University of Southampton

Faculty of Engineering and Physical Sciences
School of Physics and Astronomy
Southampton High Energy Physics Group

Revisiting New and Old Jet Clustering Algorithms for Beyond the Standard Model Higgs Searches in the Final States with b-jets

by

Shubhani Jain

Masters in Theoretical Physics

ORCiD: 0000-0002-0964-879X

A thesis for the degree of
Doctor of Philosophy

December 2023

University of Southampton

Abstract

Faculty of Engineering and Physical Sciences
School of Physics and Astronomy
Southampton High Energy Physics Group

Doctor of Philosophy

Revisiting New and Old Jet Clustering Algorithms for Beyond the Standard Model Higgs Searches in the Final States with b-jets

by Shubhani Jain

The search for novel physics Beyond the Standard Model (BSM) continues to be elusive despite the Large Hadron Collider's (LHC) many triumphs since its inception in 2008. The ultimate aim of this work is to address this issue and search for new physics using the simplest extended Higgs sector framework, the 2-Higgs Doublet Model (2HDM), manifested in cascade decays with high multiplicity b-jet final states wherever kinematically possible.

In this thesis, we compare different jet clustering algorithms to fully resolve hadronic b-jet final states arising from a decay chain of a heavy CP-even Higgs H into a pair of the lighter Higgs bosons h. We consider both scenarios where $m_H > m_h = 125$ GeV and $m_H = 125$ GeV $> m_h$ for the 2HDM Type-II framework. We provide the ideal choice of acceptance cuts, resolution parameters and reconstruction procedures in order to enhance the significance ratios and establish such a ubiquitous BSM signal using the 2HDM Type-II framework.

Furthermore, we examine the potential of detecting a cross-section at the High-Luminosity phase of the LHC (HL-LHC) for the production of SM-like h in association with a single top quark. For the illustrative example of $bg \to twh$ with $h \to b\bar{b}$ final state, the permissible benchmark points in the 2HDM Type-II are shown to yield better significance rates and distinct kinematical distributions with respect to the SM, allowing the signal to be observed at the HL-LHC.

Finally, we employ the machine learning method of image recognition to design a Convolutional Neural Network (CNN) to classify the double b-tagged fatjet final states emerging from a 2HDM Type-II signal against the leading backgrounds.

Contents

List of Figures

List of Tables

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as: [1], [2], and [3].

Signed:..... Date:.....

Acknowledgements

There are numerous people I would like to thank for their assistance and constant support during my postgraduate studies.

To begin, I want to express my gratitude to my supervisory team, Prof. Stefano Moretti and Dr. Srinandan Dasmahapatra, for their professional support, as well as the entire ML Collaboration group involved in our research: Amit Chakraborty, Giorgio Cerro, Jacan Chaplais, Henry Day-Hall, Billy Ford, Keiren Maguire, Emmanuel Olaiya, and Claire Shepherd-Themistocleous.

I am extremely fortunate to have worked with so many amazing people during my time as a PhD student, especially Alessandro Barone, Mauricio Diaz, Maria Georganti, Jack Mitchell, Raj Mukherjee, Vlad Mandric, Giovanna Salvi, Dalius Stulga, Michele Santagata, Matthew Ward, and countless others. I appreciate all the chats, lunches, card games, and Team calls during the lockdown.

Away from the university, I express my deepest gratitude to my Southampton family. Thanks to the Flat 29 girls for being a continual source of laughter and introducing me to the Great British Bake Off. To Samyak, Krishna, Hemangi, and Sukrity for all the nonsensical physics questions, living room dances, dumb charades, and, of course, for making Southampton enjoyable!

I am deeply indebted for the support I have received from my joint family. Thank you for your constant support and words of encouragement.

Finally, to my parents and sister, without your unwavering love, strength, support, counselling, and understanding, nothing would have been possible. From Haldwani to Edinburgh and finally Southampton, you have been by my side for every journey, and I eagerly await our next family adventure!

In loving memory of Amma, Baba, Nani, and Nana!

Part I

Introduction and Background Theory

Chapter 1

Introduction

The goal of particle physics is to explore hidden corners of the universe and study the nature of the fundamental forces and elements that hold matter together. The idea that matter is made up of smaller particles can be traced back to the 6th century BC. It was in the 19th century when these philosophical hypotheses were transformed into scientific reality, resulting in an onslaught of particle discoveries, right from the discovery of the electron in 1897 to the Higgs boson in 2012 at the Large Hadron Collider (LHC).

Our best understanding of the fundamental interactions of the discovered particles is given by an elegant string of equations embodied in the Standard Model (SM), which was developed in the 1960s by Sheldon L. Glashow, Steven Weinberg, and Abdus Salam. Despite its many triumphs, the SM has its well-known theoretical flaws and fails to explain many experimental data. The pursuit for a unified field theory still continues, with active research topics spanning from encapsulating neutrino mass to explaining dark matter, the hierarchy problem, and string theory.

The ultimate goal of this thesis is to address the inadequacy of the SM by searching for physics Beyond the SM (BSM). To achieve this, we employ the most straightforward extension of the SM Higgs sector, namely, the 2-Higgs Doublet Model (2HDM). The reason behind using the 2HDM is that it allows for additional Higgs boson states manifested in cascade decays. These states can be detected at the LHC through the dominant $b\bar{b}$ decay channel wherever kinematically possible. In particular, we determine whether different jet clustering algorithms, with different resolution and reconstruction parameters, might be more or less suited to disentangle such fully hadronic final states, specifically for topologies derived from the 2HDM. Furthermore, we use the 2HDM to establish the production of the SM-like

Higgs boson in association with a single top quark as one of the primary methods of production at the LHC.

The layout of the thesis is as follows. Chapter 2 provides an overview of the functioning and shortcomings of the SM, with a focus on the Higgs sector. In Chapter 3, we briefly summarise the phenomenology of the 2HDMs. In Chapter 4, we review jet physics and its relevance in the context of the LHC.

Chapters 5-7 present the results published in [1], [2], and [3]. Chapter 5, based on [1], investigates the performance of different jet-clustering algorithms, in particular, we compare the variable-R algorithm with the traditional fixed cone algorithms in the context of potential 2HDM searches.

In Chapter 6, based on [2], we compare different jet-clustering algorithms for producing fully hadronic final states, particularly in events leading to boosted topologies where particles tend to merge into a single, fat jet. This chapter aims to determine the best clustering method for establishing such a ubiquitous BSM signal using a 2HDM Type-II scenario.

In Chapter 7, based on [3], we examine the potential of establishing detectable 2HDM Type-II cross-sections at the High-Luminosity phase of the LHC (HL-LHC) for the production of the SM-like Higgs boson (h) in association with a single top (anti)quark in the 'wrong-sign solution' scenario of the bottom (anti)quark Yukawa coupling.

In Chapter 8, we discuss ongoing research on jet visualisation techniques, specifically the development of a classifier using Convolutional Neural Networks (CNNs) to differentiate the 2HDM signals from the leading backgrounds when represented as jet images in detector (η, ϕ) space.

Finally, Chapter 9 presents our conclusions for various research projects and discusses potential future works for the LHC iterations to come.

Chapter 2

The Standard Model

The SM has been a reigning champion of theoretical particle physics offering the most comprehensive description of all fundamental forces and their associated particles. It is a particular type of renormalisable Quantum Field Theory (QFT) with a local gauge $SU(3)_C \times SU(2)_L \times U(1)_Y$ symmetry group, corresponding to three fundamental forces: Quantum Chromodynamics (QCD), weak interactions and, Quantum Electro-Dynamics (QED). The particle content of the SM is detailed in Tab. 2.1. In this chapter, we will cover the Higgs and QCD sectors of the SM, which are relevant to this thesis.

Gauge Bosons	Elementary Fermions				Scalar Bosons
	Quarks		Leptons		
	Charge: 2/3	Charge: -1/3	Charge: -1	Neutral	
γ W^+ W^- Z^0 g	u c t	d s b	e^- μ^- τ^-	ν_e ν_μ ν_τ	h

TABLE 2.1: Particle content of the SM.

2.1 The Higgs Mechanism

Within the SM, the unified theory of Electro-Magnetic (EM) and weak interactions, commonly known as Electro-Weak (EW) theory, is characterised by a

symmetric Lagrangian under local weak isospin and hypercharge gauge transformations. Although the EW theory is renormalisable, it contradicts experimentally verified fermions and gauge bosons masses. The reason is that once the masses are included, the Lagrangian density will no longer be gauge invariant and thus becomes non-renormalisable.

To address this issue and establish a correct EW theory, the Higgs Mechanism was introduced by Peter Higgs [4, 5], François Englert, and Robert Brout [6]. The mechanism breaks the symmetry spontaneously, allowing the generation of masses while preserving the gauge invariance of the Lagrangian density. In 1967, Steven Weinberg utilised the concept of Spontaneous Symmetry Breaking (SSB) and the Higgs Mechanism to explain the origin of weak gauge boson masses (except the photon, which remains massless) [7]. This groundbreaking work was independently carried out by Abdus Salam in 1968 [8].

2.1.1 Spontaneous Symmetry Breaking

To begin, we first consider a Lagrangian that describes a single complex scalar field ϕ and can be expressed as [9]

$$\mathcal{L}_H = \partial_\mu \phi^*(x) \partial^\mu \phi(x) - V(\phi), \tag{2.1}$$

with the scalar potential defined as

$$V(\phi) = \mu^2 |\phi(x)|^2 + \lambda |\phi(x)|^4. \tag{2.2}$$

This Lagrangian has a global $U(1)$ invariance under gauge transformation

$$\phi(x) \to \phi'(x) = e^{i\alpha} \phi(x), \tag{2.3}$$

$$\phi^*(x) \to \phi^{*\prime}(x) = e^{-i\alpha} \phi^*(x), \tag{2.4}$$

where α is some phase rotational angle.

When we minimise the potential $V(\phi)$, we get two possible minima depending on the sign of μ^2. If $\mu^2 > 0$, a single vacuum state emerges, with $V(\phi)$ having an absolute minimum at $\phi = 0$. In this case, the global symmetry of the Lagrangian remains preserved. However, if $\mu^2 < 0$, the potential assumes the shape of a "Mexican Hat", illustrated in Fig. 2.1. The geometry of the potential is such that there is no longer an absolute minimum at $\phi = 0$, but rather a whole circle of

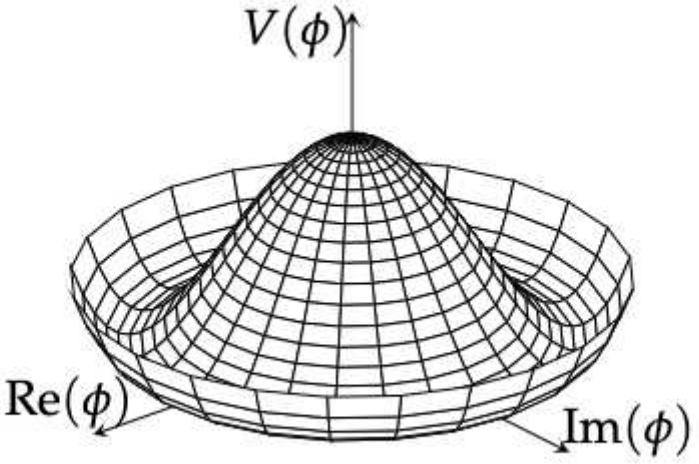


FIGURE 2.1: The 'Mexican Hat' shaped Higgs potential when $\mu^2 < 0$.

absolute minima. In that event, the global $U(1)$ symmetry is non-linearly realised and the system undergoes SSB. We, therefore, select a particular ground state with a Vacuum Expectation Value (VEV) of ϕ given by

$$\langle 0|\phi|0\rangle = \sqrt{\frac{-\mu^2}{2\lambda}} = \frac{v}{\sqrt{2}} > 0. \tag{2.5}$$

It is important to note that no perturbation computation is now allowed around the unstable point $\phi = 0$. To rectify this situation, we rewrite the complex scalar field $\phi(x)$ as

$$\phi(x) = \frac{1}{\sqrt{2}}\Big(v + h(x) + i\xi(x)\Big), \tag{2.6}$$

where $h(x)$ and $\xi(x)$ are real fields. The Lagrangian density then becomes

$$\begin{aligned}\mathcal{L}_H = \frac{1}{2}\partial^\mu h(x)\partial_\mu h(x) + \frac{1}{2}\partial^\mu \xi(x)\partial_\mu \xi(x) - \frac{\mu^2}{2}h^2 - \frac{\lambda}{2}\Big[(v+h)^2 + \xi^2\Big] \\ - \mu^2 v h - \frac{\mu^2}{2}\xi^2 - \frac{1}{2}\mu^2 v^2.\end{aligned} \tag{2.7}$$

It is clear from the above equation that $h(x)$ is a real and massive boson, whereas $\xi(x)$ corresponds to a massless boson, commonly known as a Goldstone boson. The

existence of a Goldstone boson can be traced back to SSB of the global symmetry [10].

Next, we move on to investigate the concept of SSB in the presence of a local $U(1)$ gauge field. We first replace the partial derivative with a covariant derivative [11]

$$D_\mu = \partial_\mu + ie\mathcal{A}_\mu, \tag{2.8}$$

where the gauge field $\mathcal{A}_\mu$ is introduced to ensure that the Lagrangian density is invariant under a local $U(1)$ gauge transformation. The Lagrangian density can be re-written as

$$\mathcal{L}_H = (D^\mu\phi^*)(D_\mu\phi) - \mu^2|\phi|^2 - \lambda|\phi|^4. \tag{2.9}$$

We then minimise the potential using Eq.(2.6) and the Lagrangian density in Eq.(2.9) becomes

$$\begin{aligned}\mathcal{L}_H = &\left(\frac{1}{2}\Big[(\partial^\mu h)(\partial_\mu h) - \mu^2 h^2\Big] + \frac{1}{2}(\partial^\mu\xi)(\partial_\mu\xi) + \frac{1}{2}e^2v^2\mathcal{A}^\mu\mathcal{A}_\mu\right) + ve^2\mathcal{A}^\mu\mathcal{A}_\mu h \\ &+ \frac{e^2}{2}\mathcal{A}^\mu\mathcal{A}_\mu h^2 + e(\partial^\mu\xi)\mathcal{A}_\mu(v+h) - e(\partial^\mu h)\mathcal{A}_\mu\xi - \mu^2 vh - \frac{\mu^2}{2}\xi^2 - \frac{\mu^2 v}{2} \\ &- \frac{\lambda}{2}\Big[(v+h) + \xi^2\Big]^2.\end{aligned} \tag{2.10}$$

The first three terms in Eq.(2.10) correspond to a Klein-Gordon field that describes the spin-0 particle $h(x)$ with mass $\sqrt{-\mu^2}$ and a massless Goldstone boson $\xi(x)$. The fourth term describes the free $U(1)$ gauge field. All other terms in Eq.(2.10) represent interactions between the fields. To understand the correct interpretation of Eq.(2.10), we further introduce

$$\mathcal{A}_\mu \to \mathcal{A}'_\mu = \mathcal{A}_\mu + \frac{1}{ev}\partial_\mu\xi, \tag{2.11}$$

with local unitary gauge transformation

$$\phi \to \phi' = e^{-i\xi(x)/v}\phi = \frac{(v + h(x))}{\sqrt{2}}. \tag{2.12}$$

The simplified version of the Lagrangian is given by

$$\mathcal{L}_H = \frac{1}{2}\Big[(\partial^\mu h)(\partial_\mu h) - \mu^2 h^2\Big] + \frac{1}{2}e^2 v^2 \mathcal{A}^{\mu\prime}\mathcal{A}_\mu{}' + ... \tag{2.13}$$

The $\xi(x)$ term disappears from the Lagrangian and can be interpreted as being absorbed by the $\mathcal{A}_\mu$ field to generate mass. This mechanism makes the Higgs field $h(x)$ and gauge field $\mathcal{A}_\mu{}'$ physical!

Moving on to the SM, let us have a look at how SSB is theorised for $SU(2)_L \times U(1)_Y$ gauge group. The Higgs Lagrangian is written as [12]

$$\mathcal{L}_H = (D^\mu\phi)^\dagger(D_\mu\phi) - \mu^2\phi^\dagger\phi - \lambda(\phi^\dagger\phi)^2, \tag{2.14}$$

where the $SU(2)$ complex doublet ϕ is given by

$$\phi = \begin{pmatrix}\phi^+\\ \phi^0\end{pmatrix}, \tag{2.15}$$

and the covariant derivative by, $D_\mu = \partial_\mu + i\frac{g'}{2}\mathcal{A}_\mu Y + i\frac{g}{2}\tau.\mathcal{W}_\mu$, where τ are the Pauli isospin matrices. The Higgs potential is similar to Eq.(2.2)

$$V(\phi^\dagger\phi) = \mu^2(\phi^\dagger\phi)^2 + \lambda(\phi^\dagger\phi)^2. \tag{2.16}$$

The same reasoning as before applies for both $\mu^2 > 0$ and $\mu^2 < 0$ cases: minimising the potential again gives

$$\big\langle\phi\big\rangle_0 = \begin{pmatrix}0\\ v/\sqrt{2}\end{pmatrix}, \tag{2.17}$$

with $v = \sqrt{-\mu^2/\lambda}$. This choice of non-zero VEV spontaneously breaks the $SU(2)_L$ and $U(1)_Y$ gauge symmetries to ensure $U(1)_{\rm EM}$ remains an exact symmetry. Also, the theory will contain three Goldstone bosons since three out of four generators are broken spontaneously due to the choice of VEV. To understand how these three Goldstone bosons are absorbed by gauge bosons, we rewrite our complex scalar field $\phi(x)$ as

$$\phi = e^{\frac{i\tau^i\xi^i}{2v}}\begin{pmatrix}0\\ \frac{v+h}{\sqrt{2}}\end{pmatrix}, \tag{2.18}$$

where $\xi^i(x)$ $(i = 1, 2, 3)$ and $h(x)$ are four real fields.

Performing a unitary gauge transformation, one can write

$$\phi \to \phi' = \begin{pmatrix} 0 \\ \frac{v+h(x)}{\sqrt{2}} \end{pmatrix}, \tag{2.19}$$

with $h(x)$ being identified as a physical Higgs field.

2.1.2 Gauge Boson and Fermion Masses

We begin by expanding $(D^\mu\phi)^\dagger(D_\mu\phi)$ using Eq.(2.19) to see the effect of SSB on $\mathcal{L}_H$ [13]

$$(D^\mu\phi)^\dagger(D_\mu\phi) = \frac{1}{2}(\partial_\mu h)(\partial^\mu h) + \frac{g^2}{8}(v+h)^2|\mathcal{W}^1_\mu - i\mathcal{W}^2_\mu|^2 + \frac{1}{8}(v+h)^2(g'\mathcal{A}_\mu - g\mathcal{W}^3_\mu)^2. \tag{2.20}$$

To simplify the equation we define the charge gauge fields as linear combinations of the massless $\mathcal{W}^1_\mu$ and $\mathcal{W}^2_\mu$:

$$W^\pm_\mu = \frac{\mathcal{W}^1_\mu \mp i\mathcal{W}^2_\mu}{\sqrt{2}}, \tag{2.21}$$

The superscript $\pm$ on $W^\pm_\mu$ represents the electric charge of the gauge boson. Furthermore, due to the charge conjugation matrix, the gauge field should transform into minus its transpose, which sends $W^\pm_\mu \to W^\mp_\mu$, indicating that the particle is sent into its antiparticle.

Similarly, the neutral gauge boson eigenstates

$$Z_\mu = \frac{-g'\mathcal{A}_\mu + g\mathcal{W}^3_\mu}{\sqrt{g^2+g'^2}}, \tag{2.22}$$

$$A_\mu = \frac{g\mathcal{A}_\mu + g'\mathcal{W}^3_\mu}{\sqrt{g^2+g'^2}}. \tag{2.23}$$

Plugging Eq.(2.21), (2.22), and (2.23) into Lagrangian density $\mathcal{L}_H$, we get:

$$\mathcal{L}_H = \left[\frac{1}{2}(\partial^\mu h)(\partial_\mu h) - \frac{\mu^2}{2}h^2\right] + \frac{v^2g^2}{8}W^{+\mu}W^+_\mu + \frac{v^2g^2}{8}W^{-\mu}W^-_\mu + \frac{(g^2+g'^2)v^2}{8}Z^\mu Z_\mu + ... \tag{2.24}$$

From the above equation, it is clear that $\mathcal{A}_\mu$ is massless and can be identified as the photon (γ) with $M_\gamma = 0$. The second term in Eq.(2.24) corresponds to the mass term for the Higgs boson

$$M_h = -2\mu^2. \tag{2.25}$$

The remaining terms in Eq.(2.24) correspond to mass terms for $W^\pm$ and Z^0 and can be read off as

$$M_{W^\pm} = \frac{vg}{2}, \qquad M_{Z^0} = \frac{v}{2}\sqrt{g^2 + g'^2}. \tag{2.26}$$

To see how fermions acquire mass due to SSB, we first consider the electron and its neutrino as an example. The Lagrangian density in the unitary gauge for this case is given by

$$\begin{aligned} \mathcal{L}_e &= -y_e \left[\bar{e}_r \ \phi^\dagger \begin{pmatrix} \nu_l \\ e_l \end{pmatrix} + (\bar{\nu}_l \ \bar{e}_l)\phi \ e_r \right] \\ &= -y_e \frac{(v+h)}{\sqrt{2}} (\bar{e}_r e_l + \bar{e}_l e_r) = -y_e \frac{(v+h)}{\sqrt{2}} \bar{e}e \end{aligned} , \tag{2.27}$$

where ν_l and e_l are left handed neutrino and lepton respectively, e_r is right handed lepton, $\bar{e} \equiv (\bar{e}_r, \bar{e}_l)$ and $e \equiv (e_l, e_r)^T$. The mass of the electron can be read as

$$M_e = y_e v/\sqrt{2}. \tag{2.28}$$

Similarly, for the case of up and down quarks, the Lagrangian densities can be written as [14]

$$\mathcal{L}_u = -\lambda_u \bar{Q}_l \phi^C u_r + h.c. \ , \tag{2.29}$$

$$\mathcal{L}_d = -\lambda_d \bar{Q}_l \phi d_r + h.c. \tag{2.30}$$

which gives rise to masses for up and down quarks

$$M_u = \frac{v\lambda_u}{\sqrt{2}}, \qquad M_d = \frac{v\lambda_d}{\sqrt{2}}. \tag{2.31}$$

2.2 QCD

QCD is a non-Abelian gauge theory of strong interactions modeled by $SU(N_C)$ gauge symmetry, with $N_C = 3$ and C representing the colours. The N_C corresponds to the number of colours characterised by red, green and blue respectively. In other words, the theory consists of 3×3 unitary matrices and a unit determinant. It has $(3)^2 - 1 = 8$ different generators corresponding to 8 independent directions in the matrix space. The wave function of the theory is a triplet and is given as [9, 15]

$$\psi_q = (\psi_{qR}, \psi_{qG}, \psi_{qB})^T, \tag{2.32}$$

with gauge transformations

$$\begin{aligned} \psi_q(x) \to \psi'_q(x) &= \mathcal{U}\psi_q(x), \\ \bar{\psi}_q(x) \to \bar{\psi}'_q(x) &= \bar{\psi}_q(x)\mathcal{U}^\dagger, \end{aligned} \tag{2.33}$$

where $\mathcal{U}(\theta) = e^{-ig_s\theta^a t_a}$ is the unitary gauge matrix with g_s, θ^a and t_a representing the coupling constant, constant parameters, and generators of the $SU(3)_C$, respectively. The generators t_a satisfy the commutation relation

$$[t_a, t_b] = if_{abc}t_c, \tag{2.34}$$

where f_{abc} is the structure constant.

2.2.1 Lagrangian density of QCD

The Lagrangian density of the theory is given by [16]

$$\mathcal{L}_{\text{QCD}} = \sum_q \bar{\psi}^i_q(i\not{D} - m)_{ij}\psi^j_q - \frac{1}{4}F^a_{\mu\nu}F^{a\mu\nu}, \tag{2.35}$$

where q denotes the flavour of the quark (i.e. u,d,c,s,t,b). The ψ_{qi} is a quark field with a colour index i corresponding to the three colours discussed above. $\not{D}$ corresponds to $\gamma^\mu D_\mu$ with γ^μ being a gamma matrix and covariant derivative D_μ defined as

$$(D_\mu)_{ij} = \partial_\mu\delta_{ij} - ig_s t^a_{ij}A^a_\mu, \tag{2.36}$$

where g_s is the strong coupling constant, A^a_μ is a gluon field and t^a_{ij} are matrices defined as

$$t^a_{ij} = \frac{1}{2}\lambda^a_{ij}, \tag{2.37}$$

with λ^a being hermitian and traceless Gell-Mann matrices of $SU(3)_C$. This definition dictates the normalisation of coupling g_s and sets the values of Casimir factors of $SU(3)_C$ and the structure constants of the theory. The interaction of gluons is defined by a gluon field strength tensor

$$F^a_{\mu\nu} = \partial_\mu A^a_\nu - \partial_\nu A^a_\mu - g_s f^{abc} A^b_\mu A^c_\nu, \tag{2.38}$$

and the label *a, b, c* run over $(N_C)^2 - 1 = 1, ..., 8$. The last term of Eq.(2.38) is what differentiates QCD from QED and leads to quartic and triple gluon self-interactions.

2.2.2 Colour Factors of QCD

In practice, it is not allowed to extract colour information from observables due to colour confinement. Instead, we sum over all the outgoing colours and average over all incoming ones. This leads to scattering amplitudes always having a sum over all quark fields contracted with λ^a matrices. Traces are then produced by these contractions, resulting in *colour factors* that are related to QCD processes. The colour factors basically take into account the trajectories or paths a process can take in colour space.

While computing squared Matrix Elements (ME), one might find the task of taking traces over t^a to be very tedious. Tab. 2.2 summarises various relations that can help us reduce this complexity. The Casimir factors for $SU(3)_C$ generators appearing in Tab. 2.2 are defined as [9, 15]:

$$T_F = \frac{1}{2} \qquad C_A = N_C = 3 \qquad C_F = \frac{4}{3}. \tag{2.39}$$

The last property in Tab. 2.2 is Fierz transformation, which exhibits the matrices in terms of δ functions and is frequently used in QCD calculations such as Monte Carlo shower implementation.

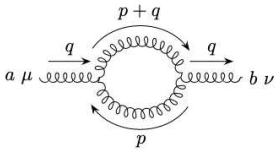
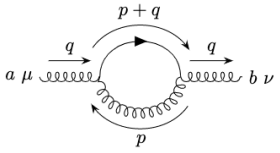
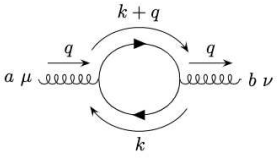
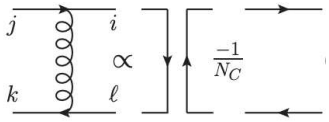
Diagram	Traces Relation	Indices
	$\sum_{c,d} f^{acd} f^{bcd} = C_A \delta^{ab}$	$a, b, c, d \in [1, ..., 8]$
	$\sum_c t^c_{aj} t^c_{jb} = C_F \delta_{ab}$	$a \in [1, ..., 8]$ $i, j, k \in [1, .., 3]$
	$Tr(t^a t^b) = T_F \delta^{ab}$	$a, b \in [1, ..., 8]$
	$t^a_{ij} t^a_{kl} = T_F \left(\delta_{jk}\delta_{il} - \frac{1}{N_C}\delta_{ij}\delta_{kl} \right)$	$i, j, k, l \in [1, ..., 3]$

TABLE 2.2: Relations for taking traces over t^a_{ij} matrices for ME calculations [9, 15, 17].

2.2.3 Running of Strong Coupling

As mentioned before, the last term in Eq.(2.38) is what distinguishes QCD from QED and leads to gluon self-interaction triple and quartic vertices, shown in Fig. 2.2. These self-interactions are responsible for the strange behavior of the running of the QCD coupling α_s ($\alpha_s = \frac{g_s}{4\pi}$).

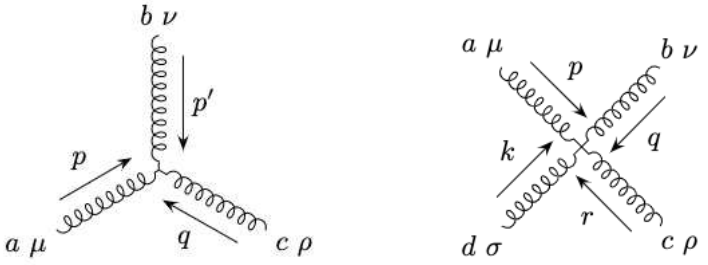


FIGURE 2.2: The 3-point (left) and 4-point (right) gluon vertex.

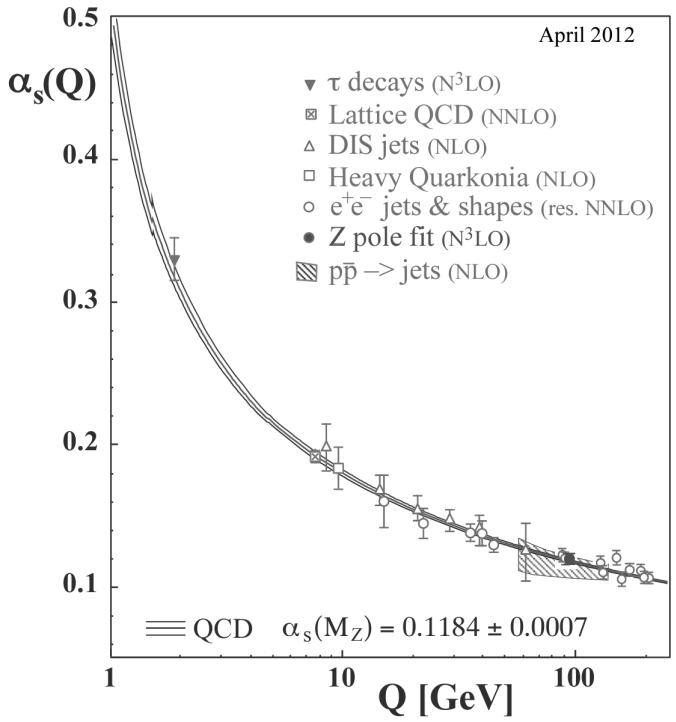


FIGURE 2.3: The running of QCD α_s in relation with energy scale Q, taken from [20].

The coupling constant α_s is logarithmic with energy and is controlled by an energy-dependent function, known as the β-function [9, 15]

$$\beta(\alpha_s) = -\alpha_s^2(b_0 + b_1\alpha_s + b_3\alpha_s^2 + ...). \tag{2.40}$$

Technically, we define coupling α_s value at a certain reference point, e.g, $Q^2 = M_Z^2$,

$$\alpha_s(Q^2) = \alpha_s(M_Z^2)\frac{1}{1 + b_0\alpha_s(M_Z^2)\ln\frac{Q^2}{M_Z^2} + \mathcal{O}(\alpha_s^2)}, \tag{2.41}$$

where $b_0 = (11C_A - 4T_F n_f)/12\pi$, is the leading order (one loop) coefficient of β-function. One can always calculate other relations by simply replacing M_Z^2 with some other scale factor.

Going back to Eq.(2.40), we see that the coupling α_s decreases with energy due to the overall negative sign of the function with coefficient $b_0 > 0$. This phenomenon is known as asymptotic freedom [18, 19]. As a direct outcome of asymptotic freedom, perturbative theories perform better at higher energies due to decreasing α_s. Fig. 2.3 shows the running of the QCD coupling constant α_s.

Another peculiar property of the QCD α_s is that it is low at smaller distances but increases at larger distances. As a result, gluons and quarks can only exist in bound states in nature, called hadrons. If we increase the separation between two quarks, it results in a growing strong force, which ultimately leads to the generation of new quark pairs. This is an important feature of the theory observed at the LHC and other particle physics experiments.

2.3 Unsolved Mysteries of the SM

To wrap up, we discuss some of the well-known unsolved mysteries of the SM, which serve as an incentive for the continuous quest for BSM physics.

2.3.1 Neutrino Masses

Due to the simple structure of the SM, neutrinos are considered massless. However, there is no conservational law or symmetry in nature that guarantees this. For many years, this has been the case, until recent experimental observations by Super-Kamiokande [21] and SNO [22] have suggested that neutrinos do have tiny masses. These experiments concluded that neutrinos change flavor over long distances and for such oscillations to occur, neutrinos must have mass [23]. As a result, there is a need to extend the SM to incorporate the neutrino masses, for which many candidates exist, including Majorana mass, Dirac mass, and the seesaw mechanism. There are other methods for incorporating neutrino masses into the 2HDMs [24].

2.3.2 Dark Matter

It is a well-known fact that the SM does not explain the origin of dark matter. There is an abundance of evidence from the physics community that there exists an invisible matter that provides additional velocities to the galaxies, and the visible matter masses will not be able to justify this phenomenon alone. Some important examples of this observation include galaxy rotations [25], velocity dispersions [26], and cosmic microwave background [27, 28].

2.3.3 CP-Violation

Another issue with the SM is the presence of Charge inversion (C) and Parity (P) violations. Both C and P are preserved separately as well as together (CP), for strong and EM interactions. However, this is not the case for the weak forces. Some examples where CP-violation is observed are: neutral charm and B-mesons [29], and neutral kaon decays [30].

The imbalance between matter and antimatter in the universe is one of the major unsolved mysteries associated with CP-violation. While experimental observations, particularly in decays of particles such as B-mesons, are consistent with Cabibbo-Kobayashi-Maskawa (CKM) matrix predictions aligning with the Sakharov conditions [31], they still do not fully explain the matter-antimatter distinction. This disparity raises serious concerns about the origin of CP-violation and whether the CKM matrix, within the framework of the SM, provides a complete explanation for the observed asymmetry. As far as we know, there are numerous candidate theories [32] that explain the phase in the CKM mixing matrix, but no concrete proof exists to establish the source of CP-violation and matter-antimatter asymmetry.

2.3.4 Gravity

One of the most famous unsolved problems within the SM is the failure to integrate QFT for gravity. The gravitational force is best described by an extraordinary theory known as General Relativity. Over the years, there have been many attempts to quantise gravity and integrate it into the SM, but all in vain as the theory is perturbatively non-renormalisable. Additionally, there is no experimental confirmation that there exists a massless, chargeless, and spin-2 boson, i.e., a graviton, that acts as a force carrier for the theory.

2.3.5 The Hierarchy Problem

The SM Higgs boson mass is surprisingly low when compared to the gravitational energy scale of $\mathcal{O}(10^{19})$ GeV. This is problematic as the first-order corrections to the Higgs mass due to renormalisation are quadratically proportional to the

hypothetical new physics energy scale, Λ, at which the SM becomes untenable:

$$\begin{aligned}\delta m_h^2 &\equiv m_h^2 - (m_h^0)^2 \\ &\propto \Lambda_{\mathrm{NP}}^2\Big(\frac{1}{4}(9g^2 + 3g'^2) - L_t^2\Big),\end{aligned}$$

where m_h is the Higgs mass, m_h^0 is the bare mass and L_t^2 is the Yukawa coupling to the top quark. From the above equation, it is clear that the Higgs mass is extremely sensitive to parameters such as Yukawa couplings. If no new physics is discovered up to $\frac{m_h^2}{\Lambda^2}10^{-34}$, then the production of measured Higgs mass requires an extremely huge bare mass. If the mass and mass correction are of similar magnitudes, then this is addressed through "finely tuned" cancellations between the quadratic radiative corrections and the bare mass. The problem is unnatural and cannot be formulated within the strict context of the SM. In some ways, the issue boils down to a concern that a future theory, in which the Higgs mass will be calculable, cannot have too many "fine tuning". To tackle this hierarchy problem, new physics at the TeV scale is required.

The examples above are just a few of the numerous unsolved problems within the SM, and there are many more, such as baryon asymmetry, the absence of coupling unification, and dark energy. To address these issues, it becomes necessary to investigate the BSM frameworks that will introduce new physics at the TeV scale. One such BSM framework is the 2HDM, the simplest possible extension to the Higgs sector. The 2HDMs can be embedded in fundamental theories such as Supersymmetry (SUSY), compositeness, and Grand Unified Theories (GUTs), making it one of the most studied frameworks in the search for new physics. In the following chapter, we will introduce the 2HDMs, with an emphasis on their relevance to this thesis.

Chapter 3

The 2HDM

In this chapter, we will briefly discuss the phenomenology of the 2HDM. We will be using the 2HDM Type-II model in our result section to study BSM physics. Refs. [33, 34] contains extensive reviews of the 2HDM. We take into account a second Higgs doublet, giving us a generic 2HDM. While 2HDMs alone cannot always fully explain the SM's inconsistencies, a second Higgs doublet embedded in many BSM models can. For example, the additional CP-violation sources found in this type of enlarged Higgs sector could explain the apparent matter-antimatter asymmetry. In particular, realisations of the 2HDM also have the appealing ability to explain neutrino mass generation [35], to provide dark matter candidates [36] or to accommodate the muon $g-2$ anomaly [37, 38, 39].

3.1 Addition of Second Higgs doublet

Starting with the SM, we have a single Higgs doublet that looks like this

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}. \tag{3.1}$$

The simplest extension possible is to add another Higgs doublet such that we have $SU(2)$ complex doublet fields given by

$$\Phi_a = \begin{pmatrix} \phi_a^+ \\ \phi_a^0 \end{pmatrix}, \tag{3.2}$$

with hypercharge $Y_i = 1$ and $a = 1, 2$. The most general form of the gauge-invariant scalar potential for the new fields Φ_1 and Φ_2 can be written as [33]

$$\begin{aligned} V(\Phi_1, \Phi_2) =& m_{11}^2 \Phi_1^\dagger \Phi_1 + m_{22}^2 \Phi_2^\dagger \Phi_2 - m_{12}^2 \Big(\Phi_1^\dagger \Phi_2 + \ h.c. \Big) + \frac{\lambda_1}{2} \Big(\Phi_1^\dagger \Phi_1 \Big)^2 + \frac{\lambda_2}{2} \Big(\Phi_2^\dagger \Phi_2 \Big)^2 \\ &+ \lambda_3 \Big(\Phi_1^\dagger \Phi_1 \Big) \Big(\Phi_2^\dagger \Phi_2 \Big) + \lambda_4 \Big(\Phi_1^\dagger \Phi_2 \Big) \Big(\Phi_2^\dagger \Phi_1 \Big) + \left[\frac{\lambda_5}{2} \Big(\Phi_1^\dagger \Phi_2 \Big)^2 + h.c. \right] \\ &+ \left[\lambda_6 \Big(\Phi_1^\dagger \Phi_1 \Big) (\Phi_1^\dagger \Phi_2 \Big) + h.c. \right] + \left[\lambda_7 \Big(\Phi_2^\dagger \Phi_2 \Big) (\Phi_1^\dagger \Phi_2 \Big) + h.c. \right], \end{aligned} \tag{3.3}$$

where λ_is (i $= 1, ..., 7$) are dimensionless parameters representing couplings of $\mathcal{O}(4)$ interactions and $m_{11,22,12}^2$ are the mass squared parameters. In the above equation, m_{11}^2, m_{22}^2 and $\lambda_{1,2,3,4}$ are real, and m_{12}^2 and $\lambda_{5,6,7}$ are complex in general. Altogether, we have 14 degrees of freedom (d.o.f) in $V(\Phi_1, \Phi_2)$.

After minimising the potential, each doublet acquires a VEV

$$\langle \Phi_a \rangle_0 = \begin{pmatrix} 0 \\ \frac{v_a}{\sqrt{2}} \end{pmatrix}, \tag{3.4}$$

with $a = 1, 2$. The real parameters v_a are defined by

$$v_1 = v \cos(\beta); \qquad v_2 = v \ e^{i\epsilon} \sin(\beta), \tag{3.5}$$

where ϵ is the phase that can be rotated away without affecting the other terms in the potential. For further calculations, ϵ is set to zero, so that we have neutral, CP conserving $v = \sqrt{v_1^2 + v_2^2} = 246$ GeV.

