

AIX-MARSEILLE UNIVERSITÉ

Faculté Des Sciences de Luminy

Ecole Doctorale 352: Physique et Sciences de la Matière

Ecole Doctorale 184: Mathématiques et Informatique

Centre de Physique des Particules de Marseille

Laboratoire d'Informatique et Systèmes

Thèse présentée pour obtenir le grade universitaire de docteur

Discipline: Physique et Sciences de la Matière

Spécialité: Physiques des Particules et Astroparticules

Ziyu GUO

Search for the Higgs boson in the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) channel in the ATLAS experiment at the LHC using machine learning methods and synchronization of the ITk geometry description for simulation and radiation studies for the HL-LHC ATLAS upgrade

Soutenue le 05/11/2019 devant le jury composé de:

Anne-Catherine Le Bihan	IPHC, Strasbourg	Rapporteur
David Rousseau	LAL, Orsay	Rapporteur
Mossadek Talby	CPPM	Examineur
Elisa Fromont	IRISA, Université Rennes 1	Examineur
Cristinel Diaconu	CPPM	Directeur de thèse
Thierry Artières	LIS & Ecole Centrale Marseille	Co-directeur de thèse
Yann Coadou	CPPM	Co-encadrant, invité

Numéro national de thèse/suffixe local: 2019AIXM0350/056ED352

Résumé

Mots-clés : LHC, ATLAS, boson de Higgs, $t\bar{t}H$, $H \rightarrow b\bar{b}$, apprentissage automatique, arbres de décision boostés, apprentissage profond, trajectographe interne

Dans le Modèle Standard (MS) les particules élémentaires obtiennent une masse par couplage au champ de Higgs. Ce mécanisme prédit l'existence du boson de Higgs, découvert en 2012 par les collaborations ATLAS et CMS. Les couplages entre le Higgs et les bosons vecteurs sont depuis bien établis, alors que les mesures des couplages entre le Higgs et les fermions (couplages de Yukawa) sont encore incomplètes. Le couplage entre le Higgs et le quark top est le plus intéressant car proche de un et beaucoup plus fort que celui des autres quarks. Tout écart entre sa mesure et la prédiction théorique du MS serait un signe de nouvelle physique. La production associée d'un boson de Higgs et d'une paire de quarks top ($t\bar{t}H$) permet une mesure directe de ce couplage de Yukawa.

Cette thèse documente une recherche d'événements $t\bar{t}H$ où le boson de Higgs se désintègre en une paire de quarks beaux ($H \rightarrow b\bar{b}$) dans les données collectées avec le détecteur ATLAS auprès du LHC durant le Run 2, avec des collisions de protons à une énergie dans le centre de masse de 13 TeV. Les nombreux jets et b -jets dans l'état final sont un défi pour cette analyse, qui nécessite l'utilisation de techniques avancées. Les grandes incertitudes théoriques sur les bruits de fond $t\bar{t}$ +jets ont un impact majeur sur la sensibilité.

Impliquée dans l'analyse des données collectées en 2015 et 2016, avec une luminosité intégrée de $36,1 \text{ fb}^{-1}$, l'auteure a étudié et optimisé des arbres de décision boostés (BDT), pour résoudre l'association entre les jets et les partons dans la reconstruction du signal $t\bar{t}H$, puis pour distinguer les événements du signal de ceux de $t\bar{t}$ +jets. Le BDT de classification est une des variables discriminantes utilisées dans l'ajustement aux données du modèle utilisant les échantillons simulés basés sur les prédictions du MS. Dans l'hypothèse de la seule présence du bruit de fond, la signification observée (attendue) de la présence du signal $t\bar{t}H$ en plus du bruit de fond du MS est de 1,4 (1,6) déviations standard, résultat consistant à la fois avec l'hypothèse du seul bruit de fond ou de la présence du signal $t\bar{t}H$.

Afin de proposer de nouvelles approches pour l'analyse de l'ensemble des données du Run 2, différentes techniques d'apprentissage profond ont été explorées. L'utilisation de réseaux de neurones récurrents permet de répondre, en une seule passe, à la double problématique de la reconstruction et de la classification. Cette méthode est d'abord testée à partir du même type de variables discriminantes, motivées par la physique et calculées explicitement à partir des propriétés cinématiques des objets de l'état final. Afin d'utiliser directement ces propriétés de base, un réseau de neurones profond est proposé, en y intégrant des connaissances sur la physique pour permettre au modèle de déterminer automatiquement des caractéristiques discriminantes. Pour améliorer la sensibilité tout en contraignant mieux les bruits de fond,

une nouvelle façon de catégoriser les événements en régions enrichies en $t\bar{t}H$ et en différentes sous-catégories de $t\bar{t}$ est implémentée, à l'aide d'un réseau de neurones profond à sorties multiples, et comparée au schéma manuel de partitionnement utilisé précédemment. Pour finir, des réseaux antagonistes ont été étudiés pour réduire la dépendance aux incertitudes de modélisation de $t\bar{t}$.

Le High-Luminosity LHC (HL-LHC), version améliorée de l'actuel LHC qui collectera dix fois plus de luminosité que lui, devrait entrer en opération au milieu des années 2020 avec un potentiel accru de découverte de nouvelle physique. Pour tenir le choc de cette nouvelle phase, le détecteur ATLAS sera mis à niveau, avec un nouveau trajectographe (ITk) dont la géométrie est en cours d'optimisation. Les descriptions de la géométrie utilisées dans la simulation du détecteur et les études de radiation sont implémentées de façon indépendante. L'auteure a contribué à la synchronisation des deux géométries, étape importante pour valider les estimations de radiation.

Abstract

Keywords: LHC, ATLAS, Higgs boson, $t\bar{t}H$, $H \rightarrow b\bar{b}$, machine learning, boosted decision trees, deep learning, inner tracker

According to the Standard Model (SM), elementary particles obtain their mass by coupling with the Higgs field. This mechanism predicts the existence of the Higgs boson, which was discovered by the ATLAS and CMS collaborations in July 2012. The couplings between the Higgs and vector bosons are completely established, while the measurements of the strength of the Higgs-fermion couplings, the so-called Yukawa couplings, are incomplete. Being much stronger than the ones for the other quarks and close to unity, the Higgs-top quark Yukawa coupling is the most interesting one. It can be a hint for new physics if any measurement shows deviation from the theory. The associated production of a Higgs boson with a top quark pair ($t\bar{t}H$) allows for a direct measurement of the top Yukawa coupling.

This thesis presents a search for $t\bar{t}H$ events with the Higgs boson decaying into a bottom quark pair ($H \rightarrow b\bar{b}$) based on data collected with the ATLAS detector at the LHC during Run 2 operation, with proton pair collisions at a center of mass energy of 13 TeV. The multiple jets and b -jets in the final state make the analysis challenging and heavily relying on advanced analysis techniques. The large modelling uncertainties of the $t\bar{t}$ + jets backgrounds are a driving factor of the sensitivity.

In the analysis round using datasets collected in years 2015 and 2016 for an integrated luminosity of 36.1 fb^{-1} , the author studies and optimizes boosted decision trees (BDT) to firstly solve the jet-parton assignment in the reconstruction of the $t\bar{t}H$ signal, and, in a second step, separate the signal from the $t\bar{t}$ + jets backgrounds. The classification BDT output is one of the discriminants to fit experimental data with simulated samples based on the SM expectation. Through this process, the existence of the $t\bar{t}H$ signal is tested. Under the background only hypothesis, the observed (expected) significance of the $t\bar{t}H$ signal over the expected SM background is 1.4 (1.6) standard deviations, which is consistent with both SM background-only and $t\bar{t}H$ prediction.

Targeting a contribution to the new analysis round using the full Run 2 data, deep learning techniques are also explored. Recurrent neural networks are exploited, aiming to, in one step, better solve the event reconstruction and classification. This method still uses as input the physics motivated features manually calculated from the basic kinematics of the final-state objects. To use these raw features as input directly, a deep neural network model is designed while incorporating the physics domain knowledge to automatically extract more discriminating features. To improve the signal sensitivity while constraining backgrounds, an event categorization, having separately the $t\bar{t}H$ and different $t\bar{t}$ components enriched regions, is implemented using a deep neural network based multi-classifier, which is compared with the manual splitting in the previous analysis. Finally adversarial networks are studied in order to

decrease the $t\bar{t}$ modelling uncertainty.

The High-Luminosity Large Hadron Collider (HL-LHC), the upgrade of the current LHC, with an enhanced luminosity by a factor of ten beyond the LHC's design value, is expected to start operation in the mid-2020's with increased potential for physics discovery. To cope with this new phase, the ATLAS detector will be upgraded, including a new inner tracker (ITk) which geometry is under design. For the detector simulation and radiation study, the geometry descriptions are independently implemented. The author contributes to the synchronization of the two geometries, which is important to validate a proper radiation estimation.

Acknowledgements

Firstly, I would like to express sincere gratitude to my three supervisors. Throughout my Ph.D research and thesis writing, most guidance and help was from Dr. Yann COADOU who gave me continuous and solid support. He taught me immense knowledge of physics, machine learning, and English as well. More importantly, he really guided me to find and solve problems independently in research, but was always willing to provide patient guidance and work together with me when I was stuck. Without his precious support, I could not have gotten such a great opportunity to conduct and finalize this Ph.D study. On a personal level, he also inspires me how to be a great researcher, by his self-discipline, hardworking, and scientific integrity. Prof. Thierry Artières, an expert in machine learning, gave me detailed guidance in research when it is related with deep learning knowledge and tools, which really helped me entering this field very fast. I was enlightened by discussions with him, by re-considering the physics problems from a completely different point of view. As a non-physicist, he has a modest and curious communication style concerning physics problems. This inspires me to be open-minded towards different fields and to interact with people from different groups. Thanks to Dr. Cristinel Diaconu, my two supervisors mentioned above were introduced to each other and finally launched this inter-disciplinary project.

I would also like to thank my colleagues and friends from both particle physics and machine learning laboratories: Nguyen Hoang Dai Nghia, Zongde Chen, Dr. Nihal Brahimi and Dr. Grigore Tarna from CPPM, Tom Dagens, Luc Giffon and Dr. Riikka Huusari from LIS. They gave me a lot of help on physics and machine learning problems, respectively. Outside of laboratories, we also have lots of happy times.

Contents

Résumé	3
Abstract	5
Acknowledgements	6
1 Theoretical introduction	12
1.1 The Standard Model of particle physics	12
1.1.1 Elementary particles in the Standard Model	12
Fermions	13
Bosons	13
1.1.2 Electrodynamic interaction	14
1.1.3 Electroweak interaction	15
1.1.3.1 Weak interaction	15
1.1.3.2 The electroweak unification	15
1.1.4 The Higgs mechanism	16
1.1.5 Strong interaction	18
Confinement	19
Asymptotic freedom	19
1.2 Higgs physics	20
1.2.1 Higgs production and decay	20
1.2.2 Higgs discovery	23
1.2.3 Higgs coupling measurement	23
2 The ATLAS experiment	26
2.1 The Large Hadron Collider	26
2.1.1 The accelerator system	26
2.1.1.1 The LHC ring setup	28
2.1.1.2 Luminosity	28
2.1.1.3 Pile-up	29
2.1.2 The physics experiments on the LHC	29
2.2 The ATLAS experiment	30
2.2.1 Coordinate system	31
2.2.2 Inner detector	32
2.2.2.1 Pixel detector	35
2.2.2.2 The semi-conductor tracker	35
2.2.2.3 Transition radiation tracker	36

2.2.3	Calorimeters	37
2.2.3.1	Electromagnetic calorimeter	37
2.2.3.2	Hadronic calorimeter	39
2.2.4	Muon spectrometer	39
2.2.5	Trigger and data acquisition system	42
2.3	Monte Carlo simulated data	43
	Event generation	43
	Detector simulation	44
2.4	Object reconstruction and identification	44
2.4.1	Tracks and primary vertex	44
2.4.1.1	Tracking algorithm	44
2.4.1.2	Vertexing algorithm	45
2.4.1.3	Track parameters	45
2.4.2	Muons	46
	Muon identification	47
	Muon isolation	47
2.4.3	Electrons	47
	Identification	48
	Isolation	49
2.4.4	Jets	49
	Jet energy calibration	49
	b -tagging	50
2.4.5	Missing transverse energy	52
3	Synchronizing the ITk geometry description for simulation and radiation study	53
3.1	The High Luminosity LHC	53
3.2	The ATLAS inner tracking system	54
3.2.1	Pixel Module	55
3.2.2	Local support structure, services and cooling system	55
3.2.3	Global support structure, moderator, and beam-pipe	58
3.3	Simulation of the ITk detector	60
3.3.1	ITk geometry in Geant4 for simulation	60
3.3.2	ITk geometry in fluence and radiation dose studies	61
3.4	Methods to synchronize ITk layout descriptions	63
3.5	Layout description synchronization	66
3.5.1	Step 1.9	66
3.5.1.1	Geometry simplification	66
3.5.1.2	Geometry comparison and synchronization	67
3.5.2	Step 2.2	72
3.5.2.1	Differences with step 1.9	72
3.5.2.2	Simplified version	72
3.5.2.3	Detailed version	75
3.6	Summary	83
4	Search for the Higgs boson in the single lepton $t\bar{t}H$ ($H \rightarrow b\bar{b}$) channel	84
4.1	Introduction	84
4.1.1	Motivation for the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis	84
4.1.2	Overview of the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis	85

	Previous analysis rounds of $t\bar{t}H$ ($H \rightarrow b\bar{b}$)	85
	Analysis channels	86
	Contributions	86
	Organization	86
4.2	Object and event selection	86
4.2.1	Object selection	87
4.2.2	Event selection	88
4.3	Simulation samples	88
4.3.1	Signal modeling	89
4.3.2	$t\bar{t}$ + jets background	89
4.3.3	Other backgrounds	89
4.4	Analysis strategy	90
4.4.1	Categorization	90
4.4.2	Multivariate analysis techniques	92
4.5	Using boosted decision trees	94
4.5.1	Boosted decision trees	94
	Adaboost	95
	Gradboost	96
4.5.2	BDT performance measure	96
4.5.3	BDT-based event reconstruction	97
4.5.3.1	Truth matching	97
4.5.3.2	Reconstruction algorithm	98
4.5.3.3	BDT technique to resolve combinatorics	99
	Using b -tagging weights	103
	BDT parameter optimization	105
4.5.3.4	Performance	107
	Training regions	107
	Reconstruction performance	107
4.5.4	Discrimination between signal and background	109
4.5.4.1	Discriminating variables	110
4.5.4.2	Using b -tagging as BDT input features	112
4.5.4.3	Reducing the BDT input variable list	120
4.6	Systematic uncertainties	126
4.6.1	$t\bar{t}$ + jets uncertainties	126
4.6.1.1	Uncertainties in the $t\bar{t} + \geq 1b$ modeling	127
4.6.1.2	Uncertainties in the $t\bar{t} + \geq 1c$ modeling	128
4.6.2	Non- $t\bar{t}$ uncertainties	129
4.6.2.1	$t\bar{t}H$ modeling uncertainties	130
4.6.2.2	Non- $t\bar{t}$ background	130
4.6.2.3	Experimental uncertainties	130
4.7	Statistical analysis	131
4.7.1	Profile likelihood fit	131
4.7.2	Fitted discriminant variables	132
4.8	Results	135
4.8.1	BDT classifier choice	135
4.8.2	Fit results	137
4.9	Summary	143

5	Deep learning methods in the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis	144
5.1	Particle physics analysis for non-physicists	145
5.1.1	Experiments and data	145
5.1.2	Signal and background	145
5.1.3	Simulated data	146
5.1.4	The default $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis pipeline	147
5.2	Introduction to deep neural networks	148
5.2.1	Deep neural network components	148
	Training and testing samples	150
	Loss function	150
	Optimizer	150
	Activation function	151
	Data pre-processing: centralization and normalization	151
	Batch normalization	152
	Regularizer	152
	Early stopping	152
5.2.2	Recurrent neural networks	153
	Bidirectional RNN	154
	Vanilla RNN and long-term dependence issues	154
	Long-short term memory RNN, and Gated Recurrent Unit	155
5.2.3	Hyper-parameter optimization	156
	Grid search	156
	Tree-structured Parzen Estimators	157
5.2.4	Tools and software	157
5.3	Using recurrent neural networks to classify $t\bar{t}H$ signal and $t\bar{t}$ background	158
5.3.1	RNN in $t\bar{t}H$ ($H \rightarrow b\bar{b}$)	158
	5.3.1.1 Number of combinations	159
	5.3.1.2 Adding combination-independent and b -tagging variables	160
	5.3.1.3 Sequence order	161
	5.3.2 RNN performance as a function of training sample size	162
	5.3.3 Hyper-parameter grid search	162
	5.3.4 Summary	164
5.4	Event classification using RNN-based multi-classification	166
	5.4.1 Model architecture and training setups	166
	5.4.2 Performance	167
	5.4.3 Event categorization	169
	5.4.4 Finer classification of the $t\bar{t} + \geq 1b$ background	170
	5.4.5 Summary	171
5.5	Low-level input features using deep neural networks	173
	5.5.1 Using low-level features with classical networks	174
	DNN + low level features	174
	RNN + low level features	175
	5.5.2 Parse tree model: encoding domain knowledge into neural networks	176
	5.5.3 Using single tree to classify $t\bar{t}H$ and $t\bar{t}$ with one combination	178
	5.5.4 Using single tree embedding to classify $t\bar{t}H$ and $t\bar{t}$ with all combinations	180
	Model architecture	180
	DNN embedding	181

Performance	181
5.5.5 Two-tree model	182
5.5.5.1 Signal tree vs. background tree	182
5.5.5.2 Combination of two trees by brute force	183
5.5.5.3 Siamese training	185
Two-phase training	185
Three-phase training	187
5.5.6 Summary	189
5.6 Reducing systematic uncertainties using adversarial training	190
5.6.1 Differences between nominal and systematic samples	190
DNN classification of the two samples	190
Classification performance with systematic samples	190
5.6.1.1 Metric to compare models: AMS	193
5.6.2 Adversarial networks	194
Generative Adversarial Networks	194
Constraining the classifier with adversarial training	195
5.6.3 Adversarial trainings	197
5.6.3.1 Validation of implementation	197
5.6.3.2 Adversarial training on $t\bar{t}H$ ($H \rightarrow b\bar{b}$)	199
5.6.4 Summary	202
5.7 Summary	202
Conclusion	205
Bibliography	206
Appendix	217
Glossary	222

Theoretical introduction 1

1.1 The Standard Model of particle physics

The *Standard Model* (SM) of particle physics is a theory that classifies elementary particles of the universe and describes three kinds of basic interactions between them: the electromagnetic, weak, and strong interactions. This will be illustrated in Section 1.1.1. The gravitational force, the fourth basic interaction, is so weak at the sub-atomic scale compared with the other three forces that it can safely be neglected. Developed as a classical theory, it is still being searched to be quantized and encapsulated in the SM which uses the physics language of *quantum field theory* (QFT) that combines special relativity and quantum mechanics. QFT explains the evolution of the system at relativistic energy scale involving particle creation and annihilation.

The establishment of the SM involved a lot of work of both experimental and theoretical physicists in the second half of the 20th century. One of its theoretical cores is the Yang-Mills theory [1] seeking for the underlying symmetry, the gauge group, to describe the particles behavior. It extends the $U(1)$ Abelian gauge group in the electromagnetic interaction (Section 1.1.2) into the non-Abelian gauge group providing ideas for other interactions. The gauge group of the SM is $SU(3)_C \otimes SU(2)_L \times U(1)_Y$. The $SU(2)_L \otimes U(1)_Y$ is the gauge group of unified electromagnetic and weak interactions, where L refers to left-handed weak isospin, and Y to weak hypercharge, described in Section 1.1.3. The $SU(3)_C$ is the gauge group of the strong interaction, where C denotes color, a kind of charge described in Section 1.1.5. These groups reflect the internal symmetries of the particle system in addition to its space-time symmetry.

Starting from symmetries, the SM can predict the existence of new particles, while encapsulating particles and interactions discovered by the time of its development. As for its predictions, several particles were discovered and show consistent properties, e.g., the W and Z bosons [2] in 1985, and the top quark [3, 4] in 1995. Then in 2012, the Higgs boson, as described in Sections 1.1.4 and 1.2, was discovered [5, 6], which is the last piece of this theory. The SM is, however, not complete, and further studies of the Higgs boson properties are thus necessary as a way to test new physics.

1.1.1 Elementary particles in the Standard Model

The elementary particles included in the SM with their properties are illustrated in Figure 1.1. There are two types of particles: fermions with half integer spin and following Pauli-Dirac statistics, and bosons with integer spin and following Bose-Einstein statistics. Fermions are building blocks of matter, and their interactions are mediated by gauge bosons. The Higgs boson is related with the Higgs field through the interaction with which elementary particles acquire their mass.

Fermions Fermions consist of two types, leptons (l) and quarks (q). Both are grouped into three generations: (e, ν_e) , (μ, ν_μ) , (τ, ν_τ) for leptons; (u, d) , (c, s) , (t, b) for quarks. The first generation is the lightest and most stable while the third is the heaviest and least stable. The six leptons have three flavors, e , μ and τ each with electric charge of -1 and with their corresponding neutrino which carries no electric charge. The up-type quarks u , c , t have electric charge $+2/3$ and down-type ones $-1/3$. Each lepton and quark have their corresponding anti-particles that have exactly the same properties as particles, except with opposite charge.

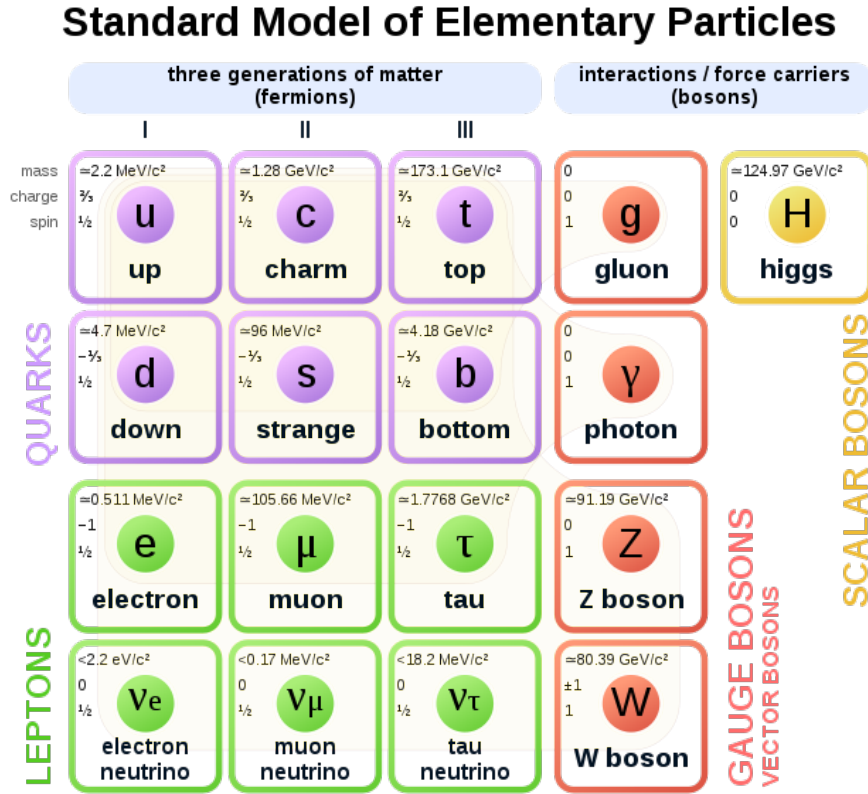


Figure 1.1: Elementary particles included in the Standard Model.

Bosons Fundamental interactions between particles with corresponding charges are carried by different gauge bosons. Massless and electrically neutral photons (γ) mediate electromagnetic interactions between charged particles. The eight massless gluons take color and anti-color, and mediate the strong interaction. Massive gauge bosons include W^+ and W^- each carrying electric charge $+1$ and -1 , respectively, and Z being electrically neutral. They carry the weak interaction. All these bosons are also referred to as vector bosons with spin equal to 1. The Higgs boson predicted by the Higgs mechanism has spin 0. Its mass is not predicted by the SM and the combined Run 1 measurement by ATLAS and CMS is $125.09 \pm 0.24 \text{ GeV}$ [7].

The three fundamental interactions and related particles are introduced in following sections using the language of quantum field theory. The *field* is the function of space-time coordinates which is built in a 4-dimensional manifold with Minkowski metric. In QFT, the field is quantized, and the particle is the excited state of the field. The interaction between particles is derived

from the interaction between various fields. The law of motion of the field is determined by the *principle of least action*. The action of the particle system is defined as $S = \int \mathcal{L}(x)dx$, where $\mathcal{L}(x)$ is the Lagrangian expressed in terms of fields.

1.1.2 Electrodynamic interaction

This section uses the electromagnetic interaction described by quantum electrodynamics (QED) as an example to show the principle of gauge theory, and coupling between particles [8].

The dynamics of a Dirac fermionic field that does not participate in any interaction can be described as:

$$\mathcal{L}_D = \bar{\psi}(i\cancel{D} - m)\psi \quad \text{with} \quad \cancel{D} = \gamma^\mu \partial_\mu, \quad (1.1)$$

where γ^μ are the Dirac 4×4 constant matrices satisfying $\{\gamma^\mu, \gamma^\nu\} = 2g^{\mu\nu}$, μ or ν are index of space-time coordinate x , and $\bar{\psi} \equiv \psi^\dagger \gamma^0$. The Lagrangian keeps invariant under the $U(1)_{em}$ global¹ transform on ψ : $\psi(x) \rightarrow \psi'(x) = e^{-iq\theta}\psi(x)$.

The above phase θ can be a function of space-time coordinates, $\psi(x) \rightarrow \psi'(x) = e^{-iq\theta(x)}\psi(x)$. This transform is local, and is referred to as the $U(1)_{em}$ gauge transform. To make the free Dirac fermion field Lagrangian invariant under the $U(1)_{em}$ local transform, a new gauge field needs to be introduced, $A_\mu(x)$, transformed as:

$$A_\mu(x) \rightarrow A'_\mu(x) = A_\mu(x) + \partial_\mu \theta(x). \quad (1.2)$$

This cancels the space-time derivative terms of $\theta(x)$.

As long as one replaces the derivative ∂_μ with $D_\mu \equiv \partial_\mu + iqA_\mu$, the Lagrangian form can stay consistent with the free fermion one (Equation 1.1):

$$\begin{aligned} \mathcal{L}'_D &= \bar{\psi}(i\gamma^\mu D_\mu - m)\psi \\ &= \mathcal{L}_D - q\bar{\psi}\gamma^\mu\psi A_\mu. \end{aligned} \quad (1.3)$$

The second term $\mathcal{L}_{int} = -q\bar{\psi}\gamma^\mu\psi A_\mu$ shows the *coupling* between the fermion field $\psi(x)$ and the gauge field $A_\mu(x)$, meaning the interaction between the fermion and the photon. For electrons, $q = -e$, and $\|q\| = \|e\| \simeq 0.3 = \sqrt{4\pi/137}$ in natural units. It is the coupling constant between electrons or positrons and photons.

The full QED Lagrangian, which has $U(1)_{em}$ gauge symmetry, is:

$$\mathcal{L} = \mathcal{L}_D + \mathcal{L}_\gamma + \mathcal{L}_{int} = \bar{\psi}(i\cancel{D} - m)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} - q\bar{\psi}\gamma^\mu\psi A_\mu. \quad (1.4)$$

It contains the term of free photon $\mathcal{L}_\gamma$ and Dirac fermion Lagrangian $\mathcal{L}_D$. The former itself has the $U(1)_{em}$ gauge symmetry and $F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$ is the EM field tensor. These two terms represent the dynamics of the free fermion and photon.

The EM interaction can happen when scattering fermions. The asymptotically free in-coming and out-going particles are represented as $|i\rangle$ and $|f\rangle$ state in the Fock space of QFT. In QFT one can calculate the transition rate (probability per unit time) from in-state $|i\rangle$ to out-state $|f\rangle$. The transition rate calculation, the measurable quantities, requires the calculation of *matrix element* (ME), $\mathcal{M}_{fi} = \langle f|S|i\rangle$, the transition amplitude between the two states, where $S = U(+\infty, -\infty)$ is the scattering matrix, and $U(t, t_0)$ is the evolution operation calculating

¹Global means the phase θ is not a function of space-time coordinates

the state at any time t knowing the state at a given time t_0 . The ME is related with the coupling strength in this way:

- $S = U(+\infty, -\infty) = T e^{-i \int_{-\infty}^{+\infty} d^4x \mathcal{H}_{int}}$, where the Hamiltonian density $\mathcal{H}_{int}$ can be calculated from $\mathcal{L}_{int}$, and T is the time ordering operator. For QED,

$$\mathcal{H}_{int} = -\mathcal{L}_{int} = q \bar{\psi} \gamma^\mu \psi A_\mu. \quad (1.5)$$

- S can be calculated perturbatively and up to the desired order: $S = \sum_{n=0}^{\infty} S^{[n]}$, where the n^{th} order is:

$$S^{[n]} = \frac{(-i)^n}{n!} \int \cdots \int dx_1 \cdots dx_n T(\mathcal{H}_{int}(x_1) \cdots \mathcal{H}_{int}(x_n)). \quad (1.6)$$

Using Equation 1.5 in Equation 1.6, the n^{th} order contribution has a factor of $\|e\|^n$ in the amplitude.

To facilitate the ME calculation, R.P. Feynman introduced the Feynman diagrams as symbolic representations of particle interactions with corresponding calculation rules according to the diagram.

1.1.3 Electroweak interaction

1.1.3.1 Weak interaction

To explain the discovered so-called β decay, $n \rightarrow pe^- \bar{\nu}$ with the neutron decaying into a proton, and electron and an anti-neutrino², Fermi proposed four-fermion contact interaction in 1934 to describe this process. It gives a very small coupling constant³ and thus he named this new interaction as the weak interaction⁴. This theory could describe well the β -decay, until parity violation in weak interaction was discovered [10, 11]. To include parity violation, the $V - A$ current [12, 13] is proposed to modify Fermi's theory. But it was realized soon that the theory is an effective theory at low energy, and diverges at higher order calculation.

1.1.3.2 The electroweak unification

In 1961 the unification of the electromagnetic and weak interactions was achieved by Glashow, Salam, Ward and Weinberg into the so-called electroweak interaction [14–16]. It provides solutions to the previous weak theory, and gives the intrinsic symmetry $SU(2)_L \otimes U(1)_Y$ of the electroweak interaction.

Due to parity violation, the weak interaction does not act in the same way on left- and right-handed fermions, noted as ψ_L and ψ_R , respectively. Related to this, in each generation of fermions, the theory of weak interaction classifies fermions as left-handed doublets and right-handed singlets. For instance for the first generation, it is returned as:

$$\begin{pmatrix} \nu_L \\ e_L \end{pmatrix}, \quad e_R \quad \begin{pmatrix} u_L \\ d_L \end{pmatrix}, \quad u_R, \quad d_R.$$

²The neutrino was not discovered by then

³ $G_F = 1.6637 \times 10^{-5} \text{GeV}^{-2}$ [9].

⁴Actually, the coupling is weak at low energy. At high energy, the weak and EM interaction converge towards the same value, hence the unification.

$SU(2)_L$ is the symmetry group describing the weak interaction only acting on left-handed doublets, giving particles a new quantum number, the *weak isospin* $I = \frac{1}{2}$ and its projection $I_3 = \pm \frac{1}{2}$. The right handed singlet has $I = 0$. To recover QED, the hypercharge Y is introduced which corresponds to the $U(1)_Y = e^{-iY\theta(x)}$. It has the relation with electric charge $Q = I_3 + \frac{Y}{2}$.

The invariance under $SU(2)_L \otimes U(1)_Y$ is obtained by the definition of two covariant derivatives:

$$D_{\mu,L} = \left(\partial_\mu + ig \frac{\sigma_i}{2} W_\mu^i + ig' \frac{Y}{2} B_\mu \right), \quad D_{\mu,R} = \left(\partial_\mu + ig' \frac{Y}{2} B_\mu \right). \quad (1.7)$$

W_μ^i are gauge fields of $SU(2)_L$, and only couple to left-handed fermions. The g and g' are the corresponding coupling constants, respectively.

The Lagrangian of fermions using the covariant derivatives is:

$$\begin{aligned} \mathcal{L}_{EW} &= \sum_f \bar{L} i \not{D} L + \bar{\psi}_R i \not{D} \psi_R \\ &= \sum_f \bar{L} i \not{\partial} L + \bar{\psi}_R i \not{\partial} \psi_R + \mathcal{L}_{int} \\ &= \mathcal{L}_{free}^{EW} + \mathcal{L}_{int} \end{aligned} \quad (1.8)$$

where L is the isospin doublet and ψ_R is the singlet, the summation index f goes for all fermions including all generations of leptons and quarks, and $\not{D} = \gamma^\mu D_\mu$. The $U(1)_Y$ symmetry introduces the gauge field B_μ coupling to both left- and right-handed components of the fermions. The interaction term between the fermions and gauge bosons is:

$$\mathcal{L}_{int} = - \sum_f g \bar{L} \gamma^\mu \frac{\sigma_i}{2} L W_\mu^i + g' \bar{L} \gamma^\mu \frac{Y}{2} L B_\mu + g' \bar{\psi}_R \gamma^\mu \frac{Y}{2} \psi_R B_\mu \quad (1.9)$$

where σ_i are the Pauli matrices.

The fields related with observable bosons ($A_\mu \leftrightarrow \gamma$, $W_\mu^\pm \leftrightarrow W^\pm$ and $Z_\mu \leftrightarrow Z$) are obtained via:

$$\begin{aligned} A_\mu &= B_\mu \cos \theta_W + W_\mu^3 \sin \theta_W \\ Z_\mu &= -B_\mu \sin \theta_W + W_\mu^3 \cos \theta_W \\ W_\mu^\pm &= \frac{W_\mu^1 \mp i W_\mu^2}{\sqrt{2}} \end{aligned} \quad (1.10)$$

where θ_W is the Weinberg angle and can be expressed as:

$$\cos \theta_W = \frac{g}{\sqrt{g^2 + g'^2}}, \quad \sin \theta_W = \frac{g'}{\sqrt{g^2 + g'^2}}. \quad (1.11)$$

However, all the particles discussed above have to be massless, since any mass term in the Lagrangian violates the gauge invariance. Obviously, this is not true. This issue can be solved when the Higgs field is introduced, which will be explained in the next section.

1.1.4 The Higgs mechanism

The Brout-Englert-Higgs mechanism was introduced to generate masses to fundamental particles [17–19]. Introduced by the mechanism, the Higgs field exists anywhere in the vacuum. The idea is that the vacuum state does not need to preserve its intrinsic symmetries of the

Lagrangian. This is called the spontaneous symmetry breaking.

The Lagrangian of the Higgs is also required to be invariant under the $SU(2)_L \otimes U(1)_Y$ symmetry:

$$\mathcal{L}_H = D^\mu \Phi^\dagger D_\mu \Phi - V(\Phi) \quad (1.12)$$

where the Higgs field is an $SU(2)$ doublet composed of complex scalar fields, given by:

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix} = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}. \quad (1.13)$$

where the $\phi_{1,2,3,4}$ are real scalar fields. As the other $SU(2)_L$ doublet discussed before, the most positive charged component is the upper one ϕ^+ with $I_3 = +\frac{1}{2}$. The neutral field ϕ^0 has $I_3 = -\frac{1}{2}$. The symmetry breaking is obtained through the form of the scalar potential V :

$$V(\Phi) = \frac{1}{2}\mu^2 \Phi^\dagger \Phi + \frac{1}{4}\lambda(\Phi^\dagger \Phi)^2 \quad (1.14)$$

where it will be assumed that $\lambda > 0$ to ensure potential stability. For $\mu^2 > 0$, the potential shape is shown in Figure 1.2(a). In this case, the lowest potential is 0, reached at $\Psi = 0$, which is called the ground state and it keeps the $SU(2)_L \otimes U(1)_Y$ symmetry. If choosing the parameter $\mu^2 < 0$, the potential shape is as Figure 1.2(b), where the minimum of the potential $V(\Phi)$ occurs when $|\Phi| = \pm\sqrt{-\frac{\mu^2}{\lambda}}$.

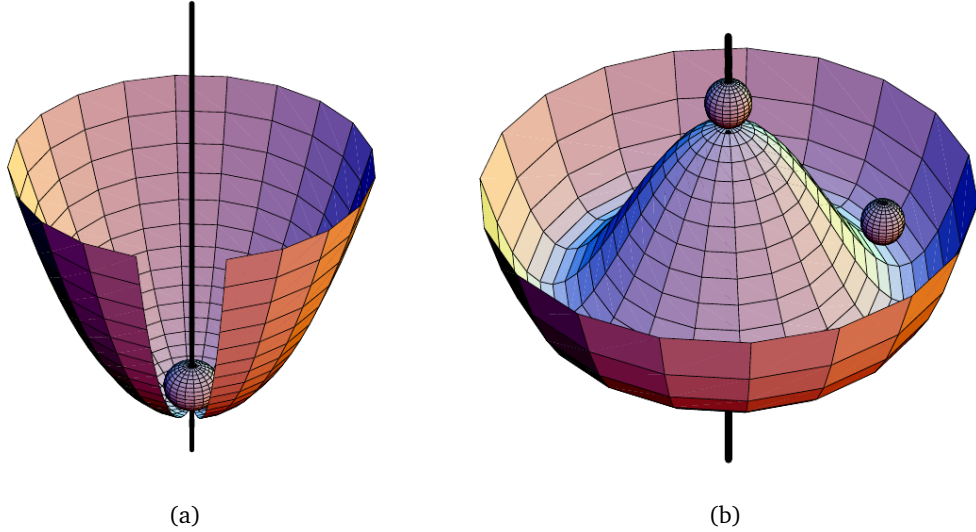


Figure 1.2: Higgs potential $V(\Phi)$ (a) before ($\mu^2 > 0$) where the minimum of the potential is at $\Psi = 0$ and (b) after ($\mu^2 < 0$) electroweak symmetry breaking where the minimum of the potential is at $|\Phi| = \pm\sqrt{-\frac{\mu^2}{\lambda}}$ [8].

Defining $v = \sqrt{-\mu^2/\lambda}$, the ground state can be chosen as:

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}.$$

However, this ground state violates both $SU(2)_L$ and $U(1)_Y$, but respects $U(1)_{em}$ symmetry. If

there is excitation from the ground state, it becomes:

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

Here the parameter v is the vacuum expectation value and H is a physical scalar field, which quantum excitation is the so-called Higgs boson.

Now vector bosons can have mass. Replacing the definition of Φ after electroweak symmetry breaking and using the covariant derivative from equation 1.7, the Lagrangian in equation 1.12 can be expended. Then W and Z mass terms can be derived as:

$$m_Z = \frac{v}{2} \sqrt{g^2 + g'^2}, \quad m_W = \frac{gv}{2}. \quad (1.15)$$

What is more, the Higgs mass can also be expressed as : $m_H = \sqrt{-2\mu^2} = \sqrt{2\lambda v^2}$. Since λ is a free parameter, the Standard Model does not predict the Higgs boson mass.

Fermion masses can be given in a similar way: as a Yukawa interaction with the Higgs field. It corresponds to a Lagrangian as follows (taking down-type quark as an example), which is invariant under $SU(2)_L \otimes U(1)_Y$:

$$\begin{aligned} \mathcal{L}_{yukawa} &= -y_d(\bar{\psi}_L \Phi \psi_R) + h.c. \\ &= -y_d \frac{1}{\sqrt{2}} (\bar{u}_L, \bar{d}_L) \begin{pmatrix} 0 \\ v + H \end{pmatrix} d_R + h.c. \\ &= -\frac{1}{\sqrt{2}} y_d v \bar{d}_L d_R - \frac{1}{\sqrt{2}} y_d \bar{d}_L H d_R + h.c. \end{aligned} \quad (1.16)$$

where y_d is the fermion Yukawa coupling denoting the interaction between the Higgs boson and fermions. The first item gives the mass of the down-type quark $m_d = y_d v / \sqrt{2}$, while the second item shows how fermions couple to the Higgs field H with strength y_d . Only quarks and charged leptons receive mass, since they have both right and left components.

1.1.5 Strong interaction

The strong interaction found between quarks and gluons, together referred to as partons, are described by the theory of quantum chromodynamics (QCD), based on the $SU(3)_C$ non-Abelian gauge group. C refers to the color charge, a quantum number carried by quarks and gluons. Each quark takes one color while a gluon takes a color and a different anti-color. Like the electric charge in QED, the strong force between quarks depends on their color charges. Three colors are considered, labeled as red, green and blue, and they belong to a $SU(3)_C$ triplet. One defines the color and hence a quark field of flavor f :

$$|r\rangle = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, |g\rangle = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, |b\rangle = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad q_f = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} \quad (1.17)$$

where $f_{1,2,3}$ represent the quark fields for red, green, and blue color. $SU(3)_C$ transforms the quark field as $q'_f = e^{-i\vec{\alpha}(x) \cdot \vec{T}}$ where $T_i = \frac{\lambda_i}{2}$. λ_i are the eight Gell-Mann matrices, and α_i are functions of space-time coordinates.

To have a quark field Lagrangian following the $SU(3)_C$ symmetry, like in the previous

sections, the partial derivative should be replaced with covariant derivative:

$$D_\mu = \partial_\mu + ig_s \frac{\lambda_a}{2} G_\mu^a \quad (1.18)$$

where eight gauge fields G_μ^a , the gluon fields, are introduced. Then the quark field Lagrangian becomes:

$$\mathcal{L} = \sum_f \bar{q}_f (i\gamma^\mu \partial_\mu - m) q_f - g_s \bar{q}_f \gamma^\mu \frac{\lambda_a}{2} q_f G_\mu^a \quad (1.19)$$

where f goes through all six quark flavors. The interaction term:

$$\mathcal{L}_{int} = \sum_f -g_s \bar{q}_f \gamma^\mu \frac{\lambda_a}{2} q_f G_\mu^a \quad (1.20)$$

describes the coupling between the gluons and the quarks with coupling constant g_s , or equivalently defining $\alpha_s = \frac{g_s^2}{4\pi}$.

The Lagrangian for the free gluon fields is:

$$\mathcal{L} = -\frac{1}{4} G_a^{\mu\nu} G_{\mu\nu}^a. \quad (1.21)$$

$G_a^{\mu\nu}$ has an additional term compared with the form of the electromagnetic tensor, in order to keep invariance under local $SU(3)_C$ transform:

$$G_a^{\mu\nu} = \partial_\mu G_\nu^a - \partial_\nu G_\mu^a - g_s f_{abc} G_\mu^b G_\nu^c \quad (1.22)$$

where f_{abc} is the structure constant of the $SU(3)_C$ group. This term means the gluons can interact with each other.

Confinement Partons carrying a specific color have never been observed as free particles, and this is referred to as color confinement. Instead, they are always grouped together by the strong interaction forming colorless particles. A particle formed of even number of quarks, usually one quark and one anti-quark, is called meson. Odd number of quarks, usually three quarks, are bounded as baryons, while anti-quarks, usually three, as anti-baryons. *Hadron* refers to either of them⁵.

Asymptotic freedom Due to the gluon self-interaction, the coupling strength α_s of the strong interaction obtains special properties [20]:

$$\alpha_s(Q^2) = \frac{12\pi}{(33 - 2n_f) \log(\frac{Q^2}{\Lambda^2})} \quad (1.23)$$

where n_f is the number of quark flavors, Q^2 is the momentum transfer, Λ is an intrinsic energy scale of QCD. As Q^2 is becoming very large, corresponding to a very restricted space scale, α_s becomes quite small, leading to quasi-free behavior of the partons inside hadrons. Thus in proton-proton collisions, the inside partons can be considered as free.

⁵The proton is made of three quarks, u , u and d , and is a baryon, or generally, a hadron. The proton-proton collider is thus also called hadron collider.

On the other hand, when the distance increases (correspondingly Q^2 decreasing), α_s is increased and gives enhanced potential energy between partons. The energy may exceed the threshold needed to create another quark-antiquark pair. The so-produced quark pairs could still have enough kinematic energy to produce another quark pair until the new produced one does not have enough kinematic energy and is bounded in hadrons. All hadrons produced this way are more or less collimated and form a cluster, called *jet*. With large α_s , this process can not be described with perturbation theory by calculating ME.

1.2 Higgs physics

1.2.1 Higgs production and decay

There are four Higgs production processes with large cross sections in ATLAS:

- Gluon–gluon fusion, referred to as ggF , as shown in Figure 1.3(a), has the largest cross-section among all Higgs processes as presented in Figure 1.4. This is due to the preferred gluon–gluon interaction during proton–proton collisions at the LHC energy scale. The two gluons interact via a quark loop and produce the Higgs boson. The top and bottom quarks, the heaviest quarks, give the largest contributions because of the large Yukawa coupling to the Higgs boson.
- Vector boson fusion, called VBF, is presented in Figure 1.3(b). The process with a vertex having W/Z and the Higgs boson connected allows to directly measure the coupling strength between them. On the other hand, the process has the signature of two boosted light jets in the forward and backward regions.
- Associated production with either W or Z boson is simply called VH , or Higgs strahlung. Figure 1.3(c) shows the leading order Feynman diagram for this process. This mode is used to study the $H \rightarrow b\bar{b}$ process with good signature requiring the W/Z decaying leptonically.
- Associated production with top-quark pair ($t\bar{t}H$) or bottom-quark pair ($b\bar{b}H$). Both processes contribute around 1% of the total cross-section of Higgs production. From the Feynman diagram shown in Figure 1.3(d), these processes allow direct measurement of the coupling between the Higgs boson and the top or bottom quarks. However, the $b\bar{b}H$ process has two b -jets produced close to the beam line which is hard to study. The $t\bar{t}H$ process allows to directly measure the top Yukawa coupling.

The production cross sections for all these processes as a function of the pp collision center-of-mass energy are shown in Figure 1.4.

The Higgs boson is a short-lived particle and decays to massive particle-antiparticle pairs at tree level. Figures 1.5(a) shows the decay to W/Z , and Figure 1.5(b) shows the decay to b -quark, τ and μ leptons, the three channels with leading branch ratios of the decay to massive fermions. The Higgs boson does not couple to massless particles. Thus the decays to $\gamma\gamma$ and gg are only via loops of massive particles. The Higgs decay to $\gamma\gamma$ via loops are shown in Figure 1.5(c) where the loop contains heavy quarks, and 1.5(d) and 1.5(e) where the loops are mediated by W/Z bosons. The Higgs decay branching ratios as a function of the Higgs boson mass are shown in Figure 1.6. At the Higgs mass of 125 GeV, $H \rightarrow b\bar{b}$ has the largest branching ratio, of around 58%.

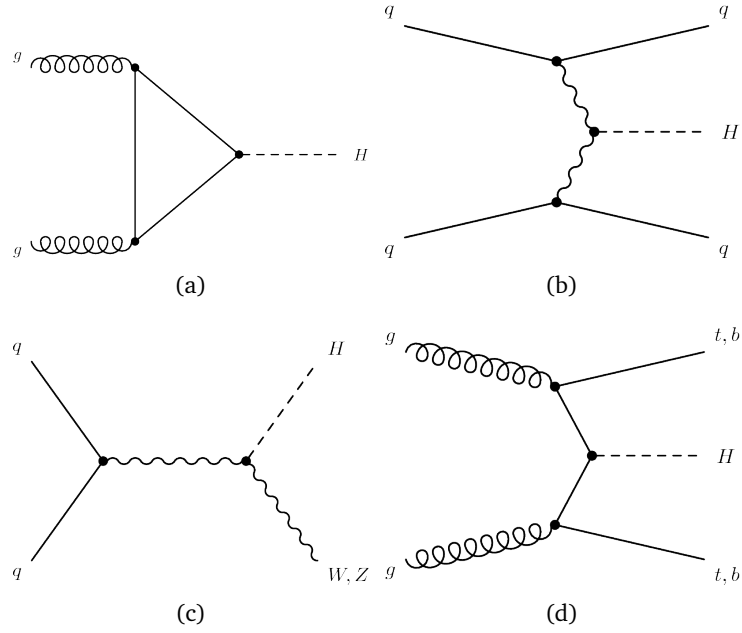


Figure 1.3: The lowest order Feynman diagrams of Higgs production processes: (a) gluon–gluon fusion, (b) vector boson fusion, (c) associated production with vector bosons, (d) associated production with heavy-quark pair. [21]

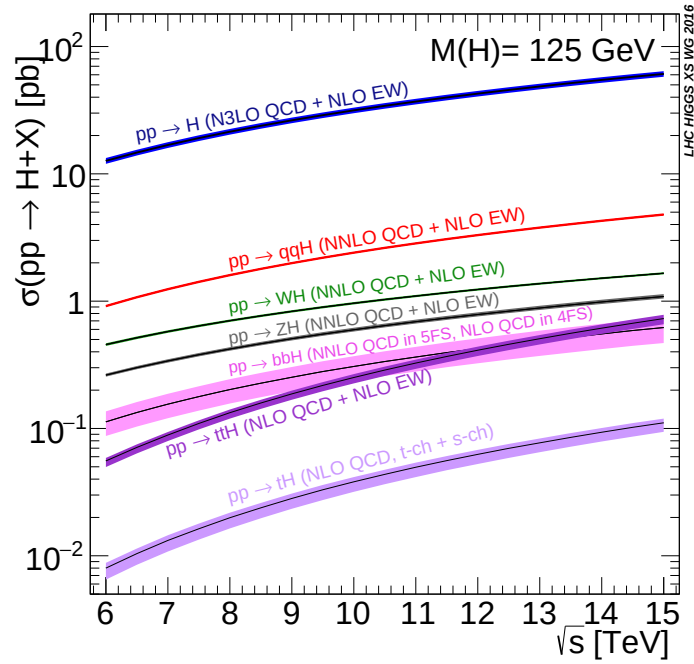


Figure 1.4: Standard Model Higgs boson production as a function of the proton–proton collision center-of-mass energy. The Higgs boson mass is set to 125 GeV [22]. $gg \rightarrow H$ is the ggF process, $qq \rightarrow Hqq$ represents the VBF mode, $q\bar{q} \rightarrow WH$ and $q\bar{q} \rightarrow ZH$ are the Higgs strahlung processes. And finally are the $b\bar{b}H$ and $t\bar{t}H$ production modes.

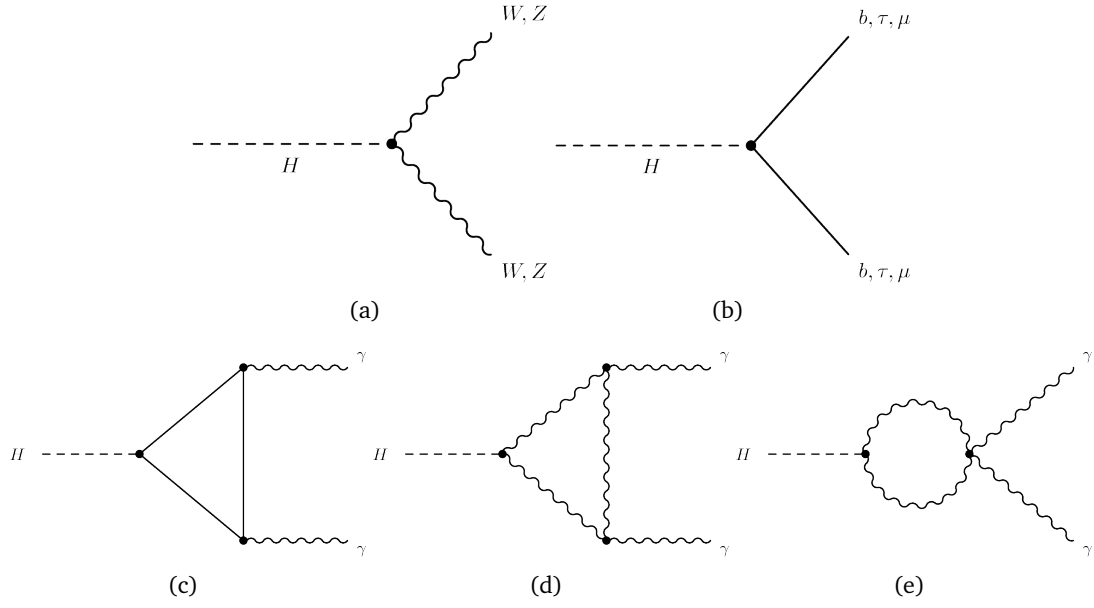


Figure 1.5: Standard Model Higgs boson decay processes [21].

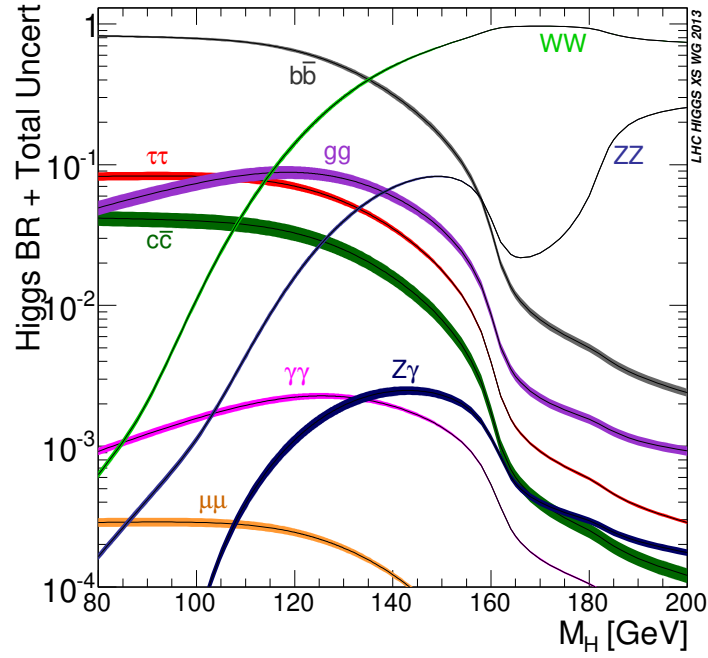


Figure 1.6: Branching ratios for various SM Higgs decay processes, as a function of the Higgs mass [23].

1.2.2 Higgs discovery

The Higgs boson discovery was announced in July 2012 by both the ATLAS and CMS collaborations [5, 6] as the last missing piece predicted by the SM. This led to the award of the Nobel Prize for physics in 2013 jointly to François Englert and Peter W. Higgs. The observation from the ATLAS experiment [5] is based on the datasets collected at $\sqrt{s}=7$ TeV in 2011 and 5.8 fb^{-1} at $\sqrt{s}=8$ TeV in 2012. Searches in individual channels are combined: $H \rightarrow ZZ^{(*)}$, WW^* , $b\bar{b}$, $\tau^+\tau^-$ and $H \rightarrow \gamma\gamma$ channels in the 7 TeV data, and the $H \rightarrow ZZ^{(*)} \rightarrow 4l$, $H \rightarrow \gamma\gamma$, and $H \rightarrow WW^* \rightarrow e\nu\mu\nu$ at 8 TeV. The observed significance is 5.9 standard deviations (σ) which corresponds to the probability of a statistical fluctuation from the background of 1.7×10^{-9} . The mass of the particle was measured as $126.0 \pm 0.4(\text{stat}) \pm 0.4(\text{sys})$ GeV. In parallel, the searches done by CMS give an observed significance of 5.8σ , and a mass of $125.3 \pm 0.4(\text{stat}) \pm 0.5(\text{sys})$ GeV [6]. The results of the two driving channels in ATLAS, $H \rightarrow \gamma\gamma$ and $H \rightarrow ZZ^{(*)} \rightarrow 4l$ are shown in Figure 1.7. The bumps around the Higgs mass of 125 GeV indicates the existence of this new particle.

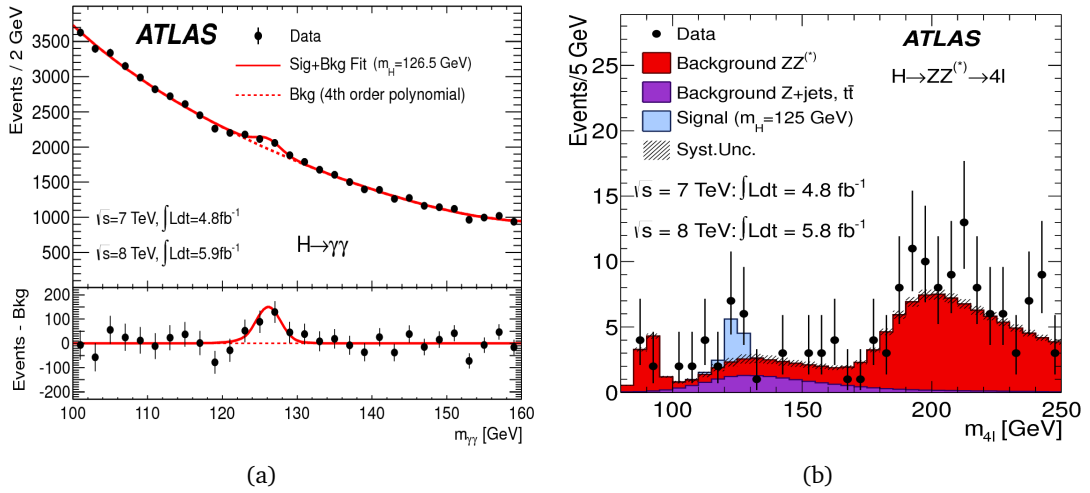


Figure 1.7: The invariant mass of (a) photon pairs reconstructed in the $H \rightarrow \gamma\gamma$ analysis and (b) four leptons reconstructed in the $H \rightarrow ZZ^{(*)} \rightarrow 4l$ channel [5].

1.2.3 Higgs coupling measurement

After the Higgs boson discovery, its properties need to be studied, which is an important way to challenge the SM and thus to test new physics. Being correlated with particle mass, the coupling between particles and the Higgs boson is an important property, the measurement of which is a combination study [24] including various Higgs production and decay modes. In the first operation period of the LHC (Run 1), from 2010 to 2012, the coupling between W/Z and the Higgs boson could be measured precisely showing consistency with the SM. This benefits from the observed $H \rightarrow ZZ^*$ and $H \rightarrow WW^*$ decay channels which allow to do direct and precise measurement of the Higgs and vector boson couplings. By then, none of the processes providing direct Yukawa coupling are observed, and thus they are indirectly measured with large uncertainty.

Upgraded to the center-of-mass energy of 13 TeV and to produce more Higgs events, the LHC

experiences a second operation period (Run 2), from 2015 to 2018, and measuring the Yukawa couplings is one of the priorities. Failed to be observed in Run 1, $t\bar{t}H$ production mode was very tempting by the start of this thesis, as it provides a direct approach to measure the top Yukawa coupling. This thesis contributes to the search for $t\bar{t}H$ in the $H \rightarrow b\bar{b}$ decay mode which is the dominant Higgs decay channel and was not observed by then. The result using part of the Run 2 data was published in April of 2018 [25]. In the second half of 2018, the $t\bar{t}H$ and $H \rightarrow b\bar{b}$ observations were published [26, 27] using data collected up to 2017 by combining several Higgs decay channels including the results of above mentioned $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis. These allow the direct measurement of both top and bottom quark Yukawa couplings.

Similarly to Run 1, the measurement involves multiple production and decay channels [28] using the κ framework. Figure 1.8 shows the results of one measurement assuming SM structure of the loops and no BSM contributions in decays. Couplings to W and Z are assumed to independently scale with modifiers κ_W and κ_Z , respectively. Coupling scales for top and bottom quarks, τ and μ leptons are introduced separately as κ_F ($F = t, b, \tau, \mu$). The plot shows the reduced coupling strength as a function of mass $\kappa_F \frac{m_F}{v}$ for fermions and $\sqrt{\kappa_V} \frac{m_V}{v}$ for vector bosons. The bottom and top quark Yukawa couplings are measured with good precision already and show agreement with the SM prediction.

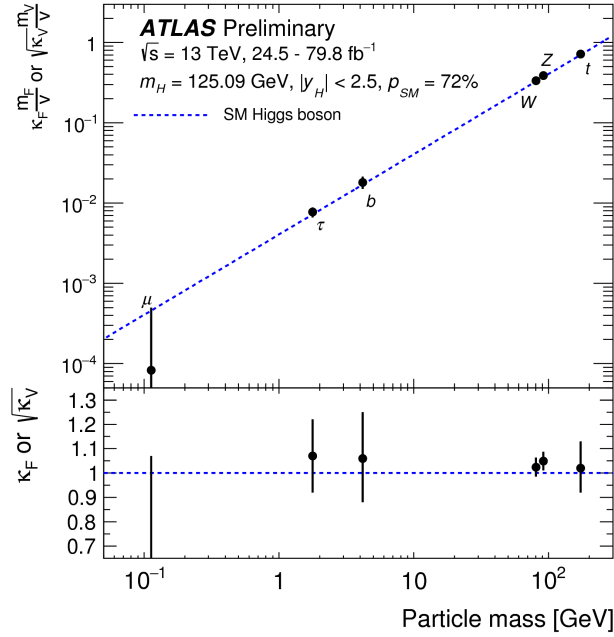


Figure 1.8: Reduced coupling strength modifiers $\kappa_F \frac{m_F}{v}$ for fermions ($F = t, b, \tau, \mu$) and $\sqrt{\kappa_V} \frac{m_V}{v}$ for weak gauge bosons ($V = W, Z$) as a function of their masses m_F and m_V , respectively. The vacuum expectation value of the Higgs field is $v = 246$ GeV. The SM prediction for both cases is also shown as dotted line. The coupling modifiers κ_F and κ_V are measured assuming no BSM contributions to the Higgs boson decays, and the SM structure of loop processes such as ggF , $H \rightarrow \gamma\gamma$ and $H \rightarrow gg$ [28].

More studies on the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis are still being carried out after the $t\bar{t}H$ observation exploring the full Run 2 data and new ideas and methods. Improvement in the result of this

channel could largely contribute to the precise measurement of the $t\bar{t}H$ cross section and top Yukawa coupling. What is more, exploring the CP structure of the $t\bar{t}H$ interaction [29] would be a way to further test the properties of the Yukawa coupling. The $t\bar{t}H$ ($H \rightarrow b\bar{b}$) has more statistics compared with $t\bar{t}H(H \rightarrow \gamma\gamma)$ to do differential measurements through variables such as Higgs p_T which could be better reconstructed than the $t\bar{t}H(H \rightarrow \text{multi-leptons})$.

The ATLAS experiment 2

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC), the most powerful accelerator ever constructed by human beings, is built in and operated by the European Organization for Nuclear Research (CERN¹). Located at the French-Swiss border near Geneva, CERN runs a complex of accelerators. It has 23 member states and around 2500 staff members working as scientific and technical researchers. What is more, CERN gets cooperation with over 12 000 scientists of 110 nationalities from institutes in more than 70 countries. Such a large-scale collaboration of scientists makes CERN a driving force of advancing human knowledge boundaries.

The LHC [30] was approved by the CERN Council in 1994 and built from 1998 to 2008 to replace the Large Electron-Positron (LEP) collider decommissioned in 2000. It collides beams of protons or ions accelerated at high energies to produce and study the properties of various particles and their interactions. To reach a high energy of particle acceleration, the LHC is a 26.7-kilometer ring installed in the same size tunnel 100 meters below the surface which was initially constructed to install LEP.

The first beam was circulated in 2008 and the first operational run (referred to as Run 1) started in March of 2010 with successful collisions of proton beams obtained at a record breaking center-of-mass energy of 7 TeV. Run 1 lasted until the early 2013 and the collision energy was successfully increased to 8 TeV in 2012. Then the LHC went through an upgrade towards higher energy and collision rate during 2012 to 2015 (Long Shutdown 1 referred to as LS1). After that, collisions were resumed as a second run (Run 2) period operating from mid 2015 to the end of 2018 at an energy of 13 TeV. Right now (summer 2019) the LHC is undergoing another upgrade (LS2) which will be followed by Run 3 with energy increased to 14 TeV. Then a new phase of the LHC (Phase 2) will be commissioned until 2035. Detailed information can be seen in Chapter 3 where the study contributes to the ATLAS upgrades for this new LHC era.

2.1.1 The accelerator system

The LHC is mainly designed to accelerate protons, as they are easy to obtain, stable and have low energy loss of synchrotron radiation compared to e^+ or e^- . As shown in Figure 3.1, starting from a simple bottle of hydrogen gas at LINAC 2 (linear accelerator 2), the protons are firstly produced by stripping hydrogen atoms of their electrons. Then they are injected into a succession of accelerators where the protons are shot into the next accelerator when they reach the maximum energy that the current accelerator can achieve:

¹CERN is the acronym of the organization's name in French, Conseil Européen pour la Recherche Nucleaire

1. LINAC 2 accelerates the protons to an energy of 50 MeV, amounting to one third of the speed of light c .
2. The Booster gives a boost to 1.4 GeV ($0.916c$).
3. The Proton Synchrotron (PS) pushes the protons to 25 GeV, 0.999 times c .
4. The Super Proton Synchrotron (SPS) gives a further boost to 450 GeV.
5. Finally, the protons are transferred to the two beam pipes of the LHC. One pipe circulates the protons clockwise while the other anti-clockwise. In 20 minutes protons in each pipe will reach their maximum energy of 6.5 TeV (in Run 2). In the end, the two beams are led to collision at four points which are enclosed by four detectors.

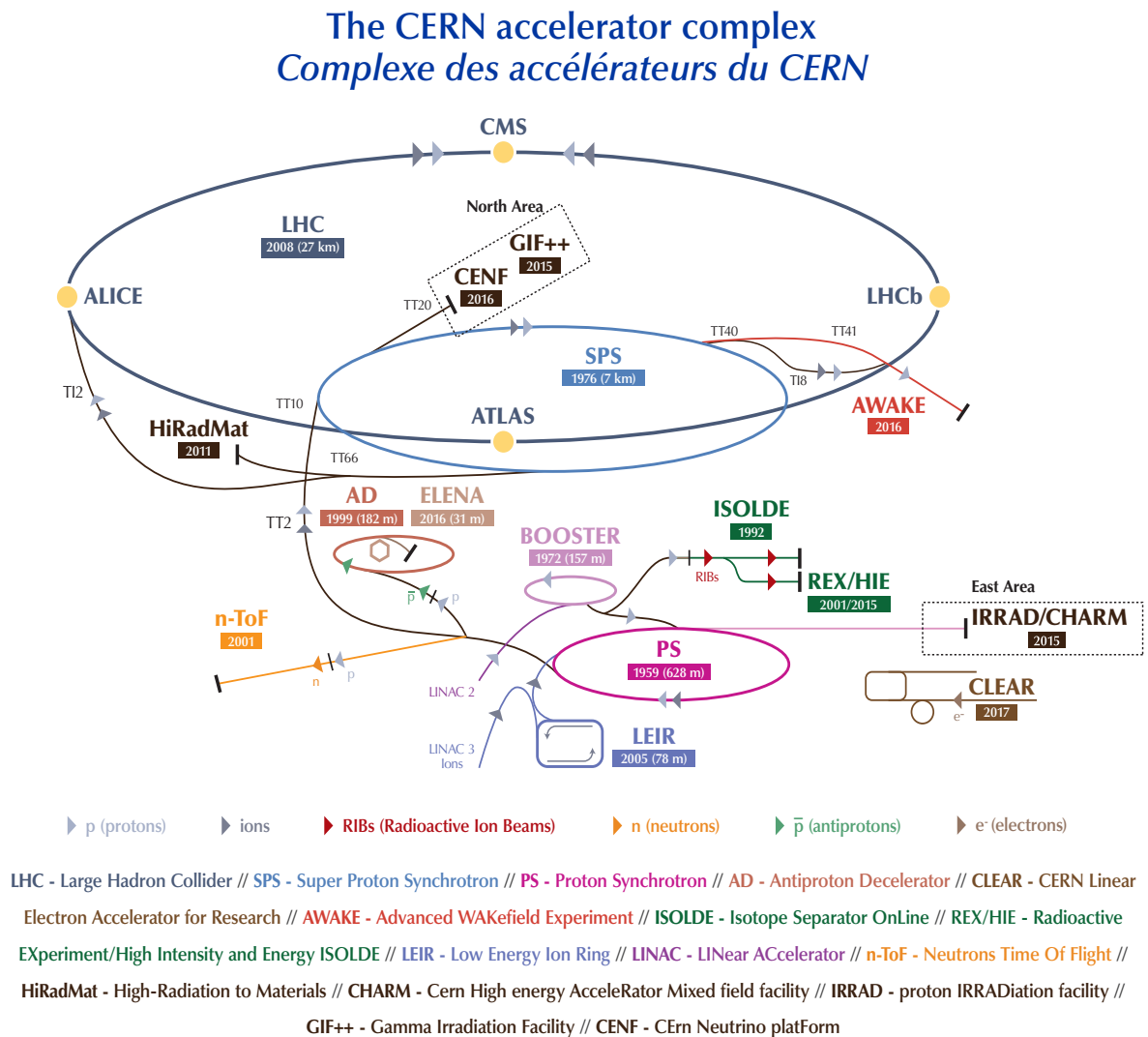


Figure 2.1: Schematic view of the CERN hadron accelerator chain [31]. The LHC is the last element of this system, in which the beams reach their highest energies.

The LHC can also perform heavy lead ion collisions at a center-of-mass energy up to 5 TeV per colliding nucleon pair, with the pre-acceleration of LINAC 3, PS and SPS.

2.1.1.1 The LHC ring setup

Particles in the LHC are accelerated by the so-called radiofrequency (RF) cavities [30] which are installed on the LHC ring as shown in Figure 2.2. An oscillating electromagnetic field with a frequency of 400 MHz is produced in the RF cavity which cannot only accelerate the protons, but also clump the protons in the beam into separated bunches where protons stay close to the desired energy. In each beam, 2808 effective bunches are formed. The bunches are 7 m (25 ns) apart and each contains around 10^{11} protons. The LHC accelerates, squeezes, focuses and crosses pairs of proton bunches, and each bunch pair crossing is referred to as an *event*.

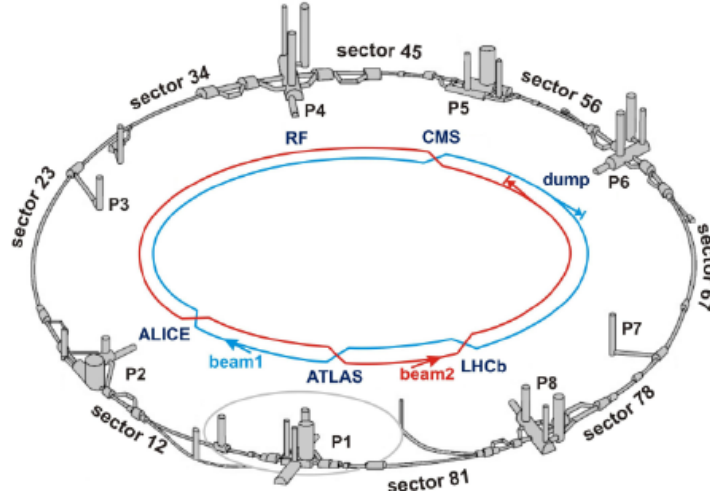


Figure 2.2: Schematic of the LHC ring showing the radiofrequency cavities located between CMS and ALICE.

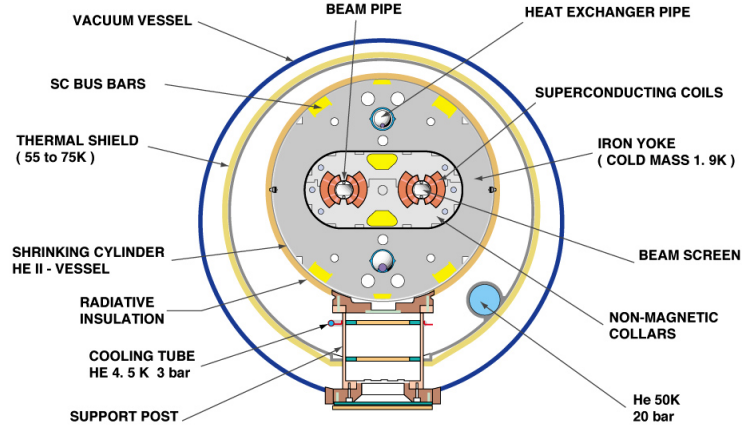
Thousands of electromagnets of various types and sizes are used to control the LHC high energy beams which usually require intense magnetic field. To bend the high energy protons in the limited-size ring, the magnetic field of 8.3 Tesla is generated by the dipole magnets, superconducting electromagnets chilled down to -271.3°C by superfluid helium. The cross section of the LHC main dipole is shown in Figure 2.3. In total there are 1232 dipole magnets each being 15 meters in length, and they fill more than $2/3$ of the ring [32]. Beams should be squeezed in order to maximize the chance of proton pair collision. To achieve this, 392 quadrupole magnets are deployed to focus the beam either vertically or horizontally. Dipoles are also equipped with sextupole, octupole and decapole magnets, which correct for small imperfections in the magnetic field at the extremities of the dipoles.

2.1.1.2 Luminosity

Luminosity is a measure of how many collisions are happening, or more precisely how squeezed the beams are in a given space and time. Assuming the equal shaped beams with Gaussian profiles in all dimensions, the instantaneous luminosity is:

$$\mathcal{L} = \frac{N_1 N_2 f N_b}{4\pi\sigma_x\sigma_y} \quad (2.1)$$

CROSS SECTION OF LHC DIPOLE



CERN AC_HE107A_V02/02/98

Figure 2.3: The cross section of the LHC main dipole.

where N_1 and N_2 are the number of particles per bunch, f is the beam revolution frequency, N_b is the number of bunches per beam, and σ_x and σ_y are the transverse beam dispersion perpendicular to the beam direction. The integrated luminosity (normally referred to as luminosity) is $L = \int \mathcal{L} dt$ which is used to denote the amount of data recorded over a period of time.

2.1.1.3 Pile-up

In each bunch-crossing, several pairs of protons can be actually collided and interact where the one with the highest energy, referred to as the *hard scattering*, is of interest. Such multiple pp collisions in one bunch-crossing is referred to as *in-time pile-up*. The mean number of interactions per crossing is denoted as $\langle \mu \rangle$. The LHC Run 2 operation gives a wide range of $\langle \mu \rangle$ for ATLAS in the years from 2015 to 2018, as shown in Figure 2.4. For a fixed time period and number of bunches, the more squeezed the bunches are, the higher luminosity and pile-up is. This produces a large number of events in processes of interest considering $N = L\sigma\epsilon$ with σ the process cross-section and ϵ the event selection efficiency.

Additional pp collisions occurring in bunch-crossings before and after the collision of interest is called *out-of-time pile-up*. This is due to the limited read-out time of sub-detectors.

2.1.2 The physics experiments on the LHC

The physics goal of the LHC is to test the SM of particle physics, and help to explore new physics that is not encapsulated by the SM. There are four main experiments located at the four collision points on the LHC ring:

- ATLAS (A Toroidal LHC ApparatuS) [34] is the largest detector operated at LHC and is about 45 meters long, more than 25 meters high, and weighs about 7,000 tons. It is designed for a wide range of physics purposes, and is aiming to detect and measure the

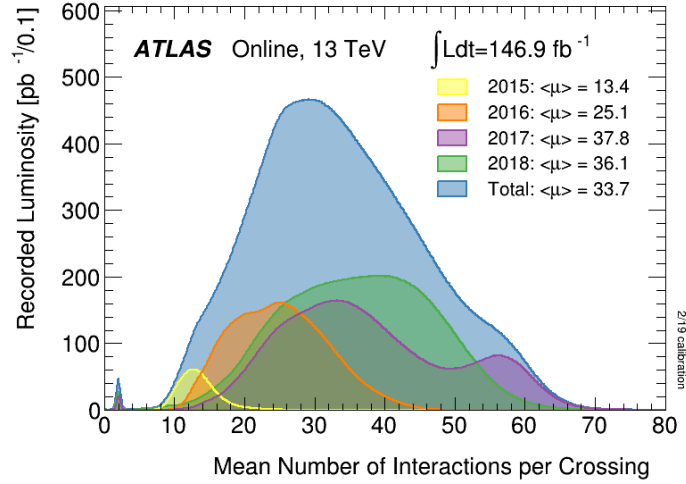


Figure 2.4: Luminosity-weighted distribution of the mean number of interactions per crossing for the Run 2 pp collision data at 13 TeV center-of-mass energy [33].

particles over a large mass range making full use of the large luminosity of the LHC. The details of this detector can be found in Section 2.2.

- CMS (Compact Muon Solenoid) [35] has the same physics goals as ATLAS, but is built with different techniques and is run by an independent collaboration for operation and data analysis. Together with ATLAS, they can cross-check each other on new results, and combine their datasets to further improve the precision.
- LHCb (Large Hadron Collider beauty) [36] is a single-arm apparatus aimed to detect particles thrown forward by the collision in one direction. This is different from ATLAS and CMS which surround the entire collision point. It is dedicated to studying the differences between matter and antimatter by studying b -quarks, which have large production cross-section at LHC and thus allows LHCb working at relatively low luminosity.
- ALICE (A Large Ion Collider Experiment) [37] is a heavy-ion detector which is designed to study the physics of strongly interacting matter produced in heavy ion collisions. With hundreds of protons and neutrons contained in the collided ion pairs, matter at extreme energy densities and high temperature forms quark–gluon plasma. It is of interest as believed to be like the conditions of the early universe.

2.2 The ATLAS experiment

ATLAS investigates a wide range of physics, which imposes strong requirements on the detector performance. For example, the search for and study of the SM Higgs boson needs high resolution of e, γ detection, excellent secondary vertex detection for τ -leptons and b -quarks. The VBF production channel with forward jet production requires the good reconstruction of jets close to the beam pipe. The precise mass measurement of W or top quarks needs good knowledge of the absolute energy scale of the calorimeters. One signature of searching for dark matter is the missing transverse energy (E_T^{miss}) which needs hermeticity. The particles with high momentum up to several TeV could be the signature for many analyses like searching for

new heavy bosons and extra dimensions. This requires the large bending power of the detector. This makes ATLAS a giant and complicated apparatus.

ATLAS is a cylindrical forward-backward symmetric detector with end-caps installed to provide almost full spatial coverage around the pp collision point. As shown in Figure 2.5, from the innermost beam pipe (BP) to the outside, various sub-detectors are installed to detect different particles generated from the collision:

1. Nearest to the collision point is the inner detector (ID), aimed to perform the charged particle tracking. Surrounded by a superconducting solenoid magnet providing a 2 T field, the charged particles leave bent tracks in the detector. The curvature is used to calculate the charge and momentum of the particles.
2. The electromagnetic (ECAL) and hadronic (HCAL) calorimeters, installed beyond the solenoid, are used to measure particle energies.
3. The Muon Spectrometer (MS) sits within the 0.2–3.5 T superconducting toroid magnets to track muons which are poorly measured in the ID and calorimeters.

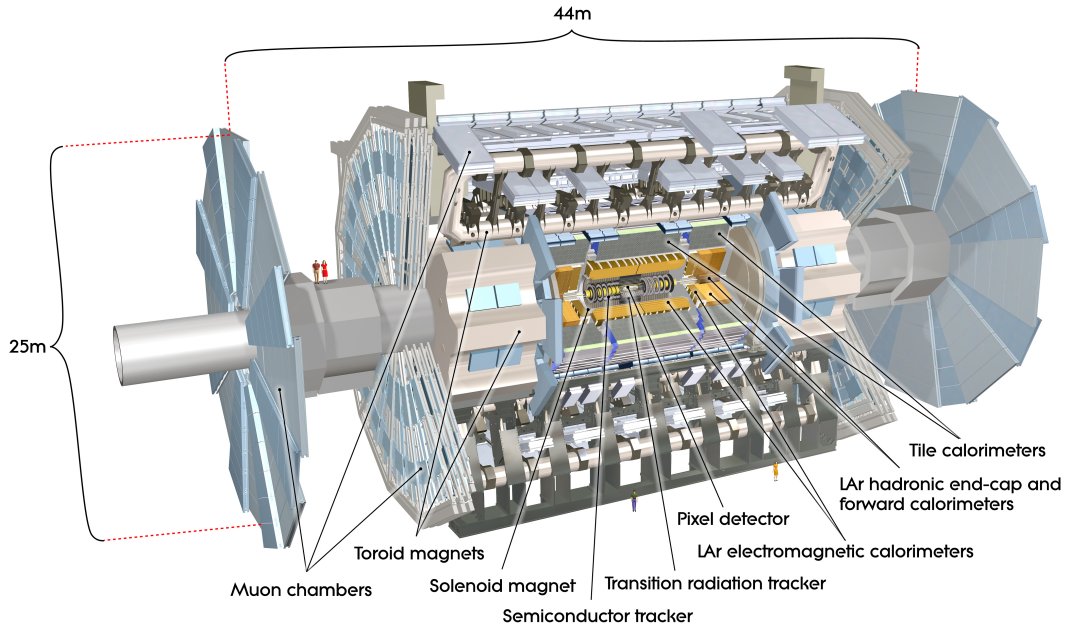


Figure 2.5: Section view of the ATLAS detector showing sub-detectors and the magnetic system.

Different particles leaves corresponding information in different sub-detectors as illustrated in Figure 2.6. A high frequency trigger and data acquisition system (TDAQ) reads out and chooses the interesting events to save them to disk.

2.2.1 Coordinate system

ATLAS uses a right-handed coordinate system with the interaction point (IP) as the origin, illustrated in Figure 2.7. The z -axis is defined by the beam direction, the x -axis directs into the center of the LHC ring and the y -axis points upward. ϕ is the azimuthal angle with respect to

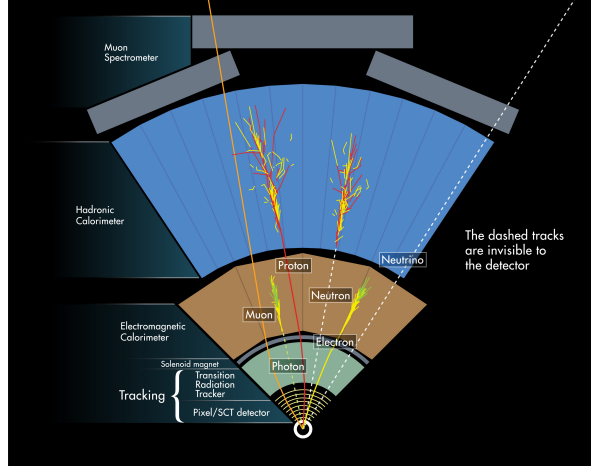


Figure 2.6: An overview of ATLAS interaction with various particles. [38]

the x -axis in the $x - y$ plane, in a range of $(-\pi, +\pi)$. The polar angle with respect to the z -axis is denoted as θ , ranging from 0 to π . The momentum component on this $x - y$ plane is called transverse momentum $p_T = p \sin \theta$, and the transverse energy is defined as $E_T = E \sin \theta$. The vector sum of transverse momentum of all collision products is conserved to be zero.

The collisions' center-of-mass momentum along the beam line p_z is non-zero as the interacting partons independently carry a fraction of the proton momentum. This leads to Lorentz boost along the beam direction. p_T and E_T are invariant under this boost while the azimuthal angle θ is not. Instead, the pseudo-rapidity η , which is a function of θ (Equation 2.2), or equivalently expressed as Equation 2.3, is usually used in particle physics. When $E \gg m$, almost always the case at the LHC, the pseudo-rapidity is close to the rapidity (Equation 2.4). The difference in rapidity Δy is invariant under the Lorentz boost caused by p_z , so as the $\Delta \eta \approx \Delta y$.

$$\eta = -\ln\left(\tan \frac{\theta}{2}\right) \quad (2.2)$$

$$\eta = -\frac{1}{2} \ln\left(\frac{|\vec{p}| + p_z}{|\vec{p}| - p_z}\right) \quad (2.3)$$

$$y = -\frac{1}{2} \ln\left(\frac{E + p_z}{E - p_z}\right) \quad (2.4)$$

The spatial separation of two objects is often defined as $\Delta R = \sqrt{(\Delta \eta)^2 + (\Delta \phi)^2}$, where $\Delta \eta$ and $\Delta \phi$ are differences of η and ϕ , respectively.

2.2.2 Inner detector

The inner detector (ID) is made up of the Pixel Detector including the Insertable B-layer (IBL), Semiconductor Tracker (SCT) and Transition Radiation Tracker (TRT), as shown in Figure 2.8. They occupy the innermost part of ATLAS up to around 1 m. This closeness to the beam pipe allows for precise tracking of a charged particle with less interaction with non-active material like cables and support structures. A particle track is built from the many hits it leaves in different layers of the ID. Based on tracks of particles in a collision final state, the interaction point of proton pairs can be measured, along with the momentum and charge of those particles

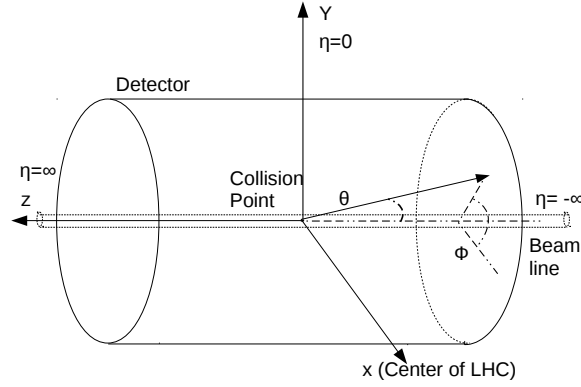


Figure 2.7: ATLAS coordinate system

bent by the magnetic field. A Run 2 event display is shown in Figure 2.9 which illustrates the challenging density of tracks produced by a collision in the ID.

