Inclusive nonresonant multilepton probes of new phenomena

A thesis
Submitted in partial fulfillment of the requirements of the degree of
Doctor of Philosophy

by
Angira Rastogi
Registration ID - 20152034

Department of Physics
Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road
Pashan, Pune - 411008, INDIA.

March 2022

Supervisor: Dr. Sourabh Dube

“We cannot solve our problems with the same thinking we used when we created them.”
Albert Einstein

Certificate

Certified that the work incorporated in the thesis entitled "Inclusive nonresonant multilepton probes of new phenomena" submitted by Angira Rastogi was carried out by the candidate, under my supervision. The work presented here or any part of it has not been included in any other thesis submitted previously for the award of any degree or diploma from any other university or institution.

Date: March 21st, 2022

Prof. Sourabh Dube
(Supervisor)

Declaration

I declare that this written submission represents my ideas in my own words and where others ideas have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Date: March 21st, 2022

Angira Rastogi
(20152034)

Acknowledgements

This thesis is a significant milestone in my evolution to being a particle physicist, starting six years ago. However, this would not have been possible without the presence of some extraordinary people in my life and their cumulative efforts in shaping me to my current being. Throughout my life, I have heard from my closed ones that I am not so great with words, when it comes to expressing my love and gratitude for them. So, here is an attempt at acknowledging them all to the best of my abilities!

First and foremost, I would like to thank my supervisor, Prof. Sourabh Dube, without whom I would not be anywhere close to where I am today. I joined IISER in 2015 with huge dreams but with little knowledge about how to achieve them. I was always curious about how the universe works, and wanted to learn more for my own understanding and then passing it on to the world. Sourabh has helped me fulfill this life-long dream of mine, and turning it to hopefully a permanent reality. He has been the most amazing person to learn from and work with through his unique style. You have believed in my abilities and pushed my boundaries constantly, be it research or outreach, even when I wasn't very confident. So many times you were just a sounding board for me to bounce off ideas from, and together we have accomplished so many great achievements. You have worked hard day after day on me and with me; fought for me and with me; and I cherish all those moments equally. You knew when to appreciate me and when to be harsh, which has always kept me grounded. I had never expected that I would find such an advisor with the real freedom of expressing thoughts, putting forth agreements or disagreements about work, important academic decisions, and life in general. You have been a true friend and a guardian at a place so far from home, and I am indebted to you for that forever.

I cannot help myself but get flashbacks from my journey that imprinted a great mark in my development. I want to thank all my collaborators, Prof. Sunil Somalwar, Dr. Halil Saka, Dr. Maximilian Heindl, Dr. Olena Karacheban, Dr. Sezen Sekmen, Prof. Petar Maksimovic, and Jay Vora, for all the insightful discussions, exchange of innovative ideas, and fun gossips with important life lessons. I want to especially thank Sunil, Halil, Maxi, and of course Sourabh, for broadening my vision about what a multilepton analysis can be, and how powerful tool it is. I also thank my analysis review committee chair, Prof. Ritchie Patterson, and all the conveners

from CMS for shaping this multilepton analysis into such an impressive result. I thank Sezen and Petar from the days I was learning to be a computing wizard through CMS fast simulation, and understanding how useful it is for the future of particle physics. I especially thank Sezen for making me a part of her celebrations once at her home, bonding with other collaborators from tracking and simulation group in CMS, over the amazing turkish dinner and swiss wine! I thank my research advisory committee members, Prof. Lea Caminada and Prof. Sunil Mukhi, for nudging me in the right direction and giving me an outside perspective of the field for my betterment. I thank Lea for her amazing hospitality at PSI Zurich, by arranging a visit and a seminar talk for me. She showed me the disassembled CMS pixel tracker, and the behind-the-scenes of upgrades of the detector, which gave me the tabletop experience of being an experimentalist in HEP. I would also like to thank Prof. Seema Sharma, Prof. Arun Thalapillil, Dr. Diptimoy Ghosh, and Prof. Mukhi for all the invaluable discussions at various times in IISER, often in front of our offices' corridor. This has helped me in my academic as well as personal growth.

I thank my former lab-mates, Dr. Kunal Kothekar, Dr. Shubhanshu Chauhan, Dr. Anshul Kapoor, Sai Neha Santpur, Divya Gadkari, Steenu Johnson, and Vipul Pawar. I may have physically parted ways with them a few years ago, but they will always have a special place in my heart for sharing their skills, knowledge, and experiences with me. Kunal and Shubhanshu have also extended their selfless support to me at various different occasions, hearing me out when I was at my brightest, but also when I needed to vent my frustration about the struggles of PhD and life. I thank all my current lab-mates: Arnab Laha, Prachurjya Hazarika, Kumar Yash, Shubham Pandey, Bhumika Kansal, and Alpana Sirohi, who have been a delight to be around. With Arnab, Prachu, and Yash, I was lucky enough to relive my journey to a particle physicist and learn even more through their mentoring. Arnab has been both, a great colleague embarking good physics discussions, and a very close companion whenever I needed one. Your passion towards your work and committment to personal life inspires me everyday. I also had the good fortune of working and interacting with so many undergraduate students from IISER and the summer interns in our group. I am thankful for their trust and patience in me during the learning process, and for always treating me with utmost respect and admiration. In their success, I feel immense joy to have contributed in some way. I wish you all a very good luck from the bottom of my heart to achieve what you aspire to be.

I want to thank my friend circle at IISER – Sunny, Shruti, Deepak, Naveen, Shailendra, Avisikta, and Bhumika. They have been my closest allies in both good and bad times. Without the uncountable fun excursion days and nights full of deep conversations, my journey at IISER would have been empty and incomplete. My lawn tennis group, especially the faculties Vaidhya,

Harinath, and Arjun, who let me make them run across the court and smash on them without being offended! They provided me the much needed physical workout in my otherwise sedentary lifestyle, and made some days even brighter and full of excitement. I also want to take a moment and thank my childhood friends, my best friend Saumya, Tanmay, Utkarsh, my dearest sister Divisha, who have been there by my side every single day when I wanted to escape all the harsh realities of adulthood. They have been my backbone, my conscience, my critics, and my cheerleaders at each step. My friends from undergraduate days, Arshia, Vinee, and Devina, with whom I had the chance of still being connected even after so many years through my work! It was a great pleasure meeting them abroad during my visits to CERN, and create nuisance on the international grounds. I am very lucky to have them all for their unique presence and contribution in my life.

Saving the best for the last, my parents, who have loved me unconditionally. They have believed in me every single day, through my transitions as their beloved kid to an aspiring scientist, celebrated my every success story, big or small, and didn't let my tough times tear me apart even when no words were communicated. Not only have I learned a lot from them, they are the reason of my sanity and my courage. Their constant pride has motivated me to thrive to do more and more every single day in my capacity. More importantly, it has also led me to set an example for my younger brother, Akshar, to show one can achieve anything with sheer discipline and a lot of passion. Akshar, 19, has kept the little child in me alive, allowing me to be my silly and goofy self by finding joys of life in even the smallest and nonsensical things. To the flip side, this relationship has also kept me grounded, taught me the sense of responsibility from time to time in so many different ways, and also how much I want to make every soul around me happy and comfortable all the time.

This thesis is dedicated to my family.

Abstract

The Standard Model (SM) of particle physics is a very successful breakthrough in our current understanding of the universe. It is a framework built on quantum field theory approach, unifying the three out of the four fundamental forces of nature: the strong nuclear, the weak nuclear, and the electromagnetic interactions. Within the SM, the matter is composed of fundamental fermions whereas the interactions between them are mediated via force-carrier fundamental bosons. The SM has provided strong reasoning of how the universe evolved ever since the Big Bang. It has also paved the way for future discoveries by predicting the existence of new particles, such as the Higgs boson which was the latest and the final addition to the SM particles' group. However, it is just a little short of being a complete theory of the universe.

The missing explanations behind the matter-antimatter asymmetry, the origin and smallness of the neutrino masses, the observed flavor anomalies in the b-hadron decays, and the particle nature of dark matter are a few shortcomings of the SM. Aside from these, there are also additional questions such as 'why just three generations of matter, and the mass hierarchy among them?', 'why do neutrinos occur as a left-handed singlet, but charged leptons do not?', 'why does the Higgs boson have a mass of 125 GeV, despite the radiative corrections from all the particles it couples to?', and 'why is gravity not part of the unified quantum field formulation, and is so many orders of magnitude weaker than other fundamental forces?'. There are many proposed theories of beyond-the-Standard Model (BSM) phenomena which attempts to address one or more of these open questions of the universe. This is done with the help of new hypothesized particles interacting with the SM particles. Thus, any unusual signature beyond the expectations of SM in particle physics experiments, consistent with the predictions of a particular theory, can provide strong evidence in its support.

The Large Hadron Collider (LHC) at CERN is the world's largest and highest energy particle accelerator, carrying out proton-proton or proton-lead or lead-lead collisions. Through these high energy collisions, a state equivalent to the universe just after the Big Bang is created for a very brief period of time. During this time, new particles are created from the plasma of quarks and gluons which then decay instantly to SM particles. Dedicated physics searches are designed to perform measurements of the SM free parameters like mass and decay widths of particles, coupling co-

efficients, and so on. Searches for BSM phenomena are conducted by utilizing the kinematic regions where the theory will manifest itself. If the proposed theory of BSM phenomena holds, then the hope is that with enough data, new hypothesized particles predicted by the theory will also be created similarly to the SM particles, provided the energy scales are the same. Through their unique topogical signatures, one can confirm the theory behind that BSM phenomena.

In this thesis, I have presented an inclusive search for new phenomena in the nonresonant multilepton final states. The search targets three different models: vector-like lepton in the doublet and singlet scenario, type-III seesaw mechanism, and scalar leptoquarks with the top-philic couplings. These three models target different open questions of the SM, such as the existence of vector-like leptons may provide a dark matter candidate and also account for the mass hierarchy between the different generations of matter particles in the SM, the origin and smallness of the neutrino masses can be explained by the production of heavy seesaw fermions, and scalar leptoquarks could provide an explanation for the observed b-anomalies. The primary reason behind this particular selection of BSM phenomena is that they are generators of complementary nonresonant multilepton signatures.

The search is designed with multiple electrons, muons, and hadronically decaying tau leptons, utilizing the combined proton-proton collisions data set collected by the CMS experiment at the LHC between 2016–2018, corresponding to an integrated luminosity of $\mathcal{L} = 138$ fb^{-1}. With a total of seven orthogonal final states covering almost the entire multilepton landscape, this analysis is a benchmark result with a huge sensitivity for a variety of BSM signals. The model-dependent part of the analysis employs the boosted decision trees algorithm to enhance the sensitivity to the probed BSM scenarios. No significant deviations from the background expectations are observed. Lower limits are set at 95% confidence level on the mass of the vector-like τ lepton in the doublet and singlet extensions of the SM, and are excluded for masses below 1045 GeV and in the mass range 125–150 GeV, respectively. Type-III seesaw heavy fermions are excluded in the mass range 845–1065 GeV for various decay branching fraction combinations to SM leptons. Scalar leptoquarks decaying exclusively to a top quark and a lepton are excluded for masses below 1.12–1.42 TeV, depending on the lepton flavor. For the vector-like lepton doublet as well as the type-III seesaw model, these constraints are the most stringent to date. For the vector-like lepton singlet model, these are the first constraints from the LHC experiments.

To ensure the longevity of this multilepton analysis, a model-independent component based purely on the expected SM predictions and observations is also designed, allowing the results to be reinterpretable for other BSM theories. Detailed results are also provided to facilitate these alternative theoretical interpretations.

Table of Contents

List of Tables

List of Figures

Let’s begin...

Chapter 1

Introduction

There are four fundamental forces which governs all the interactions in the universe, and has resulted in the creation of the world we live in. These are – the strong nuclear, the weak nuclear, the electromagnetic, and the gravitational forces, in the decreasing order of relative strengths. Together, these forces have constituted less than 5% [1, 2] of the total mass-energy content of the universe, in the form of ordinary matter and energy. This means that more than 95% of the universe is an uncharted territory of which humans have very little or no understanding.

The four forces are acting continuously upon us, whether we acknowledge it or not. The closest and most relatable example is that of gravity, which keeps us from falling off the Earth, or the planets revolving around the Sun forming the solar system, or even the light from escaping the black holes! Like cells are the smallest unit of life, atoms are the basic building blocks of matter. An atom is a million times smaller than the thickness of a human hair and for many years it was thought to be fundamental. But nuclear physics experiments, in the late 19^{th} and early 20^{th} century, unveiled atomic substructure and established the existence of even smaller constituents - electrons, protons, and neutrons. These fundamental particles bind together through the electromagnetic force to form visible matter. Furthermore, through deep inelastic scattering experiments between high energy electrons and protons/neutrons revealed the substructure of the latter. Protons and neutrons are hadrons and are composed of quarks and a sea of gluons bound together via the strong force, which is in fact the strongest interaction in nature at subatomic length scales. The final force, which is the weak interaction, is also very crucial for sustaining life on Earth. It powers our Sun through the thermonuclear process taking place at its core.

The Standard Model (SM) of elementary particles is a pillar of scientific efforts to harmonize the three out of the four fundamental forces, excluding gravity, of the nature in a common theoretical framework (See Ref. [3] for a review). This has helped in understanding the laws that govern

the various fundamental interactions and evolution of the universe to its modern form since the Big Bang explosion. In the SM, there are three generations of fermions as matter particles, each composed of a charged lepton (electron, muon, and tau), a lepton neutrino and a pair of quarks (up and down; charm and strange; top and bottom). Thus, there are six leptons and six quarks; each of these twelve particles has its own antiparticle. In addition, there are in total five force-carrier bosons ($W^{\pm}$, Z, γ, g) of spin=1, corresponding to the three forces of nature. Lastly, the scalar higgs boson interacts through its field with all the particles and assigns them the masses that they have. However, the higgs boson does not interact with neutrinos and thus neutrinos are massless in the SM.

The SM is a very successful milestone in our current understanding of the various phenomena happening in our universe. But, the complete framework of the SM, that we know as of today, has evolved over the past 50 years of experiments, standing strong through innumerable rigourous tests of the various assumptions of the model. From successfully describing the electron magnetic dipole moment up to a precision of 11 significant figures, to predicting the existence of new particles, the SM has continued to hold its ground in the field of particle physics when it comes to describing the laws of the universe. However, it is still not the full picture.

The evidence of neutrino oscillations from the Super Kamiokande experiment [4] established that at least two out of the three neutrinos must have a very tiny non-zero mass. The mechanism by which the SM neutrinos attain their masses, as well as their smallness, is a very compelling mystery. The other 27% of the mass-energy content of the universe is in the form of dark matter, for which there are indirect cosmological evidences from the rotational curves of the galaxies or gravitational lensing, but no direct explanation or a potential candidate in the SM. And, the remaining 68% of the universe is collectively termed as dark energy, for which there are neither direct nor indirect inferences about its origin and properties. It could be a completely new, fifth fundamental force of the nature, with a very suppressed interaction with the SM particles. Interestingly, the SM also doesn't incorporate the fourth fundamental interaction, the gravitational force, in its quantum field theory framework, and therefore cannot explain why gravity is $\mathcal{O}(10^{38})$ times weaker than the strong force. Aside from these, there are also incomplete and unsatisfactory explanations behind the matter-antimatter asymmetry of the universe, the fact that there are only three generations of matter fermions and the mass hierarchy among them, the recent observations of the anomalous magentic moment of the muon [5–11] and flavor anomalies in the b-hadron decay violating the argument for lepton flavor universality [12–19], and the fine-tuning of the higgs mass [20].

Science is driven by curiosity and scientific curiosity comes from the desire to learn the facts of nature. We can understand the natural phenomena by building a model based on certain as-

sumptions, and test for the validity of these theories in dedicated experiements. These models can describe the phenomena as well as predict what lies ahead. To this end, there are many proposed theories beyond-the-SM (BSM) which attempts to provide explanations to the open questions of the SM. These theories hypothesize new particles, which ultimately decay to SM particles, creating unique topological signatures for a clear detection. Hence, by producing these new proposed particles in favorable environments of high energy collisions, the evidence for theory for a given BSM phenomena can be established with the help of physics searches designed to probe that model.

The Large Hadron Collider (LHC) [21, 22], at the CERN, is the European Laboratory for Particle Physics research, outside Geneva in Switzerland. It is the world's largest circular and highest energy particle accelerator colliding proton-proton beams for the maximum part of its operation, followed by shorter periods of proton-lead and lead-lead collisions. The center-of-mass energy of the collisions reaches up to 13 TeV with the proton beams, which is the current highest record for a hadron collider. Hence, the LHC is a very powerful machine for fundamental physics research, breaking new grounds for the BSM phenomena. Along the circumference, there are four points at which the two particle beams are made to collide head-on with each other. The debris from the collision is captured by state-of-the-art particle detectors – ATLAS, CMS, ALICE, and LHCb. The ATLAS and CMS experiments are general purpose detectors to explore a wide spectrum of physics at an unprecedented precision. The primary goal is to understand the electroweak symmetry breaking mechanism via the higgs boson by performing precision measurements of the SM particles' properties, and also to explore new theories beyond the SM. Through the partons of the protons in the colliding beam, the LHC is simultaneously a Z-boson factory, a W-boson factory, a b-quark factory, a top-quark factory, a higgs boson factory, and will also produce any new particles with $\mathcal{O}(100)$ GeV mass.

As a particle physicist and a member of the CMS collaboration, I am looking for evidence of new physics that lies beyond the SM. I'm trying to discover answers to some big mysteries of the universe like hierarchy in the masses of the fundamental matter particles, origin and smallness of the neutrino masses, and finally the observed flavor anomalies in the b-hadron decays. A good discovery relies on the impeccable reproducibility of the detection and a good detection method is often associated with an elegant yet novel technique. In my research, I use leptons as a basic tool and construct a multilepton search strategy which has the advantage of having a high signal to noise ratio. The multilepton final state is a good probe because of the clean signature in the detector and less contamination from the known SM processes. Three scenarios of BSM phenomena are probed through these multilepton final states – the vector-like lepton model, the type-III seesaw mechanism, and the third generation of scalar leptoquarks. These three models target different open

questions of the SM, such as the vector-like lepton model may provide a dark matter candidate and also account for the mass hierarchy between the different generations of matter particles in the SM, the smallness of the neutrino masses can be explained by the production of heavy seesaw fermions, and scalar leptoquarks can directly provide an explanation for the observed b-anomalies. The primary reason behind this particular selection of BSM phenomena is that they are generators of complementary nonresonant multilepton signatures.

An inclusive nonresonant multilepton search is designed with three or more electrons, muons, and hadronically decaying tau leptons in the final state, utilizing the combined proton-proton collisions data set collected by the CMS experiment at the LHC between 2016–2018, corresponding to an integrated luminosity of $\mathcal{L} = 138$ fb^{-1}. With a total of seven orthogonal final states covering almost the entire multilepton landscape, this analysis is a benchmark result with a huge sensitivity for a variety of BSM signals. The model-dependent part of the analysis employs the boosted decision trees algorithm to enhance the sensitivity to the probed BSM scenarios. To ensure the longevity of this multilepton analysis, a model-independent component based purely on the expected SM predictions and observations is also performed, allowing the results to be reinterpretable for other BSM theories. Detailed results are also provided to facilitate these alternative theoretical interpretations.

The rest of the thesis is organized as follows. Chapter 2 describes the SM and its inadequacies in detail, followed by three scenarios of BSM phenomena to explain the shortcomings of the SM. Chapter 3 presents the motivation and strategy used for designing this multilepton analysis, and Chapter 4 describes the experimental setup (CMS, LHC) used for performing this analysis. In Chapter 5, the generation and simulation of the SM phenomena, that goes hand in hand with the analysis of collected data, and then the most essential element of reconstructing the physics objects is outlined. Chapter 6 describes the SM background estimation techniques for this multilepton analysis. Chapter 7 summarizes the model-specific search using the boosted decision algorithm, and Chapter 8 describes the model-independent component of this multilepton analysis. Finally, Chapter 9 presents a roadmap for reinterpretation of the multilepton results for any other BSM phenomena, not probed in this analysis.

Understanding the universe around us...

Chapter 2

The Standard Model and Beyond

The field of particle physics has had a tremendous winning streak for a century or so, where we have come an enormously long way in trying to build an understanding of the laws of the universe, through the Standard Model (SM) of elementary particles. But, despite the glorious successes of over the past 50 years, we still don't have a theory of everything that we see around us.

Nevertheless, it is fair to say that the golden age of particle physics experiments is taking place right now. Not only have we recently discovered the Higgs boson, and are busy in checking that it conforms to the predictions of the SM, we have strong indications that there should be new physics beyond the SM, and the LHC and other experiments are comprehensively searching for it in every corner of the kinematic phase space that is accessible. So far, no new phenomena has been concretely found, but the searches are continuously going on, and the LHC is also being upgraded to run at even higher energies to expand its physics reach.

Let us begin with understanding what we know of first – the SM.

2.1 The Standard Model

The Standard Model (SM) is the name given to a theory of fundamental particles and how they interact via the strong nuclear interaction, the weak nuclear interaction and the electromagnetic interaction, back in the 1970s. It incorporated all that was known about subatomic particles at the time and predicted the existence of additional particles as well. As of today, there are a total of 61 particles and anti-particles in the SM particles' group, divided into two main families: fermions and bosons. Fermions, classified into two types – quarks and leptons, are the fundamental particles (with a corresponding antiparticle) that are the building blocks of the matter. On the other hand, bosons are the mediators of the interactions.

Every elementary particle in the SM is characterized by a few quantum numbers which are conserved in the fundamental interactions. These are unique invariant masses, an electric charge (in the units of e), and a spin quantum number, which is equal to half integral ($\frac{1}{2}$,$\frac{3}{2}$,$\frac{5}{2}$,etc.) for fermions, whereas a whole integer (0,1,2,etc.) for bosons. The modern-day visualization of the SM, where all the fundamental particles are strategically placed, according to their designated roles in the nature, is shown in Figure 2.1.

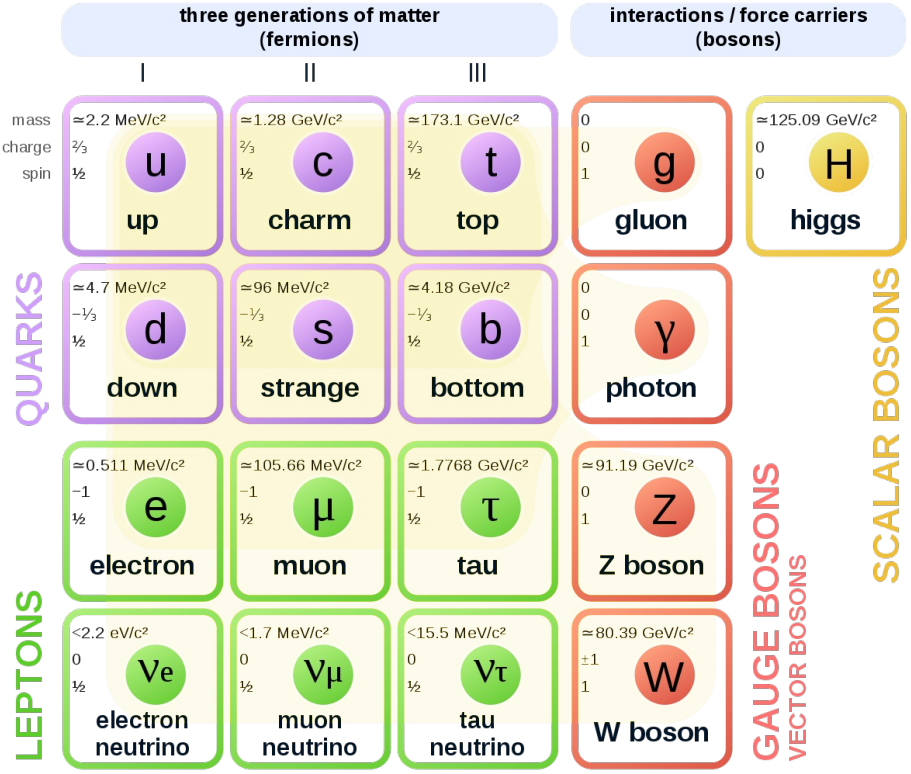

Figure 2.1: The Standard Model of elementary particles. (*Image Courtesy: Wikipedia*)

In Figure 2.1, the first, second, and third column under the block of matter particles refer to the first-, second-, and third-generation of fermions. All the properties of fermions across the three generations are same, except for the increasing masses. The first generation is mostly responsible for creating all the ordinary matter of the universe, while interesting behavior starts to emerge as we move up the generations, for e.g. particles with longer lifetimes (b quark, τ lepton) or the unique ability of the top quark to decay to lighter particles, instead of hadronizing to form color neutral baryons.

There are two types of bosons in the SM: vector bosons ($W^{\pm}$,Z,γ) with spin = 1 and a scalar

Higgs boson (H) with spin = 0. The vector bosons are the force-carrier particles of the fundamental interactions, viz. γ for electromagnetic interaction, eight types of gluons (g) for the strong nuclear interaction, and $W^{\pm}$ or Z boson for the weak nuclear interaction. Last but not the least, the Higgs boson is responsible for generating masses to all the SM particles (including itself). This is known as the Higgs mechanism, as described in Section 2.4.1.

Quarks interact with each other to form bound states which results in composite particles, known as hadrons. Hadrons are classified into two types: baryons (made up of three quarks) and mesons (made up of quark-antiquark pair). There are also recent observations of exotic baryons, i.e. tetraquarks ($qq\bar{q}\bar{q}$) [23–25] and pentaquarks ($qqqq\bar{q}$) [26, 27] bound states. Due to the additive property of spin quantum numbers, baryons (for e.g. protons, neutrons) are also fermions while mesons (for e.g. pions) are bosons. No such property is exhibited by leptons.

The mathematical foundation of the SM is based on quantum field theories (QFTs) which describe the fundamental interactions. This is explained in the subsequent sections. In a QFT, particles are treated as excited states (or quanta) of their underlying quantum fields. The interactions between SM particles are described by a Lagrangian involving the corresponding quantum fields. The SM Lagrangian can be described as a sum of Lagrangians for the three interactions:

$$\mathcal{L}_{SM} = \mathcal{L}_{QED} + \mathcal{L}_{QCD} + \mathcal{L}_{Weak} \tag{2.1}$$

Most of the discussion follows from Ref. [28]. Throughout this thesis, natural units ($\hbar = c = 1$) are used.

2.2 The Electromagnetic interaction and QED

Quantum Electrodynamics (QED) is an Abelian U(1) gauge theory. It is the relativistic analogue of classical electrodynamics, and describes how light, i.e. photon and matter, i.e. electrically-charged fermions interact with each other. It is the first theory where full agreement between quantum mechanics and special relativity is achieved. The most accurate predictions using QED include quantities like the anomalous magnetic moment of the electron and the Lamb shift of the energy levels of hydrogen.

The Lagrangian of the QED is given as,

$$\mathcal{L}_{QED} = \bar{\psi}(i\not{D} - m)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} \tag{2.2}$$

where the covariant derivative $D_\mu = \partial_\mu + ieA_\mu$ is used under Feynman's slash notation ($\not{a} \equiv$

$a_\mu \gamma^\mu$) and $F_{\mu\nu}$ is given in terms of the potential A_μ as $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. The spin-half Dirac field ψ (for e.g. an electron) and gauge field A_μ (photon) poses local gauge symmetry, i.e. under the operation $\psi \to e^{ie\alpha(x)}\psi$ and $A^\mu \to A^\mu - \partial^\mu \alpha$, respectively. The free parameters in Eqn. 2.2 are the mass of the electron, m, and the electron charge, e. The term "$A_\mu A^\mu$" which would describe the mass of the photons is forbidden by gauge invariance, thus making the photons massless.

2.3 The Strong Nuclear interaction and QCD

Quantum Chromodynamics (QCD) is a non-Abelian SU(3) gauge theory, which describes the strong nuclear interaction. There are eight mediator particles of QCD, the gluons. Gluons couple to quarks and transform under the 3-dimensional representation of SU(3). The three different values for the gluon index are labelled by the three primary colors: red (r), green (g), and blue (b). Different generations of quarks (or different flavors) also transform as color triplets. Hence, quarks and gluons exist in composite color-neutral states only in the "low" temperature realm below 10^{12} K.

The Lagrangian for the QCD is given as,

$$\mathcal{L}_{QCD} = -\frac{1}{4} G^a_{\mu\nu} G^{a\mu\nu} + \sum_{f\in\{u,d,c,s,t,b\}} \bar{\psi}\left(i\partial\!\!\!/ - g_s A\!\!\!/^a \frac{\lambda^a}{2} - m_f \right)\psi \tag{2.3}$$

where $G^a_{\mu\nu}$ is the gluon field strength, g_s is the strong coupling constant, gauge fields A^a corresponding to eight gluons, and m_f is the mass of the quark of respective flavor f in summation. The *Gell-Mann matrices*, λ^a provide the basis for defining the triplet representation.

2.4 The Weak Nuclear interaction and QFD

Quantum flavordynamics (QFD), as the theory of weak interactions in occasionally known, is a non-Abelian SU(2) gauge theory. In QFD only the left (right) handed fermions (antifermions) take part in the interactions, and thus it violates parity.

The Lagrangian for the weak interaction is given as,

$$\mathcal{L}_{Weak} = i(\bar{\psi}_L \partial\!\!\!/ \psi_L + \bar{\psi}_R \partial\!\!\!/ \psi_R) - m(\bar{\psi}_L \psi_R + \bar{\psi}_R \psi_L) \tag{2.4}$$

Here, the $\psi_{L,R}$ represent the left- and right-handed fermions of rest mass m.

The weak interaction acts over a short range ($\mathcal{O}$(size of atomic nucleus)) and this suggests that

the corresponding gauge bosons are massive. The two-fold problems of a finite mass gauge boson field, as well as massive fermions with different symmetry representations (left- and right-handed) are addressed by the Higgs mechanism. From the β decay, the Fermi constant, i.e. $G_F \sim 10^{-5}$ GeV^{-2} implying a mass scale of $\mathcal{O}(100)$ GeV.

2.4.1 Spontaneous symmetry breaking and Higgs mechanism

To incorporate the interactions with the right-handed fermions, we can describe the U(1) interaction of SM by means of a boson B_μ. Then, the physical eigenstates (A_μ and Z_μ) are superposition of W^3_μ and B_μ as given in Eqn. 2.5. Hence, we have a complete theory for the weak interaction. In addition to this, the introduction of A_μ in the admixture implies that the electromagnetic and the weak interactions are the two manifestations of the same interaction, especially at some particular energy scale. This unification is known as the unified electroweak theory.

$$\begin{aligned} W^3_\mu &= cos\theta_W Z_\mu + sin\theta_W A_\mu \\ B_\mu &= -sin\theta_W Z_\mu + cos\theta_W A_\mu \end{aligned} \tag{2.5}$$

Here, θ_W is the *Weinberg angle*, with a value of roughly $sin^2\theta_W = 0.231$.

A complex scalar field, with the Klein-Gordan Lagrangian is given as,

$$\mathcal{L} = \partial\phi^*\partial\phi - m^2|\phi|^2 \tag{2.6}$$

The Eqn. 2.6 has global symmetry under $\phi \rightarrow e^{i\alpha}\phi$. The symmetry is also not broken if we add a term of the form $-\lambda|\phi|^4$, known as *phi-to-the-fourth* theory. The terms in the lagrangian which do not involve derivatives can be thought of as a potential of the field. Hence, the potential for the field in Eqn. 2.6 can be extracted as,

$$V(\phi) = m^2|\phi|^2 + \lambda|\phi|^4 \tag{2.7}$$

For $m^2 < 0$, the global minima of the potential would lead to $|\phi| = \sqrt{\frac{-m^2}{2\lambda}} \equiv \frac{v}{\sqrt{2}}$. This defines a circle of points in the complex ϕ-plane, where all the points are degenerate in energy and could be the minimum. However, as soon as the theory picks a point out of this circle, the global symmetry is broken. This is known as the phenomena of *spontaneous symmetry breaking*.

Introducing the scalar Higgs field as H, the Higgs potential takes the form, same as that of

Eqn. 2.7,

$$V = -\mu^2 H^\dagger H + \lambda (H^\dagger H)^2 \tag{2.8}$$

The potential is minimized at $\sqrt{H^\dagger H} \equiv \frac{v}{\sqrt{2}} = \sqrt{\frac{\mu^2}{2\lambda}}$. Hence, choosing for real-valued v:

$$\langle H \rangle = \begin{bmatrix} 0 \\ \frac{v}{\sqrt{2}} \end{bmatrix} \tag{2.9}$$

Substituting the expressions from Eqn. 2.5 together with $cos\theta_W = \frac{g}{\sqrt{g^2+g'^2}}$, and $sin\theta_W = \frac{g'}{\sqrt{g^2+g'^2}}$, and extracting the terms for the complex field $W_\mu^\pm$ ($m^2\phi^*\phi$) and the real field Z_μ ($\frac{m^2}{2}\phi^2$), we get predictions for the relations between the masses of the W and the Z boson, and we get a massless photon as follows:

$$\begin{aligned} m_W &= \frac{gv}{2} \text{ from the term } \frac{(gv)^2}{4} W_\mu^+ W^{-\mu}, \\ m_Z &= v\frac{\sqrt{g^2+g'^2}}{2} = \frac{m_W}{cos\theta_W} \text{ from the term } v^2 \frac{g^2+g'^2}{8} Z_\mu Z^\mu, \\ \text{and } m_A &= 0. \end{aligned} \tag{2.10}$$

2.5 Inadequacies of the Standard Model

Despite the elegant and coherent formalism of the three out of the four fundamental interactions of the nature under one overarching framework, the SM, there are many contradictions with the experimental observations. To discuss a few, especially those which motivated the search for new phenomena in this thesis, are as follows:

1. **Mass hierarchy among the fermion generations:** The masses of various particles and their couplings with the fields in the SM framework are free parameters of the model. They are only experimentally determined, such as from the LHC experiments. It has been established that masses of the fermions (both quarks and leptons) among the three generations are not identical, and in fact have some arbitrary hierarchy. The difference is as great as $\mathcal{O}(10^6)$ between the lightest (electron) and the heaviest (top quark) particles!

It is true that fermions have to have at least one distinguishing quantum property (mass, charge, spin or color in case of quarks) to abide by the Pauli's exclusion principle. And while there is a strong case for the values of charge, spin, and color that are allowed from theory, there is no definite calculation for the values of masses and couplings of particles. Hence, we have no explanation for the intrinsic mass hierarchy among the generations in the SM, and thus requires additional implementation.

2. **Neutrino oscillations and origin of their masses:** There was a long-standing puzzle about the mismatch in the flux of solar neutrinos reaching the Earth with the prediction from models of the nuclear reaction that fuels the Sun. This was put to rest with the observation of neutrino oscillations among different flavor eigenstates while traveling over long distances by the Super Kamiokande experiment in 1998 [4], followed by another experiment at Sudbury Neutrino Observatory between 1999–2006 [29].

   The explanation of the neutrino oscillations lies in the mixing between the flavor and mass eigenstates of neutrinos. Neutrinos are emitted and absorbed in the weak interaction in flavor eigenstates but they propagate as mass eigenstates. At all times, they are a superposition of the three mass or flavor eigenstates. When the neutrino superposition state travels through space, the quantum mechanical phases of the three neutrino mass states advance at slightly different rates, which is only possible due to the differences in their respective masses. This results in changing the superposition mixture of mass eigenstates as the neutrino travels, but a different mixture of mass eigenstates corresponds to a different mixture of flavor states. Hence, an electron neutrino produced in the fission at the Sun's core may sometimes reach the Earth as a muon or a tau neutrino! The probability of transition of a neutrino between two-flavor states is given by,

$$P_{\alpha\to\beta} = |\langle\nu_\beta(L)|\nu_\alpha\rangle|^2 = |\sum_i U^*_{\alpha i}U_{\beta i}e^{-i\frac{m_i^2 L}{2E}}|^2 \tag{2.11}$$

   where $U_{\alpha i}$ are the terms of Pontecorvo-Maki-Nakagawa-Sakata matrix [30, 31], L is the path length over which neutrino travels, m_i is the mass of the neutrino of flavor i, and E corresponds to its energy. Despite the perfectly consistent description that blends in with the observation of neutrino oscillations, the origin of neutrino masses is not defined by the SM, and hence indicating that SM is an incomplete theory of universe.

3. **Flavor anomalies in the b-hadron decays:** Recently we have gathered enough evidence for unequal branching fraction in the heavy flavor decays. The most significant result is from the

LHCb experiment [16], with the smallest statistical uncertainty in the measurement. This is a scenario of Lepton Flavor Universality Violation (LFUV) in case of heavy hadron decays, which is in direct contradiction to the very construction of the SM with identical couplings between fermions and gauge fields.

4. **Dark matter candidate:** So far LHC experiments have not demonstrated the presence of dark matter in the universe, but we have enough proof of the same from the cosmological phenomena, such as in the rotational curves of the galaxy where the spiral arms are seen to be moving at a faster velocity than what the calculation predicts, based on the total visible mass of the galaxy, indicating the presence of invisible "dark matter". Other key aspect of the dark matter detection, that pervades the universe ubiquitously, is in the gravitational lensing of the galaxies, where the light rays are deflected from their straight line path by this invisible mass, forming rings around the galaxy images.

   This overwhelming evidence of the existence of dark matter lacks a description in the SM, in fact it is also not certain if it indeed has a particle nature! If that's the case, then we have no particle in the SM that exhibits the properties of dark matter particles.

5. **Anomalous muon magnetic dipole moment:** Recent experimental results from the Muon g-2 Collaboration [10, 11] at the Fermi National Laboratory have shown that while the measured electron magnetic dipole moment matches very well with the theoretical value calculated from SM, within a precision of up to 11 significant figures, such is not the case for muons which is alike an electron, only 200 times heavier in mass. This could indicate the presence of potentially new fields or bosonic particles, giving rise to additional interactions with the heavier cousin of electron. This, in turn, would imply new interactions and radiative corrections to every other SM particles, including the mass of the Higgs boson.

Apart from the list above, there are some other discrepancies in the SM with the reality. These are the presence of more matter than anti-matter in the universe, why only three generation of fermions in the SM, why the Higgs mass came out to be so small despite the radiative corrections to its bare mass from all the particles it couples to, and missing quantum field description of gravity or accelerating expansion of the universe, which cannot be ignored at the Planck scale (10^{19} GeV). Moreover, the SM theory also hints that an extension of Glashow, Salam, and Weinberg's work should be possible i.e. at higher energies, QCD should unite with QFD in much the same way that the electromagnetism unites with the weak interaction to create QFD. Such a theory has been called the grand unified theory (GUT).

All of these mysteries cry out for extensions beyond the SM, and the next section describes three such models that addresses the various shortcomings of the SM.

2.6 Beyond the Standard Model

Many proposed theories of beyond-the-SM (BSM) extensions try to address various aforementioned shortcomings of the SM, in order to provide a more complete picture of the laws of the nature. In this thesis, I have focused on three such BSM scenarios: vector-like leptons, the type-III seesaw mechanism, and leptoquarks. There are discussed in the following sections.

2.6.1 Vector-like leptons

Vector-like fermions (or leptons) are hypothetical particles whose left- and right-handed components transform under conjugate representations of the SM gauge symmetries [32–36], and hence their masses are independent of the SM Higgs mechanism and are not constrained by electroweak precision measurements [37, 38]. They arise in a wide variety of BSM scenarios, including, but not limited to, supersymmetric models [5, 39–41], models with extra spatial dimensions [42, 43], and grand unified theories [44–46].

Extensions of the SM with one or more vector-like fermion families may provide a dark matter candidate [47–50], and account for the mass hierarchy between the different generations of particles in the SM via their mixings with the SM fermions [51–53]. Furthermore, vector-like leptons (VLLs) are also among the proposed solutions to the observed tensions between the experimental measurements and the SM prediction of the anomalous magnetic moment of the muon [5–11].

In this thesis, two distinct models are considered in which the VLLs couple to the SM τ lepton [54, 55]. The Doublet VLL model contains an SU(2) doublet (τ',ν'), where the τ' and ν' are mass-degenerate at tree level and can be produced in pairs ($pp \to \tau'^{+}\tau'^{-}$/ $\nu'\overline{\nu'}$) or in association ($pp \to \tau'\nu'$). The total production cross section for the Doublet VLL model decreases from 20 pb to 1 fb for m'_τ between 100 GeV to 1000 GeV. The decay modes are $\tau' \to Z\tau$ or $H\tau$, and $\nu' \to W\tau$, with the branching fractions of the τ' dependent on the mass m'_τ. Typically, the branching fraction of $\tau' \to Z\tau$ is 100% at low masses, which then reduces with increasing m'_τ as the decay to Higgs boson is allowed, and reaches 50% asymptotically for $m'_\tau = 1000$ GeV. An example of the complete decay chain for the associated production would be $\nu'\tau'^{\pm} \to W^{\pm}\tau^{\mp}H\tau^{\pm} \to \ell^{\pm}\nu\tau^{\mp}b\bar{b}\tau^{\pm}$ and for the pair production would be $\nu'\overline{\nu'} \to W^{\pm}\tau^{\mp}W^{\pm}\tau^{\mp} \to \ell^{\pm}\nu\tau^{\mp}\ell^{\pm}\nu\tau^{\mp}$. Thus it is possible to produce up to seven leptons in the final state.

In the Singlet VLL model, only a charged lepton (τ') is present. Hence, they can be produced only in pairs via a Z boson, with a cross section exponentially decaying from 1 pb to 0.1 fb for m'_τ between 100 GeV and 1000 GeV. The τ' can decay to either $Z\tau$ or $H\tau$, or $W\nu$, with the branching fractions similarly governed by m'_τ. At low mass the τ' brancing fraction to $W\nu$ and $Z\tau$ is 80% and 20% respectively. At approximately $m'_\tau = 1000$ GeV, the branching fraction to $W\nu$ decay mode is 50%, with the other two decay modes ($Z\tau$, $H\tau$) becoming equally probable at 25% branching fraction each.

In the Singlet VLL model, the Lagrangian is given by,

$$-\mathcal{L} = m'_\tau \tau' \bar{\tau'} + \epsilon H L \bar{\tau'} + y_\tau H L \bar{\tau} + c.c. \tag{2.12}$$

where H is the Higgs boson, $L = (\tau, \nu_\tau)$ is the lepton doublet of the third generation in SM, y_τ is Yukawa coupling with the SM τ, and ϵ is the mixing parameter of the Yukawa coupling. The charged fermion mass matrix in the gauge eigenstate basis is,

$$\mathcal{M} = \begin{bmatrix} y_\tau & \epsilon v \\ 0 & m'_\tau \end{bmatrix} \tag{2.13}$$

For the Doublet VLL model, the Lagrangian is,

$$-\mathcal{L} = m_{\tau'} L' \bar{L'} + \epsilon H L' \bar{\tau} + y_\tau H L \bar{\tau} + c.c. \tag{2.14}$$

so that the charged fermion mass matrix is,

$$\mathcal{M} = \begin{bmatrix} y_\tau & 0 \\ \epsilon v & m'_\tau \end{bmatrix} \tag{2.15}$$

If we assume Yukawa coupling ϵ to be small, then the charged lepton mass eigenstates gives a τ' of mass $M'_\tau = m'_\tau$, and the SM tau lepton with mass $M_\tau = y_\tau v$.

Figure 2.2 shows two processes from the Doublet and Singlet VLL models, which exemplify the production and decay of vector-like τ lepton pairs that result in multilepton final states. Electroweak precision data constrain the mixing angle between vector-like leptons and SM leptons to be less than about 10^{-2}, permitting prompt decays for mass values that are close to the electroweak scale [56, 57].

In this thesis, we will assume prompt decays of vector-like τ leptons; aside from this assump-

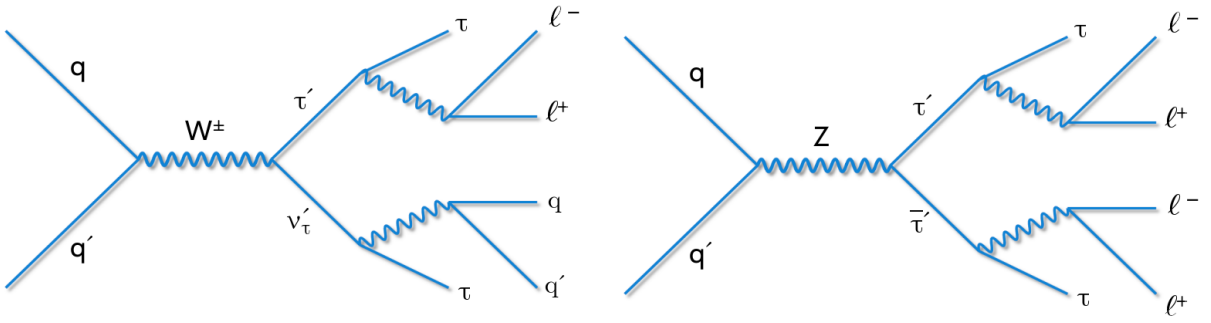

Figure 2.2: Example processes illustrating production and decay of doublet vector-like τ lepton pairs at the LHC that result in multilepton final states. The right diagram also illustrates the singlet scenario.

tion, the multilepton analysis is insensitive to the precise values of the mixing angles.

The most stringent constraints on models with vector-like τ lepton doublets are from a search conducted by the CMS Collaboration [58] with 77 fb^{-1} of data collected in 2016–2017, which excludes them in the mass range of 120–790 GeV. The search is performed with multilepton final states consisting of up to four electrons and muons, and also an additional final state with two light leptons along with one hadronically decaying τ lepton. There are, so far, no direct constraints on the vector-like τ lepton singlet model from any of the LHC experiments. The L3 Collaboration at the LEP placed a model-independent lower bound of $\sim$100 GeV on the mass of additional heavy leptons [59].

2.6.2 Type-III seesaw mechanism

The observed nonzero neutrino masses and mixing among lepton flavors can be explained by a seesaw mechanism, which introduces new heavy particles coupled to the SM leptons [60–68]. In these models, the neutrino is a Majorana particle, and the neutrino mass arises via mixing with new massive fermions. We consider the type-III seesaw model [69] in this thesis, which introduces an SU(2) triplet of heavy leptons, including Dirac charged leptons ($\Sigma^{\pm}$) and a Majorana neutral lepton (Σ^{0}). The mass relation between the neutrino and the degenerate heavy seesaw fermions is given as,

$$m_{\nu} = \frac{\Lambda^2 v^2}{M} \tag{2.16}$$

where, m_{ν} and M are the masses of the neutrino and seesaw fermions, respectively, Λ is the Yukawa coupling parameter, and v is the Higgs vev. The heavier the mass of the seesaw fermions

is, the lighter will be the mass of SM neutrinos, according to the seesaw mechanism.

At the LHC, these heavy fermions may be pair-produced through electroweak interactions in both charged-charged ($\Sigma^{\pm}\Sigma^{\mp}$) and charged-neutral ($\Sigma^{\pm}\Sigma^{0}$) modes. The total production cross section ranges between 100 pb to 0.01 fb, for m_{Σ} ranging from 100 GeV to 2000 GeV. The seesaw fermions are assumed to mix with SM leptons, and decay to a W, Z, or Higgs boson (H) and an SM lepton ($\nu,$ or $\ell = e, \mu, \tau$), such that the sum of the branching fraction of seesaw fermions to all the SM lepton flavors is always equal to unity, at all masses. The three production modes, combined with the nine possible combinations of boson-SM lepton decay yield 27 distinct signal production and decay modes. An example of the complete decay chain is $\Sigma^{\pm}\Sigma^{0} \to W^{\pm}\nu W^{\mp}\ell^{\pm} \to \ell^{\pm}\nu\nu\ell^{\mp}\nu\ell^{\pm}$.

Two diagrams exemplifying the production and decay of Σ pairs that result in multilepton final states are shown in Fig. 2.3. Electroweak and low-energy precision measurements enforce an upper limit on the mixing angles of 10^{-4} across all lepton flavors [70, 71]. This bound allows for prompt decays of heavy fermions in the mass ranges accessible to collider experiments [72–76].

In this thesis, the $\Sigma^{\pm,0}$ are assumed to be degenerate in mass and their decays are assumed to be prompt. The effects of the radiative mass splitting between the neutral and charged heavy fermions are negligible. The Σ decay branching fractions to different bosons are determined solely by their masses. The free parameters are the Σ mass, m_{Σ} and the Σ decay branching fractions to the SM lepton flavors: $\mathcal{B}_e$, $\mathcal{B}_\mu$, and $\mathcal{B}_\tau$, with the requirement that $\mathcal{B}_e + \mathcal{B}_\mu + \mathcal{B}_\tau = 1$.

The most stringent limits on the type-III seesaw model come from a search conducted by the ATLAS Collaboration using the combined LHC data set from 2016–2018 at $\sqrt{s} = 13$ TeV in multilepton final states with up to four electrons and muons [77]. The search excluded at 95% confidence level (CL) type-III seesaw fermions with masses below 910 GeV in the lepton-flavor-democratic scenario. Previous constraints in the same scenario by the CMS Collaboration from a cut-based search using a comparable data set and in similar final states excluded type-III seesaw fermions with masses below 880 GeV at 95% CL [78]. The best constraints on the type-III seesaw model in the $\mathcal{B}_\tau = 1$ scenario is set by the CMS Collaboration using the 2016 LHC data set at $\sqrt{s} = 13$ TeV in multilepton final states with up to four electrons and muons [79], excluding seesaw fermions with masses below 390 GeV.

2.6.3 Leptoquarks

Leptoquarks are color-triplet scalar or vector bosons that carry nonzero baryon and lepton quantum numbers and fractional electric charge [80]. Such particles commonly emerge in grand unified

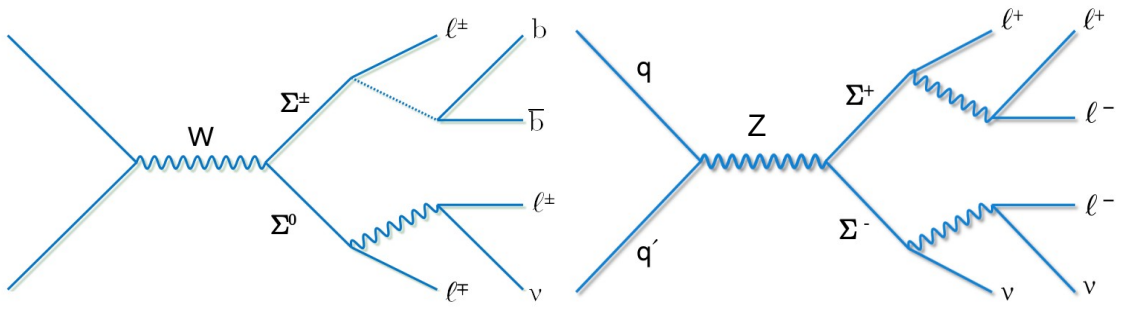

Figure 2.3: Example processes illustrating production and decay of type-III seesaw heavy fermion pairs at the LHC that result in multilepton final states.

theories, e.g., based on $SU(4)$ [81], $SU(5)$ [82], or $SO(10)$ [83] schemes, models with compositeness [84, 85], and *R*-parity violating supersymmetry models [86, 87].

In proton-proton collisions at the LHC, scalar leptoquarks (S) could be pair-produced via strong interactions, with the production cross section depending only on the leptoquark mass, m_S, but not on the unknown Yukawa coupling. Depending on the nature of the Yukawa coupling, such leptoquarks are expected to decay either to an up-type quark and a charged lepton or to a down-type quark and a neutrino, with branching fractions β and $1-\beta$, respectively. We assume that the Yukawa couplings involve only one generation of quarks or leptons. The simultaneous coupling of leptoquarks to more than one generation of quarks or leptons that are not aligned with the SM Yukawa couplings may lead to quark or lepton flavor violation [88, 89].

In this thesis, we consider scalar leptoquarks [90] with electric charge of $-1/3|e|$, and a 100% Yukawa coupling ($\beta = 1$) to the top quark and a single flavor of SM charged lepton. In a supersymmetric theory, these leptoquarks are right handed down-type squarks that couple to the top quark and charged leptons through leptonic-hadronic *R* parity violating interactions, where the down-type squarks are the scalar partners of the SM down-type quarks. We assume that only one flavor of charged lepton coupling dominates at a time, and hence consider leptoquark branching fractions $\mathcal{B}_e = 1$, $\mathcal{B}_\mu = 1$, or $\mathcal{B}_\tau = 1$, for leptoquarks decaying into a top quark and a charged lepton of the

first-, second-, or third-generation, respectively. We target the mass range from just above the top quark mass up to the TeV scale. Furthermore, the leptoquark decays are assumed to be prompt, and the coupling is assumed to satisfy $\lesssim 0.1$, within the bounds on such Yukawa couplings from leptonic Z boson decays [90, 91]. As with the vector-like lepton and type-III seesaw models, the further analysis is independent of the magnitude of the leptoquark Yukawa couplings aside from the assumption of prompt decays.

Figure 2.4 shows two processes exemplifying the production and decay of leptoquark pairs that result in multilepton final states. Leptoquarks with preferential couplings to third-generation fermions have been suggested among the possible extensions of the SM [92–96] motivated by a series of anomalies recently observed in charged- and neutral-current B meson decays, $b \to c\ell\nu$ [12–16] and $b \to s\ell\ell$ [17–19], respectively.

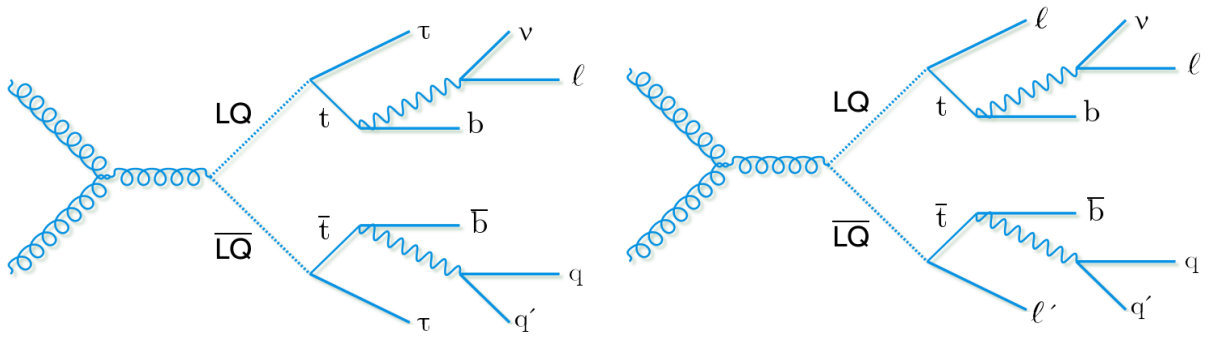

Figure 2.4: Example processes illustrating the production and decay of scalar leptoquark pairs in proton-proton collisions at the LHC that result in multilepton final states.

The ATLAS and CMS Collaborations have conducted a number of searches for leptoquarks with flavor-diagonal and cross-generational couplings involving third-generation fermions [97–105]. The most stringent constraints on scalar leptoquarks with 100% branching fraction to a top quark and first-, second-, or third-generation lepton are set by ATLAS, excluding such particles with masses below 1.48, 1.47 TeV [97] and 1.43 TeV [98], respectively. Similarly, CMS has excluded scalar leptoquarks decaying to a top quark and a τ lepton or a bottom quark and a neutrino with equal branching fractions ($\beta = 0.5$) with masses below 950 GeV [102]. The final states include hadronically decaying top quark and τ lepton, b-tagged jet, and significant missing energy.

Devising the search for beyond...

Chapter 3

The Multilepton Analysis

3.1 Why leptons?

BSM phenomena can manifest itself in various different forms. It can appear either as a narrow resonance in the invariant mass distribution of some particles or it can be nonresonant due to the invisible decays or because of the effective field realization of the theory. Nonresonant signatures can be observed as an excess of data events over the expected contribution from the SM processes in certain kinematic distributions. BSM phenomena can take place in the collision experiments, such as the LHC, and then decay to SM particles either directly or via the SM gauge bosons. In any case, we will witness the direct production of leptons and quarks, or indirect production of neutrinos and any other new invisible particles in the detector.

While leptons have a distinct and isolated footprint in the detector, quarks do not exist in the free state due to the QCD color confinement, and hadronize to form jets of particles. Neutrinos or any other new invisible particles, which are either sterile or only weakly interacting with matter, can only be interpreted as missing transverse momentum in an event, and therefore lack the information about all degrees of freedom. Hence, one of the purest handle, with large signal-to-background ratio, for finding BSM phenomena is leptons. Lepton production at a proton-proton collider machine is also a rare phenomena, mainly because of the dominant strong interaction among the incoming partons. Additionally, identifying the signature of leptons of the desired origin in the detector is much more efficient than tagging the jets to the correct quark or gluon. So, the overall signal to background ratio is much better for leptons than jets.

3.2 Why multileptons?

The BSM signals under study in this thesis give a variety of leptonic signatures i.e. single-, di-, and multileptons in the final state. However, choosing a multilepton final state has an added advantage of significantly reducing SM background contamination over the monumental W+jets, Drell-Yan (DY), and $t\bar{t}$ production. This can be realized from Figure 3.1 which shows a summary of the production cross section of all the SM processes as measured by the CMS Collaboration [106], with up to latest measurements done using the combined data set from 2016–2018 at $\sqrt{s} = 13$ TeV. There is a sharp decrease, of $\mathcal{O}(10^3)$, in the cross section of the single production of Z boson versus the diboson (ZZ) production; similarly the single production of W boson versus the associated production with the Z boson (WZ) results in a decline in the cross section by $\mathcal{O}(10^4)$. The $t\bar{t}V$ and VH processes, where V = W/Z, have cross sections even smaller by a factor of 10 than the ZZ production. Leptonic decays of WZ, ZZ, $t\bar{t}V$, and VH result in multilepton final states. Hence, multilepton probes serve as a powerful tool for the BSM searches.

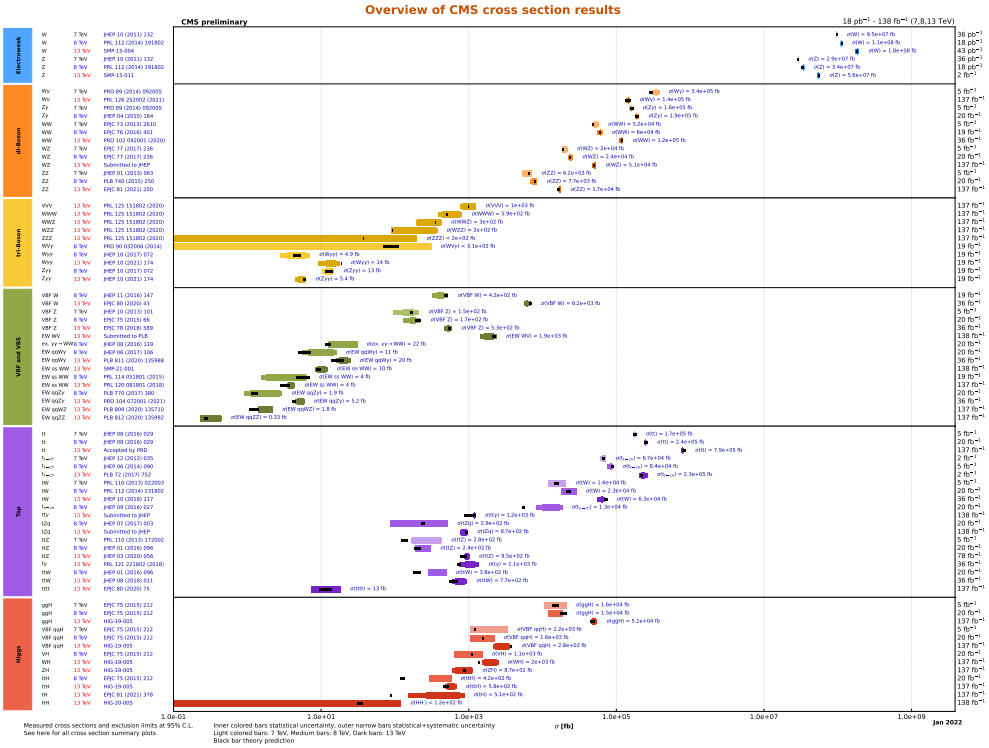

Figure 3.1: A summary of production cross section of the SM processes as measured by the CMS Collaboration [106].

3.3 The multilepton analysis

There have been many BSM searches with multileptons in the past, both by the CMS [58, 78, 79] and the ATLAS [77, 107, 108] Collaborations. However, there are always some caveats. Often the analyses are designed for only high mass BSM searches, and would lack the sensitivity to the low energy scattering processes. Most of the multilepton analyses have been carried out with only the first (e) and second (μ) generation of leptons in the final state. This is because τ leptons differ in many properties from their lighter counterparts in the SM family. They have a lifetime of $\sim 10^{-13}$ s, which means τ leptons travel upto a mean distance of 10 microns and then decay into leptons or hadrons. As a result, reconstructed tau leptons are slightly displaced with respect to the primary interaction vertex. Electrons are universally stable particles, and muons are stable up to the length of the CMS detector. Hence, they can be traced back all the way to the primary vertex. Hadronic decay modes of the τ leptons are reconstructed as jets composed of one and three tracks for the 1-prong and 3-prong decays, respectively, accompanied with zero or more neutral pions. This "multi-prong" nature of the tau decay impacts its reconstruction efficiency as well as the identification against quark/gluon jets. Hence, final states enriched with τ leptons are underexplored due to these difficulties. Consequently, sensitivity to models with preferential couplings to the third generation such as the VLLs, heavy neutral leptons, or to the extended higgs sector, becomes poor. A few multilepton results also lack inclusivity in the search channels, by sculpting only a desired signal region from the entire phase space.

In this thesis, I describe an inclusive multilepton analysis [109], with upto three hadronically-decaying τ leptons. The analysis is sensitive to nonresonant excesses arising from any BSM physics model yielding multiple leptons in the final state.

3.4 Final states

We consider seven distinct final states or channels, based on the number of light lepton (L = e or μ) and hadronic tau (T = τ_{h}) candidates. These seven channels are orthogonal selections, and are listed below:

1. ≥ 4 light leptons and any number of τ_{h} candidates (4L),
2. exactly 3 light leptons and ≥ 1 τ_{h} candidates (3L1T),
3. exactly 3 light leptons and no τ_{h} candidates (3L),

4. exactly 2 light leptons and ≥ 2 τ_{h} candidates (2L2T),

5. exactly 2 light leptons and exactly one τ_{h} candidates (2L1T),

6. exactly one light lepton and ≥ 3 τ_{h} candidates (1L3T), and

7. exactly one light lepton and exactly 2 τ_{h} candidates (1L2T).

The distribution of events in the seven multilepton channels is visualized in Table 3.1. Exclusive multilepton channels such as the 3L, 2L1T and 1L2T are exact in number of leptons, and all the leptons defining the channel are used for further SM background estimation and signal search. Inclusive channels, such as the 4L, only uses the leading four light leptons in p_{T} for the subsequent analysis. Likewise, the 3L1T, 2L2T, and 1L3T channels use only the leading one, two, and three τ_{h} candidates, respectively.

Table 3.1: Analysis channels, based on the electron, muon, and tau multiplicities per event.

	1 e/μ	2 e/μ	3 e/μ	≥ 4 e/μ
0 τ_{h}	–	–	3L	4L
1 τ_{h}	–	2L1T	3L1T	4L
2 τ_{h}	1L2T	2L2T	3L1T	4L
≥ 3 τ_h	1L3T	2L2T	3L1T	4L

Due to the trigger considerations, each event is required to have at least one muon with $p_{\mathrm{T}} > 26(29)$ GeV in 2016 and 2018 (2017) or at least one electron with $p_{\mathrm{T}} > 30(35)$ GeV in 2016 (2017 and 2018) that matches to a corresponding trigger object with $\Delta R < 0.2$. In order to remove overlapping events selected from both muon and electron triggers, we prioritize selection of events with a muon trigger first, attributed to higher trigger efficiency.

3.5 Major SM backgrounds

The multilepton landscape at the LHC is dominated by a variety of SM processes with one or more SM gauge bosons (W, Z, h) or top quarks, as seen in Figure 3.1. These include diboson processes WZ and ZZ; triboson processes WWW, WWZ, WZZ, and ZZZ; pair production of top quarks in association with a vector boson ($t\bar{t}$V where V = W,Z); and higgs boson processes such as VH and $t\bar{t}$H. All these processes give rise to leptons which are energetic, non-displaced from the production vertex, and isolated from the surrounding event activity.

Processes with fewer than three or four non-displaced and isolated leptons also pass the multilepton event selection at a rate lower than their production cross section. This happens via the presence of extra non-isolated or slightly displaced leptons originating from semi-leptonic heavy flavor decays within jets or from other misidentified detector signatures. Examples of such SM processes include W+jets, WW+jets, DY+jets, $t\bar{t}$+jets and other single-top production.

3.6 Analysis workflow

The broad strategy used for designing the multilepton analysis, and to perform the search for BSM phenomena is outlined below:

- We consider all possible multilepton final states, with the minimum requirement of a single light lepton (e/μ) to trigger the events. We select leptons with stringent quality criteria to make sure they come from the desired origin i.e. SM gauge bosons W, Z, h or leptonically decaying τ leptons, and the signal particles. We require extra custom selections on the lepton properties, to reduce some contamination from the SM processes. These are defined in Chapter 5.

- We want to explore the entire multilepton phase space to find evidence of new phenomena. Hence, we do not reject any event from the potential "signal regions (SRs)", pertaining to any particular kinematic requirement. However, in order to estimate the SM backgrounds and optimize our prediction methods, we reserve a few events as "control regions (CRs)" from the multilepton landscape. These CRs are predominantly populated with the SM backgrounds. The procedure and validation of our background estimation methods, along with the effect of important experimental uncertainties are covered in Chapter 6 in detail.

- Once the SM backgrounds are determined and estimated, we perform the search for new phenomena in the SRs. We have employed advanced machine learning (ML) techniques, trained with exhaustive information about the event kinematics. The ML methods enhance the sensitivity of the BSM search for the type-III seesaw mechanism, vector-like leptons, and scalar leptoquark models. The training strategy and the results are in Chapter 7.

- In addition to the ML approach, we also consider an approach based on focusing solely on the SM backgrounds composition, extracting SRs of varying sensitivity for performing an unbiased search. This model-independent strategy differs in ideology and performance from the ML approach, as discussed in Chapter 8. We demonstrate the usefulness of the

search using such SRs by comparing the performance of the three BSM signals probed in this thesis, and also justify where it outperforms the ML training. The best constraints on the three probed models combining the two approaches are also presented in the same chapter.

- Finally, the prospect of reinterpreting our results from the standpoint of future BSM searches is described in Chapter 9.

Conducting the experiment...

Chapter 4

The Experimental Setup

4.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [21, 22] is the world's largest and the highest energy particle accelerator. It is located at the international facility for nuclear research in Europe, the CERN, in Geneva, Switzerland, and started first collisions in 2010. The LHC is a big circular ring with a circumference of 27 km, and is built at a depth ranging from 50 to 175 metres beneath the France-Switzerland border near Geneva. The deep underground LHC tunnel along with the various experiments located at various points is shown in Figure 4.1.

The LHC tunnel houses two parallel beam pipes at ultrahigh vacuum, both of which contains proton beams but also occasionally proton-lead and lead-lead beams, one moving in clock-wise direction and another one in anti-clockwise direction. These beams are traveling with almost the speed of light and are made to collide head-on with each other at four different collision points. The center-of-mass energy of the collision, denoted as $\sqrt{s}$ where s is one of the Mandelstam variables, reaches a maximum of 13 TeV in case of proton-proton collisions, while around 5 TeV for proton-lead and lead-lead collisions. At these collision points, four different particle detectors – CMS, ATLAS, ALICE, and LHCb, are situated. Out of these, CMS and ATLAS are general multi-purpose detectors designed to study a range of phenomena, while ALICE is dedicated for heavy ion collisions, and LHCb to study the forward decays of the heavy bottom and charm hadrons.

Run-I of the LHC is the period between 2010–2013 when the proton collisions happened at $\sqrt{s}$ = 7 TeV in 2010/2011, and at 8 TeV in 2012. After that, there was a long shutdown of LHC complex when the magnets were significantly improved and the collisions restarted in 2015 at $\sqrt{s}$ = 13 TeV. Run-II of the LHC consists of data collection in the years between 2015–2018, with many technical stops in between for scheduled upgrades and maintenance. Figure 4.2 shows the

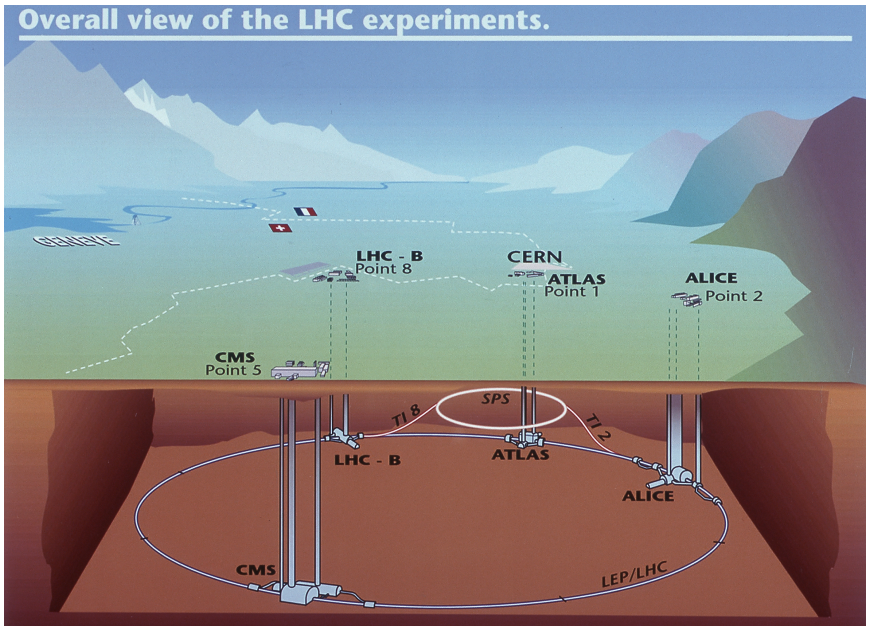

Figure 4.1: The Large Hadron Collider at CERN, Geneva. (*Image Courtesy: CERN*)

distribution of integrated delivered luminosity from the proton-proton collisions in the Run-I and Run-II of LHC, as measured by the CMS Collaboration [110].

4.1.1 Reaching the collision energy

The CERN complex provides a multi-stage acceleration unit to achieve the desired energy of collision in the LHC tunnel. This is explained as follows:

1. First, electrons are stripped off from slow moving hydrogen atoms in a linear accelerator (LINAC2), resulting in a proton beam (H^+ ions) at 50 MeV.

2. These proton beams are fed into a booster ring of 157 m in length, which provides the output proton beams at 1.4 GeV.

3. Next comes a Proton Synchrotron (PS) of circumference 628 m, which elevates the energy of the proton beams to 25 GeV. Also, the beams are accumulated in the PS to form a train of bunches with 25 ns ($\sim$7 m in circumference) spacing.

4. The Super Proton Synchrotron (SPS), 7 m in length, is used to increase the energy of the proton bunches to 450 GeV over a period of few minutes, before they are injected into the main LHC ring.

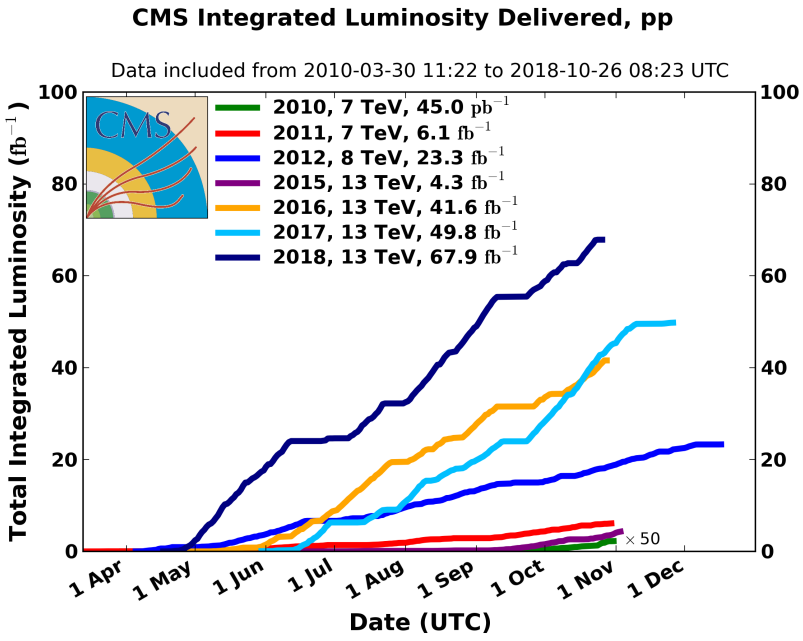

Figure 4.2: Cumulative luminosity per year of Run-I and Run-II delivered to CMS during stable beams for proton-proton collisions at nominal center-of-mass energy. This is measured by the CMS Collaboration [110]. The plots are shown for data-taking periods in 2010 (green), 2011 (red), 2012 (blue), 2015 (purple), 2016 (orange), 2017 (light blue), and 2018 (navy blue), and use the best available offline calibrations for each year.

5. Finally, the proton bunches are circulated for around 20 minutes after which they reach the peak energy of 13 TeV, and are ready for collision.

There are around 1,232 main dipole magnets along the LHC ring to keep the beams moving in circular direction, while an additional 392 main quadrupole magnets are used to keep the beams collimated. There are even stronger quadrupole magnets placed close to the intersection points to squeeze the beams further in order to maximize the chances of interaction at the collision points. Magnets of higher multipole orders are used to correct for smaller imperfections in the magnetic field geometry of the beams, expelling out rogue protons. In total, there are about 10,000 superconducting magnets and approximately 96 tonnes of superfluid helium-4 to keep these magnets at their operating temperature of 1.9 K.

4.1.2 The beam parameters

Instead of having continuous proton beams, the protons are bunched together into 2808 bunches, with 115 billion protons in each bunch. Every bunch crossing is termed as a 'collision event',

or simply an 'event'. The spatial distance between two bunch trains is set such that the collision takes place at an interval of 25 ns, making the collision frequency as 40 MHz. This is done for the purposes of synchronization, acquiring calibration data, and to counter the dead times of the front-end electronics of the detectors, so that they can be reset before the next collision to collect fresh data.

The $\sqrt{s}$ = 13 TeV energy of the collision corresponds to an energy of 6.5 TeV per proton beam. At such energies, protons are moving at the speed of about 0.999999990c, with a Lorentz factor (γ) of 6930. Hence, they complete one revolution around the LHC ring in 90 μs, resulting in 11,245 revolutions per second. The size of the proton bunch near the collision point is $\sim$10 μm in the transverse direction ($\sigma_{x,y}$) and 20 mm in the longitudinal direction (σ_z).

4.1.3 Luminosity of collisions

Luminosity in the scattering theory is a measure of the number of collisions. The instantaneous luminosity, $\mathcal{L}$, is described as the number of events detected in a particle accelerator, in a certain period of time over the cross section (σ):

$$\mathcal{L} = \frac{1}{\sigma}\frac{dN}{dt} \tag{4.1}$$

However, luminosity isn't just the collision rate, rather it measures how many particles were squeezed through a given space in a given time. This doesn't necessarily mean that all those particles will collide with each other, since the size of the particles is very small (proton radius $\sim 10^{-15}$m). The more we can squeeze into a given space, the more likely it is that they will collide. In particle physics, a cross-section is a measure of the probability of some interaction happening, and is measured in the units of area – barns, b (1 b = 10^{-28}m^2).

There are a multitude of possibilities whenever proton bunches collide in the LHC: the protons can just glance off each other or they can undergo hard scattering producing new resonances decaying to a range of other particles. Each of these processes have their own cross-section. The smaller the cross section of a process, the more rare it is to take place in a collision. The only way to increase the chances of a process happening is by increasing the number of collisions or the luminosity.

At the LHC, the instantaneous luminosity is calculated as:

$$\mathcal{L} = \frac{\gamma f k_B N_p^2}{4\pi\epsilon_n \beta^*} F \tag{4.2}$$

Table 4.1: Parameters used in the luminosity calculation for proton-proton collisions at the LHC.

Term	Definition	Value
γ	Lorentz factor	6930
f	Revolution frequency	11,245
k_B	Number of proton bunches	2808
N_p	Number of protons in a bunch	1.15×10^{11}
ϵ_n	Normalized transverse emittance	3.75 μm
β^*	Betatron function at collision point	0.55
F	Reduction factor due to crossing angle	285 μrad

where the definition and value of each of the terms is given in Table 4.1.

With the above value of the parameters, Eqn. 4.2 yields an instantaneous luminosity of $\mathcal{L} \sim 10^{34}$ cm^{-2}s^{-1}. This is the highest luminosity to be achieved by any hadron collider in the world. The integrated luminosity, $\mathcal{L}_{int}$, is calculated by integrating the instantaneous luminosity over the total time for which the collisions were happening, for the given set of parameters, and the unit of $\mathcal{L}_{int}$ is barns inverse (b^{-1}).

The large number of protons squeezed in a small cross-sectional area, i.e. high instantaneous luminosity of LHC, gives rise to additional inelastic proton-proton interactions. Hence, in addition to the hard scattering in the event, there are more concurrent low-energy proton-proton scatterings from either the same bunch crossing or from the adjacent bunches. These unwanted overlapping collisions per bunch crossing are known as "pileup". Figure 4.3 shows the distribution of the average number of pileup interactions per bunch crossing for the proton-proton collisions in the Run-I and Run-II of the LHC, as measured by the CMS Collaboration [110].

4.1.4 The Worldwide LHC Computing Grid

The Worldwide LHC Computing Grid (WLCG) [111] is a global collaboration of computer centres and storage systems to store, distribute, and analyze the 15 petabytes of data generated by LHC every year, as well as LHC-related simulation with near real-time access. This was made possible by combining computer facilities from CERN funding with the national or regional resources brought in by the member institutes and laboratories across the world. The WLCG was constructed as part of the LHC design to handle such significant volume of data.

The WLCG is composed of four levels or "Tiers" - 0,1,2 and 3, where each level provides a specific set of services. Around 20% of the computing facitlity is Tier 0 or central hub. It stores all the raw data (digital information from subdetectors), performs first pass at reconstruction and

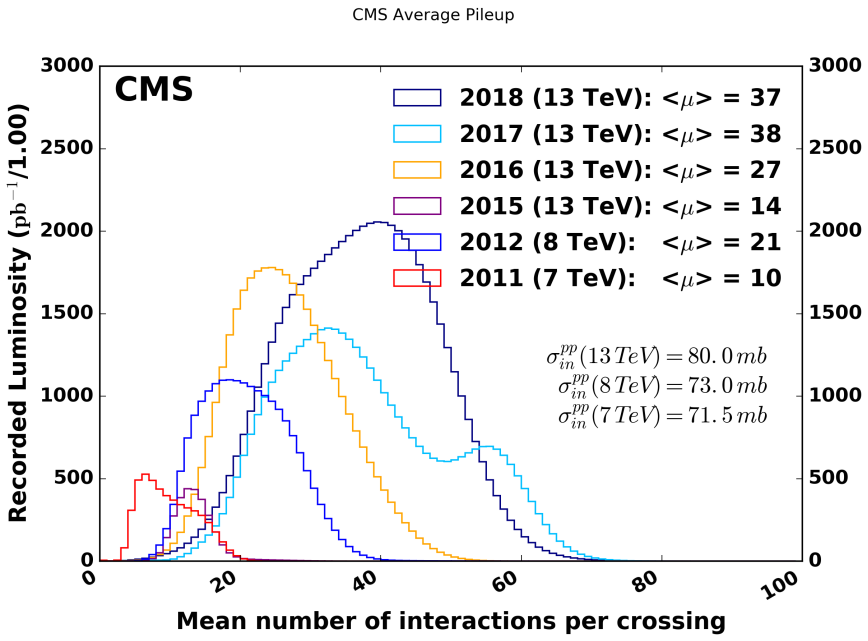

Figure 4.3: The distribution of the average number of interactions per bunch crossing (pileup) for proton-proton collisions in 2011 (red), 2012 (blue), 2015 (purple), 2016 (orange), 2017 (light blue), and 2018 (navy blue). The overall mean values and the minimum bias cross sections are also shown. These are measured by the CMS Collaboration [110]. The plots use only data that passed the "golden" certification (i.e., all CMS sub-detectors were flagged to be ok for any kind of usage in physics analysis), and the "LHC standard" values for the minimum bias cross sections, which are taken from the theoretical prediction from Pythia and should be used to compare to other LHC experiments.

passes on the output to Tier 1, through actual optical-fibre links (10 GB/s). Later, it reprocesses when LHC is not running. Tier 1 (13 centres) is responsible for round-the-clock support for grid, large scale reprocessing, and storing the corresponding output. Also, it distributes the data to Tier 2, and stores the official simulations produced at Tier 2. Tier 2s are typically universities or scientific institutes to store sufficient data locally and provide computing power for analysis tasks. Users can access the grid from one of the many entry points through proper credentials, using Tier 3s.

4.2 The CMS detector

The name of the CMS experiment [112] is inspired by its design [113] incorporating a compact solenoid magnet and its physics goals [114]. The central feature of the CMS is a superconducting

solenoid of 6 m internal diameter, providing a magnetic field of 3.8 Tesla. This strong magnetic field was chosen for the precise measurement of momentum of the muons, with a large bending power for other charged particles as well. In addition to momentum resolution, CMS also aspires for good electromagnetic energy resolution for discrimination between other particles, and good transverse momentum resolution to account for missing energy from the collisions. Hence, within the solenoid volume are a silicon pixel and strip tracker, a lead tungstate crystal electromagnetic calorimeter (ECAL), and a brass and plastic scintillator hadron calorimeter (HCAL), each composed of a barrel and two endcap sections. Forward calorimeters extend the coverage provided by the barrel and endcap detectors. Muons are detected in gas-ionization chambers embedded in the steel flux-return yoke outside the solenoid. The CMS detector geometry is illustrated in Figure 4.4. The detailed design of the CMS subdetectors, alongwith the coordinate conventions is described in the following subsections.

