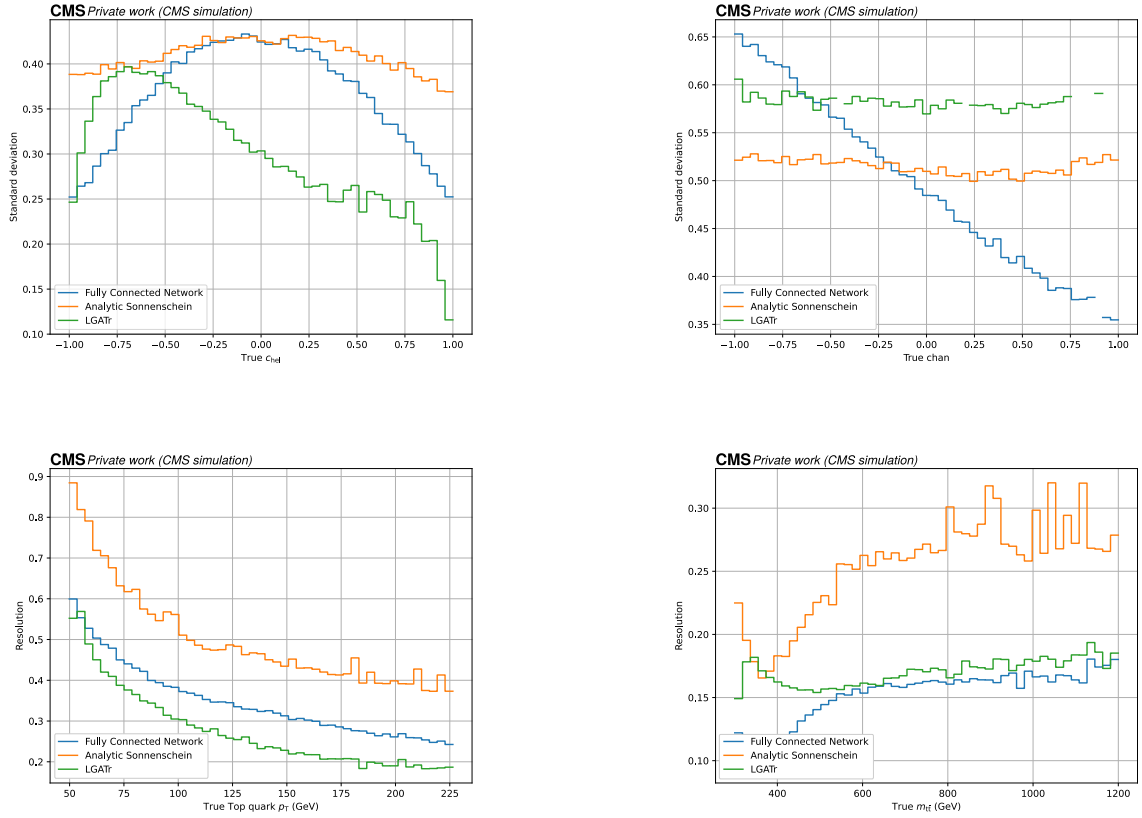


FIGURE 6.16: Resolution and standard deviation results of LGATR (green) vs DNN [98] (blue) and the analytic Sonnenschein algorithm [170] (orange) shown for some high level observables derived from the bare quark fourmomenta. The upper two panels show spin correlation results in c_{hel} and c_{han} . These plots are not showing relative values as these strongly scale with $1/x_{true}$ close to 0 and thus absolute standard deviation values give more insight. The lower panels show the top quarks p_T and $m_{t\bar{t}}$ in the desired range for an analysis.

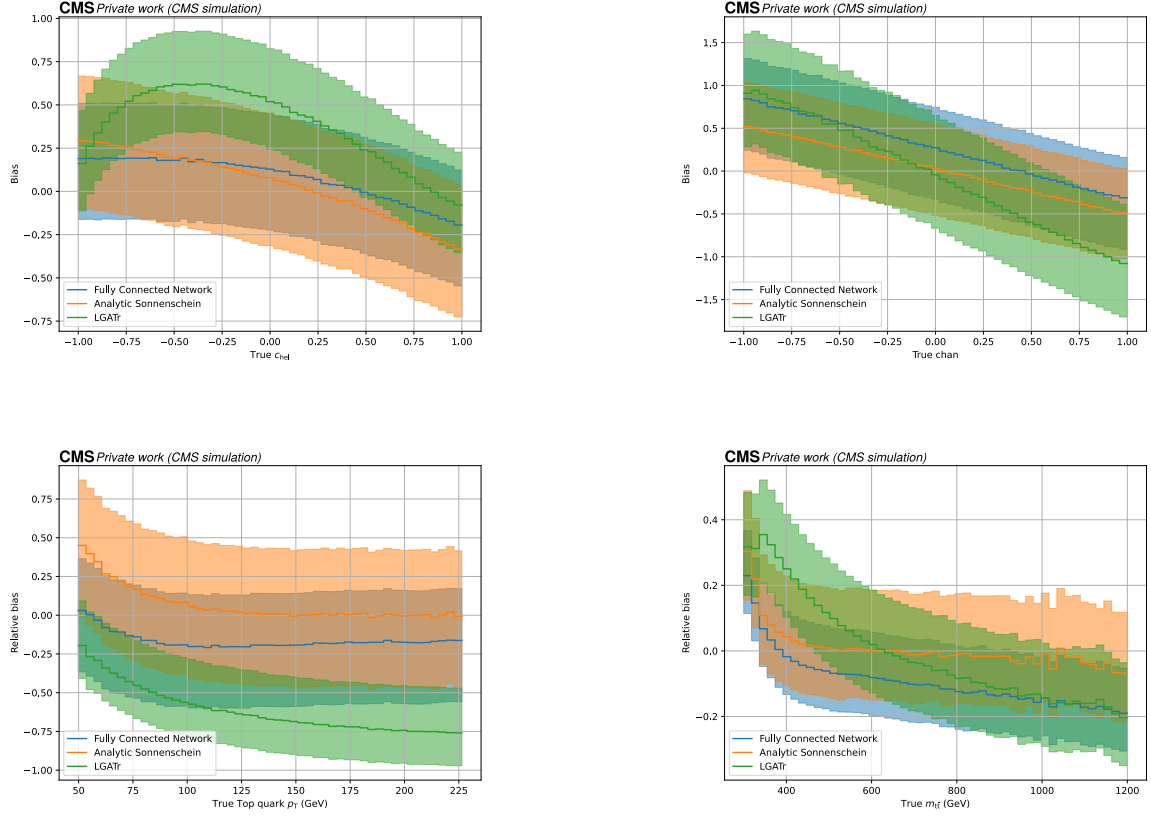


FIGURE 6.17: Bias results of LGATR (green) vs DNN [98] (blue) and the analytic Sonnenschein algorithm [170] (orange) shown for some high level observables derived from the bare quark fourmomenta. The upper two panels show spin correlation results in c_{hel} and c_{chan} . These plots are not showing relative values as these strongly scale with $1/x_{\text{true}}$ close to 0 and thus absolute bias values give more insight. The lower panels show the relative bias for top quarks p_T and $m_{t\bar{t}}$ in the desired range for an analysis.

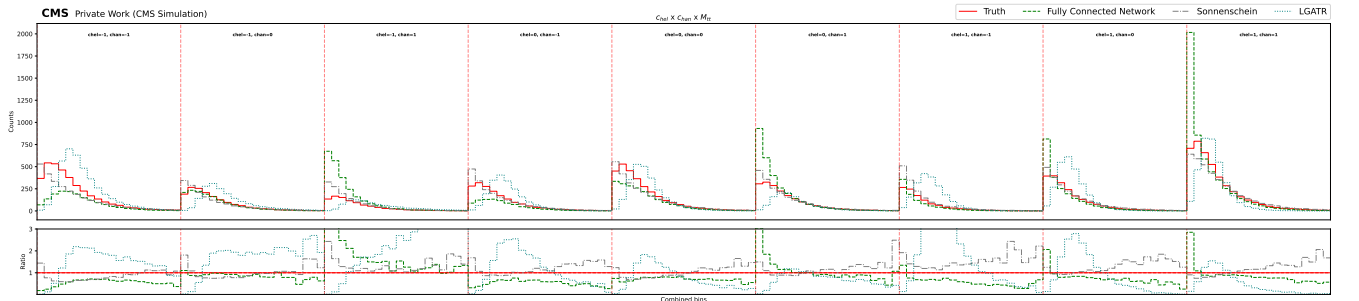


FIGURE 6.18: Reconstructed values of the search template of the A/H analysis, comparing LGATR (blue), DNN [98] (green) and the analytic Sonnenschein method [170] (grey) to the truth (red). The histogram is binned in 3 c_{hel} (-1 to -0.3, -0.3 to 0.3 and 0.3 to 1) and 3 c_{chan} bins each and 20 $m_{t\bar{t}}$ bins from 350 to 1200 GeV. The overall x-axis is given in combined bin indices. The upper panel shows the counts per bins whereas the lower panel shows the ratio to the simulated truth.

need to involve a prescaling method which keeps the full algorithm lorentz equivariant so that could overcome the magnitude difference in the momentum components. Solving this issue will remedy the angular issues that are apparent in the reconstructed spin correlation variables. Also loss functions beyond the MSE loss which is calculated as a per event difference could improve the performance on the energy component of the top momenta. Either the inclusion of further regulatory loss terms based on distributions or the full replacement of MSE with a more suitable target metric are possible ways forward.

SEARCH FOR AXION LIKE PARTICLES WITH THE CMS EXPERIMENT

The strong CP problem as well as the problem of DM have sparked the development of models with axion particles (or more generally, without necessarily solving the strong CP problem, axion like particles (ALPs)). A theoretical motivation has been given in section 2.4.2. Centrally, ALPs can be approximated as top-philic pseudoscalars. Their signature is thus very similar to the generic pseudoscalar boson A. As the only differentiation, they feature a direct coupling to gluons $c_{\tilde{G}}$ in the framework considered in Ref. [53]. For $c_{\tilde{G}} = 0$ the limits on the ALP coupling will only be rescalings of the constraints to A bosons.

These parallels justify utilizing the same analysis setup as described in section 6 to set experimental limits on the ALP model.

This chapter describes how the analysis setup from earlier section was used to infer limits on the ALP couplings to gluons and tops. In the case of vanishing gluon couplings, the results with 2016 data only are published in Ref. [53], so this work marks an extension of the available data. Additionally, limits are inferred for non-vanishing ALP-gluon couplings, which marks a new analysis result.

7.1 CREATING TEMPLATES VIA REWEIGHTING

Following the discussion in Ref. [53], it can be assumed that the final state phase space for an ALP is similar to the pseudoscalar A discussed above even in scenarios with non-vanishing $c_{\tilde{G}}$. This motivates the usage of a "post-mortem" reweighting algorithm over the generation of new signal MC samples for this search. Reweighting is a procedure where new weights for a dataset are computed and applied to represent a target distribution with events which previously have had a different context. This method is well established for ME level calculations [190]. The process can also be executed at the detector level for MC samples, provided that the relevant generator weights at ME level have been stored. The parton shower and detector simulation can then be reused, thereby saving the necessary, large computational effort to redo the calculation. The nanoAOD files utilized in this analysis are equipped with all the requisite information. The target weights w_a for the Pseudoscalar A signal samples thus explicitly read

$$w_a = \frac{|M_a(p_i, s_i)|^2}{|M_A(p_i, s_i)|^2} \quad (7.1)$$

for the ME level fourvectors p_i and spins s_i . This equation explains that the previous weight distribution making up the $|M_A(p_i, s_i)|^2$ cancels out and the reweighted MC sample follows the weight distribution of the ALPs $|M_a(p_i, s_i)|^2$. Figure 7.1 shows control plots to validate the reweighting method for the search variables $m_{t\bar{t}}$ and c_{hel} as well as the bottom quark p_T distribution for two values of $c_{\tilde{G}}/f_a$ and a low width, low mass ALP. The ALPs sample obtained from reweighting the A sample is shown against an ALP sample generated standalone with the MadGraph5_aMC@NLO generator with the same parameter values of mass, width and couplings c_t and $c_{\tilde{G}}$ to validate that their shapes match. A full set of reweighting validation plots is shown in the appendix A.4. Both samples agree within statistical uncertainties and no systematic problems can be diagnosed. The statistical error of the samples after reweighting is larger than before and thus consequently propagated through the rest of the analysis steps.

The analysis configuration outlined in 6.1.1.1 is applied to the reweighted data samples. The event selection, corrections and top reconstruction are computed and the output is binned in $m_{t\bar{t}} \times c_{\text{hel}} \times c_{\text{han}}$ to create signal templates similar to the H/A templates.

Some results for the 2018 data taking era are shown in figure 7.2. A full set of templates for all eras is attached to the appendix A.4.

At $c_{\tilde{G}}/f_a = 0 \text{ TeV}^{-1}$, the signal is generally exhibiting a dip at the ALP mass. The dip is more pronounced for smaller ALP masses. Higher masses have a less pronounced dip and a broader peak. The difference between the peak maximum and the dip minimum decreases with larger ALP masses.

Higher absolute values of the gluon coupling enhance the signal. $c_{\tilde{G}}/f_a = 0.1 \text{ TeV}^{-1}$ shows a yield approximately $100\times$ larger than $c_{\tilde{G}}/f_a = 0.0 \text{ TeV}^{-1}$.

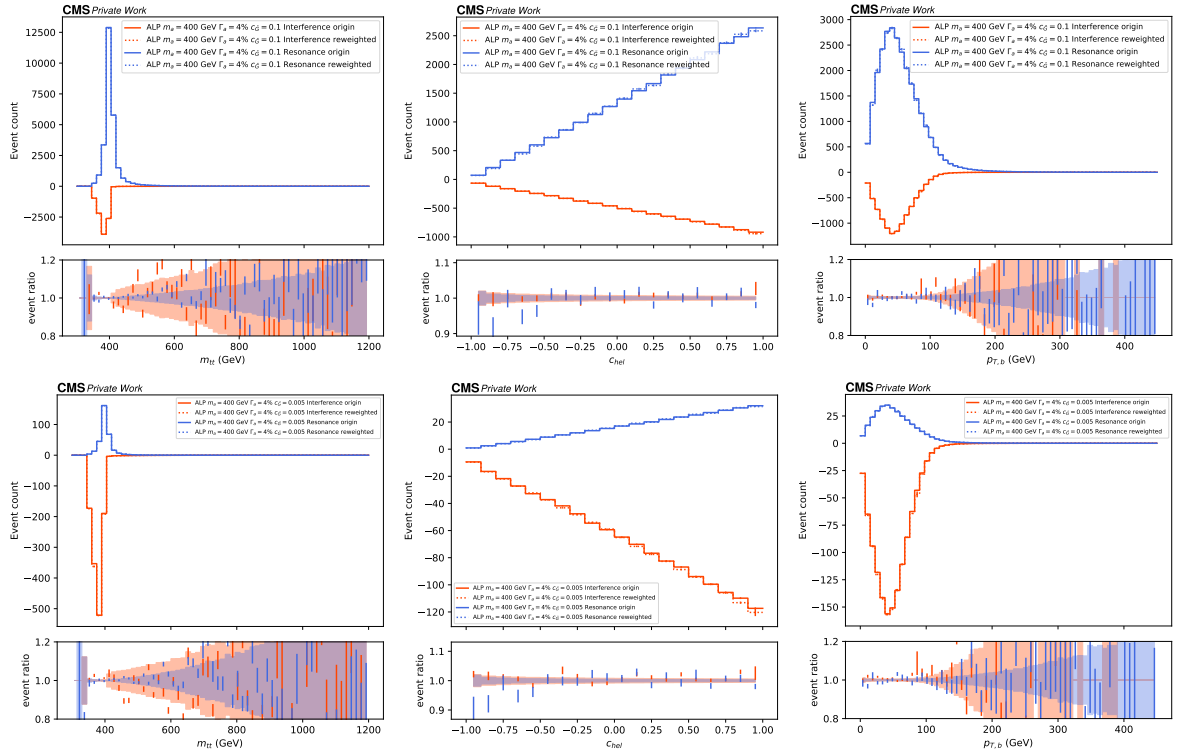


FIGURE 7.1: Distributions of $m_{t\bar{t}}$, c_{hel} and the transverse momentum of the b-quarks, $p_{T,b}$, are shown for $t\bar{t}$ production via ALPS exchange with a mass of $m_a = 400$ GeV and a gluon coupling of $c_{\tilde{G}} = 0.1$ (upper row), $c_{\tilde{G}} = 0.005$ (lower row), c_t is set to 1 for all plots. The standalone sample (solid) is compared the respective sample derived by reweighting an $A \rightarrow t\bar{t}$ sample (dotted). The panels below the plots show the ratio of the reweighted sample over the standalone generated sample. Statistical uncertainties are indicated as shaded areas here.

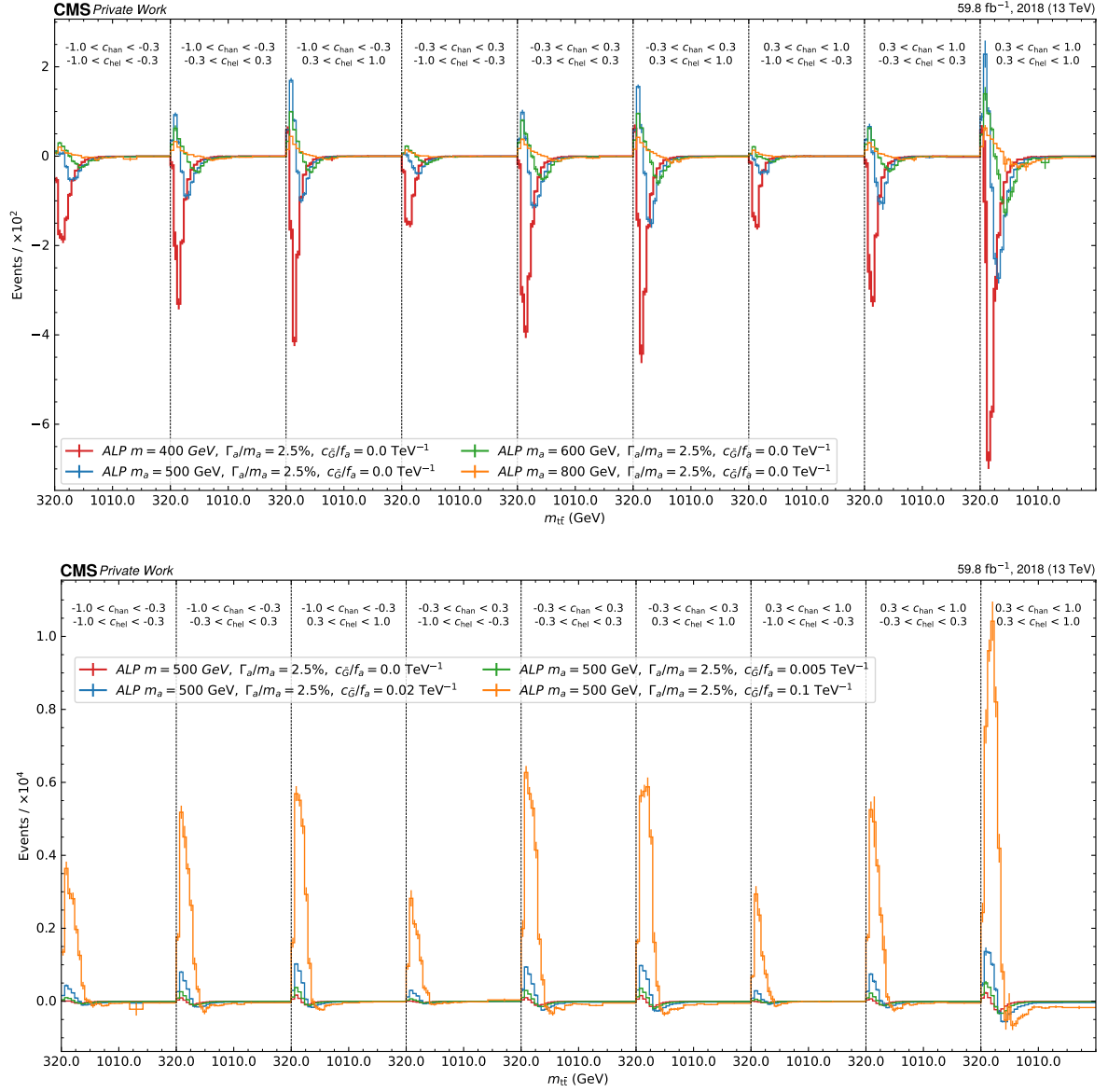


FIGURE 7.2: A set of ALP signal templates on detector level after analysis selection and corrections for the 2018 data era. The resonant and interference contributions are combined, but the SM $t\bar{t}$ background is not depicted. The upper panel shows different values for the ALP mass m_a at $c_G = 0 \text{ TeV}^{-1}$ and the lower panel shows different values of the gluon coupling modifier c_G for $m_a = 500 \text{ GeV}$.

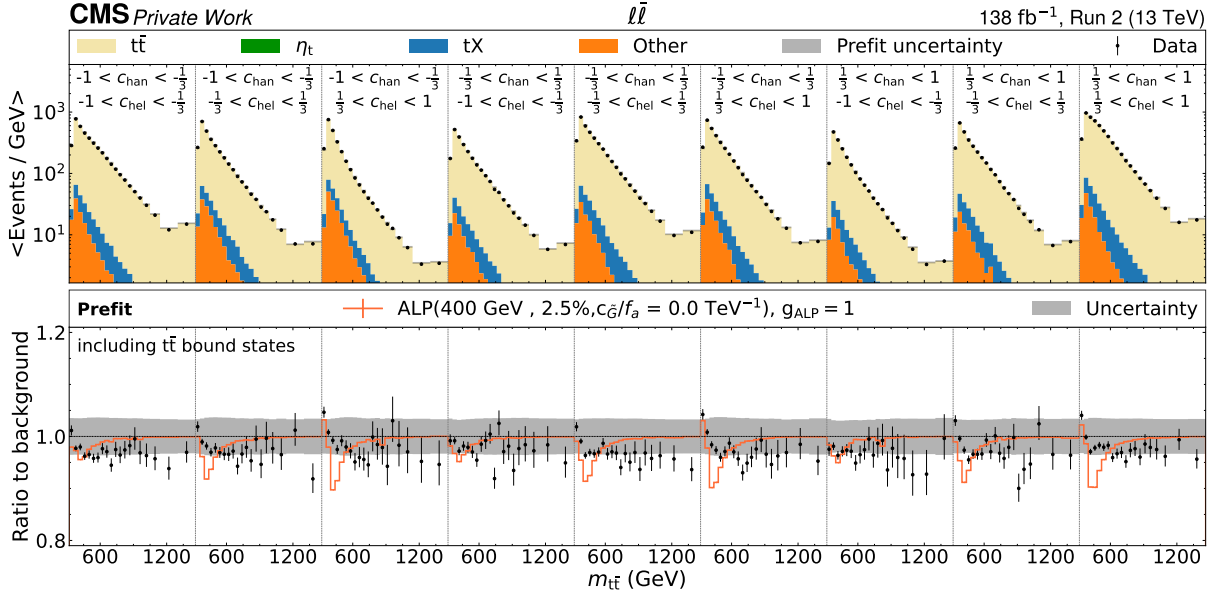


FIGURE 7.3: Observed and expected $m_{t\bar{t}}$ distribution in bins of c_{hel} and c_{han} , shown in dilepton final states channel summed over lepton flavors and eras. In the upper panel, the data (points with statistical error bars) are compared to $t\bar{t}$ production in FO pQCD and other sources of background (colored histograms) before the fit to the data. The ratio of data to the prediction is shown in the lower panel, where the ALPs signal ($m_a = 400$ GeV, $\Gamma_a/m_a = 2.5\%$, $c_{\tilde{G}} = 0$, $c_t/f_a = 0.25$ TeV^{-1}) is overlaid. The background model includes the $t\bar{t}$ bound state η_t as discussed in section 6.1.2 but the contribution is 0 before the fit, the normalization of the η_t is chosen by the fit. The gray band shows the prefit uncertainty.

7.2 DERIVING ALP LIMITS

The statistical methodology for the derivation of Φ limits described in section 6.1.3 is reused. Background treatment is directly adopted including the addition of $t\bar{t}$ bound states simulated through a simplified model based on a pseudoscalar η_t , as described in section 6.1.2, as a background with a free floating normalization. Figure 7.3 shows a prefit distribution in the search variables $m_{t\bar{t}} \times c_{\text{hel}} \times c_{\text{han}}$. The background model from the A/H analysis is shown, the majority of events passing the selection are $t\bar{t}$ events, with minor contribution from the single top (tX) and the "other" (mostly DY events) channel. The η_t contribution starts at a signal strength of 0 and is only set in the fit. As the background model is similar, the prefit distributions look comparable to the results of the search for the heavy pseudoscalar A as shown in Ref. [130]. The low $m_{t\bar{t}}$ excess as discussed in chapter 6.1.4 is clearly visible. During the fit, the contribution of the η_t is estimated, which then absorbs the excess at low $m_{t\bar{t}}$. The lower panel shows the ratio of data to background MC prediction with the prefit uncertainty overlaid as a grey band. A general normalization shift for data to slightly below 1 can be observed as well, which was already present in the search for heavy (pseudo-)scalars A/H. Since the analysis is sensitive to shape and the data points are well covered by the prefit uncertainty, the disagreement is acceptable. A possible explanation for the overprediction of the background normalization can be the normalization of the $t\bar{t}$ prediction to NNLO + NNLL which leads to a too large cross section.

As the signal features one further theory parameter in $c_{\tilde{G}}$, the signal model has to be adapted for inference. The fit is carried out for different values in m_a , Γ_a and $c_{\tilde{G}}/f_a$. On any of these points, the signal yield is nonlinearly dependent on the coupling modifier between ALPs and top quarks $g_{at\bar{t}}$ which is defined by the ALP UFO model to be $g_{at\bar{t}} = \frac{c_t \alpha_s}{f_a}$. The signal model reads

$$S_i(p_a, \nu) = \mu \left(g_{at\bar{t}}^4 S_{R,i}(m_a, \Gamma_a, c_{\tilde{G}}, \nu) + g_{at\bar{t}}^2 S_{I,i}(m_a, \Gamma_a, c_{\tilde{G}}, \nu) \right). \quad (7.2)$$

Here, ν denotes the vector of nuisance parameters. The fit is carried out with the modified LHC test statistic $\tilde{q}$ and exclusion boundaries are set using the asymptotic approximation to obtain the CL_s value as explained in section 3.3. Just like in the A/H case, the signal fit employs $\mu = 1$ to check a grid of values in $g_{at\bar{t}}$ and the background only hypothesis is tested with $\mu = 0$.

The results are presented in c_t/f_a and $c_{\tilde{G}}/f_a$ with the ALP scale f_a .

7.3 ALP LIMIT RESULTS

The results show no significant excess when η_t is included into the background model. This is expected, as the ALP scenario is comparable to the A result outlined in section 6.1.4. The derived limits are shown as a function of the ALP mass in figure 7.4 for a low width ALP hypothesis with $\Gamma_a/m_a = w_a = 2.5\%$. The sensitivity is highest for

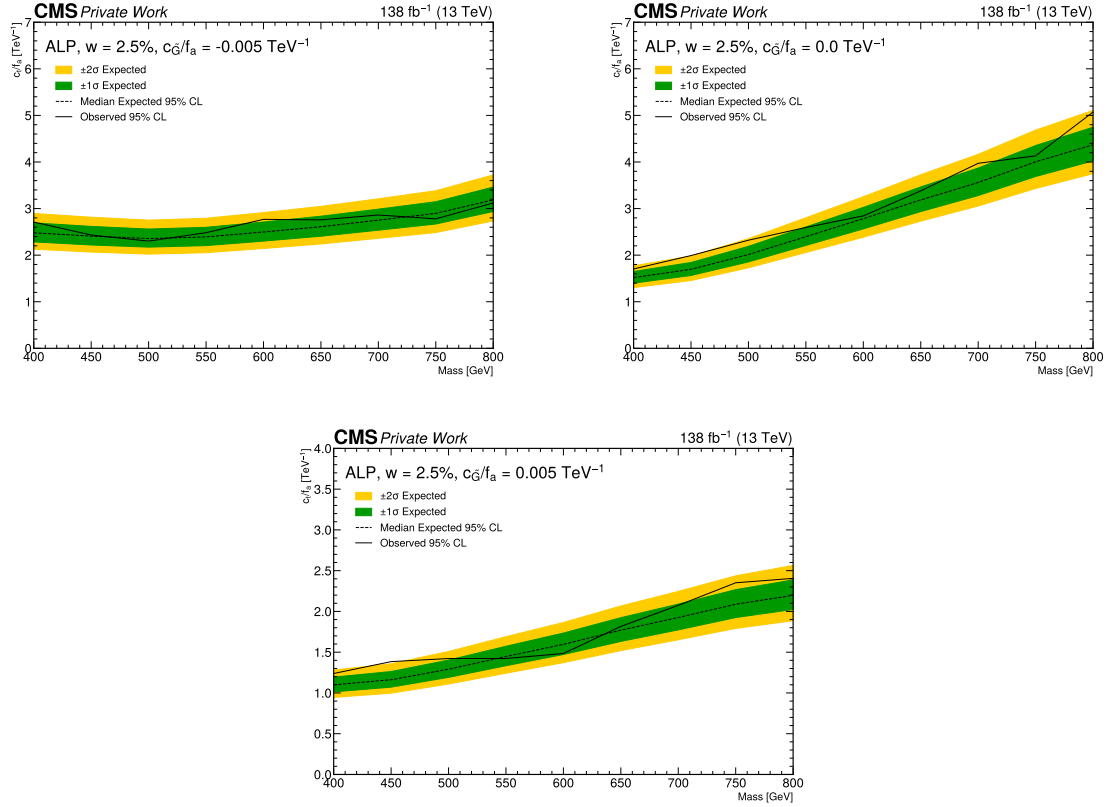


FIGURE 7.4: Model-independent constraints on c_t/f_a as functions of the ALP mass where the η_t contribution is included into the background, for Γ_a/m_a of 2.5%, and c_G/f_a of -0.005 TeV^{-1} (upper left), 0.0 TeV^{-1} (upper right), and 0.005 TeV^{-1} (lower). The observed constraints are indicated by the solid black line. The inner green and outer yellow bands indicate the regions containing 68% and 95%, respectively, of the distribution of constraints expected under the background-only hypothesis.

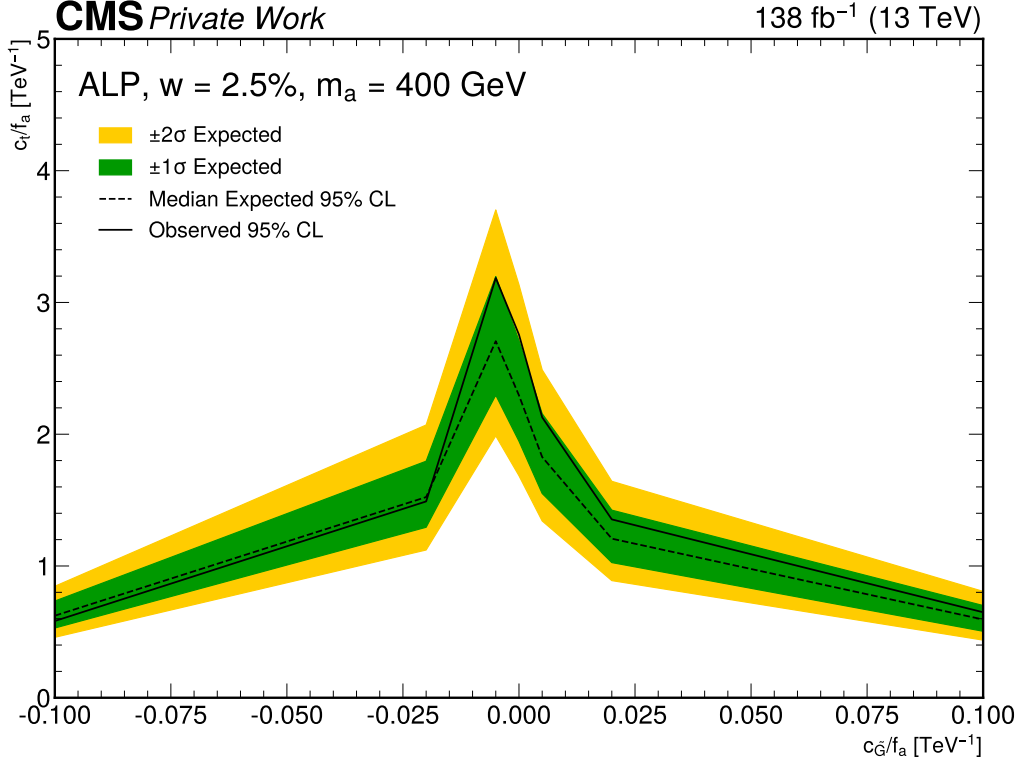


FIGURE 7.5: Model-independent constraints on c_t/f_a as function of the ALP-gluon coupling $c_{\tilde{G}}$ where the η_t contribution is included into the background, for Γ_a/m_a of 2.5%, and $m_a = 400$ GeV. The observed constraints are indicated by the solid black line. The inner green and outer yellow bands indicate the regions containing 68% and 95%, respectively, of the distribution of constraints expected under the background-only hypothesis.

small ALP masses. In the templates it is visible, that the cross section of ALPs is largest here, thus allowing for the most stringent exclusion.

Following this finding, it is particularly interesting to compare the results with the predicted sensitivity from Ref. [53]. In the phenomenological study, the exclusion was given in c_t/f_a as a function of $c_{\tilde{G}}/f_a$. The corresponding result from this analysis is shown in figure 7.5.

For higher absolute gluon coupling values, the sensitivity to ALPs increases due to the higher ALP cross section, which leads to a stricter limit. Small absolute values of $c_{\tilde{G}}/f_a$ are thus more difficult to exclude. Accordingly, a mountain-shape in the exclusion contour can be observed. The peak is shifted to small negative values of $c_{\tilde{G}}/f_a$ as the two ALP production diagrams (shown in figure 2.7) destructively interfere at this value. Subsequently, the cross section at this value of $c_{\tilde{G}}/f_a$ is lower and the sensitivity is worse. The result compares to the predicted exclusion limits for Run 2 from Ref. [53]. Shape and quantitative excluded regions agree, while the actual experimental exclusions are slightly stricter than the extrapolation in the phenomenological study likely due to an improved uncertainty modelling from the early Run 2 analysis (as used in the phenomenological study) to the full Run 2 CMS analysis.