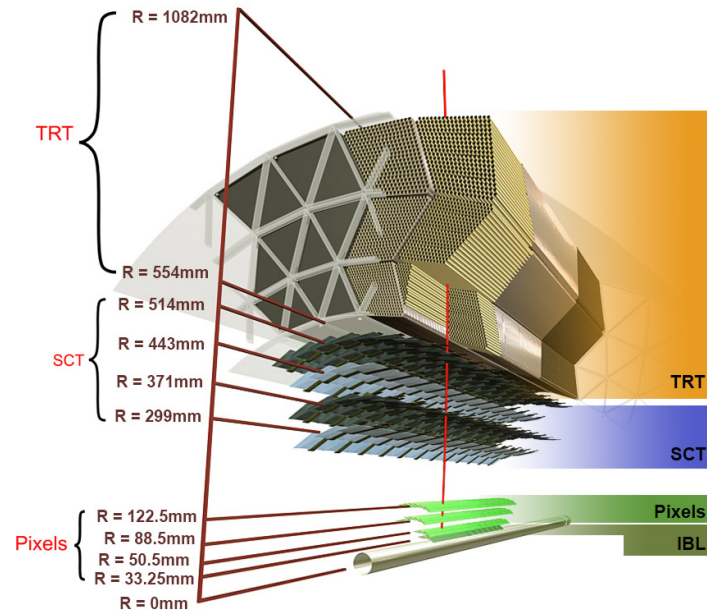


Figure 2.8: Schematic of the ATLAS inner detector in the barrel region.

All three ID sub-detectors are divided in two regions as shown in Figure 2.10. The barrel region is at low η and is build with concentric cylinder structures around the beam pipe. The high η region is also covered by all sub-detectors which are supported by disk structures perpendicular to the beam pipe, and this region is referred to as endcaps.

Being close to the IP, the ID is also confronted with hard radiation. Considering the high spatial resolution and the ability to be radiation resistant, the Pixel Detector and SCT are all semiconductor detectors while the TRT is a gas drift tube detector. In the LHC phase 2 operation, the ID will be replaced with an all-silicon detector to cope with the high track multiplicity and radiation harsh environment.

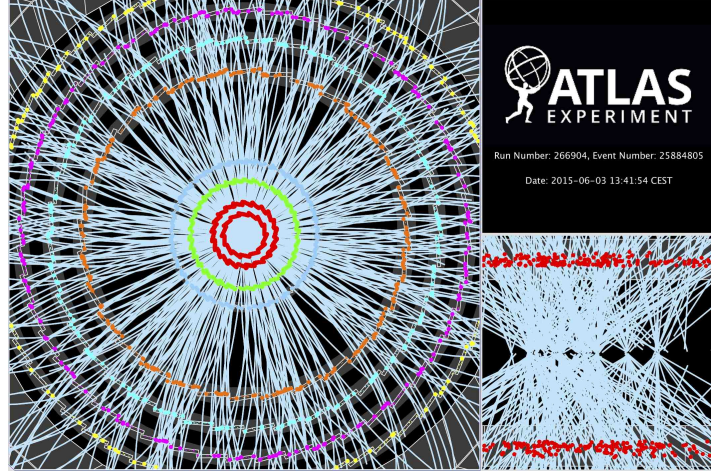


Figure 2.9: An event display showing the dense tracking environment in the ID [39].

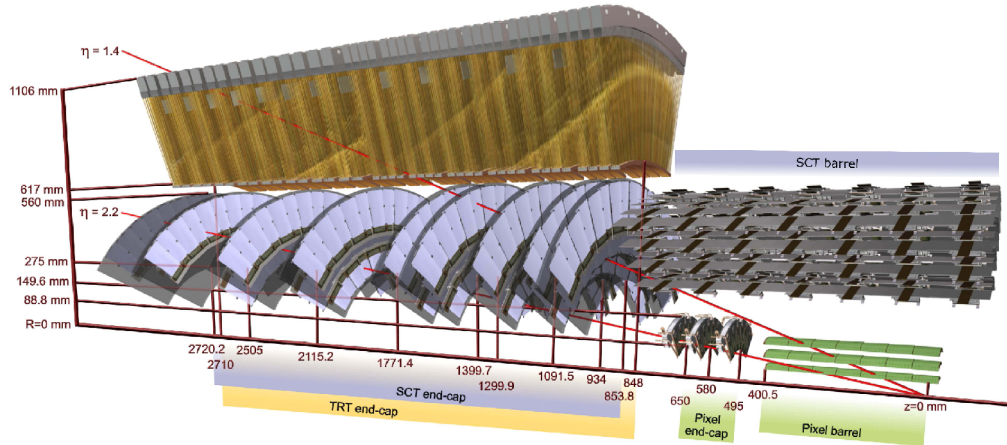


Figure 2.10: Schematic of the ATLAS inner detector showing an endcap region. The IBL is not included

2.2.2.1 Pixel detector

The pixel detector uses silicon semi-conductors as the active sensor for particle detection. When a charged particle transverses the semi-conductor, it ionizes the atoms on its path, producing electrons and holes as shown in Figure 2.11(a). Under a bias voltage, these charge carriers will be collected by electrodes as the signal. The unit element of the pixel detector is modules, each consisting of a pixel sensor bump-bonded to 16 front-end read-out chips (FEs), and some other electronics as shown in Figure 2.11(b). A module contains 46080 pixels each of a size of $50 \times 400 \mu\text{m}^2$ giving a track resolution of $14 \mu\text{m} \times 115 \mu\text{m}$.

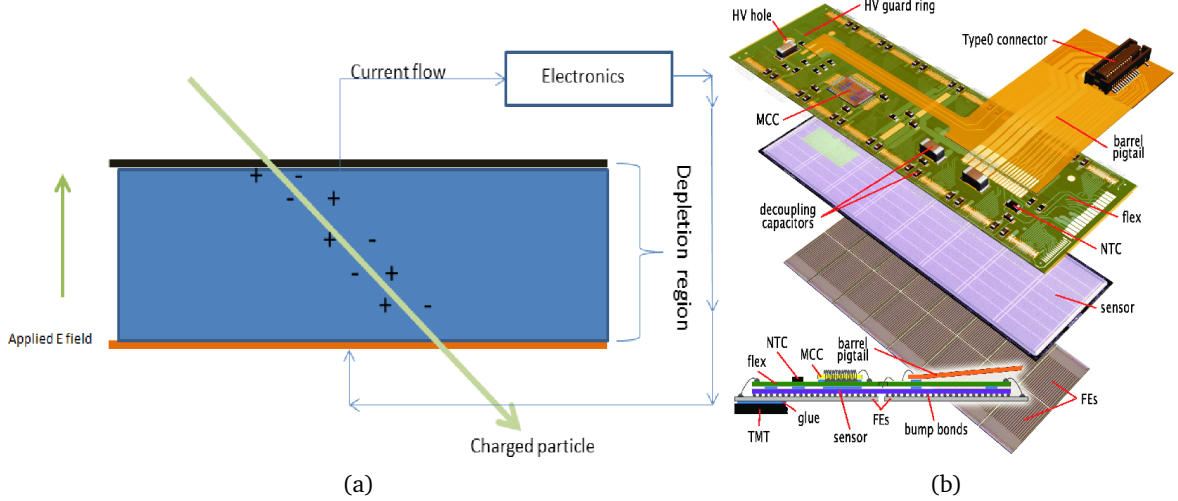


Figure 2.11: (a) Sketch showing the working principle of the pixel detector. (b) The ATLAS ID pixel module layout [34].

The pixel detector is originally made up of three barrel layers composed of staves and three endcap disks as shown in Figure 2.12(a). The modules are mounted on these staves and disks on which cooling system, cables are carefully routed. In total there are around 67×10^6 and 13×10^6 electronic channels to read out all pixels in the barrel and endcap regions respectively [34].

The IBL [40, 41] can be thought of as a fourth pixel layer at a radius of 33.25 mm, added as the last piece at the innermost of the ID. It was inserted into the ID during LS1 in 2014 after Run 1 to improve tracking precision and b -tagging (see Section 2.4.4) performance while preparing for the higher luminosities of Run 2. Figure 2.12(b) shows in reality how the IBL is inserted. The IBL contributes 12×10^6 additional pixel readout channels, and together with other pixel layers the ATLAS vertex can be reconstructed with a resolution of $11 \mu\text{m}$ in x and y directions and $24 \mu\text{m}$ in z .

2.2.2.2 The semi-conductor tracker

The SCT is located further away from the IP with lower requirements on the tracking performance. As a compromise to reduced costs and readout bandwidth, the SCT sensor is made of 768 p-n junction microstrips of 12.8 cm length, each being coupled with their readout. An SCT barrel module contains four sensors, two paired on the top, and two paired on the bottom side as shown in Figure 2.13(a). The two pairs are set at a 40 mrad stereo angle to each other allowing for 2D reconstruction in the module plane. The same strategy is also used for the

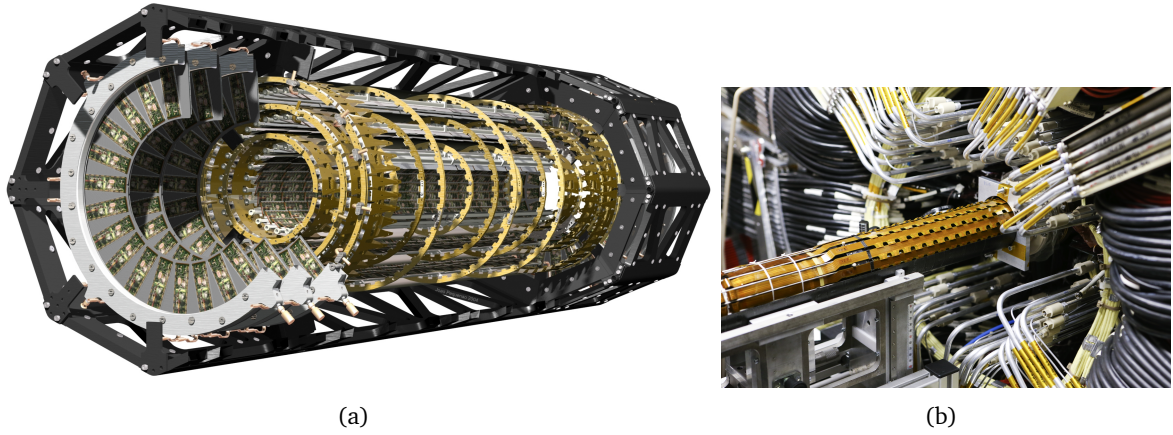


Figure 2.12: (a) Layout of the pixel detector showing modules mounted on supporting structures [42]. (b) Photo of the insertion of the IBL layer in the ATLAS ID [43].

endcap modules.

Similar to the pixel detector, modules are mounted on staves and disks. In total there are four cylindrical barrel layers and 18 endcap disks. Figure 2.13(b) shows the outermost SCT layer in reality with 672 modules mounted on. Compared to the pixel detector, the total number of readout channels is lowered to 6×10^6 , at the cost of a reduced resolution of $17 \mu\text{m}$ in $r\phi$ and $580 \mu\text{m}$ in z for the barrel and $17 \times 580 \mu\text{m}^2$ in $r\phi \times r$ for the disks.

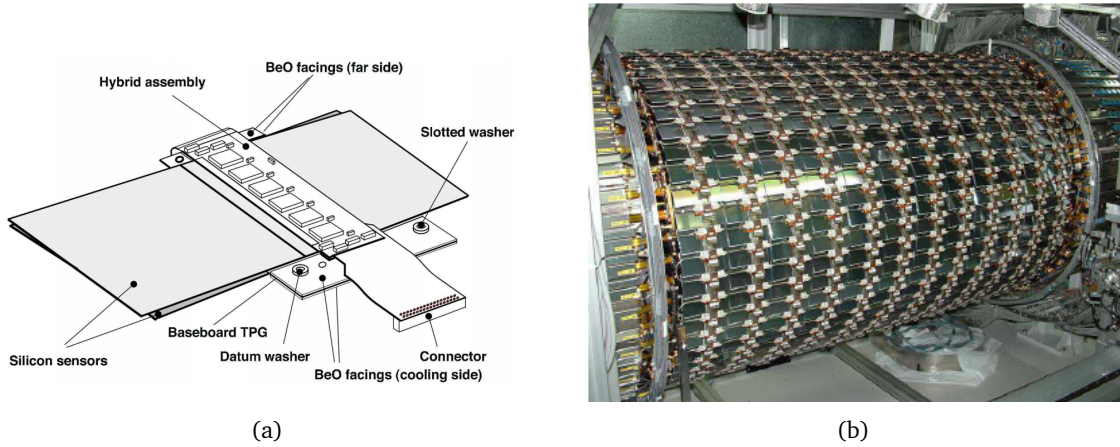


Figure 2.13: (a) Drawing of an SCT module [34]. (b) Photo of the SCT outermost barrel layer with modules mounted on [44].

2.2.2.3 Transition radiation tracker

The fundamental elements of TRT are 4 mm diameter carbon-polyimide drift tubes filled with gas, usually called straws. The straw surface acts as the cathode and the anode is a central wire under high voltage. When a charged particle travels through a straw, it ionizes the gas along its track. The so-produced negative charges are drifted to and collected by the wire, generating hits signals. This process is demonstrated in Figure 2.14. By measuring the drift

time distribution of collected charges, the distance from the track to the wire can be estimated.

Along with measuring hits, the TRT can also perform electrons identification. Relativistic charged particles emit transition radiation (TR) photons as they pass through the boundary of materials of different dielectric constants. The intensity of the TR and the resulting ionization is proportional to E/m where E and m are the energy and rest mass of the incoming particle. Thus for a particular amplitude of collected charges the electron can be identified due to their very light mass. Polymer radiator foils are installed in between the straw tubes and are meant to create TR as shown in Figure 2.14. The gas in the tubes is Xenon, which is a very good TR absorber. Thus the TR photons deposit much more energy than a normal track and can be distinguished by requiring a much higher signal threshold.

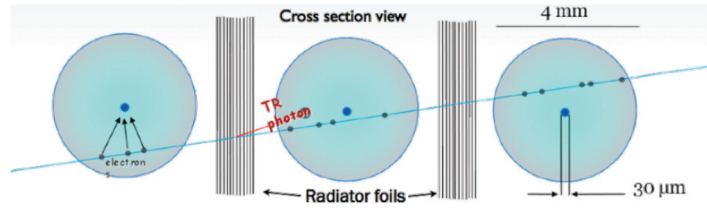


Figure 2.14: Schematic of the TRT working principle [45].

In the barrel, the TRT is made up of 73 layers of straws each being 144 cm long. These modules are then arranged to form a cylinder with the straws parallel to the beam axis, as shown in Figure 2.8. In the endcaps, straws are arranged radially on the disks as shown in Figure 2.10. Due to the length of the straws, the TRT only provides $R - \phi$ position measurement at an intrinsic accuracy of $130 \mu\text{m}$.

2.2.3 Calorimeters

The ATLAS calorimeter system measures the particles energy and stops almost all known particles by intensively interacting with them except for muons and neutrinos. It is composed of electromagnetic calorimeters (ECal), hadronic calorimeters (HCal) in both barrel and endcaps, and forward calorimeters (FCal). This provides hermetic coverage up to $|\eta| < 4.9$, which is very important for measurements of missing energy in various analyses induced by non-interacting particles.

All these sub-detectors are sampling calorimeters, composed of alternating layers of dense absorbing material, such as lead, and active detector material. The absorbing layers interact intensively and slow down an incoming particle generating a shower of particles in itself. The active material samples the energy of the resulting showers. This process is repeated several times through the layers with repeating structures until the particles are fully stopped. When being sampled the particle energy has been lost in the absorbing layers. So calibration of the detector response is important to recover the original energy of the incoming particles.

2.2.3.1 Electromagnetic calorimeter

The ECal is the first calorimeter encountered by particles beyond the ID and the solenoid. It can measure the energy of photons and electrons. The ECal in the barrel covers up to $|\eta| < 1.475$ and in the two endcaps $1.375 < |\eta| < 3.2$.

The ECal uses an accordion geometry as shown in Figure 2.15 where lead as absorber and liquid argon (LAr) as activating material are interleaved as detailed in the right plot in

Figure 2.16. Near an atom nucleus, the high energy photons undergo e^+e^- pair production while the electrons being decelerated radiate bremsstrahlung photons. This effect is enhanced when the nearby atom nucleus is heavy. Due to these effects an electromagnetic (EM) shower is induced once a photon or electron interacts with the absorber. The EM shower ionizes Ar atoms, and the charges are drifted and recorded by the readout electrodes as demonstrated in the right plot in Figure 2.16. Electrons and photons are finally stopped and deposit their energy in the ECal through several lead-LAr layers, forming an overall shower along their way, as shown in the left plot in Figure 2.16.

The accordion style geometry is repeated, providing ϕ symmetry without azimuthal cracks. Along the R direction, there are three layers with different cell granularity as shown in Figure 2.15.

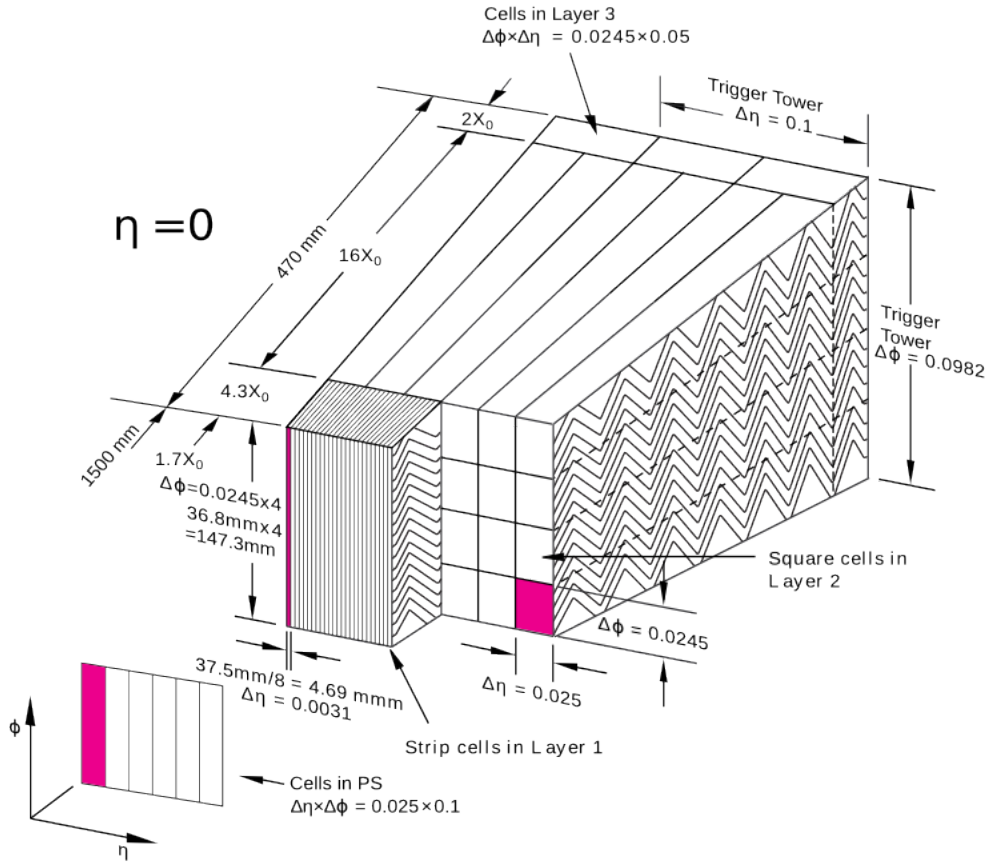


Figure 2.15: Drawing of the ECal barrel module consisting of three layers with different cell granularity [34].

The measured energy resolution is expressed as:

$$\frac{\sigma(E)}{E} = \frac{10\%}{\sqrt{E(\text{GeV})}} \oplus 0.17\% \quad (2.5)$$

which shows that the energy resolution is improved with increasing energy E . The resolution consists of an energy dependent term due to EM shower fluctuations, and a 0.17% constant term resulting from local non-uniformities. For 100 GeV electrons the energy resolution is around 1% [34].

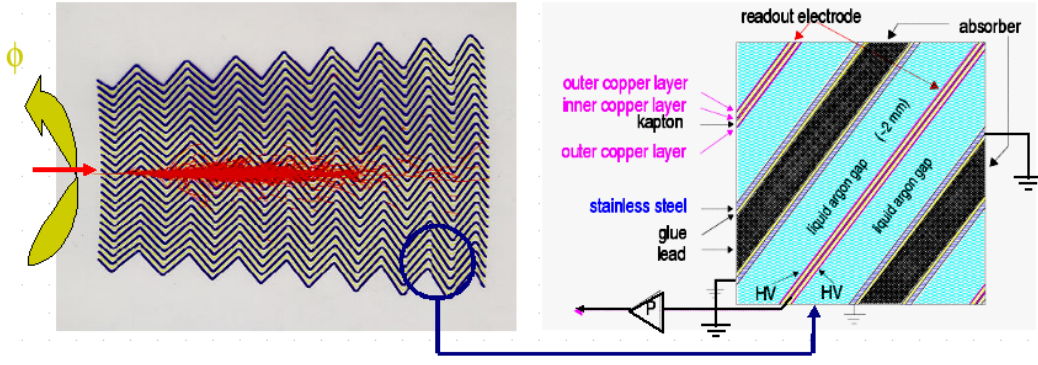


Figure 2.16: The accordion geometry of the ECal detector which uses lead-LAr layers. The left drawing shows the EM showers produced in an EM module (adapted from Reference [46]).

2.2.3.2 Hadronic calorimeter

The HCal surrounds the ECal, and can stop and measure the energy of hadronic particles. These particles have longer interaction length λ , and are hard to stop in the ECal. The HCal consists of the barrel tile calorimeter covering $|\eta| < 1.7$ and two LAr endcap calorimeters covering $1.5 < |\eta| < 3.2$.

The barrel region HCal is a tile calorimeter which uses scintillating plastic tiles as active material and steel as absorber, as illustrated in Figure 2.17. A hadronic shower traversing the polystyrene tiles excites molecules in it. As the excited molecules return to the ground state, ultra-violet scintillating light is released, which is then wavelength-shifted to the visible spectrum, collected in optic fiber cables and measured by photomultiplier tubes installed on the edges of the calorimeter. The tile HCal is organized into three layers. The first two are segmented into $\Delta\eta \times \Delta\phi \approx 0.1 \times 0.1$ cells while the third is coarser at 0.2×0.1 . The endcaps HCal use similar technology as the Ecal using LAr as active material and copper as absorbing layers.

The showers induced by hadrons in material are more complex and thus more challenging to measure. This shower contains both the EM component which deposits its energy in the EM calorimeter, and the hadronic component. The latter contains fraction of energy that does not contribute to signals. This invisible energy mainly consists of the binding energy of nucleons released in the nuclear interactions. The long-lived neutral decay particles, like neutrinos produced from the non-EM shower, will escape the detector carrying away another fraction of energy which cannot be detected. Additionally the proportions of the EM and hadronic components of a shower vary with energy. All these effects limit the HCal resolution compared with the ECal one. For pions at 350 GeV and $\eta = 0.25$, the HCal resolution is 6.6% [34].

2.2.4 Muon spectrometer

The Muon Spectrometer (MS), installed as the outermost sub-detector of ATLAS (shown in Figure 2.18), provides tracking and energy information for muons. The muon is a very non-interacting lepton which can penetrate the whole ATLAS detector. Due to its high mass, 200 times heavier than the electron, it loses much less energy through EM interaction than electrons. Also, it rarely undergoes interaction with atomic nuclei losing energy, which is the opposite case to hadrons. Thus the muon can penetrate the ID and calorimeters and arrive at the

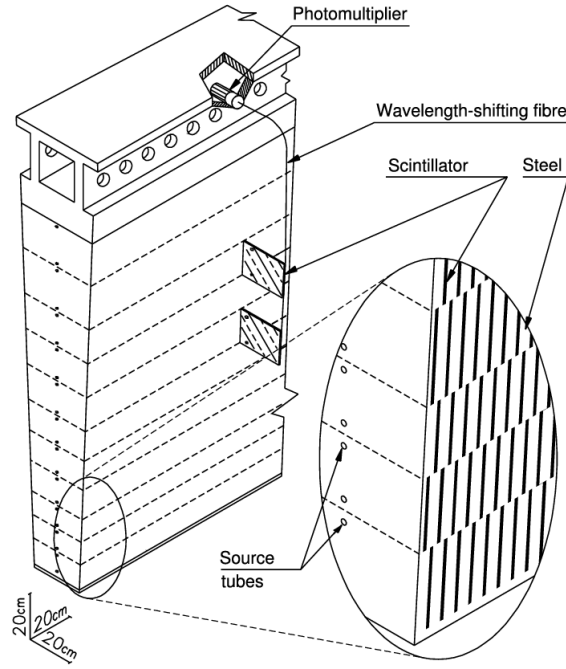


Figure 2.17: Drawing of the ATLAS HCal tile module [34].

MS retaining almost all its original energy. In the ID, it leaves tracks that are less curved because of its larger mass, negatively impacting the momentum measurement resolution. And in calorimeters the energy cannot be well measured if the muon does not interact much with the detector. The MS is designed to address both issues, and some of its readouts are also employed by the trigger system.

The MS consists of various components of different detector technologies as shown in Figure 2.19: monitored drift tubes chambers, resistive plate chambers, thin gap chambers, and cathode strip chambers. It provides coverage up to $|\eta| < 2.7$.

Monitored drift tubes (MDT) chambers are the primary detection technology and are used for precision tracking, providing coverage of $|\eta| < 2.7$ (the light blue blocks in Figure 2.19). In each chamber three or four layers of 3 cm diameter aluminum tubes are installed on either side of the support structures oriented in the azimuthal direction. The tubes are filled with an Ar and CO₂ gas mixture with a central wire at high voltage. An optical alignment system is built to monitor deformations of the chambers during the detector operation. MDT gives a mean resolution of 80 μm in the η direction.

Cathode strip chambers (CSC) are used for precision tracking in place of MDTs in forward regions of $2 < |\eta| < 2.7$ in the innermost tracking layer where there is a high particle flux near the interaction point. These chambers are multi-wire proportional detectors which give higher granularity, rate capability, and time resolution.

The resistive plate chambers (RPC) in the barrel and thin gap chambers (TGC) in the endcaps are used for triggering. Both give lower resolution 5–10 mm but can quickly provide track information to the trigger system with a few tens of nanoseconds. RPC also help to reject cosmic rays coming from outside of the detector.

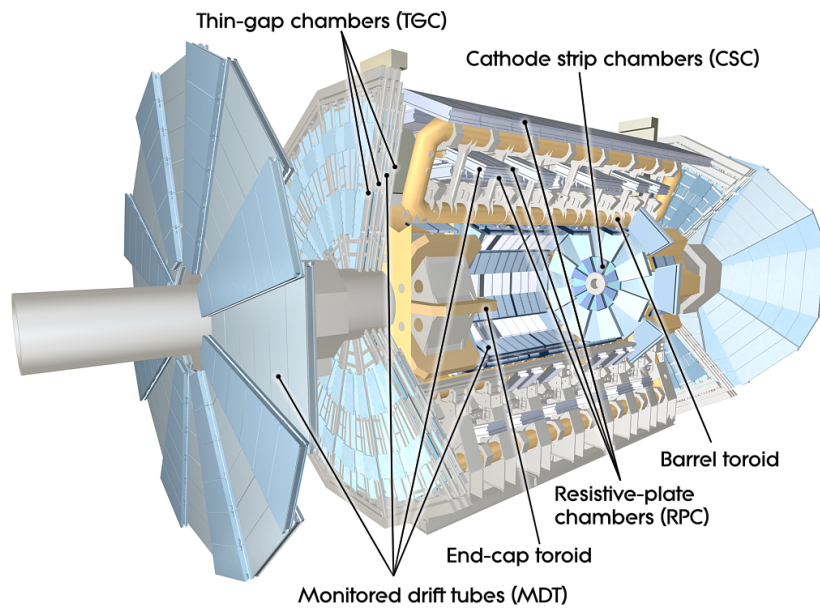


Figure 2.18: Section view of ATLAS showing components of the MS.

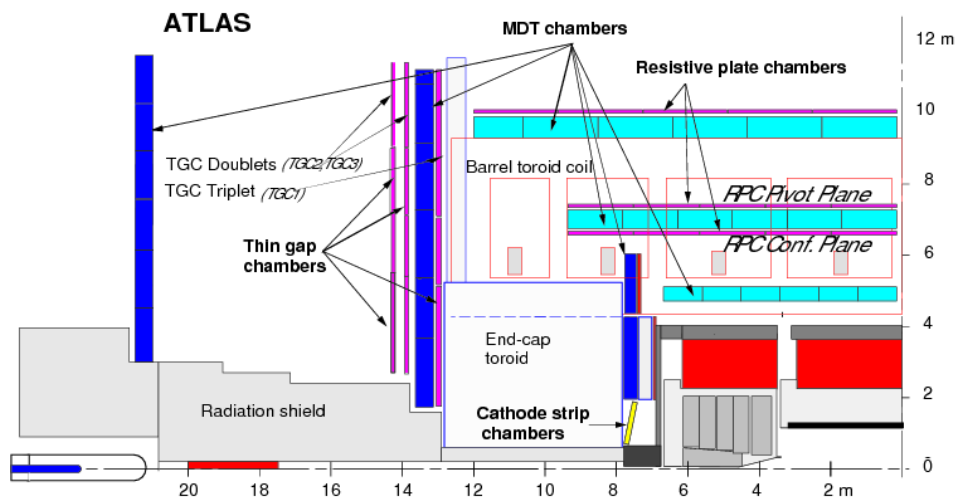


Figure 2.19: Drawing of the MS structure from side view [47].

2.2.5 Trigger and data acquisition system

In the LHC, beams with 25 ns apart bunches give a crossing rate of 40 MHz. Collectively there are around 10^8 readout channels for all sub-detectors in ATLAS in order to “watch” these events. Practically recording and storing all these data is not possible. Instead, a Trigger and Data Acquisition System (TDAQ) system is utilized to select interesting events for physics in real-time to conserve readout bandwidth and storage space.

The TDAQ in Run 2 consists of the level 1 (L1) trigger based on hardware reducing the event rate to 100 kHz, and high level trigger (HLT) based on software further decreasing the rate to 1.5 kHz. Figure 2.20 illustrates how the TDAQ system functions.

The L1 trigger selects events based on decisions made by logical circuits. It processes the inputs from both the calorimeters and MS to rapidly make an L1 acceptance decision and build regions of interest (RoI) in $\eta - \phi$ for the HLT. The L1 calorimeter trigger (L1Calo) identifies candidate e, γ and τ energy deposits as well as jets and E_T^{miss} . Muons are triggered in the MS (L1Muon) using simplified track finding algorithms while requiring levels of coincidence between MS layers. The information from L1Calo and L1Muon is further analyzed in the topological trigger module (L1Topo) to take into account the associations between triggered objects.

The HLT can access all the readout systems of ATLAS and retrieves more precise information from calorimeters and MS beyond L1. Tracking information from the ID is also exploited by the HLT. For each event, the HLT gives thousands of trigger chains, each being focused on a particular physics signature. Events passing at least one HLT chain are then permanently saved to storage for later offline reconstruction and analysis.

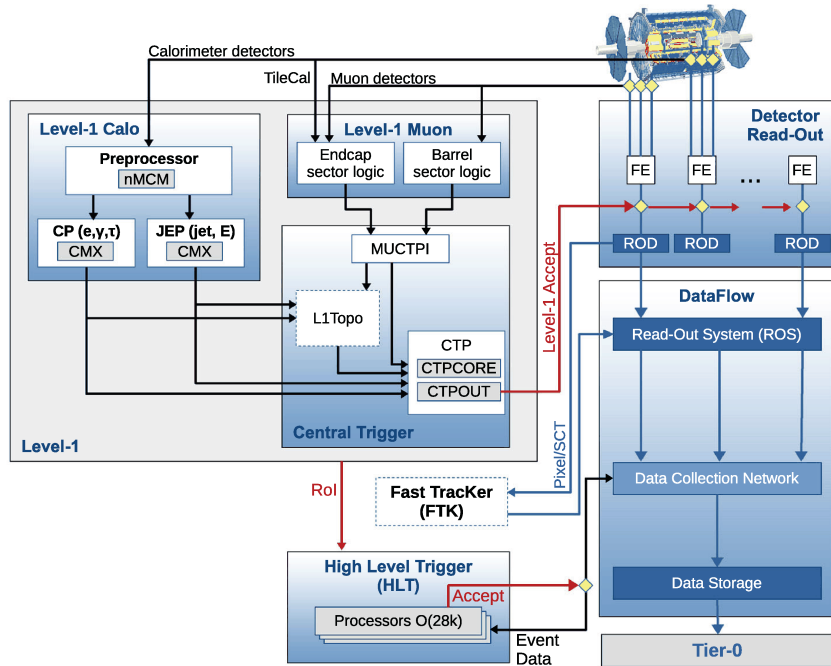


Figure 2.20: Schematic plot of the ATLAS Run 2 TDAQ system [48].

2.3 Monte Carlo simulated data

Most of ATLAS physics analyses, including the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) channel studied in this thesis use a model dependent strategy, i.e., aim to find signal excess in experimental data in comparison with the simulation based on theoretical prediction. Therefore a careful study and choice of simulations are important in order to design an unbiased physics analysis. Using the Monte Carlo (MC) method, the real data production and acquisition system is simulated. The simulated data is usually referred to as predictions or *Monte Carlo (MC) simulation*.

Event generation The first step in the simulation chain is called event generation. It simulates the event production, and the decays of the produced particles up to stable particles before interaction with the detector. Figure 2.21 schematically shows the collision of a pair of protons and the following steps.

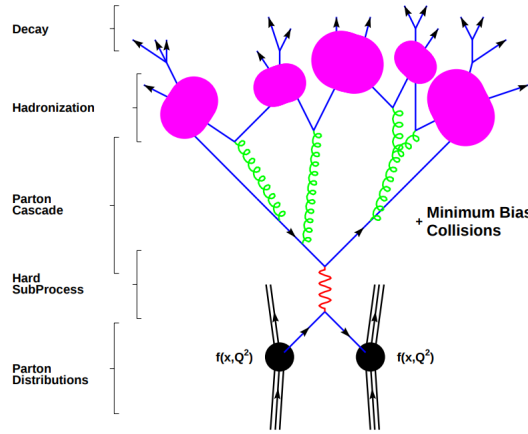


Figure 2.21: Schematic plot showing event generation steps for a proton-pair collision [49].

- The interaction between the colliding protons happens among a pair of component partons, referred to as hard scattering. As an important interaction parameter, the momentum of a parton takes part of the proton momentum, decided by parton distribution function (PDFs). This distribution is extracted from a fit to data in many experiments.
- The hard scattering process is simulated based on perturbative QCD which calculates the cross-section of a process up to a fixed order of precision. The corresponding process can be represented by Feynman diagrams which facilitate both the process representation and calculation. This step of calculation is referred to as matrix element (ME) calculation for which several tools have been developed. These tools are referred to as event generators. Among the ones used in this thesis are POWHEG-BOX [50–52], SHERPA [53] and MADGRAPH5_aMC@NLO [54].
- The partons both before (initial state radiation) and after (final state radiation) their interaction can radiate gluons which can further split into gluons or quark pairs. This process, referred to as *parton shower*, gives corrections to the ME calculation. The two calculations need to be combined which required algorithms (e.g. ME+PS method [55]) to remove the overlaps between their calculations.

- The partons in showers combine into hadrons as they reach the non-perturbative energy scale. Such a process is referred to as *hadronization* which is hard to calculate with perturbative QCD theory and is simulated with empirical models. The unstable hadrons at this stage could further decay into other particles before getting in touch with the detector. There are several tools to implement the parton shower and hadronization, and the ones used in this thesis are PYTHIA [56, 57], HERWIG [58, 59] and SHERPA.

The interactions of the remaining proton components (referred to as *underlying event*) and pile-up events (i.g., the effect of multiple pp interactions in the same and neighboring bunch crossings) are also simulated. These processes are usually at low energy and are simulated with empirical models with parameters which are tuned by fitting to data in dedicated measurements.

Detector simulation The stable particles from the above generated events are then propagated through the full ATLAS detector [60]. The GEANT 4 [61] simulation toolkit is used to simulate the particle interactions with detectors, and this process is referred to as *full simulation*. This process requires large computing resources as it is based on complex detector geometry and physics calculation, especially for the particle showering in calorimeters. Thus a *fast simulation* algorithm, which simplifies the showering calculation by a parameterization of the shower shapes, is also widely used at the cost of precision. Alternative samples as an estimation of theoretical modeling uncertainties are often produced this way.

The digitization step converts the simulated energy deposits in the detector into voltages and currents stored in an output format that is interchangeable with that of the ATLAS readout of real data. Then the HLT and reconstruction are performed on those simulated outputs in an identical way as for experimental data. For the simulated samples, all the truth information is recorded and available including, for example, the location of a particle decay within the ID and the tracks of all decay products.

2.4 Object reconstruction and identification

2.4.1 Tracks and primary vertex

The tracking and vertexing are based on information from the ID, and are important for the jet flavor tagging and thus for the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis.

2.4.1.1 Tracking algorithm

The main tracking algorithm follows an inside-out strategy [62]. A charged particle hitting a sensor of the ID fires a cluster of pixels which are first translated into a space-point. Within the pixel detector, one cluster provides enough position information of the hit position in three dimensional space while in the SCT space-points are defined using the clusters from both sides of a strip module. The so-called seed tracks are then formed as helix trajectories, each connecting three space-points in either the pixel detector or SCT. The space-points can also be paired and used to compute an inclusive estimation of the primary vertex position, called beamspot. Under the constraint of the beamspot, the seeds are grown into candidate tracks adding in more space-points through the use of a combinatorial Kalman filter.

At this stage, a very high number of track candidates are built but contaminated with a significant fraction of fake tracks. A score-based ranking scheme is employed to deal with

this issue. In this algorithm, the candidate track's quality is ranked based on several inputs. The number and type of hits in the track are important factors taken into account for the ranking. When a normal hit is associated to one track, a bonus is given to the score of this track. Sometimes a hit can also be associated to several tracks and is called shared-hit. This could be caused by the presence of fake track or when the detector granularity is insufficient to resolve close-by particles. A neural network is trained to differ these two cases [63]. Hits satisfying a quality cut on the NN output are identified as coming from multi-particle hits, labeled as split hits, while the others are labeled as merged hits. In order to enter the ranking, tracks are required to have at most one merged-hit which reduces the corresponding track scores.

Tracks passing the quality requirements on the score are then extrapolated towards the TRT associating the particle hit signal built in the TRT.

2.4.1.2 Vertexing algorithm

Multiple vertices can be reconstructed using a selected set of tracks and the beamspot. First, one vertex seed is formed. The x - and y - coordinates of the seed are defined as the center of the beamspot. The z coordinates of the points of closest approach from each track to the beam pipe are used to construct the longitudinal position of the seed vertex. Then the vertex reconstruction follows an adaptively iterative fitting approach. At each iteration, tracks are weighted according to their compatibility with the new seed in the fit, and then the fit is re-launched. Once the vertex is decided, the tracks being incompatible with the vertex at more than 7σ are removed and used to fit another vertex seed. This is repeated until all tracks are associated to a vertex.

The vertex with the largest sum of p_T of associated tracks is selected as the primary vertex, which is assumed to be from the hardest scattered proton pair. The others are treated as pile-up interactions.

2.4.1.3 Track parameters

After tracks are found through the above procedure, they are re-fitted to get their precise impact parameter. A minimum set of five parameters is necessary to describe helix tracks as shown in Figure 2.22:

- d_0 , called the transverse impact parameter, is the closest distance between the track and the beam axis in the x - y (transverse) plane.
- z_0 , longitudinal impact parameter, is the z -axis coordinate of the point, which is closest to the beamline, on the track.
- ϕ , the azimuthal angle, is defined as between the $\vec{p}_T$ and x -axis, ranging $0 < \phi < \pi$.
- θ , polar angle, is defined as between the $\vec{p}$ and the z -axis, ranging $0 < \theta < \pi$.
- q/p_T , the electric charge q [e] over the transverse momentum p_T [GeV], is calculated using the magnetic field strength B [T], and track curvature R_{curv} [m]⁻¹.

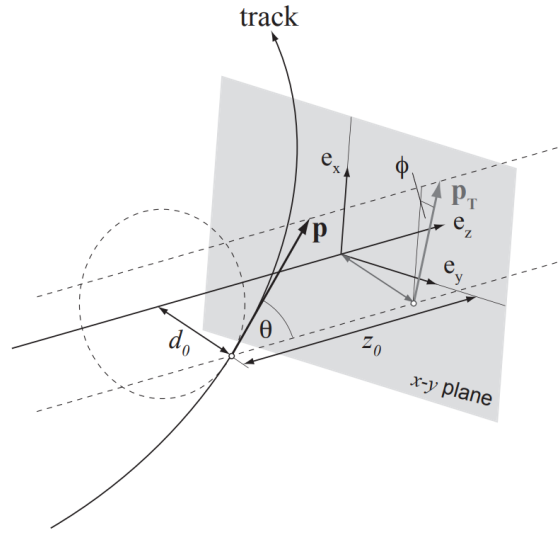


Figure 2.22: ATLAS track impact parameters [64].

2.4.2 Muons

Muons leave hits in the ID, deposit a small amount of energy and penetrate the calorimeters, and then produce signals in the MS. Thus the muon reconstruction [65] is based on the information from both the MS and the ID with additional information from calorimeters.

The process starts with the MS. In the MDTs, hits are translated into straight line segments per chamber on each MS layer. Combining these segments in all layers, track candidates are then built in the MS. In the next step, the MS tracks are extrapolated towards the primary vertex to build full muon tracks. This procedure involves the ID tracks which are reconstructed in the standard way described in Section 2.4.1.1, or calorimeter energy deposits:

- In most of the cases, muon candidates can be reconstructed by fitting hits from matched tracks of both ID and MS. This is referred to as *combined muons*.
- Muons not categorized as combined muons may have their tracks built as a combination of the ID tracks and at least one additional segmentation from the MS. This is referred to as *segment-tagged muons* which are usually muons with low p_T or traversing regions where the MS has reduced acceptance.
- If the ID tracks are not used in the above two scenarios and are in the region $|\eta| < 0.1$ where the MS has a gap, they are used to define muons tracks if they match energy deposits in the calorimeters. This is referred to as *calorimeter-tagged muons*.
- When the MS tracks cannot be combined with any ID information, the muon tracks are defined based on these MS tracks extrapolated inward employing the energy loss in the calorimeters. These muons are therefore called *standalone muons*. They tend to fall in the region $2.5 < |\eta| < 2.7$ which is not covered by the ID.

Once muon objects are reconstructed, higher quality muons are selected and then used to select interesting events for physics analysis. The identification and isolation are two important criteria.

Muon identification The identification procedure applied on the above reconstructed tracks is aimed to discriminate prompt muon tracks against other tracks induced by particles acting like muons by accident (fakes). Several identification working points, loose, medium, tight, high p_T , and low p_T , are provided, each having different requirements on basic muon qualities, like number of hits on the track, p_T and $|\eta|$. Different working points have different muon reconstruction efficiency and rate of fakes, and can be used in different analyses. Medium muons are used in the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis presented in this thesis. It requires combined or standalone muons. The former are further required to have at least 3 hits (1 hit and at most 1 hole) in at least two muon chambers for $|\eta| > 0.1$ ($|\eta| < 0.1$). The standalone muons are only used in the $2.5 < |\eta| < 2.7$ region. A further requirement is added on the ID and MS p_T compatibility. Figure 2.23 shows reconstruction efficiency for the medium muon selection as a function of p_T .

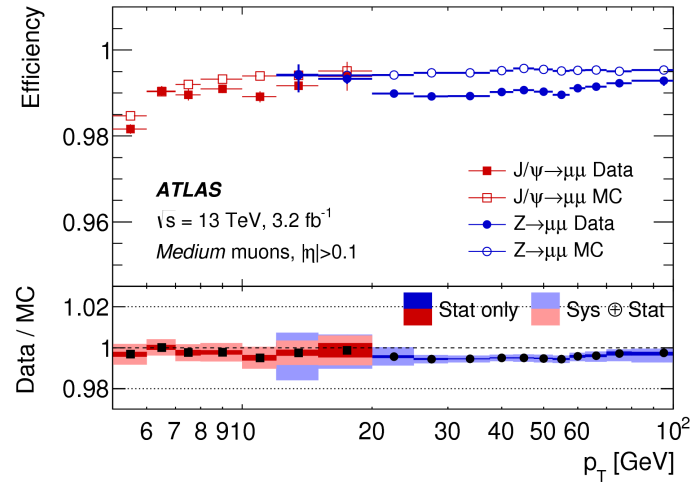


Figure 2.23: Reconstruction efficiency for the Medium muon selection as a function of the p_T of the muon, in the region $0.1 < |\eta| < 2.5$ [65].

Muon isolation The isolation procedure [65] is aimed to reject muons which come from heavy-flavored decays in jets, referred to as non-prompt muons. Tracking isolation is based on the isolation variable $p_T^{\text{varcone30}}$, defined as the sum of the transverse momentum of the tracks in a cone of size $\Delta R = \min(10 \text{ GeV}/p_T^\mu, 0.3)$ around the muon candidate. The calorimeter isolation is based on $E_T^{\text{varcone20}}$, defined as the sum of the transverse energy in calorimeter clusters in a cone of size $\Delta R = 0.2$ around the muon candidate. Several working points are provided by adding requirement on the relative isolation variables which are defined as the ratio of the track- or calorimeter-based isolation variables to the p_T of the muon. The *Gradient isolation* is used in this analysis, where the isolation efficiency is required to satisfy more stringent criteria as the lepton p_T drops.

2.4.3 Electrons

An electron leaves a track in the ID, reaches and finally stops in the ECal depositing all its energy there. Photons interact with the sub-detectors in a similar way, except without leaving a

track in the ID as neutral particles if it does not convert into e^+e^- . Thus the information from the ID and ECal are used to reconstruct electrons and photons.

Firstly, clusters of the EM cells with energy deposit are searched [66]. Each cluster seed is of size corresponding to $\eta \times \phi = 3 \times 5$ cells in the middle layer of the ECal. A sliding window algorithm [67] is used to scan the full ECal acceptance and the cluster seed is registered if their total energy deposit is larger than the background noise.

In parallel, the tracking built in the standard ATLAS algorithm is considered. The EM clusters are matched to those tracks by considering the η and ϕ distance between the track and the center of the cluster. If no ID tracks can be matched, the EM cluster is registered as unconverted photon candidate. If the cluster is matched to two tracks which are compatible with a photon conversion secondary vertex, it is saved as a converted photon. Otherwise an electron candidate is registered. The photon does not enter the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis in this thesis.

Similarly to muons, the reconstructed electrons are further selected to have high quality for physics analysis, adding criteria of identification and isolation.

Identification Electron identification employs a likelihood method [66] to identify the prompt electrons as signal and reject as background the hadrons and electrons from heavy-flavor decays in the jet or electrons from photon conversion that pass the cluster-track step. The method uses as input the information of the shower shape, track properties and matching quality between track and cluster, and gives an output discriminating the prompt electrons against others. By cutting on a fixed threshold of the likelihood discriminant, four working points are defined, VeryLoose, Loose, Medium, and Tight with increasing threshold and background rejection. The signal selection efficiency of three working points are shown in Figure 2.24 with respect to both E_T and η , measured with $Z \rightarrow ee$ and $J/\Psi \rightarrow ee$ events. The efficiency is degraded in the transition region of $1.37 < |\eta| < 1.52$ between the barrel and end-cap ECal where there is more inactive material. The electrons used in the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis in this thesis are required to pass the Tight working point.

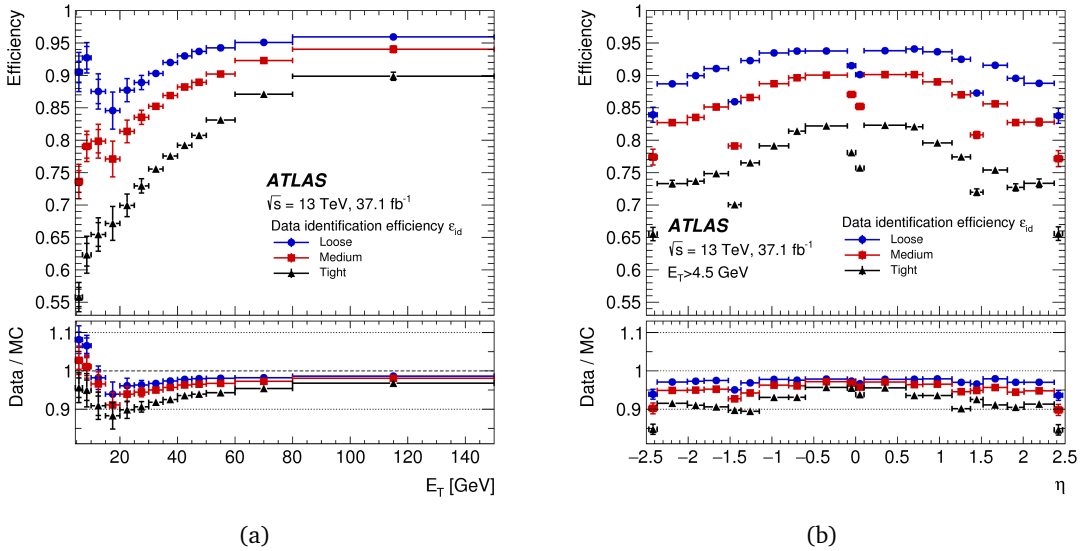


Figure 2.24: Efficiency of electron likelihood identification working points with respect to (a) E_T and (b) η [66].

Isolation The isolation requirement on electrons aims to further disentangle the prompt electrons from non-prompt candidates. And in analogy to muons, isolation variables based on track and calorimeter quantify the energy of the particles produced around the electron candidate [68]. The electrons are required to pass the Gradient isolation criteria, which is then used for the final event selection of channel studied in this thesis.

2.4.4 Jets

The jet, as spray of collimated hadrons, comes from the hadronization of a quark or gluon as soon as they are produced. In the detector, it leaves sprays of tracks and pockets of energy in the calorimeters. The jets used in the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis in this thesis are reconstructed with the anti- k_t jet algorithm [69] implemented in the FastJet package [70].

The HCal cells with significant energy deposits compared with noise are grouped to seed three-dimensional topological energy clusters (topo-clusters) [71]. Then the sequential anti- k_t jet-finding algorithm combines these topo-clusters (or any objects in other scenarios) to form conical jets of radius R . Jets reconstructed this way are referred to as calorimeter jets. $R = 0.4$ for the jets used in this thesis, while jets with larger radii of $R = 1$ could also be used in the analysis where boosted neighboring jets coming from energetic decays could overlap and form a large- R jet.

The anti- k_t algorithm gives the order to combine clusters based on the distance defined as d_{ij} between clusters i and j , and d_{iB} between the cluster i and the beam line:

$$d_{ij} = \min\left(\frac{1}{p_{Ti}^2}, \frac{1}{p_{Tj}^2}\right) \frac{\Delta R_{ij}^2}{R^2}, \quad (2.6)$$

$$d_{iB} = \frac{1}{p_{Ti}^2}. \quad (2.7)$$

Cluster i is merged with j if they have the minimum d_{ij} , and is registered as a jet if d_{iB} is the minimum. This process is repeated until all objects have been assigned to a jet. This algorithm has the advantage of giving stable shape and properties of the reconstructed jets which are not easily impacted by the appearance of soft radiation around (infrared safety) and are insensitive to the appearance of boosted collinear jets (collinear safety).

Instead of topo-clusters, various objects can be used as inputs to the anti- k_t algorithm. The inner detector tracks can be used and the jet reconstructed this way is referred to as a track jet. Both the calorimeter jets and track jets are reconstructed based on the detector response and are called *reco-jets*. Jets are referred to as *truth-jets* when being reconstructed using stable, final-state particles from MC generators as input. These candidate particles are required to have a lifetime of $c\tau > 10$ mm and muons, neutrinos, and particles from pile-up activity are excluded.

Jet energy calibration Due to the complex hadronic showers described in Section 2.2.3.2 and the experimental complications such as the non-compensating nature of the calorimeters, energy loss to dead material and extra energy from pile-up, the reco-jets require careful calibration such that the measured energy matches that of the original parton (referred to as particle-level energy scale) as close as possible.

Topo-clusters are initially reconstructed at the EM scale measuring the energy deposited by electromagnetically interacting particles. Then, reconstructed jets using topo-clusters go

through several stages of calibration. At first the jet is corrected to point back to the primary vertex and the extra energy coming from pile-up is removed. Then using truth jets in MC events [72] the absolute reco-jet energy scale (JES) is corrected to the truth jet energy. Further corrections are also implemented to account for the difference of the quark and gluon initiated jets using information from the calorimeters, the ID and MS as well. Moreover, the global sequential calibration is implemented using calorimeter, MS, and track-based variables, in order to further improve the reconstructed energy and related uncertainties. Finally the *in situ* energy calibration is derived from data using well-measured reference objects [72] to account for mismodelings in simulations and their calibration steps.

***b*-tagging** Jets originating from the fragmentation and hadronization of a *b* quark and thus containing *B*-hadrons are referred to as *b*-jets. They can be identified against *c*-jets and *light*-jets (referred to as *b*-tagging). This identification performance is crucial to the analysis studied in this thesis searching for the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) process which has multiple *b*-jets in the final state.

The identification is mainly based on the properties of *B*-hadrons in *b*-jets: the long lifetime $\tau \approx 1.5$ ps, thus long flight path $c\tau \approx 450$ μm before decay, resulting in a secondary vertex as illustrated in Figure 2.25. The *B*-hadron can further decay to a *c*-hadron giving a tertiary vertex in the jet. Also the impact parameter of the constituent tracks stemming from the secondary vertex tends to be relatively large. All these features can be detected and reconstructed using the ID and the IBL in Run 2.

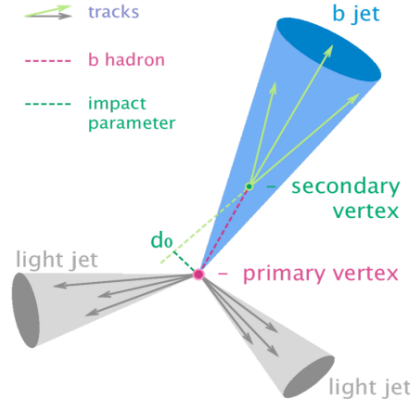


Figure 2.25: Schematic view of a *B*-hadron decay which produces a secondary vertex with decay length L_{xy} and displaced tracks. The impact parameter d_0 is shown for one of the displaced tracks.

The *b*-tagging algorithm, used in the analysis in this thesis, is the MV2c10 *b*-tagging [73] which shows how possible a jet contains a *b*-hadron, and it was developed at CPPM. This two-stage algorithm exploits the Boosted Decision Trees (BDT) technique which combines the variables provided by three other basic *b*-tagging methods: (1) a likelihood-based combination of the transverse and longitudinal impact parameters; (2) an algorithm reconstructing a single inclusive secondary vertex; (3) another approach reconstructing two vertices in the whole *B*-hadron decay chain searching for a common direction connecting the primary vertex to both the secondary bottom and the tertiary charm decay vertices. The MV2c10 *b*-tagging used in

this thesis is trained on simulated $t\bar{t}$ events to separate the b -jet signal from the backgrounds consisting of 7% c -jets and 93% light-jets [74]. The output is shown in Figure 2.26.

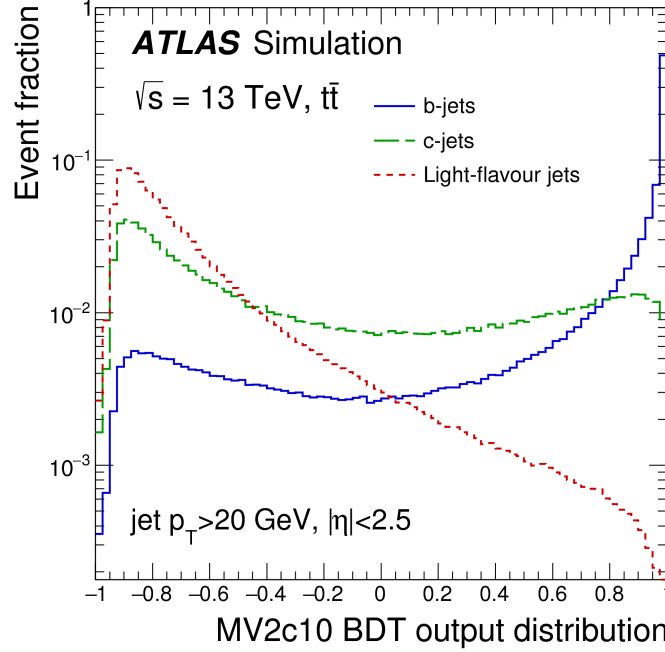


Figure 2.26: The MV2c10 output for b -jets (solid blue line), c -jets (dashed green line) and light-jets (dotted red line) in simulated $t\bar{t}$ events [73].

By cutting on the *b-tagging weight* (MV2c10 score in this thesis) at a threshold, any jet with a b -tagging output higher than the threshold is tagged as b -jet. A threshold corresponds to a *b-tagging working point* (b -tagging WP) which is chosen to provide to a b -tagging efficiency. The b -tagging performance is expressed in terms of c -jets and light-jet rejection in $t\bar{t}$ simulation samples, defined as $1/\epsilon_c$ and $1/\epsilon_{\text{light}}$, respectively, where ϵ_c (ϵ_{light}) is the fraction of c -jets (light-jets) passing the threshold of the corresponding WP. Four WPs are supported by the ATLAS Flavor Tagging Group, corresponding to 85, 77, 70, 60% b -jet efficiency in the above mentioned $t\bar{t}$ simulation. The latest performance is shown in Table 2.1 The b -tagging is calibrated independently for each WP to correct the b -tagging efficiency in simulation to data. To to this, data-MC scale factors (SF) are estimated as the ratio between data and simulated efficiencies and are applied on the MC result.

BDT cut value	b -jet efficiency	c -jet rejection	Light-jet rejection
0.9349	60%	34	1538
0.8244	70%	12	381
0.6459	77%	6	134
0.1758	85%	3.1	33

Table 2.1: Working points for the MV2c10 b -tagging [74] algorithm. These values are extracted from $t\bar{t}$ events, the main requirement being jet p_T above 20 GeV and $|\eta| < 2.5$.

Previous $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analyses used a cumulative fixed WP, where a jet is tagged as b -jet if it passes the corresponding b -tagging score threshold. In the analysis studied in this

thesis, the *pseudo-continuous (PC) b-tagging* score is exploited, using the pseudo-continuous calibration combining the independently measured SFs for each WP. The full-range MV2c10 score distribution for each jet is used, but is split into five bins. The bin edges are the thresholds of the four WPs. The end points, -1 and 1, of the MV2c10 distribution are defined as 100% and 0% WP, respectively. These five bins are defined with integer values, ranging from 1 for the bin between 100% to 85% thresholds, to 5 for the bin between 60% to 0% thresholds.

2.4.5 Missing transverse energy

Weakly interacting stable particles can transverse and escape the detectors without leaving signals, such as neutrinos, and some neutralinos hypothesized by SUSY. Thus direct detection with the ATLAS detector is not possible. However, proton pairs are accelerated and collided along the z -axis, and the total momentum in the transverse plane carried by partons is thus negligible and assumed to be zero. With momentum conservation, the missing $\vec{p}_T$ (with magnitude missing transverse energy E_T^{miss}) can be measured as the negative sum of the $\vec{p}_T$ of detectable particles.

$$E_T^{\text{miss}} = - \left\| \sum \vec{p}_T \right\| \quad (2.8)$$

The p_T summation contains two contributions [75]. The first one is from the fully reconstructed and calibrated objects including electrons, photons, τ -leptons, muons, and jets. The second contribution, referred to as *soft term*, is from the reconstructed charged-particles tracks in the ID associated to the primary vertex but not included in any reconstructed objects.

Synchronizing the ITk geometry description for simulation and radiation study 3

3.1 The High Luminosity LHC

The High Luminosity LHC (HL-LHC) program, approved in 2016 [76], is the upgrade of the current LHC with a much higher luminosity. As shown in Figure 3.1, this new phase (Phase 2) of the LHC operation will start data taking from 2026, aiming to collect a dataset of 3000 fb^{-1} by 2037 which is 10 times larger than the LHC originally designed integrated luminosity of 300 fb^{-1} collected from 2011 to 2023. While several possible future large-scale projects are being studied to reach higher energy [77], HL-LHC is the top priority for the exploitation of the full potential of the LHC.

The peak levelled instantaneous luminosity is foreseen as $7.5 \times 10^{34} \text{ cm}^{-2}\text{s}^{-1}$ for a further 10 years of operation from 2026, while it just touched $2.1 \times 10^{34} \text{ cm}^{-2}\text{s}^{-1}$ before the end of Run 2 [78]. To achieve this, during the second and third long shutdowns (LS2 and LS3), new pioneering technologies will be implemented to get more focused bunch, ending up with higher pile-up.



Figure 3.1: LHC/HL-LHC Plan (last update May 2018) [79]

Brought with the HL-LHC were not only opportunities, but also challenges for detectors. One challenge comes from the intense radiation damage on detectors, due to the high integrated

luminosity. The main source of radiation damage is from the large dose of particles produced in the pp collisions and their interactions with detectors. For the silicon tracking detector, the radiation could cause inelastic displacement of nuclei which can change the bulk electrical properties of the silicon, ending up with lower signals, larger power dissipation, and then cooling challenge. Ionizing effects on read-out chips may as well lower or even corrupt the signal. On the other hand, the augmented pile-up events are also challenging. An increased amount of data need to be read out in each bunch crossing. The current Inner Detector read-out bandwidth would easily be saturated due to high instantaneous luminosity. Also the higher number of hits will illuminate most pixels with current limited granularity (high occupancy), leading to mis-measured or misidentified tracks.

Accordingly, ATLAS upgrade studies are carried out. The basic goal is to maintain the excellent performance of the detector in terms of efficiency, resolution, and background rejection for all the physics objects used in the data analysis [80]. To be compatible with high data rate, a new two-level hardware trigger architecture is being developed which uses calorimeter, muon and track information. Thus all the trigger electronics of the calorimeters and muon spectrometer require upgrades to cope with high trigger rates. Moreover, a new all-silicon Inner Tracker (ITk) will replace the current Inner Detector. The layout, expected performance, and requirements of the proposed ITk will be discussed in the next Section.

3.2 The ATLAS inner tracking system

The above mentioned challenges and limitations of the ID necessitate the complete replacement with ITk, an all-silicon tracker, with challenging performance requirements. High track reconstruction efficiency for different particle types, muon, pion and electron are required [81]. Tracking resolution close to the interaction point is extremely important for the performance of vertex reconstruction, jet flavor tagging and converted photons. Thus maintaining a high track resolution is necessary and requires minimising the amount of material of the pixel detector's inner layers. What is more, the tracking coverage will be extended to $|\eta| = 4$. This will benefit a wide range of physics analyses, for example vector boson fusion (VBF), which requires track reconstruction beyond the $|\eta| = 2.5$ of the current ID. The studies presented [82] show that apart from the increased physics acceptance, forward jet reconstruction, E_T^{miss} resolution and pile-up jet rejection could all benefit significantly from the extended coverage. Generally the ITk is required to deliver equal or better tracking performance, to be radiation hard and high-rate capable, and to be at low complexity and cost.

Still under development, the ITk design is evolving and has been presented in a series of documents, from the "Letter of Intent for the Phase-II Upgrade of the ATLAS Experiment" [83] and the "ATLAS Phase-II upgrade Scoping Document" [82], to the two Technical Design Reports (TDR) for ITk's two subdetectors, the strip detector one [81] published at the end of 2016, and the pixel detector one [84], published at the beginning of 2018. While the Strip Detector remains unchanged since its TDR, the Pixel Detector Layout is still evolving as of Summer 2019. This study is a contribution to the ITk pixel TDR, so the ITk geometry and design illustrated in this section is consistent with it.

The ITk layout and the zoom into the pixel system is schematically shown in Figure 3.2. The strip detector has four barrel layers and six end-cap petal-design disks, both having double modules each with a small stereo angle with respect to the beam line (radial direction) to add $z(R)$ resolution in the barrel (end-caps), respectively. It covers only up to $|\eta| = 2.7$, and is complemented by a 5-layer pixel detector extending the coverage to $|\eta| = 4$. The pixel and

strip detector volumes are separated by a Pixel Support Tube (PST). Confronted with the harsh radiation, the inner two pixel layers are designed to be replaceable. They are separated from the outer three layers by an Inner Support Tube (IST). Each pixel layer consists of the endcap ring section, barrel flat section, and the barrel inclined section at the transition region between them in order to be more orthogonal to the particle trajectory. The positions and size of the ITk important volumes are shown in Figure 3.3.

Since this study contributes to the pixel detector TDR, the following detailed introduction to ITk subcomponents will focus on the pixel subsystem.

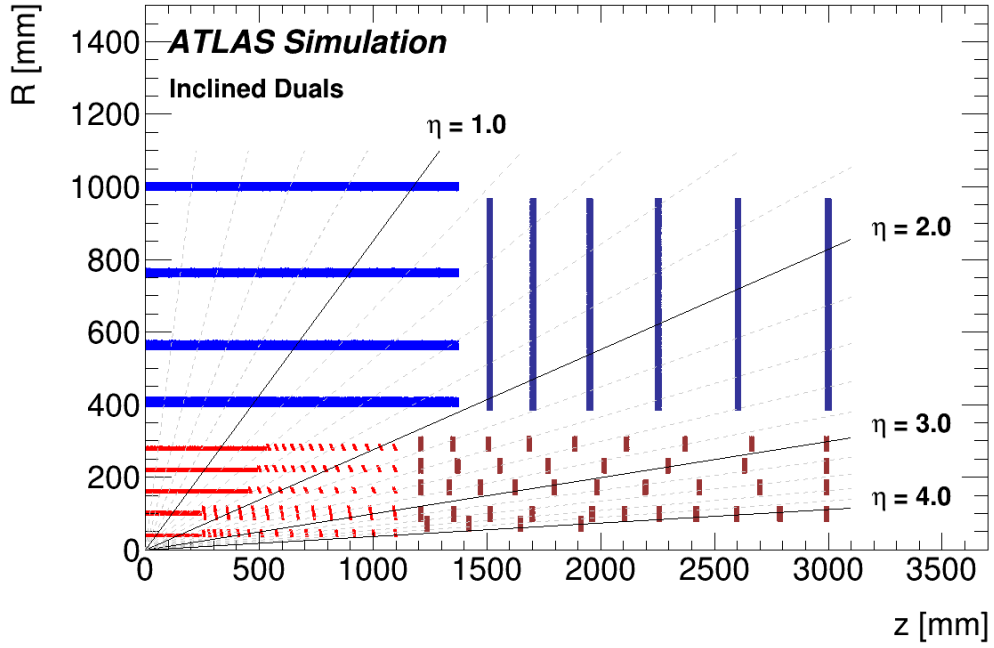
3.2.1 Pixel Module

The ITk pixel detector mainly adopts hybrid pixel modules, a similar design to the present ATLAS pixel detector and Insertable B-Layer. This kind of module contains a bare module consisting of silicon sensor and a front-end read-out chip joined together using a high density connection technique called “flip-chip bump-bonding”. A schematic view of a bare pixel module is shown in Figure 3.4. To improve intrinsic resolution and two-track separation, the pixel size has been reduced to $50 \times 50 \mu\text{m}^2$ or $25 \times 100 \mu\text{m}^2$. The largest bare module (so as the largest sensor) could reach around 4 cm^2 in area, with 4 read-out chips bump-bonded to the sensor. This is called quad module which will be used in the outer flat barrel layers and the outer end-cap rings. Smaller sensors are bump-bonded to a single (single module) or two (dual module) same-sized chips. The former will be used in the innermost barrel inclined section, and the latter in the innermost barrel layer and the inclined part of the outer barrel. This layout from the ITk Pixel TDR is referred to as the “Inclined Duals” layout. A flexible printed circuit board, called a module flex, is glued on top of the modules and interconnected to the bare module via wirebonds.

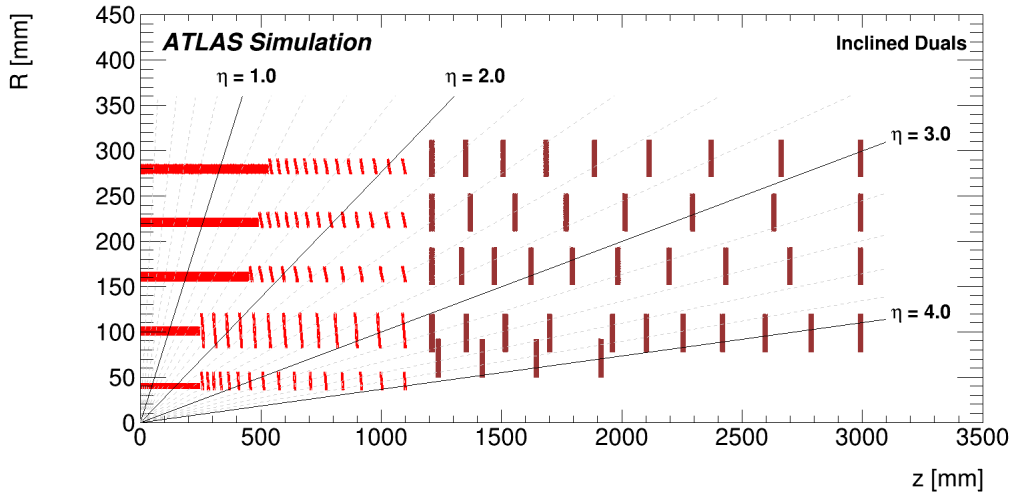
3.2.2 Local support structure, services and cooling system

Local supports provide mechanical support, alignment, electronic services routing and thermal management of the pixel modules. They are made of low mass, high rigidity and high thermal conductivity materials. All the services are routed out along staves or rings, away from the beam line to the end of the structure (EOS) as soon as feasible. Thus the number of cables will increase along the stave towards its ends. Services laid out in various sections are interconnected by patch panels (PPs), including PP0 (edge of structure) located at the end of the stave (in the outer rings system). From there, the cables run to PP1 (at the end of the Pixel package), and then further through PP2 (in the region of the muon chambers) until the services caverns where the power supplies will be located. For the cooling system, titanium cooling pipes are embedded in the support structure for CO_2 coolant circulation, to allow heat extraction.

The local supports vary for different pixel detector subsections. The Outer Barrel consists of so-called “longerons”, 2 m long staves with truss structure, oriented in the beam line direction. Each stave sustains both the central part flat modules parallel to the beam direction and the inclined modules farther away from the collision point. Figure 3.5 shows its schematic representation. The straight cooling pipe runs along the stave, as shown in Figure 3.5, and it cools the module via a highly conductive cooling block and a pyrolytic graphite plate to which the pixel module is attached. This structure is used in both the flat and inclined regions. The end-caps support structure comprises three pairs of concentric carbon half-rings, as shown in Figure 3.6. For each half-ring, modules are mounted on alternating sides of the structure



(a)



(b)

Figure 3.2: A schematic layout of the ITk Inclined Duals layout. (a): the whole ITk layout including the strip and pixel subsystems. (b): a zoom on the ITk Pixel Detector. The active elements of the strip detector are shown in blue. For the pixel detector the sensors are shown in red for the barrel layers and in dark red for the end-cap rings. Here only one quadrant and only active detector elements are shown. The horizontal axis is the axis along the beam line with zero being the interaction point. The vertical axis is the radius measured from the interaction region. The outer edge is set by the inner radius of the barrel cryostat that houses the solenoid and the electromagnetic calorimeter [84].

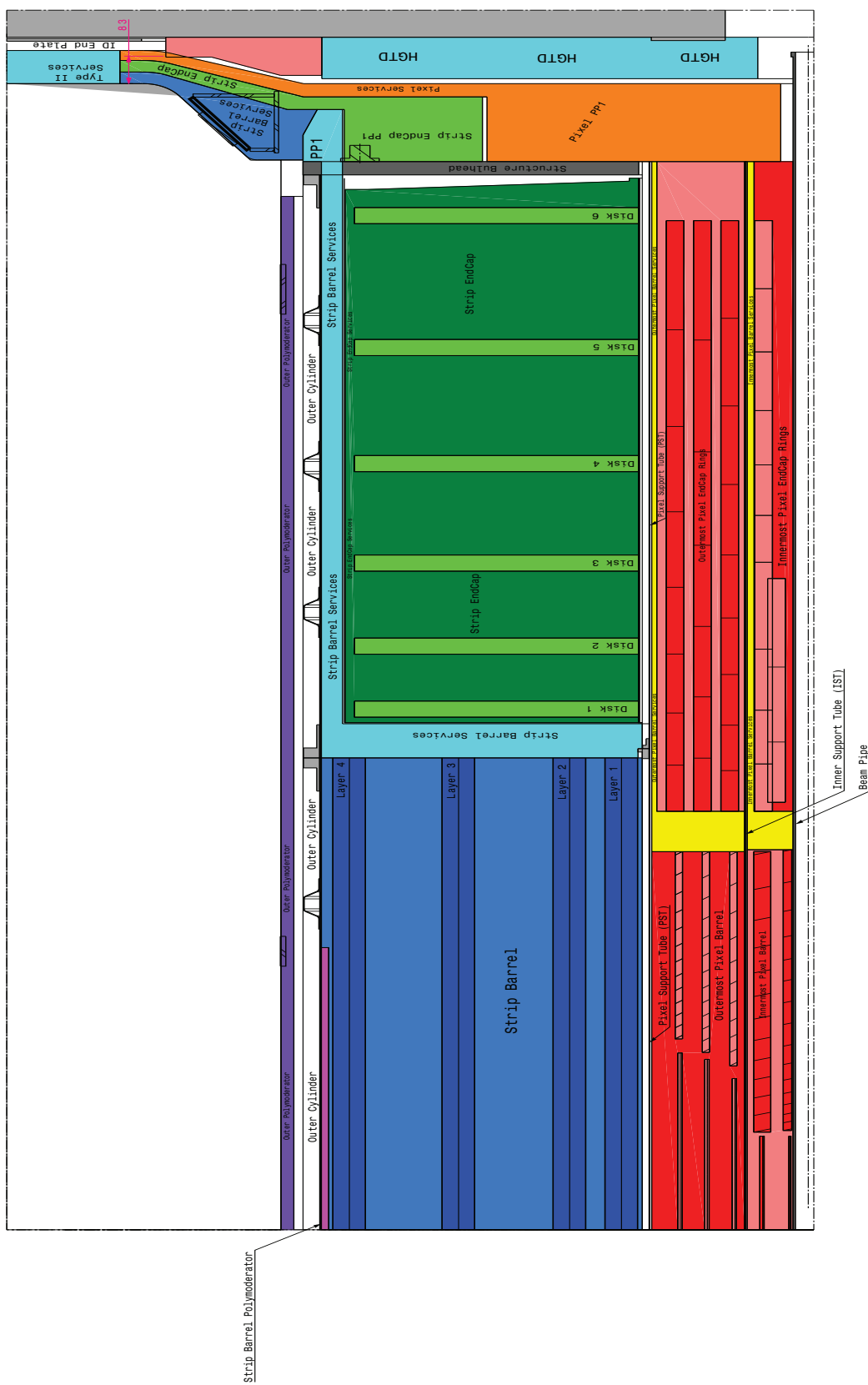


Figure 3.3: Envelope model of ITk, adapted from Reference [84].

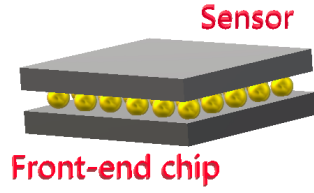


Figure 3.4: Schematic view of a bare pixel module, adapted from Reference [84].

within which is a cooling pipe embedded. For the two inner layers, two proposals are still under study as it is more difficult to accommodate more services in a more limited space due to the small radius. Generally both proposals consist of stave support for the flat modules, ring structure for end-caps and inclined modules in order to reduce the path to cooling pipes. The baseline layout used in the simulation study for the ITk TDR is shown in Figure 3.7 on the left, the “Inclined Inner”. This proposal uses a double-shell support structure (one shell for each of layer 0 and 1) with inclined rings on which are placed modules. An equivalent design is used for the 90° inclined rings used in the end-cap regions. An “Inner Alternative” proposal shown in Figure 3.7 on the right was also a promising choice although its simulation was not ready for the TDR. It uses a single-shell support structure which spans layers 0 and 1, with modules placed on it perpendicular to the beam axis.

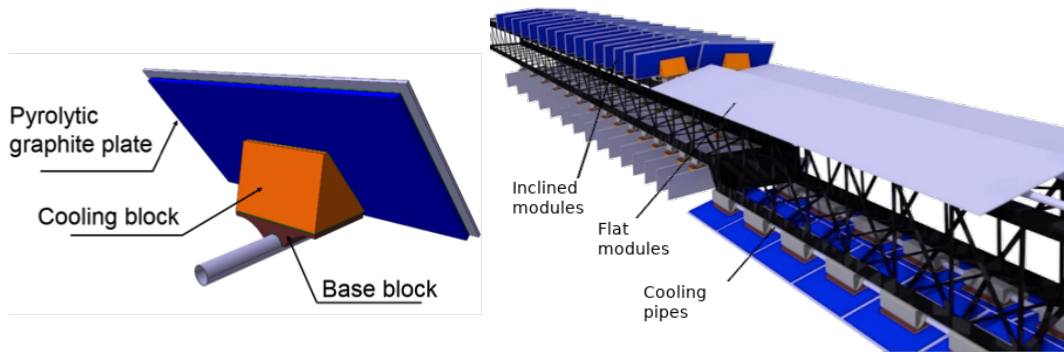


Figure 3.5: Left: Schematic representation of an inclined module mounting and cooling structure. Right: A typical longeron (truss construction) supporting four cooling lines [84].

3.2.3 Global support structure, moderator, and beam-pipe

The global support structure supplies mount for the local structure, and also separates various ITk subsystem volumes. The ITk package is assembled in the “Outer Cylinder” (OC), and at its extremity is enclosed and supported by “Structural Bulkheads” (SB) which also support the PST ends and PP1. These 3 structures define the gas vessel for the Strip system. The pixel detector is split by IST into two separately sealed volumes: the outer system, bounded between the PST and the IST, and the inner insertable system bounded by the IST and the VI (Vacuum Inner) beam pipe which is the same as the current Run 2. Thus the replacement of the insertable

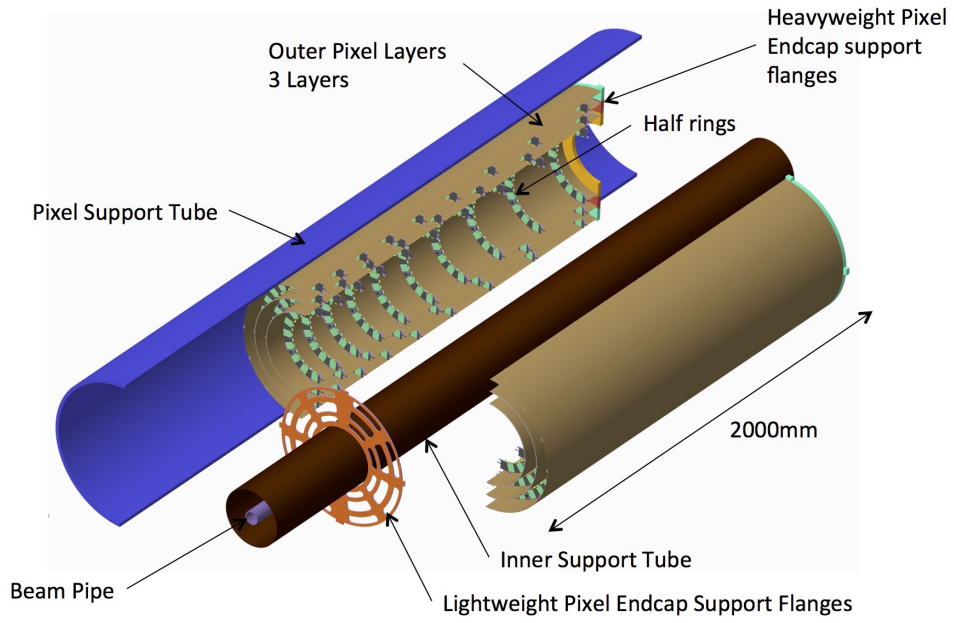


Figure 3.6: Exploded 3D drawing of one Outer End-cap [84].



Figure 3.7: Ring layouts in the “Inclined Inner” (left) and “Inner Alternative” (right) proposals. Modules are depicted as orange rectangles [84].

system does not disturb the sealed outer pixel layers. Dry nitrogen is flushed in these detector volumes to control humidity, thus mitigating galvanic corrosion and thermal effects. Same as for Run 2, around the ITk are installed borated polyethylene (BPE) moderators which reduce the neutron albedo from the calorimeter. Two pieces covering the barrel part are located at the inner OC and Inner Warm Vessel (IWV) of the calorimeter (outer OC). The other two are installed vertically on the ITk side of the calorimeter end-caps.

3.3 Simulation of the ITk detector

The layout evolves following an iterative design approach for its optimisation. For candidate layouts proposed to meet the requirements described in Section 3.2, in depth simulation and reconstruction studies are carried out with Geant4 [61] under the Athena framework, based on a dedicatedly implemented geometry. These performance studies may necessitate or inspire a new layout to some extent. In parallel, fluence and radiation dose simulation is conducted with FLUKA [85], based on a separately implemented geometry description, to make sure the design and its irradiation are within technical specifications and to estimate their damage effects. In these geometry descriptions, the dimensions, location and material of all detector elements are built corresponding to realistic engineering drawings although with necessary simplification.

3.3.1 ITk geometry in Geant4 for simulation

ITk detector simulation uses Geant4 to model the detector and particle interactions. This simulation is part of the simulated samples production procedure, which is the same as the ATLAS simulation based on the Run 2 offline software chain described in Section 2.3: event generation, detector simulation using Geant4, digitisation of simulated energy deposits, and event reconstruction. Normally the detector geometry for the Geant4 simulation is built via the GeoModel [86] database. For ITk, the simulation uses XML inputs to define and build the Geant4 models of real layouts, via the FastGeoModel framework, which allows to build the ITk pixel layout independently from the usual database system. Figure 3.8 shows a display of the overall ATLAS Phase 2 tracker ITk Geant4 geometry.

In the simulation geometry, each pixel module is implemented as an active sensor volume attached to a front-end chip. The front-end chip is modeled as a 150 μm thick silicon wafer, with a 1 μm thick copper layer to model its circuitry, and one 20 μm diameter Sn-Ag bump-bond per pixel channel. Throughout each volume, the material is homogeneously distributed. The pixel sensors are emulated as volumes with silicon in them. Modules are then placed on the different local support structures.

The outer pixel barrel is modeled as shown in Figure 3.9. The longeron truss structures are approximated by modeling them as thin sheets of carbon fiber. The four rails at each stave's four corners (containing 80% of the stave mass) are individually modeled. The total mass agrees with the engineering estimation. Services and cooling pipes mixed in one volume are running inside the longeron. For each inclined support structure, a carbon foam wedge is used as simulation to support modules. The inner barrel support structure is modeled as carbon fibre truss double shells and is approximated by one sheet of carbon fibre behind each row of modules. The layer 1 inclined quad module supports, deeply-angled at 75°, are modeled as blocks with more conservative material budget to match the corresponding engineering design. In layer 0, the inclined section mouting blocks are the same as the outer barrel. The outer pixel endcaps are modeled as rings. Each layer of rings is supported by a 0.4 mm carbon fibre

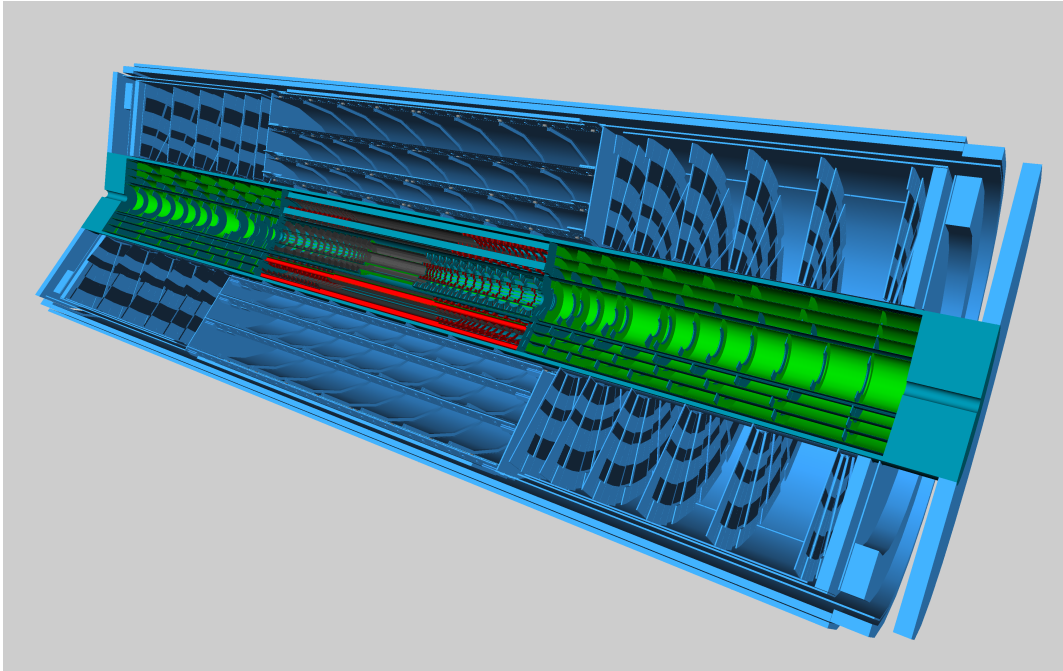


Figure 3.8: Display of the ATLAS Phase 2 Inner Tracker ITk Geant4 geometry with the Inclined Duals detector layout.

cylinder shell, along which services for the rings are routed. Each of these rings is constructed from two half-rings as in the real case. For the inner pixel end-caps, the same ring structures are used to model the detector. Figure 3.9 shows the geometry model of the inner and outer end-cap ring layers.

PP1 is modeled as a solid, homogeneous 40 kg cylinder at each end of the ITk pixel detector. The PP1 for the strip detector is modeled for each barrel end, and each end-cap end. A poly-moderator is modeled as 25 mm thick cylinder along the outer part of the cryostat bore that houses the solenoid and the electromagnetic calorimeter. A second piece, 20 mm thick, is modeled inside the outer support cylinder, while moderator disks with thickness 70 mm are placed in front of the forward calorimeter volume. Although mixed with services, the titanium cooling pipes material distribution are taken care of to be compatible with the engineering estimates, and so are the services. Its cable numbers, masses and chemical composition are well calculated at different z positions to be consistent with engineering drawings.