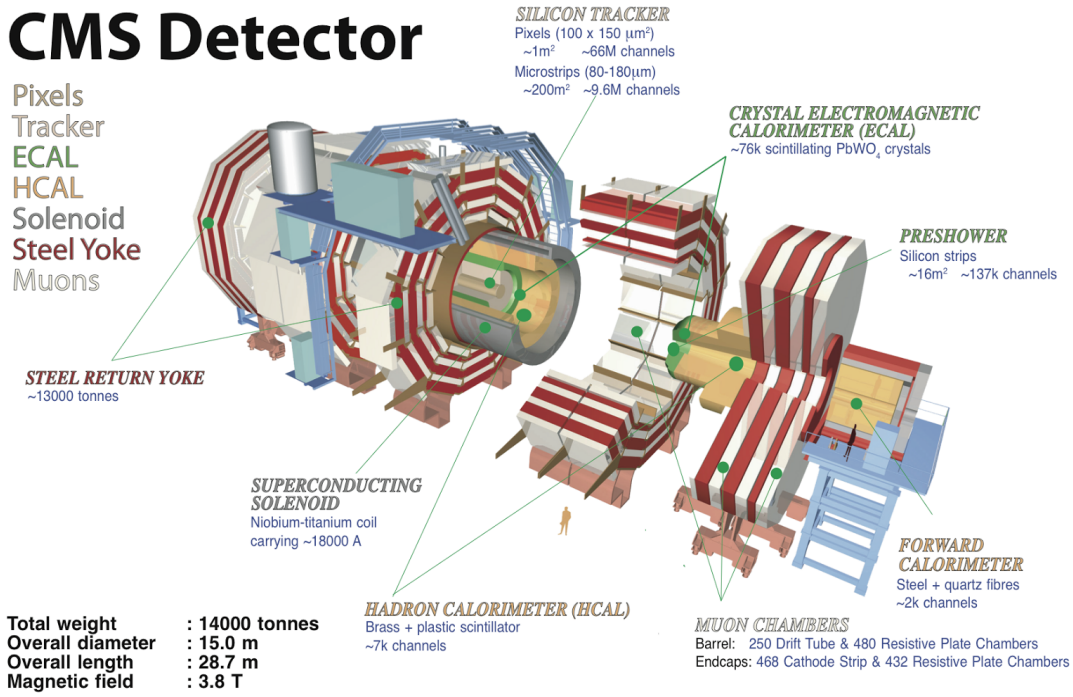

Figure 4.4: The CMS detector at the LHC, CERN. (*Image courtesy: CMS*)

This thesis is based on the data collected by the CMS detector during the Run-II of the LHC. The total integrated luminosity recorded by CMS in proton-proton collisions at $\sqrt{s}$ = 13 TeV corresponds to 138 fb^{-1}, with 36.3, 41.5, and 59.8 fb^{-1} recorded in the years 2016, 2017, and

2018, respectively.

4.2.1 The CMS Coordinate system

Inside the experiment, the origin is centered at the nominal collision point, with y-axis pointing vertically upward, and the x-axis pointing radially inward towards the center of the LHC. Thus, from the right hand curl rule, the z-axis points along the beam in the anti-clockwise direction (towards Jura mountains from LHC). The azimuthal angle ϕ is measured from the x-axis towards the y-axis in the transverse plane. This can be realized in Figure 4.5 left.

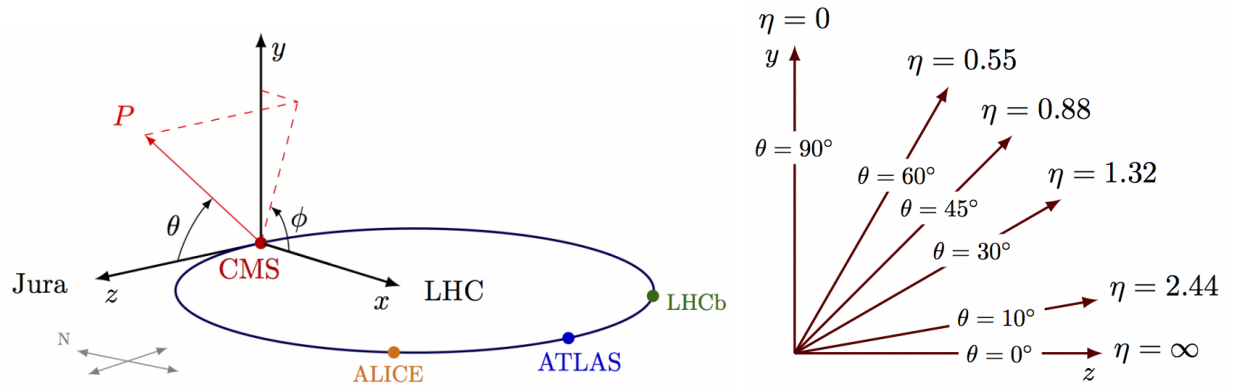

Figure 4.5: The CMS coordinate system (left) and the various planes corresponding to different pseudorapidity (η) values (right).

The polar angle θ is the angle between the particle three-momentum vector and the positive direction of the beam axis. In hadron collider phyics, rapidity (y) is preferred over polar angle θ, since the difference in rapidity of two particles is Lorentz invariant under boost along the longitudinal direction. This is important, especially since collider partons carry different longitudinal momentum fractions. This means that the rest frames of the parton-parton collisions will have different longitudinal boosts with respect to each other. Additionally, in most high-energy hadronic collisions, the number distribution of final-state hadrons is nearly uniform in rapidity in the central regions. The rapidity (y) is given as,

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

In the limit that particles are moving close to speed of light, or that they are massless, pseudo-

rapidity (η) is used instead of rapidity. It is given as,

$$\eta = -ln[tan\left(\frac{\theta}{2}\right)] \tag{4.4}$$

The value of η corresponding to the various polar angles is shown in Figure 4.5 right. Determining η requires only the trajectory of the particle (i.e. θ) while determining y requires us to measure E and p_z. Hence, pseudorapidity is a much simpler observable to measure in a high energy collision experiment.

4.2.2 Inner tracker

The CMS inner tracker [115] provides measurement of the trajectories of charged particles and reconstruction of primary interaction and secondary decay vertices. It is composed of fine granular silicon pixels near the collision point and silicon microstrip detectors afterwards. The length of the tracker barrel is 5.8 m with a radius of 1.3 m. The size of the pixel modules is 100 $\times$ 150 μm^2 corresponding to 128 million readout channels, while that of the strip sensors is 10 cm $\times$ 80 μm with a total of 9.3 million readout strips. The inner tracker has very little passive material so as to minimize particle interactions with the tracker material. The thickness in terms of the radiation length (X_0) is 0.4X_0 in the barrel region ($|\eta| < 1.1$) to a maximum material budget of 1.8X_0 in the transition region ($\sim |\eta| < 1.4$), and then 1.0X_0 in the endcap region ($|\eta| \sim 2.5$).

The transverse momentum resolution ($\frac{\sigma(p_T)}{p_T}$) provided by the inner tracker is $<$1% at 10 GeV and 2% at 100 GeV in the barrel region ($|\eta| < 1.2$), whereas it is 2% at 10 GeV and 10% at 100 GeV in the endcap region ($|\eta| \sim 2.5$). The position resolution in the transverse direction ($\sigma(d_{xy})$) is 20-100 μm in the barrel region while 20-200 μm in the endcap region at 10 GeV, and around 10 μm at 100 GeV in both the regions. Similarly, the position resolution in the longitudinal direction ($\sigma(d_z)$) is 40 μm at 10 GeV and 100 μm at 100 GeV in the barrel region, whereas it is 100 μm at 10 GeV and 1000 μm at 100 GeV in the endcap region. The tracking efficiency is around 99% for muons, and around 75–90% for charged hadrons ($\pi^{\pm}$).

During the Run-II operation of the LHC, the inner pixel tracker went through a major upgrade in March 2017 [116]. Among other minor adjustments, new pixel layers were added in the barrel and endcap section which increased the pseudorapidity coverage from $|\eta|$=2.5 to $|\eta|$=3. The CMS tracker geometries upto the year 2016 (Phase 0), and after the upgrade in 2017 (Phase 1) are illustrated in Figure 4.6. More details about the Phase 1 tracker upgrade will follow in Section 5.2.3.

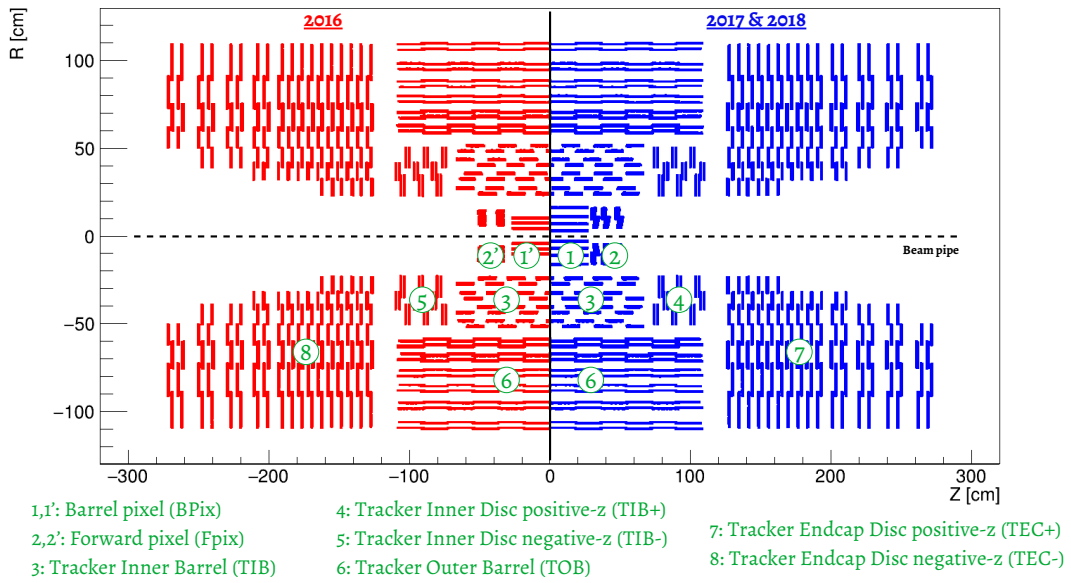

Figure 4.6: An illustration of the various layers of the CMS inner tracker in 2016 (left) and 2017–2018 (right).

4.2.3 Electromagnetic calorimeter

The electromagnetic calorimeter, ECAL [117], is a hermetic, homogeneous, transparent, fine-grained lead tungstate ($PbWO_4$) crystal calorimeter. The homogeneous medium provides better energy resolution by minimizing fluctuations. Particles undergoing electromagnetic interactions are detected in ECAL. As a result, the crystal scintillates and produces light signal in proportion to the incident particle's energy. This energy then gets collected via photodiodes, and the characteristics of the shower and through that, the incident particle's energy are estimated.

In total, there are 75,848 crystals arranged in barrel ($|\eta|$ <1.47) and two endcap sections ($|\eta|$ <3.0). The crystal length in the ECAL barrel section (EB) is 230 mm, while in the ECAL endcap section (EE) is 220 mm. These crystal lengths correspond to a radiation length of $\sim$26X_0 in EB and $\sim$25X_0 in EE, where 1 X_0 = 0.89 cm. The transverse size of crystals at the front face in EB (EE) is 2.2 $\times$ 2.2 cm^2 (2.86 $\times$ 2.86 cm^2) or 0.0174 $\times$ 0.0174 in (η, ϕ). The first ECAL layer starts at a radius of r = 1.29 m. Another characteristic property of the ECAL crystals is the Moilére radius, R_M, which gives the scale of transverse dimension of the fully contained EM showers from $e^{\pm}$ or photon. A smaller R_M results in better position resolution of the showers and also better shower separation due to less overlaps. For the CMS ECAL, R_M = 2.2 cm, related to the X_0 and

atomic number (z) as $R_M = 0.0265X_0(1.2+z)$. The total ECAL energy resolution $\left(\frac{\sigma(E)}{E}\right)$ is around 90% at 20 GeV and 30% at 250 GeV, i.e. exponentially falling with energy.

The preshower (PS) detector, composed of lead block and silicon sensor strips (4288 sensors, 137216 strips), is placed in front of endcaps at $1.65 < |\eta| < 2.6$, with a thickness of 20 cm ($\sim 3X_0$). The lead radiator initiates the EM shower and silicon sensor strips, placed behind the lead block, are used for the output readout. The PS detector improves the separation between photon and neutral pions (π^0), and can also distinguish $e^{\pm}$ and photon against minimum ionizing particles (MIPs).

4.2.4 Hadron calorimeter

The hadron calorimeter, HCAL [118], measures the energy of charged and neutral hadrons i.e. particles made up of quarks and gluons. The deposited energy is rendered measurable by ionization or excitation of atoms of the active medium. It is a sampling calorimeter, with dense absorber (brass) sandwiched with light active planes (plastic scintillator). The output is readout by the wavelength-shifting fibres embedded in scintillator tiles, and is channeled to photodetectors via clear fibres. The first layer of HCAL starts at a radius of r = 1.77 m, and extends radially up to 2.95 m.

The HCAL is organized into four major sections: barrel (HB) covering the pseudorapidity range $|\eta| \leq 1.4$, endcap (HE) covering the pseudorapidity range $1.3 \leq |\eta| \leq 3.0$, forward calorimeters (HF) covering the pseudorapidity range $2.9 \leq |\eta| \leq 5.0$, and outer calorimeter (HO) which is installed just outside the solenoid as a "tail-catcher" detects escaped particles from the inner calorimeter. The thickness of HCAL is represented in terms of the interaction lengths, λ_I, which is the mean distance travelled by hadronic particles before undergoing inelastic nuclear interaction. The total thickness of HCAL layers is around 7–11λ_I upto HF, and around 10–15λ_I including the HO. The jet energy resolution as a function of its transverse energy is around 50% in the barrel region, 30% in the endcap region, and 20% in the very forward region for E_{T} = 20 GeV, while it is around 5–10% at 300 GeV in all the regions.

4.2.5 Muon system

Muons are 200 times heavier than electrons. This is why the amount of synchroton radiation ($\propto \frac{1}{m^4}$) emitted by muons in the presence of magnetic field is very less. Hence, they are MIPs in our detector and lose very little energy while traversing through the inner tracker and both the

calorimeters. Dedicated muon chambers [119] are placed outside the solenoidal magnet, embedded in the steel flux-return yoke, to detect the presence of muons in the events, measuring their position and momentum. The momentum resolution from the muon system alone is impacted by multiple scattering in the detector material before reaching the first muon station, until the muon p_{T} reaches values of 200 GeV. This is when the chamber spatial resolution starts to dominate over the inner tracker measurement.

There are three types of gas ionization detectors which are used to identify and measure muons. The detector design and placement is driven according to the need of covering the large area of detection as well as exposure to different levels of radiation. In the barrel region ($|\eta| < 1.2$) where residual magnetic field and muon flux due to neutron-induced background is low, drift tube (DT) chambers are used. In the endcap regions, where magnetic field is strong and the muon flux is also high, cathode strip chambers (CSC) are deployed and covers the pseudorapidity range upto $|\eta| < 2.4$. Resistive plate chambers (RPC) are used both in the barrel and the endcap regions as they provide a fast response with good time resolution but coarser position resolution. RPCs can therefore identify the correct bunch crossing without ambiguity.

There are 1400 muon chambers, out of which there are 250 DTs, 540 CSCs, and 610 RPCs. In the barrel region, four stations of detectors are arranged in cylinders interleaved with the iron yoke. Then there are 5 wheels of the yoke (labeled YB-2 for the farthest wheel in -z, and YB+2 for the farthest is +z) along the beam direction. In each of the endcaps, the CSCs and RPCs are arranged in 4 disks perpendicular to the beam, and in concentric rings, with three rings in the innermost station, and two in the remaining ones. In total, the muon system contains nearly 1 million electronic readout channels.

4.2.6 Trigger and Data Acquisition System

The rate of collisions (40 MHz) and the overall data (15 PB/year) both are much higher than the rate at which it can be written to mass storage. At the LHC's instantaneous luminosity of $\sim 10^{34}$ $\mathrm{cm}^{-2}\mathrm{s}^{-1}$, each bunch crossing results in an average of 20 inelastic proton-proton collision events, with approximately 1 MB of zero-suppressed data being produced in all the subdetectors together. The current archival storage capacity is of the order of 10^2 Hz and at the data rates of $\mathcal{O}(10^2$ MB/s).

The CMS Trigger [120] and Data Acquisition System [121] (TriDAS) is designed to scrutinize the physics behind the collision events at every bunch crossing, so that it can make real-time decisions about storing only the interesting events for further analysis. The required rejection of $\mathcal{O}(10^5)$, i.e. from 40 MHz to 100 Hz is too large to be achieved efficiently in a single processing

step. Hence, the CMS triggering system is split into two levels: Level-1 Trigger and High-Level Trigger.

4.2.6.1 Level-1 Trigger

The first level, i.e. Level-1 Trigger (L1T) is designed of custom hardware processors using the information from calorimeters and muon detectors, as well as some correlation of information between these systems. It reduces the rate of events accepted for further processing to less than 100 kHz within a fixed latency of about 4 μs. The L1T decision is based on the presence of primitive objects such as electrons, photons, muons, and jets above some predefined transverse energy or momentum thresholds, and also global sums like total transverse energy or missing energy. Much of this logic is encoded in custom Application Specific Integrated Circuits (ASICs) or Gate Arrays (e.g. FPGAs), and also some static RAMs that are used as libraries of preloaded look up tables for pattern recognition. While the L1T decision making is in progress, all the other high-level information about the event is stored in buffer pipelines.

4.2.6.2 High-Level Trigger

The second level, i.e. High-Level Trigger (HLT) is designed to reduce this maximum L1 accept rate of 100 kHz to a final output rate of 100 Hz. This is done using large processor farms which performs a quick reconstruction of the full event by combining information from all the subdetectors (including track reconstruction), and then takes a software-level decision about whether to keep the event for data storage and further analysis or to discard them forever. There are many dedicated streams through which the qualifying events are ultimately stored. These streams, or the so-called trigger paths, are then used for event selection at the analysis-level.

The list of trigger paths used in this thesis for the multilepton events selection are the lowest unprescaled isolated single muon and single electron paths across the three years of data-taking. The list of full trigger path names is given in Table 4.2. The efficiency of selecting an event from these muon and electron trigger paths in the three years of data-taking are given in Appendices A.1 and A.2, respectively. Typically, the trigger efficiencies around the plateau region range between 80–90% and 60–90% for the single isolated muon and electron paths, respectively, across the three years of data-taking.

Table 4.2: List of trigger paths used in this multilepton analysis in the years 2016, 2017, and 2018.

Year	Trigger path
Single muon	
2016	HLT_IsoMu24_v or HLT_IsoTkMu24_v*
2017	HLT_IsoMu27_v*
2018	HLT_IsoMu24_v*
Single electron	
2016	HLT_Ele27_WPTight_Gsf_v*
2017	HLT_Ele27_WPTight_Gsf_L1DoubleEG_v* or HLT_Ele32_WPTight_Gsf_v*
2018	HLT_Ele32_WPTight_Gsf_v*

4.2.6.3 Detector Control System

Another crucial role of the DAQ system is the functioning of a Detector Control System (DCS) for the operation and supervision of all detector components and the general infrastructure of the experiment. The DCS is a key element for the operation of CMS, and guarantees its safe operation to obtain high-quality physics data. Figure 4.7 shows an image from the control room of CMS at LHC, CERN where I undertook these DCS shifts. The various screens in the front display the control systems and the dynamic status of the different subdetector layers.

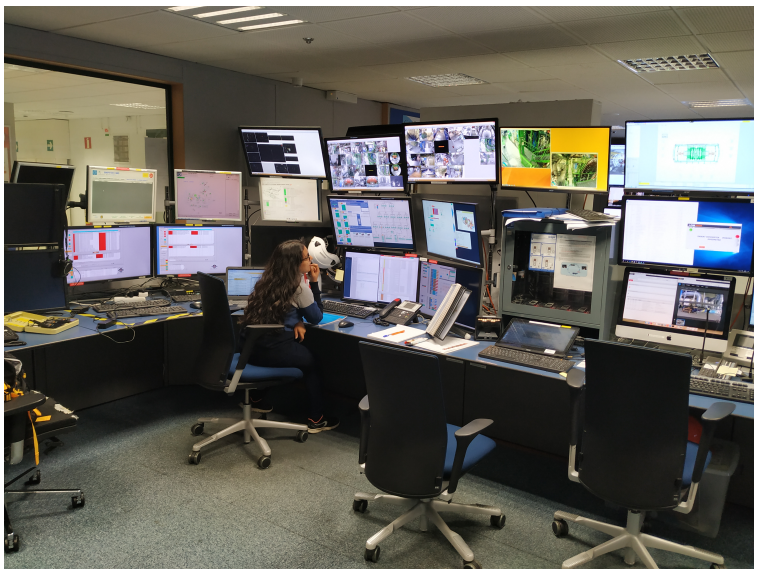

Figure 4.7: An image from the control room of the CMS at LHC, CERN with me undertaking the Detector Central System (DCS) shifts.

4.2.6.4 CMS software and data formats

The CMS data analysis software [122] is built on the Event Data Model, which is a framework centered around the concept of an event. An event is a container of products of type C++, and most of these products are containers of physics objects like tracks, clusters, particles etc. The output storage from DAQ is based on ROOT [123], and is then processed in various data formats for different use-cases for the analysts. These output formats with their contents and size per event (MB) are listed in Table 4.3.

Table 4.3: List of CMS data formats with their contents and event size (MB).

Data format	Contents	Event size (MB)
DAQ-RAW	Detector data from front end electronics + L1T result. This is the primary record of physics event and is used as input to the HLT reconstruction.	1–1.5
RAW	Detector data after the result of the HLT selections, with potentially some of the higher level quantities calculated during HLT processing. This is used as input to Tier-0 reconstruction.	0.70–0.75
RECO	Detailed reconstruction output from Tier-0 with objects (electrons, muons, photons, tracks, vertices, jets, hit clusters), reprocessed after applying detector calibrations and alignments.	1.3–1.4
FEVT	Full event output (RAW+RECO) containing complete information of all the reconstructed objects (cells, clusters, hits, electrons, muons etc).	1.75
AOD	A subset of RECO with all physics objects and some localized hit information. This is used for a large fraction of analysis studies.	0.5
MINIAOD	A subset of AOD, with only high level physics objects such as electrons, muons, photons, and jets.	0.05
NANOAOD	A skimmed version of MINIAOD to reduce event size and information, but sufficient to do final analysis.	0.009

In this thesis, I have used data and simulation samples processed in the MINIAOD [124] format.

Identifying the detector signatures...

Chapter 5

Simulation and Reconstruction

One key ingredient in data analysis is an understanding of the detector performance by estimating efficiencies, predicting the background contribution from SM processes, and also the estimation of systematic uncertainties. The most important tool for these steps is simulation of SM and BSM events in the CMS detector. Hence, we need simulation for an unbiased prior understanding, effective planning for designing physics searches, understanding how specific signals manifest themselves to devise pure selections for signal or orthogonal regions, and understanding underlying effects from boosted topologies or higher-order cross sections on important kinematic variables used for final discrimination between SM and BSM phenomena.

The reliability of prediction from simulation depends on the goodness of physics models used to create the simulation and realistic description of the CMS detector for precise emulation of the geometry and particle-material interactions. For the latter part, the quality of models used in evaluating the propagation of particles through the detector is also very important to preserve.

5.1 Generation and Simulation

Data begins with proton beams, while simulation begins with generators which use a Lagrangian. The simulation chain consists of generation of four-momenta based on matrix-level calculations, hadronization of colored particles, followed by the processes involved in particles interacting with the detector material resulting in digital signals. The process of taking the digital signals and reconstructing physics objects such as electrons, muons is then identical (as far as possible) between the collected data and the simulation. A good simulation chain will mimic the various energy-loss mechanisms of the actual detector, along with producing the correct multiplicities of secondary and tertiary particles.

5.1.1 Event generators

Event generators are software libraries that generate simulations of high energy particle physics events. They generate events from various SM or BSM processes equivalent to those produced in the collisions. The tree-level perturbative quantum chromodynamics (QCD) that describes the physics of collisions are quite simplified. However, these processes are often accompanied by photon or gluon bremsstrahlung, and loop-level corrections, which are usually too complex to be solved analytically. Furthermore, the non-perturbative QCD is presently beyond the ability of computation in the lattice QCD framework. Hence, the event generators are based on Monte Carlo (MC) methods which rely entirely on repeated random sampling to obtain the desired numerical results.

The main physics components behind the design of the modern event generators are the following:

1. Hard subprocesses and resonance decays, which are described by matrix elements.
2. Initial- and final-state parton showers or radiation producing either photons or gluons.
3. Multiple parton-parton interactions, beam remnants and other outgoing partons.
4. Hadronization via color confinement strings to produce primary hadrons, and their probabilistic decays.

The generators used in this analysis are MADGRAPH5_AMC@NLO [125] and PYTHIA [126]. While PYTHIA is independently capable of handling all of the above steps, MADGRAPH5_AMC@NLO is best suited for tree-level matrix calculations, also known as leading-order (LO), and next-to-leading order (NLO) corrections. The parton output of the latter is fed to PYTHIA for further showering and hadronization. The POWHEG [127–129] software is used for simulating samples for processes produced via quark-antiquark annihilation or gluon-gluon fusion at NLO, while MCFM [130] is used for the LO generation of the same.

In this thesis, we have used event samples generated through MC full simulation, to estimate the yields of signal and irreducible SM background processes in the multilepton final states. The list of SM processes and the corresponding event generators as used in the analysis is summarized in Table 5.1.

Table 5.1: List of all the irreducible SM background processes and the corresponding event generators, as used in the analysis.

Process	Event generator	Order
SM irreducible backgrounds		
$Z\gamma$, WZ, $t\bar{t}Z$, VVV	MADGRAPH5_AMC@NLO	NLO in pQCD
Top quark processes	MADGRAPH5_AMC@NLO	NLO, LO
Drell-Yan (DY)	MADGRAPH5_AMC@NLO	NLO in pQCD
Higgs processes	MADGRAPH5_AMC@NLO, POWHEG, JHUGEN	NLO
ZZ ($qq \to ZZ$)	POWHEG	NLO
$t\bar{t}$	POWHEG	NLO
ZZ ($gg \to ZZ$)	MCFM	LO
BSM signals		
Type-III Seesaw, Vector-like leptons	MADGRAPH5_AMC@NLO	LO in pQCD
Scalar Leptoquarks	PYTHIA	LO in pQCD

5.1.2 Simulation software

After the production and hadronization of the incoming particles from an event, we need to emulate the detector response. This includes simulating the particle trajectories through the silicon tracker and showering in the calorimeters due to particle-material interactions. This will result in the production of hits or energy deposits in the various subdetector layers in the form of analog signals. These are then digitized, similarly to the collision data, and reconstruction is performed as described in Section 5.2.

To simulate the CMS detector, we have two different dedicated software – Full Simulation [131] and Fast Simulation [132].

5.1.2.1 Full Simulation

Full Simulation or FullSim, is the primary tool for generating simulation events from SM processes, as well as from the BSM phenomena. It uses GEANT4 [133] libraries, tuned precisely to the detector design and magnetic field profile, to emulate the propagation of particles through the active and passive parts of the CMS subdetectors. Different particles lose energy via different mechanisms. All charged particles (electrons, muons, charged hadrons) undergo ionization and multiple scattering upon interacting with the tracker layers. Electrons, in addition, emit bremsstrahlung photons and these radiated photons convert to electron-positron pair. Both charged and neutral hadrons experience elastic and inelastic nuclear interactions due to strong forces from

the nucleus of the detector material. Once the generator particles have passed through the detector, the information about their simulated trajectory and energy deposits is stored for further digitization.

Digitization is the process of converting the simulated information into the electronic readout response output, close to the collision data as acquired by the DAQ systems. For example, the simulated clusters of electric charges in the tracker layers are modeled for Landau fluctuations, drift, and diffusion effects. The energy deposits in the ECAL are modeled for the efficiency and non-uniformity of the light collection by the $PbWO_4$ crystals. Similarly, the number of photoelectrons from the energy losses in the HCAL are generated after taking into account the internal non-uniformities and electronic noise. Finally, the digitization in the muon subdetectors aim to achieve a resolution greater than the dead time of the front end electronics.

After the processing of the simulated data to digitized signal, a standard reconstruction is performed. Particles are tracked in very small steps in the tracker layers, following various hit permutations. Similarly, a local reconstruction is performed in the calorimeters to collect the compatible hits and form towers of energy deposits belonging to individual particles. These tracks and energy deposits are used for further object and event reconstruction, just like in collision data, as described in Section 5.2. The performance of the FullSim software is regularly validated using test beam data and previous simulation results. Though very precise, FullSim is a highly computing-intensive and time-consuming simulation.

5.1.2.2 Fast Simulation

Computing resources are limited. So, it is difficult to meet the production needs of huge MC samples using FullSim with the increasing luminosity of data collected by the CMS. FastSimulation or FastSim, is a faster alternate to event simulation and reconstruction in CMS, without much compromise on the physics performance. The speed of the simulation step is increased as compared to FullSim due to two factors: faster particle propagation in the inner tracker and parametrized models (instead of full GEANT4-based simulation) for particle-material interaction in the calorimeters. The standard tracking of FullSim is also modified to make use of generator-level information in FastSim, for reconstructing the trajectories of charged particles faster by getting rid of resource-heavy hit combinatorics.

Fast Simulation of an event takes place through the following steps:

1. **Particle propagation -** Tracker geometry is approximated as infinitely thin concentric cylinders and disks, where material resides only on the surface. In FastSim, a simplified magnetic

($\vec{B}$) field map, parametrized in pseudorapidity, is used for particle propagation. The field exists only on the tracker layers, and is zero in between. Particles are propagated in a helical trajectory from one layer to another, using the $\vec{B}$ field on the current layer. The four-momenta is updated after every crossing with the tracker layer. The intersection points of the traversing particles are determined in this simplified geometry, and are projected onto the real tracker modules to determine the simulated position or "SimHits".

2. **Particle-material interaction -** Material interactions in the inner tracker are emulated according to the thickness of the layers, in terms of radiation lengths (X_0), and this occurs only when particles cross a layer. The development of calorimetric showers in the transverse and longitudinal direction are done through simple analytical functions, parametrized for the properties of the incoming particles. As a result of interactions, the energy and direction of the particles changes; they might disappear when the energy goes below the threshold for detection, and new particles might emerge from the decays.

3. **Emulating reconstructed hit position -** Unlike FullSim, in FastSim the SimHit positions are directly smeared according to the resolution of standard hit reconstruction. The smearing resolution depends on the subdetector and layer. For the pixel tracker, parametrized templates for hit resolution and hit merging probabilities are generated from PIXELAV, depending on dimensions in local coordinates, pixel type, incident angle, and number of pixels that are hit. For the strip tracker, simple Gaussian-based smearing around the strip center is employed, the width of which is obtained from FullSim.

The charged-particle tracking in FastSim is also simplified by the use of generator-level information. In principle, it is performed on a per-particle basis by first selecting all the hits belonging to a particle and then applying track quality cuts same as in the standard iterative tracking [see Section 5.2.2]. This not only saves the hit combinatorics, but also has negligible misreconstruction rate at similar tracking efficiency. Consequently, FastSim is faster than FullSim by a factor of 100 alone in the simulation step, and overall 20 times together with the reconstruction step. This has made FastSim a popular choice of framework for the production of large sets of simulation scanning the complex parameter space of models such as supersymmetry, or exotic BSM searches at high p_{T}, or other private physics studies. With the increasing LHC luminosity and pileup in events, FastSim will have an even wider usage across the CMS.

5.2 Object and Event Reconstruction

For each proton-proton collision, a comprehensive list of final-state particles along with their kinematic properties (momentum, pseudorapdity η, and azimuthal angle ϕ) are decoded. This is done with the help of an algorithm, known as "Particle-flow" (PF) [134] within the CMS community. It is a holistic approach to provide a global event description to improve performance by effectively identifying physics objects. The PF algorithm targets electrons, muons, hadronic tau decays, photons, jets originating from quark or gluon hadronization, and missing transverse momentum from the neutrinos. The PF algorithm also allows an efficient mitigation of the pileup interactions.

5.2.1 The PF algorithm

The secret behind the high success rate of the PF algorithm is embedded in the detector design itself. Modern general-purpose detectors at high-energy colliders, such as the CMS, are inspired from the cylindrical detection layers, nested around the beam pipe. This allows for a maximum coverage, exposing much of the active material directly in the path of the traversing particles. As a result, different physics objects leave behind distinct signatures in each subdetector layers. A visual representation of signatures from different particles can be seen in the transverse slice of CMS detector in Figure 5.1. Information from the various subdetector elements is then combined coherently to reconstruct the final physics object. The working principle of PF for the various physics objects at the CMS is described in the following subsections.

5.2.1.1 Tracks

The first encounter of all charged particles is with the tracker layers, where they ionize the silicon-based tracker modules leaving behind "hits" along their trajectory. The charged particles deflect in the presence of magnetic field in the tracker due to the Lorentz force, whereas the neutral particles (photons and neutral hadrons) pass through undeflected and without depositing any charges. The curvature of the track helps in determining the initial momentum of the charged particles. More details about the track reconstruction algorithm will follow in Section 5.2.2.

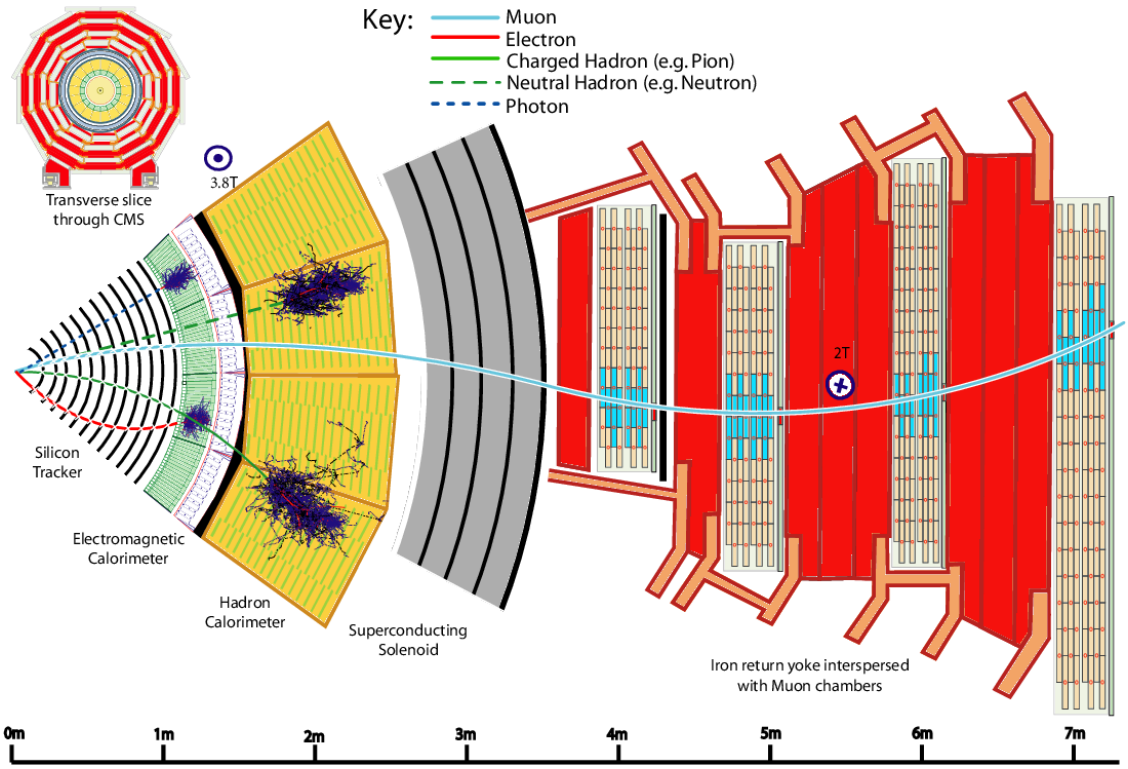

Figure 5.1: A sketch of the various particle signatures in a transverse slice of the CMS detector, shown from the beam interaction region and all the way to the muon detectors. (*Image Courtesy: CMS*)

5.2.1.2 Muons

They are minimum ionizing particles in the CMS calorimeters. Hence, muons only give rise to signatures in the inner silicon tracker and the outside muon gaseous detectors. The geometric matching of two compatible tracks, one from the inner tracker and one from the muon detectors, without any energy deposit in the ECAL and HCAL results in an unmistakable muon candidate reconstruction [135].

There are three different approaches for muon reconstruction: inside-out tracking, standalone tracking, and outside-in or global reconstruction. In the inside-out case, a tracker track is propagated to the muon system with a loose matching criteria to DT or CSC segments. If at least one segment geometrically matches the extrapolated track, then it is termed as a tracker muon. These muons have significant misidentification, caused by the energetic hadron shower remnants that enter the innermost muon station (punch-through). In the standalone approach, a track is built only from the muon subdetectors, gathering all the CSC, DT, and RPC information. These muons are

known as standalone muons, and have poor momentum resolution and highest contamination from the cosmic sources. Finally, the third and the most efficient approach is the global reconstruction of muons based on outside-in tracking. A standalone-muon track is matched to a tracker-muon track by checking the compatibility of the two tracks' parameters.

5.2.1.3 Electrons and photons

These are absorbed in the ECAL, owing to the electromagnetic showers generated upon the interaction with heavy nucleus. Electrons radiate bremsstrahlung photons under the influence of the electric field and the photons convert into electron-positron pair. This cascade of secondary particles continues to grow till the energy of the photons goes below the threshold of pair production or photoelectric effect, and Compton scattering becomes the dominant mode of energy loss mechanism for electrons. The characterstic amount of matter traversed in these interactions, before the particle is completely absorbed, can be represented in terms of the radiation length (X_0) of the material. These electromagnetic showers are then recorded as clusters of energy in ECAL from which the total energy of the particles are determined. The ECAL cluster geometrically matched to a tracker "track" from the primary interaction point is then reconstructed as an electron candidate, whereas clusters not pointing to any track are reconstructed as a photon candidate. An example reconstruction scenario for electrons and photons in a toy detector model of CMS is shown in Figure 5.2.

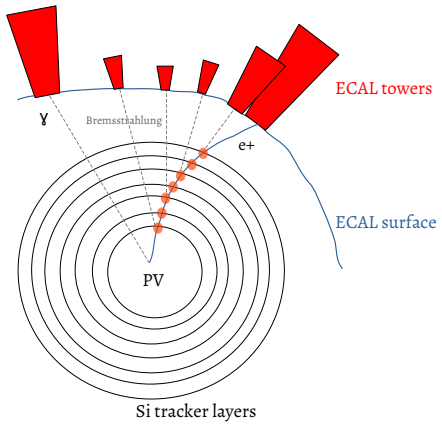

Figure 5.2: An example reconstruction scenario for electrons and photons in a toy detector model of CMS.

5.2.1.4 Hadronic taus and jets

Charged and neutral hadrons, arising either from hadronic decay products of a tau lepton or from quark/gluon hadronization, may initiate a similar hadronic shower in the ECAL, which then escalates to inelastic nuclear interactions in the HCAL producing pions and other hadrons. These are then fully absorbed in the HCAL, after a few nuclear interaction lengths (λ). The reconstruction of a charged hadron therefore is performed by the combined measurement of tracker, ECAL and HCAL clusters, without any signal in the muon detectors. The same applies on the neutral hadrons, except for the requirement of a matching track. Figure 5.3 shows a typical decay of hadron into EM and hadronic showers while traversing through dense material.

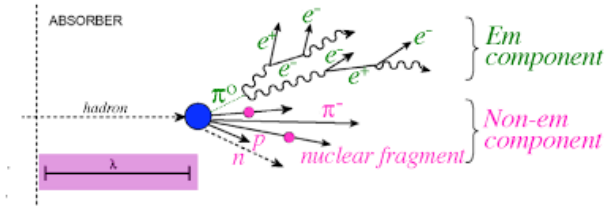

Figure 5.3: An illustration of hadron decay in a dense material, forming electromagentic and hadronic showers. (*Image courtesy: IOPscience*)

Jets are clustered using "anti-k$_T$" algorithm [136] for various cone radii (R = $\sqrt{\Delta\eta^2 + \Delta\phi^2}$). The physics objects used in the clustering of jets are the identified PF objects i.e. electrons, muons, photons, charged and neutral hadrons. The anti-k$_T$ algorithm is a type of sequential recombination around the hardest energy deposit, parametrized by the inverse of the energy scale (p_{T}) in the distance metric. Jets are composite objects made up of several particles. Hence, the momentum is determined as the vectorial sum of all particle momenta, and is found from simulation to be, on average, within 5–10% of the true momentum over the whole p_{T} spectrum and detector acceptance. Pileup contamination is removed with the help of "charged hadron subtraction" (CHS) scheme [134] where the energy of all the charged hadrons not originating from primary vertex is removed. However, impact of the neutral hadrons from PU is mitigated through an event-by-event jet-based-area correction of the jet four-momenta [137–139].

Hadronic taus differ from quark or gluon jets in the multiplicity of its constituents, the collimation, and the isolation. They decay either via one charged hadron (1-prong mode) i.e. h$^{\pm}$, and zero, one or two neutral pions (π^0) or via three charged hadrons (3-prong mode) i.e. h$^{\pm}$h$^{\pm}$h$^{\pm}$, and zero or one neutral pion (π^0). Hadronic taus are reconstructed through the "hadrons-plus-strips" (HPS) algorithm [140] using anti-k$_T$ with radius parameter of 0.4 (AK4) PF jets of p_{T} >14 GeV

and $|\eta|$ <2.5 as the seed. The jet is deconstructed and an intermediate meson resonance (ρ or a_1) is sought after, by combining the constituent particles. If found, they are reconstructed as hadronic taus. An illustration of the two decay modes of hadronic taus via the intermediate meson resonances is shown in Figure 5.4.

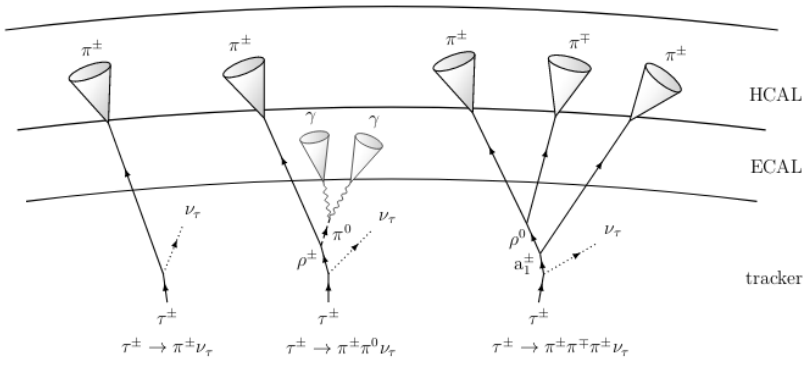

Figure 5.4: An illustration of the two decay modes of hadronic taus via the intermediate meson resonances is shown in the various subdetector layers of the CMS. (*Image courtesy: CMS Tau POG*)

The HPS algorithm tries to reconstruct neutral pions from its decay to two photons, which often converts in the tracker material itself before reaching ECAL. The electrons or positrons from photon conversion bend in the azimuthal direction due to the presence of magnetic field of the CMS solenoid. Hence, the calorimeter signature for the neutral pions from tau decay are extended in ϕ-direction. To reconstruct them, a dynamical "strip" of size 0.05×0.20 in $\eta - \phi$ plane centered around the most energetic electromagnetic particle from AK4 PF jet is built. All other particles falling within that window are recombined and the strip four-momentum is recalculated. In the end, strips with $p_{\mathrm{T}} > 1$ GeV are combined with the charged hadrons to construct the individual τ_{h} decay modes. The following decay topologies are considered by the HPS algorithm:

1. Single charged hadron without any accompanying neutral pion, or that it is too soft to be reconstructed as a strip.

2. Single charged hadron and one strip, in which photons from the π^0 decay are collimated.

3. Single charged hadron and two strips, in which photons from the π^0 decay are well-resolved.

4. Three charged hadrons without any accompanying neutral pion. All the three charged hadrons are required to originate from the same secondary vertex.

Each τ_{h} candidate is required to have a mass compatible with its decay mode and a unit charge. Collimated τ_{h} candidates are selected by requiring all charged hadrons and neutral pions from its decay to be within a circle of radius ΔR = (3.0 GeV)/p_{T} in the (η, ϕ) plane i.e. the signal cone. The size of the signal cone is capped above 0.1 at low p_{T}, and below 0.05 at high p_{T}. The radius of signal cone decreases with p_{T} to account for the boost of the τ decay products. The isolation cone of the τ_{h} lepton is defined outside the signal cone, up to a radius of 0.5.

5.2.1.5 Neutrinos

They interact only weakly with the matter, therefore doesn't produce any signature and are like "'ghost" particles. The presence of neutrinos in the collision is interpreted as a resultant missing transverse momentum in the event, calculated following the conservation of momentum in the x-y plane. This is demonstrated in Figure 5.5.

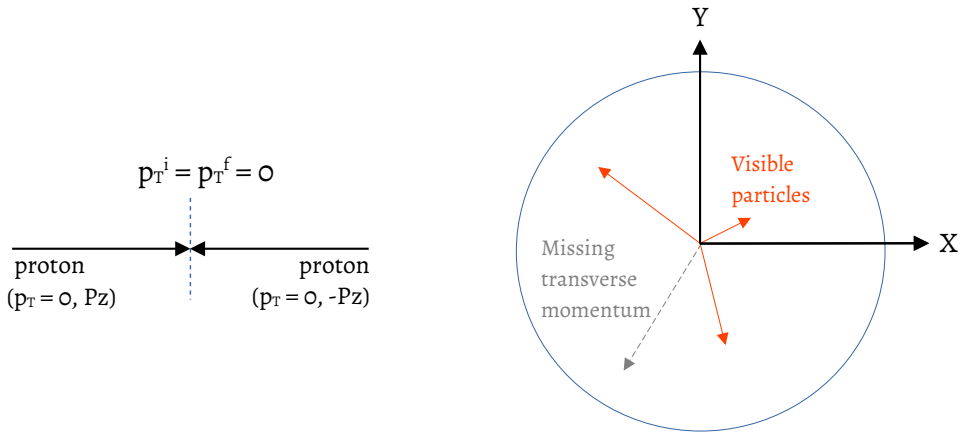

Figure 5.5: Calculation of missing transverse momentum following the conservation of momentum in the x-y plane.

5.2.1.6 Primary vertex

In each event, the candidate vertex which has the largest total sum of the physics-object p_{T}^2 is taken to be the primary proton-proton interaction vertex (PV). The physics objects considered in the sum

are the jets clustered using the anti-k$_T$ algorithm, and the associated $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$, which is the negative of the vector sum of the p_{T} of those jets.

5.2.2 The tracking algorithm

Charged particles in CMS are tracked in small steps through successive hits combination. A combinatorial track finder (CTF) based on Kalman filtering algorithm [141] is used to reconstruct the tracks. This is carried out in three broad steps:

1. Formation of initial seed, which are a collection of two (doublets) or three (triplets) hits compatible with a charged-particle trajectory, and also satisfying a few quality criteria on tranverse momentum (p_{T}), and distance from the collision point.

2. Trajectory building along the direction of seed by gathering all other hits from the tracker layers. This is also termed as pattern recognition.

3. Final track fit to determine the charge particle properties such as origin, transverse momentum, and direction of motion.

A simplified illustration of the above steps is presented in Figure 5.6.

To improve the tracking efficiency while keeping a control on the misreconstructed tracks, the CTF is performed in several iterations, known as *iterative tracking algorithm*. Each iteration differs in the type of the initial seed and quality of the final track parameters, thereby targeting charged particles of different origin. The reconstructed hits belonging to a particle once consumed in the track formation are then removed from the entire collection. This is known as "hits masking", and it prevents the false assignment of hits to another track, thereby saving the memory and time consumption by the CTF.

The first tracks to be reconstructed are the ones which give the cleanest signature in the tracker. These are high transverse momentum tracks passing through many silicon tracker layers and are mostly produced in the decays of W, Z, and Higgs boson. Thus, these tracks also tend to be close to the interaction point, due to the very short lifetimes of the gauge bosons. This property is commonly known as "promptness" of the particle. The "InitialStep" iteration is designed with the requirement of a triplet seed from pixel tracker layers, with the minimum track p_{T} of 600 MeV and within a radius (R) of 2 cm from the interaction point. The tracks arising from the decays of b hadron (with a finite lifetime) can be displaced with respect to the interaction point. These are reconstructed in a separate iteration called "DetachedTripletStep" which requires a pixel triplet

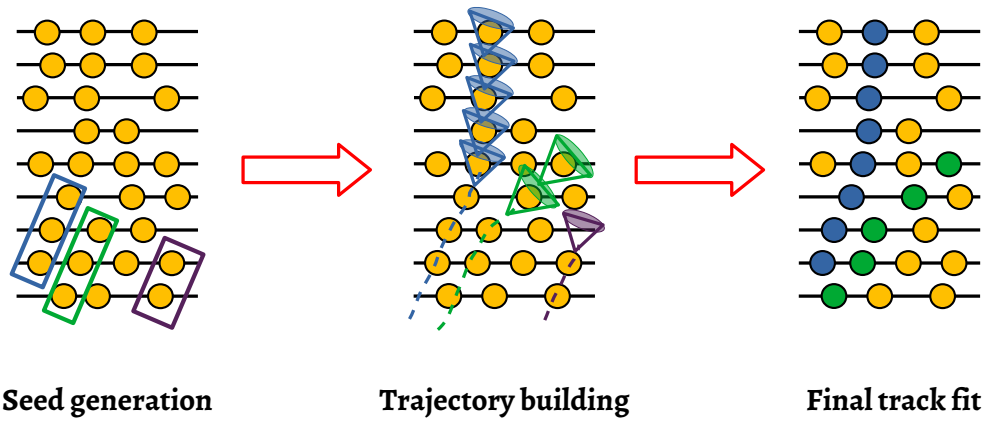

Figure 5.6: A simplified illustration of the CTF algorithm for the charged-particle tracking. The left figure is the first step of seed generation which is a combination of two or three hits. The middle figure is the trajectory building where all other hits compatible in direction with the initial seed are found. Finally, the right figure is the last step of performing a χ^2-fit to determine all the track candidate parameters.

seed of $p_T > 300$ MeV and R = 1.5 cm. To recover high p_{T} tracks which fail to form triplet seed are reconstructed from pixel doublet pair, even from nonconsecutive layers to allow for the missing hits in between. Tracks which are very displaced tend to be reconstructed with a doublet or triplet initial seed from a combination of both pixel and strip tracker layers, or just strip tracker layers. A summary of all the iterations for the Phase 0 of CMS tracker is provided in the Table 5.2.

The number of hits in the CMS tracker layers is directly proportional to the number of incoming charged particles produced in an event. At the LHC, every proton-proton collision produces approximately a thousand charged particles. As a result, the number of hit permutations for tracks reconstruction by the CTF algorithm in an event is extremely large, and also ends up using a lot of CPU time and power. More realistically, the time taken to simulate for e.g. a single top quark pair production event using a rigorous GEANT4-based full detector simulation (FullSim), and then to reconstruct the event using standard tracking and PF algorithm as explained above is $\mathcal{O}$(100s). This is quite a large number considering we produce simulation of various SM processes with a million events or more!

Table 5.2: A summary of the tracking iterations for the charged-particle track reconstruction for Phase 0 of CMS tracker.

Iteration	Name	Initial seed	Track parameters	Targeted charged-particles
1	InitialStep	Pixel triplet	$p_{\mathrm{T}} > 600 MeV$, R $\lesssim$ 0.02 cm	Prompt, high p_{T}
2	DetachedTripletStep	Pixel triplet	$p_{\mathrm{T}} > 300 MeV$, R $\lesssim$ 5 cm	From b hadron decays
3	LowPtTripletStep	Pixel triplet	$p_{\mathrm{T}} > 200 MeV$, R $\lesssim$ 0.02 cm	Prompt, low p_{T}
4	PixelPairStep	Pixel pair	$p_{\mathrm{T}} > 600 MeV$, R $\lesssim$ 0.02 cm	Recover high p_{T}
5	MixedTripletStep	Pixel & TEC Strip triplet	$p_{\mathrm{T}} > 400 MeV$, R $\lesssim$ 7 cm	Displaced
6	PixelLessStep	TIB, TID & TEC Strip triplet/pairs	$p_{\mathrm{T}} > 400$MeV, $R \lesssim$ 25 cm	Very displaced
7	TobTecStep	TOB & TEC Strip triplet/pairs	$p_{\mathrm{T}} > 550 MeV$, R $\lesssim$ 60 cm	Very displaced
8	JetCoreRegionalStep	Pixel & TIB Strip pairs	$p_{\mathrm{T}} > 10 GeV$	Inside high p_{T} jets

5.2.3 Phase 1 tracking developments in FastSim

In 2017, the CMS pixel tracker was upgraded and a single new layer was added, both in the barrel (BPix) and in the endcap or forward (FPix) section. Figure 5.7 shows a comparison of the pixel tracker geometry in Phase 0 (red points) and Phase 1 (blue points), in the longitudinal and transverse planes. This resulted in increasing the total pseudorapidity range in the tracker volume from 2.5 to 3, thus increasing the acceptance for the charged particle tracking.

More importantly, along with the addition of new BPix layer in the far end, the innermost BPix layer was also moved closer to the beam pipe by 10 mm to increase the probability for detection of very short lived particles before their decay. Hence, it became possible to reconstruct prompt and high p_{T} tracks in the InitialStep iteration from a seed of pixel hit quadruplets with much lower misreconstruction rate.

The older FPix layers were all replaced by new endcap discs with smaller inner radius (6 cm $\rightarrow$ 4.5 cm) and larger outer radius (15 cm $\rightarrow$ 16 cm), thereby increasing the coverage for charged particles. The new FPix layers consists of an assembly of an inner and an outer ring, tilted at an angle of 12° with respect to each other, as opposed to the simple geometry in Phase 0 with only radial modules oriented at 90°. This allows a simultaneous measurement of the z-position of the charged particle track, and also efficient replacement of modules (with earlier radiation damage) of the inner ring.

Together with a new FPix layer closer to the barrel section, it then also became possible to reconstruct the low p_{T} tracks through a new iteraion "LowPtQuadStep" with a pixel quadruplet seed. For the same reasons, dedicated new iterations for the displaced tracks with a pixel quadruplet seed ("DetachedQuadStep") and to recover high p_{T} tracks through a pixel triplet seed ("HighPtTriplet-Step") were also added.

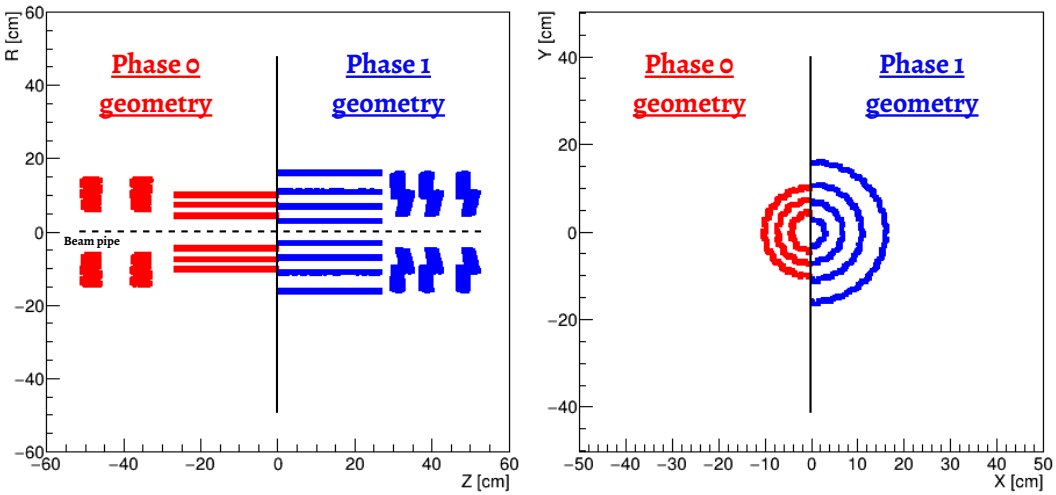

Figure 5.7: A comparison of the Phase 0 (red points) and Phase 1 (blue points) CMS pixel tracker geometry in the longitudinal (left) and transverse (right) planes. The (0,0) co-ordinate is the geometric center of the CMS detector. For illustration purposes, only one side of the detector from Phase 0 and Phase 1 is shown here. The points represents hits of particles on the various tracker layers, produced from a Fast Simulation of 1000 top quark pair production events at $\sqrt{s} = 13$ TeV without any pileup vertices.

I implemented the new pixel tracker geometry and modified the track reconstruction algorithms in the FastSim package of CMS. The changes were validated and ultimately merged in the CMSSW package [142].

Figure 5.8 shows the initial results of the track reconstruction efficiency as a function of p_{T} and average number of hits per track as a function of η of the track candidates for the Phase 0 and Phase 1 pixel tracker geometry, in both FastSim and FullSim. As can be seen from left figure, the tracking efficiency in the standard tracking of FullSim improves by 50-100%, especially in the low p_{T} region, with the new geometry. While roughly similar improvement is also observed in FastSim, it has always been slightly over-efficient than FullSim, more so at very low ($p_{\mathrm{T}} < 1$ GeV) and very high ($p_{\mathrm{T}} > 100$ GeV) regions. The higher efficiency at the low p_{T} values is due to the fact that in FastSim, equipped with the truth information, we do not lose tracks due to the multiple scattering. Nuclear interaction models in the tracker material which leads to either a kink in the original hadron trajectory, or to the production of a number of secondary particles are not implemented in FastSim. As a result, there are no efficiency losses, unlike in FullSim, especially

at the high p_{T} values. In the right figure, average number of hits per track increases in the new geometry, more so in the forward regions due to the addition of double-layered FPix discs. The higher number of hits per track in FastSim, especially in the barrel region, can be again attributed to the use of truth information in the tracking which allows to preserve all the hits belonging to a track.

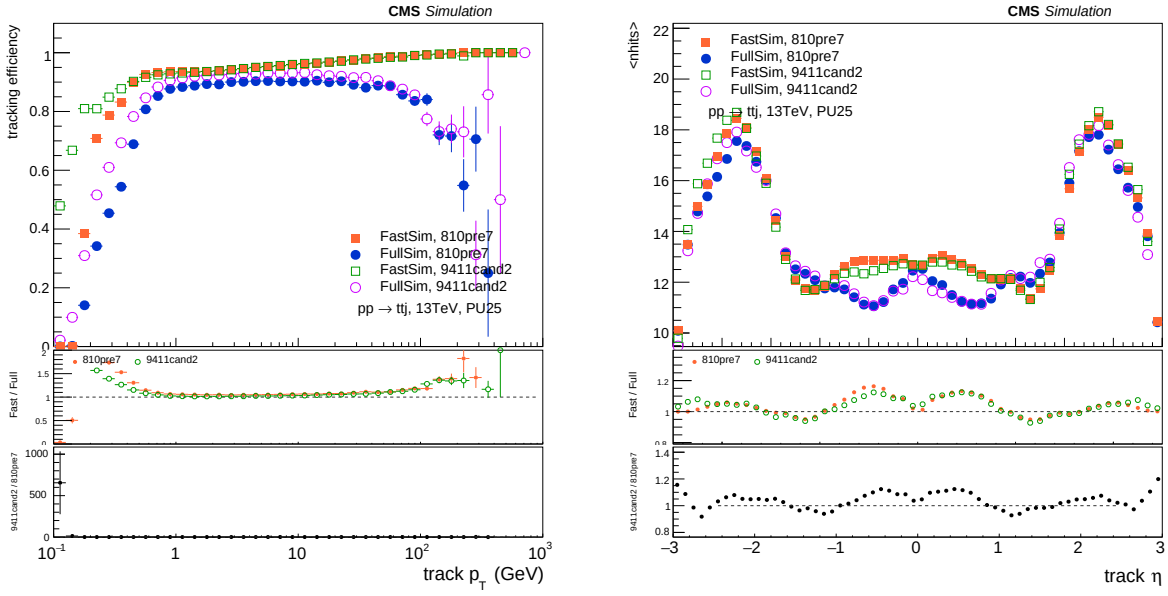

Figure 5.8: Track reconstruction efficiency as a function of p_{T} (left) and average number of hits per track as a function of η (right) of the track candidates, measured from a Fast Simulation of 1000 top quark pair production events at $\sqrt{s} = 13$ TeV with 25 pileup vertices. The solid circle and solid square points are the performance of FullSim and FastSim, respectively, in the Phase 0 geometry (CMS software release 8_1_0_pre7) of CMS tracker. The open circle and open square points are the performance of FullSim and FastSim, respectively, in the Phase 1 geometry (CMS software release 9_4_11_cand2) of CMS tracker. The lower panel shows the ratio of the FastSim to the FullSim performance in the Phase 0 and Phase 1. The uncertainties on the data points are purely statistical.

5.2.3.1 Bringing FastSim closer to FullSim

After the initial implementation of Phase 1 tracking, I did several developments to improve the performance of FastSim with respect to FullSim. These included a missing quality check on the quadruplet seeds [143] due to which FastSim was being over-efficient, especially at the low p_{T}. This can be seen in the Figure 5.9. Another significant improvement in the infrastructure

of FastSim was fixing the presence of dereferenced unique pointers which was causing memory leaks [144]. Although there wasn't much impact on the tracking efficiency after this fix, there was a huge reduction in the duplication rates of the tracks, as can seen be seen in the Figure 5.9.

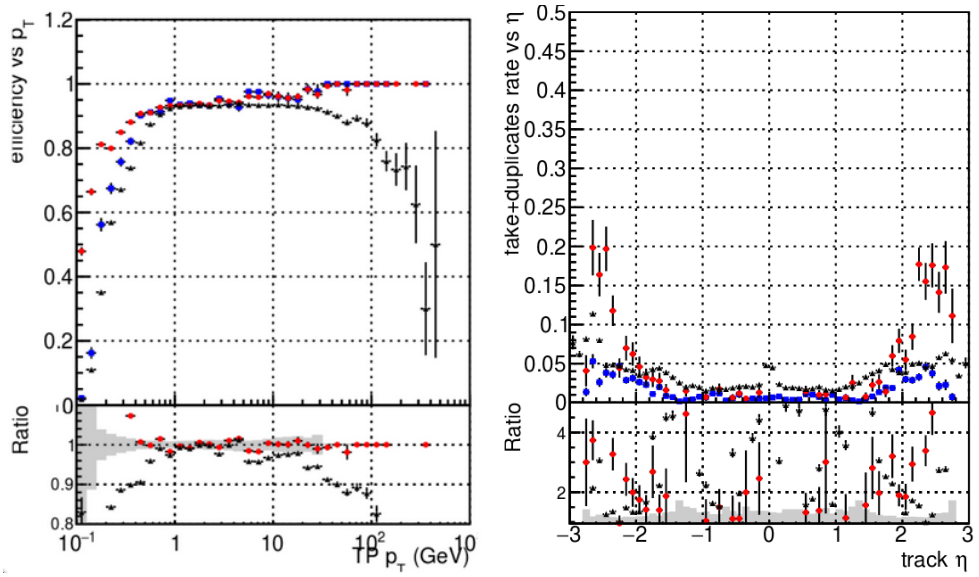

Figure 5.9: Track reconstruction efficiency as a function of p_{T} (left) and misreconstruction rate as a function of η (right) of the track candidates, measured from a Fast Simulation of 1000 top quark pair production events at $\sqrt{s} = 13$ TeV without any pileup vertices. The black, red, and blue curves are the performance of FullSim, FastSim before the fix, and FastSim after the fix, respectively, in the Phase 1 geometry (CMS software release 10_2_X) of CMS tracker. The lower panel shows the ratio of the FastSim to the FullSim performance before the fix (red) and after the fix (black). The uncertainties on the data points are purely statistical.

5.2.3.2 Configuring FastSim to switch between geometry

A crucial feature to ensure smooth functionality of FastSim while generating the simulation samples is the ability to switch between the two geometries, i.e. Phase 0 and Phase 1. I configured the FastSim package of the CMS to include this functionality. This was done with the help of python modifiers, called "Eras", which are always used in the generation commands. Hence, the end user need not to worry about setting other flags in FastSim. The developments were validated and ultimately merged in the CMSSW package [145, 146].

5.2.4 Muon identification

Muons originating from SM gauge bosons (W,Z,h) or from leptonic decays of tau or from the BSM models targeted in this thesis are energetic, relatively isolated from other event activity, and much closer to the PV since the lifetimes of all these mother particles are very small. Such muons are efficiently reconstructed from the global tracking ("outside-in" approach), as explained in Section 5.2.1.2.

Muons with $p_{\mathrm{T}} > 10$ GeV and $|\eta| < 2.4$ are chosen in this multilepton analysis. In addition, they must satisfy $|d_{\mathrm{z}}| < 0.1$ cm and $|d_{\mathrm{xy}}| < 0.05$ cm, where d_{z} and d_{xy} are the longitudinal and transverse impact parameters of the muon track with respect to the PV. The relative isolation is defined as the scalar p_{T} sum, normalized to the muon p_{T}, of photon and hadron PF objects within a cone of radius $\Delta R = 0.4$ around the muon, plus a correction term for pileup mitigation. A delta-beta corrected relative PF-isolation of maximum 15% only is allowed for the muons in this thesis. Table 5.3 summarizes the identification criteria of the muons of the desired origin, used across the three years of data-taking. These are mainly applied on the track-quality properties.

Table 5.3: Summary of muon identification requirements in 2016, 2017, and 2018.

Variable	Requirement
Loose ID (i.e. IsPFmuon() && (IsGlobalMuon() || IsTrackerMuon()))	True
Fraction of valid tracker hits	>0.8
In addition, either of the following two sets of conditions must be satisfied:	
Good global muon	IsGlobalMuon = True Normalized global-track χ^2 <3 Tracker-Standalone position χ^2 <12 Number of kinks in the track <20 Compatibility of inner track with the muon segment >0.303
Tight segment compatibility	Compatibility of inner track with the muon segment >0.451

In order to suppress multilepton background contributions due to misidentified sources, we have also applied additional custom selections on the muon $\mathrm{SIP}_{\mathrm{3D}}$ and DeepCSV neural network (used for b-tagging and described in Section 5.2.7) score of the mother jet of muon. These requirements are summarized in Table 5.4. The $\mathrm{SIP}_{\mathrm{3D}}$ is the 3-dimensional distance from the PV divided by the uncertainty of the position resolution. The mother jet is defined as the AK4 CHS jet with $p_{\mathrm{T}} > 10$ GeV, which falls within a distance of $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ equal to 0.4 to the muon. By restricting the DeepCSV score of mother jet, we eliminate muons coming from b-hadron decay being falsely selected under the above identification criteria.

Table 5.4: Muon, electron and τ_{h} lepton displacement cuts in 2016, 2017, and 2018.

		2016	2017	2018
$\mathrm{SIP_{3D}}$	e/μ	< 10.0	< 12.0	< 9.0
LeptonJetDeepCSV	e/μ	< 0.6	< 0.4	< 0.3
	τ_h	< 0.8	< 0.8	< 0.8

The cuts on the $\mathrm{SIP_{3D}}$ are tuned in each year of data-taking separately so that we obtain not less than 95% efficiency for the WZ process (taken as a standard multilepton candle) with muons and electrons. The difference in the $\mathrm{SIP_{3D}}$ cut value results from combination of multiple sources such as the different performance of the the DeepCSV b-tagging algorithm in different years, as well as different detector (tracker upgrade) and data-taking conditions (pileup profile). Figure 5.10 shows the impact of these custom selections on both muons and electrons, before and after the application, through the yield of WZ, DY+jets, $t\bar{t}$+jets, and type-III seesaw fermions of $\mathrm{m_{\Sigma}} =$ 550 GeV in the flavor-democratic scenario. These distributions are produced in the 3L channel, normalized to data luminosity in 2018. As can be seen from the figure, the WZ yield decreases only by 3–5% after the implementation of these custom selections. On the other hand, decrease in the DY+jets and $t\bar{t}$+jets yield is around 35% and 70%, respectively. For the example signal scenario shown in these distributions, the decrease in yield is about 10%.

In addition to the above working point, a less-stringent version of the identification criteria is also designed for muons, which is necessary for the data-driven background estimation method as described in Section 6.2.1. This includes relaxing the relative isolation to 100%, keeping all other selections exactly the same. All loose muons are required to lie outside the cone of radius ΔR = 0.05 from each other, to suppress contributions of tracks splitting from pions.

The efficiency of the custom selections on the muon identification in data and simulation are given in the Appendix A.3.

5.2.5 Electron identification

As described earlier, electrons are reconstructed from an inner tracker track geometrically matched to a ECAL cluster [147]. Electrons also emit bremsstrahlung photons in the tracker material at the various intersection points. Hence, a "supercluster" (SC) combining many neighbouring ECAL cells around the seed cluster ($E_{\mathrm{T}}>1$ GeV) is created, to account for the energy taken away by the radiated photons, and a dedicated tracking algorithm, based on the "Gaussian sum filter" (GSF) [148] is employed to reconstruct the matching track and its parameters.

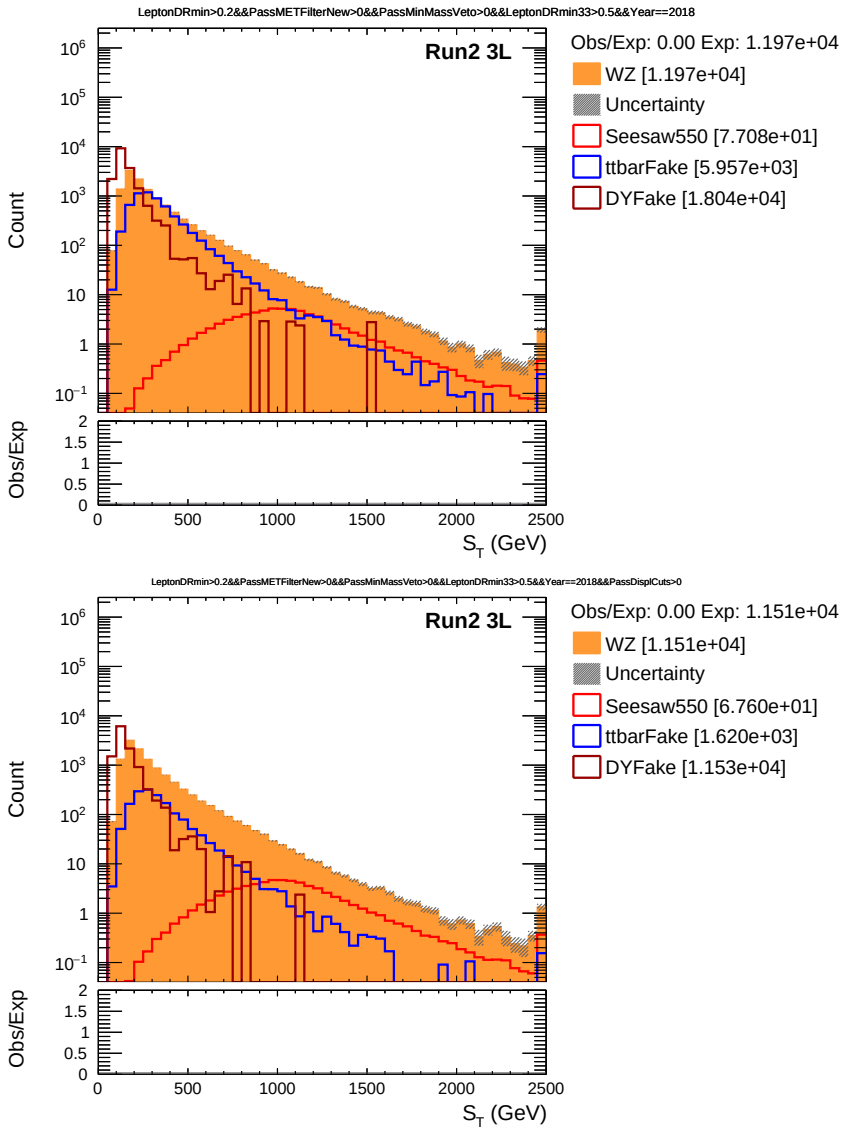

Figure 5.10: Impact of custom DeepCSV requirement of both electrons and muons, before (left) and after (right) the application, on the yield of WZ, DY+jets, $t\bar{t}$+jets, and type-III seesaw fermions of $m_{\Sigma} = 550$ GeV in the flavor-democratic scenario. These distributions are produced in the 3L channel, normalized to data luminosity in 2018.

Similar to muons, electrons originating from SM gauge bosons (W,Z,h) or from leptonic decays of tau or from the BSM models are considered. In this thesis, we use electrons of $p_{\mathrm{T}} > 10$ GeV and $|\eta| < 2.4$. In addition, electrons must satisfy $|d_{\mathrm{z}}| < 0.1$ cm and $|d_{\mathrm{xy}}| < 0.05$ cm in the ECAL barrel region ($|\eta| < 1.479$), and $|d_{\mathrm{z}}| < 0.2$ cm and $|d_{\mathrm{xy}}| < 0.1$ cm in the ECAL endcap region ($|\eta| > 1.479$). To further improve the quality of electrons used in the SM or BSM searches, we

place extra requirements on its detector signatures. These are typically applied on the transverse and longitudinal profile of the shower (full5×5_sigmaIetaIeta), absolute difference between the SC and track (η, ϕ) at the position of closest approach (dEtaSeed, dPhiIn), hadronic to electromagnetic energy fraction (H/E), effective isolation with pileup correction (relIsowithEA), absolute difference of total energy (E) to total momentum (p), number of expected missing hits in the inner tracker, and finally whether the electron is consistent with a photon conversion or not. Table 5.5 summarizes the list of selections for such electrons as used in this thesis.

Table 5.5: Summary of electron identification requirements.

Variable	Pseudorapidity region	
	$\lvert\eta_{SC}\rvert \leq 1.479$	$\lvert\eta_{SC}\rvert > 1.479$
full5x5_sigmaIetaIeta $<$	0.0106	0.0387
abs(dEtaSeed) $<$	0.0032	0.00632
abs(dPhiIn) $<$	0.0547	0.0394
H/E $<$	0.046+1.16/E_{SC}+0.0324$*\rho/E_{SC}$	0.0275+2.52/E_{SC}+0.183$*\rho/E_{SC}$
relIsowithEA $<$	0.0478+0.506/p_{T}	0.0658+0.963/p_{T}
abs(1/E - 1/p) $<$	0.184	0.0721
expected missing inner hits $\leq$	1	1
pass conversion veto	yes	yes

Table 5.4 summarizes custom selections on the electron $\mathrm{SIP_{3D}}$ and DeepCSV score of the mother jet of electron, similarly to muons. The loose identification working point for electrons only differ in the isolation requirement, which is relaxed to 100%. All selected loose electrons within a cone of radius ΔR = 0.05 of a selected loose muon or other electrons are discarded to suppress contributions due to bremsstrahlung.

The efficiency of the custom selections on the electron identification in data and simulation are given in the Appendix A.4.

5.2.6 Hadronic tau lepton identification

Hadronic tau candidates of $p_{\mathrm{T}} > 20$ GeV, $|\eta| < 2.3$, and $|d_{\mathrm{z}}| < 0.2$ cm, which are reconstructed from the HPS algorithm as discussed in Section 5.2.1.4, are required to satisfy multivariate criteria based identification. The identification of hadronic tau relies on the isolation of the reconstructed decay products with respect to any other hadronic activity in their vicinity. The isolation cone in ΔR is 0.5 and is built-in the definition of the tau ID discriminators. In CMS, we have two

different algorithms for the identification of hadronic taus, MVA-based discriminant [149] (now obsolete) and DeepNN-based discriminant [150]. The older MVA ID was designed to be used in conjuction with dedicated discriminators against electrons and muons. The newer DeepNN ID, on the other hand, is a convolutional multi-classification neural network which provides simultaneous classifiers to discriminate genuine hadronic tau decays from jets, electrons, and muons.

In this thesis, the DeepNN-based tau ID is used for the identification of hadronic taus. The neural network training is performed with a total of 129 input variables. These include low-level features from the PF candidates inside the tau signal and isolation cones such as tracks and energy deposits, and high-level features such as transverse momenta, decay mode, impact parameter etc. of tau candidate and general event properties such as average energy density. Finally, all these input features are combined, after some pre-processing, in a five-layer dense network which ends at four output neurons, carrying a likelihood for each tau candidate to be a genuine tau or a fake from light leptons or a fake from jet. Different working points can be chosen in combination for these output neurons to achieve the good quality τ_{h} leptons with desired signal-to-background ratio.

This analysis uses the very tight working point (byVTightDeepTau2017v2p1VSjet) for discrimination against the jets, and loose working points for discrimination against electrons (byLooseDeepTau2017v2p1VSe) and against muons (byLooseDeepTau2017v2p1VSmu). Table 5.6 shows a comparison of yield in the 2L1T and 1L2T channels using 2018 data, for the tight working point of the DeepNN ID as used in the analysis vs a tight working point criteria from the MVA-based identification. As can be seen from the table, number of events in the 2L1T channel reduces by $\sim$50% and by $\sim$70% in the 1L2T channel, when switching from MVA to DeepNN ID.

Table 5.6: Comparison of yield in the 2L1T and 1L2T channels using 2018 data with the use of MVA-based vs DeepNN-based identification.

Channel	No. of events with MVA-based τ_{h} ID	No. of events with DeepNN-based τ_{h} ID
2L1T	83730	35986
1L2T	2724	783

Similarly, Table 5.7 shows a comparison of yield in the 2L1T and 1L2T channels using a signal MC simulation sample of right-handed τ neutrino of mass = 200 GeV, for the tight working point of the DeepNN ID as used in the analysis vs a tight working point criteria from the MVA-based identification. The τ_{h} lepton selected in the events are matched to a generator-level τ lepton within a cone of ΔR=0.02, to make sure genuine taus are selected. As can be seen from the table, there is a small increase (3–5%) in the number of events in both the 2L1T and 1L2T channel, when switching from MVA to DeepNN ID.

Table 5.7: Comparison of yield in 2L1T and 1L2T channels with the use of MVA-based vs DeepNN-based identification using a signal MC simulation sample of right-handed τ neutrino of mass = 200 GeV.

Channel	No. of events with MVA-based τ_h ID	No. of events with DeepNN-based τ_h ID
2L1T	504	521
1L2T	279	289

To conclude from the above two tables, we can say that while the efficiency of selecting genuine taus from decays of SM gauge bosons (W,Z,h) or prompt signal particles increases only by a small amount, at the same time there is a large reduction in the number of fake τ_h lepton candidates with the use of DeepNN-based ID over the MVA ID.

For τ_h leptons, additional requirement on only the DeepCSV score of the mother jet is placed other than the above selections. The DeepCSV requirement for the three years is summarized in Table 5.4. Figure 5.11 shows the impact of custom DeepCSV requirement, before and after the application, on the yield of WZ, DY+jets, $t\bar{t}$+jets, and type-III seesaw fermions of $m_\Sigma = 550$ GeV in the flavor-democratic scenario. These distributions are produced in the 2L1T channel, normalized to data luminosity in 2018. As can be seen from the figure, the decrease in the WZ yield is $\sim$5%, whereas the fakes from $t\bar{t}$+jets reduces by around 35%.

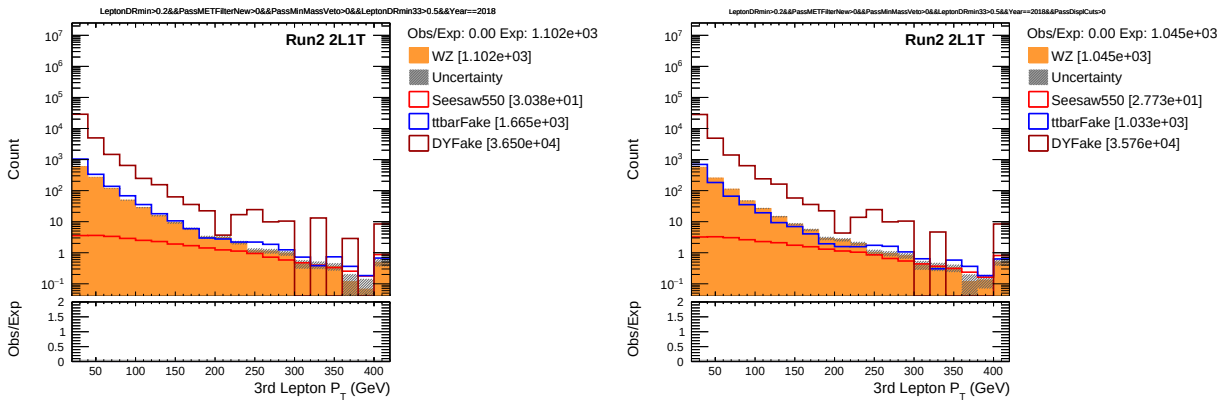

Figure 5.11: Impact of custom DeepCSV requirement of the τ_h candidates, before (left) and after (right) the application, on the yield of WZ, DY+jets, $t\bar{t}$+jets, and type-III seesaw fermions of $m_\Sigma = 550$ GeV in the flavor-democratic scenario. These distributions are produced in the 2L1T channel, normalized to data luminosity in 2018.

All selected taus within a cone of $\Delta R < 0.5$ of a selected loose muon or loose electron are discarded to suppress $\ell \to \tau$ fake tau contributions. A loose working point for discrimination

against jets (byVLooseDeepTau2017v2p1VSjet), keeping all other selections the same, is also defined for τ_{h} leptons. This is used for the data-driven misidentified lepton background estimation, with ample statistics in the sidebands region.

The efficiency of the custom selections on the tau identification in data and simulation are given in the Appendix A.5.