One of the most important components of the 2HDM model is

$$\tan(\beta) = \frac{v_1}{v_2}, \tag{3.6}$$

where parameter β is the rotation angle that diagonalises the mass-squared matrices of the pseudoscalars and charged scalars.

3.2 New Physical Higgs states

After expanding around the minimum of $V(\Phi_1, \Phi_2)$, we are left with eight fields in total

$$\Phi_1 = \begin{pmatrix} \phi_1^+ \\ \frac{v_1+\rho_1+i\eta_1}{\sqrt{2}} \end{pmatrix}, \qquad \Phi_2 = \begin{pmatrix} \phi_2^+ \\ \frac{v_2+\rho_2+i\eta_2}{\sqrt{2}} \end{pmatrix}. \tag{3.7}$$

This definition makes it easier to remove Goldstone bosons and deduce new physical Higgs states [33, 34]. From Eq.(3.7), we can deduce the Goldstone modes as

$$G^0 = \eta_1 \cos\beta + \eta_2 \sin\beta, \qquad G^\pm = \phi_1^\pm \cos\beta + \phi_2^\pm \sin\beta. \tag{3.8}$$

The physical neutral pseudoscalar and charged states, orthogonal to G^0 and $G^\pm$, respectively, are given as

$$A = \eta_1 \sin\beta - \eta_2 \cos\beta, \qquad H^\pm = -\phi_1^\pm \sin\beta + \phi_2^\pm \cos\beta. \tag{3.9}$$

The physical scalar states orthogonal to ρ_1 and ρ_2 can be obtained by performing rotation governed by α

$$H = -\rho_1 \cos\alpha - \rho_2 \sin\alpha, \qquad h = \rho_1 \sin\alpha - \rho_2 \cos\alpha. \tag{3.10}$$

Analogous to the SM, the Goldstone bosons G^0 and $G^\pm$ are once again "eaten" to give mass to the Z^0 and $W^\pm$ gauge bosons. The remaining five physical Higgs states give rise to

- CP-even, neutral, light Higgs, h.
- CP-even, neutral, heavy Higgs, H.
- CP-even, a pair of charged Higgs, $H^\pm$.
- CP-odd, neutral, (pseudoscalar) Higgs, A.

In the alignment limits of the 2HDM, where $\cos(\beta-\alpha) \to 0$ and $\sin(\beta-\alpha) \to 0$, one has the freedom to identify either lighter h or heavier H as the SM Higgs boson discovered in 2012.

Another important feature of the 2HDM is the number of d.o.f that the constituent fields have. These d.o.f. can be counted before and after the spontaneous breaking of the EW symmetry, based on the shape of the Higgs potential. Initially, we have a pair of complex doublets, Φ_1 and Φ_2, giving 8 d.o.f. in total. After EWSB, the spectrum contains two CP-even scalars h and H, one pseudoscalar A, and two

charged Higgs bosons $H^{\pm}$ (i.e. 5 d.o.f.). The Goldstone bosons of the theory will then become the longitudinal components of the weak $W^{\pm}$ and Z bosons (3 d.o.f). Consequently, the total number of d.o.f. remains unchanged.

3.2.1 Extracting Higgs Masses

Finally, we can determine the Higgs masses emerging in the 2HDM theory. We start with charged Higgs masses [40]

$$m^2_{H^{\pm}} = \frac{m^2_{12}}{\sin\beta\,\cos\beta} - \frac{v^2}{2}\Big(\lambda_4 + \lambda_5 + \lambda_6\,\cot\beta + \lambda_7\,\tan\beta\Big). \tag{3.11}$$

For pesudoscalar A, we have

$$m^2_{A} = \frac{m^2_{12}}{\sin\beta\,\cos\beta} - \frac{v^2}{2}\Big(2\lambda_5 + \lambda_6\,\cot\beta + \lambda_7\,\tan\beta\Big). \tag{3.12}$$

The CP-even neutral scalar states masses are given by [41]

$$m^2_{h,H} = \frac{1}{2}\Big(m^2_{11} + m^2_{22} \pm \sqrt{(m^2_{11} + m^2_{22})^2 + 4m^4_{12}}\Big). \tag{3.13}$$

3.2.2 A Choice of Basis

The doublets used to define the potential in Eq.(3.3) have the same quantum numbers and cannot be differentiated from one another. One can, therefore, easily rewrite the potential in terms of the linear combination of the original doublets. This change is known as a change of basis [42]. We can change the basis using

$$\bar{\Phi}_i = \sum_j U_{ij}\Phi_j, \tag{3.14}$$

where U_{ij} is 2×2 unitary matrix.

Multiple bases are allowed in the 2HDM that can be described as follows: the *general parametrisation* (as given above in terms of m^2_{ij} and λ_is), the *Higgs basis*, where one of the doublets gets zero VEV, and the *physical basis*, where the physical masses of the (pseudo)scalars are used.

One of the simplest choices of basis to work with is the Higgs basis [40, 43, 44]. We can rewrite the doublets as [33]

$$\begin{aligned} H_1 &= \cos\beta\ \Phi_1 + \sin\beta\ \Phi_2, \\ H_2 &= -\sin\beta\ \Phi_1 + \cos\beta\ \Phi_2. \end{aligned} \tag{3.15}$$

In the wake of SSB, we find that one of the doublets H_1 has a real and positive VEV of $\frac{v}{\sqrt{2}}$ with $v = \sqrt{v_1^2 + v_2^2}$, while the other has a null VEV.

However, in the light of the discovery of the 125 GeV Higgs boson, herein referred to as the h state, it is common to parametrise the theory using the *hybrid basis* [45], where the parameters allow for direct control on both the CP-even and CP-odd Higgs masses, the hVV couplings ($V = W^{\pm}, Z$), the $Aq\bar{q}$ vertices and the Higgs quartic couplings. The parameters in this basis are

$$\begin{aligned} &m_h, m_H \ = \ masses\ for\ the\ CP\ even\ Higgses \\ &cos(\beta - \alpha) \ = \ determine\ the\ g_{hVV}\ and\ g_{HVV}\ couplings \\ &tan\ \beta \ = \ given\ by\ the\ ratio\ of\ the\ VEVs \\ &Z_4,\ Z_5,\ Z_7 \ = \ self\ couplings\ parameters\ for\ the\ Higgses, \end{aligned}$$

with $m_H \geq m_h$, $0 \leq \beta \leq \pi/2$ and $0 \leq |\sin(\beta - \alpha)| \leq 1$. The quartic scalar couplings in the Higgs basis are used to express remaining (pseudo)scalar masses:

$$m_A^2 = m_H^2 \sin^2(\beta - \alpha) + m_h^2 \cos^2(\beta - \alpha) - Z_5 v_1^2, \tag{3.16}$$

$$m_{H^{\pm}}^2 = m_A^2 - \frac{1}{2}(Z_4 - Z_5)v^2. \tag{3.17}$$

By swapping the self-couplings Z_4 and Z_5 with the scalar masses given above, the 7 free parameters can be recast into 4 physical masses and 3 parameters related to the couplings of the (pseudo)scalars to gauge bosons, fermions, and scalars themselves, respectively:

$$m_h,\ m_H,\ m_A,\ m_{H^{\pm}},\ \cos(\beta - \alpha),\ \tan(\beta),\ Z_7. \tag{3.18}$$

It is worth noting that Z_7 only affects the triple and quartic Higgs interactions, so it does not appear in the tree-level diagrams for our process. Since m_h has been accurately measured at the LHC, the total number of d.o.f globally reduces to 6.

Model	h			H			A		
	u	d	l	u	d	l	u	d	l
2HDM type-I	$\frac{\cos\alpha}{\sin\beta}$	$\frac{\cos\alpha}{\sin\beta}$	$\frac{\cos\alpha}{\sin\beta}$	$\frac{\sin\alpha}{\sin\beta}$	$\frac{\sin\alpha}{\sin\beta}$	$\frac{\sin\alpha}{\sin\beta}$	$\cot\beta$	$-\cot\beta$	$-\cot\beta$
2HDM type-II	$\frac{\cos\alpha}{\sin\beta}$	$-\frac{\sin\alpha}{\cos\beta}$	$-\frac{\sin\alpha}{\sin\beta}$	$\frac{\sin\alpha}{\sin\beta}$	$\frac{\cos\alpha}{\sin\beta}$	$\frac{\cos\alpha}{\sin\beta}$	$\cot\beta$	$\tan\beta$	$\tan\beta$

TABLE 3.1: Couplings of the neutral Higgs bosons to fermions, normalised to the corresponding SM value (m_f/v) in the 2HDM Type-I and II.

3.3 Flavour Changing Neutral Currents (FCNCs)

The general Yukawa sector Lagrangian with two Higgs doublets is given as

$$\mathcal{L}_Y = \sum_{i=1}^{2}\Big(\bar{L}_{L,i}Y_l^i\Phi n_R + \bar{Q}_{L,i}Y_u^i\tilde{\Phi}n_R + \bar{Q}_{L,i}Y_d^i\tilde{\Phi}p_R\Big) + h.c.\ . \tag{3.19}$$

In the SM, the Yukawa interactions are naturally diagonalised by the mass matrix such that FCNCs are absent due to the Glashow-Weinberg-Paschos (GWP) mechanism [46]. However, the Yukawa matrices present in the above equation are not diagonalisable, and non-zero terms lead to tree-level Higgs-mediated FCNCs. The emergence of FCNCs has been contradicted by current experimental observations, and one needs to find a way to suppress them. All we need to do is enforce discrete $\mathbb{Z}_2$ symmetry

$$\Phi_1 \to \Phi_1, \qquad \Phi_2 \to -\Phi_2. \tag{3.20}$$

This allows the suppression of FCNCs that appeared after the inclusion of the second Higgs doublet in a natural way. The Higgs potential then becomes

$$\begin{aligned} V(\Phi_1,\Phi_2) =& m_{11}^2\Phi_1^\dagger\Phi_1 + m_{22}^2\Phi_2^\dagger\Phi_2 + \frac{\lambda_1}{2}\Big(\Phi_1^\dagger\Phi_1\Big)^2 + \frac{\lambda_2}{2}\Big(\Phi_2^\dagger\Phi_2\Big)^2 + \lambda_3\Big(\Phi_1^\dagger\Phi_1\Big)\Big(\Phi_2^\dagger\Phi_2\Big) \\ &+ \lambda_4\Big(\Phi_1^\dagger\Phi_2\Big)\Big(\Phi_2^\dagger\Phi_1\Big) + \left[\frac{\lambda_5}{2}\Big(\Phi_1^\dagger\Phi_2\Big)^2 + h.c.\right] - m_{12}^2\Big(\Phi_1^\dagger\Phi_2 + \ h.c.\Big). \end{aligned} \tag{3.21}$$

We can see that the $\mathbb{Z}_2$ symmetry is softly broken by the mass parameter m_{12}^2. In fact, if all the fermions in the theory have one standard $\mathbb{Z}_2$ quantum number, then we have four types of 2HDMs that can invoke softly broken $\mathbb{Z}_2$ symmetry-Type-I, Type-II, Type-III, and Type-IV [33, 47]. Throughout this thesis, we will only focus on the 2HDM Type-I (used in Chapter 7) and Type-II scenarios. For completeness, we detail the 2HDM Type-I and Type-II and respective couplings of the neutral Higgs scalars to fermions (relative to the SM value of m_f/v) in Tab. 3.1.

3.4 Theoretical Constraints

Before diving into the 2HDM phenomenology, it's important to understand the constraints imposed on some 2HDM parameters.

3.4.1 Stability of the Vacuum

In the 2HDM, a positive potential is required for the vacuum to be stable. It is imperative to maintain the stability of the scalar potential to minimize the possibility of creating a vacuum that is not the true minimum and could result in vacuum instability and the theory's collapse. The stability of the vacuum imposes constraints on the 2HDM potential parameters, ensuring that the potential is bounded from below.

At large field values, in a general 2HDM, the potential is dominated by the quartic terms and the stability constraint requires that specific combinations of these quartic couplings be positive. This can be achieved by enforcing following constraints on λ_i $(i = 1, ..., 7)$ [40]:

$$\lambda_1 > 0; \quad \lambda_2 > 0; \quad \lambda_3 > -\sqrt{\lambda_1 \lambda_2}. \tag{3.22}$$

We have an additional condition when both λ_6 and λ_7 are zero,

$$\lambda_3 + \lambda_4 - |\lambda_5| > -\sqrt{\lambda_1 \lambda_2}. \tag{3.23}$$

Both of these constraints ensure the positivity of the potential necessary for the theory. If either λ_6 or λ_7 are non-zero, then we replace $|\lambda_5|$ with λ_5 in the above equation to maintain the stability of the vacuum.

The stability constraints are determined by the type of 2HDM used and the specific potential parameters. As a result, understanding vacuum stability for a specific type of 2HDM is critical for scanning permissible parameter space for analysis and predicting possible experimental signatures and outcomes in collider experiments.

3.4.2 Oblique Parameters

The oblique parameters [48], called S, T, and U (and their higher-order extensions V, W, and X [49]) are a set of measurable quantities that combine EW precision

data to quantify potential new physics contributions. The S parameter represents corrections to custodial symmetry of the EW interactions, the T parameter represents the shift in the scattering of longitudinally polarized W bosons and corrections to transversely polarized W bosons are represented by the U parameter. At the loop level, these parameters can be calculated using EW diagrams.

A non-zero value of any of these parameters suggests the existence of BSM physics, in contrast to the SM, where $S = T = U = 0$. As a result, these parameters have been subjected to stringent constraints from experimental measurements, providing a powerful tool for probing and constraining various extensions of the SM. The 2HDM is a suitable candidate for examination using the oblique correction formalism because the additional $SU(2)_L$ doublet does not significantly contribute to the oblique parameters, as scalar doublets or singlets do not disrupt the custodial symmetry that safeguards the tree-level relation $\rho \equiv M_W/(M_Z \cos\theta_W) = 1$. However, the large mass splittings of the new extra Higgs states [50] can lead to some major contributions to these parameters.

The oblique parameters are crucial in constraining the parameter space of the 2HDMs and identifying viable regions suitable for new physics search analysis that can help us better understand the interaction between the extended Higgs sector and precision EW measurements.

3.4.3 Tree-Level Unitarity

In addition to the above-mentioned constraints, one can also obtain limits by requiring the tree-level unitarity for the scattering of the Higgs boson and EW gauge bosons. It is essential for the complete all-order scattering matrix (S) to be unitarity. To achieve this in the general 2HDM, necessary and sufficient conditions on the eigenvalues of the S-matrices must be met by imposing the eigenvalues L_i to obey $L_i \leq 8\pi$ [51]. The nine eigenvalues L_i are then given by

$$
\begin{aligned}
p_1 &= 2(\lambda_3 + \lambda_4) - \frac{\lambda_5}{2} - \frac{\lambda_6}{2}, \\
e_1 &= 2\lambda_3 - \lambda_4 - \frac{\lambda_5}{2} + \frac{5\lambda_6}{2}, \\
e_2 &= 2\lambda_3 + \lambda_4 - \frac{\lambda_5}{2} + \frac{5\lambda_6}{2}, \\
f_1 &= f_2 = 2\lambda_3 + \frac{\lambda_5}{2} + \frac{\lambda_6}{2}, \\
f_+ &= 2\lambda_3 - \lambda_4 + \frac{5\lambda_5}{2} - \frac{\lambda_6}{2},
\end{aligned}
$$

$$
\begin{aligned}
f_- &= 2\lambda_3 + \lambda_4 + \frac{\lambda_5}{2} - \frac{\lambda_6}{2}, \\
a_\pm &= 3(\lambda_1 + \lambda_2 + \lambda_3) \pm \sqrt{9(\lambda_1 - \lambda_2)^2 + \Big(4\lambda_3 + \lambda_4 + \frac{1}{2}(\lambda_5 + \lambda_6)^2\Big)}, \\
b_\pm &= \lambda_1 + \lambda_2 + 2\lambda_3 \pm \sqrt{(\lambda_1 - \lambda_2)^2 + \frac{1}{4}\Big(-2\lambda_4 + \lambda_5 + \lambda_6\Big)^2}, \\
c_\pm &= \lambda_1 + \lambda_2 + 2\lambda_3 \pm \sqrt{(\lambda_1 - \lambda_2)^2 + \frac{1}{4}\Big(\lambda_5 - \lambda_6\Big)^2}.
\end{aligned}
\tag{3.24}
$$

Tree level unitarity is critical in the context of a 2HDM due to the potential emergence of resonances or unphysical behaviors in Higgs boson scattering processes, which requires that the amplitudes do not exceed the unitarity bounds. This constraint limits the range of masses and couplings of the Higgs bosons within the 2HDM, allowing for the thorough scan of a parameter space that respects both theoretical coherence and experimental viability.

Other important constraints come from the magnetic moment of the muon [52, 53] and Higgs searches at the LHC. As previously stated, the 2HDM parameter space has numerous input parameters based on the basis selected. As a result, it is critical to scan the allowed parameter space and generate points that satisfy these theoretical constraints in order to be valid for further BSM analysis. Tab. 3.2 shows some example points that passed the theoretical constraints discussed above. We will use these benchmark points in our analysis later on.

Label	m_h (GeV)	m_H (GeV)	$\tan\beta$	$\sin(\beta-\alpha)$	m_{12}^2	BR($h \to b\bar{b}$)	σ(pb)
Point1	125	700.668	2.355	-0.999	1.46×10^5	6.164×10^{-1}	1.870×10^{-2}
Point2	60	125	1.6	0.1	4×10^3	8.610×10^{-1}	6.688
Point3	125	867.095	5.63678	0.915004	1.24×10^5	7.1629×10^{-1}	0.00203

TABLE 3.2: The 2HDM parameters and cross sections for some example benchmark points that passed the theoretical constraints.

3.5 Phenomenology of the 2HDMs at the LHC

In this section, we discuss the phenomenology of the 2HDMs and ways to probe the extended 2HDM Higgs family at the LHC after the discovery of the Higgs boson in 2012.

3.5.1 Scalar (Pseudoscalar) Sector Decay Processes

One of the most important processes that may be explored at the LHC is the $H \to hh$ channel. Assuming that $m_H > 2m_h$, it is possible for heavier Higgs H to decay into two lighter Higgses hh. If we consider h to be SM like Higgs with $m_h = 125$ GeV, then there exist various well-defined decay channels [33, 54], with $h \to b\bar{b}$ being the dominant one with a 57% branching ratio. Throughout this thesis, we will be focusing on $H \to hh$ with the $h \to b\bar{b}$ decay channel.

Another crucial process to look for at the LHC is the decay of either h or H into two pseudoscalars A. There have already been many experimental studies published for this process, for example, $h/H \to AA \to \gamma\gamma\gamma$ [55, 56] and $h/H \to AA \to \tau^+\tau^- b\bar{b}$ [57].

3.5.2 Charged Sector Decay Processes

Other searches at the LHC involve looking for the charged Higgs $H^\pm$ production. Depending on the mass of the charged Higgses, one can take different routes to search for $H^\pm$ production at the LHC [33].

For a higher mass range, most of the time h^+ decays into $t\bar{b}$. It is also possible for H^+ to couple with other Higgses via $H^+ \to A/hW^\pm$. However, for a lower mass range, the dominant production method is for t to decay into H^+b with $H^+ \to \tau^+\nu$. For a complete review of charged Higgs production at the LHC, please refer to [58].

Chapter 4

Review of Jet Physics

In this chapter, we will review jet physics, which is crucial for mapping underlying hard interactions in high-energy physics experiments as well as the cluster of particles we see in the detectors. As we will see, jet physics is critical for the SM measurements and detecting BSM physics.

4.1 Jets Formation

Due to QCD colour confinement, quarks and gluons cannot be isolated and can only exist as hadronic bound states. In the high-energy regime of experimental colliders, these partons undergo numerous processes and are observed as sprays of colourless hadrons known as jets.

The first step in the production of the jets is parton showering, which is a sequence of small-angle splits from a parton. As a simple picture, we first consider the probability of a parton (labeled X) emitting a quark or gluon, denoted by

$$\mathcal{P}(X \to Xg) \sim \alpha_s \int \frac{dE}{E} \frac{d\theta}{\theta} \tag{4.1}$$

where α_s is the coupling, θ is the angle of emission and E is the outgoing energy. As we can see, the equation diverges at low θ and thus the probability of emitting a gluon at a small angle will outweigh that of large angle emissions. As a result, a series of emissions will be in collimated flows which is the starting point for the jet formation.

Generalising to various kinds of splittings, the Dokshitzer-Gribov-Lipatov-Altarelli-Parisi (DGLAP) equation [59, 60, 61, 62], which encodes the behavior of partons in hadron collisions via the parton distribution functions (PDF) $f(x, \mu)$, at some energy scale μ, can be written down

$$\mu \frac{\partial}{\partial \mu} f_j(x, \mu) = \sum_j \int_x^1 \frac{dz}{z} \frac{\alpha_s}{2\pi} P_{ij}(z) f_j(\frac{x}{z}, \mu) \tag{4.2}$$

The sum is a generalisation to multiple parton splittings, and the $P_{ij}(z)$ are the splitting functions for a $j \to ik$ splitting with i taking a fraction z of the total momentum of j. To start, we first consider the splitting function of the quark further radiating quark(antiquark) [16]

$$P_{qq}(z) = C_F \left(\left[\frac{1+z^2}{1-z} \right]_+ + \frac{3}{2} \delta(1-z) \right). \tag{4.3}$$

where z and $(1-z)$ are the energy fractions. Similarly, a quark can also split into a gluon and a quark

$$P_{gq}(z) = C_F \left(\frac{1+(1-z)^2}{z} \right). \tag{4.4}$$

However, this is not the full picture, as sometimes a gluon can split into a quark-antiquark pair

$$P_{gq}(z) = T_R \Big(z^2 + (1-z)^2 \Big), \tag{4.5}$$

and also radiate another gluon

$$P_{gg}(z) = C_A \left(\frac{z}{1-z} + \frac{1-z}{z} + z(1-z) \right) + \delta(1-z) \frac{11 C_A - 4 n_f T_R}{6}. \tag{4.6}$$

In the preceding equations, z and $(1-z)$ are again the energy fractions, with $C_F = \frac{4}{3}$, $C_A = 3$, and $T_R = \frac{1}{2}$ being the QCD 'colour factors' and n_f being the number of fermions coupling to the gluons.

This repetitive splitting is what leads to the aforementioned parton shower, where the partons are collinear and soft, causing the final partons to be collimated in the direction of the initial ones. Once the average energy of the initial collision approaches Λ_{QCD}, quarks and gluons can no longer exist as separate entities and perturbative theory no longer holds due to the running of the QCD coupling constant α_s. This results in the generation of stable colorless hadrons (kaons, pions, etc.) from colored partons, a process known as hadronisation. The end

result of parton showers and hadronisation is collimated sprays of hadrons known as jets.

4.2 Jet Clustering Algorithms

In a few ways, the above-mentioned explanation for the formation of jets is over-simplified. To begin with, partons are ill-defined entities, and then there's the question of whether two particles belong to the same jet or two separate jets, which is relevant to what we mean by "collimated".

As a result, a mere understanding of what a jet is meant to represent is really not sufficient to distinguish jets in an event. A *jet definition* is thus used to correctly define a jet, providing a mapping between the hadronic sprays in the final state and the initial hard interactions that occurred.

A jet definition is made up of several key components, including the *jet algorithm* and a series of parameters such as *jet radius*, which is the angular distance between two particles in the rapidity-azimuth ($y - \phi$) plane. In addition to this, a jet definition also employs a *recombination scheme* to explain how constituents can be used to derive the kinematics of the jets.

A jet definition is a complex structure that must be regulated by a set of rules. In 1990, a group of renowned theorists and experimentalists came together to form the "Snowmass Accord" [63], which outlines some basic requirements that any jet definition should satisfy

Several important properties that should be met by a jet definition are [64]:

- Simple to implement in an experimental analysis;
- Simple to implement in the theoretical calculation;
- Defined at any order of perturbation theory;
- Yields finite cross sections at any order of perturbation theory;
- Yields a cross-section that is relatively insensitive to hadronisation.

The first and second conditions are largely practical considerations, where a jet definition can satisfy both theoretical calculations and experimental analysis. The third and fourth criteria are related to infrared and collinear safety which are crucial for QCD calculations. According to the fifth condition, a jet definition should be insensitive to both theoretical factors like Underlying Events (UE) and hadronisation, as well as experimental factors like Pile-Up (PU) and detector effects. We will see how these five points become relevant when defining new jet algorithms.

Moving on to the main ingredient of a jet definition, there is indeed a long history linked with the development of jet clustering algorithms. There are two main classes of jet algorithms in use: *sequential recombination algorithms* and *cone algorithms*. We only focus on sequential recombination algorithms in this thesis.

In the next section, we will briefly explore cone algorithms before delving into the main jet clustering algorithms currently used at the hadron colliders.

4.2.1 Cone Algorithms

In 1977, Sterman and Weinberg developed the first jet clustering algorithm in the context of e^+e^- collisions leading to hadrons scattering [65]. At high energies, $e^+e^- \rightarrow jj$ was found to dominate all others, with an angular distribution of $(1 + cos^2\theta)$ similar to that of charged spin-half particles.

In this jet algorithm, an event is grouped into two jets if the fraction of its total energy, $(1-\epsilon)$, can be contained into two cones of half-angle δ (hence, it is known as the "cone" algorithm), shown in Fig. 4.1. The basic idea was to define jets based on two input parameters measuring the energy and angle of radiation, given by ϵ and δ. Specifically, the radiation from one of the initial partons must be hard enough

$$\epsilon < E_{rad}, \tag{4.7}$$

and at a wider angle from any of the other jets to be classified as a different jet

$$\delta < \theta_{min}. \tag{4.8}$$

The exact values of parameters ϵ and δ depend hugely on the physics analysis being studied and ultimately decide the number of jets in an event. The existence of energy and angular parameters to define the properties of the jets is a typical feature of cone algorithms. In other words, cone algorithms define *stable cones*

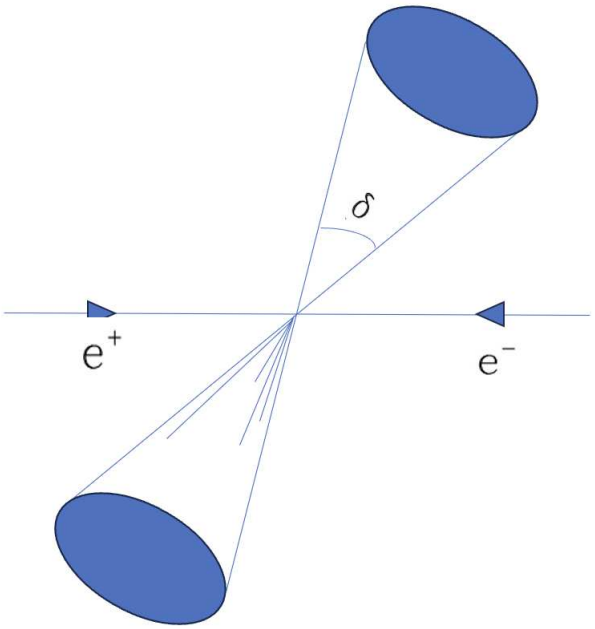


FIGURE 4.1: Diagrammatic representation of Sterman-Weinberg cone jets.

using the final state radiations without actually clustering the particles, which can be seen as self-sufficient criteria.

Since [65], cone algorithms have advanced substantially. These were altered to better suit the hadron collider's environment since it is not always evident, either computationally or physically, where to construct the cones for an event with more than two jets. Some of the famous cone algorithms in use consist of JetClu [66], midpoint-type [67], and SISCone [68].

While cone algorithms give a fair representation of radiation coming from initial partons, they can be difficult in practice, especially when the number of jets increases in an event. This is one of the main reasons why they are not used in most of the high-energy physics experiments today. Other algorithms, such as sequential recombination algorithms, are far more flexible to adapt to high-jet multiplicity events and are currently deployed at the hadron colliders.

4.2.2 Sequential Recombination Algorithms

The history of sequential recombination algorithms can be traced back to e^+e^- experiments [69]. However, it was the LUCLUS algorithm [70] that established the underlying ideas corresponding to these jet algorithms. All of today's sequential

algorithms are rather straightforward to describe when compared to cone algorithms, which is one of the reasons why they are favored at the hadron colliders.

Sequential algorithms are based on the idea that jets are the results of sequential parton branchings. Therefore, these algorithms attempt to invert the parton shower/hadronisation procedure in order to reduce the complexity of the final states. Instead of classifying the entire event, each particle in the event is considered, and all are iteratively combined together based on some inter-particle distance measure until and unless all the particles in the event are gathered into stable jets. The diagrammatic representation is shown in Fig. 4.2.

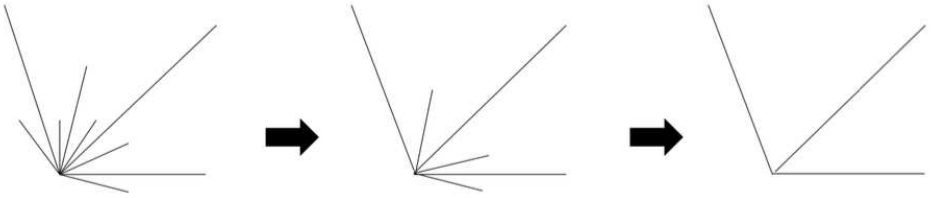


FIGURE 4.2: Diagrammatic representation of particles being combined into jets in a sequential recombination algorithm.

Another reason these algorithms are favored is that they are infrared and collinear-safe. In the next few sections, we will look at some of the well-known sequential clustering algorithms that are currently deployed at the hadron colliders, particularly at the LHC.

4.2.2.1 JADE Algorithm

In the mid-1980s, the JADE Collaboration introduced the first basic example of sequential recombination algorithms [71]. The JADE algorithm creates jets by iteratively combining the final state particles. We start with defining an inter-particle distance measure d_{ij} between two particles i and j

$$d_{ij} = \frac{y_{ij}^2}{E_{tot}^2} = \frac{2E_iE_j(1 - cos\ \theta_{ij})}{E_{tot}^2}, \tag{4.9}$$

where E_{tot} is the total energy of the entire event, θ_{ij} is the angle between two particles i and j, and E_i is the energy of particle i. Following that, the algorithm proceeds by calculating the distance d_{ij} for all possible pairs and then

- Find the pair with the minimum d_{min} of all the d_{ij}

- Define a fixed parameter that acts as a jet resolution cut-off distance measure, d_{cut}
- If $d_{min} < d_{cut}$, then combine i and j into a new pseudojet and repeat the procedure until no particles are left
- If $d_{min} > d_{cut}$, then we stop the iteration and declare all the remaining particles as jets.

Going back to the distance measure d_{ij}, we can see that it vanishes both for soft particles ($E_i \to 0$ or $E_j \to 0$) and for collinear pairs ij ($cos\ \theta_{ij} \to 1$). In practice, the algorithm progresses the clustering of particles from smaller values of d_{ij} to larger ones, prioritising the region predominantly cluttered by soft and collinear singularities. This, however, leads to the combining of two widely separated soft particles into a single object in the initial phase of the clustering. As a result, the notion that a jet's angular reach is restricted no longer holds true, resulting in higher-order correction issues.

Despite this, the JADE algorithm is far more flexible than cone algorithms as it is governed by one single parameter, d_{cut}. The reliance on one parameter makes it easier to handle multi-jet events, which was a potential issue with the cone algorithms.

4.2.2.2 $k_\perp$ Algorithm in e^+e^- Experiments

The next sequential recombination algorithm, the $k_\perp$ algorithm [72], is the direct descendant of the JADE algorithm. This algorithm was also designed for e^+e^- experiments with hadrons in the final state.

The algorithm was developed primarily to address the issue of clustering of wide-angle soft particles before more rational options at small d_{ij} values due to the presence of E_iE_j in the d_{ij} measure. As a result, a minor modification was made to the distance measure

$$d_{ij} = \frac{2\ min(E_i^2, E_j^2)\ (1 - cos\ \theta_{ij})}{E_{tot}^2}, \tag{4.10}$$

where we have replaced E_iE_j (see Eq.4.9) with $\min(E_i^2, E_j^2)$. The $\min(E_i^2, E_j^2)$ function ensures that the softer particle energy (between i and j) is considered

and they get clustered with its nearest neighbor instead of other wide-angle soft particles.

Again, we can see that the distance metric disappears when $E_i \to 0$ or $E_j \to 0$. For $\theta_{ij} \ll 1$, the numerator $(2\ min(E_i^2, E_j^2)\ (1 - cos\ \theta_{ij}))$ reduces to $(min(E_i, E_j)\theta_{ij})^2$. This can be further rewritten as $k_\perp^2$, which represents the squared transverse momentum of particle i with respect to particle j. Apart from a minor tweak to the distance metric, the algorithm then follows the same steps as JADE to cluster all the particles in an event.

4.2.2.3 Generalised k_T Algorithm

When studied in the context of colliders with incoming hadrons, the k_T algorithm gives rise to two key issues. First, the total energy of an entire event E_{tot} is unknown. Second, the QCD branching probability suffers from divergences coming from pairs of outgoing particles as well as from the incoming beam direction. So, in order to have a version that works well at the hadron colliders, the generalised k_T algorithm was introduced [73]. The modified distance measure for this algorithm is given by

$$d_{ij} = min(p_{Ti}^{2a}, p_{Tj}^{2a}) \frac{\Delta R_{ij}^2}{R^2}, \tag{4.11}$$

where a is a free parameter and R is the jet radius acting as a cutoff for any particle pairing. The ΔR_{ij}^2 is the angular distance between the two particles in the rapidity-azimuth $(y - \phi)$ plane, given by

$$\Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2. \tag{4.12}$$

In addition to this, the "beam measure" "d_{iB}", which is just the measure of separation between particle i and beam B is also introduced due to the inclusion of the beam's splittings

$$d_{iB} = p_{Ti}^{2a}. \tag{4.13}$$

The algorithm then proceeds as follows

- Calculate all possible d_{ij} and d_{iB} using Eqs.(4.11) and (4.13)
- The smallest distance d_{min} of all possible d_{ij} and d_{iB} is taken

- If d_{min} is a d_{ij}, recombine particles i and j into a new object and take it off the list
- If d_{min} is a d_{iB}, then particle i is declared a jet and erased from the list
- This procedure is then repeated until all the particles are merged and declared as jets.

Similarly to the JADE algorithm, the generalised k_T algorithm also depends on one single parameter, R. If particles i and j are closer in the $y - \phi$ plane, the distance measure d_{ij} becomes small and the pair are clustered into a jet. If, on the other hand, $\Delta R_{ij} > R$, then the d_{iB} distance measure becomes smaller than d_{ij} and the clustering is no longer possible. Thus, R plays an important role in deciding what can be declared a jet.

A more refined version of the generalised k_T algorithm is the so-called "k_T algorithm", where a minor change is made to d_{ij} and d_{iB} definitions by inputting $a = 1$. The k_T algorithm is infrared and collinear safe, which is one of the main reasons why the theoretical community supports it. However, in recent times, the algorithm has not been as widely used because of its irregular jet construction and dependence on soft particle emissions. This led to several other iterations of generalised k_T, each with its own set of rules, benefits, and disadvantages.

4.2.2.4 The Cambridge-Aachen Algorithm

Another example of the sequential recombination algorithm and a direct descendant of the generalised k_T algorithm is the "Cambridge-Aachen Algorithm" (C/A) [74]. The modified distance measures for C/A are obtained by setting $a = 0$ in the Eqs.(4.11) and (4.13)

$$\begin{aligned} d_{ij} &= \frac{\Delta R_{ij}^2}{R^2}, \\ d_{iB} &= 1. \end{aligned} \tag{4.14}$$

From the definitions of the distance measures, one can see that the C/A clustering is based on the angular separation between the particles and has no reliance on transverse momentum. The distance measures then become purely geometrical and suffer fewer soft particle emissions than the k_T algorithm.

4.2.2.5 The Anti-k_T algorithm

Perhaps the most important and widely used algorithm at the LHC is the "anti-k_T algorithm" [75]. The algorithm can be obtained by setting $a = -1$ in the definition of the distance measures of the generalised k_T algorithm. The modified distance measures are given as

$$\begin{aligned} d_{ij} &= min\left(\frac{1}{p_{Ti}^2}, \frac{1}{p_{Tj}^2}\right)\frac{\Delta R_{ij}^2}{R^2}, \\ d_{iB} &= \frac{1}{p_{Ti}^2}. \end{aligned} \tag{4.15}$$

Here, the algorithm chooses to cluster the hard particles first rather than the soft particles, unlike the k_T and the energy-independent C/A algorithms. The anti-k_T algorithm thus tends to cluster regular, well-defined circular hard jets in the $(y-\phi)$ plane. These hard jets are not sensitive to soft particle emissions and are easy to calibrate at the experiment.

4.2.2.6 The Variable-R Algorithm

All the sequential recombination algorithms require the declaration of a jet radius parameter R. This parameter acts as the cutoff for any particle pairing and decides what can be declared as a jet. However, one must exercise caution when selecting a particular R value because not all jets will fit into a single cone size.

We know that the angular separation of the given jet constituents hugely depends on jets p_T

$$\Delta R \propto \frac{1}{p_T}. \tag{4.16}$$

For high p_T objects, the constituents are accumulated into compact collimated narrow jets, for softer objects, one can expect the constituents to be more spread out over some wider angle. Therefore, one needs to carefully choose the right R in order to accommodate the clustering of high and low p_T objects.

The variable-R algorithm [76] eliminates the need to specify one fixed cone size. As the name suggests, the variable-R alters the sequential algorithm scheme by replacing the fixed-R parameter in the distance metric with a p_T dependent dimensionless parameter $R_{eff}(p_T)$. As a result, the distance metric d_{ij} can be rewritten as

$$d_{ij} = min(p_{Ti}^{2a}, p_{Tj}^{2a})\Delta R_{ij}^2, \tag{4.17}$$

with the beam distance measure

$$d_{iB} = p_{Ti}^{2a} R_{eff}^2(p_{Ti}). \tag{4.18}$$

The $R_{eff}(p_T)$ parameter is given by

$$R_{eff}(p_T) = \frac{\rho}{p_T}, \tag{4.19}$$

where ρ is a dimensionful object taken to be $\mathcal{O}(p_T)$. The modified d_{iB} plays an important role here. The soft particles get clustered with their nearest neighbors, enhancing d_{iB}, whereas for hard particles, d_{iB} is suppressed, making these particles more likely to be clustered into jets.

In the variable-R approach, the process is modified in such a way that one can avoid events with very wide jets at low p_T. The dimensionful parameter ρ can be scanned over a range to optimise the maximum desired sensitivity. This can also be done for other parameters such as $R_{\text{min/max}}$ (cut-offs for the minimum and maximum allowed R_{eff}), respectively, i.e., if a jet has $R_{\text{eff}} < R_{\text{min}}$, it is overwritten and set to $R_{\text{eff}} = R_{\text{min}}$ and equivalently for R_{max}.

In multijet signal events where one might expect signal jets with a large range of p_T's, a variable-R reconstruction procedure could outperform the traditional fixed-R routines. In particular, using a variable-R alleviates the balancing act of finding a single fixed cone size that suitably engulfs all of the radiation inside a jet without sweeping up too much outside 'junk'.

To illustrate the benefit of using the variable-R algorithm, we map the constituents of b-tagged jets in the same event, which have been clustered using both variable-R and fixed-$R = 0.8$ approaches, shown in Fig. 4.3. We can see that the fixed cone approach resolves only three b-jets, whereas the variable-R approach was able to reconstruct all four expected b-jets forming the signal. While fixed-R sweeps radiation from a nearby jet into the leading b-jet, variable-R is able to resolve both due to the larger p_T of the leading b-jet (hence a smaller effective radius R_{eff}) and also by adapting R_{eff} to a larger value to suitably reconstruct the lower p_T jets.

Apart from these minor modifications, the variable-R algorithm works in a similar fashion as the generalised k_T algorithm. The re-appearance of a in the d_{ij} and d_{iB} helps the variable-R to imitate the C/A and the anti-k_T algorithms with varying cone sizes.