3.3.2 ITk geometry in fluence and radiation dose studies

The FLUKA [87] framework is used to study the radiation simulation. The standalone FLUKA code is well-established for studies of hadronic and electromagnetic cascades induced by high-energy particles and can transport particles over a wide energy range. More simplified than the Geant4 geometry, the layout description of FLUKA is intended to be equivalent to it in order to do the proper radiation estimation, as a separately implemented description. Shown in Figure 3.10 are the fluence and dose distributions for the pixel detector simulated with FLUKA.

The Radiation Simulation Working Group uses different frameworks for various detector subsystems. Apart from FLUKA, FLUGG (FLUKA + Geant4 Geometry) [88], an interface between G4 and FLUKA physics is used for muon systems. The GCALOR [89] and Geant4

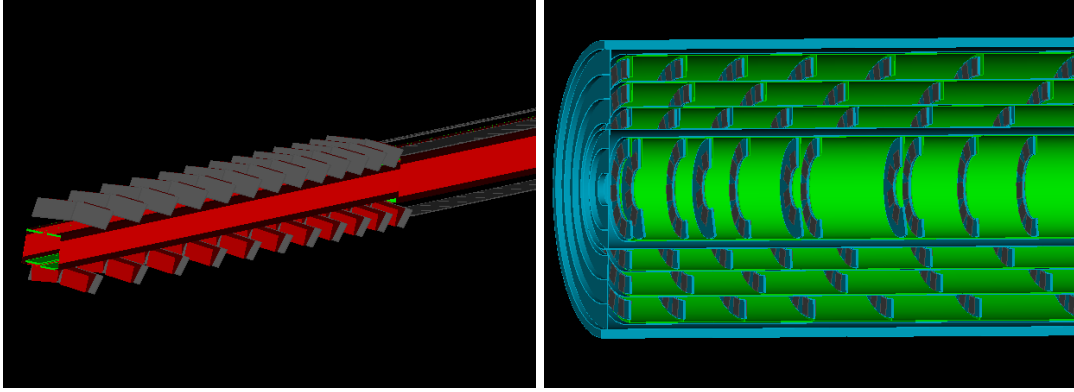


Figure 3.9: Left: A display of the Geant4 geometry model of the outer barrel longeron stave with inclined and flat modules mounted and services running inside. Right: A display of the Geant4 geometry model of the inner and outer end-cap shell structures with its ring supports for modules.

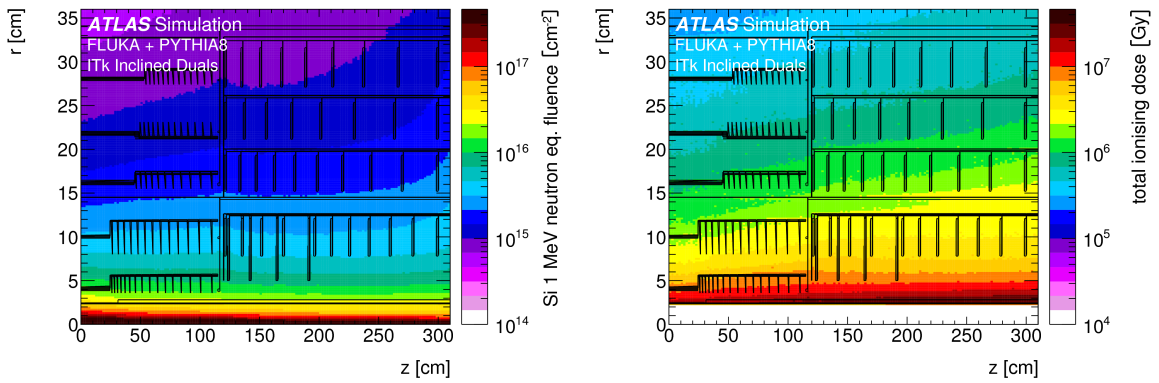


Figure 3.10: The fluence and dose distributions for the Pixel Detector. Left: 1 MeV neutron equivalent fluence. Right: Total ionising dose [84].

studies have focused more on radiation background issues in the calorimeter systems. Their native geometry description formats are developed in order to provide more flexibility for these individual studies. To facilitate the physics validation, visualization and debugging, an exchangeable geometry description between different studies was put in place, using GDML (Geometry Description Markup Language). Implemented by the same author as the FLUKA geometry, the GDML ITk geometry is intended to be equivalent to the FLUKA one, and could be imported into FLUKA for simulation by FLUGG. Figure 3.11 presents the visualization of the ITk GDML layout. The radiation study geometry used in this section is the GDML description.

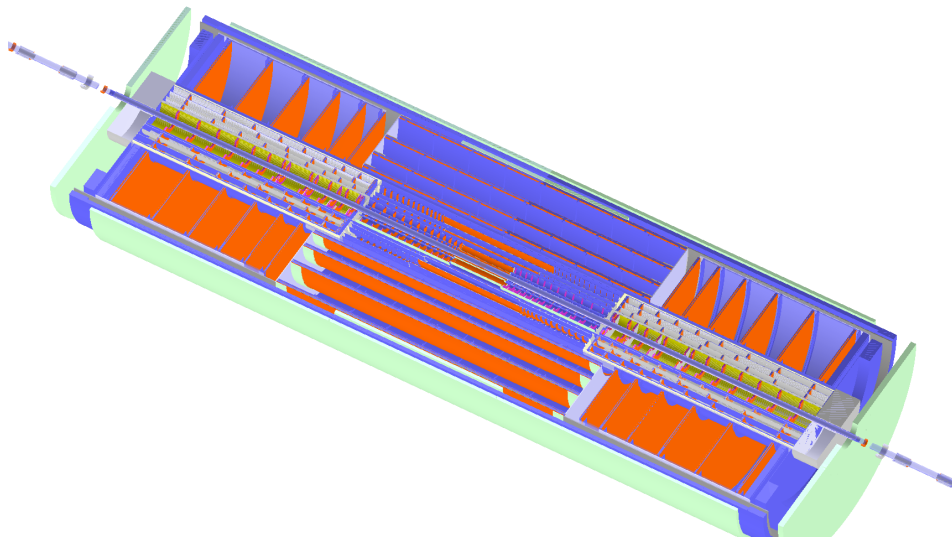


Figure 3.11: Display of the ATLAS Phase 2 Inner Tracker ITk GDML geometry with the Inclined Duals detector layout.

The GDML is simplified, while intended to be equivalent to the engineering drawings. Similarly to the Geant4 geometry, all sub-components are modeled as volumes. A material consisting of various elements with specific proportions is homogeneously distributed in each volume. For each of the pixel barrel flat sections, detector components are represented as concentric cylinders. In layer 0, closest to the beam pipe is the silicon sensor cylinder, and then closely attached to it, at larger radius, is the read-out chips, then support structure, Titanium cooling pipes and services come in order as going farther away from the beamline. Such structure composes the other layers but with the reverse order for layer 3. The inclined section is modeled with a similar structure but additionally with inclined annuli (cones) representing the inclined module mount attached to the support tubes. In end-caps, annuli are supported by cylinders consisting of cooling system tube, services tube, and then support cylinder, as radius increases. Each annulus has a sandwich structure: the support structure sandwiched by two chips annuli, and then two sensor annuli layers. Another annulus layer is attached to each ring on the outward facing side, modeling the EOS services and the cooling system. Figure 3.12 shows the zoomed-in views of pixel barrel and end-cap sections of the GDML layout.

3.4 Methods to synchronize ITk layout descriptions

The radiation study could address the radiation background issues and verify the radiation hardness of the ITk design. The correct study necessarily requires an equivalent input geometry

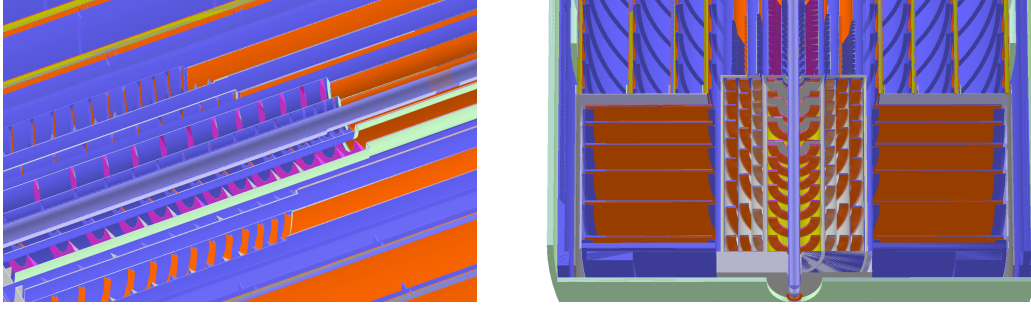


Figure 3.12: Left: A display of the GDML geometry model of the pixel barrel section. Right: A display of the GDML geometry model of the ITk end-caps.

to the Geant4 detector simulation layout description. As mentioned above, the geometries used by the two studies are independently implemented, and have a different level of simplification of engineering drawings. Thus it is highly motivated to study the equivalence between them.

The material distribution in the detector could efficiently reflect the geometry and the component deployment in a layout description. An X_0 mapping tool is already implemented based on Geant4 in Athena to show material distributions in several forms: radiation length ¹ (X_0), nuclear interaction length ² (L_0), and density. It also allows for the production of 2D plots ($R - z$, for example) that are very useful for layout debugging, and working out what materials are where in the ITk. The so-called “geantinos” are shooted in various directions. This kind of artificial particle does not interact with the detector and goes straightly from the shooting point to the outside. All submaterials properties of the detector geometry encountered by the geantino along its trajectory are recorded and used to produce the material distribution later. The study in this chapter is based on this tool. Figure 3.13 gives material distributions in both 1D and 2D produced by the mapping tool with Geant4 simulation geometry. One event in the histograms corresponds to one geantino event.

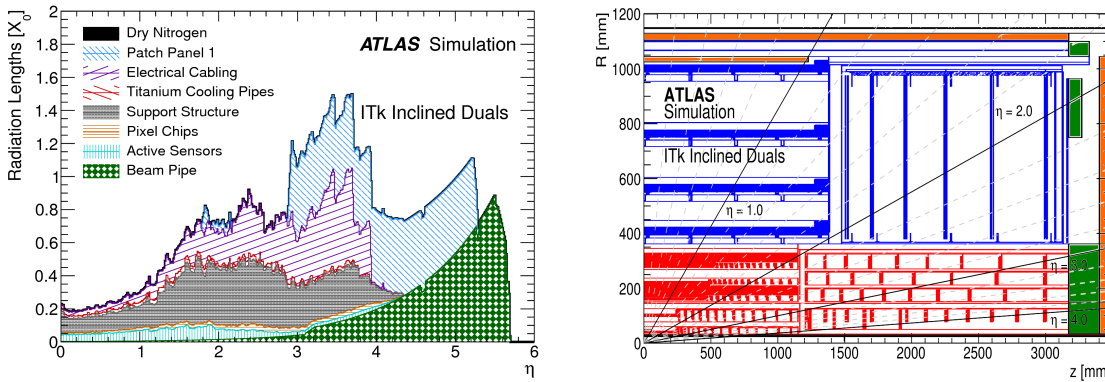


Figure 3.13: Left: Radiation length X_0 versus the pseudorapidity η based on Geant4 ITk geometry. Right: Location of the material for one quadrant of the Inclined Duals layout based on Geant4 ITk geometry [84].

¹Mean distance over which a high-energy electron loses all but $1/e$ of its energy by bremsstrahlung. Or $\frac{7}{9}$ of the mean free path for pair production by a high-energy photon.

²The mean distance travelled by a hadronic particle before undergoing an inelastic nuclear interaction.

Considering the framework capacity and availability of various geometry descriptions of the ITk at the beginning of this qualification task, three methods were considered. One potential approach is to use the existing mapping tool based on Geant4 by porting the GDML into Athena and Geant4. This kind of interface was started to be developed at the beginning of this task, but not publicly advertised by then. Another possible choice is to employ the tracking geometries which are simplified equivalents of full Geant4 geometries and are needed for tracking study and reconstruction due to much lower CPU time consumption. The tracking geometry is proposed to be exported from the Athena into GDML description. Then this transformed description could be fed into FLUKA studies via FLUGG to be compared with the local radiation geometry under the same framework.

The finally adopted proposal is to implant the Geant4-based mapping tool from Athena into FLUGG. Figure 3.14 shows the strategy of this approach and how the new implemented tools are connected with other frameworks. As FLUGG plays an interface among Geant4, FLUKA and GDML geometry, it could read in the GDML geometry and call the mapping tool to produce material distributions. To map the ITk Geant4 geometry, the ATLAS simulation in Athena calls the Geant4 simulation and generates the X_0 distribution information. Then the geometry comparison is facilitated by comparing the material distributions from both geometries.

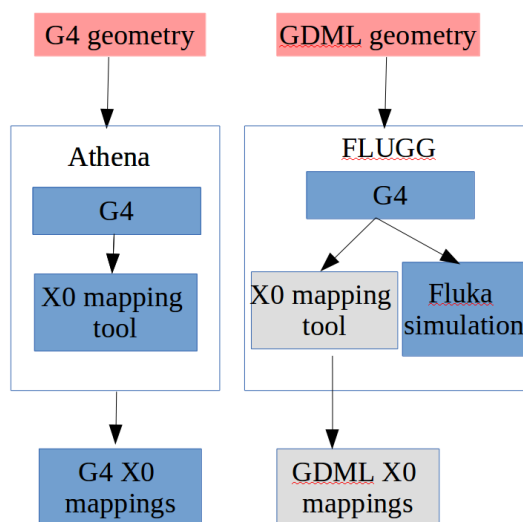


Figure 3.14: Workflow of the comparison method. Blue boxes are the frameworks used. Grey boxes are the functionalities implemented in this study.

Challenges exist for this task with the chosen approach. On technical aspect, it is not easy to port mapping code to Geant4 under the FLUGG framework where there is only a basic interface to ROOT histograms. In my implementation, similar to in Athena, the geantinos are produced as probe, but are generated on the fly as the material information is collected and recorded. Thus the produced geantino event statistics are limited due to the restricted memory and thus end up with slightly larger statistical fluctuations in the histograms of material distributions. Another difficulty comes from the difference of the two geometry descriptions and their levels of simplification. A substructure, for example, the titanium cooling pipes, could be included into a volume mixed with other subcomponents. And such a mixture could be different for

the two geometries and thus end up with different volume definitions. The span or position of a detector substructure, e.g., a pixel layer, could even be displaced from the engineering drawing due to gross but equivalent simplification. In order to compare the geometry in terms of eight categories as shown in Figure 3.13, close examination of each geometry about subcomponent volume naming conventions, and volume shapes and chemical compositions is necessary. Hundreds of volumes are defined in each geometry, and all of them need to be properly categorized. Material subtractions from volumes and transfers from one to another are calculated if necessary to follow this categorisation. Finally, when disparities are noticed in material distributions, the origin of such difference should be traced.

3.5 Layout description synchronization

Different layout versions (called steps) are proposed and studied during the ITk design iteration. Starting from the Letter of Intent layout, step 1.x development culminated in step 1.9 layout, a new reference layout with 5 pixel layers + 4 strip layers recommended by the ITk Layout Task Force. It also has inclined modules for the pixel barrel regions and four layers of pixel rings for pixel endcaps. And these ring structures are re-optimized to correspond to the inclined sections. The step 2.2 duals geometry is proposed with various updates, for example, a further inner ring layer was added to the inner pixel end-cap systems ending up with five pixel ring layers. The results for the step 2.2 duals geometry is used in the Pixel TDR. All the ITk design and modeling geometries described in previous sections are from the step 2.2 version. In this study, step 1.9 is used to develop tools without that in-depth geometry comparison, since the whole group was preparing to evolve to step 2.2 at the beginning of this study. Then detailed comparisons and discussions will be focused on step 2.2.

3.5.1 Step 1.9

3.5.1.1 Geometry simplification

The GDML geometries for radiation studies are much more simplified with step 1.9. In that description, each annulus pixel end-cap system is implemented as a single volume including many components: sensors, read-out chips, support structure, cooling pipes, and the EOS services. In the strip detector, similar level of component merging also exists in both barrel and end-cap systems. To follow the prescribed categorisation, these subcomponents need to be splitted. To achieve this, new volumes with finer granularity need to be defined, while the adaptation of relevant volumes around may also be invoked in order to be consistent with the overall engineering estimation. This requires work on geometry implementation and is beyond this qualification task. Thus, to cope with the current GDML definition, a simplified material category is used.

The following categories are adopted to do the geometry comparison study. All the strip detector materials enter into a category called strips, including its support structures, services, etc. This strip category is marked by the volume positions in z and R directions, considering that the R - z boundaries for each strip layer in GDML and Geant4 geometries are the same. Table 3.1 gives boundaries in R and z directions of the strip category. This definition is also supposed to include the two moderators in the barrel region while the moderator close to the calorimeter end-caps is in its own category called moderator. Then everything in the pixel end-cap system bounded by R - z limits, make up for another category, as defined in Table 3.2.

For the rest of the detector in the pixel barrel section, materials are divided into: chips and active sensor, electrical services with cooling system included, detector support structure, Patch Panel 1 (PP1), moderator, dry nitrogen, and the beam pipe.

Category	R [mm]	z [mm]
Barrel layer 0	$362 < R$	$ z < 1400$
Barrel layer 1	$515 < R$	$ z < 1400$
Barrel layer 2	$715 < R < 810$	$ z < 1400$
Barrel layer 3	$950 < R < 1035$	$ z < 1400$
Outer support cylinder	$R > 1050$	-
Endcap layer 0	$379 < R < 975.6$	$1495 < z < 1525$
Endcap layer 1	$379 < R < 975.6$	$1685 < z < 1715$
Endcap layer 2	$379 < R < 975.6$	$1938 < z < 1968$
Endcap layer 3	$379 < R < 975.6$	$2230 < z < 2270$
Endcap layer 4	$379 < R < 975.6$	$2584 < z < 2614$
Endcap layer 5	$379 < R < 975.6$	$2984 < z < 3014$

Table 3.1: Strip categories R- z definition

Category	R [mm]	z [mm]
Pixel Endcap 0	$34 < R < 122.15$	$1301 < z < 3174$
Pixel Endcap 1/2/3	$145 < R < 320$	$814 < z < 3174$

Table 3.2: Pixel end-cap categories R- z definition

3.5.1.2 Geometry comparison and synchronization

Differences between GDML and Geant4 geometries are noticed in the X_0 distributions. In Figure 3.15, the X_0 distributions of all categories stacked are shown for both geometries. A more spiky pattern in GDML is due to the lower statistics. There is more material in the GDML than in Geant4. Shapes and scales for various component categories are also inconsistent. Especially the moderator section is missing in the GDML. Figure 3.16 presents the category wise geometry comparison. Remarkable disparity exists in strip, support structure, and the services sections.

The 2-D X_0 plot for the strip category in R- z directions is checked to study the discrepancy, and presented in Figure 3.17 where only the edges of volumes are shown. It reveals that the moderator outside the outer cylinder is missing in the GDML. On the other hand, the vertical moderator in GDML is also found absent when checking the corresponding positions in this geometry file. These two moderators are expected to form the moderator category under the current material categorization. And due to their absence, the moderator category is missing. Shown in Figure 3.18 are moderators in Geant4, in which the outermost horizontal and vertical moderators are missing in the GDML file. The Radiation group was informed about these disparities, and proper geometries were confirmed in the full FLUKA geometry and were promised in the next version of the GDML geometry, step 2.2.

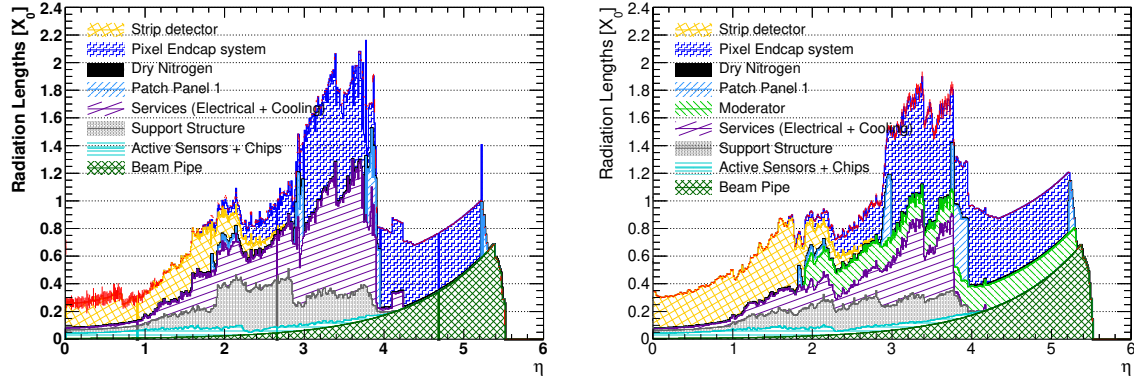


Figure 3.15: Radiation length X_0 versus the pseudorapidity η in step 1.9. The left plot is based on GDML with 5000 events, while the right one is based on the Geant4 geometry with 100k events.

All the support structures in ITk are presented in Figure 3.19. Extra vertical structures at $|z|=3152$ mm are found, and it comes from improper thickness of the disk volume with a value of 36 mm in the GDML geometry that models the strip end-cap bulkhead. The Geant4 geometry models it as 6 mm carbon fibre reinforced plastic (CFRP), and in the FLUKA geometry it is two 3 mm thick disks with 3 cm space in-between. They both match the proper amount of material but the GDML volume shape is closer to engineering drawings. What is more, the pixel end-cap ring support flanges, installed at $|z| = 1200$ mm throughout the five layers but not shown in the Figure, are found not modeled in the Geant4 geometry in step 1.9. Therefore they are also removed from the GDML in order to have a close comparison for other components.

The discrepancy in the service category between GDML and Geant4 is consistent with the Geant4 geometry mistake found by the simulation group for step 1.9. The amount of cabling routed around IST and PST should keep increasing towards the endpoint, as properly described in the GDML geometry shown in the left plot of Figure 3.20. An obvious increasing can be seen at $1200 < |z| < 3000$ mm at R around 145 and 341 mm. However, these cables stay unchanged in the Geant4 description shown in the right plot of Figure 3.20, ending up with underestimates of the real case. This bug will be fixed in step 2.2.

With all of the above disparities traced, the strip and support structure categories are shown in Figure 3.21. The moderators are removed from both GDML and Geant4 geometry for this plot while realising that they are partially missed in GDML (where the missing two moderators are treated as being outside of the tracker volume). Then the two geometries agree very well in the strip category. The improper strip endcap bulkhead disk in GDML is fixed, and the flanges of pixel end-cap rings modeling are removed from it. Then good agreement is achieved in the support structure section. All the above disparities are communicated with both simulation and radiation groups, and the corresponding updates will be assured in step 2.2.

With all category wise corrections, Figure 3.22 shows the overall material distributions except moderators, as a function of η for both GDML and Geant4 geometries. Apart from the services deficit in Geant4 which is traced, the overall and the category wise distributions are in good agreement. Figure 3.23 shows the location of the materials for one quadrant ITk of two geometries, including moderators. The more solid structures in the Geant4 geometry comes from its higher modeling granularity and the statistics.

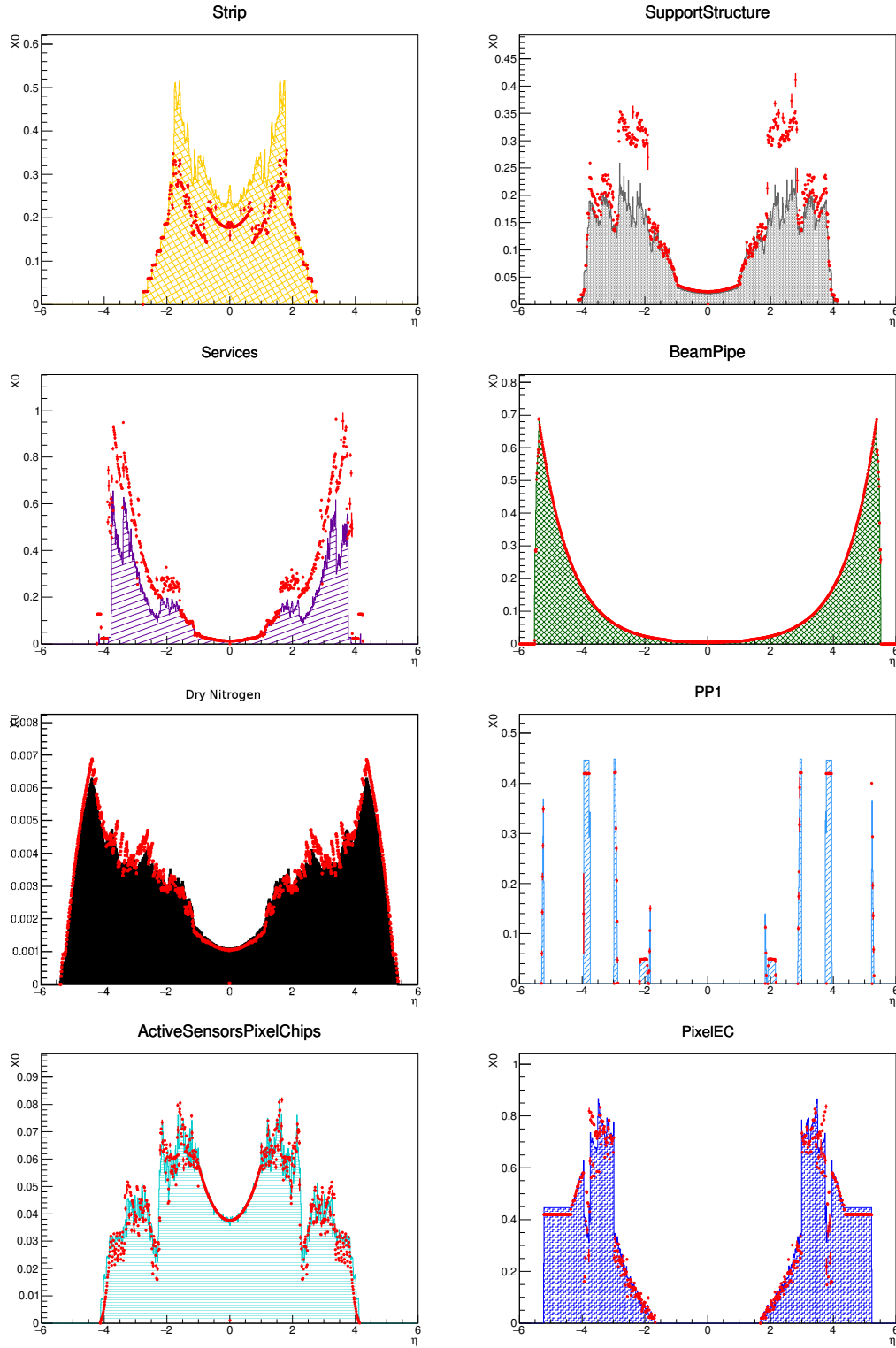


Figure 3.16: The category wise geometry comparison in step 1.9 between G4ML (red dots) and Geant4 (histograms).

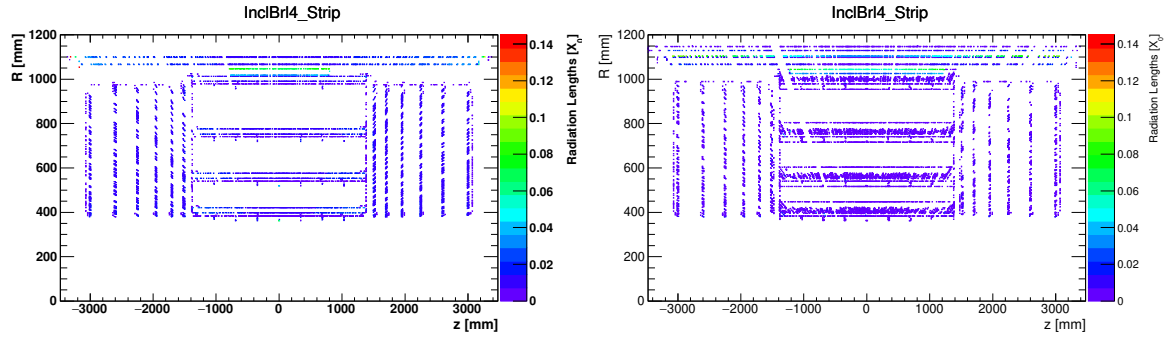


Figure 3.17: X_0 versus R - z directions for strip detector category in step 1.9. The left is based on GDML while the right on Geant4.

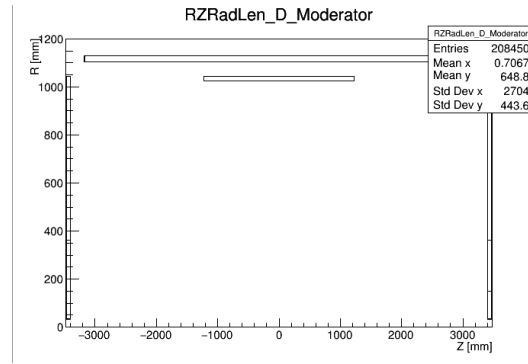


Figure 3.18: Moderator volumes in R - z space in Geant4 geometry for step 1.9.

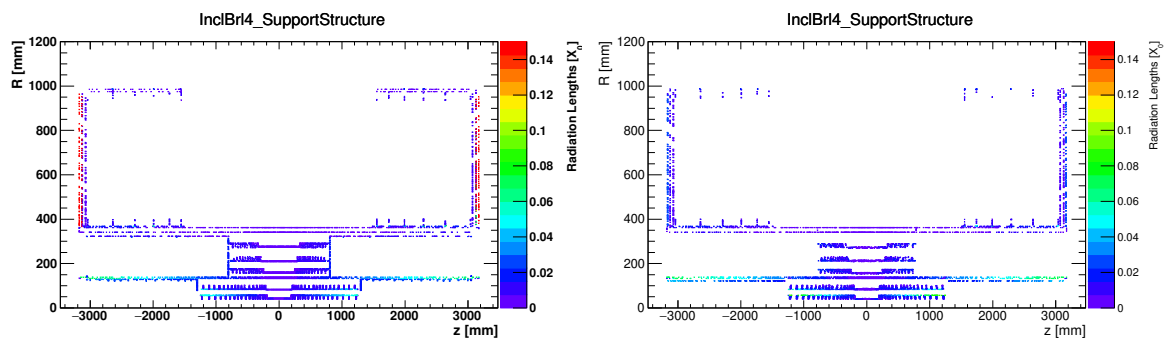


Figure 3.19: X_0 versus R - z directions for support structure category in step 1.9, for GDML (left) and Geant4 geometry (right).

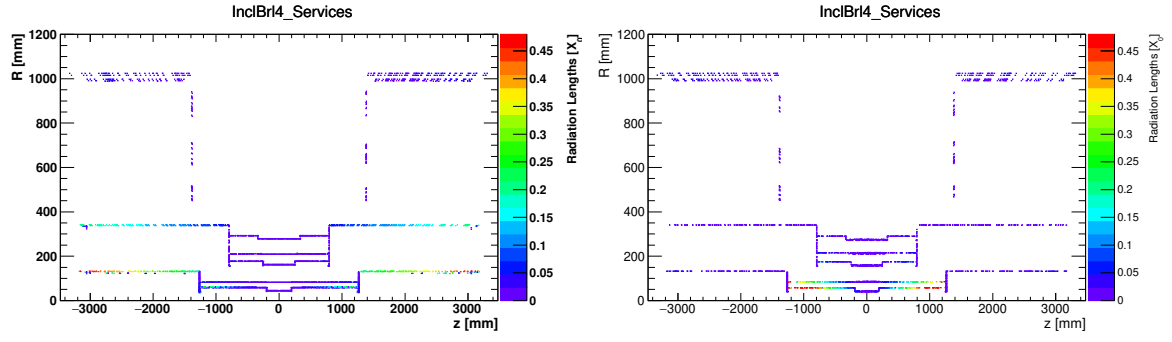


Figure 3.20: X_0 versus R - z directions for services category in step 1.9, for GDML (left) and Geant4 geometry (right).

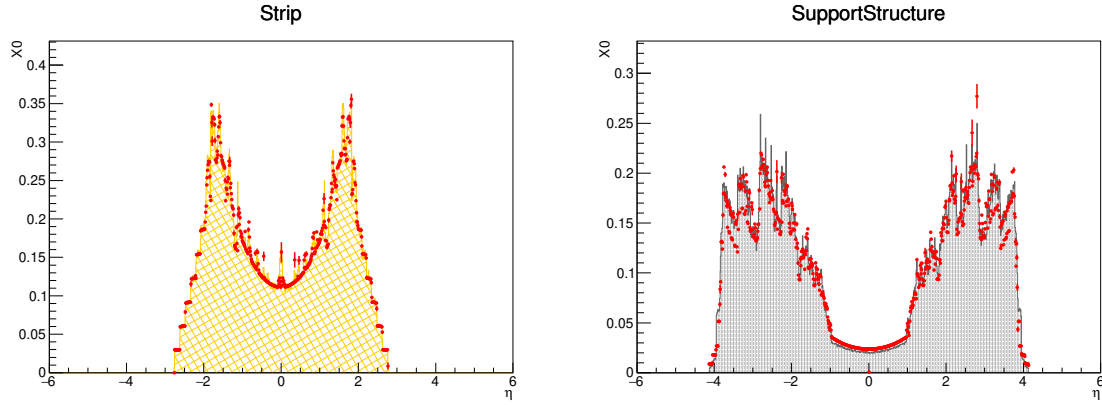


Figure 3.21: The category wise geometry comparison in step 1.9 after geometry synchronization, for strip category (left) and for support structure (right). The red dots are GDML-based distributions, while the histograms are based on Geant4 geometry.

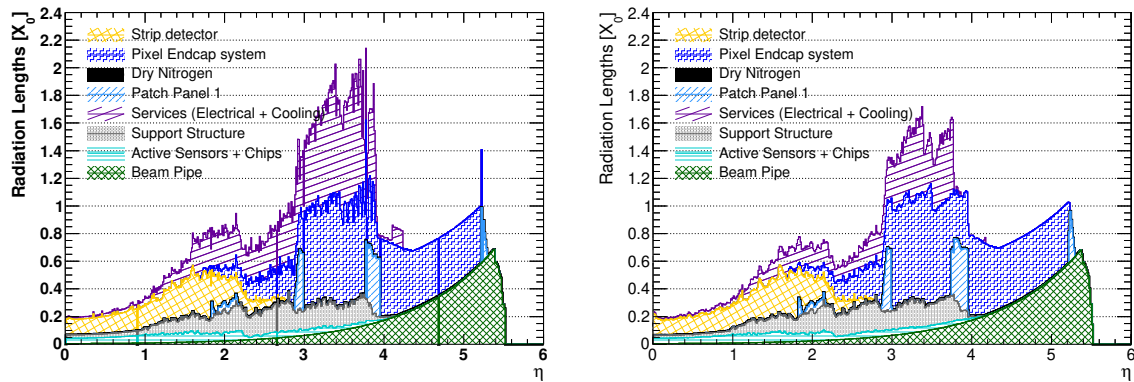


Figure 3.22: X_0 versus the pseudorapidity η after geometry synchronization in step 1.9 for GDML (left) and Geant4 (right).

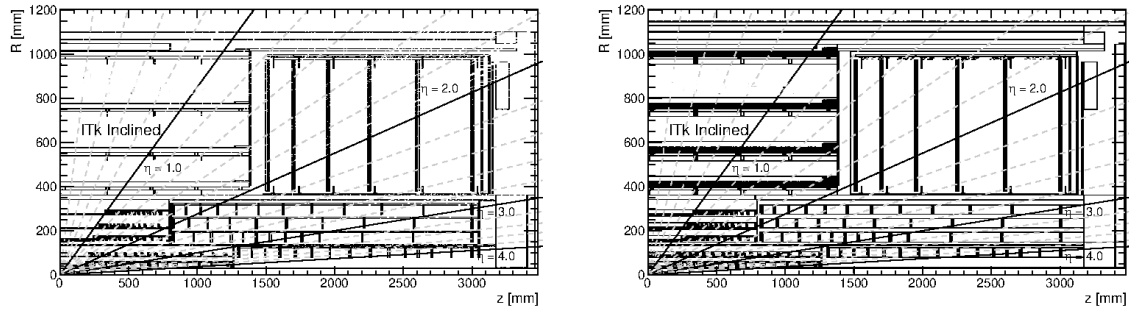


Figure 3.23: Location of material for one quadrant of step 1.9 after geometry synchronization for GDML (left) and Geant4 (right).

3.5.2 Step 2.2

3.5.2.1 Differences with step 1.9

The step 2.2 GDML and Geant4 are implemented with all of the above stated problems adapted. The service deficit in Geant4 is added back with proper estimation of cabling materials on different z positions. The moderators and strip bulkhead disks are properly modeled in the GDML file with this new version. What is more, the step 2.2 reverts to the Run 2 beam pipe instead of the new beam pipe geometry used by step 1.9. The step 2.2 has two versions of geometry: one is the version simplified to the same extent as step 1.9, and the other is implemented in details by its original author although still more simplified than the Geant4 geometry.

3.5.2.2 Simplified version

The category definition is the same as step 1.9, except that all moderators in barrel and end-caps are tagged as one category. The X_0 distribution comparison as a function of η for step 2.2 simplified version is shown in Figure 3.24. It shows that the services remarkable deficit is corrected in this new version. Most categories are in good agreement except for beampipe and moderator.

The beampipe in GDML is found to be improperly described. Presented in Figure 3.25 is the proper Run 2 beampipe geometry extracted from the GeoModel and implemented in FLUKA, which consists of the Beryllium pipe wrapped by kapton, aerogel, aluminum foil, etc. However, it is found that only the Beryllium pipe is modeled in GDML. So the GDML file is re-implemented following this beampipe description, and the result is shown to be consistent with Geant4, as proved in Figure 3.26.

To probe the origin of the difference in moderator, the X_0 distributions in R - z dimensions are compared in Figure 3.27. The material components in geometry volumes are also traced. The feedback from both simulation and radiation groups confirms that the GDML (and the FLUKA) geometries are closer to the engineering prescription while improper modeling exists in the Geant4 geometry. In the barrel area of the Geant4, neither the outer moderator cylinder radii nor the inner moderator z -end length could match the ITk envelope. For the end-cap moderator, its inner end radius in Geant4 is too small to be able to clear the ion pump when retracting the endcap. The thickness of the moderator in GDML is also confirmed to be consistent with the engineering prescription while the Geant4 geometry gives improper estimation. If the

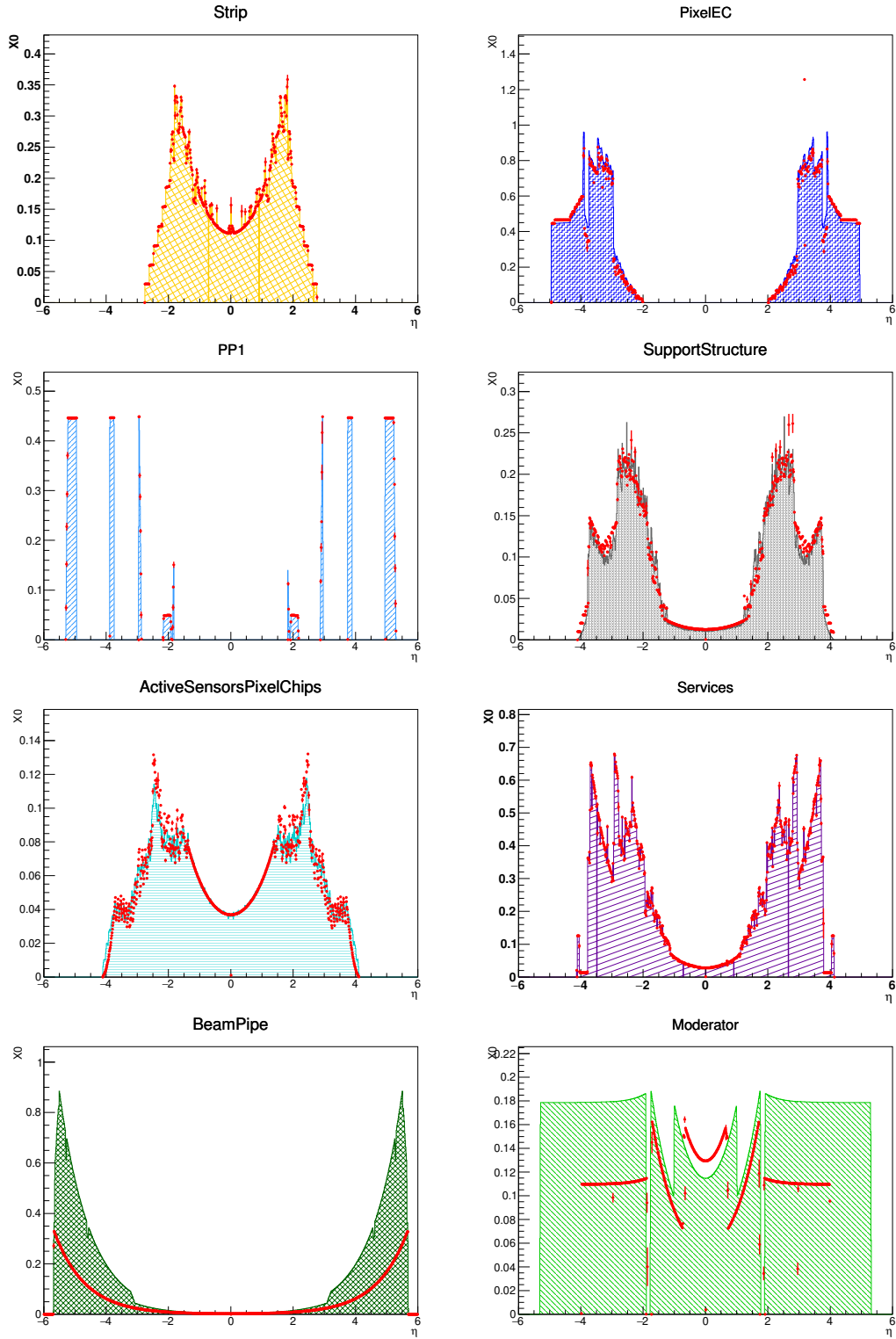


Figure 3.24: The category wise geometry comparison in step 2.2 between G4ML (red dots) and Geant4 (histograms).

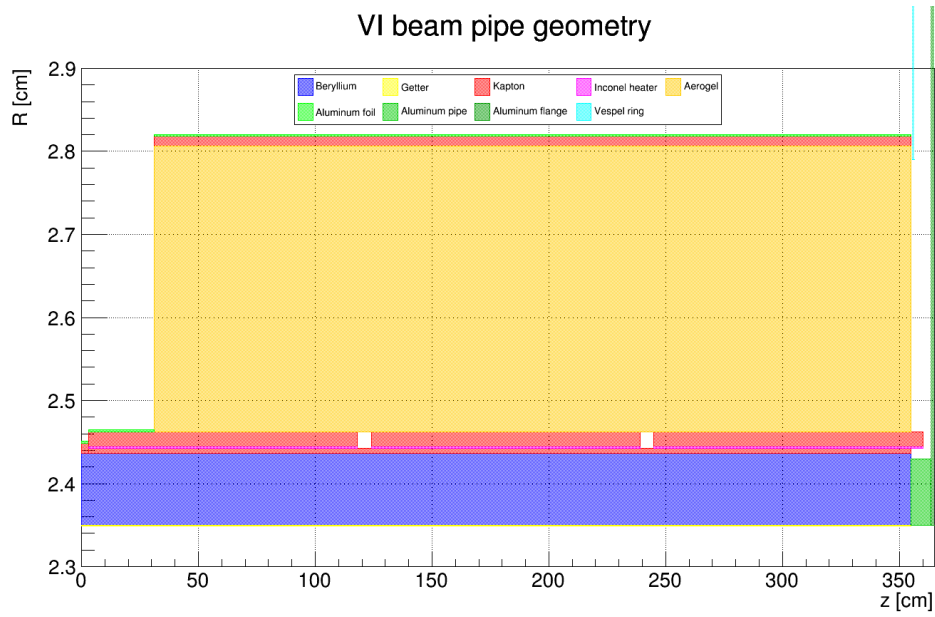


Figure 3.25: Run 2 beampipe geometry extracted from GeoModel and implemented in FLUKA.

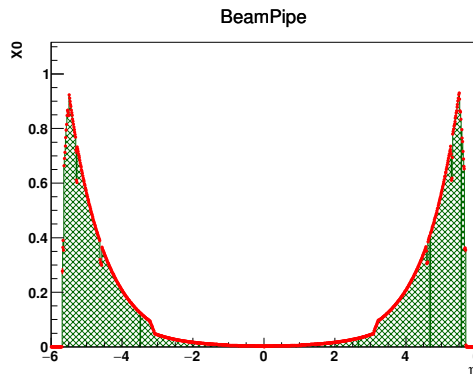


Figure 3.26: X_0 versus η comparison for beampipe in step 2.2 between GDML (red dots) and Geant4 (histograms), after correcting the GDML file.

GDML moderator volume shapes and positions are hacked to be Geant4-like while keeping the material components invariant, the distribution disparity is mitigated as shown in Figure 3.28. The remaining discrepancy can be expected to be fixed in the new geometry version by giving proper estimation of the material composition. This confirms that the difference is correctly traced, and re-modeling of Geant4 moderator is necessary.

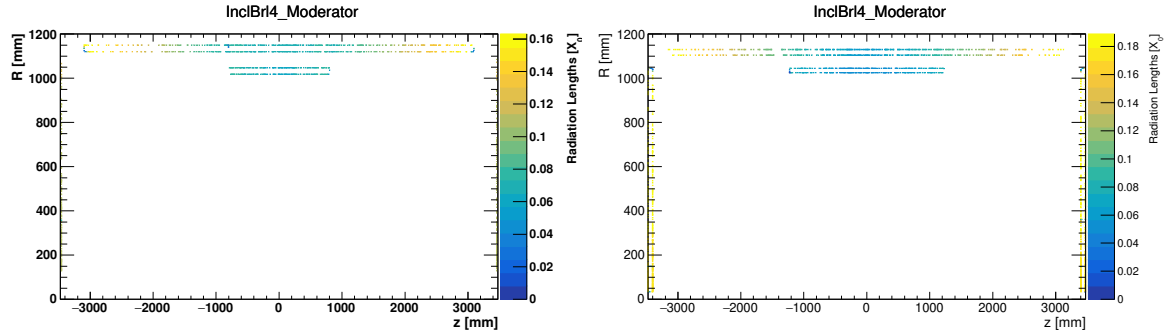


Figure 3.27: X_0 versus R - z directions for moderator category in step 2.2 for GDML (left) and Geant4 geometry (right)

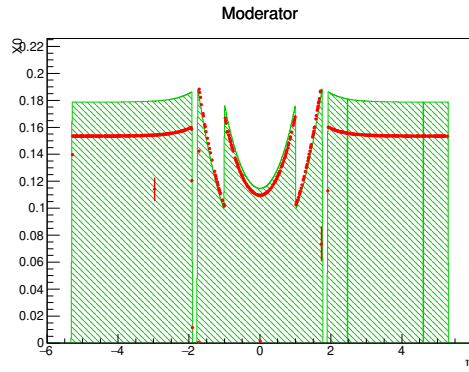


Figure 3.28: X_0 versus η comparison for moderator in step 2.2 between GDML (red dots) and Geant4 (histograms), after hacking the GDML file to make it Geant4-like.

3.5.2.3 Detailed version

For step 2.2, the simplified version is derived from the detailed one, and thus the two versions are expected to be mutually equivalent. Following the category definition for the simplified step 2.2, the material distribution consistency in each category between the detailed and the simplified geometries is demonstrated in Figure 3.29.

The detailed GDML description, with relatively higher model granularity, makes it possible to do comparison following the splitting strategy illustrated in Section 3.4. A good agreement level can be seen in Figure 3.30, where the total X_0 distributions of the entire ITk are compared between the GDML and the Geant4 geometries. Figure 3.31 presents the stacked X_0 distributions of all categories for both geometries, demonstrating the similar distributions between the two models. Figure 3.32 shows the location of the materials for one quadrant ITk of the two geometries, excluding moderators. The more solid structures in the Geant4 geometry come from the higher modeling granularity and the statistics.

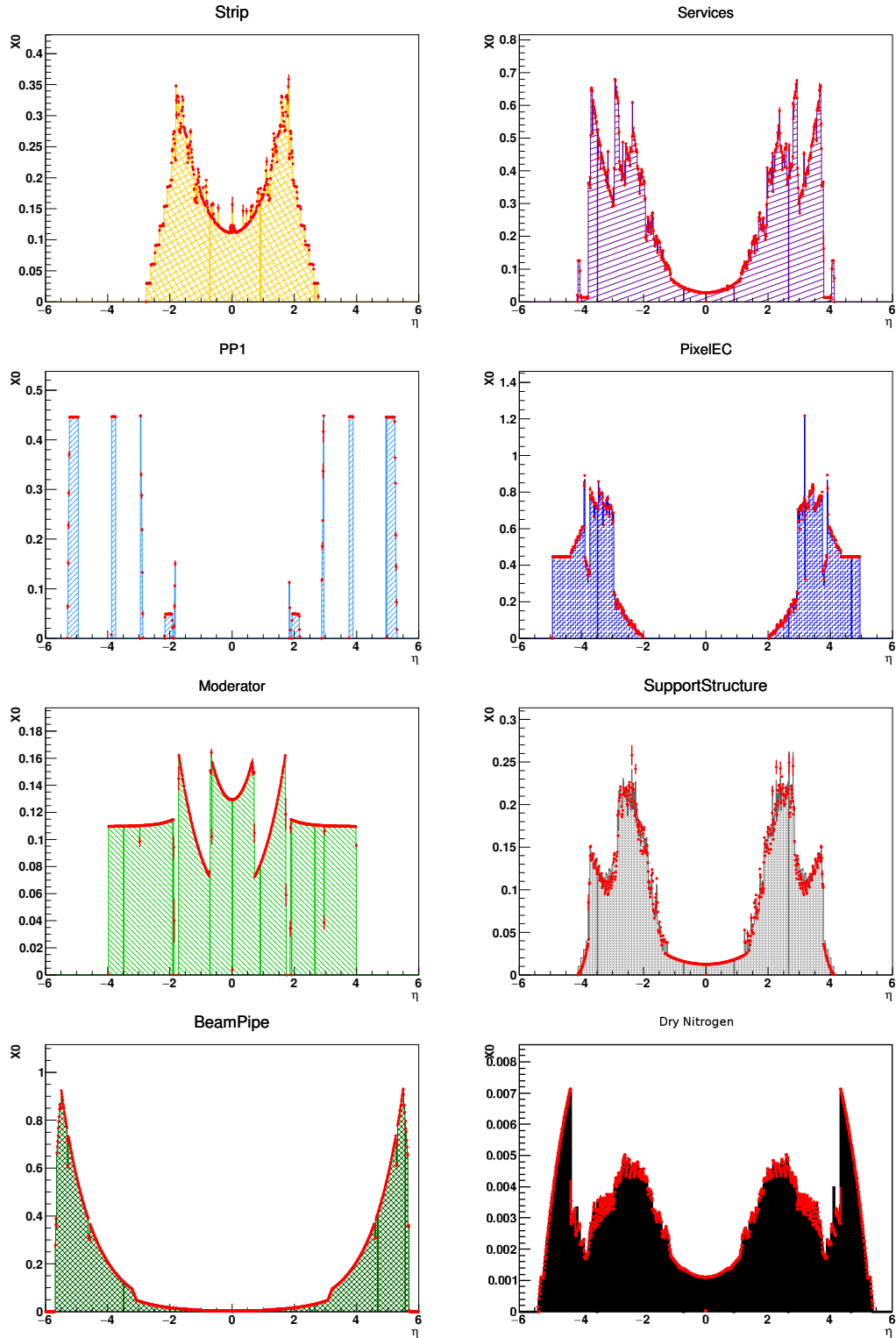


Figure 3.29: The category wise geometry comparison in step 2.2 between the detailed (red dots) and simplified (histograms) GDML.

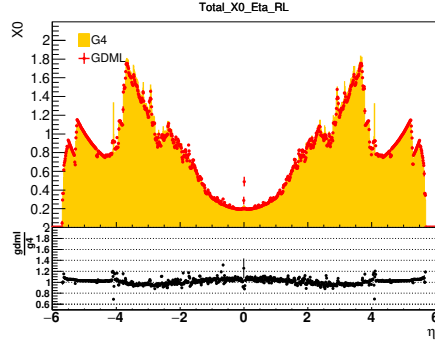


Figure 3.30: The total X_0 comparison versus η between GDML (red dots) and Geant4 (yellow histogram) in detailed step 2.2.

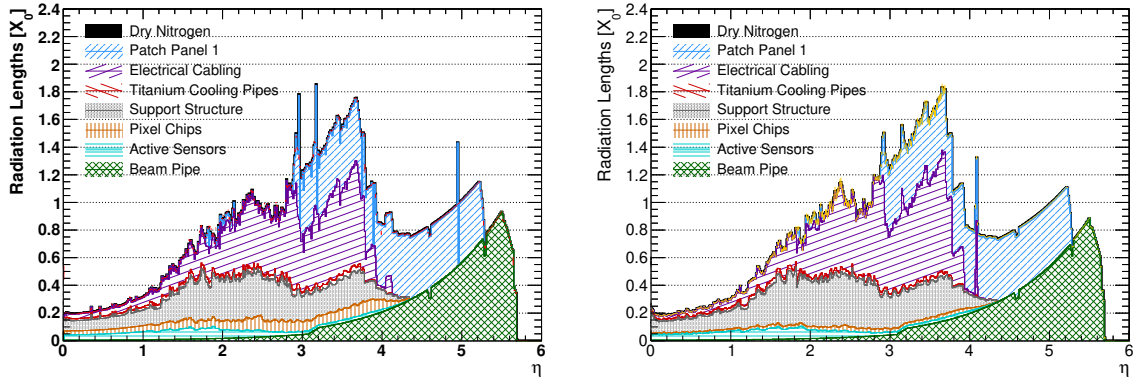


Figure 3.31: X_0 versus pseudorapidity η after geometry modifications in step 2.2 for GDML (left) and Geant4 (right).

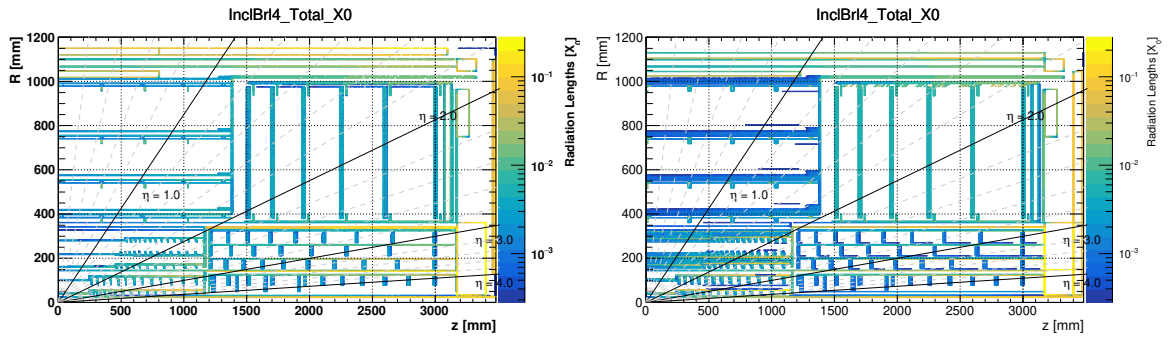


Figure 3.32: X_0 versus R - z directions of step 2.2 after geometry modifications for GDML (left) and Geant4 (right).

The categorical level comparison is shown in Figure 3.33 except moderators. The support structure, beam pipe and the active sensor sections achieve relative good agreement. Good consistency is almost everywhere except at $|\eta|$ around 5 in the nitrogen category which contribution to the overall material is anyway almost negligible. Remarkable disparities exist in the chip and titanium cooling sections at $|\eta|$ around 4, and in the service category at $|\eta|$ close to 0. Nevertheless, the materials in all of these regions contribute only modestly to the overall distributions. And such disparity impact is expected to be further mitigated once the safety factor of 1.5 is applied for the radiation study. Although the agreement at this level is enough to validate the radiation estimates, both geometries inconsistency in these three categories are scrutinized.

In the chips category, the X_0 distribution differences for the two geometries start to show up as $|\eta|$ close to 4. Figure 3.34 reveals that GDML has more material for the pixel end-cap chips than the Geant4 ones, throughout all the rings. The chip volume z -end length in Geant4 is smaller than the GDML one, ending up with $L_{\text{step}}^3/L_{X_0}^4 = 7.6 \times 10^{-5}$ for Geant4, which is two orders of magnitude smaller than the GDML one, 4.0×10^{-3} . Such ratios of the barrel chips, however, are consistent between the two geometries and has the same order of magnitude as the GDML pixel end-cap chips. The X_0 distribution along η for pixel end-cap chips are presented in Figure 3.35, and their shape discrepancy is compatible with the difference in the inclusive chip category, confirming that the pixel end-cap chips are the origin of the difference in this category.

Titanium cooling pipes are modeled in various volumes, and thus require calculation to be splitted out as a category. Some titanium pipes and the inside CO_2 coolant are modeled in one standalone volume as a mixture of carbon, oxygen and titanium elements. Most of the time, such cooling system is merged with the electrical cables into one volume filled with mixtures consisting of relevant elements with proper fractions. The inclined section in the pixel barrel layer 1 is mounted with quad modules which are difficult to cool. Therefore the material estimation is conservative where the cooling system is mixed with the mounting structure in one volume for both geometries. The top row in Figure 3.36 shows these cooling pipes merged with mounting structure in R - z space. The 2D distributions of the other pipes mixed with cables are shown in the bottom row, where more titanium pipe material estimation can be seen in GDML in the regions close to PST and IST positions (especially in the range $2000 < |z| < 3000$ mm, and $100 < R < 300$ mm). The X_0 distribution along η is shown in Figure 3.37, for the so-splitted pipes respectively. It confirms that the disparity in the global titanium distributions comes from the inconsistency in the regions close to PST and IST.

The differences in the service section are also further traced. More contributions in GDML are found in pixel barrel layers, compared with the Geant4 geometry. The layer wise X_0 distributions are presented in Figure 3.38, showing that the disparities exist in almost all layers except for the innermost one.

Although the discrepancies are at small scale and the radiation estimation for step 2.2 is validated, they will be taken care of by both the simulation and radiation study groups for the next version of geometry descriptions.

³The geometric distance that a particle traverses in a volume

⁴The distance converted into radiation length.

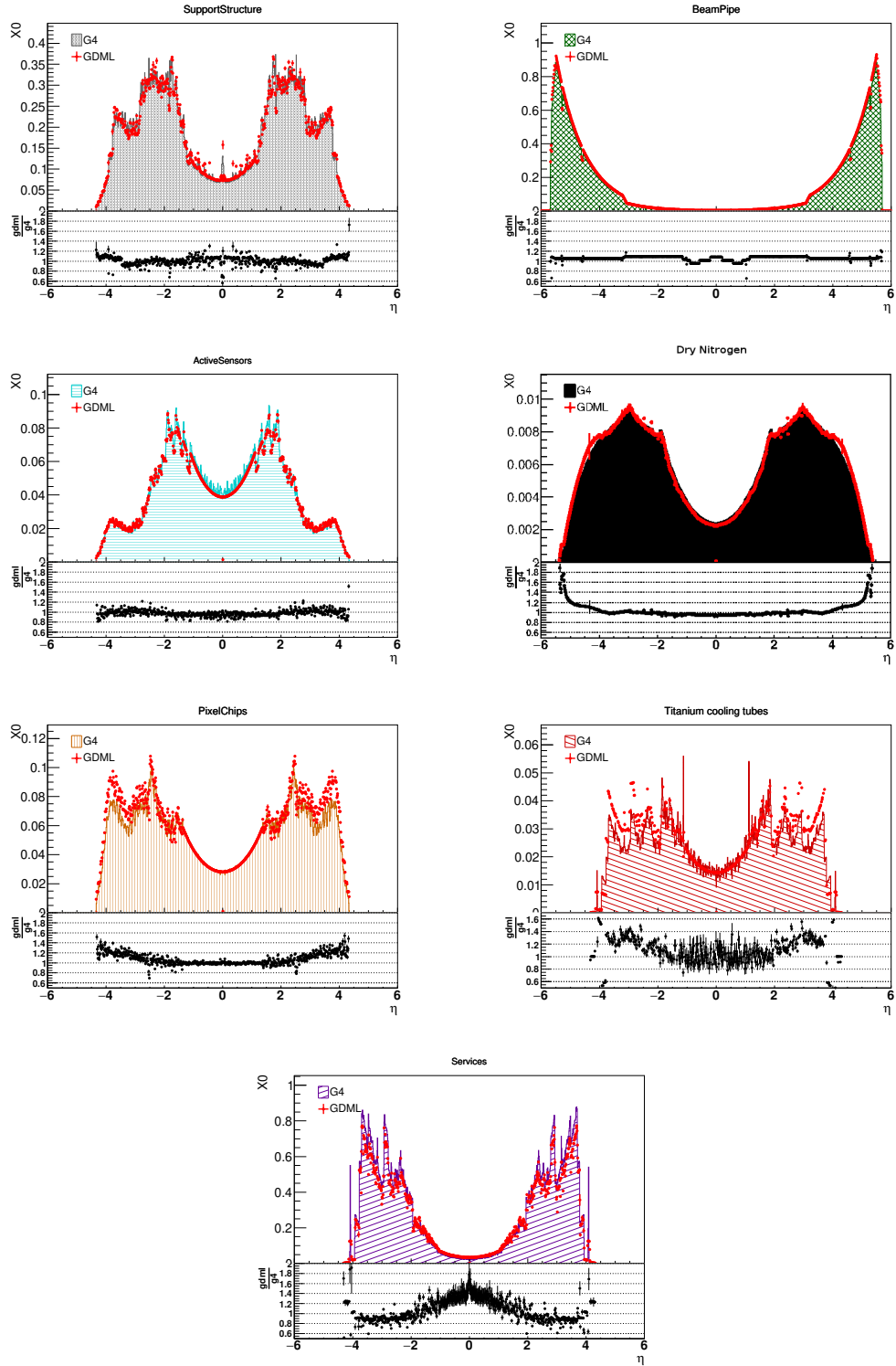


Figure 3.33: The category wise geometry comparison in step 2.2 between GDML (red dots) and Geant4 (histograms), after geometry modification.

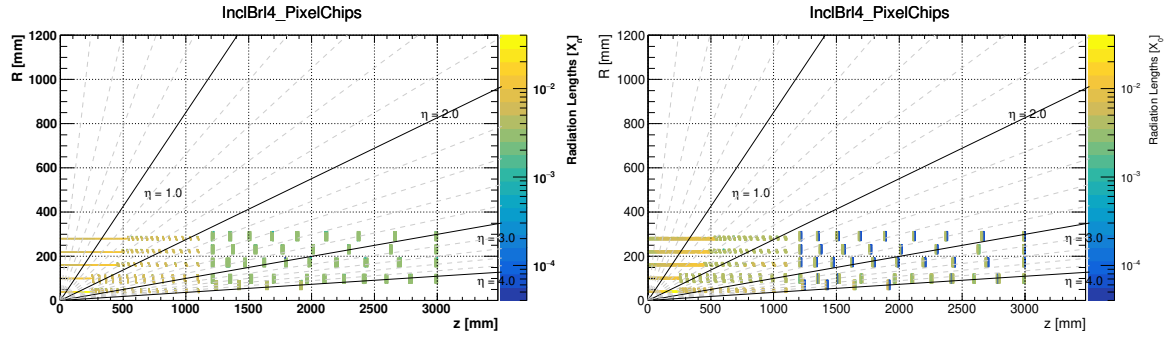


Figure 3.34: X_0 versus R - z directions for pixel detector chips of step 2.2 for GDML (left) and Geant4 (right).

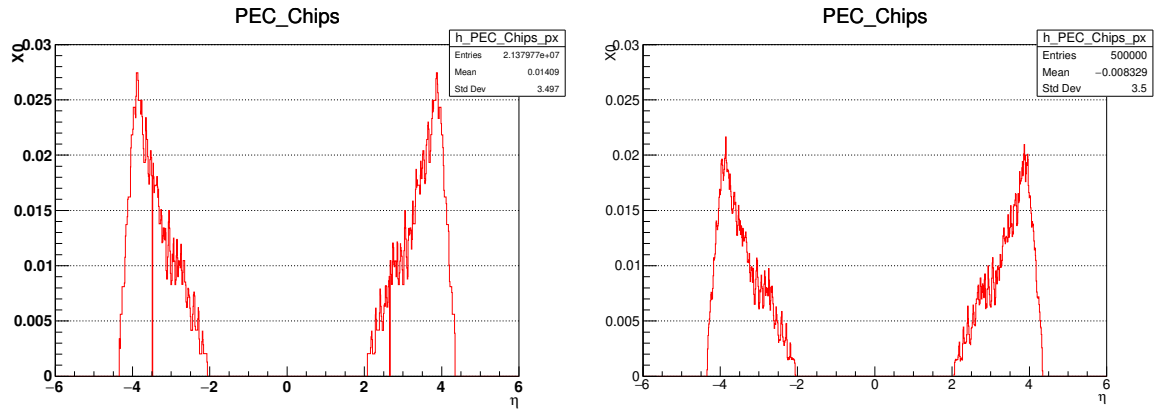


Figure 3.35: X_0 versus η direction for pixel end-cap chips in step 2.2 for GDML (left) and Geant4 (right).

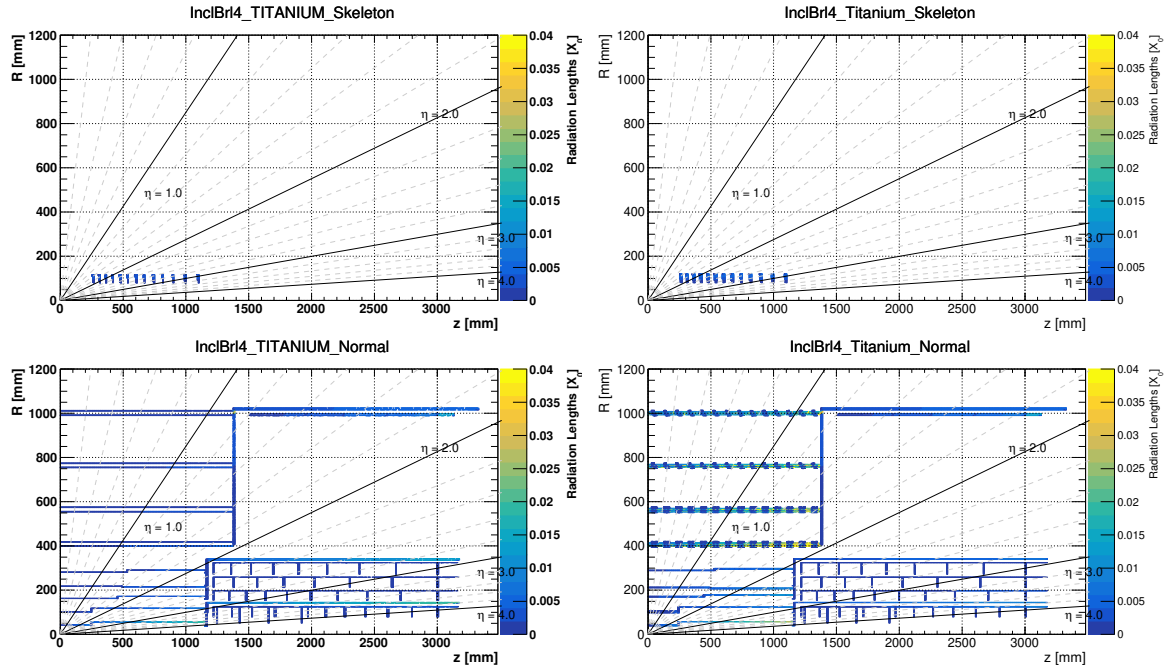


Figure 3.36: X_0 versus R - z directions for cooling pipes merged with mounting structure in layer 1 (top row), and cooling pipes in other volumes (bottom row) in step 2.2 for GDML (left) and Geant4 (right).

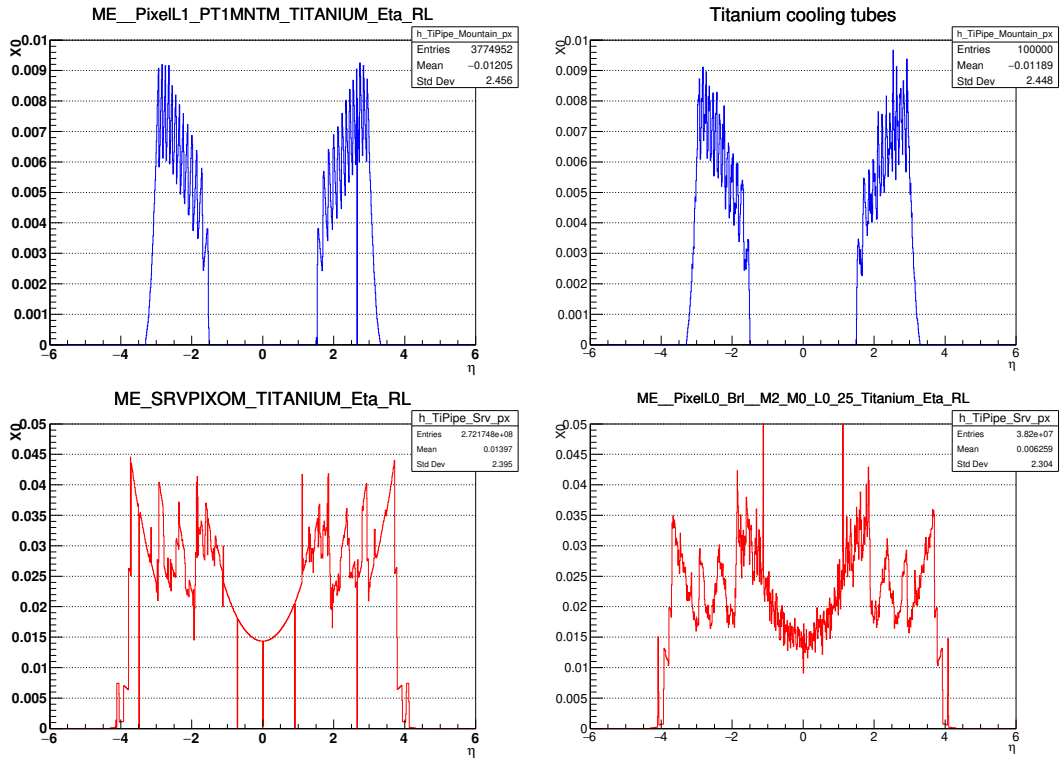


Figure 3.37: X_0 versus η direction for cooling pipes merged with mounting structure in layer 1 (top row), and cooling pipes in other volumes (bottom row) in step 2.2 for GDML (left) and Geant4 (right).

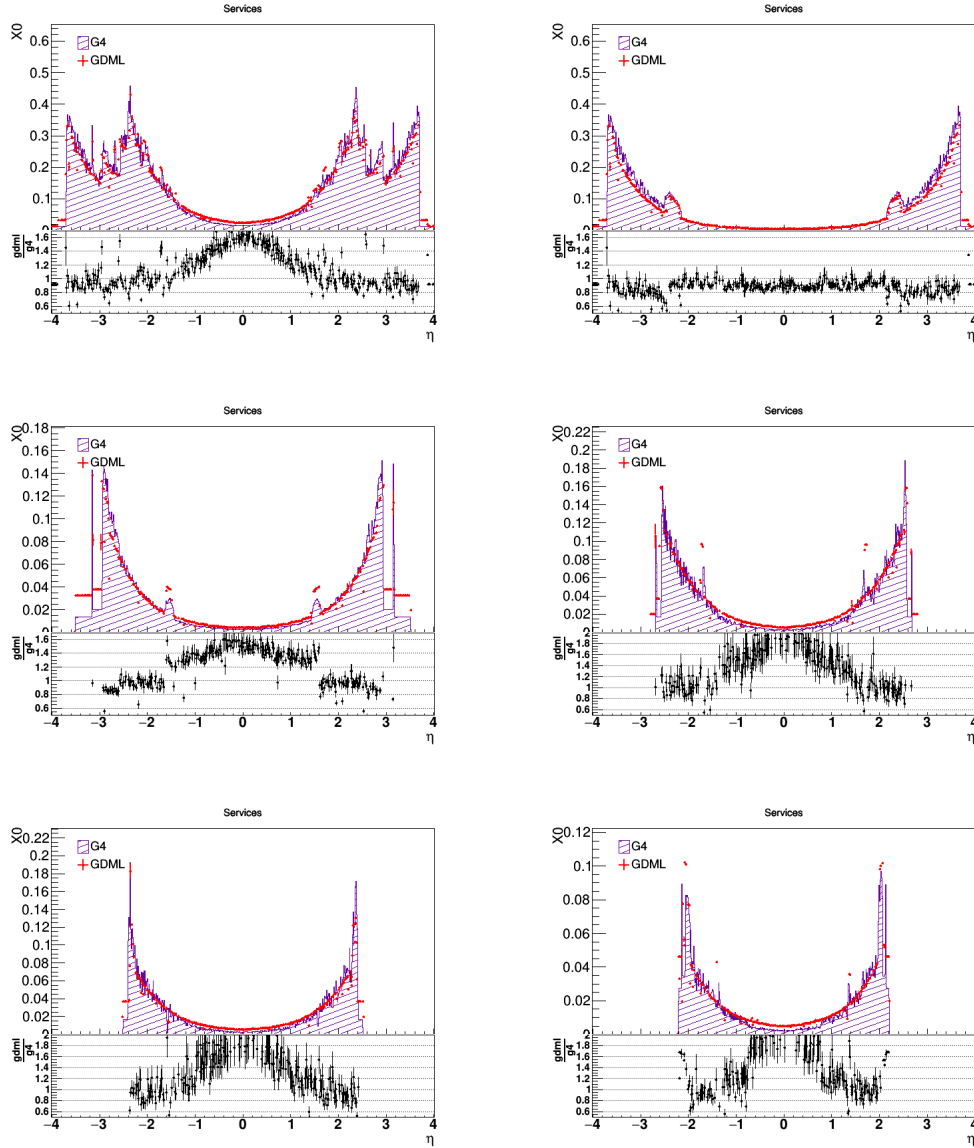


Figure 3.38: X_0 versus R- z directions for services in pixel barrel layers in step 2.2. The top left plot shows the inclusive layers. Plots for individual layer 1 (top right), layer 2 (middle left), layer 3 (middle right), layer 4 (bottom left), and layer 5 (bottom right) are also shown.

3.6 Summary

The ITk design evolves iteratively, aiming to cope with the operation of HL-LHC. Both detector simulation and radiation studies are carried out for the proposed layouts, based on independently implemented modeling geometries with different simplification schemes. Thus validation studies on the equivalence between the two geometries are needed. In my study geometries are compared by calculating the material distributions in terms of the radiation length X_0 . Although existing in Athena, such functionality is for the first time realized in FLUGG, the framework that allows to perform radiation studies with FLUKA. The material distributions are compared for various ITk subsystem categories, such as services and support structures, and the discovered disparities are further traced down to the elementary volumes of the modeling. Such a categorization and scrutinization require detailed studies on both the Geant4 geometry used by simulation and the GDML geometry by radiation studies which models the ITk at different levels of simplification.

Geometric differences are found in both step 1.9 and step 2.2 versions of GDML and Geant4 geometries. GDML geometries in both the simplified and the detailed versions of step 2.2, which are used for the pixel TDR, are found to be in good agreement with the Geant4 geometry, after the mismodeling correction for the beampipe in GDML and services in Geant4. Differences in parts of pixel end-caps, cooling pipes and services categories still exit, but at very small scale compared to the overall material distribution. Therefore the radiation estimates for step 2.2 are validated by this study, as a contribution to the ITk pixel TDR [84]. The method that is explored by this study for the geometry comparison is still being used by the radiation study group to validate the latest iteration of the pixel detector layout.

Search for the Higgs boson in the single lepton $t\bar{t}H$ ($H \rightarrow b\bar{b}$) channel 4

4.1 Introduction

4.1.1 Motivation for the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis

The observation and measurement of the Higgs boson coupling to top quarks, the top Yukawa coupling, is a key approach to test the prediction of the SM, after the Higgs boson discovery[5, 6]. Such measurements through processes such as gluon-gluon fusion production and the decay to photons depend on the contribution of a loop structure where there is a top quark, and possibly unknown particles as well. Thus the results from such measurement must be based on the assumption that no BSM particle couples to the Higgs boson. However, the Higgs production associated with a top quark pair ($t\bar{t}H$), as a tree level structure according to its Feynman diagram (see Figures 4.1(a) and 4.1(b)), makes it possible to directly measure this coupling. Although the $t\bar{t}H$ mode only represents around 1% of the total Higgs-boson production cross section as described in Section 1.2.1, the top quarks in the final state offer a distinctive signature and allow many Higgs-boson decay modes to be accessed: $H \rightarrow b\bar{b}$, $H \rightarrow WW$, $H \rightarrow \gamma\gamma$, and $H \rightarrow ZZ^* \rightarrow 4l$.

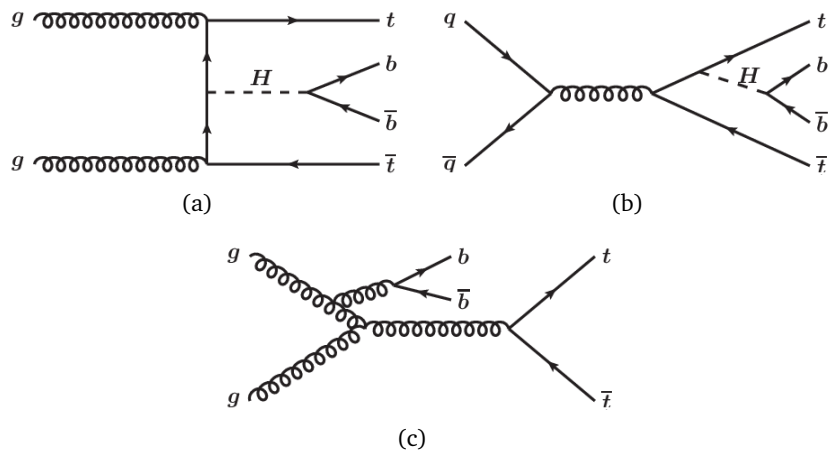


Figure 4.1: Representative tree-level Feynman diagrams for the (a) and (b) $t\bar{t}H$ ($H \rightarrow b\bar{b}$) process, and (c) the main background $t\bar{t} + b\bar{b}$ process.

The $H \rightarrow b\bar{b}$ final state is the dominant decay mode (58%) in the SM for the Higgs boson with a mass of 125 GeV. The $t\bar{t}H$ ($H \rightarrow b\bar{b}$) channel is therefore sensitive to the b -quark's Yukawa coupling, the second largest in the SM, in the decay. The $t\bar{t}H$ ($H \rightarrow b\bar{b}$) channel has a complicated final state including multiple jets with various b -flavored ones. This makes the analysis heavily depend on advanced analysis strategies. The studies in this chapter contribute to the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis using Run 2 data collected in 2015 and 2016, which was published in April 2018 [25].

4.1.2 Overview of the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis

Apart from the low rate of the $t\bar{t}H$ production, searching for the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) process is challenging in various aspects. This channel includes multiple b -jets among various jets in the final state. Thus the combined efficiency to reconstruct and identify all these jets would be low. Then combinatorial ambiguity exists when associating jets to partons which makes it difficult to reconstruct the Higgs boson. The dominant background comes from $t\bar{t}$ + jets, especially when the jets stem from b - or c -quarks, the so-called *heavy flavor* (HF) jets (see Figure 4.1(c)). The $t\bar{t}$ + jets has not only a much larger cross section, but also large theoretical uncertainties in Monte Carlo (MC) modeling.

Previous analysis rounds of $t\bar{t}H$ ($H \rightarrow b\bar{b}$) Both ATLAS and CMS searched for $t\bar{t}H$ ($H \rightarrow b\bar{b}$) using the full Run 1 data of around 20 fb^{-1} collected at 8 TeV. The signal amount is characterized by the signal strength parameter $\mu = \sigma/\sigma_{\text{SM}}$, which is the ratio of the observed cross section to the SM expectation. The observed μ in both experiments are shown in Figure 4.2 for independent studies in single lepton and dilepton channels and their combination. No evidence of a $t\bar{t}H$ signal over the background is observed in either experiment. In the ATLAS experiment $\mu = 1.5 \pm 1.1$ in the combined fit [90], and $\mu = 1.2 \pm_{-1.5}^{+1.6}$ in the CMS analysis [91]. In the ATLAS analysis, $t\bar{t}H$ is separated from $t\bar{t}$ using a Neural Network (NN) which uses variables based on the Matrix Element Method (MEM) exploited to deal with jet-parton association.

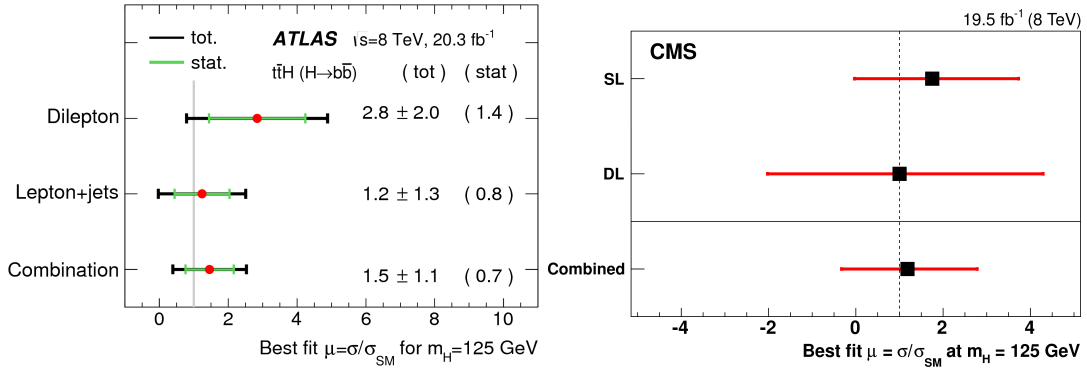


Figure 4.2: Summary of the signal strength measurement in Run 1 from (a) the ATLAS and (b) CMS experiments.

As a preliminary study with Run 2 data, the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis of the ATLAS experiment was carried out using the 3.2 fb^{-1} of data from the 2015 run, and 10 fb^{-1} of data from early 2016. Compared to the Run 1 strategy, the main change in the analysis is with the multi-variate analyses (MVA). A new technique is implemented aimed to solve the jet parton

assignment ambiguity, by using the Boosted Decision Trees (BDT) method. This step is called the reconstruction BDT. The discriminating variables are then built based on this step and are used by another BDT, a classification BDT, to extract the $t\bar{t}H$ signal. The result was presented at the ICHEP 2016 conference [92]. The signal sensitivity is similar to the Run 1 analysis with luminosity being two third of Run 1. The signal strength is found to be $\mu = 2.1^{+1.0}_{-0.9}$. Same as Run 1 results, the dominant systematic uncertainty is $t\bar{t}$ modeling.

Analysis channels Generally the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) study is split into three channels depending on the W -boson decay mode. The all-hadronic channel with both W -bosons decaying into quarks has the largest branching ratio of 46%, but is hard to be distinguished from the QCD multi-jets background. As a standalone analysis [93] using the data collected at 8 TeV, no evidence of the $t\bar{t}H$ production was found with an observed signal significance of 0.6 standard deviation (σ). The dilepton channel with both W -bosons decaying into leptons (electrons or muons) and the single-lepton channel with only one W -boson decaying leptonically and the other hadronically, follow the same analysis strategy but are studied separately and combined in the final fit in both Run 1 and Run 2 analyses. The results are driven by the single-lepton channel with a branching ratio of 44% while the dilepton study is limited by statistics (branching ratio of 10%).

Contributions The study in this chapter contributes to the single-lepton channel in the Run 2 analysis. The author explores the way to use jet tagging information and makes optimization studies in both reconstruction and classification BDTs. The latter in signal-enriched regions are fitted to get the final results. The dilepton analysis chain is not detailed in the following sections, it is implemented following a similar analysis chain.

Organization Using the full 2015 + 2016 datasets for an overall integrated luminosity of 36.1 fb^{-1} , an updated analysis [25] is conducted and presented in this chapter. Sections 4.2 and 4.3 describe the object and event selection and the event simulation respectively, followed by sections of the analysis chain with main updates. The event categorization, as shown in Section 4.4.1, is updated by employing not only the jet and b -jet multiplicity, but also the jet tagging information. Based on the Run 1 reconstruction and classification BDTs, the MVA techniques, shown in Section 4.4.2 and detailed in Section 4.5, to solve jet-parton association and signal extraction, are also updated to increase the discrimination between $t\bar{t}H$ and $t\bar{t}$ background. Section 4.6 presents the systematic models, the fit is presented in Section 4.7 and results are shown in Section 4.8.

4.2 Object and event selection

Events are selected from pp collisions at $\sqrt{s} = 13 \text{ TeV}$ recorded by the ATLAS detector in 2015 and 2016. Only events for which all relevant subsystems were operational are considered. This corresponds to an integrated luminosity of $3.2 \pm 0.1 \text{ fb}^{-1}$ recorded in 2015 and $32.9 \pm 0.7 \text{ fb}^{-1}$ in 2016. Events are required to have at least one vertex with two or more tracks with $p_T > 0.4 \text{ GeV}$.

4.2.1 Object selection

The physics objects used for event selection in this analysis are electrons, muons, jets and b -jets. The reconstruction of those objects is described in Chapter 2, and here is the set of criteria on these objects aimed to select $t\bar{t}H$ ($H \rightarrow b\bar{b}$) events.