7.3.1 Nuisance Parameter Pulls and Impacts

To validate the fit results, the pulls and impacts of nuisance parameters are evaluated. Figure 7.6 shows the 20 leading impacts and nuisance pulls in a limit setting fit for an ALP hypothesis with $m_a = 400$ GeV, $w_a = 2.5\%$ and with no gluon coupling, $c_{\tilde{G}}/f_a = 0$ TeV^{-1} .

Similar to the fit in the A/H scenario, the most impactful nuisances are related to the modelling of the $t\bar{t}$ contribution. By far the most impactful uncertainty is the normalization of the η_t contribution. Since η_t is a pseudoscalar just like the ALP and they populate the same part of the $m_{t\bar{t}}$ spectrum for $m_a = 400$ GeV, it is expected that a lower η_t contribution would necessitate more ALP contribution.

Less impactful, but pulled and constrained are the top mass nuisance and the two-point nuisance assigned to the difference between the BB4L generator and the nominal HVQ $t\bar{t}$ prediction. Both nuisances are known to have an intricate relationship to the η_t contribution. Since they are both pulled by less than one standard deviation and their impact is moderate, the behavior can be regarded as not problematic.

CMS
Private Work

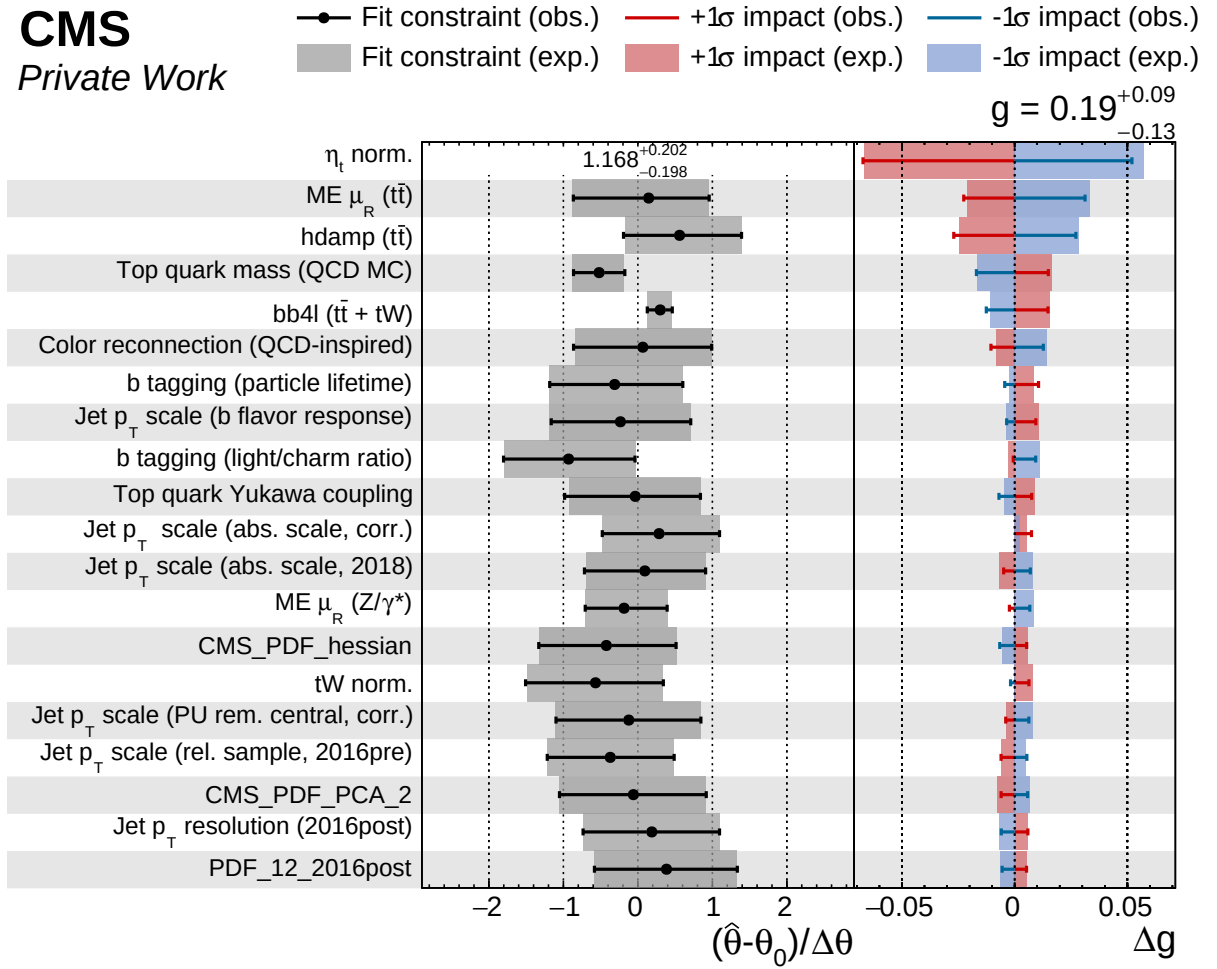


FIGURE 7.6: The 20 leading nuisance parameter pulls and impacts for the maximum likelihood fit with an ALP model at $m_a = 400$ GeV, $w_a = 2.5\%$ with $c_G/f_a = 0.0$ TeV^{-1} . The uncertainty modelling is similar to Ref. [43].

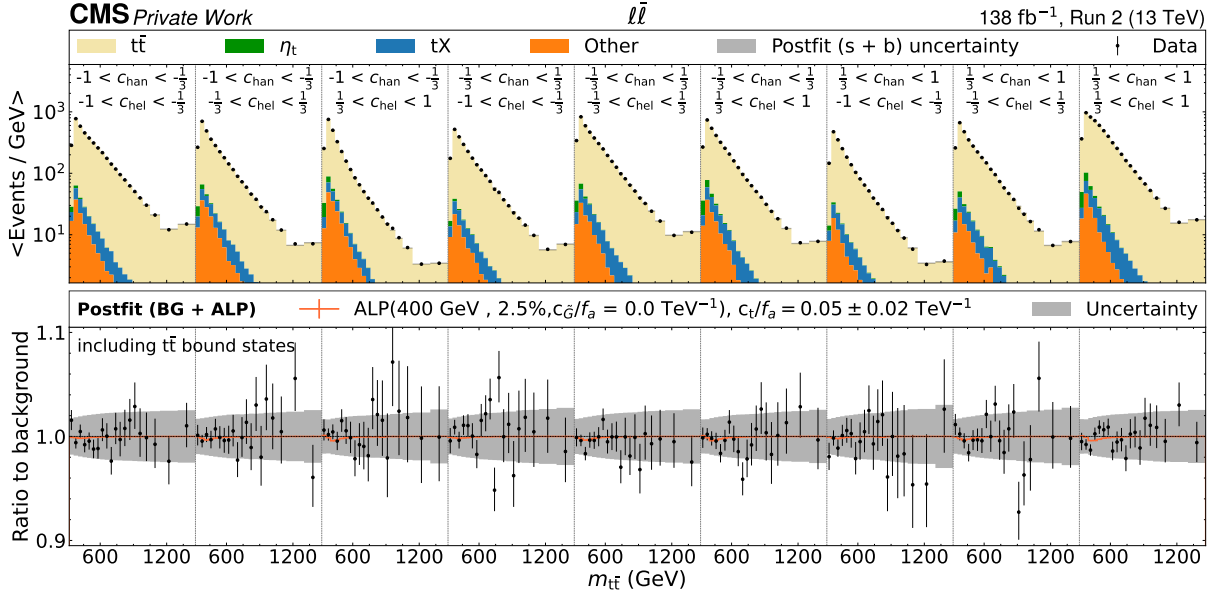


FIGURE 7.7: Postfit distribution for a fit of a $m_a = 400$ GeV, $w_a = 2.5\%$, $c_{\tilde{G}}/f_a = 0.0$ TeV^{-1} ALP hypothesis. It is evident how the η_t contribution absorbs the prefit excess for low $m_{t\bar{t}}$ and how the ALP best fit scenario is compatible with $c_t/f_a = 0$ TeV^{-1} . Other notations are similar to figure 7.3.

An important second sanity check of the results is the postfit distribution. For a low mass ALP ($m_a = 400$ GeV, $w_a = 2.5\%$, $c_{\tilde{G}}/f_a = 0.0$ TeV^{-1}), the results is shown in figure 7.7. It can be seen, that the η_t contribution was estimated in the fit. It absorbs the low $m_{t\bar{t}}$ excess that was visible in the prefit distribution fully. The best fit ALP scenario is contained in the grey postfit error band and compatible with 0. The limit that can then be set excludes ALP top coupling higher than $c_t/f_a > 1.7$ TeV^{-1} for this parameter point.

7.3.2 Discussion and Comparison

It can be concluded that there is no further evidence for an ALP at the $t\bar{t}$ threshold if one includes an η_t contribution to model $t\bar{t}$ bound state effects. One can then proceed to set exclusion limits for ALP masses between 400 and 800 GeV in the ALP top coupling c_t . These agree with the projections from earlier phenomenological studies. The c_t coupling limits are also derived for different hypothetical values of the ALP gluon coupling $c_{\tilde{G}}$. For the scenario with $c_{\tilde{G}} = 0$ the excluded range is from $c_t/f_a > 1.7$ TeV^{-1} at 400 GeV to 4.4 TeV^{-1} at 800 GeV. For $m_a = 400$ GeV the excluded region ranges from the strictest limit at $c_t/f_a > 0.6$ TeV^{-1} for $c_{\tilde{G}}/f_a = -0.1$ TeV^{-1} to the weakest limit $c_t/f_a > 3.2$ TeV^{-1} for $c_{\tilde{G}}/f_a = -0.005$ TeV^{-1} . These limits are the strictest experimental limits as of the time of writing this text for the vanishing gluon coupling and the first limits with a non-vanishing gluon coupling. Another interesting aspect is to combine these limits into a greater picture. CMS has previously published a DM search based on $t + \text{DM}$ and $t\bar{t} + \text{DM}$ [191]. This CMS search utilizes 138 fb^{-1} of $\sqrt{s} = 13$ TeV Run 2 proton-proton collision data to investigate dark matter produced in association with single or pairs of top quarks across all-hadronic, single-lepton, and dileptonic final states. Crucially, the analysis interprets these results through simplified models of scalar and pseudoscalar mediators analogous to this work, explicitly providing new limits for axion-like particles (ALPs) that couple to top quarks and dark matter fermions. Despite a 1.9 standard deviations signal-like excess, the study excludes mediator masses below 310 (320) GeV for the scalar (pseudoscalar) hypothesis at the 95% confidence level. It also derives a set of ALP top coupling exclusions as a function of the ALP mass m_a analogous to this work for a scenario with a vanishing gluon coupling $c_{\tilde{G}} = 0$.

Figure 7.8 illustrates the synthesis of these two analyses. Because the search regions are kinematically distinct, the combined exclusion limit is formed by the union of the two sensitivity regions. This complementarity achieves a wide-range coverage of the ALP parameter space, spanning more than an order of magnitude in mass, where the $t/t\bar{t} + \text{DM}$ approach covers the lower mass range and the $t\bar{t}$ search covers the high mass range.

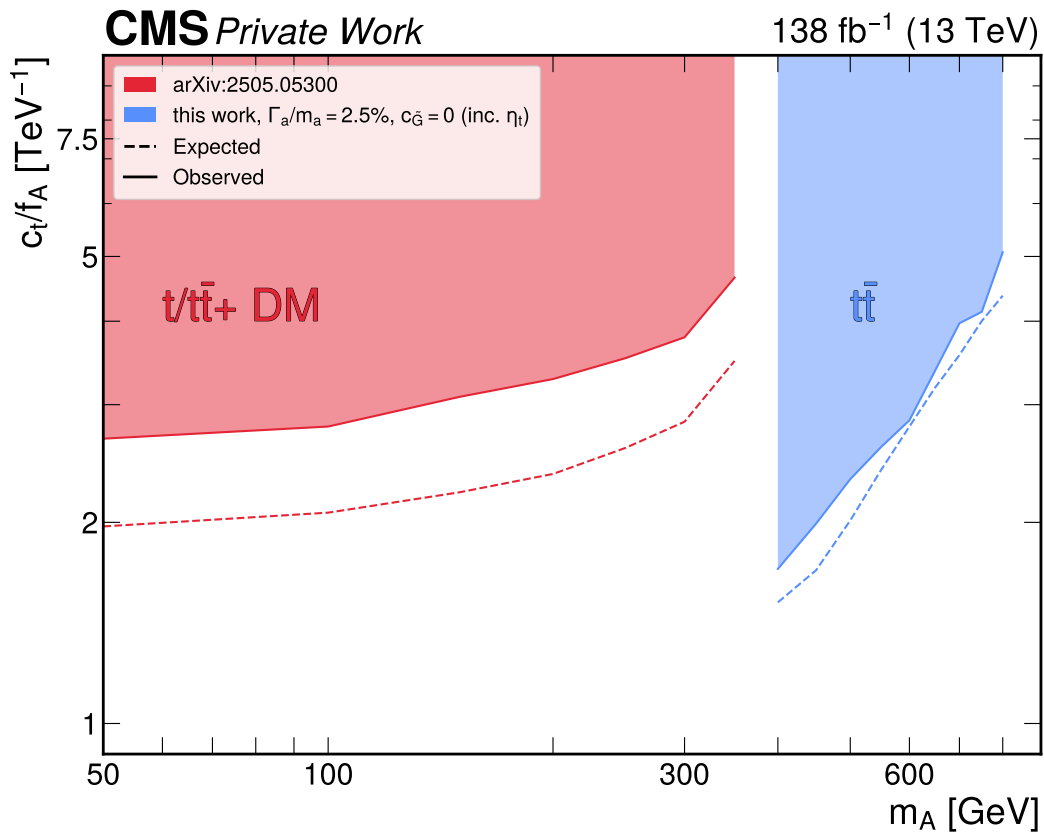


FIGURE 7.8: Upper CMS limits to the ALP- $t\bar{t}$ coupling c_t/f_a from $t/\bar{t} + \text{DM}$ [191] (red) and $t\bar{t}$ (this work, blue) analyses.

Previous chapters already highlighted diverse use cases for ML models in data analysis for collider physics experiments. A very common application are classification models which can be used for object identification or directly as an analysis discriminant that separates signal-like from background-like events in statistical inference. Typically, these models are trained on MC data, which serves as an understood and accessible proxy for the underlying physical processes and detector response. However, the reliance on MC simulations also introduces potential problems such as the propagation of systematic imperfections or artifacts inherent in the simulation to the model and into its predictions. Therefore, it is essential to quantify and mitigate the influence of such simulation-induced effects to ensure that the trained ML models remain robust and transferable to real experimental data.

One remarkable simulation-specific feature is the occurrence of negative event weights. In modern MC event generators, events are not always produced with equal statistical importance. Instead, each generated event is assigned a weight that determines its relative contribution to observable distributions. This primary weight, referred to as the generator (GEN) weight, reflects the theoretical cross section normalization. Other weights might be applied in the simulation chain to account for uncertainties (PS weights, PDF weights and scale weights) but are not further considered in this work as they are generally non-negative.

In many cases, especially in NLO calculations, the event generation can also lead to negative weights. These negative event weights are caused by multiple reasons outlined in detail in 8.1. Major contributions come from the subtraction formalism for NLO calculations but also from interference effects. As a result, simulated datasets often contain a substantial fraction of negatively weighted events.

In standard histogram-based analysis these weights pose no conceptual issue as they can simply be summed to obtain unbiased cross section estimates. However, in ML applications they introduce nontrivial challenges. Most ML algorithms, including popular classifiers such as BDTs and DNNs, are formulated under the assumption of positive-definite training weights. Negative weights can therefore distort the optimization objective, destabilize the training process, and bias the learned decision boundary.

This chapter investigates in detail the impact of negative event weights on the training and performance of ML classifiers in collider physics analyses. It explores strategies based on modifications to the training objective and sample handling. In doing so, it aims to provide a framework for developing stable and physically consistent ML models that can be trained directly on MC datasets with substantial subsets of negatively weighted events without introducing artifacts into the final analysis discriminants.

8.1 NEGATIVE EVENT WEIGHTS

There are two distinct reasons for negative event weights in MC simulation data.

8.1.1 Negative Weights in fixed higher-order Monte Carlo Simulations

Fixed order MC calculations of SM scattering processes are not necessarily finite in all areas of phase space. Ultraviolet (UV) and Infrared (IR) divergences commonly occur if one breaks the perturbative series at a given order. IR-finite observables can only be guaranteed by the cancellation of real corrections and soft loop contributions which is given by the KLN theorem as explained in section 2.1.2. UV divergences on the other hand need to be regularized to achieve convergent results. The IR cancellation is technically challenging to implement for numerical calculations as individually computed diagrams are divergent. A solution is given in form of subtraction schemes such as the Catani-Seymour Dipole Subtraction Formalism [192]. For NLO contributions, the NLO part of the cross section takes the form

$$\sigma^{\text{NLO}} = \int d\sigma^{\text{NLO}} = \int_{m+1} d\sigma^{\text{R}} + \int_m d\sigma^{\text{V}} \quad (8.1)$$

where m is the number of partons in the final state. The integral is divided into the real corrections $d\sigma^{\text{R}}$ and the virtual corrections $d\sigma^{\text{V}}$ both of which have divergences. By introducing a counterterm $d\sigma^{\text{A}}$ with the same pointwise singular behavior of $d\sigma^{\text{R}}$ and subtracting them from each other, a finitely defined result can be derived. If one can then analytically go from an integration over $m+1$ partons to m partons in the counterterm, the poles of the counterterm and the virtual corrections will cancel

$$\sigma^{\text{NLO}} = \int_{m+1} [d\sigma^{\text{R}} - d\sigma^{\text{A}}] + \int_m [d\sigma^{\text{V}} + \int_1 d\sigma^{\text{A}}]. \quad (8.2)$$

The exact form of the $d\sigma^{\text{A}}$ term can be derived from the dipole formalism introduced in Ref. [192].

Adding $d\sigma^{\text{A}}$ on one integral and subtracting it from the other is what introduces the negatively weighted events

in this scheme. However, the overall result of this cross section calculation will be physical and as such positive definite. Reweighting methods have thus been developed to reweight events into (almost) positive definite weight distributions. These will be further discussed in section 8.3.2.

8.1.2 Interferences and Signal-Background Splits

Equation 2.4 shows the decomposition of a cross section into distinct amplitude contributions that lead to an indistinguishable final state. Besides two clearly identifiable resonance terms the expression also contains a cross term that arises from the product of both amplitudes. This cross term represents the physical interference between the processes and can locally take either positive or negative values depending on the relative phase of the amplitudes. In regions of destructive interference, the corresponding contribution to the event yield becomes negative.

Despite such local cancellations, the total cross section is strictly positive, as required by unitarity. The full differential cross section, too, remains positive definite in all regions of phase space. It thus represents a probability density for the expected event distribution and encodes the physically measurable rate.

When a classification model is trained to detect the presence of an additional amplitude (such as a BSM resonance) one must provide distinct samples representing the signal and background hypotheses. For this purpose, it is natural to define the background as the known SM resonance alone and to group all modifications caused by the new amplitude into the signal class. The signal sample therefore includes not only the pure resonant contribution from the new amplitude but also its interference with the background. This definition aligns with the physical interpretation of a “signal hypothesis” as the full change in the expected distribution that arises once the additional amplitude is introduced.

The resulting split into signal and background datasets is, however, not a physical divide. It can only be introduced on labeled MC data. In the true physical process, the interference term cannot be uniquely assigned to either class, since it depends on both amplitudes. Nevertheless, for the purpose of training a classifier, this artificial separation provides a practical framework for approximating the likelihood ratio between the “background-only” and “signal-plus-background” hypotheses.

An immediate consequence of this labeling choice is that the signal sample inherits the negative event weights originating from destructive interference. In certain regions of phase space, these negative contributions can dominate, resulting in a signal distribution that is not positive definite. As a result, the signal sample does not correspond to a probability density in the strict sense but rather to a quasiprobability distribution. When the signal and background quasiprobabilities are summed, they recover the physically meaningful, positive-definite probability density that represents the full expected event yield.

From a phenomenological standpoint, this means that the presence of the signal amplitude can reduce the total number of expected events in certain areas of phase space compared to the background-only case. This reduction is itself a physical signature of the signal.

In this framework, a classifier trained on such samples effectively learns to distinguish between two hypotheses: the background-only scenario, characterized by the baseline probability distribution and the scenario in which a second amplitude modifies this distribution through both its direct and interfering effects.

The inclusion of interference in the signal class therefore ensures that the model becomes sensitive to the full physical consequence of the signal presence, both enhancements and suppressions, rather than only to resonance contributions.

8.2 ASSESSING THE EFFECT OF NEGATIVELY WEIGHTED EVENTS ON ML CLASSIFICATION TASKS

Minimizing a loss over a weighted dataset is typically implemented as a multiplication of the batch loss by the average weight in that batch. This means that this average weight has a certain probability to become zero once there are more negatively weighted events than half the batch size. This can lead to a negative training step, the training gradient is inverted. Therefore, gradient descent becomes gradient ascent, maximizing error rather than minimizing it. The minimization of a loss function, which is now unbounded from below, will lead to a divergence to $-\infty$. The actual minimization of the error on the positively weighted batches will have a minor contribution and the performance of the model suffers.

To assess the probability of this catastrophic divergence, it is first mandatory to understand at which level of negative weight contamination an effect is expected.

Naturally, one would expect negative training updates at the threshold where it is possible to draw a batch with a negative mean weight. One can estimate the probability of not having a single negative batch with a MC sampling as described in algorithm 1.

Algorithm 1 Monte Carlo estimate of $P_1(\bar{W}_{N_{\text{batch}}} < 0)$ and $P_2(\bar{W}_{\text{all } N_{\text{batch}}} > 0)$

Require: Weighted events $\{w_1, w_2, \dots, w_N\}$, batch size N_{batch} , number of simulations N_{sim}

```

1: count  $\leftarrow 0$ 
2: for  $i = 1$  to  $N_{\text{sim}}$  do
3:   Draw a random subset  $S_i \subset \{1, \dots, N\}$  of size  $N_{\text{batch}}$  without replacement
4:   Compute sample mean  $\bar{w}_i = \frac{1}{N_{\text{batch}}} \sum_{j \in S_i} w_j$ 
5:   if  $\bar{w}_i < 0$  then
6:     count  $\leftarrow$  count + 1
7:   end if
8: end for
9: Estimate probability  $\hat{P}_1 = \frac{\text{count}}{N_{\text{sim}}}$ 
10: Estimate probability of not receiving a negatively weighted batch  $\hat{P}_2 = (1 - \hat{P}_1)^{\text{Floor}(N/N_{\text{batch}})}$ 
11: return  $\hat{P}_1$  and  $\hat{P}_2$ 

```

The sampling here calculates the probability $P_1(\bar{W}_{N_{\text{batch}}} < 0)$ to draw a negative batch when having a batch size N_{batch} . From $\hat{P}_1$ one can deduce the probability of not drawing a negatively weighted batch across one epoch by taking the exponent of the counterprobability to the power of training steps. Drawing a random subset without replacement ensures that each epoch is exactly one pass of the full dataset.

This drawing behavior is reflected in the hypergeometric distribution. This makes it possible to derive an analytic estimate of the probability to draw a negative batch for the case when the weights have only two distinct values a and b , with $w_{\text{negative}} = -b$. The mean weight for a batch can be expressed as

$$\hat{W} = \frac{1}{N_{\text{batch}}} (K(-b) + (N_{\text{batch}} - K)a) = a - \frac{K(a+b)}{N_{\text{batch}}}. \quad (8.3)$$

K denotes the number of negatively weighted events drawn in the batch. Requiring the mean weight $\hat{W}$ to be negative and solving for K

$$a - \frac{K(a+b)}{N_{\text{batch}}} < 0 \Leftrightarrow K > \frac{aN_{\text{batch}}}{a+b}. \quad (8.4)$$

So the batch sample becomes negative in mean weight when at least $K_{\min} = \lfloor \frac{aN_{\text{batch}}}{a+b} \rfloor$. The hypergeometric distribution to draw exactly k negatively weighted samples is given as

$$P(K = k) = \frac{\binom{N_-}{k} \binom{N_+}{N_{\text{batch}} - k}}{\binom{N}{N_{\text{batch}}}}, \quad (8.5)$$

where N_- denotes the overall number of negative events and N_+ denotes the number of positive events. To get the probability to draw any batch with a negative weight-mean, one has to calculate the correct tail of the hypergeometric distribution. All events that allow for $K_{\min}$ negative events up to all possible negative events are

$$P(\bar{W}_{N_{\text{batch}}} < 0) = \sum_{k=K_{\min}+1}^{\min(N_-, N_{\text{batch}})} \frac{\binom{N_-}{k} \binom{N_+}{N_{\text{batch}} - k}}{\binom{N}{N_{\text{batch}}}}. \quad (8.6)$$

The validity of this calculation can be tested with an experiment. Batches are drawn from a dataset with 60000 events with different ratios of negatively weighted events to positively weighted ones. The results are plotted in figure 8.1 and show that the MC method and the analytic calculation agree. It is also evident that larger batch sizes have curves which are offset in direction of higher negative weight ratios. To study the scaling of this offset, the thresholds at which the 50% probability to not draw a single negatively weighted batch are crossed are plotted as a function of the batch size. It can be seen that the curve exhibits a steep incline for small batch sizes but flattens for large batch sizes. Larger batch sizes can cover the increase in negative weights to some degree but at some point, given by the overall sample size, the coarser batch sampling can not cover the amount of negatively weighted events.

This particular case of weight with a defined absolute value is not just an academic exercise. It is relevant to particle physics as, for example, MadGraph5_aMC@NLO unweights up to a sign, ultimately producing a weight distribution of -1 and 1.

This has established the presence and likelihood to encounter a negatively weighted batch in model training. But since losses are multiplied by their weights and then averaged, the batch loss will be dynamically influenced by the convergence in the samples in the current batch.

For ML applications, it is important to understand at which level of contamination with negative weights this problem hinders the convergence of ML classifiers.

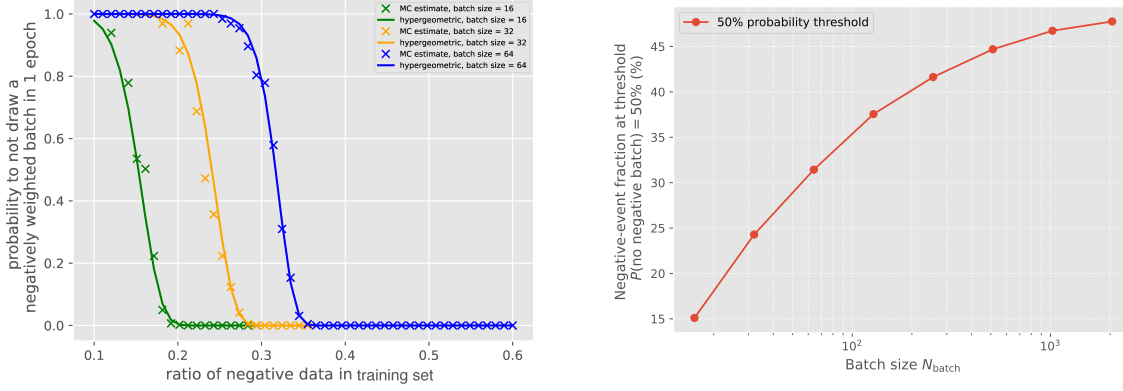


FIGURE 8.1: The left panel shows the ratio of negative data versus the probability to not draw a batch with an average negative weight. This probability is evaluated with the MC method shown in 1 and with the hypergeometric distribution tail evaluated in equation 8.6 for different batch sizes. The right panel shows the scaling of the 50% threshold of the discussed probability as a function of the batch size.

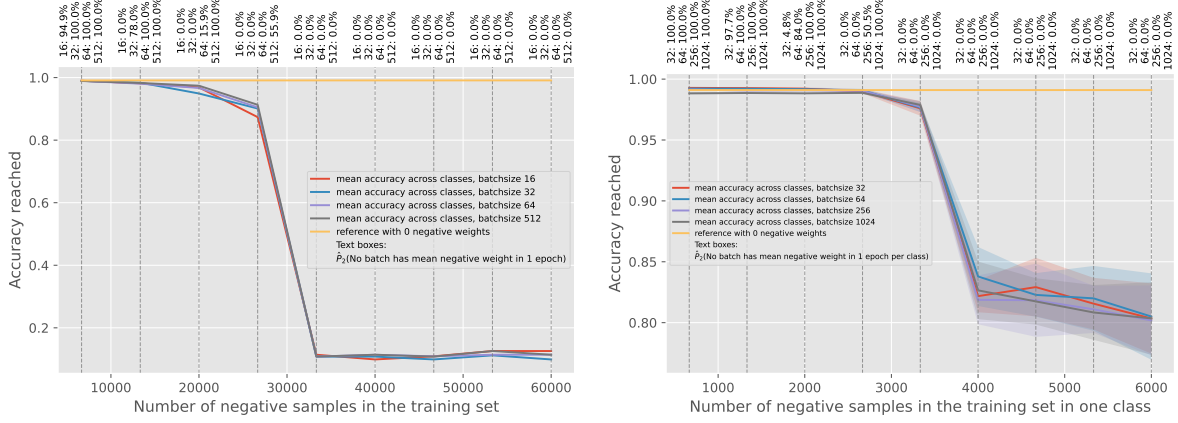


FIGURE 8.2: CNN accuracy on MNIST test set as a function of number of negative weights present during training. The left plot shows the performance after random samples were selected in the training set across all classes. The right plot shows the same experimental setup with negative samples chosen from a single class. The accuracy results were then evaluated on the positive classes in the test set only. The mean over all classes is shown and the shaded area shows the standard deviation across different classes chosen as negatively weighted.

For this purpose a toy model is built from the MNIST dataset [193]. This dataset offers a set of handwritten numbers which are to be classified to their values. Convolutional Neural Networks can achieve accuracies beyond 99% on this problem [194]. Subsequently, a simple CNN was trained which achieves 99.1% accuracy. Two experiments were carried out to quantify the impact of negative weights.

First, random data points across all classes were picked and assigned a negative weight of $w = -1$ while the rest of the data was weighted with $w = 1$. The same CNN was trained from a reinitialized state with an increasing percentage of negative weights. The left panel in Figure 8.2 illustrates the classification accuracy achieved on the test set after 15 epochs of training as a function of the number of negatively weighted training samples, evaluated for a range of batch sizes.

Each vertical line in the plot, accompanied by a text box, indicates the probability $\hat{P}_2$ of not encountering a single batch with a negative mean weight during one full epoch of training. This probability is computed for each batch size at a given fraction of negatively weighted samples in the dataset. In other words, this would provide a probabilistic interpretation of the onset of instability in the training process.

Remarkably, the results show that the choice of batch size has almost no effect on the point at which catastrophic divergence occurs. Across all tested configurations, the models exhibit a sharp breakdown in performance once approximately 50% of the training data consists of negatively weighted samples. This behavior is consistent with the expectation that, beyond a certain threshold, the effective signal-to-noise ratio in the loss gradient becomes dominated by the sign-flipped samples, leading to unstable parameter updates and the loss of meaningful convergence. Before this critical point, models trained with smaller batch sizes tend to reach slightly lower accuracies by a few percent, suggesting that stochastic effects play only a secondary role in the overall degradation pattern. This observation provides strong evidence that the underlying issue lies primarily in the optimization dynamics rather than in stochastic sampling fluctuations. Once the proportion of negatively weighted samples becomes large

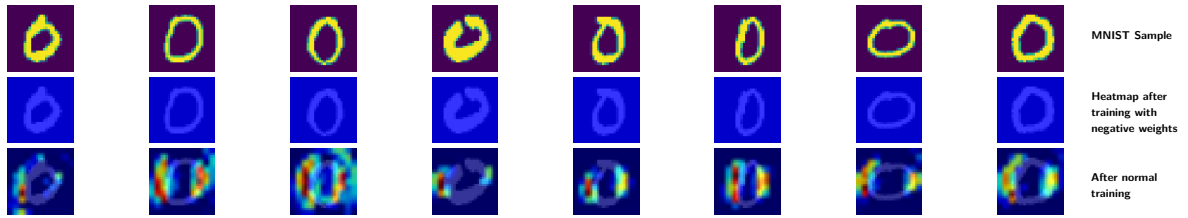


FIGURE 8.3: Grad-CAM heatmaps [195] for a DNN training with negative weights in class "0" (center row) and without any negative weights (lower row). The upper row shows the sample for reference.

enough, the model begins to overfit on those samples, effectively minimizing the loss in a direction that contradicts the intended learning objective. This results in the apparent collapse of the model's predictive performance.

Interestingly, the batch size appears to have only a minor moderating influence in the semi-stable regime preceding catastrophic divergence.

For negative weights which are concentrated in one class, the scenario is further altered. In the case of multiple classes, if the loss is larger on the negatively weighted classes, they will dominate the overall batch loss. This is expected as the negative weight does not actually make the model converge on those samples. Assuming convergence on the positive samples, the loss on all positive samples will be small and the loss on classes with negative weights will dominate.

In this limit, the prior findings can be reevaluated on the classes with negatively weighted events rather than on the full dataset.

The results shown in the right panel of Figure 8.2 follow a similar structure and analysis approach as those in the left panel. In this case, however, the probabilities $\hat{p}_2$ were evaluated with respect to a single class being negatively weighted, rather than across the entire dataset. For each class in the dataset, an independent experiment was conducted in which only samples in that class were assigned subsequently more negative weights, while all others retained positive weights. The plotted curve represents the average performance over all such experiments, and the shaded band indicates the standard deviation across different classes, thereby visualizing the class-to-class variation in the model's sensitivity to negative weighting.

Once again, a pronounced drop in performance is observed at more than 50% of the samples within the affected class carrying negative weights. Interestingly, this threshold is almost identical to the critical point identified in the previous experiment, suggesting that the instability is primarily determined by the fraction of negative weights within the relevant training subset, rather than by their absolute number in the dataset.