5.2.7 Jet identification

Jets used in this thesis are AK4 PF CHS jets with $p_{\mathrm{T}} > 30$ GeVand $|\eta| < 2.4$. They are required to satisfy the loose working point of the PF jet ID in 2016 and tight working point in 2017 and 2018 [151]. These selections are summarized in Table 5.8 for 2016 and 2017, and in Table 5.9 for 2018. All selected jets are required to be outside a cone of $\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2} = 0.4$ around a selected loose electron, loose muon, or loose tau as defined above, where $\Delta\phi$ is the azimuthal distance.

Table 5.8: Summary of jet identification requirements in 2016 and 2017.

Variable	2016	2017
For $-2.7 \leq \eta \geq 2.7$,		
Neutral Hadron Fraction	<0.99	<0.9
Neutral EM Fraction	<0.99	<0.9
Number of Constituents	>1	>1
In addition for $-2.4 \leq \eta \geq 2.4$,		
Charged Hadron Fraction	>0	>0
Charged Multiplicity	>0	>0
Charged EM Fraction	<0.99	–

Jet energy corrections are derived from simulation studies so that the average measured energy of jets matches that of particle level jets. In situ measurements of the p_{T} balance in dijet, photon+jet, leptonically decaying Z+jet, and multijet events are used to determine any residual differences between the jet energy scale in data and in simulation, and appropriate corrections are made to the jet p_{T} [139].

A subset of these AK4 PF jets originating from b-quarks are identified using the medium working point of the DeepCSV b-tagging algorithm [152], where the term "CSV" stands for combined secondary vertex. The DeepCSV is a multiclassifier deep neural network trained to discriminate the heavy-flavor jets, i.e. originating from bottom or charm quarks from the light flavor jets. The

Table 5.9: Summary of jet identification requirements in 2018.

Variable	$\|\eta\| \leq 2.6$
Neutral Hadron Fraction	<0.9
Neutral EM Fraction	<0.9
Number of Constituents	>1
Muon fraction	<0.8
Charged Hadron Fraction	>0
Charged Multiplicity	>0
Charged EM Fraction	<0.8
Number of neutral particles	–

algorithm can also identify jets containing two b-hadrons in Lorentz-boosted event topologies. The input variables to the training include properties of selected tracks, the secondary vertices and jets in the event. The medium working point corresponds to minimum DeepCSV cuts of 0.6321, 0.4941, and 0.4184, in 2016, 2017, and 2018, respectively. This results in an efficiency of 68% for the correct identification of a b-jet, with a 1% probability of misidentifying a light-flavor jet. Furthermore, b-tagging efficiency scale factors are applied to all jets using *method 1(a)* as per the physics object group BTV recommendations [153, 154].

5.2.8 Missing transverse energy

The vector $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$ is defined as the negative vector p_{T} sum of all the PF candidates in an event, and its magnitude is denoted as $p_{\mathrm{T}}^{\mathrm{miss}}$ [155]. The pileup-per-particle identification (PUPPI) algorithm [156] is applied to reduce the pileup dependence of the $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$ observable. This is done by assigning probabilities to each PF candidate to originate from the PV, according to local shape variable that can distinguish between parton shower-like radiation from pileup-like particles.

All charged particles originating from pileup vertices are assigned a weight of zero, and all charged particles from the PV are assigned a weight of one. The four-momentum of all particles are rescaled by its weight, i.e. $\mathrm{p}_i \rightarrow \mathrm{w}_i \times \mathrm{p}_i$. Particles with value of weights below a threshold (w_{cut}) or with small rescaled p_{T} are discarded. The surviving rescaled particles are then used in the calculation of the $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$. The $p_{\mathrm{T}}^{\mathrm{miss}}$ is also modified to account for corrections to the energy scale of the reconstructed jets in the event.

5.3 Important kinematic quantities

Armed with identified objects and their four-momenta, several kinematic quantities per event can be constructed. Here I describe quantities that are used in this multilepton analysis:

- **Scalar momentum sums:** We define L_{T} as the scalar p_{T} sum of all charged leptons that constitute the channel. For example, in the 4L channel, L_{T} is calculated from the leading four light leptons in p_{T}, while for the 3L1T channel, it is calculated from the three light leptons and the leading τ_{h}. We define H_{T} as the scalar p_{T} sum of all jets. Additionally, the scalar sum of L_{T}, H_{T}, and $p_{\mathrm{T}}^{\mathrm{miss}}$ is defined as S_{T}. The quantity L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ is also of interest.

- **Charge and flavor combinations:** We count the number OSSFn as distinct opposite-sign (electric charge) same-flavor lepton pairs in an event. Specific lepton pairs are labeled as OSSF (opposite-sign, same-flavor) and OSDF (opposite-sign, different-flavor).

- **Invariant and transverse masses:** We define M_{ℓ} as the invariant mass of all leptons in the event, and M_{min} as the minimum invariant mass of all dilepton pairs in the event, irrespective of charge or flavor. Additionally, the invariant mass of leptons i and j is defined as M^{ij}. The transverse mass for a single lepton i is defined as $M_{\mathrm{T}}^{\mathrm{i}} = (2p_{\mathrm{T}}^{\mathrm{miss}}p_{\mathrm{T}}^{\mathrm{i}}[1 - \cos(\vec{p}_{\mathrm{T}}^{\mathrm{miss}}, \vec{p}_{\mathrm{T}}^{\mathrm{i}})])^{1/2}$, where $p_{\mathrm{T}}^{\mathrm{i}}$ is the p_{T} of lepton i. Similarly, $M_{\mathrm{T}}^{\mathrm{ij}}$ is defined as the transverse mass calculated with the $p_{\mathrm{T}}^{\mathrm{miss}}$ and the resultant 4-momentum sum of lepton i and j. The lepton indices run over up to 4 leptons, in descending p_{T} order.

  We define the M_{OSSF} variable in a given event as the OSSF dielectron or dimuon mass closest to the Z boson mass at 91 GeV, and label events with M_{OSSF} within 15 GeV of the Z boson mass (76–106 GeV mass window) as OnZ.

  Additionally, the p_{T} of the M_{OSSF} lepton pair is defined as $p_{\mathrm{T}}^{\mathrm{OSSF}}$.

- **Angular quantities:** We define ΔR_{min} as the minimum ΔR between all the dilepton pairings in an event, irrespective of charge or flavor. Similarly, $\Delta R_{\mathrm{min}}^{\tau_{\mathrm{h}}}$ is defined as the minimum ΔR between any dilepton pair, where at least one of the leptons is a τ_{h} candidate. The quantities $\Delta\phi^{\mathrm{ij}}$ and $\Delta\eta^{\mathrm{ij}}$ are defined as the azimuthal angle or pseudorapidity difference between the i^{th} and j^{th} lepton, whereas $\Delta\phi^{\mathrm{i}}$ is defined to denote the opening azimuthal angle between lepton i and $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$.

- **Counts:** We define N_{j} as the multiplicity of jets and N_{b} as the multiplicity of b-tagged jets satisfying the selection criteria defined earlier.

Estimating the knowns...

Chapter 6

SM Backgrounds

As seen from the previous chapter, the two key features of leptons which helps in determining the source of their origin are its properties of "relative isolation" and "displacement". Figure 6.1 outlines a broad classification of leptons on the basis of these two properties.

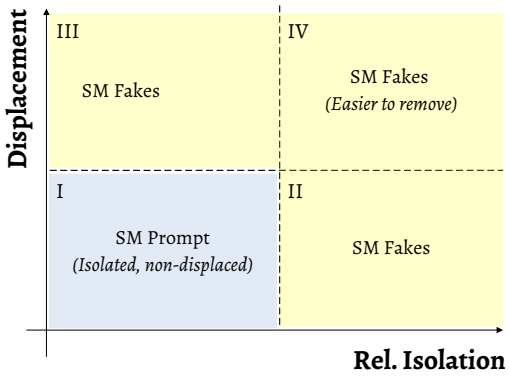

Figure 6.1: Classification of leptons from different sources. The x-axis is the relative isolation of leptons which is calculated as the ratio of the scalar sum of p_{T} of other particles in a cone of certain radius to the lepton p_{T}. The y-axis is the displacement of lepton with respect to the PV.

Leptons falling in box I, i.e. with small values of relative isolation and displacement, are typically produced from the decay of SM gauge bosons or from the leptonic decays of the τ lepton, or from promptly decaying BSM particles. The identification criteria described in Chapter 5 is specifically designed to target the selection of these *prompt* leptons. The SM processes such as WZ, ZZ, and $t\bar{t}Z$ production, which gives rise to prompt leptons in multilepton events are termed as **irreducible** backgrounds. A subdominant contribution to the irreducible background arises

from leptons originating from initial-state radiation (ISR) or final-state radiation (FSR) photons that convert asymmetrically (internal or external conversion) such that only one of the resultant leptons is reconstructed in the detector, or where an on-shell photon is misidentified as an electron. Such contributions, labelled as conversion background, are also considered to be a part of the irreducible background. These are primarily estimated using simulation, after correcting the MC for any residual differences with data. More details will follow in Section 6.1.

Leptons which have very large values of displacement and relative isolation, i.e. falling in box IV, are mostly removed by the stringent identification criteria. However, leptons which have either moderate values of displacement but poorly isolated (box II) or are isolated but have large displacements (box III) barely pass selections. W+jets, DY+jets, $t\bar{t}$+jets, and other such processes contribute via leptons originating from semi-leptonic heavy flavor decays within jets or from other misidentified detector signatures. Misidentification of muons often happens from hadron punch-throughs in the muon spectrometers, whereas photons produced from the decay of pions and associated to a nearby charged particle track gets misidentified as an electron. Hadron jets with low number of constituents or those that are boosted often resembles a hadronically-decaying τ lepton. All these are collectively labelled as *fake* or *misidentified* leptons, and such contributions constitute the **reducible** backgrounds. Reducible backgrounds may suffer from low effective luminosity of the simulation and therefore may not adequately describe data in the desired phase space. In this thesis, a data driven approach is used to estimate reducible background contributions that contain misidentified leptons, and is described in Section 6.2.1.

6.1 Irreducible background

The simulations are heavily tuned to mimic the observations, by emulating the full detector geometry and the particle-material interaction. However, there can still be some inconsistencies due to instantaenous detector conditions while data-taking, or a mismodeling of the generator-level kinematic properties, or simply a mismatch between the theory and experimental cross sections. To account for this, we followed a standard procedure of applying corrections to the MC as described below. Note that to suppress contributions to the multilepton phase space from low-mass quarkonia resonances such as J/ψ and Υ, and final-state radiation from leptons, events with $M_{\min} < 12$ GeV, $\Delta R_{\min} < 0.2$, and $\Delta R^{\tau_h}_{\min} < 0.5$ are rejected.

For a given dominant irreducible SM process (WZ, ZZ, $t\bar{t}Z$, $Z\gamma$), we normalize the yield in a specific CR dominated by that process. The measured normalization factor is then used to scale the yield in all other CRs and SRs. This correction helps in modeling the important kinematic

quantities of the simulation.Various other processes, such as $t\bar{t}W$, triboson (WWW, WWZ, WZZ, ZZZ), and top or vector boson associated Higgs production (tH, VH, tHW, $t\bar{t}HH$, and so on) can also yield prompt multilepton signatures. These contributions are generally suppressed due to lower production cross-sections, and are estimated from the MC simulation without normalizing in dedicated CRs, by simply assigning a relative 50% uncertainty on the theory cross section.

The CRs are designed such that they have fractionally very little contamination from the BSM signals, so that we don't normalize any potential signatures of new phenomena along with the SM backgrounds. Furthermore, all the CRs are explicitly removed from the subsequent search for BSM phenomena. A summary of all the CR definitions used in this analysis is provided in Table 6.1.

Table 6.1: A summary of control regions for the irreducible SM processes ZZ, WZ, $Z\gamma$, and ZZ. The $p_{\mathrm{T}}^{\mathrm{miss}}$, M_{T}, the minimum 3L lepton $p_{\mathrm{T}}(p_{\mathrm{T}}^{3})$, and S_{T} are in units of GeV. The last column "Purity" is the relative percentage of the desired background contribution from the total number of events.

CR name	OSSFn	M_{OSSF}	N_{b}	$p_{\mathrm{T}}^{\mathrm{miss}}$	M_{T}	p_{T}^{3}	Other selections	Purity (%)
4L ZZ	OSSF2	Double-OnZ	0	–	–	–	–	99
3L WZ	OSSF1	Single-OnZ	0	< 125	50–150	> 20	–	75
3L $Z\gamma$	OSSF1	BelowZ	0	–	–	–	Trilepton mass Single-OnZ	70
3L $t\bar{t}Z$	OSSF1	Single-OnZ	≥ 1	< 125	< 150	> 20	$N_{\mathrm{j}} > 2$, $S_{\mathrm{T}} > 350$	60

6.1.1 ZZ CR

For the ZZ CR, we select events where each Z boson decays to an opposite-sign (OS) and same flavor (SF) pair of light lepton (e^+e^- or $\mu^+\mu^-$). Thus, ZZ CR is defined as an 4L OSSF2 event with an additional requirement on the mass (M_{OSSF}) of both the OSSF pairs to improve the purity of the selection. In case of ambiguity in 4L OSSF2 events with four electrons or muons i.e. $e^+e^-e^+e^-$ or $\mu^+\mu^-\mu^+\mu^-$ events, M_{OSSF} is chosen such that it gives the maximum number of nonoverlapping OSSF pairs with masses within the Z boson mass window. 4L OSSF2 events with M_{OSSF} of both pairs within 15 GeV of the Z boson mass (91 GeV) i.e. falling in the window 76–106 GeV are labeled as Double-OnZ events, and are used as final ZZ CR along with the requirement of no b-tagged jet (N_{b}=0). The invariant mass distribution of the best OSSF pair, i.e. with M_{OSSF} closest to the Z boson window is shown in Figure 6.2 left, while the L_{T} distribution is shown on the right. These are made with the combined 2016–2018 data set in the 4L ZZ CR, and have statistical uncertainties only.

As can be seen from the last column of the Table 6.1, CR for the ZZ background is the purest in

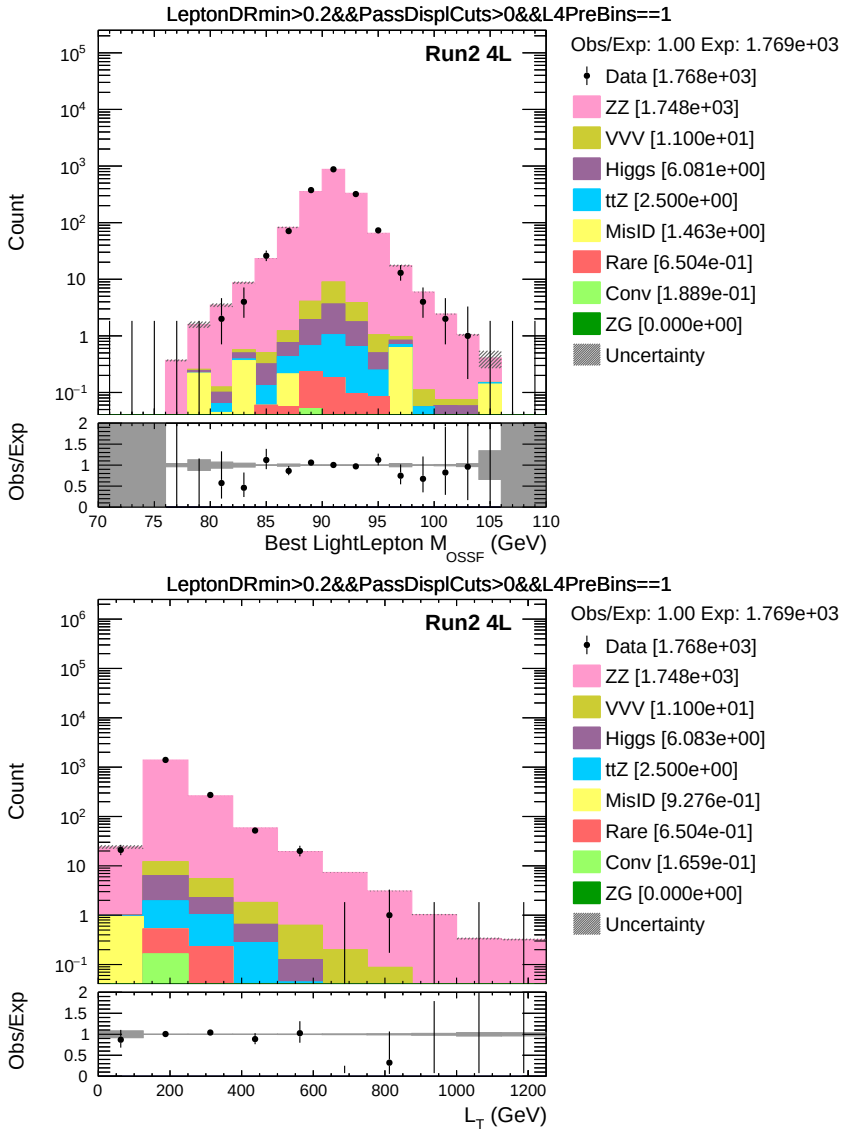

Figure 6.2: The distributions of invariant mass of the best OSSF pair (left), i.e. with M_{OSSF} closest to the Z boson window and L_{T} (right) in 4L ZZ CR events for the combined 2016–2018 data set. The rightmost bin contains the overflow events in each distribution. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represent statistical uncertainties only.

terms of contamination from other SM processes. Hence, we determine the normalization for the ZZ process first, which is then propagated to all other CRs. The normalization factors are measured to be 1.05 ± 0.05, 0.97 ± 0.04, and 1.00 ± 0.04 in 2016, 2017, and 2018, respectively. Differences in the description of the jet multiplicity distribution are used to reweight the ZZ samples in 0, 1, 2

and $\geq$3 jet bins, because of a known deficiency of POWHEG samples that include only up to 1 hard jets at the matrix level. The ZZ samples are also reweighted as a function of the generator-level visible diboson p_{T} to match the MC distribution to that of the data, as this yields a better agreement across different generators (POWHEG versus MADGRAPH5_AMC@NLO) as well as an improved description of other leptonic and hadronic quantities of interest. Typical inclusive normalization uncertainties, and those due to the diboson $p_{\mathrm{T}}$and jet multiplicity modeling are 4–5% and 5–30%, respectively.

Figure 6.3 shows the distribution of the visible diboson p_{T} in the 4L ZZ CR from the combined 2016–2018 data set. For better visualization, WZ and ZZ backgrounds are combined into one bundle as "VV", $t\bar{t}Z$ and $t\bar{t}W$ are combined into one bundle as "ttV", all the triboson and top or vector boson associated Higgs production processes are combined into one bundle as "Rare", and finally $Z\gamma$ and all other irreducible processes giving rise to conversion leptons are combined into one bundle as "Conv.". The "MisID" bundle in the figures refer to the reducible misidentified background, and the estimation is explained in Section 6.2. The figure is shown after applying the derived normalization constants, along with the systematic uncertainties as will be explained in Section 6.3; and thus an excellent agreement can be seen with the observed data.

6.1.2 WZ CR

Following the treatment of the ZZ MC samples, similar data driven corrections are applied to the WZ MC sample to improve the modeling of this background component. These are measured in 3L OSSF1 Single-OnZ events, with additional requirements on the M_{T} variable. For the WZ CR, we choose M_{T} for the lepton which is not part of the M_{OSSF} pair. In events with three electrons or three muons, the M_{OSSF} and M_{T} variables are chosen simultaneously so that the event is Single-OnZ, and M_{T} is in the range 50–150 GeV, where this is kinematically possible. We only select events with N_{b}=0, and those which have low missing energy i.e. $p_{\mathrm{T}}^{\mathrm{miss}} < 125$ GeV, in order to not exploit potential signal regions. Finally, we also increase the p_{T} threshold of the softest lepton to 20 GeV, which reduces the contamination from misidentified lepton backgrounds, and raises the purity of WZ CR to 75%. The M_{T} of the non-OnZ lepton and the distribution of number of jets is shown in Figure 6.4 left and right, respectively. These are made with the combined 2016–2018 data set in the 3L WZ CR, and have statistical uncertainties only.

The normalization factors are measured to be 0.89 ± 0.03, 0.89 ± 0.05 0.91 ± 0.03, in 2016, 2017, and 2018, respectively. The WZ MC is reweighted in 0, 1, 2 and $\geq$3 jet bins to account for the known deficiency of MADGRAPH5_AMC@NLO samples that include only up to 2 hard

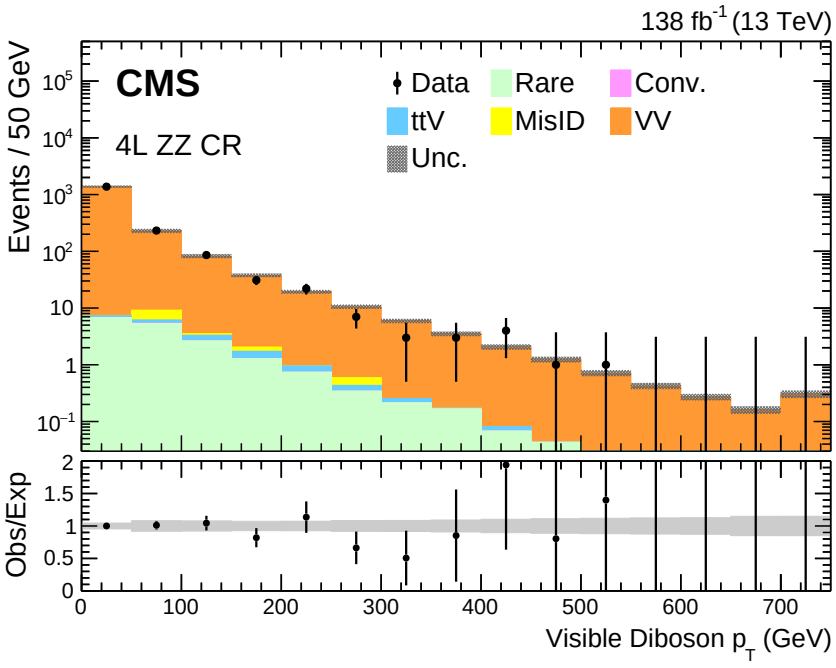

Figure 6.3: The distribution of visible diboson p_{T} in 4L ZZ CR events for the combined 2016–2018 data set. The rightmost bin contains the overflow events in each distribution. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represent the sum of statistical and systematic uncertainties, covered in Sec 6.3, in the SM background predictions.

jets at the matrix level. The WZ samples are also reweighted as a function of the generator-level visible diboson p_{T} to match the MC distribution to that of the data. Typical inclusive normalization uncertainties, and those due to the diboson $p_{\mathrm{T}}$and jet multiplicity modeling are 3–5% and 5–15%, respectively. Figure 6.5 shows the S_{T} distribution in the 3L WZ CR from the combined 2016–2018 data set, after applying the derived normalization constants.

6.1.3 $Z\gamma$ CR

An average correction factor as a ratio of the data to simulation yield is measured in a dedicated CR for the $Z\gamma$ process. The CR events are chosen from the $Z\gamma$ simulation which consists of the matrix-level as well as the ISR and FSR photons. These are events where a FSR photon is produced in association with a leptonically-decaying Z boson i.e. $\mathrm{Z} \to \ell\ell$+γ, such that $M_{\mathrm{OSSF}} < 76$ GeV (also known as "BelowZ" events). The photon converts asymmetrically to a pair of leptons, where the softer lepton is not reconstructed in the event, and the trilepton mass falls in the Z window. The

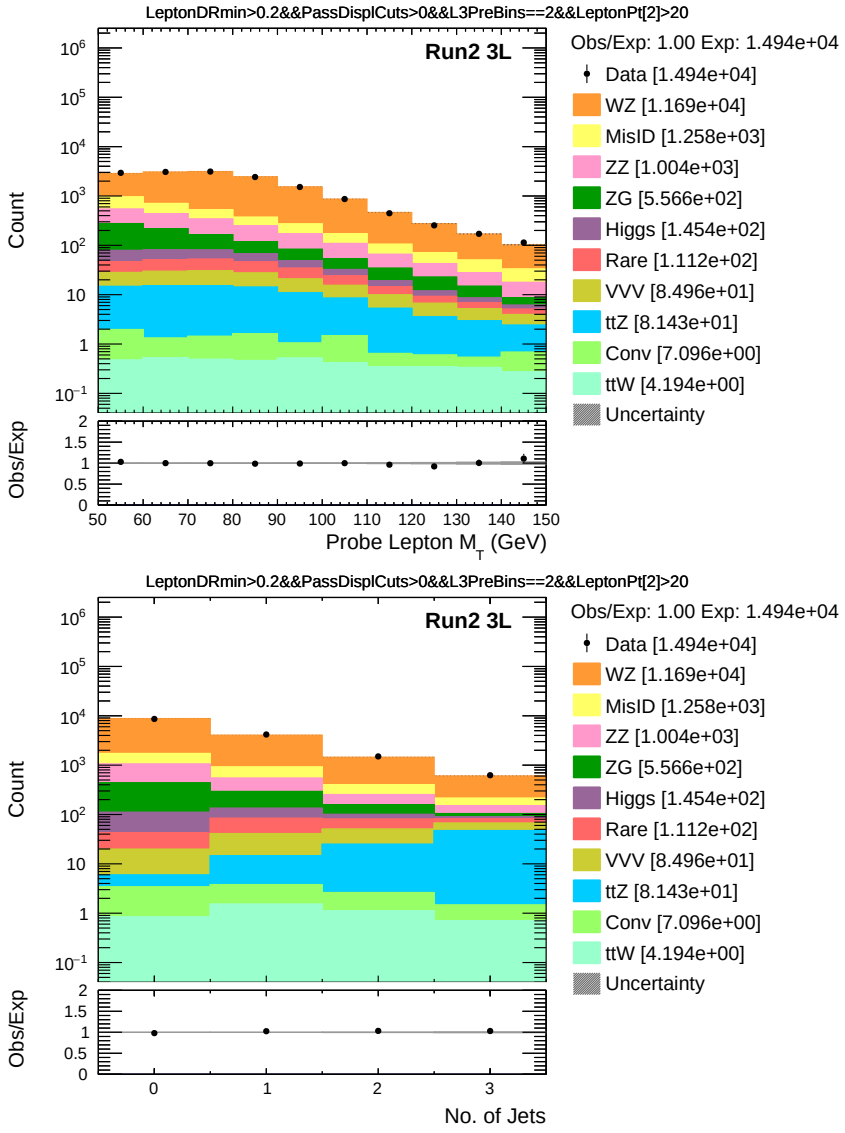

Figure 6.4: The distributions of M_{T} of the non-OnZ lepton (left) and number of jets (right) in 3L WZ CR events for the combined 2016–2018 data set. The rightmost bin contains the overflow events in each distribution. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represent statistical uncertainties only.

distributions of $p_{\mathrm{T}}^{\mathrm{miss}}$ and number of electrons are shown in Figure 6.6 left and right, respectively. These are made with the combined 2016–2018 data set in the 3L $Z\gamma$ CR, and have statistical uncertainties only.

The normalization factors are measured to be 0.81 ± 0.03, 0.87 ± 0.06, and 0.96 ± 0.05, in 2016, 2017, and 2018, respectively. Typical inclusive normalization uncertainty for $Z\gamma$ background is

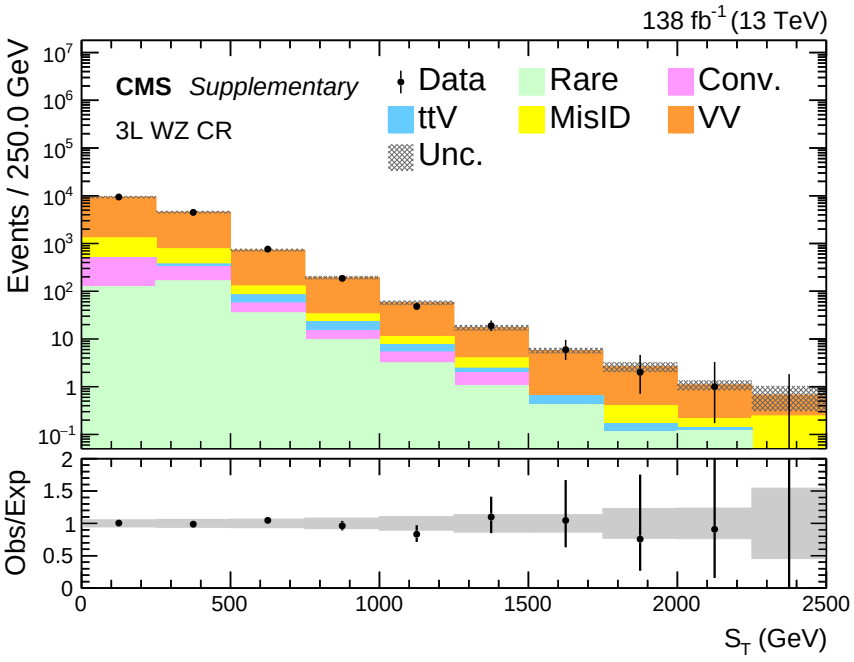

Figure 6.5: The S_{T} distribution in 3L WZ CR events for the combined 2016–2018 data set. The rightmost bin contains the overflow events in each distribution. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represent the sum of statistical and systematic uncertainties, covered in Sec 6.3, in the SM background predictions.

10%. Figure 6.7 shows the ΔR_{min} distribution in the 3L $Z\gamma$ CR from the combined 2016–2018 data set, after applying the derived normalization constants.

6.1.4 $t\bar{t}Z$ CR

Finally, an average correction factor, similar to $Z\gamma$ normalization, is measured for $t\bar{t}Z$ process in 3L OSSF1 Single-OnZ events with a requirement of N_{b} $\geq$1 coming from the decay of the two top quarks, and total number of jets (N_{j}) to be greater than 2 to account for the hadronic decay of one of the W's from top quark. There is a lot of leptonic and hadronic activity in 3L OSSF1 Single-OnZ, N_{b} $\geq$1, and N_{j} >2 events, plus missing energy attributing to the neutrino from the leptonically-decaying W boson. Hence, to improve the purity of $t\bar{t}Z$ CR, we require $S_{\mathrm{T}} > 350$ GeV but restrict to events where $p_{\mathrm{T}}^{\mathrm{miss}} < 125$ GeV and $M_{\mathrm{T}} < 150$ GeV to keep it orthogonal to signal regions. Increasing the p_{T} threshold of the softest lepton to 20 GeV reduces the contamination from misidentified lepton backgrounds, as in the WZ CR. The distributions of M_{T} of the non-OnZ

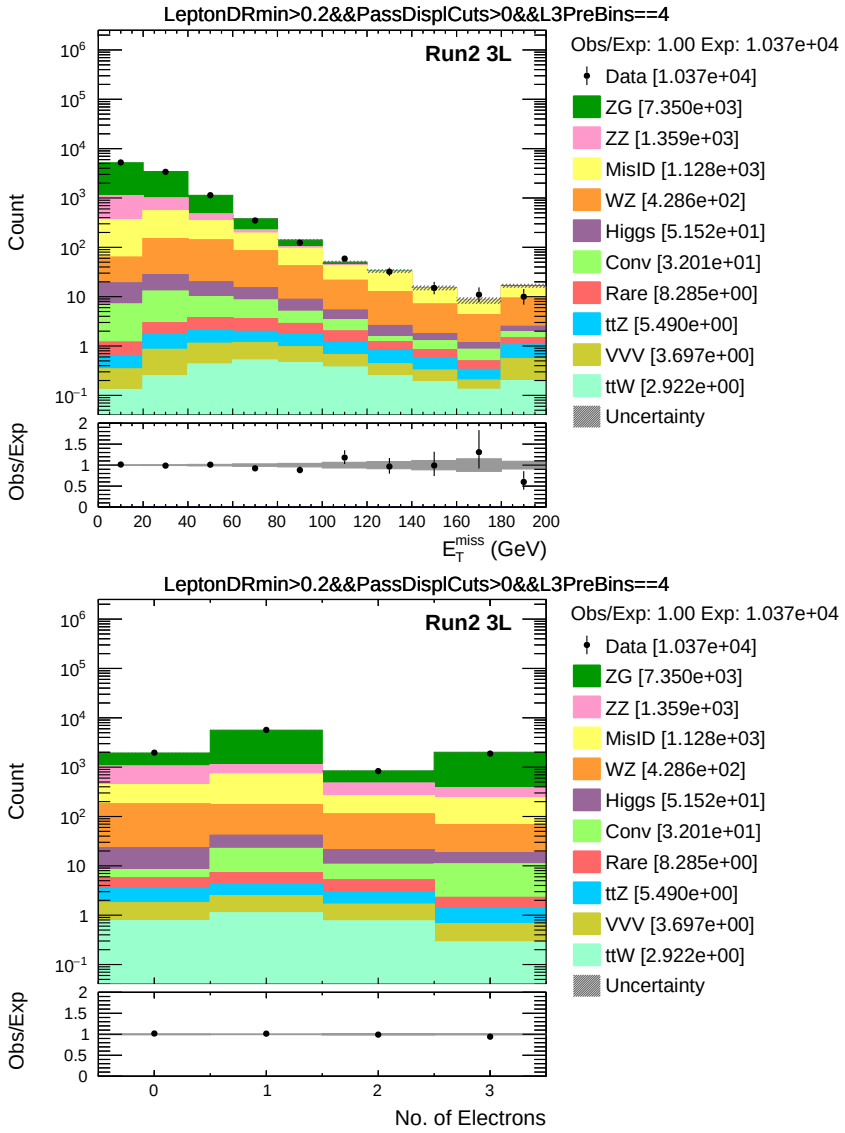

Figure 6.6: The distributions of $p_{\mathrm{T}}^{\mathrm{miss}}$ (left) and number of electrons (right) in 3L $Z\gamma$ CR events for the combined 2016–2018 data set. The rightmost bin contains the overflow events in each distribution. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represent statistical uncertainties only.

lepton and number of b-tagged jets are shown in Figure 6.8 left and right, respectively. These are made with the combined 2016–2018 data set in the 3L $t\bar{t}Z$ CR, and have statistical uncertainties only.

The normalization factors are measured to be 0.80 ± 0.22, 1.35 ± 0.22, and 1.28 ± 0.21, in 2016, 2017, and 2018, respectively. Typical inclusive normalization uncertainty for $t\bar{t}Z$ background is

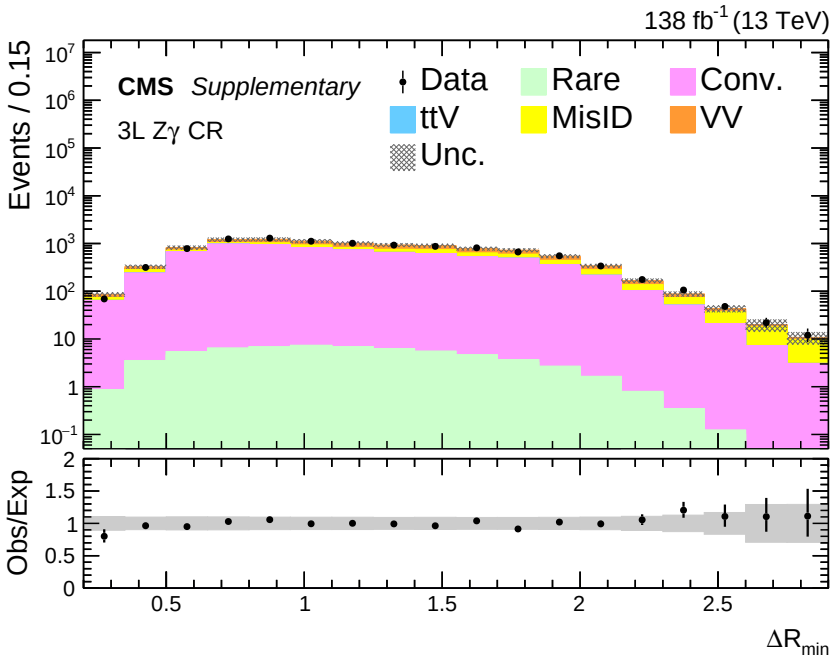

Figure 6.7: The ΔR_{min} distribution in 3L $Z\gamma$ CR events for the combined 2016–2018 data set. The rightmost bin contains the overflow events in each distribution. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represent the sum of statistical and systematic uncertainties, covered in Sec 6.3, in the SM background predictions.

15–25%. Figure 6.9 shows the H_{T} distribution in the 3L $t\bar{t}Z$ CR from the combined 2016–2018 data set after applying the derived normalization constants.

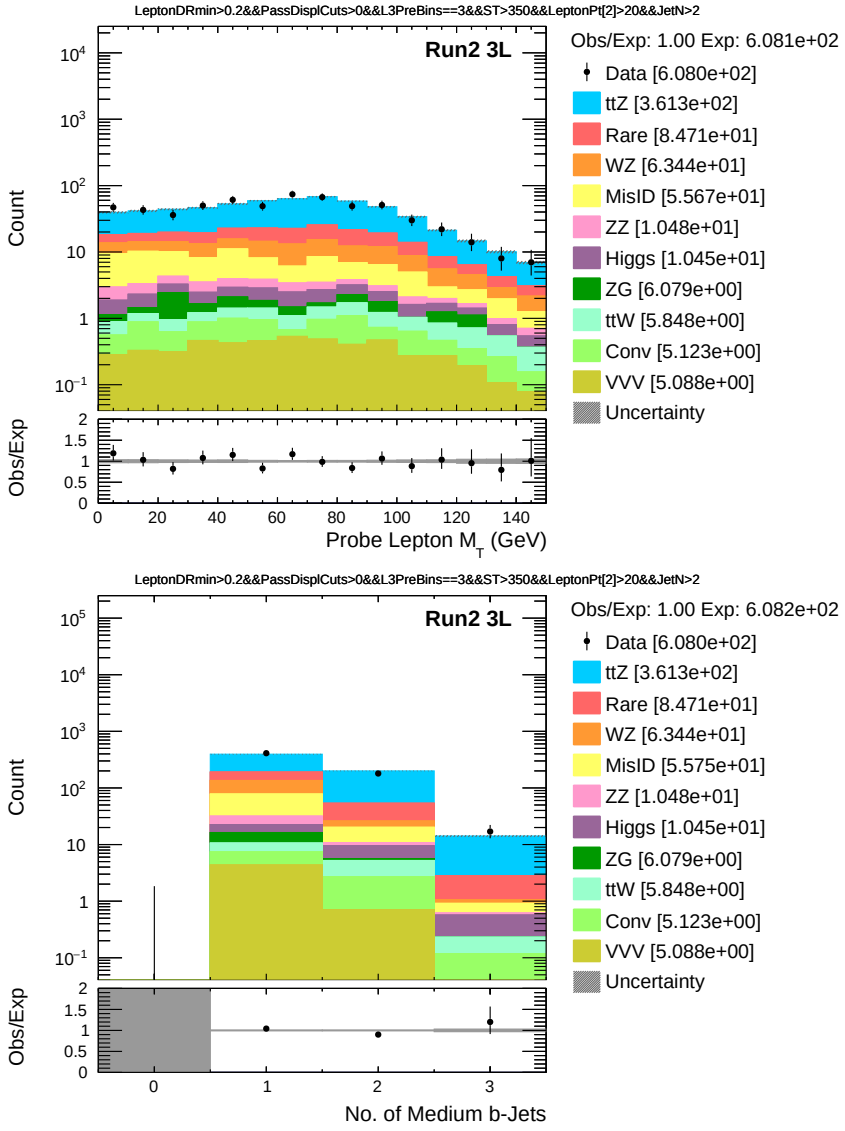

Figure 6.8: The distributions of M_{T} (left) of the non-OnZ lepton and number of b-tagged jets (right) in 3L $t\bar{t}Z$ CR events for the combined 2016–2018 data set. The rightmost bin contains the overflow events in each distribution. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represent statistical uncertainties only.

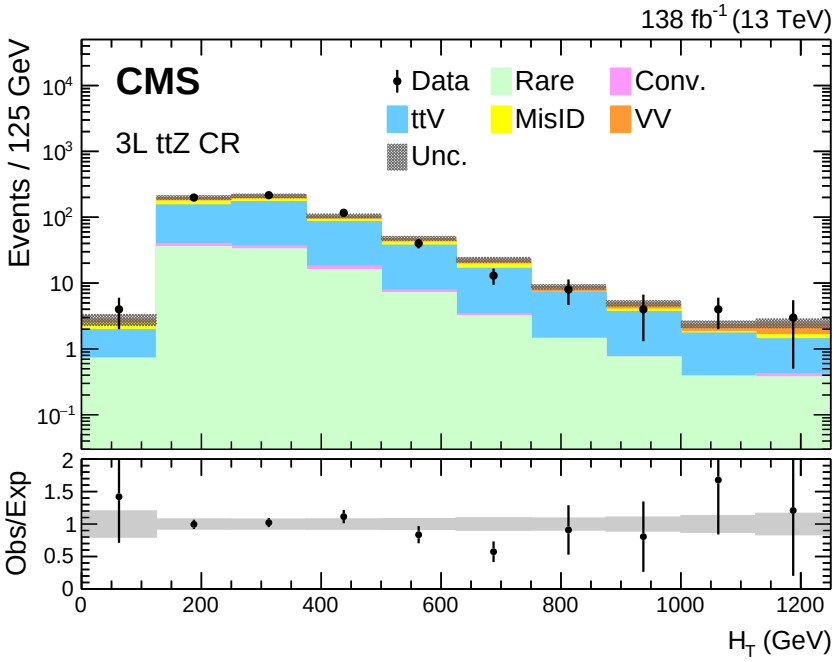

Figure 6.9: The H_{T} distribution in 3L $t\bar{t}Z$ CR events for the combined 2016–2018 data set. The rightmost bin contains the overflow events in each distribution. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represent the sum of statistical and systematic uncertainties, covered in Sec 6.3, in the SM background predictions.

6.2 Reducible background

6.2.1 Data driven matrix method

Misidentified lepton backgrounds (MisID) are estimated via a matrix method [157] using fully deterministic quantities from observed data. It is a background estimation technique that connects the underlying reality of the leptons with their observed properties, under a simple assumption that the probabilities with which prompt and fake leptons pass a tight ID selection given that they satisfy a loose ID selection, prompt (p) and fake (f) rates respectively, are universal and can be described as a function of the lepton and event dependent parameters. This assumption allows the measurement of these rates in background dominated control regions and then their application to a signal region.

To understand the mathematical formulation of the matrix method, let's consider a simpler example of a single lepton event first. Let N_P and N_F represent the true number of prompt and fake leptons, respectively, and N_L and N_T represent the observed number of leptons passing the loose and tight ID selection, respectively. Then, according to the definition of prompt (p) and fake (f) rates, as explained in the previous paragraph, we can write a relation between N_P, N_F, N_L, and N_T as follows:

$$\begin{aligned} N_T &= p \times N_P + f \times N_F \\ N_L &= \hat{p} \times N_P + \hat{f} \times N_F \end{aligned} \tag{6.1}$$

where $\hat{p}$=(1-p) and $\hat{f}$=(1-f). Writing the Eqn. 6.1 in a one-dimensional matrix form, we get:

$$\begin{bmatrix} N_T \\ N_L \end{bmatrix} = \begin{bmatrix} p & f \\ \hat{p} & \hat{f} \end{bmatrix} \begin{bmatrix} N_P \\ N_F \end{bmatrix} \tag{6.2}$$

Here, N_T and N_L are the observables of an event. While N_P is known from simulation, N_F is the unknown quantity that is estimated. Hence, inverting the matrix Eqn. 6.2:

$$\begin{bmatrix} N_P \\ N_F \end{bmatrix} = \frac{1}{(p-f)} \begin{bmatrix} \hat{f} & -f \\ -\hat{p} & p \end{bmatrix} \begin{bmatrix} N_T \\ N_L \end{bmatrix} \tag{6.3}$$

Following the same analogy as in the matrix Eqn. 6.2, we can write the two-dimensional matrix for dilepton events as:

$$\begin{bmatrix} N_{TT} \\ N_{TL} \\ N_{LT} \\ N_{LL} \end{bmatrix} = \begin{bmatrix} p_1p_2 & p_1f_2 & f_1p_2 & f_1f_2 \\ p_1\hat{p}_2 & p_1\hat{f}_2 & f_1\hat{p}_2 & f_1\hat{f}_2 \\ \hat{p}_1p_2 & \hat{p}_1f_2 & \hat{f}_1p_2 & \hat{f}_1f_2 \\ \hat{p}_1\hat{p}_2 & \hat{p}_1\hat{f}_2 & \hat{f}_1\hat{p}_2 & \hat{f}_1\hat{f}_2 \end{bmatrix} \begin{bmatrix} N_{PP} \\ N_{PF} \\ N_{FP} \\ N_{FF} \end{bmatrix} \tag{6.4}$$

Hence, the observed events with two tight leptons can be broken down into the following equation with four independent underlying terms:

$$N_{TT} = p_1p_2 \times N_{PP} + p_1f_2 \times N_{PF} + f_1p_2 \times N_{FP} + f_1f_2 \times N_{FF} \tag{6.5}$$

The first term of the Eqn. 6.5 is none other than the irreducible background contribution estimated from simulation. The other three terms can be derived by the inverse of the matrix Eqn. 6.4:

$$\begin{bmatrix} N_{PP} \\ N_{PF} \\ N_{FP} \\ N_{FF} \end{bmatrix} = \frac{1}{(p_1 - f_1)(p_2 - f_2)} \begin{bmatrix} \hat{f}_1\hat{f}_2 & -\hat{f}_1f_2 & -f_1\hat{f}_2 & f_1f_2 \\ -\hat{f}_1\hat{p}_2 & -\hat{f}_1p_2 & f_1\hat{p}_2 & -f_1p_2 \\ -\hat{p}_1\hat{f}_2 & \hat{p}_1f_2 & p_1\hat{f}_2 & -p_1f_2 \\ \hat{p}_1\hat{p}_2 & -\hat{p}_1p_2 & -p_1\hat{p}_2 & p_1p_2 \end{bmatrix} \begin{bmatrix} N_{TT} \\ N_{TL} \\ N_{LT} \\ N_{LL} \end{bmatrix} \tag{6.6}$$

Substituting the values of N_{PF}, N_{FP}, and N_{FF} from the matrix Eqn. 6.6 to the Eqn. 6.5, and solving for N_{TT} gives the total misidentified lepton background contribution in dilepton events.

To summarize, the misidentified lepton backgrounds in the three-lepton and four-lepton channels can be estimated by similarly expanding the above matrix equations to three-dimensional and four-dimensional matrices, respectively. Since this analysis relies on the use of isolated single light lepton triggers, it is implicitly assumed that at least one triggering lepton is a prompt lepton. Although background contributions with multiple fake leptons are rare, the matrix method can efficiently predict such contributions with up to two (three) simultanous misidentified leptons in three (four) lepton events.

The primary ingredient of the matrix method is the determination of the prompt and fake rates for each lepton flavor. Next, I will proceed onto describing the measurement of these prompt and fake rates in detail, starting with the tau leptons.

6.2.2 Misidentified tau leptons

In this multilepton analysis, a major challenge was the construction of final states enriched with τ_h leptons. In final states with upto three tau multiplicity, the background composition changes drastically. This is shown in Figure 6.10 with the help of pie charts for the 2L1T (left) and 1L2T (middle) channels, and a combined pie chart (right) for the inclusive 4-lepton channels with hadronic taus i.e. 3L1T, 2L2T, and 1L3T. We will call 3L1T, 2L2T, and 1L3T channels collectively as "rare tau" channels from this point onwards.

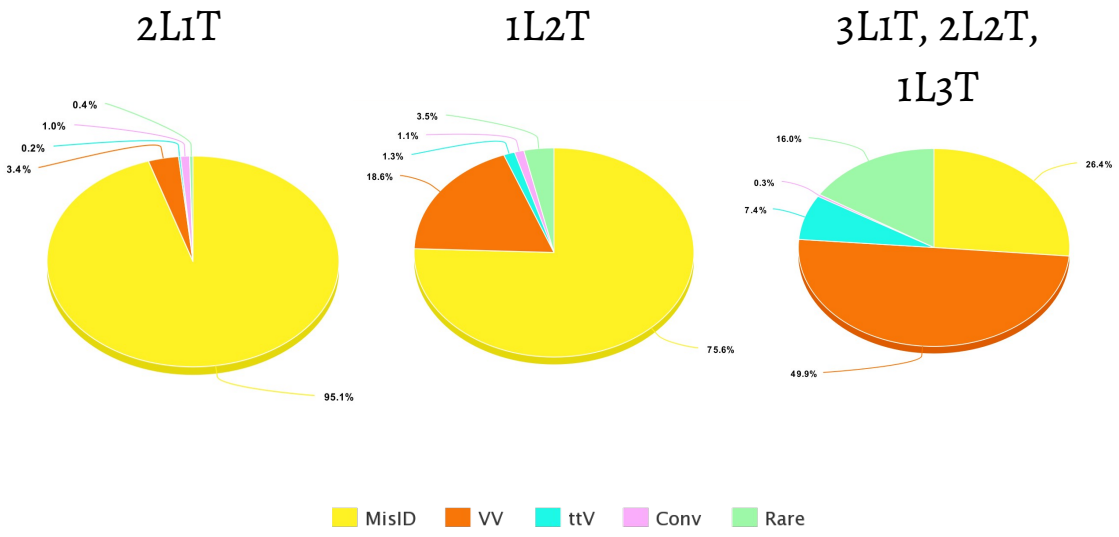

Figure 6.10: Pie charts illustrating the background composition in the 2L1T (left), 1L2T (middle), and the 4-lepton channels with τ_h leptons i.e. 3L1T, 2L2T, and 1L3T (right).

As can be seen from the Figure 6.10 (left), 2L1T channel is almost completely dominated by the misidentified lepton background. This is primarily arising from the processes DY+jets, $t\bar{t}$+jets and WW+jets, where the light leptons are prompt while the jet object is misidentified as a single fake τ_h lepton.

In the 1L2T channel, the contribution of misidentified background is slightly reduced with respect to the 2L1T channel, and the nature of the misidentified lepton events also vary significantly. While the lepton multiplicity of the two channels is exactly the same, the majority of single τ_h fakes from DY+jets process in 2L1T is now subdued for the 1L2T channel due to the leptonic decay of one of the taus in $Z \rightarrow \tau\ \tau$ events. However, the cross section of DY+1jet is almost the

same as that of W+2jets (Figure 3.1), and given the requirement of a prompt light lepton for the trigger purposes, we now get events with double τ_h fakes from the W+2jets process. Hence, the misidentified background for the 1L2T channel is an admixture of roughly an equal contribution of single τ_h fakes from DY+jets process and double τ_h fakes from W+jets process. The contribution of W+2jets process in 2L1T channel is not as appreciable as DY+1jet since the probability of a jet misidentified as tau lepton is much larger than that of the light leptons.

Finally, for the rare tau channels, we can see from Figure 6.10 (right) that 50% of the events are coming from the irreducible ZZ background which gives rise to prompt τ_h leptons. In this case, the leading misidentified background contribution arises from WZ+jets events where the jet gives rise to a single fake τ_h lepton.

An example event topology from the DY+jets (left), W+jets (middle), and ZZ (right) processes in the 2L1T, 1L2T, and the rare tau channels, respectively, is shown in Figure 6.11.

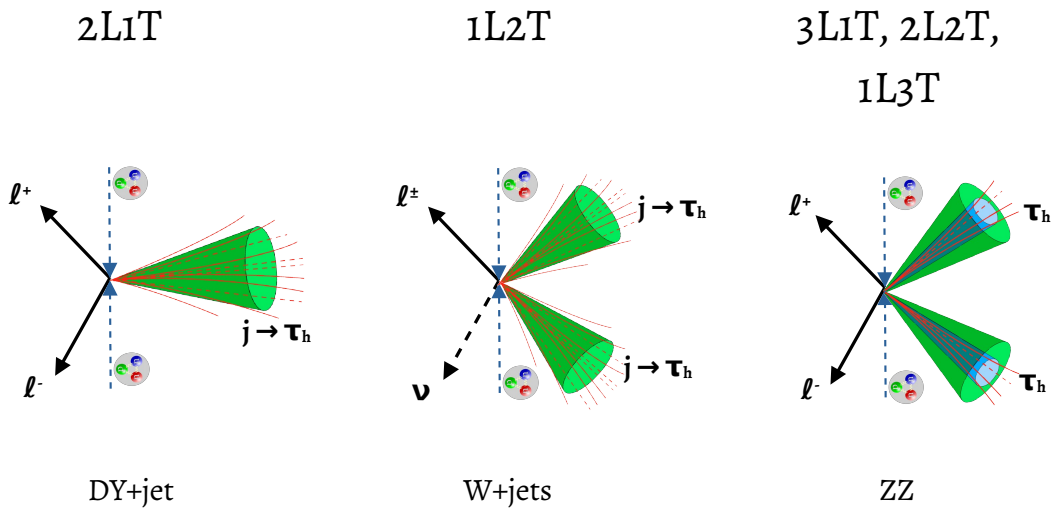

Figure 6.11: An example event topology from the DY+jets (left), W+jets (middle), and ZZ (right) processes in the 2L1T, 1L2T, and the rare tau channels, respectively.

6.2.2.1 Measurement of tau lepton prompt rates

Prompt rate of a τ lepton is the probability to pass the tight ID selection, given that it satisfies the loose ID selection when produced from the decay of SM gauge bosons (W,Z,h). It is measured using the "tag-and-probe" method in 1L1T OS ($e\tau_h$ and $\mu\tau_h$) events in DY MC, where reconstructed

Table 6.2: Prompt rate parametrizations for all lepton flavors. Individual rates and corrections are measured in orthogonal bins as specified by the binning schemes and as functions of the given variables. Corrections are normalized in order not to affect the mean rates.

	Primary		Correction	
	Binning scheme	Variable	Binning scheme	Variable
e prompt rate	$\|\eta\| : \{0, 1.5, 2.4\}$	p_{T}		
μ prompt rate	$\|\eta\| : \{0, 1.2, 2.4\}$	p_{T}		
τ_h prompt rate	$\|\eta\| : \{0, 1.5, 2.3\}^{\dagger}$	p_{T}	$\|\eta\| : \{0, 1.5, 2.3\}^{\dagger}$	$\|\eta\|$
	† in 1- and 3-prong separately			

leptons are kinematically matched to generator level prompt leptons ($\Delta R < 0.2$). The light lepton is chosen as the tag satisfying the tight ID selection as well as matches with the trigger object ($\Delta R < 0.2$), and the τ_{h} lepton is the probe.

The τ_{h} prompt rates are measured separately for the three years of data-taking, and are parametrized in tau hadronic decay modes i.e. 1-prong and 3-prong. They are also measured separately in the barrel ($|\eta| \leq 1.5$) and the endcap ($|\eta| > 1.5$) regions of the detector, for both the decay modes. The prompt rates are fitted linearly in τ_{h} p_{T}, to minimize statistical fluctuations, and are corrected for any η dependencies using a quadratic polynomial fit divided by the average tau prompt rate. Tau prompt rates for the three years of data-taking are shown in Figure 6.12 for the 1-prong τ_{h}, and in Figure 6.13 for the 3-prong τ_{h}.

We have measured the tau prompt rates in data using 1L1T events with $M_{\mathrm{T}} < 40$ GeV, $\Delta R_{\mathrm{min}} < 3.5$, $p_{\mathrm{T}}^{\mathrm{miss}} < 100$ GeV, and the mass of the opposite-sign opposite-flavor (OSOF) dilepton pair (e.g. $\mathrm{e}^{+}\tau_{\mathrm{h}}^{-}$ or $\mu^{-}\tau_{\mathrm{h}}^{+}$) in the range of 40-80 GeV. The M_{T} is computed with the light lepton and $p_{\mathrm{T}}^{\mathrm{miss}}$ vectors. The contributions of fake taus are estimated and subtracted using MC samples, and the dominant W+jets contribution is normalized to data in-situ using the $M_{\mathrm{T}} > 40$ GeV region in each bin where a prompt rate measurement is peformed. The final prompt rate measurements are found to be compatible with the DY MC based prompt rates within 15% in 2016, and within 5% in 2017 and 2018; this is taken as a systematic uncertainty on the measurement of prompt rate for the matrix-method.

Details of the parametrization for the τ_{h} leptons is given in Table 6.2. Prompt rates for τ_{h} leptons are about $50 - 70\%$ ($30 - 70\%$) for 1-prong (3-prong) taus, across all years and bins.

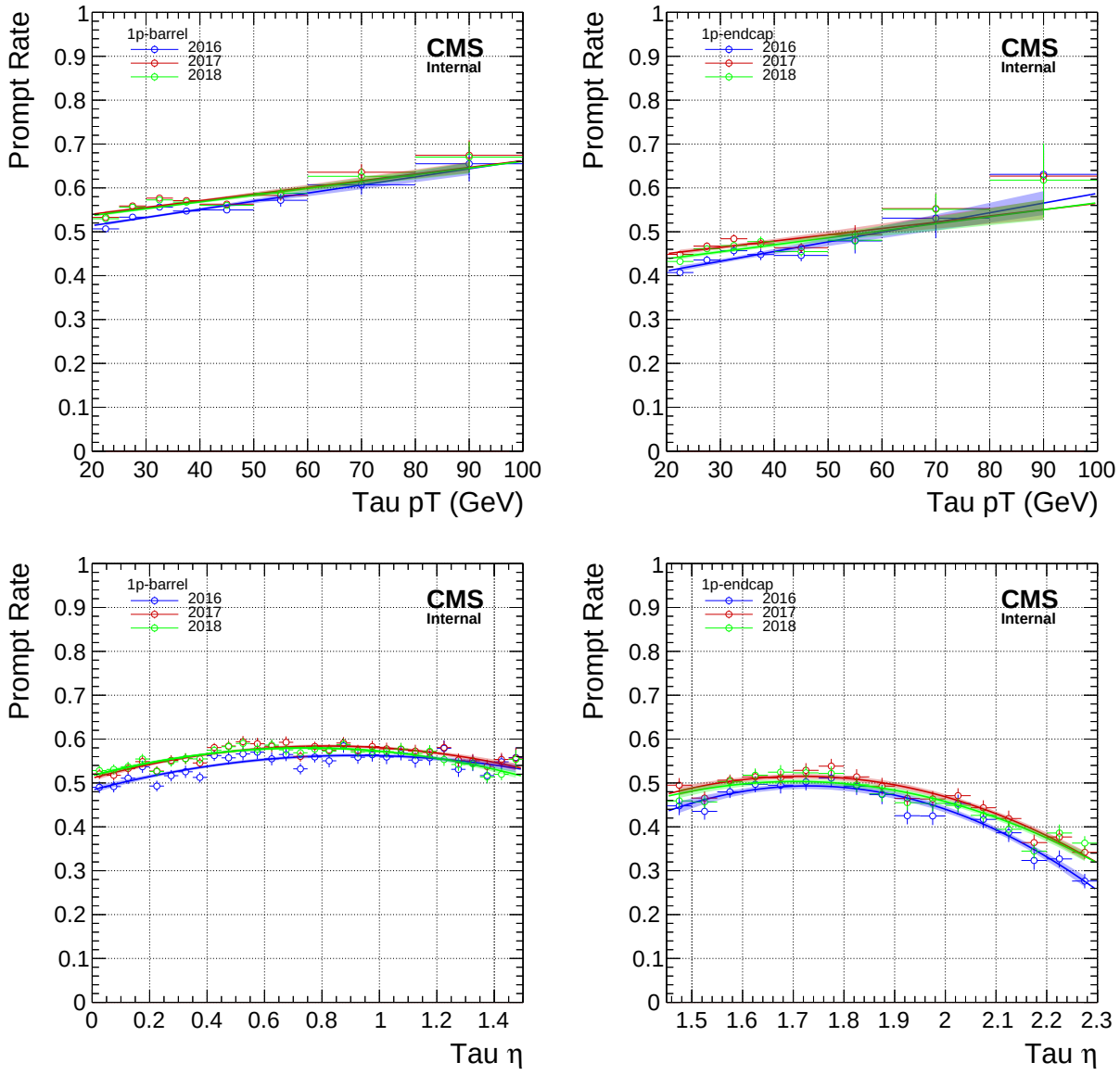

Figure 6.12: 1-prong τ_h prompt rates in the three years of data-taking. Rates for 2016, 2017, and 2018 are shown in blue, red, and green, respectively. The measurements are conducted in the 1L1T OS events in DY MC. The prompt rates have been parametrized as a function of the τ_h p_T, and additional correction factors are derived as a function τ_h $|\eta|$ for the same. The uncertainties are statistical only.Highest p_T bins include over-flows, and constant rate values are extrapolated beyond these.

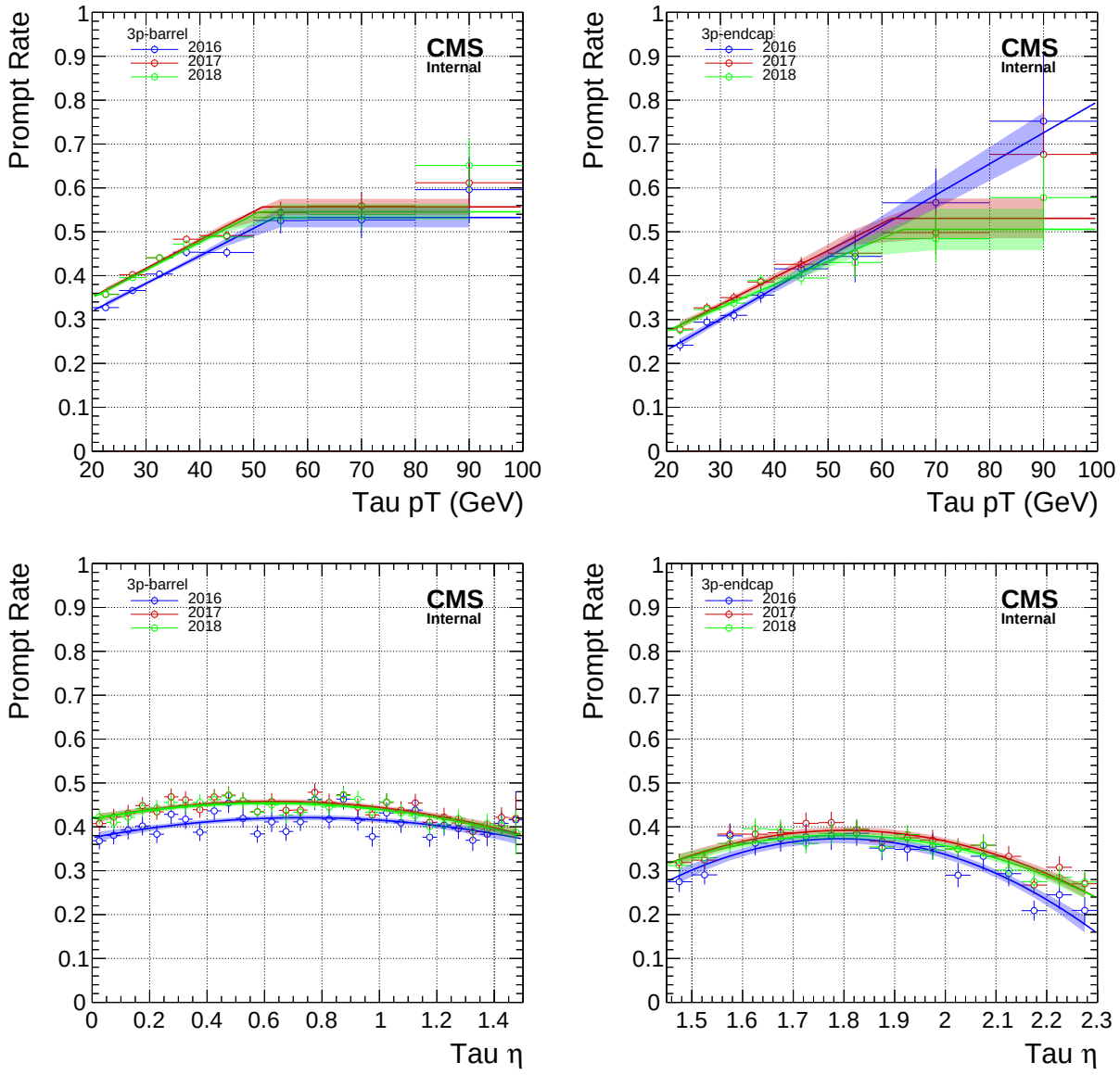

Figure 6.13: 3-prong τ_{h} prompt rates in the three years of data-taking. Rates for 2016, 2017, and 2018 are shown in blue, red, and green, respectively. The measurements are conducted in the 1L1T OS events in DY MC. The prompt rates have been parametrized as a function of the τ_{h} p_{T}, and additional correction factors are derived as a function τ_{h} $|\eta|$ for the same. The uncertainties are statistical only.Highest p_{T} bins include over-flows, and constant rate values are extrapolated beyond these.

6.2.2.2 Recoil-based parametrization

Fake rates of leptons are more closely related to the event topologies and the properties of the mother jet giving rise to the fake lepton than its own. This is because of the SM processes involved

behind the production of the fake leptons, as described in Section 6.2.2. As a result, the nature of the fake lepton changes with the multiplicity (single vs double fakes) and the hadron-jet flavor (fakes from light flavor such as u, d, s, c, g vs fakes from b-hadron decays). Hence, we need to devise a more universal strategy for the parametrization of the fake rates.

The multilepton landscape is dominated by the huge number of events from the three-lepton channels, specifically from 3L and 2L1T events. The most prevalent mechanism for the production of fake leptons in these channels are the DY+jets and $t\bar{t}$ processes. In DY+jet events, we get two prompt leptons from the decay of Z boson and the jet either gives rise to a fake lepton (electron or muon) or is misidentified as a τ_{h} lepton. This jet is topologically arranged in such a way that it balances against the dilepton system from Z boson. Consequently, the Lorentz boost of the dilepton system directly impacts the profile of the jet i.e. a collimated jet for highly boosted dilepton system whereas a more spread out jet for softer dilepton system. This, in turn, plays a major role in the properties of the fake lepton from its mother jet. Figure 6.14 shows an event topology of the 2L1T event from the DY+jets process, where the two light leptons (ℓ_1 and ℓ_2) are produced from the DY decay, and the jet j_1 (leading in p_{T}) is recoiling against the dilepton system. Together, the event is well-balanced, when the properties of jet j_1 are very similar to the reconstructed fake τ_{h} lepton.

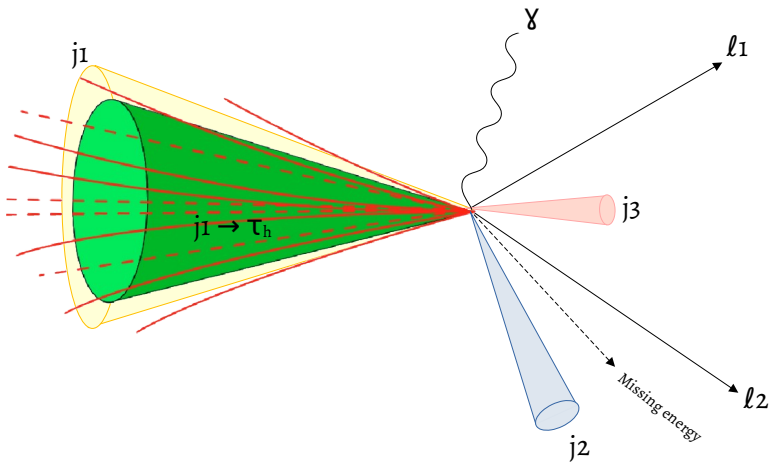

Figure 6.14: 2L1T event topology from DY+jets process. The leptons, ℓ_1 and ℓ_2, are produced from the decay of the Z boson, while the leading p_{T} jet j_1 is misreconstructed as a τ_{h} lepton. The jet j_1 balances against the dilepton system.

Due to the dependence of the properties of the fake lepton on the boost of the dilepton system, as well as small relative dependencies from other physics objects in the event; we need to be able to determine all the activity around the fake lepton fairly well in all the multilepton events. Most of this information is encoded in the properties of the mother jet of the lepton. Hence, we need to predict the source of origin of the leptons, in a way that the strategy applies for all the channels. To do this, we define a custom lepton *recoil* variable. The transverse recoil vector for any given lepton, $\vec{p}_T^{R}$, is calculated as the two-dimensional (xy-plane) vector sum of momenta of all other physics objects in the event ("non-lepton" p_{T}) i.e. excluding the lepton itself. The physics objects include all loose leptons, AK4 PF jets with $p_{\mathrm{T}} > 10$ GeV cleaned against the selected loose leptons ($\Delta R > 0.4$), and $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$. Then, the projection of the non-lepton p_{T} vector along the lepton transverse momentum axis is defined as the recoil variable $r_T \equiv -\vec{p}_T^{R} \cdot \vec{p}_T / p_{\mathrm{T}}$, such that it is positive when in opposite direction to the lepton. The magnitude of the recoil vector is basically a proxy for the p_{T} of the mother jet. Figure 6.15 illustrates two scenarios for the resultant non-lepton p_{T} and the calculated recoil vector.

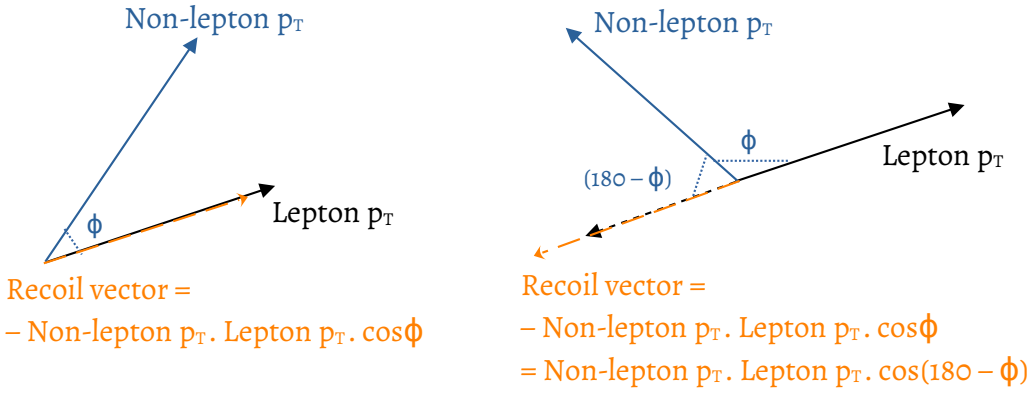

Figure 6.15: Custom recoil vector for any given lepton in two scenarios: when the non-lepton p_{T} is in the same direction as the lepton p_{T} (left) and when the non-lepton p_{T} is in the opposite direction as the lepton p_{T} (right). The non-lepton p_{T} is defined as the resultant vector sum of momenta of all other physics objects in the event, excluding the lepton itself. Recoil vector is negative in the left scenario, whereas it is positive in the right scenario.

6.2.2.3 Measurement of tau lepton fake rates

As explained earlier, DY+jets and $t\bar{t}$+jets processes are the most dominant SM contributions to the total misidentified lepton background in multilepton events. However, different light/heavy quark and gluon composition as well as different event kinematics of these two processes yield fake rates that may differ up to 50% from each other for the same lepton flavor. Therefore, dedicated measurements using a variant of the tag-and-probe method are performed in both processes.

The τ_{h} DY fake rates are measured in 2L1T OSSF Single-OnZ events with $p_{\mathrm{T}}^{\mathrm{miss}}$ <100 GeV, which constitutes the MisID CR for the fake taus. The OSSF Single-OnZ light leptons are taken as the tag leptons, and the τ_{h} is the probe lepton. The tau fake rates are primarily parametrized as a function of the delta transverse recoil, $\Delta R_T = r_T - p_{\mathrm{T}}$, where the calculation of r_T does not include $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ unlike described in Section 6.2.2.2. This is because the $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ calculation from the PF algorithm, using physics objects electrons, muons, photons, charged hadrons and neutral hadrons, is very sensitive to the difference between the properties of the fake tau and the mother jet. As a result, the effective $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ of the event based on the final global event kinematics changes with respect to the PF calculation. To understand this better, let's consider two scenarios: (a) events where the reconstructed fake τ_{h} lepton has larger p_{T} and different direction than the mother jet, and (b) events where the reconstructed fake τ_{h} lepton has same direction, but smaller in p_{T} than the mother jet. These two scenarios are shown in Figure 6.16.

In case of scenario (a), the dilepton system did not have a lot of boost and thus the recoiling jet j_1 is also very soft, making it a well-balanced system. Now let's suppose the fake τ_{h} lepton in the 2L1T event was reconstructed in a different direction than the jet axis, and has much larger p_{T} than the mother jet. This means that there is more energy on the left hand side of the system than the right hand side. According to the conservation of momentum, the new effective $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ has to be on the same side as dilepton system for the event to be well-balanced. Similarly, in scenario (b), the reconstructed fake τ_{h} lepton has much smaller p_{T} than the mother jet j_1. This means that in the final 2L1T event, there should be more missing energy on the left hand side of the system to conserve the momentum. Both these cases contradict the $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ calculated by the PF algorithm. Hence, we excluded the $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ from the calculation of the transverse recoil vector for the fake taus. This is different for events with only light leptons, since all the light leptons go in the calculation of PF $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$, therefore there is no mismatch in the initial and final $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ of the event.

The DY fake rates for τ_{h} leptons are measured in data in five orthogonal tau p_{T} regions ($20-30, 30-50, 50-80, 80-150, >150$ GeV) for the 1- and 3-prong decay modes, and a fit is performed in the first four of these p_{T} regions to minimize statistical fluctuations. Correction factors are used

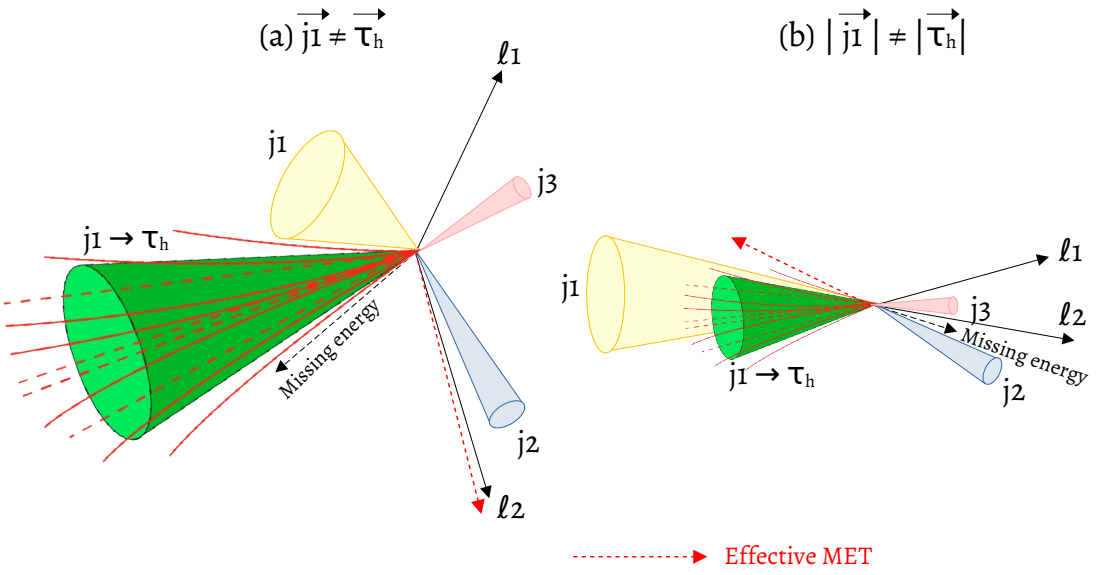

Figure 6.16: 2L1T DY+jets event topology for two scenarios: (a) events where the reconstructed fake τ_{h} lepton has larger p_{T} and different direction than the mother jet (left), and (b) events where the reconstructed fake τ_{h} lepton has same direction, but smaller in p_{T} than the mother jet (right). The black dotted line represents the $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$ calculated by the PF algorithm using physics objects electrons, muons, photons, charged and neutral hadrons; whereas the red dotted line in the effective direction of $\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$ after the tau reconstruction.

instead of orthogonal multidimensional measurements in order to maintain sufficient event yields in each bin. These are fits based on polynomial of order 2, as a function of tau $|\eta|$ in barrel ($|\eta| \leq 1.5$) and endcap ($|\eta| > 1.5$) regions for 1- and 3-prongs separately. Finally, we apply an additional correction as a function of the multiplicity of tracks originating from the PV (N_{trk}) in the event inclusively, for a better modeling of the H_{T} distribution. Tau data DY fake rates as a function of the ΔR_T for the 1-prong decays are shown in Figure 6.17-6.18, while for the 3-prong decays are shown in Figure 6.19-6.20 for the three years of data taking. Additional correction factors as a function of tau η and N_{trk} are shown in Figure 6.21-6.22-.

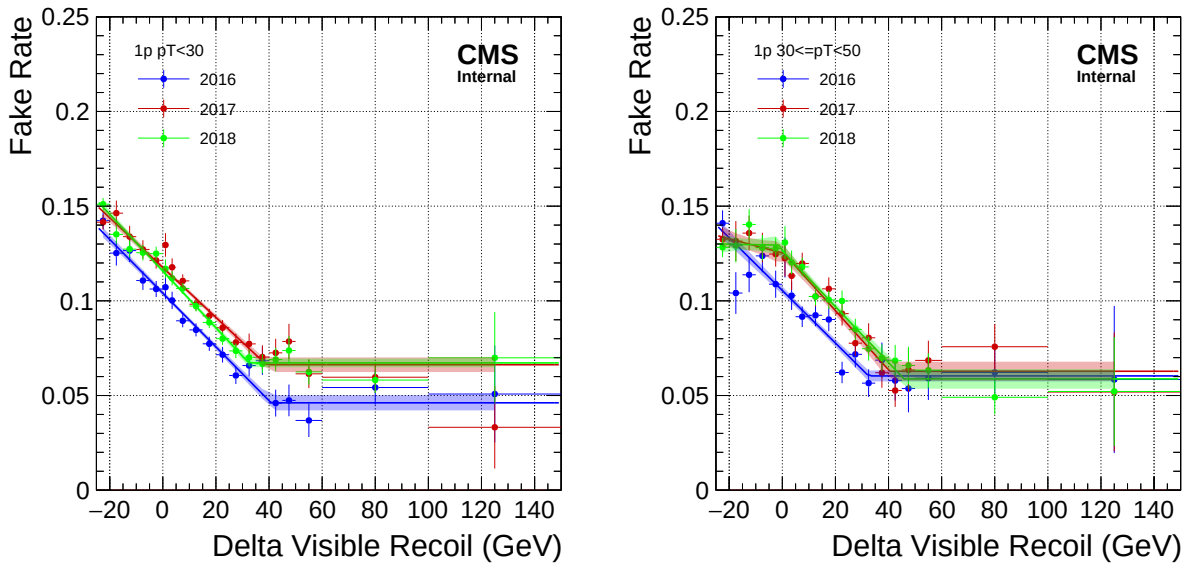

Figure 6.17: 1-prong τ_{h} data DY fake rates in the three years of data-taking. Rates for 2016, 2017, and 2018 are shown in blue, red, and green, respectively. The measurements are conducted in the 2L1T OSSF OnZ, $p_{\mathrm{T}}^{\mathrm{miss}}$ <100 GeV control region in data. The fake rates have been parametrized as a function of the lepton ΔR_T in tau p_{T} regions, shown here for $20-30$ GeV (left) and $30-50$ GeV (right). MC rates are shown for comparison-only purposes. The uncertainties include prompt subtraction effects (negligible), and the fit uncertainty bands are used for the systematic uncertainties on the MisID background estimate.

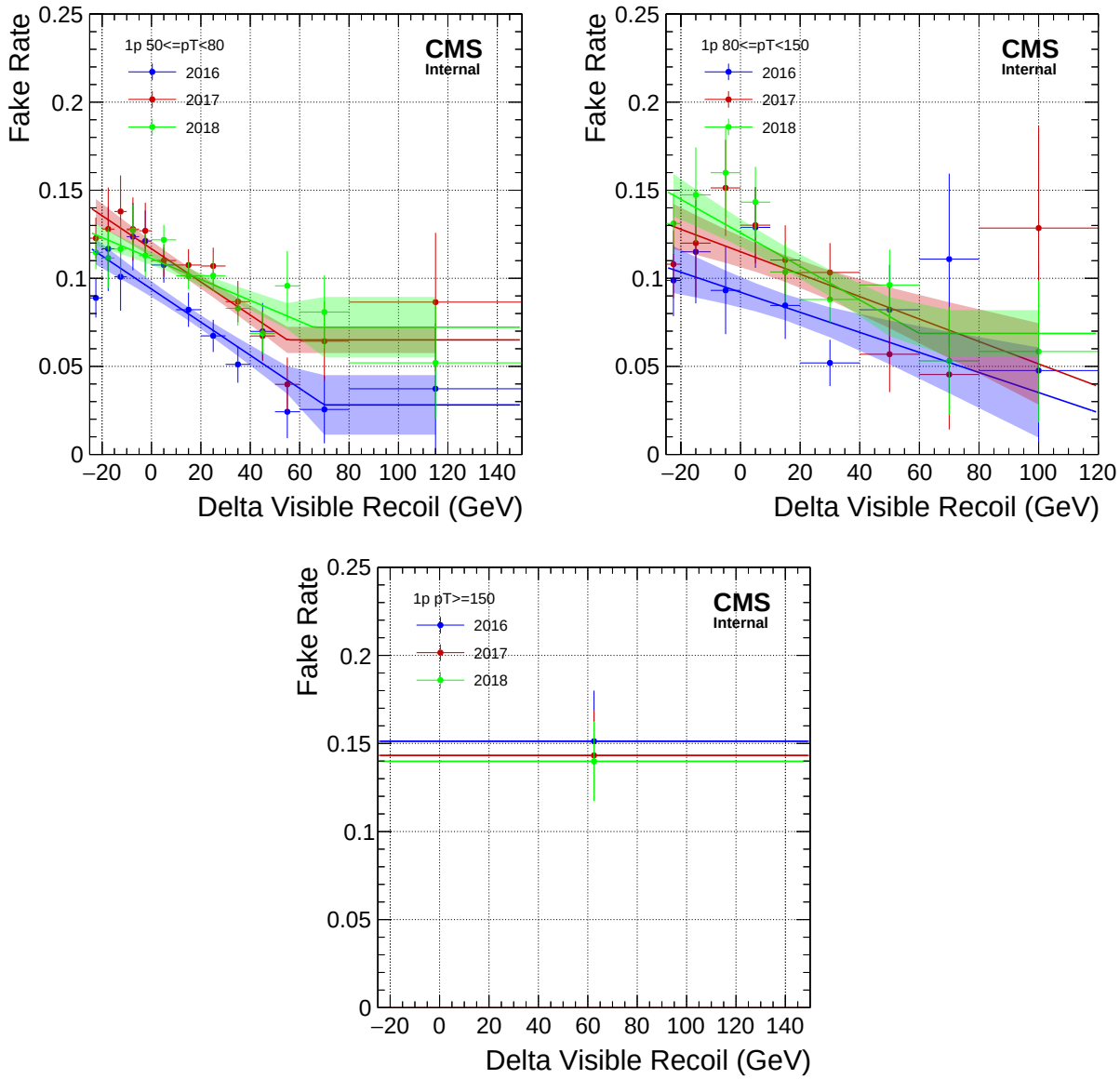

Figure 6.18: 1-prong τ_{h} data DY fake rates in the three years of data-taking. Rates for 2016, 2017, and 2018 are shown in blue, red, and green, respectively. The measurements are conducted in the 2L1T OSSF OnZ, $p_{\mathrm{T}}^{\mathrm{miss}}$ <100 GeV control region in data. The fake rates have been parametrized as a function of the lepton ΔR_T in tau p_{T} regions, shown here for $50-80$ GeV (upper left), $80-150$ GeV (upper right), and an inclusive overflow measurement is made for $p_{\mathrm{T}} > 150$ GeV (lower). MC rates are shown for comparison-only purposes. The uncertainties include prompt subtraction effects (negligible), and the fit uncertainty bands are used for the systematic uncertainties on the MisID background estimate.