Electrons and muons

Electrons (muons) are required to have $p_T > 10$ GeV and $|\eta| < 2.47(2.5)$. Electron candidates in the calorimeter barrel-endcap transition region ($1.37 < |\eta| < 1.52$) are excluded, electrons must satisfy the loose identification criterion. To reduce the contribution of leptons from hadronic decays (non-prompt leptons), the lepton isolation is required at the loose working point. To select leptons from the primary vertex, the longitudinal impact parameter is required to satisfy $|z_0 \sin \theta| < 0.5$ mm while the transverse impact parameter significance must have $|d_0|/\sigma_{d_0} < 5$ (3) for electrons (muons).

Jets

Jets are reconstructed from topological energy clusters in the calorimeter using the anti- k_t algorithm with a radius parameter of 0.4. The reconstructed jets are required to have $p_T > 25$ GeV and $|\eta| < 2.5$ after energy calibration. Quality criteria are imposed to identify jets arising from non-collision sources or detector noise, and any event containing such a jet is removed. To ensure the jets are from the primary vertex, additional requirement on the Jet Vertex Tagger (JVT) is also made for jets with $p_T < 60$ GeV in the $|\eta| < 2.4$ region of the detector, requiring $JVT > 0.59$. During the object reconstruction, a single detector response can be used to account for both jets and leptons, and an overlap removal procedure is thus also implemented to avoid double counting.

The pseudo-continuous (PC) b -tagging distribution is employed, as illustrated in Section 2.4.4.

Boosted objects

Boosted objects are also considered in the single lepton channel. Being boosted from the collision with high transverse momenta, the Higgs boson (top quark) tend to decay into collimated b -jets (b -jets and light-jets from W) forming into a larger jet cluster. Thus the so-called large- R jets are constructed by reclustering the selected jets through an anti- k_t algorithm with a radius parameter of $R = 1.0$. Boosted Higgs boson candidates are selected to have $p_T > 200$ GeV and contain at least two constituent jets among which at least two are b -tagged at 85% WP. The candidate with the highest sum of constituent jet b -tagging weight is selected. Additional large- R jets are considered as boosted top quark candidates requiring $p_T > 250$ GeV, exactly one constituent jet at 85% WP plus at least one additional constituent jet at 100% WP. The candidate with the highest mass is selected.

Missing transverse energy

The missing transverse energy is reconstructed as shown in Chapter 2. It is not used for the event selection but is exploited for the event reconstruction as described in Section 4.5.3.2.

4.2.2 Event selection

Using the single-lepton trigger, events with any of the leptons in Table 4.1 will be selected for further study.

lepton	p_T [GeV], 2015 (2016)	ID, 2015 (2016)	Isolation, 2015 (2016)
electron	$\geq 24(26)$	Medium (Tight)	Gradient
	≥ 60	Medium	None
	≥ 120	Loose	None
muon	$\geq 20(26)$	Loose	Loose (Gradient)
	≥ 50	None	None

Table 4.1: Single lepton trigger menu. Most criteria are the same for years 2015 and 2016 while the different ones used in 2016 are shown in parenthesis.

The previously mentioned lepton selections are relatively loose, to facilitate the background estimation with the data driven method and to maintain the orthogonality with other $t\bar{t}H$ channels. To improve the purity of $t\bar{t}H$ ($H \rightarrow b\bar{b}$), the events passing the previous criteria are further selected by imposing tighter criteria on the selected leptons. The electrons (muons) are required to pass the tight (medium) identification criterion. Both the electrons and muons are required to satisfy the Gradient isolation criteria which are more stringent as the lepton p_T drops.

For the single-lepton channel, events are selected to have exactly one lepton with $p_T > 27$ GeV and no other selected leptons with $p_T > 10$ GeV. This is orthogonal to the dilepton channel that requires to have exactly two leptons. In the single-lepton channel, a boosted category is firstly selected aiming to exploit the boosted signature of the $t\bar{t}H$ where events are required to have: at least one boosted Higgs-boson candidate, at least one boosted hadronic top-quark candidate and at least one additional b -jet tagged at 85% WP. Events that do not enter the boosted category are then considered for the so-called “resolved category” selected to have at least five jets, with at least two of them tagged as b -jets at 60% or three of them tagged at 77% WP.

4.3 Simulation samples

The $t\bar{t}H$ and all background process samples are detailed in this section. Almost all the processes are modeled by Monte Carlo (MC) simulations produced with the standard ATLAS simulation chain as illustrated in Section 2.3, except for the background from events with fake or non-prompt leptons (referred to as fakes in what follows). The MC samples described in this section are the so-called *nominal samples*, predicted with the full detector simulation to achieve the best precision of data modeling. Some other samples produced with fast simulation (see Section 2.3) are used to estimate modeling systematic uncertainties. To account for pile-up events, minimum bias interactions are generated using PYTHIA 8.186 and overlaid onto the hard process. All these samples are processed through the same reconstruction algorithms and analysis chain as the real data. The MVA trainings and the building of the fit template for signal extraction are only performed with MC events.

4.3.1 Signal modeling

MADGRAPH5_aMC@NLO [54] (referred to in the following as MG5_aMC@NLO) is used to calculate the matrix element (ME) for $t\bar{t}H$, at next-to-leading-order (NLO) accuracy in QCD. The NNPDF3.0NLO parton distribution function (PDF) set [94] was used. This generator is interfaced to the PYTHIA 8.210 parton shower (PS) and hadronization model using the A14 set of tuned parameters [95] for the underlying event (UE tune). The $t\bar{t}H$ cross-section is computed [22, 96–100] at NLO accuracy in QCD.

4.3.2 $t\bar{t}$ + jets background

The dominant background for the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis is the $t\bar{t}$ production with additional HF jets (especially b -jets) which are not from the $t\bar{t}$ decay products. The nominal simulation sample of the $t\bar{t}$ + jets background was generated using the POWHEG-BOX v2 NLO event generator [50–52, 101], referred to as POWHEG in the remainder of this document, with the NNPDF3.0NLO PDF set. The parton shower and the hadronization were modeled by PYTHIA 8.210 with the A14 set of tuned parameters. The sample is normalized using the predicted cross-section of 832^{+46}_{-51} pb, calculated with the Top++2.0 program [102] at next-to-next-to-leading order (NNLO) in perturbative QCD including resummation of next-to-next-to-leading logarithm (NNLL) soft gluon terms [103–106]. Alternative $t\bar{t}$ simulation samples are also produced to estimate systematic uncertainties as described in Section 4.6.1.

To be better understood and controlled by the fitting with data, the $t\bar{t}$ + jets samples are divided into subsets according to the flavors of additional jets. This jet flavor-labeling is performed with the truth jets, that is, the generator-level particle jets which are reconstructed from stable particles using the anti- k_t algorithm with a radius parameter of 0.4, and are required to have $p_T > 15$ GeV and $|\eta| < 2.5$. The flavor of such a jet is determined by counting the number of b - or c -hadrons within $\Delta R < 0.4$ of the truth jet axis (ΔR truth match), including single- b -jet with jet matched to exactly one b -hadron, and B -jets matched to at least two b -hadrons. The $t\bar{t}$ + $\geq 1b$ events are the ones required to have at least one of the additional jets ΔR truth matched to at least one b -hadron. At least one b -hadron in each jet is required to have p_T above 5 GeV. The events with only one single- b -jet are labeled as $t\bar{t}$ + b , and the others as $t\bar{t}$ + $\geq 2b$. Similarly, $t\bar{t}$ + $\geq 1c$ events are labeled from non- $t\bar{t}$ + $\geq 1b$ events if at least one of the additional jets is truth matched to at least one c -hadron with $p_T > 5$ GeV. Other events are labeled as $t\bar{t}$ + light.

4.3.3 Other backgrounds

Other backgrounds considered in the analysis represent almost 8% of the total amount of background in the single-lepton channel, after the event selection.

Processes that produce top quarks take around 4% of the total background. The electroweak single-top quark production, including W -associated (Wt), t - and s -channel production, are all generated with POWHEG-BOX v1 at NLO accuracy using the CT10 PDF set [107, 108]. The t -channel sample generation uses the four-flavor PDF set CT10 4F. All of these samples were interfaced to PYTHIA 6.428 [56] with the Perugia 2012 set of tuned parameters [109]. Their normalisation uses the approximate NNLO theoretical cross-sections [110–112]. Other top-quark production modes including four-top production ($t\bar{t}t\bar{t}$), $t\bar{t}WW$, tZW and tZ production are also taken into account. They are all generated with MG5_aMC@NLO and interfaced with PYTHIA 6 for tZ and with PYTHIA 8 for the rest.

Samples of W/Z production in association with jets ($W/Z + \text{jets}$), accounting for less than 3% of the total background, are generated, showered and hadronized with SHERPA 2.2.1. The matrix elements were calculated for up to two partons at NLO and four partons at leading order (LO) using the COMIX [113] and OPENLOOPS matrix element generators, and merged with the SHERPA parton shower [114] using the ME+PS@NLO prescription [55]. The $W/Z + \text{jets}$ events are normalized using the NNLO cross-sections [115]. And the $Z + \text{jets}$ heavy-flavor component normalization is corrected by a factor of 1.3 to match data, in a defined control regions with relatively pure $Z + \text{jets}$. The diboson + jets samples were generated using SHERPA 2.1.1 as described in Ref. [116].

$t\bar{t}$ production in association with a W or Z boson, $t\bar{t}W$ and $t\bar{t}Z$ ($t\bar{t}V$), representing around 0.4% of all backgrounds, are generated with the same setup as the $t\bar{t}H$ signal.

Higgs-boson production in association with a single top quark is rare in the SM, but is included in the analysis and treated as background though almost negligible. The tWH process is produced with MG5_aMC@NLO interfaced to HERWIG++ [58] with the CTEQ6L1 PDF set [117]. Samples of single top quarks plus Higgs boson plus jets ($tHqb$) were produced at LO with MG5_aMC@NLO interfaced to PYTHIA 8, using the CT10 4F scheme PDF set. Other Higgs-boson production modes are found to be negligible and are not considered.

In the single-lepton channel, around 1% of background is from the fake lepton which is actually a jet or a photon but misidentified as a lepton, or a non-prompt lepton. None of them is the lepton from W decay and will be referred to as fakes hereafter. Such processes with an important contribution from multi-jet production are difficult to simulate, and thus are estimated from data using a matrix method [118]. By removing the lepton isolation requirements and loosening the identification criteria on electrons, the current tight event space (N^t) is enlarged to have more fake leptons (N^f). Next, the efficiency to pass the tight criteria for both fake (ε_f) and real leptons (ε_r) are estimated from other well-understood data sets. Based on these efficiencies, the fake leptons in our current event space (the tight event space) can be estimated.

4.4 Analysis strategy

After the preselection as described in Section 4.2.2, the total event space consists of $t\bar{t}H$ and large contributions from various backgrounds, mainly $t\bar{t} + \text{jets}$. To have constraints on the background simulation, the overall event space is split into *regions*, including the background *control regions (CRs)* which are almost pure in backgrounds and the *signal regions (SRs)* having low but relatively larger signal purity than other regions. When fitting the MC predictions to data simultaneously across all regions, the systematic uncertainties of the background can be constrained by data in CRs. Such a background control would be translated to the SRs, improving the signal significance.

In the SRs, due to the low signal to background ratio, a multivariate analysis variable separating signal from the background is used as discriminant for the profile likelihood fit to further improve the signal sensitivity.

4.4.1 Categorization

The selected events are firstly divided according to jet multiplicity: one with exactly 5 jets (5j) and one with 6 or more jets ($\geq 6j$). The following describes event categorization and the

multivariate analysis techniques will be considered and performed separately for the 5j and $\geq 6j$ cases.

To gain more control over the main $t\bar{t}$ + jets background, events are categorized aiming to have control regions enriched in different jet flavors: $t\bar{t} + b\bar{b}$, $t\bar{t} + b$, $t\bar{t} + \geq 1c$ and $t\bar{t}$ + light. Signal regions are defined to have more $t\bar{t}H$ events. In this analysis, the pseudo-continuous b -tagging distributions, which are discriminating for different flavors of the $t\bar{t}$ + jets background, are exploited for the event categorization.

The categorization starts from splitting the full event space into sub-regions according to the PC b -tagging bin the jets fall in in each event, where only the four jets with the highest MV2c10 score are considered. As a first step, the *ultra pure signal region* (UPSR or SR1) was defined requiring that all the four jets are in the 5th bin of the PC b -tagging, meaning passing the WP at 60%. Then a sequential merging is performed to compose regions one by one, each region being enriched in one background component. This process is optimized and finally defined 12 categories for the single-lepton channel. Figure 4.3 shows these categories merged from sub-regions split according to jets' b -tagging.

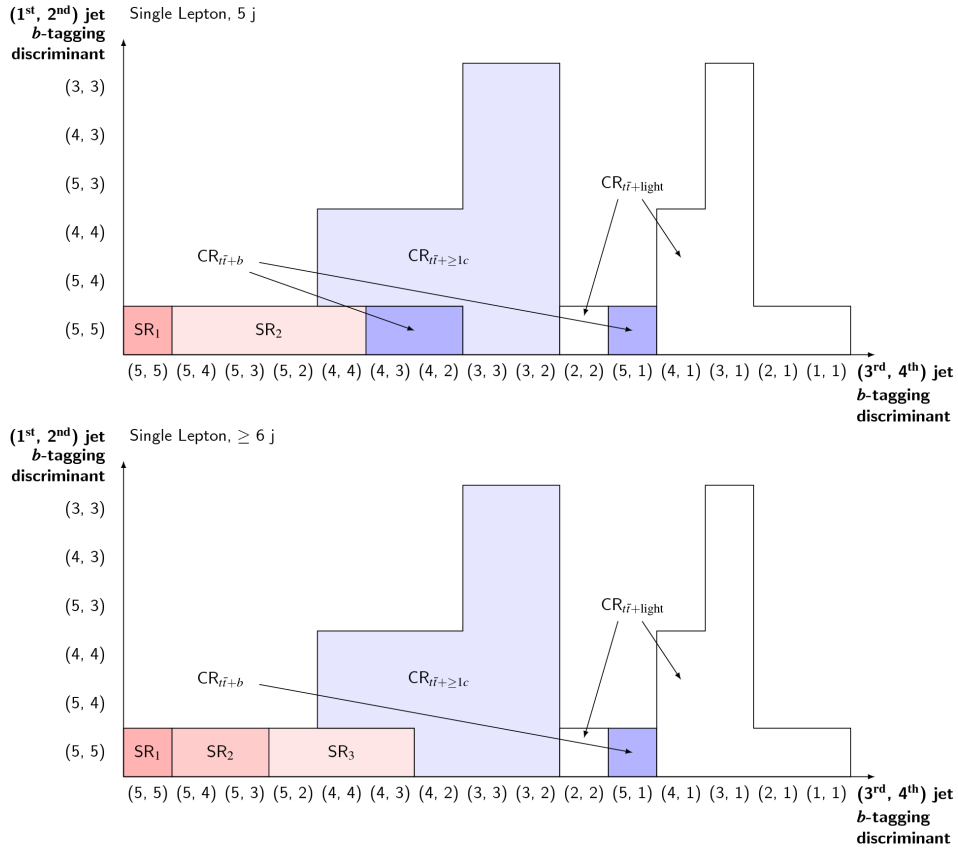


Figure 4.3: Definition of the (a) five-jet and (b) six-jet signal and control regions, as a function of the PC b -tagging discriminant. The vertical axis shows the values of the b -tagging discriminant for the first two jets, while the horizontal axis shows these values for the third and fourth jets. The jets are ordered according to their value of the b -tagging discriminant in descending order [25].

The $t\bar{t}H$ purity for each of the 12 regions are shown in Figure 4.4. The categories with $S/B > 1.5\%$ are considered as SRs, and the UPSRs are the purest signal regions.

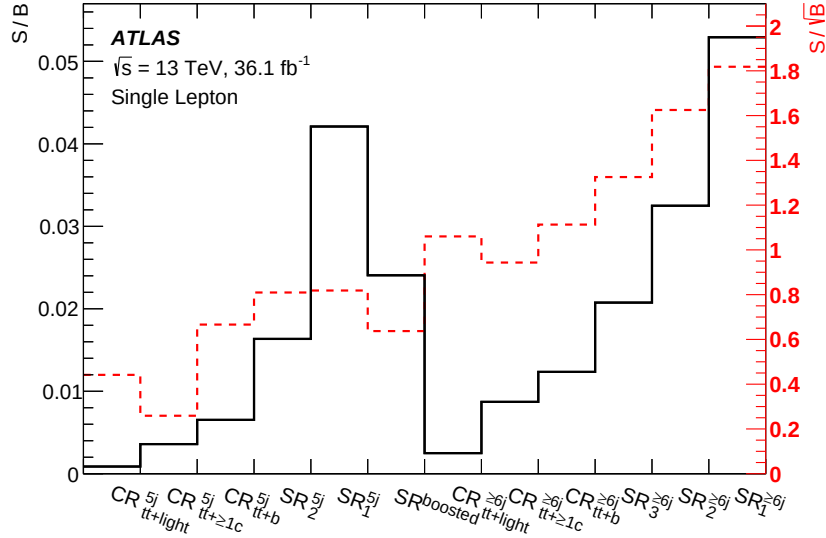


Figure 4.4: The ratios S/B (black solid line, referring to the vertical axis on the left) and $S/\sqrt{B}$ (red dashed line, referring to the vertical axis on the right) for each of the analysis categories in the single-lepton channel, where S (B) is the number of selected signal (background) events predicted by the simulation or through the data-driven estimates [25].

The control regions are enriched in $t\bar{t} + \text{light}$, $t\bar{t} + \geq 1c$ and $t\bar{t} + b$, referred to as $\text{CR}_{t\bar{t}+\text{light}}^{\geq 6j}$, $\text{CR}_{t\bar{t}+\geq 1c}^{\geq 6j}$, $\text{CR}_{t\bar{t}+b}^{\geq 6j}$ for $\geq 6j$ and $\text{CR}_{t\bar{t}+\text{light}}^{5j}$, $\text{CR}_{t\bar{t}+\geq 1c}^{5j}$ and $\text{CR}_{t\bar{t}+b}^{5j}$ for $5j$, respectively. However, it is not easy to find CRs pure in individual flavors, as presented in Figure 4.5 showing the fraction of the different backgrounds in all regions. The $t\bar{t} + \text{light}$ CR is dominated by $t\bar{t} + \text{light}$, where a noticeable contribution is from $t\bar{t} + \geq 1c$. The $t\bar{t} + \geq 1c$ CR has larger fraction of $t\bar{t} + \geq 1c$ compared to other regions, but is still dominated by $t\bar{t} + \text{light}$. Thus $t\bar{t} + \geq 1c$ is not easy to control. Enhanced in $\text{CR}_{t\bar{t}+b}^{5j}$ and $\text{CR}_{t\bar{t}+b}^{\geq 6j}$, the $t\bar{t} + \geq 1b$ can be controlled, to some extent, in these regions as well as in the SRs which are overwhelmed by it.

4.4.2 Multivariate analysis techniques

The Higgs boson and top quark are the signature of the $t\bar{t}H$ signal, and thus need to be reconstructed in order to extract the signal. However, all of them decay into b -quark and thus b -jets in the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) channel. Therefore, the final state of the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) is composed of many jets and b -jets, which is confusing without knowing whether the jet is from the Higgs, top quark, or additional radiation. There are many possible assignments between jets and quarks. Therefore advanced analysis methods are important to deal with this ambiguity.

Three intermediate multivariate techniques are implemented: (a) the ‘reconstruction BDT’ used to select possibly the most correct jet-parton assignment in each event and to build the Higgs-boson and top-quark candidates, (b) a likelihood discriminant (LHD) method that combines the signal and background probabilities of all possible combinations in each event, (c) a matrix element method (MEM) that exploits the full matrix element calculation to separate the signal from the background [25]. These methods explore similar input information, but

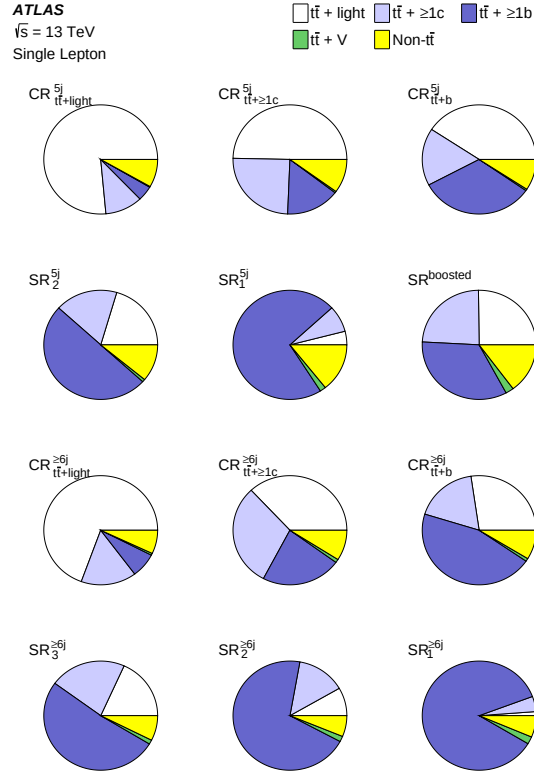


Figure 4.5: Fractional contributions of the various backgrounds to the total background prediction in each analysis category in the single-lepton channel. The predictions for the various background contributions are obtained through the simulation and the data-driven estimates [25].

from different aspects.

The results of all above intermediate MVAs together with global kinematics are then employed by another BDT, a so-called classification BDT, to extract the $t\bar{t}H$ signal from the $t\bar{t} + \text{jets}$ backgrounds. The b -tagging weights, for the first time, are also explored as input variables in all of the above mentioned algorithms.

The main contribution of the study during this thesis is to further study and develop the reconstruction and classification BDT for this analysis round using the TMVA toolkit [119].

4.5 Using boosted decision trees

This section details the study of using BDTs in the analysis, starting from an introduction to BDT, and then describing the BDT for signal reconstruction and extraction.

4.5.1 Boosted decision trees

Decision trees, a machine learning technique, firstly developed in the context of data mining and pattern recognition [120], was then adopted in various fields including high energy physics [121]. Starting from the MiniBooNe experiment where the boosted decision trees method was firstly explored for particle identification [122], the D0 experiment was the first to use the BDT for a search, which led to the first evidence (and then observation) of single top quark production [123].

A decision tree

The decision tree method can be used for classification and regression. Exemplified with the binary classification task as in Figure 4.6, the general idea is to classify events by learning a set of decision rules. The learning (referred to as *training* in the following) procedure starts from the root node containing all training events which will be split into two branches by cutting at a chosen value of a feature. Such a feature is chosen from a set of variables which are discriminating for the two classes according to expertise and are calculated manually beforehand. All the features will be considered and the one with a corresponding threshold will be chosen if they give the best separation between signal and background. The events passing the cutting point will enter one daughter node expected to be enriched in signal and the failing ones will go into a second background-enriched node. Analogously, these child nodes will be split again selecting from the full feature set. Compared with the simple cut-based method, in this way events that failed a particular criterion will not be dropped immediately but rather examined again with other features or even the same features in a subset of events. The terminal nodes where the splitting stops are called leaf nodes. A score is then assigned to each event denoting if it is more signal or background like, depending on which leaf it falls in, and it is usually the signal purity of the corresponding leaf.

Among the relatively limited number of parameters of the decision tree, the stopping condition parameters are important. They decide when the tree is not split any more, and include: (a) a minimum leaf size to ensure enough statistical precision in each node, (b) insufficient separation improvement, (c) a maximal tree depth. In addition, these stopping conditions may also help to limit the tree's over-growth which makes the tree too specialized on the training sample and lose the generalization ability when doing predictions on testing samples. This is called the *over-fitting* in the following.

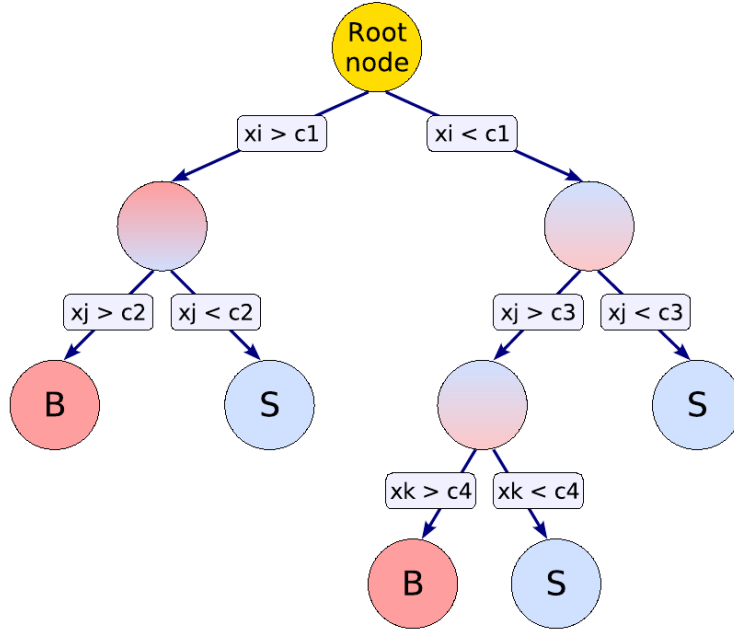


Figure 4.6: Schematic view of a binary decision tree which learns the splitting rule using variables x_i , x_j and x_k based on the training events entering the tree from the root node. The tree stops splitting at the leaf nodes which are labeled “S” for signal, and “B” for background, depending on the majority of the events in the nodes.

Boosting

Usually a single decision tree suffers from the discrete output and high sensitivity to the training sample (over-fitting). A way out is to use the boosting method. Boosted decision trees are an ensemble of decision trees, each of which is produced during the boosting iteration. The goal is to get a compound classifier combining various weaker classifiers which are trained with different samples of varied properties.

Adaboost One boosting method used in this study is adaptive boosting (Adaboost), and it is also one of the most commonly used methods. At each boosting iteration, Adaboost changes the sample distribution by modifying the weight assigned to each event. The mis-classified events will be given a larger weight during the training of the next tree which will then focus on these tough events. The weight update depends on the current tree k 's mis-classification rate ϵ_k , defined as:

$$\epsilon_k = \frac{\sum_j w_j^k}{\sum_{i=1}^{N_k} w_i^k} \quad (4.1)$$

where j runs over mis-classified training events, and N_k is the total number of training events. After training of the current tree, the weights of the mis-classified events are modified from $w_j^k \rightarrow w_j^{k+1} = w_j^k \times e^{\alpha_k}$ while the weights of the properly classified events remain unchanged. α_k is the weight associated to each tree T_k depending on its mis-classification rate ϵ_k , and could

be strengthened by a tunable factor β :

$$\alpha_k = \beta \times \ln \frac{1 - \epsilon_k}{\epsilon_k}. \quad (4.2)$$

The final score for event i , $T(i)$, is a weighted average of all trees where the first tree dominates the final score and the subsequent trees from boosting give corrections to it:

$$T(i) = \frac{1}{\sum_{k=1}^{N_{\text{tree}}} \alpha_k} \sum_{k=1}^{N_{\text{tree}}} \alpha_k T_k(i) \quad (4.3)$$

Gradboost The boosting procedure can be fully determined by a procedure [119] of minimizing a loss function which is the metric showing the deviation of the model prediction from the truth label. Training is actually a process aimed to minimize the value of loss function calculated based on training events. Equivalently, Adaboost is determined by an exponential loss function [119] $L(p, y) = e^{-p(x)y}$, where p denotes the model prediction and y is the truth label, and x represents the input features. However, the exponential loss of Adaboost has the shortcoming of being degraded in noisy setting lacking of robustness to outliers [119].

The Gradient boosting method, referred to as Gradboost, is tested in this study. It allows for potentially more robust forms of loss function. The TMVA Gradboost method uses the binominal log-likelihood loss $L = \ln(1 + e^{-2p(x)y})$ for classification. The boosting corresponding to this loss function can not be obtained straightforwardly as the Adaboost procedure. Instead, the implementation is the iterative calculation of the loss function gradient and growing of a regression tree. A set of trees are thus produced which minimize the loss function. Like the β parameter in the Adaboost method, the shrinkage in Gradboost is the learning rate to minimize the loss function as it controls the weights of individual trees.

The BDT has now become very popular in high energy physics. It is indeed a powerful machine learning method exploiting multiple variables compared with a cut-based analysis, and has a physicist-readable procedure and results. On the practical aspect, the model could be trained fast, fed in with both discrete and continuous variables and shows resistance to over-fitting.

4.5.2 BDT performance measure

The performance measure used is the Receiver Operating Characteristic (ROC), which tells how much a model is capable of distinguishing between two classes. For each decision threshold on the output of the classifier, the signal efficiency (how often signal events are correctly selected) is obtained against the background rejection (how often the background is not mis-classified, i.e. one minus background efficiency). By varying the threshold, the ROC is formed with a set of efficiency pairs. For the same signal efficiency, the higher the curve is, the better performing is the classifier, with higher background rejection. The integrated Area Under the Curve (AUC) quantifies the performance, although it is not the full story when the two curves cross each other.

For the reconstruction BDT, the efficiency of correctly reconstructing certain objects is used when comparing different models. It is the ratio of the number of events having the specific objects correctly reconstructed over the total number of events in the considered subset.

4.5.3 BDT-based event reconstruction

The full reconstruction of the $t\bar{t}H(H \rightarrow b\bar{b})$ semileptonic system is implemented by the reconstruction BDT detailed in this section. Section 4.5.3.1 defines the truth match which is used to define signal and background for the reconstruction BDT. Then to build the discriminating features for MVAs, the Higgs boson, top quark and W boson are reconstructed following the algorithm presented in section 4.5.3.2. The implementation of the BDT technique and its performance are described in section 4.5.3.3 and 4.5.3.4.

4.5.3.1 Truth matching

At truth level, jets are deemed to be correctly associated to quarks from the hard scatter if the spatial distance between them satisfies $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} < 0.3$, referred to as truth matching. According to the final state of the $t\bar{t}H(H \rightarrow b\bar{b})$ process, the matching of six jets are considered where four of them are b -jets. Each jet can only be associated to one parton. The lepton is not considered since it is correctly selected in most cases.

The fraction of selected events satisfying the different matching requirements are calculated separately for two event topologies: ≥ 6 jets where ≥ 4 of them are b -tagged at the 85% WP, referred to as 6ji4bi@85%, and $= 5$ jets where ≥ 4 of them are b -tagged, referred to as 5je4bi@85%. They are shown in Figure 4.7. It corresponds to the maximum achievable correct matching efficiency for the reconstruction BDT. The bins in the figure are defined as:

- all: all matched, the six selected jets in the event are matched to the six partons in the final state of the $t\bar{t}H(H \rightarrow b\bar{b})$ process. For 5j events, this bin is not calculated.
- b+1W: four selected jets are matched to the four b -quarks from the top and Higgs decays and one selected jet matches to one quark from the W decay.
- allb: four selected jets are matched to the four b -quarks from the top and Higgs decays.
- H: two selected jets are matched to the two b -quarks from Higgs decay.
- btop: two selected jets are matched to the two b -quarks from $t\bar{t}$ decays.
- W: two selected jets are matched to the two quarks from hadronic W decay.
- Hb1: one jet is matched to the leading p_T b -quark from Higgs decay.
- Hb2: one jet is matched to the sub-leading p_T b -quark from Higgs decay.
- blt: one jet is matched to the b -quark from leptonic top decay.
- bht: one jet is matched to the b -quark from hadronic top decay.
- wj1: one jet is matched to the leading p_T quark from hadronic W decay.
- wj2: one jet is matched to the sub-leading p_T quark from hadronic W decay.

For events with ≥ 6 jets and ≥ 4 jets tagged at 85%, around 36% of the events have six selected jets all truth matched to the six partons for the $t\bar{t}H(H \rightarrow b\bar{b})$ process. This low fraction can be explained by the fact that some products from the $t\bar{t}H(H \rightarrow b\bar{b})$ decay are absent in the selected events, due to the acceptance of the detector and selection criteria on the objects.

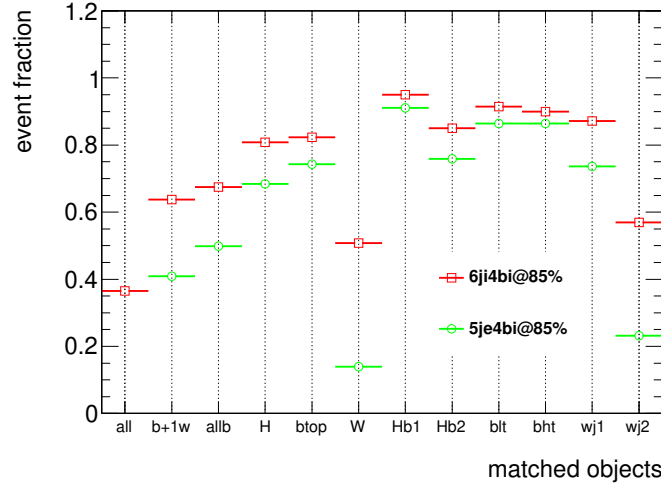


Figure 4.7: Fraction of selected events where the corresponding objects are truth matched. These are the maximum achievable efficiency for the reconstruction BDT.

4.5.3.2 Reconstruction algorithm

To solve the ambiguity of jet origin as shown in Figure 4.8, all jet-parton permutations are considered in the following reconstruction method, which is also using the lepton and E_T^{miss} in order to match the $t\bar{t}H(H \rightarrow b\bar{b})$ semileptonic final state. Each possible combination will end up with a corresponding reconstruction of objects, e.g. the hadronic top, the Higgs boson, etc.

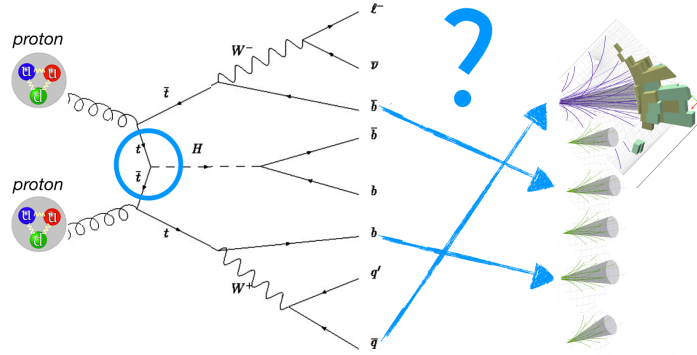


Figure 4.8: A schematic plot showing the ambiguity to associate multiple jets to the partons in the final state of $t\bar{t}H$ production.

Reconstruction of the leptonic W boson

To reconstruct the leptonic W boson, the full neutrino momentum is needed. Its transverse momentum is identified with the event E_T^{miss} , but the longitudinal component of the neutrino momentum ($p_{z\nu}$) is not measurable. It can be inferred by assuming the lepton and E_T^{miss} are originating solely from a W boson decay, and constraining the invariant mass to that of the W boson. It leads to a quadratic equation with two possible solutions:

$$p_{z\nu}^{\pm} = \frac{1}{2} \frac{p_{z\ell}\beta \pm \sqrt{\Delta}}{E_{\ell}^2 - p_{z\ell}^2} \quad (4.4)$$

where:

$$\beta = m_W^2 - m_{\ell}^2 + 2p_{x\ell}p_{x\nu} + 2p_{y\ell}p_{y\nu} \quad (4.5)$$

$$\Delta = E_{\ell}^2(\beta^2 + (2p_{z\ell}p_{T\nu})^2 - (2E_{\ell}p_{T\nu})^2). \quad (4.6)$$

In case where the quadratic equation does not have a real solution (around 20% for the $t\bar{t}H$ sample), the discriminant of the quadratic equation is set to zero ($\Delta = 0$) [124]. If there are two solutions, two different leptonic W bosons are considered.

Reconstruction of the hadronic W boson

The hadronic W is reconstructed using all combinations of 2 jets that are not considered as b -tagged jets in the selection. If the event contains less than two non- b -tagged jets a b -tagged jet is allowed to be used for reconstructing the hadronic W boson. This happens in a very small fraction of events.

The hadronic W boson is not reconstructed in the region with 5 selected jets, as the missing jet is usually originating from the sub-leading quark of the W decay. For simplicity variables may still refer to the hadronic W (as in the 6-jet scenario), which in this case is only constructed from a single non- b -tagged jet.

Reconstruction of the top quarks and the Higgs boson

The top quarks are reconstructed by associating a W boson and a b -tagged jet. The remaining b -tagged jets are used to reconstruct the Higgs boson. Since we are using an inclusive training starting from the ($\geq 6j, \geq 4b$) and ($5j, \geq 4b$) regions at 85% b -tagging efficiency, there are always at least four b -tagged jets.

In the region with 5 selected jets the hadronic top quark is reconstructed using a b -tagged jet and a light jet. This top is referred to as the incomplete hadronic top.

4.5.3.3 BDT technique to resolve combinatorics

The *reconstruction BDT* is trained only on the $t\bar{t}H$ sample.

Signal and background For the BDT training, “signal” refers to all the correct combinations in the $t\bar{t}H$ sample, as well as those where only one of the jets associated with the hadronic W is properly matched to a light quark from the W decay. All other combinations are treated as background.

The reconstruction BDT is trained to separate the signal combinations against the background ones. Then for each $t\bar{t}H$ ($H \rightarrow b\bar{b}$) event, the combination with the highest BDT score is selected as the correct combination, and will be used in the following MVA method.

Cross train and test In the following studies, cross training is used. Training and testing samples are defined based on the event number: the BDT trained on even events is then applied on odd events, and vice versa. In this way, the full MC statistics are kept for the analysis, being applied on models not trained on them.

Two types of BDT For each of 5j and $\geq 6j$ case, two types of BDT are trained : (a) the BDT relying only on kinematic variables related to decay products of the $t\bar{t}$ system, referred to as recoBDT or recoBDT_basic in the following, (b) one with additional variables associated with the Higgs boson decay products, referred to as recoBDT_withH. This is considered because the reconstruction BDT is only trained on $t\bar{t}H$ ($H \rightarrow b\bar{b}$) events. When applying it on $t\bar{t}$ background, the jets combination chosen by the BDT with the Higgs information may be biased to be $t\bar{t}H$ ($H \rightarrow b\bar{b}$) like, ending up with less discriminating kinematics against the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) when fed into the classification BDT.

The complete list of variables used in these two BDT is shown in Table 4.2. The distributions of these variables are shown for signal and background in Figures 4.9 and 4.10 for the ($\geq 6j, \geq 4b$) and (5j, $\geq 4b$) regions, respectively.

Variable	Region	
	$\geq 6j$	5j
Topological information from $t\bar{t}$		
t_{lep} mass	✓	✓
t_{had} mass	✓	–
Incomplete t_{had} mass	–	✓
W_{had} mass	✓	–
Mass of W_{had} and b from t_{lep}	✓	✓
Mass of W_{lep} and b from t_{had}	✓	✓
$\Delta R(W_{\text{had}}, b \text{ from } t_{\text{had}})$	✓	✓
$\Delta R(W_{\text{had}}, b \text{ from } t_{\text{lep}})$	✓	✓
$\Delta R(\text{lep}, b \text{ from } t_{\text{lep}})$	✓	✓
$\Delta R(\text{lep}, b \text{ from } t_{\text{had}})$	✓	✓
$\Delta R(b \text{ from } t_{\text{lep}}, b \text{ from } t_{\text{had}})$	✓	✓
$\Delta R(q_1 \text{ from } W_{\text{had}}, q_2 \text{ from } W_{\text{had}})$	✓	–
$\Delta R(b \text{ from } t_{\text{had}}, q_1 \text{ from } W_{\text{had}})$	✓	–
$\Delta R(b \text{ from } t_{\text{had}}, q_2 \text{ from } W_{\text{had}})$	✓	–
min. $\Delta R(b \text{ from } t_{\text{had}}, q \text{ from } W_{\text{had}})$	✓	–
min. $\Delta R(b \text{ from } t_{\text{had}}, q \text{ from } W_{\text{had}}) - \Delta R(\text{lep}, b \text{ from } t_{\text{lep}})$	✓	✓
Topological information from Higgs boson candidate		
Higgs candidate mass	✓	✓
Mass of Higgs candidate and q_1 from W_{had}	✓	✓
$\Delta R(b_1 \text{ from Higgs candidate}, b_2 \text{ from Higgs candidate})$	✓	✓
$\Delta R(b_1 \text{ from Higgs candidate}, \text{lep})$	✓	✓
$\Delta R(b_1 \text{ from Higgs candidate}, b \text{ from } t_{\text{lep}})$	–	✓
$\Delta R(b_1 \text{ from Higgs candidate}, b \text{ from } t_{\text{had}})$	–	✓

Table 4.2: Input variables to the reconstruction BDT in the single lepton channel. The subscript had (lep) indicates the hadronically (leptonically) decaying W or t and q_i refers to quarks from W .

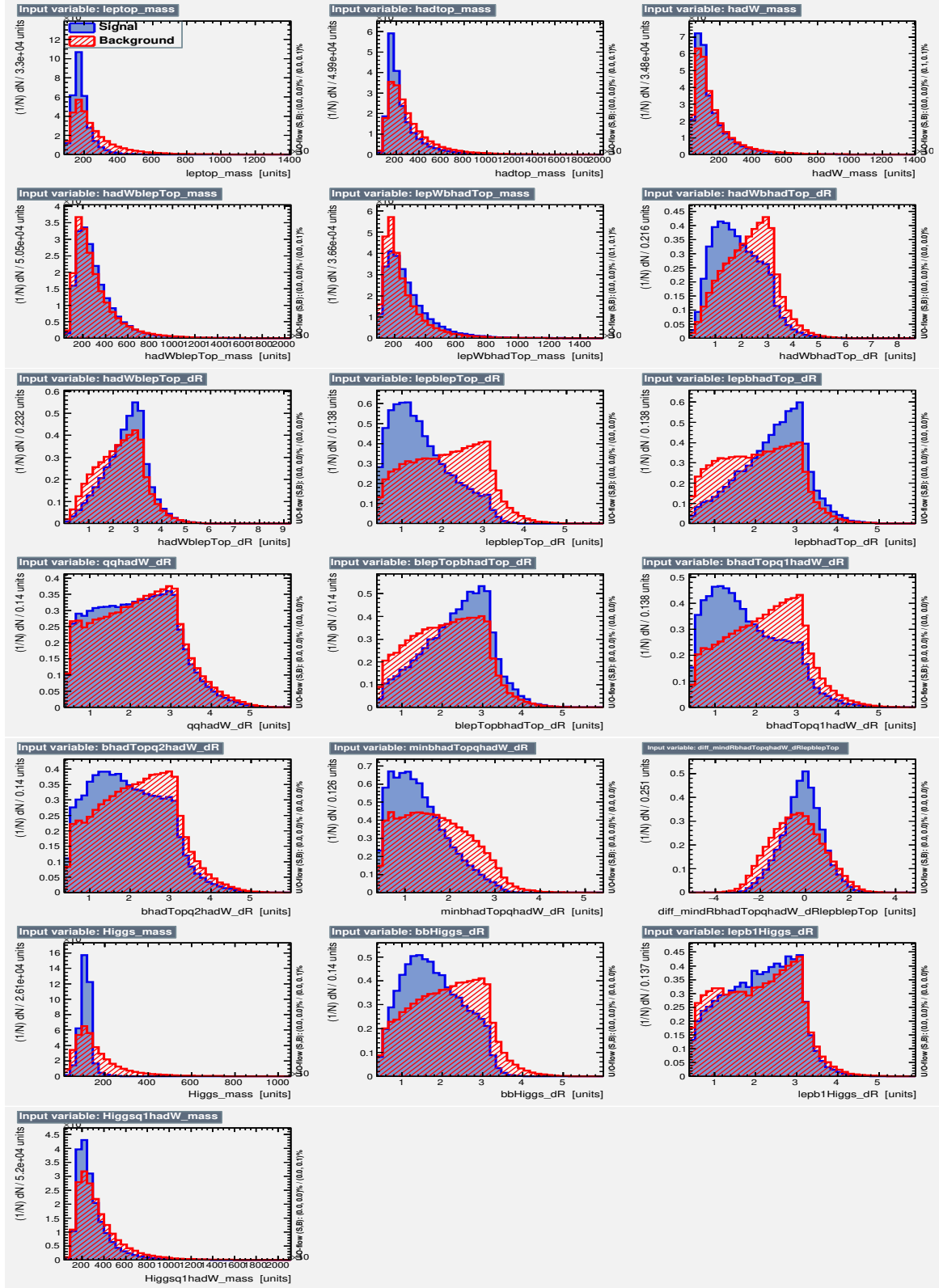


Figure 4.9: Input variables to the ($\geq 6j, \geq 4b$) reconstruction BDT. The signal is in blue while the background in red.

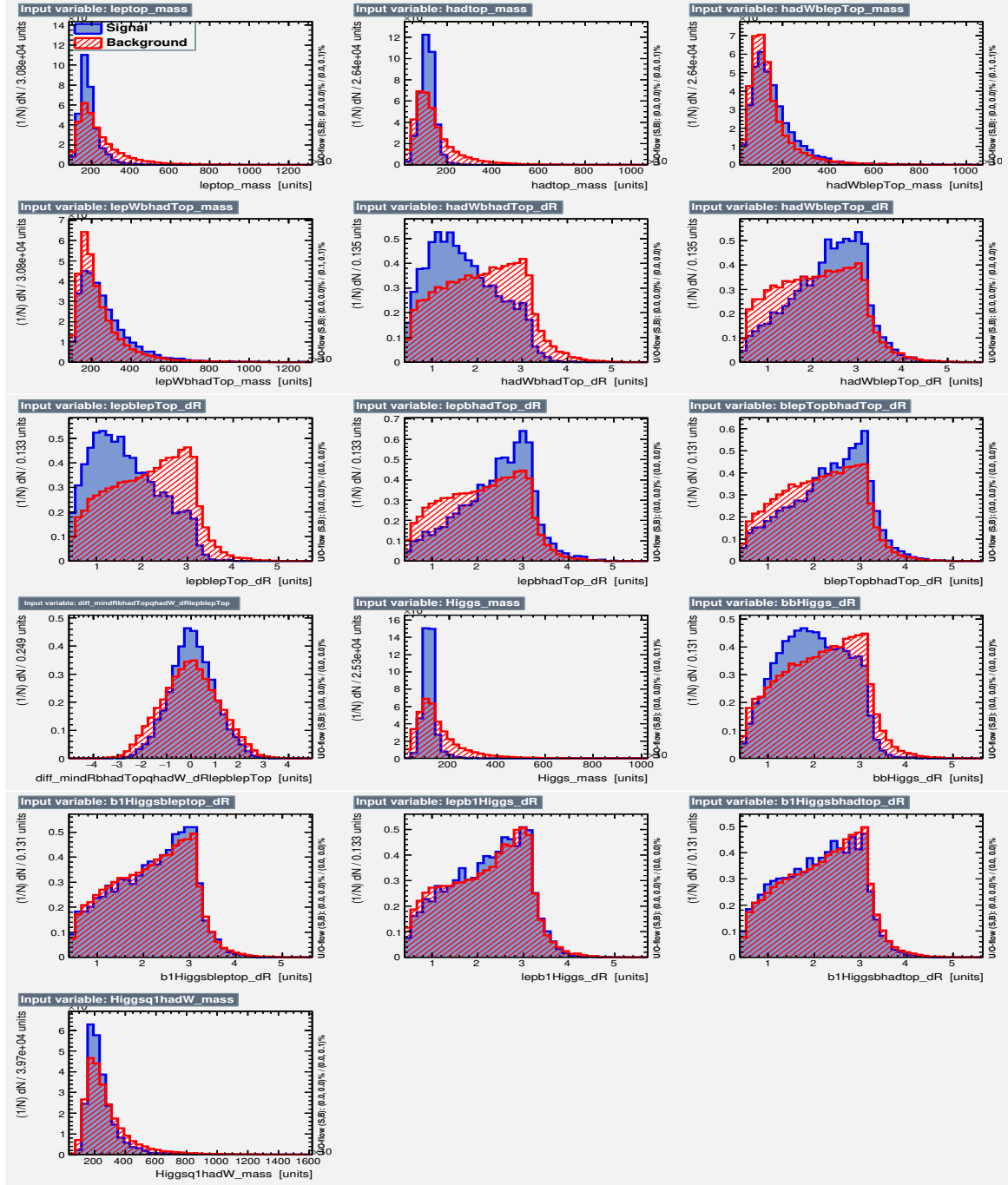


Figure 4.10: Input variables to the $(5j, \geq 4b)$ reconstruction BDT. The signal is in blue while the background in red.

Using b -tagging weights

To reduce the number of combinations, b quarks can only be associated to b -jets. In the previous iteration of the analysis [90] such jets were defined as passing a certain b -tagging efficiency threshold, and other jets were used for W decay products. However, it does happen relatively often that a light quark leads to a jet that has a higher b -tagging probability than some of the real b -jets. This effect is shown in Figure 4.11 where the jet ranked 4th sorted by b -tagging score has large probability of being a light jet from W . If an event is in this case, the light jet will be mis-used as b -jet, and no correct combination could be labeled for training the BDT. The effect is shown for both 85% and 70% tagging efficiency, and is much more mitigated in the latter case, as expected.

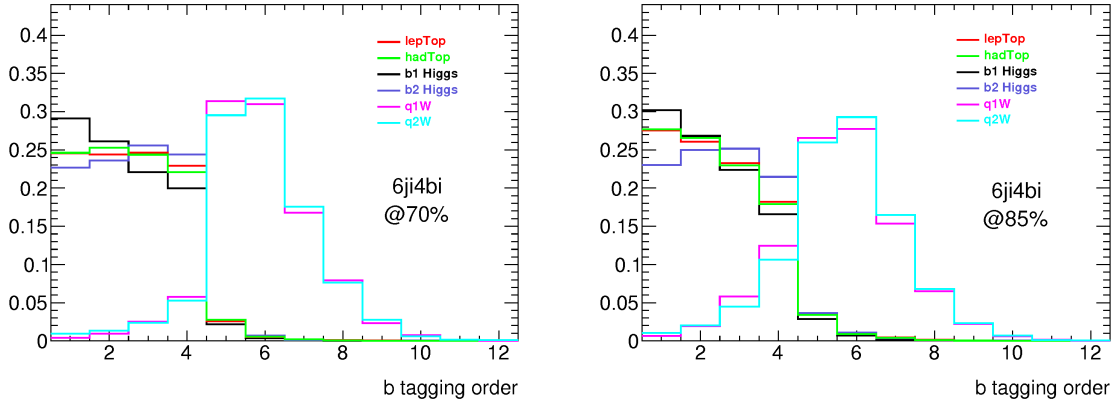


Figure 4.11: Origin of the jets, sorted in decreasing order of their MC2c10 b -tagging weight, in the ($\geq 6j, \geq 4b$) selection at 70% (left) and 85% (right) b -tagging efficiency.

Profiting from the pseudo-continuous b -tagging and inclusive training in the 6ji4bi@85% region, a new strategy is tested with the aim to have high signal statistics while limiting the background one for the reconstruction BDT. The b -tagging score is used in this new strategy: ranking jets by their MV2c10 b -tagging weight, the leading n jets are considered as b -jets, and the leading m ones cannot be used as light jets. Three scenarios were considered: ($n = 4, m = 4$), ($n = 5, m = 3$) and ($n = 6, m = 2$), in increasing number of possible combinations. Figure 4.12 shows the reconstruction performance comparison between the “b4l4” ($n = 4, m = 4$) and the “b5l3” ($n = 5, m = 3$) strategy. The former gives higher reconstruction efficiency almost for all objects.

The PC b -tagging weights of the six final-state jets are also explored, in addition to kinematics listed in Table 4.2. This BDT is referred to as recoBDT_withBTag. Figure 4.13 shows three BDTs including recoBDT_withBTag, and the two BDTs mentioned before. For the “b4l4” scenario, no improvement is brought in the reconstruction efficiency (in red) by using b -tagging weights, compared with the recoBDT_basic (in black) using kinematics only. In “b5l3” scenario, light jets can be used as b -jets and the b -tagging can help to identify the truth matching, the recoBDT_withBTag (in light blue) brings much more improvement than the recoBDT_basic (in magenta). In training samples of this scenario, the four b -jets are already fixed and the light-jets are not allowed to be used as b -jets, from which the b -tagging cannot help further. Comparing the “b4l4” and “b5l3” cases, the recoBDT_withH of “b4l4” gives the highest matching efficiency almost for all objects. The “b5l3” recoBDT_withBTag using b -tagging weight recovers the performance of the “b4l4” one, but no more improvement from it.

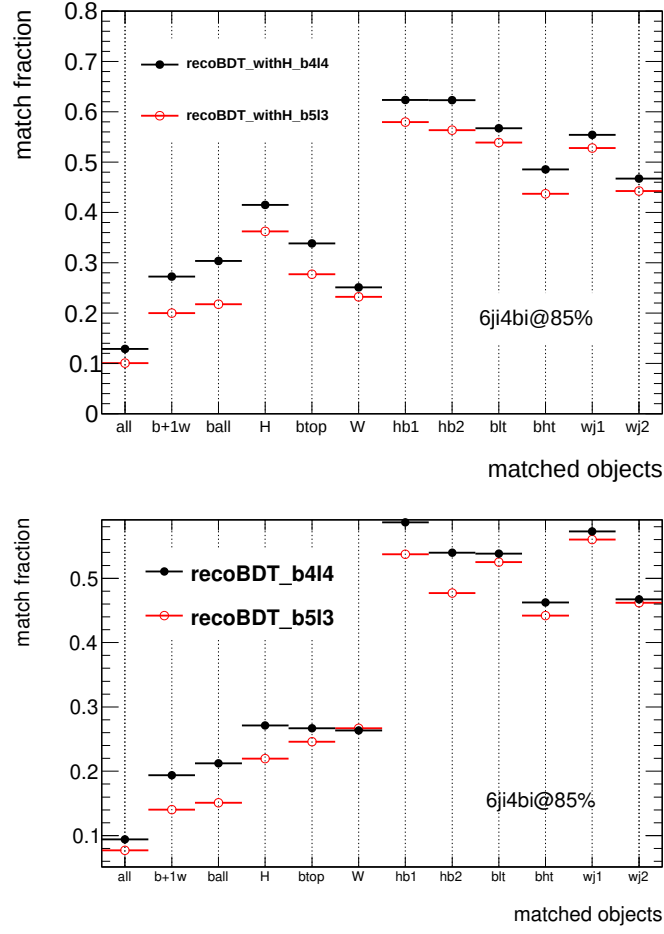


Figure 4.12: Performance of $(n = 4, m = 4)$ [black] and $(n = 5, m = 3)$ [red] reconstruction scenarios for BDT with (top) and without (bottom) Higgs-related input variables in the $(\geq 6j, \geq 4b)$ region at 85% b -tagging efficiency.

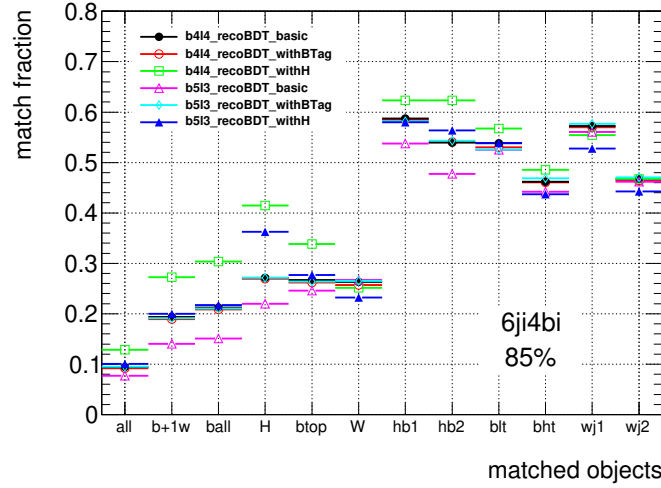


Figure 4.13: Performance of “b4l4” (“b5l3”) reconstruction scenarios for reconstruction BDTs in the $(\geq 6j, \geq 4b)$ region at 85% b -tagging efficiency. Three BDTs are trained using various input features: with Higgs-related input variables shown in green squares (blue solid triangles), without Higgs-related input variables shown in black disks (magenta triangles), and with b -tagging weights shown in red circles (light blue rhombuses).

The recoBDT_withBTag is also tested with even looser requirement of b -tagging ordering when selecting b -jets, the “b6l2” where the leading 6 jets are considered as b -jets, and the leading 2 ones cannot be used as light jets. The matching efficiency is compared with the recoBDT_withBTag of “b4l4” and “b5l3” scenarios in Figure 4.14. It shows that all three scenarios have similar performance, using b -tagging weights as inputs in addition to the kinematics.

BDT parameter optimization

The BDT parameters are optimized in order to further improve the reconstruction performance. Parameters optimized for the ICHEP [92] analysis are used as the default values, and variations from them are made. All the optimized parameters are shown in Table 4.3, including:

- The two boosting methods and their corresponding learning rate, AdaBoostBeta for AdaBoost, and Shrinkage for Gradboost,
- Ntrees, the number boosted trees in the forest,
- Ncuts, the number of grid points of input variables used to find optimal cuts in node splitting,
- MaxDepth and MinNodeSize, stopping conditions illustrated in Section 4.5.1.

For each of the boosting methods, combinatorics of all these variations are made, and the BDT is trained and tested with each of these configurations. The matching efficiencies of all the objects listed in Section 4.5.3.1 are used as metric to choose the best configuration. None of these variations lead to noticeable improvements compared to the default setting which is given in bold in Table 4.3.

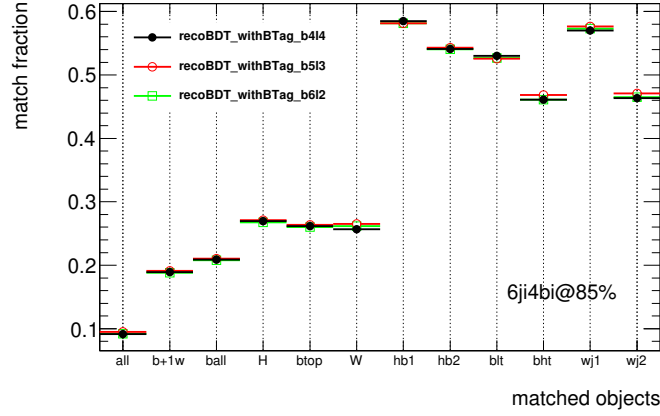


Figure 4.14: Performance of $(n = 4, m = 4)$, $(n = 5, m = 3)$ and $(n = 6, m = 2)$ reconstruction scenarios for BDT with MV2c10 weights as input variables in the $(\geq 6j, \geq 4b)$ region at 85% b -tagging efficiency.

Parameter	Variation
BoostType	AdaBoost, GradBoost
AdaBoostBeta/Shrinkage	(0.15 , 0.5)/(0.1, 1)
Ntrees	400 , 100
MinNodeSize	4% , 1%
MaxDepth	3, 5 , 10
Ncuts	80 , 10

Table 4.3: The variation of the reconstruction BDT parameters for optimization. The AdaBoost-Beta is the parameter varied with AdaBoost method, while the Shrinkage is the one corresponding to the GradBoost boosting algorithm. The chosen configuration is in bold.

4.5.3.4 Performance

Training regions The analysis is split into several regions (see Sec. 4.4.1), possibly leading to many different BDTs. With the hope to simplify the analysis, the reconstruction performance was compared between dedicated trainings (e.g. using only events in the UPSR) and inclusive trainings on all events with at least 4 b -tagged jets at 85% efficiency (one training for $\geq 6j$, one for $5j$). Results are shown in Figure 4.15 when applying these trainings in the UPSR. Other dedicated trainings on events with exclusively 6 jets were also studied, which results are presented in Figure 4.16 when applying these trainings in the exclusive 6j region. In both cases, exclusive training performance is equivalent to or worse than the inclusive training, which has larger statistics in the training set. Therefore the inclusive training is chosen as the default.

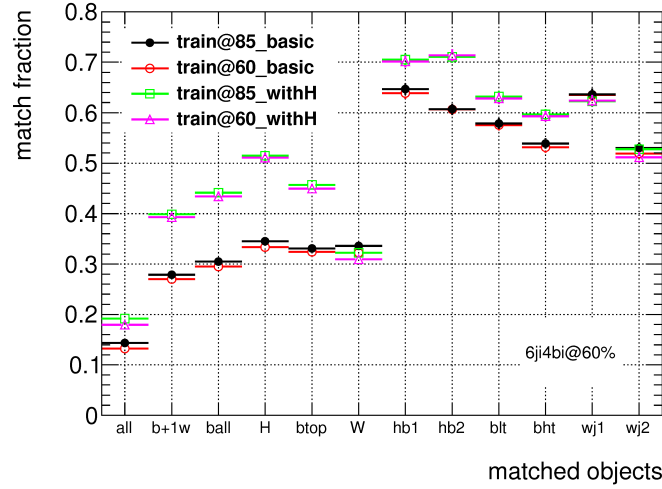


Figure 4.15: Performance of reconstruction in the ($\geq 6j$, $\geq 4b$) region, evaluated in the UPSR region, for BDT with (green squares and magenta triangles) and without (black disks and red circles) Higgs-related variables, trained in the UPSR (red and magenta) or inclusively (black and green) in 85% b -tagging efficiency events.

Reconstruction performance The reconstruction BDT using the “b4l4” strategy is adopted in this analysis and is trained in the inclusive 6ji4bi@85% and 5je4bi@85% regions. Figure 4.17 shows the matching efficiencies given by this new reconstruction BDT in the $\geq 6j$ region, together with the reconstruction BDT using the fixed b -tagging WP at 85% without the “b4l4” jet sorting strategy introduced in this thesis as was done in the ICHEP analysis. The new BDT brings around 5% improvement of the Higgs boson truth matching fraction, and almost 10% for all b -jets truth matching. Figure 4.18 shows the reconstruction efficiencies for the 5-jet case.

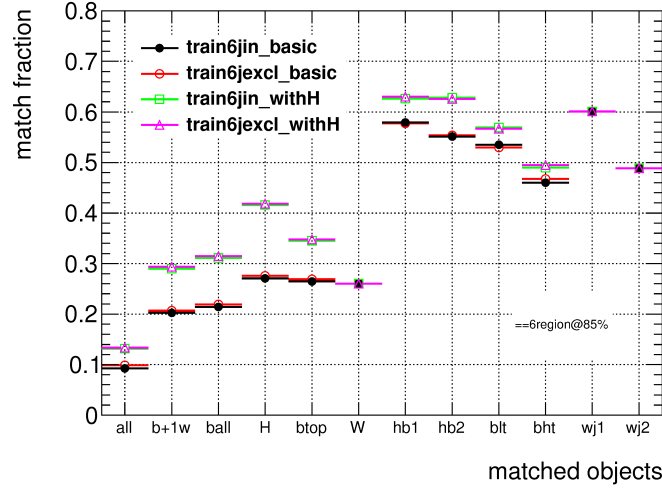


Figure 4.16: Performance of reconstruction in the $(\geq 6j, \geq 4b)$ region, evaluated in the $= 6j$ region, for BDT with (green squares and magenta triangles) and without (black disks and red circles) Higgs-related variables, trained in the $= 6j$ region (red and magenta) or inclusively $\geq 6j$ region (black and green) in 85% b -tagging efficiency events.

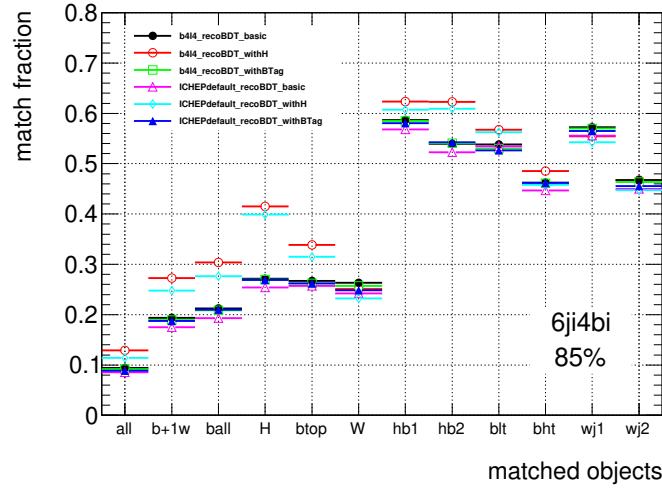


Figure 4.17: Performance of the $(n = 4, m = 4)$ reconstruction scenario for BDT with (red circles) and without (black disks) Higgs-related, and with (green squares) b -tagging weights input variables in the $(\geq 6j, \geq 4b)$ region at 85% b -tagging efficiency. The reconstruction BDT trained without the $(n = 4, m = 4)$ reconstruction strategy using the fixed b -tagging WP at 85% is also shown, with (light blue rhombuses) and without (magenta triangles) Higgs-related, and with (blue solid triangles) b -tagging input variables.

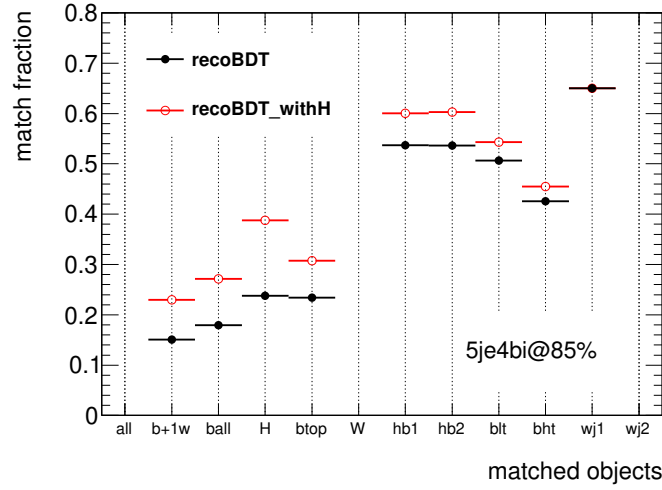


Figure 4.18: Performance of the $(n = 4, m = 4)$ reconstruction scenario for BDT with (red) and without (black) Higgs-related input variables in the $(5j, \geq 4b)$ region at 85% b -tagging efficiency.

4.5.4 Discrimination between signal and background

The BDT to discriminate the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) and $t\bar{t} + \text{jets}$ is trained including information from the reconstruction MVAs. The full strategy is illustrated in Figure 4.19 where it shows that the *classification BDT* learns from various information:

- Each of the two reconstruction BDT trained to learn the $t\bar{t}H$ signature chooses the best combination and this allows to build two sets of variables based on these particular combinations.
- The output of intermediate MVAs are also used as inputs, which are the most discriminating variables.
- In addition, the combination-independent kinematics are exploited as they are discriminant between $t\bar{t}H$ and $t\bar{t}$, and they help to further improve the performance.
- The b -tagging variables help the classification BDT to get the same good performance everywhere in the event space in which the b -tagging properties of both $t\bar{t}H$ and $t\bar{t}$ are different. This is firstly introduced in this thesis.

Training parameters for the classification BDT are given in Table 4.4 which is optimized from the last round of analysis.

BoostType	AdaBoost
AdaBoostBeta	0.15
NTrees	250
MaxDepth	5
nCuts	80
MinNodeSize	4%

Table 4.4: TMVA parameters for the classification BDT.

The input list used in the published analysis are detailed in Section 4.5.4.1. For the first time, the b -tagging variables are used and the study is shown in Section 4.5.4.2, where the default

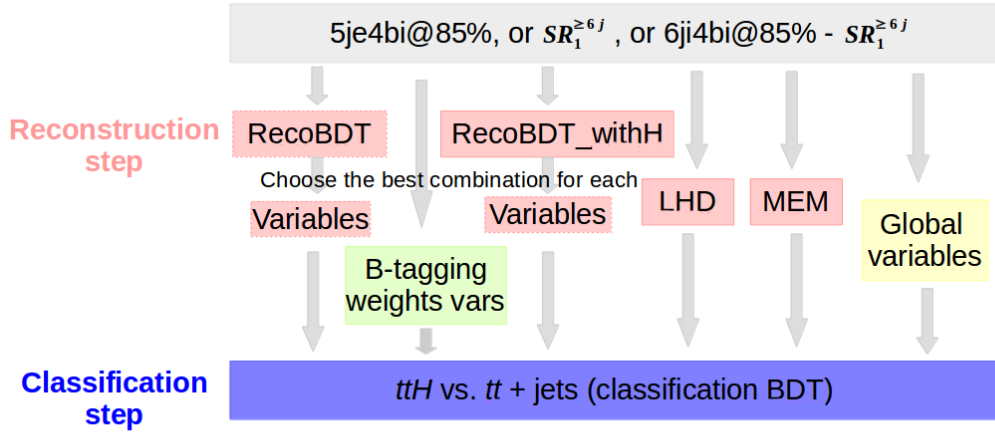


Figure 4.19: The two-step MVA strategy.

input list of the classification BDT is presented. Then a study to reduce the input list while keeping the classification performance is presented in Section 4.5.4.3. The BDT classification performance is measured using the ROC curves and the Area Under the Curve (AUC).

4.5.4.1 Discriminating variables

The input features of the BDT are based on the ones optimized for the previous iteration of the analysis which is shown in the first two sections of Table 4.5. The first section is the combination-independent kinematics, and the second is the variables built on the best combination which has the highest score of either reconstruction BDTs with or without using Higgs related features. The output of the recoBDT with Higgs information is also used as an important feature. Then the third and the fourth sections are the outputs of the likelihood discriminant and the matrix element method, the latter being only available in the UPSR due to the large computing source consumption.

Table 4.5: Input variables to the classification BDTs. For variables from the reconstruction BDT, those with a * are from the BDT using Higgs-boson information, those with no * are from the BDT without Higgs-boson information. The MEM_{D1} variable is only used in $\text{SR}_1^{\geq 6j}$, while variables based on the b -tagging discriminant are not used in this region.

Variable	Definition	$\text{SR}_{1,2,3}^{\geq 6j}$	$\text{SR}_{1,2}^{5j}$
General kinematic variables			
$\Delta R_{bb}^{\text{avg}}$	Average ΔR for all b -tagged jet pairs	✓	✓
$\Delta R_{bb}^{\text{max } p_T}$	ΔR between the two b -tagged jets with the largest vector sum p_T	✓	–
$\Delta \eta_{jj}^{\text{max}}$	Maximum $\Delta \eta$ between any two jets	✓	✓
$m_{bb}^{\text{min } \Delta R}$	Mass of the combination of two b -tagged jets with the smallest ΔR	✓	–
$m_{jj}^{\text{min } \Delta R}$	Mass of the combination of any two jets with the smallest ΔR	–	✓
$N_{bb}^{\text{Higgs } 30}$	Number of b -tagged jet pairs with invariant mass within 30 GeV of the Higgs-boson mass	✓	✓
H_T^{had}	Scalar sum of jet p_T	–	✓
$\Delta R_{\ell,bb}^{\text{min}}$	ΔR between the lepton and the combination of the two b -tagged jets with the smallest ΔR	–	✓
Aplanarity	$1.5\lambda_2$, where λ_2 is the second eigenvalue of the momentum tensor [125] built with all jets	✓	✓
H_1	Second Fox–Wolfram moment computed using all jets and the lepton	✓	✓
Variables from reconstruction BDT			
BDT output	Output of the reconstruction BDT	✓*	✓*
m_{bb}^{Higgs}	Higgs candidate mass	✓	✓
$m_{H,b_{\text{lep top}}}$	Mass of Higgs candidate and b -jet from leptonic top candidate	✓	–
$\Delta R_{bb}^{\text{Higgs}}$	ΔR between b -jets from the Higgs candidate	✓	✓
$\Delta R_{H,t\bar{t}}$	ΔR between Higgs candidate and $t\bar{t}$ candidate system	✓*	✓*
$\Delta R_{H,\text{lep top}}$	ΔR between Higgs candidate and leptonic top candidate	✓	–
$\Delta R_{H,b_{\text{had top}}}$	ΔR between Higgs candidate and b -jet from hadronic top candidate	–	✓*
Variables from likelihood and matrix element method calculations			
LHD	Likelihood discriminant	✓	✓
MEM_{D1}	Matrix element discriminant (in $\text{SR}_1^{\geq 6j}$ only)	✓	–
Variables from b -tagging (not in $\text{SR}_1^{\geq 6j}$)			
$w_{b\text{-tag}}^{\text{Higgs}}$	Sum of b -tagging discriminants of jets from best Higgs candidate from the reconstruction BDT	✓	✓
B_{jet}^3	3 rd largest jet b -tagging discriminant	✓	✓
B_{jet}^4	4 th largest jet b -tagging discriminant	✓	✓
B_{jet}^5	5 th largest jet b -tagging discriminant	✓	✓

The intermediate MVAs output are dominant discriminants of the classification BDT. In Table 4.6, the performance of the classification BDT trained and tested in the (6ji4bi@60%) region are shown, with or without intermediate MVA outputs as input features. Using additionally the output of the reconstruction BDT (LHD) gives 2.4% (3.4%) AUC improvement. Adding MEM brings not much improvement, since it is correlated with the LHD. The linear correlation matrix of input variables for $t\bar{t}H$ signal and $t\bar{t}$ background are shown in Figure 4.20, where the largest correlation is between the LHD and MEM. The correlation between the LHD and reconstruction BDT output is less and therefore there is extra information that brings performance improvement when using both variables.

The input discriminating variables are removed if they are badly modeled. Figures 4.21 and 4.22 present some important input discriminating variable distributions, which shows the consistency between data and MC simulations.

Some variations on top of these features are also studied. The tests are made using Higgs boson mass and the angular distance between the two b -jets from the Higgs candidate built based on the recoBDT instead of the recoBDT_withH. Although the latter could better reconstruct the Higgs mass, it may bias the background and lose the discrimination from the signal. The test shows a slightly worse classification performance (less than 1%) using the recoBDT to build the Higgs related kinematics, and the recoBDT_withH is thus used.

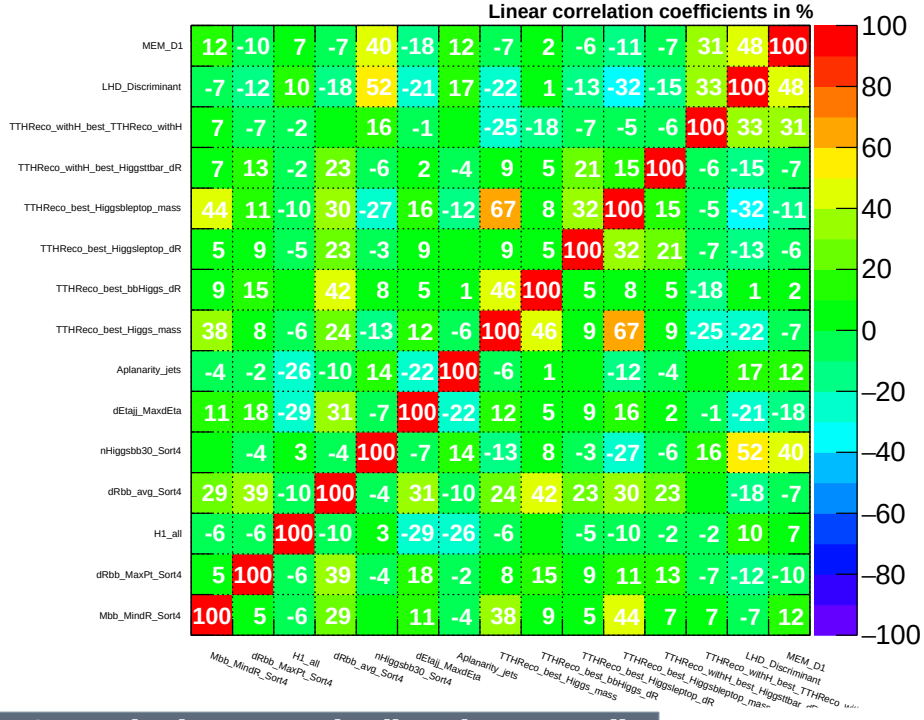
4.5.4.2 Using b -tagging as BDT input features

The b -tagging variables are also explored as BDT input features with the aim to further improve the classification performance. Four b -tagging variables are used as listed in the fifth section of Table 4.5. The inclusive training (one training in events with $\geq 4b$ with 85% b -tagging efficiency) performance of the classification BDT is shown in Figure 4.23 and 4.24 for the ($\geq 6j, \geq 4b$) and ($5j, \geq 4b$) regions, respectively, after evaluation in each of the event categories defined in Section 4.4.1. The ROC curves using the b -tagging weights (with full input list shown in Table 4.5 except MEM) gives better AUC than the curves without using b -tagging variables. All input variables in the inclusive regions are shown in Figures 4.25 and 4.26.

	AUC	improvement
Kinematics	0.738	-
Kins + RecoBDT	0.756	2.4%
Kins + LHD	0.763	3.4%
Kins + RecoBDT + LHD	0.768	4.1%

Table 4.6: The classification BDT performance improvement using either outputs of the reconstruction BDT and LHD. The BDT is trained and performed in the UPSR (6ji4bi@60% region).

Correlation Matrix (signal)



Correlation Matrix (background)

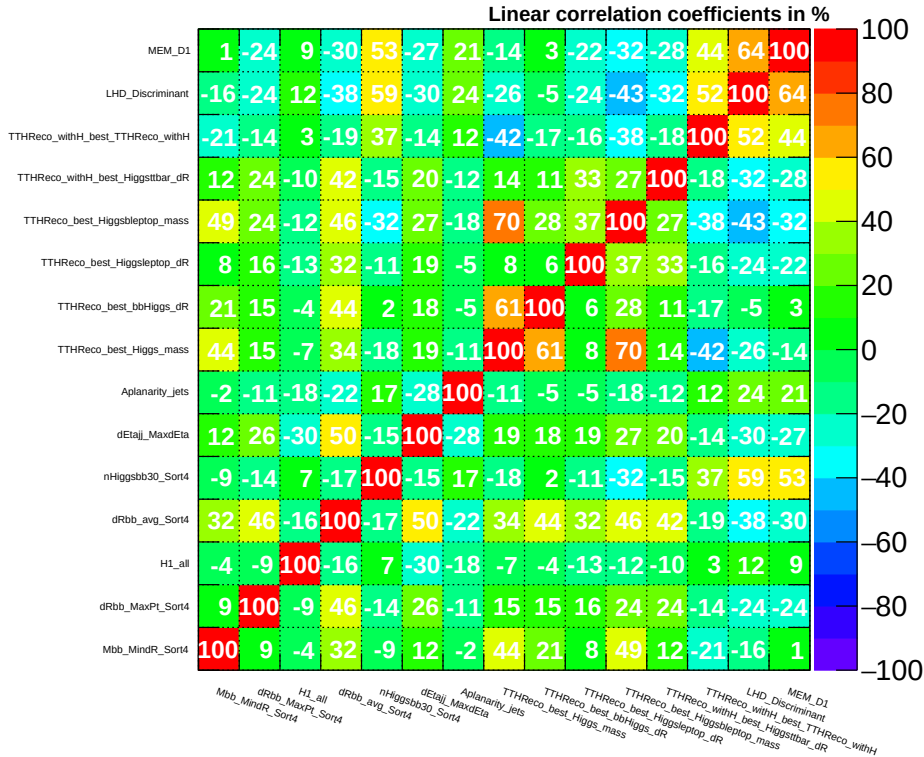
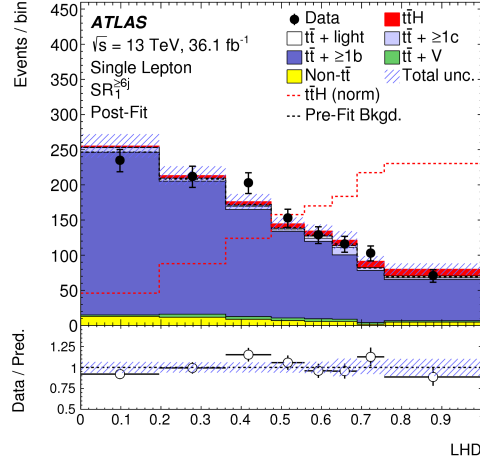
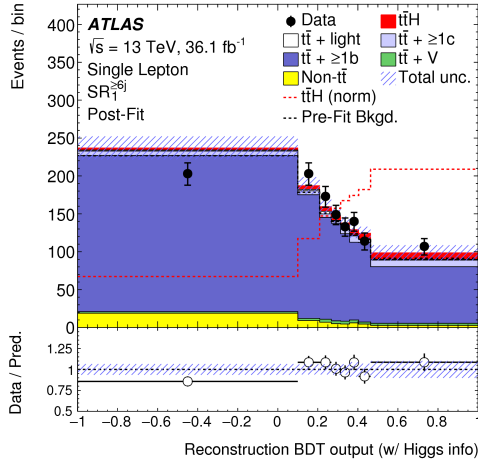


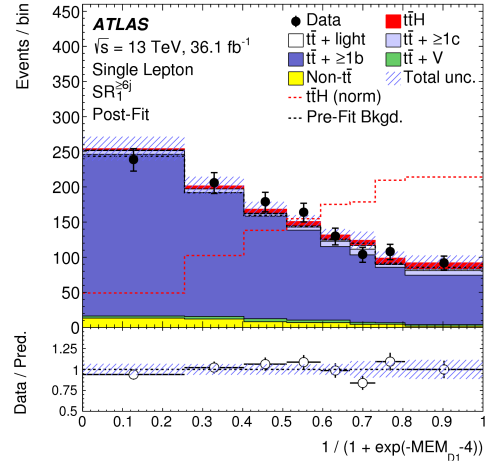
Figure 4.20: Correlation matrix for signal (top) and background (bottom) of input variables in the ($\geq 6j$, $\geq 4b$) UPSR classification BDT. The LHD, MEM, and the reconstruction BDT (with Higgs information) output are denoted as LHD_Discriminant, MEM_D1, and TTHReco_withH_best_TTHReco_withH, respectively, at the top right corner.



(a)



(b)



(c)

Figure 4.21: Comparison between data and prediction for important discriminating input variables to the classification BDT: (a) LHD, (b) RecoBDT_withH, and (c) MEM, in $SR_1^{\geq 6j}$, after the combined fit to the data (see Section 4.8.2). The $t\bar{t}H$ signal yield (solid red) is normalized to the SM cross-section before the fit and to the fitted μ after the fit. The dashed black line shows the pre-fit total background prediction. The dashed red line shows the $t\bar{t}H$ signal distribution normalized to the total background prediction.

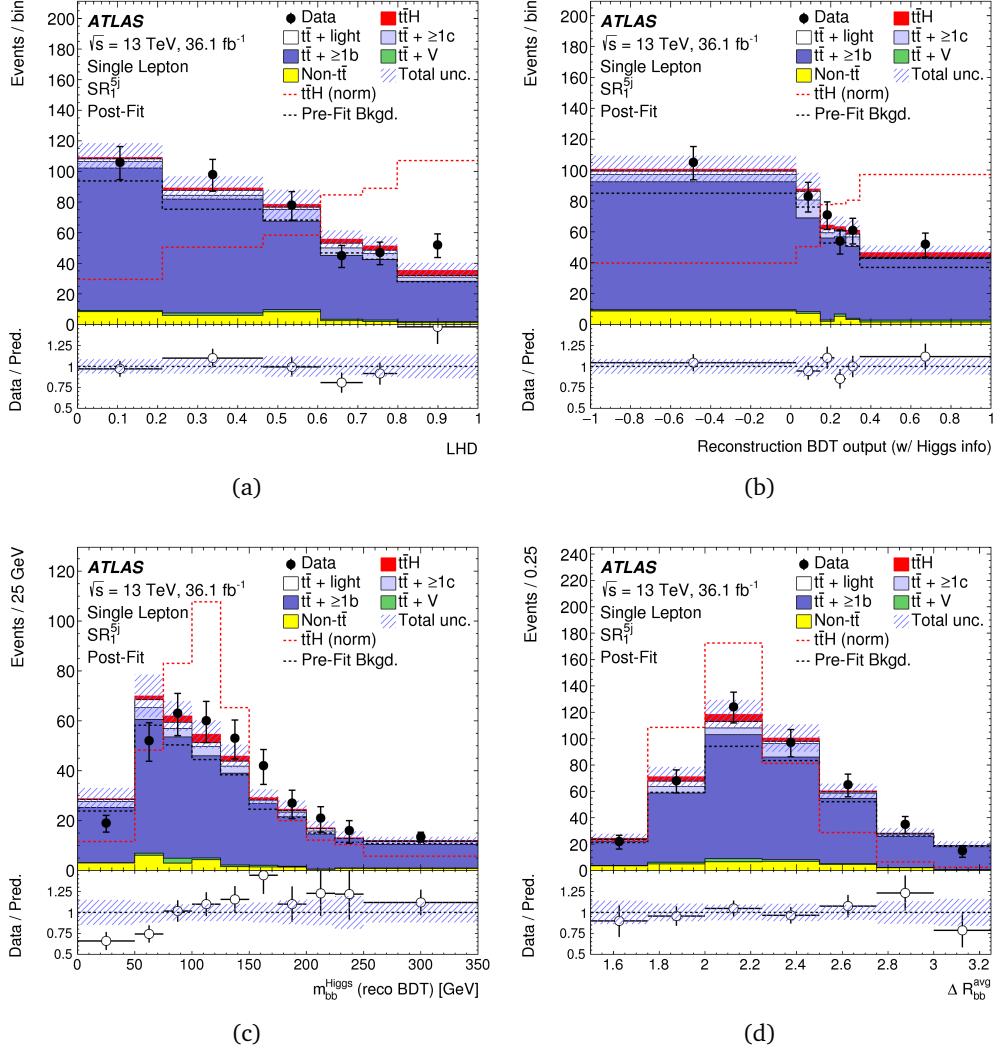


Figure 4.22: Comparison between data and prediction for important discriminating input variables to the classification BDT: (a) LHD, (b) RecoBDT_withH, and (c) MEM, in SR_1^{5j} , after the combined fit to the data (see Section 4.8.2). The $t\bar{t}H$ signal yield (solid red) is normalized to the SM cross-section before the fit and to the fitted μ after the fit. The dashed black line shows the pre-fit total background prediction. The dashed red line shows the $t\bar{t}H$ signal distribution normalized to the total background prediction.