However, since the negative weights are now confined to a single class, the overall degradation in accuracy is less severe. After catastrophic divergence, the average test accuracy remains above 80%, indicating that the model retains reasonable discriminative power for the unaffected classes.

From a quantitative perspective, assigning 50% of one class's samples negative weights corresponds to only about 5% of all samples in the full dataset being negatively weighted. This places the system within the semi-stable regime identified in the previous experiment. Training is still possible but partial degradation occurs.

Overall, these results show that the onset of instability due to negative weights depends not only on their global fraction in the training data but also on how they are distributed across classes. Concentrating negative weights within a single class localizes the instability, reducing its global impact on accuracy but still triggering catastrophic behavior once the negative-to-positive ratio within that class exceeds roughly 50%.

The Grad-CAM algorithm [195] allows to study the feature maps that a CNN learns. By extracting the outputs of the convolutional layers and their gradients, a heatmap can be computed that allows to interpret which features the network focuses on. Figure 8.3 shows Grad-CAM heatmaps for the MNIST experiment discussed above. In this instance, a network was trained with class "0" negatively weighted. The lower row shows the training without negative weights and the center row shows the scenario where the full class is negatively weighted. The heatmaps for the negative training indicate that no feature survived the training procedure. The effect on the classification that one intends to reach with negative weights is apparently accompanied by a full feature collapse within the model. Contrasting that are the features shown in the last row. The main indicator here is a focus on round edges on the left and the right of the image. Seeing this feature missing in the case of negative training means that the model can not make use of it for other classes as well. Performance degradation is a direct consequence of that.

8.3 MITIGATING NEGATIVE EVENT WEIGHTS

8.3.1 Naive Legacy Methods

In scenarios that feature enough negatively weighted events to observe problems with convergence while not relying on negatively weighted events to express important features, it can be justified to rely on basic eliminations of said events. It is straightforward to leave these events out of the training or replace their negative weight with the absolute value. Both solutions obviously impact the cross section normalization in the training data but may still result in a well trained classifier, especially if compared to a classifier that fails to converge. It is a priori not

possible to tell if that scenario is realistic, so studies in the individual case are necessary.

Other methods include the usage of the MSE loss, which can become negative but will not diverge to $-\infty$ on a limited model output [196]. For small contributions of negative weights it can also be possible to use large batch sizes to average out those destabilizing contributions as discussed above.

8.3.2 Reweighting

Research has focused on reweighting methods to eliminate negatively weighted events [197, 198]. Usually these works focus on the computational overhead that negative events introduce in the detector simulation. The idea here is, that is more efficient to carry out weight cancellations at the GEN stage and only propagate necessary, positive events. Naturally, this can solve the issue with negative event weights coming from NLO subtraction schemes and these methods are employed in practice to minimize computational overhead from negative weights.

There are two ways to carry out a reweighting. Binned methods use a distance metric to define bins in which sufficient events can be summed to yield a mean event with a positive weight. This can be done in the variables used for analysis with a naive distance metric in this observable [197]. It is found to preserve the resulting distributions in those variables within statistical uncertainties. In order to not only preserve chosen observables, binnings in more general spaces can be chosen such as the cell resampler proposed in [199]. Clustering events based on nearest neighbours can approximately preserve derived distributions in the whole phase space. However, performance depends on the chosen metric and its parameters.

Reweighting based on DNN classifiers allows to overcome limitations of the binned procedure [198]. For this, the likelihood ratio trick as described in section 4.2.1.1 is used. A classifier is trained between the phase space points $\{x_i\}$ sampled from the phase space distribution X with the according GEN weights $\{w_i\}$ sampled from the weight distribution W and the same points $\{x_i\}$ with all weights set to 1. The likelihood ratio derived from the converged classifier $g(x)$ approximates the weight distribution dependent on the phase space distribution

$$\frac{g(x)}{1 - g(x)} \approx \langle W|X \rangle. \quad (8.7)$$

The same can be done for a classifier trained on squared weights to preserve the variance $\langle W^2|X \rangle$. With these approximations a positive definite weight distribution follows when sampling the weights in these steps

1. Estimate $\hat{W}(X) \approx \langle W|X \rangle$ and $\hat{W}^2(X) \approx \langle W^2|X \rangle$ from the trained classifiers
2. Calculate $\hat{K}(X) = \hat{W}^2(X) / \hat{W}(X)^2$
3. Keep the event with probability $p_X = 1/K(X)$
4. Surviving events get a reassigned weight of $w_i \rightarrow \tilde{w}_i = \hat{W}(x_i) \hat{K}(x_i)$.

This method ensures that N_{Patch} events in a patch (defined by the classifier) get replaced with N_{Patch}/K events while keeping the overall weight and variance in that patch the same $\sum_{j=1}^{N_{\text{Patch}}/K} \tilde{w}_j = \sum_{i=1}^{N_{\text{Patch}}} w_i$ and similarly for $\tilde{w}_j^2$. All negative events will be eliminated as $\langle W|X \rangle$ is positive definite by construction. Also, no active decision on the patches and their size is necessary as the trained classifiers allow for a smooth estimation of the correct replacement. However, any failure in convergence of the classifier will propagate into the result. Also, any reweighting reduces the effective statistical power of the dataset.

8.3.3 Pull by Partial Labels

Another strategy to deal with negative events is to modify the classifier training in a way to be stable with respect to the negatively weighted events. For this, Pull by Partial Labels (PPL) is introduced.

In a supervised classification task, the learning objective is defined over a closed set of mutually exclusive output classes. Consequently, the model prediction corresponds to a discrete output node, representing one of the predefined categories. The softmax activation function, commonly employed to map the raw network outputs (logits) to class probabilities [200], enforces normalization of the output vector and enables interpretation as a probability. This normalization implies that a decrease in the predicted probability for one class inherently leads to an increase in the relative probabilities of the remaining classes.

When introducing negatively weighted samples within the training objective, these samples can be interpreted as exerting an inverse influence on the model output. A sample with label y_i and negative weight $w_i < 0$ can be understood either as reducing the network's confidence for class y_i , or equivalently, as increasing the confidence assigned to all other output nodes. This interpretation allows a reformulation in which negatively weighted samples act as a repulsive term in the loss landscape, effectively pushing the model output away from the corresponding label y_i .

To achieve numerical stability and interpretability, one can replace this repulsive interaction with an equivalent

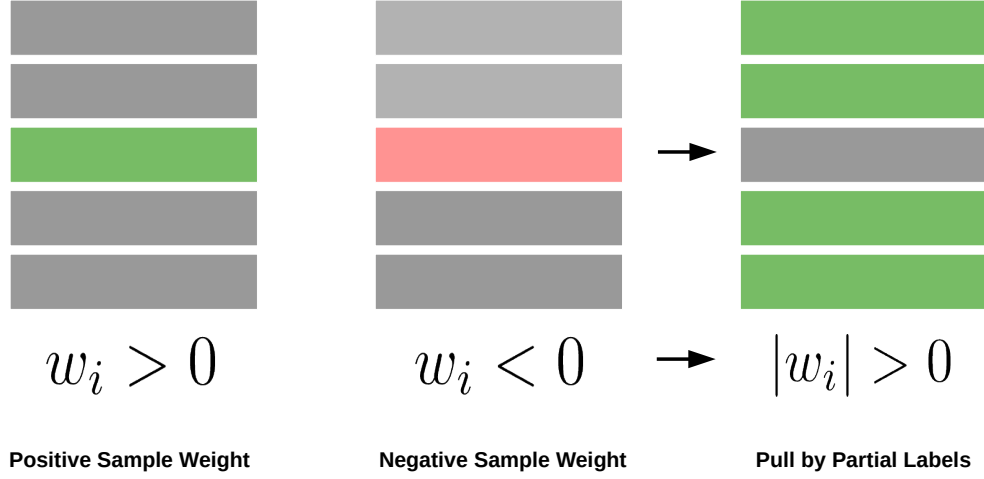


FIGURE 8.4: Schematic of the relabelling done as part of the PPL method described in the text. The left column shows the scheme for a normal, positively weighted sample. The center introduces the label vector for a negatively weighted sample, with the uniquely defined label shown in red. The relabelling is shown in the right column with all label vector components positively weighted except for the former negatively weighted label component.

attractive one acting on all other labels. Formally, this is a reinterpretation of the negative samples y_i by redefining their target label as the complement of the original label

$$Y_i \leftarrow Y \setminus y_i, \quad (8.8)$$

and replacing the negative weight by its rescaled absolute value:

$$w_i \leftarrow \frac{|w_i|}{N-1}, \quad (8.9)$$

where N denotes the total number of output classes in the training set.

This rescaling preserves the total effective contribution of the sample to the overall dataset weight, ensuring that the relative importance of each training instance remains consistent.

The transformation thus converts negative quasi-probabilities into a stable, attractive contribution distributed over all non-target classes.

In practice, this procedure is applied to all negatively weighted samples prior to training. The model is then optimized using the classifier consistent loss (CCL), defined analogously to Equation 4.4, with each sample weighted by its respective (transformed) weight.

The main advantage of this approach lies in its numerical stability. The transformation replaces the repulsive dynamics introduced by negative weights with an equivalent attractive force distributed over the remaining output classes, thereby ensuring stable and interpretable gradient updates during optimization.

8.4 PROOF OF CONCEPT FOR PPL: REVISITING MNIST

To prove the qualities of PPL introduced in the last section, the MNIST experiment from section 8.2 is revisited. A good learning algorithm should retain the accuracy level of $> 99\%$ that CNNs achieve on the positively weighted classes. The results for all 10 classes are shown in table 8.1 comparing training with negative weights and using PPL. The accuracies are averages over 5 trainings and evaluated on the MNIST test set without the relevant class. The performance losses discussed above are remedied by PPL. The method manages to restore the accuracy from the original training. The test accuracy on the negatively weighted class is close to 0 in both cases across all classes. An explanation can be found in the loss curves shown in figure 8.5. The catastrophic divergence discussed in the section above is evident from the loss curve of optimization with negative weights. The PPL loss curve however is convergent. In fact, the convergence happens to a minimum that has a comparable accuracy on the positively weighted classes.

The method was also checked for stability with increasing dataset size for negatively weighted samples in a similar setup as in section 8.2. The results are shown in the right plot of figure 8.5. It shows results with different percentages of the class 0 weighted negatively. PPL shows consistent accuracy over all positive samples which

Neg. Class	Vanilla Training	Trained with PPL
None	99.3%	99.3%
0	77.69%	99.15%
1	71.61%	99.12%
2	78.82%	99.18%
3	78.99%	99.16%
4	80.44%	99.21%
5	81.90%	99.26%
6	76.31%	99.18%
7	77.59%	99.24%
8	83.44%	99.22%
9	81.85%	99.25%

TABLE 8.1: Comparing the accuracy of a CNN trained on MNIST with the given class weighted negatively (Vanilla Training) vs. the negativity mitigated by PPL. Evaluations were done leaving the designated class out. Results are averages over 5 trainings. Table published in Ref. [201].

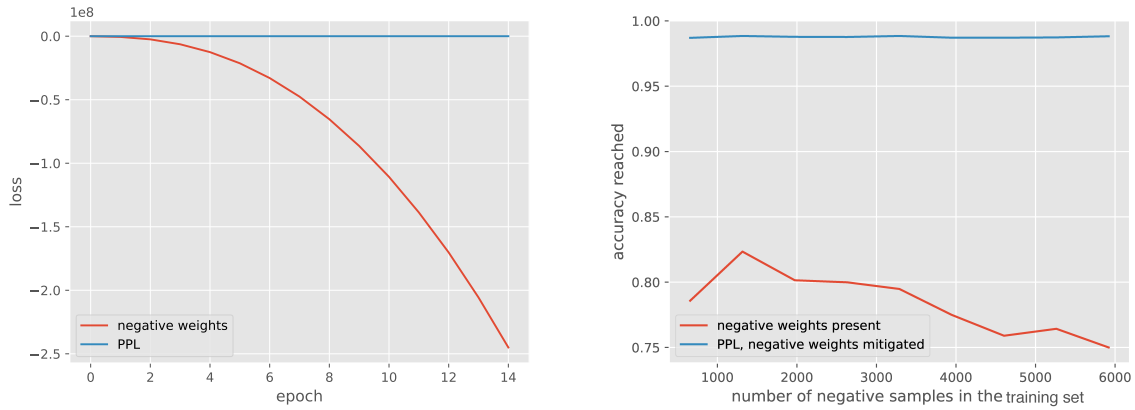


FIGURE 8.5: Training behavior of training with negative weights vs. PPL for the MNIST experiment with negative sample quasi-probabilities in class 0. The left plot shows the loss functions over 15 epochs and the right plot illustrates the accuracy outcomes for training scenarios where different fractions of class 0 were assigned negative weights. All evaluations were done on the test set without the negatively weighted class 0. Figures published in Ref. [201].

proves its stability. The problem of divergence is solved and the fraction of negative samples does not have an impact on this solution.

To understand the network response, the Grad-CAM [195] heatmaps of a training with PPL in class 0 are shown in figure 8.6. The heatmaps reveal that the model trained with PPL continues to activate on the relevant image regions that correspond to meaningful class features, even though the corresponding class (in this case, class 0) was negatively weighted. This result indicates that the model successfully retains its internal feature representations, avoiding the destructive impact on shared features that typically occurs in negatively weighted training. Such retention is critical in multi-class settings, where feature overlap between classes can significantly influence generalization performance. Hence, PPL does not merely suppress the targeted class but ensures that the remaining network structure continues to encode semantically consistent and transferable representations. It can be concluded that PPL fulfills the desired qualities

- It is stable in convergence.
- It finds a minimum which performs well on data points not impacted by negative weights.
- It actually leads to desired degradation of performance on the negatively weighted samples.
- It retains learned features even on the negatively weighted class.

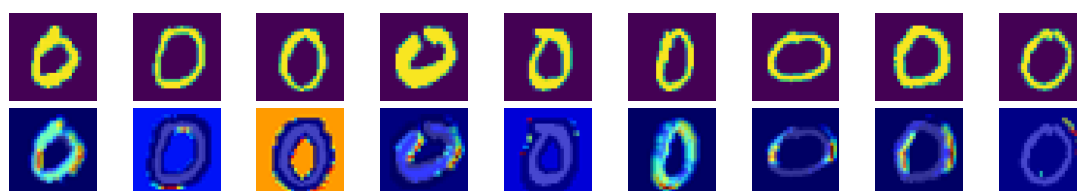


FIGURE 8.6: Grad-CAM [195] maps for a CNN trained with class "o" having gone through the PPL relabelling. The first row shows the corresponding samples without a heatmap overlay. Figure published in Ref. [201].

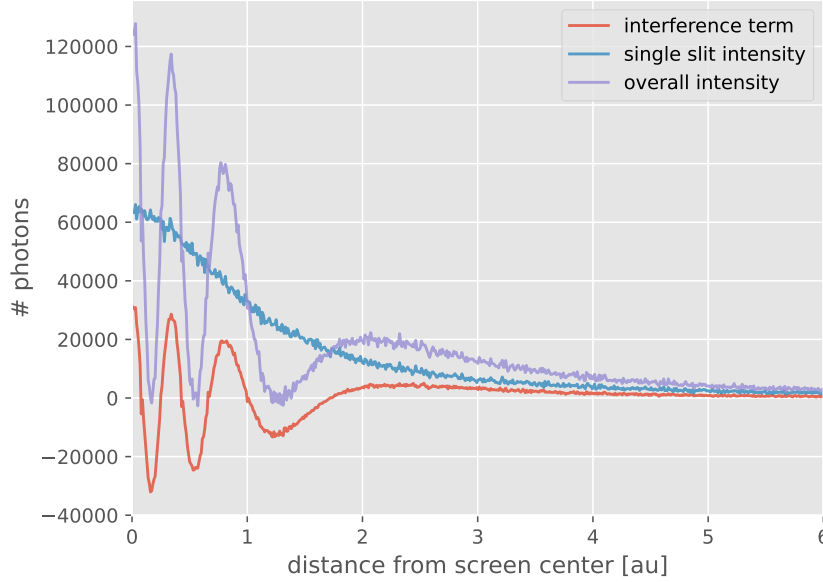


FIGURE 8.7: Datasets constructed according to equations 8.11 and 8.12. The amount of photons seen at distance x from the screen center is plotted in arbitrary units of length.

8.5 EVALUATING STRATEGIES FOR DEALING WITH NEGATIVE WEIGHTS IN INTERFERENCE SCENARIOS: THE DOUBLESPLIT EXPERIMENT

Benchmarking the performance of training strategies in the presence of negative training weights poses a conceptual and practical challenge, as there exists a fundamental divide between the training and application domains. While negatively weighted events appear naturally in the training data, the application in the CMS experiment (or any particle detector) is ultimately based on event counting and hence strictly positive event yields. To systematically investigate the behavior of different training strategies under such conditions, therefore a controlled toy model is constructed. It reproduces the essential characteristics of the physical problem: a training sample containing negative contributions and a corresponding “physical” test sample that is strictly positive definite.

A canonical analogue of interference phenomena in particle physics can be found in the double-slit experiment, which provides a conceptually clean and analytically tractable setup. In this scenario, the particle wavefunction of for example electrons or photons gives rise to an interference pattern on a detection screen, following Huygens’ principle. The total wave amplitude at a position x on the screen is given by the coherent sum of the contributions from the two slits, $\Psi_1(x)$ and $\Psi_2(x)$. The observable intensity is then

$$I(x) = |\Psi_1(x) + \Psi_2(x)|^2 = \underbrace{|\Psi_1(x)|^2 + |\Psi_2(x)|^2}_{\text{resonance}} + \underbrace{2\text{Re}(\Psi_1(x)\Psi_2(x))}_{\text{Interference}}, \quad (8.10)$$

with the resonant contributions amounting to

$$|\Psi_{1/2}(x)| = \frac{1}{2} \frac{dN}{dt} \frac{d/\pi}{d^2 + (x \pm a/2)^2}, \quad (8.11)$$

and the interference being

$$2\text{Re}(\Psi_1(x)\Psi_2(x)) = \frac{dN}{dt} \frac{d/\pi}{\sqrt{d^2 + (x - a/2)^2} \sqrt{d^2 + (x + a/2)^2}} \cos \left[(h/\lambda) (\sqrt{d^2 + (x + a/2)^2} - \sqrt{d^2 + (x - a/2)^2}) \right]. \quad (8.12)$$

The rate of incoming particle is $\frac{dN}{dt}$ and the slit geometry is encoded by the distance a between the slits. The distance to the detection screen is d and λ is the wave length. The total intensity $I(x)$ is, by construction, a positive-definite function and can be interpreted as a probability density from which a physical, positively weighted test dataset can be sampled. Samples drawn exclusively from the resonant terms are assigned the label 0, while samples including the full coherent sum (resonance plus interference) are labelled 1. A dataset constructed with this procedure is shown in figure 8.7. The $\frac{dN}{dt}$ was chosen as 1, $d = 1$, $a = 0.1$ and $\lambda = 0.005$. These are arbitrary choices that produce a visually distinct interference pattern. The events are referenced as photons from now on for simplicity.

For the training process, additionally datasets based on the individual quasi-probability distributions corresponding to the resonant and interference contributions are constructed. Since the interference term can locally take on negative values in phase space, events drawn from regions where the interference contribution is negative are assigned a weight of -1 . This setup thus emulates the situation encountered in Monte Carlo event generation, where destructive interference leads to negatively weighted events despite the overall cross section remaining

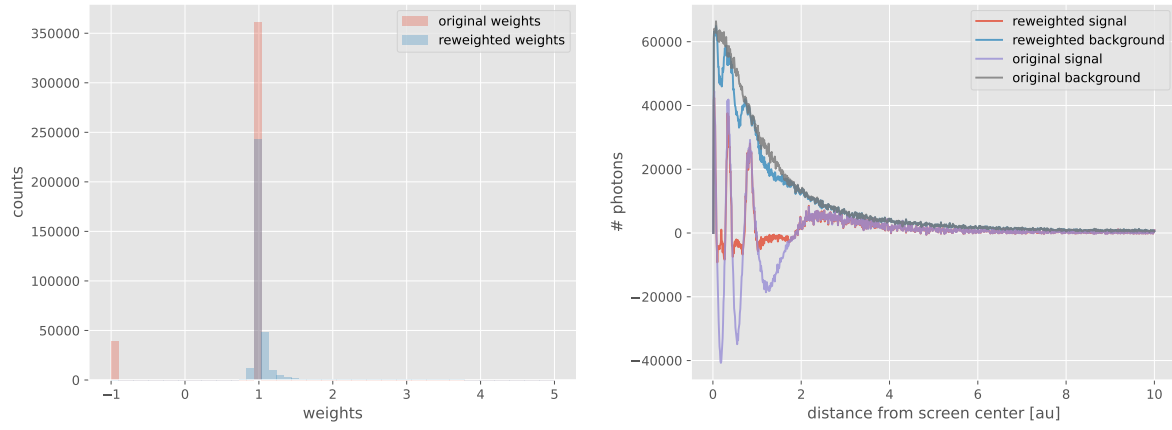


FIGURE 8.8: The plot shows the reweighting results based on the algorithm in Ref. [202]. The left plot shows the weight distribution before and after reweighting. The right plot shows the distribution of photons on the screen before and after reweighting split into the signal (= interference) and background (= resonance) distribution defined in the text.

positive.

A fully connected DNN is trained to distinguish between events originating from the interference pattern and those from the pure resonance contribution under different data treatment strategies. Specifically, the following training configurations are evaluated:

1. A weighted dataset including negative event weights.
2. A reweighted dataset in which negative weights are removed through the neural reweighting technique proposed in Ref. [202].
3. A weighted dataset with manipulated labels using the PPL method.
4. A dataset with an additional output class explicitly assigned to negatively weighted events, allowing for unweighted training after class reassignment.
5. A dataset with negative weights set to their absolute value.
6. A dataset with negative weights set to 0.

This experiment enables a controlled, quantitative comparison of how different approaches handle negatively weighted training data in a setup that closely mirrors interference effects in realistic scattering processes.

Since this task is very challenging on the single input of the screen position x alone, a second feature is added. This helps with accessing the performance of the tested modes of training, as meaningful convergence is possible. The second feature carries the same weights given by the quasi-probability distributions shown above. It is sampled from a normal distribution, where the mean value of the signal data is moved to the right with respect to the background distribution which is centered around 0. One can tune the power of this second feature by scaling this shift. For the tests below, a value of $x \rightarrow x + 2$ was chosen.

To validate the reweighting procedure, a DNN classifier was trained to distinguish the overall weighted signal and background sample from an unweighted sample with the same weights. The results can be seen in figure 8.8. The left plot shows the weight distribution before reweighting, featuring a contribution of negative samples with $w = -1$. The weight distribution after reweighting is positive, as expected from the definition of the algorithm. The weights cluster around the value of $w = 1$, with a spread smaller than 1. The reweighting can thus be considered successful.

The output distributions of the reweighting are shown in the right plot.

The background distribution agrees with some disagreements at x values which feature a destructive interference in the signal distribution. The signal distribution agrees up to those x ranges as well. Since the reweighting is done on the whole dataset some migration of the destructive interference feature into the positively weighted background is expected. This is a potential issue when using classifiers on this reweighted data. A feature that was unique to the signal dataset before reweighting is now present in both classes, making classification harder. The magnitude of this effect is important. In this case, the features of signal and background are still very distinct.

To judge the classification performance, a DNN was trained to distinguish the signal from background. The model consists of 4 fully connected layers with 128 nodes each. An ReLU activation function and a dropout of 0.1 are applied to each layer. The output between two classes is normalized with a softmax function. Models were trained for 20 epochs with a learning rate of 0.001 and the Adam optimizer [203].

The classification results are shown in figure 8.9. The ROC curve in the figure is evaluated with respect to the background (single slit term) as the signal definition varies for the method utilizing the extra class. The figure shows

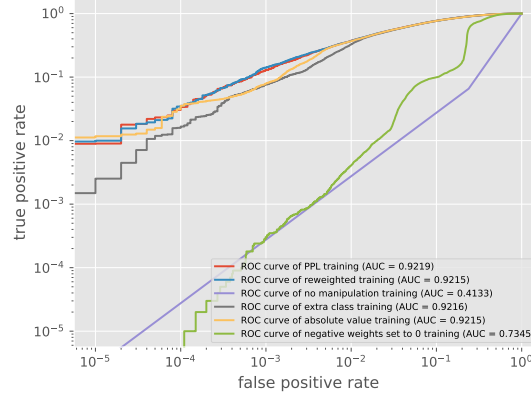


FIGURE 8.9: ROC curves for the classification task of doubleslit photons with and without interference evaluated for the background dataset drawn from a single slit. Different ROC curves are drawn for the discussed methods of dealing with negative weights during training: PPL method, reweighting, no manipulation (training including negative weights) and training with an extraclass for negatively weighted events.

that without a sufficient treatment of negative weights, no amount of learning is possible. The ROC curve of the training including negative weights falls even below the straight line, which would indicate performance on par with guessing. Also just letting the negative weights out of training does not achieve the performance of the other methods. The area under curve (AUC) of 0.73 falls way below the 0.92 that the other methods can reach.

All remaining four approaches to mitigate the negative weights are comparable in performance, reaching AUC scores of ~ 0.92 indicating that the instabilities have been overcome.

To further interpret the classifier learning behavior, figure 8.10 presents the distributions of samples predicted as signal for each method. The upper-left panel shows the true signal distribution, serving as a reference. The lower panels display, for each mitigation approach, the deviation of the predicted signal fraction from this true distribution. This visualization provides direct insight into the regions of phase space where each model deviates from the correct signal classification.

It is evident that the model trained with the original negatively weighted samples collapses to background-only predictions, which fully explains its poor overall performance.

The reweighting approach shows the largest deviations in regions dominated by constructive interference. This behavior can be attributed to the feature migration between signal and background that is intrinsic to the method as discussed above.

The model trained with negative weights set to zero exhibits its largest errors in regions characterized by destructive interference. This outcome is expected since these regions correspond to signal contributions with negative weights, which are effectively removed from training, leaving the classifier blind to these parts of phase space.

The model trained with the PPL method shows errors primarily around the x -values where the interference term crosses zero, corresponding to points with no net signal contribution. At these locations, the gradient vanishes, preventing effective learning. The classifier trained using absolute values of the weights shows a similar pattern, consistent with the absence of gradient information in these regions. In both cases, the resulting performance can therefore be understood as a natural consequence of the training rather than an artifact of the mitigation of negative weights.

No distribution is shown for the classifier trained with an additional extraclass to represent negatively weighted events. The added class is not physically interpretable and must be regarded as a methodological artifact by construction. Attempts to mitigate this issue, such as computing the Softmax probability only over the physically meaningful signal and background classes, lead to a substantial degradation in performance. Additionally, excluding events assigned to this artificial class would result in a loss of usable samples, thereby worsening the statistical precision in a realistic particle physics analysis context. This can be considered as a substantial simulation-experiment divide.

The experiment was repeated for different values of the mean shift μ in the Gaussian distribution of the second feature to test robustness against varying discriminative power. Excluding this second feature entirely leads to model collapse across all methods, rendering results meaningless. Additional results, shown in the appendix, correspond to a weaker second feature with $\mu = 0.1$ compared to $\mu = 2$ in the main setup. As illustrated in figures A.11 and A.12, the overall performance decreases for all methods under these conditions. However, both the PPL approach and the model trained with absolute weight values remain the most robust, showing the best generalization and feature stability. In the more challenging classification scenario, their advantage over other methods becomes even more pronounced.

In summary, the PPL and absolute-weight training strategies demonstrate the highest stability and accuracy across varying feature configurations. These results suggest that both approaches can effectively bridge the gap between

simulation and experiment. The interpretable nature of their learned feature distributions provides an additional layer of validation.

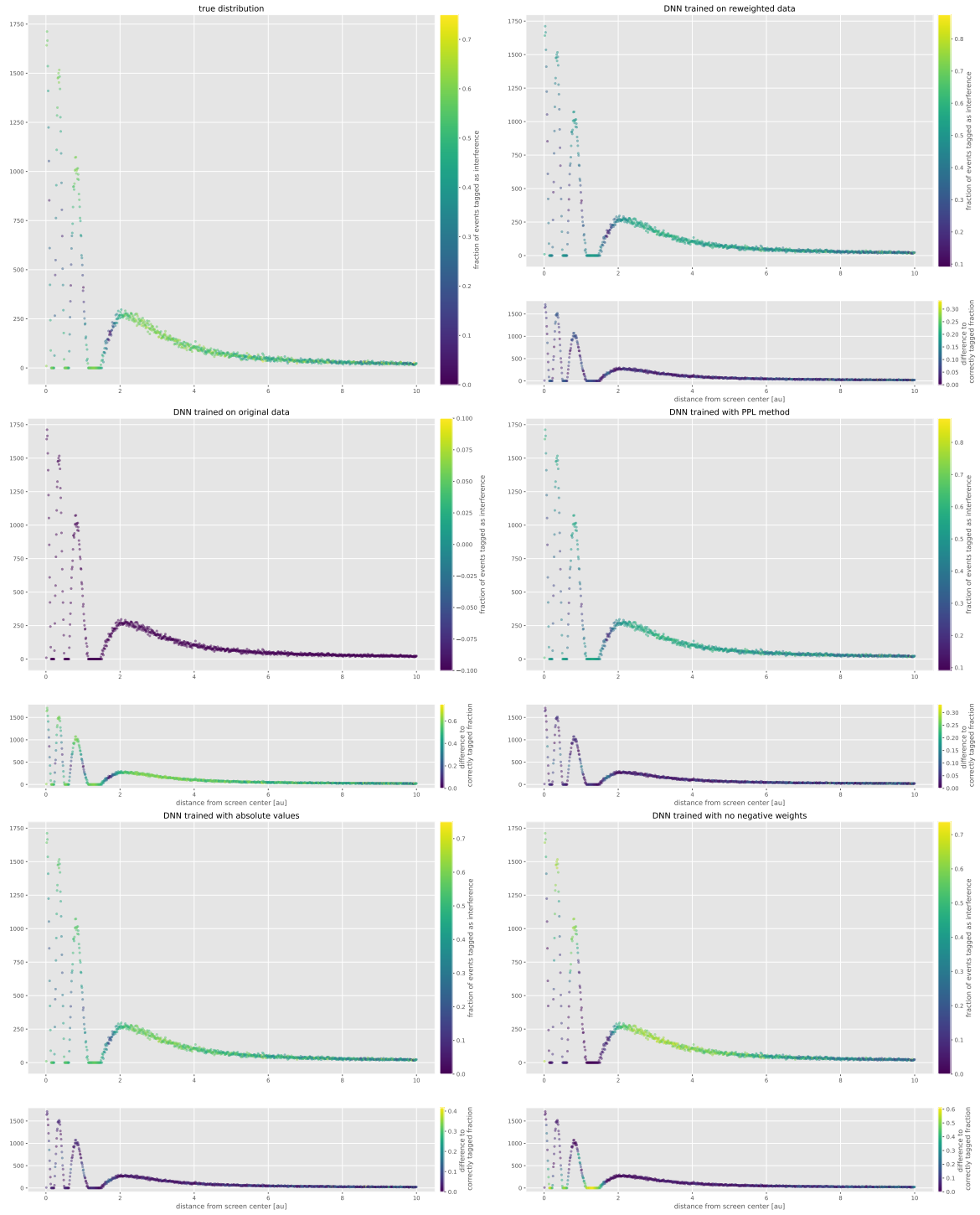


FIGURE 8.10: Distributions of percentages predicted as coming from a doubleslit. The lower panels show the difference to the correct percentage in the analytical distribution. Different negative weight mitigation methods are plotted in their respective panels.

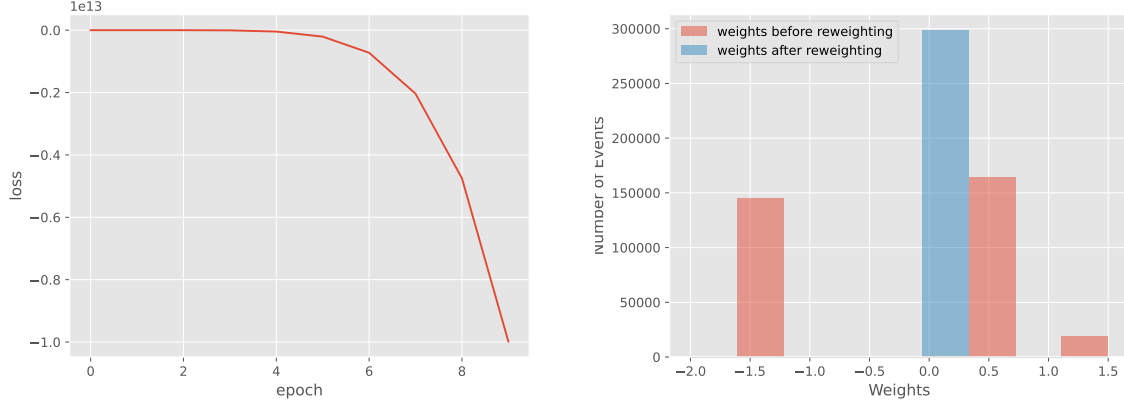


FIGURE 8.11: The left panel shows the loss of the classification PFN [85] network used on the scenario with a pseudoscalar A with 365 GeV and a width of 10%. It is evident that the presence of negative weights induces catastrophic divergence. The right panel shows the resulting reweighted weight distribution against the original MC weights. The catastrophic divergence leads the reweighting to collapse to only predict weights of 0. These are non-negative but also unusable for further physics analysis.

8.6 ANALYZING NEGATIVE WEIGHT MITIGATION METHOD IN A MOCK ANALYSIS SETUP

Having studied the issue of negatively weighted events in classification tasks and established the performance of different mitigation approaches, it is necessary to find a conclusion on the problem that started this work. For this reason, a mock analysis in the vein of the heavy Higgs search of chapter 6 was set up and the performance of different mitigation methods was again evaluated. The classification task in this scenario is the discrimination of a heavy Higgs (pseudo)scalar BSM signal A or H with different masses and widths against a SM top quark pair production background.

While the experimental-simulation divide cannot be studied as cleanly as in the double slit toy model, it is possible to set up a pipeline that allows for a close-to-experimental evaluation.