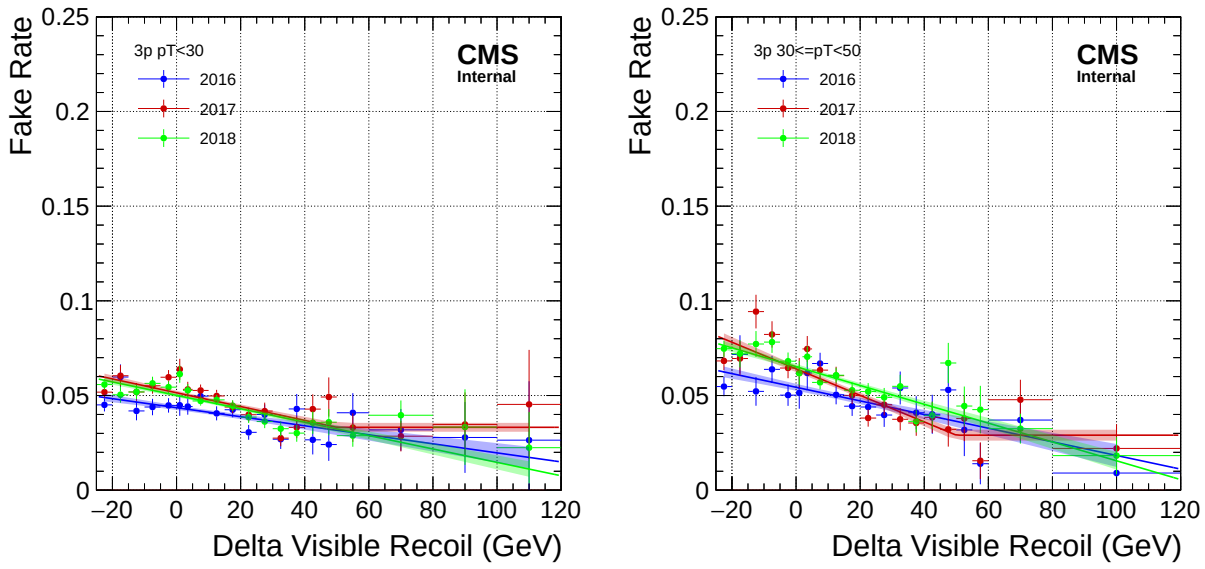

Figure 6.19: 3-prong τ_{h} data DY fake rates in the three years of data-taking. Rates for 2016, 2017, and 2018 are shown in blue, red, and green, respectively. The measurements are conducted in the 2L1T OSSF OnZ, $p_{\mathrm{T}}^{\mathrm{miss}}$ <100 GeV control region in data. The fake rates have been parametrized as a function of the lepton ΔR_T in tau p_{T} regions, shown here for $20-30$ GeV (left) and $30-50$ GeV (right). MC rates are shown for comparison-only purposes. The uncertainties include prompt subtraction effects (negligible), and the fit uncertainty bands are used for the systematic uncertainties on the MisID background estimate.

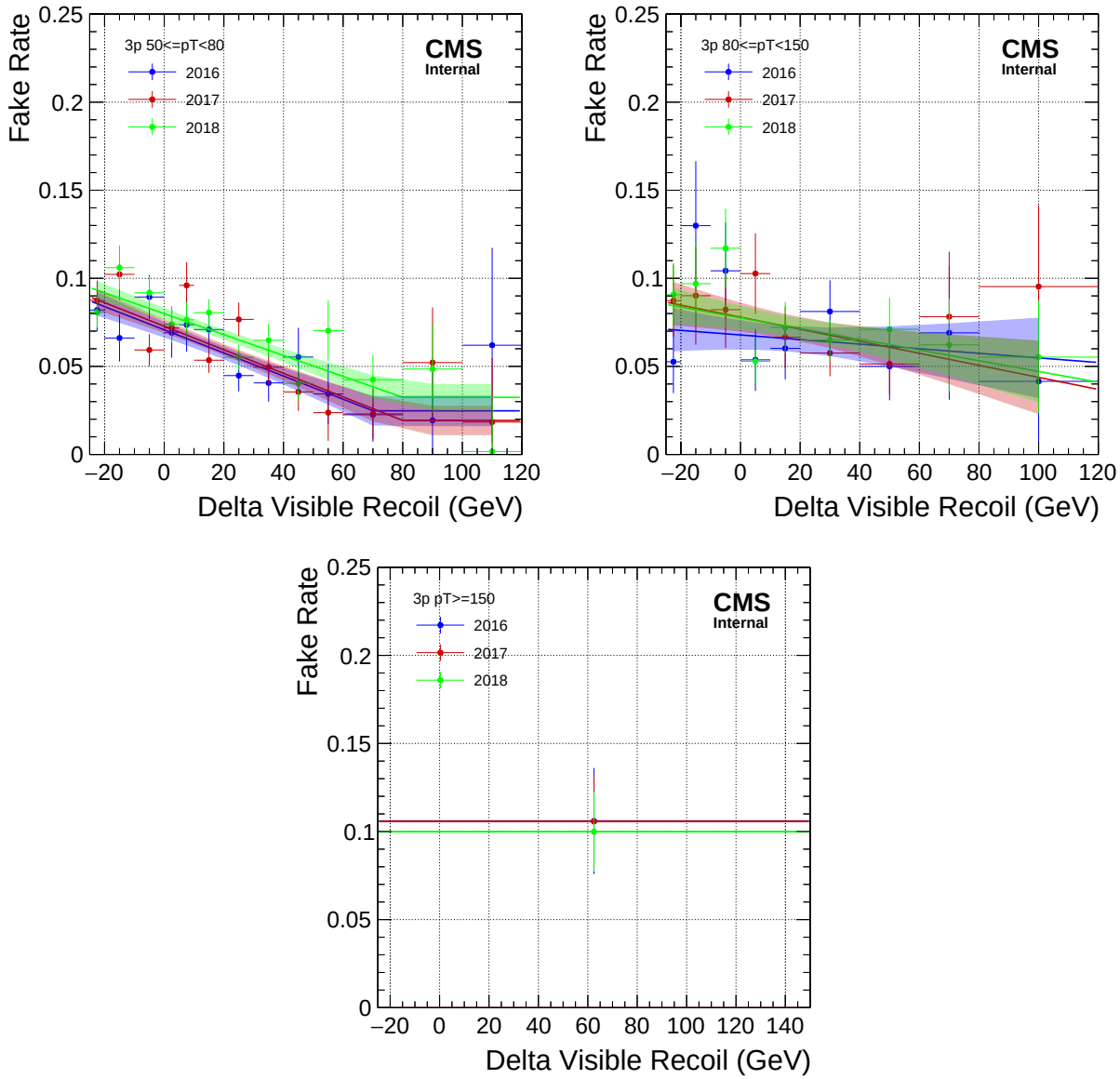

Figure 6.20: 3-prong τ_{h} data DY fake rates in the three years of data-taking. Rates for 2016, 2017, and 2018 are shown in blue, red, and green, respectively. The measurements are conducted in the 2L1T OSSF OnZ, $p_{\mathrm{T}}^{\mathrm{miss}}$ <100 GeV control region in data. The fake rates have been parametrized as a function of the lepton ΔR_T in tau p_{T} regions, shown here for $50-80$ GeV (upper left), $80-150$ GeV (upper right), and an inclusive overflow measurement is made for $p_{\mathrm{T}} > 150$ GeV (lower). MC rates are shown for comparison-only purposes. The uncertainties include prompt subtraction effects (negligible), and the fit uncertainty bands are used for the systematic uncertainties on the MisID background estimate.

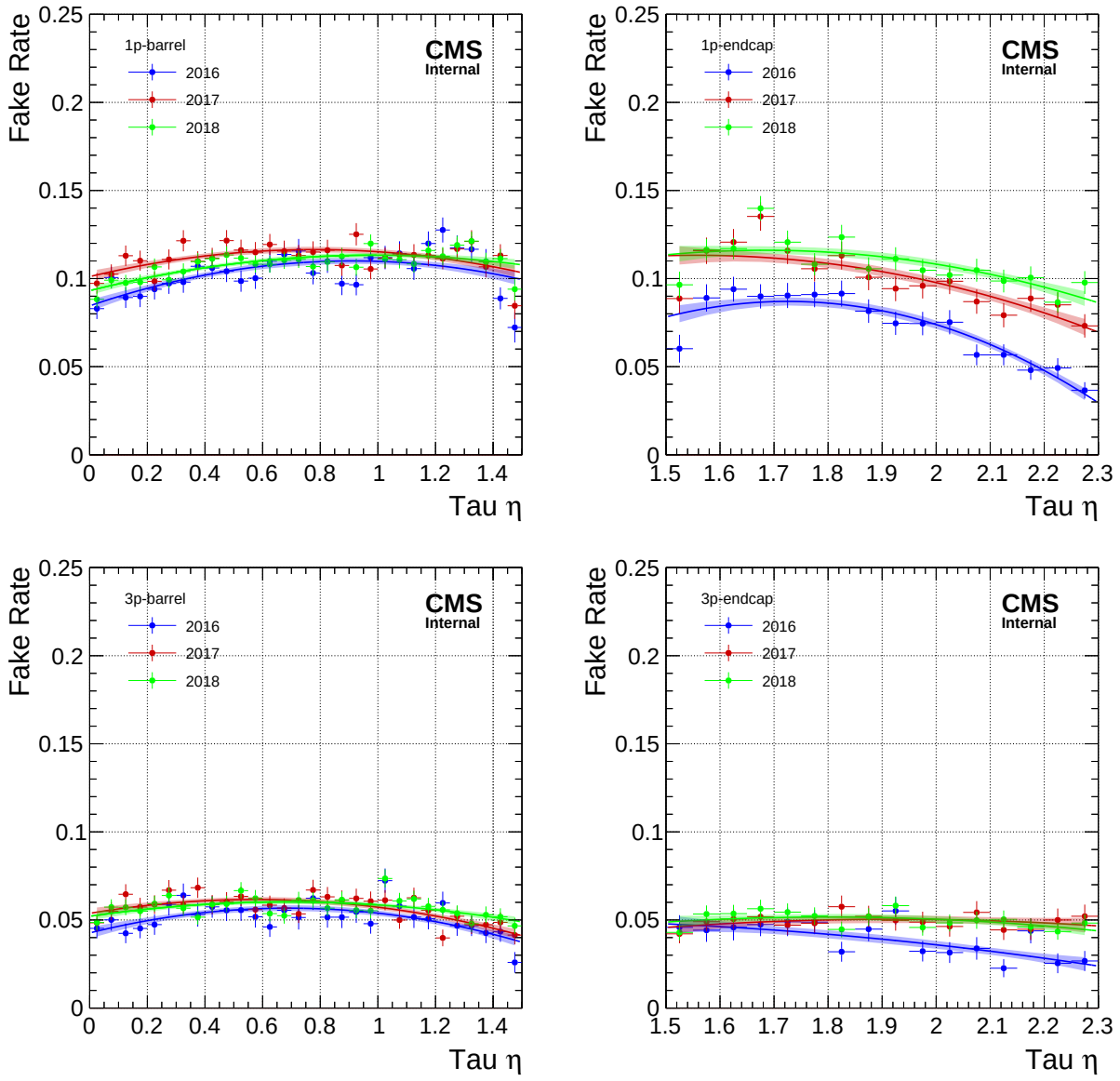

Figure 6.21: τ_{h} data DY fake rate correction factors in the three years of data-taking. Rates for 2016, 2017, and 2018 are shown in blue, red, and green, respectively. The measurements are conducted in the 2L1T OSSF OnZ, $p_{\mathrm{T}}^{\mathrm{miss}}$ <100 GeV control region in data. The correction factors are paramterized as a function of lepton $|\eta|$ in 1- and 3-prong, barrel and endcap regions. MC rates are shown for comparison-only purposes. The uncertainties include prompt subtraction effects (negligible), and the fit uncertainty bands are used for the systematic uncertainties on the MisID background estimate.

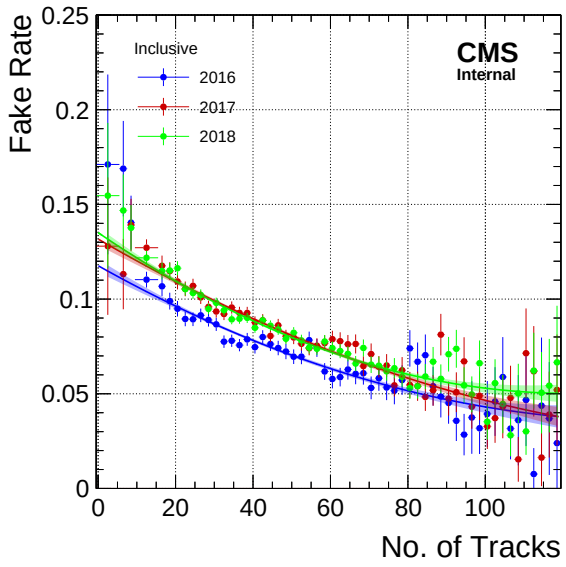

Figure 6.22: τ_{h} data DY fake rate correction factors in the three years of data-taking. Rates for 2016, 2017, and 2018 are shown in blue, red, and green, respectively. The measurements are conducted in the 2L1T OSSF OnZ, $p_{\mathrm{T}}^{\mathrm{miss}}$ <100 GeV control region in data. The correction factors are paramterized inclusively as a function of N_{trk}. MC rates are shown for comparison-only purposes. The uncertainties include prompt subtraction effects (negligible), and the fit uncertainty bands are used for the systematic uncertainties on the MisID background estimate.

Similarly, a dileptonic $t\bar{t}$ MC sample is used for the τ_{h} $t\bar{t}$ fake rate measurement following the exact same parametrization in 2L1T events with OS same- or opposite-flavor light leptons (e.g. $\mathrm{e}^{+}\mathrm{e}^{-}$, $\mu^{+}\mathrm{e}^{-}$). The "tag" light leptons are required to be matched to generator-level leptons ($\Delta R < 0.2$), while the "probe" fake τ_{h} lepton is anti-matched with the generator-level lepton ($\Delta R > 0.2$). Fake leptons from $t\bar{t}$ process are coming mainly from b-hadron decays. Those decays are rather well modeled in MC, and the fake rates can be measured quite accurately in the high statistics $t\bar{t}$ MC samples. Hence, we primarily measured them using simulation. However, we did perform various closure tests of those MC fake rates in data to asses any residual systematic differences between $t\bar{t}$ MC and data fake rates. We defined a "semi-tight" $t\bar{t}$-like CR in data where the probe lepton passes loose but fails tight ID. We check closure of misID prediction via "loose-not-tight" MC fake rates against observation. This provides us good faith in the fake rates for tight objects obtained using MC only.

The fake rates for the 1-prongs and 3-prongs, and the additional corrections for the three years of data-taking are shown in Figure 6.23-6.24, Figure 6.25-6.26, and Figure 6.27-6.28, respectively.

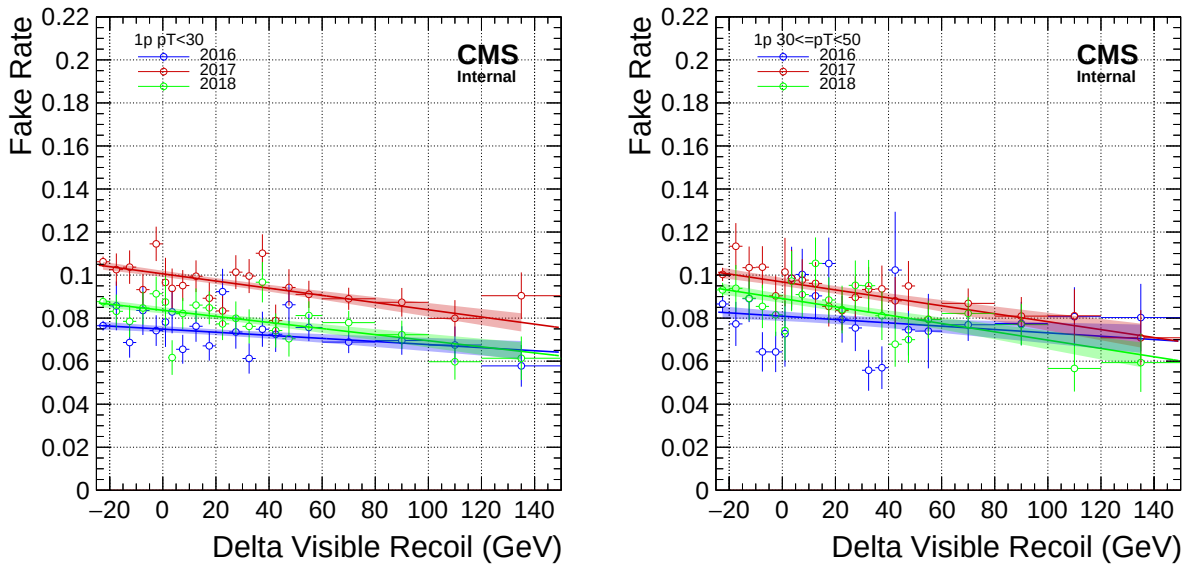

Figure 6.23: 1-prong τ_{h} $t\bar{t}$ MC fake rates in the three years of data-taking. Rates for 2016, 2017, and 2018 are shown in blue, red, and green, respectively. The measurements are conducted in a 2L1T inclusive selection in MC. The fake rates have been parametrized as a function of the lepton ΔR_T in tau p_{T} regions, shown here for $20-30$ GeV (left) and $30-50$ GeV (right). The uncertainties are statistical only, and the fit uncertainty bands are taken as the lowest bound for the systematic uncertainties on the MisID background estimate.

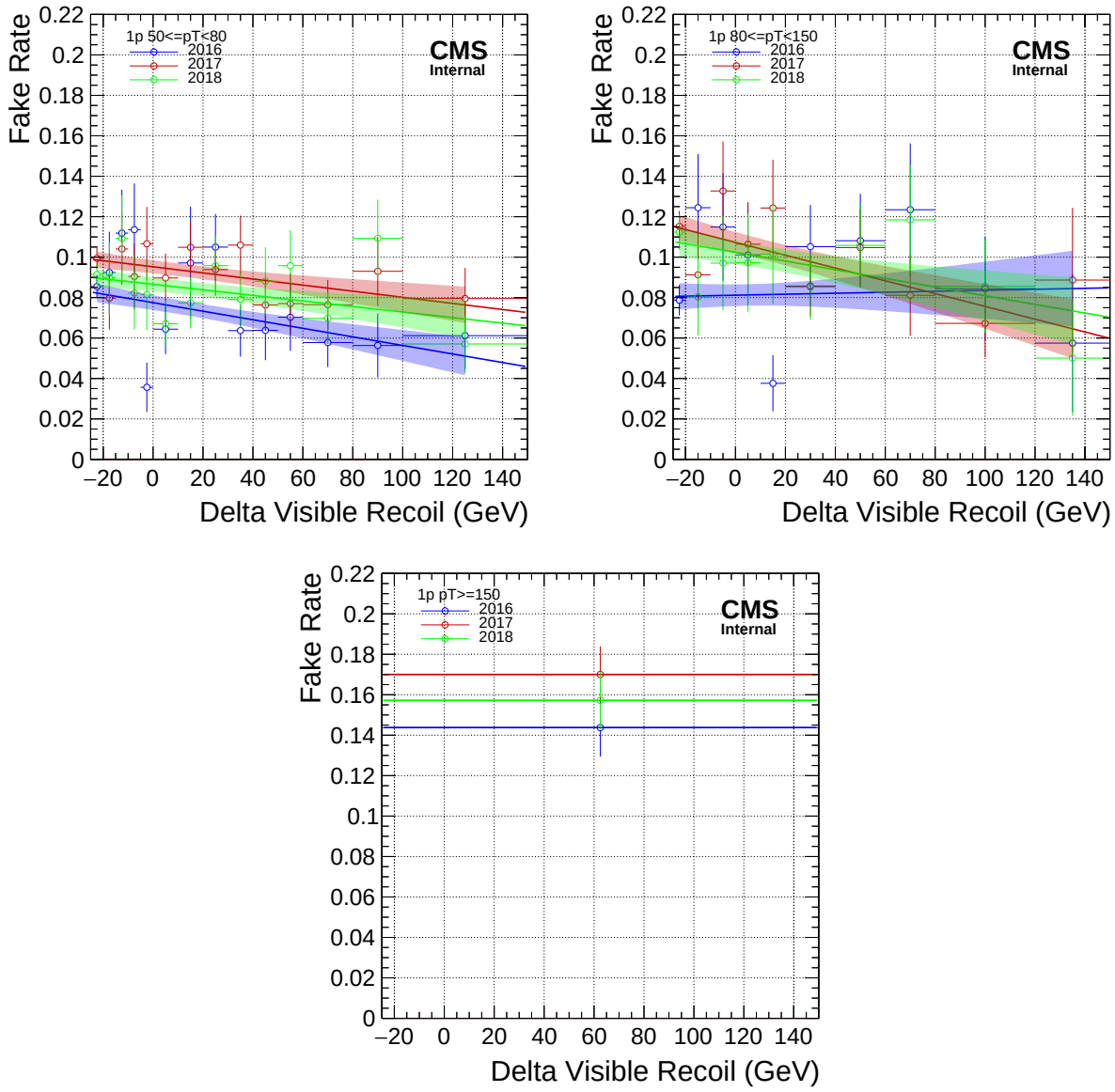

Figure 6.24: 1-prong τ_{h} $t\bar{t}$ MC fake rates in the three years of data-taking. Rates for 2016, 2017, and 2018 are shown in blue, red, and green, respectively. The measurements are conducted in a 2L1T inclusive selection in MC. The fake rates have been parametrized as a function of the lepton ΔR_T in tau p_{T} regions, shown here for $50 - 80$ GeV (upper left), $80 - 150$ GeV (upper right), and an inclusive overflow measurement is made for $p_{\mathrm{T}} > 150$ GeV (lower). The uncertainties are statistical only, and the fit uncertainty bands are taken as the lowest bound for the systematic uncertainties on the MisID background estimate.

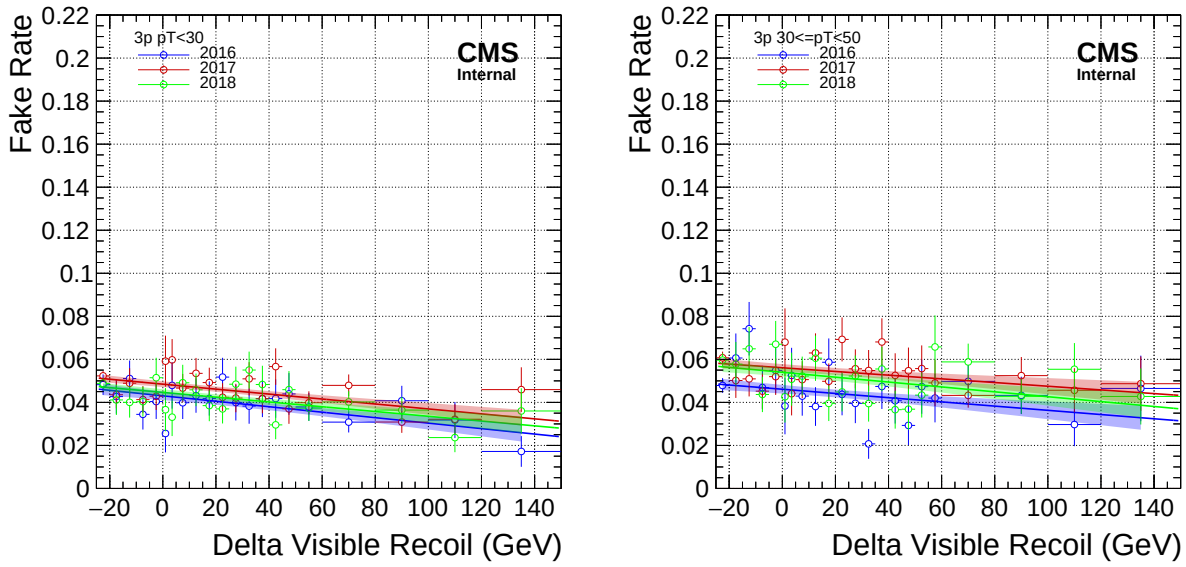

Figure 6.25: 3-prong τ_{h} $t\bar{t}$ MC fake rates in the three years of data-taking. Rates for 2016, 2017, and 2018 are shown in blue, red, and green, respectively. The measurements are conducted in a 2L1T inclusive selection in MC. The fake rates have been parametrized as a function of the lepton ΔR_T in tau p_{T} regions, shown here for $20-30$ GeV (left) and $30-50$ GeV (right). The uncertainties are statistical only, and the fit uncertainty bands are taken as the lowest bound for the systematic uncertainties on the MisID background estimate.

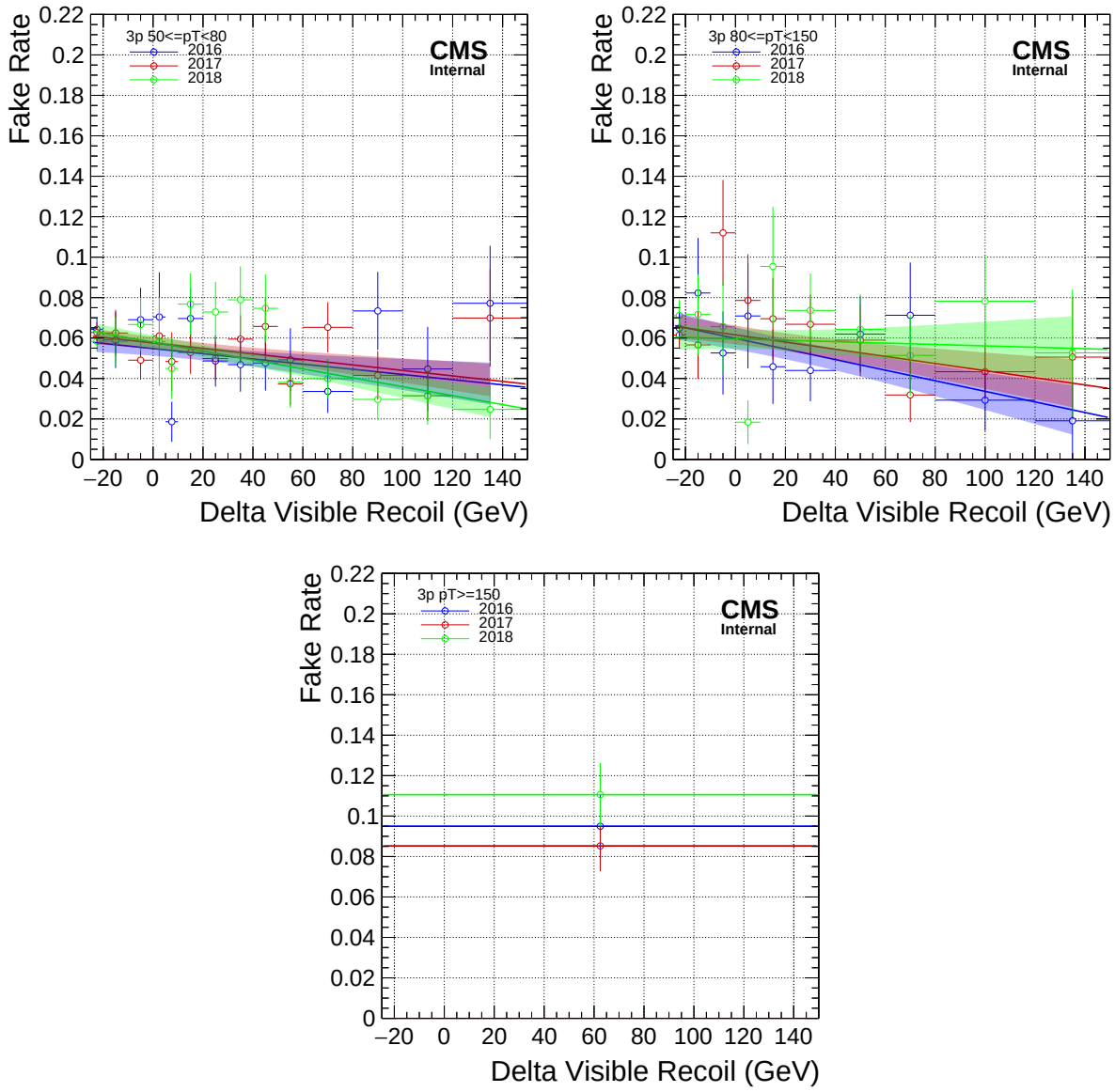

Figure 6.26: 3-prong τ_{h} $t\bar{t}$ MC fake rates in the three years of data-taking. Rates for 2016, 2017, and 2018 are shown in blue, red, and green, respectively. The measurements are conducted in a 2L1T inclusive selection in MC. The fake rates have been parametrized as a function of the lepton ΔR_T in tau p_{T} regions, shown here for $50 - 80$ GeV (upper left), $80 - 150$ GeV (upper right), and an inclusive overflow measurement is made for $p_{\mathrm{T}} > 150$ GeV (lower). The uncertainties are statistical only, and the fit uncertainty bands are taken as the lowest bound for the systematic uncertainties on the MisID background estimate.

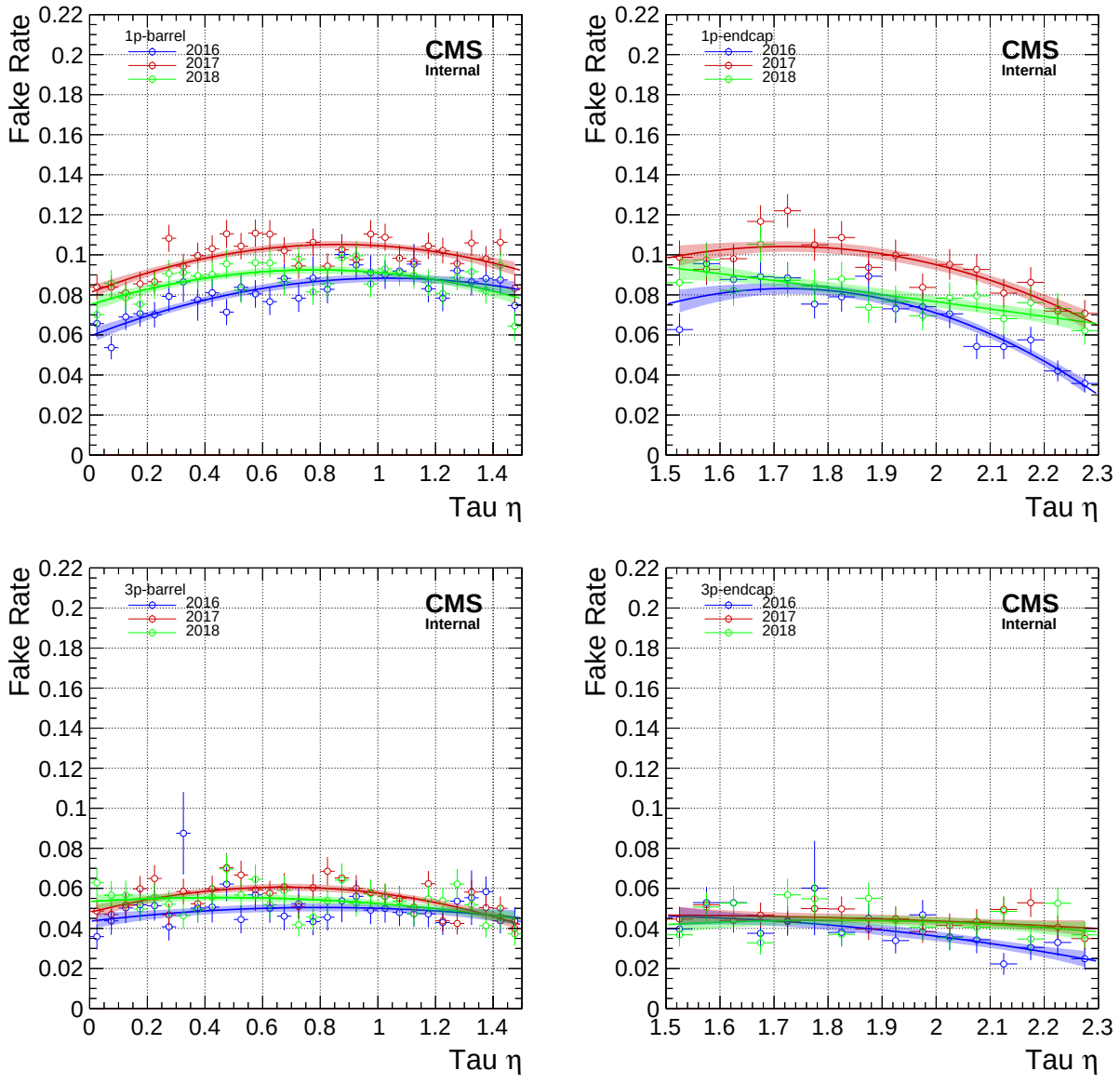

Figure 6.27: τ_h $t\bar{t}$ MC fake rate correction factors in the three years of data-taking. Rates for 2016, 2017, and 2018 are shown in blue, red, and green, respectively. The measurements are conducted in a 2L1T inclusive selection in MC. The correction factors are paramterized as a function of lepton $|\eta|$ in 1- and 3-prong, barrel and endcap regions. The uncertainties are statistical only, and the fit uncertainty bands are taken as the lowest bound for the systematic uncertainties on the MisID background estimate.

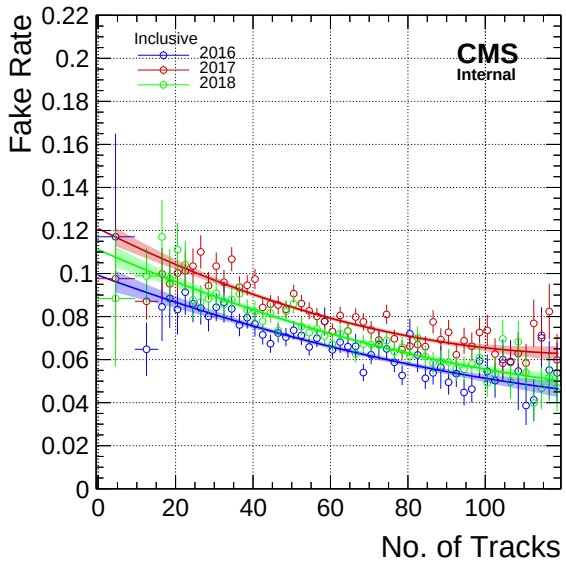

Figure 6.28: τ_{h} $t\bar{t}$ MC fake rate correction factors in the three years of data-taking. Rates for 2016, 2017, and 2018 are shown in blue, red, and green, respectively. The measurements are conducted in a 2L1T inclusive selection in MC. The correction factors are paramterized inclusively as a function of N_{trk}. The uncertainties are statistical only, and the fit uncertainty bands are taken as the lowest bound for the systematic uncertainties on the MisID background estimate.

A summary of the fake rate parametrization for the τ_{h} leptons is given in Table 6.3.

Table 6.3: Fake rate parametrizations for all lepton flavors. Individual rates and corrections are measured in orthogonal bins as specified by the binning schemes and as functions of the given variables. Corrections are normalized in order not to affect the mean rates.

	Primary		Correction	
	Binning scheme	Variable	Binning scheme	Variable
e fake rate	$\lvert\eta\rvert : \{0, 1.5, 2.4\}$	p_{T}	$p_{\mathrm{T}} : \{10, 15, 40\}$ GeV, $N_{\mathrm{j}} : \{0, \geq 1\}$	R_T (for DY) or N_{j}^{15} (for $t\bar{t}$)
μ fake rate	$\lvert\eta\rvert : \{0, 1.2, 2.4\}$	p_{T}	$p_{\mathrm{T}} : \{10, 15, 40\}$ GeV, $N_{\mathrm{j}} : \{0, \geq 1\}$	R_T (for DY) or N_{j}^{15} (for $t\bar{t}$)
τ_h fake rate	$p_{\mathrm{T}} : \{20, 30, 50, 80, 150, \infty\}$ GeV†	ΔR_T	$\lvert\eta\rvert : \{0, 1.5, 2.3\}^{\dagger}$	$\lvert\eta\rvert$
			Inclusive	N_{trk}
	† in 1- and 3-prong separately			

Typical τ_{h} fake rates are in the range of $1 - 15\%$, across all years and bins. These are found to be smaller than the fake rates of light leptons in our analysis. The reason why tau fake rates are smaller than light lepton fake rates, despite the intuition that hadronic taus are more closer look alike of jet objects is because these are relative fake rates, defined with respect to an appropriate but arbitrary denominator. These do not reflect how "fakeable" leptons are in CMS for the specific lepton selection criteria. If the denominator object had been the same in the rate measurement (e.g. loose ID jets), then one would indeed expect to see higher misID rates for taus than light leptons. In principle, we want to use the loosest denominator definition for each lepton flavor, with good closure properties in the background estimation. These denominators were chosen so that they are not too different from the isolated objects in the numerator, but also loose enough to populate the sidebands for a reliable background estimate.

The final fake rate per lepton, as used in the matrix method estimation, is calculated as the weighted sum of the DY-based and $t\bar{t}$-based measurement, as given in Eqn. 6.7. Here, the terms f_{DY} and $f_{t\bar{t}}$ represents the relative contribution of the DY and $t\bar{t}$ MC, respectively, measured in high statistics SM background dominated regions. Similarly, FR_{ℓ}^{DY} and $FR_{\ell}^{t\bar{t}}$ are the DY and $t\bar{t}$ fake rates, respectively, per lepton flavor ℓ.

$$FR_{\ell}^{final} = f_{DY} \times FR_{\ell}^{DY} + (1 - f_{DY}) \times FR_{\ell}^{t\bar{t}} \tag{6.7}$$

6.2.3 Misidentified electrons and muons

Similar to the tau leptons, we measure the prompt and fake rates for the light leptons, i.e. electrons and muons.

Prompt rates for electrons and muons are studied in a DY enriched set of OS ee and $\mu\mu$ OnZ events in data, respectively. In MC samples, prompt rates have been measured in DY and $t\bar{t}$ MC

samples. In prompt rate measurements conducted in data, contributions due to fake probe leptons are estimated and subtracted using MC methods, which is only a minor correction for electrons and muons. The final prompt rate is based on the DY enriched data measurements, whereas differences between DY and $t\bar{t}$ MC based rates is taken as a prompt rate systematic uncertainty to account for varying levels of hadronic activity.

For the measurement of fake rates of electrons and muons, a DY enriched selection of data events with a fake lepton is created by having a trilepton selection with an OnZ pair, $p_{\mathrm{T}}^{\mathrm{miss}} < 100$ GeV $M_{\mathrm{T}} < 50$ GeV and $N_{\mathrm{b}} = 0$. The OnZ leptons are taken as the tag leptons, and the additional lepton is taken as the fake probe lepton, i.e. in DY enriched dataset, eeμ and $\mu\mu\mu$ events are used to measure muon fake rates, and eee and $\mu\mu$e events are used to measure the electron fake rates. In each bin where a fake rate measurement is performed, the low M_{T} ($M_{\mathrm{T}} < 50$ GeV) region is used to compute the fake rate (to increase the purity of fakes), whereas the high M_{T} ($50 < M_{\mathrm{T}} < 150$ GeV) region is used to measure the per-bin in-situ WZ normalization, followed by subtraction in each bin due to high prompt contamination.

For the DY fake rates of light leptons, these correction factors are measured as a function of the relative transverse recoil, $R_T \equiv r_T/p_{\mathrm{T}}$, in bins of low and high lepton p_{T} as well as low and high jet multiplicity (N_{j}) with respect to the average fake rate of given bin. For the $t\bar{t}$ fake rates, the N_{j}^{15} variable is used for the correction factor parametrization instead. The final fake rates are obtained by the product of the initial fake rates and the corresponding correction factors. The operational differences in the treatment of light lepton vs tau fake rate parametrizations originate from differences in available statistics (heavily favoring taus) as well as the isolation characteristics (light leptons use a relative PF isolation, whereas taus use a multivariate discriminator) among different lepton flavors.

Details of the parametrization for the prompt rate measurement of the electrons and muons are given in Table 6.2. Prompt rates for electrons and muons vary from about 65% at $p_{\mathrm{T}} \sim 10$ GeV to about 95% at 40 GeV and beyond, across all years and bins. Similarly, a summary of the fake rate parametrization for the electrons and muons is given in Table 6.3. Typical light lepton fake rates are in the range of $5 - 30\%$.

6.2.4 Application of matrix method

Substituting all the prompt and fake rates, measured for all the lepton flavors, in the three-dimensional analogue of the two-dimensional matrix Eqn. 6.6 and the Eqn. 6.5, we get the prediction of the misidentified lepton backgrounds in the multilepton events. We perform the closure test of the matrix method in the 2L1T MisID CR events for the misidentified taus, and in 3L MisID CR events for the misidentified light leptons.

The distributions of L_{T} and H_{T} in the 3L MisID CR for the combined 2016–2018 data set are shown in Figure 6.29. The distributions of L_{T} and number of b-tagged jets in the 2L1T MisID CR for the combined 2016–2018 data set are shown in Figure 6.30. All these distributions are shown with statistical uncertainties only.

After the application of systematic uncertainties, as described in Section 6.3, the distributions of $p_{\mathrm{T}}^{\mathrm{miss}}$ in 2L1T MisID CR and distribution of trailing lepton p_{T} in 3L MisID CR are shown in Figure 6.31. An excellent agreement between the data and total predicted background in both these CRs for the combined 2016–2018 data set can be seen.

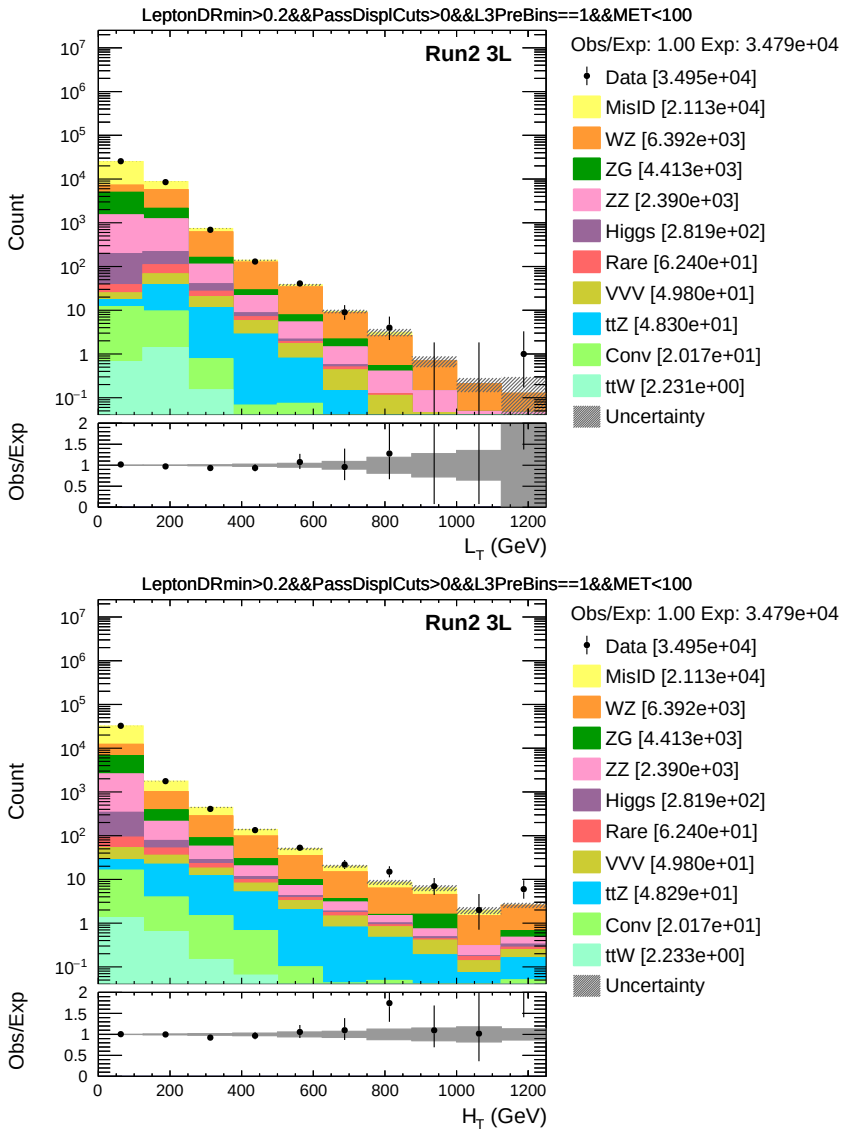

Figure 6.29: The distributions of L_{T} (left) and H_{T} (right) in the 3L MisID CR events for the combined 2016–2018 data set. The rightmost bin contains the overflow events in each distribution. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties, as explained in Section 6.3, in the SM background prediction.

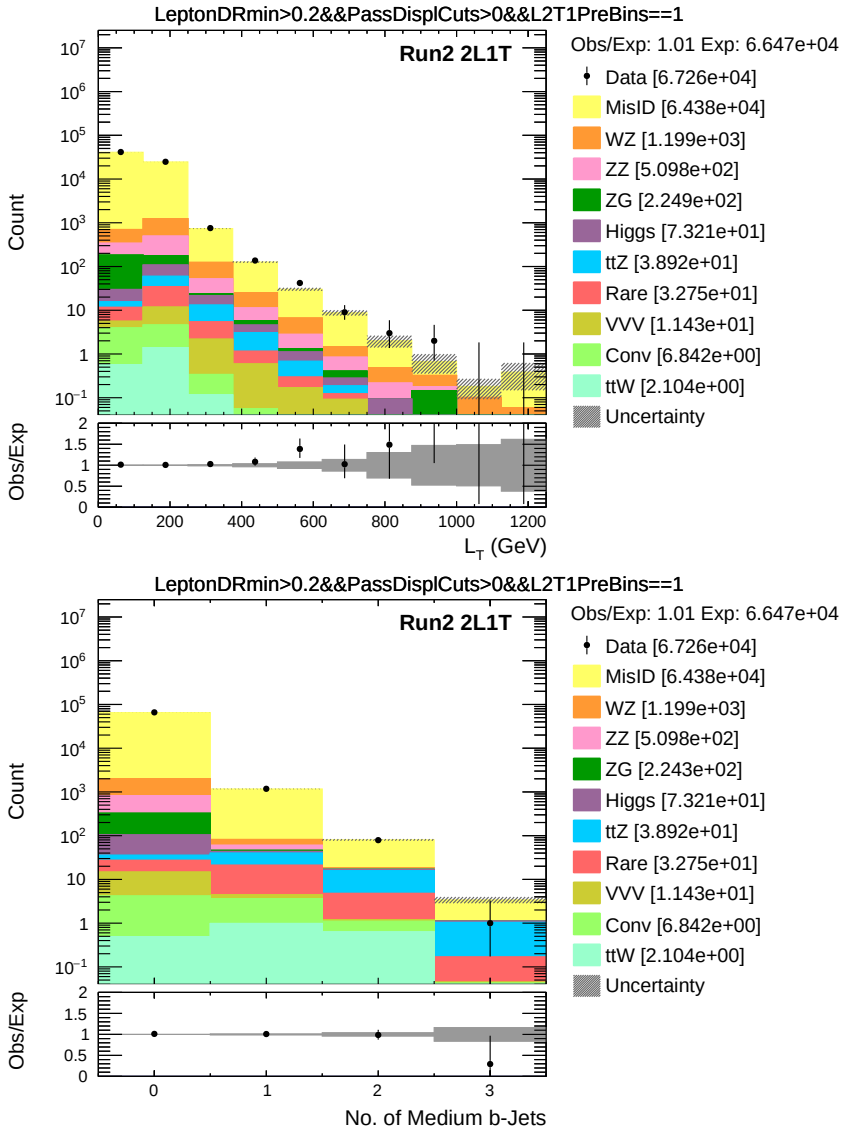

Figure 6.30: The distributions of L_{T} (left) and number of b-tagged jets (right) in the 2L1T MisID CR events for the combined 2016–2018 data set. The rightmost bin contains the overflow events in each distribution. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties, as explained in Section 6.3, in the SM background prediction.

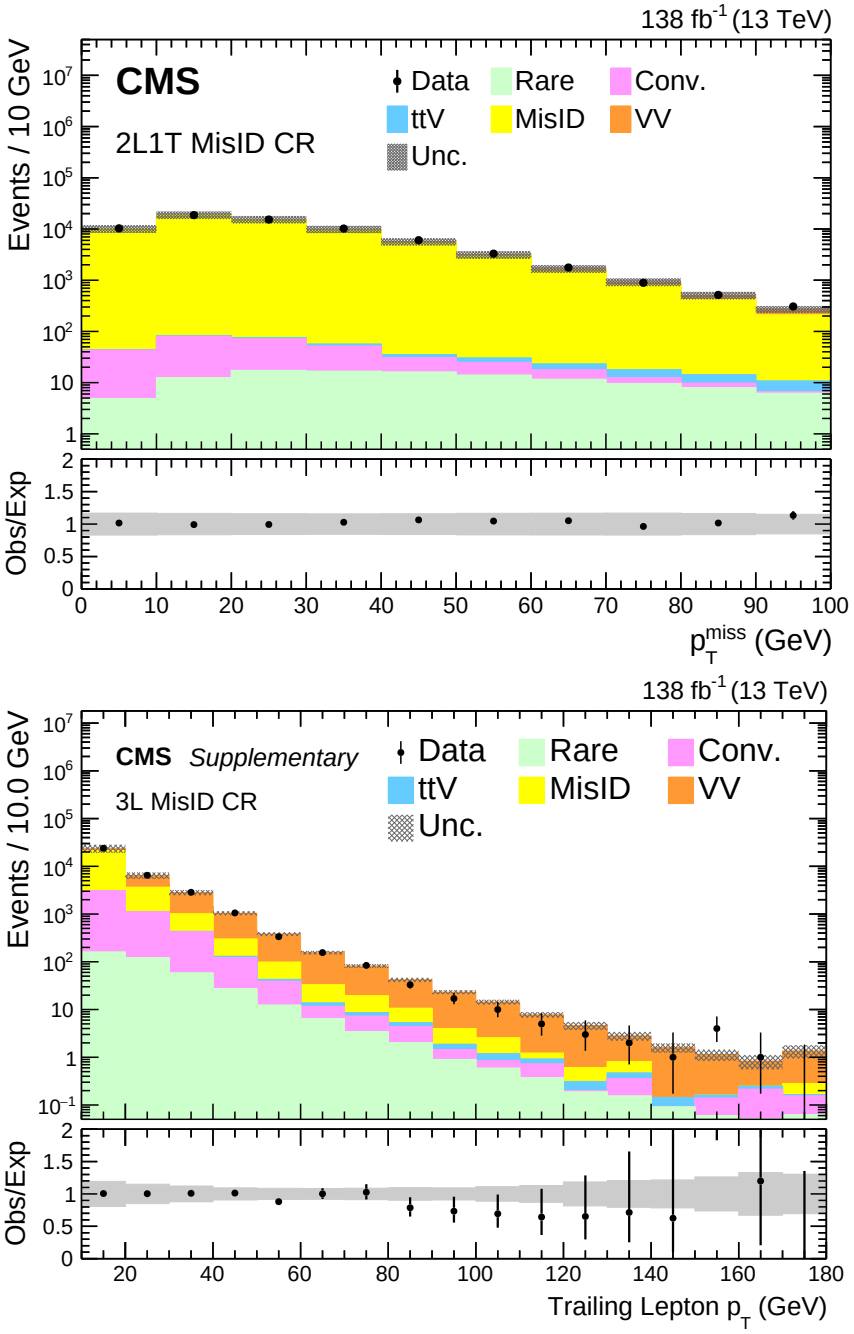

Figure 6.31: The distributions of $p_{\mathrm{T}}^{\mathrm{miss}}$ (left) and the softest or trailing lepton p_{T} (right) in the 2L1T MisID CR and 3L MisID CR events, respectively, for the combined 2016–2018 data set. The rightmost bin contains the overflow events in each distribution. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties, as explained in Section 6.3, in the SM background prediction.

6.3 Sources of systematic uncertainties

The precision in the SM background estimation is limited by two major uncertainties: statistical and systematic. The simulation samples of irreducible background are often rich in statistics, with much larger luminosity than the collected data. However, there are some corners of the signal regions such as the tails of the kinematic distributions L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ and S_{T}, which are not evenly populated due to the mismodeling of higher order cross section calculations in MC event generators. The misidentified background, on the other hand, is estimated from the loose sidebands in the data which may also fall short of statistics in those tails. Coincidentally, those regions with low expected yields offer highest sensitivity in a BSM search. In such cases, statistical uncertainty becomes the most dominant factor in constraining the new physics models in the statistical analysis.

For the bulk of the kinematic distributions, statistical uncertainty does not play a major role. This is where the shape of the distributions become highly susceptible to the systematic sources. While the statistical uncertainties are completely uncorrelated across the three years of the data-taking, the same does not apply for the systematic uncertainties. The most dominant source of systematic uncertainties are those arising from the misidentfied background estimation, and from the normalization of the MCs of the irreducible SM backgrounds.

The uncertainty in the misidentified lepton background, which is estimated from data via the matrix method, is driven by the uncertainties in the lepton misidentification rates. Lepton misidentification rates have typical relative uncertainties of 10, 30, and 60% in the low, medium, and high lepton p_{T}regions, respectively, where low is defined as ($10 < p_{\mathrm{T}} < 20$ GeV for light leptons, $10 < p_{\mathrm{T}} < 30$ GeV for τ_{h}), medium is ($20 < p_{\mathrm{T}} < 50$ GeV for light leptons, $30 < p_{\mathrm{T}} < 80$ GeV for τ_{h}), and high is ($p_{\mathrm{T}} > 50$ GeV for light leptons, $p_{\mathrm{T}} > 80$ GeV for τ_{h}). These result in variations in the range of 20–50% of the misidentified lepton background contribution estimates, and these nuisances are also kept uncorrelated in each of the three data-taking periods. In addition, we consider process-dependent uncertainties in the lepton misidentification rates. These are estimated by comparing the misidentification rates observed in the DY- and $t\bar{t}$-enriched measurements, and are typically in the range of 5–25% and correlated across the data-taking periods.

Aside from this, subdominant contribution to systematic uncertainties also arise from the corrections applied to the background and signal simulation. These include lepton reconstruction, isolation, and trigger efficiencies; b tagging efficiency; pileup modeling; electron and jet energy resolution; electron, muon, τ leptons, jet, and unclustered energy scale measurements; and due to choices of factorization and renormalization scales, and PDFs.

All the uncertainty sources, the affected processes, the resulting uncertainty in the yield of

Table 6.4: Sources, magnitudes, effective variations, and correlation properties of systematic uncertainties in the SRs. Uncertainty sources marked as "Yes" under the correlation column have their nuisance parameters correlated across the 3 years of data collection.

Uncertainty source	Magnitude	Type	Processes	Variation	Correlation
Statistical	1–100%	per event	All MC samples	1–100%	No
Integrated luminosity	1.2–2.5%	per event	Conv./Rare/Signal	1.2–2.5%	Yes
Electron/Muon reco., ID, and iso. efficiency	1–5%	per lepton	All MC samples	2–5%	No
τ_{h}reco., ID, and iso. efficiency	5–15%	per lepton	All MC samples	5–25%	No
Lepton displacement efficiency	1–2%	per lepton	All MC samples	3–5%	No
Trigger efficiency	1–4%	per lepton	All MC samples	<3%	No
b tagging efficiency	1–10%	per jet	All MC samples	2–5%	No
Pileup	5%	per event	All MC samples	<3%	Yes
PDF, fact./renorm. scale	<20%	per event	All MC samples	<10%	Yes
Jet energy scale	1–10%	per jet	All MC samples	<5%	No
Unclustered energy scale	1–25%	per event	All MC samples	<2%	No
Electron energy scale and resolution	<2%	per lepton	All MC samples	<5%	Yes
Muon energy scale and resolution	2%	per lepton	All MC samples	<5%	No
τ_{h}energy scale	<10%	per lepton	All MC samples	<5%	No
Electron charge misidentification	30%	per lepton	All MC samples	<25%	No
WZ normalization	3–5%	per event	WZ	3–5%	No
ZZ normalization	4–5%	per event	ZZ	4–5%	No
$t\bar{t}Z$ normalization	15–25%	per event	$t\bar{t}Z$	15–25%	No
Conversion normalization	10–50%	per event	$Z\gamma$/Conv.	10–50%	No
Rare normalization	50%	per event	Rare	50%	No
Prompt and misidentification rates	20–60%	per lepton	MisID	20–50%	No
DY-$t\bar{t}$ process dependence	5–25%	per lepton	MisID	5–25%	Yes
Diboson jet multiplicity modeling	<30%	per event	WZ/ZZ	5–30%	No
Diboson p_{T}modeling	<30%	per event	WZ/ZZ	5–15%	No

those processes, and the correlations across the data-taking periods are summarized in Table 6.4.

6.4 Validation in the entire multilepton phase space

Although the analysis is conducted in final states with three or more leptons, SM candles in dilepton events, such as leptonic DY or $t\bar{t}$ processes, provide high statistics region of events where the performance of the object and trigger selections as well as the scale factors can be commissioned. These results are shown in Appendix B.

Since our SM background estimation techniques and the associated uncertainties are in place, we can now look at the entire multilepton events from all the seven channels to find evidence of new physics. Figure 6.32 shows the distribution of the four most-important kinematic variables used to perform an inclusive nonresonant search with multileptons. These are L_{T} (upper left), $p_{\mathrm{T}}^{\mathrm{miss}}$ (upper right), H_{T} (lower left), and N_{b} (lower right), in the seven multilepton channels for all

events, including the ones from the CRs.

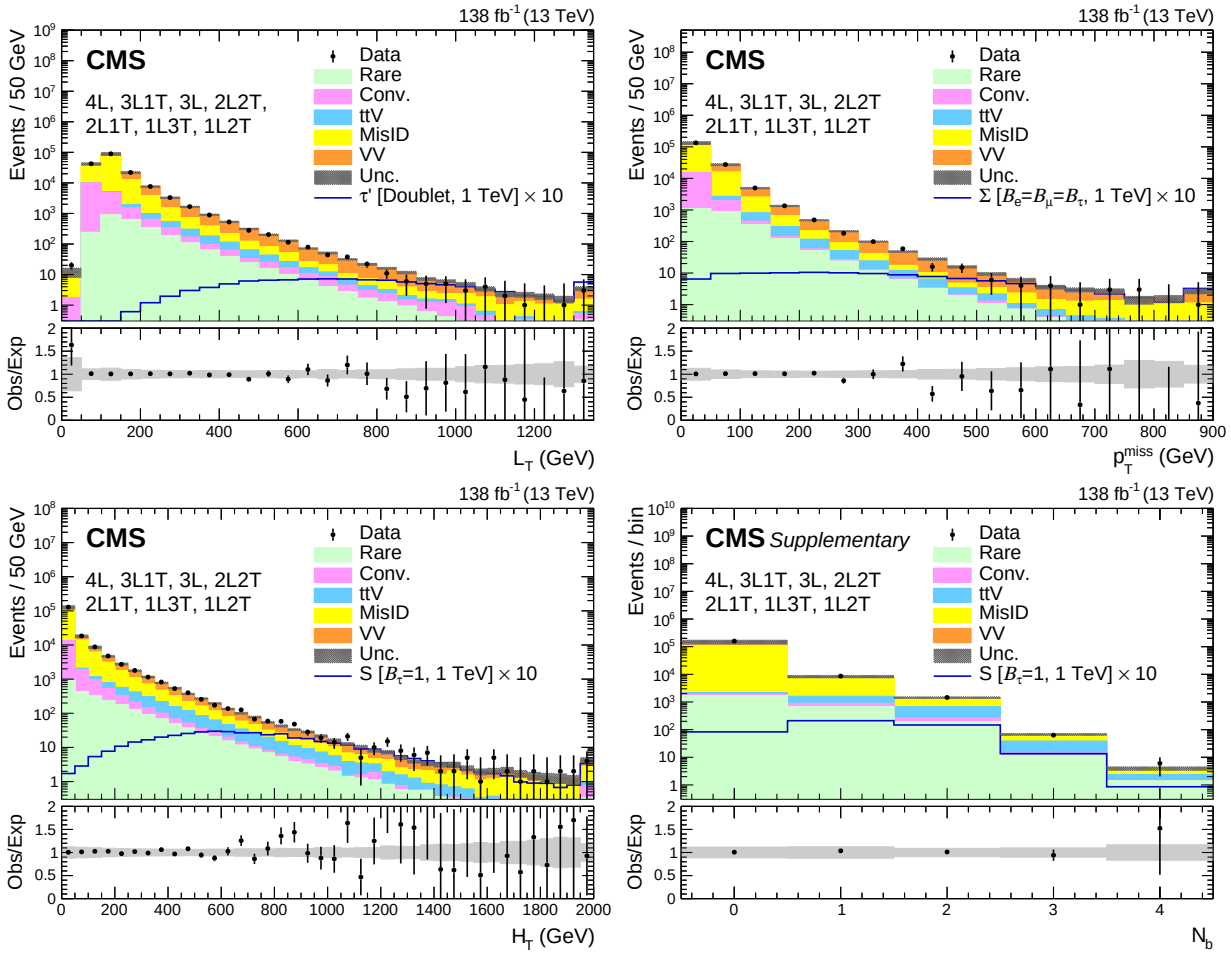

Figure 6.32: Upper left to lower right: The distributions of L_{T}, $p_{\mathrm{T}}^{\mathrm{miss}}$, H_{T}, and N_{b} in all seven multilepton channels, for the combined 2016–2018 data set. The rightmost bin contains the overflow events in each distribution. As illustrative examples, a signal hypothesis for the production of the vector-like τ leptons of $m_{\tau'} = 1$ TeV in the doublet scenario, and a signal hypothesis for the production of the type-III seesaw fermions of $m_{\Sigma} = 1$ TeV in the flavor-democratic scenario are overlaid in the L_{T} and $p_{\mathrm{T}}^{\mathrm{miss}}$ distributions, respectively. Similarly, a signal hypothesis for the production of scalar leptoquark of $m_S = 1$ TeV coupled to a top quark and a τ lepton is overlaid in the H_{T} and N_{b} distributions. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represent the sum of statistical and systematic uncertainties in the SM background predictions.

Looking at these distributions, we can conclude that there is a very nice agreement between the data and the total SM background prediction across the entire range. Our background estimation techniques are working really well, for all lepton flavors and in all the channels. To take

an example, we estimate the fake backgrounds using the lepton and its mother jet properties, but nonetheless we have a great agreement in the N_{b} distribution which is completely an event-based quantity. Not only that, we have good agreement across the bins of N_{b} which means that the changing fake background composition, from DY in N_{b}=0 to $t\bar{t}$ in N_{b}>0, is also very coherently taken care of.

So now the real question is, are there potential signs of new physics? If no, are we sure about that? If yes, how can we see them?

Hunting for beyond using advanced techniques...

Chapter 7

Searches using Machine learning

Machine learning (ML) is becoming an integral part of modern HEP research, with numerous applications in object, event and physics classification. It is a method of data analysis that automates analytical model building, by rigorously learning the features of the input data set and constantly improving via a feedback mechanism until the desired accuracy is achieved.

ML is based on the idea that systems can identify patterns from data and make decisions with minimal human intervention. To that front, there are four types of machine learning algorithms: supervised, semi-supervised, unsupervised, and reinforcement. Supervised learning is a ML algorithm in which models are trained using labeled data and takes direct feedback to check if its predicting correct output or not. The goal of supervised learning is to train the model so that it can predict the output when it is given new data. Classification and regression are the two main problems solved using supervised learning, and can be achieved using algorithms such as linear regression, multi-class classification, decision trees, etc. Unsupervised learning, on the other hand, takes places through unlabeled data and does not take any feedback. The primary goal of unsupervised learning is to find the hidden patterns to build useful insights from the unknown data set. It can be classified into problems such as clustering using K-Nearest Neighbour (KNN) algorithm with some distance metric and anomaly detection. Semi-supervised learning, as the name suggests, uses a small amount of labeled data set and a large amount of unlabeled data set during training. It can result in considerable improvement in the learning accuracy over the unsupervised learning. Reinforcement learning, positive or negative, is a reward-based way of training a model. The reward is imposed in terms of the ML training parameters, such as decrease in entropy or loss function and increasing accuracy of predicting the truth.

Traditionally, multivariate analysis techniques (MVA) such as Boosted Decision Trees (BDTs) and neural networks (NN) have been the HEP-wide favorite method for carrying out machine learn-

ing in physics analysis. For this, the ROOT-integrated environment for multivariate techniques, i.e. TMVA package [158] is used as the tool. However, the set of methods and tools commonly used in HEP has grown significantly in recent years as a result of the deep learning revolution. With the rapid development of research at the intersection of machine learning and HEP, it is difficult to keep track of the latest developments. A brilliant and extremely useful effort from the HEP community to consolidate all the results and ongoing developments can be found in this living review of machine learning [159].

MVA techniques are regularly used to enhance the sensitivity for BSM phenomena in physics searches, especially in difficult topologies where signal-to-background ratio is very small. However, there are many challenges in this multilepton analysis:

1. Three different BSM phenomena are probed which have diverse physics properties.

2. For each BSM signal, a large parameter space for the masses of the signal particles is probed. Also, there are many coupling scenarios in each BSM model. For example, the seesaw fermions can couple only to electrons ($\mathrm{B_e} = 1$) or to muons ($\mathrm{B}_\mu = 1$) or to τ_h leptons ($\mathrm{B}_\tau = 1$) or in the flavor-democratic scenario ($\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$).

3. Three years of data-taking period, which could have minor differences in the input features due to detector conditions.

4. Seven multilepton channels with different relative importance for different signals.

5. Changing background composition in the multilepton channels needs to be considered as well while designing the MVA SRs.

7.1 Boosted Decision Trees

A decision tree is a machine learning model that builds upon iterative decision-making process to answer a question (classification) or providing probabilities (regression) for a particular decision. It tries to partition the data recursively into true or false category on the basis of it's input features at each node. The split at each node is chosen to maximize the information gain, and the process is repeated until some stop condition set by the user is met. The terminal nodes are known as leaf nodes, which denotes probability for a class. A simple structure of a decision tree for discriminating between a cat and a dog is shown in Figure 7.1.

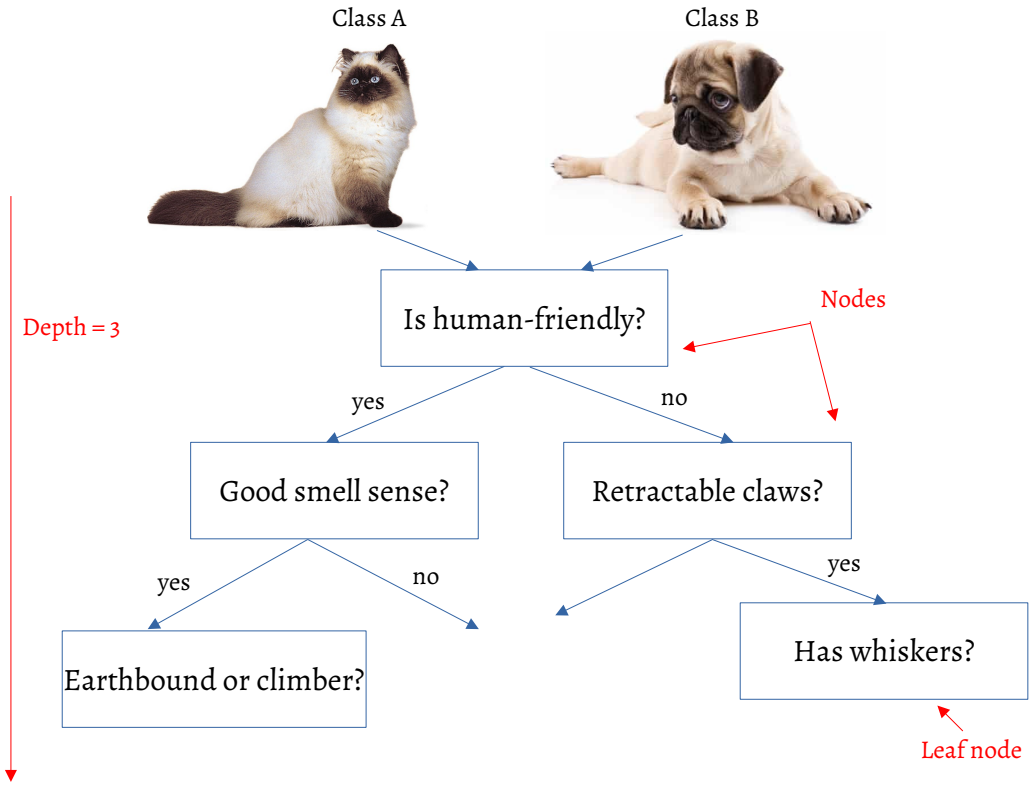

Figure 7.1: A simple structure of a Decision Tree for a classification task.

A single decision tree is prone to overfitting due to the presence of noise or the nature of the problem to be solved. Hence, different methods of combining decision trees are employed to reduce the biases in the training, and to improve the accuracy. This is known as ensemble learning, and the two most popular methods are – Random Forests and Gradient Boosting. While random forests are known to have higher accuracy, gradient boosting works better for cases with imbalanced data set i.e. when the input data set from signal and background in the classification are skewed.

In random forests, many decision trees are created in parallel, independently of the performance of other trees. Many random subsets of the same input data set are created and are trained on, before arriving at the final results by taking the ensemble average to do classification.

"Boosting" is the method of sequential learning by creating one decision tree at a time, using the feedback from the weak learners to make a stronger decision tree. Each tree attempts to minimize the errors of the previous ones, with the help of loss functions which defines the difference between truth and prediction. Usually, logarithmic loss is used for the classification task whereas

mean squared error loss is used for the regression. The learning can be reinforced by adding weights to stress on the difficult classification instances. Another way to reinforce the learning is via the "Gradient boosting" method, where instead of using weighted average of individual outputs from the decision trees as final output, a loss function is optimized to converge to the final result. This is done by minimizing the loss function using gradient descent in small steps.

Hyperparemeters are an essential part of the learning which effects the performance and the accuracy of a model. For BDTs, the most relevant hyperparameters that need to be tuned for a better learning are as follows:

1. **NTrees:** Number of trees to be created for the ensemble learning. More is better, but it could lead to overfitting.
2. **Maximum depth:** Number of nodes per decision tree before coming to a stop.
3. **Minimum node size:** Sets a limit on the node whether to split the input data further or not. This is done to maintain statistical robustness of the learning.
4. **NCuts:** Describes the density of grid size used to scan the best cut on the splitting parameter. It is a TMVA parameter.

7.2 Discriminant training strategy

In this analysis, my goal was to design an MVA algorithm which can discriminate between signal and total background shape, and not per SM process. Hence, a binary classification was found to be better suited than multi-class classification approach. Also, there is a huge disparity in the number of trainable events between background and signal. Within the class of backgrounds itself, the composition changes within and across the channels. BDTs have proven to be slightly beneficial over DNNs when it comes to low training statistics, as it is basically an iterative cut and classification approach unlike a DNN which tries to find a global minimum (among many local minima) of the loss function from the hyperparameters space.

Nevertheless, two sets of trainings were performed, one with a BDT and other one with a multiclassification-based DNN, using an optimized architecture for both. In case of BDT, the training was performed for discrimination between the vector-like lepton model in the doublet scenario versus the major backgrounds (WZ, ZZ, $t\bar{t}Z$,$Z\gamma$, DY, $t\bar{t}$), combined according to luminosity-based event weights, into one process. For the multiclassifier DNN training, the same processes were used in the training, but this time with one output neuron per process. The performance of

both the trained models was tested on a statistically independent sample, and is represented with the help of Receiver Operating Characteristic (ROC) curve. The ROC is plotted as a graph between the efficiency of selecting the signal, i.e. the true positive rate vs the efficiency of selecting the background as signal, i.e. false positive rate. Figure 7.2 shows two individual ROC curves for the testing performance from the BDT-based trained model as well as the multiclassifier DNN-based trained model on vector-like leptons of $m_{\tau'}$ = 300 GeV and $m_{\tau'}$ = 700 GeV. Clearly, the two performances are very similar. Hence, for simplicity reasons, BDTs were chosen as the choice of MVA for all the signal models considered in this analysis. The trainings are performed in ROOT 6.20/02 TMVA software package.

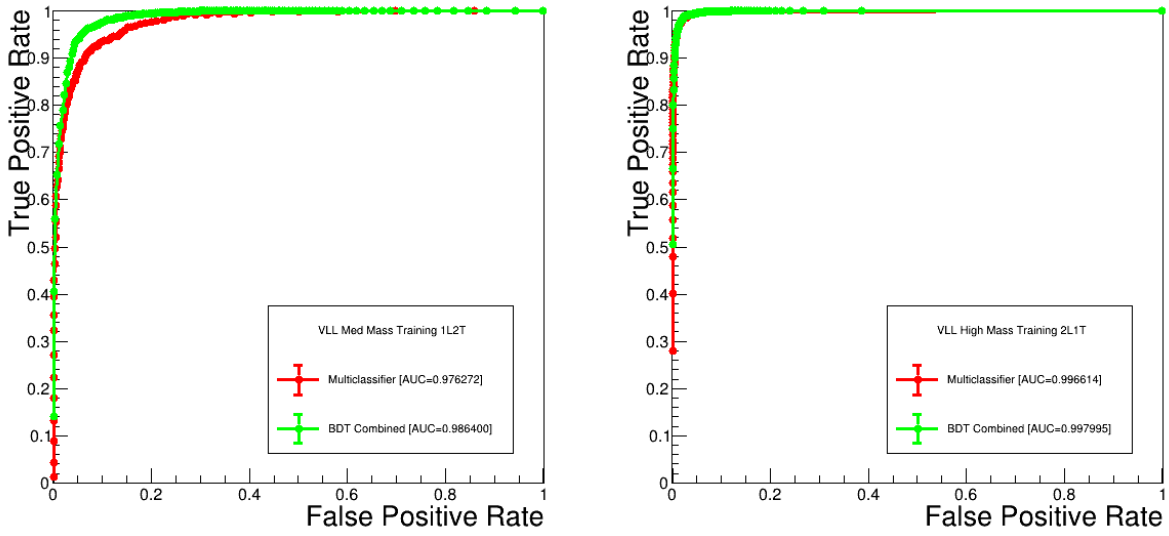

Figure 7.2: The ROC curves for the testing performance from the BDT-based trained model (green) as well as the multiclassifier DNN-based trained model (blue) on vector-like leptons of $m_{\tau'}$ = 300 GeV (left) and $m_{\tau'}$ = 700 GeV (right).

To mitigate the other challenges listed in the previous section, the following training strategy is employed:

1. **Model specific BDTs -** Each BSM signal will be trained for separately. This is important considering the differences in the kinematic properties of the three models. We denote signal-specific BDTs by VLL, SS, and LQ for vector-like tau lepton, type-III seesaw mechanism, and scalar leptoquark models, respectively.

2. **Mass- and flavor-wise BDTs -** For a given BSM model, small windows in signal particle masses as well as coupling scenarios are grouped together for training purposes, as appli-

cable. This results in typically three to four mass BDTs and one or two flavor BDTs per signal model. The mass-wise BDT splitting is done since the properties of signal vary considerably across the $\sim$ 2 TeV range, and therefore one BDT training cannot work across the mass spectrum considered for the different models. The signal properties, however, for electron- and muon-specific couplings are similar while different for τ-specific coupling in the input variables of interest. Hence the chosen splitting in flavor is not only necessary, but the simplication also increases the available number of signal events per training, yielding more performant trainings. The various mass ranges are denoted by VL (very low), L (low), M (medium), and H (high) for each signal, and the flavor BDTs are denoted by the branching ratio ($\mathcal{B}$) terms. Precise details about the mass and flavor splitting in the BDT trainings for each signal model are in Sections 7.5.1, 7.5.2, and 7.5.3.

To demonstrate that combining neighbouring masses in the MVA training improves or is at least similar to the performance in the application region by individual mass-wise trainings, a test-case training was performed using multi-class classification DNN[1] with vector-like lepton model as signal and major SM backgrounds. The neural network was trained with Doublet and Singlet VLL samples of mass $m_{\tau'} = 300$ GeV from 2016 and 2017, and the performance of the Doublet VLL was tested for application in 2018 using this trained model. A separate DNN was also trained with two neighbouring signal masses, i.e. $m_{\tau'} = 300$ GeV and 500 GeV from 2016 and 2017, and the performance of the Doublet VLL at both these masses was tested for application in 2018 using this trained model. Figure 7.3 left shows the DNN output from the signal neuron for the testing performance in 2018 evaluated from the combined training of 300 and 500 GeV in 2016 and 2017.

Figure 7.3 right shows the ROC curves of the performance at 300 GeV for the training in 2016 and 2017 with only 300 GeV (red), training in 2016 and 2017 with masses 300 and 500 GeV combined (green), and testing in 2018 from the combined mass training model (blue). Clearly, the combined training and testing performances are much closer to each other, than the individual training on one signal mass.

3. **BDTs per year of data-taking and statistical independence of training data set -** Individual BDTs are trained to discriminate a given signal process from the major SM backgrounds (WZ, ZZ, DY, $t\bar{t}$, $Z\gamma$) for each year of data-taking. For a training of the BDTs in a given year, signal and background MC samples of the other two years are used, i.e. for training a

[1]Four output neurons corresponding to four classes: Signal, DY+jets, $t\bar{t}$+jets, and WZ.
Network architecture: Input layer(256) + 4 Dense layers (128,64,32,14) + Output (4); Nepochs=100 and Batch_size=256

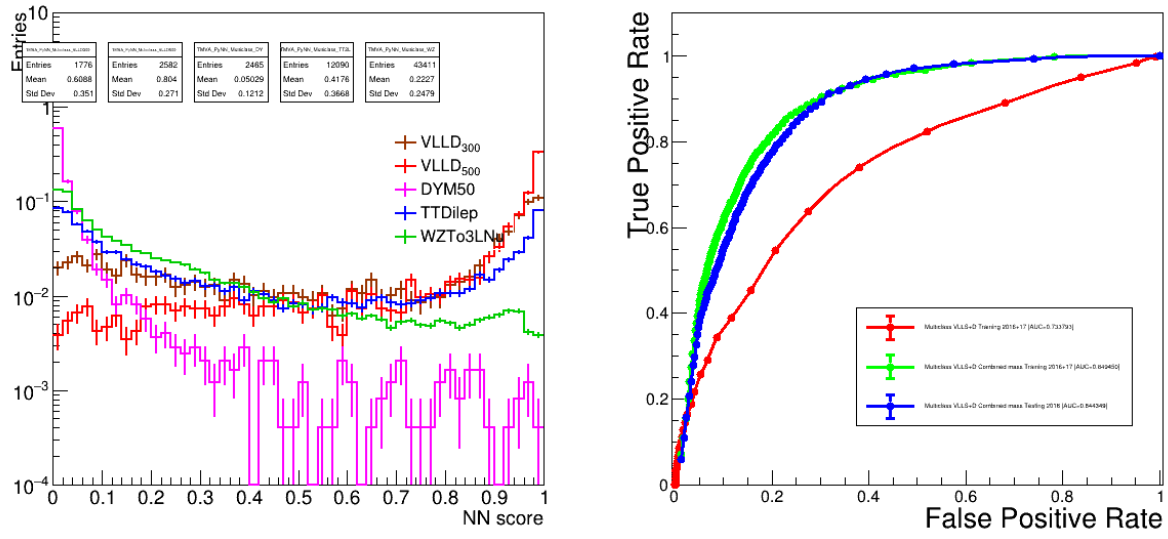

Figure 7.3: The DNN output score from the signal neuron (left) of the testing performance in 2018, evaluated from the combined training of vector-like leptons in the doublet and singlet scenario with masses $m_{\tau'} = 300$ and 500 GeV in 2016 and 2017. The ROC curves corresponding to the performance at 300 GeV for the training in 2016 and 2017 with only 300 GeV (red), training in 2016 and 2017 with masses 300 and 500 GeV combined (green), and testing in 2018 from the combined mass training model (blue) are shown in the right.

BDT to be used in the 2018 data set, the training is done using samples generated for 2016 and 2017 data sets. This ensures the statistical independence of the training and application samples, and minimizes the possibility of overtraining. Also, there is a four-fold increase in training statistics wrt to a 50–50% split of training and testing events, and no sacrifice of events from the application region.