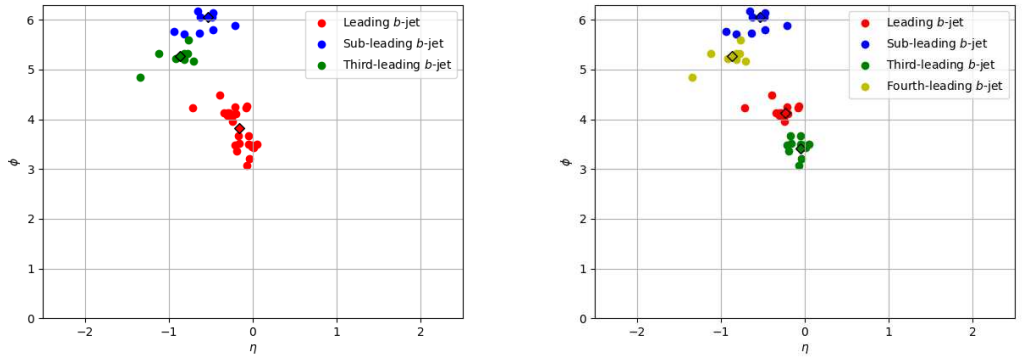


FIGURE 4.3: An event is clustered using a fixed cone size $R = 0.8$ (left) and the variable-R algorithm (right). The anti-k_T clustering algorithm was used for both cases.

We will further discuss the importance and performance of the variable-R algorithm in comparison to traditional jet clustering algorithms in the result section of this thesis by using a 2HDM Type-II benchmark point in a high jet multiplicity scenario.

4.3 Jets at the LHC

So far we have been discussing jets and their reconstruction from a theoretical point of view. However, it is important to understand how jets originate in an experimental setup since they hold the key to discovering new physics and particles at the LHC.

In this section, we will review the concept of jets at the LHC, with particular emphasis on the CMS detector phenomenology.

4.3.1 The CMS Detector

The LHC is the largest and one of the most powerful particle colliders in the world, with a 27 km circular ring of superconducting magnets. Inside the detector, two high-energy proton beams are made to collide, resulting in outgoing hadrons following parton showers and hadronisation, dispersed in every direction around the primary vertex. The cylindrical shape of the detector around the beamlines then aids in squeezing the particles and covering emission as close to the beamline as possible.

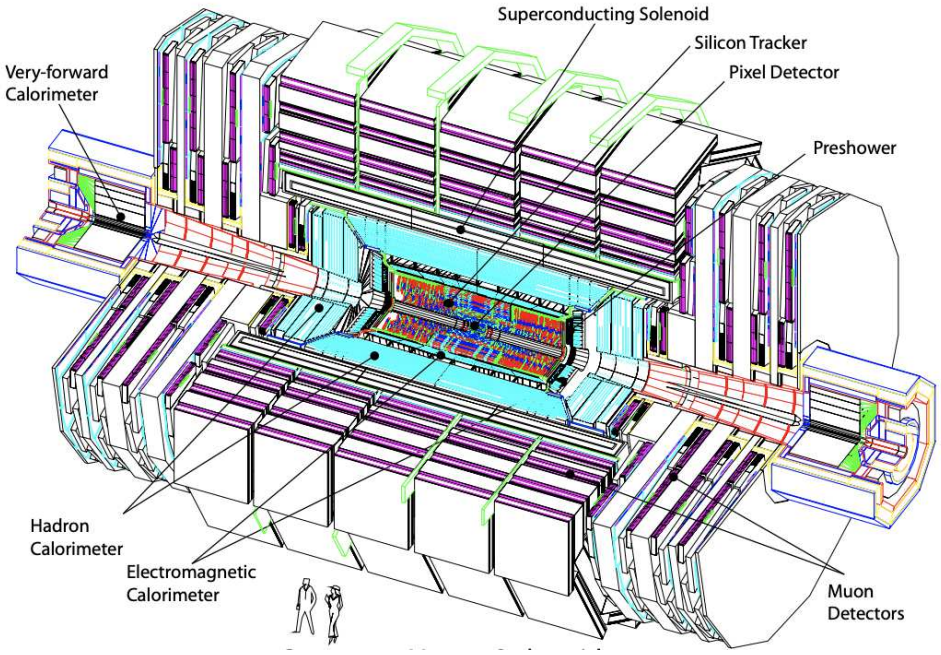


FIGURE 4.4: A pictorial representation of the CMS detector [77].

The LHC comprises of seven detectors: CMS, ATLAS, LHCb, ALICE, LHCf, TOTEM, and MoEDAL, each designed to probe different physics studies. For the purpose of this thesis, we only cover the CMS (Compact Muon Solenoid) detector, shown in Fig. 4.4.

The CMS experiment is a general-purpose detector designed with the goal of addressing a variety of SM questions and looking for new physics at the LHC. Its purpose is very similar to that of the high-resolution camera: it captures all stable particles after each collision, measures their kinematical properties, and glues all of the information into an "image" for investigating the underlying physics. To prepare these "images", the CMS experiment employs five major components, each of which determines the attributes of the particles created in a collision, namely, the magnets, the trackers, the Electromagnetic Calorimeter (ECAL), the Hadron Calorimeter (HCAL), and the muon detector.

Moving on to the geometry of the detector, we take the beamline to be the z-direction. The angular coordinates are defined as [78]

$$y = \frac{1}{2} \ln\left(\frac{\mathrm{E} + \mathrm{p_z}}{\mathrm{E} - \mathrm{p_z}}\right), \tag{4.20}$$

and

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right) = \frac{1}{2}\ln\left(\frac{|\mathrm{p}| + \mathrm{p_z}}{|\mathrm{p}| - \mathrm{p_z}}\right), \tag{4.21}$$

where y and η are the rapidity and pseudorapidity respectively. In the pseudorapidity definition, we have simply replaced E with |p| making it a massless approximation of the rapidity. Both quantities measure the amount of momentum an object has in the z-direction. The angle θ in Eq.(4.19) is the angle in the direction of the beamline, i.e., the z-direction. Another important coordinate around the beamline is the rotation angle ϕ, which starts from the x-axis and rotates in the $x - y$ plane.

4.3.2 Boosted Jets Topology

The high-energy regime of the LHC is often populated by boosted jets emerging from an intermediate heavy particle, and they are of great interest for numerous reasons [79, 80]. To highlight these reasons for studying boosted topologies at the LHC, we consider a simple example of the SM Higgs boson at rest decaying into bottom-antibottom quarks.

Following the conservation of momentum, it is clear to deduce that the decaying b-quarks will be back-to-back, forming two distinct b-jets after clustering. However, if we consider a Higgs with large momentum (what we mean by "boosted"), the resulting momenta of the $b\bar{b}$ pair would be directed towards the Higgs causing them to be mashed together. As the decaying Higgs is boosted more and more, the two b-jets become so close that they cannot be separated into two distinct objects by most of the clustering algorithms.

A way around this is to use a large R value to cluster all the final state particles into a single object, a "fat jet". The reconstruction of the invariant mass of this fat b-jet can then be targeted to deduce the Higgs mass. A diagrammatic representation of the formation of a fat jet is shown in Fig. 4.5.

Another reason to favor the fat jets study at the LHC is the reduction of backgrounds, as the jet substructure in this regime is quantitative enough to tell us whether the jet is signal-like or background-like. This is lucrative because the LHC environment suffers from large backgrounds making it difficult to search for new particles and underlying physics.

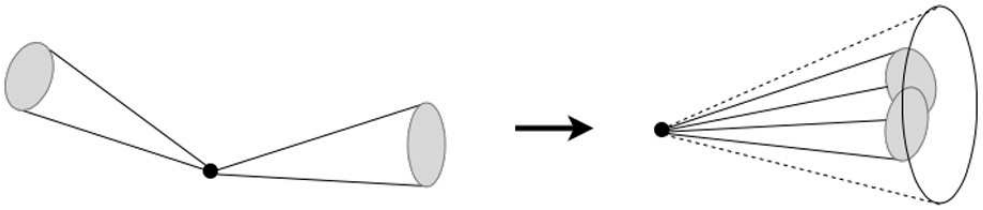


FIGURE 4.5: A diagrammatic representation of two jets merging into a fat-jet for boosted particle decay.

Apart from the boosted $h \to b\bar{b}$ example, there are several other processes in the literature that are susceptible to boosted topology and use fat jet analysis to target the intermediate heavy particles [79].

4.3.3 Choice of Clustering Algorithm and Jet Radius

After discussing the details of the CMS detector, boosted topologies, and mapping of the final state, it is important to identify which jet clustering algorithm would be best suited to cluster these particles into jets. In addition, the correct fixed-R value must also be determined [81].

To answer these questions, the LHC experiments have conducted a thorough investigation to determine the best-suited jet algorithms and R-values for multiple scenarios. At the LHC, the anti-k_T algorithm has become the default choice, with $R = 0.5, 0.7$ for the CMS and $R = 0.4, 0.6$ for the ATLAS. The purpose of using these R values is to cluster all the radiation from a given parton while avoiding sweeping in unwanted background junk.

Later on in this thesis, we will carry out an analysis and explore in detail the working of the anti-k_T algorithm with different fixed-R values as well as integrated variable-R clustering for the 2HDM Type-II scenario.

4.3.4 Jet Grooming and Substructure

So far, we have covered the reconstruction of the jets at the LHC for the simplest case possible, where the incoming radiations are from the hard interaction of interest. However, the LHC environment is crowded with busy hadronic events and unwanted radiation coming from:

- Multiple Parton Interactions (MPI), additional parton collisions coming from same protons collision of interest
- UE, the scattering proton remains
- PU, additional parton collisions coming from other protons.

When reconstructing jets at the LHC, these extra radiations play an important role in concealing the features of the analysis we are interested in. As a result, we need some sort of cleaning technique to get the relevant jets and mitigate these unwanted contributions. At the LHC, jet grooming and PU subtraction methods are utilised to mitigate the radiation inside a jet so that only particles of interest relevant to the jet substructure study are retained. In this section, we briefly discuss a few grooming and PU subtraction methods.

4.3.4.1 Jet Trimming

Jet trimming was introduced in [82] for cleaning up fat jets. This method reclusters the fat jet constituents with a smaller R and keeps a subset of subjets. After that, we only keep those subjets that pass a certain p_T threshold

$$p_{T,i} > f_{cut}\ \Lambda_{hard}, \tag{4.22}$$

where f_{cut} is a cut-off parameter and Λ_{hard} is a hard momentum scale. The sum of all these subjets makes up the final trimmed jet of interest. The notion is that the radiation coming from hard interactions of interest should be confined to clusters, such that reclustering them with a smaller R solves the issue of unwanted radiation. Usually, the k_T algorithm is preferred, but C/A and anti-k_T algorithms can also be used to recluster the constituents into subjets.

The other grooming method known as jet filtering [83] is quite similar to trimming in the sense that it recluster the constituents of a given jet with a smaller jet radius R_{filt} but instead of using p_T cut-off, it keeps only n_{filt} hardest of the subjets.

4.3.4.2 Jet Pruning

Another method for jet cleaning is pruning [84]. Pruning is based on the observation that removing soft, wide-angled emissions from a fat jet at the end of the

process helps improve the mass resolution and ability to extract important jet characteristics.

Pruning adapts the splitting algorithm to look for soft and wide-angled splittings and discards them by reclustering the constituents of a given jet with the C/A algorithm. The method then proceeds as follows

- For each splitting $r \to ij$, compute z and check whether that splitting is soft

$$z = \frac{min(p_{Ti}, p_{Tj})}{p_{Tr}} < z_{cut}, \tag{4.23}$$

  where z_{cut} is the threshold value.

- Next, check whether that splitting is at a wide angle,

$$\Delta R_{ij} > D_{cut}, \tag{4.24}$$

- If both conditions are satisfied, we drop the softer of i, j and continue unwinding the clustering until a suitable hard splitting is detected.

For each analysis, the input parameters z_{cut} and D_{cut} should be optimised according to the situation. The most common values for z_{cut} and D_{cut} are 0.1 and $\frac{m}{p_T}$ respectively.

4.3.4.3 Soft Drop Method

Similar to the pruning method, the soft drop algorithm [85] is commonly used to groom and study jet substructures. The soft drop method works in the same way as pruning but with a different particle removal condition

- Recluster a given jet into subjets j_1 and j_2 using the C/A algorithm.

- Keep that jet whose subjets satisfy

$$\frac{min(p_{T1}, p_{T2})}{p_{T1} + p_{T2}} > z_{cut} \left(\frac{\Delta R_{12}}{R_0} \right)^{\beta}. \tag{4.25}$$

- If a jet cannot be de-clustered, one can either keep the jet (grooming mode) or remove the jet from the list (tagging mode).

The z_{cut} and β input parameters can be tuned to improve performance. The method in tagging mode is infrared and collinear safe only when $\beta \leq 0$ whereas the grooming mode is infrared and collinear safe when $\beta > 0$. For $\beta = 0$, the soft drop method can be generalised to a (modified) mass-drop tagger (mMDT) [86].

4.3.4.4 PU Mitigation Techniques

Over the years, many PU mitigation methods have been developed by the theoretical and experimental communities. Early LHC approaches relied on reconstructing the PU vertices using charged tracks and making adjustments proportional to the number of vertices detected. During Run I and II of the LHC, the "area-median method" [87] was used to mitigate the PU impact. With the increase in the center of mass energy, the LHC environment has become much busier. As a result, many groups have developed techniques to fit the current situation. Some of these techniques are Pile-Up Per Particle Identification (PUPPI) [88], cleansing [89], SoftKiller [90], and Pile-Up Mitigation with Machine Learning (PUMML) [91]. Jet grooming techniques combined with the above-mentioned methods can also assist in limiting the sensitivity of jets to PU.

We will restrict our discussion to PUPPI, as it is widely utilised by CMS for PU reduction and substructure study. PUPPI uses interaction vertices to trace back the charged tracks and remove those that move away from the point of interest while keeping a check on the neutral radiation from PU events.

The method begins by defining a shape parameter α to estimate the likelihood of a particle originating from the PU event

$$\alpha_i = log \sum_{j \in event} \frac{p_{Tj}}{\Delta R_{ij}} \Theta(R_{min} \leq \Delta R_{ij} \leq R_0), \tag{4.26}$$

where Θ is the Heaviside function, R_{min} is the minimum cut-off that governs the collinear splittings from i and R_0 defines the cone surrounding the particle i. A α distribution can be plotted to identify all the particles coming from PU events. The neutral PU particles usually follow the same pattern as the charged ones allowing them to be distinguished from charged particles. This procedure then minimises the effect of the PU events to a great extent and allows for the search of particles of interest.

4.3.5 Jet Tagging

To properly analyse the underlying physics at the LHC, jet taggers are used to extract details about the original parton from which the final state jets originated. Here, we will limit our discussion to some of these taggers with a main focus on b-jet tagging.

Over the years, many taggers have been put forward by the theoretical and experimental communities. CMS and ATLAS extensively use these taggers to unravel final-state physics. For quark/gluon discrimination, CMS employs implicit and explicit taggers with a variety of variables, calibration methods, and validation techniques [92]. CMS has well-established taggers for $W^{\pm}$ bosons tagging [93] and top quark tagging [94]. Top tagging is very important at the LHC as it can be a powerful tool for accessing the dark matter candidate for some BSM scenarios. In their performance analysis, the CMS also included the b-tagging procedure.

Various jet tagging methods also combine machine-learning techniques to define the supervised classification problem and generate tagged training data using MC simulations. For b-tagging, one can feed the displaced vertices information to the ML model and then perform jet tagging. Recent CMS results show a 15% increase in tagging efficiency after including ML techniques [95]. Boosted top taggers have also used the jet substructure information by using a series of parameters [96, 97]. A recent review of ML approaches implemented on top taggers can be found in [98]. Apart from boosted top taggers, much work has gone into tagging boosted H, Z, and $W^{\pm}$ bosons [99] produced by new TeV heavy particle decays.

Since we will be exploring the impact of applying b-tagging approaches, we briefly outline the methodology behind it. The fundamental aspect of b-tagging methods is the lifetime of B-hadrons (hadrons with b-quark) [100]. In a b-quark decay, the lifetime is long enough for the B-hadrons to move away from the primary interaction vertex (known as the impact parameter). Using this information, one can trace back the tracks in the jet to check whether it originated from a b-quark or not; if they coincide with the secondary vertex, the jet is most likely to be classified as a b-jet. In our thesis, we will be using a much simpler version of b-tagging by using Monte-Carlo (MC) truth-level b-quarks information to tag jets for both boosted and non-boosted Higgs topologies; more on this later in the results section.

4.3.6 Boosted SM Higgs Boson and New Physics Searches at the LHC

Now that we have covered all the ingredients for clustering, grooming, and tagging jets, let us take a look at some of the CMS and ATLAS studies highlighting the importance of selecting the right jet clustering algorithm, jet radius R value, and grooming technique to unravel the particles of interest in the analysis.

The discovery of the Higgs Boson via $b\bar{b}$ decay mode was never considered at the LHC due to large QCD backgrounds. Nonetheless, this decay channel is of great importance and a significant contributor to the Higgs boson's total width. CMS conducted an extensive study to search for the boosted Higgs boson decaying into the $b\bar{b}$ pair using the anti-k_T algorithm with $R = 0.8$ and $p_T \geq 450$ GeV [101]. They have also used the soft drop grooming method to further enhance the sensitivity of the Higgs boson mass reconstruction. ATLAS also performed a similar study with the anti-k_T algorithm with $R = 1.0$ and $p_T \geq 480$ GeV [102]. Instead of using the soft drop procedure, ATLAS used trimmed jets to better reconstruct the mass of the Higgs boson.

In the BSM scenario, it is quite common to have a heavy resonance decaying into a pair of W/Z or into a pair of gauge bosons (W'/Z') coming from the extended sector. ATLAS used the anti-k_T algorithm with $R = 1.0$ and the trimming method to mitigate the effect of PU, whereas CMS used $R = 0.8$ and PUPPI to clean up jets for better kinematic distribution reconstruction. CMS and ATLAS have also invested in studying the decay of heavy resonances into top quarks, which can be a great source for developing jet substructure classification tools. Both of them focus on purely hadronic and dileptonic decay modes, with ATLAS again using the anti-k_T algorithm with $R = 1.0$ and the trimming method, whereas CMS uses $R = 0.8$ and PUPPI [103, 104]. ATLAS has also looked into $W' \to t\bar{b} \to q\bar{q}b\bar{b}$ with the same configuration and $p_T > 450$ GeV [105].

The takeaway conclusion from these studies is that the anti-k_T algorithm is the preferred choice for clustering the jets at the LHC. Given that the LHC environment is swamped with "extra junk", the grooming and PU mitigation techniques are actively used at the LHC. We also witnessed the importance of performing jet tagging to correctly identify the intermediate particles. In our results section, we will investigate the use of the anti-k_T algorithm with fixed R values (as used by CMS/ATLAS) for both resolved and boosted topologies. We will also use new

algorithms to check whether there is any improvement in obtaining the invariant mass peaks of the particles of interest in the presence of different resolution parameters, jet tagging, and reconstruction procedures.

Part II

Research and Results

Chapter 5

Revisiting Jet Clustering Algorithms for New Higgs Boson Searches in Hadronic Final States

This chapter is based on the work released in [1]. This work was co-written by Amit Chakraborty, Srinandan Dasmahapatra, Henry Day-Hall, Billy Ford, Stefano Moretti, Emmanuel Olaiya, and Claire Shepherd-Themistocleous.

5.1 Introduction

Following the general theme of the thesis, this chapter investigates methods for resolving signals from some BSM physics using the 2HDM. In particular, we investigate which jet clustering algorithm is best suited to resolve the specific final states of interest, notably from topologies with an extended Higgs sector coming from some 2HDM scenario in cascade decays. In such scenarios, high b-jet final states are anticipated, and it is important to address which experimental jet reconstruction procedure is in fact optimal for these kinds of searches.

As mentioned, BSM scenarios with an extended Higgs sector permit the presence of additional neutral Higgs states, CP-odd and CP-even. These resonances have the potential to be lighter or heavier than the SM-like Higgs boson discovered at the LHC in 2012 with a mass of 125 GeV [106]. These new physics frameworks are prevalent in both minimal and non-minimal models of SUSY [107], in particular, but not extensively, in the Next-to-Minimal Supersymmetric Standard Model

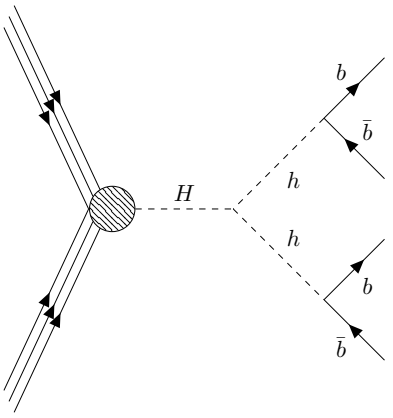


FIGURE 5.1: The 2HDM process of interest, where a heavier Higgs state H produced from gluon-gluon fusion decays into a pair of lighter scalar Higgs states, hh, each, in turn, decaying into $b\bar{b}$ pairs giving a $4b$ final state.

(NMSSM) [108]. However, if one deviates from SUSY and sticks to low-energy models, then the 2HDM is a rather simple BSM framework including light states in its particle spectrum [33, 34, 109].

As covered in Chapter 3, in 2HDM, the discovered 125 GeV Higgs boson can be identified as either h or H. In both cases, the decay $H \to hh$ and/or $H \to AA$ may occur if the conditions $m_h < m_H/2$ or $m_A < m_H/2$ are satisfied, respectively. These processes are commonly known as Higgs cascade decays. Then, taking $H(h)$ as the discovered 125 GeV Higgs boson, the dominant decay mode in a 2HDM for a $H(h)$ state with a mass order of 250(60) GeV bottom-antibottom quark pairs [110, 111], i.e., $h \to b\bar{b}$ such that the final states arising from the hard scattering $pp \to H \to hh$ comprises of four (anti)quarks, at the partonic level[1], see Fig. 5.1. However, due to the confinement properties of QCD, the partonic stage is not available to experiment, only the hadronic "jets" arising at the end of the parton shower and hadronisation phase are visible.

In order to determine the source of these hadronic showers, "jet clustering algorithms" are currently employed. Jet clustering algorithms reduce the complexity of such final states by attempting to rewind the showering back to the parton it originated, reducing a large sample of particles to a smaller number, each of which represents a state emerging from the hard interaction of interest in a given event. In other words, we consider a sample of particles arising from a single parton as an object - a jet. Needless to say, there are numerous jet clustering algorithms available, and we have discussed them at length in Chapter 4.

[1]Notice that the same argument can be made for the case of $pp \to H \to AA \to b\bar{b}b\bar{b}$ when $m_A < m_H/2$.

The goal of this chapter is to determine whether alternative jet reconstruction tools, specifically a modification to traditional sequential combinations algorithms using variable inter-jet distance measures [76] (so-called 'variable-R' algorithm, where R represents a typical cone size characterising the jet), are better suited for the $4b$ final state emerging from 2HDMs. We tackle this problem from a theoretical point of view in order to thoroughly examine a range of alternative combinations to determine whether a detailed experimental investigation is worthwhile. Additionally, the $4b$ final state that we are invoking here is a ubiquitous signal of BSM Higgs boson pairs[2], crucially providing access (through the extraction of the h/H state properties) to important aspects of the underlying BSM scenario, such as the shape of the Higgs potential and, consequently, its vacuum stability and perturbative phases.

While the problem of the optimal jet reconstruction is definitely an experimental endeavor, we emphasise that this study is conducted at a theoretical level. We perform a simple yet sophisticated MC event generation-based analysis to examine the relative performance of traditional fixed-R jet clustering against a variable-R method. A more extensive, realistic experimental investigation is left for a future study. For example, a major aspect of the hadronic final state initiated by b-quarks that we aim to analyse is that the emerging jets can be "tagged" as such, unlike lighter (anti)quarks and gluons, which are generally indistinguishable from each other. In this chapter, we implement a simplified tagging method based on MC truth information on b-partons, along with a probabilistic implementation of inefficiencies. For a more in-depth discussion on b-tagging at the detector level, we direct the reader to [112].

The layout of the chapter is as follows. In the next few sections, we detail how we performed b-tagging, the tools used for our simulations, and the cutflow adopted. Next, we present our results for both signal and background. Then we conclude.

[2]Here, ubiquitous refers to the fact that this signal is very typical of a variety of BSM scenarios so that we effectively use the 2HDM for illustration purposes. Our results can therefore be applied to the case of other new physics models.

5.2 Methodology

5.2.1 Implementation of b-Tagging

In this study, we implement a simplified MC-informed b-tagger. For events clustered with a fixed-R cone size, we search jets within angular distance R from a parton level b-(anti)quark and tag them accordingly. In cases where multiple jets are found, we select the closest and assign it a b-tag. When the variable-R approach is used, we set the size of the tagging cone is taken to be the effective size R_{eff} of the jet.

Furthermore, we account for the finite efficiency of detecting a b-jet as well as the non-zero probability that c-jets and light-flavour plus gluon jets are mis-tagged as b-jets. We apply the variable mis-tag rates and tagging efficiencies from a specific `Delphes` CMS detector card[3].

5.2.2 Simulation Details

We investigate two sample Benchmark Points (BPs): BP1 and BP2. In BP1, we consider the lighter Higgs to be the SM-like Higgs boson with $m_h = 125$ GeV and the heavier Higgs to be $m_H = 700$ GeV. For BP2, the heavier Higgs has $m_H = 125$ GeV with the lighter Higgs mass m_h set to 60 GeV. Both benchmarks are in a 2HDM Type-II and have been tested (and passed as not currently excluded) against theoretical and experimental constraints by using 2HDMC [40], HiggsBounds [113], HiggsSignals [114] as well as checking flavor constraints with SuperISO [115]. We generate samples of $\mathcal{O}(10^5)$ events, with $\sqrt{s} = 13$ TeV for the LHC energy. In SuperISO, we test our BPs against the following flavor constraints on meson decay Branching Ratios (BRs) and mixings, all to the 2σ level: BR($b \to s\gamma$), BR($B_s \to \mu\mu$), BR($D_s \to \tau\nu$), BR($D_s \to \mu\nu$), BR($B_u \to \tau\nu$), $\frac{BR(K\to\mu\nu)}{BR(\pi\to\mu\nu)}$, BR($B \to D_0\tau\nu$) and $\Delta_0(B \to K^*\gamma)$.

The production and decay rates for the subprocesses $gg, q\bar{q} \to H \to hh \to b\bar{b}b\bar{b}$ are presented in Tab. 3.2, alongside the 2HDM Type-II input parameters (see Point1 and Point2 in Tab. 3.2). (Notice that the H and h decay widths are of order MeV; since this is much smaller than the detector resolutions in two-jet and four-jet invariant masses, the Higgs states can essentially be treated as on-shell.) In the

[3]See `https://github.com/delphes/delphes/blob/master/cards/delphes_card_CMS.tcl`.

calculation of the overall cross-section, the renormalisation and factorisation scales were both set to be $H_T/2$, where H_T is the sum of the transverse energy of each parton. The PDF set used was `NNPDF23_lo_as_0130_qed` [116]. Finally, in order to carry out a realistic MC simulation, the toolbox described in Fig. 5.2 was used to generate and analyse events [117, 118, 119, 120, 121][4].

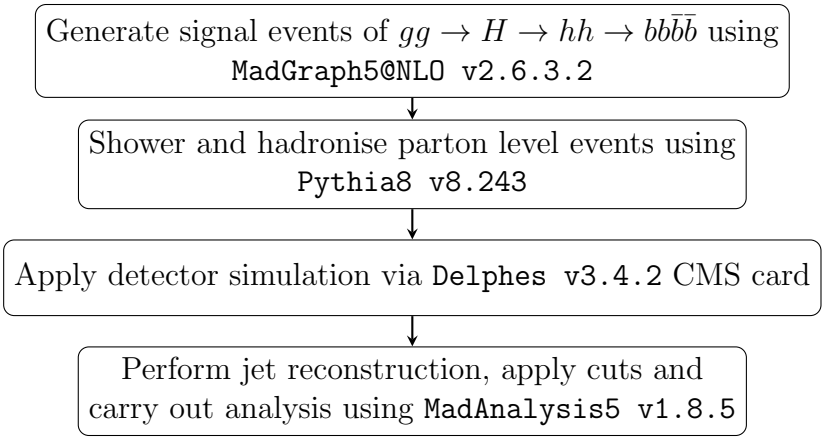


FIGURE 5.2: Description of the procedure used to generate and analyse MC events.

5.3 Cutflow

The introduction of the entire series of cuts that we have used here demands some justification. Existing four b-jet analyses by the ATLAS and CMS collaborations seek to isolate the chain decays of Higgs bosons from the background by adopting restrictive cuts at the trigger level for the resulting fully hadronic signature. For BP1, the p_T cuts informed by choices made at CMS by [122] of all four b-jets satisfying $p_T > 50$ GeV are used (at trigger level). Upon enforcing the same trigger level p_T cuts as in CMS [122] on BP2 for Run 2 and 3 luminosities, we discovered that the signal selection efficiency was too low to provide an acceptable MC sample for phenomenological analysis. To get a visible signal at the LHC, we use a flat cut on all four b-tagged jets of 20 GeV. It remains to be seen whether this is feasible at the LHC, but it produces samples of a useful size for comparing the behavior of different jet clustering algorithms. We indeed provide results in this

[4]Note that to be consistent with the Leading Order (LO) implementation of the background cross sections below, we use the LO normalisation for the signal cross sections here. Even though this affects our final results on event rates and significances, the main purpose of our paper is to assess the performance of different jet clustering algorithms, which should be unaffected by the exact values of signal and background rates.

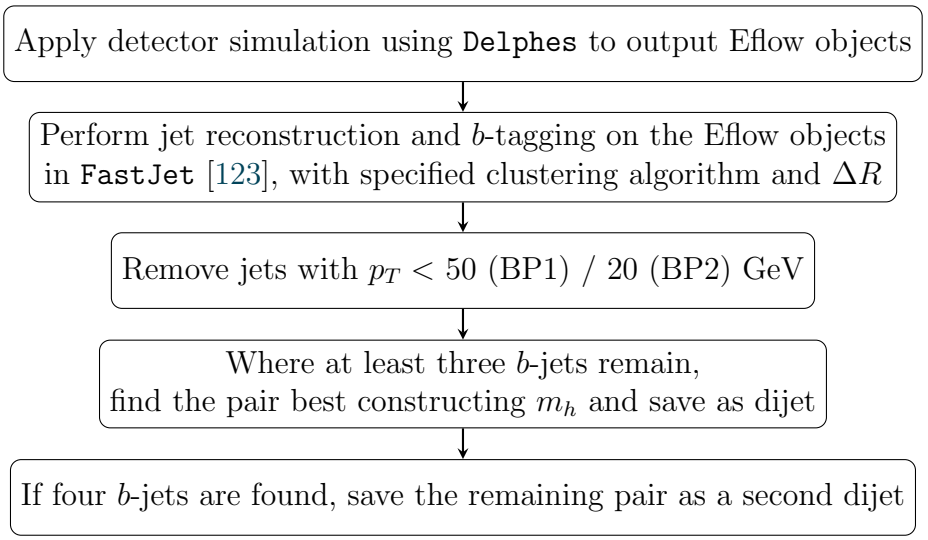


FIGURE 5.3: Description of our initial procedure for jet clustering, b-tagging and selection of jets. Note that the bulk of our analysis is performed at particle rather than detector level, so MC truth information is used for cuts on jet constituents.

regime to demonstrate the utility of using a variable-R jet reconstruction algorithm on low-p_T jets from 2HDM-II decays into $b\bar{b}b\bar{b}$ final states. Fig. 5.3 describes our initial procedure for jet clustering, b-tagging, and selection of jets.

5.4 Results

In this section, we present the results for our signal at both the parton and detector levels. In the latter case, we also discuss the dominant backgrounds, due to QCD $4b$ production, $gg, q\bar{q} \to Zb\bar{b}$ and $gg, q\bar{q} \to t\bar{t}$[5].

5.4.1 Parton Level Analysis

All the events at the matrix element level have four b-quarks originating from the decay of the two light Higgs bosons (h). We plot the R separation between the b-quarks originating from the same light Higgs state (see the upper panel of Fig. 5.4). The two distributions associated with BP1 and BP2 are labeled differently. In general, the angular separation between the decay products a and b

[5]In fact, we have checked that the additional noise due to $t\bar{t}b\bar{b}$ events as well as hadronic final states emerging from W^+W^-, $W^\pm Z$ and ZZ production and decay is negligible, once mass reconstruction around m_h and m_H is enforced.

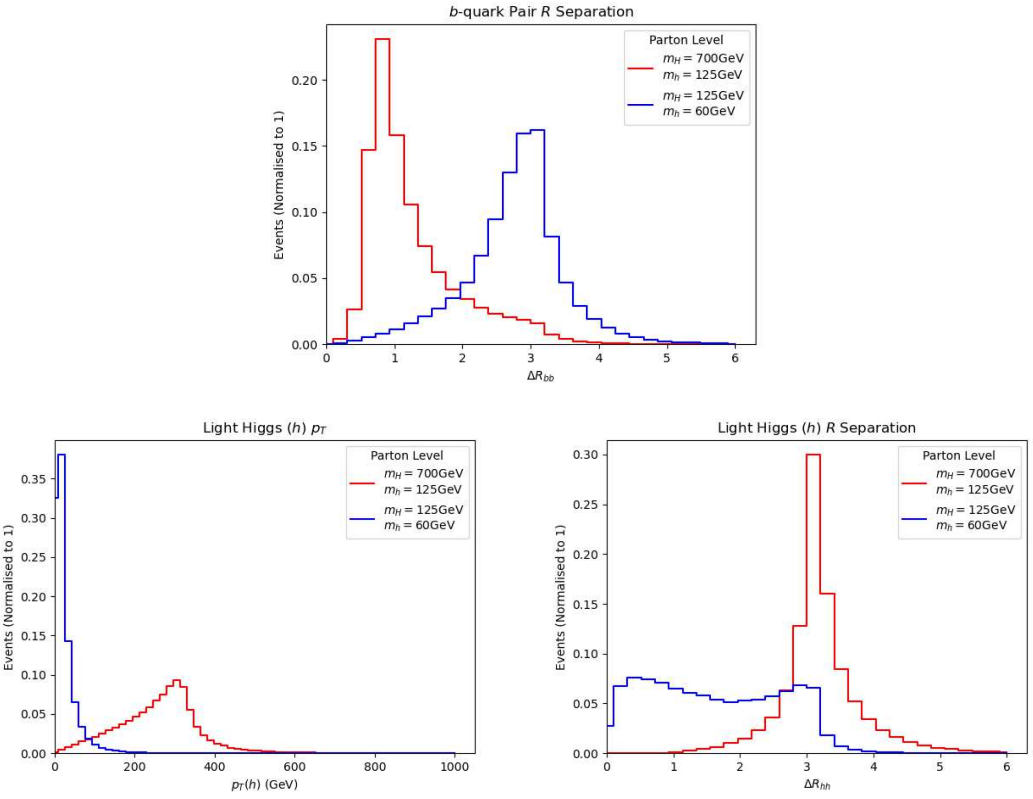


FIGURE 5.4: Upper panel: the ΔR distribution between the two b-partons originating from the same h. Lower panel: the p_T distribution of the light Higgs boson h originating from H decay (left) and the ΔR distribution between the two h states originating from the H decay (right). No (parton level) cuts have been enforced here.

in the resonant process $X \to ab$ can be approximated as $\Delta R(a,b) \sim \frac{2m_X}{p_T^X}$. Hence, we plot in the lower panel of Fig. 5.4 the transverse momentum of each of the h bosons. For $m_h = 60$ GeV, the light Higgs boson has less p_T than for lower values of m_h due to the smaller $m_H - m_h$ mass difference. This leads to b-quarks being widely separated in comparison to $m_H = 700$ GeV. In light of this, we can already conclude that there is a strong correlation between the mass difference $m_H - m_h$ and the cone size of the jet clustering algorithm that we want to use. Particularly, to maximise the number of jets[6] formed by a clustering algorithm for different choices of the mass of the light Higgs boson, it may be necessary to vary the jet radius parameter instead of having a fixed radius. In the lower panel of Fig. 5.4, we plot the ΔR separation between the two light Higgs states. For the configuration in BP1, it is clear (since $\Delta R \approx \pi$) that the $H \to hh$ decay is predominantly back-to-back (in the laboratory frame). However, for $m_h = 60$ GeV, there is a double peak structure due to a recoil effect from Initial State Radiation (ISR), which

[6]This is done also with a view to background rejection.

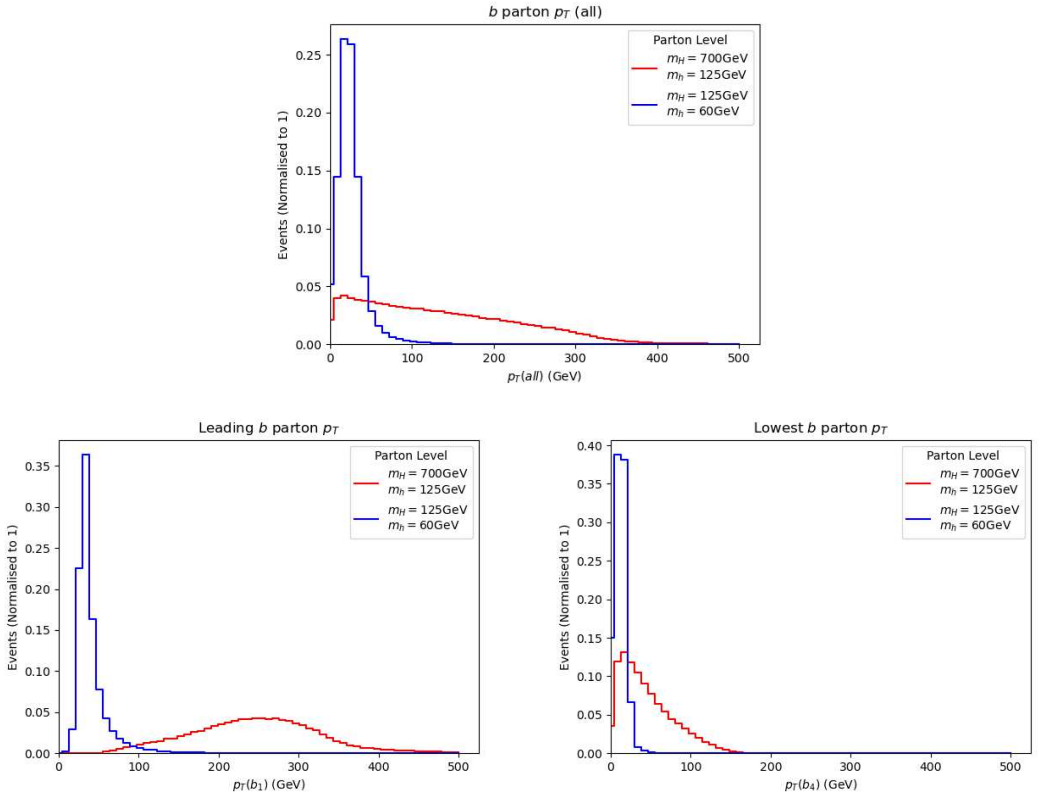


FIGURE 5.5: Upper panel: the p_T distribution for all b-quarks. Lower panel: highest p_T amongst the b-quarks (left) and lowest p_T amongst the b-quarks (right). No (parton level) cuts have been enforced here.

only becomes apparent at the mass boundary where $m_H \simeq 2m_h$. The inability of the two emerging h states to fly apart implies the overlapping of momenta from b-quark showers. Hence, we expect that the output of the clustering algorithm will have a high b-jet multiplicity as long as the two b-jets stemming from h decays are resolved depending on detector acceptance and signal selection cuts. Since m_h is small compared to typical jet p_T thresholds used in applying b-tagging, the multiplicity of jets can be reduced. We will investigate this later. As a final study, we plot b-quark p_T distributions in Fig. 5.5. From the top histogram, we can see that the p_T's of b-quarks span the range of possible values for both mass configurations and expect the resulting jets to have a similar kinematic spread. In particular, we also plot the highest and lowest p_T's amongst the b-quarks in a given event (lower frame): notice a stark difference in both cases. Further to the discussion in Section 5.2.2, one would therefore expect the resulting spread of radiation from each signal b-quark to vary in solid angle and hence the resulting jets to be of differing sizes. This motivates the need for a jet reconstruction sequence that adapts to jets of various cone sizes. Therefore, in the next section, we first test how jet clustering with fixed-R input behaves and then introduce the

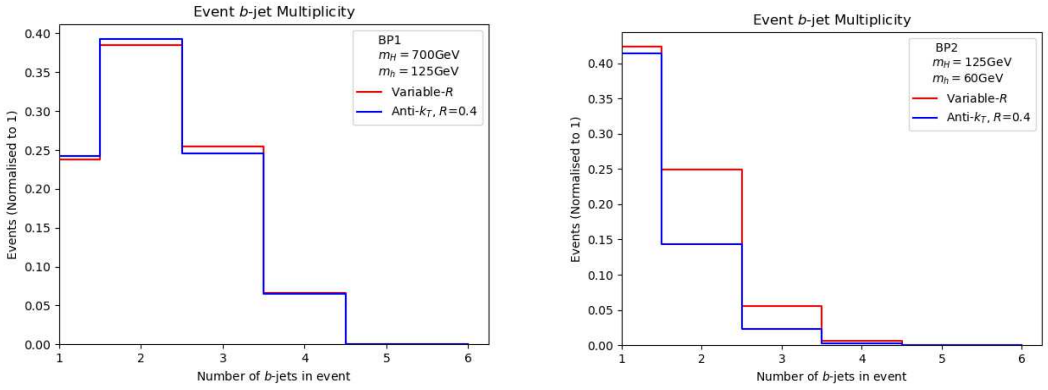


FIGURE 5.6: Left panel: The b-jet multiplicity distributions for BP1. Right panel: For BP2.

variable-R algorithm's performance.