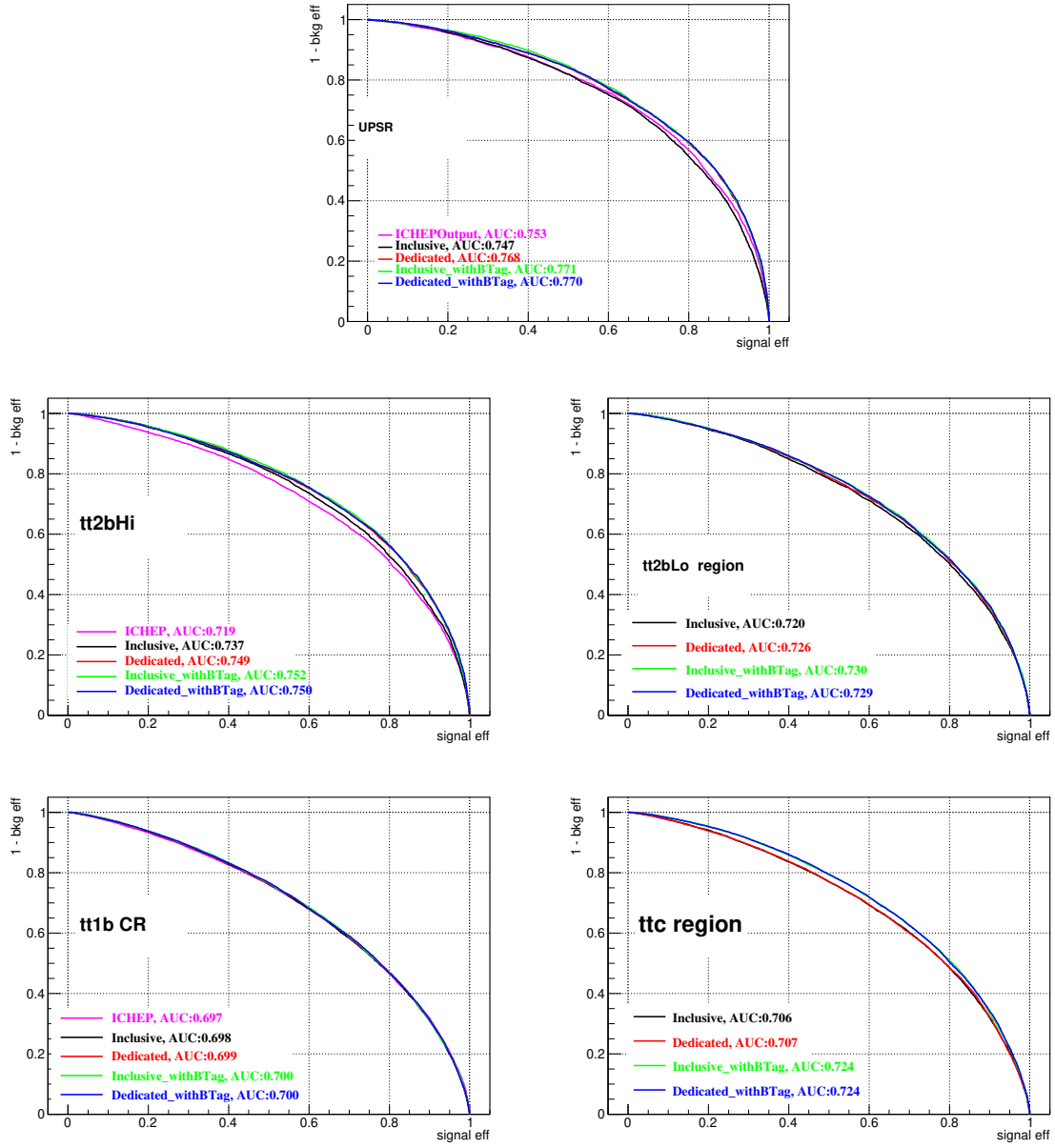


Figure 4.23: Performance of classification BDT, comparing dedicated and inclusive trainings, with or without b -tagging input variables, in the various categories within the $(\geq 6j, \geq 4b)$ region. Top: UPSR; middle left: $t\bar{t} + \geq 2b$ Hi SR; middle right: $t\bar{t} + \geq 2b$ Lo SR; bottom left: $t\bar{t} + 1b$ CR; bottom right: $t\bar{t} + \geq 1c$ CR.

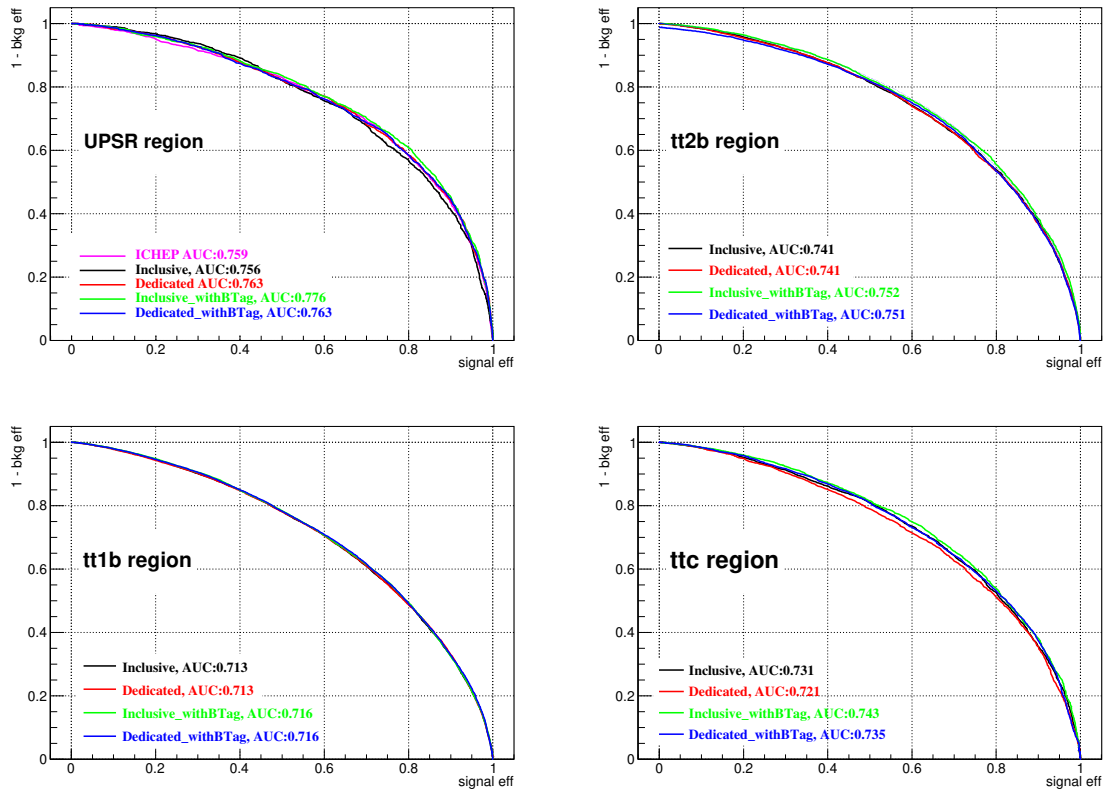


Figure 4.24: Performance of classification BDT, comparing dedicated and inclusive trainings, with or without b -tagging input variables, in the various categories within the $(5j, \geq 4b)$ region. Top left: UPSR; top right: $t\bar{t} + \geq 2b$ SR; bottom left: $t\bar{t} + 1b$ CR; bottom right: $t\bar{t} + \geq 1c$ CR.

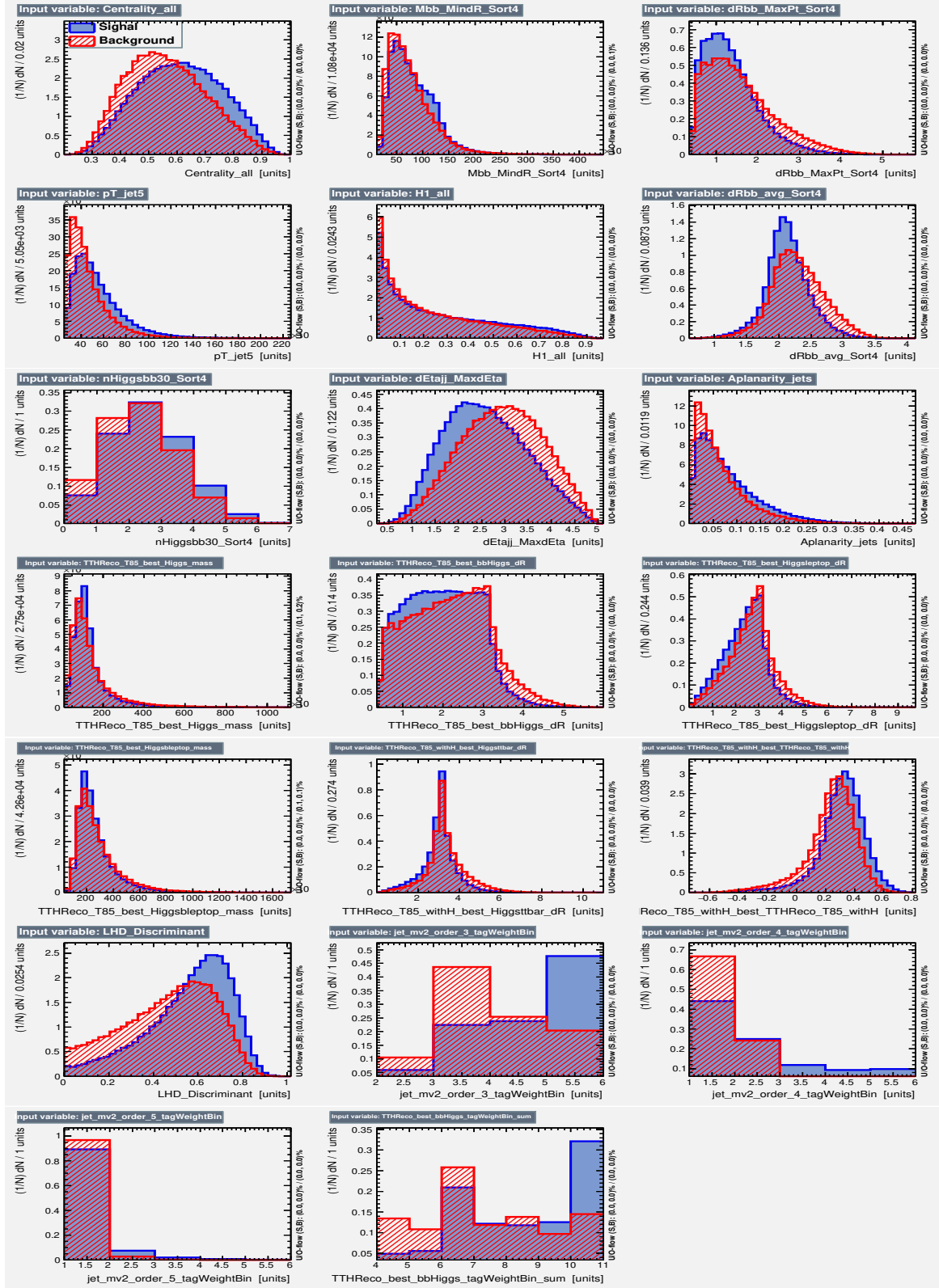


Figure 4.25: Input variables to the ($\geq 6j, \geq 4b$) classification BDT. The signal is in blue while the background in red.

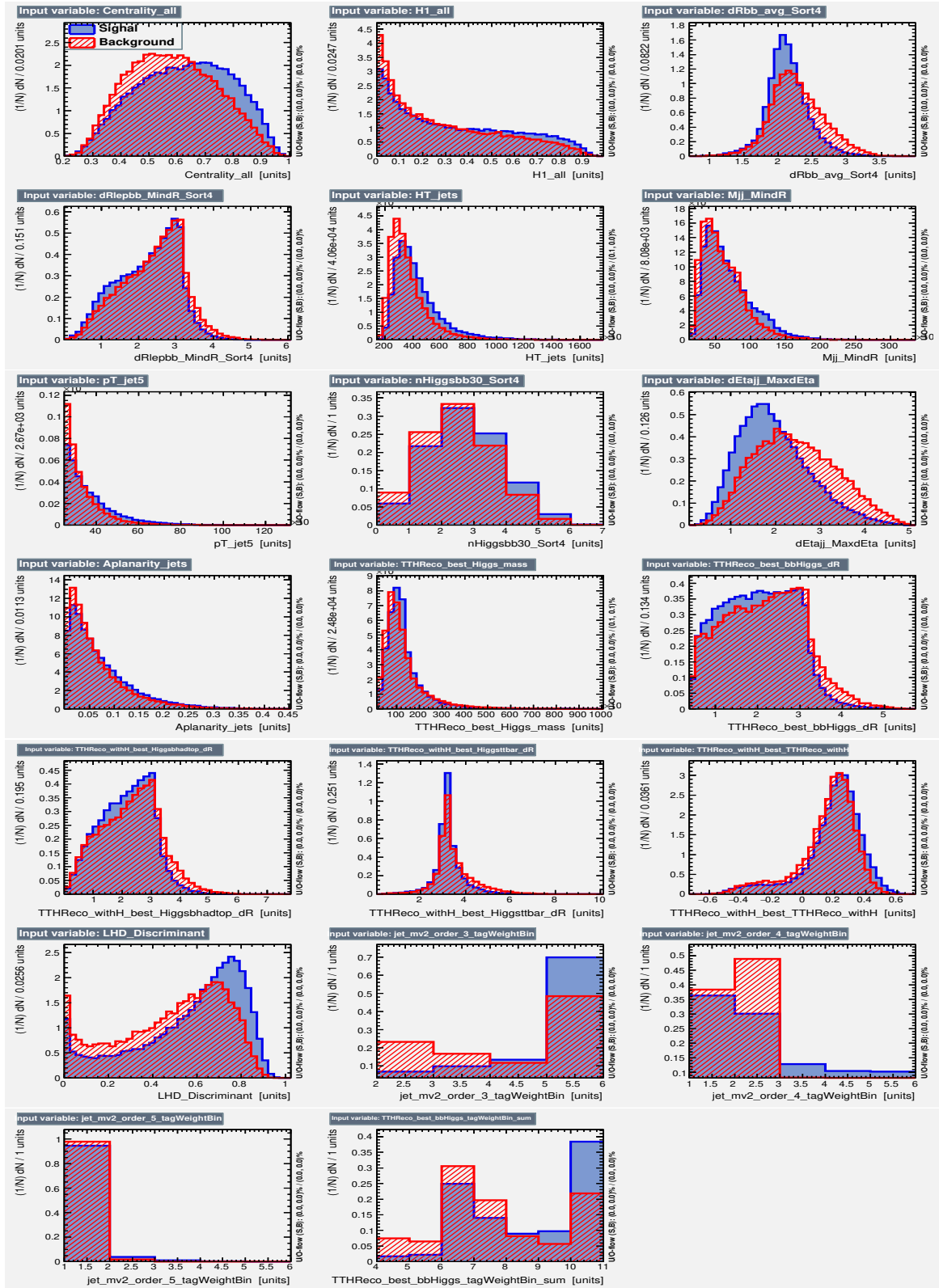


Figure 4.26: Input variables to the $(5j, \geq 4b)$ classification BDT. The signal is in blue while the background in red.

Inclusive vs exclusive training

In each of the event categories, dedicated trainings are also tested. This leads to different BDTs. The corresponding results are also presented and compared to the inclusive ones in Figure 4.23 and 4.24. For the dedicated training, using b -tagging weights in addition does not bring improvements. This is due to the fact that the categories are defined by having requirement on the jet b -tagging in each event. Therefore the b -tagging weights are not discriminating anymore for signal and background in each region.

Comparing the dedicated to the inclusive training, the latter performs as well or even better in each region, on the condition that MV2c10 b -tagging variables of some of the jets are included in the training. In inclusive regions, the training benefits from large statistics while b -tagging scores are discriminating for $t\bar{t}H$ and $t\bar{t}$ and have almost no correlation with other variables as shown in Figure 4.27 for $\geq 6j$ events. It was therefore decided to use inclusive trainings, except for the $SR_1^{\geq 6j}$ region where one can still gain a little bit of performance if adding the MEM output (see Section 4.5.4.1).

Negative weight events

In order to maximise the size of the training sample, the absolute value of the event weight is taken, so that events with negative weights are also used. Input variables are checked to be in agreement between events with positive and negative weights. The performance is also similar when assessed on the full sample with positive and negative weights, using in the training all events with their absolute weight or only events with a positive weight (see Figure 4.28).

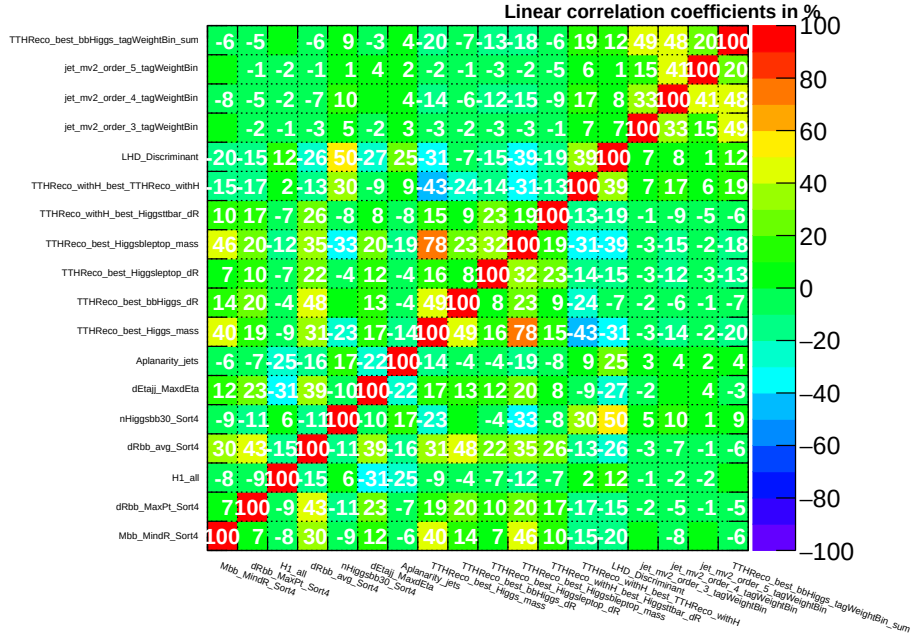
Performance

This section reports the performance of the BDT using the full list of variables from Table 4.5, trained and tested on the simulations used for the final fit. The training is inclusive in the $(5j, \geq 4b)$ regions and in all $(\geq 6j, \geq 4b)$ regions except in the UPSR, where the MEM variable is added and a separate BDT is retrained. The b -tagging weight variables are not used in the dedicated training where all events are in the same b -tagging weight bin (This is true for b -tagging variables: w_b^H , B_{j3} , B_{j4} . For B_{j5} , not all but most events are in the fifth bin.). Figures 4.29 and 4.30 show the ROC curves for all regions in the $(\geq 6j, \geq 4b)$ and $(5j, \geq 4b)$ categories, respectively.

4.5.4.3 Reducing the BDT input variable list

With the goal to have good performance but with as few as possible variables for the classification BDT, an optimization procedure was designed, and performed only in the UPSRs of ≥ 6 jets and $= 5$ jets events respectively. Starting from the most discriminant variables (reco BDT output, likelihood discriminant), each variable from a list of 30 (26) pre-selected ones for $\geq 6j$ ($= 5j$) shown in appendix Table 0.12 (Table 0.14) expected to be discriminating, was added one by one. The one leading to the best improvement in AUC is retained, and so on iteratively until the same performance as the baseline list that contains 15(14) variables plus 4 b -tagging variables for $\geq 6j$ ($= 5j$) as shown in Table 4.5 was achieved. The evolution of the AUC performance of the BDT using more and more input variables in addition to the two intermediate MVA outputs are shown in Figure 4.31. For the $\geq 6j$ scenario, the optimal list of variables in the UPSR was chosen to be with 7 variables (or 8 when adding the MEM) in

Correlation Matrix (signal)



Correlation Matrix (background)

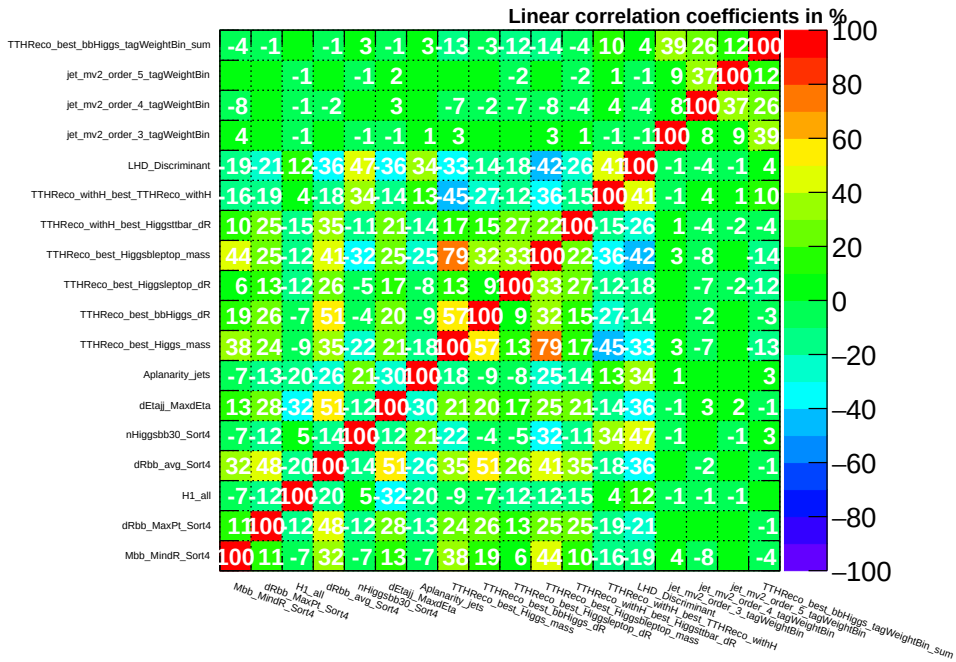


Figure 4.27: Correlation matrix for signal (top) and background (bottom) of input variables to the ($\geq 6j, \geq 4b$) inclusive classification BDT, with b -tagging variables. The four b -tagging variables are denoted as jet_mv2_order_3_tagWeightBin, jet_mv2_order_4_tagWeightBin, jet_mv2_order_5_tagWeightBin, and TTHReco_best_bbHiggs_tagWeightBin_sum.

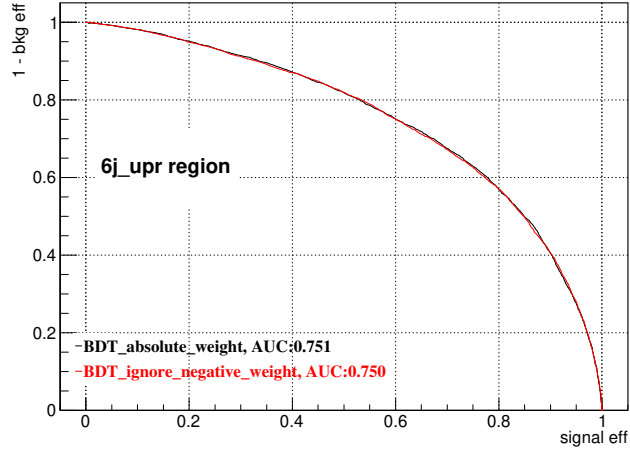


Figure 4.28: Classification BDT ROC curves in the ($\geq 6j, \geq 4b$) UPSR, when trained on events with positive weights only (red) or on all events using the absolute value of the weight (black).

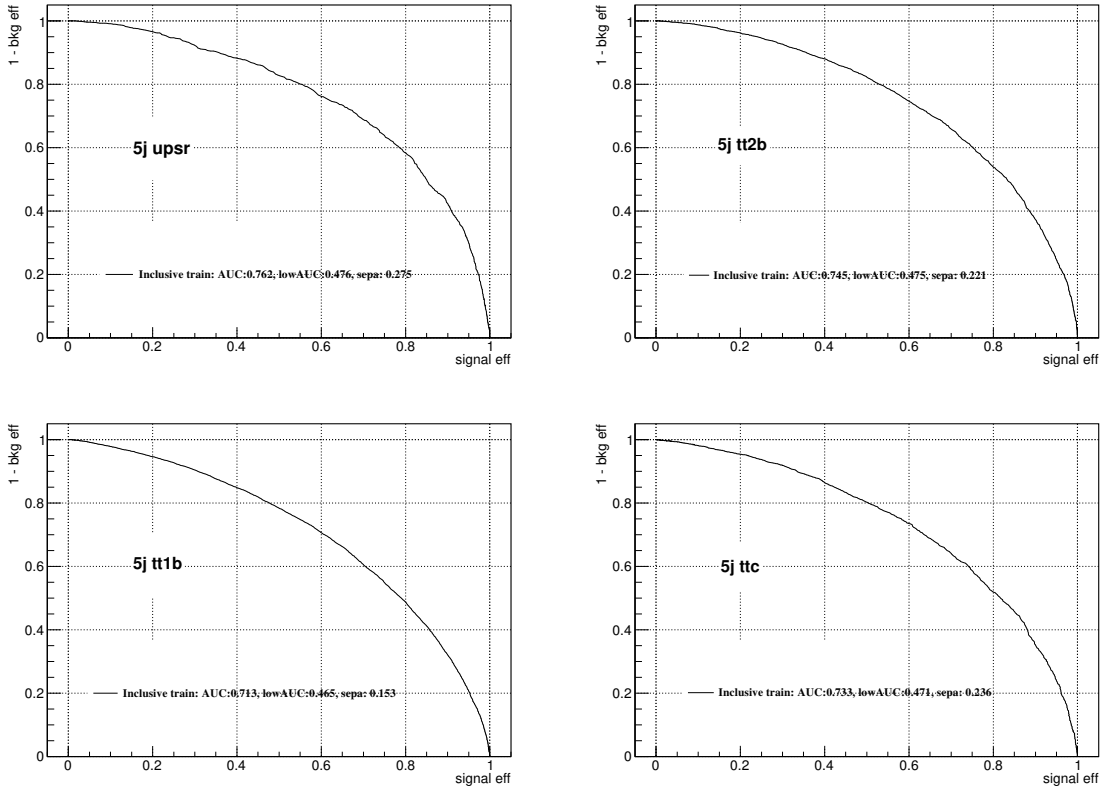


Figure 4.29: Performance of classification BDT in the various categories within the ($5j, \geq 4b$) region. Top left: UPSR; top right: $t\bar{t} + \geq 2b$ SR; bottom left: $t\bar{t} + 1b$ CR; bottom right: $t\bar{t} + \geq 1c$ CR.

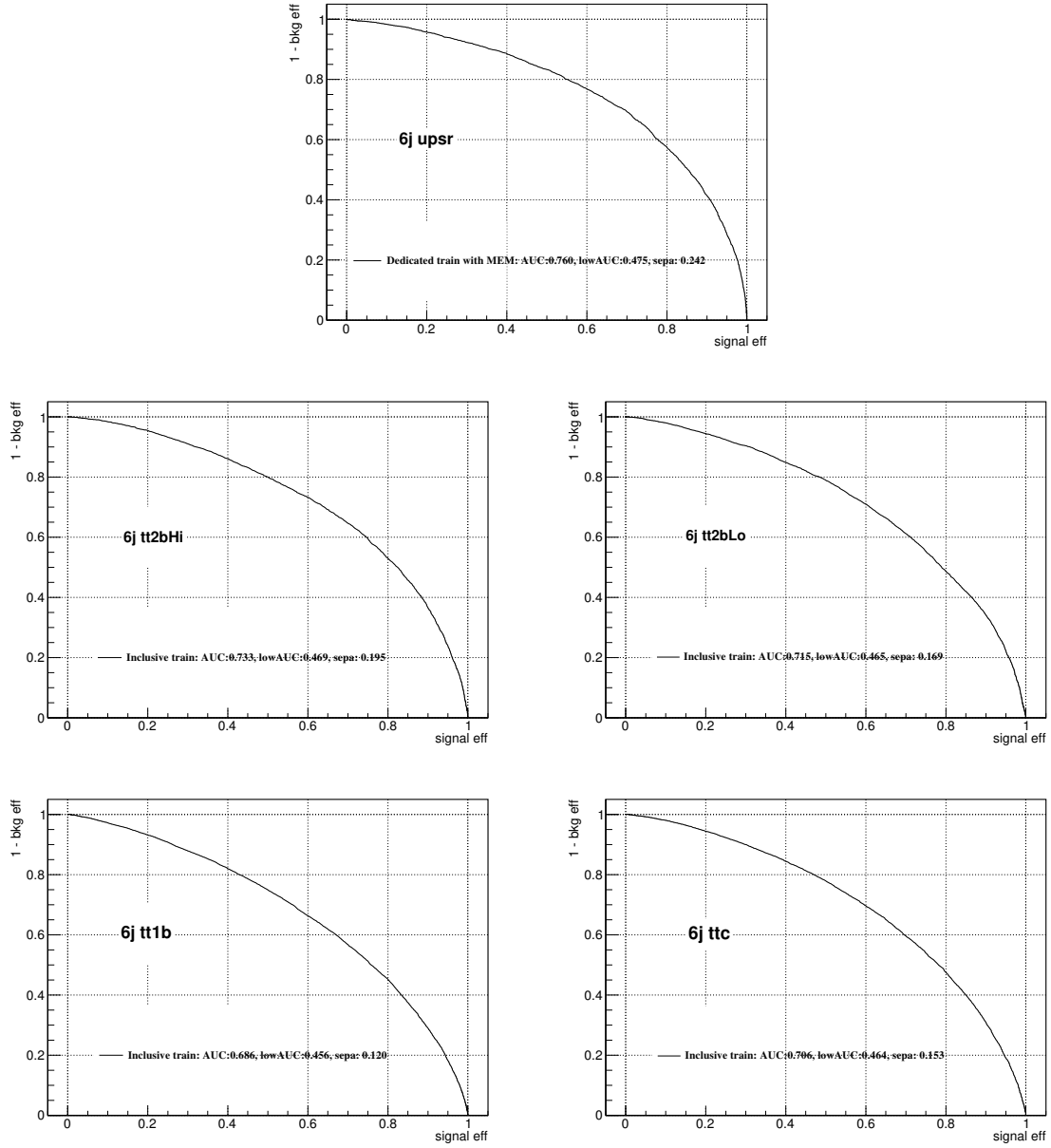


Figure 4.30: Performance of classification BDT in the various categories within the $(\geq 6j, \geq 4b)$ region. Top: UPSR (with MEM); middle left: $t\bar{t} + \geq 2b$ Hi SR; middle right: $t\bar{t} + \geq 2b$ Lo SR; bottom left: $t\bar{t} + 1b$ CR; bottom right: $t\bar{t} + \geq 1c$ CR.

total which gives comparable performance to the BDT using baseline input list. The results of such optimization in the UPSR are used directly for the inclusive BDT in other regions. In these regions, the baseline performance can be recovered picking up to 11 variables (plus four b -tagging variables as used in the baseline inclusive training). Similarly in the 5j scenario, the input variables in the UPSR regions can be reduced to 11, and these 11 variables plus the 4 b -tagging weights can recover the performance of the inclusive BDT in other regions. The reduced input list for both $\geq 6j$ and 5j events are shown in Table 4.7.

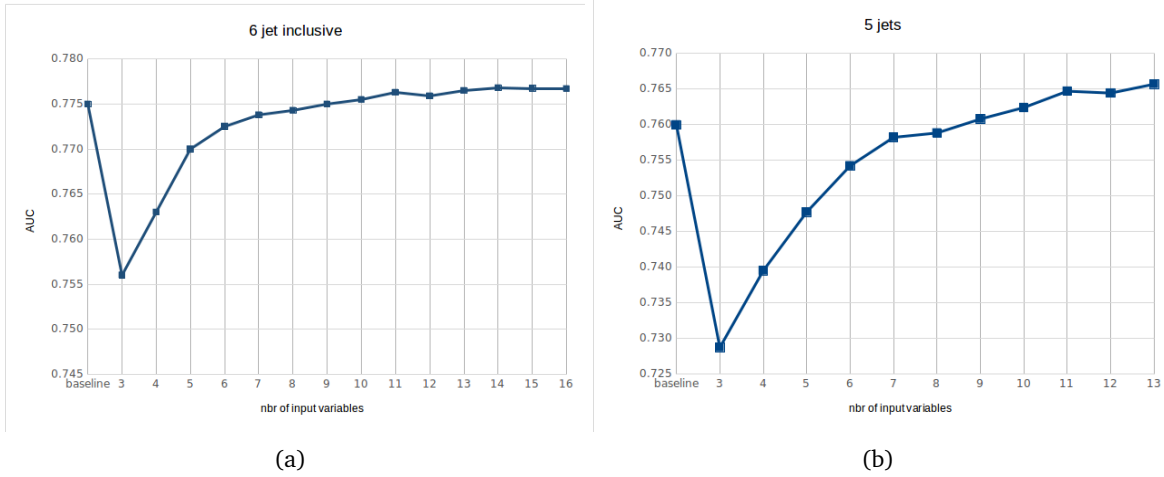


Figure 4.31: AUC of classification BDT with an increasing number of input variables in the (a) ($\geq 6j$, $\geq 4b$) and (b) ($5j$, $\geq 4b$) UPSR regions. The “baseline” is the result using the input variables shown in Table 4.5.

Using these optimized variables studied in the UPSR, the inclusive BDT are then retrained for 5jets and ≥ 6 jets events separately, with this reduced input list. The performance of the BDT trained with the final version of the simulation samples used in the analysis are shown in Table 4.8 where the results of using the baseline and reduced input lists are shown. For some categories, the AUC performance is marginally worse than the baseline BDT.

The next section describes the statistic model of fitting the data to MC simulations, where the classification BDT studied in this section will be used as fitting discriminant for signal regions. Whether or not to use the reduced-list BDT will be finally decided by fit results.

Variable	Definition	$\geq 6j$		5j	
		UPSR	Incl.	UPSR	Incl.
General kinematic variables					
$\Delta R_{bb}^{\text{avg}}$	Average ΔR for all b -tagged jet pairs	✓	✓	✓	✓
$\Delta R_{bb}^{\text{max } p_{\text{T}}}$	ΔR between the two b -tagged jets with the largest vector sum p_{T}	✓	✓	–	–
$\Delta R_{\ell,bb}^{\text{min}}$	Minimum ΔR between the lepton and any pair of b -tagged jets	–	–	✓	✓
$\Delta \eta_{jj}^{\text{max}}$	Maximum $\Delta \eta$ between any two jets	✓	✓	✓	✓
$m_{bb}^{\text{min } \Delta R}$	Mass of the combination of the two b -tagged jets with the smallest ΔR	✓	✓	✓	✓
$m_{jj}^{\text{min } \Delta R}$	Mass of the combination of any two jets with the smallest ΔR	–	–	✓	✓
$p_{\text{T}}^{\text{jet5}}$	p_{T} of the fifth leading jet (testing full removal)	–	✓	–	–
N_{30}^{Higgs}	Number of b -jet pairs with invariant mass within 30 GeV of the Higgs boson mass	–	–	✓	✓
Aplanarity	$1.5\lambda_2$, where λ_2 is the second eigenvalue of the momentum tensor [125] built with all jets	–	✓	✓	✓
Aplanarity $_{b\text{-jet}}$	$1.5\lambda_2$, where λ_2 is the second eigenvalue of the momentum tensor built with b -jets	✓	✓	–	–
$H1$	Second Fox–Wolfram moment computed using all jets and the lepton	–	–	✓	✓
Variables from reconstruction BDT output					
BDT	BDT output	✓*	✓*	✓*	✓*
m_H	Higgs boson mass	–	–	✓	✓
$\Delta R_{\text{lep top,had top}}$	ΔR between leptonic and hadronic tops	–	✓*	–	–
$\Delta R_{H,\text{had top}}$	ΔR between Higgs boson and hadronic top	–	✓	–	–
Variable from Likelihood calculation					
D	Likelihood discriminant	✓	✓	✓	✓
Variable from Matrix Element calculation					
MEM_D1	MEM discriminant	✓	–	–	–
Variables from b -tagging					
w_b^H	Sum of binned b -tagging weights of jets from best Higgs candidate	–	✓	–	✓
B_{j3}	3 rd jet binned b -tagging weight (sorted by weight)	–	✓	–	✓
B_{j4}	4 th jet binned b -tagging weight (sorted by weight)	–	✓	–	✓
B_{j5}	5 th jet binned b -tagging weight (sorted by weight)	–	✓	–	✓

Table 4.7: Optimized input variables to the classification BDT in the single-lepton channel, for the dedicated training in the UPSR and for the inclusive training in all other regions. For variables from the reconstruction BDT, those with a * are from the BDT using Higgs boson information, while those with no * are from the BDT without Higgs boson information.

Category	AUC (reduced/baseline)	Category	AUC (reduced/baseline)
UPSR (SR ₁)	0.760/0.760	UPSR (SR ₁ ^{5j})	0.760/0.762
tt2bHi (SR ₂ ^{≥6j})	0.734/0.733	tt2b (SR ₂ ^{5j})	0.738/0.745
tt2bLo (SR ₃ ^{≥6j})	0.714/0.715	tt1b CR (SR ₃)	0.714/0.713
ttc CR	0.703/0.706	ttc CR	0.716/0.733

(a)

(b)

Table 4.8: The classification BDT performance using the reduced input list, compared to the results using the baseline list, for categories in (a) ($\geq 6j$, $\geq 4b$) and (b) ($5j$, $\geq 4b$) events.

4.6 Systematic uncertainties

The $t\bar{t}H$ ($H \rightarrow b\bar{b}$) signal sensitivity is affected by many sources of *systematic uncertainties* in addition to the statistical uncertainty. The former is more dominant than the latter according to Reference [25]. On the experimental side, the sources include the uncertainties in the luminosity of the collider, the lepton and jets reconstruction and identification, the missing transverse energy and b -tagging. On the other hand, large theoretical uncertainties can also drive the signal sensitivity if the signal and background processes cannot be simulated precisely. Different uncertainties may affect only the overall normalization of the samples or also the distribution used to categorize events and to build the final discriminant.

The corresponding *systematic samples* are produced individually by each time varying a single source by one standard deviation. A nuisance parameter (NP) is then assigned to each source, and enters the fit model as described in Section 4.7.1. Some of the systematic uncertainties, in particular most of the experimental ones are decomposed into several independent sources which are then assigned with individual NPs. Across all event categories of signal and background samples, these NPs are shared in the fit model.

4.6.1 $t\bar{t}$ + jets uncertainties

The $t\bar{t}$ + jets, as the dominant background, especially the heavy flavor component, is difficult to model and predict, and thus has a large theoretical uncertainty. The $t\bar{t}$ components including $t\bar{t} + \geq 1b$, $t\bar{t} + \geq 1c$, and $t\bar{t}$ + light, are affected by different types of uncertainties. $t\bar{t}$ + light has additional diagrams and profits from relatively precise measurements in data. However, the $t\bar{t}$ + HF background modeling is challenging with large theoretical uncertainty but weak constraints from data measurement. Thus independent uncertainty sources of $t\bar{t} + \geq 1b$, $t\bar{t} + \geq 1c$, and $t\bar{t}$ + light modeling are considered and assigned with independent NPs, except that an uncertainty of $\pm 6\%$ is assumed for the inclusive $t\bar{t}$ NNLO+NNLL cross-section. All the sources of systematic uncertainty for $t\bar{t}$ + jets are summarized in Table 4.9, and will be described in the following.

Due to the differences between the nominal sample (using the POWHEG generator and PYTHIA 8 for parton shower and hadronization, therefore referred to as PP8) and other MC simulations, three independent $t\bar{t}$ + jets shape uncertainties are considered uncorrelated for $t\bar{t}$ components. Firstly, the uncertainty related with the choice of $t\bar{t}$ inclusive NLO generator is derived by comparing the prediction from PP8 with the prediction by SHERPA version 2.1.1 interfaced with OPENLOOPS. Both of the parton shower and hadronization model are varied with this SHERPA prediction which is over-conservative indeed. Secondly, the uncertainties related with the choice of parton shower and hadronization model are also derived. The alternative sample is generated by POWHEG with the same settings as the nominal sample, but interfaced with HERWIG 7, a different parton shower and hadronization model than the nominal sample. Finally, the uncertainty in the modeling of initial- and final-state radiation (ISR / FSR) is assessed with two alternative PP8 samples [126]. In one sample, the amount of radiation is increased by varying the renormalization and factorization scales, the A14 set, and h_{damp} parameter which controls the transverse momentum of the first additional gluon emission beyond the $t\bar{t}$ system. In the other sample, the radiation amount is decreased in a similar way.

From the Run 1 analysis, the $t\bar{t} + \geq 1b$ and $t\bar{t} + \geq 1c$ normalizations are found to be underestimated in MC simulation by around 40%. In this round, the individual normalizations of the two components are left free to float and will be determined by the fit. To avoid double

counting on the uncertainty of these $t\bar{t} + \text{jets}$ component fractions, in the above mentioned systematic samples, these fractions are corrected to match the PP8 sample.

4.6.1.1 Uncertainties in the $t\bar{t} + \geq 1b$ modeling

As the overwhelming component in signal-enriched regions, the $t\bar{t} + \geq 1b$ background has a large impact on the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) signal extraction but with large modeling uncertainty. A finer classification on this background is then implemented in order to assign correction factors and estimate uncertainties: events with exactly two single- b -jets (as defined in Section 4.3.2) are labeled as $t\bar{t} + b\bar{b}$, those with only one single- b -jet are labeled as $t\bar{t} + b$, and those with only one B -jet (see Section 4.3.2) are labeled as $t\bar{t} + B$, the rest of the $t\bar{t} + \geq 1b$ events being labeled as $t\bar{t} + \geq 3b$. Events with additional b -jets entirely originating from multi-parton interactions (MPI) or b -jets from final-state radiation (FSR), i.e. originating from gluon radiation from the top-quark decay products, are considered separately in the $t\bar{t} + b$ (MPI/FSR) subcategory.

To model $t\bar{t} + \geq 1b$ with the highest available precision, the contributions of the above defined different b -jet topologies in the PP8 sample are scaled to match those predicted by SHERPA+OPENLOOPS [53, 127], except $t\bar{t} + b\bar{b}$ (MPI/FSR) which is not present in the SHERPA sample. This SHERPA sample uses the CT10 four-flavor (4F) scheme PDF set to generate the $t\bar{t} + b\bar{b}$ sample, and gives NLO accurate description of the additional b -jets while the PP8 simulation can only produce additional b -jets from the parton shower. Figure 4.32 shows the predicted fractions for each of the $t\bar{t} + \geq 1b$ subcategories, with the PP8 inclusive $t\bar{t}$ sample

Table 4.9: Summary of the sources of systematic uncertainty for $t\bar{t} + \text{jets}$ modeling. The systematic uncertainties listed in the second section of the table are induced by the inclusive $t\bar{t}$ simulation. The systematic uncertainties listed in the third section of the table affect only the fractions of the various $t\bar{t} + \geq 1b$ subcategories. The last column of the table indicates the $t\bar{t}$ category to which a systematic uncertainty is assigned. In the case where all three categories ($t\bar{t} + \text{light}$, $t\bar{t} + \geq 1c$ and $t\bar{t} + \geq 1b$) are involved (marked with ‘all’), the last column also specifies whether the uncertainty is considered as correlated (“corr.” in the table) or uncorrelated (“uncorr.” in the table) across them.

Systematic source	Description	$t\bar{t}$ categories
$t\bar{t}$ cross-section	Up or down by 6%	All, corr.
$k(t\bar{t} + \geq 1c)$	Free-floating $t\bar{t} + \geq 1c$ normalization	$t\bar{t} + \geq 1c$
$k(t\bar{t} + \geq 1b)$	Free-floating $t\bar{t} + \geq 1b$ normalization	$t\bar{t} + \geq 1b$
SHERPA5F vs. nominal	Related to the choice of NLO event generator	All, uncorr.
PS & hadronization	POWHEG+HERWIG 7 vs. POWHEG+PYTHIA 8	All, uncorr.
ISR / FSR	Variations of μ_R , μ_F , h_{damp} and A14 Var3c parameters	All, uncorr.
$t\bar{t} + \geq 1c$ ME vs. inclusive	MG5_aMC@NLO+HERWIG++: ME prediction (3F) vs. incl. (5F)	$t\bar{t} + \geq 1c$
$t\bar{t} + \geq 1b$ SHERPA4F vs. nominal	Comparison of $t\bar{t} + b\bar{b}$ NLO (4F) vs. POWHEG+PYTHIA 8 (5F)	$t\bar{t} + \geq 1b$
$t\bar{t} + \geq 1b$ renorm. scale	Up or down by a factor of two	$t\bar{t} + \geq 1b$
$t\bar{t} + \geq 1b$ resumm. scale	Vary μ_Q from $H_T/2$ to μ_{CMMPs}	$t\bar{t} + \geq 1b$
$t\bar{t} + \geq 1b$ global scales	Set μ_Q , μ_R , and μ_F to μ_{CMMPs}	$t\bar{t} + \geq 1b$
$t\bar{t} + \geq 1b$ shower recoil scheme	Alternative model scheme	$t\bar{t} + \geq 1b$
$t\bar{t} + \geq 1b$ PDF (MSTW)	MSTW vs. CT10	$t\bar{t} + \geq 1b$
$t\bar{t} + \geq 1b$ PDF (NNPDF)	NNPDF vs. CT10	$t\bar{t} + \geq 1b$
$t\bar{t} + \geq 1b$ UE	Alternative set of tuned parameters for the underlying event	$t\bar{t} + \geq 1b$
$t\bar{t} + \geq 3b$ normalization	Up or down by 50%	$t\bar{t} + \geq 1b$
$t\bar{t} + \geq 1b$ MPI	Up or down by 50%	$t\bar{t} + \geq 1b$

compared to the $t\bar{t} + b\bar{b}$ SHERPA4F sample.

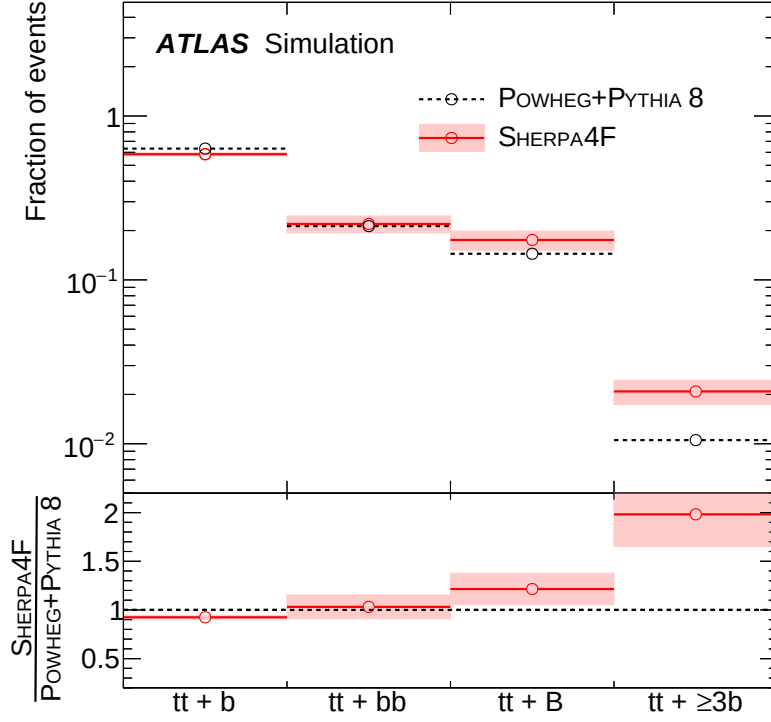


Figure 4.32: The relative predicted fractions of the $t\bar{t}+b$, $t\bar{t}+b\bar{b}$, $t\bar{t}+B$ and $t\bar{t}+\geq 3b$ subcategories before any event selection. The prediction from the inclusive POWHEG+PYTHIA 8 sample is compared to the four-flavor $t\bar{t}b\bar{b}$ calculation from SHERPA4F, with its uncertainties from a combination of the sources shown as the shaded area. The fractions are normalized to the sum of the four contributions shown here, without considering the $t\bar{t} + b$ (MPI/FSR) subcategory as part of the total [25].

The uncertainties in sub-component fractions of the SHERPA 4F simulation, the error band in Figure 4.32 are assessed by comparing with alternative SHERPA 4F samples which are shown in the third section of Table 4.9. And these uncertainties will propagate to the scaled PP8 sub-component fractions. The “ $t\bar{t} + \geq 1b$ MPI” does not exist in the SHERPA4F and thus does not enter this MC-MC scale. More details could be found in Reference [25].

For the inclusive $t\bar{t} + \geq 1b$ process, the shape difference between the prediction from PP8 5F and SHERPA4F is considered as one additional source of uncertainty after correcting the PP8 sample to match the $t\bar{t} + \geq 1b$ subcategory fractions predicted in SHERPA. And the two samples are normalised to the same cross-section.

4.6.1.2 Uncertainties in the $t\bar{t} + \geq 1c$ modeling

In the signal-enriched regions, the $t\bar{t} + \geq 1c$ component is the subleading but much less dominant background compared to the $t\bar{t} + \geq 1b$. For the $t\bar{t} + \geq 1c$ simulation, similarly to $t\bar{t} + \geq 1b$, in addition to the $t\bar{t} + \text{jets}$ inclusive PP8 prediction, an NLO prediction using the 3F scheme is

produced with additional c -jets described by $t\bar{t} + c\bar{c}$ ME calculation. And an uncertainty is assigned to the difference between these two predictions.

4.6.2 Non- $t\bar{t}$ uncertainties

All the non- $t\bar{t}$ uncertainties are listed in Table 4.10, including both the $t\bar{t}H$ signal and other background simulation and the experimental uncertainties.

Table 4.10: List of systematic uncertainties included in the analysis. An ‘N’ means that the uncertainty is taken as normalization-only for all processes and channels affected, whereas ‘SN’ means that the uncertainty is taken on both shape and normalization. Some of the systematic uncertainties are split into several components for a more accurate treatment referred to as “Comp” in the table. The last row refers to the cross-sections for the background processes tZ , tWZ , $t\bar{t}WW$, $tHjb$ and WtH [25].

Systematic uncertainty	Type	Comp.
<i>Experimental uncertainties</i>		
Luminosity	N	1
Pileup modeling	SN	1
Physics Objects		
Electron	SN	6
Muon	SN	15
Taus	SN	3
Jet energy scale	SN	20
Jet energy resolution	SN	2
Jet vertex tagger	SN	1
E_T^{miss}	SN	3
b-tagging		
Efficiency	SN	30
Mis-tag rate (c)	SN	15
Mis-tag rate (light)	SN	80
Mis-tag rate (extrapolation $c \rightarrow \tau$)	SN	1
<i>Signal and non-$t\bar{t}$ background modeling</i>		
Signal		
$t\bar{t}H$ cross-section	N	2
H branching fractions	N	3
$t\bar{t}H$ modeling	SN	1
Other Backgrounds		
$t\bar{t}W$ cross-section	N	2
$t\bar{t}Z$ cross-section	N	2
$t\bar{t}W$ modeling	SN	1
$t\bar{t}Z$ modeling	SN	1
Single top cross-section	N	3
Single top modeling	SN	5
W + jets normalization	N	3
Z + jets normalization	N	3
Diboson normalization	N	1
Fakes and non-prompt normalization	N	7
$t\bar{t}t\bar{t}$ cross-section	N	1
Small background cross-sections	N	9

The uncertainties on the non- $t\bar{t}$ background modeling have marginal impact on the signal sensitivity, but the $t\bar{t}H$ modeling uncertainty has a non-negligible effect.

4.6.2.1 $t\bar{t}H$ modeling uncertainties

The $t\bar{t}H$ cross-section prediction uncertainty is $^{+5.8\%}_{-9.2\%}(\text{scale}) \pm 3.6\%(\text{PDF})$. The two terms, the first one accounting for the QCD scale uncertainty and the second the PDF + α_s [22, 96–100], are treated uncorrelated. Uncertainties in the Higgs boson branching fractions are also considered where the one for the $H \rightarrow b\bar{b}$ process is 2.2% [22]. An uncertainty associated with the choice of parton shower and hadronization model is also introduced by comparing the nominal simulation from MG5_aMC@NLO+PYTHIA 8 to the alternative one from MG5_aMC@NLO interfaced to HERWIG++.

4.6.2.2 Non- $t\bar{t}$ background

The W + jets cross section is assumed with an uncertainty of 40%. An additional 30% normalization uncertainty for W + HF jets is taken into account as uncorrelated between events with 2 and > 2 HF jets. For the Z + jets, a normalization uncertainty of 35% is introduced, uncorrelated across jet bins, to account for variations of the SHERPA parameters. An uncertainty on the correlation factor for the HF components is also considered, which is derived from data.

An uncertainty of $^{+5\%}_{-4\%}$ is considered for cross section in each of the three single-top production modes [110–112]. For the Wt and t -channel production modes, modeling uncertainties associated with the choice of parton shower and hadronization model, and with initial- and final-state radiation are taken into account by comparing to alternative samples, as for the $t\bar{t}$ process. The uncertainty in the assessment of the interference between Wt and t -channel production at NLO [128] are also taken into account in a comparison with alternative samples produced using a different scheme.

The uncertainty of the $t\bar{t}V$ NLO cross-section is 15%, split into PDF and scale uncertainties as for $t\bar{t}H$ [129]. The modeling uncertainty, related to the choice of event generator, parton shower and hadronization model, is also introduced by comparing to alternatively generated samples. These uncertainties are treated uncorrelated between $t\bar{t}Z$ and $t\bar{t}W$. For the diboson and $t\bar{t}t\bar{t}$ background, a 50% normalization uncertainty is assumed individually. Each of the small backgrounds from tZ , $t\bar{t}WW$, $tHj\bar{b}$ and WtH is assigned with two uncorrelated cross section uncertainties accounting for PDF and scale uncertainties, while tWZ is assigned with one uncertainty accounting for both.

The estimated yield of non-prompt lepton events is assigned with an uncertainty of 50%, considered uncorrelated across the e and μ channel, 5 jets and ≥ 6 jets regions, and the resolved and boosted regions.

4.6.2.3 Experimental uncertainties

The uncertainty of the combined 2015+2016 integrated luminosity is 2.1% derived following a methodology similar to that detailed in Ref. [130]. It is applied to the normalization of all processes of MC simulations.

In total 20 independent sources of uncertainty are assigned to the jet energy scale including the jet flavor, pileup corrections, η dependence and high- p_T jets [72]. Although the uncertainties are not large, totaling 1%–6% per jet (depending on the jet p_T), the effects are amplified by the large number of jets in the final state. The uncertainty in JVT is also considered, as well as in the jet energy resolution.

The efficiency to correctly tag b -jets, and mis-tag rate for the c -jets and light jets are measured in MC and then corrected to data [73]. The b -tagging discriminant from MC is calibrated combining these measured efficiencies and rates extracted from each of the four WP used in the analysis as a function of jet kinematics. An error matrix is assigned to this calibration procedure, and contains uncertainties that correctly describe correlations across multiple working points. A diagonalization procedure of the matrix gives uncertainties that can be considered uncorrelated. The uncertainties associated with b -tagging are factorized into 30 independent sources, and the ones associated with the mis-tag rate for the c -jets (light-jets) are factorized into 15 (80) independent sources.

Uncertainties assigned to leptons arise from the trigger, reconstruction, identification and isolation efficiencies, as well as the lepton momentum scale and resolution. Uncertainties of these measurements account for a total of 24 independent uncertainty sources but have only a small impact on the result.

All uncertainties in energy scales or resolutions are propagated to the missing transverse momentum. Additional uncertainties in the scale and resolution of the soft term are considered, for a total of three additional sources of systematic uncertainty for the missing transverse momentum.

4.7 Statistical analysis

The distributions of the discriminants from each of the analysis categories are combined in a profile likelihood fit to test for the presence of the signal. The NP and normalization factors are incorporated into the fit template providing the degrees of freedom so that the fit will simultaneously determine the normalization and constrain the distribution of the most important background components.

4.7.1 Profile likelihood fit

The statistical test of the signal presence is defined as the ratio of the profile likelihood functions. The general form of a likelihood function is $\mathcal{L}(\theta) = \prod_{i=1}^n f(x_i; \theta)$ where $x_1, \dots, x_n$ are the measurement of random variable x repeated n times, and $f(x; \theta)$ is the probability density function of x known a priori except θ which denotes a set of parameters with unknown values. In this analysis, the unknown parameters include the signal strength μ , the background composition normalization factors k , and all of the NPs referred to as θ .

In this analysis, a binned likelihood function is constructed as a product of Poisson-probability terms in each bin i of the discriminant distribution, for each event category j . Data are expected to follow a Poisson probability in each bin. MC simulation follows the SM prediction and is used as expectation of data:

$$\mathcal{L}_{\text{main}}(\mu, k, \theta) = \prod_{j \in \text{cats}} \prod_{i \in \text{bins}} \frac{\left(N_{i,j}^{\text{exp}}(\mu, k, \theta) \right)^{N_{i,j}^{\text{data}}}}{N_{i,j}^{\text{data}}!} \cdot e^{-N_{i,j}^{\text{exp}}(\mu, k, \theta)} \quad (4.7)$$

In each bin, $N_{i,j}^{\text{data}}$ denotes the number of events observed in data, while $N_{i,j}^{\text{exp}}(\mu, k, \theta)$ is the number of MC simulation events:

$$N_{i,j}^{\text{exp}}(\mu, k, \theta) = \mu \cdot N_{i,j,\text{sig}}^{\text{exp}}(\theta_{\text{sig}}^1, \dots, \theta_{\text{sig}}^{n_{\text{sig}}}) + \sum_{b \in \text{bkg}} k_b \cdot N_{i,j,b}^{\text{exp}}(\theta_b^1, \dots, \theta_b^{n_b}) \quad (4.8)$$

where n_{sig} (n_b) and θ_{sig}^i (θ_b^i) denote respectively the number of NPs and the NP for signal (background component b). And k_b is the normalization factor of background component b .

Each of the NPs $\theta_{\text{sig}}^1, \dots, \theta_{\text{sig}}^{n_{\text{sig}}}$, and $\theta_b^1, \dots, \theta_b^{n_b}$ is discrete. $\theta = 0$ corresponds to the nominal samples, while ± 1 corresponds to the systematic samples generated as mentioned in Section 4.6. The continuous NPs are defined by extrapolation for $|\theta| > 1$ and interpolation for $|\theta| < 1$. A priori the NPs are treated as following a Gaussian distribution function and are multiplied to the main likelihood function:

$$\mathcal{L}(\mu, \mathbf{k}, \boldsymbol{\theta}) = \mathcal{L}_{\text{main}}(\mu, \mathbf{k}, \boldsymbol{\theta}) \cdot \prod_{u=1}^M \frac{1}{\sqrt{2\pi}} e^{-\frac{\theta_u^2}{2}} \quad (4.9)$$

where M is the total number of NPs.

The best estimate for the parameters μ , $\mathbf{k}$ and $\boldsymbol{\theta}$ are obtained by maximizing the likelihood function. More statistical inference on μ is made using a test statistic t_μ , described as the profile likelihood ratio: $t_\mu = -2 \ln(\mathcal{L}(\mu, \hat{\boldsymbol{\theta}}_\mu) / \mathcal{L}(\hat{\mu}, \hat{\boldsymbol{\theta}}))$, where $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}$ are the values of the parameters which maximize the likelihood function, and $\hat{\boldsymbol{\theta}}_\mu$ are the values of the nuisance parameters which maximize the likelihood function for a given value of μ . Based on t_μ , the upper limit of μ is inferred using the CL_s method [131, 132].

4.7.2 Fitted discriminant variables

The input variable for fitting varies in different event categories. The total event yield is used in the control regions, with the exception of $\text{CR}_{t\bar{t}+\geq 1c}^{5j}$ and $\text{CR}_{t\bar{t}+\geq 1c}^{\geq 6j}$, where the distribution of the scalar sum of jets' p_T , H_T^{had} , is used. Figure 4.33 shows the H_T^{had} distributions in the $t\bar{t} + \geq 1c$ enriched control regions.

In signal regions, to further separate the signal from the background, the output of the classification BDT, as described in Section 4.5, is used as the discriminant. The BDT distribution is rebinned with an automatic algorithm aiming to avoid bins with large statistical errors while still keeping a good separation between the $t\bar{t}H$ signal and the backgrounds. Figure 4.34 shows the distributions of the classification BDTs in all of the signal regions before the fit.

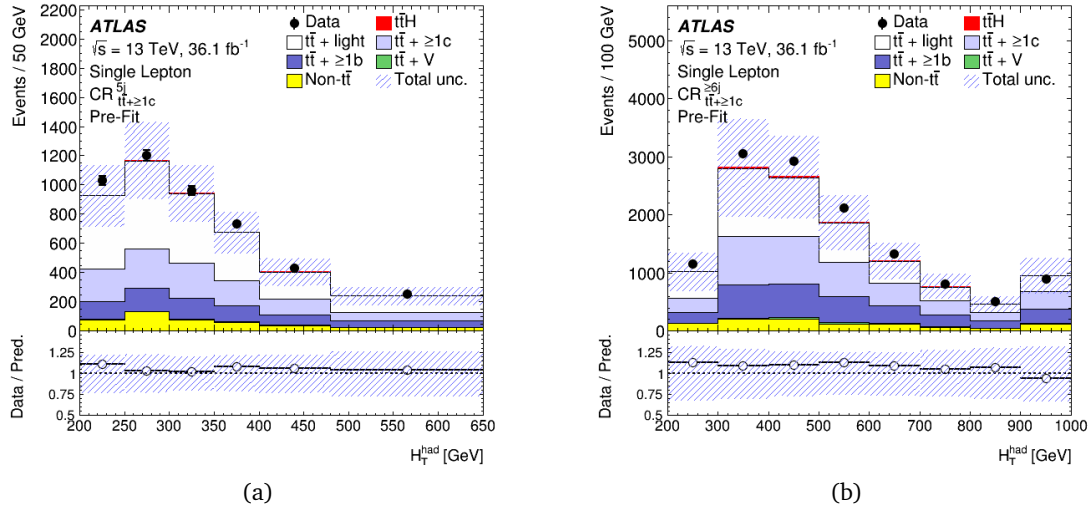
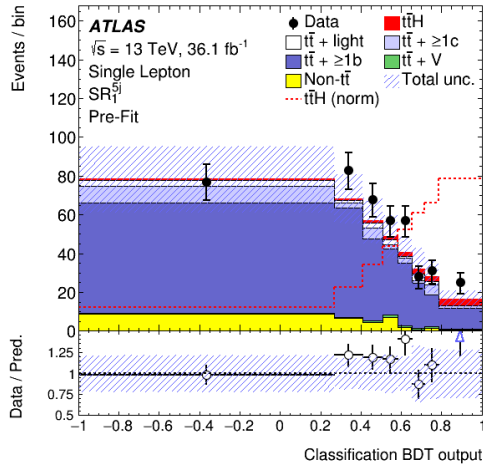
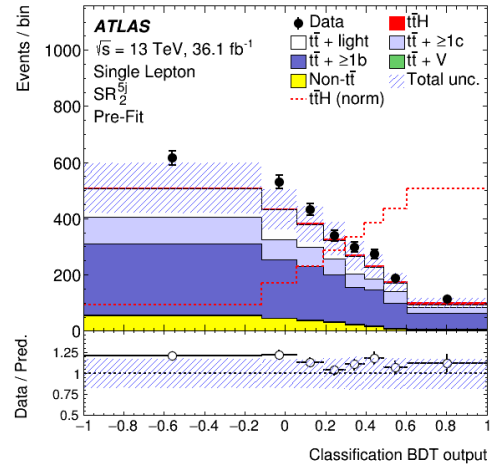


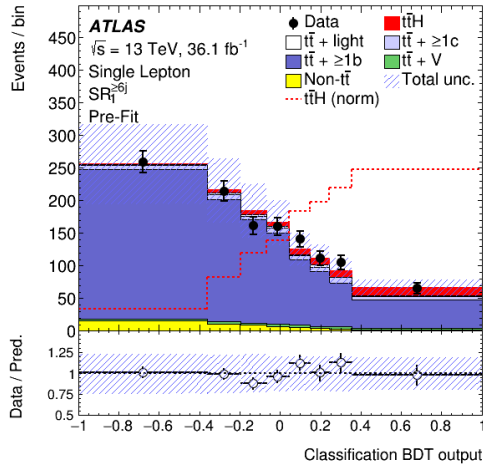
Figure 4.33: Pre-fit comparison between data and prediction for the H_T^{had} distributions in the single-lepton $t\bar{t} + \geq 1c$ -enriched control regions for (a) 5jets and (b) ≥ 6 jets events. The $t\bar{t}H$ signal prediction is shown stacked at the top of the background prediction, normalized to the SM cross-section. The $t\bar{t} + \geq 1b$ and $t\bar{t} + \geq 1c$ fraction are from PP8 nominal samples, and no uncertainty to them are taken into account.



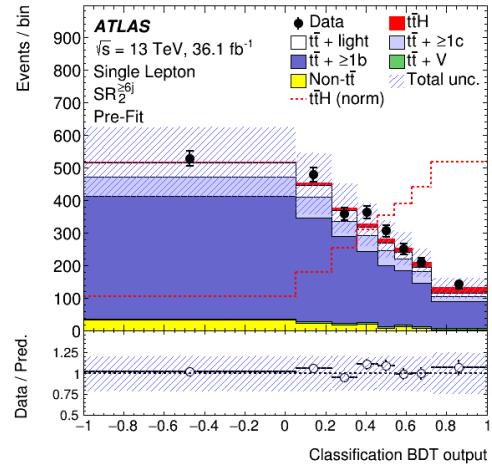
(a)



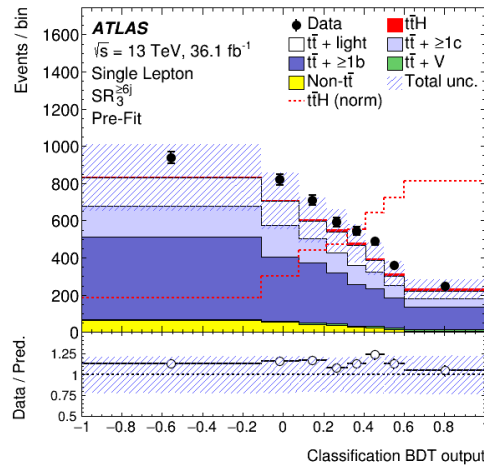
(b)



(c)



(d)



(e)

Figure 4.34: Pre-fit comparison between data and prediction for the BDT discriminant in the single-lepton channel of the 5jet ((a) and (b)) and ≥ 6 jets ((c), (d), and (e)) signal regions. The $t\bar{t}H$ signal yield is normalized to the SM cross-section (shown in solid red) and to the total background prediction (shown in dashed lines). The uncertainty for the $t\bar{t} + \geq 1b$ or $t\bar{t} + \geq 1c$ normalization is not included.

4.8 Results

The MC systematic models need to be validated before fitting to the observed data. This is studied based on the so-called Asimov data [133] which is built from the predicted distribution assuming Poisson error in each bin. As an estimation of the signal sensitivity and background constraining, many optimization and decisions on the analysis strategy are also based on the Asimov fit, including the choice of the two BDT options: the baseline and the reduced BDT as shown in section 4.8.1.

To avoid statistical fluctuations in the systematic model, *averaging* and *pruning* procedures are applied to the fit discriminants before fitting. During the averaging process, the systematic *symmetrization* is firstly applied to make the $\pm 1\sigma$ systematic variations symmetric. The systematic uncertainties induced by comparing two samples from different MC simulations has only the variation in one direction. Symmetric operation of this variation is implemented around the nominal distribution to obtain the variation in the other direction. The second step in the averaging procedure is the *smoothing* which averages systematic uncertainties across bins of the discriminant in each region. It aims to remove statistical fluctuations in the systematic model, particularly for uncertainties introduced by comparison between MC simulations with limited statistics. Then the pruning procedure removes small systematic components that have no impact on the result, in order to make the fit robust and speed it up. The final result when fitting the model to the observed data is presented in Section 4.8.2.

4.8.1 BDT classifier choice

Compared with the default BDT using the baseline list of features as presented in Table 4.5, the reduced BDT with the slimmed input list shown in Table 4.7 gives only marginally worse performance in the non-UPSR regions. The expected signal strength from the Asimov fit for both options are shown in Table 4.11. It turned out that the reduced BDT brings a deterioration in sensitivity with larger systematic uncertainties, although the AUC difference is tiny. Therefore the baseline BDT is adopted for final fit to data.

However, the sensitivity deterioration using the reduced-list BDT is driven by statistical fluctuation. Figure 4.35 shows the leading systematic uncertainty in this analysis induced by the systematic sample using NLO SHERPA4F generator instead of the PP8 5F scheme, before the fit for both BDT options. The systematic variation between the two options are similar before the smoothing, while more difference happen to be introduced after applying the smoothing which is a procedure to remove statistical fluctuation. Therefore the sensitivity decrease may not be genuine.

Configuration	best fit μ	tot	stat
BDT using full list	1.00	+0.59 -0.56	+0.33 -0.33
BDT using reduced list	1.00	+0.70 -0.66	+0.34 -0.33

Table 4.11: The Asimov fits with the S+B hypothesis based on the two BDT choices.

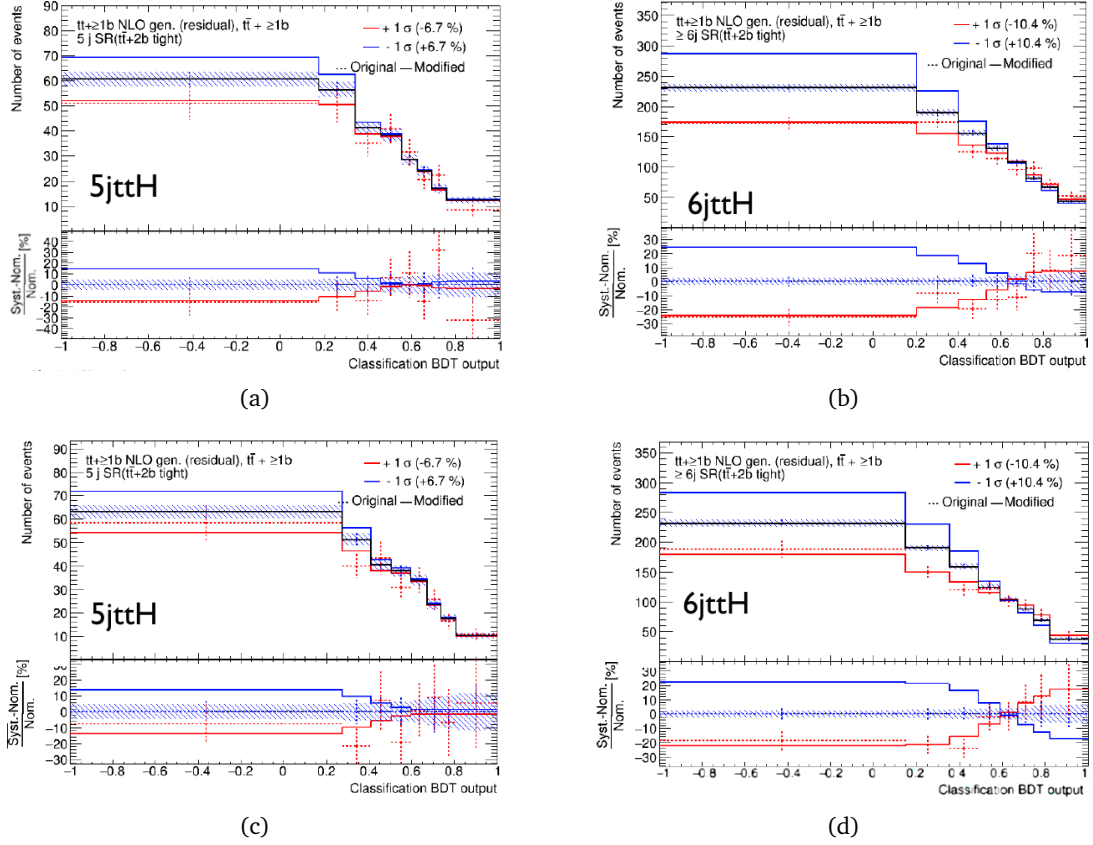


Figure 4.35: The MC BDT distribution variations induced by using a different NLO generator, the SHERPA4F with the PP8 5F scheme, for the $t\bar{t} + \geq 1b$ component. (a) (full BDT) and (c) (reduced BDT) are for the SR_2^{5j} , while (b) (full BDT) and (d) (reduced BDT) are for the $SR_2^{\geq 6j}$. The dashed points are before smoothing, while the solid lines are after smoothing.

4.8.2 Fit results

When fitting the single-lepton data alone, the observed signal strength is $0.67^{+0.71}_{-0.69}$, compatible with the standard model expectation within one standard deviation.

The fit result shown in this section is instead the combined fit, where one signal strength is determined by the simultaneous fit in all signal and control regions in both single-lepton and dilepton channels. The best-fit μ value is $0.84^{+0.64}_{-0.61}$ as shown in Figure 4.36, consistent with the SM in 1σ . An alternative fit is also performed where the dilepton and single-lepton channels are assigned two independent signal strengths and all the nuisance parameters are kept correlated across both channels as well. The dilepton channel is complementary to the single lepton one as it has orthogonal events, but suffers from low statistics. The corresponding fitted values of μ are $-0.24^{+1.02}_{-1.05}$ in the dilepton channel and $0.95^{+0.65}_{-0.62}$ in the single-lepton channel, as shown in Figure 4.36. Compared to the fitting in single-lepton channel along, the μ in the single-lepton channel is increased while the uncertainty decreased, due to the background correlation with the dilepton channel.

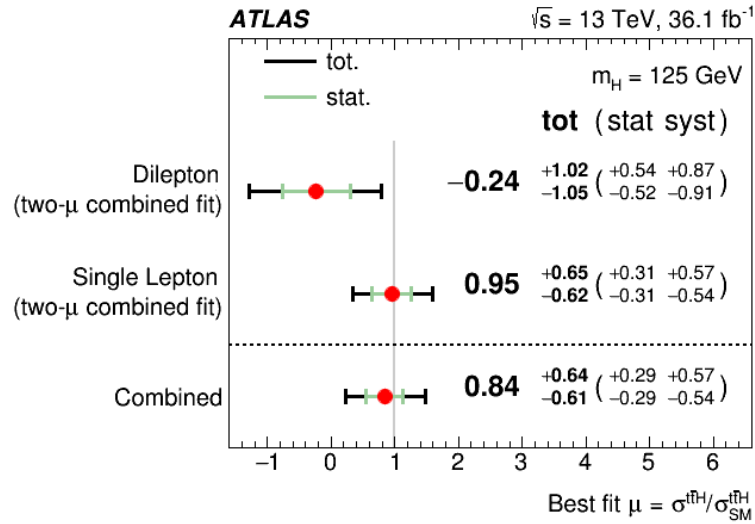
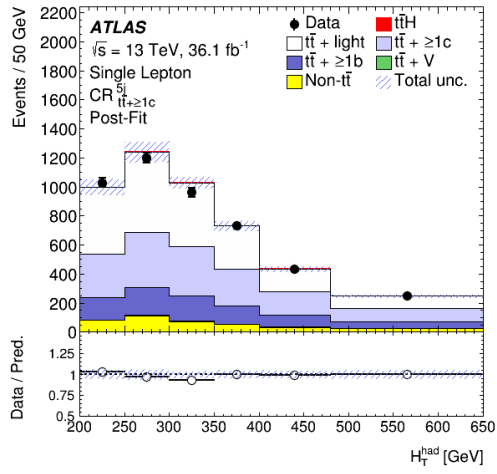
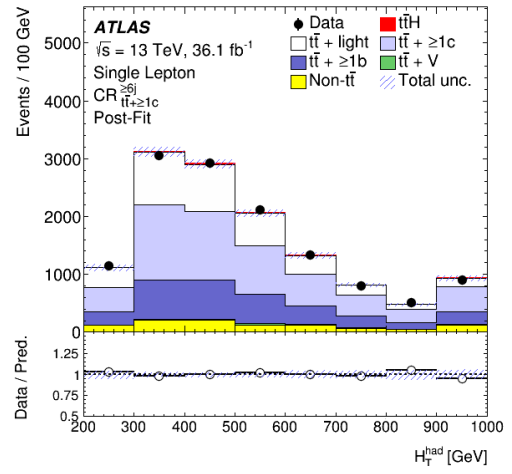


Figure 4.36: Summary of the signal strength measurements in the individual channels and for the combination. The numbers are obtained in a combined fit [25].

Simultaneously the background normalization and constraining are also achieved in the fit. In particular, the best-fit values of $k(t\bar{t} + \geq 1b)$ and $k(t\bar{t} + \geq 1c)$ are 1.24 ± 0.10 and 1.63 ± 0.23 respectively. The normalization factors of 1 correspond to the pre-fit $t\bar{t} + \geq 1b$ and $t\bar{t} + \geq 1c$ fractions which are the prediction of the POWHEG+PYTHIA 8 nominal sample for total $t\bar{t}$ production. This means that the $t\bar{t} + \text{HF}$ fractions are under-estimated by the MC predictions, as also observed by CMS [91, 134] and in previous ATLAS analyses [90, 92] as well. Figure 4.37 and Figure 4.38 show the H_T^{Had} distributions in the $t\bar{t} + \geq 1c$ -enriched control regions and the classification BDT in the signal regions, after the fit to data. Compared with the corresponding pre-fit Figures 4.33 and 4.34, the post-fit distributions show improved consistency between the MC prediction and the observed data as well as reduced uncertainties.

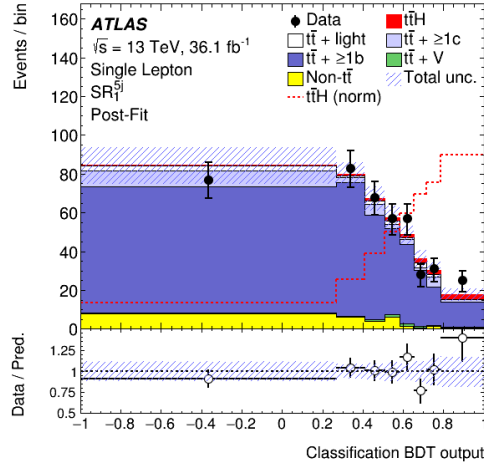


(a)

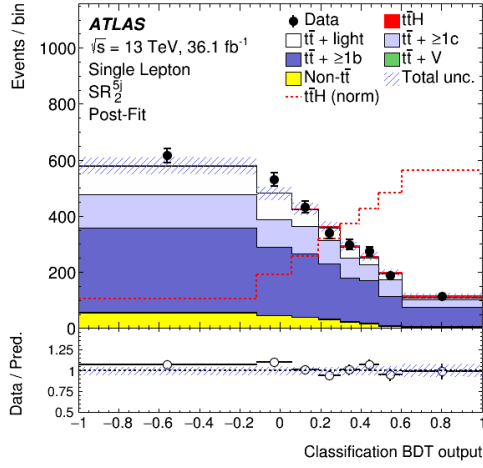


(b)

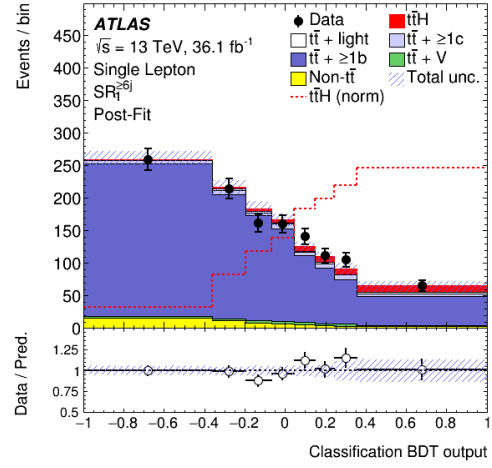
Figure 4.37: Post-fit comparison between data and prediction for the H_T^{had} distributions in the single-lepton $tt + \geq 1c$ -enriched control regions for (a) 5jets and (b) ≥ 6 jets events. The ttH signal prediction is shown stacked at the top of the background prediction, normalized to the fitted μ [25].



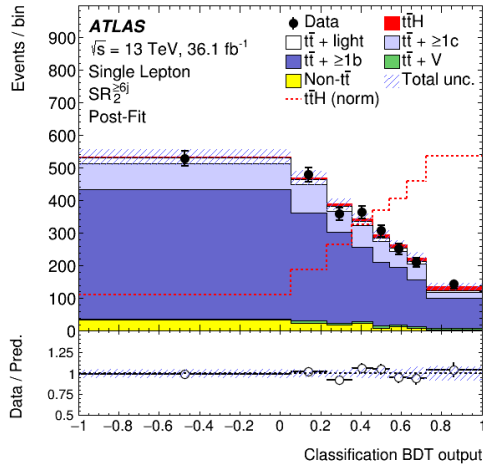
(a)



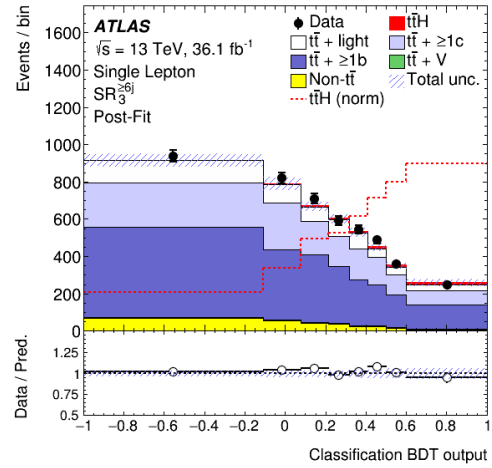
(b)



(c)



(d)



(e)

Figure 4.38: Post-fit comparison between data and prediction for the BDT discriminant in the single-lepton channel of the 5jet ((a) and (b)) and ≥ 6 jets ((c), (d), and (e)) signal regions. The $t\bar{t}H$ signal yield is normalized to the fitted μ (shown in solid red) and to the total background prediction (shown in dashed lines) [25].

The systematic uncertainty dominates the final error on μ . The most dominant source is the modeling uncertainties of the $t\bar{t} + \geq 1b$ background. The low statistics in the alternative simulated samples is the sub-leading factor. Other important sources are the uncertainties in the b -tagging efficiency and the jet energy scale and resolution, which is not surprising due to the large number of jets and b -jets in the final state. The $t\bar{t}H$ process modeling is also an important source.

The 20 nuisance parameters which describe the systematic uncertainties with the largest contribution to the total uncertainty are shown in Figure 4.39, where the NP related to MC statistical uncertainties are not included. For each NP, the best-fit value, and the impact on μ of both prefit and postfit are presented. The NP having the largest impact on the signal strength corresponds to the comparison between the SHERPA5F and the nominal prediction for the $t\bar{t} + \geq 1b$ process, related to the choice of the NLO event generator for this background component. The three sub-leading uncertainties also relate to the modeling of the $t\bar{t} + \geq 1b$ background. After the fit, the uncertainties corresponding to most of these NPs are reduced. The best-fit value of some NPs are shifted in the fit from their nominal values, which is found to correct mainly the predictions of the $t\bar{t}$ background to the observed data in various regions.

An excess of the $t\bar{t}H$ events over the expected SM background is found with an observed (expected) significance of 1.4 (1.6) standard deviation under the background only hypothesis without being able to claim $t\bar{t}H$ observation. Figure 4.40 displays the observed and expected upper limit on the signal strength where μ larger than 2.0 is excluded at the 95% confidence level combining two channels.

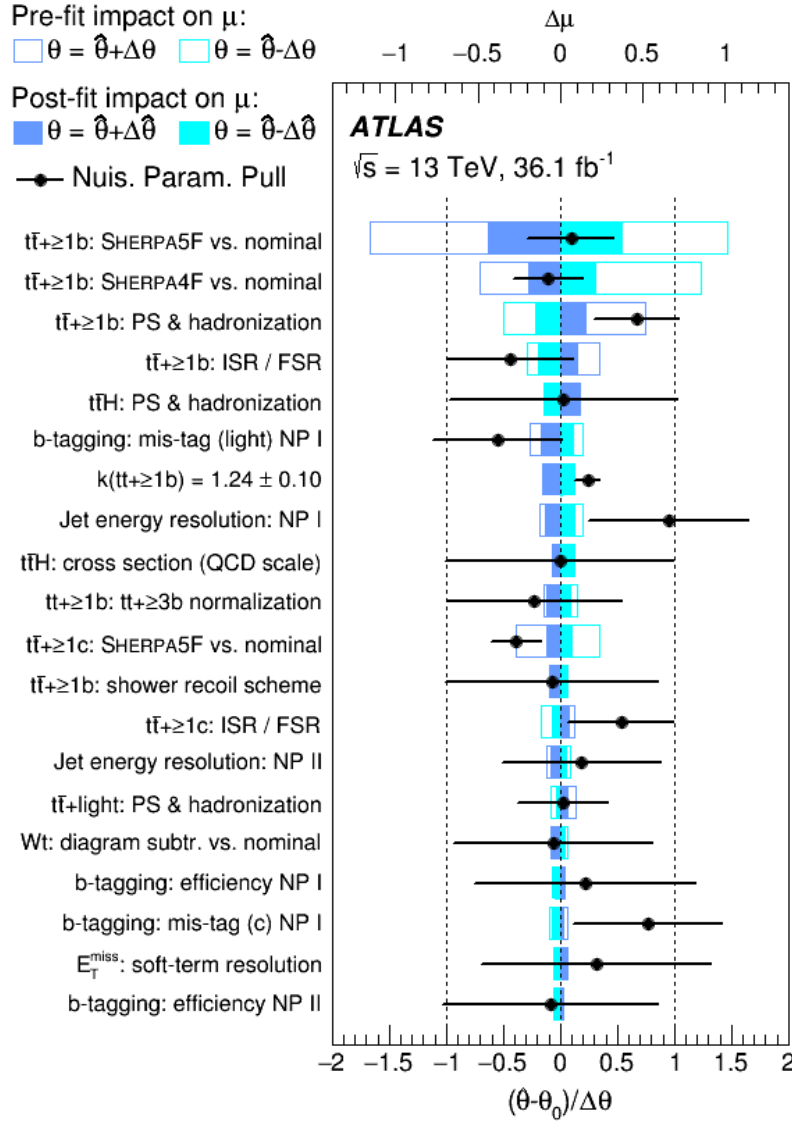


Figure 4.39: Ranking of the nuisance parameters included in the fit according to their impact on the measured signal strength μ . Only the 20 most highly ranked parameters are shown. The empty blue rectangles correspond to the pre-fit impact on μ and the filled blue ones to the post-fit impact on μ , both referring to the upper scale. The impact of each nuisance parameter, $\Delta\mu$, is computed by comparing the nominal best-fit value of μ with the result of the fit when fixing the considered nuisance parameter to its best-fit value, $\hat{\theta}$, shifted by its pre-fit (post-fit) uncertainties $\pm\Delta\theta$ ($\pm\Delta\hat{\theta}$). The black points show the pulls of the nuisance parameters relative to their nominal values, θ_0 . These pulls and their relative post-fit errors, $\Delta\hat{\theta}/\Delta\theta$, refer to the lower scale. The parameter $k(t\bar{t} + \geq 1b)$ refers to the floating normalization of the $t\bar{t} + \geq 1b$ background, for which the pre-fit impact on μ is not defined, and for which both θ_0 and $\Delta\theta$ are set to 1. For experimental uncertainties that are decomposed into several independent sources, NP I and NP II correspond to the first and second nuisance parameters, ordered by their impact on μ , respectively.