Overtraining was checked by comparing training and application performance and if the ROC curves give close to similar area-under-curve (AUC) values. Hence, no major signs of overtraining were found. In any case, minor overtrainings that maybe present in the BDT training is not an issue in the analysis. Additionally, it was also checked that this choice of using samples from independent years in the training does not compromise the performance while evaluation. To demonstrate this, 2016 signal and background samples were split into two equal halves, one of which was used for BDT training and the other independent sample was used for evaluation. The evaluated performance is compared against the performance from 2017 and 2018 BDT training applied on all 2016 events. The resulting output BDT shapes and ROC curves are shown in Figure 7.4. It is clear by looking at the

AUC values of the two ROC curves as well as the shapes of signal and background that their is no compromise in performance due to this choice of training.

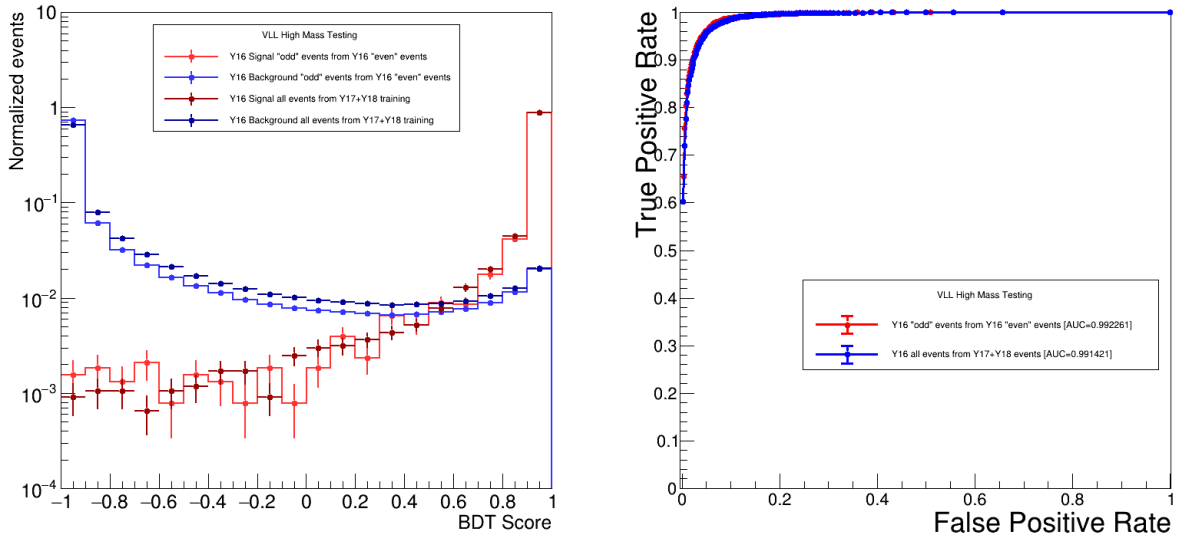

Figure 7.4: Figure (left) shows the BDT output score from two different trainings: the bright red and bright blue distributions show the testing performance by training with an independent half of 2016 events while the bright red and bright blue distributions show the testing performance of all the 2016 events by training with 2017 and 2018 samples. Figure (right) shows the testing ROC curves from the two different trainings. The red curve is the ROC from half of 2016 training while the blue curve is the ROC from 2017 and 2018 training.

The misidentified lepton background contributions are taken from the dileptonic DY and $t\bar{t}$ MC samples in the training, thus the training samples for the BDTs are completely independent from the samples used to make predictions, and from observations. In this way, no data events were sacrificed for the training, since that would reduce the number of events from the application region, and therefore causing a loss in sensitivity for the BSM search.

To check the validity of using dileptonic DY and $t\bar{t}$ MC samples as proxy for the data-driven misidentified lepton backgrounds in the training, two sets of test-case trainings were performed. In one training, DY+jets was trained against a prompt BSM phenomena (right-handed neutrinos of mass = 400 GeV) using a neural network, and the trained model was tested on a very small fraction (10%) of MisID background estimated from 2018 data in the 2L1T N_{b}= 0 events. The results are shown in Figure 7.5 left, where we can see that MisID reproduces the shape of the DY+jets background. Similarly, another neural network was trained against the same signal sample but with $t\bar{t}$+jets as background. The trained model

was applied on a sample of 10% of MisID background estimated from 2018 data in the 2L1T $N_{\mathrm{b}} > 0$ events. The results are shown in Figure 7.5 right, where again we can see that MisID reproduces the shape of the $t\bar{t}$+jets background.

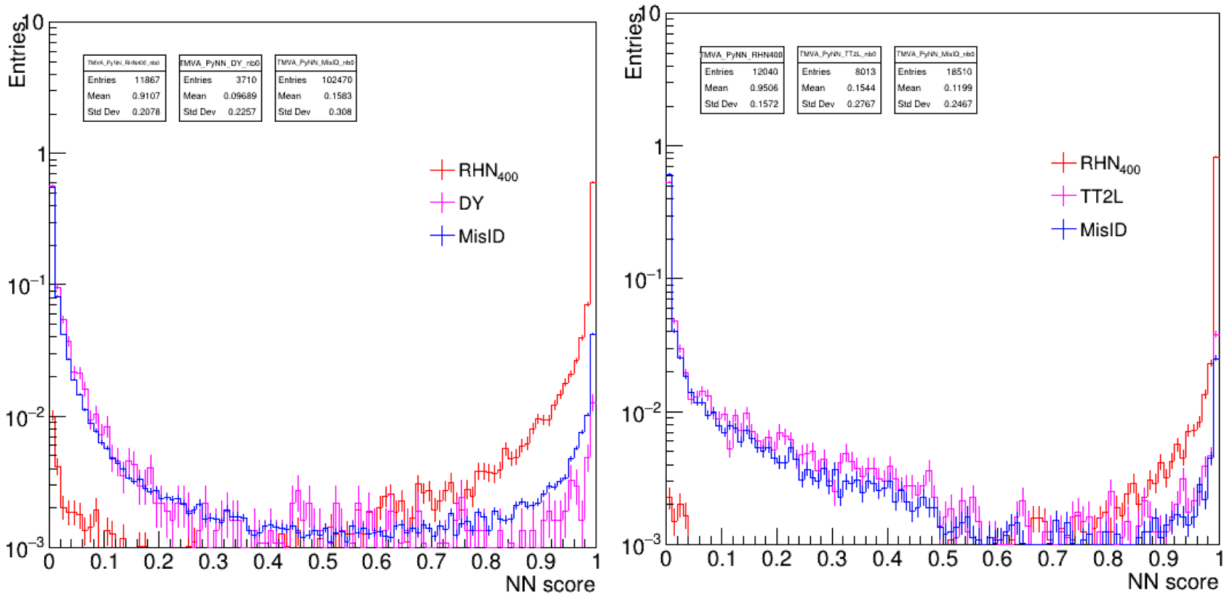

Figure 7.5: The performance of the DNN model trained on DY+jets vs signal (left) and $t\bar{t}$+jets vs signal (right) on the MisID background estimated from 2018 data in 2L1T N_{b}= 0 and $N_{\mathrm{b}} > 0$ events, respectively. The red and pink distributions shows the training performance on the signal and dedicated background, respectively, while the distribution in blue is the testing performance on the MisID background. Only statistical uncertainties are shown.

4. **One channel-inclusive training -** The BDT trainings consider all multilepton channels in a combined way, but separately for every year. Events from all the channels are fed simultaneously to the training so that BDT can learn about the implicit importance of each channel through the relative acceptances of the signal and background processes. This reduces the number of trainings by a factor of five, without any compromise in the performance.

   To test this, test-case trainings were performed using type-III seesaw fermions of $\mathrm{m}_{\Sigma} = 200$ GeV as signal and major SM backgrounds separately for the 3L channel as well as a combined training with all channels. Figure 7.6 left shows the ROC curves of the testing performance for the 3L events using dedicated 3L BDT training (blue), channel-inclusive BDT training (green), and a channel-inclusive multi-class classification DNN training (red). Similarly, another set of test-case trainings were performed for medium mass seesaw sample i.e. $\mathrm{m}_{\Sigma} = 500$ GeV in the 2L1T channel, and the ROC curves of the testing performance

are shown in Figure 7.6 right. From both the figures, we can conclude that one channel-inclusive training works just as well as the channel-wise training, with the hindsight benefit in the former case of a huge reduction in the number of trainings to be performed ultimately.

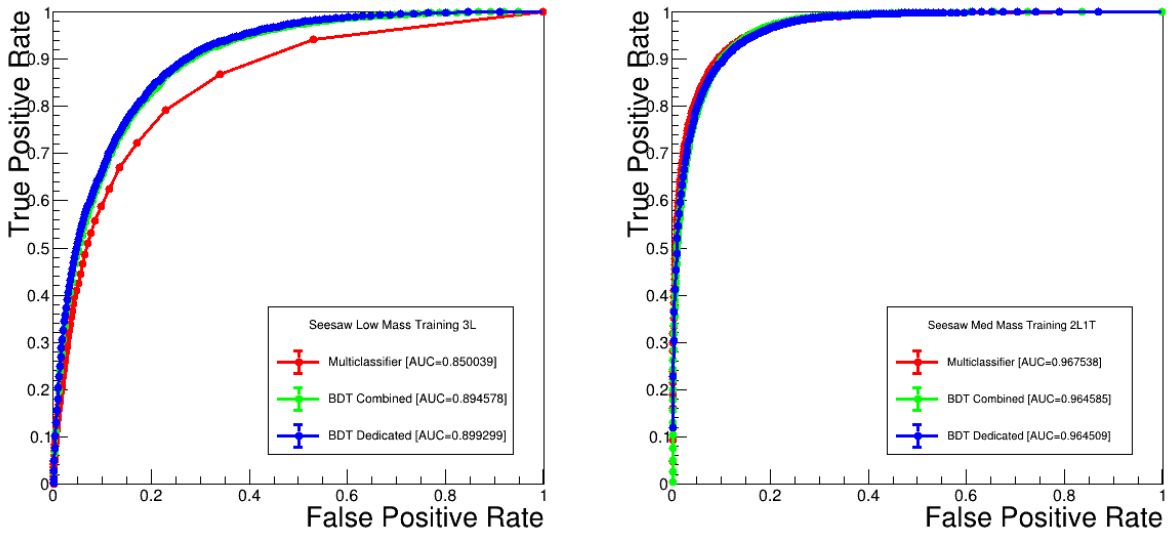

Figure 7.6: ROC curves for the testing performance of the channel-dedicated BDT training (blue), channel-inclusive BDT training (green), and channel-inclusive multi-class classification DNN training (red) for low mass seesaw fermions in 3L events (left) and medium mass seesaw fermions in 2L1T events (right).

5. **Lumi-weighted training and novel input variables -** To account for the changing background composition, luminosity-based weights are applied on the background processes in the training so as to take care of the relative contribution of each background in the seven channels. Also, novel input variables are used in the training which highlights the different features of events with light leptons and hadronic taus.

   There are a total of 48 training input variables for the vector-like lepton model BDT trainings, and 23 input variables for each of the type-III seesaw and the leptoquark model BDT trainings, as summarized in Table 7.1. The input variables consist of object- and event-level quantities such as transverse momenta, invariant (transverse) masses, and angular variables. The 48 training variables for VLL also include 19 additional categorical variables that are based on a simplified version of the fundamental scheme 8.1.

Following the strategy above, the multilepton analysis has 57 distinct BDT trainings, with one BDT output spectra per training. This is summarized in the flowchart in Figure 7.7.

Table 7.1: Input variables used for the BDTs trained for the various BSM models. Note that the indices i, j run over the leptons of all flavors ($i, j = 1, 2, 3, 4$) in a given event. If a given variable is not defined in a given channel, the variable is set to a nonphysical default value for signal and background processes, and plays no role in training.

Variable type	Used for		
	All signals	Vector-like lepton	Seesaw and leptoquarks
Event	H_{T}, $p_{\mathrm{T}}^{\mathrm{miss}}$, N_{b}, M_{ℓ}	Q_{ℓ}	L_{T}, $p_{\mathrm{T}}^{\mathrm{i}}/L_{\mathrm{T}}$, $L_{\mathrm{T}}/S_{\mathrm{T}}$, $H_{\mathrm{T}}/S_{\mathrm{T}}$, $p_{\mathrm{T}}^{\mathrm{miss}}/S_{\mathrm{T}}$
Lepton	$p_{\mathrm{T}}^{\mathrm{i}}$, $p_{\mathrm{T}}^{\mathrm{OSSF}}$		
Angular	ΔR_{min}	Max, Min: $\Delta\phi^{\mathrm{i}}$, Max, Min: $\Delta\phi^{\mathrm{ij}}$	Max: $\Delta\eta^{\mathrm{ij}}$
Mass	$M_{\mathrm{T}}^{\mathrm{i}}$	M^{ij}, M_{T}^{12}, M_{T}^{13}, M_{T}^{23}	

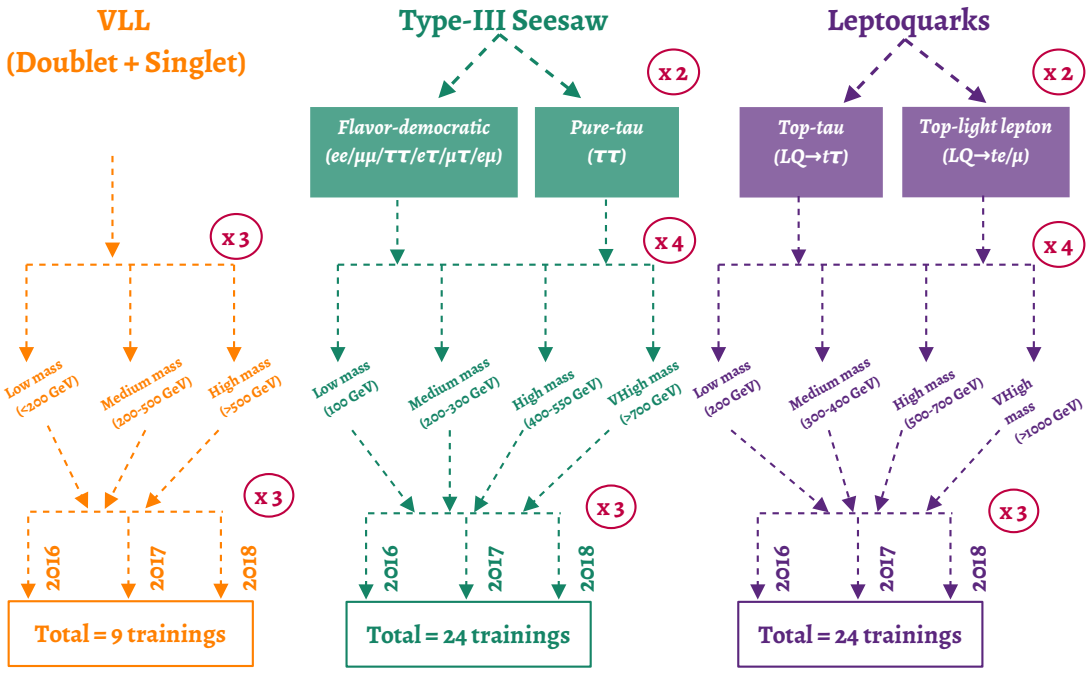

Figure 7.7: Summary flowchart of the 57 distinct BDT trainings in this multilepton analysis.

In addition to the strategy outline above, all control region selections are vetoed for the selection of events used in the BDT trainings. These control regions thus serve as a cross-check while evaluating the performance of the trained BDTs. Several distributions of various training input variables are illustrated in Figure 7.8 from different control regions. The SM backgrounds in all considered input variables are well modelled and found to be in good agreement with the data.

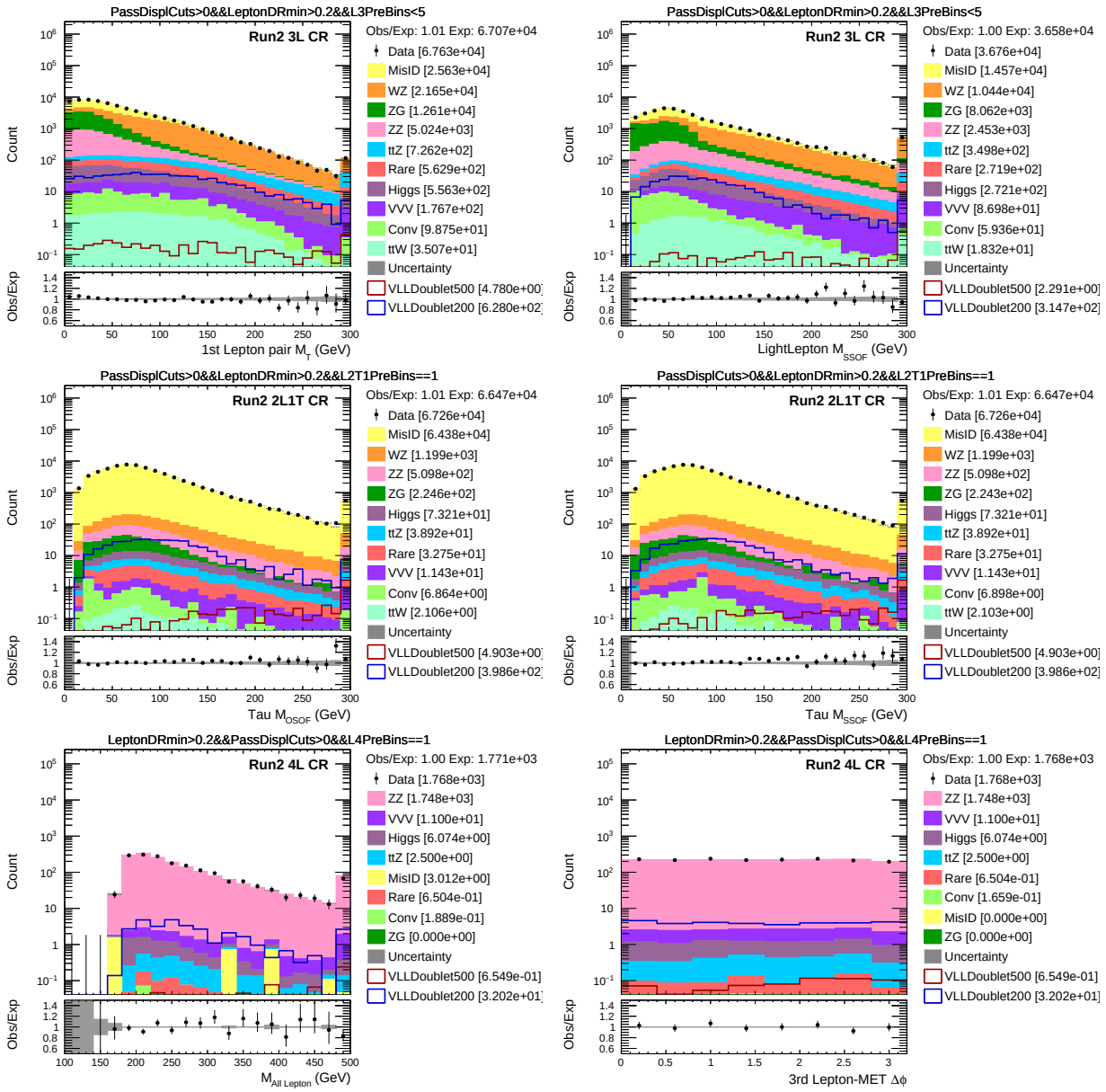

Figure 7.8: A few BDT training input variables in the control regions. The upper row shows the $M_{\mathrm{T}}(\ell'_{12} \cdot \vec{p}_{\mathrm{T}}^{\,\mathrm{miss}})$ (left) and the $M_{\mathrm{SS}}^{e\mu}$ (right) in the 3L control region. The middle row shows the $M_{\mathrm{OS}}^{\ell\tau}$ (left) and the $M_{\mathrm{SS}}^{\ell\tau}$ (right) in the 2L1T control region. The lower row shows M_{ℓ} (left) and the $\Delta\phi(\ell'_3 \cdot p_{\mathrm{T}}^{\mathrm{miss}})$ (right) in the 4L control region. The figures are shown with statistical uncertainties only.

All BDTs used in this analysis have 800 trees (NTrees), with a maximum depth of 10. The minimum node size is 1.5% with 10 steps (NCuts) during the node cut optimization. The *GradientBoost* algorithm is chosen for boosting the trees. The choices made in the training of the BDTs

have been systematically examined to be robust and well-performing. The BDT hyperparameters are varied and the performance of the BDT is assessed by evaluating the ROC curve for signal against background. Figure 7.9 shows the effect of these variations for a representative *LQ-M* training, where no significant change is observed in the performance of the BDT.

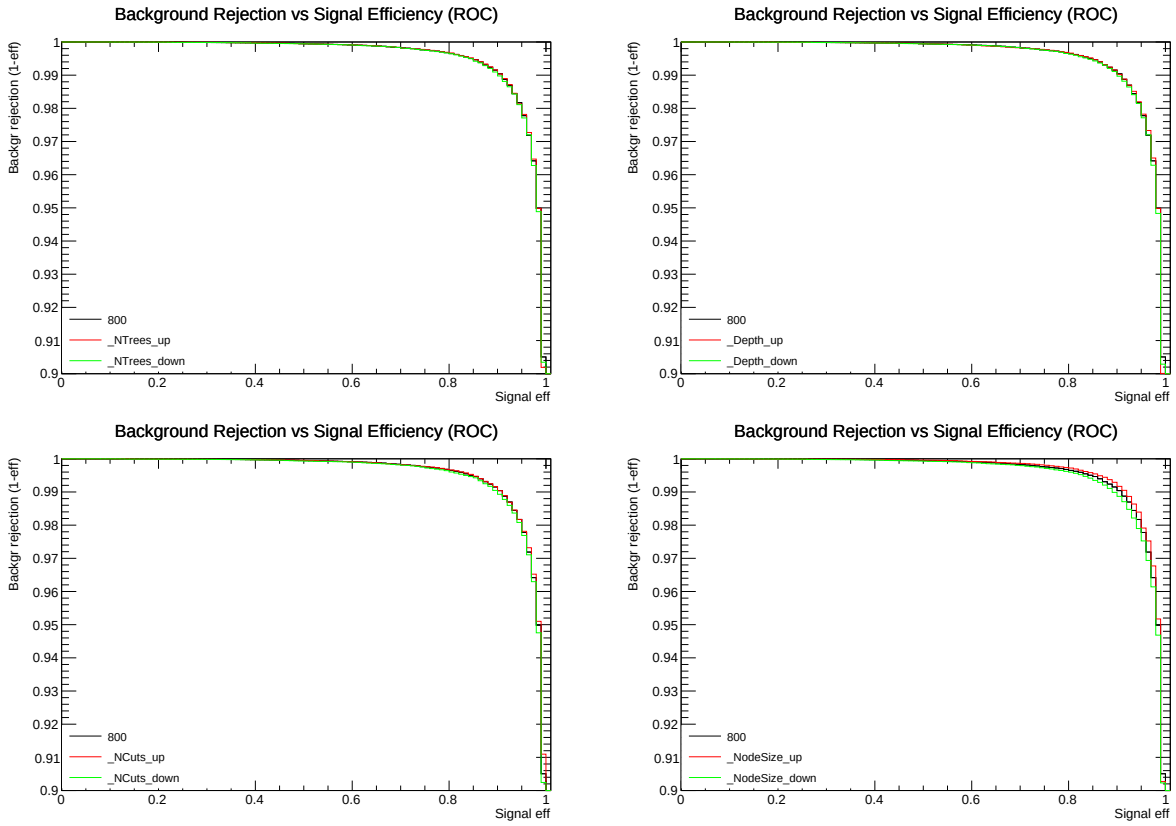

Figure 7.9: ROC curves for a representative low-mass LQ training as a function of variations in hyperparameters. Ntrees is varied from 800 (nominal) to 600 (down) and to 1000 (up) (upper left), depth varied from 10 (nominal) to 6 (down) and to 14 (up) (upper right), NCuts varied from 10 (nominal) to 6 (down) and 14 (up) (lower left), an the node size varied from 1.5% (nominal) to 2.5% (up) and to 0.5% (down) (lower right).

7.3 Discriminant application

Each of the 57 distinct BDT trainings provide an output score in the range of (-1,1) for each event to be background- or signal-like. This BDT score is then treated as the primary discriminating variable to perform counting experiments for the BSM search in this multilepton analysis. As

illustrated in the score distribution of any of the BDTs, the region with highest sensitivity to the signal is typically very close to the maximum BDT output. On the other hand, the region around the low end of the BDT output is background dominated, and has very little sensitivity to the signal.

To increase the sensitivity for the BSM search, a number of regions of variable widths across the BDT spectrum are defined. More boundaries are defined on the high score side to achieve good signal-to-noise ratio, maintaining a smooth and well-behaved expected background yield. This is achieved by extracting the first bin at the right most end of the BDT score with a total background yield of approximately 1 event, and then defining further bins towards the left in increasing step size of multiples of 0.0005, such that the event yield is smoothly increasing in each bin. This variable binning strategy can be realized through the schematic in Figure 7.10.

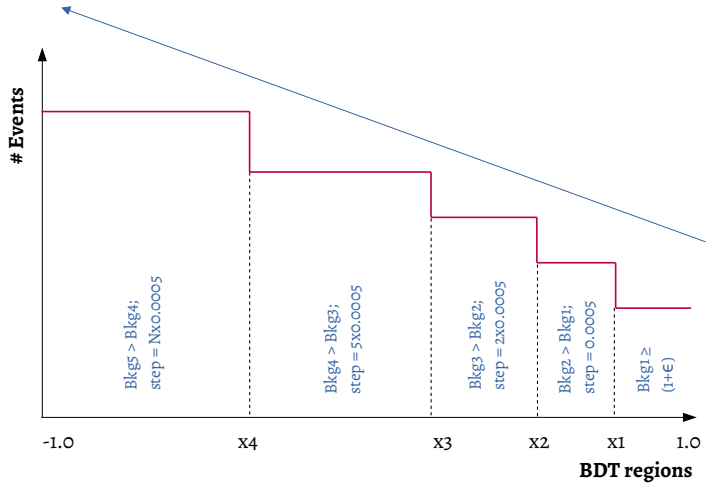

Figure 7.10: Explanation of the variable binning strategy for the BDT output score distribution through a simple schematic diagram.

These transformed distributions are referred to these as the *BDT regions* in order to distinguish them from the uniform width binning of the BDT spectra. Also, the 3-object channels (3L, 2L1T and 1L2T) and the 4-object channels (4L, 3L1T, 2L2T, 1L3T) are combined together so that we have a more uniform BDT spectra from channels with lower expected background yields, satisfying the minimum SM background yield requirement. As an example, Figure 7.11 shows the *VLL-H* BDT output score in an uniformly binned distribution and the corresponding BDT regions for the 3-object channels in the combined 2016–2018 data set.

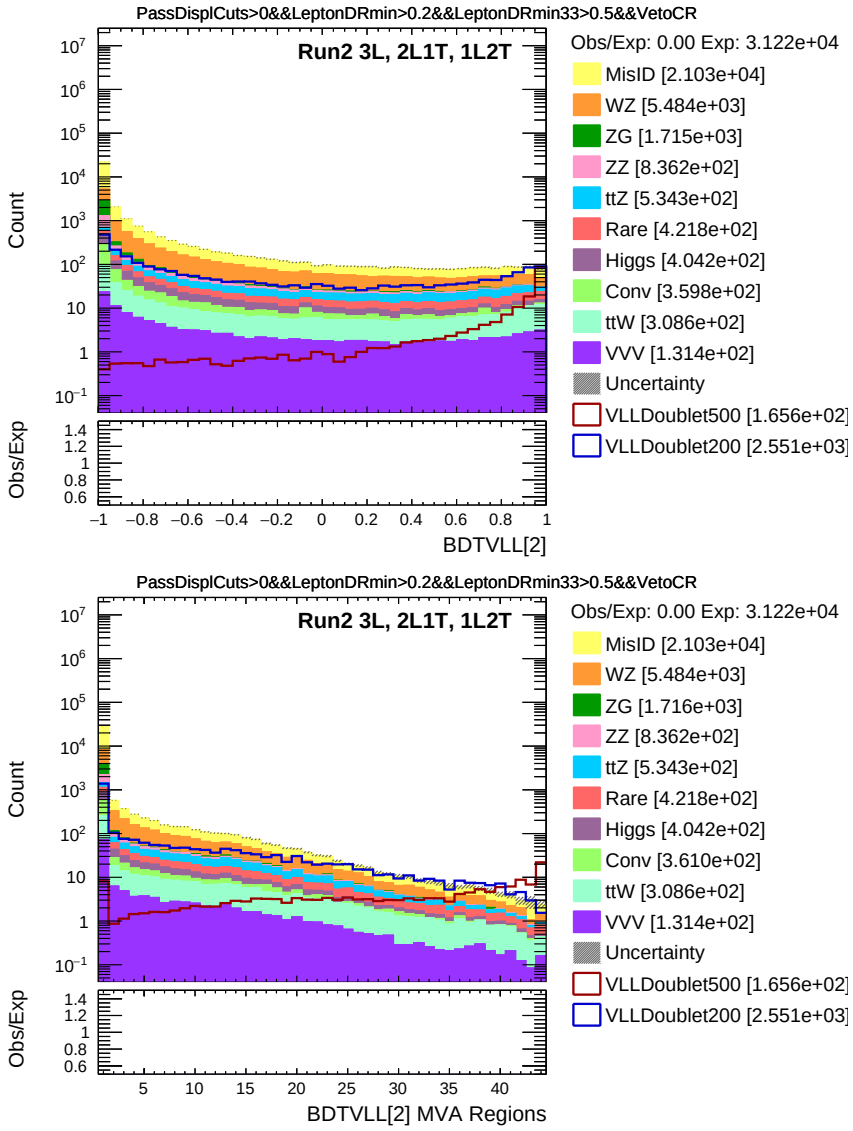

Figure 7.11: The *VLL-H* BDT output score in an uniformly binned distribution (left) and the corresponding BDT regions (right) for the combined 2016–2018 data set in the 3-object channels.

It is quite noticeable from the Figure 7.11 how the signal-to-background ratio improves in the most sensitive bins of the BDT regions, as compared to the uniformly binned BDT output score distribution. To quantify this, Table 7.2 presents a comparison of the signal significance, i.e. $S/\sqrt{B}$ in the most sensitive bins of the uniformly binned BDT output score vs the BDT regions for the *VLL-H* BDT training in 3-object channels as shown in Figure 7.11. Overall, there is an improvement in the signal sensitivity by around 20–40% using this transformation scheme on the output score.

Table 7.2: Comparison of the signal significance, i.e. $S/\sqrt{B}$ in the most sensitive bins of the uniformly binned BDT output score versus the BDT regions for the *VLL-H* BDT training in the 3-object channels as shown in Figure 7.11.

Bin number	Uniform BDT output		BDT regions	
	Bin boundary	$S/\sqrt{B}$	Bin boundary	$S/\sqrt{B}$
Nbin-3	0.8–0.85	0.7	0.9685–0.974	3
Nbin-2	0.85–0.9	1.1	0.974–0.9795	5.2
Nbin-1	0.9–0.95	2.1	0.9795–0.9845	4.2
Nbin	0.95–1	9.5	0.9845–1	12.6

Thus, to perform the search for the BSM signal with highest sensitivity, this transformation procedure is applied to all the mass- and flavor-BDT trainings per model separately, as well as separately for the 3-object and 4-object channels in the three years of the data-taking. This results in an $\mathcal{O}(100)$ independent BDT-based search bins per mass parameter of the model. Each of these bins are treated as counting experiments, and are fitted simultaneously for each of the three years of data collection to derive the final results.

7.3.1 Systematic uncertainties

To assess the effect of systematic uncertainties, as covered in Section 6.3, on the BDT output, the uncertainties were propagated from each of the sources on the BDT regions per year corresponding to every BSM signal for the 3-object and 4-object channels separately. The impact on the major backgrounds (WZ, ZZ, $t\bar{t}Z$ and $Z\gamma$) was analyzed and the relative variation wrt the nominal scenario was taken as the final systematics band along with the statistical uncertainties, while computing the constraints on the BSM phenomena. Below are some examples of the impact of major systematic sources on the BDT regions.

Figure 7.12 shows example variations of the electron (upper), muon (middle), and jet (lower) energy scale systematic uncertainties on WZ in 2016, ZZ in 2017, and $t\bar{t}Z$ in 2018, respectively. The upper and lower variations are shown for the *VLL-H* BDT regions in 3-object and 4-object channels, respectively. The middle variations are shown for the *SS-H* $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ BDT regions. The nominal variation is set at unity and the up variation (red) and down variation (blue) are shown wrt the nominal. The gray band is the statistical uncertainty per bin.

Figure 7.13 shows example variations of the misidentified lepton background systematic uncertainties for low p_T electrons in 2016 (upper), medium p_T τ_h in 2017 (middle), and high p_T muons in 2018 (lower). All the variations are shown for the *SS-H* $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ BDT regions in the 3-object channels. The nominal variation is set at unity and the up variation (red) and down variation (blue) are shown wrt the nominal. The gray band is the statistical uncertainty per bin.

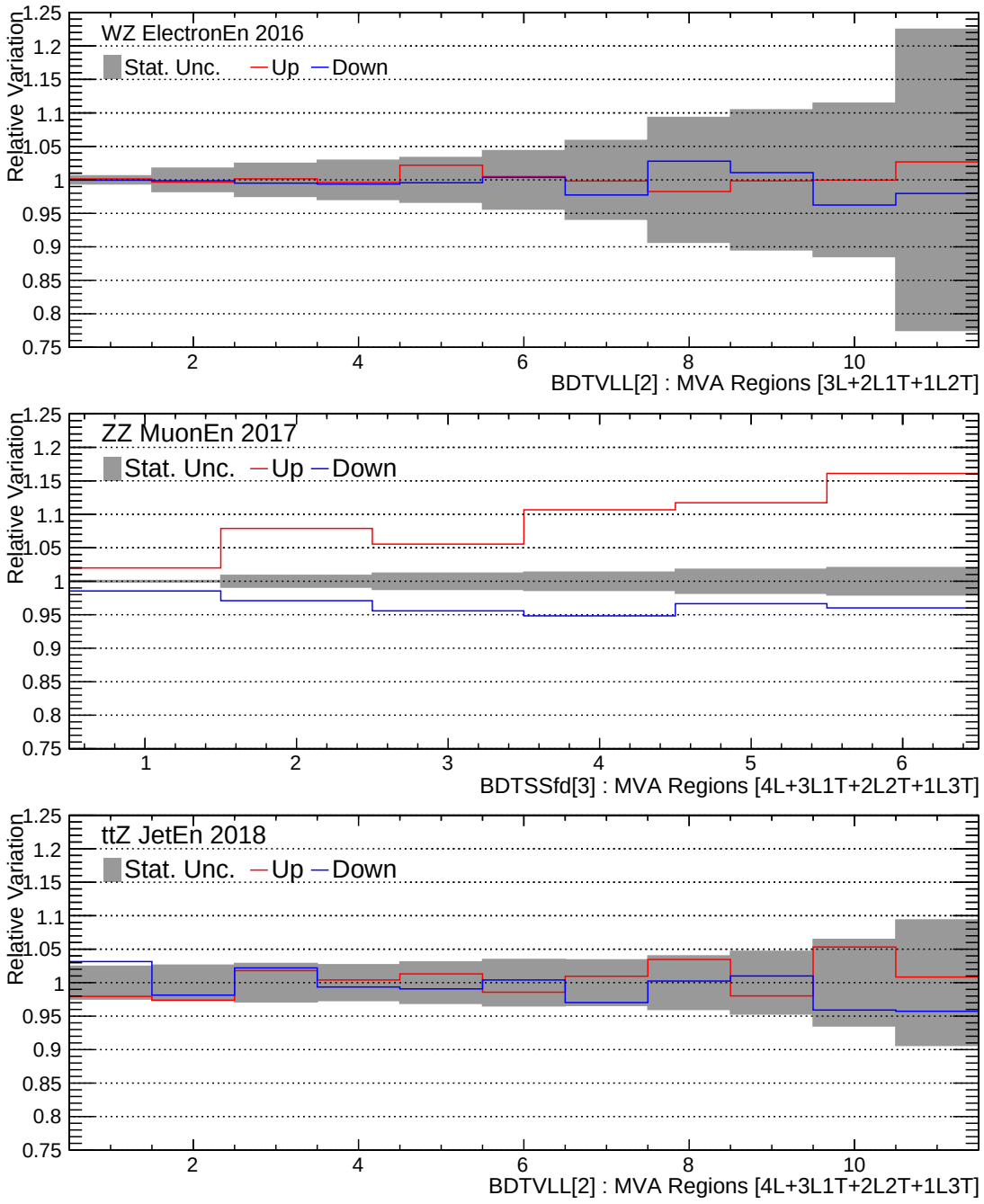

Figure 7.12: Example variations of the electron (upper), muon (middle), and jet (lower) energy scale systematic uncertainties on WZ in 2016, ZZ in 2017, and $t\bar{t}Z$ in 2018, respectively. The upper and lower variations are shown for the *VLL-H* BDT regions in 3-object and 4-object channels, respectively. The middle variations are shown for the *SS-H* $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ BDT regions. The nominal variation is set at unity and the up variation (red) and down variation (blue) are shown wrt the nominal. The gray band is the statistical uncertainty per bin.

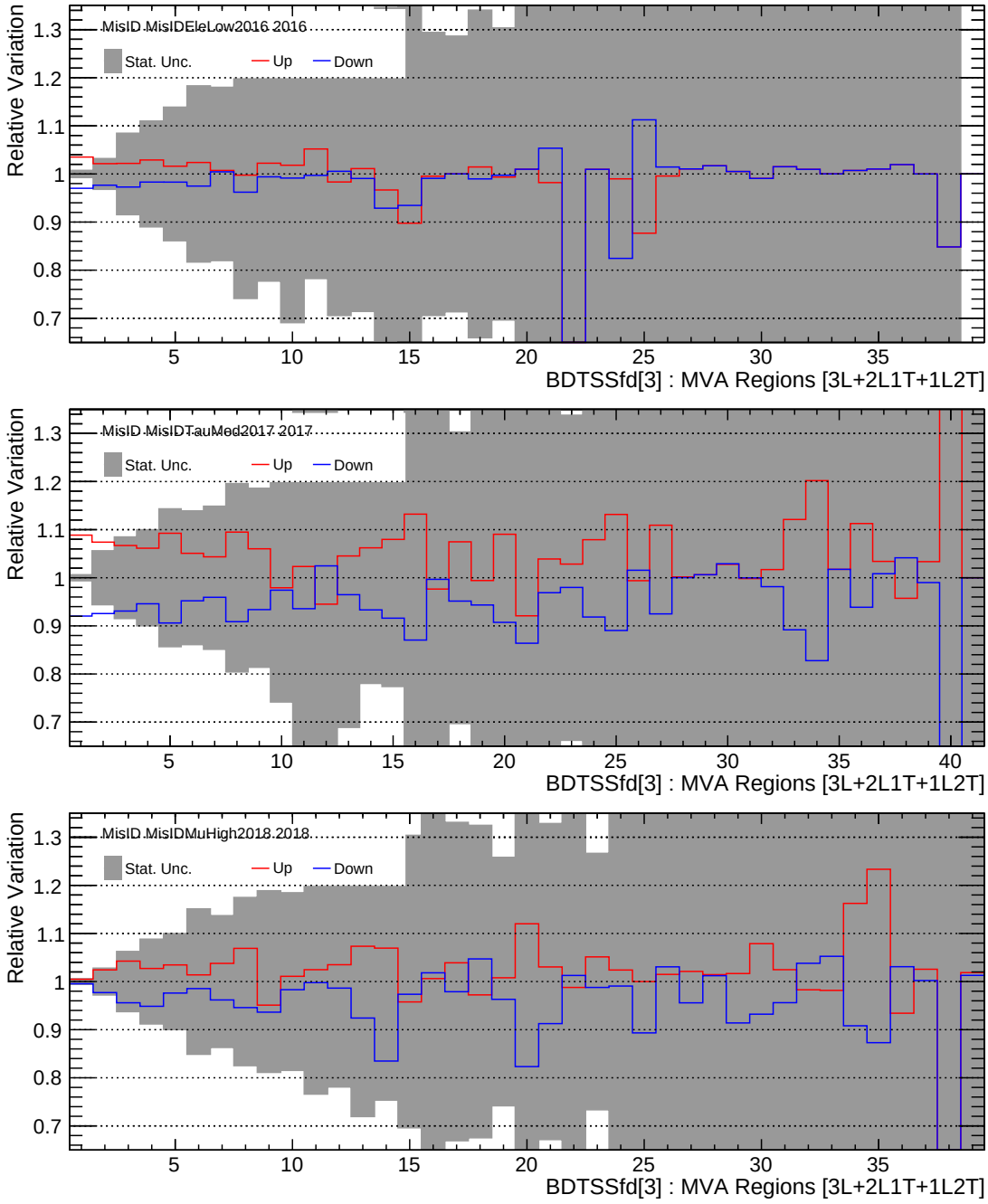

Figure 7.13: Example variations of the misidentified lepton background systematic uncertainties for low p_{T} electrons in 2016 (upper), medium p_{T} τ_{h} in 2017 (middle), and high p_{T} muons in 2018 (lower). All the variations are shown for the *SS-H* $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ BDT regions in the 3-object channels. The nominal variation is set at unity and the up variation (red) and down variation (blue) are shown wrt the nominal. The gray band is the statistical uncertainty per bin.

Figure 7.14 shows example variations of the diboson jet multiplicity modeling systematic uncertainties on the ZZ (upper) and WZ (lower) backgrounds in 2018. The ZZ variations are shown for the *VLL-H* BDT regions in the 4-object channels and the WZ variations are shown for the *SS-H* $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ BDT regions in the 3-object channels. The nominal variation is set at unity and the up variation (red) and down variation (blue) are shown wrt the nominal. The gray band is the statistical uncertainty per bin.

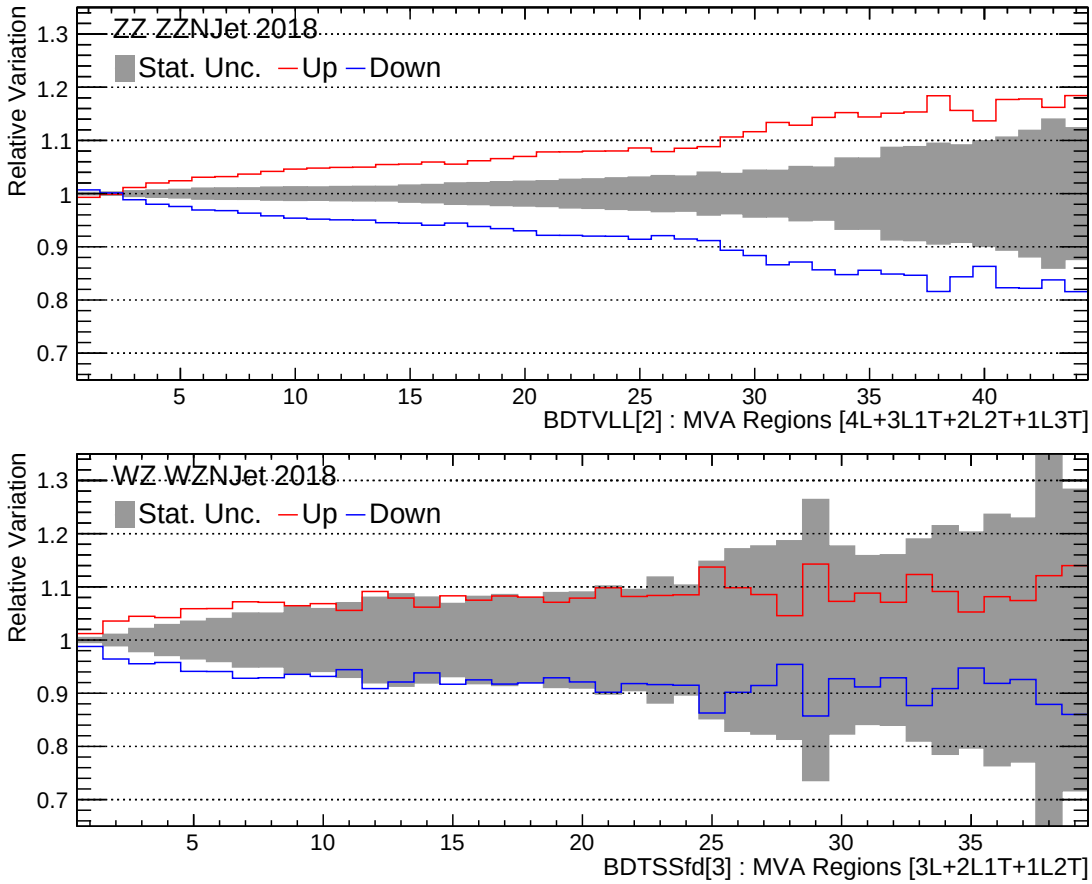

Figure 7.14: Example variations of the diboson jet multiplicity modeling systematic uncertainties on the ZZ (upper) and WZ (lower) backgrounds in 2018. The ZZ variations are shown for the *VLL-H* BDT regions in the 4-object channels and the WZ variations are shown for the *SS-H* $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ BDT regions in the 3-object channels. The nominal variation is set at unity and the up variation (red) and down variation (blue) are shown wrt the nominal. The gray band is the statistical uncertainty per bin.

Figure 7.15 shows example variations resulting from the τ identification uncertainties from the VSjet discrimination on the WZ in 2018 (upper) and ZZ in 2016 (lower) backgrounds. The WZ variations are shown for the *VLL-H* BDT regions in the 3-object channels and the ZZ variations are shown for the *VLL-H* BDT regions in the 4-object channels. The nominal variation is set at unity and the up variation (red) and down variation (blue) are shown wrt the nominal. The gray band is the statistical uncertainty per bin.

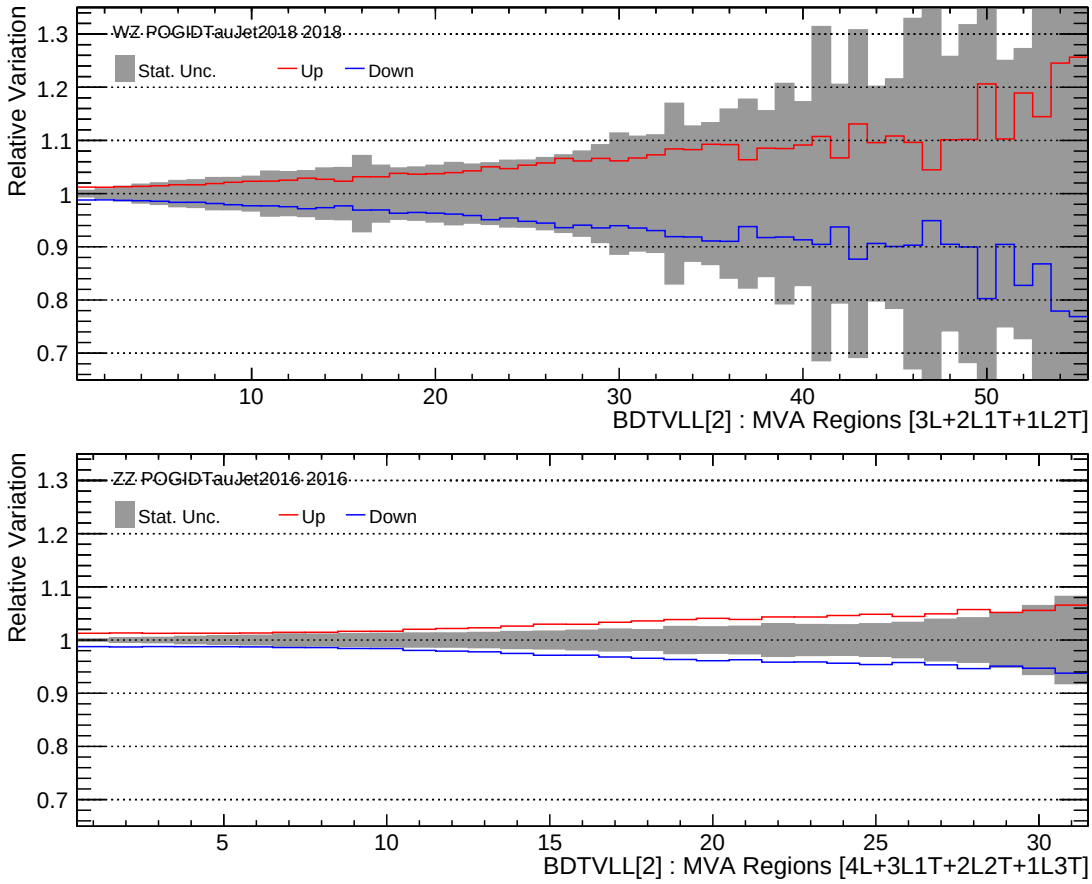

Figure 7.15: Example variations resulting from the τ identification uncertainties from the VSjet discrimination on the WZ in 2018 (upper) and ZZ in 2016 (lower) backgrounds. The WZ variations are shown for the *VLL-H* BDT regions in the 3-object channels and the ZZ variations are shown for the *VLL-H* BDT regions in the 4-object channels. The nominal variation is set at unity and the up variation (red) and down variation (blue) are shown wrt the nominal. The gray band is the statistical uncertainty per bin.

7.3.2 Validation in CRs

Figure 7.16 shows the output from the *SS-M* BDT in the flavor-democratic scenario, with statistical and systematic uncertainties in the SM background prediction. The BDT output is shown in the 4L ZZ CR, and in the combined 3L OnZ, 3L $Z\gamma$ and 2L1T MisID CRs, and the data are observed to be in good agreement with the expected SM background prediction.

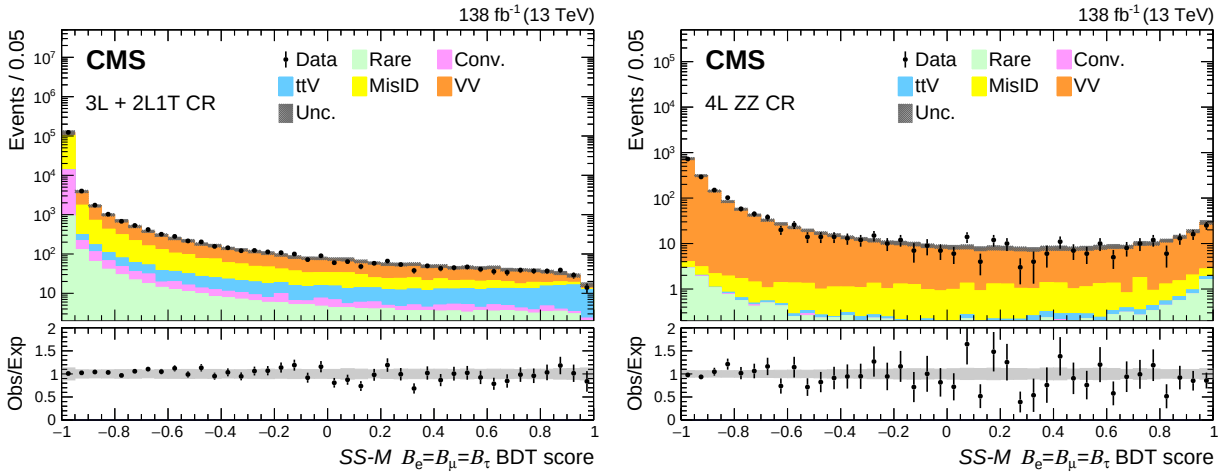

Figure 7.16: Distributions of BDT score from the *SS-M* $\mathcal{B}_e = \mathcal{B}_\mu = \mathcal{B}_\tau$ BDT are shown for the 3L+2L1T CR (left), and the 4L ZZ CR (right). The 3L+2L1T CR consists of the 3L OnZ, 3L $Z\gamma$, and 2L1T MisID CRs. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction.

7.4 Limit setting using statistical analysis

The advancement of the scientific knowledge is a two-way street between elegant theories and novel experiments. Theories are built on few assumptions and are mathematically consistent with a minimal number of arbitrary parameters to be determined from the experiments. However, these theories can be completely rejected or modified if the results are inconsistent with the observations. To build an understanding of the need for statistical methods in this process of accepting or rejecting the theory, let us take the following example.

In particle physics, we observe events (e.g. proton-proton collisions at LHC) and measure a set of properties of each of those events. These can be multiplicity of the particles, four-momentum per particle, and some other global event property like number of secondary vertices. We then

compare the observed distributions of these properties to the predictions from the theory known up to certain free parameters, e.g. α, m_{Z} or m_{H}. By doing so, we can estimate the free parameters of the theory, quantify the uncertainties in the measured parameter(s), and finally assess if the theory stands in agreement with the data.

There are many challenges in particle physics research. First of all, there are a multitude of BSM theories which could potentially account for the various unexplained phenomena of the universe. But these theories have many free parameters, some of which may not be exactly determined from the experiments. Also, there are a lot of uncertainties resulting from the the experiments, both random and systematic, which plays a role in the determination of the parameters of the theory. Hence, in order to make precise conclusions about the nature and existence of the theory, we need to build "probabilities". The subsequent discussion follows Ref. [160, 161].

7.4.1 Probability theory and inferences

There are two ways of defining the probability and the applicability of each definition depends on the kind of claim we are going after. These are as follows:

1. **Frequentist probability -** Relates probability to the fraction of times a favorable event occurs, in the limit of very large number of repeated trials. Thus, if A and B are outcomes of a repeatable experiement, then:

$$P(A) = \lim_{N\to\infty} \frac{\text{Number of times outcome is A}}{\text{Number of repeated trials (N)}} \tag{7.1}$$

$$G(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{7.2}$$

A frequentist inference procedure determines a central value (μ) and an uncertainty interval (σ) that depends on the observations. The measurement is modeled by a Gaussian distribution G(x; μ, σ), as given in Eqn. 7.2. For a large number of hypothetically repeatable experiments, the interval [x-σ,x+σ] would contain the true value (μ) in 68% of the cases. Hence, in the frequentist approach, we do not make probabilistic statements about the true value, it is what it is.

The function that returns the central value given an observed measurement is called an estimator. The most frequently adopted is the "maximum likelihood" estimator. The parameter value provided by an estimator is also called the best fit value.

2. **Bayesian or subjective probability -** Expresses one's degree of belief that the claim is true, with a probability equal to 1 expresses with certainty that the claim is true while 0 expresses with certainty that the claim is false. This definition is applicable to all unknown events/claims, and not only on repeatable experiments. Thus, if A is a hypothesis, then:

$$P(A) = \text{Degree of belief that A is true} \tag{7.3}$$

These subjective probability can be modified after learning about some observation or evidence. Those rules descend from the Bayes theorem, hence the name Bayesian probability. Thus, starting from a prior claim following some observation, the posteriori probability can be constructed from the law of conditional probability:

$$P(H|x) = \frac{P(x|H)\pi(H)}{\int P(x|H)\pi(H)dH} \tag{7.4}$$

Here, P(H|x) is the posterior probability of the hypothesis H after looking at the data, P(x|H) is the probability of the data assuming hypothesis H i.e. the likelihood, $\pi(H)$ is the prior probability of the hypothesis before looking at the data, and the denominator is the normalization term which involves summing over all possible hypothesis. Hence, the bayesian probability tells about the evolution of prior probability after looking at the data.

7.4.2 Likelihood

The likelihood is defined as,

$$\mathcal{L}(\vec{\alpha}) \propto p(data|\vec{\alpha}) \tag{7.5}$$

where the likelihood parameter, $\vec{\alpha} = (\vec{\mu}, \vec{\theta})$. Here, $\vec{\mu}$ are the Parameters Of Interest (POIs) that we want to measure and $\vec{\theta}$ are the nuisance parameters. The nuisance parameters $\vec{\theta}$ are often constrained from external measurements, for e.g. luminosity measurement and jet energy resolution. The vector notation implies that the parameters are a set with many components. Hence, introducing the constraint term in the prior probability as $\pi(\vec{\theta_0}|\vec{\theta})$, where θ_0 is the measured nominal value of the given nuisance parameter. Rewriting the likelihood from Eqn. 7.5 as,

$$\mathcal{L}(\vec{\alpha}) \propto p(data|\vec{\alpha}).\pi(\vec{\theta_0}|\vec{\theta}) \tag{7.6}$$

To determine the probability and the constraint terms, consider the following example. Let's assume we have an analysis which counts the number of events in proton-proton collisions. The data is just the number of events N that we observe. We also have a model for the number of events, n_{exp}, that we expect. Typically, we have a reference cross section for our signal process and known cross sections for our background processes - then POI might be the signal cross section relative to the reference cross section: $\sigma_{sig}^{lim} = \mu\sigma_{sig}^{ref}$. Then, n_{exp} can be written as:

$$n_{exp} = \mu\sigma_{sig}\epsilon_{sig}A_{sig}L^{int} + \sigma_{bkg}\epsilon_{bkg}A_{bkg}L^{int} \tag{7.7}$$

where ϵ, A, and L^{int} are selection efficiency, detector acceptance, and integrated luminosity of the data set. Hence, the probability term in the likelihood $p(\mathrm{data}|\mu, \vec{\theta})$ becomes the Poissonian probability of observing N data events, given the expected distribution n_{exp}.

$$p(N|n_{exp}) = \frac{n_{exp}^{N}e^{-n_{exp}}}{N!} \tag{7.8}$$

7.4.3 Treatment of nuisance parameters

Any of the terms in Eqn. 7.8 can have uncertainties associated with them. Hence, n_{exp} is affected by the presence of nuisance parameters. Let's consider an example of uncertainty of 2.5% on the integrated luminosity. Hence, the number of observed events could increase by 2.5% (multiplication by 1.025) or decrease by 2.5% (division by 1.025). So, $\mathcal{L}^{int} \rightarrow \mathcal{L}^{int}(1 + 0.025)^{\theta}$. For $\theta = 0$, $\mathcal{L}^{int}$ doesn't change. For $\theta = \pm 1$, we get $\pm 1\sigma$ uncertainty. Applying a gaussian constraint on the nuisance parameter, $\pi(\theta_0|\theta) = \pi(0|\theta) = e^{\frac{-\theta^2}{2}}$. Hence, the nuisance parameter is log-normally distributed. Hence, the simple likelihood becomes:

$$\mathcal{L}(\mu, \theta) = \frac{n_{exp}^{N}e^{-n_{exp}}}{N!}e^{\frac{-\theta^2}{2}} \tag{7.9}$$

where, n_{exp} is modified from Eqn. 7.8 as,

$$n_{exp} = \mu\sigma_{sig}\epsilon_{sig}A_{sig}L^{int}1.025^{\theta} + \sigma_{bkg}\epsilon_{bkg}A_{bkg}L^{int}1.025^{\theta} \tag{7.10}$$

Equations 7.9 and 7.10 can be expanded accordingly for multiple bins with same nuisance parameter as product of poisson probabilities and/or for multiple nuisance parameters.

7.4.4 Profiled likelihood

In order to maximise the likelihood, we "profile" over the nuisance parameters from the Eqn. 7.9 so that they are removed from the equation. Profiling the nuisance parameters refers to finding the values of these parameters which maximise the likelihood for each value of the POI. The profiled likelihood, $\mathcal{L}(\mu) = \mathcal{L}(\mu, \hat{\theta}(\mu)) \equiv \max_{\theta}\mathcal{L}(\mu, \theta)$.

To avoid dealing with large or small values of the profiled likelihood, we take the Negative Log of the Likelihood (NLL), and minimise that instead and assume that the minimum value of the curve is at $\hat{\mu}$. Since the value of the likelihood curve at the minimum is not relevant, we subtract the value at the minimum to obtain, for each value of μ the $\Delta\mathrm{NLL}$ as:

$$\begin{aligned} -\Delta ln\mathcal{L} &= -ln\mathcal{L}(\mu, \hat{\theta}(\mu)) - (-ln\mathcal{L}(\hat{\mu}, \hat{\theta})) \\ -\Delta ln\mathcal{L} &= -ln\frac{\mathcal{L}(\mu, \hat{\theta}(\mu))}{\mathcal{L}(\hat{\mu}, \hat{\theta})} \end{aligned} \tag{7.11}$$

Note that -2$\Delta ln\mathcal{L}$ is known as profile likelihood ratio. It is used as a test statistic for hypothesis testing (e.g. calculating a significance or setting an upper limit).

7.4.5 Hypothesis testing, p-values and significances

A hypothesis H specifies the probability for the data, i.e. the outcome of the observation. Consider a hypothesis H_0 and an alternative hypothesis H_1. A frequentist statistical test of H_0 is defined by specifying a critical region 'w' of the data space such that there is no more than some (small) probability α, assuming H_0 is correct, to observe the data there, i.e. $P(x \in w|H_0) \leq \alpha$. Here, α is called the size or significance level of the test. If x is observed in the critical region w, then hypothesis H_0 is to be rejected. But in general there are an infinite number of possible critical regions that give the same significance level α. So, the choice of the critical region for a test of H_0 needs to take into account the alternative hypothesis H_1. Hence, the critical region is chosen where there is a low probability for data to be found if H_0 is true, but high if H_1 is true.

Note that rejecting H_0 is not necessarily equivalent to the statement that we believe it is false and H_1 is true. Frequentist statistics only associates a probability with outcomes of repeatable observations (the data). The usefulness of the frequentist test lies in the fact that we can compute the probability to accept or reject a hypothesis assuming that it is true, or assuming some alternative hypothesis is true. This is unlike the case for bayesian statistics which depends a lot on the prior

probability, $\pi(H)$.

Suppose hypothesis H predicts some functional form of the PDF, $f(\vec{x}|H)$, and we observe one single point in this critical region. Then, to express the validity of hypothesis H in light of the data, a p-value is defined as the probability, under assumption of H, to observe data with equal or lesser compatibility with H relative to the data. This is not the probability that H is true, but tells what part of data space constitutes lesser compatibility with H than the observed data (implicitly this means that region gives better agreement with some alternative).

The p-value is a function of the data, and is thus itself a random variable with a given distribution. The p-value of hypothesis H can be found from a test statistic t(x) as,

$$p_H = \int_t^\infty f(t'|H)dt' \tag{7.12}$$

In general, for continuous data, under the assumption of H, $p_H \sim$ Uniform[0,1]. Suppose we observe n events, which can consist of n_b events from known SM processes (background) and n_s events from a BSM process (signal). If n_s and n_b are poisson random variables with mean s and b, respectively, then n = n_s + n_b is also a Poisson with mean = s+b. This can be given as,

$$P(n; s, b) = \frac{(s+b)^n}{n!}e^{-(s+b)} \tag{7.13}$$

Suppose b=0.5 and we observe n_{obs}=5, then p-value for hypothesis s=0 can be calculated as, p-value = P(n$\geq$5; b=0.5, s=0) = 1.7×10^{-4}. The primary role of the p-value is to quantify the probability that the background-only model gives a statistical fluctuation as big as the one seen or bigger. It is not intended as a means to protect against hidden systematics or the high standard required for a claim of an important discovery. In the processes of establishing a discovery there comes a point where it is clear that the observation is not simply a fluctuation, but an "effect", and the focus shifts to whether this is new physics or a systematic.

The significance, Z is defined as the number of standard deviations that a Gaussian variable would fluctuate in one direction to give the same p-value. It can be calculated as,

$$\begin{aligned} p &= \int_Z^\infty \frac{1}{\sqrt{2\pi}} e^{1\frac{x^2}{2}} dx \\ p &= 1 - \Phi(Z) \\ Z &= \Phi^{-1}(1-p) \end{aligned} \tag{7.14}$$

Conventionally, a discovery is claimed if the p-value of the no-signal hypothesis is below

2.9×10^{-7}, corresponding to a significance Z=5 (a 5σ effect).

7.4.6 Obtaining a confidence interval

In addition to a 'single-value' estimate of a POI, we should also report an interval reflecting the statistical uncertainty on the POI. Such an interval should be able to communicate objectively the result of the experiment and have a given probability of containing the true parameter.

Frequentist confidence intervals for a parameter θ can be found by defining a test of the hypothesized value θ (for all θ), by specifying values of the data that are 'disfavored' by the parameter (critical region) such that P(data in critical region) $\leq \alpha$, for a prespecified value of α, e.g. 0.05 or 0.1. If data is observed in the critical region, then reject the value θ. Now, inverting the test to define a confidence interval as: set of θ values that would not be rejected in a test of size α. Confidence level (CL) is given as 1 - α, thus α = 0.05 corresponds to 95% CL.

To establish a relationship between confidence interval and p-value, we can consider a significance test for each hypothesized value of parameter θ, resulting in a p-value, p_θ. If $p_\theta < \alpha$, then θ is rejected. The confidence interval at CL = 1 - α consists of those values of θ that are not rejected. Thus, an upper limit on θ is the greatest value for which $p_\theta \geq \alpha$. In practice, we find the upper limits by setting $p_\theta = \alpha$ and solving for θ.

According to Wilks' theorem, in the limit of large sample sizes, the profile likelihood ratio is distributed as a χ^2 with N degrees of freedom, where N is the difference in number of free parameters between the numerator and denominator of the likelihood ratio (only 1 in the example in Sec 7.4.3). Then, using the quantile function of the χ^2 distribution, it can be seen that for a 68% confidence interval, $-2\Delta\mathrm{NLL} < 1 \rightarrow -\Delta\mathrm{NLL} < 0.5$. So, extracting a 68% confidence interval from the region for which $-\Delta\mathcal{L}(\mu) < 0.5$. In the frequentist paradigm, this means that if the interval covers, 68% of intervals constructed via this method should contain the true value of the POI.

7.4.7 Analysis-specific procedure

To calculate the upper limits on the production cross section for the three BSM models considered, a modified frequentist approach with the CLs [162–165] criterion is used, with a test statistic based on the binned profile likelihood, in the asymptotic approximation. The upper limits are calculated at 95% C.L. The systematic uncertainties and their correlations are incorporated in the likelihood as nuisance parameters with log-normal probability density functions. The statistical uncertainties

in the signal and background estimates are modeled with gamma functions. These are described in detail below.

7.4.7.1 Automatic MC statistical uncertainties

The nuisance parameters arising from most of the systematic uncertainties are independent and thus multiplicative, except for the statistical uncertainties in each bin with multiple background processes. Hence, in a signal region with binned spectra such as in this multilepton analysis, a Barlow-Beeston light approach [166, 167] is used to assign a single nuisance parameter to scale the sum of the process yields in each bin, constrained by the total uncertainty, instead of requiring separate parameters, one per process. This is useful as it minimises the number of parameters required in the maximum-likelihood fit.

One significant advantage of the Barlow-Beeston light approach is that the maximum likelihood estimate of each nuisance parameter has a simple analytic form that depends only on total expected background, combined uncertainty from all the background processes, and the observed number of data events in the relevant bin. Therefore when minimising the negative log-likelihood of the whole model it is possible to remove these parameters from the fit and set them to their best-fit values automatically. For models with large numbers of bins this can reduce the fit time and increase the fit stability.

7.4.7.2 Correlation model and impact on nuisances

The misidentified lepton background has the most dominant contribution to systematic uncertainties. Hence, its important to be taken care of appropriately while performing the multidimensional fit across the bins of the BDT spectra. Moreover, the composition of fake leptons changes between different kinematic regions (e.g. 4L vs 3L, BelowZ vs AboveZ, low vs high p_{T}). Hence, the constraints on the misID nuisances are needed to be considered separately, to allow the parameters float independently of each other, especially not be affected from regions of high sideband statistics (low BDT score) to low statistics (high BDT score).

To do this, uncorrelated nuisances between different BDT regions i.e. low, medium, and high, are created for misID per channel (3-channel and 4-channel) per year (2016/2017/2018). This allows the profiling to change the yields and constrain the nuisances independently based on the underlying events from that part of the BDT spectrum. The structure of the misID nuisances broken down after this correlation model in 3-lepton channel in an example BDT region plot is

shown in Figure 7.17. Separate panels for the BDT region distributions in the years 2016, 2017, and 2018 are shown in the same figure. The characteristic background and signal shapes, after the BDT training, are clearly visible where background is peaking at the low BDT score and signal is populating the high BDT score end. Each bin is used for performing counting experiments, which are then statistically combined in the end, taking into account all the proper correlations to arrive at the final results.

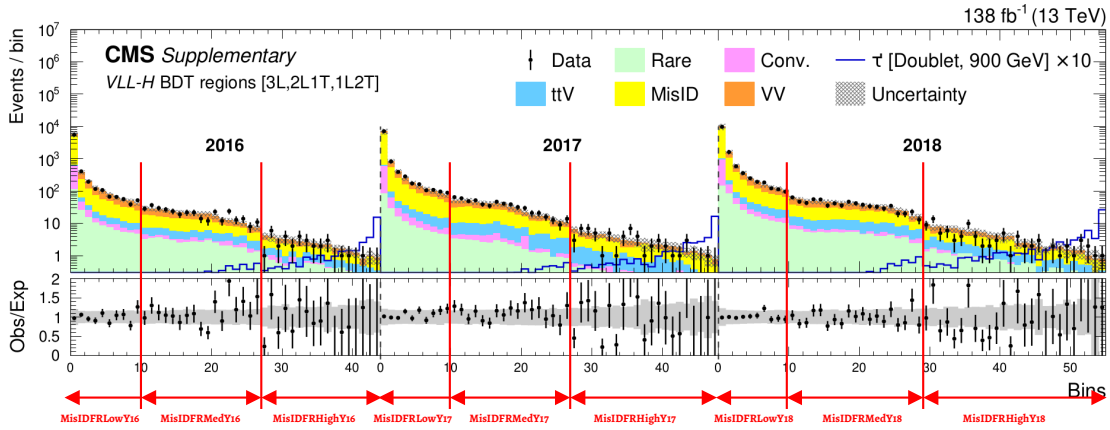

Figure 7.17: Correlation model of the misidentified lepton background nuisances in a BDT region distribution of 3-lepton channel.

To assess the effect on the prefit versus postfit nuisance parameters, we have produced "impact" plots per BDT training which, for each of those parameters, shows the shift in value and the post-fit uncertainty, both normalized to the input values, and the linear correlation between the parameter and the signal strength (r-value). The r-value is defined as the ratio of upper limit on the production cross section of the BSM signal to its theory cross section. An r-value > 1 implies that the theory cross section is much larger than the constraints from the experiments, which is then comfortably ruled out since it has not been observed. An example impact plot for the corresponding BDT training is shown in Figure 7.18.

In the low BDT region, signal is non-existent and the post-fit agreement between data and total predicted background (dominantly misID) is exceptionally well, almost in all bins. The pre-fit MisID uncertainties are typically 20% for low pT and can go upto 50% for high pT leptons.

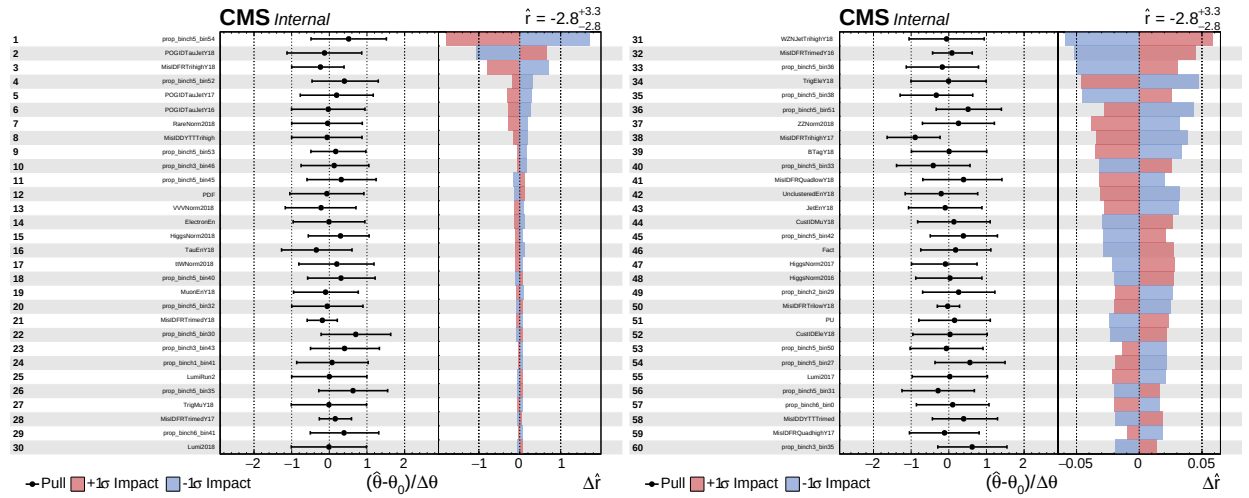

Figure 7.18: Impact, pulls and constraints of the leading systematic uncertainties for the signal plus background hypothesis with observed data for the VLL-H BDT. The fit is done for the VLL doublet signal with a mass of 900 GeV and a signal strength multiplier of 10%.

Therefore, the post-fit nuisances look severely-constrained as they shrink down to $1/\sqrt{N}$, where N is the MisID yield in the bins. In the medium BDT region, MisID statistics are still reasonably high with almost vanishingly small signal. Hence, the agreement between data and prediction after the fit is relatively good. As a result, the nuisances shrink again, but not by that much as compared to those in the low BDT region. Finally, in the high BDT region where total background is down to one-event level and signal presence is high, the nuisances do not get over constrained.

7.4.7.3 Asymptotic Frequentist Limits

The Asymptotic Limits method allows to compute quickly an estimate of the observed and expected limits, which is fairly accurate when the event yields are not too small and the systematic uncertainties don't play a major role in the counting experiments with highest sensitivity to BSM phenomena. The latter part is true since the most sensitive bins are of the order of one event yield, where statistical uncertainties dominate the measurement.

The limit calculation relies on an asymptotic approximation of the distributions of the LHC test-statistic, which is based on a profile likelihood ratio, under signal and background hypotheses to compute two p-values p_μ, p_b and therefore CLs = $\frac{p_\mu}{(1-p_b)}$, i.e. it is the asymptotic approximation of computing limits with frequentist toys. We use the CLs (which itself is not a p-value) criterion often in High energy physics as it is designed to avoid excluding a signal model when the sensitivity

is low (and protects against excluding due to underfluctuations in the data).

7.5 Search results using BDTs

7.5.1 Vector-like tau lepton

For the VLL model, three mass ranges are considered. The VLL doublet and singlet models are considered together while training, since the specific events for a given mass from doublet or singlet have similar kinematic distributions. Thus a total of 9 BDTs (3 bundles $\times$ 3 years) are trained to be used for the VLL doublet and singlet models. Table 7.3 summarizes the 9 BDTs.

Table 7.3: VLL signal mass points as used in the trainings of BDTs and as used in the evaluation in the signal regions according to the best expected limit. Separate BDTs are used for 2016, 2017, and 2018 to give a total of 9 trainings. Doublet and singlet signal mass points are used together in the trainings.

BDT	Trained masses (GeV) [Doublet+Singlet]	Applied masses (GeV) [Doublet]	Applied masses (GeV) [Singlet]
VLL-H	650, 700, 800	450 and higher	300 and higher
VLL-M	300, 500	250, 300, 350, 400	200, 250, 300
VLL-L	100, 150, 200	100, 150, 200	100, 125, 150

Figure 7.19 shows the calculated expected limits including the complete set of uncertainties for the Doublet VLL model in the combined 2016–2018 data set, evaluated by using a specific BDT for the whole mass range. For a particular signal mass hypothesis, the appropriate BDT training that yields the best upper cross-section limits is chosen. This informs the choices outlined in Table 7.3.

The BDT region distributions in the 3-object and 4-object channels for the *VLL-L*, *VLL-M*, and *VLL-H* BDT training in the three years of data-taking are shown in Figures 7.20-7.21, respectively.

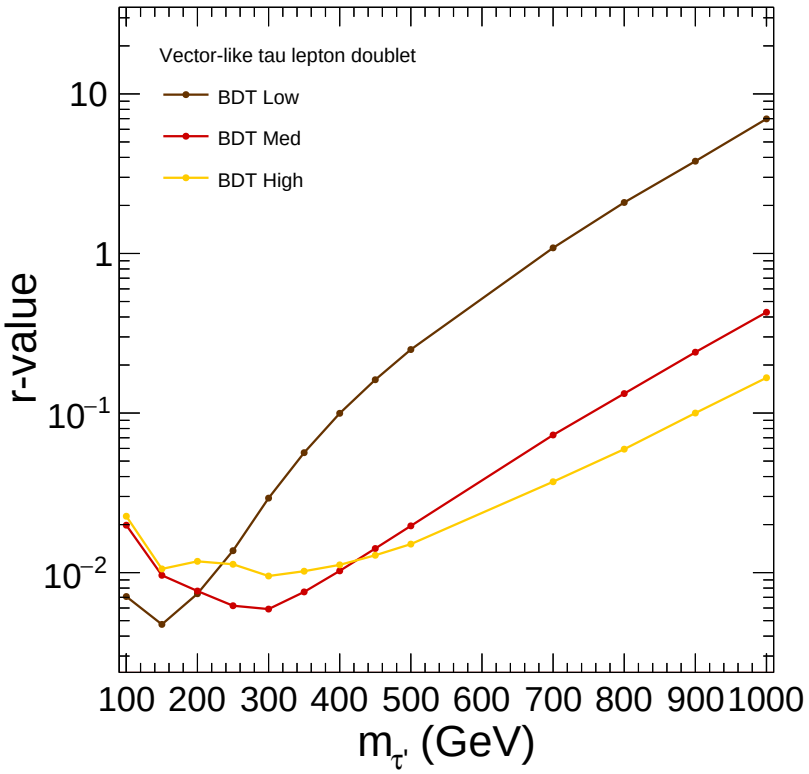

Figure 7.19: The expected limits including the complete set of uncertainties for the Doublet VLL model using a given BDT for the whole mass range. This test informs the choice of BDT for a particular mass point.

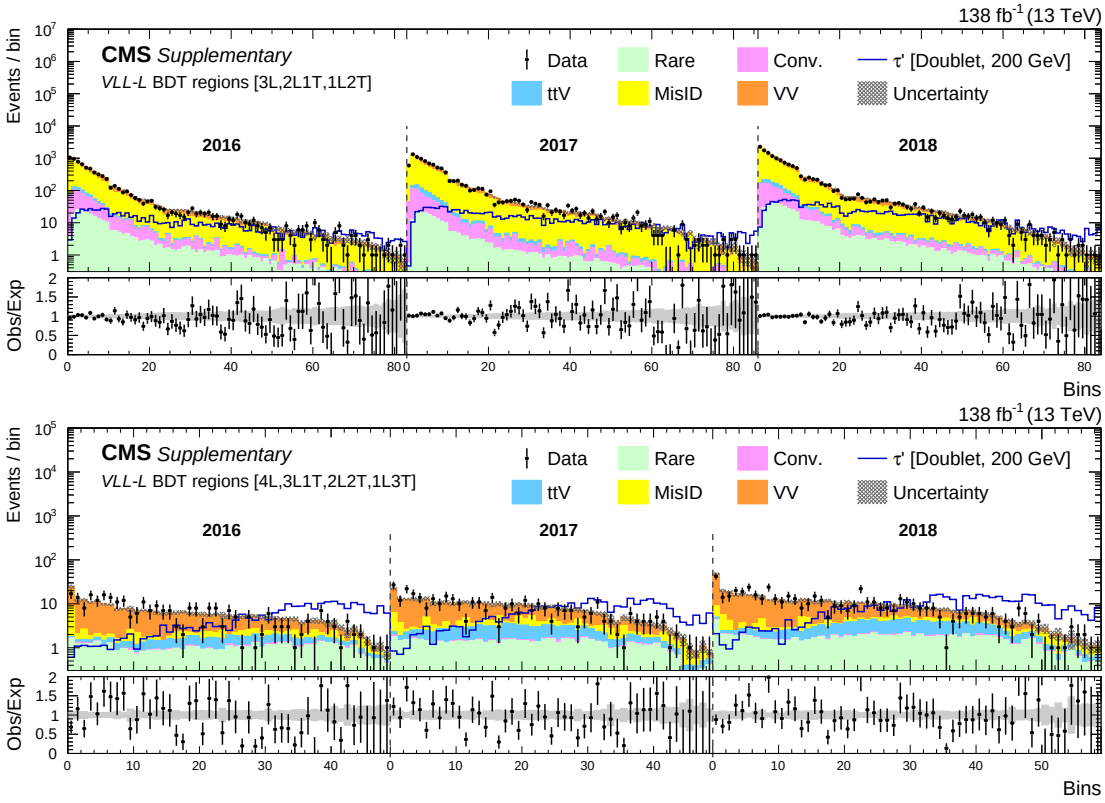

Figure 7.20: *VLL-L* (upper) BDT regions in the 3-object channels (upper) and in the 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the vector-like τ lepton in the doublet scenario, before the fit, is also overlaid.