5.4.2 Jet Level Analysis

In this section, we consider a jet-level analysis using hadronised parton showers that have been run through detector simulation and clustered into jets. We will compare the kinematic distributions of final state b-jets, when clustered with a fixed cone as well as with the variable-R, for both mass configurations in BP1 and BP2. In particular, we will be interested in the b-jet multiplicity, that is, the number of b-tagged jets in a given event. This is of course, indicative of how well our clustering is performing, in that we know the final state has four b-quarks and a good algorithm should recover all four. We will also investigate the mass distributions of b-dijets and four b-jet masses, which indicate our ability to observe the signals containing BSM Higgs bosons.

We first consider the effect of a variable versus fixed cone algorithm by observing kinematic variables from signal events for each BP. We choose a value of $R = 0.4$ and use the anti-k_T algorithm throughout. (The results for the C/A scheme are very similar, so we refrain from presenting them.) For variable-R, we use $\rho = 100$ GeV for BP1 and $\rho = 20$ GeV for BP2. These values are informed by the p_T scale of the fixed cone b-jets. Finally, we use $R_{min} = 0.4$ and $R_{max} = 2.0$ throughout wherever variable-R is used.

Fig. 5.6 depicts the b-jet multiplicity for each of the benchmarks/algorithms. The stark contrast between the two plots is caused by the relative kinematics of the final state b-jets. Due to the different mass configurations, b-jets from BP2 have

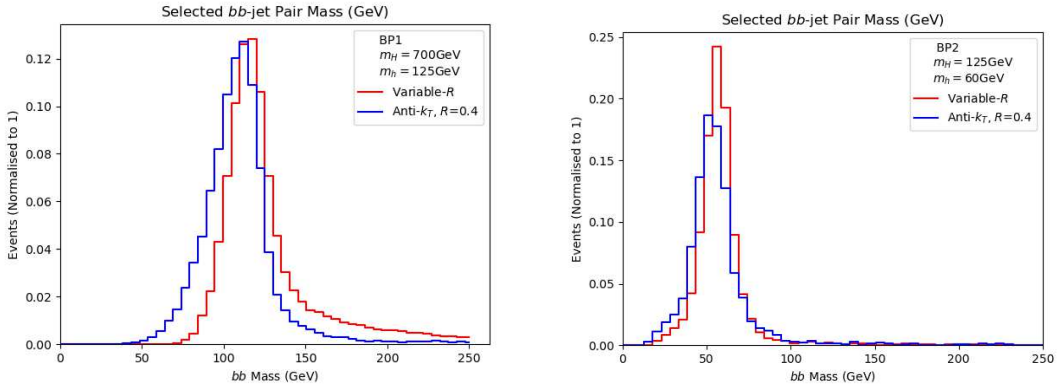


FIGURE 5.7: The invariant mass m_h distributions from dijets for BP1 (left panel) and BP2 (right panel). The peak of the mass distribution obtained from the variable-R algorithm is closer to the MC truth value of the corresponding Higgs.

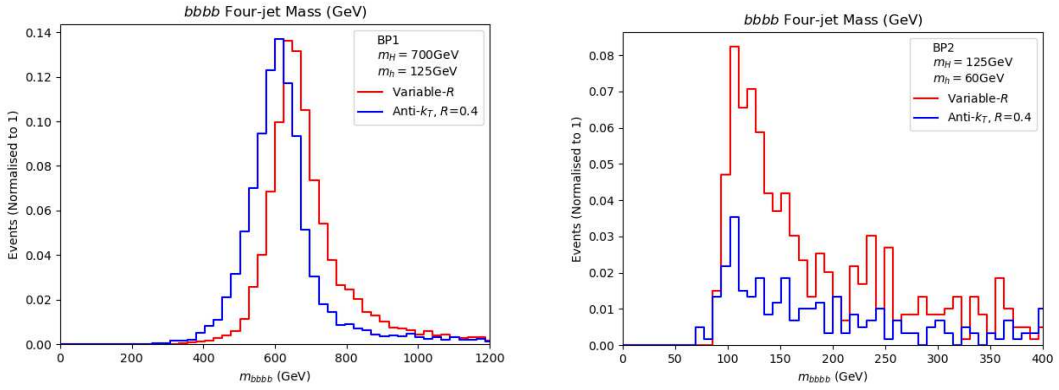


FIGURE 5.8: The four b-jet invariant mass m_H distributions from four jets obtained from jet clustering for BP1 (left panel) and BP2 (right panel).

significantly lower p_T than those from BP1, so significantly more are lost to the trigger as well as from the (p_T dependent) b-tagging efficiencies. The variable-R algorithm resulted in a small increase in events reconstructed with a higher b-jet multiplicity for BP1 but with a more significant shift for BP2.

In order to extract evidence of new physics from b-jet signals, we look at the invariant mass of dijets in order to reconstruct the mass of the resonance from which they originated.

We can see from Fig. 5.7 more definitively the benefits of using a variable-R jet clustering algorithm. The invariant mass m_h distributions from jet clustering for BP1 (left panel) and BP2 (right panel) are shown. The peak of the mass distribution obtained from the variable-R algorithm is closer to the MC truth value of the corresponding Higgs resonance. The same behavior for four b-jets

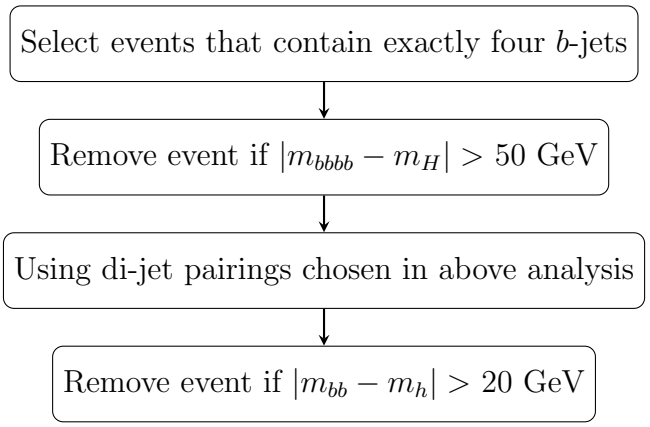


FIGURE 5.9: Event selection used to compute the signal-to-background rates.

masses can be seen in Fig. 5.8; events clustered with the variable-R have the $4b$ invariant mass more closely aligned with the expected positions at m_H.

5.4.3 Signal-to-Background Analysis

A good algorithm should not just boost the signal but also avoid sculpting the backgrounds. As a last exercise, we perform a calculation of the signal-to-background rates, so as to compare the various jet reconstruction procedures mentioned in this paper in connection with their performance in dealing with events not coming from our BSM process. To do this, we employ the selection procedure described in Fig. 5.9. We use the anti-k_T measure throughout, but conclusions would not change in the case of the C/A one.

5.4.3.1 Jet Quality Cuts

Prior to assessing the significance of the signal-to-background characteristics of the two methods, we employ jet quality cuts [76]. We start by defining the energy and the p_T center of the jets

$$P_E = \sum_{i \in jet} E_i \hat{p}_i, \;\; P_{p_T} = \sum_{i \in jet} p_{Ti} \hat{p}_i, \tag{5.1}$$

where i defines the constituents of the jet and $\hat{p}_i$, E_i and p_{Ti} are the four-momenta (normalised to unity), energy, and transverse momentum of the i^{th} constituent, respectively. The distance between P_E and P_{p_T} is then required to be within

$$\Delta R(P_E, P_{p_T}) < \delta, \tag{5.2}$$

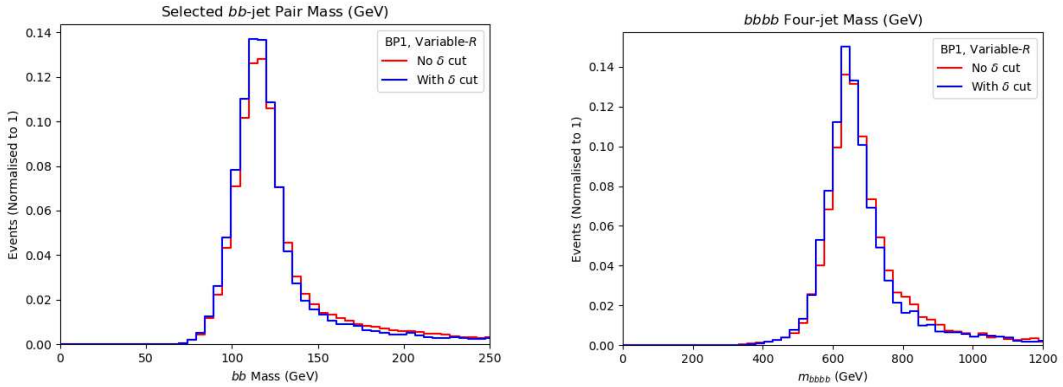


FIGURE 5.10: Left panel: The b-dijet invariant masses for BP1, with and without the addition of jet quality cuts as defined in Eqs.(5.1)–(5.2). Right panel: The four b-jet invariant mass. Here we have used a value of $\delta = 0.05$ for BP1.

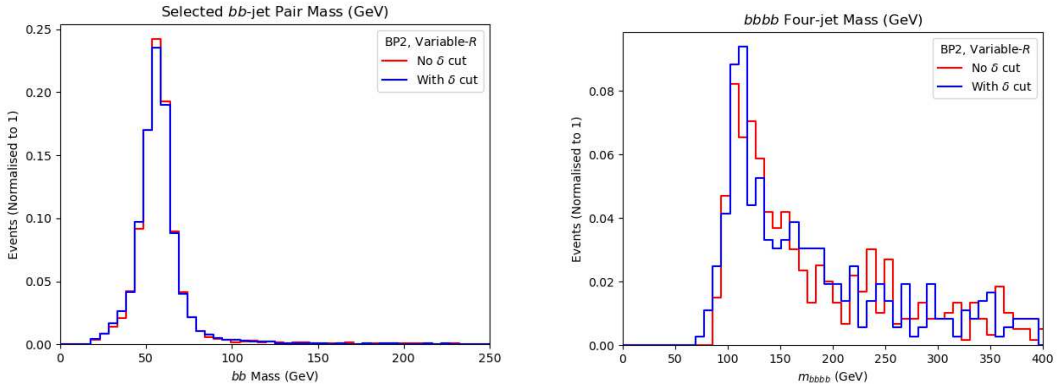


FIGURE 5.11: Left panel: The b-dijet invariant masses for BP2, with and without the addition of jet quality cuts as defined in Eqs.(5.1)–(5.2). Right panel: The four b-jet invariant mass. Here we have used a value of $\delta = 0.1$ for BP2.

where δ is a user-defined cutoff. To gain an idea of how useful jet quality cuts are, we plot b dijet and four b-jet invariant mass peaks corresponding to m_h and m_H with or without quality cuts in Fig. 5.10–5.11 (Here, we have used a value of $\delta = 0.05$ for BP1 and $\delta = 0.1$ for BP2).

Indeed, we can see that using jet quality cuts resulted in higher mass peaks for both cases (see Fig. 5.10 for BP1 and Fig. 5.11 for BP2). Finally, we further note that, while there is a hint of signal modification with jet quality cuts (see the lower panel of Fig. 5.10), the main gains come from the reduction of backgrounds, thereby obtaining higher significances, which we will see in the following section. In this work, we choose the values of the jet quality cut parameter δ for our BPs following Ref. [76], however, we suggest optimisation of δ for individual heavy

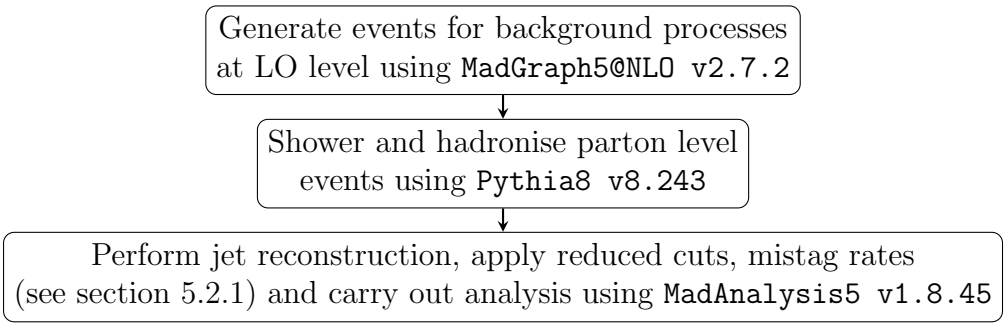


FIGURE 5.12: Description of the procedure used to generate and analyse MC events for background processes.

resonance masses to obtain higher significances. Note that one can also demand the jets to lie in the central region of the detector only or else use the method related to the catchment area of a jet [124] as outlined in reference [76], in order to have better control of the backgrounds.

5.4.3.2 Signal Selection

To carry out this exercise, we generate and analyse $pp \to b\bar{b}b\bar{b}$, $pp \to Zb\bar{b}$ and $pp \to t\bar{t}$ background processes using the toolbox described in Fig. 5.12 [117, 118, 120, 121][7]. Tab. 5.1 contains the cross sections in pb for signal and the various background processes upon applying the aforementioned cuts and mass selections, including the jet quality cuts.

It is clear from the data obtained that the QCD-induced $pp \to b\bar{b}b\bar{b}$ process is the dominant background channel[8], followed by $pp \to t\bar{t}$ and $pp \to Zb\bar{b}$. Our next step is then to calculate the event rates in order to get the significances for two values of (integrated) luminosity, e.g., $\mathcal{L} = 140$ and 300 fb^{-1}, corresponding to full Run 2 and 3 data samples, respectively. The event rate (N) for the various processes is given by

$$N = \sigma \times \mathcal{L}. \tag{5.3}$$

[7]We have also checked the non-resonant $pp \to hh \to b\bar{b}b\,barb$ background for both the BPs. For BP1, in the presence of the described kinematical selections and mass selection criteria described in Fig. 5.9, the number of events surviving is negligible. For BP2, we get a very small contribution in comparison to the other three backgrounds. Hence, for our MC studies, we do not consider this background.

[8]In fact, we have computed the full four-jet sample produced by QCD, i.e., including all four-body partonic final states, yet, in presence of the described kinematical selections and b-tagging performances, the number of non-$b\bar{b}b\bar{b}$ events surviving is negligible [125, 126, 127].

Process	variable-R		$R = 0.4$	
	BP1	BP2	BP1	BP2
$pp \to H \to hh \to bbbb$	2.077×10^{-4}	8.962×10^{-3}	1.254×10^{-4}	3.210×10^{-3}
$pp \to b\bar{b}b\bar{b}$	3.798×10^{-3}	2.131	1.651×10^{-3}	9.470×10^{-1}
$pp \to t\bar{t}$	7.973×10^{-4}	2.850×10^{-2}	1.595×10^{-3}	2.217×10^{-2}
$pp \to Zb\bar{b}$	9.689×10^{-6}	2.627×10^{-2}	3.876×10^{-6}	9.695×10^{-3}

TABLE 5.1: Cross sections (in pb) of signal and background processes upon enforcing the initial cuts plus the mass selection criteria $|m_{bbbb} - m_H| < 50$ GeV and $|m_{bb} - m_h| < 20$ GeV for the various jet reconstruction procedures.

	variable-R	$R = 0.4$
BP1	1.145	0.823
BP2	2.268	1.214

	variable-R	$R = 0.4$
BP1	1.881	1.366
BP2	3.707	1.984

TABLE 5.2: Upper panel: Final Σ values calculated for signal and backgrounds for $\mathcal{L} = 140$ fb^{-1} upon enforcing the initial cuts plus the mass selection criteria. Lower panel: Final Σ values calculated for signal and backgrounds for $\mathcal{L} = 140$ fb^{-1} with K-factors upon enforcing the initial cuts plus the mass selection criteria.

After the event rates have been calculated, we simply evaluate the significance, Σ, which is given by (as a function of signal S and respective background B rates)

$$\Sigma = \frac{N(S)}{\sqrt{N(B_{b\bar{b}b\bar{b}}) + N(B_{Zb\bar{b}}) + N(B_{t\bar{t}})}}. \tag{5.4}$$

Tabs. 5.2– 5.3 contains the significances before and after the K-factors have been applied [9]. We have used $K = 2$ for the signal [128, 129], $K = 1.5$ for the $pp \to b\bar{b}b\bar{b}$ process [130], $K = 1.4$ for $pp \to t\bar{t}$ [131] and $K = 1.4$ for $pp \to Zb\bar{b}$ [132]. It is then clear from Tabs. 5.2–5.3 that the variable-R approach provides better significance compared to those obtained from a fixed-R, for all choices of R evaluated with and without K-factors. The increase in significance is indeed considerable. This is not surprising given the ability of the variable-R to outperform the fixed-R approach in terms of kinematics. Again, we have used the anti-k_T algorithm here, but the conclusion remains the same for the C/A case.

[9] Recall that the cross-section after the cuts is contained in Tab. 5.1. These values are then used to compute the final event rate using Eq. (5.3), which is then used to calculate the $S/\sqrt{B}$ ratios in Tabs. 5.2– 5.3. Of course, the ratio is small, so an attempt to increase it would be necessary.

	variable-R	$R = 0.4$
BP1	1.676	1.205
BP2	3.320	1.777

	variable-R	$R = 0.4$
BP1	2.753	2.000
BP2	5.426	2.905

TABLE 5.3: Upper panel: Final Σ values calculated for signal and backgrounds for $\mathcal{L} = 300$ fb^{-1} upon enforcing the initial cuts plus the mass selection criteria. Lower panel: Final Σ values calculated for signal and backgrounds for $\mathcal{L} = 300$ fb^{-1} with K-factors upon enforcing the initial cuts plus the mass selection criteria.

5.4.4 Variable-R and PU

It has been noted that the nature of variable-R, combined with our reduced p_T restrictions, allows for wider cone signal b-jets. We, therefore, perform an analysis of events with PU and MPIs, using the variable-R algorithm. As briefly mentioned, in order to perform such a study, a proper detector simulation is required. We therefore now employ the use of `Delphes` [119], passing our hadronised events (simulated in `Pythia8`) through the CMS card (with the same b-tagging efficiencies and mistag rates as before). Specifically, for PU simulations, we have used `Pythia8` to generate soft QCD events. Mixing of these PU events with the signal events is then done within the Delphes CMS card, where we have used $\langle N_{PU} \rangle = 50$ for each hard scattering. We also perform the same exercise with a fixed cone (anti-k_T) of $R = 0.4$ to compare.

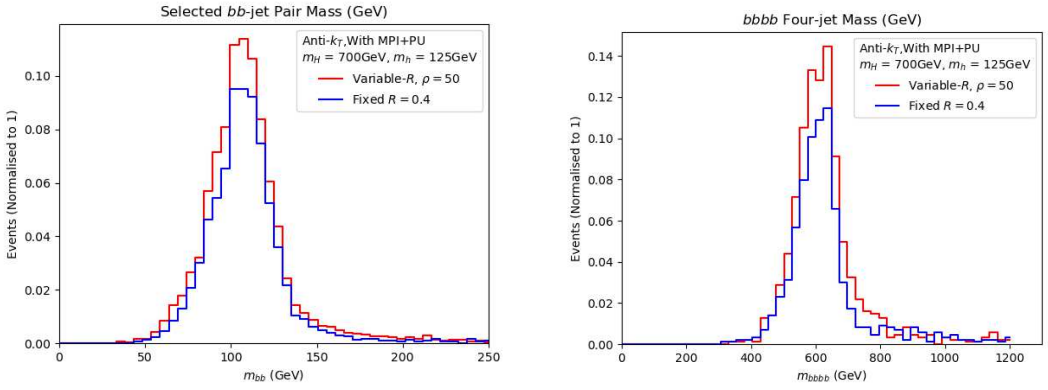


FIGURE 5.13: Left panel: The b-dijet invariant masses for BP1, using variable-R and fixed-R clustering, when considering the effect of PU and MPIs. Right panel: The same for the 4b-jet invariant mass.

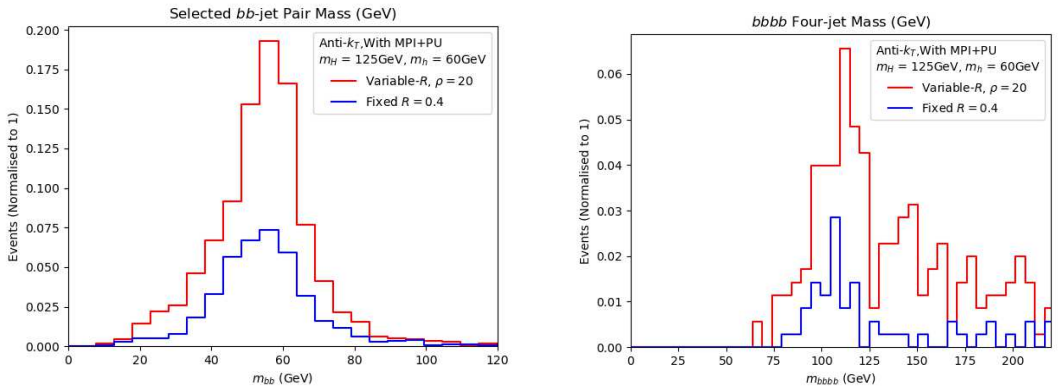


FIGURE 5.14: Left panel: The b-dijet invariant masses for BP2, using variable-R and fixed-R clustering, when considering the effect of PU and MPIs. Right panel: The same for the 4b-jet invariant mass.

In Figs. 5.13– 5.14, we present the dijet mass m_{bb} and four-jet mass m_{bbbb} spectra for the signal with PU and MPI for the fixed $R = 0.4$ and the variable-R jet clustering techniques. We note that, with the addition of PU, we have had to use a different value for the variable-R parameter ρ, i.e., $\rho = 50$. We do not consider jet quality cuts for PU-enhanced events here, as they were used for background reduction in the previous subsection. We see that, with PU added on top of our signal events, more events are selected following a variable-R jet reconstruction in comparison to fixed-R.

As a final point, we note that a further PU mitigation technique is possible in variable-R, which is in the values chosen for the $R_{min/max}$ variables. Clearly, if for some particular process, one discovers that using a variable-R reconstruction sweeps in too much extra 'junk' into the jets, a simple reduction of R_{max} is always possible. (Notice that, in order to reduce the contamination due to PU and MPIs, one can always use grooming techniques such as filtering trimming [82], filtering [83], pruning [84], mass-drop [83], soft drop [85] and modified mass-drop [86], however, this is beyond the scope of this work).

5.4.5 Other Variable-R Studies

Before concluding, we review here some other studies from the literature that use a variable-R reconstruction procedure.

We note that, while the leading b-jet has an R_{eff} roughly in line with expected values ($R_{eff} \simeq 0.5$), the lowest p_T b-jets have large cone sizes ($R_{eff} > 1.0$), risking potential contamination from additional radiation. This effect is discussed in [133]. Despite not implementing any vetoes, our results suggest that the variable-R clustering algorithm shows an improvement over other traditional clustering methods.

There have been other studies using the variable-R approach for physics searches, such as in the highly boosted object tagging of $hh \to b\bar{b}b\bar{b}$ decays in [134]. The variable-R algorithm was also used to analyse heavy particle decays in [133]. In both examples, an improvement over current fixed-R methods is present when using variable-R, which is in line with our findings.

As a final word on using variable-R jet reconstruction in experiments, we discuss its use in relation to b-tagging performance. In particular, the studies of Refs. [135, 136], explore the possibility of Higgs to b-jet tagging at ATLAS using variable-R techniques. Specifically, since these studies deal with boosted topologies, focusing on fat b-jet substructure, the advantage of applying these techniques in a non-boosted regime is yet to be determined.

5.5 Conclusions

In this chapter, we have assessed the potential scope of the LHC experiments in accessing BSM Higgs signals induced by cascade decays of the 125 GeV SM-like Higgs state discovered in July 2012 into two lighter Higgs states or indeed of a heavier one into pairs of it. The prototypical production and decay channel that we have used is $gg, q\bar{q} \to H \to hh$, where h is the lighter Higgs state and H is a heavier Higgs state, with mass greater than $m_H/2$, so as to induce resonant production and decay within 2HDM Type-II framework, thereby enhancing the overall rate. Either light Higgs boson would decay to $b\bar{b}$ pairs, eventually leading to a four b-jet signature, largely independently of the BSM construct hosting it.

The four-jet signature is extremely difficult to detect at the LHC, owing to the large hadronic background. Thus, b-tagging techniques need to be exploited in order to make such a signal visible. However, the conflict between tagging efficiency and signal retention poses a problem because these taggers are most efficient when b-jets have a large transverse momentum, say, at least 20 GeV, and at this scale, there is a significant loss of signal events if the BSM Higgs mass is in the sub-60 GeV

range. Hence, if one intends to maximise sensitivity to this hallmark signature of BSM physics, a thorough reassessment of the current Run 2 approaches is mandated, especially in view of the upcoming Run 3.

We observe that with current p_T cuts on the final state b-jets, using a fixed-R jet reconstruction and tagging procedure will lead to poor signal visibility, with a majority of signal b-jets being lost. We instead presented a reduced cutflow, based on existing $b\bar{b}\mu^+\mu^-$ analyses, and showed that this indeed provides a window onto $gg \to H \to hh \to b\bar{b}b\bar{b}$ signals with $m_H = 125$ GeV and $m_h < \frac{m_H}{2}$.

Additionally, and perhaps more remarkably, we also tested the variable-R algorithm approach on events with this reduced cutflow and showed a significant improvement in signal yield as well as signal-to-background rates. We notice that in final states of this kind, the signal b-jets have a wide range of p_T and hence a varied spread of constituents. Using a fixed cone of a standard size ($R = 0.4$) constructs well-higher p_T jets in an event but does not capture much of the wider angle radiation from lower p_T jets. This leads to two issues. Firstly, it will prove difficult to accurately construct m_h and m_H in the two- and four-jet invariant masses. Secondly, these jets will more often be lost due to kinematic cuts.

We have obtained all of the above in the presence of a sophisticated MC event simulation based on exact scattering MEs, state-of-the-art parton showers, hadronisation, and B-hadron decays, as well as a detector simulation. Given the results of our analysis, we recommend a more thorough detector-level analysis which is being undertaken for a variety of different high b-jet multiplicity scenarios, to explore whether a shift to variable-R jet clustering could be implemented and improve current signal significance limitations using fixed-R jet reconstruction. In fact, while we have quantitatively based our case on the example of the 2HDM Type-II, our procedure can identically be used in other BSM constructs featuring the same Higgs cascade decay.

Chapter 6

Fat b-Jet Analyses Using Old and New Clustering Algorithms in New Higgs Boson Searches at the LHC

This section is based on the work released in [2]. This paper was co-authored by Amit Chakraborty, Srinandan Dasmahapatra, Henry Day-Hall, Billy Ford, and Stefano Moretti.

6.1 Introduction

In Chapter 5, we studied the process $gg, q\bar{q} \to H \to hh \to 4b$, for $m_H = 125$ GeV and m_h between 40 and 60 GeV, which would be a striking signal of, for example, a 2HDM [33, 34, 109] in the so-called 'inverted hierarchy' scenario, i.e., when the discovered Higgs state is not the lightest one. We evaluated the ability of various jet clustering techniques, with varying resolution parameters and reconstruction procedures, to resolve such fully hadronic final states. We demonstrated that the efficiency of selecting the hadronic states as well as the ability to reconstruct Higgs masses from them are highly influenced by the choice of the jet clustering algorithm and its settings. We specifically highlighted that the variable-R algorithm [76] was more effective in obtaining the signal sensitivity as well as in reconstructing the light and heavy Higgs mass peaks than those based on a fixed cone radius R [75, 137].

Those results were obtained for slim b-jets with no merging (so we looked at typical four b-jet configurations) [1]. In this chapter, we want to instead study the case of fat b-jets, i.e., when two b-partons emerging from a h decay are not resolved as individual jets but are merged into a fat b-jet containing both. This is most likely to occur when the H state is significantly heavier than the h, $m_H \gg m_h = 125$ in the usual 2HDM in the 'standard hierarchy' scenario. Again, we will assess which of the two types of jet-clustering algorithms, fixed or variable cone size, is better able to extract the signal from the backgrounds and yield the sharpest rendition of the Breit-Wigner mass peaks. For this purpose, we will implement a simplified (MC truth-informed) double b-tagger. It is worthwhile to mention that one can use other boosted jet tagging methods based on the jet substructure technique to further enhance the signal significances, e.g., N-subjettiness variables and their ratios [138], Energy Correlation Functions (ECF) and their ratios [96, 139], or a combination of substructure based observables and cutting edge machine learning techniques [140]. Many experimental studies have been done on the fat jets analysis [141, 142, 143, 144].

This chapter is structured as follows. In the next section, we outline the MC analysis that we will perform (i.e., simulation tools, cutflow, b-tagger, etc.). Following that, we will present our results. Finally, in the last section, we will draw our conclusions.

6.2 Methodology

In this section, we describe the tools and selection strategy to pursue our analysis.

6.2.1 Simulation Details

We investigate a suitable BP in the context of the 2HDM Type-II. We consider the lightest CP-even Higgs state to be the SM-like Higgs boson with $m_h = 125$ GeV and set the heavier CP-even Higgs mass as $m_H = 700$ GeV. The BP has been tested against theoretical and experimental constraints using `2HDMC` [40] interfaced with `HiggsBounds` [113] and `HiggsSignals` [114] and also against flavor constraints using `SuperISO` [115]. Specifically, concerning the latter, the following flavor constraints on meson decay Branching Ratios (BRs) and mixings, all to the 2σ level, are used in our analysis: BR($b \to s\gamma$), BR($B_s \to \mu\mu$), BR($D_s \to \tau\nu$), BR($D_s \to \mu\nu$), BR($B_u \to \tau\nu$), $\frac{\text{BR}(K\to\mu\nu)}{\text{BR}(\pi\to\mu\nu)}$, BR($B \to D_0\tau\nu$) and $\Delta_0(B \to K^*\gamma)$.

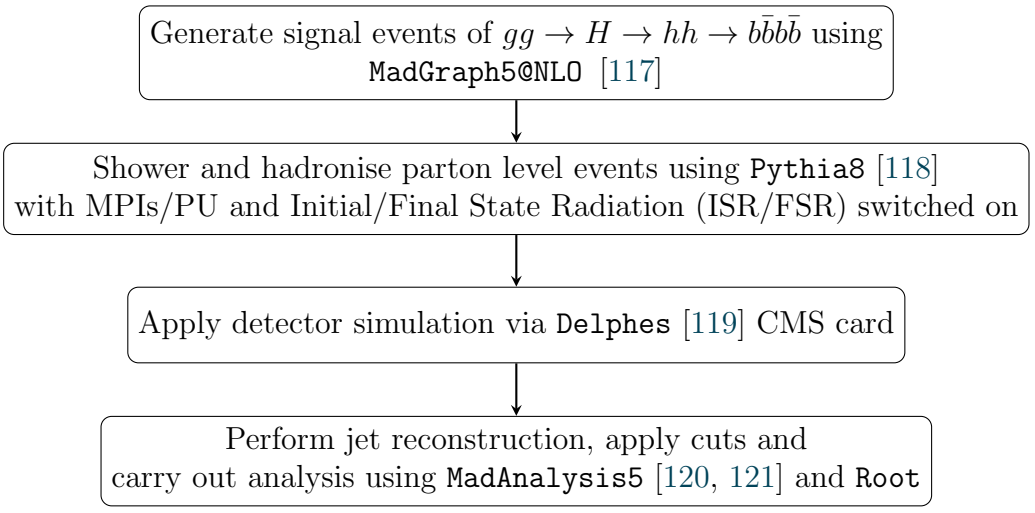


FIGURE 6.1: Description of the procedure used to generate and analyse MC events.

Our study assumes $p-p$ collisions at a $\sqrt{s} = 13$ TeV and integrated luminosities of 140 and 300 fb^{-1}, corresponding to full Run 2 and Run 3 datasets. The production cross sections at LO[1] and decay rates for the sub-processes gg, $q\bar{q} \to H \to hh \to b\bar{b}b\bar{b}$ are presented in Tab. 3.2, alongside the 2HDM Type-II input parameters (see Point1 in Tab. 3.2). In the calculation of the overall cross-section, the renormalisation and factorisation scales were both set to be $H_T/2$, where H_T is the sum of the transverse energy of each parton. The PDF set used was `NNPDF23_lo_as_0130_qed` [145].

In order to carry out a realistic MC analysis, we use the toolbox described in Fig. 6.1 to generate and analyse events.

To generate samples of the leading SM backgrounds, the same toolkit (see Fig. 6.1) is used. The background processes we consider are the following: the QCD $4b$ background, $gg, q\bar{q} \to t\bar{t}$ and $gg, q\bar{q} \to Zb\bar{b}$ [1]. Due to the kinematic differences between the signal process and leading backgrounds, we apply generation-level cuts within `MadGraph5` to improve the selection efficiency at the jet level

$$gg, q\bar{q} \to t\bar{t}: \; p_T^{\text{gen}}(t) > 250 \text{ GeV},$$

$$gg, q\bar{q} \to b\bar{b}b\bar{b}: \; p_T^{\text{gen}}(b) > 100 \text{ GeV},$$

$$gg, q\bar{q} \to Zb\bar{b}: \; p_T^{\text{gen}}(Z) > 250 \text{ GeV}, \; p_T^{\text{gen}}(b) > 200 \text{ GeV}.$$

[1]This is also the perturbative level at which MC events are generated.

This ensures that our signal and background events fall in the same p_T window to perform a sensible signal-to-background analysis later in the study.

6.2.2 Cutflow and b-tagging Implementation

The inclusion of the full sequence of cuts that we have used here requires some justification. Existing b-jet analyses that seek to extract chain decays of Higgs bosons from the background have used restrictive trigger level cuts to ensure the extraction of a fully hadronic signature. As informed by the trigger level cuts in Chapter 5, we place a loose p_T cut of 200 GeV for the boosted regime so that signal selection efficiency can provide an acceptable MC sample for phenomenological analysis. A full description of the cutflow is given in Fig. 6.2.

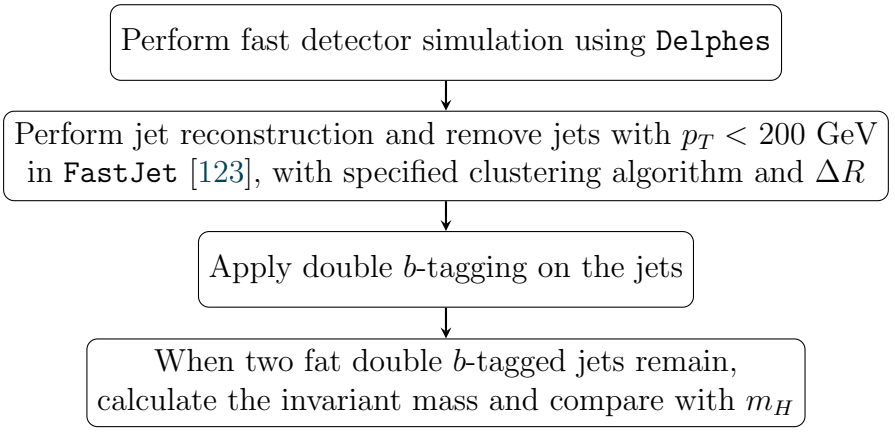


FIGURE 6.2: Description for jet clustering, b-tagging and selection of jets.

We implement a simplified (MC truth-informed) double b-tagger. For events clustered using the anti-k_T algorithm (C/A algorithm) with fixed-R cone size, parton level b-(anti)quarks within angular distance ΔR from jets are searched for, and if there are two b-quarks present within that separation, jets are tagged as double b-tagged fat jets as appropriate. When the variable-R approach is used, the size of the tagging cone is taken as the effective size R_{eff} of the jet.

Additionally, we take into account the finite efficiency of identifying a b-jet as well as the non-zero probability that c-jets, light-flavor, and gluon jets are mistagged as b-jets. We use p_T-dependent tagging efficiencies and mistag rates from a `Delphes` CMS detector card[2]. We have validated that the conclusion remains the same if we employ a modified b-tagging procedure in which the b-partons are replaced with b-hadrons produced after the hadronisation of the b-quarks.

[2]See `https://github.com/delphes/delphes/blob/master/cards/delphes_card_CMS.tcl`.

6.3 Results

In this section, we report our results for both the signal and dominant SM backgrounds, first at the parton level and then at the detector level. The dominant backgrounds, such as the QCD $4b$, $gg, q\bar{q} \to t\bar{t}$ and $gg, q\bar{q} \to Zb\bar{b}$ are taken into account later in the study for the signal-to-background analysis.

6.3.1 Parton Level Analysis

Before proceeding with the detector level analysis, we examine the parton level information of the events in order to fine-tune certain parameters for jet clustering as well as for sensibly using the selected kinematic cuts. In fact, the final state b-partons p_T will inform us which value of ρ to use for the variable-R clustering algorithm.

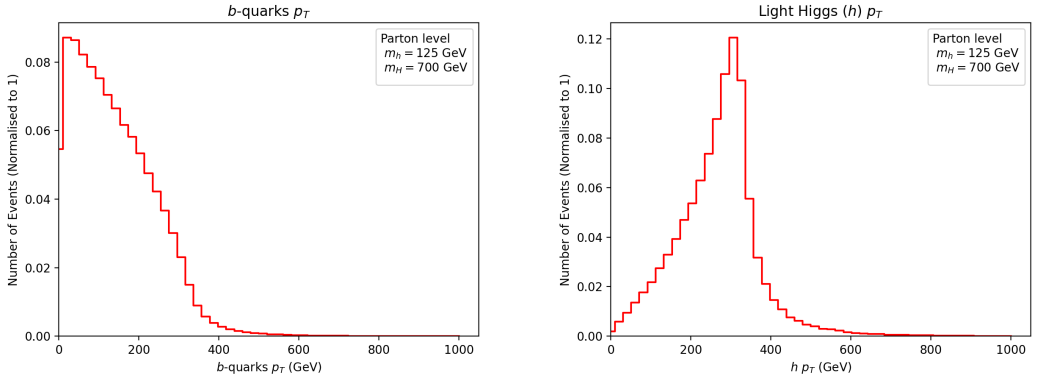


FIGURE 6.3: Left panel: Transverse momenta of the final state b-quarks. Right panel: Transverse momenta of the lights Higgses.