The training and evaluation data are generated with MADGRAPH5_AMC@NLO [184] interfaced with PYTHIA 8 [186] and an UFO model file that implements a pseudoscalar A and a scalar H [204]. For this, a set of signal samples (including resonance and interference) were generated on a grid of masses $\times$ width for A and H

$$[A, H] \times [365, 400, 500, 600, 800, 1000] \text{ GeV} \times [5, 10]\%. \quad (8.13)$$

A set of high level observables was computed on GEN information after parton shower and hadronization which are used as input features to the DNN classifier, no detector simulation was performed. The input feature distributions are shown in the appendix in figure A.13 for one mass-width scenario. It is evident that for the chosen binning some bins in the signal (which consists of resonance and interference) are negative in some kinematic observables. The ratio of negative to positively weighted events varies between the signal hypotheses and reaches $\sim 31\%$ at its maximum for the pseudoscalar A with a mass of 365 GeV and a width of 10%.

Again, the reweighting is validated by comparing the weight distributions and the feature distributions of the data after reweighting. In addition to the issue previously identified in the double slit concerning the compensation of negative contributions in the signal by less positive contributions in the background, fundamental issues with the training of the reweighting classifier emerge. Given that a classifier is trained between a weighted distribution and a flat-weighted distribution with $w = 1$ of the dataset, convergence problems caused by negative weights carry over into this training as visualized in figure 8.11. It is clear that for the pseudoscalar A at 365 GeV with 10% width training is not possible. The loss diverges to negative infinity and the resulting network is collapsed to only predict weights of 0. Steps for mitigating this problem have been taken by ensuring larger batch sizes ($n_{\text{batch}} = 1000$) and using the more powerful and stable ParticleFlow Network architecture as described in section 4.1.5 and proposed in the DNN reweighting paper [198]. It is evident that for some signal scenarios, the reweighting can not offer a solution as the negative weight fraction exceeds the threshold to catastrophic divergence.

The performance of the other methods has to be judged on the output of the trained classifiers. A classifier was trained for each signal scenario against the background $t\bar{t}$ production with each of the following methods:

- PPL,
- normal training with negative weights,
- normal training with negative weights set to 0,
- normal training with negative weights set to their absolute value,
- training of a classifier with an extra output class for the negatively weighted events, weights set to their absolute value,
- training on reweighted data where reweighting was successful.

The initial performance of the classifiers can be judged from their ROC curves. This is a pure evaluation of the performance in the context of training objective and does not take into account physical implications for further analysis. Two signal hypotheses ROC curves are shown in figure 8.12 and the mean ROC result over all scenarios. The upper left panel shows a scenario with a high ratio of negative weights to positive weights (0.27) and the upper right panel shows a signal scenario with few negatively weighted events (0.12). As expected, the largest performance difference between the two scenarios is seen in the basic training with negative weights which is most sensitive to additional negative weights. Also the reweighting succeeded for the two scenarios but it can be seen how a larger ratio of negative weights degrades the performance. All the other methods showing comparable performance between the two signal hypotheses indicating stability with respect to the ratio of negative weights in the training set. The AUC ordering of all methods is consistent across scenarios and is also observed in the mean result across all trainings. PPL, setting negative weights to their absolute value or 0 score the highest, training with an extraclass degrades the performance slightly and leaving negative weights in only works in selected scenarios. This is indicated by the large standard deviation across different trainings for this method. Reweighting consistently performs the worst due to the instabilities observed and the feature migration between classes in training data.

This results is consistent with the findings of the double slit experiment and gives further confirmation that from a ML mechanistic perspective the PPL method as well as leaving out negatively weighted events or setting the weights to their absolute values (and to some degree in extension classifying them in a separate class) cures the problem.

8.6.0.1 Impact of Negative Weights and their Mitigation on Statistical Inference for Collider Physics

To study physical implications of the methods, one has to evaluate them on the goal of an analysis in particle physics: Making a statement about an underlying theoretical process. This can be a measurement, a detection or exclusion limits. For this, a maximum likelihood fit is performed, as described in section 3.3. The fit templates are created from the trained DNNs and validation data. The data is passed through them in a forward pass and the distribution of the output node 0 is binned into 10 evenly spaced bins.

From classical, histogram based analyses there are methods available to circumvent the negative weights in the likelihood fit and hypothesis testing steps. Equation 2.4 shows how an interference enters a cross section calculation. Suppose one is interested in finding the exclusion limit on a signal strength μ on the cross section of process one with amplitude $\mathcal{A}_1$, one can write it as

$$\text{yield} = |\sqrt{\mu} * \mathcal{A}_1 + \mathcal{A}_2|^2 = \mu[S] + \sqrt{\mu}[I] + [B]. \quad (8.14)$$

Here, S describes the signal yield ($[S] = |\mathcal{A}_1|^2$), [I] the interference yield and [B] the background yield respectively. Completing the square for $\sqrt{\mu}[S + I + B] = \sqrt{\mu}|\mathcal{A}_1 + \mathcal{A}_2|^2$ gives

$$\begin{aligned} \text{yield} &= \sqrt{\mu}[I] + \sqrt{\mu}[S] + (\mu - \sqrt{\mu})[S] + \sqrt{\mu}[B] + (1 - \sqrt{\mu})[B] \\ &= \sqrt{\mu}[S + B + I] + (\mu - \sqrt{\mu})[S] + (1 - \sqrt{\mu})[B]. \end{aligned} \quad (8.15)$$

This algebraic trick effectively transforms the yield into a function of positive terms. It is recommended by CMS for setting limits with the COMBINE code if the signal interferes with the background [205, 206].

To create the signal, background and signal+background+interference template for the fit, the trained classifiers are evaluated on the respective simulated validation samples. For the [S+I+B] sample, the process was simulated without passing intermediate particles to MADGRAPH5_AMC@NLO. The ME generator creates a positive definite sample in this scenario accounting for all contributions. To single out the contribution of either A or H, the coupling parameter for the respective other particle is set to 0.

To gain insight into the stability of the methods with respect to perturbations and systematic effects, the DNNs were also evaluated with a "varied" input dataset. For this, a normally distributed value was chosen around the input values with a 5% standard deviation, simulating statistical fluctuations. Evaluation results are shown for the different methods in figures 8.13 and 8.14 for a low mass pseudoscalar hypothesis. Training with negative weights,

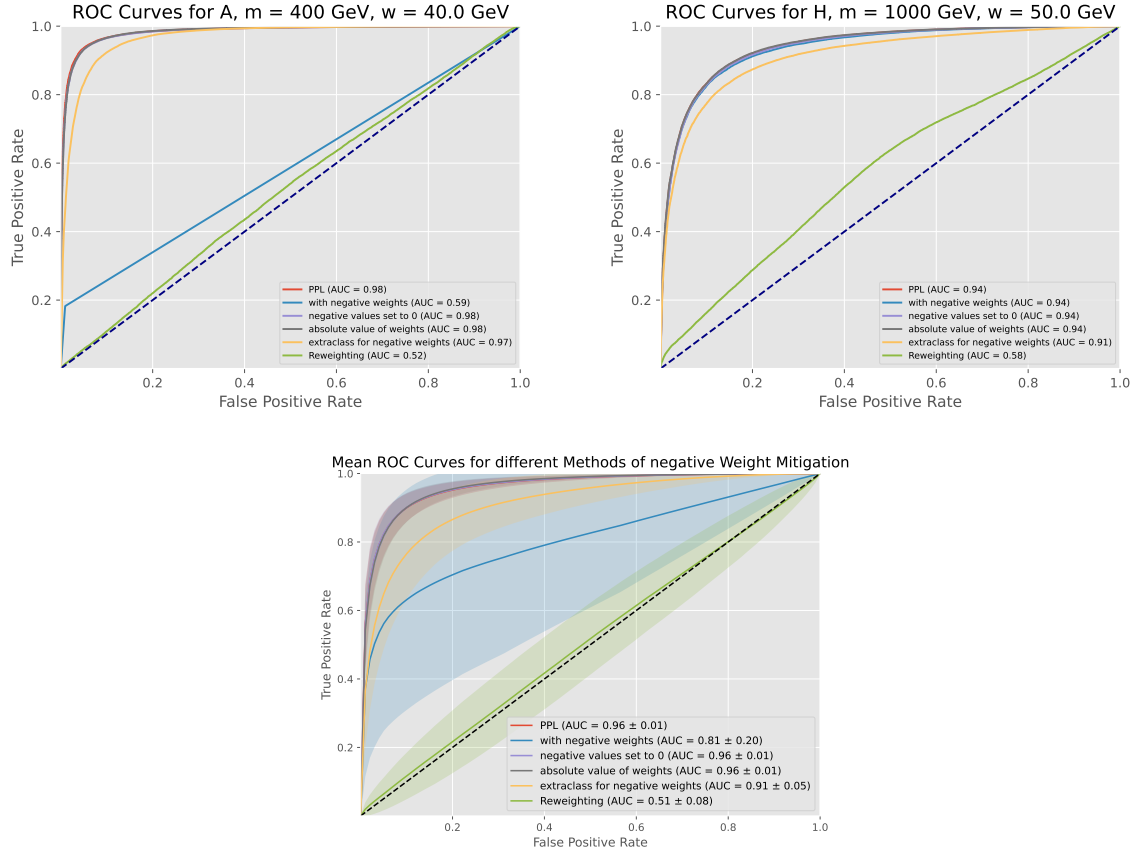


FIGURE 8.12: Performance of classifier networks with different methods of treating negative weights. The upper left panel shows a signal scenario with a large contribution of negative weights (Pseudoscalar A, $m = 400$ GeV and 10% width) and the upper right shows a signal scenario with few negative weighted samples in the training data (Scalar H, $m = 1000$ GeV, $w = 5\%$). The lower panel is a summary plot that shows the mean ROC curves for all signal scenarios. The shaded area gives the standard deviation from that mean curve for every method. Each plots has a diagonal line indicating the ROC curve of a random classifier. All plots show all lines as indicated in the legends. In scenarios where these are not visible, lines with highly similar AUC scores overlap.

Method	χ^2 -distance on [S+B+I]	χ^2 -distance on [S]	χ^2 -distance on [B]
PPL	0.05 $\pm$ 0.02	0.06 $\pm$ 0.04	0.09 $\pm$ 0.05
Absolute Value for negative Weights	0.05 $\pm$ 0.03	0.06 $\pm$ 0.04	0.09 $\pm$ 0.04
Negative Weights set to 0	0.06 $\pm$ 0.06	0.06 $\pm$ 0.01	0.09 $\pm$ 0.11
Training with negative Weights	0.04 $\pm$ 0.02	0.05 $\pm$ 0.04	0.05 $\pm$ 0.06
Extra Class for negative Weights	0.03 $\pm$ 0.03	0.00 $\pm$ 0.04	0.09 $\pm$ 0.05
Reweighting	0.34 $\pm$ 0.35	0.33 $\pm$ 0.34	0.34 $\pm$ 0.35

TABLE 8.2: χ^2 -distance between the classifier output distribution of node 0 for nominal and varied data per method presented as means across all signal hypotheses. The methods are divided into stable methods (upper three methods) and the methods which exhibit collapsed output distributions (lower three methods).

training with an extra class and reweighting show collapsed output distributions with little discrimination power due to similar distributions in [S] and [B] templates. The methods that have proven stable in the ROC benchmark (absolute value, negative weights set to 0 and PPL) also predict sensible output distribution, showing a peak at low values for signal and a peak at high values for background events.

The [S+B+I] is dominated by background events so the expected behavior would mirror the [B]-only template with more events at low values. Only PPL fulfills this expectation, setting negative events to 0 and taking the weights absolute value lead to a signal-like peak at low values. This shows the missing or misrepresented influence of the negatively weighted events. They contribute only to the signal dataset and omitting them yields an overclassification of ambiguous events as signal. This effect is even stronger when they influence the model positively during training by setting the weights to their absolute value.

The collapsed classifiers from instable methods do not have an interpretable difference between varied and nominal data. For all other methods the variation generally leads to less confident predictions, shifting the signal or background peaks from the boundaries to more central values of output activation. This is consistent across the methods and expected. The added noise should make the signal and background harder to distinguish. To quantify the impact of the variation, the χ^2 -distance between the classifier output on nominal data and varied data is calculated. The results are presented as a mean across all signal hypotheses in table 8.2. Low values indicate a larger similarity of output distributions on varied and nominal data. Using the weights absolute value, setting negative weights to 0 and PPL score very comparably, indicating that the optima chosen in training are comparably stable with respect to statistical fluctuations in the training data. It is meaningless to compare the collapsed output of the other methods.

The varied output was symmetrized and chosen as a NP template in the fit. Additionally, a 10% normalization uncertainty was added. The fit was then performed with the COMBINE package [205]. The results are shown in figure 8.15. The performance of the classifiers propagates into the limit results. Using the absolute value, leaving the negative weights out and the PPL method perform comparably well, the other methods fall short in comparison. For some signal hypotheses with fewer negative weights, the classical training including the negative weights is competitive.

This shows that any correct stabilizing method can be sufficient to achieve good physics performance. The heavy Higgs scenario discussed here is sufficiently simple to classify at GEN level to arrive at this conclusion. The nominal templates shown in figure 8.13 and 8.14 show very clearly separated resonant signal and background for the aforementioned best performing methods. Subsequently, the negative-weight-dominated interference contribution is not the driving factor in sensitivity for these. Including a full systematic treatment and detector effects into this highly sensitive classifier output region will deteriorate this performance. While the statistical limits are comparable for this simple topology, the absolute value method and leaving negative weights out fundamentally misrepresent the interference physics by flipping the sign of destructive interference, whereas PPL preserves the feature semantics.

The explainable PPL behavior with respect to the [S+B+I] template is the strongest finding here, indicating that the method is the most physically robust.

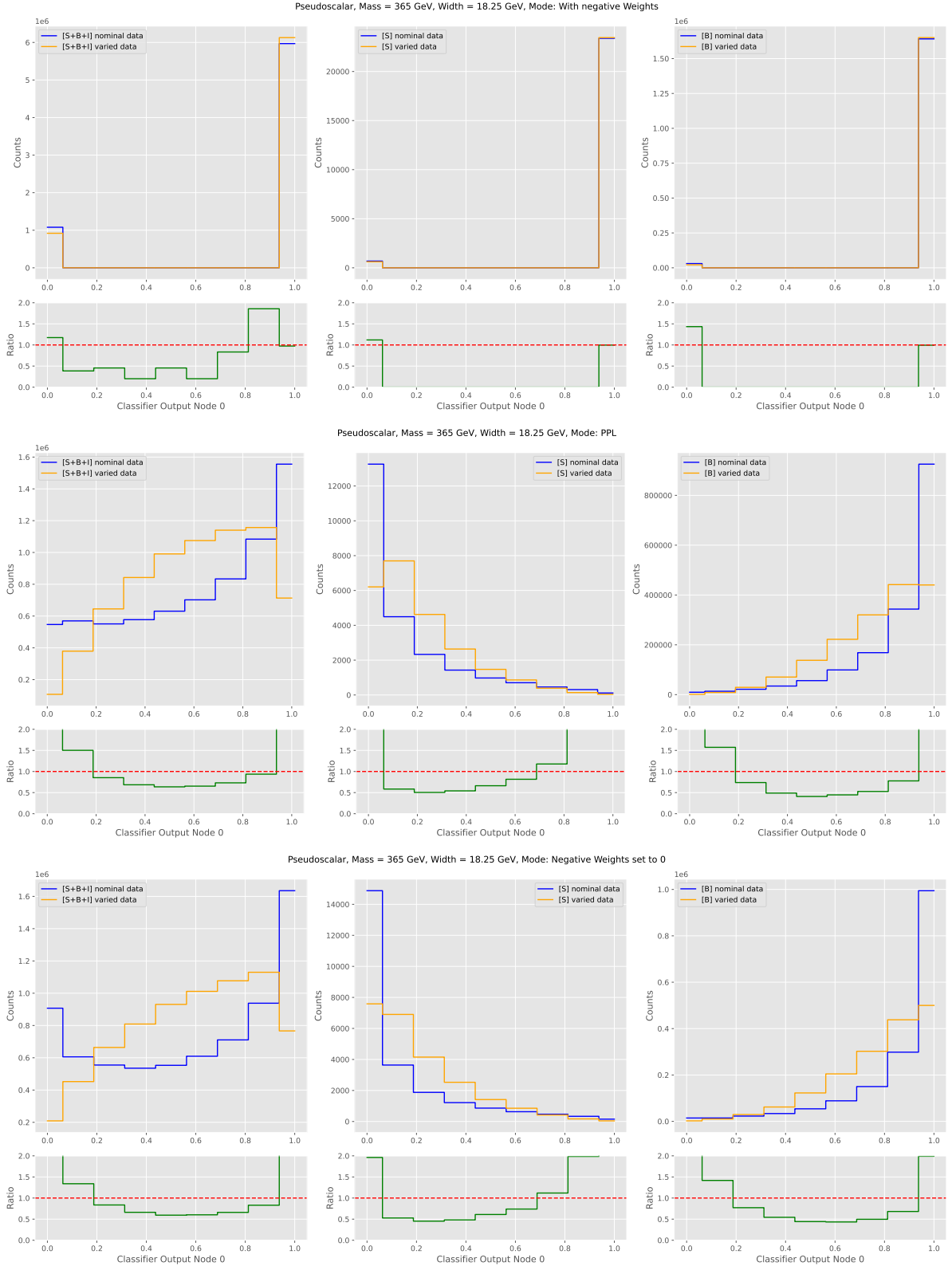


FIGURE 8.13: The panels show the distribution of output node 0 of the DNNs trained with negative weights (upper panel), with the PPL method (middle panel) and with the negative weights set to $w = 0$ during training (lower panel) on the pseudoscalar $m = 365$ GeV, $w = 18.25$ GeV hypothesis. A nominal line is shown as well as a varied evaluation which was produced with a random normal distribution of 5% around the nominal value. The respective lower panel shows the ratio of the varied histogram to the nominal result.

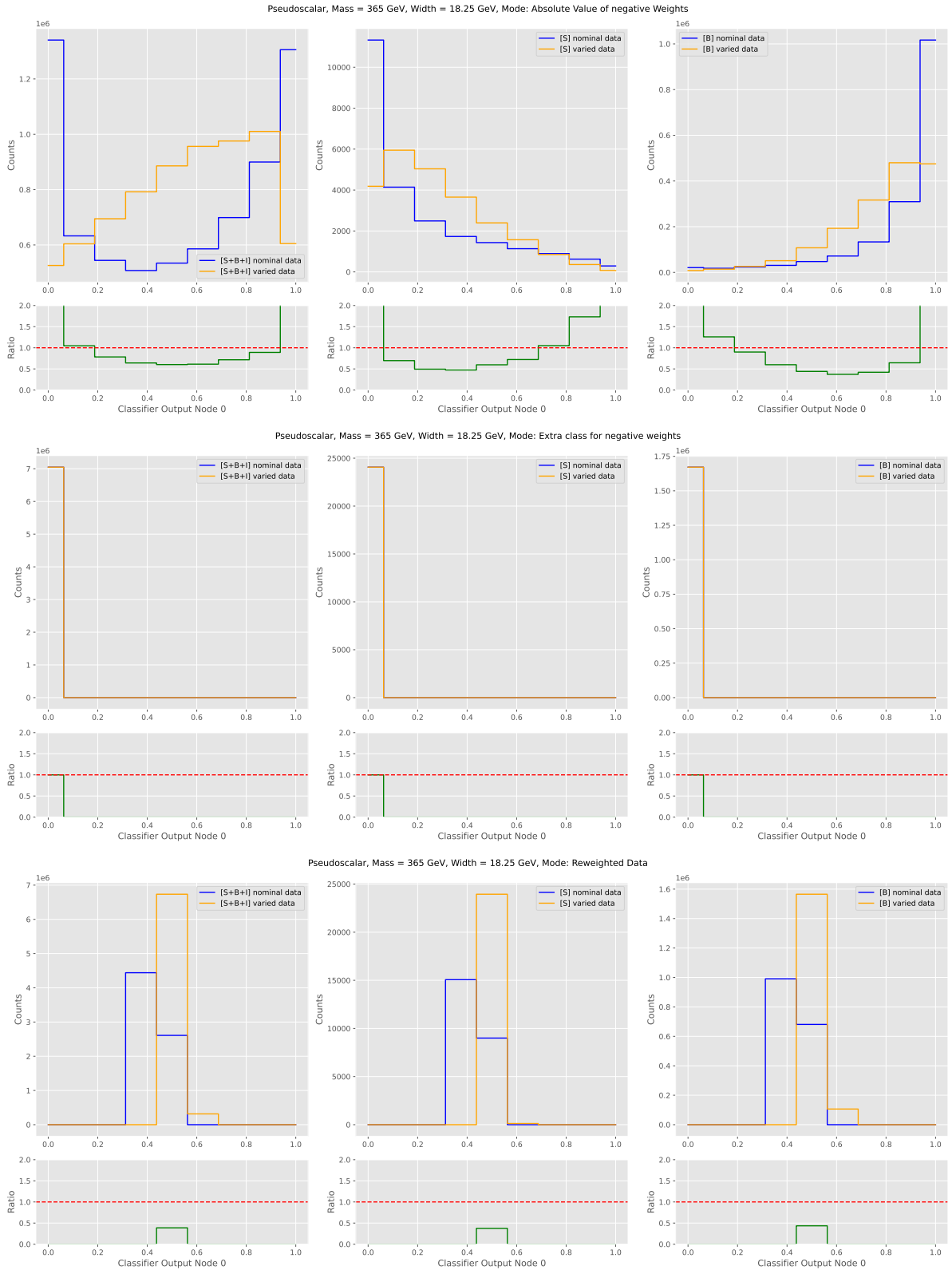


FIGURE 8.14: The panels show the distribution of output node 0 of the DNNs trained with the absolute value of the negative weights (upper panel), an extra class for the negative weights (middle panel) and with reweighting (lower panel) on the pseudoscalar $m = 365$ GeV, $w = 18.25$ GeV hypothesis. A nominal line is shown as well as a varied evaluation which was produced with a random normal distribution of 5% around the nominal value. The respective lower panel shows the ratio of the varied histogram to the nominal result.

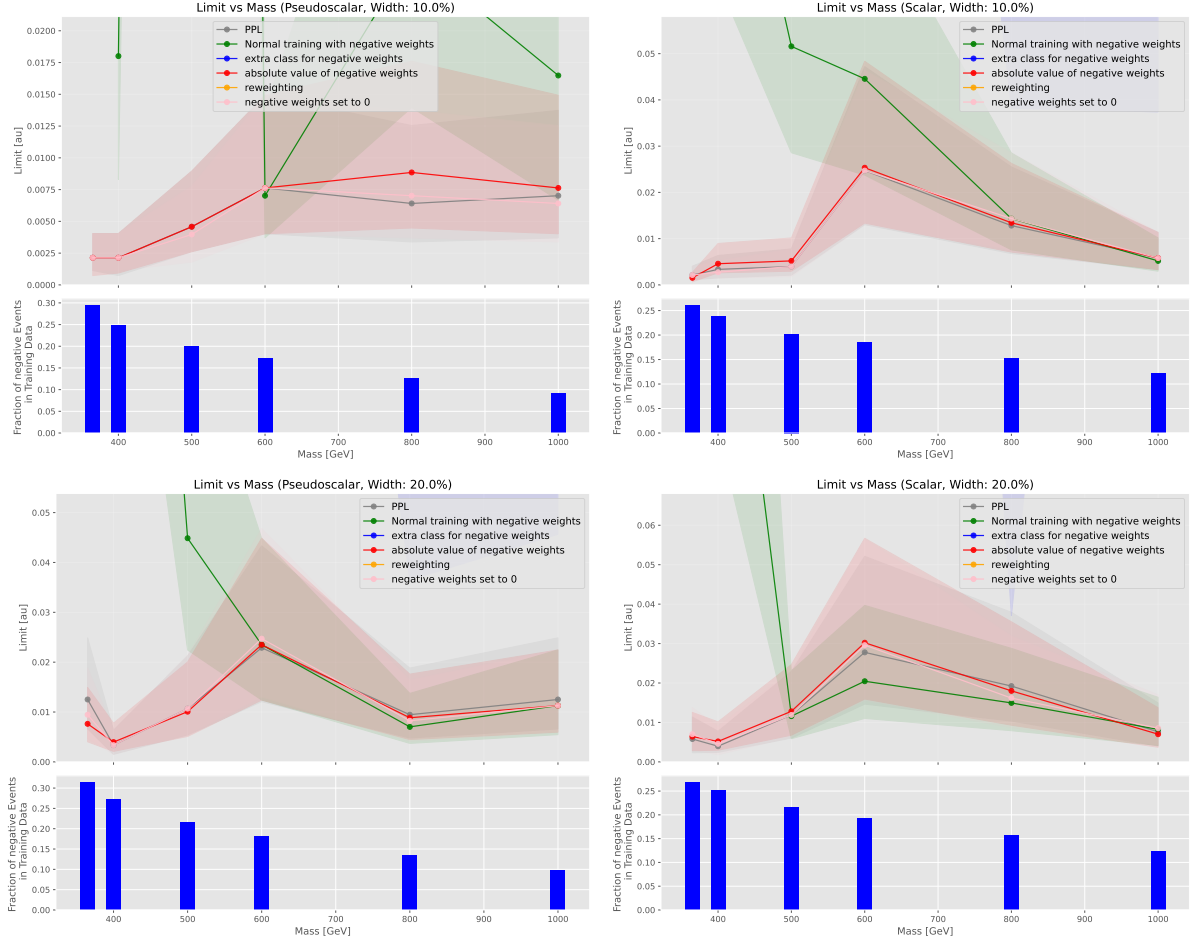


FIGURE 8.15: Limit results derived with different classifier training methodologies. Each panel shows the limit as a function of the hypothesis particles mass. The left column shows limits for two pseudoscalar scenarios with 10% and 20% width. The right panel shows the same results for scalar particles. The bar plot in the respective lower panels shows the fraction of negatively weighted events in the training dataset.

8.6.1 *Note on Likelihood Ratio Estimation*

It should be noted that Ref. [207] discusses the topic of negative weights for training of neural networks for the application of likelihood ratio estimation in particular for reweighting data. They reformulate the problem of likelihood estimation in the presence of negatively weighted data as subproblems which are positive definite by building a mixture model between signed ratio estimators, estimating the LR between positive and positive subgroups of data, the LR between positive and negative subgroups, and vice versa. This solves the problem elegantly and promises strong performance as found in their experiments. It is however not a general solution to classifier training and only applicable when an LR is ultimately computed as the sub-LR-estimators can not be interpreted as classifiers that could be used in a classification based analysis. On the other hand, the PPL method can be used in LR estimation application through the likelihood ratio trick. A proof of concept showing the applicability of PPL to the likelihood ratio problem is given in appendix A.7.

UNLEARNING WITH PARTIAL LABEL LEARNING

Studying the impact of negatively weighted training samples in particle physics naturally asks for other applications of similar features in training data.

Splitting a training set in positively (D) and negatively weighted data ($\tilde{D}$) with weights w and $\tilde{w}$ respectively, the parameter optimization of a ML model with parameters θ and loss $\mathcal{L}$ can be rewritten as

$$\theta^* = \underset{\theta}{\operatorname{argmin}} (w\mathcal{L}(\theta, D) - \tilde{w}\mathcal{L}(\theta, \tilde{D})). \quad (9.1)$$

This formulation bares resemblance of the formulation of ‘forgetting’ a single datapoint by replacing it with another new datapoint in Ref. [208]. They formalize the problem by asking how the optimal classifier θ^* behaves under a small perturbation δ of a data point with features $\mathbf{x}$ and label y in the forget set $z(\mathbf{x}, y) \rightarrow \tilde{z} = z(\mathbf{x} + \delta, y)$. They find for $\epsilon = 1$

$$\theta^* = \underset{\theta}{\operatorname{argmin}} L(\theta; D) + \epsilon l(\tilde{z}, \theta) - \epsilon l(z, \theta), \quad (9.2)$$

where $L(\theta; D)$ denotes the loss on the full dataset and $l(z, \theta)$ denotes the loss on the respective data point. Note, that ϵ works effectively as a training weight for the data point $\tilde{z}$ and as a negative weight for the data point z if we incorporate the sign into the weight. There is thus an obvious parallel between the problem of negatively weighted training data in particle physics and the class of “unlearning” problems in ML literature.

The following chapter is discussing this parallel and consequently applying the PPL method to solve the problem of class unlearning. This work has been published by the author and the supervisors of this thesis in Ref. [201].

9.1 MACHINE UNLEARNING

Machine Unlearning is a branch of ML that deals with the forgetting of information. While classical ML aims to extract and generalize knowledge from large-scale datasets, machine unlearning focuses on the *selective removal of previously learned information* from a trained model [209]. The overarching objective is to ensure that, after unlearning, the model behaves as if selected data - referred to as the *forget set* - had never been part of the training process, while still preserving its performance on the *retain set*. This selective forgetting must be achieved ideally without compromising model accuracy, stability, or generalization on unaffected data.

Formally, unlearning can be viewed as an inverse operation to training: Given a model f_θ trained on a dataset $D = D_{\text{retain}} \cup D_{\text{forget}}$, the goal is to transform f_θ into a new model $f_{\theta'}$ such that its behavior approximates that of a hypothetical model trained only on D_{retain} . This conceptual formulation shows that the unlearning problem is both technically challenging and computationally nontrivial, as modern ML models (particularly deep neural networks) store information in highly distributed and entangled representations.

Existing work in the field shows point-wise unlearning, where the forget set consists of individual data samples [210]. In contrast, the research presented in this thesis addresses the demanding problem of class unlearning, where entire classes are to be forgotten. Class unlearning poses a complex problem because removing all knowledge related to one or several classes must not degrade performance on other, potentially similar classes. This requires disentangling shared representations in latent space and mitigating unwanted interference effects between overlapping features.

The motivation for machine unlearning extends well beyond algorithmic challenge alone. In practice, unlearning addresses ethical, legal, and operational challenges associated with modern data-driven ML systems. Real-world datasets often contain sensitive, erroneous, or outdated information, including personal or private data. Regulations such as the General Data Protection Regulation (GDPR) in the European Union explicitly establish a “right to be forgotten”, which obliges data controllers to delete personal data upon user request [211]. However, while deletion of raw data from storage systems is straightforward and provable, removing the implicit representations of that data from large ML models remains an unresolved technical challenge. For instance, models may retain latent associations or statistical patterns that indirectly encode information about deleted records, posing both privacy and compliance risks. From a systems perspective, simply retraining a model from scratch on a modified dataset would yield a clean solution at a prohibitive cost. Retraining large models can be computationally expensive, time-consuming, and energy-intensive, particularly in domains such as modern large language models where trainings make use of billions of samples and often require computational power of large supercomputers. Consequently, practical unlearning methods seek to provide efficient alternatives to full retraining by modifying existing model parameters in a targeted way that approximates the desired forgetting behavior.

An effective machine unlearning approach should ideally satisfy the following properties:

- **Model reusability:** It should operate directly on pretrained models, without requiring access to the full original training data.
- **Effective forgetting:** It should remove or neutralize the influence of the forget set.
- **Computational efficiency:** It should be significantly less resource-intensive than retraining.

Ref. [212] gives a review of benchmarking methods to test for these goals.

Machine unlearning thus represents a critical step towards responsible, adaptive, and regulation-compliant AI systems. It provides a technical foundation for combining data-driven innovation with individual data rights and societal expectations of transparency and reversibility. Within this work, the focus is placed on class unlearning particularly, contributing to solving the challenges of unlearning selected classes.

9.1.1 *Related Work*

Unlearning for machine learning models has been studied for almost a decade [209] in different flavors. Unlearning methods can be categorized into exact unlearning and approximate unlearning [210, 213]. Exact unlearning reaches the distributions of a retrained model that had no information about the forget set in an analytically provable way. Unfortunately, exact unlearning (besides retraining) can rarely be applied in real-world scenarios because modern models are too complex [214]. Therefore, one often has to deal with approximate unlearning, where the model output is evaluated without reaching the native retraining performance.

Another way to categorize methods is based on the goal

- Class unlearning, concerned with unlearning the information of one or many class assignments [215, 216].
- Instance unlearning, concerned with unlearning the information of selected, single data points [217].
- Feature unlearning, concerned with unlearning semantic features in the training data [208].

For all of these goals, methods can also be categorized into data reorganization methods and model manipulation methods, depending on whether they operate at the data level or directly on the trained model parameters.

Data reorganization methods modify the training data or its influence on the model to induce forgetting indirectly. This class includes techniques such as data obfuscation, data pruning, and data replacement. In data obfuscation [218], the idea is to transform or corrupt specific samples in a way that renders them uninformative or misleading for the model. The modified data are then used to retrain or finetune the model, effectively erasing the unwanted information while preserving general model performance. Data pruning [219], by contrast, involves removing data points or subsets of data that are strongly associated with the forget set, potentially followed by a short retraining phase to restore performance on the retain set. Finally, data replacement [209] substitutes samples from the forget set with alternative examples that preserve the statistical properties of the retain set, thereby counteracting the model's prior exposure to the unwanted data. Data-centric methods have the advantage of conceptual simplicity and interpretability but usually require access to the original training data, which may not always be feasible due to privacy, storage, or licensing constraints. Another concern can be the computational overhead of additional training steps and it is often hard to derive guarantees for successful unlearning. However, these methods are generally feasible to use even with large and unstructured ML models such as DNNs and do not rely on adding model features at the time of training.

Model manipulation methods, on the other hand, operate directly on the model parameters or internal representations without revisiting the training data [220–222]. These methods aim to selectively adjust, replace, or prune parts of the network in order to remove the influence of the forget set. For instance, parameter shifting adjusts specific weights or biases in the model to suppress features associated with forgotten information. Parameter replacement involves reinitializing parts of the network or substituting learned representations with neutral alternatives. Parameter pruning physically removes neurons, filters, or submodules responsible for encoding features linked to the forget set. Model-centric methods are generally more computationally efficient, as they avoid any retraining. However, they face challenges in ensuring complete forgetting, since information can be redundantly encoded across multiple layers and representations and identifying model parameters linked to information is a highly non trivial task.

The approach in this work falls under class unlearning via data obfuscation. Specifically, this strategy manipulates labels associated with entire classes to induce forgetting at the model level.