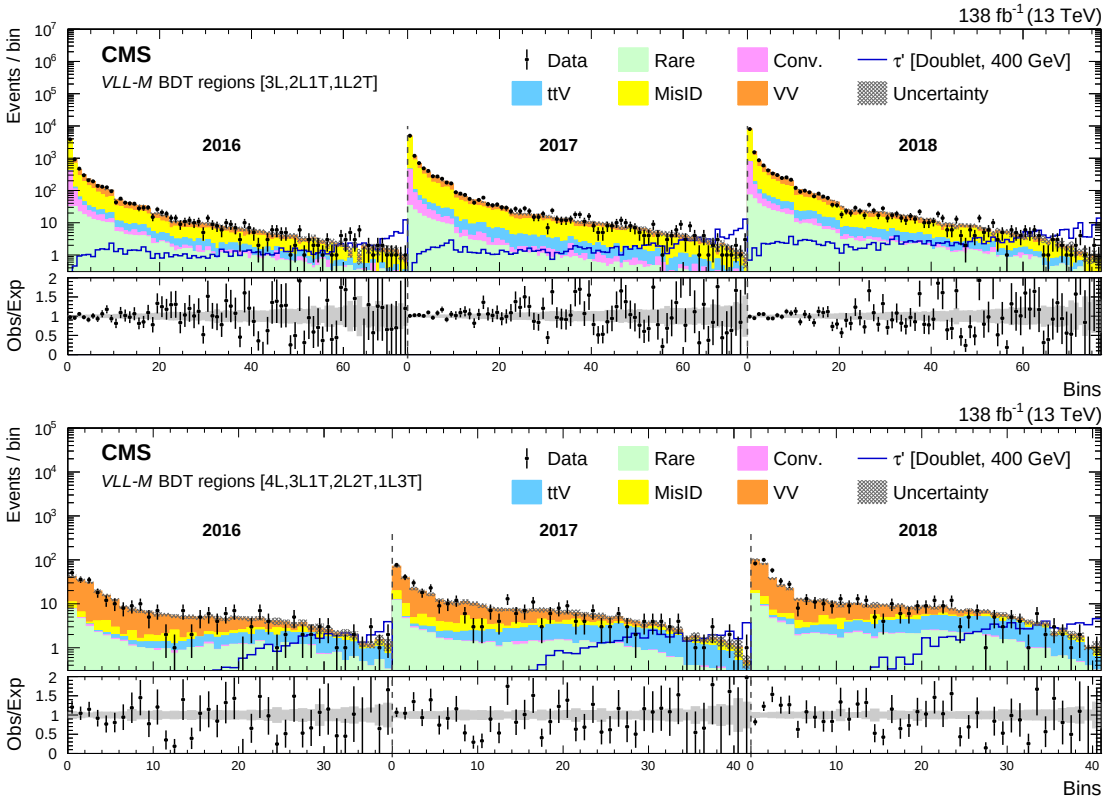

Figure 7.21: *VLL-M* BDT regions in the 3-object channels (upper) and in the 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the vector-like τ lepton in the doublet scenario, before the fit, is also overlaid.

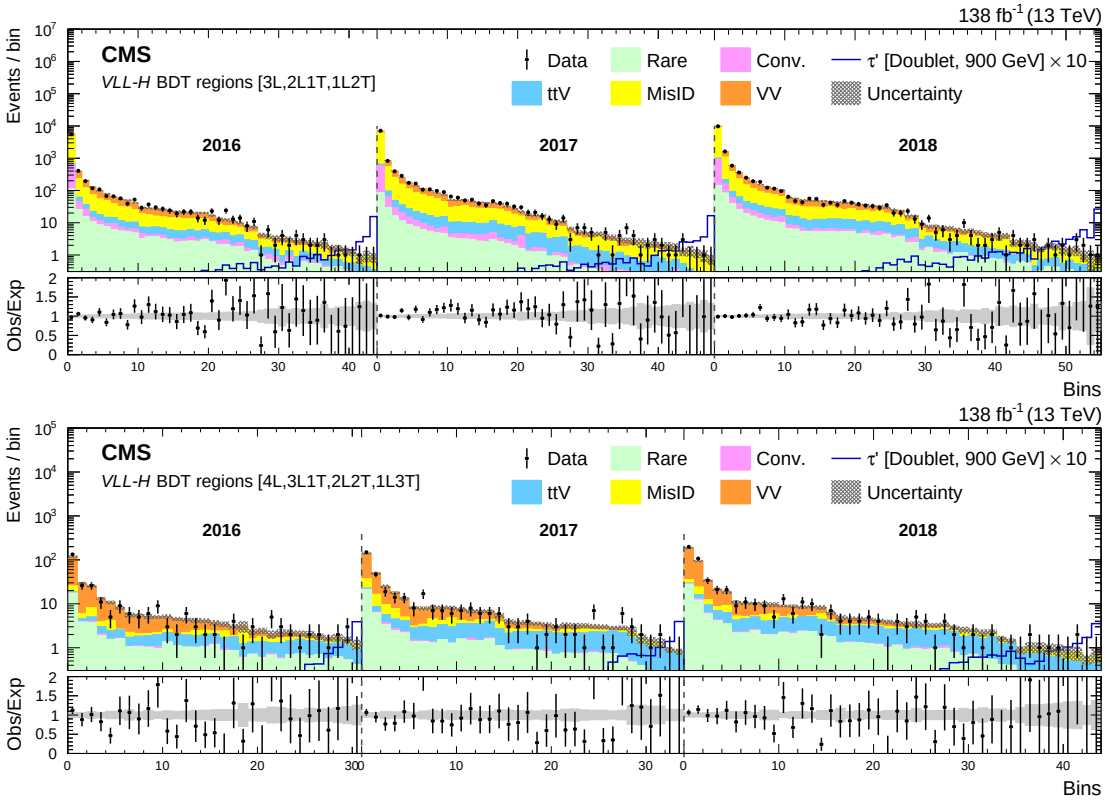

Figure 7.22: *VLL-H* BDT regions in the 3-object channels (upper) and in the 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the vector-like τ lepton in the doublet scenario, before the fit, is also overlaid.

No significant deviations in data wrt the expected SM background prediction was observed in any bin of the VLL BDT regions from both 3-object and 4-object channels in the three years of the data-taking. To summarize the per-bin agreement between the observed data and the expected SM backgrounds, pull distributions for the several BDT regions were made. The pull is defined as the ratio of the difference between the number of events observed in the data and the predicted background, over the quadratic sum of the systematic uncertainty in the background and the statistical uncertainty in the data. Since a significant number of bins have low background yields, the pull distribution is not expected to fully follow a Gaussian distribution, but it is a quick check performed for outliers. Figure 7.23 shows the pull distributions for *VLL-L* (upper left), *VLL-M* (upper-right), and *VLL-H* (lower) BDT regions for the combined 2016–2018 data set in the background-only hypothesis. These plots include all the bins from both 3-object and 4-object channels. As can be seen from the figures, all the significances are comfortably within ± 3 sigma deviation.

Consequently, the observed and expected upper limits at 95% CL were calculated on the production cross section of the doublet vector-like lepton model and are shown in Figure 7.24. The vector-like τ leptons in the doublet scenario are excluded up to a mass $m_{\tau'}$ of 1045 GeV, where the expected mass exclusion is 975 GeV.

Separate studies, as outlined in Section 8.3, have shown that due to the very small cross section of the vector-like τ lepton model in the singlet scenario, and also low acceptance in the multilepton channels due to high kinematic thresholds on the physics objects, the BDT regions were not as sensitive for the singlet model as for the doublet model. Hence, they are not used to compute the BDT-based constraints, and an alternative approach is taken as described in Section 8.2.

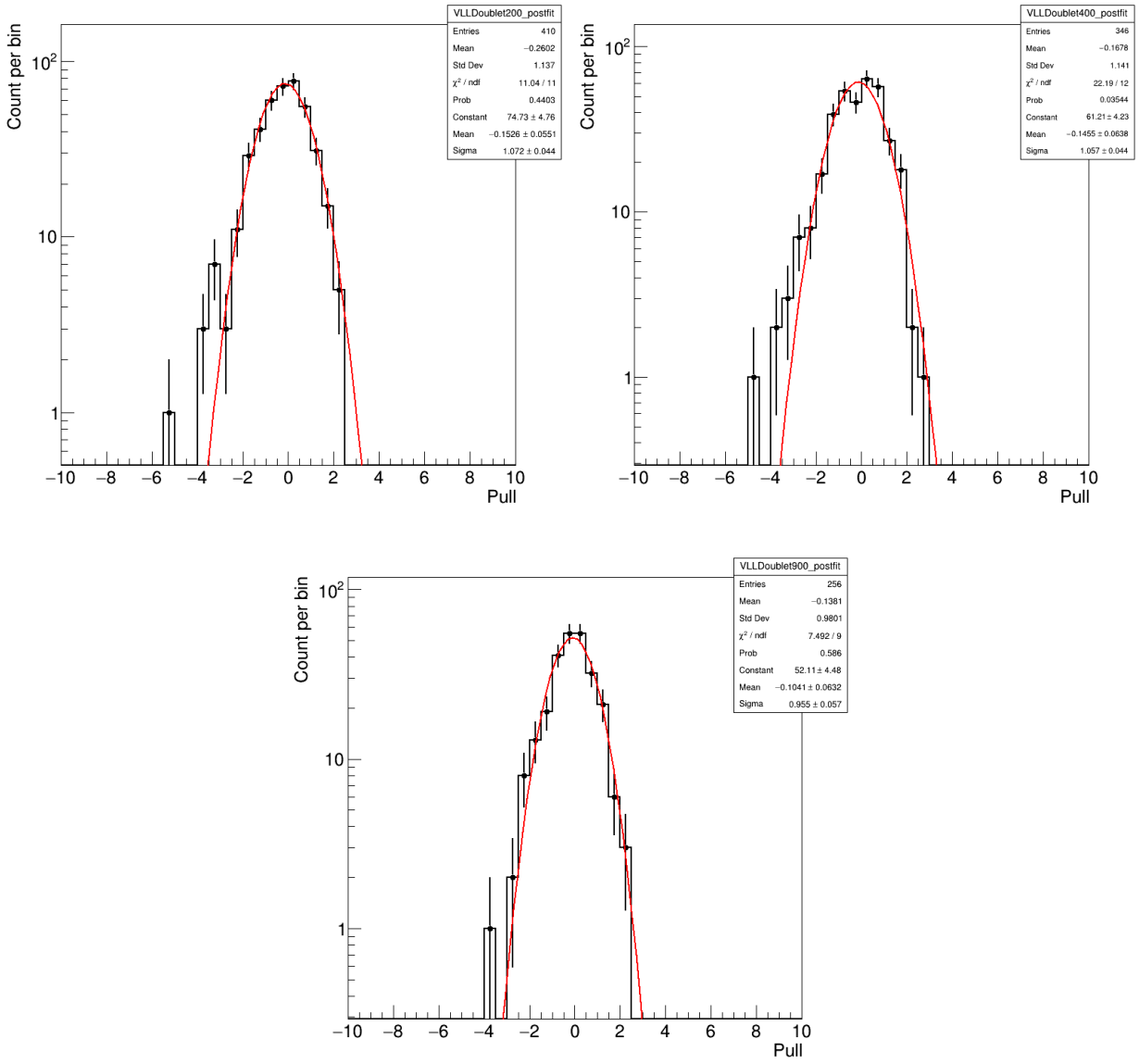

Figure 7.23: Histogram of pulls for the *VLL-L* (upper left), *VLL-M* (upper right), and *VLL-H* (lower) BDT regions for the combined 2016–2018 data set in the background-only hypothesis. These plots include all the bins from both 3-object and 4-object channels.

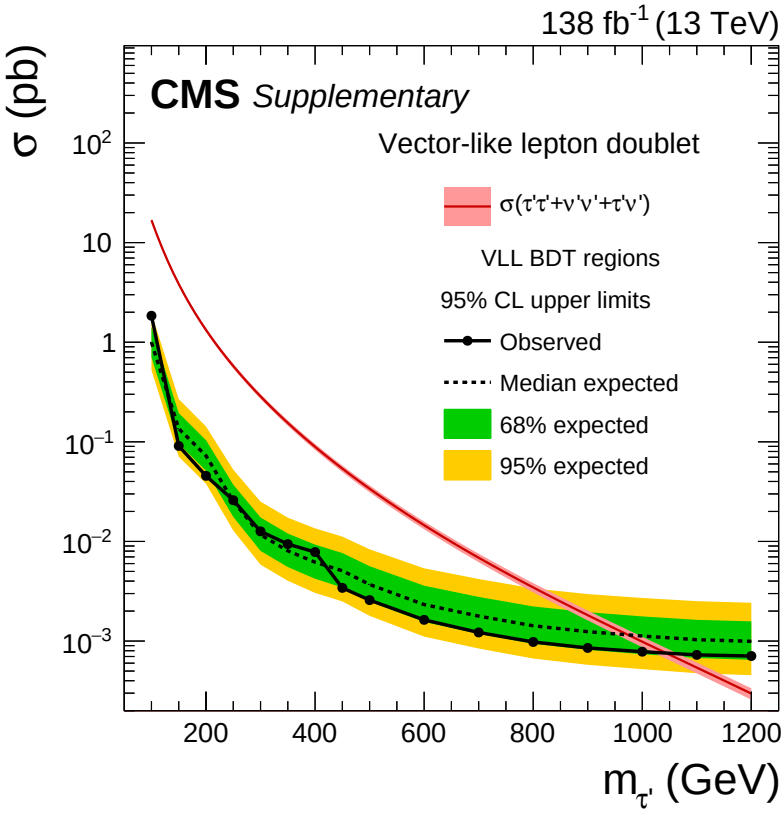

Figure 7.24: Observed and expected upper limits at 95% CL on the production cross section for the vector-like tau leptons in the doublet model using the VLL BDT regions.

7.5.2 Type-III seesaw fermions

For the type-III seesaw model, the training set up is more complex since the mixing of the seesaw fermions with the SM lepton flavor, and thus the Σ decay branching ratios to SM lepton flavors, is a free parameter. We consider four BDT trainings: (i) flavor-democratic scenario i.e. $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$, (ii) pure light lepton scenario ($ee, \mu\mu, e\mu$) i.e. $\mathrm{B_e} + \mathrm{B}_\mu = 1$, (iii) pure tau scenario ($\tau\tau$) i.e. $\mathrm{B}_\tau = 1$, and (iv) mixed scenario ($e\tau, \mu\tau$). We consider four mass ranges in each flavor-mixing scenario, giving a total of 48 BDTs (4 mass-ranges $\times$ 4 flavor scenarios $\times$ 3 years) trained for the seesaw model. Table 7.4 summarizes the details of the mass points used in training and evaluation which is same across all flavors.

Table 7.4: Seesaw signal mass points as used in the trainings of BDTs and as used in the evaluation in the signal regions according to the best expected limit. Separate BDTs are used for 2016, 2017, and 2018 for each of the four flavor-mixing scenario to give a total of 48 trainings.

BDT	Trained masses (GeV)	Applied masses (GeV)
SS-H	1000, 1250	700 and higher
SS-M	400, 550, 700, 850	400, 550
SS-L	200, 300	200, 300
SS-VL	100	100

Figure 7.25 shows the calculated expected limits including the complete set of uncertainties evaluated by using a single BDT for the whole mass range. This is done using the $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ BDT training but the same conclusions can be drawn for other flavor BDTs. This informs the choices outlined in Table 7.4.

The BDT regions are independently defined for each channel, for each of the four mass range BDTs, and for each flavor mixing scenario. For a particular signal mass and mixing hypothesis, the BDT which yields the best expected limit is chosen. After testing the performance of various BDTs on different seesaw mixing hypothesis, we find that the $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ BDT training performs the best for the flavor-democratic, pure light lepton ($ee, \mu\mu, e\mu$), and mixed ($e\tau, \mu\tau$) scenarios; whereas $\mathrm{B}_\tau = 1$ BDT training performs the best for the seesaw fermions with 100% mixing to taus.

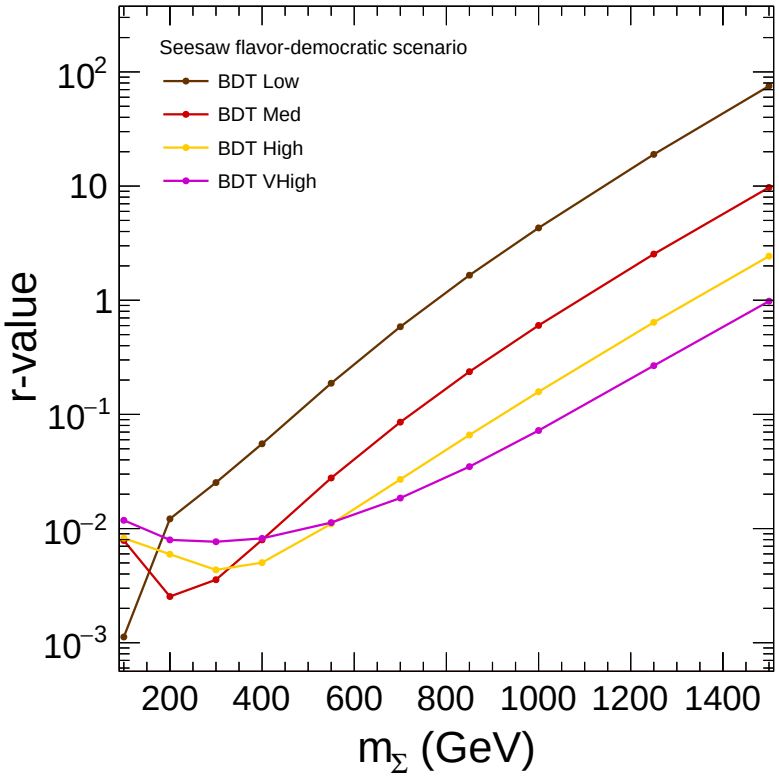

Figure 7.25: The expected limits including the complete set of uncertainties for the seesaw model in the $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ scenario using a single BDT used for the whole mass range. This test informs the choice of BDT for a particular mass point.

The BDT region distributions in the 3-object and 4-object channels for the *SS-VL*, *SS-L*, *SS-M*, and *SS-H* $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ BDT training in the three years of data-taking are shown in Figure 7.26-7.29, respectively.

The BDT region distributions in the 3-object and 4-object channels for the *SS-VL*, *SS-L*, *SS-M*, and *SS-H* $\mathrm{B}_\tau = 1$ BDT training in the three years of data-taking are shown in Figure 7.30-7.33, respectively.

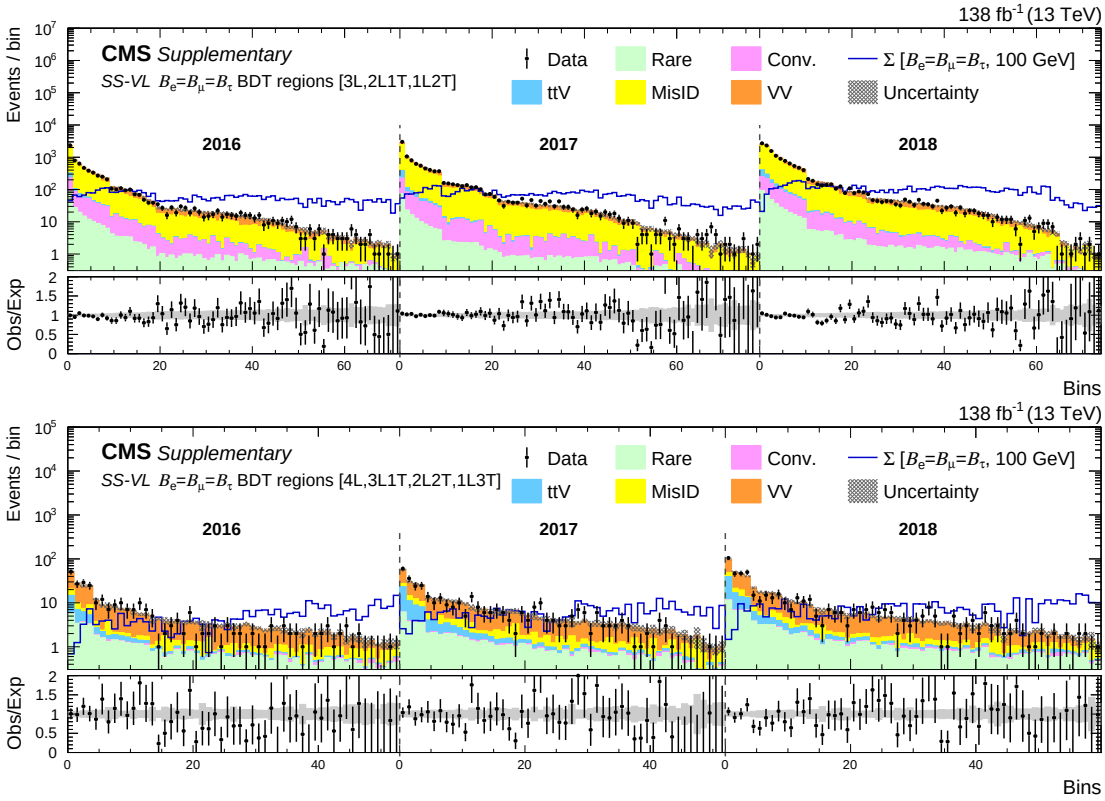

Figure 7.26: *SS-VL* BDT regions for the $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ training, in the 3-object channels (upper) and 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the type-III seesaw heavy fermions in the flavor-democratic scenario, before the fit, is also overlaid.

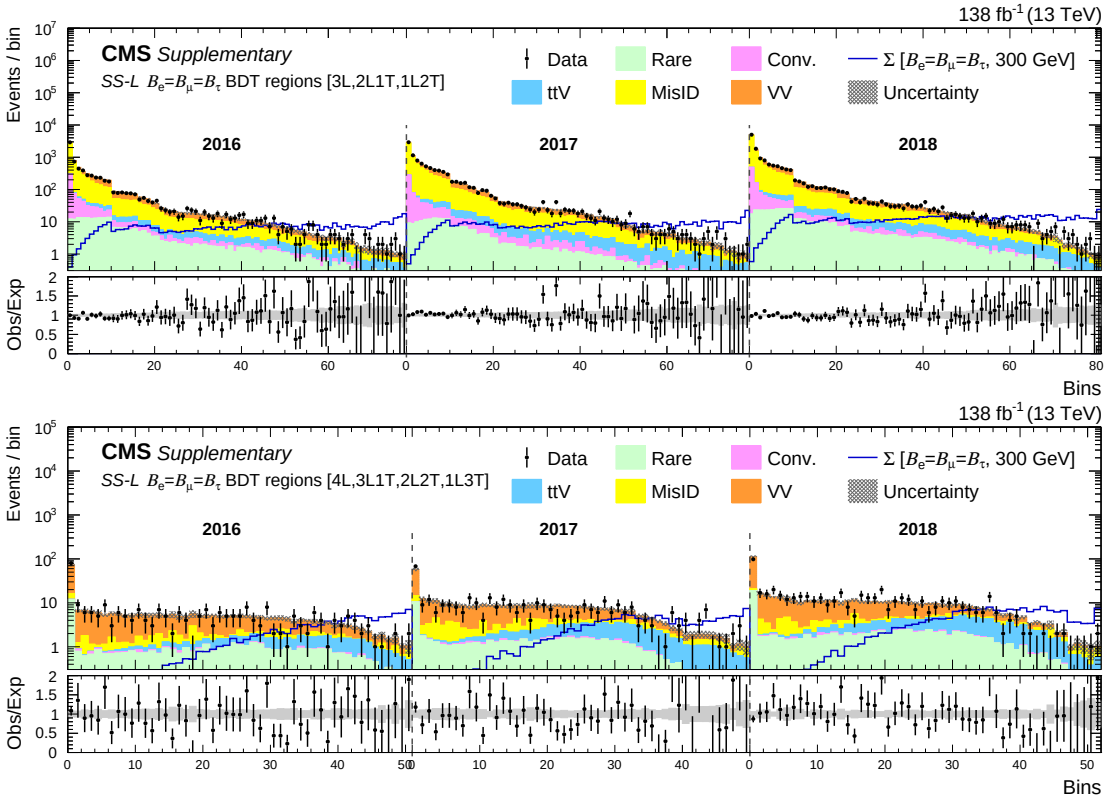

Figure 7.27: *SS-L* BDT regions for the $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ training, in the 3-object channels (upper) and 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the type-III seesaw heavy fermions in the flavor-democratic scenario, before the fit, is also overlaid.

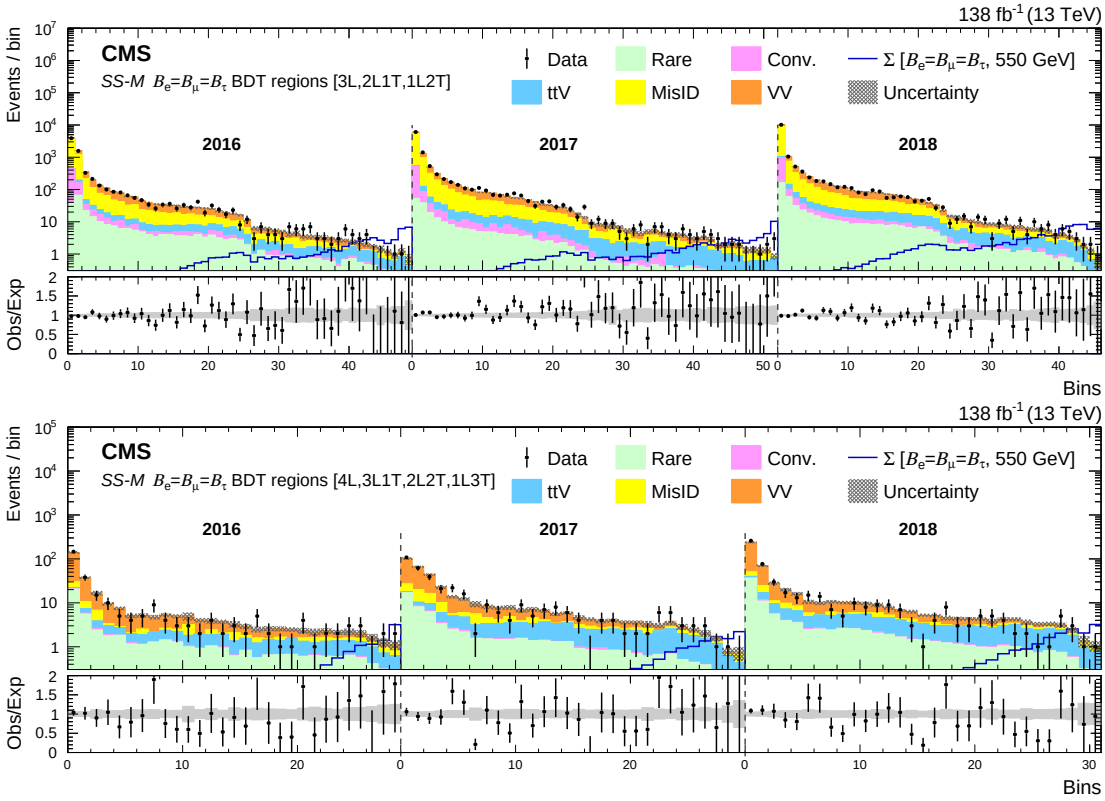

Figure 7.28: *SS-M* BDT regions for the $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ training, in the 3-object channels (upper) and 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the type-III seesaw heavy fermions in the flavor-democratic scenario, before the fit, is also overlaid.

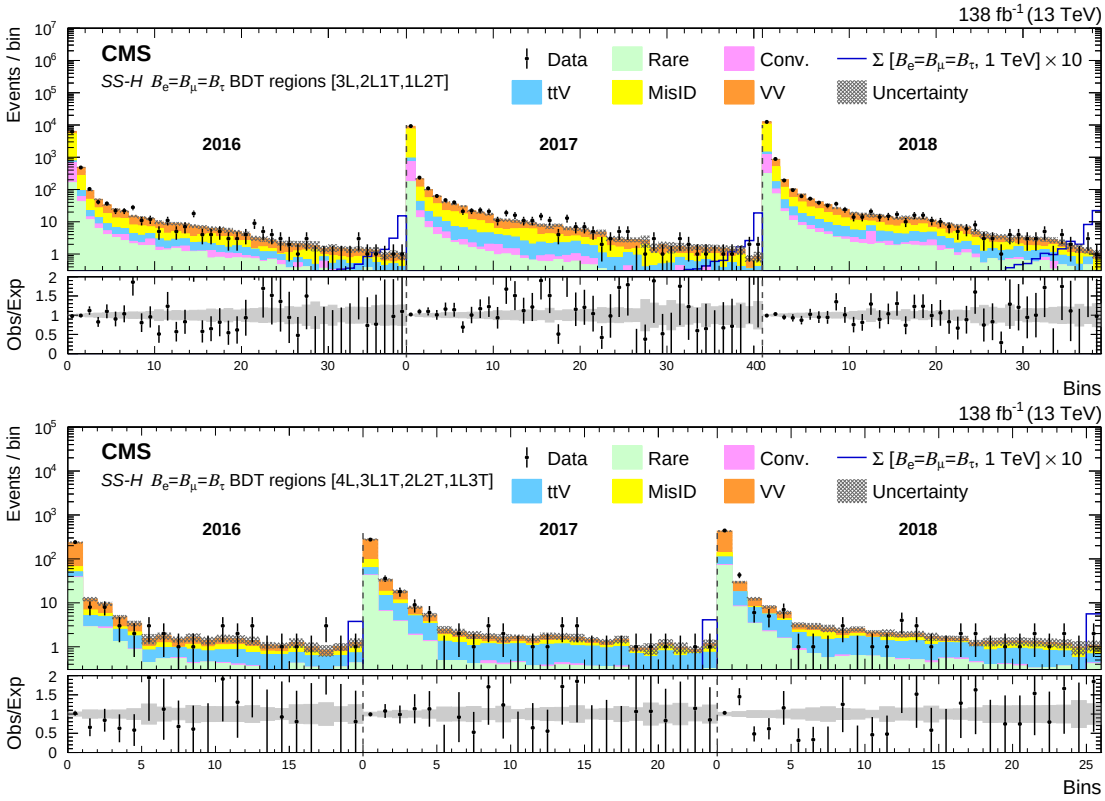

Figure 7.29: *SS-H* BDT regions for the $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ training, in the 3-object channels (upper) and 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the type-III seesaw heavy fermions in the flavor-democratic scenario, before the fit, is also overlaid.

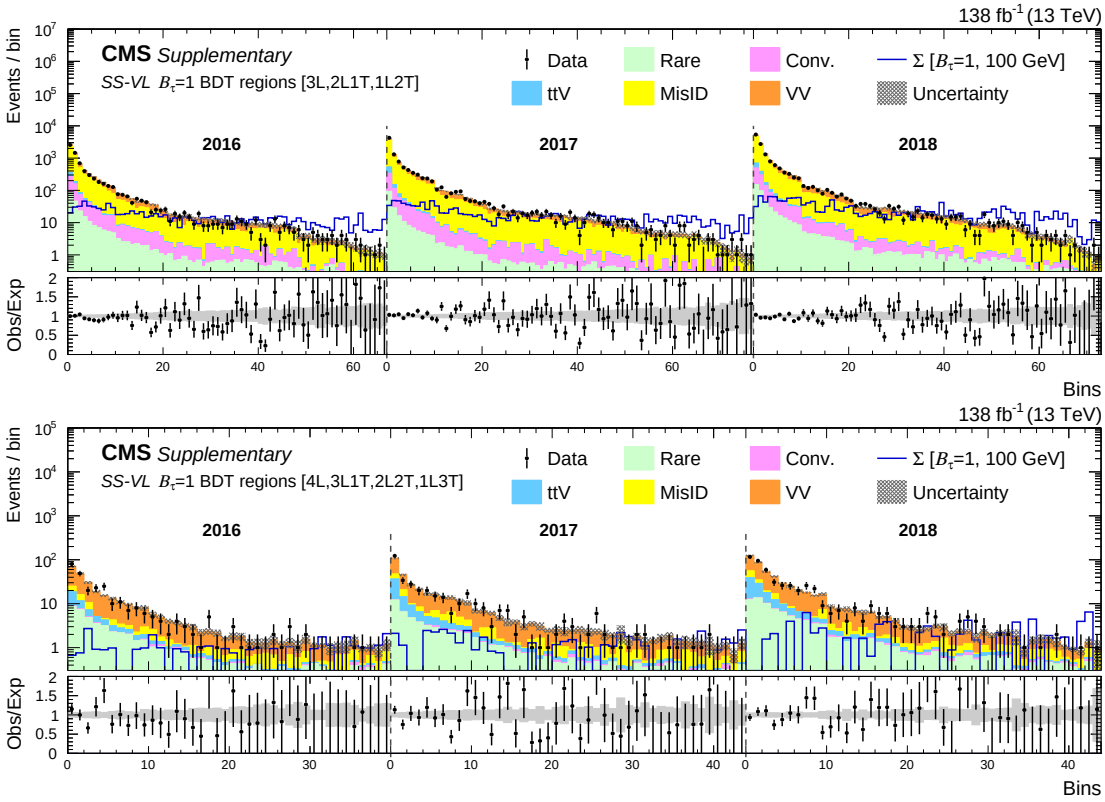

Figure 7.30: *SS-VL* BDT regions for the $\mathrm{B}_\tau = 1$ training, in the 3-object channels (upper) and 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the type-III seesaw heavy fermions in the scenario with mixing exclusively to τ lepton, before the fit, is also overlaid.

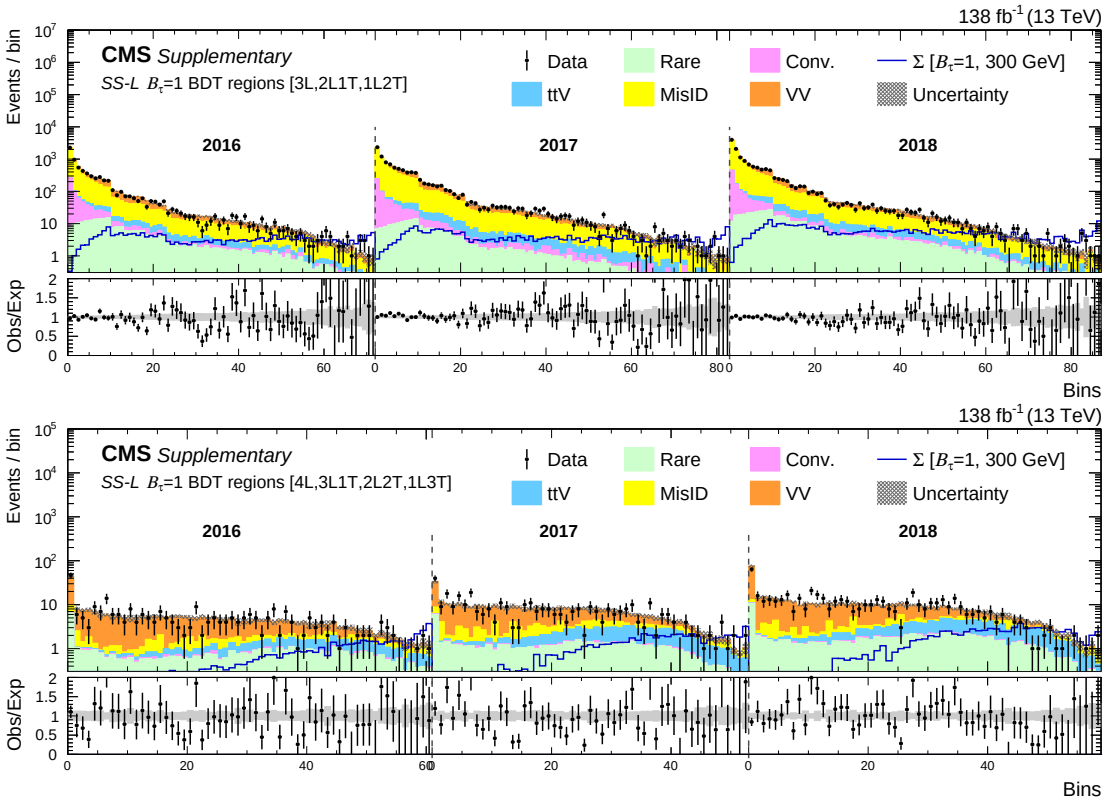

Figure 7.31: *SS-L* BDT regions for the $\mathrm{B}_\tau = 1$ training, in the 3-object channels (upper) and 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the type-III seesaw heavy fermions in the scenario with mixing exclusively to τ lepton, before the fit, is also overlaid.

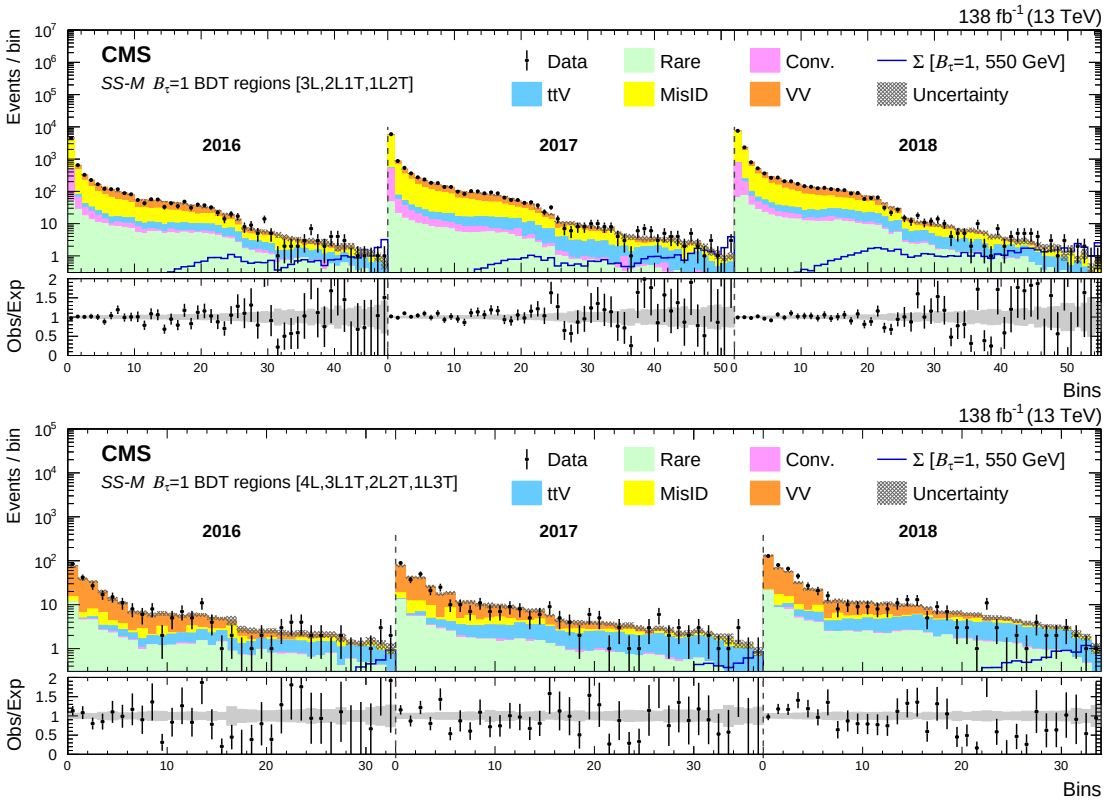

Figure 7.32: *SS-M* BDT regions for the $\mathrm{B}_\tau = 1$ training, in the 3-object channels (upper) and 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the type-III seesaw heavy fermions in the scenario with mixing exclusively to τ lepton, before the fit, is also overlaid.

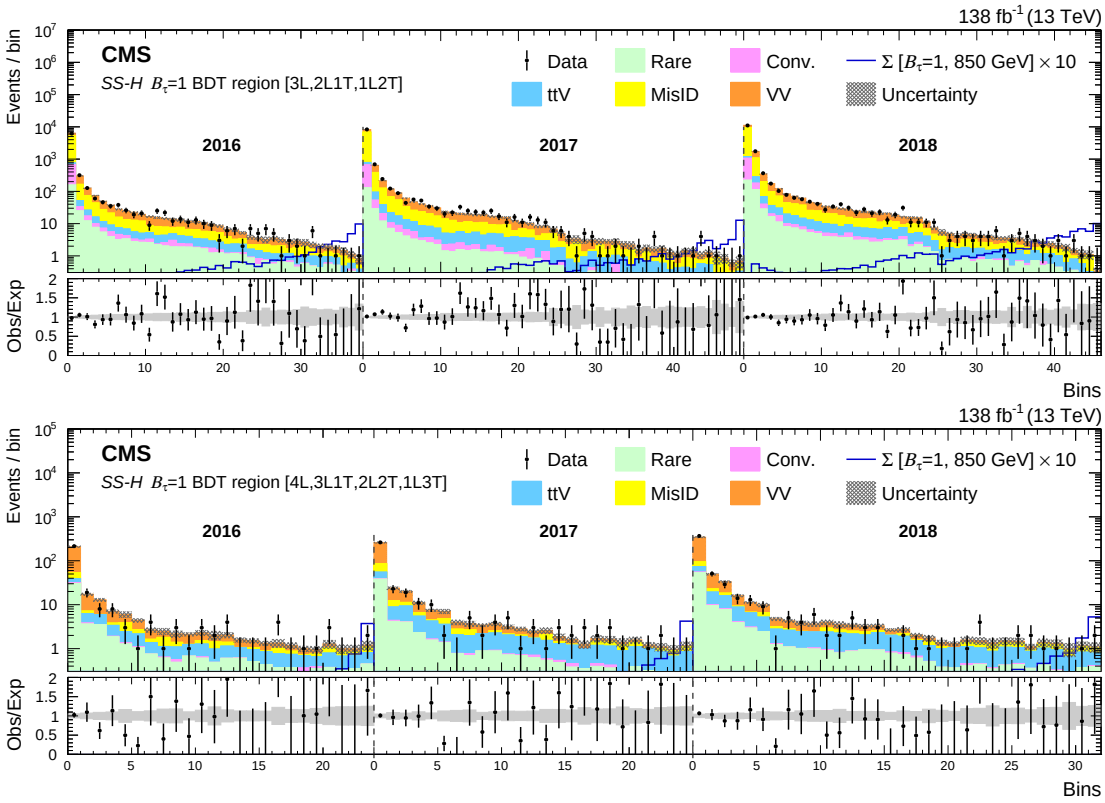

Figure 7.33: *SS-H* BDT regions for the $\mathrm{B}_\tau = 1$ training, in the 3-object channels (upper) and 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the type-III seesaw heavy fermions in the scenario with mixing exclusively to τ lepton, before the fit, is also overlaid.

No significant deviations in data wrt the expected SM background prediction was observed in any bin of the SS $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ and $\mathrm{B}_\tau = 1$ BDT regions from both 3-object and 4-object channels in the three years of the data-taking. Figure 7.34 shows example pull distributions for *SS-H* BDT regions from the $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ (left) and $\mathrm{B}_\tau = 1$ (right) trainings, for the combined 2016–2018 data set in the background-only hypothesis. These plots include all the bins from both 3-object and 4-object channels. As can be seen from the figures, all the significances are comfortably within ± 3 sigma deviation.

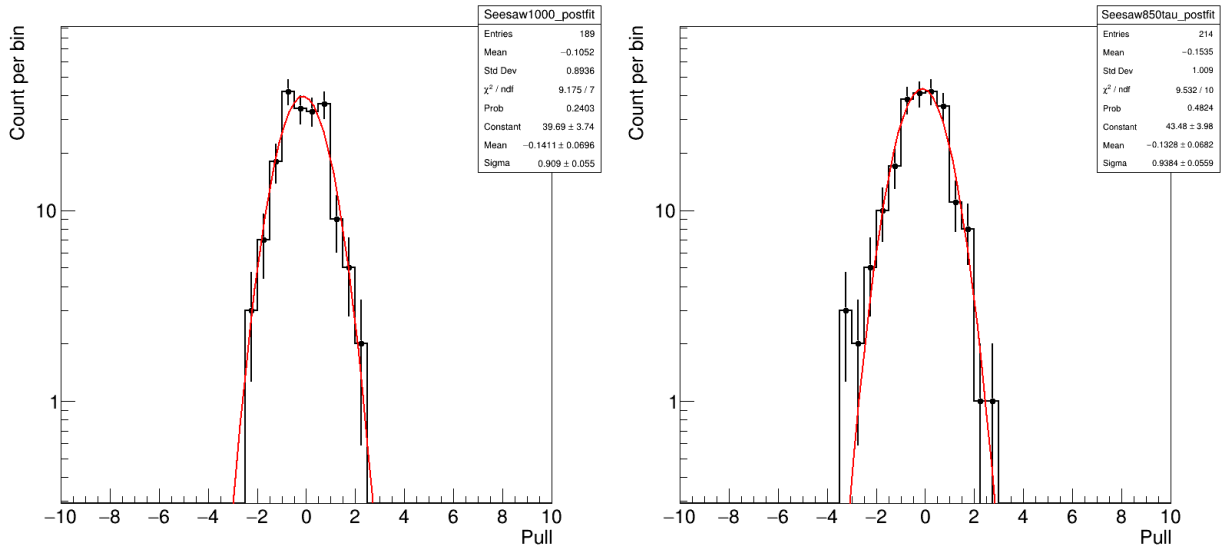

Figure 7.34: Histogram of pulls for the *SS-H* BDT regions from the $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ (left) and $\mathrm{B}_\tau = 1$ (right) trainings, for the combined 2016–2018 data set in the background-only hypothesis. These plots include all the bins from both 3-object and 4-object channels.

Consequently, the observed and expected upper limits at 95% CL were calculated on the production cross section of the type-III seesaw fermions and are shown in Figure 7.35 upper left for $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ scenario, upper right for $\mathrm{B_e} = 1$ scenario, and lower left for $\mathrm{B}_\mu = 1$ scenario using the SS $\mathcal{B}_e = \mathcal{B}_\mu = \mathcal{B}_\tau$ BDT regions, and lower right for $\mathrm{B}_\tau = 1$ scenario using the $\mathcal{B}_\tau = 1$ BDT regions.

For arbitrary Σ decay branching fractions to SM lepton flavors, subject to the constraint that $\mathcal{B}_e + \mathcal{B}_\mu + \mathcal{B}_\tau = 1$, the observed and expected lower limits on m_Σ in the plane defined by $\mathcal{B}_e$ and $\mathcal{B}_\tau$ are shown in Fig. 7.36. These limits are given by the *SS-H* $\mathcal{B}_\tau = 1$ BDT when $\mathcal{B}_\tau \geq 0.9$, and by the *SS-H* $\mathcal{B}_e = \mathcal{B}_\mu = \mathcal{B}_\tau$ BDT for the other decay branching fraction combinations. The strongest constraints are when $\mathcal{B}_\mu = 1$ ($m_\Sigma > 1065$ GeV), while the weakest are when $\mathcal{B}_\tau = 0.8$, $\mathcal{B}_e = 0.2$

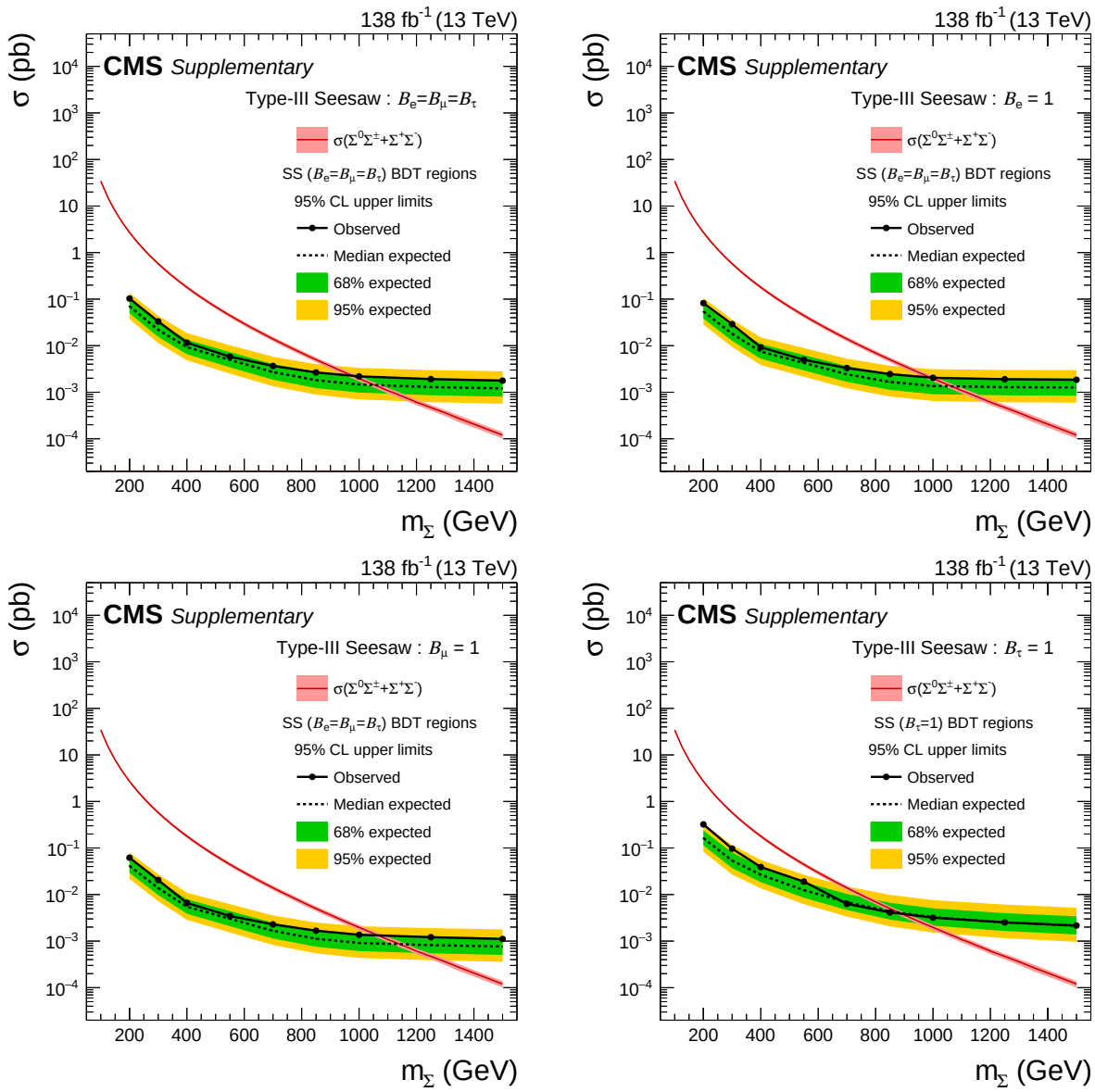

Figure 7.35: Observed and expected upper limits at 95% CL on the production cross section of the type-III seesaw fermions in the $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ scenario (upper left), $\mathrm{B_e} = 1$ scenario (upper right), and $\mathrm{B}_\mu = 1$ scenario (lower left) using the SS $\mathcal{B}_e = \mathcal{B}_\mu = \mathcal{B}_\tau$ BDT regions, and for $\mathrm{B}_\tau = 1$ scenario (lower right) using the $\mathcal{B}_\tau = 1$ BDT regions.

($m_\Sigma > 845$ GeV). This behavior is expected because of the greater efficiency of reconstructing and identifying muons versus τ_h candidates in the experiment.

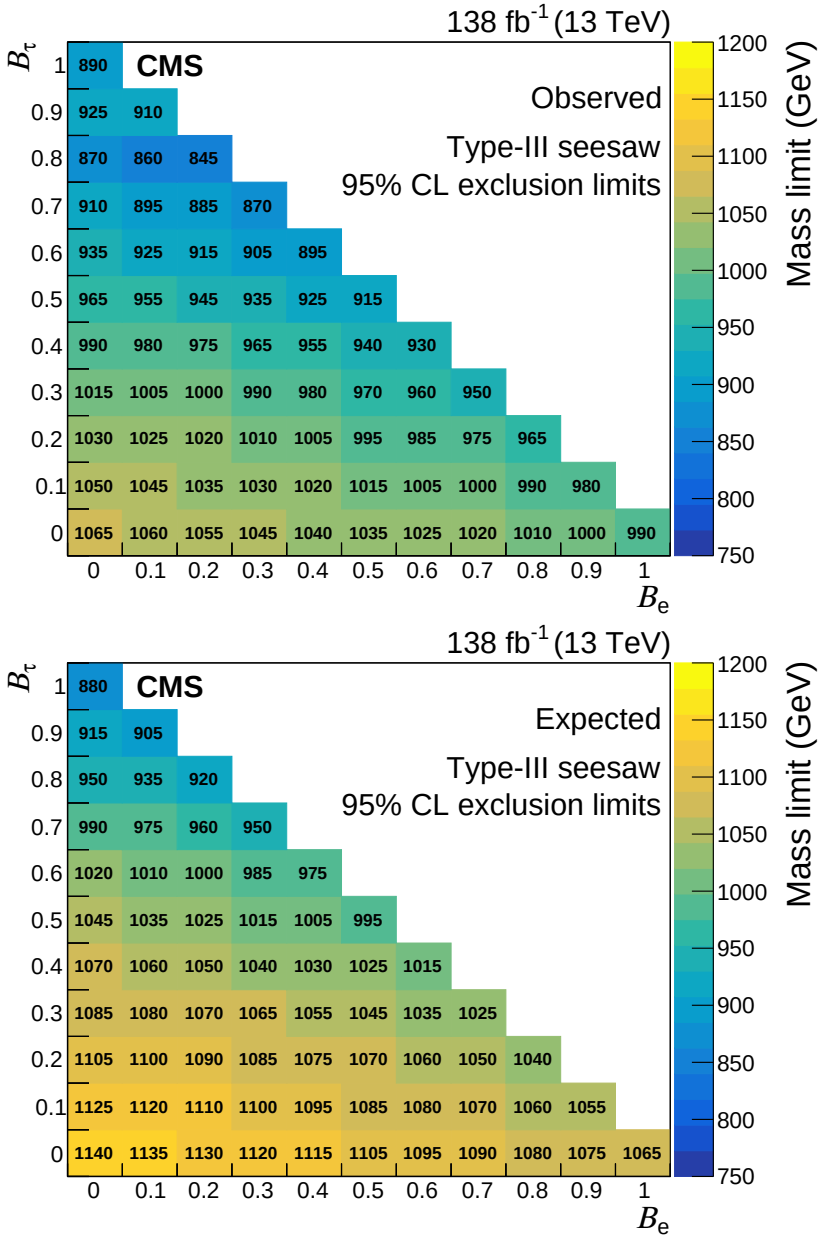

Figure 7.36: Observed (left) and expected (right) lower limits at 95% CL on the mass of the type-III seesaw fermions in the plane defined by $\mathcal{B}_e$ and $\mathcal{B}_\tau$, with the constraint that $\mathcal{B}_e + \mathcal{B}_\mu + \mathcal{B}_\tau = 1$. These limits arise from the *SS-H* $\mathcal{B}_\tau = 1$ BDT when $\mathcal{B}_\tau \geq 0.9$, and by the *SS-H* $\mathcal{B}_e = \mathcal{B}_\mu = \mathcal{B}_\tau$ BDT for the other decay branching fraction combinations.

7.5.3 Scalar leptoquarks

The leptoquark model considered in this thesis has top-philic leptoquarks with couplings to all SM lepton flavors. Similar to the seesaw model, here as well the training scheme addresses this by considering three scenarios: (i) pure light lepton scenario ($ee, \mu\mu$) i.e. $\mathcal{B}_e + \mathcal{B}_\mu = 1$, (ii) pure-$\tau$ scenario i.e. $\mathcal{B}_\tau = 1$, and (iii) mixed scenario ($e\tau, \mu\tau$). We consider four mass ranges in each flavor-mixing scenario, giving a total of 36 BDTs (4 mass-ranges $\times$ 3 flavor scenarios $\times$ 3 years) trained for the leptoquark model. Table 7.5 summarizes the details of the mass points used in training and evaluation.

Table 7.5: Leptoquark signal mass points as used in the trainings of BDTs and as used in the evaluation in the signal regions according to the best expected limit. Separate BDTs are used for 2016, 2017 and 2018 for each of the three flavor-mixing scenario to give a total of 36 trainings.

BDT	Trained masses (GeV)	Applied masses (GeV)
LQ-H	1200, 1300, 1400	800 and higher
LQ-M	500, 600, 700	500, 600, 700
LQ-L	300, 400	300, 400
LQ-VL	200	200

Figure 7.37 shows the calculated expected limits including the complete set of uncertainties evaluated by using a single BDT for the whole mass range. This is done using the $\mathcal{B}_\tau = 1$ BDT training but the same conclusions can be drawn for other BDTs. This informs the choices outlined in Table 7.5.

The BDT regions are independently defined for each channel, for each of the four mass range BDTs, and for each flavor mixing scenario. For a particular signal mass and mixing hypothesis, the BDT which yields the best expected limit is chosen. After testing the performance of various BDTs on different leptoquarks mixing hypothesis, we find that the $\mathcal{B}_e + \mathcal{B}_\mu = 1$ BDT training performs the best for the pure light lepton ($ee, \mu\mu$) and mixed ($e\tau, \mu\tau$) scenarios; whereas $\mathcal{B}_\tau = 1$ BDT training performs the best for the scalar leptoquarks with 100% mixing to taus.

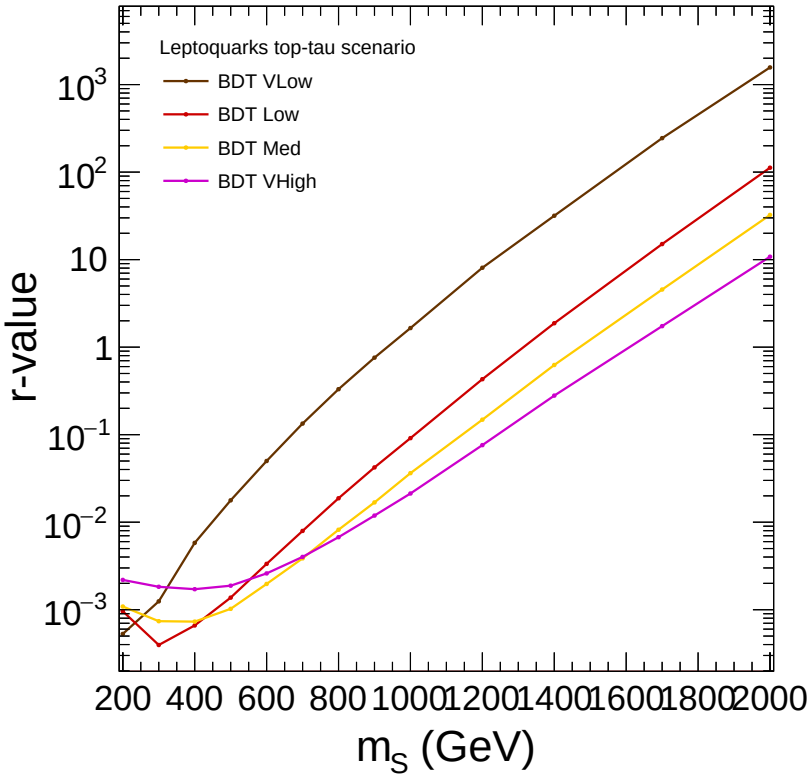

Figure 7.37: The expected limits including the complete set of uncertainties for the leptoquark model in the $\mathcal{B}_\tau = 1$ scenario using a single BDT used for the whole mass range (left). This test informs the choice of BDT for a particular mass point.

The BDT region distributions in the 3-object and 4-object channels for the *LQ-VL*, *LQ-L*, *LQ-M*, and *LQ-H* $\mathcal{B}_\tau = 1$ BDT training in the three years of data-taking are shown in Figure 7.38-7.41, respectively.

The BDT region distributions in the 3-object and 4-object channels for the *LQ-VL*, *LQ-L*, *LQ-M*, and *LQ-H* $\mathcal{B}_e + \mathcal{B}_\mu = 1$ BDT training in the three years of data-taking are shown in Figure 7.42-7.45,respectively.

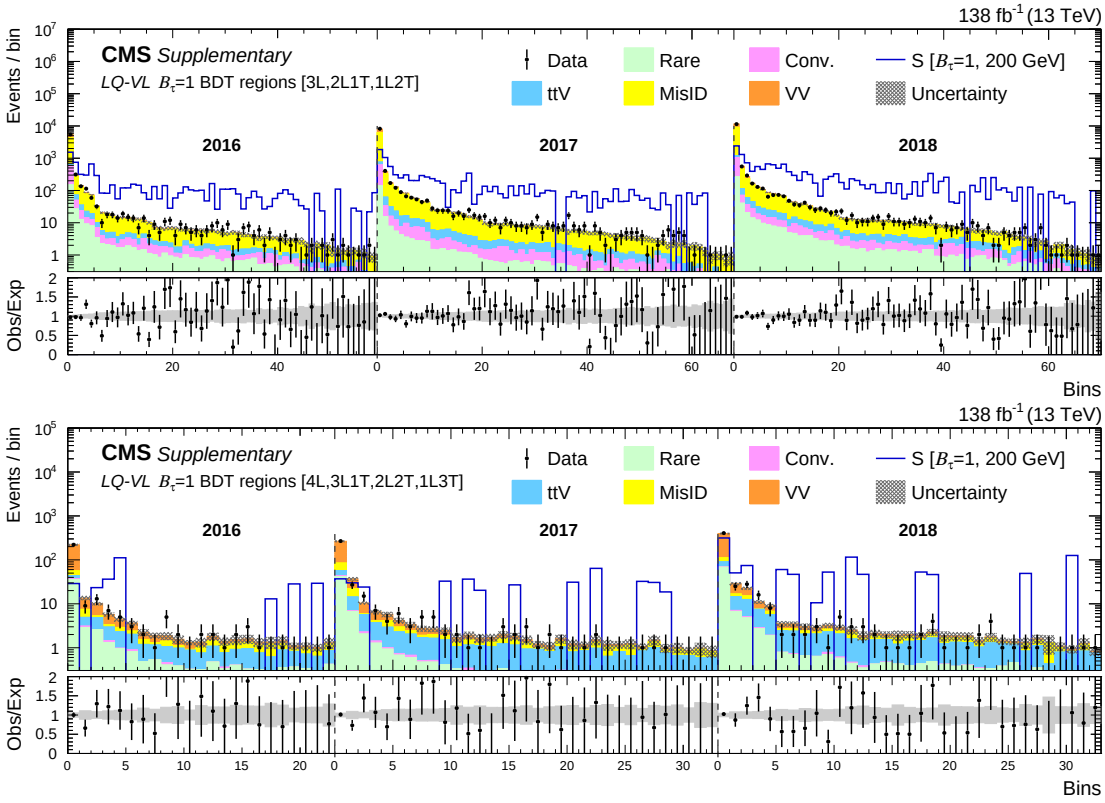

Figure 7.38: *LQ-VL* (upper) BDT regions for the $\mathrm{B}_\tau = 1$ training, in the 3-object channels (upper) and 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the scalar leptoquark coupled to a top quark and a τ lepton, before the fit, is also overlaid.

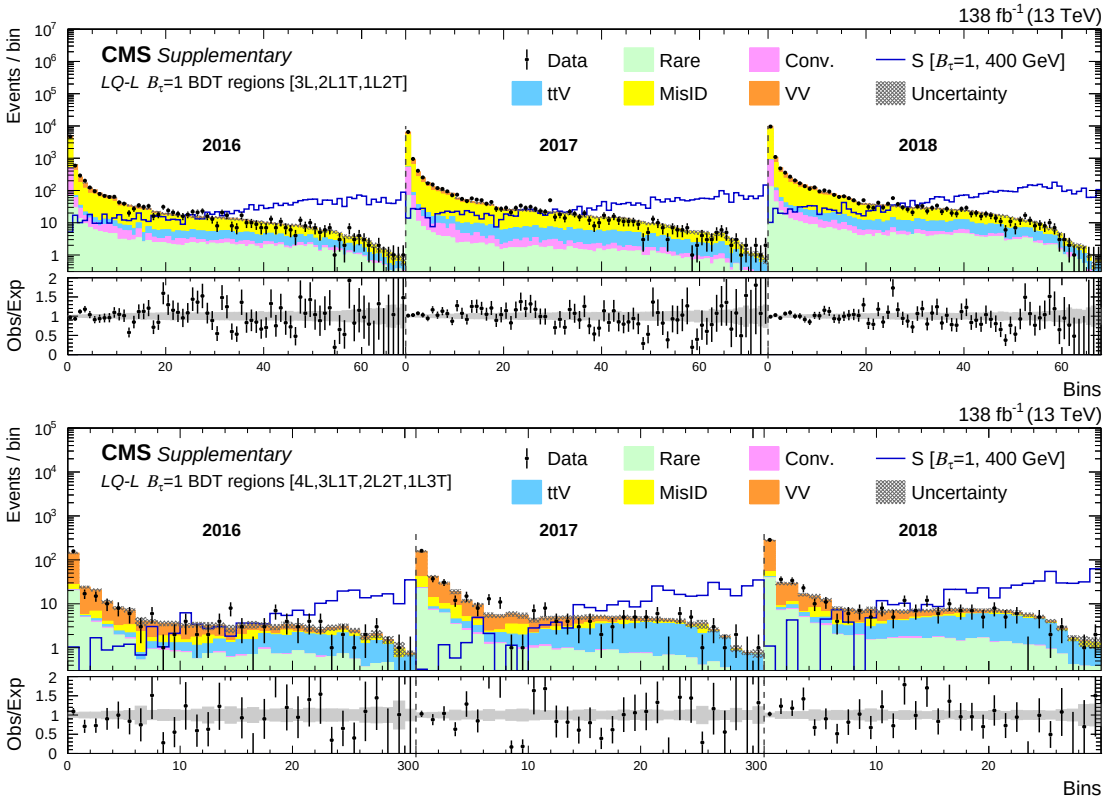

Figure 7.39: *LQ-L* BDT regions for the $\mathrm{B}_{\tau} = 1$ training, in the 3-object channels (upper) and 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the scalar leptoquark coupled to a top quark and a τ lepton, before the fit, is also overlaid.

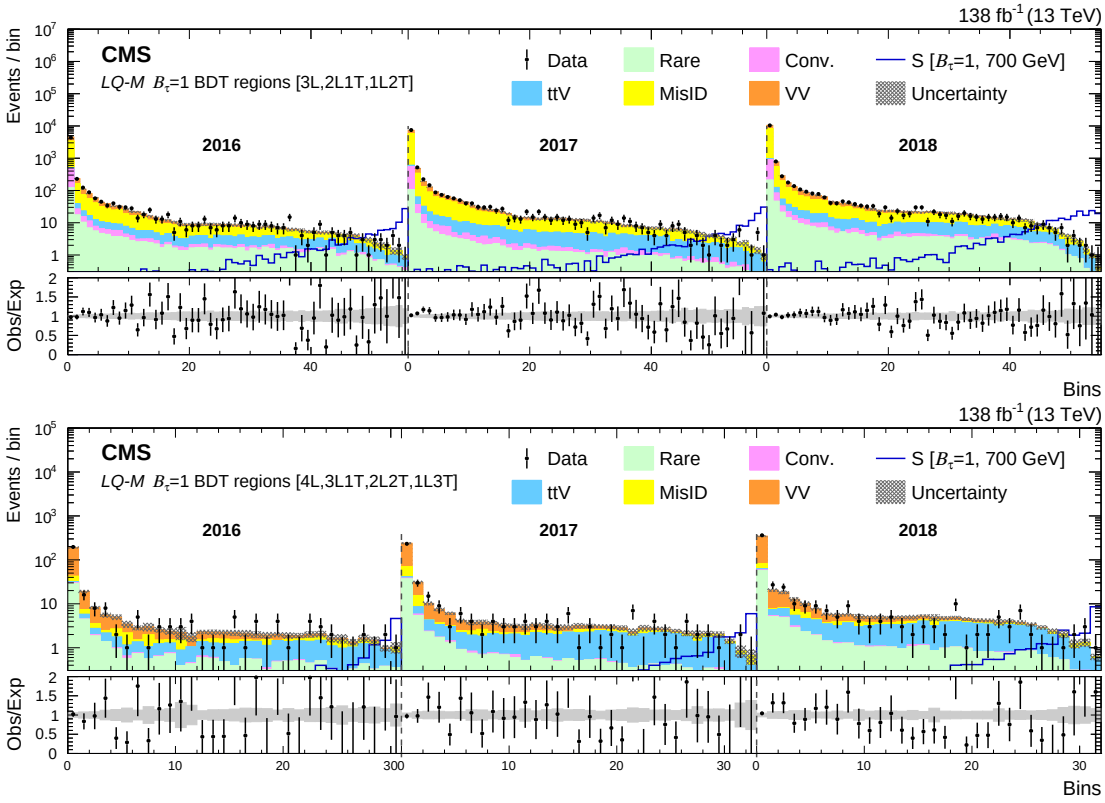

Figure 7.40: *LQ-M* BDT regions for the $\mathrm{B}_{\tau} = 1$ training, in the 3-object channels (upper) and 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the scalar leptoquark coupled to a top quark and a τ lepton, before the fit, is also overlaid.

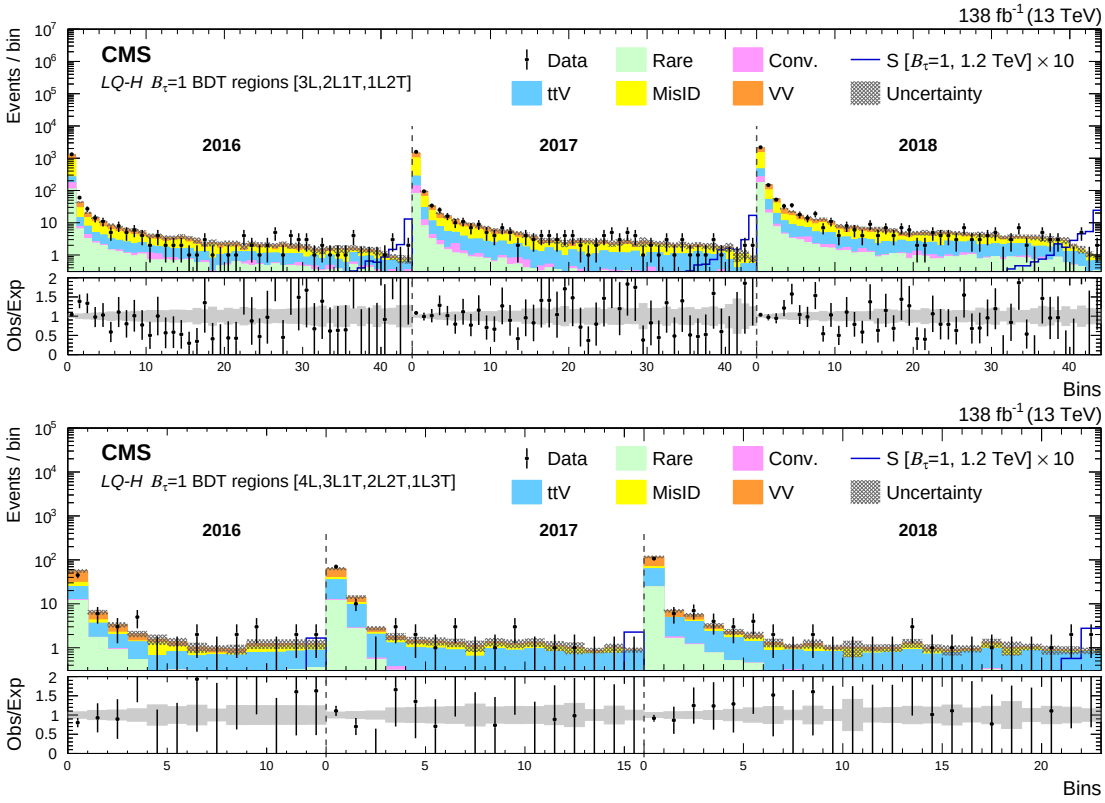

Figure 7.41: *LQ-H* BDT regions for the $\mathrm{B}_{\tau} = 1$ training, in the 3-object channels (upper) and 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the scalar leptoquark coupled to a top quark and a τ lepton, before the fit, is also overlaid.

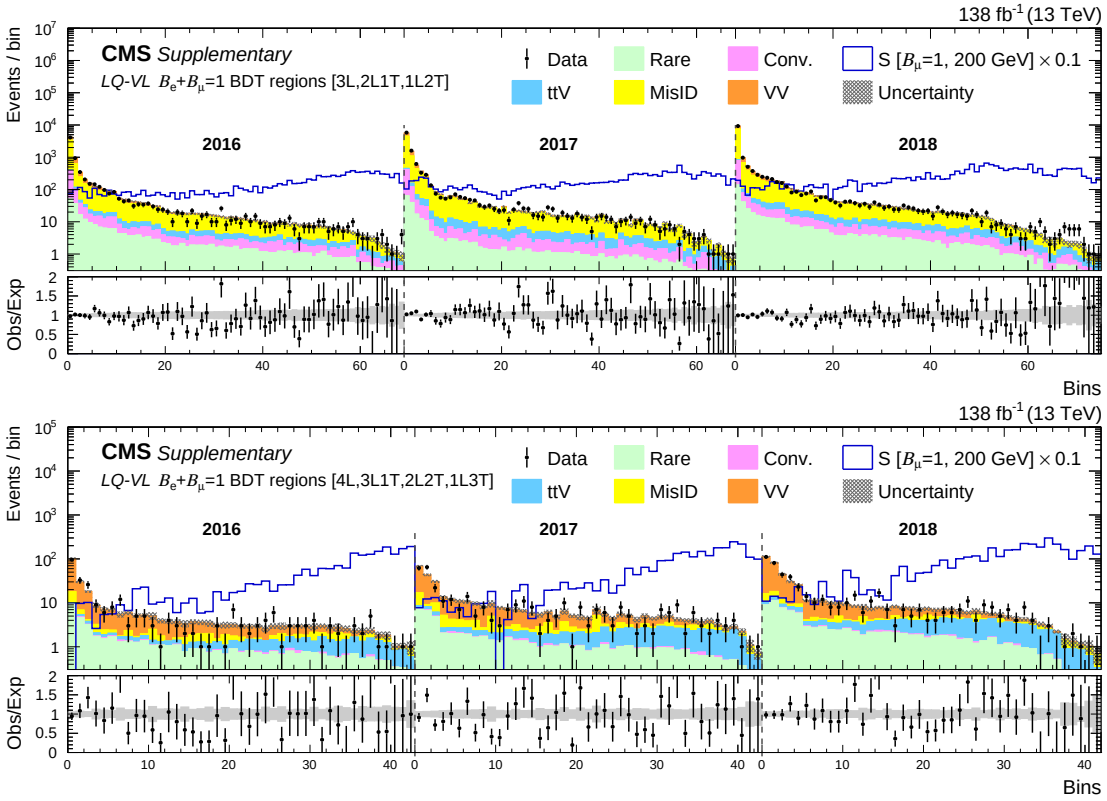

Figure 7.42: *LQ-VL* BDT regions for the $\mathrm{B}_e + \mathrm{B}_\mu = 1$ training, in the 3-object channels (upper) and 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the scalar leptoquark coupled to a top quark and an electron or a muon, before the fit, is also overlaid.

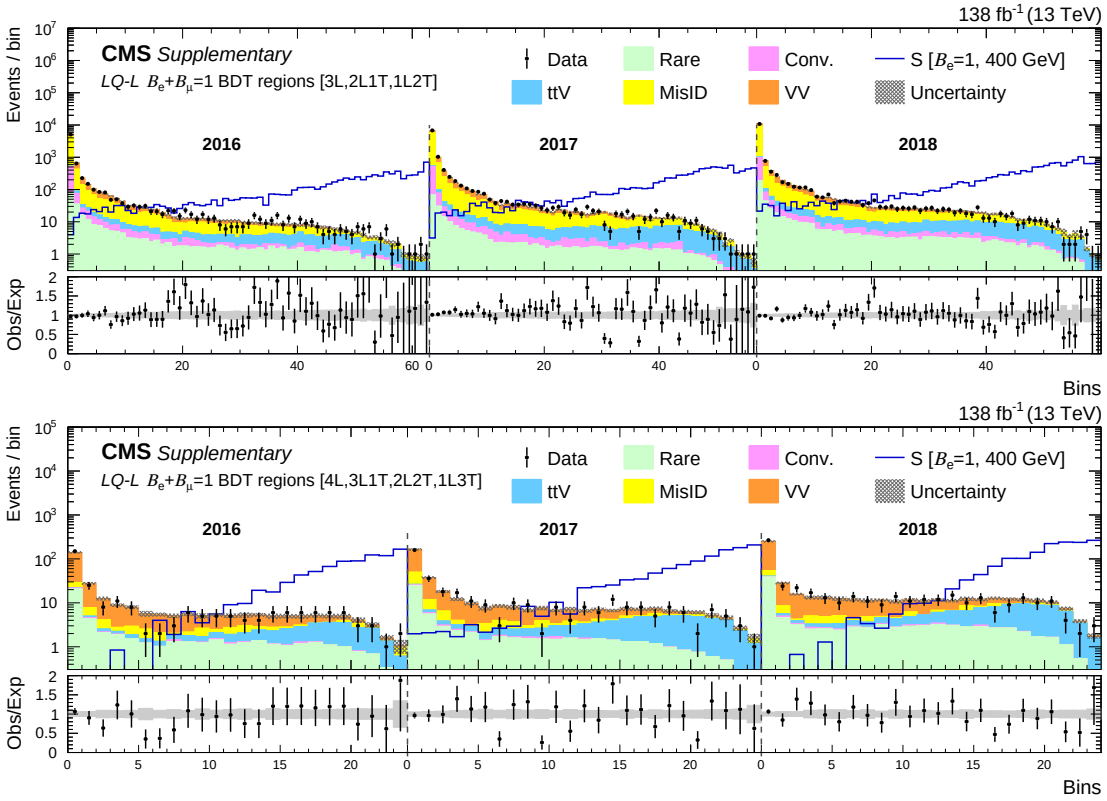

Figure 7.43: *LQ-L* BDT regions for the $\mathrm{B}_e + \mathrm{B}_\mu = 1$ training, in the 3-object channels (upper) and 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the scalar leptoquark coupled to a top quark and an electron or a muon, before the fit, is also overlaid.

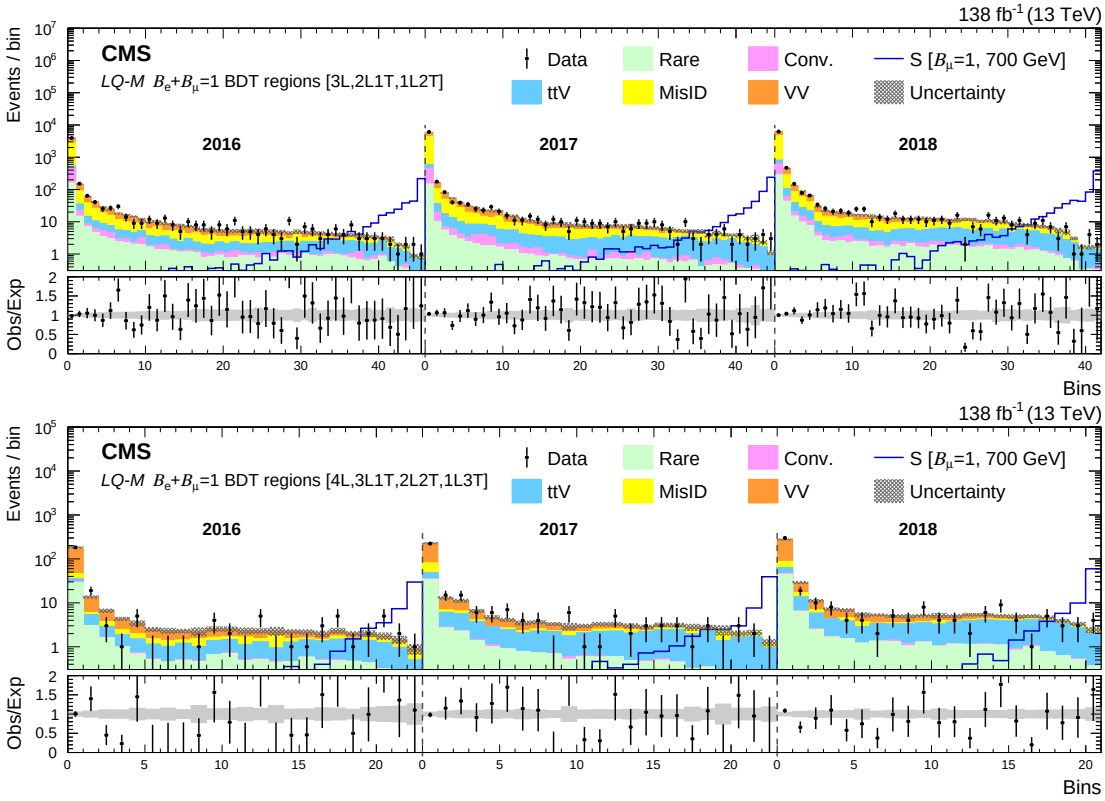

Figure 7.44: *LQ-M* BDT regions for the $\mathrm{B}_e + \mathrm{B}_\mu = 1$ training, in the 3-object channels (upper) and 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the scalar leptoquark coupled to a top quark and an electron or a muon, before the fit, is also overlaid.

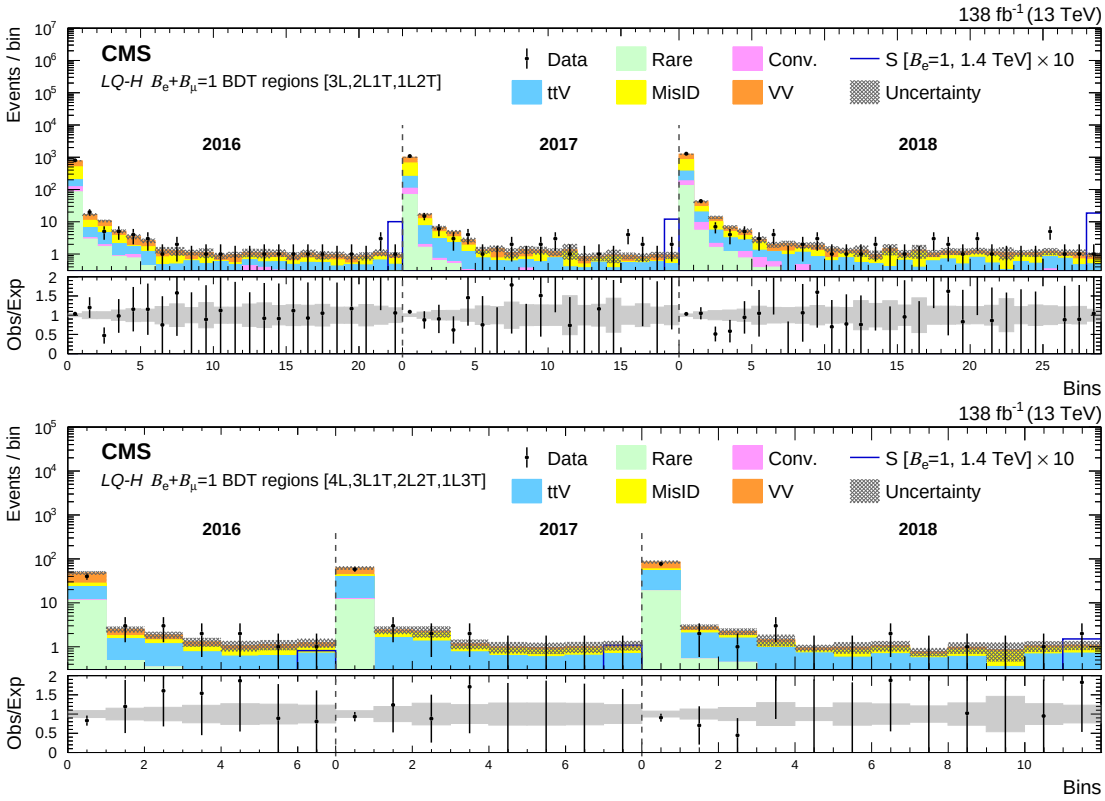

Figure 7.45: *LQ-H* BDT regions for the $\mathrm{B}_e + \mathrm{B}_\mu = 1$ training, in the 3-object channels (upper) and 4-object channels (lower) for the combined 2016–2018 data set. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction. The expected SM background distributions and the uncertainties are shown after fitting the data under the background-only hypothesis. For illustration, an example signal hypothesis for the production of the scalar leptoquark coupled to a top quark and an electron or a muon, before the fit, is also overlaid.