We can see that the final state b-quarks have a wide range of momenta, well into $O(10^2)$ GeV, as shown in Fig. 6.3 (left panel). The value of ρ, the variable-R specific parameter, is typically set to be of the same order of magnitude as the jet p_T. However, looking at the p_T distribution of the b-quarks, we perform a scan for ρ over the region $[100, 500]$ GeV to find an optimal value. Another point to note here is that the light Higgs bosons are significantly boosted, as shown in Fig. 6.3 (right panel). The angular separation in the $\eta - \phi$ plane between the pairs of Higgs bosons and b-quark pairs coming from the same Higgs boson is critically dependent on the p_T of the heavier and the lighter Higgs bosons.

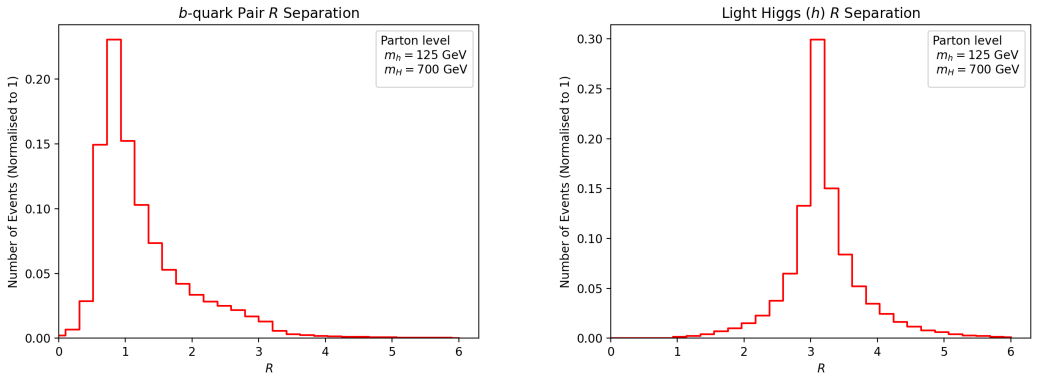


FIGURE 6.4: Left panel: ΔR separation of the $b\bar{b}$ pair from a given Higgs. Right panel: ΔR separation between the two Higgses.

The two light Higgses are generally always back-to-back with angular separation peaking around π, implying that the heavier Higgs is mostly produced at rest, see Fig. 6.4. Despite the fact that the heavier Higgs has negligible p_T due to the mass configuration of this BP, the two SM-like Higgses have a large momentum transfer from the heavy Higgs. The b-quarks originating from the lighter Higgs bosons, in contrast, tend to be closer together, i.e., collimated, which is an artifact of boosting.

As a result, the final jets from these b-partons will be closer together in the detector space. We can exploit this, and instead of trying to lower the values of R in the jet clustering algorithm to 'pick out' and tag all four signal b-jets, we can instead use a deliberately large cone in order to capture two fat (and back-to-back) jets, each containing both b-quarks coming from the same decayed SM-like Higgs boson.

6.3.2 Jet Level Analysis

We can now proceed to analyse the topology at the jet level, guided by the parton level kinematics of the events. We cluster EFlow objects obtained after the fast detector simulation using `Delphes` into wide cone jets with the anti-k_T algorithm [75]. We select those jets that have a $p_T > 200$ GeV before we proceed to tag them, as described in Fig. 6.2[3].

Here, we compare two different methods of jet clustering for these double b-tagged fat jets. To begin, we employ a large fixed cone size $R = 1.0$ to construct two

[3]We have switched on ISR and MPI in `Pythia8` to investigate the results for the two types of algorithms in Section 6.4.2.

(nearly) back-to-back fat jets from each h decay, each of which should reveal the mass of the SM-like Higgs boson[4]. Secondly, we do the same thing but use the variable-R jet clustering algorithm [76]. To obtain the best-reconstructed resonance mass peaks, we optimise the choice of ρ parameter. For variable-R, we use $\rho = 300$ with $R_{\text{min}} = 0.4$ and $R_{\text{max}} = 2.0$. These values were influenced by the p_T scale of the fixed cone b-jets as well as the aforementioned scan of ρ parameter.

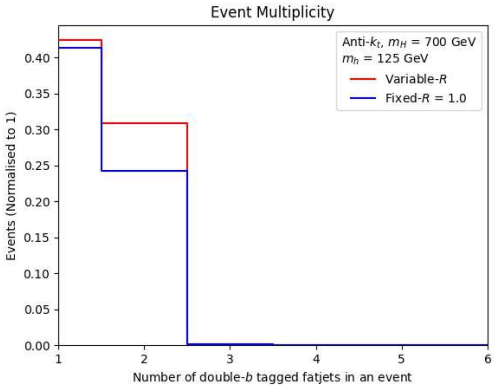


FIGURE 6.5: The double b-tagged fat jets multiplicity distribution for our BP.

We compare the b-jet multiplicity of the signal events for both fixed-R and variable-R algorithms in Fig. 6.5. It is evident from the figure that we get more events with double-b tagged fat jets for variable-R than for fixed-$R = 1.0$. The existence of more events from the signal that contain double-b tagged fat jets allows us to better reconstruct the Higgs resonance peaks in multi-jet mass distributions.

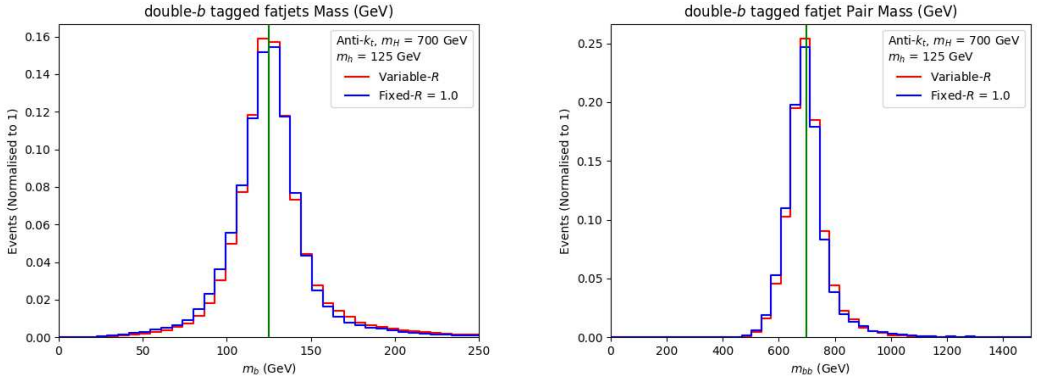


FIGURE 6.6: Left panel: The double b-tagged fat jets invariant mass m_h for our BP. Right panel: The two double b-tagged fat jets invariant mass m_H for our BP.

[4]We did optimise the fixed cone size value, and $R = 1.0$ was found to be the best choice for the reconstruction of mass peaks.

To show the evidence of new physics, we reconstruct the mass of the resonances, notably the light and heavy Higgs bosons. In Fig. 6.6, we present the invariant masses of individual double b-tagged fat jets and the pair of double-b tagged fat jets. We select the average of all double b-tagged jets for the m_h mass resonance. For the m_H resonance, we select events with two double b-tagged jets in order to recover a heavy Higgs peak. The variable-R algorithm mass distributions clearly show that the peaks are closer to the MC truth value of the corresponding Higgs boson masses, namely $m_h = 125$ GeV and $m_H = 700$ GeV.

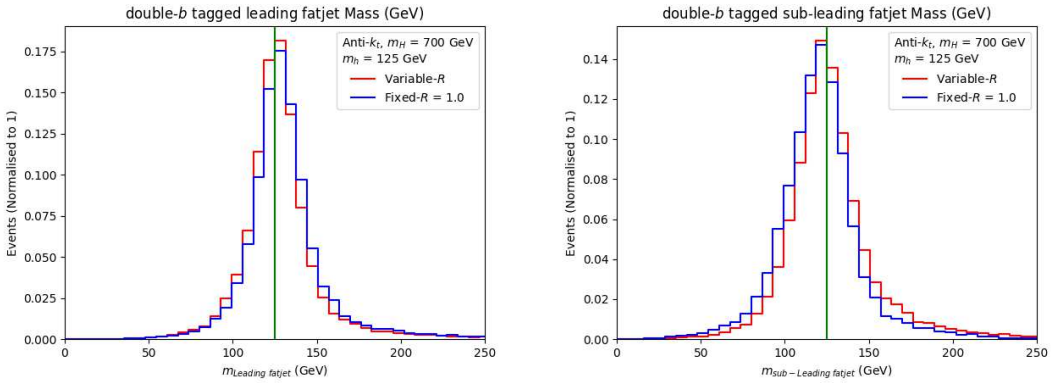


FIGURE 6.7: Left panel: The double b-tagged leading fat jet invariant mass m_h for our BP. Right panel: The double b-tagged sub-leading fat jet invariant mass m_h for our BP.

In Fig. 6.7, we also present mass distributions for the leading and subleading fat jets (double b-tagged). The same behavior can be seen here with the variable-R jet algorithm, with results being more aligned towards the corresponding MC truth value of the light Higgs boson mass. Next, we compare the two jet reconstruction algorithms mentioned in terms of signal-to-background rates.

6.3.3 Signal-to-background Analysis

In this section, we describe the performance of our final cuts employed in extracting the signal from the backgrounds and compute the final significances in the presence of both MPI and PU effects.

Process	Variable-R	$R = 1.0$
	$m_h = 125$ GeV, $m_H = 700$ GeV	$m_h = 125$ GeV, $m_H = 700$ GeV
$pp \to H \to hh \to b\bar{b}b\bar{b}$	147.56	104.874
$pp \to t\bar{t}$	166.633	111.088
$pp \to b\bar{b}b\bar{b}$	592.336	435.139
$pp \to Zbb$	0.067	0.063

TABLE 6.1: Event rates of signal and backgrounds for $\mathcal{L} = 140$ fb^{-1} upon enforcing the initial cuts plus the mass selection criteria of Fig. 6.8 for the two jet reconstruction procedures.

6.3.3.1 Signal-to-background Analysis with MPIs

In order to quantify the performance of the variable-R algorithm against the fixed-R one, we calculate signal-to-background rates and signal significances for the aforementioned two choices of integrated luminosity. To carry out this exercise, we apply the additional selection procedure described in Fig. 6.8.

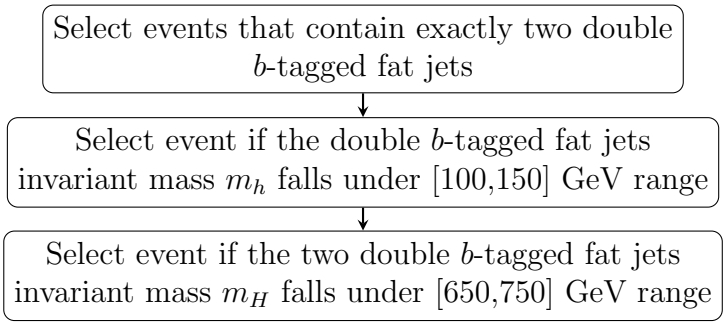


FIGURE 6.8: Additional event selection used to compute the final signal-to-background rates.

The event rates (N) for the various processes are given by:

$$N \ = \text{Cross section } (\sigma) \times \text{Luminosity } (\mathcal{L}). \tag{6.1}$$

Tabs. 6.1 and 6.2 clearly reflect that $pp \to b\bar{b}b\bar{b}$ is the dominant background process followed by $pp \to t\bar{t}$ and $pp \to Zb\bar{b}$. The next step is to calculate the significance ratio (Σ) as a function of Signal (S) and Background (B) rates for two values of integrated luminosities $\mathcal{L} = 140$ fb^{-1} and 300 fb^{-1}, given by:

$$\Sigma = \frac{N(S)}{\sqrt{N(B_{b\bar{b}b\bar{b}}) + N(B_{t\bar{t}}) + N(B_{Zb\bar{b}})}}. \tag{6.2}$$

The significances for both choices of the jet clustering algorithm without and with QCD K-factors are presented in Tab. 6.3. The QCD K-factors describe the ratio

Process	Variable-R	$R = 1.0$
	$m_h = 125$ GeV, $m_H = 700$ GeV	$m_h = 125$ GeV, $m_H = 700$ GeV
$pp \to H \to hh \to bbbb$	316.2	224.73
$pp \to t\bar{t}$	357.071	238.047
$pp \to b\bar{b}b\bar{b}$	1269.292	932.441
$pp \to Zbb$	0.145	0.135

TABLE 6.2: Event rates of signal and backgrounds for $\mathcal{L} = 300$ fb^{-1} upon enforcing the initial cuts plus the mass selection criteria of Fig. 6.8 for the two jet reconstruction procedures.

between the leading and higher order cross sections. We have used $K = 2$ (at NNLO level) for the signal [128, 129], $K = 1.5$ (at NLO level) for $pp \to b\bar{b}b\bar{b}$ [130], $K = 1.4$ (at NLO level) for $pp \to t\bar{t}$ [131] and $K = 1.4$ (at NLO level) for $pp \to Zb\bar{b}$ [132]). It is clear that the variable-R approach is more efficient compared to the fixed-R method. Even if we account for a typical 10% effect of systematic uncertainties in our signal significances calculations, the conclusion remains the same.

In Tab. 6.4, we also discuss the implications of using the trimming technique [82] to mitigate the effect of ISR and MPI and present the significances for both choices of jet clustering algorithms without and with QCD K-factors. We have used default CMS values of $R_{Trim} = 0.2$ and $p_{T_{FracTrim}} = 0.05$, taken from the `Delphes` CMS detector card. It is evident that the variable-R approach is more efficient compared to the fixed-R method and our conclusions still hold even after the jets are groomed (one can always use other grooming techniques such as filtering [83], pruning [84], mass-drop [83], modified mass-drop [86] and soft drop [85], however, this is beyond the scope of this work).

	Variable-R	$R = 1.0$
$\mathcal{L} = 140$ fb^{-1}	5.355	4.487
$\mathcal{L} = 300$ fb^{-1}	7.840	6.568
	Variable-R	$R = 1.0$
$\mathcal{L} = 140$ fb^{-1}	8.810	7.377
$\mathcal{L} = 300$ fb^{-1}	12.897	10.799

TABLE 6.3: Upper panel: Final Σ values calculated upon enforcing the initial cuts plus the mass selection criteria of Fig. 6.8 for the two jet reconstruction procedures. Lower panel: The same in the presence of K- factors.

	Variable-R	$R = 1.0$
$\mathcal{L} = 140$ fb^{-1}	5.753	4.861
$\mathcal{L} = 300$ fb^{-1}	8.421	7.116
	Variable-R	$R = 1.0$
$\mathcal{L} = 140$ fb^{-1}	9.513	8.022
$\mathcal{L} = 300$ fb^{-1}	13.926	11.743

TABLE 6.4: Upper panel: Final Σ values calculated upon enforcing the initial cuts plus the mass selection criteria of Fig. 6.8 for the two jet reconstruction procedures using Trimming grooming techniques. Lower panel: The same in the presence of K-factors.

6.3.3.2 Signal-to-background Analysis with PU

As a last exercise, we want to compare the performance of the two clustering algorithms employed in this study to reconstruct jets with PU. As previously stated, one needs to apply proper detector simulation using `Delphes`, to perform such a study. Specifically, generated events after hadronisation are passed through a `Delphes` CMS PU card[5]. We have used `Pythia8` to generate the PU simulations. Mixing of these PU events with the signal events is then done with $< N_{\rm PU} > = 50$ for each hard scattering. Next, `FastJet` is implemented for both the variable-R and fixed-R algorithms within the same card, to finally output jet information into a `Root file`. Finally, we carry out the analysis using a `Root` macro code and the same cutflow described in Section 6.3.2 in the presence of the additional selection procedure mentioned in Fig. 6.8.

We again calculate the signal-to-background rates, and consequent significances, in the presence of the usual luminosities, specifically to compare the performance of the variable-R jet clustering algorithm against the fixed-R in extracting the signal from the dominant backgrounds. Tabs. 6.5 and 6.6 shows the event rates (N) (described by Eq.(6.1)) for the various processes.

Tab. 6.7 provides the final significance rates (as per Eq.(6.2)) with and without K-factors. Even with PU effects, it is evident that the variable-R approach is much better compared to the fixed-R method.

[5]See `https://github.com/recotoolsbenchmarks/DelphesNtuplizer/blob/master/cards/CMS_PhaseII_200PU_Snowmass2021_v0.tcl#L1039-L1067`.

Process	Variable-R	$R = 1.0$
	$m_h = 125$ GeV, $m_H = 700$ GeV	$m_h = 125$ GeV, $m_H = 700$ GeV
$pp \to H \to hh \to b\bar{b}b\bar{b}$	76.655	55.239
$pp \to t\bar{t}$	111.088	166.633
$pp \to b\bar{b}b\bar{b}$	423.748	282.498
$pp \to Zb\bar{b}$	0.0180	0.0270

TABLE 6.5: Event rates of signal and backgrounds with PU for $\mathcal{L} = 140$ fb^{-1} upon enforcing the initial cuts plus the mass selection criteria of Fig. 6.8 for the two jet reconstruction procedures.

Process	Variable-R	$R = 1.0$
	$m_h = 125$ GeV, $m_H = 700$ GeV	$m_h = 125$ GeV, $m_H = 700$ GeV
$pp \to H \to hh \to b\bar{b}b\bar{b}$	164.260	118.371
$pp \to t\bar{t}$	238.047	357.071
$pp \to b\bar{b}b\bar{b}$	908.032	605.354
$pp \to Zb\bar{b}$	0.038	0.0580

TABLE 6.6: Event rates of signal and backgrounds with PU for $\mathcal{L} = 300$ fb^{-1} upon enforcing the initial cuts plus the mass selection criteria of Fig. 6.8 for the two jet reconstruction procedures.

	Variable-R	$R = 1.0$
$\mathcal{L} = 140$ fb^{-1}	3.314	2.606
$\mathcal{L} = 300$ fb^{-1}	4.851	3.815
	Variable-R	$R = 1.0$
$\mathcal{L} = 140$ fb^{-1}	5.450	4.309
$\mathcal{L} = 300$ fb^{-1}	7.978	6.309

TABLE 6.7: Upper panel: Final Σ values calculated upon enforcing the initial cuts plus the mass selection criteria of Fig. 6.8 for the two jet reconstruction procedures with PU. Lower panel: The same in the presence of K-factors.

6.4 Summary and Conclusions

In this chapter, we investigated the performance of two types of jet clustering algorithms at the LHC in accessing BSM signals induced by the cascade decays of a heavy Higgs boson H (with a mass of 700 GeV) into a pair of SM-like Higgs states, hh. Given the mass difference between the two Higgs masses involved, the lighter Higgs bosons are obtained with a substantial boost, causing their decay products, notably a pair of b-quarks in our study, to become highly collimated. As a result, we reconstruct these events into two fat jets and perform a double b-tagging on them. For illustration purposes, a 2HDM Type-II setup was assumed, by adopting a BP over its parameter space that was fully compliant with both theoretical and experimental constraints.

The two types of jet clustering algorithms are a variable-R one (where the cone size is not fixed but rather adapts to the resonant kinematics of the signal) and a more standard one, with a fixed cone size ($R = 1.0$). These are used twice to reconstruct the mass of the lighter (SM-like) Higgs boson. Furthermore, we pick events with a pair of such double b-tagged fat jets, where total invariant mass reproduces the heavy Higgs mass. We further discover that for a cut-based signal-to-background analysis, the variable-R method not only provides better-reconstructed peaks of both Higgs boson masses than the traditional algorithm but also improves the signal-to-background rates, resulting in higher signal significances at the LHC (altogether leading to potential discovery at both Run 2 and 3 of the LHC). Thus, we advocate the use of the former in establishing $pp \to H \to hh \to b\bar{b}b\bar{b}$ events in boosted topologies, in line with similar results previously obtained for the case of the same channel and different mass spectra yielding four slim b-jets. Finally, it is worth noting that we have used the anti-k_T algorithm as representative of the fixed cone size kind throughout but the results are the same for the C/A jet clustering algorithm.

Chapter 7

Exploring SM-like Higgs Boson Production in Association with Single-Top at the LHC Within a 2HDM

This chapter is based on the work released in [3]. This paper was co-authored by Ciara Byers, Stefano Moretti and Emmanuel Olaiya.

7.1 Introduction

Following the discovery of a Higgs boson in 2012, the couplings to the weak bosons and t, b, c, τ, ν fermions have proven difficult to measure directly. In fact, the ability to access the sign of the Higgs boson-bottom (anti)quark coupling is conspicuously missing. We can see this because the Yukawa-type SM Higgs to Fermion couplings display a clear hierarchy in their strength; the $ht\bar{t}$ coupling is substantially larger than the $hb\bar{b}$. Thus, in the aforementioned decay processes the role played by the second coupling is negligible in comparison to that of the former (This is also true for the production process $gg \to h$)[1].

However, if we consider some 2HDM constructs and the resulting interactions, our access changes, and the $hb\bar{b}$ coupling can have the opposite sign to that of the SM.

[1] A fairly up-to-date review of the current LHC status in establishing the nature of the SM-like Higgs boson can be found in Ref. [9].

	$\sigma(bq)$ (pb)	$\sigma(bg)$ (pb)	$\sigma(qq)$ (pb)	σ(total) (pb)
SM	0.036	0.011	0.0023	0.049[148]

TABLE 7.1: The tree-level cross-sections for the bq, bg and qq sub-processes of SM-like Higgs boson production in association with a single top (anti)quark at the LHC with 13.6 TeV of Centre-of-Mass (CM) energy. (These values have been calculated by `MadGraph-3.1.0` [151] for the default SM implementation that comes with the package.

On the one hand, this would not affect the current measurements; but it could result in a significant boost to the cross-section of some alternative ones. In light of this, it is crucial to test other h boson production channels in addition to the conventional ones that have already been established at the LHC[2]: gluon-gluon fusion ($gg \to h$), vector-boson fusion ($qq \to q'q'h$) and associated production with weak gauge bosons ($q\bar{q}' \to Zh(W^{\pm}h)$) or top (anti)quark pairs ($q\bar{q}, gg \to t\bar{t}h$) (see Ref. [146] for a review).

Specifically, we concentrate on SM-like Higgs boson production in association with a single top (anti)quark. The process is mediated by the following sub-processes: $bq \to tq'h$ + c.c. (hereafter, bq), $bg \to tW^{-}h$ + c.c. (hereafter, bg) and $q\bar{q}' \to t\bar{b}h$ + c.c. (hereafter, qq), see Figs. 7.1–7.2 [3]. In the SM, the corresponding cross-sections at the LHC are listed (in decreasing order of importance) in Tab. 7.1. Altogether, their production cross-section is smaller than, but of the same order as, that of $q\bar{q}, gg \to t\bar{t}h$, so some sensitivity presently exists to this additional h production mechanism. Furthermore, in all such analyses, the bq, bg, and qq channels are treated inclusively [148, 149, 150].

Unfortunately, recent search results have only been able to exclude cross-sections for SM-like Higgs boson production in association with a single top (anti)quark for values higher than the SM predictions. The primary cause of this is the existence of cancellations between the topologies in Figs. 7.1–7.2. This, in turn, is the result of the hWW coupling and $hb\bar{b}/ht\bar{t}$ couplings simultaneously entering at amplitude level and being able to interfere. As a result, we come to the conclusion that a mechanism such as this can be used as a privileged probe of some BSM dynamics. It is evident that any alteration of these three Higgs couplings could result in a larger cross-section than predicted by the SM, which would be detectable at the HL-LHC earlier than one might otherwise expect with the SM.

[2]Hereafter, "q'" refers to a light quark (d, u, s or c).
[3]We work in a five-flavour scheme [147].

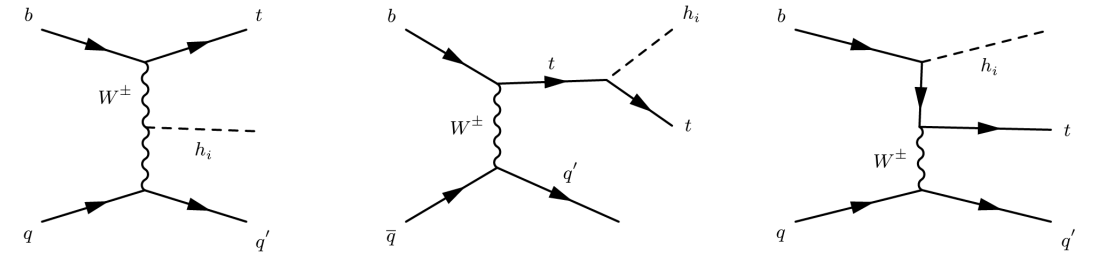


FIGURE 7.1: Feynman diagrams for the bq sub-process, assuming time flowing rightwards, wherein we ignore the contribution of a charged Higgs boson ($H^\pm$), which we set as heavy enough to give an eligible correction. Notice that the same diagrams appear in the qq sub-process when time is flowing upwards.

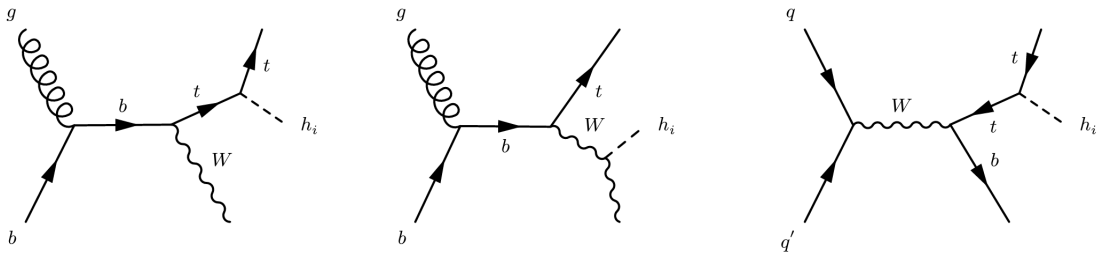


FIGURE 7.2: Feynman diagrams for the bg sub-process, assuming time flowing rightwards, wherein we ignore the contribution of a charged Higgs boson ($H^\pm$), which we set to be heavy enough so as to give a negligible correction.

Furthermore, it is possible that kinematical distributions in some BSM scenarios are also different from those produced in the SM case since the Feynman diagrams carrying such couplings are topologically distinct (i.e. they would produce different kinematics in the final state).

With this in mind, in this chapter, we investigate the possibility of such a phenomenology being realised in the simplest extension of the SM Higgs sector using a second doublet field (akin to the one in the SM), i.e. the one embedded in a generic 2HDM [33]. In particular, we focus on the $h \to b\bar{b}$ decay channel, as it strikes a good balance between the expectation of a substantial number of events to be reconstructed and a manageable background when combined with the other signal final state particles. This is one of the dominant Higgs decay channels in the SM, yet it is poorly measured due to the large background that arises when searching via the standard four production processes[4].

[4]Indeed, current estimates point to the extraction of a signal for SM-like Higgs boson production in association with a single top (anti)quark in the SM being possible with a minimum of 1500 fb^{-1} of integrated luminosity in the $h \to b\bar{b}$ channel[152, 153], which is only achievable at the HL-LHC.

We will show that it is possible, particularly for the bg sub-process, to significantly enhance the cross-section in the 2HDM Type-II. Meanwhile, the bq and qq rates remain comparable to their SM counterparts. Remarkably, this occurs exactly when the sign of the $hb\bar{b}$ coupling changes with respect to the SM. In fact, given that the kinematics arising from the diagram carrying the $hb\bar{b}$ coupling are different from those involving the $ht\bar{t}$ and hW^+W^-, as has been intimated, we will also be able to demonstrate that there are notable differences between the SM and 2HDM Type-II cases in a variety of differential distributions. Altogether, this creates the ideal environment for vigorously pursuing this additional h production channel experimentally, with a dual goal in mind. For starters, we have the opportunity to demonstrate the existence of a BSM Higgs sector. Second, there is a chance to show that it should be possible to separate the underlying structure of this model, at least for the 2HDM Type-II.

The structure of this chapter is as follows: in the next section, we will introduce 2HDM and the wrong-sign (Yukawa) coupling scenario, then describe `Magellan` [154] and how we have used it to perform parameter space scans, followed by our detector-level analysis. In the last section, we present our conclusions.

7.2 The 2HDM and the Wrong-Sign (Yukawa) Coupling Scenario

Following the detailed discussion of the 2HDM in Chapter 3, we know that under the discrete $\mathbb{Z}_2$-symmetry, fermions must also have a definite charge other than the (pseudo)scalar fields. The various assignments of the $\mathbb{Z}_2$-charge in the fermion sector result in four different types of the 2HDM. Exclusively focusing on 2HDM Type-I and -II only, the couplings of the neutral Higgses to fermions, normalised to the corresponding SM value (m_f/v, henceforth, denoted by κ_{hqq} or simply κ_{qq} for the case of the SM-like Higgs state coupling to a quark q, where $q = d, u$), can be found in Tab. 3.1. There are two limiting scenarios for Type-II cases which give rise to two distinct regions in the $(\cos(\beta-\alpha), \tan\beta)$ parameter plane [155]. These can be better understood by looking at how κ_{hqq} behaves as a function of the angles α and β. Taking the limits $\beta - \alpha \to \frac{\pi}{2}$ and $\beta + \alpha \to \frac{\pi}{2}$, the couplings

become (recall Tab. 3.1)

$$\begin{aligned}
\kappa_{hdd} = -\frac{\sin\alpha}{\cos\beta} &= \sin(\beta-\alpha) - \cos(\beta-\alpha)\tan\beta \xrightarrow[\beta-\alpha=\frac{\pi}{2}]{} 1 \text{ (middle-region)}, \\
&= -\sin(\beta+\alpha) + \cos(\beta+\alpha)\tan\beta \xrightarrow[\beta+\alpha=\frac{\pi}{2}]{} -1 \text{ (right-arm)}, \\
\kappa_{huu} = \frac{\cos\alpha}{\sin\beta} &= \sin(\beta-\alpha) + \cos(\beta-\alpha)\cot\beta \xrightarrow[\beta-\alpha=\frac{\pi}{2}]{} 1 \text{ (middle-region)}, \\
&= \sin(\beta+\alpha) + \cos(\beta+\alpha)\cot\beta \xrightarrow[\beta+\alpha=\frac{\pi}{2}]{} 1 \text{ (right-arm)}.
\end{aligned} \tag{7.1}$$

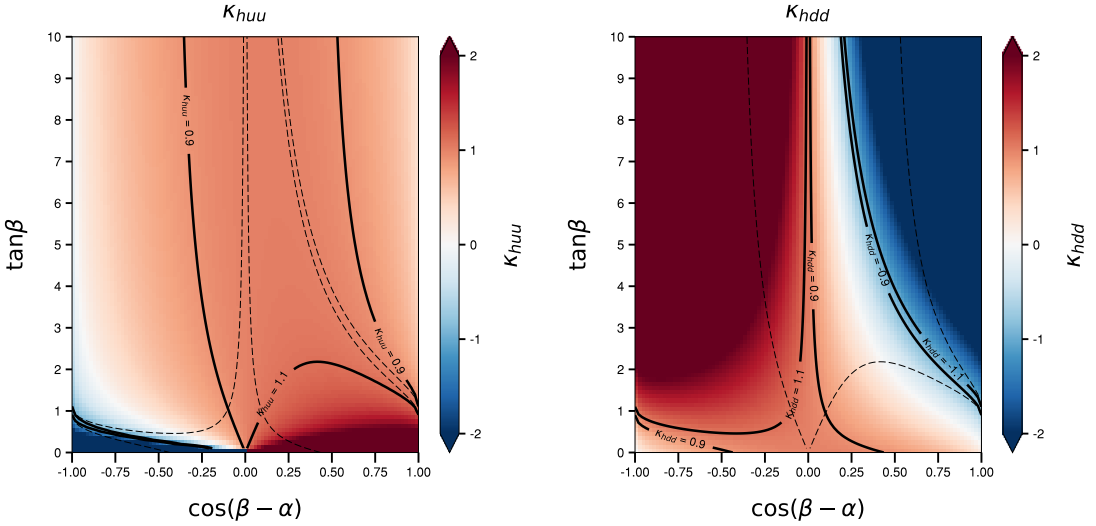


FIGURE 7.3: Light CP-even Higgs couplings to the up-type (left) and down-type (right) quarks, normalised to the corresponding SM value, in the $(\cos(\beta-\alpha), \tan\beta)$ plane. Plots are taken directly from [154].

Fig. 7.3 shows the dependence of κ_{hdd} and κ_{huu} on $\cos(\beta-\alpha)$ and $\tan(\beta)$. The $\beta-\alpha \to \frac{\pi}{2}$ case corresponds to the "middle-region" and the SM-limit of the theory. The right-hand side plot of Fig. 7.3, is identified by the contour region where $0.9 \leq \kappa_{hdd} \leq 1.1$, assuming a 10% difference from the SM values. The $\beta+\alpha \to \frac{\pi}{2}$ case corresponds to the "right-arm", where the coupling of the SM-like Higgs h and the down-type quarks has an opposite sign relative to the SM value. This is known as the wrong-sign (Yukawa) coupling scenario. The region represented by the narrow arm where the coupling is negative is illustrated on the right-hand side plot of Fig. 7.3. Again, this has a 10% displacement from the SM value: $-1.1 \leq \kappa_{hdd} \leq -0.9$. As shown in the left-hand plot of Fig. 7.3, both the alignment and the wrong-sign regions are well within the O(10%) deviation from their corresponding SM values which are allowed for the coupling of the SM-like Higgs to the up-type quarks, κ_{huu}.

The additional four states of a generic 2HDM [33, 34] give rise to a range of observables through which various theoretical models might, in principle, be tested. Therefore, it is worthwhile to investigate in depth the scope of the LHC to discover new Higgs bosons as described within 2HDMs. There has been no experimental evidence for a 2HDM, but a vast array of literature exists on phenomenological analyses that set bounds on the parameter space of such models. In the past two decades, there has been considerable development and implementation of global fits, which collect the data coming from different experiments and perform rigorous statistical analyses to extract limits on BSM theories. The package `Gfitter` [156] was a pioneer in releasing a global EW fit to constrain the new physics predicted by a variety of models, including the 2HDM. Other such toolkits have been published in the literature, with their main focus centered on SUSY and its variants (the 2HDM Type-II in the decoupling limit being one such variant).

The standard techniques used by global fitting packages integrate relevant experimental data and theoretical arguments that can confine the parameter space of the new physics model. These constraints can be divided into three main sources: measurements of the discovered 125 GeV Higgs boson properties (i.e., production and decay signal strengths), searches for the additional Higgs bosons that come within the model, both direct and indirect, and, finally, theory considerations based on perturbativity, unitarity, triviality, and vacuum stability. The likelihood function is then used to indicate the plausibilities of different parameter values for the given samples of data in statistical analysis. The 2HDM parameter space is 6-dimensional (after enforcing m_h reconstruction), so the conventional way of extracting bounds is to project the full parameter space onto 2-dimensional planes determined by any two model parameters. Typically, the statistical procedure is used to maximise the (log) likelihood of the four remaining parameters.

7.3 Parameter Space

Here, we will describe how theoretical and experimental constraints were applied to the 2HDM Type-I and -II parameter spaces using `Magellan`, as well as how cross-sections were computed over them.

7.3.1 Tools

As previously shown, the 2HDM parameter space is made up of six input parameters, namely.

$$m_H,\ m_{H^\pm},\ m_A,\ \cos(\beta-\alpha),\ \tan\beta \text{ and } Z_7$$

It is possible that the allowed regions of this parameter space are not all interconnected, but rather contain a sequence of pockets of valid points. In such an instance, a straightforward grid-type scan was found to be exceedingly inefficient and was abandoned early on. To improve the efficiency of our parameter space scans we used the software package `Magellan` integrated with tools such as (`HiggsBounds` [157], `HiggsSignals` [114, 158, 159], `2HDMC` [40] and `T3PS` [160]), which allow users to generate points in the allowed parameter space of a given 2HDM. The points are found more efficiently using `T3PS`, a Markov Chain Monte Carlo (MCMC) generator [160], with packages `HiggsBounds`, `HiggsSignals` and `2HDMC` contributing to the MCMC's likelihood and checking whether the points are valid before being outputted for phenomenological use. Following that, `Magellan` calls `MadGraph5` [117] directly, which computes the cross-sections for the three contributing sub-processes individually. This was done for both the Type-I and -II realisations of the 2HDM. We used the model file "THDM_type1_UFO", created using `FeynRules` [161, 162] for Type-I, while for Type-II, we used the easily available "2HDMtII_NLO" [163], both of which we used at LO. Finally, we have used a default PDF, the NNPDF 2.3 one [164], integrated within `MadGraph5`[5].

7.3.2 Constraints

From Fig. 7.4, we can clearly see that the constraints for the Type-II case are much tighter than those for Type-I, thus there is a larger parameter space to be scanned for potentially high cross-sections. We see a very similar picture when we look at equivalent plots from CMS in Fig. 7.5. (Our results based on `Magellan` are very similar.)

It has been demonstrated that, in order to find a wrong-sign solution in the 2HDM, $\sin(\beta-\alpha) > 0$ [167], as the available parameter space for negative values has been essentially ruled out. Moreover, in Ref. [168], a lower bound for the charged Higgs boson mass in the Type-II was found to be $m_{H^\pm} > 580$ GeV rendering the

[5]We are interested here in relative effects between the 2HDM and the SM in our three processes of reference, for which QCD corrections are essentially the same.

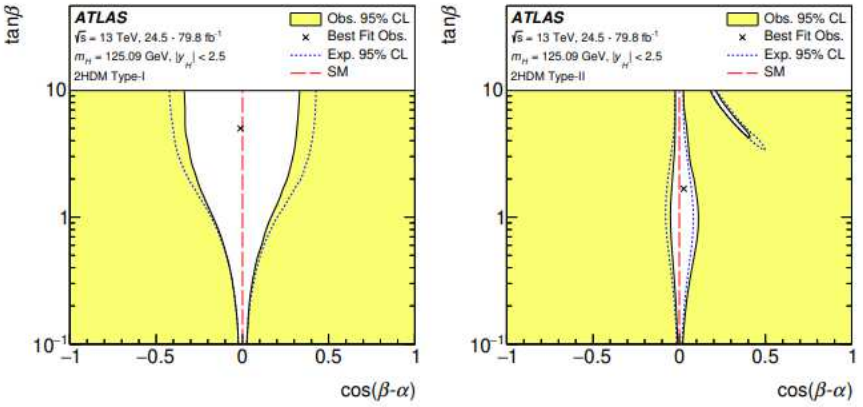


FIGURE 7.4: Allowed regions for the $cos(\beta - \alpha)$ and $tan\beta$ parameters in the 2HDM models Type-I and -II, on the left and right, respectively, for observations made by ATLAS. These are obtained assuming that the 125 GeV boson is the light, CP-even Higgs boson, h, of the 2HDM. Constraints are seen to be tighter on Type-II than on Type-I. Plots are taken directly from [165].