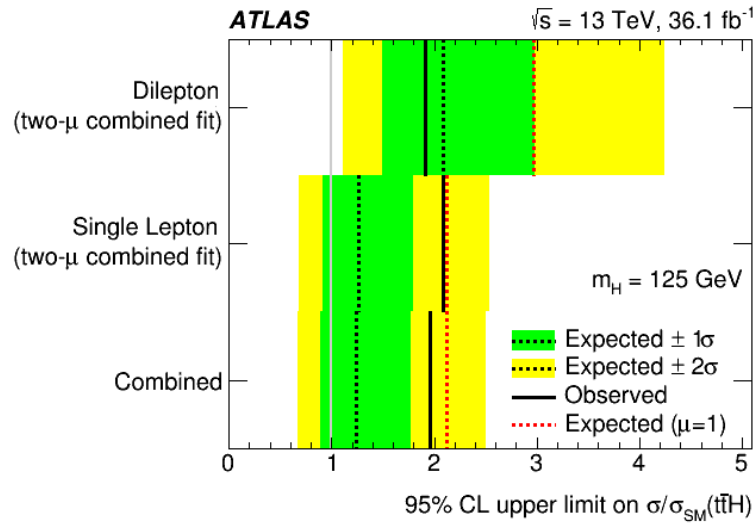


Figure 4.40: Summary of the 95% confidence level (CL) upper limits on $\sigma(t\bar{t}H)$ relative to the SM prediction in the individual channels and for the combination. The observed limits are shown, together with the expected limits both in the background-only hypothesis (dotted black lines) and in the SM hypothesis (dotted red lines). In the case of the expected limits in the background-only hypothesis, one- and two-standard-deviation uncertainty bands are also shown. All these numbers are obtained based on events in both single- and dilepton channels [25].

4.9 Summary

Searching for the Higgs boson produced in association with a top quark pair with the Higgs boson decaying into a bottom quark pair, $t\bar{t}H$ ($H \rightarrow b\bar{b}$), is presented. The analysis uses the data sets collected with the ATLAS detector in 2015 and 2016, corresponding to an integrated luminosity of 36.1 fb^{-1} of pp collisions at $\sqrt{s} = 13 \text{ TeV}$. The results are published in Reference [25], where the author contributes to the single-lepton channel for the MVA study.

The events are categorized into signal rich and background control regions, with requirement on the pseudo-continuous b -tagging of the leading four jets sorted by b -tagging score.

To separate the $t\bar{t}H$ signal and the $t\bar{t}$ background, two-step MVAs are studied where the author developed and optimized the reconstruction and classification BDT. The reconstruction BDT is trained to find the correct matching between jets and partons. A strategy of using pseudo-continuous b -tagging weights is firstly exploited on events having loose b -tagging requirement. This allows to have more truth matching as the signal of the reconstruction BDT, while the number of improper matching as the background is limited. Studies are also carried on to fully optimize the BDT and the state-of-the-art result based on this new method brings up to 5% improvement for the efficiency of correctly reconstructing the Higgs boson.

The classification BDT is trained to separate the $t\bar{t}H$ signal from the $t\bar{t} + \text{jets}$ background, using variables based on the best combination sorted by the reconstruction BDT score. The outputs of the intermediate MVAs, the reconstruction BDT, the likelihood discriminant, and the matrix element method (in the ultra pure signal rich region of ≥ 6 jets only), are dominant discriminants. By using b -tagging weights, the analysis strategy can be simplified by using one classification BDT trained inclusively across event categories (for ≥ 6 and 5 jets separately), because it gives as good performance as dedicated training in each category. The BDT trained with a reduced input kinematics list is proposed as a candidate model by the author. It gives comparable AUC performance to the baseline model, and small deterioration in the Asimov fit due to statistical fluctuations. Studies are also carried out and result in an optimal BDT that shows better performance than the one in ICHEP analysis.

The classification BDT output is used as the final discriminant for signal rich regions in the fit. Combined with the dilepton channel, the simultaneous fit across all regions gives a signal strength of $\mu = 0.84^{+0.64}_{-0.61}$ for a Higgs boson mass of 125 GeV. The observed (expected) signal significance is 1.4 (1.6) standard deviation, without being able to claim $t\bar{t}H$ observation. A signal strength greater than 2.0 is excluded at 95% confidence level.

Deep learning methods in the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis 5

Up to now the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis shows no excess of the Higgs boson signal in this channel. The studies in this chapter, for the first time, explore the usage of cutting edge deep learning techniques, on various aspects of the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis. In Section 5.1, the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) physics analysis is recapped in more common language for readers with non-particle physics background. Then basics of deep neural networks are illustrated in Section 5.2. The following sections present the author's work and contributions:

- Section 5.3: the analysis in the last publication [25] uses two-step traditional machine learning models to separate the $t\bar{t}H$ signal and $t\bar{t}$ background, where three different intermediate MVAs are exploited to solve the signal reconstruction. This section uses RNN to optimize the signal and background binary classification in one model, combining the information used by various models before.
- Section 5.4: to improve the $t\bar{t}H$ signal significance, the analysis adapted the strategy to split the whole event space into multiple categories, each being enriched in $t\bar{t}H$ or one of a $t\bar{t}$ sub-component. The categorization was implemented manually previously, and this section explores deep neural networks to do multi-classification.
- Section 5.5: so far all the discriminating features used for machine learning models are engineered by physicists. This section explores the usage of raw input features and tries to reach better $t\bar{t}H$ against $t\bar{t}$ binary classification performance by considering prior knowledge on the data:
 - Investigating the power of DNN to learn high level features.
 - Embedding the collision events via a parse tree that has topology in analogy with the $t\bar{t}H$ process, for classification. The model is expected to learn from low-level features more efficiently.
 - Using two embedding trees for classification, having topologies analogous with $t\bar{t}H$ and $t\bar{t}$ processes, respectively. Two approaches of exploiting two trees are tested: direction concatenation, and Siamese training.
- Section 5.6: exploring adversarial training as a domain adaptation technique, to account for differences in the distribution of two $t\bar{t}$ background data sets: nominal and systematic samples.

5.1 Particle physics analysis for non-physicists

In this section, ingredients of particle physics experiment and specifically the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis are illustrated in common language. The important concepts can also be found in [Glossary](#).

5.1.1 Experiments and data

In the LHC, proton bunches are accelerated and then directed to cross in the ATLAS detector. In each bunch crossing, an average of 25 proton pairs interact and lead to the production of the Higgs boson or many other particles. Only at most one of these proton-proton collisions may be of interest in each bunch crossing. The other interacting pairs, called [pile-up](#), will add noise on the interesting one. Each bunch crossing is called an [event](#), corresponding to a sample in a dataset in the machine learning (ML) field, while in this document a [sample](#) refers to a set of events. The produced particles from a bunch crossing could be stable as they are produced, or [decay](#) to other particles. When a quark is produced, it will immediately evolve into a cluster of particles called [jet](#) (see Section 1.1.5), except for top quark, the heaviest one, that will decay into other particles first. Jets evolving from b -quarks are called b -jets, which can be distinguished with special properties (called [b-tagging](#)) from c -jets and *light jets* evolving from c -quarks (lighter version of b -quark) or light quarks (even lighter). All surviving particles and jets are collectively called [final state](#) of the event and are detected. Each event is endowed with an index, called [event number](#).

ATLAS measures three properties of the final state: the type of each particle (electron, photon, jet, etc.), its energy, and the 3D direction of the particle (momentum in three directions $\vec{p}$ or equivalently η , ϕ , p_T , see Section 2.2.1). These are called [raw features](#) or [low-level features](#) of events. Variables that are functions of these features are called [high-level features](#), and they are calculated with physics expertise. For example, the angle between the directions of two particles is a high-level feature which tends to have small value if the two particles are decay products of another particle moving at high speed. Events with measured variables are called [data](#).

5.1.2 Signal and background

Events with interesting physics processes to be precisely measured or with processes to be searched for are called [signal events](#). This corresponds to the positive samples in ML field. In this study, the signal is the so-called $t\bar{t}H$ ($H \rightarrow b\bar{b}$) process, where the Higgs boson is produced together with two top quarks (called $t\bar{t}H$) and then is annihilated immediately and produces two b -quarks, called the Higgs *decay* into b -quarks, or simply $H \rightarrow b\bar{b}$ (see Section 1.2.1, and Figures 4.1(a) and 4.1(b)). However, the signal signature seen by the detector is different, as the top quarks decay as well. In this analysis, the signal final state is a *lepton* (the general term for the electron and similar particles), a *neutrino* (a neutral particle rarely interacting with matter) and one b -jet from one top quark, two b -jets from the Higgs boson, and one b -jet and two light jets from another top quark. Ideally there are in total six jets expected to be seen by the detector, four of which are identified as b -jets, but it is not possible to directly know whether they are from a top quark or the Higgs boson, nor from which quark they evolve from.

The b -jets are identified using a machine learning model, and its output score is called [b-tagging weight](#) which represents the probability of being a b -jets. Each threshold on it gives a b -tagging criterion (called a [b-tagging working point](#), or b -tagging WP): 85% WP is a loose

criterion, meaning the efficiency to correctly identify a jet from a b -quark as a b -jet is 85%, and 60% WP is a very tight criterion. The b -jets identification also comes with a false-positive rate for a jet not issued from a b -quark but identified as a b -jet. This mis-tag rate becomes small with decreasing tagging efficiency: at the 85% (60%) WP, a c -jet is mistaken for a b -jet once for every 3.1 (34) jets, while a light jet is mis-identified as a b -jet once in every 33 (1538) jets (see Section 2.4.4). The b -tagging weights are exploited in this study as a binned distribution due to the calibration technique between the data and simulation. This is called *pseudo-continuous (PC) b -tagging*. Five bins are defined by cutting on the continuous b -tagging weight, which corresponds to four different WPs, 85%, 77%, 70%, and 60%.

However, the Higgs boson has small *cross-section* (a measure of the probability for that process to occur during any proton–proton collision, proportional to event counting) as theoretical expectation. A large majority of events are processes that were discovered and studied in previous generations of experiments, but have very similar final state to the signal. This is called *backgrounds*, corresponding to the negative samples in ML. The main background in the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis is the $t\bar{t} + \text{jets}$ process (see Figure 4.1(c)), where additional jets are produced together with top-quark pair without producing a Higgs boson. The jets can evolve from light-, c - or b -quarks, and therefore the background can be further split into sub-components, denoted as $t\bar{t} + \text{light}$, $t\bar{t} + \geq 1c$, and $t\bar{t} + \geq 1b$, respectively. Both $t\bar{t}H$ ($H \rightarrow b\bar{b}$) and $t\bar{t} + \text{jets}$ have a top-quark pair and two b -jets produced: the two b -jets are from the Higgs in the signal, while the two b -jets in the background can be from $t\bar{t} + \geq 1b$, or $t\bar{t} + \geq 1c$ and $t\bar{t} + \text{light}$ since the c - or light-jets can sometimes be mis-identified as a b -jet by the b -tagging process. In data, it is not possible to know whether a particular event is signal or background, all measurements are always statistical.

5.1.3 Simulated data

Most of ATLAS physics analyses, including the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) channel, use a model dependent strategy, i.e., aim to find signal excess in experimental data in comparison with the simulated data based on theoretical prediction.

The real data production, detection, and acquisition system is simulated using the Monte Carlo (MC) method. The so-produced simulated data is usually called *predictions* or *Monte Carlo (MC) simulation*. The event *class*, signal or background, is labeled in the simulated data. The known background processes are deemed to be correctly described by the standard model (SM) theory (see Section 1.1). The SM theory also gives predictions to the signal process, although the signal existence or property requires further studies from experiments. All the machine learning models are trained on MC simulations.

In the simulation, as many events as needed can be simulated given enough computational resources (although it is not always the case due to limited computing resources). The signal or background sample with large number of simulated events is said to have large *statistics*. To match to the prior production probability of the physics process, *sample weights* w are assigned for each event. The weights are also normalized such that the sum of the weights $\sum_{i \in \text{signal}} w_i$ and $\sum_{i \in \text{background}} w_i$ are the expected total number of signal and background events, respectively, during the time interval of data collection.

Apart from the *statistical uncertainty*, the *systematic uncertainties* are given in MC simulations, providing estimation of unknowns about what we know. For instance, due to the stochastic nature of the interaction between the particle and the detector material, the measured jet energy may be dispersed from the central value, ending up with limited resolution. One of the systematic uncertainties that drives the signal significance is from the imprecise modeling

of the $t\bar{t} + \geq 1b$ process. It is estimated by comparing the *nominal sample* (as the central value) to the *systematic sample* which is simulated using different frameworks and approaches. Currently it is difficult to well model the $t\bar{t} + \text{jets}$ background, and the systematic samples gives different feature shapes from the nominal one. Therefore, when comparing to the background simulation, the very small number of $t\bar{t}H$ ($H \rightarrow b\bar{b}$) signal will be easily overwhelmed by the large prediction uncertainty, deteriorating its significance. It is thus mandatory to optimize the signal to background ratio (thanks to ML techniques in particular) to maximize the analysis sensitivity.

5.1.4 The default $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis pipeline

The search for the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) process exploits statistical tests to find a significant excess of data compared to the background only prediction. The *significance* is a measure of how unusual the data are if the background only hypothesis is true. It is expressed in terms of σ , or standard deviation¹. The higher the number of σ , the more incompatible the data are with the background only hypothesis. 5σ is the requirement to claim a discovery, and it means one chance out of 2 million that the observed signal is from a background fluctuation. Another test is to find the signal amount characterized by the *signal strength* μ , which is the ratio of the observed number of events over the one from theoretical prediction. These statistical studies where the model is adjusted to the observed data are usually called *fit*.

Due to the low signal over background ratio, the events are firstly filtered in order to discard uninteresting backgrounds. One of the jets from $t\bar{t}H$ ($H \rightarrow b\bar{b}$) may not be seen by the detector, while more than one extra jet could be found due to pile-up in the collision process. Therefore, events are selected if they have ≥ 6 jets among which ≥ 4 are tagged as b -jets, at the working point of 85%. This is referred to as *6ji4bi@85%*. And events are also selected if they have exactly 5 jets, referred to as *5je4bi@85%*.

For each of these jet multiplicities, events are split into categories in feature space, referred to as *regions*. These regions are tuned and formed with careful examination by people, in order to have signal enriched regions, called *signal regions (SRs)*, with more signal events. The region with the purest signal events is called *ultra pure signal region (UPSR)*. The other regions are formed in order to have enhanced purity of $t\bar{t} + \geq 1b$, $t\bar{t} + \geq 1c$, and $t\bar{t} + \text{light}$ sub-components individually in each of them, and are called *control regions (CRs)*. The fit machinery is implemented simultaneously in these regions, where SRs contribute to the signal significance while the CRs provide constraints on the modeled background sub-components and therefore reduce the systematic uncertainties originating from them.

In SRs, to further enhance the signal significance, machine learning methods are widely used, referred to as *multi-variate analysis (MVA)*. To further separate $t\bar{t}H$ from the $t\bar{t}$ background, a two-stage MVA is used in the analysis documented in Reference [25] and detailed in Chapter 4:

- The signal signature in the detector is multiple jets and b -jets coming from the Higgs boson or top quarks. To build features of the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) process, jets need to be correctly associated to where they are from: the goal is to find the correct matching between jets and quarks. This is called *event reconstruction*, and is also referred to as *jet-parton association*². Three MVAs are adopted to solve this problem, boosted decision

¹The significance can be estimated as $S/\sqrt{B}$ under the condition that $S \ll B$, $B \gg 1$, and the systematic uncertainties are neglected, where $S(B)$ is the sum of weights of simulated signal (background) selected as signal (background). This quantity and the signal over background ratio S/B are usually used to optimize the signal region definitions.

²Parton is a general term for quarks and gluons (another elementary particle that can produce jets in the detector).

trees (BDT) (called *reconstruction BDT* as it is trained to solve the reconstruction problem and to be distinguished from another BDT), likelihood, and matrix element method. Details can be seen in Section 4.5.3.

- Once the signal is reconstructed, it is separated from the $t\bar{t}$ background with another BDT, called *classification BDT*, using features based on the reconstruction and other event properties. Details can be found in Section 4.5.4.

Classification BDT distributions in SRs and the number of events in CRs are used as inputs to the final fit machinery.

5.2 Introduction to deep neural networks

Deep learning is a sub-field of Machine Learning and it typically refers to deep neural networks, which have more hidden layers than the shallow neural networks, and their training algorithms. The networks with deeper layers and thus larger complexity make it possible to learn beneficial representations automatically of input data for various tasks like classification or detection. For example, in image processing where the inputs are the raw image pixels, hierarchical features are learned through the network. Starting from the object edge or segmentation of different motifs acquired by first layers, more abstract features like shapes or familiar parts of objects, and then combinations of parts are extracted by deeper and deeper layers. In conventional machine learning approaches, amounts of work related with the domain expertise are required to engineer the input features that can be used by the model like a classifier. As discussed in Section 4.5.4, the BDT performance largely depends on the discriminating power of the input list of features which are carefully chosen and optimized. With feature abstraction ability, the deep models could possibly get the most useful features automatically. They benefit from large data size.

The learned representations on one task can be adapted to different tasks which could then reach higher performance in shorter period of training. In natural language processing (NLP), a better word representation as continuous numeric vectors is first learned (embedding) by pre-training a deep model on a huge corpus. This universal embedding can capture the context and semantics features of the training text, and bring improvements in the downstream NLP tasks problems on other text data [135] like sentiment classification.

Although a large amount of breakthroughs are emerging in recent years, we are still confronted with many challenges. Firstly, the deep neural networks have many parameters to be optimized. Performing a complete search in the parameter space is not always possible, and even a limited search for a large number of parameters ends up with many combinatorics consuming many computing resources. Although deep models do work well in practice, the basic understanding of their nature is far from satisfying, and we do not know well how the networks make their decisions. For example, the DNN can be easily fooled [136] by adversarial examples which have just tiny modification of the original ones for training. Some networks that are trained with images are found to identify objects based on their texture instead of shapes. The interpretability of deep learning is becoming more and more of a concern.

5.2.1 Deep neural network components

Artificial *neural networks* (NN) are a machine learning model inspired by the way information is processed in biological neural networks. They are made of layers of connected units called

neurons. The simplest neural network model is the shallow neural network as shown in Figure 5.1(a). The information is fed into the input layer mapped to the output layer through one hidden layer in the form of:

$$\begin{aligned}
 \mathbf{z}^{(1)} &= \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)} \\
 \mathbf{h} &= f^{(1)}(\mathbf{z}^{(1)}) \\
 \mathbf{z}^{(2)} &= \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)} \\
 \mathbf{p} &= f^{(2)}(\mathbf{z}^{(2)})
 \end{aligned} \tag{5.1}$$

where $\mathbf{W}^{(1)}$ and $\mathbf{W}^{(2)}$ are encoding weight matrices of the network, each element of $\mathbf{W}$ being associated to one connection arrow. $\mathbf{b}^{(1)}$ and $\mathbf{b}^{(2)}$ are the corresponding bias vectors. This transforms the input $\mathbf{x}$ into $\mathbf{z}$ linearly. For the hidden layer, represented by vector $\mathbf{h}$, non-linear transforms are added into the mapping via an *activation function* f . From the input layer, the calculation proceeds towards the output, giving prediction $\mathbf{p}$. This is called a *feed-forward* neural network.

Through the so-called *training* procedure on a set of events, referred to as *training sample*, the model can extract patterns of the sample by adjusting the connection weights and then is used for various tasks, for example, classification. Such a trained model is expected to be applied and give good performance on another set of events, referred to as *test sample*.

The basic feed-forward deep neural networks (DNN, referred to as the classical neural networks in the following) have a similar structure except with more than two hidden layers as shown in Figure 5.1(b). Going from this, various types of deep neural networks are widely used nowadays, for instance Convolutional Neural Networks (CNN) [137], Recurrent Neural Networks (RNN) [138], Auto-Encoders (AE) [139], recursive neural networks [140], Generative Adversarial Networks (GAN) [141], etc. Many recent breakthroughs in various fields are achieved with models exploiting one or a combination of these deep neural network structures.

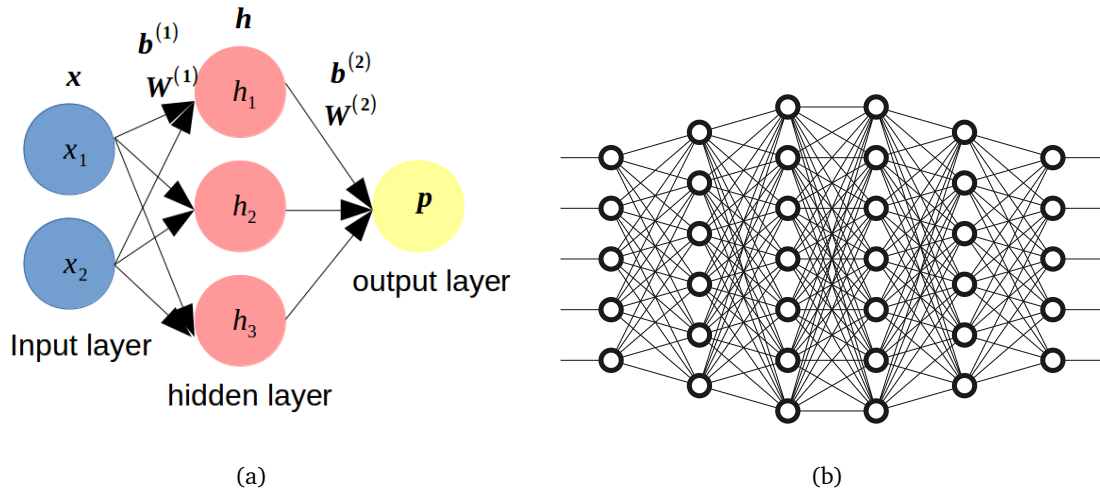


Figure 5.1: The feed-forward (a) neural network with one hidden layer and (b) deep neural network with four hidden layers.

Training and testing samples The cross training (similar to Section 4.5.3.3) is also exploited as shown in the schematic Figure 5.2. Events with only even (odd) event number form fold 0 (fold 1). In model 0 (model 1), fold 0 (fold 1) are used to learn the model, and are referred to as learning sample. Events in fold 1 (fold 0) are held out for performance testing. Among the learning samples in each model, 20% are sampled as validation sample and the rest are used for training (referred to as training sample). The validation set provides the trained model an unbiased evaluation based on which the model's hyper-parameters are tuned.

The two models provided by such a two-fold training are used when applying the NN models in the full $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis, while only model 0 is implemented when studying, optimizing or comparing different NN models.

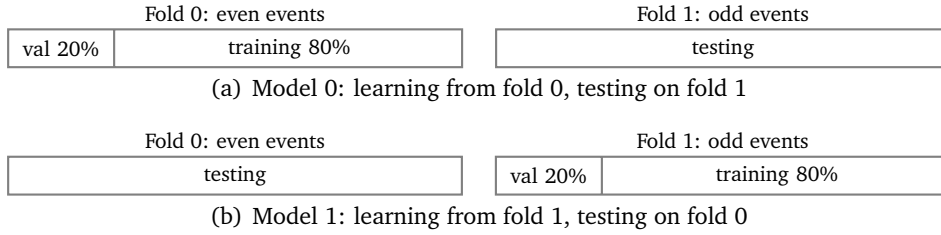


Figure 5.2: Two-folded cross training. In model 0 (1) as shown in (a) ((b)), even (odd) events are used for learning the model which is trained on 80% of events and validated on 20%. The odd (even) events are held out for testing.

Loss function The model is trained to minimize an error function referred to as the loss function, which is a measure of the difference between the model's prediction and the truth label. The binary cross-entropy function used for binary classification is defined as $L_i = -(y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$ for event i where p_i is the prediction and y_i is the label which could be defined as 0 (1) for background (signal). The multi-classification task uses categorical cross-entropy for event i and is a sum of the terms of all N classes $L_i = -\sum_{j=1}^N y_{ij} \log(p_{ij})$, with the output layer which consists of N output.

Optimizer The training of the network is a process of updating the connection weight to minimize the loss. Optimizers are algorithms telling how weights are updated, which usually depends on the gradient of the loss function. The stochastic gradient descent (SGD) [142] is a popular optimizer. According to it, the weights in each layer l are updated towards the opposite direction of the gradient of loss L :

$$\mathbf{W}^l \rightarrow \mathbf{W}^l - \eta \nabla_{\mathbf{W}^l} L \quad (5.2)$$

where η is the so-called *learning rate*, a tunable parameter which has an impact on the convergence speed. Such weight updating process can be implemented for each event, or a bunch of events referred to as *mini-batch*. Each weight update is called a *batch training*. The process is repeated for all the samples of the training dataset, and a round of updating over the full training dataset is called an *epoch*. For each epoch, the total training set is randomly shuffled and therefore the batch sampling are different from the previous epoch.

The method of computing loss gradients is called *back propagation* (BP). In hidden layer l ,

the gradient of the loss $\nabla_{\mathbf{W}^l} L$ is calculated following the chain rule of derivatives:

$$\delta^l = \begin{cases} \Delta_{f^l} L \odot f'^l(\mathbf{z}^l) & l \text{ is the output layer} \\ ((\mathbf{w}^{l+1})^T \delta^{l+1}) \odot f'^l(\mathbf{z}^l) & l \text{ is the hidden layer} \end{cases} \quad (5.3)$$

$$\frac{\partial L}{\partial b_j^l} = \delta_j^l \quad (5.4)$$

$$\frac{\partial L}{\partial w_{jk}^l} = f_k^{l-1} \delta_j^l \quad (5.5)$$

where $f'^l(\mathbf{z}^l)$ is the derivative of the activation function f^l in layer l about $\mathbf{z}^l = \mathbf{W}^l \mathbf{x} + \mathbf{b}^l$. And $\odot$ is the elementwise product³. In each hidden layer l , the derivatives in Equations 5.4 and 5.5 are calculated with respect to the weights in the current layer. According to Equation 5.3, it requires the knowledge of activation function derivatives in all subsequent layers $l+1, l+2, \dots$, till the output. Therefore the calculation should be processed from the output layer, and back towards the input layer.

Although many variations of optimizers are available, SGD and two other algorithms shown to be performant in many deep learning publications, are tested and grid-searched in this study. Considering that sometimes the proper fixed learning rate in SGD is not easy to find, alternative algorithms with adaptive learning rate are widely used. Adagrad [143] has parameter-specific learning rate where the more updates a parameter receives the smaller the rate is. Adam [144, 145] is a very popular optimizer nowadays which achieves good performance in many experiments as it integrates the benefits of various algorithms including Adagrad.

Activation function The activation function does the non-linear mapping of the weighted input and it provides richer capability in the functions the network can model. The popular activation functions which are employed in this study are shown in Figure 5.3. The sigmoid function is one of the first used. It shrinks the output between 0 and 1 and has an s-shape and a well-defined derivative everywhere. The hyperbolic tangent function (tanh) has similar properties. The rectified linear unit (ReLU) [146], a piece-wise linear function, was proposed in 2010 and works well in practice, having the advantage of being fast to compute and helping reduce the issues of vanishing gradient in BP.

The multiclassifier usually has N multi-output neurons if there are N categories, and uses the so-called *softmax* as activation function in the output layer: $S_j(\mathbf{o}) = \frac{e^{o_j}}{\sum_i e^{o_i}}$, where $\mathbf{o}$ is the original output in N dimensions. This function transfer the output vector into positive continuous values smaller than 1 and the sum of all elements equals 1. Thus it can be seen as the output probabilities for each class and can be trained with BP method with loss function of categorical cross-entropy.

Data pre-processing: centralization and normalization The input features have different concepts and units ranging from angular distance to the invariant mass with a mixture of scales. The very large scale may submerge the impact from the discriminating features in small scale, or may push some neurons to go to the boundaries of the activation function where the derivative is almost zero. Thus data pre-processing is implemented where the mean

³The elementwise product of two vectors $\mathbf{a} \odot \mathbf{b}$ gives a third vector $\mathbf{c}$ with the same dimension and $c_i = a_i b_i$

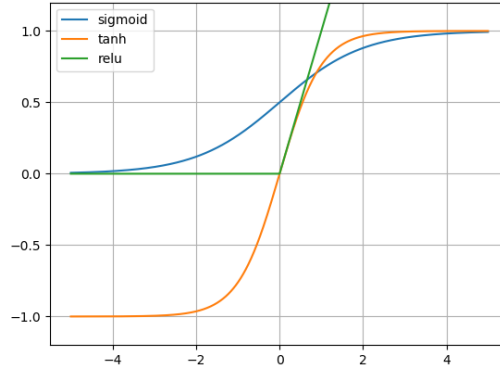


Figure 5.3: Activation functions.

value of each feature is moved to 0 (centralization) and its standard deviation is scaled to 1 (normalization).

Batch normalization Since the network benefits from the pre-processing of the input, similarly the output of the neural network layers can also be pre-processed before feeding into the next layer, which is called batch normalization [147]. During the training stage it normalizes the output of a previous activation layer of the mini-batch. When doing predictions, the global mean and standard deviation is deduced from those of all mini-batches and used to do normalization. Additional shift and scale parameters are also added to denormalize, which could be decided during the training if the normalization is not useful.

Regularizer The complexity of the deep model may lead to the over-fitting problem as described in Section 4.5.1. At the stage of over-fitting, the model fits exceptionally well with the training set, learning even the random noise on the border between two event categories. This may end up with larger contributions of non-linearity components of the border, ending up with large weights of the network. The regularization term added in the loss function $L = L_{\text{loss}} + \lambda \|\mathbf{W}\|_p$ is aimed to mitigate the over-fitting by restricting the connection weights $\|\mathbf{W}\|_p = (w_1^p + w_2^p + \dots + w_n^p)^{\frac{1}{p}}$ where n is the total number of weights. It is called L_p regularization with p often equal to 1 or 2.

Dropout [148] is another technique for regularization. The idea is to randomly drop a portion of neuron connections (dropout rate) during each batch training iteration where only the un-dropped neuron weights are updated while the dropped ones are kept unchanged. Then in a new batch training, among all the neurons another set is chosen to be dropped. Such a process is repeated. In this way, the total result is like an average of various different networks. And each neuron is forced to not rely too much on any neurons from the previous layer, thus the weights related to these neurons are controlled which is similar as $L1$ or $L2$ regularization.

Early stopping The training is monitored and stopped at some epoch as the performance on the validation set starts to deteriorate, which is a sign of over-fitting (described in Section 4.5.1). In this chapter, the training is early stopped if the AUC on the validation set does not improve in 40 epochs before reaching the specified training epoch, and the model with the maximum

validation AUC is finally saved for application. This stopping condition is applied on all the following trainings if not specified.

5.2.2 Recurrent neural networks

Neural networks with recurrent connections are called *recurrent neural networks* as shown in Figure 5.4(a) which is the simplest RNN used for classification task. The output of the hidden layer, $\mathbf{h}_t = (h_1, h_2, \dots, h_K)$, represented by K neurons, is re-directed back into itself together with the input layer. Such a structure can fit well with sequential data where the order of the elements or frames in the data is useful information, and as well, adapt to the data with variable-size sequences. The input of the simple RNN in Figure 5.4 is a sequence of vectors through time t as $\{\dots, \mathbf{x}^{t-1}, \mathbf{x}^t, \mathbf{x}^{t+1}, \dots\}$ where there are J features in each time step $\mathbf{x}^t = (x_1^t, x_2^t, \dots, x_J^t)$.

An equivalent unfolding of this cyclical structure is shown in Figure 5.4(b) where the hidden recurrent layer is extended to repetitive cells, each reading in one frame of the sequence data. The connections between these modules and inputs share the same weights for frames in the sequence, all of which share the same connection weights $\mathbf{W}_{IH}$. $\mathbf{W}_{HO}$ is the connection weight between the hidden and output layers, and $\mathbf{W}_{HH}$ is associated with connections of hidden layers at different time steps.

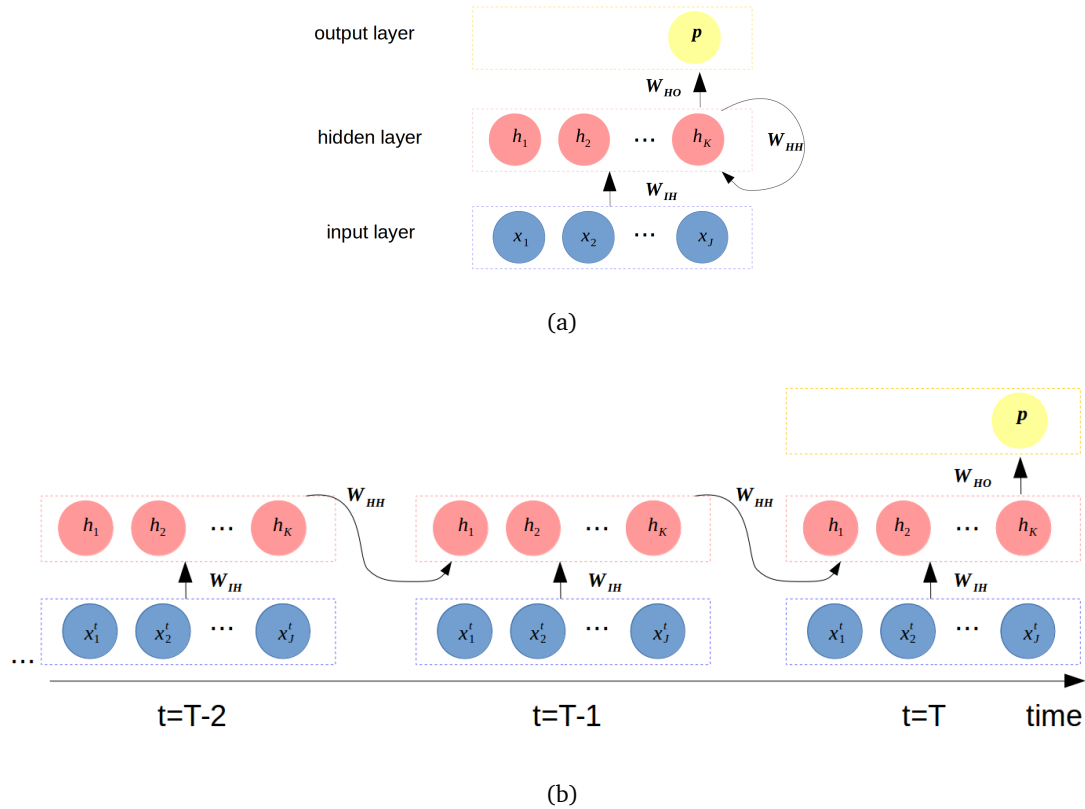


Figure 5.4: (a) Recurrent neural network with a very basic architecture. (b) The unfolded equivalence of the RNN. Only three repetitive cells are shown.

For a classification task, only the last time step hidden layer is routed out, as shown in

Figure 5.4(b), referred to as many-to-one model. It recursively nests the information from the previous steps in order, as representations for further classification layers. This is the case for the studies in this thesis. More than one time step of the hidden layer could be routed out in other tasks (many to many), for instance, in machine translation where RNN are used to convert a sentence from French to English. In other tasks like labeling an image by a sentence, the one-to-many model is used where there is only one frame in the input.

Bidirectional RNN The above described RNN is uni-directional, as it only considers the previous context of the sequence data. Instead, *bi-directional RNN (BRNN)* [149] can consider the information in both the past and future in sequence. This model has two hidden layers, one processing the sequence in the forward time direction h^f , another in the backward direction h^b as shown in Figure 5.5. And the two hidden layers are not directly interconnected.

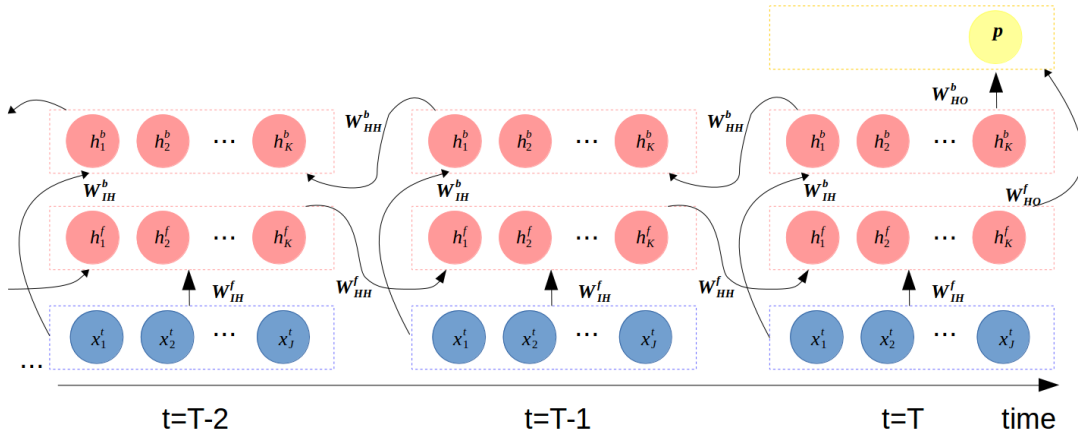


Figure 5.5: Basic structure of bidirectional RNN

The forward hidden sequence is iterated over $t = (1, \dots, T)$ and is computed as

$$h_t^f = f_H(W_{IH}^f x_t + W_{HH}^f h_{t-1}^f + b_h^f) \quad (5.6)$$

The backward hidden layer, iterated over $t = (T, \dots, 1)$ is:

$$h_t^b = f_H(W_{IH}^b x_t + W_{HH}^b h_{t-1}^b + b_h^b) \quad (5.7)$$

The output p is:

$$y = W_{HO}^f h_t^f + W_{HO}^b h_t^b + b_o \quad (5.8)$$

where for a classification task only the last time $t = T$ step is output.

Vanilla RNN and long-term dependence issues The structure shown in Figure 5.4 is the very basic RNN (Vanilla RNN). To process the sequence data, the hidden layer acts as the state of the system, $h_t = f_H(o_t)$, where $o_t = W_{IH}x_t + W_{HH}h_{t-1} + b_h$, the output $y = f_O(W_{HO}h_t + b_o)$, and f_H is the hidden layer activation function [150]. W are connection weights as above illustrated and b_h (b_o) is additional weight giving bias to the transformation performed by W_{IH} and W_{HH} (W_{HO}),

Due to the fact that the latest hidden state nests all the previous ones in sequence, the earlier time step hidden state only enters the gradient through a long chain product of derivatives. This chain has the risk to be very close to zero if most of derivatives have values smaller than one. In this case, the update of the connection weights are rarely impacted by the earlier states meaning that they are out of memory. This is the *long-term dependence problem*.

Long-short term memory RNN, and Gated Recurrent Unit Many RNN variations have been designed to tackle the above stated memory issue, and one of the most popular and efficient methods is the long-short term memory (LSTM) RNNs [151]. Instead of having a single neural network layer in the repetitive cells, controlling gates that regulates information flow are added in as shown in the left plot in Figure 5.6. The cell gate g_t^c ,

$$g_t^c = g_t^i \tanh(\mathbf{W}_{Igc} \mathbf{x}_t + \mathbf{W}_{Hgc} \mathbf{h}_{t-1} + \mathbf{b}_{gc}) + g_t^f g_{t-1}^c \quad (5.9)$$

plays as the state of repetitive cells and transports memory of the network all the way down the sequence chain. Other gates which are different neural layers decide which information is allowed to be added or dumped. The input gate g_t^i controls whether the information from a new time step should enter the memory:

$$g_t^i = \sigma(\mathbf{W}_{Igi} \mathbf{x}_t + \mathbf{W}_{Hgi} \mathbf{h}_{t-1} + \mathbf{W}_{g^i g^c} g_{t-1}^c + \mathbf{b}_{gi}) \quad (5.10)$$

and the forget gate g_t^f decides whether the information from the old steps should be forgotten:

$$g_t^f = \sigma(\mathbf{W}_{Igf} \mathbf{x}_t + \mathbf{W}_{Hgf} \mathbf{h}_{t-1} + \mathbf{W}_{g^f g^c} g_{t-1}^c + \mathbf{b}_{gf}) \quad (5.11)$$

This approach changes the hidden layer output $\mathbf{h}_t = f_H(\mathbf{o}_t)$ to $\mathbf{h}_t = g_t^o \tanh(g_t^c)$ which depends on the cell state g_t^c but is further controlled by the output gate g_t^o :

$$g_t^o = \sigma(\mathbf{W}_{Igo} \mathbf{x}_t + \mathbf{W}_{Hgo} \mathbf{h}_{t-1} + \mathbf{W}_{g^o g^c} g_{t-1}^c + \mathbf{b}_{go}) \quad (5.12)$$

The Gated Recurrent Unit (GRU) [152] is one of the LSTM variations and similarly has gating units that control the information flow, as shown in Figure 5.6. GRU gets rid of the cell gate g_t^c , and uses the hidden state $\mathbf{h}_t$ to transfer information. With a simpler repetitive cell, only two gates are used. The update gate z_t acts similarly to the forget and input gates, and decides which information to dumped or add in:

$$z_t = \sigma(\mathbf{W}_z \mathbf{x}_t + \mathbf{U}_z \mathbf{h}_{t-1}). \quad (5.13)$$

The reset gate r_t regulates how much previous information to forget:

$$r_t = \sigma(\mathbf{W}_r \mathbf{x}_t + \mathbf{U}_r \mathbf{h}_{t-1}) \quad (5.14)$$

via a candidate activation:

$$\tilde{\mathbf{h}}_t = \tanh(\mathbf{W}_h \mathbf{x}_t + \mathbf{U}_h (r_t \odot \mathbf{h}_{t-1})), \quad (5.15)$$

The activation in a GRU is linearly modeled as:

$$\mathbf{h}_t = (1 - z_t) \mathbf{h}_{t-1} + z_t \tilde{\mathbf{h}}_t. \quad (5.16)$$

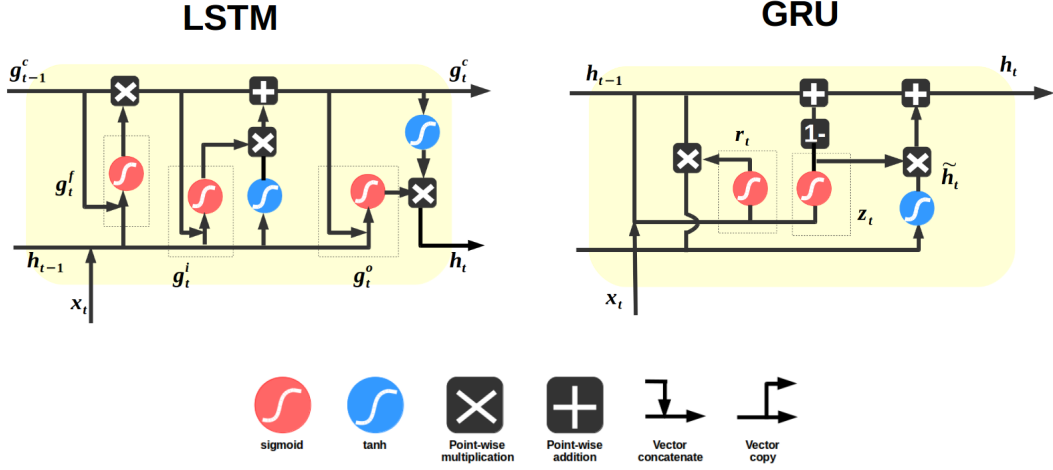


Figure 5.6: The repetitive cell architecture of the LSTM and GRU.

Many recent results based on RNN are achieved with LSTMs, and the best choice varies from task to task. The GRU and the standard LSTM are tested in this study.

5.2.3 Hyper-parameter optimization

Many of the above stated components of the neural networks have to be optimized, for example, the model architecture including the number of layers and the number of hidden units, the type of activation function, the usage of regularization or batch normalization, the various optimizers and their technical parameters, and so on. Even though the list is not complete, there is already an impressive number of configurations to test in order to optimize the network, which requires a thorough procedure.

Grid search A simple and straightforward algorithm is grid search. For each variable parameter, the set of candidate values is firstly defined. Then the models for all possible parameter combinations are trained, and the one with the best testing performance is chosen.

It is very computing-expensive to do parameter optimization in this way if we would like to search a large enough hyper-parameter space. The average training time for each configuration in this study is around two hours on GPU, which varies with respect to different parameter settings, e.g. dropout rate and the complexity of the network. The number of combinations could go up to thousands or millions. In order to get a statistical fluctuation resistant result, usually the *k-fold cross-validation*⁴ is implemented on the learning sample to get *k* different splittings of training and validation samples. The averaged *k* validation results are then used to choose the best model. This will further explode training time.

⁴During the *k*-fold cross validation, the data set *D* firstly gets split into mutually exclusive subsets, $D = D_1 \cup D_2 \cup \dots \cup D_k$, $D_i \cap D_j = \emptyset (i \neq j)$. Each time the union of $k - 1$ subsets is used for training, and the remaining subset for validation. In total, *k* different splittings are implemented, giving *k* trainings. The averaged result of these *k* models is used as the final result.

Tree-structured Parzen Estimators The Tree-structured Parzen Estimators (TPE) method [153] aims to find a better way for selection of the best parameters when searching in the hyper-parameter space. Similar to the grid search, the candidate parameters are pre-defined. In each iteration, the parameter configuration is selected by learning from the previous experiments.

For the first few iterations, the TPE algorithm needs to warm up. Several parameter configurations are randomly selected, trained and evaluated. Among these observations, a fraction of the best observations are chosen and are used to build the probability density distribution, referred to as good *pdf* $g(x)$ where x is all hyper-parameters. The rest of the observations are grouped and the corresponding bad *pdf* $b(x)$ is fitted. Starting from these initial *pdfs*, the iterative steps are:

- Sampling M parameter configurations according to the good *pdf*.
- Finding the best among these M configurations that is more likely to be in the group following good *pdf* and less likely to be in the one following the bad *pdf*. The metric is a ratio of *pdfs* defined as expected improvement, $EI(x) = \frac{g(x)}{b(x)}$. The one with the highest $EI(x)$ score is chosen for the next iteration
- Dividing the M samples into good and bad groups according to the evaluation performance, and updating the *pdfs* with them respectively.

Such steps are repeated until the desired maximum number of iterations.

The algorithm to model the *pdfs* of good and bad observations is the Parzen-window density estimation [153] which is a non-parametric way to build a probability density function from empirical data. And the tree-structured means that parameters have tree-structured dependencies. For example, the number of neurons in the second hidden layer is only defined when the second hidden layer exists in a configuration.

5.2.4 Tools and software

All the training, testing, and performance checks of the deep learning models before sending to the ATLAS framework are implemented with Python and its relative libraries⁵, including Keras [155] (version 2.0.8) as a wrapper library on top of TensorFlow [156] (version 1.2.0) which is for fast numerical computing, scikit-learn [157] (version 0.19.1) that is a machine learning library, and uproot [158] (version 2.8.16) to read branches in ROOT files into the Python environment directly as Numpy arrays [159]. The TPE method is implemented in this study using hyperas [160] which is a wrapper around hyperopt [161], a library for hyper-parameter optimization with Keras models.

The study is conducted on the LIS⁶ cluster which gathers 12 hosts with 24 CPUs (188 cores), and 23 GPUs. The deep models are trained on these GPUs, of which the type and corresponding number are listed in Table 5.1.

Type	Tesla K80	Tesla K40c	GeForce GTX 1080 Ti	Tesla K20m	Tesla K40m
Number	10	2	8	1	2

Table 5.1: GPU type and number on LIS cluster.

⁵Under Adaconda, an open source package and environment management system [154]

⁶Laboratoire d'Informatique et Systèmes

5.3 Using recurrent neural networks to classify $t\bar{t}H$ signal and $t\bar{t}$ background

In the last round of analysis [25], the $t\bar{t}H$ and $t\bar{t}$ classification is a two-step MVA model. Intermediate MVAs are firstly trained to find the proper matching between jets and partons, to first build the features of the $t\bar{t}H$ signal. Three MVA methods are used, where the reconstruction BDT and LHD are trained with similar input features, but they exploit information from different aspects, as described in Chapter 4.

The reconstruction BDT (see Section 4.5.3) are trained to choose the best combination (with the highest BDT output) among all possible assignments. Then the high-level features of the $t\bar{t}H$ signal are built based on this combination. However, this method gives limited efficiency to have the correct event reconstruction. With the dataset containing events of ≥ 6 jets and $\geq 4b$ -jets tagged as 85% (referred to as 6ji4bi@85%), only around 30% of events are reconstructed with the correct association. In fact, around 26% of events have the correct combination ranked as the 2nd largest BDT output, and around 14% and 11% as the 3rd and 4th, respectively. Therefore, the $t\bar{t}H$ discrimination features are not well built if only the best combination from the reconstruction BDT is used. Consequently, the performance of the classification BDT using these features as input is also impacted.

On the other hand, although the likelihood discriminant (LHD) (see Section 4.4.2) combines the probability density functions under signal and background hypotheses individually incorporating each of their combinations separately, the relationships between input features are not taken into account. The matrix element method (MEM) is only available in the ultra pure signal region as it is expensive to compute.

Studies in this section show the possibility to incorporate the merits of both LHD and reconstruction BDT, considering both more jet-parton combinations and the correlations between reconstruction input features at the same time as inputs to the classification model. Multiple combinations make up a set, like a sequence but without an obligatory order between elements. The Recurrent Neural Network (RNN) model is exploited, because of its good behavior dealing with sequence data.

5.3.1 RNN in $t\bar{t}H$ ($H \rightarrow b\bar{b}$)

The RNN is trained to classify the $t\bar{t}H$ signal and the $t\bar{t}$ + jets background benefiting from its cyclic structure. Each signal or background event can be seen as a sequence, the frames of which are each possible jet-parton combination, as sketched in Figure 5.7. Such a frame is represented by a set of input features consisting of variables related to the jets assigned by the corresponding combination and variables independent from combinations. In this way, both the multiple combinations as well as the correlation between features are considered as inputs to the machine learning model, realizing the combination of both reconstruction BDT and LHD.

This section shows the study of using RNN in the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis. The same samples as in the BDT section are always used for performance evaluation on the testing set, which allows to do fair performance comparison directly. All the RNN models are trained on events of 6ji4bi@85% (inclusive training, see Section 4.5.3.4), and this is also the default training set if not otherwise specified for the other machine learning models explored in this chapter.

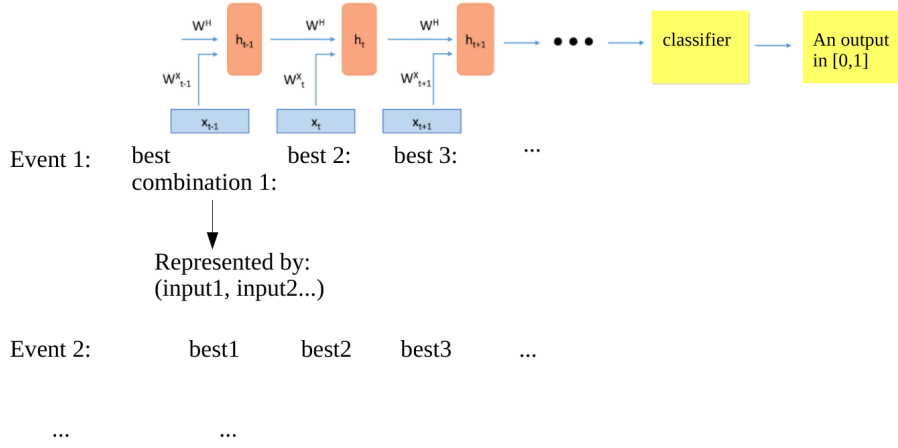


Figure 5.7: Sketch of the RNN reading in $t\bar{t}H$ events, each being sequence made of possible jet-parton combinations.

5.3.1.1 Number of combinations

Initially, the RNN was envisioned as a better intermediate MVA more discriminating than the reconstruction BDT and the LHD as input to the classification BDT. The first structure comprises the input layer, an LSTM layer with 100 neurons, and a one-dimensional output layer. It is trained using the binary cross-entropy loss function, using the same combination-dependent features without Higgs topological information as the reconstruction BDT listed in the first section in Table 4.2.

Figure 5.8 shows the ROC curves of LHD, reconstruction BDT and the RNN as discriminant for $t\bar{t}H$ and $t\bar{t}$. For an specific value of the background selection efficiency (the x -axis), the higher the curve is, the better the model performs as it gives higher signal selection efficiency. In Figure 5.8(a), each input event of the RNN contains only the three leading combinations sorted by the recoBDT score (the one without using Higgs information), and the RNN shows very close performance to the LHD. Then it gives a better performance as shown in Figure 5.8(b) after adding up to 12 leading combinations. Although being comparable or worse compared with the RNN, the LHD uses more information including all possible combinations, b -tagging weights, and a Higgs mass built from two b -jets. This shows that the RNN performs better with less information. The recoBDT performs slightly better than random guessing, which is not surprising as it only sees combinations from the $t\bar{t}H$ sample during training and is trained to separate the correct matchings from the improper ones on $t\bar{t}H$ events only (detailed in Section 4.5.3.3). Therefore it is not expected to have very good discrimination power between $t\bar{t}H$ and $t\bar{t}$.

For the events with ≥ 6 jets and $\geq 4b$ -jets, the number of combinations is at least 12 and has variable size. The study in this thesis exploits only 12 leading combinations sorted by the recoBDT, but there is no substantial difficulty moving to the full size-variable combinations, apart from disk and memory space to store the information and output computation when applying the RNN, as the number of combinations can go in the thousands.

When using the 12 combinations, it was checked that using the order sorted by the recoBDT or the recoBDT_withH gives similar AUC performance. The following studies will always employ the 12 combinations ordered by the recoBDT_withH.

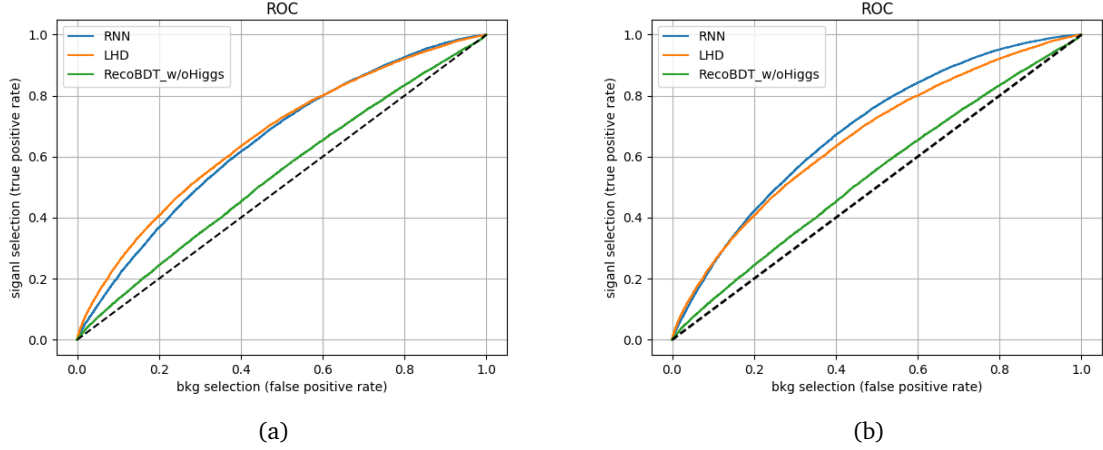


Figure 5.8: ROC curves in 6ji4bi@85% of reconstruction BDT (without Higgs information), LHD and RNN which only uses 15 reco-level inputs without Higgs-related features as shown in the first section in Table 4.2 with (a) leading 3 combinations and (b) 12 leading combinations sorted by recoBDT score.

5.3.1.2 Adding combination-independent and b -tagging variables

In addition to the above mentioned 15 features, more variables as used by the classification BDT are added as inputs to the RNN, to have their performance fairly compared. The Higgs-related input variables, as listed in the second section in Table 4.2, are firstly added in each combination of the sequence. As shown in Figure 5.9(a), the RNN (in blue, named RNN_withH_allComb in the legend) starts to be comparable with the state-of-the-art BDT without b -tagging (in red, named classBDT_Inclusive in the legend).

Feeding in each combination more features similar to the ones from the classification BDT, including the same global kinematics, listed in the first section of Table 4.5, and the b -tagging variables, the RNN (in blue, named RNN_withH_allComb_withBTag) finally becomes equivalent to the inclusively trained classification BDT (in purple, named classBDT_Inclusive_withBTag) in 6ji4bi@85% as shown in Figure 5.9(b). The global variable values are the same for each combination, and therefore they are also tried to be fed into the RNN in another way, separately from combinations: concatenating the last hidden layer of RNN with the global variables and then connecting this concatenation to the following layers until the classification output. The two feeding methods give the same classification performance.

The pseudo-continuous b -tagging weights of the six leading jets (sorted by b -tagging score) are used for the RNN. This is different from the b -tagging variables employed by the BDT which includes b -tagging scores of three of the jets and the sum of binned b -tagging weights of jets from the best Higgs candidate as listed in the last section in Table 4.5.

These inclusively trained MVAs are also tested in the UPSR, and their performance in terms of AUC is summarized in Table 5.2. With around one to two hours of training, the RNN gives comparable performance in both inclusive regions and the UPSR where the dedicated BDT is trained using the MEM discriminant. This makes it a promising substitute for the two-step MVAs composed of multiple techniques, of which the MEM is much more expensive to compute.

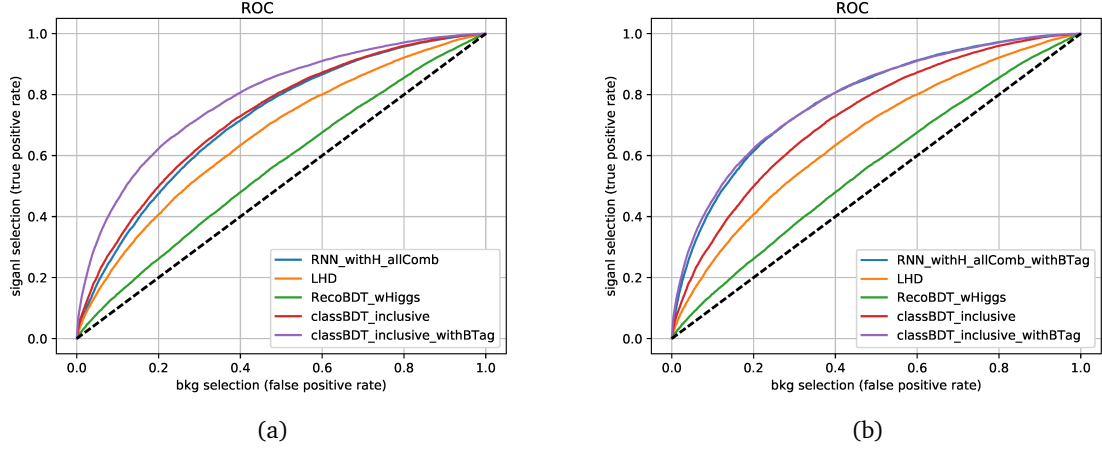


Figure 5.9: ROC curves in 6ji4bi@85% of reconstruction BDT (with Higgs information), LHD, classification BDTs with and without b -tagging variables, and RNN which only uses reconstruction-related inputs including the ones with Higgs related features as shown in the second section in Table 4.2. The RNN input in (a) does not include any b -tagging variables and in (b) includes the b -tagging scores of all six jets in each combination, and all the event-level kinematics used by the BDT.

	AUC (inclusive)	AUC (UPSR)
RNN	0.786	0.758
LHD	0.664	0.727
Best BDT	0.789	0.760

Table 5.2: AUC performance of MVAs in 6ji4bi@85% inclusive region and 6ji4bi@60% UPSR. The MVAs are trained in the inclusive region.

5.3.1.3 Sequence order

The RNN is a neural architecture for processing sequence data where special information lies in the order of the sequence. One example is natural language processing where sequence of words in a text has a natural ordering. What is more, the learning performance optimization can be further impacted by the order of input sequence [162]. When the order of the input sentence is reversed, the machine translation performance is found to be improved [163]. Thus the order of input data requires careful study.

For the $t\bar{t}H$ combination, different orders of combination are tested. The RNN is trained in the inclusive region with either increasing, decreasing, or random order, according to the ranking of reconstruction BDT score. They are tested in this same region and all give similar performance, as shown in Figure 5.10(a). Since only 12 best combinations sorted by the recoBDT are used, the impact of ordering may be under-estimated in the inclusive region where more events have more than 12 combinations. Thus the above trained models are also tested in the region where events have exactly 6 jets and exactly 4 b -jets (6je4be@85%) and thus have exactly 12 combinations. The results are shown in Figure 5.10(b) and the three orders again give similar performance.

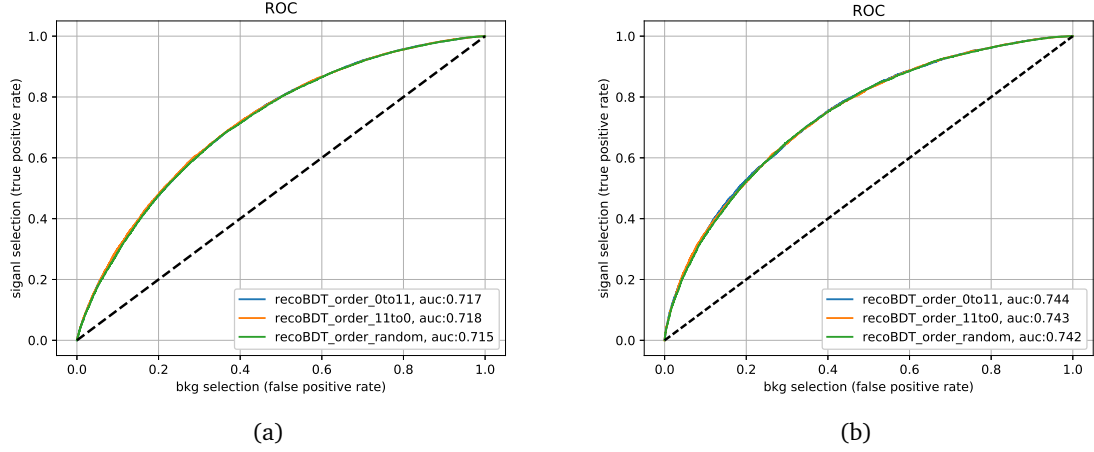


Figure 5.10: ROC curves of RNN using various sequence orders tested in the region of (a) 6ji4bi@85% and (b) 6je4be@85%. The model is trained in the 6ji4bi@85% region.

5.3.2 RNN performance as a function of training sample size

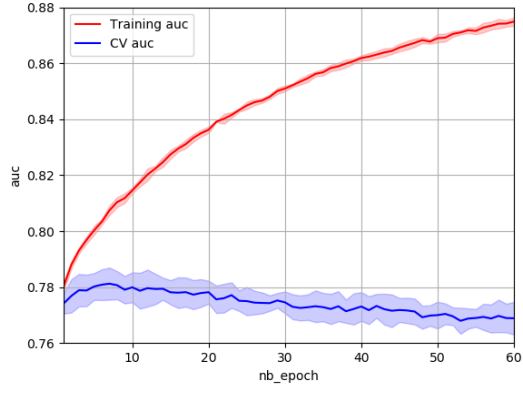
Aiming to check the impact of MC statistics on the RNN performance, the full $t\bar{t}H$ and $t\bar{t}$ nominal samples are exploited and then are decreased by hand. Respectively, 100%, 90%, ..., until only 10% of the $t\bar{t}$ events are sampled, and then used together with the full $t\bar{t}H$ sample. A 5-fold cross validation is then implemented: firstly the sampled $t\bar{t}$ and $t\bar{t}H$ sample are randomly partitioned into five equal-size sub-samples, then each time one of the sub-samples is held-out as the test sample to evaluate the model and the remaining four partitions are mixed together as training set. This ends up with five testing results which are averaged as a final estimation.

The model deteriorates with low statistics of 10% of $t\bar{t}$ size. Figure 5.11 shows the AUC evolution during the training with full and 10% of $t\bar{t}$ sample. With low statistics, the model has larger difference between training and testing averaged AUC. With low statistics, the testing performance has a lower AUC peak value of 0.781 culminating at epoch 7 and then deteriorating, which means it starts to suffer from over-fitting at this epoch. On the contrary, the high statistics starts test starts to suffer from over-fitting much later, with an averaged peak AUC of 0.796 reached after 20 training epochs.

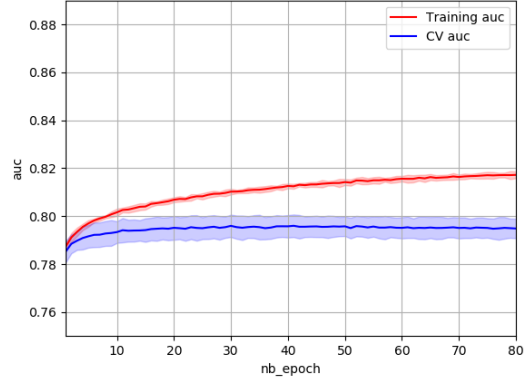
For the training with each of the above mentioned $t\bar{t}$ sample portions, the maximum testing AUC and at its corresponding epoch the training AUC, are picked, depicting the AUC trend impacted by the statistics. The $t\bar{t}H$ sample goes through the same statistical check procedure: using from 10% to the full $t\bar{t}H$ sample together with the full $t\bar{t}$ sample and 5-fold cross validation. Both of the statistics impact are shown in Figure 5.12. For our odd-even cross training routine, the training sample in each fold takes a portion of 40% (excluding validation sample) of both $t\bar{t}H$ and $t\bar{t}$, respectively. According to the curves in Figure 5.12, the RNN performance is not limited by the current statistics of the training sample.

5.3.3 Hyper-parameter grid search

Some hyper-parameters of the LSTM model are grid-searched. Table 5.3 shows the searched parameters and their values. There are two kinds of internal dropouts inside the LSTM layer,

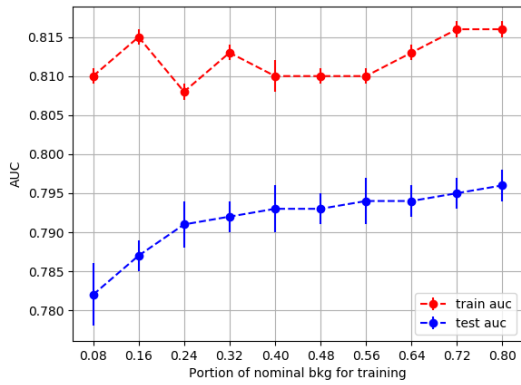


(a)

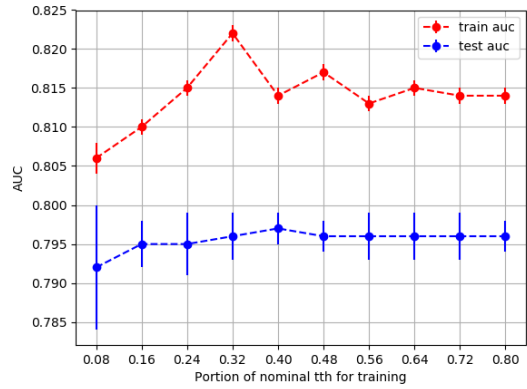


(b)

Figure 5.11: The AUC evolution with respect to training epoch for (a) 10% and (b) 100% of $t\bar{t}$ sample used with the full $t\bar{t}H$ sample. The solid line is the central value of AUC averaged with 5-fold cross validation results at each epoch, and the error band is the corresponding standard deviation.



(a)



(b)

Figure 5.12: The statistics impact on AUC using various portions of (a) $t\bar{t}$ sample and (b) $t\bar{t}H$ sample.

dropout_W, fraction of the input neurons to drop for input gates, and dropout_U, fraction of the input neurons to drop for recurrent connections. The optimizer is Adam for which the learning rate is also searched. For each neuron layer, the activation function ReLU is used without grid search.

parameter name	values
number of neurons in the RNN layer	30, 60 , 100, 500
dropout_U, and RNN dropout_W in the RNN layer	0., 0.2, 0.4
number of FC hidden layers in addition to RNN layer	0 , 1, 2, 3
number of neurons in FC hidden layers	30, 60, 100, 500
dropout rate of FC hidden layers	0, 0.2, 0.4
learning rate	0.05, 0.01, 0.001 , 0.0001, 0.00005

Table 5.3: Grid-searched hyper-parameters for the RNN. The optimized values are in bold. None of the values in the rows of number of neurons and the dropout rate of fully connected (FC) hidden layers are in bold, because the optimized number of FC hidden layers equals zero.

During the grid search, the model learns with even events only, where 20% are used as validation events and 80% as training sample. The model providing the maximum AUC on the validation set is chosen as the best model. The odd events are not touched in this grid search process. The best model structure is shown in Figure 5.13(a). The model has no hidden layers in addition to the LSTM one which has 60 neurons with both internal dropout rates optimized at 0.4. Figure 5.13(b) shows the AUC evolution as the training of the best model evolves. The model stops at epoch 61 with the early stopping where the model does not suffer from over-fitting. The model is then tested on odd events, and the ROC curve is shown in Figure 5.14(a). Using the optimized architecture, the model is re-trained on odd events, and then applied on even events. Figure 5.14(a) shows that the ROC curves of the two trainings are very similar. The AUC of both folds is better than the state-of-the-art BDT and the un-optimized LSTM, which is summarized in Table 5.4.

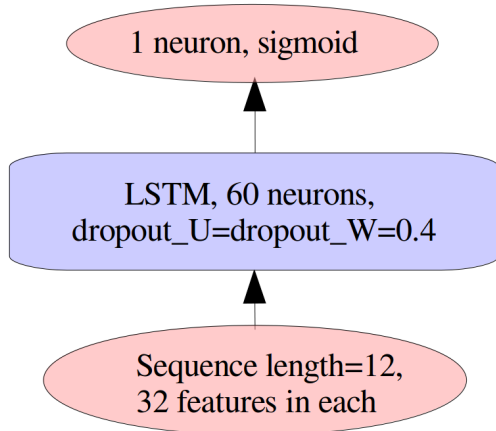
	BDT	un-optimized RNN	optimized RNN
even_odd	0.789	0.788	0.790
odd_even	0.791	0.793	0.793

Table 5.4: AUC of testing sample with state-of-the-art BDT and the optimized RNN.

The optimized LSTM architecture is then adopted for the models trained with events having exactly 5 jets. The ROC curve performance is shown in Figure 5.14(b) and the models will be used in the following fit study.

5.3.4 Summary

In this section, recurrent neural networks are exploited to do the binary classification of $t\bar{t}H$ signal and $t\bar{t}$ background, dealing with the jet-parton association problem at the same time. It can learn more information from the input than the reconstruction BDT and the likelihood discriminant in the previous two-step MVA. Using similar input features, the RNN gives even slightly better AUC performance than the two-step MVA which consists of various techniques.

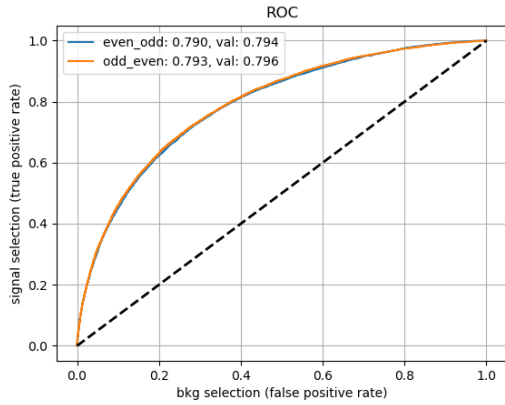


(a)

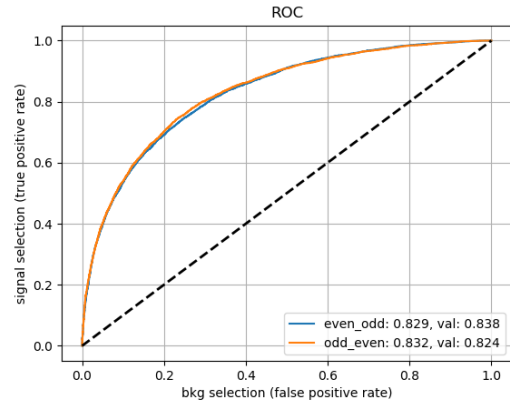


(b)

Figure 5.13: (a) The structure of the best LSTM model. (b) The AUC evolution during the training of the best model.



(a)



(b)

Figure 5.14: ROC curves of the LSTM models with two-fold training and testing, for both events with (a) at least 6 jets, and (b) exactly 5 jets. The legend shows the AUC value of both testing, and validation denoted as val.

The statistics of the current dataset is not a limiting factor for the RNN performance with optimized architecture. Only the leading 12 combinations sorted by the reconstruction BDT score are used in this study, and their order in the sequence has not impact on the performance. In the future, it would be possible to switch to using the full variable-size list of combinations as input without ranking from the reconstruction BDT.

5.4 Event classification using RNN-based multi-classification

One of the important strategies of the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis is the event categorization, in order to improve the signal significance. The full event space is split into categories, called regions, with the goal to have signal regions with as much signal as possible, and to have background regions with enhanced background, as described in Sections 4.4.1 and 5.1. The fit is performed across all regions. The signal rich regions give the main contribution to the final significance. In the background regions, there is expected to have almost only background, and when fitting the MC simulation to data, the uncertainties of the prediction can be mitigated.

In reference [25], the whole event space is scrutinized by people, as discussed in Section 4.4.1, by manually tuning the requirements imposed on the b -tagging weights of multiple jets. Only four jets with the leading b -tagging weights are considered. The obtained $t\bar{t} + \geq 1b$ enriched regions are still heavily contaminated by other $t\bar{t} + \text{jets}$ components, so are the $t\bar{t} + \geq 1c$ enriched regions.

This motivates the exploration of using deep learning models to categorize events. The deep neural networks can naturally deal with multi-classification tasks by having multiple neurons in the output layer. By using RNN, multiple jet-parton combinations can be taken into account. And as input variables, the b -tagging weights of more jets can be employed. In this section, a new event categorization method based on a deep neural network is explored.

5.4.1 Model architecture and training setups

The neural network is trained to do multi-classification to classify events into $t\bar{t}H$, $t\bar{t} + \geq 1b$, $t\bar{t} + \geq 1c$, and $t\bar{t} + \text{light}$ categories, each forming the region enriched in the corresponding component. The neural network consists of the LSTM stacked by feed-forward fully connected layers, referred to as *FC layers*, with an output layer having four neurons each corresponding to one category. The model architecture is illustrated in Figure 5.15. The LSTM is the optimized structure from Section 5.3.3. The FC layers are finally connected to the output layer via the softmax function. The truth label for each event is a one hot coding vector composed of four elements which in sequence represent $t\bar{t}H$, $t\bar{t} + \geq 1b$, $t\bar{t} + \geq 1c$, and $t\bar{t} + \text{light}$, respectively. For example, the $t\bar{t}H$ event is labeled as (1,0,0,0), and $t\bar{t} + \geq 1c$ as (0,0,1,0). The categorical cross-entropy is used as the loss function.

Training setup

In the MC simulations of training set under the assumption of the standard model, there is roughly $O(10^2)$ more $t\bar{t}$ events than $t\bar{t}H$ events. The $t\bar{t} + \geq 1b$, $t\bar{t} + \geq 1c$ and $t\bar{t} + \text{light}$ cross-sections are also different from each other. All these categories are balanced by multiplying a class factor to the sample weight of events in each category, re-weighting each background component to the $t\bar{t}H$ category. This gives the same weighted sum of events for all categories which are treated equally in this way during the training procedure.

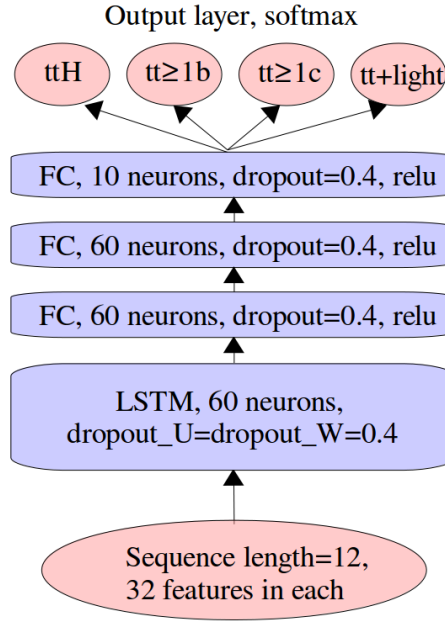


Figure 5.15: The DNN multi-classifier architecture.

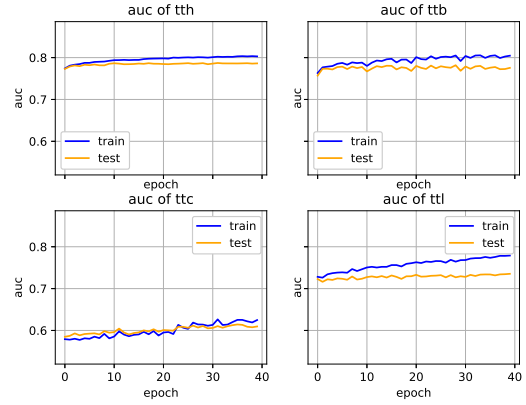
With balanced categories, the so-call macro-averaging ROC curve is also monitored during the training. For each class, a ROC curve can be obtained from the corresponding output neuron which is used as a discriminant of the events from this class against all the others (one-vs-all). The macro-averaged curves are obtained by averaging all these ROCs with equal weights. Figure 5.16 shows, on both training and validation samples, the macro-AUC evolution during the training as well as the ones from each of the outputs used as one-vs-all binary discriminant. The training is early stopped at epoch 28 according to the validation loss. At this point, the macro-AUC also reaches a stage without obviously increasing. Both the $t\bar{t}H$ and $t\bar{t} + \geq 1b$ AUCs reach a plateau at this stage, while the $t\bar{t} + \geq 1c$ and $t\bar{t} + \text{light}$ AUCs are still increasing. In future studies, more training epochs could be tried to have potential improvements on the $t\bar{t} + \geq 1c$ and $t\bar{t} + \text{light}$ separation, if the $t\bar{t}H$ and $t\bar{t} + \geq 1b$ AUCs stay on plateaus.

5.4.2 Performance

To measure the performance of the multiclassifier on testing samples with imbalanced classes, the confusion-matrix is used to present how the model behaves on various classes. For each class of events in ground truth, it gives the number (5.17(a)) or the portion (5.17(b)) of events that are classified into the proper category according to the model's prediction. An event is labeled as a certain class if its corresponding output neuron is the maximum among the four output neurons. A diagonal matrix means a perfect classifier giving proper predictions for all categories. The results in Figure 5.17 are obtained with the testing samples where the sample weights are also balanced. The multiclassifier is shown to be confused by the similarities between some classes: the $t\bar{t} + \geq 1b$ is more often predicted as $t\bar{t}H$ than as $t\bar{t} + \geq 1b$ itself while the $t\bar{t} + \geq 1c$ are categorized as $t\bar{t} + \text{light}$ very often.

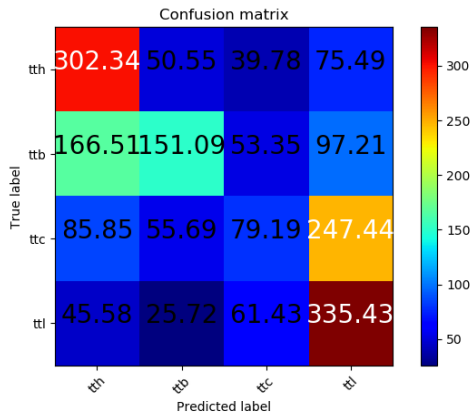


(a)

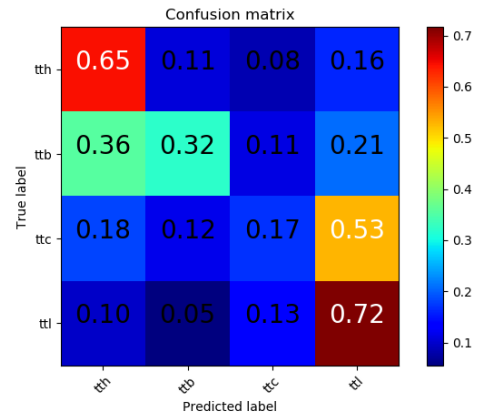


(b)

Figure 5.16: The AUC evolution with respect to epoch during the training of multiclassifier. (a) is for the macro-averaging AUC as described in the text, and (b) shows the evolution of AUCs derived from each output used as one-vs-all binary discriminant.



(a)



(b)

Figure 5.17: Confusion matrices of the multi-classifier with four categories: $t\bar{t}H$ denoted as tth, $t\bar{t} + \geq 1b$ as ttb, $t\bar{t} + \geq 1c$ as ttc, and $t\bar{t} + \text{light}$ as ttl. (a) is the number of properly classified events for each category in truth, and (b) is the fraction of events that enter the proper category after the prediction of the classifier. (b) is obtained by normalizing each element in (a) by the sum of the corresponding row.

5.4.3 Event categorization

The multi-classifier outputs can be employed in the same way as in reference [134] where the events are categorized in the same way as the confusion matrix. For instance, the $t\bar{t}H$ enriched region is formed with events for which the $t\bar{t}H$ neuron gives the highest score among the four. The same goes for other regions designed to be rich in $t\bar{t} + \geq 1b$, $t\bar{t} + \geq 1c$, and $t\bar{t} + \text{light}$. In each region, the output distribution of the corresponding neuron will be used as final discriminant in the simultaneous fit.

Figure 5.18 shows the fraction of the $t\bar{t}H$ signal and the different background components in each region defined with the multiclassifier. Each of these regions is indeed enriched in signal or one of the background sub-components. The $t\bar{t}H$ fraction is enhanced in the so-defined $t\bar{t}H$ enriched region, or called signal rich region as shown in Figure 5.18(a), where S/B is 4.9% and the $S/\sqrt{B}$ is 2.41, and the dominant background is $t\bar{t} + \geq 1b$ with the largest purity of 55.9%. The $t\bar{t} + \geq 1b$ region is also overwhelmed by the $t\bar{t} + \geq 1b$ component but also has non-negligible contamination from $t\bar{t} + \geq 1c$ and $t\bar{t} + \text{light}$. And the same case goes for the $t\bar{t} + \text{light}$ region. In the $t\bar{t} + \geq 1c$ region, the three background components are almost comparable where the $t\bar{t} + \geq 1c$ still takes the largest fraction. These background decomposition results need to be compared with the ones in Reference [25] (referred to as default definition) obtained with the algorithm described in Section 4.4.1.

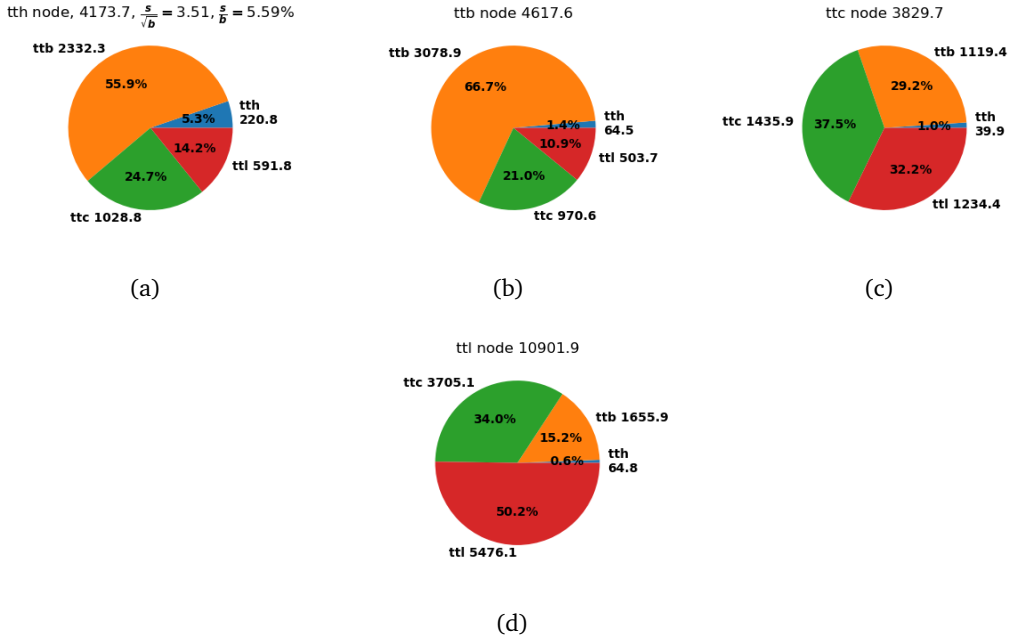


Figure 5.18: Fractional contributions of all background components and the $t\bar{t}H$ signal for all regions defined with the deep learning multi-classifier with four categories, for events with ≥ 6 jets and $\geq 4b$ -jets tagged at 85%.