9.2 APPROPRIATING THE PULL BY PARTIAL LABELS APPROACH FOR UNLEARNING

Equation 9.2 suggests a parallel between negative weights in training and unlearning. Indeed, a common unlearning method is gradient ascent [223]. It suffers from the issues discussed in chapter 8 but can outperform methods like finetuning in tasks where the forget set is small enough for these problems to not occur. Stabilizing the gradient

unlearned class	Gradient Ascent		PPL		Finetune	
	Retain Acc.	Forget Acc.	Retain Acc.	Forget Acc.	Retain Acc.	Forget Acc.
0	0.63	0.00	0.88	0.00	0.88	0.21
1	0.78	0.00	0.88	0.00	0.88	0.42
2	0.67	0.00	0.89	0.00	0.89	0.14
3	0.70	0.01	0.90	0.00	0.90	0.02
4	0.68	0.00	0.88	0.00	0.88	0.18
5	0.77	0.00	0.90	0.00	0.90	0.05
6	0.69	0.00	0.88	0.00	0.88	0.42
7	0.69	0.00	0.88	0.00	0.88	0.26
8	0.72	0.00	0.87	0.00	0.87	0.54
9	0.75	0.00	0.87	0.00	0.88	0.36

TABLE 9.1: Retain and Forget Accuracy for unlearning CIFAR10 classes with the PPL method, finetuning and gradient ascent. The full training set was used for finetuning. To gain comparability in the PPL and gradient ascent method also the full retain set was used. The highest retain accuracy per class is highlighted.

unlearned class	Gradient Ascent		PPL	
	Retain Acc.	Forget Acc.	Retain Acc.	Forget Acc.
0	0.29	0.00	0.78	0.00
1	0.23	0.06	0.75	0.01
2	0.29	0.00	0.80	0.01
3	0.28	0.00	0.82	0.00
4	0.35	0.00	0.76	0.01
5	0.21	0.00	0.77	0.00
6	0.23	0.00	0.77	0.01
7	0.25	0.00	0.78	0.01
8	0.22	0.07	0.75	0.00
9	0.21	0.00	0.76	0.01

TABLE 9.2: Retain and Forget Accuracy for unlearning CIFAR10 classes with the PPL method and gradient ascent. Both unlearning methods were restricted to the forget set. The highest retain accuracy per class is highlighted.

ascent can generalize this performance improvement and could provide a further tool at the hand of unlearners. Motivated by this, PPL (as described in section 8.3.3) was taken and applied in a finetuning scenario where the pretrained network was trained for some further optimization steps with the forget set relabelled. During these steps, some data from the retain set can be utilized to stabilize the retain accuracy.

9.2.1 Comparing PPL Unlearning Performance on the CIFAR datasets

In order to evaluate the performance and stability of different machine unlearning strategies, three distinct approaches are compared: the PPL method, gradient ascent unlearning, and conventional finetuning on the retain set. The experiments are performed on the CIFAR10 [224] and CIFAR100 datasets using a pretrained ResNet18 architecture [82]. CIFAR10 is a standard benchmarking color image dataset with labeled objects in 10 classes such as "boat", "horse" or "car" with 50,000 training data samples and 10,000 testing sample. CIFAR100 is a dataset of the same size but it contains 100 classes. It thus poses a bigger challenge for classification algorithms.

For the CIFAR10 experiments, the network was trained for 30 epochs, achieving a baseline test accuracy of 80.3%. The CIFAR100 model was trained for 100 epochs and reached a baseline accuracy of 48.6%. Both models exhibited stable convergence of the training and validation loss curves, indicating that the pretrained weights were sufficiently optimized for evaluating unlearning dynamics. Hence, further pretraining optimization was deemed unnecessary. The unlearning procedures were applied for 10 epochs using a reduced learning rate of 1×10^{-4} , a value empirically chosen to ensure gradual convergence. Gradient ascent was implemented by inverting the sample weights ($w = -1$) for the forget set samples.

The results of the experiments are summarized in Tables 9.1 and 9.2. The first table reports results when both the retain and forget sets were used during unlearning (referred to as stabilized unlearning), while the second presents the scenario where only forget samples were used.

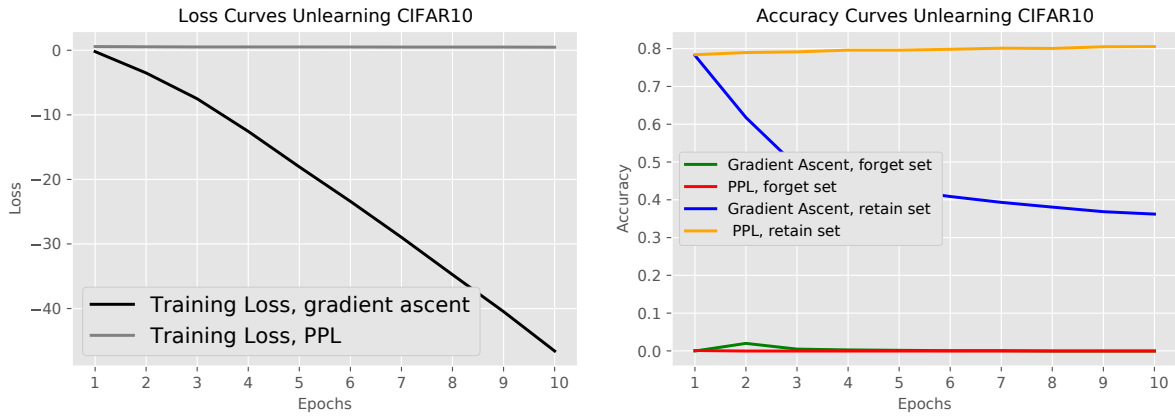


FIGURE 9.1: Unlearning performance comparison between gradient ascent and PPL on the CIFAR10 dataset. The unlearning performance on the class 7: 'horse' is shown with retain set data in the ratio 1:5. The left plot shows the loss across the unlearning epochs while the right plot shows the accuracy on the retain and forget test sets.

The comparison to finetuning can only fairly be drawn with the retain set, as finetuning is just a further training on the retain set. One can observe across all classes a low degree of forget accuracy and a high retain accuracy. The experiment was also carried out on the forget set only, where it was found that the accuracy on the retain set suffered a drop of a few percent but no catastrophic forgetting was observed. The forget accuracy was similar to the case with stabilizing support of retain samples. Using the retain set during unlearning achieves comparable retain accuracies as finetuning, while offering the superior forgetting performance. This proves the stability of the method.

To provide a detailed understanding of the unlearning dynamics, the temporal evolution of loss and accuracy during the unlearning of CIFAR10 class 7 (horse) is examined. This particular class was selected because its baseline misclassification rate lies closest to the dataset-wide average, making it a representative case for assessing the typical difficulty of class unlearning. Consequently, the observed behavior can be seen as the average unlearning complexity among the CIFAR10 classes.

The unlearning process was conducted over 10 epochs with a learning rate of 1×10^{-10} , using a 1:5 ratio between retain and forget samples. Higher learning rates can compress the unlearning into one epoch. The chosen learning rate allows for a clearer observation of the trend behavior in both loss and accuracy, as illustrated in Figure 9.1.

The loss curve reveals the instability inherent to the gradient ascent method consistent with the earlier findings on the MNIST dataset. Both the PPL and gradient ascent approaches successfully reduce the forget accuracy to approximately zero, confirming that both achieve the desired erasure of information associated with the forgotten class. However, the two methods differ significantly in their impact on the retain set: the PPL method maintains a nearly constant retain accuracy throughout the unlearning epochs, whereas gradient ascent converges to a lower equilibrium. After approximately five epochs, the competing influence of the forget and retain gradients stabilizes, and the retain accuracy for gradient ascent converges to about 37%, compared to 80% for PPL. This substantial difference highlights the robustness of PPL against gradient instability and catastrophic forgetting.

For the sake of comparison with Shah et al. [215] the forget accuracy against the retain accuracy of a ResNet50 [82] for forgetting class 1 in the right plot in figure 9.2 is shown. They report for 100% unlearning of the forget set a maximum performance degradation of 0.07% on the retain set. In this case, due to the stabilization with 1:5 retain to forget samples, one can even observe an increase in retain accuracy until 4% forget set accuracy.

The experiment is repeated on the more challenging CIFAR100 dataset, once again with the forget class chosen to represent the average number of misclassifications. The right panel of figure 9.2 shows a similar retain vs forget accuracy performance after each batch of unlearning. Once again, the final evaluation point can be obtained with a higher learning rate after only one epoch but the results are deliberately shown with a learning rate of 1×10^{-7} to visualize intermediate unlearn steps.

The results show the stability of the PPL method. While the retain performance suffers for the gradient ascent due to the instability of the loss, it is evident that PPL keeps the retain accuracy at any step within 2% of the nominal performance.

9.2.2 Membership Inference Analysis

A low forget set accuracy does not tell the full story for any unlearning method. Another important privacy concern is the ability to infer whether a training point was part of the training distribution.

As a consistency check one thus has to observe how the model reacts to unlearned samples. A good consistency check is given by the migration of the unlearned classes in figure 9.3b. The plot shows the classes that the network predicted for the members of class "horse" after unlearning. While PPL follows a distribution across multiple classes, one can see that gradient ascent only classifies the unlearned class members into one other class, failing the criteria for reliable unlearning.

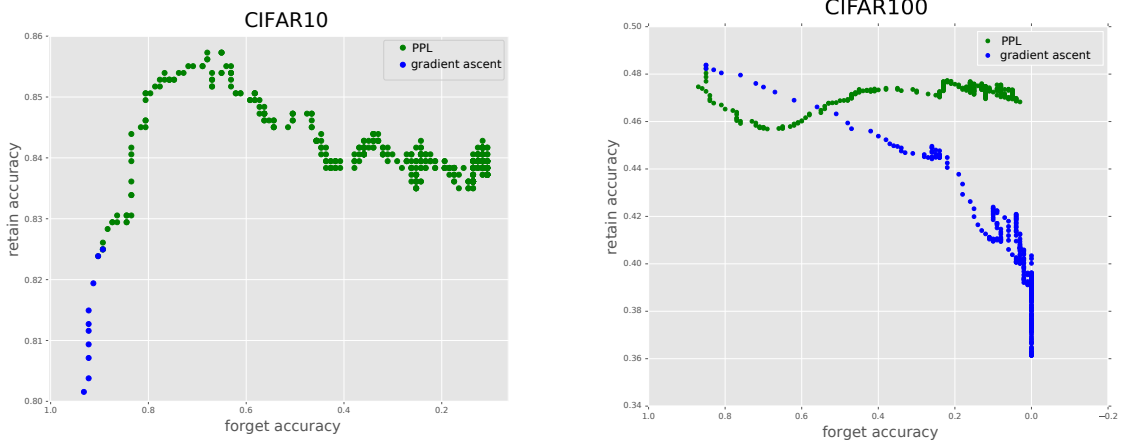


FIGURE 9.2: The left plot shows the forget to retain accuracy evaluated after every batch of batched unlearning with a retain to forget set ratio of 1:5 for the comparison experiment to Shah et al. [215]. This showcases the unlearning behavior of PPL on a pretrained ResNet50 on the CIFAR10 dataset, unlearning the class 1. The gradient ascent results are unstable and mostly contained outside the plotting range, the shown points do not show a trajectory but later training batches. The right plot shows the same experiment for the CIFAR100 experiment.

Moreover, the resulting distribution resembles the feature-based similarities between the different classes as the unlearned horses are more similar to other animals, especially deer and dogs, than to means of transportation. This suggests that the PLL unlearned network retains a notion of feature closeness on the forget class and classifies these accordingly. This is expected behavior when one compares to the shape of the native retraining on the retain set. The behavior of the gradient ascent unlearning suggests a form of misclassification overfitting that is also evident in the diverging loss curve. These findings are consistent with the MNIST Grad-CAM experiment from the previous section where PPL showed feature retention as well.

Another handle to check for training set membership information on unlearned classes is the loss. If the loss on a set of samples is distinctly shaped after unlearning and allows to classify these data samples with respect to their membership in the training set, the unlearning technique has to be considered flawed. This can be studied in membership inference attacks (MIA) [225]. A basic MIA is implemented with a linear regression model on the loss of samples of the forget set vs. the losses of a retrained model without the forget set.

Figure 9.3a shows the loss (here computed with categorical crossentropy) distributions and MIA results for different unlearning methods on the already discussed case of unlearning class 7 of CIFAR10. The finetuning, PPL and gradient ascent method are compared. Also the trivial method of marginalizing an output class is added. This is accomplished by setting its bias in the output layer to a large negative value. When setting the bias of the forget class to a large negative value, one can naively easily obtain 0% forget accuracy at the full model accuracy on the retain set. However, this does not delete the class information out of the model and one can see a very similar loss behavior as the gradient ascent. A MIA can infer training set membership in this scenario with a high accuracy of above 98%, effectively rendering this method useless.

PPL on the other hand shows a loss distribution that is much more similar to the loss of a model that never saw the forget class. The migration analysis from above gives a reason for this behavior, showing that feature similarity moves these samples to believable wrong classes. This reflects in the accuracy of the MIA, where PPL scores a much lower 66% whereas the other methods all fall above 90%.

9.2.3 Limitations of the Method

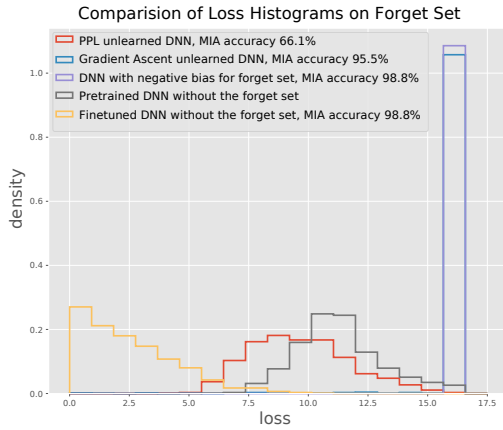
PPL has performed well in the empirical evaluations done in this chapter. They act as a proof of concept and show consistent results. For industry-adoption, studies on larger and more complex problems are needed. While the translation of results from CIFAR10 to CIFAR100 hints at scalability of the approach, modern image classification algorithms such as vision transformers [226] can deal with much larger applications.

Another important aspect in the context of regulations are unlearning guarantees. An analytical guarantee can not be established in this work because the method relies on stochastic convergence of (arbitrarily) complex models. Proper evaluation on the application at hand, as conducted here, can prove the stability of PPL but will always introduce an overhead in using the method.

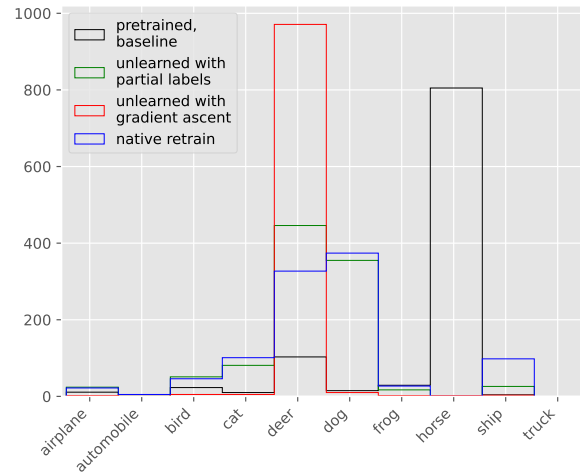
9.3 DISCUSSION OF UNLEARNING VS. NEGATIVE WEIGHTS IN PHYSICS

While sharing mathematical similarity, as shown in equation 9.2, as well as successfully being solved by PPL, the problem of unlearning and negative event weights in particle physics are distinct challenges.

Negative weights in particle physics have two distinct goals.



(a) Loss distribution of different unlearning methods on CI-FAR10 class 7. The legend also states the MIA accuracy values for the different methods obtained according to the attack described in the text.



(b) Migration analysis of the CIFAR10 "horse" unlearning task. The baseline shows the pretrained model without any unlearning applied. It is compared with gradient ascent and PPL, as well as a native retraining without the horse class.

- Negative weights coming from MC subtraction schemes are an algorithmic artifact. One wants to minimize their impact and their statistical contributions in a full dataset averages out.
- Negative weights coming from an interference are part of a signal feature. They contain information for a training that ideally contributes to a classifiers performance.

In unlearning, negative weights are an expression of the target. Their impact is the information erasure of a full class.

A major difference is the time at which they are applied. In HEP the negative weights are present during the initial training and their influence is exerted at the same time that the opposite impact of positively weighted events is applied. In unlearning however, the classifier is converged and the negatively weighted samples will (apart from stabilizing contributions) act on their own.

Another difference is the ratio of negatively weighted events. Unlearning has a full designated class that is negatively weighted. In HEP a part of some classes can be negative. This contribution can further even out as in some input feature the negative impact will be bundled up in an area of phase space and in other input features they are evenly spread. The contribution of the negative events carries meaningful information and PPL has proven that it can learn particular features from these events. This is most evident in the study of negative features in the double slit toy model (figure 8.10).

In summary, PPL demonstrates a remarkable instance of cross-domain applicability. The method can stabilize training in the presence of negative sample weights and improve performance by correctly treating negative event weights as a signal feature. It was also found to be a robust and effective solution for the stable erasure of information from a trained model. The unlearning can, based on this reasoning, be considered an edge case of negative contributions as a feature. While the feature in HEP is "less signal" the ultimate goal of unlearning is "no signal" or put in unlearning words: "no contribution of this class".

SUMMARY

This thesis presents a multi-faceted advancement in top quark physics, spanning from high-precision theoretical implementations to the development of novel machine learning frameworks for Beyond the Standard Model (BSM) searches.

First, the precedent for future inclusion of NNLO-accuracy into CMS analyses was set by implementing the `ttJ_MiNNLO` MC generator into CMS software. A physical validation was performed and while questions (e.g. on the modelling of the top mass m_t) remain open at this point, a good agreement with the originally published results was found. This MC implementation of NNLO $t\bar{t}$ production was in particular observed to solve a long standing mismodelling of the p_T spectrum of the top quark in previous NLO-only simulations.

Building on the work of a CMS search for heavy (pseudo)scalar particles with subsequent dileptonic and semileptonic $t\bar{t}$ decays, areas of future research were explored and improvements identified. The search uncovered a significant excess at the $t\bar{t}$ production threshold that opens up large opportunities for future study.

The search originally suffered the problem of limit setting in multidimensional theory parameter spaces in scenarios where Wilks theorem is violated. To solve this, a new approximate method was explored that allows to keep the advantages of toy based inference while scaling down the computational cost that one would need to pay with, for example, the Feldman-Cousins method. The method was found to be accurate and reliable in cases where nuisance parameters are not significantly pulled. With larger pulls, it can only allow for approximate results.

Future searches, especially in the light of the excess at the $t\bar{t}$ threshold, will need to consider scenarios that might not preserve CP symmetry. Discovery of CP violation would undoubtedly connect the excess to exotic mixing heavy Higgs bosons. Groundlaying work on searches for CP violating heavy scalar particles was carried out building on the framework of simulation based inference. The very high dimensional theory parameter space of CP violating models necessitates the usage of optimal DNN based observables. With cancellations from signal-signal interference, low dimensional histogram based analyses can suffer from collapsing signatures in the chosen observables. It was shown that simulation based inference can resolve ambiguities. Also, the information content of observables was studied and sensitivity predictions for future analyses with LHC run 3 or high-luminosity LHC data were given. Another important factor for the study of the excess is the $m_{t\bar{t}}$ resolution. It is largely influenced by the choice of $t\bar{t}$ reconstruction algorithm. This work suggests the usage of an advanced transformer architecture building on the lorentz algebra - LGATR. A first performance study uncovered potential for gains in reconstruction resolution.

On top of future studies, already with the present analysis results one can deduct interesting insights. The excess is pseudoscalar in nature and can thus fit a hypothetical pseudoscalar axion like particle which are of particular interest as they might play a role in solving multiple outstanding problems of the SM such as the missing CP violation of the QCD and the existence of dark matter. To study this, a full reinterpretation of the CMS search was conducted. Templates were derived with a post-mortem reweighting method and the background model was reused to derive exclusion limits for an ALP with a coupling c_t to top quarks as well as a direct gluon coupling c_G . The results are largely in agreement with a phenomenological sensitivity study. No significant excess over the background model was discovered if $t\bar{t}$ bound states were included. It was validated that the top bound state hypothesis covers the excess also in scenarios with a hypothetical ALP. Subsequently, exclusion limits on c_t were set for ALPs between 400 and 800 GeV. For a vanishing ALP gluon coupling, these exclusions are the most stringent to date. Additionally, for the first time exclusion limits on c_t were set in scenarios where the gluon ALP coupling is non-vanishing.

All the above results rely on physical observables from which they derive their sensitivity to new phenomena to be uncovered. A promising future avenue is the usage of ML based classifiers for separation of signal from background. This was previously impossible due to large parts of the simulation being negatively weighted. Negative weights can arise from technical reasons in MC generator software, namely NLO subtraction schemes. A second source are interference effects. These are particularly important in searches for heavy (pseudo)scalar particles in top quark pair production since the signal and background interfere. Training ML classifiers on MC datasets which include negative weights leads to instable behavior. This effect was studied in detail in this work. A solution was proposed in the form of the Pull by Partial Labels (PPL) method. Its effectiveness and feature-correct behavior was validated in multiple experiments with the MNIST dataset, in a scenario derived from the doubleslit experiment and in a mock analysis building on the search for heavy (pseudo)scalars in dileptonic $t\bar{t}$ decays.

Lastly, the method was shown to provide strong cross domain applicability to the field of class unlearning. In multiple experiments the method shows competitive performance in unlearning classes. Additionally, the resulting unlearned network shows desirable behavior when questioned about information from the unlearned classes in a membership inference attack.

10.0.1 *Outlook*

With the discovery of an significant excess at the $t\bar{t}$ threshold the field of top quark physics has opened up a new field of study. Understanding this excess will require many avenues of exploration, from the usage of advanced new ML algorithms that can deal with negative weights to new top reconstruction techniques and exploration of new, more complex theory models.

This thesis has layed out many possibilities that can shed a light on the physical nature of the excess. The run 3 and High-Luminosity era of the LHC will have more sensitivity, employ better $t\bar{t}$ reconstruction methods and explore new models. More accurate theory predictions will go hand in hand with better analyses techniques to answer the question if there are, as suspected, top quark bound states, if there are further contributions from new physics or if the signature is a new particle entirely. This thesis has contributed to many methods in bettering their resolution, advancing their scope or minimizing their computational burden. Beyond the LHC era, in a possible FCC-ee or linear collider, analyses will also benefit from these findings.

Additionally, this thesis has also shown the possibility of using partial labels for shaping the output of ML classifiers. This method thus has the possibility to have a broader range of applications in scenarios where one wants to interfere with the models trained predictions, not only in this analysis of effects at the $t\bar{t}$ productions threshold, but for applications in many analyses of particle physics data.

APPENDIX

A.1 IMPLEMENTATION OF ttj_minnlo INTO CMS SOFTWARE

Input Parameters to ttJ_MiNNLO

```

#-----
# INCOMING BEAMS PARAMETER
#-----
ih1      1  ! hadron 1 (1 for protons, -1 for antiprotons)
ih2      1  ! hadron 2 (1 for protons, -1 for antiprotons)
ebeam1   6500d0 ! energy of beam 1
ebeam2   6500d0 ! energy of beam 2

#-----
# PDF & ALPHAS SETTINGS
#-----
# To be set only if using LHA pdfs
#
lhans1    306000 ! pdf set for hadron 1 (LHA numbering)
lhans2    306000 ! pdf set for hadron 2 (LHA numbering)

# ALPHAS AND PDF EVOLUTION
alphas_from_pdf 1 ! (default 0) if 1, uses alphas from PDF evolution tool (e.g. lhafdf or hoppet)
                  ! This is required for running MiNNLO!

#-----
# RENORMALIZATION/FACTORIZATION SCALE SETTINGS
#-----
# For MiN(N)LO these scale settings are relevant only to set the scales of the overall two powers of alphas.
# The values of the other scales in that case are set according to the MiNNLO procedure (and according to largeptscales).
#
# (Note that for NLO+PS ttJ runs these scale settings set the values of all scales.)
#
fixedscale 0 ! (default 1), if 0 use dynamical scale for overall two powers of alphas (set by whichscale),
#           ! if 1 scale for overall two powers of alphas is fixed to the top-quark mass
whichscale 4 ! (default 0), dynamical scale of overall two powers of alphas, only used if fixedscale=0
#           ! 0 - H.T^tt/2 (sum of transverse masses of the top quarks over 2)
#           ! 1 - m_tt/2 (invariant mass over 2)
#           ! 2 - m_tt (invariant mass)
#           ! 3 - H.T^tt+jets/2 (sum of transverse masses of all final states over 2)
#           ! 4 - H.T^tt/4 (sum of transverse masses of the top quarks over 4)
#           ! 5 - H.T^tt+jets/4 (sum of transverse masses of all final states over 4)

#
# This is used only for MiN(N)LO to set the scale at large pT (i.e. in the fixed-order part)
#
largeptscales 0 ! (default 0) dynamical scale used at large pT in fixed-order part
#           ! 0 - m_tt (invariant mass of top-quark pair)
#           ! 1 - pT_tt (transverse momentum of top-quark pair)
#           ! 2 - H.T^tt (sum of transverse masses of the top quarks)

#-----
# MiNLO'/MiNNLO_PS SETTINGS
#-----
minlo      1 ! (default 1) if 1, activate MiNLO'
minnlo     1 ! (default 0) if 1, activate MiNNLO_PS
modlog_p   -1d0 ! (default -1d0; only works when minlo and/or minnlo is set to 1)
#           ! if >0d0, activate modified logs and set the exponent
#           ! if -1d0, activate piecewise modified log version 1
#           ! if -2d0, activate piecewise modified log version 2
#           ! piecewise modified log: log(Q/pT) for pT<Q/2
#           ! zero for pT>Q
#           ! smoothly interpolated inbetween
#           ! if 0d0, -log(Q/pT) * theta(Q-pT) (not recommended)
kappaQ     0.5d0 ! (default 1.) factor of scale Q inside the (modified) logarithm:
#           ! Q = Qref * kappaQ (Qref=invariant mass of the system)
#           ! Set =< 0.5d0 for ttbar !
Q0         2d0 ! (default 2.) cutoff of profiled scales

#-----
# SCALE VARIATION & REWEIGHTING
#-----
#
renscfact   1d0 ! (default 1d0) ren scale factor: muren = muref * rensfact
facscfact   1d0 ! (default 1d0) fac scale factor: mufact = muref * facscfact
storeinfo_rwgt 1 ! store info to allow for reweighting
flg_debug   1 ! store extra event info for debugging

rwl_file '-'
lhrwgt_id 'st4'
lhrwgt_descr 'st4 NNLOPS'

<initrwgt>
<weightgroup name='scale_variation' combine='envelope'>
<weight id='1001'> rensfact=1.0 facscfact=1.0 </weight>

```

```

<weight id='1002'> rensfact=1.0 facscfact=2.0 </weight>
<weight id='1003'> rensfact=2.0 facscfact=1.0 </weight>
<weight id='1004'> rensfact=2.0 facscfact=2.0 </weight>
<weight id='1005'> rensfact=1.0 facscfact=0.5 </weight>
<weight id='1006'> rensfact=0.5 facscfact=1.0 </weight>
<weight id='1007'> rensfact=0.5 facscfact=0.5 </weight>
<weight id='1008'> rensfact=0.5 facscfact=2.0 </weight>
<weight id='1009'> rensfact=2.0 facscfact=0.5 </weight>
<weight id='1010'> rensfact=2.0 facscfact=0.5 </weight>
<weight id='1011'> lhpdf=306000 rensfact=1.0 facscfact=1.0 </weight>
<weight id='1012'> lhpdf=303600 rensfact=1.0 facscfact=1.0 </weight>
</weightgroup>
</initrwtg>

#-----
# PROCESS PARAMETERS
#-----
topmass 172.5d0 ! value of top-quark mass
topwidth 1.31d0 ! value of top-quark width (non-zero value for topdecaymode/=0, even for zerowidth=1)
#topwidth 1.285d0 ! value of top-quark width (non-zero value for topdecaymode/=0, even for zerowidth=1)

# for the EW couplings/parameters (entering only in the decay matrix elements) the (real) Gmu scheme is used
gfermi 1.16637d-5 ! Fermi constant used as input for Gmu scheme
Wmass 80.4d0 ! W-boson mass (input for Gmu scheme and relevant for top decays)
Wwidth 2.141d0 ! W-boson width (input for Gmu scheme and relevant for top decays)
Zmass 91.1876d0 ! Z-boson mass (input for Gmu scheme)
Zwidth 2.4952d0 ! Z-boson width (input for Gmu scheme)

# the top-quark decays are included at tree-level with spin correlations using the following setting
topdecaymode 0 ! decay mode of the top quarks
# ! 0: stable top quarks (no decay)
# ! set maximum number of particles from top decays e,mu,tau,u,c; for instance:
# ! 22222: all top decays allowed (including tau)
# ! 22022: all top massless decays allowed (no tau)
# ! 22000: fully leptonic massless (no tau)
# ! 22200: fully leptonic (with tau)
# ! 00022: hadronic (with tau)
# ! note: only settings allowed that sum up to at least 2 (unless 0 for stable tops)
semileptonic 0 ! if 1 keep only semi-leptonic top decays, ie. one to leptons and one to hadrons
zerowidth 0 ! if 0 include off-shell effects for top quarks (and W bosons) approximately
# ! if 1 keep top quarks (and W bosons) in the decay matrix elements on-shell

# all leptons and light quarks are treated as massless at the level of the matrix elements; however, at the
# level of the event generation the momenta of the charged leptons, charm and bottom quarks are reshuffled
# to their masses specified below (the given values are the default values used by the code)
lhfm/emass 0.511d-3 ! (default 0.511d-3) electron mass for reshuffling momentum in event generation
lhfm/mumass 0.1057d0 ! (default 0.1056d0) muon mass for reshuffling momentum in event generation
lhfm/taumass 1.777d0 ! (default 1.777d0) tau mass for reshuffling momentum in event generation
lhfm/cmss 1.5d0 ! (default 1.40d0) charm-quark mass for reshuffling momentum in event generation
lhfm/bmass 4.8d0 ! (default 4.92d0) bottom-quark mass for reshuffling momentum in event generation

# OpenLoops settings
use-OLP-Interface 1 ! (default 1) use OpenLoops interface (don't turn this off when using OpenLoops!)

#-----
# Generation cuts
#-----
bornktmin 0.6d0 ! Generation cut; minimum kt in underlying born

#-----
# POWHEG PARAMETERS: Monte Carlo Integration
#-----
# Parameters to allow or not the use of stored data
use-old-grid 1 ! if 1 use old grid if file pwggrids.dat is present (<= 1 regenerate)

# General MC integration settings
ncall1 50000 ! number of calls for initializing the integration grid
itmx1 1 ! number of iterations for initializing the integration grid
ncall2 100000 ! number of calls for computing the integral and finding upper bound
itmx2 1 ! number of iterations for computing the integral and finding upper bound

# Folding used in the integration of the real contributions
foldcsi 5 ! number of folds on csi integration
foldy 5 ! number of folds on y integration
foldphi 5 ! number of folds on phi integration
testplots 0 ! (default 0, do not) do NLO and PWHG distributions

#-----
# POWHEG PARAMETERS: Upper bound evaluation
#-----
use-old-ubound 1 ! if 1 use norm of upper bounding function stored in pwgubound.dat, if present; <= 1 regenerate
nubound 10000 ! number of bbarra calls to setup norm of upper bounding function
xupbound 2d0 ! increase upper bound for radiation generation
icsimax 3 ! <= 100, number of csi subdivision when computing the upper bounds
iymax 3 ! <= 100, number of y subdivision when computing the upper bounds
storemintupb 1 ! (default 0) if 1 improve upper bounding computation
ubexcess.correct 1 ! corrects for error in upper bound estimate

#-----
# POWHEG PARAMETERS: Event generation
#-----
numevts 10000 ! number of events to be generated

# To correct for UB violations detected during stage 4