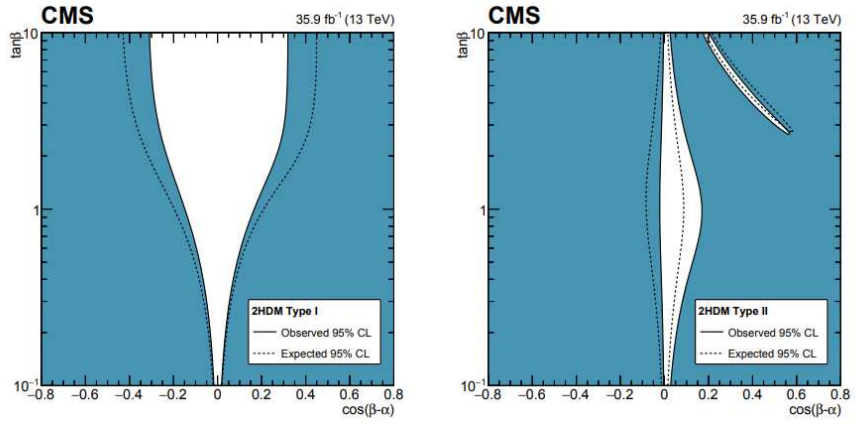


FIGURE 7.5: Allowed regions for the $cos(\beta - \alpha)$ and $tan\beta$ parameters in the 2HDM models Type-I and -II, on the left and right, respectively, for observations made by CMS. These are obtained assuming that the 125 GeV boson is the light, CP-even Higgs boson, h, of the 2HDM. Constraints are seen to be tighter on Type-II than on Type-I Plots are taken directly from [166].

cross-section contributions of $H^{\pm}$ propagators (replacing the $W^{\pm}$ ones in Fig. 7.1) negligible, so that we have completely ignored them in the calculation. Although such a severe bound on $m_{H^{\pm}}$ does not strictly apply to the 2HDM Type-I, for consistency, we have omitted the related Feynman diagrams in the corresponding cross-section calculations (by making the $H^{\pm}$ state heavy enough). Finally in Ref. [169], a general limit on 2HDMs with a (softly-broken) Z_2 symmetry is given as $\tan\beta \geq 1$, which is the constraint used here on the Type-I case. For the Type-II

one, there is a stronger restriction on $\tan\beta$ from the mass of the charged Higgs boson itself, as was shown in Ref. [168] (c.f. Fig. 4 therein). Herein, the lower limit for the Type-II scenario was then chosen as $\tan\beta \geq 5$.

In addition to these measures, the tool `HiggsTools` [170] was released during our research and this was used to check our points against the most recent LHC data. A large amount of data was excluded through this, showing how much progress the LHC has made during run 2, but there was still interesting data left for us to investigate.

7.3.3 Cross-sections at the LHC

The initial output by `MadGraph` for representative points passing all constraints, split across the three sub-processes, namely, bq, bg, and qq are shown in Fig. 7.6. The underlying horizontal lines in Fig. 7.6 represent the SM cross-sections for these channels.

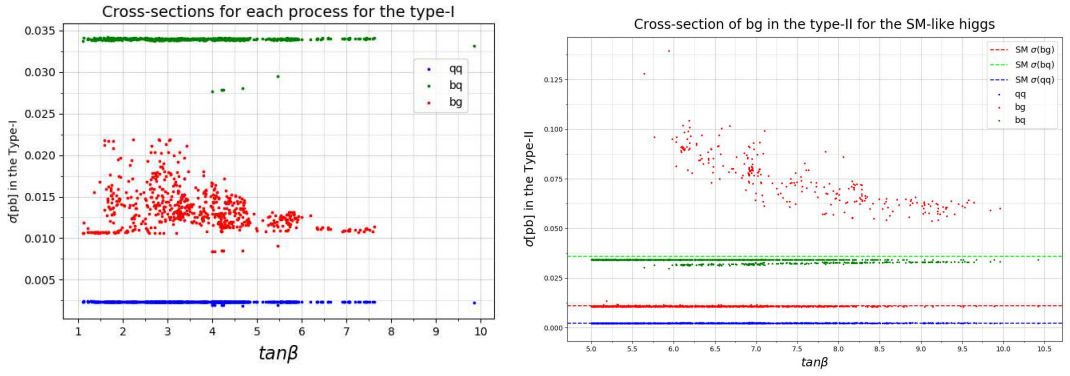


FIGURE 7.6: Cross-sections of points obtained in our scans over the parameter space for the 2HDM Type-I (left) and -II (right) plotted against the value of $\tan\beta$. (Note that these two plots are not to the same scale as the highest cross-sections in Type-II are considerably larger than in Type-I.)

The Type-I points appear to behave very much like the SM predictions, with some points even having a lower cross-section. While there is a slight increase in the bg process, this does not appear to show any significant excess that would distinguish it from the SM at the LHC. This follows from the fact that the wrong-sign solution does not exist in Type-I as the Yukawa couplings for the up and down type quarks. Indeed, we do not find any points with wrong-sign solutions, as expected; this helps in demonstrating that `Magellan` maps out the parameter space effectively and does not allow invalid points through. However, when we

look at Type-II results, we notice that they look very different. While the bq and qq sub-processes for the Type-II behave similarly to the SM, with similar cross-sections and the expected hierarchy between themselves, the bg sub-process does not follow the same pattern. Instead, we notice that the latter becomes dominant over the SM-leading process (bq): in some places, by a very large margin. The highest point in the Type-II distribution has a cross-section for bg that is over 4 times the size of the bq one, allowing us the possibility of quickly extracting the bg channel at the LHC as well as ascribing it to the 2HDM Type-II hypothesis.

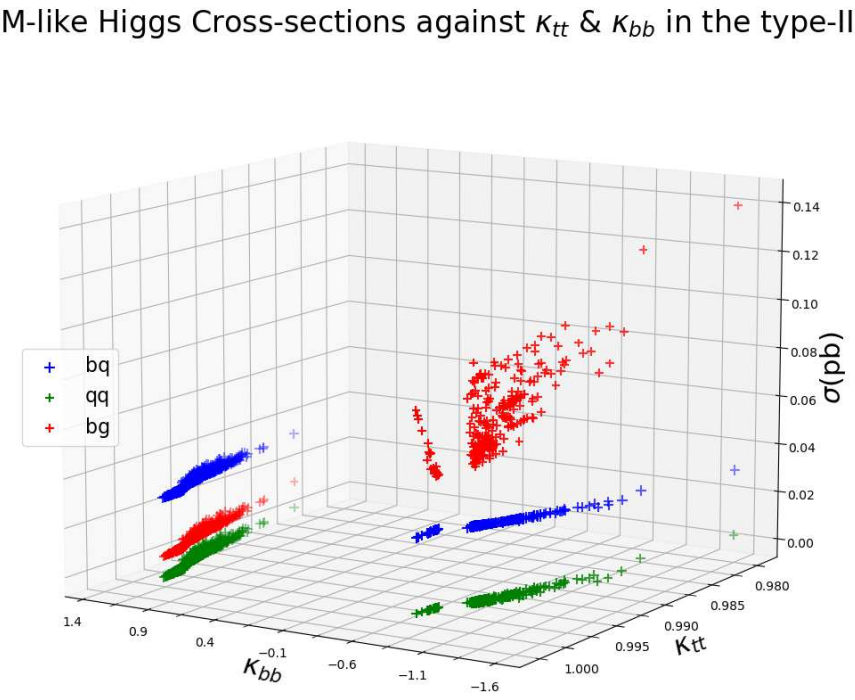


FIGURE 7.7: Cross-sections of points obtained in our described scans over the parameter space for the 2HDM Type-II plotted against κ_{bb} and κ_{tt}. Note that the κ_{bb},κ_{tt} planes are tilted to provide a better view of the points in the 3D space, in particular, the highest point of the plot has a cross-section of ≈ 0.14pb. We see a distinctive spread of high cross-section points for the bg process in the Type-II in the wrong-sign region. The magnitude of these increases with decreasing κ_{tt} and κ_{bb}.

In Fig. 7.7, we plot the cross-sections versus the (rescaled) Yukawas entering the three sub-processes. We can see that in the bg process, both wrong-sign and alignment points offer variation in cross-section. However, only in the wrong-sign region do we witness a large increase in the process, which clearly increases as the value of κ_{bb}. For both the bq and qq sub-processes in Type-II, the two regions of parameter space are similar in terms of (mild) variations in cross-section with respect to the SM, but there is no notable boost as seen in the bg process.

Surprisingly, there is a definite gap in the wrong-sign region, around $\kappa_{bb} = -1$. It is obvious that the more we deviate from the SM the greater the cross-section becomes (i.e. with decreasing κ_{tt} and κ_{bb}).

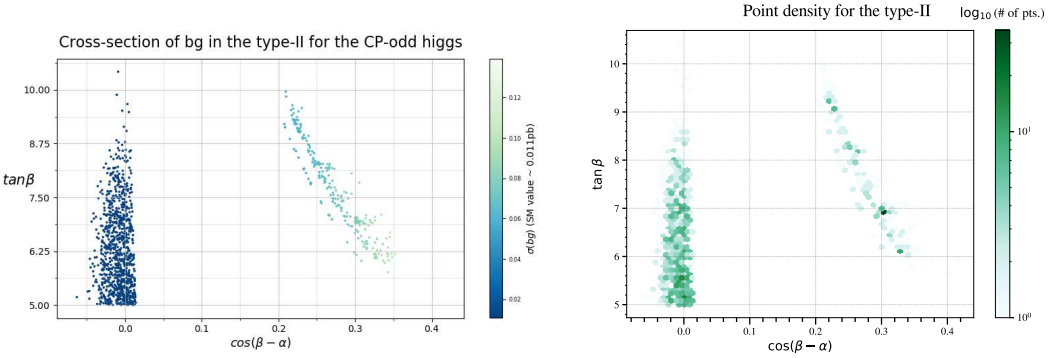


FIGURE 7.8: Cross-sections (left) and $\log_{10}$ of the number of points (right) obtained in the previously described scan of the Type-II parameter space for the bg sub-process mapped over the $(\cos(\beta - \alpha), \tan\beta)$ plane. (Recall that the SM cross-section is 0.011 pb.)

We investigate the cross-section for the bg sub-process across the $(\cos(\beta - \alpha), \tan\beta)$ plane in Fig. 7.8. The colour of a point in the left plot shows how large the cross-section is, with the corresponding colour scale on the right. It is critical to remember that the SM value is ≈ 0.011 pb, which indicates that only the very darkest points correspond to this value. Due to the cross-sections being around the same value as the SM, the central region is solidly plotted in the dark blue colour. The highest points are seen in what remains of the right arm, which has a distinctly different hue. Moving on to the right plot, the colour represents the number of points that have been binned into each coloured hexagon. The colour scale is logarithmic and shown on the right-hand side of the plot. We can detect that relatively low values of $\tan\beta$ are favored, in particular, between 5 and 10. Despite how much of the right arm has been removed by recent LHC data, we detect several dark spots indicating regions where a large number of points have been generated. This is significant as using MCMC analysis, `Magellan` signifies that the density of points in a given region is directly proportional to the likelihood of those points.

In Fig .7.9, we project the points over the plane of κ_{tt} and κ_{bb}. We use a colour gradient to represent the cross-section of each point once again. κ_{tt} is SM-like throughout, constrained to a very small region of $0.975 \leq \kappa_{tt} \leq 1.0025$, the same cannot be said of κ_{bb}, for which the region $-1.6 \leq \kappa_{bb} \leq 1.25$ is still clearly

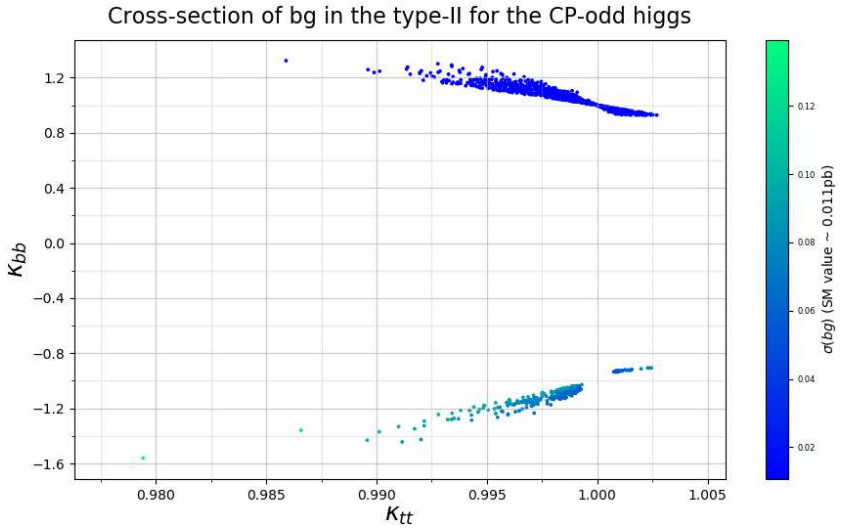


FIGURE 7.9: Cross-section of points obtained in our described scan of Type-II for the bg sub-processes mapped over the (κ_{tt}, κ_{bb}) plane. (Recall that the SM cross-section is 0.011 pb.)

accessible. As soon as we deviate from the SM, we have the highest cross-section points corresponding to the lowest value found for κ_{bb}.

We examine the (m_A, m_H) plane of the Type-II in Fig. 7.10. The alignment points are shown in the bottom-left frame, and their bg cross-sections are mainly of similar size to the SM expectation. While these cross-sections show some variance, it is only about 10% of the expected value, which is far too low to be interesting. Our points are largely clustered in a region of high mass for both particles, with a few points with a high mass for the A state and medium-to-high mass for the H state. The plot points in the bottom-right corner are all from the wrong-sign limit, whereas the top frame combines the two regions. There is greater variation overall in the size of the wrong-sign cross-sections, although they are substantially much larger than those from the alignment region. Notice that neither the H nor the A state enters the cross-sections, we present this figure for the sole purpose of completing the full mapping of the phenomenologically interesting region of (Type-II) parameter space for the bg channel. (As for the $H^{\pm}$ state, we simply reiterate here that its mass is constrained to be above 580 GeV and has no direct implications for the cross-section of the bg process in the two-parameter space regions of reference.)

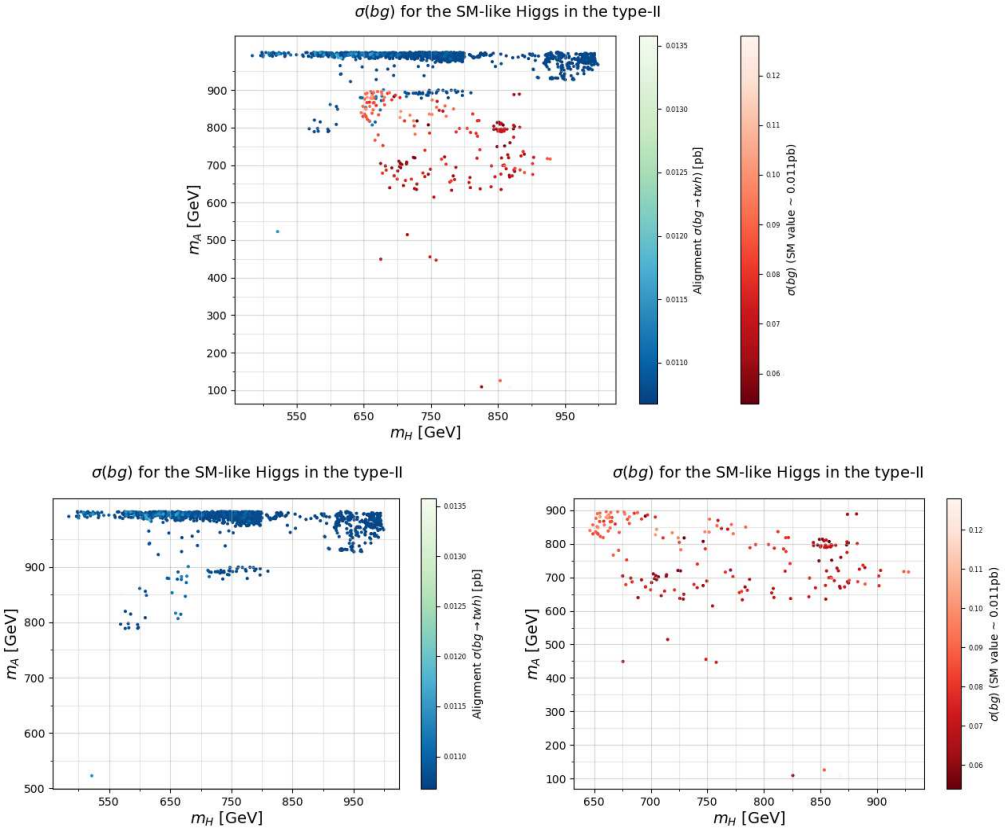


FIGURE 7.10: Cross-section of points obtained in our described scan of Type-II for the bg sub-processes mapped to the (m_H, m_A) plane. The top plot is split into two colour palettes, the blue-green ones are the alignment points while the red-orange ones are wrong-sign points. Both sets are coloured in a gradient shown in their respective colour bars. The gradient indicates the size of the cross-section. The lower plots show the two solutions separately in their own plots. These are coloured according to the size of the bg cross-section, as indicated by the associated colour bars.

7.4 Analysis

The goal of this section is to perform an analysis to assess the cross-sections mentioned in the last section. We investigate the possibility of detecting a 2HDM Type-II cross-section at the HL-LHC for the production of h in association with a single-top and a $W^{\pm}$ via the bg sub-process for the 'wrong-sign solution' of the bottom (anti)quark Yukawa coupling. For this, we conduct a realistic MC analysis in which $h \to b\bar{b}$ decay channel is used to prove that this phenomenology would clearly be visible at the detector level. We specifically test whether the signal emerging from the 2HDM Type-II scenario presents any differences in differential distributions for the final state particles (e.g. invariant and/or transverse masses, transverse momenta, etc.) when compared to the SM case. We approach this by

first considering the main backgrounds that will be applicable. We then generate MC events and implement a dedicated cutflow on these events, followed by the signal-to-background significance calculation for both models in order to perform the aforementioned comparison.

The plan of this section is as follows: we describe the MC analysis we performed (i.e. simulation tools, cutflow, etc.), after which we present our parton and hadron level results then finally, in the last sub-section, we evaluate signal-to-background significances to compare both models.

7.4.1 Methodology

In this section, we provide the simulation details and cutflow information used in our analysis.

7.4.1.1 Simulation Details

We generate samples of events with $\sqrt{s} = 13.6$ TeV as the LHC energy. Our study uses an integrated luminosity of 3000 fb^{-1}, which is expected to be attainable at the HL-LHC. First, we generate events for the bg sub-processes SM implementation. Next, we analyse a sample BP in the 2HDM Type-II framework where the light CP-even Higgs boson is fixed as the SM-like Higgs boson with $m_h = 125$ GeV.

Our research is focused on events where $h \to b\bar{b}$ and the top (anti)quark decays into leptons plus b-jet while the (primary) $W^{\pm}$ boson in the bg sub-process decays leptonically. Tab. 3.2 shows the production cross-section at the LO for the 2HDM Type-II BP, as well as the input parameters of the BP for the full decay chain of the process (see Point3 in Tab. 3.2). The corresponding SM cross-section value is 0.000187 pb. We have used the `NNPDF23_lo_as_0130_qed` [164] set to model the PDF (with default settings).

Fig. 7.11 illustrate the toolbox used to generate and analyse signal events to carry out our realistic MC simulation. The same procedure was also used to generate the background events we needed. For the background, we considered the following SM processes: $gg, q\bar{q} \to t\bar{t}$, $gg, q\bar{q} \to t\bar{t}h$, $gg, q\bar{q} \to t\bar{t}b\bar{b}$, $gg, q\bar{q} \to t\bar{t}t\bar{t}$, $q\bar{q} \to W^+W^-h$, $q\bar{q} \to ZZh$, and $q\bar{q} \to ZW^+W^-$.

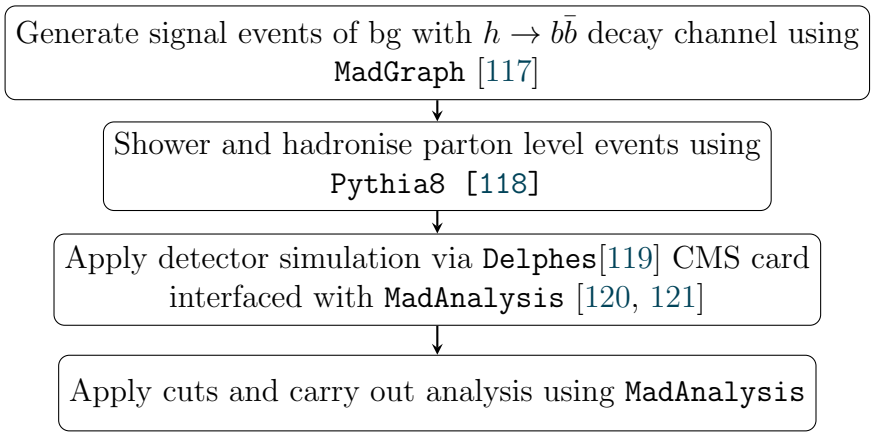


FIGURE 7.11: Illustration of the procedure used to generate and analyse MC events.

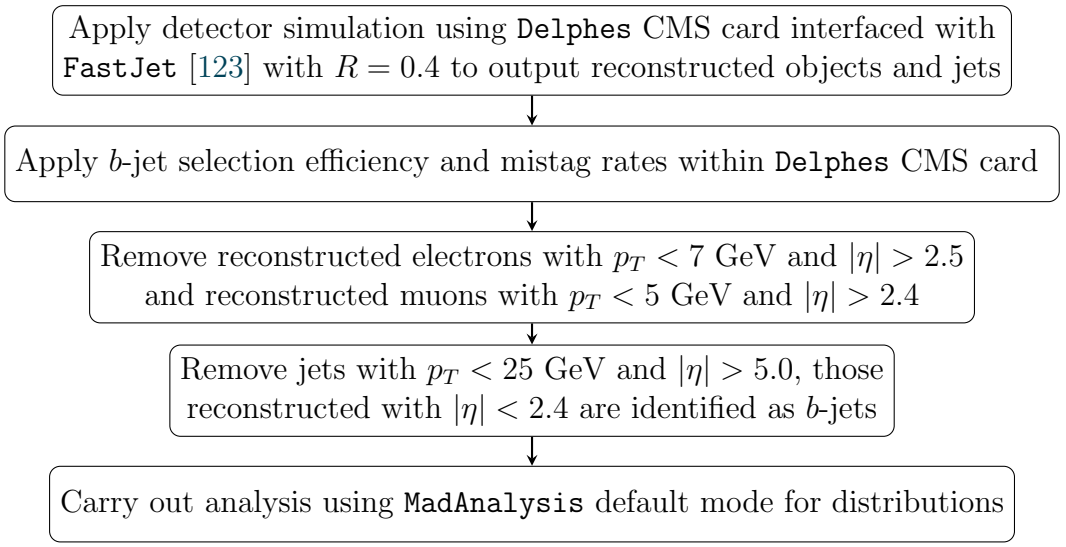


FIGURE 7.12: Illustration of the initial procedure for event reconstruction and jet clustering.

7.4.1.2 Cutflow

In this section, we discuss the cuts used to identify our SM/2HDM signals and reduce the relevant background

- Electrons were reconstructed and required to satisfy $p_T > 7$ GeV and $|\eta| < 2.5$.
- reconstructed muons are required to satisfy $p_T > 5$ GeV and $|\eta| < 2.4$.
- The jets were clustered using the anti-k_T algorithm [75] with a fixed cone size of $R = 0.4$.

- The jets considered in the analysis are additionally required to satisfy $p_T > 25$ GeV and $|\eta| < 5.0$. Jets reconstructed within $|\eta| < 2.4$ are identified as b-jets, in the presence of b-tagging.

- The b-tagging and mistagging rates were adopted from [148], where we apply a loose b-jet selection efficiency of 84% and mistag rates of 1.1% and 11% for gluon jets (c-jets) and light-quark jets respectively.

The mistagging rates may appear somewhat counterintuitive, with the c-jet rate being 10 times smaller than the light-quark rate, primarily due to the tagging's ability to reject c-jets and the high rate of light quarks. Further details on our cutflow analysis can be found in Fig. 7.12[6].

7.4.2 Results

We now present our results at both parton and hadron levels. In addition, we also present a signal-to-background analysis at the detector level comparing the 2HDM Type-II with the SM.

7.4.2.1 Parton Level Analysis

In this section, we perform a parton level analysis for both 2HDM Type-II and the SM, seeking to look for differences in their kinematical distributions. These differences can later be pursued at the hadron level to establish a BSM signal at the LHC, providing evidence beyond the large difference in the integrated cross-section yields of the two scenarios. As a result, to extract the shapes, we will only look at the kinematical distributions.

The p_T distributions for the 2HDM Type-II and SM events for the three heaviest objects in the final state of the bg sub-process are given in Fig. 7.13. The p_T distributions for the h and $W^{\pm}$ states in the two scenarios are clearly different: in the 2HDM Type-II, they are significantly harder than in the SM due to the different relative kinematics of the two models. This difference is driven by the signs of the κ_b values entering the cross-section. When compared to the SM, the p_T distribution for the top (anti)quark remains largely the same in the 2HDM Type-II as well. Thus, assuming an efficient reconstruction of h at the detector

[6]Only CMS cards are used in our analysis.

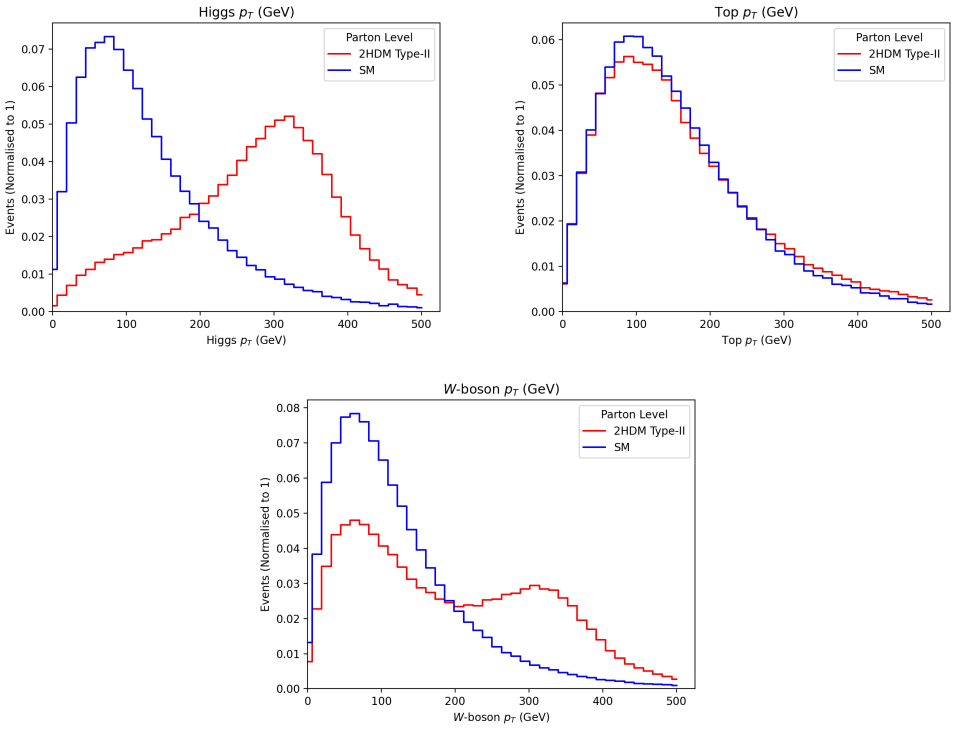


FIGURE 7.13: Upper panel: The p_T distributions of the Higgs boson (left) and top (anti)quark (right) at the parton level. Lower panel: The p_T distribution of the $W^{\pm}$ bosons at the parton level.

level (recall that the top (anti)quark decay would also produce a b-jet), we would expect equivalent differences in the p_T of the b-jet pair assigned to the h. We would also expect the different p_T's of the (prompt) $W^{\pm}$ bosons to transfer efficiently into the lepton spectra at the detector level, despite the dilution due to the near indistinguishability of leptons emerging from the top (anti)quark leptonic decays via (secondary) $W^{\pm}$'s.

Fig. 7.14 displays the p_T distribution of the b-quarks for both models. It is again clear that there are differences between the two scenarios; especially, the b-quarks have a wider range in p_Ts in the 2HDM Type-II when compared to the SM. As a result, we would expect the resulting b-jets to have a similar kinematic spread of p_Ts at the detector level.

7.4.2.2 Hadron Level Analysis

We now perform a detector-level analysis of the hadronised events in order to identify kinematic differences between our scenarios resulting from a legacy of those found at the parton level.

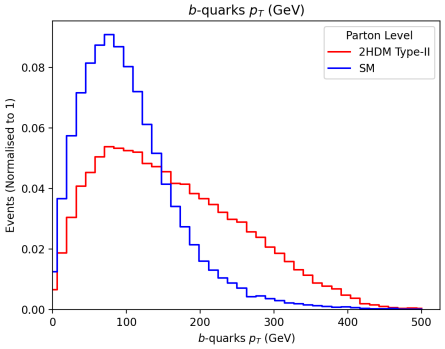


FIGURE 7.14: The p_T distribution for all b-quarks at the parton level.

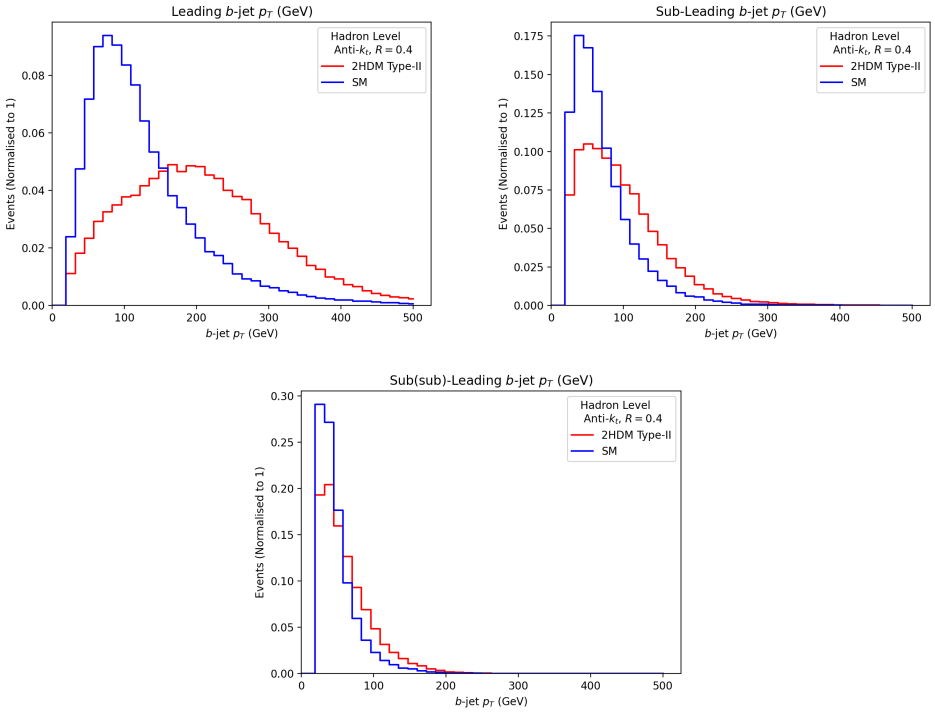


FIGURE 7.15: Upper panel: The p_T distributions of the leading b-jet (left) and sub-leading b-jet (right). Lower panel: The p_T distribution of the sub-sub-leading b-jet.

To begin, in Fig. 7.15, we present the ordered transverse momenta for the three b-jets expected to emerge from the decays of h and t (or indeed, $\bar{t}$). Notably, the leading b-jet's p_T coming from our 2HDM Type-II BP is quite distinct and much harder than that produced by the SM. This disparity arises from the fact that the b-(anti)quarks produced in h decays have significantly higher momentum in the BSM scenario than in the SM. Furthermore, there is a minor difference for sub-leading b-jet p_T, and the sub-sub-leading b-jet p_T's are nearly identical for both

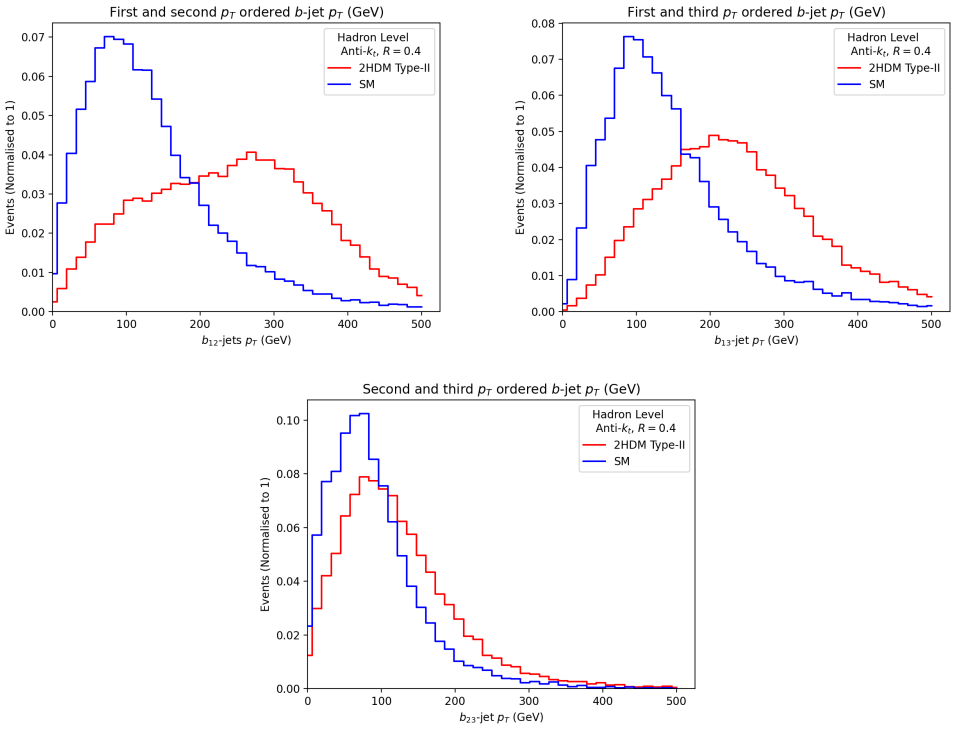


FIGURE 7.16: Upper panel: The p_T distributions of the leading plus sub-leading b- jets pair, b_{12} (left), and leading plus sub-sub-leading b-jets pair, b_{13} (right). Lower panel: The p_T distribution of the sub-leading plus sub-sub-leading b-jets pair, b_{23}.

models.

Our detector-level events should feature three b-jets with two of them coming from the h decay and the third one coming from the top (anti)quark one. We plot the combined transverse momentum distributions of these three b-jets permutated in pairs in Fig. 7.16. The p_T for the leading plus sub-leading b-jet pair (b_{12}) and for the leading plus sub-sub-leading b-jet pair (b_{13}) for the two models are quite different. The reason behind this is the differences in the p_T of the leading b-jet (as seen in Fig. 7.15) for both the models are more pronounced in comparison to those between the other pair of b-jets. This is also the reason we observe only a slight difference in p_T for the sub-leading plus sub-sub-leading b-jets pair (b_{23}).

Next, to reconstruct and identify the SM Higgs mass resonance, we analyse the invariant b-dijet mass, m_{bb}. From Fig. 7.17, we can see that the mass reconstruction in the 2HDM Type-II framework is somewhat sharper than in the SM. This indicates that the combinatorial effect is milder in the former case than in the latter. However, both distributions are somewhat misaligned with respect to the true MC value of the corresponding Higgs boson resonance.

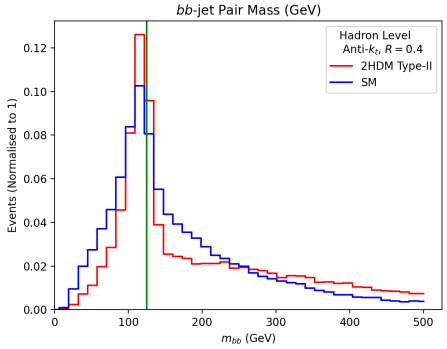


FIGURE 7.17: The invariant b-dijet mass distribution. The vertical green line represents the MC truth value of the h mass, $m_h = 125$ GeV.

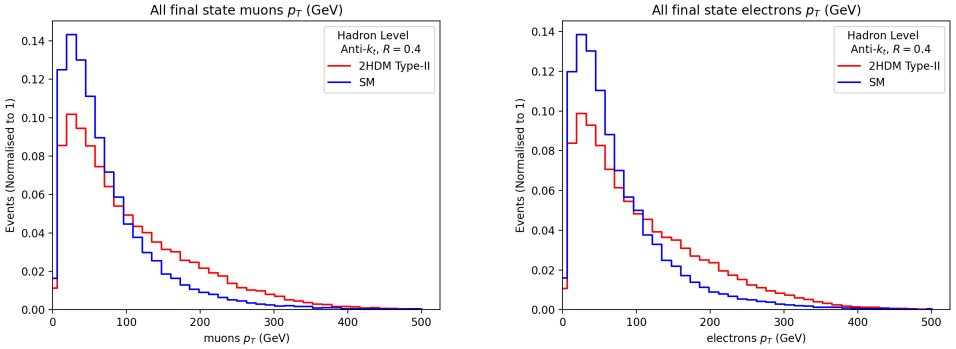


FIGURE 7.18: The p_T distributions of all muons (left) and electrons (right).

Finally, in Fig. 7.18, we examine the transverse momentum distributions for all muons and electrons coming from the top (anti)quark and $W^{\pm}$ boson decays. There is a small change in the shape of these kinematic distributions between models, with 2HDM Type-II events tending to have somewhat higher momenta than the SM events.

As the differences between the kinematical observables in the BSM and SM scenarios persist at the hadron level, we proceed to construct a suitable cutflow for our analysis. The purpose of this cutflow is to preserve the regions of phase space where differences are manifested simultaneously suppressing the background relative to the signal as much as possible.

7.4.2.3 Signal-to-background Analysis

We now compute the signal-to-background significance rates for both the SM and the 2HDM BP by enforcing the additional selection designed in Fig. 7.19.

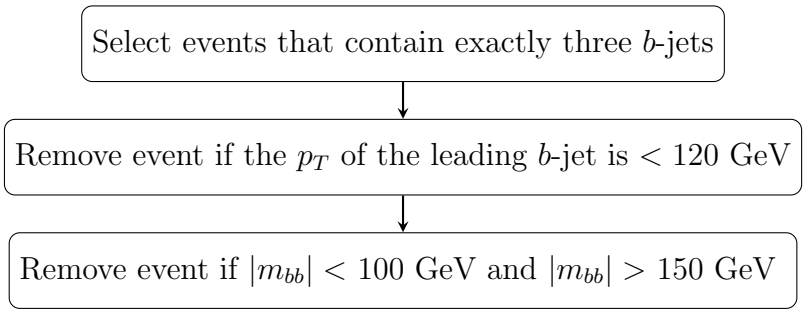


FIGURE 7.19: Additional event selection used to compute the final significances of the signal.

We calculate the expected event rates for the various signal and background processes, assuming an integrated luminosity of $\mathcal{L} = 3000$ fb^{-1} (corresponding to the one expected at the end of the HL-LHC stage), as follows

$$N = \sigma \times \mathcal{L}. \tag{7.2}$$

Tab. 7.2 contains the event rates for the signal (both models) and relevant backgrounds. It is clear that $gg, q\bar{q} \to t\bar{t}$ make up the dominant background followed by $gg, q\bar{q} \to t\bar{t}b\bar{b}$ and $gg, q\bar{q} \to t\bar{t}h$. Contributions from other backgrounds, including $gg, q\bar{q} \to t\bar{t}t\bar{t}$, $q\bar{q} \to W^+W^-h$, $q\bar{q} \to ZZh$ and $q\bar{q} \to ZW^+W^-$, are negligible.

The significance, Σ, is then calculated and given as a function of Signal (S) and sum of all Backgrounds (B) event rates:

$$\Sigma = \frac{N(S)}{\sqrt{N(B)}}. \tag{7.3}$$

Tab. 7.3 clearly shows that the signal within the 2HDM Type-II framework provides far better significance compared to the SM case. This indicates that the BSM signal would be detected at the HL-LHC significantly sooner than the SM one. Additionally, given the differences identified between the two theoretical scenarios at the detector level which in turn stem from rather different partonic behaviors, it may be possible to identify such a signal as being specifically due to the 2HDM Type-II.

This is definitely a preliminary conclusion, aiming solely at alerting the experimental community to the fact that 'SM-like Higgs boson production in association with a single top via the bg channel can be used to test the presence of an extended Higgs sector. In reality, much more could be done to put this on more solid footing. To begin, it should be emphasised that our analysis used a relatively simple

Process	
bg (2HDM)	290.415
bg (SM)	7.613
$pp \to t\bar{t}$	8119.259
$pp \to t\bar{t}h$	115.121
$pp \to t\bar{t}b\bar{b}$	1261.920
$pp \to t\bar{t}t\bar{t}$	0.167
$pp \to W^+W^-h$	0.436
$pp \to ZZh$	0.057
$pp \to ZW^+W^-$	0.169

TABLE 7.2: Event rates of signal (in both models) and backgrounds for $\mathcal{L} = 3000$ fb^{-1} upon enforcing all cuts.