No significant deviations in data wrt the expected SM background prediction was observed in any bin of the LQ $\mathrm{B}_\tau = 1$ and $\mathrm{B}_e + \mathrm{B}_\mu = 1$ BDT regions from both 3-object and 4-object channels in the three years of the data-taking. Figure 7.46 shows example pull distributions for *LQ-H* BDT regions from the $\mathrm{B}_\tau = 1$ (left) and $\mathrm{B}_e + \mathrm{B}_\mu = 1$ (right) trainings, for the combined 2016–2018 data set in the background-only hypothesis. These plots include all the bins from both 3-object and 4-object channels. As can be seen from the figures, all the significances are comfortably within ± 3 sigma deviation.

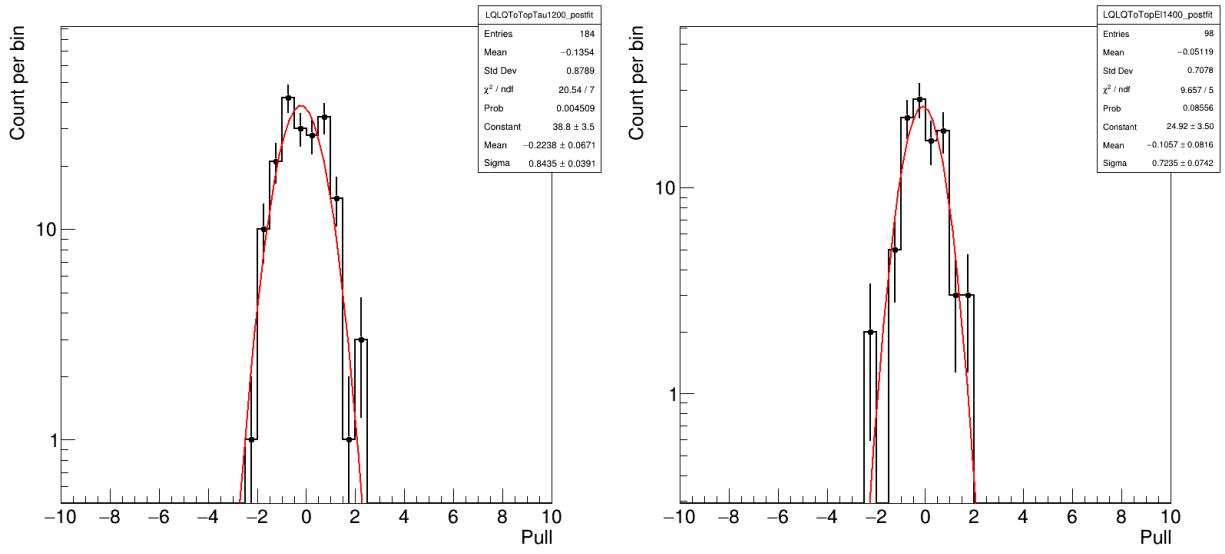

Figure 7.46: Histogram of pulls for the *LQ-H* BDT regions from the $\mathrm{B}_\tau = 1$ (left) and $\mathrm{B}_e + \mathrm{B}_\mu = 1$ (right) trainings, for the combined 2016–2018 data set in the background-only hypothesis. These plots include all the bins from both 3-object and 4-object channels.

Consequently, the observed and expected upper limits at 95% CL were calculated on the production cross section of the scalar leptoquarks and are shown in Figure 7.47 upper left for $\mathcal{B}_e = 1$ scenario and upper right for $\mathcal{B}_\mu = 1$ scenario using the LQ $\mathcal{B}_e + \mathcal{B}_\mu = 1$ BDT regions, and lower for $\mathcal{B}_\tau = 1$ scenario using the LQ $\mathcal{B}_\tau = 1$ BDT regions.

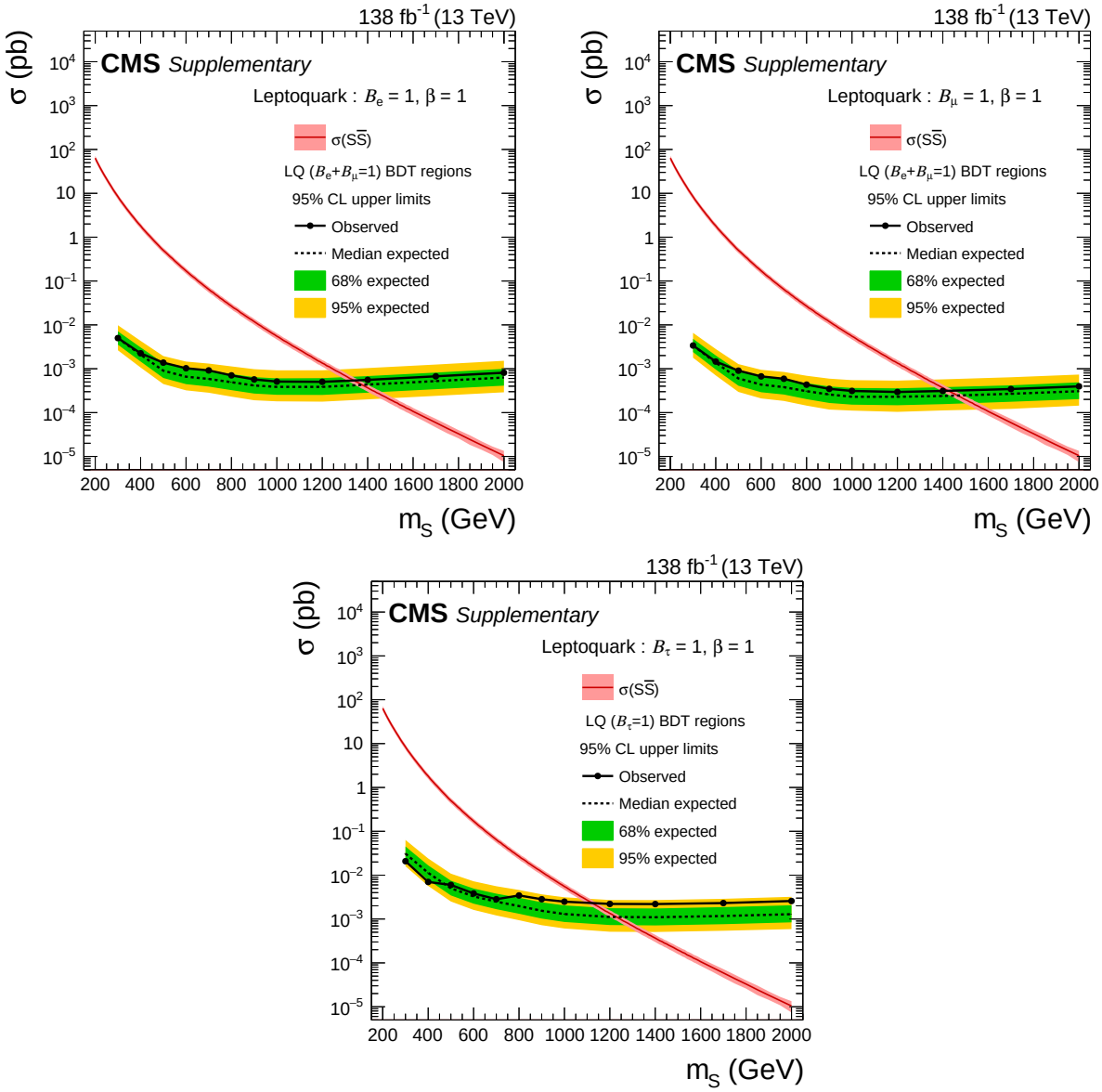

Figure 7.47: Observed and expected upper limits at 95% CL on the production cross section of the scalar leptoquarks with $\mathcal{B}_e = 1$ coupling (upper left) and $\mathcal{B}_\mu = 1$ coupling (upper right) using the LQ $\mathcal{B}_e + \mathcal{B}_\mu = 1$ BDT regions and with $\mathcal{B}_\tau = 1$ coupling (lower) using the LQ $\mathcal{B}_\tau = 1$ BDT regions.

Finding physics the model-agnostic way...

Chapter 8

Model-Independent Search

8.1 Strategy

In addition to performing the dedicated BDT-based search for the BSM signals, we have also designed alternate signal regions using the cut-based method to isolate the regions with high SM background contamination to potential regions with new physics signatures. To do this, we exploited the important features of the SM processes with multileptons in the final state. The basic strategy is outlined as follows:

- Each of the three- or four-lepton channels are split into various lepton charge and flavor combinations, mass, and kinematic regions depending on the dominant SM background processes.
- The primary event classification is performed based on the number of distinct OSSF dilepton pairs in the event considering all lepton flavors. The allowed values for 4 lepton events are OSSF0, OSSF1, or OSSF2, whereas OSSF2 is disallowed for 3 lepton events.
- Subsequently, each OSSF1 and OSSF2 category is further split based on the number of distinct OSSF dielectron or dimuon pairs whose mass is consistent with that of the Z boson in a predetermined window (76-106 GeV), yielding the OnZ and OffZ categories. It is also possible to have two distinct OnZ pairs in a 4L event, labelled as double-OnZ. If the event has no OnZ candidates, then it is classified as an OffZ event. If the event has more than one flavor of OSSF pair, dielecton and dimuon pairs are prioritized over the ditau ones.
- If the OSSF pair is a dielectron or a dimuon pair, BelowZ (< 76 GeV) and AboveZ (> 106 GeV) cateogories are defined, depending on the masses of all light lepton OSSF candidates

with respect to the Z window. A 3L event is classified as MixedZ if there are non-distinct light lepton OSSF pairs in both BelowZ and AboveZ regions. If the OSSF pair is a ditau pair (e.g. in 1L2T), no resonance is sought after due to the invisible component of the tau decays, and such pairs are only categorized as BelowZ or AboveZ with respect to M_{Z} (91 GeV).

- In OSSF0 events, the mass of the OSOF dilepton pair with the largest mass is chosen for classification purposes as BelowZ or AboveZ. Similar to OSSF ditau pairs, OSOF dilepton pairs are also only categorized as BelowZ or AboveZ. OSSF0 events with no OSOF pairs are classified as SS (same-sign) events.

- The 3L and 2L1T channels are further split into two, based on the values of either the M_{T}, or the minimum light lepton p_{T}, or the tau p_{T} variables. In the 3L OnZ channel, an $M_{\mathrm{T}} > 150$ GeVcriterion is used for this binary low/high classification, whereas a minimum light lepton $p_{\mathrm{T}} > 25$ GeVcriterion is used for the rest of the 3L channel and a tau $p_{\mathrm{T}} > 50$ GeVcriterion is used in the 2L1T channel.

This categorization scheme, detailed below in Table 8.1 and labelled as the fundamental scheme, yields 43 orthogonal selections labelled A1-G1. All control regions, which are used in the estimation of major SM backgrounds as described in Chapter 6, are explicitly not used in the fundamental categorization scheme. This scheme allows the complete utilization of multilepton events collected during 2016–2018, such that any event that does not populate a control region is a part of the signal regions.

Figure 8.1 displays the changing composition of the various SM backgrounds in the 43 fundamental categories. As can be seen clearly, some bins are populated by the diboson production (e.g. A8–A12, D4–G1), whereas many of the categories in the tau channels are dominated by the misidentified background (e.g. B11–C5).

In order to gain sensitivity to a large class of BSM scenarios, L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ and S_{T} variables have been chosen as the final discriminating variables in each of the 43 categories, producing the fundamenal L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ table, and the fundamental S_{T} table, respectively. The individual S_{T} and L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ spectra are evaluated in 200 GeV wide bins to ensure smooth and mostly monotonically falling expected background behavior. L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ variable is best suited for BSM models like type-III seesaw with low to moderate hadronic activity, whereas S_{T} variable is expected to be more performant for VLL and leptoquark models with high intrinsic hadronic activity resulting from energetic jets.

A second categorization scheme, the so-called advanced scheme, is also defined building on the fundamental scheme. Each of the 43 fundamental scheme categories is first split, background

Table 8.1: Fundamental scheme of event categorization, as a function of lepton charge combinations and mass variables. The mass categorizations refer to masses of OSSF pairs if present, and of OSDF pairs otherwise. For categorization purposes, all possible opposite-sign dielectron and dimuon pair masses in the event are considered, whereas only the largest mass in the event is considered for all other opposite-sign pairs. Only the dielectron and dimuon pairs are considered to tag events as OnZ. The 1L3T OSSF0 and OSSF1 events are combined into a single category. Disallowed categories are marked with "–".

		OSSF0			OSSF1				OSSF2		
		BelowZ	AboveZ	SS	OnZ	BelowZ	AboveZ	MixedZ	Single-OnZ	Double-OnZ	OffZ
3L	Low $p_{\mathrm{T}}/M_{\mathrm{T}}$	A1	A1	A2	A3	A4	A5	A6	–	–	–
	High $p_{\mathrm{T}}/M_{\mathrm{T}}$	A7	A7	A8	A9	A10	A11	A12	–	–	–
2L1T	Low p_{T}	B1	B2	B3	B4	B5	B6	–	–	–	–
	High p_{T}	B7	B8	B9	B10	B11	B12	–	–	–	–
1L2T		C1	C2	C3	–	C4	C5	–	–	–	–
4L		D1	D1	D1	D2	D3	D3	D3	D4	D5	D6
3L1T		E1	E1	E1	E2	E3	E3	E3	–	–	–
2L2T		F1	F1	F1	F2	F2	F2	–	F3	–	F4
1L3T		G1	G1	G1	–	G1	G1	–	–	–	–

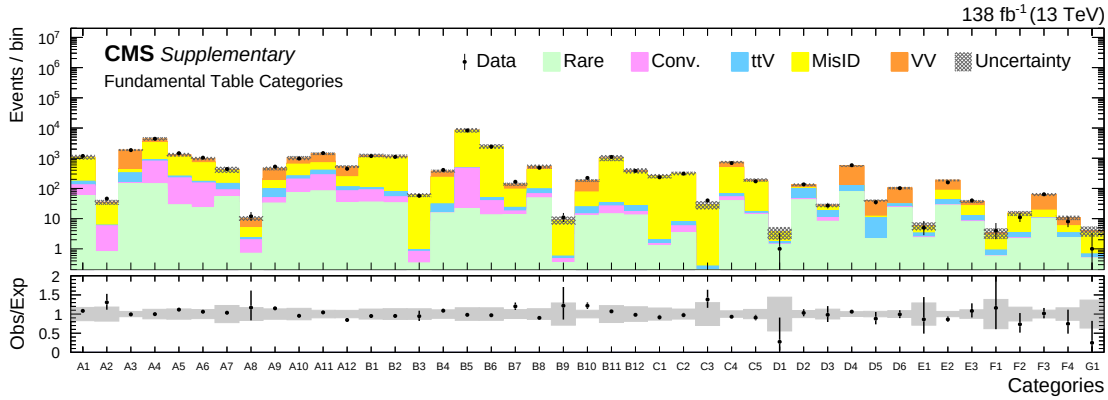

Figure 8.1: The model independent fundamental scheme categories, as defined in Table 8.1. The lower panel shows the ratio of observed events to the total expected background prediction. The gray band on the ratio represents the sum of statistical and systematic uncertainties in the SM background prediction.

statistics permitting, in up to three b-tag multiplicities regions. Furthermore, each category in a given b-tag multiplicity region is split, background statistics permitting, in up to four bins, using binary low or high $p_{\mathrm{T}}^{\mathrm{miss}}$ and H_{T} selection criteria. This results in a total of 204 orthogonal categories. The S_{T} variable is used as the final discriminating variable, producing the advanced S_{T} table, and this table is particularly useful for BSM signals with masses at the electroweak scale.

8.2 Comparison of cut-based vs BDT performance

In order to scrutinize the expected improvements due to machine learning techniques over cut-based approaches, we calculated the upper limits on the production cross-section at 95% CL, for all the signal models that are considered in this analysis. We compare the performance of the three model-independent tables i.e. fundamental L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$, fundamental S_{T}, and advanced S_{T} tables, with the BDT results from Chapter 7.

Consequently, we present the comparison of performance as the ratio of r-values calculated from each of the tables, and divided by the r-value from the BDT results. The MVA limit including the complete set of uncertainties is taken as the denominator in the ratio, and we plot the limits from the three model-independent tables (also including the complete set of uncertainties) in Figure 8.2-8.4. We also show the impact of the systematic uncertainties over the statistical-only MVA limit at all signal masses in the same.

From Figure 8.2 (left), we observe that for the Doublet VLL model, the three tables perform better than the MVA limit up to 280 GeV. The S_{T} tables are particularly more sensitive than the fundamental L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ table because of the rich signal topology with leptons, jets and $p_{\mathrm{T}}^{\mathrm{miss}}$ in the final state. The lowest mass BDT training, VLL-L, is less efficient due to the intrinsic similarity between the signal and SM kinematics at the low masses, and also due to the overall low acceptance of the signal. At masses above 300 GeV, the MVA techniques become much more sensitive than the cut-based tables as the training benefits from the nature of input variables which are more suited for discrimination at higher masses. The impact of systematic uncertainties on the statistical-only MVA limit is larger at the low masses as expected, up to 30-40%, and it asymptotically reaches to less than 10% at the highest probed masses.

For the Singlet VLL in Figure 8.2 (right), the advanced scheme provides the best limits over the complete mass range. This is due to the larger number of counting experiments in the advanced table as compared to the MVA spectrum, especially in the regions sensitive to this particular signal model. Hence, the r-value ratio for this model is calculated with respect to the advanced table, and the MVA limit shown for comparison is statistical-only.

From Figure 8.3, we observe that the Seesaw MVA training is more sensitive for the entire mass range from both the BDTs i.e. the $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ and the $\mathrm{B}_\tau = 1$ scenario. There is a small cross-over in performance between the advanced S_{T} table and the MVA at the 200-300 GeVfrom the SS-M $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ BDT. While the exact reason is difficult to speculate, it could be attributed to the slightly sub-optimal training with the desired combination of signal mass hypothesis. Among the three tables, the fundamental L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ table is the next best after advanced S_{T} up to 400 (500)

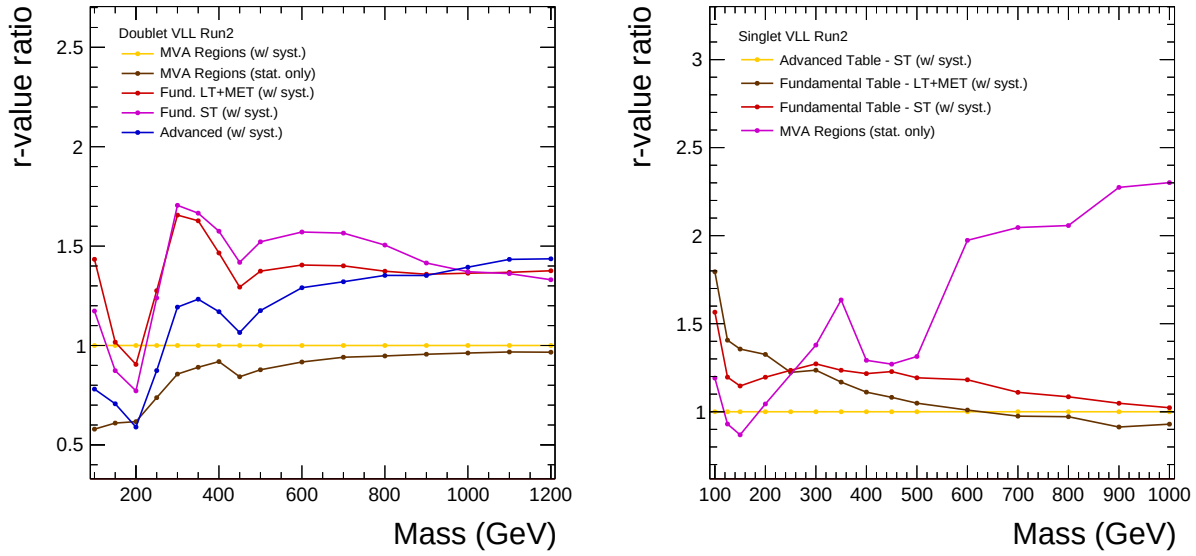

Figure 8.2: The impact of systematic uncertainties on the statistical only MVA limit and the comparison with the different tables on the upper cross-section limits computed as a function of mass for the VLL doublet and singlet signal models are shown. The results are presented as an r-value (signal strength) ratio with respect to the MVA limit with systematic uncertainties for VLL doublet (left). For VLL singlet (right), the r-value ratio is calculated with respect to the advanced table limit with systematic uncertainties and the MVA limit shown is calculated with statistical nuisances only.

GeVin the $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ ($\mathrm{B}_\tau = 1$) signal scenario. This is due to the higher branching ratio of Seesaw fermions to the W and Z gauge bosons, thus giving rise to very high $p_\mathrm{T}$leptons and $p_\mathrm{T}^\mathrm{miss}$ in the final state and lesser hadronic activity. The impact of systematic uncertainties on the statistical-only MVA limit for the BDT training with $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ is around 30-40% at the lower masses, and it rapidly decreases to less than 5% at the highest probed masses. For the BDT training with $\mathrm{B}_\tau = 1$, the impact of systematics uncertainties on the statistical-only MVA limit is around 30% at the lower masses and it asymptotically reaches to around 5% at the highest seesaw mass.

From Figure 8.4, we observe that the leptoquark BDT training with $\mathrm{B_e} + \mathrm{B}_\mu = 1$ is more sensitive than the three cut-based approaches at all the signal mass hypothesis, whereas for the $\mathrm{B}_\tau = 1$ scenario, only the LQ-M and LQ-H BDTs give stronger constraints than the tables. The BDT training for the low mass $\mathrm{B}_\tau = 1$ scenario (up to 500 GeV) is poor than the advanced table due to lower signal acceptance (smaller mass difference between LQ and top quark mass) and similar kinematics of the final states with the SM processes, where the latter table wins due to a higher dimensionality in devising the search regions and larger number of counting experiments. The

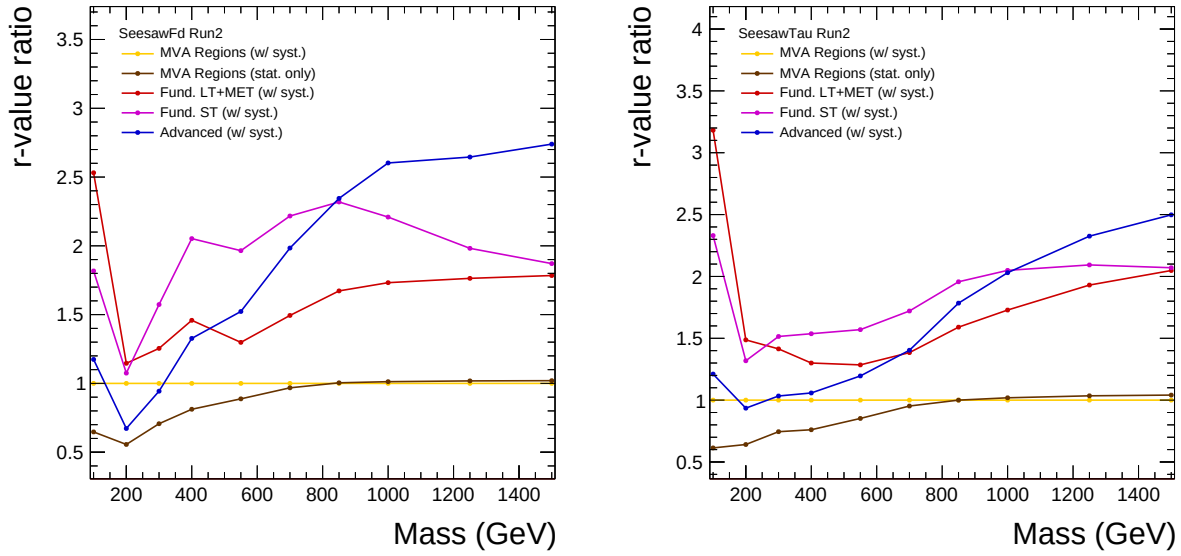

Figure 8.3: The impact of systematic uncertainties on the statistical only MVA limit and the comparison with the different tables on the upper cross-section limits computed as a function of mass for the type-III seesaw model is shown in two scenarios: flavor-democratic (left) and pure-τ (right). The results are presented as an r-value (signal strength) ratio with respect to the MVA limit with systematic uncertainties.

fundamental S_{T} table is consistently better than the L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ table due to much higher hadronic activity from the decays of the two top quarks in each event. The fundamental S_{T} table also performs better than the advanced table above masses 700 GeVas the signal significance goes down in each of the counting experiments due to higher number of bins in the latter scheme. The impact of systematic uncertainties on the statistical-only MVA limit for the BDT training with $\mathrm{B_e} + \mathrm{B}_\mu = 1$ is around 20-30% at the lower masses, and it rapidly decreases to less than 5% at the highest probed masses. For the BDT training with $\mathrm{B}_\tau = 1$, the impact of systematics uncertainties on the statistical-only MVA limit is much larger at the low masses, typically 30-40%, and it asymptotically reaches to around 5% at the highest LQ mass.

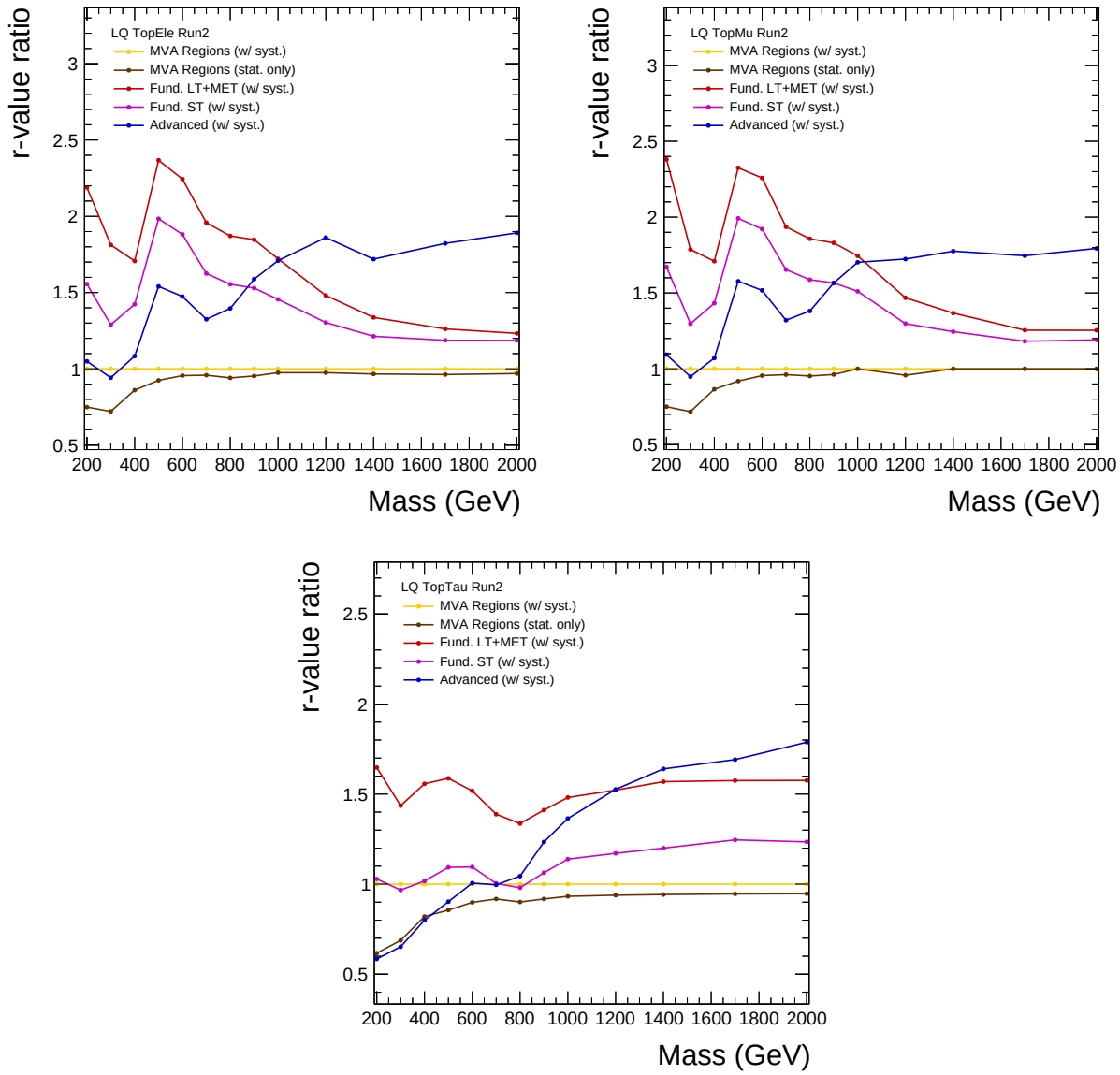

Figure 8.4: The impact of systematic uncertainties on the statistical only MVA limit and the comparison with the different tables on the upper cross-section limits computed as a function of mass for the scalar leptoquarks model is shown in three scenarios: $\mathcal{B}_e = 1$ (upper left), $\mathcal{B}_\mu = 1$ (upper right), and $\mathcal{B}_\tau = 1$ (bottom) couplings. The results are presented as an r-value (signal strength) ratio with respect to the MVA limit with systematic uncertainties.

To summarize, with the use of MVA techniques, the limits on the signal cross-section at higher masses improve typically by 25-50% wrt the fundamental tables, whereas the improvements with respect to the advanced table ranges from 50-100% between different signal models.

8.3 Suboptimal performance of low-mass BDTs

Effective training of the BDT requires sufficient training statistics. At low signal masses, the signal acceptance is lower, and thus fewer events are available for training. The BDT has to "learn" particular selections, and also must learn it robustly (the "boosted" in BDT). The simple cut-based classification we use, is in fact still fairly complex and benefits from a thorough and intensive binning. The best sensitivity comes from the advanced scheme with 204 categories, where the cuts are imposed and therefore do not need to be learned. Hence, they can be thought of as one of the decision trees during BDT training, among the dense forest, with a reasonable signal-to-background ratio but poor reproducibility (accuracy). As a general remark, the model-independent approach is more sensitive than the lowest-mass BDT trainings for all the models. This is because at low signal masses, the BDT training is impacted by the low signal yield, and similar kinematics of signal and SM processes.

Furthermore, we try to address and demonstrate the above laid facts behind why the MVA provides weaker constraints for low signal mass hypotheses. This feature is present in almost all signal models considered, as seen in Figures 8.2-8.4. We do this study by taking the example of one signal model, for a single mass hypothesis viz. Doublet VLL of mass 200 GeV.

1. **BDT training at low mass:** The BDT training is one of the most important factors determining the MVA performance. At low signal masses, the kinematics of signal are very similar to the SM processes; thus the BDT finds it difficult to discriminate between the signal and the background. Aside from the kinematics, the overall signal yield is also lower at low masses (lower acceptance), which impacts training as well as final performance in terms of constraints.

   Figure 8.5 shows the ROC curves for the low, medium, and high mass BDT trainings in 2018 for the 3-object (upper row) and 4-object (lower row) channels for the VLL doublet and VLL singlet model (using signals with mass 150 GeV, 300 GeV, 700 GeV). Clearly, the training improves as we go from low to high masses. It should be noted that while the training of the doublet and singlet models is not so different in terms of performance, the constraints from BDT for the singlet model are significantly worse than the advanced table approach

(across the entire mass range) - thus highlighting the role of low overall signal yield in the final constraints.

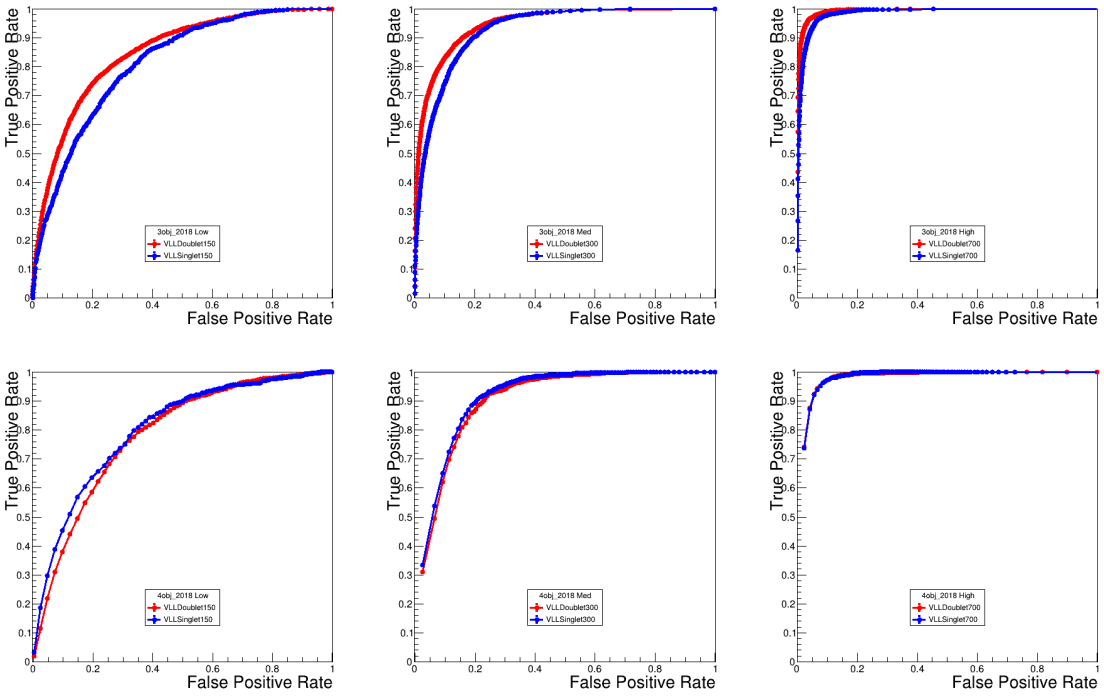

Figure 8.5: The ROC curves for the VLL-L (left), VLL-M (middle) and VLL-H (right) BDT trainings in 2018. The upper row is for 3-object channels and the lower row is for 4-object channels.

2. **Higher dimensionality of the cut-based tables:** The model-independent cut-based approaches provide sensitivity for specific signal processes by virtue of smart binning in many variables. The higher dimensionality in the advanced table scheme makes it more sensitive than the fundamental table scheme; since even at the low end of the S_{T} spectra, additional divisions into several regions improve sensitivity.

   The MVA regions, on the other hand, provide sensitivity mainly at high values of the BDT score. If a signal does not populate the extreme high BDT score, then there are no other selections to enhance sensitivity. Indeed, this approach is not favored for the MVA, since the cut-based tables are precisely this approach which already gives us the desired sensitivity.

   The MVA approach provides stronger constraints in extracting sensitivity, while reducing the overall dimensionality. Figure 8.6 shows the comparison of number of bins with different total background yields between the advanced table scheme, and VLL-L (left), and VLL-M BDTs (right) for all the counting experiments (bins) we do to extract the final Run 2 limits.

Figure 8.7 shows the $S/\sqrt{B}$ for all the bins (using the Doublet VLL with mass 200 GeV). From these two figures, we see that the advanced table has many more low-background yield bins. In terms of high sensitivity bins, the numbers are comparable, with the advanced table having slightly more significant bins; which results in better sensitivity. At higher signal masses, the total signal yield gets divided into many bins in the advanced table, while it remains integrated in the MVA regions. This results in higher sensitivity from the MVA at high signal masses.

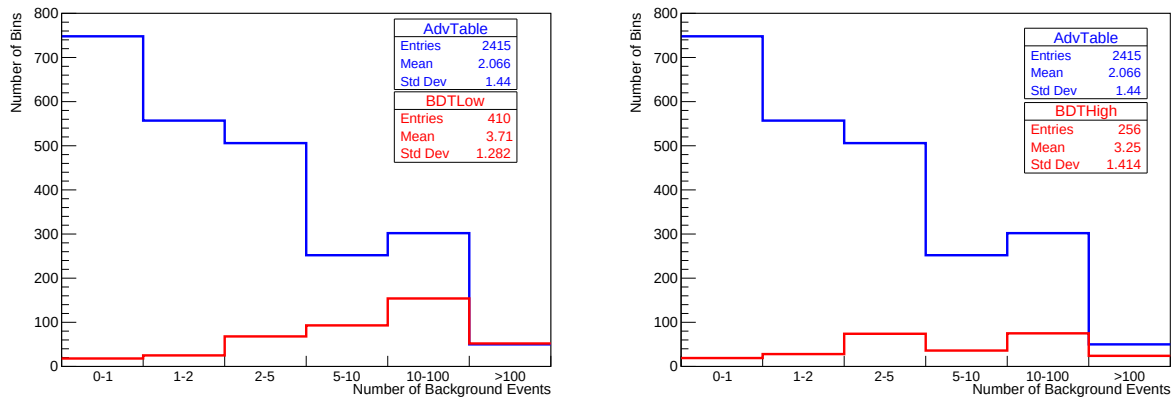

Figure 8.6: The distribution of number of bins with different total background yields in the advanced table scheme and VLL-L BDT (left) and the VLL-H BDT (right) for the full Run-2 dataset. The blue and red curves represents the advanced table scheme bins and the nominal MVA regions respectively.

3. **Binning scheme for the MVA regions:** The final comparative performance of the constraints from the MVA approach and the cut-based approach depend on several inter-dependent factors. Aside from the MVA training parameters (variables and mass ranges used for each BDT training, overall lower signal acceptance at low masses), the choices of boundaries to make MVA regions also play some role in affecting the sensitivity of the MVA approach. Given the extremely large number of signal models and mass hypotheses considered in this analysis, we pick these boundaries based on the SM background yield alone.

   To check if the choice of boundaries plays a major role, we perform the following illustrative study: For the VLL-L BDT, in the 4-object channels, we redesign the MVA regions - increasing the number of low background yield bins by almost a factor of 6. Figure 8.8 shows the redesigned MVA regions (and can be compared to Figure **??** lower one). We recalculate the upper limit for VLL doublet mass 200 GeVwith these redesigned 4-object regions (combin-

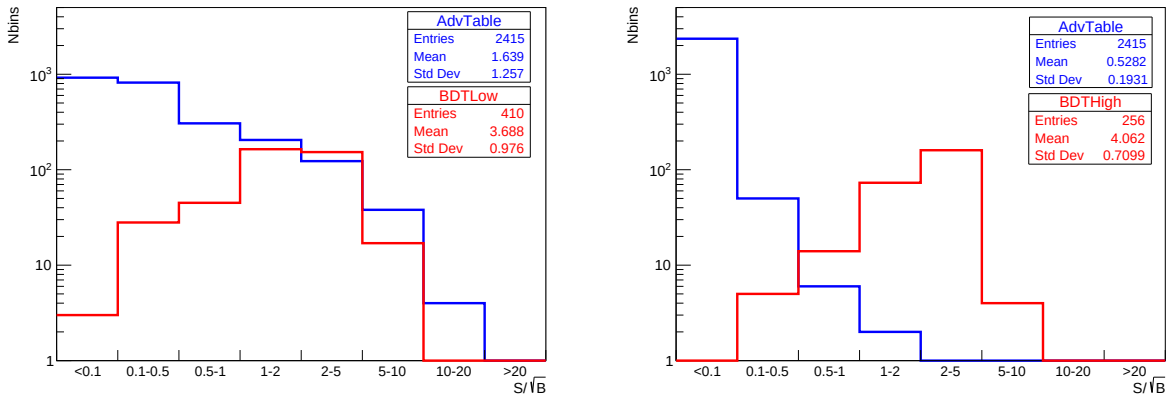

Figure 8.7: The distribution of number of bins with varying signal significance (for Doublet VLL with mass 200 GeV) in the advanced table scheme and VLL-L BDT (left) and the VLL-H BDT (right) for the full Run-2 dataset. The blue and red curves represents the advanced table scheme bins and the nominal MVA regions respectively.

ing them with the orignal 3-object regions), and find that the improvement in upper limits is a meagre 3% (5% considering limits from 4-object channels alone). Figure 8.9 shows the comparison of the redesigned MVA regions to the nominal approach; we see that even with a significant increase in number of bins, the specific number of high sensitivity bins does not change drastically.

It is worthwhile to note here that the BDT score is not a physical variable such as L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ or S_{T}. The BDT score has a complex dependence on the input variables. The SM (and signal) events that populate the non-extreme values of the BDT score could arise from several different combinations of the input variables - and having a large number of bins in the MVA is not justified without significant gains in sensitivity - in fact, those gains are already “in the bank” from our advanced table approach.

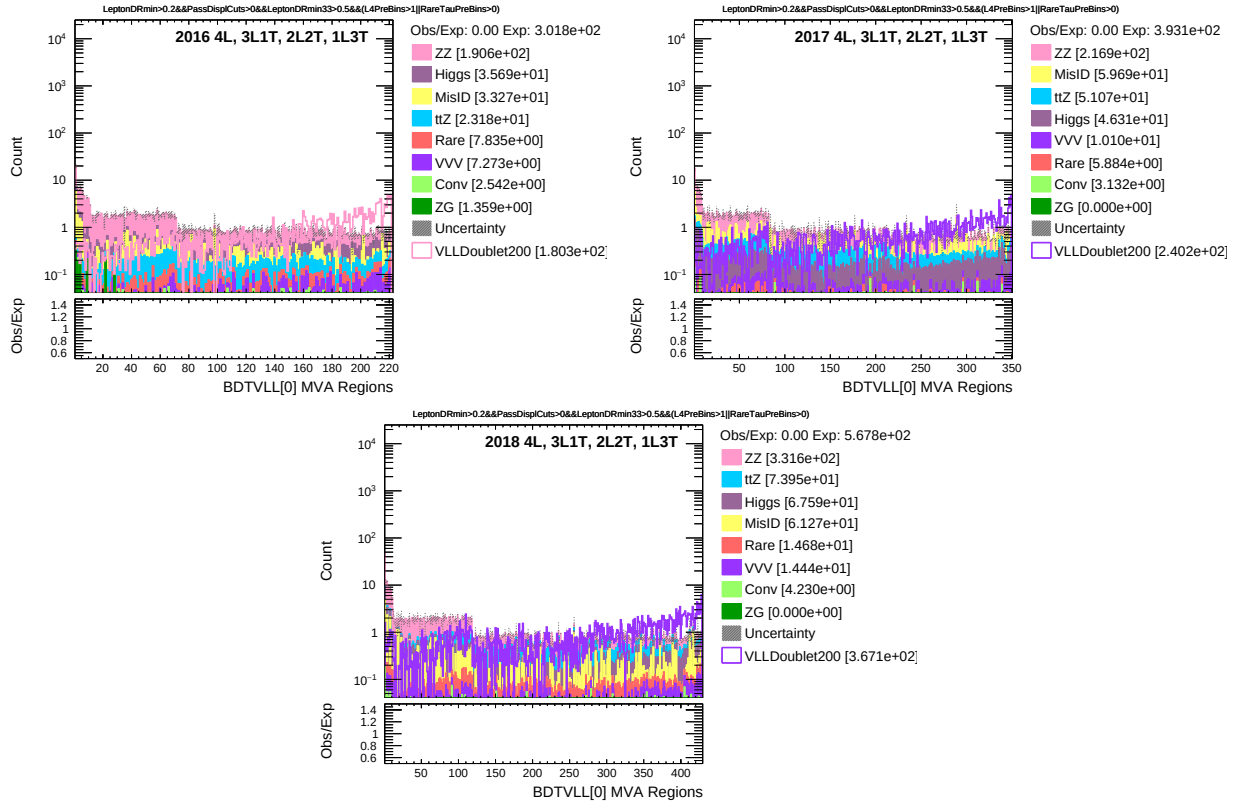

Figure 8.8: The MVA regions for the VLL-L BDT in 2016, 2017 and 2018 for the 4-object channels with a modified binning scheme designed to yield many low background bins.

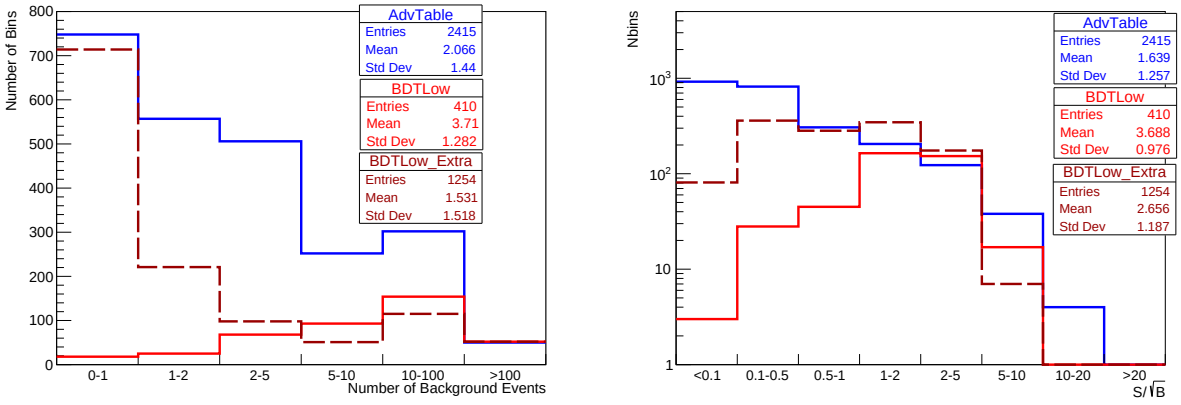

Figure 8.9: The distribution of number of bins with different total background yields (left) and signal significance (right) in the advanced table scheme and VLL-L BDT for the full Run-2 dataset. The blue, red, and maroon dotted curves represents the advanced table scheme bins, nominal MVA regions and new MVA regions with many low background yield bins respectively.

8.4 Best constraints on the probed models

Observed and expected upper limits at 95% CL on the production cross section of the vector-like leptons in the doublet and singlet scenario are shown in Figure 8.10 left and right, respectively. For the doublet model, vector-like τ leptons are excluded up to a mass $m_{\tau'}$ of to 1045 GeV, where the expected mass exclusion is 975 GeV. The best expected limit for $m_{\tau'} < 280$ GeV is given by the advanced S_{T} table scheme, and by the BDT regions for larger masses. For the singlet model, the best expected limits are given by the advanced S_{T} table over the entire mass range. Singlet vector-like τ leptons are excluded in the mass interval from 125 to 150 GeV, while the expected exclusion range is from 125 to 170 GeV.

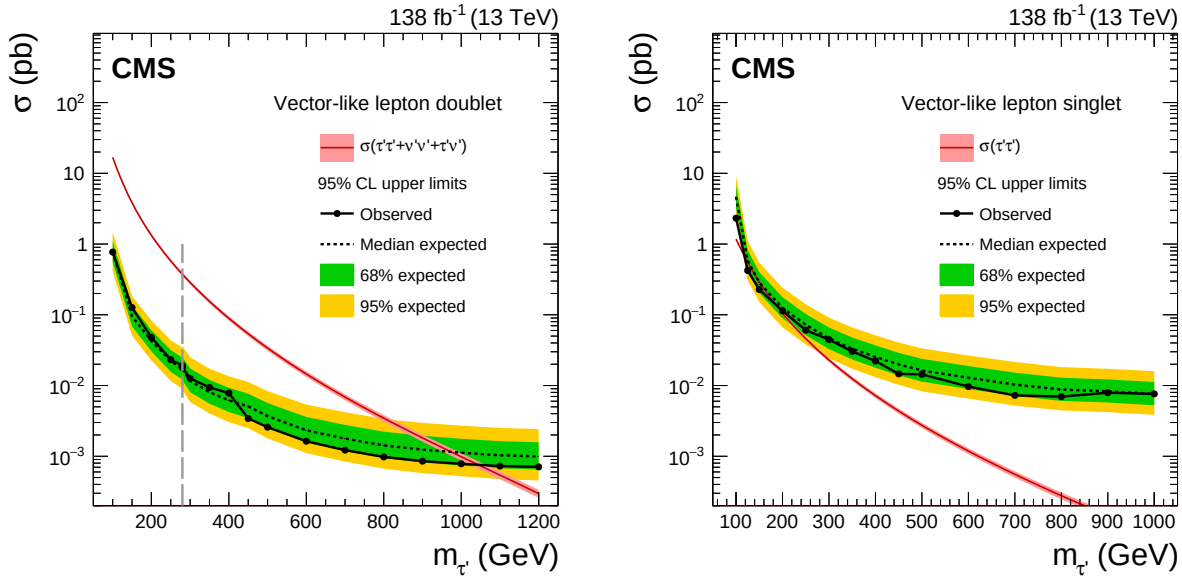

Figure 8.10: Observed and expected upper limits at 95% CL on the production cross section of the vector-like τ leptons: doublet model (left), and singlet model (right). For the doublet vector-like lepton model, to the left of the vertical dashed gray line, the limits are shown from the advanced S_{T} table, while to the right the limits are shown from the BDT regions. For the singlet vector-like lepton model, the limit is shown from the advanced S_{T} table for all masses.

Observed and expected upper limits at 95% CL on the production cross section of the type-III seesaw heavy fermions in the flavor-democratic scenario are shown in Figure 8.11. The observed (expected) lower limit on m_{Σ} in this scenario is 980 (1060) GeV. The best expected limit is given by the advanced S_{T} table scheme for $m_{\Sigma} < 350$ GeV, and by the BDT regions for higher signal mass values.

Observed and expected upper limits at 95% CL on the production cross section of the scalar

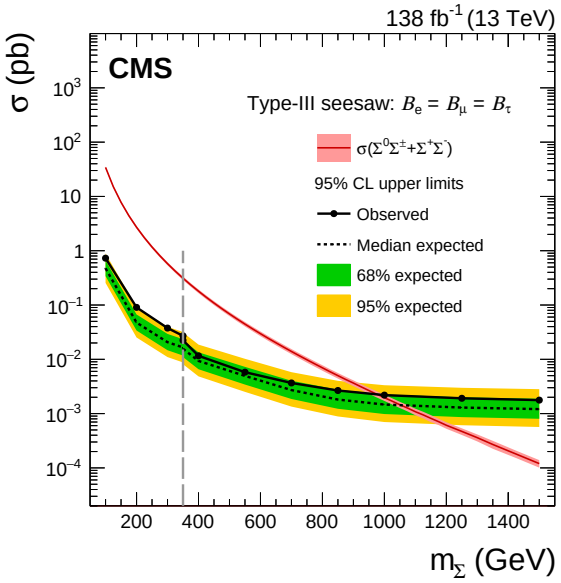

Figure 8.11: Observed and expected upper limits at 95% CL on the production cross section of the type-III seesaw fermions in the flavor-democratic scenario using the table schemes and the BDT regions of the *SS-M* and the *SS-H* $\mathcal{B}_e = \mathcal{B}_\mu = \mathcal{B}_\tau$ BDTs. To the left of the vertical dashed gray line, the limits are shown from the advanced S_{T} table, and to the right the limits are shown from the BDT regions.

leptoquarks exclusively coupling to top quark and a muon, top quark and an electron, and top quark and a τ lepton are shown in Figure 8.12 upper left, upper right, and lower, respectively. For a leptoquark S exclusively coupling to a top quark and a muon, the observed (expected) lower limit on the mass of pair produced leptoquarks is 1420 (1460) GeV. For the top quark and electron decay scenario, the observed (expected) lower limit on m_S is 1340 (1370) GeV, while for the top quark and τ lepton decay scenario, the lower limit is 1120 (1235) GeV. The advanced S_{T} table gives the best expected limit for m_S less than 400, 400, and 500 GeVfor the $\mathcal{B}_\mu = 1$, $\mathcal{B}_e = 1$, and $\mathcal{B}_\tau = 1$ scenarios, respectively.

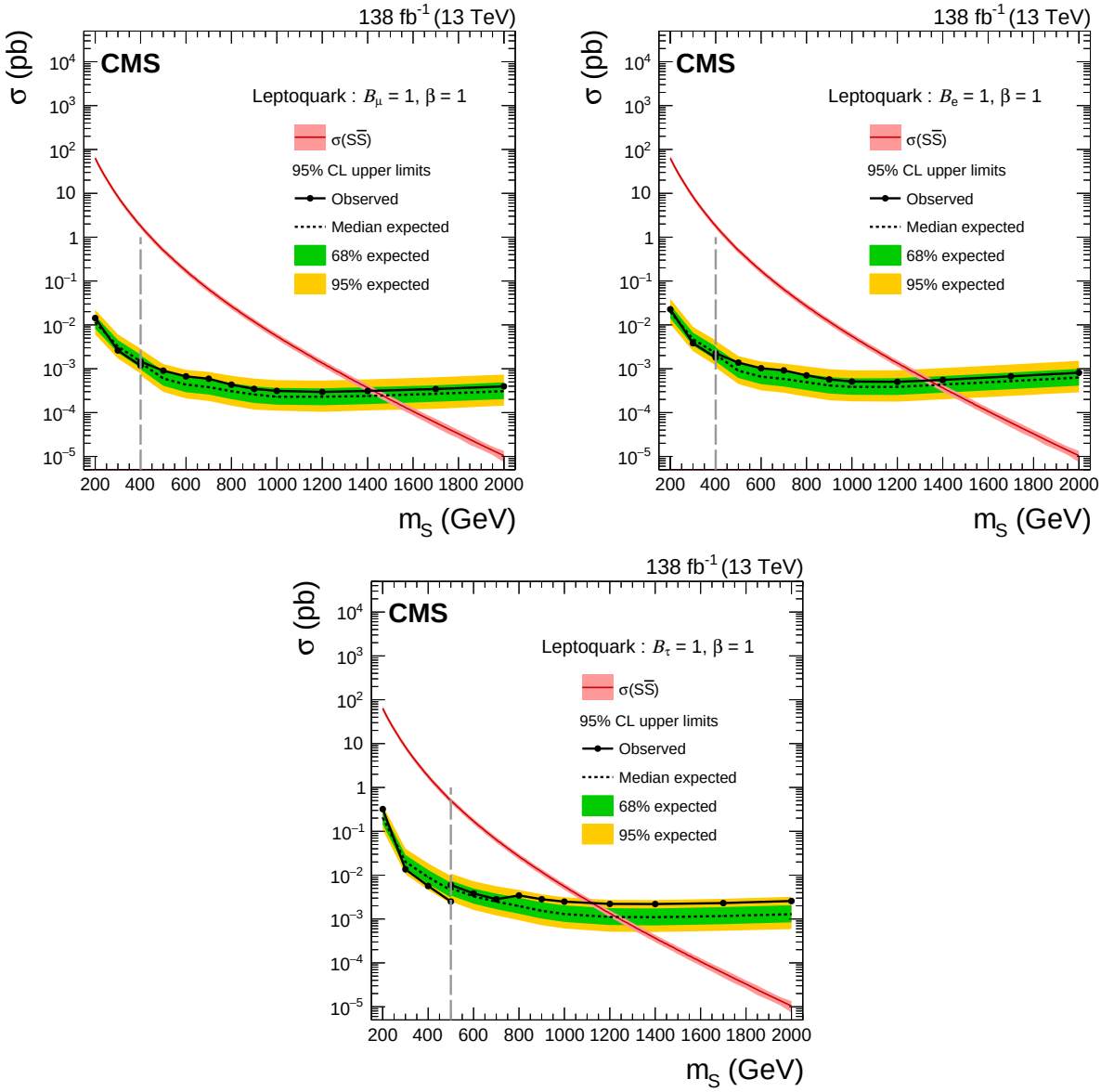

Figure 8.12: Observed and expected upper limits at 95% CL on the production cross section of the scalar leptoquarks: $\mathcal{B}_\mu = 1$ (upper left), $\mathcal{B}_e = 1$ (upper right), and $\mathcal{B}_\tau = 1$ (lower). In each figure, the limits to the left of the vertical dashed gray line are shown from the advanced S_{T} table, and to the right are shown from the BDT regions.

Paving the way for future BSM searches...

Chapter 9

Reinterpreting the search results

The LHC experiments have played a crucial role over the years in understanding the SM phenomena via precision measurements of the various fundamental matter particles and force-carrier bosons. Not only that, there are many direct as well as indirect searches for evidence of BSM physics in a vast variety of final states from these experiments. Still, there are a multitude of compelling BSM theories with large parameter spaces which are unexplored, either because of many unknowns of the theory or because it is beyond the physics reach of the experiments. Hence, the searches are sensitive to only a small subset of possible theories and their phase spaces. Often the subjects of analysis interpretations are simplified models, designed to facilitate searches, but ultimately are unable to address the full phenomenology.

In order to determine the implications of LHC data for a broad range of theories, the experimental collaborations are actively encouraged to provide supplementary information in addition to the primary results (the observed and the expected SM background yields along with their associated uncertainty in the signal regions). This opens the gateway to reinterpret the theories, following a statistical analysis, not probed in the original analysis. The feedback from reinterpretation also allows the phenomenology community to tweak the theories appropriately, and to better suggest promising BSM scenarios within the physics reach of the experiments.

This multilepton analysis is a benchmark result covering almost the entire multilepton arena, with the exception of minimum one light lepton in the final state. The analysis is performed with the combined 2016–2018 data set, which corresponds to an integrated luminosity of $\mathcal{L} = 138$ fb^{-1}. This is a significant amount of data to be collected and analyzed so far, and it would take a few upcoming years of the LHC operation until we collect enough data, or do significant developments (use of ML techniques, lower trigger thresholds etc.) to improve the current sensitivity. Inspired by the need for reinterpretation, and considering the vast applicability and importance of

our benchmark result, we did extra measurements and designed a roadmap to help the end users reinterpreting the results. This is described in the subsequent sections.

9.1 Procedure

The starting point for reinterpretation of any BSM model is the selection of a signal region, whose acceptance can be reproduced from the generator-level properties of an event. The model-independent signal regions, as described in Chapter 8, which are purely designed using a cut-based method are an ideal choice. Hence, for a given a specific BSM model, a particular model-independent scheme (fundamental or advanced) should be selected. The fundamental $L_{\mathrm{T}} + p_{\mathrm{T}}^{\mathrm{miss}}$ table will be sensitive to BSM models that produce primarily leptons and missing energy, while the S_{T} tables will be more sensitive for models which populate final states with several jets, which may or may not arise from b-quarks. The reconstructed yield for the model should then be derived in the various categories of the chosen scheme.

To obtain the reconstructed yield from generator-level kinematic properties in the SRs, the acceptance and efficiency for the model is required. The acceptance is defined as the fraction of generated events passing the analysis-level selections, which is then multiplied with the various lepton and global efficiencies to account for detector effects. Figure 9.1 shows the product of acceptance and efficiency for the vector-like τ lepton model in the doublet scenario (left) and for the type-III seesaw fermions in the $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ scenario (right). These are calculated in the inclusive signal regions of all seven multilepton channels separately. The product is defined as the ratio of the total reconstructed yield in a given channel (after all the corrections and scale factor implementation) to the product of luminosity and the production cross section of the given simulation sample.

Every BSM model has its own phenomenology, and therefore different acceptance for the analysis-level selections. There are only a finite number of possible models for which these product of acceptance and efficiency can be provided from experimentalists. Hence, in order to target a broader audience reinterpreting our results, we also provide reconstruction efficiency maps for the leptons selected in this analysis, so that the yield of any given BSM process can be predicted using purely generator-level information. Section 9.2 describes the measurement of the lepton reconstruction efficiency maps in detail, and Section 9.3 discusses the procedure to derive the signal yields in our signal regions.

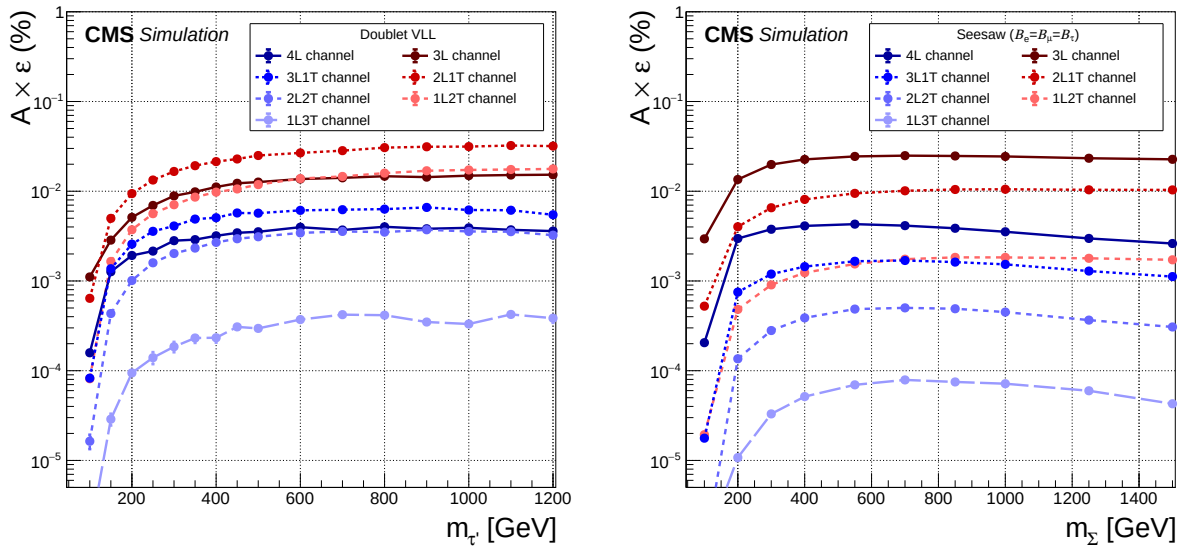

Figure 9.1: The product of acceptance and efficiency with statistical uncertainty for the vector-like τ lepton model in the doublet scenario (left) and for the type-III seesaw fermions in the $\mathrm{B_e} = \mathrm{B}_\mu = \mathrm{B}_\tau$ scenario (right) in the signal regions of all seven multilepton channels.

9.2 Measurement of lepton efficiency maps

The lepton reconstruction efficiency maps are obtained from a simulation of the ZZ process. Since the leptons from SM gauge bosons (W, Z, h) and promptly-decaying signal particles share similar properties, these ZZ maps can be used for all those processes. For a given input generator-level p_T, the efficiency map provides the probability distribution of the reconstructed p_T, accounting for reconstruction and identification efficiency, and the p_T resolution.

1. Separate maps are produced for electrons, muons, 1-prong τ_h, and 3-prong τ_h. Further, separate maps are produced for light leptons arising from τ decay and from gauge boson decay.

2. For muons, maps are produced in two regions of pseudorapidity: Barrel ($|\eta| \leq$1.2) and Endcap (1.2$< |\eta| <$2.4). For electrons and τ_h, maps are defined in three regions of pseudorapidity: Barrel ($|\eta| \leq$1.1), Transition (1.1$< |\eta| \leq$1.6), and Endcap ($|\eta| >$1.6), to account for the efficiency losses in the region with overlap between tracker and ECAL.

3. In addition, the topology of the event plays an important role in the efficiency determination. To account for this, we produce each map in two cases.

- N_j maps: N_j is the number of generator level jets passing selection criteria outlined in item 4 above. We produce maps in separate N_j regions ($N_j \leq 1$ and $N_j \geq 2$) for each lepton flavor.
- dRmin maps: We define dRmin as the minimum angular separation (dR) between any pair of selected light leptons at generator level in the event. We produce maps in separate dRmin regions ($0.2 <$dRmin< 0.4 and dRmin> 0.4) for electrons and muons (for τ_{h}, a single map inclusive in dRmin is sufficient).

The N_j maps are useful over dRmin maps for signals where the multilepton final state is usually accompanied by jets from gauge boson decay. For example, we recommend using N_j maps in decays such as for the VLL model $\tau'^{+}\nu'(\tau'^{-}\bar{\nu}') \rightarrow Z\tau W\tau \rightarrow \ell\ell\tau\tau qq$ or the leptoquarks model $SS \rightarrow t\tau t\tau \rightarrow Wb\tau Wb\tau \rightarrow \ell\ell\nu\nu\tau\tau bb$. A different example is for using the dRmin maps in decays such as for the Seesaw model $\Sigma^{\pm}\Sigma^{0} \rightarrow W\nu W\ell \rightarrow \ell\ell\ell\nu\nu\nu$ or models with sterile right handed neutrinos (N) $WN \rightarrow \ell\ell\ell\nu\nu$.

In total, we have 58 lepton reconstruction efficiency maps with the corresponding statistical uncertainty maps. These are all available on the public repository for high energy physics data, or HEPDATA record [168]. A few sample efficiency maps are shown below.

Figure 9.2 shows a few examples of lepton reconstruction efficiency dRmin maps measured from a simulation of the ZZ process with leptons produced from the decay of gauge bosons. The x-axis and the y-axis represents bins in the reconstructed and generated p_{T}, respectively.

Figure 9.3 shows a few examples of lepton reconstruction efficiency N_j maps measured from a simulation of the ZZ process with leptons produced from the decay of gauge bosons. The x-axis and the y-axis represents bins in the reconstructed and generated p_{T}, respectively.

Finally, Figure 9.4 shows a few of examples of light lepton reconstruction efficiency dRmin and N_j maps measured from a simulation of the ZZ process, with leptons produced from the decay of τ lepton. The x-axis and the y-axis represents bins in the reconstructed and generated p_{T}, respectively.

To conclude, all these lepton reconstruction and selection requirements result in typical efficiencies of 40–85%, 65–90%, and 20–50% for electrons, muons, and τ_{h} leptons, respectively, depending on the lepton p_{T}, η, the source of their origin, and the global event topology.

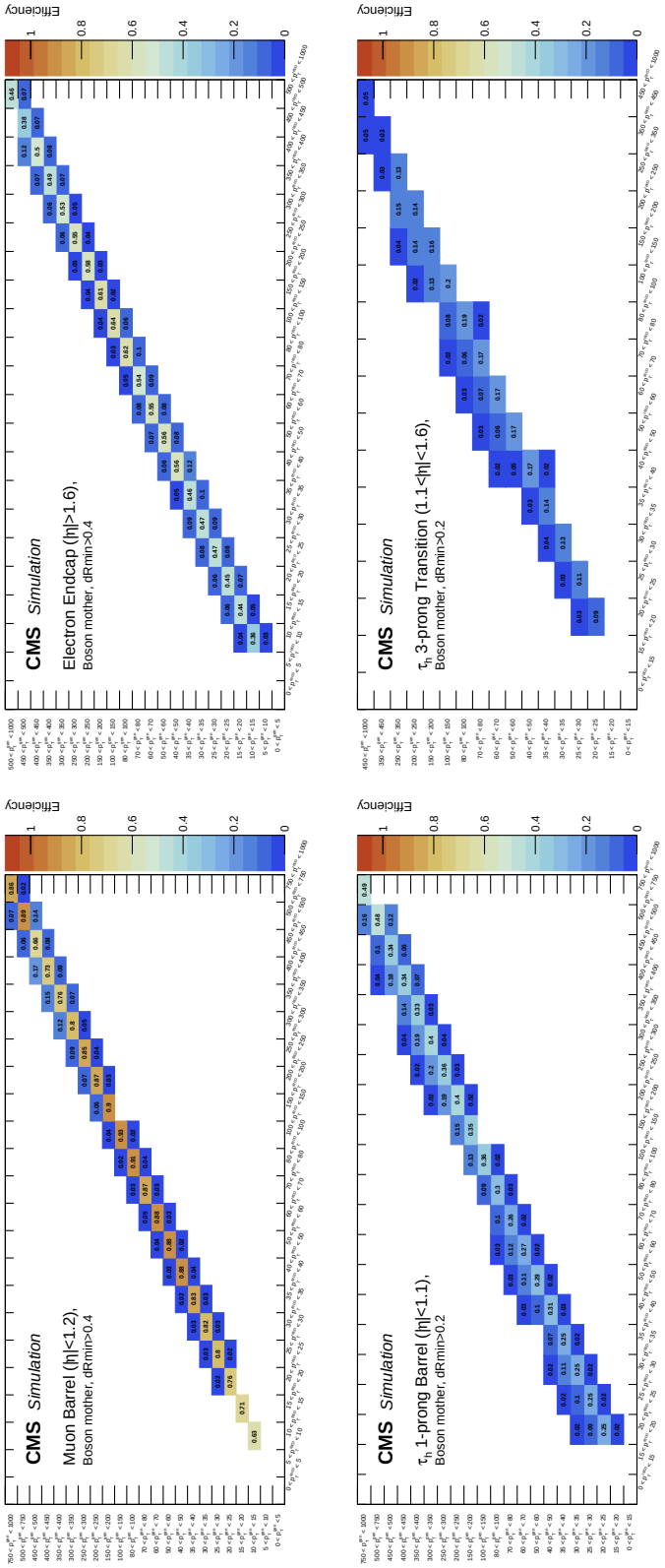

Figure 9.2: Example reconstruction efficiency dRmin maps for barrel muons (upper), endcap electrons (upper middle), 1-prong τ_h in the barrel region (lower middle), and 3-prong τ_h in the transition region (lower). The leptons are produced from the decay of gauge bosons. The lepton efficiency is measured from a simulation of the ZZ process. The x-axis and the y-axis represents bins in the reconstructed and generated p_T, respectively.

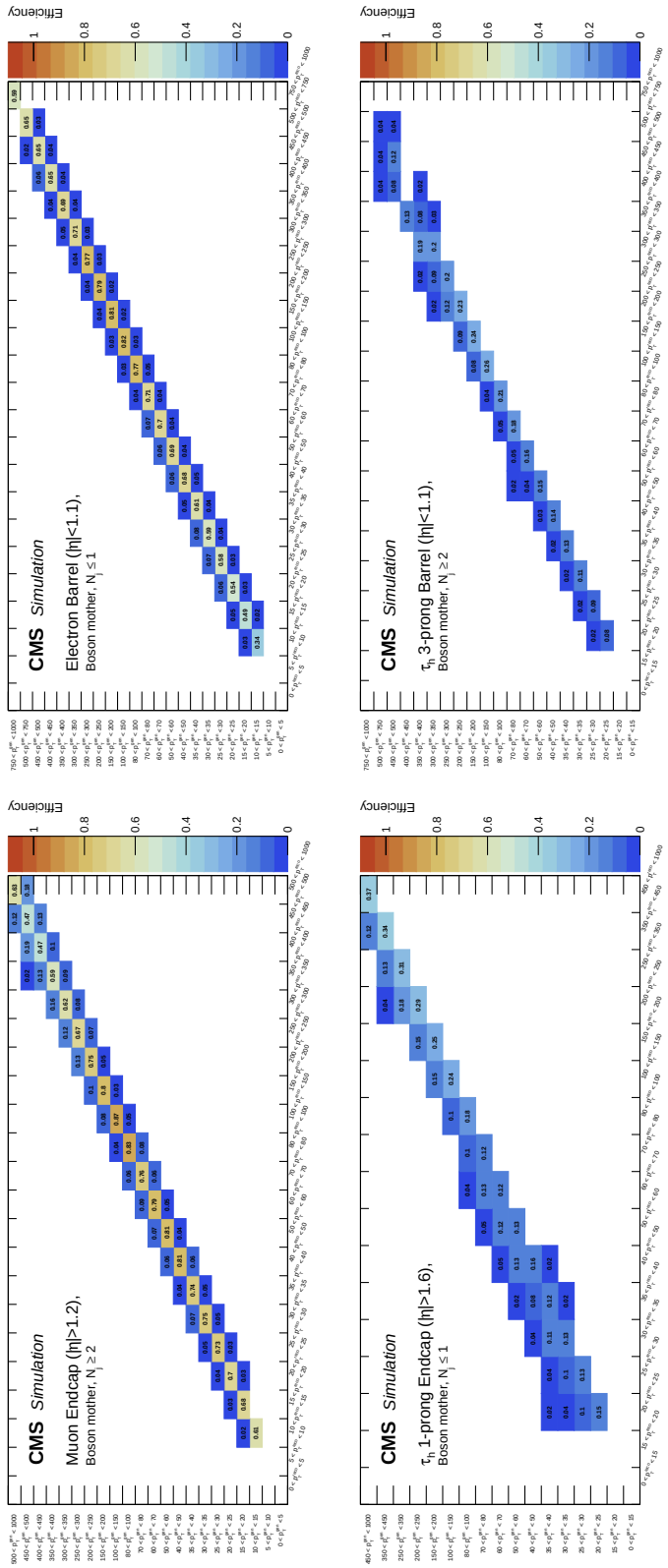

Figure 9.3: Example reconstruction efficiency N_j maps for endcap muons (upper), barrel electrons (upper middle), 1-prong τ_{h} in the endcap region (lower middle), and 3-prong τ_{h} in the barrel region (lower). The leptons are produced from the decay of gauge bosons. The lepton efficiency is measured from a simulation of the ZZ process. The x-axis and the y-axis represents bins in the reconstructed and generated p_{T}, respectively.

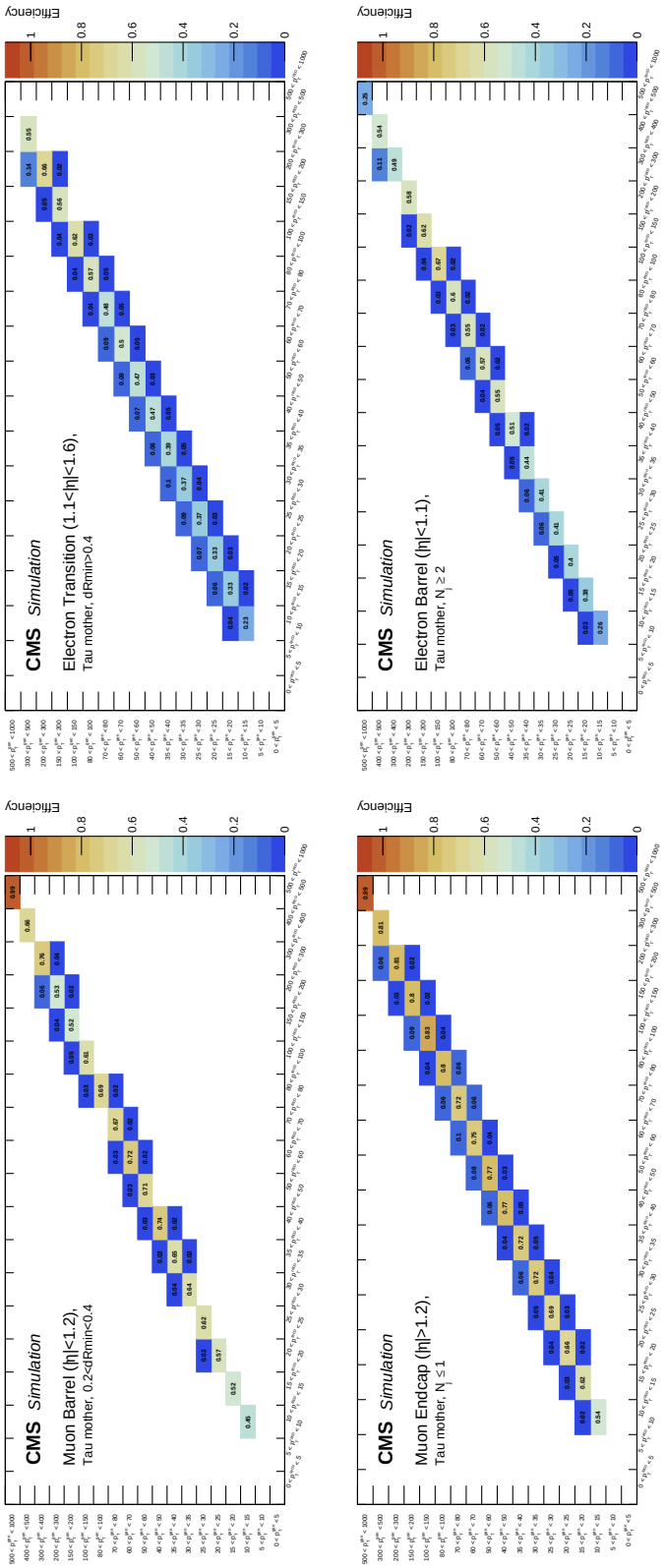

Figure 9.4: Example reconstruction efficiency dRmin maps for barrel muons (upper) and transition electrons (upper middle), and N_j maps for endcap muons (lower middle) and barrel electrons (lower). The light leptons are produced from the decay of τ lepton. The lepton efficiency is measured from a simulation of the ZZ process. The x-axis and the y-axis represents bins in the reconstructed and generated p_{T}, respectively.

9.3 Workflow for deriving yield in signal regions

To calculate the signal yield in the various categories of the model-independent schemes, one should proceed as follows:

1. Leptons should be selected at the generator level passing the following criteria: $p_{\mathrm{T}} > 5$ GeV, $|\eta| < 2.4$ for electrons and muons, and $p_{\mathrm{T}} > 15$ GeV, $|\eta| < 2.3$ for hadronically decaying taus. The leptons must originate from an appropriate source: either the mother is a SM gauge boson (W,Z,H) or a signal particle, or additionally the mother is a τ lepton in case of electrons and muons. Leptons from any other source should be rejected.

2. The provided efficiency maps should be applied to each generator level lepton, thus giving a predicted p_{T} for each lepton ($= -1$ in cases where the lepton is predicted to fail reconstruction/identification). Henceforth, only the predicted p_{T} should be used in subsequent analysis. In the rest of this document, we use p_{T} to refer to the predicted lepton p_{T} after application of efficiency map.

3. The number and flavor of the leptons should now be used to determine the channel in which the event falls. In addition the p_{T} should be used to calculate the L_T. An example of this would be as follows. Suppose at the generator level, an event has two electrons (e_1, e_2), two muons (μ_1, μ_2) and two τ_{h} (τ_{h1}, τ_{h2}).

   - Case 1: suppose the two electrons and two muons pass and have a $p_{\mathrm{T}} > 0$, and say τ_{h1} does as well. This event is thus predicted to be a $e_1 e_2 \mu_1 \mu_2 \tau_{h1}$ event - and thus should be classified as a 4L event.
   - Case 2: suppose e_1 and μ_2 and both τ_{h} have $p_{\mathrm{T}} > 0$. This event is thus predicted to be a $e_1 \mu_1 \tau_{h1} \tau_{h2}$ event - and thus should be classified as a 2L2T event.

   Thus it may also happen that an event which has sufficient leptons at generator level, does not have enough leptons predicted to pass reconstruction to be selected in the analysis.

4. The $p_{\mathrm{T}}^{\mathrm{miss}}$ should be calculated from a 4-vector sum of all neutrinos and other invisible particles specific to the model. The H_T should be calculated from a scalar sum of all generator level jets [jet clustering algorithm implemented at the particle level]. These jets should pass the requirement of $p_{\mathrm{T}} > 30$ GeVand $|\eta| \leq$2.4, with a minimum angular separation of 0.4 against the selected leptons in the event. We recommend no further corrections for the $p_{\mathrm{T}}^{\mathrm{miss}}$ and H_T. The number of reconstructed b-tagged jets (N_b) can be estimated by first selecting

all generator level jets containing at least one b hadron, and then applying the efficiencies measured by the CMS collaboration [152].

5. Any remaining analysis selections can be imposed on the leptons (such as additional p_{T} or invariant mass selections), and along with using $p_{\mathrm{T}}^{\mathrm{miss}}$ and H_T one can thus determine the signal yield in any particular signal region of the analysis.

We find that the procedure described here typically predicts the analysis level yields within 20 to 25% and we recommend a conservative uncertainty of 25% on the yield predicted from this procedure to account for the choices of parameterization for the different lepton flavors and event topologies.

9.4 Closure and tests of yield prediction

In this section, we show the results of closure tests, as well as comparisons of the signal yield obtained using the lepton efficiency map approach with the actual analysis yields.

Figure 9.5 shows a selection of the lepton p_{T} distributions from Seesaw $m_{\Sigma} = 200$ GeV sample in the $\mathrm{B_e} = \mathrm{B}_{\mu} = \mathrm{B}_{\tau}$ scenario, where the blue curve is the generated-level lepton p_{T}, red curve is the prediction-level lepton p_{T} derived after applying the correct efficiency map on the generated-level lepton p_{T}, and the green curve is the actual reconstructed-level lepton p_{T}. The red dots in the ratio panel of the distributions is a representation of the measured lepton efficiency, and is calculated as the ratio of predicted p_{T} and generated p_{T}. The green dots are the ratio between prediction-level lepton p_{T} and reconstructed-level lepton p_{T} which is found to be close to unity in most of the cases across the p_{T} range.

Next, we demonstrate the reproducibility of the kinematic variables in the multilepton channels using the lepton efficiency maps on different signals. For the purposes of this exercise, we show the agreement in the L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ distribution, where $p_{\mathrm{T}}^{\mathrm{miss}}$ is the generated-level quantity calculated as the p_{T} of the resultant of the vector sum of the neutrinos in the event. Note that the lepton efficiency maps work in signals with intrinsic $p_{\mathrm{T}}^{\mathrm{miss}}$ in the event, or in signals with no more than 10% neutrino-less events.

Figure 9.6 shows the L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ distribution in the four lepton channels: 4L (left) and 3L1T (right) for Seesaw $m_{\Sigma} = 850$ GeV sample in the $\mathrm{B_e} = \mathrm{B}_{\mu} = \mathrm{B}_{\tau}$ scenario. The hatched region in the ratio panel is the flat 25% uncertainty band which mostly encapsulates the overall agreement between the predicted and the reconstructed L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ distribution.