```

```

# (see file POWHEG-BOX-V2/Docs/BoundViolationFixes.pdf)
#
# in essence, before analysing the .lhe files/running Pythia, run:
#
# POWHEG-BOX-RES/Scripts/FindReweightFromCounters.py pwgcounters-st4-*.dat
#
# uncomment lines below, and set corr_btilde and corr_remnant to the provided values
#
#btildeviol 1 ! if 1, turn on correction of UB violations
#corr_btilde 1.01624045621 ! correction factor for btilde
#corr_remnant 1.0 ! correction factor for remnant

#-----
# PARTON SHOWER PARAMETERS
#-----
# Settings for shower with pythia8
py8tune 14 ! (default 21 -> A14 tune) tune for pythia8 (14 -> Monash2013 tune)
nohad 1 ! (default 1) if 1 switch off hadronization
nompi 1 ! (default 1) if 1 switch off multi-parton interactions (should only be 0, if also nohad 0)
noqed 1 ! (default 1) if 1 switch off QED shower effects

#-----
# Soft/Collinear limit TESTS
#-----
# Check soft/coll limits -> Just for debugging
colltest 1 ! (default 1), if 0 turn off collinear test
softtest 1 ! (default 1), if 0 turn off collinear test

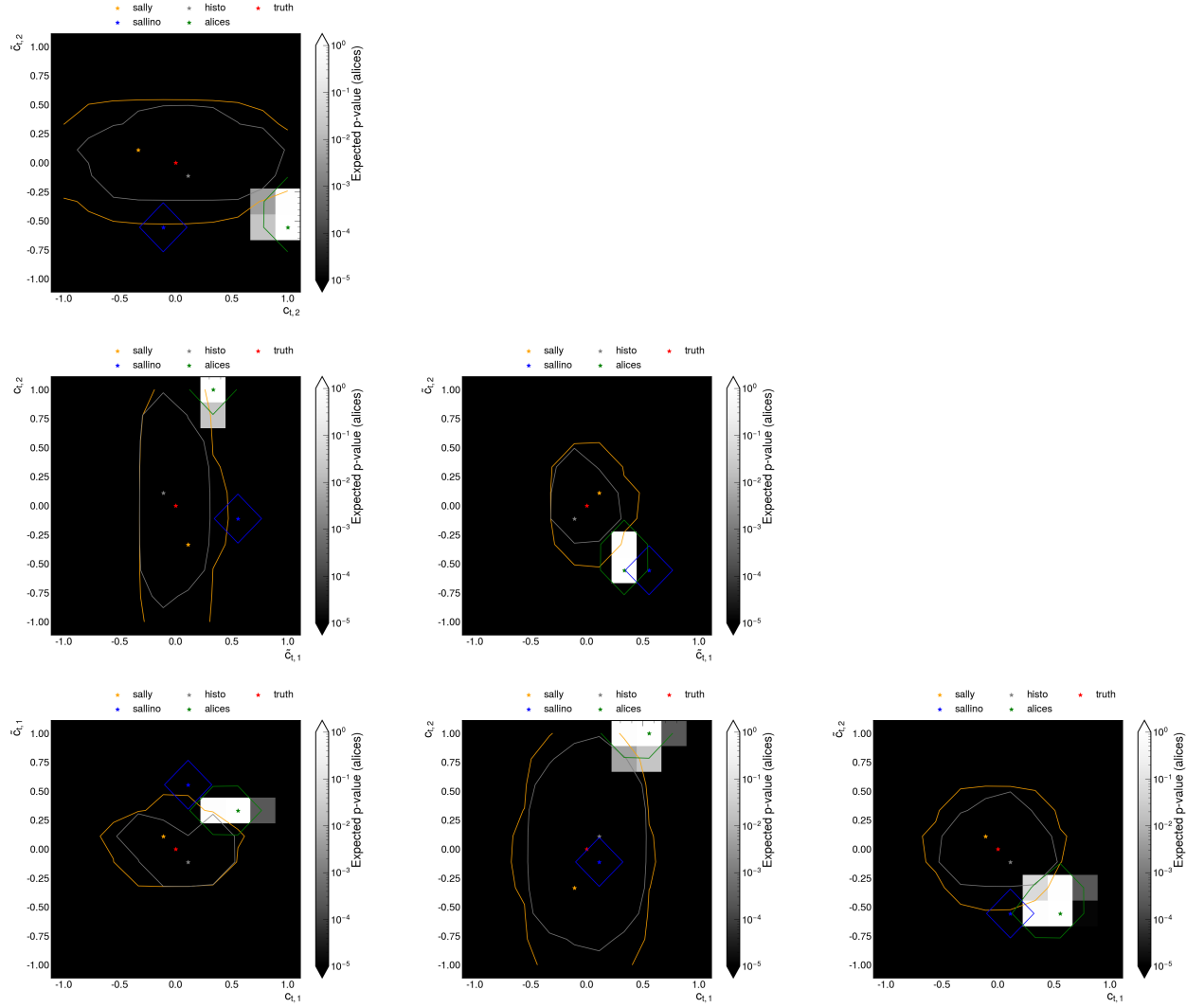
#-----
# OPTIMIZATION OPTIONS
#-----
smartsig 1 ! (default 1)
fastbtlbound 0 ! (default 0), if 1 speed up computation of upper bound of btilde

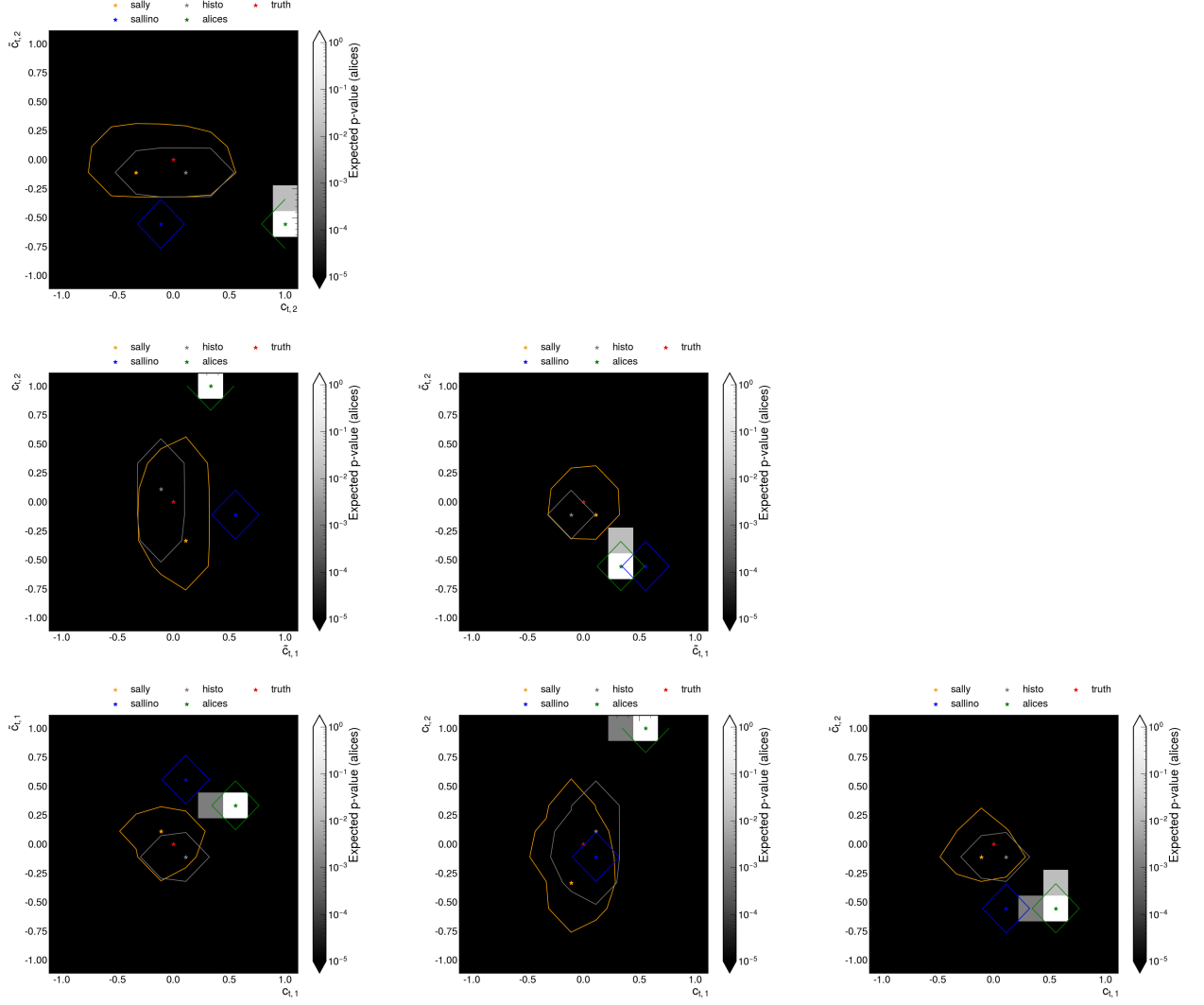
#-----
# OPTIONS TO CATCH INSTABILITIES
#-----
check_bad_st1 1 ! (default 0), if 1 takes care of removing bad grids during stage 1
check_bad_st2 1 ! (default 0), if 1 takes care of removing bad grids during stage 2

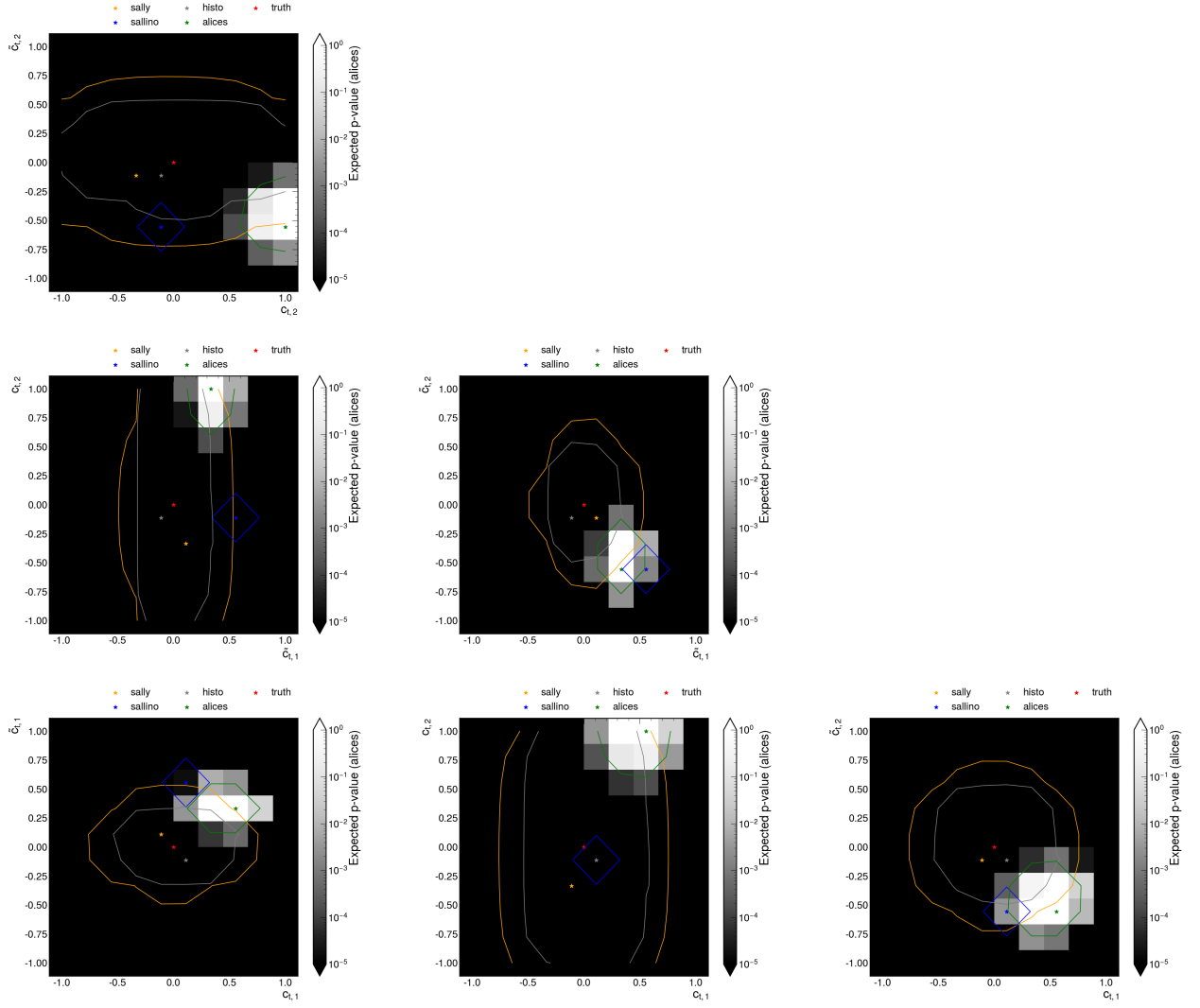
#-----
# PARALLEL RUNS
#-----
# The following is for parallel runs (see manual in Z2jet/Docs/)
manyseeds 1
parallelstage 1
xgriditeration 1
maxseeds 400

```

A.2 ADDITIONAL LIMIT RESULTS FOR CP VIOLATING HEAVY HIGGS SCENARIOS WITH MADMINERS MORPHING

FIGURE A.1: Limit projections for 300 fb⁻¹. The notations are similar as in figure 6.11.

FIGURE A.2: Limit projections for 3000 fb⁻¹. The notations are similar as in figure 6.11.

FIGURE A.3: Limit projections for 129 fb^{-1} . The notations are similar as in figure 6.11.

SALLY Fisher Information Matrix for Mixed Benchmark

The full Sally Fisher information matrix at the mixed benchmark $c_{t,1} = \tilde{c}_{t,2} = c_{t,2} = \tilde{c}_{t,1} = 0.5$ is

$$\mathcal{I}_{ij}(c_{t,1} = \tilde{c}_{t,2} = c_{t,2} = \tilde{c}_{t,1} = 0.5) = \begin{pmatrix} 0.5250 & 0.0470 & 0.0382 & 0.1046 \\ 0.0470 & 1.0000 & 0.1515 & 0.1195 \\ 0.0382 & 0.1515 & 0.5143 & -0.0347 \\ 0.1046 & 0.1195 & -0.0347 & 0.6734 \end{pmatrix}. \quad (\text{A.1})$$

For easier interpretation, these values have been normalized to the maximum. The first row relates to the $c_{t,1}$ parameter, the second row to $\tilde{c}_{t,1}$ and vice versa for the second scalar particle. The eigenvalue decomposition of equation A.1 is

$$\lambda_1 = 1.0837, \quad v_1 = \begin{pmatrix} 0.1467 \\ 0.9173 \\ 0.2367 \\ 0.2847 \end{pmatrix} \quad (\text{A.2})$$

$$\lambda_2 = 0.6973, \quad v_2 = \begin{pmatrix} 0.3807 \\ -0.2486 \\ -0.2866 \\ 0.8433 \end{pmatrix} \quad (\text{A.3})$$

$$\lambda_3 = 0.4152, \quad v_3 = \begin{pmatrix} -0.5442 \\ -0.2212 \\ 0.6940 \\ 0.4163 \end{pmatrix} \quad (\text{A.4})$$

$$\lambda_4 = 0.5166, \quad v_4 = \begin{pmatrix} -0.7331 \\ 0.2187 \\ -0.6166 \\ 0.1858 \end{pmatrix} \quad (\text{A.5})$$

A.3 MORE DISTRIBUTIONS FOR DILEPTONIC TOP-ANTITOP PAIR RECONSTRUCTION WITH LGATR

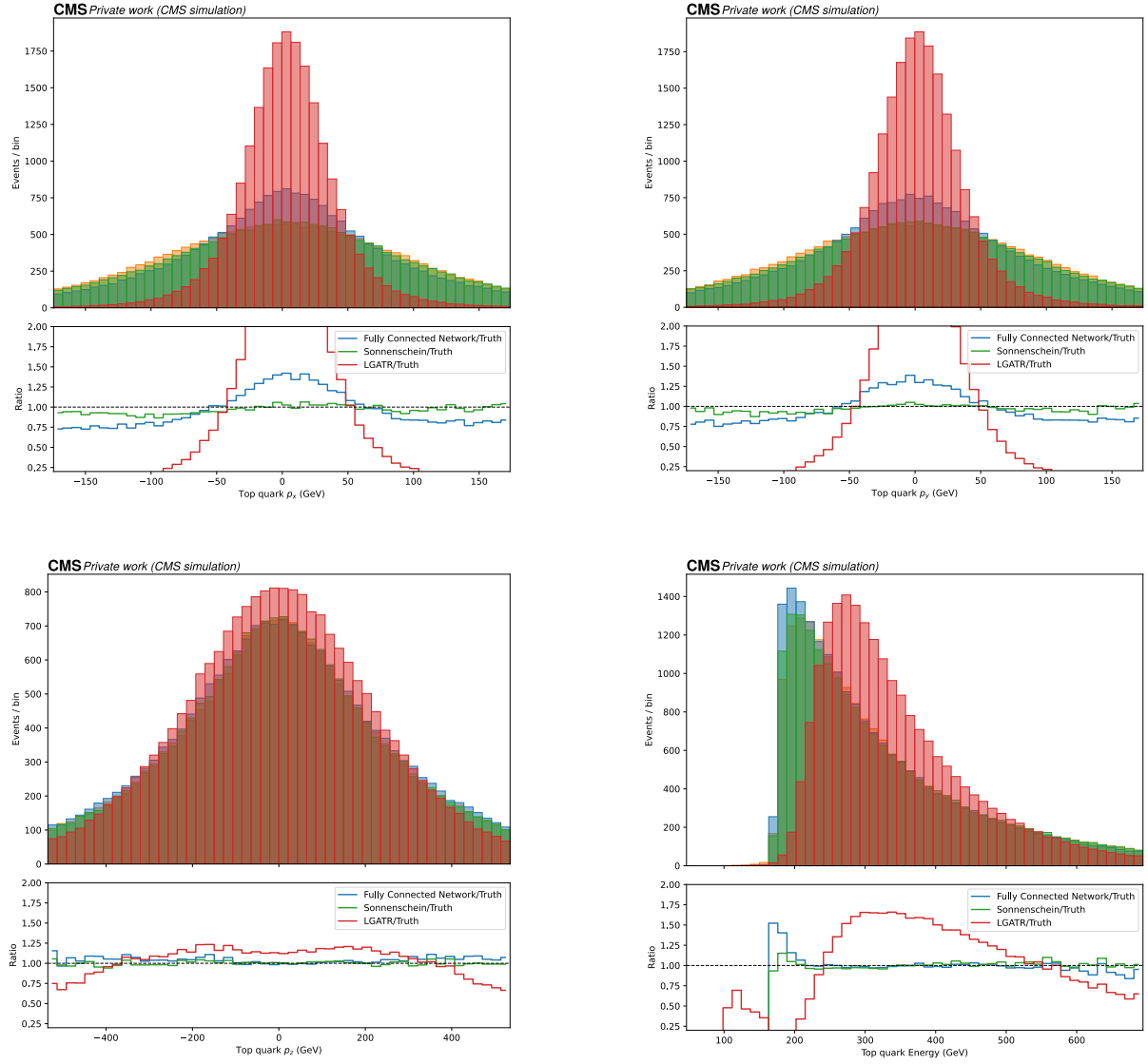


FIGURE A.4: Prediction vs. truth of the momentum fourvector components for the reconstructed top quarks. The MC truth (orange) is shown against the analytic prediction and the DNN from [98] as well as the LGATR trained in this work.

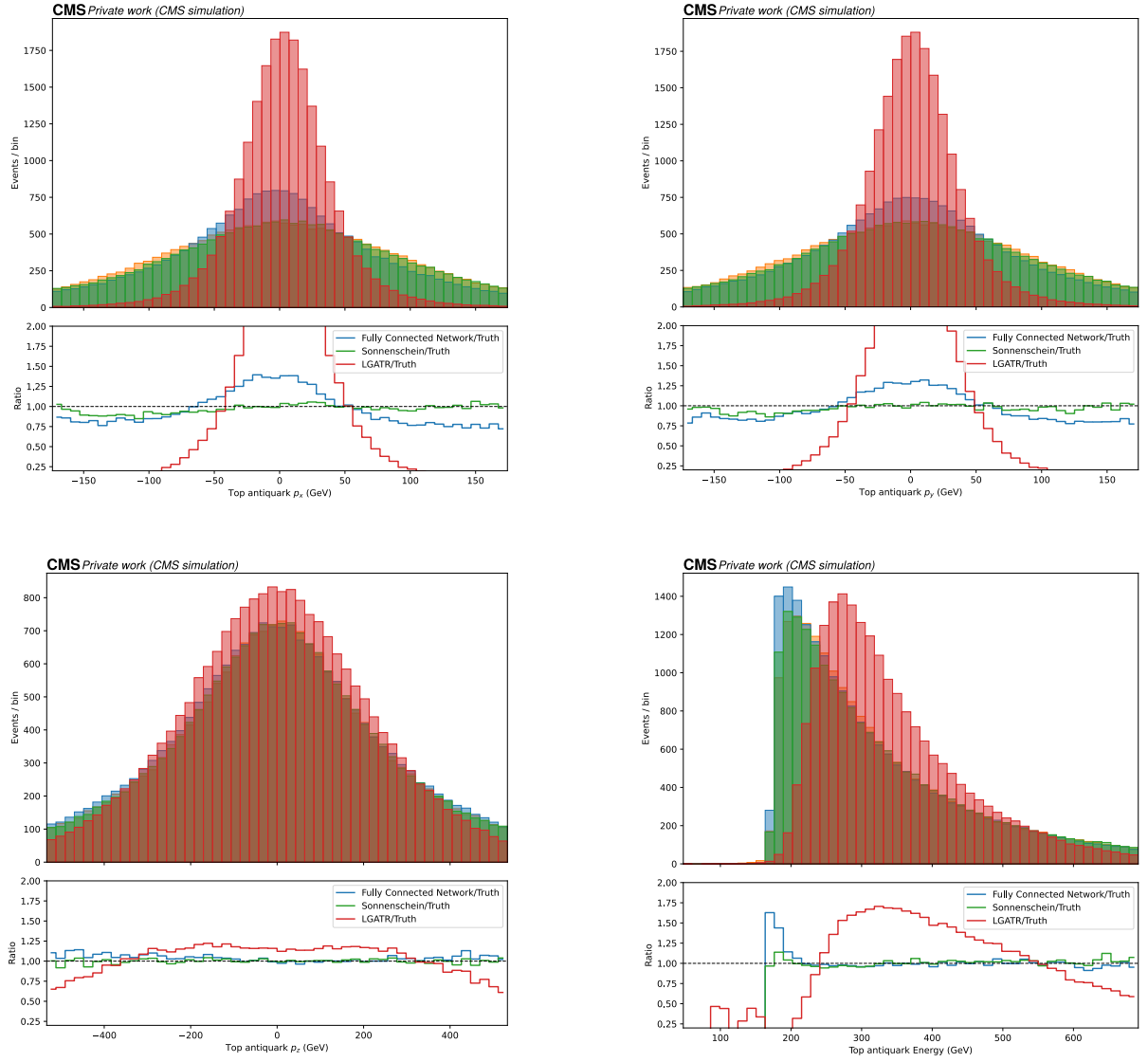


FIGURE A.5: Prediction vs. truth of the momentum fourvector components for the reconstructed antitop quarks. The MC truth (orange) is shown against the analytic prediction and the DNN from [98] as well as the LGATR trained in this work.

A.4 ALP REWEIGHTING VALIDATION & SIGNAL TEMPLATES

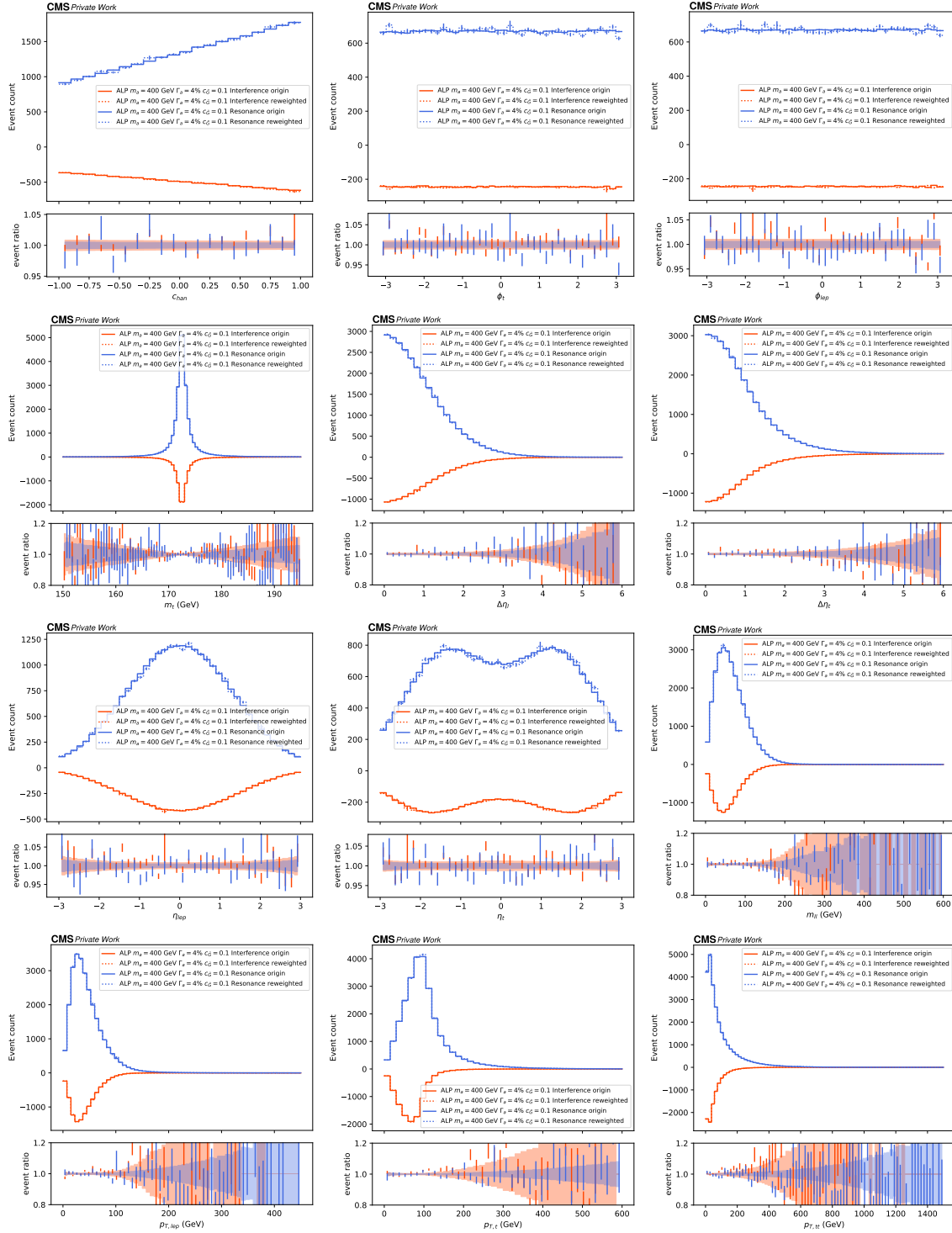
Validation Plots for $c_{\tilde{G}} = 0.1$ 

FIGURE A.6: ALP reweighting validation following the notation of plot 7.1.

Validation Plots for $c_{\tilde{G}} = 0.02$

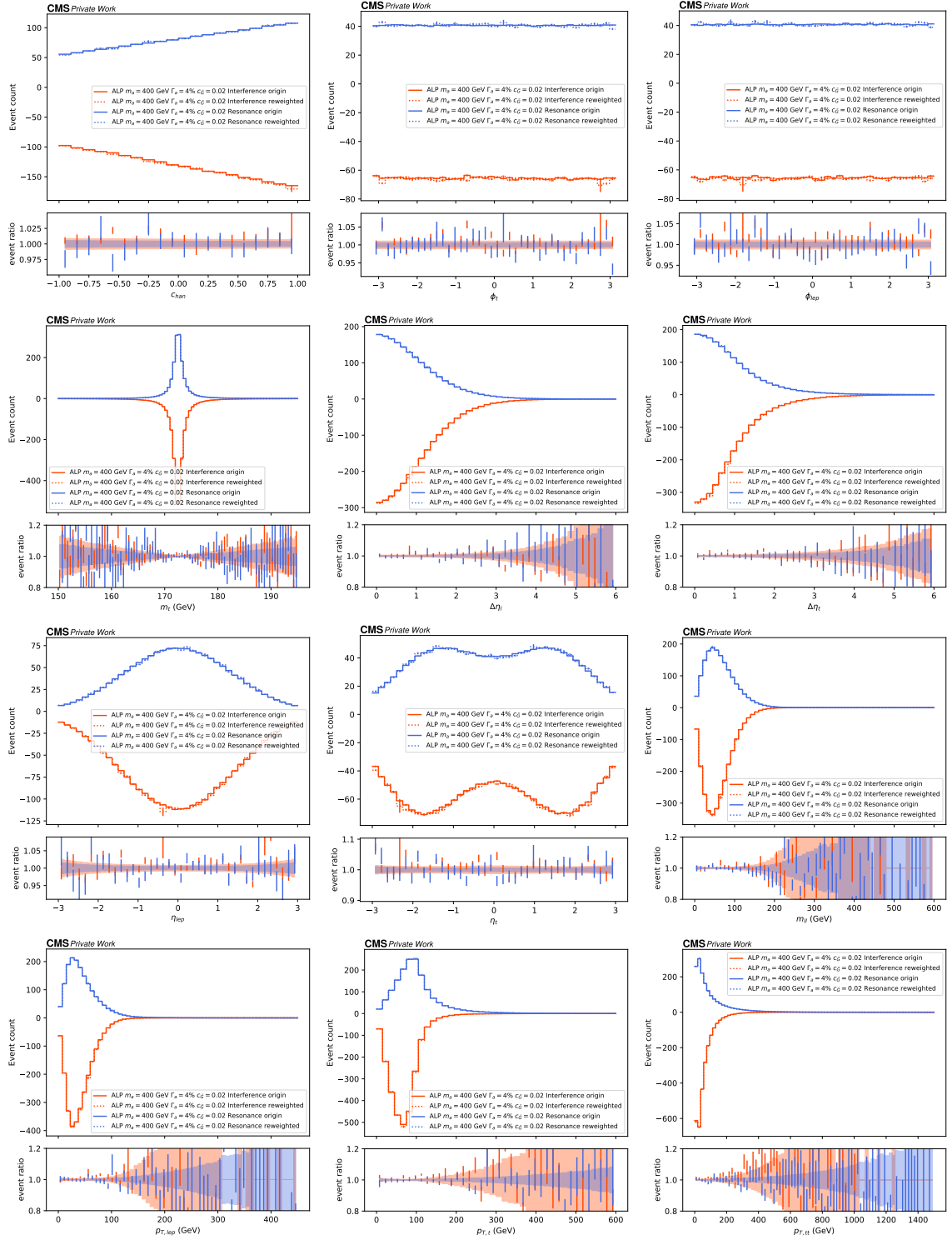


FIGURE A.7: ALP reweighting validation following the notation of plot 7.1.

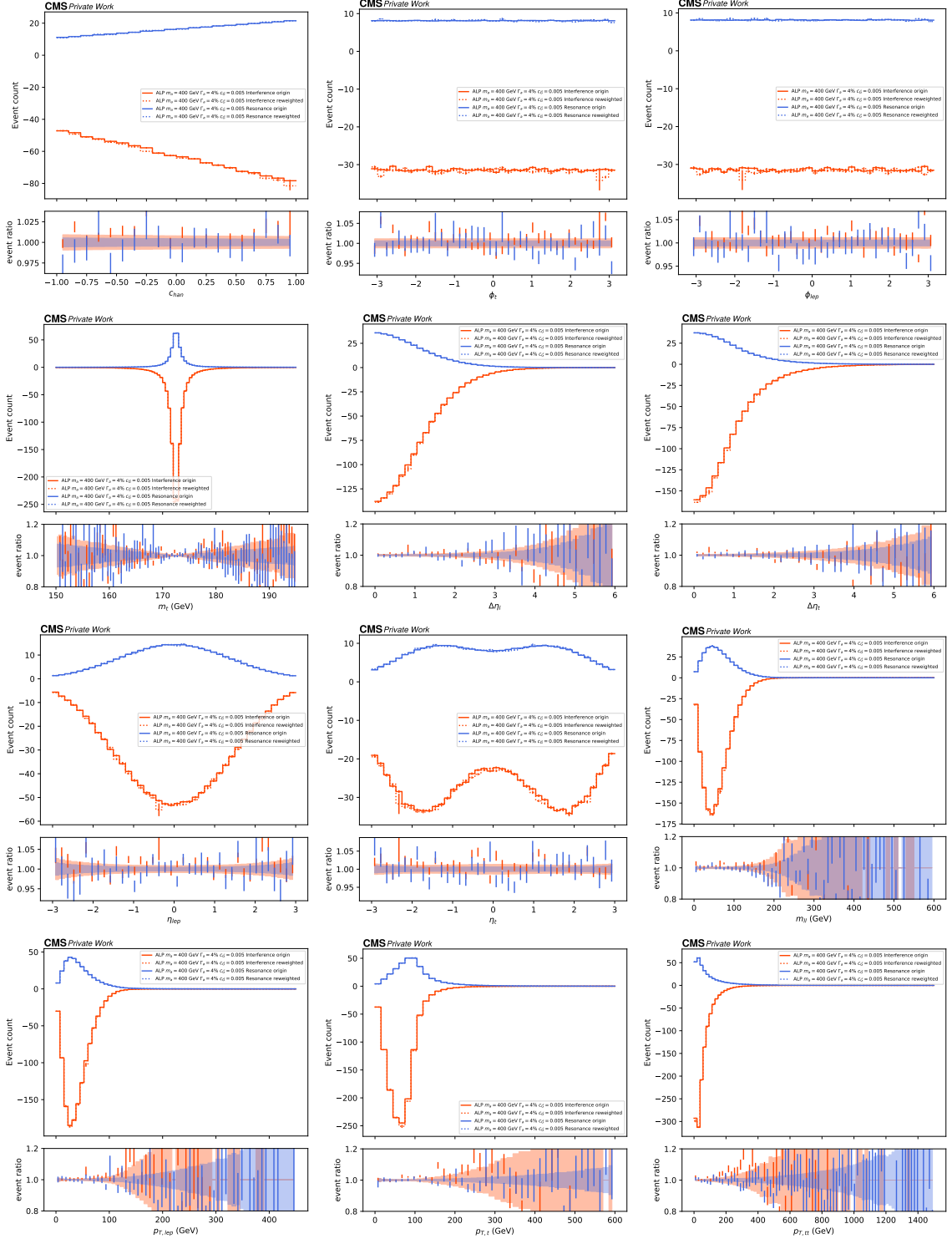
Validation Plots for $c_{\tilde{G}} = 0.005$ 

FIGURE A.8: ALP reweighting validation following the notation of plot 7.1.