	2HDM	SM
$\mathcal{L} = 3000$ fb^{-1}	2.980	0.078

TABLE 7.3: Final Σ values calculated for $\mathcal{L} = 3000$ fb^{-1} after enforcing all cuts.

cut-and-count method to calculate the signal significance; and only considered one final state for the signal ($h \to b\bar{b}$). Using a more advanced method for determining the signal significance, such as maximum-likelihood fit or machine learning approaches as well as analysing further signal final states through consideration of alternative h and W decays, we ultimately anticipate a substantial boost to the expected signal significance.

7.5 Additional Results using variable-R

Continuing the thesis theme, in this section, we briefly present the signal-to-background significance ratios for both the SM and the 2HDM BP by enforcing the additional selection designed in Fig. 7.19 for both the fixed-$R = 0.4$ and the variable-R [76] jet reconstruction procedures. For variable-R, we use $\rho = 50$ with $R_{\text{min}} = 0.4$ and $R_{\text{max}} = 2.0$.

Using Eqs.(7.2) and (7.3), we calculate the expected event rates and significance ratios for an integrated luminosity of $\mathcal{L} = 3000$ fb^{-1} for both reconstruction procedures, given in Tabs. 7.4 and 7.5 respectively.

Process	$R = 0.4$	Variable-R
bg (2HDM)	290.415	392.442
bg (SM)	7.613	8.865
$pp \to t\bar{t}$	8119.259	8007.459
$pp \to t\bar{t}h$	115.121	114.746
$pp \to t\bar{t}b\bar{b}$	1261.920	1240.800
$pp \to t\bar{t}t\bar{t}$	0.167	0.165
$pp \to W^+W^-h$	0.436	0.636
$pp \to ZZh$	0.057	0.060
$pp \to ZW^+W^-$	0.169	0.291

TABLE 7.4: Event rates of signal (in both models) and backgrounds for $\mathcal{L} = 3000$ fb^{-1} upon enforcing all initial cuts for the fixed-R and the variable-R jet reconstruction procedures.

	2HDM	SM
Fixed-R	2.980	0.078
Variable-R	4.055	0.091

TABLE 7.5: Final Σ values calculated for $\mathcal{L} = 3000$ fb^{-1} after enforcing all cuts for both the reconstruction procedure.

From Tab. 7.5, we can see that the variable-R approach is more efficient than the fixed-R method. A more thorough scan of the ρ parameter best suited for experimental setup could yield even higher significance ratios, increasing the likelihood of detecting such a signal at the LHC much beyond the results presented in Section 7.4.2.3. Again, these are preliminary results, aimed mainly at informing the experimental community that this process can be probed to test the presence of the extended Higgs sector, especially with the appropriate choice of jet-clustering algorithms, reconstruction procedure, and parameter settings.

7.6 Summary

To conclude, we investigated the parameter spaces of the 2HDM Type-I as well as II and discovered that, while the Type-I does not appear to contain any significantly larger cross-sections than in the SM for the 'SM-like Higgs boson production in association with single-top', the Type-II does. In this research, parameter space points for both the so-called 'wrong-sign solution' (of the bottom (anti)quark Yukawa coupling) and 'alignment limit' (most notably so in the former than parameter space configuration) were found, with a large proportion of them having a

cross-section far greater than that expected in the SM. This applies to the bg sub-process. A representative BP was then chosen within the region of 2HDM Type-II parameter space realising the 'wrong-sign solution' for detailed MC analysis. This analysis was performed in the presence of the most significant backgrounds (carried out in parallel to the SM case). This used a simple cut-and-count method at the detector level, which has led to a twofold result. First, the 2HDM Type-II signal may be established at the HL-LHC much before the SM one. Second, the corresponding excess events in the 2HDM Type-II would have kinematic features notably different from the SM case, offering a diagnostic scope of the underlying Higgs dynamics. We also presented a brief comparison between the variable-R and fixed-R jet reconstruction procedure, demonstrating that using variable-R further improved our significance ratios when employed for 2HDM Type-II configuration.

Our analysis was preliminary and aimed at drawing the attention of the experimental community to the potential of the bg sub-process. Specifically through triggering SM-like Higgs boson production in association with a single top to test a possible non-standard nature of EWSB and doing so better than the alternative two channels (bq and qq) can do, given that the production cross-sections for these are essentially the same in both these scenarios. In fact, only inclusive approaches have been used in pursuing this signature, i.e. capturing all three sub-processes at the same time, an approach that may be better avoided in the future, at least in the hunt for this very peculiar configuration of the 2HDM Type-II. The aforementioned 'wrong-sign solution', has withstood rigorous experimental scrutiny in the context of the SM-like Higgs boson signals to date.

Chapter 8

Image recognition for BSM searches in the hadronic final states with b-jets

The study presented in this chapter is not yet published, instead, it serves as a concept for future research paper once additional research is finished. I am the primary author of the work presented here.

8.1 Introduction

Since its inception, machine learning (ML) has evolved and been applied to countless problems, including those in particle physics. The applications of ML are numerous, ranging from enhancing jet tagging efficiencies [96, 171, 172], and designing alternative jet clustering techniques [173] to exploring parameter space for BSM searches [174, 175]. In the case of LHC searches, more advanced ML techniques have been used, in particular, mapping the final state of the detector into an image.

Many of these jet physics studies, such as jet tagging and identification, are substantially independent of the specific physics from which they originated and can therefore be applied to other specialised LHC searches. ML's exceptional flexibility to adapt to different situations is a special benefit. Without much alteration, an ML model can be easily trained on other datasets from various physics processes.

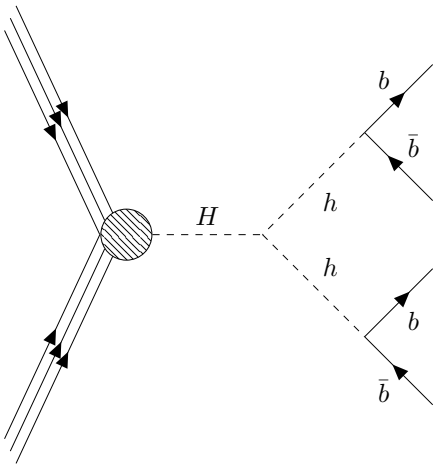


FIGURE 8.1: The 2HDM process of interest for this work.

In order to attempt to create a classification algorithm using ML, we need to organise the remnants of an event into a dataset that can be trained on. There are two main routes one takes in doing this, firstly using numerical information about the event, such as the jet η, ϕ, p_T, m, etc., as in [176]. Secondly one can build so-called jet images, where the constituents of final state jets found in a detector are mapped into the (η, ϕ) plane such that we can unravel the barrel-shaped detector into an image. A neural network can then be trained on this jet image data to perform the desired classification.

There is an extensive catalog of studies investigating jet image recognition [177, 178, 179, 180, 181]. A general feature of such studies is for the algorithm to learn on a jet substructure. Image recognition allows this to be achieved without the need for calculations and analysis of jet variables such as N-subjettinness.

In this study, we seek to utilize ML techniques to build a classifier for finding signs of new physics at the LHC using 2HDM Type-II. In particular, we will incorporate the advanced technique of image recognition by designing a CNN and visualizing what is 'seen' in a detector in particle physics experiments. Of course, there are well-established LHC searches using traditional techniques and alternative clustering algorithms for evaluating 2HDM final states. Here, we will deploy an image recognition-informed jet tagger's exceptional ability to map the jet-level information to an image and distinguish signals from relevant backgrounds.

We will be dealing with multijet events, it might therefore be advantageous to include as much of the jet information as possible in order to maximise our learning capability. Therefore, we will employ jet-level image recognition studies performed

in the context of boosted decays such that the entire signal event can be clustered into a single fat jet using a large cone size, and the big discriminatory feature for signal and background is the presence of a two-prong jet substructure, as done in [178] in the context of $H_{SM} \to b\bar{b}$ decay. For our study, we focus on b-jet final states from 2HDM Type-II with decay chains of the form $gg \to H \to hh \to b\bar{b}b\bar{b}$, see Fig. 8.1.

The layout of the chapter is as follows. In the next section, we will briefly provide an overview of ML techniques. We then outline the event generation toolbox and MC analysis details. Following that, we will present the pre-processing steps for image generation and the CNN model used for training purposes. Finally, we will present our results and draw our conclusions for further investigation.

8.2 Overview of ML techniques in High Energy Physics

In this section, we will review the principles of ML techniques and their uses in tackling real-world problems. Tom M. Mitchell provided a fundamentally operational definition of machine learning and widely quoted [182]: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E."

Generally speaking, ML is an umbrella term for solving prohibitively expensive problems by assisting machines in developing their own algorithms without any explicit human interference. ML techniques have been applied to text filtering, large language models in agriculture, computer vision, speech recognition, and medicine. However, in this context, we emphasize its application in high-energy physics relevant to our study.

8.2.1 ML categories

Modern ML techniques are broadly classified into three groups based on the nature of the problem to be solved, which we will quickly review here.

8.2.1.1 Supervised Learning

Supervised learning algorithms work on a dataset with features, where each sample is also tagged with a target or label. The term supervised learning comes from the idea that target "T" informed by the user, instructs the ML system on what to do. For example, a supervised learning algorithm can analyse the melon dataset and learn to classify melons into different species based on measurements.

There are two types of supervised problems:

Classification: Classification requires the prediction of class labels in order to determine which of x categories some input belongs to. For example, object recognition is a classification problem where an image is the input and a numeric code is an output for identifying the object in the image.

A well-established jet physics classification problem is jet tagging, where we attempt to categorise two target classes, i.e., a $b-$jet or not a $b-$jet.

Regression: Regression asks the machine to predict a numerical value for some given input. The regression task is similar to the classification method, however, the output format differs. An example of a regression task is the measurement of physical quantities such as prices of financial assets.

Going back to the jet tagging problem, a regression task can be used to correctly assign a b-tag to the jets we are certain about, resulting in the reduction of b-tag jets that were incorrectly allocated.

8.2.1.2 Unsupervised Learning

Unsupervised learning learns important information about the structure of the dataset containing input data without labels. An excellent application of unsupervised learning in physics is jet clustering techniques. In this situation, we have a dataset with relevant information, but we do not know how to split them into jets. Unsupervised learning can be used to cluster the particles into jets instead of traditional methods.

Spectral clustering [173] is one such example of unsupervised clustering learning techniques.

8.2.1.3 Reinforcement Learning

Reinforcement Learning educates the machine through trial and error to take the best action by developing a feedback loop between its environment and learning system. Reinforcement Learning can be used to drive autonomous vehicles by informing the machine when it made the correct judgments or train models to play games, allowing it to learn what actions should be performed over time.

8.2.2 Importance of Data in ML

As we have seen, the input data is a crucial component of ML. Together with the hyperparameters of the ML model, input data determines how useful the outcome will be. ML algorithms use data to understand the correlations and patterns between input variables and target labels that can be used for classification or prediction tasks. Data can be numerical, time series, or categorical and can come from a variety of sources. Numerical data consists of values that can be measured and ordered, such as income, house price, or age. Categorical data consists of the values of categories, such as tree type or gender. Typically, data can be divided into two branches

- **Labelled Data**: consists of label or target variables for which the model is attempting to predict.
- **Unlabelled Data**: does not consist of label or target variables for model predictions.

A basic notion for developing an ML model is dividing the input data into training and testing sets. The reason behind this is that models are developed by using a minimising function that symbolises the error in the model's predictions, and using the input data to train the model itself to assess its performance is unreliable.

Instead, we utilise a training subset to train the model and a testing subset to assess the model's performance, while keeping in mind that the input data is split between representative and random manner.

8.2.3 ML Models

In this section, we will briefly review some of the ML models, with a particular emphasis on deep learning models relevant to our study.

8.2.3.1 Supervised Learning Models

In literature, there are many supervised learning models; here, we briefly outline some of the most commonly used models:

- **Logistic Regression**: It is used to determine whether an input belongs to a specific group and to estimate the probabilities for each output class [183]. It is used to represent a binary dependent variable, which has two values, 0 and 1, to represent outcomes.
- **Linear Regression**: The most basic sort of regression is linear regression [184]. This model is used to identify correlations between two continuous variables.
- **Decision Trees**: These are flow-chart-like classifiers that are used to determine a branching approach to illustrate every possible consequence of a decision [185]. Each node in the tree demonstrates a test on a variable, with each branch indicating the result of that test.
- **Support Vector Machine (SVM)**: SVM [186] simply filters data into classes by giving a set of training samples, with each set flagged as falling into either of the two classes. SVM then constructs a model that allocates new values to one of the two classes.
- **K Nearest Neighbours (KNN)**: The KNN algorithm [187] groups the nearest objects in a dataset and determines the most average or frequent attributes among the objects.

Other supervised learning models include Naive Bayes [188], Random Forest [189], and Boosting algorithms [190].

8.2.3.2 Unsupervised Learning Models

In this subsection, we briefly outline some of the most commonly used unsupervised learning models:

- **K-Means Clustering Algorithm**: The K-Means algorithm [191] is used to classify unlabelled data. The algorithm works to detect similarities between objects and categorise them into K clusters.
- **Hierarchical Algorithm**: Hierarchical clustering [192] creates a tree of nested clusters without specifying the exact number of clusters.

8.2.4 Deep Learning Models

In this section, we summarize modern deep learning models used to solve practical problems, with a special emphasis on CNN.

Deep learning neural networks (NN) are a set of algorithms modeled after biological neurons. Despite their very simple core function, neurons can send and receive electrical impulses and can be organised in large arrays to solve complex problems. NN recognises numerical patterns stored in vectors such that all the real-world data (such as images, text, time series, or sound) must be translated into the relevant form. We can use NN to cluster the unlabeled data based on similarities between given inputs and also classify the labeled data for model predictions.

There are several types of NN that exist in the literature. Here, we discuss some of them that are relevant to this study.

8.2.4.1 Perceptron

A neural network can be built using a sequence of neurons, in which each neuron is fed an input y_i and associated weight w_i, which are then coupled with a static bias value. This information is fed to subsequent neurons until reaching an output z, which is given by:

$$z = \sum_{i}^{n} w_i y_i = \vec{y}^T \vec{w}. \tag{8.1}$$

This intermediate output is then subsequently delivered to an appropriate activation function, which determines the neuron's final output value given by:

$$h_w(\vec{y}) = \mathcal{H}(z). \tag{8.2}$$

This is an example of a single-layer perceptron (SLP) which can be generalised to build a multi-layer perceptron for tackling complex problems.

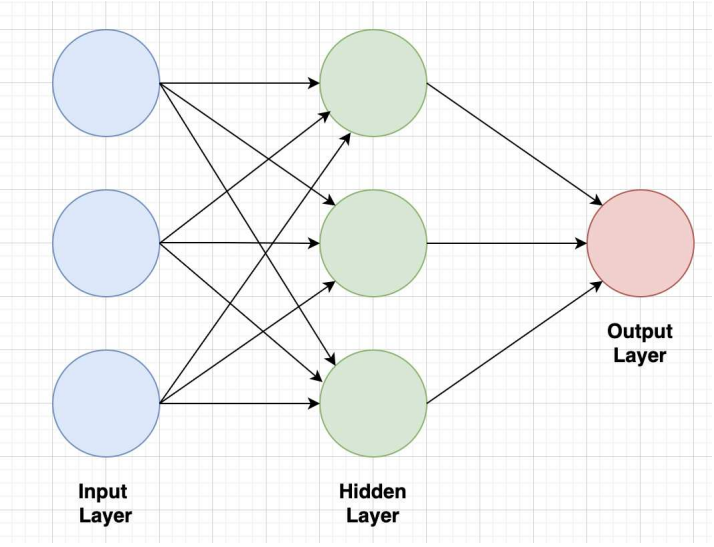


FIGURE 8.2: Simple MLP visual diagram with three characteristics, including one hidden layer.

8.2.4.2 Multi-Layer Perceptron (MLP)

MLPs are probably the most prevalent type of NN, consisting of the input layer, an output layer, and additional hidden layers missing in SLPs. The visual representation of MLP is shown in Fig. 8.2. The three layers' functions are as follows:

- `Input layer`- This is the first layer of nodes in the NN and defines the dimensions of the input vector.
- `Hidden layer`- Hidden layers are intermediary nodes that partition the input space into soft boundary regions. They are fed weighted inputs and generate outputs using an activation function.
- `Output layer`- This layer provides the output of the NN.

Each neuron in a particular layer is connected to every neuron in the previous and next layers, making MLPs a fully connected network.

To train an MLP NN, in addition to the method outlined in the previous section, one must compute the loss function and use a backpropagation algorithm to adjust the weight to minimise the loss. The mathematical nuances of the loss function and backpropagation are beyond the scope of this chapter.

Another distinction between the two perceptrons is the choice of activation function. To model non-linear data, activation functions like sigmoid, rectified linear

unit (ReLU) and tan are used with softmax acting as an output layer activation function.

MLPs have an advantage over SLPs in that they can be used for deep learning but are complex to design and may take longer to train depending on the number of hidden layers.

8.2.4.3 CNN

In this section, we will discuss our final deep-learning model, CNN [193], which we will use later on in the study. CNNs are a type of neural network that have a grid-like structure for processing data and are highly effective for time-series data and image data recognition tasks.

Images can be used as inputs by translating each pixel to a numerical scale expressing color values. Traditionally, NN layers would have to employ matrix multiplication of parameters, having a separate parameter characterising the interaction between each input and output unit, resulting in every output unit interacting with every input unit. However, it has been found that, in certain circumstances, the performance of more complex images is highly limited.

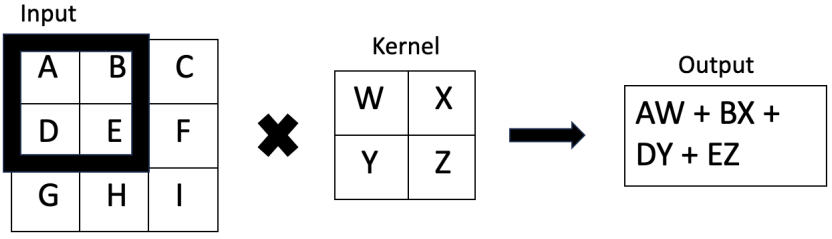


FIGURE 8.3: An example of a subset of an image being reduced to an element of the output tensor by the convolutional kernel layer. The black box illustrates the appropriate upper-left region of the input being used by the kernel layer to create the element of the output tensor.

CNN, on the other hand, uses a convolutional layer, which applies a small filter (also known as the kernel) to a subset of the input image and performs element-wise multiplication representing them as a single element of the output tensor (see Fig. 8.3). This procedure assists the network in learning the edges, texture, local patterns, and high-level visual features from the data. This method also reduces the amount of data being used for model training, resulting in better computational performance.

After convolutional layers, CNNs frequently employ pooling layers to downsample the dimensionality of the feature maps and reduce computational complexity but in a more elementary fashion. Pooling methods that are commonly used include max pooling and average pooling, shown in Fig. 8.4.

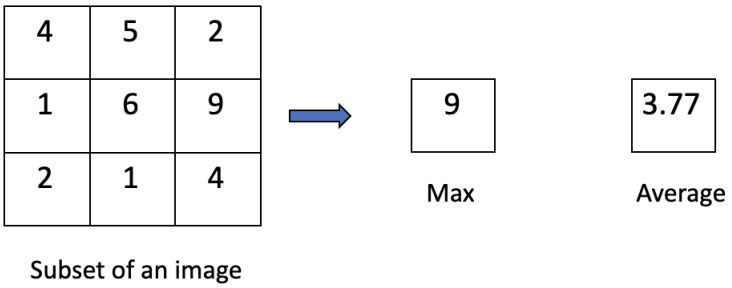


FIGURE 8.4: An example of a subset of an image being reduced to a single element using max pooling and average pooling.

CNNs have shown exceptional performance in various computer vision tasks and have been developed and adapted for use in other domains. These architectures differ in depth, layer arrangement, and network design choices to optimise performance for various applications and processing resources.

In the next sections, we will use CNNs to build a classifier for detecting signs of new physics beyond the standard model (BSM) at the LHC and visualising what is 'seen' in particle physics detectors.

8.3 Methodology

The goal of this research is threefold: select a suitable benchmark and generate events for training purposes, pre-process the relevant data and map them into a sample of images, and adapt these methods to produce a suitable classifier for BSM final states from 2HDM Type-II.

8.3.1 Simulation Details and Cutflow

We first select the phenomenologically preferred 2HDM Type-II benchmark point where the discovered 125 GeV Higgs is the lighter of the two scalars with a heavier

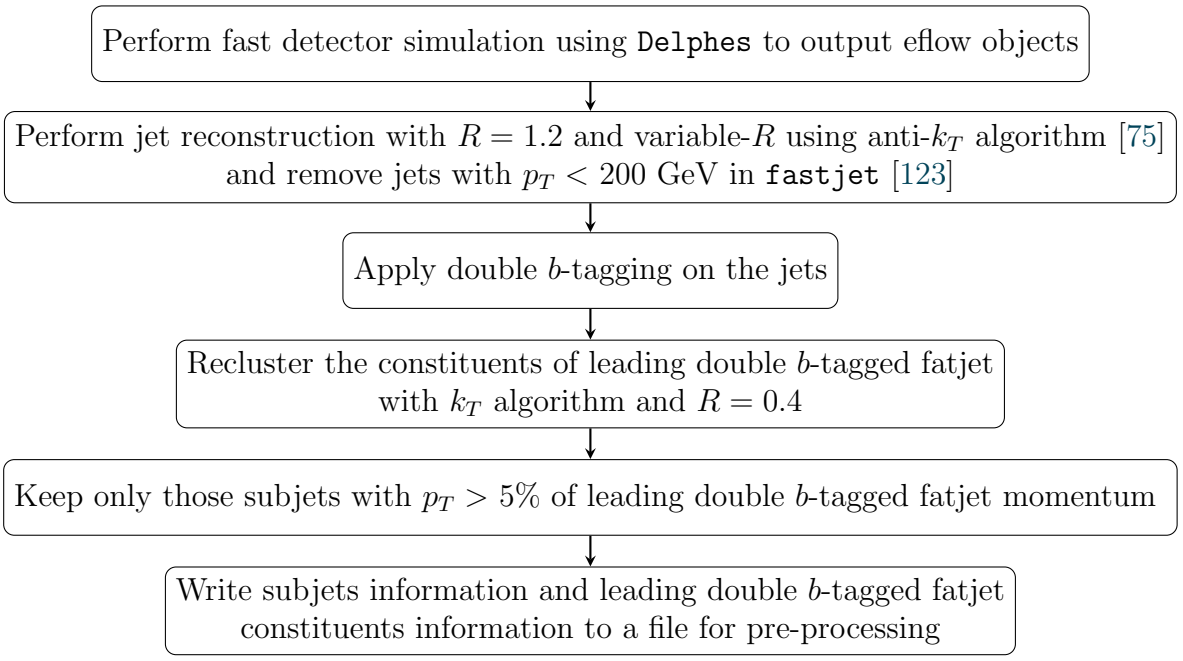


FIGURE 8.5: Description of the procedure used to analyse generated events for ML training.

CP-even Higgs boson mass set as $m_H = 700$ GeV. We test the benchmark against 2HDMC [40], HiggsBounds [113], HiggsSignals [114] as well as checking flavor constraints with SuperISO [115]. The 2HDM parameters and selected benchmark point information are given in Tab. 3.2 (see Point1 for the reference).

We generate samples of $\mathcal{O}(10^5)$ at $\sqrt{s} = 13$ TeV for the process $gg \to H \to hh \to b\bar{b}b\bar{b}$ using Madgraph5 [117]. We also consider leading SM backgrounds, such as:

$$gg, q\bar{q} \to t\bar{t} : \; p_T^{\text{gen}}(t) > 250 \text{ GeV},$$

$$gg, q\bar{q} \to b\bar{b}b\bar{b} : \; p_T^{\text{gen}}(b) > 100 \text{ GeV},$$

$$gg, q\bar{q} \to Zb\bar{b} : \; p_T^{\text{gen}}(Z) > 250 \text{ GeV}, \; p_T^{\text{gen}}(b) > 200 \text{ GeV}.$$

Here, we apply the generation level cuts within Madgraph5 to improve the selection efficiency and ensure a sensible signal to background analysis. We shower and hadronise the generated events using Pythia8 [118] with MPIs and ISR/FSR switched on. To perform a realistic MC analysis we use Delphes [119] to apply detector simulations and perform jet reconstruction and analysis using MadAnalysis5 [120, 121]. A full description of the cutflow is given in Fig. 8.5. For variable-R, we use $\rho = 300$ with $R_{min} = 0.4$ and $R_{max} = 2.0$. These values are influenced by the

p_T scale of the b-jets as well as the scan on ρ. For fixed-R, we have used a wider cone size of $R = 1.2$ to capture the particles into fat b-jets.

We also implement a simplified (MC truth-informed) double b-tagger. For events clustered with a fixed-R cone size, we look for jets that contain two b-quarks present within the angular distance ΔR and tag them as double b-tagged fat jets. For the variable-R approach, we take the size of the tagging cone as the effective size R_{eff} of the jet.

8.3.2 Construction of Jet Images

In this section, we describe in detail the generation of jet images as well as the reasoning for certain preprocessing steps. Our procedure substantially resembles that reported in [181].

8.3.2.1 Input Data

Following event generation using `Pythia8`, jets are reconstructed with the anti-k_T algorithm [75, 137] using EFlow objects information sourced from photons, neutral, and charged hadrons in the detector simulation (refer to Fig. 8.5 for full detail). Subsequently, we select the leading double b-tagged fatjet and recluster the constituents into subjets with the k_T algorithm using the fixed size of $R = 0.4$. Only subjets with more than 5% of the leading double b-tagged fatjet momentum are kept for image preprocessing. In particular, we store:

- Subjets information:
  $n^{event},\ n^{subjet},\ p_T^{subjet},\ m^{subjet},\ \eta^{subjet},\ \phi^{subjet}$
- Leading double b-tagged fatjet information:
  $n^{event},\ n^{jetcons},\ p_T^{jetcons},\ m^{jetcons},\ \eta^{jetcons},\ \phi^{jetcons}$

This data is in raw form and to successfully convert it into jet images for visualisation, we apply the preprocessing steps of translation, pixelisation, rotation, reflection, cropping, and normalisation. These steps are designed to eliminate spatial symmetries of sorts so that CNN can readily pick out the subjets and learn the substructure, thereby distinguishing Higgs jets from the relevant backgrounds.

8.3.2.2 Preprocessing Steps

The jet images are preprocessed to enable ML techniques to learn the distinguishing features between signal and background while avoiding learning symmetries of space-time. This approach can significantly enhance performance while also reducing the size of the sample used for testing.

In this sub-section, we describe the complete jet image production and preprocessing stages as follows:

- **Translation**:
  The first step is the translation where we translate all constituents of the leading double b-tagged fatjet in (η, ϕ) space such that the leading subjet is placed at the origin.
- **Pixelisation**:
  The next step is to construct a pixel grid of size (0.1×0.1) in (η, ϕ) space, and then the transverse momentum measured within each pixel is used to create a jet image.
- **Rotation**:
  The third step is to rotate the jet image such that the subleading subjet is placed directly beneath the leading subjet. In cases where no subjets are found, rotate the image to align the principal component along the vertical axis.
- **Reflection**:
  The jet image is then reflected to place the third leading subjet on the right-hand side. If only two subjets are present, the image is reflected to make sure that the total intensity of the image is highest on the right side. A cubic spline interpolation is utilised whenever the modified pixels do not line up with the original image pixels.
- **Cropping and normalisation**:
  The last step is to crop the jet image to a fixed size of 24×24 and then normalise the pixel intensities such that their squared sum evaluates to one.

After completing all the preprocessing steps, we obtain a jet image for each event based on the leading double-b tagged fatjet constituent and subjet information. Next, we aim to classify the signal from backgrounds using the stack of jet images represented as an array of shape $(24 \times 24 \times 1)$.

8.3.3 Average Jet Images

After laying out the methodology for the jet image generation, we will proceed to examine the distinguishing features of the signal and background images that we intend to exploit for training our model. To avoid the impracticality of scrolling through a gallery containing a large number of images, we present a representative average jet image of N events.

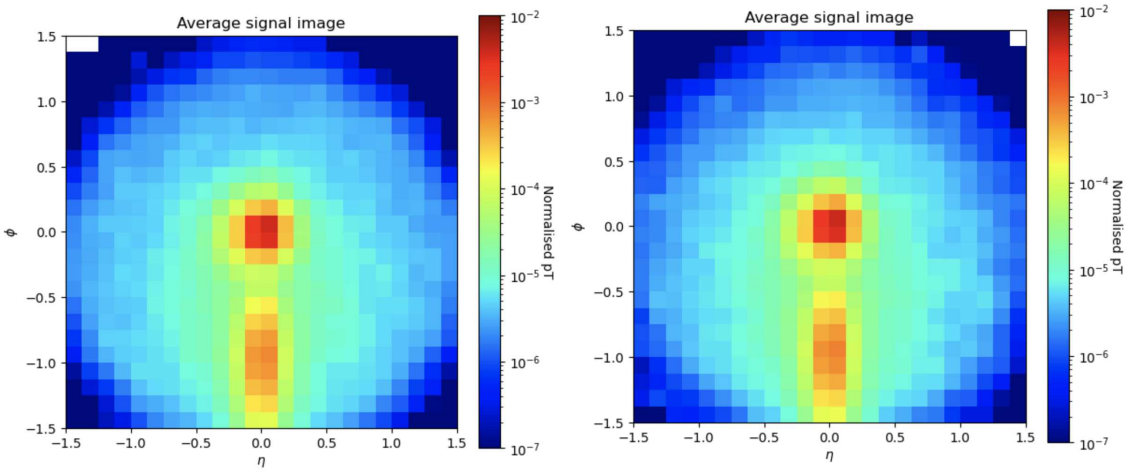


FIGURE 8.6: Left panel: The average signal image for leading double b-tagged jets coming from the process $gg \to H \to hh \to b\bar{b}b\bar{b}$ for fixed $R = 1.2$. Right panel: The average signal image using the variable-R approach.

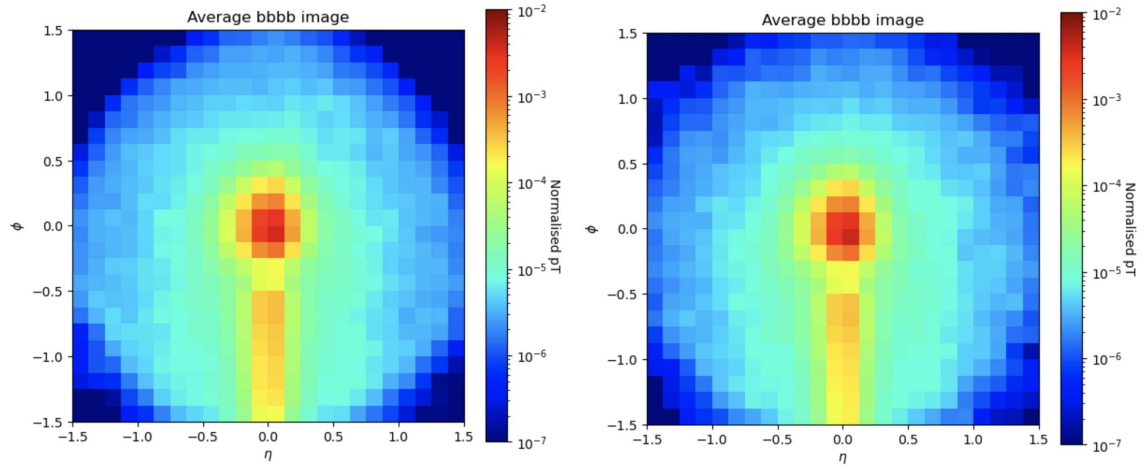


FIGURE 8.7: Left panel: The average background image for leading double b-tagged jets coming from the process $pp \to b\bar{b}b\bar{b}$ for fixed $R = 1.2$. Right panel: The average background using the variable-R approach.

Figs. 8.6 – 8.9 contains the average images for the signal and relevant backgrounds for the jet reconstruction procedures. The figures depict the general substructure of the leading double-b tagged wide cone jets for each process. In the signal image,

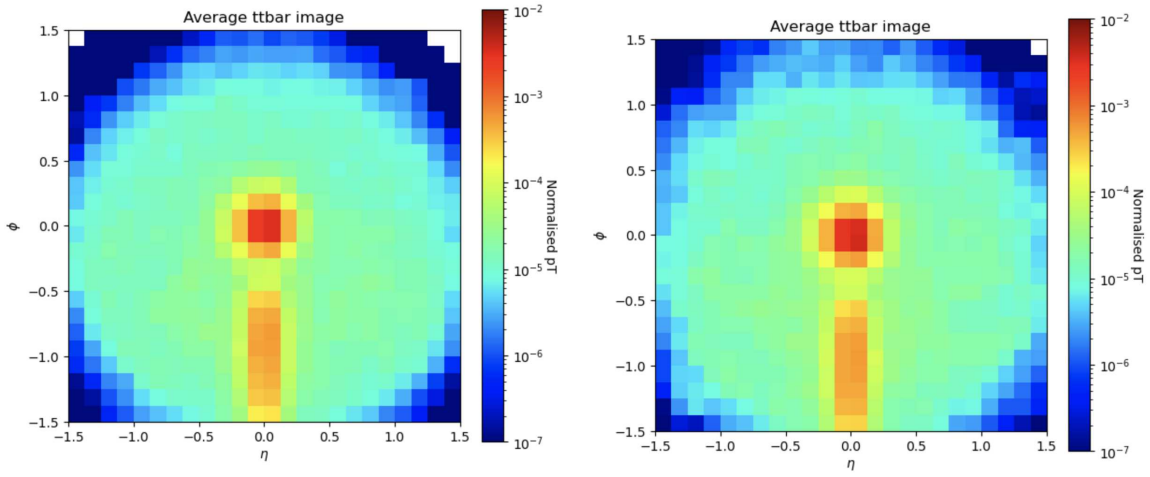


FIGURE 8.8: Left panel: The average background image for leading double b-tagged jets coming from the process $pp \to t\bar{t}$ for fixed $R = 1.2$. Right panel: The average background using the variable-R approach.

we can see a prominent two-prong structure where the second subjet is visible and spatially defined as expected.

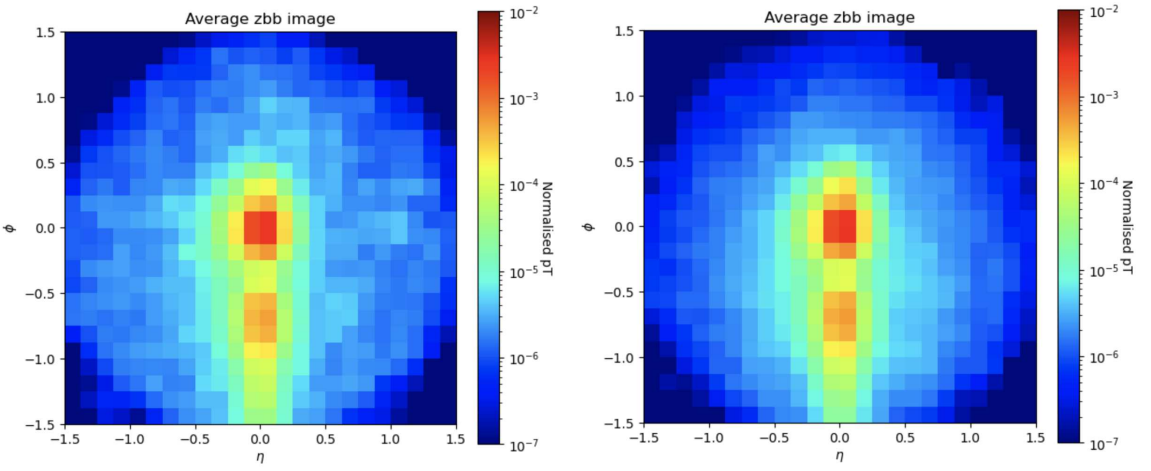


FIGURE 8.9: Left panel: The average background image for leading double b-tagged jets coming from the process $pp \to z b\bar{b}$ for fixed $R = 1.2$. Right panel: The average background using the variable-R approach.

For $pp \to b\bar{b}b\bar{b}$ background, we observe that the leading fatjets do not have a well-defined substructure. This is due to the fact that leading fatjets are generated by QCD processes with no boosting, hence we rarely see the two-prong structure visible in the signal image. For the $pp \to t\bar{t}$ background, there is a hint of a second subjet but it is softer and somewhat more smeared in comparison to the signal. We can also see a lot of low p_T activity spread across the image owing to the fact that there is a more intricate substructure associated with the b-jets and jet from $W^{\pm}$ boson originating from $t\bar{t}$ decays. Visually, the $pp \to z b\bar{b}$ background

closely resembles the signal substructure, with two different subjets. However, when compared to the signal, the average *zbb* image looks more compact.

Returning to the differentiation between the fixed-R and variable-R average images for signal and three backgrounds, we find little difference between the two sets of images. This is due to the fact that these images are based on the constituents and subjet information of the leading double-b tagged jets, which may differ slightly due to the different reconstruction procedures but still preserve the true substructure of the particles involved in the processes.

8.3.4 CNN Model Architecture

After taking a look at the jet images we're training on, in this section, we will describe the CNN architecture utilised to train the classifier and present preliminary results. This CNN architecture can be thought of as a toy CNN prototype used to test the viability of distinguishing the signal from the relevant background images.

The CNN architecture comprises of following layers:

- 2D Convolutional layer, (3×3)
- 2D Convolutional layer, (3×3)
- Max Pooling layer, (2×2)
- 0.5 Dropout layer
- Flattening layer, length 128
- 0.25 Dropout layer
- 2 node output layer.

The convolutional and pooling layers aid in extracting meaningful information from the images. To avoid overfitting the model, we use the dropout layer to drop specific nodes, while a flattening layer is used to feed the data into a fully connected MLP-like structure. Finally, we design a two-node output layer defining the probability of distinguishing a jet image as signal (1) or relevant background (0). To train the toy model, we will use a simple 50/50 split between signal and background data and 20 epochs. Of course, a final CNN model based on

prospective improvements from the next section would be built by scanning over the tunable hyperparameters to improve the classifier's efficiency.

8.4 Results

In this section, we present the performance of the CNN model described in Section 8.3.4.

To assess the model's training, we plot the progress of both loss and accuracy on both the training and validation datasets. Next, we examine the CNN output score for each image in the validation dataset, which is a score ranging from 0 and 1. This can be interpreted as the probability of the production of a specific instance from the signal. When the output score is close to 1, the model predicts it is a signal, and when it's close to 0, the model predicts it is a background. Finally, a receiving output characteristic (ROC) is plotted to evaluate the area under the curve (AUC).

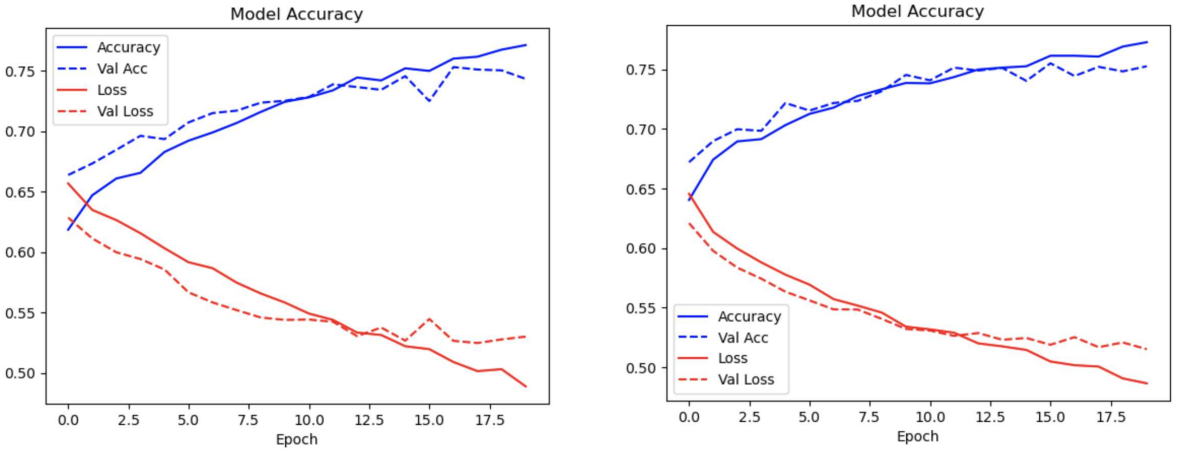


FIGURE 8.10: Left panel: The accuracy and loss progression training across 20 epochs for the fixed-R case. Right panel: The accuracy and loss progression training across 20 epochs for the variable-R case.