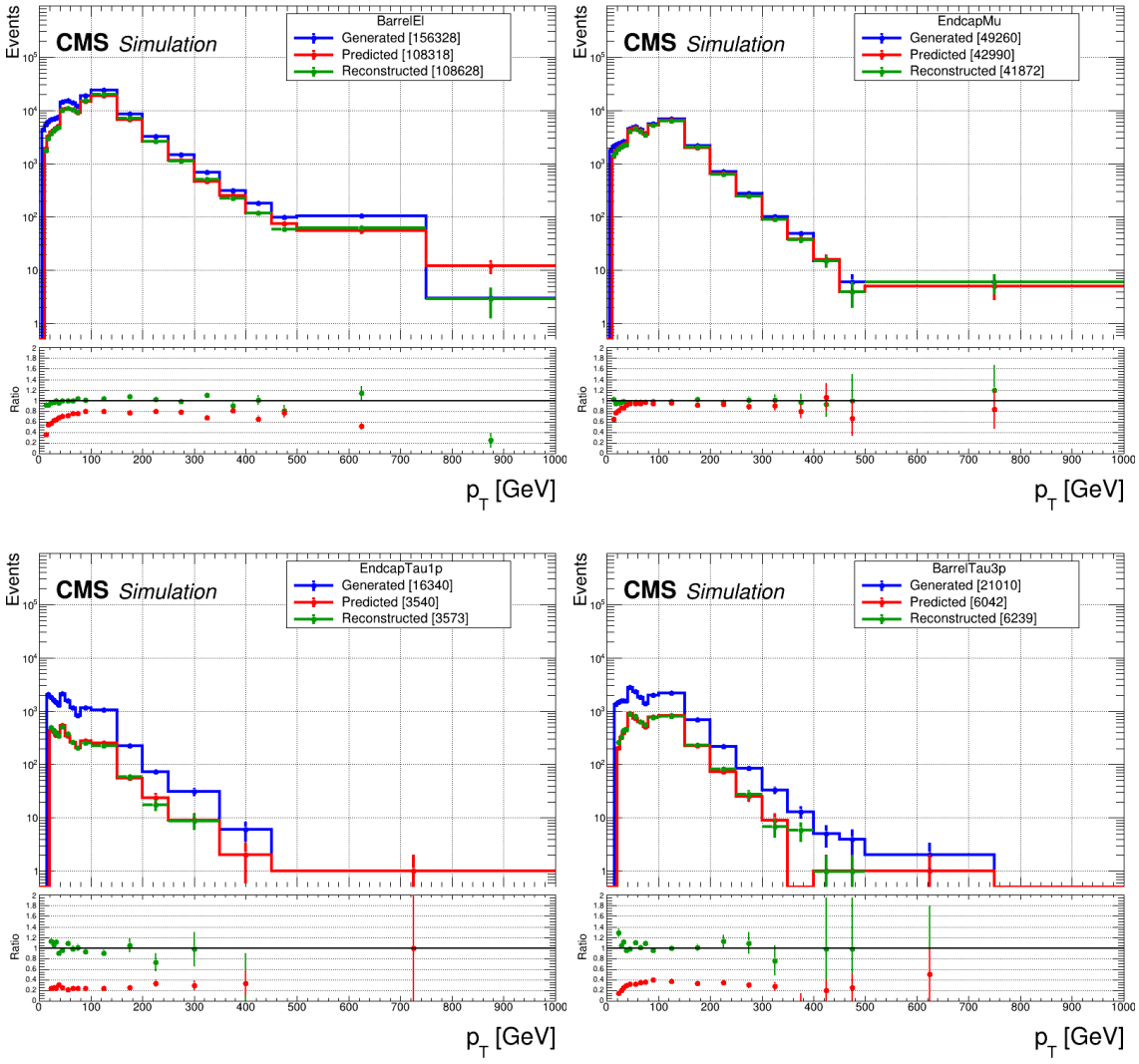

Figure 9.5: Lepton p_{T} distributions from Seesaw $m_{\Sigma} = 200$ GeV sample in the $\mathrm{B_e} = \mathrm{B}_{\mu} = \mathrm{B}_{\tau}$ scenario for barrel electrons (upper-left), endcap muons (upper-right), endcap 1-prong τ_{h} (lower-left), and barrel 3-prong τ_{h} (lower-right). The blue curve is the generated-level lepton p_{T}, red curve is the prediction-level lepton p_{T}, and the green curve is the actual reconstructed-level lepton p_{T}. The red dots in the ratio panel is the ratio of predicted p_{T} and generated p_{T}. The green dots are the ratio between prediction-level lepton p_{T} and reconstructed-level lepton p_{T}.

Figure 9.7 shows the L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ distribution in the three lepton channels: 3L (left) and 2L1T (right) for Leptoquarks $m_S = 400$ GeV sample coupled to a top quark and a τ lepton. The hatched region in the ratio panel is the flat 25% uncertainty band which mostly encapsulates the overall agreement between the predicted and the reconstructed L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ distribution.

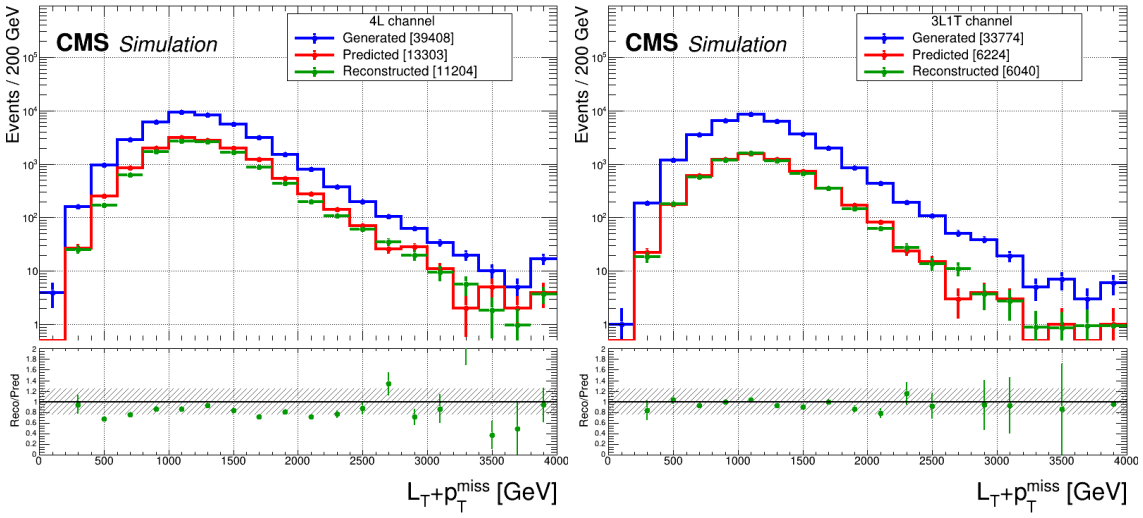

Figure 9.6: L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ distributions from Seesaw $m_{\Sigma} = 850$ GeV sample in the $\mathrm{B_e} = \mathrm{B}_{\mu} = \mathrm{B}_{\tau}$ scenario in 4L (left) and 3L1T (right) channels. The blue, red, and green curves are the generated-level, predicted-level, and the actual reconstructed-level L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ distributions, respectively. The green dots in the ratio panel is the ratio of predicted-level and reconstructed-level L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$. The hatched region in the ratio panel is the flat 25% uncertainty band.

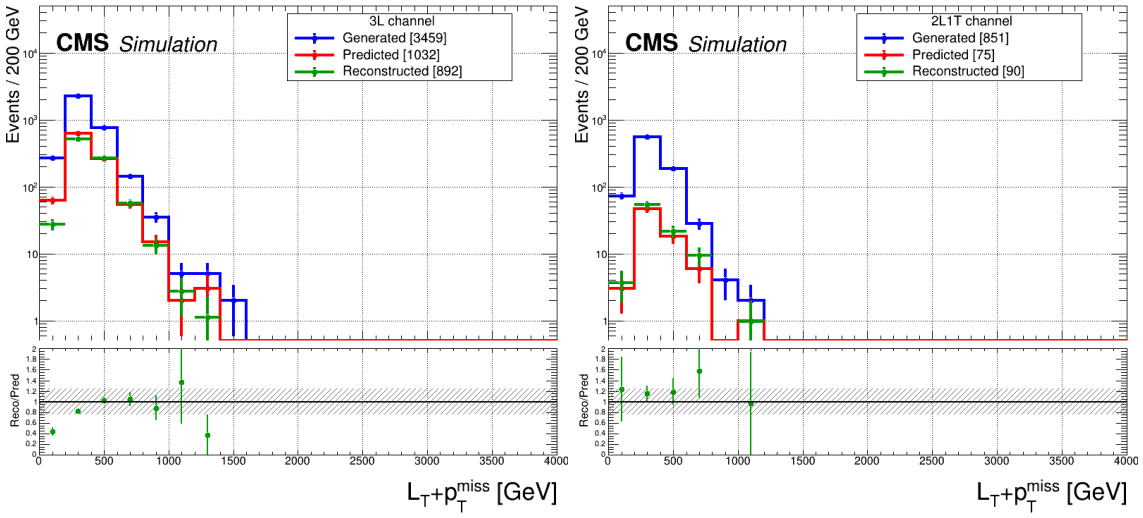

Figure 9.7: L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ distributions from Leptoquarks $m_S = 400$ GeV sample coupled to a top quark and a τ lepton in 3L (left) and 2L1T (right) channels. The blue, red, and green curves are the generated-level, predicted-level, and the actual reconstructed-level L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ distributions, respectively. The green dots in the ratio panel is the ratio of predicted-level and reconstructed-level L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$. The hatched region in the ratio panel is the flat 25% uncertainty band.

Figure 9.8 illustrates the importance of choosing the correct global paramaterization for the lepton efficiencies. For example, the dominant production and decay chain for Seesaw in the three lepton channels is $\Sigma^{\pm}\Sigma^{0} \rightarrow W^{\pm}\nu W^{\pm}\ell^{\mp} \rightarrow \ell^{\pm}\nu\nu\ell^{\pm}\nu\ell^{\mp}$ i.e. not primarily accompanied by jets. Hence, the dRmin-based efficiency parameterization works much better than the N_j-based parameterization, as shown in the upper panel of Figure 9.8 for Seesaw $m_{\Sigma} = 200$ GeV sample in the $\mathrm{B_e} = \mathrm{B}_{\mu} = \mathrm{B}_{\tau}$ scenario in 2L1T channel. Another example is the vector-like leptons which is primarily accompanied with jets via a production and decay chain such as $\tau'^{+}\nu'(\tau'^{-}\overline{\nu}') \rightarrow Z\tau^{\pm}W^{\pm}\tau^{\mp} \rightarrow \ell^{\pm}\ell^{\mp}\tau^{\pm}qq\tau^{\mp}$. Hence, the N_j-based efficiency parameterization works much better than the dRmin-based parameterization, as shown in the lower panel of Figure 9.8 for Vector-like lepton $m_{\tau'} = 900$ GeV sample in the doublet scenario in 4L channel.

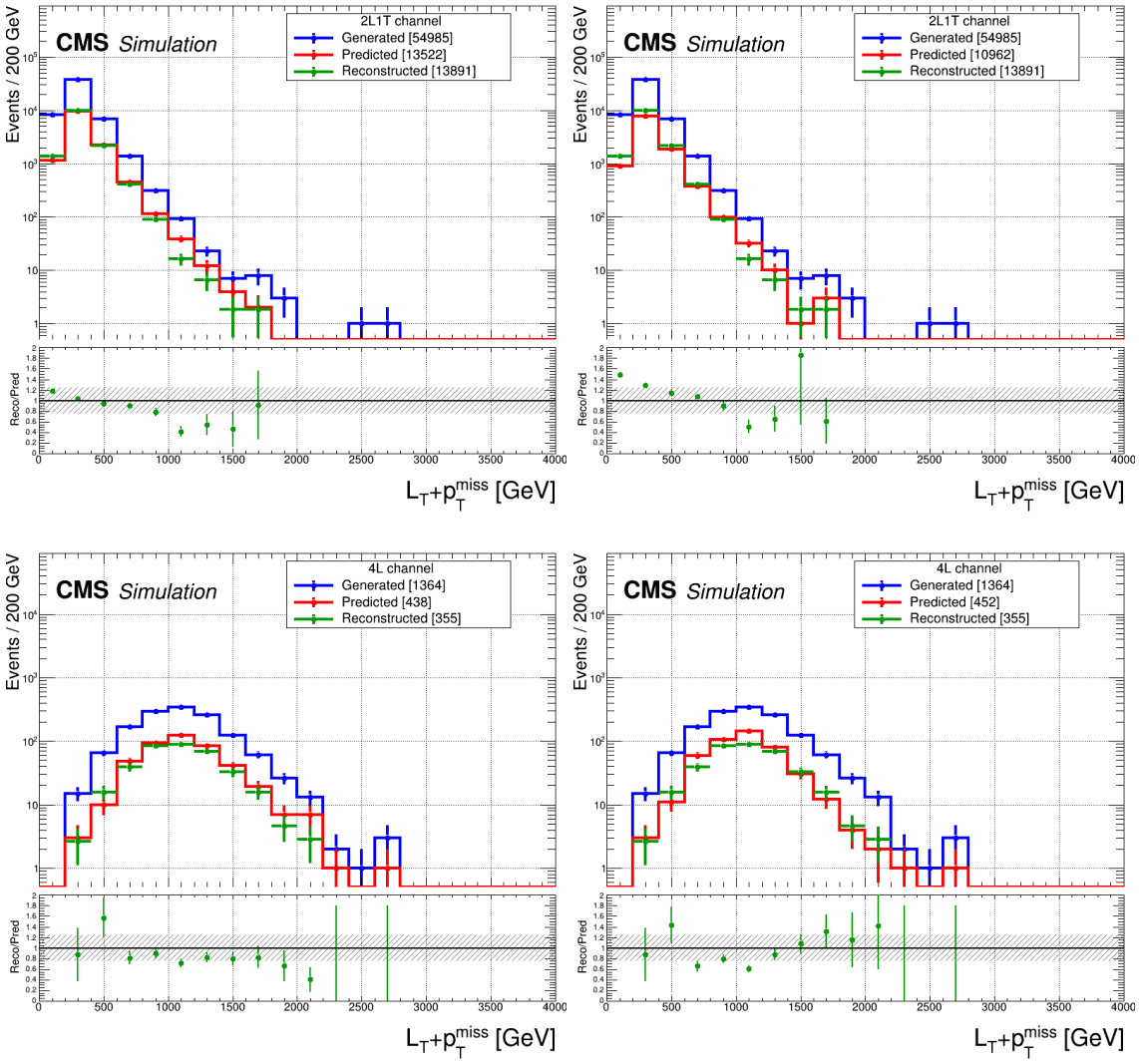

Figure 9.8: L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ distributions from Seesaw $m_{\Sigma} = 200$ GeV sample in the $\mathrm{B_e} = \mathrm{B}_{\mu} = \mathrm{B}_{\tau}$ scenario in 2L1T channel (upper) and Vector-like lepton $m_{\tau'} = 900$ GeV sample in the doublet scenario in 4L channel (lower). For Seesaw, L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ distribution with dRmin-based parametrization (upper-left) works better than the N_j-based parametrization (upper-right) in the three lepton channels. For Vector-like leptons, L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ distribution with N_j-based parametrization (lower-left) works better than the dRmin-based parametrization (lower-right) in the four lepton channels. The blue, red, and green curves are the generated-level, predicted-level, and the actual reconstructed-level L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ distributions, respectively. The green dots in the ratio panel is the ratio of predicted-level and reconstructed-level L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$. The hatched region in the ratio panel is the flat 25% uncertainty band.

Finally, we have also checked that the ZZ efficiencies work on other SM processes such as WZ as shown in Figure 9.9 for 3L (left) and 2L1T (right) channels.

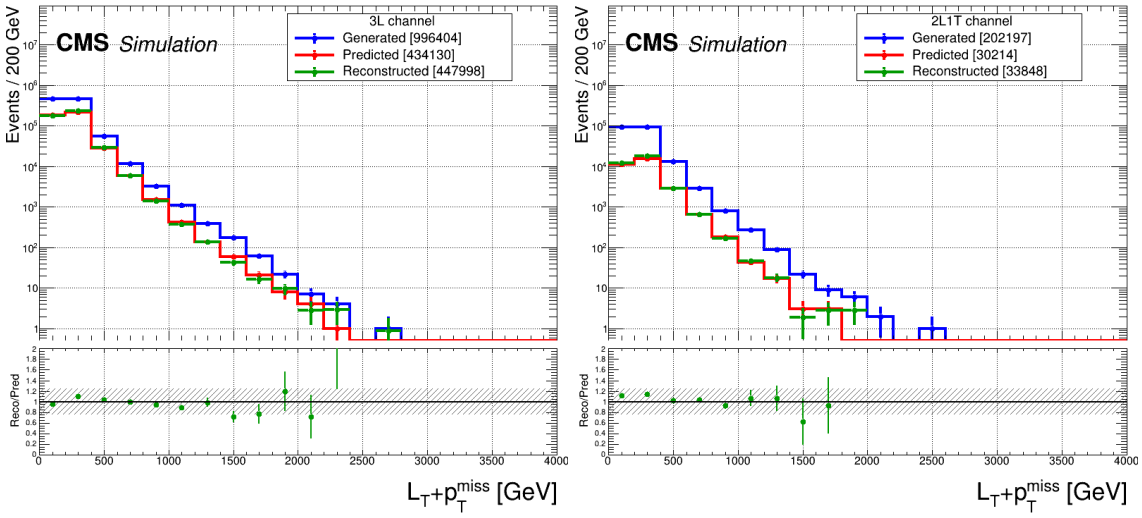

Figure 9.9: L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ distributions from WZ sample in 3L (left) and 2L1T (right) channels. The blue, red, and green curves are the generated-level, predicted-level, and the actual reconstructed-level L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$ distributions, respectively. The green dots in the ratio panel is the ratio of predicted-level and reconstructed-level L_{T}+$p_{\mathrm{T}}^{\mathrm{miss}}$. The hatched region in the ratio panel is the flat 25% uncertainty band.

Chapter 10

Summary

This thesis presents a search for inclusive nonresonant multilepton probes of new phenomena beyond the Standard Model (SM) using proton-proton collisions data at $\sqrt{s} = 13$ TeV, collected in 2016–2018 by the CMS experiment at the LHC, corresponding to an integrated luminosity of 138 fb^{-1}. The search is carried out in seven orthogonal multilepton final states characterized according to the number of light leptons, i.e. electrons and muons, and hadronically decaying tau leptons. For the first time in LHC, tau-enriched channels with up to a multiplicity of three are designed to carry out a coherent search for beyond-the-SM (BSM) phenomena, with high sensitivity to the models coupling primarily to the third generation of leptons.

Three scenarios of BSM phenomena are probed: type-III seesaw mechanism, vector-like lepton in the doublet and singlet extensions of the SM, and scalar leptoquarks with top-philic couplings. These three models target different open questions of the SM, such as a potential dark matter candidate and an explanation for the mass hierarchy among the three generations by vector-like leptons, the smallness of the neutrino masses by the seesaw, and an explanation for the observed b-anomalies by the leptoquarks. The primary reason behind this particular selection of BSM models is that they are generators of complementary nonresonant multilepton signatures.

The analysis employs boosted decision trees algorithm to enhance the sensitivity for each of the probed BSM scenarios. No significant deviations from the background expectations are observed in any signal regions. To obtain upper limits at 95% confidence level (CL) on the production cross section of the probed models, a modified frequentist approach is used with a test statistic based on the profile likelihood in the asymptotic approximation and the CLs criterion.

In the vector-like lepton doublet model, vector-like τ leptons are excluded at 95% CL with masses below 1045 GeV, with an expected exclusion of 975 GeV. These are the most stringent constraints on the doublet model. For the singlet model, vector-like τ leptons are excluded from

125 to 150 GeV, while the expected exclusion range is from 125 to 170 GeV. These are the first constraints from the LHC on the singlet model.

Type-III seesaw heavy fermions are excluded at 95% confidence level (CL) with masses below 980 GeV(expected 1060 GeV), assuming flavor-democratic mixings with SM leptons, and below 990 GeV(expected 1065 GeV), 1065 GeV (expected 1140 GeV), and 890 GeV(expected 880 GeV), assuming mixings exclusively with electron, muon, and τ lepton flavors, respectively. Lower limits on the masses of the heavy fermions are also presented for various decay branching fractions of the heavy fermions to the different SM lepton flavors. These are the most stringent constraints on the type-III seesaw heavy fermions to date.

Scalar leptoquarks coupled to top quarks and individual lepton flavors are also probed. In the scenario with the leptoquark coupling to a top quark and a τ lepton, leptoquarks with masses below 1120 GeV are excluded at 95% CL (expected 1235 GeV). For the decay to a top quark and an electron, leptoquarks are excluded with masses below 1340 GeV (expected 1370 GeV), and for the decay into a top quark and a muon, masses below 1420 GeV (expected 1460 GeV) are excluded.

To ensure the longevity of this multilepton analysis, a model-independent component based purely on the expected SM predictions and observations is also performed, allowing the results to be reinterpretable for other BSM theories. Detailed results are also provided to facilitate these alternative theoretical interpretations. This includes detailed efficiency maps for electrons, muons, and τ_{h}, where the provided efficiency is that for a generator level lepton to be both reconstructed and identified as described in this analysis. In addition, the product of acceptance and efficiency for each probed signal model in this thesis are also provided, for a quick back-of-the-envelope calculation of selection efficiency for other BSM scenarios. Finally, the obtained BSM model yields in the various categories can then be used along with the SM backgrounds, the background covariance matrix, and the observations, also provided as part of the detailed results, to arrive at constraints for the model in the simplified likelihood framework.

APPENDICES

Appendix A

Trigger and lepton efficiency

A.1 Single isolated muon trigger efficiency

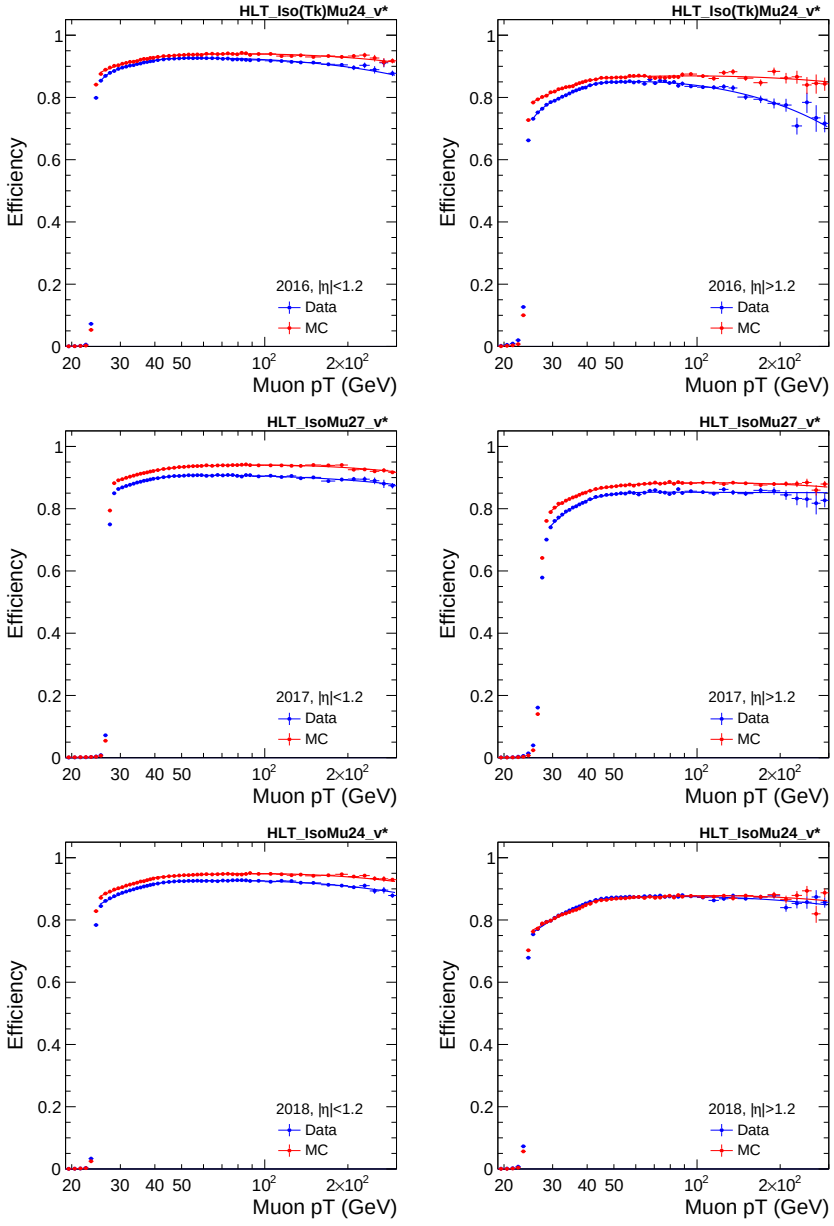

Figure A.1: Single isolated muon trigger efficiencies in 2016 (upper row), 2017 (middle row), and 2018 (lower row) as measured by the tag-and-probe method in Z$\rightarrow \mu\mu$ enriched data and DY MC samples, and described by a 7-parameter ad-hoc, continuous fit.

A.2 Single isolated electron trigger efficiency

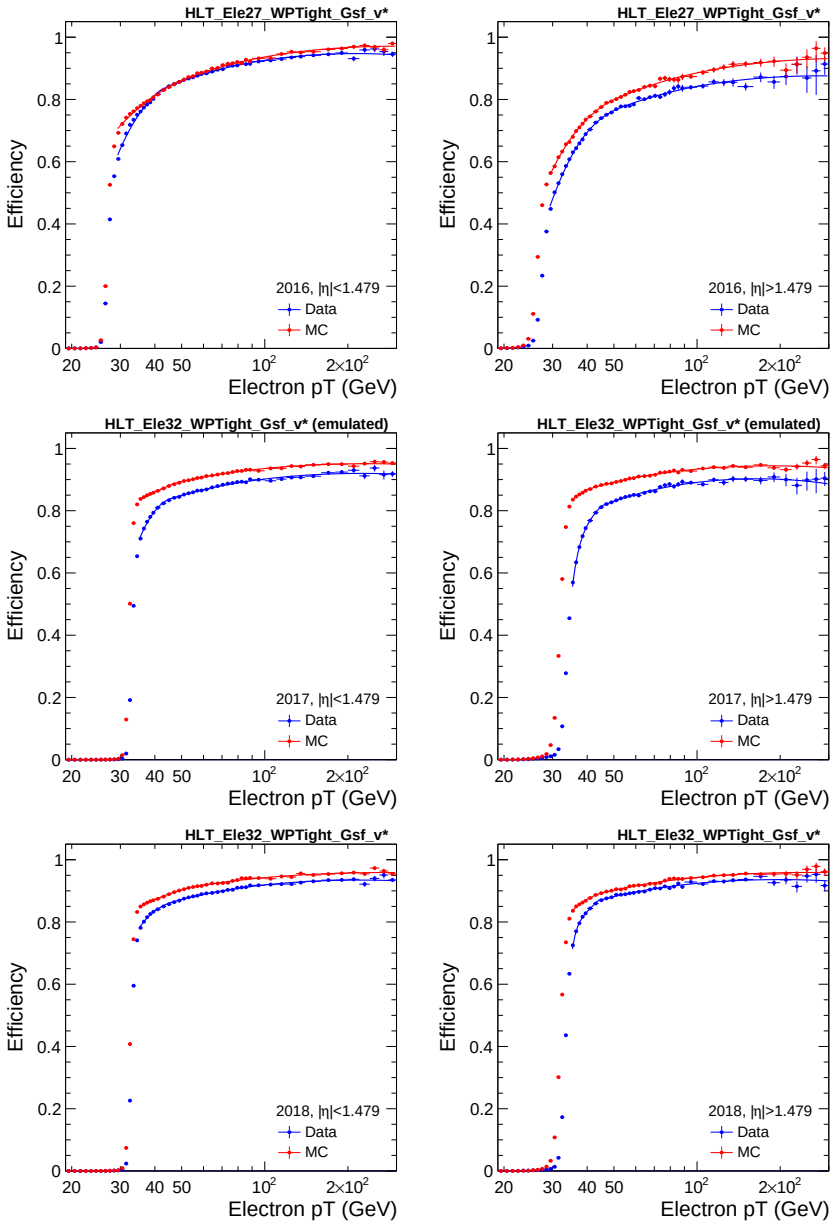

Figure A.2: Single isolated electron trigger efficiencies in 2016 (upper row), 2017 (middle row), and 2018 (lower row) as measured by the tag-and-probe method in Z→ee enriched data and DY MC samples, and described by a 7-parameter ad-hoc, continuous fit.

A.3 Muon custom identification efficiency

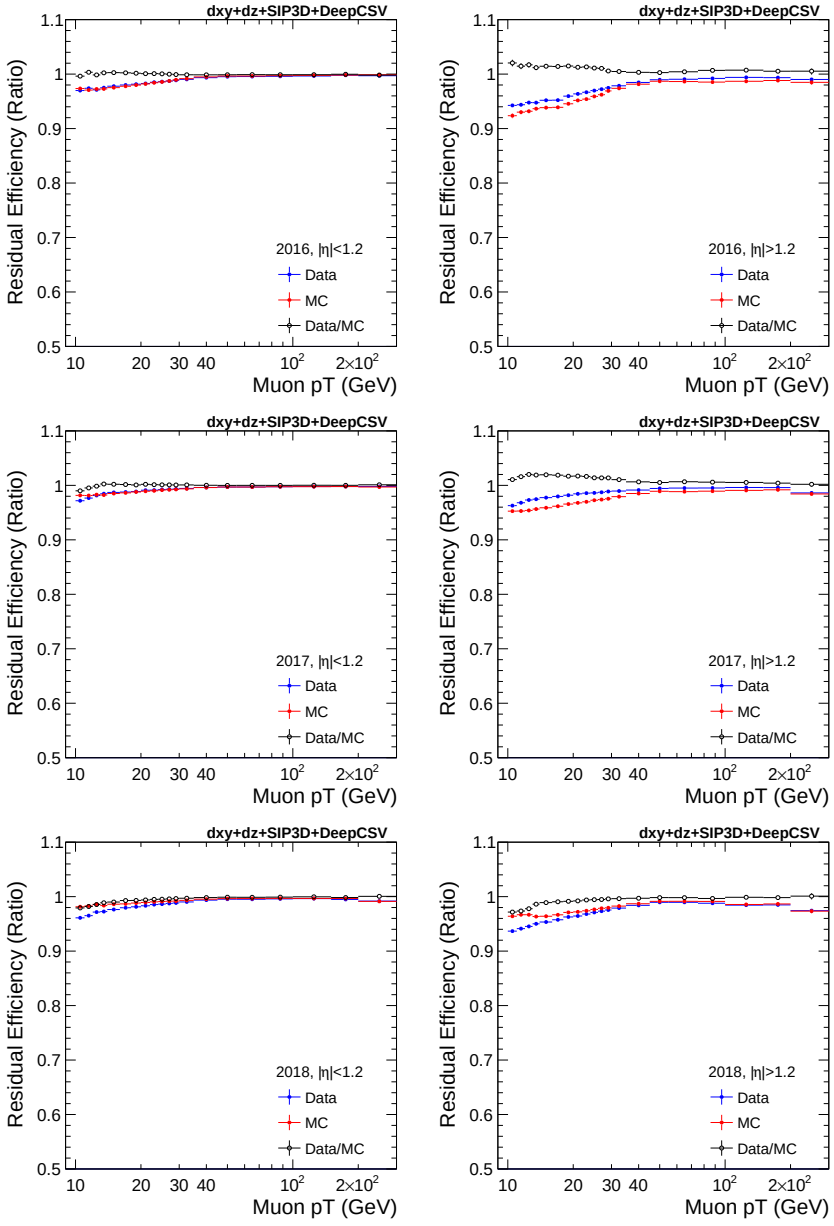

Figure A.3: Custom muon ID efficiency and scale factor (data/MC) in 2016 (upper row), 2017 (middle row), and 2018 (lower row) as measured by the tag-and-probe method in Z$\rightarrow \mu\mu$ enriched OSSF 2L events in data and DY MC samples. The custom ID requirements refer to the d_{xy}, d_{z}, $\mathrm{SIP}_{\mathrm{3D}}$, and DeepCSV criteria applied to muons that already satisfy the medium working point of the cut-based muon ID and the tight working point of the PF-based relative isolation criteria.

A.4 Electron custom identification efficiency

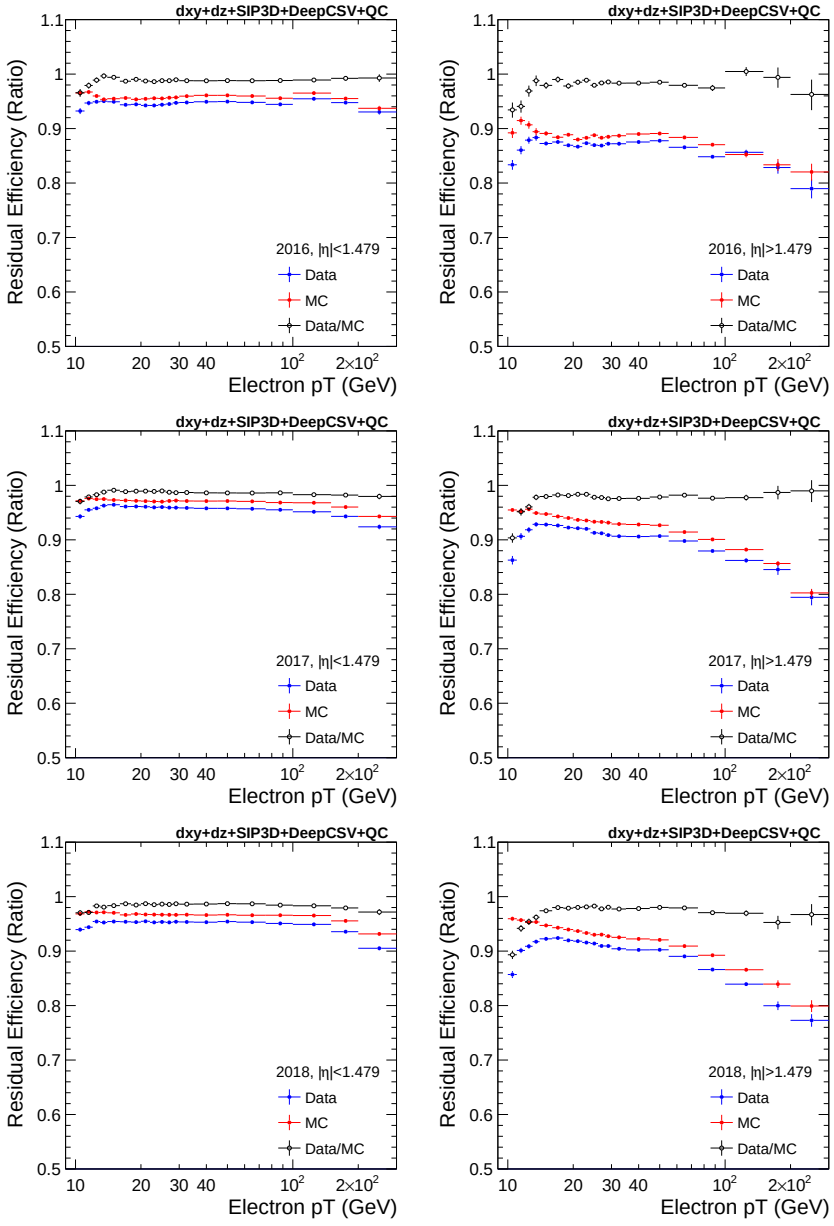

Figure A.4: Custom electron ID efficiency and scale factor (data/MC) in 2016 (upper row), 2017 (middle row), and 2018 (lower row) as measured by the tag-and-probe method in Z→ee enriched OSSF 2L events in data and DY MC samples. The custom ID requirements refer to the d_{xy}, d_z, SIP_{3D}, DeepCSV, and charge consistency (QC) criteria applied to electrons that already satisfy the medium working point of the cut-based electron ID.

A.5 Tau custom identification efficiency

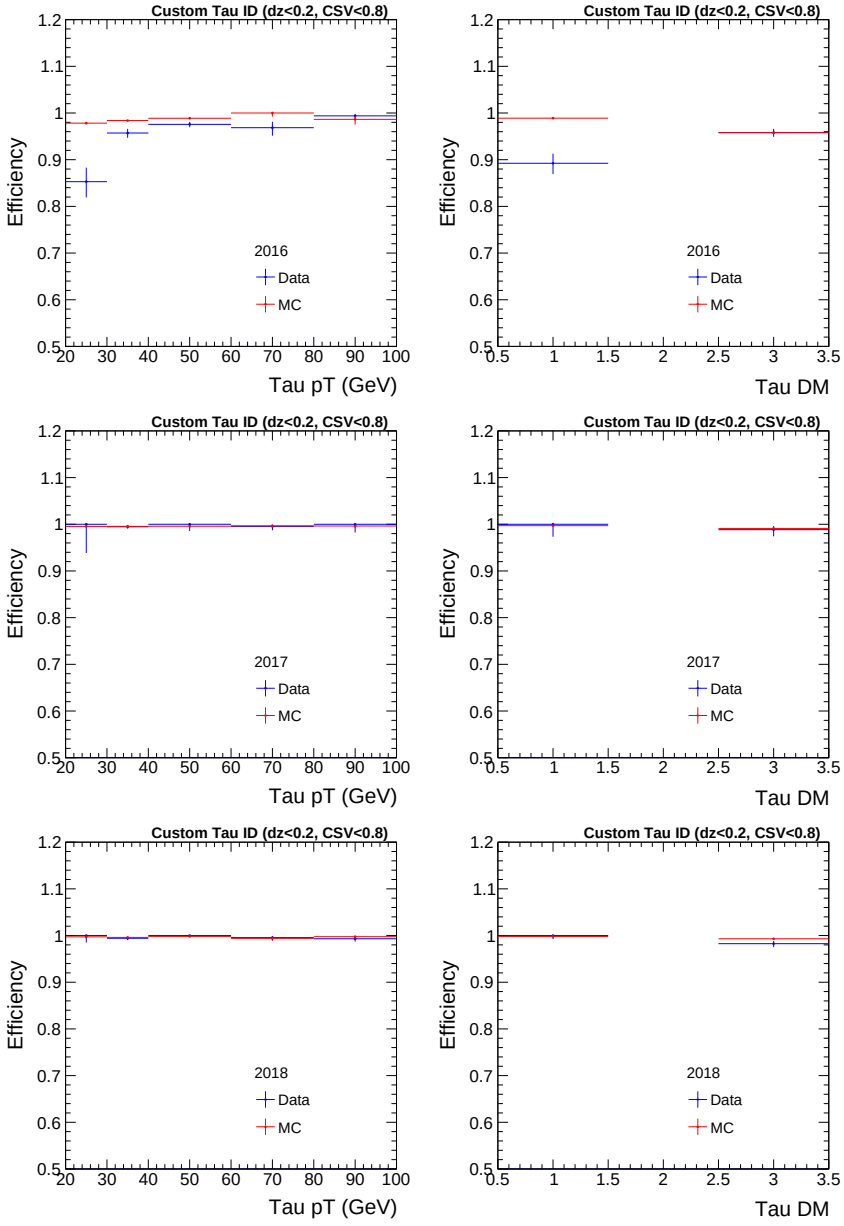

Figure A.5: Custom tau ID efficiency in 2016 (upper left), 2017 (upper right), and 2018 (lower left) as measured by the tag-and-probe method in Z$\rightarrow \tau\tau$ enriched OSOF 1L1T events in data and DY MC samples. The custom ID requirements refer to the d_z and DeepCSV criteria applied to taus that already satisfy the byVTightDeepTau2017v2p1VSjet, byLooseDeepTau2017v2p1VSe, and byLooseDeepTau2017v2p1VSmu discriminators of the DeepTau ID.

Appendix B

Dilepton control regions

B.1 DY 2LOS control region

A set of events enriched in Z(→ee)+jets and Z(→ $\mu\mu$)+jets processes are created in a 2L OnZ selection. The contributions due to fake leptons is estimated from MC samples, where at least one reconstructed lepton is not matched ($\Delta R > 0.2$) to a generator level prompt lepton. The NLO MADGRAPH5_AMC@NLO DY MC sample yield is normalized to the observed data, and the Z-p_{T} shape, as well as the jet multiplicity of the MC sample is corrected to agree with that of the data shape using the events in the dimuon channel. We observe that this Z-p_{T} correction is also valid in the dielectron channel, and therefore conclude that this is a deficiency of the MC sample.

The distributions of key kinematic and event variables in the Z(→ $\mu\mu$) and Z(→ ee) control region are given in Figures B.1 and B.2, respectively. Good agreement with respect to predictions is observed across all other variables of interest.

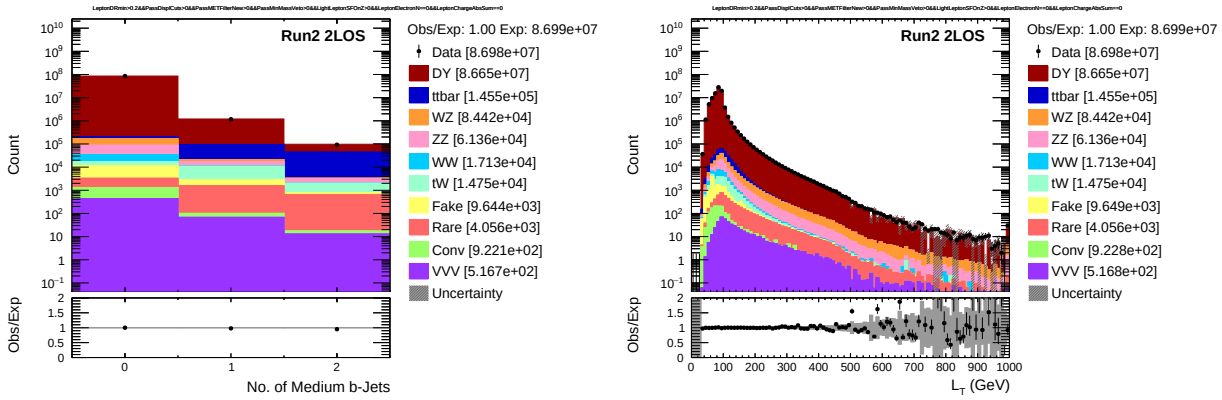

Figure B.1: The distributions of number of b-tagged jets (left) and L_{T} (right) in 2L Z($\rightarrow \mu\mu$) control region in Run2. Only statistical uncertainties are shown.

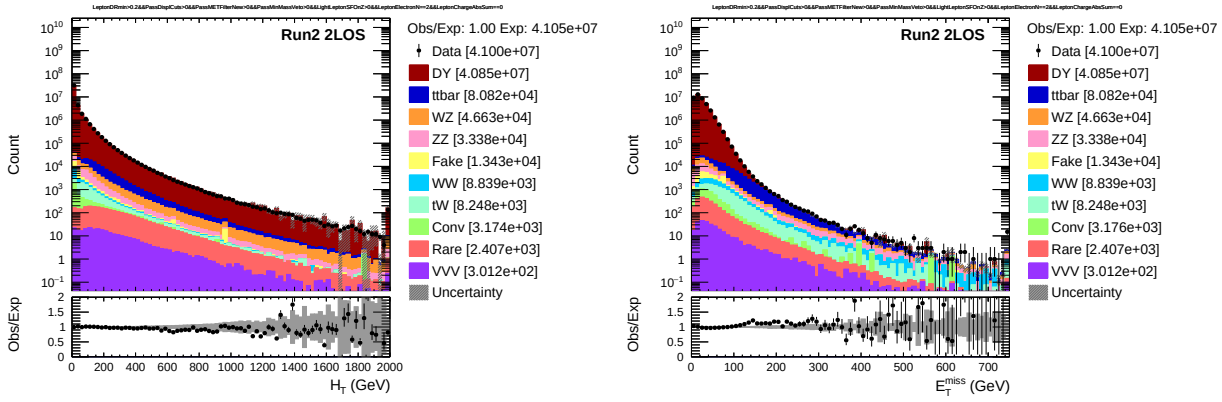

Figure B.2: The distributions of H_{T} (left) and $p_{\mathrm{T}}^{\mathrm{miss}}$ (right) in 2L Z($\rightarrow$ee) control region in Run2. Only statistical uncertainties are shown.

B.2 DY 1L1T control region

A set of events enriched in prompt taus originating from $Z \to \tau\tau$ decays is created by an opposite-sign 1L1T selection, where events are required to have $M_{\mathrm{T}} < 40$ GeV computed with the light lepton and $p_{\mathrm{T}}^{\mathrm{miss}}$, $\Delta R < 3.5$ between the lepton pair, $p_{\mathrm{T}}^{\mathrm{miss}} < 100$ GeV, and a dilepton mass of 40-120 GeV. The triggering light lepton is originating from a leptonic decay of one of the prompt taus coming from the Z boson.

A 2D implementation of the matrix method, as described in Section 6.2.1, is used to estimate the contributions due to fake leptons. The normalization and Z-p_{T} correction of the NLO MADGRAPH5_AMC@NLO DY MC are taken from the higher purity 2L control regions and are observed to be valid in the 1L1T channel as well. The distributions of key kinematic and event variables in the $Z \to \tau\tau$ control region are given in Figure B.3 illustrating overall good agreement between observations and the predictions.

The prompt τ_{h} contributions originating from $Z \to \tau\tau$ decays correspond to the mass peak approximately below 80 GeV, whereas those above 80 GeV are mostly originating from $Z \to \ell\ell$ contributions where a light lepton is then misidentified as a hadronic tau.

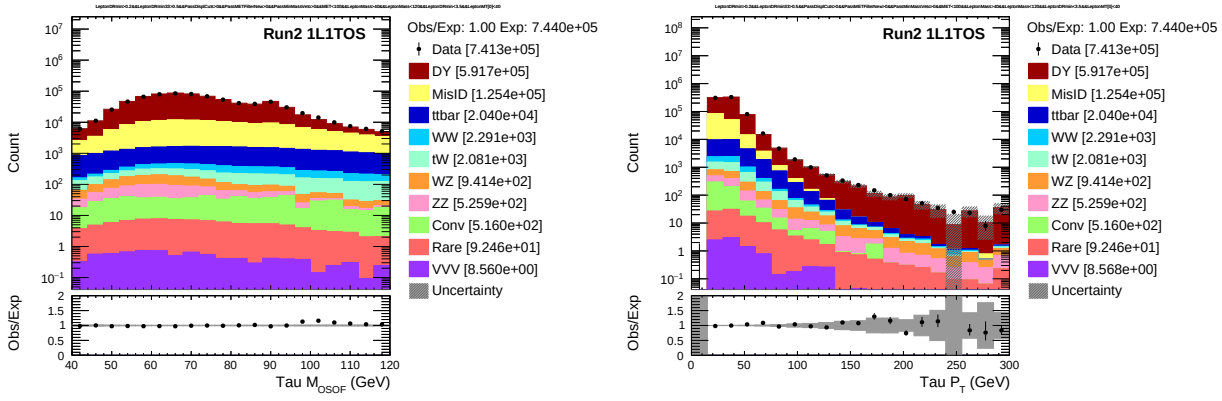

Figure B.3: The distributions of invariant mass of the opposite-sign light lepton and tau pair (left) and τ_{h} p_{T} (right) in 1L1T $Z \to \tau\tau$ control region in Run2. Only statistical uncertainties are shown.

B.3 $t\bar{t}$ control region

A set of events enriched in $t\bar{t}$ process is created in the 2L channel. A selection with an opposite-sign eμ pair and $N_{\mathrm{j}} > 1$ as the main control region is defined. In addition, a separate region enriched with OffZ opposite-sign same flavor pairs, $p_{\mathrm{T}}^{\mathrm{miss}} > 50$ GeV, $N_{\mathrm{b}} > 0$ and $N_{\mathrm{j}} > 2$ is defined.

A 2D implementation of the matrix method is used to estimate the contributions due to fake leptons. The NLO POWHEG $t\bar{t}$ MC sample is normalized to the observed data in the main opposite-sign eμ control region. The distributions of key kinematic and event variables in the $t\bar{t}$ control region are given in Figure B.4 illustrating overall good agreement between observations and the predictions.

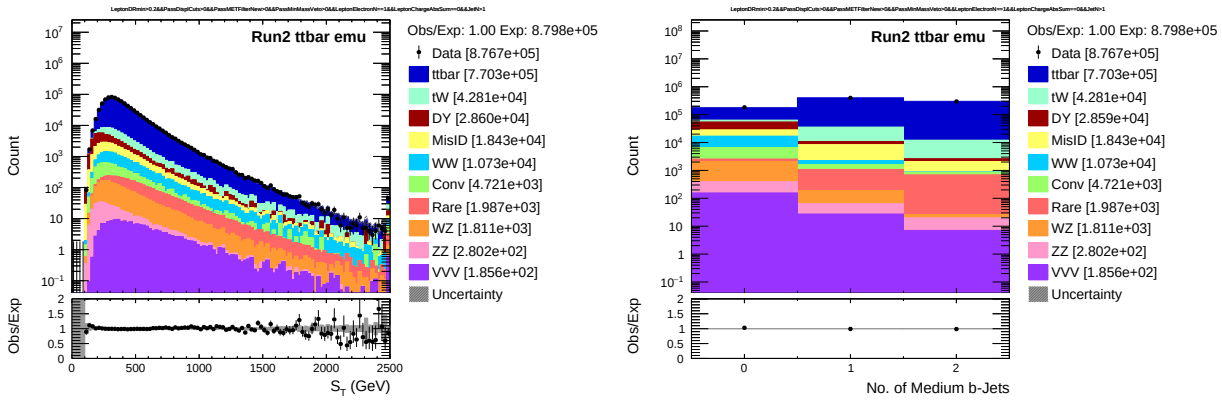

Figure B.4: 2L $t\bar{t}$ control region in Run2. Statistical uncertainties only.

References

[1] Planck Collaboration. "*Planck 2013 results. I. Overview of products and scientific results*" (2014). DOI: 10.1051/0004-6361/201321529.

[2] N. Jarosik et al. "*Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Sky maps, systematic errors, and basic results*" (2011). DOI: 10.1088/0067-0049/192/2/14. URL: https://doi.org/10.1088/0067-0049/192/2/14.

[3] Particle Data Group Collaboration. "*Review of Particle Physics*" (2020). DOI: 10.1093/ptep/ptaa104.

[4] Super Kamiokande Collaboration. "*Evidence for oscillation of atmospheric neutrinos*" (1998). DOI: 10.1103/PhysRevLett.81.1562.

[5] Motoi Endo et al. "*Higgs Mass and Muon Anomalous Magnetic Moment in Supersymmetric Models with Vector-like Matters*" (2011). DOI: 10.1103/PhysRevD.84.075017.

[6] Radovan Dermíšek and Aditi Raval. "*Explanation of the Muon g-2 Anomaly with Vector-like Leptons and its Implications for Higgs Decays*" (2013). DOI: 10.1103/PhysRevD.88.013017.

[7] Eugenio Megias, Mariano Quiros, and Lindber Salas. "*Muon g-2 from Vector-Like Leptons in Warped Space*" (2017). DOI: 10.1007/JHEP05(2017)016.

[8] Junichiro Kawamura, Stuart Raby, and Andreas Trautner. "*Complete vectorlike fourth family and new $U(1)'$ for muon anomalies*" (2019). DOI: 10.1103/PhysRevD.100.055030.

[9] Gudrun Hiller et al. "*Model Building from Asymptotic Safety with Higgs and Flavor Portals*" (2020). DOI: 10.1103/PhysRevD.102.095023.

[10] Muon g-2 Collaboration. "*Final Report of the Muon E821 Anomalous Magnetic Moment Measurement at BNL*" (2006). DOI: 10.1103/PhysRevD.73.072003.

[11] Muon g-2 Collaboration. "*Measurement of the Positive Muon Anomalous Magnetic Moment to 0.46 ppm*" (2021). DOI: 10.1103/PhysRevLett.126.141801.

[12] BaBar Collaboration. "*Measurement of an Excess of $\bar{B} \to D^{(*)}\tau^-\bar{\nu}_\tau$ Decays and Implications for Charged Higgs Bosons*" (2013). DOI: 10.1103/PhysRevD.88.072012.

[13] Belle Collaboration. "*Measurement of $\mathcal{R}(D)$ and $\mathcal{R}(D^*)$ with a semileptonic tagging method*" (2020). DOI: 10.1103/PhysRevLett.124.161803.

[14] LHCb Collaboration. "*Measurement of the ratio of branching fractions $\mathcal{B}(B_c^+ \to J/\psi\tau^+\nu_\tau)$/$\mathcal{B}(B_c^+ \to J/\psi\mu^+\nu_\mu)$*" (2018). DOI: 10.1103/PhysRevLett.120.121801.

[15] LHCb Collaboration. "*Measurement of the ratio of the $B^0 \to D^{*-}\tau^+\nu_\tau$ and $B^0 \to D^{*-}\mu^+\nu_\mu$ branching fractions using three-prong τ-lepton decays*" (2018). DOI: 10.1103/PhysRevLett.120.171802.

[16] LHCb Collaboration. "*Test of lepton universality in beauty-quark decays*" ((2021)). arXiv: 2103.11769 [hep-ex].

[17] Belle Collaboration. "*Lepton-Flavor-Dependent Angular Analysis of $B \to K^*\ell^+\ell^-$*" (2017). DOI: 10.1103/PhysRevLett.118.111801.

[18] LHCb Collaboration. "*Test of lepton universality with $B^0 \to K^{*0}\ell^+\ell^-$ decays*" (2017). DOI: 10.1007/JHEP08(2017)055.

[19] LHCb Collaboration. "*Search for lepton-universality violation in $B^+ \to K^+\ell^+\ell^-$ decays*" (2019). DOI: 10.1103/PhysRevLett.122.191801.

[20] Andrew Fowlie. "*CMSSM, naturalness and the "fine-tuning price" of the Very Large Hadron Collider*" (2014). DOI: 10.1103/PhysRevD.90.015010.

[21] "*LHC Design Report Vol.1: The LHC Main Ring*" (2004). Ed. by Oliver S. Bruning et al. DOI: 10.5170/CERN-2004-003-V-1.

[22] Lyndon Evans and Philip Bryant. "*LHC Machine*" (2008). DOI: 10.1088/1748-0221/3/08/s08001. URL: https://doi.org/10.1088/1748-0221/3/08/s08001.

[23] LHCb Collaboration. "*Observation of structure in the J/ψ -pair mass spectrum*" (2020). DOI: 10.1016/j.scib.2020.08.032.

[24] LHCb Collaboration. "*Observation of New Resonances Decaying to $J/\psi K^+$+ and $J/\psi\phi$*" (2021). DOI: 10.1103/PhysRevLett.127.082001.

[25] LHCb Collaboration. "*Observation of an exotic narrow doubly charmed tetraquark*" (2021). arXiv: 2109.01038 [hep-ex].

[26] LHCb Collaboration. "*Observation of $J/\psi p$ Resonances Consistent with Pentaquark States in $\Lambda_b^0 \to J/\psi K^- p$ Decays*" (2015). DOI: 10.1103/PhysRevLett.115.072001.

[27] LHCb Collaboration. "*Observation of a narrow pentaquark state, $P_c(4312)^+$, and of two-peak structure of the $P_c(4450)^+$*" (2019). DOI: 10.1103/PhysRevLett.122.222001.

[28] Ben Gripaios. "*Lectures: From quantum mechanics to the Standard Model*" (2020). arXiv: 2005.06355 [hep-ph].

[29] SNO Collaboration. "*The Sudbury Neutrino Observatory*" (2016). DOI: 10.1016/j.nuclphysb.2016.04.035.

[30] B. Pontecorvo. "*Inverse beta processes and nonconservation of lepton charge*". *Zh. Eksp. Teor. Fiz.* (1957).

[31] Ziro Maki, Masami Nakagawa, and Shoichi Sakata. "*Remarks on the Unified Model of Elementary Particles*". *Progress of Theoretical Physics* (1962). DOI: 10.1143/PTP.28.870. URL: https://doi.org/10.1143/PTP.28.870.

[32] F. del Aguila and Mark J. Bowick. "*The Possibility of New Fermions with ΔI = 0 Mass*" (1983). DOI: 10.1016/0550-3213(83)90316-4.

[33] Paul M. Fishbane, Richard E. Norton, and Michael J. Rivard. "*Experimental Implications of Heavy Isosinglet Quarks and Leptons*" (1986). DOI: 10.1103/PhysRevD.33.2632.

[34] Paul M. Fishbane and Pham Quang Hung. "*Lepton Masses in a Dynamical Model of Family Symmetry*" (1988). DOI: 10.1007/BF01624371.

[35] I. Montvay. "*Three Mirror Pairs of Fermion Families*" (1988). DOI: 10.1016/0370-2693(88)91671-1.

[36] Kazuo Fujikawa. "*A vector-like extension of the standard model*" (1994). DOI: 10.1143/PTP.92.1149.

[37] F. del Aguila et al. "*Vector-like Fermion and Standard Higgs Production at Hadron Colliders*" (1990). DOI: 10.1016/0550-3213(90)90655-W.

[38] F. del Aguila, J. de Blas, and M. Peréz-Victoria. "*Effects of new leptons in electroweak precision data*" (2008). DOI: 10.1103/PhysRevD.78.013010.

[39] Stephen P. Martin. "*Extra vector-like matter and the lightest Higgs scalar boson mass in low-energy supersymmetry*" (2010). DOI: 10.1103/PhysRevD.81.035004.

[40] Peter W. Graham et al. "*A Little Solution to the Little Hierarchy Problem: A Vector-like Generation*" (2010). DOI: 10.1103/PhysRevD.81.055016.

[41] Sibo Zheng. "*Minimal Vectorlike Model in Supersymmetric Unification*" (2020). DOI: 10.1140/epjc/s10052-020-7843-8.

[42] Kyoungchul Kong, Seong Chan Park, and Thomas G. Rizzo. "*A vector-like fourth generation with a discrete symmetry from Split-UED*" (2010). DOI: 10.1007/JHEP07(2010)059.

[43] Gui-Yu Huang, Kyoungchul Kong, and Seong Chan Park. "*Bounds on the Fermion-Bulk Masses in Models with Universal Extra Dimensions*" (2012). DOI: 10.1007/JHEP06(2012)099.

[44] Roman Nevzorov. "E_6 *inspired supersymmetric models with exact custodial symmetry*" (2013). DOI: 10.1103/PhysRevD.87.015029.

[45] Ilja Doršner, Svjetlana Fajfer, and Ivana Mustać. "*Light vector-like fermions in a minimal SU(5) setup*" (2014). DOI: 10.1103/PhysRevD.89.115004.

[46] Aniket Joglekar and Jonathan L. Rosner. "*Searching for signatures of* E_6" (2017). DOI: 10.1103/PhysRevD.96.015026.

[47] Pedro Schwaller, Tim M. P. Tait, and Roberto Vega-Morales. "*Dark Matter and Vectorlike Leptons from Gauged Lepton Number*" (2013). DOI: 10.1103/PhysRevD.88.035001.

[48] James Halverson, Nicholas Orlofsky, and Aaron Pierce. "*Vectorlike Leptons as the Tip of the Dark Matter Iceberg*" (2014). DOI: 10.1103/PhysRevD.90.015002.

[49] Sahar Bahrami et al. "*Dark matter and collider studies in the left-right symmetric model with vectorlike leptons*" (2017). DOI: 10.1103/PhysRevD.95.095024.

[50] Subhaditya Bhattacharya et al. "*Mini Review on Vector-Like Leptonic Dark Matter, Neutrino Mass, and Collider Signatures*" (2019). DOI: 10.3389/fphy.2019.00080.

[51] Kaustubh Agashe, Takemichi Okui, and Raman Sundrum. "*A Common Origin for Neutrino Anarchy and Charged Hierarchies*" (2009). DOI: 10.1103/PhysRevLett.102.101801.

[52] Michele Redi. "*Leptons in Composite MFV*" (2013). DOI: 10.1007/JHEP09(2013)060.

[53] Adam Falkowski, David M. Straub, and Avelino Vicente. "*Vector-like leptons: Higgs decays and collider phenomenology*" (2014). DOI: 10.1007/JHEP05(2014)092.

[54] Nilanjana Kumar and Stephen P. Martin. "*Vectorlike Leptons at the Large Hadron Collider*" (2015). DOI: 10.1103/PhysRevD.92.115018.

[55] Prudhvi N. Bhattiprolu and Stephen P. Martin. "*Prospects for vectorlike leptons at future proton-proton colliders*" (2019). DOI: 10.1103/PhysRevD.100.015033.

[56] Radovan Dermíšek, Aditi Raval, and Seodong Shin. "*Effects of vectorlike leptons on $h \to 4\ell$ and the connection to the muon g-2 anomaly*" (2014). DOI: 10.1103/PhysRevD.90.034023.

[57] Radovan Dermíšek et al. "*Limits on Vectorlike Leptons from Searches for Anomalous Production of Multi-Lepton Events*" (2014). DOI: 10.1007/JHEP12(2014)013.

[58] CMS Collaboration. "*Search for vector-like leptons in multilepton final states in proton-proton collisions at $\sqrt{s}$ = 13 TeV*" (2019). DOI: 10.1103/PhysRevD.100.052003.

[59] L3 Collaboration. "*Search for heavy neutral and charged leptons in e^+e^- annihilation at LEP*" (2001). DOI: 10.1016/S0370-2693(01)01005-X.

[60] Peter Minkowski. "*$\mu \to e\gamma$ at a Rate of One Out of 10^9 Muon Decays?*" (1977). DOI: 10.1016/0370-2693(77)90435-X.

[61] Rabindra N. Mohapatra and Goran Senjanović. "*Neutrino Mass and Spontaneous Parity Nonconservation*" (1980). DOI: 10.1103/PhysRevLett.44.912.

[62] M. Magg and C. Wetterich. "*Neutrino Mass Problem and Gauge Hierarchy*" (1980). DOI: 10.1016/0370-2693(80)90825-4.

[63] Rabindra N. Mohapatra and Goran Senjanović. "*Neutrino Masses and Mixings in Gauge Models with Spontaneous Parity Violation*" (1981). DOI: 10.1103/PhysRevD.23.165.

[64] J. Schechter and J. W. F. Valle. "*Neutrino Masses in SU(2) x U(1) Theories*" (1980). DOI: 10.1103/PhysRevD.22.2227.

[65] J. Schechter and J. W. F. Valle. "*Neutrino Decay and Spontaneous Violation of Lepton Number*" (1982). DOI: 10.1103/PhysRevD.25.774.

[66] R. N. Mohapatra. "*Mechanism for Understanding Small Neutrino Mass in Superstring Theories*" (1986). DOI: 10.1103/PhysRevLett.56.561.

[67] R. N. Mohapatra and J. W. F. Valle. "*Neutrino Mass and Baryon Number Nonconservation in Superstring Models*" (1986). DOI: 10.1103/PhysRevD.34.1642.

[68] Robert Foot et al. "*Seesaw Neutrino Masses Induced by a Triplet of Leptons*" (1989). DOI: 10.1007/BF01415558.

[69] Carla Biggio and Florian Bonnet. "*Implementation of the Type-III Seesaw Model in FeynRules/MadGraph and Prospects for Discovery with Early LHC Data*" (2012). DOI: 10.1140/epjc/s10052-012-1899-z.

[70] Carla Biggio et al. "*Global Bounds on the Type-III Seesaw*" (2020). DOI: 10.1007/JHEP05(2020)022.

[71] Arindam Das and Sanjoy Mandal. "*Bounds on the triplet fermions in Type-III seesaw and implications for collider searches*" (2021). DOI: 10.1016/j.nuclphysb.2021.115374.

[72] A. Abada et al. "*Low energy effects of neutrino masses*" (2007). DOI: 10.1088/1126-6708/2007/12/061.

[73] A. Abada et al. "$\mu \to \mathrm{e}\gamma$ *and* $\tau \to \ell\gamma$ *decays in the fermion triplet seesaw model*" (2008). DOI: 10.1103/PhysRevD.78.033007.

[74] Roberto Franceschini, Thomas Hambye, and Alessandro Strumia. "*Type-III seesaw at LHC*" (2008). DOI: 10.1103/PhysRevD.78.033002.

[75] Yi Cai et al. "*Lepton Number Violation: Seesaw Models and Their Collider Tests*" (2018). DOI: 10.3389/fphy.2018.00040.

[76] Saiyad Ashanujjaman and Kirtiman Ghosh. "*Type-III seesaw: Phenomenological implications of the information lost in decoupling from high-energy to low-energy*" (2021). DOI: 10.1016/j.physletb.2021.136403.

[77] ATLAS Collaboration. "*Search for type-III seesaw heavy leptons in leptonic final states in* pp *collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*" (2021). Submitted to *Eur. Phys. J. C*. arXiv: 2202.02039 [hep-ex].

[78] CMS Collaboration. "*Search for physics beyond the standard model in multilepton final states in proton-proton collisions at* $\sqrt{s} = 13$ *TeV*" (2020). DOI: 10.1007/JHEP03(2020)051.

[79] CMS Collaboration. "*Search for Evidence of the Type-III Seesaw Mechanism in Multilepton Final States in Proton-Proton Collisions at $\sqrt{s}$ = 13 TeV*" (2017). DOI: 10.1103/PhysRevLett.119.221802.

[80] W. Buchmüller, R. Rückl, and D. Wyler. "*Leptoquarks in Lepton-Quark Collisions*" (1987). [Erratum: 10.1016/S0370-2693(99)00014-3]. DOI: 10.1016/0370-2693(87)90637-X.

[81] Jogesh C. Pati and Abdus Salam. "*Lepton Number as the Fourth Color*" (1974). [Erratum: https://doi.org/10.1103/PhysRevD.11.703.2]. DOI: 10.1103/PhysRevD.10.275.

[82] H. Georgi and S. L. Glashow. "*Unity of All Elementary Particle Forces*" (1974). DOI: 10.1103/PhysRevLett.32.438.

[83] Harald Fritzsch and Peter Minkowski. "*Unified Interactions of Leptons and Hadrons*" (1975). DOI: 10.1016/0003-4916(75)90211-0.

[84] Ben Gripaios, Marco Nardecchia, and S. A. Renner. "*Composite leptoquarks and anomalies in B-meson decays*" (2015). DOI: 10.1007/JHEP05(2015)006.

[85] Leandro Da Rold and Federico Lamagna. "*Composite Higgs and leptoquarks from a simple group*" (2019). DOI: 10.1007/JHEP03(2019)135.

[86] Steven Weinberg. "*Supersymmetry at Ordinary Energies. Masses and Conservation Laws*" (1982). DOI: 10.1103/PhysRevD.26.287.

[87] R. Barbier et al. "*R-parity violating supersymmetry*" (2005). DOI: 10.1016/j.physrep.2005.08.006.

[88] Rusa Mandal and Antonio Pich. "*Constraints on scalar leptoquarks from lepton and kaon physics*" (2019). DOI: 10.1007/JHEP12(2019)089.

[89] Bastian Diaz, Martin Schmaltz, and Yi-Ming Zhong. "*The leptoquark Hunter's guide: Pair production*" (2017). DOI: 10.1007/JHEP10(2017)097.

[90] Sacha Davidson and Patrice Verdier. "*Leptoquarks decaying to a top quark and a charged lepton at hadron colliders*" (2011). DOI: 10.1103/PhysRevD.83.115016.

[91] J. K. Mizukoshi, Oscar J. P. Éboli, and M. C. Gonzalez-Garcia. "*Bounds on scalar leptoquarks from Z physics*" (1995). DOI: 10.1016/0550-3213(95)00162-L.

[92] Ezequiel Alvarez et al. "*A composite pNGB leptoquark at the LHC*" (2018). DOI: 10.1007/JHEP12(2018)027.

[93] A. Angelescu et al. "*Closing the window on single leptoquark solutions to the B-physics anomalies*" (2018). DOI: 10.1007/JHEP10(2018)183.

[94] Andreas Crivellin, Dario Müller, and Francesco Saturnino. "*Flavor Phenomenology of the Leptoquark Singlet-Triplet Model*" (). DOI: 10.1007/JHEP06(2020)020.

[95] Shaikh Saad and Anil Thapa. "*Common origin of neutrino masses and $R_{D^{(*)}}$, $R_{K^{(*)}}$ anomalies*" (2020). DOI: 10.1103/PhysRevD.102.015014.

[96] Ulrich Haisch and Giacomo Polesello. "*Resonant third-generation leptoquark signatures at the Large Hadron Collider*" (2021). DOI: 10.1007/JHEP05(2021)057.

[97] ATLAS Collaboration. "*Search for pair production of scalar leptoquarks decaying into first- or second-generation leptons and top quarks in proton-proton collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*" (2021). DOI: 10.1140/epjc/s10052-021-09009-8.

[98] ATLAS Collaboration. "*Search for pair production of third-generation scalar leptoquarks decaying into a top quark and a τ-lepton in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*" (2021). DOI: 10.1007/JHEP06(2021)179.

[99] ATLAS Collaboration. "*Searches for third-generation scalar leptoquarks in $\sqrt{s}$ = 13 TeV pp collisions with the ATLAS detector*" (2019). DOI: 10.1007/JHEP06(2019)144.

[100] CMS Collaboration. "*Search for leptoquarks coupled to third-generation quarks in proton-proton collisions at $\sqrt{s} = 13$ TeV*" (2018). DOI: 10.1103/PhysRevLett.121.241802.

[101] CMS Collaboration. "*Search for third-generation scalar leptoquarks decaying to a top quark and a τ lepton at $\sqrt{s} = 13$ TeV*" (2018). DOI: 10.1140/epjc/s10052-018-6143-z.

[102] CMS Collaboration. "*Search for singly and pair-produced leptoquarks coupling to third-generation fermions in proton-proton collisions at $\sqrt{s} = 13$ TeV*" (2021). DOI: 10.1016/j.physletb.2021.136446.

[103] CMS Collaboration. "*Constraints on models of scalar and vector leptoquarks decaying to a quark and a neutrino at $\sqrt{s} = 13$ TeV*" (2018). DOI: 10.1103/PhysRevD.98.032005.

[104] CMS Collaboration. "*Search for heavy neutrinos and third-generation leptoquarks in hadronic states of two τ leptons and two jets in proton-proton collisions at $\sqrt{s} = 13$ TeV*" (2019). DOI: 10.1007/JHEP03(2019)170.

[105] CMS Collaboration. "*Search for a singly produced third-generation scalar leptoquark decaying to a τ lepton and a bottom quark in proton-proton collisions at $\sqrt{s} = 13$ TeV*" (2018). DOI: 10.1007/JHEP07(2018)115.

[106] *Summary of the cross section measurements of Standard Model processes.* https://twiki.cern.ch/twiki/pub/CMSPublic/PhysicsResultsCombined/. (2022).

[107] ATLAS Collaboration. "*Search for new phenomena in three- or four-lepton events in pp collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*" (2021). DOI: 10.1016/j.physletb.2021.136832.

[108] ATLAS Collaboration. "*Search for supersymmetry in events with four or more charged leptons in 139 fb^{-1} of $\sqrt{s}$ = 13 TeV pp collisions with the ATLAS detector*" (2021). DOI: 10.1007/JHEP07(2021)167.

[109] CMS Collaboration. "*Inclusive nonresonant multilepton probes of new phenomena at $\sqrt{s}$ = 13 TeV*" (2022). Submitted to *Phys. Rev. D.*. arXiv: 2202.08676 [hep-ex].

[110] *CMS Luminosity public results.* https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

[111] *Worldwide LHC Computing Grid.* https://wlcg-public.web.cern.ch/.

[112] CMS Collaboration. "*The CMS Experiment at the CERN LHC*" (2008). DOI: 10.1088/1748-0221/3/08/S08004.

[113] CMS Collaboration. "*CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*" (2006). URL: http://cds.cern.ch/record/922757.

[114] CMS Collaboration. "*CMS technical design report, volume II: Physics performance*" (2007). DOI: 10.1088/0954-3899/34/6/S01.

[115] CMS Collaboration. "*Operation and performance of the CMS tracker*" (2014). Ed. by Pietro Govoni et al. DOI: 10.1088/1748-0221/9/03/C03005.

[116] CMS Collaboration. "*The CMS Phase-1 Pixel Detector Upgrade*" (2021). DOI: 10.1088/1748-0221/16/02/P02027.

[117] CMS Collaboration. "*The CMS electromagnetic calorimeter project*: Technical Design Report" (1997). URL: https://cds.cern.ch/record/349375.

[118] CMS Collaboration. "*Performance of the CMS hadron calorimeter with cosmic ray muons and LHC beam data*" (2010). DOI: 10.1088/1748-0221/5/03/t03012. URL:

https://doi.org/10.1088/1748-0221/5/03/t03012.

[119] CMS Collaboration. "*The CMS muon system*". *9th ICATPP Conference on Astroparticle, Particle, Space Physics, Detectors and Medical Physics Applications*. 2006. DOI: 10.1142/9789812773678_0096.

[120] CMS Collaboration. "*The CMS trigger system*" (2017). DOI: 10.1088/1748-0221/12/01/P01020.

[121] G. Bauer et al. "*The CMS Data Acquisition System Software*" (2010). Ed. by Jan Gruntorad and Milos Lokajicek. DOI: 10.1088/1742-6596/219/2/022011.

[122] Kenneth Bloom. "*CMS software and computing for LHC Run 2*" (2016). DOI: 10.22323/1.282.0185.

[123] Rene Brun et al. *root-project/root: v6.18/02*. Version v6-18-02. 2019. DOI: 10.5281/zenodo.3895860. URL: https://doi.org/10.5281/zenodo.3895860.

[124] CMS Collaboration. "*Mini-AOD: A New Analysis Data Format for CMS*" (2015). DOI: 10.1088/1742-6596/664/7/072052.

[125] J. Alwall et al. "*The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*" (2014). DOI: 10.1007/JHEP07(2014)079.

[126] Torbjörn Sjöstrand et al. "*An Introduction to PYTHIA 8.2*" (2015). DOI: 10.1016/j.cpc.2015.01.024.

[127] Paolo Nason. "*A new method for combining NLO QCD with shower Monte Carlo algorithms*" (2004). DOI: 10.1088/1126-6708/2004/11/040.

[128] Stefano Frixione, Paolo Nason, and Carlo Oleari. "*Matching NLO QCD computations with Parton Shower simulations: the POWHEG method*" (2007). DOI: 10.1088/1126-6708/2007/11/070.

[129] Simone Alioli et al. "*A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*" (2010). DOI: 10.1007/JHEP06(2010)043.

[130] John M. Campbell and R. K. Ellis. "*MCFM for the Tevatron and the LHC*" (2010). DOI: 10.1016/j.nuclphysbps.2010.08.011.

[131] M. Hildreth et al. "*CMS Full Simulation for Run-2*" (2015). DOI: 10.1088/1742-6596/664/7/072022.

[132] Andrea Giammanco. "*The Fast Simulation of the CMS Experiment*" (2014). Ed. by D. L. Groep and D. Bonacorsi. DOI: 10.1088/1742-6596/513/2/022012.

[133] GEANT4 Collaboration. "*GEANT4—a simulation toolkit*" (2003). DOI: 10.1016/S0168-9002(03)01368-8.

[134] CMS Collaboration. "*Particle-flow reconstruction and global event description with the CMS detector*" (2017). DOI: 10.1088/1748-0221/12/10/P10003.

[135] CMS Collaboration. "*Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV*" (2018). DOI: 10.1088/1748-0221/13/06/P06015.

[136] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "*The anti-k_T jet clustering algorithm*" (2008). DOI: 10.1088/1126-6708/2008/04/063.

[137] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "*The catchment area of jets*" (2008). DOI: 10.1088/1126-6708/2008/04/005.

[138] Matteo Cacciari and Gavin P. Salam. "*Pileup subtraction using jet areas*" (2008). DOI: 10.1016/j.physletb.2007.09.077.

[139] CMS Collaboration. "*Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV*" (2017). DOI: 10.1088/1748-0221/12/02/P02014.

[140] CMS Collaboration. "*Performance of reconstruction and identification of τ leptons decaying to hadrons and ν_τ in pp collisions at $\sqrt{s} = 13$ TeV*" (2018). DOI: 10.1088/1748-0221/13/10/P10005.

[141] R. Fruhwirth et al. "*Vertex reconstruction and track bundling at the LEP collider using robust algorithms*" (1996). DOI: 10.1016/0010-4655(96)00040-9.

[142] *Phase1 in FastSim of CMS.*. https://github.com/cms-sw/cmssw/pull/23363. (2018).

[143] *Fixing the seed check in CA iterations of FastSim.* https://github.com/cms-sw/cmssw/pull/25758. (2019).

[144] *FastSim memory fix in TrajectorSeedProducer.* https://github.com/cms-sw/cmssw/pull/25824. (2019).

[145] *New FastSim modifier skipping GEM sequence.* https://github.com/cms-sw/cmssw/pull/23363/commits/1a4a7437afbf144e8fa980c2ff81d81d0e54fc59. (2018).

[146] *Phase1 configurable geometry*. https://github.com/cms-sw/cmssw/pull/23363/commits/4ddf7ca981e3a436906248ec37bdb78f1e0774cb. (2018).

[147] CMS Collaboration. "*Electron and photon reconstruction and identification with the CMS experiment at the CERN LHC*" (2021). DOI: 10.1088/1748-0221/16/05/P05014.

[148] R. Fruhwirth and T. Speer. "*A Gaussian-sum filter for vertex reconstruction*" (2004). Ed. by S. Kawabata and D. Perret-Gallix. DOI: 10.1016/j.nima.2004.07.090.

[149] CMS Collaboration. "*Identification of hadronic tau decay channels using multivariate analysis (MVA decay mode)*" (2020). URL: https://cds.cern.ch/record/2727092.

[150] CMS Collaboration. "*Identification of hadronic tau lepton decays using a deep neural network*" (2022). Submitted to *JINST*. arXiv: 2201.08458 [hep-ex].

[151] CMS Collaboration. *Jet algorithms performance in 13 TeV data*. CMS Physics Analysis Summary CMS-PAS-JME-16-003. (2017). URL: http://cds.cern.ch/record/2256875.

[152] CMS Collaboration. "*Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV*" (2018). DOI: 10.1088/1748-0221/13/05/P05011.

[153] *Methods to apply b-tagging efficiency scale factors*. https://twiki.cern.ch/twiki/bin/viewauth/CMS/BTagSFMethods?rev=35.

[154] *Recommendation for Using b-tag Objects in Physics Analyses*. https://twiki.cern.ch/twiki/bin/view/CMS/BtagRecommendation?rev=28.

[155] CMS Collaboration. "*Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector*" (2019). DOI: 10.1088/1748-0221/14/07/P07004.

[156] Daniele Bertolini et al. "*Pileup Per Particle Identification*" (2014). DOI: 10.1007/JHEP10(2014)059.

[157] CMS Collaboration. "*Search for Third-Generation Scalar Leptoquarks in the $t\tau$ Channel in Proton-Proton Collisions at $\sqrt{s}$ = 8 TeV*" (2015). [Erratum: 10.1007/JHEP11(2016)056]. DOI: 10.1007/JHEP07(2015)042.

[158] H. Voss et al. "*TMVA, the Toolkit for Multivariate Data Analysis with ROOT*". DOI: 10.22323/1.050.0040.

[159] Matthew Feickert and Benjamin Nachman. "*A Living Review of Machine Learning for Particle Physics*" ((2021)). arXiv: 2102.02770 [hep-ph].

[160] Luca Lista. "*Practical Statistics for Particle Physicists*". *2016 European School of High-Energy Physics*. 2017. DOI: 10.23730/CYRSP-2017-005.213.

[161] *Statistical Data Analysis for Particle Physics*. http://www.pp.rhul.ac.uk/~cowan/stat_aachen.html. (2014).

[162] Thomas Junk. "*Confidence level computation for combining searches with small statistics*" (1999). DOI: 10.1016/S0168-9002(99)00498-2.

[163] Alexander L. Read. "*Presentation of search results: The CL_s technique*" (2002). Ed. by M. R. Whalley and L. Lyons. DOI: 10.1088/0954-3899/28/10/313.

[164] Glen Cowan et al. "*Asymptotic formulae for likelihood-based tests of new physics*" (2011). [Erratum: 10.1140/epjc/s10052-013-2501-z]. DOI: 10.1140/epjc/s10052-011-1554-0.

[165] ATLAS and CMS Collaborations, and LHC Higgs Combination Group. *Procedure for the LHC Higgs boson search combination in Summer 2011*. Tech. rep. CMS-NOTE-2011-005, ATL-PHYS-PUB-2011-011. CERN, (2011). URL: http://cds.cern.ch/record/1379837.

[166] Roger Barlow and Christine Beeston. "*Fitting using finite Monte Carlo samples*" (1993). DOI: https://doi.org/10.1016/0010-4655(93)90005-W. URL: https://www.sciencedirect.com/science/article/pii/001046559390005W.

[167] J. S. Conway. "*Incorporating Nuisance Parameters in Likelihoods for Multisource Spectra*". *PHYSTAT 2011*. 2011. DOI: 10.5170/CERN-2011-006.115. arXiv: 1103.0354 [physics.data-an].

[168] *HEPData record for arXiv:2202.08676*. 10.17182/hepdata.110691. (2022).

The work described in this thesis forms part of the following publications:

- **Inclusive nonresonant multilepton probes of new phenomena at $\sqrt{s} = 13$ TeV.** The CMS Collaboration, arXiv:2202.08676 [hep-ex] (2022).
- **Search for new physics in multilepton final states in pp collisions at $\sqrt{s} = 13$ TeV.** The CMS Collaboration, JHEP **03**, 51 (2020).
- **Search for vector-like leptons in multilepton final states in proton-proton collisions at $\sqrt{s}$ = 13 TeV.** The CMS Collaboration, Phys. Rev. D **100** 052003 (2019).

The full list of publications can be obtained from this URL, courtesy of InspireHEP.