Signal Templates for ALPs with different Masses

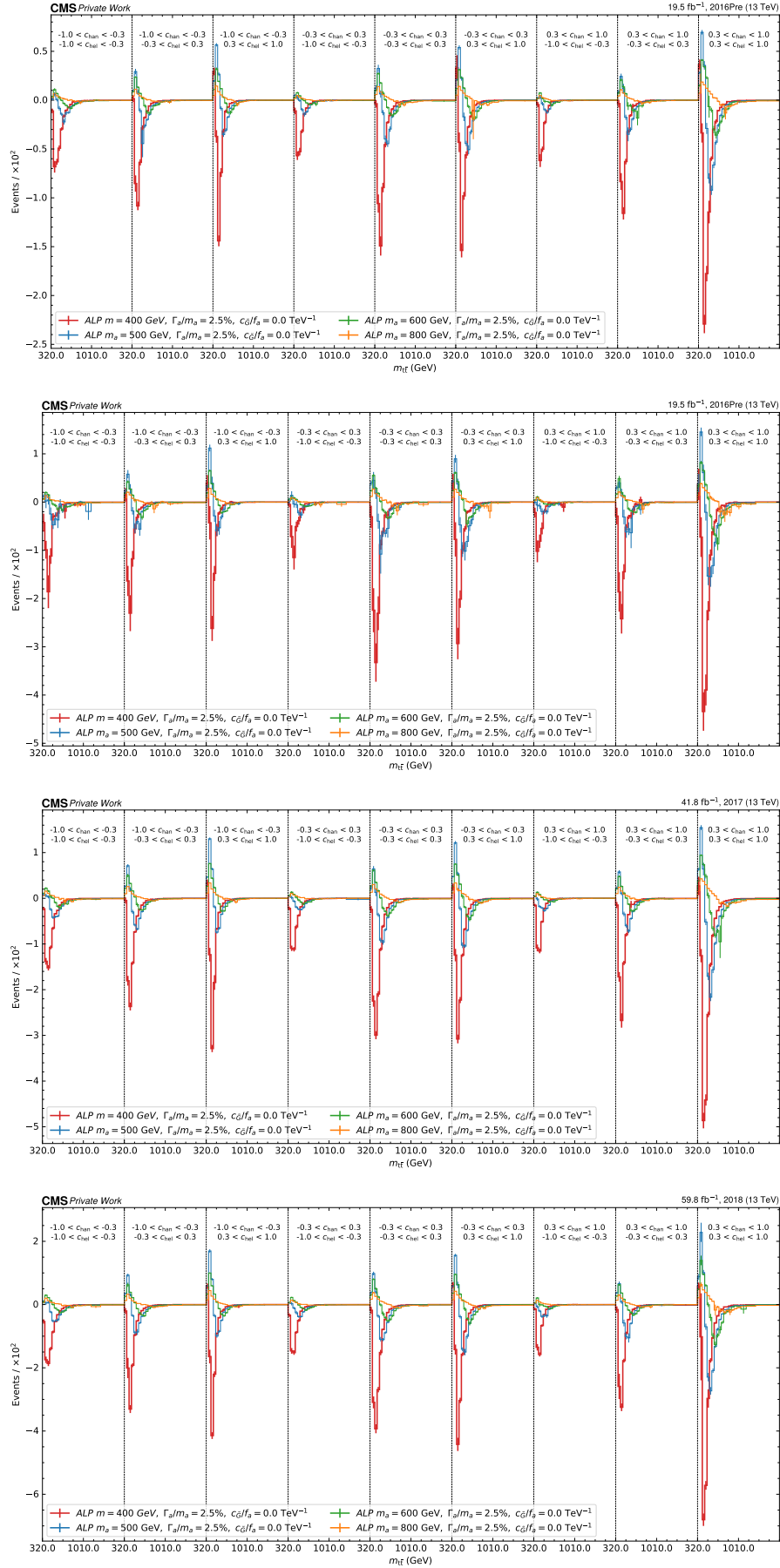


FIGURE A.9: Signal templates for ALPs with different masses and a fixed gluon coupling $c_{\tilde{G}}/f_a = 0$. The notation follows figure 7.2.

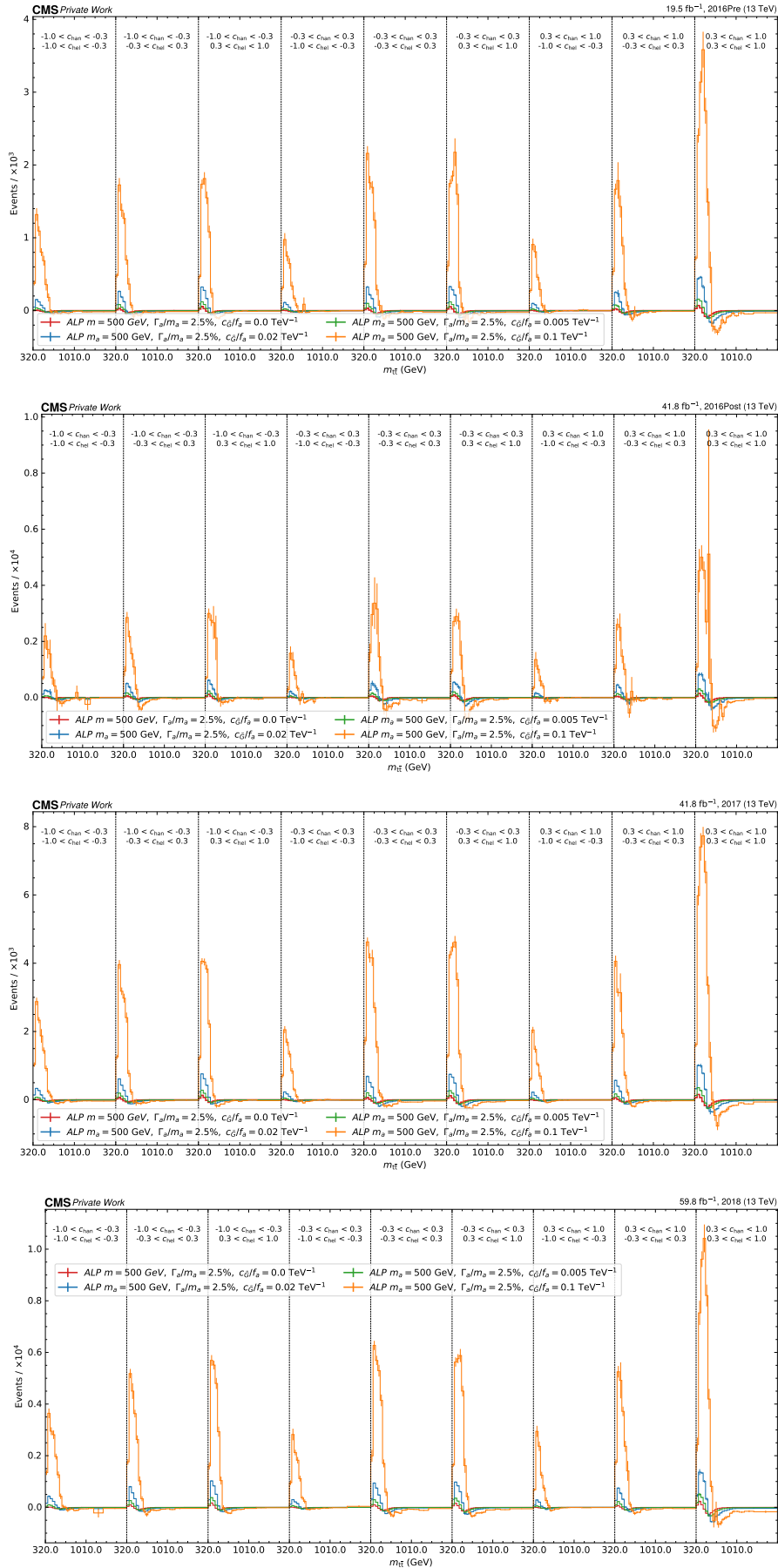
Signal Templates for ALPs with different $c_{\tilde{g}}$ 

FIGURE A.10: Signal templates for ALPs with different values of the gluon coupling $c_{\tilde{g}}/f_a$ and a fixed mass of $m_a = 500$ GeV. The notation follows figure 7.2.

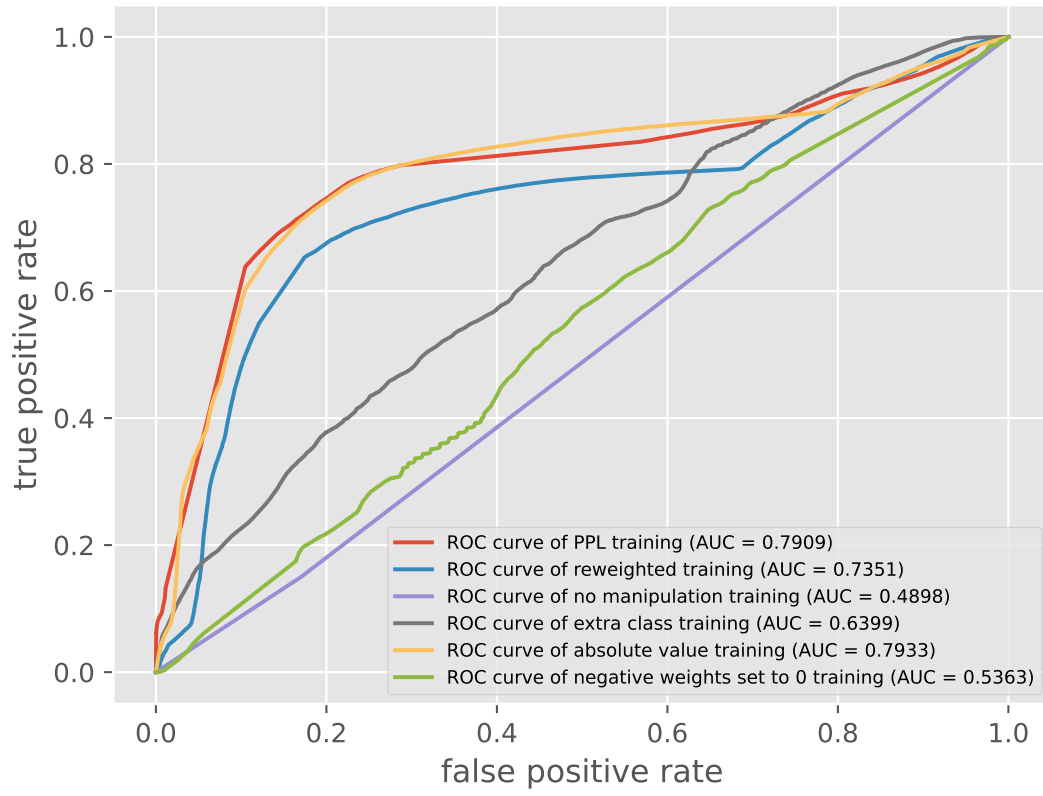


FIGURE A.11: Comparison of ROC curves for a classifier that classifies events from a single slit against events from a doubleslit. Different methods mitigating the negative weights are compared. The auxiliary feature is a normal distribution with a shift of 0.1 for the signal versus the shift of 2 in the results in the main text body.

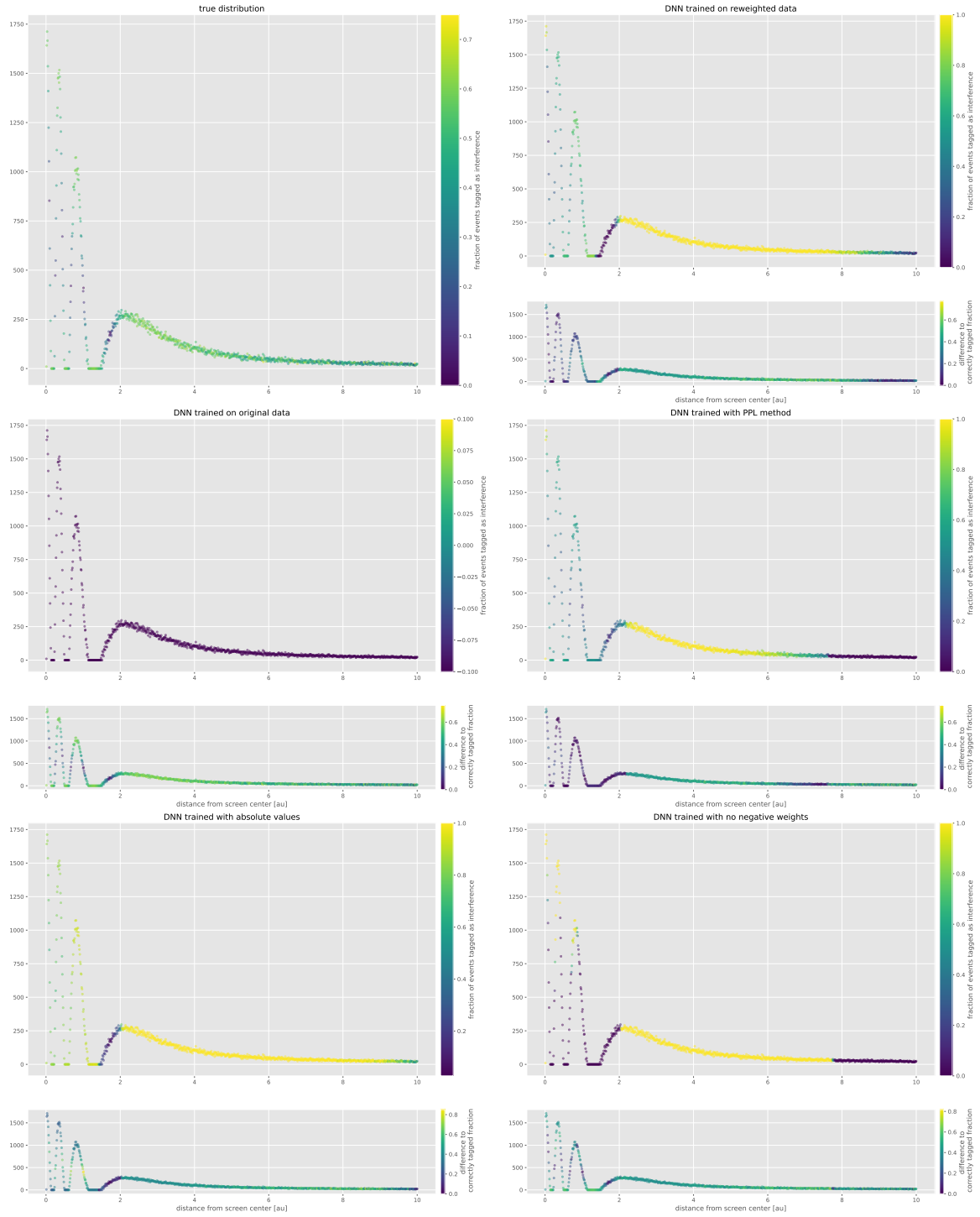


FIGURE A.12: Distributions of percentages predicted as coming from a doubleslit. The lower panels show the difference to the correct percentage in the analytical distribution. Different negative weight mitigation methods are plotted in their respective panels.

A.6 HEAVY HIGGS MOCK ANALYSIS WITH DIFFERENT NEGATIVE WEIGHT MITIGATION SCHEMES - PLOTS OF FEATURES

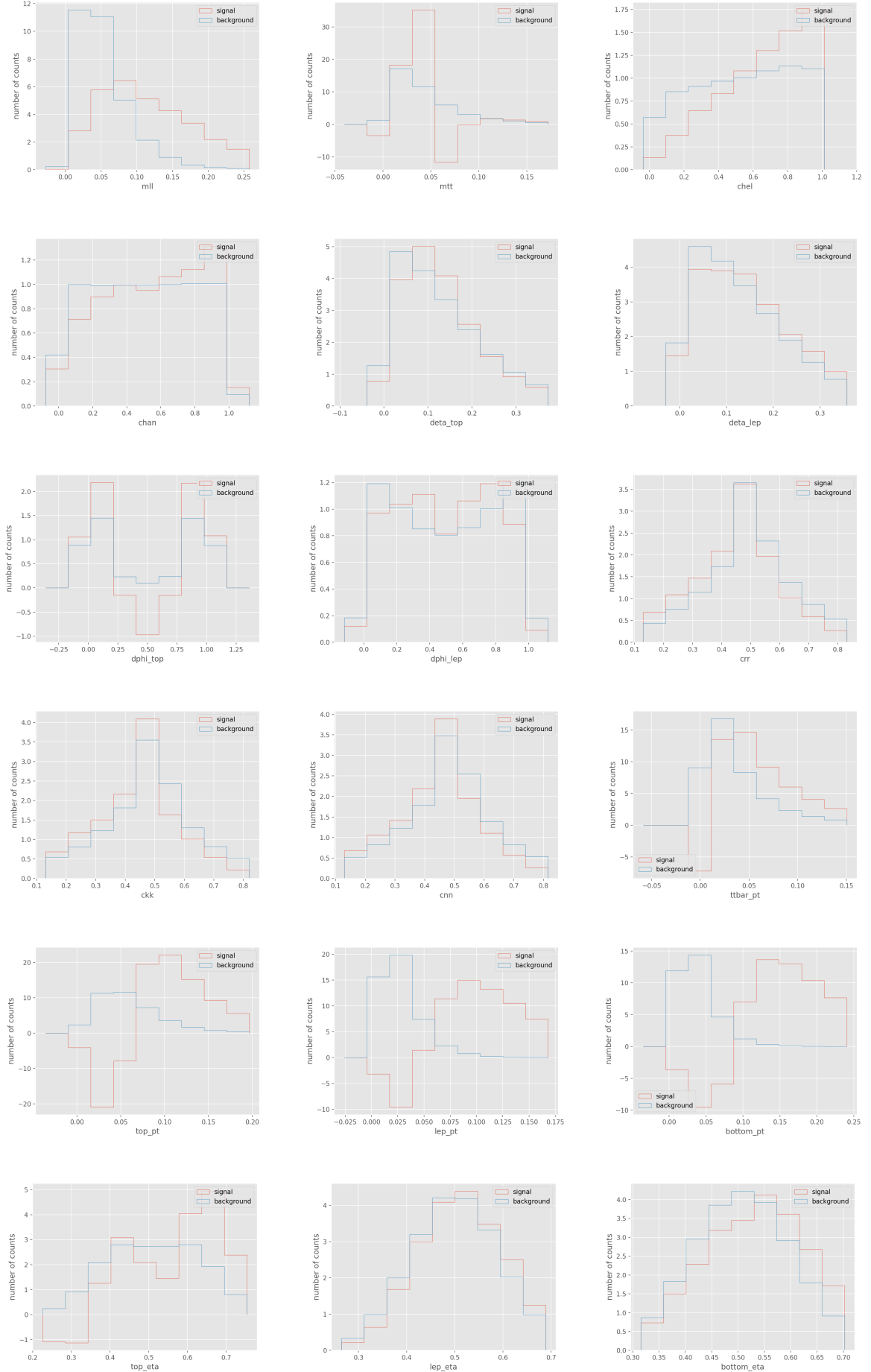


FIGURE A.13: Feature distributions for the analysis toy model. The signal scenario shown is a pseudoscalar A with a mass of 365 GeV and a width of 10% and the background description follows the explanation in section 8.6.

A.7 USING PPL AND THE LIKELIHOOD RATIO TRICK TO REWEIGHT A TOY DISTRIBUTION

The following shows that the PPL method is capable of estimating likelihood ratios in scenarios with negatively weighted contributions. To evaluate the performance of the likelihood ratio estimation in a quasiprobabilistic setting, the signed Gaussian mixture model setup described by Drnevich et al. [207] is used. This pedagogical example involves a two-dimensional problem where the target distribution is defined as a signed mixture of two centered Gaussian distributions. This setup allows for the controlled introduction of negative probability densities and negative weights.

Mathematical Formulation

The quasiprobability density function $q(\mathbf{x})$ is constructed as a weighted sum of two zero-mean, two-dimensional Gaussian distributions, $p_1(\mathbf{x}; \sigma_1)$ and $p_2(\mathbf{x}; \sigma_2)$, parameterized by their standard deviations $\sigma_{1,2}$. The mixture is controlled by a scalar coefficient $c \in (1, \infty)$, ensuring that the second component contributes with a negative weight while preserving the normalization of the total density (integrating to 1).

The density is defined as:

$$q(\mathbf{x}; c, \sigma_1, \sigma_2) = c \cdot p_1(\mathbf{x}; \sigma_1) + (1 - c) \cdot p_2(\mathbf{x}; \sigma_2), \quad (\text{A.6})$$

where the individual components are given by:

$$p_k(\mathbf{x}; \sigma_k) = \frac{1}{2\pi\sigma_k^2} \exp\left(-\frac{1}{2} \frac{\|\mathbf{x}\|^2}{\sigma_k^2}\right). \quad (\text{A.7})$$

In this construction, the term $(1 - c)$ is strictly negative. Depending on the choice of σ_1 and σ_2 , the local density $q(\mathbf{x})$ may become negative, particularly near the origin where the amplitude of the subtracted component p_2 is largest.

Dataset Configuration

A test case is used to evaluate the model with a *Signed LR*. The reference distribution (labeled $Y = 0$) is generated using parameters that ensure a strictly nonnegative density, despite the signed formulation. The target distribution (labeled $Y = 1$) is controlled by the width (σ_2), which determines whether the true likelihood ratio contains negative regions.

The specific parameters for the reference and target distribution are detailed in Table A.1.

Dataset	Mixture Coeff. (c)	Width 1 (σ_1)	Width 2 (σ_2)
<i>Common Reference</i>			
Reference ($Y = 0$)	4/3	2.5	2.3
<i>Signed LR</i>			
Target ($Y = 1$)	2.0	2.5	1.20

TABLE A.1: Parameters for the Signed Gaussian Mixture Datasets. The Reference distribution is constant across both test cases. The *Signed LR* target is distinguished by a narrower σ_2 , creating a region of negative density near the origin.

Data Generation

The training and validation datasets are generated using a method that ensures all importance weights are either +1 or -1. This is achieved by sampling from the two Gaussian components with probabilities proportional to the magnitude of their mixture coefficients, rather than uniformly.

The sampling procedure utilizes a Bernoulli auxiliary variable B to select the component:

$$B \sim \text{Bernoulli}\left(p = \frac{c}{2c - 1}\right). \quad (\text{A.8})$$

The realizations of the input features $\mathbf{X}$ and the weights W are then generated as follows:

$$(\mathbf{X}, W) \sim \begin{cases} (G_1, +1) & \text{if } B = 1 \\ (G_2, -1) & \text{if } B = 0 \end{cases} \quad (\text{A.9})$$

where $G_1 \sim \mathcal{N}(0, \sigma_1^2)$ and $G_2 \sim \mathcal{N}(0, \sigma_2^2)$.

For the reference dataset ($c = 4/3$), samples are drawn from the first component with probability $p = 0.8$. For the target datasets ($c = 2$), samples are drawn from the first component with probability $p \approx 0.67$.

Technical Realization and Network Setup

The implementation of the likelihood ratio estimation models is performed using the PyTorch deep learning framework. A model with no further treatment of negative weights and a model using the PPL method are trained.

Network Architecture

Both methodologies utilize an identical MLP architecture to ensure a fair comparison of their representational capabilities. The network consists of:

- 2 input nodes (corresponding to the Cartesian coordinates x and y).
- Two fully connected hidden layers with 64 units each, utilizing Rectified Linear Unit (ReLU) activation functions.
- A single output node.

Training Configuration

The models are trained using the Adam optimizer with a learning rate of 10^{-4} . The training procedure runs for a fixed duration of 15 epochs with a batch size of 256. The dataset consists of 1×10^6 training samples and 3×10^5 validation samples per class (reference and target).

Evaluation

Model performance is evaluated by examining the closure of the re-weighted distributions. The estimated likelihood ratio $r(x)$ is used to re-weight the reference distribution ($Y = 0$) to match the target distribution ($Y = 1$). The quality of this matching is visualized by comparing the radial density profiles $\frac{1}{N} \frac{dN}{dR}$. The results are shown in figure A.14. It is evident that PPL performs competitively while the training with negative weights shows inadequate reweighting resulting in a distribution that is closer to the reference.

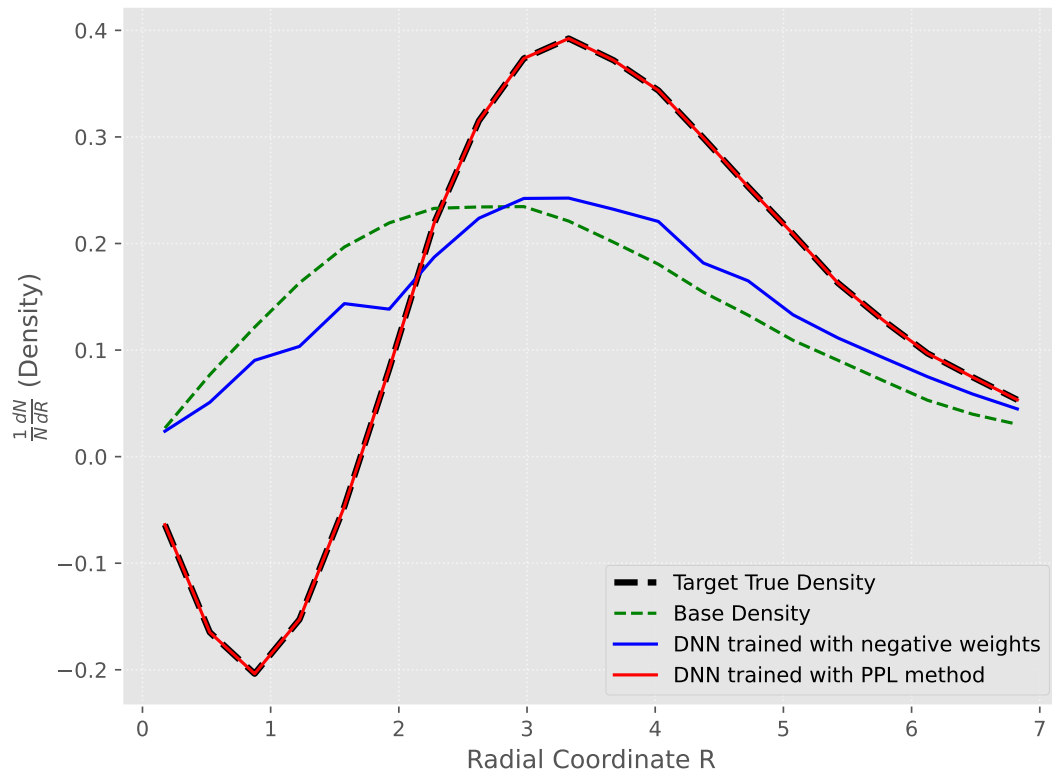


FIGURE A.14: Likelihood ratio based reweighting results for a setup derived from [207]. The PPL results are shown against a normal training of a classifier with negative weights during training.

BIBLIOGRAPHY

1. Aad, G. *et al.* Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B* **716**, 1 (2012).
2. Chatrchyan, S. *et al.* Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC. *Phys. Lett. B* **716**, 30 (2012).
3. Noether, E. Invariante Variationsprobleme. *ger. Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse* **1918**, 235 (1918).
4. Luders, G. On the Equivalence of Invariance under Time Reversal and under Particle-Antiparticle Conjugation for Relativistic Field Theories. *Kong. Dan. Vid. Sel. Mat. Fys. Med.* **28N5**, 1 (1954).
5. Pauli, W. in *Wolfgang Pauli: Das Gewissen der Physik* (eds Enz, C. P. & v. Meyenn, K.) 459 (Vieweg+Teubner Verlag, Wiesbaden, 1988).
6. Wu, C. S., Ambler, E., Hayward, R. W., Hoppes, D. D. & Hudson, R. P. Experimental Test of Parity Conservation in β Decay. *Phys. Rev.* **105**, 1413 (1957).
7. Christenson, J. H., Cronin, J. W., Fitch, V. L. & Turlay, R. Evidence for the 2π Decay of the K_2^0 Meson. *Phys. Rev. Lett.* **13**, 138 (1964).
8. Kinoshita, T. Mass Singularities of Feynman Amplitudes. *Journal of Mathematical Physics* **3**, 650 (1962).
9. Lee, T. D. & Nauenberg, M. Degenerate Systems and Mass Singularities. *Phys. Rev.* **133**, B1549 (6B 1964).
10. Fritzsche, H., Gell-Mann, M. & Leutwyler, H. Advantages of the Color Octet Gluon Picture. *Phys. Lett. B* **47**, 365 (1973).
11. Plehn, T. Lectures on LHC Physics. *Lect. Notes Phys.* **844**, 1 (2012).
12. Higgs, P. W. Broken Symmetries and the Masses of Gauge Bosons. *Phys. Rev. Lett.* **13**, 508 (16 1964).
13. Englert, F. & Brout, R. Broken Symmetry and the Mass of Gauge Vector Mesons. *Phys. Rev. Lett.* **13**, 321 (9 1964).
14. Guralnik, G. S., Hagen, C. R. & Kibble, T. W. B. Global Conservation Laws and Massless Particles. *Phys. Rev. Lett.* **13**, 585 (20 1964).
15. Sirunyan, A. M. *et al.* Measurement of the top quark Yukawa coupling from $t\bar{t}$ kinematic distributions in the dilepton final state in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. D* **102**, 092013 (2020).
16. Fukuda, Y. *et al.* Evidence for Oscillation of Atmospheric Neutrinos. *Phys. Rev. Lett.* **81**, 1562 (8 1998).
17. Aghanim, N. *et al.* Planck 2018 results. VI. Cosmological parameters. *Astron. Astrophys.* **641**. [Erratum: *Astron. Astrophys.* 652, C4 (2021)], A6 (2020).
18. Navas, S. *et al.* Review of particle physics. *Phys. Rev. D* **110**, 030001 (2024).
19. Bernreuther, W., Heisler, D. & Si, Z.-G. A set of top quark spin correlation and polarization observables for the LHC: Standard Model predictions and new physics contributions. *JHEP* **12**, 026 (2015).
20. Aaltonen, T. *et al.* Measurement of $t\bar{t}$ Spin Correlation in $p\bar{p}$ Collisions Using the CDF II Detector at the Tevatron. *Phys. Rev. D* **83**, 031104 (2011).
21. Abbott, B. *et al.* Spin correlation in $t\bar{t}$ production from $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV. *Phys. Rev. Lett.* **85**, 256 (2000).
22. Aad, G. *et al.* Observation of spin correlation in $t\bar{t}$ events from pp collisions at $\sqrt{s} = 7$ TeV using the ATLAS detector. *Phys. Rev. Lett.* **108**, 212001 (2012).
23. Sirunyan, A. M. *et al.* 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. *Phys. Rev. D* **100**, 072002 (2019).
24. Catani, S., Devoto, S., Grazzini, M., Kallweit, S. & Mazzitelli, J. Top-quark pair production at the LHC: Fully differential QCD predictions at NNLO. *JHEP* **07**, 100 (2019).
25. Maltoni, F., Severi, C., Tentori, S. & Vryonidou, E. Quantum detection of new physics in top-quark pair production at the LHC. *JHEP* **03**, 099 (2024).
26. Bernreuther, W., Brandenburg, A., Si, Z. G. & Uwer, P. Top quark pair production and decay at hadron colliders. *Nucl. Phys. B* **690**, 81 (2004).
27. Anuar, A. A. *Top Quark Spin and Polarization Properties in Searches for New Phenomena with the CMS Detector at the LHC* 2020.
28. Aad, G. *et al.* Observation of quantum entanglement with top quarks at the ATLAS detector. *Nature* **633**, 542 (2024).

29. Hayrapetyan, A. *et al.* Observation of quantum entanglement in top quark pair production in proton–proton collisions at $\sqrt{s} = 13$ TeV. *Rept. Prog. Phys.* **87**, 117801 (2024).
30. Hook, A. TASI Lectures on the Strong CP Problem and Axions. *PoS TASI2018*, 004 (2019).
31. Pendlebury, J. M. *et al.* Revised experimental upper limit on the electric dipole moment of the neutron. *Phys. Rev. D* **92**, 092003 (2015).
32. Rubin, V. C., Ford Jr., W. K. & Thonnard, N. Rotational properties of 21 SC galaxies with a large range of luminosities and radii, from NGC 4605 ($R=4\text{kpc}$) to UGC 2885 ($R=122\text{kpc}$). *Astrophysical Journal* **238**, 471 (1980).
33. Zwicky, F. On the Masses of Nebulae and of Clusters of Nebulae. *Astrophysical Journal* **86**, 217 (1937).
34. Clowe, D., Bradac, M., Gonzalez, A. H., Markevitch, M., Randall, S. W., Jones, C. & Zaritsky, D. A direct empirical proof of the existence of dark matter. *Astrophys. J. Lett.* **648**, L109 (2006).
35. Sakharov, A. D. Violation of CP Invariance, C asymmetry, and baryon asymmetry of the universe. *Pisma Zh. Eksp. Teor. Fiz.* **5**, 32 (1967).
36. Gavela, M. B., Hernandez, P., Orloff, J. & Pene, O. Standard model CP violation and baryon asymmetry. *Mod. Phys. Lett. A* **9**, 795 (1994).
37. Branco, G., Ferreira, P., Lavoura, L., Rebelo, M., Sher, M. & Silva, J. P. Theory and phenomenology of two-Higgs-doublet models. *Physics Reports* **516**. Theory and phenomenology of two-Higgs-doublet models, 1 (2012).
38. Haber, H. & Kane, G. The search for supersymmetry: Probing physics beyond the standard model. *Physics Reports* **117**, 75 (1985).
39. Kim, J. E. Light pseudoscalars, particle physics and cosmology. *Physics Reports* **150**, 1 (1987).
40. Karan, A., Miralles, V. & Pich, A. Updated global fit of the aligned two-Higgs-doublet model with heavy scalars. *Phys. Rev. D* **109**, 035012 (2024).
41. Aad, G. *et al.* Search for heavy neutral Higgs bosons decaying into a top quark pair in 140 fb^{-1} of proton–proton collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector. *JHEP* **08**, 013 (2024).
42. Hayrapetyan, A. *et al.* Search for heavy neutral Higgs bosons A and H in the $t\bar{t}Z$ channel in proton–proton collisions at 13 TeV. *Phys. Lett. B* **866**, 139568 (2025).
43. Hayrapetyan, A. *et al.* Search for heavy pseudoscalar and scalar bosons decaying to a top quark pair in proton–proton collisions at $\sqrt{s} = 13$ TeV. *Rept. Prog. Phys.* **88**, 127801 (2025).
44. Bernreuther, W., Flesch, M. & Haberl, P. Signatures of Higgs bosons in the top quark decay channel at hadron colliders. *Phys. Rev. D* **58**, 114031 (1998).
45. Bahl, H., Kumar, R. & Weiglein, G. Impact of interference effects on Higgs-boson searches in the di-top final state at the LHC. *JHEP* **05**, 098 (2025).
46. Fuchs, E. & Weiglein, G. Breit-Wigner approximation for propagators of mixed unstable states. *JHEP* **09**, 079 (2017).
47. Fuchs, E. & Weiglein, G. Impact of CP-violating interference effects on MSSM Higgs searches. *Eur. Phys. J. C* **78**, 87 (2018).
48. Wilczek, F. Problem of Strong P and T Invariance in the Presence of Instantons. *Phys. Rev. Lett.* **40**, 279 (5 1978).
49. Weinberg, S. A New Light Boson? *Phys. Rev. Lett.* **40**, 223 (4 1978).
50. Peccei, R. D. & Quinn, H. R. Constraints imposed by CP conservation in the presence of pseudoparticles. *Phys. Rev. D* **16**, 1791 (6 1977).
51. Peccei, R. D. & Quinn, H. R. CP Conservation in the Presence of Pseudoparticles. *Phys. Rev. Lett.* **38**, 1440 (25 1977).
52. Di Luzio, L., Giannotti, M., Nardi, E. & Visinelli, L. The landscape of QCD axion models. *Phys. Rept.* **870**, 1 (2020).
53. Anuar, A., Biekötter, A., Biekötter, T., Grohsjean, A., Heinemeyer, S., Jeppe, L., Schwanenberger, C. & Weiglein, G. ALP-ine quests at the LHC: hunting axion-like particles via peaks and dips in $t\bar{t}$ production. *JHEP* **12**, 197 (2024).
54. LHC Machine. *JINST* **3** (eds Evans, L. & Bryant, P.) So8001 (2008).
55. The ATLAS Collaboration. *ATLAS detector and physics performance* (CERN, Geneva, 1999).
56. Chatrchyan, S. *et al.* The CMS Experiment at the CERN LHC. *JINST* **3**, So8004 (2008).
57. The LHCb Collaboration. The LHCb Detector at the LHC. *JINST* **3**. Also published by CERN Geneva in 2010, So8005 (2008).