Fig. 8.10 shows the accuracy and loss progression training across 20 epochs for the reconstruction procedures. We can see that both the algorithm's validation accuracy and loss closely mirror the training sets. This is a good sign as it depicts that our model is not overfitting and can be generalised to the new set of images.

In Fig 8.11, we assess the performance of the model by splitting the signal and background datasets to examine whether the model's predictions match the truth information for both fixed-R and variable-R algorithms. Background events clearly

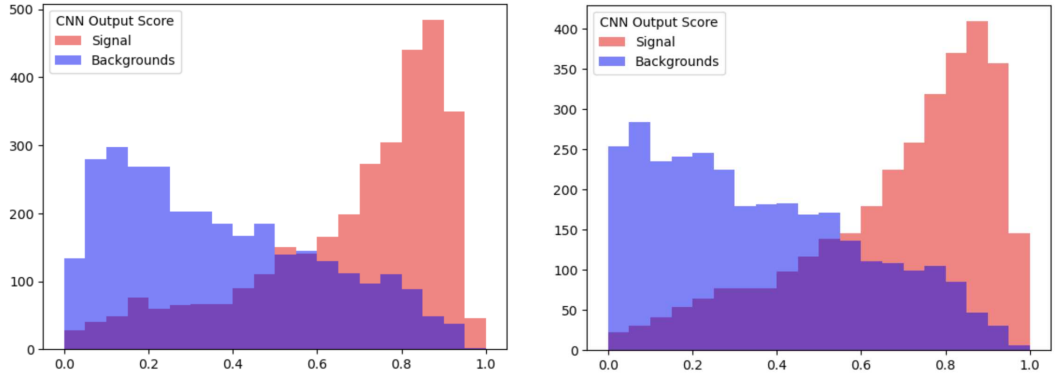


FIGURE 8.11: Left panel: The CNN model output score in the validation set for fixed R case. Right panel: The CNN model output score in the validation set for the variable-R case.

forecast values closer to zero, whereas the signal events peak around 1.0. However, we can see some overlap, more for variable-R than fixed-R, indicating that the model is somewhat struggling to correctly identify the signal from backgrounds and requires additional refinement.

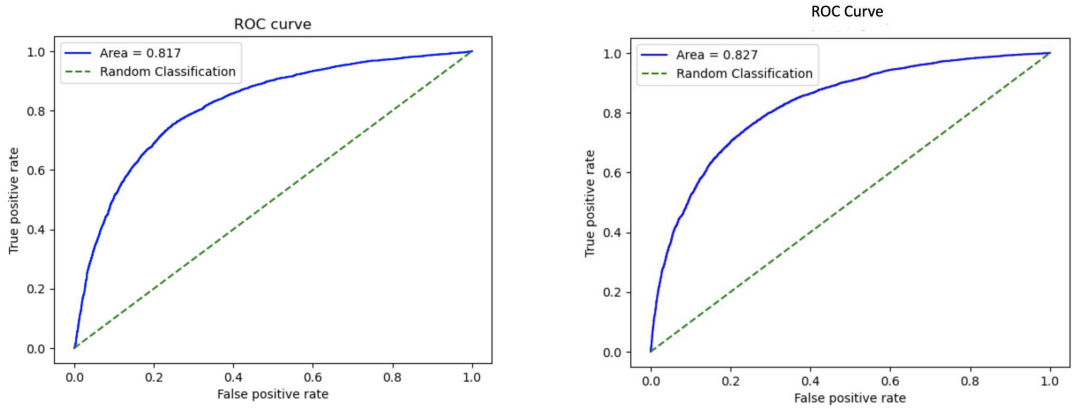


FIGURE 8.12: Left panel: The ROC curve plot to show the performance of the model's final iteration in training for the fixed R case. Right panel: The ROC curve plot to show the performance of the model's final iteration in training for the variable-R case.

Lastly, we plot the ROC to evaluate the AUC for the data in the validation set, see Fig 8.12. The ROC plot represents the true positive rate (TPR) and false positive rate (FPR) given by:

$$\text{TPR} = \frac{\text{TP}}{\text{P}} \quad ; \quad \text{FPR} = \frac{\text{FP}}{\text{N}}, \tag{8.3}$$

where TP represents true positives, P is the condition positive, FP denotes false positives, and N is the condition negative. Points on the AUC correspond to

a distinct decision boundary for the model. With an adequately set decision boundary, we can see that the model can reach a TPR of 0.81 for fixed-R and 0.82 for variable-R with a background rejection of less than 0.3 for both cases. Another key point is that once the variable-R parameter ρ is fine-tuned, we may expect considerably better results when compared to fixed-R, increasing the efficiency of distinguishing the signal from backgrounds.

8.5 Conclusions and Future Work

In this chapter, we presented a working toy model for jet visualisation using CNN for classifying the 2HDM Type-II signal $gg \to H \to hh \to b\bar{b}b\bar{b}$ from the relevant backgrounds $pp \to b\bar{b}b\bar{b}$, $pp \to t\bar{t}$ and $pp \to zb\bar{b}$ using fixed-R and variable-R jet reconstruction procedures. We showed that a simple ML CNN model can obtain an ROC area under the curve of up to 0.81 for fixed-R and 0.82 for variable-R procedures when trained for the signal jet images against a balanced mixture of the relevant backgrounds. These preliminary results further inspire us to create a more robust model to classify two sets of jet images.

The potential model improvement would be to incorporate more information, like in [178]. In our model, we solely train on leading double-b tagged fatjet constituents and subjets information. To include more information, we can add a second stream to the CNN model that uses global event information covering more of the detector picture. Another adjustment would be to filter our event samples using a more sensible selection procedure guided by the experimental setup. The inclusion of p_T window and masses of fatjets in the cutflow can help minimise the backgrounds to a level comparable to the signal, hence improving the model's performance.

To train our binary model, we employed a simple 50/50 split between signal and background data; if the data is not evenly split, the network will be biased towards the class with more samples. This is, however, not true for the underlying physics, where each process has different cross-sections and event rates. If the event rate for the signal is very rare when compared to the backgrounds, one might use a different ML method, such as anomaly detection, or continue to train the model using a 50/50 split. Another aspect to consider is the grouping of background data into a single class, despite knowing that each background has unique characteristics distinguishing it from the signal. Although this is a useful solution, more research should be done to examine the separation of power between each background separately.

These are some of the suggestions that can help address the input and physics problems. Of course, a final model based on prospective improvements would be constructed by scanning over the tunable hyperparameters to improve the classifier's efficiency.

Part III

Summary and Final Comments

Chapter 9

Conclusions

The goal of high-energy physics particle detectors such as LHC and HL-LHC, is to unravel new BSM physics. The particle collisions inside these detectors often involve crowded final state hadronic events of interest, along with unwanted radiation coming from UE, MPI, and PU. This extra radiation obscures the features of the relevant final state and must be carefully removed by using mapping methodologies to extract the relevant physics.

The final state hadronic events of interest can be mapped into jets, which are collimated sprays of hadrons formed after parton shower, hadronisation, and heavy flavor decays. The precise reconstruction of jets using jet clustering algorithms is, therefore, crucial for tracing the initial hard interactions of interest as well as looking for new physics signs.

In this thesis, we evaluate the prospective scope of the LHC experiments in accessing BSM Higgs signals with high multiplicity b-jet final states using the simplest possible extension to the Higgs sector known as the 2HDM. The $b\bar{b}$ decay channel is the dominant production method for the SM-like Higgs boson at the LHC. Therefore, it is crucial to assess the current state of phenomenological methods for extracting high multiplicity b-jet final states and to develop new studies for Higgs boson pair production in the BSM framework.

In Chapter 5, we compared the performance of fixed-R and variable-R algorithms (interfaced with the anti-k_T algorithm), in the presence of different resolution parameters, acceptance cuts, and reconstruction procedures to fully resolve high multiplicity b-jet final states resulting from $gg, q\bar{q} \to H \to hh \to b\bar{b}b\bar{b}$ decay chain. We considered both $m_H > m_h = 125$ GeV and $m_H = 125$ GeV $> m_h$ scenarios for

investigation. We found that using the variable-R reconstruction approach when combined with quality cuts resulted in a considerable improvement in signal yield and signal-to-background significance rates.

In Chapter 6, we revisited the topic of comparing the performance of fixed-R and variable-R reconstruction procedures in resolving fully hadronic final states derived from $gg, q\bar{q} \to H \to hh \to b\bar{b}b\bar{b}$. In Chapter 5, we obtained results for slim b-jets, with no merging, but here, we investigate the scenario with boosted topology, in which $b\bar{b}$ pairs coming from 125 Higgs boson merge into a fat b-jet. We discover that variable-R not only yields better reconstructed Higgs boson mass peaks but also enhances the signal-to-background significance ratios. Furthermore, the performance of the two clustering procedures was also tested to reconstruct jets with PU with variable-R emerging as the winner.

In Chapter 7, we tested the possibility of detecting cross-section at the HL-LHC for the production of $bg \to twh$, $bq \to tqh$, and $qq \to tbh$ processes. These processes allow direct access to Yukawa coupling values and have the potential to yield significantly larger cross-sections than those from SM in the wrong-sign scenario of the 2HDM. In the 2HDM Type-II model, $bg \to twh$ benchmark points achieved much larger cross-sections than SM despite limited available parameter space. An analysis using the 2HDM Type-II highest cross-section BP revealed that it yields different kinematical distributions and higher significance rates than the SM. This gives compelling evidence for testing the 2HDM Type-II model instead of SM for detecting SM Higgs boson production in association with a single top quark at the LHC.

In Chapter 8, we evaluated more advanced strategies for 2HDM Type-II Higgs decays with high b-jet final states. By mapping the constituents of leading b-jet and subjet information onto images, we used deep learning methodologies to build and train a classifier to assess whether the jets in an event are coming from the signal or relevant backgrounds. In this chapter, we offered a proof-of-concept demonstrating CNN's ability to learn from jet information.

In all four of these initiatives, we observe that fine-tuning and modernising the traditional jet clustering algorithms in the presence of different resolution parameters, cutflow, and reconstruction procedures leads to better performance in unraveling the final states at the LHC. With the development of new ML techniques, we expect to see significantly improved signal efficiency and visibility at

the LHC. Furthermore, the development of new tools, such as `Magellan` (mentioned in Chapter 7), can provide more precise benchmark points for exploring Higgs production channels at the LHC.

With the HL-LHC set to commence operations in 2027, we anticipate witnessing more innovative approaches and the application of our findings to handle large-scale data, pushing the existing boundaries of our current understanding of nature's fundamental forces.

References

[1] A. Chakraborty, S. Dasmahapatra, H. Day-Hall, B. Ford, S. Jain, S. Moretti, E. Olaiya and C. Shepherd-Themistocleous, Eur. Phys. J. C **82** (2022) no.4, 346 [arXiv:2008.02499 [hep-ph]].

[2] A. Chakraborty, S. Dasmahapatra, H. Day-Hall, B. Ford, S. Jain and S. Moretti, Eur. Phys. J. C **83** (2023) no.4, 347 [arXiv:2303.05189 [hep-ph]].

[3] C. Byers, S. Jain, S. Moretti and E. Olaiya, [arXiv:2303.09225 [hep-ph]].

[4] P. W. Higgs, Phys. Lett. **12** (1964), 132-133.

[5] P. W. Higgs, Phys. Rev. Lett. **13** (1964), 508-509.

[6] F. Englert and R. Brout, Phys. Rev. Lett. **13** (1964), 321-323.

[7] S. Weinberg, Phys. Rev. Lett. **19** (1967), 1264-1266.

[8] A. Salam, Conf. Proc. C **680519** (1968), 367-377.

[9] S. Khalil and S. Moretti, CRC Press, 2022, ISBN 978-1-138-33643-8.

[10] J. Goldstone, Nuovo Cim. **19** (1961), 154-164.

[11] J. Ellis, [arXiv:1312.5672 [hep-ph]].

[12] C. Quigg, Ann. Rev. Nucl. Part. Sci. **59** (2009), 505-555 [arXiv:0905.3187 [hep-ph]].

[13] M. Bustamante, L. Cieri and J. Ellis, [arXiv:0911.4409 [hep-ph]].

[14] S. Dawson, AIP Conf. Proc. **1116** (2009) no.1, 11-34 [arXiv:0812.2190 [hep-ph]].

[15] P. Skands, [arXiv:1207.2389 [hep-ph]].

[16] R. K. Ellis, W. J. Stirling and B. R. Webber, Camb. Monogr. Part. Phys. Nucl. Phys. Cosmol. **8** (1996), 1-435 Cambridge University Press, 2011.

[17] M. E. Peskin and D. V. Schroeder, Addison-Wesley, 1995, ISBN 978-0-201-50397-5.

[18] D. J. Gross and F. Wilczek, Phys. Rev. Lett. **30** (1973), 1343-1346.

[19] H. D. Politzer, Phys. Rev. Lett. **30** (1973), 1346-1349.

[20] S. Bethke, Nucl. Phys. B Proc. Suppl. **234** (2013), 229-234 doi:10.1016/j.nuclphysbps.2012.12.020 [arXiv:1210.0325 [hep-ex]].

[21] Y. Fukuda *et al.* [Super-Kamiokande], Phys. Rev. Lett. **82** (1999), 2644-2648 [arXiv:hep-ex/9812014 [hep-ex]].

[22] Q. R. Ahmad *et al.* [SNO], Phys. Rev. Lett. **87** (2001), 071301 [arXiv:nucl-ex/0106015 [nucl-ex]].

[23] B. Pontecorvo, Zh. Eksp. Teor. Fiz. **53** (1967), 1717-1725.

[24] D. A. Camargo, A. G. Dias, T. B. de Melo and F. S. Queiroz, JHEP **04** (2019), 129 [arXiv:1811.05488 [hep-ph]].

[25] E. Corbelli and P. Salucci, Mon. Not. Roy. Astron. Soc. **311** (2000), 441-447 [arXiv:astro-ph/9909252 [astro-ph]].

[26] S. M. Faber and R. Jackson, The Astrophysical Journal **204** (1976), 668683.

[27] A. McKellar, Publications of the Dominion Astrophysical Observatory Victoria 7, **251** (1941).

[28] A. A. Penzias and R. W. Wilso, The Astrophysical Journal **142** (1965), 419421.

[29] M. Romero Lamas, J. Phys. Conf. Ser. **1526** (2020), 012007.

[30] J. H. Christenson, J. W. Cronin, V. L. Fitch and R. Turlay, Phys. Rev. Lett. **13** (1964), 138-140.

[31] A. D. Sakharov, Pisma Zh. Eksp. Teor. Fiz. **5** (1967), 32-35.

[32] M. Kobayashi and T. Maskawa, Prog. Theor. Phys. **49** (1973), 652-657.

[33] G. C. Branco, P. M. Ferreira, L. Lavoura, M. N. Rebelo, M. Sher and J. P. Silva, Phys. Rept. **516** (2012), 1-102 [arXiv:1106.0034 [hep-ph]].

[34] J. F. Gunion, H. E. Haber, G. L. Kane and S. Dawson, Front. Phys. **80** (2000), 1-404 SCIPP-89/13.

[35] M. Aoki, S. Kanemura and O. Seto, Phys. Rev. Lett. **102** (2009), 051805 [arXiv:0807.0361 [hep-ph]].

[36] P. Ko, Y. Omura and C. Yu, JHEP **11** (2014), 054 [arXiv:1405.2138 [hep-ph]].

[37] J. Cao, P. Wan, L. Wu and J. M. Yang, Phys. Rev. D **80** (2009), 071701 [arXiv:0909.5148 [hep-ph]].

[38] A. Broggio, E. J. Chun, M. Passera, K. M. Patel and S. K. Vempati, JHEP **11** (2014), 058 [arXiv:1409.3199 [hep-ph]].

[39] L. Wang and X. F. Han, JHEP **05** (2015), 039 [arXiv:1412.4874 [hep-ph]].

[40] D. Eriksson, J. Rathsman and O. Stal, Comput. Phys. Commun. **181** (2010), 189-205 [arXiv:0902.0851 [hep-ph]].

[41] J. F. Gunion and H. E. Haber, Phys. Rev. D **67** (2003), 075019 [arXiv:hep-ph/0207010 [hep-ph]].

[42] P. M. Ferreira, B. Grzadkowski, O. M. Ogreid and P. Osland, JHEP **02** (2021), 196 [arXiv:2010.13698 [hep-ph]].

[43] H. Georgi and D. V. Nanopoulos, Phys. Lett. B **82** (1979), 95-96.

[44] A. Barroso, P. M. Ferreira and R. Santos, Phys. Lett. B **652** (2007), 181-193 [arXiv:hep-ph/0702098 [hep-ph]].

[45] H. E. Haber and O. Stål, Eur. Phys. J. C **75** (2015) no.10, 491 [erratum: Eur. Phys. J. C **76** (2016) no.6, 312] [arXiv:1507.04281 [hep-ph]].

[46] S. L. Glashow, J. Iliopoulos and L. Maiani, Phys. Rev. D **2** (1970), 1285-1292.

[47] V. D. Barger, J. L. Hewett and R. J. N. Phillips, Phys. Rev. D **41** (1990), 3421-3441.

[48] M. E. Peskin and T. Takeuchi, Phys. Rev. Lett. **65** (1990), 964-967.

[49] I. Maksymyk, C. P. Burgess and D. London, Phys. Rev. D **50** (1994), 529-535 [arXiv:hep-ph/9306267 [hep-ph]].

[50] S. Kanemura, M. Kikuchi and K. Yagyu, Nucl. Phys. B **896** (2015), 80-137 [arXiv:1502.07716 [hep-ph]].

[51] A. G. Akeroyd, A. Arhrib and E. M. Naimi, Phys. Lett. B **490** (2000), 119-124 [arXiv:hep-ph/0006035 [hep-ph]].

[52] F. Jegerlehner and A. Nyffeler, Phys. Rept. **477** (2009), 1-110 [arXiv:0902.3360 [hep-ph]].

[53] D. Chang, W. F. Chang, C. H. Chou and W. Y. Keung, Phys. Rev. D **63** (2001), 091301 [arXiv:hep-ph/0009292 [hep-ph]].

[54] G. Aad *et al.* [ATLAS and CMS], JHEP **08** (2016), 045 [arXiv:1606.02266 [hep-ex]].

[55] [CMS], [arXiv:2209.06197 [hep-ex]].

[56] [CMS], [arXiv:2208.01469 [hep-ex]].

[57] A. M. Sirunyan *et al.* [CMS], Phys. Lett. B **785** (2018), 462 [arXiv:1805.10191 [hep-ex]].

[58] A. G. Akeroyd, M. Aoki, A. Arhrib, L. Basso, I. F. Ginzburg, R. Guedes, J. Hernandez-Sanchez, K. Huitu, T. Hurth and M. Kadastik, *et al.* Eur. Phys. J. C **77** (2017) no.5, 276 [arXiv:1607.01320 [hep-ph]].

[59] V. N. Gribov and L. N. Lipatov, Sov. J. Nucl. Phys. **15**, 675-684 (1972).

[60] V. N. Gribov and L. N. Lipatov, Sov. J. Nucl. Phys. **15**, 438-450 (1972) IPTI-381-71.

[61] G. Altarelli and G. Parisi, Nucl. Phys. B **126**, 298-318 (1977).

[62] Y. L. Dokshitzer, Sov. Phys. JETP **46**, 641-653 (1977).

[63] J. E. Huth, N. Wainer, K. Meier, N. Hadley, F. Aversa, M. Greco, P. Chiappetta, J. P. Guillet, S. Ellis and Z. Kunszt, *et al.* FERMILAB-CONF-90-249-E.

[64] S. D. Ellis, Z. Kunszt and D. E. Soper, Phys. Rev. D **40** (1989), 2188-2222.

[65] G. F. Sterman and S. Weinberg, Phys. Rev. Lett. **39** (1977), 1436.

[66] F. Abe *et al.* [CDF], Phys. Rev. D **45** (1992), 1448-1458.

[67] G. C. Blazey, J. R. Dittmann, S. D. Ellis, V. D. Elvira, K. Frame, S. Grinstein, R. Hirosky, R. Piegaia, H. Schellman and R. Snihur, *et al.* [arXiv:hep-ex/0005012 [hep-ex]].

[68] G. P. Salam and G. Soyez, JHEP **05** (2007), 086 [arXiv:0704.0292 [hep-ph]].

[69] S. Moretti, L. Lonnblad and T. Sjostrand, JHEP **08** (1998), 001 [arXiv:hep-ph/9804296 [hep-ph]].

[70] T. Sjostrand, Comput. Phys. Commun. **28** (1983), 229

[71] W. Bartel *et al.* [JADE], Z. Phys. C **33** (1986), 23

[72] S. Catani, Y. L. Dokshitzer, M. Olsson, G. Turnock and B. R. Webber, Phys. Lett. B **269** (1991), 432-438

[73] S. D. Ellis and D. E. Soper, Phys. Rev. D **48** (1993), 3160-3166 [arXiv:hep-ph/9305266 [hep-ph]].

[74] Y. L. Dokshitzer, G. D. Leder, S. Moretti and B. R. Webber, JHEP **08** (1997), 001 [arXiv:hep-ph/9707323 [hep-ph]].

[75] M. Cacciari, G. P. Salam and G. Soyez, JHEP **04** (2008), 063 [arXiv:0802.1189 [hep-ph]].

[76] D. Krohn, J. Thaler and L. T. Wang, JHEP **06** (2009), 059 [arXiv:0903.0392 [hep-ph]].

[77] S. Chatrchyan *et al.* [CMS], JINST **3** (2008), S08004

[78] G. L. Bayatian *et al.* [CMS], CERN-LHCC-2006-001.

[79] S. Marzani, G. Soyez and M. Spannowsky, Lect. Notes Phys. **958** (2019), pp. Springer, 2019, [arXiv:1901.10342 [hep-ph]].

[80] J. Haller, R. Kogler and F. Tackmann, PUBDB-2018-03593, **155-168** (2018).

[81] J. Shelton, [arXiv:1302.0260 [hep-ph]].

[82] D. Krohn, J. Thaler and L. T. Wang, JHEP **02** (2010), 084 [arXiv:0912.1342 [hep-ph]].

[83] J. M. Butterworth, A. R. Davison, M. Rubin and G. P. Salam, AIP Conf. Proc. **1078** (2009) no.1, 189-191 [arXiv:0809.2530 [hep-ph]].

[84] S. D. Ellis, C. K. Vermilion and J. R. Walsh, Phys. Rev. D **80** (2009), 051501 [arXiv:0903.5081 [hep-ph]].

[85] A. J. Larkoski, S. Marzani, G. Soyez and J. Thaler, JHEP **05** (2014), 146 [arXiv:1402.2657 [hep-ph]].

[86] M. Dasgupta, A. Fregoso, S. Marzani and G. P. Salam, JHEP **09** (2013), 029 [arXiv:1307.0007 [hep-ph]].

[87] M. Cacciari and G. P. Salam, Phys. Lett. B **659** (2008), 119-126 [arXiv:0707.1378 [hep-ph]].

[88] D. Bertolini, P. Harris, M. Low and N. Tran, JHEP **10** (2014), 059 [arXiv:1407.6013 [hep-ph]].

[89] D. Krohn, M. D. Schwartz, M. Low and L. T. Wang, Phys. Rev. D **90** (2014) no.6, 065020 [arXiv:1309.4777 [hep-ph]].

[90] M. Cacciari, G. P. Salam and G. Soyez, Eur. Phys. J. C **75** (2015) no.2, 59 [arXiv:1407.0408 [hep-ph]].

[91] P. T. Komiske, E. M. Metodiev, B. Nachman and M. D. Schwartz, J. Phys. Conf. Ser. **1085** (2018) no.4, 042010

[92] [CMS], CMS-PAS-JME-16-003.

[93] V. Khachatryan *et al.* [CMS], JHEP **12** (2014), 017 [arXiv:1410.4227 [hep-ex]].

[94] [CMS], CMS-PAS-JME-15-002.

[95] A. M. Sirunyan *et al.* [CMS], JINST **13** (2018) no.05, P05011 [arXiv:1712.07158 [physics.ins-det]].

[96] A. Chakraborty, S. H. Lim, M. M. Nojiri and M. Takeuchi, JHEP **07** (2020), 111 [arXiv:2003.11787 [hep-ph]].

[97] J. A. Aguilar-Saavedra, Eur. Phys. J. C **81** (2021) no.8, 734 [arXiv:2102.01667 [hep-ph]].

[98] G. Kasieczka, T. Plehn, A. Butter, K. Cranmer, D. Debnath, B. M. Dillon, M. Fairbairn, D. A. Faroughy, W. Fedorko and C. Gay, *et al.* SciPost Phys. **7** (2019), 014 [arXiv:1902.09914 [hep-ph]].

[99] A. Cagnotta, F. Carnevali and A. De Iorio, Appl. Sciences **12** (2022) no.20, 10574.

[100] I. R. Tomalin [CMS], J. Phys. Conf. Ser. **110** (2008), 092033.

[101] A. M. Sirunyan *et al.* [CMS], Phys. Rev. Lett. **120** (2018) no.7, 071802 [arXiv:1709.05543 [hep-ex]].

[102] [ATLAS], ATLAS-CONF-2018-052.

[103] M. Aaboud *et al.* [ATLAS], Eur. Phys. J. C **78** (2018) no.7, 565 [arXiv:1804.10823 [hep-ex]].

[104] A. M. Sirunyan *et al.* [CMS], JHEP **04** (2019), 031 [arXiv:1810.05905 [hep-ex]].

[105] M. Aaboud *et al.* [ATLAS], Phys. Lett. B **781** (2018), 327-348 [arXiv:1801.07893 [hep-ex]].

[106] G. Aad *et al.* [ATLAS], Phys. Lett. B **716** (2012), 1-29 [arXiv:1207.7214 [hep-ex]].

[107] S. Khalil and S. Moretti,, "Supersymmetry Beyond Minimality: from Theory to Experiment", CRC Press (Taylor & Francis), 2017.

[108] U. Ellwanger, C. Hugonie and A. M. Teixeira, Phys. Rept. **496** (2010), 1-77 [arXiv:0910.1785 [hep-ph]].

[109] J. F. Gunion, H. E. Haber, G. L. Kane and S. Dawson, [arXiv:hep-ph/9302272 [hep-ph]].

[110] S. Moretti and W. J. Stirling, Phys. Lett. B **347** (1995), 291-299 [erratum: Phys. Lett. B **366** (1996), 451] [arXiv:hep-ph/9412209 [hep-ph]].

[111] A. Djouadi, J. Kalinowski and P. M. Zerwas, Z. Phys. C **70** (1996), 435-448 [arXiv:hep-ph/9511342 [hep-ph]].

[112] L. Scodellaro [ATLAS and CMS], [arXiv:1709.01290 [hep-ex]].

[113] P. Bechtle, O. Brein, S. Heinemeyer, O. Stål, T. Stefaniak, G. Weiglein and K. E. Williams, Eur. Phys. J. C **74** (2014) no.3, 2693 [arXiv:1311.0055 [hep-ph]].

[114] P. Bechtle, S. Heinemeyer, O. Stål, T. Stefaniak and G. Weiglein, Eur. Phys. J. C **74** (2014) no.2, 2711 [arXiv:1305.1933 [hep-ph]].

[115] F. Mahmoudi, Comput. Phys. Commun. **180** (2009), 1718-1719

[116] R. D. Ball *et al.* [NNPDF], JHEP **04** (2015), 040 [arXiv:1410.8849 [hep-ph]].

[117] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli and M. Zaro, JHEP **07** (2014), 079 [arXiv:1405.0301 [hep-ph]].

[118] T. Sjostrand, S. Mrenna and P. Z. Skands, Comput. Phys. Commun. **178** (2008), 852-867 [arXiv:0710.3820 [hep-ph]].

[119] J. de Favereau *et al.* [DELPHES 3], JHEP **02** (2014), 057 [arXiv:1307.6346 [hep-ex]].

[120] E. Conte, B. Fuks and G. Serret, Comput. Phys. Commun. **184** (2013), 222-256 [arXiv:1206.1599 [hep-ph]].

[121] E. Conte and B. Fuks, Int. J. Mod. Phys. A **33** (2018) no.28, 1830027 [arXiv:1808.00480 [hep-ph]].

[122] A. M. Sirunyan *et al.* [CMS], JHEP **04** (2019), 112 [arXiv:1810.11854 [hep-ex]].

[123] M. Cacciari, G. P. Salam and G. Soyez, Eur. Phys. J. C **72** (2012), 1896 [arXiv:1111.6097 [hep-ph]].

[124] M. Cacciari, G. P. Salam and G. Soyez, JHEP **04** (2008), 005 [arXiv:0802.1188 [hep-ph]].

[125] B. Tannenwald, C. Neu, A. Li, G. Buehlmann, A. Cuddeback, L. Hatfield, R. Parvatam and C. Thompson, [arXiv:2009.06754 [hep-ph]].

[126] J. K. Behr, D. Bortoletto, J. A. Frost, N. P. Hartland, C. Issever and J. Rojo, Eur. Phys. J. C **76** (2016) no.7, 386 [arXiv:1512.08928 [hep-ph]].

[127] J. Amacker, W. Balunas, L. Beresford, D. Bortoletto, J. Frost, C. Issever, J. Liu, J. McKee, A. Micheli and S. Paredes Saenz, *et al.* JHEP **12** (2020), 115 [arXiv:2004.04240 [hep-ph]].

[128] V. Ravindran, J. Smith and W. L. van Neerven, Nucl. Phys. B **665** (2003), 325-366 [arXiv:hep-ph/0302135 [hep-ph]].

[129] R. V. Harlander and W. B. Kilgore, Phys. Rev. Lett. **88** (2002), 201801 [arXiv:hep-ph/0201206 [hep-ph]].

[130] N. Greiner, A. Guffanti, T. Reiter and J. Reuter, Phys. Rev. Lett. **107** (2011), 102002 [arXiv:1105.3624 [hep-ph]].

[131] T. Binoth *et al.* [SM and NLO Multileg Working Group], [arXiv:1003.1241 [hep-ph]].

[132] F. Febres Cordero, L. Reina and D. Wackeroth, Phys. Rev. D **80** (2009), 034015 [arXiv:0906.1923 [hep-ph]].

[133] T. Lapsien, R. Kogler and J. Haller, Eur. Phys. J. C **76** (2016) no.11, 600 [arXiv:1606.04961 [hep-ph]].

[134] [ATLAS], ATL-PHYS-PUB-2016-013.

[135] [ATLAS], ATL-PHYS-PUB-2017-010.

[136] [ATLAS], ATLAS-CONF-2020-007.

[137] G. P. Salam, Eur. Phys. J. C **67** (2010), 637-686 [arXiv:0906.1833 [hep-ph]].

[138] J. Thaler and K. Van Tilburg, JHEP **03** (2011), 015 [arXiv:1011.2268 [hep-ph]].

[139] B. Bhattacherjee, C. Bose, A. Chakraborty and R. Sengupta, [arXiv:2212.11606 [hep-ph]].

[140] A. J. Larkoski, I. Moult and B. Nachman, Phys. Rept. **841** (2020), 1-63 [arXiv:1709.04464 [hep-ph]].

[141] G. Aad *et al.* [ATLAS], Phys. Lett. B **800** (2020), 135103 [arXiv:1906.02025 [hep-ex]].

[142] A. M. Sirunyan *et al.* [CMS], Phys. Lett. B **781** (2018), 244-269 [arXiv:1710.04960 [hep-ex]].

[143] M. Aaboud *et al.* [ATLAS], JHEP **01** (2019), 030 [arXiv:1804.06174 [hep-ex]].

[144] V. Khachatryan *et al.* [CMS], Eur. Phys. J. C **76** (2016) no.7, 371 [arXiv:1602.08762 [hep-ex]].

[145] R. D. Ball *et al.* [NNPDF], JHEP **04** (2015), 040 [arXiv:1410.8849 [hep-ph]].

[146] Z. Kunszt, S. Moretti and W. J. Stirling, Z. Phys. C **74** (1997), 479-491 [arXiv:hep-ph/9611397 [hep-ph]].

[147] F. Maltoni, Z. Sullivan and S. Willenbrock, Phys. Rev. D **67** (2003), 093005 [arXiv:hep-ph/0301033 [hep-ph]].

[148] [CMS], CMS-PAS-HIG-19-008.

[149] A. M. Sirunyan *et al.* [CMS], Phys. Rev. D **99** (2019) no.9, 092005 [arXiv:1811.09696 [hep-ex]].

[150] J. Chang, K. Cheung, J. S. Lee and C. T. Lu, JHEP **05** (2014), 062 [arXiv:1403.2053 [hep-ph]].

[151] O. Mattelaer and K. Ostrolenk, Eur. Phys. J. C **81** (2021) no.5, 435 [arXiv:2102.00773 [hep-ph]].

[152] A. Kobakhidze, L. Wu and J. Yue, JHEP **10** (2014), 100 [arXiv:1406.1961 [hep-ph]].

[153] M. Farina, C. Grojean, F. Maltoni, E. Salvioni and A. Thamm, JHEP **05** (2013), 022 [arXiv:1211.3736 [hep-ph]].

[154] E. Accomando, C. Byers, D. Englert, J. Hays and S. Moretti, Phys. Rev. D **105** (2022) no.11, 115004.

[155] P. M. Ferreira, R. Guedes, M. O. P. Sampaio and R. Santos, JHEP **12** (2014), 067 [arXiv:1409.6723 [hep-ph]].

[156] M. Baak *et al.* [Gfitter], Eur. Phys. J. C **72** (2012), 2003 [arXiv:1107.0975 [hep-ph]].

[157] P. Bechtle, D. Dercks, S. Heinemeyer, T. Klingl, T. Stefaniak, G. Weiglein and J. Wittbrodt, Eur. Phys. J. C **80** (2020) no.12, 1211 [arXiv:2006.06007 [hep-ph]].

[158] P. Bechtle, S. Heinemeyer, T. Klingl, T. Stefaniak, G. Weiglein and J. Wittbrodt, Eur. Phys. J. C **81** (2021) no.2, 145 [arXiv:2012.09197 [hep-ph]].

[159] P. Bechtle, S. Heinemeyer, O. Stål, T. Stefaniak and G. Weiglein, JHEP **11** (2014), 039 [arXiv:1403.1582 [hep-ph]].

[160] V. Maurer, Comput. Phys. Commun. **198** (2016), 195-215 [arXiv:1503.01073 [cs.MS]].

[161] C. Degrande, C. Duhr, B. Fuks, D. Grellscheid, O. Mattelaer and T. Reiter, Comput. Phys. Commun. **183** (2012), 1201-1214 [arXiv:1108.2040 [hep-ph]].

[162] A. Alloul, N. D. Christensen, C. Degrande, C. Duhr and B. Fuks, Comput. Phys. Commun. **185** (2014), 2250-2300 [arXiv:1310.1921 [hep-ph]].

[163] C. Durh, and M. Herquet and C. Degrande, FeynRules (2018).

[164] C. S. Deans [NNPDF], [arXiv:1304.2781 [hep-ph]].

[165] G. Aad *et al.* [ATLAS], Phys. Rev. D **101** (2020) no.1, 012002 [arXiv:1909.02845 [hep-ex]].

[166] A. M. Sirunyan *et al.* [CMS], Eur. Phys. J. C **79** (2019) no.5, 421 [arXiv:1809.10733 [hep-ex]].

[167] X. F. Han and H. X. Wang, Chin. Phys. C **44** (2020) no.7, 073101 [arXiv:2003.06170 [hep-ph]].

[168] A. Abdesselam *et al.* [Belle], [arXiv:1608.02344 [hep-ex]].

[169] P. Sanyal, Eur. Phys. J. C **79** (2019) no.11, 913 [arXiv:1906.02520 [hep-ph]].

[170] H. Bahl, T. Biekötter, S. Heinemeyer, C. Li, S. Paasch, G. Weiglein and J. Wittbrodt, Comput. Phys. Commun. **291** (2023), 108803 [arXiv:2210.09332 [hep-ph]].

[171] J. Bielčíková, R. Kunnawalkam Elayavalli, G. Ponimatkin, J. H. Putschke and J. Sivic, JINST **16** (2021) no.03, P03017 [arXiv:2005.01842 [hep-ph]].

[172] G. Kasieczka, T. Plehn, M. Russell and T. Schell, JHEP **05** (2017), 006 [arXiv:1701.08784 [hep-ph]].

[173] G. Cerro, S. Dasmahapatra, H. A. Day-Hall, B. Ford, S. Moretti and C. H. Shepherd-Themistocleous, JHEP **02** (2022), 165 [arXiv:2104.01972 [hep-ph]].

[174] J. Brehmer, K. Cranmer, G. Louppe and J. Pavez, Phys. Rev. Lett. **121** (2018) no.11, 111801 [arXiv:1805.00013 [hep-ph]].

[175] J. Brehmer, K. Cranmer, G. Louppe and J. Pavez, Phys. Rev. D **98** (2018) no.5, 052004 [arXiv:1805.00020 [hep-ph]].

[176] M. Romão Crispim, N. F. Castro, R. Pedro and T. Vale, Phys. Rev. D **101** (2020) no.3, 035042 [arXiv:1912.04220 [hep-ph]].

[177] A. Schwartzman, M. Kagan, L. Mackey, B. Nachman and L. De Oliveira, J. Phys. Conf. Ser. **762** (2016) no.1, 012035.

[178] J. Lin, M. Freytsis, I. Moult and B. Nachman, JHEP **10** (2018), 101 [arXiv:1807.10768 [hep-ph]].

[179] J. Cogan, M. Kagan, E. Strauss and A. Schwartzman, JHEP **02** (2015), 118 [arXiv:1407.5675 [hep-ph]].

[180] L. de Oliveira, M. Kagan, L. Mackey, B. Nachman and A. Schwartzman, JHEP **07** (2016), 069 [arXiv:1511.05190 [hep-ph]].

[181] J. Barnard, E. N. Dawe, M. J. Dolan and N. Rajcic, Phys. Rev. D **95** (2017) no.1, 014018 [arXiv:1609.00607 [hep-ph]].

[182] T. Mitchell, ISBN 978-0-07-042807-2.

[183] D.R. Cox, Journal of the royal statistical society series b-methodological, **20** (1958).

[184] D.H. Maulud, A.M. Abdulazeez, Journal of Applied Science and Technology Trends, **1** (2020).

[185] Y. Izza, A. Ignatiev, J.M. Silva, [arXiv:2010.11034 [cs.LG]].

[186] C. Cortes and V. Vapnik, Machine Learning, Springer, **20** (1995).

[187] Z. Zhang, Annals of Translational Medicine, **4** (2016).

[188] Vikramkumar, V. B and Trilochan, [arXiv:1404.0933 [cs.LG]].

[189] T.K. Ho, Proceedings of 3rd international conference on document analysis and recognition, **1** (1995).

[190] R.E. Schapire and Y. Freund, The MIT Press, (2012).

[191] Y. Zhao and X. Zhao, Journal of Physics: Conference Series, **1873** (2021).

[192] D. Müllner, [arXiv:1109.2378 [stat.ML]], (2011).

[193] Y. LeCun, Y. Bengio and G. Hinton, 521(7553), pp.436-444.