The default definition is represented in Figure 5.19, in the same event space of ≥ 6 jets and $\geq 4b$ -jets tagged at 85% WP as the results of the new classifier. This space is smaller than the one selected in Reference [25] which requires $\geq 2b$ -jets tagged at 60% or $\geq 3b$ -jets tagged at 77% with ≥ 6 jets. Thus compared with the pie-charts in Figure 4.5, the background control regions in Figure 5.19 are not complete: the $t\bar{t} + \geq 1b$ enriched region is missing and the

$t\bar{t}$ + light enriched region is smaller. The $t\bar{t}$ + $\geq 1c$ control region and the three signal regions are almost the same except that the $t\bar{t}V$ and non- $t\bar{t}$ samples are not taken into account. An inclusive signal region, referred to as the default inclusive signal region, is also shown merging the three signal regions to facilitate the comparison with the new method.

In Figure 5.18 and Figure 5.19, the signal region and the $t\bar{t}$ + $\geq 1c$ region are more or less comparable. The fractions of all the background components and the $t\bar{t}H$ signal are quite similar in the $t\bar{t}$ + $\geq 1c$ enriched region, although the new definition has lower statistics. The newly defined signal region with the multiclassifier gives similar number of $t\bar{t}H$ events but better S/B and $S/\sqrt{B}$ compared with the default inclusive signal region.

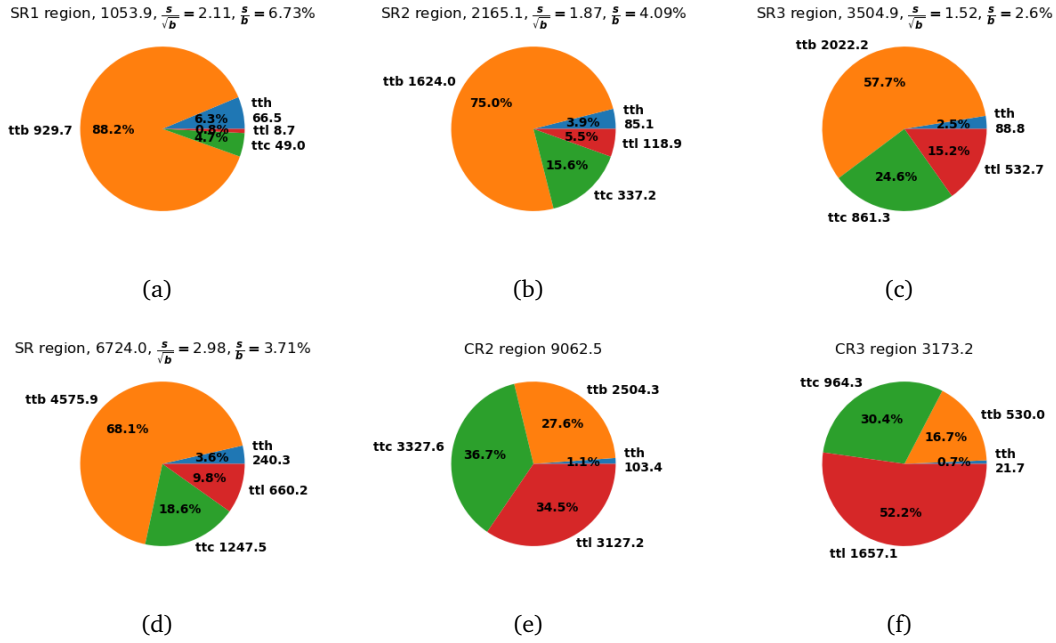


Figure 5.19: Fractional contributions of all background components and the $t\bar{t}H$ signal for regions defined in the analysis described in Section 4.4.1. Only the events of 6ji4bi@85% are shown so that the $t\bar{t}$ + $\geq 1b$ enriched region is missing, and the $t\bar{t}$ + light enriched region is incomplete.

5.4.4 Finer classification of the $t\bar{t}$ + $\geq 1b$ background

The $t\bar{t}$ + $\geq 1b$ background could be further categorized into $t\bar{t} + B$, $t\bar{t} + b$ and $t\bar{t} + \geq 2b$ as defined in Section 4.6.1.1, ending up with six classes together with the $t\bar{t}H$, $t\bar{t}$ + $\geq 1c$, and $t\bar{t}$ + light. A finer RNN-based multi-classifier having six outputs is also studied following the same training strategy. The $t\bar{t}$ + $\geq 1b$ inclusive, $t\bar{t}$ + $\geq 1c$ and $t\bar{t}$ + light backgrounds are individually re-weighted to the $t\bar{t}H$ category. Then among the $t\bar{t}$ + $\geq 1b$ events, all sub-components are scaled to $t\bar{t} + B$. The confusion matrix is presented in Figure 5.20. Being more visible in Figure 5.20(b), the $t\bar{t} + b$ component is more often predicted as $t\bar{t}$ + $\geq 1c$ than as $t\bar{t} + b$ itself, which is understandable because the two components can be confused if a c -jet is mis-tagged as a b -jet. The $t\bar{t} + \geq 2b$ has the closest final state to the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) and is therefore more difficult to be separated from it. 26% of $t\bar{t} + \geq 2b$ are mis-predicted as $t\bar{t}H$ which is comparable to the fraction (30%) of the correctly predicted $t\bar{t} + \geq 2b$ events.

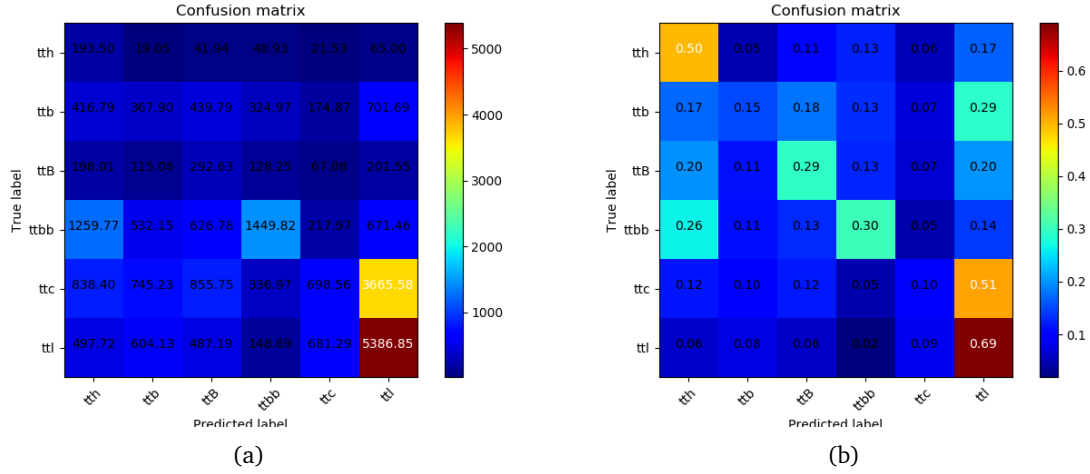


Figure 5.20: Confusion matrices of the multi-classifier with six categories: $t\bar{t}H$ denoted as tth, $t\bar{t} + B$ denoted as ttb, $t\bar{t} + b$ denoted as ttB, and $t\bar{t} + \geq 2b$ denoted as ttbb, respectively, $t\bar{t} + \geq 1c$ denoted as ttc, and $t\bar{t} + \text{light}$ as tti. (a) is the number of properly classified events for each category in truth, and (b) is the fraction of events that enter the proper category after the prediction of the classifier. (b) is obtained by normalizing each element in (a) by the sum of the corresponding row.

Fractional contributions of all six classes are shown in Figure 5.21. In the $t\bar{t}H$ rich region, $t\bar{t} + \geq 2b$ is the dominant background. In each of the background rich regions, the corresponding component is also enhanced compared with other regions.

In order to be compared with the 4-category multi-classification, the $t\bar{t} + \geq 1b$ sub-components are merged when checking the performance. The confusion matrix of the training is presented in Figure 5.22. With finer classification during the training, more $t\bar{t} + \geq 1b$ inclusive backgrounds are correctly labeled compared with the four-category multi-classifier, while less fraction of $t\bar{t}H$ is properly classified. The background decomposition with this classifier is shown in Figure 5.23. In the $t\bar{t}H$ rich region, it gives slightly higher signal purity, 5.7%, but marginally less significance, 3.42, compared with the four-category classifier having purity of 5.3% and significance of 3.51. For the $t\bar{t} + \geq 1b$ control region, the finer classifier gives lower $t\bar{t} + \geq 1b$ purity (56.5%) compared with the four-category one (66.7%).

The performance related with $t\bar{t} + \geq 1c$ and $t\bar{t} + \text{light}$ backgrounds are very similar for the two classifiers. Further studies should be carried out to use new variables that are sensitive to c -jets to improve the $t\bar{t} + \geq 1c$ categorization. More training epochs should also be tested to see if the $t\bar{t} + c$ can be further separated without deterioration in the $t\bar{t}H$ and $t\bar{t} + b$ performance.

5.4.5 Summary

In this section, the deep learning model is studied as a four-category multi-classifier to classify events into regions, each being rich in $t\bar{t}H$ ($H \rightarrow b\bar{b}$) signal, $t\bar{t} + \geq 1b$, $t\bar{t} + \geq 1c$, or $t\bar{t} + \text{light}$ subcategories. The region definition is based on model's output which is a vector with each element corresponding to one category.

Without optimization, this new region definition gives already as good performance as the scrutinized manual splitting in terms of background decomposition and signal purity. In parallel, the element of output vector, that corresponds to the signal category, can be used as a

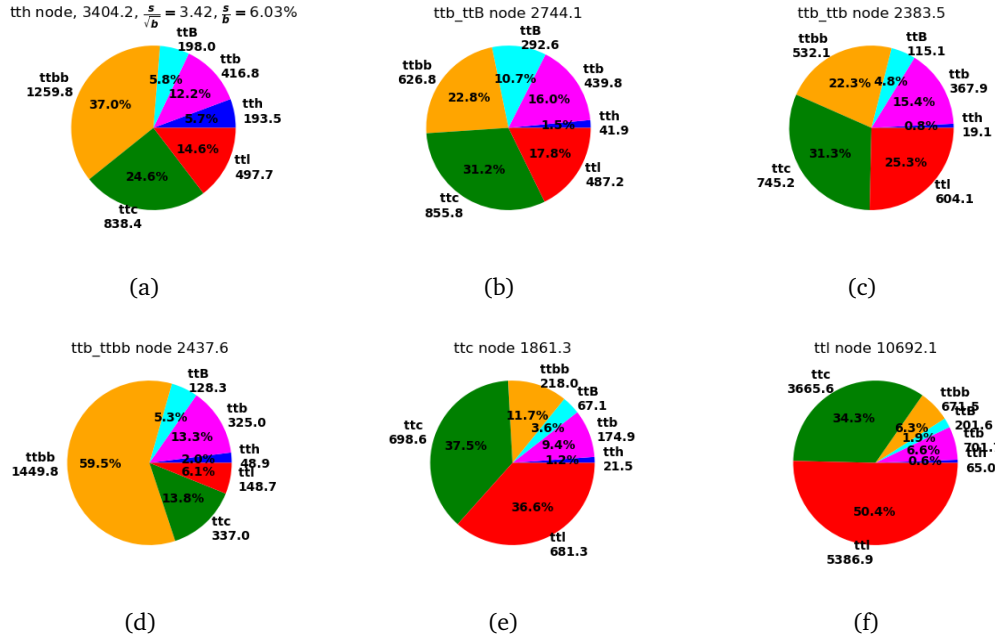


Figure 5.21: Fractional contributions of all background components and the $t\bar{t}H$ signal for all regions defined with deep learning multi-classifier with six categories, for events with ≥ 6 jets and $\geq 4b$ -jets tagged at 85%.

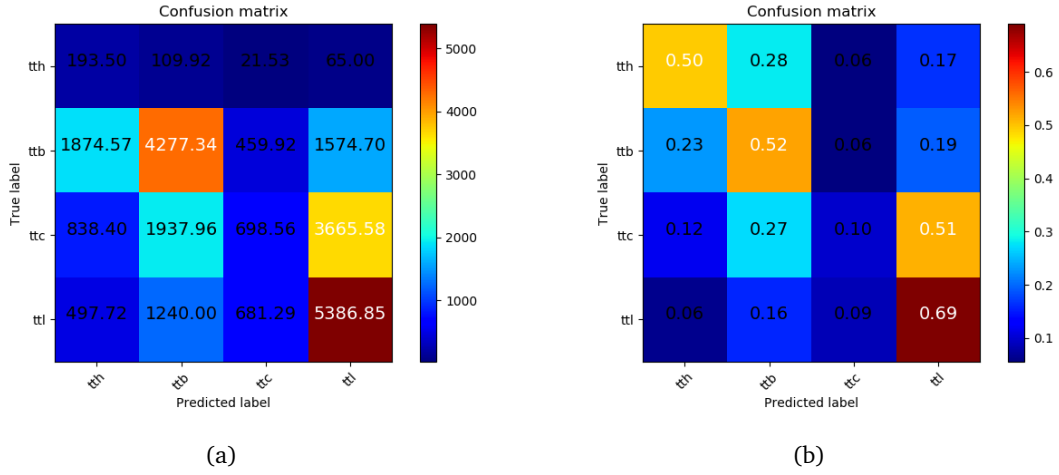


Figure 5.22: Confusion matrices of the multi-classifier with six categories with merged $t\bar{t} + \geq 1b$ sub-categories. $t\bar{t}H$ is denoted as $t\bar{t}h$, $t\bar{t} + \geq 1b$ as $t\bar{t}b$, $t\bar{t} + \geq 1c$ as $t\bar{t}c$, and $t\bar{t} + \text{light}$ is denoted as $t\bar{t}l$. (a) is the number of properly classified events for each category in truth, and (b) is the fraction of events that enter the proper category after the prediction of the classifier. (b) is obtained by normalizing each element in (a) by the sum of the corresponding row.

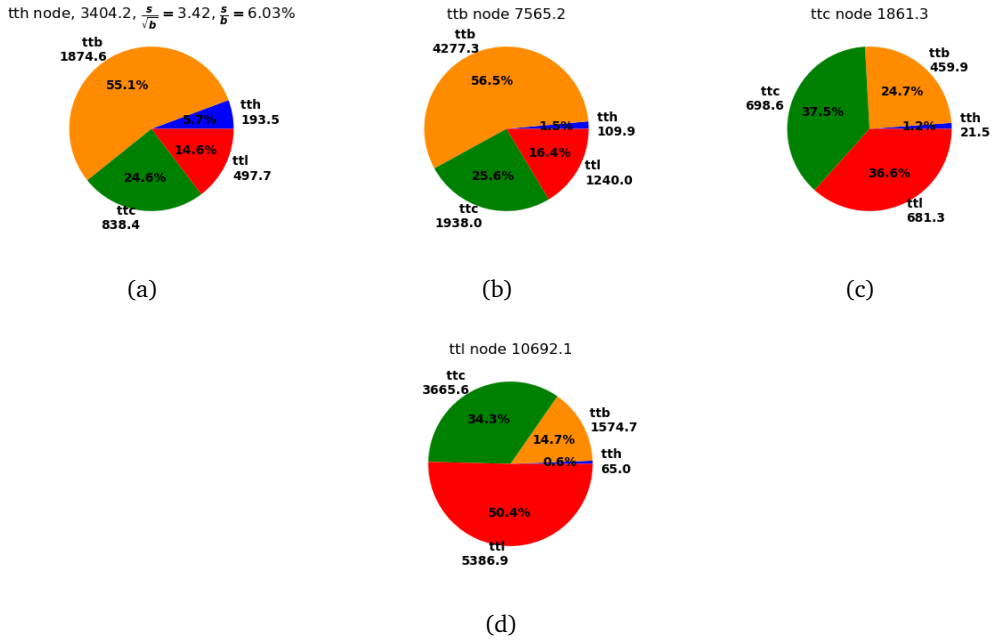


Figure 5.23: Fractional contributions of all background components and the $t\bar{t}H$ signal for all regions defined with deep learning multi-classifier with six categories, for events with ≥ 6 jets and $\geq 4b$ -jets tagged at 85%.

binary discriminant for $t\bar{t}H$ against the inclusive $t\bar{t}$ background. And in SR it performs as well as the binary classifier trained for $t\bar{t}H$ and $t\bar{t}$ separation.

The $t\bar{t} + \geq 1b$ sub-components and $t\bar{t}H$ signal are difficult to be separated by the DNN classifier, and also the separation between $t\bar{t} + \geq 1c$ and $t\bar{t} + \text{light}$. This is consistent with the previous definition. The finer categorization on $t\bar{t} + \geq 1b$ sub-components gives similar performance as the four-category classifier. Apart from extending to finer splitting, the DNN classifier is also convenient to exploit more features for categorization in the future, e.g., c -jets tagging variables. Potentially a purer $t\bar{t} + \geq 1c$ region can also be obtained by cutting on the output of the neuron corresponding to this component.

5.5 Low-level input features using deep neural networks

This section explores the usage of low-level input features of the dataset. Up to now, all deep learning models in this thesis use high level features that have been engineered and optimized. In order to possibly further improve the performance of the $t\bar{t}H$ -vs- $t\bar{t}$ binary classifier, the deep network is trained with low-level features, to learn useful information by itself. Contributions in this section includes:

- Investigating the ability of DNN to catch information from low level features
- Combining the low- and high-level features as inputs to DNN
- Incorporating the domain knowledge of the structure of the data into the neural network. The classical DNN shows limited ability to learn from low-level features directly. By adding physics domain knowledge into the neural network topology, it is expected to

learn from low-level features more efficiently, so as to learn more useful features than current engineered high-level ones.

- A parse tree model is designed that has topology in analogy with the $t\bar{t}H$ process. Events are embedded via this model for classification.
- Classification models are studied to use two parse trees having topologies analogous with $t\bar{t}H$ and $t\bar{t}$ processes, respectively. Two models are tested: direction concatenation, and Siamese training.

5.5.1 Using low-level features with classical networks

Manual calculation is a necessary step for popular machine learning methods in particle physics, like BDT and likelihood method. It is based on different expertise for different physics analyses, aiming to provide discriminating variables between signal and background. Lots of time is spent on the calculation and optimization of the input list, since the BDT performance is largely impacted by different input features. Up to now, all the machine learning models explored for the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis including the BDT and RNN are based on selected input features that are manually derived from the 4-vectors of the final state objects. RNN, using similar input features as the BDT, gives just slightly better AUC after hyper-parameter optimization.

This motivates the exploration of raw features as input using deep learning methods. The deep neural networks are expected to learn, from raw inputs, the discriminating features automatically, without heavy investigation on the choice of input features. Also, these self-learned features may help achieving even better classification performance, as the high-level feature engineering is not necessarily optimal, even with intense labor of optimization.

Several studies with raw inputs in particle physics have shown promising results compared with their state-of-the-art results using high level features. An early study [164] applied the DNN in a physics analysis, which aims to distinguish a cascading decay of new exotic particles from the dominant background. It compared the performance of shallow networks and BDT using physics-inspired high-level features and deep networks using low-level features. The results showed that deep networks using the low-level data significantly outperformed the shallow networks, meaning that the high-level features capture real insights but lose some useful information. Studies [165] are also implemented to use a DNN multi-classifier to classify different kinds of jets, i.e., to identify the jets that are initiated by different quarks. Traditional classification is based on hand-crafted features which are usually based on a subset of the particles, a subset of the information available for each particle, or only information from a single sub-detector. The new DNN classifier can exploit much more information which is enabled by including convolutional and LSTM layers. This new deep classifier outperforms significantly the standard classification approach.

DNN + low level features The ability of the classical deep neural network to extract information from low-level features is firstly investigated. The 4-vectors and the b -tagging weight of the final state objects, i.e. $\mathbf{o} = [p_x, p_y, p_z, E, b\text{-tagging}]$, are used as input features, consisting of 6 jets, the lepton, and the neutrino. The full neutrino 4-vector is built using the missing transverse energy with the W -boson mass constraint, as described in Section 4.5.3.2. b -tagging weights of leptons are set to zero. These objects are chosen by the best combination sorted by the reco_BDT, and other sub-leading combinations are not taken into account in this experiment.

The DNN structure is optimized using the TPE (illustrated in Section 5.2.3) method with 150 optimization runs. Table 5.5 lists all the searched hyper-parameters and their value variations.

The optimized DNN architecture is shown in Figure 5.24.

Parameter name	Values
Number of hidden layers	3,4,5,6
Number of neurons in 1 st and 2 nd (3 rd and 4 th , 5 th and 6 th) hidden layers	50, 100, 200, 300
Dropout rate in 1 st and 2 nd (3 rd and 4 th , 5 th and 6 th) hidden layer	Uniform(0,1)
Batch normalization in first three hidden layers varying independently	true, false

Table 5.5: Grid-searched hyper-parameters for the DNN. To reduce the number of parameter combinations, some parameters share the same value. The number of neurons in 1st and 2nd hidden layers are set to be identical. The same goes for 3rd and 4th, as well as 5th and 6th if the layers exist. Each of these pairs varies independently in the same range. The dropout rate variation follows a similar strategy.

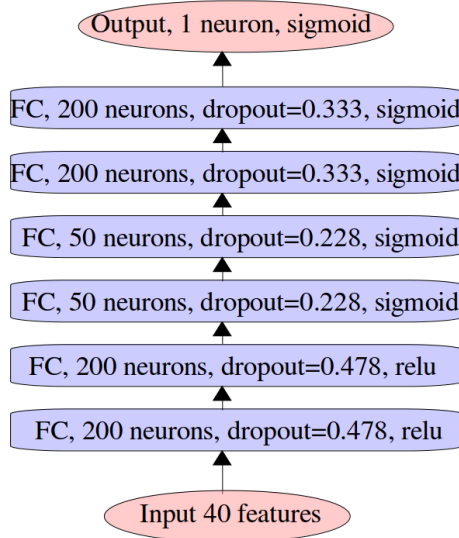


Figure 5.24: The optimized DNN architecture using TPE algorithm.

The optimized model AUC performance is shown as the red curve in Figure 5.25 where the AUC is 0.767 on the testing sample. This model is also re-trained with the high-level features including both the global kinematics and combination-dependent variables which are the same as inputs for the classification BDT shown in Table 4.5. The performance in terms of AUC is 0.787, presented as the blue curve, which is much better than that of the model using the low level features only, and is similar to the classification BDT in the Reference [25].

RNN + low level features To add the information of more than one combination, the RNN with the optimized architecture described in Section 5.3.3 is retrained with the low-level

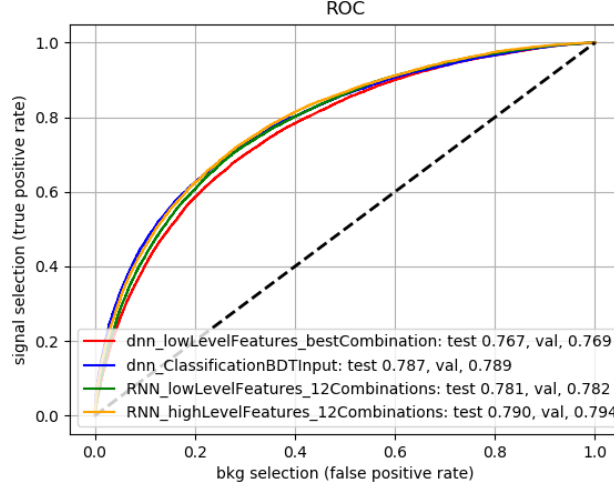


Figure 5.25: ROC curves of DNN and RNN models trained with low-level features including 4-vectors and b -tagging of final state objects, and with the high-level features as described in Section 5.3.1.2.

features $\mathcal{o}$ for the eight final-state objects as mentioned above. Again for each event 12 combinations are used and each of them gives six jets associated to b - and *light*-quarks. As shown in Figure 5.25, the AUC performance on the testing sample is 0.781 which is better than the DNN using exactly the same features. This performance is still worse than the RNN trained with high-level features as described in Section 5.3.3 which gives an AUC of 0.790.

Up to now using the raw input features with both the RNN and classical DNN does not bring better results than using directly the high-level features. It is dependent on the specifics of the classification task [166]. As pointed by Reference [164], their experiments on another set of data have similar performance whether using the deep or shallow networks with either low-level or engineered features. Deep neural networks are not necessary for a task with less complexity where the useful features are quite easy to extract, e.g., as linear functions of the low-level features. All information can be easily included in the simply engineered features, and there is no more information to be learned by the DNN. On the other hand, the conclusion also depends on how optimized the high-level features are, as a reference for the DNN. If the high-level features are carefully scrutinized and well optimized enough, it is possible that all information has been taken into account. Then the DNN can not bring more either. The currently used high level features for the BDT and the RNN are indeed optimized with lots of work (as presented in Section 4.5.4.1). However, it is possible that there is still more information to gain from the dataset of the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis, and the classical DNN trained in this section has difficulties to learn discriminating features from the raw information. Therefore, more deep learning models are studied in the following sections.

5.5.2 Parse tree model: encoding domain knowledge into neural networks

Deep neural network models are usually good at processing specific input samples. For example, the convolutional neural network is designed to work with data that has a spatial structure, like images. The RNN works well with sequence data processing. One technique used for the natural language processing (NLP) is parse tree [140], or grammar tree which represents the

syntactic structure of a sentence. Each word in the sentence corresponds to a leaf node in the tree. Going from the leaves to the root node, the full sentence can be embedded into a vector space. By construction, the design of these models takes the data structure into account.

Quite a few studies in particle physics have started to build physics-aware deep learning models for their tasks. The work on jet physics [167] provides an idea using the analogy between the jet and natural language for jets classification. Each particle in the jet is like a word in the sentence, and the jet is identified in the detector by clustering many particles. Therefore a parse tree for jet embedding can be built, where each particle is represented by a tree node and the tree topology reflects how particles are clustered. Similar ideas also exist in the study for quark/gluon tagging [168], jet charge reconstruction [169], and boosted top quark tagging [170].

Inspired by the above mentioned references, the idea in this study is based on the analogy between the physics process represented as Feynman diagram and the binary tree model. The $t\bar{t}H$ ($H \rightarrow b\bar{b}$) process is incorporated into the tree topology. Figure 5.26(a) shows a tree developed following the tree-level t -channel $t\bar{t}H$ production Feynman diagram as shown in Figure 5.26(b). The model exploits the raw level features of all the final-state particles including two leptons and six jets: momentum, energy and jets' b-tagging weight. Going through from the leaves until the root node which is an n-dimensional vector, the low-level information is embedded into a new space.

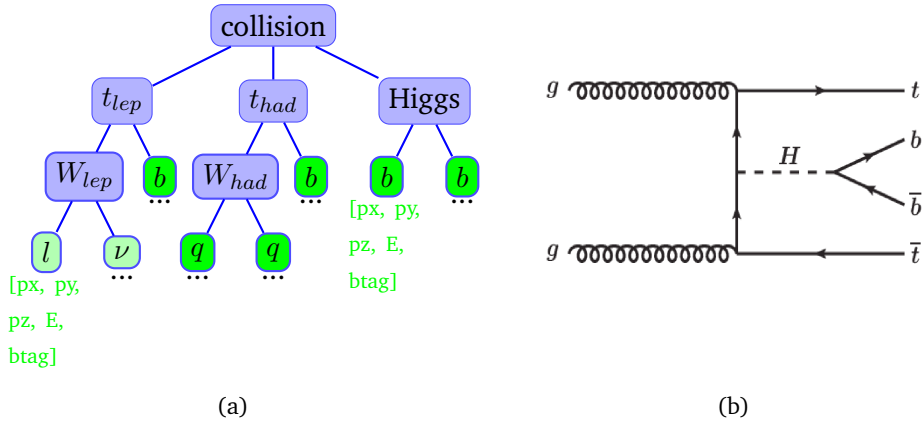


Figure 5.26: A parse tree model (a) developed following the tree-level t -channel $t\bar{t}H$ ($H \rightarrow b\bar{b}$) production Feynman diagram (b).

As input, each final-state object (leaf node) is represented by the 4-momentum vector concatenated with b -tagging, $\mathbf{o}_l = [p_x, p_y, p_z, E, b\text{-tagging}]$, the b -tagging of leptons being set to zero. Each of the leaves (l) is then embedded into a new space of dimension q , via a neural network layer (referred to as transfer layer):

$$\mathbf{u}_l = \sigma(W_u g(\mathbf{o}_l) + b_u) \quad (5.17)$$

where $W_u \in \mathbb{R}^{q \times 5}$ and $b_u \in \mathbb{R}^q$ are the connection weights of the transfer layer which are shared for all the nodes. $g(\mathbf{o})$ is a function extracting the kinematic features p , η , θ , ϕ , E and p_T from $\mathbf{o}$, and is realized by a lambda layer, a layer object in keras that wraps arbitrary expression of an input tensor. σ denotes the rectified linear unit (ReLU).

Each intermediate node (k)'s embedding ($\mathbf{h}_k$) is:

$$\mathbf{h}_k = \sigma \left(W_{h_k} \begin{bmatrix} \mathbf{h}_{k_L} \\ \mathbf{h}_{k_R} \\ \mathbf{u}_k \end{bmatrix} + b_{h_k} \right) \quad (5.18)$$

$$\mathbf{u}_k = \sigma(W_u g(\mathbf{o}_k) + b_u) \quad (5.19)$$

$$\mathbf{o}_k = \mathbf{o}_{k_L} + \mathbf{o}_{k_R} \quad (5.20)$$

where $W_{h_k} \in \mathbb{R}^{q \times 3q}$ and $b_{h_k} \in \mathbb{R}^q$ are different neural network layers (referred to as embedding layers) for each intermediate node k . $\mathbf{h}_{k_L}$ ($\mathbf{h}_{k_R}$) is the embedding of the hidden node k 's left (right) child, and it becomes $\mathbf{u}_l$ (Equation 5.17) if it is a leaf. The $\mathbf{o}_k$ transferring to $\mathbf{u}_k$ is the vector sum of the raw-features of the node k 's two children. In this way, the interesting collision is represented by the re-clustering of the final-state objects following the physics process, adding up momentum and b -tagging weights of all final-state objects.

5.5.3 Using single tree to classify $t\bar{t}H$ and $t\bar{t}$ with one combination

The tree model is firstly tested on a simplified task, where only the best combination according to the reconstruction BDT is obtained for $t\bar{t}H$ as signal and $t\bar{t}$ as background. This is referred to as best-vs-best tree in the following. To do the classification, the embedding representation from the tree is fed into a fully-connected network which has an output layer with one neuron using the sigmoid activation function. The model is trained with the binary cross-entropy loss function using the adagrad optimizer.

Model architecture

The model parameter is grid-searched using 3-fold cross-validation where only even-events are used split into training (80%) and validation (20%) samples. The searched hyper-parameters and their value variations are listed in Table 5.6.

Parameter name	Values
Embedding space dimension	30 , 50, 100
Batch normalization after lambda layer	True , False
Batch normalization after embedding layer	True , False
number of layers for the embedding layers	1 , 2
dropout rate of the embedding layers	0, 0.01, 0.2
learning rate of adagrad	0.01, 0.001 , 0.05
ϵ of adagrad	1e-07 , 1e-08, 2e-08

Table 5.6: Grid-searched hyper-parameters for the best-vs-best tree. The optimized values are in bold.

The embedding space dimension q is optimized to be 30. The batch normalization is applied both after the lambda and the embedding layers, the latter having the dropout rate optimized

as 0.2 and activation function fixed as ReLU. The adagrad is fixed as optimizer for which the learning rate and epsilon value is grid-searched to be 0.001 and 1e-07 respectively. The fully connected layer is fixed to be a single structure with one hidden layer having 30 neurons with ReLU activation function, to facilitate the comparison with other models later. The AUC evolution during the training of this tree model is shown in Figure 5.27(a).

This tree model, supposed to be $t\bar{t}H$ -like, will be referred to as the *signal tree* in the best-vs-best experiment in the following sections. Its architecture and corresponding weights are saved to be used as initial values of the trees which are nested in complicated models described in the following sections.

Performance

The signal tree model is trained to classify the best combination of $t\bar{t}H$ signal and $t\bar{t}$ background, and gives an AUC of 0.781. The ROC curve of this model is presented in Figure 5.27(b). It also shows the comparison with the classical feed-forward DNN (AUC of 0.772) (discussed in Section 5.5.1) trained with the same signal and background as well as low-level input features. By using the tree model, the AUC performance is improved by 1.2%, while much fewer network connection weights (19,037) are required than for the DNN (112,501). The BDT performance is still better than both neural network models, but it is based on intermediate MVAs that actually use all combinations.

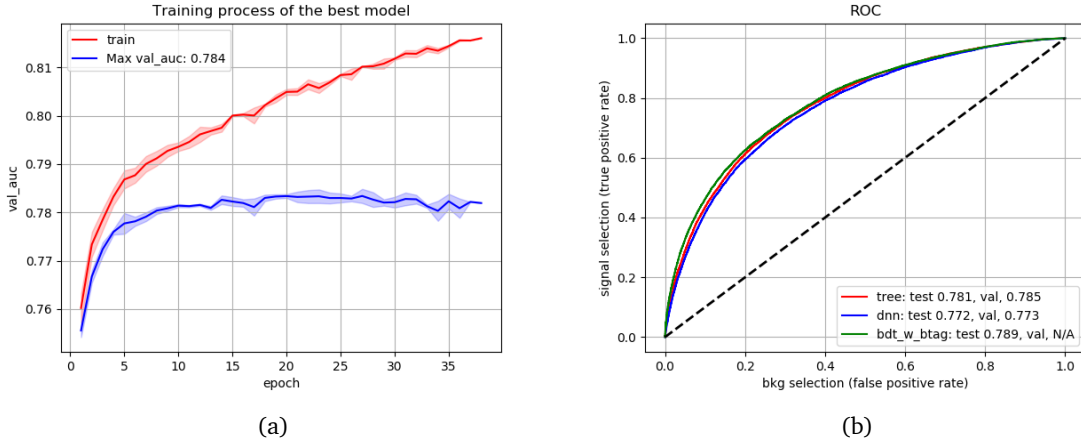


Figure 5.27: The tree model performance. (a): the AUC evolution with respect to the training epoch for both training and validation sets. (b): the ROC curves of the tree model, the DNN model, and the classification BDT. Shown in the legend are the AUCs on both test and validation samples.

Visualizing the event separation embedded by the trees

To visualize the signal and background separation in the high-dimensional embedding space, the technique called t-distributed Stochastic Neighbor Embedding (t-SNE) [171] is used. This technique aims to take a set of points in high-dimensional space and find a faithful representation of those points in a lower-dimensional space, typically the 2D plane.

Using this method, the signal and background representations after being embedded by the signal tree are shown in Figure 5.28. One tuneable parameter of this method is “perplexity” which represents how to balance attention between local and global aspect of the data. The performance of resulting projections could be affected by this parameter, although it is said to be robust to the perplexity ranging from 5 to 50 [171]. In Figure 5.28, for each tree, the projections on 2D are shown with perplexity varying from 2 to 100, and it shows that the patterns start to be stable with perplexity larger than 30. It shows that the two classes are indeed projected into relatively separated representations in the embedded space.

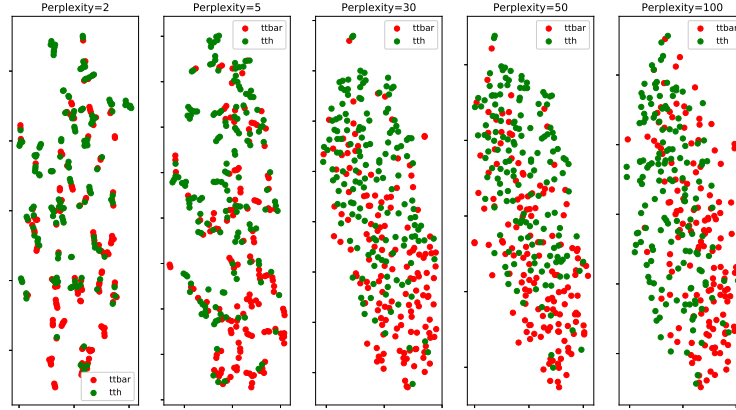


Figure 5.28: The visualization of the (a) signal tree embedding in 2-dimension reduced from 30-dimension using t-SNE algorithm.

5.5.4 Using single tree embedding to classify $t\bar{t}H$ and $t\bar{t}$ with all combinations

To employ more combinations, the tree model is combined with a recurrent neural network. As with the plain RNN, only the 12 best combinations are used, as sorted by the reconstruction BDT. The overall model structure is shown in Figure 5.29(a), where combinations of each event are embedded by the tree and then form a sequence for the RNN. Following the RNN are the fully connected feed-forward layers (the classifier module in the figure) used to further classify the $t\bar{t}H$ and $t\bar{t}$ events. The tree part has the same architecture as the best-vs-best tree (see Section 5.5.3) from which the connection weight of the new tree is initialized. Then during the training of the full model, the embedding tree is shared by all combinations and its weights are updated from the initial values.

Model architecture A hand tuning of the hyper-parameters is implemented. The GRU layer is found to perform better having 30 neurons in it which are then connected to four FC layers having individually 60, 100, 60 and 30 neurons. For each FC layer, a dropout layer with a rate of 0.2 is applied, and the activation function is ReLU. Batch normalization is applied after each of the FC layers. The model has an output layer of one neuron with sigmoid activation and is trained with the binary cross-entropy loss function and the Adam optimizer with default setting of Keras. The best AUC performance of the tuned model with testing sample (odd events) is 0.788.

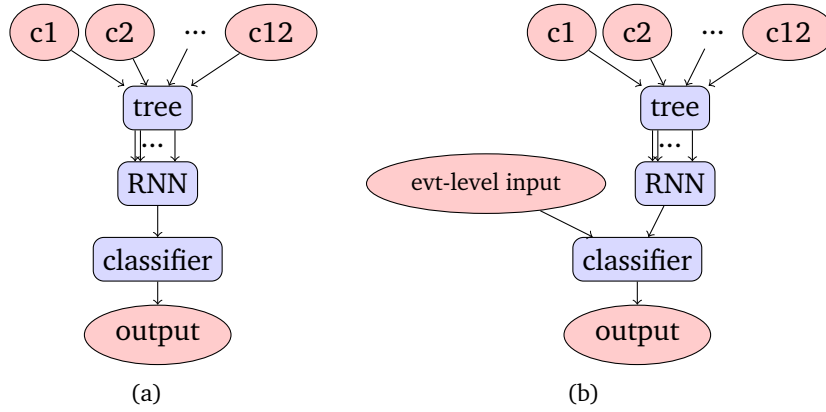


Figure 5.29: Classifiers of $t\bar{t}H$ against $t\bar{t}$ using recurrent neural networks and the parse tree embedding. The 12 combinations are represented as pink ellipses denoted as $c_1, c_2, \dots, c_{12}$ which are embedded via the same tree into 12 new combinations that comprise a sequence event into the RNN. Model (b) has the same architecture (but separately trained ending up with different connection weight) as (a), except that event-level kinematics are fed as another set of inputs concatenated with the output of the RNN into the classifier layer.

In addition to the low-level features, the event-level kinematics can also be exploited as inputs to the model shown in Figure 5.29(b) where the additional global variables are concatenated with the RNN layers and fed into the FC layer (the classifier module in the figure) while the architecture of the other parts are the same. The added kinematics are the same ones as listed in the first section of Table 4.5. The model is retrained with the tree initialized in the same way as before.

DNN embedding As a baseline to be compared to the tree embedding, another model is also trained, which is similar to the previous tree-based model in Figure 5.29(a) but the tree embedding is replaced with the DNN architecture that is optimized in the best-vs-best experiment in Section 5.5.1. The whole model is trained, of which the connection weights of the DNN part are initialized with the DNN model trained in the best-vs-best experiment. Analogously to the tree model in Figure 5.29(b), the event-level kinematics are also added for the DNN embedding and the whole network is retrained.

Performance The AUC results of all the above mentioned RNN models using tree or DNN embedding are compared in Figure 5.30. Together with the RNN, the tree model using low level features alone gives an AUC of 0.788, and adding the global kinematic features gives an AUC of 0.789, as good as the classification BDT, but with almost no gain compared to low level alone. The model with DNN embedding gives worse AUC both using low level features alone (0.776) or having in addition the global kinematics (0.783). Its performance is more improved having the global kinematics in addition compared with the tree model.

These results show that the tree structure is helpful to learn information efficiently from the low level input features.

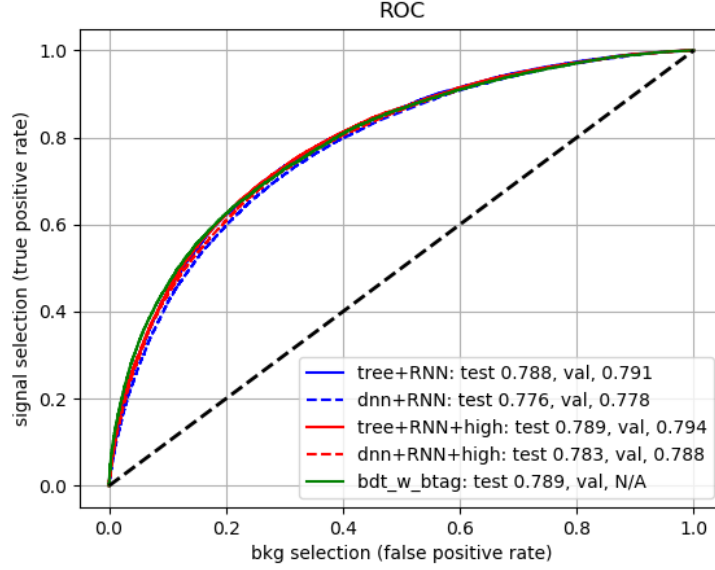


Figure 5.30: The ROC curves of various models on the testing sample. The state-of-the-art classification BDT is in green as reference. The other solid (dashed) curves are from the RNN models using tree embedding (DNN embedding). The red (blue) curves are the results using (without using) the event-level kinematics as input.

5.5.5 Two-tree model

Up to now, only the signal-like tree shown in Figure 5.26(a) is used to provide the embedded representations of both $t\bar{t}H$ signal and $t\bar{t}$ background to the FC classifier. However, the $t\bar{t}$ event representation may be biased to be $t\bar{t}H$ -like after being embedded by the signal-like tree. Thus this section explores approaches to use two trees to do the binary classification of $t\bar{t}H$ signal and $t\bar{t}$ background. Only the best combination is used for both the signal and the background, to make the task easier without being entangled with the recurrent structure.

Tree mutations The tree topology can be mutated to make it more signal-like or background-like. Four different tree structures are tested where either one top quark and Higgs or two top quarks are merged into one node before going to the root node, as presented in Figures 5.31 and 5.32. Together with the topology shown in Figure 5.26(a), the signal-like tree, these four trees are individually stacked to a one-layer classifier containing 30 neurons, and are trained individually. These trees give highly superimposed ROC curves and similar AUC values, although no grid search is done for all the mutations. It seems that providing a signal-like embedding or a background-like embedding leads to the same performance.

5.5.5.1 Signal tree vs. background tree

The tree topology in Figure 5.32(a) (referred as $t\bar{t} + b\bar{b}$ -like tree) is supposed to be more like the Feynman diagram of background $t\bar{t} + b\bar{b}$ in Figure 5.32(b) as the top-quark pair is merged directly into one intermediate node. Stacked with a classification layer of the same architecture, the correctly classified signal and background events of both the $t\bar{t} + b\bar{b}$ -like tree and the signal

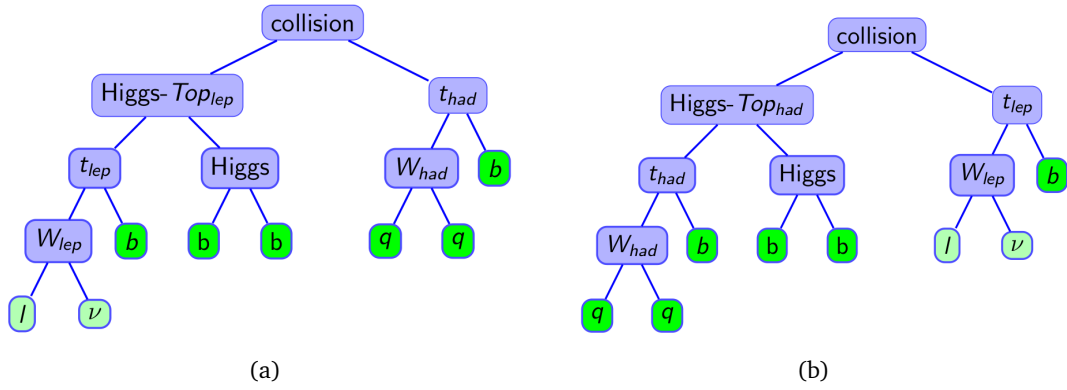


Figure 5.31: Tree topology mutations. Before going to the root node, the (a) Higgs and leptonic top candidates, and (b) Higgs and hadronic top candidates are merged first.

like tree are listed in Table 5.7. It shows that the $t\bar{t}H$ events are more often correctly labeled by the classification model using the signal-like tree embedding, with an efficiency of 0.774, higher than the efficiency of 0.761 of the model using the $t\bar{t} + b\bar{b}$ -like tree embedding. On the other hand, the correct-labeling efficiency of the $t\bar{t}$ events is higher for the $t\bar{t} + b\bar{b}$ -like tree (0.588) compared to the signal-like tree (0.570). These show that the tree with a specific topology is helpful to extract events of the corresponding category, although the performance improvement is small.

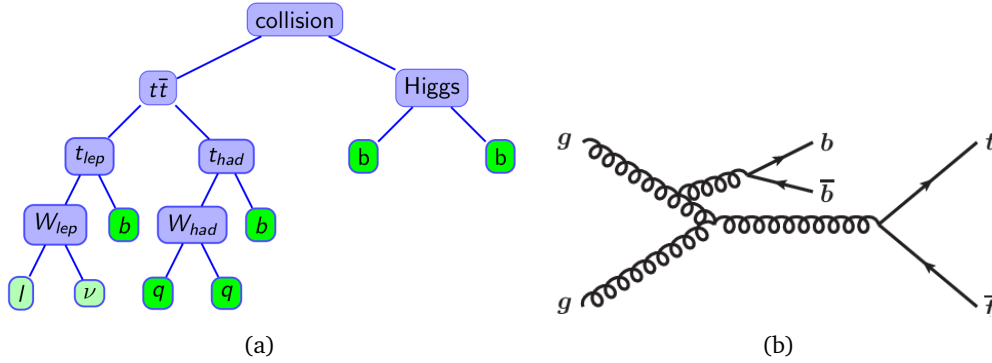


Figure 5.32: A parse tree model (a) designed following the $t\bar{t} + b\bar{b}$ Feynman diagram (b).

Comparing the second and third columns in each table, the events correctly labeled by the two models are highly overlapped. Among the correctly labeled signal events by the signal-like tree, around 95% of them are also properly labeled by the $t\bar{t} + b\bar{b}$ -like tree. And around 91% of the correctly labeled $t\bar{t}$ events by the $t\bar{t} + b\bar{b}$ -like tree are also properly classified by the signal-like tree.

5.5.5.2 Combination of two trees by brute force

A simple way to employ the embedding of both the signal and $t\bar{t} + b\bar{b}$ -trees is to feed the concatenation of two trees' output into one classifier as sketched in Figure 5.33. The motivation is to check whether complementary discriminating information is extracted in the embedded space via the $t\bar{t} + b\bar{b}$ -tree, in addition to the signal tree. Compared with the single tree model

Model	correctly labeled	correctly labeled by both
signal tree	135102 (.774)	128075 (.734)
$t\bar{t} + b\bar{b}$ -like tree	132714 (.761)	

(a) signal-tree vs. $t\bar{t} + b\bar{b}$ -like tree, signal events

Model	correctly labeled	correctly labeled by both
signal tree	33242 (.570)	31355 (.537)
$t\bar{t} + b\bar{b}$ -like tree	34283 (.588)	

(b) signal tree vs. $t\bar{t} + b\bar{b}$ -like tree, background events

Table 5.7: Number of (a) $t\bar{t}H$ signal events and (b) $t\bar{t}$ background events that are correctly labeled by signal-like or $t\bar{t} + b\bar{b}$ -like tree model, and by both models.

described in Section 5.5.3, the FC classifier part of this model has more neurons and layers, in order to learn from the embedding of two trees.

Three ways of training two-tree models are tested and the results are listed in Table 5.8: the trees are fixed to their initial weights which are the results of best-vs-best training, while only the FCs are trainable; the trees are initialized from the best-vs-best results and are made trainable as well; the whole network is trained from scratch without tree weights being initialized. The two models with initialization give better AUC than the model trained from scratch. The model with all trees fixed during training gives marginally better AUC among all two-tree models and single-tree models, thus the following two-tree models will be trained this way.

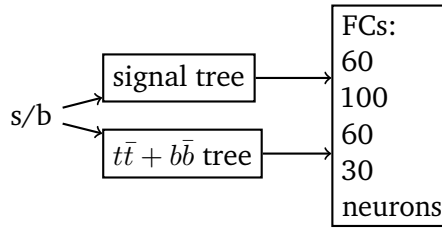


Figure 5.33: The classification model architecture combining two trees. Through the model, each event is embedded by two trees into new representations which are then concatenated and fed into a classifier consisting of fully-connected layers.

Models	test AUC	val. AUC
single tree + 1 FC		
signal tree	0.781	0.785
$t\bar{t} + b\bar{b}$ tree	0.784	0.787
two tree + FCs		
trees fixed	0.785	0.788
trees trainable	0.784	0.786
all from scratch	0.782	0.786

Table 5.8: The AUC performance of the single-tree models, and of the two-tree models where the classifier is trained with the trees either fixed, initialized or not as detailed in the text.

Although the two-tree model gives slightly better AUC for the best-vs-best task, no more improvement is observed when nesting the two-tree embedding into the RNN for the 12-combination experiment as the single-tree+RNN model. Among the two training strategies with initialization as mentioned above, the model with two trees fixed during the training gives the highest AUC of 0.788 on testing sample, while the trainable-tree model gives 0.784.

5.5.5.3 Siamese training

Each tree topology is desired to be more efficient to learn a specific class giving discriminant representations for the signal and background categories. At the same time, for each class, the representations from the two trees should be less correlated, providing non-redundant features for the classifier. To achieve these goals, the method explored in this section is the so-called Siamese training [172]. The architecture of such a model includes three parts as shown in Figure 5.34(a). Two of them are of the same signal tree topology which are the same tree and share weights. The two trees are fed in with a dataset consisting of event pairs which could be homogeneous containing two different signal events, or heterogeneous having one signal and one background event. The embedding of the two events via the twin trees is fed into a third module which calculates the distance between the events in the new space. The embedded representations of events from the same class are trained to be closer while the events from different classes should be far away from each other.

Two-phase training The two-phase Siamese training strategy is sketched in Figure 5.34 where in addition to Figure 5.34(a) as illustrated above, another Siamese network is also built similarly but replacing the signal tree with the $t\bar{t} + b\bar{b}$ tree as shown in Figure 5.34(b). Correspondingly, the dataset for this phase is made of event pairs which are homogeneous having two different background events or heterogeneous containing one background and one signal event. These two training phases are processed iteratively, where each phase is trained on a batch of corresponding dataset for each iteration.

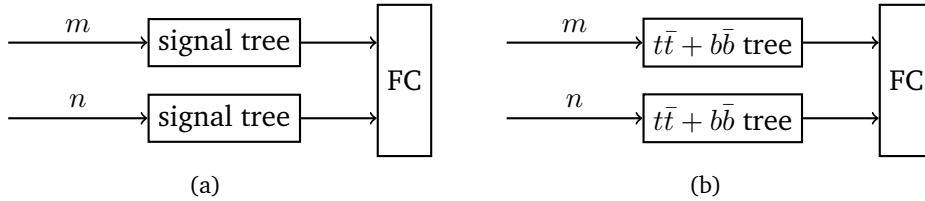


Figure 5.34: Sketches of the two-phase Siamese training. In each iteration of batch training, phase (a) is firstly processed to train the signal tree model and then phase (b) to train the $t\bar{t} + b\bar{b}$ tree. In both phases, the FC module here has no extra connection weights and is used to calculate the L1 distance in the embedding space between the two events m and n .

The distance between the two events in the embedding space is the L1 distance:

$$p(x_{c1}^{(m)}, x_{c2}^{(n)}) = \sum_{l=1}^{l=d} |h^l(x_{c1}^{(m)}) - h^l(x_{c2}^{(n)})|. \quad (5.21)$$

Here h denotes the shared weights of the twin trees and d is the embedding dimension.

$(x_{c1}^{(m)}, x_{c2}^{(n)})$ is the input event pair consisting of event m in class $c1$ and event n in class $c2$ where $m \neq n$ when $c1$ and $c2$ are the same class. $c1$ and $c2$ can be signal (S) or background (B).

The model is trained to minimize the loss function designed as $\mathcal{L} = \mathcal{L}_{\text{sig_tree}} + \mathcal{L}_{\text{bkg_tree}}$ where the first term $\mathcal{L}_{\text{sig_tree}}$ (Equation 5.22) is minimized during the phase (a) training, and the second term $\mathcal{L}_{\text{bkg_tree}}$ (Equation 5.23) minimized during phase (b):

$$\begin{aligned} \mathcal{L}_{\text{sig_tree}} = & \sum_{i=1}^N y^i(x_S^{(m)}, x_S^{(n)}) \log p^i(x_S^{(m)}, x_S^{(n)}) \\ & + f_S \sum_{j=1}^N y^j(x_S^{(m)}, x_B^{(n)}) \log p^j(x_S^{(m)}, x_B^{(n)}), \end{aligned} \quad (5.22)$$

$$\begin{aligned} \mathcal{L}_{\text{bkg_tree}} = & \sum_{i=1}^N y^i(x_B^{(m)}, x_B^{(n)}) \log p^i(x_B^{(m)}, x_B^{(n)}) \\ & + f_B \sum_{j=1}^N y^j(x_B^{(m)}, x_S^{(n)}) \log p^j(x_B^{(m)}, x_S^{(n)}). \end{aligned} \quad (5.23)$$

$y^i(x_{c1}^{(m)}, x_{c2}^{(n)})$ is the truth label of the input pair. It is labeled 1 if $c1$ and $c2$ are the same class, and -1 if $c1$ and $c2$ are different classes. f_S and f_B are extra factors weighting hetero-class pairs for the signal-tree and $t\bar{t} + b\bar{b}$ -tree models, respectively. A batch of signal events with size $2N$ and of background events with the same size are first sampled separately for each training iteration. Then they are paired to form N inputs for the Siamese models, consisting of $(x_S^{(m)}, x_S^{(n)})$, $(x_S^{(m)}, x_B^{(n)})$, $(x_B^{(m)}, x_S^{(n)})$ and $(x_B^{(m)}, x_B^{(n)})$ where $m \neq n$ if the two elements are from the same class. The sample weight of each event pair is the product of the sample weights of its two elements. The four types of event pairs are re-weighted to have equal weight. so that they are treated equally during training. This is realized by requiring the $t\bar{t}H$ signal and $t\bar{t}$ background in the original training set to be balanced beforehand. Firstly the background events (whose size is one third of the signal sample one) are duplicated to be of the same size as signals. Then the sum of the two classes' sample weights are scaled to be the same. The maximum number of Siamese training iterations is set to 50 while the training is early stopped when the validation loss of both phases does not decrease in 5 epochs.

Such Siamese-trained signal and $t\bar{t} + b\bar{b}$ -like trees are subsequently used for binary classification. To do this, a compound neural network is trained as a classifier, feeding the concatenation of the two trees' embedding into four fully-connected layers consisting of 60, 100, 60 and 30 neurons respectively. The ReLU activation function and dropout layers with a rate of 0.4 are applied for each of these hidden layers. The tree model weights can be frozen to the Siamese-trained result, or allowed to be trainable during the training of this classifier.

The values of additional factors f_S and f_B are grid-searched with 0.05, 0.5, 1, 5, 30 and 100. The AUC performance of all these experiments are shown in Figure 5.35 for the whole classifier trained with trees frozen to the Siamese results as well as the one with trees allowed to be trainable as well. For all the models with various values of additional factors, the performance can be recovered using a flexible tree during the training, although the model with a frozen Siamese-trained tree gives the best AUC on both validation and test sets when both f_S and f_B are equal to 1. The results of the same classification experiments but starting from the classifier based on the concatenation of the two trees' embedding without Siamese training are also

shown in the plots. Compared to this normal training, the Siamese trained embedding does not bring better results in this binary classification.

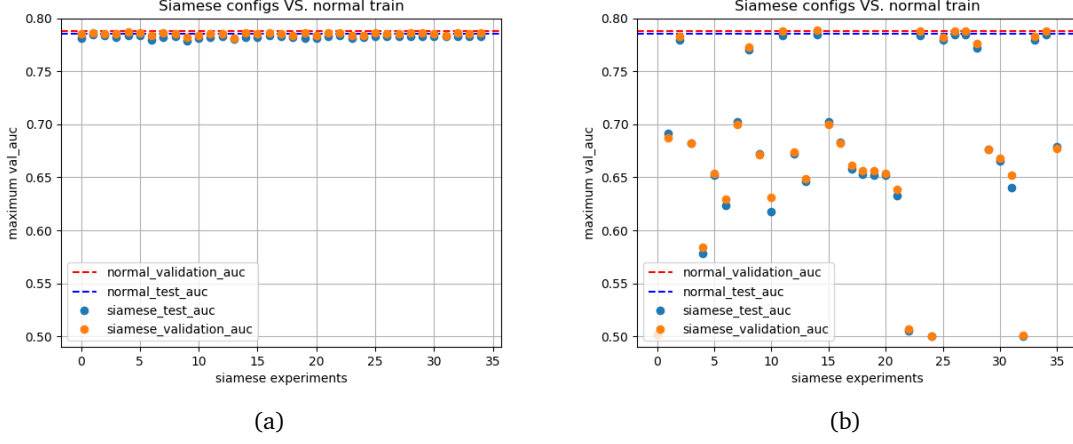


Figure 5.35: The classifier performance using the Siamese-trained trees which are grid-searched varying f_S and f_B . The x -axis is the index of experiments with different f_S and f_B combinations, and y -axis is the AUC performance on the validation sample. (a) is the results of the classifier trained together with the trees whose weights are initialized with results of the Siamese training, and (b) is similar results but with trees frozen to their Siamese-trained results. The classification performance using the concatenation of the two trees without Siamese training is also shown as dashed lines.

The evolution of the loss function (Equations 5.22 and 5.23) and the averaged L1 distance (Equation 5.21) of paired events during the Siamese training are shown in Figure 5.36, for the Siamese-trained model ($f_S = f_B = 1$). This model is picked because the compound binary classifier using this embedding gives the best AUC in the grid-search above. The validation loss of the signal ($t\bar{t} + b\bar{b}$ -like) tree model stops decreasing at epoch 8 (6) and the models at that stage are obtained for the binary classification, while the training loss of the $t\bar{t} + b\bar{b}$ tree model continuously decreases. For the averaged L1 distance in the embedding space of the signal tree, the one from the homo-class pairs decreases by epoch 8, while the one from hetero-class pairs is fluctuating and does not show a clear trend. For the $t\bar{t} + b\bar{b}$ tree embedding, the distance is decreasing for both the homo-class and hetero-class pairs although the latter is expected to increase. Therefore, the Siamese training does not evolve and converge as expected, and this could explain why it does not bring the improvement in the performance of the subsequent binary classifier.

Three-phase training The Siamese experiment in this section adds one more training phase in each iteration as shown in Figure 5.37. Figures 5.37(a) and 5.37(b) are the same training phases as employed for the 2-phase experiment. Figure 5.37(c) is the new added training phase where fed into the distance calculator layer are the signal tree and the $t\bar{t} + b\bar{b}$ tree that share the weights with those trees in the other two phases. Corresponding to this new phase, the loss function $\mathcal{L} = \mathcal{L}_{\text{sig_tree}} + \mathcal{L}_{\text{bkg_tree}} + \mathcal{L}_{\text{two_trees}}$ has a third term which is:

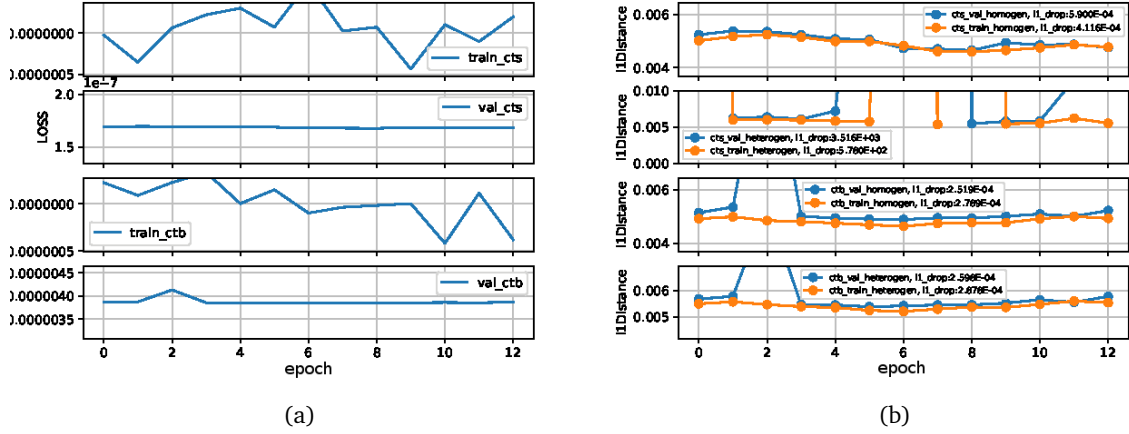


Figure 5.36: The evolution of the loss function in (a) and the averaged L1 distance between paired events in (b) for both the signal and $t\bar{t} + b\bar{b}$ -like trees. The upper two pads are for the signal tree, while the lower two for the $t\bar{t} + b\bar{b}$ -like tree.

$$\mathcal{L}_{\text{two_trees}} = f_{SB} \sum_{i=1}^N -1 \cdot \log p^i(x_S^{(m')}, x_B^{(n')}) \quad (5.24)$$

where the event pair $(x_S^{(m')}, x_B^{(n')})$ is always hetero-class with signal entering the signal tree and background entering the $t\bar{t} + b\bar{b}$ tree. In each iteration of batch sampling, this type of pairs are also cooked from the sampled batch of signal and background events, the event overlaps between $(x_S^{(m')}, x_B^{(n')})$ and each of $(x_S^{(m)}, x_S^{(n)})$, $(x_S^{(m)}, x_B^{(n)})$, $(x_B^{(m)}, x_S^{(n)})$ and $(x_B^{(m)}, x_B^{(n)})$ are controlled to be 50%. f_{SB} is an additional factor giving more or less importance to this new loss term, and is grid-searched together with f_B and f_S .

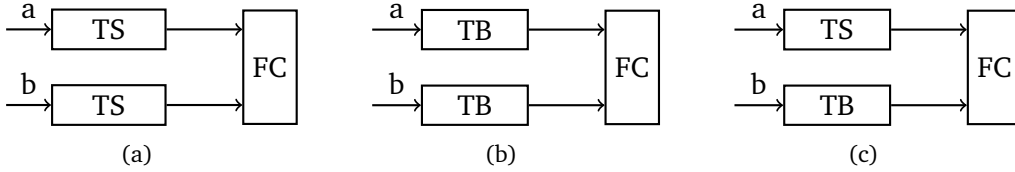


Figure 5.37: Sketches of the three-phase Siamese training. In each iteration of batch training, phase (a) is firstly processed to train the signal tree (TS) model, then is phase (b) to train the $t\bar{t} + b\bar{b}$ tree (TB), and in the end the phase (c) which mixes TS and TB. In all phases, the FC module has no connection weights and is used to calculate the L1 distance in the embedding space between the two events. The signal and the $t\bar{t} + b\bar{b}$ -like trees are shared among all phases.

Similarly to Section 5.5.5.3 for the 2-phase training, these Siamese-trained embeddings are used for a subsequent classifier to separate $t\bar{t}H$ and $t\bar{t}$ events. The embeddings are concatenated as input to fully-connected layers having the same structure as for the two-phase scenario. The performance of binary classifiers that use the Siamese-trained models with various f_S , f_B and f_{SB} configurations is shown in Figure 5.38. Only the experiments of the classifiers trained with

the fixed trees are implemented. It shows that using the Siamese-trained trees does not bring improvement compared to the classifiers using normally-trained trees.

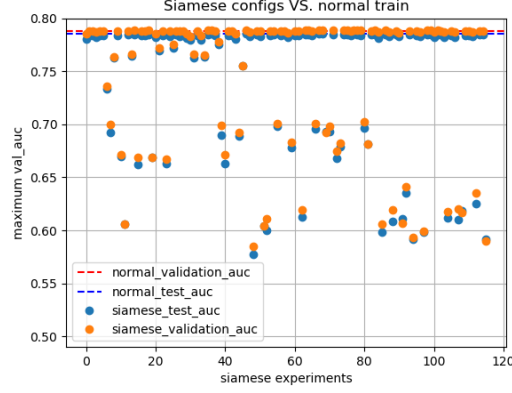


Figure 5.38: The classifier performance using the Siamese-trained trees which are grid-searched varying f_S , f_B and f_{SB} . The trees are frozen to their Siamese-trained results during the training. The x -axis is the index of experiments with different f_S , f_B and f_{SB} combinations, and y -axis is the AUC performance on the validation sample. The classification performance using the concatenation of the two trees without Siamese training is also shown as dashed lines.

5.5.6 Summary

This section explores the usage of low-level features as input to deep learning models, aiming to further improve the performance of the binary classifier to separate the $t\bar{t}H$ signal from the $t\bar{t}$ background. The classical DNN and the RNN are found difficult to learn enough discriminating information from the low level features, and give worse performance than the classification BDT designed in Chapter 4.5.4 or the RNN (Section 5.3) trained with high-level features. The classical DNN's performance can be recovered to be as good as (but no better than) the BDT if both the low- and high-level features are exploited.

In order to learn enough information more efficiently from low-level features, a parse tree model is designed, which has a topology analogous with the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) signal process. Each event, from its low-level feature space made of four-vectors and b -tagging weights, is embedded through the tree until the root node denoting the proton-pair collision which is represented by a neural network layer. The embedded representation is then used for the classification combined with the RNN to exploit multiple jet-parton combinations. This tree-based binary classifier is shown to be able to learn more efficiently than the classical DNN. It gives as good performance as the classification BDT that exploits high-level features. In turn, this shows that the classification BDT are well optimized.

Considering that background events may be biased to be signal-like when embedded by the signal tree, a $t\bar{t}$ + jets process-like tree is also considered for the embedding of the background. Two approaches of exploiting two trees are tested, including direct concatenation and Siamese training. Although the background tree alone is shown to be slightly more sensitive for the background classification accuracy, the combination of the two-tree embedding does not bring improvements. This does not preclude a more positive outcome with other physics analysis in

which signal and background have more different decay chains (thus a more different Feynman diagram topology, leading to more different embedding trees).

5.6 Reducing systematic uncertainties using adversarial training

In Reference [25], the BDT is trained on the PP8 (using the POWHEG and PYTHIA 8 for simulation) nominal sample only as a classifier for $t\bar{t}H$ and $t\bar{t}$, and then applied on all MC predictions (and also data) including systematic ones, ending up with BDT shape variations with respect to the nominal distribution. All of these are the description of uncertainties which will decrease the signal sensitivity in the final fit. The results after the fit show that the dominant uncertainties are the modeling of $t\bar{t} + \geq 1b$ components as described in Figure 4.39. The leading one is the $t\bar{t} + \geq 1b$ related modeling two-point uncertainty when comparing the $t\bar{t}$ samples generated by PP8 (referred to as the nominal sample) and Sherpa 2.2.1 @NLO with 5F scheme (referred to as the systematic sample). The study in this section aims to reduce the impact of this systematic source, and the default binary classifier is RNN.

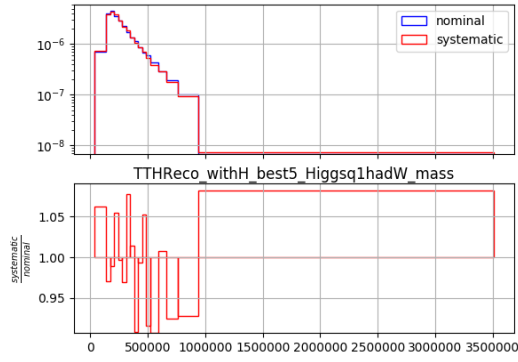
5.6.1 Differences between nominal and systematic samples

Firstly, the differences in the $t\bar{t}$ distributions between the nominal and systematic samples are checked. In Figure 5.39, the comparison of the nominal and the systematic samples are shown for some RNN input features. The comparison shows that these features are quite similar for the two samples. This is not surprising as the features that have large discrepancy between MC predictions were dropped during the BDT optimization (see Section 4.5.4.1) since they may not be well modeled by one of the simulations.

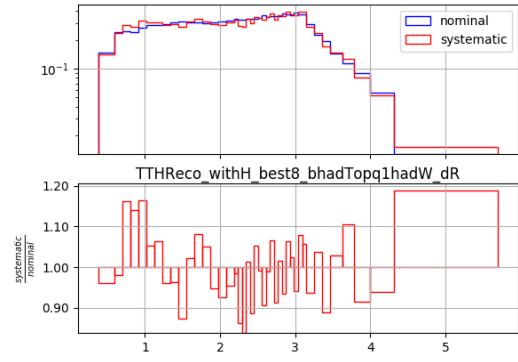
DNN classification of the two samples To further study how the neural network is sensitive to the difference between the two samples, a classical deep neural network is trained to separate the nominal and systematic $t\bar{t}$ samples. The input variables are exactly the same as the RNN inputs. Using binary cross entropy loss and the Adam optimizer, the DNN has six layers with 100 neurons and a dropout rate of 0.3 in each. Fast hand tuning of parameters like dropout rate and the number of neurons and hidden layers does not bring improvement. Due to their similarity, the DNN cannot separate the two samples very well.

The AUC evolution during the training is shown in Figure 5.40. Only events with even event number are used for the 5-fold cross-validation. The solid line in Figure 5.40 is the averaged result of the five trainings and the uncertainty band is the standard deviation. The DNN performs just better than random guessing when trying to separate the nominal and systematic $t\bar{t}$ samples.

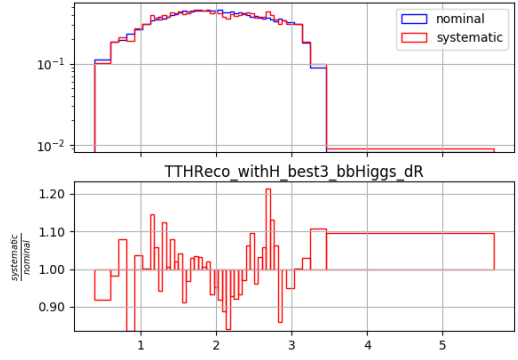
Classification performance with systematic samples More studies are carried out to check the difference of MVA distributions and performance induced by the discrepancy between the nominal and systematic samples. Table 5.9 shows the AUC of RNN and BDT applied to nominal and systematic samples separately but trained on different samples. For each RNN model training, the experiment is repeated ten times with different random seeds and both the averaged value and RMS are shown. The first section in the table is the results of RNN and BDT trained on the nominal sample only, while the second section shows the results of inclusive training on nominal and systematic samples. The models are also trained on the systematic sample only and the results are shown in the third section, where the nominal sample is halved



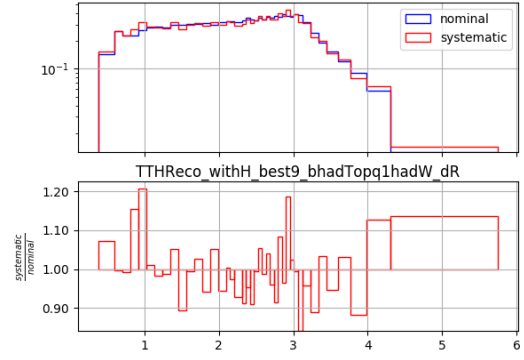
(a)



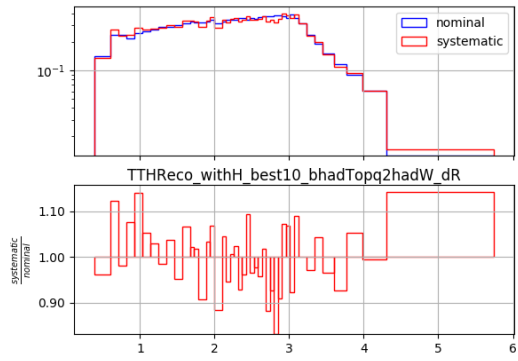
(b)



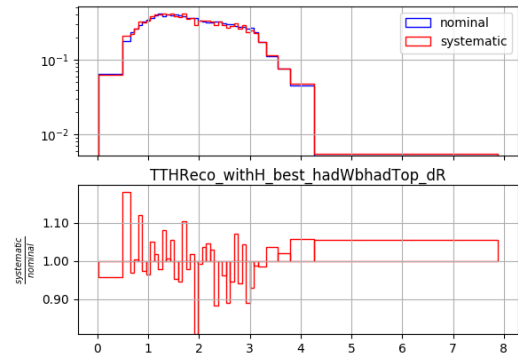
(c)



(d)



(e)



(f)

Figure 5.39: RNN input features, both with nominal and systematic samples.

and compared to the systematic sample result, as its statistics are twice that of the systematic one (in this way avoiding a purely statistical effect).

The results show that when training on the nominal sample only, the AUC testing performance



Figure 5.40: Evolution of AUC during the training of DNN to separate the nominal and systematic samples, using the same input features as the binary RNN.

	AUC
trained on nominal only	
RNN nominal	0.790 ± 0.001
RNN syst.	0.787 ± 0.001
BDT nominal	0.788
BDT syst.	0.784
trained on nominal+syst.	
RNN nominal	0.785 ± 0.001
RNN syst.	0.785 ± 0.001
trained on syst. only	
RNN nominal	0.778 ± 0.002
RNN syst.	0.781 ± 0.002

Table 5.9: AUC performance of RNN and BDT models on nominal and systematic samples. The models are trained with nominal samples only in the first section, and with both nominal and systematic events in the second. The experiments in the third section are trained on systematic only and half of the nominal sample is used for testing to use comparable statistics.

deteriorates with the systematic sample, and the degradation is similar for RNN and BDT. When training inclusively on the nominal and systematic samples, the AUC decreases with both but not that much. This shows that there is indeed some small difference between the RNN outputs of the two samples, as the test performance deteriorates when the training sample is not from the same source as the test one. Comparing the first and third sections of Table 5.9, the MVA performance decreases with nominal samples when trained on the systematic sample only, and even so does the performance with the systematic itself. This could be due to the low statistics of the systematic training sample which will limit the model performance. As presented in Section 5.3.2 where the impact of sample size is studied, a obvious decrease happens when the sample size is seriously cut. In Figure 5.12, the size of the full $t\bar{t}$ nominal

sample corresponds to the high-value end of the x -axis, the full systematic sample is half-sized. The systematic training sample is further halved with the 2-fold cross-validation, and in the end its size corresponds to the value around 0.24 on the x -axis where there is obvious drop of the AUC performance.

5.6.1.1 Metric to compare models: AMS

As discussed before, the AUC is not the full story and the signal significance is also affected by the difference between the nominal and systematic discriminant distributions. One of the difficulties in the study of this method is to find a metric to measure whether the adversarial training is actually working. Feeding the model into the final fit machinery would be a decisive test, but the model requires to be tuned and optimized by varying the hyper-parameters of the method and it is not practical to do the fit for each of the hyper-parameter configurations. The metric used in this study is the so-called AMS1_{sum} , similar to the approximate median significance (AMS1) defined in Reference [173] taking into account both signal presence and nominal-systematic difference. The signal and background distribution histograms used for the AMS calculation are rebinned to eliminate the shape effects of backgrounds. AMS1 is calculated based on the signal and background events in each bin, and then the results of all bins are summed up:

$$\text{AMS1}_{\text{sum}} = \sqrt{\sum_i^N \left(2 \left((s_i + b_i) \ln \left(\frac{s_i + b_i}{b_{0i}} - s_i - b_i + b_{0i} \right) + \frac{(b_i - b_{0i})^2}{\sigma_{bi}^2} \right) \right)}, \quad (5.25)$$

$$b_{0i} = \frac{1}{2} \left(b_i - \sigma_{bi}^2 + \sqrt{(b_i - \sigma_{bi}^2)^2 + 4(s_i + b_i)\sigma_{bi}^2} \right), \quad (5.26)$$

where s_i and b_i are the number of nominal signal and background events in each bin, and $\sigma_{bi} = \|b_i - b_{\text{syst}_i}\|$ denotes the difference between nominal and systematic background events in each bin i . A higher AMS1_{sum} means a better result.

Another AMS metric is also explored:

$$\text{AMS2}_{\text{diff}} = \frac{\text{AMS2}_{\text{nominal}} - \text{AMS2}_{\text{syst}}}{\text{AMS2}_{\text{nominal}}} \quad (5.27)$$

where AMS2 is defined as:

$$\text{AMS2} = \sqrt{\sum_i^N 2 \left((s_i + b_i) \ln \left(1 + \frac{s_i}{b_i} \right) - s_i \right)}. \quad (5.28)$$

It accounts for the signal presence significance against the background. This formula is the same as AMS2 in Reference [174] which is an approximation of AMS1 for σ_{bi} close to zero, except being calculated bin by bin to take into account the discriminating MVA output shape. Such bin-by-bin AMS2 is computed using nominal signal against nominal background ($\text{AMS2}_{\text{nominal}}$), or against systematic background ($\text{AMS2}_{\text{syst}}$). The difference between the two is calculated to represent the difference between the two background samples. A lower $\text{AMS2}_{\text{diff}}$ means a better result.

Both measures are found to be unstable, where the AMS2 is more fluctuating as shown in Table 5.10. 100 experiments are implemented with the RNN binary classifier, where only the random seeds are varied. The models are trained on the nominal sample only, and are

only applied on the systematic samples to calculate AMS values. When doing the bin-by-bin calculations, the binning is redone for each experiment separately using the optimized method described in Section 4.7.2. AMS2_{diff} shows much larger standard deviation on validation and testing samples compared with the AMS1_{sum}.

	AMS1 _{sum} (mean±std(%))	AMS2 _{diff} (mean±std(%))
train	0.76±0.15(19.3)	0.12±0.02(13.9)
val	1.08±0.14(12.6)	0.02±0.02(94.1)
test	1.77±0.30(17.2)	0.03±0.01(31.0)

Table 5.10: The performance metric AMS1_{sum} and AMS2_{diff} on train, validation, and test samples. The mean and standard derivation are calculated from 100 experiments where only the random seeds are varied.

Thus AMS1_{sum} will be used in the following study as a measure of the classifier performance taking into account both signal against background separation and the resistance to nominal-systematic sample variations. However, the unstability of AMS1_{sum} makes it partially unreliable to compare various classifiers with a single training.

The AMS1_{sum} evolution during training is shown in Figure 5.41 where the AUC evolution is also shown for the same training process. The RNN model is trained to classify $t\bar{t}H$ and $t\bar{t}$ on nominal samples only. Systematic samples are only used to calculate AMS. The general trend of the validation (in orange) curves are the same for AMS1_{sum} and AUC. The training (in blue) and testing (in green) curves of AMS go down as more epochs are trained, which is not surprising as the model is trained specifically to fit better and better the nominal sample and loses the generalization power on a different sample as the training evolves.

AMS1_{sum} is also more fluctuating than the AUC, and it is even sensitive to randomness caused by different GPUs. The most important contribution to the fluctuation is the RNN predictions variations or the systematic samples in each bin in experiments with random seed. For a specific experiment, whether using the re-binning based on its proper distribution or the one from another experiment with a different random seed does not have very large impact on the evolution curve.

Figure 5.42 shows the MVA distributions of nominal and systematic samples for the $t\bar{t}$ background and their relative difference. Nominal $t\bar{t}H$ signal is also shown and normalized to the nominal background. The histograms are rebinned to have the relative statistical error of the systematic background in each bin to be at most 5%. Three MVA distributions are shown which are trained on nominal sample only where the RNN and BDT use high level features while the tree+RNN uses low level features only. From the distributions and their AMS values, no clear evidence shows that one or two MVAs has more consistent nominal and systematic distributions.

5.6.2 Adversarial networks

Generative Adversarial Networks Adversarial networks were invented for a generative model, the *Generative Adversarial Networks (GAN)* [141]. Without relying on any assumptions about the distribution in the given data, GAN shows the potential to learn high-dimensional and complex data distribution. As its name implies, the model involves two components: the generator G trained in competition with the discriminator D to produce the desired samples. G is a continuous transformation function trained to map a random vector space into another

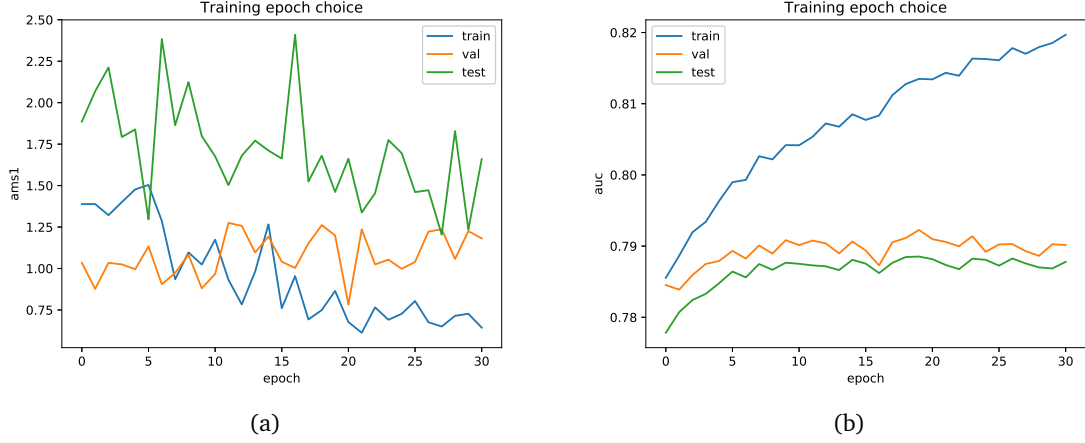


Figure 5.41: The AMS1_{sum} and AUC evolution during the training of the RNN to classify $t\bar{t}H$ and $t\bar{t}$. The training is on nominal sample only, and the systematic sample just enters the calculation of AMS1_{sum} .

space that is similar to the desired data. D distinguishes its input about whether it comes from the real data distribution or the generator. Such a pitted training iteration of G and D is repeated and can theoretically converge to stability where D can not classify the real and generated samples any more.

GAN has received wide attention and many variants have emerged applied to various tasks [175]. In the realm of particle physics, this deep learning technique is being explored as an alternative fast simulation approach where it is trained to learn from full simulation to generate new samples. The simulation of EM showers in ECal [176, 177] and jet images [178] show much less computational consumption than the full simulation, and promising accuracy and fidelity.

Constraining the classifier with adversarial training Based on the idea of adversarial network training of GAN, a method is proposed [179] to train and add constraints on the classifier. As discussed at the beginning of this chapter, the classification BDT, as a discriminant for fitting, has different shapes for $t\bar{t}$ nominal and systematic samples, as a representation of the background modeling uncertainty. The large shape difference will decrease the $t\bar{t}H$ signal sensitivity. This method aims to train a discriminating classifier while forcing it to produce more consistent distributions on nominal and systematic samples through adversarial training, so as to improve the signal sensitivity.

The idea is as shown in Figure 5.43. The whole network system is trained inclusively on the nominal and systematic samples as the input to the classifier which is the plain binary RNN described in Section 5.3 of this study. The output of the classifier is fed into another neural network, called the discriminator, which is trained to restrict the output of the classifier.

In each batch training iteration, two training phases are processed successively and in an adversarial way trying to find the minimax solution:

$$\hat{\theta}_c, \hat{\theta}_d = \arg \min_{\theta_c} \max_{\theta_d} \left(L_c(\theta_c) - \lambda L_d(\theta_c, \theta_d) \right), \quad (5.29)$$

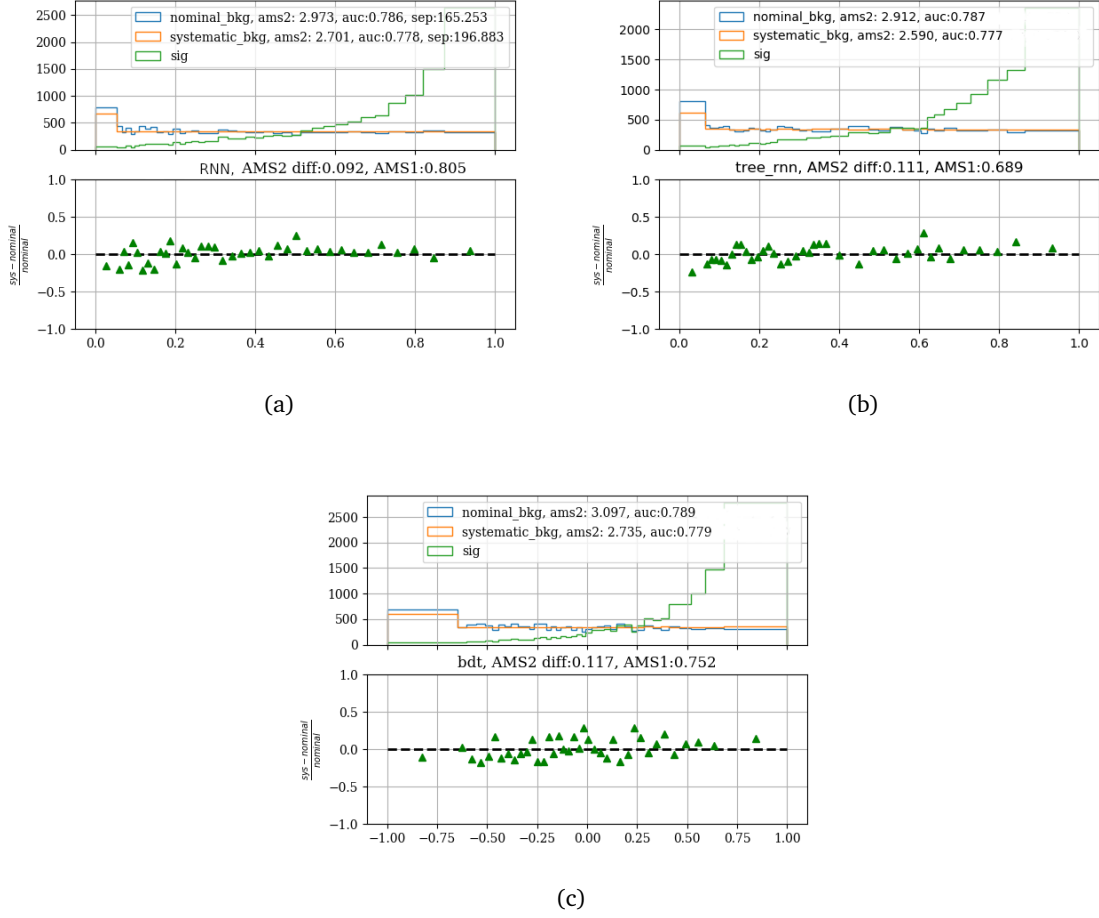


Figure 5.42: The RNN (a), tree+RNN (b) and BDT (c) distributions of nominal and systematic samples for $t\bar{t}$ background and their relative difference. Nominal $t\bar{t}H$ signal is also shown and normalized to the nominal background. Models are trained on nominal only.

where $L_c(\theta_c)$ and $L_d(\theta_c, \theta_d)$ are the loss functions of the network connection weights of the classifier (θ_c) and discriminator (θ_d), and λ is a tunable positive real number. The training is performed as follows:

- With the classifier θ_c fixed, the discriminator θ_d is trained to discriminate the classifier output of the nominal sample from that of the systematic sample.
- With the discriminator θ_d fixed, the classifier θ_c is trained to minimize $L_c(\theta_c) - \lambda L_d(\theta_c, \theta_d)$, meaning that it tries to discriminate the $t\bar{t}H$ signal from the $t\bar{t}$ background of the inclusive samples, and at the same time to give as similar as possible outputs for the nominal and systematic samples.