58. Aamodt, K. *et al.* The ALICE experiment at the CERN LHC. *JINST* **3**, So8002 (2008).
59. Collaboration, H.-L. *High-Luminosity Large Hadron Collider (HL-LHC): Technical design report* (CERN, Geneva, 2020).
60. CMS-Collaboration. *CMS Schematic* <https://cms.cern/index.php/detector>. [Online; accessed 25.08.2025].
61. Sirunyan, A. M. *et al.* Particle-flow reconstruction and global event description with the CMS detector. *JINST* **12**, P10003 (2017).
62. Cacciari, M., Salam, G. P. & Soyez, G. The anti- k_t jet clustering algorithm. *JHEP* **04**, 063 (2008).
63. CMS-Collaboration. *CMS Software (CMSSW)* <https://cms-sw.github.io/index.html>. [Online, accessed 27.08.2025].
64. Cranmer, K. Practical Statistics for the LHC. *CERN Proceedings of the 2013 CERN-Latin-American School of High-Energy Physics* (2013).
65. Brun, R. & Rademakers, F. ROOT: An object oriented data analysis framework. *Nucl. Instrum. Meth. A* **389** (eds Werlen, M. & Perret-Gallix, D.) 81 (1997).
66. Neyman, J. & Pearson, E. S. IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **231**, 289 (1933).
67. Wilks, S. S. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Annals Math. Statist.* **9**, 60 (1938).
68. Bernlochner, F. U., Fry, D. C., Menary, S. B. & Persson, E. Cover your bases: asymptotic distributions of the profile likelihood ratio when constraining effective field theories in high-energy physics. *SciPost Phys. Core* **6**, 013 (2023).
69. Read, A. L. Presentation of search results: the CLs technique. *Journal of Physics G: Nuclear and Particle Physics* **28**, 2693 (2002).
70. Feldman, G. J. & Cousins, R. D. A Unified approach to the classical statistical analysis of small signals. *Phys. Rev. D* **57**, 3873 (1998).
71. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* <http://www.deeplearningbook.org> (MIT Press, 2016).
72. Nair, V. & Hinton, G. Rectified Linear Units Improve Restricted Boltzmann Machines Vinod Nair. *Proceedings of ICML* **27**, 807 (2010).
73. Maierov, V. & Pinkus, A. Lower bounds for approximation by MLP neural networks. *Neurocomputing* **25**, 81 (1999).
74. Cour, T., Sapp, B. & Taskar, B. Learning from Partial Labels. *Journal of Machine Learning Research* **12**, 1501 (2011).
75. Liu, W., Wang, H., Shen, X. & Tsang, I. W. The Emerging Trends of Multi-Label Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**, 7955 (2022).
76. Feng, L., Lv, J., Han, B., Xu, M., Niu, G., Geng, X., An, B. & Sugiyama, M. *Provably Consistent Partial-Label Learning in Proceedings of the 34th International Conference on Neural Information Processing Systems* (Curran Associates Inc., Red Hook, NY, United States, 2020), 10948.
77. Tian, Y., Yu, X. & Fu, S. Partial Label Learning: Taxonomy, Analysis and Outlook. *Neural Networks* **161**, 708 (2023).
78. Zhang, M., Yu, F. & Tang, C. Disambiguation-Free Partial Label Learning. *IEEE Transactions on Knowledge and Data Engineering* **29**, 2155 (2017).
79. Zhang, M. *Disambiguation-Free Partial Label Learning in Proceedings of the 14th SIAM International Conference on Data Mining* (2014), 37.
80. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278 (1998).
81. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* **60**, 84 (2012).
82. He, K., Zhang, X., Ren, S. & Sun, J. *Deep Residual Learning for Image Recognition in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 770.
83. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. *Attention is all you need in Proceedings of the 31st International Conference on Neural Information Processing Systems* (Curran Associates Inc., Long Beach, California, USA, 2017), 6000.
84. Butter, A. *et al.* The Machine Learning landscape of top taggers. *SciPost Phys.* **7** (eds Kasieczka, G. & Plehn, T.) 014 (2019).

85. Komiske, P. T., Metodiev, E. M. & Thaler, J. Energy Flow Networks: Deep Sets for Particle Jets. *JHEP* **01**, 121 (2019).
86. Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R. R. & Smola, A. J. Deep sets. *Advances in neural information processing systems* **30** (2017).
87. Brehmer, J., Bresó, V., de Haan, P., Plehn, T., Qu, H., Spinner, J. & Thaler, J. A Lorentz-Equivariant Transformer for All of the LHC. *SciPost Physics* **19**, 108 (2024).
88. Spinner, J., Bresó, V., de Haan, P., Plehn, T., Thaler, J. & Brehmer, J. Lorentz-Equivariant Geometric Algebra Transformers for High-Energy Physics (2024).
89. Brehmer, J., De Haan, P., Behrends, S. & Cohen, T. S. Geometric algebra transformer. *Advances in Neural Information Processing Systems* **36**, 35472 (2023).
90. Edelen, A. *et al.* Opportunities in Machine Learning for Particle Accelerators. *FERMILAB-PUB-19-017-AD* (2018).
91. Plehn, T., Butter, A., Dillon, B., Heimel, T., Krause, C. & Winterhalder, R. Modern Machine Learning for LHC Physicists (2022).
92. Macaluso, S. & Shih, D. Pulling Out All the Tops with Computer Vision and Deep Learning. *JHEP* **10**, 121 (2018).
93. Roe, B. P., Yang, H.-J., Zhu, J., Liu, Y., Stancu, I. & McGregor, G. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **543**, 577 (2005).
94. Metodiev, E. M., Nachman, B. & Thaler, J. Classification without labels: Learning from mixed samples in high energy physics. *JHEP* **10**, 174 (2017).
95. Blance, A., Spannowsky, M. & Waite, P. Adversarially-trained autoencoders for robust unsupervised new physics searches. *JHEP* **10**, 047 (2019).
96. Belis, V., Odagiu, P. & Arrestand, T. K. Machine learning for anomaly detection in particle physics. *Reviews in Physics* **12**, 100091 (2024).
97. Andreassen, A. & Nachman, B. Neural Networks for Full Phase-space Reweighting and Parameter Tuning. *Phys. Rev. D* **101**, 091901 (2020).
98. Rubenach, J. *Search for heavy Higgs bosons in conjunction with neural-network-driven reconstruction and upgrade of the Fast Beam Condition Monitor at the CMS experiment 2023*.
99. Ehrke, L., Raine, J. A., Zoch, K., Guth, M. & Golling, T. Topological reconstruction of particle physics processes using graph neural networks. *Phys. Rev. D* **107**, 116019 (2023).
100. Leigh, M., Raine, J. A., Zoch, K. & Golling, T. ν -flows: Conditional neutrino regression. *SciPost Phys.* **14**, 159 (2023).
101. Cranmer, K., Brehmer, J. & Louppe, G. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences* **117**, 30055 (2020).
102. Brehmer, J., Kling, F., Espejo, I. & Cranmer, K. MadMiner: Machine learning-based inference for particle physics. *Comput. Softw. Big Sci.* **4**, 3 (2020).
103. Brehmer, J., Louppe, G., Pavez, J. & Cranmer, K. Mining gold from implicit models to improve likelihood-free inference. *Proc. Nat. Acad. Sci.* **117**, 5242 (2020).
104. Stoye, M., Brehmer, J., Louppe, G., Pavez, J. & Cranmer, K. Likelihood-free inference with an improved cross-entropy estimator (2018).
105. Fisher, R. A. On the Mathematical Foundations of Theoretical Statistics. *Phil. Trans. Roy. Soc. Lond. A* **222**, 309 (1922).
106. Rao, C. Information and the Accuracy Attainable in the Estimation of Statistical Parameters. *Bull. Calcutta Math. Soc.* **37** (1945).
107. Cramer, H. *Mathematical Methods of Statistics*. Princeton University Press (1946).
108. Collins, J. C., Soper, D. E. & Sterman, G. F. Factorization of Hard Processes in QCD. *Adv. Ser. Direct. High Energy Phys.* **5**, 1 (1989).
109. Oleari, C. The POWHEG BOX. *Nuclear Physics B - Proceedings Supplements* **205–206**, 36 (2010).
110. Frixione, S., Nason, P. & Ridolfi, G. A Positive-weight next-to-leading-order Monte Carlo for heavy flavour hadroproduction. *JHEP* **09**, 126 (2007).
111. Frixione, S., Laenen, E., Motylinski, P. & Webber, B. R. Angular correlations of lepton pairs from vector boson and top quark decays in Monte Carlo simulations. *JHEP* **04**, 081 (2007).
112. Monni, P. F., Nason, P., Re, E., Wiesemann, M. & Zanderighi, G. MiNNLO_{PS}: a new method to match NNLO QCD to parton showers. *JHEP* **05**. [Erratum: *JHEP* **02**, 031 (2022)], 143 (2020).

113. Hamilton, K., Nason, P. & Zanderighi, G. MINLO: Multi-Scale Improved NLO. *JHEP* **10**, 155 (2012).
114. Frederix, R. & Hamilton, K. Extending the MINLO method. *JHEP* **05**, 042 (2016).
115. Mazzitelli, J., Monni, P. F., Nason, P., Re, E., Wieseemann, M. & Zanderighi, G. Next-to-Next-to-Leading Order Event Generation for Top-Quark Pair Production. *Phys. Rev. Lett.* **127**, 062001 (2021).
116. Mazzitelli, J., Monni, P. F., Nason, P., Re, E., Wieseemann, M. & Zanderighi, G. Top-pair production at the LHC with MINNLO_{PS}. *JHEP* **04**, 079 (2022).
117. Bucciioni, F., Lang, J.-N., Lindert, J. M., Maierhöfer, P., Pozzorini, S., Zhang, H. & Zoller, M. F. OpenLoops 2. *Eur. Phys. J. C* **79**, 866 (2019).
118. Alioli, S., Nason, P., Oleari, C. & Re, E. A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX. *JHEP* **06**, 043 (2010).
119. Ball, R. D. *et al.* Parton distributions from high-precision collider data. *Eur. Phys. J. C* **77**, 663 (2017).
120. Bierlich, C. *et al.* A comprehensive guide to the physics and usage of PYTHIA 8.3. *SciPost Phys. Codeb.* **2022**, 8 (2022).
121. Sirunyan, A. M. *et al.* Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements. *Eur. Phys. J. C* **80**, 4 (2020).
122. Czakon, M. & Mitov, A. TOP++: a program for the calculation of the top-pair cross-section at hadron colliders. *Comput. Phys. Commun.* **185**, 2930 (2014).
123. Bierlich, C., Buckley, A., Butterworth, J. M., Gutsche, C., Lonnblad, L., Procter, T., Richardson, P. & Yeh, Y. Robust independent validation of experiment and theory: Rivet version 4 release note. *SciPost Phys. Codeb.* **36**, 1 (2024).
124. Sirunyan, A. M. *et al.* Measurement of differential cross sections for the production of top quark pairs and of additional jets in lepton+jets events from pp collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. D* **97**, 112003 (2018).
125. Serkin, L. *Treatment of top-quark backgrounds in extreme phase spaces: the "top p_T reweighting" and novel data-driven estimations in ATLAS and CMS in 13th International Workshop on Top Quark Physics* (2021).
126. Skands, P., Carrazza, S. & Rojo, J. Tuning PYTHIA 8.1: the Monash 2013 Tune. *Eur. Phys. J. C* **74**, 3024 (2014).
127. Aad, G. *et al.* Measurement of the $t\bar{t}$ production cross-section and lepton differential distributions in $e\mu$ dilepton events from pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Eur. Phys. J. C* **80**, 528 (2020).
128. Ježo, T., Lindert, J. M., Nason, P., Oleari, C. & Pozzorini, S. An NLO+PS generator for $t\bar{t}$ and Wt production and decay including non-resonant and interference effects. *Eur. Phys. J. C* **76**, 691 (2016).
129. Hayrapetyan, A. *et al.* Observation of a pseudoscalar excess at the top quark pair production threshold. *Rept. Prog. Phys.* **88**, 087801 (2025).
130. Jeppe, L. *Measurement of the inclusive $t\bar{t}$ cross section and search for additional scalars in $t\bar{t}$ final states at CMS experiment*. PhD thesis (Universität Hamburg, Hamburg, 2025).
131. Simulation of the Silicon Strip Tracker pre-amplifier in early 2016 data (2020).
132. Bols, E., Kieseler, J., Verzetti, M., Stoye, M. & Stakia, A. Jet Flavour Classification Using DeepJet. *JINST* **15**, P12012 (2020).
133. Sirunyan, A. M. *et al.* Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV. *JINST* **13**, P05011 (2018).
134. Cacciari, M., Salam, G. P. & Soyez, G. The anti- k_t jet clustering algorithm. *JHEP* **04**, 063 (2008).
135. Sirunyan, A. M. *et al.* Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC. *JINST* **16**, P05014 (2021).
136. Sirunyan, A. M. *et al.* Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV. *JINST* **13**, P06015 (2018).
137. Chatrchyan, S. *et al.* Measurement of the $t\bar{t}$ production cross section and the top quark mass in the dilepton channel in pp collisions at $\sqrt{s} = 7$ TeV. *JHEP* **07**, 049 (2011).
138. Grazzini, M., Kallweit, S. & Wieseemann, M. Fully differential NNLO computations with MATRIX. *Eur. Phys. J. C* **78**, 537 (2018).
139. Aliev, M., Lacker, H., Langenfeld, U., Moch, S., Uwer, P. & Wiedermann, M. HATHOR: Hadronic top and heavy quarks cross section calculator. *Comput. Phys. Commun.* **182**, 1034 (2011).
140. Kühn, J. H., Scharf, A. & Uwer, P. Electroweak corrections to top-quark pair production in quark-antiquark annihilation. *Eur. Phys. J. C* **45**, 139 (2006).
141. Kühn, J. H., Scharf, A. & Uwer, P. Electroweak effects in top-quark pair production at hadron colliders. *Eur. Phys. J. C* **51**, 37 (2007).

142. Kühn, J. H., Scharf, A. & Uwer, P. Weak interactions in top-quark pair production at hadron colliders: An update. *Phys. Rev. D* **91**, 014020 (2015).
143. Frixione, S., Nason, P. & Ridolfi, G. The POWHEG-hvq manual version 1.0 (2007).
144. Re, E. Single-top Wt -channel production matched with parton showers using the Powheg method. *Eur. Phys. J. C* **71**, 1547 (2011).
145. Kidonakis, N. & Yamanaka, N. Higher-order corrections for tW production at high-energy hadron colliders. *JHEP* **05**, 278 (2021).
146. Monni, P. F., Nason, P., Re, E., Wiesemann, M. & Zanderighi, G. MiNNLO_{PS}: a new method to match NNLO QCD to parton showers. *JHEP* **05**. [Erratum: *JHEP* **02**, 031 (2022)], 143 (2020).
147. Monni, P. F., Re, E. & Wiesemann, M. MiNNLO_{PS}: optimizing $2 \rightarrow 1$ hadronic processes. *Eur. Phys. J. C* **80**, 1075 (2020).
148. Melnikov, K. & Petriello, F. Electroweak gauge boson production at hadron colliders through $\mathcal{O}(\alpha_s^2)$. *Phys. Rev. D* **74**, 114017 (2006).
149. Li, Y. & Petriello, F. Combining QCD and electroweak corrections to dilepton production in FEWZ. *Phys. Rev. D* **86**, 094034 (2012).
150. Alioli, S., Nason, P., Oleari, C. & Re, E. NLO single-top production matched with shower in POWHEG: s- and t-channel contributions. *JHEP* **09**. [Erratum: *JHEP* **02**, 011 (2010)], 111 (2009).
151. Artoisenet, P., Frederix, R., Mattelaer, O. & Rietkerk, R. Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations. *JHEP* **03**, 015 (2013).
152. Kant, P., Kind, O. M., Kintscher, T., Lohse, T., Martini, T., Mölbitz, S., Rieck, P. & Uwer, P. HATHOR for single top-quark production: Updated predictions and uncertainty estimates for single top-quark production in hadronic collisions. *Comput. Phys. Commun.* **191**, 74 (2015).
153. Gehrmann, T., Grazzini, M., Kallweit, S., Maierhöfer, P., von Manteuffel, A., Pozzorini, S., Rathlev, D. & Tancredi, L. W^+W^- Production at Hadron Colliders in Next to Next to Leading Order QCD. *Phys. Rev. Lett.* **113**, 212001 (2014).
154. Campbell, J. M. & Ellis, R. K. MCFM for the Tevatron and the LHC. *Nucl. Phys. B Proc. Suppl.* **205-206** (eds Blümlein, J., Moch, S.-O. & Riemann, T.) 10 (2010).
155. Sirunyan, A. M. *et al.* Combined measurements of Higgs boson couplings in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Eur. Phys. J. C* **79**, 421 (2019).
156. Sirunyan, A. M. *et al.* Measurement of the single top quark and antiquark production cross sections in the t channel and their ratio in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Phys. Lett. B* **800**, 135042 (2020).
157. Aaboud, M. *et al.* Measurement of the inclusive cross-sections of single top-quark and top-antiquark t -channel production in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *JHEP* **04**, 086 (2017).
158. Sirunyan, A. M. *et al.* Measurement of the cross section for top quark pair production in association with a W or Z boson in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP* **08**, 011 (2018).
159. Aaboud, M. *et al.* Measurement of the $t\bar{t}Z$ and $t\bar{t}W$ cross sections in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Rev. D* **99**, 072009 (2019).
160. Aaboud, N. *et al.* Measurements of top-quark pair to Z -boson cross-section ratios at $\sqrt{s} = 13, 8, 7$ TeV with the ATLAS detector. *JHEP* **02**, 117 (2017).
161. Sirunyan, A. M. *et al.* Precision luminosity measurement in proton-proton collisions at $\sqrt{s} = 13$ TeV in 2015 and 2016 at CMS. *Eur. Phys. J. C* **81**, 800 (2021).
162. CMS Collaboration. CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV CMS Physics Analysis Summary CMS-PAS-LUM-17-004 (2018).
163. CMS Collaboration. CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV CMS Physics Analysis Summary CMS-PAS-LUM-18-002 (2019).
164. Sirunyan, A. M. *et al.* Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 13$ TeV. *JHEP* **07**, 161 (2018).
165. Khachatryan, V. *et al.* Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV. *JINST* **12**, P02014 (2017).
166. Sirunyan, A. M. *et al.* Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector. *JINST* **14**, P07004 (2019).
167. Sirunyan, A. M. *et al.* Performance of the CMS Level-1 trigger in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JINST* **15**, P10017 (2020).
168. Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.* **74**, 829 (1979).

169. Cleveland, W. S. & Devlin, S. J. Locally weighted regression: an approach to regression analysis by local fitting. *J. Am. Stat. Assoc.* **83**, 596 (1988).
170. Sonnenschein, L. Analytical solution of $t\bar{t}$ dilepton equations. *Phys. Rev. D* **73**, 054015 (2006).
171. Betchart, B. A., Demina, R. & Harel, A. Analytic solutions for neutrino momenta in decay of top quarks. *Nucl. Instrum. Meth. A* **736**, 169 (2014).
172. Demina, R., Harel, A. & Orbaker, D. Reconstructing $t\bar{t}$ events with one lost jet. *Nucl. Instrum. Meth. A* **788**, 128 (2015).
173. Kiyo, Y., Kühn, J. H., Moch, S., Steinhauser, M. & Uwer, P. Top-quark pair production near threshold at LHC. *Eur. Phys. J. C* **60**, 375 (2009).
174. Sumino, Y. & Yokoya, H. Bound-state effects on kinematical distributions of top quarks at hadron colliders. *JHEP* **09**. [Erratum: 10.1007/JHEP06(2016)037], 034 (2010).
175. Ju, W.-L., Wang, G., Wang, X., Xu, X., Xu, Y. & Yang, L. L. Top quark pair production near threshold: single/double distributions and mass determination. *JHEP* **06**, 158 (2020).
176. *Observation of a cross-section enhancement near the $t\bar{t}$ production threshold in $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector* tech. rep. <https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2025-008> (CERN, Geneva, 2025).
177. Fuks, B., Hagiwara, K., Ma, K. & Zheng, Y.-J. Simulating toponium formation signals at the LHC. *Eur. Phys. J. C* **85**, 157 (2025).
178. *Procedure for the LHC Higgs boson search combination in Summer 2011* tech. rep. (CERN, Geneva, 2011).
179. Sirunyan, A. M. *et al.* Search for heavy Higgs bosons decaying to a top quark pair in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP* **04**. [Erratum: *JHEP* **03**, 187 (2022)], 171 (2020).
180. Dvoretzky, A., Kiefer, J. & Wolfowitz, J. Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator. *The Annals of Mathematical Statistics* **27**, 642 (1956).
181. Massart, P. The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *The Annals of Probability* **18** (1990).
182. Degrande, C., Duhr, C., Fuks, B., Grellscheid, D., Mattelaer, O. & Reiter, T. UFO - The Universal FeynRules Output. *Comput. Phys. Commun.* **183**, 1201 (2012).
183. Brehmer, J., Cranmer, K., Louppe, G. & Pavez, J. A Guide to Constraining Effective Field Theories with Machine Learning. *Phys. Rev. D* **98**, 052004 (2018).
184. Alwall, J., Herquet, M., Maltoni, F., Mattelaer, O. & Stelzer, T. MadGraph 5 : Going Beyond. *JHEP* **06**, 128 (2011).
185. Ball, R. D., Bertone, V., Carrazza, S., Del Debbio, L., Forte, S., Guffanti, A., Hartland, N. P. & Rojo, J. Parton distributions with QED corrections. *Nucl. Phys. B* **877**, 290 (2013).
186. Bierlich, C., Chakraborty, S., Desai, N., Gellersen, L., Helenius, I., Ilten, P., Lönnblad, L., Mrenna, S., Prestel, S., Preuss, C. T., Sjöstrand, T., Skands, P., Uthmeim, M. & Verheyen, R. *A comprehensive guide to the physics and usage of PYTHIA 8.3* 2022.
187. Bahl, H., Bresó-Pla, V., De Crescenzo, G. & Plehn, T. Advancing tools for simulation-based inference. *SciPost Phys. Core* **8**, 060 (2025).
188. Cluster, D. M. *Maxwell Documentation* <https://docs.desy.de/maxwell/>. [Online; accessed 17.09.2025].
189. Defazio, A., Yang, X. A., Mehta, H., Mishchenko, K., Khaled, A. & Cutkosky, A. *The Road Less Scheduled* 2024.
190. Artoisenet, P., Lemaitre, V., Maltoni, F. & Mattelaer, O. Automation of the matrix element reweighting method. *JHEP* **12**, 068 (2010).
191. Chekhovsky, V. *et al.* Search for dark matter produced in association with one or two top quarks in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP* **08**, 085 (2025).
192. Catani, S. & Seymour, M. A general algorithm for calculating jet cross sections in NLO QCD. *Nuclear Physics B* **485**, 291 (1997).
193. Deng, L. The MNIST Database of Handwritten Digit Images for Machine Learning Research. *IEEE Signal Processing Magazine* **29**, 141 (2012).
194. Hasanpour, S. H., Rouhani, M., Fayyaz, M. & Sabokrou, M. Lets keep it simple, using simple architectures to outperform deeper and more complex architectures. *arXiv preprint arXiv:1608.06037* (2016).
195. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. & Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization in *Proceedings of the IEEE International Conference on Computer Vision* (2017), 618.
196. CMS Collaboration. Reweighting simulated events using machine-learning techniques in the CMS experiment (2024).

197. Andersen, J., Gütschow, C., Maier, A. & Prestel, S. A Positive Resampler for Monte Carlo Events with Negative Weights. *The European Physical Journal C* **80** (2020).
198. Nachmann, B. & Thaler, J. Neural Resampler for Monte Carlo Reweighting with Preserved Uncertainties. *Physical Review D* **102**, 076004 (2020).
199. Andersen, J. R., Cueto, A., Jones, S. P. & Maier, A. A Cell Resampler study of Negative Weights in Multi-jet Merged Samples. *CERN-TH-2024-200, IPPP/24/73* (2024).
200. Bridle, J. in *Neurocomputing: Algorithms, Architectures and Applications* 227 (Springer, Berlin, Heidelberg, 1989).
201. Bach, J., Grohsjean, A., Schwanenberger, C. & Stelldinger, P. *Unlearning with Partial Label Learning* in (Association for Computing Machinery, PETRA '25: Proceedings of the 18th ACM International Conference on Pervasive Technologies Related to Assistive Environments, 2025).
202. Nachman, B. & Thaler, J. Neural resampler for Monte Carlo reweighting with preserved uncertainties. *Physical Review D* **102** (2020).
203. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)* (2015).
204. Artoisenet, P. *et al.* A framework for Higgs characterisation. *JHEP* **11**, 043 (2013).
205. Hayrapetyan, A. *et al.* The CMS Statistical Analysis and Combination Tool: Combine. *Comput. Softw. Big Sci.* **8**, 19 (2024).
206. CMS-Collaboration. *CMS Combine Documentation* <https://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/v10.3.3/part2/physicsmodels/?h=interf#multisignalmodel-ready-made-model-for-multiple-signal-processes>. [Online; accessed 21.11.2025].
207. Drnevich, M., Jiggins, S., Katzy, J. & Cranmer, K. Neural quasiprobabilistic likelihood ratio estimation with negatively weighted data. *Mach. Learn. Sci. Tech.* **6**, 045023 (2025).
208. Warnecke, A., Pirch, L., Wressnegger, C. & Rieck, K. *Machine Unlearning of Features and Labels in Network and Distributed System Security (NDSS) Symposium* (2021).
209. Cao, Y. & Yang, J. Towards Making Systems Forget with Machine Unlearning in *2015 IEEE Symposium on Security and Privacy* (2015), 463.
210. Xu, H., Zhu, T., Zhang, L., Zhou, W. & Yu, P. S. Machine Unlearning: A Survey. *ACM Computing Surveys* **56**, Article 9, 1 (2023).
211. Hoofnagle, C., van der Sloot, B. & Zuiderveen Borgesius, F. The European Union General Data Protection Regulation: What It Is and What It Means. *Information & Communications Technology Law* **28**, 65 (2019).
212. Cadet, X. F., Borovykh, A., Malekzadeh, M., Ahmadi-Abhari, S. & Haddadi, H. *Deep Unlearn: Benchmarking Machine Unlearning for Image Classification* IEEE, 2025.
213. Wang, W., Tian, Z., Zhang, C. & Yu, S. *Machine Unlearning: A Comprehensive Survey* 2024.
214. Schelter, S. "Amnesia" - Towards Machine Learning Models that Can Forget User Data Very Fast in *10th Conference on Innovative Data Systems Research, CIDR* (2020).
215. Shah, V., Träuble, F., Malik, A., Larochelle, H., Mozer, M., Arora, S., Bengio, Y. & Goyal, A. *Unlearning via Sparse Representations* 2023.
216. Chen, M., Gao, W., Liu, G., Peng, K. & Wang, C. *Boundary Unlearning: Rapid Forgetting of Deep Networks via Shifting the Decision Boundary* in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2023), 7766.
217. Kurmanji, M., Triantafillou, P., Hayes, J. & Triantafillou, E. *Towards Unbounded Machine Unlearning* in *Proceedings of the 37th International Conference on Neural Information Processing Systems* (2023), 1957.
218. Graves, L., Nagisetty, V. & Ganesh, V. *Amnesiac Machine Learning* in *35th AAAI Conference on Artificial Intelligence, 33rd Conference on Innovative Applications of Artificial Intelligence, 11th Symposium on Educational Advances in Artificial Intelligence* (2021), 11516.
219. Bourtoutle, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D. & Papernot, N. *Machine Unlearning* in *42nd IEEE Symposium on Security and Privacy* (2021), 141.
220. Guo, C., Goldstein, T., Hannun, A. & van der Maaten, L. *Certified Data Removal from Machine Learning Models* in *37th International Conference on Machine Learning (Proceedings of Machine Learning Research)* **119** (PMLR, 2021), 3832.
221. Schelter, S., Grafberger, S. & Dunning, T. *HedgeCut: Maintaining Randomised Trees for Low-Latency Machine Unlearning* in *International Conference on Management of Data (ACM, 2021)*, 1545.
222. Wang, J., Guo, S., Xie, X. & Qi, H. *Federated Unlearning via Class-Discriminative Pruning* in *ACM Web Conference (ACM, 2022)*, 622.

- 223. Golatkar, A., Achille, A. & Soatto, S. *Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks* in *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Computer Vision Foundation/IEEE, 2020), 9301.
- 224. Krizhevsky, A. *Learning Multiple Layers of Features from Tiny Images* 2009.
- 225. Shokri, R., Stronati, M., Song, C. & Shmatikov, V. Membership Inference Attacks Against Machine Learning Models. *2017 IEEE Symposium on Security and Privacy (SP)*, 3 (2017).
- 226. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv e-prints*, arXiv:2010.11929 (2020).

ACKNOWLEDGEMENTS

Space echoes like an immense tomb, yet the stars still burn. Why does the sun take so long to die ?

— Nick Land

Working on humanity's bleeding edge of knowledge is a gigantic privilege that I acknowledge as extraordinary. The people that made this possible and who were by my side I owe my deepest thankfulness.

I would like to thank my supervisors Alexander, Christian and Peer. I am very grateful for valuable insights into academia, the fields of particle physics and computer science as well as a lot of personal guidance in handling the ups and downs in a PhD project. You made this interesting interdisciplinary project come alive by stepping outside of your domain comfort zones and you made the effort to understand the other discipline.

Additionally I want to thank the CMS group of DESY, first and foremost the wonderful Exotica(+friends) group. Thank you Laurids, Afq, Dominic, Mads, Sam, Andre and Jonas for crazy "sidequest", stellar support with technical problems and understanding particle physics. A/H o'clock sessions might've been lengthy and challenging but it also felt very special that I was allowed to be part of this outstanding team.

I also want to thank the HAW ML AG for an interesting insight into another domain. The topics might've been apples, trash, or stock price prediction, but they were always interesting ML challenges.

Another thank you is owed to the DASHH team, especially Christiane, Heike and my Panel Chair Prof. Matthias Rarey, who enabled this work to be truly interdisciplinary.

I also want to thank my flatmates over the years, David and later Mathis, who also happen to be particle physicists. You provided some grounding, some hype or just some physics chatter. This personal thanks expands to all my friends from my physics studies, from my Abitur time and from Handball - cheers to Teutonia Altona. Thank you to Malte and Buschi who are always there and always have been there.

And of course a special thank you to my parents, my brother, my grandparents and Alina. I tried and tested you with terrible work-life balance decisions and rambling about NNLO QCD but you always supported me unconditionally.

CURRICULUM VITAE

PERSONAL DATA

Name	Jörn Bach
Date of Birth	August 21, 1998
Place of Birth	Darmstadt, Germany
Citizen of	Germany

EDUCATION

2022 – Now	Universität Hamburg Hamburg, Germany <i>Target degree:</i> Doctorate in Physics
2019 – 2022	Universität Hamburg Hamburg, Germany <i>Final degree:</i> Master of Science Physik
2016 – 2019	Universität Hamburg Hamburg, Germany <i>Final degree:</i> Bachelor of Science Physik
2008 – 2016	Gymnasium am Mühlenberg, Bad Schwartau, Germany <i>Final degree:</i> Abitur

EMPLOYMENT

July 2022 – December 2025	PhD Researcher <i>Deutsches Elektronen-Synchrotron DESY,</i> Hamburg, Germany
------------------------------	---

PUBLICATIONS

Articles in peer-reviewed journals:

1. Hayrapetyan, A. *et al.* Search for heavy pseudoscalar and scalar bosons decaying to a top quark pair in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Rept. Prog. Phys.* **88**, 127801 (2025).
2. Hayrapetyan, A. *et al.* Observation of a pseudoscalar excess at the top quark pair production threshold. *Rept. Prog. Phys.* **88**, 087801 (2025).

Conference contributions:

3. Bach, J., Grohsjean, A., Schwanenberger, C. & Stelldinger, P. *Negative event weights in Deep Neural Network training for heavy Higgs searches* in 86th Annual Conference of the DPG and DPG Spring Meeting 2023 (Deutsche Physikalische Gesellschaft, Dresden, Germany).
4. Bach, J., Grohsjean, A., Schwanenberger, C. & Stelldinger, P. *Dealing with negatively weighted Events in DNN-based LHC Analyses* in 87th Annual Conference of the DPG and DPG Spring Meeting 2024 (Deutsche Physikalische Gesellschaft, Karlsruhe, Germany).
5. Bach, J., Grohsjean, A., Schwanenberger, C., Jeppe, L., Baxter, S. & Anuar, A. *Search for heavy Higgs bosons and Axion-Like Particles in the $t\bar{t}$ final state at CMS* in 87th Annual Conference of the DPG and DPG Spring Meeting 2024 (Deutsche Physikalische Gesellschaft, Karlsruhe, Germany).
6. Bach, J., Grohsjean, A., Schwanenberger, C. & Stelldinger, P. *Unlearning with Partial Label Learning* in (Association for Computing Machinery, PETRA '25: Proceedings of the 18th ACM International Conference on Pervasive Technologies Related to Assistive Environments, 2025).
7. Bach, J. *CMS Results at the $t\bar{t}$ threshold* in 13th Large Hadron Collider Physics Conference (2025).