In the end, such an alternating procedure is proven [179] to converge at the point where the discriminator is totally confused by the classifier, giving random guessing about the nominal and systematic discrimination, and the classifier is trained to separate $t\bar{t}H$ and $t\bar{t}$ and be insensitive to the different samples.

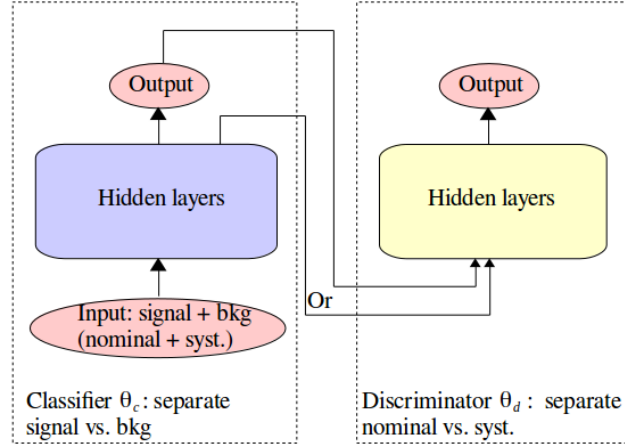


Figure 5.43: Sketch of the adversarial training network.

5.6.3 Adversarial trainings

As many other adversarial models, such a game requires parameter fine tuning in order to converge. Both the classifier and the discriminator are pre-trained individually before adversarial training iterations, in order to find the proper initial states for the adversarial game. During the pre-training, the classifier is trained to separate two classes, e.g., $t\bar{t}H$ from $t\bar{t}$ events, while the discriminator is trained alone to classify domains, e.g., nominal against systematic samples. Then in each adversarial training iteration, the discriminator can be trained with more batches for one batch training of the classifier. Such a number of batches together with the importance factor λ of the adversarial restriction are grid-searched. Lots of experiments with varied discriminator architecture and the optimizer of the two training phases are conducted as well. Instead of feeding the classifier's output into the discriminator, experiments are also carried out to feed the last hidden layer which may be more different between the nominal and systematic samples, as it contains more information than the classifier output. Lots of experiments have been implemented considering all these possible variations of the adversarial model.

5.6.3.1 Validation of implementation

The results in Reference [179] are firstly repeated using the same public high energy physics simulated data, in order to validate the code that will be used later for the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis. On the other hand, this also helps to learn how the loss function and AUC of both the classifier and discriminator evolve during the adversarial training, if the adversarial training is working.

In this experiment, the training dataset has two hundred thousand samples with two domains, nominal and systematic samples, for both signal and background. And the discriminator is sensitive to these two domains of both classes. This is slightly different from the adversarial training for the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis in this thesis, where the task is simplified and the adversarial model is only trained to be sensitive to the two domains of the background. Among the training set, four thousand events are used as the validation sample.

Both the classifier and the discriminator in this model are simple classical DNN. The classifier,

according to Reference [179], has three hidden layers with 64 neurons in each. The first hidden layer uses the tanh activation function, the other hidden layers use ReLU and the output layer uses sigmoid. The discriminator has three hidden layers as well with 60 neurons and ReLU activation function in each. The classifier is pre-trained until it reaches the maximum validation AUC before early stopping, while the discriminator is pre-trained with 11 epochs (half way before reaching the maximum validation performance). Five epochs are trained with λ set to 10 and batch size to 400. For each batch training of the classifier, the discriminator is trained with the full training set.

The distributions of the nominal signal, the nominal and systematic background, and the relative difference between the two background samples, are shown in Figure 5.44, both before and after adversarial training. A significant inconsistency between the background shapes can be seen. The absolute value of $\text{AMS2}_{\text{diff}}$ is decreased, while the AMS1_{sum} is largely increased, from around 32 to 68, while the AUC performance is maintained.

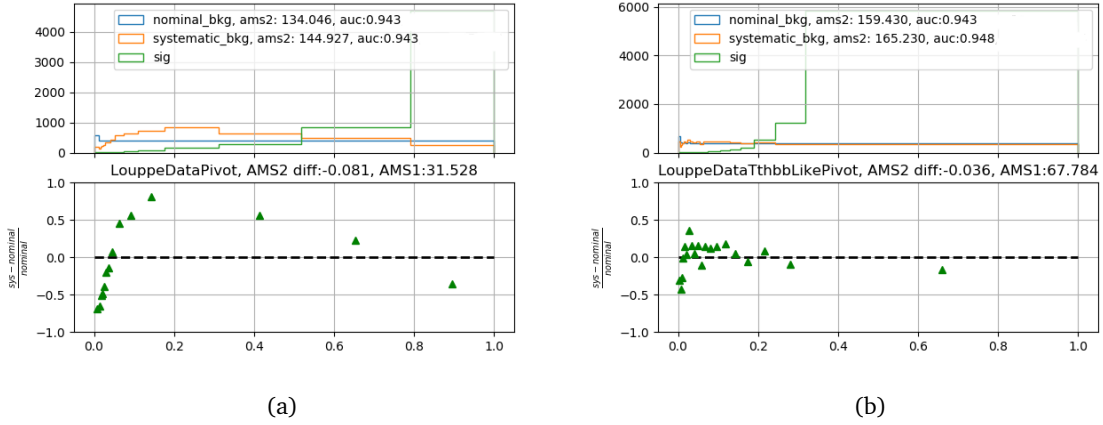


Figure 5.44: Distributions of the DNN classifier output for nominal signal and background, as well as the systematic background sample of the public data used in Reference [179], both (a) before and (b) after the adversarial training.

The evolution of the loss and AUC of the classifier and the discriminator are shown in Figure 5.45. In total five epochs are trained without early stopping and the model at the end of the training are obtained as the adversarial training result. For every 50 batches training of the discriminator, the loss and AUC are calculated on the mini-batch of that iteration for the training curves in blue and on the separate validation samples for the test curve in orange. Thus the training curve is calculated on different sample batch while the test curve is calculated on the same samples.

The loss of classifier curve is the evolution of the first term in equation 5.29 which is the binary cross-entropy of the classification of the nominal signal and the inclusive background of nominal plus systematic samples. This classification performance is shown as the classifier AUC evolution curve in the figure. The overall scale of the two curves does not change as the training evolves, while in some other trials where different parameter configurations are used and this algorithm also works, they could go up first and then go down to the previous level.

The total loss curve is the evolution of the full loss shown in equation 5.29 trained to be minimized during the training phase where the discriminator is fixed. The curve shows an expected decreasing trend as the training evolves.

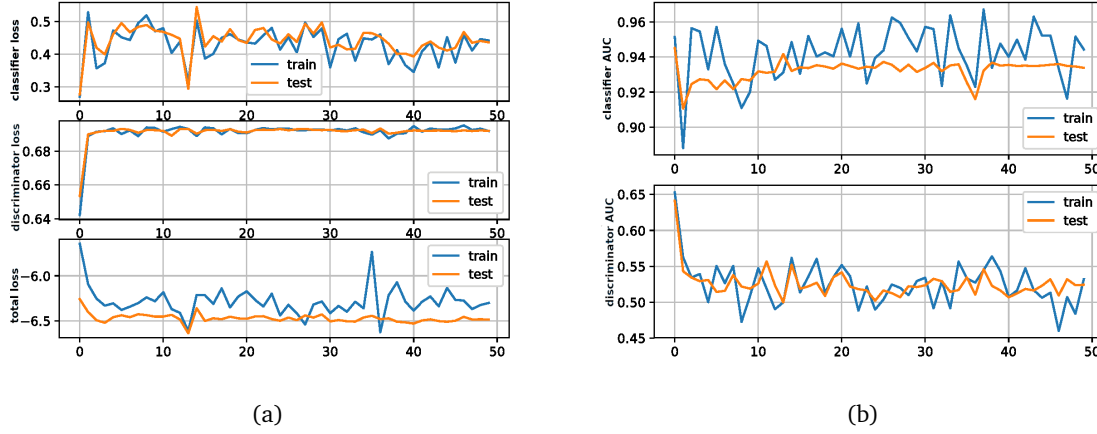


Figure 5.45: Training evolution of the adversarial network system on the public data used in Reference [179]. The x -axis is training epochs. The y -axis includes (a) several loss functions and (b) AUC of both classifier and discriminator. Definitions are given in the text.

The loss of the discriminator corresponds to the second term $L_d(\theta_c, \theta_d)$ in equation 5.29 which is the binary cross-entropy when classifying the nominal and systematic samples. The loss curves keep increasing and the AUC decreasing until the random guessing line at 0.5. This means that the discriminator is confused by the classifier at the end of the training and the algorithm converges.

In order to learn how the loss and AUC values evolve during the training in an experiment that is more close to the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) scenario, another adversarial training experiment is also implemented using the same two hundred thousand events in the public data, but only two domains (nominal and systematic samples) of the background events enter the adversarial game. The tuned model has the same architecture, with the same batch training pipeline and λ value as before. The only difference is that the discriminator is pre-trained with one epoch. The difference of the two background domains is again mitigated as shown in Figure 5.46. The evolution of the loss and AUC values of both classifier and discriminator are shown in Figure 5.47, and they have the similar trend as before.

These curves show how the loss and AUC values evolve in the experiments where the adversarial training can actually bring improvement, and can help to monitor or diagnose the adversarial training in the following experiments for the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis.

5.6.3.2 Adversarial training on $t\bar{t}H$ ($H \rightarrow b\bar{b}$)

The classifier architecture uses an LSTM consisting of 30 neurons. Both internal dropout rates are 0.2 (one for input gates, and another for recurrent connections). The LSTM layer is connected to a hidden layer having 30 neurons using ReLU activation function. Both the classifier and discriminator reach their optimal performance during the pre-training. Variations on these initial states brings no improvement.

Many experiments are also implemented to optimize the adversarial training, especially in order to make the discriminator sensitive to the difference between the two background samples. Firstly, the discriminator is hand-tuned starting from a simple FC DNN with three layer

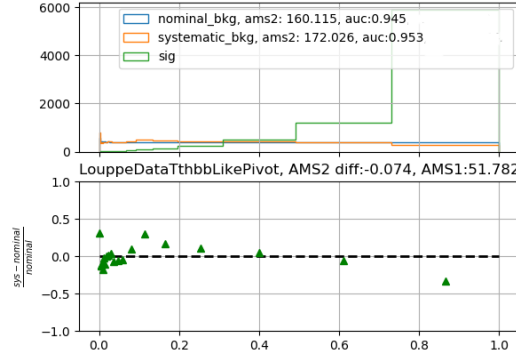


Figure 5.46: Distributions of the DNN classifier output for nominal signal and background, as well as the systematic background sample after the adversarial training where only the background events have its two domains, nominal and systematic samples, enter the adversarial training.

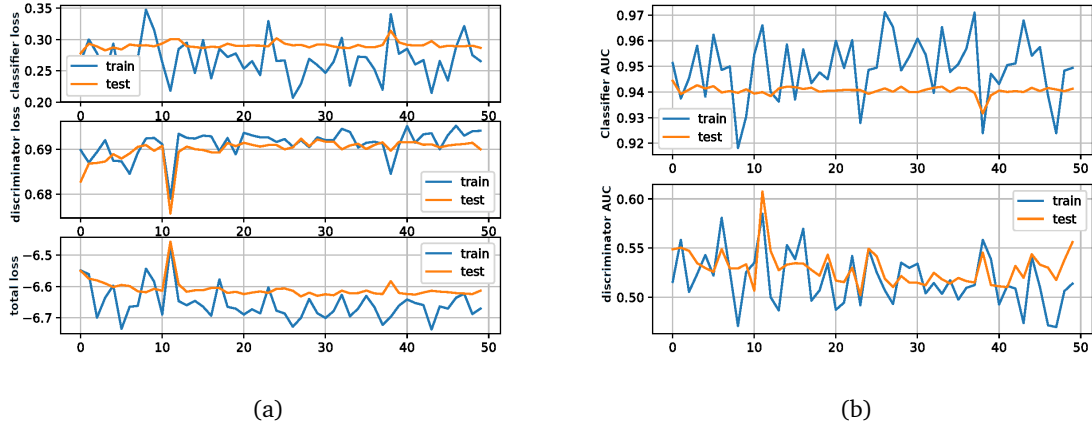


Figure 5.47: Training evolution of the adversarial network system on the public data used in Reference [179], where only the background events have its two domains, nominal and systematic samples, enter the adversarial training. The x -axis is training epochs. The y -axis includes (a) several loss functions and (b) AUC of both classifier and discriminator. Definitions are given in the text.

each having 30 neurons. More layers, neurons and dropout rate are added to discriminate the two background samples. The increased complexity of the network is tested in both pre-training and the adversarial training, and it does not bring much help to the sample discrimination and the functionality of the adversarial training. Furthermore, the discriminator is also set to be only sensitive to a subset of background: (a) the $t\bar{t} + \geq 1b$ sub-component of the inclusive background as the systematic sample, (b) the events having classifier output larger than 0.5. The discriminator in the pre-training does not give better separation performance on these subsets, and does not give improvement in the adversarial training either.

Since the one-dimensional classifier output has only small differences for the two background

samples, feeding the classifier's hidden layer into the discriminator is also tested. This gives only marginal improvement during the pre-training of the discriminator. In the adversarial training, both feeding options are tested, together with variations on optimizers choosing from Adam, Adagrad, and SGD, and on λ value varying as 10, 100 or 1000. Among many experiments, the most promising adversarial training on the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis is shown in this section. The corresponding model is trained with $\lambda = 1000$. For each batch training of the classifier, half of the total batches are trained for the discriminator. The batch size is 400. The hidden layer of the classifier is fed into the discriminator which has five hidden layers with 50 neurons and dropout rate of 0.3 in each. Both training phases use the SGD optimizer.

The adversarial training results are shown in Table 5.11(a) and Figure 5.48(a). The results without adversarial training are also shown in Table 5.11(b) and Figure 5.48(b), where the classifier is trained inclusively on the nominal and systematic samples. In Table 5.11, experiments are repeated ten times with different random seeds. Both the average AUC and AMS1_{sum} and their RMS are shown. The distributions in Figure 5.48 are one result from the ten. The adversarial training gives increased central value of AMS1 but with large uncertainty, when the AUC pays the price, which is expected. And the RNN distribution difference between samples does not show obvious mitigation. There is no clear evidence of an improvement.

	AUC	AMS1
nominal	0.771 ± 0.004	0.993 ± 0.189
syst.	0.762 ± 0.005	

(a) With adversarial training

	AUC	AMS1
nominal	0.784 ± 0.001	0.942 ± 0.149
syst.	0.778 ± 0.001	

(b) Without adversarial training

Table 5.11: The AUC and AMS1 performance of the RNN classifier trained inclusively with nominal and systematic samples, with (a) and without (b) the adversarial training in the configuration described in the text.

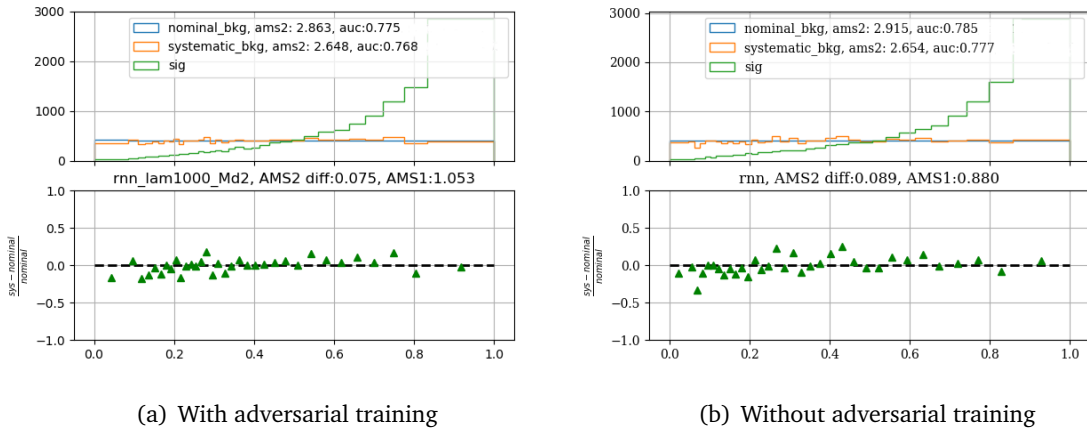


Figure 5.48: The RNN distributions of nominal and systematic samples for $t\bar{t}$ background and their relative difference with (a) and without (b) adversarial training in the configuration described in the text. Nominal $t\bar{t}H$ signal is also shown and normalized to the nominal background.

The diagnosing curves shown in Figure 5.49 for this experiment do not show sensitive trends.

The classifier loss increases while the AUC decreases as described above. This could be due to the large emphasis weight λ on the loss term of the discriminator. However, the choice of the large λ could be triggered by the fact that the background nominal and systematic samples are too similar ending up with a weak impact from the discriminator. The discriminator AUC is shown at the bottom of Figure 5.49(b) where the curve floats around the 0.5 random guessing line even from the beginning of the training.

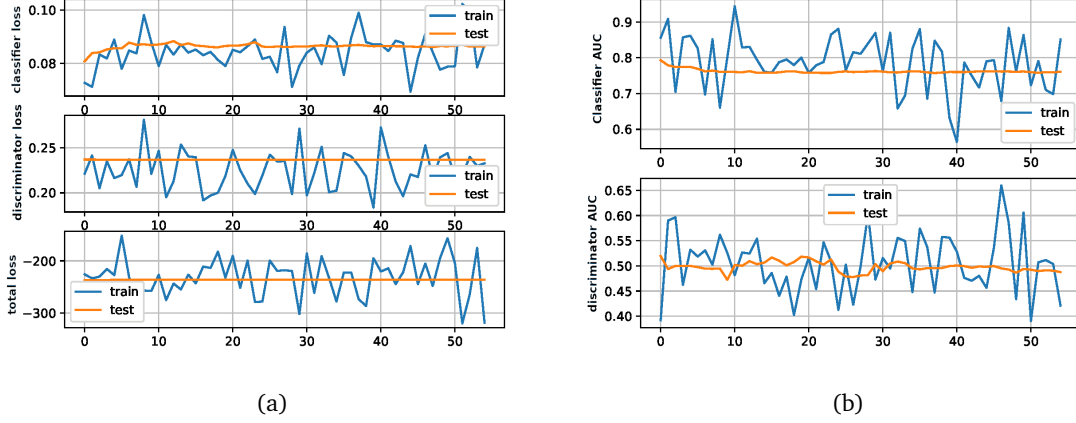


Figure 5.49: Training evolution of the adversarial network system on $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis samples. The x -axis is training epochs. The y -axis includes (a) several loss functions and (b) AUC of both classifier and discriminator. Definitions are given in the text.

5.6.4 Summary

This section exploits the adversarial training to solve a two label domain adaptation task, i.e., to train a binary classifier that is insensitive to the difference between two background samples, the nominal and the systematic samples. Although the systematic sample, giving estimation about the modeling uncertainty of the $t\bar{t} + \geq 1b$ background component, is relatively large compared to the very rarely produced $t\bar{t}H$ ($H \rightarrow b\bar{b}$) signal, studies show that the difference by itself is quite small and is very difficult to be discriminated by the DNN model. This makes the classifier difficult to gain anything from the adversarial training, as the discriminator is rarely able to give feedback. There is no clear evidence showing whether this method is working without having fit results. Another difficulty is to find a metric to measure this method's functionality before fitting. Two AMS variations, AMS1_{sum} and $\text{AMS2}_{\text{diff}}$, are studied to take into account both the signal presence and the background uncertainty. The AMS1_{sum} is more stable than $\text{AMS2}_{\text{diff}}$, but both are much more fluctuating compared with the AUC, when random seeds vary from one experiment to another.

5.7 Summary

In this chapter, cutting edge deep learning techniques are exploited in the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis for the first time, targeting a contribution to the new analysis round using the full

Run 2 data.

Recurrent neural networks are used to do binary classification of $t\bar{t}H$ signal and $t\bar{t}$ background. The two-step MVAs from the analysis described in Chapter 4 use intermediate models to solve the jet-parton assignment, and, in a second step, use a BDT for classification. The intermediate models use similar input features but each can only exploit partial information from them. The RNN instead provides an elegant approach to solve the reconstruction and the classification problem in one step, being able to exploit full information from the input. It shows slightly better AUC performance than the two-step MVAs in the previous analysis, trained on the same events and using similar high-level features. It also provides the possibility to exploit the full variable-size combinations between jets and partons.

Deep neural networks are also studied using low-level inputs. In the analysis in Chapter 4, the performance of the classification BDT closely depends on the choice of the input features. Lots of work and computing resources are required to calculate and optimize input high-level features. When trained with low-level features only, the classical feed-forward DNN shows worse performance. A parse tree model is designed to incorporate physics domain knowledge into the model architecture. The study shows that the tree can learn from the low-level features more efficiently than the traditional DNN. The combination of the RNN and parse tree models gives comparable AUC performance as the two-step MVA, but only using low-level features, showing that the tree can automatically learn as much information as high-level features from low-level ones only. In turn, it also shows that the two-step MVAs are well optimized (see Chapter 4).

Deep neural networks are also exploited to categorize the full event space into regions. In Chapter 4, regions are defined by having requirement on the b -tagging weight of jets. The requirement is optimized manually in order to have signal rich and background control regions. The deep neural network with RNN provides a new region definition approach, as a multi-classifier to classify the whole dataset into $t\bar{t}H$ signal and background sub-components. This gives as good performance as manual splitting in terms of background decomposition and signal purity. However, the splitting with DNN could conveniently benefit from c -tagging variables as additional inputs in the future, and from tunable selections on the various outputs, to further improve the separation between the $t\bar{t} + \geq 1c$ and $t\bar{t} + \text{light}$ backgrounds. Trainings with more epochs of the multiclassifier can also be tested, since the current model is stopped before the $t\bar{t} + \geq 1c$ AUC reaches a plateau.

In order to decrease the $t\bar{t}$ modeling uncertainty, a driving factor of the signal sensitivity, adversarial training is studied. The classifier is trained to pit against a discriminator to give outputs insensitive to the difference between the nominal and systematic samples. It is not clear that this approach brings any benefit to this analysis. One difficulty is that the nominal and systematic samples are very similar and are hard to be separated by the discriminator. Fitting to data will help to conclude, as it is also difficult to find a metric to measure the model's performance before fit.

Conclusion

The Higgs boson discovery in 2012, by the ATLAS and CMS collaboration during the LHC Run 1 operation, completes the last piece on the particle table of the standard model. However, property measurements of this new particle are fundamental as well, as an important way to further challenge the SM and therefore to test new physics. The measurement of the couplings between the Higgs boson and fermions, the so-called Yukawa couplings, is one of the priorities on the LHC Run 2 experiments. By the start of this thesis in the second half of 2016, this coupling strength was only indirectly measured with large uncertainty. The $t\bar{t}H$ process instead provides a direct approach to measure the Yukawa coupling between the Higgs boson and top quark which is much heavier than other quarks. The $t\bar{t}H$ ($H \rightarrow b\bar{b}$) process is searched in this thesis, where the $H \rightarrow b\bar{b}$ is the dominant Higgs decay mode and is also sensitive to the b -quark Yukawa coupling. On the other hand, this channel is challenged by the complicated final state including multiple jets and b -jets, and therefore heavily depends on advanced analysis algorithms for signal extraction, such as the multivariate analysis techniques.

As presented in Chapter 4, the author contributes to the MVA models in the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis published in April 2018 [25], using datasets collected in years 2015 and 2016 for an integrated luminosity of 36.1 fb^{-1} . In the first step, reconstruction BDT is studied and optimized to reconstruct the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) process, which tries to find the correct association between multiple jets in the final state and partons. The author develops a strategy to exploit b -tagging weight for the first time. Implemented during the training sample preparation, this strategy allows to have more signal events which are correct associations while the increment of backgrounds, the improper associations, is limited. The strategy brings up to 5% improvement for the efficiency of correctly reconstructing the Higgs boson. In the second step, boosted decision trees are studied and optimized to separate $t\bar{t}H$ ($H \rightarrow b\bar{b}$) from the $t\bar{t}$ events, the main dominant background. This classification BDT uses features built from the combination that is selected by the reconstruction BDT. Event-level kinematics and the output of two other intermediate MVAs are also used as input, as well as the b -tagging variables which are exploited for the first time. It is found that the analysis can be simplified by training one classification BDT in the full event space by using the b -tagging variables. The study also shows the possibility to reduce the input list while keeping the classification performance. The output of this classification BDT is used as the final discriminant in the final fit.

The fit result gives a signal strength of $\mu = 0.84^{+0.64}_{-0.61}$ for a Higgs boson mass of 125 GeV. Under the background only hypothesis, the observed (expected) significance of the $t\bar{t}H$ signal over the expected SM background is 1.4 (1.6) standard deviation, without being able to claim $t\bar{t}H$ observation. In the second half of 2018, the $t\bar{t}H$ and $H \rightarrow b\bar{b}$ observations were published where combined studies were implemented using several Higgs decay channel and production modes, both including the result of this $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis. Further studies are still being carried out in the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) channel, the improvement of which can largely contribute to

the precise measurement of the top Yukawa coupling. Also, the well reconstructed variables thanks to the reconstruction BDT can help to test the CP structure of the $t\bar{t}H$ interaction.

Targeting an improved result in the new analysis round using the full Run 2 data, the author exploits the deep learning techniques on various aspects of the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis, as presented in Chapter 5. Recurrent neural networks provide an elegant approach to solve the reconstruction and the classification problems in one step, being able to exploit full information from the input. It gives better AUC performance of the classification between the $t\bar{t}H$ signal and the $t\bar{t}$ background than the two-step MVAs which consist of various techniques. It also provides the possibility to exploit the full variable-size jet-parton combinations. The ability of the deep models to extract information from low-level features is firstly investigated, where the classical deep neural network shows worse performance when training with raw features only. A parse tree model, that is designed in analogy to the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) decay chain, is found to be able to learn from the low-level features more efficiently than the classical DNN. Combined with the RNN, this model shows that the tree can automatically learn as much information as high-level features. This provides the possibility to save computing resources taken by the feature engineering, since the traditional MVA performance depends on the discrimination power of the input high-level features. In turn, this also shows that the two-step MVAs in Chapter 4 are well optimized. A new region definition is also provided by RNN doing multiclassification of $t\bar{t}H$ signal and $t\bar{t}$ sub-components. This gives as good performance as manual splitting, but it can conveniently benefit from c -tagging variables additionally in the future, and from tunable selections on its various outputs, and therefore potentially further separate $t\bar{t} + \geq 1c$ and $t\bar{t} + \text{light}$ backgrounds. The adversarial networks are also studied aiming to reduce the systematic uncertainties caused by the bad modeling of $t\bar{t} + \geq 1b$ background. The difference between the nominal and systematic samples are difficult to be discriminated by neural networks. There is no clear evidence that this method is improving the analysis sensitivity.

The author also contributes to the upgrades of the ATLAS Inner Tracker, an important part of the whole detector upgrades intended to cope with the high-luminosity LHC expected to start operation in the mid-2020's. As the closest sub-detector to the beam line, the ITk is expected to withstand much harsh radiation environment. As a key point to have the correct radiation estimation in the ITk, the author works to synchronize its geometry descriptions that are independently implemented by radiation estimation and simulation groups. The radiation estimation for the ITk pixel TDR [84] is validated. The method that is explored by the author is still being used by the radiation study group to validate the latest iteration of the pixel detector layout.

Bibliography

- [1] C. N. Yang and R. L. Mills, *Conservation of Isotopic Spin and Isotopic Gauge Invariance*, [Phys. Rev. **96** \(1 1954\) p. 191](#) (cit. on p. 12).
- [2] P. M. Watkins, *Discovery of the W and Z bosons*, [Contemp. Phys. **27** \(1986\) p. 291](#) (cit. on p. 12).
- [3] F. Abe et al., *Observation of top quark production in $\bar{p}p$ collisions*, [Phys. Rev. Lett. **74** \(1995\) p. 2626](#), arXiv: [hep-ex/9503002 \[hep-ex\]](#) (cit. on p. 12).
- [4] S. Abachi et al., *Observation of the top quark*, [Phys. Rev. Lett. **74** \(1995\) p. 2632](#), arXiv: [hep-ex/9503003 \[hep-ex\]](#) (cit. on p. 12).
- [5] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, [Phys. Lett. B **716** \(2012\) p. 1](#), arXiv: [1207.7214 \[hep-ex\]](#) (cit. on pp. 12, 23, 84).
- [6] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, [Phys. Lett. B **716** \(2012\) p. 30](#), arXiv: [1207.7235 \[hep-ex\]](#) (cit. on pp. 12, 23, 84).
- [7] ATLAS and CMS Collaborations, *Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments*, [Phys. Rev. Lett. **114** \(2015\) p. 191803](#), arXiv: [1503.07589 \[hep-ex\]](#) (cit. on p. 13).
- [8] F. Boudjema and D. Zeppenfeld, *The Standard Model of particle physics: an introduction to the theory*, [Comptes Rendus Physique **3** \(2002\) p. 1097](#) (cit. on pp. 14, 17).
- [9] M. Tanabashi et al., *Review of Particle Physics*, [Phys. Rev. D **98** \(2018\) p. 030001](#) (cit. on p. 15).
- [10] T. D. Lee and C. N. Yang, *Question of Parity Conservation in Weak Interactions*, [Phys. Rev. **104** \(1956\) p. 254](#) (cit. on p. 15).
- [11] C. S. Wu, E. Ambler, R. W. Hayward, et al., *Experimental Test of Parity Conservation in Beta Decay*, [Phys. Rev. **105** \(1957\) p. 1413](#) (cit. on p. 15).
- [12] E. C. G. Sudarshan and R. E. Marshak, *Chirality Invariance and the Universal Fermi Interaction*, [Phys. Rev. **109** \(5 1958\) p. 1860](#) (cit. on p. 15).
- [13] S. Okubo, R. E. Marshak, E. C. G. Sudarshan, et al., *Interaction Current in Strangeness-Violating Decays*, [Phys. Rev. **112** \(1958\) p. 665](#) (cit. on p. 15).
- [14] S. L. Glashow, *Partial Symmetries of Weak Interactions*, [Nucl. Phys. **22** \(1961\) p. 579](#) (cit. on p. 15).
- [15] S. Weinberg, *A Model of Leptons*, [Phys. Rev. Lett. **19** \(1967\) p. 1264](#) (cit. on p. 15).

- [16] A. Salam, *Weak and Electromagnetic Interactions*, Proceedings of the 8th Nobel symposium, Ed. N. Svartholm, Almqvist & Wiskell, 1968, Conf. Proc. **C680519** (1968) p. 367, URL: <http://inspirehep.net/record/53083> (cit. on p. 15).
- [17] F. Englert and R. Brout, *Broken symmetry and the mass of gauge vector mesons*, *Phys. Rev. Lett.* **13** (1964) p. 321 (cit. on p. 16).
- [18] P. W. Higgs, *Broken symmetries, massless particles and gauge fields*, *Phys. Lett.* **12** (1964) p. 132 (cit. on p. 16).
- [19] P. W. Higgs, *Broken symmetries and the masses of gauge bosons*, *Phys. Rev. Lett.* **13** (1964) p. 508 (cit. on p. 16).
- [20] F. Halzen and A. D. Martin, *quarks and leptons: an introductory course in modern particle physics*, New York, Usa: Wiley (1984) 396p, 1984, ISBN: 0471887412, 9780471887416 (cit. on p. 19).
- [21] C. Mariotti and G. Passarino, *Higgs boson couplings: measurements and theoretical interpretation*, *Int. J. Mod. Phys. A* **32** (2017) p. 1730003, arXiv: 1612.00269 [hep-ph] (cit. on pp. 21, 22).
- [22] D. de Florian et al., *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*, 2016, arXiv: 1610.07922 [hep-ph] (cit. on pp. 21, 89, 130).
- [23] J. R. Andersen et al., *Handbook of LHC Higgs Cross Sections: 3. Higgs Properties*, ed. by S. Heinemeyer, C. Mariotti, G. Passarino, et al., 2013, arXiv: 1307.1347 [hep-ph] (cit. on p. 22).
- [24] ATLAS and CMS Collaborations, *Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV*, *JHEP* **08** (2016) p. 045, arXiv: 1606.02266 [hep-ex] (cit. on p. 23).
- [25] ATLAS Collaboration, *Search for the standard model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, *Phys. Rev. D* **97** (2018) p. 072016, arXiv: 1712.08895 [hep-ex] (cit. on pp. 24, 85, 86, 91–93, 126, 128, 129, 137–139, 142–144, 147, 158, 166, 169, 175, 190, 204).
- [26] ATLAS Collaboration, *Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector*, *Phys. Lett. B* **784** (2018) p. 173, arXiv: 1806.00425 [hep-ex] (cit. on p. 24).
- [27] ATLAS Collaboration, *Observation of $H \rightarrow b\bar{b}$ decays and VH production with the ATLAS detector*, *Phys. Lett. B* **786** (2018) p. 59, arXiv: 1808.08238 [hep-ex] (cit. on p. 24).
- [28] ATLAS Collaboration, *Combined measurements of Higgs boson production and decay using up to 80fb^{-1} of proton–proton collision data at $\sqrt{s} = 13$ TeV collected with the ATLAS experiment*, ATLAS-CONF-2019-005, 2019, URL: <https://cds.cern.ch/record/2668375> (cit. on p. 24).
- [29] F. Boudjema, R. M. Godbole, D. Guadagnoli, et al., *Lab-frame observables for probing the top-Higgs interaction*, *Phys. Rev. D* **92**, 015019 (2015) (2015), arXiv: 1501.03157 (cit. on p. 25).
- [30] L. Evans and P. Bryant, *LHC Machine*, *JINST* **3** (2008) S08001 (cit. on pp. 26, 28).
- [31] *Facts and figures about the LHC*, 2019, URL: <https://home.cern/resources/faqs/facts-and-figures-about-lhc> (cit. on p. 27).

- [32] L. Rossi, *The LHC superconducting magnets*, [Conf. Proc. C030512](#) (2003) p. 141 (cit. on p. 28).
- [33] *Luminosity Public Results Run 2*, 2019, URL: <https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2> (cit. on p. 30).
- [34] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, [JINST 3 \(2008\) S08003](#) (cit. on pp. 29, 35, 36, 38–40).
- [35] CMS Collaboration, *The CMS Experiment at the CERN LHC*, [JINST 3 \(2008\) S08004](#) (cit. on p. 30).
- [36] A. Augusto Alves Jr et al., *The LHCb Detector at the LHC*, [JINST 3 \(2008\) S08005](#) (cit. on p. 30).
- [37] K. Aamodt et al., *The ALICE experiment at the CERN LHC*, [JINST 3 \(2008\) S08002](#) (cit. on p. 30).
- [38] *How ATLAS detects particles: diagram of particle paths in the detector*, 2019, URL: <https://cds.cern.ch/record/1505342> (cit. on p. 32).
- [39] *Event Displays from Collision Data*, 2015, URL: https://twiki.cern.ch/twiki/bin/view/AtlasPublic/EventDisplayRun2Collisions#13_TeV_Stable_Beam_Collisions (cit. on p. 34).
- [40] ATLAS Collaboration, *ATLAS Insertable B-Layer Technical Design Report*, ATLAS-TDR-19, 2010, URL: <https://cds.cern.ch/record/1291633> (cit. on p. 35), Addendum: ATLAS-TDR-19-ADD-1, 2012, URL: <https://cds.cern.ch/record/1451888>.
- [41] B. Abbott et al., *Production and integration of the ATLAS Insertable B-Layer*, [JINST 13 \(2018\) T05008](#), arXiv: 1803.00844 [[physics.ins-det](#)] (cit. on p. 35).
- [42] J. Pequeno, *Computer generated images of the Pixel, part of the ATLAS inner detector*, CERN-GE-0803013, 2008, URL: <http://cds.cern.ch/record/1095925> (cit. on p. 36).
- [43] C. Marcelloni De Oliveira, *IBL installation into the inner detector of the ATLAS Experiment side C*, ATLAS-PHO-COLLAB-2014-008, 2014, URL: <https://cds.cern.ch/record/1702006> (cit. on p. 36).
- [44] Mikulec, Bettina, *Tests and final integration of the ATLAS semiconductor tracker*, [3 \(2005\) p. 1480](#) (cit. on p. 36).
- [45] A. Bingül, *The ATLAS TRT and its performance at LHC*, [Journal of Physics: Conference Series 347 \(2012\) p. 012025](#) (cit. on p. 37).
- [46] ATLAS Collaboration, *Drift Time Measurement in the ATLAS Liquid Argon Electromagnetic Calorimeter using Cosmic Muons*, [Eur. Phys. J. C 70 \(2010\) p. 755](#), arXiv: 1002.4189 [[hep-ex](#)] (cit. on p. 39).
- [47] ATLAS Collaboration, *Commissioning of the ATLAS Muon Spectrometer with Cosmic Rays*, [Eur. Phys. J. C70 \(2010\) p. 875](#), arXiv: 1006.4384 [[physics.ins-det](#)] (cit. on p. 41).
- [48] ATLAS Collaboration, *Performance of the ATLAS Trigger System in 2015*, [Eur. Phys. J. C77 \(2017\) p. 317](#), arXiv: 1611.09661 [[hep-ex](#)] (cit. on p. 42).
- [49] M. Dobbs and J. B. Hansen, *The HepMC C++ Monte Carlo event record for High Energy Physics*, [Comput. Phys. Commun. 134 \(2001\) p. 41](#) (cit. on p. 43).
- [50] P. Nason, *A new method for combining NLO QCD with shower Monte Carlo algorithms*, [JHEP 11 \(2004\) p. 040](#), arXiv: [hep-ph/0409146](#) (cit. on pp. 43, 89).

- [51] S. Frixione, P. Nason, and C. Oleari, *Matching NLO QCD computations with parton shower simulations: the POWHEG method*, [*JHEP* **11** \(2007\) p. 070](#), arXiv: 0709.2092 [[hep-ph](#)] (cit. on pp. 43, 89).
- [52] S. Alioli, P. Nason, C. Oleari, et al., *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, [*JHEP* **06** \(2010\) p. 043](#), arXiv: 1002.2581 [[hep-ph](#)] (cit. on pp. 43, 89).
- [53] T. Gleisberg et al., *Event generation with SHERPA 1.1*, [*JHEP* **02** \(2009\) p. 007](#), arXiv: 0811.4622 [[hep-ph](#)] (cit. on pp. 43, 127).
- [54] J. Alwall, R. Frederix, S. Frixione, et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, [*JHEP* **07** \(2014\) p. 079](#), arXiv: 1405.0301 [[hep-ph](#)] (cit. on pp. 43, 89).
- [55] S. Höche, F. Krauss, M. Schönherr, et al., *QCD matrix elements + parton showers: The NLO case*, [*JHEP* **04** \(2013\) p. 027](#), arXiv: 1207.5030 [[hep-ph](#)] (cit. on pp. 43, 90).
- [56] T. Sjöstrand, S. Mrenna, and P. Z. Skands, *PYTHIA 6.4 physics and manual*, [*JHEP* **05** \(2006\) p. 026](#), arXiv: [hep-ph/0603175](#) (cit. on pp. 44, 89).
- [57] T. Sjöstrand, S. Mrenna, and P. Z. Skands, *A brief introduction to PYTHIA 8.1*, [*Comput. Phys. Commun.* **178** \(2008\) p. 852](#), arXiv: 0710.3820 [[hep-ph](#)] (cit. on p. 44).
- [58] M. Bahr et al., *Herwig++ physics and manual*, [*Eur. Phys. J. C* **58** \(2008\) p. 639](#), arXiv: 0803.0883 [[hep-ph](#)] (cit. on pp. 44, 90).
- [59] J. Bellm et al., *Herwig 7.0/Herwig++ 3.0 release note*, [*Eur. Phys. J. C* **76** \(2016\) p. 196](#), arXiv: 1512.01178 [[hep-ph](#)] (cit. on p. 44).
- [60] ATLAS Collaboration, *The ATLAS Simulation Infrastructure*, [*Eur. Phys. J. C* **70** \(2010\) p. 823](#), arXiv: 1005.4568 [[physics.ins-det](#)] (cit. on p. 44).
- [61] S. Agostinelli et al., *GEANT4—a simulation toolkit*, [*Nucl. Instrum. Meth. A* **506** \(2003\) p. 250](#) (cit. on pp. 44, 60).
- [62] ATLAS Collaboration, *Reconstruction of primary vertices at the ATLAS experiment in Run 1 proton–proton collisions at the LHC*, [*Eur. Phys. J. C* **77** \(2017\) p. 332](#), arXiv: 1611.10235 [[hep-ex](#)] (cit. on p. 44).
- [63] ATLAS Collaboration, *A neural network clustering algorithm for the ATLAS silicon pixel detector*, [*JINST* **9** \(2014\) P09009](#), arXiv: 1406.7690 [[hep-ex](#)] (cit. on p. 45).
- [64] A. Salzburger, *The ATLAS Track Extrapolation Package*, ATL-SOFT-PUB-2007-005, 2007, URL: <http://cds.cern.ch/record/1038100> (cit. on p. 46).
- [65] ATLAS Collaboration, *Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s} = 13$ TeV*, [*Eur. Phys. J. C* **76** \(2016\) p. 292](#), arXiv: 1603.05598 [[hep-ex](#)] (cit. on pp. 46, 47).
- [66] ATLAS Collaboration, *Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton–proton collision data at $\sqrt{s} = 13$ TeV*, [*Eur. Phys. J.* \(2019\)](#), arXiv: 1902.04655 [[hep-ex](#)] (cit. on p. 48).
- [67] W. Lampl, S. Laplace, D. Lelas, et al., *Calorimeter clustering algorithms: Description and performance*, ATL-LARG-PUB-2008-002, 2008, URL: <https://cds.cern.ch/record/1099735> (cit. on p. 48).

- [68] ATLAS Collaboration, *Electron efficiency measurements with the ATLAS detector using the 2015 LHC proton-proton collision data*, ATLAS-CONF-2016-024, 2016, URL: <https://cds.cern.ch/record/2157687> (cit. on p. 49).
- [69] M. Cacciari, G. P. Salam, and G. Soyez, *The anti- k_t jet clustering algorithm*, **JHEP** **04** (2008) p. 063, arXiv: 0802.1189 [hep-ph] (cit. on p. 49).
- [70] M. Cacciari, G. P. Salam, and G. Soyez, *FastJet User Manual*, **Eur. Phys. J. C** **72** (2012) p. 1896, arXiv: 1111.6097 [hep-ph] (cit. on p. 49).
- [71] ATLAS Collaboration, *Topological cell clustering in the ATLAS calorimeters and its performance in LHC Run 1*, **Eur. Phys. J. C** **77** (2017) p. 490, arXiv: 1603.02934 [hep-ex] (cit. on p. 49).
- [72] ATLAS Collaboration, *Jet energy scale measurements and their systematic uncertainties in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, **Phys. Rev. D** **96** (2017) p. 072002, arXiv: 1703.09665 [hep-ex] (cit. on pp. 50, 130).
- [73] ATLAS Collaboration, *Measurements of b-jet tagging efficiency with the ATLAS detector using $t\bar{t}$ events at $\sqrt{s} = 13$ TeV*, **JHEP** **08** (2018) p. 089, arXiv: 1805.01845 [hep-ex] (cit. on pp. 50, 51, 131).
- [74] ATLAS Collaboration, *Optimisation of the ATLAS b-tagging performance for the 2016 LHC Run*, ATL-PHYS-PUB-2016-012, 2016, URL: <https://cds.cern.ch/record/2160731> (cit. on p. 51).
- [75] ATLAS Collaboration, *Performance of missing transverse momentum reconstruction with the ATLAS detector using proton-proton collisions at $\sqrt{s} = 13$ TeV*, **Eur. Phys. J. C** **78** (2018) p. 903, arXiv: 1802.08168 [hep-ex] (cit. on p. 52).
- [76] *The High-Luminosity LHC Project. 298th Meeting of Scientific Policy Committee*, 2016, URL: <https://cds.cern.ch/record/2199189> (cit. on p. 53).
- [77] F. Zimmermann, *High-energy Physics Strategies and Future Large-scale Projects*, **Nucl. Instrum. Meth. B** **355** (2015) p. 4 (cit. on p. 53).
- [78] *Record luminosity: well done LHC*, 2018, URL: <https://home.cern/news/news/accelerators/record-luminosity-well-done-lhc> (cit. on p. 53).
- [79] *The HL-LHC project*, 2018, URL: <https://hilumilhc.web.cern.ch/about/hl-lhc-project> (cit. on p. 53).
- [80] B. Schmidt, *The High-Luminosity upgrade of the LHC: Physics and Technology Challenges for the Accelerator and the Experiments*, **J. Phys. Conf. Ser.** **706** (2016) p. 022002 (cit. on p. 54).
- [81] ATLAS Collaboration, *Technical Design Report for the ATLAS Inner Tracker Strip Detector*, CERN-LHCC-2017-005, 2017, URL: <https://cds.cern.ch/record/2257755> (cit. on p. 54).
- [82] ATLAS Collaboration, *ATLAS Phase-II Upgrade Scoping Document*, CERN-LHCC-2015-020, 2015, URL: <https://cds.cern.ch/record/2055248> (cit. on p. 54).
- [83] ATLAS Collaboration, *Letter of Intent for the Phase-II Upgrade of the ATLAS Experiment*, CERN-LHCC-2012-022, 2012, URL: <https://cds.cern.ch/record/1502664> (cit. on p. 54).

- [84] ATLAS Collaboration, *Technical Design Report for the ATLAS Inner Tracker Pixel Detector*, CERN-LHCC-2017-021, 2017, URL: <https://cds.cern.ch/record/2285585> (cit. on pp. 54, 56–59, 62, 64, 83, 205).
- [85] A. Ferrari, P. R. Sala, A. Fasso, et al., *FLUKA: A multi-particle transport code (Program version 2005)*, (2005) (cit. on p. 60).
- [86] J. Boudreau and V. Tsulaia, *The GeoModel Toolkit for Detector Description*, (2005), URL: <https://cds.cern.ch/record/865601> (cit. on p. 60).
- [87] S. Baranov, M. Bosman, I. Dawson, et al., *Estimation of Radiation Background, Impact on Detectors, Activation and Shielding Optimization in ATLAS*, ATL-GEN-2005-001, 2005, URL: <https://cds.cern.ch/record/814823> (cit. on p. 61).
- [88] *FLUGG: FLUKA + Geant4 Geometry for Simulation in HEP*, 2003, URL: <http://www.fluka.org/content/tools/flugg/> (cit. on p. 61).
- [89] T. G. C. Zeitnitz, *The GEANT-CALOR Interface User's Guide*, (2001), URL: <https://www.atlas.uni-wuppertal.de/zeitnitz/gcalor> (cit. on p. 61).
- [90] ATLAS Collaboration, *Search for the Standard Model Higgs boson produced in association with top quarks and decaying into $b\bar{b}$ in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, *Eur. Phys. J. C* **75** (2015) p. 349, arXiv: 1503.05066 [hep-ex] (cit. on pp. 85, 103, 137).
- [91] CMS Collaboration, *Search for a standard model Higgs boson produced in association with a top-quark pair and decaying to bottom quarks using a matrix element method*, *Eur. Phys. J. C* **75** (2015) p. 251, arXiv: 1502.02485 [hep-ex] (cit. on pp. 85, 137).
- [92] ATLAS Collaboration, *Search for the Standard Model Higgs boson produced in association with top quarks and decaying into $b\bar{b}$ in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, 2016, URL: <https://cds.cern.ch/record/2194416> (cit. on pp. 86, 105, 137).
- [93] ATLAS Collaboration, *Search for the Standard Model Higgs boson decaying into $b\bar{b}$ produced in association with top quarks decaying hadronically in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, *JHEP* **05** (2016) p. 160, arXiv: 1604.03812 [hep-ex] (cit. on p. 86).
- [94] R. D. Ball et al., *Parton distributions for the LHC Run II*, *JHEP* **04** (2015) p. 040, arXiv: 1410.8849 [hep-ph] (cit. on p. 89).
- [95] ATLAS Collaboration, *ATLAS Pythia 8 tunes to 7 TeV data*, ATL-PHYS-PUB-2014-021, 2014, URL: <https://cds.cern.ch/record/1966419> (cit. on p. 89).
- [96] R. Raitio and W. W. Wada, *Higgs-boson production at large transverse momentum in quantum chromodynamics*, *Phys. Rev. D* **19** (1979) p. 941 (cit. on pp. 89, 130).
- [97] W. Beenakker et al., *NLO QCD corrections to $t\bar{t}H$ production in hadron collisions*, *Nucl. Phys. B* **653** (2003) p. 151, arXiv: hep-ph/0211352 (cit. on pp. 89, 130).
- [98] S. Dawson, C. Jackson, L. H. Orr, et al., *Associated Higgs boson production with top quarks at the CERN Large Hadron Collider: NLO QCD corrections*, *Phys. Rev. D* **68** (2003) p. 034022, arXiv: hep-ph/0305087 (cit. on pp. 89, 130).
- [99] Y. Zhang, W.-G. Ma, R.-Y. Zhang, et al., *QCD NLO and EW NLO corrections to $t\bar{t}H$ production with top quark decays at hadron collider*, *Phys. Lett. B* **738** (2014) p. 1, arXiv: 1407.1110 [hep-ph] (cit. on pp. 89, 130).
- [100] S. Frixione, V. Hirschi, D. Pagani, et al., *Electroweak and QCD corrections to top-pair hadroproduction in association with heavy bosons*, *JHEP* **06** (2015) p. 184, arXiv: 1504.03446 [hep-ph] (cit. on pp. 89, 130).

- [101] J. M. Campbell, R. K. Ellis, P. Nason, et al., *Top-pair production and decay at NLO matched with parton showers*, *JHEP* **04** (2015) p. 114, arXiv: 1412.1828 [hep-ph] (cit. on p. 89).
- [102] M. Czakon and A. Mitov, *Top++: a program for the calculation of the top-pair cross-section at hadron colliders*, *Comput. Phys. Commun.* **185** (2014) p. 2930, arXiv: 1112.5675 [hep-ph] (cit. on p. 89).
- [103] M. Cacciari, M. Czakon, M. Mangano, et al., *Top-pair production at hadron colliders with next-to-next-to-leading logarithmic soft-gluon resummation*, *Phys. Lett. B* **710** (2012) p. 612, arXiv: 1111.5869 [hep-ph] (cit. on p. 89).
- [104] M. Czakon and A. Mitov, *NNLO corrections to top-pair production at hadron colliders: the all-fermionic scattering channels*, *JHEP* **12** (2012) p. 054, arXiv: 1207.0236 [hep-ph] (cit. on p. 89).
- [105] M. Czakon and A. Mitov, *NNLO corrections to top pair production at hadron colliders: the quark-gluon reaction*, *JHEP* **01** (2013) p. 080, arXiv: 1210.6832 [hep-ph] (cit. on p. 89).
- [106] M. Czakon, P. Fiedler, and A. Mitov, *Total Top-Quark Pair-Production Cross Section at Hadron Colliders Through $O(\alpha_S^4)$* , *Phys. Rev. Lett.* **110** (2013) p. 252004, arXiv: 1303.6254 [hep-ph] (cit. on p. 89).
- [107] M. Guzzi, P. Nadolsky, E. Berger, et al., *CT10 parton distributions and other developments in the global QCD analysis*, (2011), arXiv: 1101.0561 [hep-ph] (cit. on p. 89).
- [108] J. Gao, M. Guzzi, J. Huston, et al., *CT10 next-to-next-to-leading order global analysis of QCD*, *Phys. Rev. D* **89** (2014) p. 033009, arXiv: 1302.6246 [hep-ph] (cit. on p. 89).
- [109] P. Z. Skands, *Tuning Monte Carlo generators: The Perugia tunes*, *Phys. Rev. D* **82** (2010) p. 074018, arXiv: 1005.3457 [hep-ph] (cit. on p. 89).
- [110] N. Kidonakis, *Two-loop soft anomalous dimensions for single top quark associated production with a W^- or H^-* , *Phys. Rev. D* **82** (2010) p. 054018, arXiv: 1005.4451 [hep-ph] (cit. on pp. 89, 130).
- [111] N. Kidonakis, *NNLL resummation for s-channel single top quark production*, *Phys. Rev. D* **81** (2010) p. 054028, arXiv: 1001.5034 [hep-ph] (cit. on pp. 89, 130).
- [112] N. Kidonakis, *Next-to-next-to-leading-order collinear and soft gluon corrections for t-channel single top quark production*, *Phys. Rev. D* **83** (2011) p. 091503, arXiv: 1103.2792 [hep-ph] (cit. on pp. 89, 130).
- [113] T. Gleisberg and S. Höche, *Comix, a new matrix element generator*, *JHEP* **12** (2008) p. 039, arXiv: 0808.3674 [hep-ph] (cit. on p. 90).
- [114] S. Schumann and F. Krauss, *A Parton shower algorithm based on Catani-Seymour dipole factorisation*, *JHEP* **03** (2008) p. 038, arXiv: 0709.1027 [hep-ph] (cit. on p. 90).
- [115] ATLAS Collaboration, *Measurement of $W^\pm$ and Z Boson Production Cross Sections in pp Collisions at $\sqrt{s} = 13$ TeV with the ATLAS Detector*, ATLAS-CONF-2015-039, 2015, URL: <https://cds.cern.ch/record/2045487> (cit. on p. 90).
- [116] ATLAS Collaboration, *Multi-boson simulation for 13 TeV ATLAS analyses*, ATL-PHYS-PUB-2016-002, 2016, URL: <https://cds.cern.ch/record/2119986> (cit. on p. 90).
- [117] J. Pumplin, D. R. Stump, J. Huston, et al., *New generation of parton distributions with uncertainties from global QCD analysis*, *JHEP* **07** (2002) p. 012, arXiv: hep-ph/0201195 [hep-ph] (cit. on p. 90).

- [118] ATLAS Collaboration, *Estimation of non-prompt and fake lepton backgrounds in final states with top quarks produced in proton–proton collisions at $\sqrt{s} = 8$ TeV with the ATLAS Detector*, ATLAS-CONF-2014-058, 2014, URL: <https://cds.cern.ch/record/1951336> (cit. on p. 90).
- [119] A. Höcker et al., *TMVA - Toolkit for Multivariate Data Analysis*, 2007, arXiv: [physics/0703039](https://arxiv.org/abs/physics/0703039) [[physics.data-an](https://arxiv.org/abs/physics/0703039)] (cit. on pp. 94, 96).
- [120] L. Breiman, J. H. Friedman, R. A. Olshen, et al., *Classification and Regression Trees*, Statistics/Probability Series, Wadsworth Publishing Company, 1984 (cit. on p. 94).
- [121] Y. Coadou, *Boosted Decision Trees and Applications*, [EPJ Web of Conferences](https://arxiv.org/abs/1303.0004) **55** (2013) p. 02004 (cit. on p. 94).
- [122] H.-J. Yang, B. P. Roe, and J. Zhu, *Studies of boosted decision trees for MiniBooNE particle identification*, [Nucl. Instrum. Meth. A](https://arxiv.org/abs/nucl-instrum-meth-a/0505370) **555** (2005) p. 370 (cit. on p. 94).
- [123] V. M. Abazov et al., *Evidence for production of single top quarks and first direct measurement of $|V_{tb}|$* , [Phys. Rev. Lett.](https://arxiv.org/abs/hep-ex/0612052) **98** (2007) p. 181802, arXiv: [hep-ex/0612052](https://arxiv.org/abs/hep-ex/0612052) [[hep-ex](https://arxiv.org/abs/hep-ex/0612052)] (cit. on p. 94).
- [124] G. Aad, “Mise en service du détecteur à pixels de l’expérience ATLAS auprès du LHC et étude du canal $t\bar{t}H, H \rightarrow b\bar{b}$ pour la recherche du boson de Higgs”, PhD thesis, 2009, URL: <http://cds.cern.ch/record/1233933> (cit. on p. 99).
- [125] V. Barger, J. Ohnemus, and R. Phillips, *Event shape criteria for single lepton top signals*, [Phys. Rev. D](https://arxiv.org/abs/hep-ph/9308216) **48** (1993) p. 3953, arXiv: [hep-ph/9308216](https://arxiv.org/abs/hep-ph/9308216) (cit. on pp. 111, 125, 218, 220).
- [126] ATLAS Collaboration, *Studies on top-quark Monte Carlo modelling with Sherpa and MG5_aMC@NLO*, ATL-PHYS-PUB-2017-007, 2017, URL: <https://cds.cern.ch/record/2261938> (cit. on p. 126).
- [127] F. Cascioli, P. Maierhofer, and S. Pozzorini, *Scattering Amplitudes with Open Loops*, [Phys. Rev. Lett.](https://arxiv.org/abs/hep-ph/1111.5206) **108** (2012) p. 111601, arXiv: [1111.5206](https://arxiv.org/abs/1111.5206) [[hep-ph](https://arxiv.org/abs/hep-ph/1111.5206)] (cit. on p. 127).
- [128] S. Frixione, E. Laenen, P. Motylinski, et al., *Single-top hadroproduction in association with a W boson*, [JHEP](https://arxiv.org/abs/hep-ph/0805.3067) **07** (2008) p. 029, arXiv: [0805.3067](https://arxiv.org/abs/0805.3067) [[hep-ph](https://arxiv.org/abs/hep-ph/0805.3067)] (cit. on p. 130).
- [129] J. M. Campbell and R. K. Ellis, *$t\bar{t}W^\pm$ production and decay at NLO*, [JHEP](https://arxiv.org/abs/hep-ph/1204.5678) **07** (2012) p. 052, arXiv: [1204.5678](https://arxiv.org/abs/1204.5678) [[hep-ph](https://arxiv.org/abs/hep-ph/1204.5678)] (cit. on p. 130).
- [130] ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector at the LHC*, [Eur. Phys. J. C](https://arxiv.org/abs/hep-ex/1608.03953) **76** (2016) p. 653, arXiv: [1608.03953](https://arxiv.org/abs/1608.03953) [[hep-ex](https://arxiv.org/abs/hep-ex/1608.03953)] (cit. on p. 130).
- [131] A. L. Read, *Presentation of search results: The CL_s technique*, [J. Phys. G](https://arxiv.org/abs/hep-ph/0004066) **28** (2002) p. 2693 (cit. on p. 132).
- [132] T. Junk, *Confidence level computation for combining searches with small statistics*, [Nucl. Instrum. Meth. A](https://arxiv.org/abs/hep-ex/9902006) **434** (1999) p. 435, arXiv: [hep-ex/9902006](https://arxiv.org/abs/hep-ex/9902006) (cit. on p. 132).
- [133] G. Cowan, K. Cranmer, E. Gross, et al., *Asymptotic formulae for likelihood-based tests of new physics*, [Eur. Phys. J. C](https://arxiv.org/abs/hep-ph/1007.1727) **71** (2011) p. 1554, [Erratum: [Eur. Phys. J. C](https://arxiv.org/abs/hep-ph/1007.1727) **73** (2013) 2501], arXiv: [1007.1727](https://arxiv.org/abs/1007.1727) [[physics.data-an](https://arxiv.org/abs/physics/1007.1727)] (cit. on p. 135).
- [134] CMS Collaboration, *Search for $t\bar{t}H$ production in the $H \rightarrow b\bar{b}$ decay channel with leptonic $t\bar{t}$ decays in proton–proton collisions at $\sqrt{s} = 13$ TeV*, [JHEP](https://arxiv.org/abs/hep-ex/1804.03682) **03** (2019) p. 026, arXiv: [1804.03682](https://arxiv.org/abs/1804.03682) [[hep-ex](https://arxiv.org/abs/hep-ex/1804.03682)] (cit. on pp. 137, 169).

- [135] M. E. Peters, M. Neumann, M. Iyyer, et al., *Deep contextualized word representations*, CoRR (2018), arXiv: [1802.05365 \[cs\]](#) (cit. on p. 148).
- [136] A. M. Nguyen, J. Yosinski, and J. Clune, *Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images*, CoRR (2014), arXiv: [1412.1897](#) (cit. on p. 148).
- [137] K. Fukushima, *Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position*, *Biological Cybernetics* **36** (1980) p. 193 (cit. on p. 149).
- [138] H. Williams Ronald J., *Learning representations by back-propagating errors*, *Nature* **323** (1986) p. 533 (cit. on p. 149).
- [139] D. H. Ballard, *Modular Learning in Neural Networks*, AAAI’87 (1987) p. 279, URL: <http://dl.acm.org/citation.cfm?id=1863696.1863746> (cit. on p. 149).
- [140] A. Goller C.; Küchler, *Learning task-dependent distributed representations by backpropagation through structure*, **1** (1996) p. 347 (cit. on pp. 149, 176).
- [141] I. Goodfellow, J. Pouget-Abadie, M. Mirza, et al., *Generative Adversarial Nets*, (2014) p. 2672, ed. by Z. Ghahramani, M. Welling, C. Cortes, et al., URL: <http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf> (cit. on pp. 149, 194).
- [142] H. Robbins and S. Monro, *A Stochastic Approximation Method*, *The Annals of Mathematical Statistics* **22** (1951) p. 400, URL: <http://www.jstor.org/stable/2236626> (cit. on p. 150).
- [143] J. Duchi, E. Hazan, and Y. Singer, *Adaptive Subgradient Methods for Online Learning and Stochastic Optimization*, *J. Mach. Learn. Res.* **12** (2011) p. 2121, URL: <http://dl.acm.org/citation.cfm?id=1953048.2021068> (cit. on p. 151).
- [144] D. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization*, *International Conference on Learning Representations* (2014), arXiv: [1412.6980 \[cs\]](#) (cit. on p. 151).
- [145] S. J. Reddi, S. Kale, and S. Kumar, *On the Convergence of Adam and Beyond*, (2018), URL: <https://openreview.net/forum?id=ryQu7f-RZ> (cit. on p. 151).
- [146] V. Nair and G. E. Hinton, *Rectified Linear Units Improve Restricted Boltzmann Machines*, (2010) p. 807, URL: <http://dl.acm.org/citation.cfm?id=3104322.3104425> (cit. on p. 151).
- [147] S. Ioffe and C. Szegedy, *Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift*, CoRR (2015), arXiv: [1502.03167 \[cs\]](#) (cit. on p. 152).
- [148] N. Srivastava, G. Hinton, A. Krizhevsky, et al., *Dropout: A Simple Way to Prevent Neural Networks from Overfitting*, *Journal of Machine Learning Research* **15** (2014) p. 1929, URL: <http://jmlr.org/papers/v15/srivastava14a.html> (cit. on p. 152).
- [149] M. Schuster and K. K. Paliwal, *Bidirectional recurrent neural networks*, *IEEE Transactions on Signal Processing* **45** (1997) p. 2673, ISSN: 1053-587X (cit. on p. 154).
- [150] M. R. Minar and J. Naher, *Recent Advances in Deep Learning: An Overview*, CoRR (2018), arXiv: [1807.08169 \[cs\]](#) (cit. on p. 154).
- [151] S. Hochreiter and J. Schmidhuber, *Long Short-Term Memory*, *Neural Comput.* **9** (1997) p. 1735, ISSN: 0899-7667 (cit. on p. 155).
- [152] J. Chung, Ç. Gülçehre, K. Cho, et al., *Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling*, CoRR (2014), arXiv: [1412.3555 \[cs\]](#) (cit. on p. 155).

- [153] J. S. Bergstra, R. Bardenet, Y. Bengio, et al., *Algorithms for Hyper-Parameter Optimization*, (2011) p. 2546, ed. by J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, et al., URL: <http://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization.pdf> (cit. on p. 157).
- [154] *Anaconda Software Distribution*, 2016, URL: <https://anaconda.com> (cit. on p. 157).
- [155] F. Chollet et al., “Keras”, <https://keras.io>, 2015 (cit. on p. 157).
- [156] Martín Abadi, Ashish Agarwal, Paul Barham, et al., “TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems”, 2015, URL: <http://tensorflow.org/> (cit. on p. 157).
- [157] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research **12** (2011) p. 2825, URL: <https://scikit-learn.org/stable/index.html> (cit. on p. 157).
- [158] *uproot*, 2019, URL: <https://uproot.readthedocs.io/en/latest/index.html> (cit. on p. 157).
- [159] T. Oliphant, *NumPy: A guide to NumPy*, (2006), URL: <http://www.numpy.org/> (cit. on p. 157).
- [160] *Keras + Hyperopt: A very simple wrapper for convenient hyperparameter optimization*, 2019, URL: <https://github.com/maxpumperla/hyperas> (cit. on p. 157).
- [161] *Hyperopt: Distributed Asynchronous Hyper-parameter Optimization*, 2019, URL: <https://github.com/hyperopt/hyperopt> (cit. on p. 157).
- [162] O. Vinyals, S. Bengio, and M. Kudlur, *Order Matters: Sequence to sequence for sets*, arXiv e-prints (2015), arXiv: [1511.06391](https://arxiv.org/abs/1511.06391) [stat.ML] (cit. on p. 161).
- [163] I. Sutskever, O. Vinyals, and Q. V. Le, *Sequence to Sequence Learning with Neural Networks*, arXiv e-prints (2014), arXiv: [1409.3215](https://arxiv.org/abs/1409.3215) [cs.CL] (cit. on p. 161).
- [164] P. Baldi, P. Sadowski, and D. Whiteson, *Searching for Exotic Particles in High-Energy Physics with Deep Learning*, *Nature Commun.* **5** (2014) p. 4308, arXiv: [1402.4735](https://arxiv.org/abs/1402.4735) [hep-ph] (cit. on pp. 174, 176).
- [165] M. Stoye, *Deep learning in jet reconstruction at CMS*, *J. Phys. Conf. Ser.* **1085** (2018) p. 042029 (cit. on p. 174).
- [166] D. Guest, K. Cranmer, and D. Whiteson, *Deep Learning and Its Application to LHC Physics*, *Annual Review of Nuclear and Particle Science* **68** (2018) p. 161 (cit. on p. 176).
- [167] G. Louppe, K. Cho, C. Becot, et al., *QCD-Aware Recursive Neural Networks for Jet Physics*, *JHEP* **01** (2019) p. 057, arXiv: [1702.00748](https://arxiv.org/abs/1702.00748) [hep-ph] (cit. on p. 177).
- [168] T. Cheng, *Recursive Neural Networks in Quark/Gluon Tagging*, *Comput. Softw. Big Sci.* **2** (2018) p. 3, arXiv: [1711.02633](https://arxiv.org/abs/1711.02633) [hep-ph] (cit. on p. 177).
- [169] K. Fraser and M. D. Schwartz, *Jet Charge and Machine Learning*, *JHEP* **10** (2018) p. 093, arXiv: [1803.08066](https://arxiv.org/abs/1803.08066) [hep-ph] (cit. on p. 177).
- [170] S. Egan, W. Fedorko, A. Lister, et al., *Long Short-Term Memory (LSTM) networks with jet constituents for boosted top tagging at the LHC*, (2017), arXiv: [1711.09059](https://arxiv.org/abs/1711.09059) [hep-ex] (cit. on p. 177).
- [171] L. Maaten and G. Hinton, *Visualizing High-Dimensional Data using t-SNE*, Journal of Machine Learning Research **9** (2008) p. 2579, URL: <http://www.jmlr.org/papers/v9/vandermaaten08a.html> (cit. on pp. 179, 180).

- [172] J. Bromley, J. W. Bentz, L. Bottou, et al., *Signature Verification using a "Siamese" Time Delay Neural Network*, [International Journal of Pattern Recognition and Artificial Intelligence](#) **7** (1993) p. 25 (cit. on p. 185).
- [173] C. Adam-Bourdarios, G. Cowan, C. Germain-Renaud, et al., *The Higgs Machine Learning Challenge*, [Journal of Physics: Conference Series](#) **664** (2015) p. 072015 (cit. on p. 193).
- [174] G. Cowan, K. Cranmer, E. Gross, et al., *Asymptotic formulae for likelihood-based tests of new physics*, [Eur. Phys. J. C](#) **71** (2011) p. 1554, arXiv: 1007.1727 [[physics.data-an](#)], Erratum: [Eur. Phys. J. C](#) **73** (2013) 2501 (cit. on p. 193).
- [175] Y. Hong, U. Hwang, J. Yoo, et al., *How Generative Adversarial Nets and its variants Work: An Overview of GAN*, CoRR (2017), arXiv: 1711.05914 [[cs](#)] (cit. on p. 195).
- [176] M. Paganini, L. de Oliveira, and B. Nachman, *Accelerating Science with Generative Adversarial Networks: An Application to 3D Particle Showers in Multilayer Calorimeters*, [Phys. Rev. Lett.](#) **120** (2018) p. 042003, arXiv: 1705.02355 [[hep-ex](#)] (cit. on p. 195).
- [177] ATLAS Collaboration, *Deep generative models for fast shower simulation in ATLAS*, ATL-SOFT-PUB-2018-001, 2018, URL: <http://cds.cern.ch/record/2630433> (cit. on p. 195).
- [178] L. de Oliveira, M. Paganini, and B. Nachman, *Learning Particle Physics by Example: Location-Aware Generative Adversarial Networks for Physics Synthesis*, [Comput. Softw. Big Sci.](#) **1** (2017) p. 4, arXiv: 1701.05927 [[stat.ML](#)] (cit. on p. 195).
- [179] G. Louppe, M. Kagan, and K. Cranmer, *Learning to Pivot with Adversarial Networks*, (2016), arXiv: 1611.01046 [[stat.ME](#)] (cit. on pp. 195–200).

Appendix

Table 0.12: The list of 30 input variables as candidates, from which a reduced input list of the classification BDT is selected for $\geq 6j$ regions, as described in Section 4.5.4.3.

Variable	Definition
General kinematic variables	
Centrality	Scalar sum of the p_T divided by sum of the E for all jets and the lepton
$m_{bb}^{\min \Delta R}$	Mass of the combination of two b -tagged jets with the smallest ΔR
$\Delta R_{bb}^{\max p_T}$	ΔR between the two b -tagged jets with the largest vector sum p_T
p_T^{jet5}	p_T of the fifth leading jet
$H1$	Second Fox–Wolfram moment computed using all jets and the lepton
$\Delta R_{bb}^{\text{avg}}$	Average ΔR for all b -tagged jet pairs
$N_{bb}^{\text{Higgs } 30}$	Number of b -tagged jet pairs with invariant mass within 30 GeV of the Higgs-boson mass
$\Delta \eta_{jj}^{\max}$	Maximum $\Delta \eta$ between any two jets
Aplanarity	$1.5\lambda_2$, where λ_2 is the second eigenvalue of the momentum tensor [125] built with all jets
Aplanarity $_{b\text{-jet}}$	$1.5\lambda_2$, where λ_2 is the second eigenvalue of the momentum tensor [125] built with b -tagged jets
$m_{bj}^{\max p_T}$	Mass of the combination of a b -tagged jet and any jet with the largest vector sum p_T .
$\Delta R_{\ell,bb}^{\min}$	ΔR between the lepton and the combination of the two b -tagged jets with the smallest ΔR
$m_{uu}^{\min \Delta R}$	Mass of the closest in ΔR pair of untagged jets
Variables from likelihood and matrix element method calculations	
LHD	Likelihood discriminant
MEM $_{D1}$	Matrix element discriminant (in $\text{SR}_1^{\geq 6j}$ only)

Table 0.13: Continuation of Table 0.12. All variables in this table are based on reconstruction BDT. The recoBDT_withH is used if not specified.

Variable	Definition
Variables from reconstruction BDT	
BDT output	Output of the recoBDT_withH
m_{bb}^{Higgs}	Higgs candidate mass, using recoBDT
$m_{H,b_{\text{lep top}}}$	Mass of Higgs candidate and b -tagged jet from leptonic top candidate, using recoBDT
$\Delta R_{bb}^{\text{Higgs}}$	ΔR between b -tagged jets from the Higgs candidate, using recoBDT
$\Delta R_{H,t\bar{t}}$	ΔR between Higgs candidate and $t\bar{t}$ candidate system
$\Delta R_{H,\text{lep top}}$	ΔR between Higgs candidate and leptonic top candidate, using recoBDT
$\Delta R_{H,b_{\text{had top}}}$	ΔR between Higgs candidate and b -tagged jet from hadronic top candidate.
$\Delta R_{H,\text{had top}}$	ΔR between Higgs candidate and hadronic top candidate, using recoBDT.
$m_{H,q1}$	Mass of Higgs candidate and the jet with leading p_T from hadronic W candidates, using recoBDT
$m_{H,b_{\text{had top}}}$	Mass of Higgs candidate and the b -tagged jet from hadronic top candidates, using recoBDT
$\Delta R_{b1H,b_{\text{had top}}}$	ΔR between the b -tagged jet with leading p_T from the Higgs candidate and b -tagged jet from the hadronic top decay.
$\Delta R_{b1H,b_{\text{lep top}}}$	ΔR between the b -tagged jet with leading p_T from the Higgs boson and b -tagged jet from the leptonic top decay.
$\Delta R_{b2H,b_{\text{had top}}}$	ΔR between the b -tagged jet with subleading p_T from the Higgs candidate and b -tagged jet from the hadronic top decay.
$\Delta R_{b2H,b_{\text{lep top}}}$	ΔR between the b -tagged jet with subleading p_T from the Higgs boson and b -tagged jet from the leptonic top decay.
$\Delta R_{\text{lep top},\text{had top}}$	ΔR between the leptonic and hadronic top candidates.

Table 0.14: The list of 26 input variables as candidates, from which a reduced input list of the classification BDT is selected for = 5j regions, as described in Section 4.5.4.3.

Variable	Definition
General kinematic variables	
Centrality	Scalar sum of the p_T divided by sum of the E for all jets and the lepton
$m_{bb}^{\min \Delta R}$	Mass of the combination of two b -tagged jets with the smallest ΔR
$\Delta R_{bb}^{\max p_T}$	ΔR between the two b -tagged jets with the largest vector sum p_T
p_T^{jet5}	p_T of the fifth leading jet
$H1$	Second Fox–Wolfram moment computed using all jets and the lepton
$\Delta R_{bb}^{\text{avg}}$	Average ΔR for all b -tagged jet pairs
$N_{bb}^{\text{Higgs } 30}$	Number of b -tagged jet pairs with invariant mass within 30 GeV of the Higgs-boson mass
$\Delta \eta_{jj}^{\max}$	Maximum $\Delta \eta$ between any two jets
Aplanarity	$1.5\lambda_2$, where λ_2 is the second eigenvalue of the momentum tensor [125] built with all jets
Aplanarity _{b-jet}	$1.5\lambda_2$, where λ_2 is the second eigenvalue of the momentum tensor [125] built with b -tagged jets
$m_{bj}^{\max p_T}$	Mass of the combination of a b -tagged jet and any jet with the largest vector sum p_T .
$\Delta R_{\ell,bb}^{\min}$	ΔR between the lepton and the combination of the two b -tagged jets with the smallest ΔR
$m_{jj}^{\min \Delta R}$	Mass of the combination of any two jets with the smallest ΔR
N_{40}^{jet}	Number of jets with $p_T \geq 40$ GeV
$m_{jj}^{\max p_T}$	Mass of the combination between any two jets with the largest vector sum p_T
H_T^{had}	Scalar sum of jet p_T
Variables from likelihood calculations	
LHD	Likelihood discriminant

Table 0.15: Continuation of Table 0.14. All variables in this table are based on reconstruction BDT. The recoBDT_withH is used if not specified.

Variable	Definition
Variables from reconstruction BDT	
BDT output	Output of the recoBDT_withH
m_{bb}^{Higgs}	Higgs candidate mass, using recoBDT
$m_{H,b_{\text{lep top}}}$	Mass of Higgs candidate and b -tagged jet from leptonic top candidate, using recoBDT
$\Delta R_{bb}^{\text{Higgs}}$	ΔR between b -tagged jets from the Higgs candidate, using recoBDT
$\Delta R_{H,t\bar{t}}$	ΔR between Higgs candidate and $t\bar{t}$ candidate system
$\Delta R_{H,\text{lep top}}$	ΔR between Higgs candidate and leptonic top candidate, using recoBDT
$\Delta R_{H,b_{\text{had top}}}$	ΔR between Higgs candidate and b -tagged jet from hadronic top candidate.
$\Delta R_{H,\text{had top}}$	ΔR between Higgs candidate and hadronic top candidate, using recoBDT.
$\Delta R_{\text{lep top,had top}}$	ΔR between the leptonic and hadronic top candidates.

Glossary of physics terms for non-physicists

This part gives a glossary of physics jargon for non-physicists. At the end of each term, the first page number (if more than one page is there) is where they are firstly introduced in this thesis. And the second page corresponds to the context of Section 5.1 where the term is illustrated for non-physicists.

pile-up The effect of multiple proton–proton interactions in the same and neighboring bunch crossings. 29, 145

event A crossing of two bunches of protons recorded in the detector. This corresponds to a sample in a dataset in the machine learning field. 145

sample A set of events. 145

decay A physics process through which an unstable particle annihilates and produces other particles. 145

final state Particles produced from a bunch crossing may be unstable and decay before reaching the detector. Final state means collectively surviving particles that are decays from production in a bunch crossing, and their properties will be measured by the detector. 145

jet A cluster of collimated particles evolving from a quark or a gluon. Jets issued from different quarks have special properties and can be identified, or tagged. *b*-jets are jets initiated by *b*-quarks, *c*-jets by *c*-quarks, and light jets are issued from light quarks (*u*-, *d*-, *s*-quark) or gluons. Jets are important final state objects. 19, 145

***b*-tagging** The identification of *b*-jets against *c*-jets and light-jets. 50, 145

***b*-tagging weight** The output score of a machine learning model which is trained on a specific sample consisting of *b*-, *c*- and light-jets to identify *b*-jets. The score represents how likely a jet is from a *b*-quark. A jet is tagged as *b*-jet if its *b*-tagging weight passes a certain threshold. 51, 145

event number An index of events. 145

low-level features The final state object features that are directly measured from the detector, including the energy E , and the momentum in three Euclidean spatial directions $\vec{p} = (p_x, p_y, p_z)$ or equivalently η, ϕ and p_T (see Section 2.2.1). 145

raw features Same as low-level features. 145

high-level features Variables as functions of low-level features. They are calculated with physics expertise aimed to be discriminating input features for a classifier. 145

signal events Events from interesting physics processes to be precisely measured, or from unknown processes to be searched for according to certain theories. They correspond to positive samples in the machine learning field. In this analysis, it is the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) process. 145

backgrounds Events having similar final state or properties to signal events. They are known processes but usually have much larger production rate than signal. 146

b -tagging working point Also referred to as b -tagging WP. These are different thresholds on the b -tagging weight: 85% WP is a loose criterion, meaning the efficiency to correctly identify a jet from a b -quark as a b -jet is 85% (truth positive rate). Correspondingly a c -jet (light-jet) is mistaken for a b -jet once for every 3.1 (33) jets (false positive rate). There are four WPs: 85%, 77%, 70% and 60%, with decreasing truth positive rate, and faster decreasing false positive rate as well. For each of these WP, the b -tagging efficiency in simulation is corrected to match data. 51, 145

pseudo-continuous (PC) b -tagging The full-range b -tagging score but split into five bins defined by the four WPs. It is calibrated bin by bin for each WP. 51, 145

cross-section A measure of the probability for a process to occur during any proton–proton collision, proportional to the corresponding event counting. 146

$t\bar{t} + \text{jets}$ The main background in the $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis, where a top-quark pair is produced accompanied by additional jets (see Figure 4.1(c)), but without Higgs boson production. 146

$t\bar{t} + \geq 1b$ A sub-component (subset) of the $t\bar{t} + \text{jets}$ background, where at least one additional jet is initiated by a b -quark. 146

$t\bar{t} + \geq 1c$ A sub-component (subset) of the $t\bar{t} + \text{jets}$ background, where at least one additional jet is initiated by a c -quark and no jet was initiated by a b -quark. 146

$t\bar{t} + \text{light}$ A sub-component (subset) of the $t\bar{t} + \text{jets}$ background, where at least one additional jet is initiated by a light-quark and no jet was initiated by a b -quark. 146

data Events with features that are collected from experiments. In data, the event truth label is not known. 145

Monte Carlo (MC) simulation Events produced from theoretical calculations and proceeded to mimic how they would look like in the detector if they were produced in real collisions. This is also referred to as prediction. The event truth label is known. [42](#), [146](#)

class The truth label of events, it can be $t\bar{t}H$ ($H \rightarrow b\bar{b}$) signal, $t\bar{t}$ background, or $t\bar{t}$ sub-components. [146](#)

statistics The size of a sample. A sample with large number of events is referred to as having large statistics. [146](#)

nominal sample The simulated sample of a particular process considered as the best description. [88](#), [146](#)

systematic sample Systematic samples are also simulated samples, produced as estimation of systematic uncertainties. They represent variations on the nominal sample, accounting for the various systematic effects. [126](#), [146](#)

systematic uncertainties All uncertainties that are not directly due to the statistics of the data. They arise from uncertainties associated with the nature of the measurement apparatus, statistical uncertainty of simulated samples due to their limited size, and uncertainties induced by imprecise theoretical predictions. [126](#), [146](#)

statistical uncertainty An estimation of statistical fluctuation of data counting, arising from the fact that the number of collected data are finite. [146](#)

signal strength The ratio of the observed number of signal events over the one from theoretical prediction. It is referred to as μ . [147](#)

significance A measure of how unusual the data are if the background only hypothesis is true. It is expressed in terms of standard deviation σ . The higher the number of σ , the more incompatible the data are with the background only hypothesis. 5σ is the requirement to claim a discovery. It means the probability that the observed signal is faked by a background fluctuation is 2.87×10^{-7} , which is the area above the quantile that is five standard deviations larger than the mean of a Gaussian distribution. [147](#)

regions Orthogonal categories splitting all events used for a physics analysis based on some features. They are also referred to as categories. In the baseline $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis, the categorization is based on the multiplicity of jets and b -tagging weights of jets. [90](#), [147](#)

signal regions (SRs) The regions that are enriched in signal events, and also referred to as SR. In the baseline $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis studied in this analysis (focusing on events having exactly one lepton), there are three signal regions for events having at least six jets, and two signal regions for events with exactly five jets. [90](#), [147](#)

ultra pure signal region The region with purest signal events, also referred to as UPSR or SR1. In the baseline $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis, there are two ultra pure signal regions, one for events with at least 6 jets and one for events with exactly 5 jets. These two UPSRs require at least four b -jets tagged at 60%. [91](#), [147](#)

control regions (CRs) Regions that are formed in order to have enhanced purity of background events. In the baseline $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis, there are three control regions for events with at least six jets and for events with exactly five jets, enriched in $t\bar{t} + \geq 1c$, $t\bar{t} + \text{light}$, and $t\bar{t} + \geq 1b$ respectively.. [90](#), [147](#)

6ji4bi@85% Events with at least 6 jets among which at least 4 are tagged as b -jets at the working point of 85%. This is the full training sample of the machine learning model for ≥ 6 jets events, on which the so-called inclusive training is implemented (see Section [4.5.3.4](#)). [147](#)

5je4bi@85% Events with exclusively 5 jets among which at least 4 are tagged as b -jets at the working point of 85%. This is the full training sample of the machine learning model for events with exclusively 5 jets, on which the so-called inclusive training is implemented (see Section [4.5.3.4](#)). [147](#)

multi-variate analysis (MVA) Classification or regression methods based on machine learning techniques. These methods have become a fundamental ingredient in many high energy physics analyses. [147](#)

event reconstruction The procedure to correctly associate final state jets to quarks or gluons that initiated them. The jet-parton association is also referred to as combination, or matching. [147](#)

reconstruction BDT The boosted decision trees model used to solve jet-parton association for $t\bar{t}H$ ($H \rightarrow b\bar{b}$) events, in the default $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis. Two reconstruction BDT are trained: the one using the Higgs boson related information as additional input is referred to as recoBDT_withH, and another without using Higgs information is referred to as recoBDT. [99](#), [147](#)

classification BDT The boosted decision trees model used to solve the classification between $t\bar{t}H$ ($H \rightarrow b\bar{b}$) signal and $t\bar{t} + \text{jets}$ background, in the baseline $t\bar{t}H$ ($H \rightarrow b\bar{b}$) analysis. [109](#), [148](#)